Réseau de neurones récurrent à attention pour la détection de lésions intestinales

R. Vallée [1] H. Mouchère [1] N. Normand [1] A. Coutrot [1]
A. De Maissin[2] A. Boureille [2]

[1] Laboratoire des Sciences du Numérique de Nantes
[2] Centre Hospitalier et Universitaire de Nantes

remi.vallee@univ-nantes.fr

Résumé

La maladie de Crohn (MC) est une maladie inflammatoire chronique intestinale, atteignant les sujets jeunes, provoquant des lésions muqueuses de l'intestin grêle : des érosions, des ulcérations, de l'œdème et des sténoses. La vidéo-capsule endoscopique (VCE) est le meilleur examen permettant leur détection. La VCE génère environ 50000 images dont l'analyse par les gastroentérologues est consommatrice de temps. L'objectif de notre travail est donc de développer un outil permettant la reconnaissance automatique des lésions muqueuses de MC dans l'intestin grêle. L'algorithme est basé sur un réseau de neurones convolutifs à attention. Celui-ci a été entraîné sur une base de données publique, GIANA, contenant des images de VCE normales et avec des lésions inflammatoires et vasculaires. Une autre base a été utilisée séparément, CROHN-IPI, constituée d'images normales et de lésions de MC annotées par des gastro-entérologues du CHU de Nantes. Les résultats préliminaires montrent que l'algorithme entraîné sur les 1800 images de GIANA, est capable de détecter avec une précision de 99,77% les images pathologiques des images non pathologiques. Concernant CROHN-IPI, la précision obtenue sur les 2500 images provenant de 39 patients est de 80.36%. Cet écart peut s'expliquer par la façon dont été sélectionner les images de la base (images de lésions plus évidentes dans GIANA) ou encore par la sous-représentation de certaines pathologies dans CROHN-IPI. A l'avenir, une application d'annotation de VCE à plus grande échelle sera développée pour enrichir CROHN-IPI.

Mots Clef

Apprentissage profond, réseau récurrent à attention, maladie de Crohn, classification d'images médicales.

Abstract

Crohn's disease (CD) is a chronic inflammatory bowel disease, affecting young subjects, causing mucosal damage to the small intestine : erosions, ulcerations, edema and stenosis. The wireless capsule endosopy (WCE) is the best examination for their detection. The WCE generates approximately 50,000 images that are time-consuming for gastroenterologists to analyse. The objective of our work is therefore to develop a tool for the automatic recognition of mucosal lesions of CD in the small intestine. The algorithm is based on a network of convolutional neurons for attention. This was trained on a public database, GIANA, containing images of normal WCEs and with inflammatory and vascular lesions. Another database was used separately, CROHN-IPI, consisting of normal images and MC lesions annotated by gastroenterologists from Nantes University Hospital. Preliminary results show that the algorithm trained on the 1800 GIANA images, is able to detect with an accuracy of 99,77% the pathological images of non-pathological images. Concerning CROHN-IPI, the accuracy obtained on the 2500 images from 39 patients is 80.36%. This difference can be explained by the way the images in the database were selected (images of more obvious lesions in GIANA) or by the under-representation of certain pathologies in CROHN-IPI. In the future, a larger scale WCE annotation application will be developed to enrich CROHN-IPI.

Keywords

Deep learning, recurrent attention model, Crohn's disease, medical images classification.

1 Introduction

Depuis le milieu des années 1990, le développement de la VCE a permis une avancée dans le diagnostique des maladies gastro-intestinale. Cette technologie peu invasive présente des performances de diagnostic importantes du fait de sa capacité à rendre possible l'exploration complète des 3 à 4 mètres d'intestin grêle. Elle est maintenant bien acceptée par l'ensemble de la communauté scientifique [1]. La maladie de Crohn et l'angiodysplasie, sont deux maladies dont le diagnostic repose en grande partie sur la détection de lésions intestinales qui peuvent grâce à la VCE être plus facilement détectées.

Le diagnostique de la maladie de Crohn repose sur deux

scores : *The Capsule Endoscopy Crohn's Disease Activity Index* (CECDAI) [2] et le score de Lewis [3]. Ces indicateurs dépendent du nombre de lésions, de leurs types et de leurs localisations dans l'intestin. Ces scores dépendant du nombre et de l'identification précise des lésions, il est nécessaire que les experts gastro-entérologues visionnent l'ensemble des 50 000 images générées par la VCE pour chaque patient, ce qui peut s'avérer être très consommateur en temps.

Différents systèmes ont été développés dans le but de limiter le temps d'analyse des données par les gastro-entérologues en identifiant les images potentiellement pathologiques. De nombreux efforts ont été réalisés dans le but de détecter les lésions intestinales à partir d'images obtenues par VCE [4–10]. Ces travaux peuvent être classés en deux catégories principales. La première, les algorithmes non-basés sur l'apprentissage profond, certains se basant sur des caractéristiques liées au couleurs des images en partant de l'idée que la couleur des ulcères est majoritairement blanche/jaune et d'autre se basant sur une analyse des caractéristiques liées à la texture. La classification des images une fois cette extraction des caractéristiques réalisée peut être effectuée par un SVM ou par des réseaux de perceptrons multicouches. Le problème de cette première catégorie est qu'elle est peu robuste aux variations d'éclairage, d'angle de vue et souvent spécifique à un type de lésion, la rendant donc plus difficilement applicable en situation réelle.

La deuxième catégorie, présenté dans la section 2 repose sur des algorithmes d'apprentissage profond. Ces algorithmes se basent sur des réseaux de neurones convolutifs (CNN) permettant l'extraction de caractéristiques automatiquement par le réseau de neurones, les rendant plus résilients, mais complexifiant leur entraînement du fait du nombre important de données nécessaires pour ajuster les très nombreux paramètres de ces réseaux.

Grâce au partenariat entre le Centre Hospitalier et Universitaire de Nantes et le Laboratoire des Sciences du Numérique, une importante quantité de données a pu et pourra être récuppérées. De plus, à l'heure actuelle, une plateforme d'annotation collaborative est en cours de réalisation, ce qui permettra dans le futur d'obtenir une plus grande quantité de données afin d'améliorer l'entraînement d'un réseau de neurones profond. Le choix à donc été fait de réaliser un réseau de neurones récurrent à attention à la manière de Mnih et al. [11]. Ce réseau, plutôt que de prendre en entrée la totalité de l'image, choisit grâce à un apprentissage non-supervisé plusieurs patchs successivement extraits dans l'image d'origine qui permettront de prendre une décision de classification à l'issue de ces multiples "regards" sur la même image. Ainsi le fonctionnement du réseau se rapproche du comportement oculaire de l'être humain. L'objectif de ce réseau est donc double. Dans un premier temps, il permet la classification des images endoscopiques pour l'aide au diagnostic et dans un second temps il permettra la comparaison entre le fonctionnement de ce réseau de neurones artificielles et le comportement des experts face à des images issues de VCE. Les résultats obtenus par cette comparaison permettront alors sans doute l'amélioration du réseau initialement conçu et présenté dans cet article.

2 État de l'art

La deuxième catégorie, [12–14]. Les différents réseau se démarquent de par leur différentes architectures et les données utilisées pour les entrainer. Aoki et al.[12] propose un réseau basé sur une architecture nommé Single Shot multibox Detector [15]. Ce réseau apprend à générer des boites de délimitation autour des zones considérées comme pathologiques.

3 Architecture du réseau

3.1 Réseau de neurones à attention

Dans cet article on considère l'attention comme un processus de décision séquentiel d'un agent interagissant avec un environnement visuel. [11] On fourni au réseau d'entrée une image endoscopique X. Le Glimpse Sensor va ensuite extraire un patch $\rho(X)$ de l'image d'origine fonction de $l_t = (x, y, z)$ avec x et y les coordonnées normalisées entre $[-1; 1]$ et $(0, 0)$ le centre de l'image. z est un coefficient de zoom compris entre $]0; 1]$. Avec $s_x \times s_y$ la résolution de l'image d'origine, on obtient un patch de résolution $(s_x \times z) \times (s_y \times z)$ aux coordonnées $(s_x \times x, s_y \times y)$. Ce patch est ensuite redimensionné de façon bilinéaire pour conserver une taille fixe en entrée du réseau. Le *What ? Network* basé sur VGG16 [16] pré-entrainé sur ImageNet [17] permet d'extraire les caractéristiques du patch $\rho(X)$. Seules les couches convolutives du VGG et la première couche *fully-connected* ont été conservées. En parallèle les informations relatives à l'extraction du patch l_t, traversent le *Where ? Network* composé de 2 couches *fully-connected*, permettant ainsi l'extraction des caractéristiques relatives à la position d'extraction du patch $\rho(X)$. Les deux vecteurs de caractéristiques produits par ce réseau sont alors de taille identique et seront sommés ensemble avant de subir une 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 *Glimpse Network* au temps t. Une Gated Recurrent Unit [18] (GRU) permet ensuite de fusionner les caractéristiques extraite au temps t par le réseau avec celles extraites au temps précédent contenu dans l'état interne précédent de la GRU h_{t-1}. Cet état interne de la GRU sera réutilisé au temps suivant.

A partir du nouvel état interne produit par la GRU, l'*action network* va produire un vecteur associant un score à chacune des classes. Le *Baseline Network* va lui permettre de calculer la récompense associée à une prédiction afin de pouvoir entraîner le *Localisation Network* par renforcement. Ainsi le réseau va augmenter la probabilité

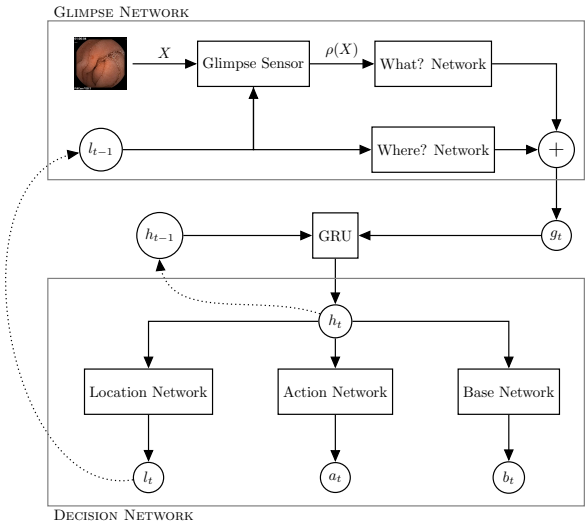

FIGURE 1 – **Architecture du réseau à attention** : À chaque temps t, on fourni au *Glimpse sensor* une image endoscopique X et la localisation l_{t-1} du patch à extraire de l'image d'origine. Le patch $\rho(X)$ et la localisation l_{t-1} sont ensuite traités par deux réseaux de neurones indépendants afin de produire un vecteur caractéristique g_t. Une Gated Recurrent Unit (GRU) va ensuite permettre de fusionner les caractéristiques extraites précédemment par le réseau contenu dans le vecteur h_{t-1} avec g_t afin de produire l'état actuelle du système h_t. A partir de cet état trois sous réseau vont produire indépendamment l_t la position du prochain patch à extraire, a_t un vecteur contenant un score associé à chacune des classes et b_t, la *baseline* à partir de laquelle est calculée la récompense de l'apprentissage par renforcement.

des localisations maximisant la fonction récompense. Si le réseau classifie correctement l'image, la récompense vaut alors le nombre de regards placés sur l'image auquel est soustraite la baseline calculée par le réseau.

3.2 Fonction de coût

La fonction de coût du réseau est une combinaison linéaire de trois sous-fonctions de coût (1). Elle est similaire à celle présentée par Mnih et al. 2014 [11].

$$\mathcal{L} = \mathcal{L}_{action} + \mathcal{L}_{baseline} + \mathcal{L}_{reinforce} \tag{1}$$

Premièrement $\mathcal{L}_{action}$ (2) la fonction d'entropie croisée permettant de calculer l'erreur de classification du réseau lors du dernier regard sur l'image avec $\hat{Y}_i$ la prédiction du réseau et Y_i la vérité terrain.

$$\mathcal{L}_{action} = -\frac{1}{n}\sum_{i=1}^{n} Y_i log(\hat{Y}_i) \tag{2}$$

Ensuite la fonction $\mathcal{L}_{baseline}$ (3), calculant l'erreur quadratique moyenne entre la récompense R_t^i et la baseline b_t établie par le réseau. La récompense est égale au nombre de regards posés sur l'image si elle a été correctement classifié, 0 sinon. Cette *baseline* dépendant du contexte présent et passé, permet d'ajuster la récompense de sorte à pousser le réseau à s'améliorer par rapport à ses résultats précédents.

$$\mathcal{L}_{baseline} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}(b_t - R_t^i)^2 \tag{3}$$

La $\mathcal{L}_{reinforce}$ (4) elle permet de mettre en place l'apprentissage non-supervisé par renforcement. Le principe est le suivant : une stratégie $\pi(\tau^j;\theta)$, dépendante des paramètres du réseau θ associe une probabilité au trajet τ^j. Un trajet est une succession d'actions u_t, ici le changement de la localisation de l'extraction d'un patch, entraînant un nouvel état s_t, ici un nouveau patch en entrée du système. Ce nouvel état influence alors le système qui produira une nouvelle action qui conduira à un nouvel état. Ainsi l'objectif de l'algorithme de renforcement est de maximiser la probabilité des trajets qui maximise la récompense $(R_t - b_t)$ et a contrario de diminuer celle des trajet conduisant à une récompense faible. En réalisant la somme des récompenses associées aux différents trajets possibles M coefficientée par leurs probabilités respectives on obtient la fonction de coût de l'algorithme de d'apprentissage par renforcement.

$$\mathcal{L}_{reinforce} = -\frac{1}{n}\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{M}\sum_{t=1}^{T} log\pi(u_t^j|s_{i:t}^j;\theta)(R_t^j - b_t) \tag{4}$$

3.3 Initialisation du premier regard

A chaque nouvelle image la matrice d'état h_0 de la GRU est remise à zéro afin que chacune des prédictions soit indépendante. Il est proposé ici d'étudier deux possibilité d'initialisation du premier regard sur l'image. La première, dite "libre", laisse le réseau apprendre à positionner lui même sont premier regard. Se premier regard est alors dépendant de la dernière image vu. La deuxième possibilité, dite "centrée", s'inspire du comportement humain.D'après B. W. Tatler [13], quelle que soit la répartition de l'information dans l'image, lors de l'apparition d'une nouvelle image, l'homme a tendance à regarder son centre. Ce phénomène est appelé biais de fixation centrale. Ainsi à chaque nouvelle image, le réseau est forcé de procéder à un regard central englobant toute l'image ($(x, y, z) = (0, 0, 1)$).
La comparaison entre le comportement des premières zones du cortex visuel humain et celui des réseau neurones est possible d'après Cichy et al. [19]. Ainsi les expériences menés sur l'effet du biais de fixation centrale artificielle sont un premier pas sur l'étude des stratégies d'exploration humaine à travers l'étude du comportement d'un réseau de neurones artificielle.

4 Résultats et discussion

Les différents algorithmes de segmentation et de détection d'images médicales cités précédemment ont chacun été testé sur des bases de données différentes, ne contenant pas toutes les mêmes types de lésions et avec des images ayant des lésions plus ou moins évidentes. Cette hétérogénéité des bases de données ne permet pas de pouvoir bien comparer les algorithmes entre eux et leur efficacité en condition réelle. C'est pour pallier ce problème que nous avons choisi, en plus de tester et entraîner notre algorithme sur notre propre base de données CROHN-IPI, d'également l'entraîner et l'évaluer sur GIANA, une base de données utilisée lors d'un challenge à MICCAI 2017 [20].

4.1 Résultats sur Giana

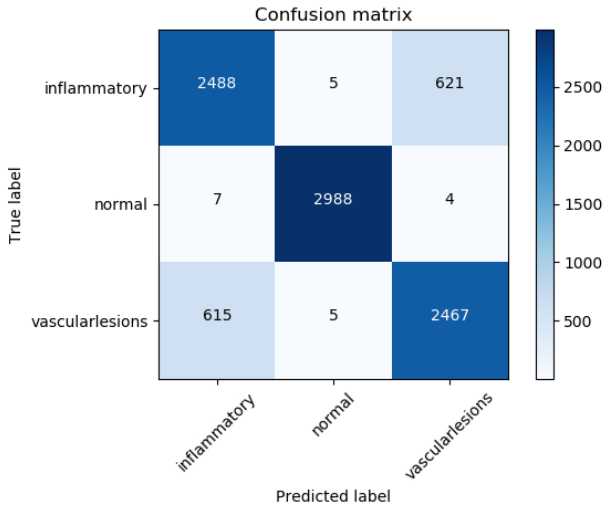

FIGURE 2 – Matrice de confusion sur pour 10 entrainements/tests sur la base de données GIANA

Le dataset contient 1800 images obtenues par VCE réparties en 3 classes : 600 images avec des lésions vasculaires, 600 avec des inflammatoires et 600 ne contenant aucune pathologie. Les images de la base d'entraînement subissent des transformations aléatoires lors de l'entraînement afin d'augmenter la diversité des exemples permettant ainsi au réseau de généraliser plus facilement. 50% des images de la base sont utilisées pour entraîner le réseau et 50% pour l'évaluation. Pour obtenir des résultats statistiquement significatifs, on entraîne et évalue le réseau sur 10 sous-datasets. La matrice de confusion présentée figure 2 montre les résultats cumulés sur les 10 différents sous-dataset. On obtient alors une précision de 86,33% sur la tâche de classification entre les trois classes. On remarque que la majorité des erreurs commises par le réseau sont principalement concentrées dans la classification des lésions inflammatoires en lésions vasculaires et inversement. Partant de ce constat, les classes lésions vasculaires et lésions inflammatoires ont été regroupées en une classe "pathologique". Les résultats obtenus pour les différentes configuration du réseau sont présenté dans le tableau 1. Deux méta-paramètres ont varié lors de ces expériences, l'initialisation du premier regard sur l'image et le nombre de regards placé sur chaque image. Les meilleurs résultats sont obtenus pour 4 regards sur l'image. Une précision maximum de 87,51% est obtenue pour la tâche de classification à trois classes (lésion inflammatoire, lésion vasculaire et non pathologique) en laissant la localisation du premier regard libre. La précision maximum sur ma tâche de classification à deux classes (pathologique, non pathologique) est de 99,77% soit une erreur toutes les 434 images. Sur la deuxième tâche le meilleur résultat est obtenu en forçant le premier regard à englober l'ensemble de l'image. Des exemples de regards sur quatre images de GIANA sont présent sur la figure 3.

4.2 Résultats sur CROHN-IPI

La base de données de CROHN-IPI comprend 2 590 images réparties en 8 classes différentes, 7 correspondant à des types de lésions pathologiques différentes (ulcération aphtoïde, ulcération de 3 à 10 mm, ulcération de plus de 10 mm, œ]dème, sténose, érythème, pseudo-polype) et une classe contenant des images intestins non-pathologiques. Ces données ont été annotées par deux gastro-entérologues et proviennent de 39 patients différents.
Les images n'étant pas équitablement réparties entre les classes, les classes pathologiques ont été regroupées ensemble. L'entraînement du réseau sur cette base se fait sur 970 images équitablement réparties en 2 classes : intestins sains et intestins pathologiques. Le test se fait sur les 1 620 images restantes comprenant 70% d'images pathologiques et 30% de non-pathologiques. Les résultats obtenus sont regroupé dans les tableau 2 et 3. Une précision maximum de 80,36 est obtenue après cross-validation sur 10 *sous-datasets* différents, pour une architecture procédant à 4 regards de résolution 80x80 sur l'image issue de VCE. Lors de ces différentes expériences les trois principaux méta-paramètres du réseau ont subi des variations : la résolution du patch d'entrée du réseau, le nombre de regards posés sur l'image et l'initialisation du premier regard.

4.3 Évaluation du biais de fixation centrale artificiel

Comme présenté précédemment une expérience à été mené dans le but d'évaluer l'efficacité du biais de fixation centrale artificielle sur une tache de classification d'images endoscopiques. En cumulé, 22 expériences différentes ont été réalisées dans le but de vérifier l'hypothèse que le biais de fixation central permet d'améliorer les scores du réseau. Ces 22 expériences étant cross-validées sur 10 sous data-set différents d'en moyenne 1300 images. Un test de permutation à donc été effectué sur les résultats des 220 sous-expériences. L'écart entre la moyenne du groupe de contrôle contenant les précision associé au 110 expériences sans biais centré avec celle du groupe de test contenant les 110 expériences avec le biais centrale est de 1,1%. Suite à cela 10 000 permutations aléatoires sont réalisées entre

Nombre de regards	Précision 3 classes	Précision	Spécificité	Sensibilité	F1	F2
3 regards	85.42%	98.67%	98.53%	98,97%	99,23%	99,50%
3 regards centrés	84,37%	99,60%	99,77%	99,23%	99,43%	99,63%
4 regards	**87,51%**	99,30%	99,44%	99,03%	99,28%	99,53%
4 regards centrés	86,34%	**99,77%**	**99,84%**	**99,63%**	**99,73%**	**99,82%**
5 regards	86,46%	99,01%	99,10%	98,83%	99,13%	99,43%
5 regards centrés	86,46%	98,75%	98,58%	99,10%	99,33%	99,56%

TABLE 1 – Tableau récapitulatif des expériences menées sur la base de donnée GIANA

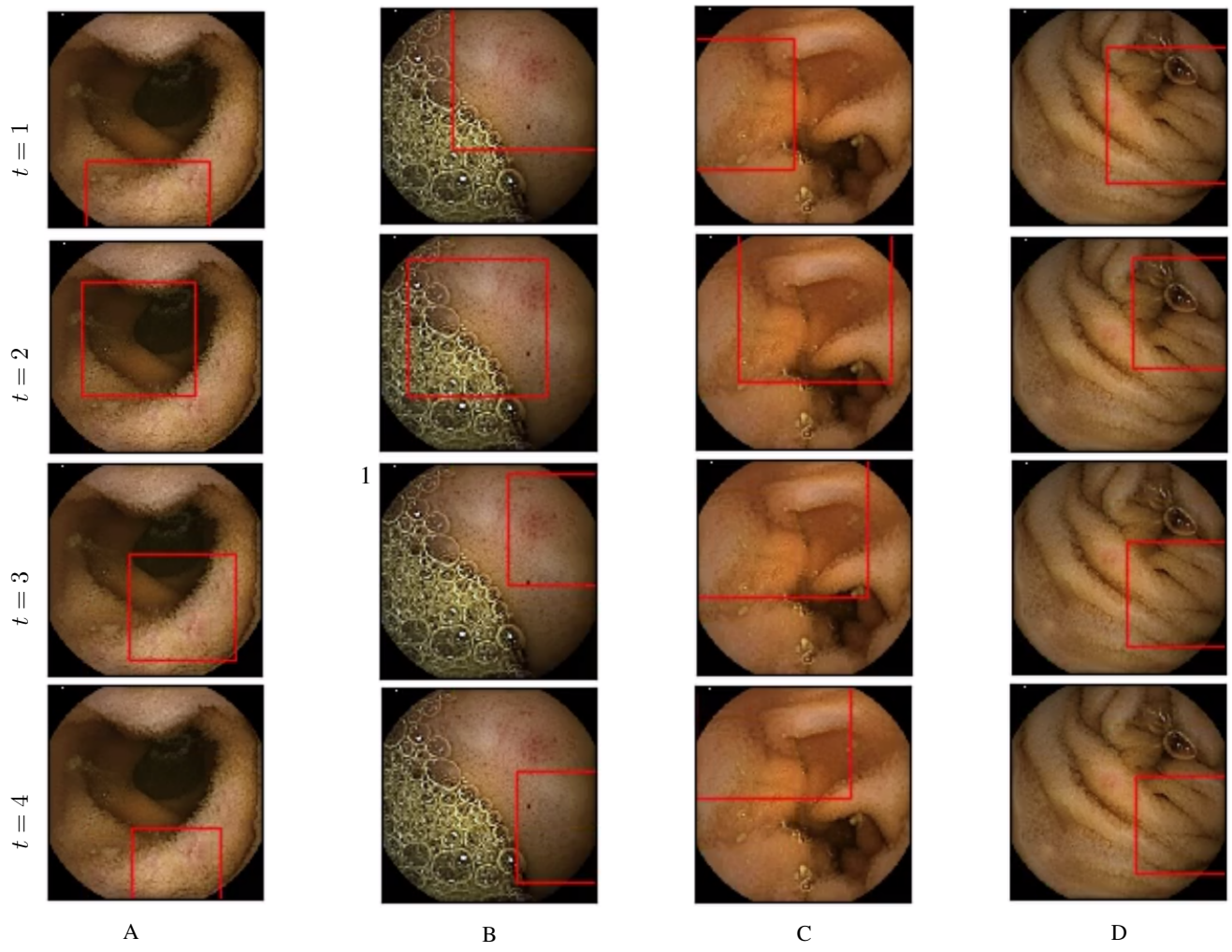

FIGURE 3 – Exemples de regards sur 4 images de GIANA. La colonne A présente un cas de lésion inflammatoire correctement classifiée, la colonne B un cas de lésion vasculaire correctement classifiée, la colonne C un cas d'image non pathologique correctement classifiée, un la colonne D un cas de lésion inflammatoire classé en lésion vasculaire

les différents groupes pour chaque permutation une nouvelle moyenne est calculé. On obtient à l'issue de ce test $p = 0,0894$ ce qui ne permet pas de valider l'hypothèse que le biais centré améliore la précision du réseau. Les résultats obtenues pour les différentes métriques en fonction des différents datasets sont présenté dans le tableau 4. A partir de ce tableau il est possible de valider l'hypothèse de départ sur le dataset GIANA, la moyenne des p_{value} étant nettement inférieur à 5%. Cependant la généralisation de l'hypothèse de départ n'est pas possible. Sur le dataset CROHN-IPI, les p_{value} obtenues sont proche de 50%, significatif du fait que les moyennes sur l'ensemble des métriques entre les deux groupes sont très proche.

Nombre de regards	Précision	Spécificité	Sensibilité	F1	F2
3 regards	76,66%	71,68	88,14%	90,65%	93,30%
3 regards centrés	78,55%	74,40%	88,10%	90,72%	93,51%
4 regards	**80,03%**	**77,23%**	86,47%	89,59%	92,93%
4 regards centrés	77,86%	73,30%	**88,36%**	**90,88%**	**93,53%**
5 regards	78,24%	74,24%	87,47%	90,23%	93,17%
5 regards centrés	79,77%	76,81	86,61%	89,68%	92,97%

TABLE 2 – Tableau récapitulatif des expériences menées sur la base de donnée CROHN-IPI

Taille des patchs	Précision	Spécificité	Sensibilité	F1	F2
Patchs 112x112	80,03%	77,23%	86,47%	89,59%	92,93%
Patchs 112x112 centrés	77,86%	73,30%	88,36%	90,88%	93,53%
Patchs 96x96	79,45%	76,22%	86,86%	89,84%	93,04%
Patchs 96x96 centrés	77,48%	72,94	87,93	90,54	93,30
Patchs 80x80	**80,36%**	**77,34%**	87,32%	90,24%	93,36%
Patchs 80x80 centrés	78,34%	74,47%	87,24%	90,07%	93,08%
Patchs 64x64	78,22%	73,36%	**89,42%**	**91,71%**	**94,11%**
Patchs 64x64 centrés	79,83%	76,47%	87,57%	90,40%	93,41%
Patchs 48x48	78,17%	73,84%	88,16%	90,75%	93,49%
Patchs 48x48 centrés	78,56%	74,42%	87,14%	90,00%	93,06%
Patchs 32x32	75,82%	70,08%	89,05%	91,30%	93,65%
Patchs 32x32 centrés	77,16%	72,41%	88,10%	90,64%	93,34%

TABLE 3 – Tableau récapitulatif des expériences menées sur la base de donnée CROHN-IPI (A mettre en annexe je pense)

Métrique	p_{GIANA}	p_{CI}	p
Taille de l'échantillon	60	160	220
Précision	0,014	0.226	0.185
Spécificité	0.009	0.217	0.245
Sensibilité	0.898	0.377	0.594

TABLE 4 – Tableau des différentes p_{value} obtenue par test de permutations sur les différentes métriques selon les différents datasets

4.4 Discussion

Les résultats obtenus sont largement inférieurs sur la base de données CROHN-IPI que sur la base de données GIANA (Une erreur sur toutes les 5 images sur Crohn-IPI contre une erreur toutes les 400 images sur GIANA). Cet écart peut s'expliquer de plusieurs manières. Premièrement, les lésions inflammatoires de l'angiodysplasie sont nettement plus reconnaissables de par leur couleur rouge vif que les lésions de la maladie de Crohn. Deuxièmement, le dataset GIANA comprends 600 images de chacune des lésions et 600 images non-pathologiques, ce qui permet au réseau de pouvoir s'entraîner sur 300 exemples avant d'être testé. Pour CROHN-IPI, certaines classes lésions comme les sténoses ou les érythèmes sont sous-représentées dans le dataset (moins de 70 exemples). Ensuite, cela peut aussi s'expliquer par la façon dont ont été choisies les images de GIANA par rapport à celles de CROHN-IPI. De façon qualitative, les images de GIANA semblent nettement plus "propres", contenant moins de bruit lié à la capture de l'image (flou, faible luminosité) et de bruit lié à l'activité intestinale (moins de bulles sur les images).
Les travaux à venir tenteront de résoudre les problèmes de la base de données grâce au développement d'un outil d'annotation pour faciliter la récupération de nouvelles données.

5 Conclusion

Références

[1] Marisol Luján-Sanchis, Laura Sanchis-Artero, Laura Larrey-Ruiz, Laura Peño-Muñoz, Paola Núñez-Martínez, Génesis Castillo-López, Lara González-González, Carlos Boix Clemente, Cecilia Albert Antequera, Ana Durá-Ayet, and Javier Sempere-Garcia-Argüelles. Current role of capsule endoscopy in Crohn's disease. *World journal of gastrointestinal endoscopy*, 8(17) :572–83, sep 2016.

[2] Eyal Gal, Alex Geller, Gerald Fraser, Zohar Levi, and Yaron Niv. Assessment and Validation of the New Capsule Endoscopy Crohn's Disease Activity Index (CECDAI). *Digestive Diseases and Sciences*, 53(7) :1933–1937, jul 2008.

[3] I. M. Gralnek, R. Defranchis, E. Seidman, J. A. Leighton, P. Legnani, and B. S. Lewis. Development of a capsule endoscopy scoring index for small bowel mucosal inflammatory change. *Alimentary Pharmacology & Therapeutics*, 27(2) :146–154, oct 2007.

[4] Baopu Li and Max Q.-H. Meng. Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments. *Computers in Biology and Medicine*, 39(2) :141–147, 2009.

[5] Leontios J Hadjileontiadis. Use of adaptive hybrid filtering process in Crohn&apos ;s disease lesion detection from real capsule endoscopy videos. *Healthcare Technology Letters*, 3(1) :27–33(6), 2016.

[6] Y. Yuan, J. Wang, B. Li, and M. Q. . Meng. Saliency based ulcer detection for wireless capsule endoscopy diagnosis. *IEEE Transactions on Medical Imaging*, 34(10) :2046–2057, Oct 2015.

[7] W. S. L. Jebarani and V. J. Daisy. Assessment of crohn's disease lesions in wireless capsule endoscopy images using svm based classification. In *2013 International Conference on Signal Processing , Image Processing Pattern Recognition*, pages 303–307, Feb 2013.

[8] A. Eid, V. S. Charisis, L. J. Hadjileontiadis, and G. D. Sergiadis. A curvelet-based lacunarity approach for ulcer detection from wireless capsule endoscopy images. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 273–278, June 2013.

[9] Yingju Chen and Jeongkyu Lee. Ulcer detection in wireless capsule endoscopy video. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1181–1184, New York, NY, USA, 2012. ACM.

[10] L. Yu, P. C. Yuen, and J. Lai. Ulcer detection in wireless capsule endoscopy images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 45–48, Nov 2012.

[11] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of Visual Attention. jun 2014.

[12] Tomonori Aoki, Atsuo Yamada, Kazuharu Aoyama, Hiroaki Saito, Akiyoshi Tsuboi, Ayako Nakada, Ryota Niikura, Mitsuhiro Fujishiro, Shiro Oka, Soichiro Ishihara, Tomoki Matsuda, Shinji Tanaka, Kazuhiko Koike, and Tomohiro Tada. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointestinal Endoscopy*, 2018.

[13] Shanhui Fan, Lanmeng Xu, Yihong Fan, Kaihua Wei, and Lihua Li. Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. *Physics in Medicine & Biology*, 63(16) :165001, aug 2018.

[14] Vladimir I Iglovikov, Selim S Seferbekov, Alexander V Buslaev, and Alexey A Shvets. TernausNetV2 : Fully Convolutional Network for Instance Segmentation. Technical report.

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD : single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

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

[18] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 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*, pages 103–111. Association for Computational Linguistics, 2014.

[19] Radoslaw M Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of Deep Neural Networks to Spatio-temporal Cortical Dynamics of Human Visual Object Recognition reveals Hierarchical Correspondence. *Scientific Reports*, 2016.

[20] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sanchez, Bogdan J. Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjorn Rustad, Ilangko Balasingham, Konstantin Pogorelov, Sungbin Choi, Quentin Debard, Lena Maier-Hein, Stefanie Speidel, Danail Stoyanov, Patrick Brandao, Henry Córdova, Cristina Sánchez-Montes, Suryakanth R. Gurudu, Gloria Fernández-Esparrach, Xavier Dray, Jianming Liang, and Aymeric Histace. Comparative validation of polyp detection methods in video colonoscopy : Results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans. Med. Imaging*, 36(6) :1231–1249, 2017.