SELECTIVE SCALE CASCADE ATTENTION NETWORK FOR BREAST CANCER HISTOPATHOLOGY IMAGE CLASSIFICATION

Bowen Xu, Wenqiang Zhang

Academy for Engineering and Technology, Fudan University, Shanghai, China

ABSTRACT

Convolutional Neural Networks (CNNs) approaches are widely applied to histopathological image analysis due to the breakthrough performance achieved. However, it remains challenging because complex backgrounds obscure the most discriminative region. In this paper, we propose selective scale cascade attention network (SSCA) to learning discriminative features for breast histopathological image classification. Based on the backbone, SSCA mainly consists of three modules: 1) cross scale attention module leverage deeper level features to generate an attention map that locates the discriminative part in high-resolution features and conduct multi-scale classification. 2) cascade attention modules gradually identify fine-grained cues and increase their weight through a cascade of attention. 3) selective scale fusion module dynamically adjusts the weight of each scale feature, i.e., the select scale depends on the characteristics of the cancer subtypes. Extensive experiments show that our proposed consistently outperforms the existing state-of-the-art methods on the public BreakHis dataset.

Index Terms— Histopathological Image Classification, Multi-Scale, Cascade Network, Attention Network

1. INTRODUCTION

Breast cancer is one of the highest mortality rate cancers in women[1]. Identifying the subtype of cancer is crucial to selecting the appropriate treatment, and identifying the subtype of benign lesions can help assess a patient's future cancer risk in the pathological examination. However, it is highly time-consuming, depends on the skill and experience of the pathologist. Therefore, there is a significant demand to develop computer-aided diagnosis (CADx) to automatically detect and categorize the pathology.

In recent years, Convolutional Neural Networks (CNNs) approaches have been widely applied to the automatic breast histopathological images classification. Spanhol et al. [3] used the modified AlexNet to breast histopathological image classification and achieved significantly better results than previous work, which employed various hand-crafted features[2, 4, 5]. MuDeRN[6] propose a framework that classifying images either as benign or cancer and then categorizing cancer and benign cases into subtypes each. FE-BkCapsNet[7] propose a deep feature fusion and enhanced routing framework, combining CNN that usually highlights semantics and capsule network that focus on detailed position information. DSoPN[8] present a robust global covariance pooling module based on matrix power normalization to explore second-order statistics of deep features. To leverage multi-scale features, MCUa[9] extract features from several patches in each scale and exploit dynamic model ensemble.

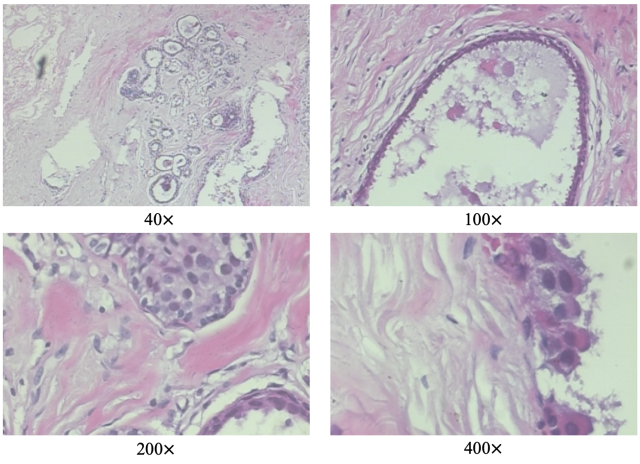

Fig. 1. Examples of breast histopathological images at different magnification factors from BreakHis dataset[2]

Although they achieved impressive performance, there are two challenging perspectives in the more fine-grained cancer subtype classification. First, the most discriminative region are obscure by complex backgrounds as shown in Fig.1. Second, there are significant intra-class fluctuation and inter-class similarities in the histopathological images. Recently, attention is widely used in various fields as making the network focuses on discriminative regions and less on distractors. SagaNet[10] propose a novel gated network to constrain the model pay attention to the valid areas in the pathological

This work supported by National Key RD Program of China (No. 2020AAA0108300, 2019YFC1711800), National Natural Science Foundation of China (No.62072112), Fudan University-CIOMP Joint Fund (No. FC2019-005)

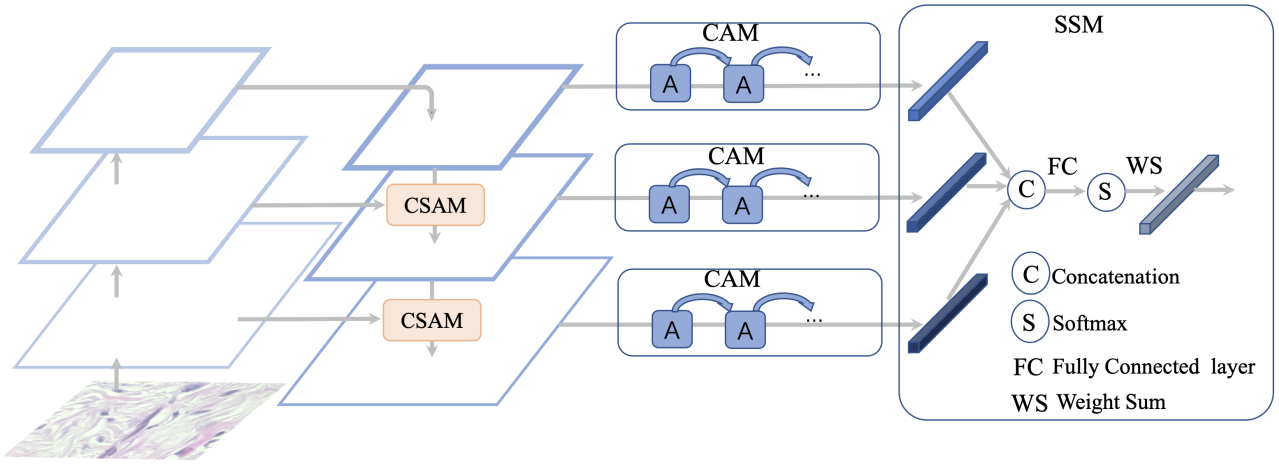

Fig. 2. The detailed of the proposed Selective Scale Cascade Attention Network. The whole pipeline can be divided into four steps, backbone network extraction of features, Cross Scale Attention Module (CSAM) aided feature pyramid generation, Cascade Attention Module (CAM) mining discriminative feature at each scale, and Selective Scale Module (SSM) fuses multi-scale features to final classify. A denotes a single attention block

image with a masking mechanism. GuSA[11] adapt addition region level segmentation and explicitly guides the focus of the network on diagnostically relevant regions.

In this paper, we aim to improve discriminative feature further from two aspects: 1) pathologists constantly adjust image resolution in the diagnosis of cancer subtypes. The assumption is that the importance of the magnifications depends on the image characteristics such as cancer subtypes[12]. 2) subtle differences in local structures need to be exploited to distinguish intermediate types. A single attention module generates a noisy attention map still because of heterogeneity in the histopathological image. Specifically, we propose Selective Scale Cascade Attention Network (SSCA) to exploit the above two aspects. First, we propose Cross Scale Attention Module(CSAM) for feature fusion and plug it into FPN[13], where the complementary information is aggregated to realize the pyramid feature. Second, Cascade Attention Module(CAM) is utilized in each scale to constantly improve the quality of the attention map to capture the most discriminative regions. Finally, we propose Selective Scale Module(SSM) for multi-scale fusion, which consists of Select and Fuse. The Select operator aggregates multiple-scale features to obtain a global representation and learn selection weights, and the Fuse operator aggregates the feature of different scales according to the selection weights.

2. METHODS

In this section, we describe the proposed Selective Scale Cascade Attention Network (SSCA). The overall pipeline of the proposed network is shown in Fig.2. We adapt ResNet-50 as the backbone and send $[F_3, F_4, F_5]$ to the next stage. Cross

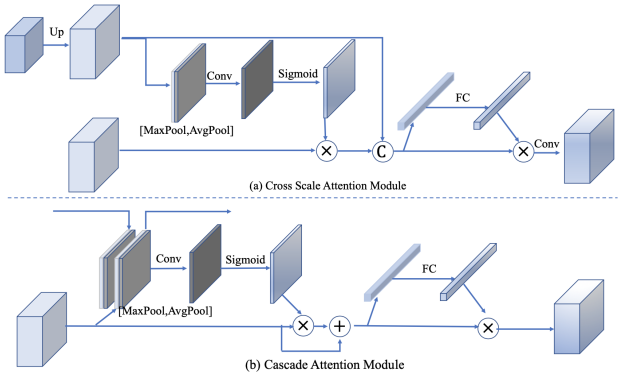

Fig. 3. The details of Cross Scale Attention Module and Cascade Attention Module.

Scale Attention Module(CSAM) empowers FPN to extract discriminative regions feature at multiple scales. Then Cascade Attention Module(CAM) is utilized to continually mining scale-aware local subtle differences. Finally, Selective Scale Module(SSM) dynamically fuses different scales according to subtypes like pathologists.

2.1. Cross Scale Attention Module

Multi-scale information is helpful in diagnosing cancer subtypes[12]. To realize the multi-scale processing, Feature Pyramid Network (FPN) [13] is widely adopted in existing frameworks. FPN leverages the inherent feature hierarchy fusing adjacent features through lateral connections in a top-down pathway to construct feature pyramid. However, high resolution tends to introduce more irrelevant areas, and such simple designs inhibit aggregate more discriminative fea-

tures. Thus, we propose CSAM to deal with this issue, which provides feature pyramid that has stronger semantics at all scales.

Figure 3 (a) shows that the CSAM can be divided into spatial attention and channel attention inspired by CBAM[14]. We first leverage high-level feature guides low-level feature to avoid introducing noise from low-level features. Specifically, we upsample high-level feature and apply average-pooling and max-pooling operations to highlight informative regions. Then, we concatenate them as spatial feature F^s and employ convolution layer to generate a spatial attention map $M_s \in R^{H\times W}$. The 2D spatial attention map guides low-level feature focuses on 'where' are informative regions. In short, the spatial attention is computed as:

$$F_i = \sigma(Conv^{7\times7}([AvgPool(F_{i+1}), MaxPool(F_{i+1})]))F_i \quad (1)$$

where σ denotes the sigmoid function. Afterwards, we concatenate $[F_i, F_{i+1}]$ as F and explore inter-channel relationship by channel attention. We conduct both average-pooling and max-pooling to get a global information embedding and employ a bottleneck with two fully-connected (FC) layers to produce channel attention $M_c \in R^C$. Then, we reweight each channel and apply a convolution layer to reduce channel to origin. The final output of the module is computed as:

$$F_i = Conv^{3\times3}(\sigma(W_1(W_0(F^c_{avg} + F^c_{max}))F) \quad (2)$$

where σ denotes the sigmoid function, $\mathbf{W_0} \in R^{C/r\times C}$ and $\mathbf{W_1} \in R^{C\times C/r}$denotes the FC weights.

2.2. Cascade Attention Module

To address the challenge of learning high quality attention map from a single attention module, we apply cascade mechanism to continually refine the attention map and aggregate discriminative information, inspired by the works of cascade object detection[15]. The Cascade Attention Module is composed of multiple single attention modules. Each attention module is consists of spatial attention and channel attention. Differently, we connect the previous spatial feature F^s_{i-1} to generate a spatial attention map, and residual is added to avoid value degrading. As shown in Figure 3 (b). The spatial attention can be described as:

$$M_i = \sigma(Conv^{7\times7}([F^s_{i-1}, AvgPool(F_i), MaxPool(F_i)])) \quad (3)$$

This connection ensures the current spatial attention map learns from both extracted features and previous attention information. Then, we recalibrate channel like previous CSAM by channel attention.

2.3. Selective Scale module

As mentioned previously, Pathologists scale images to diagnose. Thus, we propose selective scale module to dynamically fuse different scales' features according to the learned scale weights. Specifically, we implement the SSM via two operators, Select and Fuse, as illustrated in Fig. 1.

Select: Given the input feature map $F_i \in R^{C\times H\times W}$ at each scale, global average pooling is utilized for feature compression. Then, all compressed features are concatenated to obtain global feature representation F_g. We conduct a FC layer, with Batch Normalization is adopted as feature transformation and another FC layer to project dimension. Finally, a softmax operator is applied to generate scale weights. The select operator is formulated as: $Ws = Softmax(W_1(\alpha(W_1(F_g))))$ where α represent the batch normalization. $W_1 \in^{C\times(3\times C)}$, $W_2 \in^{3\times C}$ are the parameters of FC layer.

Fuse: With the learned sacle weight, we make an element-wise summation to obtain the final descriptor F_f as: $F_f = \sum_{i=0}^{S} Ws_i F_i$.

Table 1. Histopathological image distribution of BreaKHis divided by magnification and class

Class	Subclass	Magnification factors				Total
		40x	100x	200x	400x	
B	A	114	113	111	106	444
	F	253	260	264	237	1014
	TA	109	121	108	115	453
	PT	149	150	140	130	569
M	DC	864	903	896	788	3451
	LC	156	170	163	137	626
	MC	205	222	196	169	792
	PC	145	142	135	138	560
Total		1995	2081	2013	1820	7909

3. EXPERIMENT

3.1. Dataset and Evaluation Metrics

We adopt a commonly used breast cancer histopathological image dataset, i.e. BreaKHis dataset[2]. According to the different magnification factors, each image can be classified into four groups of 40x, 100x, 200x and 400x, respectively. Table 1 shows the image distributions of eight classes of breast cancer which include four distinct histological types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenona (TA); and four malignant tumors (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC). The original dataset is randomly divided into training set and testing set for each magnification at a ratio of 7: 3 follow previous work.

Image-level recognition rate(IRR) and patient-level recognition rate(PRR) are employed to evaluation performance. The IRR can be calculated as $IRR = \frac{N_{rec}}{N_{all}}$, where N_{all} represents the number of images and N_{rec} represents the number of images correctly classified. Meanwhile, the score

Table 2. Compare results with state-of-the-art on both image level and patient level

Methods	Years	Image Level (%)				Patient Level (%)			
		40x	100x	200x	400x	40x	100x	200x	400x
AlexNet variant[3]	2016	85.60	83.50	82.70	80.70	90.00	88.40	84.60	86.10
Inception V3[16]	2018	90.20	85.60	86.10	82.50	91.50	85.10	86.80	82.90
SE-ResNet variant[17]	2019	94.43	94.45	92.27	91.15	-	-	-	-
Independent framework[18]	2019	-	-	-	-	96.81	95.26	93.78	90.76
VGGNET16-SVM[19]	2020	-	-	-	-	94.11	95.12	97.01	93.40
DSoPN[8]	2020	96.00	96.16	**98.01**	95.97	95.01	96.84	**97.92**	96.28
FE-BkCapsNet[7]	2021	92.71	94.52	94.03	93.54	-	-	-	-
SagaNet[10]	2021	-	-	-	-	96.2	96.0	94.4	92.7
SSCA	-	**96.93**	**97.32**	95.31	**96.24**	**97.24**	**97.02**	96.42	**96.64**

Table 3. Ablation study of cascade attention on 40x

model	Image Level	Patient Level
baseline	93.42	93.85
+CAN (1)	94.16	94.62
+CAN (2)	94.66	95.35
+CAN (3)	95.16	96.15
+CAN (4)	95.43	96.74

Table 4. Ablation study of generation and fusion of multi-scale features on 40x

Generation	Fusion	Image Level	Patient Level
FPN	Concat	95.54	96.85
CSAM	Concat	96.32	97.13
FPN	SSM	96.53	97.06
CSAM	SSM	**96.93**	**97.24**

on the patient level P_s computed as $P_s = \frac{N_{rec}}{N_p}$, where N_p is the number of images of patient P. Let N be the number of patients, the PRR can be calculated by $PRR = \frac{P_s}{N}$.

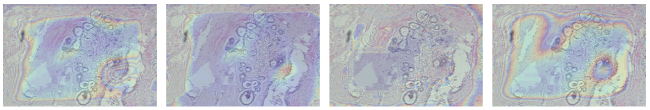

Fig. 4. Attention map in CAM.

3.2. Implementation Details

We ramdom resize and crop images into the resolution of $224 \times 224 \times 3$. For data augmentation, we deploy random flip, rotation, affine, color jitter in the training dataset. Stochastic gradient descent (SGD) is used to train model with a mini-batch size of 32, momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is 0.05. Warm-up and CosineAnneal decay strategies are used to adjust the learning rate.

3.3. Comparison with State-of-the-art

We compare our SSCA with nine state-of-the-art CNNs based model as shown in Table 2. The results illustrate that, SSCA reaches the best performance in 40x, 100x and 400x among all competing models (0.93% / 0.36% / 0.3% higher than the second best) in IRR and competitive performance in 200X. Especially in 40x which have the maximum field of view but also have a greater need to find discriminative areas.

3.4. Ablation study

We use 40x to conduct the ablation studies to validate the effectiveness of each component in our model. As illustrate in Table 3, the results show cascading multiple attention modules can significantly improve performance. We draw the attention maps in CAM(4) as show in Fig.4. We find they are concentrate gradually on the marginal areas which provides discriminative information. Beside, we validate multi-scale feature generation and fusion methods as illustrate in Table 4. For different generation methods, compare with CSAM, directly apply FPN bring slight improvement, because the high resolution features have many unrelated features. For different fusion methods, we compare SSM with directly concat. Better results can be obtained with SSM, especially when CSAM is used together.

4. CONCLUSION

In this paper, we exploit subtle discriminative regions feature extraction through cascade attention and efficient feature fusion by dynamically selecting weights of different scales for breast histology image classification. Specifically, we propose selective scale cascade attention network, which consists of three modules: 1) Cross scale attention module that promotes the feature pyramid to has stronger semantics at all scales. 2) Cascade attention modules is applied to aggregate discriminative feature at each scale continually. 3) Selective scale fusion module which dynamically fuses multi-scale features.

5. REFERENCES

[1] Freddie, Bray, Jacques, Ferlay, Isabelle, Soerjomataram, Rebecca, L, Siegel, and Lindsey, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries.," *CA: a cancer journal for clinicians*, 2018.

[2] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, 2016.

[3] Fabio Alexandre Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 International Joint Conference on Neural Networks*, 2016, pp. 2560–2567.

[4] Carlos Andrés Peña-Reyes and Moshe Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," *Artif. Intell. Medicine*, 1999.

[5] Mehmet Fatih Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, 2009.

[6] Ziba Gandomkar, Patrick C. Brennan, and Claudia Mello-Thoms, "Mudern: Multi-category classification of breast histopathological image using deep residual networks," *Artif. Intell. Medicine*, vol. 88, pp. 14–24, 2018.

[7] Pin Wang, Jiaxin Wang, Yongming Li, Pufei Li, Linyu Li, and Mingfeng Jiang, "Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing," *Biomed. Signal Process. Control.*, vol. 65, pp. 102341, 2021.

[8] Jiasen Li, Jianxin Zhang, Qiule Sun, Hengbo Zhang, Jing Dong, Chao Che, and Qiang Zhang, "Breast cancer histopathological image classification based on deep second-order pooling network," in *2020 International Joint Conference on Neural Networks*, 2020, pp. 1–7.

[9] Zakaria Senousy, Mohammed Abdelsamea, Mohamed Medhat Gaber, Moloud Abdar, Rajendra U Acharya, Abbas Khosravi, and Saeid Nahavandi, "Mcua: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2021.

[10] Yuhan Liu and Shiliang Sun, "Saganet: A small sample gated network for pediatric cancer diagnosis," in *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. 2021, vol. 139, pp. 6947–6956, PMLR.

[11] Heechan Yang, Ji-Ye Kim, Hyongsuk Kim, and Shyam Prasad Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE Trans. Medical Imaging*, vol. 39, no. 5, pp. 1306–1315, 2020.

[12] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise, "Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12597–12606.

[13] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944.

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: convolutional block attention module," in *ECCV 2018*, 2018, vol. 11211, pp. 3–19.

[15] Zhaowei Cai and Nuno Vasconcelos, "Cascade R-CNN: delving into high quality object detection," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*.

[16] Yassir Benhammou, Siham Tabik, Boujemâa Achchab, and Francisco Herrera, "A first study exploring the performance of the state-of-the art CNN model in the problem of breast cancer," in *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications 2018*, 2018, pp. 47:1–47:6.

[17] Yun Jiang, Li Chen, Hai Zhang, and Xiao Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module," *PLoS ONE*, vol. 14, 03 2019.

[18] Vibha Gupta and Arnav Bhavsar, "Partially-independent framework for breast cancer histopathological image classification," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1123–1130.

[19] Abhinav Kumar, Sanjay Kumar Singh, Sonal Saxena, K. Lakshmanan, Arun Kumar Sangaiah, Himanshu Chauhan, Sameer Shrivastava, and Raj Kumar Singh, "Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer," *Inf. Sci.*, vol. 508, pp. 405–421, 2020.