EVET: Enhancing Visual Explanations of Deep Neural Networks Using Image Transformations

Youngrock Oh
Technology Research, Samsung SDS
Seoul, Republic of Korea
y52.oh@samsung.com

Hyungsik Jung
jhyungsik89@gmail.com

Jeonghyung Park
jeonghyung.park@gmail.com

Min Soo Kim
minsoo07.kim@samsung.com

Abstract

Numerous interpretability methods have been developed to visually explain the behavior of complex machine learning models by estimating parts of the input image that are critical for the model's prediction. We propose a general pipeline of enhancing visual explanations using image transformations (EVET). EVET considers transformations of the original input image to refine the critical input region based on an intuitive rationale that the region estimated to be important in variously transformed inputs is more important. Our proposed EVET is applicable to existing visual explanation methods without modification. We validate the effectiveness of the proposed method qualitatively and quantitatively to show that the resulting explanation method outperforms the original in terms of faithfulness, localization, and stability. We also demonstrate that EVET can be used to achieve desirable performance with a low computational cost. For example, EVET-applied Grad-CAM achieves performance comparable to Score-CAM, which is the state-of-the-art activation-based explanation method, while reducing execution time by more than 90% *on VOC, COCO, and ImageNet.*

1. Introduction

Although supervised deep neural network models achieve impressive performance in terms of the prediction accuracy, their complexities obscure the underlying process from which inferences are derived. Accordingly, methods that render the model behavior human-interpretable [14, 25] are important for improving the model performance and for assessing the acceptability of model decisions. A number of interpretability methods [21, 23, 8, 17, 19] provide visual explanations that identify important regions of the input image on which a model bases its decision. By helping understand the inner workings behind the model's prediction, visual explanations have shown useful for achieving better performance in tasks such as segmentation [12], classification [9, 10], and visual question answering [15, 16, 20].

In this work, we propose a pipeline of enhancing visual explanations using image transformations (EVET) that is compatible with any existing visual explanation methods and can be used in conjunction with them. A brief overview of EVET is illustrated in Fig. 1. In EVET, we transform the input image by applying image transformations such as rotation and horizontal flip. We then compute the pixel-wise importance maps of the transformed images with a given explanation method and inversely transform them into the original image space. We claim that overlapping regions produced by the inverted importance maps are essential to the model's prediction. The underlying rationale is that the region estimated to be important in variously transformed inputs is more important than the region that is not. To be more specific, we determine the weight of each inverted importance map as the target class probability of the corresponding perturbed image and reweigh for parts that are deemed important consistently across multiple inverted importance maps. This refined weighted sum is the final map of EVET and we show that it results in improved performance in terms of faithfulness, localization, and stability.

We demonstrate that EVET can also be used for computational efficiency. By applying EVET to an existing computationally cheap explanation method, we can achieve desirable performance while reducing computational cost surprisingly. For example, EVET-applied Grad-CAM with 7 transformations achieves performance comparable to Score-CAM, which is the state-of-the-art activation-based explanation method, in terms of faithfulness while reducing execution time by more than 90% on VOC, COCO, and ImageNet.

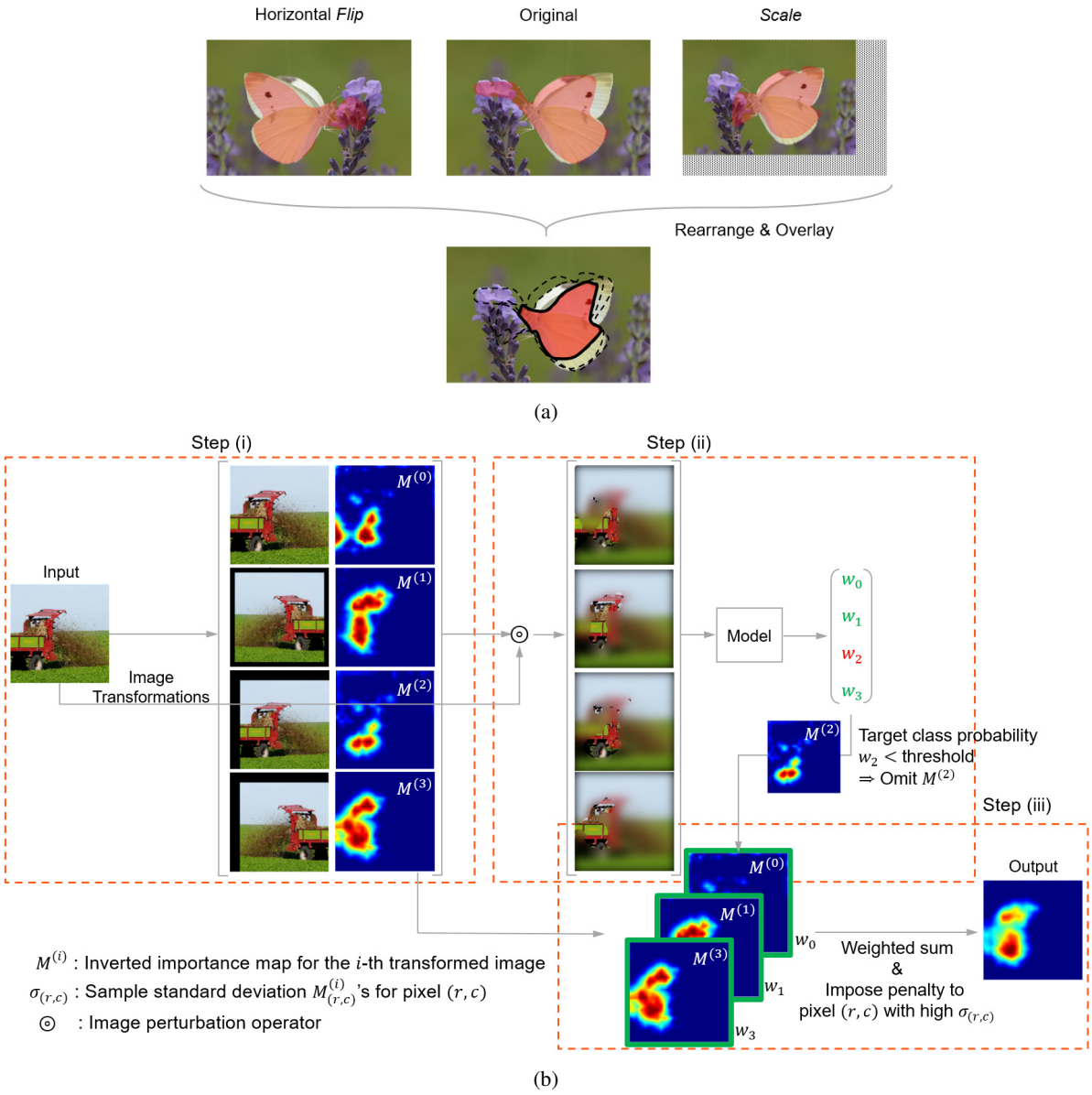


Figure 1: An overview of our proposed EVET. (a) Conceptual figure of our motivation. (b) Pipeline of our proposed EVET. Visual explanations are obtained by applying inverse transformation to the explanations of the transformed images from the original method in step (i). For each importance map, its weight is determined by the target class probability of the corresponding perturbed image, and ones with low target class probabilities are omitted in step (ii). The filtered explanations are linearly combined with the weights and penalties are imposed to pixels with unreliable saliencies in step (iii).

In short, our contributions are as follows:

- Method to improve existing visual explanation methods. We propose a way of enhancing explanation methods, EVET, by considering image transformations. We validate EVET qualitatively and quantitatively, and the results show that it brings considerable improvement.
- Computational efficiency of enhancing the explanation methods. By applying EVET to a computationally cheap explanation method, we can achieve desirable performance while reducing computational cost surprisingly.
- Evaluation metric for visual explanation methods. We propose a new metric to evaluate explanation methods in terms of stability, which captures robustness of

the explanation method by comparing the importance maps of the original input image and transformed image.

2. Related Work

Since interpretability has no objective ground-truth answers, diverse methods have been proposed to quantify the contribution of a pixel or a group of pixels to the model's prediction. There are three main categories of explanation methods widely used: backpropagation-based, perturbation-based, and activation-based methods. First, backpropagation-based methods [21, 24, 23, 1] utilize the backpropagated gradient of the target class score with respect to the input and formulate that a large absolute value of the gradient of the input is equivalent to high importance in the prediction. These methods provide fine-grained and yet often noisy importance maps relatively fast and easily. Second, perturbation methods [3, 8, 7, 17, 26] observe how the model reacts when a certain part of the input is altered, under the rationale that the more important a pixel is to the model, the more the prediction fluctuates when the particular pixel is perturbed. Perturbation methods are intuitive and generally produce high-quality importance maps, but they require hyper-parameter tuning for weight coefficients of additional regularization terms and are time-consuming due to iterations for optimization for each instance. Lastly, activation-based methods [29, 19, 4, 5, 28] generate an importance map as a linear combination of the activation maps from the convolutional layers, hence focusing on how to determine the weight coefficients of the activation maps. While rapidly computable, the resulting importance maps tend to be coarse-grained and can be applied only to the model which contains a convolutional layer.

On the top of these findings, there have been some efforts to attain enhanced visual explanations using the above explanation methods as input methods. Specifically, [1, 23, 27] employ extra importance maps of the altered input images to achieve better visual performance. They generate random noisy copies of the input image and use the average [23] or variance [1] of the resulting importance maps, respectively. [27] introduces a smooth operation on Score-CAM [28] to determine the weights of activation maps. Our approach differs from these schemes in two aspects. First, EVET uses deterministic image transformations, which can be favorable since it does not require computing many additional importance maps to guarantee stable performance. Second, it uses the weighted sum of the additional importance maps and adjusts the result considering the sample variance of them, thereby producing more reliable importance maps.

3. Image Transformations in Explanation Methods

We utilize image transformations for the following two purposes. First, we enhance explanation methods by considering importance maps obtained from the transformed images (Sec 3.1). We also introduce a new evaluation metric that measures stability of importance maps with respect to image transformations (Sec 3.2).

3.1. Methodology: EVET

Our main idea is that a pixel estimated to be important simultaneously in the inverted importance maps obtained from the transformed images is more important than those that are not. As illustrated in Fig 1b, EVET consists of three major steps: (i) computation of importance maps from the original and transformed images; (ii) assessment of the importance maps; (iii) determination of the final importance map. Our goal is to enhance a explanation method J where the importance map of an input image x is denoted by $J(x)$. The details of each step are described below.

Preparation step We first list the requirements for an image transformation to be used in EVET.

- We use the transformations that preserve most of the information in the input image; the model's prediction should be insensitive to the transformation of the original image.
- It should be ensured that the importance map of the transformed image can be inversely transformed into the original image space in an accurate manner for a fair comparison with the original importance map.

In this work, we consider a set of geometric image transformations such as *scale*, *rotate*, *shear*, and horizontal *flip*. Although we focus on geometric image transformations here, the same logic can be applied to other types of transformations satisfying the above requirements.

Step (i) For a given set of image transformations $\{T_i\}_{i=1,\dots,n}$, we compute the inverted importance maps $\{M^{(i)}\}_{i=1,\dots,n}$ of the transformed image given by

$$M^{(i)} = T_i^{-1}\left(J\left(T_i(x)\right)\right) \tag{1}$$

for $i = 1, \dots, n$. For brevity, let $M^{(0)}$ be the importance map of the original image and define $\boldsymbol{M} = \{M^{(i)}\}_{i\in I}$ where $I = [1, \dots, n]$. In case where T_i is not invertible, one can use a pseudo inverse instead.

Step (ii) Note that EVET is based on the premise that the model's prediction is robust with respect to the transformation of x given class k of interest. However, there is a possibility that T_i might affect the model's inference considerably for some $i \in I$. To address this problem, we

filter out suspicious importance maps whose perturbed image has a lower target class probability compared to that of the original map. More specifically, we exclude $M^{(j)}$ from $\boldsymbol{M}$ given threshold Γ if

$$f_k\left(\Phi(x, M^{(j)})\right) \leq \Gamma f_k\left(\Phi(x, M^{(0)})\right) \tag{2}$$

where the perturbed image $\Phi(x, M)$ is defined by

$$\Phi(x, M) := x \otimes M + x_d \otimes (1 - M) \tag{3}$$

and x_d, $\otimes$, and f_k denote the distorted version of x (e.g. a blurred or black image), element-wise multiplication, and softmax output in the model f of class k, respectively. Let $\tilde{I}$ denote the set of the indices of the maps left after filtering out. Here, we upsample M into the input size if necessary. In our experiments, Γ is set as 0.99 and black images are used as x_d.

Step (iii) Next, we consider a weighted sum $\hat{M}$ to form a single importance map which inherits condensed information from $\{M_i\}_{i\in\tilde{I}}$. The weights are set in proportion to the target class probabilities of the perturbed images, so that each weight represents the estimated importance of each map. That is, $\hat{M}$ is given by $\hat{M} = \sum_{i\in\tilde{I}} w_i M^{(i)}$ where

$$w_i = \frac{f_k\left(\Phi(x, M^{(i)})\right)}{\sum_{j\in\tilde{I}} f_k\left(\Phi(x, M^{(j)})\right)} \text{ for } i \in \tilde{I}. \tag{4}$$

Furthermore, we assess the importance of each pixel in terms of stability. We maintain that for a given pixel location (r, c), a small sample standard deviation $\sigma_{(r,c)}$ of $\{M^{(i)}_{(r,c)}\}_{i\in\tilde{I}}$ implies that the estimated saliencies of the corresponding pixel are stable, and therefore reliable. Based on the above perspective, we adjust $\hat{M}_{(r,c)}$ using $\sigma_{(r,c)}$ such that $\hat{M}_{(r,c)}$ with small $\sigma_{(r,c)}$ is boosted for $(r, c) \in \{1, \ldots, R\} \times \{1, \ldots, C\}$. For instance, if a pixel (r, c) is estimated to be important (i.e., $\hat{M}_{(r,c)} > 0.5$) and $\sigma_{(r,c)}$ is small, we increase $\hat{M}_{(r,c)}$. On the other hand, we decrease $\hat{M}_{(r,c)}$ if $\hat{M}_{(r,c)} < 0.5$ and $\sigma_{(r,c)}$ is small. Formally, given $\hat{M}_{(r,c)}$, we define the boosted saliency $\tilde{M}_{(r,c)}$ as

$$\tilde{M}_{(r,c)} = \max\left(\min\left(\hat{M}_{(r,c)} + \alpha(\hat{M}_{(r,c)} - 0.5), 1\right), 0\right) \tag{5}$$

and set the final $M^*_{(r,c)}$ as a weighted average of $\hat{M}_{(r,c)}$ and $\tilde{M}_{(r,c)}$ as follows:

$$M^*_{(r,c)} = \sigma_{(r,c)}\hat{M}_{(r,c)} + (1 - \sigma_{(r,c)})\tilde{M}_{(r,c)}. \tag{6}$$

One can see that a smaller standard deviation makes the final $M^*_{(r,c)}$ closer to the boosted saliency. In (5), the max and min functions are to ensure that $M^*_{(r,c)}$ stays within $[0, 1]$. We also suppose that $\sigma_{(r,c)}$ is normalized by the upper bound for the given sample size (i.e., $\sigma_{(r,c)} \in [0, 1]$). Here, α is a hyper-parameter which controls the amount of the boost. As a final step, we perform min-max normalization of M^*. Fig. 1b describes the procedure drawing out the final importance map M^* from $\boldsymbol{M}$.

Let J^* denote the resulting explanation method when EVET is applied to J. Combining the above steps together, we rewrite the final importance map in matrix notation:

$$J^*(x) = V \otimes \hat{M} + (1 - V) \otimes \tilde{M} \tag{7}$$

where V represents the standard deviation matrix.

In summary, EVET outputs the importance map from the auxiliary maps derived from the transformed images in the following two perspectives. Each inverted importance map $M^{(i)} \in \boldsymbol{M}$ is assessed in terms of the target class probability of its perturbed image (focusing on a single importance map), then enforced to boost the estimated importance of a pixel presumed to be stable (observing multiple importance maps).

3.2. Evaluation Metric: Stability Index

If a transformation is not influential to the model's prediction, it is reasonable to assume that the high dissimilarity between the original and the inverted importance map implies that the explanation method is unfavorable. From this viewpoint, we propose a new evaluation metric for explanation methods, called the stability index (SI). We first define a similarity function $\Psi_d(M_1, M_2)$ of two importance maps M_1, M_2 for a given difference measure $d(\cdot, \cdot)$ as follows:

$$\Psi_d(M_1, M_2) = 1 - \frac{d(M_1, M_2)}{d_{max}} \tag{8}$$

where d_{max} is the maximum difference, i.e., $\max_{M_1,M_2} d(M_1, M_2)$, so the SI is normalized into the range $[0, 1]$. We use Frobenius norm as $d(\cdot, \cdot)$, and accordingly d_{max} is the square root of the number of elements in M_1.

For a given explanation method J, transformation T and dataset D, we define SI by:

$$SI = \frac{\sum_{x\in D} \Psi_d(J(x), T^{-1}(J(T(x))))}{|D|}. \tag{9}$$

Therefore, a larger value of SI indicates that the explanation method is more invariant with respect to the transformation.

4. Experiments

In this section, we validate the effectiveness of EVET both qualitatively and quantitatively. The results demonstrate that EVET generates improved explanation methods which produce more accurate and stable importance maps. We also confirm that EVET can be used to achieve desirable performance efficiently. For instance, EVET-applied

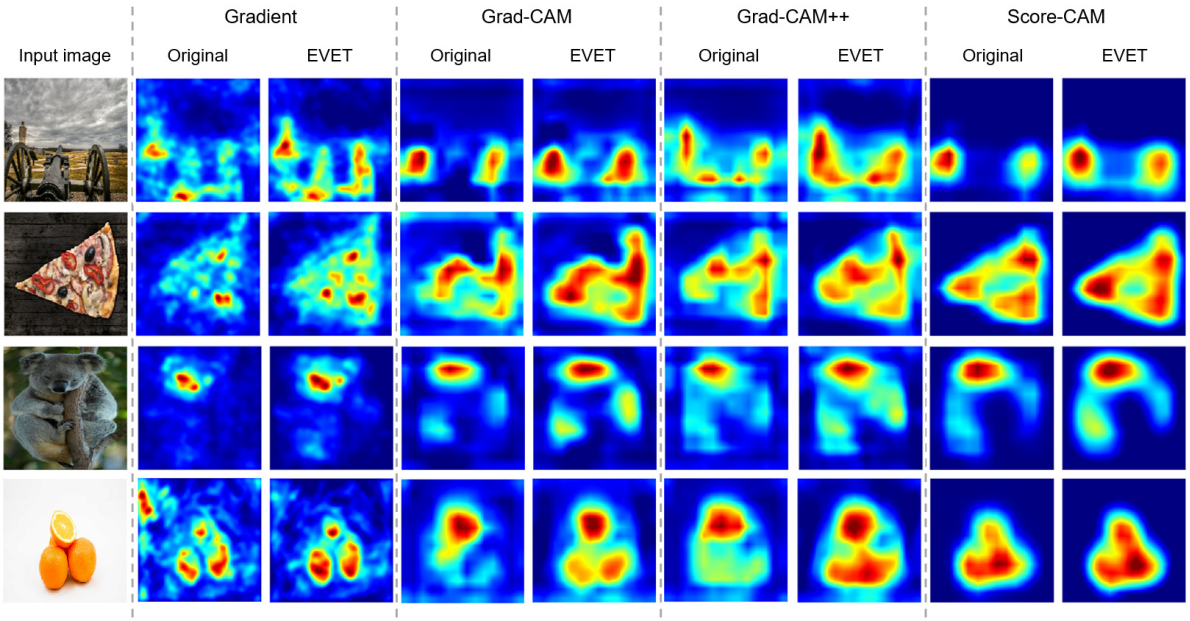


Figure 2: Qualitative comparison before and after applying EVET. Given an explanation method, the left and right columns represent the original and EVET-applied results, respectively. Gradient [21], Grad-CAM [19], Grad-CAM++ [4], and Score-CAM [28] are used with a VGG-16 network trained on ImageNet.

Grad-CAM [19] achieves performance comparable to the state-of-the-art activation based method, Score-CAM [28], with shorter execution times.

Experimental setup Our experiments are conducted on the following datasets: PASCAL VOC [6] (2007 segmenation task test set of 210 images, abbreviated as VOC), COCO [13] (2014 validation set $\simeq 50k$ images), and ImageNet [18] (2012 validation set of $50k$ images). For COCO and ImageNet, we use 1000 randomly sampled images for quantitative evaluation. We use the ResNet-50 [11] models provided in Torchray [1] for VOC and COCO. Pre-trained VGG-16 [22] network from the torchvision models [2] is used for ImageNet. For explanation methods to apply EVET, we consider Gradient [21], Grad-CAM [19], Grad-CAM++ [4], and Score-CAM [28]. Hyper-parameter α is chosen for each case separately and the detailed procedure is provided in Section C. of the supplementary materials.

For implementation, we use a variant of Score-CAM, called Faster-Score-CAM [3], instead of the original Score-CAM unless mentioned otherwise. It is designed to reduce execution time by utilizing activation maps with large variances only.

Choice of transformations As candidates of transformations to be used for EVET, we consider *scale*, *rotate*, *shear*, and horizontal *flip* since they are widely used in image data augmentation and easy to apply. Specifically, we use horizontal flip, scale (by 0.9), rotate (by $10°$), shear (by 0.2), and the transformations that apply these scale, rotate, shear after horizontal flip. Given the set of transformations, we evaluate the performance when its subsets are used for EVET. Empirically, we observe that EVET tends to show better performance when more transformations are used, so we decide to use all 7 transformations. The more details are provided in Section B. of the supplementary materials.

4.1. Qualitative Evaluation of EVET

Fig. 2 provides a qualitative comparison between before and after applying EVET to existing explanation methods. It indicates that EVET generally removes redundant noise and generate importance maps that cover more pixels belonging to the target object compared to the original method. Especially, the improvement is significant for Gradient. The results in Table 5 provide useful clues to explain the reason. Note that the low values of SI imply that there are large differences between the importance maps in $\boldsymbol{M}$. Therefore, there is a plenty of room for the final map, which is a combination of many different inverted maps, to be improved from the original one.

[1] We use the code from https://github.com/facebookresearch/TorchRay for explanation methods: Gradient, Grad-CAM, Guided Backpropagation [24], and extremal perturbation [7].

[2] https://github.com/pytorch/vision/blob/master/torchvision/models/vgg.py

[3] https://github.com/tabayashi0117/Score-CAM

		Gradeint			Grad-CAM			Grad-CAM++			Score-CAM		
		VOC	CO	Img	VOC	CO	Img	VOC	CO	Img	VOC	CO	Img
Average	Original	24.8	24.4	40.6	11.0	15.3	27.8	9.0	13.2	26.2	3.4	7.7	19.0
drop (%)	w/ EVET	16.6	17.4	30.0	7.8	12.2	20.4	7.0	10.7	18.2	2.0	5.3	14.4
Increase in	Original	29.5	40.1	21.6	50.0	57.5	32.7	41.4	52.4	28.7	53.3	59.8	36.5
confidence (%)	w/ EVET	36.7	47.5	29.4	55.2	60.4	38.6	47.1	55.4	38.0	60.0	64.6	41.4
Win (%)	Original	25.7	29.8	25.2	40.5	42.0	26.2	38.1	37.3	27.5	40.5	31.6	32.7
	w/ EVET	70.5	69.5	73.3	51.9	56.0	72.0	55.2	61.3	71.0	54.8	66.6	65.3

Table 1: Faithfulness evaluation of EVET on VOC, COCO, and ImageNet. Results before and after applying EVET are presented. Lower is better for average drop; higher is better for increase in confidence and win.

4.2. Quantitative Evaluation of EVET

We examine the importance maps generated by EVET on the object recognition task by faithfulness tests adopted in [4, 5, 28] which evaluate the faithfulness of the explanations using three metrics: (i) average drop; (ii) increase in confidence; (iii) win. We also assess the generated importance maps in terms of localization ability based on the energy-based pointing game.

Faithfulness evaluation Generally, occluding parts of an image decreases its target class probability estimated by the model. However, if we occlude only unimportant regions while maintaining most of the important regions for the prediction of the target class, this fall would be small. From this perspective, average drop is formulated by:

$$\text{Average drop}(\%) = \frac{1}{N}\sum_{i=1}^{N}\frac{max(0, Y_i^c - O_i^c)}{Y_i^c} \times 100 \tag{10}$$

where N is number of images in the dataset, and Y_i^c and O_i^c denote the softmax outputs of the image x_i and the perturbed image for class c, respectively, i.e., $Y_i^c = f_c(x_i)$ and $O_i^c = f_c(\Phi(x_i, J(x_i)))$. Complementary to the above, highlighting the important regions of an image can increase the target class probability of the perturbed image. Increase in confidence is defined by the ratio of the images having such an increase:

$$\text{Increase in confidence}(\%) = \sum_{i=1}^{N}\frac{1_{\{Y_i^c < O_i^c\}}}{N} \times 100. \tag{11}$$

We also explicitly compare the target class probabilities of the perturbed images obtained before and after applying EVET. Win (%) is defined by the ratio of the images where the target class probability of one method is higher than that of the other.

Table 1 shows that EVET achieves improved results in all cases; applying EVET lowers average drop, elevates increase in confidence and target class probabilities (higher win). This demonstrates that EVET enhances explanation methods in terms of more accurately finding out the most distinguishable region of the target class.

Localization evaluation The energy-based pointing game aims to quantify the degree in which the importance map is focused on the target object. It measures how much energy of the importance map falls into the target object as follows:

$$E_{point}(x, J) = \frac{\sum_{(r,c)\in object} J(x)_{(r,c)}}{\sum_{(r,c)\in\{1,\dots,R\}\times\{1,\dots,C\}} J(x)_{(r,c)}} \tag{12}$$

Instead of using the bounding box [28], we use segmentation maps of VOC and COCO for more exact evaluation. The results are reported in Table 2. It indicates that applying EVET increases the proportion of the importance map belonging to the target object for every tested case.

4.3. Computational Efficiency

In order to illustrate that EVET can be used to achieve desirable performance with a low computational cost, we provide comparative evaluation of the EVE-applied Grad-CAM [19], denoted by Grad-CAM*, and Score-CAM. Note that here we consider the original version of Score-CAM for a fair comparison because Faster-Score-CAM might be degraded due to not utilizing the entire activation maps. It was shown that Score-CAM outperforms other perturbation-based and activation-based methods in terms of faithfulness by large scale [28]. However, Score-CAM is more computationally expensive than Grad-CAM because it requires forward propagations as many as the number of the activation maps N_A while Grad-CAM requires only a single backward pass.

For instance, VGG-16 entails 15.3 billion floating-point operations (FLOPs) for a single forward pass [2] and $N_A = 512$; the last convolutional layer is chosen as recommended in [28]. Accordingly, Score-CAM needs about $15.3\times512 =$

Metric		Gradeint		Grad-CAM		Grad-CAM++		Score-CAM	
		VOC	COCO	VOC	COCO	VOC	COCO	VOC	COCO
Energy-based	Original	28.0	27.4	47.5	44.9	35.5	33.6	34.8	31.7
pointing game (%)	w/ EVET	30.9	30.5	50.0	47.2	39.9	36.2	38.4	33.8

Table 2: Localization evaluation of EVET on VOC and COCO; higher is better.

Score-CAM vs	Grad-CAM			Grad-CAM*		
	VOC	COCO	ImageNet	VOC	COCO	ImageNet
Average difference	-0.0640	-0.0436	-0.0731	-0.0191	-0.0183	-0.0089
Win (%)	49.0	50.6	37.6	47.8	**52.6**	**50.4**

Table 3: Comparative evaluation of EVET for Grad-CAM on the target class probabilities versus Score-CAM; higher is better for average difference and win. Grad-CAM* represents EVET-applied Grad-CAM.

	VOC	COCO	ImageNet
Grad-CAM*	0.130 ± 0.030	0.134 ± 0.023	0.124 ± 0.007
Score-CAM	4.755 ± 0.474	4.811 ± 0.514	2.045 ± 0.020

Table 4: Execution time of Grad-CAM* and Score-CAM (seconds) where Grad-CAM* represents EVET-applied Grad-CAM.

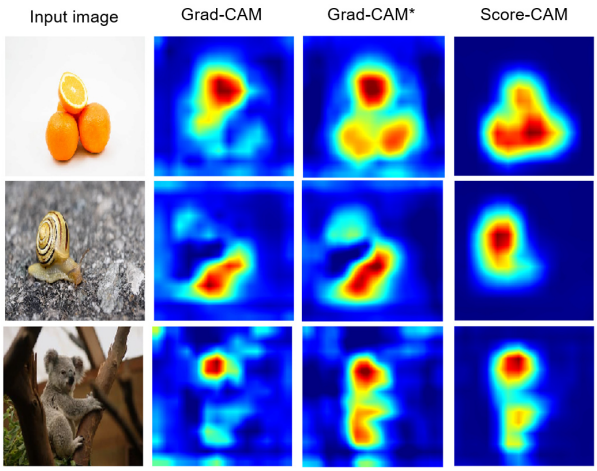


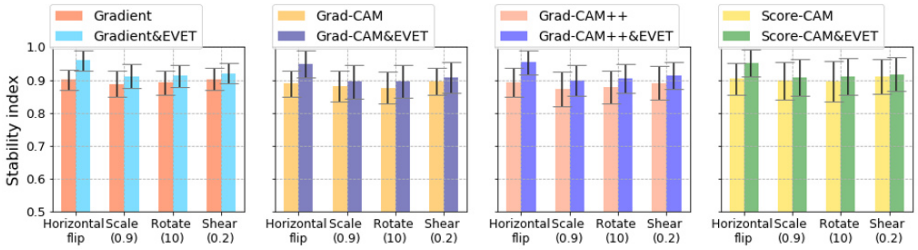
Figure 3: Qualitative comparison of Grad-CAM, Grad-CAM* and Score-CAM. A VGG-16 network trained on ImageNet is used.

7833.6 G-FLOPs. On the other hand, Grad-CAM* involves (number of transformations + 1)×(FLOPs for image transformations and a backward pass). Considering the case of dense and convolutional layer, we assume that the backward pass is twice as compute-intensive as the forward pass. Since the cost of image transformations is negligible compared to the backward pass, Grad-CAM* has roughly $8 \times 15.3 \times 2 = 244.8$ G-FLOPs. Therefore, GRAD-CAM* requires approximately $7833.6/244.8 \simeq 32.0$ times less computations compared to Score-CAM.

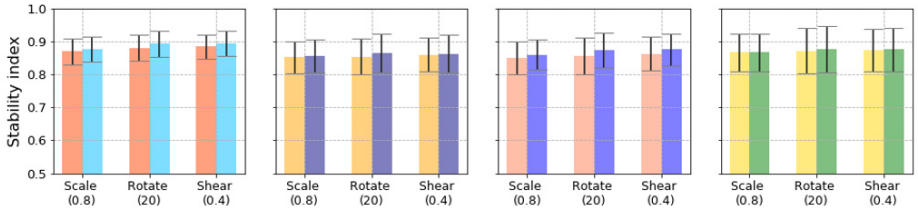
Applying EVET to Grad-CAM, the resulting importance maps appear to be closer to those of Score-CAM (see Fig. 3). For a concrete comparison, we calculate the average difference of the target class probabilities and the ratio of images where Grad-CAM* has higher target class probability. We then measure the execution time of those methods to compare computational efficiency. The results are reported in Table 3 and 4, respectively. It can be seen that Grad-CAM* is comparable to Score-CAM on VOC, COCO, and ImageNet with much shorter execution time (reduced by 97.3 %, 97.2 %, and 94.0% , respectively).

4.4. Stability Evaluation with Respect to Image Transformations

Using the proposed SI, we examine the image transformation robustness of explanations generated by existing explanation methods [21, 19, 4, 28, 8, 7] (see Table 5). It shows that activation-based methods have high SI in general while perturbation-based methods do not. We interpret this result as follows. Since perturbation-based methods go through many iterations to converge to the optimal solution, the result can be sensitively changed to the variation of the



(a) Stability index for transformations used in EVET.



(b) Stability index for transformations not used in EVET.

Figure 4: Stability evaluation of EVET on ImageNet. We compare SIs before and after applying EVET with respect to both the transformations used in EVET (a) and those not (b).

Method	SI
Gradient	0.879
Grad-CAM	0.909
Grad-CAM++	0.907
Score-CAM	0.891
Mask [8]	0.754
Extremal perturbation [7]	0.863

Table 5: Stability evaluation of explanation methods on VOC. The average of the SI calculated for scale (by 0.9), rotate (by 10°), shear (by 0.2), and horizontal flip is used.

input image. Meanwhile, Extremal perturbation [7] which resolves the balancing issues in [8] by regularizing perturbations has higher SI compared to Mask [8].

We investigate SI to quantify how much performance is improved by applying EVET in terms of stability. In the following, we calculate SI for transformations used in EVET to derive the final importance map and transformations that are not. Considering that EVET deems the parts estimated important consistently in the inverted importance maps more influential, it is reasonable to assume that EVET increases SI compared to the original explanation method for the former case. In fact, it turns out that applying EVET leads to an evident increase in SI (see Fig. 4a). This trend is observed in common for VOC and COCO (refer to Section D. of the supplementary materials).

Moreover, it is worthwhile to note that SI also increases for transformations other than those used in EVET (see Fig. 4b). It suggests that EVET makes explanation methods better at finding parts which are consistently important in the transformed images over a range of transformations rather than just those used in EVET.

5. Conclusion and Future Research

In this paper, we proposed a new method which can be applied to enhance existing explanation methods. We conjecture that the parts of the input image estimated to be important consistently in the transformed images are essential to the model's prediction as long as the prediction is robust to the transformations. EVET uses importance maps from the transformed images to yield the final importance map by considering their importance and stability. We highlight that EVET is easily applicable to various types of explanation methods without violating the criterion for estimating the importance. The effectiveness of EVET is validated both qualitatively and quantitatively. In particular, it is shown that EVET-applied explanation methods return more stable and trustworthy results compared to existing explanation methods. Future work includes considering different ways of modifying the input image other than geometric transformations and alternative methods of combining inverted importance maps.

References

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

[2] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep

neural network architectures. *IEEE Access*, 6:6427064277, 2018.

[3] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *arXiv preprint*, arXiv:1807.08024, 2018.

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

[5] S. Desai and H. G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 972–980, 2020.

[6] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 06 2010.

[7] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2950–2958, Oct 2019.

[8] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3429–3437, Oct 2017.

[9] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10697–10706, 2019.

[10] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang. Visual attention consistency under image transforms for multi-label image classification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–739, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] K. Li, Z. Wu, K. Peng, J. Ernst, and Y. Fu. Tell me where to look: Guided attention inference network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9215–9223, 2018.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *Lecture Notes in Computer Science*, page 740755, 2014.

[14] Grgoire Montavon, Wojciech Samek, and Klaus-Robert Mller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:115, Feb 2018.

[15] Badri Patro, Anupriy, and Vinay Namboodiri. Explanation vs attention: A two-player game to obtain attention for vqa. volume 34, pages 11848–11855, 04 2020.

[16] B. Patro, M. Lunayach, S. Patel, and V. Namboodiri. U-cam: Visual explanation using uncertainty based class activation maps. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7443–7452, 2019.

[17] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. *CoRR*, abs/1806.07421, 2018.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211252, Dec. 2015.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[20] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2591–2600, 2019.

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

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv: 1409.1556, 09 2014.

[23] D. Smilkov, N. Thorat, B. Kim, F. Vigas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

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

[25] Brian J Taylor. *Methods and procedures for the verification and validation of artificial neural networks*. Springer Science & Business Media, 2006.

[26] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9097–9107, June 2019.

[27] Haofan Wang, Rakshit Naidu, Joy Michael, and Soumya Snigdha Kundu. Ss-cam: Smoothed score-cam for sharper visual feature localization. *arXiv preprint*, arXiv: 2006.14255, 2020.

[28] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 24–25, June 2020.

[29] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.