Apprentissage de relations floues pour l'annotation sémantique expliquée avec peu de données

Régis Pierrard, Jean-Philippe Poli, Céline Hudelot

Apprentissage de relations floues pour l'annotation sémantique expliquée avec peu de données

Régis Pierrard[1,2] Jean-Philippe Poli[1] Céline Hudelot[2]

[1] CEA, LIST, 91191 Gif-sur-Yvette cedex, France.
[2] Mathématiques et Informatique pour la Complexité et les Systèmes, CentraleSupélec, Université Paris-Saclay, 91190, Gif-sur-Yvette, France.

{regis.pierrard, jean-philippe.poli}@cea.fr, celine.hudelot@centralesupelec.fr

Résumé

Malgré son récent succès, l'apprentissage profond semble loin d'égaler certaines capacités humaines comme l'apprentissage à partir de peu d'exemples, le raisonnement ou encore l'explicabilité. Dans cet article, nous mettons l'accent sur l'annotation d'objets dans des images et nous présentons un cadre de raisonnement s'appuyant sur l'apprentissage de relations floues qui permet notamment de générer des explications. Étant donné un ensemble de relations, les plus pertinentes d'entre elles sont induites et combinées pour construire des contraintes permettant de définir un problème de satisfaction de contraintes floues. La résolution de ce probléme nous permet d'annoter l'image et de générer des explications à l'aide des contraintes utilisées. Nous testons notre approche sur un jeu de données public avec deux objectifs : effectuer une annotation d'organes et fournir une explication à chaque annotation. Nous montrons que le modèle que nous obtenons peut générer des explications tout en atteignant des performances élevées malgré un jeu de données pouvant être constitué de seulement deux exemples.

Mots Clef

Explicabilité, annotation, logique floue, peu de données.

Abstract

Despite the recent successes of deep learning, such models are still far from some human abilities like learning from few examples, reasoning and explaining their decisions. In this paper, we focus on object annotation in images and we introduce a reasoning framework that is based on learning fuzzy relations on a dataset for generating explanations. Given a catalogue of relations, it efficiently induces the most relevant relations and combines them for building constraints in order to both solve the target task and generate explanations. We test our approach on a publicly available dataset on which the goal was both to perform multiple organ annotation and to provide explanations. We show that our model can generate explanations and achieve high performance despite being trained on a small dataset containing as few as a couple of examples.

Keywords

Explainability, annotation, fuzzy logic, few-data learning.

1 Introduction

Au cours de ces dernières années, l'explicabilité est devenue un enjeu majeur pour les intelligences artificielles [12, 17]. Cela fait notamment écho à la popularité des réseaux de neurones profonds, qui peuvent atteindre des performances très élevées dans de nombreuses tâches mais ne sont pas adaptés à la génération d'explication [14, 27]. Être capable d'expliquer les résultats produits par les intelligences atificielles est utile non seulement pour comprendre leur raisonnement, mais aussi pour les rendre plus crédibles et dignes de confiance aux yeux des utilisateurs [34]. Dans des domaines en lien direct avec l'être humain tels que l'analyse d'images médicales [25] ou les voitures autonomes, on ne peut pas prendre de décisions en faisant une confiance aveugle à un modèle car les conséquences peuvent être désastreuses.

Plusieurs définitions d'*interprétabilité* et d'*explicabilité* ont été données dans la litérature [10, 16, 24, 29], mais aucun consensus ne se dégage et ces deux notions sont mêmes parfois utilisées de manière interchangeable. Globalement, il émerge que l'interprétabilité est la capacité à présenter des informations sur le fonctionnement du sytème en des termes intelligibles, alors que l'explicabilité est la capacité à décrire de manière précise et logique le fonctionnement d'un système. Dans cet article, nous nous focalisons sur l'explicabilité et donc sur la restitution du raisonnement effectué par notre modèle. Pour obtenir des explications, une première famille de méthodes consiste à apprendre une approximation locale et interprétable autour de la prédiction fournie par un modèle de type boîte noire [26, 34]. Ces approches peuvent traiter tout type de modèle et sont donc adaptées aux réseaux de neurones profonds. Cependant, bien qu'elles soient capables d'extraire les caractéristiques principales qui ont servi à générer une prédiction, elles ne peuvent pas exactement reconstituer le

raisonnement effectué par le modèle cible. La seconde famille d'approches repose sur l'utilisation de modèles qui sont propices à la génération d'explications, tels que les arbres de décisions, les règles de décision ou l'obtention d'un modèle explicable à partir d'un modèle de type boîte noire par distillation [18]. Leur avantage principal est qu'ils permettent de facilement reproduire le raisonnement qui a mené à un résultat donné, ce qui est très utile pour la génération d'explications. Cependant, ces modèles risquent de ne pas être aussi performants que des modèles de type boîte noire car l'explicabilité a souvent un coût. En effet, il existe un compromis entre performance et explicabilité [17]. Dans ce papier, nous proposons une approche basée sur cette seconde famille de méthodes qui compense ce compromis en n'apprenant que sur une faible quantité de données étiquetées dont l'acquisition est coûteuse. Cela repose notamment sur des études cognitives sur l'interprétation humaine des images : (1) l'importance des relations spatiales et contextuelles en reconnaissance d'objets et de scènes [1], (2) la capacité des humains à n'avoir besoin que de quelques exemples pour apprendre à accomplir une tâche [23, 37]. Ce second point est notamment abordé par les zero-shot [21], one-shot et few-shot learning [13]. Ils visent à effectuer cela en utilisant les précédents exemples dans la base d'apprentissage ou les représentations disponibles. Cependant, de telles méthodes nécessitent toujours d'avoir des informations entre les classes connues et les classes inconnues [39].

Notre objectif est de proposer une nouvelle approche qui permet d'apprendre à raisonner et de générer à la fois des annotations et des explications à partir d'un faible nombre d'exemples d'apprentisage. Elle s'appuie notamment sur l'utilisation de relations floues, qui permettent de prendre en compte à la fois des informations quantitatives et des informations qualitatives. Pour un exemple donné, le système cherche les régions de l'image qui satisfont au mieux les relations entre les objets d'intérêt. Nous modélisons cela sous la forme d'un problème de satisfaction de contraintes. Dans la section 3, nous décrivons la démarche globale qui consiste en trois étapes principales : l'évaluation des relations, l'extraction des relations les plus pertinentes et la génération des contraintes pour résoudre un problème de satisfaction de contraintes et produire des explications. Dans la section 4, l'approche proposée est évaluée sur un exemple d'annotation d'images médicales. Ceci est un bon exemple de raisonnement spatial car la reconnaissance des différents organes repose essentiellement sur leur disposition spatiale. De plus, l'utilisation d'un jeu de données médicales présente plusieurs défis, tels que sa faible taille et l'importance des explications dans ce domaine. Nous avons testé et comparé notre modèle à l'état de l'art et nous avons montré que notre approche est capable d'obtenir de bonnes performances et de générer des explications malgré un faible nombre d'exemples d'apprentissage.

2 Rappels

2.1 Logique floue

La logique floue et la théorie des ensembles flous ont été introduits par Zadeh [41]. Il s'agit d'une extension de la logique Booléenne qui permet notamment de gérer l'imprécision des données. Alors qu'une variable est soit vraie soit fausse en logique Booléenne, elle peut prendre n'importe quelle valeur entre 0 (fausse) et 1 (vraie) en logique floue. Cette gamme de valeurs, qu'on appelle degrés, sert à gérer l'imprécision. Dans un univers A, un ensemble flou F est défini par une fonction d'appartenance $\mu_F : A \rightarrow [0; 1]$. Cette fonction, intitulée *fonction d'appartenance de* F, représente dans quelle mesure chaque $a \in A$ appartient à F. Si F est un ensemble non-flou, alors $\mu_F(a)$ vaut soit 0 quand a n'appartient pas à F, soit 1 si a appartient à F.

Il est également possible d'exprimer des relations entre deux univers différents. Pour deux univers A et B, une relation floue dyadique $\mathcal{R}$ est un ensemble flou défini par une fonction d'appartenance $\mu_{\mathcal{R}} : A \times B \rightarrow [0, 1]$. Cette fonction attribue à chaque paire $(a, b) \in A \times B$ un degré d'appartenance à la relation $\mathcal{R}$. Les relations floues n-aire sont définies de la même façon. Un ensemble flou peut caractériser une expression du langage naturel lorsqu'il est associé à une description linguistique, ce qui est courant en logique floue.

2.2 Problème de satisfaction de contraintes floues

Un problème de satisfaction de contraintes est un problème dans lequel l'objectif est d'attribuer des valeurs à un ensemble de variables qui doivent vérifier un ensemble de contraintes. Les problèmes de planification en sont un exemple [30]. [11] en présente une extension au cadre de la logique floue afin de pouvoir gérer des paramètres et des contraintes flous lors de la résolution de problèmes de satisfaction de contraintes floues. Un tel problème est défini par un ensemble de variables X, un ensemble de domaines D et un ensemble de contraintes floues C. Par exemple, dans le cadre de l'annotation d'images, X correspond aux étiquettes que l'on souhaite attribuer aux objets de l'image, D représente les différentes régions à annoter et C est un ensemble de relations floues liant ces objets. Les problèmes de satisfaction de contraintes floues nous intéressent tout particulièrement car ils nous permettent d'effectuer l'annotation (en résolvant le problème) ainsi que de générer des explications à partir des contraintes utilisées.

Pour résoudre un tel problème, on utilise un algorithme de backtracking pour parcourir l'ensemble des solutions possibles. Avant chaque assignation d'une valeur à une variable, l'algorithme FAC-3 [11, 38] est utilisé pour maintenir la cohérence de l'ensemble des valeurs possibles. Cette opération de propagation des contraintes permet de réduire la taille de l'espace de recherche. À la fin, la solution retenue est celle qui vérifie au mieux les contraintes.

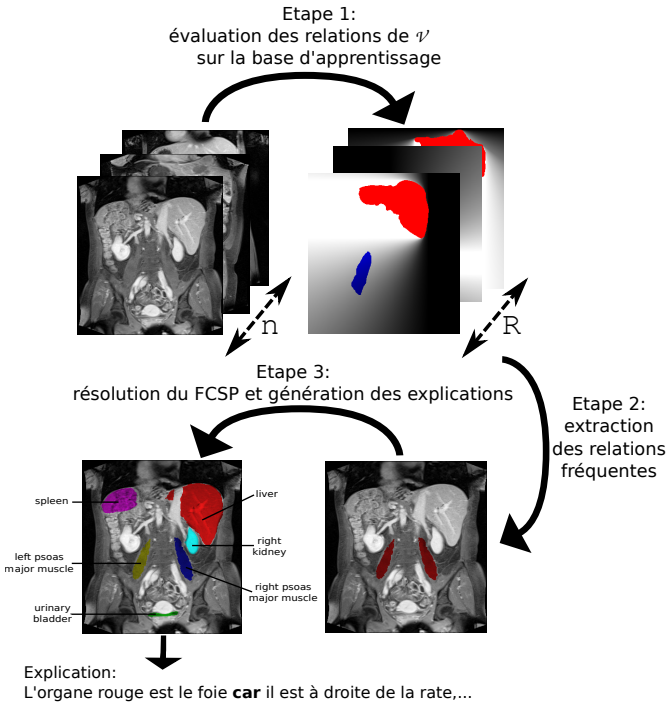


FIGURE 1 – Représentation de notre approche qui est constituée de trois étapes. Dans l'étape 1, les n_r relations floues du vocabulaire $\mathcal{V}$ sont évaluées sur le jeu de données $\mathcal{T}$ qui est composé de n images. Ensuite, les relations les plus fréquentes sont extraites sur la base d'apprentissage $\mathcal{T}_{\text{app}}$ et elles sont utilisées pour déterminer les contraintes de notre problème. Enfin, un problème de satisfaction de contraintes floues est défini et résolu afin d'étiqueter les objets cibles pour chaque image de $\mathcal{T}_{\text{test}}$. Une explication s'appuyant sur les contraintes utilisées est fournie pour chaque labellisation d'objet.

3 Approche proposée

Dans cette partie, nous décrivons notre approche dont l'objectif est d'annoter les régions d'intérêt et de fournir une explication pour chaque annotation. Elle se divise en trois étapes : l'évaluation des relations entre les régions d'intérêt (Sec. 3.1), l'apprentissage des relations les plus pertinentes (Sec. 3.2) et la résolution d'un FCSP pour trouver les régions qui satisfont au mieux les relations apprises tout en générant des explications à partir des contraintes (Sec. 3.3). L'approche globale est représentée sur la Figure 1.

3.1 Étape 1 : évaluation des relations

L'objectif de cette première étape est d'évaluer différentes relations entre les régions d'intérêt (les organes par exemple) afin de pouvoir plus tard (dans l'étape suivante) déterminer celles qui sont les plus pertinentes.

Soit un ensemble d'apprentissage $\mathcal{T}_{\text{train}}$ composé de n images $\{\mathbf{i}_1, \ldots, \mathbf{i}_n\}$ et un ensemble d'étiquettes $\mathcal{Y}$ composé de N éléments $\{y_1, \ldots, y_N\}$ tels que chaque image $\mathbf{i} \in \mathcal{T}_{\text{train}}$ est divisée en un ensemble $\mathcal{O}$ de N régions d'intérêt $\mathcal{O} = \{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\}$ qui sont associées aux étiquettes par la fonction suivante :

$$\begin{aligned} f : \mathcal{Y} &\to \mathcal{O} \\ y_i &\mapsto o_j \end{aligned} \tag{1}$$

Soit $\mathcal{V} = \{\mathcal{R}_1, \ldots, \mathcal{R}_R\}$ un ensemble de R relations. Nous appelons cet ensemble un *vocabulaire*. Il est défini par un expert et contient des relations qui sont potentiellement d'intérêt. Par exemple, il peut s'agir de relations directionnelles, comme *à gauche de*, ou de relations de distance, comme *près de*. Un vocabulaire plus riche devrait rendre le système plus expressif et améliorer les annotations et explications produites. Les relations de notre vocabulaire $\mathcal{V}$ sont automatiquement évaluées sur les régions d'intérêt de chaque image de $\mathcal{T}_{\text{train}}$. La manière dont elles sont évaluées dépend de leur définition (cf. Sec. 4.2).

Pour chaque relation $\mathcal{R} \in \mathcal{V}$, nous appelons $\alpha(\mathcal{R})$ son arité. $\mathcal{R}$ est évaluée pour tous les $\alpha(\mathcal{R})$-uplet de régions d'intérêt possibles. Il est important de distinguer le vocabulaire $\mathcal{V}$ de l'ensemble $\mathcal{V}_{\mathcal{Y}}$ des relations à évaluer entre les régions d'intérêt. Par exemple, pour une relation binaire $\mathcal{R} \in \mathcal{V}$, il faut évaluer l'ensemble $\{\mathcal{R}\big(f(y), f(y')\big) \mid y, y' \in \mathcal{Y}, y \neq y'\}$. Ainsi, le nombre d'évaluations à effectuer est :

$$\text{Card}(\mathcal{V}_{\mathcal{Y}}) = \sum_{j=1}^{R} \frac{N!}{(N - \alpha(\mathcal{R}_j))!} \tag{2}$$

À la fin de cette étape, nous avons donc une base de données $\mathcal{B} : \mathcal{T}_{\text{train}} \times \mathcal{V}_{\mathcal{Y}} \to [0; 1]$ de taille $n \times \text{Card}(\mathcal{V}_{\mathcal{Y}})$.

3.2 Étape 2 : apprentissage des relations pertinentes

Lors de cette étape, l'objectif est d'extraire les relations les plus pertinentes parmi celles évaluées à l'étape précédente. Pour une étiquette $y \in \mathcal{Y}$, nous faisons le postulat que les relations pertinentes impliquant les régions étiquetées comme y sont les plus fréquentes car elles devraient être vérifiées par la plupart, si ce n'est tous, des exemples associés à cette classe. Ainsi, l'apprentissage des relations les plus pertinentes revient à extraire les relations les plus fréquentes. Cette opération est effectuée classe par classe, il s'agit d'une stratégie *one vs all*. En effet, les relations pertinentes pour une classe d'objet ne le sont pas forcément pour une autre classe. Au sein d'une même classe, les exemples devraient être fortement corréllés entre eux. C'est pourquoi nous utilisons un algorithme d'extraction de motifs fréquents qui tire profit de cette propriété [32].

Un sous-ensemble J de relations est un ensemble appartenant à $2^{\mathcal{V}_{\mathcal{Y}}}$. Soit $J = \{r_1, \ldots, r_p \mid r_j \in \mathcal{V}_{\mathcal{Y}}, 1 \leq j \leq p \leq \text{Card}(\mathcal{V}_{\mathcal{Y}})\}$. La fréquence de J dans la base de données B s'appelle le *support* et est défini par :

$$\text{support}(J) = \frac{1}{n} \sum_{k=1}^{n} \min\big(\mathcal{B}(\mathbf{i}_k, r_1), ..., \mathcal{B}(\mathbf{i}_k, r_p)\big) \tag{3}$$

Étant donné un seuil, J est dit *fréquent* si, et seulement si, son support est supérieur à ce seuil. L'algorithme que nous

utilisons s'appuie sur un opérateur de fermeture que nous définissons comme $h : 2^{\mathcal{V}_\mathcal{Y}} \rightarrow 2^{\mathcal{V}_\mathcal{Y}}$. Une première phase consiste à déterminer l'ensemble des sous-ensembles de relations vérifiant $h(J) = J$. De tels sous-ensembles de relations sont dits *fermés*. Ensuite, parmi les ensembles fermés, seuls ceux qui sont fréquents sont conservés. Enfin, l'ensemble des sous-ensembles fermés et fréquents de relations permet d'engendrer l'ensemble des sous-ensembles fréquents de relations.

Puisque nous suivons une approche one-vs-all, chaque classe d'objets y a son propre seuil de fréquence S_y dont la valeur est déterminée lors d'une phase de validation. La valeur de ce seuil a une influence directe sur le nombre d'ensembles de relations fréquents qui sont extraits. Si elle est trop élevée, alors pas ou peu d'ensembles de relations seront considérés comme fréquents. On risque alors de ne pas pouvoir discriminer les différentes classes et on se retrouve donc dans une situation d'*underfitting*. Au contraire, si la valeur du seuil est trop basse, alors des ensembles de relations non pertinents seront appris. Ceci mènerait à une situation d'*overfitting*.

À la fin de cette étape, pour chaque étiquette $y \in \mathcal{Y}$, nous avons un ensemble F_y de sous-ensembles fréquents de relations tel que $F_y \subseteq 2^{2^{\mathcal{V}_\mathcal{Y}}}$.

3.3 Étape 3 : résolution du problème de satisfaction de contraintes floues et génération des explications

Pour un exemple de test **i**, on peut obtenir par segmentation un ensemble de régions potentiellement d'intérêt. L'objectif de cette étape est de déterminer les étiquettes des régions qui satisfont au mieux les relations apprises à létape précédente. Cela peut être modélisé par un problème de satisfaction de contraintes floues. De plus, puisque les relations sont associées à une description linguistique, on peut générer une explication pour chaque annotation.

Pour chaque étiquette $y \in \mathcal{Y}$, nous avons obtenu à l'issue de l'étape précédente un ensemble F_y. Soit $F_y^{\max}$ un ensemble défini de la façon suivante :

$$F_y^{\max} = \{z \in F_y \mid \text{Card}(z) = \max_{v \in F_y} \big(\text{Card}(v)\big)\} \quad (4)$$

$F_y^{\max}$ correspond donc à l'ensemble des sous-ensembles fréquents de relations de taille maximale. Chaque relation $\mathcal{R}\big(f(y), f(y')\big)$ présente dans les sous-ensembles de relations de $\bigcup_{y \in \mathcal{Y}} F_y^{\max}$ est directement convertie en une contrainte floue $c_\mathcal{R}\big(f(y), f(y')\big)$. Nous pouvons donc désormais construire un modèle défini par les contraintes qui ont été apprises ainsi que par les valeurs des seuils de fréquence. Reposant sur l'apprentissage de symboles (les relations) et n'ayant pas recours à la minimisation d'une fonction de perte, notre modèle est adapté à des jeux de données de faible taille.

Après segmentation, l'exemple de test **i** est divisé en K régions potentiellement d'intérêt $\{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\}$. Le problème de satisfaction de contraintes floues que nous cherchons à résoudre est donc le suivant :

$$X = \{x_y \mid y \in \mathcal{Y}\} \quad (5)$$

$$D = \{D_y \mid y \in \mathcal{Y} \text{ and } D_y = \{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\}\} \quad (6)$$

$$C = \{c_\mathcal{R}(x_y, x_{y'}) \mid \mathcal{R}\big(f(y), f(y')\big) \in U \text{ such as } U \subseteq \bigcup_{y \in \mathcal{Y}} F_y^{\max}\} \quad (7)$$

Chaque élément D_y de D est l'ensemble de toutes les régions qui pourraient être associées à la variable x_y. Au début, tous les D_y sont identiques. Ils sont ensuite modifiés par l'algorithme FAC-3 qui écarte les solutions incohérentes. Chaque contrainte de C ayant été évaluée, le problème peut alors être résolu et les étiquettes correspondant à chaque variable determinées. Ensuite, pour chaque variable $x_y \in X$, une explication est générée en utilisant les contraintes obtenues à partir de $F_y^{\max}$. Par exemple, la contrainte $c_{\mathcal{R}_{\text{à gauche de}}}(x_W, x_Z)$ engendre "W est à gauche de Z". Ainsi, l'utilisation de contraintes générées à partir de $F_y^{\max}$ permet d'exprimer une explication sous la forme : "classe y CAR cause$_1$,...,cause$_n$". Pour une classe d'objets donnée, la contrainte la moins satisfaite nous donne un facteur de certitude pour modérer l'explication [5], e.g. "L'annotation de cet organe est *probablement* le foie...". Les contraintes apprises et ce facteur de certitude sont envoyés à un *surface realiser* comme simpleNLG [15] afin de les aggréger en une phrase syntaxiquement correcte.

L'évaluation de la qualité d'une explication est une tâche délicate. Une bonne explication est une explication cohérente [29], simple et générique [33], utile et pertinente [28]. Ces différents critères sont difficiles à évaluer de manière automatique et ils dépendent de la connaissance et des attentes de l'utilisateur. Trois types de solution sont proposés dans [10] : faire évaluer les explications par un expert, poser plusieurs questions simples à un groupe de personnes n'étant pas expert dans le domaine ciblé ou utiliser un modèle proxy dont l'explicabilité a déjà été évaluée pour juger l'explicabilité du modèle en cours d'évaluation.

4 Expériences

Dans cette partie, nous décrivons les expériences qui ont été menées sur un jeu d'images médicales. L'objectif est d'effectuer une annotation d'organes justifiée en apprenant un modèle à partir d'un faible nombre de données. Tandis que la détection d'organes est un sujet qui a été régulièrement abordé dans la litérature [36, 9, 31, 22], l'annotation d'organes n'a été traitée que dans [40]. Cette dernière méthode consiste à trouver dans le jeu de données des images qui ont des caractéristiques visuelles similaires à l'image étudiée et à l'étiqueter de la même manière que celles-ci. [22] aborde le problème de la détection d'organes sur des coupes abdominales. Leur méthode s'appuie sur des règles spatiales floues. Cependant, ces règles sont trop spécifiques et ne sont pas adaptées à d'autres jeux de données.

(a) CTwb (b) CTce (c) MRwb (d) MRce

FIGURE 2 – Exemples correspondant aux quatre types d'images composant le jeu de données utilisé.

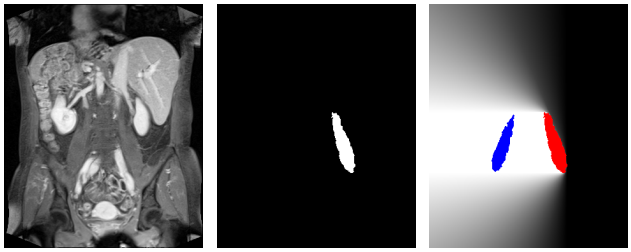
La classification d'organes a été traitée par [35] à l'aide d'un réseau de neurones convolutionnel. Leur méthode repose également sur des techniques d'augmentation de données permettant de répondre au problème de la faible taille du jeu de données utilisé.

4.1 Jeu de données

Partant du constat qu'il n'existe actuellement pas de jeu de données dédié à l'évaluation de l'explicabilité des modèles, nous avons utilisé un jeu de données initialement créé pour évaluer des méthodes de segmentation supervisée. Il s'intitule *Anatomy3* et a été présenté dans [19]. Il contient 391 images provenant de deux modalités différentes (CT et IRM) ainsi que les segments correspondant à certains organes de ces images. Certaines images représentent le corps entier (CTwb et MRwb) et les autres l'abdomen (CTce et MRce). Ce sont des images à trois dimensions obtenues par superposition de tranches en deux dimensions. Puisque nous travaillons sur des données à deux dimensions, nous avons sélectionné une tranche 2D par patient afin de construire un jeu d'images à deux dimensions. La figure 2 montre quatre tranches, chacune correspondant à un des quatre types d'images (CTwb, MRwb, CTce et MRce).
L'ensemble $\mathcal{Y}$ d'organes auxquel nous nous intéressons est composé du *foie*, de la *rate*, de la *vessie*, des *reins gauche* et *droit*, des *poumons gauche* et *droit* et des *muscles psoas gauche* et *droit*. Nous avons conservé toutes les images contenant ces neuf organes et leurs segments correspondants, ce qui nous donne un jeu de 35 images et 315 segments.

4.2 Paramètres expérimentaux

Apprentissage du modèle. Le modèle que nous construisons repose sur l'extraction des sous-ensembles de relations fréquents. Il y a autant d'hyperparamètres que d'organes. Ils correspondent tous au seuil utilisé pour juger la fréquence d'une relation. Afin d'optimiser la recherche de ces hyperparamètres, nous avons utilisé une validation croisée imbriquée [6] : (1) une validation croisée externe est réalisée pour obtenir un ensemble d'apprentissage et un ensemble de test à chaque itération, (2) une validation croisée interne est menée sur chaque base d'apprentissage



(a) Image d'origine (b) Organe segmenté (c) Relation

FIGURE 3 – (Affichage optimal en couleur) Exemple d'évaluation d'une relation. L'objectif est de déterminer la relation *l'organe bleu à gauche de l'organe rouge*. Étant donné une image en entrée (3a), un organe segmenté (3b) est considéré comme l'objet de référence. Cet organe est utilisé pour générer un paysage flou (3c) qui représente le degré d'appartenance de chaque pixel à la relation *à gauche de l'organe rouge*. La relation est déterminée en évaluant le degré d'intersection entre le paysage flou et l'organe bleu.

de la validation croisée externe pour obtenir un ensemble d'apprentissage interne et un ensemble de validation permettant d'ajuster la valeur des hyperparamètres. Cela permet d'obtenir une estimation de l'erreur non-biaisée par la construction de l'ensemble de validation.
Dans la boucle interne, l'optimisation des hyperparamètres est effectuée par optimisation bayesienne sur 20 itérations avec un processus Gaussien comme a priori. La fonction d'acquisition utilisée ici est l'*expected improvement* [20].

Relations. De nombreuses relations spatiales floues ont été étudiées dans la litérature [4]. Dans nos expériences, nous utilisons des relations directionnelles, de distance et de symétrie. Les relations directionnelles et de distance [2, 3] sont obtenues en générant un paysage flou et en utilisant une approche de type reconnaissance de forme floue [7]. Comme on peut le voir sur la figure 3, le paysage flou est obtenu par la dilatation floue d'un objet de référence par un élément structurant dont la définition détermine le type de relation à évaluer. Soit S l'espace de l'image. Soit A un objet de référence dans S et $\mu_{A,\mathcal{R}}$ la fonction d'appartenance associée au paysage flou représentant la relation $\mathcal{R}$ dont l'objet de référence est A. Soit μ_B la fonction d'appartenance associée à un objet B dans S. La relation $\mathcal{R}$ entre A et B est alors obtenue par évaluation du degré d'intersection flou μ_{int} [4] entre $\mu_{A,\mathcal{R}}$ et μ_B :

$$\mu_{int}(\mu_{A,\mathcal{R}}, \mu_B) = \frac{\sum_{x \in S} \min\Big(\mu_{A,\mathcal{R}}(x), \mu_B(x)\Big)}{\min\Big(\sum_{x \in S} \mu_{A,\mathcal{R}}(x), \sum_{x \in S} \mu_B(x)\Big)} \quad (8)$$

Sur la figure 3, la relation $\mathcal{R}$ est *à gauche de*, l'objet de référence A est l'organe rouge et l'objet B est l'organe bleu. Afin d'obtenir un vocabulaire fini, nous n'avons utilisé que certaines valeurs pour les paramètres de ces relations afin de n'exprimer que des relations telles que *au dessus de* ou

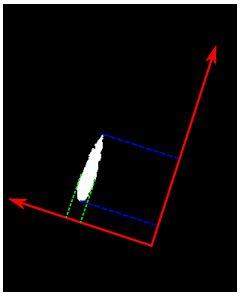


FIGURE 4 – (Affichage optimal en couleur) Illustration décrivant l'évaluation de la propriété *allongé*. Une analyse en composantes principales est effectuée afin d'obtenir les deux axes principaux. Ensuite, l'organe étudié est projeté sur ces deux axes. Soit p_g la prjection verte (la plus petite) et p_b la projection bleue (la plus grande). Le degré d'appartenance est déterminé par $1 - \frac{p_g}{p_b}$.

près de.
La relation de symétrie axiale [8] que nous utilisons repose sur la détermination de l'axe qui maximise une mesure de symétrie entre deux organes. Puisque cette mesure n'est pas différentiable, ce problème d'optimisation est résolu à l'aide d'une méthode telle que le *downhill simplex*.
Nous utilisons également une propriété. Elle peut être vue comme une relation unaire puisqu'elle ne caractérise qu'un seul organe. Cette propriété évalue à quel point la forme d'un organe est allongée. Étant donné un organe segmenté, une analyse en composantes principales est réalisée afin d'en obtenir les deux axes principaux. Ensuite, comme le décrit la figure 4, l'organe est projeté sur chacun de ces deux axes. Le rapport entre ces deux projections est alors utilisé pour déterminer le degré d'appartenance à cette propriété. Cependant, cette définition ne permet pas de bien juger l'allongement des objects concaves.
Notre vocabulaire de relations $\mathcal{V}$ contient les relations suivantes : *à gauche de*, *à droite de*, *en dessous de*, *au-dessus de*, *près de*, *symétrique à* et *est allongé*. $\mathcal{V}$ comporte donc six relations binaires et une relation unaire. Puisque $\mathrm{Card}(\mathcal{Y}) = 9$, le nombre de relations à évaluer pour chaque image est $\mathrm{Card}(\mathcal{V}_\mathcal{Y}) = 441$. Cela contribue à rendre notre modèle expressif, mais le temps de calcul nécessaire à l'évaluation de toutes ces relations pour chaque image peut être important.

4.3 Initialisation du problème

Comme nous l'avons montré dans la section 3, la méthode complète est composée de trois étapes. Notre jeu de données contient pour chaque image les segments associés à chaque organe. Ces segments ne sont pas flous, mais la méthode resterait inchangée s'ils l'étaient.
L'objectif intermédiaire est de générer des contraintes afin de définir un problème de satisfaction de contraintes floues pour chaque exemple à évaluer. Une fois ce problème résolu, nous obtenons les segments correspondant à chaque organe et les contraintes peuvent désormais être utilisées pour générer des explications.

Organe	Valeur du seuil
Foie	0.96
Rate	0.86
Vessie	0.80
Rein droit	0.92
Rein gauche	0.89
Poumon droit	0.98
Poumon gauche	0.97
Muscle psoas droit	0.92
Muscle psoas gauche	0.88

TABLE 1 – Valeurs des seuils correspondant à chaque organe. Ce sont les hyperparamètres de notre modèle.

L'ensemble de variables X et l'ensemble de domaines D de notre problème de satisfaction de contraintes sont les suivants :

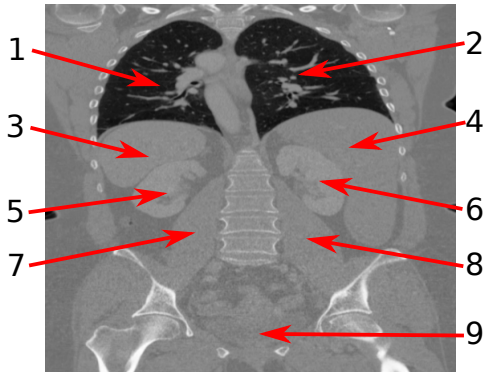
$$X = \{x_{\text{foie}}, x_{\text{rate}}, x_{\text{vessie}}, x_{\text{rein\_d}}, x_{\text{rein\_g}}, x_{\text{poumon\_d}}, x_{\text{poumon\_g}}, x_{\text{psoas\_d}}, x_{\text{psoas\_g}}\}$$

$$D = \{D_{\text{foie}}, D_{\text{rate}}, D_{\text{vessie}}, D_{\text{rein\_d}}, D_{\text{rein\_g}}, D_{\text{poumon\_d}}, D_{\text{poumon\_g}}, D_{\text{psoas\_d}}, D_{\text{psoas\_g}}\}$$

avec D_i l'ensemble des segments possibles pour la variable x_i. Il s'agit des segments fournis dans le jeu de données. Pour chaque organe $y \in \mathcal{Y}$, les contraintes floues sont générées à partir de l'ensemble des sous-ensembles fréquents de relations de taille maximale $F_y^{\max}$. De plus, comme chaque organe est unique, il ne peut pas y avoir deux annotations identiques sur la même image. Il faut donc compléter C avec des contraintes imposant la différence entre chaque variable deux à deux. C'est pourquoi nous ajoutons la contrainte globale *AllDifferent*.
La définition du problème de satisfaction de contraintes floues est donc effectuée automatiquement. Ensuite, pour un exemple donné, le problème est résolu comme décrit dans la section 2.2.

4.4 Résultats

Un exemple est présenté sur la figure 5. Il y a neuf organes à annoter et donc neuf explications à fournir.
La mesure d'évaluation de notre modèle est la *précision*. Il s'agit du rapport entre le nombre d'annotations correctes pour tous les organes et le nombre total d'annotations effectuées. Nous obtenons une précision de 100% avec un modèle ne contenant que des relations directionnelles. La validation croisée externe contient trois plis (23/24 exemples dans la base d'apprentissage et 12/11 dans la base de test à chaque itération). La validation croisée interne comporte quatre plis. Comme il y a neuf organes à annoter, il y a neuf hyperparamètres à déterminer pour extraire les ensembles de relations fréquents. Les valeurs



L'annotation de l'organe 1 est très probablement le poumon gauche **car** il est *à gauche du* poumon droit (organe 2), il est *symétrique au* poumon droit et il est *au-dessus de* la rate (3).
L'annotation de l'organe 9 est probablement la vessie **car** elle est *allongée*, elle est *en dessous du* rein droit (6) et *en dessous du* rein gauche (5).
L'annotation de l'organe 4 est le foie **car**...

FIGURE 5 – Exemple d'annotations justifiées.

obtenues pour ces hyperparamètres sont affichées dans la table 1. Des contraintes pourraient être ajoutées sur les valeurs que peuvent prendre ces hyperparamètres pour spécifier la longueur des explications.

Nous observons que les explications reprennent bien les relations qui ont été extraites puis traduites en contraintes. Pour l'exemple proposé sur la figure 5, l'ensemble des contraintes associées au rein droit est :
$C_{rein\_d} = \{(x_{rein\_d}, x_{rein\_g}, R_{\text{symétrique à}}),$
$(x_{poumon\_d}, x_{rein\_d}, R_{\text{au-dessus de}}), (x_{rein\_d}, x_{foie}, R_{\text{à gauche de}}),$
$(x_{vessie}, x_{rein\_d}, R_{\text{en dessous de}}), (x_{rein\_d}, x_{rein\_g}, R_{\text{à droite de}}),$
$(x_{rein\_g}, x_{rein\_d}, R_{\text{à gauche de}})\}$. Certaines de ces contraintes peuvent sembler redondantes, comme les deux dernières par exemple. Ce genre de situation peut arriver car le résultat d'une dilatation floue dépend de la forme de l'objet de référence. Comme deux organes différents ne peuvent pas avoir exactement la même forme, il peut y avoir de légères différences entre deux contraintes réciproques. Chaque organe est lié à un tel ensemble de contraintes. C est l'union de tous ces ensembles.

Bien que le modèle entraîné avec uniquement des relations directionnelles n'ait pas besoin d'autres types de relation pour obtenir une précision de 100%, ajouter de nouvelles relations peut permettre de construire de meilleures explications. Comme on peut le voir sur la figure 5, la relation *symétrique à* est utilisée dans l'explication de l'annotation du poumon gauche. Cette relation n'est pas nécessaire puisque nous atteignons 100% de précision sans elle, mais elle rend l'explication plus convaincante.

Nous avons aussi étudié le nombre d'exemples nécessaires dans la base d'apprentissage pour obtenir de bonnes performances. Pour une base d'apprentissage comportant uniquement deux images (et 33 exemples dans la base de test, par validation croisée imbriquée), nous obtenons au moins 99% de précision. Dans le cas extrême d'une base d'apprentissage contenant un seul exemple, les relations fréquentes seront celles dont le degré d'appartenance est plus grand que les seuils considérés (cf. section 3.2). Ainsi, tout exemple qui n'est pas une aberration devrait permettre d'obtenir de bonnes performances.

Enfin, nous observons que notre modèle obtient de meilleures performances que le réseau de neurones convolutionnel présenté dans [35]. Ce modèle a été entraîné sur un jeu de données plus grand que le nôtre et ne fournit pas d'explications. La méthode la plus proche de la nôtre, présentée dans [40], ne donne pas de mesure de précision comme référence. Son désavantage est qu'elle peut ne pas assigner certaines classes d'objet, ce qui arrive en moyenne une fois tous les cinq exemples. Avec notre approche, ce genre de situation est impossible car chaque variable de notre problème doit être associée à une région de l'image.

5 Conclusion et perspectives

Dans cet article, nous présentons une nouvelle méthode d'apprentissage et de raisonnement visuel dont le but est d'annoter des objets dans des images et d'expliquer les résultats obtenus. Cette tâche est formalisée comme un problème de satisfaction de contraintes floues. Cette méthode est basée sur des relations spatiales floues qui sont apprises sur un ensemble d'objets annotés et qui sont ensuite traduites en contraintes. Nous avons validé notre approche sur un jeu d'images médicales et nous avons montré qu'elle tire profit de l'apprentissage symbolique et de sa capacité à raisonner pour expliquer les résultats qu'elle produit et atteindre une précision de 99% pour une base d'apprentissage de taille 2.

Nous envisageons de développer une stratégie pour réaliser la première étape de manière plus rapide. Une première idée est de déterminer une structure hiérarchique des relations spatiales afin de pouvoir utiliser un tri topologique. Une deuxième idée est de mettre à jour l'ordre d'évaluation des relations à partir des exemples précédemment étudiés. De plus, puisque l'utilisation de la logique floue nous permet de gérer des segments imprécis, nous souhaitons utiliser une méthode de segmentation non-supervisée sur les images de tests pour initialiser les domaines de D.

Plus globalement, il s'agit d'une première étape vers une combinaison de l'apprentissage statistique et de l'apprentissage symbolique afin de créer une intelligence artificielle explicable.

Références

[1] I. Biederman. *On the Semantics of a Glance at a Scene*. 1981.

[2] I. Bloch. Fuzzy relative position between objects in image processing : a morphological approach. *IEEE transactions on pattern analysis and machine intelligence*, 21(7) :657–664, 1999.

[3] I. Bloch. On fuzzy distances and their use in image processing under imprecision. *Pattern Recognition*, 32(11) :1873–1895, 1999.

[4] I. Bloch. Fuzzy spatial relationships for image processing and interpretation : a review. *Image and Vision Computing*, 23(2) :89–110, 2005.

[5] D. V. Budescu, H-H. Por, and S. B. Broomell. Effective communication of uncertainty in the ipcc reports. *Climatic Change*, 113(2) :181–200, Jul 2012.

[6] G. C. Cawley and N. L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul) :2079–2107, 2010.

[7] M. Cayrol, H. Farreny, and H. Prade. Fuzzy pattern matching. *Kybernetes*, 11(2) :103–116, 1982.

[8] O. Colliot, A. V. Tuzikov, R. M. Cesar, and I. Bloch. Approximate reflectional symmetries of fuzzy objects with an application in model-based object recognition. *Fuzzy Sets and Systems*, 147(1) :141 – 163, 2004.

[9] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. Regression forests for efficient anatomy detection and localization in ct studies. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pages 106–117, 2011.

[10] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. In *eprint arXiv :1702.08608*, 2017.

[11] D. Dubois, H. Fargier, and H. Prade. Possibility theory in constraint satisfaction problems : Handling priority, preference and uncertainty. *Applied Intelligence*, 6(4) :287–309, 1996.

[12] European Council. The general data protection regulation, 2016.

[13] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4) :594–611, April 2006.

[14] M. Garnelo and M. Shanahan. Reconciling deep learning with symbolic artificial intelligence : representing objects and relations. *Current Opinion in Behavioral Sciences*, 29 :17 – 23, 2019.

[15] A. Gatt and E. Reiter. Simplenlg : A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90–93, 2009.

[16] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining Explanations : An Approach to Evaluating Interpretability of Machine Learning. *ArXiv e-prints*, 2018.

[17] D. Gunning. Explainable artificial intelligence (xai). 2017.

[18] G. Hinton and N. Frosst. Distilling a neural network into a soft decision tree. 2017.

[19] O. Jimenez-del Toro, H. Müller, M. Krenn, K. Gruenberg, A. A. Taha, M. Winterstein, I. Eggel, A. Foncubierta-Rodríguez, O. Goksel, A. Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms : Visceral anatomy benchmarks. *IEEE transactions on medical imaging*, 35(11) :2459–2475, 2016.

[20] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4) :455–492, 1998.

[21] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, pages 646–651, 2008.

[22] C-C. Lee, P-C. Chung, and H-M. Tsai. Identifying multiple abdominal organs from ct image series using a multimodule contextual neural network and spatial fuzzy rules. *IEEE Transactions on Information Technology in Biomedicine*, 7(3) :208–217, Sep. 2003.

[23] F-F. Li, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14) :9596–9601, 2002.

[24] Z.C. Lipton. The mythos of model interpretability. *Queue*, 16(3) :30 :31–30 :57, June 2018.

[25] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. Adiyoso Setio, F. Ciompi, M. Ghafoorian, J. A.W.M. van der Laak, B. van Ginneken, and C. I. SÃąnchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42 :60 – 88, 2017.

[26] S.M. Lundberg and S-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017.

[27] G. Marcus. Deep learning : A critical appraisal. *CoRR*, abs/1801.00631, 2018.

[28] J. McClure. Goal-based explanations of actions and outcomes. *European review of social psychology*, 12(1) :201–235, 2002.

[29] T. Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artifical Intelligence*, 267 :1–38, February 2019.

[30] S. Minton, M. D. Johnston, A. B. Philips, and P. Laird. Minimizing conflicts : a heuristic repair method for constraint satisfaction and scheduling problems. *Artificial Intelligence*, 58(1) :161 – 205, 1992.

[31] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A.M. Möller, S. Nekolla, and N. Navab. Fast multiple organ detection and localization in whole-body mr dixon sequences. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, pages 239–247, 2011.

[32] R. Pierrard, J-P. Poli, and C. Hudelot. A fuzzy close algorithm for mining fuzzy association rules. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, pages 88–99, 2018.

[33] S. J. Read and A. Marcus-Newhall. Explanatory coherence in social explanations : A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3) :429, 1993.

[34] M.T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you ? : Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[35] H.R. Roth, C.T. Lee, H-C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, and R.M. Summers. Anatomy-specific classification of medical images using deep convolutional nets. *arXiv preprint arXiv :1504.04003*, 2015.

[36] H. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8) :1930–1943, Aug 2013.

[37] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582) :520, 1996.

[38] M. C. Vanegas, I. Bloch, and J. Inglada. Fuzzy constraint satisfaction problem for model-based image interpretation. *Fuzzy Sets and Systems*, 286 :1 – 29, 2016.

[39] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot Learning – a Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[40] Z. Xue, S. Antani, L. R. Long, and G. R. Thoma. Automatic multi-label annotation of abdominal ct images using cbir. In *Medical Imaging 2017 : Imaging Informatics for Healthcare, Research, and Applications*, volume 10138, page 1013807, 2017.

[41] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3) :338 – 353, 1965.