Learning Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition

Link to Paper:
http://openaccess.thecvf.com/content_iccv_2017/html/Zheng_Learning_Multi-Attention_Convolutional_ICCV_2017_paper.html

1.0 Summary of the Paper

1.1 Main Technical Contributions

This paper seeks to introduce a multi-attention convolutional neural network (MA-CNN) that allows the reinforcement of the learning of discriminative parts location and the learning of fine-grained features, which will be described further in section 1.2 below. The MA-CNN is divided into three different sub-network parts – base network, channel grouping layer and the part-classification layer. At the beginning, when an image is fed through the base network, feature channels containing regionally-based features will be generated. Since the peak response in each feature channel may be similar and overlap with one another, these similar feature channels will be clustered into the same group followed by a sigmoid function to produce probabilities, which is equivalent to the formation of multiple part attention maps. Subsequently, by pooling each part attention map, it will result in the generation of its respective part representations, which will be connected to a fully-connected layer with a softmax function to perform classification prediction.

1.2 Main Empirical Contributions

The current approaches for fine-grained recognition are typically based on two main methods, which are (i) discriminative part localization, and (ii) part-based fine-grained feature learning. However, current approaches tend to separate both methods without considering the mutual correlation between them. In contrast, this paper proposed that both methods of part generation and feature learning should reinforce each other.

This mutual reinforcement is performed based on the two loss functions introduced in the paper – Channel Grouping Loss and Part Classification Loss. The channel grouping loss is composed of two parts, DIS and DIV, which is the distance and diversity function respectively. The distance function aims to aggregate the coordinates in the same part, while the diversity function aims to encourage greater distance between different parts. By optimizing both losses alternately in an iterative manner, mutual reinforcement of the channel grouping layers and the classification network can thus be achieved.

2.0 Three Main Strengths of this Paper

2.1 Strength One

Since MA-CNN learns part proposals, which are multiple attention areas in an image that has strong discrimination ability, it does not rely on human-annotated bounding box / part annotations which is expensive and subjective. Therefore, this allows a great increase in applicability and scalability via MA-CNN.

2.2 Strength Two

Since MA-CNN is able to localize parts based on channel clusters of high intra-class similarity and inter-class separability, it further takes advantage of the part localization to predict fine-grained categories based on each part representation, which is optimized by learning a classification loss over the part representation and a channel grouping loss over the part attention map. By taking advantage of the inherent correlation, it increases its capability to localize multiple parts with discriminative features, while also be able to differentiate between fine-grained labels that are similar.

2.3 Strength Three

With multiple part proposals, it adds robustness to fine-grained recognition due to its greater generalization ability. This benefits the model greatly in cases where the objects have high pose variation, or when there is occlusion.

3.0 Three Main Weaknesses of this Paper

3.1 Weakness One

According to the paper, MA-CNN with four parts outperforms MA-CNN with two part in each of the CUB-Birds, Stanford-Cars and FGVC-Aircraft dataset. However, the performance level will saturate after reaching a certain point. For example, upon increasing four parts to six parts, it became difficult for the model to learn more discriminative parts from the object. Additionally, according to Zheng *et al* [1], the number of part attentions are pre-defined, which decreases the effectiveness and flexibility of the model.

3.2 Weakness Two

According to the implementation details in the paper in section 4.2, MA-CNN requires a higher resolution input with size of 448 x 448, and the 'larger resolution inputs in MA-CNN can benefit discriminative part learning'. However, according to Zheng *et al*, the requirement of input with higher resolution translate to the requirement of a higher computational cost

3.3 Weakness Three

According to section 3.3, 'Joint Part-based Feature Representation' of the paper, each region are taken as input by each part-CNNs to classify each part. However,

the training of CNNs for each part representation is inefficient. According to Zheng *et al* [1], this would result in bottlenecks that would impede 'the study on attention-based fine-grained recognition'.

4.0 Proposed Idea

4.1 To Overcome Weakness One

According to Hu, T., & Qi, H [2], the number of object parts generated by the MA-CNN is limited and proposed that the classification accuracy can be increased if more object parts can be introduced. In order to do so, Hu, T., & Qi, H proposed a Weakly-Supervised Data Augmentation Network (WS-DAN), which performs an attention guided data augmentation. The network extracts the local features from different object parts by erasing one region of objects' part out of the image, which in turn encourages the network to extract discriminative features from other object parts. Please see Figure 1 in Appendix I for an illustration from the authors' paper, 'See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification'.

Additionally, according to Zheng *et al* [1], the proposed trilinear attention sampling network (TASN) with a trilinear attention module can be introduced to learn rich feature representations from hundreds of part proposals. The trilinear attention module allows the network to generate multiple attention maps by modelling the inter-channel relationships of each feature map. Please see Figure 2 in Appendix I for an illustration from the authors' paper on the trilinear attention module.

Therefore, both methods indicated above allows the network to have greater flexibility in the learning of more discriminative parts.

4.2 To Overcome Weakness Two

According to Zheng *et al* [1], an attention-based sampler can be introduced to allow the network to attend to parts with a higher resolution so that discriminative parts can be better represented with a higher resolution. The attention-based sampler works by obtaining both structure-preserved and detail-preserved images by conducting non-uniform sampling over different attention maps. The structure-preserved image is obtained by taking the average of all attention maps. This is done as to take all discriminative parts into consideration. On the other hand, the detail-preserved image is obtained by randomly selecting one attention map, which preserves the fine-grained details of the attended area with high resolution. Since each attention map has equal probability to be selected during the training process, different fine-grained details can be refined sequentially. Additionally, based on the authors' observations, regions with large attention values are allocated with more sampling points. Therefore, instead of having an image input of a higher resolution, we can allow the network to select image parts that have higher resolutions.

4.3 To Overcome Weakness Three

According to Zheng *et al* [1], a feature distiller can be introduced to allow the distillation of part features into a global feature under a single stream, instead of ensembling multiple part CNNs. The feature distiller works by taking a detail-preserved image and a structure-preserved image (as described in section 4.2 above) as input, and transfers the details learned from a part-net to a master-net in a teacher-student manner in multiple iterations during the training process. The master-net learns the features of the structure-preserved image, while the part-net learns the fine-grained features of the detail-preserved image and distils the fine-grained features into the master-net during each iteration of the training process. This process is optimized by a classification loss in the part-net and a soft target cross entropy loss in the master-net.

4.3 To Improve the Quality of this Work

In this paper, the evaluation of the model performance is based on three datasets – Caltech-UCSD Birds, FGVC-Aircraft and Stanford Cars. However, despite being widely used, each of this dataset are relatively small. On the other hand, another fine-grained recognition dataset, the iNaturalist-2017 contains a large amount of data, with 675,170 training and validation images and 5,089 fine-grained categories under 13 super classes. Therefore, the quality of this work can be improved by conducting the performance evaluation on the iNaturalist-2017 dataset, which allows for a greater convincing evaluation on how generalizable the MA-CNN model is.

Appendix I

Figure 1

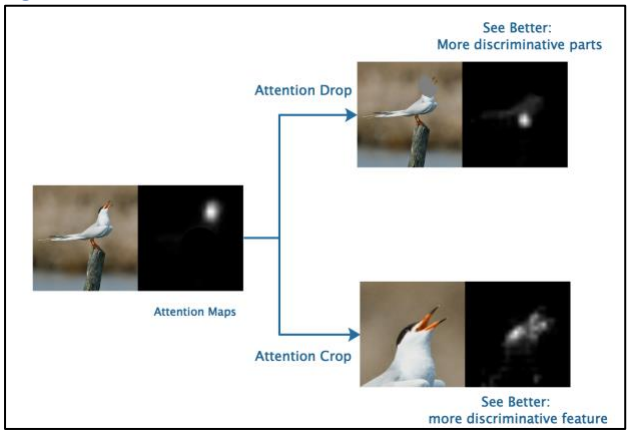

Figure 1: Illustration of the proposed attention guided data augmentation to enhance local feature representation and to extract feature from multiple object discriminative regions.

Figure 2

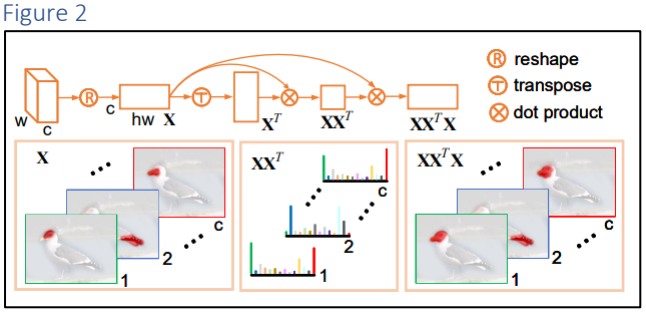

Figure 2: Illustration of the proposed trilinear attention module to integrate related feature maps.

References

[1] Zheng, Heliang & Fu, Jianlong & Zha, Zheng-Jun & Luo, Jiebo. (2019). Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-grained Image Recognition.

[2] Hu, T., & Qi, H. (2019). See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. *CoRR, abs/1901.09891*.