

Worapon Yingyong¹, Shaoqing Wu¹, Hiroyuki Yamauchi²¹ Graduate School of Engineering, Department of Intelligent System Engineering,² Department of Information Engineering, Faculty of Information Engineering,
Fukuoka Institute of Technology, Fukuoka, Japan

1. Introduction

Although convolutional neural networks (CNNs) achieve high accuracy in image classification tasks, the rationale behind their predictions is often unclear. Attention Branch Network (ABN) [1] attempts to visualize model focus using attention maps, yet these maps do not always align with human expectations. In this study, we propose a fine-tuning approach that adjusts the attention map locations using human-edited heatmaps. By refining only the attention branch weights, our method aims to improve interpretability while preserving classification accuracy.

2. Proposed Method

Our approach incorporates human knowledge by guiding correct attention maps. The attention branch is then fine-tuned to align the model's focus with human-guided attention maps. Furthermore, we analyze the learned channel weights to identify which feature maps contribute most to human-guided attention.

2.1 Embedding Human Knowledge via Edited Attention Map

The flow of the proposed method is illustrated in Fig. 1. We begin by training an ABN model using a labeled image dataset. After training, we visualize the attention maps generated by the attention branch for each input image. Among these, we identify a subset of samples whose attention maps do not align with human expectations, even though they are classified correctly. These maps are then manually edited using a custom graphical interface, similar to Attention Editor ABN [2]. The resulting corrected attention maps serve as ground truth for the fine-tuning phase.

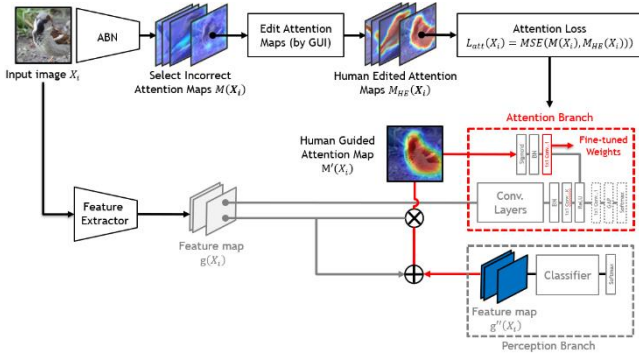


Figure 1: Overview of the proposed method: training ABN, editing attention maps, and fine-tuning with human guidance.

2.2 Fine-Tuning of Attention Branch

During the fine-tuning phase, we aim to adjust only the internal weights in the attention branch so that the model's generated attention maps closely resemble the human-guided ones, while maintaining the original classification performance. We define an attention loss $L_{att}(X_i)$ that measures the discrepancy between the predicted attention map $M(X_i)$ and the human-edited target attention map $M_{HE}(X_i)$ for a given input image X_i . The attention loss is calculated as the mean squared error (MSE):

$$L_{att}(X_i) = MSE(M(X_i), M_{HE}(X_i)) \quad (1)$$

This loss is backpropagated to update only the weights of the attention branch. All other parameters, including the feature extractor and the perception branch, are kept fixed. This ensures that the fine-tuning process focuses solely on improving the model's interpretability without significantly affecting the original accuracy.

2.3 Identify Important Channels

After fine-tuning the attention branch, we analyze the updated weights to determine which channels significantly contribute to generating the human-guided attention map. Each channel in the feature map corresponds to a learned weight in the final 1×1 convolution layer. The magnitude of each weight reflects its importance in forming the final attention map. We compare the channel-wise weights before and after fine-tuning by visualizing bar charts. Channels with high weight values after fine-tuning are interpreted as being more aligned with human-indicated focus areas. This analysis provides insight into which features the model prioritizes when guided by human knowledge.

3. Result

To evaluate the effectiveness of our human-guided attention refinement, we compared attention maps generated by the initial model, the human-guided attention maps, and the attention maps after fine-tuning. As shown in Fig. 2, the initial attention maps often focused on irrelevant or broad regions, despite the model achieving high classification accuracy. After incorporating human guidance and fine-tuning the attention branch, the refined attention maps aligned more closely with the human annotations. This demonstrates the model's improved interpretability while maintaining its predictive performance.

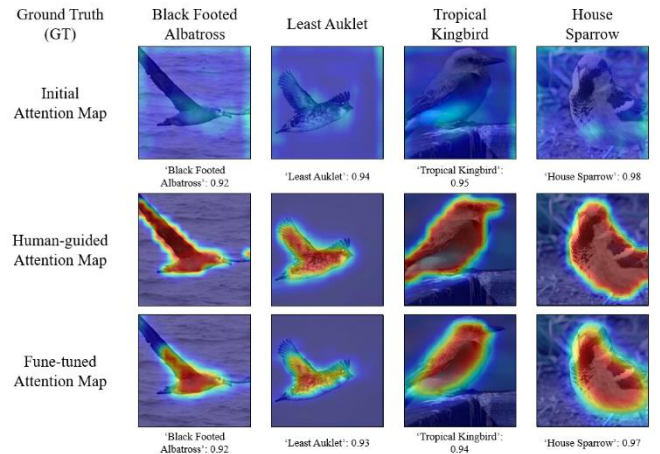


Figure 2: Comparison of initial, human-guided, and fine-tuned attention maps.

4. Conclusion

In this research, we demonstrate a human-guided fine-tuning framework to refine attention map localization in ABN. By updating only final weights in the attention branch, our method shifts attention toward human-designated regions while preserving classification accuracy.

References

- [1] T. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention Branch Network: Learning of Attention Mechanism for Visual Explanation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
- [2] M. Mitsuhashi, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Tuning Class-specific Attention with Additive Attention Map Supervision," in Proceedings of the Asian Conference on Computer Vision (ACCV), 2022.