HAL
open science

A New Approach for Explainable Multiple Organ Annotation with Few Data

Régis Pierrard, Jean-Philippe Poli, Céline Hudelot

To cite this version:

Régis Pierrard, Jean-Philippe Poli, Céline Hudelot. A New Approach for Explainable Multiple Organ Annotation with Few Data. IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI), Aug 2019, Macao, Macau SAR China. hal-02418480

HAL Id: hal-02418480

https://hal.science/hal-02418480v1

Submitted on 30 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A New Approach for Explainable Multiple Organ Annotation with Few Data

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

[1]CEA, LIST, 91191 Gif-sur-Yvette cedex, France.
[2]Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, 91190, Gif-sur-Yvette, France.
{*regis.pierrard, jean-philippe.poli*}*@cea.fr, celine.hudelot@centralesupelec.fr*

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 decisions. In this paper, we focus on organ annotation in medical images and we introduce a reasoning framework that is based on learning fuzzy relations on a small 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 organ annotation task and generate explanations. We test our approach on a publicly available dataset of medical images where several organs are already segmented. A demonstration of our model is proposed with an example of explained annotations. It was trained on a small training set containing as few as a couple of examples.

1 Introduction

In the last few years, explaining outputs returned by Artificial Intelligence (AI) algorithms has become more and more important [14, 20]. This echoes the dominance of deep neural networks, which reach very high performance in several visual recognition tasks but lack of explainability [29, 16]. Explaining decisions returned by intelligent systems is not only helpful for understanding their reasoning process, it is also essential for gaining acceptance and becoming trustworthy to humans [34]. In human-centered fields like medical image analysis [27], decisions cannot be made relying blindly on a model since the consequences could be disastrous.

While several definitions of interpretability and explainability exist in the literature [30, 18, 26, 12], there is no consensus among them and these two notions are sometimes used interchangeably. Overall, it emerges that interpretability is the ability to present insight into how a system works in understandable terms, whereas explainability is the ability to describe how a system works in an accurate and logical way. In this paper, we focus on rendering the reasoning process of our model to explain its decisions. To get explanations, a first family of methods consists in learning a local interpretable approximation model around the prediction returned by a black-box model [28, 34]. Those approaches can deal with any model, so they are well-suited for deep neural networks. However, although they aim at extracting key characteristics that led to the output, they cannot exactly replicate the reasoning the black-box model performed. The second possibility is to use models that are propitious for generating explanations, such as decision trees, decision rules or by distilling an unexplainable model into an explainable one [21]. Their main advantage is that the reasoning leading to a specific output is easy to track, so it can be used for generating an explanation. However, those models may not be as effective as black-box models, since explainability usually comes at a cost. Indeed, there is a well-known trade-off between accuracy and explainabil-

ity [20]. In this paper, we propose to rely on this second family of approaches by counterbalancing this trade-off with very little need for labelled data whose acquisition is costly. Our approach is based on two conclusions from human image interpretation studies: (1) the importance of contextual and spatial relations in object and scene recognition [2], and (2) the ability of humans to learn from few examples [37, 25]. Several approaches focus on few data learning [23, 15] but they need side information. We propose to mix statistical and symbolic learning to train a model that learns to manipulate spatial relations from few examples.

Our goal is to build a novel approach that can learn to reason and generate both annotations and explanations from just few examples. In our experiments, the organs to annotate all have properties and they are all linked by spatial relations. Thus, learning these relations and properties should help us to recognize them. Our approach relies on using fuzzy relations that take into account both quantitative and qualitative information, which enables to have a linguistic and thus understandable description of each relation. Learning fuzzy relations has already been proposed in [11] and in [19] to achieve higher classification performance but not for explaining the reasoning as we propose. Given an unknown example, the system looks for the set of objects that best satisfies the relations between the objects of interest. We model this as a constraint satisfaction problem. In Section 3, we describe the whole pipeline that consists in three main steps: assessing relations, extracting the most relevant ones and generating constraints for solving a constraint satisfaction problem and producing explanations. In Section 4, a demonstration of this approach is shown on a task of multiple organ recognition on medical images. This task is a good example of spatial reasoning since the spatial arrangement of the organs plays an important role in their recognition. In addition, working on medical images presents several challenges, including a need for explainability and the fact that datasets are usually small. We tested and compared our model to the state of the art and showed that our approach is able to achieve high accuracy and generate explanations in spite of a low number of training data.

2 Background

The approach we present in the next section relies on learning relevant fuzzy relations between objects for defining a constraint satisfaction problem. All the notions that are involved are reminded in this section.

2.1 Fuzzy Logic

Fuzzy logic and fuzzy set theory [40] can be seen as an extension of Boolean logic that enables to manage imprecision. In a universe A, a fuzzy set F is characterized by a mapping such as $\mu_F : A \rightarrow [0, 1]$. This mapping specifies in what extent each $a \in A$ belongs to F and it is called *the membership function of* F. If F is a non-fuzzy set, $\mu_F(a)$ is either 0, i.e. a is not a member of F, or 1, i.e. a is a member of F. This range of degrees is useful for dealing with vagueness.

The fuzzy logic framework is also convenient for expressing relations between two sets. Given two universes A and B, a binary fuzzy relation $\mathcal{R}$ is characterized by a mapping defined as $\mu_{\mathcal{R}} : A \times B \rightarrow [0, 1]$. It assigns a degree of relationship to any $(a, b) \in A \times B$. n-ary fuzzy relations are defined identically. Another advantage is that fuzzy logic allows using words instead of mathematical symbols.

2.2 Fuzzy Constraint Satisfaction Problem

A constraint satisfaction problem (CSP) consists in assigning some values to a set of variables that must respect a set of constraints, such as scheduling problems [31] for instance. [13] presents an extension of CSPs to the fuzzy logic framework to deal with imprecise parameters and flexible constraints. This is called a fuzzy constraint satisfaction problem (FCSP). A FCSP is defined by a set of variables X, a set of domains D and a set of flexible constraints C. It is an appealing framework in the context of explainable annotation since it enables to both solve the annotation task (getting each variable assignment) and generate explanations using the constraints.

To solve a FCSP, the FAC-3 algorithm [13, 38] is usually applied to prune the search space. Then, a backtracking algorithm explores every possible solution. Finally, we get the best solution by picking the one that is the most consistent with the set of constraints C.

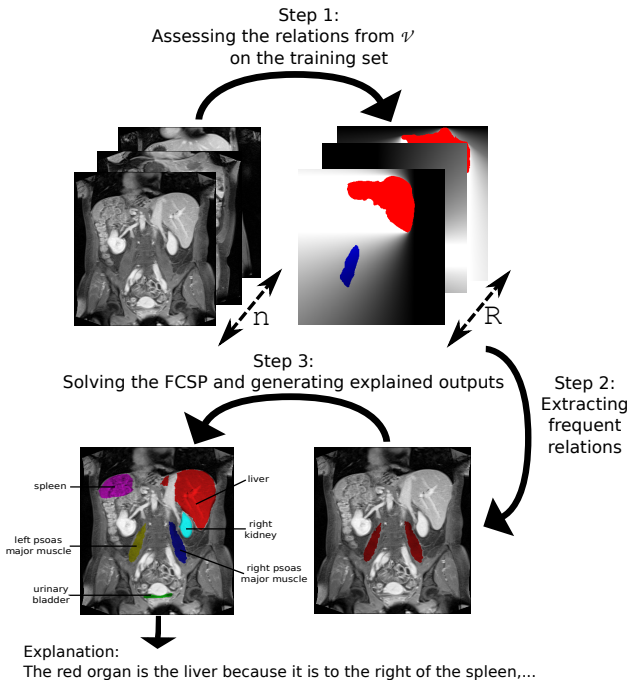


Figure 1: Illustration of our explainable multiple organ annotation system. In step 1, the R fuzzy relations from the vocabulary $\mathcal{V}$ are evaluated on a training set of n images. In step 2, the most frequent of them are extracted to set constraints. In step 3, for each test image, a FCSP is defined and solved to label the different regions. An explanation is provided for each labeling based on the constraints that are used.

3 Proposed Approach

In this section, we describe our new approach that aims at annotating regions of interest in images and at providing an explanation for each annotation. It consists of three steps: the assessment of fuzzy relations from a given vocabulary between the organs we are looking for (Sec. 3.1), the learning of the most relevant relations between the organs (Sec. 3.2) and the solving of a FCSP providing explanations for finding the regions that are the most consistent with the relevant relations and explaining the reasoning behind it (Sec. 3.3). An overview of the whole approach is illustrated in Figure 1.

3.1 Step 1: Assessing Relations

This step aims at evaluating several relations between the regions of interest (the organs) so that we can later (in the following step) find the most relevant of them.

Let us consider a training set $\mathcal{T}_{\text{train}}$ that contains n images $\{\mathbf{i}_1, \ldots, \mathbf{i}_n\}$ and a set of labels $\mathcal{Y}$ that contains N labels $\{y_1, \ldots, y_N\}$ such as each image $\mathbf{i} \in \mathcal{T}_{\text{train}}$ is divided into K regions of interest $\{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\}$ that are mapped to labels by the following function:

$$\begin{aligned} f : \{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\} &\rightarrow \mathcal{Y} \\ o_i \quad &\mapsto y_j \end{aligned} \tag{1}$$

Let us consider a set $\mathcal{V} = \{\mathcal{R}_1, \ldots, \mathcal{R}_R\}$ of relations. We call this set a vocabulary. It is set by an expert in the target task and it is composed of would-be relevant relations. For example, one relation can be a directional relation like *to the left of* or a distance relation like *close to*. The richer the vocabulary, the more expressive the system which should help to produce better annotations and explanations. Relations in $\mathcal{V}$ are automatically evaluated on the regions of interest of each image in $\mathcal{T}_{\text{train}}$. The way they are computed depends on the definition of the relation, as shown in Sec. 4.2.2.

For any relation $\mathcal{R} \in \mathcal{V}$, let $\alpha(\mathcal{R})$ denote its arity. $\mathcal{R}$ is evaluated for each possible $\alpha(\mathcal{R})$-tuple of regions of interest. It is important to distinguish $\mathcal{R}$ from its evaluations on the different regions. The number of evaluations to perform is:

$$\sum_{p=1}^{n} \sum_{j=1}^{R} \frac{K!}{(K - \alpha(\mathcal{R}_j))!} \tag{2}$$

At the end of this step, we have a set of evaluated relations between organs $\{\mathcal{R}(f(o_{\mathbf{i},v}), f(o_{\mathbf{i},w})) \mid \mathcal{R} \in \mathcal{V}, \mathbf{i} \in \mathcal{T}_{\text{train}}, o_{\mathbf{i},v}, o_{\mathbf{i},w} \in \mathbf{i}\}$ that can be seen as features.

3.2 Step 2: Learning Relevant Fuzzy Relations

In this step, the objective is to extract among the previously assessed relations the most relevant of them. For a label $y \in \mathcal{Y}$, our postulate is that the relevant relations involving the regions labelled as y are the most frequent

ones since they should be verified by most, if not all, examples of these regions. Thus, learning the relevant relations is performed by mining the most frequent ones. It is done in a one-vs-all way since the relevant relations for one class of organs are not the same as for a different class. As each example from one class should be correlated to each other, we use a fuzzy mining algorithm that takes advantage of that [33].

Let $E(\mathcal{V})$ be the set of all the evaluations of relations from $\mathcal{V}$ on the labeled regions of interest. A subset of relations J is a set belonging to $2^{E(\mathcal{V})}$. The mining algorithm we use is based on a fuzzy closure operator $h : 2^{E(\mathcal{V})} \rightarrow 2^{E(\mathcal{V})}$ that enables to find all the closed sets of relations [33]. All the frequent closed sets of relations are computed and the frequent sets of relations can be derived from them. A set of relations is said to be frequent when its frequency in the dataset is larger than a given threshold. Since this step is performed in a one-vs-all way, each class has its own threshold whose value is an hyperparameter determined during a validation phase. The value of this threshold has a direct impact on the number of frequent subsets of relations that are extracted. If it is too high, it is likely that no or few subsets of relations are seen as frequent, which may be not enough for discriminating classes. This would be a case of underfitting. On the other hand, if the threshold is too low, some irrelevant features will be kept. That would lead to overfitting. At the end of this step, for each label $y \in \mathcal{Y}$, we have a set of frequent subsets of evaluated relations F_y such as $F_y \subseteq 2^{2^{E(\mathcal{V})}}$.

3.3 Step 3: Solving the FCSP and Generating Explanations

Given a test example **i**, we can obtain a set of potential regions of interest by segmentation. The goal of this step is to find the labels of the regions that best satisfy the relations between organs that were learnt in the previous step. This can be modelled as a FCSP. Also, since these relations are associated to a linguistic description, we can generate an explanation for each annotation.

For each label $y \in \mathcal{Y}$, we got at the end of the previous step a set F_y. Let us define $F_y^{\max}$ such as :

$$F_y^{\max} = \{z \in F_y \mid \text{Card}(z) = \max_{v \in F_y} \big(\text{Card}(v)\big)\} \quad (3)$$

This set corresponds to the set of the frequent subsets of relations of maximal size. Each evaluated relation $\mathcal{R}\big(f(o_{\mathbf{i},v}), f(o_{\mathbf{i},w})\big)$ in the subsets of relations in $\bigcup_{y \in \mathcal{Y}} F_y^{\max}$ is directly translated into a constraint $c_{\mathcal{R}}\big(f(o_{\mathbf{i},v}), f(o_{\mathbf{i},w})\big)$. We can now build a model that is defined by the constraints that have been learned and its frequency thresholds. No iterative optimization process is needed, which makes it well suited to small training sets.

The test example **i** is divided into K regions of interest $\{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\}$ that we want to annotate. The FCSP we get is the following :

$$X = \{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\} \quad (4)$$

$$D = \{D_j \mid D_j = \mathcal{Y}, 1 \leq j \leq K\} \quad (5)$$

$$C = \{c_{\mathcal{R}}(f(o_{\mathbf{i},v}), f(o_{\mathbf{i},w})) \mid \mathcal{R}(f(o_{\mathbf{i},v}), f(o_{\mathbf{i},w})) \in U \text{ such as } U \subseteq \bigcup_{y \in \mathcal{Y}} F_y^{\max}\} \quad (6)$$

Then, each constraint in C is evaluated, the FCSP is solved and the first part of the output, the labels, are returned. We obtain a new mapping $f_{\mathbf{x}}$ such as :

$$\begin{aligned} f_{\mathbf{i}} : \{o_{\mathbf{i},1}, \ldots, o_{\mathbf{i},K}\} &\rightarrow \mathcal{Y} \\ o_i &\mapsto y_j \end{aligned} \quad (7)$$

Then, for each variable $o_{\mathbf{i},j} \in X$, an explanation is generated using the constraints in C. This is possible because the relations (and so the constraints) that we use are associated to a linguistic description. For instance, the constraint $c_{\mathcal{R}_{\text{to the left of}}}(A, B)$ (represented as a tuple $(A, B, \mathcal{R}_{\text{to the left of}})$) leads to: "$A$ is to the left of B". Thus, using the constraints generated from $F_y^{\max}$ enables to express an explanation in the form of "*output* BECAUSE *cause*$_1$,...,*cause*$_n$". For a given label y, all the constraints related to y are extracted. The least satisfied constraint gives us a certainty factor to moderate the explanation [6], e.g. "*This organ is likely to be annotated as the liver...*". The constraints and the certainty factor are then sent to a surface realiser like simpleNLG [17] to aggregate them into a syntactically correct sentence.

4 Case study

In this section, we detail the experiments we have performed on a dataset of medical images. The task is to perform explained multiple organ annotation by learning a model from few data. While multiple organ detection has been a regularly tackled topic in the literature [36, 10, 32, 24], multiple organ annotation has only been tackled in [39]. The principle of this method is to find images in the dataset that share visual characteristics with the image under study, and then to label it based on the labels from visually similar images. However, it cannot provide any explanation. In [24], abdominal organ detection is performed using fuzzy spatial rules, but these rules are not suited to other datasets and they have to be set by an expert before learning. Organ classification has been addressed in [35] using data augmentation to dodge the problem of having a small training set.

4.1 Dataset

It is important to note that the field of XAI is currently lacking a dataset that mainly focuses on explanations. This is why we carried out our experiments on a segmentation dataset that we used for assessing the accuracy of our model and the reliability of the explanations it produces. This dataset is named *Anatomy3* and has been presented in [22]. It contains 391 CT and MR images and their corresponding segmented organs. Images can be scans of the whole body (referred as CTwb and MRwb) or enhanced images of the abdomen (referred as CTce and MRce). Those are all 3D images that are actually the superposition of 2D slices. As we work on 2D images, we consider only slices in the following. We selected the slices to build a 2D image dataset. Figure 2 displays one example for each type of scan.

The set $\mathcal{Y}$ of organs (labels) we study is composed of the *liver*, the *spleen*, the *urinary bladder*, the *left* and *right kidneys*, the *left* and *right lungs* and the *left* and *right psoas major muscles*. We kept all the images that contain these 9 organs (and their corresponding segments), for a total of 35 examples and 315 segments in our dataset.

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

Figure 2: Examples of the four types of scans in the dataset.

4.2 Experimental Settings

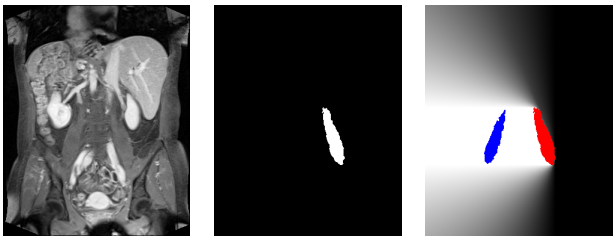
4.2.1 Model Training

The model we build with our approach consists in the frequent subsets of relations that are extracted. There are as many hyperparameters as labels and they correspond to the thresholds used for assessing the frequency of a subset of relations. Model selection is necessary to get optimized thresholds, which is why we used *nested cross-validation* [7]: (1) an outer cross-validation is performed in which we get a training set and a test set for each iteration, (2) an inner cross-validation is performed on the training set of the outer cross-validation to get an inner training set and a validation set for tuning hyperparameters. This enables to get an unbiased error prediction.

In the inner cross-validation, hyperparameter tuning is performed using bayesian optimization over 20 iterations with a Gaussian process prior. The acquisition function is the expected improvement.

4.2.2 Relations

Many fuzzy spatial relations have been studied in the literature [5]. In our experiments, we use directional, distance and symmetry relations. Directional and distance relations [3, 4] are computed as a fuzzy landscape and assessed using a fuzzy pattern matching approach [8]. As shown in Figure 3, the fuzzy landscape is generated by computing the fuzzy morphological dilation of a reference object by a structuring element whose shape determines the kind of relation. Let S be the space of the images. Let A be a reference object in S and $\mu_{A,\mathcal{R}}$ the membership



(a) Input (b) Segmented organ (c) Relation

Figure 3: (Best viewed in color) Example of how an input is used to compute a specific relation. Here, the goal is to compute the relation *blue organ to the left of the red organ*. Given an input (3a), a segmented organ is considered (3b) as the reference object. This organ is used to compute a fuzzy landscape (3c) that represents the degree to which each pixel verifies the relation *to the left of the red organ*. Finally, the relation is assessed by evaluating the degree of intersection between this fuzzy landscape and the blue organ.

function associated to the fuzzy landscape representing the relation $\mathcal{R}$ whose reference object is A. Let μ_B be the membership function corresponding to an object B in S. The relation $\mathcal{R}$ between A and B is the result of the fuzzy degree of intersection μ_{int} between $\mu_{A,\mathcal{R}}$ and μ_B such as [5]

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

For instance, in Figure 3, the relation $\mathcal{R}$ is *to the left of*, the reference object A is the red organ and the object B is the blue organ.

To get a finite catalogue of relations, we constrained the parameters of these relations to express only relations such as *above* or *close to*.

The symmetry relation [9] we use consists in finding the line that maximizes a symmetry measure between two organs. Since this measure is not differentiable, a direct search method is used to solve this optimization problem, such as the downhill simplex method.

We also use one property that can be seen as a unary relation since it characterizes just one organ. It evaluates how stretched an organ is. Given a segmented organ, a PCA is performed to get its two principal axes. Then, the organ is projected on both axis and the ratio of these projections is used to compute the degree corresponding to this property. However, this does not manage concave shapes well.

Our vocabulary of relations $\mathcal{V}$ contains: *to the left of*, *to the right of*, *below*, *above*, *close to*, *symmetrical to* and *stretched*. That makes 6 binary and one unary relations. As we consider 9 organs, the number of relations to evaluate for one image is equal to 441, which contributes to make our model expressive. There is however a trade-off between the expressivity of the system and the computation time needed for assessing all these relations.

4.3 Problem initialization

As stated in in Sec. 3, the whole process consists in three main steps. The inputs we deal with are segments provided in the datasets. They are not fuzzy, but the process is exactly the same whether we deal with fuzzy or crisp objects.

The intermediary goal is to generate constraints for defining a FCSP. Once solved, the FCSP returns the labels and constraints are used for generating explanations.

The variables are the segments provided in the dataset. Each of them corresponds to an organ. We have the following FCSP:

$$X = \{o_{\text{liver}}, o_{\text{spleen}}, o_{\text{bladder}}, o_{\text{r_kidney}}, o_{\text{l_kidney}}, o_{\text{r_lung}}, o_{\text{l_lung}}, o_{\text{r_psoas}}, o_{\text{l_psoas}}\}$$

$$D = \{D_{\text{liver}}, D_{\text{spleen}}, D_{\text{bladder}}, D_{\text{r_kidney}}, D_{\text{l_kidney}}, D_{\text{r_lung}}, D_{\text{l_lung}}, D_{\text{r_psoas}}, D_{\text{l_psoas}}\}$$

where D_i is equal to $\mathcal{Y}$. For each organ y, the flexible constraints are generated from the set of the frequent subsets of relations of maximal size $F_y^{\max}$ to build a set of constraints C. Furthermore, since every organ is unique, there cannot be identical annotations in this problem. That means C has to be extended with constraints representing that two variables cannot be the same, which is the *AllDifferent* global constraint.

The definition of the FCSP is thus made automatically. Then, once the FCSP is defined, for a given example, it can be solved as described in Sec. 2.2.

Organ	Value of the corresponding threshold
Liver	0.96
Spleen	0.86
Bladder	0.80
Right kidney	0.92
Left kidney	0.89
Right lung	0.98
Left lung	0.97
Right psoas muscle	0.92
Left psoas muscle	0.88

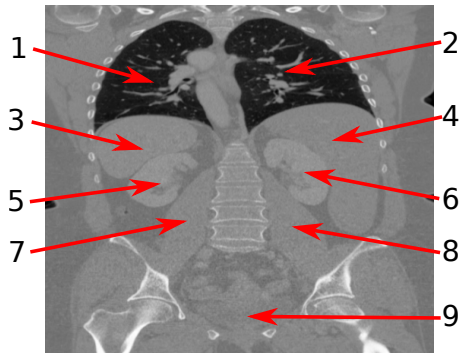
Table 1: Values of the thresholds that are the hyperparameters of our model. Each threshold is associated to one organ.

4.4 Results

Fig. 4 shows an example of output for an input image with 9 organs to annotate and thus 9 explanations to provide.

We evaluate our model using the accuracy, which is the ratio for all organs of the number of correct annotations over the total number of annotations. We got an accuracy of 100% for a model containing only directional relations. The outer cross-validation is actually a 3-fold cross-validation (23/24 training examples for 12/11 test examples in each iteration) and the inner one is a 4-fold cross-validation. As there are 9 organs to annotate, there are 9 hyperparameters that need to be set for extracting frequent relations (Table 1). Constraints could be added to the hyperparameter optimization process to make explanations longer or shorter. We observe the explanations rightfully rely on the relations that have been extracted and later turned into constraints. For example in Fig. 4, the set of constraints associated to the right kidney is: $C_{r_kidney} = \{(x_{r_kidney}, x_{l_kidney}, \mathcal{R}_{\text{symmetrical to}}), (x_{r_lung}, x_{r_kidney}, \mathcal{R}_{\text{above}}), (x_{r_kidney}, x_{liver}, \mathcal{R}_{\text{to the left of}}), (x_{bladder}, x_{r_kidney}, \mathcal{R}_{\text{below}}), (x_{r_kidney}, x_{l_kidney}, \mathcal{R}_{\text{to the right of}}), (x_{l_kidney}, x_{r_kidney}, \mathcal{R}_{\text{to the left of}})\}$. Some of these constraints may seem redundant, like the last two constraints in C_{r_kidney}. That can happen because fuzzy morphological dilations depend on the shape of the reference object. As two different organs are never exactly the same, there are slight differences between those two constraints. Each organ is linked to such a set of constraints. The final set of constraints C is the union of all these sets.

Assessing the quality of the explanations is tricky.



Organ 1 is very likely to be annotated as the left lung **because** it is *to the left of* the right lung (organ 2), it is *symmetrical* to the right lung and it is *above* the spleen (3).
Organ 9 is likely to be annotated as the bladder **because** it is *stretched*, it is *below* the right kidney (6) and *below* the left kidney (5).
Organ 4 is very likely to be annotated as the liver **because**...

Figure 4: Example of explained annotations.

What makes a good explanation ultimately depends on the knowledge and expectation of the end-user. Criteria like the coherence, the simplicity and the relevancy of the explanation are good indicators [30, 1], but they may not be easy to assess. Three evaluation methods are proposed in [12]: asking an expert, asking simple questions to a group of non-expert people or using a proxy model that has been proved to be explainable to assess the model under study.

We also investigated on the number of training examples that are required by our model to perform well. We get an accuracy of 99% at worst for a couple of training images (so 33 test examples). Actually, when dealing with just one training example, since our model looks for frequent relations to set the constraints, it will extract the relations whose evaluation is larger than the thresholds we talked about in Sec. 3.2. Any example that is not an outlier should then allow the model to perform well. Thus, we show that our approach can perform spatial reasoning and achieve high accuracy from just a pair of training examples.

We observe that our model outperforms the CNN classifier presented in [35], which does not achieve perfect accuracy. That model was trained on a bigger training set and does not provide any kind of explanation. The closest method to ours, which was presented in [39], does not

give any accuracy as a baseline. Its drawback is that it can miss labels, which happens at least once every five examples. In our approach, a label cannot be missing since every variable of the FCSP has to be associated to a domain.

On a side note, the generalisability of our approach depends on how well images are segmented (although fuzzy logic helps to deal with imprecision), how expressive the vocabulary is and how many outliers are in the dataset. Applications where one of this is missing may lead to a drop in performance regarding both annotations and explanations.

5 Conclusion and Prospects

In this article, we present a novel visual learning and reasoning framework whose goal is to explain and annotate relevant objects in images. The problem is formalized as a fuzzy constraint satisfaction problem. It is based on fuzzy spatial relations, which are learned on a set of annotated objects in images and then translated into constraints. We demonstrated our approach on a medical image dataset and showed that our method takes advantage of symbolic learning and reasoning so that it explains its results and it only needs a couple of training examples to achieve 99% accuracy.

In the future, we would like to work on a strategy that makes the first step of the process faster. A first idea is to determine a hierarchical structure of the spatial relations to apply a topological sort. Moreover, since fuzzy logic enables to manage imprecise segments, the goal is to insert an unsupervised segmentation model before the model we presented here. This would enable to adapt to different kinds of images.

Finally, this is a first step in mixing statistical machine learning (especially deep learning) for perception with symbolic learning and reasoning for higher level intelligence in order to create an explainable artificial intelligence.

References

[1] I. Baaj and J-P. Poli. Natural language generation of explanations of fuzzy inference decisions. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019.

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

[3] 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.

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] I. Donadello, L. Serafini, and A. d'Avila Garcez. Logic tensor networks for semantic image interpretation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1596–1602, 2017.

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

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] A. González, R. Pérez, Y. Caises, and E. Leyva. An efficient inductive genetic learning algorithm for fuzzy relational rules. *International Journal of Computational Intelligence Systems*, 5(2):212–230, 2012.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017.

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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.

[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] 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.

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