FINE-GRAINED MULTI-INSTANCE CLASSIFICATION IN MICROSCOPY THROUGH DEEP ATTENTION

Mengran Fan[1] *Tapabrata Chakraborti*[1] *Eric I-Chao Chang*[2] *Yan Xu* [2,3] *Jens Rittscher*[1]

[1] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2] Microsoft Research, Beijing, China
[3] Department of Biology and Medicine, Beihang University, Beijing, China

ABSTRACT

Fine-grained object recognition and classification in biomedical images poses a number of challenges. Images typically contain multiple instances (e.g. glands) and the recognition of salient structures is confounded by visually complex backgrounds. Due to the cost of data acquisition or the limited availability of specimens data sets tend to be small. We propose a simple yet effective attention based deep architecture to address these issues, specially improved background suppression and recognition of multiple instances per image. Attention maps per instance are learnt in an end-to-end fashion. Microscopic images of fungi and a publicly available Breast Cancer Histology benchmark data set are used to demonstrate the promise of the proposed approach. Our algorithm comparison suggests that our approach advances the state of the art.

Index Terms— attention models, fine-grained classification, object recognition, medical image analysis, deep learning, convolutional neural networks

1. INTRODUCTION

Fine-grained image classification, which focuses on localizing discriminative regions and recognizing subtle visual differences between sub-classes, is an open problem in biomedical image analysis. Recently, a number of deep learning classification models have been proposed on large-scale open datasets of natural images. However, there are several challenges that currently limit the adoption of these methods to biomedical domain, especially for fine-grained problems.

Firstly, a large and diverse dataset is essential for deep learning models to effectively recognize informative patterns and suppress the confounding effect of background. However, publicly available benchmark datasets in this domain tend to be small due to restrictions in data sharing and high image acquisition costs. To address this problem we design a small and lightweight module to reject the deep feature maps that represent background clutter or other irrelevant regions. With more informative feature maps, we can then force the network to make the predictions that are based on biologically or medically interpretable features.

Secondly, in contrast to open datasets of natural images, where there is mostly single object in each image, target objects in these images may have multiple instances and a wide variety of sizes and densities. Consequently, automatic classification systems for microscopy images would be expected to capture the global structure of multiple instances as well as the low-level details of each instance. After filtering out irrelevant feature maps, we adopt the non-local self-attention mechanism to further optimize the boundary information of instances. This enables the network to automatically discover and localize the whole/complete objects without any bounding box or extra information.

Microscopy images are usually characterized by high resolution, which has not been fully taken advantage of traditional approaches and thus have caused a loss of information after extracting patches or performing down-sampling. Instead, we first downscale the original images and feed these lower-resolution images to a simple network for capturing global structural features such us object densities or the number of objects. With the output of the global attention mechanism, we then crop and amplify the corresponding patches directly from original images in order to utilise the relevant high-resolution information.

While there is a substantial body of work in fine-grained classification in computer vision, few groups have addressed challenges in biomedical applications. The model we propose provides a number of improvements when compared with existing fine-grained recognition methods. The popular NTS model [1] relies on a large amount of predefined region proposals to select the most informative regions. Zheng et al. [2] proposed a channel grouping strategy that groups the spatially correlated channels to generate part-based attention maps. Rodriguez et al. [3] designed a gated attention mechanism to highlight discriminative parts. Similarly, in [4], the authors incorporated self-attention mechanisms to perform the object and part localization. Apart from fine-grained classification, attention mechanisms have been extensively studied in other vision tasks [5, 6, 7].

Contributions. Firstly, our approach effectively sup-

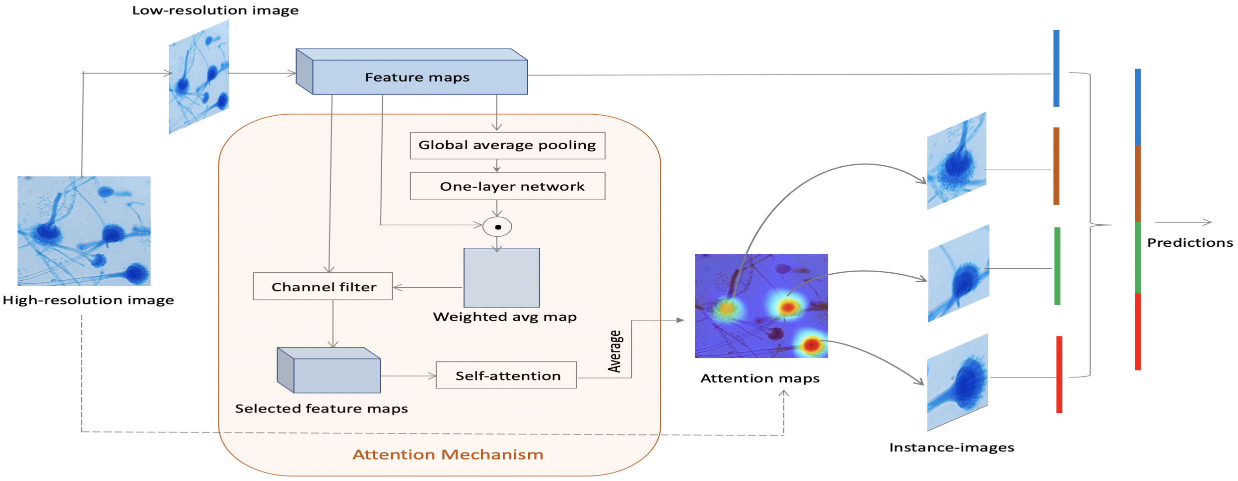

Fig. 1. **Framework for the proposed multiple instance fine-grained classification pipeline.** The network consists of three modules: feature extraction, attention module and feature fusion.

presses the background and other uninformative regions. Secondly, our model allows for the detection of multiple instances. The attention maps per instance are learnt end-to-end without the need for any additional annotations. Our results on the breast cancer histology data set (BACH) are an improvement on previously reported results.

2. NETWORK ARCHITECTURE

We propose a novel fine-grained multi-instance classification scheme that consists of three main modules: (i) a simple and lightweight CNN that extracts global structural features from images; (ii) a novel attention mechanism for localizing multiple instances and (iii) a feature learning framework that consolidates the global and local features to facilitate the final predictions.

2.1. Attention mechanism

The proposed attention mechanism aims at automatically localizing complete and discriminative instances in fine-grained images without any extra supervision or redundant region proposals. As shown in Figure 1, the whole model can be divided into three procedures: (a) semantic modelling framework, which groups the channels via weighed averages with learnable weights; (b) channel filter for rejecting the feature maps highlighting the irrelevant regions; (c) lightweight self-attention model, which globally optimizes the selected feature maps.

Semantic Modeling Framework: It is aimed at generating a aggregated global feature map with strong semantic meaning (roughly highlighting the main objects but discarding the noisy background). We directly build it by calculating a weighted average of original feature maps obtained from the low-resolution images. This approach rests on the assumption that if many channels focus on the same region, we could expect this region to be part of objects rather than noise or background. We opt for the simplest channel-wise aggregation technique, introduced in the squeeze-excitation (SE) block [6], noting that more comprehensive strategies could be adapted.

To learn the importance weight for each channel, we construct channel descriptors by global average pooling. Subsequently we feed it to a one-layer network. The global semantic map is then constructed by computing a weighted average using the learnt weights. With features maps $\mathbf{X} \in \mathcal{R}^{H\times W\times C}$ the aggregated semantic map can be written as:

$$w = \sum_{p=1}^{C} F(x_p, \delta(\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_p(i,j))), \qquad (1)$$

where $\delta(.)$ denotes the feature transform with fully-connected layers and $F(.,.)$ refers to a rescaling function to fuse the importance weights and original feature maps (broadcast element-wise multiplication in this case). It is important to note that the main difference between our design and SE block lies in its intention. We aim at aggregating a global weighted semantic map for subsequent transformations (channel filtering) rather than directly emphasizing good channels/features as an attention module.

Channel Filter : In terms of the large-scale image dataset, background or irrelevant parts can be naturally discarded in a deep neural network. However, for the limited microscopy imaging data, an extra operation selecting effective deep descriptors would be necessary for noisy information removal and improved object localization. Here, we propose a simple yet effective method to filter out the noisy feature maps. With the global semantic map w that roughly indicates the position of main objects, we calculate the pairwise similarities between itself and each original feature map x_p based on their activation responses. Consequently, we sort these similarity values in the descending order and only select first $N (N < C)$ channels for subsequent transformations.

Self-attention Module : After filtering, the selected low-dimensional feature descriptors are obtained. As mentioned before, it is important to attend all relevant parts of the object to keep the sufficient fine-grained information. Therefore, we further recalibrate the activation responses by employing a self-attention module [4], where global long-range channel interactions can be captured explicitly to enhance the object boundaries and highlight the useful regions. The self-attention module can be formulated as:

$$\mathcal{M} := \mathcal{N}(\mathcal{N}(X)X^T)X, \qquad (2)$$

where $(N)(.)$ indicates $softmax$ normalization, $(N)(X)X^T$ refers to the pairwise inter-channel similarities. The attention maps are obtained by performing a dot product of $(N)(X)X^T$ and X. Finally, all generated attention maps are added along the depth direction for the aggregated global attention map, which can used to localize the main objects. The computational burden of this self-module is very low as it is only applied in the selected low-dimensional feature maps.

2.2. Feature Learning

When multiple informative regions are proposed, there are three different feature learning approaches for fine-grained classification: (i) hard sampling that crops the patches from original image, (ii) soft sampling where regions with high responses are sampled more densely and (iii) deep descriptors selection, which directly selects the useful deep descriptors in high-level feature maps. Compared to deep descriptors selection, resampling from original images can make good use of the high-resolution information, and directly enhance the reliability and interpretability by providing extracted patches to a human expert as well as final predictions. Therefore, in this work, we focus on sampling strategies ((i) and (ii)), which aim to sample the informative pixels from original image. By empirical analysis, we have found that most soft sampling methods result in spatial distortions and fail to keep the shape information of instances as illustrated in Fig. 2. To that end, we adopt a hard sampling strategy to generate multiple informative patches. In particular, we perform thresholding and connected component analysis on the global attention map. With the bounding box positions and corresponding size, N sampled images containing main objects are generated by cropping squares from original high-resolution images. Each of the cropped patches are amplified into a larger resolution (e.g., 336 × 336) and fed to a CNN-based network for capturing local instance-level details. Finally, we consolidate these with high-level structural features for final predictions.

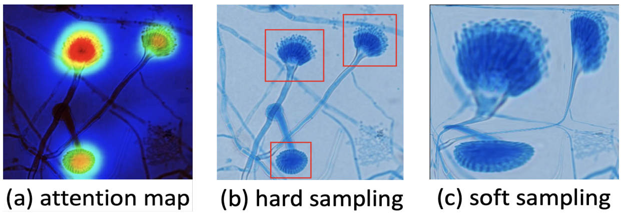
(a) attention map (b) hard sampling (c) soft sampling

Fig. 2. A comparison of hard sampling and soft sampling. It can be seen that soft sampling may result in spatial distortion. To keep the shape information of multiple instances, we conduct hard sampling to generate informative regions.

3. EVALUATION

3.1. Implementation details

The framework was implemented using open-sourced MXNet [8] libraries. In all our experiments, images were first resized to 224 x 224, and a pre-trained Resnet-18 [9] was used to extract global structural information from these down-sampled images. Note that other CNN architectures could also be used here. After resampling from original high-resolution images, images were rescaled to a size of 336 x 336, and we fix N=2, which means 2 regions were used to feed to pre-trained Resnet-50 [9] for local instance-level features. We employ standard cross entropy loss for our classification task. Batch size is set to 16 and and SGD optimizer is used with an initial learning rate of 0.05 that multiplied by 0.1 after 50 epochs.

3.2. Experimental Results

Method	top-1 % accuracy
Residual Attention Network (RAN) [5]	0.867
Attend and Rectify [3]	0.871
Trilinear Attention Module[4]	0.883
NTS Module [1]	0.914
Our Method	**0.943**

Table 1. Fungal dataset. The proposed methods nearly achieves a 3% improvement when compared to other state of the art methods.

Attention Mechanism	top-1 % accuracy
Trilinear channel-wise self-attention [4]	0.883
Non-local spatial self-attention [10]	0.859
Dual self-attention [11]	0.901
SE Net [6]	0.939
GC Net[7]	0.937
Our Method	**0.943**

Table 2. Comparison of global attention mechanisms.

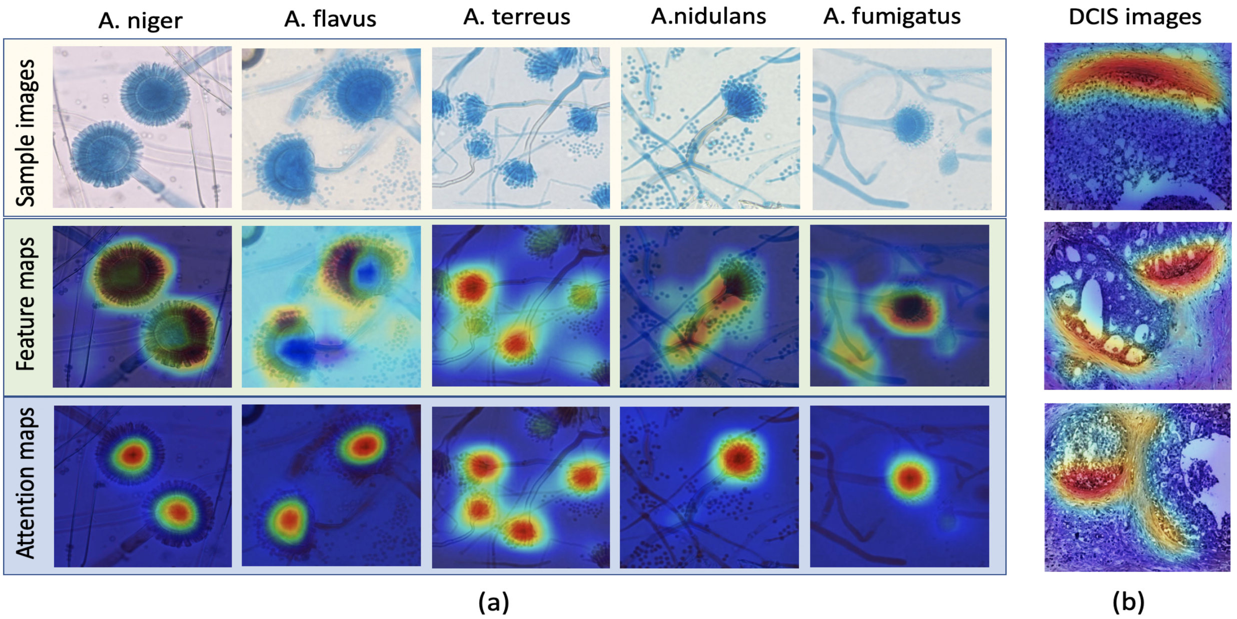

Fig. 3. **Visualization of global attention maps.** (a) Each column shows a sample image from each class represented in the fungal dataset. Our attention maps suppress features associated to background clutter and focus on whole objects. (b) In histology images of ductal carcinoma in situ (DCIS), our attention maps highlight disease relevant features at tissue interfaces.

Datasets. We evaluated our proposed framework on two datasets: microscopic images of fungi collected at the Peking Union Hospital and the benchmark dataset from the 2018 grand challenge in Breast Cancer Histology images (BACH) [12].

Aspergillosis is a group of diseases resulting from aspergillus infection fungal growth as well as allergic responses. The automatic and precise classification of species is crucial for improving survival rates in patients with life-threatening fungal infections. Here, in collaboration with the Peking Union Hospital we collected 2000 clinical images representing 5 very common Aspergillus species: A. fumigatus, A. flavus, A. niger, A. terreus and A.nidulans (400 images for each class). We conduct comprehensive experiments on this dataset and demonstrate that our proposed model achieves best performance compared with the state-of-the-art fine-grained classification algorithms. The classification accuracy are summarized in Table 1, and a comparison of feature maps and obtained attention maps are shown in Figure 3. Furthermore, we conduct quantitative comparison on different attention mechanisms. For fair comparison, all compared methods use the same backbone network and feature learning strategy, but with different attention algorithms.

In addition, we demonstrate the generalization ability of this work by applying to a benchmark breast cancer dataset, which provides 400 Hematoxylin and eosin stained breast histology microscopy images. Microscopy images are labeled as normal, benign, in situ carcinoma or invasive carcinoma according to the predominant cancer type in each image. As shown in Table 3, we compare our framework with five recently reported methods and our model also improve the accuracy on histology dataset.

Method	top-1 % accuracy
CNN+SVM [13]	0.925
DenseNet-161[14]	0.940
ResNet-152[15]	0.830
Model Fusion[16]	0.925
Ensemble with refinement[17]	0.875
RFSVM-All[15]	0.930
Our Method	**0.960**

Table 3. **Experimental results in breast cancer dataset.**

4. CONCLUSION

Advancing the start of art in the classification of fine-grained structures has important implications for biomedical imaging. Our effectively suppreses confounding effects of irrelevant regions and localizes multiple objects per image even if the size of the dataset is limited. Importantly, no additional annotations are required. Here, high-level structure information and local details of multiple instances are combined to improve classification. Quantitative and qualitative experimental results demonstrate the promise of our approach.

5. REFERENCES

[1] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 420–435.

[2] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[3] Pau Rodríguez, Josep M Gonfaus, Guillem Cucurull, F XavierRoca, and Jordi Gonzalez, "Attend and rectify: a gated attention mechanism for fine-grained recovery," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 349–364.

[4] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5012–5021.

[5] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[6] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[7] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," *arXiv preprint arXiv:1904.11492*, 2019.

[8] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[12] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al., "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, 2019.

[13] Yaqi Wang, Lingling Sun, Kaiqiang Ma, and Jiannan Fang, "Breast cancer microscope image classification based on cnn with image deformation," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 845–852.

[14] Matthias Kohl, Christoph Walz, Florian Ludwig, Stefan Braunewell, and Maximilian Baust, "Assessment of breast cancer histology using densely connected convolutional networks," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 903–913.

[15] Hongliu Cao, Simon Bernard, Laurent Heutte, and Robert Sabourin, "Improve the performance of transfer learning without fine-tuning using dissimilarity-based multi-view learning for breast cancer histology images," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 779–787.

[16] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 737–744.

[17] Yeeleng S Vang, Zhen Chen, and Xiaohui Xie, "Deep learning framework for multi-class breast cancer histology image classification," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 914–922.