ProtoMIL: Multiple Instance Learning with Prototypical Parts for Whole-Slide Image Classification

Dawid Rymarczyk, Adam Pardy, Jarosław Kraus, Aneta Kaczyńska, Marek Skomorowski, and Bartosz Zieliński

Abstract. The rapid development of histopathology scanners allowed the digital transformation of pathology. Current devices fastly and accurately digitize histology slides on many magnifications, resulting in whole slide images (WSI). However, direct application of supervised deep learning methods to WSI highest magnification is impossible due to hardware limitations. That is why WSI classification is usually analyzed using standard Multiple Instance Learning (MIL) approaches, that do not explain their predictions, which is crucial for medical applications. In this work, we fill this gap by introducing ProtoMIL, a novel self-explainable MIL method inspired by the case-based reasoning process that operates on visual prototypes. Thanks to incorporating prototypical features into objects description, ProtoMIL unprecedentedly joins the model accuracy and fine-grained interpretability, as confirmed by the experiments conducted on five recognized whole-slide image datasets.

Keywords: Multiple Instance Learning · Digital Pathology · Interpretable Deep Learning.

1 Introduction

A typical supervised learning scenario assumes that each data point has a separate label. However, in Whole Slide Image (WSI) classification, only one label is usually assigned to a gigapixel image due to the laborious and expensive labeling. Because of the hardware limitations, the direct application of supervised deep learning methods to WSI two highest magnification is impossible. That is why recent
Fig. 1: ProtoMIL divides the whole slide image into patches and analyzes their similarity to the reference prototypical parts that describe the given data class. As a result, it can provide a visual explanation of its prediction. One can observe that ProtoMIL identifies the most important patches with attention weights, that can appear both inside and outside a cancer region (marked as green and blue areas, respectively). Moreover, these patches are described by cancer or healthy tissue prototypes (corresponding to patches in green and red frames, respectively), showing their resemblance to the training examples.
similarities over instances. As a result, we obtain bag-level representation classified with an additional neural layer. This approach significantly differs from non-MIL approaches because it applies an aggregation layer and introduces a novel regularization technique that encourages the model to derive prototypes from the instances responsible for the positive label of a bag. The latter is a challenging problem because those instances are concealed and underrepresented. Lastly, the prototypical parts are pruned to characterize the data classes compactly. This results in detailed interpretation, where the most important patches according to attention weights are described using prototypes, as shown in Fig. 1.

To show the effectiveness of our model, we conduct experiments on five WSI datasets: Bisque Breast Cancer [16], Colon Cancer [41], Camelyon16 Breast Cancer [13], Lung cancer subtype identification TCGA-NSCLC [5] and Kidney cancer subtype classification [2]. Additionally, in the Supplementary Materials, we show the universal character of our model in different scenarios such as MNIST Bags [20] and Retinopathy Screening (Messidor dataset) [11]. The results we obtain are usually on par with the current state-of-the-art models. However, at the same time, we strongly enhance interpretation capabilities with prototypical parts obtained from the training set. We made our code publicly available at https://github.com/apardyl/ProtoMIL.

The main contributions of this work are as follows:

– Introducing the ProtoMIL method, which substantially improves the interpretability of existing MIL models by introducing case-based reasoning.
– Developing a training paradigm that encourages generating prototypical parts from the underrepresented instances responsible for the positive label of a bag.

The paper is organized as follows. In Section 2, we present recent advancements in Multiple Instance Learning and deep interpretable models. In Section 3, we define the MIL paradigms and introduce ProtoMIL. Finally, in Section 4, we present the results of conducted experiments, and Section 5 summarizes the work.

2 Related works

Our work focuses on classification of whole slide images which is described using Multiple Instance Learning (MIL) framework. Additionally, we develop an interpretable method which relates to eXplainable Artificial Intelligence (XAI). We briefly describe both fields in the following subsections.

2.1 Multiple instance learning

Before the deep learning era, models based on SVM, such as MI-SVM [3], were used for MIL problems. However, currently, MIL is addressed with numerous deep models. One of them, Deep MIML [14], introduces a sub-concept layer that is learned and then pooled to obtain a bag representation. Another example
is mi-Net [44], which pools predictions from single instances to derive a bag-level prediction. Other architectures adapted to MIL scenarios includes capsule networks [45], transformers [38] and graph neural networks [43]. Moreover, many works focus on the attention-based pooling operators, like AbMILP introduced in [20] that weights the instances embeddings to obtain a bag embedding. This idea was also extended by combining it with mi-Net [24], clustering similar instances [27], self-attention mechanism [34], and sharing classifier weights with pooling operator [39]. However, the above methods either do not contain an XAI component or only present the importance of the instances. Hence, our ProtoMIL is a step towards the explainability of the MIL methods.

2.2 Explainable artificial intelligence

There are two types of eXplainable Artificial Intelligence (XAI) approaches, post hoc and self-explaining methods [4]. Among many post hoc techniques, one can distinguish saliency maps showing pixel importance [32,36,37,40] or concept activation vectors representing internal network state with human-friendly concepts [9,17,21,46]. They are easy to use since they do not require any changes in the model architecture. However, their explanations may be unfaithful and fragile [1]. Therefore self-explainable models were introduced like Prototypical Part Network [8] with a layer of prototypes representing the activation patterns. A similar approach for hierarchically organized prototypes is presented in [18] to classify objects at every level of a predefined taxonomy. Moreover, some works concentrate on transforming prototypes from the latent space to data space [26] or focus on sharing prototypical parts between classes and finding semantic similarities [35]. Other works [28] build a decision tree with prototypical parts in the nodes or learn disease representative features within a dynamic area [22]. Nonetheless, to our best knowledge, no fine-grained self-explainable method, like ProtoMIL, exists for MIL problems.

3 ProtoMIL

Due to the large resolution of whole slide images, which should not be scaled down due to loss of information, we first divide an image into patches. However, we do not know which patches correspond to the given disease state. Therefore, this problem boils down to Multiple Instance Learning (MIL), where there is a bag of instances (in our case patches) and only one label for the whole bag. This bag is passed through the four modules of ProtoMIL (see Fig. 2): convolutional network \( f_{\text{conv}} \), prototype layer \( f_{\text{proto}} \), attention pooling \( a \), and fully connected last layer \( g \). Convolutional and prototype layers process single instances, whereas attention pooling and the last layer work on a bag level. More precisely, given a bag of patches \( X = \{x_1, \ldots, x_k\} \), each \( x \in X \) is forwarded through convolutional layers to obtain low-dimensional embeddings \( F = \{f_{\text{conv}}(x_1), \ldots, f_{\text{conv}}(x_k)\} \). As \( f_{\text{conv}}(x) \in H \times W \times D \), for the clarity of description, let \( Z_x = \{z_j \in f_{\text{conv}}(x) : \)
Fig. 2: ProtoMIL passes a bag of patches through four modules. First, convolutional layer $f_{conv}$ generates embeddings for each patch. Then, the prototype layer $f_{proto}$ calculates similarities between patches representations and its prototypes. The similarities are aggregated using the attention pooling $a$ to obtain the bag similarity scores classified using the last layer $g$. Notice that particular colors in vectors $h_i$ and $h_{bag}$ correspond to prototypes similarities.
z_j \in \mathbb{R}^D, j = 1..HW \}$. Then, the prototype layer computes vector $h$ of similarity scores [8] between each embedding $f_{conv}(x)$ and all prototypes $p \in P$ as

$$h = \left( g(Z_x, p) = \max_{z \in Z_x} \log \left( \frac{\|z-p\|}{\|z-p\|+\varepsilon} \right) \right)_{p \in P} \quad \text{for } \varepsilon > 0.$$ 

This results in a bag of similarity scores $H = \{h_1, \ldots, h_k\}$, which we pass to the attention pooling [20] to obtain a single similarity scores for the entire bag

$$h_{bag} = \sum_{i=1}^k a_i h_i,$$

where $a_i = \frac{\exp\{w^T(\tanh(V h_i^T) \odot \text{sigm}(U h_i^T))\}}{\sum_{j=1}^k \exp\{w^T(\tanh(V h_j^T) \odot \text{sigm}(U h_j^T))\}}$, (1)

$w \in \mathbb{R}^{L \times 1}, V \in \mathbb{R}^{L \times M},$ and $U \in \mathbb{R}^{L \times M}$ are parameters, tanh is the hyperbolic tangent, sigm is the sigmoid non-linearity and $\odot$ is an element-wise multiplication. Note that weights $a_i$ sum up to 1, and thus the formula is invariant to the size of the bag. Such representation is then sent to the last layer to obtain the predicted label $\hat{y} = g(h_{bag})$ as in [8].

**Regularization.** In MIL, the instances responsible for the positive label of a bag are underrepresented. Hence, training ProtoMIL without additional regularizations can result in a prototype layer with only prototypes of a negative class. That is why we introduce a novel regularization technique that encourages the model to derive positive prototypes. For this purpose, we introduce the loss function composed of three components

$$\mathcal{L}_{CE}(\hat{y}, y) + \lambda_1 \mathcal{L}_{Clst} + \lambda_2 \mathcal{L}_{Sep},$$

where $\hat{y}$ and $y$ denotes respectively the predicted and ground truth label of bag $X$, $\mathcal{L}_{CE}$ corresponds to cross-entropy loss, while

$$\mathcal{L}_{Clst} = \frac{1}{|X|} \sum_{x_i \in X} a_i \min_{p \in P^y} \min_{z \in Z_{x_i}} \|z-p\|_2^2,$$

$$\mathcal{L}_{Sep} = -\frac{1}{|X|} \sum_{x_i \in X} a_i \min_{p \in P^y} \min_{z \in Z_{x_i}} \|z-p\|_2^2,$$

where $P^y$ is a set of prototypes assigned to class $y$. Comparing to [8], components $\mathcal{L}_{Clst}$ and $\mathcal{L}_{Sep}$ additionally use $a_i$ from Equation 1. As a result, we encourage the model to create more prototypes corresponding to positive instances, which usually have higher $a_i$ values.

### 4 Experiments

We test our ProtoMIL approach on five datasets, for which we train the model from scratch in three steps: (i) warmup phase with training all layers except
the last one, (ii) prototype projection, (iii) and fine-tuning with fixed $f_{\text{conv}}$ and $f_{\text{proto}}$. Phases (ii) and (iii) are repeated several times to find the most optimal set of prototypes. All trainings use Adam optimizer for all layers with $\beta_1 = 0.99$, $\beta_2 = 0.999$, weight decay 0.001, and batch size 1. Additionally, we use an exponential learning rate scheduler for the warmup phase and a step scheduler for prototype training. All results are reported as an average of all runs with a standard error of the mean. In the subsequent subsections, we describe experiment details and results for each dataset.

Across all datasets we use convolutional block from ResNet-18 followed by two additional $1 \times 1$ convolutions as the convolutional layer $f_{\text{conv}}$. We use ReLU as the activation function for all convolutional layers except the last layer, for which we use the sigmoid activation function. The prototype layer stores prototypes shared across all bags, while the attention layer implements AbMILP. The last layer is used to classify the entire bag. Weights between similarity scores of prototypes corresponding class logit are initialized with 1, while other connections are set to $-0.5$ as in [8]. Together with the specific training procedure, such initialization results in a positive reasoning process (we rather say “this looks like that” instead of saying “this does not look like that”).

### 4.1 Bisque Breast Cancer and Colon Cancer datasets

#### Experiment details. We experiment on two histological datasets: Colon Cancer and Bisque Breast Cancer. The former contains 100 H&E images with 22,444 manually annotated nuclei of four different types: epithelial, inflammatory, fibroblast, and miscellaneous. To create bags of instances, we extract $27 \times 27$ nucleus-centered patches from each image, and the goal is to detect if the bag contains one or more epithelial cells, as colon cancer originates from them. On the other hand, the Bisque dataset consists of 58 H&E breast histology images of size $896 \times 768$, out of which 32 are benign, and 26 are malignant (contain at least one cancer cell). Each image is divided into $32 \times 32$ patches, resulting in 672 patches per image. Patches with at least 75% of the white pixels are discarded, resulting in 58 bags of various sizes.

We apply extensive data augmentation for both datasets, including random rotations, horizontal and vertical flipping, random staining augmentation, staining normalization, and instance normalization. We use ResNet-18 convolutional parts with the first layer modified to $3 \times 3$ convolution with stride 1 to match the size of smaller instances. We set the number of prototypes per class to 10 with a size of $128 \times 2 \times 2$. Warmup, fine-tuning, and end-to-end training take 60, 20, and 20 epochs, respectively. 10-fold cross-validation with 1 validation fold and 1 test fold is repeated 5 times.

#### Results. Table 4 presents our results compared to both traditional and attention-based MIL models. On the Bisque dataset, our model significantly outperforms all baseline models. However, due to the small size of the Colon Cancer dataset, ProtoMIL overfits, resulting in poorer AUC than attention-based models. Nev-
Table 1: Results for small histological datasets, where ProtoMIL significantly outperforms baseline methods on the Bisque dataset. However, it achieves worse results for the Colon Cancer dataset, probably due to its small size. Additionally, interpretability of the methods is noted and further discussed in Section 4.6. Notice that values for comparison indicated with * and ** comes from [20] and [34], respectively.

| Method              | Bisque Accuracy | Bisque AUC  | Colon Cancer Accuracy | Colon Cancer AUC | Inter. |
|---------------------|-----------------|-------------|------------------------|------------------|--------|
| instance+max*       | 61.4% ± 2.6%    | 0.612 ± 0.026 | 84.2% ± 2.1% | 0.914 ± 0.010 | +      |
| instance+mean*      | 67.2% ± 2.6%    | 0.719 ± 0.019 | 77.2% ± 1.2% | 0.866 ± 0.008 | -      |
| embedding+max*      | 60.7% ± 1.5%    | 0.650 ± 0.013 | 82.4% ± 1.5% | 0.918 ± 0.010 | -      |
| embedding+mean*     | 74.1% ± 2.3%    | 0.796 ± 0.012 | 86.0% ± 1.4% | 0.940 ± 0.010 | -      |
| AbMIL**             | 71.7% ± 2.7%    | 0.856 ± 0.022 | 88.4% ± 1.4% | 0.973 ± 0.007 | ++     |
| SA-AbMILP***        | 75.1% ± 2.4%    | 0.862 ± 0.022 | 90.8% ± 1.3% | 0.981 ± 0.007 | +++    |
| ProtoMIL (our)      | 76.7% ± 2.2%    | 0.886 ± 0.033 | 81.3% ± 1.9% | 0.932 ± 0.014 | +++    |

Nevertheless, in both cases, ProtoMIL provides finer explanations than all baseline models (see Fig. 8 and Supplementary Materials).

4.2 Camelyon16 dataset

**Experiment details.** The Camelyon16 dataset [13] consists of 399 whole-slide images of breast cancer samples, each labeled as normal or tumor. We create MIL bags by dividing each slide 20x resolution image into 224 × 224 patches, rejecting patches that contain more than 70% of background. This results in 399 bags with a mean of 8,871 patches and a standard deviation of 6,175. Moreover, 20 largest bags are truncated to 20,000 random patches to fit into the memory of a GPU. The positive patches are again highly imbalanced, as only less than 10% of patches contain tumor tissue.

Due to the size of the dataset, we preprocess all samples using a ResNet-18 without two last layers, pre-trained on various histopathological images using self-supervised learning from [10]. The resulting embeddings are fed into our model to replace the feature backbone net. ProtoMIL is trained for 50, 40, and 10 epochs in warmup, fine-tuning, and end-to-end training, respectively. The number of prototypes per class is limited to 5 with no data augmentation. The experiments are repeated 5 times with the original train-test split.

**Results.** We compare ProtoMIL to other state-of-the-art MIL techniques, including both traditional mean and max MIL pooling, RNN, attention-based MIL pooling, and transformer-based MIL pooling [38]. ProtoMIL performs on par in terms of accuracy and slightly outperforms other models on AUC metric (Table 2) while providing a better understanding of its decision process, as presented in Fig. 4 and Supplementary Materials.
Fig. 3: Similarity scores between five crucial instances of a bag (columns) and ten prototypical parts (rows) for a positive and negative bag (left and right side, respectively) from the Colon Cancer bags. Each prototypical part is represented by a part of the training image and three nearest training patches, and each instance is represented by the patch and the value of its attention weight \( a_i \). Moreover, each cell contains a similarity score and a heatmap corresponding to prototype activation. One can observe that instances of a negative bag usually activate prototypes of a negative class (four upper prototypes in red brackets), while the instances of positive bags mostly activate positive prototypes (four bottom prototypes in green brackets).

4.3 TCGA-NSCLC dataset

Experiment details. TCGA-NSCLC includes two subtype projects, i.e., Lung Squamous Cell Carcinoma (TCGA-LUSC) and Lung Adenocarcinoma (TCGA-LUAD), for a total of 956 diagnostic WSIs, including 504 LUAD slides from 478 cases and 512 LUSC slides from 478 cases. We create MIL bags using WSI Segmentation and Patching from [27] with default parameters, except patch-level parameter set to 1. Each slide image is cropped into a series of 224×224 patches. This results in 1,016 bags with a mean of 3,961 patches. We randomly split the data in the ratio of train:valid:test equal 60:15:25 and assure that there is no case overlap between the sets, and use the same ProtoMIL settings as in the Camelyon16 dataset are used. The results are reported for 4-fold cross-validation.

Results. Results for the TCGA-NSCLC dataset are presented in Table 2 alongside results of other state-of-the-art approaches from [38]. ProtoMIL performs slightly lower on the Area Under the ROC Curve (AUC) and accuracy metrics than the powerful transformer-based model TransMIL but still is competitive to other CNN-based approaches. However, the advantage of ProtoMIL is its capability to provide a detailed explanation of predictions as presented in Fig. 5 and Supplementary Materials.
Table 2: Our ProtoMIL achieves state-of-the-art results on the Camelyon16 dataset in terms of AUC metric, surpassing even the transformer-based architecture. Moreover, it is competitive on TCGA-NSCLC and slightly worse on TCGA-RCC, with a small drop of accuracy and AUC compared to TransMIL. Additionally, interpretability of the methods is noted and further discussed in Section 4.6. Notice that values for comparison marked with “*” and “**” are taken from [24] and [38], respectively.

| METHOD      | Camelyon16 | TCGA-NSCLC | TCGA-RCC | INTER. |
|-------------|------------|------------|----------|--------|
|             | ACCURACY   | AUC        | ACCURACY | AUC    |        |
| INSTANCE+MEAN* | 79.84%     | 0.762      | 72.82%   | 0.840  | -      |
| INSTANCE+MAX* | 82.95%     | 0.864      | 85.93%   | 0.946  | +      |
| MILRNN**     | 80.62%     | 0.807      | 86.19%   | 0.910  | -      |
| ABMILP**     | 84.50%     | 0.865      | 77.19%   | 0.865  | +      |
| DSMIL*       | 86.82%     | 0.894      | 80.58%   | 0.892  | ++     |
| CLAM-SB**    | 87.60%     | 0.881      | 81.80%   | 0.881  | +      |
| CLAM-MB**    | 83.72%     | 0.868      | 84.22%   | 0.937  | +      |
| TransMIL**   | 88.37%     | 0.931      | 88.35%   | 0.960  | ++     |
| ProtoMIL (ou)| 87.29%     | 0.935      | 81.66%   | 0.918  | ++     |

4.4 TCGA-RCC dataset

Experiment details. TCGA-RCC consists of three unbalanced classes: Kidney Chromophobe Renal Cell Carcinoma (TGCA-KICH, 111 slides from 99 cases), Kidney Renal Clear Cell Carcinoma (TCGA-KIRC, 489 slides from 483 cases), and Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP, 284 slides from 264 cases) for a total of 884 WSIs. We create MIL bags using WSI Segmentation and Paching from [27] with default parameters and a patch-level parameter set to 1. Each slide image is cropped into a series of 224×224 patches. This results in 884 bags with a mean of 4,309 patches. A separate model is trained for each class, and scores are averaged for all classes. Other experiment settings are identical as for TCGA-NSCLC described above.

Results. We compare ProtoMIL to other state-of-the-art MIL techniques, including both traditional mean and max MIL pooling, attention-based MIL pooling, and transformer-based MIL pooling [38]. ProtoMIL performs on par in terms of accuracy and AUC metric (Table 2) while providing a better understanding of its decision process, as presented in Supplementary Materials.

4.5 Pruning

Experiment details. We run prototype pruning experiments on all the datasets to remove not class-specific prototypical parts and check their influence on the model performance. For each of them, we use the model trained in the previously described experiments. As pruning parameters, we use \( k = 6 \) and \( l = 40\% \) and fine-tuned for 20 epochs. Details about pruning operation are described in the Supplementary Materials.
Fig. 4: Similarity scores between five crucial instances of a bag (columns) and eight prototypical parts (rows) for a negative bag from the Camelyon16 dataset. One can observe that ProtoMIL strongly activates only one prototype and focuses mainly on nuclei when analyzing the healthy parts of the tissue. Please refer to Fig. 8 for a detailed description of the visualization.

Table 3: The influence of ProtoMIL pruning on the accuracy and AUC score. One can notice that even though the pruning removes around 30% of the prototypes, it usually does not noticeably decrease the AUC and accuracy of the model.

| Dataset      | Before Pruning | After Pruning |
|--------------|----------------|---------------|
|              | Proto. # | Accuracy | AUC   | Proto. # | Accuracy | AUC   |
| Bisque       | 20±0     | 76.7%±2.2% | 0.886±0.031 | 13.6±0.25 | 73.0%±2.4% | 0.867±0.022 |
| Colon Cancer | 20±0     | 81.3%±1.9% | 0.932±0.014 | 15.69±0.34 | 81.8%±2.4% | 0.880±0.022 |
| Camelyon16   | 10±0     | 87.3%±1.2% | 0.935±0.007 | 6.4±0.24  | 85.9%±1.5% | 0.937±0.007 |
| TCGA-NSCLC   | 10±0     | 83.66%±1.6% | 0.918±0.003 | 7.6±1.2   | 81.1%±1.4% | 0.880±0.003 |
| TCGA-RCC     | 10±0     | 94.66%±1.0% | 0.988±0.009 | 6.2±1.2   | 91.5%±1.2% | 0.955±0.006 |

Results. The accuracy and AUC in respect to the number of prototypes before and after pruning are presented in Table 6. For all datasets, the number of prototypes after pruning has decreased around 30% on average. However, it does not result in a noticeable decrease in accuracy or AUC, except for Colon Cancer, where we observe a significant drop in AUC. Most probably, it is caused by the high visual resemblance of nuclei patches (especially between epithelial and miscellaneous) that after prototype projection may be very close to each other in the latent space.

4.6 Interpretability of MIL methods

Column Inter. in Tables 4, and 2 indicates how interpretable are the considered models. Instances and embeddings-based methods, except instance-max, are not interpretable, similarly to MILRNN, since they lose information about instances crucial for the prediction. On the other hand, the AbMILP [20] identifies crucial instances within a bag and can present the local explanation to the users. However,
other attention-based methods, such as SA-AbMIL \cite{34}, TransMIL \cite{38} and CLAMs \cite{27} perform additional operations, like self-attention, requiring more effort from the user to analyze the explanation. That is why those methods have been assigned with lower interpretability. Moreover, DS-MIL \cite{24} finds a decision boundary on the bag level and can produce a more detailed explanation than AbMILP, but only for a single prediction (local explanations). In contrast, the ProtoMIL can produce both local (see Figure 8) and global explanations (see Supplementary Materials).

5 Discussion and conclusions

In this work, we introduce Prototypical Multiple Instance Learning (ProtoMIL), a method for Whole Slide Image classification that incorporates a case-based reasoning process into the attention-based MIL setup. In contrast to existing MIL methods, ProtoMIL provides a fine-grained interpretation of its predictions. For this purpose, it uses a trainable set of prototypical parts correlated with data classes. The experiments on five datasets confirm that introducing fine-grained interpretability does not reduce the model’s effectiveness, which is still on par with the current state-of-the-art methodology. Moreover, the results can be presented to the user with a novel visualization technique.

The experiments show that ProtoMIL can be applied to a challenging problem like Whole-Slide Image classification. Therefore, in future works, we plan to generalize our method to multi-label scenarios and multimodal classification problems since WSI often comes with other medical data like CT and MRI.

5.1 Limitations

ProtoMIL limitations are inherited from the other prototype-based models, such as non-obvious prototype meaning. Ergo, prototype projection might still result
in uncertainty on which attributes it represents. However, there are methods mitigating these, e.g. explainer defined in [29].

5.2 Negative impact

Our solution is based on prototypical parts that are susceptible to different types of adversarial attacks such as [19]. That is why practitioners shall address this risk in a deployed system with ProtoMIL. What is more, it may be used in information war to disinform societies when prototypes are obtained with spoiled data or are shown without appropriate comment, especially in fields like medicine.

6 Acknowledgments

This work was founded by the POIR.04.04.00-00-14DE/18-00 project carried out within the Team-Net programme of the Foundation for Polish Science cofinanced by the European Union under the European Regional Development Fund. This research was funded by the National Science Centre, Poland (research grant no. 2021/41/B/ST6/01370). For the purpose of Open Access, the authors have applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in neural information processing systems. pp. 9505–9515 (2018)
2. Akin, O., Elnajjar, P., Heller, M., Jarosz, R., Erickson, B., Kirk, S., Filippini, J.: Radiology data from the cancer genome atlas kidney renal clear cell carcinoma [tcga-kirc] collection. The Cancer Imaging Archive (2016)
3. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in neural information processing systems. vol. 2, p. 7 (2002)
4. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012 (2019)
5. Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Benson, J.A., Zhang, W., Leung, A.N., et al.: A radiogenomic dataset of non-small cell lung cancer. Scientific data 5(1), 1–9 (2018)
6. Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y., Rudin, C.: Iiaia-bl: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. arXiv preprint arXiv:2103.12308 (2021)
7. Borowa, A., Rymarczyk, D., Ochońska, D., Brzuchczy-Wloch, M., Zieliński, B.: Classifying bacteria clones using attention-based deep multiple instance learning interpreted by persistence homology. In: International joint conference on neural networks (2021)
8. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: This looks like that: deep learning for interpretable image recognition. arXiv preprint arXiv:1806.10574 (2018)
9. Chen, Z., Bei, Y., Rudin, C.: Concept whitening for interpretable image recognition. Nature Machine Intelligence 2(12), 772–782 (2020)
10. Ciga, O., Martel, A.L., Xu, T.: Self supervised contrastive learning for digital histopathology. arXiv preprint arXiv:2011.13971 (2020)
11. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al.: Feedback on a publicly distributed image database: the messidor database. Image Analysis & Stereology 33(3), 231–234 (2014)
12. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence 89(1-2), 31–71 (1997)
13. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., , the CAMELYON16 Consortium: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318(22), 2199–2210 (12 2017). https://doi.org/10.1001/jama.2017.14585, https://doi.org/10.1001/jama.2017.14585
14. Feng, J., Zhou, Z.H.: Deep miml network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
15. Foulds, J., Frank, E.: A review of multi-instance learning assumptions. The Knowledge Engineering Review 25(1), 1–25 (2010)
16. Gelasca, E.D., Byun, J., Obara, B., Manjunath, B.: Evaluation and benchmark for biological image segmentation. In: 2008 15th IEEE International Conference on Image Processing. pp. 1816–1819. IEEE (2008)
17. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Advances in neural information processing systems. pp. 9277–9286 (2019)
18. Hase, P., Chen, C., Li, O., Rudin, C.: Interpretable image recognition with hierarchical prototypes. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 7, pp. 32–40 (2019)
19. Hoffmann, A., Fanconi, C., Rade, R., Kohler, J.: This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. arXiv preprint arXiv:2105.02968 (2021)
20. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
21. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
22. Kim, E., Kim, S., Seo, M., Yoon, S.: Xprotonet: Diagnosis in chest radiography with global and local explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15719–15728 (2021)
23. Kolodner, J.: Case-based reasoning. Morgan Kaufmann (2014)
24. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
25. Li, G., Li, C., Wu, G., Ji, D., Zhang, H.: Multi-view attention-guided multiple instance detection network for interpretable breast cancer histopathological image diagnosis. IEEE Access (2021)
26. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
27. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering 5(6), 555–570 (2021)
28. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14933–14943 (2021)
29. Nauta, M., Jutte, A., Provoost, J., Seifert, C.: This looks like that, because... explaining prototypes for interpretable image recognition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 441–456. Springer (2021)
30. Quellec, G., Lamard, M., Abr` amoff, M.D., Decenci` ere, E., Lay, B., Erginay, A., Cochener, B., Cazuguel, G.: A multiple-instance learning framework for diabetic retinopathy screening. Medical image analysis 16(6), 1228–1240 (2012)
31. Rani, P., Elagiri Ramalingam, R., Rajamani, K.T., Kandemir, M., Singh, D.: Multiple instance learning: Robust validation on retinopathy of prematurity. Int J Ctrl Theory Appl 9, 451–459 (2016)
32. Rebuffi, S.A., Fong, R., Ji, X., Vedaldi, A.: There and back again: Revisiting backpropagation saliency methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8839–8848 (2020)
33. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5), 206–215 (2019)
34. Rymarczyk, D., Borowa, A., Tabor, J., Zielinski, B.: Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1721–1730 (2021)
35. Rymarczyk, D., Struski, L., Tabor, J., Zielinski, B.: Protopshare: Prototype sharing for interpretable image classification and similarity discovery. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’21) (2021). https://doi.org/https://doi.org/10.1145/3447548.3467245
36. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 618–626 (2017)
37. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2591–2600 (2019)
38. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. arXiv preprint arXiv:2106.00908 (2021)
39. Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., Yang, L.: Loss-based attention for deep multiple instance learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5742–5749 (2020)
40. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 (2013)
41. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.W., Snead, D.R., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE transactions on medical imaging 35(5), 1196–1206 (2016)
42. Straehle, C., Kandemir, M., Koethe, U., Hamprecht, F.A.: Multiple instance learning with response-optimized random forests. In: 2014 22nd International Conference on Pattern Recognition. pp. 3768–3773. IEEE (2014)
43. Tu, M., Huang, J., He, X., Zhou, B.: Multiple instance learning with graph neural networks. arXiv preprint arXiv:1906.04881 (2019)
44. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition 74, 15–24 (2018)
45. Yan, Y., Wang, X., Guo, X., Fang, J., Liu, W., Huang, J.: Deep multi-instance learning with dynamic pooling. In: Asian Conference on Machine Learning. pp. 662–677. PMLR (2018)
46. Yeh, C.K., Kim, B., Arik, S.O., Li, C.L., Pfister, T., Ravikumar, P.: On completeness-aware concept-based explanations in deep neural networks. Advances in neural information processing systems (2019)
47. Zhao, Z., Fu, G., Liu, S., Elokely, K.M., Doerksen, R.J., Chen, Y., Wilkins, D.E.: Drug activity prediction using multiple-instance learning via joint instance and feature selection. In: BMC bioinformatics. vol. 14, p. S16. Springer (2013)

Supplementary Materials

In this Supplementary Materials, we present additional details on the ProtoMIL model and similarity scores visualizations with more instances and prototypes for all datasets considered in our experiments.

7 ProtoMIL

7.1 Prototypes projection.

Prototypes projection is an important step in the training procedure because it visualizes the prototypes using training patches. For this purpose, it replaces every learned prototype with the nearest training patch from the bag with the same label as the prototype class. The prototype \( \mathbf{p}^c \) of class \( c \) (negative or positive) can be replaced using the following formula

\[
\mathbf{p}^c \leftarrow \arg \min_{\mathbf{z} \in Z} \| \mathbf{z} - \mathbf{p}^c \|_2,
\]

where \( Z = \{ \mathbf{z} \in Z_x | x \in X \land y = c \} \) and \( y \) is a label of bag \( X \).

7.2 Pruning.

During the prototype projection, every prototype is replaced with the representation of the nearest training patch from the bag with the same label. Generally,
the representations of the nearest training patches correspond to the same label. However, in some cases, the nearest patches of a prototype correspond to more than one class. It is especially problematic in highly unbalanced datasets, frequently occurring in MIL tasks. To remove such misleading prototypes, we extend the prototype pruning algorithm from [8] to work in the MIL scenario. More precisely, we find $k$-nearest training patches for each prototype $p_c$ belonging to class $c$. If out of those $k$ patches less than $r$ belong to bags labeled with class $c$, we assume that this prototype is not determinant and remove it. Moreover, in contrast to [8], we automatically select $r$ to remove up to $l\%$ of prototypes ($l$ and $k$ are selected so that both classes still contain prototypes, and the drop in training accuracy is minimal). Finally, we fine-tuned attention and the final layers to compensate for the prototype removal.

8 Additional results

8.1 MNIST Bags

Experiment details. We experiment with the MNIST dataset, for which we generate the bags like proposed in [20]. Namely, a single bag contains grayscale images randomly sampled from the MNIST dataset. The bags’ sizes are chosen using a normal distribution with a mean of 100 and a standard deviation of 20. A bag is considered positive if it contains at least one image labeled as “9”. There are equal numbers of positive and negative bags. Notice that even though such dataset is class-balanced, it contains only 5% of images labeled as “9” (10% instances in the positive bags). We test ProtoMIL for different size of dataset (50, 100, 200, 300, 400, 500 bags). Every experiment is run with random 10-fold cross-validation and repeated five times with a different seed to obtain mean AUC as the evaluation metric. We train a model for 30, 20, and 10 epochs for warmup, fine-tuning, and end-to-end training, respectively. The number of prototypes per class is set to 10, with prototype size $64 \times 2 \times 2$ (determined experimentally).

Results. We compare our model to baseline MIL pooling methods from [20]. As shown, our ProtoMIL approach requires slightly more samples to achieve AUC scores competitive to the regular models (Figure 6). However, as presented in Figure 8a, it increases model interpretability by finding distinct parts of images and match them with intuitive positive and negative prototypes (see Figure 7).

We experiment on two histological datasets as out toy task: Colon Cancer and Bisque breast cancer. The former contains 100 H&E images with 22,444 manually annotated nuclei of four different types: epithelial, inflammatory, fibroblast, and miscellaneous. To create bags of instances, we extract $27 \times 27$ nucleus-centered patches from each image, and the goal is to detect if the bag contains one or more epithelial cells, as colon cancer originates from them. On the other hand, the Bisque dataset consists of 58 H&E breast histology images of size $896 \times 768$, out of which 32 are benign, and 26 are malignant (contain at least one cancer cell). Each image is divided into $32 \times 32$ patches, resulting in 672 patches per
Fig. 6: Results for ProtoMIL and baseline MIL approaches on the MNIST Bags dataset depending on the number of training bags (x axis) using the AUC metric (y axis). One can observe that ProtoMIL achieves state-of-the-art results with a larger number of samples.

Fig. 7: Sample positive and negative prototypes of ProtoMIL trained on the MNIST Bags dataset. Notice that the positive prototypes correspond to parts of “9” while the negative prototypes contain parts of the other digits (like “8” or “4”). It is expected because a bag is considered positive if it contains at least one image of “9”.

image. Patches with at least 75% of the white pixels are discarded, resulting in 58 bags of various sizes.

We apply extensive data augmentation for both datasets, including random rotations, horizontal and vertical flipping, random staining augmentation, staining normalization, and instance normalization. We use ResNet-18 convolutional parts with the first layer modified to $3 \times 3$ convolution with stride 1 to match the size of smaller instances. We set the number of prototypes per class to 10 with a size of $128 \times 2 \times 2$. Warmup, fine-tuning, and end-to-end training take 60, 20, and 20 epochs, respectively. 10-fold cross-validation with 1 validation fold and 1 test fold is repeated 5 times.

Results. Table 4 presents our results compared to both traditional and attention-based MIL models. On the Bisque dataset, our model significantly outperforms all baseline models. However, due to the small size of the Colon Cancer dataset, ProtoMIL overfits, resulting in poorer AUC than attention-based models. Nevertheless, in both cases, ProtoMIL provides finer explanations than all baseline models (see Figure 8b and Supplementary Materials).
Fig. 8: Similarity scores between five crucial instances of a bag (columns) and eight or ten prototypical parts (rows) for a positive and negative bag (left and right side, respectively) from the MNIST Bags (a) and Colon Cancer datasets (b). Each prototypical part is represented by a part of image and three nearest training patches, and each instance is represented by the image and the value of attention weight $a_i$. Moreover, each cell contains a similarity score and a heatmap corresponding to prototype activation. One can observe that instances of a negative bag usually activate negative prototypes (four upper prototypes in red brackets), while the instances of positive bags mostly activate positive prototypes (four bottom prototypes in green brackets).

| Method          | Colon Cancer  |
|-----------------|---------------|
|                 | Accuracy      | AUC            |
| instance+max*   | 84.2% ± 2.1%  | 0.914 ± 0.010  |
| instance+mean*  | 77.2% ± 1.2%  | 0.866 ± 0.008  |
| embedding+max*  | 82.4% ± 1.5%  | 0.918 ± 0.010  |
| embedding+mean* | 86.0% ± 1.4%  | 0.940 ± 0.010  |
| AbMILP*         | 88.4% ± 1.4%  | 0.973 ± 0.007  |
| SA-AbMILP***    | 90.8% ± 1.3%  | 0.981 ± 0.007  |
| ProtoMIL (our)  | 81.3% ± 1.9%  | 0.932 ± 0.014  |

Table 4: Results for Colon Cancer dataset. ProtoMIL achieves slightly worse results for the Colon Cancer dataset, probably due to its small size. Notice that values for comparison indicated with "*" and "***" comes from [20] and [34], respectively.
Table 5: Results for the Messidor dataset show that in terms of F-score, our ProtoMIL method is comparable with methods based on attention (AbMILP) or graph convolutions (MIL-GNN-ATT). Notice that values for comparison marked with '*' and '**' are taken from [43] and [34], respectively.

| Method            | Accuracy | F-score |
|-------------------|----------|---------|
| MI-SVM*           | 54.5%    | 0.70    |
| MI-SVM*           | 54.5%    | 0.71    |
| EMDD*             | 55.1%    | 0.69    |
| Citation k-NN*    | 62.8%    | 0.69    |
| MILBoost*         | 64.1%    | 0.66    |
| mi-Graph*         | 72.5%    | 0.75    |
| MIL-GNN-Att*      | 72.9%    | 0.75    |
| MIL-GNN-DP*       | 74.2%    | 0.77    |
| AbMILP**          | 74.5%    | 0.74    |
| SA-AbMILP**       | 75.2%    | 0.76    |
| LSA-AbMILP**      | 76.3%    | 0.77    |
| ProtoMIL (our)    | 70.0%    | 0.75    |

8.2 Messidor dataset

Experiment details. The Messidor dataset contains 1200 retinal images: 654 with a positive label (diabetic retinopathy) and 546 with a negative one. To create bags of instances, we crop overlapping patches of size 224 x 224 from each of 700 x 700 images, and patches with more than 70% black pixels are dropped as in [43]. Additionally, we apply extensive data augmentation, including random rotations, horizontal and vertical flipping, Gaussian noise, and patch normalization. We use ResNet-18 convolutional layers learned from scratch with 10 prototypes per class and prototype size of 1 x 1 x 128. Warmup, fine-tuning, and end-to-end training take 30, 20, and 10 epochs, respectively. We perform 10 fold cross-validation repeated two times as in [43].

Results. Results of ProtoMIL in the case of F-score are comparable with the ones achieved in [43] and [34] (see Table 5). However, the accuracy is significantly lower, most possibly due to the data class imbalance. Nevertheless, our model provides a fine-grained interpretation of its decision, as presented in Figure 9.

8.3 Additional pruning results

9 Additional visualizations
Fig. 9: Similarity scores between four crucial instances of a bag (columns) and four prototypical parts (rows) for a positive bag from the Messidor dataset. One can observe that ProtoMIL focuses on the disease factors, which are the brightest yellow spots on the image. Moreover, both positive and negative prototypes are activated since the retina with pathological changes still shows healthy features, such as veins. Please refer to Figure 8 for a detailed description of the visualization.

Table 6: The influence of ProtoMIL pruning on the accuracy and AUC score. One can notice that even though the pruning removes around 30% of the prototypes, it usually does not noticeably decrease the AUC and accuracy of the model.
Fig. 10: Similarity scores for a positive bag from MNIST Bags.
Fig. 11: Similarity scores for a negative bag from MNIST Bags.
Fig. 12: Similarity scores for a positive bag from Bisque dataset.
Fig. 13: Similarity scores for a negative bag from Bisque dataset.
Fig. 14: ProtoMIL analysis matrix for a positive example from Colon Cancer dataset.
Fig. 15: ProtoMIL analysis matrix for a negative example from Colon Cancer dataset.
patches in bag: 31, positive patches: 1, class label: 1

Fig. 16: Similarity scores for a positive bag from Messidor dataset.

Fig. 17: Similarity scores for a positive bag from Camelyon16 dataset.
Fig. 18: Similarity scores for a negative bag from Camelyon16 dataset.

Fig. 19: Similarity scores for a LUAD bag from TCGA-NSCLC dataset.

Fig. 20: Similarity scores for a LUSC bag from TCGA-NSCLC dataset.
Fig. 21: Similarity scores for a positive bag from TCGA RCC dataset.

Fig. 22: Similarity scores for a negative bag from TCGA RCC dataset.