International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Deep Attention Networks for Fine-Grained Image Classification in Low-Resource Settings

Mrs M.Ramya[1], Mrs.P.Gnanapriya[2], Mr.Murugaraja[3], Dr.Gajendran[4]

Assistant Professor, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India

ramya.m@shanmugha.edu.in

Assistant Professor, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India

gnanapriya@shanmugha.edu.in

Assistant Professor, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India

murugaraja@shanmugha.edu.in

Professor, Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India

gajendran@shanmugha.edu.in

Abstract- Fine-grained image classification is a challenging task due to the large inter-class difference and small intra-class difference. In this paper, we propose a novel Cascade Attention Model using the Deep Convolutional Neural Network to address this problem. Our method first leverages the Spatial Confusion Attention to identify ambiguous areas of the input image. Two constraint loss functions are proposed: the Spatial Mask loss and the Spatial And loss; Second, the Cross-network Attention, applying different pre-train parameters to the two stream architecture. Also, two novel loss functions called Cross-network Similarity loss and Satisfied Rank loss are proposed to make the two-stream networks reinforce each other and get better results. Finally, the Network Fusion Attention merges intermediate results with the novel entropy add strategy to obtain the final predictions. All of these modules can work together and can be trained end to end. Besides, different from previous works, our model is fully weak-supervised and fully paralleled, which leads to easier generalization and faster computation. We obtain the state-of-the-art performance on three challenge benchmark datasets (CUB-200-2011, FGVC-Aircraft and Flower 102) with results of 90.8%, 92.1%, and 98.5%, respectively.

Keywords – Deep Convolutional Neural Network, Spatial Confusion, Spatial Mask loss, novel loss functions.

INTRODUCTION

Unlike the coarse-grained image classification e.g. Image Net (Deng et al., 2009), which only needs to classify the general category of the objects (e.g., a flower, a bird or an aircraft), the fine-grained image classification needs to further discriminate subtle differences among various sub-

classes of a given object category[1]. While we have observed a significant performance improvement for the coarse-grained classification (He, Zhang, Ren, & Sun, 2016; Huang, Li, Van Der Maaten and Weinberger, 2017; Simonyan & Zisserman, 2014; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), thanks to the advance of deep learning, the fine-grained image classification remains a challenging problem[2]. Due to the background noise in the real-world photographs, it is important to isolate the object/part before performing the classification. To this end, many part based attention methods have been proposed. Strongly supervised methods (Huang, Xu, Tao, & Zhang, 2016; Lin, Shen, Lu and Jia, 2015; Zhang, Donahue, Girshick, & Darrell, 2014; Zhang et al., 2016) are often built on certain object detection models like R-CNN (Girshick, Donahue, Darrell, & Malik, 2014) or YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016), which need the user to label object's bounding box manually. On the other hand, weakly supervised methods (Fu, ∗ Corresponding author. E-mail address: yening@njfu.edu.cn (N. Ye). Zheng, & Mei, 2017; He & Peng, 2017b; Simonelli, De Natale, Messelodi, & Bulo, 2018; Sun, Yuan, Zhou, & Ding, 2018; Xiao et al., 2015; Yang et al., 2018; Zheng, Fu, Mei, & Luo, 2017), only require an image-level label for the training[3]. However, these methods are often serial, meaning the classification can be carried unless the previous object localization has been completed successfully. It leads to two kinds of problems. The first is the slower computation, as this method cannot be parallelized with modern GPUs at both stages simultaneously. The second issue is the error propagation if the object of interest is mislocated, the following classification will be most likely to fail. In addition, existing methods only focus on the max activation class, and if there is a tie in softmax outputs[4], it will make the error propagation problem even worse. Another important factor to improve the fine-grained image classification is to learn a better feature representation. It can be done with better backbone networks (Hu, Shen, & Sun, 2017; Huang, Li, Van Der Maaten et al., 2017), better feature extraction structure (Shi, Gong, Tao, Cheng, & Zheng, 2018; Wang, Morariu, & Davis, 2018), better data augumentation (He & Peng, 2018), or better transfer learning (Cui, Song, Sun, Howard, & Belongie, 2018; Qiu et al., 2018). Despite the success of these methods, there are two problems to be solved. First, a powerful method should be designed to fuse different intermediate information to the final output and fully take advantage of different outputs under different circumstances. Second, general pre-trained transfer learning may perform worse while domain-specific transfer learning performs better but needs to take much time to pre-train on large scale datasets for each fine-grained dataset[5].

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

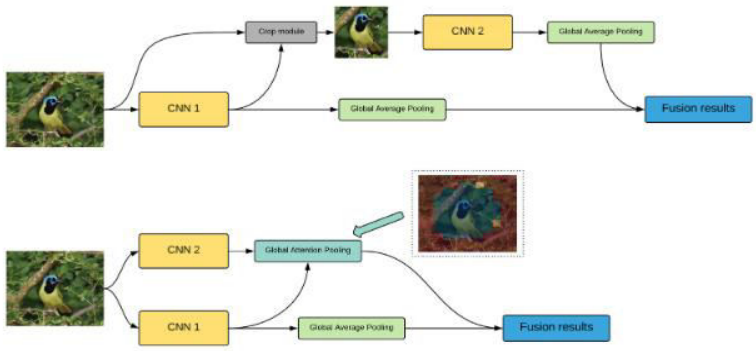

Figure 1- Comparison between previous serial attention methods (upper) and our parallel attention model (lower)[5]

RELATED WORK

Feature representation learning With the advancement of deep learning, automatic feature extraction with modern CNNs has become the mainstream instead of the previous hand-craft method (e.g., Perronnin, Sánchez, & Mensink, 2010; Yao, Brad- ski, & Fei-Fei, 2012) in fine-grained image classification. Hence, learning powerful representation is crucial for improving the performance of the fine-grained image classification[6]. For developing powerful backbone network, DenseNet (Huang, Li, Van Der Maaten et al., 2017) improves residual connection (He et al., 2016) to the densely connected network and gains 77.85% accuracy on the ImageNet test set; SENet (Hu et al., 2017) recalibrates channel-wise feature representation by the Squeeze-and-Excitation block, and it can work with both the residual (He et al., 2016) and inception (Szegedy et al., 2016) module. For better feature exaction, Wang et al. (2018) learned a discriminative filter bank and obtained the state-of-the-art result on CUB bird dataset (Wah, Branson, Welinder, Perona, & Belongie, 2011); Shi et al. (2018) worked with the hierarchical label by the cascaded softmax and generalized large-margin losses[7]. He and Peng (2017a) combined vision and language that introduces the text description as the secondary supervision. Also, Cui et al. (2018) explored the large-scale domain-specific transfer learning for fine-grained image classification. Spatial part attention To locate the discriminative part of the input images, a large amount of attention based methods are proposed. Early works (Huang et al., 2016; Lin, Shen et al., 2015; Zhang et al., 2014; Zhang, Xu et al., 2016) are heavy involvement in human labor that needs bounding box annotations, which are unfit for real-world application. To address this, many weak-supervised methods are proposed, which only needs image level annotation. Recurrent attention network (Fu et al., 2017) iteratively generates region attention from the original image to discriminative part with proposed attention proposal sub-network[8]. The multi-attention convolutional neural network (Zheng et al., 2017) further improves the attention mechanism which can localize multiple parts. HSnet (Lam, Mahasseni, & Todorovic, 2017) searches informative image parts sequentially using LSTM (Hochreiter & Schmidhuber, 1997). Adversarial complementary learning (Zhang, Wei, Feng, Yang and Huang, 2018) discovers new and complementary object regions by erasing its discovered regions from the feature maps. Navigator–Teacher–Scrutinizer Network (Yang et al., 2018) use multi-agent cooperation mechanism to locate the discriminative parts [9-14]. Increasingly specialized ensemble network (Simonelli et al., 2018)

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

exploits a nearly cost-free method to locate the image part for ensemble learning. StackDRL (He, Peng, & Zhao, 2018a) localizes discriminative regions via attention-based deep reinforcement Learning. Compared with these methods, ours has the following advantages. First, we propose the confusion attention mechanism which can pass the confusion of the previous stage to the next stage. It effectively alleviates the error propagation problem. Second, our model can run parallelly, which results in fast computation with modern GPUs. Third, we propose the network fusion attention to take advantage of multiple intermediate outputs and fuse them to the final output [7].

PROPOSED CASCADE ATTENTION MODEL

(i) Spatial confusion attention- The class activation map (Zhou et al., 2016) is a useful way to locate the object in the image by generating the attention heatmap. Existing studies generate the heatmap M as follows:

$$M = max_C(A) \in \mathbb{R}^{h*w}$$

where maxC takes the max along the C axis, that is, only taking the max activation into account. However, the max activation is not always located in the right area, and it may lead to the error propagation problem in the following stage[6]. Based on the fact that the top 5 accuracy in fine-grained image classification is often larger than 98%, one naive approach is to take the top 5 average of the class activation map:

$$M = \frac{\sum_i^5 max_C^5(A)_i}{5} \in \mathbb{R}^{h*w}$$

However, it becomes less efficient when the top 1 prediction is confident. Alternately, based on the property that $\sum_i^C y_i = 1,$ we use a weighted sum to model the confusion attention heatmap: Also, the elements in A can be negative, so we preprocess A with,

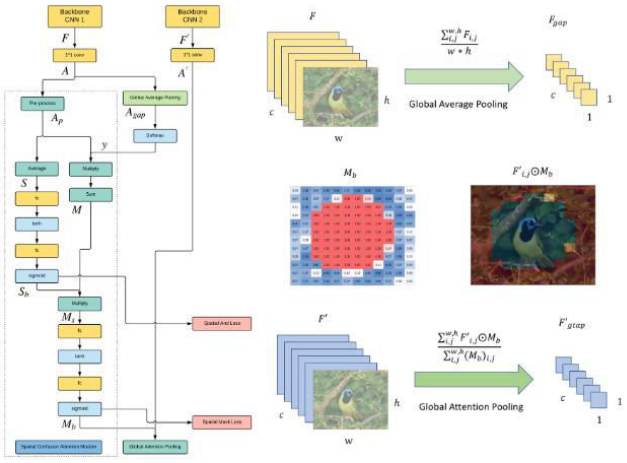

Figure 2- Left: Overview of the spatial confusion attention module. The ‘‘conv’’ stands for the convolutional layer(s) and ‘‘fc’’ stands for the fully connected layer(s)[9]

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Where, $min_{-1}()$ returns the minimum elements in A along the last channel (depth).

To utilize the attention map M, a scale transform operation should be done. That is, we need to convert M to a binary mask. However, the value of M is changing during training, so we cannot manually assign the function which maps each element in M to {0, 1}. Alternately, we consider a small trainable network to do the scale transform[8].

$$M = \sum^{C} y * A_p \in \mathbb{R}^{h*w}$$

$$A_{min} = min_{-1}(A) \in \mathbb{R}^{h*w}$$

$$A_p = A - A_{min} \in \mathbb{R}^{h*w*c}$$

The network should satisfy the following properties: First, the output should be in range of [0, 1], and the most of elements should be either 0 or 1; Second, the 1s should be located at the discriminative region of the input image, neither the whole image nor none of the image. To satisfy the property one, a two-layer 1*1 convolutional network is needed, each layer has one neuron, and the first one uses the tanh activation, and the second uses the sigmoid activation[9].

EXPERIMENTS

To evaluate our model, we conduct experiments on three datasets(CUB-200-2011, FGVC-Aircraft and Flower 102). The detail statistics of the datasets are shown in Table 1.

The statistics of datasets used in this paper.

Datasets	# Classes	# Training	# Testing
CUB-200-2011 (Wah et al., 2011)	200	5994	5794
Flower 102 (Nilsback & Zisserman, 2008)	102	2040	6149
FGVC-Aircraft (Maji, Rahtu, Kannala, Blaschko, & Vedaldi, 2013)	100	6667	3333

Table 1- The statistics of datasets used in this paper[3]

(i) Implementation details- We implement our model by keras (Chollet et al., 2015) library with tensorflow (Abadi et al., 2015) backend. Inception V3 (Szegedy et al., 2016) is used as our backbone CNN. For the ImageNet pre-train parameters, we use build-in parameters provided by keras; for the iNaturalist pre-train parameters, we use parameters trained by Cui et al. (2018). Note that the original iNaturalist pre-train parameters are trained with TensorFlow-Slim (Guadarrama, 2016); we convert it to fit the keras format. We follow the standard evaluate protocol that sets the network input to $448 * 448$. A dynamic data augmentation strategy is used that randomly rotates the image every 15° from 0° to 90° and from 270° to 360° (from 0° to 30° and from 330° to 360° for FGVC-Aircraft dataset). As for training, we only utilize the image-level label and optimize it with the RmsProp algorithm (Tieleman & Hinton, 2012). We set the batch size to 14 (mainly because it is the largest batch size we can set with single GTX-1080 Ti) and learning rate ranging from 1e−4 to 1e−6 (depend on the training data). We set bl = 1/4, bh = 1/2, mr = 0.05, ms = 0.8 and mcns = 0.15.

(ii) CUB-200-2011 experiments- CUB-200-2011 (Wah et al., 2011) is a real-world bird dataset with 200 bird species. Note that to demonstrate the effectiveness of our model, we re-implement the following architectures using inception V3 backbone as the baseline for comparison: ImageNet baseline: Fine-tune inception V3 pre-trained by ImageNet dataset directly. iNaturalist baseline: Fine-tune inception V3 pre-trained by iNaturalist dataset directly. Cross-ConcatNet: Concatenate the two-stream feature after the global average pooling layer and then feed it into full connected layers. One stream is using the ImageNet pre-train parameters and the other is using the iNaturalist pre-train parameters. Cross-DualNet (Hou et al., 2017): Our re-implementation of the DualNet architecture. One stream is using the ImageNet pretrain parameters and the other is using the iNaturalist pre-train parameters. Cross-Bilinear (Lin, RoyChowdhury et al., 2015): Our re-implementation of the Bilinear architecture. One stream is using the ImageNet pre-train parameters and the other is using the iNaturalist pre-train parameters.

(iii) Flower 102 experiments- Flower 102 (Nilsback & Zisserman, 2008) is a plant dataset that contains 102 categories of flowers commonly occurring in the United Kingdom. Note that these methods do not use any part annotations. The inception V3 pre-trained on ImageNet dataset achieves 96.3% accuracy, while the inception V3 pre-trained on iNaturalist dataset achieves 97.7% accuracy. The accuracy further improved to 98.5% through Cascade attention model.

CONCLUSION

In this paper, we propose the Cascade Attention Model in deep convolutional neural network for fine-grained image classification, which consists of three parts: The spatial confusion attention, the cross-network attention, and the network fusion attention. The proposed framework makes two-stream CNN to not only run parallelly but also reinforce each other. The whole model can be trained end to end with the only image-level label. Extensive experiments demonstrate the effectiveness of our method. This study was supported by the National Key Research and Development Plan of China (2016YFD0600101), Jiangsu Provincial Department of Housing and Urban–Rural Development, PR China (2016ZD44), in part by the Practice Innovation Training Program Projects for Jiangsu College Students, PR China under Grant 201810298052Z.

REFERENCES

[1]. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2023). TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org. URL https://www.tensorflow.org/.

[2]. Branson, S., Van Horn, G., Belongie, S., & Perona, P. (2022). Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952.

[3]. Chen, T., Wu, W., Gao, Y., Dong, L., Luo, X., & Lin, L. (2023). Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. arXiv preprint arXiv:1808.04505.

[4]. Chollet, F., et al. (2022). Keras. https://keras.io. Cui, Y., Song, Y., Sun, C., Howard, A., & Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer

learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4109–4118).

[5]. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2022). ImageNet: A large-scale hierarchical image database. In CVPR09.

[6]. Feng, H., Wang, S., & Ge, S. S. (2023). Fine-grained visual recognition with salient feature detection. arXiv preprint arXiv:1808.03935.

[7]. Fu, J., Zheng, H., & Mei, T. (2022). Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In Conf. on computer vision and pattern recognition.

[8]. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2023). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580–587).

[9]. Guadarrama, N. S. S. (2023). Tensorflow-slim: a lightweight library for defining, training and evaluating complex models in tensorflow. URL https://github. com/tensorflow/tensorflow/tree/master/tensorflow/contrib/slim.

[10]. Sanjay Kumar Suman, Dhananjay Kumar and L. Bhagyalakshmi, "Non Cooperative Power Control Game with New Pricing for Wireless Ad hoc Networks", International Review on Computers and Software, vol. 9, no. 1, pp. 18-28, 2014. ISSN: 1828-6003,

[11]. S. Porselvi, Sanjay Kumar Suman and L. Bhagyalakshmi, "Harvesting RF energy for mobile charging", Australian Journal of Basic and Applied Science, vol. 9, no. 20, pp. 454-465, June 2015.

[12]. K. Swapna, P. Rajalakshmi and Sanjay Kumar Suman, "Security Enhancement in MANET using Game Theory", Middle East Journal of Scientific Research, vol. 23, pp. 190-195, 2015.

[13]. VinaySrivatsan, Sanjay Kumar Suman, L. Bhagyalakshmi and S. Porselvi, "Non radiative wireless power transfer", Journal of Advances in Natural and Applied Sciences, vol. 10, no. 16, pp. 147-153, Nov. 2016.

[14]. Sujeetha Devi, Bhagyalakshmi L and Sanjay Kumar Suman, "Cluster based energy efficient joint routing algorithm for delay minimization in wireless sensor networks", International Journal of Pure and Applied Mathematics, vol. 119, no. 15, 307-313, 2018