Loss-Based Attention for Deep Multiple Instance Learning

Xiaoshuang Shi,[1] Fuyong Xing,[2] Yuanpu Xie,[1] Zizhao Zhang,[1] Lei Cui,[3] Lin Yang[1]
[1]University of Florida, Gainesville, FL, USA
[2]University of Colorado Denver, Denver, CO, USA
[3]Northwestern University, Xi'an, China
{xsshi2015, shampool, zizhaozhang}@ufl.edu,
fuyong.xing@ucdenver.edu, cuilei1989@163.com, lin.yang@bme.ufl.edu

Abstract

Although attention mechanisms have been widely used in deep learning for many tasks, they are rarely utilized to solve multiple instance learning (MIL) problems, where only a general category label is given for multiple instances contained in one bag. Additionally, previous deep MIL methods firstly utilize the attention mechanism to learn instance weights and then employ a fully connected layer to predict the bag label, so that the bag prediction is largely determined by the effectiveness of learned instance weights. To alleviate this issue, in this paper, we propose a novel loss based attention mechanism, which simultaneously learns instance weights and predictions, and bag predictions for deep multiple instance learning. Specifically, it calculates instance weights based on the loss function, e.g. softmax+cross-entropy, and shares the parameters with the fully connected layer, which is to predict instance and bag predictions. Additionally, a regularization term consisting of learned weights and cross-entropy functions is utilized to boost the recall of instances, and a consistency cost is used to smooth the training process of neural networks for boosting the model generalization performance. Extensive experiments on multiple types of benchmark databases demonstrate that the proposed attention mechanism is a general, effective and efficient framework, which can achieve superior bag and image classification performance over other state-of-the-art MIL methods, with obtaining higher instance precision and recall than previous attention mechanisms. *Source codes are available on* https://github.com/xsshi2015/Loss-Attention.

Introduction

Multiple instance learning is a significant research topic in machine learning and computer vision communities, and it has been widely used in many real-world applications, such as image categorization or retrieval, gene expression, face detection and medical imaging (Wei and Zhou 2016) (Wang et al. 2018) (Hou et al. 2016) (Shi et al. 2017) (Shi et al. 2018) , where only a general statement of the category is given for multiple instances (Ilse, Tomczak, and Welling 2018), e.g. one bag contains tens or hundreds of instances, while it is usually described by a single bag label and there is no label information associated with instances.

The main goal of MIL is to learn a classification model with training bag labels in order to predict a test bag label. An additional challenge task is to interpret the significance of each instance for bag classification. Numerous algorithms have been applied to MIL, and they can be roughly classified into three paradigms (Amores 2013): bag-space, instance-space and embedded-space. Bag-space paradigm (Cheplygina, Tax, and Loog 2015) treats each bag as a whole and exploits their relations for classification. Instance-space paradigm (Ramon and De Raedt 2000) trains instance-level classifiers and aggregates their responses for bag classification. Embedded-space paradigm (Andrews, Tsochantaridis, and Hofmann 2003) (Chen, Bi, and Wang 2006) first embeds all instances in one bag into a compact low-dimensional representation and then feeds it to a bag-level classifier. Among these three paradigms, only instance-space paradigm (Ramon and De Raedt 2000) (Zhang, Platt, and Viola 2006) is able to interpret the contribution of each instance to bag classification. Unfortunately, instance-space paradigm often exhibits inferior performance to the other paradigms (Kandemir and Hamprecht 2015).

In oder to interpret the significance of instances and meanwhile achieve desired bag classification accuracy, several attention based MIL algorithms (Pappas and Popescu-Belis 2017) (Ilse, Tomczak, and Welling 2018) embed attention mechanisms into neural networks to learn instance weights and classify bags. These algorithms usually introduce auxiliary layers to learn instance weights and then utilize another fully connected layer to produce bag predictions, and thus the bag prediction is largely determined by the effectiveness of learned instance weights. Unfortunately, the attention mechanism often assigns a large weight to the instance, which has a different label from that of the bag, thereby misleading the network and decreasing its performance on bag prediction and instance interpretation.

To alleviate this issue, in this paper, we propose a novel loss based attention mechanism, which connects the attention mechanism with the loss function to simultaneously learn instance weights and predictions, and bag predictions. To the best of our knowledge, the proposed method is the first work to directly connect the attention mechanism with the loss function for multiple instance learning. The major

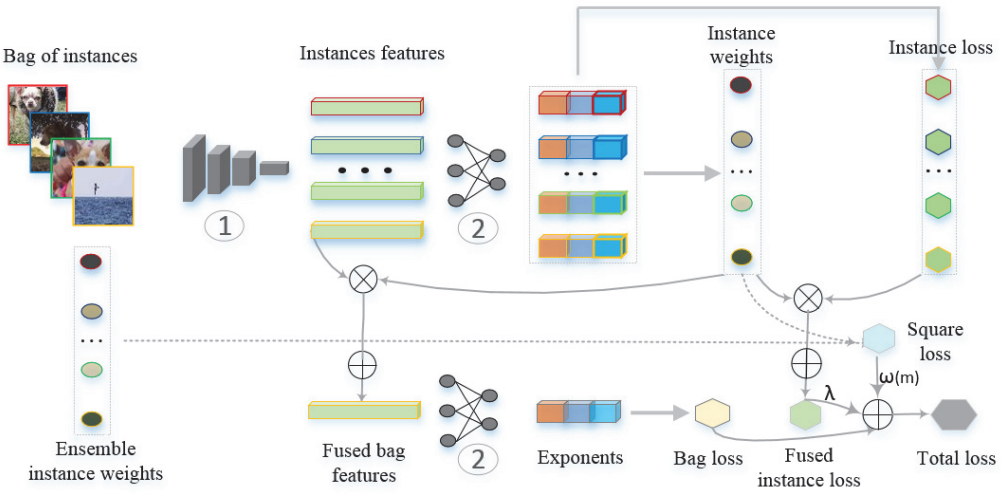


Figure 1: An example of the structure pass of the proposed loss based attention mechanism. Given a bag with multiple instances, ① represents the layers of a traditional or convolutional neural network in order to obtain instance features, ② denotes a fully-connected layer and then an exponential function. *Note that the instance and bag features utilize the same parameters to obtain their exponents.* The fused bag features is the sum of dot products between instance features and their corresponding attention probabilities, and similar definitions are used for the fused instance loss. *Bag loss is obtained by the softmax+cross-entropy functions to predict bag labels, fused instance loss is a regularization term to predict instance labels, and square loss is the consistency cost to smooth the training process.*

contributions of this paper are summarized as follows:

- We connect the attention mechanism with the loss function, e.g. softmax+cross-entropy, by calculating the instance weights based on the loss function and sharing the same parameters with the fully connected layer, to simultaneously learn instance weights and predictions, and bag predictions. Additionally, we propose a regularization term composed of learned instance weights and cross-entropy functions to further enhance the connection between instance weights and the loss, and introduce a consistency cost to smooth the training process. For clarity, we show an example of bag classification to illustrate the main structure pass of the proposed attention mechanism in Figure 1.
- We theoretically prove: (i) Only using the attention mechanism with the softmax and cross-entropy functions for bag classification will produce low instance recall; (ii) The newly introduced regularization term can boost the instance recall. These two statements have also been verified in our experiments (please refer to Figure 3).
- Extensive experiments on multiple types of datasets demonstrate the generality, effectiveness and efficiency of the proposed loss based attention mechanism, which not only outperforms recent state-of-the-art MIL methods on bag and image classification, but also achieves higher precision and recall of instances with large weights than previous approaches.

Related Work

In this section we briefly introduce the related work: MIL-based neural networks including traditional neural networks and convolutional neural networks (CNNs), and attention algorithms for MIL.

MIL based neural networks. MIL-based traditional neural networks (Li, Gondra, and Liu 2012) usually utilize the feature representation as instance given, while MIL-based CNNs (Pathak et al. 2014) (Pinheiro and Collobert 2015) can learn feature representations through multiple convolutional layers to further improve the prediction accuracy. Most of MIL-based neural networks (Feng and Zhou 2017) adopt max-pooling to perform back propagation along the instance with the maximum response. BP-MIP (Zhou and Zhang 2002) performs the back propagation on the instance with the maximum training error, (Oquab et al. 2014) uses global max-pooling to search the best-scoring candidate object position, and (Pathak et al. 2014) computes the multi-class logistic loss at maximum predictions for semantic segmentation. Since max-pooling leads to one instance per bag being trained in one iteration, it might be not robust to search the significant instance and even predict bag labels. To alleviate this issue, some alternative pooling functions, such as Noisy-or (Zhang, Platt, and Viola 2006), ISR (Keeler, Rumelhart, and Leow 1991), generalized mean and LSE (Ramon and De Raedt 2000) (Kraus, Ba, and Frey 2016), have been embedded into neural networks. However, the flexibility of these functions is restricted, because they are pre-defined and not trainable. To attain learnable pooling, adaptive pooling function (Zhou et al. 2017) and a fully-connect CRF (Chen et al. 2014) have been developed to smooth the prediction. In addition to the pooling functions, expectation-maximization (EM) methods (Papandreou et al. 2015) (Hou et al. 2016) combining with CNNs have been used for weakly supervised semantic segmenta-

tion and whole slide pathology image classification.

Attention for MIL. The attention mechanism with deep learning has been widely applied to many tasks, such as image captioning (Xu et al. 2015) (Zhang et al. 2019), classification (Wang et al. 2017), and model interpretation (Zhang and Zhu 2018) (Xu et al. 2018) (Xu et al. 2019). However, very little effort focuses on attention mechanisms for MIL. Multiple instance regression (MIR) (Pappas and Popescu-Belis 2014) learns the weight of instances by using them as parameters of an auxiliary linear regression model. Weighted multiple instance regression (WMIR) (Pappas and Popescu-Belis 2017) follows the idea in MIR but learns instance weights via a single neural network layer. Attention based deep multiple instance learning (ADMIL) (Ilse, Tomczak, and Welling 2018) proposes a two-layered neural network to learn instance weights and uses the sigmoid function to predict bag probability. Unlike previous methods using the parameters of auxiliary layers to learn instance weights and then employing another layers to learn bag predictions, our proposed attention mechanism directly exploits the parameters in a fully connected layer to connect with the loss function, for learning instance weights and predictions, and bag predictions simultaneously.

Loss-based Attention Mechanism for Deep MIL

In this section, we present the proposed loss based attention mechanism derived from the softmax and cross-entropy functions, and then theoretically analyze its several important characteristics.

Cross-entropy in Neural Networks

Given one training image $\mathbf{x}$ and its corresponding label $y \in \{0, 1, \cdots, K-1\}$, where K is the number of classes. Let $f(\cdot)$ represent a neural network and $\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^K$ be the final output, i.e. the prediction class vector of $\mathbf{x}$. By using the softmax function, the estimated class probability of $\mathbf{x}$ belonging to the k-th class is:

$$q_k = \frac{exp(z_k)}{\sum_{c=0}^{K-1} exp(z_c)}, \tag{1}$$

where $exp(\cdot)$ represents the exponential function.

Suppose that $p_c \in \{0, 1\}$ denotes the true class probability of $\mathbf{x}$ belonging to the c-th class, we utilize the cross-entropy function to measure the dissimilarity between the true class probability $\mathbf{p} \in \{0, 1\}^K$ and the estimated class probability $\mathbf{q} \in \{0, 1\}^K$ (De Boer et al. 2005):

$$L(\mathbf{p}, \mathbf{q}) = -\sum_{c=0}^{K-1} p_c \, log \, q_c. \tag{2}$$

Because of $\mathbf{p} \in \{0, 1\}^K$ and $\sum_{c=0}^{K-1} p_c = 1$, when $\mathbf{x}$ belongs to the k-th class, i.e. $p_k = 1$ and $\sum_{c=0, c\neq k}^{K-1} p_c = 0$, Eq. (2) equals:

$$L(\mathbf{p}, \mathbf{q}) = -log \, \frac{exp(z_k)}{\sum_{c=0}^{K-1} exp(z_c)}. \tag{3}$$

Loss-based Attention Mechanism

Given a set of training images $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ representing n bags, each bag $\mathbf{X}_i = \{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,n_i}\}$ consists of n_i instances and $y_i \in \{0, 1, \cdots, K-1\}$ is the corresponding bag label, where $\mathbf{x}_{i,t}$ is the t-th instance in the i-th bag and $y_{i,t} \in \{0, 1, \cdots, K-1\}$ is the corresponding instance label. Similar to (Ilse, Tomczak, and Welling 2018), we suppose that a bag contains at most two kinds of instance labels including $y_{i,t} = 0$ and one kind of other labels $y_{i,t} \in \{1, \cdots, K-1\}$, and the relationship between y_i and $y_{i,t}$ is:

$$y_i = \max_{1 \le t \le n_i} y_{i,t}. \tag{4}$$

Suppose that each instance of the i-th bag has the same significance, and an L-layer neural network applies a mean operator to low-dimensional representations of instances. Given an instance $\mathbf{x}_{i,t}$, let $\mathbf{h}_{i,t}^l$ $(1 \le l \le L-1)$ be its feature representation at the l-th layer. For example, $\mathbf{h}_{i,t}^{L-1} = g(\mathbf{h}_{i,t}^{L-2}) \in \mathbb{R}^d$ is the output of the $L-1$-th layer and the input of the L-th layer, where $g(\cdot)$ is an activation function and $\mathbf{h}_{i,t}^{L-2}$ is the input of the $L-1$-th layer. Suppose that $\mathbf{W} \in \mathbb{R}^{d \times K}$ is a projection matrix and $\mathbf{b} \in \mathbb{R}^K$ is a bias vector in the L-th layer, the final output of the neural network is $\mathbf{z}_i = \mathbf{h}_i^{L-1}\mathbf{W} + \mathbf{b}$, where $\mathbf{h}_i^{L-1}$ is obtained by a mean operator, e.g. $\mathbf{h}_i^{L-1} = \frac{1}{n_i}\sum_{t=1}^{n_i} \mathbf{h}_{i,t}^{L-1}$, and $\mathbf{z}_i \in \mathbb{R}^K$ represents the prediction class vector of the i-th bag. After obtaining $\mathbf{z}_i$, the loss function Eq. (3) can be utilized to learn model parameters.

When the significance of instances in the i-th bag is different, we introduce the proposed attention mechanism as follows:

$$\begin{gathered} \alpha_{i,j} = \frac{\sum_{c=0}^{K-1} exp(\mathbf{h}_{i,j}^{L-1}\mathbf{w}_c + b_c)}{\sum_{t=1}^{n_i}\sum_{c=0}^{K-1} exp(\mathbf{h}_{i,t}^{L-1}\mathbf{w}_c + b_c)} \\ \mathbf{h}_{i,j}^{L-1} \leftarrow \alpha_{i,j}\mathbf{h}_{i,j}^{L-1} \\ \mathbf{h}_i^{L-1} = \sum_{t=1}^{n_i} \mathbf{h}_{i,t}^{L-1}, \end{gathered} \tag{5}$$

where $\alpha_{i,j}$ is the weight of the j-th $(1 \le j \le n_i)$ instance in the i-th bag, $\mathbf{w}_c \in \mathbb{R}^d$ is the c-th column vector of $\mathbf{W}$ and $b_c \subset \mathbf{b}$ is a bias. After calculating $\mathbf{h}_i^{L-1}$ by using Eq. (5), $\mathbf{z}_i$ can be calculated by using $\mathbf{z}_i = \mathbf{h}_i^{L-1}\mathbf{W} + \mathbf{b}$. Based on Eq. (3), suppose that the i-th bag belongs to the k-th class, we present the following loss function:

$$\begin{gathered} L = L_1 + L_2 \\ = -log\frac{exp(\mathbf{h}_i^{L-1}\mathbf{w}_k + b_k)}{\sum_{c=0}^{K-1} exp(\mathbf{h}_i^{L-1}\mathbf{w}_c + b_c)} \\ -\lambda\sum_{t=1}^{n_i} \alpha_{i,t} log\frac{exp(\mathbf{h}_{i,t}^{L-1}\mathbf{w}_k + b_k)}{\sum_{c=0}^{K-1} exp(\mathbf{h}_{i,t}^{L-1}\mathbf{w}_c + b_c)}, \end{gathered} \tag{6}$$

where the first term (bag loss) is the main objective function L_1 to predict bag labels, the second term (fused instance loss) is the regularization term L_2 to predict instance labels, and λ is a non-negative constant to balance the bag and instance predictions. Note that because of $\mathbf{z}_i = \mathbf{h}_i^{L-1}\mathbf{W} + \mathbf{b}$ and $\mathbf{h}_i^{L-1} = \sum_{t=1}^{n_i} \mathbf{h}_{i,t}^{L-1}$, we have $z_{i,k} = \mathbf{h}_i^{L-1}\mathbf{w}_k + b_k$,

$z_{i,t,k} = \mathbf{h}_{i,t}^{L-1}\mathbf{w}_k + b_k$ and $z_{i,k} = \sum_{t=1}^{n_i} z_{i,t,k}$. Similar definitions are applied to $z_{i,c}$ and $z_{i,t,c}$ for any $t \in \{1, 2, \cdots, n_i\}$ and $c \in \{0, 1, 2, \cdots, K-1\}$.

Eqs. (5)-(6) are mainly inspired by: (i) The weights of instances with different labels from the bag label should be approximately equal to zeros; (ii) When the loss of the regularization term in Eq. (6) is close to zero, i.e. $L_2 \to 0$, if the j-th instance has a large weight ($\alpha_{i,j} >> 0$) and belongs to the k-th class, there must exist $exp(z_{i,j,k}) \approx \sum_{c=0}^{K-1} exp(z_{i,j,c})$. Additionally, if the r-th instance has a very small weight ($\alpha_{i,r} \to 0$), i.e. $\frac{\sum_{c=0}^{K-1} exp(z_{i,r,c})}{\sum_{t=1}^{n_i}\sum_{c=0}^{K-1} exp(z_{i,t,c})} \approx 0$, which means that $\sum_{c=0}^{K-1} exp(z_{i,r,c})$ and $exp(z_{i,r,k})$ can be neglected. They suggest that for the j-th instance with a weight $\alpha_{i,j} >> 0$, it can be calculated by: $\alpha_{i,j} = \frac{\sum_{c=0}^{K-1} exp(z_{i,j,c})}{\sum_{t=1}^{n_i}\sum_{c=0}^{K-1} exp(z_{i,t,c})} \approx \frac{exp(z_{i,j,k})}{\sum_{t=1}^{n_i} exp(z_{i,t,k})}$.

Recently, many self-ensembling based semi-supervised deep classification methods (Laine and Aila 2016) (Miyato et al. 2018) illustrate that smoothing the training process of neural networks can boost the model generalization performance. To smooth the training process, similar to (Laine and Aila 2016), we first create an ensemble target for each instance weight. Specifically, for $\alpha_{i,t}$, we accumulate it into an ensemble weight $\tilde{\alpha}_{i,t}$ by $\tilde{\alpha}_{i,t} = \beta\tilde{\alpha}_{i,t} + (1-\beta)\alpha_{i,t}$ in each training epoch, where $\beta \geq 0$ is to determine how far the ensemble weight reaches into training history. Then we utilize a consistency cost (square loss) $\|\alpha_{i,t} - \tilde{\alpha}_{i,t}\|_2^2$ to form a consensus prediction of the t-th instance weight in the i-th bag. Combining the consistency cost with Eq. (6), we obtain the proposed loss function as follows:

$$\begin{aligned} L_p &= L_1 + L_2 + L_3 \\ &= -log\frac{exp(\mathbf{h}_i^{L-1}\mathbf{w}_k + b_k)}{\sum_{c=0}^{K-1} exp(\mathbf{h}_i^{L-1}\mathbf{w}_c + b_c)} \\ &-\lambda\sum_{t=1}^{n_i}\alpha_{i,t} log\frac{exp(\mathbf{h}_{i,t}^{L-1}\mathbf{w}_k + b_k)}{\sum_{c=0}^{K-1} exp(\mathbf{h}_{i,t}^{L-1}\mathbf{w}_c + b_c)} \\ &+\omega(m)\sum_{t=1}^{n_i}\|\alpha_{i,t} - \tilde{\alpha}_{i,t}\|_2^2. \end{aligned} \tag{7}$$

where $\omega(m)$ is an unsupervised ramp-up function depending on the epoch number m to gradually enhance the weight of the consistency cost L_3.

Attention Mechanism Analysis

In this subsection, we first analyze the relationship between the proposed attention mechanism and the popular max-pooling operator in Theorem 1. Next, we discuss a lower bound of the main objective function L_1 of Eq. (7) in Theorem 2 and explain the motivation of adding a regularization term L_2. Finally, we present a lower bound of the regularization term L_2 in Theorem 3 and show the relationship between the instance weight and the loss of the regularization term in Theorem 4, which demonstrates that the regularization term can boost the instance recall. All proofs of Theorems 1-4 are provided in the supplemental material.

Theorem 1. *In one bag, when the weight of one instance is close to 1, e.g. $\alpha_{i,j} \to 1$, the probability of a bag being the k-th class is approximately equal to that of this instance belonging to the k-th class.*

Theorem 1 suggests that when the weight of one instance is close to 1 in one bag, the attention mechanism can have almost the same effect as max-pooling to select only one instance. However, different from the max-pooling operator that usually updates parameters through only one instance, the attention mechanism updates the weight and model parameters through all instances, and thus it is more smooth than the max-pooling operator, thereby yielding better prediction performance. .

Theorem 2. *Suppose that the i-th bag belongs to the k-th class and contains n_i instances, $q_{i,t,k} = \frac{exp(z_{i,t,k})}{\sum_{c=0}^{K-1} exp(z_{i,t,c})}$ denotes the estimated class probability of the t-th instance belonging to the k-th class. For the main objective function L_1 in Eq. (7), there exists:*

$$L_1 \geq \frac{\sum_{c=0,c\neq k}^{K-1}\prod_{t=1}^{n_i}(\frac{q_{i,t,c}}{q_{i,t,k}})^{\alpha_{i,t}}}{1+\sum_{c=0,c\neq k}^{K-1}\prod_{t=1}^{n_i}(\frac{q_{i,t,c}}{q_{i,t,k}})^{\alpha_{i,t}}}. \tag{8}$$

Eq. (8) suggests that when $L_1 \to 0$, at least one instance in a bag belongs to the k-th class. Specifically, for any one of instances, if it has $\frac{q_{i,t,c}}{q_{i,t,k}} \to 0$ ($c \in \{0, 1, 2, \cdots, K-1\}$ and $c \neq k$) and $\alpha_{i,t} >> 0$, then $L_1 \to 0$. *However, $L_1 \to 0$ cannot theoretically guarantee that more than one instance belong to the k-th class, thereby leading to the low recall of instances.* To address this issue, i.e. to ensure that more instances with large weights share the labels with the bag, we propose the attention mechanism and add a regularization term L_2 in Eq. (7). Because the effect of the term L_2 depends on the value of λ, we analyze the relations between its value and the instance recall by the following theorems.

Theorem 3. *Suppose that the i-th bag belongs to the k-th class and contains n_i instances, for the regularization term L_2 in Eq. (7), there exists:*

$$L_2 \geq \lambda\frac{\sum_{t=1}^{n_i}\sum_{c=0,\ c\neq k}^{K-1} exp(z_{i,t,c})}{\sum_{t=1}^{n_i}\sum_{c=0}^{K-1} exp(z_{i,t,c})}. \tag{9}$$

Theorem 4. *Suppose that $\alpha_{i,j}$ is the j-th instance weight in the i-th bag, which belongs to the k-th class, if $\alpha_{i,j} > \frac{2L_2}{\lambda}$, the j-th instance will be predicted to the k-th class.*

Theorem 4 suggests that λ plays a significant role in adjusting the number of instances, which share the labels with its corresponding bag. *Specifically, given fixed values of L_2 and $\alpha_{i,j}$, the larger λ, the more instances belonging to the same class as its bag, i.e. the higher recall of instances.*

Experiments

We evaluate the proposed method, referred to as Loss-Attention, on multiple benchmark MIL datasets (MUSK1, MUSK2, FOX, TIGER and ELEPHANT), MNIST-based and CIFAR-10-based MIL datasets, CIFAR-10 and Tiny ImageNet image databases. Following (Ilse, Tomczak, and Welling 2018) (Wang et al. 2018), we adopt 10-fold-cross-validation and repeat five times per experiment for MIL and histopathology datasets. For experiments on MNIST-bags, we utilize a fixed division into training and test sets. We compare Loss-Attention to recent state-of-the-art methods and basic algorithms: "Instance+max/mean" and "Embedding+max/mean'. They denote the instance-level and

Table 1: Results on benchmark MIL databases. We run each experiment five times and report the average classification accuracy (mean ± standard deviation). We bold the best accuracy on each database and highlight the second best results via underlines.

Method	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
mi-SVM	0.874±N/A	0.836±N/A	0.582±N/A	0.784±N/A	0.822±N/A
MI-SVM	0.779±N/A	0.843±N/A	0.578±N/A	0.840±N/A	0.843±N/A
MI-Kernel	0.880±N/A	0.893±N/A	0.603±N/A	0.842±N/A	0.843±N/A
EM-DD	0.849±0.098	0.869±0.108	0.609±0.101	0.730±0.096	0.771±0.097
mi-Graph	0.889±0.073	0.903±0.086	0.620±0.098	0.860±0.083	0.869±0.078
miVLAD	0.871±0.097	0.872±0.095	0.620±0.098	0.811±0.087	0.850±0.080
miFV	0.909±0.089	0.884±0.094	0.621±0.109	0.813±0.083	0.852±0.081
mi-Net	0.889±0.088	0.858±0.110	0.613±0.078	0.824±0.076	0.858±0.083
MI-Net	0.887±0.091	0.859±0.102	0.622±0.084	0.830±0.072	0.862±0.077
MI-Net with DS	0.894±0.093	0.874±0.097	0.630±0.080	0.845±0.087	0.872±0.072
MI-Net with RC	0.898±0.097	0.873±0.098	0.619±0.104	0.836±0.083	0.857±0.089
Attention	0.892±0.090	0.858±0.106	0.615±0.096	0.839±0.054	0.868±0.054
Gated-Attention	0.900±0.088	0.863±0.094	0.603±0.068	0.845±0.046	0.857±0.064
Loss-Attention	**0.917**±0.066	**0.911**±0.063	**0.712**±0.074	**0.897**±0.065	**0.900**±0.069

embedding-level neural networks with MIL pooling layers using the max or mean operator, respectively. For fairness, they utilize the same architectures as the proposed method. To evaluate the performance of MIL methods, we adopt the following metrics: classification accuracy, precision, recall, F-score, and the area under the receiver operator operating characteristic curve (AUC).

MIL datasets classification

We conduct experiments on five popular MIL datasets: MUSK1, MUSK2, FOX, TIGER and ELEPHANT. Because these databases contain precomputed features belonging to two classes and only a small number of instances and bags, it is usually difficult for neural networks to attain the same good performance as traditional state-of-the-art methods. The detailed information about features, instances and bags in each dataset is shown in Table A1 of the supplemental material. MUSK1 and MUSK2 are used to predict drug activity, and the molecule has the drug effect if and only if one or more of the conformations of one molecule bind to the target binding site. One molecule contains multiple shapes, and a bag is composed of shapes belonging to the same molecule (Dietterich, Lathrop, and Lozano-Pérez 1997). The remaining three datasets, FOX, TIGER and ELEPHANT, consist of features extracted from images. Each bag contains a set of segments obtained from one image. Positive bags are constituted by images with the animal of interest, and negative bags are made up of images with other animals (Andrews, Tsochantaridis, and Hofmann 2003). Following (Ilse, Tomczak, and Welling 2018) (Wang et al. 2018), we utilize the same architecture as the MI-NET model (Wang et al. 2018) except the attention layer and the final layer for the proposed loss function. The details of architectures, the parameter λ, ramp-up function $\omega(m)$, optimizer and hyperparameters are shown in the supplemental material (Tables A2, A3 and A4).

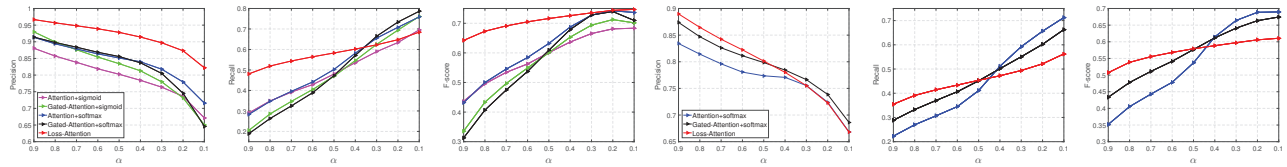
Results and discussion: Table 1 shows the classification accuracy of Mi-SVM and MI-SVM (Andrews, Tsochantaridis, and Hofmann 2003), MI-Kernel (Gärtner et al. 2002), EM-DD (Zhang and Goldman 2002), mi_Graph (Zhou, Sun, and Li 2009), miVLDA and miFV (Wei, Wu, and Zhou 2017), mi-Net, MI-Net, MI-Net with DS, MI-Net with RC (Wang et al. 2018), Attention and Gated-Attention (Ilse, Tomczak, and Welling 2018), and the proposed Loss-Attention method. It illustrates that Loss-Attention consistently achieves the best average accuracy among all algorithms on the five datasets. These results demonstrate the effectiveness and efficiency of the proposed method.

Table 2: Results on MNIST-bags with different numbers of training bags. Each experiment is repeated 50 times and average results are reported.

# of training bags	50	100	150	200
Binary (AUC)				
Attention+sigmoid	0.858	0.901	0.942	0.961
Gated-Attention+sigmoid	0.869	0.912	0.966	0.968
Instance+max	0.904	0.947	0.952	0.954
Instance+mean	0.800	0.851	0.913	0.939
Embedding+max	0.805	0.943	0.962	0.975
Embedding+mean	0.794	0.847	0.904	0.934
Attention+softmax	0.914	0.963	0.977	**0.984**
Gated-Attention+softmax	0.908	0.959	0.973	0.979
Loss-Attention	**0.931**	**0.969**	**0.978**	**0.984**
Multi-class (Accuracy)				
Instance+max	0.477	0.750	0.846	0.887
Instance+mean	0.587	0.774	0.865	0.917
Embedding+max	0.635	0.796	0.879	0.918
Embedding+mean	0.582	0.774	0.869	0.920
Attention+softmax	0.753	0.885	**0.923**	0.938
Gated-Attention+softmax	0.720	0.869	0.911	0.930
Loss-Attention	**0.765**	**0.892**	0.917	**0.939**

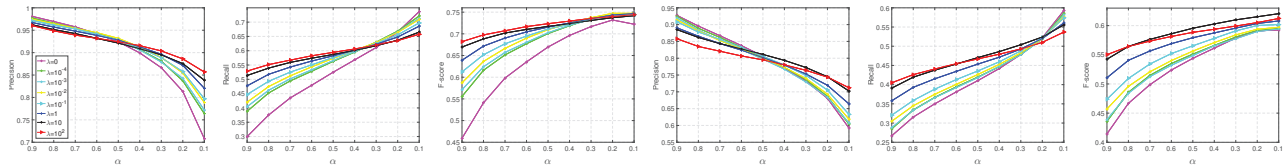
MNIST-based MIL datasets classification

Here, we create challenging datasets for binary and multi-class classification using images from the popular MNIST dataset to evaluate Loss-Attention, the basic algorithms: Instance+max/mean and Embedding+max/mean, and the comparative ones: Attention and Gated-Attention (Ilse, Tomczak, and Welling 2018). Note that all basic algorithms utilize the softmax+cross-entropy functions for bag classification. Unlike classic MIL datasets using precomputed features to represent instances, created bags consist of a random number of 28×28 grayscale images selected from the MNIST dataset. The number of images in a bag is Gaussian-distributed, with the mean bag size and the variance being 10 and 2, respectively. We build training sets with 50, 100, 150 and 200 bags, respectively, and a test set containing 1,000



(a) @Bi/Precision (b) @Bi/Recall (c) @Bi/F-score (d) @Multi/Precision (e) @Multi/Recall (f) @Multi/F-score

Figure 2: The instance precision, recall and F-score of different attention based MIL algorithms using 50 training bags from MNIST-based MIL datasets for binary and multi-class classification.



(a) @Bi/Precision (b) @Bi/Recall (c) @Bi/F-score (d) @Multi/Precision (e) @Multi/Recall (f) @Multi/F-score

Figure 3: The instance precision, recall and F-score of Loss-Attention with different values of λ on MNIST-bags for binary and multi-class classification.

bags. For a binary classification scenario, following (Ilse, Tomczak, and Welling 2018), a bag is labeled as positive if it contains one or more images with the digit '9', otherwise negative. For a multi-class classification scenario, target numbers are the digits '3', '5' and '9' and one bag at most contains one of these three digits. The bags containing these digits of '3', '5' and '9' are given labels '1', '2' and '3', respectively. If one bag is not composed of anyone of these three digits, the bag is labeled as '0'. For binary and multi-class classification experiments, we utilize AUC and classification accuracy as evaluation metrics, respectively. To evaluate the models' interpretation capability, we quantitatively investigate their retrieval performance on target digits using precision, recall and F-score metrics. The architectures used in this experiment are on the basis of a LeNet5 model (LeCun et al. 1998). The details of architectures, the parameter λ, ramp-up function $\omega(m)$, optimizer and hyperparameters are shown in the supplemental material (Tables A8 and A9).

Results and discussion: Table 2 displays the AUC and classification accuracy of the basic algorithms and three attention mechanisms on MNIST-bags. Attention and Gated-Attention using the softmax function can obtain higher AUC than that using the sigmoid function. Additionally, all attention mechanisms have superior performance over the basic algorithms. Moreover, Loss-Attention performs better than Attention and Gated-Attention on binary and multi-class classification tasks in most of the cases. To evaluate the interpretation capability of attention based MIL algorithms, Figure 2 presents their instance precision, recall and F-score on different values of α with 50 training bags, where α denotes the weight of instances. When $\alpha > 0.5$, Loss-Attention can achieve higher precision, recall and F-score of instances than Attention and Gated-Attention. It suggests that Loss-Attention can better interpret the instance with a large weight, e.g. $\alpha > 0.5$, which are more attractive in practice. When using other numbers of training bags, we can observe similar findings.

Ablation study and parameter analysis: Because λ in Loss-Attention plays a significant role in instance interpretation, here we present the influence of different values of λ on instance interpretation and verify the proposed theorems. We conduct binary and multi-class classification experiments on MNIST based MIL datasets, by using a training set with 50 bags and a test set with 1,000 bags. Figure 3 displays the instance precision, recall and F-score of Loss-Attention with $\lambda \in \left[0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\right]$. Note that when $\lambda = 0$, it means that the regularization term L_2 is removed. Figure 3(a)-(b) and (d)-(e) show that for large α, e.g. $\alpha > 0.5$, the smaller value of λ, the higher precision but the much lower recall. Because the loss of the main objective in Eq. (7) can be decreasing to small even when only one instance is predicted to share the label with the bag. We have theoretically proved this statement in Theorem 2. Figure 3(c) and (f) illustrate that Loss-Attention with $\lambda > 0$ can achieve higher F-score than that with $\lambda = 0$. They demonstrate the effectiveness of the regularization term, which can boost the recall of instances with large α, e.g. $\alpha > 0.5$, by increasing the value of λ (please refer to Theorem 4). Similar findings can be observed when using other numbers of training bags. Additionally, here we do not analyze $\omega(m)$, because it is often used on large-scale databases and deep neural networks, upon which its effectiveness has been demonstrated by previous literature (Laine and Aila 2016).

CIFAR-10-based MIL datasets

To better evaluate Loss-Attention, we create more challenging MIL datasets for multi-class classification using images selected from the CIFAR-10 database, whose images belong to 10 categories. Similar to Section 4.3, we build training

Table 3: Results on CIFAR10-bags with different numbers of training bags. Each experiment is repeated 20 times and average results are reported.

Mean bag size	10		50	
# of training bags	500	5000	500	5000
Attention+softmax	0.384	0.507	0.610	0.823
Gated-Attention+softmax	0.370	0.476	0.596	0.769
Loss-Attention	**0.395**	**0.519**	**0.615**	**0.839**

sets with 500 and 5,000 bags, respectively, and a test set consisting of 1,000 bags. The training and test sets contain two types of bags, one type is with the mean bag size and the variance being 10 and 2, respectively, the second type is with the mean bag size and the variance being 50 and 10, respectively. Additionally, the target classes are '3', '5' and '9' and one bag at most contains images from one of these three classes. The details of architectures, the parameter λ, ramp-up function $\omega(m)$, optimizer and hyperparameters are shown in the supplemental material (Tables A10 and A11). We present the bag classification results of Attention, Gated-Attention and Loss-Attention in Table 3. It further suggests that Loss-Attention can achieve superior classification accuracy over Attention and Gated-Attention on multi-class MIL tasks.

Image classification and localization

Here, we conduct experiments on two popular multi-class single-label databases, CIFAR-10 and tiny ImageNet (Le and Yang 2015), to evaluate the performance of Loss-Attention on image classification and localization only using image labels. The CIFAR-10 database consists of a training set with 50,000 images and a test set containing 10,000 images. The tiny ImageNet database has 200 categories and each class contains 500 training images, 50 validation images and 50 test images. We adopt training images for training and validation images for test. We compare Loss-Attention against mean-pooling, max-pooling, Attention and Gated-Attention. We adopt them to replace the global average pooling (mean-pooling) layer in ResNet18 (He et al. 2016), because each point in the feature map obtained by convolutional layers can be viewed as an instance. To evaluate the localization ability, we first rescale the feature map to the original image size, and then each point in the feature map will correspond to one patch in the original image. If half of the patch with the maximum weight falls within the ground truth bounding box of an object, we label the predicted location as correct; otherwise, we count the prediction as wrong. Then we calculate the average precision (AP) to describe the localization prediction accuracy. The details of architectures, the parameter λ, ramp-up function $\omega(m)$, optimizer and hyperparameters are displayed in the supplemental material (Tables A12 and A13).

Table 4 presents the image classification accuracy of five methods on the CIFAR-10 database with 1,000, 5,000 and all images selected from the training set. As we can see, Loss-Attention can achieve 2.1%, 1.1% and 0.2% higher accuracy than the best competitor mean-pooling when using 1,000, 5,000 and all training images, respectively. Table 5 shows their image classification and localization accuracy on the tiny ImageNet database. It suggests that Loss-Attention can achieve the best top-1 classification accuracy and localization accuracy among five methods. Although mean-pooling can obtain the best top-5 classification accuracy, its localization accuracy is significantly worse than Loss-Attention. Tables 4-5 illustrate that Loss-Attention can obtain better classification and localization accuracy than Attention and Gated-Attention. This might be caused by that Loss-Attention can learn patch (instance) weights and predictions, and image (bag) predictions simultaneously, and smooth the training process, thereby largely reducing the possibility and effect of assigning large weights to wrong instances, which are out of the bounding box.

Table 4: Image average classification accuracy on the CIFAR-10 database with 1000, 5000 and all training images selected (10 runs for 1000 and 5000 training images, and 5 runs for all training images).

Methods	1000	5000	All
max-pooling	0.571	0.815	0.944
mean-pooling	0.584	0.819	0.946
Attention+softmax	0.557	0.791	0.933
Gated-Attention+softmax	0.557	0.794	0.933
Loss-Attention	**0.605**	**0.830**	**0.948**

Table 5: Image and localization classification accuracy of five methods on the tiny ImageNet database.

Methods	classification		localization
	top-1	top-5	AP
max-pooling	0.575	0.780	0.748
mean-pooling	0.592	**0.791**	0.704
Attention+softmax	0.554	0.776	0.499
Gated-Attention+softmax	0.557	0.777	0.499
Loss-Attention	**0.598**	0.785	**0.756**

Conclusions

In this paper, we present a novel loss based attention mechanism to simultaneously learn instance weights and predictions, and bag predictions for deep multiple instance learning, by connecting the attention mechanism with the softmax and cross-entropy loss functions. The proposed attention mechanism learns instance weights by using the parameters of the fully connected layer for bag predictions, and directly calculates instance weights based on the loss function. Additionally, a regularization term, consisting of instance weights and cross-entropy functions, is proposed to further boost the instance recall. And a consistency cost, forming a consensus prediction of learned instance weights, is introduced into the final loss to smooth the training process of neural networks. Furthermore, we theoretically analyze the proposed loss based attention mechanism and prove that the regularization term can boost the instance recall. Experiments on multiple small and large-scale databases demonstrate that the proposed method outperforms state-of-the-art methods. Although our method can achieve promising performance on multi-class single-label tasks, it cannot be directly applied to multi-label tasks because of the inferior performance of the softmax function. Therefore, we will extend the proposed method to handle multi-label tasks in the future.

References

Amores, J. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201:81–105.

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2003. Support vector machines for multiple-instance learning. In *NIPS*, 577–584.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, Y.; Bi, J.; and Wang, J. Z. 2006. Miles: Multiple-instance learning via embedded instance selection. *TPAMI* 28(12):1931–1947.

Cheplygina, V.; Tax, D. M.; and Loog, M. 2015. Multiple instance learning with bag dissimilarities. *Pattern Recognition* 48(1):264–275.

De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of operations research* 134(1):19–67.

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

Feng, J., and Zhou, Z.-H. 2017. Deep miml network. In *AAAI*, 1884–1890.

Gärtner, T.; Flach, P. A.; Kowalczyk, A.; and Smola, A. J. 2002. Multi-instance kernels. In *ICML*, volume 2, 179–186.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hou, L.; Samaras, D.; Kurc, T. M.; Gao, Y.; Davis, J. E.; and Saltz, J. H. 2016. Patch-based convolutional neural network for whole slide tissue image classification. In *CVPR*, 2424–2433.

Ilse, M.; Tomczak, J. M.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *ICML*, 1884–1890.

Kandemir, M., and Hamprecht, F. A. 2015. Computer-aided diagnosis from weak supervision: a benchmarking study. *Computerized medical imaging and graphics* 42:44–50.

Keeler, J. D.; Rumelhart, D. E.; and Leow, W. K. 1991. Integrated segmentation and recognition of hand-printed numerals. In *NIPS*, 557–563.

Kraus, O. Z.; Ba, J. L.; and Frey, B. J. 2016. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* 32(12):i52–i59.

Laine, S., and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Le, Y., and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Li, C. H.; Gondra, I.; and Liu, L. 2012. An efficient parallel neural network-based multi-instance learning algorithm. *The Journal of Supercomputing* 62(2):724–740.

Miyato, T.; Maeda, S.-i.; Ishii, S.; and Koyama, M. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*.

Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J.; et al. 2014. Weakly supervised object recognition with convolutional neural networks. In *NIPS*.

Papandreou, G.; Chen, L.-C.; Murphy, K.; and Yuille, A. L. 2015. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*.

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

Pappas, N., and Popescu-Belis, A. 2017. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research* 58:591–626.

Pathak, D.; Shelhamer, E.; Long, J.; and Darrell, T. 2014. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*.

Pinheiro, P. O., and Collobert, R. 2015. Weakly supervised semantic segmentation with convolutional networks. In *CVPR*, volume 2, 6.

Ramon, J., and De Raedt, L. 2000. Multi instance neural networks.

Shi, X.; Xing, F.; Xu, K.; Xie, Y.; Su, H.; and Yang, L. 2017. Supervised graph hashing for histopathology image retrieval and classification. *MIA* 42:117–128.

Shi, X.; Sapkota, M.; Xing, F.; Liu, F.; Cui, L.; and Yang, L. 2018. Pairwise based deep ranking hashing for histopathology image classification and retrieval. *Pattern Recognition* 81:14–22.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. *CVPR*.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition* 74:15–24.

Wei, X.-S., and Zhou, Z.-H. 2016. An empirical study on image bag generators for multi-instance learning. *Machine learning* 105(2):155–198.

Wei, X.-S.; Wu, J.; and Zhou, Z.-H. 2017. Scalable algorithms for multi-instance learning. *TNNLS* 28(4):975–987.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Xu, K.; Liu, S.; Zhao, P.; Chen, P.-Y.; Zhang, H.; Fan, Q.; Erdogmus, D.; Wang, Y.; and Lin, X. 2018. Structured adversarial attack: Towards general implementation and better interpretability. *arXiv preprint arXiv:1808.01664*.

Xu, K.; Liu, S.; Zhang, G.; Sun, M.; Zhao, P.; Fan, Q.; Gan, C.; and Lin, X. 2019. Interpreting adversarial examples by activation promotion and suppression. *arXiv preprint arXiv:1904.02057*.

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

Zhang, Q.-s., and Zhu, S.-C. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19(1):27–39.

Zhang, Z.; Chen, P.; McGough, M.; Xing, F.; Wang, C.; Bui, M.; Xie, Y.; Sapkota, M.; Cui, L.; Dhillon, J.; et al. 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence* 1(5):236.

Zhang, C.; Platt, J. C.; and Viola, P. A. 2006. Multiple instance boosting for object detection. In *NIPS*, 1417–1424.

Zhou, Z.-H., and Zhang, M.-L. 2002. Neural networks for multi-instance learning. In *ICIIT*, 455–459.

Zhou, Y.; Sun, X.; Liu, D.; Zha, Z.; and Zeng, W. 2017. Adaptive pooling in multi-instance learning for web video annotation. In *ICCV*.

Zhou, Z.-H.; Sun, Y.-Y.; and Li, Y.-F. 2009. Multi-instance learning by treating instances as non-iid samples. In *ICML*, 1249–1256. ACM.