Fair and Efficient Alternatives to Shapley-based Attribution Methods

Charles Condevaux, Sébastien Harispe, Stéphane Mussard

Fair and Efficient Alternatives to Shapley-based Attribution Methods

Charles Condevaux[✉][1], Sébastien Harispe[2], and Stéphane Mussard[1]

[1] Univ. Nîmes CHROME, France
{charles.condevaux, stephane.mussard}@unimes.fr
[2] EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, France
sebastien.harispe@mines-ales.fr

Abstract. Interpretability of predictive machine learning models is critical for numerous application contexts that require decisions to be understood by end-users. It can be studied through the lens of local explainability and attribution methods that focus on explaining a specific decision made by a model for a given input, by evaluating the contribution of input features to the results, e.g. probability assigned to a class. Many attribution methods rely on a game-theoretic formulation of the attribution problem based on an approximation of the popular Shapley value, even if the underlying rationale motivating the use of this specific value is today questioned. In this paper we introduce the FESP - Fair-Efficient-Symmetric-Perturbation - attribution method as an alternative approach sharing relevant axiomatic properties with the Shapley value, and the Equal Surplus value (ES) commonly applied in cooperative games. Our results show that FESP and ES produce better attribution maps compared to state-of-the-art approaches in image and text classification settings.

Keywords: Machine learning interpretability · XAI · local interpretability · Attribution method

1 Introduction

Deep learning models are today state-of-the-art to tackle a large variety of machine learning problems in image or natural language processing (NLP) to cite a few. The use of these efficient models is however still limited due to their intrinsic black-box nature, i.e. deciphering the complex input-output mapping performed by trained deep learning models -sometimes involving billions of parameters- is still an open problem [21]. Indeed, many application contexts require not only models with good average performance, but also significant explanations allowing to fully understand and interpret predictor outputs. This is not only true for obvious critical use cases, e.g. in the medical field, in which sensitive decisions have to be supported by evidence [10,21,29]. More generally, legitimate concerns about potential harmful bias of inscrutable models are more expressed. Due to those issues, regulators introduce more and more legal requirements imposing life-impacting automated decision making to be explainable [15,28].

In this context, numerous works analyze approaches contributing to deep learning model explainability, in particular through the notions of global and local interpretability while dealing with predictive tasks [16,4,21]. Global interpretability sheds some light on the general model behavior, e.g. global decision rules, while local interpretability focuses on explaining a specific decision (output) for a given input instance. This paper is concerned with local interpretability, and in particular with *attribution methods* (AMs). These methods aim at explaining the prediction made by a predictor for an input instance by assigning a scalar *attribution value* to each input feature. The purpose is therefore, for a given instance, to distinguish the features best explaining a model output prediction. The core problem is thus to define an AM that assigns a relevant attribution value to each feature in a specific predictive context.

Assigning attribution values in a meaningful way is an open question that has been studied through different angles. A large body of works focuses on AMs backed on axiomatic motivations defining supposedly intuitive properties that these methods should respect [36,38,24,22,2]. A wide range of contributions are in particular considering the Shapley value [32] as ground truth since it defines the unique way to solve the game-theoretic formulation of the attribution problem considering admitted axioms in coalition games - attribution is made considering a cooperative game between model features; attribution among the features is then made based on the Shapley value [2,38,22]. In that context, several approaches have been proposed to approximate the prohibitive computation of the Shapley value which requires evaluating 2^N feature subsets considering inputs of N features (NP-hard). Nevertheless, even if contributions stress that AMs based on the Shapley value seem to agree with human intuitive expectations [2], no clear agreement on that matter has been reached and the ground truth status of the Shapley value is today questioned [20,12]. Axiomatically grounded and algorithmically efficient, AMs have still to be investigated.
The contributions proposed in this paper are threefold:

1. We introduce FESP (Fair-Efficient-Symmetric-Perturbation value), an axiomatically grounded and algorithmically efficient AM that shares some properties with the Shapley value.

2. We propose the use of the equal surplus (ES) value, an $O(N)$ AM employed in cooperative games, which is linear, efficient and symmetric.

3. We show that FESP and ES achieve good accuracy on image and text classification compared with usual AMs.[3][4] The results outline their benefits with different benchmarks, e.g. issued from SHAP [22] and gradients [33,38].

The paper is structured as follows: Section 2 introduces existing AMs and discussions about the Shapley value. Section 3 presents FESP and ES. Section 4 evaluates ES, FESP and existing AMs on image and text classification tasks, with discussions on performances with respect to different protocols. Section 5 discusses our findings before mentioning perspectives they open.

[3] Our experiments: https://github.com/ccdv-ai/fesp_es.git. This work used HPC resources of IDRIS (allocation 2022-AD011011309R2) made by GENCI.

[4] This work has benefited from LAWBOT (ANR-20-CE38-0013) grant.

2 State of the art

In this section we present the attribution problem focusing on a multiclass classification setting, as well as state-of-the art AMs proposed to solve it.

2.1 The attribution problem

Considering a predictor, the attribution problem consists in attributing a scalar value to each input feature characterizing an instance with respect to (w.r.t.) a predicted value (e.g. class probability in a classification setting or real value in a regression setting). This value represents the contribution of a specific feature to the prediction, e.g. in a classification setting, this value may be useful to understand which input features support a given class.

Without loss of generality, a multiclass classification setting is considered with a set of classes $\mathcal{C} := [\![1, C]\!]$, with $[\![a, b]\!]$ denoting the interval of all integers between a and b included. In that context, a predictor f takes an N-dimensional feature input $\mathbf{x} := [x_1, \ldots, x_N] \in \mathbb{R}^N$ and produces a probability distribution $f(\mathbf{x}) := [f_1(\mathbf{x}), \ldots, f_C(\mathbf{x})] \in [0, 1]^C$, with $f_i(\mathbf{x})$ the probability assigned to class $i \in \mathcal{C}$ by f for $\mathbf{x}$. $\mathcal{N} := [\![1, N]\!]$ is the set of feature indices.

Considering this setting, given predictor f and an input $\mathbf{x} \in \mathbb{R}^N$, an AM φ aims at computing a contribution vector $\varphi(\mathbf{x}, f_i)$ for any class $i \in \mathcal{C}$ such as $\varphi(\mathbf{x}, f_i) = [\varphi_1(\mathbf{x}, f_i), \ldots, \varphi_N(\mathbf{x}, f_i)] \in \mathbb{R}^N$, with $\varphi_j(\mathbf{x}, f_i)$ the attribution value of feature $j \in \mathcal{N}$ w.r.t. $f_i(\mathbf{x})$. Otherwise stated, considering the AM φ, $\varphi_j(\mathbf{x}, f_i)$ is the contribution of feature j to the probability assigned by the predictor f to class i for the input $\mathbf{x}$.

The two main classes of approaches studied in the literature to solve the attribution problem are introduced hereafter. They are both based on the evaluation of a perturbation of the input features on the predictive value under study.

2.2 Attribution using feature coalisation analysis

In the local interpretability setting, numerous perturbation-based approaches define an AM φ by evaluating the contribution $\varphi_j(\mathbf{x}, f_i)$ of a specific feature $j \in \mathcal{N}$ (to $f_i(\mathbf{x})$) as its contribution to coalitions of features. Considering a coalition including all features except j (i.e. $\mathcal{N} \setminus \{j\}$), the contribution of j to that coalition is assessed by evaluating the impact of a perturbation of x_j on $f_i(\mathbf{x})$. Such a perturbation aims at mimicking the removal of the studied feature, e.g. by naively setting its value to zero or a baseline value.

For any $\mathcal{S} \subseteq \mathcal{N}$, $\mathbf{x}(\mathcal{S})$ refers to the vector $\mathbf{x}$ in which all feature values x_k, $k \in \mathcal{N} \setminus \mathcal{S}$ have been substituted by a baseline value. As the input $\mathbf{x}$ is implicitly fixed in our discussions, $f_i(\mathcal{S})$ is used to denote $f_i(\mathbf{x}(\mathcal{S}))$, which is the probability assigned by f to class $i \in \mathcal{C}$ w.r.t. $\mathbf{x}(\mathcal{S})$.

The *marginal contribution* of a feature $j \in \mathcal{N}$ to a coalition $\mathcal{S}$ ($j \notin \mathcal{S}$) is thus defined by $f_i(\mathcal{S} \cup \{j\}) - f_i(\mathcal{S})$. Numerous AMs based on this notion of marginal contribution have been studied [13,6,41,39]. Game theory allows us to obtain such contributions, through the (least) core [18] but also through the Shapley

value often considered as the ground truth value to explain the role of a given variable [2].

Attribution value as the Shapley value: The Shapley value averages marginal contributions over all possible feature coalitions:

$$\varphi_j^{Sh}(\mathbf{x}, f_i) := \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \{j\}} P(\mathcal{S}) \Big(f_i(\mathcal{S} \cup \{j\}) - f_i(\mathcal{S}) \Big),$$

for all $j \in \mathcal{N}$; $f_i(\emptyset) := 0$ for all $i \in \mathcal{C}$ by convention, and $P(\mathcal{S}) := (N - S - 1)!S!/N!$ ($S := |\mathcal{S}|$).

The Shapley value implies (and is implied by) four axioms: *efficiency, additivity, symmetry* and the *null player axiom*, see [32].[5] These axioms make the Shapley value appealing from a theoretical point of view, and have motivated the *de facto* ground truth status given to this value.

However, considering N features, 2^N coalitions have to be evaluated which makes the Shapley value prohibitively expensive to compute. A natural way to reduce computation complexity is to rely either on coalition sampling to compute the marginal contributions [8], on local coalitions [9] or on Boolean circuits [3]. The first approach can however be slow to converge when the number of features is large. Instead of directly modifying original inputs, DASP [2] relies on distribution propagation using an auxiliary network based on Lightweight Probabilistic Deep Networks [14]. This model sequentially produces an estimate for each coalition size, thus allowing to greatly reduce the complexity from $O(2^N)$ to $O(N^2)$. Although this approximation is accurate, building an additional network is cumbersome, especially when fine tuning a pretrained model (as it requires rewriting each layer and activation function).

Attribution based on Occlusion: In order to determine whether a feature or a group of features impacts a prediction, occlusion models measure the effect of removing them from the input (marginal contribution). In computer vision, these feature coalitions generally take the form of a sliding block [42], of a predefined size, inside which pixels are disturbed or replaced by a specific value (e.g. 0). Although such perturbation and occlusion models can accurately measure the marginal contribution of a variable, they tend to be slower than other AMs since they require multiple forward passes to fully cover the input and are thus dependent on the number of features. The size of the block is also an additional hyperparameter which can have a significant impact on overall performances.

2.3 Attribution based on gradient analysis

Gradient-based approaches rely on various gradient computations through backpropagation evaluations. They compute the attribution value of a feature evaluating the partial derivative of the studied predicted value with regard to the

[5] It is noteworthy that *additivity* implies *linearity* but the converse does not hold. Invoking *linearity* enlarges the class of admissible AMs, see Theorem 1 below.

feature value, e.g., $\varphi_j(\mathbf{x}, f_i)$ is defined as a function of $\partial f_i(\mathbf{x})/\partial x_j$. In this context $\varphi_j(\mathbf{x}, f_i)$ is then evaluated based on the impacts on $f_i(\mathbf{x})$ induced by a local change of x_j. The function φ_j should be carefully chosen to respect some properties or specific behaviors. For instance, multiplying the gradient by the input [34] increases the sharpness of the attribution map but fails to handle specific functions like ReLU, which can produce zero values. More sophisticated models like DeepLift [33] and Integrated Gradient [38] satisfy a desirable axiom called completeness which is closely related to the efficiency axiom in cooperative game theory: for a baseline $\mathbf{x}'$ we have $\sum_{j\in\mathcal{N}} \varphi_j(\mathbf{x}, f) = f(\mathbf{x}) - f(\mathbf{x}')$.

To compute the contribution map, DeepLift takes all neurons and compares their activations after feeding a true sample and a reference input which can depend on the task and on the dataset. This model is inspired by Layer-wise Relevance Propagation which relies on a similar idea without the use of a reference [5]. Integrated Gradient averages different gradients: the input is modified multiple times along a linear path between itself and a baseline often set to zero. This continuous setting has been connected to another branch of the literature based on coalisation analysis, such as the Aumann-Shapley value [37].

3 Fair-Efficient-Symmetric Perturbations-based AMs

3.1 The Equal Surplus Value

It is well established that the Shapley value is easy to interpret since it displays the average of all marginal contributions of each feature; in this respect, it is a marginalist value. It shares some common properties with other marginalist values which form the Linear-Efficient-Symmetric values family (LES values) [31]. To our knowledge, this family has not been studied in the context of the attribution problem. The axioms respected by LES values are introduced hereafter.

Axiom 1 Linearity: *For all predictors f, g, an AM φ satisfies linearity if, $\varphi(\mathbf{x}, \alpha_1 f_i + \alpha_2 g_i) = \alpha_1\varphi(\mathbf{x}, f_i) + \alpha_2\varphi(\mathbf{x}, g_i)$, for all $\alpha_1, \alpha_2 \in \mathbb{R}$ and for all classes $i \in \mathcal{C}$.*

Axiom 2 Efficiency: *For all predictors f, an AM φ satisfies efficiency if, $\sum_{j\in\mathcal{N}} \varphi_j(\mathbf{x}, f_i) = f_i(\mathcal{N})$, for all classes $i \in \mathcal{C}$.*

Axiom 3 Symmetry: *For all predictors f, an AM φ satisfies symmetry if, for all features $j \in \mathcal{N}$, $\varphi_j(\mathbf{x}, f_i) = \varphi_{\pi(j)}(\mathbf{x}_\pi, f_i)$ for all permutations π over the set of $N!$ permutations on $\mathcal{N}$ and for all classes $i \in \mathcal{C}$.*

LES values have been extensively characterized outside the machine learning literature first by [31], then by [17,25,27] through the following theorem:

Theorem 1. *For all predictors f and all classes $i \in \mathcal{C}$, an AM φ satisfies linearity, efficiency and symmetry if and only if there exists a unique sequence of $N-1$ real numbers $\{b_s\}_{s=1}^{N-1}$ such that for each $j \in \mathcal{N}$ with $b_0 = 0$ and $b_N = 1$:*

$$\varphi_j(\mathbf{x}, f_i) = \sum_{\mathcal{S}\subseteq\mathcal{N}\setminus\{j\}} P(\mathcal{S})\Big(b_{s+1} f_i(\mathcal{S}\cup\{j\}) - b_s f_i(\mathcal{S})\Big).$$

LES values are all based on marginal contributions, therefore they provide feature contributions and interpretations very close to the usual Shapley value. The Shapley value φ^{Sh} is indeed a particular case of the LES family considering all marginal contributions equally weighted ($b_s = 1$ for all $s = 1, \ldots, N-1$). Other well-known LES values, studied in the cooperative game literature are: the Equal Surplus value (ES) [11], the Solidarity value [26], the Prenucleolus [30], and the Consensus value [19]. The ES value φ_j^{ES} ($b_s = 0$ if $1 < s < N$, $b_s = 1$ if $s = N$, $b_s = N-1$ if $s = 1$) is a peculiar member of the LES family since it is of complexity $O(N)$ whereas the others are $O(2^N)$:

$$\varphi_j^{ES}(\mathbf{x}, f_i) = f_i(\{j\}) + \frac{f_i(\mathcal{N}) - \sum_{k=1}^{N} f_i(\{k\})}{N}. \tag{1}$$

The first term of the right-hand side of Equation (1) is the contribution of feature x_j alone: its individual marginal contribution compared to a model composed of all features with baseline values $f_i(\{j\}) - f_i(\emptyset)$. The second term is the equal surplus: $f_i(\mathcal{N}) - \sum_{j=1}^{N} f_i(\{j\})$, i.e. the additional gain produced by the grand coalition in excess of the sum of the individual marginal contributions of features x_j, which evolve independently of the others.[6]

3.2 FESP

An AM grounded on the individual marginal contributions of each feature $f_i(\{j\}) - f_i(\emptyset)$ as in the ES is welcome since it outlines the role of each feature independently of the others. However, the equal surplus term is a constant for all features, consequently it cannot display the interaction of each feature with the grand coalition. In order to capture this specific effect, the exclusion of one feature from the whole set of features is employed, which consists in the occlusion technique. Occlusion related to feature x_j over class i may be simply characterized by $f_i(\mathcal{N} \setminus \{j\})$ instead of the equal surplus $f_i(\mathcal{N}) - \sum_{j=1}^{N} f_i(\{j\})$.

Then, two extreme feature coalitions could be considered for an AM: the one with the feature itself (such as Fig. 1 on the right-hand side - considering features as superpixels), and the one associated with occlusion, i.e. the entire image minus a given feature (center of Fig. 1). On this basis, we propose the following family of AMs based on extreme feature coalitions:

Definition 1. Family of AMs based on extreme feature coalitions:

$$\varphi_j(\mathbf{x}, f_i) = w_i f_i(\{j\}) + (1 - w_i)(-f_i(\mathcal{N} \setminus \{j\})), \tag{2}$$

with $w_i \in [0, 1]$ *a weight associated with class* $i \in \mathcal{C}$.

The first component of the family, $w_i f_i(\{j\})$, is grounded on the individual marginal feature contribution, which is always positive. Then, as far as the feature is discriminant, its contribution to the classification in class i increases. The

[6] The study of the independence is of importance for the tractability of the Shapley value, this is the case with fully factorized data distributions [7].

Fig. 1. Extreme feature coalitions

second component, $(1-w_i)(-f_i(\mathcal{N}\setminus\{j\}))$, is the contribution of occlusion, it is always negative. Occlusion of a discriminant feature x_j for class i entails that the probability f_i collapses, implying that the second component tends to zero. If an AM does not lie in the family of extreme feature coalitions, anything guaranties that bad features would be penalized by occlusion. Indeed, whenever a feature x_j is not discriminant for the classification in class i, the second component becomes negative, and the attribution value $\varphi(\mathbf{x}, f_i)$ can also become negative so that feature x_j is considered non-explanatory for the task. Furthermore, in order to gauge whether a feature is more *relevant* than another, the *fair treatment* axiom must be respected.

A feature x_k is said to be more relevant compared to feature x_ℓ when the association of x_k with all feature coalitions $\mathcal{S}\setminus\{k,\ell\}$ provides a greater attribution value compared to that of x_ℓ [27]. This property is welcome for all classification tasks such as image and text classifications. For instance, in an image classification setting, if a pixel x_k is more relevant compared to another one, because it allows some important shapes to be outlined, then the AM provides a higher contribution for x_k.

Axiom 4 Fair treatment: *For all models f, and two given features x_k, x_ℓ, an AM φ satisfies fair treatment if, whenever feature x_k is more relevant compared to feature x_ℓ, i.e. $f_i(\mathcal{S}\cup\{k\}) \geq f_i(\mathcal{S}\cup\{\ell\})$ for all $\mathcal{S}\setminus\{k,\ell\}$, then $\varphi_k(\mathbf{x}, f_i) \geq \varphi_\ell(\mathbf{x}, f_i)$, for any given class $i \in \mathcal{C}$.*

FESP is an $O(N)$ complexity AM that shares a common structure with members of the LES family: it respects *efficiency*, *symmetry* and *fair treatment* (see Appendix A and B).

Proposition 1. *If an AM φ lies in the family of AMs based on extreme feature coalitions, and if it satisfies efficiency, then it is the FESP (Fair-Efficient-Symmetric-Perturbation) value given by, for $j \in \mathcal{N}$ and for $i \in \mathcal{C}$,*

$$\varphi_j^{FESP}(\mathbf{x}, f_i) = w_i f_i(\{j\}) + (1-w_i)(-f_i(\mathcal{N}\setminus\{j\})), \tag{3}$$

$$w_i = \frac{f_i(\mathcal{N}) + \sum_{k=1}^{N} f_i(\mathcal{N}\setminus\{k\})}{\sum_{k=1}^{N} f_i(\{k\}) + \sum_{k=1}^{N} f_i(\mathcal{N}\setminus\{k\})}. \tag{4}$$

4 Experiments

This section presents results and evaluation protocols defined for comparing AMs on image and text classification tasks.We report experiments running ES and FESP along-side Integrated Gradients [38], DeepLIFT rescale [33], GradientShap [22] and Occlusion model. We also compare to the SHAP library using the DeepExplainer model [22] for vision tasks and the NLP pipeline named ShapExplainer for language tasks.

Local explainability. A model is first trained to solve the predictive task under consideration. In order to focus on AM evaluation and to avoid any interpretative bias, we consider *simple* predictive tasks for which good performances are today easily achieved. Based on the predictor obtained, an AM is then evaluated regarding the features it brings out as important to explain the prediction obtained for a given input (only predictions are performed, no training phase is involved while evaluating AMs).

Top-k model accuracy. This metric consists in evaluating how the predictor accuracy evolves only using top-k input-dependent contributing features according to an AM φ. If φ identifies the features that best explain an input classification, the predictor should keep achieving good performances only considering those features, i.e., the more φ performs correctly, the better should be the predictor accuracy only considering a subset of features provided by φ. Unselected features for a given input are simply masked during prediction; the shape of the predictor input is not modified. For a given task, the same predictor is therefore employed independently of the features considered during prediction.

4.1 Image classification: protocols and results

A pretrained VGG16 model [35] is fine tuned on a binary classification task related to Oxford-IIIT Pet Dataset[7] (`dog` vs `cat`, fine-tuning: 3 epochs over 6325 images). It achieves 99% accuracy (1024 images, features are pixels).

Masking strategy. An image segmentation dataset gives for an image, the pixels of the shape of interest (segmentation mask), as well as its label, e.g. for an image labeled `dog`, the pixels of the dog are known. As focus is put on a simple classification task, it is assumed that pixels inside the mask should be relevant and have high attribution values (Appendix C presents a similar experiment modifying the pixels outside the segmentation mask with a random value). Considering an AM φ and a given image $\mathbf{x}$, the top-k contributing pixels of $\mathbf{x}$ are computed w.r.t. φ, with k a fixed number of pixels set based on the size of the segmentation mask of $\mathbf{x}$. AM-*Precision*@k is computed: AM-*Precision*@20$(\varphi, \mathbf{x}, i)$ is the precision of φ on $\mathbf{x}$ over class i, only considering the top-k pixels, k being here equal to 20% of the size of the segmentation mask for $\mathbf{x}$ on class i. The precision of an AM is set to the average of the precision obtained for each image.

[7] https://www.robots.ox.ac.uk/$\sim$vgg/data/pets/

Averaging ES and FESP. Despite their $O(N)$ complexity, computing ES and FESP is slower than gradient based attribution methods since a forward pass is required for each feature. For a 224×224 RGB image, 50,176 passes would be in theory necessary to compute attribution values. In practice, removing or inferring a class by modifying a single pixel has little to no impact on the prediction of the VGG16. 56×56 superpixels are considered for ES, FESP and Occlusion, so that the image becomes a grid of 16 superpixels. These methods are run on the superpixels, and the process is repeated by moving the grid with a stride of 8. All pixels inside a given superpixel get the same attribution score φ_j for the current pass; these scores are then averaged resulting in an overlapping process (each pixel gets masked the same number of times in order to get a balanced average).[8] This approach is similar to Occlusion and DeepExplain implementation.[9]

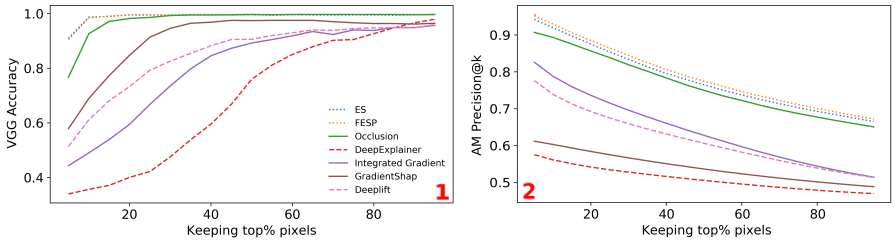

Fig. 2. Effect of feature selection on the predictor accuracy/precision.

Plot 1 (Figure 2) shows the accuracy of the pretrained model (VGG16) while only considering the top-k features (pixels) evaluated as important by each AM. Considering the top 10% of the features, ES and FESP provide good predictive performances (90% accuracy). Except Occlusion, the other AMs must consider almost 90% of the selected features in order to reach the full input predictor accuracy (99%). This accuracy is reached only using 15% to 20% of the features identified as important by FESP and ES.

Plot 2 (Figure 2) presents the AM-$Precision@k$, i.e. the capacity of each AM to outline expected informative pixels (pixels of the segmentation mask). AMs generally tend to consider the most important pixels to be inside the segmentation mask at first. FESP, ES and Occlusion achieves very good performances compared to other methods according to that test.

Figure 3 shows which image parts are recognized as relevant by the different AMs to explain the network prediction (top-10%).

We observe very different behaviors. AMs based on backpropagation independently treat pixels and therefore may return a noisy representation that is

[8] Good tradeoff between performance and time complexity since large superpixels lead to higher performances while small ones tend to be noisier.

[9] https://github.com/slundberg/shap

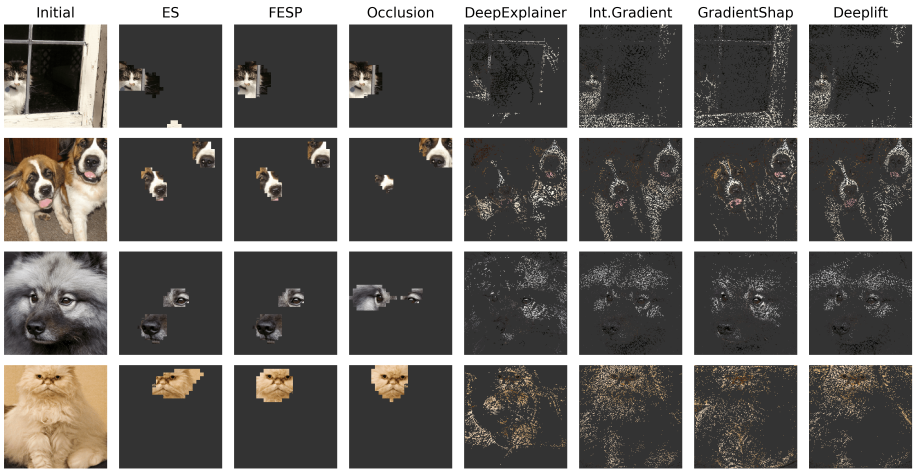

Fig. 3. Top 10% highest contributing pixels.

difficult to understand even for a relatively simple task where the discriminating criteria are fairly high level. The other AMs choose very localized areas, ignoring the rest of the image since they take benefit from convolutional layers that rely on local information. FESP and ES have a similar behavior but we observe in practice a noisier selection using ES, sometimes resulting in small artifacts. This behavior partly explains the performances obtained in the prediction task on partially masked inputs. Indeed FESP and ES tend to quickly identify very discriminant groups of features enabling to achieve good predictive performances even with a very limited set of features (Figure 2). Thus, normalizing and merging the best performing AMs can be a good solution to improve the overall selection as shown in Appendix D.

Robustness of ES and FESP. An additional evaluation protocol is conducted based on recent contributions on AM evaluations [40] (refining [1]). It relies on a binary classification setting involving fictive composite images, each one being composed of 2×2 images from the Oxford-IIIT Pet Dataset. Each composite image is labeled `cat` or `dog` and only contains a single image among four corresponding to its label. Considering a specific composite image and a good predictive model, we assume that an efficient AM should make it easy to distinguish the single image corresponding to its label. Each composite image is a random mix of: (i) a labeled image (`cat` or `dog`), (ii) 3 unrelated additional images, (iii) the locations of the 4 images in the 2×2 grid. The train and test sets are generated using the same approach compared to disjoint subsets of images. The same pretrained VGG16 architecture is fine-tuned on 6325 images over 4 epochs (96% accuracy on the test set).

Figure 4 shows that ES and FESP display the highest top-k model accuracy, indeed an accuracy greater than 90% is reached with 3% of top pixels (plot 1).

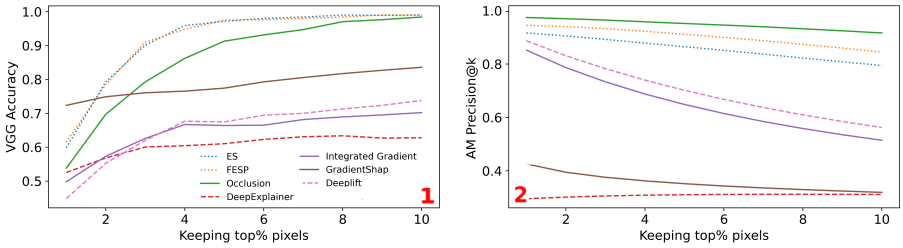

Fig. 4. Effect of feature selection on accuracy and precision.

The AM-*Precision@k* (plot 2) gives the percentage of pixels located within the labeled subimage (among the four) given the selection of the top-k contributing pixels. According to this metric, FESP is between Occlusion and ES with 95% for only 2% of top pixels.

As shown in Figure 5 with 2×2 images, Occlusion, FESP and ES bring out relevant areas of the classes `dog` and `cat` and tend to be less noisy than gradient-based techniques (more images are available in Appendix D).

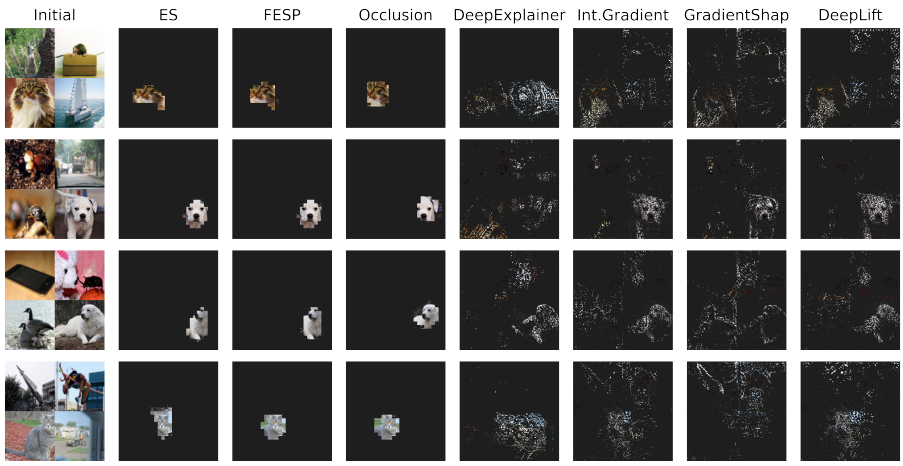

Fig. 5. Top 5% highest contributing pixels.

4.2 Text classification: protocols and results

A binary text classification task is performed using IMDB dataset [23]. The model is a pretrained RoBERTa fine tuned on IMDB dataset, for which features

are words (95.5% accuracy).[10] Testing is made on a subset of 1024 samples of the official testing set.

Masking strategy. The masking strategy is task dependent. Transformers can take benefit from the softmax function inside the self-attention mechanism to fully mask a token and avoid all connections. This is not possible with convolution layers used in vision tasks.

Averaging ES and FESP. A block of size 1 with a stride of 1 is used, consequently ES and FESP are directly estimated without an averaging strategy.

Compared to the image classification task, the same accuracy performances are obtained (Figure 6). On the one hand, the top 5% of words yield 95% accuracy for ES and FESP, and 90% for Occlusion.

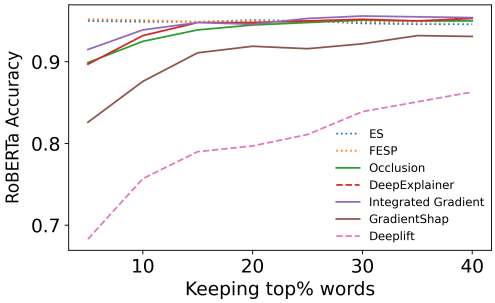

Fig. 6. Effect of feature selection on the predictor accuracy.

Finally, Figure 7 depicts words selected by the seven AMs over one example of the IMDB testing set. Additional examples are provided in Appendix D. For words chunked into several subwords by the tokenizer, the maximum score is used. AMs are normalized in such a way that each feature contribution takes value between 0 and 1, with 1 the highest contributing tokens being colored red. As expected, all AMs easily capture positive words such as "love" and "sexy", but these are not necessarily associated to the highest contribution. For instance Occlusion assigns the most important contribution to "this" and "I".

4.3 Discussions

Occlusion, ES and FESP. FESP and ES behave similarly most of the time although ES being slightly noisier since each feature evolves independently (of the grand coalition). In order to remedy this problem, FESP straddles the line between ES and Occlusion since it can be considered as a weighted mean of the two methods. In terms of interpretability, these models differ greatly. Occlusion is unable to determine the sign of the feature contributions unlike FESP & ES. This difference makes interpretability difficult in many cases, especially for word

[10] https://huggingface.co/textattack/roberta-base-imdb

Fig. 7. Word importance normalized scores.

importance tasks as can be shown in Figure 7. In the case of image classifications a feature with a high contribution does not mean that it contributes positively to the prediction in the case of Occlusion.

From local to global explainability. Although we focused on local explainability, ES and FESP can also be used in a global explainability context. This can be achieved by using specific metrics (e.g accuracy, coefficient of determination) to measure the average impact of a feature on the predictions of a given predictor (see Appendix E for examples).

5 Conclusion

We have presented Equal Surplus (ES) and FESP (Fair-Efficient-Symmetric-Perturbation), two AMs based on marginalist values that can be used for local explainability of deep supervised learning models. These AMs compute attribution values that share relevant axiomatic properties with the Shapley value while ensuring an $O(N)$ time complexity for N-dimensional inputs.

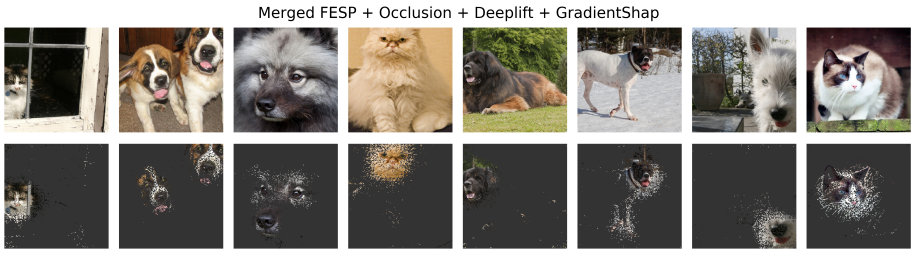

Fig. 8. Merged top 30% highest contributing pixels relative to mask size.

According to the proposed evaluations based on two image and text classification tasks, FESP, ES and Occlusion seem to be more suited for tasks with spatial or temporal dependencies such as computer vision and NLP. Indeed, in these contexts, backpropagation and gradient-based approaches tend to be noisy and generally more difficult to interpret for humans. Additionally, our results also corroborate literature findings highlighting that backpropagation gradient-based approaches tend to act like shape detectors, and therefore achieve good results in distinguishing the global shape of an object of interest in a local attribution setting [1]. This paves the way to the study of various AMs mixing both approaches highlighting different but often complementary features (see Figure 8).

Finally, the quantitative and qualitative results achieved by ES and FESP motivate the study of fast and axiomatically grounded AMs derived from LES values, which could reveal more, for example with the employ of AMs issued from the least square prenucleolus [30,31], or in the global attribution setting.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292 (2018)
2. Ancona, M., Öztireli, C., Gross, M.: Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In: International Conference on Machine Learning. pp. 272–281. PMLR (2019)
3. Arenas, M., Barceló, P., Bertossi, L.E., Monet, M.: The tractability of shap-score-based explanations for classification over deterministic and decomposable boolean circuits. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 6670–6678. AAAI Press (2021)
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020)
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wiserelevance propagation. PloS one **10** (2015)

6. Brink, R., Funaki, Y., Ju, Y.: Reconciling marginalism with egalitarianism: consistency, monotonicity, and implementation of egalitarian shapley values. Social Choice and Welfare **40**, 693–714 (2013)
7. den Broeck, G.V., Lykov, A., Schleich, M., Suciu, D.: On the tractability of shap explanations. Proceedings of the AAAI Conference on Artificial Intelligence **35**(07), 6505–6513 (2021)
8. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the shapley value based on sampling. Computers & Operations Research **36**(5), 1726–1730 (2009)
9. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: L-shapley and c-shapley: Efficient model interpretation for structured data. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=S1E3Ko09F7
10. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al.: Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface **15**(141), 20170387 (2018)
11. Driessen, T.S.H., Funaki, Y.: Coincidence of and collinearity between game theoretic solutions. Operations-Research-Spektrum **13(1)**, 15–30 (1991)
12. Frye, C., Rowat, C., Feige, I.: Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. Advances in Neural Information Processing Systems **33** (2020)
13. Funaki, Y., Hoede, K., Aarts, H.: A marginalistic value for monotonic set games. International Journal of Game Theory **26**, 97–111 (1997)
14. Gast, J., Roth, S.: Lightweight probabilistic deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
15. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI magazine **38**(3), 50–57 (2017)
16. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
17. Hernández-Lamoneda, L., Juárez, R., Sánchez-Sánchez, F.: Dissection of solutions in cooperative game theory using representation techniques. International Journal of Game Theory **35**, 395–426 (2007)
18. Heskes, T., Sijben, E., Bucur, I.G., Claassen, T.: Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 4778–4789. Curran Associates, Inc. (2020)
19. Ju, Y., Borm, P., Ruys, P.: The consensus value: a new solution concept for cooperative games. Social Choice and Welfare **28**, 685–703 (2007)
20. Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.: Problems with shapley-value-based explanations as feature importance measures. In: International Conference on Machine Learning. pp. 5491–5500. PMLR (2020)
21. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2021)
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
23. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/P11-1015
24. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
25. Nembua, C.C., Andjiga, N.G.: Linear, efficient and symmetric values for TU-games. Economics Bulletin **3**, 1–10 (2008)
26. Nowak, A.S., Radzik, T.: A solidarity value for n-person transferable utility games. International Journal of Game Theory **23**, 43–48 (1994)
27. Radzik, T., Driessen, T.: On a family of values for tu-games generalizing the shapley value. Mathematical Social Sciences **65**, 105–111 (2013)
28. Ras, G., van Gerven, M., Haselager, P.: Explanation methods in deep learning: Users, values, concerns and challenges. In: Explainable and Interpretable Models in Computer Vision and Machine Learning, pp. 19–36. Springer (2018)
29. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
30. Ruiz, L.M., Valenciano, F., Zarzuelo, J.M.: The least square prenucleolus and the least square nucleolus. two values for tu games based on the excess vector. International Journal of Game Theory **25**, 113–34 (1996)
31. Ruiz, L.M., Valenciano, F., Zarzuelo, J.M.: The family of least square values for transferable utility games. Games and Economic Behavior **24**, 109–130 (1998)
32. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)
33. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. PMLR (06–11 Aug 2017)
34. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. CoRR **abs/1605.01713** (2016)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
36. Sun, Y., Sundararajan, M.: Axiomatic attribution for multilinear functions. In: Proceedings of the 12th ACM conference on Electronic commerce. pp. 177–178 (2011)
37. Sundararajan, M., Najmi, A.: The many shapley values for model explanation. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 9269–9278. PMLR (13–18 Jul 2020)
38. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning. pp. 3319–3328. PMLR (2017)
39. Wang, J., Zhang, Y., Kim, T.K., Gu, Y.: Shapley q-value: A local reward approach to solve global reward games. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 7285–7292 (Apr 2020)
40. Yona, G., Greenfeld, D.: Revisiting sanity checks for saliency maps (2021). https://doi.org/10.48550/ARXIV.2110.14297, https://arxiv.org/abs/2110.14297
41. Young, P.: Monotonic solutions of cooperative games. International Journal of Game Theory **29**, 65–72 (1985)
42. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. CoRR **abs/1311.2901** (2013)