DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE NANTES

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

Rémi VALLÉE

Apprentissage profond pour l'aide au diagnostic et comparaison des mécanismes d'explicabilité avec l'attention visuelle humaine : application à la détection de la maladie de Crohn

Thèse présentée et soutenue à Nantes, le 18/05/2022
Unité de recherche : LS2N, Laboratoire des Sciences du Numérique de Nantes, équipe IPI

Composition du Jury :

Président et rapporteur :	Michel DOJAT	Professeur des universités, Grenoble Institut des Neurosciences
Rapporteure :	Isabelle BLOCH	Professeure des universités, Sorbonne Université
Examinateur :	Clément CHATELAIN	Maître de conférences, INSA Rouen
Dir. de thèse :	Harold MOUCHÈRE	Professeur des universités, Nantes Université
Co-dir. de thèse :	Nicolas NORMAND	Professeur des universités, Nantes Université
Co-encadrant :	Antoine COUTROT	Chargé de recherche, CNRS, Université de Lyon

Invité(s) :

Arnaud BOURREILLE	Professeur des universités - Praticien hospitalier, CHU de Nantes

Remerciements

Ces trois dernières années ont été les plus denses en émotions de toute ma vie. J'ai partagé, rires, pleurs, sourires, tristesse, mélancolie et joies avec de nombreuses personnes qui sont chères à mon cœur.

J'aimerais tout d'abord remercier mes 3 *senseis* qui m'ont encadré avec beaucoup de bienveillance. Merci beaucoup à toi Harold. Tu as été mon maître de stage avant d'être mon directeur de thèse, merci d'avoir été toujours à l'écoute, de m'avoir reboosté quand j'en avais besoin et de m'avoir tant de fois fait confiance. Antoine, j'aimerais te remercier pour ta gentillesse et pour la qualité de tes conseils scientifiques et méthodologiques (malgré nos différends en terme de goûts musicaux). Nicolas, merci pour ta convivialité et ta rigueur qui m'ont beaucoup apporté et aussi pour les moments partagés en conférence.

Merci aux IPIs pour cette ambiance d'équipe si chaleureuse, pour toutes les trouspinettes, repas de Noël, mots croisés et pour toutes nos discussions qui m'ont fait si plaisir.

Merci à Arnaud et Astrid avec qui nous avons bien collaboré pendant ces trois années. Merci pour leur confiance, leur temps et leur amitié. Merci également aux gastro-entérologues des CHU de Nantes, Angers et Cholet pour avoir accepté de participer à l'expérience d'eye-tracking.

J'aimerais dire un grand merci à tous mes amis qui m'ont soutenu et encouragé pendant ces trois années. À ma base, aux Guys, à nos si belles amitiés qui durent pour certaines depuis plusieurs décennies. À Alex, Quentin, Quentin, Pierre, Pierre, Kiki, Colin et Hélène. Aux Nantais, aux anciens, avec qui j'ai tant ri, tant échangé, tant de fois refait le monde, tant partagé. À Quentin, Hugo, Manu, Julie, Brewal et Meggane. Aux nouveaux, avec qui j'ai traversé beaucoup de choses, pour toutes les fois où on s'est serré les coudes. À Arthur, Étienne, Julie, Thibaut, Tom, Victor et à Louise. Aux Toulousains, avec qui j'ai terminé ma course, pour leurs sourires, leur bienveillance, pour cette petite vie de famille qui m'a réchauffé le cœur. À Gaby, Brieuc, Lise, Manon, Alix et Onyx.

Et surtout, merci à ma famille. Merci du fond du coeur. Merci pour tout. Merci pour tout votre amour. Merci d'avoir toujours été là. Merci à mon papa de m'avoir aidé à faire

jusqu'à mon dernier devoir. Merci à ma maman de m'avoir appris à écouter les autres, et pour m'avoir tant de fois écouté. Merci Gaby pour ta joie si musicale. Merci Yann d'avoir grandi à mes côtés. Merci à nous 5 pour notre union qui n'a pas grand chose à envier aux doigts de la main.

TABLE DES MATIÈRES

Introduction

Définitions et contexte

Le début du troisième millénaire est marqué par le développement d'une nouvelle technologie appelée intelligence artificielle par les organes médiatiques, et apprentissage profond par les scientifiques. Grâce à ces algorithmes, de nombreuses tâches considérées jusqu'alors comme irréalisables par la machine sont devenues possibles et progressivement, l'apprentissage profond a permis l'automatisation de tâches nécessitant la capacité d'analyse cognitive humaine. Dans ce manuscrit, nous limiterons l'appellation « Intelligence Artificielle ». En effet, de notre point de vue, ces algorithmes ne sont pas « intelligents » du fait qu'ils n'apprennent pas par eux-mêmes, mais plutôt à partir d'exemples choisis par le concepteur de l'algorithme. Le terme intelligence est lui-même difficile à définir, et s'attaquer à sa définition relève davantage d'un travail de réflexion philosophique que de la démarche technique proposée dans ce manuscrit. Nous utiliserons plutôt le terme d'apprentissage profond pour désigner ces algorithmes qui, grâce à des phases d'optimisation successives impliquant des millions de calculs unitaires, apprennent à extraire les caractéristiques les plus pertinentes pour résoudre une tâche clairement définie. Ils sont à notre sens d'excellents copieurs, permettant de répliquer avec précision des comportements d'analyse cognitifs humains, quand le nombre d'exemples utilisés pour les entraîner est adapté à la complexité de la tâche. Ils sont capables d'appréhender un environnement complexe à partir de stimuli comme les vidéos, les images et le son.

Cet environnement est si riche et complexe que notre cerveau humain ne peut l'analyser dans sa globalité. Afin d'optimiser son appréciation et conduire aux meilleures prises de décision, nous utilisons un processus appelé attention, filtrant ainsi les informations les plus pertinentes pour leur allouer un maximum de ressources cognitives. Ce mécanisme, pouvant être aussi bien conscientisé qu'inconscient, a été étudié depuis plus d'un siècle chez l'humain ce qui a permis une meilleure compréhension de nos prises de décision. Inspirés par ce mécanisme extrêmement efficient, les chercheurs en apprentissage profond ont essayé depuis la dernière décennie de mettre ce principe attentionnel en application pour optimiser et comprendre les réseaux de neurones artificiels. L'objectif de ce mimé-

tisme est à la fois de rendre les algorithmes plus efficients, mais également de rendre leur comportement plus compréhensible. L'objectif d'une partie des travaux de recherches qui seront présentés ici est d'étudier le parallèle possible entre le traitement cognitif humain de l'information visuelle et le traitement de l'information grâce aux réseaux de neurones profonds.

Dans ce manuscrit, nous nous concentrerons sur un cas applicatif concret, la détection des lésions de la maladie de Crohn dans des images issues de vidéos capsules endoscopiques. La maladie de Crohn est une maladie chronique provoquant des lésions sur les muqueuses de l'intestin grêle. Afin de la distinguer d'autres pathologies intestinales, il est nécessaire de comptabiliser ainsi que d'évaluer la sévérité de ces lésions. Au début des années 2000, le développement de la vidéo capsule endoscopique a permis l'examen complet de l'intestin grêle [1] et l'amélioration du diagnostic de la maladie de Crohn. Le volume conséquent de données produites par ces vidéos, ainsi que le temps d'analyse associé à leur visionnage, a naturellement poussé les chercheurs en vision par ordinateur et les gastro entérologues à collaborer. Ces travaux s'inscrivent donc dans le cadre de cette collaboration, dont le but est d'aider à l'élaboration d'outils permettant à ces experts médicaux d'accélérer et de fiabiliser leur diagnostic.

Motivations

Ces travaux de recherches s'inscrivent dans une dynamique bien spécifique. Les algorithmes d'apprentissage profond et le traitement cognitif humain sont régis par des processus bien trop nombreux et trop complexes pour être appréciés dans leur globalité. Des travaux antérieurs ont déjà montré quelques parallèles pouvant être faits entre l'un et l'autre. L'objectif de cette thèse est de creuser ce parallèle en se concentrant sur les principes attentionnels de chacun d'entre eux. En les comparant, en améliorant la compréhension de l'un et de l'autre, nous espérons ouvrir une voie de recherche nous permettant d'améliorer la compréhension des mécanismes cognitifs humains et de ceux des réseaux de neurones profonds. En apprentissage profond, la course aux performances commence peu à peu à ralentir pour laisser place à une course à l'explicabilité, enjeu majeur de par le fait que l'on souhaite utiliser ces derniers dans des secteurs critiques impliquant des vies humaines. Au travers l'établissement de ce parallèle entre attention artificielle et attention humaine, l'objectif serait de développer des outils pour aider les humains dans l'apprentissage de tâches spécifiques et également de s'inspirer du mécanisme décisionnel humain

afin d'améliorer les performances et l'explicabilité des réseaux de neurones artificiels.

La dimension inter-disciplinaire de ces travaux de recherche est également centrale. Il nous semble que créer des passerelles entre le domaine médical et le domaine de la recherche en informatique est un enjeu majeur de notre temps. Les algorithmes décisionnels sont de plus en plus présents dans le diagnostic médical et il est important que les chercheurs en informatique intègrent les problématiques de la routine clinique. Le dialogue avec les médecins est indispensable pour permettre le développement de solutions qui leurs seront « réellement » utiles hors des conditions expérimentales. Il est, à notre sens, également essentiel que les professions médicales comprennent mieux ces outils de diagnostic afin de pouvoir apprécier leurs possibilités et leurs limites. Les algorithmes d'apprentissage profond ne sont pas des « baguettes magiques » permettant de résoudre tous les problèmes, mais bien des outils qu'il convient de maîtriser avant de les utiliser.

Plan

Ce manuscrit s'organise en quatre chapitres décrits ci-après. Le choix a été fait de répartir l'état de l'art dans les différents chapitres afin de permettre au lecteur de suivre le fil des différentes réflexions qui nous ont animé pendant ces trois années.

Chapitre 1

Le premier chapitre présente le contexte applicatif de la recherche de cette thèse : l'aide au diagnostic de la maladie de Crohn. Premièrement, les enjeux de l'aide au diagnostic assisté par ordinateur, et plus particulièrement ceux de la maladie de Crohn, seront présentés. La revue des différentes méthodes existantes développées au cours du temps montrera l'évolution progressive vers des techniques d'apprentissage supervisé, et plus particulièrement vers des méthodes utilisant des réseaux de neurones profonds. Le constat sera fait que chacune des méthodes présentées utilise des bases d'entraînement et de test privées ne permettant pas une bonne comparaison des différentes méthodes, malgré les performances encourageantes de chacune d'entre elles. Nous introduirons donc CrohnIPI, notre base de données publique réalisée dans le cadre de ces travaux en partenariat avec le CHU de Nantes. Elle permet l'évaluation des différentes méthodes de détection des lésions de l'intestin grêle. Les différentes étapes de création du jeu de données seront clairement détaillées de façon à rendre le processus de création le plus transparent possible afin de

limiter les biais d'apprentissage. Des résultats de réseaux de neurones à convolution de l'état de l'art seront présentés afin de servir de base de référence aux autres méthodes, et l'effet de la qualité d'annotation sur les performances de ces réseaux sera étudié.

Chapitre 2

Le deuxième chapitre dépeint l'état de l'art en introduisant les concepts d'attention humaine et artificielle. Le chapitre s'ouvre sur la question d'interprétabilité des algorithmes d'apprentissage profond. Les différentes raisons nous poussant à les considérer comme des boites noires seront définies ainsi que les axes d'amélioration permettant de les rendre plus explicables. Avant de présenter le concept d'attention artificielle rendant les réseaux de neurones profonds plus transparents et leurs décisions plus interprétables, nous présenterons l'attention humaine. C'est de cette dernière dont s'inspire l'attention artificielle et elle est étudiée depuis le XIX^{e} siècle. Les notions d'attention montante et descendante seront définies afin de mieux comprendre le processus attentionnel mesuré par l'expérience oculométrique présentée dans le chapitre 4. Par ailleurs, des applications de l'étude du comportement attentionnel humain seront présentés afin de permettre au lecteur de comprendre que l'explicabilité de nos processus cognitifs est un enjeu important dans le développement des technologies liées au multimédia et à l'imagerie médicale. La quantification des comportements attentionnels étant relative, nous définirons les métriques de comparaison utilisées tout au long du manuscrit, ainsi que les modèles linéaires mixtes qui permettront d'évaluer la significativité de nos mesures. Dans une dernière section, sera présenté le concept d'attention artificielle. L'attention artificielle est découpée en deux grandes familles d'algorithmes, ceux à attention apprise, et ceux à attention post-hoc. Les algorithmes à attention apprise impliquent l'apprentissage d'un processus attentionnel lors de l'entraînement de l'algorithme. Ils répondent au besoin de rendre les algorithmes plus transparents. Les algorithmes à attention post-hoc, eux, interviennent sur des réseaux de neurones profonds déjà entraînés. Ils permettent de visualiser les zones ayant permis à l'algorithme de faire sa prédiction. Ces derniers répondent au besoin d'avoir des informations supplémentaires à la prédiction afin de permettre à l'utilisateur de faire confiance à l'algorithme.

Chapitre 3

Dans le troisième chapitre, nous verrons comment ces algorithmes ont été adaptés à notre problématique de détection des lésions de la maladie de Crohn. Dans un premier temps, il sera présenté un algorithme à attention apprise que nous avons réalisé. Ce réseau de neurones récurrents apprend par renforcement à extraire les parties de l'image qui permettent de réaliser les meilleures prédictions. À la différence des réseaux de neurones à convolution classiques, il n'observe pas l'entièreté de l'image mais seulement les parties d'intérêt, occultant le reste de l'image. Les performances de ce réseau seront comparés aux réseaux à convolution plus classiques présentés dans le premier chapitre. Dans un second temps, les résultats d'attention des méthodes d'extraction de l'attention post-hoc seront présentés. Nous verrons comment, d'une méthode à l'autre, les zones définies comme ayant permis de prendre la décision varient. Une étude de la stabilité de ces méthodes en fonction des conditions d'entraînement sera réalisée. Elle montrera que les résultats obtenus par les réseaux sont plus stables pour la méthode des gradients, et que pour n'importe quelle méthode, appliquée à n'importe quel réseau, les résultats sont plus stables sur les images pathologiques que sur les images non-pathologiques. Finalement, ces résultats seront discutés, et il sera soulevé la nécessité d'avoir une vérité terrain permettant de comparer les résultats obtenus par l'attention artificielle.

Chapitre 4

Dans ce dernier chapitre sera présentée une comparaison entre l'attention humaine et l'attention artificielle. Dans un premier temps, nous établirons un état de l'art de la comparaison entre ce mécanisme cognitif humain et son implémentation pour les réseaux de neurones profonds. Afin de pouvoir réaliser cette comparaison, nous avons créé un jeu de données oculométriques impliquant 22 participants, de niveaux d'expertises différents. Ce jeu de données sera donc décrit, ainsi que le concept de placer les humains dans des conditions expérimentales aussi proches que possible de celles d'un réseau de neurones réalisant des prédictions sur des images issues de vidéos capsules endoscopiques. Les différentes méthodes d'attention artificielle présentées dans le chapitre 3 seront alors comparées avec ce jeu de données oculométriques. Au travers de différents tests statistiques, nous évaluerons alors l'influence du niveau d'expertise des participants sur cette comparaison et nous prendrons également en compte le label des images. Nous étudierons également l'évolution de cette corrélation au fur et à mesure de l'entraînement des réseaux de neurones

profonds. Finalement, nous discuterons des différences de résultats obtenus pour chacune des méthodes d'extraction de l'attention post-hoc, et proposerons l'utilisation de notre jeu de données oculométriques pour évaluer les méthodes d'explicabilité.

Liste des publications

Cette thèse a donné lieu à la publication de plusieurs articles avec acte :

— ORASIS 2019 : Réseau de neurones récurrent à attention pour la détection de lésions intestinales. (Présentation orale [2], HAL)

— MMSP 2019 : Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. (Poster [3], HAL)

— SPIE 2020 : CrohnIPI : An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms. (Poster [4], HAL)

— Endoscopy international 2021 : Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network. (Journal [5], HAL)

— MICCAI 2022 : Influence of training and expertise on Deep Neural Network and human attention during a medical image classification task. (En cours de publication)

Et également sans acte :

— Folle journées de l'imagerie Nantaise 2019 : Réseau de neurones récurrent à attention pour la détection de lésions intestinales. (Poster, introduction orale)

— United European Gastroenterology week 2019 : Crohn's disease lesion detection by small-bowel capsule endoscopy : an automatic deep learning method. (Abstract, Best abstract presentation price [6])

— ECVP 2021 : Influence of expertise on human and machine visual attention in a medical image classification task. (Présentation orale [7], HAL)

— Journée Bio-informatique 2022 : Influence de l'expertise sur l'attention visuelle humaine et machine dans une tâche de classification d'images médicales. (présentation orale)

Chapitre 1

L'aide au diagnostic médical de la maladie de Crohn

Sommaire

Avec le développement des algorithmes de vision par ordinateur, on peut observer une demande croissante d'outil d'aide au diagnostic médical afin de pouvoir permettre aux experts médicaux de se focaliser sur le traitement des maladies, où leur expertise et leur expérience sont indispensables. Pour les chercheurs en vision par ordinateur, travailler main dans la main avec des experts médicaux est une chance. Premièrement, cela leur permet de se confronter à des problèmes réels, et ainsi d'évaluer plus facilement la pertinence de leurs algorithmes aujourd'hui de plus en plus transposables d'une tâche à une autre. Deuxièmement, les tâches d'analyse médicale sont complexes et exigeantes. Complexes, de par le faible volume des jeux de données mis à disposition du fait de la rareté des experts capables d'annoter les données. Exigeantes, de par l'importance des décisions impliquant des vies humaines. Face à ces contraintes, les chercheurs en vision par ordinateur sont obligés de repousser les limites de l'état de l'art pour résoudre ces problèmes en développant de nouvelles formes d'algorithmes afin d'obtenir de meilleures performances, et en comprenant mieux le fonctionnement de ceux existants pour leur faire confiance.

La maladie de Crohn est une maladie chronique inflammatoire de l'intestin grêle. Le diagnostic de cette maladie est complexe puisqu'il repose sur la comptabilisation et l'évaluation de la sévérité des lésions qu'elle produit sur les muqueuses de l'intestin grêle. Afin d'accélérer et fiabiliser son diagnostic, un partenariat a été mis en place avec l'Institut des Maladies de l'Appareil Digestif (IMAD) et l'équipe Image Perception Interaction (IPI) du Laboratoire des Sciences du Numérique de Nantes (LS2N). Cette association est doublement bénéfique, car elle permet le développement de nouveaux outils utiles aux gastro-entérologues dans leur pratique clinique. Elle permet également de fournir un cadre applicatif concret à la recherche en vision par ordinateur.

Cette section s'organisera de la façon suivante : premièrement, le concept d'Aide au Diagnostic Assisté par Ordinateur (ADAO) sera présenté, son évolution pendant le dernier siècle et plus particulièrement l'importance des données dans l'élaboration de telles techniques. Deuxièmement, il sera présenté le cas particulier du diagnostic de la maladie de Crohn, ses enjeux, les différentes techniques d'aide au diagnostic existantes et les jeux de données associés à ces dernières. Troisièmement, une fois l'accent porté sur l'importance

des jeux de données permettant l'entraînement et l'évaluation des méthodes d'ADAO, le jeu de données CrohnIPI créé avec le CHU de Nantes sera alors présenté ainsi que les différentes étapes associées à son développement. Finalement, dans une quatrième partie, l'influence de la qualité des annotations sur les performances de réseaux de neurones profonds à convolution de l'état de l'art sera évaluée.

1.1 L'aide au diagnostic assisté par ordinateur

1.1.1 Définitions et contexte

Le diagnostic est un concept de raisonnement abstrait qui constitue la base des directives cliniques, de la pratique fondée sur des preuves et des interventions en matière de soins. Maitland 2010 en propose la définition suivante : c'est « le processus de détermination des mécanismes par lesquels l'état de santé du patient se produit et les conclusions qui en découlent »[8].

Il se déroule en plusieurs étapes. Tout d'abord, un patient rencontre un problème de santé. C'est alors, la plupart du temps, le premier à considérer les symptômes et peut choisir à ce moment-là de consulter afin d'obtenir l'avis d'un expert médical. Une fois le patient pris en charge, un processus itératif de collecte d'informations est mis en place, ainsi qu'une interprétation de ces dernières. Au travers de l'anamnèse[1], de l'entretien clinique et de l'examen physique, les cliniciens accumulent des informations pouvant être utiles pour comprendre le problème de santé du patient. Les approches de collecte d'informations peuvent être utilisées à différents moments. Le processus continu de collecte, d'intégration et d'interprétation des informations implique la formulation d'hypothèses et la mise à jour des probabilités antérieures au fur et à mesure de l'acquisition d'informations [9].

Afin de répondre à ce besoin de données variées permettant d'assurer un diagnostic de plus en plus précis, de nombreuses technologies d'imagerie médicale ont vu le jour. Ces techniques non invasives se sont très vite répandues et font aujourd'hui partie intégrante du quotidien des médecins. Elles permettent de révéler des pathologies, de les observer et ainsi de mieux comprendre les symptômes des patients. On peut citer les techniques les plus connues telles que la radiologie, la tomographie, l'endoscopie, la thermographie, la photographie médicale et la microscopie [10].

1. Ensemble des renseignements fournis au médecin par le malade ou par son entourage sur l'histoire d'une maladie ou les circonstances qui l'ont précédée.

Bien que toutes ces nouvelles informations à disposition des experts médicaux permettent d'affiner leur diagnostic, elles restent inutiles tant qu'elles n'ont pas été correctement interprétées. C'est afin de permettre la fiabilisation et l'accélération de cette étape d'interprétation, très chronophage pour les praticiens, que des outils informatisés d'aide au diagnostic médical ont commencé à être développés [11].

1.1.2 Le XX^{è} siècle et le début de l'ADAO

La recherche en analyse informatisée des images médicales a commencé dans les années 1960 avec plusieurs tentatives dans le domaine de la radiologie [12, 13, 14, 15, 16, 17]. C'est dans les années 1980, avec un changement de concept que la recherche dans ce domaine prend de l'ampleur. L'idée de créer des algorithmes capables de réaliser le diagnostic informatiquement s'estompe pour laisser place à la recherche d'outils informatisés pour assister les experts médicaux dans leur pratique clinique : c'est le début du diagnostic assisté par ordinateur (DAO) [18]. Les premiers systèmes développés se basent alors sur des algorithmes séquentiels réalisant des opérations à l'échelle des pixels (détection de lignes, de bordures, de couleurs ou de textures par exemple). Grâce à des modèles mathématiques, ils réalisent l'association entre les caractéristiques de bas niveau extraites et un ensemble de règles bien définies, permettent la résolution de nombreux problèmes spécifiques. On peut par exemple trouver un algorithme de détection de calcification dans des mammographies, qui, malgré des performances faibles (spécificité $< 30\%$), permettait l'amélioration des performances des radiologistes [19] dès la fin des années 1990.

Le début des années 1990 est marqué par l'essor des techniques dites supervisées. À la différence des algorithmes précédemment présentés, ces derniers apprennent à réaliser des prédictions à partir d'exemples annotés. On peut trouver des exemples d'application pour la mesure des formes des vertèbres (diagnostic de l'ostéoporose) [20], pour la réalisation d'un atlas 3D du cerveau (planification de neurochirurgies) [21], ou encore pour la segmentation d'IRM [22]. Bien qu'apprenant à partir d'exemples, ils nécessitaient l'expertise humaine pour la définition des caractéristiques à extraire afin de réaliser leur prédiction. C'est afin d'optimiser cette étape d'extraction de caractéristiques que la recherche s'est orientée vers les réseaux de neurones à convolution (RNC), permettant l'extraction des caractéristiques de bas et de haut niveau de façon autonome. On peut trouver les premiers essais de RNC appliqués à l'imagerie médicale en 1995 pour la détection de nodules pulmonaires [23]. Ils ne connaîtront leur réel premier succès qu'avec LeNet5 [24] en 1998

pour la reconnaissance des chiffres manuscrits et seront ensuite largement démocratisés et appliqués au domaine médical.

1.1.3 Le XXI^{è} siècle et l'essor de l'apprentissage profond

L'apprentissage automatique s'est très rapidement développé depuis le début du XXI^{è} siècle. L'idée derrière ces algorithmes est de donner aux ordinateurs la capacité de résoudre des problèmes en tirant des enseignements de leurs expériences. Les algorithmes sont «nourris» avec des données et conçus pour généraliser leurs connaissances afin de réaliser des prédictions sur des données qui leur sont inconnues. L'apprentissage profond, une sous-catégorie de l'apprentissage automatique, est basé sur des réseaux neuronaux artificiels dans lesquels plusieurs couches de traitement sont utilisées pour extraire des données des caractéristiques de niveau de plus en plus élevé. Ces derniers ont été capables de dépasser les performances humaines dans certaines tâches, comme les jeux vidéo Atari [25], divers jeux de société, dont le jeu de go et les échecs [26] ou encore la caractérisation de bactéries précises [27].

La puissance de ces modèles mathématiques n'est à l'heure actuelle plus à démontrer, tant il est communément admis qu'elles surpassent les performances des anciens algorithmes. Assez naturellement, les problèmes d'ADAO se sont orientés vers eux. Le secteur médical a été un des premiers secteurs d'application, de par le nombre potentiel de vies que cette technologie peut sauver. On trouve de nos jours des algorithmes surpassant les performances des experts médicaux dans quelques tâches précises comme la classification des cancers de la peau [28, 29], le diagnostic du diabète [30] ou encore dans l'analyse de radiographies [31]. On peut également trouver des centaines d'autres applications permettant d'assister les médecins dans leur pratique clinique [11, 32].

De nombreux problèmes cependant nécessitent encore l'investissement des efforts des communautés scientifiques médicales et d'analyse d'images par ordinateur. La limitation principale de ce genre d'algorithmes est la quantité de données qu'ils nécessitent pour s'entraîner avant de parvenir à des performances permettant leur utilisation dans la pratique clinique des experts. La question de la qualité des données est, elle aussi, centrale. En effet, afin que l'algorithme puisse généraliser correctement ses connaissances et réalise des prédictions précises sur des cas inconnus, il est nécessaire qu'il soit testé dans des conditions proches du réel, en évitant tout biais d'apprentissage.

1.2 Le diagnostic de la maladie de Crohn

1.2.1 La maladie de Crohn

La maladie de Crohn est une maladie chronique intestinale. Les maladies chroniques, globalement responsables de 63% des décès et première cause de mortalité dans le monde selon l'OMS [33], sont des maladies de longue durée évoluant lentement. Cette pathologie peut provoquer des lésions dans les muqueuses de l'intestin grêle, pouvant alors nécessiter une intervention chirurgicale. Bien que le signe déterminant pour établir le diagnostic de la maladie de Crohn soit la présence de ces dernières, ce n'est pas la seule pathologie qui provoque de telles lésions dans l'intestin grêle. Son diagnostic repose sur le calcul de deux scores : *The Capsule Endoscopy Crohn's Disease Activity Index* (CECDAI) [34] et le score de Lewis [35], qui nécessitent d'établir le nombre de lésions présentes dans le tube digestif ainsi que leur sévérité. La sévérité est définie en fonction du type des lésions. Ces deux scores dont le calcul est détaillé ci-dessous prennent en compte la sévérité, le type, le nombre et la localisation des lésions présentes dans l'intestin grêle du patient.

CECDAI

The Capsule Endoscopy Crohn's Disease Activity Index (CECDAI) [34], est calculé à l'aide des 3 scores suivants pour les différents segments de l'intestin grêle (proximal et distal)

- A : Score d'inflammation, évalué sur une échelle de 0 (aucune) à 5, (gros ulcère, >2 cm)
- B : Score d'étendue de la maladie, évalué sur une échelle de 0 (aucune) à 3 (diffuse)
- C : Score de sténose, noté de 0 (aucune) à 3 (obstruction)

Une fois les différents scores calculés pour les deux segments de l'intestin grêle le CEDCAI score est calculé avec la formule présentée dans l'équation 1.1, l'indice 1 et 2 représentant respectivement les scores obtenus pour le segment proximal et distal. La limite entre l'intestin proximal et distal est établi en utilisant le temps de transit de l'intestin grêle.

$$CEDCAI = (A_1 \times B_1 + C_1) + (A_2 \times B_2 + C_2) \tag{1.1}$$

Score de Lewis

Pour le calcul du score de Lewis [35], un $Score_t$ (voir équation 1.2) est associé à chacun des tertiles t de l'intestin. Pour l'aspect des villosités[1] V (normal ou œdémateuse) et les ulcères U est associé un score de paramètre p de la lésion, ainsi qu'un score d'étendue e et un score de taille s.

$$Score_t = (V_t^p \times V_t^e \times V_t^s) + (U_t^p \times U_t^e \times U_t^s) \tag{1.2}$$

Le score maximum obtenu pour un tertile t de l'intestin est conservé et on y ajoute un score pour de sténose S prenant en compte le nombre n de sténoses, si la capsule a réussi à la traverser tr et si la sténose est ulcérée u. On obtient ainsi le score de Lewis 1.3.

$$Score_{Lewis} = \max_t Score_t + (S^n \times S^u \times S^{tr}) \tag{1.3}$$

1.2.2 Développement de la vidéo capsule endoscopique et application à la maladie de Crohn

En 2000, les avancées dans la miniaturisation des systèmes électroniques rendent enfin possible le développement de la vidéo capsule endoscopique [1] (voir figure 1.1). D'un point de vue clinique, ce système permet de répondre à de nombreux besoins. Jusqu'à lors, il était impossible d'observer la majeure partie de l'intestin grêle avec l'endoscopie traditionnelle. Il était seulement possible d'observer l'intestin grêle proximal et distal, soit seulement 10% de l'entiéreté de l'intestin, induisant une faible spécificité et sensitivité pour les tâches de détection des lésions [36]. De plus, les technologies dites traditionnelles comme l'endoscopie par fibre optique impliquent une sédation du patient, et sont très invasives, pouvant impliquer des effets secondaires tels que la perforation de la paroi intestinale, des difficultés à respirer et des infections [37].

Lors de la digestion du patient, la capsule endoscopique émet, grâce à une transmission radio ultrahaute fréquence, le flux vidéo à des antennes accrochées au corps du patient. Les images sont stockées sur un enregistreur portable accroché à la ceinture du patient. L'autonomie est d'environ 5 heures, ce qui permet l'exploration complète de l'intestin grêle. L'avantage également d'une telle technologie est qu'elle n'astreint pas le patient à

1. Les villosités intestinales sont des structures de l'intestin grêle permettant d'amplifier la surface d'échange entre l'intestin et le sang. Elles donnent un aspect "poilu" à l'intestin. Pour le calcul du score de Lewis, on quantifie si ces villosités sont anormalement gonflées (œdémateuse) ou non.

(a) Image de la capsule endoscopique

(b) Image du boîtier à la ceinture du patient

FIGURE 1.1 – Image d'une capsule vidéo endoscopique (a), image issue de [38]. Elle capte des images à une vitesse de 2 à 6 images la seconde en fonction de la vitesse de la capsule lors de son déplacement dans l'intestin grêle. Elle mesure environ 1cm de diamètre afin de pouvoir être ingérée facilement. Les images capturées par la capsule sont ensuite transmises par radiofréquence à des antennes positionnées sur le corps du patient (b), image issue de [39].

rester à l'hôpital pendant la durée de l'examen et lui permet de poursuivre son activité quotidienne.

C'est assez naturellement que l'on a voulu intégrer cette nouvelle technologie à la pratique clinique des gastro-entérologues, pour aider au diagnostic des maladies intestinales. Diverses études ont démontré sa supériorité pour le diagnostic de la maladie de Crohn, face à d'autres technologies comme l'entérographie[1] par imagerie à résonance magnétique [40, 41, 42] ou l'entérographie par tomographie [43, 41], qui, bien qu'également non-invasives, ne permettent pas un aussi bon diagnostic que l'analyse des images enregistrées par la vidéo capsule endoscopique.

Bien que la vidéo capsule ait fait ses preuves en termes de performance et que son acceptation dans le domaine médical soit prouvée [44], son usage n'est pas encore généralisé pour le diagnostic de la maladie de Crohn. Une des raisons principales est le temps d'analyse induit par cet examen. En effet, chaque examen de vidéo endoscopique génère entre 50000 et 60000 images. En moyenne, le temps de relecture d'une vidéo est estimé entre 30 et 60 min [45]. L'effort cognitif que nécessite la revue des images endoscopiques peut induire un risque de manquer des lésions en raison de a fatigue du praticien en charge de leur interprétation.

1. La réalisation d'un enregistrement d'images délimitant l'activité musculaire.

1.2.3 Premiers essais d'aide au diagnostic de la maladie de Crohn

Du fait de la nature chronophage de la tâche de relecture des vidéos issues de capsules endoscopiques, et de par les récentes avancées en matière d'analyse d'images par ordinateur, de nombreuses équipes de recherche ont cherché à relever le défi de la détection des lésions issues de la maladie de Crohn. Vers la fin des années 2000, les premiers algorithmes basés sur des règles sont élaborés pour résoudre cette tâche. On peut trouver plusieurs exemples de méthodes basées sur des décisions réalisées à l'aide de machines à vecteurs de support (SVM) à partir de caractéristiques extraites dans l'image. Plusieurs de ces méthodes sont regroupées dans le tableau 1.1 ainsi que les jeux de données utilisés pour leur entraînement.

Référence	**Lésions étudiées**	**Classifieur**	**Nombre d'images (P/NP)**
Bejakovic 2009 [46]	Lésions de la MC	SVM	8605 (1725/6880)
Karargyris 2009 [47]	Ulcères	MLP	50 (20/30)
Girgis 2010 [48]	Lésions de la MC	SVM	999 (474/525)
Haji-Maghsoudi 2012 [49]	Diverses lésions	Filtre de Canny	206 (206/0)
Chen 2012[50]	Ulcères	SVM	272 (108/164)
Charisis 2012[51]	Ulcères	DAC+MLP+SVM	174 (87/87)
Jebarani 2013 [52]	Érosions et ulcères	SVM	1828 (100/1728)
Eid 2013 [53]	Ulcères	SVM	260 (130/130)
Szczypiski 2014 [54]	Ulcères, saignements	SVM	613(113/500)
Jinn-Yi 2014 [55]	Ulcères, saignements	SVM	896 (410/486)
Iakovidis 2014 [56]	Diverses lésions	MLP+SVM	1370 (1370/0)
Yuan 2015 [57]	Ulcères	SVM	240 (170/170)
Charisis 2016 [58]	Ulcères	SVM	800 (400/400)
Liaqat 2018 [59]	Ulcères, saignements	Diverses	442 (187/255)
Souaidi 2019 [60]	Ulcères	SVM	452 (213/239)

TABLE 1.1 – Tableau récapitulatif des premières méthodes n'utilisant pas d'apprentissage profond pour la détection de maladies intestinales. Chacune des méthodes présentées dans ce tableau repose sur l'extraction de caractéristiques propres à l'image avant l'utilisation d'un classifieur. Chaque méthode est associée au nombre d'images de la base de données ayant permis leur entraînement et leur test. Les performances obtenues par chacune des méthodes n'ont pas été synthétisées volontairement dans ce tableau, de par l'absence d'une évaluation commune permettant d'obtenir des résultats comparables. « P » correspond aux nombre d'images pathologiques et « NP » au nombre d'images non pathologiques.

Comme le montre le tableau 1.1, chacune des méthodes précédemment citées utilise des bases de données internes privées, ne permettant pas une bonne comparaison des méthodes entre elles. Les modalités d'annotations sont pour la grande majorité inconnues ou peu détaillées. Les résultats, peu comparables entre eux du fait de l'absence de base de données publique, atteignent jusqu'à 98% de précision [59].

1.2.4 Apprentissage profond et vidéo capsule endoscopique

C'est assez logiquement que les algorithmes d'apprentissage profond ont été confrontés au problème de détection des pathologies dans des images issues de vidéos capsules endoscopiques. On peut trouver dans le tableau 1.2 des exemples d'algorithmes utilisés pour des tâches de détection d'ulcères et de lésions inflammatoires.

Référence	**Lésions étudiés**	**Classifieur**	**Nombre d'images (P/NP)**
Georgakopoulos 2016 [61]	Lésions inflammatoire	RNC	876 (227/599)
Fan 2018 [62]	Ulcères et érosions	RNC (Alexnet)	21160 (8160/13000)
Aoki 2019 [63]	Ulcères et érosions	RNC	15800 (5800/10000)
Haya 2019 [64]	Ulcères	RNC (Alexnet et googLeNet)	1878(1525/250)
Barash 2021 [65]	Ulcères	RNC (ResNet)	17 640(7391/10249)

TABLE 1.2 – Tableau récapitulatif des méthodes utilisant des réseaux de neurones profonds pour détecter des maladies intestinales à partir d'images issues de VCE (RNC : Réseau de neurones à convolution).

L'efficacité de ces algorithmes basés sur des réseaux de neurones profonds est bien supérieure aux méthodes dites traditionnelles. Bien que les performances semblent bonnes, et validées sur une quantité bien plus importante, le problème de l'absence de base publique reste d'actualité. On peut remarquer que chacune des méthodes a été testée et entraînée sur un jeu de données différent. Ce manque de jeu de données permettant une évaluation commune des algorithmes empêche une comparaison égalitaire des modèles entre eux. Pour la plupart des méthodes présentées précédemment, les phases de création des jeux de données, et les jeux de données eux-mêmes ne sont pas présentés. Cela implique un potentiel risque de biais qui ne permet pas de garantir la capacité de généralisation des algorithmes à des cas inconnus et donc l'utilisation de ces derniers dans la pratique clinique des gastro-entérologues.

1.3 Étapes et enjeux de la création de jeux de données médicaux

Le développement des algorithmes d'apprentissage supervisé pose le problème de la collecte des données d'apprentissage. En effet, pour maximiser la justesse des prédictions, ces algorithmes doivent pouvoir apprendre sur un jeu de données suffisamment large, avec des données correctement annotées et représentatives du problème. De plus, dans les images médicales, à la différence des images naturelles, la phase d'annotation requiert l'expertise de spécialistes du domaine, ce qui complexifie la phase d'annotation. On peut également identifier une difficulté supplémentaire pour les tâches de détection d'images pathologiques. Les images ne contenant pas de pathologies sont souvent bien plus faciles à collecter que les images dites pathologiques présentes chez un nombre de patients beaucoup plus faible.

Le principal problème de l'entraînement de ces algorithmes réside en la petite taille des jeux de données disponibles, et le manque de diversité dans ces derniers, qui entravent la capacité de généralisation des réseaux de neurones. Les jeux de données disponibles publiquement sont souvent de taille limitée et manquent d'explication quant aux processus d'annotation. Dans cette partie seront présentées les différentes étapes nécessaires à la création d'un jeu de données médicales d'après Willemink et al. 2020 [66].

1.3.1 Préparation des données

La première étape de la création d'un jeu de données est la préparation des données. L'utilisation de données pour la recherche informatique nécessite de les sortir de leur contexte clinique et requiert l'approbation du conseil éthique du centre hospitalier, ainsi que celle des patients. Les données médicales sont des données sensibles et protégées par les lois liées au secret médical. Ainsi, avant toute utilisation dans la recherche en intelligence artificielle, ces données doivent être anonymisées, et stockées en toute sécurité. Afin de garantir leur utilité pour l'apprentissage, la structure des données doit également être homogénéisée.

De plus, la quantité de paramètres entraînables des réseaux de neurones à convolution est dépendante de la taille des images d'entrée. Plus le nombre de paramètres du réseau est important, plus les ressources en calcul demandées ainsi que le temps d'inférence

seront importants. Ainsi de permettre l'utilisation des données en entrée d'un réseau de neurones profond il est nécessaire de les redimensionner. La résolution des images en entrées des réseaux de neurones est souvent inférieure à 300 × 300 pixels afin de limiter la complexité de ces derniers. Cela peut induire certaines contraintes pour les prédictions à partir d'images haute résolution [66].

1.3.2 Annotation des données

Une fois la préparation réalisée, il est nécessaire de choisir les labels appropriés pour définir la vérité terrain associée à chaque image. Cela impose au concepteur d'avoir une idée bien conçue de l'application future de l'algorithme. Cela nécessite de trouver un bon compromis entre des labels suffisamment discriminants et suffisamment précis pour être utilisés dans le processus clinique. Une fois les labels définis, la phase d'annotation peut commencer. On peut identifier deux grandes familles de techniques d'annotation. La technique prospective où les labels sont associés aux images pendant l'examen clinique, ce qui nécessite des outils permettant à la fois d'annoter et de diagnostiquer, et la technique rétrospective, la plus couramment utilisée. Elle est plus simple à mettre en place et consiste à étudier les précédents diagnostics réalisés afin de constituer un jeu de données.

En ce qui concerne les techniques rétrospectives, outre l'annotation manuelle réalisée par des humains à partir des rapports préalablement établis lors du diagnostic, on trouve également des exemples d'applications [67, 68, 69] de réseaux récurrents permettant le traitement automatique du texte contenu dans ces rapports, l'analysant afin d'établir la vérité terrain. Le temps des experts en médecine étant limité, la collecte de données pose un problème fondamental. Vaut-il mieux un large jeu de données provenant d'un seul expert, et dont la qualité du diagnostic ne peut être vérifiée ? Ou, vaut-il mieux un jeu de données plus réduit, où les annotations ont été soigneusement apposées sur les images et revues par plusieurs experts afin de minimiser le nombre de faux négatifs et de faux positifs ? [70]

1.4 CrohnIPI : une base publique de données d'entraînement et d'évaluation

Comme il est décrit dans les parties 1.2.3 et 1.2.4, les algorithmes permettant l'analyse issue de VCE sont entraînés et évalués sur des bases privées et indépendantes. Afin de

développer notre propre solution d'aide au diagnostic de la maladie de Crohn assistée par ordinateur, nous avons créé notre propre base de données. Au vu du récent intérêt de la communauté scientifique pour cette tâche, nous avons souhaité rendre publique la base de données créée. En rendant cette base de données publique, nous espérons pouvoir fournir aux chercheurs un moyen d'évaluer et de comparer leurs algorithmes. Cela implique de mettre à disposition des utilisateurs des données de qualité, afin de leur permettre d'entraîner et tester ces derniers sur une base de données sans biais. Cette base de données est disponible sur `http://crohnipi.ls2n.fr/`. Dans cette partie, il sera décrit tout son processus de création.

1.4.1 Description générale du jeu de données

Le jeu de données CrohnIPI est un jeu de données multicentrique approuvé par la Commission nationale informatique et libertés et par le groupe nantais d'éthique dans le domaine de la santé, le comité d'éthique de Nantes, France. Trois unités endoscopiques françaises ont participé à l'ensemble des données. Quatre lecteurs ont été impliqués dans l'annotation des images : une interne en gastro-entérologie et trois experts. Le premier lecteur, expérimenté en endoscopie conventionnelle, avait été formé à la lecture des vidéos de capsules pour les besoins de l'étude et était responsable de la sélection des images. Les trois experts avaient plus de 10 ans d'expérience dans la prise en charge des patients atteints de maladies inflammatoires de l'intestin et avaient lu plus de 200 vidéos capsules endoscopiques chez des patients atteints de la maladie de Crohn. Trois tours d'annotation ont été effectués afin d'obtenir une annotation consensuelle aussi proche que possible de la « vérité ».

1.4.2 Acquisition des images

Les images sont issues de vidéos capsules endoscopiques de troisième génération (système Pillcam SB3, Medtronic, Minnesota, États-Unis) acquises entre 2014 et 2018 auprès de patients atteints de la maladie de Crohn ou ayant subi une exploration pour suspicion de la maladie de Crohn. Elles ont été enregistrées dans les trois unités d'endoscopie participantes, ont été recueillies rétrospectivement et anonymisées. Seules les vidéos de patients atteints de la maladie de Crohn et présentant des signes de présence de lésions de l'intestin grêle ont été sélectionnées. Les images fixes successives, au format JPEG, ont été extraites des vidéos (sans perte de qualité), permettant leur annotation. Les données

cliniques et démographiques au moment de la capsule ont été enregistrées.

1.4.3 Sept différentes classes

Les lésions pathologiques ont été définies comme présentées dans le tableau 1.8 [71]. On peut en identifier six différents types, que l'on peut classer par sévérité de la plus sévère à la moins sévère comme suit : sténose, ulcération > 10 mm, ulcération entre 3 et 10 mm, ulcération aphtoïde, œdème, érythème [35]. En plus de ces six classes pathologiques, une classe non pathologique est également présente de façon à permettre aux algorithmes supervisés de faire la distinction entre images saines et images pathologiques. La proportion d'images pathologiques et d'images non pathologiques est proche, de façon à faciliter la généralisation des réseaux de neurones profonds. Cette proportion n'est pas représentative de la réalité, où les images non pathologiques sont bien plus nombreuses.

1.4.4 Outil d'annotation

Afin de permettre les différentes phases d'annotations, nous avons travaillé en partenariat avec l'équipe médicale sur le développement d'un outil d'annotation. Une capture d'écran de l'application est visible figure 1.2. Cette interface utilisateur permet dans un premier temps de charger des vidéos au format GVF (format en sortie de l'application utilisateur de la Pillcam SB3). Une fois les vidéos chargées, il est possible de parcourir les images à l'aide de la souris ou des touches directionnelles pour procéder à l'annotation grâce à des boutons représentant chacune des sept classes présentées dans le tableau 1.8. La possibilité de répondre « Non concluante » était proposée pour signifier que, soit une image est peu lisible, et donc très difficilement annotable, soit un doute entre plusieurs classes est présent. Les images classées comme « Non concluante » impliquent la relecture de l'image par des tiers. Cette application produit en sortie un fichier JSON récapitulant les annotations par vidéo et par annotateur. Bien que cette application ait été utilisée en local lors des différentes phases d'annotation présentées dans la section 1.4.5, cette dernière a pour ambition d'évoluer afin de pouvoir être utilisée dans la pratique clinique des experts médicaux. De nouvelles caractéristiques sont d'ores et déjà présentes, comme l'utilisation en ligne ou l'évaluation des images par un réseau de neurones profond.

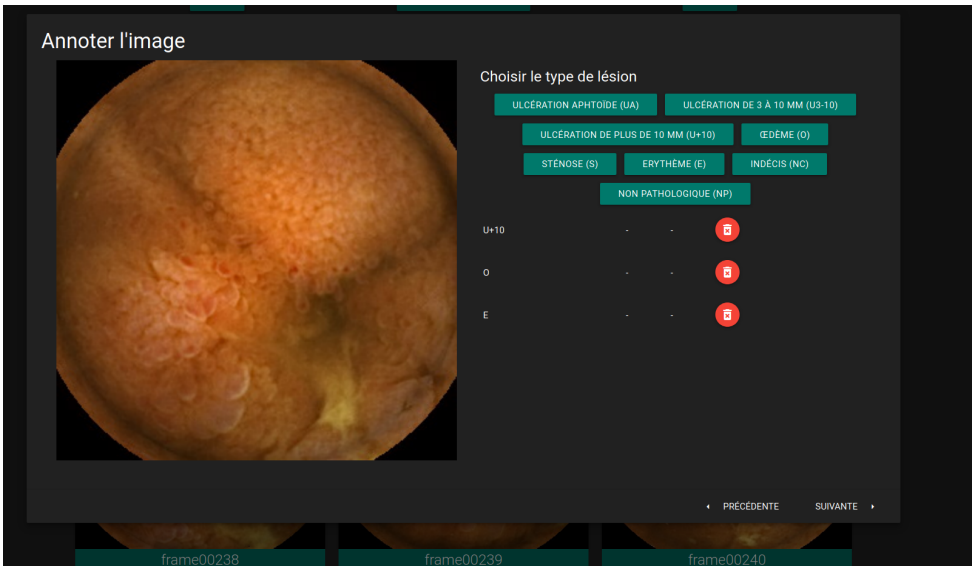


FIGURE 1.2 – Capture d'écran de l'application d'annotation

1.4.5 Phases d'annotation

L'annotation s'est déroulée en trois phases illustrées dans la figure 1.3 et décrites ci-dessous.

Première étape : sélection des images

Les images ont été sélectionnées et annotées par le lecteur initial, Astrid de Maissin interne en gastro-entérologie. Toutes les images sélectionnées par le lecteur, pathologiques ou non, ont été considérées comme des images d'intérêt. Chaque image d'intérêt a été extraite et incluse dans le jeu de données CrohnIPI. Lorsqu'une image contenait plus d'une lésion, toutes les lésions ont été annotées. Les images présentant une lésion douteuse (type ou présence) ont été étiquetées comme «Non concluantes». Les images pathologiques ont été sélectionnées indépendamment de la position de la lésion dans le cadre.

La même proportion d'images de contrôle non pathologiques, ne contenant aucune lésion, a été extraite du même ensemble de vidéos. Les images de contrôle ont été sélectionnées de manière aléatoire et non successive, indépendamment de la présence de bulles, de résidus ou de liquide trouble, afin d'éviter tout biais d'apprentissage, sur des images trop parfaites et donc loin de la réalité de la routine clinique.

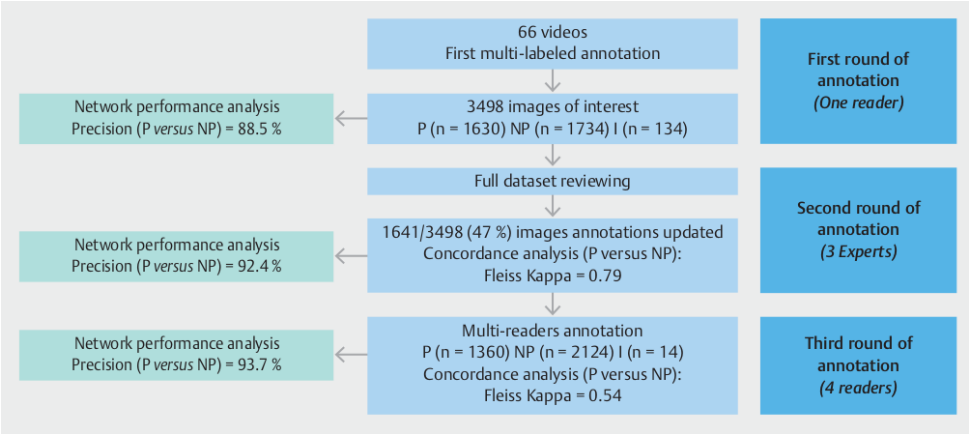


FIGURE 1.3 – Le jeu de données final a été obtenu après la sélection des images non pathologiques (NP) et pathologiques (P) d'intérêt extraites de 66 vidéos capsules réalisées chez des patients atteints de maladie de Crohn par un premier lecteur. Toutes les images ont été revues et annotées par trois experts. Les images discordantes ont été relues par les quatre gastro-entérologues pour obtenir une annotation consensuelle. Les images non concluantes ont été exclues de l'ensemble de données. La performance du réseau neuronal a été testée à chaque étape du processus ainsi que la concordance entre les lecteurs.

Deuxième étape : annotations indépendantes

Toutes les images sélectionnées par le lecteur initial, non pathologiques et pathologiques, ont été revues par trois experts en maladies chroniques intestinales et en capsule endoscopique Arnaud Boureille, Caroline Trang et Mathurin Flamant. Toutes les images ont été attribuées à chaque lecteur dans un ordre aléatoire. Les experts étaient aveugles aux annotations des autres experts et n'avaient accès qu'aux images fixes. Les mêmes définitions des lésions ont été utilisées pour annoter les images. Lorsque les images contenaient plus d'une lésion, toutes les lésions étaient annotées par les lecteurs. À la fin du processus, les annotations obtenues du lecteur initial et des experts ont été codées pour ne garder qu'une seule annotation par image. Si plusieurs lésions étaient annotées sur une même image, seule la plus sévère était retenue. Ensuite, l'analyse de la concordance entre les experts a été calculée et les images discordantes entre les quatre lecteurs ont été identifiées et enregistrées pour une analyse consensuelle.

Troisième étape : consensus

Toutes les images avec une annotation discordante ont été revues. À ce stade, une seule lésion par image a été étiquetée, la plus grave. Les quatre lecteurs se sont réunis en trois sessions pour obtenir une annotation consensuelle des images, considérée comme la « vérité » pour chaque image. Chacun des lecteurs devait d'abord donner son avis, sur 20 images successives. Pour éviter qu'un lecteur n'influence les autres, j'ai animé les débats, aidé à la répartition de la parole sans prendre part à la lecture des images. Si nécessaire, une courte séquence vidéo adjacente pouvait être récupérée, comprenant 10 images en amont et 10 images en aval de l'image à annoter. À la fin de ce processus, les images discordantes étaient classées comme non concluantes et étaient exclues de l'ensemble des données pour une analyse ultérieure.

1.4.6 Analyse statistique

L'accord inter-observateur pour la classification des images dans l'ensemble de données CrohnIPI a été évalué à l'aide de la *p-value* de l'accord inter-observateur kappa de Fleiss [72]. La valeur du kappa varie de -1 à 1, la valeur 0 indiquant l'indépendance statistique et la valeur 1 indiquant un accord parfait entre les observateurs. Un kappa de Fleiss compris entre 0,41 et 0,60 peut être interprété comme un accord modéré, entre 0,61 et 0,80 comme un accord substantiel et au-dessus de 0,81 comme un accord quasi parfait.

Soixante-six vidéos contenant au moins une image pathologique obtenues chez 63 patients atteints de la maladie de Crohn ont été incluses dans l'ensemble de données. Les caractéristiques cliniques et démographiques des patients sont détaillées dans le tableau 1.3. Six patients ont été explorés pour une suspicion de maladie de Crohn, et à la fin du bilan, ont eu un diagnostic définitif de la maladie de Crohn basé sur les résultats de l'endoscopie conventionnelle et de l'histologie ; les lésions de l'intestin grêle détectées par la capsule ont été considérées comme des lésions de la maladie de Crohn. À partir des vidéos, 3498 images ont été extraites et annotées par le premier lecteur. Selon cette première lecture, 1630 images (46,6%) contenaient au moins une lésion, 1734 (49,6%) ont été considérées comme non pathologiques, et 134 n'étaient pas concluantes (3,8%).

Après la deuxième série d'annotations, lors de la distinction entre les images pathologiques et non pathologiques, 537 images (15%) ont été étiquetées différemment par au moins un expert parmi trois. Parmi les images, 2345 (67%) ont été codées au moins une fois comme non pathologiques, 1614 (46%) comme pathologiques et 94 (2%) comme

Caractéristiques des patients	
Patients, n	63
Femmes, n (%)	37 (59)
Age, médiane (EI)	39 (28-49)
Durée de la maladie en années, médiane (EI)	9,7 (2-16)
Raison du diagnostic, n (%)	
Diagnostic initial	6 (10)
Suivi de routine	45 (71)
Maladie non contrôlée	12 (19)

TABLE 1.3 – Caractéristiques des patients et raisons du diagnostic par vidéo capsule endoscopique. Ici EI signifie Écart interquartile.

non concluantes. Le tableau 1.4 présente l'accord entre les experts selon trois modalités différentes de classification des images : non pathologique versus pathologique ; non pathologique versus sténose ou tous types d'ulcérations ou œdème et érythème ; et non pathologique versus chaque type de lésion. L'accord entre les experts était bon pour distinguer les images pathologiques et non pathologiques avec un coefficient kappa de 0,79 ($p < 0{,}0001$). Avec un codage intermédiaire des lésions, la concordance globale inter observatrice était substantielle ($k = 0{,}68$, $p < 0{,}0001$). Avec le codage le plus fin des lésions, la concordance globale inter observatrice était modérée ($k = 0{,}57$, $P < 0{,}0001$).

En considérant les annotations du lecteur initial et des trois experts, 1641 des 3498 images (47 %) ont été annotées différemment par au moins l'un d'entre eux, ce qui correspond à un accord modéré (kappa de Fleiss = 0,54 ; $P < 0{,}0001$). Les 1641 images ont été revues au cours du troisième tour d'annotation afin d'obtenir une annotation consensuelle aussi proche que possible de la « vérité ». À la fin du processus, 2124 des 3498 images ont été considérées comme non pathologiques (60,7 %), 1360 pathologiques (38,9 %) et 14 (0,4 %) non concluantes.

1.5 Influence de la qualité d'annotation sur les performances des algorithmes d'aide au diagnostic

Ici, nous émettons l'hypothèse que la qualité d'annotation permet la réduction du nombre de faux positifs dans la base d'entraînement. Nous pensons que cela devrait permettre aux réseaux neuronaux entraînés sur une telle base de données de mieux généraliser leurs connaissances à de nouveaux cas. Dans cette partie, nous vérifierons cette hypothèse

		Images classées comme pathologiques ou non pathologiques	Images classées comme non pathologiques ou contenant une sténose ou des ulcérations ou un œdème/érythème	Images classées comme non pathologiques ou contenant tout type de lésions
		Rapport entre les images concordantes et les images totales N/N (%)		
Non-pathologic		1827/2345 (78)	1827/2345 (78)	1827/2345 (78)
Pathologique	S	1134/1614 (70)	80/323 (25)	80/323 (25)
	U>10		658/1300 (51)	74/369 (20)
	U3-10			117/850 (14)
	UA			103/555 (19)
	O		39/406 (10)	16/250 (6)
	E			10/197 (5)
Non concluante		0/94 (0)	0/94 (0)	0/94 (0)
Total		2961/3498 (85)	2604/3498 (74)	2227/3498 (64)
Coefficient de Fleiss Kappa		0,79	0,68	0,57

TABLE 1.4 – Proportion d'images classées de manière identique par les experts par rapport au nombre d'images avec désaccord, selon le type de lésions. Trois niveaux d'étiquettes sont considérés avec leur accord inter observateur respectif (coefficient Kappa de Fleiss)

en observant les résultats obtenus par 3 réseaux de neurones profonds de l'état de l'art sur un même nombre d'images avec les étiquettes correspondant aux différentes phases d'annotation.

1.5.1 Méthode

Afin d'observer l'effet de la qualité d'annotation sur les performances des réseaux de neurones, nous avons sélectionné trois réseaux de neurones convolutifs profonds de l'état de l'art : ResNet34 [73], VGG16 et VGG19[74]. Ces modèles neuronaux ont connu un franc succès dans des tâches de classification d'images naturelles et sont utilisés dans de nombreuses applications. Afin d'entraîner les réseaux, dans l'optique d'évaluer l'influence de la qualité d'annotation, nous avons construit 4 jeux de données avec les mêmes images étiquetées selon les différentes phases présentées dans la partie 1.4.5. Chacun des sous-jeux de données est composé de 3 sous-jeux de données servant respectivement à réaliser l'entraînement, la validation et le test du réseau. Pour chacune des bases, 70% est alloué à l'entraînement, 10% à la validation et 20% pour le test. Les trois premiers jeux de données

sont constitués des images recueillies lors des trois phases d'annotations décrites dans la partie 1.4.5. Dans ces jeux de données, les labels pathologiques ont été regroupés entre eux afin de constituer une seule et unique classe pathologique, les autres appartenant à la classe non pathologique. Dans le dernier jeu de données, les labels pathologiques n'ont pas été regroupés, et correspondent à la lésion la plus sévère décrite par les gastro-entérologues, comme défini dans la partie 1.4.5. Pour chacun des jeux de données, les images contenant l'étiquette « Non concluante » (3,83 %) ont été exclues de chacun des jeux de données présentés précédemment. Exclure les images « Non concluantes » nous permet de garantir une comparaison équitable entre les différents tests en conservant des jeux de données contenant le même nombre d'images, l'influence de la quantité de données sur les performances des réseaux de neurones profonds n'étant, elle, plus à démontrer. Lors de la phase d'entraînement, les images ont subi de façon aléatoire 3 types de rotation différents, permettant une augmentation de la taille du jeu de données par 8 (2^3). Chacun des entraînements/tests a été réalisé 5 fois en validation croisée, de sorte que chacune des images ait été vue une unique fois en phase de test (*K-Fold cross-validation*).

1.5.2 Résultats

Dans cette partie sont présentés les résultats associés à chacune des expériences précédemment décrites dans les tableaux 1.5,1.6 et 1.7.

	Précision	Spécificité	Sensitivité	F1 score
Phase 1 (Étiquettes du sélecteurs)	93.58	94.35	92.76	93.33
Phase 2 (Étiquettes des experts indépendant)	93,30	93,63	**92,79**	91,54
Phase 3 (Étiquettes du consensus)	94,26	96,19	91,25	92,54
Phase 3 multi-labels (Étiquettes du consensus)	**94.58**	**97.98**	89.26	**92.78**

TABLE 1.5 – Résultats des classifications N/NP (2 classes) ou multi-label (7 classes) obtenues avec ResNet34 [73], sur les quatre versions d'étiquettes.

On peut observer que les meilleures décisions sont prises à partir des réseaux entraînés sur les jeux de données de la phase 3. On obtient en moyenne pour les trois réseaux 92,73% de précision pour la phase 1, 93,41% pour la phase 2, 94,33% pour la phase 3 et 94,34% pour la phase 3 multi-label. On peut ainsi observer que, plus les réseaux ont été entraînés à partir d'images soigneusement annotées, plus les performances des réseaux sont bonnes. Il est rappelé ici que chacun des modèles a été entraîné, validé et testé

	Précision	Spécificité	Sensitivité	F1 score
Phase 1	92.60	93.54	**91.60**	92.30
Phase 2	93,93	95,80	91,02	92,14
Phase 3	**94,40**	96,42	91,25	**92,72**
Phase 3 multi-labels	94.17	**97.93**	88.31	92.21

TABLE 1.6 – Résultats des classifications N/NP (2 classes) ou multi-label (7 classes) obtenues avec VGG16 [74], sur les quatre versions d'étiquettes.

	Précision	Spécificité	Sensitivité	F1 score
Phase 1	92.03	92.79	91.23	91.73
Phase 2	93,01	95,80	88,67	90,84
Phase 3	**94,35**	95,20	**93,01**	**92,78**
Phase 3 multi-labels	94.29	**97.98**	88.53	92.37

TABLE 1.7 – Résultats des classifications N/NP (2 classes) ou multi-label (7 classes) obtenues avec VGG19 [74], sur les quatre versions d'étiquettes.

avec le strict même nombre d'images, présentées dans le même ordre et ayant subi les mêmes transformations aléatoires. Ces résultats peuvent être attribués à une diminution du nombre de faux négatifs et de faux positifs dans la base d'apprentissage, permettant aux réseaux de mieux généraliser leurs connaissances à de nouveaux cas. Les résultats de la phase 3 multi-labels ont été obtenus en entraînant les réseaux avec le label de la lésion la plus sévère contenue dans l'image. Afin de permettre une comparaison équitable avec les autres modèles, les prédictions ont été regroupées *a posteriori* en deux classes : pathologique et non-pathologique. La précision du réseau pour la classification de la lésion la plus sévère contenue dans les images est de 87,54% pour ResNet34, 86,85% pour VGG16 et de 86,71% pour VGG19. La meilleure performance est obtenue avec le réseau ResNet34 sur la base de données multi-annotation, avec une précision de 94,56% et une AUC (Area Under ROC Curve) de 98,23%.

1.6 Conclusion

Le jeu de données CrohnIPI est un jeu de données spécifique et dédié à la maladie de Crohn, composé d'images pathologiques et non pathologiques soigneusement examinées par plusieurs experts afin d'obtenir une annotation consensuelle aussi proche que possible

de la « vérité » pour l'entraînement, la validation et le test des outils de diagnostic assisté par ordinateur. Après trois séries d'annotations, l'ensemble de données contient 3498 images bien codées avec une grande variété de lésions de la muqueuse de la maladie de Crohn et des images non pathologiques choisies indépendamment de la qualité de la préparation de l'intestin, reflétant aussi fidèlement que possible les conditions réelles. Les performances des réseaux sont très bonnes, avec une précision atteignant 94%. De plus, nous avons démontré que les performances de ces derniers augmentent lorsque l'algorithme est testé sur le jeu de données multi-expert plutôt que sur le jeu de données annoté par le premier lecteur, soulignant l'importance majeure d'une annotation de haute qualité.

La principale force de cette étude provient du processus d'annotation multi-lecteurs décrit. Le processus que nous avons utilisé pour créer l'ensemble de données CrohnIPI a corroboré que les lecteurs étaient d'accord pour la classification binaire des images pathologiques et non pathologiques. Cependant, l'accord concernant les différents types de lésions de l'intestin grêle fréquemment observées chez les patients atteints de la maladie de Crohn était plus faible. Cette divergence n'a pas été réduite par le fait que tous les lecteurs ont appris l'endoscopie par capsule dans la même unité d'endoscopie et par l'utilisation d'une définition standardisée de chaque lésion. Une partie du désaccord pourrait être due à l'absence de formation commune avant l'annotation des images sélectionnées par le premier lecteur. Les divergences concernant l'étiquetage de l'érythème et de l'œdème pourraient également s'expliquer par la grande variabilité et la non-spécificité de ces lésions. Ceci a été mis en évidence par la difficulté de sélectionner des images typiques pour l'établissement d'un consensus sur la nomenclature et la description des lésions muqueuses de l'intestin grêle de la maladie de Crohn. La difficulté à classer ces lésions a eu un impact négatif sur les performances de notre réseau neuronal et souligne la nécessité de disposer d'un jeu de données plus rigoureusement développé pour entraîner correctement les systèmes d'apprentissage profond.

Les autres points forts de cette étude proviennent de l'utilisation d'images natives sur lesquelles aucune transformation et aucun filtre n'ont été appliqués, ce qui nous a permis de nous couvrir des biais colorimétriques et, contrairement à d'autres jeux de données, d'utiliser toutes les lésions typiques de la maladie de Crohn sans se limiter aux ulcérations. De plus, les images non pathologiques ont été extraites des mêmes vidéos que les images pathologiques et les images n'ont pas été sélectionnées en fonction de la qualité de la préparation, de la présence ou non de bulles, de résidus, ou de la luminosité, afin de simuler au mieux la pratique clinique réelle. De plus le jeu de données CrohnIPI

contient autant d'images normales que pathologiques, alors que dans les jeux de données plus importants, la plupart des images sont non pathologiques, même si la tâche assignée au réseau est de détecter les images contenant au moins une lésion.

Pour améliorer les performances des modèles, un plus grand nombre d'images peut être nécessaire ou une stratégie d'apprentissage différente (par exemple, l'échantillonnage ou la pondération des exemples difficiles). Des solutions de grands ensembles de données ont été explorées dans d'autres études pour la détection des érosions et des ulcérations avec des ensembles de données de plus de 15 000 images, atteignant des taux de précision proches de 90% [63, 62]. En revanche, aucune donnée sur les œdèmes et les érythèmes n'a encore été publiée.

Une autre limite de l'étude provient de la source unique des images, à savoir la Pillcam SB3, empêchant la généralisation de cet outil de CAO à d'autres appareils. Malgré la conception multicentrique de l'étude, l'apprentissage de l'interprétation des images issues de vidéos capsules endoscopiques a été réalisé dans la même unité clinique, ce qui pourrait introduire un biais par rapport aux études complètement multicentriques. Le développement futur des algorithmes doit être testé sur différentes sources d'images pour être utilisé dans la pratique clinique.

L'ensemble de données CrohnIPI a été construit pour être partagé gratuitement avec la communauté scientifique afin de faciliter et d'accélérer le développement de tels outils, qui seront également accessibles aux gastro-entérologues à l'avenir. La base de données a déjà été partagée avec plusieurs scientifiques et permet déjà à ces derniers de travailler sur l'élaboration d'outils d'annotation automatique. L'ensemble de données CrohnIPI peut être téléchargé, à la demande, à l'adresse `http://crohnipi.ls2n.fr/`. L'ensemble de données est destiné uniquement à la recherche et est protégé par la licence Creative Commons CC BY-NC-ND[1]. Il sera enrichi au fil du temps en incluant des images pathologiques et non pathologiques représentant tous les types de lésions. Un nouveau processus d'annotation sera testé en sélectionnant les images d'intérêt à partir des résultats fournis par les réseaux entraînés avec une validation *a posteriori* par un groupe d'experts en capsule endoscopique et maladie chronique intestinale. Ce processus devrait faciliter l'enrichissement du jeu de données en limitant le nombre nécessaire d'images analysées par les experts. À moyen terme, l'enrichissement du jeu de données devrait permettre de classer chaque type de lésion et non plus seulement comme pathologique ou non pathologique.

Ce chapitre a fait l'objet de deux publications, une dans un journal médical [5] et une

1. `https://creativecommons.org/licenses/by-nc-nd/2.0/`

dans une conférence d'informatique orientée imagerie médicale [4].

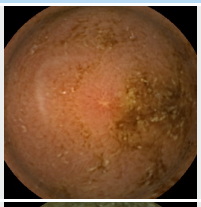
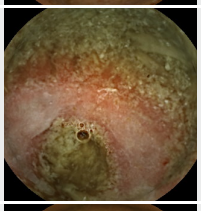
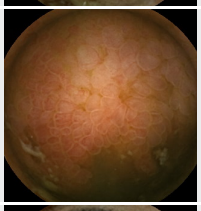
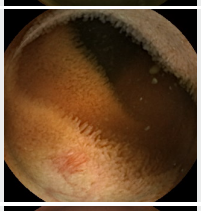
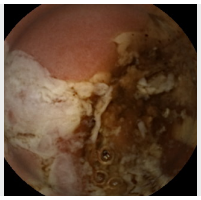
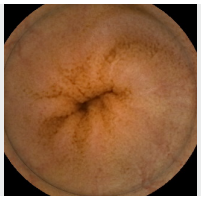
Nomenclature	Classe	Description	Exemple	Part de la base de données
Ulcération aphtoïde	UA	Perte de substance minime de la couche épithéliale, au fond blanc, entourée d'un halo érythémateux, posée sur une muqueuse normale		7,2%
Sténose	S	Rétrécissement de la lumière intestinale retardant ou empêchant le passage de la capsule		3,7%
Œdème	O	Aspect grossi, gonflé, engorgé des villosités		4,3%
Érythème	E	Zone de villosités rougeâtres.		3,6%
Ulcération de plus de 10mm	U+10	Perte de substance déprimée par rapport à la muqueuse environnante qui est gonflée/œdémateuse et à un fond blanchâtre dont la taille est supérieure à 10mm.		8,2%
Ulcération de moins de 10mm	U3-10	Perte de substance déprimée par rapport à la muqueuse environnante qui est gonflée/œdémateuse et à un fond blanchâtre dont la taille est inférieur à 10mm		11,9%
Non pathologique	NP	Image d'intestin normal		60,7%

TABLE 1.8 – Description des différentes classes utilisées lors des différentes phases d'annotation. La dernière colonne correspond au pourcentage de la base de données finale contenant la classe.

Chapitre 2

De l'attention humaine à l'attention artificielle, une quête d'explicabilité

Sommaire

Les réseaux de neurones profonds entraînés sur la base de données CrohnIPI et présentés dans le premier chapitre nous proposent des résultats encourageants permettant une possible utilisation dans la pratique clinique des gastro-entérologues.

Cependant, et comme dans de nombreux autres domaines, une critique fondamentale leur est adressée entravant leur expansion et leur intégration dans la pratique clinique des experts médicaux : les réseaux de neurones profonds sont considérés comme des « boîtes noires »[75].

La phase d'optimisation des algorithmes d'apprentissage profond nécessite de réaliser de millions de calculs nombreux et complexes dont l'inférence est encore mal comprise. De plus, l'influence des différents méta paramètres des réseaux et l'influence des données d'entrée sur ces calculs est également difficile à prévoir, et il est impossible de prévoir à l'avance les valeurs des jeux de poids résultant de l'apprentissage malgré l'approche stochastique de la descente de gradients des approches supervisées. Comment réussir à faire confiance à ces algorithmes que nous admettons ne pas comprendre complètement ? Quelles sont les solutions à mettre en place pour y remédier ?

Au travers de cette partie, nous essaierons de comprendre dans un premier temps les fondements du besoin d'interprétabilité. Dans un second temps, il sera présenté le principe d'attention chez les humains, afin d'introduire et comprendre ce concept. Et finalement nous verrons comment cette notion d'attention peut être traduite en terme algorithmique, comment l'attention permet de créer des explications visuelles pour comprendre le fonctionnement des réseaux de neurones profonds.

2.1 Interprétabilité des réseaux de neurones profonds

Les algorithmes d'apprentissage profond sont souvent considérés comme des boîtes noires de par leur manque d'explicabilité. Mais qu'entend-on par explicabilité ? Les premières définitions de l'explicabilité des algorithmes d'apprentissage sont apparues au milieu des années 1980 [76, 77]. Avec le récent développement des algorithmes d'apprentissage profond, leurs applications à des domaines impliquant des vies humaines et les nouveaux

enjeux légaux les encadrant, les définitions de notre besoin d'interprétabilité et celles d'un bon modèle interprétable se sont multipliées. Dans cette partie, nous nous baserons sur la définition de Lipton en 2018, avec son article « The Mythos of Model Interpretability » [78]. D'autres excellents articles proposent également leur définition d'un bon modèle interprétable comme Confalonieri et al. 2021 [79], Miller 2017 [80] ou encore Hoffman et al. 2018 [81].

Dans un premier temps dans cette partie, seront présentées nos attentes en terme d'explicabilité des modèles décisionnels ainsi que les enjeux légaux associés à l'explicabilité. Ensuite, nous présenterons les caractéristiques pouvant être rattachées à un modèle interprétable.

2.1.1 Le besoin d'interprétabilité

Les deux premières questions qu'il est essentiel de se poser sont : « Pour quelle raison les algorithmes d'apprentissage profond ont-ils besoin d'être explicables ? » et « Pourquoi sont-ils considérés comme inexplicables ? ». Le besoin d'explicabilité peut facilement s'expliquer par le fait que récemment les chercheurs ont commencé à s'intéresser à leur application pour assister des décisions critiques impliquant des vies humaines telles que la santé [82, 83], l'évaluation de risque pénal [84] ou la conduite autonome [85].

L'aspect boîte noire de l'intelligence artificielle vient du fait que le nombre de calculs menant à la décision réalisée par l'algorithme est bien trop important pour être produit à la main et interprété par un humain. Cependant, nous ne considérons pas nos décisions humaines comme inexplicables, bien que nos processus cognitifs soient très difficilement interprétables et ne soient pas toujours clairement compréhensibles. L'idée du manque d'interprétabilité des réseaux de neurones profonds provient donc de plusieurs facteurs entremêlés. Lipton [78] en décode plusieurs.

Premièrement, il est difficile de faire pleinement confiance aux prédictions des modèles. Il est essentiel de comprendre que les modèles d'apprentissage profond ont été entraînés sur des données précises, correspondant à une tâche bien définie. Il serait inconcevable d'utiliser un algorithme d'apprentissage profond entraîné à identifier des lésions intestinales dans des images de vidéos capsules endoscopiques pour remplacer un ophtalmologiste dans une tâche d'identification de lésions rétiniennes. Les algorithmes d'apprentissage profond réalisent leur prédiction à partir de connaissances acquises sur des exemples donnés qui peuvent contenir des biais et qui ne peuvent pas représenter la totalité des cas présents dans la nature. Il est donc bien nécessaire de comprendre dans quel cas l'algorithme réalise

des prédictions justes et de les comparer aux prédictions humaines pour pouvoir quantifier à quel point la supervision de l'être humain est nécessaire lorsque celui-ci est déployé.

Deuxièmement, les réseaux de neurones profonds ne permettent pas d'établir des relations causales facilement vérifiables. Les algorithmes d'apprentissage supervisé sont entraînés à établir des relations d'associations entre des caractéristiques localisées et des caractéristiques plus générales. Cependant, les chercheurs essaient souvent d'établir grâce à eux des relations causales pour pouvoir établir des lois générales nous permettant de mieux appréhender le monde qui nous entoure. Bien qu'il soit possible d'utiliser ces algorithmes pour formuler des hypothèses, il est cependant dangereux de les utiliser pour établir des relations causales qui risquent de ne refléter que les postulats utilisés pour les entraîner.

Troisièmement, on peut discerner chez l'être humain une fantastique capacité à transférer ses connaissances d'une expérience passée à une toute nouvelle situation. Cette capacité de généralisation, pour les algorithmes d'apprentissage supervisé, se limite souvent à la capacité de maintenir un niveau de performance similaire à celui obtenu sur le jeu de test. Il a été démontré que les réseaux de neurones peuvent être trompés par des perturbations indiscernables par l'humain et entraînant des erreurs de classification [86], bien que le réseau ait prouvé une bonne capacité de généralisation du jeu d'entraînement au jeu de test. Nous avons donc besoin que les algorithmes soient capables de transférer leur connaissance afin de pouvoir leur faire confiance.

Quatrièmement, les algorithmes décisionnels ne fournissent souvent à l'utilisateur qu'une prédiction, sans explication supplémentaire. Dans certains cas, les prédictions sans le raisonnement conduisant à la prédiction sont inutilisables, car ils ne permettent pas à l'utilisateur de combiner les connaissances de la machine avec les siennes. Pour les algorithmes d'aide à la décision, il est souvent bien plus utile pour l'utilisateur d'obtenir des informations supplémentaires qu'une prédiction dont le cheminement lui est inconnu.

2.1.2 Enjeux légaux

Au fur et à mesure que les systèmes d'IA évoluent, la question de leur autonomie et de leurs interconnexions avec d'autres systèmes d'IA soulève plusieurs questions liées au degré d'autonomie qui leur est réellement accordé ou aux mesures à prendre lorsqu'une solution d'IA est en désaccord avec un opérateur humain. De même, à mesure que l'omniprésence des systèmes d'IA augmente, les méthodes d'interprétabilité peuvent contribuer à atténuer l'augmentation du biais d'automatisation, par lequel les opérateurs humains

ne remarquent pas ou ne tiennent pas compte des défaillances de l'IA ou acceptent à tort la décision de la machine malgré des preuves contraires. C'est pour cela que l'explicabilité des algorithmes décisionnels est de plus en plus réglementée, notamment dans les secteurs critiques énumérés précédemment. On peut par exemple citer la Food and Drug Administration américaine qui propose de nouveaux cadres réglementaires afin de permettre le déploiement des technologies d'IA dans les soins et la santé [87]. On peut également citer le Règlement Général sur la Protection des Données (RGPD) en Europe, qui place l'explicabilité des algorithmes décisionnels au cœur des futurs défis du développement des algorithmes d'intelligence artificielle [88].

2.1.3 Les caractéristiques d'un modèle interprétable

Comment rendre ces algorithmes plus interprétables ?

Un algorithme d'apprentissage automatique interprétable peut être décrit comme un algorithme dans lequel le lien entre les caractéristiques utilisées par le système d'apprentissage automatique et la prédiction elle-même peut être compris par un humain [89]. Deux concepts phares sont alors essentiels pour définir les caractéristiques d'un bon modèle interprétable : la transparence et l'interprétabilité post-hoc. La transparence met en jeu plusieurs concepts. Premièrement, celui de pouvoir comprendre simplement le fonctionnement général du modèle. Bien que l'idée de simplicité soit subjective, il est essentiel de pouvoir expliquer rapidement le fonctionnement du modèle à l'aide de phrases ou de schémas. Deuxièmement, la transparence passe par la compréhension de chacune des sous-parties et des éléments individuels d'un modèle. Et finalement par l'explication du fonctionnement d'apprentissage des algorithmes.

Ensuite, un bon modèle interprétable se doit de pouvoir « justifier » sa décision en fournissant à l'utilisateur des informations supplémentaires à sa prédiction, permettant ainsi de comprendre le cheminement le conduisant à cette dernière. Cette idée de développer des méthodes permettant d'expliquer la prédiction d'un modèle *a posteriori* de son entraînement appartient au concept d'interprétabilité post-hoc. L'interprétabilité post-hoc est importante, car même si elle ne permet pas de clarifier de façon précise le fonctionnement du modèle, elle permet de donner à l'utilisateur des informations utiles pour l'assister dans sa décision. On trouve diverses méthodes d'interprétabilité post-hoc comme l'explication par texte, souvent réalisée à l'aide d'un réseau récurrent, la visualisation des caractéristiques des images, l'explication par l'exemple, en regroupant les cas les plus proches selon le réseau, où encore, et ce qui nous intéressera le plus dans ce manuscrit, le calcul des

zones d'intérêt ayant le plus contribué à la prédiction du réseau.

2.2 Attention humaine

Le fonctionnement du cerveau humain recèle de nombreux mystères quant à son fonctionnement et au peu d'énergie qu'il consomme afin de produire de si complexes abstractions. Afin de se frayer un chemin à travers ce mystère, l'étude de l'attention humaine a été un enjeu majeur. En effet, les avancées en neurosciences ayant démontré qu'une grande partie de notre cortex était allouée au traitement des stimuli visuels, il a été naturel de porter notre attention sur la compréhension des mécanismes impliqués.

Afin de pouvoir répondre aux enjeux de l'interprétabilité et également améliorer les performances des réseaux de neurones profonds, les chercheurs en vision par ordinateur ont développé des algorithmes dits à « attention ». Ces algorithmes, divisés en plusieurs sous-familles qui seront présentées dans la section 2.7, permettent de répondre pour certains aux besoins d'informations supplémentaires accompagnant la prédiction, et pour les autres au besoin de transparence des réseaux de neurones profonds.

Le principe des réseaux à attention s'inspire directement de l'attention humaine. Ainsi, avant d'expliquer comment fonctionne l'attention artificielle, il est nécessaire de décrire l'attention humaine, ainsi que son fonctionnement. Comprendre le fonctionnement attentionnel est également important, car dans le chapitre 4 l'attention humaine sera quantifiée de façon à pouvoir être comparée avec l'attention artificielle.

2.2.1 Définition

Notre cerveau reçoit en permanence une quantité considérable d'informations lui permettant d'appréhender la richesse du monde qui nous entoure. Malgré sa grande capacité de traitement, il ne traite pas l'entièreté des stimuli entrants. En effet, à partir de la quantité d'informations virtuellement illimitée présente dans l'environnement, 10^{10} bits/sec sont captés par la rétine, 10^4 bit/sec arrivent jusque à la couche 4 de V1 (la première zone du cortex visuel), et la bande passante de la conscience (ce qu'on l'on perçoit vraiment) est de l'ordre de 100 bits/sec. [90]

En effet, il existe un processus appelé attention, nous permettant de filtrer les informations les plus pertinentes afin de leur associer un maximum de ressources cognitives. Dès la fin du XIX$^{\text{è}}$ siècle, ce processus cognitif tente d'être compris et théoriser. On peut

en lire la première définition par William en 1890 [91] : « L'attention est la prise de possession par l'esprit, sous une forme claire et vive, d'un objet ou d'une suite de pensées parmi plusieurs qui sont présents simultanément [...]. Elle implique le retrait de certains objets afin de traiter plus efficacement les autres [...]. Seuls les éléments que je remarque façonnent mon esprit. Sans intérêt sélectif, l'expérience est un chaos total ».

L'attention est une propriété au cœur de toutes les opérations perceptives et cognitives. Elle régit le traitement cognitif de nos cinq sens (la vue [92], l'ouïe [93], le toucher [94], l'odorat [95] et le goût [96]), en nous permettant d'activer la région appropriée de notre cortex afin d'optimiser la compréhension de notre environnement. Malgré la présence d'un processus attentionnel orchestrant les perceptions liées à nos cinq sens, l'attention visuelle est à l'heure actuelle la plus étudiée des types d'attention. Premièrement parce que le sens de la vue est prédominant dans notre prise de décision, comme le montre l'importante place que prennent les aires visuelles dans notre cerveau [97]. Deuxièmement parce qu'elle est la plus facile à mesurer, car contrairement aux 4 autres sens, la vue possède un organe bien délimité, orientable et exogène : l'œil.

Comme nous l'avons vu précédemment seule une faible partie de l'information mis à disposition par l'oeil est traitée par le cortex visuel. Le centre de notre rétine, la fovéa, présente une densité de photorécepteur 40 fois plus importante que dans sa zone périphérique. Naturellement l'humain, afin d'optimiser l'analyse de son environnement, place les objets à étudier au centre de la fovéa.

On peut considérer deux types d'attention visuelle distincts, l'attention dite *covert* (cachée) et *overt* (flagrante)[98]. L'attention *covert* correspond à l'action de porter mentalement notre attention vers une région particulière de notre champ visuel, de la même manière que nous porterions notre attention auditive pour écouter la conversation de la table voisine dans un café (effet appelé « Cocktail party » [99]). La deuxième, l'attention *overt* se traduit par le déplacement des yeux vers la zone d'intérêt. En 1987, on démontre que l'attention *overt* et les mouvements oculaires utilisent les mêmes processus cognitifs [100].

Dans le reste de ce manuscrit, nous assimilerons le terme d'attention visuelle à l'attention *overt*, cette dernière étant bien plus facile à quantifier via l'analyse des mouvements oculaires.

2.2.2 Fixations et saccades

Dès 1903 Dodge [101] identifie 5 types de mouvements oculaires différents : des mouvements réflexes permettant de compenser les perturbations induites par le corps afin de stabiliser l'image rétinienne, les saccades, des mouvements rapides de l'œil permettant de passer d'un point de fixation à un autre, les poursuites, des mouvements lents permettant de maintenir la fovéa sur un objet en mouvement, les vergences, également des mouvements lents permettant d'orienter nos deux fovéas vers le même objet quelle que soit la distance de fixation, et les fixations, où le regard est relativement stable. Lors de l'exploration d'une scène, notre perception repose principalement sur l'enchaînement successif de saccades et de fixations, les autres mouvements n'intervenant que pour compenser des perturbations de notre environnement.

Durant les fixations, l'objet d'intérêt est placé au centre de la fovéa de sorte à maximiser la prise d'information. En effet, il a été montré que l'information visuelle était dégradée si la cible fixée se situait à plus d'un degré d'angle visuel de la fovéa. Les fixations sont caractérisées par leur position et leur durée. La durée de fixation permet d'obtenir des informations sur les processus cognitifs de traitement de l'information visuelle [102]. On peut remarquer que la durée des fixations est impactée par la tâche en cours. Par exemple, elles sont plus longues lors d'une tâche de mémorisation que lors d'une tâche de recherche visuelle [103]. Les saccades, elles, sont principalement caractérisées par leur amplitude.

2.2.3 Attention descendante et ascendante

Le comportement attentionnel des êtres humains peut être influencé par plusieurs facteurs pouvant être classés en deux grandes familles : les facteurs descendants (*top-down* en anglais) et les facteurs ascendants (*bottom-up*). Dans cette partie sera présentée succinctement l'attention ascendante, influencée par les facteurs ascendants, puis plus en détail l'attention descendante, influencée par les facteurs descendants, qui sera quantifiée et étudiée dans le chapitre 4.

Attention ascendante

Également appelée attention exogène, il s'agit d'un processus dirigé par les stimuli, où les caractéristiques saillantes sont automatiquement sélectionnées par le système visuel. En effet, l'attention humaine est influencée par la nature des stimuli. On peut observer des différences notables entre notre façon d'observer des scènes dynamiques et des images fixes

[104]. Notre attention est également influencée par le contenu des stimuli. Il a été établi par de nombreuses études [105] que des caractéristiques (contrastes, couleurs, luminance, intensité des contours) propres aux images influencent notre comportement attentionnel.

Comme ce processus est uniquement basé sur l'image, il est beaucoup plus simple à modéliser et de nombreux modèles de saillance visuelle ascendants ont été proposés dans la littérature, voir [106, 107, 108] pour des revues de la littérature.

Attention descendante

Également appelée attention endogène, ce processus est dirigé par l'observateur et influencé par ses connaissances préalables, son centre d'intérêt, la tâche à accomplir ou son état cognitif [109]. Elle fait référence à l'allocation volontaire de l'attention à certains objets, caractéristiques ou régions de l'espace [110]. On peut trouver des preuves que la tâche influence le comportement attentionnel humain dès 1967 avec l'expérience de Yarbus sur la toile de « Le visiteur inattendu » (1888) du peintre russe Ilya Repin. Il montre qu'en fonction de la tâche donnée au sujet observant le tableau, la façon d'explorer le tableau change profondément comme il est montré sur la figure 2.1. Des études ont montré que l'attention descendante mettait plus longtemps que l'attention montante à se déployer, 300ms contre 100-120ms [111, 112].

Les facteurs descendants étant plus difficiles à modéliser, ils ont reçu moins d'attention de la part de la communauté de vision par ordinateur, bien que le travail de leur modélisation gagne de l'intérêt dans la communauté [114, 115]. Pour le reste de ce manuscrit, nous nous intéresserons principalement à l'attention descendante, de par le fait que lors d'une tâche de classification d'image intestinale, cette dernière est largement plus mobilisée que l'attention ascendante.

2.3 Quantification de l'attention visuelle humaine

La quantification de l'attention visuelle s'organise en plusieurs étapes. Premièrement, il est essentiel de pouvoir mesurer les zones d'attention des sujets. Pendant les deux derniers siècles, la recherche en oculométrie a permis d'aboutir à des oculomètres permettant d'enregistrer les positions des fixations oculaires sur un écran. Cette technologie est indispensable pour pouvoir évaluer l'attention visuelle, car il a été montré que les sujets étaient inefficaces pour savoir quelles parties d'une image ils avaient regardées, même après une courte période [116, 117]. Une fois les données oculométriques récupérées, ces dernières

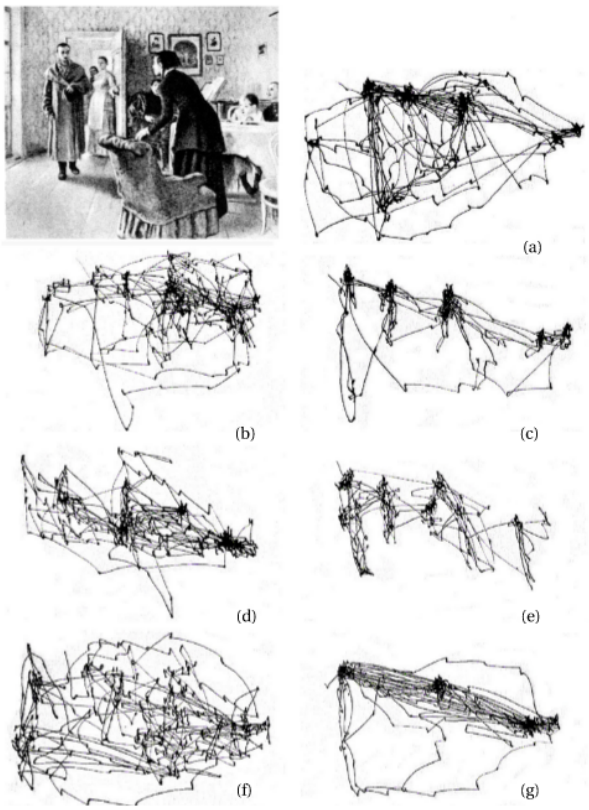


FIGURE 2.1 – Examiner une image (Le visiteur inattendu) avec différentes questions en tête. Chaque enregistrement durait 3 minutes. (a) Examen libre. (b) Estimez la situation matérielle de la famille représentée sur la photo. (c) Donnez l'âge des personnes. (d) Supposez ce que la famille faisait avant l'arrivée du « Le visiteur inattendu ». (e) Rappelez-vous les vêtements portés par les personnes. (f) Rappelez-vous la position des personnes et des objets dans la pièce. (g) Estimez combien de temps le visiteur inattendu a été éloigné de la famille. (Illustration et légende adaptées de Yarbus 1967, extrait de Tatler 2010 [113])

peuvent être analysées de façon temporelle en s'intéressant aux durées des fixations et aux amplitudes des saccades à l'aide des *scanpaths* ou en observant les zones d'intérêt des sujets pour une image donnée avec les cartes de saillance. Plusieurs métriques permettant de quantifier les points communs et les différences entre les comportements attentionnels seront présentées dans cette partie. Finalement, afin de valider l'influence de tel ou tel facteur sur des comportements attentionnels, l'utilisation de modèle statistique est essentiel. Dans ce manuscrit, nous utiliserons les modèles linéaires mixtes, ces derniers étant présentés dans la partie 2.6.

2.3.1 Historique de l'oculométrie

L'étude de l'attention visuelle peut sembler liée au développement de nouvelles technologies comme l'eye-tracking tant ces dernières facilitent la quantification des processus attentionnels. Cependant, on peut trouver des traces des premières recherches dès la fin des années 1870 avec les travaux d'un ophtalmologiste français, Louis Émile Javal, qui concluait que lire ne requérait pas de mouvements continus de l'œil, mais plutôt des mouvements rapides de l'œil (saccades). Il plaçait un miroir sur les pages du livre lu par le sujet testé, derrière lequel se tenait l'expérimentateur qui observait le mouvement des yeux du sujet dans ce miroir [118].

On considère que le premier eye-tracker fut inventé en 1908 par Edmund Huey. Le dispositif était composé d'une lentille de contact avec une ouverture pour l'iris, cette dernière était reliée à un pointeur changeant de position suivant les mouvements de l'œil. Cette méthode bien qu'efficace était extrêmement invasive, nécessitant l'administration de cocaïne aux sujets afin que la douleur puisse être supportable. Il montrera à l'aide de cette technique que chaque mot d'une phrase ne revêt pas la même importance [119].

En 1935, un dispositif moins invasif a été mis au point par Guy Thomas Buswell pour réfléchir des faisceaux lumineux sur les yeux et les enregistrer sur un film. Ce dernier montra que différents sujets avaient des comportements assez similaires lors de l'observation d'œuvres d'art complexes. [120]

Aujourd'hui les eye-trackers utilisent toujours le même principe. Le dispositif de mesure est composé d'une caméra équipée d'une diode électroluminescente produisant une lumière infrarouge dirigée vers les yeux du participant. En calculant le vecteur entre le centre de la pupille et le reflet cornéen (la lumière projetée dans l'œil du participant produisant un reflet sur sa cornée) [121], il est possible de définir le point d'attention de l'œil avec une grande précision[122]. Ici, la lumière infrarouge est utilisée afin de ne pas éblouir le participant et donc affecter la mesure. L'avantage d'utiliser cette méthode est qu'elle permet de dissocier les mouvements de tête des mouvements oculaires [123]. Ces appareils de mesure doivent être calibrés en fonction des mouvements oculaires de chacun des participants. Cela implique une étape préalable au cours de laquelle le participant regarde une série de points qui apparaissent à des endroits prédéfinis, formant généralement une grille de 9 points couvrant la totalité de l'écran. Pour chaque point fixé, l'ordinateur enregistre la relation centre pupillaire/réflexion coréenne correspondante aux coordonnées x, y spécifiques de l'emplacement de ce point sur l'écran. À partir de ce processus d'étalonnage, l'ordinateur peut ensuite interpoler afin de déterminer les coordonnées x, y de n'importe

quel point d'observation.

2.3.2 Analyse des données oculométriques

On peut distinguer deux types d'analyse des données oculométriques, les analyses diachroniques et les analyses synchrones [124]. Soit l'on considère que chacun des événements attentionnels se produit à des moments précis du temps en réalisant une analyse synchrone, soit on considère que les événements se déroulent au cours d'un flux temporel en réalisant une analyse diachronique. Le travail d'analyse synchrone de l'attention réside dans l'étude des saccades (de leur amplitude et de leur vitesse) et des fixations (leur nombre et leur durée).

L'analyse diachronique de l'attention réside dans l'étude des cartes de saillance et des scanpaths. Être capable de mesurer les points communs et les différences entre plusieurs comportements attentionnels est fondamental pour comprendre l'impact des différents facteurs montant et descendant sur notre processus cognitif. Dans cette partie, nous présenterons rapidement le concept de scanpath, et nous nous intéresserons en profondeur à celui de carte de saillance, essentielle à la comparaison entre attention humaine et attention artificielle.

Scanpath

On peut définir les scanpaths (ou séquences de balayage en français) comme « une alternance idiosyncrasique[1] et répétitive d'aperçus et de sauts rapides de la position des yeux vers diverses régions d'intérêt dans la scène observée » [125]. Cette théorie d'analyse se base sur le fait qu'un modèle cognitif spatial interne contrôle à la fois la perception et le mécanisme des mouvements oculaires.

Cette représentation d'une suite de mouvements oculaires permet de mettre en avant la dimension temporelle de l'attention humaine. Afin de permettre la comparaison des scanpaths, de nombreuses méthodes utilisant leur dimension temporelle ont été développées [126, 127, 128].

La comparaison de deux scanpaths nécessite de prendre en compte un certain nombre de facteurs, tels que la dimension temporelle ou la procédure d'alignement. L'étude des cartes de saillances et des cartes de fixations permet de surmonter ces problèmes, tout en permettant la comparaison entre différents comportements attentionnels.

1. Qui a trait aux caractères propres du comportement d'un individu particulier

Cartes de saillance et de fixations

La définition des cartes de saillance et des cartes de fixations varie dans la littérature. Ici, et pour le reste du manuscrit, nous les définirons comme suit.

La différence entre ces deux types de cartes bidimensionnelles est que les cartes de fixations sont discrètes quand les cartes de saillance sont continues. Les cartes de fixations sont donc des cartes discrètes, correspondant à la répartition des regards sur une image. Si un pixel a été observé par l'utilisateur, il aura la valeur 1, sinon 0. La carte de saillance est la version continue de la carte de fixation. Pour l'obtenir, on réalise une opération de convolution entre la carte de fixation et un noyau gaussien isotropique à deux dimensions. La taille du noyau est définie de façon à représenter 1° d'angle visuel, soit l'estimation de la taille de la fovéa, et dépend des conditions de l'expérience (taille de l'écran et distance entre le participant et l'écran). L'approximation d'une fixation avec une gaussienne est couramment admise par de nombreuses études [129, 130].

Avec l'utilisation de cartes de saillance et de cartes de fixation, la dimension temporelle de l'attention humaine n'est pas prise en compte car on agrège les mouvements oculaires sur l'ensemble de l'exploration. L'objectif de ce manuscrit étant de comparer l'attention humaine et l'attention artificielle, nous utiliserons principalement les cartes de saillances et de fixations, du fait que l'attention artificielle ne possède pas de dimension temporelle comme l'attention humaine. Nous utiliserons le terme de carte d'attention pour désigner les cartes de saillances obtenues lors d'une tâche mobilisant principalement des facteurs descendants.

2.3.3 Métrique de comparaison

Nous avons employé deux métriques fréquemment utilisées dans la litérature pour comparer les distributions spatiales de l'attention [131].

Premièrement, nous utiliserons le Normalized Scanpath Saliency (NSS). Cette métrique implique une carte de fixations et une carte d'attention. Le NSS prend les valeurs de la carte d'attention aux emplacements de fixation [132]. Elle est définie par l'équation suivante, avec une carte d'attention donnée P, sa normalisation centrée normée $\overline{P}$, une carte de fixation binaire Q^B, et i l'indice du i-ème pixel :

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \overline{P_i} \times Q_i^B,$$

où

$$N = \sum_i Q_i^B,$$

et

$$\overline{P} = \frac{P - \mu(P)}{\sigma(P)}.$$

NSS=0 signifie que les fixations et la carte d'attention ne sont pas liées, NSS>0 signifie qu'elles sont positivement liées, NSS<0 signifie qu'elles sont négativement liées.

Nous utiliserons également le Coefficient de Corrélation de Pearson (CC). Cette métrique quantifie la relation linéaire entre deux variables. Elle est définie par l'équation suivante, avec deux cartes d'attention P et Q.

$$CC(P, Q) = \frac{\sigma(P, Q)}{\sigma(P) \times \sigma(Q)}$$

De nombreuses autres métriques peuvent être utilisées afin de comparer l'attention visuelle. On peut trouver d'excellentes comparaisons de ces différentes méthodes dans [124] et [131].

2.4 Applications

Mais pourquoi étudier l'attention visuelle ? Quels sont les comportements qui peuvent être caractérisés ? Quels sont les domaines d'applications de ces connaissances ? Dans cette partie seront présentés différents exemples de domaines d'application de l'étude de l'attention visuelle. Dans un premier temps, nous présenterons les applications au multimédia, puis nous verrons comment ce savoir peut être utilisé dans le domaine du diagnostic de troubles neurocomportementaux, et finalement, nous verrons de façon un peu plus profonde les applications dans l'imagerie médicale.

2.4.1 Dans les multimédia

La recherche de la compréhension des mécanismes de l'attention visuelle peut permettre l'amélioration de notre perception des contenus multimédias. La connaissance des zones saillantes d'une image peut permettre de réaliser des algorithmes de compression sélective [133]. Ces algorithmes permettent d'allouer plus de ressources à l'encodage des zones considérées comme saillantes, et permettent ainsi une meilleure qualité visuelle.

Cette méthode peut également être utilisée pour s'assurer que les pertes d'informations liées à la transmission de flux vidéo affectent le moins possible les régions saillantes. Avec la multiplication des supports de visionnage des contenus multimédias (tablettes, smartphones, grands écrans, ...), la stimulation du champ visuel de l'utilisateur peut varier énormément. Ainsi en comprenant mieux les tâches réalisées par les utilisateurs et les conditions de visionnage, il est possible de réadapter les paramètres de visualisation (taille de l'image, résolution, fréquence d'images) de sorte à maximiser l'expérience de l'utilisateur [134].

2.4.2 Dans le diagnostic clinique

Le déploiement de l'attention visuelle, et donc des mouvements oculaires associés, est lié à de nombreuses fonctions cérébrales [135, 136]. De nombreux troubles neurologiques affectent le comportement oculaire. En effet, il a été démontré que le contrôle de l'attention impliquait les cortex occipitaux, temporels, frontaux, pariétaux ainsi que le système limbique, les systèmes de récompense, et des noyaux cérébraux profonds, dont le thalamus et le colliculus supérieur [137, 138, 139]. On peut observer des différences de comportement oculaire chez les personnes atteintes de troubles neurocomportementaux, notamment au niveau du temps de réaction saccadique dans des tâches où le comportement du sujet est guidé par la vision [140]. On peut trouver un exemple de détection de comportement attentionnel atypique chez les personnes atteintes d'autisme [141], de schizophrénie [142], de dépression [75], ou encore de la maladie d'Alzheimer [143]. En identifiant ces comportements spécifiques, il est alors possible d'aider le diagnostic de ces différentes pathologies [144].

2.4.3 Dans l'imagerie médicale

Avec les récents développements de l'imagerie médicale (voir partie 1.1), la pratique clinique a beaucoup évolué. Lors d'un diagnostic basé sur des informations issues de l'imagerie médicale, les experts opèrent dans un premier temps une phase d'inspection visuelle, suivie d'une phase d'interprétation. L'inspection visuelle met en jeu des processus de perception visuelle, notamment des tâches de détection et de localisation, faisant intervenir des facteurs descendants. La phase d'interprétation, elle, fait intervenir des processus cognitifs, pouvant impliquer des erreurs dont les conséquences peuvent être importantes pour la prise en charge du patient. Il a été montré que de nombreux cas de cancer n'avaient pas

été détectés malgré la présence de tumeurs dans les images médicales [145, 146] et que le taux de faux négatifs et de faux positifs restait élevé [147]. L'étude des mouvements oculaires des experts lorsqu'ils effectuent des tâches de perception d'images médicales est un moyen d'identifier les points faibles éventuels des processus de perception d'images médicales, et ainsi d'envisager des méthodes afin d'améliorer leurs performances [148].

La compréhension de la perception de l'imagerie 3D est également un enjeu phare de l'oculométrie dans l'imagerie médicale. Le passage de la 2D à la 3D modifie nécessairement les analyses et les performances, se traduisant par de nouvelles stratégies d'exploration [149].

L'étude des scanpaths et du champ de vision utile est également intéressante. Le champ de vision utile est l'espace de notre champ visuel où notre attention nous permet d'interpréter ce que nous voyons. La taille de ce champ de vision utile est spécifique à la tâche et à l'image. Il a été établi que lorsqu'une anomalie est présente trop loin d'une fixation, elle n'est pas vue par un expert [150], ce qui impose un temps de visionnage et une taille d'image spécifique à la tâche.

L'étude des mouvements oculaires nous renseigne aussi sur la nature des erreurs. On peut en distinguer trois sortes : les erreurs de recherche, de reconnaissance et de décision [151]. Des exemples de scanpaths associés à chacun des types d'erreurs sont répertoriés dans la figure 2.2. Les erreurs de recherche sont caractérisées par une absence de fixation sur la cible. L'erreur de reconnaissance est identifiable à une fixation brève sur la cible. À la suite de cette fixation brève, le lecteur continue de balayer l'image sans que rien n'indique qu'il ait perçu un élément de diagnostic. Finalement l'erreur de décision est caractérisée par de multiples et/ou longues fixations sur la cible. Le lecteur reconnaît que la cible est un élément qui mérite d'y porter attention, mais a finalement pris la mauvaise décision. Dans leur étude sur les nodules pulmonaires, Kundel et al. [151] ont constaté que les cliniciens faisaient environ 30 % d'erreurs de recherche, 25 % de reconnaissance et 45 % d'erreurs de décision. Ces proportions ont été validées par d'autres études [152, 153, 154].

Les recherches en perception ont également permis de mettre en valeur le phénomène de « cécité attentionnelle ». Inspirée de la célèbre vidéo où on ne remarque pas un gorille danser au milieu de personnes jouant à la balle [155], une expérience a été réalisée visant à montrer le même phénomène chez les radiologistes [156]. Ce phénomène est défini par une attention saturée par certains stimuli, empêchant la détection d'un autre stimuli, la plupart du temps inattendu, qui devrait pourtant être perçu. Sur une des images tomographiques, revue par 24 radiologistes, avait été rajouté un gorille. Seul quatre d'entre

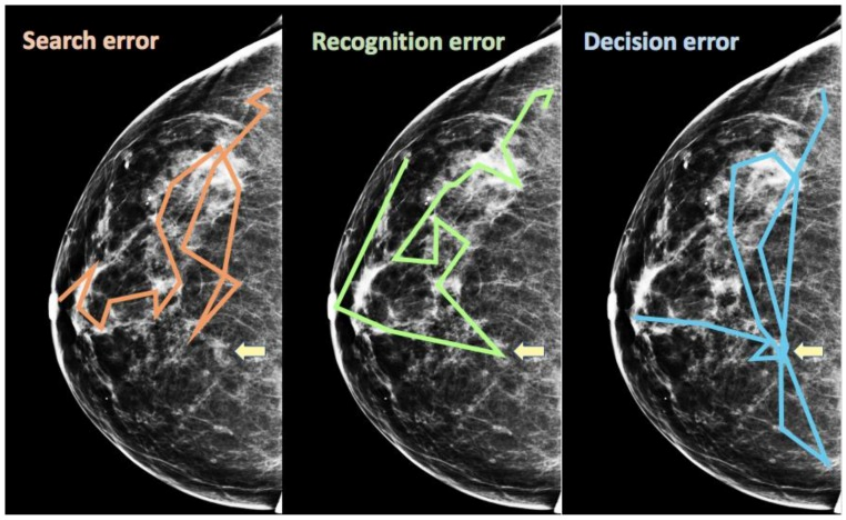


FIGURE 2.2 – Figure extraite de [148]. Ces différents types de scanpath sont caractéristiques des types d'erreurs pouvant être réalisées par les experts lors de leur diagnostic.

eux l'ont repéré, réalisant principalement des erreurs de reconnaissance, montrant que ce phénomène de « cécité attentionnelle » est également présent lors des phases de diagnostic et qu'il est important d'être vigilant.

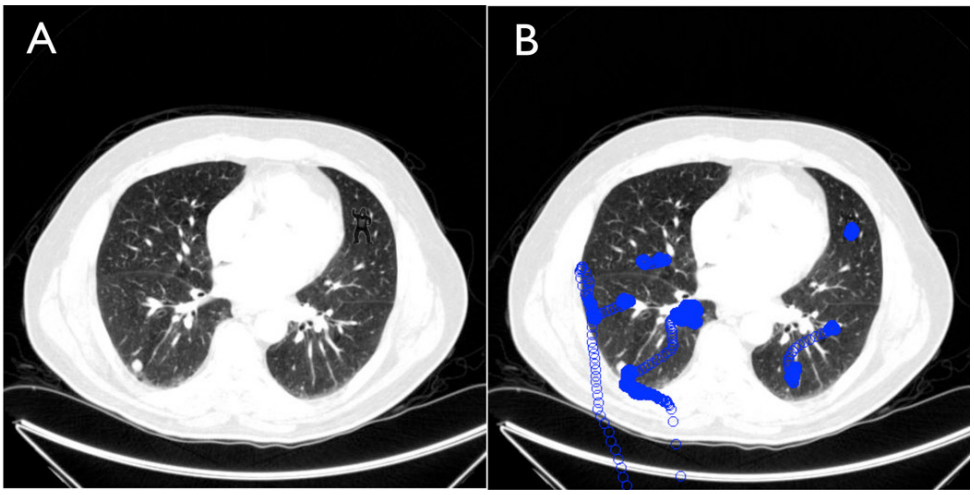


FIGURE 2.3 – Figure extraite de [156]. Cet exemple illustre le concept de « cécité d'inattention ». Dans cette image de radio de poitrine, un gorille a été rajouté. L'image A correspond à l'image originale et l'image B correspond aux résultats de l'expérience d'eye-tracking, chacun des cercles correspondant à une position oculaire pendant 1ms.

Toutes ces recherches et ces nouvelles connaissances, permettent d'accélérer la formation des novices, par exemple à l'aide de méthodes basées sur de l'eye-tracking et un

protocole de réflexion à voix haute voix, donnant des résultats encourageants [157]. Elles permettent également d'évaluer de manière plus objective le développement des compétences chez les novices en formation [158]. Des recherches ont également montré que les capacités d'observation visuelle peuvent être transmises aux novices [159].

2.5 Influence de l'expertise sur l'attention humaine

Le développement de l'eye-tracking a également permis d'étudier l'effet de l'expertise sur notre comportement attentionnel. Diverses études ont montré des stratégies attentionnelles différentes entre les experts et les novices. Il a pu être établi que les experts sont plus rapides et plus précis que les novices pour identifier les régions suspectes d'une image médicale. Ils sont également moins distraits par les régions de l'image ne permettant pas de prendre la décision de diagnostic [160, 161, 162].

Une théorie propose que la capacité des experts à réaliser des classifications précises dans des images médicales repose sur des stratégies perceptives spécifiques [163]. Les experts encodent des informations visuelles plus larges que les novices en développant rapidement une représentation relativement holistique d'une configuration globale [164]. On peut prendre l'exemple des grands joueurs d'échecs étant capable en quelques coups d'oeil précis de reproduire les positions des pièces sur un échiquier. À la différence des joueurs moins expérimentés, ces derniers opèrent moins de fixations, pour de meilleures performances. Leurs fixations à la différence des novices sont réalisées entre les pièces de sorte à se rappeler d'une configuration de pièces plutôt que de mémoriser la position individuelle de chacune [165].

2.6 Outil de validation statistique : le modèle linéaire mixte

2.6.1 Intérêt du modèle

Dans le chapitre 4, nous nous intéresserons à une comparaison entre les zones d'attention humaines et les zones d'attention artificielles. Afin d'évaluer à quel point l'attention artificielle est une modélisation correcte de l'attention humaine, nous utiliserons des modèles linéaires mixtes. Ces modèles sont parfaitement adaptés aux données biologiques qui peuvent être complexes, désordonnées et dont le nombre de mesures peut être assez faible.

Mais l'intérêt principal de ces modèles réside dans le fait qu'ils permettent de prendre en compte dans la modélisation des facteurs de groupe influençant les mesures. En effet, dans nos expériences, nous avons réalisé plusieurs mesures par participant, ce qui viole l'hypothèse d'indépendance nécessaire à la mise en place d'un modèle de régression linéaire simple. Grâce à ces modèles, nous pourrons évaluer si les labels des images et le niveau d'expertise des humains ont un effet significatif sur la comparaison entre l'attention humaine et artificielle en permettant au modèle de prendre en compte des effets aléatoires pour prendre en compte les comportements attentionnels intrinsèques des participants.

Pour valider statistiquement les résultats obtenus par les différentes comparaisons qui seront présentées dans les chapitres 3 et 4, nous utiliserons des modèles linéaires mixtes.

Afin de bien comprendre ce type de modèle et ses cas d'application, nous allons prendre un exemple fictif. Imaginons que nous souhaitons mesurer l'influence de l'ensoleillement sur la taille de concombres provenant de plusieurs exploitations de différentes régions de la France. Nous aurions donc mesuré la taille de n concombres au détour des différentes exploitations concombricoles.

Dans un premier temps, nous allons considérer la régression linéaire simple pour les n observations d'une variable réponse y (la taille des concombres) et d'un prédicteur x (le temps d'ensoleillement de chacune des régions). Selon ce modèle, l'observation y_k (pour $k = 1, 2, ..., n$) suit une distribution normale $N(\mu_k, \sigma_y)$ avec une moyenne $\mu_k = \beta_0 + \beta_1 x_k$.

Dans notre cas, les observations sont groupées par l'exploitation de provenance ainsi que par leur région. Ici nous nous attendons à ce que la variation résiduelle de la réponse, soit la partie de la variance de la variable y non expliquée par x, ne soit pas indépendante d'une observation à l'autre. En effet, il est logique d'imaginer que bien que le temps d'ensoleillement soit proche pour les exploitations d'une même région, la nature des sols et les techniques agricoles varient d'une exploitation à l'autre. La taille des concombres d'une même exploitation (d'un même groupe) tend à être plus proches que les observations de groupes différents en raison de facteurs non mesurés qui varient au niveau du groupe plutôt qu'au niveau de l'observation individuelle.

Le modèle linéaire mixte permet de représenter cette situation en permettant aux coefficients du modèle linéaire de varier d'un groupe à l'autre selon une distribution normale. Ainsi, si β_0 et β_1 varient d'un groupe à l'autre et que $j[k]$ désigne le groupe (l'exploitation) j contenant l'observation k alors la valeur moyenne selon le modèle linéaire mixte est égale à :

$$\mu_k = \beta_0 + \beta_{j[k]} + \beta_1 \times x_k$$

Dans ce modèle, y_k suit toujours une distribution normale $N(\mu_k, \sigma_y)$ ainsi que le paramètres $\beta_{j[k]}$.

Les modèles linéaires mixtes sont ainsi nommés du fait qu'ils combinent des effets fixes spécifiés par le prédicteur x et des effets aléatoires représentant la variation entre les groupes. En ajustant le modèle linéaire mixte, nous permettons l'estimation de la valeur moyenne du coefficients β_0, de son écart-type ainsi que de l'écart-type des observations individuelles par rapport à la moyenne des groupes.

Le modèle linéaire mixte contracte les effets de chaque groupe en direction de l'effet moyen. Cette idée est basée sur le fait qu'une partie des différences observées entre les groupes est due au hasard de l'échantillonnage plutôt qu'à une réelle différence entre eux. Cet effet est encore plus fort lorsqu'il y a peu d'observations dans le groupe.

Ces modèles permettent donc de prendre en compte à la fois des effets fixes et des effets aléatoires. On note l'équation du modèle de la façon suivante :

réponse $\sim$ prédicteur 1 * prédicteur 2 + (1 | effet aléatoire 1) + (1 | effet aléatoire 2)

La notation (1 | effet aléatoire 1) correspond bien à estimer un intercept par groupe. Le produit prédicteur 1 * predicteur lui correspond à estimer le coefficient β_1. Dans notre cas :

Taille du concombre $\sim$ Temps d'ensoleillement + (1 | exploitation de provenance)

2.6.2 Application à notre problématique

Dans notre cas d'étude, plusieurs sujets ont été confrontés à plusieurs images. Nous souhaitons observer l'influence du niveau d'expertise et du label des images sur les comportements attentionnels humains et artificiels à l'aide des métriques décrites dans la partie 2.3.3. Ces mesures ne sont pas indépendantes des images et des sujets. Un même sujet pourrait avoir une façon d'observer des images propre à lui-même indépendamment de son niveau d'expertise et une même image peut induire des comportements attentionnels similaires chez différents sujets, indépendamment de la pathologie contenue ou non dans l'image.

Ainsi les modèles linéaires mixtes peuvent prendre en compte l'influence de ces facteurs

de groupe. Dans notre cas les effets aléatoires sont ceux de l'observateur et de l'image montrée. Les effets dont on souhaite mesurer l'influence sont appelés effets fixes, dans notre cas, le label de l'image et le niveau d'expertise du sujet.

2.7 L'attention artificielle

L'attention artificielle a été développée depuis une dizaine d'années afin d'améliorer les performances des réseaux de neurones profonds en s'inspirant de comportements humains et surtout afin de les rendre plus explicables. Les modèles à attention apprise permettent d'avoir des modèles plus transparents, se rapprochant plus d'un comportement humain facilement compréhensible et interprétable. On peut en distinguer deux grandes familles, les algorithmes à attention douce assignant un poids à chacun des éléments de l'entrée et les algorithmes à attention dure déterminant quelle partie de l'entrée doit être considérée ou non. Les modèles à attention post-hoc, quant à eux permettent de visualiser les zones d'intérêt d'un réseau de neurones profonds déjà entraîné. Ce mécanisme permet de fournir aux utilisateurs des informations utiles, grâce aux explications visuelles, sur la décision du système et de repérer d'éventuels biais. Dans cette partie seront présentées ces deux types d'attention et des exemples d'application.

2.7.1 Attention apprise

Définition

L'attention apprise, correspond à une portion d'architecture de réseaux de neurones profonds ayant pour consigne d'apprendre à considérer plus où moins certaines parties de l'image. Lors de ces 6 dernières années, l'intérêt pour ce genre d'algorithme a considérablement augmenté. À travers ces nouvelles architectures, les chercheurs essaient à la fois de gagner en explicabilité, en rendant les modèles plus transparents, et également de gagner en performance en s'inspirant du comportement humain. D'après [166], une excellente revue, analysant plus de 6500 articles scientifiques traitant des réseaux de neurones profonds à attention, quelques articles ont permis à la communauté scientifique d'avancer dans l'élaboration et le test de nouvelles architectures de plus en plus utilisées aujourd'hui. La fig. 2.4, nous indique les articles scientifiques les plus importants du domaine. Pour le traitement du langage naturel, c'est *RNNsearch* [167] et *Transformer* [168] qui semblent réellement avoir révolutionné le domaine. En ce qui concerne le traitement d'image, ce

qui nous intéresse plus particulièrement dans cette partie, on peut noter l'influence de « Show, attend and tell » [169], RAM [170], DRAM[171] et Image GPT [172].

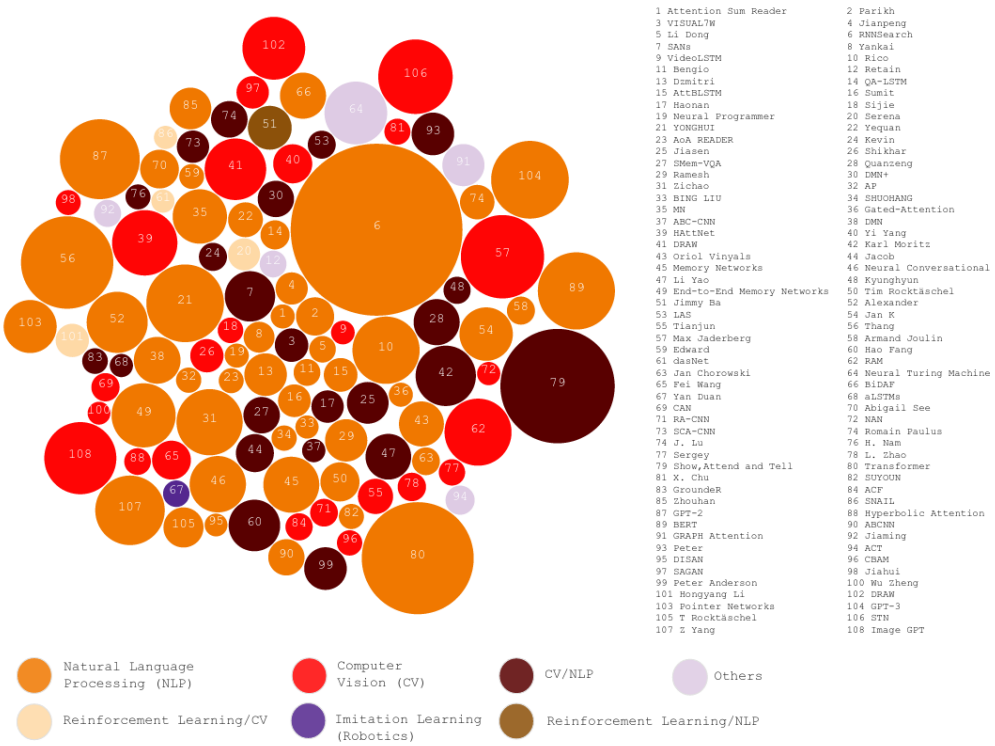


FIGURE 2.4 – Figure extraite de [166]. Chacun des cercles correspond à une architecture et le rayon des cercles à leur impact sur le développement des réseaux de neurones à attention. Cet impact est défini par le nombre de citations et le degré d'innovation de l'architecture. L'analyse est réalisée à partir de plus de 6500 articles scientifiques traitant des réseaux de neurones profonds à attention apprise.

On peut discerner deux types principaux d'algorithmes à attention : les réseaux à attention dure, apprenant à sélectionner uniquement certaines parties des images en entrée en ignorant le reste, et les réseaux à attention douce, apprenant à pondérer l'importance des différentes parties de l'image d'entrée pour la décision finale. Ces réseaux participent à rendre les modèles plus transparents et plus explicables, car ils permettent de comprendre quelles sont les parties de l'entrée permettant au réseau de réaliser sa prédiction.

Attention douce

L'objectif central de l'attention douce est de mettre en valeur les caractéristiques importantes à la prise de décision et d'atténuer celles induisant du bruit. Elle fait corres-

pondre à chaque élément de l'entrée une pondération permettant de leur accorder plus ou moins d'importance dans le processus décisionnel. Ces modèles sont purement déterministes et différenciables, le mécanisme d'attention étant souvent réalisé à l'aide de fonction softmax, rendant les modèles complètement différenciables. Ce processus permet de modéliser l'interdépendance entre les mécanismes d'entrée des réseaux et la cible.

Les premiers réseaux de neurones à attention douce voient le jour au début des années 2010, et obtiennent leurs premières réussites pour des algorithmes de traitement automatique du langage naturelle avec [167]. Le fonctionnement des algorithmes de traduction se basaient sur des structures encodeuses/décodeuses, codant la phrase d'entrée en un vecteur de caractéristiques de taille fixe et le décodant en sa traduction. Le problème de ces algorithmes était que plus la taille du vecteur de la phrase d'entrée était grande, plus les performances diminuaient [173]. C'est afin de pallier ce problème que le concept d'attention a fait pour la première fois ses preuves dans le monde des réseaux de neurones profonds. Le réseau [167] apprenait à aligner les mots de la sortie et de l'entrée à chaque itération en cherchant à chaque étape le mot le plus important de la phrase d'entrée. C'est ce concept de chercher les parties de l'information disponible la plus importante et de moduler la sortie en fonction qui peut être assimilé au principe d'attention humaine.

On peut trouver ensuite de nombreux exemples d'application de cette idée d'attention douce à la vision par ordinateur. Ces réseaux sont souvent utilisés pour générer des légendes d'images [169], générer des réponses par rapport à des questions posées sur des images [174] et classifier des images [175, 176]. L'intérêt de ce type d'attention est que les réseaux dotés d'un module d'attention douce sont complètement dérivable, à la différence des modules d'attention dure présentés dans la partie suivante qui eux nécessitent souvent d'utiliser un apprentissage par renforcement, ne permettant pas d'utiliser les techniques traditionnelles de descente de gradient.

Attention dure

L'attention dure accorde à chaque élément de l'entrée un poids binaire, soit 0 soit 1. Cela signifie que chaque élément est soit vu, soit non vu. Tout comme l'attention douce, ce mécanisme permet de refléter l'interdépendance entre l'entrée du réseau et la cible. Cependant, à la différence de l'attention douce, l'attention dure n'est pas différenciable, du fait de la non-linéarité des valeurs des poids du module d'attention et du processus de sélection séquentielle que cela implique. Ces algorithmes sont souvent entraînés avec de l'apprentissage par renforcement. L'apprentissage par renforcement, bien qu'ayant le

désavantage de ne pas présenter le même niveau de stabilité que l'apprentissage supervisé, permet d'entraîner des modèles non différenciables, avec un processus de maximisation d'une fonction récompense. Chaque action du modèle modifie son environnement, influant alors sur la décision finale, qui est associée à un score dont la valeur augmente lorsque le modèle réalise une action en accord avec l'objectif fixé. Ces modèles sont probabilistes, d'où le comportement plus instable que les modèles déterministes. Ces modèles possèdent l'avantage de présenter des temps d'inférence et des coûts en calcul plus faibles que les modèles à attention douce ou que les modèles traditionnels, de par le fait que l'entrée entière n'est pas stockée ou traitée.

Le premier exemple de réseau de neurones profond à attention dure fut développé par Larochelle et Hinton en 2010 [177]. Il est basé sur une machine de Boltzmann apprenant des séquences de regards sur des images. Avec le développement de l'intelligence artificielle, on peut trouver des versions différentes de cette même idée avec [170], [178] et [166]. Ces méthodes de regard inspirés du comportement humain, alternant saccades et fixations pour interpréter leur environnement seront développées plus en détail dans la partie 3.1, avec notre algorithme élaboré dans le cadre de l'identification des lésions de la maladie de Crohn. On peut trouver aussi l'algorithme DRAW qui applique ce mécanisme au problème de génération d'image, en utilisant ce processus séquentiel d'attention spatiale avec un auto-encodeur variationnel [179].

2.7.2 Attention post-hoc

L'objectif des algorithmes d'extraction de l'attention post-hoc est de permettre la visualisation des parties de l'entrée qui conduisent à la décision finale, en se concentrant sur les informations contenues dans le modèle. Ce type d'attention vient être extraite *a posteriori*, l'algorithme d'apprentissage supervisé ayant déjà été entraîné sur un jeu de données au préalable. Ces méthodes ne nécessitent pas de ré-entraînement et d'étapes d'optimisation supplémentaires. On peut trouver trois types de méthode : les méthodes dites par perturbation, les méthodes par rétropropagation et les méthodes basées sur l'activation des cartes de caractéristiques.

Grad-CAM

Grad-CAM est un algorithme basé sur les travaux de Bengio et al. 2012 [180] et Mahendran et al. 2016 [181] qui montre que plus la couche convolutive est profonde (donc

proche de la couche de sortie *i.e.* la couche de softmax pour une tâche de classification), plus elle capture des constructions visuelles de haut niveau. Ainsi en se concentrant sur les informations capturées par les dernières couches convolutives, il extrait des informations sémantiques spécifiques à chacune des classes, et de par la nature spatiale de l'opération de convolution, détermine les zones d'attention d'un réseau de neurones profond. Pour ce faire, il utilise les informations qui circulent dans la dernière couche de convolution : les cartes d'activation obtenues lors de la propagation et les gradients obtenus lors de la rétropropagation.

Les gradients sont utilisés pour calculer l'importance de chacune des cartes d'activation A_k en calculant les scalaires α_k^c. Ce coefficient est spécifique à une classe c et à une carte d'activation k. Ainsi, on calcule le gradient de la sortie y^c avant la couche de softmax par rapport à la carte d'activation A^k. Les gradients obtenus sont ensuite moyennés de façon globale (*Global Average Pooling* en anglais) comme il l'est décrit dans l'équation. 2.1

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\substack{\text{Moyenne}\\ \text{globale}}} \overbrace{\frac{\partial y^c}{\partial A_{ij}^k}}^{\text{Gradient}} \tag{2.1}$$

Une fois les coefficients d'importance α_k^c des cartes d'activation A^k calculés, on peut obtenir la carte $\mathcal{L}_{Grad-CAM}^c$ pour une classe c donnée en réalisant une combinaison pondérée entre ces cartes d'activation et leur coefficient d'importance. Afin de n'obtenir que les cartes contribuant de façon positive à la classe cible c, une ReLU est appliquée à cette combinaison pondérée. Ainsi, grâce à l'utilisation d'une ReLU, on ne capture que les caractéristiques ayant contribué à augmenter la sortie y^c en ignorant celles l'ayant diminuée. Les caractéristiques contribuant négativement à une classe sont, elles, trop susceptibles d'appartenir à une autre classe. On obtient ainsi l'équation suivante 2.2 :

$$\mathcal{L}_{Grad-CAM}^c = \overbrace{ReLu}^{\substack{\text{Seulement l'influence}\\ \text{Positive}}} \Big(\sum_k \overbrace{\alpha_k^c}^{\substack{\text{Coefficient}\\ \text{d'importance}}} \overbrace{A_k}^{\substack{\text{Cartes de}\\ \text{Caractéristiques}}}\Big) \tag{2.2}$$

Ainsi chaque carte représentera la combinaison pondérée des activations ayant contribué de façon positive à la décision finale. Les coefficients α_k^c représentent une linéarisation partielle du réseau en aval de A et capturent l'importance d'une carte de caractéristique k pour une classe cible c donnée.

La taille des cartes obtenues dépend donc de la taille d'entrée et de la profondeur du

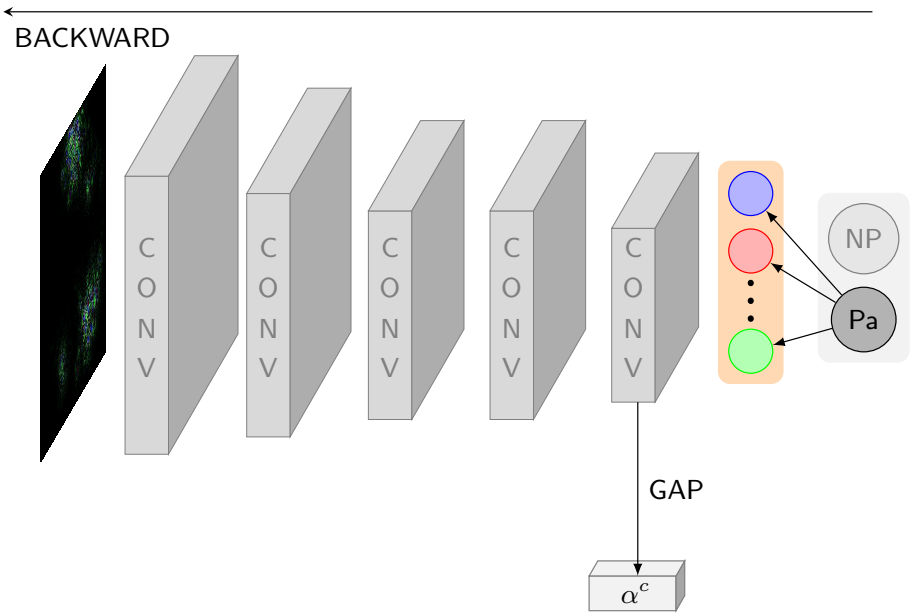


FIGURE 2.5 – Schéma de l'obtention des cartes des gradients lors de la rétropropagation, permettant le calcul des coefficients d'importances α_k^c

volume de caractéristiques cible. En effet, ces cartes possèdent la même résolution que les cartes d'activation ayant permis leur calcul et sont donc plus petites lorsque le volume cible est profond du fait de l'utilisation des max-pooling dans les réseaux de neurones profonds à convolution. Afin de pouvoir visualiser les zones d'attention sur l'image d'origine, elles sont souvent remises à l'échelle par interpolation bilinéaire, ce qui induit que leurs zones d'attention soient plus grossières.

Algorithme par rétropropagation

Dans les algorithmes par rétropropagation, on peut citer les deux plus célèbres, proches entre eux dans l'idée, mais qui possèdent quelques différences : l'algorithme dit de Gradients (ou également appelé de rétropropagation simple) [182] et l'algorithme de rétropropagation guidée [183]. Ces méthodes s'appliquent comme les autres techniques d'extraction de l'attention post-hoc à des réseaux déjà entraînés et ne modifient pas leur poids. Elles se basent sur la même idée : calculer le gradient de la prédiction du réseau par rapport à l'entrée en maintenant les poids fixes. Cela permet de déterminer quels éléments d'entrée (par exemple, quels pixels dans le cas d'une image d'entrée) doivent être

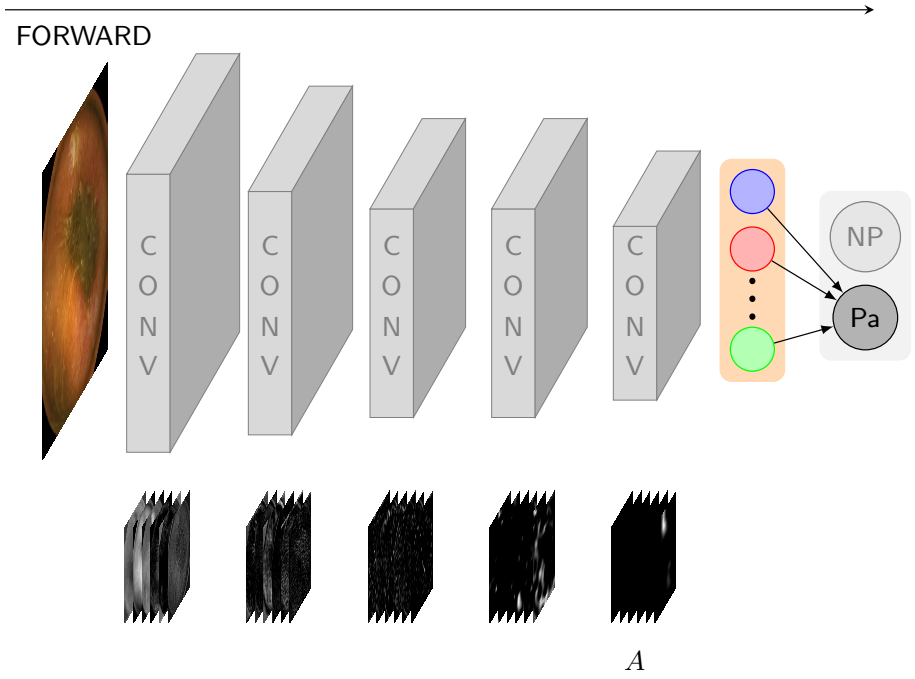


FIGURE 2.6 – Schéma de l'obtention des cartes de caractéristiques lors de la propagation

modifiés le moins possible pour affecter le plus la prédiction. La différence entre ces trois méthodes réside dans la façon dont sont calculés les gradients et comment se déroule la rétropropagation à travers le réseau.

L'idée de ces méthodes est de trouver la matrice de poids du réseau w_c associant à chaque pixel d'une entrée I une fonction de score $S_c(I)$. Ainsi, on peut définir avec b_c les biais du réseau :

$$S_c(I) = w_c^T I + b_c$$

Le problème avec les réseaux de neurones profonds est que l'équation de la fonction de score $S_c(I)$ est une fonction hautement non linéaire de I. Cependant en utilisant un développement de Taylor de premier ordre, au voisinage d'une image donnée I_0, on peut approximer l'équation précédente de la façon suivante :

$$S_c(I) \approx w_c = \frac{\partial I}{\partial S_c}\bigg|_{I_0}^T I + b,$$

avec w la dérive de S_c par rapport à I au point de fonctionnement I_0.

D'après Simmonyan : «Une autre interprétation du calcul de la saillance de classe spécifique à l'image à l'aide de la dérivée du score de classe est que la magnitude de la dérivée indique quels pixels doivent être modifiés le moins possible pour affecter le plus le score de classe. On peut s'attendre à ce que ces pixels correspondent à l'emplacement de l'objet dans l'image.» On peut aussi voir cela comme les pixels dont l'amplitude de la dérivée de la fonction de score par rapport à une image donnée I_0 est la plus grande sont ceux ayant le plus d'influence sur la décision finale, et donc ceux permettant au réseau de prendre sa décision.

Gradients

Dans la méthode des gradients [182], aucun changement n'est appliqué au réseau et on rétropropage les gradients jusqu'à la couche d'entrée. Une fois la carte de gradients obtenue, on calcule la valeur absolue élément par élément de cette carte de même dimension que l'image d'entrée. Si l'image possède 3 canaux (RGB), on prend le maximum de chacun des canaux. Des exemples de résultats issus de l'article d'origine de Simmonyan et al. sont présents figure 2.7.

FIGURE 2.7 – Exemples de résultats obtenus par la méthode Gradients issue de l'article d'origine de Simmonyan et al. 2014 [182]

Rétropropagation guidée

Dans la méthode de rétropropagation guidée, comme dans grad-CAM, on souhaite observer seulement l'influence positive des poids sur une classe afin d'éviter que les gradients d'une autre classe viennent interférer avec la classe cible. Seule l'influence positive des pixels d'entrée sur la classe cible c est conservée. Les gradients de la sortie par rapport à l'entrée sont calculés en ignorant tous les gradients négatifs. Cela se traduit dans l'implémentation de la méthode par l'application d'une ReLU sur les gradients lors de la rétropropagation. On rappelle ici que bien que l'architecture du réseau soit modifiée en remplaçant les ReLu pour des ReLu à rétropropagation guidée, les poids et les biais restent inchangés.

Grâce à cet algorithme, on obtient une carte de saillance, spécifique à la classe cible c avec les mêmes dimensions, et le même nombre de canaux que l'image d'entrée. Si l'image en entrée possède 3 canaux (RGB), on prend alors la valeur maximale de la carte de gradients sur les 3 canaux.

Guided GradCam

L'algorithme Guided GradCAM [184] est, comme son nom le laisse entendre, une combinaison entre l'algorithme de rétropropagation guidée et GradCAM. Ainsi pour calculer les coefficients d'importance α_k^c dans la couche cible k pour une classe donnée c, on n'utilise que des gradients positifs lors de la rétropropagation en appliquant une ReLU au signal d'erreur comme dans l'algorithme de rétropropagation guidée. La carte calculée à l'aide des cartes d'activation de la couche cible pondérée par les coefficients d'importance α_k^c sera utilisée comme un masque pour la carte obtenue par rétropropagation guidée. Cette technique présente l'intérêt de prendre en compte les cartes d'activation calculées pendant la propagation, tout en gardant une bonne résolution puisque la carte finale est de taille identique à celle d'entrée comme dans la technique de rétropropagation guidée.

2.8 Conclusion

Dans cette partie, nous avons pu voir comment l'enjeu de l'interprétabilité était central dans le développement et le déploiement des algorithmes d'apprentissage profond et en particulier lorsque ces derniers sont appliqués dans des domaines critiques tels que la santé. Afin de pouvoir se fier à leur décision, les utilisateurs ont besoin de mieux interpréter leur

processus décisionnel.

Afin de pouvoir mieux comprendre les mécanismes décisionnels humains, de nombreuses études ont été menées sur le processus attentionnel humain, clé de voûte de notre prise de décision. Le sens prédominant étant dans la majorité des situations la vue, de nombreuses technologies et méthodes ont été développées pour comprendre son fonctionnement et comment l'attention le régissait. Dans l'imagerie médicale, la recherche en attention visuelle peut permettre d'aider la formation des nouveaux médecins et de mieux comprendre les conditions optimales conduisant à de meilleures performances de classification.

C'est ainsi que la recherche en attention artificielle a émergé depuis les 10 dernières années afin de rendre les algorithmes d'apprentissage profond plus interprétables en s'inspirant de l'attention humaine. On distingue deux grandes familles d'algorithmes, les algorithmes à attention apprise et les algorithmes à attention post-hoc permettant à l'utilisateur de mieux comprendre les éléments induisant la décision des réseaux, et ainsi de réduire le risque de biais par exemple ou encore de permettre à l'utilisateur de combiner plus facilement ses connaissances avec celles de l'algorithme afin de prendre la décision optimale.

Dans notre cas, nous essayons d'aider les gastro-entérologues dans la tâche d'identification des lésions de la maladie de Crohn à partir d'images issues de vidéos capsules endoscopiques. L'utilisation de l'attention artificielle pourrait être une réelle plus-value pour assister les médecins dans leur diagnostic. Ces méthodes pourraient fournir des informations complémentaires à la prédiction permettant d'assurer les gastro-entérologues de la fiabilité des algorithmes d'aide au diagnostic. Les biais d'apprentissage pourraient également être limités. Lors du développement d'une des premières versions du jeu de données CrohnIPI, les images pathologiques provenaient d'examens réalisés avec un type de caméra quand les images non pathologiques, elles, provenaient d'un autre type de caméra. Cela impliquait des taux de reconnaissance excellents sur la base de test alors que le réseau n'avait appris qu'à reconnaître le type de caméra plutôt que les lésions de la maladie de Crohn, rendant impossible son utilisation dans la routine clinique des médecins. Ce genre de biais peut également se retrouver dans d'autres bases de données, d'autant plus quand le processus d'annotation n'est pas clairement détaillé. Ainsi avec des méthodes basées sur l'attention, il est possible de vérifier que les zones de l'entrée permettant à l'algorithme de réaliser sa prédiction concorde bien avec celles utilisées par les experts médicaux.

Chapitre 3

L'attention artificielle dans des images issues de vidéos capsules endoscopiques

Sommaire

L'attention dans les réseaux de neurones profonds est une source d'informations permettant de mieux comprendre les prédictions des réseaux. Dans cette partie, nous étudierons plusieurs applications de ces méthodes attentionnelles à notre problématique d'aide au diagnostic de la maladie de Crohn. Dans un premier temps, l'architecture d'un réseau

à attention dure que nous avons construit sera introduite dans la section 3.1, ainsi que les résultats obtenus sur la base de données CrohnIPI présentée dans la partie 1.4. Dans un second temps, les résultats obtenus pour différentes méthodes d'extraction post-hoc sur différents réseaux de neurones de l'état de l'art entraînés sur la base de données CrohnIPI seront expliqués dans la section 3.2. Finalement une comparaison entre les différentes méthodes post-hoc sera réalisée à l'aide des métriques de saillance présentées dans la partie 2.3.3.

3.1 Un réseau récurrent à attention dure

Ici, l'attention est considérée comme un processus de décision séquentiel d'un agent interagissant avec un environnement visuel. Le réseau est basé sur celui présenté par Mnih et al. 2014 [170] et sur l'algorithme REINFORCE de Williams 1992 [185]. À partir de ce réseau de base, l'extraction des caractéristiques locales a été améliorée grâce à un réseau VGG16 pré entraîné. Le processus de mémorisation a également été amélioré par l'ajout d'une cellule récurrente à mémoire court et long terme (ou *Long short-term memory* en anglais couramment abrégé LSTM). Nous avons ajouté la possibilité pour le réseau de choisir la taille de son champ de vision grâce à de l'apprentissage par renforcement.

3.1.1 Architecture du réseau

L'architecture de ce réseau est composée de trois sous-parties, le *Réseau de regard*, *la cellule récurrente* et le *Réseau de décision* et est présentée dans la figure 3.1. Le réseau de regard permet d'extraire les caractéristiques d'une partie de l'image. La cellule (notée LSTM sur le schéma de l'architecture) récurrente permet au réseau de se « rappeler » des précédents regards et transmet au réseau de décision les caractéristiques nécessaires pour réaliser sa prédiction ainsi que pour choisir la prochaine partie de l'image à observer.

Une image endoscopique X est fournie en entrée du réseau. Le *Capteur de regard* va alors extraire une partie rectangulaire de l'image, un patch, $\rho(X)$ de l'image originale selon les paramètres $l_t = (x, y, z)$ où les coordonnées x et y sont normalisées dans l'intervalle $[-1; 1]$, avec $(0, 0)$ le centre de l'image, et avec z le coefficient de zoom normalisé entre entre $]0; 1]$. Avec $s_x \times s_y$ la résolution de l'image originale, on obtient un patch de résolution $(s_x \times z) \times (s_y \times z)$ et centré en (x, y). Ce patch est ensuite redimensionné pour conserver une taille fixe à l'entrée du réseau. Le sous-réseau *« Quoi ? »*, basé sur VGG16

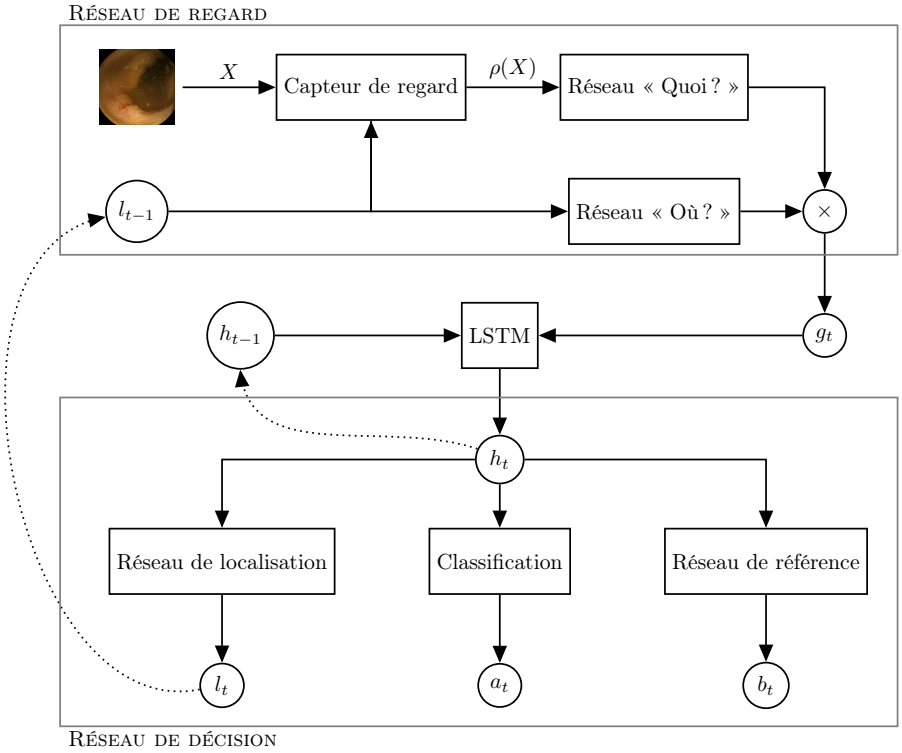


FIGURE 3.1 – **Architecture du réseau de neurones récurrent à attention** : À chaque instant t, nous fournissons au *Capteur de regard* une image endoscopique X et la localisation l_{t-1} du patch à extraire de l'image originale. Deux réseaux neuronaux indépendants, les réseaux *« Quoi ? »* et *« Où ? »*, vont ensuite extraire les informations relatives au contenu et à l'emplacement du patch. Une mémoire court et long terme (LSTM) fusionne ensuite les caractéristiques précédemment extraites par le réseau pour produire l'état actuel du système h_t. À partir de cet état, trois sous-réseaux produisent indépendamment l_t, la position du prochain patch à extraire, a_t un vecteur contenant un score associé à chaque classe, et b_t la *baseline* (ou récompense de référence) à partir de laquelle est calculée la récompense pour l'apprentissage par renforcement.

[182] avec normalisation par batch, pré entraîné sur ImageNet [186], permet d'extraire les caractéristiques du patch $\rho(X)$. Seules les 12 premières couches du VGG16 et la première couche entièrement connectée ont été conservées. En parallèle, les informations relatives à l'extraction du patch l_t traversent le *réseau « Où ? »* composé de 2 couches entièrement connectées, permettant ainsi l'extraction des caractéristiques relatives à la position du patch $\rho(X)$. Les deux vecteurs caractéristiques produits par ce réseau sont de taille identique et multipliés avant d'être soumis à la non-linéarité ReLU. Le nouveau vecteur

caractéristique g_t en sortie de la non-linéarité contient alors les informations *« Où ? »* et *« Quoi ? »* extraites par le *Réseau de regard* au temps t. Une cellule récurrente à mémoire court et long terme (LSTM) [187] permet de fusionner les caractéristiques extraites au temps t par le réseau avec celles extraites au temps précédent contenues dans l'état interne précédent h_{t-1} de la LSTM. Cet état interne de la LSTM sera réutilisé au prochain pas de temps. Ce sous-réseau prenant en entrée une image X et produisant en sortie un vecteur de caractéristiques g_t est appelé *Réseau de regard*. Son architecture détaillée est présentée dans la figure 3.2.

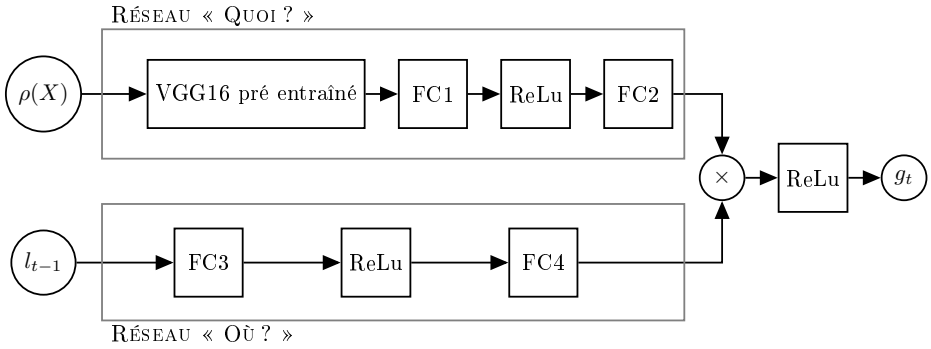


Figure 3.2 – Le réseau de regard est composé de deux sous-réseaux, le réseau *« Quoi ? »* et le réseau *« Où ? »*. Le réseau *« Quoi ? »* est composé d'un réseau convolutif, VGG16, pré-entraîné sur la base de données ImageNet et de deux couches complètement connectées *FC1* et *FC2*. Ce réseau permet d'encoder les informations nécessaires à la décision qui sont contenues dans le patch extrait $\rho(X)$. Le réseau *« Où ? »* est lui composé de deux couches complètement connectées *FC3* et *FC4* et permet d'encoder les informations relatives à la localisation d'extraction du patch $\rho(X)$ aux coordonnées l_{t-1}. Les deux vecteurs de tailles équivalentes produits par ces réseaux sont ensuite combinés par multiplication afin de produire le vecteur g_t.

À partir du nouvel état interne produit par le LSTM, le *Réseau d'action* va produire un vecteur associant un score à chaque classe. Le *Réseau de référence* va lui permettre de calculer la récompense associée à une prédiction afin de pouvoir entraîner le *Réseau de localisation* par renforcement. L'apprentissage par renforcement est non supervisé et associe à chaque action du réseau une récompense qui doit être maximisée lors de l'optimisation du réseau. Ainsi, le réseau augmentera la probabilité que les emplacements d'extraction des patchs maximisent la fonction de récompense. Si le réseau classe correctement l'image, à la suite des multiples regards dans l'image, la récompense vaut le nombre de vues placées sur l'image moins la récompense de référence calculées par le réseau.

3.1.2 Fonction de coût et optimisation

La fonction de coût du réseau est une somme de trois sous-fonctions de coût $\mathcal{L}_a$, $\mathcal{L}_b$ et $\mathcal{L}_r$. Tout d'abord, la fonction d'entropie croisée $\mathcal{L}_a$ calcule l'erreur de classification du réseau à la dernière itération de regard sur l'image avec la prédiction du réseau $\hat{Y}_i$ et la vérité terrain Y_i pour N images :

$$\mathcal{L}_a = -\frac{1}{N}\sum_{i=1}^{N} Y_i \log(\hat{Y}_i) \tag{3.1}$$

Ensuite, la fonction $\mathcal{L}_b$ calcule l'erreur quadratique moyenne entre la récompense R_t^i et la récompense de référence b_t établie par le réseau (3.2). À chaque regard sur l'image, une prédiction est réalisée par le réseau d'action. La fonction récompense R_t^i vaut alors 1 si la prédiction est correcte et 0 sinon. Ainsi, le réseau est encouragé à réaliser des prédictions correctes à chaque itération, bien que seule la dernière prédiction $\hat{Y}_i$ soit utilisée pour le calcul de l'erreur d'entropie croisée. Cette *récompense de référence*, en fonction des contextes présents et passés, permet d'ajuster la récompense afin de pousser le réseau à améliorer ses résultats par rapport aux précédentes itérations d'optimisation.

$$\mathcal{L}_b = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(b_t - R_t^i)^2 \tag{3.2}$$

La fonction $\mathcal{L}_r$ permet de mettre en place l'apprentissage par renforcement non supervisé (3.3). Le principe est le suivant : une stratégie $\pi(\tau^j;\theta)$, dépendant des paramètres du réseau θ, associe une probabilité à la trajectoire τ^j. Une trajectoire est une succession d'actions u_t, dans notre cas le changement de l'emplacement de l'extraction du patch. Chaque action résulte en un nouvel état s_t, dans notre cas un nouveau patch à l'entrée du système. Ce nouvel état influence ensuite le système qui produira une nouvelle action qui conduira à un nouvel état. L'objectif de l'algorithme de renforcement est de maximiser la probabilité des trajectoires qui maximisent la récompense $(R_t - b_t)$ et *a contrario* de diminuer la probabilité des trajectoires menant à une faible récompense. Nous utilisons la méthode de sur-échantillonnage de Monte-Carlo pour estimer le rendement attendu d'un état en faisant la moyenne des récompenses pour M tirages d'une stratégie $\pi(\tau^j;\theta)$.

$$\mathcal{L}_r = -\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\log\pi(u_t^j|s_{i:t}^j;\theta)(R_t^j - b_t) \tag{3.3}$$

3.1.3 Entraînement

L'entraînement du réseau de neurones récurrents à attention a été réalisé avec la même répartition des données que l'entraînement des réseaux présentés dans la partie 1.5. Pour rappel, 70% des images sont allouées à l'entraînement, 10% à la validation et 20% au test. Une validation croisée en cinq étapes a été réalisée de sorte que chacune des images soit utilisée une unique fois pour la phase de test. L'implémentation du réseau de neurones a été réalisée sur GPU permettant ainsi un entraînement plus rapide.

Pour chacune des expériences, le réseau extrait 4 patchs par image avec une résolution de 80×80 pixels. Bien que la résolution soit fixe, ils peuvent correspondre à des zones de l'image plus grandes que 80×80 pixels du fait de l'ajout du coefficient de zoom z. Ainsi le réseau peut extraire des patchs plus grands, mais ces derniers seront redimensionnés en 80×80 pixels avant d'être traités par le réseau de neurones convolutif.

L'entraînement se poursuit tant que le réseau n'a pas atteint ses performances maximales sur le jeu de validation. Le critère d'arrêt de l'entraînement est défini comme une diminution des performances sur ce jeu de validation pendant 50 epochs successives.

3.1.4 Performances

Comme pour les entraînements réalisés avec ResNet 34, VGG16 et VGG19, les réseaux ont été entraînés sur chacun des jeux de données correspondant aux différentes phases d'annotation présentées dans la partie 1.4.5, selon la même méthode que celle présentée dans la partie 1.5.1. Les résultats obtenus par le réseau de neurones présentés précédemment sont consignés dans le tableau 3.2.

On peut observer la même dynamique au niveau de l'évolution que pour les 3 réseaux de l'état de l'art en terme de précision au cours des différentes phases d'annotation. On peut remarquer que l'amélioration de la qualité des labels améliore principalement la spécificité. La spécificité indique la probabilité qu'un test négatif soit correct. Ainsi, plus les labels sont de qualité, plus le réseau devient performant à repérer les images non pathologiques. Une interprétation possible est que la réduction du nombre de faux négatifs dans la base de données d'entraînement permet au réseau d'être capable de mieux généraliser ses connaissances. La sensitivité du réseau, soit la proportion d'images pathologiques correctement classées, elle, semble rester constante.

Lorsque l'expérience est réalisée avec pour objectif de prévoir le label exact de l'image, en utilisant la totalité des images de la base, on obtient une performance de 83,00% et

	Précision	**Spécificité**	**Sensitivité**	**F1 score**
Phase 1 (Étiquettes du sélecteur)	90.90	91.70	90.06	90.56
Phase 2 (Étiquettes des experts indépendants)	91.83	94.00	88.45	89.43
Phase 3 (Étiquettes du consensus)	92.48	95.24	88.16	90.15
Phase 3 multi labels (Étiquettes du consensus)	92.07	96.84	84.63	89.29

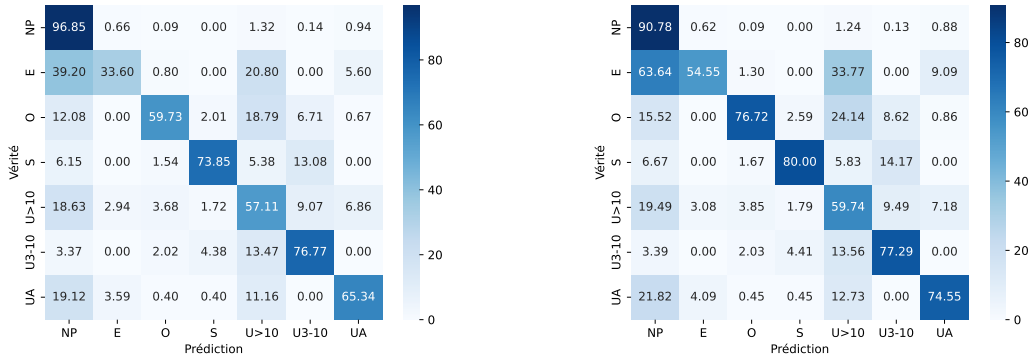
TABLE 3.1 – Résultats de classification obtenus avec le réseau récurrent à attention dure, sur les quatre versions d'étiquettes en utilisant l'entièreté de la base de données (3498 images).

	Précision	**Spécificité**	**Sensitivité**	**F1 score**
Phase 1 (Étiquettes du sélecteur)	88.53	90.67	86.24	87.89
Phase 2 (Étiquettes des experts indépendants)	92.37	93.31	90.96	90.49
Phase 3 (Étiquettes du consensus)	93.70	94.76	92.09	92.09
Phase 3 multi labels (Étiquettes du consensus)	92.64	96.95	86.14	90.32

TABLE 3.2 – Résultats de classification obtenus avec le réseau récurrent à attention dure, sur les quatre versions d'étiquettes avec le même nombre d'images pour chacune des phases en retirant les images non concluantes dans tous les sous jeux de données (3331 images).

92,64% lorsque l'on regroupe les images classées comme pathologiques dans une même classe, bien que le réseau ait été entraîné à faire la différence entre les pathologies. La matrice de confusion associée à ces résultats est présentée dans la figure 3.3.

On peut voir que le réseau réalise une bonne distinction des images non pathologiques. Au niveau de l'identification des lésions, il obtient les meilleures performances pour l'identification des sténoses, avec une précision de 80,00% et un rappel de 73,85%. À l'inverse, il est difficile pour lui d'identifier les érythèmes avec un rappel de 33,6% et une précision de 54,55%. Ce résultat parait logique. La sténose étant considérée comme la plus sévère des



(a) Matrice de confusion normalisée par le nombre d'images par classes.

(b) Matrice de confusion normalisée par le nombre de prédictions par classes.

FIGURE 3.3 – Matrices de confusion liées à l'entraînement du réseau sur l'ensemble de la base de données après la phase de consensus. La diagonale de la matrice (a) indique le rappel pour chacune des classes. La diagonale de la matrice (b), elle, indique la précision pour chacune d'elles.

lésions dues à la maladie de Crohn, elle est plus facile à identifier du fait qu'elle couvre de larges aires de l'image tandis que l'érythème, étant la lésion la moins sévère, souvent localisée dans une faible portion de l'image, est plus difficile à identifier correctement. De plus, les labels identifient la lésion la plus sévère de l'image, ce qui n'implique pas qu'elle soit la seule présente. La présence d'une lésion sévère implique souvent également la présence d'un érythème. Ainsi il est plus difficile pour le réseau de faire la distinction entre un érythème seul et un érythème accompagné d'une autre lésion, rendant la tâche de généralisation du réseau de neurones plus complexe.

3.1.5 Résultats visuels

Dans la figure 3.3, on peut trouver des exemples de cartes d'attention de l'algorithme récurrent. Afin d'obtenir des résultats plus stables et de compenser la faible quantité de données utilisées pour l'entraînement, l'état caché de la cellule du réseau récurrent a été initialisé de façon différente et aléatoire bien que les différentes copies du réseau réalisées par sur-échantillonnage de Monte-Carlo possèdent des poids identiques. En effet, les algorithmes utilisant de l'apprentissage par renforcement sont réputés pour être peu stables, du fait de la conception probabiliste de la fonction de coût de ces derniers. Ici, l'initialisation de la cellule induit directement le placement du premier regard de l'image,

depuis lequel découlent les 3 autres regards. Le sur-échantillonnage de Monte-Carlo est utilisé lors des phases de test et de validation.

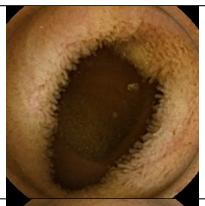
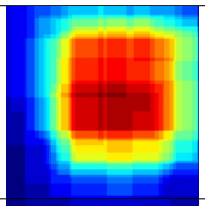
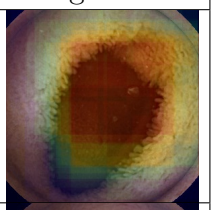
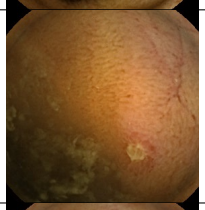
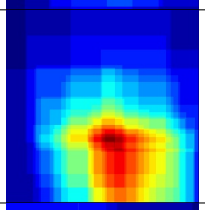
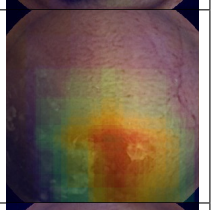
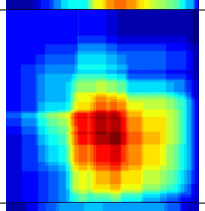
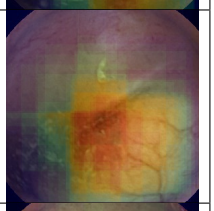
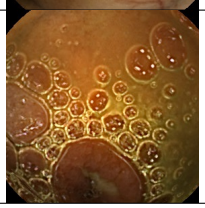
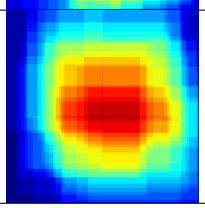
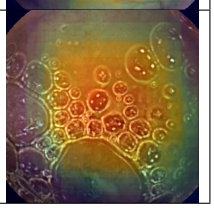
Vérité	Prédiction	Originale	Saillance artificielle	Saillance artificielle et image originale
NP	NP			
P	P			
NP	P			
P	NP			

TABLE 3.3 – Exemples de cartes de saillance artificielle en fonction du résultat de la détection (avec P pour la classe pathologique et NP pour la classe non pathologique)

3.1.6 Discussion

Les résultats obtenus par ce réseau sont encourageants. Ils montrent qu'il est possible d'obtenir des résultats proches de ceux des réseaux traditionnels avec une architecture plus proche d'un fonctionnement séquentiel comme celui des humains, auquel on peut assimiler ce processus de saccades et fixations présentées dans la partie 2.3. Bien que pour chaque image, des parties de l'image aient été complètement ignorées par le réseau, les résultats restent bons. L'amélioration des performances sur les différents jeux de données

correspondant aux différentes phases, visible avec les réseaux de l'état de l'art, reste encore vérifiable, confirmant un peu plus l'idée que les différentes phases d'annotations ont contribué à consolider le jeu de données CrohnIPI et à améliorer les labels.

Les performances du réseau sont les résultats d'un important travail de recherche de méta paramètres. Il existe des méta paramètres traditionnels liés à l'entraînement des réseaux à convolution, comme le taux d'apprentissage, le nombre de couches cachées, pour n'en citer que quelques-uns. En plus de ceux-ci, on peut citer ceux plus spécifiques à cette architecture comme le type de cellule récurrente, le nombre de réseaux clonés, le nombre de regards par image, l'initialisation du premier regard, la résolution des patchs extraits, la définition de la distribution permettant l'échantillonnage des positions des regards, et la taille de toutes les couches cachées permettant de faire les liens entre les différents sous-réseaux. Les expériences menées pour la recherche de ces méta paramètres ne sont pas présentées dans cette partie du fait d'un protocole de recherche par tâtonnement basé sur l'essai/erreur plutôt que sur une recherche exhaustive ou optimisée. Afin d'améliorer ce réseau, il serait possible d'imaginer l'utilisation de méthodes d'exploration des méta paramètres. On peut également trouver dans la littérature un réseau ayant été publié à la même période que le nôtre, proposant l'utilisation d'une cellule récurrente spécialement conçue pour faciliter l'optimisation du modèle [188].

3.2 Résultats visuels de l'attention post-hoc

Dans cette partie seront exposés les résultats obtenus par les différentes méthodes d'extraction de l'attention post-hoc présentées dans la partie 2.7.2, sur trois réseaux de neurones convolutifs de l'état de l'art : ResNet34, VGG16 et VGG19. Les méthodes d'extraction de l'attention sont les suivantes : l'algorithme des gradients, l'algorithme de rétropropagation guidée, Grad-CAM et Guided Grad-CAM. Une fois les post-traitements décrits, différents exemples d'attention seront présentés et discutés. L'objectif de cette partie est de se faire une idée du fonctionnement attentionnel artificiel en fonction des prédictions réalisées et de leur justesse. Les résultats obtenus sans traitement *a posteriori* sont présentés dans l'annexe A.

Post-traitements inspirés par l'humain

Bien que les post-traitements originaux permettent de mettre en valeur des formes et des contours d'objets détectés par le réseau, les cartes obtenues ne ressemblent pas

à des cartes de saillance humaine. Dans le troisième chapitre, nous nous concentrerons sur la comparaison entre les cartes d'attention humaine et artificielle. Afin de permettre cette comparaison, les post-traitements des cartes de gradients ont été modifiés. Comme présentées dans la partie 2.3.2, les cartes de saillance humaine sont construites à partir de cartes de fixations binaires. Ces cartes de fixations répertorient l'ensemble des pixels fixés par le sujet humain lors de la visualisation de l'image et enregistrés à l'aide d'un *eye-tracker*. Une fois ces cartes de fixations binaires obtenues, elles sont convoluées avec un noyau gaussien afin d'obtenir des cartes continues, appelées cartes de saillance. Cette opération est réalisée pour prendre en compte les imprécisions dues aux erreurs de mesure de l'*eye-tracker* et également car les humains ne regardent jamais qu'un seul pixel à chaque regard. Nous avons donc considéré les cartes de gradients comme des cartes de fixations et convolué ces dernières avec un noyau gaussien.

Dans le post-traitement traditionnel, lorsque l'image possède 3 canaux d'entrée (RGB) et donc que la carte de gradients de sortie en possède également 3, seule la valeur maximale des trois canaux pour chaque pixel est conservée. De notre point de vue, cette opération implique une perte d'information utile pour comprendre les zones d'attention de la machine. Ainsi, nous avons calculé la somme des trois canaux et finalement normalisé par le maximum.

3.2.1 Résultats visuels obtenus par les méthodes à gradients

Dans cette partie seront présentés des résultats visuels obtenus par les différentes méthodes de visualisation à gradient présentées précédemment (partie 2.7.2), à savoir la méthode des gradients (ou méthode de rétropropagation), la méthode de rétropropagation guidée, GradCAM et Guided GradCAM. Pour chacune de ces méthodes, les trois réseaux présentés dans la partie 1.5.1 ont été entraînés sur la totalité de la base de données CrohnIPI, en utilisant 80% de la base pour l'entraînement et 20% pour la validation. Chacun des réseaux a été entraîné 5 fois selon 5 découpages différents de la base, de sorte que chacune des images soit utilisée une unique fois en validation (validation croisée). Une fois cette étude qualitative présentée, une étude quantitative sera réalisée dans la partie 3.3. Les tests ont été réalisés sur 250 images hors de la base CrohnIPI et annotés comme décrit ci-après.

Les labels des images présentées ci-après sont issues d'une expérience impliquant 10 experts gastro-entérologues et 5 internes en médecine en phase d'apprentissage. Le processus d'acquisition de ces labels est décrit en détail dans la partie 4.1.4. Pour chacune

des images, le label est défini comme le vote majoritaire entre la classe pathologique et la classe non pathologique. La possibilité était donnée aux participants de s'abstenir de répondre en choisissant le label "Non concluante". Ainsi pour chacune des images, 15 humains ont voté pour définir le label de chaque image et la réponse de 15 classifieurs a été enregistrée (3 modèles avec 5 validations croisées pour chacun des modèles). On rappelle ici que les différents classifieurs ont été entraînés en utilisant la totalité du jeu de données CrohnIPI avec 80% des données pour l'entraînement et 20% pour la validation.

Ainsi, dans cette partie seront présentés différents résultats nous permettant de mieux comprendre les décisions des différents classifieurs en fonction de la justesse de leur prédiction. Premièrement nous verrons les résultats obtenus lorsque l'image est un vrai positif, ensuite lorsqu'elle est un vrai négatif, lorsque le label de l'image est non concluant, puis lorsque le classifieur commet une erreur.

Vrai positif

Premièrement, nous allons observer et comparer visuellement les résultats obtenus sur un vrai positif, lorsque les classifieurs sont en accord avec la vérité terrain associée à l'image. Pour l'exemple du tableau 3.4 on obtient 14 votes humains pour la classe pathologique et une abstention. Au niveau des classifieurs on obtient 14 classifieurs ayant répondu correctement et un classifieur ayant mal répondu. Dans ce tableau sont répertoriés les résultats d'attention post-hoc de quatre méthodes sur 3 réseaux ayant correctement répondu.

On peut observer que sur cet exemple la méthode des gradients semble rendre compte de zones attentionnelles plus larges. Bien que GradCAM semble également couvrir de larges zones, il faut garder à l'esprit que la carte d'origine est une matrice de 8×8 pixels redimensionnée en 320×320 par interpolation bilinéaire ayant pour conséquence d'étaler les zones d'attention. Bien que les modèles soient identiques, les zones d'intérêt des différentes méthodes ne semblent pas identifier la même région comme celle ayant le plus contribué à la décision finale. Pour ResNet34 par exemple, on peut observer que les GradCAM, guided-GradCAM, et rétro propagation guidée semblent identifier la zone située en bas à droite de l'image comme la plus importante, quand la méthode des gradients semble identifier la zone en bas à gauche.

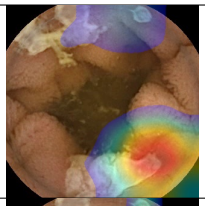
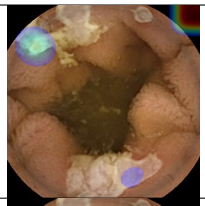
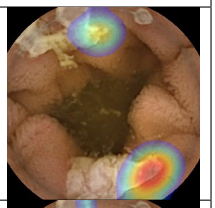
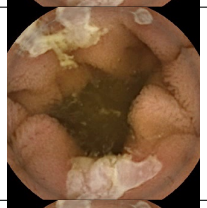
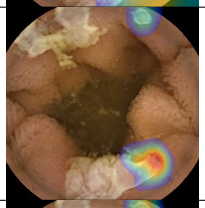
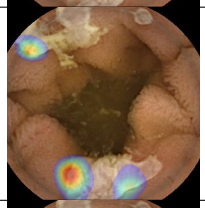
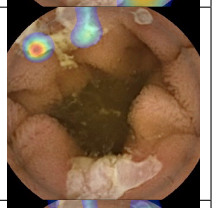
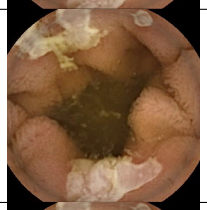
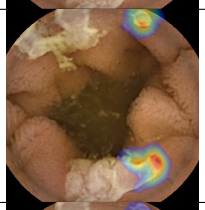
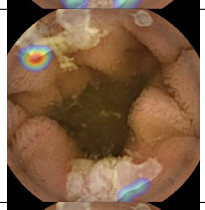
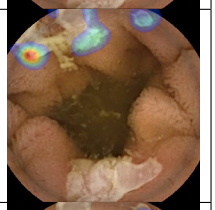
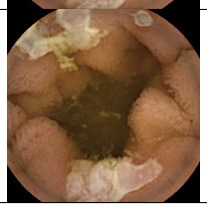
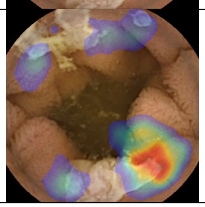
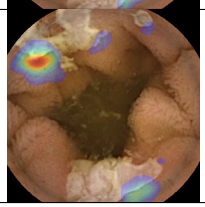
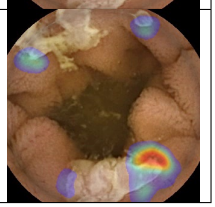
Méthode post-hoc	Originale	ResNet34	VGG16	VGG19
GradCAM				
Guided-GradCAM				
Rétro propagation guidée				
Gradients				

TABLE 3.4 – Cartes d'attention artificielle post-hoc obtenues avec 4 méthodes présentées dans la partie 2.7.2 pour 3 réseaux de neurones profonds de l'état de l'art. L'exemple choisi est un **vrai positif** avec la majorité des classifieurs et la majorité des experts votant ayant considéré l'image comme pathologique.

Vrai négatif

Nous allons maintenant observer les résultats visuels obtenus par ces différentes méthodes lorsque l'image est non pathologique et que les réseaux répondent correctement. Sur l'image suivante, 14 votes ont été enregistrés pour la classe non pathologique et une abstention. Au niveau des prédictions des classifieurs, les 15 ont correctement prédit la classe non pathologique. Les résultats sont visibles dans le tableau 3.5.

On peut observer que pour les 3 premières méthodes, les zones de détection prédites par ces dernières sont majoritairement concentrées sur les bulles, lorsque les zones prédites par la méthode des gradients indiquent au contraire les zones sans bulle et ce pour les 3 classifieurs. La présence de bulles dans l'intestin grêle est naturelle et ces zones sont rarement significatives pour la prise de décision, du fait que ces dernières cachent les muqueuses de l'intestin sur lesquelles les lésions de la maladie de Crohn sont visibles. En

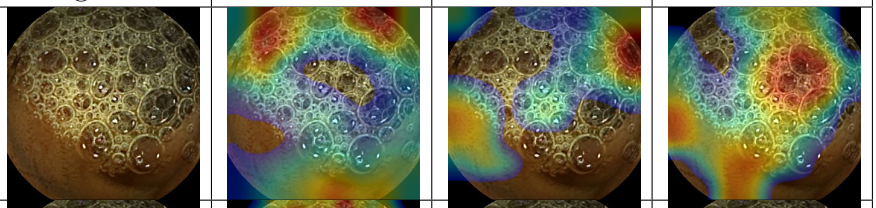
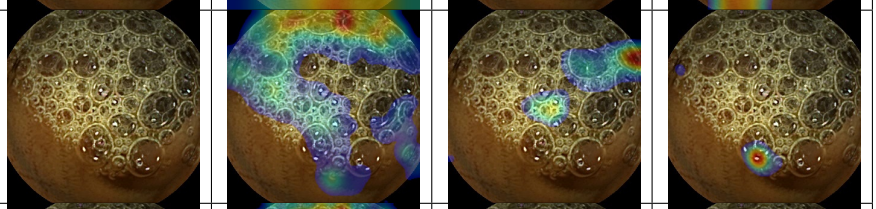
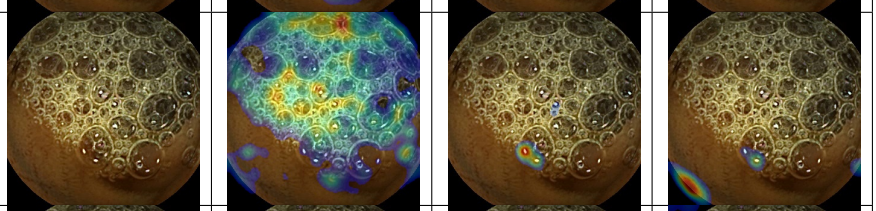
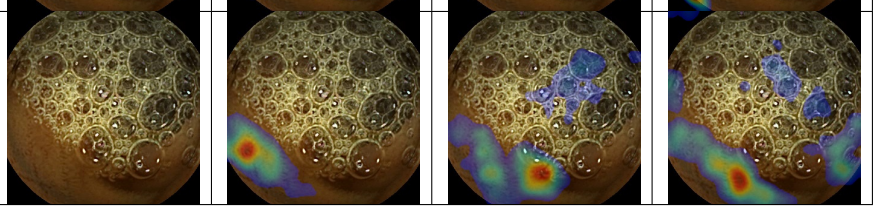
Méthode post-hoc	Originale	ResNet34	VGG16	VGG19
GradCAM				
Guided-GradCAM				
Rétro propagation guidée				
Gradients				

TABLE 3.5 – Cartes d'attention artificielle post-hoc obtenues avec 4 méthodes présentées dans la partie 2.7.2 pour 3 réseaux de neurones profonds de l'état de l'art. L'exemple choisi est un **vrai positif** avec la majorité des classifieurs et la majorité des experts votant ayant considéré l'image comme pathologique.

revanche la dernière méthode, celle des gradients, semble rendre compte d'une activité attentionnelle artificielle sur la seule partie de l'image où les muqueuses sont visibles.

Pour ce deuxième exemple, nous avons une image plus ambiguë. En effet, 8 experts médicaux ont voté pour la classe pathologique et 7 pour la classe non pathologique. Sur la même image 14 des 15 classifieurs ont voté pour la classe non pathologique. Les différents résultats ont été consignés dans le tableau 3.6.

On peut observer pour quelques méthodes, que l'attention est dirigée vers les bordures de l'image. Pour quelques autres, par exemple celle de rétropropagation guidée sur le réseau VGG16 ou encore pour la méthode des gradients sur les réseaux VGG16 et VGG19 l'attention est localisée sur un point précis de l'image. Ceci est compréhensible, l'élément blanchâtre mis en évidence par ce procédé post-hoc est sans doute celui créant débat au sein des humains, la distinction entre un résidu de digestion et une ulcération aphtoïde

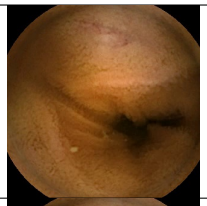
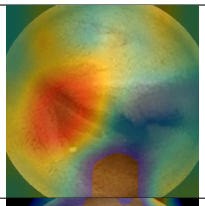
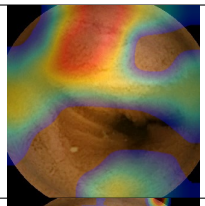
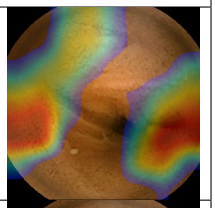
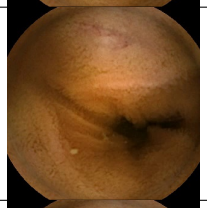
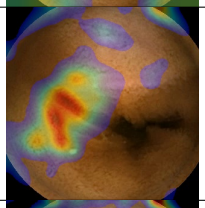
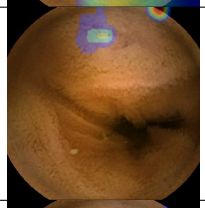
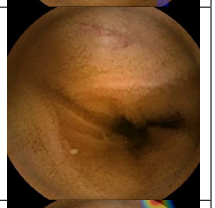
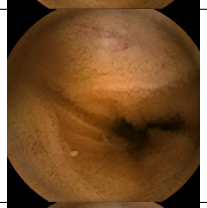
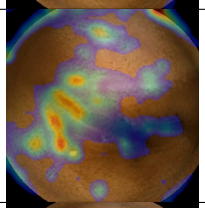
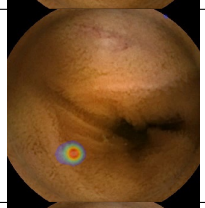
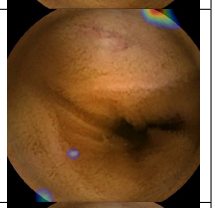
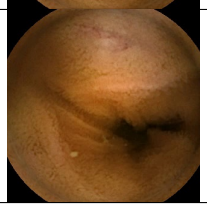
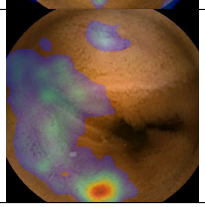
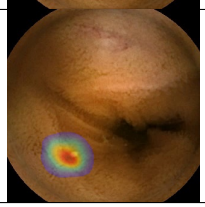
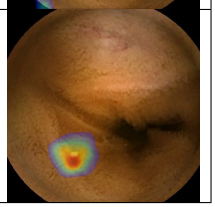
Méthode post-hoc	Originale	ResNet34	VGG16	VGG19
GradCAM				
Guided-GradCAM				
Rétro propagation guidée				
Gradients				

TABLE 3.6 – Cartes d'attention artificielle post-hoc obtenues avec 4 méthodes présentées dans la partie 2.7.2 pour 3 réseaux de neurones profonds de l'état de l'art. L'exemple choisi est un **vrai négatif** avec la majorité des classifieurs et la majorité des experts votant ayant considéré l'image comme non pathologique.

étant compliquée. On remarque également que même pour des réseaux identiques, d'une méthode à l'autre, ce ne sont pas les mêmes zones qui sont mises en valeur. Les cartes obtenues entre la méthode des gradients et GradCAM semblent même opposées.

Non concluante

Dans le tableau 3.7, nous allons nous intéresser aux résultats sur une image non concluante : sur cette image 7 humains ont voté pour pathologique, 7 pour non pathologique et une abstention. Les 15 classifieurs ont eux répondu non pathologique.

On peut observer au centre de l'image une zone portant l'ambiguïté de la décision. En effet, la zone mise en lumière par la méthode des gradients sur les réseaux VGG16 et ResNet34 est complexe à analyser. Il est difficile de dire si elle correspond à des tissus sains ou à une ulcération aphtoïde.

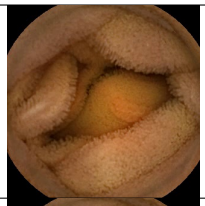
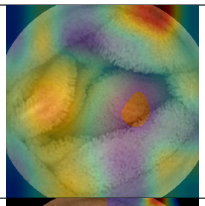
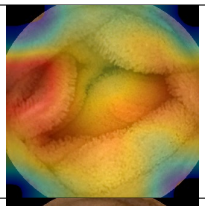
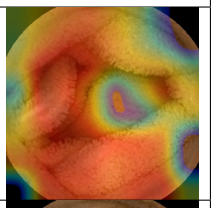
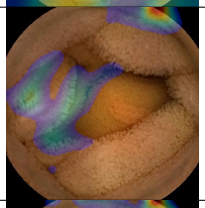
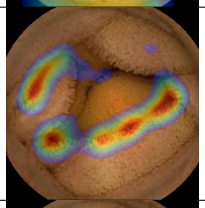
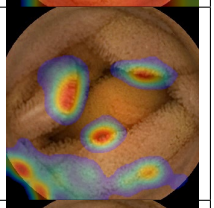
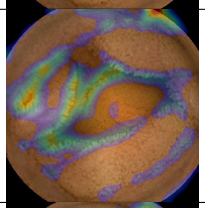
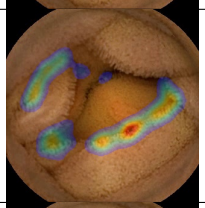
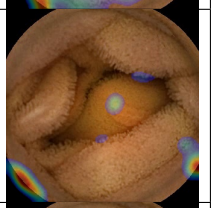
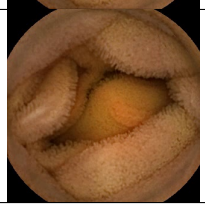
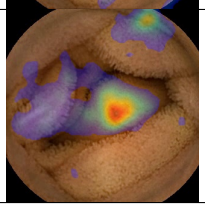
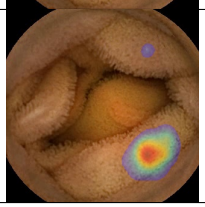
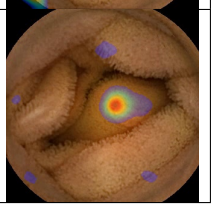
Méthode post-hoc	Originale	ResNet34	VGG16	VGG19
GradCAM				
Guided-GradCAM				
Rétro propagation guidée				
Gradients				

TABLE 3.7 – Cartes d'attention artificielle post-hoc obtenues avec 4 méthodes présentées dans la partie 2.7.2 pour 3 réseaux de neurones profonds de l'état de l'art. L'exemple choisi est un **non concluant** la moitié des votants ayant considéré l'image comme pathologique et l'autre moitié comme non pathologique.

Faux positif

Les résultats présents dans le tableau 3.8, correspondent à des cas de faux positifs. Sur cette image, 10 participants ont voté pour la classe non pathologique et 5 pour la classe pathologique. Au niveau des classifieurs, 9 d'entre eux ont répondu non pathologique et les 6 autres ont répondu pathologique.

Cette image est difficile à analyser, la différence entre les dépôts liés à la digestion et les ulcérations aphtoïdes pouvant être difficile à analyser sur une seule image. Dans le cadre du diagnostic clinique, cette image aurait été annotée à l'aide des images suivantes et précédentes, et cela aurait permis de déterminer si l'objet blanchâtre était accroché à la paroi (auquel cas l'image aurait été classifiée comme pathologique) ou si cet objet flottait dans le liquide intestinal (auquel cas elle aurait été classifiée en image non pathologique).

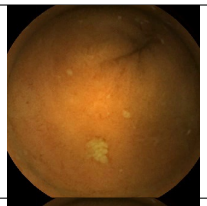
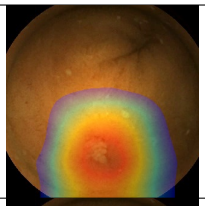
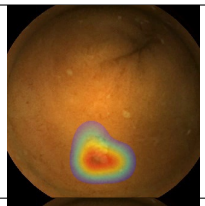
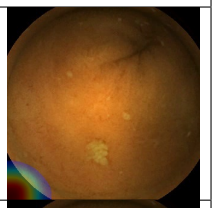
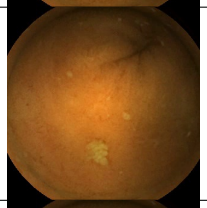
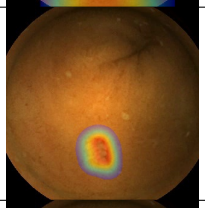
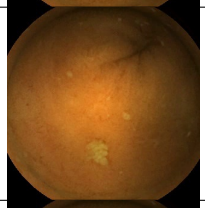
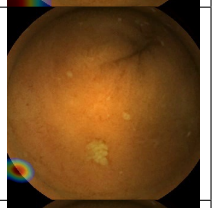
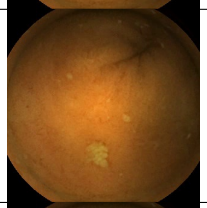
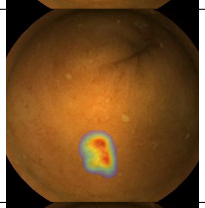
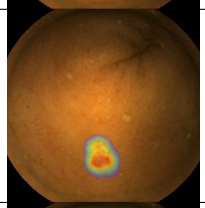
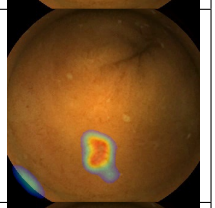
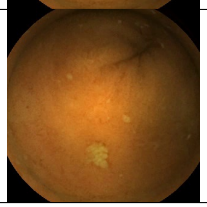
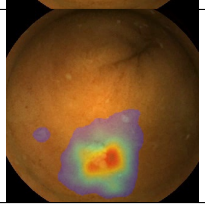
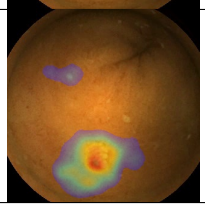
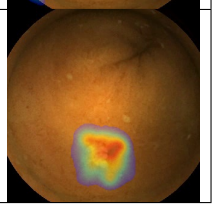
Méthode post-hoc	Originale	ResNet34	VGG16	VGG19
GradCAM				
Guided-GradCAM				
Rétro propagation guidée				
Gradients				

TABLE 3.8 – Cartes d'attention artificielle post-hoc obtenues avec 4 méthodes présentées dans la partie 2.7.2 pour 3 réseaux de neurones profonds de l'état de l'art. L'exemple choisi est un **faux positif** avec la majorité des experts votants ayant considéré l'image comme non pathologique et les exemples présentés sont issus de classifieurs ayant classé l'image comme pathologique.

Faux négatif

Le premier exemple est une image classée comme pathologique par 11 des 15 participants, deux se sont abstenus et les deux autres ont voté pour la classe non pathologiques. Sur cette image, 6 des 15 classifieurs ont prédit correctement la classe pathologique et les 9 autres ont voté pour la classe non pathologique. Dans le tableau 3.9, on peut trouver des exemples d'attention appliqués à 3 réseaux ayant voté pour la classe non pathologique.

On peut observer que pour les trois réseaux avec la méthode GradCAM, les éléments suspects situés en bas à gauche et en bas au milieu de l'image sont ignorés par les réseaux. La méthode des gradients rend bien compte d'une attention particulière pour ces éléments.

Sur l'image 3.10, 14 participants ont voté pour la classe pathologique et le dernier pour la classe non pathologique. Les classifieurs ont majoritairement voté pour la classe

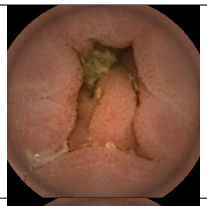
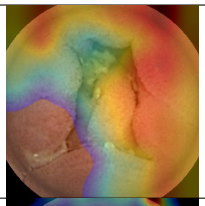
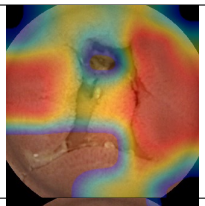
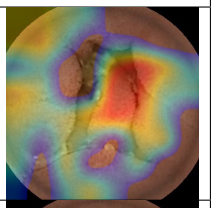
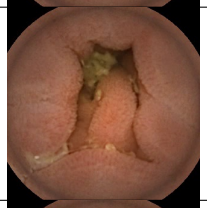
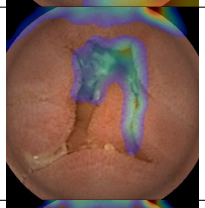
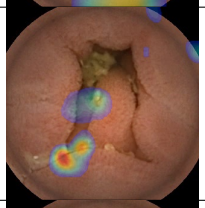
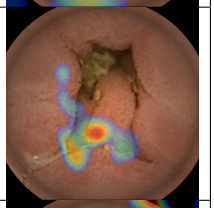
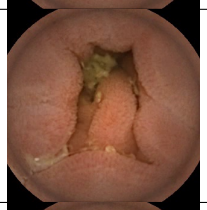
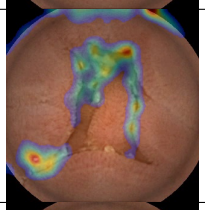
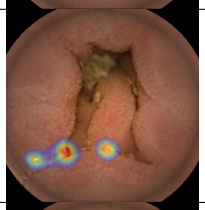
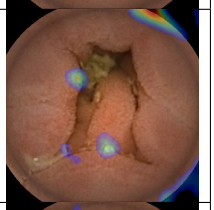
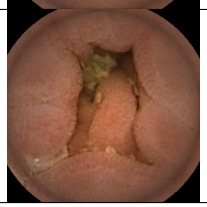
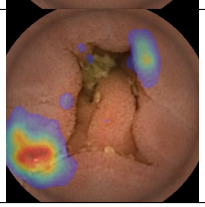
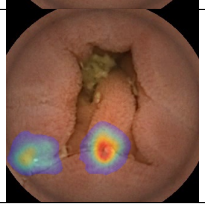
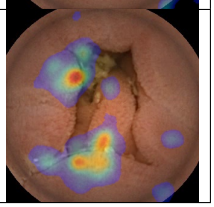
Méthode post-hoc	Originale	ResNet34	VGG16	VGG19
GradCAM				
Guided-GradCAM				
Rétro propagation guidée				
Gradients				

TABLE 3.9 – Cartes d'attention artificielle post-hoc obtenues avec 4 méthodes présentées dans la partie 2.7.2 pour 3 réseaux de neurones profonds de l'état de l'art. L'exemple choisi est un **faux négatif** avec la majorité des experts votants ayant considéré l'image comme pathologique et les exemples présentés sont issus de classifieurs ayant classé l'image comme non pathologique.

non pathologique (11 contre 4 pour la classe non pathologique).

Bien que les réseaux aient commis une erreur de classification, on peut observer que la méthode des gradients rend compte d'une activité attentionnelle sur l'élément situé en bas à gauche de l'image. Cet élément est en effet représentatif d'une lésion issue de la maladie de Crohn. On remarque que les autres méthodes ne font pas ressortir cette zone comme une zone d'intérêt (à l'exception de la méthode de rétro propagation guidée avec le réseau VGG16).

3.2.2 Discussion

Les différentes méthodes et leur application nous permettent de mieux comprendre les solutions d'explication visuelles des réseaux de neurones profonds. On observe une

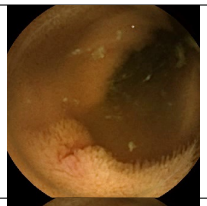
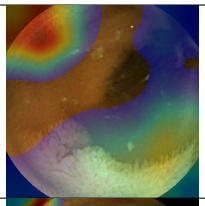
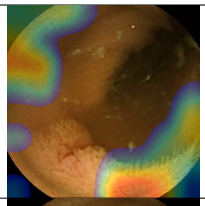
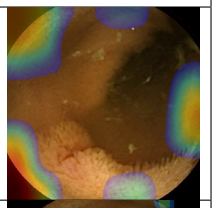
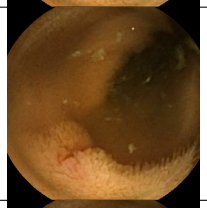
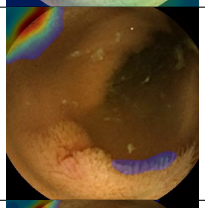
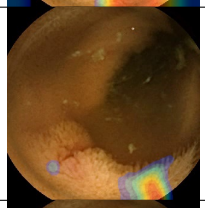
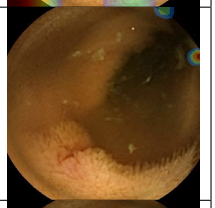
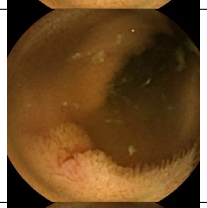
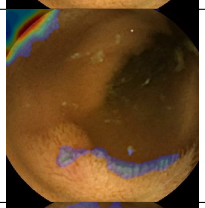
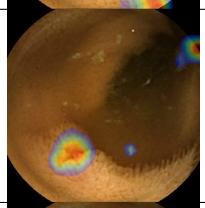
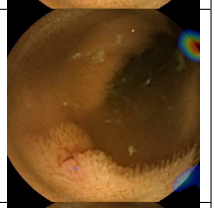
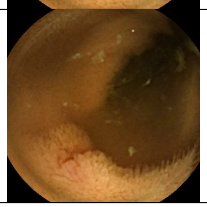
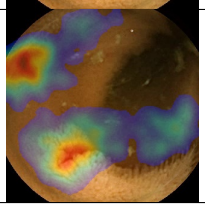
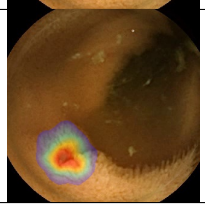
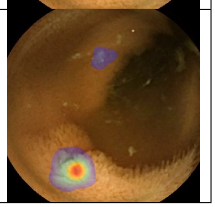
Méthode post-hoc	Originale	ResNet34	VGG16	VGG19
GradCAM				
Guided-GradCAM				
Rétro propagation guidée				
Gradients				

TABLE 3.10 – Cartes d'attention artificielle post-hoc obtenues avec 4 méthodes présentées dans la partie 2.7.2 pour 3 réseaux de neurones profonds de l'état de l'art. L'exemple choisi est un **faux négatif** avec la majorité des experts votants ayant considéré l'image comme pathologique et les exemples présentés sont issus de classifieurs ayant classé l'image comme non pathologique.

différence entre le fonctionnement de la méthode des gradients et les autres méthodes (GradCAM, guided-GradAM et rétropropagation guidée). Cette différence semble logique quand on considère ce que les cartes générées par ces méthodes représentent. La méthode des gradients, à la différence des trois autres méthodes, représente à la fois les parties de l'image ayant contribué de façon positive et négative à la décision pour telle ou telle classe, quand les autres méthodes se limite à la classe de la décision. Cela implique que lorsqu'une image est non pathologique, mais qu'un élément, comme un dépôt, peut induire une ambiguïté, cet élément sera mis en avant par la méthode des gradients et sera occulté par les autres méthodes.

Il semblerait donc que la méthode des gradients est plus représentative des différents éléments induisant la décision finale. Cependant, les méthodes utilisant les cartes d'ac-

tivations comme gradCAM et guided-GradCAM prennent en compte des informations utilisées lors de la propagation de l'image à travers le réseau.

L'étude au cas par cas des différentes méthodes de visualisation nous a permis de nous rendre compte que pour des réseaux similaires, possédant exactement les mêmes poids, les résultats des explications pouvaient être parfaitement opposés.

3.3 Comparaison de la stabilité de l'attention post-hoc

Dans la partie précédente, nous avons présenté des résultats visuels obtenus pour différentes méthodes d'extraction de l'attention post-hoc, appliquées à différents réseaux de l'état de l'art, ResNet34, VGG16 et VGG19. Dans cette partie, les points communs et les différences entre ces méthodes seront quantifiés, à l'aide d'une métrique de saillance. Nous chercherons à montrer s'il y a, oui ou non, une influence du label sur la façon d'observer les images par les réseaux.

Nous verrons ici si la répartition des images entre le jeu d'entraînement et le jeu de validation a une influence sur les zones d'attention du réseau. En d'autres termes, nous verrons si les résultats sont stables d'un apprentissage à l'autre lors de la validation croisée. Nous précisons ici que chacun des réseaux a initialement été entraîné sur ImageNet, et possède ainsi le même état de départ.

3.3.1 Méthode

Comme expliqué précédemment, les réseaux ont été entraînés 5 fois, de façon à obtenir des résultats statistiquement valides. Dans cette comparaison, nous souhaitons observer si la variabilité de l'entraînement induit une variabilité des zones d'attention. Pour cela, les cartes d'attention des réseaux ont été calculées pour chacune des 4 méthodes, pour chacun des 3 réseaux et pour chaque répartition des données d'entraînement. Une fois ces cartes calculées, nous avons utilisé le coefficient de corrélation de Pearson (CC voir partie 2.3.3) afin d'évaluer pour un réseau donné, pour une méthode donnée, et pour chaque image si les réseaux accordent leur attention aux mêmes zones en fonction de l'initialisation du réseau. On rappelle que les valeurs de cette métrique sont comprises entre [-1,1], -1 correspondant à une complète anti corrélation des cartes de saillance et 1 à leur parfaite corrélation. Ainsi, pour chacune des cartes obtenues, pour chacune des 5 répartitions, le

CC a été calculé avec les 4 autres cartes obtenues avec les autres répartitions. Plus les cartes d'attention seront stables d'une initialisation à l'autre, plus le score de corrélation (CC) sera proche de 1.

3.3.2 Résultats

L'ensemble des résultats pour les différentes méthodes d'explication visuelle pour chacun des trois réseaux de l'état de l'art sont synthétisés dans la section suivante.

GradCAM

Premièrement nous allons nous intéresser aux résultats obtenus par la méthode GradCAM. On peut observer sur la figure 3.4 que les résultats et les rapports entre la stabilité sur les images pathologiques et non pathologiques sont très dépendants du modèle. Pour le modèle ResNet34, les résultats obtenus sont nettement plus stables avec en moyenne un CC de 0.418 (std : 0.538) sur les images pathologiques contre un CC de 0.203 (std : 0.362) sur les images non pathologiques. Pour VGG19, à l'inverse le comportement est plus stable sur les images pathologiques avec un CC de 0.353 (std : 0.389) et de 0.438 (std : 0.286) sur les non pathologiques. Et finalement, avec les résultats pour VGG16 on trouve un comportement inverse proche de celui obtenu pour VGG19 avec une plus grande stabilité des zones d'attention sur les images non pathologiques (moyenne 0.323, std : 0.302) que sur les images pathologiques (moyenne 0.246, std : 0.335).

Guided GradCAM

Pour la méthode Guided GradCAM, on peut observer que pour les différents réseaux, la différence de stabilité entre les cartes d'attention sur les images pathologiques et non pathologiques est présente. Pour VGG16, on obtient une valeur de CC de 0.177 (std : 0.262) sur les images non pathologiques contre un CC de 0.370 (std : 0.390) pour les images pathologiques. Avec le réseau VGG19, les résultats sont légèrement plus faibles mais la tendance reste la même avec une moyenne de 0.262 (std : 0.256) sur les images non pathologiques contre 0.435 (std : 0.370) sur les images pathologiques. C'est avec ResNet34 que cette tendance est la plus visible. On observe un CC, de par la comparaison entre les différentes répartitions de la base d'entraînement, de 0.252 (std : 0.279) sur les images non pathologiques et de 0.455 (std : 0.421) sur les images pathologiques.

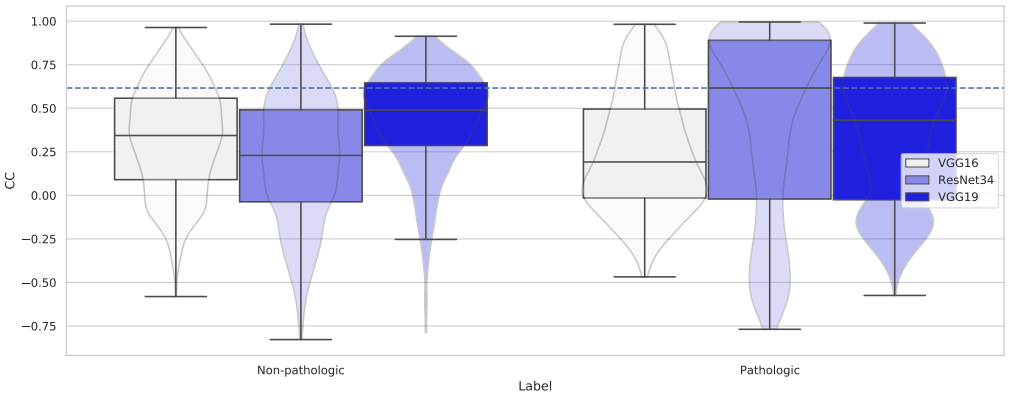


FIGURE 3.4 – Résultats de stabilité pour les réseaux de neurones profonds ResNet34, VGG19, VGG16 pour la méthode **GradCAM**. L'évaluation de la stabilité est réalisée à l'aide de la métrique CC, la valeur 1 correspondant à une parfaite similitude entre deux cartes d'attention, la valeur 0 à une non-corrélation et la valeur -1 à anti-corrélation. La médiane maximum pour l'ensemble des réseaux et pour les groupes pathologique et non pathologique est représentée par une ligne bleu en pontillés.

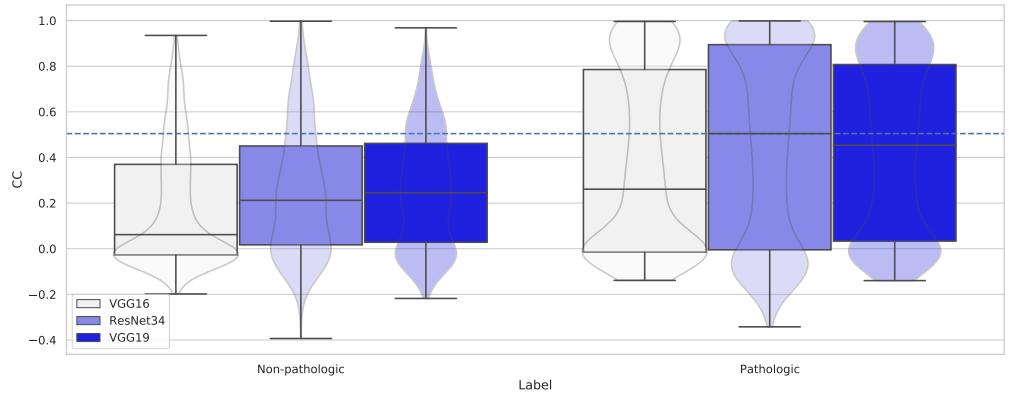


FIGURE 3.5 – Résultats de stabilité pour les réseaux de neurones profonds ResNet34, VGG19, VGG16 pour la méthode **Guided GradCAM**. L'évaluation de la stabilité est réalisée à l'aide de la métrique CC, la valeur 1 correspondant à une parfaite similitude entre deux cartes d'attention, la valeur 0 à une non-corrélation et la valeur -1 à anti-corrélation. La médiane maximum pour l'ensemble des réseaux et pour les groupes pathologique et non pathologique est représentée par une ligne bleu en pontillés.

Rétropropagation guidée

Avec la méthode de rétro-propagation guidée, la même tendance générale est encore observable. La valeur maximale de stabilité est toujours obtenue avec ResNet34, obtenant une valeur de CC de 0.564 (std : 0.276) sur les images pathologiques et une valeur de 0.442 (std : 0.234) sur les images non pathologiques.

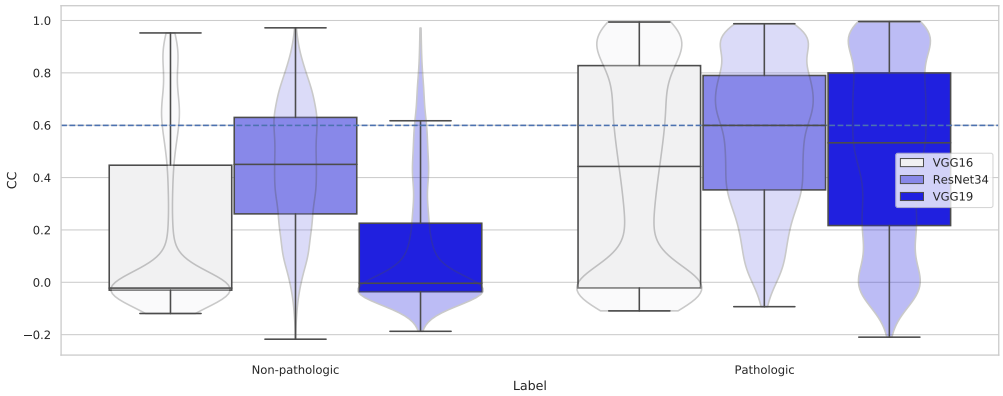


FIGURE 3.6 – Résultats de stabilité pour les réseaux de neurones profonds ResNet34, VGG19, VGG16 pour la méthode de **Rétropropagation guidée**. L'évaluation de la stabilité est réalisée à l'aide de la métrique CC, la valeur 1 correspondant à une parfaite similitude entre deux cartes d'attention, la valeur 0 à une non-corrélation et la valeur -1 à anti-corrélation. La médiane maximum pour l'ensemble des réseaux et pour les groupes pathologique et non pathologique est représentée par une ligne bleu en pontillés.

Gradients

Finalement c'est avec la méthode des gradients qu'on obtient les meilleurs résultats de stabilité sur les réseaux. Avec ResNet34, la valeur moyenne obtenue sur l'ensemble des images non pathologiques est de 0.365 (std : 0.293) contre 0.685 (std : 0.273) pour les images pathologiques. Avec VGG16 on obtient une valeur moyenne de 0.163 (std : 0.263) sur les images non pathologiques et 0.602 (std : 0.346) sur les pathologique. Finalement avec VGG19 on obtient un CC moyen de 0.150 (std : 0.319) sur les images non pathologiques et de 0.593 (std : 0.349) sur les images pathologiques.

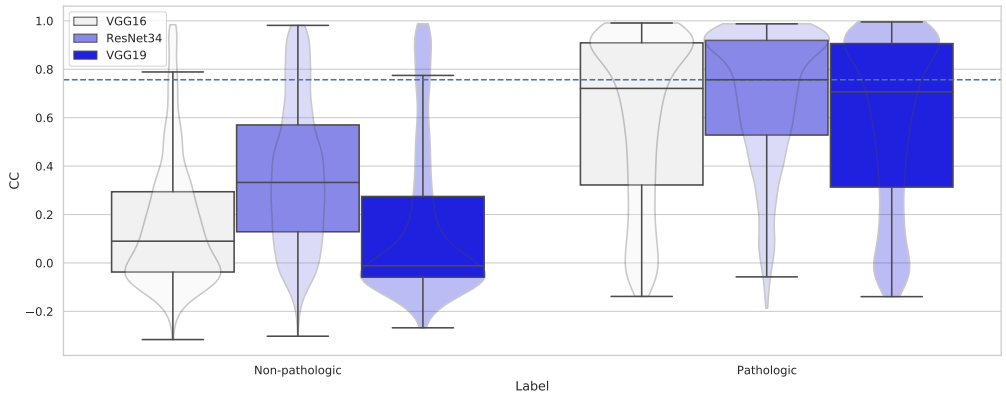


FIGURE 3.7 – Résultats de stabilité pour les réseaux de neurones profonds ResNet34, VGG19, VGG16 pour la méthode des **Gradients**. L'évaluation de la stabilité est réalisée à l'aide de la métrique CC, la valeur 1 correspondant à une parfaite similitude entre deux cartes d'attention, la valeur 0 à une non-corrélation et la valeur -1 à anti-corrélation. La médiane maximum pour l'ensemble des réseaux et pour les groupes pathologiques et non pathologique est représentée par une ligne bleu en pontillés.

3.3.3 Discussion

Dans les tableaux 3.11 et 3.12 sont résumés les résultats de stabilité pour la métrique CC pour les images non pathologiques et les images pathologiques respectivement. On y retrouve les résultats présentés précédemment. Bien que pour des réseaux équivalents, les résultats de stabilité pour les différentes méthodes soient fluctuants, pour les images pathologiques, on observe bien que la méthode des gradients est plus stable que les autres méthodes sur les images pathologiques. Une différence notable est également observable entre la stabilité sur les images pathologiques et la stabilité sur les images non pathologiques. En moyenne, les réseaux bien qu'entraînés différemment ont tendance à porter leur attention sur des zones plus proches quand l'image est pathologique. Ce résultat semble logique du fait que les images pathologiques contiennent des éléments localisés symptomatiques de la maladie de Crohn.

Les résultats présentés ici semblent nous indiquer un comportement attentionnel différent des réseaux de neurones profonds sur les images pathologiques et sur les images non pathologiques. Les lésions de la maladie de Crohn, dont l'identification est nécessaire pour la classification de l'image, homogénéisent les comportements attentionnels des différents réseaux.

Méthode	VGG16	VGG19	ResNet34
GradCAM	**0.323 (0.302)**	**0.438 (0.286)**	0.203 (0.362)
Guided GradCAM	0.177 (0.262)	0.262 (0.256)	0.252 (0.279)
Rétro propagation guidée	0.176 (0.330)	0.121 (0.238)	**0.442 (0.234)**
Gradients	0.163 (0.263)	0.150 (0.319)	0.365 (0.293)

TABLE 3.11 – Tableau récapitulatif des scores obtenus avec le coefficient de corrélation de Pearson (CC) entre les cartes d'attention des images **non pathologiques** d'un même réseau entraîné pour différentes répartitions du jeu d'entraînement et de validation.

Méthode	VGG16	VGG19	ResNet34
GradCAM	0.246 (0.335)	0.353 (0.389)	0.418 (0.538)
Guided GradCAM	0.370 (0.390)	0.435 (0.370)	0.455 (0.421)
Rétro propagation guidée	0.424 (0.393)	0.503 (0.336)	0.564 (0.276)
Gradients	**0.602 (0.346)**	**0.593 (0.349)**	**0.685 (0.273)**

TABLE 3.12 – Tableau récapitulatif des scores obtenus avec le coefficient de corrélation de Pearson (CC) entre les cartes d'attention des images **pathologiques** d'un même réseau entraîné pour différentes répartitions du jeu d'entraînement et de validation.

Bien que l'influence du label soit clairement visible sur la stabilité des comportements attentionnels des réseaux de neurones profonds, on observe que pour un même réseau, les comportements attentionnels rendus par les différentes méthodes sont différents. Bien que les poids du réseau soient parfaitement semblables, les différentes méthodes ne nous donnent pas les mêmes zones d'attention pour une même décision.

3.4 Conclusion

Dans cette partie nous avons vu plusieurs méthodes d'attention artificielle appliquées à la détection des lésions de la maladie de Crohn dans des images issues de vidéos capsules endoscopiques. Premièrement avec de l'attention apprise et un réseau récurrent inspiré de celui de Mnih et al. 2014 [170] . Ce réseau bien que ne dépassant pas les performances de l'ensemble des réseaux de neurones à convolution de l'état de l'art permet d'obtenir des performances proches de celles de ces derniers. Avec son comportement attentionnel proche de celui des humains, alternant saccades et fixations, le réseau comme les humains

apprend à positionner ses zones d'attention de façon à optimiser sa décision. Cependant, le réseau récurrent, à la différence des humains ne possède pas de vision périphérique lui permettant de contextualiser l'information extraite pendant la fixation et l'aidant à mieux choisir ou positionner son regard. On pourrait envisager, de manière à améliorer les performances de ce réseau, de créer un sous-réseau supplémentaire permettant de reproduire le comportement de cette dernière. L'avantage principal du réseau mis en place est qu'il ne dépend pas de la taille de l'image d'entrée. Ce genre d'algorithme décisionnel pourrait être utile pour traiter de larges images comme les radiographies.

En ce qui concerne les méthodes post-hoc, les différentes méthodes nous permettent de comprendre les éléments localisés conduisant à la décision. Par exemple ont peut interpréter que les erreurs commises par les algorithmes sont plus de l'ordre des erreurs de décision et des erreurs de reconnaissance, du fait que la méthode des gradients fait ressortir des zones suspectes même quand la décision finale n'est pas en accord avec la vérité terrain. Nous pouvons également observer que les méthodes ne produisent pas les mêmes explications visuelles, bien que les poids du réseau dont elles proviennent soient les mêmes. Un comportement général y est cependant observable, les pathologies présentes dans l'image semble homogénéiser les différents comportements attentionnels des réseaux quelle que soit la méthode utilisée pour extraire les explications visuelles.

Une comparaison de la stabilité avec les méthodes d'attention apprise pourrait être réalisée. Nous pourrions vérifier grâce à elle que le comportement des méthodes d'attention artificielle est également plus stable sur les images pathologiques que sur les images non pathologiques. De manière générale, dans ce manuscrit, nous nous sommes concentrés principalement sur l'étude des méthodes de l'extraction de l'attention visuelle post-hoc, une étude plus approfondie des méthodes intégrant directement de l'attention serait intéressante.

Maintenant que nous avons vu que les différentes méthodes d'explicabilité produisent des résultats différents et parfois opposés, nous nous posons la questions de savoir quel est la meilleure méthode pour rendre compte de l'attention d'un réseau de neurones profonds. Ces zones d'attention artificielle sont-elles proches de celles des humains ? Bien que le fonctionnement de l'attention humaine et artificielle soit profondément différents, les éléments localisés permettant de définir l'image comme pathologique restent les mêmes.

Chapitre 4

Comparaison entre l'attention artificielle et humaine

Sommaire

La comparaison entre l'attention humaine et l'attention machine n'a pour l'instant été que très peu explorée. On peut trouver une étude sur la comparaison entre des cartes d'attention humaine et artificielle dans Das et al. [189]. Les auteurs concluent que les modèles d'attention et les humains n'observent pas les mêmes zones lors d'une tâche de

réponse visuelle aux questions (*Visual Question Answering* en anglais). Cette étude est critiquable, les données d'attention humaine ayant été recueillies à l'aide d'une méthode de clic avec la souris. Il a été montré que cette méthode d'annotation produit des données bruitées ne reflétant pas la réalité des fixations humaines [190]. Lai et al. 2019 [191], eux, comparent les résultats obtenus avec des architectures à attention douce et des données oculométriques réelles. Ils montrent que lorsque l'attention visuelle humaine et l'attention artificielle sont toutes deux guidées par la tâche, plus les performances des réseaux sont élevées, plus l'attention artificielle est proche de l'attention humaine. Cependant, ces travaux sont limités par une forte hétérogénéité entre les conditions expérimentales de l'homme et de la machine et par le nombre limité de participants (cinq sujets).

Bien que certaines similarités entre l'attention humaine et l'attention artificielle puissent être vérifiées expérimentalement, les algorithmes d'apprentissage profond, à la différence des humains, sont sensibles à de faibles perturbations de bruit [192]. Récemment des travaux ont été réalisés pour comprendre comment fonctionnait la représentation interne des objets lors d'une tâche de classification chez des algorithmes d'apprentissage profond au travers de la comparaison avec la représentation humaine [193]. Réussir à comprendre le fonctionnement de la perception des algorithmes d'apprentissage profond est un enjeu majeur pour permettre le développement d'architectures plus résilientes.

Comparer le fonctionnement de l'attention humaine et artificielle pourrait donc être un moyen de mieux comprendre le fonctionnement des réseaux de neurones profonds. On peut trouver dans la littérature d'autres tentatives de comparaison entre attention humaine et artificielle dans différentes tâches, comme la préparation de repas [194], les jeux vidéos [195] et la conduite autonome [196].

Dans cette partie, nous présentons une comparaison entre l'attention humaine et l'attention artificielle dans une tâche de diagnostic médical. Toujours dans le contexte de l'aide au diagnostic de la maladie de Crohn, nous avons travaillé en partenariat avec des gastro-entérologues ainsi qu'avec des internes en médecine de la région Pays de la Loire. Les objectifs de cette étude sont multiples. Dans un premier temps, nous souhaitons permettre l'évaluation des différentes méthodes d'attention artificielle présentées dans le chapitre 3. Comme nous l'avons montré, ces dernières montrent des zones d'attentions différentes pour des réseaux et des stimuli similaires. Ainsi, au travers de la comparaison humain et machine on souhaite observer dans un premier temps si le comportement attentionnel humain est également influencé par le label de l'image comme il l'a été montré pour l'attention artificielle dans la partie 3.3. Dans un deuxième temps, nous souhaitons

mesurer à quel point les comportements attentionnels humain et artificiel sont semblables et influencés par leur niveau d'expertise. Et finalement nous discuterons de l'évaluation des méthodes d'explication visuelle par la comparaison avec l'attention humaine.

Ce chapitre s'organisera de la façon suivante. Dans une première section, nous présenterons le travail d'acquisition réalisé au cours des trois dernières années afin de pouvoir dresser un parallèle cohérent entre les comportements attentionnels humains et artificiels. Dans une seconde section, nous évaluerons les différentes méthodes d'attention artificielle en cherchant lesquelles présentent le plus de similarité avec l'attention humaine. Et dans une dernière section nous évaluerons l'effet de l'expertise sur la comparaison entre attention humaine et artificielle.

4.1 Étapes de création de la base de données et travail d'acquisition

4.1.1 Contexte

L'idée originale de créer une base de données oculométriques permettant la comparaison entre les comportements attentionnels humain et artificiel est apparue au milieu de ma première année de thèse. Le premier problème auquel nous avons directement été confronté était de trouver des experts médicaux volontaires pour participer à une telle expérience. En effet, les experts médicaux étant rares et leur temps précieux, il est difficile de mettre en place une expérience permettant de récupérer suffisamment de données oculométriques afin d'obtenir des analyses comportementales qui seront statistiquement valides.

Grâce au partenariat avec le service de gastro-entérologie du CHU de Nantes et plus particulièrement grâce à la confiance en notre travail du Dr. Boureille (chef du service de gastro-entérologie du CHU de Nantes) et à l'aide d'Astrid de Maissin (interne en médecine du CHU de Nantes), nous avons pu obtenir la participation de médecins pour une première expérience d'eye-tracking en juin 2019.

Afin de répondre aux contraintes de disponibilité des gastro-entérologues, nous avons choisi de travailler avec un eye-tracker portable : l'eyeTribe (voir section 4.1.4). J'ai ainsi développé une application Python permettant de réaliser une expérience oculométrique. L'eye-tracker utilisé ne possédait pas d'API Python fonctionnelle permettant une utilisation facile avec la librairie *PsychoPI*. J'ai donc dans un premier temps codé les éléments

essentiels permettant de réaliser l'expérience : la phase de calibration permettant d'associer la position des yeux du sujet aux pixels fixés de l'écran ; la phase de fixation de la croix permettant de valider qu'il n'y a pas eu de dérive de l'appareil. Dans cette phase, on demande au sujet de fixer une croix au centre de l'image avec une marge d'erreur de 5%. Ainsi, on s'assure que les stimuli présents sur l'écran et le comportement des sujets sont toujours bien en accord. Cette phase précédant la phase de visualisation de l'image assure que chacun des participants ait une première fixation au centre de l'image. La phase de visualisation, comme son nom l'indique, permet de visualiser l'image. Pendant cette phase, les positions du regard humain sont enregistrées pour être ensuite analysées. Finalement une phase de vote permet au participant de réaliser son diagnostic de l'image, en votant pour une ou plusieurs des 6 pathologies possibles, de s'abstenir de répondre, ou de considérer l'image comme saine (non pathologique). Les détails de la mise en place de l'expérience seront détaillés dans la section 4.1.4.

4.1.2 Alignement des conditions expérimentales humaines et artificielles

Notre objectif initial était de comparer le comportement attentionnel humain et artificiel. Pour limiter les biais lors de la comparaison entre l'attention humaine et celle de la machine, nous devons utiliser la même tâche et les mêmes conditions expérimentales pour les deux [197]. Tout d'abord, le nombre de calculs effectués par le réseau est fixe, puisque le nombre de paramètres est identique pour chaque image. Ainsi, en supposant que le nombre de calculs effectués quasi-simultanément pour la machine correspond au temps de visualisation des images pour les humains, nous avons également fixé le même temps de visualisation (2 secondes) pour chaque image.

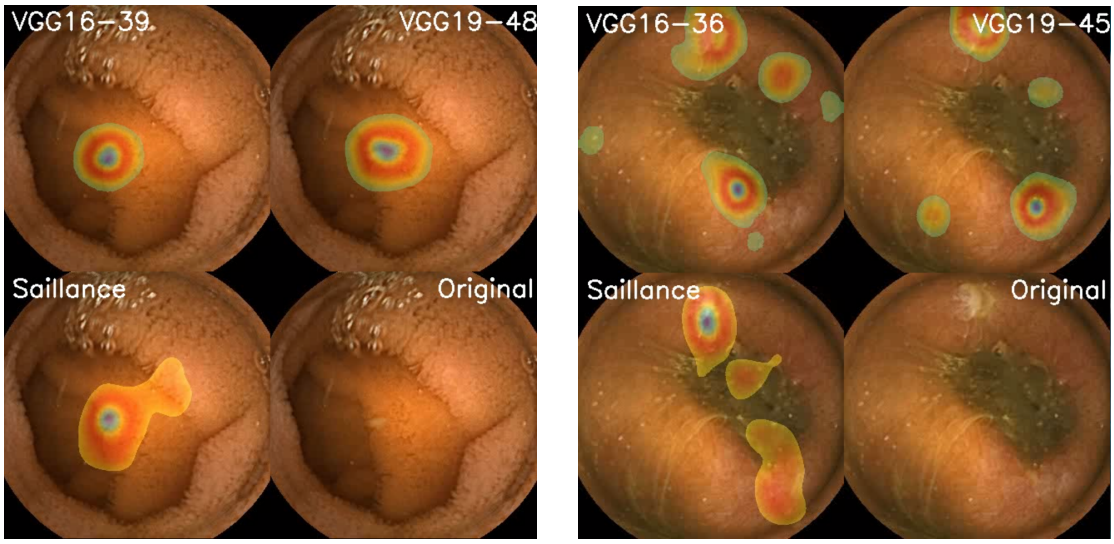
Deuxièmement, chaque image est traitée par le réseau indépendamment des images précédentes et suivantes, et sans aucune information sur le patient. Nous avons donc également présenté les images aux humains indépendamment de leur contexte clinique, et dans un ordre aléatoire différent pour chaque participant.

Troisièmement, les valeurs des paramètres ont été fixées, calculées à partir d'un entraînement précédent sur notre ensemble de données. De même, aucun feedback n'a été donné aux participants, les empêchant d'apprendre pendant l'expérience.

4.1.3 Pré-tests

Les pré-tests réalisés au CHU de Nantes sur 8 participants (4 internes en médecine et 4 gastro-entérologues) nous ont permis de valider le bien fondé d'une telle expérience et la possibilité d'une comparaison entre attention artificielle et attention humaine. Ces tests ont été réalisés sur 212 images. Le protocole expérimental décrit plus en détail dans la partie 4.1.4 a été appliqué. Pendant une journée entière, j'ai réalisé l'acquisition des données au CHU de Nantes grâce à l'aide d'A. De Maissin, qui coordonnait les différents participants.

Sur les images non pathologiques, il n'était pas possible d'observer une réelle corrélation entre les comportements attentionnels humains et artificiels. Cependant, sur les images pathologiques, comme montré sur la figure 4.1, des zones d'intérêt semblaient être partagées entre humains et machine. L'expérience comptait un total de 212 images. Malheureusement, moins d'un cinquième d'entre elles avait été annotées comme pathologiques. Le déséquilibre entre le nombre d'images pathologiques et le nombre d'images non pathologiques, ne nous a donc pas permis de réaliser une étude quantitative de qualité, le risque d'induire un biais du fait de ce déséquilibre étant trop important.



(a) (b)

FIGURE 4.1 – Exemples de résultats des prétests. Sur les deux premières lignes on peut observer les résultats obtenus par les réseau VGG16 et VGG19 avec la méthode de rétropropagation guidée. Sur la deuxième ligne sont présents les résultats de saillance humaine ainsi que l'image originale.

Cette pré-expérience nous permet de formuler l'hypothèse selon laquelle une corrélation existerait entre les zones d'attention humaine et artificielle et principalement sur les images pathologiques. Le fait que cette corrélation soit plus forte sur les images pathologiques que sur les images non pathologiques pourrait s'expliquer par le fait qu'elles ne contiennent pas d'élément discriminant faisant converger à la fois les comportements attentionnels humain et artificiel.

Afin de vérifier cette hypothèse, nous avons réalisé une nouvelle expérience oculométrique. Celle-ci est présentée dans la section 4.1. Elle compte 250 images, hors de la base de données CrohnIPI, sélectionnées au préalable par un expert en gastro-entérologie ne participant pas à l'expérience, de sorte à assurer un meilleur équilibre entres les classes. En réalisant l'expérience sur un jeu de données mieux équilibré et sur un plus grand nombre de participants, nous avons souhaité permettre la vérification statistique de notre hypothèse avec des modèles linéaires mixtes. De plus, lors de la pré-expérience, les niveaux d'expertise des différents sujets étaient hétérogènes. La nouvelle expérience sera ainsi également un moyen d'observer dans quelle mesure cette hétérogénéité influe sur la comparaison entre comportement attentionnel humain et machine.

4.1.4 Crohn Eye-Pi une base de données oculométriques multi-experts

Participants

Nous avons enregistré le regard de 23 participants. Onze d'entre eux sont des gastro-entérologues expérimentés ayant examiné plus de 100 capsules vidéos au cours de leur carrière, cinq étaient des médecins débutants ayant examiné moins de 100 capsules vidéos et les sept derniers étaient hors du domaine médical et n'avaient jamais vu d'images d'intestin. Nous avons retiré un des gastro-entérologues seniors en raison d'échecs répétés de calibration. Les analyses ont été effectuées sur un groupe final de 22 participants (13 hommes, 9 femmes). Nous sommes conscients que 11 experts peut être considérés comme un échantillon de faible taille pour effectuer des analyses statistiques inférentielles. Nous justifions ce nombre par le défi considérable que représente la recherche d'experts médicaux capables de consacrer du temps à une expérience comportementale. Pour rassembler cet ensemble de données, nous avons dû mettre en place un système d'eye-tracking itinérant et visiter plusieurs hôpitaux universitaires en France sur une période d'un an. Nous reconnaissons la taille de l'échantillon comme une limitation de notre étude dans la

discussion.

Stimuli

Les stimuli consistaient en 250 images issues de vidéos capsules endoscopiques provenant de deux patients diagnostiqués avec la maladie de Crohn. La résolution des images était de 640×640 pixels. Pour limiter le déséquilibre entre le nombre d'images pathologiques et non pathologiques, elles ont été sélectionnées par un expert qui n'a pas participé à l'expérience. Pour étiqueter les images comme pathologiques ou non pathologiques, nous avons utilisé les votes des 15 médecins seniors et juniors sur chaque image. Lorsqu'une lésion liée à Crohn était détectée, l'image correspondante était étiquetée comme pathologique, sinon elle était étiquetée comme non pathologique. L'option "Je ne sais pas" correspondait à une abstention. 58,0% des images ont été classées comme non pathologiques, 40,8% comme pathologiques et 1,2% indéterminées (3 images). Les 3 images indéterminées correspondaient aux images pour lesquelles les nombres de votes pour les classes pathologique et non pathologique étaient égaux. Elles ont été retirées des analyses suivantes.

Experts et non experts

Afin d'attester du niveau d'expertise des différents participants, nous avons choisi de les classer selon le nombre de réponses correctes. La vérité associée à chacune des images était définie par le vote de la majorité des participants. Afin de ne pas biaiser les résultats, les parfaits novices ont été retirés du vote. Les participants ayant un score de précision supérieur à 80 % ont formé le groupe des experts (11 participants). Les 11 autres participants formaient le groupe des non experts. Logiquement, tous les novices se sont retrouvés dans le groupe des non experts, mais aussi un médecin senior, avec un score de précision de 71,7% (voir tableau 4.1).

Appareil

Les stimuli ont été affichés sur un moniteur Dell P2417H à 60 Hz (1920x1080 pixels). La taille de l'écran était de 52,7 cm par 29,6 cm. Les données d'eye-tracking ont été enregistrées à 60 Hz avec un eye-tracker 'The EyeTribe' [1]. Les participants étaient assis à 60 cm de l'écran (distance tête-écran).

1. `https://theeyetribe.com/theeyetribe.com/about/index.html`

	Participant ID	Accuracy (%)	Correct answers	Total answers
Novices	Random people 7	44.3	101	228
	Random people 4	55.1	136	247
	Random people 3	58.6	143	244
	Random people 5	61.9	153	247
	Junior doctor 4	64.9	155	239
	Senior doctor 4	71.7	165	230
	Random people 1	69.3	167	241
	Junior doctor 1	69.7	170	244
	Random people 2	69.8	171	245
	Random people 6	73.2	180	246
	Junior doctor 2	76.1	185	243
Experts	Senior doctor 2	80.8	193	239
	VGG16	**81.4**	**203.6**	**250**
	VGG19	**82.2**	**205.5**	**250**
	Senior doctor 8	93.7	209	223
	ResNet34	**84.1**	**210.3**	**250**
	Senior doctor 5	89.9	213	237
	Junior doctor 3	89.1	220	247
	Senior doctor 9	92.1	220	239
	Junior doctor 5	90.2	221	245
	Senior doctor 6	93.4	225	241
	Senior doctor 7	93	225	242
	Senior doctor 10	91.5	226	247
	Senior doctor 3	93.9	230	245
	Senior doctor 1	94.7	234	247

TABLE 4.1 – Tableau récapitulatif des participants à l'expérience d'eye-tracking. Chaque participant a été assigné au groupe expert ou novice en fonction du nombre de bonnes réponses obtenues lors de la tâche de classification des images de la capsule vidéo endoscopique. ResNet34 a obtenu 210 réponses correctes sur 250 possibles, VGG16 et VGG19 en ont obtenu 203 et 205 respectivement. Les trois réseaux seraient donc classés dans la catégorie expert.

Procédure

L'expérience d'eye-tracking a eu lieu dans différents bureaux personnels, dans différents centres hospitaliers. Chaque expérience s'est déroulée dans des pièces calmes. Le déroulement de l'expérience est décrit dans la figure 4.2. L'expérience se composait de deux phases d'environ 20 minutes, entrecoupées d'une pause obligatoire. Chaque phase consistait à examiner 125 images. Les images étaient présentées dans un ordre aléatoire

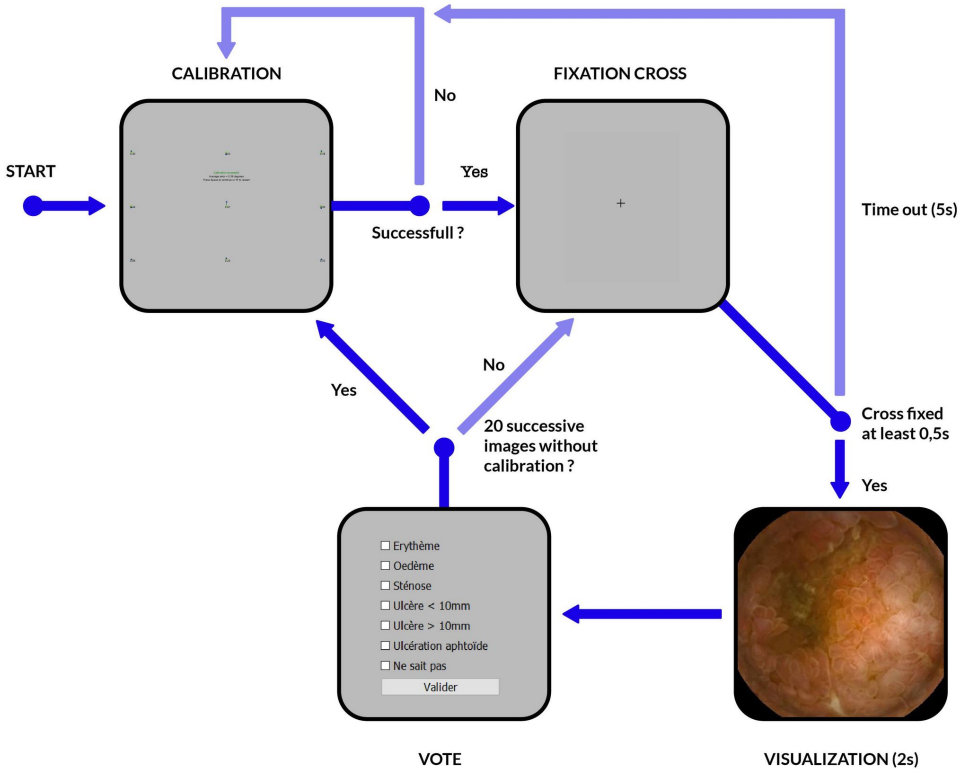


FIGURE 4.2 – Déroulement de l'expérience d'eye-tracking.

différent pour chaque participant, au centre de l'écran. Les participants pouvaient faire une pause après chaque image. Une procédure de calibration en 9 points a été réalisée au début de l'expérience, après 20 images successives, et après une pause. Entre chaque image, une correction de dérive était également effectuée, afin de prévenir une éventuelle dérive du dispositif de mesure due aux possibles changements de position du participant. La phase de correction de dérive consistait à fixer une croix de fixation centrale. Si le regard du participant se posait sur la croix pendant au moins une demi-seconde, le stimulus était affiché, sinon un nouveau calibrage était initié. Le stimulus était ensuite affiché pendant 2 secondes.

4.2 Évaluation des méthodes d'attention artificielle

On peut trouver dans la littérature de nombreux articles tentant de définir comment évaluer correctement les algorithmes d'explicabilité. Selon Doshi-Velez et al. [198], il existe trois façons différentes d'évaluer l'interprétabilité d'un modèle : l'évaluation fondée sur

l'application en utilisant des humains réels sur une tâche réelle, les mesures fondées sur l'humain en utilisant des humains réels sur des tâches de proxy simplifiées, et l'évaluation fondée sur la fonction en n'impliquant pas les humains et en utilisant simplement une certaine définition formelle de l'interprétabilité. Notre approche s'inscrit dans le cadre des évaluations fondées sur les applications, où l'on demande aux humains et aux machines d'effectuer une même tâche dans des conditions réelles afin de comparer leurs décisions et leurs performances. Ici, nous proposons d'étudier l'interprétabilité des algorithmes d'apprentissage profond dans une tâche de diagnostic médical assisté par ordinateur. La médecine est un domaine où les approches d'apprentissage automatique sont de plus en plus demandées afin d'alléger la charge de travail de médecins souvent surchargés. Cependant, les médecins ont besoin d'algorithmes qui ne sont pas seulement performants, mais qui sont également dignes de confiance, transparents, interprétables et explicables pour les experts humains [199].

Dans un premier temps, nous avons cherché à observer laquelle des différentes méthodes présentées précédemment avait un comportement attentionnel le plus proche du comportement humain. Pour ce faire, nous avons calculé les cartes de saillance artificielle pour chacune des 4 méthodes d'extraction de l'attention post-hoc pour chacune des 250 images de la base. Les cartes d'attention ont été calculées pour les trois modèles présentés dans le chapitre 3 : VGG16, VGG19 et ResNet34, et pour les 10 validations croisées. Une fois ces 10 000 cartes calculées, nous avons réalisé la comparaison entre l'attention humaine et l'attention artificielle en comparant chacune des cartes obtenues avec celles de chacun des 22 participants.

Les résultats pour les métriques CC et NSS présentés dans le chapitre 2 sont consignés dans les tableaux 4.2 et 4.3 respectivement.

On peut observer que, pour chacun des réseaux, la méthode obtenant en moyenne, sur l'ensemble des cartes d'attention extraites, le meilleur résultat est la méthode des gradients. En effet, en moyenne, la méthode des gradients obtient pour la métrique CC une valeur moyenne de 0,169 pour l'ensemble des mesures. La méthode de rétropropagation guidée obtient une moyenne de 0,088 quand la méthode de Guided GradCAM obtient une valeur moyenne de 0,117. Finalement la méthode GradCAM obtient les moins bons résultats avec une valeur moyenne de 0,099. La même dynamique de score est observable avec la métrique NSS.

On remarque également qu'en moyenne les résultats sont meilleurs sur l'ensemble des méthodes pour le réseau ResNet34.

Méthode	VGG16	VGG19	ResNet34	Moyenne
GradCAM	0.099	**0.196**	0.128	0.141
Guided GradCAM	0.117	0.187	0.137	0.147
Rétropropagation guidée	0.088	0.074	0.134	0.098
Gradients	**0.169**	0.123	**0.207**	**0.166**

TABLE 4.2 – Tableau récapitulatif de la comparaison entre attention humaine et attention artificielle. Pour chacune des méthodes d'attention post hoc, nous avons calculé la valeur moyenne du score CC sur l'ensemble des images. Ce tableau nous permet d'observer qu'en moyenne la méthode des gradients est la plus proche du comportement attentionnel humain.

Méthode	VGG16	VGG19	ResNet34	Moyenne
GradCAM	0.302	0.253	0.429	0.328
Guided GradCAM	0.382	0.273	0.512	0.389
Rétropropagation guidée	0.338	0.281	0.501	0.373
Gradients	**0.586**	**0.420**	**0.696**	**0.567**

TABLE 4.3 – Tableau récapitulatif de la comparaison entre attention humaine et attention artificielle. Pour chacune des méthodes d'attention post hoc, nous avons calculé la valeur moyenne du score NSS sur l'ensemble des images. Ce tableau nous permet d'observer qu'en moyenne la méthode des gradients est la plus proche du comportement attentionnel humain.

Nous avons vu dans le chapitre 3 que les résultats attentionnels ne semblaient pas être les mêmes lorsque l'image était non pathologique que lorsqu'elle était pathologique. Quel est l'influence du label sur ces mesures ? De plus, les différents participants à l'expérience d'eye tracking n'ont pas réalisé les mêmes performances du fait de leur différent niveau d'expertise. Quel est l'influence de l'expertise humaine sur cette comparaison ?

4.3 Influence de l'expertise des participants et du label des images sur l'attention humaine

Dans cette partie nous allons nous intéresser à l'influence du label des images et de l'expertise des participants sur les zones d'intérêt humaines et artificielles. Dans un premier temps, nous comparerons les humains entre eux selon les groupes d'expertise définis

dans la section 4.1.4. Pour ce faire, nous observerons la distance au centre et la dispersion des fixations des participants. La distance au centre permet d'évaluer si les comportements attentionnels sont éloignés ou non du biais centrée, soit la tendance à observer majoritairement le centre d'une image. La dispersion des fixations indique si les zones d'attention sont concentrées ou si plutôt elles sont aléatoirement réparti dans l'image. Dans un deuxième temps nous évaluerons de quel groupe d'expertise le comportement attentionnel artificiel est le plus proche. Cette analyse sera réalisée à l'aide des métriques présentées dans la partie 2.3.3. Finalement, dans un dernier temps, nous verrons comment la comparaison entre attention humaine et artificielle évolue au fur et à mesure que le réseau s'entraîne. Pour permettre une lecture plus facile, nous présenterons principalement les résultats obtenus par la méthode des gradients lorsque nous comparerons les stratégies attentionnelles entre humain et machine. Nous prendrons comme référence le réseau ResNet34, ce dernier ayant obtenu les meilleures performances à la tâche de classification ainsi que le comportement attentionnel moyen le plus proche de l'humain. Les différents résultats présentés dans cette section seront évalués à l'aide de modèles linéaires mixtes présentés dans la partie 2.6.

4.3.1 Dispersion et distance au centre

La distance moyenne au centre de l'écran et la dispersion moyenne du regard sont deux mesures simples et très interprétables pour quantifier le comportement du regard d'un participant. La distance par rapport au centre peut représenter le degré d'activité d'un participant pendant l'exploration d'un stimulus. Si le participant est passif, il restera probablement près du centre de l'image en attendant que l'expérience passe à autre chose, tandis qu'un participant plus proactif aura un comportement du regard plus dynamique et explorera des zones plus éloignées du centre de l'écran. La dispersion des positions des yeux quantifie la façon dont le participant a réparti son attention sur les stimuli. Si le participant a regardé en dispersant les positions de ses yeux sur toute l'image, la dispersion sera élevée. Si le participant s'est concentré sur une zone spécifique, la dispersion sera faible. Pour chaque mesure, nous avons calculé un modèle mixte linéaire avec l'expertise (expert, non expert), l'étiquette (pathologique, non pathologique) et leur interaction en tant qu'effets fixes, avec des intercepts aléatoires pour chaque participant et chaque image :

$$\text{score} \sim \text{expertise*étiquette} + (1 \mid \text{participant}) + (1 \mid \text{image})$$

Pour la distance au centre, l'effet de l'étiquette n'était pas significatif $t(1, 5994) = 1.2387, p = 0.21$, de même que l'effet de l'expertise, $t(1, 5994) = -0.896, p = 0.37$. Cependant, nous avons trouvé un effet significatif de l'interaction entre l'expertise et l'étiquette, $t(2, 5994) = -2.1866, p = 0.03$.

Pour la dispersion, l'effet de l'étiquette était significatif $t(1, 5994) = -6, 1922, p < 0, 001$, mais pas l'effet de l'expertise $t(1, 5994) = -1, 183, p = 0, 23$. Nous avons trouvé un effet significatif de l'interaction entre expertise et label, $t(2, 5994) = 5, 33, p < 0, 001$. Comme le montre la figure 4.3, la dispersion est plus élevée dans les images non pathologiques que dans les images pathologiques à la fois pour les experts et les non experts, et cet effet est plus fort chez les experts. Cela peut être interprété comme le fait que les lésions présentes dans les images pathologiques attirent l'attention, diminuant ainsi la dispersion. Les experts étant plus susceptibles de les repérer, cet effet est logiquement renforcé dans ce groupe.

La distance par rapport au centre et la dispersion du regard quantifient grossièrement le comportement du regard mais ne permettent pas de savoir si les observateurs ont spécifiquement insisté aux mêmes endroits. Dans la prochaine section, nous étudierons la comparaison de la distribution spatiale de l'attention avec les métriques introduites dans la 2.3.3.

4.3.2 Comparaison des cartes d'attention entre humains avec les métriques de saillance

Pour les images pathologiques et non pathologiques, nous avons évalué l'effet de l'expertise sur la distribution spatiale de l'attention à travers deux types de comparaison : intra-groupe et inter-groupe, voir figure 4.4. Pour la comparaison intra-groupe (voir les boxplots "novices vs novices" et "experts vs experts"), nous avons créé une carte d'attention de référence en faisant la moyenne des cartes d'attention de chaque membre d'un groupe à l'exclusion de celui qui est traité (procédure leave-one-out). Grâce aux métriques décrites dans 2.3.3, nous avons obtenu 2 scores (NSS, CC) pour chaque sujet, sur chaque image, indiquant si le sujet sur une image donnée regardait les mêmes zones que les membres de son propre groupe. Nous avons également réalisé une comparaison inter-groupe, en faisant la moyenne des cartes d'un groupe pour chaque image et en comparant cette carte avec toutes les cartes de l'autre groupe.

Comme pour la dispersion et la distance au centre, nous avons calculé pour le score

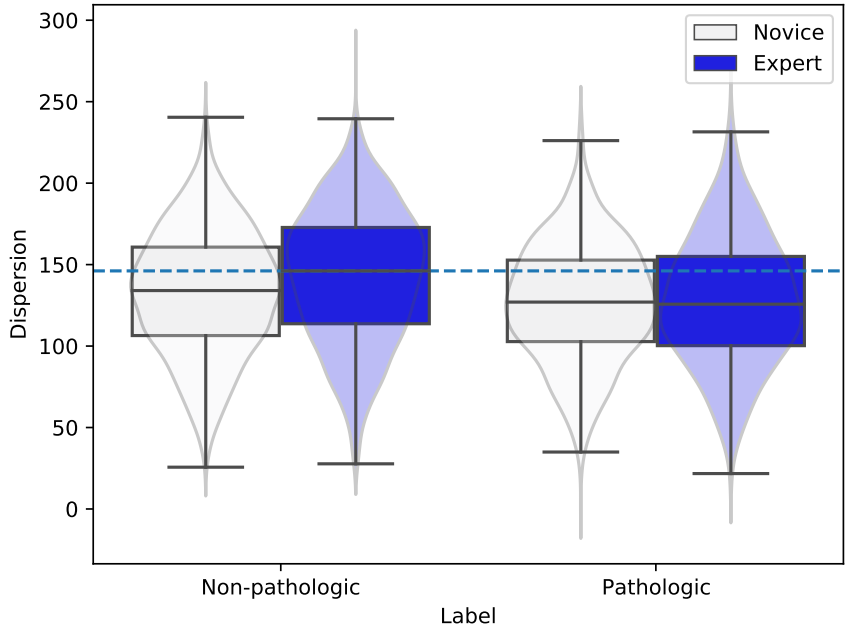


FIGURE 4.3 – Dispersion des positions des yeux par expertise et par étiquette. La ligne de référence horizontale correspond à la valeur de dispersion la plus élevée (experts regardant des images non pathologiques).

NSS un modèle mixte linéaire avec l'étiquette, l'expertise et leur interaction comme effets fixes, et les images et les observateurs comme effets aléatoires. Nous avons trouvé un effet significatif du label de l'image, $t(1, 10776) = 11.62, p < 0.001$ et de son interaction avec le niveau d'expertise des observateurs, $t(2, 10776) = -14.01, p < 0.001$. L'effet de l'expertise n'était pas significatif, $t(1, 10776) = -0,17, p = 0,86$. Comme le montre la figure 4.4, le score NSS était plus élevé pour les images pathologiques que pour les images non pathologiques. Cela montre que sur les images pathologiques, la distribution spatiale de l'attention visuelle est plus similaire entre les observateurs que sur les images non pathologiques. Cela pourrait être dû à la présence de lésions guidant l'attention vers les mêmes zones.

Pour étudier plus en détail l'interaction entre l'étiquette et l'expertise, nous avons calculé deux modèles mixtes linéaires indépendants avec l'expertise comme effet fixe et des intercepts aléatoires pour les participants et les images. Le premier modèle utilise uniquement des images pathologiques, le second uniquement des images non pathologiques. Nous avons trouvé un effet significatif de l'expertise avec les images pa-

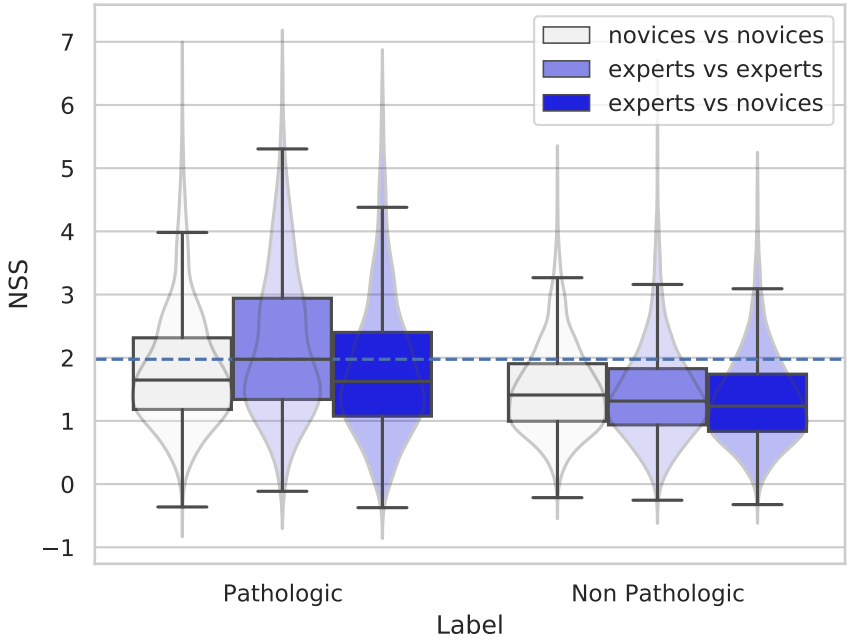


FIGURE 4.4 – Comparaison de cartes d'attention humaines. La similarité entre les cartes d'attention est évaluée par le score de salinité normalisé (NSS, plus il est élevé, plus les distributions d'attention sont proches),pour les images pathologiques et non pathologiques. Les étiquettes « novices vs novices » et « experts vs experts » correspondent aux comparaisons intra-groupes. L'étiquette "experts vs novices" correspond à la comparaison inter-groupe.

thologiques ($t(1,4442) = -2.36, p = 0.02$), mais pas sur les images non pathologiques ($t(1,6334) = -0.21, p = 0.83$). Ceci pourrait être interprété comme un effet des lésions pathologiques attirant davantage l'attention des experts sur les mêmes zones, alors que les non experts sont moins guidés par celles-ci. Sur les images non pathologiques, rien ne guide l'attention visuelle spatiale ni des experts ni des non experts, d'où l'absence de différence entre les groupes.

4.4 Cartes d'attention artificielle vs humaine

Nous avons vu dans la partie précédente que le comportement attentionnel humain était influencé par les labels des images, comme cela était le cas pour le comportement attentionnel artificiel (voir partie 3.3). Nous avons également montré que le comportement

attentionnel humain était influencé par son niveau d'expertise. Nous souhaitons donc maintenant comparer les comportements attentionnels humain et artificiel et observer si cette comparaison est également affectée par le label des images et l'expertise de la machine et de l'humain. Sur la figure 2.7.2 on peut observer que le comportement de l'attention artificielle semble plus proche de celui des experts que de celui des novices.

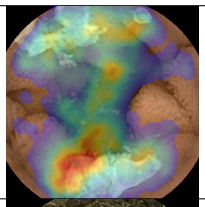
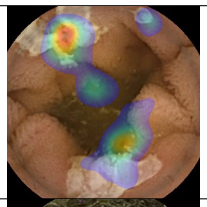
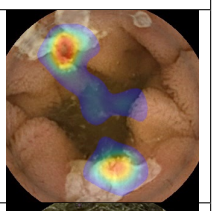
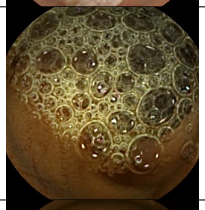
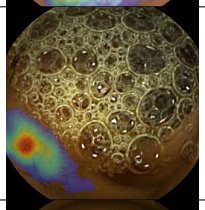
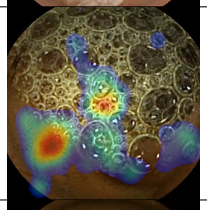
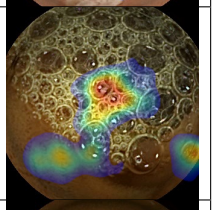
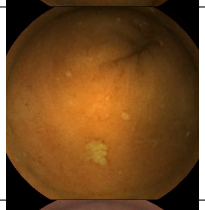
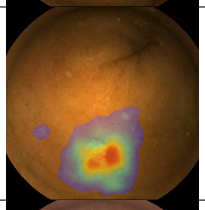
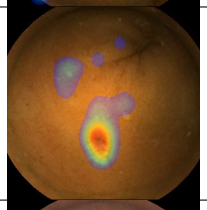
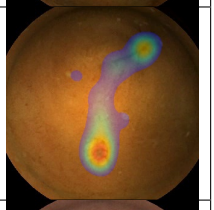
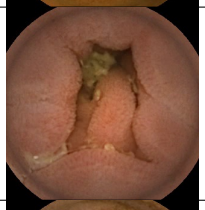
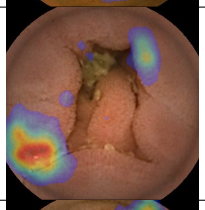
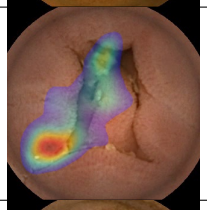
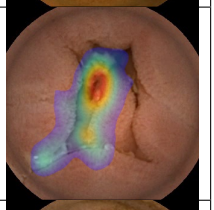
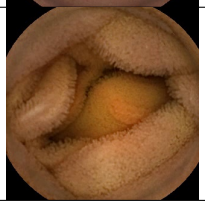
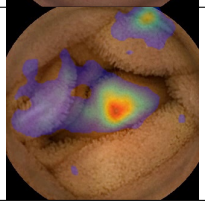
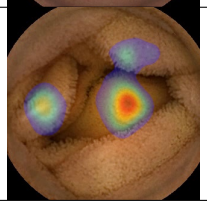
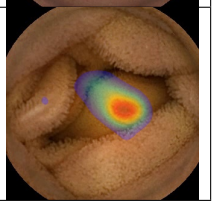
Matrice de confusion	Originale	ResNet34	Experts	Novices
Vrai positif				
Vrai Négatif				
Faux positif				
Faux Négatif				
Non concluante				

TABLE 4.4 – Exemples d'attentions humaines et artificielles sur les différents exemples présentés dans la partie 3.2. Les cartes d'attention humaine ont été moyennées en fonction de leur groupe d'expertise.

4.4.1 Avec la méthode des gradients

Ici, nous comparons les cartes d'attention humaine et artificielle. Comme décrit dans la section 2.7.2, nous avons utilisé la méthode des gradients pour obtenir pour chaque image la carte des pixels impliqués dans le processus de décision du réseau de neurones profond ResNet34.

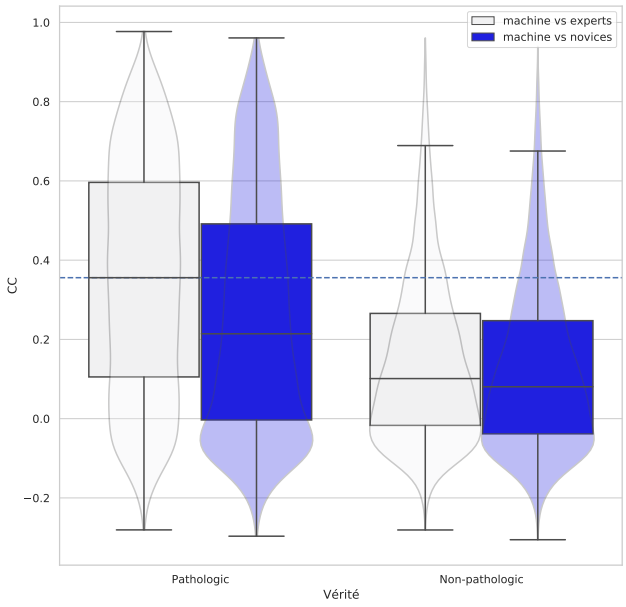


FIGURE 4.5 – Comparaison de la similarité entre la distribution spatiale de l'attention humaine et celle de la machine (CC plus haut signifie plus similaire), pour des images pathologiques et non pathologiques. Le label "machine vs non experts" correspond à la comparaison entre le groupe de non experts et l'attention artificielle extraite par la méthode des gradients. Le label "machine vs experts" correspond à la même comparaison avec le groupe d'experts.

Nous avons calculé le même modèle linéaire mixte que dans 4.3.1 et 4.3.2 (effet fixe : étiquette, expertise et leur interaction, effet aléatoire : images et observateurs), cette fois pour le score NSS entre les cartes d'attention machine et humaine. Ce modèle s'applique à 55 000 observations (250 images $\times$ 22 observateurs $\times$ 10 validations croisées). Comme le montre la figure 4.6, les mêmes effets que dans la comparaison de l'attention humaine sont visibles. Nous avons constaté que l'étiquette a un effet significatif sur le score NSS,

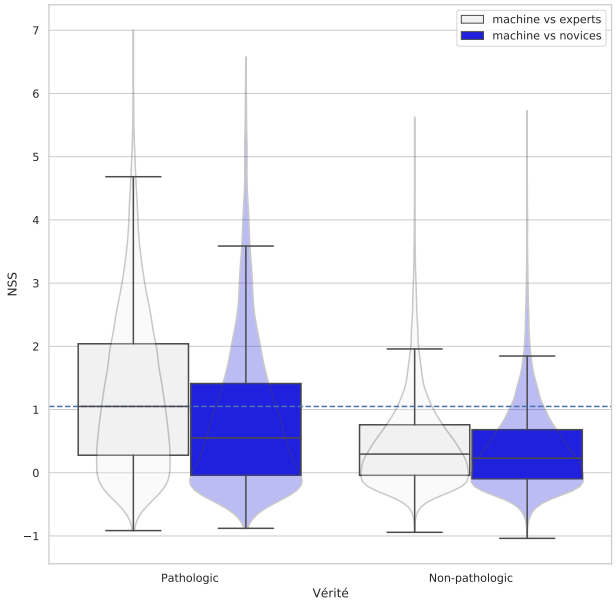


FIGURE 4.6 – Comparaison de la similarité entre la distribution spatiale de l'attention humaine et celle de la machine (NSS plus haut signifie plus similaire), pour des images pathologiques et non pathologiques. Le label "machine vs non experts" correspond à la comparaison entre le groupe de non experts et l'attention artificielle extraite par la méthode des gradients. Le label "machine vs experts" correspond à la même comparaison avec le groupe d'experts.

$t(55000) = 10.508, p = 1e-26$, avec des cartes d'attention plus similaires pour les images pathologiques que pour les images non pathologiques. Comme prévu, le contenu de l'image détermine le comportement d'attention artificielle.

Pour mieux comprendre l'interaction entre l'étiquette et l'expertise sur la comparaison entre l'attention artificielle et humaine, nous avons calculé deux modèles mixtes linéaires indépendants avec l'expertise comme effet fixe et des intercepts aléatoires pour les participants et les images. Le premier modèle utilise uniquement des images pathologiques, le second uniquement des images non pathologiques. Pour les deux modèles, nous trouvons un effet significatif de l'expertise avec $t(22440) = -2.7379, p = 0.006$ pour les images pathologiques et $t(22440) = -2.0972, p < 0.035$ sur les images non pathologiques. Les cartes d'attention de la machine sont plus similaires aux cartes d'attention des experts qu'aux cartes d'attention des non experts. Ceci est cohérent avec les performances de clas-

sification du modèle d'apprentissage profond, qui sont plus proches de celles du groupe d'experts que de celles du groupe des non experts (précision = 84.1 %, voir tableau 4.1).

Les résultats observés avec le réseau ResNet34 sont également visibles avec les réseaux VGG16 et VGG19. Un effet significatif du label est présent sur l'ensemble des données (VGG16 : $t(55000) = 12.687, p = 7e-37$, VGG19 : $t(55000) = 14.338, p = 1e-46$), et un effet de l'expertise est visible sur les images pathologiques (VGG16 : $t(22440) = -2.3538, p = 0.018$, VGG19 : $t(22440) = -2.3509, p = 0.018$).

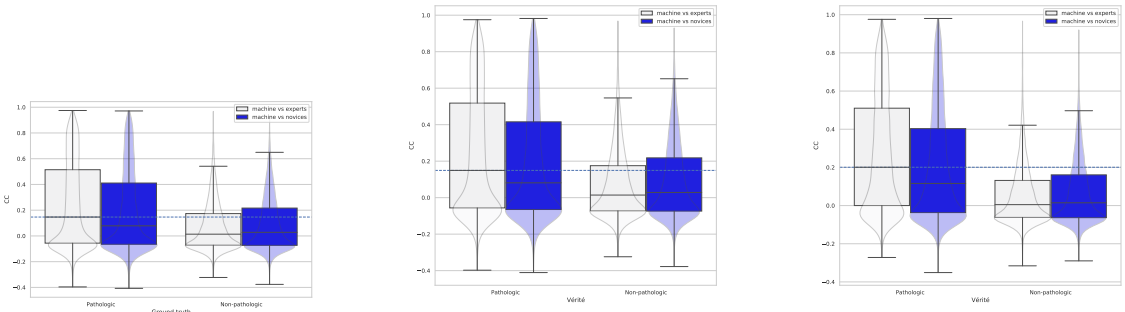
4.4.2 Pour les autres méthodes

Les différents modèles linéaires mixtes ont été calculés pour les trois réseaux de neurones profonds et pour les trois autres méthodes d'extraction de l'attention post-hoc. Les résultats des comparaisons entre l'attention machine et l'attention humaine pour ces trois autre méthodes sont présentés dans la figure 4.7. Les résultats avec la métrique NSS, eux, sont présents sur la figure B.1 annexe. Pour ces méthodes d'extraction de l'attention post-hoc, on observe un effet significatif du label sur l'ensemble des images, et un effet significatif de l'expertise sur les images pathologiques. Cependant, c'est seulement avec la méthode des gradients que l'effet de l'expertise sur les images non pathologiques est significatif. L'effet de l'expertise sur les images pathologiques est moins fort pour les autres méthodes que pour la méthode des gradients.

Avec la méthode de rétropropagation guidée, on obtient sur l'ensemble des images, un effet du label significatif sur le score NSS avec $t(55000) = 12.015, p = 3e-33$. Quand le test est réalisé seulement sur les images pathologiques, un l'effet de l'expertise est également significatif avec $t(22440) = -2.2355, p = 0.025$. Pour la méthode GradCAM, l'effet du label est significatif sur l'ensemble des images sur la métrique NSS avec $t(55000) = 9.1345, p = 6e-20$ et un effet de l'expertise visible sur les images pathologiques $t(22440) = -2.085, p = 0.0370$. Pour la méthode Guided GradCAM, l'effet du label comme pour les autres modèles est significatif sur la métrique NSS $t(55000) = 9.1059, p = 8e-20$ et un effet de l'expertise est significatif sur les images pathologiques $t(22440) = -2.085, p = 0.037$.

La dynamique des résultats présentés précédemment reste valable pour l'ensemble des modèles et également avec le score CC comme présenté dans la figure 4.7.

Les résultats des différentes méthodes sur VGG16 et VGG19 avec la métriques CC sont respectivement présents sur les figures B.2 et B.3. Pour la métrique NSS, les résultats sont présents en B.4 et B.5.



(a) GradCAM (b) Guided GradCAM (c) Rétropropagation guidée

FIGURE 4.7 – Résultats de la comparaison des différentes méthodes d'extraction de l'attention artificielle post-hoc avec l'attention humaine en fonction du niveau d'expertise et du label des images. Ces résultats ont été obtenus pour le réseau ResNet34 et la comparaison a été réalisée à l'aide de la métrique CC. Plus le score de CC est grand, plus les zones d'attention entre machine et humains sont proches.

4.5 Évolution de la comparaison attentionnelle au cours de l'apprentissage

Nous avons vu que le comportement de l'attention artificielle était plus proche du comportement attentionnel des experts sur les images pathologiques. Les performances des trois réseaux de neurones profonds étant plus proches de celles des experts que de celles des novices, nous interrogerons ici l'influence des performances du réseau sur cette comparaison. Nous avons vu dans la partie 4.3 que plus les performances à la tâche de classification des participants étaient élevées, plus les différentes stratégies attentionnelles étaient homogènes. Ainsi en réalisant une comparaison entre des cartes attentionnelles extraites à différents moments de l'entraînement et donc sur des réseaux avec différents niveaux de performances, nous cherchons à vérifier l'hypothèse selon laquelle les comportements attentionnels évoluent en fonction du niveau d'expertise de l'algorithme, et que plus ses performances sont élevées, plus son comportement attentionnel est proche du comportement d'un expert.

Pour vérifier cette hypothèse, nous avons enregistré les poids des réseaux à différents moments de leur apprentissage sur la base de données CrohnIPI. Pour chacun des réseaux, les poids ont été enregistrés pour les 10 premières epochs, ainsi que pour l'état initial où le réseau est pré entrainé sur ImageNet. Les poids ont ensuite été enregistrés toutes les 5

epochs jusqu'à 50 epochs après que le réseau ait atteint l'erreur minimum sur le jeu de validation. Pour chacune des 10 validations croisées réalisées pour chacun des 3 réseaux, les cartes attentionnelles obtenues avec chacune des 4 méthodes ont été comparées avec les cartes de l'ensemble des 22 participants à l'expérience d'eye-tracking.

Pour approfondir la relation entre l'expertise et le comportement attentionnel, nous avons calculé des cartes d'attention artificielle à différents moments de la formation du réseau et effectué la même comparaison avec les cartes d'attention humaine. Pour chaque carte d'attention artificielle, nous avons calculé $\Delta_{NSS} = NSS_{experts} - NSS_{non-experts}$ avec $NSS_{experts}$ et $NSS_{non-experts}$ respectivement la valeur NSS moyenne sur toutes les images entre la carte d'attention artificielle et les cartes d'attention des experts et des non experts. La figure 4.8 montre que plus le réseau devient précis, plus ses cartes d'attention se rapprochent des cartes d'attention des experts humains pour chacun des trois réseaux. On obtient un coefficient de corrélation de Pearson de $r = 0.438, p < 0.001$ pour ResNet34, de $r = 0.494, p < 0.001$ pour VGG16 et de $r = 0.371, p < 0.001$ pour VGG19.

Pour les autres méthodes, la même tendance est observable sur l'ensemble des images. Plus les performances du réseau sont hautes, plus le comportement attentionnel de la machine est proche de celui des experts. Les résultats pour les différentes méthodes sont présentés dans la figure 4.9. Pour ces méthodes, l'effet des performances du réseau de neurones est encore plus visible sur les images pathologiques comme il est possible de le voir sur la figure 4.11.

Les résultats obtenus pour les seules images pathologiques sont présents sur la figure 4.10. On peut remarquer que pour la méthode des gradients, en moyenne, la différence entre le score NSS obtenu entre les experts et les non experts est plus marquée. Les coefficients de corrélation obtenus pour chacun des modèles sont proches de ceux obtenus sur l'ensemble des images. Pour le modèle ResNet34 on observe un coefficient $r = 0.481, p < 0.0001$, pour VGG16 $r = 0.497, p < 0.0001$ et pour VGG19 $r = 0.354, p < 0.0001$.

Cela valide notre conclusion précédente : il existe une relation entre la précision obtenue dans notre tâche de diagnostic et le comportement attentionnel des humains et de la machine. Plus les performances du réseau de neurones profond sont élevées, plus son comportement attentionnel se rapproche de celui des humains qui ont obtenu de meilleurs résultats sur la même tâche.

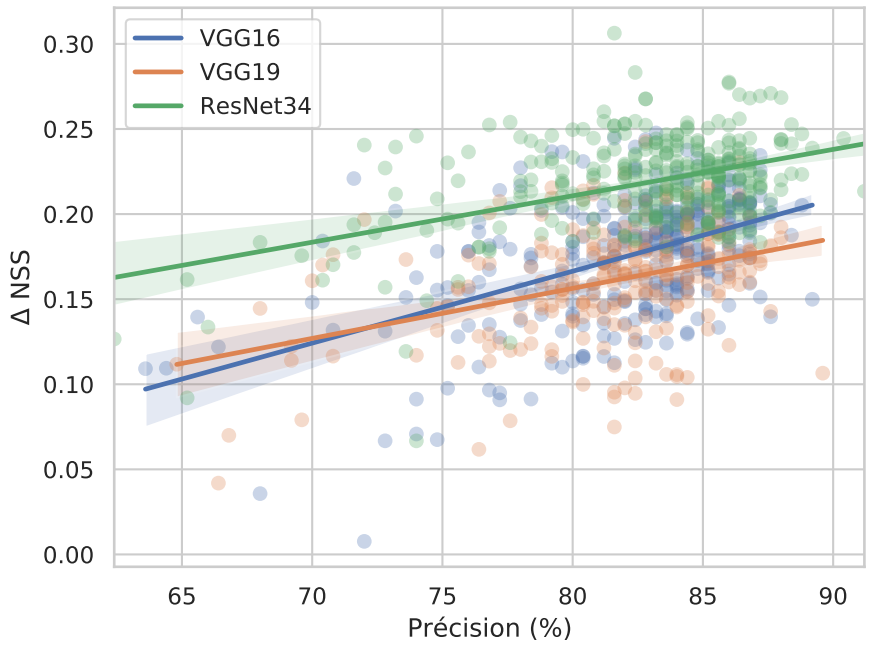
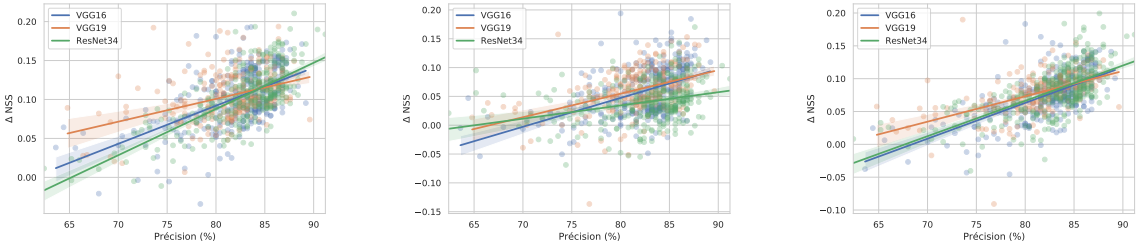


FIGURE 4.8 – Corrélation entre les performances du réseau et la similarité entre ses cartes d'attention et les cartes d'attention des experts humains. La variable Δ_{NSS} indique si les cartes d'attention du réseau sont plus proches des cartes d'attention des experts que de celles des non experts. Plus le Δ_{NSS} est élevé, plus les cartes d'attention du réseau sont proches de celles des experts. Chaque point correspond à la moyenne des Δ_{NSS} pour toutes les images en fonction de la performance atteinte lors de l'entraînement du réseau. Les courbes correspondent aux régressions linéaires pour chacun des modèles, avec $r = 0.438, p < 0.001$ pour ResNet34, $r = 0.494, p < 0.001$ pour VGG16 et $r = 0.371, p < 0.001$ pour VGG19.

4.6 Discussion

Nous avons construit un ensemble de données pour une tâche spécifique de classification d'imagerie médicale. Nous avons enregistré et comparé l'attention visuelle d'experts médicaux, de novices et de réseaux neuronaux profonds pendant la réalisation de cette classification. Un certain nombre de biais ont été identifiés dans la littérature lors de la comparaison entre la perception visuelle humaine et celle de la machine [197]. Pour les éviter, nous avons essayé de concevoir la tâche de classification humaine et machine de manière aussi proche que possible. Tout d'abord, puisque le nombre de paramètres était identique pour chaque image, nous avons pu fixer le nombre de calculs effectués par



(a) Rétropropagation guidée (b) GradCAM (c) Guided GradCAM

FIGURE 4.9 – Résultats pour les différents réseaux avec la méthode de l'évolution de la comparaison de l'attention entre humain et machine en fonction de la précision de classification du réseau sur la métrique NSS sur **l'ensemble des images**.

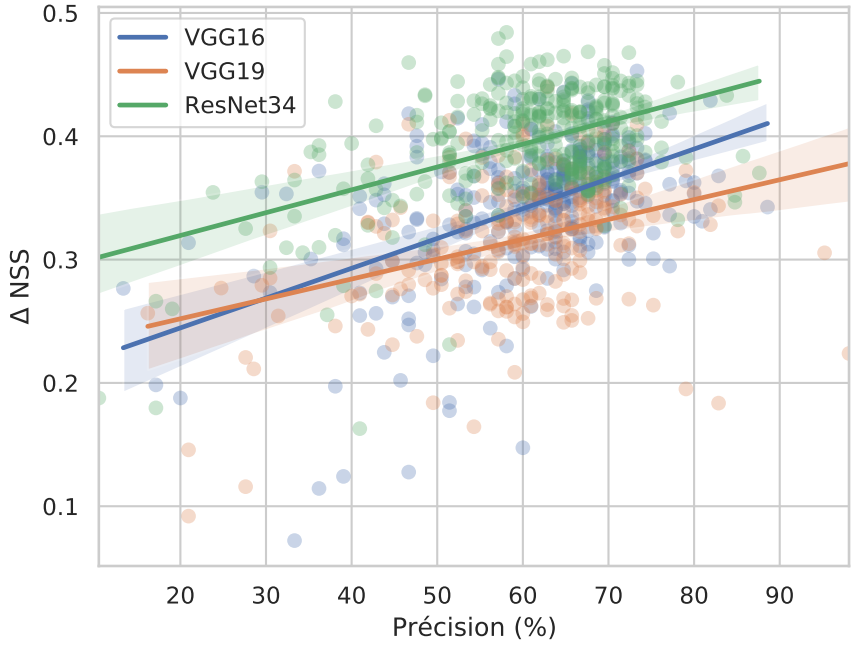
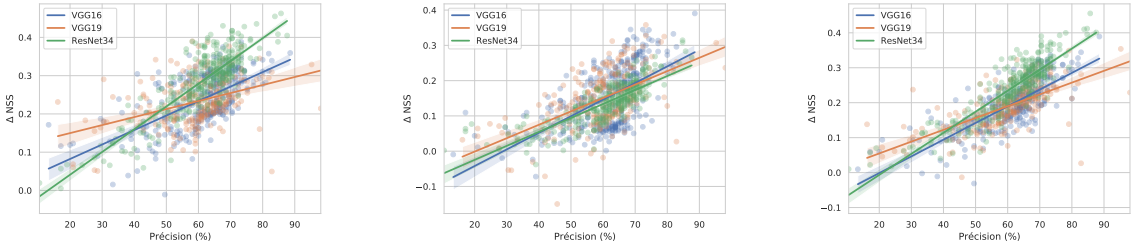


FIGURE 4.10 – Corrélation entre les performances du réseau et la similarité entre ses cartes d'attention et les cartes d'attention des experts humains sur les images **pathologiques**.

le réseau. En supposant que le nombre de calculs effectués quasi-simultanément par la machine correspond au temps de visualisation des images pour les humains, nous avons également fixé le même temps de visualisation (2 secondes) pour chaque image. Ensuite, chaque image a été visualisée par le réseau indépendamment des images précédentes et suivantes, et sans aucune information sur le patient. De même, nous avons présenté les



(a) Rétropropagation guidée (b) GradCAM (c) Guided GradCAM

FIGURE 4.11 – Résultats pour les différents réseaux avec la méthode de l'évolution de la comparaison de l'attention entre humain et machine en fonction de la précision de classification du réseau sur la métrique NSS sur les images **pathologiques**.

images aux humains indépendamment de leur contexte clinique, et dans un ordre aléatoire différent pour chaque participant. Troisièmement, les valeurs des paramètres ont été fixées, calculées à partir d'un entraînement précédent sur notre ensemble de données. De même, aucun retour n'a été donné aux participants, les empêchant d'apprendre pendant l'expérience.

Nous avons constaté que les stratégies d'attention visuelle sont plus similaires entre les experts qu'entre les novices. Cela indique qu'une performance élevée dans notre tâche de diagnostic médical est corrélée à un comportement attentionnel spécifique. Nous avons comparé l'attention des experts avec l'attention artificielle, extraite d'un réseau neuronal convolutif de l'état de l'art avec la méthode des différentes méthodes d'extraction de l'attention post-hoc. Nous avons montré que lorsque les performances du réseau se rapprochaient de celles des experts, son comportement attentionnel se rapprochait également de celui des experts, principalement sur les images pathologiques. Cette corrélation s'est renforcée lorsque le réseau s'est amélioré dans la tâche de classification.

Nous avons été surpris par la performance de certains novices bien au-dessus de la chance, même sans aucun retour des expérimentateurs. Il pourrait être intéressant d'évaluer l'évolution de leurs performances tout au long de la tâche avec un algorithme d'apprentissage profond non supervisé. Les novices, lors de l'évaluation des images endoscopiques, apprennent dans un premier temps à distinguer les images normales des images anormales puis regroupent les images anormales dans des sous-catégories. Ce fonctionnement peut faire penser au fonctionnement des algorithmes de regroupement (*clustering*).

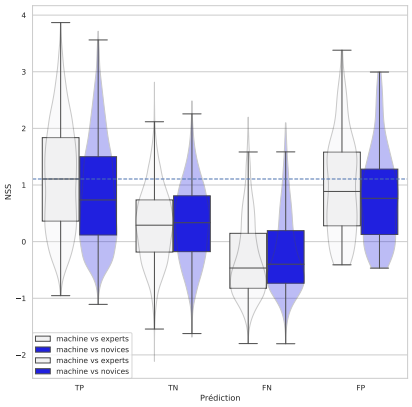
La comparaison des différentes méthodes d'extraction de l'attention post-hoc nous permet d'évaluer celle la plus susceptible d'aider les experts médicaux à la réalisation de

leur diagnostic. Face à une vérité terrain composée de données oculométriques provenant de professionnels médicaux, nous observons que les résultats obtenus par les différentes méthodes ne sont pas égaux. Il est possible de les classer de la meilleure à la moins performante au vue des différentes expériences réalisées : premièrement la méthode des gradients, puis la méthode de rétropropagation guidée, suivie de guided GradCAM et finalement GradCAM.

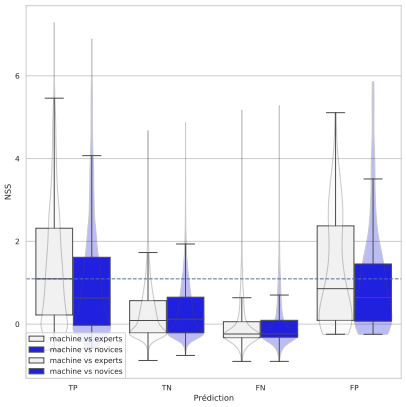
Dans l'article de Adebayo et al. 2018 [200], une comparaison des méthodes d'attention post-hoc a été réalisée au travers de deux tests. Le premier test consiste en la comparaison des cartes obtenues pour un réseau initialisé aléatoirement et pour un réseau entraîné sur une tâche précise. Si la méthode est effective, les résultats de saillance obtenus devraient être différents, montrant ainsi que la méthode est dépendante des paramètres des modèles. Le deuxième test consiste à entraîner deux modèles similaires sur un jeu de données contenant les mêmes images mais dont les labels ont été permutés de façon aléatoire pour le deuxième. Au travers de cette expérience, les auteurs cherchent à identifier si la méthode d'extraction de l'attention est bien sensible à la relation entre image et label. Ils concluent que, parmi les différents méthodes de l'extraction de l'attention artificielle, seule la méthode des gradients et la méthode GradCAM passent les deux tests.

Ici, dans ce manuscrit, nous montrons que la méthode des gradients est bien supérieure aux trois autres méthodes. En effet, elle est plus stable (voir 3.3), son comportement attentionnel est plus proche de celui des experts que de celui des non experts et ce aussi bien sur les images pathologiques que sur les images non pathologiques. Les meilleurs résultats sur les images non pathologiques peuvent s'expliquer par le fait qu'à la différence des autres méthodes qui ne rendent compte que des parties de l'image ayant contribué de façon positive à la prédiction, la méthode des gradients rend également compte des éléments ayant diminué le score de la prédiction finale. Bien que l'algorithme ait répondu « non pathologique », la méthode montre des zones l'ayant fait douté, et permet ainsi de rendre compte d'erreurs de reconnaissance (voir partie 2.4.3). Pour les autres méthodes, toutes les erreurs sont assimilées à des erreurs de détection, puisque les parties « observées » par le réseau mais n'ayant pas permis de conclure à une présence de pathologie sont effacées par l'application d'une ReLU sur les gradients. On peut observer sur la figure 4.12 que pour la méthode des gradients et à la différence des autres méthodes l'écart entre les comportements attentionnels artificiels et humains est faible sur les faux négatifs. Bien que l'image ait été mal classifiée par le réseau, son comportement attentionnel est aussi proche de celui d'un expert que si l'image est pathologique et correctement classifiée.

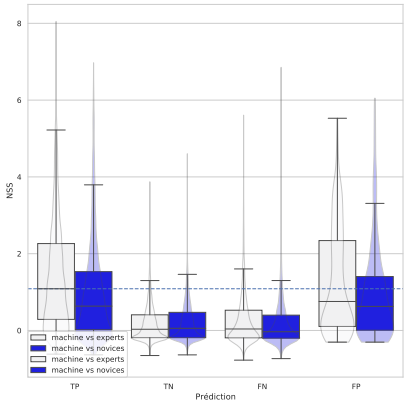
Notre jeu de données vient combler le manque de données nécessaires à l'évaluation des méthodes d'extraction de l'attention artificielle. Nous le mettons à la disposition de tout chercheur qui souhaiterait évaluer ses propres modèles d'attention. Nous allons plus loin dans la compréhension de la stratégie d'attention artificielle en montrant une forte corrélation entre le comportement de l'homme et celui du réseau de neurones profonds, corrélé avec le niveau d'expertise sur une tâche spécifique. Grâce à la compréhension des similitudes entre le processus de prise de décision visuelle des experts humains et machines, nous espérons contribuer à la formation de nouveaux médecins et au développement d'architecture de nouveaux réseaux de neurones profonds.



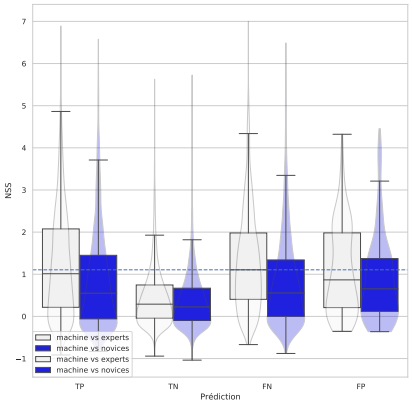
(a) GradCAM



(b) Guided GradCAM



(c) Rétropropagation guidée



(d) Gradients

FIGURE 4.12 – Résultats de la comparaison des différentes méthodes d'extraction de l'attention artificielle post-hoc avec l'attention humaine en fonction du niveau d'expertise des prédictions et des labels des images. Les abréviations TP, TN, FN et FP correspondent respectivement aux vrais positifs, vrais négatifs, faux négatifs et faux positifs. Ces résultats nous montrent que seule la méthode des gradients permet de rendre compte que lorsque l'algorithme commet une erreur de prédiction, son comportement attentionnel est proche de celui des humains ayant correctement classé l'image. Cela nous indique que les erreurs des réseaux doivent être majoritairement des erreurs de reconnaissances ou de diagnostic et non des erreurs de détection.

Conclusion

Contributions

Crohn IPI : une base de données publique

La première contribution de cette thèse est la mise à disposition d'un jeu de données public permettant l'entraînement et le test des algorithmes de détection de lésions issues de la maladie de Crohn dans des vidéos capsules endoscopiques. Les étapes de construction de ce jeu de données sont clairement décrites. Dans la littérature, il n'existe à notre connaissance aucune base de données publique permettant l'entraînement des algorithmes d'aide au diagnostic de la maladie de Crohn. La base de données a déjà été partagée à plusieurs équipes de recherche à travers le monde et permettra l'évaluation de classifieurs. Cette base de données possède l'avantage sur bon nombre de base de données médicales d'être associée à une description précise de son processus de création. Le processus d'étiquetage des images s'est déroulé en trois phases, impliquant trois experts gastro-entérologues et une interne en médecine. Les réflexions autour du processus d'annotation et de la mise à disposition du public de ce jeu de données sont une contribution majeure au domaine médical. Nous montrons qu'après chacune des phases, les performances obtenues sur des réseaux de neurones de l'état de l'art s'améliorent. Cela montre que la qualité des données s'est effectivement améliorée, le nombre de faux positifs et faux négatifs diminuant après chacune des phases, rendant l'apprentissage meilleur pour les réseaux de neurones profonds. Bien que l'effet de la quantité de données sur l'augmentation des performances des modèles prédictifs ait déjà été largement démontré [201], obtenir des jeux de données de volume conséquent est difficile dans le domaine médical. Nous montrons ici que l'influence de la qualité du jeu de données influe sur la qualité des prédictions des algorithmes d'apprentissage profond et invitons les chercheurs dans ce domaine à réfléchir davantage en terme de qualité plutôt qu'uniquement en terme de quantité.

Application de l'attention artificielle à la détection des lésions de la maladie de Crohn

Nous avons conçu une architecture de réseau de neurones à attention dure obtenant des performances similaires à des réseaux convolutifs de l'état de l'art. Avec une stratégie attentionnelle plus proche de celle des humains, car séquentielle, nous montrons qu'il est possible d'obtenir des résultats viables pouvant être adaptés à la routine clinique des experts médicaux. Au travers de l'utilisation de l'apprentissage par renforcement pour des algorithmes décisionnels appliqués à des images, nous espérons encourager les chercheurs à continuer à pousser le mimétisme humain/machine en utilisant des processus de décision séquentiels plus proches de celui des humains.

Nous étudions également l'application de 4 méthodes d'extraction de l'attention post hoc à notre problématique de détection des lésions de la maladie de Crohn. Nous montrons que chacune de ces méthodes rend compte d'une activité attentionnelle différente, bien que les poids du réseau à partir duquel ils ont été extraits soient les mêmes. La comparaison de la stabilité de ces résultats en fonction des différentes validations croisées nous fait penser que ces méthodes ne sont pas égales en terme de performances.

Comparaison entre l'attention humaine et artificielle

La comparaison entre l'attention humaine et l'attention artificielle nous a permis de tirer plusieurs conclusions. Premièrement, les comportements attentionnels artificiels et humains sont influencés par le label de l'image. Cette conclusion avait déjà été observée lors de la comparaison entre les comportements attentionnels des réseaux entre eux, et reste vraie lors de la comparaison avec des données oculométriques réelles. De plus, l'effet de l'expertise humaine peut être observé lors de la comparaison entre les humains. Deuxièmement, le niveau d'expertise influe sur notre façon de percevoir des stimuli. Nous montrons que le niveau d'expertise des algorithmes a également une influence sur leur façon de percevoir les stimuli qu'ils analysent. De plus, une corrélation est visible entre la similarité des comportements attentionnels humain et machine et leur niveau d'expertise. Plus un réseau obtient de bonnes performances sur la tâche de classification, plus son comportement attentionnel aura tendance à être proche de celui des experts.

Évaluation des méthodes d'attention artificielle

Avec la création de ce jeu de données oculométrique public, il sera possible d'évaluer les différentes méthodes rendant compte de l'attention artificielle. En effet, grâce à l'étude de la stabilité et grâce à la comparaison avec les comportements attentionnels humains, nous avons pu hiérarchiser les différentes méthodes. La méthode des gradients étant la plus stable et ses zones d'attention étant les plus proches de celles des experts, il nous semble qu'elle est supérieure en terme de rendu des zones d'intérêt d'une image donnée aux autres méthodes (Rétropropagation guidée, Guided GradCAM et GradCAM). Nous proposons aux chercheurs voulant évaluer leur modèle d'attention artificielle d'utiliser notre base de données pour évaluer à quel point les zones d'attention obtenues sont proches de celles des experts. De plus, l'étude de la stabilité nous semble être un bon indicateur de la qualité des zones d'intérêt. Enfin, si le modèle est robuste aux aléas de l'entraînement, cela garantira que ses zones d'attention sont représentatives de la perception de l'algorithme.

Travaux futurs

Élargissement de la base de données CrohnIPI

Bien que cette base de données permette de compenser le manque de données publiques et que les performances des réseaux de l'état de l'art soient bonnes, cette dernière ne compte que 3500 images, ce qui, malgré les techniques d'augmentation de données et la validation croisée, limite son utilisation pour le développement de solutions permettant d'assister les experts dans leur routine clinique. Afin d'augmenter la taille de la base de données tout en permettant d'assister les gastro entérologues dans leur tâche, une application web a été développée permettant de trier des images provenant de nouvelles vidéos capsules par les scores obtenus grâce au réseau ResNet34 entraîné sur les 3500 images de la base de données CrohnIPI. Cette application en ligne permettra à plusieurs experts d'annoter les images depuis différents sites.

Un travail de recherche pourrait être réalisé de façon à identifier les meilleures nouvelles images à annoter. En effet, le travail d'annotation est long et fastidieux, et les experts médicaux peu nombreux. Trouver des critères de sélection permettant de maximiser la généralisation de l'algorithme avec un minimum de cas d'entraînement serait un bon moyen de pallier les problématiques d'annotation. L'utilisation d'apprentissage actif semble tout à fait adaptée à cette situation [202]. On pourrait également envisager d'utiliser un réseau

pré-entrainé sur le jeu de données existant et chercher quels sont les exemples les plus éloignés de ceux déjà assimilés par l'algorithme. On calculerait la distance entre les cartes de caractéristiques de l'exemple à potentiellement annoter et la moyenne des cartes de caractéristiques pour les différentes classes. Les images obtenant les plus grandes distances avec l'une ou l'autre des classes seraient les exemples à annoter. En observant la répartition des erreurs en fonction des scores obtenus, on pourrait également trouver les scores dont la proportion d'erreur (nombre d'images incorrectement classées sur le nombre d'images classées) est la plus importante. Ainsi, en admettant que les images incorrectement classées sont les plus intéressantes pour le réseau (car les plus susceptibles de faire varier les poids lors d'un nouvel apprentissage), cette technique pourrait permettre d'améliorer la base de données en limitant le travail d'annotation.

Évolution du réseau à attention dure

Il serait intéressant d'observer les résultats sur des images trop grandes pour être utilisées sans être redimensionnées avant le passage dans le réseau de neurones profond. En effet, le nombre de poids augmentant exponentiellement avec la taille des images en entrée, les images médicales, comme les radiographies, sont souvent redimensionnées avant d'être utilisées en entrée d'un réseau de neurones convolutif. Les images en entrée du réseau récurrent n'ayant pas besoin d'être redimensionnées, du fait que la taille d'extraction du patch est fixe, cela permettrait de pallier le problème des larges images.

La comparaison entre les zones d'attention du réseau de neurones récurrent à attention dure nous montre que le comportement attentionnel de ce dernier est plus proche de celui d'un novice que de celui d'un expert. Le réseau de neurones récurrent étant composé d'un réseau de neurones à convolution, nous pourrions appliquer les méthodes d'extraction de l'attention post-hoc présentées dans la section 2.7.2 afin d'affiner les cartes d'attention. Les méthodes d'extraction de l'attention post-hoc sont basées sur l'étude de l'influence des pixels d'entrée sur la sortie, ce qui n'a pas été pris en compte dans la comparaison entre attention dure et attention humaine. L'utilisation des méthodes d'extraction de l'attention post-hoc pourrait être utilisé sur le réseau de regard afin d'affiner les explications visuelles fournies par le réseau à attention dure. Il serait également intéressant d'observer comment l'attention post-hoc du réseau de regard évolue au cours des différentes itérations.

Expérience subjective d'évaluation des méthodes d'attention artificielle

La comparaison entre l'attention humaine et l'attention artificielle nous montre que des similitudes sont observables entre les deux comportements attentionnels. Mais le fait que les méthodes produisent des cartes d'attention proches de celles des humains atteste-t-il qu'elles leurs soient utiles pour assister la prise de décision des experts ? Est-ce qu'au contraire, une carte plus éloignée de celle humaine ne leur fournirait pas des informations supplémentaires ?

Il faudrait réaliser une expérience subjective pour quantifier cela. Les sujets de différents niveaux d'expertise seraient confrontés aux zones d'attention de plusieurs méthodes, ainsi qu'à celle humaine. Les différents sujets devraient alors choisir entre les différentes propositions qui, selon eux, rendent le mieux compte des zones d'intérêt de l'image. Afin de pouvoir exprimer au mieux leur avis, les sujets établiraient un classement de la meilleure à la moins bonne. Au travers de cette expérience subjective, nous pourrions étudier l'influence du niveau d'expertise sur l'appréciation des explications. Cela nous permettrait également d'observer si les explications préférées par les sujets correspondent aux zones d'attention de leur groupe d'expertise.

Au travers de cette expérience, nous éviterons un éventuel biais de confirmation. Nous concluons que les méthodes les plus fiables sont celles dont le comportement est le plus proche de celui des experts, mais peut-être qu'au contraire, les meilleures méthodes sont celles qui montrent aux experts ce qu'ils auraient pu ne pas voir.

Apprentissage grâce à des informations privilégiées

Nous avons montré dans ce manuscrit que plus les réseaux de neurones profonds de l'état de l'art obtenaient de bonnes performances à la tâche de détection des lésions de la maladie de Crohn, plus leur comportement attentionnel se rapprochait de celui des experts. Ainsi, il serait intéressant de mettre en place une architecture utilisant les données issues de l'expérience oculométrique en entrée du réseau de neurones. En fixant un objectif visant à minimiser l'écart entre les cartes d'attention humaine et artificielle, il serait sans doute possible d'améliorer les performances du réseau ou d'accélérer la phase d'entraînement. Afin d'y parvenir, il faudrait appliquer le paradigme d'apprentissage avec des informations privilégiées [203]. Cette méthode pourrait s'appliquer aux réseaux utilisant un principe d'attention apprise. Il serait possible d'utiliser des métriques de

saillances pour réaliser cet objectif comme pour les modèles encodeur-décodeur cherchant à prédire la saillance [204, 205, 206].

Dans notre cas d'application, le nombre d'images composant la base de données oculométrique étant réduit, il faudrait réfléchir à un protocole d'entraînement utilisant également la base de données CrohnIPI composée d'un nombre d'images plus important. L'utilisation de l'information de saillance pourrait permettre une classification plus précise, en forçant le réseau à créer des filtres qui s'activeraient dans la région ciblée par les regards des experts. Forcer l'activation des filtres dans une certaine région pourrait ainsi entraîner une optimisation plus rapide et plus efficace.

En utilisant à la fois le jeu de données oculométriques en entraînement et en validation et le jeu de données CrohnIPI en test pour établir une base de référence, nous pourrions imaginer différents protocoles expérimentaux pour permettre l'apprentissage supervisé sous l'influence d'informations privilégiées. Il serait nécessaire d'envisager différentes stratégies d'alternances des phases d'apprentissage entre celles permettant l'optimisation de l'attention du réseau et celles, plus traditionnelles, permettant la généralisation sur un nombre de cas plus large.

L'application à des réseaux utilisant de l'attention apprise nous semble tout indiquée. À notre sens, il serait plus judicieux de considérer les réseaux à attention douce de par le fait qu'ils produisent des zones d'attention plus fines que celles à attention dure. Cependant, ajouter un terme à la fonction de récompense du réseau à attention dure qui serait proportionnel à la similarité entre le comportement attentionnel humain et artificiel, est également envisageable. Les méthodes d'apprentissage post-hoc, elles, semblent contre indiquées puisqu'elle permettent de rendre de compte de l'attention *a posteriori* de l'entraînement et qu'elles impliquent déjà des calculs de gradients pour la plupart.

L'explicabilité des réseaux de neurones profonds par les méthodes de saillance est-elle fiable ?

L'étude de l'explicabilité dans le domaine médical gagne de plus en plus d'intérêt [207, 87, 208]. Mais rendre les algorithmes explicables grâce aux zones d'intérêt des modèles est une question créant le débat. En effet, bien que ces méthodes d'explicabilité permettent de mieux comprendre le cheminement des algorithmes jusqu'à leur décision, il ne faut pas perdre de vue que ces derniers ne sont que des outils et qu'il n'est pas possible de substituer l'expertise humaine par l'expertise artificielle.

Dans l'article [209], les auteurs identifient les limites des approches d'explicabilité post-hoc. Ils se basent sur un exemple d'application à des radiographies de poumons dans lesquelles on cherche à identifier des marqueurs de la pneumonie grâce au réseau ChestNext [30, 210]. Afin de comprendre les décisions de ce réseau dont les performances semblent inférieures à celles des humains, les auteurs utilisent une méthode d'explicabilité par saillance post-hoc : Class Activation Map. Cette méthode permet de visualiser les activations des couches convolutives lors de la propagation de l'image à travers le réseau. Des quatre méthodes présentées dans ce manuscrit, on peut la rapprocher de la méthode GradCAM, également basée sur les cartes d'activation, à la différence que ces dernières ne sont pas pondérées par le gradient les traversant lors de la propagation de l'image.

Un premier reproche fait aux méthodes d'explicabilité post-hoc est que les zones d'activation sont trop larges, ne permettant pas d'identifier clairement quelle partie de la zone activée a réellement permis de prendre la décision. Cette remarque est tout à fait pertinente, et converge avec nos expérimentations. Les méthodes basées sur l'activation des filtres convolutifs produisent des cartes de faible résolution du fait que ces dernières sont extraites sur les dernières couches de convolution. En effet, ces dernières couches, dont le niveau d'abstraction/généralisation est le plus haut, résultent de nombreuses opérations de mutualisation (pooling). De par nos expériences, nous avons montré que la méthode GradCAM dont la résolution de la carte d'activation est proche de celle obtenue avec la méthode des activations par classes (Class Activation Map), obtient de moins bons résultats de stabilité et un comportement attentionnel plus éloigné de celui des experts que les méthodes basées sur les gradients (dont la résolution de la carte de saillance est similaire à la résolution de l'image en entrée du réseau).

Une seconde remarque est que les zones d'activation n'évoluent que très peu lorsque le réseau subit une attaque. On rappelle ici qu'une des limitations des réseaux de neurones profonds est qu'il sont susceptibles aux attaques de bruit. En ne modifiant que quelques pixels, on parvient à faire changer la décision du réseau de neurones, sans que cette perturbation ne soit visible pour un humain. Il nous semble que cette limitation s'adresse plus aux réseaux de neurones qu'aux techniques d'explicabilité.

Il est évident qu'à l'heure actuelle, ces méthodes restent plus utiles pour le dépannage et l'audit des systèmes que pour une réelle assistance aux experts médicaux. Comme le font Ghassemi et al. [209] nous invitons également ces derniers à faire preuve de prudence lorsqu'ils utilisent les méthodes d'explicabilité pour appuyer leurs décisions. De notre point de vue, il faut principalement faire attention aux biais de confirmation car ils peuvent

être une cause d'erreur quand le doute, lui, peut permettre d'appréhender les cas les plus complexes. Le travail entre les chercheurs en vision par ordinateur et les experts médicaux est un enjeu majeur des prochaines décennies et ce dernier devra faire appel à beaucoup de pédagogie des deux côtés. Pour les experts médicaux, il sera nécessaire de comprendre les limitations des outils prédictifs, et pour les chercheurs en informatique, il faudra bien comprendre l'écart entre les expérimentations de laboratoire et la réalité du terrain.

BIBLIOGRAPHIE

[1] Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. Nature. 2000 ;405(6785) :417. doi :10.1038/35013140.

[2] Vallée R, De Maissin A, Mouchère H, Boureille A, Coutrot A, Normand N. Réseau de neurones récurrent à attention pour la détection de lésions intestinales. In : ORASIS. Saint-Dié-des-Vosges, France ; 2019.Available from : https://hal.archives-ouvertes.fr/hal-02283753.

[3] Vallée R, Coutrot A, Normand N, Mouchère H. Accurate small bowel lesions detection in wireless capsule endoscopy images using deep recurrent attention neural network. In : IEEE 21st International Workshop on Multimedia Signal Processing (MMSP 2019). Proc. IEEE 21st International Workshop on Multimedia Signal Processing. Kuala Lumpur, Malaysia ; 2019.Available from : https://hal.archives-ouvertes.fr/hal-02296282.

[4] Vallée R, de Maissin A, Coutrot A, Mouchère H, Bourreille A, Normand N. CrohnIPI : An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms. In : SPIE Medical Imaging. vol. 11317 of Proc. SPIE, Medical Imaging 2020 : Biomedical Applications in Molecular, Structural, and Functional Imaging. Houston, France : SPIE ; 2020. p. 61. Available from : https://hal.archives-ouvertes.fr/hal-02518263.

[5] de Maissin A, Vallée R, Flamant M, Fondain-Bossiere M, Berre CL, Coutrot A, et al. Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network. Endoscopy International Open. 2021 ;09(07) :E1136 – E1144. doi :10.1055/a-1468-3964.

[6] UEG Week 2019 Oral Presentations. United European Gastroenterology Journal. 2019 ;7(S8) :10–188. doi :https ://doi.org/10.1177/2050640619854670.

[7] Vallée R, Coutrot A, Normand N, Mouchère H. Influence of expertise on human and machine visual attention in a medical image classification task. In : European Conference on Visual Perception. online, France ; 2021.Available from : https://hal.archives-ouvertes.fr/hal-03379161.

[8] Maitland ME. A transdisciplinary definition of diagnosis. Journal of allied health. 2010 ;39(4) :306–313.

[9] Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care ; 2015. Available from : https://www.nap.edu/catalog/21794/improving-diagnosis-in-health-care.

[10] Kasban H. A Comparative Study of Medical Imaging Techniques. International Journal of Information Science and Intelligent System,. 2015 ;4 :37–58.

[11] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Anal. 2017 ;42 :60–88.

[12] Lodwick GS, Haun CL, Smith WE, Keller RF, Robertson ED. Computer Diagnosis of Primary Bone Tumors. Radiology. 1963 ;80(2) :273–275. doi :10.1148/80.2.273.

[13] Meyers PH, Nice CM, Becker HC, Nettleton WJ, Sweeney JW, Meckstroth GR. Automated Computer Analysis of Radiographic Images. Radiology. 1964 ;83(6) :1029–1034. doi :10.1148/83.6.1029.

[14] Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W. Detection of Radiographic Abnormalities in Mammograms by Means of Optical Scanning and Computer Analysis. Radiology. 1967 ;89 :211–215.

[15] Kruger RP, Townes JR, Hall DL, Dwyer SJ, Lodwick GS. Automated Radiographic Diagnosis via Feature Extraction and Classification of Cardiac Size and Shape Descriptors. IEEE Transactions on Biomedical Engineering. 1972 ;BME-19(3) :174–186. doi :10.1109/TBME.1972.324115.

[16] Toriwaki JI, Suenaga Y, Negoro T, Fukumura T. Pattern recognition of chest X-ray images. Computer Graphics and Image Processing. 1973 ;2(3) :252 – 271. doi :https ://doi.org/10.1016/0146-664X(73)90005-1.

[17] Jagoe JR, Paton KA. Reading Chest Radiographs for Pneumoconiosis by Computer. British Journal of Industrial Medicine. 1975 ;32(4) :267–272.

[18] Doi K. Computer-Aided Diagnosis in Medical Imaging : Historical Review, Current Status and Future Potential. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2007 ;31 :198–211. doi :10.1016/j.compmedimag.2007.02.002.

[19] Nishikawa RM, Giger ML, Doi K, Vyborny CJ, Schmidt RA. Computer-aided detection of clustered microcalcifications on digital mammograms. Medical and Biological Engineering and Computing. 1995 ;33(2) :174–178. doi :10.1007/BF02523037.

[20] Smyth PP, Taylor CJ, Adams JE. Vertebral Shape : Automatic Measurement with Active Shape Models. Radiology. 1999 ;211(2) :571–578. doi :10.1148/radiology.211.2.r99ma40571.

[21] Kikinis R, Shenton ME, Iosifescu DV, McCarley RW, Saiviroonporn P, Hokama HH, et al. A digital brain atlas for surgical planning, model-driven segmentation, and teaching. IEEE Transactions on Visualization and Computer Graphics. 1996 ;2(3) :232–241. doi :10.1109/2945.537306.

[22] Warfield S. Fast k-NN classification for multichannel image data. Pattern Recognition Letters. 1996 ;17(7) :713 – 721. doi :https ://doi.org/10.1016/0167-8655(96)00036-0.

[23] Lo SCB, Lou SLA, Jyh-Shyan Lin, Freedman MT, Chien MV, Mun SK. Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Transactions on Medical Imaging. 1995 ;14(4) :711–718. doi :10.1109/42.476112.

[24] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998 ;86(11) :2278–2324. doi :10.1109/5.726791.

[25] Badia AP, Piot B, Kapturowski S, Sprechmann P, Vitvitskyi A, Guo ZD, et al. Agent57 : Outperforming the Atari Human Benchmark. In : III HD, Singh A, editors. Proceedings of the 37th International Conference on Machine Learning. vol. 119 of Proceedings of Machine Learning Research. PMLR ; 2020. p. 507–517. Available from : `http://proceedings.mlr.press/v119/badia20a.html`.

[26] Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science. 2018 ;362(6419) :1140–1144. doi :10.1126/science.aar6404.

[27] Buetti-Dinh A, Galli V, Bellenberg S, Ilie O, Herold M, Christel S, et al. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. Biotechnology Reports. 2019 ;22 :e00321. doi :https ://doi.org/10.1016/j.btre.2019.e00321.

[28] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 ;542(7639) :115–118. doi :10.1038/nature21056.

[29] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer. 2019 ;113 :47–54. doi :https ://doi.org/10.1016/j.ejca.2019.04.001.

[30] Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine. 2018 ;1(1) :39. doi :10.1038/s41746-018-0040-6.

[31] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis : A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine. 2018 ;15(11) :1–17. doi :10.1371/journal.pmed.1002686.

[32] Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Medical Physics. 2019 ;46(1) :e1–e36. doi :https ://doi.org/10.1002/mp.13264.

[33] de la Santé OM. Maladies chroniques ; 2020. Available from : `https://www.who.int/topics/chronic_diseases/fr/`.

[34] Gal E, Geller A, Fraser G, Levi Z, Niv Y. Assessment and Validation of the New Capsule Endoscopy Crohn's Disease Activity Index (CECDAI). Digestive Diseases and Sciences. 2008 ;53(7) :1933–1937.

[35] Gralnek IM, Defranchis R, Seidman E, Leighton JA, Legnani P, Lewis BS. Development of a capsule endoscopy scoring index for small bowel mucosal inflammatory change. Alimentary Pharmacology & Therapeutics. 2007 ;27(2) :146–154.

[36] Secretariat MA. Wireless capsule endoscopy : an evidence-based analysis. Ontario health technology assessment series. 2003 ;3(2) :1–35.

[37] Mylonaki M, Fritscher-Ravens A, Swain P. Wireless capsule endoscopy : a comparison with push enteroscopy in patients with gastroscopy and colonoscopy negative gastrointestinal bleeding. Gut. 2003 ;52(8) :1122–1126. doi :10.1136/gut.52.8.1122.

[38] Diagnostic, Cardiology I. Capsule Endoscopy Systems Safety in Patients With Cardiovascular Implants ; 2017. Avai-

lable from : https://www.dicardiology.com/article/capsule-endoscopy-systems-safety-patients-cardiovascular-implants.

[39] VOnews. Un nouveau moyen d'investigation endoscopique pour l'hôpital d'Argenteuil ; 2012. Available from : https://95.telif.tv/2012/10/16/un-nouveau-moyen-dinvestigation-endoscopique-pour-lhopital-dargenteuil/.

[40] Comparison of magnetic resonance imaging and video capsule enteroscopy in diagnosing small-bowel pathology : localization-dependent diagnostic yield. Scandinavian journal of gastroenterology. 2010 ;45(4) :490–500. doi :10.3109/00365520903567817.

[41] Diagnostic accuracy of capsule endoscopy for small bowel Crohn's disease is superior to that of MR enterography or CT enterography. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2011 ;9(2) :124–129. doi :10.1016/j.cgh.2010.10.019.

[42] Comparison of Capsule Endoscopy and Magnetic Resonance Enterography for the Assessment of Small Bowel Lesions in Crohn's Disease. Inflammatory bowel diseases. 2018 ;24(4) :775–780. doi :10.1093/ibd/izx107.

[43] Dionisio PM, Gurudu SR, Leighton JA, Leontiadis GI, Fleischer DE, Hara AK, et al. Capsule endoscopy has a significantly higher diagnostic yield in patients with suspected and established small-bowel Crohn's disease : a meta-analysis. The American journal of gastroenterology. 2010 ;105(6) :1240–8 ; quiz 1249. doi :10.1038/ajg.2009.713.

[44] Buisson A, Gonzalez F, Poullenot F, Nancey S, Sollellis E, Fumery M, et al. Comparative Acceptability and Perceived Clinical Utility of Monitoring Tools : A Nationwide Survey of Patients with Inflammatory Bowel Disease. Inflammatory Bowel Diseases. 2017 ;23(8) :1425–1433. doi :10.1097/MIB.0000000000001140.

[45] McAlindon ME, Ching HL, Yung D, Sidhu R, Koulaouzidis A. Capsule endoscopy of the small bowel. Annals of Translational Medicine. 2016 ;4(19).

[46] Bejakovic S, Kumar R, Dassopoulos T, Mullin G, Hager G. Analysis of Crohn's disease lesions in capsule endoscopy images. 2009 ; p. 2793–2798.

[47] Karargyris A, Bourbakis N. Identification of ulcers in Wireless Capsule Endoscopy videos. 2009 ; p. 554–557. doi :10.1109/ISBI.2009.5193107.

[48] Girgis HZ, Mitchell BR, Dassopoulos T, Mullin G, Hager G. An intelligent system to detect Crohn's disease inflammation in Wireless Capsule Endoscopy videos. 2010 ; p. 1373–1376. doi :10.1109/ISBI.2010.5490253.

[49] Hajo-Maghsoudi O, Talebpour A, Soltanian-Zadeh H, Haji-Maghsoodi N. Segmentation of Crohn, Lymphangiectasia, Xanthoma, Lymphoid Hyperplasia and Stenosis diseases in WCE. Studies in health technology and informatics. 2012 ;180 :143–7.

[50] Chen Y, Lee J. Ulcer detection in wireless capsule endoscopy video. MM 2012 - Proceedings of the 20th ACM International Conference on Multimedia. 2012 ; p. 1181–1184. doi :10.1145/2393347.2396413.

[51] Charisis VS, Hadjileontiadis LJ, Liatsos CN, Mavrogiannis CC, Sergiadis GD. Capsule endoscopy image analysis using texture information from various colour models. Computer Methods and Programs in Biomedicine. 2012 ;107(1) :61–74. doi :https ://doi.org/10.1016/j.cmpb.2011.10.004.

[52] Jebarani WSL, Daisy VJ. Assessment of Crohn's disease lesions in Wireless Capsule Endoscopy images using SVM based classification. 2013 ; p. 303–307. doi :10.1109/ICSIPR.2013.6497945.

[53] Eid A, Charisis VS, Hadjileontiadis LJ, Sergiadis GD. A curvelet-based lacunarity approach for ulcer detection from Wireless Capsule Endoscopy images. 2013 ; p. 273–278. doi :10.1109/CBMS.2013.6627801.

[54] Szczypiński P, Klepaczko A, Pazurek M, Daniel P. Texture and color based image segmentation and pathology detection in capsule endoscopy videos. Computer Methods and Programs in Biomedicine. 2014 ;113(1) :396–411. doi :https ://doi.org/10.1016/j.cmpb.2012.09.004.

[55] Yeh JY, Wu TH, Tsai WJ. Bleeding and Ulcer Detection Using Wireless Capsule Endoscopy Images. Journal of Software Engineering and Applications. 2014 ;07 :422–432. doi :10.4236/jsea.2014.75039.

[56] Iakovidis D, Koulaouzidis A. Automatic lesion detection in capsule endoscopy based on color saliency : Closer to an essential adjunct for reviewing software. Gastrointestinal Endoscopy. 2014 ;doi :10.1016/j.gie.2014.06.026.

[57] Yuan Y, Wang J, Li B, Meng MQ. Saliency Based Ulcer Detection for Wireless Capsule Endoscopy Diagnosis. IEEE Transactions on Medical Imaging. 2015 ;34(10) :2046–2057. doi :10.1109/TMI.2015.2418534.

[58] Charisis VS, Hadjileontiadis LJ. Use of adaptive hybrid filtering process in Crohn's disease lesion detection from real capsule endoscopy videos. vol. 3 ; 2016. p. 27–33.

[59] Liaqat A, Khan MA, Shah JH, Sharif M, Yasmin M, Fernandes SL. Automated ulcer and bleeding classification from wce images using multiple features fusion and

selection. Journal of Mechanics in Medicine and Biology. 2018 ;18(04) :1850038. doi :10.1142/S0219519418500380.

[60] Souaidi M, Ansari ME. Multi-scale analysis of ulcer disease detection from WCE images. IET Image Processing. 2019 ;13(12) :2233–2244. doi :10.1049/iet-ipr.2019.0415.

[61] Georgakopoulos SV, Iakovidis DK, Vasilakakis M, Plagianakos VP, Koulaouzidis A. Weakly-supervised Convolutional learning for detection of inflammatory gastrointestinal lesions. 2016 ; p. 510–514.

[62] Fan S, Xu L, Fan Y, Wei K, Li L. Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. Physics in Medicine and Biology. 2018 ;63. doi :10.1088/1361-6560/aad51c.

[63] Aoki T, Yamada A, Aoyama K, Saito H, Tsuboi A, Nakada A, et al. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointestinal Endoscopy. 2019 ;89(2) :357 – 363.e2. doi :https ://doi.org/10.1016/j.gie.2018.10.027.

[64] Haya A, Hussain A, Al-Aseem N, Liatsis P, Al-Jumeily D. Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images. Sensors. 2019 ;19 :1265. doi :10.3390/s19061265.

[65] Barash Y, Azaria L, Soffer S, Margalit Yehuda R, Shlomi O, Ben-Horin S, et al. Ulcer severity grading in video capsule images of patients with Crohn's disease : an ordinal neural network solution. Gastrointestinal Endoscopy. 2021 ;93(1) :187–192. doi :https ://doi.org/10.1016/j.gie.2020.05.066.

[66] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing Medical Imaging Data for Machine Learning. Radiology. 2020 ;295(1) :4–15. doi :10.1148/radiol.2020192224.

[67] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) : architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 ;17(5) :507–513. doi :10.1136/jamia.2009.001560.

[68] Lee C, Kim Y, Kim YS, Jang J. Automatic Disease Annotation From Radiology Reports Using Artificial Intelligence Implemented by a Recurrent Neural Network. American Journal of Roentgenology. 2019 ;212(4) :734–740. doi :10.2214/AJR.18.19869.

[69] Sugimoto K, Takeda T, Oh JH, Wada S, Konishi S, Yamahata A, et al. Extracting clinical terms from radiology reports with deep learning. Journal of Biomedical Informatics. 2021 ;116 :103729. doi :https ://doi.org/10.1016/j.jbi.2021.103729.

[70] Genta R, Sonnenberg A. Big data in gastroenterology research. Nature reviews Gastroenterology & hepatology. 2014 ;11. doi :10.1038/nrgastro.2014.18.

[71] Leenhardt R, Buisson A, Bourreille A, Marteau P, Koulaouzidis A, Li C, et al. Nomenclature and semantic descriptions of ulcerative and inflammatory lesions seen in Crohn's disease in small bowel capsule endoscopy : An international Delphi consensus statement. United European gastroenterology journal. 2020 ;8(1) :99–107. doi :10.1177/2050640619895864.

[72] Fleiss JL, et al. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971 ;76(5) :378–382.

[73] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Computing Research Repository. 2015 ;abs/1512.03385.

[74] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computing Research Repository. 2014 ;abs/1409.1556.

[75] Li J, Monroe W, Jurafsky D. Understanding Neural Networks through Representation Erasure. CoRR. 2016 ;abs/1612.08220.

[76] Buchanan B, Shortliffe E. Rule-based Expert System – The MYCIN Experiments of the Stanford Heuristic Programming Project ; 1984.

[77] Wick MR, Thompson WB. Reconstructive expert system explanation. Artificial Intelligence. 1992 ;54(1) :33–70. doi :https ://doi.org/10.1016/0004-3702(92)90087-E.

[78] Lipton ZC. The Mythos of Model Interpretability. CoRR. 2016 ;abs/1606.03490.

[79] Confalonieri R, Coba L, Wagner B, Besold TR. A historical perspective of explainable Artificial Intelligence. WIREs Data Mining and Knowledge Discovery. 2021 ;11(1) :e1391. doi :https ://doi.org/10.1002/widm.1391.

[80] Miller T. Explanation in Artificial Intelligence : Insights from the Social Sciences. CoRR. 2017 ;abs/1706.07269.

[81] Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for Explainable AI : Challenges and Prospects. CoRR. 2018 ;abs/1812.04608.

[82] Dave D, Naik H, Singhal S, Patel P. Explainable AI meets Healthcare : A Study on Heart Disease Dataset ; 2020.

[83] Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA. 2017 ;318(6) :5120177–518. doi :10.1001/jama.2017.7797.

[84] Angwin J, Larson J, Kirchner L, Mattu S. Machine Bias ; 2013. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[85] Codevilla F, Santana E, Lopez AM, Gaidon A. Exploring the Limitations of Behavior Cloning for Autonomous Driving. In : Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) ; 2019.

[86] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In : International Conference on Learning Representations ; 2014.Available from : http://arxiv.org/abs/1312.6199.

[87] Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, Tengg-Kobligk Hv, et al. On the Interpretability of Artificial Intelligence in Radiology : Challenges and Opportunities. Radiology : Artificial Intelligence. 2020 ;2(3) :e190043. doi :10.1148/ryai.2020190043.

[88] Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". AI magazine. 2017 ;38(3) :50–57.

[89] Mitchell D, Selman B, Levesque H. Proceedings of the tenth national conference on artificial intelligence. 1992 ;.

[90] Raichle ME. Two views of brain function. Trends in cognitive sciences. 2010 ;14(4) :180–190. doi :10.1016/j.tics.2010.01.008.

[91] William J. The principles of psychology, Vol I. ; 1890.

[92] Tootell RBH, Hadjikhani N, Hall EK, Marrett S, Vanduffel W, Vaughan JT, et al. The Retinotopy of Visual Spatial Attention. Neuron. 1998 ;21(6) :1409–1422. doi :https ://doi.org/10.1016/S0896-6273(00)80659-5.

[93] Woldorff MG, Gallen CC, Hampson SA, Hillyard SA, Pantev C, Sobel D, et al. Modulation of early sensory processing in human auditory cortex during auditory selective attention. Proceedings of the National Academy of Sciences of the United States of America. 1993 ;90(18) :8722–8726. doi :10.1073/pnas.90.18.8722.

[94] Johansen-Berg H, Lloyd DM. The physiology and psychology of selective attention to touch. Frontiers in bioscience : a journal and virtual library. 2000 ;5 :D894–904.

[95] Zelano C, Bensafi M, Porter J, Mainland J, Johnson B, Bremner E, et al. Attentional modulation in human primary olfactory cortex. Nature neuroscience. 2005 ;8(1) :114–120. doi :10.1038/nn1368.

[96] Veldhuizen MG, Bender G, Constable RT, Small DM. Trying to detect taste in a tasteless solution : modulation of early gustatory cortex by attention to taste. Chemical senses. 2007 ;32(6) :569–581. doi :10.1093/chemse/bjm025.

[97] Wandell BA, Dumoulin SO, Brewer AA. Visual Field Maps in Human Cortex. Neuron. 2007 ;56(2) :366 – 383. doi :https ://doi.org/10.1016/j.neuron.2007.10.012.

[98] Hoffman JE. Visual attention and eye movements. In : Attention. Hove, England : Psychology Press/Erlbaum (UK) Taylor & Francis ; 1998. p. 119–153.

[99] Bronkhorst AW. The cocktail-party problem revisited : early processing and selection of multi-talker speech. Atten Percept Psychophys. 2015 ;77(5) :1465–1487.

[100] Rizzolatti G, Riggio L, Dascola I, Umilt ? C. Reorienting attention across the horizontal and vertical meridians : evidence in favor of a premotor theory of attention. Neuropsychologia. 1987 ;25(1A) :31–40.

[101] Dodge R. FIVE TYPES OF EYE MOVEMENT IN THE HORIZONTAL MERIDIAN PLANE OF THE FIELD OF REGARD. American Journal of Physiology-Legacy Content. 1903 ;8(4) :307–329. doi :10.1152/ajplegacy.1903.8.4.307.

[102] Nuthmann A, Smith TJ, Engbert R, Henderson JM. CRISP : a computational model of fixation durations in scene viewing. Psychological review. 2010 ;117(2) :382–405. doi :10.1037/a0018924.

[103] Henderson JM, Weeks Jr PA, Hollingworth A. The effects of semantic consistency on eye movements during complex scene viewing. Journal of Experimental Psychology : Human Perception and Performance. 1999 ;25(1) :210–228. doi :10.1037/0096-1523.25.1.210.

[104] Smith TJ, Mital PK. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. Journal of Vision. 2013 ;13(8) :16–16. doi :10.1167/13.8.16.

[105] Tatler BW, Baddeley RJ, Gilchrist ID. Visual correlates of fixation selection : effects of scale and time. Vision Research. 2005 ;45(5) :643–659. doi :https ://doi.org/10.1016/j.visres.2004.09.017.

[106] Kümmerer M, Bethge M. State-of-the-Art in Human Scanpath Prediction. arXiv. 2021 ; p. 1–19.

[107] Borji A. Saliency Prediction in the Deep Learning Era : Successes and Limitations. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021 ;43(2) :679–700.

[108] Borji A, Itti L. State-of-the-art in Visual Attention Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013 ;35(1) :185–207.

[109] de Haas B, Iakovidis AL, Schwarzkopf DS, Gegenfurtner KR. Individual differences in visual salience vary along semantic dimensions. Proceedings of the National Academy of Sciences. 2019 ;116(24) :11687–11692.

[110] Pinto Y, van der Leij AR, Sligte IG, Lamme VAF, Scholte HS. Bottom-up and top-down attention are independent. Journal of Vision. 2013 ;13(3) :16–16. doi :10.1167/13.3.16.

[111] Hein E, Rolke B, Ulrich R. Visual attention and temporal discrimination : Differential effects of automatic and voluntary cueing. Visual Cognition. 2006 ;13(1) :29–50. doi :10.1080/13506280500143524.

[112] Ling S, Carrasco M. Sustained and transient covert attention enhance the signal via different contrast response functions. Vision Research. 2006 ;46(8) :1210–1220. doi :https ://doi.org/10.1016/j.visres.2005.05.008.

[113] Tatler B, Wade N, Kwan H, Findlay J, Velichkovsky B. Yarbus, Eye Movements, and Vision. i-Perception. 2010 ;1 :7–27. doi :10.1068/i0382.

[114] Tanner J, Itti L. A top-down saliency model with goal relevance. Journal of Vision. 2019 ;19(1) :1–16.

[115] Schutt HH, Rothkegel LOM, Trukenbrod HA, Engbert R, Wichmann FA. Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. Journal of Vision. 2019 ;19(3) :1–23.

[116] Kok EM, Aizenman AM, Võ MLH, Wolfe JM. Even if I showed you where you looked, remembering where you just looked is hard. Journal of vision. 2017 ;17(12) :2. doi :10.1167/17.12.2.

[117] Võ MLH, Aizenman AM, Wolfe JM. You think you know where you looked ? You better look again. Journal of experimental psychology Human perception and performance. 2016 ;42(10) :1477–1481. doi :10.1037/xhp0000264.

[118] Płużyczka M. The First Hundred Years : a History of Eye Tracking as a Research Method. Applied Linguistics Papers. 2018 ;4/2018 :101–116. doi :10.32612/uw.25449354.2018.4.pp.101-116.

[119] Huey EB. The Psychology and Pedagogy of Reading. Oxford, England : Macmillan ; 1908.

[120] Buswell GT. How people look at pictures : a study of the psychology and perception in art. Oxford, England : Univ. Chicago Press ; 1935.

[121] Goldberg J, Wichansky A. Eye tracking in usability evaluation : A practitioner's guide ; 2003.

[122] Ball L. In : Eye-tracking and reasoning : What your eyes tell about your inferences ; 2013.

[123] Jacob R, Karn K. In : Eye Tracking in Human-Computer Interaction and Usability Research : Ready to Deliver the Promises. vol. 2 ; 2003. p. 573–605.

[124] Le Meur O, Baccino T. Methods for comparing scanpaths and saliency maps : strengths and weaknesses. Behavior Research Methods. 2013 ;45(1) :251–266. doi :10.3758/s13428-012-0226-9.

[125] Privitera CM. The scanpath theory : its definition and later developments. In : Electronic Imaging ; 2006.

[126] Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady. 1966 ;10(8) :707–710.

[127] Mannan S, Ruddock KH, Wooding DS. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. Spatial vision. 1995 ;9(3) :363–386. doi :10.1163/156856895x00052.

[128] Jarodzka H, Holmqvist K, Nyström M. A vector-based, multidimensional scanpath similarity measure ; 2010. p. 211–218.

[129] Velichkovsky B, Pomplun M, Rieser J. Attention and communication : Eye-movement-based research paradigms. In : Zangemeister WH, Stiehl HS, Freksa C, editors. Visual Attention and Cognition. vol. 116 of Advances in Psychology. North-Holland ; 1996. p. 125–154. Available from : `https://www.sciencedirect.com/science/article/pii/S0166411596800744`.

[130] Bruce NDB, Tsotsos JK. Saliency Based on Information Maximization. In : Proceedings of the 18th International Conference on Neural Information Processing Systems. NIPS'05. Cambridge, MA, USA : MIT Press ; 2005. p. 155–162.

[131] Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F. What do different evaluation metrics tell us about saliency models ? CoRR. 2016 ;abs/1604.03605.

[132] Peters RJ, Iyer A, Itti L, Koch C. Components of bottom-up gaze allocation in natural images. Vision Research. 2005 ;45 :2397–2416.

[133] Lee JS, Ebrahimi T. Perceptual video compression : A survey. IEEE Journal of selected topics in signal processing. 2012 ;6(6) :684–697.

[134] Le Callet P, Niebur E. Visual Attention and Applications in Multimedia Technologies. Proceedings of the IEEE Institute of Electrical and Electronics Engineers. 2013 ;101(9) :2058–2067. doi :10.1109/JPROC.2013.2265801.

[135] Miller EK, Buschman TJ. Cortical circuits for the control of attention. Current Opinion in Neurobiology. 2013 ;23(2) :216–222. doi :https ://doi.org/10.1016/j.conb.2012.11.011.

[136] Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience. 2002 ;3(3) :201–215. doi :10.1038/nrn755.

[137] Baluch F, Itti L. Mechanisms of top-down attention. Trends in Neurosciences. 2011 ;34(4) :210–224. doi :https ://doi.org/10.1016/j.tins.2011.02.003.

[138] Petersen SE, Posner MI. The attention system of the human brain : 20 years after. Annual review of neuroscience. 2012 ;35 :73–89. doi :10.1146/annurev-neuro-062111-150525.

[139] Posner MI, Petersen SE. The Attention System of the Human Brain. Annual Review of Neuroscience. 1990 ;13(1) :25–42. doi :10.1146/annurev.ne.13.030190.000325.

[140] Itti L. New Eye-Tracking Techniques May Revolutionize Mental Health Screening. Neuron. 2015 ;88(3) :442–444. doi :10.1016/j.neuron.2015.10.033.

[141] Wang S, Jiang M, Duchesne X, Laugeson E, Kennedy D, Adolphs R, et al. Atypical Visual Saliency in Autism Spectrum Disorder Quantified through Model-Based Eye Tracking. Neuron. 2015 ;88(3) :604–616. doi :https ://doi.org/10.1016/j.neuron.2015.09.042.

[142] Holzman PS, Proctor LR, Levy DL, Yasillo NJ, Meltzer HY, Hurt SW. Eye-Tracking Dysfunctions in Schizophrenic Patients and Their Relatives. Archives of General Psychiatry. 1974 ;31(2) :143–151. doi :10.1001/archpsyc.1974.01760140005001.

[143] Wilcockson TDW, Mardanbegi D, Xia B, Taylor S, Sawyer P, Gellersen HW, et al. Abnormalities of saccadic eye movements in dementia due to Alzheimer's disease and mild cognitive impairment. Aging. 2019 ;11(15) :5389–5398. doi :10.18632/aging.102118.

[144] Woo M. Eyes hint at hidden mental-health conditions. ; 2019.

[145] Hoff SR, Abrahamsen AL, Samset JH, Vigeland E, Klepp O, Hofvind S. Breast cancer : missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography– results from a retrospective review. Radiology. 2012 ;264(2) :378–386. doi :10.1148/radiol.12112074.

[146] Boyer B, Hauret L, Bellaiche R, Gräf C, Bourcier B, Fichet G. [Retrospectively detectable carcinomas : review of the literature]. Journal de radiologie. 2004 ;85(12 Pt 2) :2071–2078. doi :10.1016/s0221-0363(04)97784-0.

[147] Le MT, Mothersill CE, Seymour CB, McNeill FE. Is the false-positive rate in mammography in North America too high ? The British journal of radiology. 2016 ;89(1065) :20160045. doi :10.1259/bjr.20160045.

[148] Wu CC, Wolfe JM. Eye Movements in Medical Image Perception : A Selective Review of Past, Present and Future. Vision (Basel, Switzerland). 2019 ;3(2). doi :10.3390/vision3020032.

[149] Drew T, Vo MLH, Olwal A, Jacobson F, Seltzer SE, Wolfe JM. Scanners and drillers : characterizing expert visual search through volumetric images. Journal of vision. 2013 ;13(10). doi :10.1167/13.10.3.

[150] Carmody DP, Nodine CF, Kundel HL. An analysis of perceptual and cognitive factors in radiographic interpretation. Perception. 1980 ;9(3) :339–344. doi :10.1068/p090339.

[151] Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Investigative radiology. 1978 ;13(3) :175–181. doi :10.1097/00004424-197805000-00001.

[152] Hu CH, Kundel HL, Nodine CF, Krupinski EA, Toto LC. Searching for bone fractures : a comparison with pulmonary nodule search. Academic radiology. 1994 ;1(1) :25–32. doi :10.1016/s1076-6332(05)80780-9.

[153] Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Academic radiology. 1996 ;3(2) :137–144. doi :10.1016/s1076-6332(05)80381-2.

[154] Kundel HL, Nodine CF, Krupinski EA. Searching for lung nodules. Visual dwell indicates locations of false-positive and false-negative decisions. Investigative radiology. 1989 ;24(6) :472–478.

[155] Simons DJ, Chabris CF. Gorillas in our midst : sustained inattentional blindness for dynamic events. Perception. 1999 ;28(9) :1059–1074. doi :10.1068/p281059.

[156] Drew T, Võ MLH, Wolfe JM. The invisible gorilla strikes again : sustained inattentional blindness in expert observers. Psychological science. 2013 ;24(9) :1848–1853. doi :10.1177/0956797613479386.

[157] Gegenfurtner A, Lehtinen E, Jarodzka H, Säljö R. Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis. Computers Education. 2017 ;113 :212–225. doi :https ://doi.org/10.1016/j.compedu.2017.06.001.

[158] Richstone L, Schwartz MJ, Seideman C, Cadeddu J, Marshall S, Kavoussi LR. Eye metrics as an objective assessment of surgical skill. Annals of surgery. 2010 ;252(1) :177–182. doi :10.1097/SLA.0b013e3181e464fb.

[159] Winter M, Pryss R, Probst T, Reichert M. Applying Eye Movement Modeling Examples to Guide Novices' Attention in the Comprehension of Process Models. Brain sciences. 2021 ;11(1). doi :10.3390/brainsci11010072.

[160] Nodine CF, Kundel HL. THE COGNITIVE SIDE OF VISUAL SEARCH IN RADIOLOGY. In : O'REGAN JK, LEVY-SCHOEN A, editors. Eye Movements from Physiology to Cognition. Amsterdam : Elsevier ; 1987. p. 573–582. Available from : https://www.sciencedirect.com/science/article/pii/B9780444701138500813.

[161] Crowley RS, Naus GJ, Stewart Jr, Friedman CP. Development of visual diagnostic expertise in pathology – an information-processing study. Journal of the American Medical Informatics Association : JAMIA. 2003 ;10(1) :39–51. doi :10.1197/jamia.m1123.

[162] Krupinski EA, Graham AR, Weinstein RS. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. Human pathology. 2013 ;44(3) :357–364. doi :10.1016/j.humpath.2012.05.024.

[163] Bukach CM, Gauthier I, Tarr MJ. Beyond faces and modularity : the power of an expertise framework. Trends in cognitive sciences. 2006 ;10(4) :159–166. doi :10.1016/j.tics.2006.02.004.

[164] Brunyé TT, Nallamothu BK, Elmore JG. Eye-tracking for assessing medical image interpretation : A pilot feasibility study comparing novice vs expert cardiologists. Perspectives on medical education. 2019 ;8(2) :65–73. doi :10.1007/s40037-019-0505-6.

[165] Reingold EM, Charness N, Pomplun M, Stampe DM. Visual span in expert chess players : evidence from eye movements. Psychological science. 2001 ;12(1) :48–55. doi :10.1111/1467-9280.00309.

[166] de Santana Correia A, Colombini EL. Attention, please ! A survey of Neural Attention Models in Deep Learning. CoRR. 2021 ;abs/2103.16775.

[167] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv :14090473. 2014 ;.

[168] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. CoRR. 2017 ;abs/1706.03762.

[169] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, Attend and Tell : Neural Image Caption Generation with Visual Attention. In : Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. vol. 37 of Proceedings of Machine Learning Research. Lille, France : PMLR ; 2015. p. 2048–2057. Available from : http://proceedings.mlr.press/v37/xuc15.html.

[170] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent Models of Visual Attention. CoRR. 2014 ;abs/1406.6247.

[171] Ba J, Mnih V, Kavukcuoglu K. Multiple Object Recognition with Visual Attention. CoRR. 2015 ;abs/1412.7755.

[172] Chen M, Radford A, Child R, Wu J, Jun H, Luan D, et al. Generative Pretraining From Pixels. In : III HD, Singh A, editors. Proceedings of the 37th International Conference on Machine Learning. vol. 119 of Proceedings of Machine Learning Research. PMLR ; 2020. p. 1691–1703. Available from : http://proceedings.mlr.press/v119/chen20s.html.

[173] Cho K, van Merriënboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation : Encoder–Decoder Approaches. In : Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar : Association for Computational Linguistics ; 2014. p. 103–111. Available from : https://aclanthology.org/W14-4012.

[174] Yang Z, He X, Gao J, Deng L, Smola AJ. Stacked Attention Networks for Image Question Answering. CoRR. 2015 ;abs/1511.02274.

[175] Seo PH, Lin ZL, Cohen S, Shen X, Han B. Hierarchical Attention Networks. ArXiv. 2016 ;abs/1606.02393.

[176] Jetley S, Lord NA, Lee N, Torr PHS. Learn To Pay Attention. CoRR. 2018 ;abs/1804.02391.

[177] Larochelle H, Hinton GE. Learning to combine foveal glimpses with a third-order Boltzmann machine. In : Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A, editors. Advances in Neural Information Processing Systems. vol. 23. Curran Associates, Inc. ; 2010.Available from : https://proceedings.neurips.cc/paper/2010/file/677e09724f0e2df9b6c000b75b5da10d-Paper.pdf.

[178] Ba J, V, Kavukcuoglu K. Multiple Object Recognition with Visual Attention. 2014 ;.

[179] Gregor K, Danihelka I, Graves A, Wierstra D. DRAW : A Recurrent Neural Network For Image Generation. CoRR. 2015 ;abs/1502.04623.

[180] Bengio Y, Courville AC, Vincent P. Unsupervised Feature Learning and Deep Learning : A Review and New Perspectives. CoRR. 2012 ;abs/1206.5538.

[181] Mahendran A, Vedaldi A. Visualizing Deep Convolutional Neural Networks Using Natural Pre-images. International Journal of Computer Vision. 2016 ;120 :233–255.

[182] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps. CoRR. 2014 ;abs/1312.6034.

[183] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity : The All Convolutional Net. 2015 ;.

[184] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM : Why did you say that ? Visual Explanations from Deep Networks via Gradient-based Localization. CoRR. 2016 ;abs/1610.02391.

[185] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning. 1992 ;8(3) :229–256.

[186] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet : A Large-Scale Hierarchical Image Database. In : CVPR09 ; 2009.

[187] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 ;9(8) :1735–1780.

[188] Elsayed GF, Kornblith S, Le QV. Saccader : Improving Accuracy of Hard Attention Models for Vision. CoRR. 2019 ;abs/1908.07644.

[189] Das A, Agrawal H, Zitnick CL, Parikh D, Batra D. Human Attention in Visual Question Answering : Do Humans and Deep Networks Look at the Same Regions ? CoRR. 2016 ;abs/1606.03556.

[190] Tavakoli HR, Ahmed F, Borji A, Laaksonen J. Saliency Revisited : Analysis of Mouse Movements versus Fixations. CoRR. 2017 ;abs/1705.10546.

[191] Lai Q, Wang W, Khan SH, Shen J, Sun H, Shao L. Human vs Machine Attention in Neural Networks : A Comparative Study. CoRR. 2019 ;abs/1906.08764.

[192] Serre T. Deep Learning : The Good, the Bad, and the Ugly. Annual review of vision science. 2019 ;5 :399–426. doi :10.1146/annurev-vision-091718-014951.

[193] Jacob G, Rt P, Katti H, Arun S. Qualitative similarities and differences in visual object representations between brains and deep networks. Nature Communications. 2021 ;12. doi :10.1038/s41467-021-22078-3.

[194] Li Y, Liu M, Rehg JM. In the Eye of the Beholder : Gaze and Actions in First Person Video. 2020 ;.

[195] Zhang R, Guo S, Liu B, Zhu Y, Hayhoe M, Ballard D, et al. Machine versus Human Attention in Deep Reinforcement Learning Tasks. 2021 ;.

[196] Leong YC, Radulescu A, Daniel R, DeWoskin V, Niv Y. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron. 2017 ;93(2) :451–463. doi :https ://doi.org/10.1016/j.neuron.2016.12.040.

[197] Funke CM, Borowski J, Stosio K, Brendel W, Wallis TSA, Bethge M. Five points to check when comparing visual perception in humans and machines. Journal of Vision. 2021 ;21(3) :1–23.

[198] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. 2017 ;.

[199] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery. 2019 ;9(4) :e1312.

[200] Adebayo J, Gilmer J, Muelly M, Goodfellow IJ, Hardt M, Kim B. Sanity Checks for Saliency Maps. CoRR. 2018 ;abs/1810.03292.

[201] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In : 2017 IEEE International Conference on Computer Vision (ICCV) ; 2017. p. 843–852.

[202] Shao J, Wang Q, Liu F. Learning to Sample : an Active Learning Framework. CoRR. 2019 ;abs/1909.03585.

[203] Vapnik V, Vashist A. A new learning paradigm : Learning using privileged information. Neural Networks. 2009 ;22(5) :544–557. doi :https ://doi.org/10.1016/j.neunet.2009.06.042.

[204] Jain S, Yarlagadda P, Jyoti S, Karthik S, Subramanian R, Gandhi V. ViNet : Pushing the limits of Visual Modality for Audio-Visual Saliency Prediction ; 2021.

[205] Pan J, Canton-Ferrer C, McGuinness K, O'Connor NE, Torres J, Sayrol E, et al. SalGAN : Visual Saliency Prediction with Generative Adversarial Networks. CoRR. 2017 ;abs/1701.01081.

[206] Borji A. Saliency Prediction in the Deep Learning Era : An Empirical Investigation. CoRR. 2018 ;abs/1810.03716.

[207] Wang F, Kaushal R, Khullar D. Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine ? Annals of internal medicine. 2020 ;172(1) :59–60. doi :10.7326/M19-2548.

[208] Gastounioti A, Kontos D. Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI ? Radiology Artificial intelligence. 2020 ;2(3) :e200088. doi :10.1148/ryai.2020200088.

[209] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021 ;3(11) :e745–e750. doi :10.1016/S2589-7500(21)00208-9.

[210] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet : Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. CoRR. 2017 ;abs/1711.05225.

Annexe A

Exemples d'attention artificielle sans traitement *a priori*

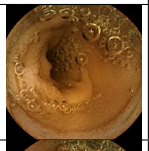
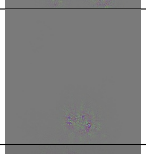
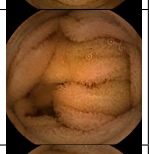
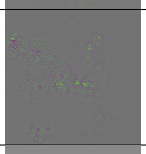
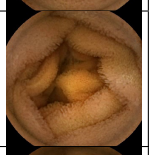
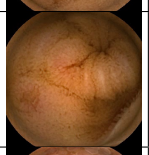
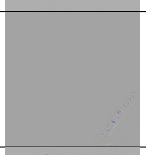
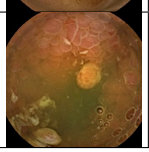
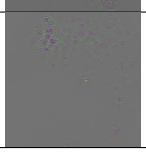
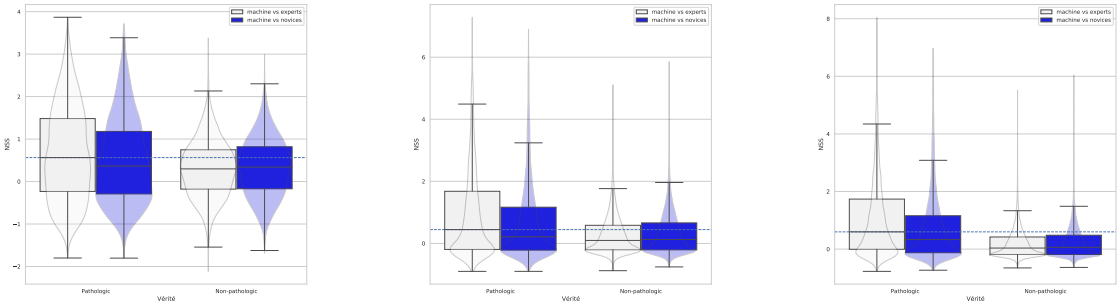
Label	Originale	Rétro-propagation	Rétro-propagation guidée	Guided GradCAM
Non-pathologique				
Pathologique				
Non-pathologique				
Non-pathologique				
Pathologique				
Pathologique				

Table A.1 – Résultats visuelles des différentes méthodes d'extraction de l'attention post-hoc avec sans traitement spécifique.

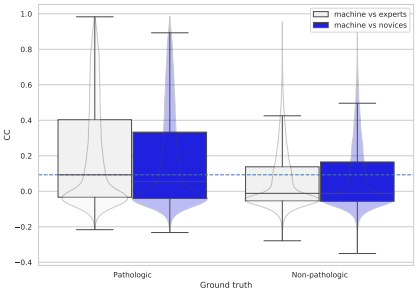
Annexe B

Comparaison attention humaine et artificielle pour les différentes méthodes pour les différents réseaux avec les métriques NSS et CC

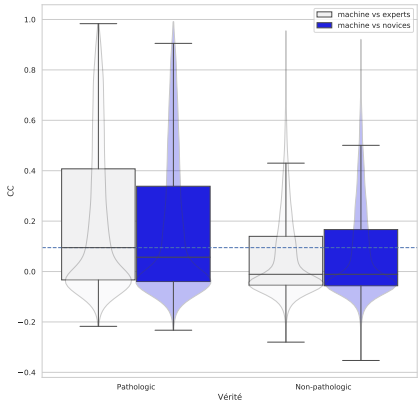


(a) GradCAM (b) Guided GradCAM (c) Rétropropagation guidée

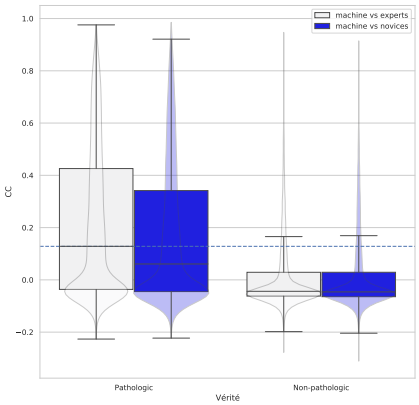
Figure B.1 – Résultats de la comparaison des différentes méthodes d'extraction de l'attention artificielle post hoc avec l'attention humaine en fonction du niveau d'expertise et du label des images. Ces résultats ont été obtenus pour le réseau ResNet34 et la comparaison a été réalisée à l'aide de la métrique NSS. Plus le score de NSS est grand, plus les zones d'attention entre machine et humains sont proches.



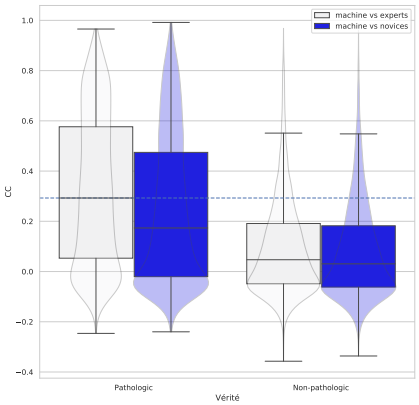
(a) GradCAM



(b) Guided GradCAM

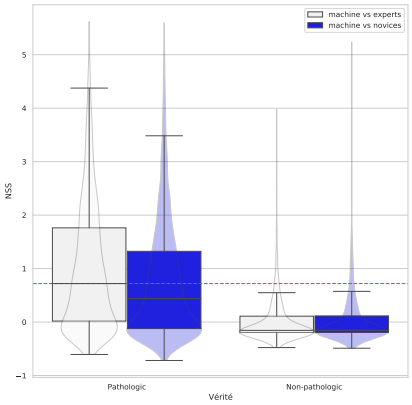


(c) Rétropropagation guidée

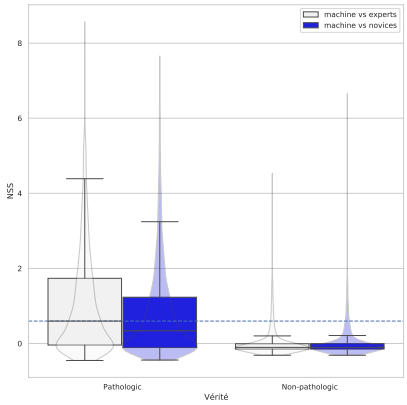


(d) Gradients

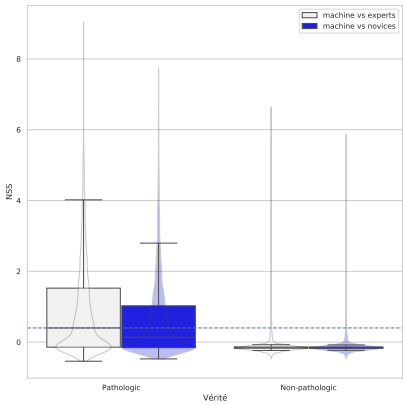
FIGURE B.2 – Résultats de la comparaison des différentes méthodes d'extraction de l'attention artificielle post hoc avec l'attention humaine en fonction du niveau d'expertise et du label des images. Ces résultats ont été obtenus pour le réseau VGG16 et la comparaison a été réalisée à l'aide de la métrique CC. Plus le score de CC est grand, plus les zones d'attention entre machine et humains sont proches.



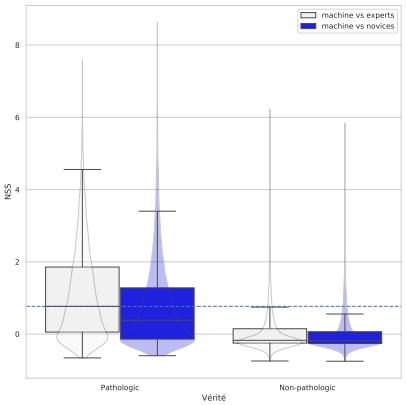
(a) GradCAM



(b) Guided GradCAM

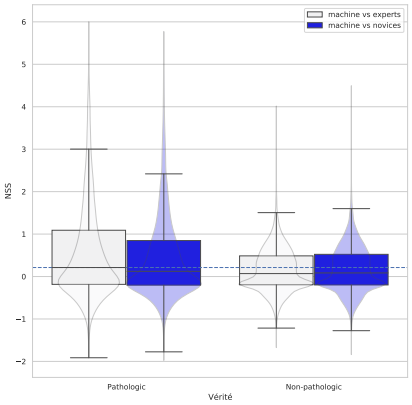


(c) Rétropropagation guidée

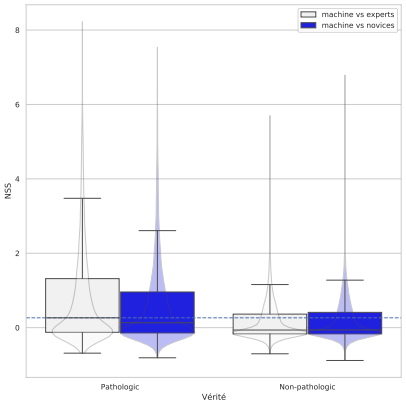


(d) Gradients

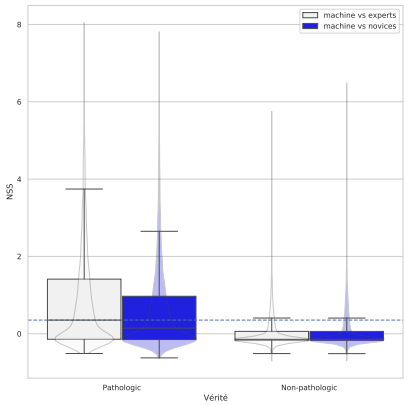
FIGURE B.3 – Résultats de la comparaison des différentes méthodes d'extraction de l'attention artificielle post hoc avec l'attention humaine en fonction du niveau d'expertise et du label des images. Ces résultats ont été obtenus pour le réseau VGG19 et la comparaison a été réalisée à l'aide de la métrique CC. Plus le score de CC est grand, plus les zones d'attention entre machine et humains sont proches.



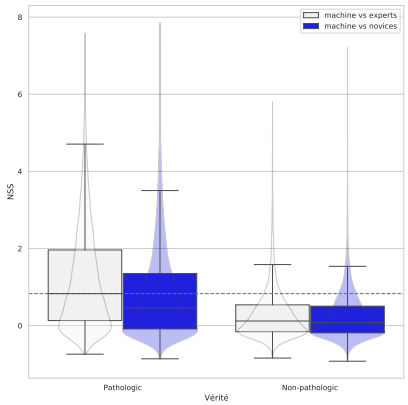
(a) GradCAM



(b) Guided GradCAM

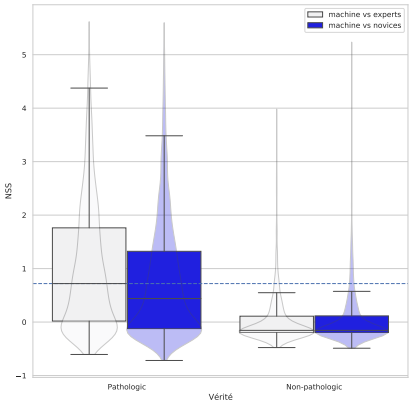


(c) Rétropropagation guidée

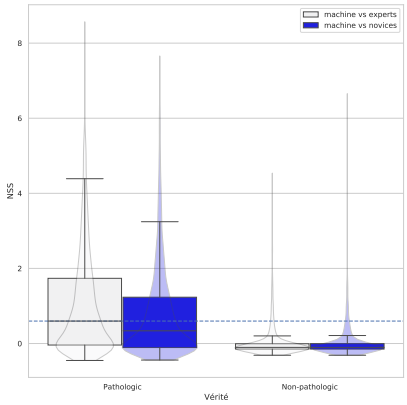


(d) Gradients

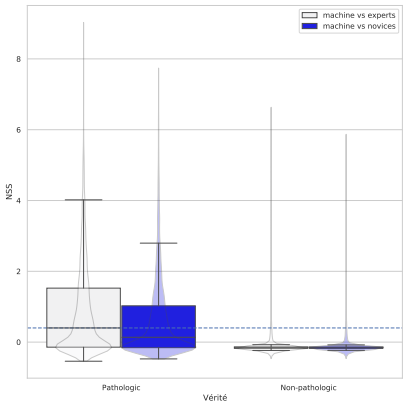
FIGURE B.4 – Résultats de la comparaison des différentes méthodes d'extraction de l'attention artificielle post hoc avec l'attention humaine en fonction du niveau d'expertise et du label des images. Ces résultats ont été obtenus pour le réseau VGG16 et la comparaison a été réalisée à l'aide de la métrique NSS. Plus le score de NSS est grand, plus les zones d'attention entre machine et humains sont proches.



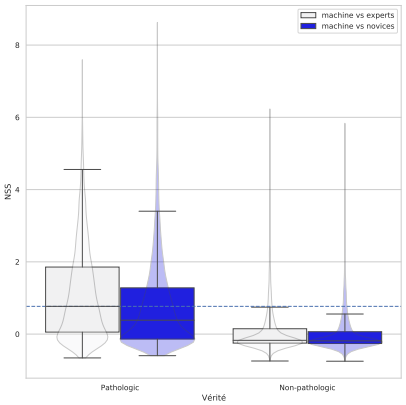
(a) GradCAM



(b) Guided GradCAM



(c) Rétropropagation guidée



(d) Gradients

FIGURE B.5 – Résultats de la comparaison des différentes méthodes d'extraction de l'attention artificielle post hoc avec l'attention humaine en fonction du niveau d'expertise et du label des images. Ces résultats ont été obtenus pour le réseau VGG19 et la comparaison a été réalisée à l'aide de la métrique NSS. Plus le score de NSS est grand, plus les zones d'attention entre machine et humains sont proches.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

Titre : Apprentissage profond pour l'aide au diagnostic et comparaison des mécanismes d'explicabilité avec l'attention visuelle humaine : application à la détection de la maladie de Crohn

Mot clés : Apprentissage profond, oculométrie, explicabilité, endoscopie, attention

Résumé : Quels sont les points communs et les différences entre notre façon de percevoir notre environnement et celles des réseaux de neurones profonds ? Nous étudions cette question au travers d'un cas d'application concret, la détection des lésions issues de la maladie de Crohn dans des vidéos capsules endoscopiques. Dans un premier temps, nous avons développé une base de données, soigneusement annotée par plusieurs experts, que nous avons rendu publique afin de compenser le manque de données permettant l'évaluation et l'entraînement des algorithmes d'apprentissage profond dans ce domaine. Dans un second temps, pour rendre les réseaux plus transparents lors de leur prise de décision et leurs prédictions plus explicables, nous avons travaillé sur l'attention artificielle et établissons un parallèle entre celle-ci et l'attention visuelle humaine. Nous avons enregistré les mouvements oculaires de sujets de différents niveaux d'expertise lors d'une tâche de classification et montrons que les réseaux de neurones profonds, dont les performances sur la tâche de classification sont plus proches de celles des experts que de celles des novices, ont également un comportement attentionnel plus proche de ces premiers. Au travers de ce manuscrit, nous espérons fournir des outils permettant le développement d'algorithmes d'aide au diagnostic, ainsi qu'un moyen d'évaluer les méthodes d'attention artificielle. Ce travail permet d'approfondir les liens entre attention humaine et artificielle, dans le but d'aider les experts médicaux dans leur formation et d'aider au développement de nouvelles architectures d'algorithmes.

Title: Deep learning for diagnostic support and comparison of explicability mechanisms with human visual attention: application to Crohn's disease detection

Keywords: Deep learning, eye tracking, explicability, endoscopy, attention

Abstract: What are the similarities and differences between the way we perceive our environment and that of deep neural networks? We study this question through a concrete application case, the detection of lesions from Crohn's disease in endoscopic video capsules. In a first step, we have developed a database, carefully annotated by several experts, which we have made public in order to compensate for the lack of data allowing the evaluation and training of deep learning algorithms in this domain. In a second step, to make the networks more transparent in their decision making and their predictions more explainable, we worked on artificial attention and establish a parallel between it and human visual attention. We have recorded the eye movements of subjects of different levels of expertise during a classification task and show that deep neural networks, whose performance on the classification task is closer to that of experts than to novices, also have an attentional behavior closer to the former. Through this manuscript, we hope to provide tools for the development of diagnostic assistance algorithms, as well as a way to evaluate artificial attention methods. This work provides a deeper understanding of the links between human and artificial attention, with the goal of assisting medical experts in their training and helping to develop new algorithm architectures.