

# Information-Theoretic Visual Explanation for Black-Box Classifiers

Jihun Yi, Eunji Kim, Siwon Kim, and Sungroh Yoon, *Senior Member, IEEE*

arXiv:2009.11150v1 [cs.CV] 23 Sep 2020

**Abstract**—In this work, we attempt to explain the prediction of any black-box classifier from an information-theoretic perspective. For this purpose, we propose two attribution maps: an information gain (IG) map and a point-wise mutual information (PMI) map. IG map provides a class-independent answer to “How informative is each pixel?”, and PMI map offers a class-specific explanation by answering “How much does each pixel support a specific class?” In this manner, we propose (i) a theory-backed attribution method. The attribution (ii) provides both supporting and opposing explanations for each class and (iii) pinpoints most decisive parts in the image, not just the relevant objects. In addition, the method (iv) offers a complementary class-independent explanation. Lastly, the algorithmic enhancement in our method (v) improves faithfulness of the explanation in terms of a quantitative evaluation metric. We showed the five strengths of our method through various experiments on the ImageNet dataset. The code of the proposed method is available online<sup>1</sup>.

**Index Terms**—Attribution map, deep learning, explainable AI, interpretation, XAI

## I. INTRODUCTION

IMAGE classifiers using deep neural networks show a high level of performance [1], [2], but they lack human interpretability; they provide accurate classification results but the reasoning behind them is not accessible. Such a black-box property hinders their prediction from convincing human users, and thus making a neural network interpretable is a prerequisite for high-stakes applications.

A popular approach for explaining the prediction of an image classifier is to generate an *attribution* map [3], [4], which is a quantified contribution of each input pixel to the prediction. Overlapping the map with an input image offers an intuitive and easily interpretable visualization that highlights a salient region of the image. Previous studies have primarily analyzed a classifier as a deterministic function. By contrast, we regard a prediction as a conditional distribution of a label and define the attribution from an information-theoretic perspective.

To obtain an attribution map, a group of recent approaches called *perturbation-based methods* [5], [6] is used to measure the amount of change in prediction when each pixel is removed individually. If removing a pixel  $x_i$  from an input image  $\mathbf{x}$  results in a significant change in the classifier output, it can

J. Yi, E. Kim, S. Kim, and S. Yoon are with Seoul National University, South Korea.

S. Yoon is also with ASRI, INMC, ISRC, and Institute of Engineering Research, Seoul National University, Seoul 08826, Korea.

Correspondence should be addressed to S. Yoon. (sryoon@snu.ac.kr).

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

<sup>1</sup><https://github.com/nuclearboy95/XAI-Information-Theoretic-Explanation>

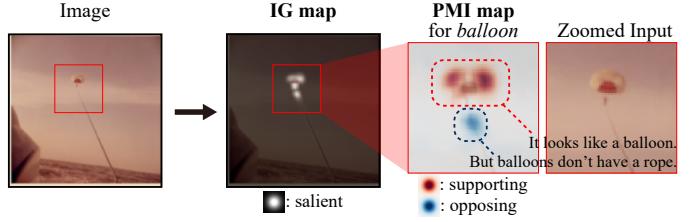


Fig. 1. The proposed IG and PMI maps. VGG19 [2] classified the parachute image as a *balloon* class. The **IG map** highlights salient regions (bright) for the classifier, and the **PMI map** indicates evidences for (red) and against (blue) the class from the viewpoint of the classifier. The orange fabric of the parachute is identified as a supporting evidence because it appears like a part of a balloon. However, the rope is highlighted as an opposing evidence because balloons don’t have a rope.

be concluded that  $x_i$  plays an important role in the classifier prediction  $p_\theta(Y | \mathbf{x})$  for a label  $Y$ . However, typical classifiers cannot make a prediction for a partially removed input,  $\mathbf{x}_{\setminus i}$ . Therefore, a prediction without  $x_i$ ,  $p_\theta(Y | \mathbf{x}_{\setminus i})$  should be calculated through marginalization as (1), where multiple values of  $\tilde{x}_i$  are sampled from the distribution of the pixel ( $X_i$  is a random variable corresponding to the pixel  $x_i$ ). Typically, generative models are used to model the pixel distribution.

$$p_\theta(Y | \mathbf{x}_{\setminus i}) = \mathbb{E}_{\tilde{x}_i \sim p(X_i | \mathbf{x}_{\setminus i})} [p_\theta(Y | \mathbf{x}_{\setminus i}, \tilde{x}_i)]. \quad (1)$$

Some previous methods have approximated  $p_\theta(Y | \mathbf{x}_{\setminus i})$  by heuristically simulating the pixel removal instead of marginalization. Pixels were replaced by gray [5], black [7], or blurred images [8], [9]. However, these techniques push the image out of the training distribution of the classifier, thereby resulting in an ill-defined behavior. Chang *et al.* [10] partially addressed the out-of-distribution (OoD) problem by replacing pixels using CA-GAN [11], which generates seemingly natural images. Nonetheless, their approach remains far from marginalization because CA-GAN in-fills the pixels deterministically without sampling. For true marginalization, we sample multiple pixel values using a powerful generative model while maintaining the perturbed image within the data distribution.

Followed by the marginalization, we make use of two information-theoretic measures to quantify the attribution. As a result, we herein propose two attribution maps: an information gain (**IG map**) and a point-wise mutual information (**PMI map**), which are class-independent and class-specific explanations, respectively. Fig. 1 presents their examples. The attributions are designed (i) to have a clear theoretic implication, (ii) to provide both supporting and opposing evidences to a class, and (iii) to indicate truly decisive parts in the image rather than whole relevant areas. The attributions include (iv)

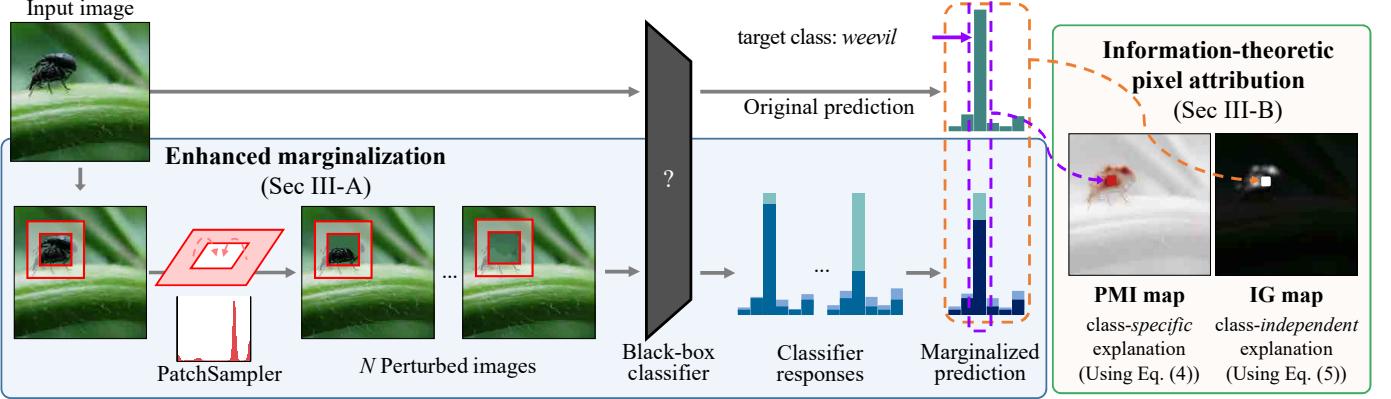


Fig. 2. **Overall flow of the proposed attribution method.** For every patch in an image, we compute a marginalized prediction (expected prediction for the image without that patch). The difference between the marginalized and original prediction is measured into PMI and IG maps. IG map indicates the *overall* salient area, and PMI map suggests the region for and against a *specific class*.

a class-independent explanation that is still convincing even when the classifier makes an unconfident prediction. Compared with the previous methods, we implement a more accurate marginalization by employing more powerful generative model. As a consequence, we (v) improved *faithfulness* [12] of the attribution map in terms of a quantitative evaluation metric.

## II. BACKGROUND

### A. Attribution method

Given an input image  $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ , the classifier predicts its label  $y$  ( $H$ ,  $W$ , and  $C$  are the height, width, and the number of channels in the image). Regarding the unknown label  $Y$  as a random variable, the classifier predicts it by calculating a posterior distribution,  $p_\theta(Y | \mathbf{x})$ . For a specific class  $y_c$  (typically the classifier's top prediction), to provide a human-interpretable reasoning for the prediction  $p_\theta(y_c | \mathbf{x})$ , the *attribution method* generates an attribution map,  $\mathbf{m} \in \mathbb{R}^{H \times W}$ . An element of  $\mathbf{m}$ ,  $m_i$  quantifies the contribution of the pixel  $x_i$  to the prediction; however, its definition and range vary among the different methods.

Starting from the Saliency map [13], the backprop-based methods [14], [15], [16] including Integrated Gradients [4] and Grad-CAM [17] have been suggested to generate an attribution map using backpropagation. Although they are sufficiently fast to facilitate realtime results, Dabkowski *et al.* [18] argued that their quality is limited and suggested the generation of a mask-like attribution map. Following the work by Dabkowski *et al.* [18], the mask-based methods including FIDO [10], Meaningful Perturbation [9], Extremal Perturbation [8], and RISE [7] aim to find a mask that covers a relevant region of the target class. They apply an image perturbation with a mask and optimize the mask with their unique objective functions. More details regarding the backprop-based and mask-based methods are provided in the Appendix.

1) *Perturbation-based method:* Since the work by Zeiler *et al.* [5], perturbation-based methods have been developed to compute  $p_\theta(y_c | \mathbf{x}_{\setminus i})$  for each pixel  $x_i$  and define an attribution as its discrepancy from the original prediction,  $p_\theta(y_c | \mathbf{x})$ . Therefore, a perturbation-based method consists of two steps:

Step 1: Calculate  $p_\theta(y_c | \mathbf{x}_{\setminus i})$

Step 2: Calculate  $m_i = \text{Diff}(p_\theta(y_c | \mathbf{x}), p_\theta(y_c | \mathbf{x}_{\setminus i}))$

Here, the  $\text{Diff}(\cdot)$  function measures the difference between two predictions. In step 1, typical classifiers cannot make a prediction for a partially unknown input, and thus a direct computation of  $p_\theta(y_c | \mathbf{x}_{\setminus i})$  is infeasible. One solution is to approximate it by substituting  $x_i$  for a reference value  $x_{\text{ref}}$ :

$$p_\theta(y_c | \mathbf{x}_{\setminus i}) \approx p_\theta(y_c | \mathbf{x}_{\setminus i}, x_{\text{ref}}). \quad (2)$$

The occlusion method [5] uses a gray value for  $x_{\text{ref}}$ . Although such a heuristic substitution has been widely used [18], [9], [3], selecting the reference value is non-trivial. This is because unlike the MNIST dataset, the background of which is black, many datasets (*e.g.* ImageNet) have no uninformative color; a black pixel indicates a black color, not “no information.” In such cases, a replacement with  $x_{\text{ref}}$  introduces a new flaw, instead of removing a pixel.

Therefore, Prediction Difference Analysis (PDA) [6] marginalizes each pixel to compute  $p_\theta(y_c | \mathbf{x}_{\setminus i})$ , as expressed in (1). They assumed a pixel,  $x_i$ , to follow a Gaussian distribution conditioned on its surrounding. The expectation is further approximated using a Monte Carlo (MC) sampling of sample number  $N = 10$ , thereby reducing the computational overhead. However, we show that conditional Gaussian does not model the pixel distribution accurately. In this work, we enhance the distribution modeling and hence improve the quality of the attribution maps.

Following various approaches in step 1, the choice of the  $\text{Diff}(\cdot)$  function differs among the methods. The occlusion method [5] uses  $\text{Diff} = p_\theta(y_c | \mathbf{x}) - p_\theta(y_c | \mathbf{x}_{\setminus i})$ , which is a direct subtraction of two probabilities. PDA [6] uses a weight of evidence:  $\text{Diff} = \text{logodds}(p_\theta(y_c | \mathbf{x})) - \text{logodds}(p_\theta(y_c | \mathbf{x}_{\setminus i}))$ , where  $\text{logodds}(p) = \log(p/(1-p))$ . We utilize point-wise mutual information and information gain, and their definitions and strengths are described in the following sections.

### B. Information-theoretic analysis

Given random variables  $X$  and  $Y$  and their realizations  $x$  and  $y$ , information theory discloses how they are interconnected. Specifically, point-wise mutual information (PMI) between two events is a measure of how much one event provokes the

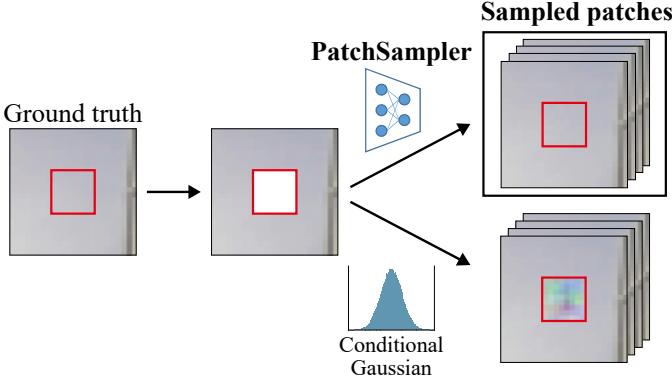


Fig. 3. **Enhanced distribution modeling.** PatchSampler models the distribution of patches more accurately than Gaussian. As a result, it prevents the OoD sample generation and thus facilitates more precise marginalization.

other:  $\text{PMI}(x; y) = \log(p(y|x)/p(y))$ . PMI has a positive or negative value when one event triggers or suppresses the other event, respectively. Therefore, by investigating the sign and magnitude of PMI, we can infer if an observation of  $x$  is an evidence for (positive) or against (negative)  $y$ .

We now consider an event  $\{X = x\}$  and a random variable  $Y$ , not a single event  $\{Y = y\}$ . If the observation of  $x$  brings about a considerable change in the probability distribution of  $Y$ , the observation is considered informative. Therefore, we quantify the information as Kullbeck—Leibler (KL) divergence between the posterior and prior distributions of  $Y$  and call it information gain (IG):  $\text{IG}(Y, x) = D_{\text{KL}}(p(Y|x) \| p(Y))$ . IG is minimized at zero when observing  $x$  provides no information about  $Y$ , while a large IG value denotes a significant amount of *Shannon's information*. Therefore, by investigating the magnitude of IG, we can estimate the informativeness of an event.

We utilize these two information-theoretic measures, PMI and IG, to quantify the contribution of each pixel.

### III. METHOD

In this study, we propose a model-agnostic *perturbation-based* method. The method visually explains the prediction of any black-box classifier by using two building blocks: enhanced marginalization (Sec III-A) and information-theoretic pixel attribution (Sec III-B). Fig. 2 shows an overview of the proposed method.

#### A. Enhanced marginalization

As shown in (1), marginalization requires a generative model  $p(X_i|\mathbf{x}_{\setminus i})$  from which a pixel is sampled. If the generative model does not model the distribution correctly, it makes the perturbed images ( $[\mathbf{x}_{\setminus i}, \tilde{x}_i]$  in (1)) OoD. Previous works [19], [20] discovered that the predicted probability of a classifier rapidly drops for the OoD images. As a consequence, uninformative regions (e.g., background) can receive high attributions because the perturbations are unrealistic, not because the regions are relevant to the class. Therefore, a faithful attribution map needs an accurate generative model.

Zintgraf *et al.* [6] used conditional Gaussian, but it does not accurately model the pixel distribution, and its samples

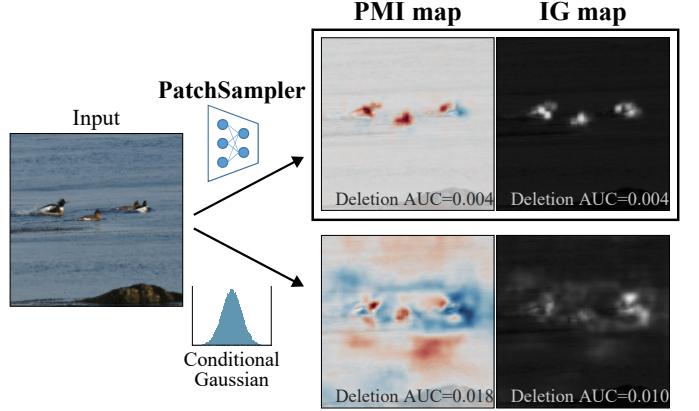


Fig. 4. **Enhanced marginalization improves attribution maps.** The enhanced marginalization by PatchSampler improves the quality of attribution maps. Besides the visual quality, a quantitative evaluation using Deletion AUC [7] (lower the better, see Sec IV-B) confirms the improvement.

are clearly OoD as shown in the bottom of Fig. 3. To model the complex distribution of pixels, we deploy a more powerful generative model consisting of a neural network and name it “PatchSampler”. PatchSampler models the distribution more accurately (Fig. 3 top), and thus it dramatically improves the quality of the attribution map, as depicted in Fig. 4.

Since a pixel is strongly dependent on its surroundings, PatchSampler approximates  $p(X_i|\mathbf{x}_{\setminus i}) \approx p(X_i|\hat{\mathbf{x}}_{\setminus i})$ , where  $\hat{\mathbf{x}}_{\setminus i}$  is the neighborhood of the pixel  $x_i$ .  $p(X_i|\hat{\mathbf{x}}_{\setminus i})$  is then modeled with  $p_\phi(X_i|\hat{\mathbf{x}}_{\setminus i})$  using a neural network comprising a series of convolutional layers with parameter  $\phi$ . We trained it in advance using the training data of the classifier. As the PatchSampler has been trained using the ImageNet training set, which is a large general image dataset, the trained PatchSampler is expected to applicable to classifiers for the other natural image datasets.

Evaluating (1) requires a classifier feedforward  $p_\theta(y_c|\mathbf{x}_{\setminus i}, x_i)$  for all possible values of  $x_i$ , which demands a large number of computations. We mitigate this computing issue using the MC approximation of sample number  $N$ :

$$p_\theta(y_c|\mathbf{x}_{\setminus i}) \approx \frac{1}{N} \sum_{\tilde{x}_i \sim p_\phi}^N p_\theta(y_c|\mathbf{x}_{\setminus i}, \tilde{x}_i). \quad (3)$$

Although a larger  $N$  yields better approximation, we empirically show that a small  $N$  provides visually and numerically indistinguishable results in Fig. 23 in the Appendix. In the perturbation-based method, a pixel  $x_i$  obtains a high attribution when the corresponding  $p_\theta(y_c|\mathbf{x}_{\setminus i})$  is far from  $p_\theta(y_c|\mathbf{x})$ . When the probable values of  $x_i$  are not sufficiently diverse, the two probabilities are close to each other. Moreover, when  $x_i$  is irrelevant to class  $y_c$ , observing  $x_i$  would not lead to any difference between the posterior and prior. Therefore, pixels receive high attribution when they are both *unpredictable* and *relevant* to  $y_c$  [6].

However, pixels are highly correlated with their neighborhood and extremely predictable. Alternatively, patches are more informative than pixels because they are more uncertain. Therefore, we conduct a multivariate analysis by calculating the attribution for patches instead of pixels. We

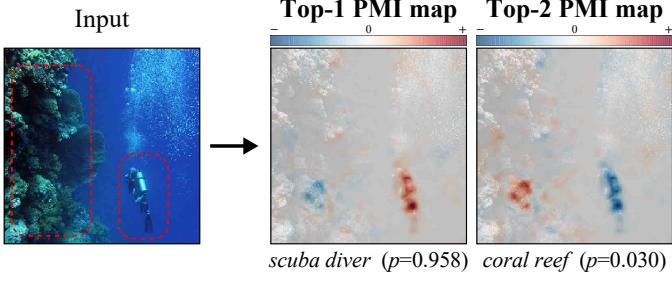


Fig. 5. **PMI maps for the top-1 (scuba diver) and top-2 (coral reef) predicted classes.** The scuba diver supports the *scuba diver* class, but is a contradicting evidence for the *coral reef*, and vice versa.

used a patch of size  $K \times K$  with a hyperparameter  $K$ . Marginalization was performed on every patch  $\mathbf{x}_r$  using  $p_\theta(y_c | \mathbf{x}_{\setminus r}) \approx \frac{1}{N} \sum_{\tilde{\mathbf{x}}_r} p_\theta(y_c | \mathbf{x}_{\setminus r}, \tilde{\mathbf{x}}_r)$ , where  $\tilde{\mathbf{x}}_r$  is sampled from a multivariate PatchSampler,  $p_\phi(\mathbf{X}_r | \tilde{\mathbf{x}}_r)$ . In our experiments, we used  $\tilde{\mathbf{x}}_r$ , a  $3K \times 3K$  patch centered at  $\mathbf{x}_r$ , and patch-wise calculated attributions were distributed to every pixel in each patch.

### B. Information-theoretic pixel attribution

Given an image  $\mathbf{x}$  and its (unknown) label, we consider their corresponding random variables,  $\mathbf{X}$  and  $Y$ . The classifier prediction  $p_\theta(Y | \mathbf{x})$  is a posterior distribution of  $Y$ . Starting from a prior distribution, the classifier makes a final prediction based on the information from the input observation. To quantify the contribution of each feature, measuring the amount of information provided by each feature is a natural choice. Accordingly, to explain the prediction, our algorithm outputs two attribution maps,  $\mathbf{m}_{\text{pmi}}$  (PMI map) and  $\mathbf{m}_{\text{ig}}$  (IG map)  $\in \mathbb{R}^{H \times W}$ , for two different questions: “How much does each pixel support a specific class  $y_c$ ?” and “How informative is each pixel?” Hereinafter, for a multivariate analysis, we consider the contribution of a patch  $\mathbf{x}_r$  instead of a pixel  $x_i$ .

1) *PMI map*: First, assume that we are interested in how much a patch  $\mathbf{x}_r$  accounts for a class  $y_c$ . In other words, given the other part of the image, how much does a patch observation  $\{\mathbf{X}_r = \mathbf{x}_r\}$  trigger the predicted event  $\{Y = y_c\}$ ? Such a notion is captured by PMI between the two events conditioned on  $\mathbf{x}_{\setminus r}$ . For a given patch, its PMI for a class  $y_c$  is computed as follows:

$$\text{PMI}(y_c; \mathbf{x}_r | \mathbf{x}_{\setminus r}) = \log \left( \frac{p_\theta(y_c | \mathbf{x}_{\setminus r}, \mathbf{x}_r)}{p_\theta(y_c | \mathbf{x}_{\setminus r})} \right). \quad (4)$$

It is noteworthy that the numerator in (4) is the prediction for the original input. We calculate the PMI for each patch and distribute it to the pixels in the patch such that the pixels receive the average attribution of the patches in which they belong. The calculated attribution map is called a **PMI map**:  $\mathbf{m}_{\text{pmi}}^i(y_c, \mathbf{x}) \stackrel{\text{def}}{=} \text{AVG}_{\mathbf{x}_r \ni x_i} [\text{PMI}(y_c; \mathbf{x}_r | \mathbf{x}_{\setminus r})]$ , where AVG indicates the average. Patches with positive PMI support  $y_c$ , whereas those with negative PMI oppose  $y_c$ . For example, the PMI maps in Fig. 5 suggests that the scuba diver supports the *scuba diver* class, but is a contradicting evidence for the rival class, *coral reef*, and vice versa.

Since a PMI map can be calculated for any class  $y_c$ , the rationales behind the prediction related to the less probable

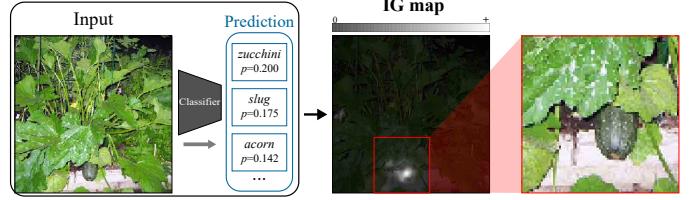


Fig. 6. **IG map highlights where the classifier mostly focused on.** VGG19 [2] failed to make a confident prediction. The IG map shows that the classifier was trying to classify the herb at the bottom.

classes can be obtained as well. Note that a supporting evidence of the top-1 class does not necessarily oppose the top-2 class, and vice versa (see Fig. 20). From the supporting and opposing evidences of two competing classes, viewers can easily understand the reason for the prediction for each class.

2) *IG map*: The importance of  $\mathbf{x}_r$  can be quantified using the IG between an observation  $\{\mathbf{X}_r = \mathbf{x}_r\}$  and  $Y$  given  $\mathbf{x}_{\setminus r}$ . Analogous to a PMI map, we compute the IG of each patch using (5) and distribute it to the pixels in the patch.

$$\begin{aligned} \text{IG}(Y, \mathbf{x}_r | \mathbf{x}_{\setminus r}) &= D_{\text{KL}}(p_\theta(Y | \mathbf{x}_{\setminus r}, \mathbf{x}_r) \| p_\theta(Y | \mathbf{x}_{\setminus r})) \\ &= \mathbb{E}_{y_c \sim p_\theta(Y | \mathbf{x})} \left[ \log \left( \frac{p_\theta(y_c | \mathbf{x}_{\setminus r}, \mathbf{x}_r)}{p_\theta(y_c | \mathbf{x}_{\setminus r})} \right) \right] \\ &= \mathbb{E}_{y_c \sim p_\theta(Y | \mathbf{x})} [\text{PMI}(y_c; \mathbf{x}_r | \mathbf{x}_{\setminus r})]. \end{aligned} \quad (5)$$

Likewise, the **IG map** is  $\mathbf{m}_{\text{ig}}^i(\mathbf{x}) \stackrel{\text{def}}{=} \text{AVG}_{\mathbf{x}_r \ni x_i} [\text{IG}(Y, \mathbf{x}_r | \mathbf{x}_{\setminus r})]$ . Notably, the IG is an expectation of the PMI among all possible  $y_c$  in (5). This allows the calculation of IG by obtaining the PMI for every class  $y_c$  and calculating the weighted sum with the predicted probability  $p_\theta(y_c | \mathbf{x})$ . Because a single feed-forward of a classifier calculates  $p_\theta(y_c | \mathbf{x})$  for every  $y_c$ , an IG map can be obtained at only a marginal cost given PMI maps. Patches with high IG attribution are informative to the classifier and thus salient. IG map is not class-specific and is complementary with a class-specific PMI map. When multiple classes are probable so that no class-specific explanations are sufficient to describe the behavior of the classifier, IG map offers an overall explanation, as provided in Fig. 6. Note that such class-independent explanation is not available in most existing methods.

In summary, our method generates two attribution maps, namely a PMI map and an IG map, which provide class-specific and class-independent explanations, respectively. The pseudo-code of our method is provided in the Appendix.

3) *Comparison to PDA* [6]: Zintgraf *et al.* [6] quantified the contribution of each pixel using weight of evidence (WoE). The authors viewed a prediction as a hypothesis, and measured the extent to which an event  $\{X_i = x_i\}$  makes the **hypothesis**  $H : Y = y_c$  more likely. By contrast, we consider a prediction as a probability distribution of a random variable  $Y$ , and PMI quantifies the amount of *Shannon's information* between an event  $\{X_i = x_i\}$  and another **event**  $\{Y = y_c\}$ . Considering  $Y$  as a **random variable**, information gain of  $Y$  from the event  $\{X_i = x_i\}$  can be further calculated, thereby identifying informative features regardless of a class.

4) *Comparison to Learning to Explain* [21]: Inspired from information-theoretic perspective, Learning to Explain (L2X)

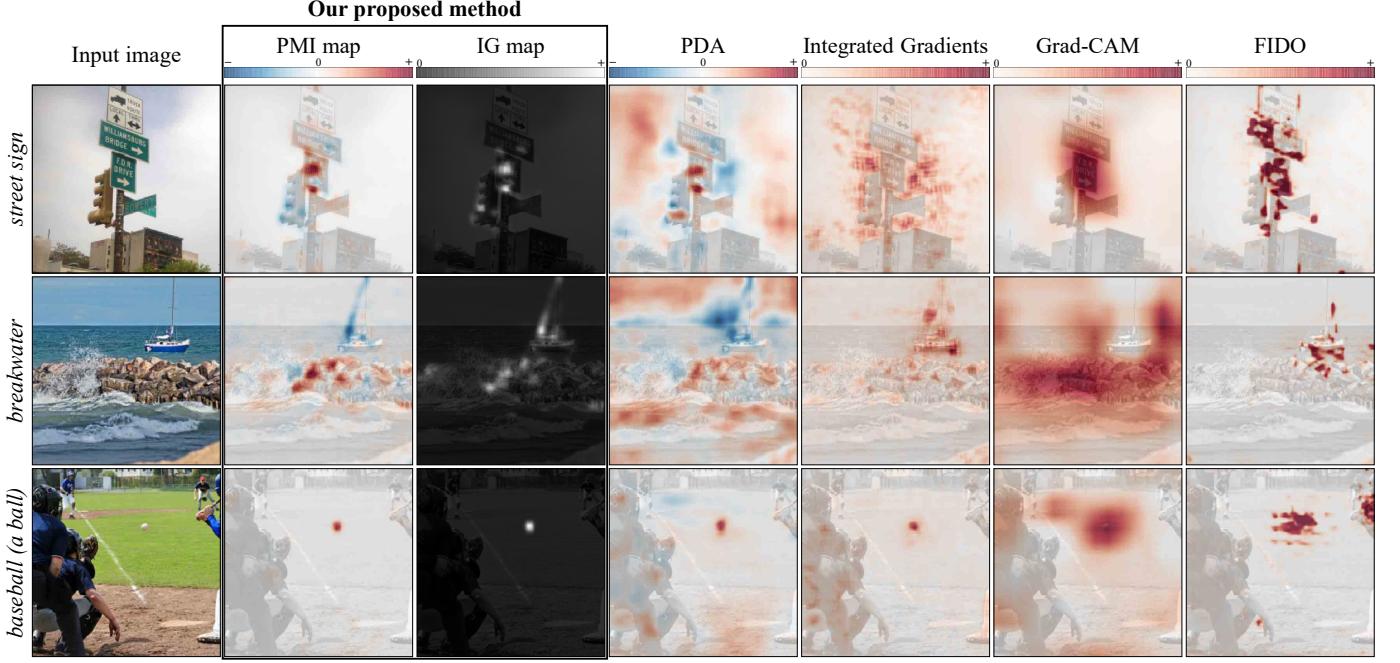


Fig. 7. **Explanations for VGG19 [2] predictions.** We compare PMI and IG maps with the other attribution maps generated by popular methods in each category (PDA [6], Integrated Gradients [4], Grad-CAM [17], and FIDO [10]). Integrated Gradients maps were smoothed using a flat kernel of size  $K = 8$  (the same  $K$  used in PDA, PMI and IG maps) due to their low visibility.

tries to explain a prediction by selecting a set of informative input features. L2X trains a selector model that predicts a set of features with high variational lower bound of mutual information. Unlike PDA [6] and our method, L2X does not consider the distribution of a feature and makes use of a variational approximation with a heuristic black replacement. For this reason, MNIST was the only image dataset that the authors demonstrated. Taking the pixel distribution into account, we attempt to explain a classifier trained with datasets of any kind. Moreover, we directly quantify the information of each feature so that our method can deduce a solution for their problem, but not vice versa.

#### IV. RESULTS AND DISCUSSION

First, we qualitatively compare our attribution method with the existing methods by providing visual assessment and the key differences (Sec IV-A). Next, the *faithfulness* [12] of our method is demonstrated through quantitative evaluations (Sec IV-B). Additional results from our method are presented afterwards. In the Appendix, more results regarding the hyperparameters and the implementation details are provided.

##### A. Qualitative comparison

1) *Visual assessment:* The street sign image in Fig. 7 (1st row) contains two objects: a street sign and a traffic light. The PMI map suggests that each object supports and denies the class *street sign*, respectively. It implies that both positive and negative evidences in the PMI map correctly capture the contributions of the objects. The IG map indicates that the two objects are informative, which coincides with the PMI map result. Note that the other methods except PDA [6], which is a perturbation-based method, only provide supporting evidences.

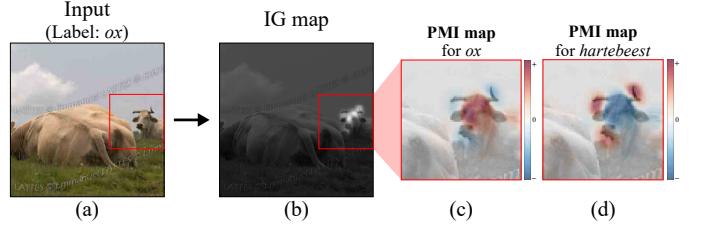


Fig. 8. **Negative evidence explains why the classifier rejected a class.** VGG19 [2] mis-classified the *ox* image (a) as a *hartebeest*. By observing its *ox*-like face, the image could be classified as an *ox*, but the horns rejected it, as described in (c). The PMI map for *hartebeest* (d) shows that the horns support the *hartebeest* class rather than the *ox*. This is because all *hartebeests* have horns in ImageNet training data, but not all *ox*.

They only highlight the street sign or show barely interpretable results.

Our method also pinpoints most decisive areas. In the baseball image (3rd row), Grad-CAM [17] and the mask-based method [10] tend to cover a large area around the object, and are less informative about which part was most distinctive. By contrast, our attribution maps indicated the most decisive object. Note that the *baseball* class is a ball, not a sport.

Compared with our method, PDA [6] assigns excessive attribution to irrelevant areas such as the background. This exemplifies the improved faithfulness attributable to the enhanced marginalization, as described in Sec III-A. This effect is quantitatively verified in Sec IV-B. Although PDA also provides negative evidence, its maps are far less informative due to relatively low visual quality.

2) *Negative evidence:* Explanation becomes more descriptive when it provides contradictory as well as advocating evidence. In Fig. 8, the PMI map explains the rationales behind the wrong prediction using the negative evidences.

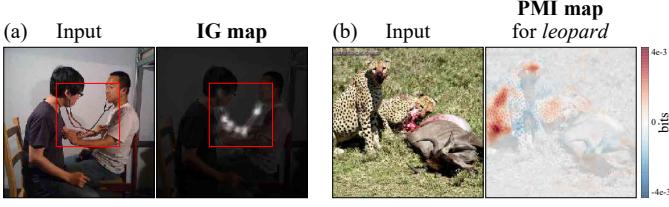


Fig. 9. IG map (a) highlights the salient area that the classifier focused on, and the attributions in PMI and IG maps are measured in bits (b).

Negative evidence facilitates more in-depth understanding about a classifier. However, all the mask-based and most backpropagation-based methods do not support such analysis because they only provide supporting regions.

To verify if the negative evidences truly have negative influence on the classifier prediction, we conducted the following experiment. Fig. 10 shows the average top-1 predicted probabilities for 4,500 ImageNet test images as portions of negative evidences are gradually removed. The result shows that the classifier gains confidence as the negative evidences are removed. This result indicates that the negative evidences provided by our method are indeed negative.

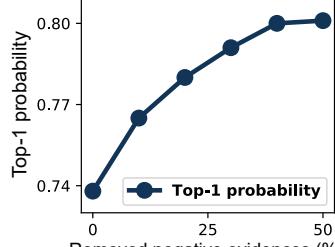


Fig. 10. Removing negative evidences makes the prediction more confident.

3) *Class-independent explanation*: Almost all existing methods [10], [18], [8], [4], [5], [6] only provide class-specific explanations. However, when multiple classes are probable, it is difficult to select an appropriate class-specific map to explain the prediction. In Fig. 9(a), VGG19 [2] predicted the image as a *bassoon* ( $p=0.08$ ), a *stethoscope* ( $p=0.07$ ), and a *horn* ( $p=0.07$ ) and failed to identify a dominant class. The IG map implies that the classifier focused on the stethoscope. Unlike the other methods, IG map can provide overall explanations about the region that the classifier made the decision based on.

4) *Implication of an attribution*: Mask-based methods [10], [18], [8], [9] propose *how to calculate* their own attribution, without a clear definition of *what* the attribution is. They quantify a relative importance of input features, but the value itself has no clear meaning. By contrast, the PMI and IG maps measure the actual amount of *Shannon's information* that each feature delivers in *bits* (Fig. 9(b)). Therefore, our method provides a theoretically meaningful yet intuitive explanation.

## B. Quantitative comparison

In this section, we examine a *faithfulness* [12] of our attribution method through a quantitative analysis. Faithful explanation is one that correctly explains the behavior of the classifier [12]. Deletion AUC [7], the metric to be used, measures the area under the prediction probability curve as pixels with high attributions are gradually removed. A low AUC implies a steep decrease in the prediction, thus indicating that the explanation correctly captures the relevant area for the predicted class. For example, given an image in Fig. 11(a), the PMI map for the *ringneck snake* class is provided in Fig. 11(b). Fig. 11(c) shows the predicted probability curve as the pixels

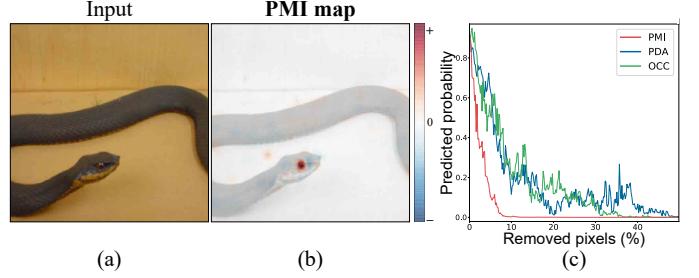


Fig. 11. An input image (a), the corresponding PMI map (b), and the deletion curves for the image (c). A sharp drop in the deletion curve (c) signifies a faithful attribution map.

with high scores are gradually removed. In this example, the deletion curve of the PMI map shows the steepest drop, and hence the PMI map provides the explanation with the highest faithfulness.

Meanwhile, each attribution map is an answer to a specific question related to the method. For example, Saliency map [13] answers “How much does the logit change as each pixel slightly changes?”, and FIDO [10] seeks to answer “What is the smallest input region sufficient to maintain the classification score?” Therefore, it is nonsense to directly compare the answers (attribution maps) of orthogonal questions using the same metric to test their faithfulness. Accordingly, quantitative evaluation can be meaningful when they are used to compare methods with similar implications (*i.e.*, methods in the same category). Thus, Deletion AUCs [7] are compared between maps from closely related methods (perturbation-based and mask-based methods) and are presented in Fig. 12.

The results confirmed the improved faithfulness of the PMI map over the perturbation-based baselines. Since the IG map is a class-independent attribution map, it does not necessarily yield a low AUC. Interestingly, the AUC for the IG map turned out to be quite low, and we postulate that this is because the regions that discriminate among the most probable classes deliver significant information to the classifier overall as well.

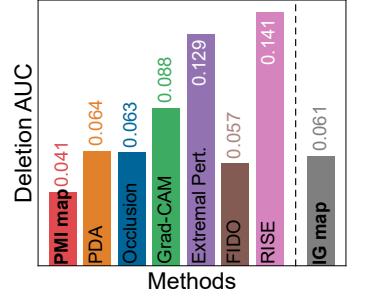


Fig. 12. Average Deletion AUCs [7] for 4,500 random ImageNet test images. The PMI map yields the lowest Deletion AUC.

## C. Additional remarks

1) *PatchSampler*: Accurate modeling of a patch distribution by a generative model is key to accurate marginalization. As presented in Fig. 14, PatchSampler models the distribution more correctly, compared to Gaussian and the heuristic replacement methods. The improved modeling accuracy enhanced the quality of attribution maps, as shown

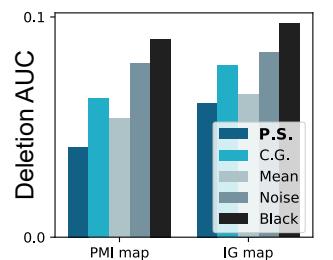


Fig. 13. Using PatchSampler (P.S.) as a generative model yields lower Deletion AUC [7] than using conditional Gaussian (C.G.) and the other heuristic replacements (replacement with a patch mean, noise, or black).

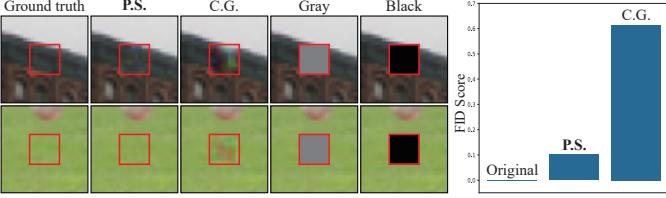


Fig. 14. **Samples from PatchSampler (P.S.) are more realistic.** Samples (left) from the other methods clearly indicate the OoD problem. Perturbed images by PatchSampler yield much lower FID score [23] (right) than conditional Gaussian (C.G.) from [6].

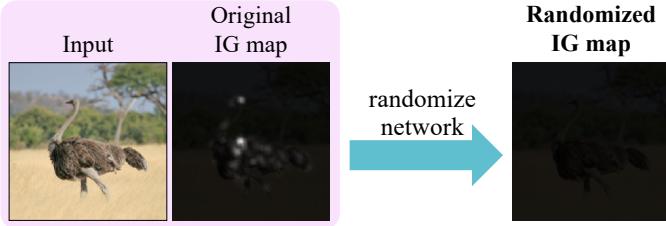


Fig. 15. **Sanity Checks** [22]. The parameters of VGG19 [2] are successively randomized from logits to conv1 layer. In the end, the IG map indicates that no pixel delivers any information to the classifier.

in Fig. 4. They are not only visually clearer, but also more correct. Fig. 13 shows the average Deletion AUCs [7] of the PMI and IG maps using PatchSampler or the existing methods. The result shows that the deployment of PatchSampler quantitatively improves the correctness of the attribution map.

2) *Sanity Checks:* Adebayo *et al.* [22] proposed two Sanity Checks that any valid attribution method should pass. They test if an attribution map is sensitive to both training data and the classifier parameters. The IG map showed a clear sensitivity to the classifier parameters, as in Fig. 15, and both PMI and IG maps passed the two Sanity Checks. More results are provided in Fig. 24 in the Appendix.

3) *Classifying animals:* ImageNet has a number of animal classes, and a well-trained classifier should capture the unique appearance of each animal class. In Fig. 16, we provided PMI maps of VGG19 [2] for several animal classes. The classifier successfully captured some distinctive features. For many animals, the facial areas are particularly discriminative than the other body parts.

4) *Biased data:* In Fig. 17 (top), the PMI maps report that the person riding a swing is more supporting feature of the class *swing* than the swing itself. This is because the majority of the training data for the class *swing* contains a swing with a person riding it, as shown in Fig. 17 (bottom). Thus, the classifier learned a strong correlation between the target class and person. For better generalization and to reduce unwanted bias towards data, examining a trained classifier using PMI maps will be beneficial.

5) *Classification based on texts:* In Fig. 19 (top), VGG19 [2] classified the test images as the *Band Aid* class, and the corresponding PMI maps suggest that the predictions are based on the red “BAND-AID” texts in the images. This observation can be attributed to the frequent occurrence of the similar texts in the training data, as shown in Fig. 19 (bottom). To verify if the text functions as a crit-

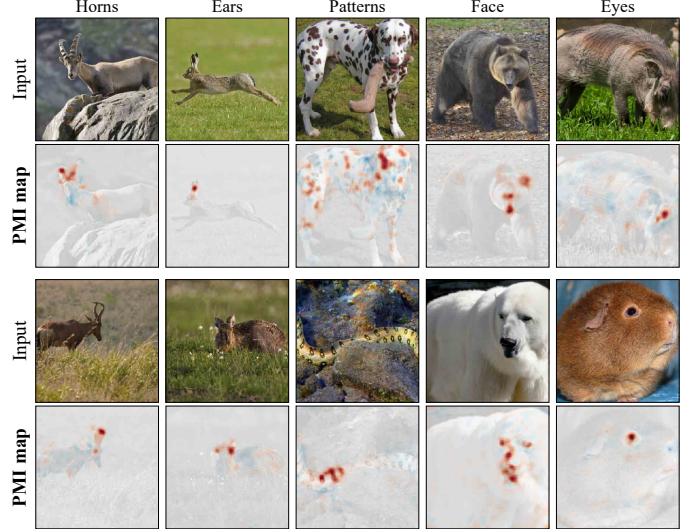


Fig. 16. **VGG19** [2] captures unique appearances. The classifier learned distinctive animal features such as horns (*ibex* and *impala*), ears (*hare*), and body patterns (*dalmatian* and *sea snake*).

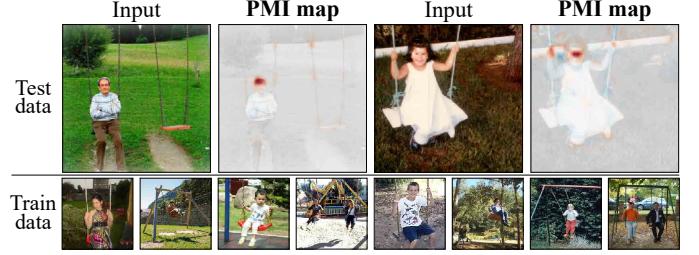


Fig. 17. **VGG19** [2] responds to a person. The majority of the training data for the *swing* class have a person riding on it. Such bias makes the classifier regard a person as a supporting evidence of the class.

ical cue for classification, we observe the predicted probability as the letters in the text are removed one-by-one. Fig. 18 shows that erasing a couple of letters yield a significant drop in the predicted probability. This result signifies that the classifier indeed responds to the text in the image.

6) *Top-1 and top-2 PMI maps:* PMI maps for the top-1 and top-2 predicted classes for some images are depicted in Fig. 20. When the subregions within a single object support different classes, PMI maps can identify the evidences for the respective classes (1st and 2nd rows). However, the supporting evidence of the top-1 class does not necessarily oppose the top-2 prediction, and vice versa. The evidences for the top-1 and top-2 classes can be analogous particularly when the two classes are closely related (3rd and 4th rows).

7) *Explaining different classifiers:* Since our method is model-agnostic, it can explain a classifier of any kind including non-differentiable one. The PMI maps for four popular image classifiers are presented in Fig. 21. The four classifiers made

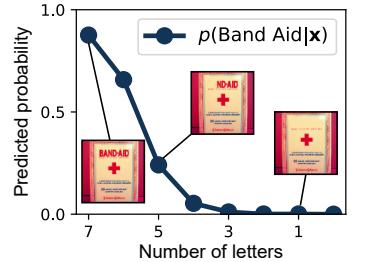


Fig. 18. Successive removal of letters in “BAND-AID” text drops the predicted probability of the *Band Aid* class.



Fig. 19. VGG19 [2] responds to texts. The PMI maps suggest that the images are classified as *Band Aid* class because of the “BAND-AID” texts in the images.

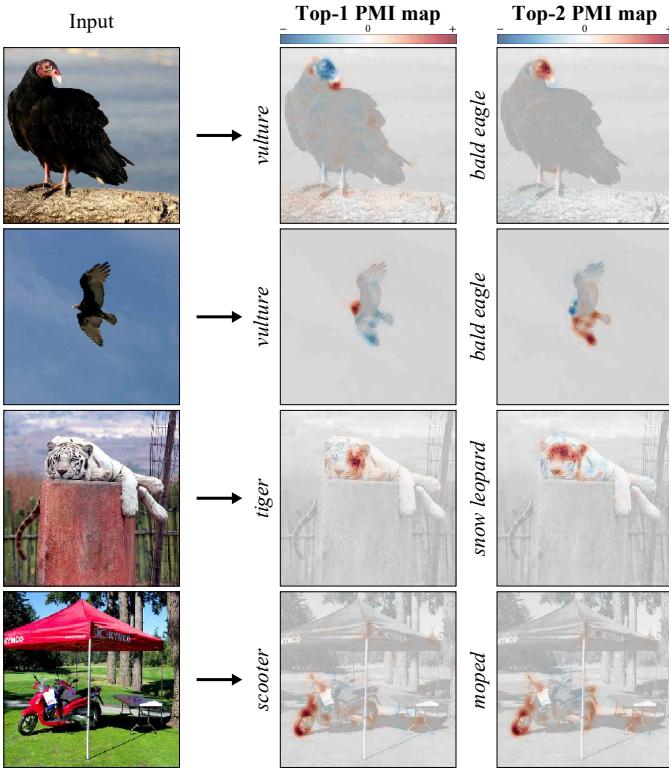


Fig. 20. Top-1 and top-2 PMI maps for VGG19 [2] predictions. The predicted top-1 and top-2 classes are presented next to each PMI map.

the same predictions for each image (provided at the left side of the input images), and the PMI maps are generated for those classes. For some images, the classifiers focused on the similar cues (1st row), while for the other images, they made decisions based upon different areas (2nd and 3rd rows).

8) *Adversarial robustness:* Ghorbani *et al.* [24] proposed two attack methods (iterative and random perturbation attacks) for feature-based interpretations. Iterative attack performs a gradient descent with respect to the input image in the direction that maximally changes the explanation, and the authors manipulated the explanation significantly while leaving the input image almost unchanged. However, gradient signals

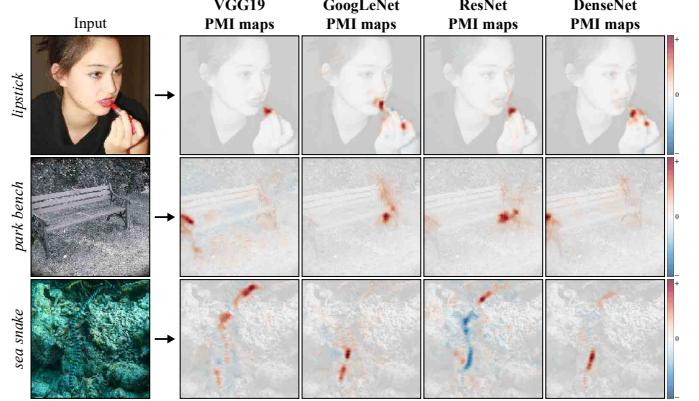


Fig. 21. Explaining different classifiers: VGG19 [2], GoogLeNet [25], ResNet50 [26], and DenseNet121 [27].

are unattainable with our method because the sampling node followed by PatchSampler blocks backpropagation. Only the random attack that provides a lower attack performance is feasible, and hence our method is more robust against adversarial attacks than backprop-based methods such as Saliency map [13] and its variants [3], [4].

## V. CONCLUSION

### A. Contributions

In this study, we proposed a new visual explanation method based on information theory, and showed its five advantages over the existing methods. We developed a novel approach in analyzing the input and its label as two random variables, thereby suggesting the use of theory-backed attribution methods, namely PMI and IG maps. The improved marginalization and the use of theory-backed attribution calculation schemes provided easily interpretable and strongly convincing attribution maps.

Ribeiro *et al.* [12] reported a trade-off between the interpretability and faithfulness of an explanation. The most faithful explanation (*i.e.*, the parameters themselves) lacks interpretability, and a more understandable explanation inevitably simplifies the classifier behavior, thus losing its faithfulness. Hence, refining an attribution map by averaging out the noise [15] and forcing attributions to cluster or suppress artifacts using a regularizer [10], [9] gains human-interpretability in exchange for faithfulness. By contrast, we performed no heuristic refinement for the visual appearance, yet giving easily interpretable results.

### B. Expandability

Although PMI and IG maps are defined and provided for image data herein, the notions of marginalizing the input feature and measuring PMI and IG are applicable to other domains such as language modeling [28] and tabular data [29] using appropriate generative models.

## APPENDIX A

### A. Pseudo-code

#### Algorithm 1 Information-Theoretic Pixel Attribution

```

1: Input image  $\mathbf{x}$ , classifier  $\theta$ , PatchSampler  $\phi$ , and class  $y_c$ 
2:  $\mathbf{m}_{\text{pmi}}, \mathbf{m}_{\text{ig}} \leftarrow \text{zeros}(H, W)$ 
3: Calculate  $p_\theta(y_c|\mathbf{x})$ 
4: for patch  $\mathbf{x}_r$  in  $\mathbf{x}$  do
5:    $\hat{\mathbf{x}}_{\setminus r} \leftarrow \text{Neighborhood}(\mathbf{x}_r)$ 
6:   Sample  $\tilde{\mathbf{x}}_r \sim p_\phi(\mathbf{X}_r|\hat{\mathbf{x}}_{\setminus r})$   $N$  times
7:   Calculate  $p_\theta(y_c|\mathbf{x}_{\setminus r})$  using Eq. (3)
8:   Calculate PMI and IG using Eq. (4) and Eq. (5)
9:   Distribute PMI and IG to each pixel  $x_i \in \mathbf{x}_r$  in  $\mathbf{m}_{\text{pmi}}$  and  $\mathbf{m}_{\text{ig}}$ 
10: end for
11: return  $\mathbf{m}_{\text{pmi}}$  and  $\mathbf{m}_{\text{ig}}$ 

```

### B. Effect of patch size

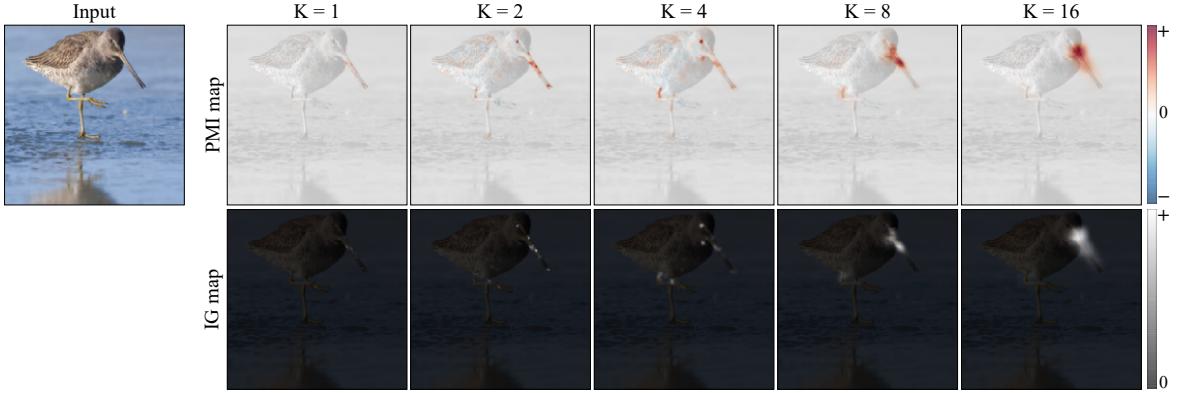


Fig. 22. **Attribution maps with varying patch size,  $K$ .** VGG19 [2] classified the image as a *dowitcher* with  $p = 0.999$ . As  $K$  increases, our method generates more interpretable attribution maps. By examining these maps with proper  $K$ , one can easily understand that the model made a decision upon the bird's beak and facial region.

### C. Effect of MC sample number

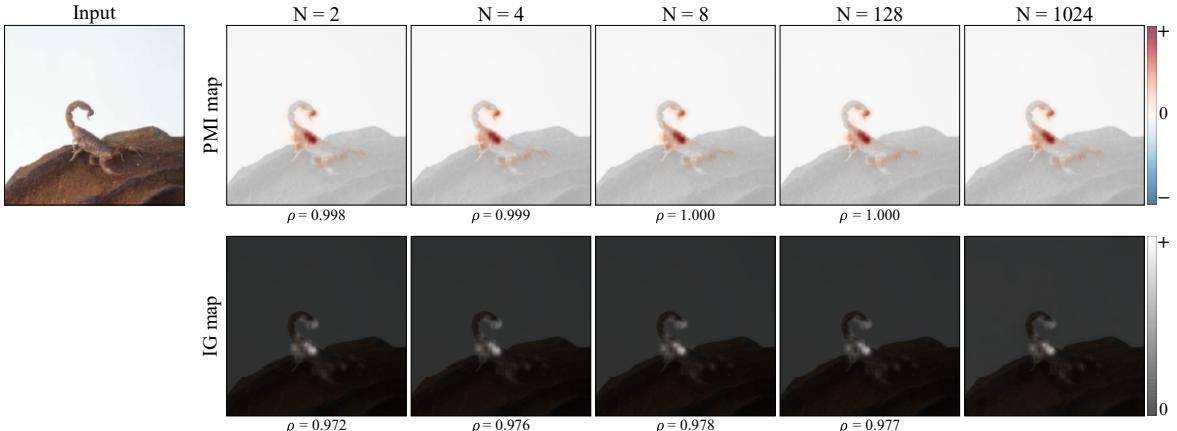


Fig. 23. **Attribution maps with varying MC sample number,  $N$ .** Attribution maps when  $N \leq 128$  are visually indistinguishable from maps generated using  $N = 1024$ . They are almost identical in terms of the Pearson correlation coefficient ( $\rho$ ) as well. Accordingly, we chose  $N = 8$  as a point of compromise between computational efficiency and approximation accuracy and used this value throughout the study.

#### D. Sanity Checks

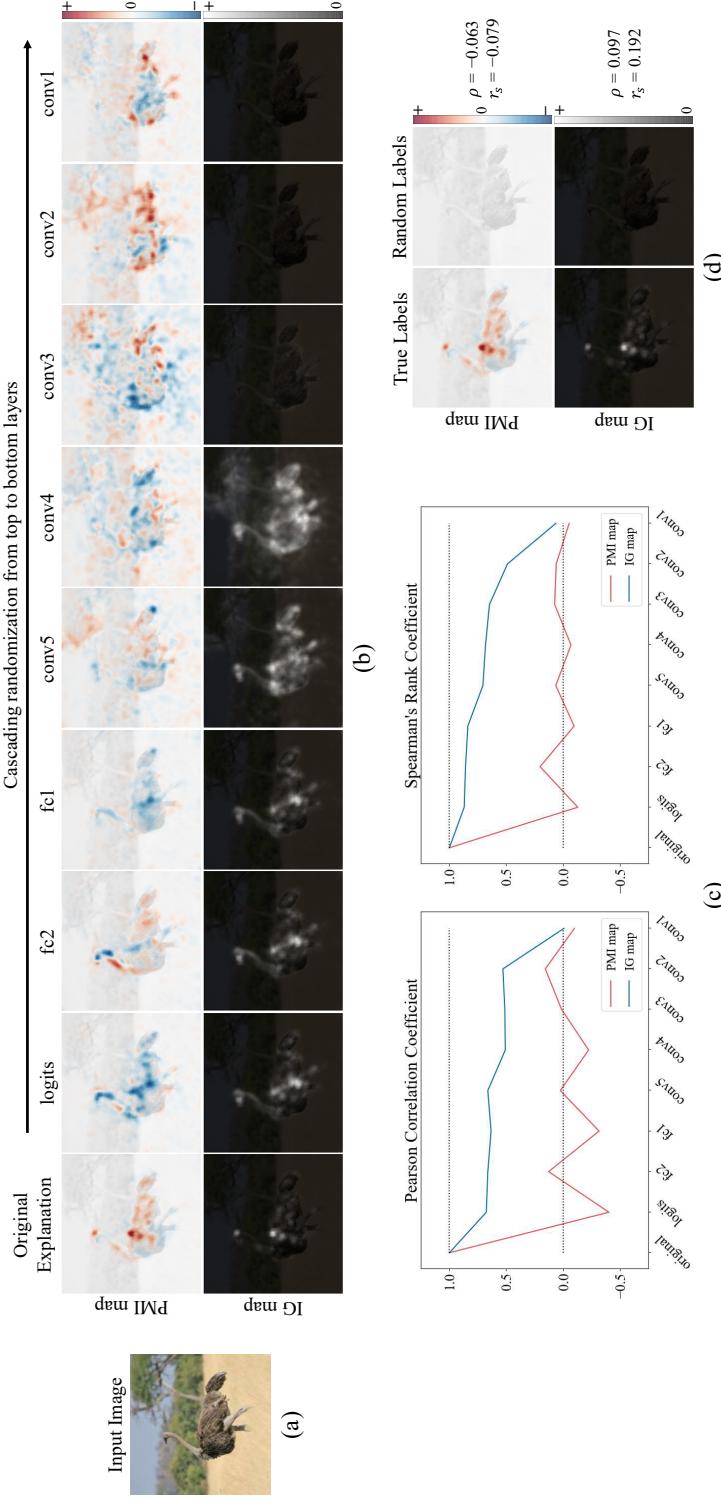


Fig. 24. **An input image (a), results from the parameter randomization test [22] (b, c), and the label randomization test [22] (d).** In a parameter randomization test, weights of VGG19 [2] are randomized from top to bottom layers. PMI and IG maps for each randomized classifier are depicted in (b). As each layer is successively randomized, both attribution maps become dissimilar to the original explanation. Pearson correlation coefficient and Spearman's rank coefficient between each map and the original maps are plotted in (c). PMI and IG maps for a fully randomized classifier are completely different from the original maps; both coefficients are close to zero. Thus they passed the first test. In a label randomization test, VGG19 is retrained to fit the ImageNet training images with random labels. Both attribution maps for the deformed classifier look completely different from the original maps. They imply that no pixels provided useful information *overall* nor supported the *ostrich* class. Indeed, numerical comparison using Pearson correlation coefficient ( $\rho$ ) and Spearman's rank coefficient ( $r_s$ ) confirms this point. Therefore, PMI and IG maps passed both Sanity Checks [22].

## APPENDIX B IMPLEMENTATION DETAILS

### A. Environments

The proposed method is implemented using TensorFlow [30] version 1.12, on a machine consisting of an Intel i7-6850K CPU and a GeForce GTX 1080 Ti GPU.

### B. Running time

The generation of a pair of PMI and IG maps of VGG19 [2] for one input image requires less than thirty minutes using a single GPU. Note that the running time is linear in  $N$ , and reducing  $N$  to a quarter of the used value (2) gives almost identical attribution maps. Moreover, additional approximations, such as increasing the patch stride, can further reduce the computation time to a couple of minutes.

### C. PatchSampler

PatchSampler predicts pixel values of a patch from a larger patch surrounding it. We model PatchSampler as a stack of the convolutional (conv.) layers as follows: conv9-256, conv9-256, maxpool2, crop out 2 pixels from each edge, conv8-256, conv8-256, conv8-256, conv8-256, conv8-256, and conv8-768. Every conv. and maxpool operations used stride size of one (conv9-256 means a conv. layer with kernel size 9 yielding 256 feature maps). The channels in the last layer are grouped into three, representing RGB channel for 256 possible values. Therefore, the softmax activations are applied for each group of channels. The output of PatchSampler is three groups of 256-class categorical distributions for each pixel and each RGB channel. PatchSampler is then trained to predict the pixel values using an ImageNet training set and multi-class classification loss. All convolutional layers except for the last layer are followed by a LeakyReLU [31] activation with  $\alpha = 0.2$ .

### D. Classifiers

The parameter weights of the analyzed classifiers were downloaded from PyTorch model zoo<sup>2</sup>.

### E. Methods

The implementations of other attribution maps are downloaded from various online sources as follows. Prediction Difference Analysis (PDA) [6]: <https://github.com/lmzintgraf/DeepVis-PredDiff>; Saliency map (Gradients) [13], Integrated Gradients [4], and  $\epsilon$ -LRP [14]: <https://github.com/marcoancona/DeepExplain>; Grad-CAM [17]: <https://github.com/jacobgil/pytorch-grad-cam>; Real Time Image Saliency [18]: <https://github.com/PiotrDabkowski/pytorch-saliency>; Meaningful Perturbation [9]: <https://github.com/jacobgil/pytorch-explain-black-box>; FIDO [10]: <https://github.com/zzzace2000/FIDO-saliency>; Extremal Perturbation [8]: <https://github.com/facebookresearch/TorchRay>; RISE [7]: <https://github.com/ecliique/RISE>. Occlusion [5] was implemented by us. SSR loss was used for a generation of FIDO maps because the loss is reported to be less susceptible to artifacts than SDR loss [10].

<sup>2</sup>[https://pytorch.org/docs/stable/model\\_zoo.html#module-torch.utils.model\\_zoo](https://pytorch.org/docs/stable/model_zoo.html#module-torch.utils.model_zoo)

## APPENDIX C ADDITIONAL REMARKS

### A. Backprop-based method

Since Simonyan *et al.* [13] first suggested Saliency map to visualize attribution by using a gradient of class score with respect to the input, successive works have improved its visual quality by reducing the noise with averaging [15] and by using integration [4]. In addition, some works have changed the back-propagation rule by using various heuristic methods such as Deconv [5], Layer-wise Relevance Propagation (LRP) [14], and Guided Back Propagation [16], thereby producing visually appealing results. Because these methods calculate an attribution map through several backpropagations, they are called backprop-based methods.

As indicated by Dabkowski *et al.* [18], the backprop-based methods are fast enough to be real-time, but their quality is limited. Moreover, some of these methods fail Sanity Checks [22] by showing invariance to classifier parameters and to the training data. Therefore, the failed methods are proved to be inadequate in many domains. Moreover, Nie *et al.* [32] theoretically showed that Deconv [5] and Guided Back Propagation [16] simply generated a map similar to the input image, rather than providing an explanation for the classifier.

### B. Mask-based method

Dabkowski *et al.* [18] attempted to find an attribution map having a different implication. Their attribution map is a mask that covers a relevant region of the target object. They generated the mask in real-time by training a mask-generating model in advance. In another approach [9], [10], [8], similar masks were optimized iteratively using a gradient descent, in which additional loss terms were used to force the mask to cluster together and suppress high-frequency components. Such masks tend to cover the entire object in the image rather than only the most decisive portion within the object. For example, given an image of a *car*, mask-based methods cover the entire car area, whereas in other approaches [5], [6], a high score will be assigned only to the most salient subregions, such as the wheels.

## REFERENCES

- [1] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [2] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [3] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3145–3153.
- [4] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.
- [5] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [6] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, “Visualizing deep neural network decisions: Prediction difference analysis,” *arXiv preprint arXiv:1702.04595*, 2017.
- [7] V. Petsiuk, A. Das, and K. Saenko, “Rise: Randomized input sampling for explanation of black-box models,” *arXiv preprint arXiv:1806.07421*, 2018.

- [8] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2950–2958.
- [9] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.
- [10] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," 2018.
- [11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.
- [12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.
- [13] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
- [14] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.
- [15] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.
- [16] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
- [17] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [18] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Advances in Neural Information Processing Systems*, 2017, pp. 6967–6976.
- [19] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.
- [20] S. Liang, Y. Li, and R. Srikanth, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.
- [21] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation."
- [22] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, 2018, pp. 9505–9515.
- [23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.
- [24] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3681–3688.
- [25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [29] U. Hwang, D. Jung, and S. Yoon, "Hexagan: Generative adversarial nets for real world classification," in *International Conference on Machine Learning*, 2019, pp. 2921–2930.
- [30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," *arXiv preprint arXiv:1603.04467*, 2016.
- [31] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.
- [32] W. Nie, Y. Zhang, and A. Patel, "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations," *arXiv preprint arXiv:1805.07039*, 2018.



**Jihun Yi** received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include deep learning, anomaly detection, and explainable AI.



**Eunji Kim** received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2018, where she is currently pursuing the integrated M.S./Ph.D. degree in electrical and computer engineering. Her research interests include artificial intelligence, deep learning, and computer vision.



**Siwon Kim** received her B.S. degrees in electrical and computer engineering from Seoul National University in Seoul, Korea in 2018. Currently, she is pursuing an integrated M.S./Ph.D. degree in electrical and computer engineering at Seoul National University. Her research interests include artificial intelligence, deep learning and biomedical applications.



**Sungroh Yoon** (S'99–M'06–SM'11) received the B.S. degree in electrical engineering from Seoul National University, South Korea, in 1996, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, CA, USA, in 2002 and 2006, respectively. From 2016 to 2017, he was a Visiting Scholar with the Department of Neurology and Neurological Sciences, Stanford University. He held research positions at Stanford University and Synopsys, Inc., Mountain View, CA, USA. From 2006 to 2007, he was with Intel Corporation, Santa Clara, CA, USA. He was an Assistant Professor with the School of Electrical Engineering, Korea University, from 2007 to 2012. He is currently a Professor with the Department of Electrical and Computer Engineering, Seoul National University. His current research interests include machine learning and artificial intelligence. He was a recipient of the SNU Education Award, in 2018, the IBM Faculty Award, in 2018, the Korean Government Researcher of the Month Award, in 2018, the BRIC Best Research of the Year, in 2018, the IMIA Best Paper Award, in 2017, the Microsoft Collaborative Research Grant, in 2017, the SBS Foundation Award, in 2016, the IEEE Young IT Engineer Award, in 2013, and many other prestigious awards.