Pattern Recognition 145 (2024) 109898

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

PAMI: Partition Input and Aggregate Outputs for Model Interpretation

Wei Shi [a,b,c], Wentao Zhang [a,c], Wei-shi Zheng [a,c], Ruixuan Wang [a,b,c,*]

[a] School of Computer Science, Sun Yat-sen University, Guangzhou, China
[b] Department of Network Intelligence, Pengcheng Laboratory, Shenzhen, China
[c] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China

ARTICLE INFO

Keywords:
Interpretation
Visualization
Post-hoc

ABSTRACT

There is an increasing demand for interpretation of model predictions especially in high-risk applications. Various visualization approaches have been proposed to estimate the part of input which is relevant to a specific model prediction. However, most approaches require model structure and parameter details in order to obtain the visualization results, and in general much effort is required to adapt each approach to multiple types of tasks particularly when model backbone and input format change over tasks. In this study, a simple yet effective visualization framework called PAMI is proposed based on the observation that deep learning models often aggregate features from local regions for model predictions. The basic idea is to mask majority of the input and use the corresponding model output as the relative contribution of the preserved input part to the original model prediction. For each input, since only a set of model outputs are collected and aggregated, PAMI does not require any model detail and can be applied to various prediction tasks with different model backbones and input formats. Extensive experiments on multiple tasks confirm the proposed method performs better than existing visualization approaches in more precisely finding class-specific input regions, and when applied to different model backbones and input formats. The source code is available at https://github.com/fuermowei/PAMI.

1. Introduction

Deep learning models have shown human-level performance in many machine learning tasks and started to be applied in real scenarios, such as face identification, medical image analysis, and language translation. However, current deep learning models often lack interpretations for their decision making, which hinders the massive deployment of intelligent systems particularly in high-risk applications like medical diagnosis and autonomous driving.

To improve interpretation of model predictions, multiple visualization approaches [1] have been proposed to localize input regions or components which are more relevant to the model prediction given any specific input to the model. For image classification task as an example, the class activation map (CAM) and its variants utilize the output (i.e., feature maps) of certain convolutional layer in the convolutional neural networks (CNNs) and their contribution weights to find the image regions which are responsible for the specific model prediction given any input image [2], and the back-propagation approaches propagate the CNN output layer-by-layer to the input image space either based on gradient information (or its modified versions) at each layer [3] or based on the relevance between input elements and output at each layer [4]. While the CAM-like approaches can only roughly localize relevant regions due to the lower resolution of feature maps, the back-propagation approaches often just find sparse and incomplete object regions relevant to the model prediction. Moreover, both types of approaches need either part of or the whole model structure and parameter details, which may be unavailable in some applications due to privacy or security concerns. When the model details are not available, the occlusion method may be utilized to roughly localize image regions relevant to the model prediction by occluding each local patch and checking the change in model output [5]. However, the occlusion method often only localizes the most discriminative object part and misses the other parts which actually also contribute to the model prediction. LIME [6] is another method without requiring model details for model interpretation by locally approximating the model decision surface for any specific input, but it often requires an optimization process for interpretation of a specific model prediction. Furthermore, most existing visualization approaches for interpretation of model predictions are developed for specific type of tasks (e.g., just for image classification), model backbone (e.g., for CNNs), and input format (e.g., just for image data). Substantial efforts are often required

* Corresponding author at: School of Computer Science, Sun Yat-sen University, Guangzhou, China.
E-mail address: wangruix5@mail.sysu.edu.cn (R. Wang).

https://doi.org/10.1016/j.patcog.2023.109898
Received 11 February 2023; Received in revised form 9 June 2023; Accepted 17 August 2023
Available online 21 August 2023
0031-3203/© 2023 Elsevier Ltd. All rights reserved.

to adapt one visualization approach to various tasks (e.g., image caption) with different model backbones (e.g., Transformer backbone) or input formats (e.g., sequence of items).

Different from existing visualization approaches, a simple yet effective visualization framework for interpretation of model predictions is proposed in this study. The proposed framework, called PAMI ('Partition input and Aggregate outputs for Model Interpretation'), is inspired by the observation that both humans and popular deep learning models extract and aggregate features of local regions for image understanding and decision making. Suppose a well-trained image classifier predicts an input image as a specific class. To find the relevant image regions and their contributions to the model prediction, the proposed framework first partitions the input into multiple parts, and then feeds only one part (with the remaining regions masked) to the model to obtain the corresponding output probability of the specific class. Aggregating the output probabilities over all the individual parts would result in an importance map representing the contribution of each input part to the original model prediction. In contrast to existing visualization approaches, the proposed PAMI framework does not require model structure and parameter details, can more likely find all possible input parts which are relevant to the model prediction, more precisely localize relevant parts, and work for various model backbones with different input formats. Such merits of the proposed PAMI framework has been confirmed by extensive experiments on multiple tasks with different model backbones and input formats.

In the following, we will first summarize relevant studies on visualization-based model interpretation (Section 2), and then provide a detailed description of the proposed framework and comparison with similar methods (Section 3). After that, extensive qualitative and quantitative evaluations are performed to demonstrate the superiority and the effectiveness of our method (Section 4), followed by the conclusion of this study and possible future work (Section 5).

2. Related work

In computer vision, post-hoc interpretation of deep learning models focuses on either understanding of model neurons (e.g., convolutional kernels, output elements) which is independent of input information, or understanding of a specific model prediction given an input image [5]. This study belongs to the latter one, i.e., trying to understand what information in the input causes the specific model prediction.

Multiple approaches have been proposed for understanding of model predictions, including the activation map approach [7], the back-propagation approach [4], the perturbation approach [8], the local approximation approach [6]. The activation map approach often obtains a class-specific activation map with the weighted sum of all feature maps often at the last convolutional layer, and considers the regions with stronger activation relevant to the specific model output [7]. Since the activation map is often much smaller than the input image, only approximate image regions corresponding to the stronger activation regions can be localized for interpretation of the specific model prediction. Different from the activation approach which often works at higher layer of the deep learning model, the back-propagation approach tries to estimate the importance of each input pixel by propagating the specific model output layer-by-layer back to the input space. This can be obtained by calculating the gradient of the specific model output with respect to input elements at each layer [3], or the relevance between output and each input element at each layer [4]. The back-propagation approach considers each input pixel as an independent component and often only a subset of disconnected pixels in the relevant regions are estimated to be relevant to the model prediction.

To find local image regions rather than disconnected pixels relevant to the model prediction, the perturbation approach has been proposed by perturbing local image regions somehow and checking the change in model output of the originally predicted class [9]. If certain perturbed local region causes large drop in the output, the local region in the input image is considered crucial to the original model prediction. Perturbation can be in the form of simply masking a local region by a constant pixel intensity, by neighboring image patches, or by blurring the original region information with a constant value or smoothing operator. This approach can often find only the most discriminative part of the relevant regions which are responsible for the model prediction, because perturbing less-discriminative part of the relevant regions often does not cause much drop in model output. Besides the perturbation approach, the local approximation approach provides another way to estimate the contribution of each meaningful image region (e.g., object parts, background region) to the model prediction [6]. This approach assumes that the model decision surface is locally linear in the region-based feature space for any specific input, and therefore can be approximated with a linear model in the feature space. The weight parameters in the linear model can directly indicate the contribution of each meaningful image region to the original model output, thus obtaining image regions most relevant to the model prediction.

While most of these visualization approaches to interpretation of model predictions were originally developed for image classification models, they have been extended or modified for other tasks or other deep learning models. Besides these approaches, prototype-based [10] and attention-based [11] approaches have also been proposed for model interpretation. Note that except the local approximation approach and part of the perturbation approach (e.g., the occlusion method [8]), most approaches require at least part of the model structure and parameter details in order to find input parts which are relevant to the model prediction.

3. Method

In this study, we aim to provide interpretation for model prediction given any specific input to a well-trained and fixed deep learning model. The interpretation is demonstrated by estimating relative contribution of each input part to the specific model prediction and correspondingly localizing input regions or elements which are relevant to the model prediction. It is worth noting that no model structure and parameter details are assumed to be known during the model interpretation process.

3.1. Motivation

Although humans can often instantly recognize objects in images, certain attention mechanism in human brain is likely involved in the process of object recognition [12]. In other words, humans often need to implicitly or explicitly attend to local regions for image understanding and object recognition. While the detailed human attention mechanism is yet to be further explored, initial studies [13] suggest that most local parts of an object in an image help humans recognize the object, and appearance of only an individual object part could help humans recall the corresponding class of the object. Consistent with the visual attention studies, recent exploration of convolutional neural networks (CNNs) shows that convolutional kernels even at higher convolutional layers (i.e., closer to the CNN output) often have smaller receptive fields than expected. Considering that a global pooling is performed at the last convolutional layer in most CNN classifier models, it is widely accepted that CNN models largely depend on the collection of local image region features for image classification. For the other type of deep learning model backbone Transformer and its variants (e.g., ViT [14], Swin Transformer [15]), since most items in the input sequence at each model layer correspond to components (e.g., words for a sentence input, image patches for an image input) of the original input, the final model prediction also largely depends on the collection of local features of the original input. With the above observation, we hypothesize that the model output response to each single component of the original input may directly imply the importance of the single input component for the specific model prediction.

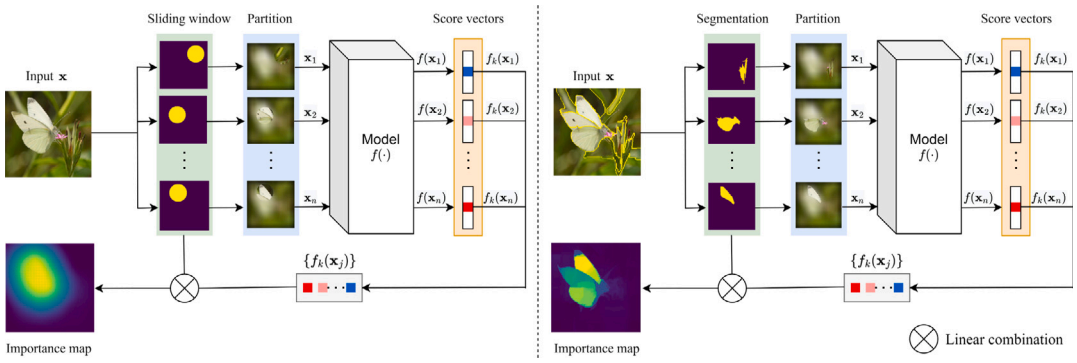

Fig. 1. The proposed PAMI framework with the sliding window based (left half) or the pre-segmentation based (right half) input partition strategy. Each time only one local part of the input is preserved and the remaining parts are masked (blurred here).

3.2. The proposed PAMI framework

The proposed interpretation framework is demonstrated in Fig. 1. For image classification task as an example, given a well-trained classifier model $f(\cdot)$ and any input image $\mathbf{x}$, denote by $f_c(\mathbf{x})$ the output prediction probability of the input image belonging to the cth class. Suppose the model predicts the input as the kth class, i.e., $f_k(\mathbf{x})$ is the maximum over all the output probabilities. To interpret the classifier's prediction, the proposed framework first partitions the input into multiple either overlapped or non-overlapped parts (see the following subsections), and then with the jth individual part preserved and all the remaining image regions masked, the output probability $f_k(\mathbf{x}_j)$ of the kth class for the majority-masked input $\mathbf{x}_j$ is used to estimate the relative contribution of the preserved jth image part to the original model prediction $f_k(\mathbf{x})$. By collecting and aggregating the output responses $f_k(\mathbf{x}_j)$'s over all the partitioned image parts, an importance map $\mathbf{h}_k(\mathbf{x})$ with the same spatial size as that of the input image $\mathbf{x}$ can be generated to represent the contribution of each input element (i.e., pixel here), i.e.,

$$\mathbf{h}_k(\mathbf{x}) = \frac{1}{\mathbf{N}} \circ \sum_{j=1}^{J} \{ f_k(\mathbf{x}_j) \cdot \mathbf{m}_j \}, \tag{1}$$

where $\mathbf{m}_j$ is a binary mask associated with the input $\mathbf{x}_j$, with value 1 for un-masked pixels and 0 for masked pixels, and J is the total number of majority-masked inputs. The symbol "$\circ$" denotes the Hadamard product, and $\mathbf{N}$ is a frequency map with the spatial size of the original input image $\mathbf{x}$, and each element in $\mathbf{N}$ corresponds one image pixel and represents the number of un-masked regions containing the pixel. Local image regions with correspondingly higher response values in the importance map $\mathbf{h}_k(\mathbf{x})$ are supposed to contribute more to the model prediction $f_k(\mathbf{x})$, thus providing visual evidence for the model prediction. It is worth noting that the interpretation framework can be applied to different tasks (e.g., image caption and sentiment analysis) with various input formats.

3.2.1. Input partition strategy I: sliding window

One simple way to partition input is to apply the sliding window strategy with a pre-defined window size and sliding step size, where the window is in certain regular shape (e.g., circular or rectangular). In this way, the original input can be easily partitioned into multiple parts, and each part can be more or less overlapped by its neighboring parts determined by window size and sliding step size. For each partitioned part, all the remaining image regions will be masked somehow (e.g., by black pixels or blurred version of the original regions; see Fig. 2), and the output probability of the originally predicted class can be directly obtained with the majority-masked image as input. Since each input element (e.g., pixel of an input image) could be covered by multiple partitioned parts, the contribution of each input element in the final importance map can be obtained by averaging the output probabilities of the originally predicted class over all the partitioned parts covering the input element (see Eq. (1)).

Note that the window size would affect the resolution level of the final importance map. Although smaller window would result in desired higher resolution, an image part with much smaller size (i.e., too small window) could contain little semantic information such that it becomes challenging for the framework to estimate the contribution of the smaller image part to a specific model prediction. In practice, users can choose one appropriate window size for interpretation of model prediction, or multiple window sizes for multi-scale interpretation of the model prediction.

3.2.2. Input partition strategy II: pre-segmentation

Another way to partition input is to pre-segment the input into multiple parts with certain segmentation strategy (Fig. 3). When the input is an image, various unsupervised segmentation algorithms can be adopted for pre-segmentation of the input. In this study, super-pixel segmentation algorithms are used for input image partition [16–19]. With a particular super-pixel segmentation method, an input image can be partitioned into multiple non-overlapped parts (i.e., super-pixels), with each part often having irregular form of region boundary and likely containing homogeneous visual information. As introduced above, the contribution of each super-pixel to the model prediction can be obtained by preserving the single super-pixel and masking the other super-pixels as the input and collecting the model output response of the predicted class to the majority-masked input.

In practice, due to the imperfect performance of any single super-pixel segmentation method, some super-pixels may contain parts of both object region and background region, resulting in the importance map where part of background regions also has relatively higher responses. To alleviate such an issue, multiple super-pixel segmentation methods are employed, and multiple importance maps based on these segmentation methods are then averaged to estimate the contribution of each input element (e.g., pixels) to the original model prediction (Fig. 4). Considering that one segmentation method with different hyper-parameters often generates different segmentation results, multiple segmentation results from each segmentation method with multiple typical hyper-parameter settings method are employed here. The average importance map may be further improved by running the above process once more (i.e., second run), in which the super-pixel segmentation methods are performed on the average importance map rather than the original input image. In addition, when generating each majority-masked input, the highly smoothed version of the original input image is used to fill the corresponding masked regions.

Compared to the sliding window strategy, the partitioned parts by the pre-segmentation strategy have more precise and reasonable region boundaries particularly for image data. This in turn often leads to the final importance map with clear boundaries between object regions and background regions, thus more precisely locating the image

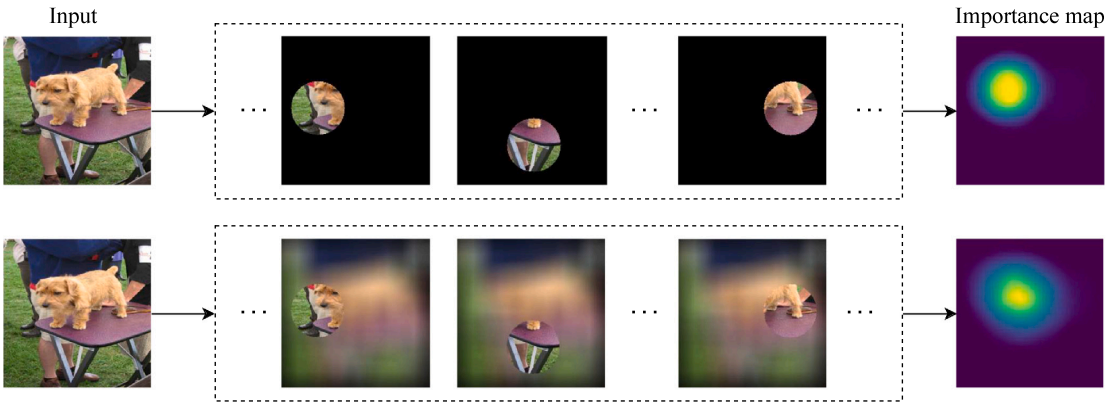

Fig. 2. Demonstration of the sliding window based input partition strategy. Upper half: image pixels inside each sliding window remain original, while pixels outside each sliding window are masked by black pixels; Lower half: pixels outside each sliding window are masked by the blurred version of the corresponding original pixels.

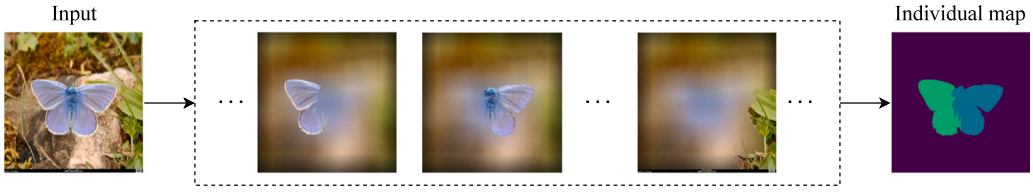

Fig. 3. Demonstration of the pre-segmentation based input partition strategy. The input image is segmented into multiple parts (i.e., super-pixels) using the felzenszwalb segmentation method. Each part together with the smoothed other parts was used as the input, and the corresponding output represents the importance of the image part to the original classifier prediction.

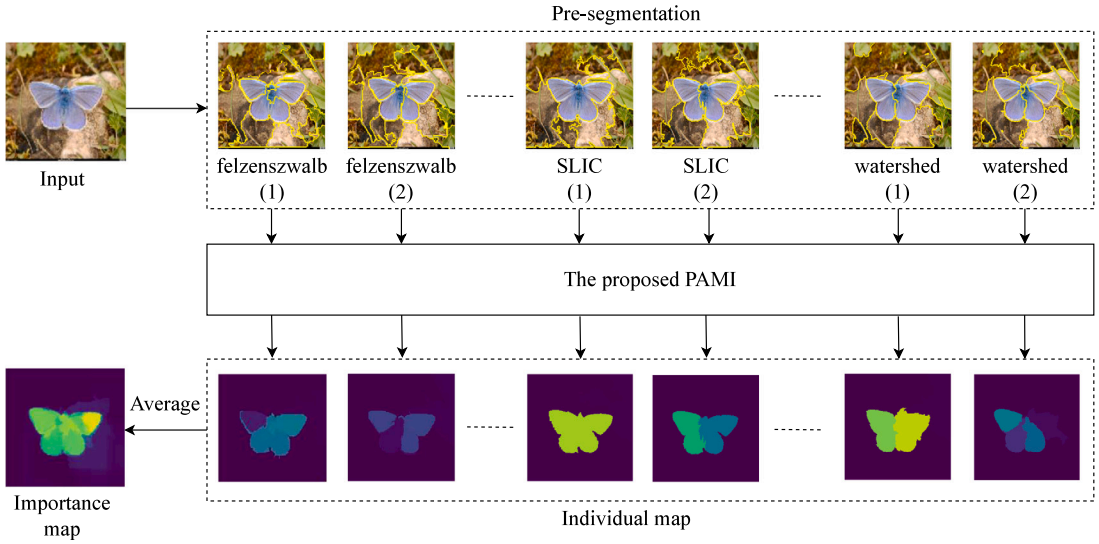

Fig. 4. Multiple segmentation methods with various typical hyperparameter settings are employed to respectively generate the importance map by the proposed PAMI method, and then the multiple importance maps are averaged as the final importance map for the original classifier prediction. The numbers under each method name in pre-segmentation indicate different hyperparameter settings.

regions which contribute to the model prediction. Note that both input partition strategies also work when input is a sequence of items. For example, when input is a sentence as in the sentiment analysis task, any input can be partitioned into words or phrases with either the sliding window strategy or appropriate pre-segmentation strategy.

3.3. Comparison with relevant studies

The proposed PAMI framework can provide interpretation of model prediction without requiring to know model structure and parameter information. In contrast, most existing interpretation methods requires either part of or the whole model details. For example, CAM and its variants need the feature maps from certain convolutional layers and part of model parameters in order to obtain the final class activation map [20], and the gradient-based methods need all the model details to calculate gradient information over model layers [3]. One exception is the occlusion method [21] which does not require model details as the proposed PAMI framework. PAMI can be considered as an opposite version of the occlusion method, i.e., only preserving an input part versus only removing or occluding an input part for estimating the contribution of the input part to the model prediction. In image classification, occluding part of the foreground object in the image may not significantly affect the model prediction because the model can

Table 1
Comparison in characteristics between representative visualization methods and ours. "Gradient*": methods requiring gradient computations [22–24]; "intermediate output": model output from intermediate layer(s); "model parameters": model structure and all parameter values at each layer; "various backbones": different backbones including CNNs and Transformers; "various tasks": different task types including computer vision and natural language processing tasks; "various input formats": different input types including image and word sequence; "High-resolution map": final importance map is in high-resolution; "Interpretation of single region": a method estimates importance of a local region during the interpretation process; "✓": a method has the relevant characteristic; "×": a method has no relevant characteristic;.

Characteristics	Gradient*	LRP [25]	Occlusion [21]	GradCAM [2]	ScoreCAM [20]	RISE [5]	Ours
Not need intermediate output	✓	×	✓	×	×	✓	✓
Not need model parameters	×	×	✓	×	✓	✓	✓
For various backbones	✓	×	✓	×	×	✓	✓
For various tasks	✓	✓	✓	✓	✓	✓	✓
With various input formats	✓	✓	✓	✓	✓	✓	✓
High-resolution map	✓	✓	×	×	×	×	✓
Interpretation of single region	×	×	✓	×	×	×	✓

use the other object parts in the image for confident prediction. As a result, occlusion method may neglect contribution of certain object parts to the original model prediction, and often performs worse than the proposed PAMI.

Because the proposed PAMI framework can consider the well-trained model as a black-box, it can potentially work for various model backbone structures (e.g., both CNN and Transformer backbones). In contrast, the majority of interpretation methods were proposed for the CNN backbone, and specific modifications are often required when applying existing interpretation methods (e.g., CAM or Grad-CAM) to other backbones like Transformer and graph neural networks. Another merit of the proposed PAMI framework is its potential usage in multiple tasks with different input formats. While this study mainly use the image classification task for evaluation of the PAMI framework, PAMI in principle can be applied to various model prediction tasks, such as sentiment analysis and image caption, where the input data can be in the format of sentence or image. In contrast, most existing interpretation methods do not work across tasks without further modifications or extensions.

The most relevant interpretation methods are RISE [5] and ScoreCAM [20] which also estimate the importance map based on linear combination of input masks with weights from model outputs. However, RISE is based on a large set of randomly generated masks and ScoreCAM is based on the feature maps at last convolutional layer of the (CNN) model, both leading to low-resolution and often inappropriate importance maps. It is also worth noting that RISE is an occlusion method and therefore is opposite to the proposed PARMI framework, and ScoreCAM requires part of the model feature maps and therefore cannot be applied to black-box models. Extensive empirical comparisons show that the proposed PAMI framework clearly outperforms RISE and ScoreCAM (also see Figs. 5, 8, 12 and Table 4). More detailed comparisons between the proposed PAMI and representative interpretation methods are summarized in Table 1.

4. Experiments

4.1. Experimental setup

In this study, three image classification datasets, ImageNet-2012 [26], Pascal VOC 2007 [27], and COCO 2014 [28], were mainly used for evaluation of the proposed PAMI method. In addition, an image caption dataset COCO [28] and sentiment analysis dataset Sentiment140 [29] were also employed to show the wide applications of the proposed method. All the models were from the publicly released resources and evaluated on the corresponding validation or test set (see Table 2 for more details).

By default, for the sliding window strategy of the proposed PAMI, circular window with radius 40 pixels and step size 6 pixels was used to generate local image regions. Under this setting, windows are large enough to include regions with discriminative features of objects for most images, and the relatively small stride helps generate importance maps with sufficient resolution. For the pre-segmentation

Table 2
Models and datasets used in experiments.

Task	Model source	Dataset
Classification	VGG19bn from PyTorch [30]	50 000 images of ImageNet-2012 [26] validation set with 1000 classes
	VGG16 from TorchRay [31]	First 1000 images of Pascal VOC 2007 [27] test set with 20 classes
	VGG16 from TorchRay [31]	First 1000 images of COCO 2014 [28] instances validation set with 80 classes
Image caption	ClipCap [32]	COCO 2014 [28]
Sentiment analysis	Transformers library [33]	Sentiment140 [29]

Table 3
Hyper-parameter configuration for each segmentation algorithm.

Method	Hyper-parameter
felzenszwalb [16]	scale = 250, 200, 150, 100, 70, 50; sigma = 0.8; min_size = 784
SLIC [17]	n_segments = 10, 20, 30, 40, 50, 60, 70, 80; compactness = 20
SEEDS [19]	num_superpixels = 10, 20, 30; num_levels = 5; n_iter = 10
Watershed [18]	markers = 10, 20, 30; compactness = 0.0001

strategy, four super-pixel segmentation algorithms, i.e., felzenszwalb [16], SLIC [17], watershed [18] from scikit-image library [34], and SEEDS [19] from the OpenCV library [35] were utilized respectively for pre-segmentation of each image into multiple regions. Considering region of interest (i.e., relevant region to the model prediction) could vary a lot over images, each segmentation algorithm was run multiple times with different hyper-parameter settings to generate sets of local regions at different scales. The hyperparameter settings for these super-pixel segmentation algorithms were listed in Table 3. The importance maps over all hyper-parameter settings and all the four pre-segmentation algorithms were averaged as the estimated importance map. A Gaussian kernel with size 49 × 49 pixels and standard deviation 100 was used to generate the smoothed (blurred) image for region masking.

The proposed PAMI was compared with widely used visualization methods for interpretation of model predictions, including Gradient [22], GradCAM [2], ScoreCAM [20], RISE [5], FullGrad [3], MASK [8], Occlusion [21], GuidedBP [23], SmoothGrad [24] and LRP [25]. The default hyper-parameter setting for each method was adopted. Besides qualitative evaluation, quantitative evaluation was also performed using the pointing game [36] and the insertion metric [5]. In the pointing game, it measures whether the pixel with the highest activation in the importance map is successfully within the image region of the object corresponding to the interpreted class, with 'hit' for success and 'miss' for failure. The average hit rate over all classes is used to measure

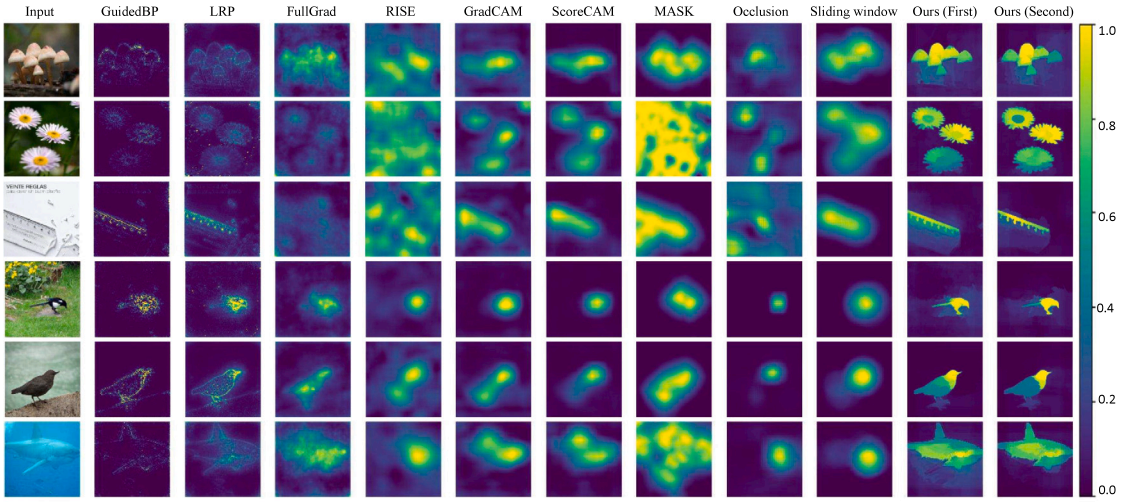

Fig. 5. Qualitative evaluation of the proposed PAMI method on the ImageNet-2012 dataset. The first column lists the input images to the classifier. The last two columns are the importance maps from the proposed PAMI method with the two strategies respectively, and all the other columns are from the representative strong baseline methods. In each importance map or heatmap, higher activation is in yellow and lower activation is in blue. For clear display of relatively different importance at different image regions, we have normalized each importance map such that the min and max importance values are respectively 0 and 1 in each importance map. Therefore, the intensity bar shows the relative importance within each image.

performance of each method. For the insertion metric, it gradually restores the original pixels in the blurred version of the original image, with pixels having higher activation in the importance map restored earlier. Higher insertion score would indicate a better performance of interpretation.

4.2. Qualitative evaluation

The efficacy of the proposed PAMI method was extensively evaluated on the ImageNet-2012 data. Fig. 5 demonstrates the visualization results on multiple representative images with the VGG19bn model. The visualization results were generated with respect to the model output of the ground-truth class for each image. It can be observed that, the proposed PAMI with the pre-segmentation strategy for input partition (last column) can often precisely and largely completely localize the object regions which are actually relevant to the specific model prediction, while existing methods roughly localize either object regions at low resolution, sparse part of object regions, disconnected pixels within object regions, or even irrelevant background regions. Multiple colors within relevant region in the importance map from the proposed PAMI method suggests that different object regions may have different degrees of contributions to the model prediction. On the other hand, the proposed PAMI simply with the sliding window strategy can often obtain similar performance as GradCAM but without requiring model structure and parameter details.

Another observation is that the proposed PAMI method can work more stably than existing methods under challenging conditions. In particular, the PAMI method can well localize small-scale objects (rows 3 & 4) and relatively large-scale objects (row 7) in images, and also can precisely localize the regions of multiple object instances of the same class (rows 1 & 2). In comparison, most existing methods often perform worse under at least some of these challenging conditions. more results from PAMI on the ImageNet dataset are shown in Fig. 6, further verifying the effectiveness of the proposed method. Similar observations were also obtained on the PASCAL-VOC dataset and the COCO dataset (see Fig. 7), consistently supporting that the proposed PAMI method is effective in providing visual evidence for interpretation of model predictions.

4.3. Quantitative evaluation

Although the non-existence of ground-truth or ideal interpretation for any specific model prediction makes it challenging to quantitatively evaluate any interpretation method, the pointing game [36] and the insertion metric [5] have been proposed to roughly evaluate the performance on correctly localizing regions relevant to model predictions. With the pointing game, Table 4 (columns 2, 4, and 6) shows that the proposed PAMI method with the pre-segmentation partition strategy (last row) has similar hitting rate on the ImageNet and COCO datasets compared with the best baseline GradCAM and higher hitting rate than all the baselines on the VOC dataset, suggesting that the local region which is considered most relevant to the model prediction by PAMI is often actually part of the object region. Similarly with the insertion metric (Table 4, columns 3, 5, and 7), PAMI has the best performance on ImageNet and VOC datasets, and is close to the best baseline RISE on COCO dataset, again supporting that PAMI can well localize image regions belonging to the interpreted class. Note that although overall competitive quantitative performance was observed from GradCAM and RISE, the above extensive qualitative evaluations show that the both methods perform worse than the proposed PAMI method in precisely and completely localizing object regions of corresponding model prediction. This also indicates that more appropriate quantitative evaluation metrics should be designed in order to more accurately compare interpretation methods.

4.4. Generality of the proposed PAMI method

To evaluate the generality of the proposed method, well-trained deep learning classifiers with multiple different backbones were employed, including VGG16 [37], ResNet50 [38], SE-ResNet [39], InceptionV3 [40], DenseNet121 [41], RegNet-X-16GF [42], ConvNext-Tiny [43], ViT-L-16 [14] and SwinT-Tiny [15]. From Fig. 8, we can see that the proposed PAMI method can robustly and precisely localize the object regions which are relevant to the model prediction for each input image, regardless of the classifier backbones. In contrast, for each representative baseline method, the importance maps often change over model backbones and even may not work for the Transformer backbone ViT and SwinT. This confirms that the proposed PAMI method is more stable and can be applied to interpretation of model predictions with various model backbones. More results from the proposed PAMI method with different model backbones were shown in Fig. 9.

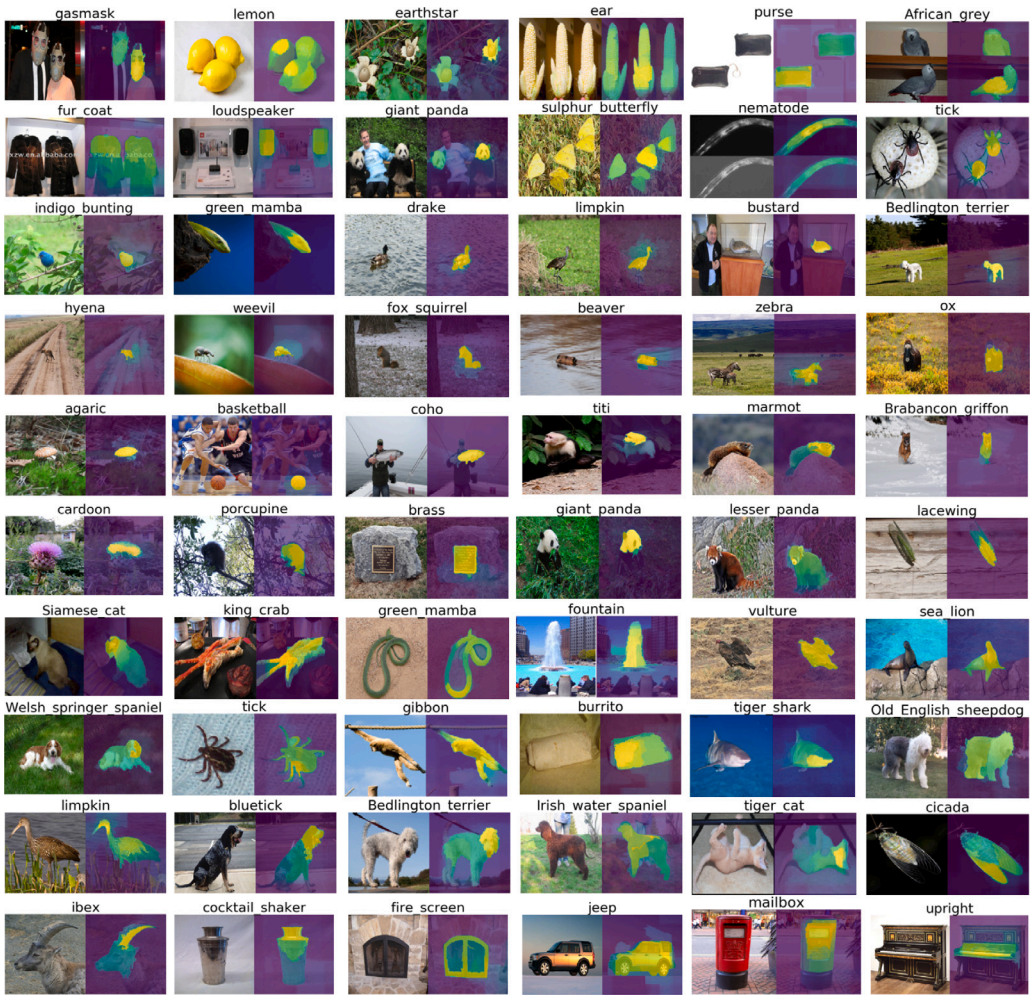

Fig. 6. More qualitative evaluation of the proposed PAMI method on the ImageNet-2012 dataset. For each pair: input image is on the left and the visualization result from the proposed PAMI is on the right.

Table 4
Quantitative evaluation of the proposed PAMI method on the three image datasets. 'Random': randomly generating a heatmap for each input image. 'Center': taking the fixed image center position as the highest activation point for each input image.

Method	ImageNet		VOC		COCO	
	Pointing	Insertion	Pointing	Insertion	Pointing	Insertion
Random	47.89	–	33.39	–	11.27	–
Center	81.96	–	70.79	–	25.97	–
Gradient [22]	83.14	0.1928	72.61	0.3321	34.65	0.1585
GuidedBP [23]	83.95	0.2632	71.14	0.4737	32.83	0.1935
Occlusion [21]	84.53	0.5741	84.49	0.6753	54.83	0.3229
MASK [8]	84.49	0.4867	76.30	0.5616	49.78	0.2664
RISE [5]	91.58	0.5460	82.43	0.6885	56.95	**0.3305**
SmoothGrad [24]	86.51	0.2494	75.38	0.3824	39.45	0.1753
GradCAM [2]	**93.22**	0.5154	87.45	0.5720	**57.95**	0.2660
ScoreCAM [20]	92.01	0.5191	86.51	0.6030	55.01	0.2656
FullGrad [3]	87.01	0.5045	77.58	0.5049	44.52	0.2362
Ours (Strategy I)	89.17	0.5566	74.95	0.6133	48.19	0.2688
Ours (Strategy II)	92.32	**0.5965**	**87.87**	**0.7213**	56.85	0.3291

4.5. Sensitivity and ablation study

In the proposed method, masking majority of the input is one necessary step to estimate the contribution of each single region or part of the input. There are multiple choices for the masking operator. When input is an image, the to-be-masked region could be replaced by constant intensity value such as 0 (i.e., becoming black) or 255 (i.e., becoming white), or by the blurred version of the input image. Fig. 10 demonstrates exemplar results with different masking operators, which shows that masking by blurred region ('Blurred') results in better importance maps with clearer boundaries between background and object regions and more accurate attention to object regions of interest. When the majority of the input is replaced with extreme black or white pixels, the modified input image becomes further from the original class distribution in the feature space compared to the modified image with blurred region, which makes the model prediction unstable and therefore may not faithfully represent the importance of the preserved local region.

In addition, the average of multiple importance maps from multiple pre-segmentation algorithms often results in better visualization than that of using a single pre-segmentation algorithm, as demonstrated in Fig. 11 (columns 2–5 vs. column 6). Fig. 11 also shows that the second run (last column) can often refine the importance map from the first run (column 6). This is probably because sometimes the adopted pre-segmentation algorithms cannot well separate background regions from object regions at the first run, but the initially estimated importance map from the first run provides alternative information for pre-segmentation algorithms to well separate background regions from object regions. Note that the proposed PAMI is independent of the adopted pre-segmentation algorithms, and better pre-segmentation algorithms can be adopted to replace the current ones in the future.

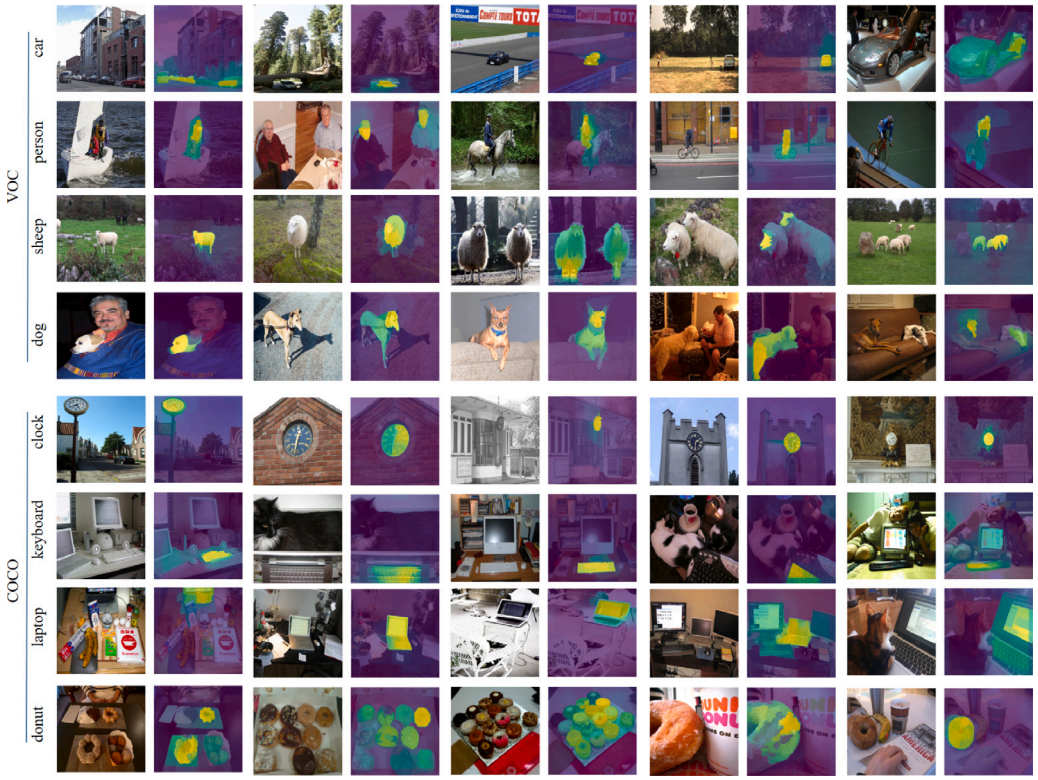

Fig. 7. Qualitative evaluation of the proposed PAMI method on the PASCAL-VOC dataset and the COCO dataset. For each pair: input image is on the left and the visualization result from the proposed PAMI is on the right.

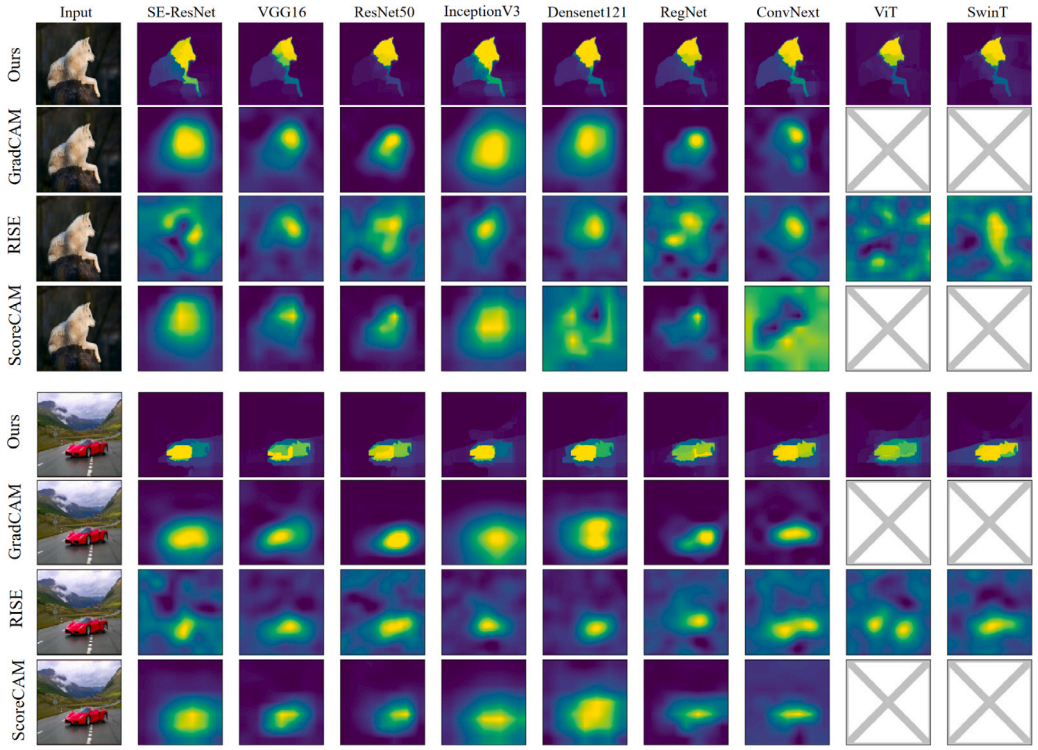

Fig. 8. Representative visualization results from the proposed method and representative baselines with different model backbones. Cross sign means the relevant baseline is not working on the corresponding model backbone.

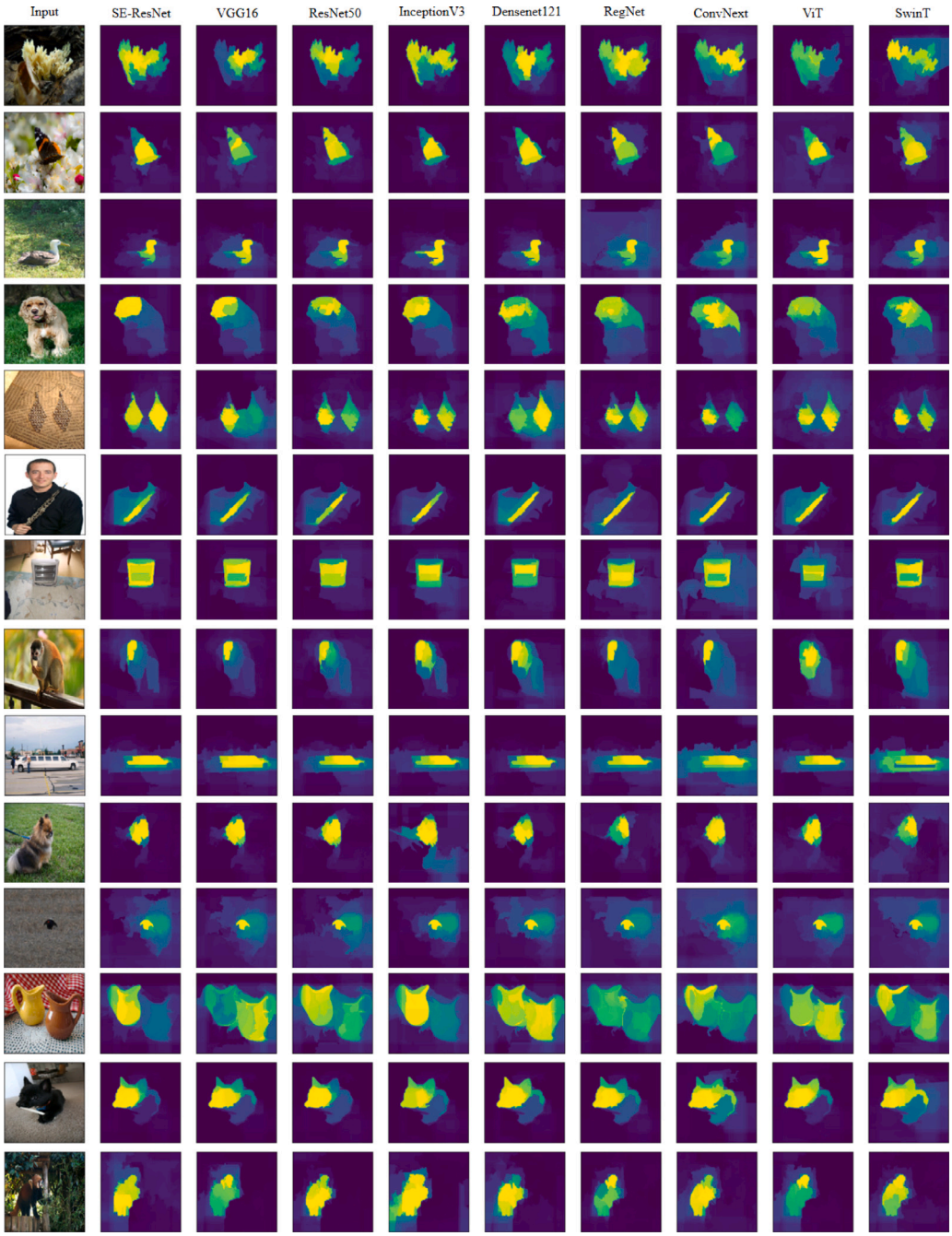

Fig. 9. Representative visualization results from the proposed method with different model backbones.

4.6. Extensive applications of the proposed PAMI

The proposed PAMI method is expected to work for multiple types of prediction tasks. For example, based on a well-trained image caption model [32], the PAMI method can well localize the image regions relevant to the predicted words (e.g., 'dog', 'laying', 'sidewalk', 'bicycle') which refer to certain object or behavior in the image (Fig. 12, rows 1, 3), while the representative baseline method RISE often cannot precisely localize relevant regions (Fig. 12, rows 2, 4). More results from the proposed PAMI method for the image caption task were shown in Fig. 13. Another example is for the sentiment analysis task, where the model tries to evaluate whether the viewpoint in an input sentence is positive or negative. With an input partition strategy similar to the pre-segmentation for images, the proposed PAMI method can directly and correctly estimate the contribution of each word to the final model prediction (Fig. 14). These results confirm that the proposed PAMI method can work for various tasks with different input modalities.

5. Conclusion

In this study, we propose a novel visualization method called PAMI for interpretation of model predictions. PAMI does not requires any model parameter details and works stably across model backbones and input formats. Compared to existing visualization approaches, PAMI can more likely and precisely find the possible local input regions which contribute to the specific model prediction to some degree. It can be used as a plug-in component and applied to multiple types of prediction tasks, which has been partly confirmed by image classification, image caption, and sentiment analysis tasks. It is worth noting that the efficacy of PAMI is partly affected by the input partition

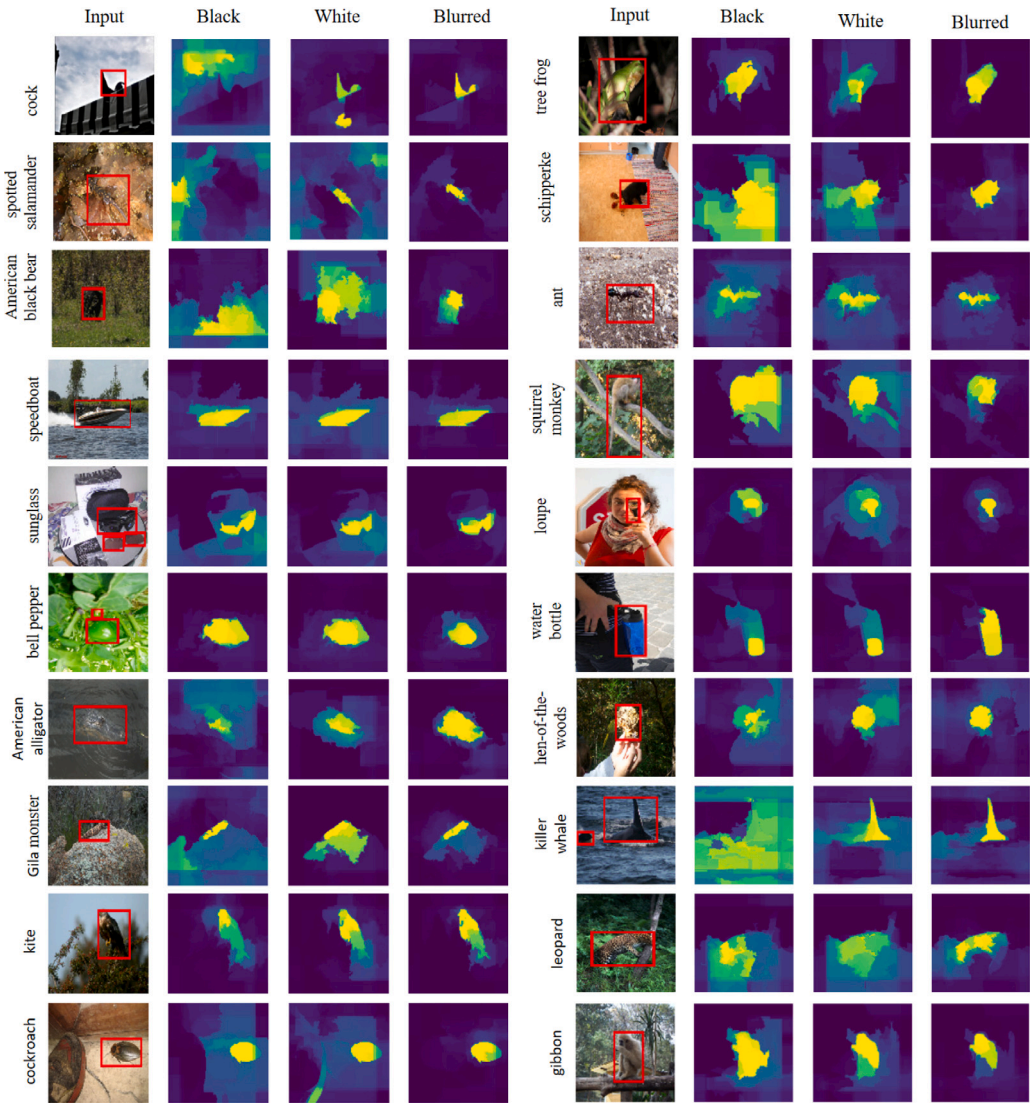

Fig. 10. Importance map of an representative input based on different masking operators. 'Black/White/Blurred': masking types. Red boxes in images: object regions relevant to model predictions.

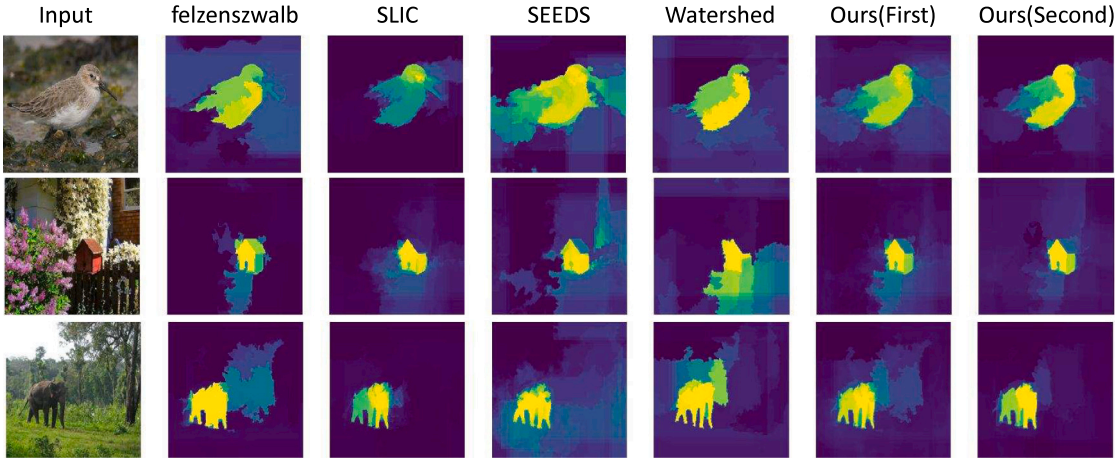

Fig. 11. Effect of applying multiple pre-segmentation algorithms. From left to right: input, importance maps from four individual pre-segmentation algorithms, and average importance maps at the first and the second run respectively.

strategy. When input is an image, the pre-segmentation based strategy often performs better than the sliding window based strategy. However, occasionally object regions may not be clearly pre-segmented from some background regions especially when they are visually similar.

Fig. 12. Two exemplar visualization results from the proposed method and the strong baseline RISE for the image caption task.

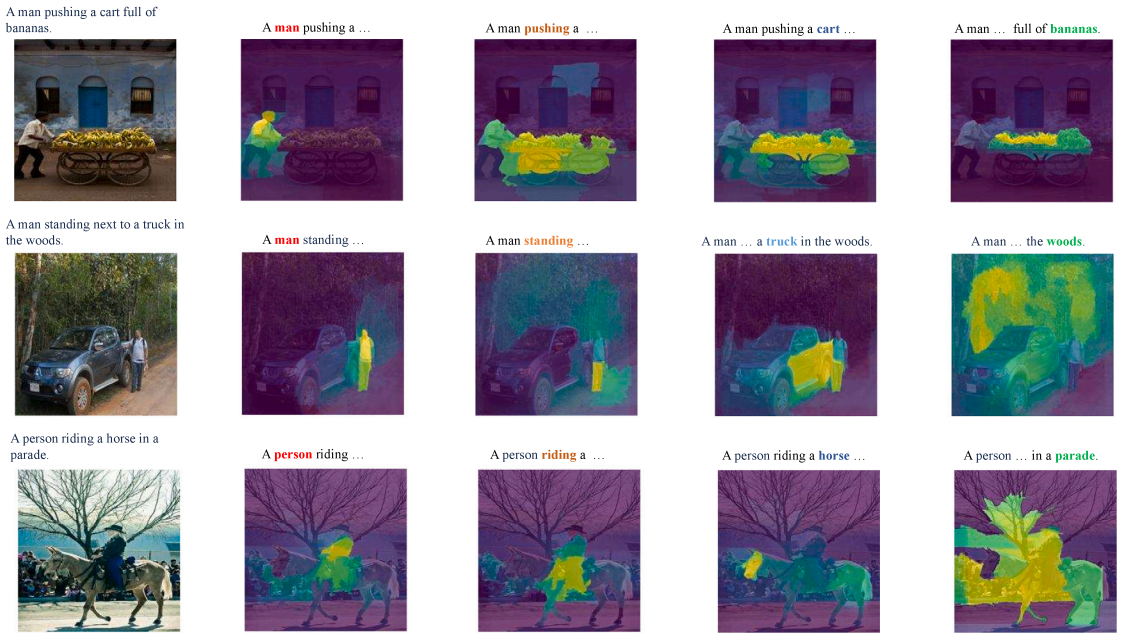

Fig. 13. More visualization results from the proposed method for the image caption task.

Multiple pre-segmentation methods are used together in this study to alleviate this issue of imperfect pre-segmentation. More advanced pre-segmentation methods like SAM (Segment Anything Model) [44] may be employed to obtain better pre-segmented local regions. In addition, the scale range of local regions often needs to be empirically pre-determined in order for PAMI to perform optimally. Automated selection of region scale could be investigated to reduce such manual effort. Alternatively, it may investigate multi-scale strategy to consider multi-scale information at each local region during estimating the contribution of the local region to the original model prediction. The proposed PAMI is only evaluated on three tasks in this study. Its utility is clearly not limited to such tasks. For example, the high-resolution importance map particularly with precise boundaries of object regions can provide much better initial estimate of object location and region, and such better initial estimate should more effectively help solve the weakly supervised (with image-level labels) object detection and region

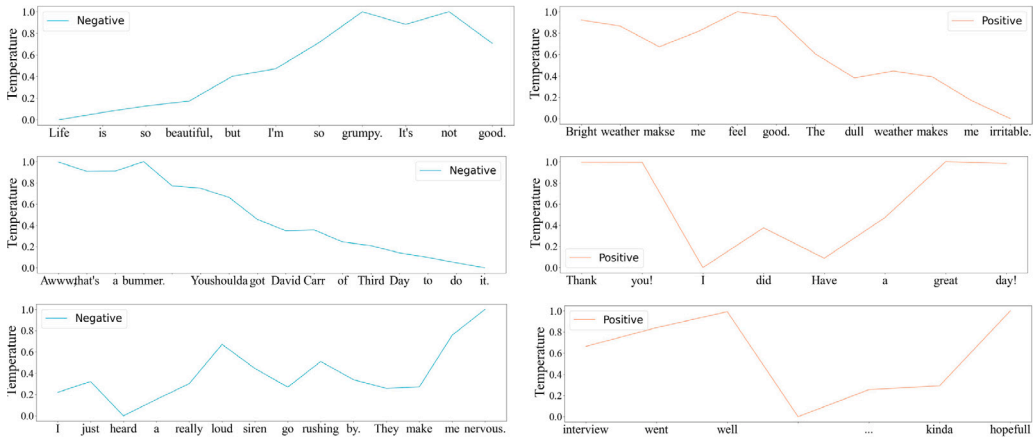

Fig. 14. Exemplar visualization results from the proposed method for the sentiment analysis task. The left column shows the contribution of each word to a negative emotion prediction, and the right column for a positive emotion prediction.

segmentation. More applications of PAMI to various computer vision and natural language processing tasks will be investigated in future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is supported in part by the Major Key Project of PCL (grant No. PCL2023AS7-1), the National Natural Science Foundation of China (grant No. 62071502), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

References

[1] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, Been Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognit. 120 (2021) 108102.

[2] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv 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.

[3] Suraj Srinivas, François Fleuret, Full-gradient representation for neural network visualization, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 4126–4135.

[4] Brian Kenji Iwana, Ryohei Kuroki, Seiichi Uchida, Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation, in: Proceedings of the IEEE International Conference on Computer Vision Workshop, 2019, pp. 4176–4185.

[5] Vitali Petsiuk, Abir Das, Kate Saenko, RISE: Randomized input sampling for explanation of black-box models, in: Proceedings of the British Machine Vision Conference, 2018.

[6] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[7] Satya M. Muddamsetty, Mohammad N.S. Jahromi, Andreea E. Ciontos, Laura M. Fenoy, Thomas B. Moeslund, Visual explanation of black-box model: Similarity Difference and Uniqueness (SIDU) method, Pattern Recognit. 127 (2022) 108604.

[8] Ruth C. Fong, Andrea Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.

[9] Jessica Cooper, Ognjen Arandjelović, David J. Harrison, Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping, Pattern Recognit. 129 (2022) 108743.

[10] Srishti Gautam, Marina M.-C. Höhne, Stine Hansen, Robert Jenssen, Michael Kampffmeyer, This looks more like that: Enhancing self-explaining models by prototypical relevance propagation, Pattern Recognit. 136 (2023) 109172.

[11] Abraham Montoya Obeso, Jenny Benois-Pineau, Mireya Saraí García Vázquez, Alejandro Álvaro Ramírez Acosta, Visual vs internal attention mechanisms in deep neural networks for image classification and object detection, Pattern Recognit. 123 (2022) 108411.

[12] Steven E. Petersen, Michael I. Posner, The attention system of the human brain: 20 years after, Annu. Rev. Neurosci. 35 (2012) 73.

[13] Robert Desimone, John Duncan, et al., Neural mechanisms of selective visual attention, Annu. Rev. Neurosci. 18 (1) (1995) 193–222.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 10012–10022.

[16] Pedro F. Felzenszwalb, Daniel P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 167–181.

[17] Xiaofeng Ren, Jitendra Malik, Learning a classification model for segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 10–17.

[18] Fernand Meyer, Color image segmentation, in: International Conference on Image Processing and its Applications, 1992, pp. 303–306.

[19] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, Luc Van Gool, Seeds: Superpixels extracted via energy-driven sampling, in: European Conference on Computer Vision, 2012, pp. 13–26.

[20] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, Xia Hu, Score-CAM: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2020, pp. 24–25.

[21] Matthew D. Zeiler, Rob Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, 2014, pp. 818–833.

[22] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, in: International Conference on Learning Representations, 2014.

[23] J. Springenberg, Alexey Dosovitskiy, Thomas Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: International Conference on Learning Representations Workshop, 2015.

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, Martin Wattenberg, SmoothGrad: removing noise by adding noise, 2017, arXiv preprint arXiv:1706.03825.

[25] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, Klaus-Robert Müller, Explaining nonlinear classification decisions with deep Taylor decomposition, Pattern Recognit. 65 (2017) 211–222.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[27] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.
[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[29] Alec Go, Richa Bhayani, Lei Huang, Twitter sentiment classification using distant supervision, 2009, http://help.sentiment140.com/home.
[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.
[31] Ruth Fong, Mandela Patrick, Andrea Vedaldi, TorchRay, 2019, https://github.com/facebookresearch/TorchRay.
[32] Ron Mokady, Amir Hertz, Amit H. Bermano, ClipCap: CLIP prefix for image captioning, 2021, arXiv preprint arXiv:2111.09734.
[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[34] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, the scikit-image contributors, scikit-image: image processing in Python, PeerJ 2 (2014) e453.
[35] G. Bradski, The OpenCV library, Dr. Dobb's J. Softw. Tools (2000).
[36] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Stan Sclaroff, Top-down neural attention by excitation backprop, Int. J. Comput. Vis. 126 (10) (2018) 1084–1102.
[37] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[39] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[41] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[42] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, Piotr Dollár, Designing network design spaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10428–10436.
[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[44] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al., Segment anything, 2023, arXiv preprint arXiv:2304.02643.

Wei Shi obtained relevant Bachelor degree from Chongqing University, He is currently a postgraduate student at Sun Yat-sen University. His research interests include robustness and interpretability in deep learning.

Wentao Zhang received the B.S. degree and the M.S. degree, in 2017 and 2020, respectively. He is currently pursuing Ph.D. degree with Sun Yat-sen University. His research interests include computer vision, medical image analysis, and deep learning.

Wei-Shi Zheng received the Ph.D. degree in applied mathematics from Sun Yat-sen University. He is currently a full Professor with Sun Yat-sen University. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm.

Ruixuan Wang obtained Ph.D. degree from National University of Singapore, followed by the participation of multiple interesting AI-relevant research programs mainly in University of Dundee. He is currently a professor with Sun Yat-sen University. His research group uses healthcare applications as the engine, driving the exploration of novel AI techniques.