Shap-CAM: Visual Explanations for Convolutional Neural Networks based on Shapley Value

Quan Zheng[1,2,3], Ziwei Wang[1,2,3], Jie Zhou[1,2,3], and Jiwen Lu[1,2,3]

[1] Department of Automation, Tsinghua University, China
[2] State Key Lab of Intelligent Technologies and Systems, China
[3] Beijing National Research Center for Information Science and Technology, China
{zhengq20,wang-zw18}@mails.tsinghua.edu.cn
{jzhou,lujiwen}@tsinghua.edu.cn

Abstract. Explaining deep convolutional neural networks has been recently drawing increasing attention since it helps to understand the networks' internal operations and why they make certain decisions. Saliency maps, which emphasize salient regions largely connected to the network's decision-making, are one of the most common ways for visualizing and analyzing deep networks in the computer vision community. However, saliency maps generated by existing methods cannot represent authentic information in images due to the unproven proposals about the weights of activation maps which lack solid theoretical foundation and fail to consider the relations between each pixels. In this paper, we develop a novel post-hoc visual explanation method called Shap-CAM based on class activation mapping. Unlike previous gradient-based approaches, Shap-CAM gets rid of the dependence on gradients by obtaining the importance of each pixels through Shapley value. We demonstrate that Shap-CAM achieves better visual performance and fairness for interpreting the decision making process. Our approach outperforms previous methods on both recognition and localization tasks.

Keywords: CNNs, Explainable AI, Interpretable ML, Neural Network Interpretability

1 Introduction

The dramatic advance of machine learning within the form of deep neural networks has opened up modern Artificial Intelligence (AI) capabilities in real-world applications. Deep learning models achieve impressive results in tasks like object detection, speech recognition, machine translation, which offer tremendous benefits. However, the connectionist approach of deep learning is fundamentally different from earlier AI systems where the predominant reasoning methods are logical and symbolic. These early systems can generate a trace of their inference

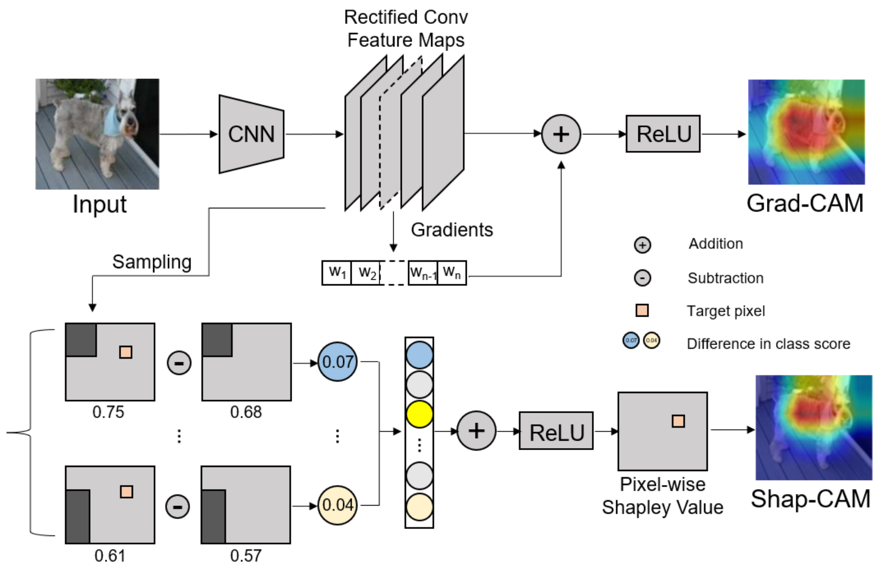

Fig. 1. Comparison between the conventional gradient-based CAM methods and our proposed Shap-CAM. The gradient-based CAM methods, taking Grad-CAM for example, combine the rectified convolutional feature maps and the gradients via backpropagation to compute the saliency map which represents where the model has to look to make the particular decision. Our Shap-CAM introduces Shapley value to estimate the marginal contribution of pixels. For a given pixel in the feature map (viewed in color), we sample various pixel combinations and compute the score difference if the given pixel is added. The score differences are synthesized to obtain pixel-wise Shapley value, which generates the finally Shap-CAM saliency map.

steps, which at that point serves as the basis for explanation. On the other hand, the usability of today's intelligent systems is limited by the failure to explain their decisions to human users. This issue is particularly critical for risk-sensitive applications such as security, clinical decision support or autonomous navigation.

For this gap, various methods have been proposed by researchers over the last few years to figure out what knowledge is hidden in the layers and connections when utilizing deep learning models. While encouraging development has been carrying this field forward, existing efforts are restricted and the goal of explainable deep learning still has a long way to go, given the difficulty and wide range of issue scopes.

In the context of understanding Convolutional Neural Networks (CNNs), Zhou et al. proposed a technique called CAM (Class Activation Mapping), and demonstrated that various levels of the CNN functioned as unsupervised object detectors [34]. They were able to obtain heat maps that illustrate which regions of an input image were looked at by the CNN for assigning a label by employing a global average pooling layer and showing the weighted combination of the resulting feature maps at the penultimate (pre-softmax) layer. However, this technique was architecture-sensitive and involved retraining a linear classifier for each class. Similar methods were examined with different pooling layers such as global max pooling and log-sum-exp pooling [19, 22]. After that, Selvaraju et al. developed Grad-CAM, an efficient version of CAM that combines the class-conditional property of CAM with current pixel-space gradient visualization techniques like Guided Back-propagation and Deconvolution to emphasize fine-grained elements on the image [24]. Grad-CAM improved the transparency of CNN-based models by displaying input regions with high resolution details that are critical for prediction. The variations of Grad-CAM, such as Grad-CAM++ [7], introduce more reliable expressions for pixel-wise weighting of the gradients. However, gradient-based CAM methods cannot represent authentic information in images due to the unproven proposals about the weights of activation maps [1, 2, 6, 8, 17, 18]. Adversarial model manipulation methods fool the explanations by manipulating the gradients without noticeable modifications to the original images [15], proving that the gradient-based CAM methods are not robust and reliable enough.

To this end, Wang et al. proposed Score-CAM which got rid of the dependence on gradients by obtaining the weight of each activation map through its forward passing score on target class [31]. Though Score-CAM discarded gradients for generating explanations, it still suffered from self designed expression of score which lacked solid theoretical foundation and failed to take the relationship between pixels into consideration. In this work, we present a new post-hoc visual explanation method, named Shap-CAM, where the importance of pixels is derived from their marginal contribution to the model output utilizing Shapley value. Our contributions are:

- We propose a novel gradient-free visual explanation method, Shap-CAM, which introduces Shapley value in the cooperative game theory to estimate the marginal contribution of pixels. Due to the superiority of Shapley value and the consideration of relationship between pixels, more rational and accurate contribution of each pixel is obtained.
- We quantitatively evaluate the generated saliency maps of Shap-CAM on recognition and localization tasks and show that Shap-CAM better discovers important features.
- We show that in a constrained teacher-student setting, it is possible to achieve an improvement in the performance of the student by using a specific loss function inspired from the explanation maps generated by Shap-CAM, which indicates that our explanations discover authentic semantic information mined in images.

The remainder of the paper is organized as follows. In Sec. 2, we introduce the related work about visual explanations and Shapley value. In Sec. 3, we develop our Shap-CAM for the generation of visual explanations based on Shapley value. In Sec. 4, we present some experimental results on recognition and localization tasks and show the effectiveness of our proposed method. We finish the paper with final conclusions and remarks.

2 Related Work

2.1 Visual Explanations

We give a summary of related attempts in recent years to understand CNN predictions in this part. Zeiler et al. provided one of the earliest initiatives in this field, developing a deconvolution approach to better grasp what the higher layers of a given network have learned [33]. Springenberg et al. extended this work to guided backpropagation, which allowed them to better comprehend the impact of each neuron in a deep network on the input image [29]. From a different perspective, Ribeiro et al. introduced LIME (Local Interpretable Model-Agnostic Explanations), an approach that uses smaller interpretable classifiers like sparse linear models or shallow decision trees to make a local approximation to the complex decision surface of any deep model [23]. Shrikumar et al. presented DeepLift, which approximates the instantaneous gradients (of the output with respect to the inputs) with discrete gradients to determine the relevance of each input neuron for a given decision [26]. Al-Shedivat et al. presented Contextual Explanation Networks (CENs), a class of models that learns to anticipate and explain its decision simultaneously [3]. Unlike other posthoc model-explanation tools, CENs combine deep networks with context-specific probabilistic models to create explanations in the form of locally-correct hypotheses.

Class Activation Mapping (CAM) [34] is a technique for discovering discriminative regions and giving understandable explanations of deep models across domains. In CAM, the authors demonstrate that a CNN with a Global Average Pooling (GAP) layer after the last convolutional layer shows localization capabilities despite not being explicitly trained to do so. The CAM explanation regards the importance of each channel as the weight of fully connected layer connecting the global average pooling and the output probability distribution. However, an obvious limitation of CAM is the requirements of a GAP penultimate layer and retraining of an additional fully connected layer. To resolve this problem, Grad-CAM [24] extends the CAM explanation and regards the importance of each channel as the gradient of class confidence w.r.t. the activation map. In Grad-CAM, the authors naturally regard gradients as the importance of each channel towards the class probability, which avoids any retraining or model modification. Variations of Grad-CAM, like Grad-CAM++ [7], use different combinations of gradients and revise the weights for adapting the explanations to different conditions.

However, gradient-based CAM methods do not have solid theoretical foundation and receive poor performances when the gradients are not reliable. Hoe et al. explored whether the neural network interpretation methods can be fooled via adversarial model manipulation, a model fine-tuning step that aims to radically alter the explanations without hurting the accuracy of the original models [15]. They showed that the state-of-the-art gradient-based interpreters can be easily fooled by manipulating the gradients with no noticeable modifications to the original images, proving that the gradient-based CAM methods are not robust and reliable enough. Score-CAM gets rid of the dependence of gradients and introduces channel-wise increase of confidence as the importance of each channel [31]. Score-CAM obtains the weight of each activation map through its forward passing score on target class, the final result is obtained by a linear combination of weights and activation maps. This approach however suffers from self-designed expression of score which lacked solid theoretical foundation. Besides, it fails to consider the relationship between different pixels.

2.2 Shapley Value

One of the most important solution concepts in cooperative games was defined by Shapley [25]. This solution concept is now known as the Shapley value. The Shapley value is useful when there exists a need to allocate the worth that a set of players can achieve if they agree to cooperate. Although the Shapley value has been widely studied from a theoretical point of view, the problem of its calculation still exists. In fact, it can be proved that the problem of computing the Shapley value is an NP-complete problem [9].

Several authors have been trying to find algorithms to calculate the Shapley value precisely for particular classes of games. In Bilbao et al. for example, where a special class of voting game is examined, theoretical antimatriod concepts are used to polynomially compute the Shapley value [11]. In Granot et al. a polynomial algorithm is developed for a special case of an operation research game [13]. In Castro et al., it is proved that the Shapley value for an airport game can be computed in polynomial time by taking into account that this value is obtained using the serial cost sharing rule [4].

Considering the wide application of game theory to real world problems, where exact solutions are often not possible, a need exists to develop algorithms that facilitate this approximation. Although the multilinear extension defined by Owen is an exact method for simple games [20], the calculation of the corresponding integral is not a trivial task. So, when this integral is approximated (using the central limit theorem) this methodology could be considered as an approximation method. In Fatima et al. , a randomized polynomial method for determining the approximate Shapley value is presented for voting games [10]. Castro et al. develop an efficient algorithm that can estimate the Shapley value for a large class of games [5]. They use sampling to estimate the Shapley value and any semivalues. These estimations are efficient if the worth of any coalition can be calculated in polynomial time.

In this work, we propose a new post-hoc visual explanation method, named Shap-CAM, where the importance of pixels is derived from their marginal contribution to the model output utilizing Shapley value. Due to the superiority of Shapley value and the consideration of relationship between pixels, more rational and accurate explanations are obtained.

3 Approach

In this section, we first present the prelimilaries of visual explanations and the background the CAM methods. Then we introduce the theory of Shapley Value

[25], a way to quantify the marginal contribution of each player in the cooperative game theory. We apply this theory to our problem and propose the definition of Shap-CAM. Finally, we clarify the estimation of Shapley value in our method. The comparison between the conventional gradient-based CAM methods and our proposed Shap-CAM is illustrated in Fig. 1.

3.1 Prelimilaries

Let function $\boldsymbol{Y} = f(\boldsymbol{X})$ be a CNN which takes $\boldsymbol{X}$ as an input data point and outputs a probability distribution $\boldsymbol{Y}$. We denote Y^c as the probability of class c. For the last convolutional layer, $\boldsymbol{A}^k$ denotes the feature map of the k-th channel.

In CAM, the authors demonstrate that a CNN with a Global Average Pooling (GAP) layer after the last convolutional layer shows localization capabilities despite not being explicitly trained to do so. However, an obvious limitation of CAM is the requirements of a GAP penultimate layer and retraining of an additional fully connected layer. To resolve this problem, Grad-CAM [24] extends the CAM explanation and regards the importance of each channel as the gradient of class confidence $\boldsymbol{Y}$ w.r.t. the activation map $\boldsymbol{A}$, which is defined as:

$$L^c_{ij,\ Grad-CAM} = ReLU\Big(\sum_k w^c_k A^k_{ij}\Big) \tag{1}$$

where

$$w^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial Y^c}{\partial A^k_{ij}} \tag{2}$$

Constant Z stands for the number of pixels in the activation map. In Grad-CAM, the explanation is a weighted summation of the activation maps $\boldsymbol{A}^k$, where the gradients are regarded as the importance of each channel towards the class probability, which avoids any retraining or model modification. Variations of Grad-CAM, like Grad-CAM++ [7], use different combinations of gradients and revise w^c_k in Eq. (1) for adapting the explanations to different conditions.

However, gradient-based CAM methods do not have solid theoretical foundation and can be easily fooled by adversarial model manipulation methods [15]. Without noticeable modifications to the original images or hurting the accuracy of the original models, the gradient-based explanations can be radically alteredby manipulating the gradients, proving that the gradient-based CAM methods are not robust and reliable enough. Score-CAM [31] gets rid of the dependence of gradients and introduces channel-wise increase of confidence as the importance score of each channel. This approach however suffers from self designed expression of the score which fails to consider the relationship between different pixels.

3.2 Definition of Shap-CAM

In order to obtain more accurate and rational estimation of the marginal contribution of each pixel to the model output, we turn to the cooperative game theory. The Shapley value [25] is useful when there exists a need to allocate the worth that a set of players can achieve if they agree to cooperate. Consider a set of n players $\mathbb{P}$ and a function $f(\mathbb{S})$ which represents the worth of the subset of s players $\mathbb{S} \subseteq \mathbb{P}$. The function $f : 2^{\mathbb{P}} \rightarrow \mathbb{R}$ maps each subset to a real number, where $2^{\mathbb{P}}$ indicates the power set of $\mathbb{P}$. Shapley Value is one way to quantify the marginal contribution of each player to the result $f(\mathbb{P})$ of the game when all players participate. For a given player i, its Shapley value can be computed as:

$$Sh_i(f) = \sum_{\mathbb{S} \subseteq \mathbb{P}, i \notin \mathbb{S}} \frac{(n-s-1)!s!}{n!} \left[f(\mathbb{S} \cup \{i\}) - f(\mathbb{S}) \right] \tag{3}$$

The Shapley value for player i defined above can be interpreted as the average marginal contribution of player i to all possible coalitions $\mathbb{S}$ that can be formed without it. Notably, it can be proved that Shapley value is the only way of assigning attributions to players that satisfies the following four properties:

Null player. If the class probability does not depend on any pixels, then its attribution should always be zero. It ensures that a pixel has no contribution if it does not bring any score changes to every possible coalition.

Symmetry. If the class probability depends on two pixels but not on their order (i.e. the values of the two pixels could be swapped, never affecting the probability), then the two pixels receive the same attribution. This property, also called anonymity , is arguably a desirable property for any attribution method: if two players play the exact same role in the game, they should receive the same attribution.

Linearity. If the function f can be seen as a linear combination of the functions of two sub-networks (i.e. $f = af_1 + bf_2$), then any attribution should also be a linear combination, with the same weights, of the attributions computed on the sub-networks, i.e. $Sh_i(\boldsymbol{x}|f) = a \cdot Sh_i(\boldsymbol{x}|f_1) + b \cdot Sh_i(\boldsymbol{x}|f_2)$. Intuitively, this is justified by the need for preserving linearities within the network.

Efficiency. An attribution method satisfies efficiency when attributions sum up to the difference between the value of the function evaluated at the input, and the value of the function evaluated at the baseline, i.e. $\sum_{i=1}^{n} Sh_i = \Delta f = f(\boldsymbol{x}) - f(\mathbf{0})$. In our problem, this property indicates that all the attributions of the pixels sum up to the difference between the output probability of the original feature map and the output of the feature map where no original pixels remain. This property, also called completeness or conservation, has been recognized by previous works as desirable to ensure the attribution method is comprehensive in its accounting. If the difference $\Delta f > 0$, there must exist some pixels assigned a non-zero attribution, which is not necessarily true for gradient-based methods.

Back to our problem on class activation mapping, we consider each pixel (i, j) in the feature map of the last convolutional layer $\boldsymbol{A}$ as a player in the cooperative game. Let $\mathbb{P} = \{(i, j)|i = 1, \ldots, h; j = 1, \ldots, w\}$ be the set of pixels in the feature map $\boldsymbol{A}$, where h, w stand for the height and width of the feature map. Let $n = h \cdot w$ be the number of pixels in the activation map. We then define the worth function f in Eq. (3) as the class confidence Y^c, where c is the class of interest. For each subset $\mathbb{S} \subseteq \mathbb{P}$, $Y^c(\mathbb{S})$ represents the output probability of class c when only the pixels in the set $\mathbb{S}$ remain and the others are set to the average value of the whole feature map. By the symbolization above, the original problem turns to an n-player game $(\mathbb{P}, Y^c)$. Naturally, the Shapley Value of the pixel (i,j) represents its marginal contribution to the class confidence. Thus, we define the Shapley Value as the saliency map of our Shap-CAM:

$$\begin{aligned} L^c_{ij,\ Shap-CAM} &= Sh_{(i,j)}(Y^c) \\ &= \sum_{\mathbb{S}\subseteq\mathbb{P},(i,j)\notin\mathbb{S}} \frac{(n-s-1)!s!}{n!}\left[Y^c(\mathbb{S}\cup\{(i,j)\}) - Y^c(\mathbb{S})\right] \end{aligned} \tag{4}$$

The obtained heatmap is then upsampled to the size of the original image, as it is done in Grad-CAM.

The contribution formula that uniquely satisfies all these properties is that a pixel's contribution is its marginal contribution to the class confidence of every subset of the original feature map. Most importantly, this formula takes into account the interactions between different pixels. As a simple example, suppose there are two pixels that improve the class confidence only if they are both present or absent and harm the confidence if only one is present. The equation considers all these possible settings. This is one of the few methods that take such interactions into account and is inspired by similar approaches in Game Theory. Shapley value is introduced as an equitable way of sharing the group reward among the players where equitable means satisfying the aforementioned properties. It's possible to make a direct mapping between our setting and a cooperative game; therefore, proving the uniqueness of Shap-CAM.

3.3 Estimation of Shapley Value

Exactly computing Eq. (3) would require $\mathcal{O}(2^n)$ evaluations. Intuitively, this is required to evaluate the contribution of each activation with respect to all possible subsets that can be enumerated with the other ones. Clearly, the exact computation of Shapley values is computationally unfeasible for real problems. Sampling is a process or method of drawing a representative group of individuals or cases from a particular population. Sampling and statistical inference are used in circumstances in which it is impractical to obtain information from every member of the population. Taking this into account, we use sampling in this paper to estimate the Shapley value and any semivalues. These estimations are efficient if the worth of any coalition can be calculated in polynomial time. Here we use a sampling algorithm to estimate Shapley value which reduces the complexity to $\mathcal{O}(mn)$, where m is the number of samples taken [5].

Following the definition of Shapley value in Eq. (3), an alternative definition of the Shapley value can be expressed in terms of all possible orders of the players. Let $O : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a permutation that assigns to each position k the player $O(k)$. Let us denote by $\pi(\mathbb{P})$ the set of all possible permutations with player set $\mathbb{P}$. Given a permutation O, we denote by $Pre^i(O)$ the set of predecessors of the player i in the order O, i.e. $Pre^i(O) = \{O(1), \ldots, O(k-1)\}$, if $i = O(k)$.

It can be proved that the Shapley value in Eq. (3) can be expressed equivalently in the following way:

$$Sh_i(f) = \frac{1}{n!} \sum_{O \in \pi(\mathbb{P})} \left[f(Pre^i(O) \cup \{i\}) - f(Pre^i(O)) \right], \quad i = 1, \dots, n \tag{5}$$

In estimation, we randomly take m samples of player order O from $\pi(\mathbb{P})$, calculate the marginal contribution of the players in the order O, which is defined in the summation of the equation above, and finally average the marginal contributions as the approximation.

Then we will obtain, in polynomial time, an estimation of the Shapley value with some desirable properties. To estimate the Shapley value, we will use a unique sampling process for all players. The sampling process is defined as follows:

- The population of the sampling process P will be the set of all possible orders of n players. The vector parameter under study is $Sh = (Sh_1, \dots, Sh_n)$.
- The characteristics observed in each sampling unit are the marginal contributions of the players in the order O, i.e.

$$\chi(O) = \{\chi(O_i)\}_{i=1}^{n}, \quad where\ \chi(O)_i = f(Pre^i(O) \cup \{i\}) - f(Pre^i(O)) \tag{6}$$

- The estimate of the parameter will be the mean of the marginal contributions over the sample M, i.e.

$$\hat{Sh} = (\hat{Sh}_1, \dots, \hat{Sh}_n), \qquad where \quad \hat{Sh}_i = \frac{1}{m} \sum_{O \in M} \chi(O)_i. \tag{7}$$

4 Experiments

In this section, we conduct experiments to evaluate the effectiveness of the proposed explanation method. We first introduce the datasets for evaluation and our implementation details. Then we assess the fairness of the explanation (the significance of the highlighted region for the model's decision) qualitatively via visualization in Sec. 4.2 and quantatively on image recognition in Sec. 4.3. In Sec. 4.4 we show the effectiveness for class-conditional localization of objects in a given image. The knowledge distillation experiment is followed in Sec. 4.5.

4.1 Datasets and Implementation Details

We first detail the datasets that we carried out experiments on: The ImageNet (ILSVRC2012) dataset consists of about 1.2 million and 50k images from 1,000 classes for training and validation respectively. We conducted the following experiments on the validation split of ImageNet. The PASCAL VOC dataset consists of 9,963 natural images from 20 different classes. We used the PASCAL

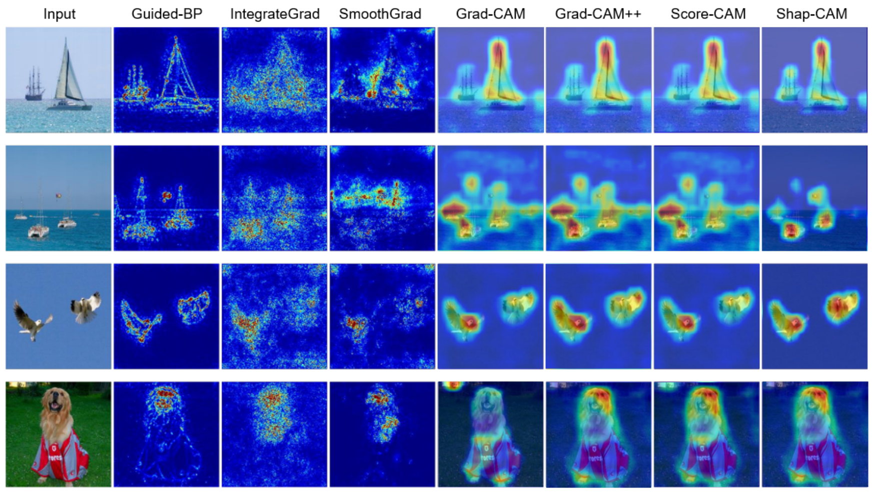

Fig. 2. Visualization results of Guided Backpropagation [29], SmoothGrad [28], IntegrateGrad [30], Grad-CAM [24], Grad-CAM++ [7], Score-CAM [31] and our proposed Shap-CAM.

VOC 2007 trainval sets which contained 5,011 images for the recognition evaluation.

For both the ImageNet and the PASCAL VOC datasets, all images are resized to $224 \times 224 \times 3$, transformed to the range [0, 1], and then normalized using mean vector [0.485, 0.456, 0.406] and standard deviation vector [0.229, 0.224, 0.225]. In the following experiments, we use pre-trained VGG16 network [27] from the Pytorch model zoo as a base model. As for the calculation of Shapley value in Eq. (4), only the pixels in the set $\mathbb{S}$ are preserved and the others are set to the average value of the whole feature map. Unless stated otherwise, the sampling number for estimating Shapley value is set to 10^4. For a fair comparison, all saliency maps are upsampled with bilinear interpolate to 224×224.

4.2 Qualitative Evaluation via Visualization

We qualitatively compare the saliency maps produced by recently SOTA methods, including gradient-based methods (Guided Backpropagation [29], IntegrateGrad [30], SmoothGrad [28]), and activation-based methods (Grad-CAM [24], Grad-CAM [24], Grad-CAM++ [7]) to validate the effectiveness of Shap-CAM. As shown in Fig. 2, results in Shap-CAM, random noises are much less than that other methods. In addition, Shap-CAM generates smoother saliency maps comparing with gradient-based methods. Due to the introduction of Shapley value, our Shap-CAM is able to estimate the marginal contribution of the pixels and take the relationship between pixels into consideration, and thus can obtain more rational and accurate saliency maps.

Table 1. Recognition evaluation results on the ImageNet (ILSVRC2012) validation set (lower is better in Average Drop, higher is better in Average Increase).

Method	Mask	RISE	GradCAM	GradCAM++	ScoreCAM	ShapCAM
Avr. Drop(%)	63.5	47.0	47.8	45.5	31.5	**28.0**
Avr. Increase(%)	5.29	14.0	19.6	18.9	30.6	**31.8**

Table 2. Recognition evaluation results on the PASCAL VOC 2007 validation set (lower is better in Average Drop, higher is better in Average Increase).

Method	Mask	RISE	GradCAM	GradCAM++	ScoreCAM	ShapCAM
Avr. Drop(%)	45.3	31.3	28.5	19.5	15.6	**13.2**
Avr. Increase(%)	10.7	18.2	21.4	19.0	28.9	**32.7**

4.3 Faithfulness Evaluation via Image Recognition

The faithfulness evaluations are carried out as depicted in Grad-CAM++ [7] for the purpose of object recognition. The original input is masked by point-wise multiplication with the saliency maps to observe the score change on the target class. In this experiment, rather than do point-wise multiplication with the original generated saliency map, we slightly modify by limiting the number of positive pixels in the saliency map. Two metrics called Average Drop, Average Increase In Confidence are introduced:

Average Drop: The Average Drop refers to the maximum positive difference in the predictions made by the prediction using the input image and the prediction using the saliency map. It is given as: $\sum_{i=1}^{N} \frac{\max(0, Y_i^c - O_i^c)}{Y_i^c} \times 100\%$. Here, Y_i^c refers to the prediction score on class c using the input image i and O_i^c refers to the prediction score on class c using the saliency map produced over the input image i. A good explanation map for a class should highlight the regions that are most relevant for decision-making. It is expected that removing parts of an image will reduce the confidence of the model in its decision, as compared to its confidence when the full image is provided as input.

Increase in Confidence: Complementary to the previous metric, it would be expected that there must be scenarios where providing only the explanation map region as input (instead of the full image) rather increases the confidence in the prediction (especially when the context is distracting). In this metric, we measure the number of times in the entire dataset, the model's confidence increased when providing only the explanation map regions as input. The Average Increase in Confidence is denoted as:$\sum_{i=1}^{N} \frac{sign(Y_i^c < O_i^c)}{N} \times 100\%$, where $sign$ presents an indicator function that returns 1 if input is True.

Our comparison extends with state-of-the-art methods, namely gradient-based, perturbation-based and CAM-based methods, including Mask [12], RISE [21], Grad-CAM [24] and Grad-CAM++ [7]. Experiment conducts on the ImageNet (ILSVRC2012) validation set, 2000 images are randomly selected. Results

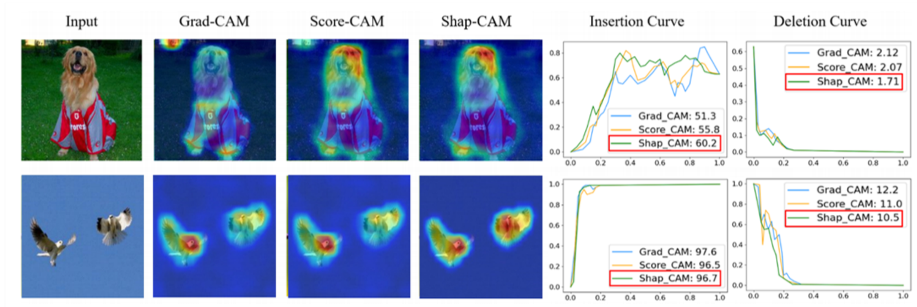

Fig. 3. Grad-CAM, Score-CAM and Shap-CAM generated saliency maps for representative images in terms of deletion and insertion curves. In the insertion curve, a better explanation is expected that the prediction score to increase quickly, while in the deletion curve, it is expected the classification confidence to drop faster.

are reported in Tab. 1. Results on the PASCAL VOC 2007 validation set are reported in Tab. 2.

As shown in Tab. 1 and Tab. 2, Shap-CAM outperforms other perturbation-based and CAM-based methods. Shap-CAM can successfully locate the most distinguishable part of the target item, rather than only determining what humans think is important, based on its performance on the recognition challenge. Results on the recognition task show that Shap-CAM can more accurately reveal the decision-making process of the original CNN model than earlier techniques. Previous methods lack solid theoretical foundation and suffer from self-designed explanations, which are prone to the manipulation of gradients or fail to consider the relationship between different pixels. Our Shap-CAM instead is able to obtain more rational and accurate contribution of each pixel due to the superiority of Shapley value and the consideration of relationship between pixels.

Furthermore, for a more comprehensive comparison, we also evaluate our method on deletion and insertion metrics which are proposed in [21]. In our experiment, we simply remove or introduce pixels from an image by setting the pixel values to zero or one with step 0.01 (remove or introduce 1% pixels of the whole image each step). Example are shown in Fig. 3, where our approach achieves better performance on both metrics compared with SOTA methods.

4.4 Localization Evaluation

In this section, we measure the quality of the generated saliency map through localization ability. Bounding box evaluations are accomplished. We employ the similar metric, as specified in Score-CAM, called the Energy-based pointing game. Here, the amount of energy of the saliency map is calculated by finding out how much of the saliency map falls inside the bounding box. Specifically, the input is binarized with the interior of the bounding box marked as 1 and the region outside the bounding box as 0. Then, this input is multiplied with

Table 3. Localization Evaluations of Proportion (%) using Energy-based Pointing Game (Higher the better).

Method	Grad-CAM	Grad-CAM++	Score-CAM	Shap-CAM
VGG-16	39.95	40.16	40.10	**40.45**
ResNet18	40.90	40.85	40.76	**41.28**

the generated saliency map and summed over to calculate the proportion ratio, which given as

$$Proportion = \frac{\sum L^c_{(i,j)\in bbox}}{\sum L^c_{(i,j)\in bbox} + \sum L^c_{(i,j)\notin bbox}}. \tag{8}$$

Two pre-trained models, namely VGG-16 [27], ResNet18 [14], are used to conduct the energy-based pointing game on the 2000 randomly chosen images from the ILSVRC 2012 Validation set.

We randomly select images from the validation set by removing images where object occupies more than 50% of the whole image. For convenience, we only consider these images with only one bounding box for target class. We experiment on 500 random selected images from the ILSVRC 2012 validation set. Evaluation results are reported in Tab. 3, which show that our method outperforms previous works. This also confirms that the Shap-CAM-generated saliency map has fewer noises. As is shown in the previous research [15], the state-of-the-art gradient-based interpreters can be easily fooled by manipulating the gradients with no noticeable modifications to the original images, proving that the SOTA methods are not robust enough to the noise. On the contrary, our Shap-CAM can alleviate the effect of this problem by estimating the importance of each pixel more accurately.

4.5 Learning from Explanations: Knowledge Distillation

Following the knowledge distillation experiment settings in Grad-CAM++ [7], we show that in a constrained teacher-student learning setting, knowledge transfer to a shallow student (commonly called knowledge distillation) is possible from the explanation of CNN decisions generated by CAM methods. We use Wide ResNets [32] for both the student and teacher networks. We train a WRN-40-2 teacher network (2.2 M parameters) on the CIFAR-10 dataset. In order to train a student WRN-16-2 network (0.7 M parameters), we introduce a modified loss L_{stu}, which is a weighted combination of the standard cross entropy loss L_{CE} and an interpretability loss L_{interp}.

$$L_{stu}(c, W_s, W_t, I) = L_{CE}(c, W_s(I)) + \alpha ||L^c_s(W_s(I)) - L^c_t(W_t(I))||^2_2 \tag{9}$$

where the first term represents the cross entropy loss and the second term represents the interpretability loss L_{interp}. In the above equations, I indicates the input image and c stands for the corresponding output class label. L^c is the

Table 4. Test error rate (%) for knowledge distillation to train a student from a deeper teacher network. L_{CE} is the normal cross entropy loss function. The Column 2-6 refer to the modified loss function L_{stu} where the explanations for images are generated using the corresponding interpreter.

Loss function	L_{CE}	GradCAM	GradCAM++	ScoreCAM	ShapCAM
w/o L_{KD}	6.78	6.86	6.74	6.75	**6.69**
w/ L_{KD}	5.68	5.80	5.56	5.42	**5.37**

explanations given by a certain interpreter. α is a hyper parameter that controls the importance given to the interpretability loss. W_s denotes the weights of the student network, and W_t the weights of the teacher network. This equation above forces the student network not only minimize standard cross-entropy loss for classification, but also learn from the most relevant parts of a given image used for making a decision from the teacher network, which is the influence of the interpretability loss L_{interp} term.

Tab. 4 shows the results for this experiment. L_{CE} is the normal cross entropy loss function, i.e. the student network is trained independently on the dataset without any intervention from the expert teacher. The following four columns refer to loss functions defined in Eq. (9) where the explanations for image I are generated using the corresponding interpreter. We further also included L_{KD}, the knowledge distillation loss introduced by Hinton et al. with temperature parameter set to 4 [16]. It is indicated from these results that knowledge distillation can be improved by considering the explanations of the teacher. The results also show that Shap-CAM provides better explanation-based knowledge distillation than existing CAM-based methods.

5 Conclusion

We propose Shap-CAM, a novel CAM variant, for visual explanations of deep convolutional networks. We introduce Shapley value to represent the marginal contribution of each pixel to the model output. Due to the superiority of Shapley value and the consideration of relationship between pixels, more rational and accurate explanations are obtained. We present evaluations of the generated saliency maps on recognition and localization tasks and show that Shap-CAM better discovers important features. In a constrained teacher-student setting, our Shap-CAM provides better explanation-based knowledge distillation than the state-of-the-art explanation approaches.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 62125603 and Grant U1813218, in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI).

References

1. Adebayo, J., Gilmer, J., Goodfellow, I., Kim, B.: Local explanation methods for deep neural networks lack sensitivity to parameter values. In: arXiv preprint arXiv:1810.03307 (2018)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems (2018)
3. Al-Shedivat, M., Dubey, A., Xing, E.P.: Contextual explanation networks. In: arXiv preprint arXiv:1705.10301 (2017)
4. Castro, J., Gomez, D., Tejada, J.: A polynomial rule for the problem of sharing delay costs in pert networks. In: Computers and Operation Research (2008)
5. Castro, J., Gomez, D., Tejada, J.: Polynomial calculation of the shapley value based on sampling. In: Computers & Operations Research (2009)
6. Chang, C.H., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: arXiv preprint arXiv:1807.08024 (2018)
7. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision. pp. 839–847 (2018)
8. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: Advances in Neural Information Processing Systems (2017)
9. Deng, X., Papadimitriou, C.: On the complexity of cooperative solution concepts. In: Mathematics of Operations Research (1994)
10. Fatima, S., Wooldridge, M., Jenniggs, N.: An analysis of the shapley value and its uncertainty for the voting game. In: Lectures notes in artificial intelligence (2006)
11. Fernández, J., Algaba, E., et al, J.B.: Generating functions for computing the myerson value. In: Annals of Operations Research (2002)
12. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV (2017)
13. Granot, D., Kuipers, J., Chopra, S.: Cost allocation for a tree network with heterogeneous customers. In: Mathematics of Operations Research (2002)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Heo, J., Joo, S., Moon, T.: Fooling neural network interpretations via adversarial model manipulation. In: NeurIPS (2019)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: arXiv preprint arXiv:1503.02531 (2015)
17. Lin, M., Chen, Q., Yan, S.: Network in network. In: arXiv preprint arXiv:1312.4400 (2013)
18. Omeiza, D., Speakman, S., Cintas, C., Weldermariam, K.: Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. In: arXiv preprint arXiv:1908.01224 (2019)
19. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? weakly-supervised learning with convolutional neural networks. In: CVPR. pp. 685–694 (2015)
20. Owen, G.: Multilinear extensions of games. In: Management Science Series B–Application (1972)
21. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: arXiv preprint arXiv:1806.07421 (2018)

22. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR. pp. 1713–1721 (2015)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)
24. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batrat, D.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. In: arXiv preprint arXiv:1610.02391 (2016)
25. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games, 2(28): 307-317 (1953)
26. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: arXiv preprint arXiv:1704.02685 (2017)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv preprint arXiv:1409.1556 (2014)
28. Smilkov, D., Thorat, N., Kim, B., Viegas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. In: arXiv preprint arXiv:1706.03825 (2017)
29. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: arXiv preprint arXiv:1412.6806 (2014)
30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (2017)
31. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: CVPR (2020)
32. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: arXiv preprint arXiv:1605.07146 (2016)
33. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833 (2014)
34. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)