Free Argumentative Exchanges for Explaining Image Classifiers

Avinash Kori
Imperial College London
United Kingdom
a.kori21@imperial.ac.uk

Antonio Rago
Imperial College London
United Kingdom
a.rago@imperial.ac.uk

Francesca Toni
Imperial College London
United Kingdom
f.toni@imperial.ac.uk

ABSTRACT

Deep learning models are powerful image classifiers but their opacity hinders their trustworthiness. Explanation methods for capturing the reasoning process within these classifiers faithfully and in a clear manner are scarce, due to their sheer complexity and size. We provide a solution for this problem by defining a novel method for explaining the outputs of image classifiers with debates between two agents, each arguing for a particular class. We obtain these debates as concrete instances of *Free Argumentative eXchanges* (FAXs), a novel argumentation-based multi-agent framework allowing agents to internalise opinions by other agents differently than originally stated. We define two metrics (*consensus* and *persuasion rate*) to assess the usefulness of FAXs as argumentative explanations for image classifiers. We then conduct a number of empirical experiments showing that FAXs perform well along these metrics as well as being more faithful to the image classifiers than conventional, non-argumentative explanation methods. All our implementations can be found at https://github.com/koriavinash1/FAX.

KEYWORDS

Argumentation, Explainable AI, Quantization

ACM Reference Format:
Avinash Kori, Antonio Rago, and Francesca Toni. 2025. Free Argumentative Exchanges for Explaining Image Classifiers. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

With the increasing complexity and widespread deployment of deep learning models in our daily lives, the interpretation and explanation of these models' decisions have become a central focus in recent eXplainable Artificial Intelligence (XAI) literature [27, 34, 40]. Many existing approaches for explaining image classification models heavily rely on heatmaps and segments to localize regions of interest in input images that contribute to the model's output [9, 37, 38], typically offering static input-output-based explanations while lacking deeper insights into the underlying model being explained. The literature has repeatedly highlighted the need for deeper and more dynamic explanations [26, 28, 29], highlighting the input-output relationships of the model but also delving into its internal mechanisms and elucidating the model's reasoning. Also, there is an ongoing debate about the necessity of interactive explanations [7, 29] and explanations that are contrastive and selected [29].

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Dialogue-based explanations have been advocated as being useful in understanding the inner working of deep learning models [26]. [45] argues that explanations are especially important when the model has high uncertainty about the output when the prediction oscillates between different classes resulting in different interpretations of the model behaviour. [33] proposes argumentative exchanges to explain models via interactions amongst agents. Motivated by these varied lines of work, the main focus of this work is to extract explanations for image classifiers as debates between two artificial agents, arguing, in the spirit of bipolar argumentation [8], for and against the classifiers' outputs for given inputs.

Example 1. *As an abstract illustration, consider two agents $\mathcal{A}^1$ and $\mathcal{A}^2$ as outlined in Figure 1 (black and grey, respectively). Each agent has an initial perception (at timestep $t = 0$) of their environment as a bipolar argumentation framework (BAF) about a topic of interest (a in the figure) which we consider as private. As agents start debating, they share their knowledge (see the exchange BAF figure 1) and may expand their private BAFs (e.g. $\mathcal{A}^2$ learns that b supports a at timestep $t = 1$ and agent $\mathcal{A}^1$ learns that c attacks a at timestep $t = 2$). As in human debates, as agents expand their knowledge they may see things differently from the other agents (e.g. at the alternative timestep $t = 2'$ $\mathcal{A}^1$ sees $\mathcal{A}^2$'s's attack from c to a as a support).*

Overall, we make the following contributions:

- we define a novel form of *free argumentative exchanges* (FAXs) to characterise explanations amongst agents as illustrated in Figure 1; differently from [33], these exchanges allow for agents to disagree on what constitutes an attack or support amongst arguments exchanged during debates (and are thus *free*); this technical novelty empowers the use of FAXs for explanation of image classifiers;
- we instantiate FAXs so that they can serve as the basis for explaining the outputs of image classifiers;
- we provide an implementation of the instantiated FAXs, by adapting the methodology of [23] to allow agents to generate their own arguments;
- we evaluate our methodology and implementation quantitatively and qualitatively, with two types of image classifiers on two datasets; for the quantitative evaluation, we use two novel metrics to assess the argumentative quality of the generated debates, and, for comparison with baselines, two existing metrics (adapted to our setting) for ascertaining the faithfulness of explanations to the explained classifiers.

2 RELATED WORK

XAI methods. There has been a recent surge in methods advocating for a shift in how we perceive explanations, emphasizing the importance of viewing them as dialogues rather than solely relying on heatmaps or feature attributions, as standard in much of

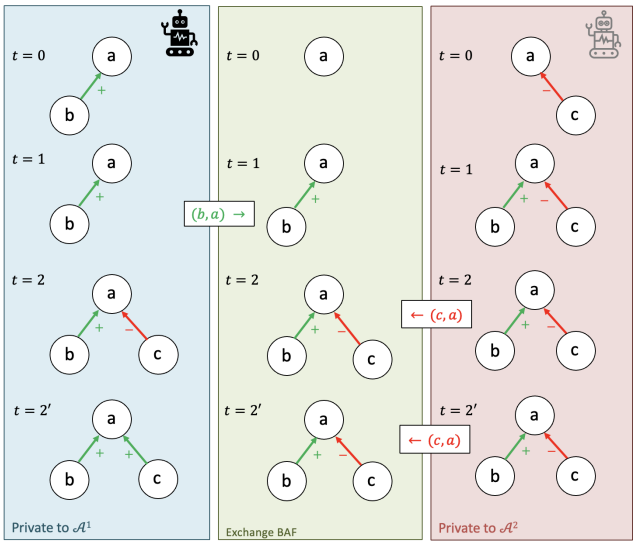

Figure 1: Illustration of FAXs (see Example 1).

the XAI literature [26, 28, 29]. Within the multi-agent setting, [33] proposed an argumentative framework for expressing (interactive) explanations in the form of dialogues. Additionally, [18] demonstrated a debate framework, which was further expanded in [23] to scale, allowing for the extraction of post-hoc explanations in the form of dialogues between two fictional agents. Inspired by [33], we define a multi-agent argumentative framework for explanation, and we adopt a variant of approach in [23] to implement the framework. Both [23] and our implementation utilize a surrogate model that faithfully represents the given classifier. The use of surrogate models is a standard practice in XAI, as seen e.g. in [15, 22, 47].

The agents in our explanations put forward arguments which can be understood as feature attributions such as LIME [34]. However, unlike [34], we do not randomly select input regions to mask; instead, our agents learn two different strategies to select regions to argue for and against a particular explanandum (input-output).

Our method is also related to [39], which argues against the rigidity of static and shallow explanations and introduces a new method for compactly visualizing how different combinations of regions in images impact the confidence of a classifier. Also, [45] aims to encourage the capturing of uncertain image regions, while [45] focus on statically capturing ambiguities in an image with respect to the given classifier, we generate both certain and uncertain regions through agent interactions in an iterative fashion [42].

Argumentation methods. Some research in XAI explores the use of computational argumentation [11]. This typically aims to assess specific claims by considering arguments that support and/or challenge the claim, as well as their relations within argumentative frameworks (AFs). These AFs may be as in [13] or Bipolar AFs (BAFs). Broadly, with our XAI approach we delve into a relatively unexplored area: explaining image classifiers through debates amounting to BAFs, which involve interactive gameplay among learning agents. Other approaches employing AFs for explainable image classification either utilize intrinsically argumentative models, e.g. as in [3], or mirror the mechanics of the model itself, e.g. as seen in [41]. In contrast, our approach focuses on explaining classifiers using latent features through (free) argumentative exchanges.

3 PRELIMINARIES

3.1 Computational Argumentation Background

We use (BAFs) [8], i.e. triples $\langle \mathbf{X}, \mathbf{A}, \mathbf{S} \rangle$ such that $\mathbf{X}$ is a finite set of *arguments*, $\mathbf{A} \subseteq \mathbf{X} \times \mathbf{X}$ is an *attack* relation and $\mathbf{S} \subseteq \mathbf{X} \times \mathbf{X}$ is a *support* relation. For all BAFs in this paper, we assume that $\mathbf{A} \cap \mathbf{S} = \varnothing$. For any $\alpha \in \mathbf{X}$, $\mathbf{A}(\alpha) = \{\beta \in \mathbf{X} | (\beta, \alpha) \in \mathbf{A}\}$ are the *attackers* of α and $\mathbf{S}(\alpha) = \{\beta \in \mathbf{X} | (\beta, \alpha) \in \mathbf{S}\}$ are the *supporters* of α.

We use the following notation from [33]: given BAFs $\mathbf{B} = \langle \mathbf{X}, \mathbf{A}, \mathbf{S} \rangle$, $\mathbf{B}' = \langle \mathbf{X}', \mathbf{A}', \mathbf{S}' \rangle$, we say that $\mathbf{B} \sqsubseteq \mathbf{B}'$ iff $\mathbf{X} \subseteq \mathbf{X}'$, $\mathbf{A} \subseteq \mathbf{A}'$ and $\mathbf{S} \subseteq \mathbf{S}'$; also, we use $\mathbf{B}' \setminus \mathbf{B}$ to denote $\langle \mathbf{X}' \setminus \mathbf{X}, \mathbf{A}' \setminus \mathbf{A}, \mathbf{S}' \setminus \mathbf{S} \rangle$. We also say that $\mathbf{B} = \mathbf{B}'$ iff $\mathbf{B} \sqsubseteq \mathbf{B}'$ and $\mathbf{B}' \sqsubseteq \mathbf{B}$, and $\mathbf{B} \sqsubset \mathbf{B}'$ iff $\mathbf{B} \sqsubseteq \mathbf{B}'$ but $\mathbf{B} \neq \mathbf{B}'$.

As conventional in the literature [5], a BAF $\langle \mathbf{X}, \mathbf{A}, \mathbf{S} \rangle$ may be equipped with *gradual semantics* $\sigma : \mathbf{X} \to \mathbb{I}$ assigning to arguments $\alpha \in \mathbf{X}$ values in $\mathbb{I}$, which is some set equipped with a pre-order $\leq$ (where, as usual $v < w$ denotes $v \leq w$ and $w \not\leq v$). In line with [33], we refer to σ as *evaluation method* and to $\mathbb{I}$ as *evaluation range*, to indicate their use by agents to evaluate arguments internally. We say σ is *dialectically monotonic* iff, as in [5]:

- given two BAFs $\mathbf{B} = \langle \mathbf{X}, \mathbf{A}, \mathbf{S} \rangle$ and $\mathbf{B}' = \langle \mathbf{X}', \mathbf{A}', \mathbf{S}' \rangle$, where $\mathbf{X}' = \mathbf{X} \cup \{a\}$, $\mathbf{A}' \cup \mathbf{S}' = \mathbf{A} \cup \mathbf{S} \cup \{(a, b)\}$, it is always the case that if $(a, b) \in \mathbf{A}'$, then $\sigma(\mathbf{B}', b) \leq \sigma(\mathbf{B}, b)$, while if $(a, b) \in \mathbf{S}'$, then $\sigma(\mathbf{B}', b) \geq \sigma(\mathbf{B}, b)$; and
- given two BAFs $\mathbf{B} = \langle \mathbf{X}, \mathbf{A}, \mathbf{S} \rangle$ and $\mathbf{B}' = \langle \mathbf{X}', \mathbf{A}', \mathbf{S}' \rangle$, where for an argument $a \in \mathbf{X} \cap \mathbf{X}'$, $\mathbf{A}'(a) = \mathbf{A}(a)$, $\mathbf{S}'(a) = \mathbf{S}(a)$, $\exists b \in \mathbf{A}'(a) \cup \mathbf{S}'(a)$ such that $\sigma(\mathbf{B}', b) > \sigma(\mathbf{B}, b)$ and $\forall c \in \mathbf{A}'(a) \cup \mathbf{S}'(a) \setminus \{b\}$, $\sigma(\mathbf{B}', c) = \sigma(\mathbf{B}, c)$ it is always the case that if $b \in \mathbf{A}'(a)$, then $\sigma(\mathbf{B}', a) \leq \sigma(\mathbf{B}, a)$, while if $b \in \mathbf{S}'(a)$, then $\sigma(\mathbf{B}', a) \geq \sigma(\mathbf{B}, a)$.

The first bullet states that an argument's strength cannot increase/decrease when a new attacker against/supporter for (respectively) the argument is added, all else being equal; the second bullet states that an argument's strength cannot increase/decrease when an attacker against/supporter for (respectively) the argument is strengthened, all else being equal.

Given a BAF $\mathbf{B} = \langle \mathbf{X}, \mathbf{A}, \mathbf{S} \rangle$, for $a, b \in \mathbf{X}$, we let a *path* from a to b be defined as $(c_0, c_1), \ldots, (c_{n-1}, c_n)$ for some $n > 0$ (the *length* of the path) where $c_0 = a$, $c_n = b$ and, for any $1 \leq i \leq n$, $(c_{i-1}, c_i) \in \mathbf{A} \cup \mathbf{S}$. We also use $\text{paths}(a, b)$ to denote the set of all paths from a to b (leaving the BAF implicit), and use $|p|$ for the length of path p. Also, we may see paths as sets of pairs. Then, as in [12, 33], we say that $\mathbf{B}$ is a *BAF for explanandum* $e \in \mathbf{X}$ iff i.) $\nexists (e, a) \in \mathbf{A} \cup \mathbf{S}$; ii.) $\forall a \in \mathbf{X} \setminus \{e\}$, there is a path from a to e; and iii.) $\nexists a \in \mathbf{X}$ with a path from a to a.

3.2 Image Classification Set-up

Consider a dataset $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \in \mathbb{R}^{h \times w \times c}$ represents a vector space with dimensions $h \times w$ corresponding to an image and c channels ($c \geq 1$). The label space $\mathcal{Y} = 1, \ldots, N$ consists of $N \geq 2$ classes. Following conventions in image classification [16, 17], we consider an image classifier trained on $\mathcal{D}$, comprising a feature extractor $f : \mathcal{X} \to \mathcal{Z}$ and a feature classifier $g : \mathcal{Z} \to \mathcal{Y}$. The feature extractor f maps the observational space $\mathcal{X}$ to a continuous latent space $\mathcal{Z} \subseteq \mathbb{R}^{m \times d}$, where each element of $\mathcal{Z}$ represents a set of

latent features with m elements, each of length d. Given an input $x \in \mathcal{X}$, the image classifier orders the classes in $\mathcal{Y}$ based on $g(f(x))$, predicting the top-class y as the output (with an abuse of notation we often write $g(f(x)) = y$). Consistent with prior works [23, 43], we assume the existence of a set $\mathcal{Z}^i \subset \mathcal{Z}$ of features associated with each class $i \in \mathcal{Y}$. These are referred to as the *class-i-specific features*. Up to Section 6, we defer discussion on how these class-specific features are obtained.

4 AGENTS AND EXPLANATIONS

Agents are represented as *private triples for explananda*, a notion adapted from [33], as follows.

Definition 1. *An agent $\mathcal{A}^i$ is a* private triple for an explanandum *e, amounting to $(\mathbb{I}^i, \mathbf{B}^i, \sigma^i)$ where:*

- $\mathbb{I}^i = \mathbb{I}^i_+ \cup \mathbb{I}^i_-$ *is an evaluation range, referred to as $\mathcal{A}^i$'s* private evaluation range*, where $\mathbb{I}^i_+$ (referred to as* positive evaluations*) and $\mathbb{I}^i_-$ (referred to as* negative evaluations*) are disjoint and for any $v_+ \in \mathbb{I}^i_+$ and $v_- \in \mathbb{I}^i_-$, $v_+ > v_-$;*
- $\mathbf{B}^i = \langle \mathbf{X}^i, \mathbf{A}^i, \mathbf{S}^i \rangle$ *is a BAF for e, referred to as $\mathcal{A}^i$'s* private BAF*;*
- σ^i *is an evaluation method, referred to as $\mathcal{A}^i$'s* private evaluation method*, such that, for any BAF $\mathbf{B} = \langle \mathbf{X}, \mathbf{A}, \mathbf{S} \rangle$ and, for any $a \in \mathbf{X}$, $\sigma^i(\mathbf{B}, a) \in \mathbb{I}^i$.*

Here, we use BAFs instead of quantitative BAFs [4] as in [33], and rely upon evaluation ranges split into two, rather than three, partitions as in [33] (so we disregard the neutral partition in [33]). The threshold between the two partitions is seen as the point where an agent "changes their mind" on the explanandum, and may correspond to a classifier's decision boundary, as we will see later. We assume, for the remainder of this section, that any evaluation method is dialectically monotonic (as defined in Section 3). We exemplify our notion of an agent in a simple, generic setting below (see Section 5 for instantiations for image classification).

Example 2. *An agent $\mathcal{A}^1$ is a private triple for an explanandum a, amounting to $(\mathbb{I}^1, \mathbf{B}^1, \sigma^1)$ where: $\mathbb{I}^1 = [0, 1]$ with $\mathbb{I}^1_- = [0, 0.6[$ and $\mathbb{I}^1_+ = [0.6, 1]$; $\mathbf{B}^1 = \langle \mathbf{X}^1, \mathbf{A}^1, \mathbf{S}^1 \rangle$ such that $\mathbf{X}^1 = \{a, b\}$ with arguments a:* we should eat at this pizzeria *and b:* it is highly recommended, *$\mathbf{S}^1(a) = \{(b, a)\}$ and $\mathbf{A}^1(a) = \emptyset$; and σ^1 is some dialectically monotonic semantics, which in this case could give, for example $\sigma^1(\mathbf{B}^1, a) = 0.75$ and $\sigma^1(\mathbf{B}^1, b) = 0.5$ (given the asymmetric set-up of a's attackers and supporters). Here we can see that $\mathcal{A}^1$'s positive reasoning for the explanandum overcomes the (absent) negative reasoning against it, resulting in its strength being above the threshold of 0.6 and thus gives a positive evaluation.*

We define a novel form of *argumentation exchanges* amongst agents, which will serve, later, as explanations:[1]

Definition 2. *A* free argumentative exchange (FAX) for an explanandum *e* amongst agents $\mathcal{A}$, *where $|\mathcal{A}| = m \geq 2$ and each agent in $\mathcal{A}$ is a private triple for e, is a tuple:*

$$\langle \mathbf{B}^x_0, \ldots, \mathbf{B}^x_n, \mathcal{A}_0, \ldots, \mathcal{A}_n, \mathcal{M} \rangle \quad (n \geq 0)$$

where, for all timesteps $t \in [n]$, $\mathbf{B}^x_t$ (referred to as the exchange BAF at step *t) is a BAF for e and $\mathcal{A}_t$ is a set of k agents, all private triples for e, such that:*

- *$\mathbf{B}^x_0 = \langle \mathbf{X}^x_0, \mathbf{A}^x_0, \mathbf{S}^x_0 \rangle$ is a BAF such that:*
  - *$\mathbf{X}^x_0 = \{e\}$; $\mathbf{A}^x_0 = \emptyset$; $\mathbf{S}^x_0 = \emptyset$;*
- *$\mathcal{A}_0 = \mathcal{A}$;*

and, at timestep $t \in]n]$, letting $\mathbf{B}^x_* = \langle \mathbf{X}^x_*, \mathbf{A}^x_*, \mathbf{S}^x_* \rangle = \mathbf{B}^x_t \setminus \mathbf{B}^x_{t-1}$:

- *$\mathbf{B}^x_t \supseteq \mathbf{B}^x_{t-1}$, where $\forall (a, b) \in \mathbf{A}^x_*$, $\exists j \in]m]$ such that $(a, b) \in \mathbf{A}^j_{t-1}$ and $\forall (c, d) \in \mathbf{S}^x_*$, $\exists k \in]m]$ such that $(c, d) \in \mathbf{S}^k_{t-1}$;*
- *$\mathcal{A}_t$ is a set of private triples $(\mathbb{I}^i, \mathbf{B}^i_t, \sigma^i)$ for e, one for each $i \in]m]$, where $\mathbf{B}^i_t \supseteq \mathbf{B}^i_{t-1}$ and $\mathbf{X}^x_* \subseteq \mathbf{X}^i_t$ and $\mathbf{A}^x_* \cup \mathbf{S}^x_* \subseteq \mathbf{A}^i_t \cup \mathbf{S}^i_t$;*

$\mathcal{M}$ is the contributor mapping*, such that, for $\mathbf{B}^x_n = \langle \mathbf{X}^x_n, \mathbf{A}^x_n, \mathbf{S}^x_n \rangle$, for every $(a, b) \in \mathbf{A}^x_n \cup \mathbf{S}^x_n$: $\mathcal{M}((a, b)) = (i, t)$ with $i \in]m]$ and $t \in]n]$.*

FAXs thus allow agents to add attacks or supports from their private to the exchange BAF, which are then incorporated to all other agents' frameworks as some form of relation. Note that it is only the BAFs which change with the timestep, not, for example, the evaluation ranges or methods. Intuitively, the contributor mapping returns, for each attack/support pair in the final exchange BAF, the agent who contributes the pair as well as the timestep at which the pair was contributed. Note that, by defining $\mathcal{M}$ as a mapping, we impose that there is a single contributor and timestep for each attack/support pair in the final exchange BAF, and therefore pairs cannot be introduced multiple times in FAXs. Note that FAXs are variants of argumentative exchanges (AXs) in [33]: whereas in AXs agents are equipped with quantitative BAFs, in FAXs they are equipped with BAFs; and, whereas in AXs agents are assumed to share a *lingua franca* of arguments' attackers and supporters (so that if an argument attacks or supports another for an agent it does so for all others), in FAXs attackers/supporters for an agent may be supporters/attackers for another, witnessing the 'free" nature of FAXs. Next, we illustrate (in the earlier simple, generic setting) why this may be useful.

Example 3. *Continuing from Example 2, a second agent $\mathcal{A}^2$ is a private triple for a, amounting to $(\mathbb{I}^2, \mathbf{B}^2, \sigma^2)$ where: $\mathbb{I}^2 = \mathbb{I}^1$; $\mathbf{B}^2 = \langle \mathbf{X}^2, \mathbf{A}^2, \mathbf{S}^2 \rangle$ such that $\mathbf{X}^2 = \{a, c\}$ with argument c:* there is pineapple on the pizza, *$\mathbf{A}^2(a) = \{(c, a)\}$ and $\mathbf{S}^2(a) = \emptyset$; and σ^2 is some dialectically monotonic semantics, which in this case could give, e.g., $\sigma^2(\mathbf{B}^2, a) = 0.25$ and $\sigma^2(\mathbf{B}^2, c) = 0.5$ (again given the asymmetric set-up of a's attackers and supporters). Here $\mathcal{A}^2$'s negative reasoning for the explanandum overcomes the (absent) positive reasoning against it, resulting in its strength being below the threshold of 0.6 and thus gives a negative evaluation. A FAX for a amongst agents $\mathcal{A} = \{\mathcal{A}^1, \mathcal{A}^2\}$ is then a tuple $\langle \mathbf{B}^x_0, \mathbf{B}^x_1, \mathbf{B}^x_2, \mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2, \mathcal{M} \rangle$ such that:*

- *$\mathbf{B}^x_1 \setminus \mathbf{B}^x_0 = \langle \{b\}, \emptyset, \{(b, a)\} \rangle$, where $\mathcal{M}((b, a)) = (1, 1)$, and thus $\mathcal{A}_1 \neq \mathcal{A}_0$, where $\mathcal{A}^1_1 = \mathcal{A}^1_0$ and $\mathcal{A}^2_1 \neq \mathcal{A}^2_0$ with $\mathbf{B}^2_1 \setminus \mathbf{B}^2_0 = \langle \{b\}, \emptyset, \{(b, a)\} \rangle$;*
- *$\mathbf{B}^x_2 \setminus \mathbf{B}^x_1 = \langle \{c\}, \{(c, a)\}, \emptyset \rangle$, where $\mathcal{M}((c, a)) = (2, 2)$, and thus $\mathcal{A}_2 \neq \mathcal{A}_1$, where $\mathcal{A}^2_2 = \mathcal{A}^2_1$ and $\mathcal{A}^1_2 \neq \mathcal{A}^1_1$ with $\mathbf{B}^1_2 \setminus \mathbf{B}^1_1 = \langle \{c\}, \emptyset, \{(c, a)\} \rangle$.*

Here, support (b, a) provided at timestep 2 by $\mathcal{A}^1$ is learnt by $\mathcal{A}^2$ (as is conventional in AXs [33]) as a support, indicating that the agents agree on the relation. Then, at timestep 2, $\mathcal{A}^2$ provides the attack (c, a), e.g. because pineapple on a pizza is anathema in Italy. However, $\mathcal{A}^1$ learns the relation (c, a) as a support, indicating that they considered c to be providing reasoning for a, e.g. because $\mathcal{A}^1$ likes pineapple on pizza. This shows how FAXs, differently to AXs, allow for differences in the

[1] Throughout, we denote with $[k]$ the set $\{0, 1, \ldots, k\}$, with $]k]$ $\{1, \ldots, k\}$, etc.

way agents interpret relations. Given that σ^1 and σ^2 are dialectically monotonic, we know that $\sigma^1(\mathbf{B}_0^1, a) = \sigma^1(\mathbf{B}_1^1, a) \leq \sigma^1(\mathbf{B}_2^1, a)$ and $\sigma^2(\mathbf{B}_0^2, a) \leq \sigma^2(\mathbf{B}_1^2, a) = \sigma^2(\mathbf{B}_2^2, a)$, respectively.

We can define a notion to restrict FAXs so that agents therein can be deemed to share a lingua franca, as follows.

Definition 3. *Given a FAX F for e amongsts $\mathcal{A}$, with $F = \langle \mathbf{B}_0^x, \ldots, \mathbf{B}_n^x, \mathcal{A}_0, \ldots, \mathcal{A}_n, \mathcal{M} \rangle$, $\mathcal{A}_n^i \in \mathcal{A}_n$ such that $\mathcal{A}_n^i = (\mathbb{I}^i, \mathbf{B}_n^i, \sigma^i)$ and $\mathbf{B}_n^i = \langle \mathbf{X}_n^i, \mathbf{A}_n^i, \mathbf{S}_n^i \rangle$, we say that F has an* effective lingua franca *iff* $(\bigcup_{\mathcal{A}_n^i \in \mathcal{A}_n} \mathbf{A}_n^i) \cap (\bigcup_{\mathcal{A}_n^i \in \mathcal{A}_n} \mathbf{S}_n^i) = \varnothing$.

It can be seen that the FAX in Example 3 does not have a lingua franca since the agents disagree on the relation (c, a). Typically, FAXs begin because there is a *conflict* between agents, amounting to a different *stance* on the explanandum:

Definition 4. *Given an agent $\mathcal{A}^i$, i.e. a private triple $(\mathbb{I}^i, \mathbf{B}^i, \sigma^i)$ (for some explanandum e) with $\mathbf{B}^i = \langle \mathbf{X}^i, \mathbf{A}^i, \mathbf{S}^i \rangle$, for any $a \in \mathbf{X}^i$, let $\mathcal{A}^i$'s* stance on *a be defined as $\Sigma^i(\mathbf{B}^i, a) = +$ (*positive stance*) iff $\sigma^i(\mathbf{B}^i, a) \in \mathbb{I}_+^i$, and $\Sigma^i(\mathbf{B}^i, a) = -$ (*negative stance*) otherwise. Then, a set $\mathcal{A}$ of agents/ private triples for explanandum e is* in conflict wrt *e iff there are two or more agents in $\mathcal{A}$ with different stances on e.*

Note that this notion of stance is adapted from [33], ignoring the neutral stance therein.

We see FAXs as means to lead to resolution of initial conflicts, by allowing agents to argue while identifying and filling any gaps in their beliefs (represented by their private BAFs). We adapt the following from [33], to characterise FAXs successfully leading to resolution (or not), from an initial conflict.

Definition 5. *Let $\mathcal{A}$ be a set of agents/private triples for explanandum e. Let $\mathcal{A}$ be in conflict wrt e. Let $F = \langle \mathbf{B}_0^x, \ldots, \mathbf{B}_n^x, \mathcal{A}_0, \ldots, \mathcal{A}_n, \mathcal{M} \rangle$ be a FAX for e amongst $\mathcal{A}$. Then:*

- *F is* unresolved at timestep t, *for $t \in]n]$, iff $\mathcal{A}_t$ is in conflict wrt e, and* resolved at t *otherwise;*
- *F is* unresolved *iff it is unresolved at every timestep $t \in]n]$;*
- *F is* resolved *iff it is resolved at timestep n.*

Example 4. *Continuing from Example 3, we know that $\sigma^1(\mathbf{B}_0^1, a) \in \mathbb{I}_+^1$ and $\sigma^2(\mathbf{B}_0^2, a) \in \mathbb{I}_-^2$, and so $\Sigma^1(\mathbf{B}_0^1, a) = +$ $\Sigma^2(\mathbf{B}_0^2, a) = -$, meaning $\mathcal{A}$ is in conflict wrt a. At timestep 1, let us assume that, although $\sigma^2(\mathbf{B}_0^2, a) \leq \sigma^2(\mathbf{B}_1^2, a)$, it remains the case that $\sigma^2(\mathbf{B}_1^2, a) \in \mathbb{I}_-^2$. Since $\sigma^1(\mathbf{B}_0^1, a) = \sigma^1(\mathbf{B}_1^1, a) \in \mathbb{I}_+^1$, we can see that the FAX is unresolved at timestep 1. Then, since $\sigma^1(\mathbf{B}_1^1, a) \leq \sigma^1(\mathbf{B}_2^1, a)$, thus $\sigma^1(\mathbf{B}_2^1, a) \in \mathbb{I}_+^1$ and $\sigma^2(\mathbf{B}_1^2, a) = \sigma^2(\mathbf{B}_2^2, a) \in \mathbb{I}_-^1$, we know that $\Sigma^1(\mathbf{B}_2^1, a) = +$, $\Sigma^2(\mathbf{B}_2^2, a) = -$ and thus the FAX is unresolved. Meanwhile, if $\mathcal{A}^1$ had interpreted (c, a) as an attack, let us say at an alternate $t = 2'$, since σ^1 is dialectically monotonic we know that it would have been the case that $\sigma^1(\mathbf{B}_1^1, a) \geq \sigma^1(\mathbf{B}_2^1, a)$, and the FAX may have been resolved if $\sigma^1(\mathbf{B}_{2'}^1, a) \in \mathbb{I}_-^1$, thus giving $\Sigma^1(\mathbf{B}_{2'}^1, a) = \Sigma^2(\mathbf{B}_{2'}^2, a) = -$.*

When FAXs are used for conflict resolution, exchange BAFs therein can be seen as *explanations*, in that they unearth the reasoning behind the resolution or otherwise of the conflict amongst the agents, with evidence that the explanandum is "correct" or not (when the FAX is resolved, depending on the final stance of all agents) or why it cannot be deemed "correct" or otherwise (when the FAX is unresolved). For illustration, in the first FAX ($t = 2$) in Example 4, the agents do not share the same stance on the explanandum due to their differing interpretations of (c, a), whereas in the second FAX ($t = 2'$), they agree on both this relation being an attack and on their final stances on the explanandum.

When using FAX for explaining image classification (from Section 5), we will restrict attention to special forms of FAXs, as follows.

Definition 6. *A* strictly interleaved FAX *for e amongst $\mathcal{A}$ is a FAX $\langle \mathbf{B}_0^x, \ldots, \mathbf{B}_n^x, \mathcal{A}_0, \ldots, \mathcal{A}_n, \mathcal{M} \rangle$ for e amongst $\mathcal{A}$ such that*

- *$\mathcal{A} = \{\mathcal{A}^1, \mathcal{A}^2\}$ (i.e. with $|\mathcal{A}| = 2$);*
- *for each timestep $t \in]n]$ there exists exactly one (a, b) such that $\mathcal{M}((a, b)) = (k, t)$ and if t is odd then $k = 1$; else $k = 2$.*

If n is even, then the FAX is equal opportunity.

The FAX in Example 3 is equal opportunity strictly interleaved. Intuitively, in a strictly interleaved FAX agents take turns, contributing one attack/support at a time and making the same number of contributions, to have the same chances at persuading the other.

5 EXPLANATIONS FOR IMAGE CLASSIFICATION

We see explanations for image classification as (exchange BAFs of) equal opportunity strictly interleaved FAXs amongst:

- $\mathcal{A}^1$ (the *proponent*, arguing for the class in $\mathcal{Y}$ predicted by the underlying image classifier for input image $x \in \mathcal{X}$ of interest), and
- $\mathcal{A}^2$ (the *opponent*, arguing against the predicted class).

For simplicity, we restrict attention to the top two classes in the ordering determined by the image classifier (g, see Section 3) for the input image, so that $\mathcal{A}^2$ argues against the predicted (top) class by arguing for the second best in the ordering. In the remainder of the paper we assume, without loss of generality, that class 1 is the predicted class for x and class 2 is the second best. In line with the machine learning literature, we also refer to class 1 as y.

In our approach to explaining image classification using FAXs, each agent argues for a particular class, and thus class-specific features (see Section 3) play the role of arguments exchanged between agents. The arguments put forward by the proponent and opponent can be seen as playing the role of positive and negative, respectively, feature attributions as in some XAI literature [27, 34, 38, 40]. However, intuitively, the exchange BAF in the FAX can also convey the reasoning of the classifier. We will provide an evaluation of FAXs as explanations of image classifiers in Section 8. Here, we focus on instantiating the generic FAXs for our purposes.

Concretely, we assume that the two agents explaining the output of an underlying image classifier by virtue of a FAX can leverage on *class-specific classifiers* $q^1 : 2^{\mathcal{Z}^1 \cup \mathcal{Z}^2} \to [0, 1]$ and $q^2 : 2^{\mathcal{Z}^1 \cup \mathcal{Z}^2} \to [0, 1]$, allowing the agents to evaluate sets of their own and other agents' arguments (in other words, amounting to their private evaluation methods). Until Section 6, we ignore how these class-specific classifiers can be learnt, alongside the class-specific features, so that they are faithful to the original image classifier being explained (see Section 7), which is crucial to guarantee that FAXs as explanations are also faithful. We also refer to the class-specific classifiers as *private classifiers* (with $\mathcal{Z}^1$ and $\mathcal{Z}^2$ also referred to as *private features*).

Finally, given that we aim at explaining the output of the image classifier, we use (x, y) as the *explanandum* e.

We now instantiate the general Definition 1, describing agents, to capture proponent and opponent for image classification.

Definition 7. *Let $i \in \{1, 2\}$. Then, the* initial image classification agent *is a private triple $(\mathbb{I}^i, \mathbf{B}^i, \sigma^i)$ for (x, y) such that:*

- $\mathbb{I}^i = [0, 1]$ *is the* agent's evaluation range, *with $\mathbb{I}^i_- = [0, \tau)$ and $\mathbb{I}^i_+ = [\tau, 1]$, for some* threshold $\tau \in [0, 1]$;
- $\mathbf{B}^i = \langle \mathbf{X}^i, \mathbf{A}^i, \mathbf{S}^i \rangle$ *is a BAF where if $i = 1$ then*
  - $\mathbf{X}^i \subseteq \{(x, y)\} \cup \mathcal{Z}^1$ *such that* $(x, y) \in \mathbf{X}^i$,
  - $\mathbf{A}^i = \emptyset$,
  - $\mathbf{S}^i \subseteq \mathcal{Z}^1 \times \{(x, y)\}$;

  and if $i = 2$ then
  - $\mathbf{X}^i \subseteq \{(x, y)\} \cup \mathcal{Z}^2$ *such that* $(x, y) \in \mathbf{X}^i$,
  - $\mathbf{A}^i \subseteq \mathcal{Z}^2 \times \{(x, y)\}$,
  - $\mathbf{S}^i = \emptyset$;
- $\sigma^i : \mathbf{X}^i \to \mathbb{I}^i$ *is such that:*
  - $\sigma^i(\mathbf{B}^i, (x, y)) = q^i(\mathbf{A}^i((x, y)) \cup \mathbf{S}^i((x, y)))$;
  - *for* $z \in \mathbf{X}^i \setminus \{(x, y)\}$, $\sigma^i(\mathbf{B}^i, z) = q^i(\{z\})$.

Corollary 1. *If $\sigma^1(\mathbf{B}^1, (x, y)) \neq \sigma^2(\mathbf{B}^2, (x, y))$, i.e. $q^1(\mathbf{S}^1((x, y))) \neq q^2(\mathbf{A}^2((x, y)))$, then $\exists \tau \in [0, 1]$ such that $\mathcal{A}$ is in conflict.*

Here, we use specialised notions compared to those in Definition 1. Specifically, the private BAFs are "shallow" and acyclic, with all attacks and supports from private features (of either agent) to the explanandum. The evaluation ranges are divided into positive and negative evaluations by a threshold τ, which, as Lemma 1 demonstrates, can be guaranteed to provide a dividing line between the two classes if the agents' class-specific features have an effect on the private classifiers' outputs. The evaluation method is defined in terms of the class-specific classifier of the agent: the evaluation of the explanandum is given by the classifier applied to its attackers and supporters (which are class-specific features), while the evaluation of a class-specific feature is given by the private classifier applied to this feature only.

To obtain explanations for image classifiers, we will use FAXs where agents update their private BAFs guided by their private classifiers, using the following "learning strategy".

Definition 8. *Let $\langle \mathbf{B}^x_0, \ldots, \mathbf{B}^x_n, \mathcal{A}_0, \ldots, \mathcal{A}_n, \mathcal{M} \rangle$ be an equal opportunity strictly interleaved FAX for explanandum (x, y) amongst initial image classification agents $\mathcal{A} = \{\mathcal{A}^1, \mathcal{A}^2\}$. Then, for $i, j \in \{1, 2\}$ with $i \neq j$, $\mathcal{A}^i$ adopts a* dialectically monotonic learning strategy *iff at timestep $t \in]n]$, if $\exists (z, (x, y))$ such that $\mathcal{M}((z, (x, y))) = (i, t)$, then $\mathbf{B}^i_t = \mathbf{B}^i_{t-1}$; otherwise $\mathcal{M}((z, (x, y))) = (j, t)$ and $\mathbf{B}^i_t = \langle \mathbf{X}^i_t, \mathbf{A}^i_t, \mathbf{S}^i_t \rangle$ is such that:*

- $(z, (x, y)) \in \mathbf{A}^i_t \setminus \mathbf{A}^i_{t-1}$ *iff* $q^i(\mathbf{A}^i_{t-1}((x, y)) \cup \mathbf{S}^i_{t-1}((x, y)) \cup \{z\}) - q^i(\mathbf{A}^i_{t-1}((x, y)) \cup \mathbf{S}^i_{t-1}((x, y))) < 0$;
- $(z, (x, y)) \in \mathbf{S}^i_t \setminus \mathbf{S}^i_{t-1}$ *iff* $q^i(\mathbf{A}^i_{t-1}((x, y)) \cup \mathbf{S}^i_{t-1}((x, y)) \cup \{z\}) - q^i(\mathbf{A}^i_{t-1}((x, y)) \cup \mathbf{S}^i_{t-1}((x, y))) \geq 0$;

and σ^i is such that:

- $\sigma^i(\mathbf{B}^i_t, (x, y)) = q^i(\mathbf{A}^i_{t-1}((x, y)) \cup \mathbf{S}^i_{t-1}((x, y)) \cup \{z\})$;
- $\forall z' \in \mathbf{X}^i_t \cap \mathbf{X}^i_{t-1}$, $\sigma^i(\mathbf{B}^i_t, z') = \sigma^i(\mathbf{B}^i_{t-1}, z')$; *and*
- $\sigma^i(\mathbf{B}^i_t, z) = |q^i(\mathbf{A}^i_{t-1}((x, y)) \cup \mathbf{S}^i_{t-1}((x, y)) \cup \{z\}) - q^i(\mathbf{A}^i_{t-1}((x, y)) \cup \mathbf{S}^i_{t-1}((x, y)))|$.

In the remainder, we refer to equal opportunity strictly interleaved FAXs where both agents adopt a dialectically monotonic learning strategy simply as FAXs. Note that, in FAXs, the initial private BAF of each agent contains all arguments representing the agent's private features. Then, any arguments learned from the other agent's contributions are such that their addition to the agent's private BAF results in dialectically monotonic behaviour (due to the agent's characterisation of the relation as an attacker or supporter), in the spirit of [32]. This naturally leads to:

Proposition 1. *In any FAX for an explanandum (x, y) amongst agents $\mathcal{A} = \{\mathcal{A}^1, \mathcal{A}^2\}$, for any $\mathcal{A}^i \in \mathcal{A}$, σ^i is dialectically monotonic.*[2]

This result sanctions that our choices in instantiating the semantics for FAXs leads to explanations that are dialectically monotonic, which has been identified as an important property by argumentation practitioners [2, 5, 31] and found to be intuitive by humans (as shown for a form of dialectical monotonicity in [1]).

6 IMPLEMENTATION

In this section, we detail our methodology (overviewed in Figure 2) for obtaining and evaluating empirically FAXs with data for image classification. Specifically, we detail how class-specific (private) features can be obtained (Section 6.1); and how class-specific agents (including their private classifiers and their policies for contributing to FAXs) are learnt and deployed (Section 6.2). Note that, while FAXs are amongst two agents only (for the top- and second-best predicted classes by the classifier), given that different inputs will result in different predictions we need to develop, at training time, all agents, with their private features and classifiers.

6.1 Class-specific discrete features

We obtain these by simultaneously training, similarly to [23]:

- N *codebooks* $\mathfrak{C}_1, \ldots, \mathfrak{C}_N$ (one per class in $\mathcal{Y}$); for each $i \in \{1, \ldots, N\}$, $\mathfrak{C}_i \in \mathbb{R}^{\tilde{n} \times d}$ corresponds to $\mathcal{Z}^i$ (see Section 3);[3] for $z = f(x)$, if $\bar{y}$ is the top-class predicted by $g(z)$, we use $\mathfrak{C}_{\bar{y}}(z)$ to stand for the specific discrete features in $\mathfrak{C}_{\bar{y}}$ corresponding to z (we use $\tilde{z}$ to stand for $\mathfrak{C}_{\bar{y}}(z)$ if clear);
- a *quantized classifier* $q : \mathfrak{C} \to \mathcal{Y}$, for $\mathfrak{C} = \mathfrak{C}_1 \cup \ldots \cup \mathfrak{C}_N$, distilling the knowledge of the feature classifier g so that $q(\tilde{z})$ approximates $g(z)$.

Intuitively, a codebook is a collection of averaged patterns or "concept representations" that summarize the key features of data [43].

Training. To get the codebooks, we draw inspiration from [43]. Intuitively, for each $(x, y) \in \mathcal{D}$, with $z = f(x)$, for $\bar{y} = \texttt{argmax}(g(z))$ (i.e. $\bar{y}$ is the top-class predicted by the feature classifier g for z), we aim at deterministically mapping the elements of z to the nearest elements of $\mathfrak{C}_{\bar{y}}$ using some convex distance function δ, as follows:

$$\tilde{z} = \{\underset{\tilde{k} \in \mathfrak{C}_{\bar{y}}}{\operatorname{argmin}} \delta(z_k, \tilde{k}) | z_k \in z\} \tag{1}$$

To learn this in an end-to-end fashion, we use the Gumble sampling procedure from [19], resulting in $\tilde{z}$ as a projection of the

[2] The proof of this result can be found in [25].

[3] We assume for simplicity that $M^i = \tilde{n}$ for all $i \in \{1, \ldots, N\}$, namely all agents have the same number of class-specific features.

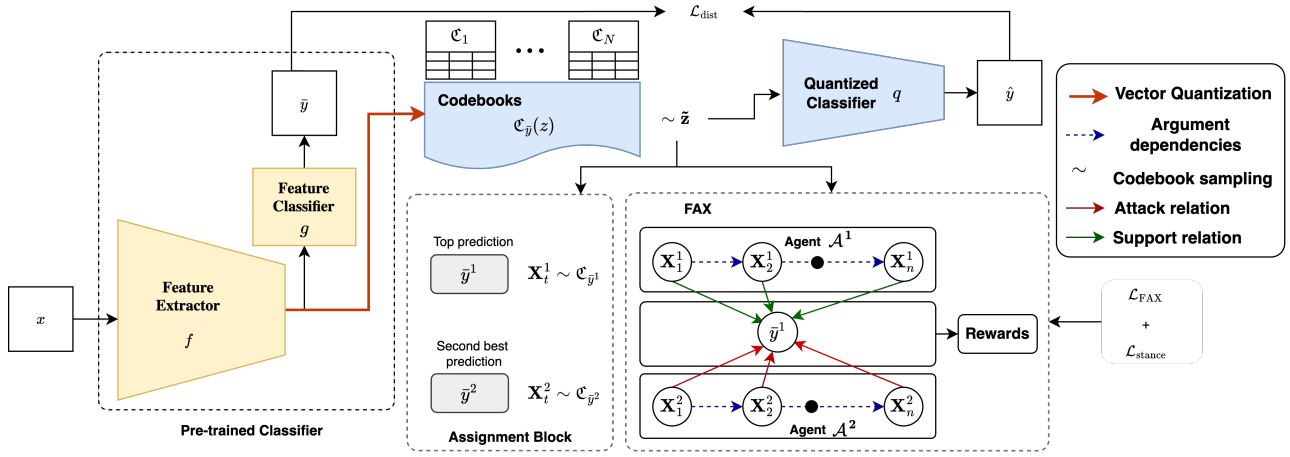

Figure 2: Overview of our implementation (argument dependencies are temporal).

continuous features in z onto every element of codebook $\mathfrak{C}_{\bar{y}}$, corresponding to pairwise similarity scores between z and all $\tilde{n}$ codebook features $\mathfrak{C}_{\bar{y}}$. The resulting quantization objective for training is described as follows (see [19] for details):

$$\mathcal{L}_{\text{quant}} = \sum \texttt{softmax}(\tilde{z})\,(\texttt{logsoftmax}(\tilde{z})) \tag{2}$$

To faithfully learn the quantized classifier q, we adopt the following *distillation* loss during training, where $\bar{y}$ is the predicted class (by g) for input x and CE correspond to cross-entropy loss:

$$\mathcal{L}_{\text{dist}} = \mathcal{L}_{\text{quant}} + \text{CE}\,(q(\tilde{z})), \bar{y}) \tag{3}$$

By using this loss, we strive towards a faithful q to the classifier.

6.2 Class-specific agents

We obtain N agents, each of which is responsible for arguing wrt a particular class in $\mathcal{Y}$. For any given input x, an instance of FAX is obtained between the two agents whose class is the top-class predicted by $g(f(x))$ (for the proponent) and the second-best class (for the opponent). Specifically, if $g(f(x)) = [\bar{y}^1, \ldots, \bar{y}^N]$, then the proponent is $\mathcal{A}^{\bar{y}^1}$ and the opponent is $\mathcal{A}^{\bar{y}^2}$, as depicted in the assignment block in Figure 2 (the assignment block selects the agents based on the estimated top two classes).

We model each agent $\mathcal{A}^i$ (arguing for class $\bar{y}^i$) as a *sequence model* ζ^i. In line with [23], we use gated recurrent units (GRUs) for realizing these sequence models. The sequence model ζ^i operates on a hidden state vector (h^i_{t-1}), which can be treated as a proxy interpretation for information drawn from the exchange BAF in the FAX till the current timestep (t). The sequence model uses a private modulator network $\mathcal{M}^i$ which encodes arguments (in $\mathbf{X}^i_t$ in the agent's private BAF $\mathbf{B}^i_t$) to update the hidden state representation:

$$h^i_t = \zeta^i(h^i_{t-1}, \mathcal{M}^i(\mathbf{X}^i_t)) \tag{4}$$

The output (hidden state vector) of ζ^i is then used to determine a *policy function* Π^i, for determining how agents contribute attacks and supports (in the agent's private BAF $\mathbf{B}^i_t$) to FAXs, by determining which argument these attacks (if the agent is the opponent $\mathcal{A}^j$) or supports (if the agent is the proponent $\mathcal{A}^i$) are drawn from:

$$\mathbf{X}^i_t \sim \Pi^i(h^i_{t-1} \mid \mathbf{X}^i_{t-1} \cup \mathbf{X}^j_{t-1}, \bar{y}^i) \tag{5}$$

Note that these arguments are from $\mathfrak{C}_{\bar{y}^i}$, corresponding to private features in $\mathcal{Z}^{\bar{y}^i}$. The output (hidden state vector) of ζ^i is also used to obtain the *private classifier* q^i, as a multi-layer perceptron with hidden state vectors as inputs, associating them to class confidence in turn used to assign values to sets of private features as in Section 5.

Training. To obtain each agent's sequence model in an end-to-end fashion, we adapt the REINFORCE learning algorithm from [30]. Specifically, we see the next argument prediction/selection as a reinforcement learning task, with agents' rewards as follows.

Definition 9. *Let $\langle \mathbf{B}^x_0, \ldots, \mathbf{B}^x_n, \mathcal{A}_0, \ldots, \mathcal{A}_n, \mathcal{M}\rangle$ be a FAX for explanandum (x, y) amongst agents $\mathcal{A} = \{\mathcal{A}^1, \mathcal{A}^2\}$. Then, for $i \in \{1, 2\}$ and $t \in \{1, \ldots, n\}$, $\mathcal{A}^i$'s* reward *at timestep t is $r^i_t = \Sigma^i(\mathbf{B}^i_0, e)\sigma^i(\mathbf{B}^i_t, e)$.*[4]

Thus, reward is a continuous-valued function modelled using the agent's "confidence" towards the explanandum, and reflecting the contributed arguments to date and the agent's original stance towards the explanandum. Note that this stance is always positive for the first agent and negative for the second, given our choice of τ in Section 5. Note also that rewards are between $[-1, 1]$.

We combine the agent's reward with REINFORCE gradients:

$$\mathcal{L}^i_{\text{FAX}} = -\sum_t \log \Pi^i_{\theta^i}(h^i_t \mid \mathbf{X}^i_{t-1} \cup \mathbf{X}^j_{t-1}, \bar{y}^i)(r^i_t - b^i_t) \tag{6}$$

where θ^i are the parameters of policy Π^i (and thus modulator $\mathcal{M}^i$) being learnt and b^i_t is a baseline value estimated by $\mathcal{A}^i$ at timstep t, which is mainly used to reduce the variance in the agent's behaviour during the exploration stage in (reinforcement) learning. Minimisation of this loss can also be viewed as maximisation of log-likelihood of the policy distribution [30].

[4] Here, we treat the agent's initial stance as a (positive/negative) sign for the explanandum's current strength in the private BAF.

Finally, to encourage agents to argue for a particular class ($\bar{y}^1$ for the proponent and $\bar{y}^2$ for the opponent) we use the *stance loss* $\mathcal{L}^i_{\text{stance}} = \text{CE}(\Sigma^i(\mathbf{B}^i_t, (x, \bar{y}^i)), \bar{y}^i)$ to obtain the combined loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dist}} + \left(\mathcal{L}^1_{\text{FAX}} + \mathcal{L}^1_{\text{stance}}\right) + \left(\mathcal{L}^2_{\text{FAX}} + \mathcal{L}^2_{\text{stance}}\right).$$

Deployment. After training, we deploy the learnt agents for generating FAXs. To determine the number of timesteps in each FAX, we adopt the following strategy. We analyse cosine similarity between arguments to evaluate the information contributed in each timestep and we model the gain in information as the average dissimilarity of the contributed argument wrt all the previous arguments. We terminate the FAX if the mean dissimilarity is less then a small amount (which is a parameter) or when the FAX is resolved.

7 EVALUATION

In this section we lay out our approach to evaluating the realization of our FAXs for explaining image classifiers. We use evaluation metrics for assessing *faithfulness* and *argumentative quality* of our FAXs. The metrics are measured on a test set $\mathcal{T} \subseteq \mathcal{X} \times \mathcal{Y}$, providing ground-truths (correct classifications) for a number of inputs (we will use concrete instances of $\mathcal{T}$ in our experiments in Section 8).

We define the metrics using the same notation $\mathfrak{C}(z)$ as in Section 6 as well as notations $\mathfrak{C}(h)$ to represent the codebook corresponding to hidden state representation h (in some sequence model) and $q^i(h)$ to represent the values assigned by q^i to the private features/arguments corresponding to h (in ζ^i).

The faithfulness metrics are adapted from the literature, and [23] in particular. The first metric measures *correctness* of q and of the codebooks by measuring accuracy wrt the ground-truth in $\mathcal{T}$:

$$|\{(x, y) \in \mathcal{T} \mid q(\mathfrak{C}(z)) = y\}| / |\mathcal{T}|$$

The second metric measures *completeness* of q on the BAFs resulting from FAXs, by measuring the accuracy of q on the codebooks corresponding to the hidden state representations of the arguments in these BAFs, wrt the classifier's predictions on $\mathcal{T}$:[5]

$$\left|\{(x, y) \in \mathcal{T} \mid q\left(\mathfrak{C}\left(\cup_{i=1}^{N} h^i_n\right)\right) = (g \circ f)(x)\}\right| / |\mathcal{T}|$$

This metric gives an indication of the faithfulness to the original classifier of q on the output FAXs (as the loss function used during training only strives towards faithfulness of q on the input to FAXs).

The argumentative quality metrics are tailored to our (implementation of) FAXs. The third metric measures *consensus* amongst the (two) agents, in terms of the number of resolved FAXs:

$$\left|\{(x, y) \in \mathcal{T} \mid q^i(h^i_n) = q^j(h^j_n), j \neq i\}\right| / |\mathcal{T}|$$

The fourth (and final) metric (*pro persuasion rate*) measures consensus again, but towards the proponent agent:

$$\left|\{(x, y) \in \mathcal{T} \mid q^1(h^1_0) = q^j(h^j_n), j \neq 1\}\right| / |\mathcal{T}|$$

Given that agents are trained to disagree (see Definition 9) both argumentative metrics can be seen as estimates of the goodness of the learnt class-specific features: high values indicate that information is leaked across different features (arguments).

[5]With an abuse of notation we use n to indicate the length of every FAX obtained from datapoints in $\mathcal{T}$, even though different FAXs will typically have different lengths.

Table 1: Accuracies of the trained classifiers

	Fair	Biased	Random
AFHQ-ResNet-18	0.95	0.39	0.31
AFHQ-DenseNet-121	0.96	0.47	0.32
FFHQ-ResNet-18	0.88	0.58	0.47
FFHQ-DenseNet-121	0.92	0.61	0.48

8 EXPERIMENTS

We analyse FAXs on the high resolution animal and human faces (AFHQ[10], FFHQ [20]) datasets, with two well known architectures ResNet-18 [16] and DenseNet121 [17] as image classifiers $g \circ f$. We consider three settings: (i) *fair*, where the classifier is trained with correct labels; (ii) *biased*, where the classifier is trained with biased labels, obtained by randomly switching the labels for 10% of the datasets, and (iii) *random*: where we use randomly initialised weights for the classifiers rather than training them.

Qualitative results. Figure 3 shows some FAXs visualised as in [23], using the approach in [6]. In all these figures, the first image is the input, while the first and second rows correspond respectively to $\mathcal{A}^1$'s and $\mathcal{A}^2$'s arguments. We can see that FAXs focus on different but semantically meaningful regions on the input images, for both agents. In the biased setting, as previously described, we expect some leakage of information across different class-specific features/arguments, as observed in the figure.

Quantitative results. Table 1 gives the classifiers' accuracy on test sets. For the DenseNet121 classifiers, Table 2 gives the faithfulness results in comparison with standard baselines (i.e. GradCAM [38], DeepLIFT [40], DeepSHAP [27] and LIME [34]) and Table 3 measures the argumentative metrics in the three settings (for FAXs only, as these metrics are not applicable to baseline methods). Further, Table 4 measure all metrics for the esNet-18 classifiers. We observe that *completeness* is high in all settings, while *correctness* is high for the fair and biased settings only; this is due to completeness reflecting FAXs behaviour wrt continuous classifier, while correctness is measured wrt ground truth. As for the last two metrics, the experiments show higher values in the biased than fair settings, as expected, given that we expect the leak of features across codebooks due to incorrect label assignment in the former, which results in higher consensus and easier persuasion. We can observe mixed behaviours with random classifiers, which results in high value in the case of overlapping features during initialisation and low in the other case. Overall, the experiments show that for more "uncertain" models (random), the agents cannot reach as much consensus as for "certain" models (fair and biased) and the persuasion rate is lower for "uncertain" models. This analysis is empowered by the argumentative nature of our explanatory framework.

9 CONCLUSIONS

We have defined explanations for image classification as (free) argumentative exchanges between two agents, aiming to demystify trained image classifiers based on the argument contribution

Table 2: Faithfulness properties of all considered methods on the DenseNet121 classifiers.

Methods → Dataset ↓	Correctness					Completeness				
	GradCAM	DeepLIFT	DeepSHAP	LIME	FAX	GradCAM	DeepLIFT	GradSHAP	LIME	FAX
FFHQ-Random	0.55	0.58	0.54	0.50	0.53	0.40	0.41	0.38	0.36	0.97
FFHQ-Biased	0.35	0.37	0.39	0.50	0.91	0.34	0.36	0.40	0.48	1.00
FFHQ-Fair	0.77	0.77	0.73	0.58	0.96	0.74	0.74	0.68	0.59	0.96
AFHQ-Random	0.30	0.32	0.28	0.25	0.55	0.35	0.36	0.33	0.31	0.72
AFHQ-Biased	0.27	0.32	0.37	0.48	0.71	0.35	0.38	0.40	0.47	0.91
AFHQ-Fair	0.75	0.71	0.71	0.52	0.78	0.70	0.66	0.67	0.54	0.99

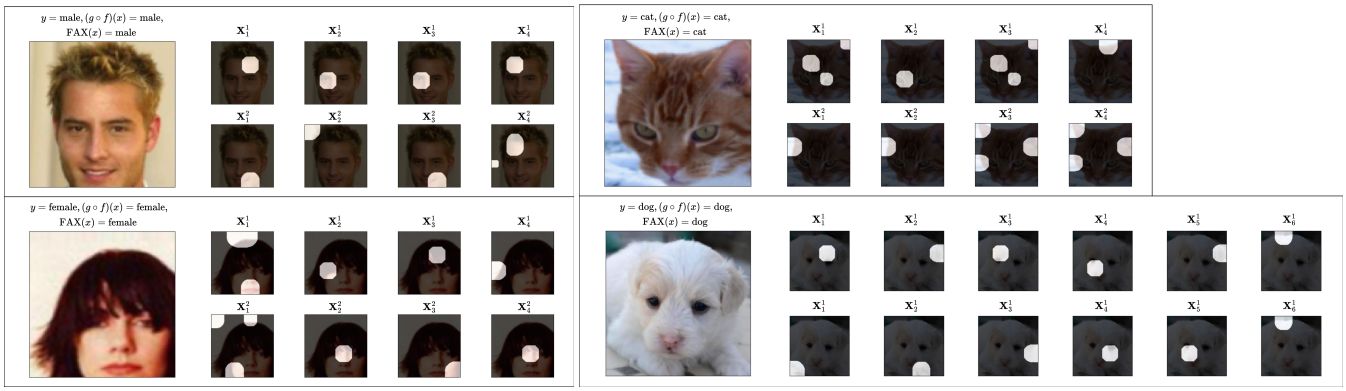

Figure 3: Arguments in FAXs (the proponent starts with the top left argument, the opponent follows with the argument below it, etc.), for classifiers trained on FFHQ (left column) and AFHQ (right column), on fair (top row) and biased (bottom row) settings.

Table 3: Argumentative metrics for DenseNet-121 classifiers.

	Fair	Biased	Random
AFHQ-Consensus	0.24	0.44	0.26
AFHQ-Pro persuasion rate	0.27	0.41	0.77
FFHQ-Consensus	0.42	0.54	0.09
FFHQ-Pro persuasion rate	0.33	0.50	0.38

Table 4: All metrics for ResNet-18 classifiers.

	Fair	Biased	Random
AFHQ-Correctness	0.77	0.73	0.29
AFHQ-Completeness	0.99	0.93	0.68
AFHQ-Consensus	0.57	0.89	0.13
AFHQ-Pro persuasion rate	0.48	0.56	0.16
FFHQ-Correctness	0.65	0.45	0.51
FFHQ-Completeness	0.99	0.99	0.91
FFHQ-Consensus	0.31	0.90	0.45
FFHQ-Pro persuasion rate	0.24	0.53	0.71

strategies by the agents. Differently from standard feature attribution methods generating heatmaps over responsible regions in images, our method generates more fine-grained composition of sub-regions, incrementally. Our work opens many opportunities for future work. We plan to investigate whether FAXs can uncover shortcuts in classifiers. Further, it would be valuable to collaborate with domain experts to attribute semantic meaning to arguments, potentially aiding alignment between human understanding and the latent knowledge of models. Also, it would be interesting to apply our approach in settings where quantized representation learning is already explored, ranging from natural images to medical data [14, 35, 43], or by targeting other (potentially more complex) architectures, e.g. transformers [44]. Methodologically, we also plan to explore the use of object identification methods such as grounded slot attention [24], instead of quantization, to improve the human understandability of arguments in our FAXs. Finally, a promising avenue for fully exploiting the capabilities of FAXs amounts to leveraging notions from [21, 36, 46] to create hierarchical concepts, giving FAXs with more interesting argument interactions.

ACKNOWLEDGMENTS

This research was partially supported by the ERC under the EU's Horizon 2020 research and innovation programme (grant no. 101020934), by J.P. Morgan and the RAEng (grant no. RCSRF2021/11/45) and by the UKRI (grant no. EP/S023356/1) via the CDT in Safe and Trusted Artificial Intelligence.

REFERENCES

[1] Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. 2020. DAX: Deep Argumentative eXplanation for Neural Networks. *CoRR* abs/2012.05766 (2020). https://arxiv.org/abs/2012.05766

[2] Leila Amgoud and Jonathan Ben-Naim. 2018. Weighted Bipolar Argumentation Graphs: Axioms and Semantics. In *IJCAI*. 5194–5198. https://doi.org/10.24963/IJCAI.2018/720

[3] Hamed Ayoobi, S. Hamidreza Kasaei, Ming Cao, Rineke Verbrugge, and Bart Verheij. 2023. Explain What You See: Open-Ended Segmentation and Recognition of Occluded 3D Objects. *CoRR* abs/2301.07037 (2023). https://doi.org/10.48550/arXiv.2301.07037

[4] Pietro Baroni, Antonio Rago, and Francesca Toni. 2018. How Many Properties Do We Need for Gradual Argumentation?. In *AAAI*. 1736–1743. https://doi.org/10.1609/aaai.v32i1.11544

[5] Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *Int. J. Approx. Reason.* 105 (2019), 252–286. https://doi.org/10.1016/J.IJAR.2018.11.019

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *CVPR*. 3319–3327. https://doi.org/10.1109/CVPR.2017.354

[7] Alison Cawsey. 1991. Generating Interactive Explanations. In *AAAI*. 86–91. http://www.aaai.org/Library/AAAI/1991/aaai91-014.php

[8] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *ECSQARU*. 378–389. https://doi.org/10.1007/11518655_33

[9] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *WACV*. 839–847. https://doi.org/10.1109/WACV.2018.00097

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *CVPR*. 8185–8194. https://doi.org/10.1109/CVPR42600.2020.00821

[11] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: A Survey. In *IJCAI*. 4392–4399. https://doi.org/10.24963/IJCAI.2021/600

[12] Louise Dupuis de Tarlé, Elise Bonzon, and Nicolas Maudet. 2022. Multiagent Dynamics of Gradual Argumentation Semantics. In *AAMAS*. 363–371. https://doi.org/10.5555/3535850.3535892

[13] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artif. Intell.* 77, 2 (1995), 321–358. https://doi.org/10.1016/0004-3702(94)00041-X

[14] Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*. 12873–12883. https://doi.org/10.1109/CVPR46437.2021.01268

[15] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. In *ICML*. 2376–2384. http://proceedings.mlr.press/v97/goyal19a.html

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *CVPR*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[18] Geoffrey Irving, Paul F. Christiano, and Dario Amodei. 2018. AI safety via debate. *CoRR* abs/1805.00899 (2018). arXiv:1805.00899 http://arxiv.org/abs/1805.00899

[19] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *ICLR*. https://openreview.net/forum?id=rkE3y85ee

[20] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. 4401–4410. https://doi.org/10.1109/CVPR.2019.00453

[21] Eoin M. Kenny, Eoin Delaney, and Mark T. Keane. 2023. Advancing Post-Hoc Case-Based Explanation with Feature Highlighting. In *IJCAI*. 427–435. https://doi.org/10.24963/IJCAI.2023/48

[22] Avinash Kori, Ben Glocker, and Francesca Toni. 2022. GLANCE: Global to Local Architecture-Neutral Concept-based Explanations. *CoRR* abs/2207.01917 (2022). https://doi.org/10.48550/ARXIV.2207.01917 arXiv:2207.01917

[23] Avinash Kori, Ben Glocker, and Francesca Toni. 2024. Explaining Image Classifiers with Visual Debates. In *DS*. 200–214. https://doi.org/10.1007/978-3-031-78980-9_13

[24] Avinash Kori, Francesco Locatello, Fabio De Sousa Ribeiro, Francesca Toni, and Ben Glocker. 2024. Grounded Object-Centric Learning. In *ICLR*. https://openreview.net/forum?id=pBxeZ6pVUD

[25] Avinash Kori, Antonio Rago, and Francesca Toni. 2025. Free Argumentative Exchanges for Explaining Image Classifiers. *CoRR* abs/2502.12995 (2025). https://doi.org/10.48550/ARXIV.2502.12995 arXiv:2502.12995

[26] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *CoRR* abs/2202.01875 (2022). arXiv:2202.01875 https://arxiv.org/abs/2202.01875

[27] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS*. 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[28] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *AAMAS*. 1033–1041. http://dl.acm.org/citation.cfm?id=3331801

[29] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38. https://doi.org/10.1016/J.ARTINT.2018.07.007

[30] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. In *NIPS*. 2204–2212. https://proceedings.neurips.cc/paper/2014/hash/09c6c3783b4a70054da74f2538ed47c6-Abstract.html

[31] Nico Potyka. 2021. Interpreting Neural Networks as Quantitative Argumentation Frameworks. In *AAAI*. 6463–6470. https://doi.org/10.1609/AAAI.V35I7.16801

[32] Antonio Rago, Pietro Baroni, and Francesca Toni. 2022. Explaining Causal Models with Argumentation: the Case of Bi-variate Reinforcement. In *KR*. https://proceedings.kr.org/2022/52/

[33] Antonio Rago, Hengzhi Li, and Francesca Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *KR*. 582–592. https://doi.org/10.24963/kr.2023/57

[34] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *SIGKDD*. 1135–1144. https://doi.org/10.1145/2939672.2939778

[35] Ainkaran Santhirasekaram, Avinash Kori, Mathias Winkler, Andrea G. Rockall, and Ben Glocker. 2022. Vector Quantisation for Robust Segmentation. In *MICCAI*. 663–672. https://doi.org/10.1007/978-3-031-16440-8_63

[36] Ainkaran Santhirasekaram, Avinash Kori, Mathias Winkler, Andrea G. Rockall, Francesca Toni, and Ben Glocker. 2023. Robust Hierarchical Symbolic Explanations in Hyperbolic Space for Image Classification. In *CVPR*. 561–570. https://doi.org/10.1109/CVPRW59228.2023.00063

[37] Sam Sattarzadeh, Mahesh Sudhakar, Konstantinos N. Plataniotis, Jongseong Jang, Yeonjeong Jeong, and Hyunwoo Kim. 2021. Integrated Grad-Cam: Sensitivity-Aware Visual Explanation of Deep Convolutional Networks Via Integrated Gradient-Based Scoring. In *ICASSP*. 1775–1779. https://doi.org/10.1109/ICASSP39728.2021.9415064

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[39] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. 2021. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *NeurIPS*. 11352–11363. https://proceedings.neurips.cc/paper/2021/hash/5e751896e527c862bf67251a474b3819-Abstract.html

[40] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *ICML*. 3145–3153. http://proceedings.mlr.press/v70/shrikumar17a.html

[41] Purin Sukpanichnant, Antonio Rago, Piyawat Lertvittayakumjorn, and Francesca Toni. 2021. Neural QBAFs: Explaining Neural Networks Under LRP-Based Argumentation Frameworks. In *AIxIA*. 429–444. https://doi.org/10.1007/978-3-031-08421-8_30

[42] Dao Thauvin, Stéphane Herbin, Wassila Ouerdane, and Céline Hudelot. 2024. Interpretable Image Classification Through an Argumentative Dialog Between Encoders. In *ECAI*. 3316–3323. https://doi.org/10.3233/FAIA240880

[43] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *NIPS*. 6306–6315. https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[45] Pei Wang and Nuno Vasconcelos. 2019. Deliberative Explanations: visualizing network insecurities. In *NeurIPS*. 1372–1383. https://proceedings.neurips.cc/paper/2019/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html

[46] Weiyan Xie, Xiao-Hui Li, Zhi Lin, Leonard K. M. Poon, Caleb Chen Cao, and Nevin L. Zhang. 2023. Two-stage holistic and contrastive explanation of image classification. In *UAI*. 2335–2345. https://proceedings.mlr.press/v216/xie23a.html

[47] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2019. INVASE: Instance-wise Variable Selection using Neural Networks. In *ICLR*. https://openreview.net/forum?id=BJg_roAcK7