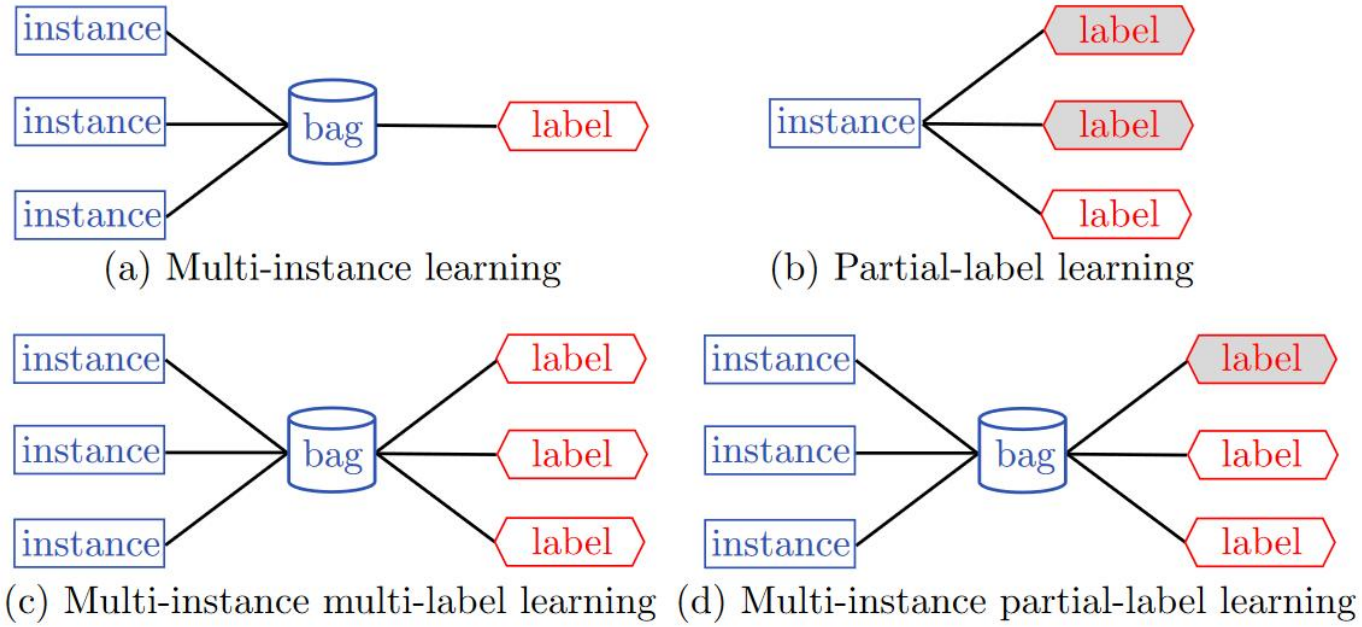
Exploiting Conjugate Label Information for Multi-Instance Partial-Label Learning

Wei Tang[1,2], **Weijia Zhang**[3], **Min-Ling Zhang**[1,2*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Lab. of Computer Network and Information Integration (Southeast University), MoE, China
[3]School of Information and Physical Sciences, The University of Newcastle, NSW 2308, Australia
tangw@seu.edu.cn, weijia.zhang@newcastle.edu.au, zhangml@seu.edu.cn

IJCAI 2024

Multi-instance Partial-label Learning (M



Each training sample is represented by a multi-instance bag associated with a bag-level candidate label set, which consists of one ground-truth label and some false positive labels. Moreover, the bag contains at least one instance that belongs to the ground-truth label while no instance pertains to the false positive labels. Therefore, inexact supervision exists both in the instance space and the label space in MIPL.

Background

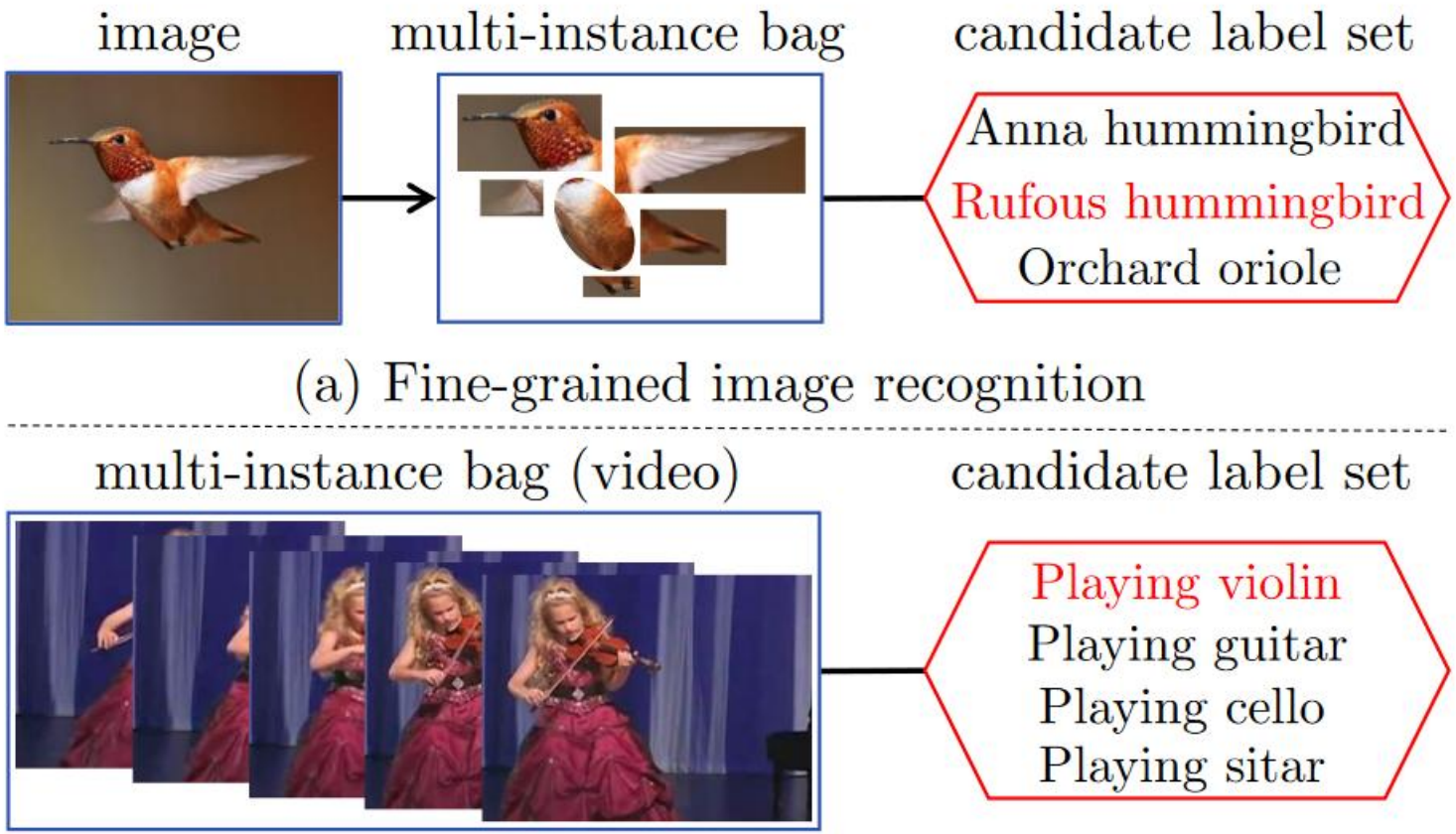


Figure 2: Potential applications of MIPL, where the red is the ground-truth label.

Related work

MIPLGP

MIPLGP begins by augmenting a negative class for each candidate label set, subsequently treating the candidate label set of each multi-instance bag as that of each instance within the bag. Finally, it employs the Dirichlet disambiguation strategy and the Gaussian processes regression model for disambiguation.

DEMIPL

DEMIPL follows the embedded-space paradigm and aggregates each multi-instance bag into a feature representation and employs a momentum-based disambiguation strategy to find true labels from candidate label sets.

However, both methods primarily depend on mapping from instances or multi-instance bags to candidate label sets for disambiguation, without considering the proposed CLI in this paper.

Method

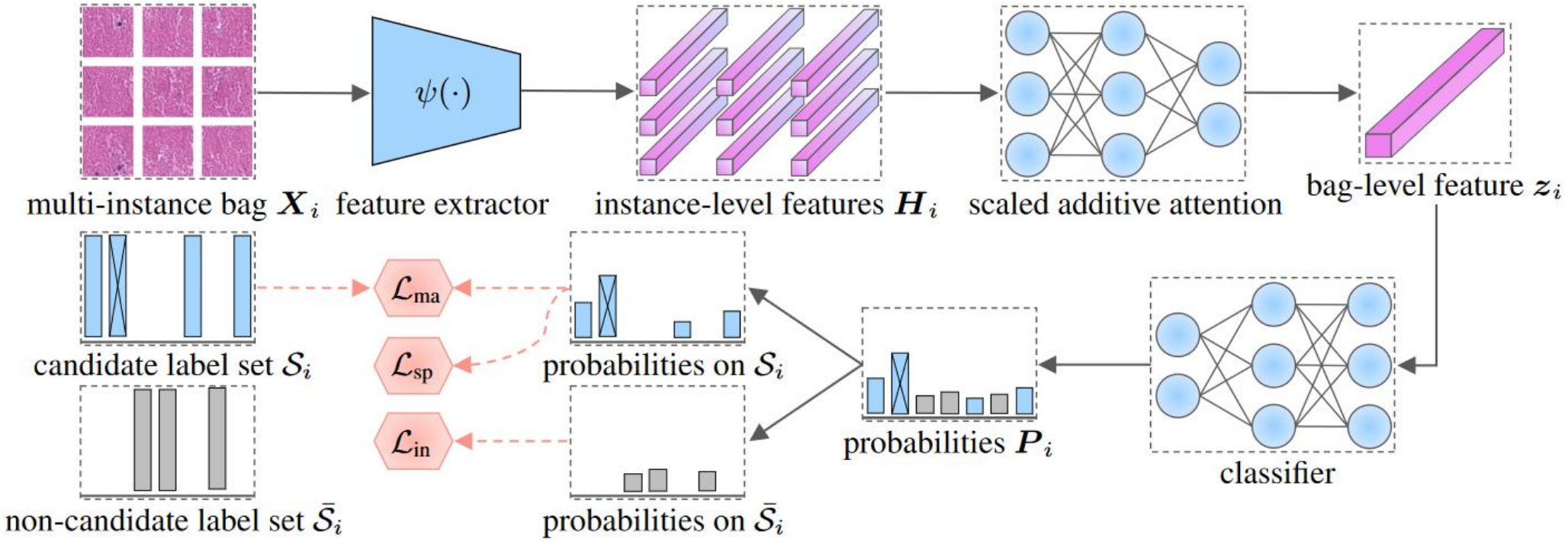
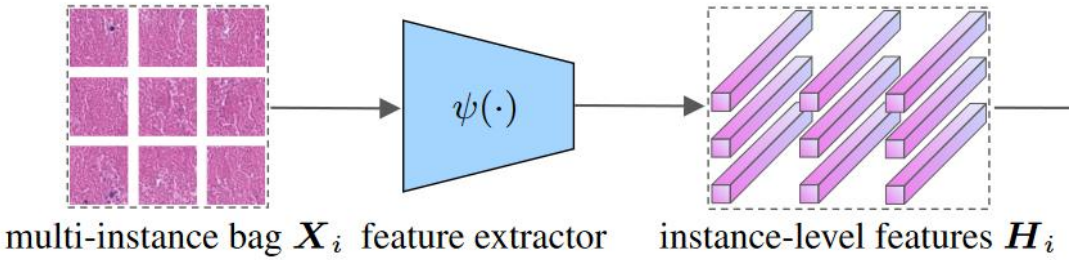


Figure 3: The pipeline of ELIMIPL, where $\mathcal{L}_{ma}$, $\mathcal{L}_{sp}$, and $\mathcal{L}_{in}$ refer to mapping loss, sparsity loss, and inhibition loss, respectively.

Instance-Level Feature Extractor

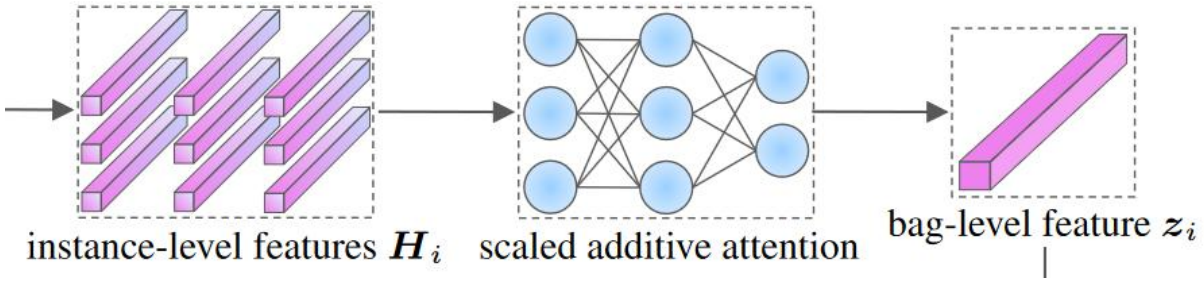


For a given multi-instance bag $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,n_i}\}$ with n_i instances, instance-level feature representations $\mathrm{H_i}$ are learned using a feature extractor $\psi(\cdot)$ as follows:

$$\boldsymbol{H}_i = \psi(\boldsymbol{X}_i) = \{\boldsymbol{h}_{i,1}, \boldsymbol{h}_{i,2}, \cdots, \boldsymbol{h}_{i,n_i}\},$$

where $h_{i,j} \in \mathbb{R}^l$ indicates the feature representation of the j-th instance within the i-th multi-instance bag, and $\psi(\cdot)$ is a neural network comprised of two components, i.e., $\psi(X_i) = \psi_2(\psi_1(X_i))$. Here, $\psi_1(\cdot)$ is a feature extractor that can be tailored to the specific characteristics of the datasets, and $\psi_2(\cdot)$ is composed of fully connected layers that map instance-level features to an embedded space of dimension l.

Scaled Additive Attention Mechanism



We first denote the output of the additive attention mechanism as $\xi(h_{i,j})$, quantifying the impact of the j-th instance on the i-th bag as follows:

$$\xi(h_{i,j}) = \boldsymbol{W}^\top(\tanh(\boldsymbol{W}_t^\top \boldsymbol{h}_{i,j} + \boldsymbol{b}_t) \odot \mathrm{sigm}(\boldsymbol{W}_s^\top \boldsymbol{h}_{i,j} + \boldsymbol{b}_s)),$$

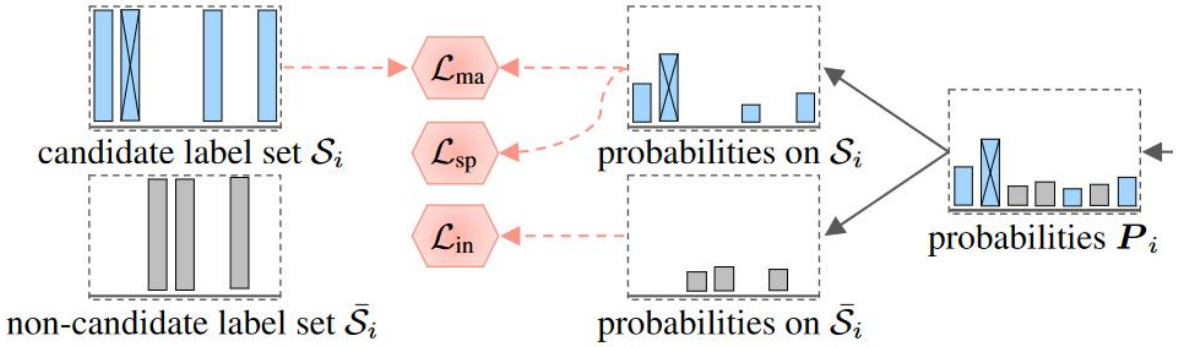
Then, we normalize $\xi(x_{i,j})$ using softmax with a scaling factor $1/\sqrt{l}$ to derive the attention score:

$$a_{i,j} = \frac{\exp\left(\xi\left(h_{i,j}\right)/\sqrt{l}\right)}{\sum_{j'=1}^{n_i} \exp\left(\xi\left(h_{i,j'}\right)/\sqrt{l}\right)},$$

Finally, we consolidate the instance-level features into a bag-level representation, as demonstrated below:

$$\boldsymbol{z}_i = \sum_{j=1}^{n_i} a_{i,j} \boldsymbol{h}_{i,j},$$

Conjugate Label Information



We employ a weighted mapping loss function:

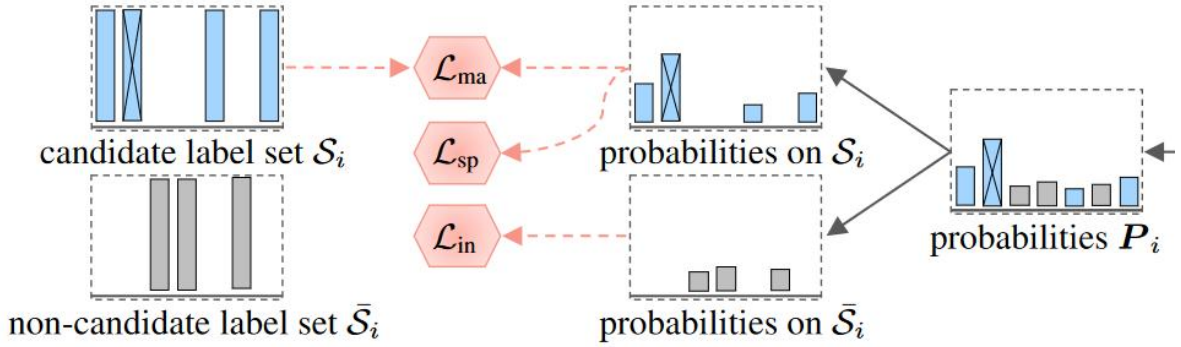
$$\mathcal{L}_{\mathrm{ma}}(\mathcal{Z},\mathcal{S}) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{c\in\mathcal{S}_i} w_{i,c}^{(t)} \log(f_c(\boldsymbol{z}_i)),$$

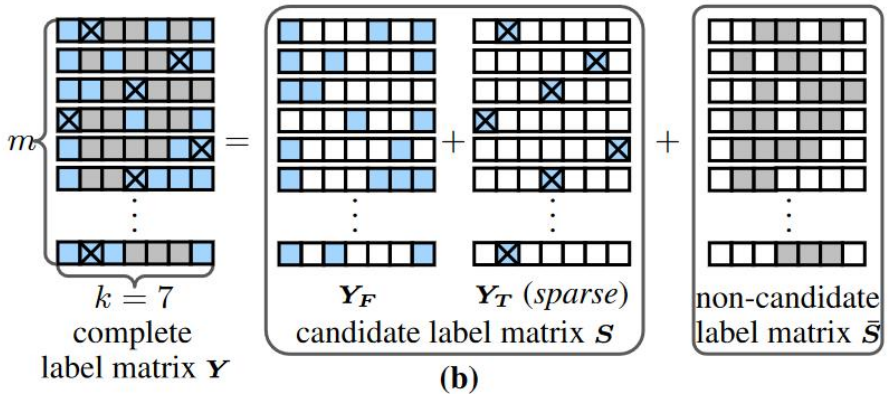
For candidate labels, we initialize $w_{i,c}^{(0)} = \frac{1}{|\mathrm{S}_i|}$ through an averaging approach. During training, we update $w_{i,c}^{(t)}$ by computing a weighted sum of the classifier's outputs at both the previous epoch and current epoch as follows:

$$w_{i,c}^{(t)} = \rho^{(t)} w_{i,c}^{(t-1)} + (1-\rho^{(t)})\frac{f_c(\boldsymbol{z}_i)}{\sum_{c'\in\mathcal{S}_i} f_{c'}(\boldsymbol{z}_i)},$$

where $\rho^{(t)} = (T-t)/T$ is dynamically adjusted across epochs, and T is the maximum of the training epochs.

Conjugate Label Information





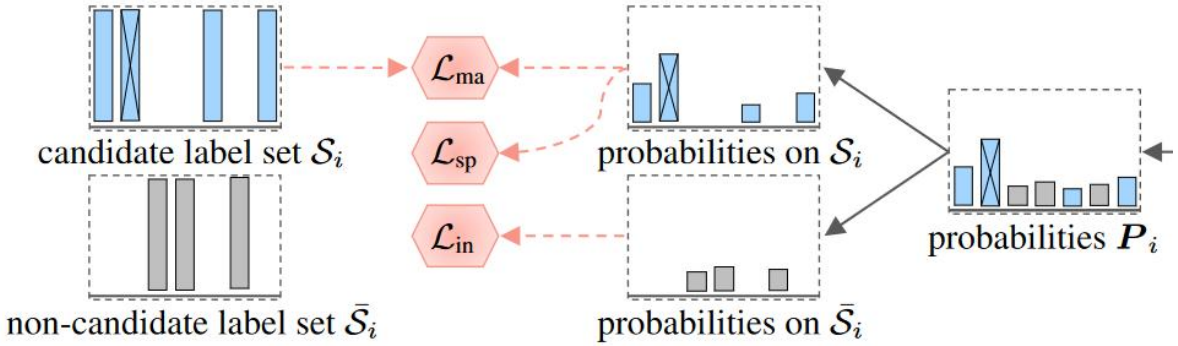
While the mapping loss can assess the relative labeling probabilities of candidate labels, it fails to capture the mutually exclusive relationships among the candidate labels.

Although the true labels remain inaccessible during the training process, we encourage the classifier to generate sparse prediction probabilities for the candidate labels. Therefore, we directly capture the mutually exclusive relationships among the candidate labels by implementing the sparsity loss, as detailed below:

$$\mathcal{L}_{\mathrm{sp}}(\mathcal{S}) = \frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{P}_i \odot \boldsymbol{S}_i\|_0,$$

Since minimizing the ℓ_0 norm is NP-hard, we employ the ℓ_1 norm as a surrogate for the ℓ_0 norm.

Conjugate Label Information



As the label space has a fixed size, an antagonistic relationship arises between the non-candidate and candidate label sets. To enhance the classifier's prediction probabilities for the candidate label set, a natural strategy is to diminish the classifier's prediction probabilities for the noncandidate label set. Motivated by this insight, we introduce an inhibition loss as follows:

$$\mathcal{L}_{\text{in}}(\mathcal{Z}, \bar{\mathcal{S}}) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{\bar{c}\in\bar{\mathcal{S}}_i}\log(1 - f_{\bar{c}}(\boldsymbol{z}_i)),$$

CLI Loss

$$\mathcal{L} = \mathcal{L}_{\text{ma}}(\mathcal{Z}, \mathcal{S}) + \mu\mathcal{L}_{\text{sp}}(\mathcal{S}) + \gamma\mathcal{L}_{\text{in}}(\mathcal{Z}, \bar{\mathcal{S}}),$$

Experiment

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
ELIMIPL	1	**.992±.007**	**.903±.018**	**.771±.018**	**.675±.022**
	2	**.987±.010**	**.845±.026**	**.745±.015**	**.616±.025**
	3	**.748±.144**	**.702±.055**	**.717±.017**	**.600±.029**
DEMIPL	1	.976±.008	.881±.021	.744±.016	.635±.041
	2	.943±.027	.823±.028	.701±.024	.554±.051
	3	.709±.088	.657±.025	.696±.024	.503±.018
MIPLGP	1	.949±.016	.847±.030	.716±.026	.669±.019
	2	.817±.030	.791±.027	.672±.015	.613±.026
	3	.621±.064	.670±.052	.625±.015	.569±.032
Mean					
PRODEN	1	.605±.023	.697±.042	.296±.014	.219±.014
	2	.481±.036	.573±.026	.272±.019	.184±.014
	3	.283±.028	.345±.027	.211±.013	.166±.017
RC	1	.658±.031	.753±.042	.362±.015	.279±.011
	2	.598±.033	.649±.028	.335±.011	.258±.017
	3	.392±.033	.401±.063	.298±.009	.237±.020
LWS	1	.463±.048	.726±.031	.265±.010	.240±.014
	2	.209±.028	.720±.025	.254±.010	.223±.008
	3	.205±.013	.579±.041	.237±.005	.194±.026
PL-AGGD	1	.671±.027	.743±.026	.353±.019	.355±.015
	2	.595±.036	.677±.028	.314±.018	.315±.019
	3	.380±.032	.474±.057	.296±.015	.286±.018
MaxMin					
PRODEN	1	.508±.024	.424±.045	.387±.014	.316±.019
	2	.400±.037	.377±.040	.357±.012	.287±.024
	3	.345±.048	.309±.058	.336±.012	.250±.018
RC	1	.519±.028	.731±.027	.390±.014	.306±.023
	2	.469±.035	.666±.027	.371±.013	.288±.021
	3	.380±.048	.524±.034	.363±.010	.267±.020
LWS	1	.242±.042	.435±.049	.225±.038	.289±.017
	2	.239±.048	.406±.040	.207±.034	.271±.014
	3	.218±.017	.318±.064	.216±.029	.244±.023
PL-AGGD	1	.527±.035	.391±.040	.383±.014	.397±.028
	2	.439±.020	.371±.037	.372±.020	.360±.029
	3	.321±.043	.327±.028	.344±.011	.328±.023

Table 2: The classification accuracies (mean±std) of ELIMIPL and comparative algorithms on the benchmark datasets with varying numbers of false positive candidate labels ($r \in \{1, 2, 3\}$).

Experiment

Algorithm	C-Row	C-SBN	C-KMeans	C-SIFT
ELIMIPL	.433±.008	.509±.007	**.546±.012**	**.540±.010**
DEMIPL	.408±.010	.486±.014	.521±.012	.532±.013
MIPLGP	.432±.005	.335±.006	.329±.012	–
Mean				
PRODEN	.365±.009	.392±.008	.233±.018	.334±.029
RC	.214±.011	.242±.012	.226±.009	.209±.007
LWS	.291±.010	.310±.006	.237±.008	.270±.007
PL-AGGD	.412±.008	.480±.005	.358±.008	.363±.012
MaxMin				
PRODEN	.401±.007	.447±.011	.265±.027	.291±.011
RC	.227±.012	.338±.010	.208±.007	.246±.008
LWS	.299±.008	.382±.009	.247±.005	.230±.007
PL-AGGD	**.460±.008**	**.524±.008**	.434±.009	.285±.009

Table 3: The classification accuracies (mean±std) of ELIMIPL and comparative algorithms on the real-world datasets.

Experiment

Dataset	r	ELIMIPL	MA+SP	MA+IN	MA
Birdsong	1	.771±.018	.742±.014	.746±.015	.733±.011
	2	.745±.015	.665±.024	.689±.020	.677±.017
	3	.717±.017	.592±.031	.674±.023	.652±.016
SIVAL	1	.675±.022	.618±.021	.626±.019	.620±.022
	2	.616±.025	.532±.041	.550±.040	.540±.038
	3	.600±.029	.545±.027	.521±.025	.521±.032

Table 4: The classification accuracies of the variants on the Birdsong-MIPL and SIVAL-MIPL datasets.

Experiment

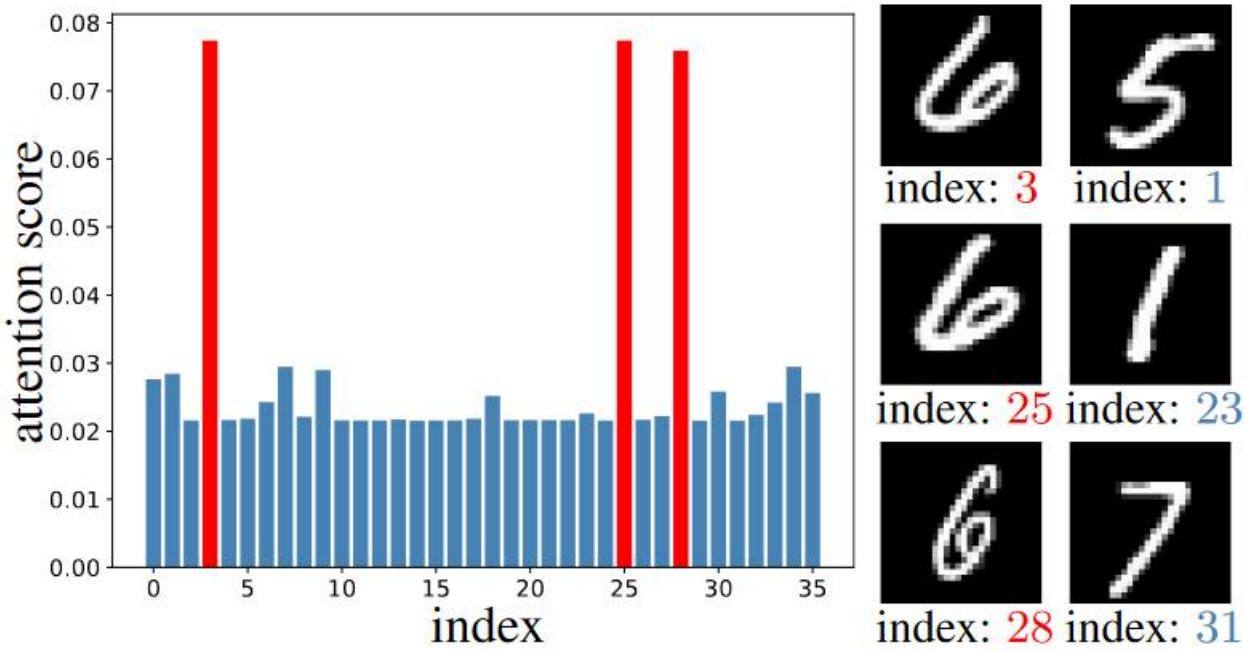


Figure 5: Attention scores for a test bag. Red and blue are the attention scores of positive and negative instances, respectively.

Thanks