17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Attention Mining Branch for Optimizing Attention Map

Takaaki Iwayoshi
Chubu University
Address
iwayoshi@mprg.cs.chubu.ac.jp

Masahiro Mitsuhara
Chubu University
Address2
mitsuhara@mprg.cs.chubu.ac.jp

Masayuki Takada
Chubu University
Address3
mosa@mprg.cs.chubu.ac.jp

Tsubasa Hirakawa
Chubu University
Address4
hirakawa@mprg.cs.chubu.ac.jp

Takayoshi Yamashita
Chubu University
Address5
takayoshi@isc.chubu.ac.jp

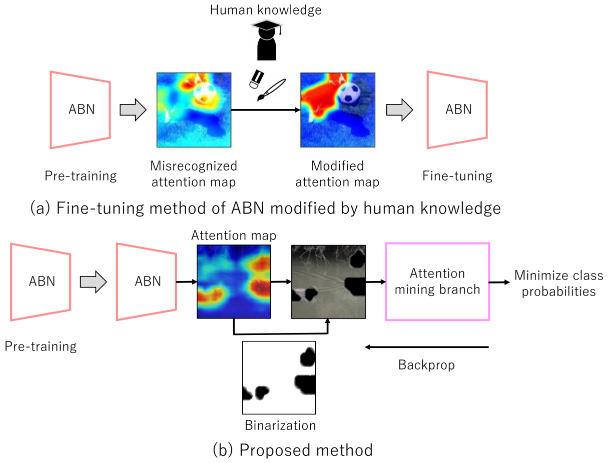
Hironobu Fujiyoshi
Chubu University
Address6
fujiyoshi@isc.chubu.ac.jp

Abstract

Attention branch networks (ABNs) can achieve high accuracy by visualizing the attention area of the network during inference and utilizing it in the recognition process. However, if the attention area does not highlight the target object to be recognized, it may cause recognition failure. While there is a method for fine-tuning the ABN using attention maps modified by human knowledge, it takes up a lot of labor and time because the attention map needs to be modified manually. In this paper, we propose a method that automatically optimizes the attention map by introducing an attention mining branch to the ABN. Our evaluation experiments show that the proposed method improves the recognition accuracy and obtains an attention map that appropriately focuses on the target object to be recognized.

1 Introduction

Visual explanation provides the attention area during the inference of a convolutional neural network (CNN) [1] as an attention map to interpret the reason of the network output. A typical visual explanation method is the attention branch network (ABN) [2], which uses the attention map as an attention mechanism that calculates the element-wise product of the attention map and the feature map. This attention mechanism enables the important area for recognition to be captured and improves the accuracy. However, the attention map may focus on objects other than the object to be recognized. Such attention maps can cause false recognition and have a negative effect on the network training and accuracy.



(a) Fine-tuning method of ABN modified by human knowledge

(b) Proposed method

Figure 1. Fine-tuning methods of ABN using attention maps modified by (a) human knowledge and (b) the proposed method.

To overcome this problem, a fine-tuning method of ABN by human knowledge has been proposed [3]. In this method, the attention maps of misclassified samples are modified for an ideal attention map through human knowledge, as shown in Fig. 1(a). The pretrained ABN is then fine-tuned using the modified attention maps. Although this approach can refine attention maps and improve the recognition accuracy, manual modification of the attention maps takes a lot of human effort and time.

In this paper, we propose a fine-tuning method of ABN while considering the effective regions for recognition. Figure 1(b) shows the overview of our method.

We introduce an *attention mining branch* for the fine-tuning that utilizes the concept of the Guided Attention Inference Network [4], which is a segmentation method, to help the attention mining branch learn to gaze only at the object to be recognized and then automatically modify the attention map. Experimental results show that the proposed method can obtain effective attention maps for recognition while reducing the human cost by automatically modifying the attention maps.

2 Related Work

Attention maps enable us to understand the reason for a network decision. Several methods for obtaining the attention map have been proposed [5, 6, 2], which can be categorized into two approaches: bottom-up and top-down. The bottom-up approach computes the attention map by using local responses of convolution [12, 13]. The top-down approach computes attention maps derived from class information of the network output [5, 6, 2]. ABN [2], which is one of the major top-down visual explanation methods, generates an attention map by using global average pooling [7] and feature maps, and then uses the map for the attention mechanism to enhance the features of the target object. This attention mechanism improves the classification accuracy. Our method utilizes the branch structure and attention mechanism for optimizing attention maps.

For optimizing attention maps, a fine-tuning method based on human-in-the-loop has been proposed [3]. This method manually edits the attention maps of misclassified images that focus on the target object or characteristic region for classification and then fine-tunes the network parameters by using the edited attention map. This enables the network to correctly focus on the same region as a human would and improves the explainability and accuracy. However, this method requires the attention maps to be manually edited, which causes an increase in human labor and time. In contrast, our fine-tuning approach can optimize attention maps without manual editing.

The most similar work to our own is GAIN, a method proposed by Li *et al.* [4]. GAIN is a weakly supervised semantic segmentation method that first makes a mask image from an attention map obtained by Gradient-weighted Class Activation Mapping (Grad-CAM) [6] and then generates an image whose highlighted region is hidden by applying the mask for the input image. It then updates the network parameters using an additional loss value calculated from the correct class classification probability for the generated image. This additional loss makes the network focus only on the target object. GAIN computes Grad-CAM attention maps that require backpropagation and then inputs masked images to the network to update the parameters. Our fine-tuning method differs in that it can generate the attention map, infer the masked sample, and update the parameters during an inference.

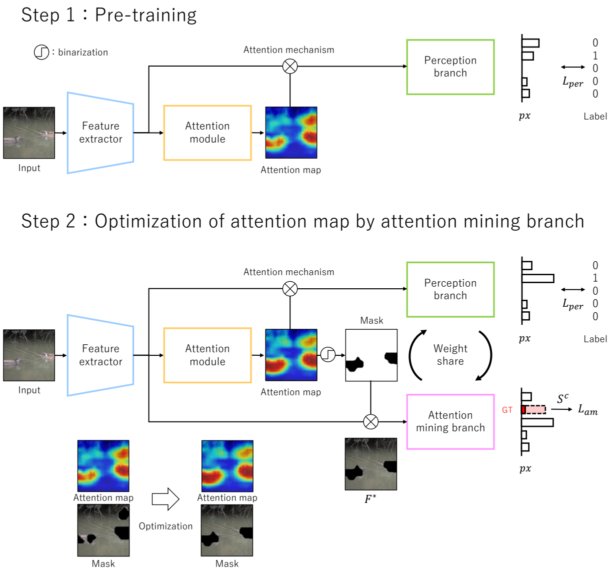


Figure 2. Process flow of proposed method. In step 1, we train the network. In step 2, we fine-tune the network by using the attention mining branch and masked feature map.

3 Proposed Method

In this section, we introduce our proposed attention mining branch and fine-tuning method while considering the regions that are effective for recognition.

The proposed method automatically optimizes the attention map by introducing the attention mining branch into the ABN. Figure 2 shows the structure of the proposed method. It first extracts a feature map from an input image by a feature extractor and then inputs the feature map into the attention module to generate an attention map. The feature map and attention map are used for the attention mechanism to enhance the features of the highlighted region and obtain classification results by the perception branch. Our method further utilizes the attention mining branch to optimize the attention map during the fine-tuning step.

3.1 Attention mining branch

The attention mining branch learns to acquire regions that are effective for recognition. Figure 2 shows the optimization flow of the attention map by the attention mining branch. The structure of the attention mining branch is the same as that of the perception branch. Also, the branch shares the weights with the perception branch and outputs class probabilities by using a masked feature map. If the class probability of the target class decreases, we can assume that the masked region hides the target objects. Therefore, by learning to minimize the class probability of the target class, the attention map is optimized to gaze only at the target object. The attention mining branch shares weights with the perception branch. This weight share allows the perception branch to reflect the weights of the attention mining branch, which has learned to gaze only at the object to be recognized.

3.1.1 Mask generation method

For generating a masked feature map, we use the attention maps obtained from the attention module. Let A be the attention map, and σ be the threshold of attention. The mask T is defined by

$$T(A) = \frac{1}{1 + exp(-100(A - \sigma))}. \tag{1}$$

By using the Sigmoid function, the process is equivalent to binarization while maintaining the gradient. Then, we multiply the feature map obtained from the feature extractor and the mask. Let F be the feature map from the feature extractor. The masked feature map F^* is defined by

$$F^* = F - (T(A) \odot F). \tag{2}$$

Consequently, we can generate a masked feature map that hides the highlighted area.

3.2 Learning algorithm

Figure 2 shows the process flow of the proposed method. The training procedure is implemented as follows.

Step 1 We first initialize the network's parameters, and train the network.

Step 2-1 We generate the mask from an attention map obtained by the attention module. Then, the output of the feature extractor is multiplied by the generated mask to obtain the masked feature map.

Step 2-2 We input the masked feature map generated in step 2-1 to the attention mining branch and obtain class probabilities as an output. Then, we compute a loss of the attention mining branch L_{am} from the output probability and the ground truth. L_{am} is the sum of the class probabilities of each sample output from the attention mining branch. This means that the smaller loss L_{am} successfully hides the object to be recognized. Let $c \in \{1, \ldots, C\}$ be class and $i \in \{1, \ldots, n\}$ be a sample in a mini-batch. We denote the classification probability of correct class c for the i-th masked feature map as S_i^c. The loss L_{am} is defined as follows:

$$L_{am} = \sum_{i=1}^{n} S_i^c. \tag{3}$$

Table 1. Top-1 and top-5 accuracy on each dataset [%]

Model	CUB-200-2010		Stanford Dogs	
	Top-1	Top-5	Top-1	Top-5
ABN [2]	31.68	57.01	71.81	**93.02**
Proposed	**33.53**	**58.68**	**71.99**	92.80
Human knowledge [3]	37.42	62.08	–	–

Step 2-3 We update the network parameters. The loss is calculated by three loss values: L_{am}, L_{att}, and L_{per}. L_{att} is a cross-entropy loss between the output of the attention module and the correct label. Likewise, L_{per} is a cross-entropy loss between the output of the perception branch and the correct label. The entire loss function L is defined as

$$L = L_{att} + L_{per} + \alpha L_{am}, \tag{4}$$

where α is a scaling parameter for L_{am}.

4 Experiments

To evaluate the effectiveness of the proposed method, we performed evaluation experiments on a fine-grained image recognition task.

4.1 Experimental settings

We used the Caltech-UCSD Birds 200-2010 (CUB-200-2010) dataset [8] and the Stanford Dogs dataset [9]. ResNet-50 [10] was utilized as the base network. The number of training updates was 300 epochs each for the ABN pre-training and the proposed method. The batch size was set to 16. The coefficient α of L_{am} was set to 0.0001. The mask threshold was set to 0.78 for the CUB-200-2010 dataset and to 0.40 for the Stanford Dogs dataset. As comparative methods, we adopted ABN [2] and the conventional fine-tuning method by human knowledge (human knowledge) [3].

4.2 Experimental results

Table 1 shows a comparison of the top-1 and top-5 accuracies for each dataset. In the results of CUB-200-2010, the recognition accuracy of the proposed method was 1.85 points better than that of ABN. In the Stanford Dogs dataset, the proposed method improved the recognition accuracy of Top-1 compared with ABN. Although the recognition accuracy of Top-1 was lower than that of the method introducing human knowledge, our method successfully improved accuracies without manually modified attention maps.

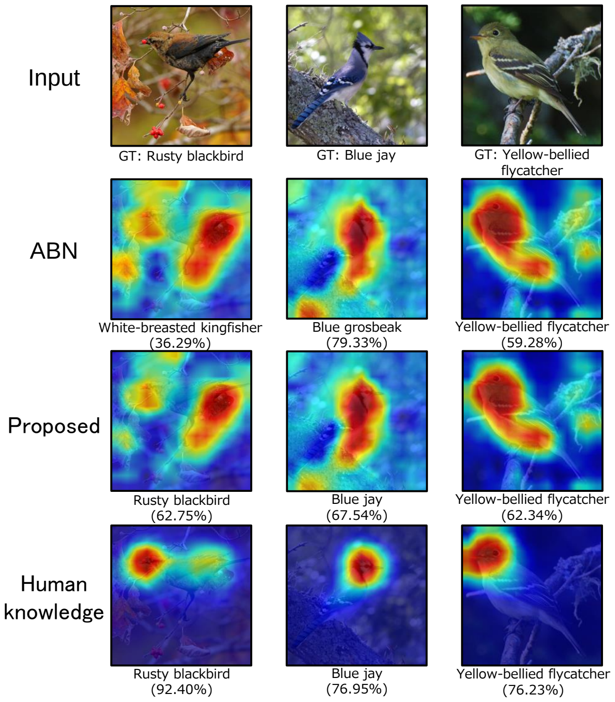


Figure 3. Examples of attention maps on CUB-200-2010.

4.3 Visualization of attention maps

We qualitatively evaluated the obtained attention maps. Figures 3 and 4 show examples of attention maps on CUB-200-2010 and Stanford Dogs, respectively.

As shown in Fig. 3, the attention maps of human knowledge-based fine-tuning could identify class objects by focusing on more localized regions. Compared with ABN, the proposed method improved the class probability by reducing the attention area outside the recognition target while gaining the effective area for recognition.

In the case of the Stanford Dogs dataset, as shown in Fig. 4, the proposed method improved the class probability by reducing the attention area outside the recognition target compared to ABN. Moreover, in the middle column results, the attention map of ABN highlighted the outside of the dog. In contrast, since the proposed method successfully refined the attention maps, the dog region was accurately highlighted.

4.4 Quantitative Evaluation of Attention Map

Next, we quantitatively evaluated the effectiveness of the attention acquired by the proposed method. As an evaluation metric, we used insertion [11]. In this evaluation, we masked images in the lower attention region and computed the accuracy for the masked images. We first evaluated the accuracy while changing the percentage of masked regions and then checked the average class probability of each sample for each percentage of insertions and evaluated them by the area under curve (AUC). The higher the AUC, the more effective the attention map is for recognition, as insertion is evaluated only in the more highlighted region in the attention map. In this experiment, we used only samples that ABN misclassified to confirm misclassification improvements.

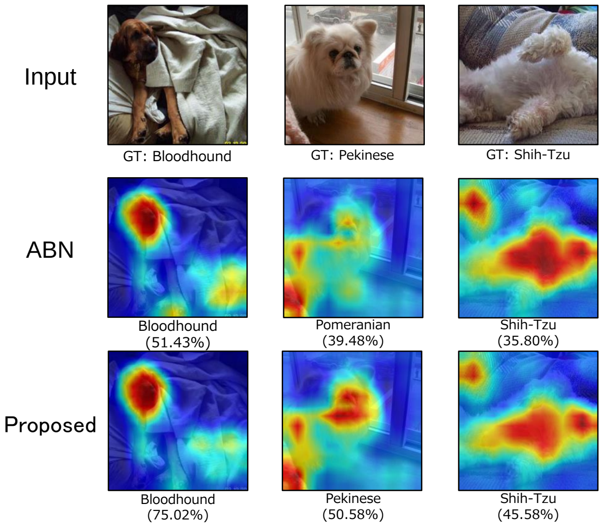
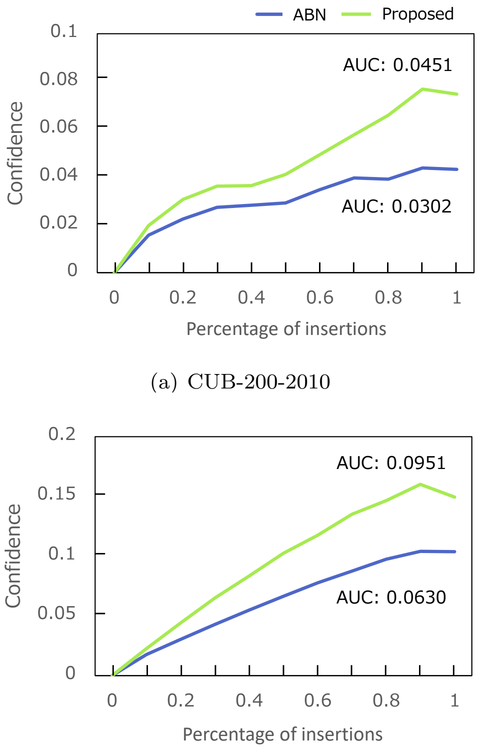


Figure 4. Examples of attention maps on Stanford Dogs.

Figure 5 shows the results of insertion for each dataset. In Fig. 5(a), we can see that the AUC of the proposed method was higher than that of ABN on the CUB-200-2010 dataset. Similarly, the AUC of the proposed method was higher than that of ABN on the Stanford Dogs dataset, as shown in Fig. 5(b). These results demonstrate that the proposed method can optimize the attention map.

5 Conclusion

In this paper, we proposed a method to optimize an attention map by introducing an attention mining branch into the ABN structure. The attention mining branch classifies samples using masked feature maps by generated attention maps during the training, which appropriately refines the attention maps to focus on the target object. Our experiments showed that the proposed method improved both the attention area and the recognition accuracy. Further, evaluation with insertion metrics demonstrated that the attention map obtained by the proposed method could capture the effective region for recognition. Our future work will include a more extensive evaluation of the proposed method on additional datasets.



(a) CUB-200-2010

(b) Stanford Dogs

Figure 5. Insertion metrics and area under curve on (a) CUB-200-2010 and (b) Stanford Dogs datasets.

References

[1] Alex, K., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, Neural Information Processing Systems, pp. 1097–1105 (2012).

[2] Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation, 2019 IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705–10714 (2019).

[3] Mitsuhara, M., Fukui, H., Sakashita,Y., Ogata,T., Hirakawa, T., Yamashita, T., and Fujiyoshi, H.: Embedding human knowledge in deep neural network via attention map, VISAPP, pp. 626–636 (2021).

[4] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu: Tell me where to look: Guided attention inference network, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9215–9223 (2018).

[5] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A.: Learning deep features for discriminative localization, 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016).

[6] Ramprasaath, R. S., Michael, C., Abhishek, D., Ramakrishna, V., Devi, P., and Dhruv, B.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Local-ization, International Conference on Computer Vision, pp. 618–626 (2017).

[7] Lin, M., Chen, Q., and Yan, S.: Network in network, in 2nd International Conference on Learning Representations, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (2014).

[8] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P.: Caltech-UCSD Birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology (2010).

[9] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li: Novel dataset for fine-grained image categorization: Stanford dogs, in Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Vol. 2, Citeseer (2011).

[10] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, Computer Vision and Pattern Recognition, pp. 770–778 (2016).

[11] Vitali Petsiuk, Abir Das, and Kate Saenko: RISE: Randomized Input Sampling for Explanation of Black-box Models, British Machine Vision Conference (BMVC), (2018).

[12] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg: SmoothGrad: Removing noise by adding noise, arXiv preprint, arXiv:1706.03825 (2017).

[13] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, and Karol Zieba: VisualBackProp: Efficient visualization of CNNs, arXiv preprint, arXiv:1611.05418 (2016).