Masked Autoencoders Pre-training in Multiple Instance Learning for Whole Slide Image Classification

Jianpeng An [1,2] anjianpeng@tju.edu.cn
Yunhao Bai [1] bai951785964@tju.edu.cn
Huazhen Chen [1] huazhenchen@tju.edu.cn
Zhongke Gao [1] zhongkegao@tju.edu.cn
Geert Litjens[2] Geert.Litjens@radboudumc.nl
[1] *School of Electrical and Information Engineering, Tianjin University, Tianjin, China*
[2] *Radboud University Medical Center, Department of Pathology, Nijmegen, the Netherlands*

Editors: Under Review for MIDL 2022

Abstract

End-to-end learning with whole-slide digital pathology images is challenging due to their size, which is in the order of gigapixels. In this paper, we propose a novel weakly-supervised learning strategy that combines masked autoencoders (MAE) with multiple instance learning (MIL). We use the output tokens of a self-supervised, pre-trained MAE as instances and design a token selection module to reduce the impact of global average pooling. We evaluate our framework on the assessment of whole-slide image classification on Camelyon16 dataset, showing improved performance compared to the state-of-the-art CLAM algorithm.

Keywords: histopathology, self-supervised learning, multiple instance learning.

1. Introduction

End-to-end classification of whole-slide-images (WSI) using deep learning is a challenging task due to the sheer size of these images, typically in the range of gigapixels. Most commonly this is circumvented by training patch-based deep learning systems using pixel-level annotations and then aggregating the patch-level results. However, obtaining such annotations is expensive and generally requires the input of skilled and experienced pathologists. As a result, weakly-supervised methods such as multiple instance learning (MIL) are popular in WSI classification. Most MIL methods use a patch-based convolutional neural network, pre-trained on ImageNet, to extract descriptive feature representations for WSI analysis. Subsequently, the resultant feature representations are used to obtain a slide level prediction, either through patch classification and aggregation (Campanella et al., 2019) or a secondary neural network (Lu et al., 2021). However, these approaches suffer from a domain shift from natural images to histopathology, making the encoded features non-optimal. Recently, a new self-supervised learning (SSL) paradigm was introduced: masked autoencoders (MAE) (He et al., 2021), which could circumvent these issues. It is a transformer-based encoder-decoder framework for reconstructing the masked part of the input, and only the encoder would be applied for downstream tasks after the training process. Motivated by this, we propose a MAE-based MIL framework (MAE-MIL) for WSI classification. In addition, a token selection module (TSM) is devised in MAE-MIL to fully explore effective feature representations of instances. We evaluate our method for weakly supervised classification of the Camelyon16 dataset. We compare against the state-of-the-art weakly-supervised method CLAM (Lu et al., 2021) and ablation study show the effectiveness of our method.

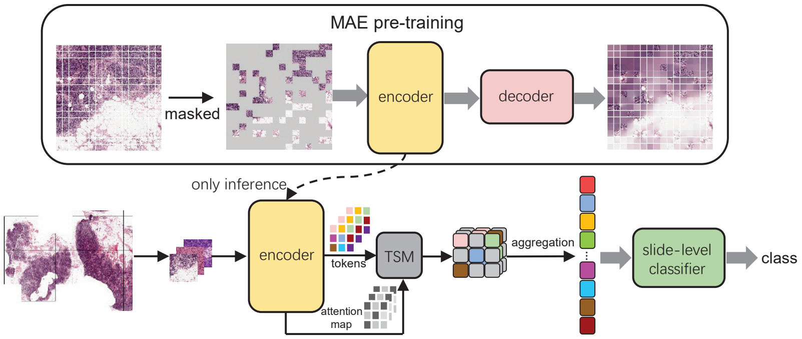

Figure 1: MAE-MIL Framework: After MAE pre-training, tile images are cropped from each slide and fed into the pre-trained encoder to extract the token features, and the top-k token features are selected by the attention map from the last transformer block of the encoder. Then the selected token features are aggregated into slide-level representations for training slide-level classifier.

2. Method

A schematic overview of our MAE-MIL framework is shown in Fig. 1. For computational efficiency, we first downsample each WSI to the $\times 5$ magnification level and randomly crop tiles to build a pre-training dataset. In the training process of MAE, we split each tile into N tokens and randomly mask out 75% of them, then after the encoder extracts latent features from unmasked tokens, the decoder takes the latent features and mask tokens to reconstructs the original tile. The loss function computes mean squared error (MSE) between the reconstructed and original tiles.

After MAE pre-training, we discard the decoder and apply the encoder to extract descriptive feature representations for WSI classification. Generally, only a small number of tiles within a WSI contains effective information for classification. Therefore, we take the output tokens features of the encoder as a bag of instances directly instead of applying a pooling operation. Further, not all instances are equally informative. Hence we propose a token selection model (TSM) that leverages the attention map from the last transformer block of the encoder as a distribution of the informative content to select tokens. Specifically, the attention map $A \in \mathbb{R}^{1\times N}$ can be defined as:

$$A = \frac{1}{N}\sum_{i=1}^{N}\left\{Softmax\left(\frac{QK^\top}{\sqrt{d}}\right)\right\}(i,j) \tag{1}$$

where $Q \in \mathbb{R}^{N\times d}$ is the query matrix and $K \in \mathbb{R}^{N\times d}$ is the key matrix from the last transformer block respectively, and N is the split number of tokens from the an input, i and j denote the row and column of the matrix after $softmax$. At the feature extraction stage, we select top-k indices in A and extract the corresponding k token features for further aggregation as a bag of instances. Then we use an attention network, similar to (Lu et al., 2021), as the slide-level classifier in our framework. It consists of several stacked fully connected layers, and based on its attention pooling function, the selected token features can be aggregated into slide-level classification.

3. Experiments and conclusion

We test our MAE-MIL framework on Camelyon16 dataset, a public lymph node dataset with and without breast cancer metastases. It consists of 270 training images and 129 testing images. The tiles cropped from WSIs are 1024×1024 size without overlapping. We follow the original MAE pipeline for the pre-training process, with each token feature embedded in a 1024-dimensional vector by the pre-trained encoder. In particular, in Fig. 2 (a), we show the image reconstruction, and Fig. 2 (b) shows the visualization of the attention maps from pre-trained encoder.

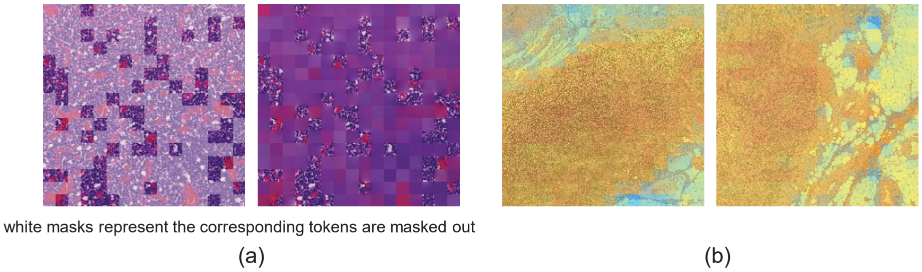
white masks represent the corresponding tokens are masked out
(a)
(b)

Figure 2: Visualization: (a) reconstruction of histopathological images (b) attention maps

As shown in Table 1, we compare MAE-MIL (top-16 selection) to the state-of-the-art CLAM approach, which the features are extracted from a CNN model pre-trained on ImageNet. It shows that the MAE pre-training process outperforms CNN pre-trained model. Moreover, we also show the result of our method without TSM but only average pooling operation over the tile-level features. It also demonstrates the effectiveness of our TSM.

Table 1: Results on Camelyon16 dataset with ×5 magnification

	Accuracy	AUC
CLAM	0.48	0.54
MAE-MIL (w/o TSM)	0.49	0.54
MAE-MIL	0.61	0.58

In conclusion, this paper provide a first insight into combining MAE and MIL on WSI classification task. Experimental results on the public WSI dataset demonstrate the effectiveness of pre-training process, which would be a promising paradigm for WSI analysis.

References

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.