Supplemental Material for "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

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.*

Proof. Based on Eq. (5), when $\alpha_{i,j} \to 1$, it is easy to obtain $\alpha_{i,t} \to 0$ ($1 \le t \le n_i$ and $t \ne j$) and $\mathbf{h}_{i,t}^{L-1} \to 0$. Hence, $\mathbf{h}_i^{L-1} \approx \mathbf{h}_{i,j}^{L-1}$, which determines the $\mathbf{z}_i$. Thus, the bag and instance prediction probabilities are approximate. Therefore, Theorem 1 is proved. □

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. (6), there exists:*

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

Proof. Because Eq. (5) contains $\mathbf{h}_{i,j}^{L-1} \leftarrow \alpha_{i,j}\mathbf{h}_{i,j}^{L-1}$ and $\mathbf{h}_i^{L-1} = \sum_{t=1}^{n_i} \mathbf{h}_{i,t}^{L-1}$, combining these two terms, it has $\mathbf{h}_i^{L-1} = \sum_{t=1}^{n_i} \alpha_{i,t}\mathbf{h}_{i,t}^{L-1}$. Then in Eq. (6), a lower bound of the main objective can be obtained by:

$$\begin{aligned}
L_1 &= -log\frac{exp(z_{i,k})}{\sum_{c=0}^{K-1} exp(z_{i,c})} \\
&= -log\frac{exp(\sum_{t=1}^{n_i} \alpha_{i,t} z_{i,t,k})}{\sum_{c=0}^{K-1} exp(\sum_{t=1}^{n_i} \alpha_{i,t} z_{i,t,c}))} \\
&= -log\frac{\prod_{t=1}^{n_i} exp(\alpha_{i,t} z_{i,t,k})}{\sum_{c=0}^{K-1} \prod_{t=1}^{n_i} exp(\alpha_{i,t} z_{i,t,c})} \\
&= -log\frac{\prod_{t=1}^{n_i} (exp(z_{i,t,k}))^{\alpha_{i,t}}}{\sum_{c=0}^{K-1} \prod_{t=1}^{n_i} (exp(z_{i,t,c}))^{\alpha_{i,t}}} \\
&= -log\frac{\frac{\prod_{t=1}^{n_i} (exp(z_{i,t,k}))^{\alpha_{i,t}}}{\prod_{t=1}^{n_i} \sum_{c=0}^{K-1} (exp(z_{i,t,c}))^{\alpha_{i,t}}}}{\frac{\sum_{c=0}^{K-1} \prod_{t=1}^{n_i} (exp(z_{i,t,c}))^{\alpha_{i,t}}}{\prod_{t=1}^{n_i} \sum_{c=0}^{K-1} (exp(z_{i,t,c}))^{\alpha_{i,t}}}} \\
&= -log\frac{\prod_{t=1}^{n_i} (q_{i,t,k})^{\alpha_{i,t}}}{\sum_{c=0}^{K-1} \prod_{t=1}^{n_i} (q_{i,t,c})^{\alpha_{i,t}}} \\
&= -log(1 - \frac{\sum_{c=0,c\ne k}^{K-1} \prod_{t=1}^{n_i} (\frac{q_{i,t,c}}{q_{i,t,k}})^{\alpha_{i,t}}}{1+\sum_{c=0,c\ne k}^{K-1} \prod_{t=1}^{n_i} (\frac{q_{i,t,c}}{q_{i,t,k}})^{\alpha_{i,t}}}) \\
&\ge \frac{\sum_{c=0,c\ne k}^{K-1} \prod_{t=1}^{n_i} (\frac{q_{i,t,c}}{q_{i,t,k}})^{\alpha_{i,t}}}{1+\sum_{c=0,c\ne k}^{K-1} \prod_{t=1}^{n_i} (\frac{q_{i,t,c}}{q_{i,t,k}})^{\alpha_{i,t}}},
\end{aligned} \tag{A1}$$

where the sixth equality is derived from $q_{i,t,k} = \frac{exp(z_{i,t,k})}{\sum_{c=0}^{K-1} exp(z_{i,t,c})}$ and the eighth inequality is derived from $log(1+a) \le a$ for all $a > -1$.
Therefore, Theorem 2 is proved. □

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. (6), there exists:*

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

Proof. For the regularization term L_2, its lower bound can be calculated as follows:

$$\begin{aligned}
L_2 &= -\lambda \sum_{t=1}^{n_i} \alpha_{i,t} log\frac{exp(z_{i,t,k})}{\sum_{c=0}^{K-1} exp(z_{i,t,c})} \\
&= -\lambda \frac{\sum_{t=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,t,c}) log(1 - \frac{\sum_{c=0,c\ne k}^{K-1} exp(z_{i,t,c})}{\sum_{c=0}^{K-1} exp(z_{i,t,c})})}{\sum_{j=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,j,c})} \\
&\ge \lambda \frac{\sum_{t=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,t,c}) \frac{\sum_{c=0,c\ne k}^{K-1} exp(z_{i,t,c})}{\sum_{c=0}^{K-1} exp(z_{i,t,c})}}{\sum_{j=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,j,c})} \\
&= \lambda \frac{\sum_{t=1}^{n_i} \sum_{c=0,\ c\ne k}^{K-1} exp(z_{i,t,c})}{\sum_{t=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,t,c})},
\end{aligned} \tag{A2}$$

where the second equality is derived from $\alpha_{i,t} = \frac{\sum_{c=0}^{K-1} exp(z_{i,t,c})}{\sum_{j=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,j,c}}$) and the third inequality is based on $log(1+a) \le a$ for all $a > -1$.
Therefore, Theorem 3 is proved. □

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.*

Proof. Because $L_2 \ge \lambda \frac{\sum_{t=1}^{n_i} \sum_{c=0,\ c\ne k}^{K-1} exp(z_{i,t,c})}{\sum_{t=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,t,c})}$, it has $\frac{L_2}{\lambda} \ge \frac{\sum_{c=0,\ c\ne k}^{K-1} exp(z_{i,j,c})}{\sum_{t=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,t,c})} \ge \frac{\max\limits_{0\le c\le K-1, c\ne k} exp(z_{i,j,c})}{\sum_{t=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,t,c})}$. Additionally, $\alpha_{i,j} = \frac{\sum_{c=0}^{K-1} exp(z_{i,j,c})}{\sum_{t=1}^{n_i} \sum_{c=0}^{K-1} exp(z_{i,t,c})} = \frac{\sum_{c=0,c\ne k}^{K-1} exp(z_{i,j,c})}{\sum_{t=1}^{n_i} \sum_{c=0,c\ne k}^{K-1} exp(z_{i,t,c})} + \frac{exp(z_{i,j,k})}{\sum_{t=1}^{n_i} \sum_{c=0,c\ne k}^{K-1} exp(z_{i,t,c})} > \frac{2L_2}{\lambda}$. They suggest $z_{i,j,k} > \sum_{c=0,c\ne k}^{K-1} z_{i,j,c} \ge \max\limits_{0\le c\le K-1, c\ne k} z_{i,j,c}$, and thus the j-th instance will be

predicted to the k-th class. Therefore, Theorem 4 is proved. □

MIL datasets classification

Table A1 shows the detailed information about features, instances and bags in each MIL dataset. In Tables A2 and A3, we list the architectures of the embedding-based and instance-based models for classical MIL datasets, respectively. These architectures are on the basis of the model in (Pappas and Popescu-Belis 2017). We utilize 'fc' to denote the fully-connected layer and adopt 'mil-max'/ 'mil-mean' to represent the max or mean pooling operator for multiple instance learning, and provide a number of output hidden units after a dash. The ReLU non-linearity is used in the networks. Table A4 provides the parameter λ, ramp-up function $\omega(m)$, the optimizer, hyperparameters and stopping criterion used in our experiments. Note that we utilize the same stopping criterion as the previous work (Pappas and Popescu-Belis 2017) (Ilse, Tomczak, and Welling 2018).

Table A1: Classic 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 A2: Classical MIL datasets: The embedding-based model architecture.

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
8	fc-2 + softmax

Table A3: Classical MIL datasets: The instance-based model architecture.

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

MNIST-based MIL datasets classification

Table A8 shows the architecture of the used models for the MNIST-bags dataset. The architecture is based on (LeCun et al. 1998). Table A9 provides the parameter λ, ramp-up function $\omega(m)$, the optimizer, hyperparameters and stopping criterion used in our binary and multi-class classification experiments.

Table A8: MNIST-based MIL datasets: The used model architecture.

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	dropout
7	mil-attention
8	fc-2 + softmax

CIFAR-10-based MIL datasets classification

Table A10 shows the architecture of the used models for the CIFAR10-bags dataset. The architecture is on the basis of (LeCun et al. 1998). Table A11 provides the parameter λ, ramp-up function $\omega(m)$, the optimizer and hyperparameters used in our experiments. Here, $\omega(m)$ is a ramp-up Gaussian function $e^{-5\|1-T_1\|_2^2}$, where T_1 linearly advances from 0 to 1 during the first 40 epochs and then keeps unchanged. We utilize the optimizer Adam (Kingma and Ba 2014) to update the network parameters, and initialize Adam momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate using $\omega(m)$ ramps up to the maximum 0.001 during the first 40 epochs and keeps unchanged from 40 to 70 epochs, but gradually decreases to 0 using a Gaussian ramp-down function $e^{-7.5\|1-T_2\|^2}$, where T_2 linearly decreases from 1 to 0 during the last 30 epochs. β_1 keeps unchanged during the first 70 epochs but decreases to 0.5 using the Gaussian ramp-down function during the last 30 epochs. β_2 becomes 0.999 after the first 40 epochs.

Image classification and localization

Table A12 shows the architecture of the used models for image classification and localization on CIFAR-10 and tiny ImageNet databases. The architecture is based on ResNet18 (He et al. 2016). Table A13 provides the parameter λ, ramp-up function $\omega(m)$, the optimizer and hyperparameters used in our experiments. Similar to (Laine and Aila 2016), $\omega(m)$ is a ramp-up Gaussian function $500 * e^{-5\|1-T_1\|_2^2}$, where T_1 linearly advances from 0 to 1 during the first 80 epochs and then keeps unchanged. We utilize the optimizer Adam (Kingma and Ba 2014) to update the network parameters, and initialize Adam momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate using $\omega(m)$ ramps up to the maximum 0.003 during the first 80 epochs and keeps unchanged from 80 to 150 epochs, but gradually decreases to 0 using a Gaussian ramp-down function $e^{-12.5\|1-T_2\|^2}$, where T_2 linearly decreases from 1 to 0 during the last 50 epochs. β_1 keeps unchanged during the first 80 epochs but decreases

Table A4: Classical MIL datasets: The optimization procedure details.

Experiment	λ	$\omega(m)$	Optimizer	Learning rate	β_1, β_2	Weight decay	Epochs	Stopping criteria
All	0.1	0	Adam	0.0001	(0.9, 0.999)	0.0005	100	lowest validation error and loss

Table A9: MNIST-based MIL datasets: The optimization procedure details.

Experiment	λ	$\omega(m)$	Optimizer	Learning rate	β_1, β_2	Weight decay	Epochs	Stopping criteria
Binary	2.0	0	Adam	0.0005	(0.9, 0.999)	0.0001	30	lowest validation error and loss
Multi-class	2.0	0	Adam	0.0005	(0.9, 0.999)	0.0001	50	lowest validation error and loss

Table A11: CIFAR-10-based MIL datasets: The optimization procedure details.

λ	$\omega(m)$	Optimizer	Learning rate	β_1, β_2	Epochs
0.01	$e^{-5\|1-T\|_2^2}$	Adam	0.001	(0.9, 0.99)	100

Table A13: Image classification and localization: The optimization procedure details.

λ	$\omega(m)$	Optimizer	Learning rate	β_1, β_2	Epochs
1	$400 * e^{-5\|1-T\|_2^2}$	Adam	0.003	(0.9, 0.99)	200

Table A10: CIFAR-10-based MIL datasets: The used model architecture.

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	dropout
7	fc-500 + tanh
8	mil-attention
9	fc-2 + softmax

Table A12: Image classification and localization: The used model architecture.

Layer	Type
1	conv(3, 1, 1)-64+ReLU
2	maxpool(3, 2)
3	conv(3, 1, 1)-64+ReLU
4	conv(3, 1, 1)-64+ReLU
5	conv(3, 1, 1)-64+ReLU
6	conv(3, 1, 1)-64+ReLU
7	conv(3, 2, 1)-128+ReLU
8	conv(3, 1, 1)-128+ReLU
9	conv(3, 1, 1)-128+ReLU
10	conv(3, 1, 1)-128+ReLU
11	conv(3, 2, 1)-256+ReLU
12	conv(3, 1, 1)-256+ReLU
13	conv(3, 1, 1)-256+ReLU
14	conv(3, 1, 1)-256+ReLU
15	conv(3, 2, 1)-512+ReLU
16	conv(3, 1, 1)-512+ReLU
17	conv(3, 1, 1)-512+ReLU
18	conv(3, 1, 1)-512+ReLU
19	mil-attention
20	fc-m+softmax

to 0.5 using the Gaussian ramp-down function during the last 50 epochs. β_2 becomes 0.999 after the first 80 epochs.

References

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

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

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

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

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.

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