Complement Objective Mining Branch for Optimizing Attention Map

Takaaki Iwayoshi[a], Hiroki Adachi[b], Tsubasa Hirakawa[c] and Takayoshi Yamashita[d] and Hironobu Fujiyoshi[e]
Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi, Japan

Keywords: Deep Learning, Attention Branch Network, Attention Mining, Attention Mechanism, Visual Explanation.

Abstract: Attention branch network (ABN) 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 requires a lot of human labor and time because the attention map needs to be modified manually. The method introducing the attention mining branch (AMB) to ABN improves the attention area without using human knowledge by learning while considering whether the attention area is effective for recognition. However, even with AMB, attention regions other than the target object, i.e., unnecessary attention regions, may remain. In this paper, we investigate the effects of unwanted attention areas and propose a method to further improve the attention areas of ABN and AMB. In the evaluation experiments, we show that the proposed method improves the recognition accuracy and obtains an attention map with more gazed objects. 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

In the field of image recognition, Deep Convolutional Neural Network (DCNN) (Alex and Hinton, 2012) has achieved high recognition performance, but it is difficult for humans to interpret the basis of decisions during recognition due to the complex network structure of DCNN. Visual explanation is a commonly used approach to overcome this difficulty.

Class activation mapping (CAM) (Zhou et al., 2016), gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017), and attention branch network (ABN) are major visual explanation methods. These methods can visualize the attention regions as the attention map during the inference of DCNN. ABN not only visualizes the attention regions during the inference but also can boost recognition performance by leveraging the attention regions during training. Specifically, it can capture an effective region for recognition by multiplying the attention regions to feature maps with the attention mechanism. However, attention maps of ABN focus on not only a recognition target but also other things or no longer focus on it. These attention maps, i.e., the attention maps including an inappropriate region, make training of ABN difficult, so they may degrade the recognition performance.

For optimizing attention maps, a fine-tuning method based on human-in-the-loop has been proposed (Mitsuhara et al., 2021). 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.

To overcome the shortcomings of Mitsuhara *et al.*'s method, attention mining branch (AMB) (Iwayoshi et al., 2021) reduces attention regions other than objects to be recognized without human knowledge. Specifically, we have succeeded in suppressing the generation of unwanted attention

[a] https://orcid.org/0000-0003-4421-3270
[b] https://orcid.org/0000-0001-5920-2633
[c] https://orcid.org/0000-0003-3851-5221
[d] https://orcid.org/0000-0003-2631-9856
[e] https://orcid.org/0000-0001-7391-4725

Iwayoshi, T., Adachi, H., Hirakawa, T., Yamashita, T. and Fujiyoshi, H.
Complement Objective Mining Branch for Optimizing Attention Map.
DOI: 10.5220/0011699000003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 105-113
ISBN: 978-989-758-634-7; ISSN: 2184-4321

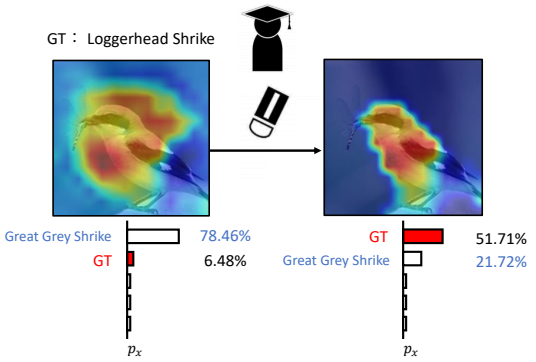

Figure 1: Investigation of the effect of attention areas outside the recognition target on inference results. We remove the attention areas outside the recognition target and investigate changes in class probability.

regions through learning that considers whether the attention regions are effective for recognition in ABN and AMB. However, attention regions other than the object to be recognized remain, and the recognition accuracy has not improved as much as human knowledge. Therefore, in this study, we first clarify the effect of attention regions other than the object to be recognized on the inference results. As shown in Figure 1, we observed that the incorrect class probability tends to decrease, and the correct class probability tends to increase when unnecessary regions are manually removed. Based on the results of this experiment, we believe it is important to intentionally reduce the incorrect answer class probability and incorporate such a learning method into ABN and AMB.

In summary, our work makes the following contributions:

- In the randomly selected examples, we observed a tendency for the incorrect class probability to decrease by arbitrarily reducing the unnecessary attention area.
- In this paper, we propose a method to introduce Complement Objective Training (COT)(Chen et al., 2019) to flatten the incorrect class probability in ABN and AMB and show that our method can reduce the unwanted attention area by decreasing the incorrect class probability and improve the recognition accuracy accordingly. Our method is able to reduce the number of unnecessary attention regions by decreasing the incorrect class probability and to improve the recognition accuracy accordingly.
- We show that the proposed method acquires attention regions better than human knowledge by using Insertion, which performs inference only on regions of high importance for each percentage of the acquired attention regions.

2 RELATED WORKS

In this section, we introduce Methods for Visual Explanation and a learning method that considers the incorrect class probability, which is considered to be one of the causes of unwanted attention areas.

2.1 Visual Explanation

Attention maps enable us to understand the reason for a network decision. Several methods for obtaining the attention map have been proposed (Fukui et al., 2019; Iwayoshi et al., 2021; Zhou et al., 2016; Chattopadhay et al., 2018; Chen et al., 2019; Fong et al., 2019; Fong and Vedaldi, 2017; Selvaraju et al., 2017; Mitsuhara et al., 2021; Ribeiro et al., 2016; Vitali Petsiuk and Saenko, 2018; Springenberg et al., 2014; Zhang et al., 2021), 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 (Smilkov et al., 2017; Bojarski et al., 2016). The top-down approach computes attention maps derived from class information of the network output. ABN (Fukui et al., 2019), which is one of the major top-down visual explanation methods, generates an attention map by using global average pooling (Lin et al., 2013) 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 (Mitsuhara et al., 2021). 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.

Moreover, AMB can reduce the attention areas other than the object to be recognized without using human knowledge by introducing AMB into ABN and learning it while considering the effective areas for recognition. However, the recognition accuracy is not improved compared with the case where human knowledge is introduced, because the attention regions that are not necessary for recognition remain.

Therefore, in this study, we investigate how attention regions unnecessary for recognition affect class probability. Then, we propose a method to improve ABN and AMB.

2.2 Complement Objective Training

COT is a learning method that considers the entropy of incorrect classes. The weights are updated so that the correct answer class probability is close to 1. Next, the weights are updated to flatten the probability of the incorrect answer class. This is done for each iteration to improve the probability distribution. From the investigation described below, we know that the attention regions other than the object to be recognized affect the probability of the incorrect answer class. Therefore, we aim to further improve the attention regions by introducing this method to ABN and AMB.

3 PROPOSED METHOD

This paper investigates the effect of incorrect answer class probability on the attention area and proposes a method to improve ABN and AMB accordingly. Specifically, this study proposes ABN + COT, which introduces COT to ABN, and ABN + COMB, which introduces Complement Objective Mining Branch (COMB), a combination of AMB and COT, to ABN.

3.1 Influence of Unnecessary Attention Areas

AMB reduces the attention regions other than the object to be recognized without human knowledge. However, attention regions other than the object to be recognized remain, and the recognition accuracy has not improved as much as human knowledge. Therefore, in this study, we first clarify the effect of attention regions other than the object to be recognized on the inference results. Specifically, as shown in Figure 2, we investigate the relationship between unnecessary attention areas and class probability by deleting unnecessary attention areas from the attention map when misrecognition occurs in ABN and performing inference again.

The ABN model trained on the CUB-200-2010 dataset was used to investigate the relationship between unwanted attention areas and class probability. First, we investigate changes in the probability distribution using three samples in the CUB-200-2010 dataset that are misrecognized by ABN. Figure 3 shows the probability distributions of the three samples. As shown in Figure 3, the incorrect class probability decreases, and the correct class probability increases in all cases. Second, we investigate the variation of the highest incorrect class probability using 10 randomly selected samples from the CUB-200-2010 dataset. Table 1 shows the average of the highest incorrect answer class probabilities for the 10 samples. As shown in Table 1, we observed that the incorrect answer class probabilities decreased when the unnecessary regions were manually removed. Based on these findings, we believe it is important to intentionally reduce the incorrect answer class probability and incorporate such a learning method into ABN and AMB.

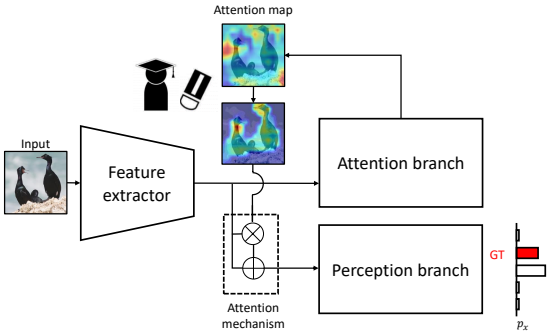

Figure 2: Flow of Correcting Attention. We remove attention areas outside the recognition target of Attention acquired by ABN and weight the feature map by the Attention mechanism. Then, the class probability of the perception branch, which is the final evaluation, is investigated.

Table 1: Average of the highest incorrect class probability in 10 randomly selected samples [%]. Bold letters indicate the lowest the incorrect class probability.

	Incorrect class probability
ABN	41.61
Reduction	**26.61**

3.2 ABN + COT

This paper aims to reduce unnecessary attention areas by introducing COT to the ABN and flattening the incorrect answer class probability.

3.2.1 Network Structure

The structure of ABN + COT. 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.

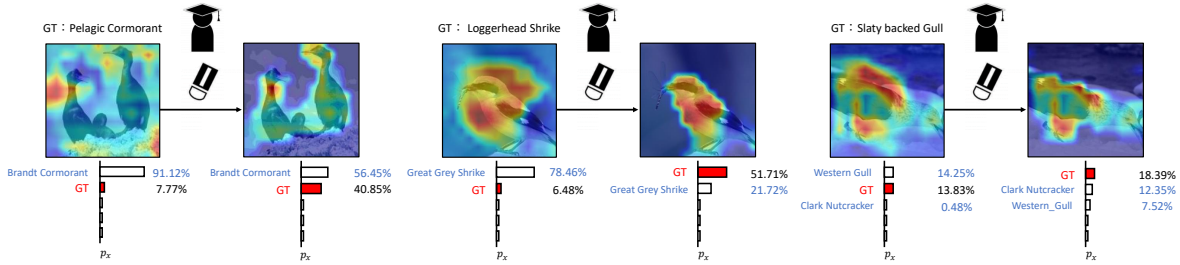

Figure 3: Probability distribution of three samples misrecognized by ABN in the CUB-200-2010 dataset. Blue letters indicate incorrect class names.

3.2.2 Learning Algorithm

Algorithm 1 alternates between minimizing the cross-entropy loss that brings the correct answer class probability close to 1 and flattening the incorrect answer class probability. The specific method is described below.

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

Step 2. We update the network parameters. The loss is calculated by two loss values: 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

$$L = L_{att} + L_{per}. \tag{1}$$

Step 3. Flatten the incorrect class probability by updating the weights to minimize the complement entropy. Let $\hat{y}$ be the predicted probability distribution for input, $\mathcal{H}(\cdot)$ be the entropy function, and g be the ground truth. Complement entropy $C(\hat{y}_{\bar{C}})$ is defined by

$$\begin{aligned} C(\hat{y}_{\bar{C}}) &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(\hat{y}_{i\bar{C}}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,j\neq g}^{K}\frac{\hat{y}_{ij}}{1-\hat{y}_{ig}}\log\left(\frac{\hat{y}_{ij}}{1-\hat{y}_{ig}}\right). \end{aligned} \tag{2}$$

Step 4. Repeat Steps 2 and 3 for each iteration.

3.3 ABN + COMB

This paper aims to reduce unnecessary attention areas by introducing COT to the AMB and flattening the incorrect answer class probability.

3.3.1 Network Structure

ABN + COMB automatically optimizes the attention map by introducing an AMB into the ABN. Figure 5 shows the structure of ABN + COMB. 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 AMB to optimize the attention map during the fine-tuning step.

Require: Total number of samples N, Class label y_i, iteration n, probability $\hat{y}$, Ground truth g, Predicted probabilities of the correct classes $\hat{y}_{ig}$, Complement entropy $\bar{C}$, Entropy function $\mathcal{H}$, Predicted probabilities of the complement (incorrect) classes $\hat{y}_{i\bar{C}}$

Initialize: Update weights to flatten incorrect answer class probability Load weights from network trained in ABN.

1: **for** $t \leftarrow 1$ to n **do**
2: At Attention branch and Perception branch Update weights so that the correct answer class probability approaches 1.:$-\frac{1}{N}\sum_{i=1}^{N}\log(\hat{y}_{ig})$
3: At Attention branch and Perception branch Update weights to flatten incorrect answer class probability:$\frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(\hat{y}_{i\bar{C}})$
4: **end for**

Figure 4: ABN + COT.

3.3.2 Attention Mining Branch

The AMB learns to acquire regions that are effective for recognition. Figure 5 shows the optimization flow of the attention map by the AMB. The structure of the AMB 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,

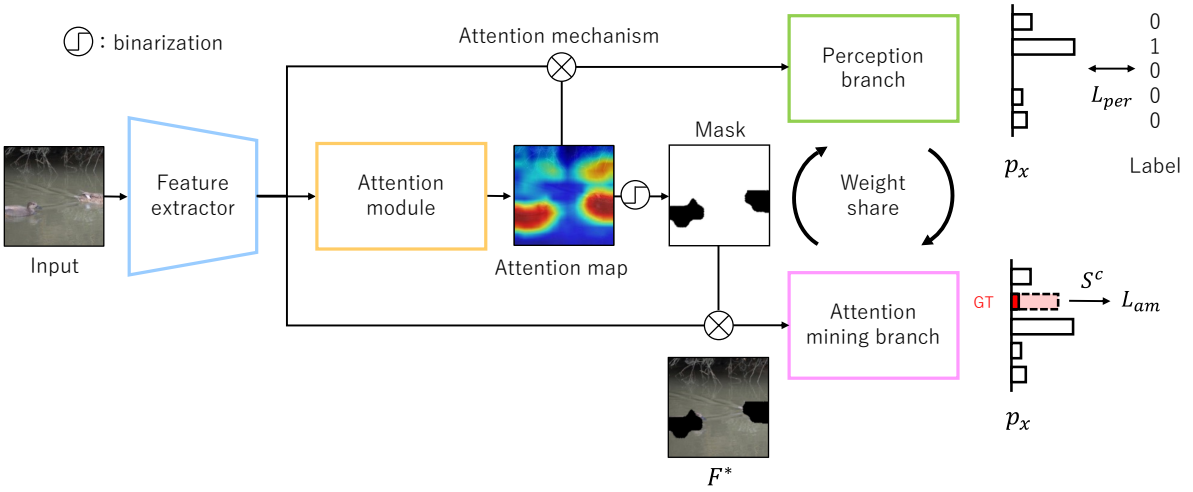

Figure 5: The structure of ABN + AMB. In our method, Attention is binarized to obtain a mask, and the mask is applied to the feature map to obtain a feature map F^* that hides the highlighted area. This is input to Attention mining branch, which shares weights with Perception branch, to obtain class probabilities. If the class probability of the target class S^c 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 map is optimized to gaze only at the target object. The AMB shares weights with the perception branch. This weight share enables the perception branch to reflect the weights of the AMB, which has learned to gaze only at the object to be recognized.

3.3.3 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

$$T(A) = \frac{1}{1+\exp(-100(A-\sigma))}. \quad (3)$$

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

$$F^* = F - (T(A) \odot F). \quad (4)$$

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

3.3.4 Learning Algorithm

Algorithm 2, minimization of the correct answer class probability and flattening of the incorrect answer class probability are alternately repeated in the AMB. The specific method is described below.

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

Require: Total number of samples N, Class label y_i, iteration n, probability $\hat{y}$, Ground truth g, Predicted probabilities of the correct classes $\hat{y}_{ig}$, Complement entropy $\bar{C}$, Entropy function $\mathcal{H}$, Predicted probabilities of the complement (incorrect) classes $\hat{y}_{i\bar{C}}$

Initialize: Update weights to flatten incorrect answer class probability Load weights from network trained in ABN.

1: **for** $t \leftarrow 1$ to n **do**
2: At Attention branch and Perception branch
Update weights so that the correct answer class probability approaches 1.: $-\frac{1}{N}\sum_{N}^{i=1}\log(\hat{y}_{ig})$
At Attention mining branch
Update weights so that the correct answer class probability approaches 0.
3: At Attention mining branch
Update weights to flatten incorrect answer class probability: $\frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(\hat{y}_{i\bar{C}})$
4: **end for**

Figure 6: ABN + COMB.

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 AMB and obtain class probabilities as an output. Then, we compute a loss of the AMB 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 AMB. 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:

$$L_{am} = \sum_{i=1}^{n} S_i^c. \tag{5}$$

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

$$L = L_{att} + L_{per} + \alpha L_{am}, \tag{6}$$

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

Step 3. Flatten the incorrect class probability by updating the weights to minimize the complement entropy. Let $\hat{y}$ be the predicted probability distribution for input, $\mathcal{H}(\cdot)$ be the entropy function, and g be the ground truth. The complement entropy $C(\hat{y}_{\bar{C}})$ is defined as

$$\begin{aligned} C(\hat{y}_{\bar{C}}) &= \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(\hat{y}_{i\bar{C}}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,j\neq g}^{K} \frac{\hat{y}_{ij}}{1-\hat{y}_{ig}} \log\left(\frac{\hat{y}_{ij}}{1-\hat{y}_{ig}}\right). \end{aligned} \tag{7}$$

By setting $\hat{y}_{ig} = 0$, the correct answer class probability is calculated to be 0 and the remaining incorrect answer class probabilities are flat.

Step 4. Repeat Steps 2 and 3 for each iteration.

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 (Welinder et al., 2010) and the Stanford Dogs dataset (Khosla et al., 2011). ResNet-50 (He et al., 2016) 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 (Fukui et al., 2019), ABN + AMB(Iwayoshi et al., 2021), and the conventional fine-tuning method by human knowledge (Mitsuhara et al., 2021). In this experiments, we call the method proposed by Mitsuhara et al. to "human knowledge".

Table 2: Top-1 and top-5 accuracy on CUB-200-2010 dataset [%]. Bold letters indicate the highest accuracy.

model	Top-1	Top-5
ABN	31.68	57.01
ABN + AMB	33.53	58.68
Human knowledge	37.42	62.08
ABN + COT	**43.98**	**66.83**
ABN + COMB	39.76	66.57

Table 3: Top-1 and top-5 accuracy on Stanford Dogs dataset [%]. Bold letters indicate the highest accuracy.

model	Top-1	Top-5
ABN	71.81	93.02
ABN + AMB	71.99	92.80
ABN + COT	72.33	91.12
ABN + COMB	**73.59**	**93.89**

4.2 Experimental Results

Table 2 compares the top-1 and top-5 accuracies for CUB-200-2010. In the results of CUB-200-2010, the recognition accuracy of the proposed method was better than that of ABN, ABN + AMB, and Human knowledge.

Table 3 compares the top-1 and top-5 accuracies for the Stanford Dogs dataset. In the Stanford Dogs dataset, the proposed method improved the recognition accuracy of Top-1 compared with ABN and ABN + AMB. Especially, the recognition accuracy of ABN + COT was 12.30 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. Especially, the recognition accuracy of ABN + COMB was 1.78 points better than that of ABN. These results show that the proposed method contributes to the recognition performance.

4.3 Visualization of Attention Maps

We qualitatively evaluated the obtained attention maps. Figures 7 and 8 show examples of attention

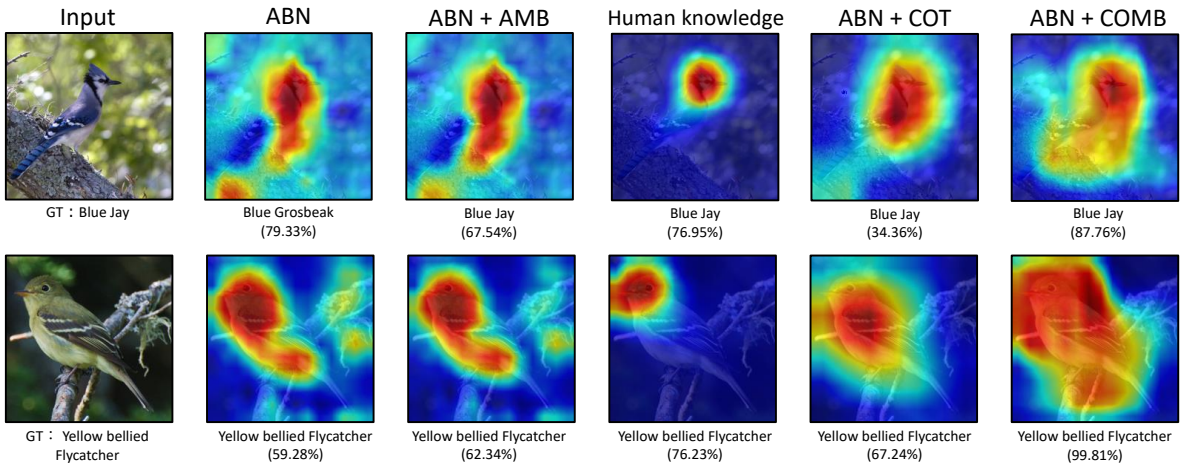

Figure 7: Examples of attention maps on CUB-200-2010. Class names and confidences of the highest class probabilities are shown below the Attention map.

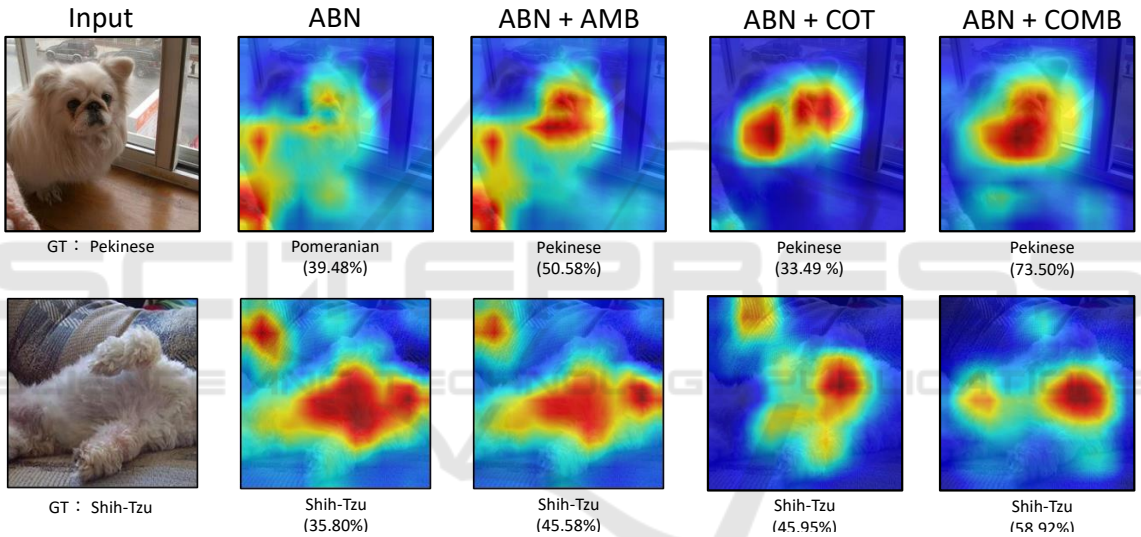

Figure 8: Examples of attention maps on Stanford Dogs. Class names and confidences of the highest class probabilities are shown below the Attention map.

maps on CUB-200-2010 and Stanford Dogs, respectively. As shown in Fig. 7, the attention maps of human knowledge-based fine-tuning could identify class objects by focusing on more localized regions. Compared with ABN and ABN + AMB, the proposed method reducing the attention area outside the recognition target while gaining the effective area for recognition. Moreover, compared with ABN, ABN + AMB, and human knowledge-based fine-tuning, ABN + COMB improved the class probability.

In the case of the Stanford Dogs dataset, as shown in Fig. 8, compared with ABN and ABN + AMB, the proposed method reduced the attention area outside the recognition target while gaining the effective area for recognition.

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 (Vitali Petsiuk and Saenko, 2018). 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 at-

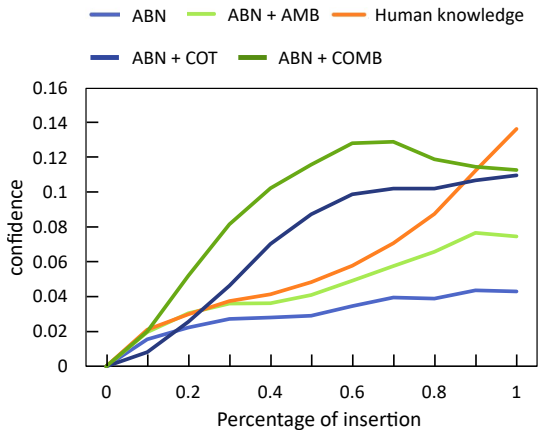

Figure 9: Insertion metrics on CUB-200-2010 dataset.

Table 4: Area Under Curve for Insertion metrics on CUB-200-2010 dataset. Bold letters indicate the highest score.

model	AUC
ABN	0.0302
ABN + AMB	0.0451
Human knowledge	0.0576
ABN + COT	0.0704
ABN + COMB	**0.0921**

tention map. In this experiment, we used only samples that ABN misclassified to evaluate misclassification improvements. Figure 9 and Table 4 show the results of insertion for each dataset. Table 4 shows that the AUC of the proposed method was higher than that of ABN and ABN + AMB, human knowledge-based fine-tuning. These results demonstrate that the proposed method can optimize the attention map.

5 CONCLUSION

In this paper, we investigated the relationship between the attention area outside the recognition target and the incorrect answer class probability and proposed a method to optimize an attention map by introducing Complement Objective Training (COT) into the attention branch network (ABN) and attention mining branch (AMB). 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 apply this technology to segmentation and multitasking.

ACKNOWLEDGEMENTS

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

Alex, K., S. I. and Hinton, G. E. (2012). Paper templates. In *ImageNet Classification with Deep Convolutional Neural Networks*. SCITEPRESS.

Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U., and Zieba, K. (2016). Visualbackprop: efficient visualization of cnns. *arXiv preprint arXiv:1611.05418*.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). 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.

Chen, H.-Y., Wang, P.-H., Liu, C.-H., Chang, S.-C., Pan, J.-Y., Chen, Y.-T., Wei, W., and Juan, D.-C. (2019). Complement objective training. *arXiv preprint arXiv:1903.01182*.

Fong, R., Patrick, M., and Vedaldi, A. (2019). Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437.

Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10705–10714.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Iwayoshi, T., Mitsuhara, M., Takada, M., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2021). Attention mining branch for optimizing attention map. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. (2011). Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Mitsuhara, M., Fukui, H., Sakashita, Y., Ogata, T., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2021). Embedding human knowledge in deep neural network

via attention map. In *The International Conference on Computer Vision Theory and Applications*, pages 626–636.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Vitali Petsiuk, A. D. and Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. (2010). Caltech-ucsd birds 200.

Zhang, Q., Rao, L., and Yang, Y. (2021). Group-cam: group score-weighted visual explanations for deep convolutional networks. *arXiv preprint arXiv:2103.13859*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.