MULTIPLE-INSTANCE LEARNING AS A FRAMEWORK TO EXPLAIN WITH SHAPLEY COEFFICIENTS

by

Jacopo Teneggi

A thesis submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Master of Science in Engineering

Baltimore, Maryland

May, 2022

Abstract

Explainability and interpretability have become questions of fundamental importance for a safe and responsible deployment of modern machine learning models in high-stakes scenarios. Many examples exist of accidental behavior of autonomous systems that systematically under perform on minorities, or emulate hateful human behavior. Notwithstanding the recent advances in fair and interpretable machine learning, several theoretical issues remain open on the validity of popular explanation methods.

In this thesis, we study multiple-instance learning as a framework to explaining model predictions with Shapley coefficients. In particular, we focus on *local* explanations, i.e. we seek to find the most important features in an input towards a model's prediction. We show that a principle approach to explainability can produce fast and exact explanation methods that provide precise mathematical guarantees on their speed and accuracy. We apply our new explanation method to a medical imaging task of clinical importance–intracranial hemorrhage detection–where the use of autonomous systems can support radiologists in their daily work, for example, by prioritizing the most severe cases or provide a second opinion for subtle ones. We find that an explainability-driven approach can significantly reduce the number of labels

needed to train a model, and therefore make collecting new datasets cheaper.

Primary readers: Jeremias Sulam (Advisor), Soledad Villar, and Adam Charles.

Acknowledgments

This thesis would have been impossible to achieve on my own, as it is impossible to list all the people who helped me reach this goal. I am extremely grateful to my advisor, Jeremias, for his tireless support and deep kindness. It is a unique privilege to be able to work closely and learn from such an inspiring role model. I want to thank Beepul, Zhenzhen, Zhenghan, Ramchandran, and Ambar for their constant excitement, motivation, and advice. Next, I want to thank Teresa, Phillip, Tommy, and Shreya for their friendship and for making JHU feel like home. Then, I want to thank Edoardo, Enrico, Meike, and Linda for their continuous, remote support. I want to thank my parents, my siblings, and my family for their unconditional love and for pushing me to pursue my dreams. Last but certainly not least, I want to thank my readers, Soledad, and Adam, for their constructive feedback. I see this thesis as a stepping stone to my PhD and I cannot wait to see what the next few years ahead hold!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Explainability has become a question of increasing relevance in machine learning, where the growing complexity of deep neural networks often renders them *opaque* to us, the humans interacting with them. This issue is commonly referred to as the *black-box problem* and comprises theoretical, technical, and regulatory questions (Zednik, 2019; Tomsett et al., 2018). As deep neural networks are applied to sensitive tasks in medical, legal, and financial settings, they need to achieve both high accuracy and high transparency for a responsible, fair, and trustworthy deployment in real-world scenarios. For example, uninterpretable predictions could mislead clinicians in their decision making rather than support it (Amann et al., 2020). Furthermore, it is sometimes required by law (Kaminski, 2019) to provide an explanation of how data lead an automated algorithm, for example, to reject a loan application (Kaminski, 2019; Kaminski and Malgieri, 2019; Hacker et al., 2020). Finally, opaque models can conceal dataset bias, and lead to socially unfair models (Shin, 2021).

In this thesis, we are particularly interested in local model explanations.

Given a model's prediction on a specific input, we look for the features that contributed the most towards it. That is, these methods act *locally* on individual predictions rather than globally across a dataset. We focus on supervised learning scenarios where we would like for a model to learn a target concept, e.g. an object or some medical findings related to a diagnosis. The goal of explainability in these settings is to first make sure a model has actually learned the target concept, and potentially gain further insights. For example, assume one has a model that predicts the presence of brain tumor in MRI scans with very high accuracy. What are the most relevant morphological features that indicate the presence of tumor, and where are they located? Can we discover new features of the disease from what the model has learned? Many important problems of this kind exist, but the necessary tools to answer these questions effectively and efficiently are still lacking.

The foundational work by Ribeiro, Singh, and Guestrin (2016a) spurred exciting advances in local feature attribution methods, such as Grad-CAM (Selvaraju et al., 2017), Integrated Gradients (Sundararajan, Taly, and Yan, 2017), and DeepLIFT (Shrikumar, Greenside, and Kundaje, 2017a). Lundberg and Lee (2017a) provide a unified framework for several different approaches under their SHAP method, which leverages Shapley coefficients–a game-theoretic measure (Shapley, 1953)–and feature removal strategies. Unlike other perturbation-based alternatives (Shah, Jain, and Netrapalli, 2021), these methods enjoy of important consistency results and theoretical properties that the resulting attributions satisfy *blue* by framing interpretability in terms of game theory. Since then, a plethora of different explanation methods has been

developed[1] for different kinds of data (tabular, sequential, imaging), both based on Shapley coefficients (Chen et al., 2018) as well as other information theoretic quantities (MacDonald et al., 2019; Heiß et al., 2020; Merrick and Taly, 2020).

Here, we focus on problems that satisfy a certain *multiple-instance learning* (MIL) assumption (Dietterich, Lathrop, and Lozano-Pérez, 1997; Weidmann, Frank, and Pfahringer, 2003), which can be found in many relevant fields. In MIL scenarios, input samples are regarded as *bags* of *instances*. For example, in computer vision, an image can be considered as a bag of regions, while in medical image analysis, a volumetric Computed Tomography (CT) scan can be seen as a bag of images. Furthermore, we assume that the label of a bag is a known deterministic function of the labels of the instances. In its simplest form: binary classification, the MIL assumption implies that the bag of a label is the logical `OR` of the labels of the instances. That is, a bag is labeled positively if and only if it contains a positive instance. Multiple-instance learning is a generalization of the classic supervised learning framework. Importantly, it provides a weaker sense of supervision: an MIL learner does not have access to the individual instance labels, and it can only learn from global, bag labels. We show that a principled approach to explainability can yield fast and exact methods for a variety of MIL problems, achieving similar or better performance compared to models trained with full-supervision in a medical imaging tasks.

We now introduce the necessary background information.

[1] To our knowledge, Covert, Lundberg, and Lee, 2021 compiled the most comprehensive review of currently available explanation methods based on feature removal.

1.1 Supervised learning

In the standard supervised learning framework, given input and output domains $\mathcal{X}$ and $\mathcal{Y}$, we are interested in predicting a response $Y \in \mathcal{Y}$ on a new input $X \in \mathcal{X}$. Hence, we search for a good predictor $\hat{h}$ in $\mathcal{H}$–a suitable family of hypotheses from $\mathcal{X}$ to $\mathcal{Y}' \supseteq \mathcal{Y}$. Herein, we will assume realizability–there exists a function $h^* \in \mathcal{H}$ such that $Y = h^*(X)$. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and let $\ell : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}^+$ be a loss function that penalizes differences between the true response and the predicted one. Then, an optimal predictor h has minimal risk $R = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(Y, h(X))]$. However, $\mathcal{D}$ is unknown in most real-world scenarios, and we search for $\hat{h}$ by minimizing the empirical risk on a training set $\{(X_i, Y_i)\}_{i=1}^m \sim \mathcal{D}^m$:

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{m} \ell(Y_i, h(X_i)), \tag{1.1}$$

where $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$ is a family of hypotheses parametrized by θ that encodes our prior knowledge on the problem. For example, $\mathcal{H}$ can be the class of Convolutional Neural Networks (CNNs) for image classification tasks.

1.2 Multiple-Instance Learning

Multiple-Instance Learning (MIL) (Dietterich, Lathrop, and Lozano-Pérez, 1997; Maron and Lozano-Pérez, 1997; Weidmann, Frank, and Pfahringer, 2003) generalizes the supervised learning framework to *bags* of inputs. Formally, let $\overline{X} = (X_1, \ldots, X_r) \in \mathcal{X}^r$, $r \in \mathbb{N}$ be a sequence of inputs, and denote $\overline{X}$ a *bag* of size r. Then, we refer to $X_1, \ldots, X_r$ as *instances*, whose order may or

may not matter depending on the specific MIL assumption. Furthermore, the bag-level response $\overline{Y} \in \mathcal{Y}$ is assumed to be a known deterministic function of the instance-level responses. That is, let $\phi : (\mathcal{Y}')^r \to \mathcal{Y}'$ be a *bag-pooling function*, then

$$\overline{Y} = (\phi \circ (h^*)^r)(\overline{X}) \coloneqq \phi(h^*(X_1), \ldots, h^*(X_r)). \tag{1.2}$$

We remark that h^* is an *instance-level* hypothesis defined on a single instance as in the classical supervised learning setting. Eq. (1.2) highlights that the true bag-level response is obtained by applying the same instance-level hypothesis to all instances in a bag. In the context of MIL literature, this setting is usually referred to as *homogeneous* (Sabato and Tishby, 2009), i.e. the bag-level response is invariant under permutations on the instances. Importantly, an MIL learner has access to bag-level labels only. Let $\overline{\mathcal{D}}$ be a distribution over $\mathcal{X}^r \times \mathcal{Y}$, $\{(\overline{X}_i, \overline{Y}_i)\}_{i=1}^m \sim \overline{\mathcal{D}}$ be a set of bags, then

$$\hat{h}_{\text{MIL}} = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{m} \ell(\overline{Y}_i, (\phi \circ h^r)(\overline{X}_i)). \tag{1.3}$$

The MIL framework encompasses a wide range of problems by making different assumptions on the bag distribution $\overline{\mathcal{D}}$, the response domain $\mathcal{Y}$, the target concept h^*, and the bag-pooling function ϕ (Blum and Kalai, 1998; Auer, Long, and Srinivasan, 1998; Andrews, Tsochantaridis, and Hofmann, 2002; Sabato and Tishby, 2009; Sabato and Tishby, 2012). For example, independent, binary classification MIL problems satisfy $\overline{\mathcal{D}} = \mathcal{D}^r$, $\mathcal{Y} = \{0, 1\}$, $\phi = \texttt{OR}$, and

$$\overline{Y} = 1 \iff \exists i \in [r] : \; h^*(X_i) = 1, \tag{1.4}$$

where $[r] := \{1, \ldots, r\}$. That is, a bag is labeled positively if and only if there is at least one positive instance in the bag.

Broadly speaking, current MIL models can be divided into two main groups: instance-level classifiers, and bag-level classifiers (Amores, 2013; Wang et al., 2018; Cheplygina, Bruijne, and Pluim, 2019; Quellec et al., 2017). The former models set out to learn good instance classifiers and then use them to predict bag labels via an *a-priori* bag-pooling function. That is, since the pooling function ϕ is known for the problem, they try to learn good instance-level hypothesis to plug into ϕ. On the other hand, the latter models seek to learn good bag representations and to classify them. Indeed, bag-level classifiers do not usually estimate instance-level responses directly, and instead learn a classifier (e.g. a fully-connected layer) on top of bag representations which are computed from instance representations. In this thesis, we study bag-level MIL classification problems.

1.3 Bag-level MIL binary classification.

Bag-level classifiers solve a more general version of Eq. (1.3). Assume the instance domain $\mathcal{X}$ has dimension D, and let $\mathcal{F}$ be a feature space of dimension $K \ll D$. Then, $\mathcal{H}_{\mathcal{F}} \subseteq \mathcal{F}^{\mathcal{X}}$ is a suitable family of hypotheses from the instance domain $\mathcal{X}$ to the feature space $\mathcal{F}$. Furthermore, let $\varphi : \mathcal{F}^r \to \mathcal{Y}'$ be a generalized bag-pooling function, then $h^r : \mathcal{X}^r \to \mathcal{F}^r$ and

$$\hat{h}_{\text{bag}-\text{MIL}} = \underset{\varphi,\, h \in \mathcal{H}_{\mathcal{F}}}{\arg\min} \sum_{i=1}^{m} \ell(\overline{Y}_i, (\varphi \circ h^r)(\overline{X}_i)). \tag{1.5}$$

Note that differently from Eq. (1.3), the optimization problem above is over both the family of hypotheses $\mathcal{H}_{\mathcal{F}}$ and the generalized bag-pooling function φ. Several approaches have been proposed to aggregate instance representations into bag representations (Keeler, Rumelhart, and Leow, 1990; Maron and Lozano-Pérez, 1997; Ramon and De Raedt, 2000; Kraus, Ba, and Frey, 2016; Ilse, Tomczak, and Welling, 2018). Importantly, as noted by (Ilse, Tomczak, and Welling, 2018), the MIL assumption in Eq. (1.4) requires for the bag-pooling function φ to be permutation invariant, i.e. independent of the the order of the instances. This property is satisfied if and only if

$$(\varphi \circ h^r)(\overline{X}) = g(z(\overline{X})), \quad z(\overline{X}) \coloneqq \sum_{X \in \overline{X}} f(h(X)), \tag{1.6}$$

for two appropriate transformations f, g (see Theorem 1 in (Ilse, Tomczak, and Welling, 2018)). A natural choice of f is, for example, $f(h(X)) = 1/r \cdot h(X)$, such that $z(\overline{X})$ is the mean instance representation, and g is a bag classifier. As noted in (Sabato and Tishby, 2012), we remark that the definition in Eq. (1.6) comprises well-known real-valued permutation invariant functions, such as the max. Indeed, $\forall p \in [1, \infty)$ define the bag p-norm as

$$\overline{\ell}_p(\overline{X}) \coloneqq \left(\frac{1}{r} \cdot \sum_{X \in \overline{X}} (h(X) + 1)^p \right)^{1/p} - 1. \tag{1.7}$$

Then, $\max(\overline{X}) \equiv \lim_{p \to \infty} \bar{l}_p(\overline{X})$ (see Definition 3 in (Sabato and Tishby, 2012)).

1.4 Bag classification via attention mechanisms

In the context of computational linguistics, machine translation is the task of transforming an input sentence from a source language into its equivalent in a target language. Attention mechanisms were introduced for Neural Machine Translation (NMT) models, which comprise deep-learning approaches to sequence to sequence translation problems (Bahdanau, Cho, and Bengio, 2014; Luong, Pham, and Manning, 2015; Kim et al., 2017). In particular, NMT systems are trained to model the conditional probability of the output sentence in the target language given the input sentence in the source language. Most NMT architectures comprise two components: (i) an encoder, which learns representations of the input sentences, and (ii) a decoder, which sequentially generates the target words. Intuitively, attention mechanisms were introduced to *align* the target and input domains by means of a *context* vector–a weighted sum of the input representations, where the weights are the output of the attention mechanism. The practical effectiveness of such mechanisms has lead to the development of several variations of the original architectures (Vaswani et al., 2017; Mishra et al., 2017; Zhang et al., 2019). Most notably, (Vaswani et al., 2017) introduced the Transformer architecture–a sequence to sequence model solely composed of self-attention mechanisms, which has lead to significant improvements in the Natural Language Processing (NLP) field (Devlin et al., 2018; Liu et al., 2019; Shoeybi et al., 2019; Clark et al., 2020; Gu et al., 2021; Fedus, Zoph, and Shazeer, 2021), and whose use is now being explored in the context of computer vision (Dosovitskiy et al., 2020; Touvron et al., 2021) and multimodal learning (Radford et al., 2021; Ramesh et al., 2021)

as well. Notwithstanding the widespread adoption of this type of architecture, we still lack of analytical results to fully understand its behavior (Edelman et al., 2021; Vidal, 2021).

We remark that ideas of trainable, attention-like mechanisms were also introduced around the same time in the Weakly Supervised Object Detection (WSOD) community (Bilen and Vedaldi, 2016) in order to rank region proposals by their classification and detection scores. Indeed, WSOD can be phrased as a multiple-instance learning problem. For example, assume we are given a dataset of labeled images, where the labels indicate the presence of certain objects in the images. Then, we are interested in detecting those objects without access to ground truth annotations, i.e. without knowing where the objects are located in the images. We can regard every image as a bag of region proposals, and the image label indicates whether any of the proposals contains the target objects. Naturally, this is a multiclass extension of the MIL binary classification problem presented in Sec. 1.2. Ilse, Tomczak, and Welling (2018) were the first to explicitly link attention mechanisms and bag-level multiple-instance learning classification, in order to overcome some of the limitations of fixed bag-pooling transformations. Indeed, they propose to replace f in Eq. (1.6) with a trainable, two-layer attention mechanism such that

$$z(\overline{X}) = Ha, \quad a = \sigma\big(\langle w, \tanh(VH)\rangle\big)^\top, \tag{1.8}$$

where $H := [h(X_1), \ldots, h(X_r)] \in \mathbb{R}^{K\times r}$ are the instance representations, $a \in \mathbb{R}^r$ are the attention weights, $V \in \mathbb{R}^{L\times K}$, $w \in \mathbb{R}^{L\times 1}$ are learned, and $\sigma : \mathbb{R}^r \to \Delta^{r-1}$ is some normalization function such that the attention weights sum up

to 1 (e.g. softmax). We note that the attention mechanism in Eq. (1.8) can be seen as a weighted sum of the instance-level feature representations, hence it is permutation invariant.

1.5 Explaining model predictions with Shapley coefficients

In game theory (Owen, 2013), we define a cooperative n-person *Transfer Utility* (TU) game with a pair $([n], v)$, where $[n] := \{1, \ldots, n\}$ is a set of n players, and $v : \ 2^{[n]} \to \mathbb{R}^+$ is the *characteristic function* of the game, which assigns a nonnegative score $v(S)$ to any nonempty coalition of players $S \subseteq [n]$. Furthermore, we usually let $v(\varnothing) = 0$. In particular, in a TU game, players can exchange parts of their utility without incurring in any penalty. As an example, consider the dummy characteristic function $v(C) = \sum_{i \in C} i$ which simply returns the sum of the indices of the players in the coalition C. Then, if $n = 2$ and there are only two players in the game, i.e. $[n] = \{1, 2\}$, we have

$$v(\varnothing) = 0, \quad v(\{1\}) = 1, \quad v(\{2\}) = 2, \quad v(\{1, 2\}) = 3. \tag{1.9}$$

One of the goals of game theory is to study *solution concepts*, i.e. formal rules that describe the strategy that each player will employ in the game. Let γ_j denote the Shapley value of the j-th player in the game. Then, the Shapley value (Shapley, 1953) is the only solution concept that satisfies the following axioms:

1. **Efficiency** In a TU game $([n], v)$, the sum of the Shapley values is equal

to the score of the *grand coalition* $[n]$:

$$\sum_{j \in [n]} \gamma_j = v([n]). \tag{1.10}$$

2. **Nullity** If player j does not contribute to any coalition $S \subseteq [n]$ in a TU game $([n], v)$, then its Shapley value is 0:

$$\forall S \subseteq [n],\ v(C \cup \{j\}) - v(C) = 0 \implies \gamma_j = 0. \tag{1.11}$$

3. **Symmetry** Given two players j, k in a TU game $([n], v)$, if their contributions to any coalition $S \subseteq [n]$ are the same, their Shapley values are the same:

$$\forall S \subseteq [n] \setminus \{j, k\},\ v(C \cup \{j\}) = v(C \cup \{k\}) \implies \gamma_j = \gamma_k. \tag{1.12}$$

4. **Additivity** Given two TU games $([n], v), ([m], v)$ with the same characteristic function, the Shapley value of the sum of the two games is equal to the sum of the Shapley values of the individual games:

$$\gamma_j^{[n] \cup [m]} = \gamma_j^{[n]} + \gamma_j^{[m]}, \tag{1.13}$$

where $\gamma_j^{[n] \cup [m]}$ denotes the Shapley value of the TU game $([n] \cup [m], v)$, and $\gamma_j^{[n]}, \gamma_j^{[m]}$ of $([n], v)$ and $([m], v)$, respectively.

We note that, formally, the Shapley value also satisfies a fifth property, usually referred to as *balanced contribution*, but axioms 1–4 are sufficient to derive the

exact formulation, which can be expressed as

$$\gamma_j = \sum_{S \subseteq [n] \setminus \{j\}} w_S \cdot [v(S \cup \{j\}) - v(S)]\,, \tag{1.14}$$

where $w_S = |S|! \cdot (n - |S| - 1)!/n!$ is an appropriate constant that depends on the size of the subset S and the number of players n. That is, γ_j is the averaged marginalized contribution of the j-th player over all possible permutations of players, i.e. over all possible coalitions of players playing the same game. We remark that the cost of computing the exact Shapley coefficients as defined in Eq. (1.14) is exponential in the number of players, which quickly renders them intractable in practical scenarios where the number of players may be large.

The SHAP framework (Lundberg and Lee, 2017b) translates Shapley coefficients to machine learning models, and it unifies previously existing explanation methods (Ribeiro, Singh, and Guestrin, 2016b; Shrikumar, Greenside, and Kundaje, 2017b). The overall idea is to consider a cooperative game where the players are the features in the input, and the characteristic function is some trained model. More precisely, let $h : \mathbb{R}^n \to \mathbb{R}$ be a model in an appropriate family of hypotheses $\mathcal{H}$. Given a new input $X \in \mathbb{R}^n$, we can interpret the model prediction $h(X)$ as the score of a TU game with n players–the features in X. Then, we can compute the score of coalitions of players $C \in [n]$ by masking the features not in C and predicting with the model h. Formally, consider the game $([n], h)$, then

$$\widetilde{\gamma}_j = \sum_{S \subseteq [n] \setminus \{j\}} w_S \cdot \Big[h(X_{S \cup \{j\}}) - h(X_S)\Big] \tag{1.15}$$

is the Shapley coefficient of the j-th feature in X. The most notable difference

between the original definition of the Shapley value in Eq.(1.14) and the machine learning interpretation above, is that the hypothesis h has fixed domain $\mathbb{R}^n$, hence it cannot predict on inputs of arbitrary size. Then, one has to devise ways to *mask* features in the input by replacing them with some value that does not contribute to the prediction. Indeed, let $X = (x_1, \dots, x_n)$, and $\forall A \subseteq [n]$ define

$$X_A \coloneqq \begin{cases} x_i & i \in A \\ b_i & \text{elsewhere,} \end{cases} \tag{1.16}$$

where $B = (b_1, \dots, b_n) \in \mathbb{R}^{n-|A|}$ is an *uninformative baseline*. That is, X_A is the same as X in the entries in A, and some reference value elsewhere. Then, one should consider how to determine the value of B. Given a new input X and a fixed subset of features $A \subseteq [n]$, a natural choice is to sample B from the (observational) conditional distribution of the features not in A, such that the model prediction is by definition independent of B (as it is done in Lundberg and Lee (2017a)). With abuse of notation, given a new sample $X = (x_1, \dots, x_n)$, assume that the exact conditional distribution of the data is known, and let

$$B \sim X_{-A} \mid X_A = x_A, \quad \widetilde{h}(X_A) = \mathbb{E}_B \left[h(X_A) \right], \tag{1.17}$$

where X_{-A} indicates the subset of features not in A. Then, the Shapley coefficients in Eq. (1.15) become random variables such that

$$\widetilde{\gamma}_j = \sum_{S \subseteq [n] \setminus \{j\}} w_S \cdot \left[\widetilde{h}(X_{S \cup \{j\}}) - \widetilde{h}(X_S) \right]. \tag{1.18}$$

The equation above highlights the second major hurdle to overcome when using Shapley coefficients–how to approximate the conditional distribution of B to estimate $\widetilde{h}$. As a result, all state-of-the-art image explanation methods

based on Shapley coefficients rely on some approximation strategy to work around both the exponential computational complexity of computing the exact Shapley coefficients and the approximation of the distribution of the baseline value, for example, using generative models. For instance, GradientExplainer (Lundberg and Lee, 2017a) extends Integrated Gradients (Sundararajan, Taly, and Yan, 2017) by sampling multiple references from the background dataset to integrate on. Similarly, DeepExplainer (Lundberg and Lee, 2017a; Chen, Lundberg, and Lee, 2021) builds upon DeepLIFT (Shrikumar, Greenside, and Kundaje, 2017a) by choosing a per-node attribution rule that can approximate Shapley coefficients when integrated over many background samples. We refer the reader to the review by Covert, Lundberg, and Lee (2021) which, to our knowledge, is the most recent survey of feature-removal-based explanation methods, both using Shapley coefficients and not. We note that current methods that try to estimate the conditional distribution of B only provide consistency results and lack finite-sample results, making it hard to understand in practice how close their results are to the true Shapley coefficients, especially in high-dimensional scenarios. Lastly, we note that under feature independence and model linearity, the conditional expectations over the model in Eq. (1.18) can be replaced with unconditional expectations over the data. That is, one can use a fixed reference value equal to the average input over the data to mask features. Although these assumptions are hardly satisfied with machine learning models, which are highly nonlinear, masking features with their unconditional expectation works well in practice in many real-world applications. We will motivate this choice more precisely in Chapter 2 for our novel explanation method.

References

Zednik, Carlos (2019). "Solving the black box problem: a normative framework for explainable artificial intelligence". In: *Philosophy & Technology*, pp. 1–24.

Tomsett, Richard, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty (2018). "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems". In: *arXiv preprint arXiv:1806.07552*.

Amann, Julia, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai (2020). "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective". In: *BMC Medical Informatics and Decision Making* 20.1, pp. 1–9.

Kaminski, Margot E (2019). "The right to explanation, explained". In: *Berkeley Tech. LJ* 34, p. 189.

Kaminski, Margot E and Gianclaudio Malgieri (2019). "Algorithmic impact assessments under the GDPR: producing multi-layered explanations". In: *U of Colorado Law Legal Studies Research Paper* 19-28.

Hacker, Philipp, Ralf Krestel, Stefan Grundmann, and Felix Naumann (2020). "Explainable AI under contract and tort law: legal incentives and technical challenges". In: *Artificial Intelligence and Law*, pp. 1–25.

Shin, Donghee (2021). "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI". In: *International Journal of Human-Computer Studies* 146, p. 102551.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). ""' 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*, pp. 1135–1144.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International Conference on Machine Learning*. PMLR, pp. 3319–3328.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017a). "Learning important features through propagating activation differences". In: *International Conference on Machine Learning*. PMLR, pp. 3145–3153.

Lundberg, Scott and Su-In Lee (2017a). "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874*.

Shapley, Lloyd S (1953). "A value for n-person games". In: *Contributions to the Theory of Games* 2.28, pp. 307–317.

Shah, Harshay, Prateek Jain, and Praneeth Netrapalli (2021). "Do Input Gradients Highlight Discriminative Features?" In: *arXiv preprint arXiv:2102.12781*.

Covert, Ian, Scott Lundberg, and Su-In Lee (2021). "Explaining by removing: A unified framework for model explanation". In: *Journal of Machine Learning Research* 22.209, pp. 1–90.

Chen, Jianbo, Le Song, Martin J Wainwright, and Michael I Jordan (2018). "L-shapley and c-shapley: Efficient model interpretation for structured data". In: *arXiv preprint arXiv:1808.02610*.

MacDonald, Jan, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok (2019). "A rate-distortion framework for explaining neural network decisions". In: *arXiv preprint arXiv:1905.11092*.

Heiß, Cosmas, Ron Levie, Cinjon Resnick, Gitta Kutyniok, and Joan Bruna (2020). "In-Distribution Interpretability for Challenging Modalities". In: *arXiv preprint arXiv:2007.00758*.

Merrick, Luke and Ankur Taly (2020). "The Explanation Game: Explaining Machine Learning Models Using Shapley Values". In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, pp. 17–38.

Dietterich, Thomas G, Richard H Lathrop, and Tomás Lozano-Pérez (1997). "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial intelligence* 89.1-2, pp. 31–71.

Weidmann, Nils, Eibe Frank, and Bernhard Pfahringer (2003). "A two-level learning method for generalized multi-instance problems". In: *European Conference on Machine Learning*. Springer, pp. 468–479.

Maron, Oded and Tomás Lozano-Pérez (1997). "A framework for multiple-instance learning". In: *Advances in neural information processing systems* 10.

Sabato, Sivan and Naftali Tishby (2009). "Homogeneous Multi-Instance Learning with Arbitrary Dependence." In: *COLT*. Citeseer.

Blum, Avrim and Adam Kalai (1998). "A note on learning from multiple-instance examples". In: *Machine learning* 30.1, pp. 23–29.

Auer, Peter, Philip M Long, and Aravind Srinivasan (1998). "Approximating hyper-rectangles: learning and pseudorandom sets". In: *Journal of Computer and System Sciences* 57.3, pp. 376–388.

Andrews, Stuart, Ioannis Tsochantaridis, and Thomas Hofmann (2002). "Support vector machines for multiple-instance learning". In: *Advances in neural information processing systems* 15.

Sabato, Sivan and Naftali Tishby (2012). "Multi-instance learning with any hypothesis class". In: *The Journal of Machine Learning Research* 13.1, pp. 2999–3039.

Amores, Jaume (2013). "Multiple instance classification: Review, taxonomy and comparative study". In: *Artificial intelligence* 201, pp. 81–105.

Wang, Xinggang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu (2018). "Revisiting multiple instance neural networks". In: *Pattern Recognition* 74, pp. 15–24.

Cheplygina, Veronika, Marleen de Bruijne, and Josien PW Pluim (2019). "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis". In: *Medical image analysis* 54, pp. 280–296.

Quellec, Gwenolé, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard (2017). "Multiple-instance learning for medical image and video analysis". In: *IEEE reviews in biomedical engineering* 10, pp. 213–234.

Keeler, James, David Rumelhart, and Wee Leow (1990). "Integrated segmentation and recognition of hand-printed numerals". In: *Advances in neural information processing systems* 3.

Ramon, Jan and Luc De Raedt (2000). "Multi instance neural networks". In: *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pp. 53–60.

Kraus, Oren Z, Jimmy Lei Ba, and Brendan J Frey (2016). "Classifying and segmenting microscopy images with deep multiple instance learning". In: *Bioinformatics* 32.12, pp. i52–i59.

Ilse, Maximilian, Jakub Tomczak, and Max Welling (2018). "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR, pp. 2127–2136.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning (2015). "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025*.

Kim, Yoon, Carl Denton, Luong Hoang, and Alexander M Rush (2017). "Structured attention networks". In: *arXiv preprint arXiv:1702.00887*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Mishra, Nikhil, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel (2017). "A simple neural attentive meta-learner". In: *arXiv preprint arXiv:1707.03141*.

Zhang, Han, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena (2019). "Self-attention generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 7354–7363.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Shoeybi, Mohammad, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro (2019). "Megatron-lm: Training multi-billion parameter language models using model parallelism". In: *arXiv preprint arXiv:1909.08053*.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning (2020). "Electra: Pre-training text encoders as discriminators rather than generators". In: *arXiv preprint arXiv:2003.10555*.

Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon (2021). "Domain-specific language model pretraining for biomedical natural language processing". In: *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1, pp. 1–23.

Fedus, William, Barret Zoph, and Noam Shazeer (2021). "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". In: *arXiv preprint arXiv:2101.03961*.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929*.

Touvron, Hugo, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou (2021). "Going deeper with image transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. (2021). "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever (2021). "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR, pp. 8821–8831.

Edelman, Benjamin L, Surbhi Goel, Sham Kakade, and Cyril Zhang (2021). "Inductive Biases and Variable Creation in Self-Attention Mechanisms". In: *arXiv preprint arXiv:2110.10090*.

Vidal, Rene (2021). "Attention: Self-Expression Is All You Need". In.

Bilen, Hakan and Andrea Vedaldi (2016). "Weakly supervised deep detection networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2846–2854.

Owen, Guillermo (2013). *Game theory*. Emerald Group Publishing.

Lundberg, Scott M and Su-In Lee (2017b). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016b). """ 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*, pp. 1135–1144.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017b). "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMLR, pp. 3145–3153.

Chen, Hugh, Scott Lundberg, and Su-In Lee (2021). "Explaining models by propagating Shapley values of local components". In: *Explainable AI in Healthcare and Medicine*. Springer, pp. 261–270.

Chapter 2

Fast Hierarchical Games for Image Explanations

Notwithstanding the recent advances in image attribution methods based on Shapley coefficients, several limitations hinder their use for "large" images–a standard image contains $\approx 10^6$ pixels, and larger images are used in several important applications. Although previous work explores structured and hierarchical approaches (Chen, Zheng, and Ji, 2020; Chen et al., 2018; Singh, Murdoch, and Yu, 2018), they remain limited for high-dimensional data. We focus on problems that satisfy a certain *multiple instance learning* assumption (Dietterich, Lathrop, and Lozano-Pérez, 1997), which can be found in many relevant fields. Following Sec. 1.2, let $f^* : \mathcal{X} \to \mathcal{Y}$ be the true target concept from the input domain $\mathcal{X}$ into some response domain $\mathcal{Y}$. Then, the classical binary classification MIL assumption implies that

$$f^*(X) = 1 \iff \exists\, C \subseteq [n] : \; f^*(X_C) = 1. \tag{2.1}$$

We show that in these problems, the computation of Shapley coefficients can be solved efficiently and without the need of approximation by exploring a

hierarchical partition of the input image. The contribution of this chapter is three-fold: first, we present a fast explanation method based on Shapley coefficients that is exponentially faster than popular SHAP methods. Second, under some distributional assumptions similar to those in multiple instance learning problems, we show that the coefficients provided by our novel explanation method, h-Shap, are exact, and can be further approximated in a controlled manner by trading off computational cost. Third, we compare h-Shap with other popular explanation methods on three benchmarks, of varied complexity and dimension, demonstrating that h-Shap outperforms the state of the art both in terms of runtime and retrieval of relevant features in all experiments. The content of this chapter received a Best Paper Award at the ICML 2021 Workshop on Interpretable Machine Learning in Healthcare and it is currently under review for publication in the IEEE Transactions on Pattern Analysis and Machine Intelligence.

Let $g = (X, f, [n])$ be an n-person cooperative game with players $[n]$ and characteristic function $f : \mathcal{X} \to \mathbb{R}$ which maps the input space $\mathcal{X}$ to a score. In particular, $f(X_C)$ is the score that the players in C would earn by collaborating in the game, with $f(X_\varnothing) = 0$ by convention. Our motivating observation is that if the problem satisfied a certain MIL assumption, then if an area of an image is uninformative (i.e. it does not contain the concept), so will be its constituent sub-areas. Therefore, the exploration of relevant areas of an image can be done in a hierarchical manner. There exists extensive literature on hierarchies of games and their properties (Faigle and Peis, 2008; Algaba and Brink, 2019). Our contribution is to deploy these ideas for the purpose of

image explanations.

We now make this more precise. Denote $\phi_i(f)$ the Shapley value of the i-th player in a TU game $(X, f, [n])$. Following Sec. 1.5, $\phi_i(f)$ can be defined as

$$\phi_i(f) = \sum_{C \subseteq [n] \setminus \{i\}} \frac{|C|!(n - |C| - 1)!}{n!} \left[f(X_{C \cup \{i\}}) - f(X_C) \right]. \tag{2.2}$$

Let $\mathcal{T}_0 = (S_0, \mathcal{T}_1, \ldots, \mathcal{T}_\gamma)$ be a recursive γ-partition tree of X, where S_0 is the root node containing all features of X, i.e. $S_0 = [n], |S_0| = n$, and $\mathcal{T}_1, \ldots, \mathcal{T}_\gamma$ are the subtrees branching off of S_0. Let $c(S_i) = \{C_1, \ldots, C_\gamma\}$ denote the children of S_i, and $h_{\hat{f}} : S_i \mapsto (X, \hat{f}, c(S_i))$ be a mapping from the node S_i of $\mathcal{T}_i$ to the γ-person cooperative game $(X, \hat{f}, c(S_i))$. Succinctly, $\mathcal{G}_0 = h_{\hat{f}}(\mathcal{T}_0)$ is a hierarchy of γ-person games, and we denote by $\phi_{i,1}(\hat{f}), \ldots, \phi_{i,\gamma}(\hat{f})$ the Shapley coefficients of $g_i \in \mathcal{G}_0$. In simpler words, we partition an image X into *a few disjoint components*, compute the Shapley coefficients ϕ_i of each component, and then partition further in a hierarchical manner. In particular, the number of such partitions per level (specified by γ) is very small: if X is a one dimensional vector, we set $\gamma = 2$ and $\mathcal{T}_0$ is a binary tree; when X is a $(\sqrt{n} \times \sqrt{n})$ image, $\gamma = 4$ and $\mathcal{T}_0$ is a quadtree. As a result, computing all 2^γ unique evaluations of $\hat{f}$ required for each game $(X, \hat{f}, c(S_i))$ is trivial. For images, each coefficient requires only 16 model evaluations. In fact, the remaining coefficients (for the same node) involve the same terms but in different permutations, so no extra model evaluations are needed. We have chosen to employ symmetric disjoint partitions in this work (i.e. halves for vectors, quadrants for images, etc) for simplicity only. More sophisticated (and potentially data-dependent) hierarchical partitions are possible as well.

We will comment on this in the discussion.

Given such nested partitions, h-Shap relies on evaluating the resulting hierarchy of games while only visiting nodes that are relevant. More precisely, beginning at S_0, it computes the coefficients $\phi_{0,1}, \ldots, \phi_{0,\gamma}$ of g_0. Under Eq. (2.1), if any $\phi_{0,i} = 0$, all features in the corresponding subtrees will also be irrelevant. As a result, they can be ignored altogether, and we only proceed by exploring the S_i for which $\phi_i > 0$. This process finishes when all relevant leaves have been visited. In practice, we introduce two parameters to add flexibility. We set a relevance tolerance, τ, which determines the threshold to be used to declare a partition relevant, and therefore expand on its subtrees. We further introduce a minimal feature size, s, that serves as a condition for termination. These two parameters are naturally motivated by application and easy to set. For example, it might not be that useful for a domain expert to know the exact pixel-level explanation of a given input. Rather, it would be more informative to have a coarser aggregation of the features that inform the model prediction. Later in this section, we will precisely characterize how the minimal feature size s affects the dissimilarity between h-Shap's attributions and the exact Shapley coefficients. On the other hand, model deviations and noise in the input may result in positive coefficients very close to 0. Requiring $\phi_i > \tau > 0$ provides control over the sensitivity of the method. Finally, when $\tau = 0, s = 1$, h-Shap simply explores all relevant nodes in $\mathcal{T}_0$ as described above.

Fixed τ and s, h-Shap explores $\mathcal{T}_0$ starting from S_0, and it visits all relevant nodes $S_i : \phi_i > \tau, |S_i| \geq s$. This tree exploration can be naturally done in a depth-first or breadth-first manner; Algorithm 1 presents dh-Shap (depth-first

Algorithm 1 Depth-first h-Shap (*d*h-Shap)

procedure *d*h-Shap($X, \mathcal{T}_0, \hat{f}$)
inputs: image X, threshold $\tau \geq 0$, trained model $\hat{f}$
    $g_0 \leftarrow (X, \hat{f}, c(S_0))$
    $\phi_{0,1}, \ldots, \phi_{0,\gamma} \leftarrow \mathsf{shap}(g_0)$
    for all ϕ_i **do**
        if $\phi_i > \tau$ **then**
            if $|S_i| \leq s$ **then**
                return S_i
            else
                return $d\mathsf{h\text{-}Shap}(X, \mathcal{T}_i, \hat{f})$
            end if
        end if
    end for
end procedure
$L \leftarrow d\mathsf{h\text{-}Shap}(X, \mathcal{T}_0, \hat{f})$

h-Shap). Please refer to Algorithm 2 in Appendix 5.2 for *b*h-Shap (breadth-first h-Shap). The only difference between the two algorithms is that the former defines τ as an absolute value (e.g. 0), whereas the latter does so relative to the pooled Shapley coefficients of all nodes at the same depth (e.g. 50^{th} percentile). Both algorithms return the set of relevant leaves $L \subseteq [n]$ with coefficients greater than τ, and the saliency map $\widehat{\Phi}_{(X,\hat{f})}$ is finally computed as

$$\widehat{\phi}_i = \begin{cases} 1/|L| & \text{if } i \in L, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

This choice will ensure that $\widehat{\Phi}_{(X,\hat{f})}$ is consistent with the exact Shapley attributions $\Phi_{(X,\hat{f})}$ under the MIL assumption, as we will formalize shortly.

To mask features out (i.e. as baseline), h-Shap uses their expected value (or *unconditional distribution* (Janzing, Minorics, and Blöbaum, 2020)) for simplicity, as done by other works (Covert, Lundberg, and Lee, 2021). As pointed out by

(Covert, Lundberg, and Lee, 2021; Lundberg and Lee, 2017), this is valid under the assumptions of model linearity and feature independence[1]. Yet, as we will argue later in Sec. 2.3, the feature independence property holds approximately in the cases we are interested in this work, whereas our MIL assumptions are enough to provide specific guarantees without requiring linearity of the model. We will also show in Sec. 2.1 that these assumptions are sufficient for h-Shap to work well in practice. More generally, our contribution is independent of the particular method employed for sampling the baseline, and follow-up work can employ better approximations of both the observational and interventional conditional distributions in appropriate tasks (Chen et al., 2020).

2.0.1 Computational analysis

The benefit of h-Shap relies in decoupling the dimensionality of the sample X (i.e. n), from the number of players in each game (i.e. γ). As we will explain in this section, this leads to an exponential computational advantage over the general expression in Eq. (2.2) in explaining $\hat{f}$. In the analysis that follows, we do not include the computation of the baseline value–which we assume fixed, see discussion in Sect. 2.3–and we refer the reader to the proofs of all the results in this section to the Appendix 5.1. Let us denote by $\hat{\mathcal{T}}_0$ the subtree of $\mathcal{T}_0$ explored by h-Shap (i.e. the one with the visited nodes only). We will also assume in this section that n is a power of γ for simplicity of the expressions[2]. We begin by making the following remark.

[1] We refer to (Chen et al., 2020; Sundararajan and Najmi, 2020; Merrick and Taly, 2020; Janzing, Minorics, and Blöbaum, 2020) for recent discussion on the use of *observational* vs *interventional* conditional distributions in the context of removal-based explanation methods.

[2] Note that it is trivial to accommodate cases where this is not true.

Remark 2.0.1 (Computational cost)**.** *Given $X \in \mathbb{R}^n$, h-Shap requires at most $2^\gamma k \log_\gamma(n)$ model evaluations, where k is the number of relevant leaves in $\hat{\mathcal{T}}_0$.*

This result follows directly by noting that the cost of splitting each node is always 2^γ, and by realizing that each important leaf takes, at most, $\log_\gamma(n)$ nodes, which is exponentially better than the cost of Eq. (2.2). The reader should recall that the number of internal nodes of a full and complete γ-partition tree is $(n-1)/(\gamma-1)$. Then, the above result is relevant whenever $k \log_\gamma n < (n-1)/(\gamma-1)$. This implies that further benefit is obtained whenever $k = \mathcal{O}(n/\log_\gamma n)$, which is only a mild requirement in the number of relevant features.

Moreover, it is of interest to know the expected computational cost, which can be significantly smaller than the upper bound above. Throughout the rest of this section, and to provide more precise results, we will let the data X be drawn from a distribution of *important* and *non-important* features. A distribution is "important" in the sense that it leads to positive responses.

Assumption A1. *The data $X \in \mathbb{R}^n$ is drawn so that each entry $x_i \sim a_i\mathcal{I} + (1-a_i)\mathcal{I}^c$, where $a_i \sim$ Bernoulli(ρ) is a binary random variable that indicates whether the feature x_i comes from an* important *distribution $\mathcal{I}$, or its* non-important *complement $\mathcal{I}^c$, so that*

$$\hat{f}(X_C) = 1 \iff \exists i \in C : \; x_i \sim \mathcal{I}, \; C \subseteq [n]. \tag{2.4}$$

For example, we can imagine $x_i \in \mathbb{R}$, $\mathcal{I} \sim \mathcal{N}(0,1)$, and $\mathcal{I}^c \sim \mathcal{N}(1,1)$. With these elements, we present the following result.

Theorem 2.0.2 (Expected number of visited nodes)**.** *Assume X and $\hat{f}(X)$ satisfy*

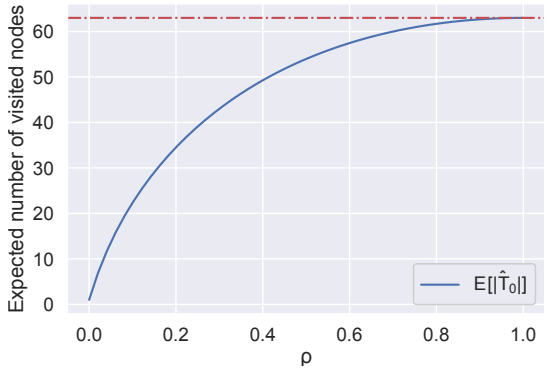


Figure 2.1: Expected number of visited nodes as a function of ρ when $n = 64, \gamma = 2, s = 1$.

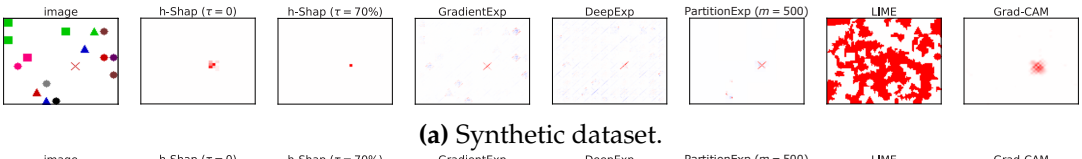
A1, $\tau = 0$, *and* $s = 1$. *Then, the expected number of visited nodes in* $\hat{\mathcal{T}}_0$ *is*

$$\mathbb{E}[|\hat{\mathcal{T}}_0|] = 1 + \gamma(1 - p(S_0))\mathbb{E}[|\hat{\mathcal{T}}_1|], \tag{2.5}$$

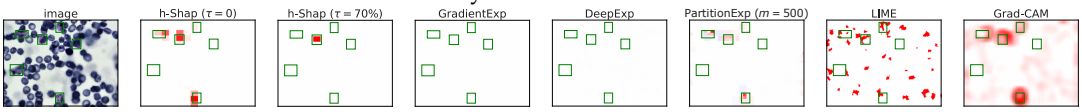
where

$$p(S_i) = \begin{cases} (1-\rho)^{\frac{|S_i|}{\gamma}} & \text{if } i = 0, \\ (1-\rho)^{\frac{|S_i|}{\gamma}} \left(\frac{1-(1-\rho)^{|S_i|\frac{\gamma-1}{\gamma}}}{1-(1-\rho)^{|S_i|}} \right) & \text{otherwise.} \end{cases}$$

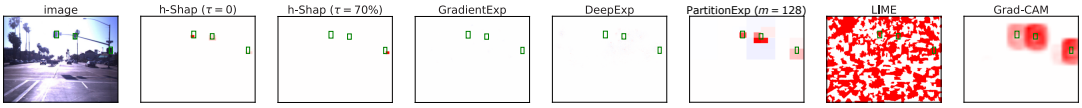
See Proof 5.1.1. This result does not provide a closed-form expression for the expected number of visited nodes (and, correspondingly, computational cost), but it does provide a simple recurrent formula that can be easily computed. Naturally, this cost depends on the Bernoulli probability ρ, the average number of important features in X. We present the resulting $\mathbb{E}[|\hat{\mathcal{T}}_0|]$ for a specific case in Fig. 2.1 as a function of ρ, showing that indeed the expected cost can be much lower than the worst-case bound. While this result (and, centrally, Assumption A1) was presented for the case where the relevant features are of size 1, similar results can be provided for the case when the minimal features size $s > 1$.



(a) Synthetic dataset.



(b) BBBC041 dataset.



(c) LISA dataset.

Figure 2.2: A few saliency maps for the three settings studied in this work, where blue pixels have negative, white pixels have negligible, and red pixels have positive Shapley coefficients. The color mapping is adapted to each saliency map and centered around 0. For h-Shap, we show the saliency map before the normalization step.

2.0.2 Accuracy and Approximation

Recall that h-Shap provides image attributions by means of a hierarchy of collaborative games. As a result, the attributions are different, in general, from those estimated by analyzing the grand coalition directly–that is, by the general Shapley approach in Eq. (2.2). We remark that computing the Shapley coefficients directly from Eq. (2.2) quickly becomes intractable in image classification tasks. For example, even for a toy-like dataset of small 10×10 pixels images, assuming that each model computation takes 1 nanosecond (which is unrealistically fast), computing the exact Shapley coefficients would take $\approx 3 \times 10^{13}$ years. Yet, we now show that under A1, h-Shap can in fact provide exact Shapley coefficients while being exponentially faster. Additionally, h-Shap can provide controlled approximations by trading computational efficiency with accuracy.

We begin by noting that under the MIL assumption, all positive features have the same importance. This agrees with intuition that the number of times the positive concept appears in the input image does not affect its label. We denote as Φ and $\widehat{\Phi}$ the exact and hierarchical Shapley coefficients, respectively, for simplicity.

Remark 2.0.3. *Under A1, and denoting $k = \|\Phi\|_0$, it holds that the exact saliency map Φ satisfies*

$$\phi_i = \begin{cases} 1/k & \text{if } x_i \sim \mathcal{I} \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

This remark follows simply from the nullity and symmetry properties of Shapley coefficients. As a result, the saliency map computed by h-Shap, $\widehat{\Phi}$, as in Eq. (2.3), coincides with Φ under the MIL assumption. We now derive a more general similarity lower bound between Φ and $\widehat{\Phi}$ that allows for minimal feature sizes $s > 1$. For simplicity, we assume that n and s are powers of γ, and $1 \leq s \leq n$. First of all, because of the MIL assumption, h-Shap will *only* keep exploring nodes that have at least one important feature in them at each level of the hierarchy. Thus, for each important feature i with $\Phi_i = 1/k$ there will be a non-zero coefficient produced by h-Shap. The following result precisely quantifies to what extent these two vectors Φ and $\widehat{\Phi}$ match.

Theorem 2.0.4 (Similarity lower bound)**.** *Assume $X \in \mathbb{R}^n$ and $\hat{f}(X)$ satisfy A1, and $k = \|\Phi\|_0$. Then*

$$\frac{\langle \Phi, \widehat{\Phi} \rangle}{\|\Phi\|_2 \|\widehat{\Phi}\|_2} \geq \max\{1/\sqrt{s}, \sqrt{k/n}\}. \tag{2.7}$$

See Proof 5.1.2. This result shows that not only does h-Shap provide faster image attributions, but it retrieves the exact Shapley coefficients defined in

Eq. (2.6) under the MIL assumption if $s = 1$. Notwithstanding, one can employ a larger minimal feature size, $s > 1$, while still providing attributions that are similar to the original ones. In light of the result in Theorem 2.0.2, the latter attributions will naturally result in improved (smaller) computational costs.

2.1 Experiments

We now move to demonstrate the performance of h-Shap and of other state-of-the-art methods for image attributions. Our objective is mainly to compare with other Shapley-based methods, such as GradientExplainer (Lundberg and Lee, 2017), DeepExplainer (Lundberg and Lee, 2017; Chen, Lundberg, and Lee, 2021), and PartitionExplainer[3]. We also include LIME[4] (Ribeiro, Singh, and Guestrin, 2016) given its relation to Shapley coefficients, and Grad-CAM[5] (Selvaraju et al., 2017) because of its popularity. We study three complementary binary classification problems of different complexity and input dimension: a simple synthetic benchmark, a medical imaging dataset, and a general computer vision task. We focus on scenarios where the ground truth of the image attributions (i.e. what defines the label) is well defined and available for evaluation. All experiments were conducted on a workstation with NVIDIA Quadro RTX 5000. Our code is made available for the purpose of reproducibility[6]. When possible, each method was set to use as much GPU memory as possible, so as to minimize their runtime. DeepExplainer and

[3]The implementation of GradientExplainer, DeepExplainer and PartitionExplainer are openly available at https://github.com/slundberg/shap.

[4]https://github.com/marcotcr/lime.

[5]https://github.com/jacobgil/pytorch-grad-cam.

[6]https://github.com/Sulam-Group/h-shap

GradientExplainer were constrained the most by memory, reflecting their limitation in analyzing large images. We use h-Shap with both an absolute threshold $\tau = 0$, and a relative threshold τ equal to the 70^{th} percentile, which we refer to as $\tau = 70\%$ with abuse of notation. Finally, we perform *full* model randomization sanity checks (Adebayo et al., 2018) on the network used in the synthetic dataset for all explanation methods. We refer the reader to Appendix 5.5 for these results.

2.1.1 Synthetic dataset

We created a controlled setting where the joint data distribution is completely known, giving us maximal flexibility for sampling. We generate images of size 100×120 pixels with a random number of non-overlapping geometric shapes of size 10×10 and of different colors, uniformly distributed across the image. Each image that contains at least one cross receives a positive label, and each image without any crosses receives a negative label. Alongside with the images, we generate the ground truth saliency maps by setting all pixels that precisely lie on a cross to 1, and every other pixel to 0. We generate 8000 positive and negative images, and we randomly sample train, validation, and test splits, with size 5000, 1000 and 2000 images, respectively. We train a simple ConvNet architecture, optimizing for 50 epochs with Adam (Kingma and Ba, 2014), learning rate of 0.001 and cross-entropy loss. We achieve an accuracy greater than 99 % on the test set–implying that the model has effectively satisfied the MIL assumption for this problem. From the true positive predictions on the test set, we choose 300 example images with 1 cross

and as many with 6 crosses to evaluate the saliency maps. Fig. 2.2a presents a qualitative demonstration of h-Shap and other related methods on this task.

2.1.2 P. vivax (malaria) dataset

Moving on to a real and high-dimensional problem, we explore the BBBC041v1 dataset, available from the Broad Bioimage Benchmark Collection[7] (Ljosa, Sokolnicki, and Carpenter, 2012). The dataset consists of 1328, 1200×1600 pixels blood smears with uninfected (i.e. red blood cells and leukocytes) and infected (i.e. gametocytes, rings, trophozoites, and schizonts) blood cells. The dataset also comprises bounding-box annotations of both healthy and sick cells. We consider the binary problem of detecting images that contain at least one trophozoite, yielding 655 positive and 673 negative samples. Given the small amount of data available, we augment the training dataset with random horizontal flips, and we randomly choose 120 positive, and equally many negative images as the testing set. We apply transfer learning to a ResNet18 (He et al., 2016) network pretrained on ImageNet. We optimize all parameters of the pretrained network for 25 epochs with Adam (Kingma and Ba, 2014)– learning rate 0.0001. We use cross-entropy loss and learning rate decay of 0.2 every 10 epochs. After training, our model achieves a test accuracy greater than 99 %. We finally aggregate all 112 true positive predictions for evaluation, without distinction on the number of trophozoites in the image. Fig. 2.2b shows a sample image and the corresponding saliency maps produced by the various methods.

[7] https://www.kaggle.com/kmader/malaria-bounding-boxes.

2.1.3 LISA traffic light dataset.

We finally look at a general computer vision dataset consisting of driving sequences collected in San Diego, CA, available from[8] (Jensen et al., 2016; Philipsen et al., 2015). The complete dataset counts 43 007 frames of size 960×1280 pixels, and 113 888 annotated traffic lights. From this set, we take daytime traffic images, and train a model to predict the presence of a green light in a sample image. We respect the original train/test splits, providing 6108 train, 3846 test positive samples, and 6667 train, 3627 test negative samples. As before, we use data augmentation and apply transfer learning on a pretrained ResNet18. We optimize all parameters of the pretrained network for 25 epochs with Adam (Kingma and Ba, 2014)–learning rate 0.0001. We use cross-entropy loss and learning rate decay of 0.2 every 10 epochs. After training, we achieve a test accuracy of $\approx 95\,\%$. Finally, we randomly sample 300 true positive examples to evaluate the different attribution methods on. Fig. 2.2c illustrates a positive sample image, and the corresponding saliency maps.

2.2 Results

Fig. 2.2 shows a visual comparison of some saliency maps obtained in the three experiments (for more examples, see Fig. 5.1). Note that while the saliency maps produced by GradientExplainer and DeepExplainer appear empty in Fig. 2.2b and 2.2c, they are not, and instead the single pixels are too small to be visible (these are large images). This illustrates how current Shapley-based

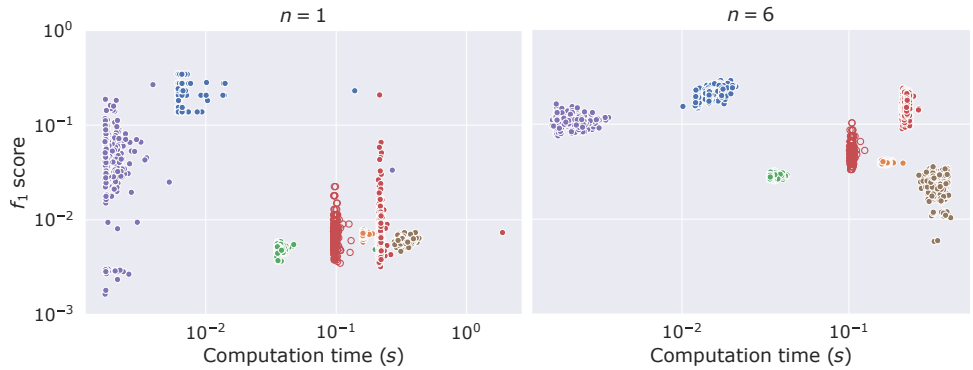
[8] https://www.kaggle.com/mbornoe/lisa-traffic-light-dataset.

Figure 2.3: Ablation examples for all explanation methods removing all important pixels from the original image 2.3a. Images are generated synthetically by placing different geometric shapes of 10×10 pixels uniformly without overlap. The ground truth binary MIL rule labels positively images that contain at least one cross. We remark that colors are sampled uniformly in order to remove any correlation with the true label.

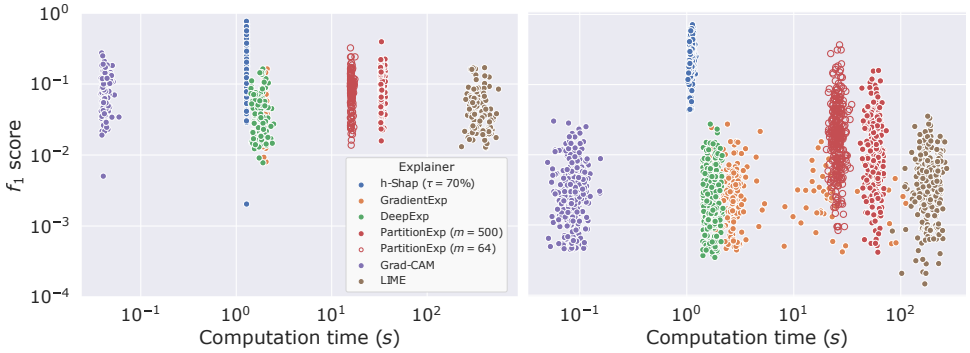
explanation methods fall short of producing informative saliency maps in problems with large images. We further evaluate the explanation methods by means of three performance measures: ablation tests, accuracy, and runtime.

2.2.1 Ablation tests

As commonly done in literature (Lundberg and Lee, 2017; Sturmfels, Lundberg, and Lee, 2020; Haug et al., 2021) we remove the top k scoring features of all methods by setting them to their expected value, and plot the logit of the prediction as a function of k. For these experiments, we use $\tau = 0$ so as to find *all* the features that are relevant for the model. Fig. 2.3 shows ablation results on one example image from the synthetic dataset for all explanation methods. We expect a perfect method to remove all crosses from the image–and only those. We can appreciate how h-Shap removes mostly only the crosses, while other methods also erase other shapes which should not be identified as important. Furthermore, removing more relevant features should produce a steeper drop of the prediction logit. We include the respective curves in Fig. 5.3, depicting that h-Shap's logit curves either quickly drop towards 0 or provide



(a) Synthetic dataset. Results for $n = 1, 6$ crosses.



(b) BBBC041 dataset. **(c)** LISA dataset.

Figure 2.4: f_1 scores as a function of runtime for all explanation methods in all three experiments. To account for noise in the explanations, we threshold saliency maps at 1×10^{-6} and compute f_1 scores on the resulting binary masks. For PartitionExplainer, m indicates the maximal number of model evaluations.

a logit ≈ 0 at complete ablation. Indeed, h-Shap quickly identifies the most relevant features in the image. Naturally, as tasks become harder, the accuracy of $\hat{f}$ decreases, and the model gets further away from the oracle function f^*. In these cases (for the real datasets), $\hat{f}$ might not satisfy Eq. (2.1), resulting in noisier saliency maps, and correspondingly, non-monotonic curves.

2.2.2 Accuracy and Runtime

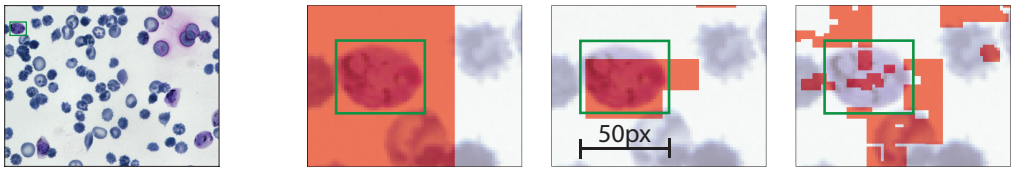
Since we have ground-truth explanations in all these cases (i.e. a cross, a sick cell, or a green traffic light), we use f_1 scores as a measure of goodness of

explanation. We argue that f_1 scores are a particularly informative measure for explanations (when ground-truth is known), and consistent with previous work (Guidotti, 2021). We note that we compute f_1 scores as the harmonic mean of precision and recall of the saliency maps at the pixel level. Fig. 2.4 depicts the f_1 scores as a function of runtime for every explanation method and experiment. The advantage of setting a relative relevance tolerance τ is clear: to detect the *most* relevant features and discard the noisy ones, taking into account the risk of the model $\hat{f}$, while also decreasing runtime. These results reflect how the computational cost and accuracy guarantees described earlier translate into application. Not only does h-Shap decrease runtime compared to current Shapley-based explanation methods–by one to two orders of magnitude–but it also increases the f_1 score. Fig. 2.4a shows that h-Shap's accuracy is not affected by the number of crosses in the image, while other methods deteriorate when there is only one cross to detect in the image. Importantly, in all experiments–both synthetic and real–h-Shap consistently provides more accurate and faster saliency maps compared to other Shapley-based methods, and it is only beaten in speed by Grad-CAM, which provides less accurate saliency maps.

2.3 Discussion

2.3.1 Limitations

Before concluding, we want to delineate the limitations of h-Shap, the most important of which is its MIL assumption on the data distribution. The methodology proposed in this work is designed to identify local *findings*



(a) Original image. **(b)** $s = 80$ pixels. **(c)** $s = 20$ pixels. **(d)** $s = 5$ pixels.

Figure 2.5: Degradation of h-Shap's maps as the minimal feature size s becomes smaller than the target concept.

that produce a positive global response, accurately and efficiently. These are precisely the important features C. This setting is controlled by the ratio of the size of the actual object that defines the label, and the minimal feature size of the algorithm. As an example, Fig. 2.5 depicts a zoomed-in version of the map produced by h-Shap for one of the samples from the P. vivax dataset, for different values of s. We see that even when s is somewhat smaller than the object, h-Shap still recognizes the important features in the image. Once s is too small, however, the resulting map breaks down, as our assumption does not hold any more. Indeed, small (5×5 pixels) image patches break Assumption A1 because a small patch of a cell is not sufficient for the model to recognize it. In practice, these failure cases can easily be identified by deploying simple conditions searching over decreasing sizes of s (which would not increase the computational cost). We note that Eq. (2.4) can also be phrased as an `OR` function across features. Intuitively, when the minimal feature size s is smaller that the concept of interest, the `OR` function is no longer appropriate.

A second limitation of h-Shap pertains the way hierarchical partitions are created. We have chosen to use quadrants for their effectiveness and elegance,

but this could be sub-optimal: important features may fall in-between quadrants, impacting performance. This limitation is minor, as it can be easily fixed by applying ideas of cycle spinning and averaging the resulting estimates. Furthermore, and more interestingly, hierarchical data-dependent partitions could also be employed. We regard this as future work.

2.3.2 Baseline and assumptions

Recall that all explanation methods based on feature removal–like Shapley-based explanation methods–are sensitive to the choice of baseline, i.e. the reference value used to mask features. Then, we now turn our attention to h-Shap's masking strategy, or alternatively, how to sample a reference. We recall that in this work we defined the variable X_C as

$$(X_C)_i = \begin{cases} X_i & \text{if } i \in C \\ R_i & \text{otherwise,} \end{cases} \tag{2.8}$$

where $R \in \mathbb{R}^{n-|C|}$ is a baseline value. Throughout this work, we have treated R as a fixed, deterministic quantity. However, more generally, reference inputs are random variables. Let this masked input be the random variable $X_C = [\bar{X}_C, R] \in \mathbb{R}^n$, where $\bar{X}_C \in \mathbb{R}^{|C|}$ is fixed, and R is a random variable. Here, we want to identify what relationships in the data distribution are important for the model, so we follow the original approach in (Lundberg and Lee, 2017). Indeed, the definition of Shapley values for the i^{th} coefficient in Eq. (2.2) can be made more precise by writing its expectation $\mathbb{E}[\hat{f}(X_{C\cup\{i\}}) - \hat{f}(X_C)]$ as

$$\mathbb{E}_R[\hat{f}([\bar{X}_{C\cup\{i\}}, R]) \mid \bar{X}_{C\cup\{i\}}] - \mathbb{E}_R[\hat{f}([\bar{X}_C, R]) \mid \bar{X}_C]. \tag{2.9}$$

As it can be seen, if the model $\hat{f}$ is linear, and the features are independent, then Eq. (2.9) simplifies to

$$\hat{f}([\bar{X}_{C\cup\{i\}}, \mathbb{E}[R]]) - \hat{f}([\bar{X}_C, \mathbb{E}[R]]), \tag{2.10}$$

where $\mathbb{E}[R]$ is an unconditional expectation which can be easily computed over the training data, and is precisely the fixed baseline we employed in this work.

How realistic are these assumptions in our case? First, the cases that we study here approximately satisfy feature independence in a local sense, and it is therefore reasonable to consider the input features as independent when s–the minimal feature size–is greater or similar to the size of the concept we are interested in detecting. Indeed, this is precisely true in the synthetic dataset, where each 10×10 pixels shape is sampled independently from the others. This assumption is still approximately valid in the other two experiments, where, for example, the presence or absence of a cell does not affect the content of the image so many pixels apart. On the other hand, while we have chosen very general models $\hat{f}$ which are far from linear, we argue that A1 is enough to obtain a weaker sense of interpretability: looking at

$$\hat{f}([X_C, \mathbb{E}[R]]), \tag{2.11}$$

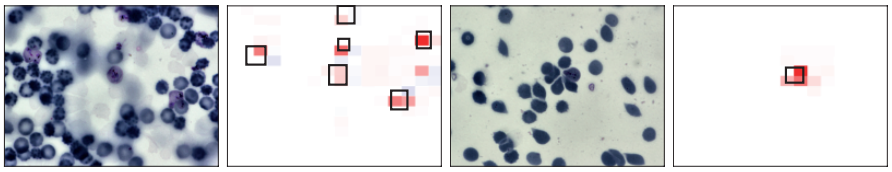
and under the MIL assumption, there are only two mutually exclusive events for the subset C: (a) C contains at least one relevant feature, and (b) C does not contain any relevant features. When event (a) occurs, Eq. (2.11) will necessarily yield a high value ≈ 1, regardless of the value of the baseline $\mathbb{E}[R]$. It follows

that if both $C \cup \{i\}$ and C contain important features, Eq. (2.10) will be ≈ 0; which agrees with intuition that all important features are equally important. As a result, because $\mathbb{E}[R]$ is fixed and A1 holds, a positive value of Eq.(2.10) is only attained if (i.e. implies that) i is an important feature (and it also implies that $\mathbb{E}[R]$ is not important).

To summarize, the choice of using the unconditional expectation as a baseline value is approximately valid because feature independence approximately holds on a local sense, and although the models we study are highly non-linear, Assumption A1 guarantees a weaker sense of interpretability. However, when these two conditions are not satisfied, one should deploy different methods to approximate the conditional distribution as in Eq. (2.9). Lastly, note that our method relies on $\hat{f}$ satisfying A1, and one should wonder when this holds. Such an assumption is true when f^*–the true classification rule $Y = f^*(X)$–satisfies A1 (which is true for a variety of problems, including the ones studied in our experiments), and $\hat{f}$ constitutes a good approximation for f^*. As demonstrated in this work, such assumptions are reasonable in practical settings.

2.3.3 Multi-class extensions

Even though we have focused on binary classification tasks in this work, h-Shap can also be applied to multi-class settings. We now briefly demonstrate this by modifying the P. vivax experiment. We let $Y \in \mathcal{Y} = \{0,1\}^2$, such that $Y = (\texttt{trophozoite}, \texttt{ring})$. Then, $\texttt{trophozoite} = 1$ if and only if there is at least one *trophozoite* in the image, and $\texttt{ring} = 1$ if and only if there is at least



(a) `trophozoite` = 1.

(b) Explanation for label `trophozoite`.

(c) `ring` = 1.

(d) Explanation for label `ring`.

Figure 2.6: Example saliency maps for different labels in a multiclass setting.

one *ring cell* in the image. Note that in this setting, these two classes are not mutually exclusive, as is typically the case for traditional image classifications problems. The latter setting is simply a particular case of the former. We randomly choose a training split that contains 80% of each class, and we finetune a ResNet18 pretrained on ImageNet. We optimize all parameters for 60 epochs with Adam (Kingma and Ba, 2014), using binary cross-entropy loss per class (as the classes are not mutually exclusive), learning rate of 0.0001, weight decay of 0.00001, and learning rate decay of 0.7 every 7 epochs. After training, the model achieves an accuracy of $\approx 87\%$ on each label across the held-out test set. Fig. 2.6 shows saliency maps for two example images from the test set, one containing 6 trophozoites, and one containing 1 ring cell. h-Shap can explain every class, and it retrieves the desired, different types of cells. We regard studying the full implications and capabilities of h-shap in multi-class MIL problems as future work.

References

Chen, Hanjie, Guangtao Zheng, and Yangfeng Ji (2020). "Generating hierarchical explanations on text classification via feature interaction detection". In: *arXiv preprint arXiv:2004.02015*.

Chen, Jianbo, Le Song, Martin J Wainwright, and Michael I Jordan (2018). "L-shapley and c-shapley: Efficient model interpretation for structured data". In: *arXiv preprint arXiv:1808.02610*.

Singh, Chandan, W James Murdoch, and Bin Yu (2018). "Hierarchical interpretations for neural network predictions". In: *arXiv preprint arXiv:1806.05337*.

Dietterich, Thomas G, Richard H Lathrop, and Tomás Lozano-Pérez (1997). "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial intelligence* 89.1-2, pp. 31–71.

Faigle, Ulrich and Britta Peis (2008). "A hierarchical model for cooperative games". In: *International Symposium on Algorithmic Game Theory*. Springer, pp. 230–241.

Algaba, Encarnacion and Rene van den Brink (2019). "The Shapley Value and Games with Hierarchies". In: *Handbook of the Shapley Value*, p. 49.

Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum (2020). "Feature relevance quantification in explainable AI: A causal problem". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2907–2916.

Covert, Ian, Scott Lundberg, and Su-In Lee (2021). "Explaining by removing: A unified framework for model explanation". In: *Journal of Machine Learning Research* 22.209, pp. 1–90.

Lundberg, Scott and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874*.

Chen, Hugh, Joseph D Janizek, Scott Lundberg, and Su-In Lee (2020). "True to the Model or True to the Data?" In: *arXiv preprint arXiv:2006.16234*.

Sundararajan, Mukund and Amir Najmi (2020). "The many Shapley values for model explanation". In: *International Conference on Machine Learning*. PMLR, pp. 9269–9278.

Merrick, Luke and Ankur Taly (2020). "The Explanation Game: Explaining Machine Learning Models Using Shapley Values". In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, pp. 17–38.

Chen, Hugh, Scott Lundberg, and Su-In Lee (2021). "Explaining models by propagating Shapley values of local components". In: *Explainable AI in Healthcare and Medicine*. Springer, pp. 261–270.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""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*, pp. 1135–1144.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2018). "Sanity checks for saliency maps". In: *arXiv preprint arXiv:1810.03292*.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Ljosa, Vebjorn, Katherine L Sokolnicki, and Anne E Carpenter (2012). "Annotated high-throughput microscopy image sets for validation." In: *Nature methods* 9.7, pp. 637–637.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Jensen, Morten Bornø, Mark Philip Philipsen, Andreas Møgelmose, Thomas Baltzer Moeslund, and Mohan Manubhai Trivedi (2016). "Vision for looking at traffic lights: Issues, survey, and perspectives". In: *IEEE Transactions on Intelligent Transportation Systems* 17.7, pp. 1800–1815.

Philipsen, Mark Philip, Morten Bornø Jensen, Andreas Møgelmose, Thomas B Moeslund, and Mohan M Trivedi (2015). "Traffic light detection: A learning algorithm and evaluations on challenging dataset". In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, pp. 2341–2345.

Sturmfels, Pascal, Scott Lundberg, and Su-In Lee (2020). "Visualizing the impact of feature attribution baselines". In: *Distill* 5.1, e22.

Haug, Johannes, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci (2021). "On Baselines for Local Feature Attributions". In: *arXiv preprint arXiv:2101.00905*.
Guidotti, Riccardo (2021). "Evaluating local explanation methods on ground truth". In: *Artificial Intelligence* 291, p. 103428.

Chapter 3

Are bags all you need? A case on weakly-supervised intracranial hemorrhage detection in brain CT examinations

Supervised learning is the most popular framework to develop machine learning solutions for a broad variety of medical imaging problems. Consequently, the curation of large datasets with high-quality labels remains one of the most difficult hurdles to overcome. Indeed, recent efforts have been focusing on semi-supervised learning approaches, for example, by extracting low-quality labels from unstructured medical records. Notwithstanding these exciting advances, ground truth labels are often required for a careful assessment of a model's performance, trustworthiness, and safety in medical settings. Then, *what kind* of labels should be sought-after? In this chapter, we focus on intracranial hemorrhage *location* (i.e. within an examination) and *detection* (i.e within images) in brain CT scans. We frame the task as a Multiple-Instance Learning (MIL) binary classification problem, and propose an attention-based model

that can be trained with both *local* (i.e. image-level) and *global* (i.e. scan-level) labels. We investigate whether, given the same architecture, full-supervision with local labels significantly improves the model's performance compared to weak-supervision with global labels. We find that strong, (i.e. local) learners, and weak, (i.e. global) learners achieve comparable performances in hemorrhage location and detection. Furthermore, we study this behavior as a function of the number of labels available to the learners. These results suggest that local labels may not be necessary at all, drastically reducing the time-consuming process of data curation.

3.0.1 Background

Unenhanced brain computed tomography (CT) is the most common imaging assessment technique for the diagnosis of Intracranial Hemorrhage (ICH) in clinical settings (e.g. traumatic brain injury). Expert radiologists can identify hemorrhage in CT images from the differences in attenuation compared to normal brain tissue, and classify lesions based on their size, shape, and location. Indeed, ICH is usually divided in five subtypes: epidural, intraparenchymal, intraventricular, subarachnoid, and subdural. Computer-Aided Diagnosis (CAD) systems are desirable for ICH given the high volumes of CT scans produced in clinical settings and the importance of a quick response in treating severe cases. For example, a CAD system may support radiologists by prioritizing the most severe cases, or by providing a second opinion for subtle ones.

The need for such autonomous systems in real-world settings has motivated the development of several deep-learning-based classification and segmentation models (Hssayeni et al., 2020; Lee et al., 2019; Salehinejad et al., 2021). Currently, the most common approach to developing such models is supervised learning–models are trained on a collection of CT images, each of which can be annotated with a binary label (e.g. "`hemorrhage`" or "`no hemorrhage`"), a multiclass label (e.g. "`epidural`", "`intraparenchymal`", "`intraventricular`", ...), or a manual segmentation (e.g. bounding boxes around the bleeds). These approaches rely on expert radiologists annotating significant amounts of data in order for models to achieve adequate performance levels. It remains unclear, however, what kind of labels should be collected. Are local, image-level labels necessary, or would global, examinations-level labels suffice? The answer to this question is of imperative performance given the time and costs associated with the construction of one or the other type of dataset. In this chapter, we study intracranial hemorrhage *location* (i.e. within examinations) and *detection* (i.e. within images) and precisely address the gap between training with full and weak supervision. We propose an attention-based, multiple-instance learning model (Ilse, Tomczak, and Welling, 2018) that can be trained with both full and weak supervision. The Multiple-Instance Learning (MIL) framework (Dietterich, Lathrop, and Lozano-Pérez, 1997; Maron and Lozano-Pérez, 1997; Weidmann, Frank, and Pfahringer, 2003) generalizes supervised learning to *bags* of inputs. Binary classification is the most common MIL setting: a bag is labeled positively if and only if it contains at least one positive instance. Importantly, an MIL learner can only access global, bag labels. Hence, it is usually considered a type of weakly-supervised

learning. We investigate whether fully-supervised models show significant improvements over weakly-supervised ones for the problem of ICH detection on three different dataset: we train and validate on the RSNA 2019 Brain CT Hemorrhage Challenge[1] dataset (Flanders et al., 2020), and test on both the CQ500[2] dataset (Chilamkurthy et al., 2018)–with bounding boxes from the BHX[3] extension (Reis et al., 2020; Goldberger et al., 2000)–and the CT-ICH[4] dataset (Hssayeni, 2020; Hssayeni et al., 2020; Goldberger et al., 2000). Furthermore, we compare learners when trained with the same number of labels. This comparison precisely answers whether local labels lead to significant improvements compared to global labels, and it informs how to build new datasets. Indeed, assume that we want to construct a new intracranial hemorrhage detection dataset. Given some resource constraints, we are only allowed to ask expert radiologists a finite number of questions, be those about images or examinations. Then, should radiologists label individual instances, or entire CT examinations? We find that, depending on the number of labels available, labeling examinations can lead to equivalent or improved performance compared to labeling images. Importantly, labeling examinations with a binary label ("`sick`" or "`healthy`") is a much faster and cheaper task than labeling single images.

[1] available at: https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection
[2] available at: http://headctstudy.qure.ai/dataset
[3] available at: https://physionet.org/content/bhx-brain-bounding-box/1.1/
[4] available at: https://physionet.org/content/ct-ich/1.3.1/

3.1 Results

Following the notation and definitions in Sec. 1.2 and 1.3, we rephrase hemorrhage location and detection as parts of an MIL binary classification problem. Let a bag $\overline{X} \in \mathcal{X}^r$, $\mathcal{X} \subseteq \mathbb{R}^D$ be a brain CT examination of length $r \in \mathbb{N}$ (i.e. each examination is in $\mathbb{R}^{\sqrt{D}\times\sqrt{D}\times r}$), and let the instances $X_1, \ldots, X_r \in \mathcal{X}$ be the images in the examination. Then, $\overline{X}$ should be labeled as "`with hemorrhage`" as soon as it contains at least one image with findings in support of the presence of hemorrhage, and "`normal`" otherwise. Let $\overline{Y}$ denote the bag label, then

$$\overline{Y} = 1 \iff \exists i \in [r] : \; X_i \text{ contains hemorrhage}, \tag{3.1}$$

where $[r] := \{1, \ldots, r\}$. The equation above defines an MIL binary classification problem where the target concept: h^*, is $h^*(X) = \mathbb{1}[X \text{ contains hemorrhage}]$. It follows that hemorrhage location can be phrased in terms of retrieving the positive instances in a positive bag. Furthermore, $h^*(X)$ is equal to 1 as soon as X contains findings in support of the presence of hemorrhage. Let $X = (x_1, \ldots, x_D)$, then

$$h^*(X) = 1 \iff \exists C \subseteq [D] : \; h^*(X_C) = 1, \tag{3.2}$$

where $X_C \in \mathcal{X}$ is defined following Sec. 1.5, i.e. it is equal to x_i for all features i in C, and some *uninformative baseline* value otherwise. We remark that, differently from Eq. (3.1), the equation above is not permutation invariant. Indeed, permuting pixels will change the features in the image. We conclude that hemorrhage detection can be phrased in terms of identifying the positive features in a positive instance.

To summarize:

- Hemorrhage *location* can be rephrased as selecting the positive instances in an MIL binary classification problem, and likewise
- hemorrhage *detection* can be rephrased as finding the positive features within positive instances.

We remark that, to our knowledge, this is the first MIL-based work addressing both positive instance retrieval and pixel-level detection concurrently with one model for a medical imaging task. Indeed, most literature addressing positive instance retrieval does not entail pixel-level segmentation, while most works that propose methods for pixel-level segmentation consider images as bags of region proposals, and do not consider bags of images.

Finally, we remark that the MIL assumption in Eq. (3.1) provides a weaker sense of supervision. Indeed, an MIL learner needs to disambiguate between those concepts different from the true concept, h^*, that may classify bags with high accuracy. For example, say that one image–either positive or negative–in every positive examination has been annotated with a "P" marker. Then, the instance-level hypothesis $\widetilde{h}(X) = \mathbb{1}\left[\text{"P" is in } X\right]$ is independent of $h^*(X)$ and will perform poorly on images. However, $\overline{Y}(\overline{X}) = (\texttt{OR} \circ (h^*)^r)(\overline{X}) = (\texttt{OR} \circ \widetilde{h}^r)(\overline{X})$, and $\widetilde{h}$ will correctly classify all bags while being independent of the true target concept. This is a well-known source of hardness for MIL problems, and several state-of-the-art architectures and regularization strategies have been proposed (Bilen and Vedaldi, 2016; Tang et al., 2017; Wan et al., 2019) in the Weakly-Supervised Object Detection (WSOD) literature (Shao et al., 2022), which often frame the task in terms of MIL.

3.1.1 Model architectures

We define fully- and weakly-supervised learners, which we refer to as strong ($\mathcal{SL}$) and weak ($\mathcal{WL}$) learners respectively, as follows. Let $\mathcal{X} \subseteq \mathbb{R}^D$, $\mathcal{F} \subseteq \mathbb{R}^K$ be the input and feature domains, such that $\mathcal{H}_\mathcal{F}$ is a suitable subset of the family of functionals from $\mathcal{X}$ to $\mathcal{F}$, i.e. $\mathcal{F}^\mathcal{X}$. Let $\mathcal{Y} = \{0, 1\}$ be the bag-level output domain, $R \subseteq \mathbb{N}$ be arbitrary bag sizes, and

$$\begin{aligned} \texttt{ENCODER} &:= h \in \mathcal{H}_\mathcal{F} \subseteq \mathcal{F}^\mathcal{X} && \text{(feature extractor)} \\ \texttt{ATTENTION} &:= z : \ (\mathbb{R}^K)^R \to \mathbb{R}^K && \text{(attention mechanism)} \\ \texttt{CLS} &:= \mathbb{R}^K \to [0, 1] \supseteq \mathcal{Y} && \text{(bag classifier).} \end{aligned} \tag{3.3}$$

Then,

$$\begin{aligned} \mathcal{SL} &:= \texttt{CLS} \circ \texttt{ENCODER} \\ \mathcal{WL} &:= \texttt{CLS} \circ \texttt{ATTENTION} \circ \texttt{ENCODER}^R. \end{aligned} \tag{3.4}$$

Specifically, we let ENCODER be a ResNet-18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) and CLS be a linear classifier with sigmoid activation, i.e. $\texttt{CLS} : \ u \mapsto S(\langle \theta, u \rangle)$, $u, \theta \in \mathbb{R}^K$, and S is the sigmoid function. Finally, we remark that the weak learner ($\mathcal{WL}$) is permutation invariant with respects to the order of the images in a bag.

3.1.2 Training strong and weak learners on the RSNA 2019 Brain CT Hemorrhage Challenge dataset.

The RSNA 2019 Brain CT Hemorrhage Challenge dataset (Flanders et al., 2020) is currently the largest publicly available collection of labeled images from unenhanced brain CT examinations. Each image was labeled by expert

Training			
Learner	Positive samples	Negative samples	Total
$\mathcal{WL}$	7100	10288	17388
$\mathcal{SL}$	86295	515635	601930
Validation			
$\mathcal{WL}$	1776	2572	4348
$\mathcal{SL}$	21489	129003	150492

Table 3.1: Number of positive and negative samples in the RSNA 2019 Brain CT Hemorrhage Challenge for strong and weak learners.

Dataset	Positive exams	Negative exams	Total images
CQ500	212	224	15156
CT-ICH	36	39	2539

Table 3.2: Number of positive and negative examinations in the CQ500 and CTICH datasets, alongside the total number of images contained in the two datasets.

radiologists depending on the type(s) of hemorrhage present: epidural, intraparenchymal, intraventricular, subarachnoid, and subdural. Furthermore, it is the first dataset containing volumetric data instead of individual images. It consists of $874\,035$, $3-5\,\text{mm}$-thick images of size 512×512 pixels which were collected from various medical institutions across several countries. In this work, we are interested in comparing performance on binary labels, so we collapse the five classes into "`healthy`" (i.e. 0, no hemorrhage) and "`sick`" (i.e. 1, any type of hemorrhage). We use $80\,\%$ of the data for training and $20\,\%$ for validation. We remark that the splits are created by randomly dividing examinations instead of images. This way, both weak and strong learners can be compared fairly on the same data. Table 3.1 shows the number of positive and negative samples for strong and weak learners in both the training and validation split. Note that the use of a weak learner reduces the number of labels by approximately 35 times.

Since data is stored in DICOM format, we first convert images to their Hounsfield Units (UH) values, window them using the typical brain setting (i.e. L = 40, W = 80), and finally normalize them with min-max normalization.

Experiments were performed on Nvidia Quadro RTX 5000 GPU's. Strong and weak learners solve very different optimization problems, so we train them with those settings that achieve the best validation accuracy in order to guarantee a fair comparison. Specifically, strong learners were trained using Adam (Kingma and Ba, 2014) with a learning rate of 10^{-5}, weight decay of 10^{-7}, and a batch size of 64. Weak learners were trained using Stochastic Gradient Descent with momentum equal to 0.9, a learning rate of 10^{-3}, weight decay of 10^{-4} and a batch size of 1. We train a strong and a weak learner for 15 epochs and keep the best performing one according to the accuracy on the validation set. Both optimizers were scheduled with a learning rate decay of 0.3 every 3 epochs. We remark that the choice of batch size equal to 1 for the weak learners comes both from memory limitations and gradient propagation imbalances through the attention mechanism for volumes with different lengths. Given the high class imbalance and the significant difference in difficulty between predicting the presence of hemorrhage compared to predicting its absence, all models were trained using *focal loss* (Lin et al., 2017)– a variation of binary cross-entropy loss that accounts for both label imbalance and difficulty gaps between classes.

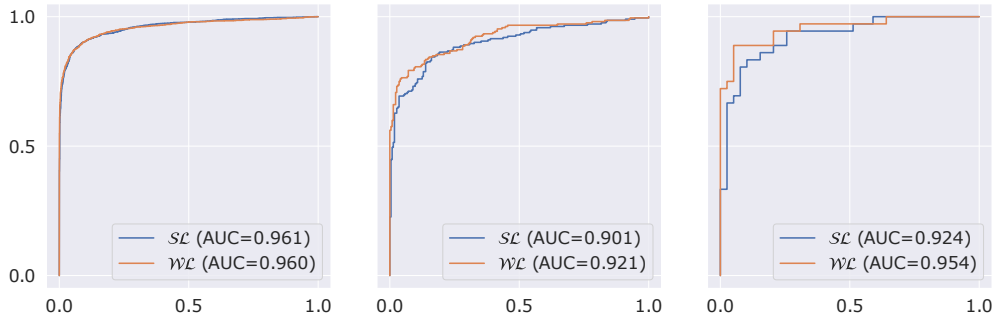
We employ both image-level and examination-level augmentation to regularize the optimization problem. For images, we use TorchIO's[5] (Pérez-García,

[5]available at: https://github.com/fepegar/torchio

Sparks, and Ourselin, 2021) library of spatial and intensity transformations. Specifically, every image is augmented independently via random flips, affine transformations, deformations and rotations, and one out of addition of random noise, random bias field, random anisotropy, random gamma transformation, or random ghosting artifact. For examinations, we randomly sample (without replacement) between 10 and r images, where r is the number of images in the examination. We remark that the sub-sampling process does not rely on knowing image labels, and can be used in practical scenarios where image labels may not be available. Sampling at least 10 images controls the probability of flipping a positive examination to a negative one. That is, sampling a subset of all negative images from a positive examination would result in a noisy label, i.e. the subset would be labeled positively even if it did not contain any positive images. Although we cannot completely rule out this event without knowing local labels, we can reduce its probability such that the level of noise in the labels is tolerable to the learner.

3.1.3 Testing strong and weak learners on the CQ500, BHX, and CT-ICH datasets.

Generalization power is one of the most important characteristics of a good machine learning model. Besides being able to transfer to data coming from a different institution, a model may face shifts within the clinical setting it is deployed in (e.g. machines could be updated, imaging methods could change, and patient demographic may evolve over time). Hence, we test both strong and weak learners on two external datasets–the CQ500 dataset (Chilamkurthy et al., 2018) and the CT-ICH dataset (Hssayeni, 2020; Hssayeni



(a) RSNA dataset. **(b)** CQ500 dataset. **(c)** CT-ICH dataset.

Figure 3.1: Comparison of strong and weak learners on the examination-level binary classification MIL problem.

et al., 2020; Goldberger et al., 2000). The preprocessing pipeline for these dataset is the same as the one used for the RSNA 2019 Brain CT Hemorrhage Challenge dataset: we first select the non-contrast brain examinations with a slice thickness between 3 and 5 mm, then convert the DICOM or NIfTI images to their HU values, window them with the typical brain setting, and finally normalize with min-max normalization. Table 3.2 shows the number of examinations contained in the two datasets. The CT-ICH dataset provides labeled examinations with manual segmentations of the bleed, while the original CQ500 dataset only provides labeled examinations. Thus, we use the bounding boxes in the BHX extension dataset (Reis et al., 2020; Goldberger et al., 2000) to evaluate hemorrhage detection. We remark that the RSNA 2019 Brain CT Detection Challenge dataset does not provide segmentations of the bleeds.

3.1.4 Comparison on bag-level retrieval.

First of all, we compare strong and weak learners on the examination-level MIL binary classification problem. Indeed, recall from Eq. (3.1) that the label of a CT examination can be phrased as the OR function of the image-level rule $h^*(X) = \mathbb{1}[X \text{ contains hemorrhage}]$. Furthermore, recall that a weak learner can predict on entire examinations, while a strong learner can predict on single images only. Let $\widehat{\mathcal{SL}}$ be the fully-supervised model obtained after training. Then, given a new CT examination $\overline{X} \in \mathcal{X}^r$, we define the bag-level prediction of $\widehat{\mathcal{SL}}$ as the maximum image-level prediction in $\overline{X}$. That is $\overline{Y}_{\widehat{\mathcal{SL}}} := \max(\widehat{\mathcal{SL}}(X_1), \ldots, \widehat{\mathcal{SL}}(X_r))$, which naturally extends the OR function to real-valued functions in the interval $[0, 1]$. Fig. 3.1 shows the ROC curves for strong and weak learners on the three dataset. In particular, Fig. 3.1a shows that there is virtually no difference in performance between fully- and weakly-supervised models on the validation split of the RSNA 2019 Brain CT Hemorrhage Challenge dataset, while Fig. 3.1b and 3.1c show that weak learners have a slight advantage on the external datasets. That is, weak learners can generalize better compared to strong learners for the examination-level MIL binary classification problem. This behavior agrees with intuition, since the weak learners are trained directly on the bag-level classification problem. Yet, it is still a surprising result given the more detailed information available to a strong learner during training.

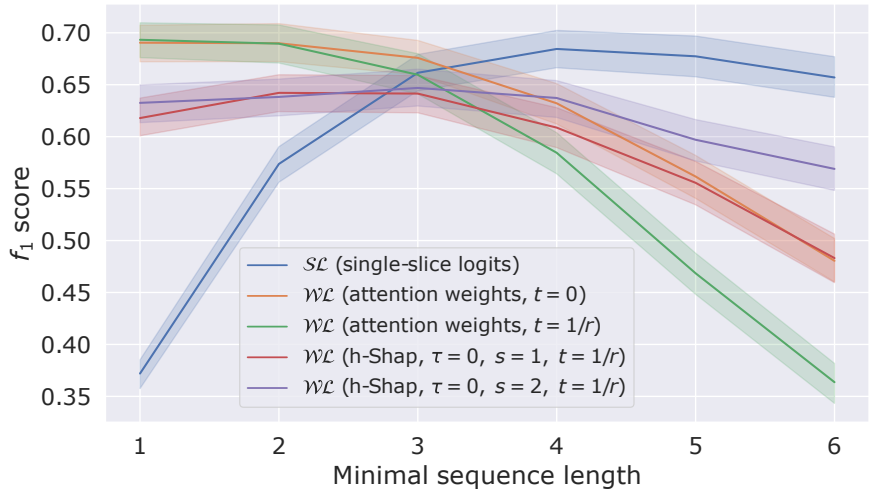


Figure 3.2: Hemorrhage location performance on the RSNA 2019 Brain CT Hemorrhage Challenge dataset. τ and s are respectively the importance tolerance and minimal feature size in h-Shap, t is the threshold used to locate predicted hemorrhage sequences.

3.1.5 Comparison on hemorrhage location.

Recall that we refer to hemorrhage location as the task of selecting the images that contain hemorrhage within an examination. For a strong learner trained on image-level labels, this task is no different than predicting on single images. Indeed, given a new CT examination $\overline{X} \in \mathcal{X}^r$, a strong learner should select the predicted positive images by making a prediction for each image in the scan independently.

On the other hand, for a weak learner as in Eq. (3.4), one can adopt different strategies. Intuitively, the attention weights should be larger for the positive instances in predicted positive bags. That is, one could select those instances whose attention weights are above a certain threshold. We propose and compare two choices of attention threshold, t: an absolute threshold of $t = 0$, and a relative threshold equal to $t = 1/r$ which corresponds to

uniform contribution of every instance in a bag of size r. Furthermore, we introduce a Shapley-coefficient based explanation method for bags. Indeed, while attention weights are extensively used in recent literature as proxies to interpretability, they currently lack of theoretical results which ensure their validity. On the contrary, Shapley coefficients satisfy several desirable theoretical properties (Shapley, 1953). In Chapter 2, we introduced *h-Shap*[6]–a fast and exact method for the computation of Shapley coefficients when data satisfies some MIL assumption in the form of Eq. (3.2). Although the method was introduced in the context of image explanations, it can be extended to bags of instances. In fact, recall that the binary classification problem on examinations satisfies the MIL assumption in Eq. (3.1). Hence, given a new CT examination $\overline{X} \in \mathcal{X}^r$, one can explore a binary-tree of $\overline{X}$ and compute the exact Shapley coefficient of every image in the scan. The symmetry axiom of Shapley values (i.e. equally important players receive equal attributions) implies that every positive instance in a positive bag will receive the same coefficient. Thus, one can use a relative threshold of $t = 1/r$ and select those images whose Shapley coefficients are equal or above t. Such an explanation method for the bag-level predictions of a weak learner is particularly attractive because it does not require to sample an uninformative baseline value. Indeed, recall that a weak learner is permutation invariant and it can predict on inputs of arbitrary size. Thus, one can simply remove images from an examination without having to replace them with an unimportant reference value.

To summarize, when using a strong learner, hemorrhage location is equivalent to selecting the predicted positive images in a CT examination. In the case

[6]available at: https://github.com/Sulam-Group/h-shap

of a weak learner, one can: (i) select the most attended images by thresholding the attention weights, or (ii) select the images with a Shapley coefficient greater or equal to $1/r$. Fig. 3.2 shows hemorrhage location performance of strong and weak learners as a function of minimal sequence length, i.e. the minimal number of consecutive images that have to be selected in order to consider them as a candidate hemorrhage sequence. Indeed, Fig. 3.2 shows that the hemorrhage location performance of both weak and strong learners is comparable. In particular, strong learners tend to have a higher false positive rate which hurts performance with a minimal sequence length of 1 and 2. Importantly, weak learners with attention weights achieve the same performance of strong learners. This result suggests that weak learners are capable of learning the true image-level concept without local labels.

We now formally describe how the f_1 score for hemorrhage location is computed. Given a CT examination $\overline{X} \in \mathcal{X}^r$, we define an *hemorrhage sequence* to be a set of consecutive images that contain hemorrhage. That is, a CT scan can be regarded as the concatenation of hemorrhage and non-hemorrhage sequences. In the presence of local labels (i.e. a ground-truth label for every image), hemorrhage location may be evaluated in terms of the area under the ROC curve of the image-level binary classification problem. However, these local-label figures of merit do not provide a sense of performance in terms of the number of true hemorrhage sequences (which might correspond to more than one positive image slice) that are retrieved by a model. Indeed, in clinical settings, we are interested in quantifying how well a machine learning model locates the hemorrhage sequences in a scan in order to support

a radiologist in their analysis. Thus, we propose an alternative method to evaluate hemorrhage location. Recall that depending on the learner, we can use different local estimators of the image-level labels, i.e. the image predictions for a strong learner, or either the attention weights or the Shapley coefficients for a weak learner. To find the predicted hemorrhage sequences, then, we can threshold the local estimator and find the sets of consecutive images that are above the threshold. Formally, let $T = \{T_1, \ldots, T_n\},\ T_i \subseteq [r]$ be the true hemorrhage sequences in $\overline{X}$ such that $T_i \cap T_j = \varnothing,\ \forall i \neq j$ and $\cup_{i=1}^{n} T_i \subseteq [r]$. That is, T_i contains the indices of the images that compose the i-th hemorrhage sequence in $\overline{X}$, and $|T|$ corresponds to the number of hemorrhages in $\overline{X}$. Let $\widehat{Y} = (\widehat{y}_1, \ldots, \widehat{y}_r) \in (\mathbb{R}^+)^r$ indicate the local estimator used to locate hemorrhages, and let $P = \{P_1, \ldots, P_m\},\ P_i \subseteq [r],\ |P_i| \geq L$ be the predicted hemorrhage sequences with at least L images obtained by thresholding $\widehat{Y}$. Analogously to T, it holds that $P_i \cap P_j = \varnothing,\ \forall i \neq j$ and $\cup_{i=1}^{n} P_i \subseteq [r]$. We define the true positives TP, false positives FP, and predicted positives PP as follows. For every true hemorrhage sequence $T_i \in T$, we count one true positive prediction if there exists a predicted hemorrhage sequence $P_j \in P$ such that the image with the largest estimator value within P_j is in T_i up to an offset of d (specifically, we show results for $d = 2$). Similarly, we count one false positive prediction for every predicted sequence P_j such that there does not exist a corresponding true one. Note that the definition of true positives above avoids double counting predicted sequences that correspond to the same true one, and that using the arg max penalizes those cases where models may predict a few long sequences that include multiple true ones.

Formally,

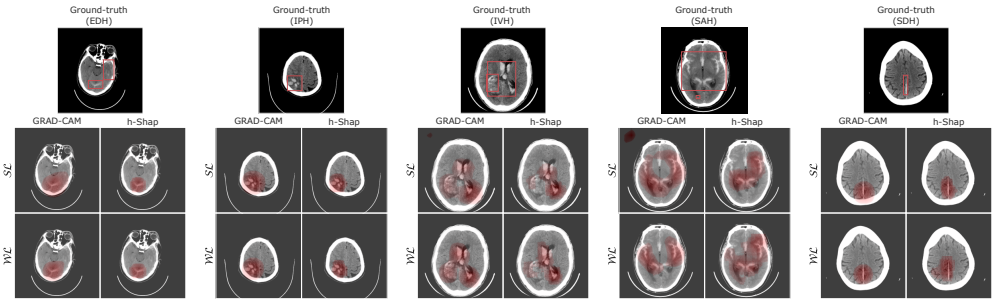
$$\begin{aligned} \text{TP} &:= \#\{i \in |T|,\ \exists P_j \in P :\ \arg\max_{k \in P_j} \widehat{y}_k \in \widetilde{T}_i\} \\ \text{FP} &:= \#\{j \in |P|,\ \nexists T_i \in T :\ \arg\max_{k \in P_j} \widehat{y}_k \in \widetilde{T}_i\} \\ \text{PP} &:= \text{TP} + \text{FP}, \end{aligned} \tag{3.5}$$

where $\widetilde{T}_i := \{\max(0, \min T_i - d), \ldots, T_i, \ldots, \min(r, \max T_i + d)\}$, and $d \in \mathbb{N}$ is a detection offset set to relax the definition of true positives to cases where a model may select images that are close to the boundaries of a true hemorrhage sequence. To conclude, the f_1 score is the harmonic average of precision and recall, that is

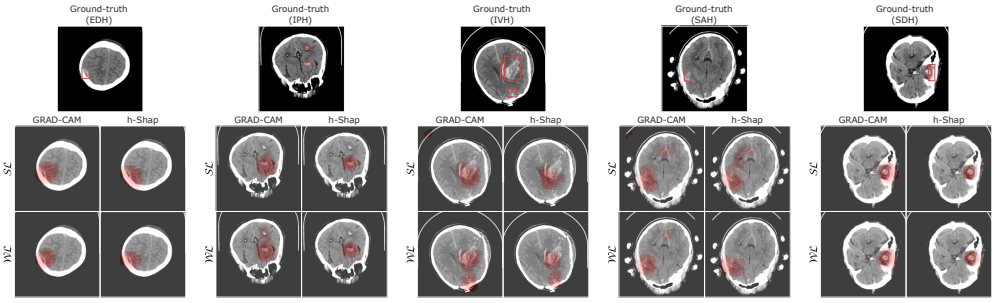
$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{PP}}, \quad \text{recall} = \frac{TP}{|T|} \\ f_1 &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{3.6}$$

3.1.6 Comparison on hemorrhage detection.

So far, we have phrased hemorrhage location and detection as parts of MIL classification problems. However, we are ultimately interested in locating hemorrhage within images. Thus, methods that bridge classification and detection are needed. This problem has been extensively studied in the computer vision literature (Girshick, 2015; Oquab et al., 2015; Zhao et al., 2019; Wang et al., 2021), which largely inspire MIL-based WSOD methods (Bilen and Vedaldi, 2016; Wan et al., 2019; Tang et al., 2017). At the same time, others (Zhou et al., 2016; Selvaraju et al., 2017; Ribeiro, Singh, and Guestrin, 2016; Lundberg and Lee, 2017) have developed explanation methods, which are designed to find the most important features towards a model's prediction. In



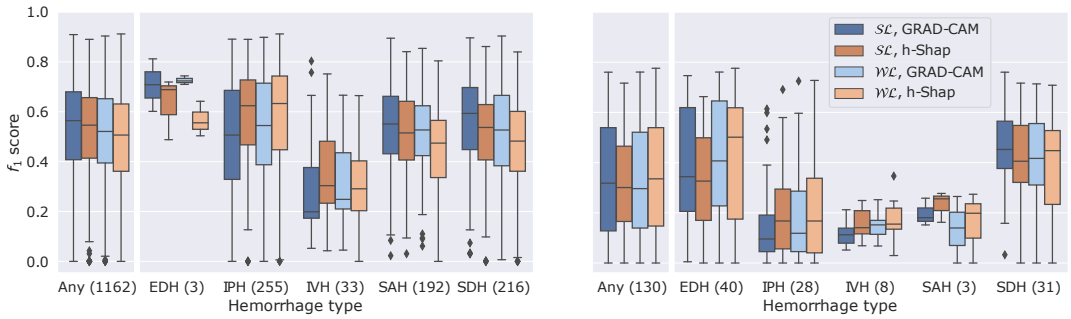
(a) CQ500 dataset.



(b) CT-ICH dataset.

Figure 3.3: Example saliency maps on some predicted positive images that contain hemorrhage. We use h-Shap with an absolute importance tolerance of $\tau = 0$ (i.e. h-Shap explores all partitions with a positive Shapley coefficient), minimal feature size $s = 64$, number of radii $\eta = 3$, and number of angles $\beta = 12$. We apply GRAD-CAM to the last convolutional layer of both strong and weak learners. All saliency maps are thresholded using Otsu's method to reduce noise.

the context of vision models, explanation methods usually produce *saliency maps*, i.e. filters that highlight those regions that contributed the most to the prediction. Even though these methods usually provide a weaker sense of detection, they have gained significant popularity in the medical image analysis field for their ability to explain *black-box* models (Guidotti et al., 2018; Zednik, 2021; Eschenbach, 2021). That is, as models keep getting more and more complex, they become *opaque* to their users–the mechanics of a model's decision making process cannot be understood intuitively. This lack of transparency raises concerns regarding the trustworthiness, fairness, and safety of



(a) CQ500 dataset. **(b)** CTICH dataset.

Figure 3.4: Hemorrhage detection performance for weak and strong learners on the CQ500 and CTICH datasets with saliency maps obtained with GRAD-CAM and h-Shap. The f_1 scores are computed between the thresholded saliency maps and the ground truth bounding box annotations. For a fair comparison, we show the f_1 score distributions of true positive images explained both by the weak and strong learners (i.e. 1162 images for the CQ500 dataset and 130 images for the CT-ICH dataset).

deep learning models in high-stakes scenarios (Rudin, 2019).

In this chapter, we perform hemorrhage detection through saliency maps. Specifically, we use GRAD-CAM (Selvaraju et al., 2017) for its popularity and h-Shap given its relation to multiple-instance problems. We compare the detection performance of strong and weak learners by means of the pixel-level f_1 score of the saliency maps with the ground truth annotations provided in the CT-ICH dataset and the BHX extension of the CQ500 dataset. We note that the use of the pixel-level f_1 score avoids double-counting in the case of overlapping ground truth annotations by only considering their union. The BHX dataset provides bounding boxes around the bleeds in an image, while the CT-ICH provides precise segmentations of the bleeds. Hence, we convert the segmentations in the CT-ICH dataset to their respective bounding boxes.

Bleeds can present complex and irregular shapes. However, h-Shap explores fixed quad-trees of the input image. Thus, we extend the original implementation with ideas of cycle spinning (Coifman and Donoho, 1995). Let s denote the minimal size of the partitions explored by h-Shap, $\rho := \{i \cdot s/\eta\}_{i=1}^{\eta}$ be η equally spaced radii between s/η and s, and let $\alpha = \{j \cdot 2\pi/\beta\}_{j=1}^{\beta}$ be β equally spaced angles between $2\pi/\beta$ and 2π. Then, $\Phi_{s,\rho,\alpha} := 1/(\eta\beta) \cdot \sum_{i\in[\eta],\, j\in[\beta]} \Phi_{s,\rho_i,\alpha_j}$ is the average over all saliency maps Φ_{s,ρ_i,α_j} obtained by rolling the partitions in the direction of the polar vector (ρ_i, α_j). As pointed out in Chapter 2, the unconditional expectation over the training dataset is a valid baseline value for MIL problems that satisfy Eq. (3.2).

Fig. 3.3 shows the saliency maps obtained with strong and weak learners using GRAD-CAM and h-Shap on one image for every type of hemorrhage in both datasets (note that for the CQ500 dataset, we group examinations labeled as "`chronic`" together with those labeled as "`subdural`"). To reduce noise in the explanations, we threshold them using Otsu's method (Otsu, 1979).

Qualitatively, we can see that there is virtually no difference in the explanations produced by strong and weak learners, and that the extension of h-Shap to flexible partitions does allow for the saliency maps to capture the complex shapes of the bleeds. This result confirms that both fully-supervised and weakly-supervised models do in fact learn the true target concept, hemorrhage, versus other findings that may be correlated with the presence of hemorrhage. For example, Fig. 3.3 shows images with signs of external brain hemotomas and skull fractures. Although a model may learn to rely on these findings to predict the presence of hemorrhage, the saliency maps highly

concentrate around the bleeds. Furthermore, even though learners are trained on binary labels, they are capable of correctly identifying all five types of hemorrhage.

Quantitatively, Fig. 3.4 shows the distribution of the f_1 scores of the saliency maps divided by hemorrhage type, learner, and explanation method. For strong learners, we explain all predicted positive images, while for weak learners, we explain all images with a Shapley coefficient greater than $1/r$, where r is the number of images in the examination. We restrict the hemorrhage detection comparison to the true positive images that are explained by both strong and weak learners. We note that in order to compare performance across types of hemorrhage, we only consider those images that contain only one type of hemorrhage. Fig. 3.4 shows that the performance of strong and weak learners is comparable, with no clear advantage for models trained with full supervision. Overall, the performance on the CT-ICH dataset is lower compared to the CQ500 dataset.

3.1.7 Label complexity results.

As noted in at the beginning of this Chapter, supervised-learning approaches to medical imaging problems are currently limited by the amount of labels needed. Indeed, expert radiologists are often asked to perform time-consuming tasks such as manually annotating or labeling large volumes of images. On the other hand, medical institutions have now accumulated significant amounts of data thanks to the widespread adoption of Electronic Health Record (EHR) systems. The potential of these systems to improve

the current state-of-the-art may often go untapped because of the expensive data-curation process. Following (Sabato, Srebro, and Tishby, 2010), we define *label complexity* as the number of labels required by a machine learning model to achieve a certain level of performance.

Until now, we have presented results for strong and weak learners trained on the entire RSNA 2019 Brain CT Hemorrhage Challenge dataset. As shown in Table 3.1, strong learners have access to $\approx$ 600 000 labels, while weak learners can only access $\approx$ 17 000. Hence, we compare the performance of strong and weak learners on both the examination-level binary classification problem and on hemorrhage location when trained on the same number of labels. Intuitively, assuming that the cost to ask for a radiologist to label an examination is the same as the one for a radiologist to label an image, label complexity represents the cost of the data curation process. In practice, however, it is a much simpler task to quickly go through the images in an examination and determine whether it contains hemorrhage, rather than having to analyze each image individually. Let m be the number of labels available to each learner, we train strong and weak learners on random subsets of size m sampled from the original training split. Note that in order to account for the variance in the training process, we train a decreasing number of duplicates–from 20 to 1–as m increases. Similarly to training on the entire RSNA 2019 Brain CT Hemorrhage Challenge dataset, for strong learners, we use Adam (Kingma and Ba, 2014) with a learning rate of 10^{-5}, weight decay of 10^{-7}, and a batch size of 64. For weak learners, we use Stochastic Gradient Descent with momentum equal to 0.9, a learning rate of 10^{-3}, weight

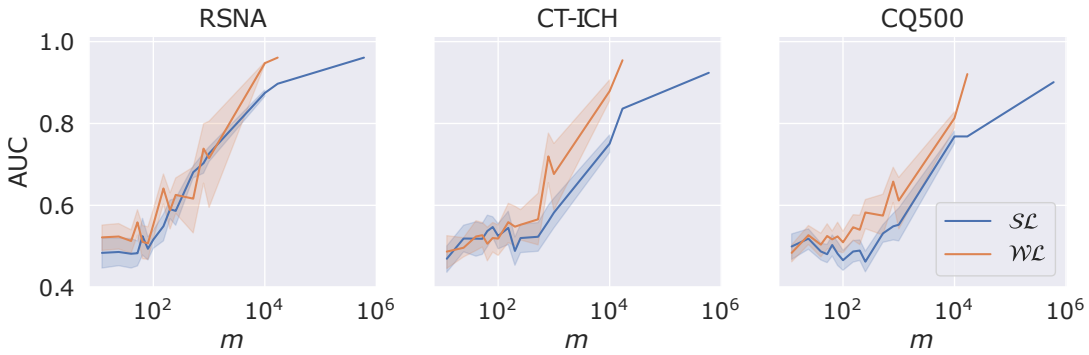


Figure 3.5: Mean performance of strong ($\mathcal{SL}$) and weak ($\mathcal{WL}$) learners on the examination-level binary classification problem as a function of number of labels m. For the RSNA 2019 Brain CT Hemorrhage Challenge dataset, we validate models on a fixed subset of 1000 examinations. We note that the confidence intervals around the mean vanish after $m = 10^4$ since we only train one duplicate for each learner.

decay of 10^{-4} and a batch size of 1. Both optimizers were scheduled with a learning rate decay of 0.3 every 3 epochs. Instead of training for a fixed number of epochs, we terminate when the accuracy on the validation split has not increased for more than 3 epochs. We evaluate all models on a fixed subset of 1000 examinations from the validation split of the RSNA 2019 Brain CT Hemorrhage Challenge dataset, and on the entire CQ500 and CT-ICH datasets.

Fig. 3.5 shows the mean performance of strong and weak learners on the examination-level binary classification problem when trained on the same number of labels m. Indeed, weakly-supervised models generalize better than fully-supervised ones. However, the examination-level performance cannot be used to compare how well models have learned the true concept of interest. As we noted in Sec. 3.1, a bag-level MIL model may classify bags with high accuracy without learning the true target hypothesis. Thus, Fig. 3.6 shows the comparison of strong and weak learners on hemorrhage location

Figure 3.6: Mean hemorrhage location performance as a function of number of labels m on a fixed subset of 1000 examinations in the RSNA 2019 Brain CT Hemorrhage Challenge dataset. For weak learners, we consider candidate sequences with at least 2 images, and we require at least 4 images for strong learners.

when trained on the same number of labels m. Interestingly, we see that weak supervision can lead to improvements in hemorrhage location performance compared to full supervision. Indeed, one should prefer labeling bags rather than single instances as the amount of data available increases. Fig. 3.6 shows that there persists an advantage in labeling individual images when very limited data is available, and that asymptotically (i.e. with virtually infinite amount of data) full supervision does not provide significant advantage over weak supervision.

References

Hssayeni, Murtadha D, Muayad S Croock, Aymen D Salman, Hassan Falah Al-khafaji, Zakaria A Yahya, and Behnaz Ghoraani (2020). "Intracranial hemorrhage segmentation using a deep convolutional model". In: *Data* 5.1, p. 14.

Lee, Hyunkwang, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H Tajmir, Claude E Guerrier, Sarah A Ebert, Stuart R Pomerantz, Javier M Romero, Shahmir Kamalian, et al. (2019). "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets". In: *Nature biomedical engineering* 3.3, pp. 173–182.

Salehinejad, Hojjat, Jumpei Kitamura, Noah Ditkofsky, Amy Lin, Aditya Bharatha, Suradech Suthiphosuwan, Hui-Ming Lin, Jefferson R Wilson, Muhammad Mamdani, and Errol Colak (2021). "A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography". In: *Scientific reports* 11.1, pp. 1–11.

Ilse, Maximilian, Jakub Tomczak, and Max Welling (2018). "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR, pp. 2127–2136.

Dietterich, Thomas G, Richard H Lathrop, and Tomás Lozano-Pérez (1997). "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial intelligence* 89.1-2, pp. 31–71.

Maron, Oded and Tomás Lozano-Pérez (1997). "A framework for multiple-instance learning". In: *Advances in neural information processing systems* 10.

Weidmann, Nils, Eibe Frank, and Bernhard Pfahringer (2003). "A two-level learning method for generalized multi-instance problems". In: *European Conference on Machine Learning*. Springer, pp. 468–479.

Flanders, Adam E, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein,

Felipe C Kitamura, Matthew P Lungren, et al. (2020). "Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge". In: *Radiology: Artificial Intelligence* 2.3, e190211.

Chilamkurthy, Sasank, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier (2018). "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study". In: *The Lancet* 392.10162, pp. 2388–2396.

Reis, Eduardo Pontes, Felipe Nascimento, Mateus Aranha, F Mainetti Secol, Birajara Machado, Marcelo Felix, Anouk Stein, and Edson Amaro (2020). *Brain hemorrhage extended (bhx): Bounding box extrapolation from thick to thin slice ct images*.

Goldberger, Ary L, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley (2000). "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *circulation* 101.23, e215–e220.

Hssayeni, Murtadha (2020). "Computed tomography images for intracranial hemorrhage detection and segmentation". In: *Intracranial Hemorrhage Segmentation Using A Deep Convolutional Model. Data* 5.1.

Bilen, Hakan and Andrea Vedaldi (2016). "Weakly supervised deep detection networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2846–2854.

Tang, Peng, Xinggang Wang, Xiang Bai, and Wenyu Liu (2017). "Multiple instance detection network with online instance classifier refinement". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2843–2851.

Wan, Fang, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye (2019). "C-mil: Continuation multiple instance learning for weakly supervised object detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2199–2208.

Shao, Feifei, Long Chen, Jian Shao, Wei Ji, Shaoning Xiao, Lu Ye, Yueting Zhuang, and Jun Xiao (2022). "Deep Learning for Weakly-Supervised Object Detection and Localization: A Survey". In: *Neurocomputing*.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár (2017). "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Pérez-García, Fernando, Rachel Sparks, and Sebastien Ourselin (2021). "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning". In: *Computer Methods and Programs in Biomedicine* 208, p. 106236.

Shapley, Lloyd S (1953). "A value for n-person games". In: *Contributions to the Theory of Games* 2.28, pp. 307–317.

Girshick, Ross (2015). "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448.

Oquab, Maxime, Léon Bottou, Ivan Laptev, and Josef Sivic (2015). "Is object localization for free?-weakly-supervised learning with convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 685–694.

Zhao, Zhong-Qiu, Peng Zheng, Shou-tao Xu, and Xindong Wu (2019). "Object detection with deep learning: A review". In: *IEEE transactions on neural networks and learning systems* 30.11, pp. 3212–3232.

Wang, Jianfeng, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng (2021). "End-to-end object detection with fully convolutional network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15849–15858.

Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba (2016). "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""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, pp. 1135–1144.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi (2018). "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5, pp. 1–42.

Zednik, Carlos (2021). "Solving the black box problem: a normative framework for explainable artificial intelligence". In: *Philosophy & Technology* 34.2, pp. 265–288.

Eschenbach, Warren J von (2021). "Transparency and the black box problem: Why we do not trust AI". In: *Philosophy & Technology* 34.4, pp. 1607–1622.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215.

Coifman, Ronald R and David L Donoho (1995). "Translation-invariant denoising". In: *Wavelets and statistics*. Springer, pp. 125–150.

Otsu, Nobuyuki (1979). "A threshold selection method from gray-level histograms". In: *IEEE transactions on systems, man, and cybernetics* 9.1, pp. 62–66.

Sabato, Sivan, Nathan Srebro, and Naftali Tishby (2010). "Reducing label complexity by learning from bags". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, pp. 685–692.

Chapter 4

Conclusions

In this thesis, we discussed multiple-instance learning (MIL)–a generalization of classical supervised learning to bags of inputs–as a framework to explain with Shapley coefficients. In Chapter 1 we presented the MIL framework, its relevance to weakly-supervised learning and its two main formulations: (i) instance-level, and (ii) bag-level MIL. We presented how modern deep learning architectures, such as attention mechanisms, are being deployed for bag-level classification tasks, and showed that these concepts had been previously introduced in the weakly supervised object detection (WSOD) literature, which is closely related to MIL. Then, we briefly introduced the Shapley value, its axioms, and how they have been translated into the machine learning literature in the context of model explainability. In Chapter 2, we deployed the MIL framework and ideas of hierarchical partitioning for images to introduce h-Shap: the first fast, scalable, and exact explanation method for images based on a hierarchical extension of Shapley coefficients. We showed that when the data distribution satisfies some multiple instance learning assumption,

the computational complexity of Shapley coefficients can be reduced exponentially while providing a precise trade-off between speed and accuracy of the results. We studied h-Shap on three representative datasets from different fields and of varying complexity, and we compared h-Shap with other popular explanation methods, both Shapley and non-Shapley based. Our results show that h-Shap consistently outperforms the current state-of-the-art on speed and/or object retrieval performance by orders of magnitude across the datasets explored. Importantly, we show that a principled and informed approach to explainability can lead to methods that provide mathematical guarantees on their performance, overcoming some of the limitations of current popular methods. Furthermore, we discuss the current limitations and assumptions of h-Shap. For example, in Chapter 2, we use fixed quadtrees of the input image for their simplicity and longstanding history in computer vision. However, ideas of semantic segmentations could be deployed in order to explore partitions that respect the underlying structure of the images. Lastly, as it is for other explanation methods based on feature removal, the choice of masking features with their unconditional expectation over the training set may not be appropriate in certain scenarios where feature independence is not satisfied, even in a local sense. Future work entails extending current approaches to efficiently sample the baseline from its conditional distribution.

In Chapter 3, we presented a case application of h-Shap and MIL for intracranial hemorrhage detection in CT examinations. Supervised learning still remains the most popular approach for deep learning solutions in healthcare, but these methods are strongly limited by the amount of labels that expert

radiologists need to label in order for models to achieve the necessary performance levels. On the other hand, it is unclear what kind of labels should be collected. Specifically, we compare the performance between models trained with local (image-level) labels and global (examination-level) labels. We proposed an attention-based, bag-level MIL classifier binary classifier that can be trained under both supervision regimes. In particular, the use of an attention mechanisms allows us to predict on examinations of arbitrary length. We introduced two extensions of the original implementation of h-Shap: (i) a bag-level explanation method that can be used to retrieve the positive instances in a positive bag, and (ii) a cycle-spinning-based strategy to capture the complex shapes of the bleeds. We showed that fully-supervised learners do not show significant performance improvements on both hemorrhage location (within examinations), and detection (within images), both on the validation split of the training dataset, and on two external test datasets. Furthermore, we compared strong and weak learners when trained with the same number of labels. Importantly, answering this questions is fundamental importance to inform how to collect and label new datasets. We found that weakly-supervised learners trained on examination-level binary labels can outperform fully-supervised ones trained on image-level labels. This result suggests that radiologists should not be recruited to label individual images but rather entire examinations, which would drastically reduce the time and costs associated with the data curation process of new datasets. These promising empirical results open several theoretical questions that we will address in future work. For example, can simple attention mechanisms express boolean functions of the inputs and extend the original MIL assumption? Is it possible

to precisely characterize the trade-off between learning from bags rather than instances for neural networks? Can we extend some known generalization bounds to attention-based bag classifiers?

The multiple-instance learning framework still offers several interesting research topics. In fact, it has been hardly studied in the context of deep learning or other modern machine learning models due to its complexity as soon as one relaxes its simple, original assumptions. We hope to shed a light on these topics in an effort to develop more reliable and trustworthy models to be deployed safely in real-world scenarios in order to improve human health.

Chapter 5

Appendix

5.1 Proofs

We summarize here the assumptions and notation used in the following results. Let $X \in \mathbb{R}^n$ be drawn so that each entry $x_i \sim a_i\mathcal{I} + (1 - a_i)\mathcal{I}^c$, where $a_i \sim \text{Bernoulli}(\rho)$ is a binary random variable that indicates whether the feature x_i comes from an *important* distribution $\mathcal{I}$, or its *non-important* complement $\mathcal{I}^c$. Let

$$\hat{f}(X_C) = 1 \iff \exists i \in C : \; x_i \sim \mathcal{I}, \; C \subseteq [n],$$

where $n \coloneqq \{1, \dots, n\}$ and $X_C \in \mathbb{R}^n$ is equal to X in the entries of C and takes value in the baseline in its complement $\bar{C}$. We denote with $\Phi_{(X,\hat{f})} = \{\phi_1(\hat{f}), \dots, \phi_n(\hat{f})\} \in \mathbb{R}^n$ the saliency map of X where $\phi_i(\hat{f})$ is the Shapley coefficient of x_i. Let $k = \|\Phi_{(X,\hat{f})}\|_0$ be the number of reported important features by the exact Shapley coefficients. We showed earlier (see Eq. (2.6))

that under these assumptions, it follows:

$$\phi_i(\hat{f}) = \begin{cases} 1/k & \text{if } x_i \sim \mathcal{I} \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, let $\mathcal{T}_0 = (S_0, \mathcal{T}_1, \ldots, \mathcal{T}_\gamma)$ be the recursive definition of a γ-partition tree of X such that $S_0 = [n]$; $\mathcal{T}_i, \ldots, \mathcal{T}_\gamma$ are the subtrees branching off of S_0; and $c(S_i)$ are the γ children of the node S_i. Recall that h-Shap explores $\mathcal{T}_0$ from S_0 and returns all relevant leaves $L \subseteq [n]$ such that their Shapley coefficient is greater than a relevance tolerance τ. We denote with $\hat{\mathcal{T}}_0$ the subtree composed of the nodes visited by h-Shap, and with $\widehat{\Phi}_{(X,\hat{f})}$ the saliency map computed by h-Shap, such that

$$\widehat{\phi}_i(\hat{f}) = \begin{cases} 1/|L| & \text{if } i \in L \\ 0 & \text{otherwise.} \end{cases}$$

Now, we will provide proof of the Theorems presented in Sec. **??**.

5.1.1 Expected number of visited nodes 2.0.2

Here, we are interested in evaluating the expected number of nodes visited by h-Shap, to better characterize its computational advantage.

Proof. Recall that S_0 contains all features of X. That is, $S_0 = [n]$. Since $x_1, \ldots, x_n \in X$ are iid, so are groups of features. Then, it suffices to analyze each child of S_0 independently. Consider the two mutually exclusive events on the child node $c_i \in c(S_0)$: **1)** it does not contain any important features, i.e. $\nexists\, j \in c_i : \hat{f}(X_j) = 1$; and **2)** it contains at least one important feature, i.e. $\exists\, j \in c_i : \hat{f}(X_j) = 1$. Let $p_1(S_0)$ be the probability of event 1, and $1 - p_1(S_0)$ be the probability of event 2. When event 2 occurs, we add one node to $\hat{\mathcal{T}}_0$,

and then we explore the subtree $\hat{\mathcal{T}}_i$ branching off of c_i. We can recursively apply this strategy to each subtree of $\hat{\mathcal{T}}_0$, which yields

$$E[|\hat{\mathcal{T}}_0|] = 1 + \gamma(1 - p_1(S_0))E[|\hat{\mathcal{T}}_1|]. \tag{5.1}$$

We are left with evaluating $p_1(S_0)$, which simply is

$$p_1(S_0) = (1-\rho)^{|S_0|/\gamma} \tag{5.2}$$

since the probability for x_k not to be important, i.e $x_k \sim \mathcal{I}^c$, is $(1-\rho)$, and all the children $c(S_0)$ have cardinality $n/\gamma = |S_0|/\gamma$ because they form a disjoint symmetric partition of S_0. When analyzing the i^{th} subtree branching off of S_0, $\hat{\mathcal{T}}_i$, one has to condition on the probability of the event that S_i contains at least one important feature. The probability $p'(S_i)$ of the event that S_i contains at least one important feature is, again, simply $1-(1-\rho)^{|S_i|}$. Therefore

$$p_1(S_i) = \frac{(1-\rho)^{|S_i|/\gamma}(1-(1-\rho)^{|S_i|(\gamma-1)/\gamma})}{1-(1-\rho)^{|S_i|}} \tag{5.3}$$

is the conditioned probability that a child of S_i does not contain any important features. □

5.1.2 Similarity lower bound 2.0.4

Here, we want to find the lower bound of the similarity between Φ and $\widehat{\Phi}$, defined as

$$\alpha = \frac{\langle \Phi, \widehat{\Phi} \rangle}{\|\Phi\|_2 \|\widehat{\Phi}\|_2}.$$

Proof. Let $k = \|\Phi\|_0$ be the number of reported important features in X as returned by the Shapley coefficients. Let $L \subseteq [n]$ be the relevant leaves

returned by h-Shap. From Eq. (2.6) and (2.3) it follows that

$$\|\Phi\|_2 = \sqrt{\frac{1}{k^2}k} = \sqrt{\frac{1}{k}}, \tag{5.4}$$

$$\|\widehat{\Phi}\|_2 = \sqrt{\frac{1}{(\ell s)^2}\ell s} = \sqrt{\frac{1}{\ell s}}, \tag{5.5}$$

where $|L| = \ell s$, ℓ is the number of relevant leaves, and s is the minimal feature size. Furthermore, we know that

$$\langle \Phi, \widehat{\Phi} \rangle = k \left(\frac{1}{k} \frac{1}{\ell s} \right) = \frac{1}{\ell s}. \tag{5.6}$$

Therefore

$$\alpha = \frac{\langle \Phi, \widehat{\Phi} \rangle}{\|\Phi\|_2 \|\widehat{\Phi}\|_2} = \frac{\frac{1}{\ell s}}{\frac{1}{\sqrt{\ell s k}}} = \sqrt{\frac{k}{\ell s}}. \tag{5.7}$$

Fixed s and k, α is a monotonically decreasing function of ℓ, which means that minimizing the similarity between Φ and $\widehat{\Phi}$ is equivalent to maximizing the number of leaves returned by h-Shap. When $k \leq n/s$, $\ell \leq k$, so $\alpha \geq \sqrt{k/(ks)} = 1/\sqrt{s}$. When $k > n/s$, $|L| = n$, therefore $\alpha \geq \sqrt{k/n}$. □

5.2 Algorithms

Algorithm 2 describes the breadth-first version of h-Shap presented in Sec. **??**. We recall that both implementations return the set of relevant leaves $L \subseteq [n] := \{1, \ldots, n\}$ such that their Shapley coefficients are greater than a relevance tolerance τ. dh-Shap uses an absolute tolerance, while bh-Shap uses a relative tolerance.

Algorithm 2 Breadth-first h-Shap

procedure bh-Shap($X, \mathcal{T}_0, \hat{f}$)
inputs: image X, threshold $\tau \geq 0$, trained model $\hat{f}$
    $L \leftarrow \emptyset$
    $l \leftarrow S_0$
    while l is not empty **do**
        $\Phi_l \leftarrow \emptyset$
        for all $S_i \in l$ **do**
            $g_i \leftarrow (X, \hat{f}, c(S_i))$
            $\phi_{i,1}, \ldots, \phi_{i,\gamma} \leftarrow \mathsf{shap}(g_i)$
            $\Phi_l \leftarrow \Phi_l \cup \phi_{i,1}, \ldots, \phi_{i,\gamma}$
        end for
        $\tau \leftarrow \tau(\Phi_l)$
        $l' \leftarrow \emptyset$
        for all $\phi_i \in \Phi_l$ **do**
            if $\phi_i \geq \tau$ **then**
                if $|S_i| \leq s$ **then**
                    $L \leftarrow L \cup S_i$
                else
                    $l' \leftarrow l' \cup S_i$
                end if
            end if
        end for
        $l \leftarrow l'$
    end while
    return L
end procedure
$L \leftarrow b\mathsf{h\text{-}Shap}(X, \mathcal{T}_0, \hat{f})$

5.3 Comparison with PartitionExplainer

PartitionExplainer and h-Shap are closely related, as they both consider coalitions of features. The former computes Shapley coefficients recursively through a hierarchy of clusters of features, in a fashion inspired by Owen coefficients (Owen, 1977)–an extension of Shapley coefficients for games with

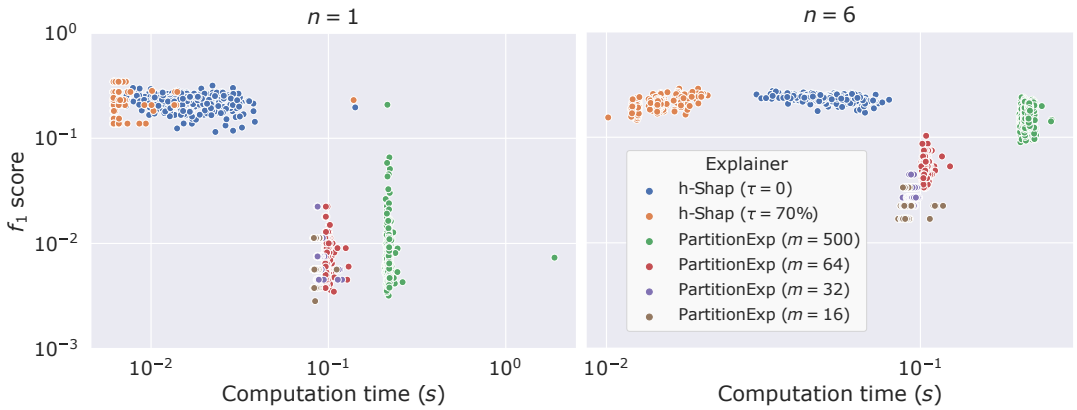


Figure 5.1: Detailed Comparison of PartitionExplainer with h-Shap in the synthetic dataset for $n = 1, 6$ crosses. We use PartitionExplainer with $m = 500, 64, 32, 16$ maximal model evaluations and h-Shap with and absolute relevance tolerance of $\tau = 0$ and a relative one of $\tau = 70\,\%$. f_1 scores are computed on binary masks obtained by thresholding the saliency maps at 1×10^{-6} to account for noisy attributions.

a-priori coalitions of players. The latter explores a quadtree of the input image, where every node corresponds to a game with 4 players, and it only computes the exact Shapley coefficients of relevant games under a certain multiple instance learning assumption (see Assumption A1). PartitionExplainer is partition- and model-agnostic. Here, we use PartitionExplainer with axis-aligned splits, i.e. at every node, the longest axis of a partition is halved in order to generate two children nodes. That is, two iterations of PartitionExplainer produce the same leaves as one iteration of h-Shap.

Both methods reduce the exponential cost of computing Shapley coefficients. PartitionExplainer, when run on a balanced partition tree, requires quadratic runtime with respects to the number of features in the input, while h-Shap only requires a number of model evaluations that is logarithmic in the number of features.

Finally, one can control the number of clusters PartitionExplainer will

explore by limiting the maximal number of model evaluations m. On the other hand, h-Shap provides two parameters, s and τ, which are informed by the problem and control the minimal feature size and relevance tolerance, respectively. Fig. 5.1 showcases a detailed comparison of the two methods in the synthetic case for a different settings of their parameters.

5.4 Experimental details

5.4.1 Synthetic dataset

Table 5.1 represents the network architecture used in the synthetic dataset experiment. We optimize for 50 epochs with Adam optimizer, learning rate 0.001 and cross-entropy loss.

Layer	Filter size	Input size
Conv_1	$6 \times (3 \times 5 \times 5)$	$3 \times 100 \times 120$
ReLU_1	–	$6 \times 96 \times 116$
MaxPool_1	2×2	$6 \times 96 \times 116$
Conv_2	$16 \times (6 \times 4 \times 4)$	$6 \times 48 \times 58$
ReLU_2	–	$16 \times 45 \times 55$
MaxPool_2	5×5	$16 \times 45 \times 55$
FC_1	1584×120	1584×1
ReLU_3	–	120×1
Dropout_1	–	120×1
FC_2	120×84	120×1
ReLU_4	–	84×1
Dropout_2	–	84×1
FC_3	84×2	84×1

Table 5.1: Network architecture for the synthetic dataset experiment

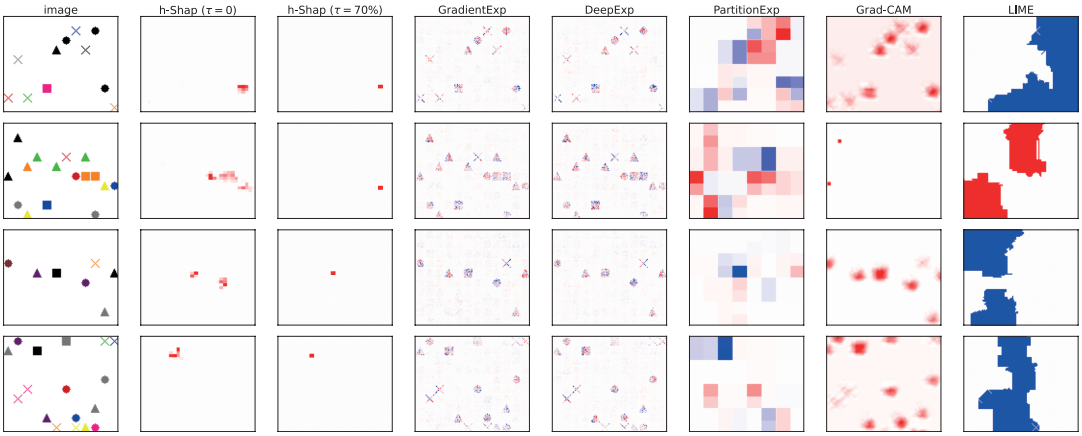


Figure 5.2: Examples of full model randomization tests in the synthetic dataset.

5.4.2 P. vivax (malaria), LISA datasets

In both experiments, we optimize all parameters of a pretrained ResNet18 for 25 epochs with Adam (Kingma and Ba, 2014)–learning rate 0.0001. We use cross-entropy loss and learning rate decay of 0.2 every 10 epochs.

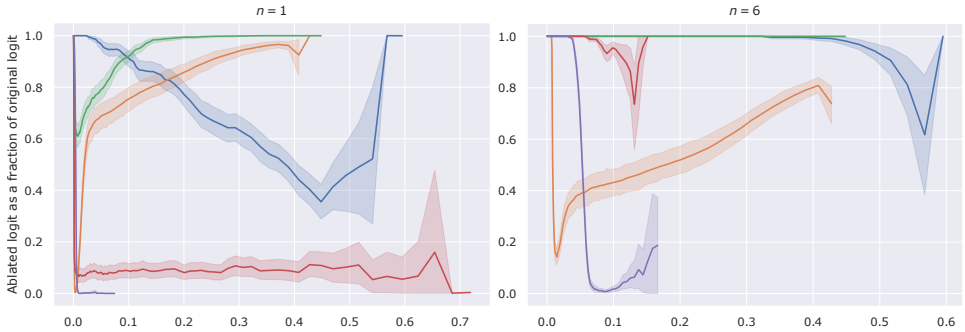
5.5 Sanity checks

Some interpretability methods have been shown (Adebayo et al., 2018) to be unreliable in that they do not truly rely on what the model has learned, i.e. the precise parametrization of $\hat{f}$. For this reason, (Adebayo et al., 2018) advocates for some *sanity checks*. Following this observation, we perform full model randomization tests on all methods compared in this work. The intuition behind model randomization tests is that if the explanation method actually depends on features learned by the model, the explanations should degrade as model weights are randomized. We perform *full* randomization tests in the sense that we randomly initialize *all* the parameters in the simple network

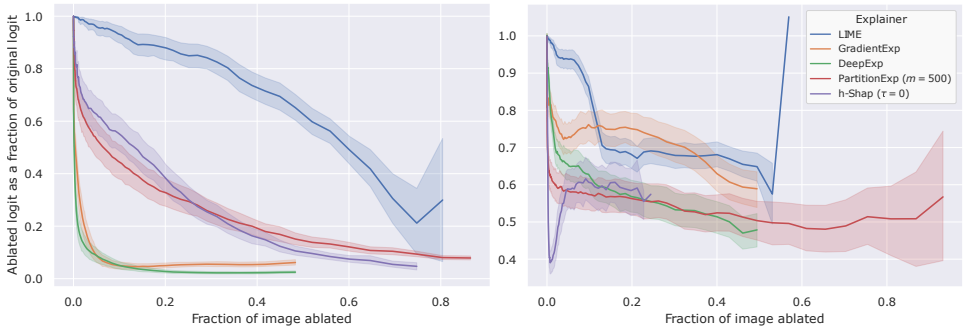
described above in Table 5.1. Fig. 5.2 shows that all explanation methods employed in this work pass the model randomization test, in the sense that the saliency maps degrade completely with a random model.

5.6 Figures

This Appendix contains supplementary figures.



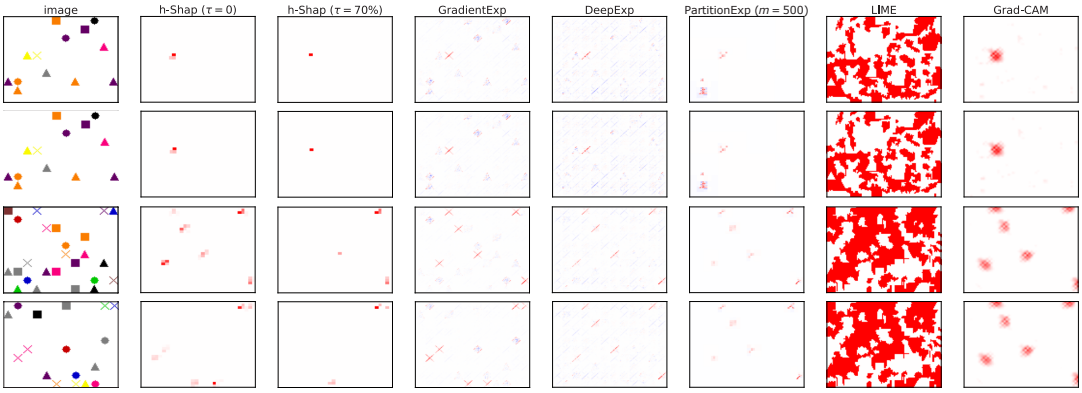
(a) Synthetic dataset. Results for $n = 1, 6$ crosses.



(b) BBBC041 dataset

(c) LISA dataset

Figure 5.3: Logit output compared to original logit output as a function of image ablation.



(a) Synthetic dataset

(b) BBBC041 dataset

(c) LISA dataset

Figure 5.1: More examples of saliency maps.

References

Owen, Guilliermo (1977). "Values of games with a priori unions". In: *Mathematical economics and game theory*. Springer, pp. 76–88.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2018). "Sanity checks for saliency maps". In: *arXiv preprint arXiv:1810.03292*.