Gigapixel Whole-Slide Image Classification Using Unsupervised Image Compression And Contrastive Training

David Tellez*, **Jeroen van der Laak, and Francesco Ciompi**
Diagnostic Image Analysis Group, Department of Pathology
Radboud University Medical Center, The Netherlands

Abstract

We propose a novel two-step methodology for entire whole-slide image (WSI) classification. First, all tissue patches in a WSI are mapped into vector embeddings using an encoder trained in an unsupervised fashion. The spatial arrangement of these embeddings is maintained with respect to the tissue patches, forming a stack of 2D feature maps representing the WSI. Second, a convolutional neural network is trained on these compact representations to predict weak labels associated with entire WSIs. We investigated several unsupervised schemes to train the encoder model: convolutional autoencoders (CAE), variational autoencoders (VAE), and a novel approach based on contrastive training. We validated the proposed methodology by predicting the existence of tumor metastasis at WSI-level using the Camelyon16 dataset. Our experimental results showed that the proposed methodology can be used to predict weak labels from entire WSIs. Furthermore, the novel contrastive encoder proved to be superior to the CAE and VAE approaches.

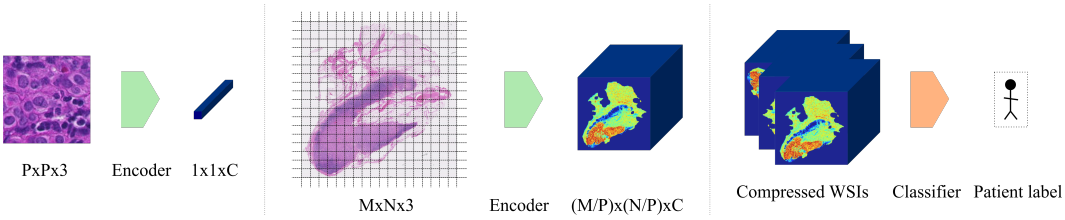

Figure 1: Overview of the proposed method to predict patient outcome from entire WSIs. Left: encoder mapping tissue patches into embedding vectors. Center: feature extraction applied in a sliding window fashion to an entire WSI. Right: training of a CNN-based classifier on compact representations of WSIs.

1 Introduction

Automatic analysis of entire whole-slide images (WSI) with convolutional neural networks (CNN) to predict patient outcome is currently an impossible task with the current GPU technology. Limiting factors include the large size of these images (typically 100,000 by 100,000 RGB pixels), and the global and weak nature of the labels (overall survival, molecular tests, cancer recurrence, etc.).

We propose a CNN-based method that can make predictions at whole-slide level by transforming gigapixel images into compact representations that fit in the GPU memory. We compress a WSI by sliding an encoder throughout the entire image, producing an embedding vector for each tissue patch,

*Corresponding author: david.tellezmartin@radboudumc.nl

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

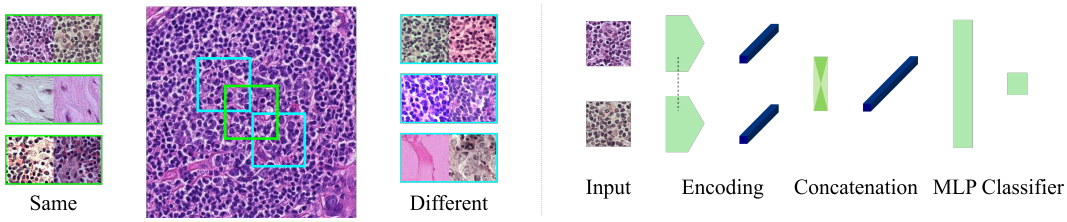

Figure 2: Training a feature extractor using an unsupervised contrastive approach. Left: building the contrastive dataset. In green, positive pairs correspond to the exact same tissue location (with different augmentation), whereas negative pairs correspond to neighbor non-overlapping tissue locations, in blue. Right: patches in a pair are encoded with the same network, the embeddings are concatenated, and an MLP is trained to distinguish between same or different tissue. After training, the encoder network alone is used as a feature extractor.

and conserving the spatial location of each vector with respect to each patch. Subsequently, these compact representations of WSIs, i.e. stacks of 2D feature maps, are used to train a regular CNN targeting image-level labels. The process is illustrated in Fig. 1.

2 Methods

The proposed method for WSI classification is divided into two steps. First, an encoder compresses gigapixel WSIs into compact representations. Second, a CNN-based classifier is trained on these compact representations to predict labels at WSI level.

2.1 Whole-slide image compression

We extracted relevant information from tissue images using a CNN-based encoder. This network mapped tissue patches into embedding vectors. We investigated the effectiveness of several types of encoders trained in an unsupervised manner, using tissue patches that were heavily augmented with the data augmentation routines detailed in [1].

First, we trained a convolutional autoencoder (CAE) by minimizing the mean squared error between the input patch and the reconstructed one. Once the training concluded, we used the encoding part of the CAE model as a feature extractor for the next step of the method.

Second, we trained a variational convolutional autoencoder (VAE) similarly as we did with the CAE model. In this case, we included a KL-divergence loss term and added a stochastic variational layer during training as indicated in [2].

Third, we proposed and trained a novel contrastive encoding scheme. We created an artificial training dataset consisting of pairs of tissue patches representing either the *same* or *different* tissue morphology. Positive pairs consisted of patches extracted from the exact same WSI location (although different augmentation). Negative pairs consisted of patches from: a) different WSI locations, and b) neighbor locations but non-overlapping tissue. A model composed of two encoders sharing weights, followed by a feature-wise concatenation operation and an MLP, was trained to distinguish between the two classes (see Fig. 2). Because two *same* patches present the same tissue morphology with heavily altered appearance, the encoder learns to extract high-level semantic features instead of low-level pixel ones, an advantage over encoders based on reconstruction error.

Finally, we trained a lower and an upper baseline encoders to provide a performance reference for the task at hand. For the lower baseline, we extracted the mean pixel intensity per color channel. For the upper baseline, we trained a fully supervised model to discern between patches with and without tumor cells and used an intermediate activation as the embedding.

2.2 Whole-slide image classification

We trained a CNN-based classifier on the set of compressed WSI representations obtained with the previous step. This classifier was optimized to predict a weak label associated with each WSI. We hypothesize that filters in the convolutional layers would learn to recognize spatial patterns in the input feature maps that are relevant to the label, integrating information from the entire WSI.

Figure 3: Evaluation on test set (AUC ROC)

Feature extractor	Patch	WSI	WSI-macro
Mean RGB (*)	0.855	0.513	0.596
VAE	0.897	0.530	0.613
CAE	0.914	0.584	0.653
Contrastive	0.899	0.607	0.745
Supervised (*)	0.969	0.672	0.956

(*) Baselines.

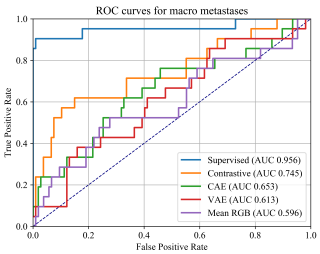

Figure 4: Classification performance for WSI-macro.

Due to the limited number of training samples, i.e. the number of patients, we took specific measures to alleviate overfitting. First, we encoded each WSI 8 times, using unique combinations of 90-degree rotation and flipping. Second, we took spatial crops of the compressed WSIs at random locations during training. Third, we used strided depthwise separable convolutions [3] to drastically reduce the number of trainable parameters while conserving the expressive capacity of the CNN.

3 Results

We used Camelyon16 data [4] to train and evaluate our methodology. We divided the set of slides into training (180), validation (90) and test (128). Each slide is associated with a binary label indicating the presence of tumor metastasis.

We trained instances of the five different encoders explained in Sec. 2.1 using a patch size of 128x128 px extracted at 0.5 um/px resolution, and compressed all WSIs in the dataset with each encoder. All CNN-based encoders shared the same architecture: five strided convolutional layers of 32, 64, 128, 256 and 512 3x3 filters, respectively, and stride of 2; followed by a dense layer of 512 units. All these layers included batch normalization and used leaky-ReLU as non-linearity. An additional linear dense layer produced the final embedding vector of size 128.

As a first experiment, we trained an MLP on top of each encoder (with frozen weights) to distinguish between healthy and tumor patches, serving as a qualitative metric of the effectiveness of each feature extractor. We measured the area under the ROC curve obtained with each model on 100,000 unseen patches extracted from the test slides without augmentation, see results in Tab. 3.

As a second experiment, we trained a CNN on each encoder's set of compressed WSI representations to predict the existence of tumor metastasis at whole-slide level using crops of 400x400. We measured the area under the ROC curve obtained on the test slides. Additionally, we measured the same metric when only WSIs with macro-metastasis were considered as positive samples, i.e. lesions larger than 2 mm, since metastases smaller than that were practically undetectable. See results in Tab. 3 and Fig. 4.

4 Conclusion

Our experimental results showed that the proposed methodology can be used to predict weak labels from entire WSIs. Furthermore, the novel contrastive encoder proved to be superior to the CAE and VAE approaches, suggesting that it was able to extract higher-level features than the other encoders. Future work could focus on investigating more advanced encoding schemes and embedding constraints to generate more similar embedding vectors for semantically similar images, e.g. involving concepts such as cosine similarity and content loss.

References

[1] David Tellez et al. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 2018.

[2] Diederik Kingma et al. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, 2016.

[4] Babak Ehteshami Bejnordi et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 2017.