DEEP ATTENTIVE FEATURE LEARNING FOR HISTOPATHOLOGY IMAGE CLASSIFICATION

Pengxiang Wu[†], *Hui Qu*[†], *Jingru Yi*[†], *Qiaoying Huang*[†], *Chao Chen*[⋆], *Dimitris Metaxas*[†]

[†]Department of Computer Science, Rutgers University, NJ, USA
[⋆]Department of Biomedical Informatics, Stony Brook University, NY, USA

ABSTRACT

In this paper, we present a new deep learning-based approach for histopathology image classification. Our method is built upon standard convolutional neural networks (CNNs), and incorporates two separate attention modules for more effective feature learning. In particular, the attention modules infer the attention maps along different dimensions, which help focus the CNNs on critical image regions, as well as highlight discriminative feature channels while suppressing the irrelevant information with respect to the classification task. The attention modules are light-weight, and enhances the feature representation with small extra computational overhead. Experimental results on the publicly available BreakHis dataset demonstrate that our method outperforms the state-of-the-arts by a large margin.

***Index Terms*—** Histopathology image analysis, breast, convolutional neural network, attention, transfer learning

1. INTRODUCTION

Microscopic histopathological examination using a tissue biopsy has been widely used in cancer diagnosis and is considered confirmatory gold standard in practice. Diagnostic report, including grading and staging, is typically completed by experienced pathologists through visually inspecting the histological samples. With the recent advances in image processing, it becomes increasingly possible to automate such histopathology analysis, thereby assisting the pathologists to be more productive and objective. As one of the primordial tasks, classification of histopathology images has gained much attention in recent years. However, such classification task is quite challenging due to the inherent complex visual patterns of histopathology images.

Early works on histopathology image classification mainly rely on handcrafted features extracted from the whole image or segmented patches [1, 2, 3]. While being interpretable, handcrafted features are typically unsatisfactory for this task due to their limited description of the images. Inspired by recent advances in deep learning, several methods have been developed to employ convolutional neural networks (CNNs) for automatic image feature learning, which has been shown to achieve better performance than the handcrafted design [4, 5]. However, one major weakness with CNN-based models is that they typically require massive data for training. To mitigate the data-intensive issue, one common strategy is to fine-tune the models pre-trained on large-scale image dataset (e.g., ImageNet) [6]. Another different class of methods simply utilizes the pre-trained CNNs as feature extractors and then applies Fisher Vector (FV) encoding for global feature representation [7, 8, 9]. While achieving the state-of-the-art results, these methods tend to generate features with redundancy and noise, which are adverse to the classification.

In this paper, we propose a new CNN architecture and improve the feature representation for histopathology image (patch) classification from a different perspective. At the core of our method is an attention mechanism, which helps the CNN focus on regions and feature channels that are critical to the classification task. Our key motivation is from the human vision system: when perceiving a scene, humans first glance at the scene and then instantly attend to the salient contents while ignoring the irrelevant information. We implement such mechanism as *attention maps* through global feature correlation analysis. Specifically, inspired by the Transformer [10] and non-local neural networks [11], we design two attention modules, which infer the attention maps along channel and spatial dimensions, respectively (see Fig. 1). The channel-wise attention (C-Attn) module allows the network to concentrate on discriminative feature channels and reduce the redundancy, while the spatial attention (S-Attn) module highlights the useful regions and suppresses the irrelevant ones for the network. These two modules enhance the discriminative learning ability collaboratively, and can be integrated into arbitrary existing CNN architectures in a plug-and-play manner. In practice, we adopt VGG19 [12] as the base model and insert the attention modules at different positions, as shown in Fig. 1. We apply our method to the task of benign and malignant breast caner classification, and on the publicly available BreakHis dataset [2] we demonstrate the superiority of our method compared to the state-of-the-arts.

In the following sections we present the details of our method, and provide the experimental results and discussions.

This work was partially supported by NSF IIS-1855759, CCF-1855760, and CCF-1733843.

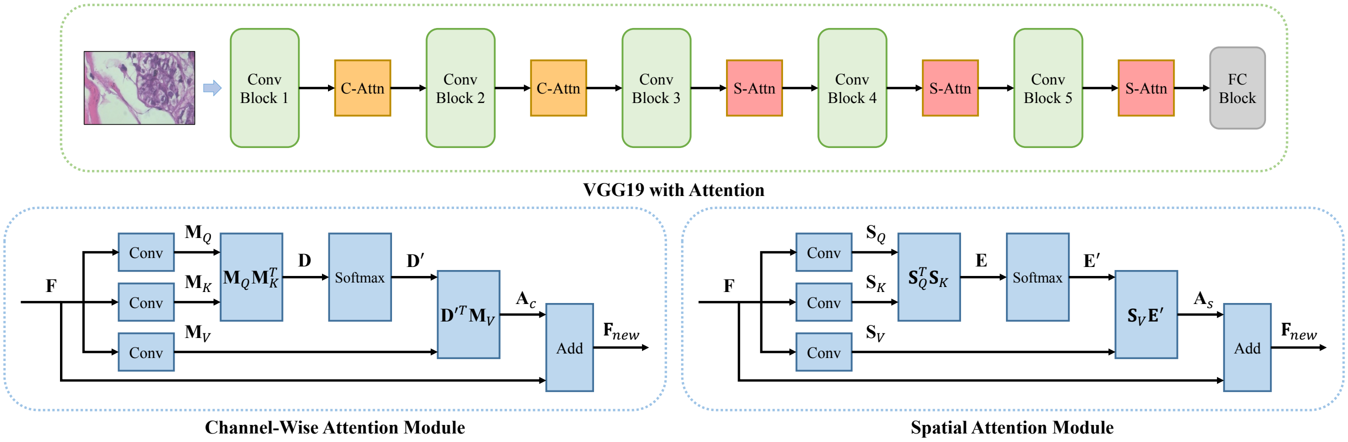

Fig. 1. Illustration of the attention modules, which are placed at different positions of VGG19 network.

2. METHOD

Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as input (C, H and W are the channel number, height and width of $\mathbf{F}$, respectively), the C-Attn and S-Attn modules infer the attention maps $\mathbf{A}_c$ and $\mathbf{A}_s$ along the channel and spatial dimensions, respectively. The generated attention maps $\mathbf{A}_c$ and $\mathbf{A}_s$ are then applied to $\mathbf{F}$ for feature refinement. Below we illustrate the details of each module.

2.1. Channel-Wise Attention Module

The C-Attn module produces the attention map by explicitly modeling the inter-channel relationships of features. The key motivation is that different channels typically correspond to different patterns, and only a portion of them are useful for the classification task. Therefore, the C-Attn module is designed to help the CNN focus on the discriminative patterns and reduce the redundancy by suppressing the non-discriminative ones. In this way, it prevents the most discriminative features from being averaged out by the background channels.

Specifically, given input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, C-Attn module first generates the channel-wise feature vectors:

$$\mathbf{M} = \mathcal{R}\left(W_c\mathbf{F}\right). \tag{1}$$

W_c represents the weights of a 2D Conv layer without bias, which produces a feature map of size $C \times H' \times W'$. $\mathcal{R}$ is the reshape operation and generates $\mathbf{M} \in \mathbb{R}^{C\times N}$, where $N = H' \times W'$. Here, $\mathbf{M}$ can be seen as a set containing C vectors of length N. In a similar spirit to Transformer [10], Eq. (1) is applied to $\mathbf{F}$ twice with different Conv weights, leading to different sets of vectors $\mathbf{M}_Q$ and $\mathbf{M}_K$. Meanwhile, by fixing the 2D Conv in Eq. (1) to be 1×1, the feature map $\mathbf{F}$ is transformed into another vector set $\mathbf{M}_V$ (see Fig. 1). Afterwards, we capture the inter-channel relationships by computing the channel-wise statistics:

$$\mathbf{D} = \mathbf{M}_Q\mathbf{M}_K^T, \tag{2}$$

$$\mathbf{D}' = \mathrm{softmax}(\mathbf{D}), \tag{3}$$

where the softmax operation is applied column-wise:

$$\mathbf{D}'_{ij} = \frac{\exp(\mathbf{D}'_{ij})}{\sum_i^C \exp(\mathbf{D}'_{ij})}. \tag{4}$$

Then we compute the attention map as weighted sum of feature vectors:

$$\mathbf{A}_c = \mathbf{D}'^T\mathbf{M}_V. \tag{5}$$

Finally, the attention map $\mathbf{A}_c$ is added to the input feature map $\mathbf{F}$ for feature refinement:

$$\mathbf{F}_{new} = \mathbf{F} + \mathcal{R}'(\mathbf{A}_c), \tag{6}$$

where the operation $\mathcal{R}'$ reshapes $\mathbf{A}_c$ back to $C \times H \times W$. In place of $\mathbf{F}$, the refined $\mathbf{F}_{new}$ is then fed forward to the subsequent layers. Since the attention map $\mathbf{A}_c$ learns the long-range semantic dependencies among feature channels, it is able to highlight the class-specific discriminative features and thus help improve the classification performance. Note that $\mathbf{A}_c$ can also be interpreted as a residual component, which has been verified to be beneficial to the feature learning [13].

2.2. Spatial Attention Module

Unlike C-Attn, the spatial attention allows the network to concentrate on useful regions and suppress the background information. It works in a similar manner to C-Attn, but with a focus on pixels instead of feature channels.

For an input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, it is first linearly transformed and reshaped as follows:

$$\mathbf{S} = \mathcal{R}\left(W_s\mathbf{F}\right), \tag{7}$$

where W_s is the weights of a 1×1 Conv layer whose output feature map is of size $C' \times H \times W$. In practice, to reduce the computational overhead, C' is set as $C' = C/r$, with reduction ratio $r \geq 2$. The convolutional output is transformed by the reshape operation $\mathcal{R}$ into a set of N feature vectors with length C', i.e., $\mathbf{S} \in \mathbb{R}^{C' \times N}$, where $N = H \times W$. Similar to C-Attn, Eq. (7) is applied to $\mathbf{F}$ three times, thereby generating $\mathbf{S}_Q$, $\mathbf{S}_K$ and $\mathbf{S}_V$, among which the vector set $\mathbf{S}_V$ is created without channel reduction, i.e., $r = 1$. Then we compute the spatial correlation between different positions as:

$$\mathbf{E} = \mathbf{S}_Q^T \mathbf{S}_K, \tag{8}$$

$$\mathbf{E}' = \mathrm{softmax}(\mathbf{E}), \tag{9}$$

where the softmax is performed column-wise:

$$\mathbf{E}'_{ij} = \frac{\exp(\mathbf{E}'_{ij})}{\sum_i^N \exp(\mathbf{E}'_{ij})}. \tag{10}$$

Note that similar feature vectors would correspond to high correlation values in $\mathbf{E}'$. Finally, we generate the attention map $\mathbf{A}_s$ and add it back to the input feature map:

$$\mathbf{A}_s = \mathbf{S}_V \mathbf{E}', \tag{11}$$

$$\mathbf{F}_{new} = \mathbf{F} + \mathcal{R}'(\mathbf{A}_s). \tag{12}$$

The attention map $\mathbf{A}_s$ learns and aggregates the global context into the refined feature map $\mathbf{F}_{new}$, thereby effectively guiding the network to concentrate on more critical region information for the classification task.

2.3. Arrangement of the Attention Modules

The two different attention modules compute complementary attention and can be inserted into CNN collaboratively. In the case of VGG19, we empirically observe that placing C-Attn and S-Attn on the bottom and top layers respectively gives the best results (see Fig. 1). We also experiment with inserting the attention modules at other positions (e.g., after all convolutional blocks), but obtain slightly inferior performance. The reason would be that the bottom layers contain more redundant channels which correspond to the low-level background signals, while the top layers have rich semantic features where the critical regions are not salient.

3. EXPERIMENTS AND RESULTS

We evaluate the proposed method on the publicly available BreakHis dataset [2]. This dataset consists of hematoxylin and eosin (H&E) stained microscopy biopsy images of benign and malignant breast tumors. The images are collected from 82 patients and captured at four different magnifications (40×, 100×, 200× and 400×). In total there are 7909 images, with 2480 benign and 5429 malignant cases. Each image is a patch from the whole slide and has size 700 × 460 × 3.

In the implementation we use VGG19 as the base model, where the two different attention modules are integrated. Note that our method is generic and not limited to VGG19. The weights of the convolutional blocks in VGG19 are fine-tuned while the remaining fully connected layers and the attention modules are trained from scratch. The input images are normalized by the data mean and variance, and are resized to 224 × 224, followed by random flip and rotation. For model training, we use Adam optimizer and set the learning rate to 0.00005, with a decay of 0.5 every 30 training epochs. The training batch size is 8, and the weight decay is chosen to be 0.0001. For C-Attn, we set both the kernel size and stride of the 2D convolution for generating $\mathbf{M}_Q$ and $\mathbf{M}_K$ to 4. For S-Attn, we set the reduction parameter in $\mathbf{S}_Q$ and $\mathbf{S}_K$ to $r = 8$. The code[1] is implemented with PyTorch and executed on a single NVIDIA GTX 1080 Ti GPU.

Table 1. The image-level classification accuracies (%) on BreakHis dataset. Our method incorporates both the spatial and channel-wise attentions, and achieves the best results. ("C": C-Attn module. "S": S-Attn module.)

Method	Magnification			
	40×	100×	200×	400×
CNN-r [4]	89.6±6.5	85.0±4.8	82.8±2.1	80.2±3.4
CNN-m [4]	85.6±4.8	83.5±3.9	82.7±1.7	80.7±2.9
FV-dr [7]	87.0±2.6	86.2±3.7	85.2±2.1	82.9±3.7
FV-ada [9]	87.5±1.6	88.6±3.6	85.5±2.0	85.0±4.6
VGG19	89.1±3.4	90.3±3.9	90.6±2.8	87.4±2.4
VGG19 + C	91.1±2.6	91.5±3.5	92.0±2.2	88.6±2.3
VGG19 + S	89.9±3.0	91.7±2.5	92.2±1.9	89.2±2.5
Ours	**91.4±3.0**	**92.2**±2.9	**93.4**±2.3	**90.0**±2.2

Table 2. The patient-level classification accuracies (%) on BreakHis dataset. Our method achieves the best performance. ("C": C-Attn. "S": S-Attn.)

Method	Magnification			
	40×	100×	200×	400×
PFTAS[2]	81.6±3.0	79.9±5.4	85.1±3.1	82.3±3.8
Vote [3]	87.2	88.2	88.9	85.8
CNN-r [4]	88.6±5.6	84.5±2.4	83.3±3.4	81.7±4.9
CNN-m [4]	90.0±6.7	88.4±4.8	84.6±4.2	86.1±6.2
CNN-st [5]	83.1	83.2	84.6	82.1
MIL-CNN[6]	89.5	89.1	88.8	87.7
FV-dr [7]	90.0±3.2	88.9±5.0	86.9±5.2	86.3±7.0
FV-ada [9]	88.5±2.7	90.8±4.4	89.2±3.2	89.2±7.9
VGG19	95.7±3.9	96.4±3.6	96.4±2.5	92.9±3.6
VGG19 + C	97.1±3.0	98.6±2.0	98.6±2.0	95.7±3.0
VGG19 + S	97.9±3.2	98.6±2.0	97.9±3.2	94.3±4.8
Ours	**97.9±3.2**	**99.3**±1.6	**98.6±2.0**	**96.4**±2.5

We perform 5-fold validation in the experiment, and follow the train/test split provided by the BreakHis dataset: 70% of the images are used for training and 30% for testing. In the evaluation, we measure the accuracies at both image and

[1] Code is available at https://github.com/pxiangwu/attn-hist-classify.

patient levels. The patient-level accuracies are obtained by majority voting using the image-level classification results. Following the setting of [9], we train our model with all the available training images regardless of the magnification factors. This makes the task more challenging due to the image heterogeneity.

We compare our method with several existing works, including PFTAS [2], Vote [3], CNN-r [4], CNN-m [4], CNN-st [5], MIL-CNN [6], FV-dr [7] and FV-ada [9]. In particular, PFTAS and Vote are based on handcrafted features, while CNN-r, CNN-m, CNN-st and MIL-CNN employ CNN for automatic feature learning. FV-dr and FV-ada follow a different route by utilizing Fisher Vector to further encode the learned convolutional features into lower-dimensional space. To validate the effectiveness of the attention modules, we also conduct several ablation studies. In particular, we fine-tune a baseline VGG19 model using the same experimental setting. Besides, we experiment with removing the S-Attn modules from our method, leading to a model with C-Attn only (i.e., VGG19 + C-Attn, as shown in Table 1 and 2). Similar ablation is also applied to the C-Attn modules, and thereby gives a model with spatial attention only (i.e., VGG19 + S-Attn).

We report the results in Table 1 and Table 2. Note that some of the existing works only provide the patient-level results, and a few of them only give the average accuracies without reporting the standard deviations. From the experimental results it can be observed that our method outperforms the existing methods by a large margin. This demonstrates the advantages of attention modules. In particular, compared to the vanilla VGG19 model, our method achieves much performance gain due to the benefits of attention. Besides, the classification accuracy decreases when removing either one type of attention modules. This shows that the two modules are complementary. However, even equipped with only one attention mechanism, the resulting model still outperforms the baseline, which verifies the effectiveness of attention. Note that such performance gain comes with a small parameter overhead, indicating that the improvement is not due to the increased network capacity brought by attention modules, but instead from the attentive feature refinement.

4. CONCLUSION

In this work, we aim to enhance the feature representation for histopathology image classification, and propose a new deep neural network based on attention mechanism. We show that convolutional feature learning could be improved by two different attention modules, thereby leading to better classification performance in practice. The attention modules are lightweight and enhance the feature discriminability with small extra overhead. We validate the proposed method on the publicly available BreakHis dataset. Experimental results demonstrate the effectiveness of our method, particularly the attention modules, in histopathology image classification task.

5. REFERENCES

[1] P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Trans. Med. Imaging*, vol. 32, no. 12, pp. 2169–2178, 2013.

[2] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Engg.*, vol. 63, no. 7, pp. 1455–1462, 2016.

[3] V. Gupta and A. Bhavsar, "Breast cancer histopathological image classification: Is magnification important?," in *CVPR Workshop*, 2017, pp. 769–776.

[4] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *IJCNN*, 2016, pp. 2560–2567.

[5] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *ICPR*, 2016, pp. 2440–2445.

[6] K. Das, S. Conjeti, A. G. Roy, J. Chatterjee, and D. Sheet, "Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification," in *ISBI*, 2018, pp. 578–581.

[7] Y. Song, J. J. Zou, H. Chang, and W. Cai, "Adapting fisher vectors for histopathology image classification," in *ISBI*, 2017, pp. 600–603.

[8] Y. Song, Q. Li, H. Huang, D. Feng, M. Chen, and W. Cai, "Low dimensional representation of fisher vectors for microscopy image classification," *IEEE Trans. Med. Imaging*, vol. 36, no. 8, pp. 1636–1649, 2017.

[9] Y. Song, H. Chang, Y. Gao, S. Liu, D. Zhang, J. Yao, W. Chrzanowski, and W. Cai, "Feature learning with component selective encoding for histopathology image classification," in *ISBI*, 2018, pp. 257–260.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 6000–6010.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.