ABSTRACT
Multiple Instance Learning (MIL) is a widely employed framework for learning on gigapixel whole-slide images (WSIs) from WSI-level annotations. In most MIL-based analytical pipelines for WSI-level analysis, the WSIs are divided into patches, and deep features for patches (i.e., patch embeddings) are extracted prior to training to reduce the overall computational cost and cope with the GPUs’ limited RAM. Because of this bottleneck, incorporating patch-level data augmentations during training adds an extra computational burden. To overcome this limitation, we present EMB AUGMENTER, a data augmentation generative adversarial network (DA-GAN) that can synthesize data augmentations in the embedding space rather than in the pixel space, thereby significantly reducing the computational requirements. Experiments on the SICAPv2 dataset show that our approach outperforms MIL without augmentation and is on par with traditional patch-level augmentation for MIL training while being substantially faster.

Index Terms—Computational Pathology, Data Augmentation, Generative Adversarial Networks.

1. INTRODUCTION
Computational pathology has made significant progress in recent years with new methods capable of classifying high-dimensional whole-slide images (WSI) of the order of 100,000 × 100,000 pixels [1, 2]. Most of these successes build on advances in weakly supervised learning, particularly on Multiple Instance Learning (MIL). MIL is an efficient way to aggregate information from all the patches that constitute the WSI into a slide-level representation that can further be used for classification. Specifically, MIL follows a two-step process: First, a pre-trained feature extractor processes each patch to derive patch embeddings, significantly reducing the dimensionality (typically from 256 × 256 patches to 1024-dimensional feature vectors). The feature extractor may be pre-trained on an auxiliary task [3] or based on self-supervised learning techniques [4, 5]. Given the large number of patches in each WSI (often > 10,000), this step is particularly computationally intensive. Then, in the second step, a neural network combines the low-dimensional patch embeddings to generate a slide-level representation used for classification, e.g., for cancer subtyping or grading.

Although MIL methods are widely employed, pre-extracting patch embeddings beforehand is often considered too com-
(a) Traditional Multiple Instance Learning (MIL) with data augmentation

(b) Training of EmbAugmenter

(c) MIL with EmbAugmenter

Fig. 2. MIL training using traditional data augmentation vs. EmbAugmenter: (a) In an off-line step, the WSI is divided into a set of patches, augmented using image processing transforms, and fed to a feature extractor to derive patch embeddings. Then, an attention mechanism aggregates the patch embeddings into a slide-level representation used for WSI classification. (b) Instead of doing pixel-space augmentation, we train a Data-Augmentation GAN to synthesize patch embeddings at a lower computational cost. (c) Augmented patch embeddings can be generated during MIL training to enhance data variability without extracting patch-level augmentations.

The computational demand and time-consuming to allow for data augmentation (a widespread computer vision technique used to reduce overfitting and improve generalization). Indeed, patch embeddings would have to be extracted as many times as a new augmentation is needed. Since patch embedding extraction is the most time-consuming part of training, the total cost quickly becomes very large [6]. Furthermore, in this scenario, there is a finite limit to the number of embeddings that can be extracted, which limits the range of augmentations that can be created. Instead, we propose to generate augmentations in the embedding space by synthesizing variations of the patch embeddings. In this way, we only need to extract patch embeddings once on the original images and augment them during training. Specifically, we employ a data augmentation generative adversarial network (DA-GAN) to learn the distribution of patch embeddings (see Figure 1). After training, the GAN generator can be re-instantiated to generate entirely new augmentations from the original embeddings.

Specifically, our contributions are: (1) We propose a novel GAN-based EmbAugmenter that learns to synthesize rich augmentations of patch embeddings at a minimal cost; (2) We show that MIL training with the EmbAugmenter outperforms MIL training without augmentation at a fraction of the computational cost of original augmentations and is comparable to traditional patch-level augmentation on the SICAPv2 dataset for ISUP grading of prostate biopsies.

2. METHOD

In this section, we present our approach for enabling embedding space augmentation during MIL training. An overview of the method is shown in Fig. 2.
2.1. Patch embedding augmentation

We first decompose an input WSI $X$ into a bag of $N$ patches, denoted as $X = \{x_1, \ldots, x_N\}$. We then employ a feature extractor $f(\cdot)$ to map each patch $x_i$ to a patch embedding as $h_i = f(x_i) \in \mathbb{R}^d$. Our goal is then to synthesize embeddings of patch augmentations. For this purpose, we introduce EMB AUGMENTER, a data augmentation GAN consisting of a generator denoted as $\mathcal{T}$ and a discriminator denoted as $D$. Given a patch embedding $h_i = f(x_i)$ and a randomly sampled latent vector $z \in \mathbb{R}^d$, we learn a network $\mathcal{T} : (z, h_i) \mapsto \tilde{h}_i$ that can synthesize patch embeddings resembling the true augmented patch embeddings. Similar to the traditional GAN training [7], the generator and discriminator play a min-max game, where the discriminator tries to identify real from fake samples while the generator iteratively learns better and better embeddings.

We propose two variants of the GAN generator and discriminator with different model expressivity. First, a model that assumes that the embedding factors are independent of each other, i.e., $\tilde{h}_i^{(j)} \perp \perp \tilde{h}_i^{(j')}$, $j \in \{1 : d\}$. The second variant models all-to-all interactions between the patch embedding factors and $z$. Formally, the generator $\mathcal{T}$ is expressed as,

$$\mathcal{T}_{\text{Exp}}(z, h_i) = \text{MLP}_{\text{Exp}}(z \parallel h_i) \quad (1)$$

$$\mathcal{T}_{\text{Ind}}(z, h_i) = \left[ \left| \left| d \right| \prod_{j=1}^d \text{MLP}_{\text{Ind}}(z^{(j)} \parallel h_i^{(j)}) \right| \right] \quad (2)$$

where $\parallel$ denotes the concatenation operation, MLP denotes a multi-layer perceptron, $\text{MLP}_{\text{Ind}}$ denotes the independent model (Ind), and $\text{MLP}_{\text{Exp}}$ denotes the expressive variant (Exp). The discriminator is defined analogously. The generator loss is composed of two terms: the cosine similarity between the true and fake patch embeddings and the discriminator binary cross-entropy (BCE). The discriminator is optimized with the BCE loss as in regular GAN training. Prior to GAN training, true patch embeddings are extracted using patch augmentations based on random rotation, color jittering, and zoom in/out. In essence, EMB AUGMENTER is similar to a Pix2Pix model [8] where the pixels would be replaced by the patch embedding factors.

2.2. Embedding space augmented MIL training

We now present how EMB AUGMENTER can be integrated into MIL training. Each WSI $X$ is associated with a label $y$ that we aim to predict. In this work, we employ an attention-based MIL model [9]. Specifically, after patch-level feature extraction, we increase the data variability by further augmenting the patch embeddings $h_i = \{h_1, \ldots, h_N\}$ using our proposed EMB AUGMENTER yielding to $h_i = \{\mathcal{T}(h_i)\}_{i=1, \ldots, N}$, where $\mathcal{T}(\cdot)$ is the GAN generator. In a second step, a neural network, denoted as $g(\cdot)$, combines the embeddings into a WSI-level embedding $h_{\text{WSI}} \in \mathbb{R}^{4\times 256}$, that is finally fed to a predictor, denoted as $c(\cdot)$, for classifying the WSI. These steps can be summarized as,

$$\hat{y} = c \left( g \left( \{ \mathcal{T}(f(x_1)), \ldots, \mathcal{T}(f(x_N)) \} \right) \right) \quad (3)$$

where $\hat{y}$ denotes the WSI prediction. The augmented patch embeddings are combined into a WSI representation using an attention mechanism as $g(\cdot) = \sum_{i=1}^N a_i h_i$, where $\{a_i\}_{i=1,N}$ are gated-attention weights [10, 9]. Using this approach, the time-consuming step of extracting features on pixel-augmented patches is replaced by an efficient embedding space augmentation. An overview of this process is depicted in Fig. 2(c).

3. EXPERIMENTAL RESULTS

We benchmark EMB AUGMENTER on the SICAPv2 dataset for ISUP grading of prostate biopsies (5-class problem). SICAPv2 comprises 155 WISIs of varying shapes and sizes processed at 10× magnification with an average of 129 256×256 patches per WSI. 5-fold cross-validation was employed with a 60%, 20%, 20% train, validation and test split. A different EMB AUGMENTER was trained for each fold to avoid transductive data leakage between training and testing. Each EMB AUGMENTER training used 48,552 $(h_i, \mathcal{T}(h_i))$ pairs. We benchmark EMB AUGMENTER in terms of classification performance – measured with the accuracy, quadratic kappa score ($\kappa^2$), and negative log-likelihood (NLL) – and computational time.

We compare EMB AUGMENTER against two baselines: First, a baseline where the MIL model is trained without augmentation (referred to as No augmentation in Table 1), and second a baseline that uses traditional pre-extracted patch-level augmentations (referred to as Patch augmentation). In order to ensure a fair comparison, we employed a consistent approach across all of our experiments. Specifically, we utilized five augmentations per patch for each experiment and employed identical model hyperparameters for all three approaches. The only difference between the approaches lay in their augmentation strategies. Additionally, we optimized all methods to determine the optimal learning rate and weight decay. The code for our experiments was implemented in PyTorch and optimized using the Adam algorithm.

The EMB AUGMENTER generator and discriminator are both based on MLPs: $\text{MLP}_{\text{Exp}}$ is a 6-layer encoder/decoder MLP with 256 bottleneck dimensions and $\text{MLP}_{\text{Ind}}$ is a 2-layer MLP with 4 hidden dimensions. The feature extractor $f(\cdot)$ is using ResNet50 [11] features pre-trained on ImageNet as proposed in [3, 12]. The attention in $g(\cdot)$ uses a gated mechanism with 2-layer MLPs to map the 1024 patch embedding dimensions to a single attention weight. The classifier $c(\cdot)$ uses a 2-layer MLP with 256 hidden dimensions.

Table 1 presents classification results on SICAPv2. Including augmentation during training (at both patch and em-
embedding levels) leads to a significant performance boost, e.g., absolute gain of +5.8% in accuracy with and without patch augmentation. While the embedding space features may appear to be independent to some extent (lower NLL loss than the baseline without augmentation), having a more expressive model that captures interactions between all the features further increases performance. Interestingly, MLP_{Exp., EMB AUGMENTER} leads to only slightly lower performance compared to traditional patch-level augmentation both in terms of classification performance and NLL. We hypothesize that while MIL training with embedding augmentations can be generated as often as needed during training time, they do not capture the entire spectrum of true patch augmentations. Despite this small decrease in performance, the computational benefits of the two proposed embedding space augmentation strategies are significant, with the cost of computing an augmentation reduced by over 300 times compared to traditional patch-level augmentation.

4. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we proposed a new technique for image data augmentation in the embedding space. The proposed EMB AUGMENTER is particularly valuable for training MIL methods that rely on pre-extracted patch embedding representations. With the EMB AUGMENTER, new augmentations can be generated during training at each epoch, thus increasing the variability of the data in an efficient manner. In the future, this method can be tested on larger datasets with thousands of WSIs. EMB AUGMENTER could also be conditioned by the augmentation type to enable more control over the synthesized augmentations.

5. REFERENCES

[1] Ming Y Lu et al., “AI-based pathology predicts origins for cancers of unknown primary,” Nature, 2021.
[2] Mohsin Bilal et al., “Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study,” The Lancet Digital Health, 2021.
[3] Ming Y Lu et al., “Data-efficient and weakly supervised computational pathology on whole-slide images,” Nature Biomedical Engineering, 2021.
[4] Navid Alemi Koohbanani et al., “Self-path: Self-supervision for classification of pathology images with limited annotations,” IEEE Transactions on Medical Imaging, 2021.
[5] Richard J Chen et al., “Scaling vision transformers to gigapixel images via hierarchical self-supervised learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[6] D. Tellez et al., “Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology,” in Medical Image Analysis, 2019.
[7] Ian Goodfellow et al., “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 2014.
[8] Phillip Isola et al., “Image-to-image translation with conditional adversarial networks,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[9] Maximilian Ilse et al., “Attention-based deep multiple instance learning,” in Proceedings of the 35th International Conference on Machine Learning, 2018.
[10] Dzmitry Bahdanau et al., “Neural machine translation by jointly learning to align and translate,” in International Conference on Learning Representations, 2015.
[11] Kaiming He et al., “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
[12] Zhuchen Shao et al., “Transmil: Transformer based correlated instance learning for whole slide image classification,” Advances in Neural Information Processing Systems, 2021.

6. ACKNOWLEDGMENTS

A large part of this work was conducted while IZ was visiting Harvard Medical School. NR is the founder, CSO, and Director of Histofy, a UK-based company.