2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE 2021)

2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) | 978-1-6654-1540-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBAIE52039.2021.9389824

Whole Slide Images Based Cervical Cancer Classification Using Self-supervised Learning and Multiple Instance Learning

Tingzhen Li*, Ming Feng, Yin Wang
Tongji University
Shanghai, China
1930791@tongji.edu.cn

Kele Xu
National University of Defense Technology
Changsha, China

***Abstract*—Currently, most whole slide images classification models rely on manual pixel-level annotations, which requires specific domain experts to annotate that is delicate and time-consuming. To overcome this problem, we propose to combine self-supervised learning with multiple instance learning to deal with large WSIs datasets only with the reported diagnoses as labels. In WSIs classification task, it's a key challenge to learn good image representation, where self-supervised learning has held tremendous potential. In our study, we propose to use self-supervised learning network Bootstrap Your Own Latent as the pre-trained network, which can be trained using unlabeled data and learn the deep domain-specific features. We evaluated our proposed framework at scale on a uterine cervical dataset of 3,063 whole slide images(720GB). Our results have shown that the combination of self-supervised learning model and multiple instance learning model can match and exceed the performance of former methods.**

Keywords-cervical cancer; whole slide images; self-supervised learning; multiple instance learning; classification;

I. INTRODUCTION

Cervical cancer is one of the most common malignant cancer among women in the world wide, which is infected by a virus known as HPV [1]. According to the report of World Health Organization, about 570,000 women were infected with cervical cancer in the world and an estimated 311,000 women died from the virus HPV in 2018. To accurately detect cervical cancer, pathologists always need to carefully examine the whole slide images(WSIs) [2], which is digitized biopsy containing all the information required to diagnose lesions as malignant or benign. Normally, WSIs are tremendously large varying from megapixel to gigapixel. It's extremely laborious and time-consuming to carefully examine the lesions of cervical tissue and distinguish lots of mimics. Besides, the examine result largely depends on the experience of the pathologist.

Computer aided diagnostics in digital pathology has made great progress recent years, inspired by the appearance of deep Convolutional Neural Networks(CNNs) and the massive scale of annotated datasets such as ImageNet [3]. Due to the tremendous size of WSIs, methods based on CNN always have to train a feature extractor by a large amount of expertise annotations [4], which output a reasonable number of features containing critical information of WSIs. Now there are two main successful methods applying CNNs into WSIs classification: 1) training the CNN from scratch directly [5], [6] 2) transferring deep features from the CNN models pre-trained on medical images or other domain images and fine-tuning on target images [7], [8].

However, training the CNN from scratch requires large expertise annotated datasets and costs lots of time for training, Gao et al [9] proved that a pre-trained CNN model is a better choice for WSIs classification tasks, for which can use common low-level features. As for pre-trained models, most of the previous papers focus on CNNs pre-trained on ImageNet and adjusting a little on target images [10]-[12], but the extracted features aren't domain-specific. What's more, it's hard to get the fine-grained pixel-level annotations of WSIs in practice, which have to be annotated by domain-specific experienced pathologist.

At present, self-supervised learning has performed as well as and even better than the current state of the art networks on transfer benchmarks [13]. Self-supervision allows networks to be trained using just little or even no labeled data and can learn representations from data effectively by pre-designed task [14]. During training, self-supervised learning evaluates the network by proposing various tasks, which are hard to solve without understanding the images semantically such as filling in image holes [15] and solving jigsaw puzzles [16].

Multiple instance learning also has shown giant potential in WSIs classification, which deals with a set of labeled bag containing multiple instances. In [17], they use multiple instance learning to train a feature extractor by the slide-level labels. In [2], they combined multiple instance learning and attention based methods to get more interpretable results.

In this paper, we apply self-supervised learning and multiple instance learning into WSIs classification task using only image-level labels. We pre-train a self-supervised learning model to extract features for downstream tasks and use MIL for classification. To the best of our knowledge, utilizing self-supervised learning and MIL into the task of cervical cancer WSIs classification has not been shown before. Our main contribution is experimenting the three most popular self-supervised models to extract features for clustering and combining self-supervised learning with Multiple Instance Learning to classify WSIs using only coarse-grained image-level labels.

This paper is structured as follows. In section II, we thoroughly introduce our proposed model. After that, we show

978-1-6654-1540-8/21/$31.00 ©2021 IEEE

our two experiments in section III, including the introduction of datasets and performance of our proposed models. Finally, we draw conclusions and propose some promising directions in section IV.

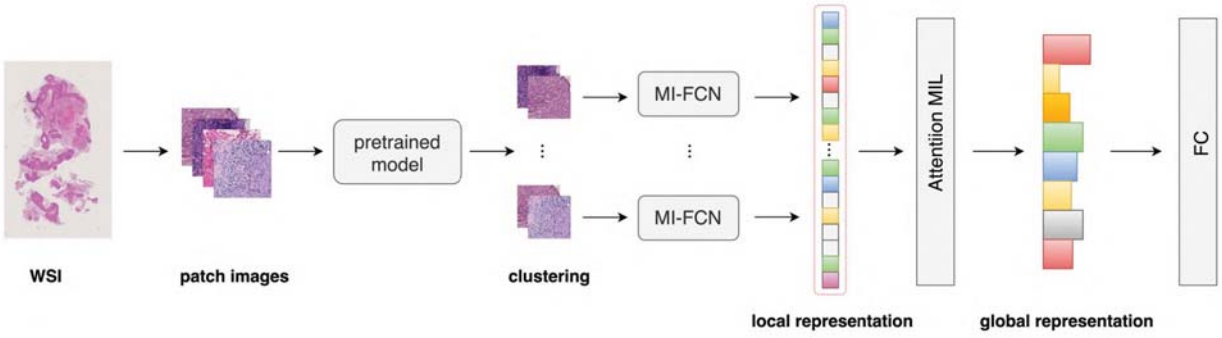

Figure 1 Overviewed model

II. Model

One whole slide image is usually at gigabyte scale and has about 50% background area, which is almost impossible directly processed by CNN. In preprocessing, we extract patches from WSIs ignoring the background region, where per whole slide image samples thousands of patches. The sampled patches don't necessary have the same labels as the original WSI. For instance, a WSI labeled as malignant could have sampled patches of benign and malignant classes. At least some sampled patches will represent the most severe class. To deal with this problem, we use Multiple Instance Learning(MIL) [18] that's a paradigm of weakly supervised learning, considering the instances are grouped in labeled bags without requiring each instance to have individual label.

In MIL term, let's set a WSI X as a bag of patch image instances, $X=\{x_1, x_2, \ldots, x_c\}$, every bag could have different number C of patch image instances. In our model, we don't directly use individual patch image as the instance of bag, but use the representation cluster.

Our proposed model is shown in Figure 1. As mentioned in [6], it's extremely time-consuming to directly train patch-based CNNs for multiple instance learning from patch, so we propose to extract features using the self-supervised learning based pre-trained model, then use the extracted features for clustering. We use K-means clustering [19] to cluster the patch images by the learned features. Every cluster is treated as the instance in multiple instance learning, which largely reduces the scale of instances comparing to directly using the patch images as instances. Then we deal the clusters with the Multiple Instance Fully Convolutional Network(MI-FCN) [20] and output the local representations in parallel. For better interpretability, we use attention-based MIL network to visual the critical parts of the local representations for detecting lesions, which finds the most useful and relevant areas for classification task and remarkably reduces the number of model parameters. After that, we get the global representation for sub-type of cervical cancer classification. Comparing with local representations that are extracted from the patch images, global representations are for the WSIs, which have more spatial information than local features. Finally, we get the prediction of the cancer type by using the fully convolutional layer.

A. Pre-trained model

Feature extraction is an extremely challenging task in WSIs classification for its immensely large size. Traditional methods are usually using hand-crafted features and machine learning based methods like SVM and K Nearest Neighbor which cost a quantity of time and have no adaptability for new datasets [21]. Inspired by fast evolving deep learning, the WSIs classification task also try to find help from CNNs based methods, which achieved better performance in practice, but it's extremely protracted and difficult to gain large quantity of pixel-level annotated data for training a multiple layers deep CNN from scratch directly.

We evaluate three most popular self-supervised learning models in experiment. SimCLR [22] repulses the views of different images and attract the same images' two views. MoCo [23] uses a slow-moving average network to attract negative pairs' view that extracted from the memory bank. While Bootstrap Your Own Latent(BYOL) [13] utilizes a moving average network to produce prediction targets as a means of maintaining consistent representations. As shown in our experiment result, BYOL performs better than the other two models, so we choose to use BYOL as pre-trained model.

BYOL also reaches 79.6% top-1 classification accuracy on ImageNet. BYOL has two neural networks interacting and learning from each other, named as online and target networks. BYOL inputs the same image into the two neural networks and apply to different augmentations, then it trains the online network to predict the target network representation. Meanwhile, it updates the target network with an average of the online network. Comparing with other pre-trained models, our proposed BYOL based pre-trained model has better generation ability.

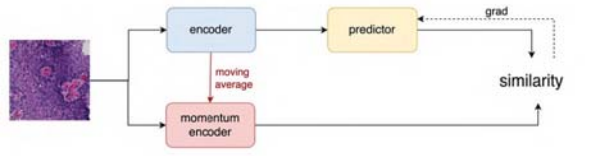

Figure 2 Pre-trained model

Our pre-trained model shown in Figure 2 has two neural networks to learn, referenced as encoder and momentum encoder. The encoder and the momentum encoder get an image as input, then separately outputs the representation and the target representation, after that the predictor outputs the prediction. We compute the similarity of the prediction and target representation to assess the performance of encoder and pass back the gradient to predictor. In the end of training, we only save the encoder that can output the representation of images for downstream tasks [24], [25].

B. MI-FCN

We use the Multiple Instance Fully Convolutional Networks (MI-FCN) [2] in Figure 3 to deal with the clustered representative patch images. The MI-FCN enables multiple sub-networks to run together in the whole deep learning network by sharing weights. The MI-FCN aims to learn the informative local representation for each clustered bag.

The input of MI-FCN is a bag containing m_i patch images, which normally as $1 \times m_i \times \mathrm{d}$ (d is the feature channel or dimension). Then we use two pairs of 1×1 conv layer and ReLu layer, which have been proposed as an effective feature extractor [26]. As the number of patch images varies in each bag, we use the Fully Convolutional Network (FCN) without including fully connected layer to make the network more flexible and enable it to handle any spatial resolution. At the end of MI-FCN, we use the global pooling layer such as max pooling and average pooling. Then the network generates local representations for the patch images.

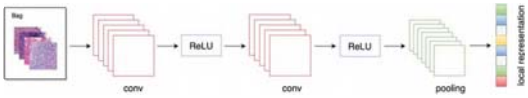

Figure 3 MI-FCN

III. EXPERIMENT

We use the dataset of public competition TissueNet [27], which contains thousands of WSIs of uterine cervical tissue from medical centers. These WSIs are classified into four classes: benign (class 0), low malignant potential (class 1), high malignant potential (class 2) and invasive cancer (class 3). The dataset includes three formats of training set images: native whole slide image formats (720 GB), pyramidal TIFs (928 GB), and downsampled JPEGs (752 MB). The JPEGs contains 5,927 patch images that sampled from 32x resolution. In our experiment, we use the native WSI formats that includes 1,015 labeled WSIs and 2.048 unlabeled WSIs for classification, and we use the 5,927 downsampled JPEGs for pre-training self-supervised learning model.

We use Top-1 accuracy and weight metrics from Competition TissueNet [27] to evaluate the performance of models, which is an evaluate metrics proposed by expert pathologists. Each prediction gets the score of 1 minus the error that is defined by the Table I set by a consensus from the scientific council. The final score is the average of all predictions.

TABLE I ERROR TABLE

Actual	Prediction			
	Class 0	*Class 1*	*Class 2*	*Class3*
Class 0	0.0	0.1	0.7	1.0
Class 1	0.1	0.0	0.3	0.7
Class 2	0.7	0.3	0.0	0.3
Class 3	1.0	0.7	0.3	0.0

TABLE II PERFORMANCE OF DIFFERENT METHODS ON PATCH IMAGES

Methods	Metrics		
	Arch.	*Accuracy*	*Weight Metrics*
Supervised	Resnet18	0.739	0.849
SimCLR	Resnet18	0.741	0.852
MoCo V2	Resnet18	0.746	0.860
BYOL	**Resnet18**	**0.747**	**0.860**

TABLE III RESULTS UNDER DIFFERENT NETWORKS AND CLUSTER NUMBERS

Cluster Number	*Network*	*Accuracy*	*Weight Metrics*
4	Resnet18+MIL	0.610	0.905
	BYOL+MIL	0.657	0.910
6	Resnet18+MIL	0.618	0.889
	BYOL+MIL	0.661	**0.922**
8	Resnet18+MIL	0.618	0.894
	BYOL+MIL	**0.681**	0.900
10	Resnet18+MIL	0.610	0.883
	BYOL+MIL	0.648	0.890

Data augmentation is critical to get more robust results for self-supervised learning. A common problem of WSIs classification is the color variation, due to the inconsistent histology stains or different scanners [28], so we use color normalization to weaken the influence of color by applying a uniformly random offset to the brightness, contrast and saturation of the images. Beyond this, we also use random cropping that picks a random scaled crop with resizing and Gaussian filter to improve the robustness of network.

In experiment, we compare the three most popular self-supervised model SimCLR, MoCo V2 and BYOL with supervised model, we evaluate these models on the JPEGs. The results are shown in Table II. According to the result, self-supervised models all are performing better than supervised method with 74.7% accuracy and 0.860 weight metrics. Based on it, we use BYOL as our pre-trained model.

To assess the effect of our proposed pre-trained model, we compare it with Resnet18 model pre-trained on ImageNet that has been evaluated to be effective for delicate whole slide images [29]. In experiment, we split the training dataset into 80% training and 20% testing. Table III shows the result. We can see in all cases our proposed BYOL based pre-trained model

performs much better than Resnet18 based pre-trained model at Top-1 accuracy and weight metrics.

To see the influence of representation kinds, we compared different cluster number varying from 4 to 10. According to the result, 8 cluster number gets the highest 68.1 % accuracy and 6 cluster number gets the highest 0.922 weight metrics. At first, the accuracy increases as more cluster number, but when the cluster number increases to 10, the accuracy drops. And the effects of cluster number to our proposed BYOL pre-trained model is more than Resnet18 pre-trained model.

IV. Conclusion

In this paper, we combine the self-supervised learning with MIL for the WSIs classification using only the coarse- grained image-level labels. To the best of our knowledge, it's the first time that a self-supervised learning based pre-trained model is implemented for cervical cancer whole slide images classification. Our experimental findings prove the self-supervision performs very well in WSIs classification. With future research, our proposed model has potential for other type of cancers and even other domain gigabytes images.

Acknowledgment

This work has been performed on data that were available during the 2020 Data Challenge of the French Society of Pathology and the Health Data Hub, with the support of Grand Defi for A.I in Health.

References

[1] Y. Jusman, S. C. Ng, and N. A. Abu Osman, "Intelligent screening systems for cervical cancer," The Scientific World Journal, vol. 2014, 2014.

[2] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," Medical Image Analysis, vol. 65, p. 101789, 2020.

[3] J.Deng, W.Dong, R.Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition.Ieee, 2009, pp. 248–255.

[4] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv:1606.05718, 2016.

[5] M. Mostavi, Y.-C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," BMC medical genomics, vol. 13, pp. 1–13, 2020.

[6] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Efficient multiple instance convolutional neural networks for gigapixel resolution image classification," arXiv preprint arXiv:1504.07947, p. 7, 2015.

[7] F. Idlahcen, M. M. Himmi, and A. Mahmoudi, "Cnn-based approach for cervical cancer classification in whole-slide histopathology images," arXiv preprint arXiv:2005.13924, 2020.

[8] T. Schlegl, J. Ofner, and G. Langs, "Unsupervised pre-training across image domains improves lung tissue classification," in International MICCAI Workshop on Medical Computer Vision. Springer, 2014, pp. 82–93.

[9] Z. Gao, L. Wang, L. Zhou, and J. Zhang, "Hep-2 cell image classification with deep convolutional neural networks," IEEE journal of biomedical and health informatics, vol. 21, no. 2, pp. 416–428, 2016.

[10] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado et al., "Detecting cancer metastases on gigapixel pathology images," arXiv preprint arXiv:1703.02442, 2017.

[11] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 en- coder pre-trained on imagenet for image segmentation," arXiv preprint arXiv:1801.05746, 2018.

[12] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multi- view mammogram analysis with pre-trained deep learning models," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 652–660.

[13] J.-B. Grill, F. Strub, F. Altche´, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar et al., "Bootstrap your own latent: A new approach to self-supervised learning," arXiv preprint arXiv:2006.07733, 2020.

[14] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2051–2060.

[15] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[16] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9359–9367.

[17] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva,K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," Nature medicine, vol. 25, no. 8, pp. 1301–1309, 2019.

[18] O. Maron and T. Lozano-Pe´rez, "A framework for multiple-instance learning," Advances in neural information processing systems, pp. 570–576, 1998.

[19] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," Pattern recognition, vol. 36, no. 2, pp. 451–461, 2003.

[20] J. Yao, X. Zhu, and J. Huang, "Deep multi-instance learning for survival prediction from whole slide images," in International Confer- ence on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 496–504.

[21] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," IEEE transactions on medical imaging, vol. 28, no. 7, pp. 1037–1050, 2009.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[24] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014, pp. 1626–1630.

[25] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," Remote Sensing, vol. 9, no. 5, p. 489, 2017.

[26] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3743–3752.

[27] Tissuenet: Detect lesions in cervical biopsies.[Online]. Available: https://www.drivendata.org/competitions/67/competition-cervical-biopsy/page/254/

[28] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, and N. Navab, "Structure-preserved color normalization for histological images," in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015, pp. 1012–1015.

[29] Z. Gandomkar, P. C. Brennan, and C. Mello-Thoms, "Mudern: Multi-category classification of breast histopathological image using deep residual networks," Artificial intelligence in medicine, vol. 88, pp. 14–24,2018