Supporting Document for Dual Space Multiple Instance Representative Learning for Medical Image Classification

BMVC 2022 Submission # 768

1 Experimental Setup

1.1 Datasets

We comprehensively evaluate DSMIRL on five well-known classical MIL benchmarks (Elephant, Fox, Tiger, Musk1, and Musk2,) and two medical image datasets (Camelyon16 and Pneumonia CT).

MIL benchmarks: The MIL benchmarks are small-scale datasets, and every sample only contains pre-extracted features. Musk1 and Musk2 are datasets used to predict drug activity [4]. Elephant, Fox, and Tiger are animal image datasets. For each category, positive bags contain at least one of the animals of interest, and negative bags contain other animals [1]. Following the conventions [7], we adopt 10-fold cross-validation to evaluate the models.

Camelyon16: Camelyon16 is a public dataset proposed for metastasis detection in breast cancer [2], which includes 399 images. The dataset provides pixel-level annotations of tumor regions, but we ignore pixel-level annotations during model training and only consider slide-level annotations (i.e., a slide is deemed positive if it contains any tumor region). Each Whole Slide Image (WSI) is cropped into a series of 256×256 non-overlapping patches while discarding the background region (saturation < 15). After pre-processing, about 3.5 million patches at 20 × magnification, with an average of 8,800 patches per bag. Following the previous work [11], the feature of each patch is embedded in a 1024-dimensional vector by a ResNet50 [6] model pre-trained on ImageNet. The Camelyon16 provides an official data split, and its testing sample ratio is $13/40 \approx 1/3$. To reduce the effects of the data split on the model evaluation, we use three-fold cross-validation to ensure that each sample participates in training and testing. Each fold approximately has 133 samples.

Pneumonia CT: We collect a Pneumonia CT dataset from hospital, including 897 patients (450 with pneumonia and 447 without pneumonia). The Ethics Committee of hospital approves this study. Pneumonia CT differs from Camelyon16, where each patient only has a gigapixel WSI, whereas the Pneumonia CT has a series of conventional pixel slices per patient. In addition, this dataset has not provided slice-level annotations and only has patient-level annotations. Unlike the Camelyon16 dataset, this dataset has not provided slice-level annotations and only has patient-level annotations. During the training, we adopt three-fold cross-validation similarly. Each fold contains about 300 patients, with an average of about 150 images per patient, each scaled to 256×256.

1.2 Implementation Details

1.2.1 Feature Extractor

Following the previous work [], we adopt a two-stage learning strategy on Camelyon16. The feature extraction and MIL aggregation are individually optimized. We employ ResNet50 to obtain 1024-dimensional instance feature vectors in the feature extraction stage. As for Pneumonia CT, we adopt the popular end-to-end training strategy due to the lack of convention to follow in the learning strategy. We also use ResNet50 as the backbone for feature extraction and stack two fully connected layers to reduce the dimension of features to 512. Each fully connected layer is followed by a ReLU activation layer and a Dropout layer. The MIL benchmarks directly gives the extracted features, requiring no feature extraction.

1.2.2 Experiments Setup

In the model optimization, we employ the Adam optimizer. On MIL benchmarks and Camelyon16 , the learning rate is 2×10^{-4} and weight-decay is 10^{-5}, while the learning rate is 2×10^{-5} and weight-decay is 5×10^{-3} on Pneumonia CT. The learning rate is decayed by a factor of 0.1 when the training loss does not decrease in 10 epochs. In AIRS, the number of clusters K is 4, 2, and 2 on Camelyon16, Pneumonia CT, and MIL benchmarks, respectively. All experiments are conducted on an RTX 3090 using PyTorch.

2 The Algorithm of DSMIRL

Algorithm 1 shows an iterative process of DSMIRL.

Algorithm 1: The DSMIRL processing flow

```
Input: bag set $B$, parameters $\varphi, \theta, u, v, \psi$, the number of bags $N$
Output: $\hat{\varphi}, \hat{\theta}, \hat{u}, \hat{v}, \hat{\psi}$
for $n$=1,2,$\cdots$,$N$ do
    /* obtain instance features and predictions*/
    obtain instance features $f \leftarrow \mathcal{F}_{\varphi}(B_n)$
    obtain instance predictions $\hat{y} \leftarrow \mathcal{J}_{\theta}(f)$
    /* instance representative selection*/
    group bag $B_n$ into $K$ sub-bags $b$;
    calculate sub-bag score $p_k \leftarrow \max(\{\hat{y}_i\}_{f_i \in b_k})$
    select optimal sub-bag $b_{\hat{k}} \leftarrow \arg\max_{b_k \in B_n} p$
    /* dual-space aggregation*/
    /*feature space*/
    aggregate instance feature
    $\hat{f} \leftarrow \underset{u,v}{attention}(\{f_i\}_{f_i \in b_{\hat{k}}})$
    produce bag label $\hat{\mathcal{Y}}^F \leftarrow \mathcal{Q}_{\psi}(\hat{f})$
    /*label space*/
    produce bag label $\hat{\mathcal{Y}}^L \leftarrow \text{mean}(\{\hat{y}_i\}_{f_i \in b_{\hat{k}}})$
    calculate the cross entropy loss $\mathcal{L}$
    update parameters $\hat{\varphi}, \hat{\theta}, \hat{u}, \hat{v}, \hat{\psi} \leftarrow \arg\min_{\varphi,\theta,u,v,\psi} \mathcal{L}$
end
```

3 Ablation Study

3.1 Discussion of different clustering strategies

In the main body, we discuss the performance of different clustering methods on the Camelyon16 dataset. Here we supplement the performance of different clustering methods on Pneumonia CT datasets. As shown in Table 1, Spectral Clustering also performs best on this dataset.

Methods	Accuracy	AUC	F1-score
K-means [8]	.909±.02	.953±.02	.908±.02
Mean-shift [3]	.898±.04	.958±.03	.895±.04
DBSCAN [5]	.864±.04	.914±.04	.866±.05
Hierarchical Clustering [9]	.846±.05	.915±.05	.846±.05
Spectral Clustering [10]	**.930±.01**	**.967±.01**	**.930±.01**

Table 1: Results on Pneumonia CT. The best ones are in bold, and the second-best ones are underlined.

3.2 Effects of the number of clusters

Table 2 shows the details of the MIL benchmarks, where most of the bags contain only five instances. For example, the Musk1 dataset has 73 bags with less than six instances. If we discuss $K = 6$ on this dataset, only 19 bags can be preserved for training. Therefore, we only discuss three cases of $K = 2$, 3, and 4 on MIL benchmarks. As shown in the figure 1, we report the variation trend of K in other two indicators (accuracy and F1-score) on two medical image datasets.

Datasets	#Bag				
	<6	<5	<4	<3	total
Musk1	73	67	**36**	32	92
Musk2	28	27	**10**	10	100
Tiger	87	48	**27**	8	199
Fox	69	32	**20**	3	200
Elephant	61	31	**14**	6	200

Table 2: Detailed characteristics of the MIL benchmarks. $\# < n$ indicates the number of bags with less than n instances.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *NeurIPS*, pages 577–584, 2003.

[2] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen,

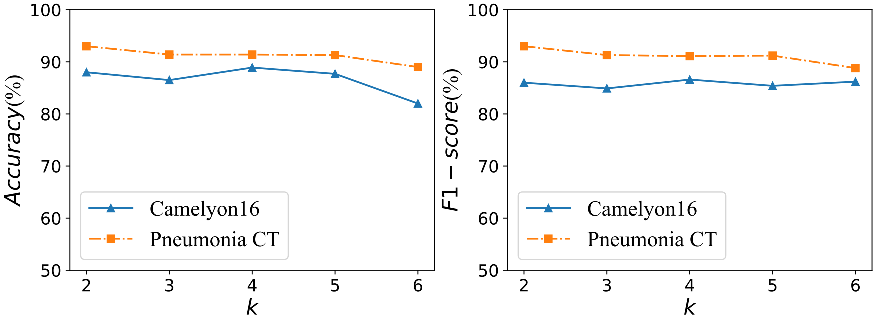

Figure 1: The effect of K in accuracy and F1-score on two medical image datasets.

Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[3] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

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

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[7] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2127–2136. PMLR, 2018.

[8] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[9] Alena Lukasová. Hierarchical agglomerative clustering procedure. *Pattern Recognition*, 11(5-6):365–381, 1979.

[10] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, pages 849–856, 2002.

[11] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *NeurIPS*, 34, 2021.