Kernel Self-Attention for Weakly-supervised Image Classification using Deep Multiple Instance Learning

Dawid Rymarczyk [1,2,*], Adriana Borowa[1, 2,*], Jacek Tabor[1,**], and Bartosz Zieliński[1, 2,**]

[1]Faculty of Mathematics and Computer Science, Jagiellonian University, 6 Łojasiewicza Street, 30-348 Kraków, Poland,
[2]Ardigen SA, 76 Podole Street, 30-394 Kraków, Poland,

[*]{dawid.rymarczyk,ada.borowa}@student.uj.edu.pl
[**]{jacek.tabor,bartosz.zielinski}@uj.edu.pl

Abstract

Not all supervised learning problems are described by a pair of a fixed-size input tensor and a label. In some cases, especially in medical image analysis, a label corresponds to a bag of instances (e.g. image patches), and to classify such bag, aggregation of information from all of the instances is needed. There have been several attempts to create a model working with a bag of instances, however, they are assuming that there are no dependencies within the bag and the label is connected to at least one instance. In this work, we introduce Self-Attention Attention-based MIL Pooling (SA-AbMILP) aggregation operation to account for the dependencies between instances. We conduct several experiments on MNIST, histological, microbiological, and retinal databases to show that SA-AbMILP performs better than other models. Additionally, we investigate kernel variations of Self-Attention and their influence on the results.

1. Introduction

Classification methods typically assume that there exists a separate label for each example from a dataset. However, in many real-life applications, there exists only one label for a bag of instances because it is too laborious to label all of them separately. This type of problem, called Multiple Instance Learning (MIL) [7], assumes that there is only one label provided for the entire bag and that some of the instances associate to this label [9].

MIL problems are common in medical image analysis due to the vast resolution of images and the weakly-labeled small datasets [4, 21]. Among others, they appear in the whole slide-image classification of biopsies [2, 3, 34], classification of dementia based on brain MRI [28], or the diabetic retinopathy screening [22, 23]. They are also used in computer-aided drug design to identify which conformers are responsible for the molecule activity [26, 36].

Recently, Ilse *et al*. [13] introduced the Attention-based MIL Pooling (AbMILP), a trainable operator that aggregates information from multiple instances of a bag. It is based on a two-layered neural network with the attention weights, which allows finding essential instances. Since the publication, this mechanism was widely adopted in the medical image analysis [18, 20, 32], especially for the assessment of whole-slide images. However, the Attention-based MIL Pooling is significantly different from the Self-Attention (SA) mechanism [35]. It perfectly aggregates information from a varying number of instances, but it does not model dependencies between them. Additionally, the SA-AbMILP is distinct from other MIL approaches developed recently because it is not modeling the bag as a graph like in [30, 33], it is not using the pre-computed image descriptors as features as in [30], and it models the dependencies between instances in contrast to [8].

In this work, we introduce a method that combines self-attention with Attention-based MIL Pooling. It simultaneously catches the global dependencies between the instances in the bag (which are beneficial [37]) and aggregates them into a fixed-sized vector required for the successive layers of the network, which then can be used in regression, binary, and multi-class classification problems. Moreover, we investigate a broad spectrum of kernels replacing dot product when generating an attention map. According to the experiments' results, using our method with various kernels is beneficial compared to the baseline approach, especially in the case of more challenging MIL assumptions. Our code is publicly available at `https:`

//github.com/gmum/Kernel_SA-AbMILP.

2. Multiple Instance Learning

Multiple Instance Learning (MIL) is a variant of inductive machine learning belonging to the supervised learning paradigm [9]. In a typical supervised problem, a separate feature vector, e.g. ResNet representation after Global Max Pooling, exists for each sample: $\mathbf{x} = \mathbf{h}, \mathbf{h} \in \mathbb{R}^{L\times 1}$. In MIL, each example is represented by a bag of feature vectors of length L called instances: $\mathbf{x} = \{\mathbf{h_i}\}_{i=1}^{n}, \mathbf{h_i} \in \mathbb{R}^{L\times 1}$, and the bag is of variable size n. Moreover, in *standard MIL assumption*, label of the bag $\mathbf{y} \in \{0, 1\}$, each instance h_i has a hidden binary label $y_i \in \{0, 1\}$, and the bag is positive if at least one of its instances is positive:

$$\mathbf{y} = \begin{cases} 0, & \text{iff } \sum_{i=1}^{n} y_i = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

This standard assumption (considered by AbMILP) is stringent and hence does not fit to numerous real-world problems. As an example, let us consider the digestive track assessment using the NHI scoring system [19], where the score 2 is assigned to a biopsy if the neutrophils infiltrate more than 50% of crypts and there is no epithelium damage or ulceration. Such a task obviously requires more challenging types of MIL [9], which operate on many assumptions (below defined as concepts) and classes.

Let $\hat{C} \subseteq C$ be the set of required instance-level concepts, and let $p : X \times C \to K$ be the function that counts how often the concept $c \in C$ occurs in the bag $\mathbf{x} \in X$. Then, in *presence-based assumption*, the bag is positive if each concept occurs at least once:

$$\mathbf{y} = \begin{cases} 1, & \text{iff for each } c \in \hat{C} : p(\mathbf{x}, c) \geq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

In the case of *threshold-based assumptions*, the bag is positive if concept $c_i \in C$ occurs at least $t_i \in \mathbb{N}$ times:

$$\mathbf{y} = \begin{cases} 1, & \text{iff for each } c_i \in \hat{C} : p(\mathbf{x}, c_i) \geq t_i, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In this paper, we introduce methods suitable not only for the standard assumption (like it was in the case of AbMILP) but also for presence-based and threshold-based assumptions.

3. Methods

3.1. Attention-based Multiple Instance Learning Pooling

Attention-based MIL Pooling (AbMILP) [13] is a type of weighted average pooling, where the neural network determines the weights of instances. More formally, if the bag $\mathbf{x} = \{\mathbf{h_i}\}_{i=1}^{n}, \mathbf{h_i} \in \mathbb{R}^{L\times 1}$, then the output of the operator is defined as:

$$\mathbf{z} = \sum_{i=1}^{n} a_i \mathbf{h_i}, \text{ where } a_i = \frac{\exp\left(\mathbf{w}^T tanh(\mathbf{V}\mathbf{h_i})\right)}{\sum_j^N \exp\left(\mathbf{w}^T tanh(\mathbf{V}\mathbf{h_j})\right)}, \tag{4}$$

$\mathbf{w} \in \mathbb{R}^{M\times 1}$ and $\mathbf{V} \in \mathbb{R}^{M\times L}$ are trainable layers of neural networks, and the hyperbolic tangent prevents the exploding gradient. Moreover, the weights a_i sum up to 1 to wean from various sizes of the bags and the instances are in the random order within the bag to prevent the overfitting.

The most important limitation of AbMILP is the assumption that all instances of the bag are independent. To overcome this limitation, we extend it by introducing the Self-Attention (SA) mechanism [35] which models dependencies between instances of the bag.

3.2. Self-Attention in Multiple Instance Learning

The pipeline of our method, which applies Self-Attention into Attention-based MIL Pooling (SA-AbMILP), consists of four steps. First, the bag's images are passed through the Convolutional Neural Network (CNN) to obtain their representations. Those representations are used by the self-attention module (with dot product or the other kernels) to integrate dependencies of the instances into the process. Feature vectors with integrated dependencies are used as the input for the AbMILP module to obtain one fixed-sized vector for each bag. Such a vector can be passed to successive Fully-Connected (FC) layers of the network. The whole pipeline is presented in Fig. 1. In order to make this work self-contained, below, we describe self-attention and particular kernels.

Self-Attention (SA). SA is responsible for finding the dependencies between instances within one bag. Instance representation after SA is enriched with the knowledge from the entire bag, this is important for the detection of the number of instances of the same concept and their relation. SA transforms all the instances into two feature spaces of keys $\mathbf{k_i} = \mathbf{W_k}\mathbf{h_i}$ and queries $\mathbf{q_j} = \mathbf{W_q}\mathbf{h_j}$, and calculates:

$$\beta_{\mathbf{j},\mathbf{i}} = \frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{N} \exp\left(s_{ij}\right)}, \text{ where } s_{ij} = \langle \mathbf{k}(\mathbf{h_i}), \mathbf{q}(\mathbf{h_j}) \rangle, \tag{5}$$

to indicate the extent to which the model attends to the i^{th} instance when synthesizing the j^{th} one. The output of the attention layer is defined separately for each instance as:

$$\hat{\mathbf{h}}_\mathbf{j} = \gamma \mathbf{o_j} + \mathbf{h_j}, \text{ where } \mathbf{o_j} = \sum_{i=1}^{N} \beta_{j,i} \mathbf{W_v} \mathbf{h_i}, \tag{6}$$

$\mathbf{W_q}, \mathbf{W_k} \in \mathbb{R}^{\bar{L}\times L}$, $\mathbf{W_v} \in \mathbb{R}^{L\times L}$ are trainable layers, $\bar{L} = L/8$, and γ is a trainable scalar initialized to 0. Parameters $\bar{L}$ and γ were chosen based on the results presented in [35].

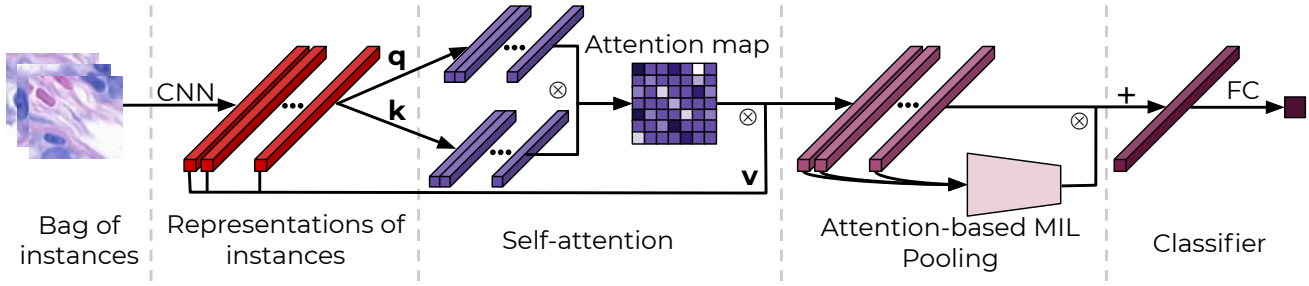

Figure 1: The pipeline of self-attention in deep MIL starts with obtaining feature space representation for each of the instances from the bag using features block of Convolutional Neural Network (CNN). In order to model dependencies between the instances, their representations pass trough the self-attention layer and then aggregate using AbMILP operator. The obtained fixed-size vector goes trough the Fully Connected (FC) classification layer.

Kernels in self-attention. In order to indicate to which extent one instance attends on synthesizing the other one, the self-attention mechanism typically employs a dot product (see s_{ij} in Eq. 5). However, dot product can be replaced by a various kernel with positive results observed in Support Vectors Machine (SVM) [1] or Convolutional Neural Networks (CNN) [31], especially in the case of small training sets.

The Radial Basis Function (RBF) and Laplace kernels were already successfully adopted to self-attention [14, 29]. Hence, in this study, we additionally extend our approach with the following standard choice of kernels (with α as a trainable parameter):

- Radial Basis Function (GSA-AbMILP): $k(x, y) = \exp\left(-\alpha \|x - y\|_2^2\right)$,
- Inverse quadratic (IQSA-AbMILP): $k(x, y) = \frac{1}{\alpha \|x-y\|_2^2 + 1}$,
- Laplace (LSA-AbMILP): $k(x, y) = -\|x - y\|_1$,
- Module (MSA-AbMILP): $k(x, y) = \|x - y\|^\alpha - \|x\|^\alpha - \|y\|^\alpha$.

We decided to limit to those kernels because they are complementary regarding the shape of tails in their distributions.

4. Experiments

We adopt five datasets, from which four were adopted to MIL algorithms [13, 30], to investigate the performance of our method: MNIST [17], two histological databases of colon [25] and breast [10] cancer, a microbiological dataset DIFaS [38], and a diabetic retinopathy screening data set called “Messidor” [5]. For MNIST, we adapt LeNet5 [17] architecture, for both histological datasets SC-CNN [25] is applied as it was in [13], for microbiological dataset we use convolutional parts of ResNet-18 [12] or AlexNet [16] followed by 1×1 convolution as those were the best feature extractor in [38], and for the "Messidor" dataset we used the ResNet-18 [12] as most of the approaches that we compare here are based on the handcrafted image features like in [30]. The experiments, for MNiST, histological and "Messidor" datasets, are repeated 5 times using 10 fold cross-validation with 1 validation fold and 1 test fold. In the case of the microbiological dataset, we use the original 2 fold cross-validation which divides the images by the preparation, as using images from the same preparation in both training and test set can result in overstated accuracy [38]. Due to the dataset complexity, we use the early stopping mechanism with different windows: 5, 25, 50 and 70 epochs for MNIST, histological datasets, microbiological, and "Messidor" datasets, respectively. We compare the performance of our method (SA-AbMILP) and its kernel variations with instance-level approaches (instance+max, instance+mean, and instance+voting), embedding-level approaches (embedding+max and embedding+mean), and Attention-based MIL Pooling, AbMILP [13]. The instance-level approaches in the case of MNIST and histological database compute the maximum or mean value of the instance scores. For the microbiological database, instance scores are aggregated by voting due to multiclassification. The embedding-level approaches calculate the maximum or mean for feature vector of the instances. For "Messidor" dataset we are comparing to the results obtained by [30, 8]. We run a Wilcoxon signed-rank test on the results to identify which ones significantly differ from each other, and which ones do not (and thus can be considered equally good). The comparison is performed between the best method (the one with the best mean score) and all the other methods, for each experiment separately. The mean accuracy is obtained as average over 5 repetitions with the same train/test divisions used by all compared methods. The number of repetitions is relatively small for statistical tests. Therefore we set the p-value to 0.1. For computations, we use Nvidia GeForce RTX 2080.

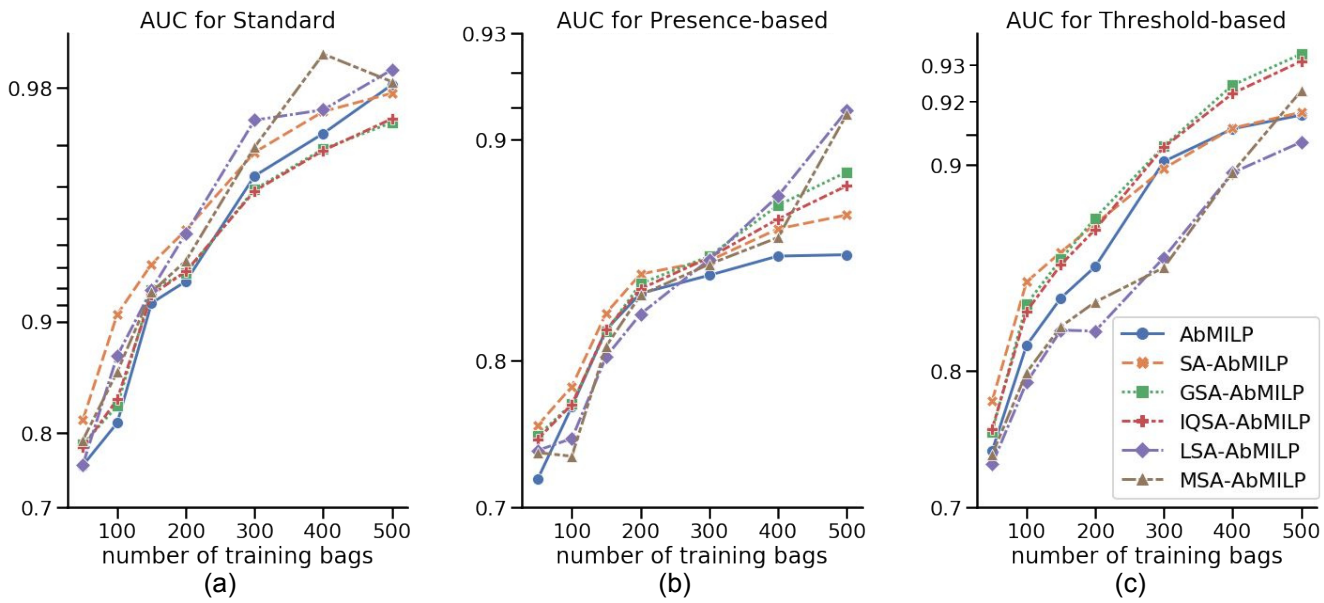

Figure 2: Results for MNIST dataset with bags generated using standard (a), presence-based (b), and threshold-based (c) assumption. In all cases, our approach, either with dot product (SA-AbMILP) or the other kernels (GSA-AbMILP, IQSA-AbMILP, LSA-AbMILP, and MSA-AbMILP) obtains statistically better results than the baseline method (AbMILP). See Section 3 for description of the shortcuts.

4.1. MNIST dataset

Experiment details. As in [13], we first construct various types of bags based on the MNIST dataset. Each bag contains a random number of MNIST images (drawn from Gaussian distributions $\mathcal{N}(10, 2)$). We adopt three types of bag labels referring to three types of MIL assumptions:

- Standard assumptions: $\mathbf{y} = 1$ if there is at least one occurrence of "9",
- Presence-based assumptions: $\mathbf{y} = 1$ if there is at least one occurrence of "9" and at least one occurrence of "7",
- Threshold-based assumptions: $\mathbf{y} = 1$ if there are at least two occurrences of "9".

We decided to use "9" and "7" because they are often confused with each other, making the task more challenging.

We investigated how the performance of the model depends on the number of bags used in training (we consider 50, 100, 150, 200, 300, 400, and 500 training bags). For all experiments, we use LeNet5 [17] initialized according to [11] with the bias set to 0. We use Adam optimizer [15] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, learning rate 10^{-5}, and batch size 1.

Results. AUC values for considered MIL assumptions are visualized in Fig. 2. One can observe that our method with a dot product (SA-AbMILP) always outperforms other methods in case of small datasets. However, when the number of training examples reaches 300, its kernel extensions work better (Laplace in presence-based and inverse quadratic in threshold-based assumption). Hence, we conclude that for small datasets, no kernel extensions should be applied, while in the case of a larger dataset, a kernel should be optimized together with the other hyperparameters. Additionally, we analyze differences between the weights of instances in AbMILP and our method. As presented in Fig. 3, for our method, "9"s and "7"s strengthen each other in the self-attention module, resulting in higher weights in the aggregation operator than for AbMILP, which returns high weight for only one digit (either "9" or "7").

4.2. Histological datasets

Experiment details. In the second experiment, we consider two histological datasets of *breast* and *colon* cancer (described below). For both of them, we generate instance representations using SC-CNN [25] initialized according to [11] with the bias set to 0. We use Adam [15] optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, learning rate 10^{-4}, and batch size 1. We also apply extensive data augmentation, including random rotations, horizontal and vertical flipping, random staining augmentation [13], staining normalization [27], and instance normalization.

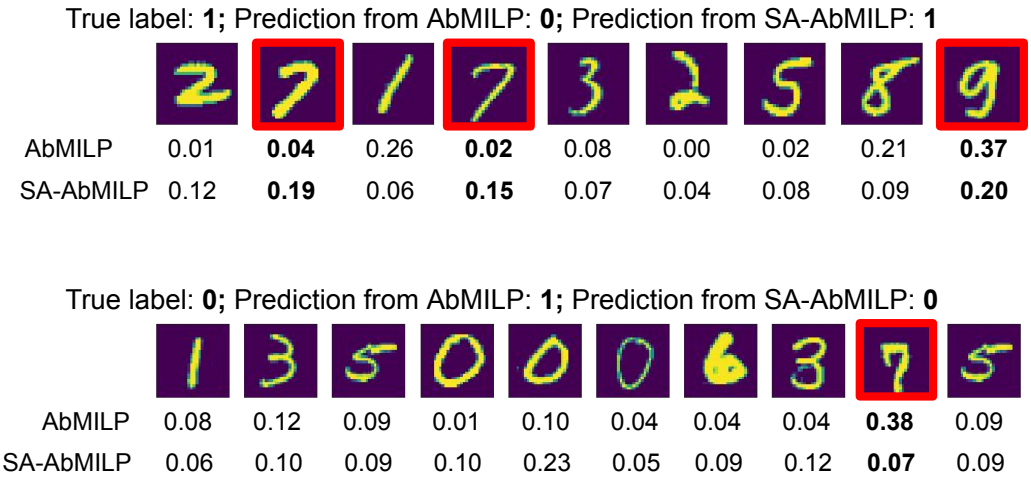

Figure 3: Example of instances' weights for a positive (top) and negative (bottom) bag in a presence-based assumption (where positive is at least one occurrence of "9" and "7") for AbMILP and our method. One can observe that for SA-AbMILP, "9"s and "7"s strengthen each other in the self-attention module, resulting in higher weights in the aggregation operator than for AbMILP.

Breast cancer dataset. Dataset from [10] contains 58 weakly labeled H&E biopsy images of resolution 896×768. The image is labeled as malignant if it contains at least one cancer cell. Otherwise, it is labeled as benign. Each image is divided into patches of resolution 32×32, resulting in 672 patches per image. Patches with at least 75% of the white pixels are discarded, generating 58 bags of various sizes.

Colon cancer dataset. Dataset from [25] contains 100 images with 22444 nuclei manually assigned to one of the following classes: epithelial, inflammatory, fibroblast, and miscellaneous. We construct bags of 27×27 patches with centers located in the middle of the nuclei. The bag has a positive label if there is at least one epithelium nucleus in the bag. Tagging epithelium nuclei is essential in the case of colon cancer because the disease originates from them [24].

Results. Results for histological datasets are presented in Table 1. For both of them, our method (with or without kernel extension) improves the Area Under the ROC Curve (AUC) comparing to the baseline methods. Moreover, our method obtains the highest recall, which is of importance for reducing the number of false negatives. To explain why our method surpasses the AbMILP, we compare the weights of patches in the average pooling. Those patches contribute the most to the final score and should be investigated by the pathologists. One can observe in Fig. 4 that our method highlights fewer patches than AbMILP, which simplifies their analysis. Additionally, SA dependencies obtained for the most relevant patch of our method are justified histologically, as they mostly focus on nuclei located in the neighborhood of crypts. Moreover, in the case of the colon cancer dataset, we further observe the positive aspect of our method, as it strengthens epithelium nuclei and weakens nuclei in the lamina propria at the same time,. Finally, we notice that kernels often improve overall performance but none of them is significantly superior.

4.3. Microbiological dataset

Experiment details. In the final experiment, we consider the microbiological DIFaS database [38] of *fungi species*. It contains 180 images for 9 fungi species (there are 2 preparations with 10 images for each species and it is a multi-class classification). As the size of images is $5760 \times 3600 \times 3$ pixels, it is difficult to process the entire image through the network so we generate patches following method used in [38]. Hence, unlike the pipeline from Section 3.2, we use two separate networks. The first network generates the representations of the instances, while the second network (consisting of self-attention, attention-based MIL pooling, and classifier) uses those representations to recognize fungi species. Due to the separation, it is possible to use deep architectures like ResNet-18 [12] and AlexNet [16] pre-trained on ImageNet database [6] to generate the representations. The second network is trained using Adam [15] optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, learning rate 10^{-5},

Table 1: Results for breast and colon cancer datasets (mean and standard error of the mean over 5 repetitions). See Section 3 for description of the acronyms.

breast cancer dataset					
method	accuracy	precision	recall	F-score	AUC
instance+max	61.4 ± 2.0	58.5 ± 3.0	47.7 ± 8.7	50.6 ± 5.4	61.2 ± 2.6
instance+mean	67.2 ± 2.6	67.2 ± 3.4	51.5 ± 5.6	57.7 ± 4.9	71.9 ± 1.9
embedding+max	60.7 ± 1.5	55.8 ± 1.3	54.6 ± 7.0	54.3 ± 4.2	65.0 ± 1.3
embedding+mean	$\mathbf{74.1 \pm 2.3}$	74.1 ± 2.3	65.4 ± 5.4	$\mathbf{68.9 \pm 3.4}$	79.6 ± 1.2
AbMILP	71.7 ± 2.7	$\mathbf{77.1 \pm 4.1}$	68.6 ± 3.9	66.5 ± 3.1	85.6 ± 2.2
SA-AbMILP	$\mathbf{75.1 \pm 2.4}$	$\mathbf{77.4 \pm 3.7}$	74.9 ± 3.7	$\mathbf{69.9 \pm 3.0}$	$\mathbf{86.2 \pm 2.2}$
GSA-AbMILP	$\mathbf{75.8 \pm 2.2}$	$\mathbf{79.3 \pm 3.3}$	74.7 ± 3.4	$\mathbf{72.5 \pm 2.5}$	$\mathbf{85.9 \pm 2.2}$
IQSA-AbMILP	$\mathbf{76.7 \pm 2.3}$	$\mathbf{78.6 \pm 4.2}$	75.1 ± 4.2	$\mathbf{66.6 \pm 4.3}$	$\mathbf{85.9 \pm 2.2}$
LSA-AbMILP	65.5 ± 2.9	62.5 ± 3.7	$\mathbf{89.5 \pm 2.6}$	$\mathbf{68.5 \pm 2.6}$	$\mathbf{86.7 \pm 2.1}$
MSA-AbMILP	$\mathbf{73.8 \pm 2.6}$	$\mathbf{78.4 \pm 3.9}$	73.8 ± 3.6	$\mathbf{69.4 \pm 3.4}$	85.8 ± 2.2
colon cancer dataset					
method	accuracy	precision	recall	F-score	AUC
instance+max	84.2 ± 2.1	86.6 ± 1.7	81.6 ± 3.1	83.9 ± 2.3	91.4 ± 1.0
instance+mean	77.2 ± 1.2	82.1 ± 1.1	71.0 ± 3.1	75.9 ± 1.7	86.6 ± 0.8
embedding+max	82.4 ± 1.5	88.4 ± 1.4	75.3 ± 2.0	81.3 ± 1.7	91.8 ± 1.0
embedding+mean	86.0 ± 1.4	81.1 ± 1.1	80.4 ± 2.7	85.3 ± 1.6	94.0 ± 1.0
AbMILP	88.4 ± 1.4	$\mathbf{95.3 \pm 1.5}$	$\mathbf{84.1 \pm 2.9}$	$\mathbf{87.2 \pm 2.1}$	97.3 ± 0.7
SA-AbMILP	$\mathbf{90.8 \pm 1.3}$	$\mathbf{93.8 \pm 2.0}$	$\mathbf{87.2 \pm 2.4}$	$\mathbf{89.0 \pm 1.9}$	$\mathbf{98.1 \pm 0.7}$
GSA-AbMILP	88.4 ± 1.7	$\mathbf{95.2 \pm 1.7}$	83.7 ± 2.8	$\mathbf{86.9 \pm 2.1}$	$\mathbf{98.5 \pm 0.6}$
IQSA-AbMILP	89.0 ± 1.9	$\mathbf{93.9 \pm 2.1}$	$\mathbf{85.5 \pm 3.0}$	$\mathbf{86.9 \pm 2.5}$	96.6 ± 1.1
LSA-AbMILP	84.8 ± 1.8	$\mathbf{92.7 \pm 2.7}$	71.1 ± 4.6	73.4 ± 4.3	95.5 ± 1.7
MSA-AbMILP	$\mathbf{89.6 \pm 1.6}$	$\mathbf{94.6 \pm 1.5}$	$\mathbf{85.7 \pm 2.7}$	$\mathbf{87.9 \pm 1.8}$	$\mathbf{98.4 \pm 0.5}$

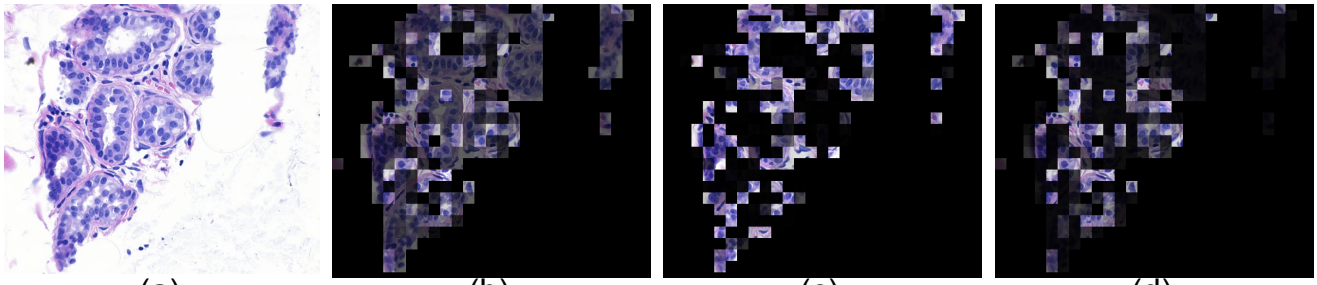

Figure 4: An example image from the breast cancer dataset (a), weights of patches obtained by AbMILP (b) and SA-AbMILP (c), and SA dependencies obtained for the most relevant patch in SA-AbMILP (d). One can observe that SA-AbMILP highlights fewer patches than AbMILP, which simplifies their analysis. Additionally, SA dependencies are justified histologically, as they mostly focus on nuclei located in the neighborhood of crypts.

and batch size 1. We also apply extensive data augmentation, including random rotations, horizontal and vertical flipping, Gaussian noise, and patch normalization. To increase the training set, each iteration randomly selects a different subset of image's patches as an input.

Results. Results for DIFaS database are presented in Table 2. Our method improves almost all of the scores for both feature pooling networks. The exception is the precision for ResNet-18, where our method is on par with maximum and mean instance representation. Moreover, we observe that while non-standard kernels can improve results for representations obtained with ResNet-18, they do not

Table 2: Results for DIFaS dataset (mean and standard error of the mean over 5 repetitions). See Section 3 for description of the shortcuts.

DIFaS (ResNet-18)					
method	accuracy	precision	recall	F-score	AUC
instance+voting	78.3 ± 2.0	78.0 ± 1.4	76.0 ± 1.7	75.8 ± 2.0	N/A
embedding+max	77.1 ± 0.7	$\mathbf{83.1 \pm 0.5}$	77.1 ± 0.7	75.5 ± 0.9	95.3 ± 0.2
embedding+mean	78.1 ± 0.8	$\mathbf{83.3 \pm 0.5}$	78.1 ± 0.8	76.4 ± 1.0	95.2 ± 0.2
AbMILP	77.5 ± 0.6	82.6 ± 0.5	77.5 ± 0.6	75.6 ± 0.8	96.1 ± 0.3
SA-AbMILP	$\mathbf{80.1 \pm 0.6}$	$\mathbf{84.6 \pm 0.6}$	$\mathbf{80.1 \pm 0.6}$	$\mathbf{78.4 \pm 0.8}$	$\mathbf{96.8 \pm 0.3}$
GSA-AbMILP	$\mathbf{79.1 \pm 0.4}$	83.5 ± 0.7	$\mathbf{79.1 \pm 0.4}$	77.2 ± 0.5	$\mathbf{97.0 \pm 0.3}$
IQSA-AbMILP	$\mathbf{79.4 \pm 0.4}$	$\mathbf{83.7 \pm 0.7}$	$\mathbf{79.4 \pm 0.4}$	$\mathbf{77.6 \pm 0.6}$	96.8 ± 0.3
LSA-AbMILP	77.6 ± 0.5	82.5 ± 0.5	77.6 ± 0.5	75.7 ± 0.7	96.2 ± 0.3
MSA-AbMILP	$\mathbf{79.2 \pm 0.4}$	83.5 ± 0.4	$\mathbf{79.2 \pm 0.4}$	$\mathbf{77.5 \pm 0.6}$	$\mathbf{96.9 \pm 0.3}$
DIFaS (AlexNet)					
method	accuracy	precision	recall	F-score	AUC
instance+voting	77.3 ± 1.9	78.4 ± 0.8	76.6 ± 1.2	76.2 ± 1.0	N/A
embedding+max	82.9 ± 1.2	87.1 ± 0.9	82.9 ± 1.2	82.3 ± 1.4	98.4 ± 0.2
embedding+mean	82.3 ± 0.7	87.2 ± 0.4	82.3 ± 0.7	81.5 ± 1.0	98.1 ± 0.3
AbMILP	83.6 ± 1.2	87.8 ± 0.7	83.6 ± 1.2	82.9 ± 1.5	98.6 ± 0.2
SA-AbMILP	$\mathbf{86.0 \pm 1.0}$	$\mathbf{89.6 \pm 0.6}$	$\mathbf{86.0 \pm 1.0}$	$\mathbf{85.7 \pm 1.3}$	$\mathbf{98.9 \pm 0.1}$
GSA-AbMILP	84.6 ± 1.0	89.1 ± 0.6	84.6 ± 1.0	84.2 ± 1.3	98.8 ± 0.2
IQSA-AbMILP	84.1 ± 1.2	88.4 ± 0.6	84.1 ± 1.2	83.4 ± 1.4	$\mathbf{98.9 \pm 0.2}$
LSA-AbMILP	83.9 ± 1.3	88.0 ± 0.7	83.9 ± 1.3	83.2 ± 1.6	98.6 ± 0.2
MSA-AbMILP	83.1 ± 0.8	87.8 ± 0.4	83.1 ± 0.8	82.4 ± 1.0	98.7 ± 0.2

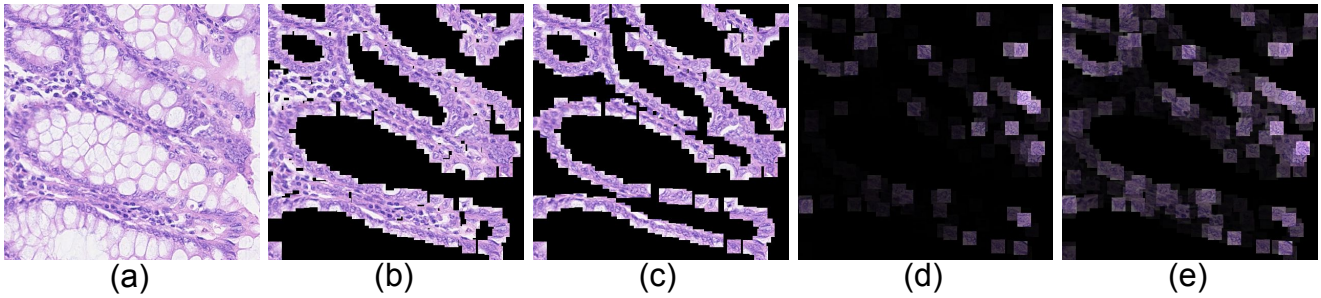

Figure 5: An example image from the colon cancer dataset (a) with annotated nuclei (b) and epithelium nuclei (c), as well as, the weights of patches obtained by AbMILP (d) and SA-AbMILP (e). One can observe that SA-AbMILP strengthens epithelium nuclei and, at the same time, weakens nuclei in the lamina propria.

operate well with those generated with AlexNet. Additionally, to interpret the model outputs, we visualize patches with the lowest and highest weights in the average pooling. As shown in Figure 6, our method properly [38] assigns lower weights to blurred patches with a small number of cells. Also, in contrast to the baseline method, it assigns high weight to clean patches (without red artifacts).

4.4. Retinal image screening dataset.

Experiment details. Dataset "Messidor" from [5] contains 1200 images with 654 positive (diagnosed with diabetes) and 546 negative (healthy) images. The size of each image is 700×700 pixels. Each image is partitioned into patches of 224×224 pixels. Patches containing only background are dropped. We are using ResNet18 [12] pretrained on the ImageNet [6] as an instance feature vectors generator and SA-AbMILP to obtain the final prediction, which is trained in an end to end fashion. The model is trained using Adam [15] optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, learning rate $5 * 10^{-6}$, and batch size 1. We also apply data augmentation as in Section 4.3.

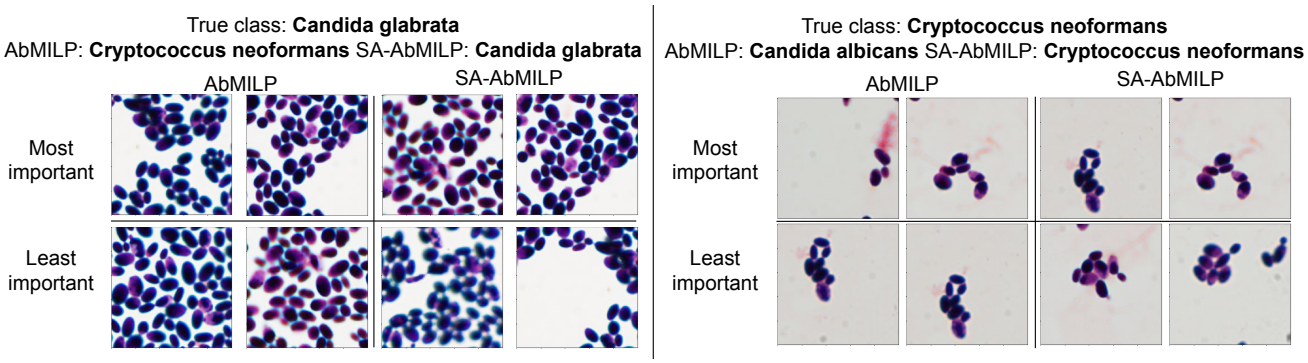

Figure 6: Example patches from the DIFaS dataset with the lowest and highest weights in the average pooling. One can observe that SA-AbMILP properly [38] assigns lower weights to blurred patches with a small number of cells. Moreover, in contrast to the AbMILP, it assigns high weight to clean patches (without red artifacts).

Table 3: Retinal image screening dataset results. *Results with asterisk are sourced from [30].

Retinal image screening dataset		
method	accuracy	F-score
LSA-AbMILP	**76.3%**	**0.77**
SA-AbMILP	75.2%	0.76
AbMILP	74.5%	0.74
GSA-AbMILP	74.5%	0.75
IQSA-AbMILP	74.5%	0.75
MSA-AbMILP	73.5%	0.74
MIL-GNN-DP*	74.2%	**0.77**
MIL-GNN-Att*	72.9%	0.75
mi-Graph*	72.5%	0.75
MILBoost*	64.1%	0.66
Citation k-NN*	62.8%	0.68
EMDD*	55.1%	0.69
MI-SVM*	54.5%	0.70
mi-SVM*	54.5%	0.71

Results. Results for "Messidor" database are presented in Table 3 alongside results of other approaches which are used for comparison and were taken from [30]. Our method improves accuracy and the highest scores are achieved using Laplace kernel variation, which obtains F1 score on par with the best reference approach. This is the only database for which Laplace kernel obtains the best results, which only confirms that the kernel should be optimized together with the other hyperparameters when applied to new problems.

5. Conclusions and discussion

In this paper, we apply Self-Attention into Attention-based MIL Pooling (SA-AbMILP), which combines the multi-level dependencies across image regions with the trainable operator of weighted average pooling. In contrast to Attention-based MIL Pooling (AbMILP), it covers not only the standard but also the presence-based and threshold-based assumptions of MIL. Self-Attention is detecting the relationship between instances, so it can embed into the instance feature vectors the information about the presence of similar instances or find that a combination of specific instances defines the bag as a positive. The experiments on five datasets (MNIST, two histological datasets of breast and colon cancer, microbiological dataset DIFaS, and retinal image screening) confirm that our method is on par or outperforms current state-of-the-art methodology based on the Wilcox pair test. We demonstrate that in the case of bigger datasets, it is advisable to use various kernels of the self-attention instead of the commonly used dot product. We also provide qualitative results to illustrate the reason for the improvements achieved by our method.

The experiments show that methods covering a wider range of MIL assumptions fit better for real-world problems. Therefore, in future work, we plan to introduce methods for more challenging MIL assumptions, e.g. collective assumption, and apply them to more complicated tasks, like digestive track assessment using the Nancy Histological Index. Moreover, we plan to introduce better interpretability, using Prototype Networks.

6. Acknowledgments

The POIR.04.04.00-00-14DE/18-00 project is carried out within the Team-Net programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund.

References

[1] Gaston Baudat and Fatiha Anouar. Kernel-based methods and function approximation. In *IJCNN'01. International Joint Conference on Neural Networks.*

Proceedings (Cat. No. 01CH37222), volume 2, pages 1244–1249. IEEE, 2001.

[2] 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.

[3] Gabriele Campanella, Vitor Werneck Krauss Silva, and Thomas J Fuchs. Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arXiv preprint arXiv:1805.06983*, 2018.

[4] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

[5] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[8] Ji Feng and Zhi-Hua Zhou. Deep miml network. In *AAAI*, pages 1884–1890, 2017.

[9] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010.

[10] Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and BS Manjunath. Evaluation and benchmark for biological image segmentation. In *2008 15th IEEE International Conference on Image Processing*, pages 1816–1819. IEEE, 2008.

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

[14] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Jiayun Li, Wenyuan Li, Arkadiusz Gertych, Beatrice S Knudsen, William Speier, and Corey W Arnold. An attention-based multi-resolution model for prostate whole slide imageclassification and localization. *arXiv preprint arXiv:1905.13208*, 2019.

[19] Katherine Li, Richard Strauss, Colleen Marano, Linda E Greenbaum, Joshua R Friedman, Laurent Peyrin-Biroulet, Carrie Brodmerkel, and Gert De Hertogh. A simplified definition of histologic improvement in ulcerative colitis and its association with disease outcomes up to 30 weeks from initiation of therapy: Post hoc analysis of three clinical trials. *Journal of Crohn's and Colitis*, 13(8):1025–1035, 2019.

[20] Ming Y Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*, 2019.

[21] Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*, 10:213–234, 2017.

[22] Gwénolé Quellec, Mathieu Lamard, Michael D Abràmoff, Etienne Decencière, Bruno Lay, Ali Erginay, Béatrice Cochener, and Guy Cazuguel. A multiple-instance learning framework for diabetic retinopathy screening. *Medical image analysis*, 16(6):1228–1240, 2012.

[23] Priya Rani, Rajkumar Elagiri Ramalingam, Kumar T Rajamani, Melih Kandemir, and Digvijay Singh. Multiple instance learning: Robust validation on retinopathy of prematurity. *Int J Ctrl Theory Appl*, 9:451–459, 2016.

[24] Lucia Ricci-Vitiani, Dario G Lombardi, Emanuela Pilozzi, Mauro Biffoni, Matilde Todaro, Cesare Peschle, and Ruggero De Maria. Identification and expansion of human colon-cancer-initiating cells. *Nature*, 445(7123):111–115, 2007.

[25] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. 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.

[26] Christoph Straehle, Melih Kandemir, Ullrich Koethe, and Fred A Hamprecht. Multiple instance learning with response-optimized random forests. In *2014 22nd International Conference on Pattern Recognition*, pages 3768–3773. IEEE, 2014.

[27] Jakub M Tomczak, Maximilian Ilse, Max Welling, Marnix Jansen, Helen G Coleman, Marit Lucas, Kikki de Laat, Martijn de Bruin, Henk Marquering, Myrtle J van der Wel, et al. Histopathological classification of precursor lesions of esophageal adenocarcinoma: A deep multiple instance learning approach. 2018.

[28] Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V Hajnal, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Multiple instance learning for classification of dementia in brain mri. *Medical image analysis*, 18(5):808–818, 2014.

[29] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*, 2019.

[30] Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. Multiple instance learning with graph neural networks. *arXiv preprint arXiv:1906.04881*, 2019.

[31] Chen Wang, Jianfei Yang, Lihua Xie, and Junsong Yuan. Kervolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 31–40, 2019.

[32] Shujun Wang, Yaxi Zhu, Lequan Yu, Hao Chen, Huangjing Lin, Xiangbo Wan, Xinjuan Fan, and Pheng-Ann Heng. Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification. *Medical image analysis*, 58:101549, 2019.

[33] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[34] Hiroshi Yoshida, Taichi Shimazu, Tomoharu Kiyuna, Atsushi Marugame, Yoshiko Yamashita, Eric Cosatto, Hirokazu Taniguchi, Shigeki Sekine, and Atsushi Ochiai. Automated histological classification of whole-slide images of gastric biopsy specimens. *Gastric Cancer*, 21(2):249–257, 2018.

[35] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[36] Zhendong Zhao, Gang Fu, Sheng Liu, Khaled M Elokely, Robert J Doerksen, Yixin Chen, and Dawn E Wilkins. Drug activity prediction using multiple-instance learning via joint instance and feature selection. In *BMC bioinformatics*, volume 14, page S16. Springer, 2013.

[37] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256, 2009.

[38] Bartosz Zieliński, Agnieszka Sroka-Oleksiak, Dawid Rymarczyk, Adam Piekarczyk, and Monika Brzychczy-Włoch. Deep learning approach to description and classification of fungi microscopic images. *arXiv preprint arXiv:1906.09449*, 2019.