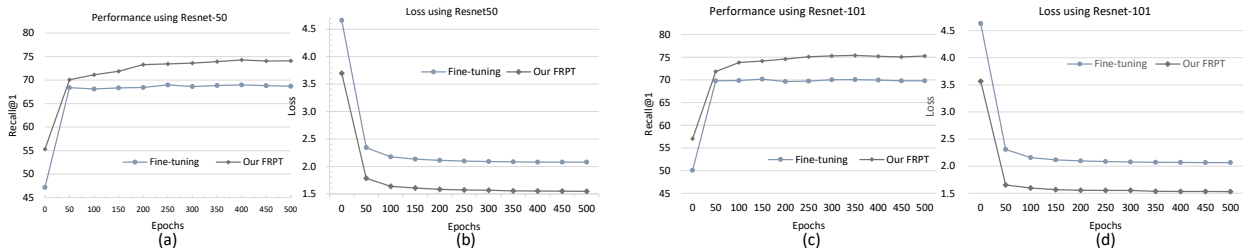

Figure 2: Curve visualization based on CUB-200-2011. (a) (c) denote the recall@1 curves about retrieval performance using Resnet-50 and Resnet-101, respectively. (b)(d) are the loss curves using Resnet-50 and Resnet-101, respectively.

Figure 3: Visualization of the object content perturbation. The first and second rows denote the original and modified images, respectively.

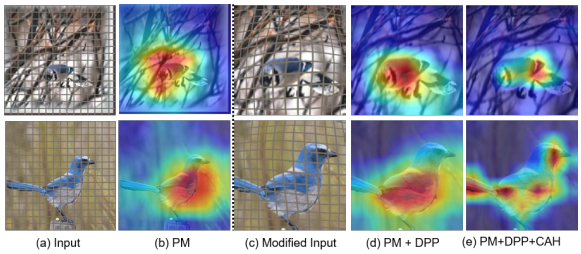

Figure 4: Illustration of class activation maps (CAM). (a)(c) are the original and modified images, respectively. (b)(d)(e) denote CAMs generated by diverse networks.

from the original images. Tab. 5 shows a significant performance improvement when we utilize more accurate localization manner to remove the background information and preserve more object regions as much as possible. However, our FRPT zooms and even exaggerates the idiosyncratic elements contributing to decision boundary rather than to simply amplify objects and remove background, thus making the FGOR task aided by the discriminative perturbation prompt close to the solved task during the original pre-training and forming a steady improvement.

What makes a network retrieve objects visually? With this question in our mind, we exhibit the visualization results of original and modified images in Fig. 3. These visualization images can interpret why and how our approach can correctly identify diverse subcategories. As shown in the second row, our sample prompting scheme can enhance the visual evidence of object parts via the dense sampling operation while suppressing the background and even non-discriminative parts, thus instructing the pre-trained model to pay more attention to discriminative details and improving the retrieval performance accordingly. It should be clarified that we manually put grid lines on the images to better display the pixel shift in the images after our prompt processing. In Fig. 4, in addition to showing the original and modified images, we present the discriminative activation maps of three representation models, i.e, pre-trained model (Fig. 4(b)) , our FRPT without CAH (Fig. 4(d)), and our FRPT (Fig. 4(e)). It is clear that using DPP module can make the network focus on the object rather than background information, thus improving the discriminative ability of feature representation. Compared to Fig. 4(d), the activation maps (e) can pay more attention to the category-specific details via introducing CAH module. Based on these visualizations, our model generates clearer object boundaries and emphasises the discriminative details, thus providing higher retrieve performance.

Conclusion

In this paper, we propose Fine-grained Retrieval Prompt Tuning (FRPT), which aims to solve the issue of convergence to sub-optimal solutions caused by fine-tuning the entire FGOR model. FRPT design the discriminative perturbation prompt (DPP) and category-specific awareness head (CAH) to steer frozen pre-trained vision model to perform fine-grained retrieval task. Technically, DPP zooms and exaggerates some pixels contributing to category prediction, which assists the frozen pre-trained model prompted with this content perturbation to focus on discriminative details. CAH optimizes the semantic features extracted by pre-trained model via removing the species discrepancies using category-guided instance normalization, which makes the optimized features sensitive to fine-grained objects within the same meta-category. Extensive experiments demonstrate that our FRPT with fewer learnable parameters achieves the state-of-the-art performance on three widely-used fine-grained datasets.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant NO. 61976038 and NO.61932020. Neque autem nulla incidunt ratione unde recusandae non, sit tenetur dolorum doloribus fuga consectetur dignissimos voluptates illo, expedita minus iste laudan-

tium voluptatem magni aliquid qui quidem, quod tempore quisquam, iste unde aperiam at optio debitis sint