Discriminative Model-Agnostic Vision Explainer for Differentiable Classifiers

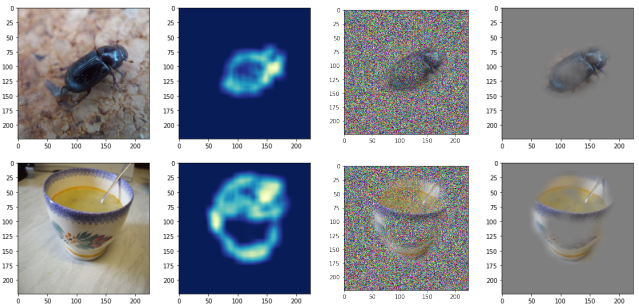
Xiaozhe Gu
FNii@ Chinese University of Hong Kong, Shenzhen
guxi0002@e.ntu.edu.sg

ABSTRACT

Machine learning models are increasingly adopted in various real-world applications (e.g., medical diagnosis or autonomous driving). However, most of they remain mostly black boxes and their decision-making process is largely unclear. To verify and validate these models, it is critical to gain insight into how their predictions are arrived.

In this work, we propose a model-agnostic vision explanation method for differentiable classifiers in case the user can only send queries and read off outputs. In order to address the shortcomings of existing perturbation based approaches, we introduce a number of technical innovations to defend against adversarial evidence by encouraging the generated explanations to have a regular structure and cover objects of interest tightly are sharply. The resulting explanations can capture the most important features that contribute to the model predictions with little background information. Finally, we also evaluate the proposed method qualitatively and quantitatively.

ACM Reference Format:
Xiaozhe Gu. 2020. Discriminative Model-Agnostic Vision Explainer for Differentiable Classifiers . In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Image (b) Mask (c) Original Explanation (d) Explanations with gray background

Figure 1: Fine-grained explanations for input images with softmax score of the most-likely class. Our method tries to find a fine-grained explanation that is responsible for a model's output. For better visulization of the explanation, in the rest of the paper, we will use gray color to represent masked regions, while the model actually sees random noise sampled from a uniform distribution $(0, 1)$ unless otherwise specified.

1 INTRODUCTION

Black box predictors such as deep neural networks (DNNs) have become the state-of-the-art techniques for a variety of computer vision benchmarks such as Imagenet [8], Caltech [9] and Cityscapes [6]. In spite of their excellent prediction accuracy, they remain mostly black boxes since it is unclear how a particular a decision/prediction is made. The black box reputation of DNNs has become a barrier to adoption especially in safety-critical applications (e.g., medical diagnosis, autonomous driving etc). Without the explanations to get insight into the decision making process, it would become difficult for uses to build up trust and credibility in the models.

In this paper, we focus on the problem of providing explanations for classification decisions made by black box differentiable classifiers on natural images. We consider the case in which the users can only send queries and read off outputs, which implies the explanation should not rely on the inner workings of the models. In fact, black-box access is a common deployment mode for many public and commercial models. The internal details, such as model architecture or training data, can be proprietary to be protected. Besides, an complete access to these models also increases their risk to attacks like adversarial examples.

According to whether internals of the model (e.g., intermediate feature maps, or the network's weights) are required to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

produce an explanation, we can classify the existing explanation methods into *model-agnostic* techniques and *model-specific* techniques. For example, activation based explanation methods [5, 23, 37] and FGVis [31] are model-specific because they utilize the inner states of the model. On the other hand, LIME [22] and RISE [19] are completely model-agnostic method because the only information they require to produce an explanation is the output $f(\mathbf{x})$ of the model. A special case are perturbation based explanation methods (e.g., [3, 7, 12]) that do not need the internals of the model, but require access to the gradients of model output with respect to the input image. Thus, here we classify them as *partially* model-agnostic techniques since they can only be applied to differentiable models. However, in order to defend against adversarial evidence generated by artefacts introduced in the computation of the explanation, these approaches can only produce coarse and low-resolution explanations.

2 RELATED WORK

Much progress in model explanation for vision tasks has been made in the last few years. For more comprehensive details of works in this topic, please refer to a survey in [36]. While some of them focus on model-specific scenarios where the users have a full ownership of the model to be diagnosed, others only considered **model-agnostic** situations where the internal details of the model is not available due to commercial or safety issues. In this section, we provide an overview of these vision explanation techniques, which roughly fall into three categories:

(1) Backpropagation Based Methods
(2) Activation Based Methods
(3) Perturbation Based Methods

2.1 Backpropagation Based Methods

Backpropagation based methods estimate the importance of each pixel by backpropagating an importance signal from an output neuron backwards through the layers to the input in one pass.

Partially Model-Agnostic: In [26], a visual explanation is generated based on computing the gradient of the class score with respect to the input image. The Integrated Gradients [29] accumulates all gradients at all points along the straightline path from a baseline input $\mathbf{x}'$ (e.g., a black image) to the given input $\mathbf{x}$. The SmoothGrad [27] samples similar images by adding noise to the image, and then take the average of the gradients as a explanation. The VarGrad [1] is similar to SmoothGrad except that it uses the variance of the gradients as an explanation.

Model-Specific: The work in [2] proposes a novel method called layer-wise relevance propagation to achieve a pixel-wise decomposition to understand the contribution of a single pixel of an image. Guided Backpropagation [28] builds on the DeConvNet explanation method [33, 34], where negative gradient entries are set to zero while back-propagating through a ReLU unit. DeepLift [25] decompose the output prediction of a neural network on a specific input by back-propagating the contributions of all neurons in the network to every feature of the input. Excitation Backpropagation [35] proposes a probabilistic Winner-Take-All formulation to model the top-down neural attentions using a novel back-propagation scheme, which integrates both top-down and bottom-up information to compute the winning probability of each neuron.

Backpropagation based methods are time efficient in generating an fine-grained explanation. However, the explanations generated by them are generally of low quality and interpretability [7, 31].

2.2 Activation Based Methods

Model-Specific: Activation based methods are belong to model-specific techniques because they typically utilize the intermediate activations from convolutional layers to generate an explanation. These works (e.g., [5, 23, 37]) use a weighted linear sum of activations from convolutional layers to estimate the importance of visual pattern at different spatial locations. In [23], the authors fuse backpropagation based methods and activation based methods to obtain high-resolution and fine-grained explanations. The major limitation of activation based methods is that they are not well suited for generating fine-grained explanations [23, 31]. Besides, the faithfulness of their resulting explanations are also not guaranteed [10, 23].

2.3 Perturbation Based Methods

Perturbation based methods explain how a model reaches a decision by perturbing around the input image $\mathbf{x}$ and observing how the model reacts to the perturbations from different regions of the image.

Model-Agnostic: Early work in [33] occludes different portions of the input image with a grey square, and estimates the importance of different regions by monitoring the change in the output of the classifier. In [19], randomly sampled occlusion masks take the places of the grey square as perturbations around the input. Lime [22] and Anchor [21] generate super-pixels for images and then use an interpretable model (e.g, linear model) to estimate the importance of each super-pixel. In [24], a region-based approach is proposed to estimate feature importance in terms of appropriately segmented regions. The works mentioned above are *completely*

model-agnostic because they only rely on the predictive outputs of the model. Despite that these methods are very flexible, they suffer from the following two major problems: 1) they may have high computational overhead s each perturbation requires a separate forward propagation through the network; and 2) they are sensitive to the occlusion strategy (e.g., Mean, Grey, Blurred or Random sampled masks).

Partially Model-Agnostic: Let $\mathbf{m}$ denote a mask for image $\mathbf{x}$ and $\mathbf{r}$ denote a reference image with little information (e.g., a blurred version of image $\mathbf{x}$ or constant values), then a perturbed image $\mathbf{x}'$ can be defined as

$$\mathbf{x}' = \mathbf{x}.\mathbf{m} + (1 - \mathbf{m})\mathbf{r}$$

In practice, the value of $\mathbf{m}$ is clamped between 0 and 1.

Some other methods [3, 4, 7, 11, 12] generate explanations by directly optimizing the perturbed image. These methods [3, 4, 7, 12] also add additional regularizations to improve the quality of the explanations and avoid adversarial evidence. In [3, 4], the authors use genebrative models like Contextual Attention GAN [32] or Variational Autoencoder [16] to fill-in the masked region. In [7], a regularizing surrogate model to produce the desired masks. The work in [12] adds total-variation (TV) norm of the masks as regularization and introduces stochasticity in the optimization. In [11], the authors introduce a new area constraint and a parametric family of smooth masks. However, all these works add some limitations on the resolution of the masks and are not fine-grained.

Model-Specific: The exceptional works [10, 31] do not belong to model-agnostic techniques because they utilize the intermediate neuron activations as regularizations to avoid adversarial evidence.

3 METHOD

In this work, we apply a similar strategy with the existing perturbation based approaches (e.g., [3, 7, 10–12, 31]), which generate explanations by optimizing perturbed images directly. These works [3, 7, 10–12, 31] follow SSR paradigm, SDR paradigm or both of them, where

(1) Smallest Supporting Region (SSR): Minimize the region of the image which must be retained in order to preserve the model output for the original image.
(2) Smallest Deletion Region (SDR): Minimize the region of the image which must be deleted in order to change the model output for the origin image.

3.1 SSR/SDR Formulation

In this section, we formalize the optimization problem of SSR and SDR paradigm adopted in the existing approaches [3, 7, 10–12, 31]. Suppose f is a pre-trained model of interest that maps an input image $\mathbf{x} \in \mathbb{R}^{3\times W\times H}$ to a output vector $\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^C$. The output vector $\mathbf{y} = (y_1, y_2, \ldots, y_C)$, $y_c \in [0, 1]$ denotes the corresponding predictive probabilities for different classes. Let c_T denotes the target class of interest, which could be either the label c^* for the input $\mathbf{x}$, or the class with the highest predictive probability $c_T = \arg\max_c y_c$ ($c_T \neq c^*$), in case we want to analyze which part of the image is most responsible for the false classification. Given a mask $\mathbf{m} \in [0, 1]^{W\times H}$, a perturbation version of the input image $\mathbf{x}$ is considered as the weighted average between the image and a reference image $\mathbf{r}$ [31].

$$\mathbf{x}_{\mathbf{r},\mathbf{m}} = \mathbf{m}.\mathbf{x} + (1 - \mathbf{m}).\mathbf{r} \tag{1}$$

The reference image $\mathbf{r}$ could be constant values, a blurred version of the original image, Gaussian noise, or sampled references of a generative model [3, 7, 10, 12, 31].

In the SSR paradigm, we want to minimize the region of the image that must be retained and meanwhile maximize the classification score for the target class c_T. By searching over the input masks, an preservation based explanation can be computed as follow:

$$\begin{aligned} \mathbf{m}^* &= \arg\max_{\mathbf{m}} f(\mathbf{m}.\mathbf{x} + (1 - \mathbf{m}).\mathbf{r})_{c_T} - \lambda||\mathbf{m}||_1 \\ &= \arg\max_{\mathbf{m}} f(\mathbf{x}_{\mathbf{r},\mathbf{m}})_{c_T} - \lambda||\mathbf{m}||_1 \\ &\Rightarrow L_{SSR}(\mathbf{m}, \mathbf{r}) = \lambda||\mathbf{m}||_1 - f(\mathbf{x}_{\mathbf{r},\mathbf{m}})_{c_T} \\ &\text{, where mask } \mathbf{m} \text{ is initialized to 0} \end{aligned} \tag{2}$$

In the SDR paradigm, the objective is to minimize the region of the image that needs to be removed in order to change the model output for the origin image. Since we are interest in the target c_T with a high classification score $f(\mathbf{x})$, a significant change in $f(\mathbf{x})$ is equivalent to minimize the classification score for the perturbation version of the image. Therefore, we can compute the corresponding deletion based explanations as follows:

$$\begin{aligned} \mathbf{m}^* &= \arg\min_{\mathbf{m}} f(\mathbf{m}.\mathbf{x} + (1 - \mathbf{m}).\mathbf{r})_{c_T} - \lambda||\mathbf{m}||_1 \\ &= \arg\min_{\mathbf{m}} f(\mathbf{x}_{\mathbf{r},\mathbf{m}})_{c_T} - \lambda||\mathbf{m}||_1 \\ &\Rightarrow L_{SDR}(\mathbf{m}, \mathbf{r}) = f(\mathbf{x}_{\mathbf{r},\mathbf{m}})_{c_T} - \lambda||\mathbf{m}||_1 \\ &\text{, where mask } \mathbf{m} \text{ is initialized to 1} \end{aligned} \tag{3}$$

3.2 Limitation

State-of-the-art deep neural networks are vulnerable to adversarial examples that are only slightly different from the original input image [13, 15, 30]. As a result, one major problem of perturbation based approaches is that, they are susceptible to adversarial noises because their optimization formulation is similar with the adversarial attack techniques.

To address the artefacts, stochastic noise and total-variation (TV) norm regularization is introduced in the optimization to encourage the mask to have a simple, regular structure which cannot be co-adapted to artefacts [7, 12]. In [11], a Gaussian or similar Kernel is used to convolve the mask to obtain a explanation regular structure. These methods [3, 4, 7, 10–12] optimize on a low-resolution mask and then upsample of the computed mask. As a result, the resulting explanations are usually coarse and n contain a lot of background information that is not responsible for the model prediction. In [10, 31], the intermediate outputs are utilized to defense adversarial evidence. However, they require the internal states of the model of interest, which may not be available sometimes.

Another common issue shared by these methods is that the hyper-parameter λ have a great impact on the final explanation. For example, in the SDR paradigm, $\mathbf{m}_{i,j}$ will stop deceasing when

$$\frac{\partial L_{SDR}(\mathbf{m}, \mathbf{r})}{\partial \mathbf{m}} = \left\{ \frac{\partial f(\mathbf{x}_{\mathbf{r},\mathbf{m}})}{\partial \mathbf{x}_{\mathbf{r},\mathbf{m}}} . (\mathbf{x} - \mathbf{r}) \right\} - \lambda < 0$$

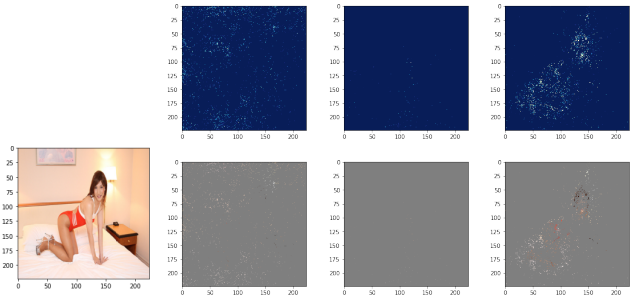
If the classifier is locally smooth around $\mathbf{x}$, which implies it has a small gradient, then a relative larger λ can usually lead the optimizer to be stuck at $m_{i,j} = 1$. Similarly, in the SSR paradigm, the optimizer may also be stuck at $\mathbf{m}_{i,j} = 0$ if λ is too small. As a result, typically, a very small λ is used in existing methods (e.g., 10^{-4} in [12] and 10^{-3} in [4]) On the other hand, if the value of λ is too small, then the optimizer requires much more iterations to converge and more background information will be kept in the explanations.

3.3 Hybrid of SDR and SSR

Based on the observation that the objective of SSR or SDR is not opposite and serves as a complement to each other, we consider the hybrid of them [7, 10]. Let $\mathbf{x}_{\mathbf{r},\mathbf{m}}$ denote the deletion based explanation with a low classification score $f(\mathbf{x}_{\mathbf{r},\mathbf{m}}) \approx 0$. If the explanation is faithful and captures the most informative region of the input image that contributes to the model output, the corresponding preservation based explanation $\mathbf{x}_{\mathbf{r},1-\mathbf{m}}$ should have a high classification score $f(\mathbf{x}_{\mathbf{r},1-\mathbf{m}}) \approx 1$. Therefore, in this work, we consider to optimise $\mathbf{x}_{\mathbf{r},1-\mathbf{m}}$ and $\mathbf{x}_{\mathbf{r},\mathbf{m}}$ together, i.e.,

$$L_{DSSR}(\mathbf{m}, \mathbf{r}) = f(\mathbf{x}_{\mathbf{r},\mathbf{m}})_{c_T} - f(\mathbf{x}_{\mathbf{r},1-\mathbf{m}})_{c_T} - \lambda ||\mathbf{m}||_1 \quad \text{where, } \mathbf{m} \text{ is initialized to } 0 \tag{4}$$

Note that, in this formulation, we initialize $\mathbf{m}$ to 0 instead of 1, so that we are able to choose a large λ and $||\mathbf{m}||_1$ would not stuck at 0.



(a) Image (b) SSR [12] (c) SDR [12] (d) DSSR

Figure 2: A high resolution mask that leads to explanations that are either based on false adversarial evidence or explanations that lack of clarity.

$$\Rightarrow \frac{\partial L_{DSSR}(\mathbf{m}, \mathbf{r})}{\partial \mathbf{m}_{i,j}} = (\mathbf{x} - \mathbf{r}) \left(\frac{\partial f(\mathbf{x}_{\mathbf{r},\mathbf{m}})}{\mathbf{x}_{\mathbf{r},\mathbf{m}}} + \frac{\partial f(\mathbf{x}_{\mathbf{r},1-\mathbf{m}})}{\mathbf{x}_{\mathbf{r},1-\mathbf{m}}} \right) - \lambda$$

$$\mathbf{m}^{t+1} = \mathbf{m}^t + \delta\lambda - \delta(\mathbf{x} - \mathbf{r}) \left(\frac{\partial f(\mathbf{x}_{\mathbf{r},\mathbf{m}})}{\mathbf{x}_{\mathbf{r},\mathbf{m}}} + \frac{\partial f(\mathbf{x}_{\mathbf{r},1-\mathbf{m}})}{\mathbf{x}_{\mathbf{r},1-\mathbf{m}}} \right)$$

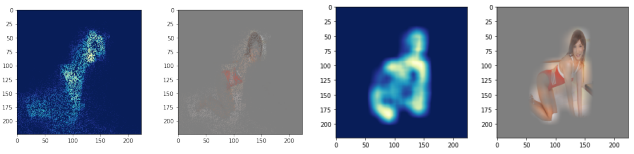
However, if we directly use a larger λ, $||\mathbf{m}||_1$ will tend to increase while $f(\mathbf{x}_{\mathbf{r},1-\mathbf{m}})_{c_T}$ may still be close to 0. Therefore, to encourage the resulting explanations to be faithful, i.e., $f(\mathbf{x}_{\mathbf{r},1-\mathbf{m}})_{c_T} \rightarrow 1$, we choose a two-stage λ. As long as $f(\mathbf{x}_{\mathbf{r},1-\mathbf{m}})_{c_T} \leq \theta$, then a smaller λ is used, and otherwise, a larger λ can be applied.

3.4 Fine-Grained Explanation with Regular Structure

Explanations generated by existing perturbation based approaches [3, 4, 7, 10–12] are limited to being low resolution, which prevents discriminative evidence from being visualized. The reason that a low resolution mask is used is that 1) pixel-level high resolution explanation are more vulnerable to adversarial evidence and 2) the high resolution explanation usually does not have a regular structure.

For example, in Figure 2, we show the explanations generated by SSR/SDR [12] and DSSR (Equation 4) when a high resolution mask is used. We set mask size $k = 224$, and $\lambda = \frac{1}{k^2}$ for SSR/SDR [12] and $\lambda = \frac{10}{k^2}$ for DSSR in Equation 4. A reference image sampled from a uniform distribution $U(0, 1)$ is used in this example. As we can observe, the explanation produced by SDR is based on adversarial evidence. On the other hand, the explanations generated by SDR and DSSR in Equation 4 have a irregular structure and poor clarity.

In order to encourage the explanation to be discriminative and clear, a new regularization term $||1 - \text{maxp}(1 - \mathbf{m}|k_1, s_1, p_1)||_1$ is used to replace the original regularization term $||\mathbf{m}||_1$, where maxp denotes a max-pooling operation



(a) (b) (c) (d)

Figure 3

with window size $k_1 \times k_1$, stride s_1 and padding p_1. In this formulation, only the smallest $m_{i,j} \in \mathbf{m}$ within the pooling window will be updated in each iteration. For example, suppose we have a 4×4 mask $\mathbf{m}$

$$\mathbf{m} = \begin{matrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{matrix}$$

, and maxpool function has a window size 2, stride 1 and pad 0, then

$$\frac{\partial ||\mathbf{m}||_1}{\partial \mathbf{m}} = \begin{matrix} 0 & 2 & 2 & 0 \\ 6 & 0 & 0 & 0 \\ 2 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}$$

Suppose pixel (i, j) (e.g, $m_{1,0}$ in this example) is around masked pixels (i.e., pixels that are unimportant to the model output), then pixel (i, j) is also likely to be irrelevant to the model output. As $\mathbf{m}_{i,j}$ has a larger $\frac{\partial ||\mathbf{m}||_1}{\partial m_{i,j}}$, pixel (i, j) is encouraged to be masked. Figure 3 (a) and (b) presents the resulting explanations when we use $||1 - \text{maxp}(1 - \mathbf{m}|12, 1)||_1$ as the regularization term. The proposed regularization term enables the explanation to focus a particular object.

Pixel level explanation has higher risk being attacked by false evidence especially for objects that are not the most-likely class of the model. We then use a auxiliary mask $\bar{\mathbf{m}} = \text{avgp}(\mathbf{m}|k_2, s_2, p_2)$ to increase the robustness and faithful of the explanations, where $\text{avgp}(\mathbf{m}|k_2, s_2, p_2)$ denotes a average pooling function with window size k_2, stride s_2 and padding p_2. Thus, the importance of a single pixel will be determined by the pixels around it. As a result, the loss function is modified as follows.

$$L_{DSSR+}(\mathbf{m}, \mathbf{r}) = f(\mathbf{x}_{\mathbf{r},1-\bar{\mathbf{m}}})_{c_T} - f(\mathbf{x}_{\mathbf{r},\bar{\mathbf{m}}})_{c_T} - \lambda||1 - \text{maxp}(1 - \mathbf{m}|k_1, s_1, p_1)||_1, \text{where, } \bar{\mathbf{m}} = \text{avgp}(\mathbf{m}|k_2, s_2, p_2) \quad (5)$$

In Figure 3 (c) and (d), we present the resulting explanations according to Equation 5.

4 EXPERIMENTS

In this section, we compare our method with other state-of-the-art approaches qualitatively and quantitatively. We implement our explanation method using Pytorch [18]. Masks are optimised using Adam [17], initializing them with all zeros. We set $k_1 = 12, s_1 = 1, p_1 = 6$ for the max pooling $\text{maxp}(1 - \mathbf{m}|k_1, s_1, p_1)$ and $k_1 = 16, s_1 = 1, p_1 = 0$ for average pooling $\text{avgp}(\mathbf{m}|k_2, s_2, p_2)$.

4.1 Qualitative Comparison

Implementation details: For the proposed method, masks are optimised using Adam with learning rate 0.1 and $\lambda = 10/224^2$ for 300 iterations. A pre-trained ResNet34 [14] is used as the base model. For MP, the mask size is set to 28. Other methods i.e., Extreme Perturbation [11], GradCam [23], where the saliency_layer is 'layer4.1.relu', and RISE [19] are implemented by TorchRay[1] with the default setting.

In Table 1, we present a qualitative comparison between our method and others. For MP [12], we show the explanations for both SSR and SDR. As we can observe, our method delivers the most discriminative explanations and the masks tend to cover the object of interest tightly with little background information.

For MP, its explanations include a lot of pixels that are far from the target object. The explanations generated by our method faithful and verifiable. The deletion of the highlighted pixels prevents the model from correctly predicting the object. On the other hand, the model can still make a correct prediction when the highlighted pixels are preserved and other pixels are replaced by with random noise.

Besides, visual explanations should also be class discriminative. In Table 3, we show explanations for images containing two objects. As we can observe, our method is able to generate class discriminative explanations and only highlights pixels of the chosen target class.

For RISE and GRAD-CAM, the saliency maps $\mathbf{m}$ is normalized as follows

$$\mathbf{m} \longleftarrow = \frac{\mathbf{m} - \mathbf{m}.\min()}{\mathbf{m}.\max() - \mathbf{m}.\min()}$$

because the variance of original mask can be very small.

4.2 Quantitative Comparison

Despite a growing body of research focusing on explaining machine learning models, there is still no consensus about how to measure the interpretability [20] quantitatively. Human evaluation has been the predominant way to assess model explanation by measuring it from the perspective of transparency, user trust or human comprehension of the

[1] https://github.com/facebookresearch/TorchRay

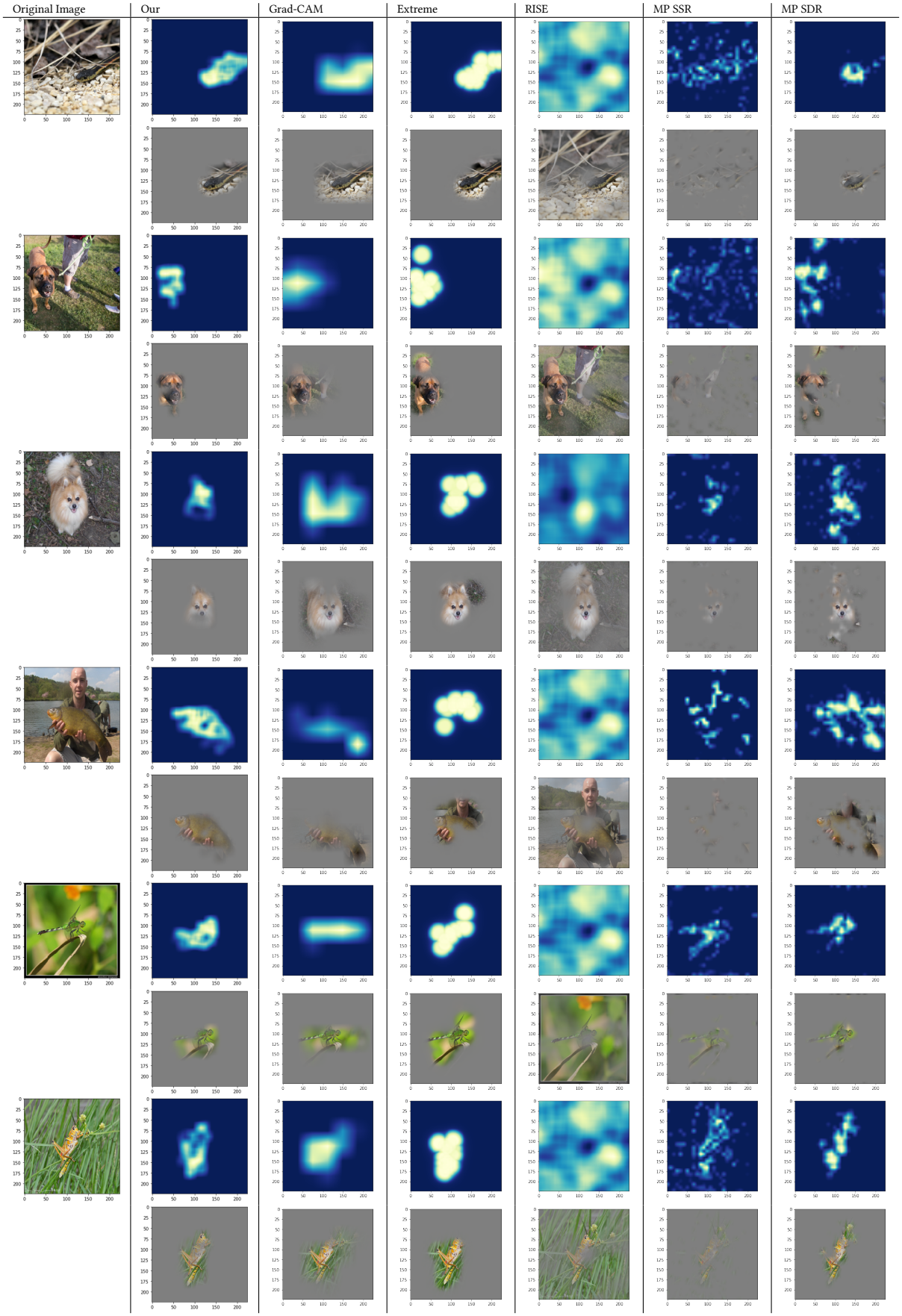


Table 1: Comparison with several state-of-the-art explanation methods.

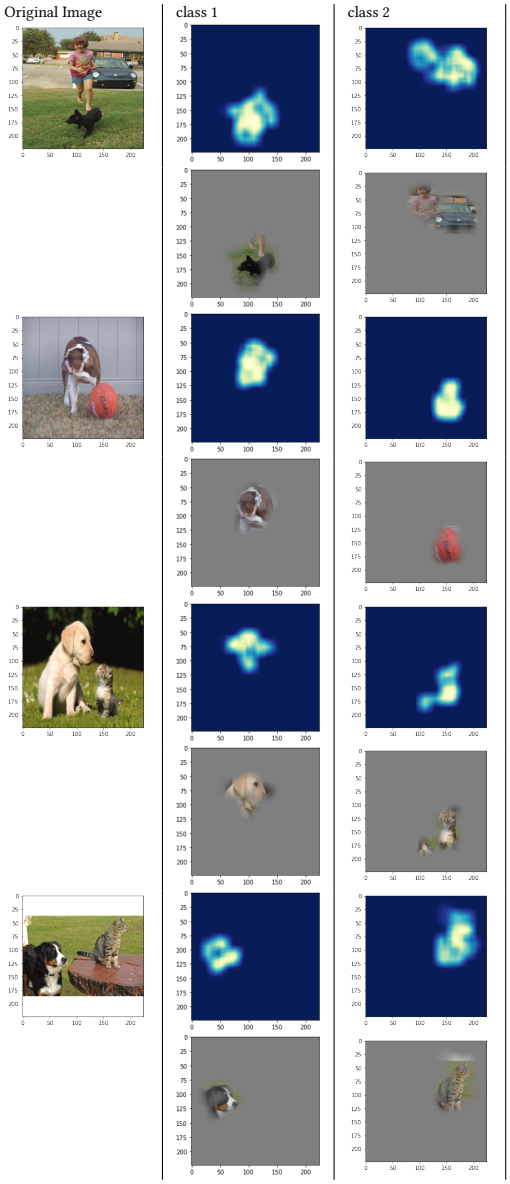


Table 2: Explanations for images with two objects

decisions made by the model. Here we also evaluate explanations in terms of a human evaluation metric, the pointing game introduced in [32]. If the highest saliency point lies inside the human annotated bounding box of an object, it is counted as a hit. The pointing game accuracy is given by $\frac{num\_hit}{num\_hit+num\_miss}$, averaged over all target categories in the dataset.

The performance in terms of pointing game accuracy is shown in Table3 for the test split of PASCAL VOC07 and the class-wise accuracy is shown in Figure 4.

Cntr	Grad	DConv	Guid	MWP	cMWP	RISE	GCAM	Extreme	our
69.6	72.3	68.6	77.9	84.4	90.7	86.4	90.4	88.9	85.3

Table 3: Pointing Accuracy

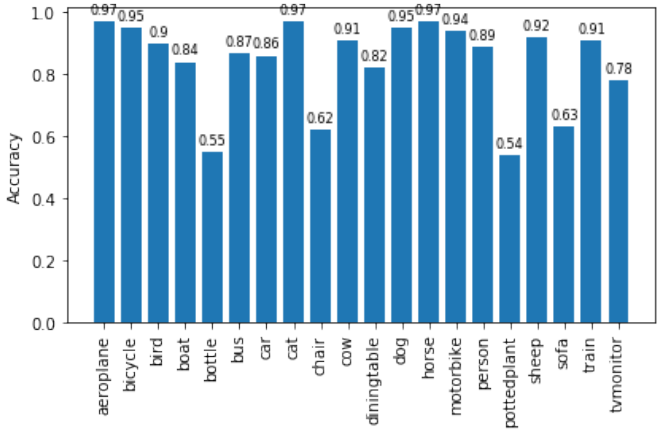


Figure 4: Class-wise hit accuracy.

ACKNOWLEDGMENTS

This work was supported by

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. 2018. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307* (2018).

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[3] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2018. Explaining Image Classifiers by Adaptive Dropout and Generative In-filling. *CoRR* abs/1807.08024 (2018). arXiv:1807.08024 http://arxiv.org/abs/1807.08024

[4] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2019. Explaining Image Classifiers by Counterfactual Generation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1MXz20cYQ

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 839–847.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[7] Piotr Dabkowski and Yarin Gal. 2017. Real Time Image Saliency for Black Box Classifiers. arXiv:stat.ML/1705.07857

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[9] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2009. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 304–311.

[10] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. 2018. Towards Explanation of DNN-based Prediction with Guided Feature Inversion. arXiv:cs.CV/1804.00506

[11] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. arXiv:cs.CV/1910.08485

[12] Ruth Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *CoRR* abs/1704.03296 (2017). arXiv:1704.03296 http://arxiv.org/abs/1704.03296

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. arXiv:stat.ML/1412.6572

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[15] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial Attacks on Neural Network Policies. *CoRR* abs/1702.02284 (2017). arXiv:1702.02284 http://arxiv.org/abs/1702.02284

[16] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 107.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[19] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).

[20] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2018. Manipulating and Measuring Model Interpretability. *CoRR* abs/1802.07810 (2018). arXiv:1802.07810 http://arxiv.org/abs/1802.07810

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[22] Sameer Singh Ribeiro, Marco Tulio and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. (2016).

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

[24] Dasom Seo, Kanghan Oh, and Il-Seok Oh. 2018. Regional Multi-scale Approach for Visually Pleasing Explanations of Deep Neural Networks. *CoRR* abs/1807.11720 (2018). arXiv:1807.11720 http://arxiv.org/abs/1807.11720

[25] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3145–3153.

[26] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[27] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[28] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3319–3328.

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[31] Jörg Wagner, Jan Mathias Köhler, Tobias Gindele, Leon Hetzel, Jakob Thaddäus Wiedemer, and Sven Behnke. 2019. Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks. arXiv:cs.CV/1908.02686

[32] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5505–5514.

[33] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[34] Matthew D Zeiler, Graham W Taylor, Rob Fergus, et al. 2011. Adaptive deconvolutional networks for mid and high level feature learning.. In *ICCV*, Vol. 1. 6.

[35] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision* 126, 10 (2018), 1084–1102.

[36] Quanshi Zhang and Song-Chun Zhu. 2018. Visual Interpretability for Deep Learning: a Survey. *CoRR* abs/1802.00614 (2018). arXiv:1802.00614 http://arxiv.org/abs/1802.00614

[37] Bolei Zhou, 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*. 2921–2929.