2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9433836

DEEP FISHER VECTOR CODING FOR WHOLE SLIDE IMAGE CLASSIFICATION

Amir Akbarnejad[†] *Nilanjan Ray*[†] *Gilbert Bigras*[⋆]

[⋆] Department of Laboratory Medicine and Pathology, University of Alberta
[†] Department of Computing Science, University of Alberta

ABSTRACT

Adopting machine learning methods for histological sections is a challenging task given the generated huge size of whole slide images (WSIs) especially using high power resolution. In this paper we propose a novel WSI classification method which efficiently predicts a WSI's label. The proposed method considers each WSI as a population of patches and computes a statistic by having some samples from the population. This statistic can be computed efficiently, and our test time on a WSI is about one tenth of that of the existing methods. Moreover, our pooling strategy on the WSI is more general than that of previous works. Further, the assumptions of our method are quite general, and therefore, it is applicable to any WSI classification task. The experiments show that the performance of our method is competitive in two different tasks, while, unlike some of the competing methods, it does not consider any prior clinical knowledge about the label to be predicted.

Index Terms— digital pathology, deep learning, whole slide image classification, Fisher vector distribution encoding

1. INTRODUCTION

Scanning glass slides and storing them as digital images have recently become prevalent in pathology clinical workflow [1]. Apart from its clinical benefits, it provides machine learning researchers with abundant data to develop models for different tasks such as diagnostic classification, prognosis and predictive response to therapy. Despite the recent success of machine learning, especially deep learning, these methods cannot be directly applied to digital pathology images, as these images are huge (e.g., 100k-by-100k pixels), and the information needed to predict the label might be located anywhere over a huge whole slide image (WSI). In literature, this problem is referred to as the *where problem* [2]. The *where problem* is further complicated when relevant tissue information is heterogeneously [3] scattered among non-diagnostic normal tissue over a WSI, or when the diagnostic tissue has intrinsic biological heterogeneity with label-relevant information segregated for instance in a subset of a tumor. On the other hand, pathology images are known to have a large amount of variability and artifacts coming from different sources: improper tissue fixation impacting the tissue morphology, improper paraffin block sectioning, variation in tissue staining and variation among image scanning devices. In literature, this issue is referred to as the *what problem* [2].

In this paper we propose a novel WSI classification pipeline based on Fisher vector distribution encoding [4]. As an embedding-based method, our approach tackles both the *what* and the *where* problems simultaneously. Moreover, our method achieves unprecedent efficacy for the following reasons: 1) predictions are based on multiple-instance learning statistics derived from a WSI and computed efficiently. 2) Model training and testing are expedited thanks to our recently developed package PyDmed [5]. By doing so, in each experiment we train our model on **$\sim$2.5 billion patches in $\sim$8 hours.** Moreover, our test **time is less than one minute per WSI of size 100k-by-100k pixels.** We ran our experiments on a desktop PC with HDD hard drive, one GPU (GeForce GTX TITAN X), Intel(R) Core(TM) i7-4790K CPU, and 32 gigabytes of RAM. Our implementation is available online [1]. Besides efficiency, our pipeline is end-to-end, and it does not have the known shortcomings of multi-stage pipelines [2]. Moreover, the proposed method is quite general, and it does not consider any prior clinical knowledge about the label to be predicted.

2. RELATED WORK

It is conventional to categorize WSI classification methods based on, e.g., their pooling strategies, their modeling assumptions, or the challenges that they address [2, 6]. However, here we adopt the terminology of multiple-instance learning (MIL) [7], which is a mature framework and encompasses most concepts that we are going to discuss in this paper. In view of MIL, it is conventional to think of each WSI as a "bag". Each instance can be a random patch extracted from the WSI, or a tuple containing the patch as well as its (i, j) location in the WSI. The so called *standard MIL assumption* and its variants have been successfully adapted for WSI classification [8, 9, 10, 11]. By making the standard assumption, one presumes that each instance in the bag has a latent

[1] `https://github.com/amirakbarnejad/code_submission_isbi2021`

label. A bag is positive if and only if some of its instances are positive. The benefit of the standard assumption is that patch-level labels become closely related to WSI-level labels. The standard assumption is a natural fit when the label can be directly assigned to patches (e.g. colon cancer phenotype) or when the label depends on the local information. However, it is not the case for all labels. To relax this assumption, inspired by embedding-based approach to MIL, it is conventional to extract some patches like $\{x_1, ..., x_n\}$ from a WSI and to feed them to a CNN (f) to get $\{f(x_1), ..., f(x_n)\}$ and then to somehow aggregate these vectors to get a vector W. One can think of W as a vector that encodes the WSI from which $\{x_1, ..., x_n\}$ are cropped. To aggregate the values, several pooling methods have been proposed: max pooling, global average pooling [12], and kmin+kmax pooling [13]. It has been shown that the choice of pooling strategy has a dramatic effect on the performance, for either natural images [14] or histopathology images [3]. The closest pooling strategy to ours is the approach of [3] where the authors encode a large image by computing different quantiles of $f(x)$ where x is a random patch. To encode a WSI, Tellez et al. [15] propose to train a generative model, e.g., a variational auto-encoder or BiGAN, on WSIs' patches. Afterwards, they split each WSI to a grid of patches and they feed each patch (i.e. each cell in the grid) to the encoder part of the trained generative model. The resulting volumetric map actually encodes the WSI, and can be fed to a follow up module like a CNN to make the final prediction. The main drawback of such an approach is that it exhaustively feeds every patch (i.e. every cell) in the grid to the encoder network. A typical WSI will contain hundreds of thousands of patches (i.e. cells) and this process comes at a huge computation cost.

3. PROPOSED METHOD

Let x be a random patch extracted from a WSI. Let $f(.)$ be a function that takes in x and produces a D-dimensional vector. The vector $f(x)$ is often referred to as a *descriptor*. We encode the WSI as the distribution $P\big(f(x)\big)$. To represent this distribution as a vector, we used Fisher vector distribution encoding [4]. This approach considers a set of fixed vectors $\{v_1, ..., v_m\}$ in the space of descriptors. Let $s_{ij} \in [0, 1]$ be the soft assignment of $f(x_i)$ to the fixed vector v_j. A set of descriptors $\{f(x_1), ..., f(x_n)\}$ are encoded as

$$\frac{1}{n}\sum_{i=1}^{N} FV\big(f(x_i)\ ;\ v_1, ..., v_m\big), \tag{1}$$

where $FV : \mathbb{R}^{D+m} \rightarrow \mathbb{R}^{2m}$ is a function defined as follows:

$$\begin{aligned} FV(.) = \Big[&\frac{s_{i1}}{c_1}\big(f(x_i) - v_1\big), ..., \frac{s_{im}}{c_m}\big(f(x_i) - v_m\big), \\ &\frac{s_{i1}}{\hat{c}_1}\big(f(x_i) - v_1\big)^2, ..., \frac{s_{im}}{\hat{c}_m}\big(f(x_i) - v_m\big)^2\Big]. \end{aligned} \tag{2}$$

In Eq.2 c_j and $\hat{c}_j$ are some constants. We encode a WSI to a vector as follows:

$$W \quad = \quad \mathop{\mathbb{E}}_{\substack{x \sim otsu \\ foreground}} \Big[FV\big(f(x)\ ;\ v_1, ..., v_m\big)\Big], \tag{3}$$

where $\mathbb{E}$ denotes mathematical expectation and x is a random patch extracted from the WSI's foreground obtained by Otsu's method [16]. After encoding a WSI to a vector by Eq.3, we feed this vector to a linear classifier to predict the WSI's label. During training, the expectation of Eq.3 is approximated by firstly extracting a few patches $x_1, ..., x_n$ from the WSI (using PyDmed[5]) and then averaging over them as in Eq.1. Accordingly, our pipeline is trained end-to-end.

We implemented the function $f(x)$ by a convolutional neural network. As mentioned above, given a patch x the function $f(x)$ may produce only one descriptor in $\mathbb{R}^D$. However, inspired by the method CBoF [17], we implemented the function $f(x)$ by a fully convolutional module that produces a volumetric map of shape $[D \times V \times V]$. This volumetric map indeed contains $V \times V$ descriptors in $\mathbb{R}^D$.

4. EXPERIMENTAL SETUP AND RESULTS

We evaluated our method on two tasks:

- Predicting Breast Cancer HER2 score from IHC HER2 WSIs preparations: we downloaded the dataset used in Warwick university HER2 contest [18]. This dataset contains 52 WSIs for training and 34 WSIs for testing. As we didn't have access to the labels for testing WSIs, we considered 60% of the training WSIs for training and the remaining 40% for testing.
- Classifying brain astrocytic (glial) tumor grade from H&E whole slide images: We downloaded H&E slides from 120 patients from TCGA [19] dataset. For each patient, we considered the H&E slide which was most recently scanned. The task was to classify astrocytomas in two groups: grade IV astrocytoma and other lower grades astrocytomas. We considered 50% of the cases for training and the remaining 50% for testing. Following the authors of [20], we used the labels reported in the supplementary material of [21].

For the fully convolutional module of Fig.1, we used a pre-trained Resnet50 (excluding the final pooling layer and the follow up fully-connected layers) followed by a convolutional layer that reduces the number of channels (i.e. the dimensionality of the descriptor space) to 10.

Finally we added a batch-normalization layer to this module. In all of our experiments we set the number of Fisher vector coding centers (i.e. the variable m in Eq.3) to 10. Moreover, with the notation of [4], we set the parameters of Fisher vector encoding as $\pi_k = 0.1$ and $\sigma_k = 0.1$. During training we set the batch size to 32, and we trained on each training set for 80000 iterations. We used a RMSprop optimizer with learning rate 0.00001. As a baseline, we replaced the Fisher

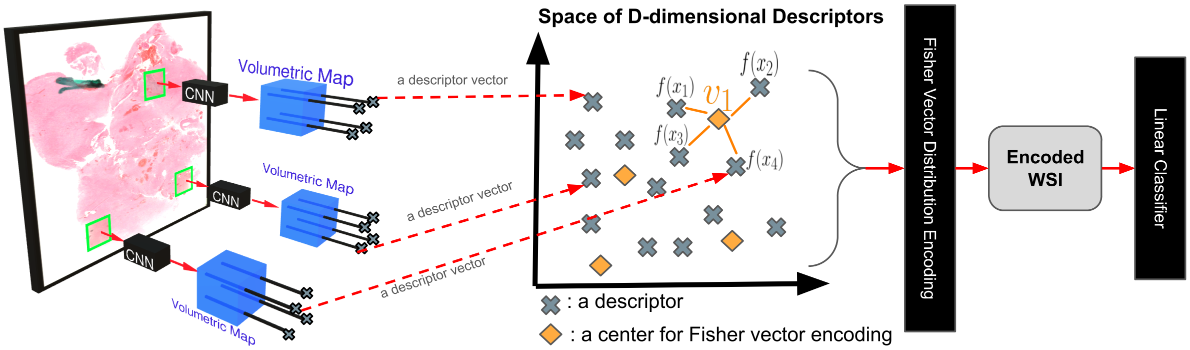

Fig. 1. The proposed pipeline.

vector encoding stage by an average pooling layer to see how the prediction performance changes. In Tabs. 1 and 2 this baseline is referred to as *Baseline Average Pooling*.

Approximating the expectation of Eq.3 by a few samples may result in a large noise in the approximate gradient, which in turn may cause the optimization to fail. One solution is to increase the batch size. However, doing so is impractical due to the limited memory of GPU. To tackle this issue, we used PyTorch's mechanism for accumulating gradients in a burst of backward passes and updating parameters at the end of each burst. By doing so, we effectively increased the batch size to 32 × 20. In the test phase, we approximate the expectation of Eq.3 by extracting 500 patches from each WSI. We implemented both the training and the testing phase by our recently developed package PyDmed [5] (Python Dataloader for MEDical imaging). PyDmed tailors the idea of PyTorch's dataloader to medical datasets. With PyDmed [5] our training time on more than 2.5 billion patches in about 8 hours. Moreover, our testing time is less than one minute per WSI. The reason behind this efficiency is to think of each WSI as a population of patches and to estimate the expectation of Eq.3 by some sampels from the population. The alternative way is to sweep over each WSI in the testing phase. In Tabs. 1 and 2 the latter approach is denoted by *Proposed Method (sweep)*. According to Tabs. 1 and 2, this approach dramatically increases the testing time although its prediction performance is almost the same.

We split each dataset to training/testing sets three times. Table 1 and 2 provide the average evaluation metric over these three runs. In Table 1 and 2 the numbers within parenthesis denote standard deviations. In the second column we report a performance measure used in Warwick HER2 contest [18]. According to the leader board of the contest [18], our method outperforms the competitors by a large margin: we obtained the highest combined score of 426.42 among 16 other teams (average = 355.83). For brain tumor grading, our prediction performance is comparable to the results reported in [20], although our dataset is much smaller (120 WSIs vs. 700 WSIs).

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted using human subject data made available in open access by [19] and upon registration by [18]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] Sanjay Mukhopadhyay and Michael D. Feldman et al., "Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study)," *The American journal of surgical pathology*, vol. 42, no. 1, pp. 39–52, Jan 2018, 28961557[pmid].

[2] Neofytos Dimitriou et al., "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine*, vol. 6, pp. 264, 2019.

[3] Heather D. Couture, J. S. Marron, and Charles M. Perou et al., "Multiple instance learning for heterogeneous images: Training a cnn for histopathology," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Cham, 2018, pp. 254–262, Springer International Publishing.

[4] K. S. Arun et al., "Enhanced bag of visual words representations for content based image retrieval: a comparative study," *Artificial Intelligence Review*, vol. 53, no. 3, pp. 1615–1653, Mar 2020.

[5] "Pydmed, python dataloader for medical imaging," `https://github.com/amirakbarnejad/PyDmed`, Accessed: 2020-09-22.

	her2 contest	accuracy	precision	recall	F-score	testing-time(s)
Proposed Method	426.42(38.31)	63.33(7.64)	0.75(0.02)	0.63(0.08)	0.61(0.08)	1439.2(21.6)
Proposed Method (sweep)	430(20.64)	62.50(0.08)	0.59(0.07)	0.62(0.08)	0.58(0.07)	12158(120.05)
Baseline Average Pooling	413.21(0.17)	57.14(0.03)	0.55(0.02)	0.57(0.03)	0.52(0.03)	1410(17.2)

Table 1. Performance of predicting breast cancer HER2 score. Rows (resp. columns) correspond to different methods (resp. performance measures).

	auc	accuracy	precision	recall	F-score	testing-time(s)
Proposed Method	0.92(1.01)	90.82(1.02)	0.88(0.02)	0.90(0.00)	0.89(0.01)	3065(10.53)
Proposed Method (sweep)	0.91(2.02)	91.2(0.89)	0.88(0.02)	0.89(0.72)	0.88(0.01)	26406(16.16)
Baseline Average Pooling	0.90(0.54)	90.38(0.32)	0.86(0.02)	0.89(0.00)	0.87(0.01)	3073(20.50)

Table 2. Performance of predicting brain astrocytic (glial) tumor grade. Rows (resp. columns) correspond to different tasks (resp. performance measures).

[6] Asmaa Ibrahim and Paul Gamble et al., “Artificial intelligence in digital breast pathology: Techniques and applications,” *The Breast*, vol. 49, pp. 267 – 273, 2020.

[7] James Foulds and Eibe Frank, “A review of multi-instance learning assumptions,” *The Knowledge Engineering Review*, vol. 25, 03 2010.

[8] Le Hou et al., “Patch-based convolutional neural network for whole slide tissue image classification,” *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016, 06 2016.

[9] Marc Combalia and Verónica Vilaplana, *Monte-Carlo Sampling Applied to Multiple Instance Learning for Histological Image Classification: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*, pp. 274–281, 09 2018.

[10] Ziqiang Li et al., “Da-refinenet:a dual input whole slide image segmentation algorithm based on attention,” *CoRR*, vol. abs/1907.06358, 2019.

[11] Hanbo Chen et al., “Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Cham, 2019, pp. 351–359, Springer International Publishing.

[12] Bolei Zhou et al., “Learning deep features for discriminative localization,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Thibaut Durand, Nicolas Thome, and Matthieu Cord, “Weldon: Weakly supervised learning of deep convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord, “Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, United States, July 2017, IEEE.

[15] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, “Neural image compression for gigapixel histopathology image analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[16] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[17] Nikolaos Passalis and Anastasios Tefas, “Bag-of-features pooling for deep convolutional neural networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[18] Talha Qaiser et al., “Her2 challenge contest: a detailed assessment of automated her2 scoring algorithms in whole slide images of breast cancer tissues,” *Histopathology*, vol. 72, no. 2, pp. 227–238, 2018.

[19] “The cancer genome atlas,” `portal.gdc.cancer.gov`, Accessed: 2020-09-22.

[20] Alexandre Momeni, Marc Thibault, and Olivier Gevaert, “Deep recurrent attention models for histopathological image analysis,” *bioRxiv*, 2018.

[21] Michele Ceccarelli, Floris P. Barthel, Tathiane M. Malta, Thais S. Sabedot, and Sofie R. Salama et al., “Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma,” *Cell*, vol. 164, no. 3, pp. 550–563, Jan 2016, 26824661[pmid].