Interpretable Histopathology Image Diagnosis via Whole Tissue Slide Level Supervision

Zhuoyue Wu, Hansheng Li, Lei Cui(✉), Yuxin Kang, Jianye Liu, Haider Ali, Jun Feng(✉), and Lin Yang(✉)

School of Information Science and Technology, Northwest University, Xi'an, China
{leicui,fengjun,linyang}@nwu.edu.cn

Abstract. The deep learning methods supervised by annotating different regions of histopathology images (patch-level labels) have achieved promising outcomes in assisting pathologic diagnosis. However, most clinical data only contains label information for the whole slide image (WSI-level labels), so the methods supervised by WSI-level labels are more necessary than the ones supervised by patch-level labels. Additionally, various methods supervised by WSI-level labels ignore the contextual relations among patches extracted from a WSI, making incorrect predictions for some patches in a WSI and further misclassifying the WSI. In this paper, we propose to utilize an interpretable dual encoder network with a context-capturing RNN module to capture the contextual relations among all patches extracted from a WSI. Besides, we propose to utilize a feature attention module to weigh the importance of each patch automatically. More importantly, visualization of weight for each patch in a WSI demonstrates that our approach matches the concerns of pathologists. Furthermore, extensive experiments demonstrate the superiority of the interpretable dual encoder network.

Keywords: Whole slide image · Contextual relations · Interpretable · Patch-level labels · WSI-level labels.

1 Introduction

Histopathology image diagnosis plays a critical role in treating disease, as it can guide clinic doctors to determine the follow-up treatment plan [1,2]. Recently, deep learning methods for whole slide images (WSIs) diagnosis have continually developed to empower pathologists' efficiency and accuracy. According to which types of labels are available, these deep learning methods can be separated into two classes: methods supervised by patch-level labels [3–6] and WSI-level ones [7–10].

These methods supervised by patch-level labels require manually annotating regions of different tissue types (patch-level labels) for WSIs. Conceivably, patch-level labels' acquisition is time-consuming and labor-intensive, making it

J. Feng and L. Yang—Joint corresponding authors.

C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 40–49, 2021.
https://doi.org/10.1007/978-3-030-87589-3_5

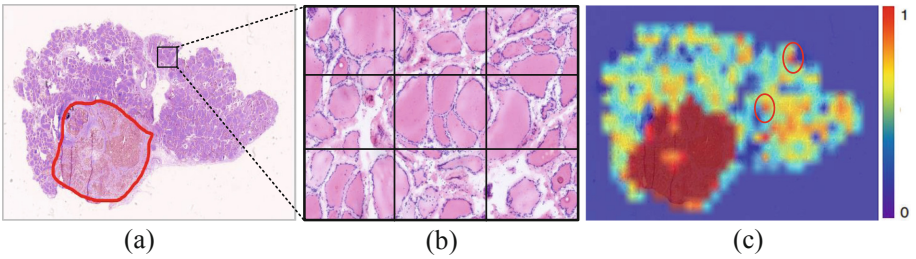

Fig. 1. The Illustration of contextual relations problem. (a) A WSI with malignant lesion outlined by an expert. (b) A benign sub-region in a WSI, which includes nine patches. It is worth noting that the center patch is more likely to have the same properties as its neighbors. (c) All patchs' malignant probabilities in the WSI are obtained by method that ignore contextual relations [10].

challenging to build large training datasets. However, patch-level labels for WSIs can further improve the model performance by modeling context relations among neighboring patches [3]. As an example of context relations, Fig. 1b reveals that the center patch may fall in the benign region with a high probability when its neighboring patches are in the benign region. To model such context relations among neighboring patches, Zanjani et al. [4] utilized Convolutional Neural Network (CNN) to extract features from neighboring patches, and then Conditional Random Field (CRF) is applied to these patches to refine the predicted probability map in the post-processing stage. In parallel to this work, Li et al. [5] proposed a neural conditional random field (NCRF) model that can be trained in an end-to-end manner, avoiding the post-processing stage. It is worth noting that the above methods' tremendous success depends on the patch-level labels for all WSIs.

In contrast, WSI-level labels can easily be obtained from pathological reports. Thus the methods supervised by WSI-level labels are more necessary than the ones supervised by patch-level labels. As a typical WSI-level supervised method, Hou et al. [7] initially took all patches in the WSI as training samples and eliminated the patches with low classification probability iteratively. In [8], the authors extended this approach by clustering patches wisely and eliminating the patches that are less discriminative for the classification task. Chen et al. [10] believe that the eliminated policy is a hard sampling technique which makes a binary decision to select samples and proposed a soft-weighted technique called rectified cross-entropy loss (L_{RCE}). Besides, they introduced an upper transition loss (L_{UT}) to improve the patches' classification accuracy further. However, these methods supervised by WSI-level labels predict each patch independently in the inference stage. From our perspective, the independent predictions ignore the context relations among neighboring patches, leading to inconsistent predictions of patches in the same region. For instance, as shown in the red circle of Fig. 1c, some patches in the benign region are misclassified as malignant ones.

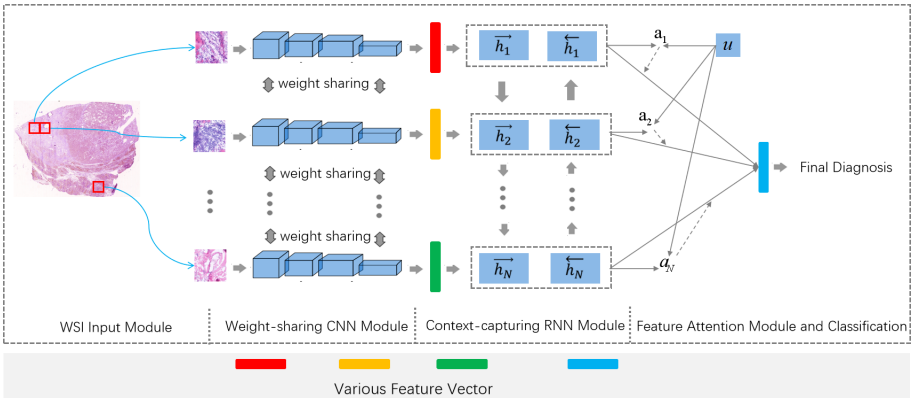

Fig. 2. Overview of the interpretable dual encoder network. It is worth noting that the blue color represents the WSI-level feature vector and other colors represent different patch-level feature vectors.

Overall, we are surprised to find that there has been little discussion about how to model contextual relations among patches when lacking patch-level labels. Thus, this paper regards patches extracted from a WSI as a sequence and proposes to utilize an RNN module to take contextual relations among all patches extracted from the WSI into account when only WSI-level labels are available. Besides, considering that not all patches extracted from a WSI contribute equally to the final diagnosis, we propose to utilize a feature attention module [12] to weigh the importance of each patch automatically. Finally, An interpretable dual encoder network is formed by combining the modules mentioned above. The significant superiority of this network is as follows.

(1) Our network merely requires WSI-level labels that are easy to obtain, which significantly saves a lot of labor cost on annotations.
(2) The context-capturing RNN module of our network is adopted to automatically model the contextual relations among all patches extracted from a WSI, which are different from previous patch-level supervised methods that just model the contextual relations among neighboring patches.
(3) Our network's feature attention module automatically captures the most critical patch for final diagnosis, making we can produce fine heatmaps by visualizing each patch's importance.

2 Methodology

Figure 2 shows an overview of the interpretable dual encoder network. It contains five primary modules: WSI input module, weight-sharing CNN module, context-capturing RNN module, feature attention module, and classification module.

(1) The WSI input module takes a WSI as input and then splits the WSI into non-overlapping patches.
(2) The weight-sharing CNN module acts as a patch's appearance encoder that takes patches extracted from a WSI as input and then codes each patch as a patch-level feature vector.
(3) The context-capturing RNN module jointly processes these patch-level feature vectors and then encodes each patch-level feature vector as context-aware feature vector. By jointly processing these patch-level feature vectors, we make sure that we capture the contextual relations among all patches extracted from a WSI.
(4) The feature attention module merges context-aware feature vectors from each time step into a WSI-level feature vector. It is worth noting that the feature attention module can weigh the importance of each patch automatically.
(5) The classification module uses the WSI-level feature vector for the final diagnosis.

2.1 WSI Input Module

A WSI from the training set is taken as X_i, and meanwhile, its WSI-level label is taken as Y_i. Thus the training set can be expressed as $\{(X_1, Y_1), ..., (X_M, Y_M)\}$, where M is the total number of training samples. The WSI input module takes a WSI X_i as input and then splits the WSI into non-overlapping $512 * 512$ patches. Before passing these non-overlapping patches to the weight-sharing CNN module, we discard non-informative background patches to reduce the computational cost and mark the rest of the foreground patches as $\{x_{i,1}, x_{i,2}, ..., x_{i,N}\}$.

2.2 Weight-Sharing CNN Module

To fully extract the foreground patches' discriminative feature, we employ a modified CNN $f(\cdot)$ as patches' appearance encoder. Specifically, we modify the Resnet [13] by replacing the classification layer with a feature reduction layer that is essentially a three-layer fully connected neural network. The convolution layers of modified Resnet encode a patch into a patch-level feature vector, and further, the feature reduction layer of modified Resnet reduces the patch-level feature vector's dimension to 256. Formally, as shown in Eq. 1, the modified Resnet can map a patch to a 256-dimensional feature vector $v_{i,t}$, where $t = 1, 2, ..., N$. Due to GPU memory limitation, we modify Resnet pre-trained on ImageNet and only update the feature reduction layer's weights during the training process.

$$v_{i,t} = f\left(x_{i,t}\right) \tag{1}$$

2.3 Context-Capturing RNN Module

After getting a sequence of patch-level feature vectors $\{v_{i,1}, ..., v_{i,N}\}$, we propose to use a Bidirectional Long Short-Term Memory (BLSTM) [14] to model the contextual relations among them. Note that the BLSTM consists of two

independent processing streams, one moving left to right($\overrightarrow{LSTM}(\cdot)$) and the other right to left($\overleftarrow{LSTM}(\cdot)$). As a matter of convenience, we use the index t ($t = 1...N$) to denote the position of the patch-level vector $v_{i,t}$. For the t-th patch-level feature vector $v_{i,t}$, the sub-network $\overrightarrow{LSTM}(\cdot)$ merges $v_{i,t}$ with its previous output $\overrightarrow{h_{i,t-1}}$ to generate the forward feature vector $\overrightarrow{h_{i,t}}$. The above calculation process is shown in Eq. 2. As same as the sub-network $\overrightarrow{LSTM}(\cdot)$, the calculation process of sub-network $\overleftarrow{LSTM}(\cdot)$ is shown in Eq. 3.

$$\overrightarrow{h_{i,t}} = \overrightarrow{LSTM}(v_{i,t}, \overrightarrow{h_{i,t-1}}) \tag{2}$$

$$\overleftarrow{h_{i,t}} = \overleftarrow{LSTM}(v_{i,t}, \overleftarrow{h_{i,t-1}}) \tag{3}$$

Next, Eq. 4 merge the forward vector $\overrightarrow{h_{i,t}}$ with the backward vector $\overleftarrow{h_{i,t}}$ to generate a context-aware feature vector $h_{i,t}$ for the t-th patch-level feature vector.

$$h_{i,t} = [\overrightarrow{h_{i,t}}, \overleftarrow{h_{i,t}}] \tag{4}$$

In this paper, the hyperparameter details of LSTM are as follows. The number of features in the hidden state is 512, and the number of recurrent layers is 3.

2.4 Feature Attention Module and Classification

Given a WSI, not all patches extracted from it contribute equally to the final diagnosis. Therefore, we embed the feature attention module in the interpretable dual encoder network to weigh the importance of each patch automatically. The feature attention module can map the context-aware feature vector set $\{h_{i,1}, ..., h_{i,N}\}$ to a WSI-level feature vector S_i by a weighted sum of these context-aware feature vectors. The details of the weighted sum are as follows. First, fully connected layer projects each patch-level feature vector $h_{i,t}$ to a hidden representation $u_{i,t}$. The fully connected layer is defined in Eq. 5:

$$u_{i,t} = \tanh(W_u h_{i,t} + b_u) \tag{5}$$

As shown in Eq. 6, the importance of context-aware feature vector $h_{i,t}$ is weighed by the similarity between the hidden representation $u_{i,t}$ and a context vector u learned during training.

$$a_{i,t} = \frac{\exp(u_{i,t}^{\top} u)}{\sum_{t=0}^{N} \exp(u_{i,t}^{\top} u)} \tag{6}$$

Next, the WSI-level feature vector S_i is formed by a weighted sum of these context-aware feature vectors $\{h_{i,1}, ..., h_{i,N}\}$.

$$S_i = \sum_{t=0}^{N} a_{i,t} h_{i,t} \tag{7}$$

Finally, we use a softmax classifier to predict the WSI-level label $\hat{Y}_i$ for the WSI X_i. The softmax classifier takes the WSI-level feature vector S_i as input and then outputs estimated probabilities $p(y|X_i)$ for all categories.

$$p(y|X_i) = softmax(W_p S_i + b_p) \tag{8}$$

$$\hat{Y}_i = \arg\max p(y|X_i) \tag{9}$$

3 Experiments

3.1 Dataset

To validate the network's effectiveness, we create a digital thyroid frozen section dataset, including 547 WSIs. According to the subsequent surgical plan, these WSIs can be categorized as benign or malignant. As shown in Table 1, the collected dataset includes 97 benign and 154 malignant WSIs for training, 19 benign and 25 malignant WSIs for validation, 85 benign and 167 malignant WSIs for testing. The slides in the training set and testing set only have WSI-level labels. However, the slides in the validation set have both WSI-level labels and patch-level labels.

Table 1. Summary of experimental data

	Benign	Malignant	Total
Train	97	154	251
Val	19	25	44
Test	85	167	252
Total	201	346	547

3.2 Implementation Details

The WSI input module throws the non-informative background patches away as Chen et al. [10]. The interpretable dual encoder network is trained with NVIDIA GeForce GTX 2080 Ti GPU. In the training process, a batch size of 1 WSI is feed into the network. Meanwhile, it is trained in an end-to-end way for 100 epochs, using adam optimizer with a learning rate of 0.0001.

3.3 Performance Comparison

In this sub-section, we conduct ablation experiments to investigate the feature attention module's effectiveness firstly. Then we implement a method supervised by WSI-level labels [10] and compare it with our solution on the different backbone. In the experiments, we used Accuracy, Precision, Recall, and F1-Score as our evaluation criteria.

Table 2. Performance of feature attention module (FA) on different benchmark models

Method				Metrics			
Resnet18 + BLSTM	Resnet50 + BLSTM	Resnet152 + BLSTM	FA	Accuracy	Precision	Recall	F1-score
✓				89.68%	100%	84.43%	91.56%
✓			✓	93.25%	96.30%	93.41%	94.83%
	✓			86.91%	100%	80.24%	89.04%
	✓		✓	90.48%	96.73%	94.27%	95.48%
		✓		90.48%	100%	85.63%	92.26%
		✓	✓	91.67%	97.40%	89.82%	93.46%

Effect of Feature Attention Module. To test whether the additional feature attention module can help improve model performance, we first implement three benchmark models and then embed the feature attention module in each benchmark model. Table 2 illustrates the experimental results with and without the feature attention module. It can be seen that these benchmark models have high precision but low recall for malignant class. Our explanation for this phenomenon is that benchmark models tend to classify some malignant WSIs as benign ones due to sizeable benign lesion regions in these malignant WSIs. In an extreme case, only less than 1% of regions on a misclassified WSI may be malignant while the rest are benign. However, when adding the feature attention module, these models can achieve high accuracy and high recall. Specifically, the feature attention module increased the three benchmark model F1-score by 3.27%, 6.44%, and 1.20%, respectively. Meanwhile, the feature attention module can also increase three benchmark model accuracy by 3.57%, 3.57%, and 1.19%, respectively. It demonstrates that the feature attention module can automatically weigh each patch's importance and further capture the patches most relevant to the final diagnosis.

Comparision with Method Supervised by WSI-level Labels. We compare our method with Chen's method [10] in Table 3. From Table 3, we can observe that our method's performance is better than Chen et al. on the different backbone. The possible reason is that our method models contextual relations among all patches in a WSI, which is different from Chen's method that predicts each patch in a WSI separately. Besides, it is worth noting that Chen's method uses patch-level labels in the validation set to select the best model for testing, whereas we only use WSI-level labels.

To further compare the method proposed by Chen et al. with our method, we visualize the expert annotation in WSI, each patch's malignant probability obtained by Chen's method and each patch's weight obtained by our network's feature attention module, respectively in Fig. 3. We can find that both Chen's model and our model focus on the malignant lesions.

Table 3. Performance of different deep learning algorithms

Methods	Accuracy	Precision	Recall	F1-score
Chen et al.(Resnet18+$L_{RCE} + L_{UT}$) [10]	85.71%	97.12%	80.83%	88.23%
Ours(Resnet18+BLSTM+FA)	93.25%	96.30%	93.41%	94.83%
Chen et al.(Resnet50+$L_{RCE} + L_{UT}$) [10]	88.49%	92.07%	90.42%	91.24%
Ours(Resnet50+BLSTM+FA)	90.48%	96.73%	94.27%	95.48%
Chen et al.(Resnet152+$L_{RCE} + L_{UT}$) [10]	87.30%	96.55%	83.83%	89.74
Ours(Resnet152+BLSTM+FA)	91.67%	97.40%	89.82%	93.46%

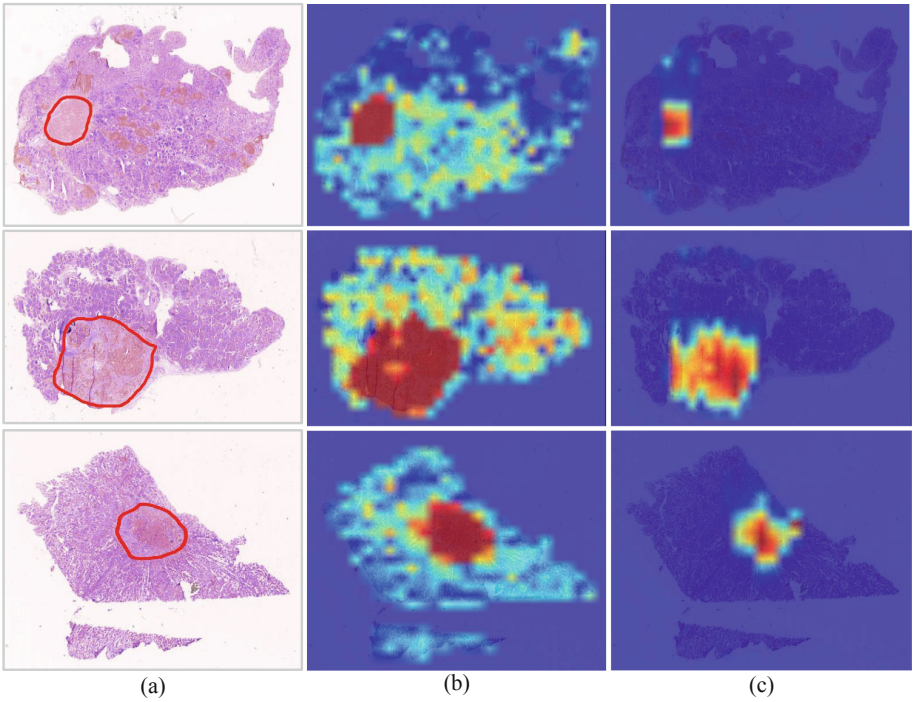

Fig. 3. Prediction visualization of selected examples. (a) WSI with malignant lesion outlined by an expert. (b) Visualizing each patch's malignant probability obtained by Chen's method [10]. (c) Visualizing each patch's weight obtained by our network's feature attention module.

4 Conclusion

In this paper, we find that there has been little discussion about how to model contextual relations among all patches extracted from a WSI when lacking patch-level labels. Thus we propose to utilize an interpretable dual encoder network

with a context-capturing RNN module to capture the contextual relations among all patches extracted from a WSI. The comparative experiments on thyroid datasets demonstrate the superiority of our method. Besides, we propose to utilize a feature attention module to weigh the importance of each patch automatically. Extensive ablation experiments demonstrated that the feature attention module could capture the patches most relevant to the final diagnosis.

Acknowledgment. This work was funded by the Natural Science Foundation of Shaanxi Province of China(2021JQ-461).

References

1. Li, Y., Chen, P., Li, Z., Su, H., Yang, L., Zhong, D.: Rule-based automatic diagnosis of thyroid nodules from intraoperative frozen sections using deep learning. Artif. Intell. Med. **108**, 101918 (2020)
2. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: a survey. Medical Image Analysis, p. 101813 (2020)
3. Kong, B., Wang, X., Li, Z., Song, Q., Zhang, S.: Cancer metastasis detection via spatially structured deep network. In: International Conference on Information Processing in Medical Imaging, pp. 236–248. Springer (2017). https://doi.org/10.1007/978-3-319-59050-9_19
4. Zanjani, F.G., Zinger, S., et al.: Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces. In: Medical imaging 2018: Digital pathology, vol. 10581. International Society for Optics and Photonics, p. 105810I (2018)
5. Li, Y., Ping, W.: Cancer metastasis detection with neural conditional random field. arXiv:1806.07064 (2018)
6. Zhang, Z., et al.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Mach. Intell. **1**(5), 236–245 (2019)
7. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433 (2016)
8. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7234–7242 (2017)
9. Wang, X., et al.: Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans. Cybern. **50**(9), 3950–3962 (2019)
10. Chen, H., et al.: Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 351–359. Springer (2019). https://doi.org/10.1007/978-3-030-32239-7_39
11. Campanella, G., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (2019)
12. Zhou, P., et al.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers), pp. 207–212 (2016)

13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Zhang, S., Zheng, D., Hu, X., Yang, M.: Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pp. 73–78 (2015)