WHOLE SLIDE IMAGE CLASSIFICATION VIA ITERATIVE PATCH LABELLING

Chaoyi Zhang[⋆], *Yang Song*[⋆], *Donghao Zhang*[⋆], *Sidong Liu*[†], *Mei Chen*[‡§], *Weidong Cai*[⋆]

[⋆] School of Information Technologies, University of Sydney, Australia
[†] Brain and Mind Centre, Sydney Medical School, University of Sydney, Australia
[‡] Electrical and Computer Engineering, State University of New York at Albany, United States
[§]Robotics Institute, Carnegie Mellon University, United States

ABSTRACT

Brain tumor can be a fatal disease in the world. With the aim of improving survival rates, many computerized algorithms have been proposed to assist the pathologists to make a diagnosis, using Whole Slide Pathology Images (WSI). Most methods focus on performing patch-level classification and aggregating the patch-level results to obtain the image classification. Since not all patches carry diagnostic information, it is thus important for our algorithm to recognize discriminative and non-discriminative patches. In this study, we propose an iterative patch labelling algorithm based on the Convolutional Neural Network (CNN), with a well-designed thresholding scheme, a training policy and a novel discriminative model architecture, to distinguish patches and use the discriminative ones to achieve WSI-classification. Our method is evaluated on the MICCAI 2015 Challenge Dataset, and shows a large improvement over the baseline approaches.

Index Terms— Iterative patch labelling, brain cancer, WSI, discriminative patches, classification

1. INTRODUCTION

In recent decades, brain tumors have become a serious health issue around the world [1]. In order to improve the survival rate of brain cancer, early and accurate diagnosis of brain cancer is essential. With the rapid development of machine learning and computer vision techniques, automated and computerized methods have emerged to assist human pathologists in the interpretation of pathology images and the diagnostic decision-making process. However, compared to other classification tasks, computational cost is an important factor to address in classifying high resolution pathology images, such as the gigapixel Whole Slide Pathology Images (WSI).

Most existing approaches perform the classification at patch-level then aggregate the patch classification outputs to obtain the image-level label [2, 3, 4, 5, 6, 7, 8]. Fig. 1 is an example of patch subsetting. The main drawback of these approaches is that the patches are assigned the same label as the image containing these patches since the datasets typically only provide image-level labels. This label propagation, however, is incorrect because only a small area of the image would exhibit tumor features and most of the patches would be normal tissue. Training a classifier by treating all the patches with the image label would thus become an ill-posed problem.

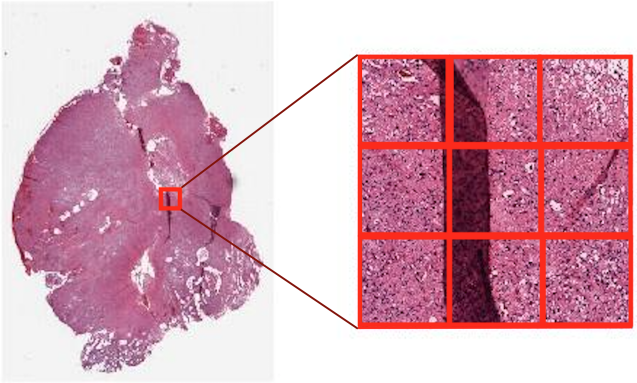


Fig. 1. Patches of size *500 by 500* extracted from a brain astrocytoma WSI of size *28500 by 19500*, at 40X resolution scale.

To address this issue, a coarse-to-fine approach has recently been proposed [9]. It generates a set of discriminative tiles containing the most representative morphological features of entire dataset by applying clustering techniques. This pipeline increased the processing speed greatly and achieved highly accurate classification. However, this method requires hand-crafted features, which might not extend well to other tumor types. Another related work is to train a patch-level convolutional neural network (CNN) to select discriminative patches by using the Expectation-Maximum (EM) algorithm [10]. The experiments show the great contribution made by discriminative patches to image-level classification. However, the network model is only trained on discriminative patches recognized, and we found based on our experiments that this design would trap the classifier into a local minimum easily, where the classifier cannot differentiate the tumor subtypes with heterogeneous features leading to difficulty in reaching convergence, especially for small datasets.

In this study, we propose an iterative patch labelling algorithm to recognize discriminative and non-discriminative patches, with the aim of addressing the image-label-only issue. To make use of non-discriminative patches, we define the non-discriminative patches first, and then propose a thresholding scheme and a training policy for our iterative algorithm. We also design a novel model architecture, which is a feedforward neural network with *N*-branch outputs corresponding to *N* tumor classes, to calculate the discriminative probabilities for branches more independently. Our method is applied to classify WSIs of brain astrocytoma and oligodendroglioma using the MICCAI 2015 Challenge Dataset and the results show large improvement over the benchmark techniques.

2. METHODS

In this section, we provide details of our proposed method. Visual illustration is shown in Fig. 2.

2.1. Patch Subsetting

Considering the massive size of a WSI, patch subsetting is an essential step to be achieved in the data preprocessing process. Otsu's method is first applied on the thumbnails, to generate the binary masks distinguishing between tissue and background [11]. Then non-overlapping patches of size 500 by 500 are extracted, at 40X (0.25 microns per pixel) resolution scale. Invalid patch is discarded if it contains less than 80% tissue sections. There would be roughly 0.3K-15K valid patches extracted per WSI.

2.2. Feature Extraction

We extract the patch-level features by fine-tuning the VGG19 model that is pre-trained on ImageNet [12, 13]. Due to the overfitting concerns, we set the first 3 convolutional blocks as non-trainable and fine-tuned the last 2 convolutional blocks on our training dataset for a few iterations. After fine-tuning, all fully-connected layers are removed and a two-dimension global MaxPooling layer is added at the end, so that it can generate the 512D-visual-feature-code for each input patch.

2.3. EM-based Iterative Patch Labelling

We define discriminative patches to class *K* (**Dis-to-*K***) as the patches containing the representative features of cancer class *K* and non-discriminative patches (**Non-Dis**) as the patches that do not cover distinguishable features of any target cancer.

Our model is expected to perform classification for *N* different classes, and each patch *X* will be associated with *N* discriminative probabilities and *P(X, K)* represents the probability that patch *X* contains representative features of class *K*. Although patch label information is missing initially, each patch *X* would eventually be assigned as either Dis-to-*K* or Non-Dis at the end. Specifically, a patch *X* will be recognized as: (1) Dis-to-*K* if *P(X, K)* is the highest value among all *P(X, Q)* for all target class *Q* and *P(X, K)* is greater than 0.5. (2) Non-Dis if *P(X, Q)* is less than or equal to 0.5 for all target classes *Q*.

Our iterative patch labeling algorithm is designed to identify the Dis-to-*K* patches while performing patch classification simultaneously in an EM construct, which is summarized in Algorithm 1, and the details are given in the following subsections.

Algorithm 1 Iterative Patch Labelling Algorithm

```
procedure
    Initialize discriminative model $D$
    Assign all patches $X$ with WSI-level labels
    while convergence $C$ is not reached do
  [M-step]
        Use $X$ and their current labels $L$ to train $D$
  [E-step]
        Use $D$ to calculate $P(X, K)$ for each $X$ with class $K$
        Generate the possibility maps $PMap$ for each WSI
        Apply Gaussian Smoothing on $PMap$
        Apply Thresholding Scheme on $PMap$
        Update $L$
    Save $D$ architecture and its weights
```

2.3.1. Initialization

We first design a feedforward neural network as the discriminative model *D* for our algorithm. It contains 3 blocks and each block includes a dense layer, a Rectified Linear Unit (ReLU) layer and a Dropout layer. The neuron number for each dense layer is 256, which is the half of the length of patch features, and the dropout rate is 50%. To produce the discriminative probabilities *P(X, K)* for *N* classes more independently, the traditional output layer is modified from *1*-branch output to *N*-branch outputs. Those *N* branches are separated after the last Dropout layer and not sharing weights with others. The loss function of each branch is binary cross-entropy, since the *K*th-branch output can only be 0 or 1, indicating whether the input patch contains representative features of class *K*.

At the initial stage, all patches *X* are assumed to be discriminative to their image-level labels. The initial ground truth label *L(X, K)* at *K*-th branch output is set to 1 if class *K* is the image-level label of *X* and to 0 otherwise. Take binary classification between class-1 (corresponds to 1th-branch) and class-2 (corresponds to 2th-branch) as an example, the initial label vector for patches from class-1 WSI and class-2 WSI would be [1, 0] and [0, 1] respectively and Non-Dis patches would be labelled as [0, 0] in the training process.

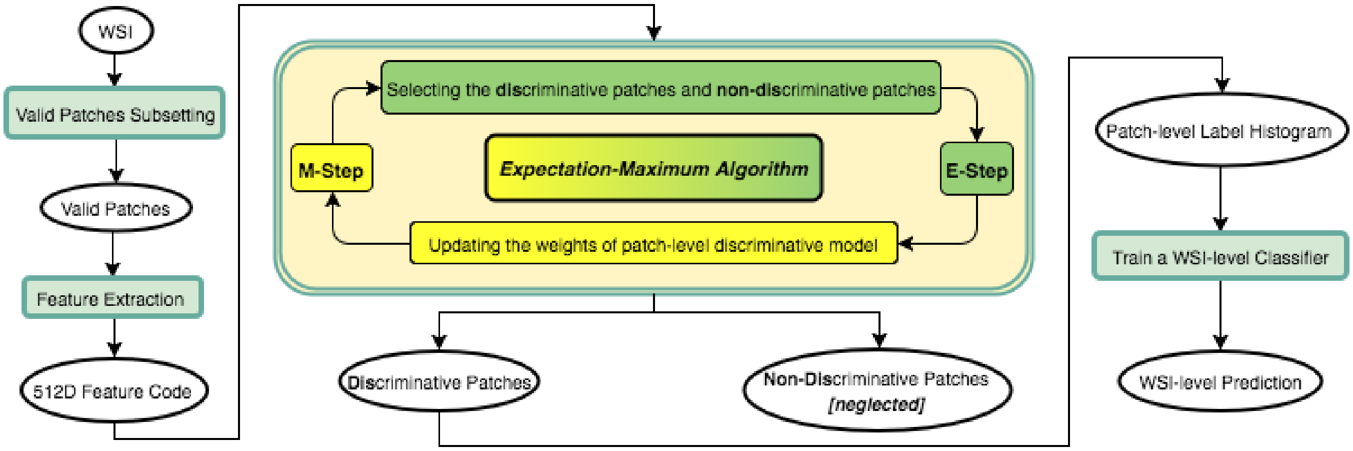

WSI
Valid Patches Subsetting
Valid Patches
Feature Extraction
512D Feature Code
Selecting the discriminative patches and non-discriminative patches
M-Step
Expectation-Maximum Algorithm
E-Step
Updating the weights of patch-level discriminative model
Discriminative Patches
Non-Discriminative Patches
[neglected]
Patch-level Label Histogram
Train a WSI-level Classifier
WSI-level Prediction

Fig. 2. Proposed pipeline.

2.3.2. *Maximization (M) and Expectation (E) Steps*

In Maximization (M) stage, all patches *X* with their current labels *L*, are used to train the discriminative model *D* for 2 epochs. The optimizer for our network is Adam and the total loss is the sum of each branch loss computed individually. In Expectation (E) step, the discriminative model *D* produces the up-to-date discriminative probability *P(X, K)* for all combination of patch *X* and class *K*. Those *P(X, K)* are used to generate the WSI-level probabilities map *PMap*. The map *PMap* includes *N* channels, corresponding to the *N* target classes. Then Gaussian smoothing is applied on each channel of *PMap*, considering the spatial relationship between patches. Finally, a special thresholding scheme is utilized on *PMap* to update their patch-level labels, which will be explained in detail in Section 2.3.3.

2.3.3. *Thresholding Scheme*

We introduce H_K as the sorted list containing the probabilities of all patches whose image-level label is class *K*. Then we define *T* as the thresholding percentage and S_K as the *T*-th percentile of H_K. Hence, given a patch *X*, if its image-level label is class *K* and correspondingly the other classes are *R*, then the thresholding scheme and its patch-level label updating would be performed by Formula 1. T is set to 0.7 in our pipeline.

$$\text{Relabel } X \text{ as } \begin{cases} \text{Non-Dis} & \textit{if } P(X,K) \leq S_K \\ & \textit{and } P(X,R) \leq 0.5 \\ \text{Dis-to-}K & \textit{else} \end{cases} \qquad (1)$$

2.3.4. *Training Policy*

We design a new training policy to use Non-Dis patches for next training iteration, along with all Dis-to-K patches. Although the Non-Dis patches will not be considered in WSI-level classification, the Non-Dis patches found in training process can benefits the recognition of Dis-to-*K* patches, by narrowing down the set of possibly discriminative patches and the algorithm can therefore concentrate on the difference between the patches discriminative to different classes. This benefit enhances the discriminative ability of the model and make it easier to reach the convergence, which will be explained in more detail in Section 3.2.

2.4. WSI-level Classification

When the convergence is reached by early stopping, the patch class distribution of WSIs can be computed easily. The Non-Dis patches are ignored at this stage and the label histograms of Dis-to-*K* patches are used to train another SVM classifier that learns the WSI-level decision fusion function [14], with the aim of increasing the classification robustness.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Setup

We performed four-fold cross validation to estimate our proposed method on CBTC dataset, which belongs to MICCAI 2015 Challenges. Lower grade glioma is a primary type of brain cancer, and the main challenge of the CBTC dataset is to classify astrocytoma and oligodendroglioma, which are the two most common subtypes of lower grade gliomas [15]. Owing to the high similarity of their morphological features, this classification task is hard to complete precisely even for pathologists [16]. Moreover, this dataset contains 32 whole slide pathology images (16 ***A****strocytoma* cases and 16 ***O****ligodendroglioma* cases), with only image-level labels released. To alleviate the imbalanced effects caused by varying number of valid patches extracted from each WSI, three data augmentation techniques are randomly applied in the training process, namely flipping tiles horizontally/vertically, shearing tiles at a random angle between 0.1 radians clockwise to 0.1 radians anti-clockwise, and rotating tiles 90, 180 and 270 degrees respectively.

3.2. Contribution of Non-Dis Patches

One of the uniquenesses of our pipeline is to make use of Non-Dis patches, whose importance is usually ignored by other approaches [9, 10]. As illustrated in Fig. 3, with the help of Non-Dis patches, the iterative training can gradually converge and identify the real discriminative patches effectively. If Non-Dis patches are simply discarded during the iterative training, the network would have difficulty converging and the patches would keep oscillating between Dis-to-*A* and Dis-to-*O*, as shown in Fig. 4.

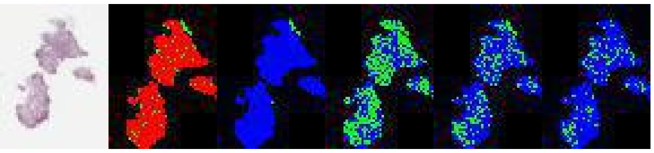


Fig. 3. The patch labelling visualization of an astrocytoma instance with our proposed training policy. The visualization scheme utilized: Blue for Non-Dis patches, red for Dis-to-*O* patches and green for Dis-to-*A* patches. The first subplot is the thumbnail of WSI, and other are captured at 1st, 5th, 6th, 100th and 293th (the last) iteration respectively.

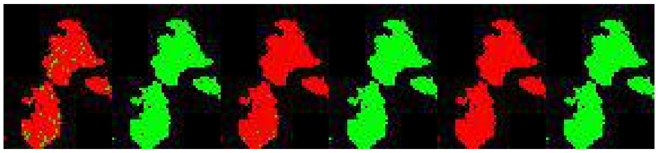


Fig. 4. The patch labelling visualization of same instance as shown in Fig.3. Same experiment setting and pipeline, with Non-Dis patches excluded from training set. They are captured at 1st, 5th, 6th, 98th and 99th and 100th (the pre-set maximum) iteration respectively.

3.3. Classification Results

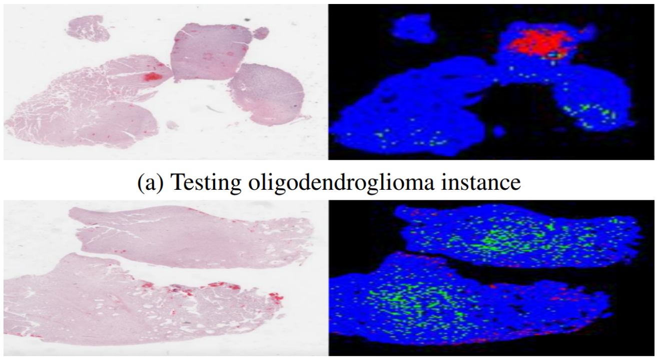
The classification results are presented in Table 1. We compared the performance on the MICCAI 2015 Challenge Dataset between the following approaches: (1) CNN-Feat-SVM: We employ a pre-trained VGG19 to extract features and produce patch labels and then train a SVM to learn the WSI-level decision fusion function. (2) Finetune-CNN-Feat-SVM: Similar to (1) except the fine-tuning for VGG19 is performed. In these two approaches above, patches are assigned the same label as the WSI containing these patches. (3) Iter-Finetune-CNN-SVM[Discriminative]: The training of VGG19 is completed in an EM construct, where the patch label updates iteratively and only the latest discriminative patches would be used for next training iteration. (4) Iter-Finetune-CNN-SVM[Both]: Our proposed pipeline, which makes use of non-discriminative patches, achieved the best performance.

Methods	Acc.
CNN-Feat-SVM	62.50%
Finetune-CNN-Feat-SVM	69.13%
Iter-Finetune-CNN-SVM[Discriminative]	76.62%
Iter-Finetune-CNN-SVM[Both]	**84.38%**

Table 1. Classification results between astrocytoma and oligodendroglioma.

3.4. ROIs Visualization

With post-processing, the regions of interest (ROIs), which consist of the discriminative patches that are identified and used in the final WSI-level classification, can be visualized in Fig. 5.



(a) Testing oligodendroglioma instance

(b) Testing astrocytoma instance

Fig. 5. Two testing WSI instances with ROIs marked.

4. CONCLUSIONS

In this study, we proposed a method to address the Whole Slide Pathology Images (WSI) classification challenge with only image-level labels provided, by designing an iterative patch labelling algorithm to recognize discriminative and non-discriminative patches. Our method achieved the best classification performance of 84.38% on the MICCAI 2015 Challenge Dataset. By identifying the discriminative patches, our model could produce a more explainable and reliable classification results, which are valuable to human pathologists in the diagnosis decision-making process. Furthermore, our approach demonstrates the importance of recognizing non-discriminative patches in image classification tasks.

5. REFERENCES

[1] A. M. Gardeck, J. Sheehan, and W. C. Low, "Immune and viral therapies for malignant primary brain tumors," *Expert Opinion on Biological Therapy*, vol. 17, no. 4, pp. 457–474, 4 2017.

[2] R. Sparks and A. Madabhushi, "Explicit shape descriptors: Novel morphologic features for histopathology classification," *Medical image analysis*, vol. 17, no. 8, pp. 997–1009, 2013.

[3] H. Chang, Y. Zhou, A. Borowsky, K. Barner, P. Spellman, and B. Parvin, "Stacked predictive sparse decomposition for classification of histology sections," *International journal of computer vision*, vol. 113, no. 1, pp. 3–18, 2015.

[4] M. Kandemir, C. Zhang, and F. A Hamprecht, "Empowering multiple instance histopathology cancer diagnosis by cell graphs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 228–235.

[5] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE transactions on medical imaging*, vol. 35, no. 3, pp. 738–751, 2016.

[6] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging 2014: Digital Pathology*. International Society for Optics and Photonics, 2014, vol. 9041, p. 904103.

[7] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, I. Eric, and C. Chang, "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 947–951.

[8] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2013, pp. 411–418.

[9] J. Barker, A. Hoogi, A. Depeursinge, and D. L. Rubin, "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles," *Medical image analysis*, vol. 30, pp. 60–71, 2016.

[10] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.

[11] 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.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[14] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Publishing Company, Incorporated, 1st edition, 2008.

[15] D. A. Forst, B. V. Nahed, J. S. Loeffler, and T. T. Batchelor, "Low-grade gliomas," *The oncologist*, vol. 19, no. 4, pp. 403–413, 2014.

[16] M. Gupta, A. Djalilvand, and D. J. Brat, "Clarifying the diffuse gliomas: an update on the morphologic features and markers that discriminate oligodendroglioma from astrocytoma," *American journal of clinical pathology*, vol. 124, no. 5, pp. 755–768, 2005.