Improving Interpretability by Information Bottleneck Saliency Guided Localization

Hao Zhou
zhouhao@stmail.ujs.edu.cn
Keyang Cheng†
kycheng@ujs.edu.cn
Yu Si
siyu@stmail.ujs.edu.cn
Liuyang Yan
liuyangyan@stmail.ujs.edu.cn

School of Computer Science and Communication Engineering
Jiangsu University
Zhenjiang, China

† corresponding author

Abstract

The saliency map produced by current deep neural network models fails to accurately focus on important regions of an image due to the influence of input noise. In this paper, we propose a deep learning interpretability method based on information bottleneck, which guides the model training by the probability distribution between the saliency map attributed by the information bottleneck and the gradient-based saliency map. This approach corrects the important regions focused by the model from an information-theoretic perspective. Meanwhile, a saliency suppression mechanism is presented to keep the saliency map of the model away from incorrect classification results and close to correct ones. Experiments show that our method can improve the saliency localization of the model while retaining its accuracy. Compared with other state-of-the-art methods, the Average Drop rate improves by 1.57% and 1.43%, and the Average Increase rate improves by 2.18% and 0.18% in the ResNet-50 model and the VGG-16 model, respectively.

1 Introduction

Deep Neural Networks (DNNs) have achieved superior performance in many real-world applications. The superior representational learning capabilities of DNNs have been demonstrated in a variety of disciplines, including deep reinforcement learning and neural machine translation. However, deep learning still has some significant disadvantages. As a complicated model with millions of free parameters, DNNs often exhibit unexpected behaviors.

To interpret deep neural networks, especially CNNs with visual inputs, visual saliency can be used to highlight important features that contribute to model prediction. However, current interpretable methods produce saliency maps that are often noisy or do not match human knowledge. Neural networks are trained using datasets can converge well, but the neural network only extracts features from the background of the image instead of the foreground, due to the noise and uncertainty in the dataset. Given an image of an airplane, for

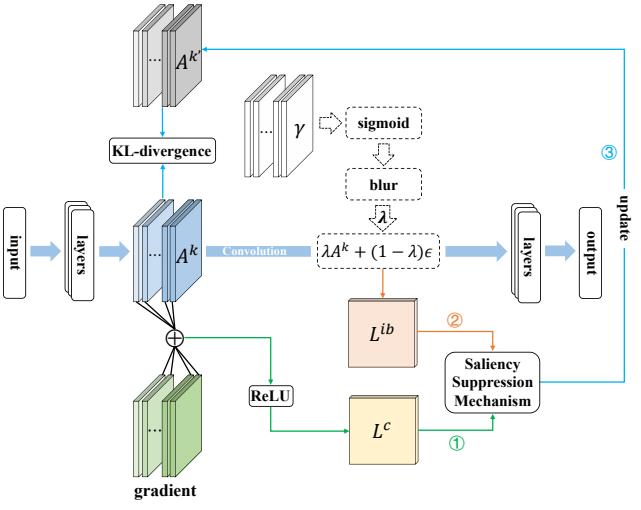

Figure 1: The proposed overview model for information bottleneck saliency-guided localization. step ① implements a gradient-based saliency map. step ② implements a saliency map based on information bottleneck attribution. step ③ implements an information bottleneck saliency map is used to guide model training.

example, the saliency map might highlight the blue sky rather than the airplane itself. Therefore, interpretable methods based on the saliency map must not only be able to produce a saliency map, but also to correct the focus of the saliency map.

In this paper, an interpretable method for saliency-guided localization based on information bottleneck is proposed to improve the focusing ability of convolutional neural networks. The contributions of our method are as follows:

- A novel interpretable method based on information bottleneck saliency-guided localization is proposed, which modifies the saliency map of the model to improve interpretability from the perspective of information theory.
- We propose a saliency suppression mechanism that constrains the focus between ground-truth and non-ground truth saliency maps to reduce saliency from non-ground truth classes.

2 Related Work

Information Bottleneck. The information Bottleneck [18] is an information-theoretic based data analysis method, which treats the pattern extraction from data as a process of data compression. Tishby *et al*. [17] and Shwartz *et al*. [14] proposed the use of information bottleneck theory to analyze decision-making within deep neural networks. Achille *et al*. [1] and Dubois *et al*. [3] used information bottleneck theory to obtain optimal representations in deep neural networks. Jeon *et al*. [7] studied decomposed representation learning for generative models by information bottleneck theory. DICE [11] asserts that to use information bottleneck theory effectively, it is vital to eliminate the redundant information exchanged between

features produced by independently trained DNNs and the needless redundancy between features and inputs. Lee *et al*. [9] argued that most of the information bottlenecks occur at the last layer of the network and reduce this information bottleneck by modifying the training scheme. Schulz *et al*. [12] restricted information transmission in the model by introducing noise to the intermediate feature maps before quantifying the information present in the images. Most of the above information bottleneck methods are used only for model analysis but are not applied to model training and manipulation.

Visualizing CNN. The Class Activation Mapping (CAM) [21] approach modifies the architecture of convolutional neural networks by replacing the fully connected layers with convolutional and average pooling layers to achieve class-specific saliency maps. CAM is a classical and intuitive method that satisfies the requirements of the benchmark in terms of faithfulness [5, 8, 20]. Grad-CAM [13] flows the gradient of a specific class to each feature graph and then uses the average gradient as the weight. Grad-CAM++ [2] considers that each element on the gradient map contributes differently, so additional weight is added to the weight of the elements on the gradient map. Score-CAM [19] does not need the gradient, but generates the weight for each feature graph through its forwarding score. Ismail *et al*. [6] used saliency-guided training to reduce the noise gradient by repeatedly masking input features of low gradient values. The above CAM methods generate saliency maps from the parameters of the model itself, and its main purpose is to understand the attention region of the model. However, the attention region of the model is not always optimal, so it is necessary to be able to modify it.

3 Method

We design a salience-guided training method based on information bottleneck. The overview model is shown in Figure 1. First, the saliency map is derived using a gradient-based approach, and the gradient information flowing into the last convolutional layer of the CNN is used to assign importance values to specific decisions of each neuron. The saliency map is derived from this importance value. Second, the quantification of the forward pass information flow [12] over the network is done. Noise is injected into the feature map of the pre-trained model, thus suppressing the information flow through it. The intensity of the noise is then optimized to minimize the information flow while maximizing the classification score of the original model. The saliency map is derived from the distribution of the noise. Last, the loss between the information bottleneck and the gradient-based saliency map is computed using our proposed saliency suppression mechanism to update the entire convolutional neural network.

3.1 Saliency Suppression Mechanism

The CAM approach produces class-specific saliency maps in the model, so that a network will have multiple saliency maps and different classes will correspond to different saliency maps. For this reason, a saliency suppression mechanism is proposed, where the focus of other classes is suppressed in the saliency map of the target class, and the focus of different classes of saliency maps is separated.

The approach in this paper uses saliency maps to guide the training and localization of the model, so a loss function between saliency maps is necessary. Therefore, we propose a new learning objective that incorporates the discrepancy between saliency maps as part of

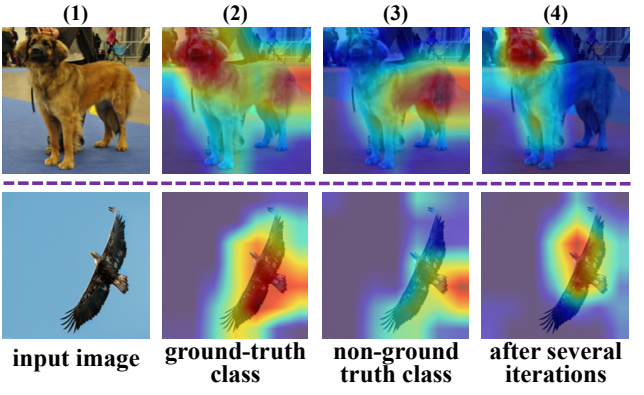

Figure 2: (1) is the input image. (2) is the saliency map of the ground-truth (target) class. (3) is the saliency map of the non-ground truth (highest probability) class. (4) is the saliency map of the ground-truth class after several iterations of saliency suppression mechanism.

the learning process. The saliency map L^c of the ground-truth class c and the saliency map L^p of the class p with the highest probability are given, where the saliency map L^p comes from the non-ground truth class with the highest classification probability. We want to keep the saliency map of the ground-truth classes away from the saliency map of the non-ground truth. The saliency suppression loss $\mathcal{L}_{SS}$ is shown in Equation 1:

$$\mathcal{L}_{SS}\left(L^c, L^p\right) = \frac{\sum_{ij}\left[\min\left(L^c, L^p\right) \cdot Mask_r\right]}{\sum_{ij}\left(L^c + L^p\right)} \tag{1}$$

where the $(i,j)^{th}$ is the pixel in saliency map, and the "$\cdot$" indicates scalar product.

In addition, to reduce the noise from the saliency map of the non-ground truth class, it is used as a mask to suppress the focus in the ground-truth class of saliency maps. The $Mask_r$ in the Equation 1 represents the target object region generated by threshold the saliency map L^c of the ground-truth class. Define $Mask_r = M_r\left(S(L^c)\right)$, where the $S(L^c)$ operation represents the ranking of the saliency map L^c and the $M_r(\cdot)$ operation takes 1 for the first $r \in (0,1)$ pixel and 0 for the others.

Visualization results of saliency suppression loss are shown in Figure 2. The model has saliency maps for each class, the most representative of which are the ground-truth class and the highest probability (non-ground truth) class. The overlap between the saliency maps of the ground-truth class and the non-ground truth class is reduced by using saliency suppression mechanism.

3.2 Information Bottleneck Guided Localization

Consider a classification problem on the input data $\{(X_i, y_i)\}_{i=1}^n$, X_i is the input image, y_i is the label. Let f_θ denote a neural network with θ as parameters. The network is trained to minimize the cross-entropy loss $\mathcal{L}_{CE}$ on the training set as follows:

$$\underset{\theta}{\text{minimize}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{CE}\left(f_\theta\left(X_i\right), y_i\right) \tag{2}$$

Saliency Map Based On The Gradient. To obtain the saliency map L^c for the class c of the input image, the gradient-based class activation mapping method is applied. Let f_{conv} denote the convolutional layer of the model. We compute the gradient of the score y^c of the class c with respect to feature map activations $A^k = f_{conv}(X)$ of a convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k}$. The gradient $\frac{\partial y^c}{\partial A^k}$ captures the "importance" of the feature map A^k for the target class c. The saliency map L^c is derived from a linear combination of feature maps and gradients followed by ReLU:

$$L^c = \text{ReLU}\left(\sum_k \frac{\partial y^c}{\partial A^k} A^k\right) \tag{3}$$

The ReLU operation retains the features that have a positive impact on a particular class. Unlike other gradient-based algorithms, this method does not use a global averaging pool because we do not want to blur the importance of the feature maps, but rather use pixel-level gradient importance.

Saliency Map Based On Information Bottleneck Attribution. The information bottleneck attribution method of Schulz *et al*. [12] is used and injected into the pre-trained network. The information in A^k is reduced by adding noise. The signal A^k is damped when noise is added, replacing the part signal with the noise. Linear interpolation between the signal and the noise is applied to obtain the variable N:

$$N = \lambda A^k + (1-\lambda)\varepsilon \tag{4}$$

where $\varepsilon \sim \mathcal{N}\left(\mu_{A^k}, \sigma^2_{A^k}\right)$ and $\lambda = \text{blur}(\sigma, \text{sigmoid}(\gamma))$ controls the signal damping and noise adding. The γ contains each element of the corresponding A^k. The parameter γ controls the amount of information that is conveyed to the next layer. According to the Equation 8, γ is optimized individually for each sample. The sigmoid(γ) operation allows γ to freely choose the size, which is restricted to $[0,1]$ during the optimisation process. The blur$(\sigma,\cdot)$ operation convolves the sigmoid output with a fixed Gaussian kernel of standard deviation σ to obtain a robust and smooth attribution graph.

If a region contains information that is useful for classification, it is considered relevant. So we evaluate how much information about A^k is contained in N. This quantity is the mutual information $I[A^k, N]$, and can be shown as:

$$\text{I}[A^k, N] = \mathbb{E}_{A^k}\left[D_{KL}[P(N \mid A^k) \| P(N)]\right] \tag{5}$$

where $P(N \mid A^k)$ and $P(N)$ denote the probability distributions, respectively. The mutual information cannot be computed exactly, so the variational approximation $Q(N) = \mathcal{N}\left(\mu_{A^k}, \sigma^2_{A^k}\right)$ is used, where all dimensions of N are assumed to be normally distributed and independent, since the activation after linear or convolution is usually Gaussian distributed. Substitute $Q(N)$ into the Equation 5:

$$\text{I}[A^k, N] = \mathbb{E}_{A^k}\left[D_{KL}[P(N \mid A^k) \| Q(N)]\right] - D_{KL}[Q(N) \| P(N)] \tag{6}$$

The KL-divergence between the first and second normal distributions is contained in the first term, which makes evaluation simple. Mutual information is approximated by the first KL-divergence term. Thus, the loss function $\mathcal{L}_1$ is:

$$\mathcal{L}_I = \mathbb{E}_{A^k}\left[D_{KL}[P(N \mid A^k) \| Q(N)]\right] \tag{7}$$

We only retain the information needed to classify correctly. Therefore, there should be little mutual information, but the classification score should be high. Let $\mathcal{L}_{II}$ be the cross-entropy of the classification and we get the loss function:

$$\mathcal{L} = \mathcal{L}_{II} + \alpha \mathcal{L}_{I} \tag{8}$$

where the parameter α balances the importance of the two optimization objectives. We evaluate the $D_{\mathrm{KL}}(P(N \mid A^k) \| Q(N))$ of per dimension to measure the importance of each feature in N. The saliency map L^{ib} is obtained by summing over the channel axis c:

$$L^{ib}_{[h,w]} = \sum_{j=0}^{c} D_{KL}\left(P\left(N_{[j,h,w]} \mid A^k_{[j,h,w]}\right) \| Q\left(N_{[j,h,w]}\right)\right) \tag{9}$$

where $[h,w]$ is the corresponding height and width.

Algorithm 1 Saliency Guided Localization algorithm

Input: Image X_i, Class c, Sample size n of data set
Output: Model after saliency-guided localization

Initialization: Let A^k be the feature map of the last convolutional layer. y^c and y^p are the prediction scores for the target classes c and p, respectively.
for i in $[0,...,n-1]$ **do**
  Get the saliency map of class c and class p,
  $L^c, L^p \leftarrow \mathrm{ReLU}\left(\sum_k \frac{\partial y^c}{\partial A^k} A^k\right), \mathrm{ReLU}\left(\sum_k \frac{\partial y^p}{\partial A^k} A^k\right)$
  Saliency map L^{ib} is given based on information bottleneck attribution.
  if $c = p$ **then**
    $\underset{\omega}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{L}_{CE}\left(L^c, L^{ib}\right) + \beta \mathcal{L}_{SS}\left(L^c, L^p\right)\right]$
  else
    $\underset{\omega}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{CE}\left(L^c, L^{ib}\right)$
  end if
  Update $A^k \rightarrow A^{k'}$ according to the above loss function.
  $\underset{\theta}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{L}\left(f_\theta\left(X_i\right), y_i\right) + \mu D_{KL}\left(A^k \| A^{k'}\right)\right]$
  Update model parameters.
end for

Saliency Guided Localization. In addition to calculating the cross-entropy loss $\mathcal{L}_{CE}$ between the saliency map L^c and L^{ib} attributed to the information bottleneck, the saliency suppression loss between L^c and L^p is minimized and A^k is updated to obtain $A^{k'}$:

$$\underset{\omega}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{L}_{CE}\left(L^c, L^{ib}\right) + \beta \mathcal{L}_{SS}\left(L^c, L^p\right)\right] \tag{10}$$

where $\omega = \frac{\partial y_i^c}{\partial A_i^k}$, L^p is the saliency map of the non-ground truth class with the highest classification probability, and β is the hyperparameter used to weigh L^c and L^p. Making feature graphs continuously learn features that have large contributions to the salience map of information bottleneck.

The saliency-guided localization minimizes the KL-divergence between A^k and $A^{k'}$, and the classification scores were retained. Therefore, the optimization problem for the guided

localization is:

$$\underset{\theta}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{L}\left(f_\theta\left(X_i\right), y_i\right) + \mu D_{KL}\left(A^k \| A^{k'}\right) \right] \tag{11}$$

where μ is a hyperparameter that trade-offs the KL-divergence term and the cross-entropy classification loss. The KL-divergence term promotes the model to produce similar results for the original feature map A^k and the updated feature map $A^{k'}$. The Saliency Guided Localization algorithm is shown as Algorithm 1.

4 Experiments

In this section, the experimental setup is firstly introduced by describing the data set, the model, and the parameter settings used for the experiments. Secondly, experiments with guided localization are conducted, comparing the before-guided and after-guided saliency maps. Finally, the model is evaluated quantitatively and the reliability of the saliency maps is illustrated by continually reducing the region of significance.

4.1 Experimental Settings

These datasets are used in our experiments: (1) **ILSVRC2012** is a subset of a large hand-labeled ImageNet dataset organized according to the WordNet hierarchy. (2) **PASCAL VOC 2007** is a target detection dataset containing 4952 test images from 20 different output classes. The presence of multiple targets in the dataset makes interpretation more challenging.

In our work, two different networks are used. The first model is the **Resnet-50** network and the second is the **VGG-16** network. Both models take as input images of size $224 \times 224 \times 3$, so all images will be resized before being input into the model.

All experiments have been carried out with $r = 0.35$, $\alpha = 1000$, $\gamma = 5$, $\sigma = 1$, $\beta = 1$ and $\mu = 2$.

4.2 Guided Localization

The experiments indicate that this method can then change the important region of the model without reducing the overall accuracy of the model, making the focus of the model more consistent with information theory.

The Figure 3 shows some saliency guided localization experiments of the ResNet-50 model trained on the ILSVRC2012 dataset. The input image is shown on the left, the before-guided saliency map is shown in the middle, and the after-guided saliency map is shown on the right. This approach allows the model's saliency map to converge on the saliency map of the information bottleneck and away from the saliency map of the non-ground truth. As can be seen from the experimental figure, the guided saliency maps are more consistent with human knowledge and have improved classification probabilities for the target classes, as well as improving the performance of the model. For example, in the third row of images on the right half, the focus region of the saliency map before the guidance is on seagrass, the focus region of the saliency map after the guidance is shifted to fish, and the prediction probability of the category fish is also improved.

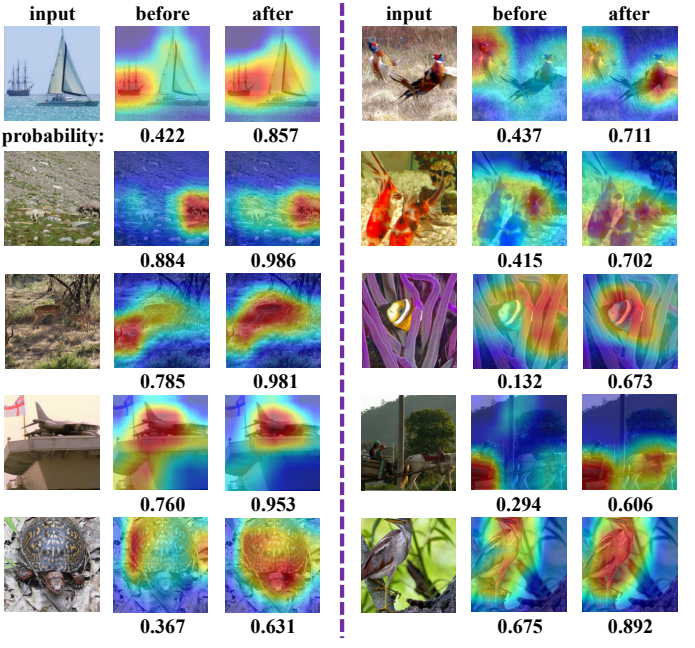

Figure 3: The model's saliency map is changed through saliency-guided localization to make it easier to understand.

4.3 Quantitative Evaluations

The accuracy of the location evaluation of salience is first measured by objective quantification. We deploy ground truth-based metrics, including **Energy-based Pointing game (EBPG)**, **mean Intersection over Union(mIoU)** and **Bounding box (Bbox)**, to assess the capability of our method in terms of accurate object localization and feature visualization compared to the baseline approaches.

Metric	EBPG	mIoU	Bbox
Grad CAM [13]	60.08	32.16	**60.25**
Grad CAM++ [2]	47.78	30.16	58.66
Extremal Perturbation [4]	63.24	26.29	52.34
RISE [10]	32.86	27.40	55.55
Score CAM [19]	35.56	31.0	60.02
Integrated Gradient [16]	40.62	15.41	34.79
FullGrad [15]	39.55	20.20	44.94
Ours method	**65.07**	**32.74**	58.88

Table 1: Results of the state-of-the-art methods compared with ours method on ResNet-50 model.

Metric	EBPG	mIoU	Bbox
Grad CAM [13]	55.44	26.52	51.70
Grad CAM++ [2]	46.29	28.10	**55.59**
Extremal Perturbation [4]	**61.19**	25.44	51.20
RISE [10]	33.44	27.11	54.59
Score CAM [19]	46.42	27.71	54.98
Integrated Gradient [16]	36.87	14.11	33.97
FullGrad [15]	38.72	26.61	54.17
Ours method	58.78	**28.32**	55.12

Table 2: Results of the state-of-the-art method compared with our method on the VGG-16 model.

Table 1 shows the results of the state-of-the-art interpretable methods and the interpretable evaluation metrics of our method on the ResNet-50 model trained on the PASCAL VOC 2007 dataset. Table 2 shows the results on VGG-16. For each metric, the best is shown

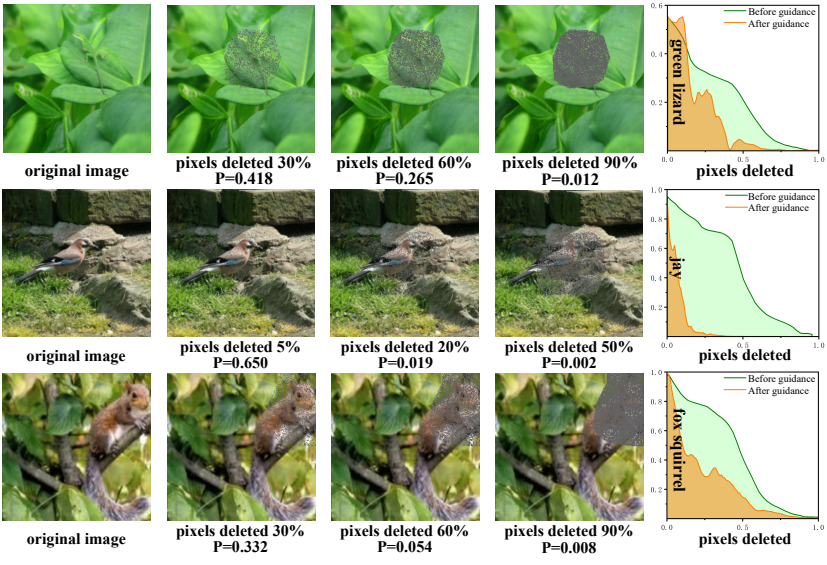

Figure 4: Effect of deleting important regions of images with results.

in bold and the second best is underlined. All values are reported as percentages. Our method achieves excellent results for different metrics and models, indicating its robustness.

Besides, the saliency maps reliability in the decision area is verified. We measure the "**Average Drop (AD)**" and "**Average Increase (AI)**". AD represents the maximum positive difference between the prediction using the input image and the prediction by the saliency map mask. The lower the drop in the AD score is, the more reliable the model will be. AI indicates that the saliency map yields a higher score. A higher score indicates a more reliable interpretation of the model generation. The faithfulness of the interpretation method is evaluated by observing the behavior of the model by inputting only the features that are represented as significant by the interpretation algorithm.

Metric	AD(%)	AI(%)
Grad CAM [13]	35.80	36.58
Grad CAM++ [2]	41.77	32.15
Extremal Perturbation [4]	39.38	34.27
RISE [10]	39.77	<u>37.08</u>
Score CAM [19]	<u>35.36</u>	<u>37.08</u>
Integrated Gradient [16]	66.12	24.24
FullGrad [15]	65.99	25.36
Ours method	**33.79**	**39.26**

Table 3: Results of the state-of-the-art methods compared with our method on ResNet-50 model.

Metric	AD(%)	AI(%)
Grad CAM [13]	49.47	31.08
Grad CAM++ [2]	60.63	23.89
Extremal Perturbation [4]	43.90	32.65
RISE [10]	<u>39.62</u>	<u>37.76</u>
Score CAM [19]	39.79	36.42
Integrated Gradient [16]	64.74	26.17
FullGrad [15]	60.78	22.73
Ours method	**38.19**	**37.94**

Table 4: Results of the state-of-the-art methods compared with our method on VGG-16 model.

In Table 3 and Table 4, the proposed method in this paper has an AD rate of 33.79% and an AI rate of 39.26% on the ResNet-50 model. The AD rate is 1.57% better than the

other methods and the AI rate is 2.18% better than the other methods. The AD rate on the VGG-16 model is 38.19% and the AI rate is 37.94%. The AD rate is 1.43% better than the other methods and the AI rate is 0.18% better than the other methods.

Part of the pixels in the important regions of our saliency maps are continuously and randomly deleted from the corresponding input images. From Figure 4, we can find that for images with partial pixel deletion, the prediction probability slowly decreases with the percentage of deletion when predicted by the model without guided localization. And the prediction probability decreases sharply when predicted by the model after guided localization. Guided localization makes the important regions of the saliency map more accurate.

5 Conclusion

In this paper, an interpretable information bottleneck saliency-guided localization method is proposed to guide model training based on information bottleneck and to improve its localization capability without degrading accuracy. In addition, a saliency suppression mechanism is introduced to suppress the saliency maps of ground-truth classes away from non-ground truth classes. The feature maps are updated by minimizing the loss between the saliency map of the information bottleneck and the saliency map of the gradient, and the KL-divergence between the feature maps is calculated to update the parameters. Our method does not restrict the structure of the network and supports any activation function and network structure. Experiments demonstrate that the synthetic saliency of our proposed method outperforms that of state-of-the-art methods.

Acknowledgement

This work was supported by National Natural Science Foundation of China [61972183] and Jiangsu Province Science and Technology Program [BE2022781].

References

[1] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[3] Yann Dubois, Douwe Kiela, David J Schwab, and Ramakrishna Vedantam. Learning optimal representations with the decodable information bottleneck. *Advances in Neural Information Processing Systems*, 33:18674–18690, 2020.

[4] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[5] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[6] Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Insu Jeon, Wonkwang Lee, Myeongjang Pyeon, and Gunhee Kim. Ib-gan: Disentangled representation learning with information bottleneck generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7926–7934, 2021.

[8] Ashkan Khakzar, Pedram Khorsandi, Rozhin Nobahari, and Nassir Navab. Do explanations explain? model knows best. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10244–10253, 2022.

[9] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021.

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

[11] Alexandre Rame and Matthieu Cord. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. *arXiv preprint arXiv:2101.05544*, 2021.

[12] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

[13] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[14] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[15] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

[16] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[17] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

[18] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[19] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[20] Yang Zhang, Ashkan Khakzar, Yawei Li, Azade Farshad, Seong Tae Kim, and Nassir Navab. Fine-grained neural network explanation by identifying input features with predictive information. *Advances in Neural Information Processing Systems*, 34:20040–20051, 2021.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.