Attention-based Deep Multiple Instance Learning

Maximilian Ilse [*1] **Jakub M. Tomczak** [*1] **Max Welling** [1]

Abstract

Multiple instance learning (MIL) is a variation of supervised learning where a single class label is assigned to a bag of instances. In this paper, we state the MIL problem as learning the Bernoulli distribution of the bag label where the bag label probability is fully parameterized by neural networks. Furthermore, we propose a neural network-based permutation-invariant aggregation operator that corresponds to the attention mechanism. Notably, an application of the proposed attention-based operator provides insight into the contribution of each instance to the bag label. We show empirically that our approach achieves comparable performance to the best MIL methods on benchmark MIL datasets and it outperforms other methods on a MNIST-based MIL dataset and two real-life histopathology datasets without sacrificing interpretability.

1. Introduction

In typical machine learning problems like image classification it is assumed that an image clearly represents a category (a class). However, in many real-life applications multiple instances are observed and only a general statement of the category is given. This scenario is called *multiple instance learning* (MIL) (Dietterich et al., 1997; Maron & Lozano-Pérez, 1998) or, *learning from weakly annotated data* (Oquab et al., 2014). The problem of weakly annotated data is especially apparent in medical imaging (Quellec et al., 2017) (*e.g.*, computational pathology, mammography or CT lung screening) where an image is typically described by a single label (benign/malignant) or a Region Of Interest (ROI) is roughly given.

MIL deals with a bag of instances for which a single class label is assigned. Hence, the main goal of MIL is to learn a model that predicts a bag label, *e.g.*, a medical diagnosis. An additional challenge is to discover *key instances* (Liu et al., 2012), *i.e.*, the instances that trigger the bag label. In the medical domain the latter task is of great interest because of legal issues[1] and its usefulness in clinical practice. In order to solve the primary task of a bag classification different methods are proposed, such as utilizing similarities among bags (Cheplygina et al., 2015b), embedding instances to a compact low-dimensional representation that is further fed to a bag-level classifier (Andrews et al., 2003; Chen et al., 2006), and combining responses of an instance-level classifier (Ramon & De Raedt, 2000; Raykar et al., 2008; Zhang et al., 2006). Only the last approach is capable of providing interpretable results. However, it was shown that the instance level accuracy of such methods is low (Kandemir & Hamprecht, 2015) and in general there is a disagreement among MIL methods at the instance level (Cheplygina et al., 2015a). These issues call into question the usability of current MIL models for interpreting the final decision.

In this paper, we propose a new method that aims at incorporating interpretability to the MIL approach and increasing its flexibility. We formulate the MIL model using the Bernoulli distribution for the bag label and train it by optimizing the log-likelihood function. We show that the application of the Fundamental Theorem of Symmetric Functions provides a general procedure for modeling the bag label probability (the bag score function) that consists of three steps: (i) a transformation of instances to a low-dimensional embedding, (ii) a permutation-invariant (symmetric) aggregation function, and (iii) a final transformation to the bag probability. We propose to parameterize all transformations using neural networks (*i.e.*, a combination of convolutional and fully-connected layers), which increases the flexibility of the approach and allows to train the model in an end-to-end manner by optimizing an unconstrained objective function. Last but not least, we propose to replace widely-used permutation-invariant operators such as the maximum operator $\max$ and the mean operator mean by a trainable weighted average where weights are given by a two-layered neural network. The two-layered neural network corre-

[*]Equal contribution [1]University of Amsterdam, the Netherlands. Correspondence to: Maximilian Ilse <m.ilse@uva.nl>, Jakub M. Tomczak <j.m.tomczak@uva.nl>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]According to the European Union General Data Protection Regulation (taking effect 2018), a user should have the right to obtain an explanation of the decision reached.

sponds to the attention mechanism (Bahdanau et al., 2014; Raffel & Ellis, 2015). Notably, the attention weights allow us to find key instances, which could be further used to highlight possible ROIs. In the experiments we show that our model is on a par with the best classical MIL methods on common benchmark MIL datasets, and that it outperforms other methods on a MNIST-based MIL problem as well as two real-life histopathology image datasets. Moreover, in the image datasets we provide empirical evidence that our model can indicate key instances.

2. Methodology

2.1. Multiple instance learning (MIL)

Problem formulation In the classical (binary) supervised learning problem one aims at finding a model that predicts a value of a target variable, $y \in \{0, 1\}$, for a given instance, $\mathbf{x} \in \mathbb{R}^D$. In the case of the MIL problem, however, instead of a single instance there is a bag of instances, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, that exhibit neither dependency nor ordering among each other. We assume that K could vary for different bags. There is also a single binary label Y associated with the bag. Furthermore, we assume that individual labels exist for the instances within a bag, *i.e.*, $y_1, \ldots, y_K$ and $y_k \in \{0, 1\}$, for $k = 1, \ldots, K$, however, there is no access to those labels and they remain unknown during training. We can re-write the assumptions of the MIL problem in the following form:

$$Y = \begin{cases} 0, & \text{iff } \sum_k y_k = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

These assumptions imply that a MIL model must be **permutation-invariant**. Further, the two statements could be re-formulated in a compact form using the maximum operator:

$$Y = \max_k \{y_k\}. \tag{2}$$

Learning a model that tries to optimize an objective based on the maximum over instance labels would be problematic at least for two reasons. First, all gradient-based learning methods would encounter issues with vanishing gradients. Second, this formulation is suitable only when an instance-level classifier is used.

In order to make the learning problem easier, we propose to train a MIL model by optimizing the log-likelihood function where the bag label is distributed according to the Bernoulli distribution with the parameter $\theta(X) \in [0, 1]$, *i.e.*, the probability of $Y = 1$ given the bag of instances X.

MIL approaches In the MIL setting the bag probability $\theta(X)$ must be permutation-invariant since we assume neither ordering nor dependency of instances within a bag. Therefore, the MIL problem can be considered in terms of a specific form of the Fundamental Theorem of Symmetric Functions with monomials given by the following theorem (Zaheer et al., 2017):

Theorem 1. *A scoring function for a set of instances X, $S(X) \in \mathbb{R}$, is a symmetric function (i.e., permutation-invariant to the elements in X), if and only if it can be decomposed in the following form:*

$$S(X) = g\Big(\sum_{\mathbf{x} \in X} f(\mathbf{x})\Big), \tag{3}$$

where f and g are suitable transformations.

This theorem provides a general strategy for modeling the bag probability using the decomposition given in (3). A similar decomposition with $\max$ instead of sum is given by the following theorem (Qi et al., 2017):

Theorem 2. *For any $\varepsilon > 0$, a Hausdorff continuous symmetric function $S(X) \in \mathbb{R}$ can be arbitrarily approximated by a function in the form $g\big(\max_{\mathbf{x} \in X} f(\mathbf{x})\big)$, where $\max$ is the element-wise vector maximum operator and f and g are continuous functions, that is:*

$$|S(X) - g\big(\max_{\mathbf{x} \in X} f(\mathbf{x})\big)| < \varepsilon. \tag{4}$$

The difference between Theorems 1 and 2 is that the former is a universal decomposition while the latter provides an arbitrary approximation. Nonetheless, they both formulate a general three-step approach for classifying a bag of instances: (i) a transformation of instances using the function f, (ii) a combination of transformed instances using a symmetric (permutation-invariant) function σ, (iii) a transformation of combined instances transformed by f using a function g. Finally, the expressiveness of the score function relies on the choice of classes of functions for f and g.

In the MIL problem formulation the score function in both theorems is the probability $\theta(X)$ and the permutation-invariant function σ is referred to as the MIL pooling. The choice of functions f, g and σ determines a specific approach to modeling the label probability. For a given MIL operator there are two main MIL approaches:

(i) *The instance-level approach*: The transformation f is an instance-level classifier that returns scores for each instance. Then individual scores are aggregated by MIL pooling to obtain $\theta(X)$. The function g is the identity function.

(ii) *The embedding-level approach*: The function f maps instances to a low-dimensional embedding. MIL pooling is used to obtain a bag representation that is independent of the number of instances in the bag. The bag representation is further processed by a bag-level classifier to provide $\theta(X)$.

It is advocated in (Wang et al., 2016) that the latter approach is preferable in terms of the bag level classification performance. Since the individual labels are unknown, there is a threat that the instance-level classifier might be trained insufficiently and it introduces additional error to the final prediction. The embedding-level approach determines a joint representation of a bag and therefore it does not introduce additional bias to the bag-level classifier. On the other hand, the instance-level approach provides a score that can be used to find *key instances i.e.*, the instances that trigger the bag label. Liu et al. (2012) were able to show that a model that is successfully detecting key instances is more likely to achieve better bag label predictions. We will show how to modify the embedding-level approach to be interpretable by using a new MIL pooling.

2.2. MIL with Neural Networks

In classical MIL problems it is assumed that instances are represented by features that do not require further processing, *i.e.*, f is the identity. However, for some tasks like image or text analysis additional steps of feature extraction are necessary. Additionally, Theorem 1 and 2 indicate that for a flexible enough class of functions we can model any permutation-invariant score function. Therefore, we consider a class of transformations that are parameterized by neural networks $f_\psi(\cdot)$ with parameters ψ that transform the k-th instance into a low-dimensional embedding, $\mathbf{h}_k = f_\psi(\mathbf{x}_k)$, where $\mathbf{h}_k \in \mathcal{H}$ such that $\mathcal{H} = [0,1]$ for the instance-based approach and $\mathcal{H} = \mathbb{R}^M$ for the embedding-based approach.

Eventually, the parameter $\theta(X)$ is determined by a transformation $g_\phi : \mathcal{H}^K \to [0,1]$. In the instance-based approach the transformation g_ϕ is simply the identity, while in the embedding-based approach it could be also parameterized by a neural network with parameters ϕ. The former approach is depicted in Figure 6(a) and the latter in Figure 6(b) in the Appendix.

The idea of parameterizing all transformations using neural networks is very appealing because the whole approach can be arbitrarily flexible and it can be trained end-to-end by backpropagation. The only restriction is that the MIL pooling must be differentiable.

2.3. MIL pooling

The formulation of the MIL problem requires the MIL pooling σ to be permutation-invariant. As shown in Theorem 1 and 2, there are two MIL pooling operators that ensure the score function (*i.e.*, the bag probability) to be a symmetric function, namely, the maximum operator:

$$\forall_{m=1,\dots,M} : z_m = \max_{k=1,\dots,K} \{\mathbf{h}_{km}\}, \tag{5}$$

and the mean operator:[2]

$$\mathbf{z} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}_k. \tag{6}$$

In fact, other operators could be used such as, the convex maximum operator (*i.e.*, log-sum-exp) (Ramon & De Raedt, 2000), Integrated Segmentation and Recognition (Keeler et al., 1991), noisy-or (Maron & Lozano-Pérez, 1998) and noisy-and (Kraus et al., 2016). These MIL pooling operators could replace $\max$ in Theorem 2 and proofs would follow in a similar manner (see Supplementary in (Qi et al., 2017) for a detailed proof for the maximum operator). All of these operators are differentiable, hence, they could be easily used as a MIL pooling layer in a deep neural network architecture.

2.4. Attention-based MIL pooling

All MIL pooling operators mentioned in the previous section have a clear disadvantage, namely, they are pre-defined and non-trainable. For instance, the $\max$-operator could be a good choice in the instance-based approach but it might be inappropriate for the embedding-based approach. Similarly, the mean operator is definitely a bad MIL pooling to aggregate instance scores, although, it could succeed in calculating the bag representation. Therefore, a flexible and adaptive MIL pooling could potentially achieve better results by adjusting to a task and data. Ideally, such MIL pooling should also be interpretable, a trait that is missing in all operators mentioned in Section 2.3.

Attention mechanism We propose to use a weighted average of instances (low-dimensional embeddings) where weights are determined by a neural network. Additionally, the weights must sum to 1 to be invariant to the size of a bag. The weighted average fulfills the requirements of the Theorem 1 where the weights together with the embeddings are part of the f function. Let $H = \{\mathbf{h}_1, \dots, \mathbf{h}_K\}$ be a bag of K embeddings, then we propose the following MIL pooling:

$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k, \tag{7}$$

where:

$$a_k = \frac{\exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\}}, \tag{8}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times M}$ are parameters. Moreover, we utilize the hyperbolic tangent $\tanh(\cdot)$ element-wise non-linearity to include both negative and positive values for proper gradient flow. The proposed construction allows to discover (dis)similarities among instances.

[2] Notice that the weight $\frac{1}{K}$ can be seen as a part of the f function.

Interestingly, the proposed MIL pooling corresponds to a version of the attention mechanism (Lin et al., 2017; Raffel & Ellis, 2015). The main difference is that typically in the attention mechanism all instances are sequentially dependent while here we assume that all instances are independent. Therefore, a naturally arising question is whether the attention mechanism could work without sequential dependencies among instances, and if it will not learn the mean operator. We will address this issue in the experiments.

Gated attention mechanism Furthermore, we notice that the $\tanh(\cdot)$ non-linearity could be inefficient to learn complex relations. Our concern follows from the fact that $\tanh(x)$ is approximately linear for $x \in [-1, 1]$, which could limit the final expressiveness of learned relations among instances. Therefore, we propose to additionally use the gating mechanism (Dauphin et al., 2016) together with $\tanh(\cdot)$ non-linearity that yields:

$$a_k = \frac{\exp\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_k^\top\right) \odot \mathrm{sigm}\left(\mathbf{U}\mathbf{h}_k^\top\right)\right)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_j^\top\right) \odot \mathrm{sigm}\left(\mathbf{U}\mathbf{h}_j^\top\right)\right)\}}, \quad (9)$$

where $\mathbf{U} \in \mathbb{R}^{L \times M}$ are parameters, $\odot$ is an element-wise multiplication and $\mathrm{sigm}(\cdot)$ is the sigmoid non-linearity. The gating mechanism introduces a learnable non-linearity that potentially removes the troublesome linearity in $\tanh(\cdot)$.

Flexibility In principle, the proposed attention-based MIL pooling allows to assign different weights to instances within a bag and hence the final representation of the bag could be highly informative for the bag-level classifier. In other words, it should be able to find key instances. Moreover, application of the attention-based MIL pooling together with the transformations f and g parameterized by neural networks makes the whole model fully differentiable and adaptive. These two facts make the proposed MIL pooling a potentially very flexible operator that could model an arbitrary permutation-invariant score function. The proposed attention mechanism together with a deep MIL model is depicted in Figure 6(c) in the Appendix.

Interpretability Ideally, in the case of a positive label ($Y = 1$), high attention weights should be assigned to instances that are likely to have label $y_k = 1$ (key instances). Namely, the attention mechanism allows to easily interpret the provided decision in terms of instance-level labels. In fact, the attention network does not provide scores as the instance-based classifier does but it can be considered as a proxy to that. The attention-based MIL pooling bridges the instance-level approach and the embedding-level approach.

From the practical point of view, *e.g.*, in the computational pathology, it is desirable to provide ROIs together with the final diagnosis to a doctor. Therefore, the attention mechanism is potentially of great interest in practical applications.

3. Related work

MIL pooling Typically, MIL approaches utilize either the mean pooling or the $\max$ pooling, while the latter is mostly used (Feng & Zhou, 2017; Pinheiro & Collobert, 2015; Zhu et al., 2017). Both operators are non-trainable which potentially limits their applicability. There are MIL pooling operators that contain global adaptive parameters, such as noisy-and (Kraus et al., 2016), however, their flexibility is restricted. We propose a fully trainable MIL pooling that adapts to new instances.

MIL with neural networks In the classical work on MIL it is assumed that instances are represented by precomputed features and there is very little need to apply additional feature extraction. Nevertheless, recent work on utilizing fully-connected neural networks in MIL shows that it could still be beneficial (Wang et al., 2016). Similarly, in computer vision the idea of MIL combined with deep learning significantly improves final accuracy (Oquab et al., 2014). In this paper, we follow this line of research since it allows to apply a flexible class of transformations that can be trained end-to-end by backpropagation.

MIL and attention The attention mechanism is widely used in deep learning for image captioning (Xu et al., 2015) or text analysis (Bahdanau et al., 2014; Lin et al., 2017). In the context of the MIL problem it has rarely been used and only in a very limited form. In (Pappas & Popescu-Belis, 2014) an attention-based MIL was proposed but attention weights were trained as parameters of an auxiliary linear regression model. This idea was further expanded and the linear regression model was replaced by a one-layer neural network with single output (Pappas & Popescu-Belis, 2017). The attention-based MIL operator was used very recently in (Qi et al., 2017), however, the attention was calculated using the dot product and it performed worse than the $\max$ operator. Here, we propose to use a two-layered neural network to learn the MIL operator and we show that it outperforms commonly used MIL pooling operators.

MIL for medical imaging The MIL seems to perfectly fit medical imaging where processing a whole image consisting of billions of pixels is computationally infeasible. Moreover, in the medical domain it is very difficult to obtain pixel-level annotations, that drastically reduces number of available data. Therefore, it is tempting to divide a medical image into smaller patches that could be further considered as a bag with a single label (Quellec et al., 2017). This idea attracts a great interest in the computational histopathology where patches could correspond to cells that are believed to indicate malignant changes (Sirinukunwattana et al., 2016). Different MIL approaches were used for histopathology data, such as, Gaussian processes (Kandemir et al., 2014; 2016) or a two-stage approach with neural networks and EM algorithm to determine instance classes (Hou et al., 2016).

Other applications of MIL methods in medical imaging are mammography (nodule) classification (Zhu et al., 2017) and microscopy cell detection (Kraus et al., 2016). In this paper, we show that the proposed attention-based deep MIL approach can be used not only to provide the final diagnosis but also to indicate ROIs in a histopathology slide.

4. Experiments

In the experiments we aim at evaluating the proposed approach: a MIL model parameterized with neural networks and a (gated) attention-based pooling layer ('Attention' and 'Gated-Attention'). We evaluate our approach on a number of different MIL datasets: five MIL benchmark datasets (MUSK1, MUSK2, FOX, TIGER, ELEPHANT), an MNIST-based image dataset (MNIST-BAGS) and two real-life histopathology datasets (BREAST CANCER, COLON CANCER). We want to verify two research questions in the experiments: (i) whether our approach achieves the best performance or is comparable to the best performing method, (ii) if our method can provide interpretable results by using the attention weights that indicate key instances or ROIs.

In order to obtain a fair comparison we use a common evaluation methodology, *i.e.*, 10-fold-cross-validation, and five repetitions per experiment. In the case of MNIST-BAGS we use a fixed division into training and test set. In order to create test bags we solely sampled images from the MNIST test set. During training we only used images from the MNIST training set. For all experiments we use modified versions of models that have shown high classification performance on the individual datasets (Wang et al., 2016; LeCun et al., 1998; Sirinukunwattana et al., 2016). The MIL pooling layers are either located before the last layer of the model (the embedded-based approach) or after last layer of the model (the instance-based approach). If an attention-based MIL pooling layer is used the number of parameters in $\mathbf{V}$ was determined using a validation set. We tested the following dimensions (L): 64, 128 and 256. The different dimensions only resulted in minor changes of the model's performance. For layers using the gated attention mechanism $\mathbf{V}$ and $\mathbf{U}$ have the same number of parameters. Finally, all layers were initialized according to Glorot & Bengio (2010) and biases were set to zero.

We compare our approach to various MIL methods on MIL benchmark datasets. On the image datasets our method is compared with instance-level and embedding-level neural networks and commonly used MIL pooling layers (max and mean). In the following, we are using 'Instance+max/mean' and 'Embedding+max/mean' to indicate networks that are build from convolutional layers and fully-connected layers. In contrast to networks purely build from fully-connected layers, referred to as 'mi-Net' and 'MI-Net' (Wang et al., 2016).

On MNIST-BAGS we include a SVM-based MIL model, called (MI-SVM). We do not present results of MI-SVM on the histopathology datasets since we could not train (including hyperparameter search and five times 10-fold-cross-validation procedure) the model in a reasonable amount of time.[3] In order to compare the bag level performance we use the following metrics: the classification accuracy, precision, recall, F-score, and the area under the receiver operating characteristic curve (AUC).

4.1. Classical MIL datasets

Details In the first experiment we aim at verifying whether our approach can compete with the best MIL methods on historically important benchmark datasets. Since all five datasets contain precomputed features and only a small number of instances and bags, neural networks are most likely not well suited. First we predict drug activity (MUSK1 and MUSK2). A molecule has the desired drug effect if and only if one or more of its conformations bind to the target binding site. Since molecules can adopt multiple shapes, a bag is made up of shapes belonging to the same molecule (Dietterich et al., 1997). The three remaining datasets, ELEPHANT, FOX and TIGER, contain features extracted from images. Each bag consists of a set of segments of an image. For each category, positive bags are images that contain the animal of interest, and negative bags are images that contain other animals (Andrews et al., 2003). For detailed information on the number of bags, instances and features in each dataset see Section 6.3 in the Appendix.

In our experiments we use the same architecture, optimizer and hyperparameters as in the MI-Net model (Wang et al., 2016).

Table 1. Results on classical MIL datasets. Experiments were run 5 times and an average of the classification accuracy (± a standard error of a mean) is reported. [1] (Andrews et al., 2003), [2] (Gärtner et al., 2002), [3] (Zhang & Goldman, 2002) [4] (Zhou et al., 2009) [5] (Wei et al., 2017) [6] (Wang et al., 2016)

METHOD	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
mi-SVM [1]	0.874±N/A	0.836±N/A	0.582±N/A	0.784±N/A	0.822±N/A
MI-SVM [1]	0.779±N/A	0.843±N/A	0.578±N/A	0.840±N/A	0.843±N/A
MI-Kernel [2]	**0.880**±0.031	**0.893**±0.015	**0.603**±0.028	0.842±0.010	0.843±0.016
EM-DD [3]	0.849±0.044	**0.869**±0.048	**0.609**±0.045	0.730±0.043	0.771±0.043
mi-Graph [4]	**0.889**±0.033	**0.903**±0.039	**0.620**±0.044	**0.860**±0.037	**0.869**±0.035
miVLAD [5]	**0.871**±0.043	**0.872**±0.042	**0.620**±0.044	0.811±0.039	**0.850**±0.036
miFV [5]	**0.909**±0.040	**0.884**±0.042	**0.621**±0.049	0.813±0.037	**0.852**±0.036
mi-Net [6]	**0.889**±0.039	**0.858**±0.049	**0.613**±0.035	0.824±0.034	**0.858**±0.037
MI-Net [6]	**0.887**±0.041	**0.859**±0.046	**0.622**±0.038	**0.830**±0.032	**0.862**±0.034
MI-Net with DS [6]	**0.894**±0.042	**0.874**±0.043	**0.630**±0.037	**0.845**±0.039	**0.872**±0.032
MI-Net with RC [6]	**0.898**±0.043	**0.873**±0.044	**0.619**±0.047	**0.836**±0.037	**0.857**±0.040
Attention	**0.892**±0.040	**0.858**±0.048	**0.615**±0.043	**0.839**±0.022	**0.868**±0.022
Gated-Attention	**0.900**±0.050	**0.863**±0.042	**0.603**±0.029	**0.845**±0.018	**0.857**±0.027

Results and discussion The results of the experiment are

[3] Learning a single MI-SVM took approximately one week due to the large number of patches.

presented in Table 1. Our approaches (Attention and Gated-Attention) are comparable with the best performing classical MIL methods (notice the standard error of the mean).

4.2. MNIST-bags

Details The main disadvantage of the classical MIL benchmark datasets is that instances are represented by precomputed features. In order to consider a more challenging scenario, we propose to investigate a dataset that is created using the well-known MNIST image dataset. A bag is made up of a random number of 28×28 grayscale images taken from the MNIST dataset. The number of images in a bag is Gaussian-distributed and the closest integer value is taken. A bag is given a positive label if it contains one or more images with the label '9'. We chose '9' since it can be easily mistaken with '7' or '4'. We investigate the influence of the number of bags in the training set as well as the average number of instances per bag on the prediction performance. During evaluation we use a fixed number of 1000 test bags. For all experiments a LeNet5 model is used (LeCun et al., 1998), see Table 8 and 9 in the Appendix. The models are trained with the Adam optimization algorithm (Kingma & Ba, 2014). We keep the default parameters for β_1 and β_2, see Table 10 in the Appendix. In addition, we compare our method with a SVM-based MIL method (MI-SVM) (Andrews et al., 2003) that uses a Gaussian kernel on raw pixel features[4].

In the experiments we use different numbers of the mean bag size, namely, 10, 50 and 100, and the variance 2, 10, 20, respectively. Moreover, we use varying numbers of training bags, *i.e.*, 50, 100, 150, 200, 300, 400, 500. These different settings allow us to verify how different number of training bags and different number of instances influence MIL models. We compare instance-based and embedding-based approaches parameterized with a neural network (LeNet5) with mean and max MIL pooling. We use AUC as the evaluation metric.

Results and discussion The results of AUC for the mean bag sizes equal to 10, 50 and 100 are presented in Figure 1, 2 and 3, respectively, and detailed results are given in the Appendix. The findings of the experiment are the following: First, the proposed attention-based deep MIL approach performs much better than other methods in the small sample size regime. Moreover, when there is a small effective size of the training set that corresponds to 50-150 bags for around 10 instances per bag (see Figure 1) or 50-100 bags in the case of on average 50 instances in a bag (see Figure 2), our method still achieves significantly higher AUC than all other methods. Second, we notice that our approach is more flexible and obtained better results than the SVM-

[4]We use code provided with (Doran & Ray, 2014): https://github.com/garydoranjr/misvm

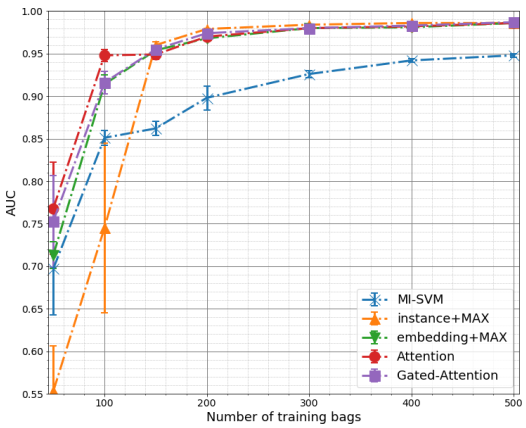


Figure 1. The test AUC for MNIST-BAGS with on average 10 instances per bag.

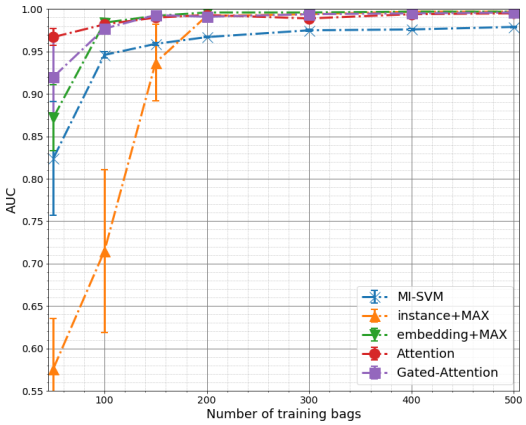


Figure 2. The test AUC for MNIST-BAGS with on average 50 instances per bag.

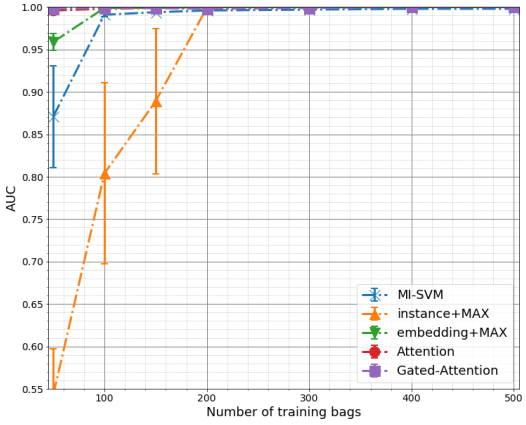


Figure 3. The test AUC for MNIST-BAGS with on average 100 instances per bag.

based approach in all cases except large effective sample sizes (see Figure 3). Third, the embedding-based models performed better than the instance-based models. However, for a sufficient number of training images (number of training bags and training instances per bag) all models achieve very similar results. Fourth, the mean operator performs significantly worse than the max operator. However, the embedding-based model with the mean operator converged eventually to the best value but always later than the one with max. See Section 6.4 in the Appendix for details.

The results of this experiment indicate that for a small-sample size regime our approach is preferable to others. Since attention serves as a gradient update filter during backpropagation (Wang et al., 2017), instances with higher weights will contribute more to learning the encoder network of instances. This is especially important since medical imaging problems contain only a small number of cases. In general, the more instances are in a bag the easier the MIL task becomes, since the MIL assumption states that every instance in a negative bag is negative. For example, a negative bag of size 100 from the MNIST-bags dataset will include about 11 negative examples per class.

Finally, we present an exemplary result of the attention mechanism in Figure 4. In this example a bag consists of 13 images. For each digit the corresponding attention weight is given by the trained network. The bag is properly predicted as positive and all nines are correctly highlighted. Hence, the attention mechanism works as expected. More examples are given in the Appendix.

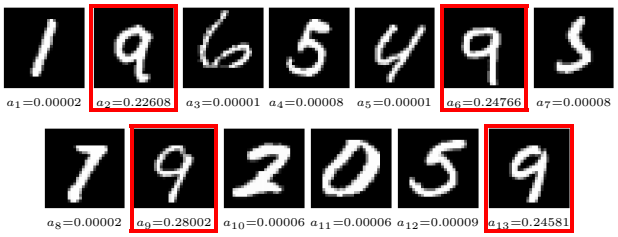


Figure 4. Example of attention weights for a positive bag.

4.3. Histopathology datasets

Details An automatic detection of cancerous regions in hematoxylin and eosin (H&E) stained whole-slide images is a task with high clinical relevance. Current supervised approaches utilize pixel-level annotations (Litjens et al., 2017). However, data preparation requires large amount of time from pathologists which highly interferes with their daily routines. Hence, a successful solution working with weak labels would hold a great promise to reduce the workload of the pathologists. In the following, we perform two experiments on classifying weakly-labeled real-life histopathology images of the breast cancer dataset (BREAST CANCER) (Gelasca et al., 2008) and the colon cancer dataset (COLON CANCER) (Sirinukunwattana et al., 2016).

BREAST CANCER consists of 58 weakly labeled 896×768 H&E images. An image is labeled malignant if it contains breast cancer cells, otherwise it is benign. We divide every image into 32×32 patches. This results in 672 patches per bag. A patch is discarded if it contains 75% or more of white pixels.

COLON CANCER comprises 100 H&E images. The images originate from a variety of tissue appearance from both normal and malignant regions. For every image the majority of nuclei of each cell were marked. In total there are 22,444 nuclei with associated class label, *i.e.* epithelial, inflammatory, fibroblast, and miscellaneous. A bag is composed of 27×27 patches. Furthermore, a bag is given a positive label if it contains one or more nuclei from the epithelial class. Tagging epithelial cells is highly relevant from a clinical point of view, since colon cancer originates from epithelial cells (Ricci-Vitiani et al., 2007).

For both datasets we use the model proposed in (Sirinukunwattana et al., 2016) for the transformation f. All models are trained with the Adam optimization algorithm (Kingma & Ba, 2014). Due to the limited amount of data samples in both datasets we performed data augmentation to prevent overfitting. See the Appendix for further details.

Results and discussion We present results in Table 2 and 3 for BREAST CANCER and COLON CANCER, respectively. First, we notice that the obtained results confirm our findings in MNIST-BAGS experiment that our approach outperforms all other methods. A trend that is especially visible in the small-sample size regime of the MNIST-BAGS. Surprisingly, the embedding-based method with the max pooling failed almost completely on BREAST CANCER but in general this dataset is difficult due to high variability of slides and small number of cases. The proposed method is not only most accurate but it also received the highest recall. High recall is especially important in the medical domain since false negatives could lead to severe consequences including patient fatality. We also notice that the gated-attention mechanism performs better than the plain attention mechanism on BREAST CANCER while these two behave similarly on COLON CANCER.

Eventually, we present the usefulness of the attention mechanism in providing ROIs. In Figure 5 we show a histopathology image divided into patches containing (mostly) single cells. We create a heatmap by multiplying patches by its corresponding attention weight. Although only image-level annotations are used during training, there is a substantial matching between the heatmap in Figure 5(d) and the ground truth in Figure 5(c). Additionally, we notice that

Table 2. Results on BREAST CANCER. Experiments were run 5 times and an average (± a standard error of the mean) is reported.

METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
Instance+max	0.614±0.020	0.585±0.03	0.477±0.087	0.506±0.054	0.612±0.026
Instance+mean	0.672±0.026	0.672±0.034	0.515±0.056	0.577±0.049	0.719±0.019
Embedding+max	0.607±0.015	0.558±0.013	0.546±0.070	0.543±0.042	0.650±0.013
Embedding+mean	**0.741**±0.023	**0.741**±0.023	0.654±0.054	0.689±0.034	**0.796**±0.012
Attention	**0.745**±0.018	0.718±0.021	**0.715**±0.046	**0.712**±0.025	0.775±0.016
Gated-Attention	**0.755**±0.016	**0.728**±0.016	**0.731**±0.042	**0.725**±0.023	**0.799**±0.020

Table 3. Results on COLON CANCER. Experiments were run 5 times and an average (± a standard error of the mean) is reported.

METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
Instance+max	0.842 ± 0.021	0.866 ± 0.017	0.816 ± 0.031	0.839 ± 0.023	0.914 ± 0.010
Instance+mean	0.772 ± 0.012	0.821 ± 0.011	0.710 ± 0.031	0.759 ± 0.017	0.866 ± 0.008
Embedding+max	0.824 ± 0.015	0.884 ± 0.014	0.753 ± 0.020	0.813 ± 0.017	0.918 ± 0.010
Embedding+mean	0.860 ± 0.014	0.911 ± 0.011	0.804 ± 0.027	0.853 ± 0.016	0.940 ± 0.010
Attention	**0.904** ± 0.011	**0.953** ± 0.014	**0.855** ± 0.017	**0.901** ± 0.011	**0.968** ± 0.009
Gated-Attention	**0.898** ± 0.020	**0.944** ± 0.016	**0.851** ± 0.035	**0.893** ± 0.022	**0.968** ± 0.010

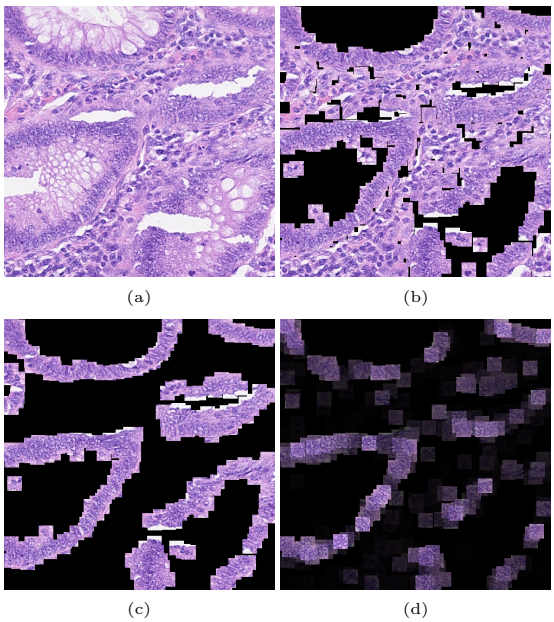


Figure 5. (a) H&E stained histology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Heatmap: Every patch from (b) multiplied by its corresponding attention weight, we rescaled the attention weights using $a'_k = (a_k - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$.

the instance-based classifier tends to select only a small subset of positive patches (see Figure 10(e) in Appendix) that confirms low instance accuracy of the instance-based approach discussed in (Kandemir & Hamprecht, 2015). For more examples please see the Appendix.

The obtained results again confirm that the proposed approach attains high predictive performance and allows to properly highlight ROIs. Moreover, the attention weights can be used to create a reliable heatmap.

5. Conclusion

In this paper, we proposed a flexible and interpretable MIL approach that is fully parameterized by neural networks. We outlined the usefulness of deep learning for modeling a permutation-invariant bag score function in terms of the Fundamental Theorem of Symmetric Functions. Moreover, we presented a trainable MIL pooling based on the (gated) attention mechanism. We showed empirically on five MIL datasets, one image corpora and two real-life histopathology datasets that our method is on a par with the best performing methods or performs the best in terms of different evaluation metrics. Additionally, we showed that our approach provides an interpretation of the decision by presenting ROIs, which is extremely important in many practical applications.

We strongly believe that the presented line of research is worth pursuing further. Here we focused on a binary MIL problem, however, the multi-class MIL is more interesting and challenging (Feng & Zhou, 2017). Moreover, in some applications it is worth to consider *repulsion points* (Scott et al., 2005), *i.e.*, instances for which a bag is always negative, or assume dependencies among instances within a bag (Zhou et al., 2009). We leave investigating these issues for future research.

Acknowledgements

The authors are very grateful to Rianne van den Berg for insightful remarks and discussions.

Maximilian Ilse was funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Grant DLMedIa: Deep Learning for Medical Image Analysis).

Jakub Tomczak was funded by the European Commission within the Marie Skodowska-Curie Individual Fellowship (Grant No. 702666, "Deep learning and Bayesian inference for medical imaging").

References

Andrews, Stuart, Tsochantaridis, Ioannis, and Hofmann, Thomas. Support vector machines for multiple-instance learning. In *NIPS*, pp. 577–584, 2003.

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Chen, Yixin, Bi, Jinbo, and Wang, James Ze. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

Cheplygina, Veronika, Sørensen, Lauge, Tax, David MJ, de Bruijne, Marleen, and Loog, Marco. Label stability in multiple instance learning. In *MICCAI*, pp. 539–546, 2015a.

Cheplygina, Veronika, Tax, David MJ, and Loog, Marco. Multiple instance learning with bag dissimilarities. *Pattern Recognition*, 48(1):264–275, 2015b.

Dauphin, Yann N, Fan, Angela, Auli, Michael, and Grangier, David. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083*, 2016.

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

Doran, Gary and Ray, Soumya. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine Learning*, 97(1-2):79–102, 2014.

Feng, Ji and Zhou, Zhi-Hua. Deep MIML Network. In *AAAI*, pp. 1884–1890, 2017.

Gärtner, Thomas, Flach, Peter A, Kowalczyk, Adam, and Smola, Alexander J. Multi-instance kernels. In *ICML*, volume 2, pp. 179–186, 2002.

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

Glorot, Xavier and Bengio, Yoshua. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pp. 249–256, 2010.

Hou, Le, Samaras, Dimitris, Kurc, Tahsin M, Gao, Yi, Davis, James E, and Saltz, Joel H. Patch-based convolutional neural network for whole slide tissue image classification. In *CVPR*, pp. 2424–2433, 2016.

Kandemir, Melih and Hamprecht, Fred A. Computer-aided diagnosis from weak supervision: a benchmarking study. *Computerized Medical Imaging and Graphics*, 42:44–50, 2015.

Kandemir, Melih, Zhang, Chong, and Hamprecht, Fred A. Empowering multiple instance histopathology cancer diagnosis by cell graphs. In *MICCAI*, pp. 228–235, 2014.

Kandemir, Melih, Haußmann, Manuel, Diego, Ferran, Rajamani, Kumar T, van der Laak, Jeroen, and Hamprecht, Fred A. Variational Weakly Supervised Gaussian Processes. In *BMVC*, 2016.

Keeler, James D, Rumelhart, David E, and Leow, Wee Kheng. Integrated segmentation and recognition of hand-printed numerals. In *NIPS*, pp. 557–563, 1991.

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

Kraus, Oren Z, Ba, Jimmy Lei, and Frey, Brendan J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

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

Lin, Zhouhan, Feng, Minwei, Santos, Cicero Nogueira dos, Yu, Mo, Xiang, Bing, Zhou, Bowen, and Bengio, Yoshua. A structured self-attentive sentence embedding. 2017.

Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, van der Laak, Jeroen A.W.M., van Ginneken, Bram, and Snchez, Clara I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017.

Liu, Guoqing, Wu, Jianxin, and Zhou, Zhi-Hua. Key instance detection in multi-instance learning. In *JMLR*, volume 25, pp. 253–268, 2012.

Maron, Oded and Lozano-Pérez, Tomás. A framework for multiple-instance learning. In *NIPS*, pp. 570–576, 1998.

Oquab, Maxime, Bottou, Léon, Laptev, Ivan, Sivic, Josef, et al. Weakly supervised object recognition with convolutional neural networks. In *NIPS*, 2014.

Pappas, Nikolaos and Popescu-Belis, Andrei. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *EMNLP*, pp. 455–466, 2014.

Pappas, Nikolaos and Popescu-Belis, Andrei. Explicit Document Modeling through Weighted Multiple-Instance Learning. *Journal of Artificial Intelligence Research*, 58:591–626, 2017.

Pinheiro, Pedro O and Collobert, Ronan. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pp. 1713–1721, 2015.

Qi, Charles R, Su, Hao, Mo, Kaichun, and Guibas, Leonidas J. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

Quellec, Gwenole, Cazuguel, Guy, Cochener, Beatrice, and Lamard, Mathieu. Multiple-instance learning for medical image and video analysis. *IEEE Reviews in Biomedical Engineering*, 2017.

Raffel, Colin and Ellis, Daniel PW. Feed-forward networks with attention can solve some long-term memory problems. 2015.

Ramon, Jan and De Raedt, Luc. Multi instance neural networks. In *ICML Workshop on Attribute-value and Relational Learning*, pp. 53–60, 2000.

Raykar, Vikas C, Krishnapuram, Balaji, Bi, Jinbo, Dundar, Murat, and Rao, R Bharat. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *ICML*, pp. 808–815, 2008.

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

Ruifrok, Arnout C and Johnston, Dennis A. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4): 291–299, 2001.

Scott, Stephen, Zhang, Jun, and Brown, Joshua. On generalized multiple-instance learning. *International Journal of Computational Intelligence and Applications*, 5(01): 21–35, 2005.

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

Wang, Fei, Jiang, Mengqing, Qian, Chen, Yang, Shuo, Li, Cheng, Zhang, Honggang, Wang, Xiaogang, and Tang, Xiaoou. Residual Attention Network for Image Classification. In *CVPR*, 2017.

Wang, Xinggang, Yan, Yongluan, Tang, Peng, Bai, Xiang, and Liu, Wenyu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2016.

Wei, Xiu-Shen, Wu, Jianxin, and Zhou, Zhi-Hua. Scalable algorithms for multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(4): 975–987, 2017.

Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron, Salakhudinov, Ruslan, Zemel, Rich, and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pp. 2048–2057, 2015.

Zaheer, Manzil, Kottur, Satwik, Ravanbakhsh, Siamak, Poczos, Barnabas, Salakhutdinov, Ruslan, and Smola, Alexander. Deep Sets. In *NIPS*. 2017.

Zhang, Cha, Platt, John C, and Viola, Paul A. Multiple instance boosting for object detection. In *NIPS*, pp. 1417–1424, 2006.

Zhang, Qi and Goldman, Sally A. Em-dd: An improved multiple-instance learning technique. In *NIPS*, pp. 1073–1080, 2002.

Zhou, Zhi-Hua, Sun, Yu-Yin, and Li, Yu-Feng. Multi-instance learning by treating instances as non-iid samples. In *ICML*, pp. 1249–1256, 2009.

Zhu, Wentao, Lou, Qi, Vang, Yeeleng Scott, and Xie, Xiaohui. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *MICCAI*, pp. 603–611, 2017.

6. Appendix

6.1. Deep MIL approaches

In Figure 6 we present three deep MIL approaches discussed in the paper.

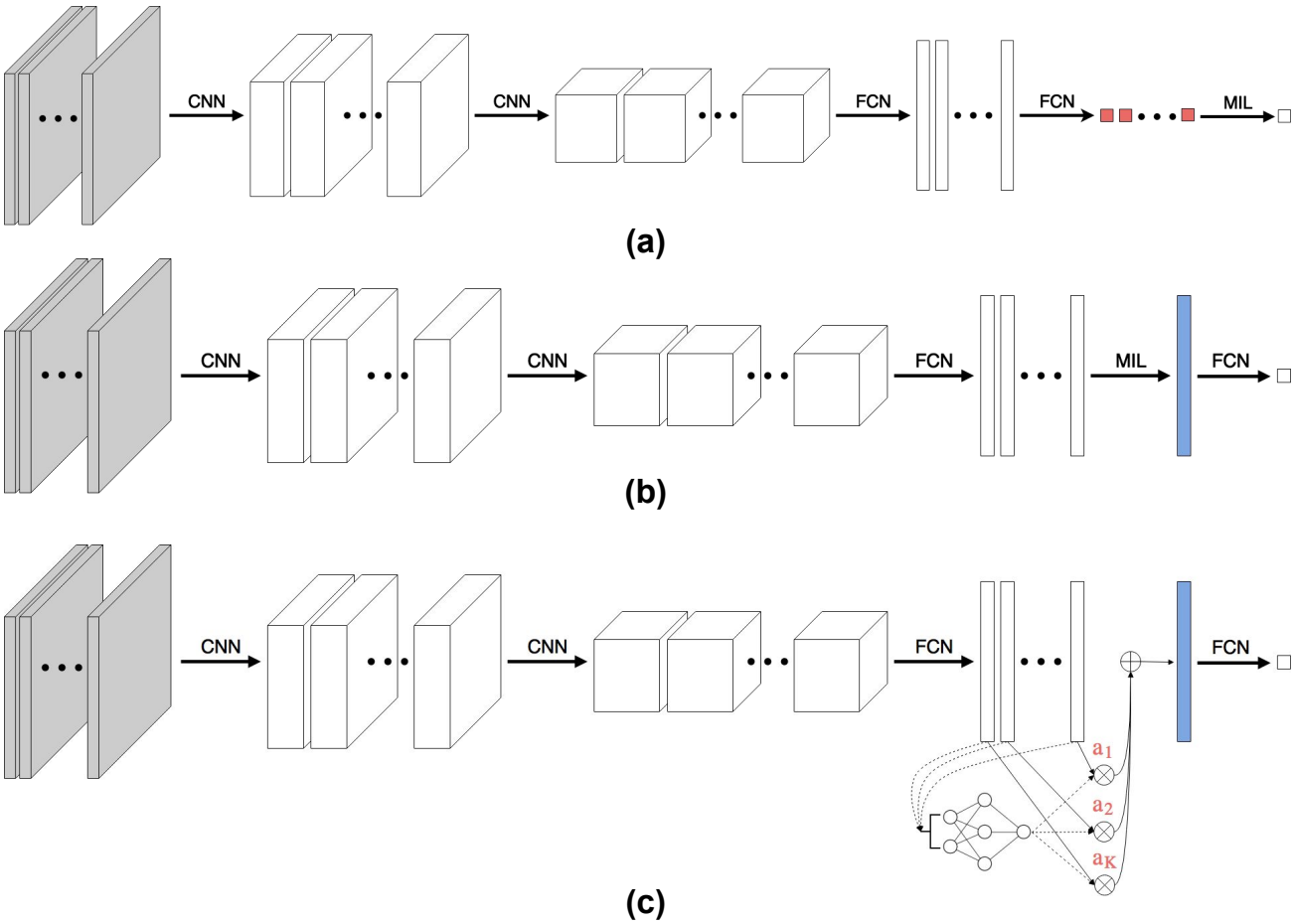


Figure 6. Deep MIL approaches: (a) the instance-based approach, (b) the embedding-based approach, (c) the proposed approach with the attention mechanism as the MIL pooling. Red color corresponds to instance scores, blue color depicts a bag vector representation. *Best viewed in color.*

6.2. Code

The implementation of our methods is available online at `https://github.com/AMLab-Amsterdam/AttentionDeepMIL`. All experiments were run on NVIDIA TITAN X Pascal with a batch size of 1 (= 1 bag) for all datasets.

6.3. Classical MIL datasets

Additional details In Table 1 a general description of the five benchmark MIL datasets used in the experiments is given. In Tables 5 and 6 we present architectures of the embedding-based and the instance-based models, respectively. We denote a fully-connected layer by 'fc' and the number of output hidden units is provided after a dash. The ReLU non-linearity was used. In Table 7 the details of the optimization (learning) procedure are given. We provide values of hyperparameters determined by the model selection procedure for which the highest validation performance was achieved.

Table 4. Overview of classical MIL datasets.

Dataset	# of bags	# of instances	# of features
Musk1	92	476	166
Musk2	102	6598	166
Tiger	200	1220	230
Fox	200	1302	230
Elephant	200	1391	230

Table 5. Classical MIL datasets: The embedding-based model architecture (Wang et al., 2016).

Layer	Type
1	fc-256 + ReLU
2	dropout
3	fc-128 + ReLU
4	dropout
5	fc-64 + ReLU
6	dropout
7	mil-max/mil-mean/mil-attention-64
8	fc-1 + sigm

Table 6. Classical MIL datasets: The instance-based model architecture (Wang et al., 2016).

Layer	Type
1	fc-256 + ReLU
2	dropout
3	fc-128 + ReLU
4	dropout
5	fc-64 + ReLU
6	dropout
7	fc-1 + sigm
8	mil-max/mil-mean

Table 7. Classical MIL datasets: The optimization procedure details (Wang et al., 2016).

Experiment	Optimizer	Momentum	Learning rate	Weight decay	Epochs	Stopping criteria
Musk1	SGD	0.9	0.0005	0.005	100	lowest validation error and loss
Musk2	SGD	0.9	0.0005	0.03	100	lowest validation error and loss
Tiger	SGD	0.9	0.0001	0.01	100	lowest validation error and loss
Fox	SGD	0.9	0.0005	0.005	100	lowest validation error and loss
Elephant	SGD	0.9	0.0001	0.005	100	lowest validation error and loss

6.4. MNIST-bags

Additional details In Tables 8 and 9 we present architectures of the embedding-based and the instance-based models for MNIST-BAGS, respectively. We denote a convolutional layer by 'conv', in brackets we provide kernel size, stride and padding, and the number of kernels is provided after a dash. The convolutional max-pooling layer is denoted by 'maxpool' and the pooling size is given in brackets. The ReLU non-linearity was used. In Table 10 the details of the optimization (learning) procedure for deep MIL approach are given. The details of the SVM are given in Table 11. We provide values of hyperparameters determined by the model selection procedure for which the highest validation performance was achieved.

Table 8. MNIST-bags: The embedding-based model architecture (LeCun et al., 1998).

Layer	Type
1	conv(5,1,0)-20 + ReLU
2	maxpool(2,2)
3	conv(5,1,0)-50 + ReLU
4	maxpool(2,2)
5	fc-500 + ReLU
6	mil-max/mil-mean/mil-attention-128
7	fc-1 + sigm

Table 9. MNIST-bags: The instance-based model architecture (LeCun et al., 1998).

Layer	Type
1	conv(5,1,0)-20 + ReLU
2	maxpool(2,2)
3	conv(5,1,0)-50 + ReLU
4	maxpool(2,2)
5	fc-500 + ReLU
6	fc-1 + sigm
7	mil-max/mil-mean

Table 10. MNIST-bags: The optimization procedure details.

Experiment	Optimizer	β_1, β_2	Learning rate	Weight decay	Epochs	Stopping criteria
All	Adam	0.9, 0.999	0.0005	0.0001	200	lowest validation error+loss

Table 11. MNIST-bags: SVM configuration.

Model	Features	Kernel	C	γ	Max iterations
MI-SVM	Raw pixel values	RBF	5	0.0005	200

Additional results In Tables 12, 13 and 14 we present the test AUC value for 10, 50 and 100 instances on average per a bag, respectively.

In Figure 7 a negative bag is presented. In Figure 8 a positive bag with a single '9' is given. In Figure 9 a positive bag with multiple '9's is presented. In all figures attention weights are provided and in the case of positive bags a red rectangle highlights positive instances.

Table 12. The test AUC for MNIST-BAGS with on average 10 instances per bag for different numbers of training bags.

# of training bags	50	100	150	200	300	400	500
Instance+max	0.553 ± 0.053	0.745 ± 0.100	0.960 ± 0.004	0.979 ± 0.001	0.984 ± 0.001	0.986 ± 0.001	0.986 ± 0.001
Instance+mean	0.663 ± 0.014	0.676 ± 0.012	0.694 ± 0.010	0.694 ± 0.017	0.709 ± 0.020	0.693 ± 0.023	0.712 ± 0.018
MI-SVM	0.697 ± 0.054	0.851 ± 0.009	0.862 ± 0.008	0.898 ± 0.014	0.926 ± 0.004	0.942 ± 0.002	0.948 ± 0.002
Embedded+max	0.713 ± 0.016	0.914 ± 0.011	0.954 ± 0.005	0.968 ± 0.001	0.980 ± 0.001	0.981 ± 0.003	0.986 ± 0.002
Embedded+mean	0.695 ± 0.026	0.841 ± 0.027	0.926 ± 0.004	0.953 ± 0.004	0.974 ± 0.002	0.980 ± 0.001	0.984 ± 0.002
Attention	0.768 ± 0.054	0.948 ± 0.007	0.949 ± 0.006	0.970 ± 0.003	0.980 ± 0.000	0.982 ± 0.001	0.986 ± 0.001
Gated Attention	0.753 ± 0.054	0.916 ± 0.013	0.955 ± 0.003	0.974 ± 0.002	0.980 ± 0.004	0.983 ± 0.002	0.987 ± 0.001

Table 13. The test AUC for MNIST-BAGS with on average 50 instances per bag for different numbers of training bags.

# of training bags	50	100	150	200	300	400	500
Instance+max	0.576 ± 0.059	0.715 ± 0.096	0.937 ± 0.045	0.992 ± 0.002	0.994 ± 0.001	0.997 ± 0.001	0.997 ± 0.001
Instance+mean	0.737 ± 0.014	0.744 ± 0.029	0.824 ± 0.012	0.813 ± 0.030	0.722 ± 0.021	0.728 ± 0.017	0.798 ± 0.011
MI-SVM	0.824 ± 0.067	0.946 ± 0.004	0.959 ± 0.002	0.967 ± 0.002	0.975 ± 0.001	0.976 ± 0.001	0.979 ± 0.001
Embedded+max	0.872 ± 0.039	0.984 ± 0.005	0.992 ± 0.001	0.996 ± 0.001	0.996 ± 0.001	0.997 ± 0.001	0.997 ± 0.001
Embedded+mean	0.841 ± 0.013	0.906 ± 0.046	0.983 ± 0.005	0.992 ± 0.001	0.996 ± 0.001	0.997 ± 0.001	0.997 ± 0.001
Attention	0.967 ± 0.010	0.982 ± 0.003	0.990 ± 0.002	0.993 ± 0.002	0.989 ± 0.003	0.994 ± 0.001	0.995 ± 0.001
Gated Attention	0.920 ± 0.042	0.977 ± 0.006	0.993 ± 0.003	0.991 ± 0.002	0.994 ± 0.002	0.995 ± 0.001	0.996 ± 0.001

Table 14. The test AUC for MNIST-BAGS with on average 100 instances per bag for different numbers of training bags.

# of training bags	50	100	150	200	300	400	500
Instance+max	0.543 ± 0.054	0.804 ± 0.107	0.899 ± 0.086	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Instance+mean	0.842 ± 0.023	0.855 ± 0.025	0.824 ± 0.014	0.896 ± 0.037	0.859 ± 0.029	0.899 ± 0.012	0.868 ± 0.016
MI-SVM	0.871 ± 0.060	0.991 ± 0.002	0.994 ± 0.002	0.996 ± 0.001	0.997 ± 0.001	0.998 ± 0.001	0.998 ± 0.001
Embedded+max	0.977 ± 0.009	0.999 ± 0.001	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Embedded+mean	0.959 ± 0.010	0.990 ± 0.003	0.998 ± 0.001	0.900 ± 0.089	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Attention	0.996 ± 0.001	0.998 ± 0.001	0.999 ± 0.000	0.998 ± 0.001	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Gated Attention	0.998 ± 0.001	0.999 ± 0.000	0.998 ± 0.001	0.998 ± 0.001	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

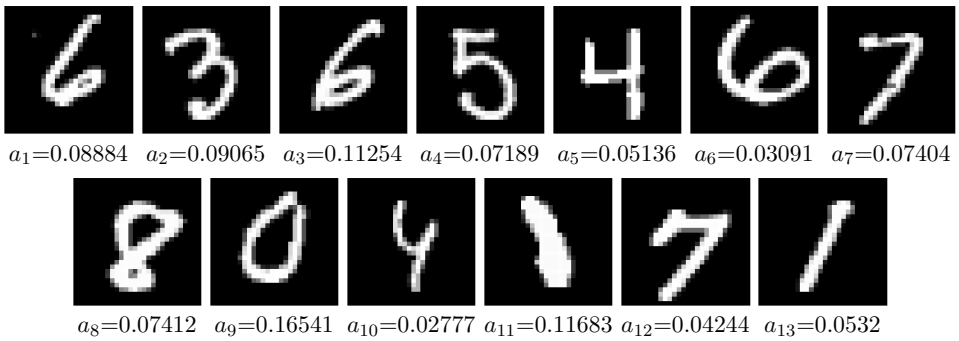


Figure 7. Example of attention weights for a negative bag.

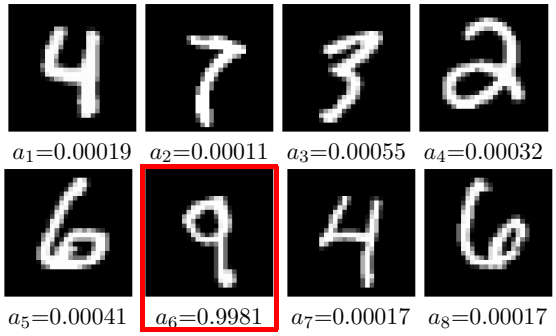


Figure 8. Example of attention weights for a positive bag containing a single '9'.

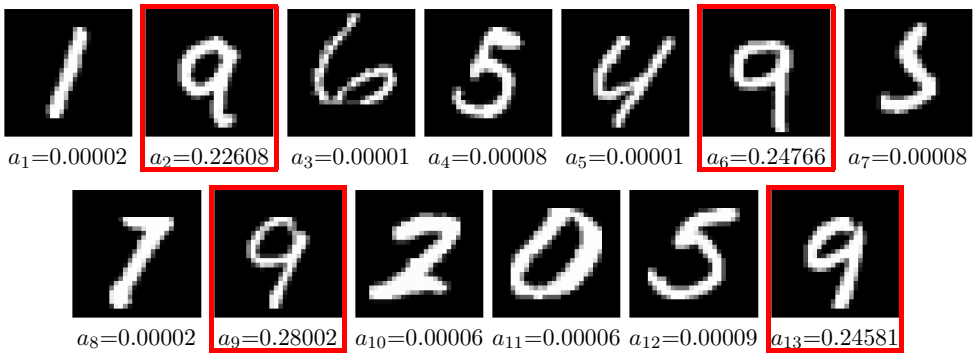


Figure 9. Example of attention weights for a positive bag containing multiple '9's.

6.5. Histopathology datasets

Data augmentation We randomly adjust the amount of H&E by decomposing the RGB color of the tissue into the H&E color space (Ruifrok & Johnston, 2001), followed by multiplying the magnitude of H&E for a pixel by two i.i.d. Gaussian random variables with expectation equal to one. We randomly rotate and mirror every patch. Lastly, we perform color normalization on every patch.

Additional details In Tables 15 and 16 we present architectures of the embedding-based and the instance-based models for histopathology datasets, respectively. In Table 17 the details of the optimization (learning) procedure for deep MIL approach are given. We provide values of hyperparameters determined by the model selection procedure for which the highest validation performance was achieved.

Table 15. Histopathology: The embedding-based model architecture (Sirinukunwattana et al., 2016).

Layer	Type
1	conv(4,1,0)-36 + ReLU
2	maxpool(2,2)
3	conv(3,1,0)-48 + ReLU
4	maxpool(2,2)
5	fc-512 + ReLU
6	dropout
7	fc-512 + ReLU
8	dropout
9	mil-max/mil-mean/mil-attention-128
10	fc-1 + sigm

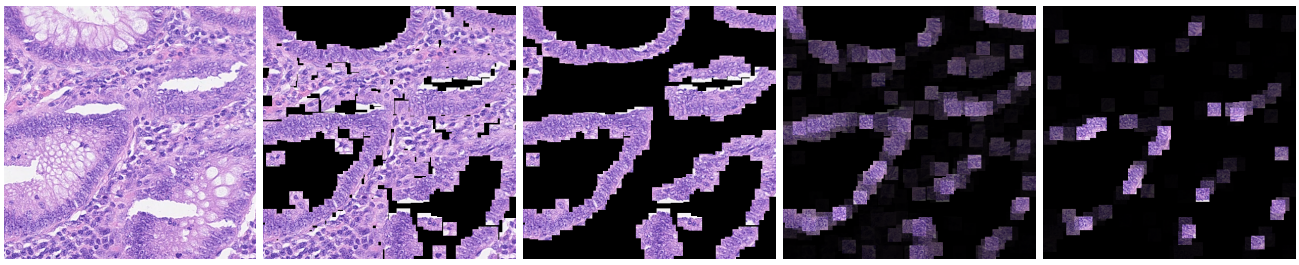
Table 16. Histopathology: The instance-based model architecture (Sirinukunwattana et al., 2016).

Layer	Type
1	conv(4,1,0)-36 + ReLU
2	maxpool(2,2)
3	conv(3,1,0)-48 + ReLU
4	maxpool(2,2)
5	fc-512 + ReLU
6	dropout
7	fc-512 + ReLU
8	dropout
9	fc-1 + sigm
10	mil-max/mil-mean

Table 17. Histopathology: The optimization procedure details.

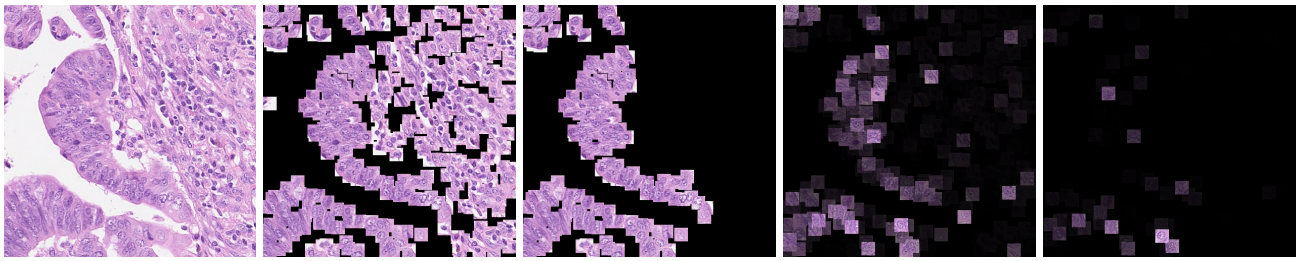
Experiment	Optimizer	β_1, β_2	Learning rate	Weight decay	Epochs	Stopping criteria
All	Adam	0.9, 0.999	0.0001	0.0005	100	lowest validation error+loss

Additional results In Figures 10, 11 and 12 five images are presented: (a) a full H&E image, (b) all patches containing cells, (c) positive patches, (d) a heatmap given by the attention mechanism, (e) a heatmap given by the Instance+max. We rescaled the attention weights and instance scores using $a'_k = (a_k - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$.



(a) (b) (c) (d) (e)

Figure 10. Colon cancer example 1: (a) H&E stained histology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Attention heatmap: Every patch from (b) multiplied by its attention weight. (e) Instance+max heatmap: Every patch from (b) multiplied by its score from the INSTANCE+max model. We rescaled the attention weights and instance scores using $a'_k = (a_k - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$.



(a) (b) (c) (d) (e)

Figure 11. Colon cancer example 2: (a) H&E stained histology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Attention heatmap: Every patch from (b) multiplied by its attention weight. (e) Instance+max heatmap: Every patch from (b) multiplied by its score from the INSTANCE+max model. We rescaled the attention weights and instance scores using $a'_k = (a_k - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$.

(a) (b) (c) (d) (e)

Figure 12. Colon cancer example 3: (a) H&E stained histology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Attention heatmap: Every patch from (b) multiplied by its attention weight. (e) Instance+max heatmap: Every patch from (b) multiplied by its score from the INSTANCE+max model. We rescaled the attention weights and instance scores using $a'_k = (a_k - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$.