Improving the Self-Supervised Pretext Task for Histopathologic Subtype Classification

Ruiwen Ding[1] RuiwenDing@mednet.ucla.edu
Anil Yadav[1] AnilYadav@mednet.ucla.edu
Erika Rodriguez[2] erikarodriguez@mednet.ucla.edu
Ana Cristina Araujo Lemos da Silva[3] anacals@gmail.com
William Hsu[1] whsu@mednet.ucla.edu
[1] *Medical & Imaging Informatics, Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA*
[2] *Department of Pathology & Laboratory Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA*
[3] *Federal University of Uberlândia, MG, Brazil*

Abstract

In computational pathology, fully-supervised convolutional neural networks have been shown to perform well on tasks such as histology segmentation and classification but require large amounts of expert-annotated labels. In this work, we propose a self-supervised learning pretext task that utilizes the multi-resolution nature of whole slide images to reduce labeling effort. Given a pair of image tiles cropped at different magnification levels, our model predicts whether one tile is contained in the other. We hypothesize that this task induces the model to learn to distinguish different structures presented in the images, thus benefiting the downstream classification. The potential of our method was shown in downstream classification of lung adenocarcinoma histologic subtypes using H&E-images from the National Lung Screening Trial.

Keywords: Self-supervised learning, pretext task, histopathology, lung adenocarcinoma

1. Introduction

Early-stage invasive lung adenocarcinoma (LUAD) exhibits heterogeneous biological behaviors within the same tumor. Patients are classified as having one of five predominant histologic subtypes (lepidic, acinar, papillary, micropapillary, solid), each associated with a different prognosis. Supervised convolutional neural networks can improve the accuracy and reduce subjectivity of LUAD histologic subtype classification (Gertych et al., 2019). However, they rely on a large amount of expert annotation. Self-supervised learning (SSL) techniques that leverage the multi-resolution nature of whole slide images (WSIs) can be used to reduce labeling effort. WSIs can be acquired at different magnification levels (low to high are 5x, 10x, 20x, 40x) with higher levels capturing local cellular features (more zoomed in) and lower levels capturing global spatial morphology (more zoomed out). As shown in Figure 1a, each subtype has a different gland architecture and cell morphology, which should be reflected in the learned embeddings. In this work, we propose a pretext task $P_{Proposed}$ that predicts whether an image cropped at a higher magnification level is contained in another image cropped at a lower magnification level. We hypothesize that this task induces the model to learn to distinguish different structures presented in WSIs, and thus benefit the downstream classification where those structures are also present.

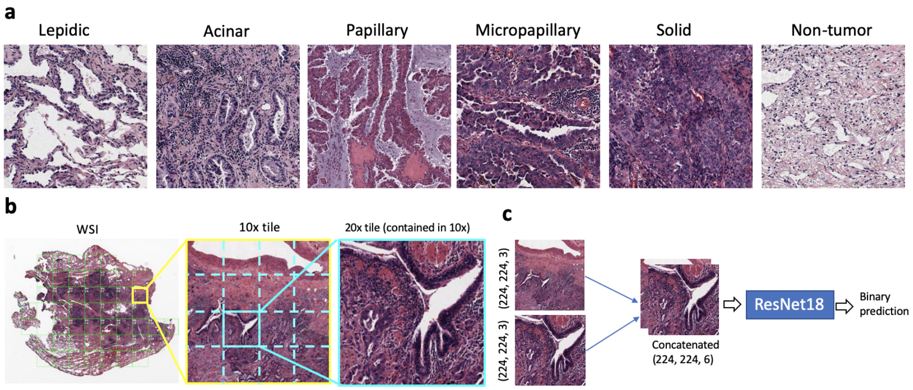

Figure 1: (a) 5 LUAD histologic subtypes plus non-tumor tiles cropped at 20x magnification. (b) Example image pair where the 20x tile is contained in the 10x tile. (c) Input and output of the pretext task $P_{Proposed}$.

2. Methods

A total of 407 H&E-stained WSIs of 146 patients diagnosed with early-stage LUAD who had computed tomography and pathology images were obtained from the National Lung Screening Trial. $P_{Proposed}$ takes a pair of 10x and 20x image tiles as input and predicts whether one of the tiles is contained in the other one. Non-overlapping tiles were cropped at 10x magnification. From each 10x tile, 16 non-overlapping tiles were generated at 20x. All tiles regardless of magnification levels were sized 512 × 512. Figure 1b summarizes the approach for generating 20x tiles from 10x WSIs (paired, class 1). For the unpaired (class 0), the 20x tile was randomly sampled from another 10x tile from the same patient. 241,640 (80%) pairs of tiles were used for training and 60,422 (20%) pairs for validation with equal numbers of paired and unpaired tiles. The RGB image pair was concatenated channel-wise such that the resulting image has 6 channels (Figure 1c). The concatenated images were fed into an ImageNet-initialized ResNet18, where the adaptive average pooling layer was followed by a dropout layer (p = 0.2), linear layer, and sigmoid activation. The network was trained using an Adam optimizer with batch size 32, a learning rate and weight decay of 0.0001, and binary cross-entropy loss. Early stopping was monitored by validation loss with patience of 5 epochs. The learned weights were transferred to the downstream task for finetuning.

We performed experiments to evaluate the impact of $P_{Proposed}$ compared against: 1)ImageNet pre-trained weights ($P_{ImageNet}$), 2) a common pretext task ($P_{MagLevel}$) that predicts the magnification level (5, 10, 20, or 40x) of a tile (Koohbanani et al., 2021), and 3) the state-of-the-art contrastive learning method SimSiam ($P_{SimSiam}$) (Chen and He, 2021). The feature extractor for both $P_{MagLevel}$ and $P_{SimSiam}$ was ImageNet-initialized ResNet18. See the code for their hyperparameters.

As for the downstream task (D), 316 20x tiles were annotated and data augmentation was applied. Stratified five-fold cross validation was used. Within each fold, there were 60% training, 20% validation, and 20% testing tiles. Performance was measured using F1-score. The network and hyperparameters were the same as in $P_{Proposed}$ except that the batch size was 16.

3. Results and Conclusions

Table 1: Average F1 score on test sets (n = 106) for downstream task D

$\mathrm{D}_{pretrain}$	Lepidic (n = 8)	Acinar (n = 23)	Papillary (n = 7)	Solid (n = 14)	Nontumor (n = 54)
$\mathrm{D}_{Scratch}$	0.487 ± 0.153	0.701 ± 0.0764	0.756 ± 0.110	0.802 ± 0.0762	0.853 ± 0.0420
$\mathrm{D}_{ImageNet}$	**0.668 ± 0.0890**	0.729 ± 0.0343	0.790 ± 0.108	0.843 ± 0.0510	**0.886 ± 0.0225**
$\mathrm{D}_{MagLevel}$	0.615 ± 0.0832	0.701 ± 0.0590	0.750 ± 0.141	0.814 ± 0.0629	0.860 ± 0.0330
$\mathrm{D}_{SimSiam}$	0.539 ± 0.173	0.724 ± 0.0805	0.833 ± 0.118	0.794 ± 0.0748	0.831 ± 0.0614
$\mathrm{D}_{Proposed-NoImageNet}$	0.630 ± 0.0623	0.718 ± 0.103	0.835 ± 0.151	0.829 ± 0.104	0.863 ± 0.0311
$\mathrm{D}_{Proposed}$	0.656 ± 0.100	**0.789 ± 0.0980**	**0.844 ± 0.129**	**0.867 ± 0.0730**	0.884 ± 0.0274

Table 1 summarizes the average F1 score and standard deviation for the downstream task of classifying LUAD subtypes using different pre-training methods. Micropapillary was excluded due to too few annotations. $P_{Proposed}$ substantially improved the downstream task $D_{Proposed}$ compared to $D_{Scratch}$. $D_{Proposed}$ achieved the highest F1 score for acinar, papillary, and solid subtypes; pre-training with a large dataset such as ImageNet achieved the best average F1 score for lepidic and non-tumor for $D_{ImageNet}$. While SSL effectively leverages unlabeled data to improve model training, large labeled datasets such as ImageNet for training is still preferred. Further, $D_{Proposed-NoImageNet}$ outperformed $D_{Scratch}$, which indicates $P_{Proposed}$ learned useful embeddings when not initialized by ImageNet weights. $D_{MagLevel}$ and $D_{SimSiam}$ improved upon $D_{Scratch}$ but did not outperform $D_{Proposed}$, demonstrating the informative value of our pretext learning task. We posit that $D_{Proposed}$ may be learning embeddings related to the macro- and micro- structures presented in various histologic subtypes. It is also possible that ImageNet weights won't have as much of a prominent benefit to $D_{ImageNet}$ once we increase the downstream sample size. Future work includes optimizing hyperparameters using a grid search.

In summary, we proposed a SSL pretext task for LUAD subtype classification and showed its effectiveness by comparing it with other pre-training methods. Our novel pretext task, which forces the model to understand tissue structures and identify features across different magnifications, can improve downstream task results. Our pre-text task can be beneficial when learning salient features from multi-resolution images.

References

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

Arkadiusz Gertych, Zaneta Swiderska-Chadaj, Zhaoxuan Ma, Tomasz Markiewicz, Szczepan Cierniak, Hootan Salemi, Samuel Guzman, Ann E Walts, Beatrice S Knudsen, et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific reports*, 9(1):1–12, 2019.

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 40(10):2845–2856, 2021.