Fast and Accurate Gigapixel Pathological Image Classification with Hierarchical Distillation Multi-Instance Learning

Supplementary Material

6. The Architecture of LIPN

We designed LIPN as a convolutional neural network with a similar structure to MobileNetV4 [6], but it is more lightweight. The specific structure and related parameters are shown in Fig. 5 and Tab. 8 respectively.

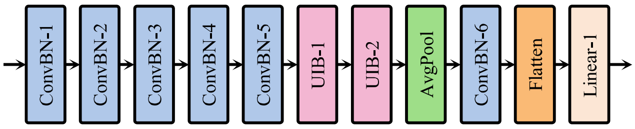

Figure 5. The architecture of LIPN. "ConvBN" is the abbreviation of a cascaded module consisting of a Conv layer, a BatchNorm layer and a ReLU layer, and "UIB" is the abbreviation of Universal Inverted Bottleneck [6].

Layer	Parameter	Output Size
ConvBN-1	k=3, s=2, in=3, out=16	$1 \times 16 \times 8 \times 8$
ConvBN-2	k=3, s=2, in=16, out=16	$1 \times 16 \times 4 \times 4$
ConvBN-3	k=1, s=1, in=16, out=16	$1 \times 16 \times 4 \times 4$
ConvBN-4	k=3, s=2, in=16, out=48	$1 \times 48 \times 2 \times 2$
ConvBN-5	k=1, s=1, in=48, out=24	$1 \times 24 \times 2 \times 2$
UIB-1	sdk=5, mdk=5, e=2, s=2, in=24, out=48	$1 \times 48 \times 1 \times 1$
UIB-2	sdk=3, mdk=3, e=2, s=2, in=48, out=64	$1 \times 64 \times 1 \times 1$
AvgPool	null	$1 \times 64 \times 1 \times 1$
ConvBN-6	k=1, s=1, in=64, out=64	$1 \times 64 \times 1 \times 1$
Flatten	null	1×64
Linear-1	in=64, out=2	1×2

Table 8. The specific parameters of layers within LIPN and the dimensions of the output features. "k", "s", "in", and "out" represent kernel size, stride, input channels, and output channels, respectively. "sdk", "mdk" and "e" represent start depth-wsie kernel size, middle depth-wsie kernel size and expand ratio respectively. The size of the input image is $1 \times 3 \times 16 \times 16$.

7. Experimental Settings.

More Dataset Details. Camelyon16 is a dataset used for diagnosing lymph node metastasis in breast cancer through binary classification. After excluding WSIs with too small tissue areas, we obtained a total of 397 WSIs, where each WSI corresponds to an individual case. In the official splitting, there are 158 normal WSIs and 111 tumor WSIs in the training set, as well as 80 normal WSIs and 48 tumor WSIs in the test set. After background removal and tissue segmentation, approximately 4.5 million 256×256 patches were obtained at 20× magnification, with an average of about 11,000 patches per WSI. It is worth mentioning that, although pixel-level annotations of tumor regions are provided in the official dataset, in all experiments, only slide-level labels were used for training.

TCGA-NSCLC is a dataset of non-small cell lung cancer, comprising two cancer subtypes: lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). After excluding WSIs with too small tissue areas and WSIs with a maximum magnification smaller than 20, a total of 510 LUAD WSIs (from 453 cases) and 500 LUSC WSIs (from 468 cases) were obtained. Different from Camelyon16, each case in TCGA-NSCLC may correspond to multiple WSIs. After background removal and tissue segmentation, approximately 13 million 256×256 patches were obtained at 20× magnification, averaging about 13,000 patches per WSI.

TCGA-BRCA is a dataset of breast cancer, mainly comprising two cancer subtypes: invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). After excluding WSIs with too small tissue areas and WSIs with a maximum magnification smaller than 20, a total of 796 IDC WSIs (from 750 cases) and 201 ILC WSIs (from 188 cases) were obtained. After background removal and tissue segmentation, approximately 11 million 256×256 patches were obtained at 20× magnification, averaging about 11,000 patches per WSI.

Implementation and Hyper-parameters. For fair comparison, the networks used for feature extraction in all experiments are ResNet-50 [1] pre-trained on ImageNet [3], whose parameters are obtained from the official pytorch [5] library. We first trained the DIMN to convergence, then froze it to train the LIPN. For both DMIN and LIPN, we use the Adam [2] optimizer with a batch size of 1. When training DIMN, different learning rates are applied on different datasets: 3e-4 for Camelyon16 and TCGA-BRCA, and 3e-5 for TCGA-NSCLC. For the LIPN training, the learning rate is set to 1e-4 across all datasets. The hyper-parameters τ and γ in formulas Eq. (3), and Eq. (4) are fixed at 0.7 and 0.5, respectively, for all datasets. The coefficients β_1 and β_2 are both set to 1.0. Due to differences in tumor area sizes within WSIs across different datasets, we set the instance retention ratios r to 0.6, 0.7, and 0.7 for Camelyon16, TCGA-NSCLC, and TCGA-BRCA, respectively, based on the results of the ablation study Tab. 9. Simliarly, we set K in Eq. (6) to 12 for for Camelyon16 and TCGA-BRCA, and 16 for TCGA-NSCLC.

8. Additional Experimental Results

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AUC	96.7	97.3	97.3	96.5	97.7	98.2	97.5	96.5	97.4	97.2
ACC	93.9	93.5	92.7	93.5	95.0	93.8	93.9	93.5	93.5	93.9
Mean	95.3	95.4	95.0	95.0	**96.4**	96.0	95.7	95.0	95.4	95.5

(1) HDMIL† performance on Camelyon16.

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AUC	93.1	97.5	96.9	96.4	96.4	97.6	98.1	97.5	97.7	97.2
ACC	91.5	91.9	93.1	93.5	92.7	95.4	94.6	94.6	94.2	93.9
Mean	92.3	94.7	95.0	94.9	94.6	**96.5**	96.3	96.1	96.0	95.5

(2) HDMIL performance on Camelyon16.

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AUC	93.9	93.1	93.3	95.1	95.4	95.6	95.6	95.6	95.2	95.2
ACC	88.4	86.9	86.7	87.9	88.9	89.6	90.3	89.9	89.1	89.7
Mean	91.1	90.0	90.0	91.5	92.1	92.6	**92.9**	92.8	92.2	92.4

(3) HDMIL† performance on TCGA-NSCLC.

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AUC	93.1	94.9	94.7	95.0	96.0	95.7	95.9	95.9	95.6	95.2
ACC	85.7	88.5	88.5	87.9	90.2	90.0	90.5	90.4	89.8	89.7
Mean	89.4	91.7	91.6	91.4	93.1	92.8	**93.2**	93.1	92.7	92.4

(4) HDMIL performance on TCGA-NSCLC.

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AUC	71.9	91.4	91.8	92.0	91.9	92.3	93.3	92.2	91.0	91.2
ACC	78.6	90.3	89.4	88.1	89.8	88.4	89.8	89.4	89.4	89.0
Mean	75.2	90.9	90.7	90.1	90.9	90.3	**91.6**	90.8	90.2	90.1

(5) HDMIL† performance on TCGA-BRCA.

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AUC	91.1	90.5	92.3	92.6	92.3	91.7	93.3	92.6	91.2	91.2
ACC	89.0	89.7	87.8	88.1	88.9	88.1	88.7	89.0	88.7	89.0
Mean	90.1	90.1	90.0	90.4	90.6	89.9	**91.0**	90.8	89.9	90.1

(6) HDMIL performance on TCGA-BRCA.

Table 9. Performance of the models with different preset instance retention rates r on the **validation** set. The average AUC and ACC scores over 10-fold cross-validation are reported. We selected the r that maximizes the mean of ACC and AUC as the final hyperparameter following DSMIL [4].

Tab. 9 illustrates the impact of the preset instance retention ratio r on the classification performance. It can be observed that as r increases, both HDMIL† and HDMIL exhibit a simliar trend in performance. On the Camelyon16 dataset, HDMIL† and HDMIL achieve their best results when r is 0.5 and 0.6, respectively. On the TCGA-NSCLC and TCGA-BRCA datasets, both methods reach optimal performance at $r = 0.7$. This may be due to the smaller proportion of tumor regions within Camelyon16 WSIs, allowing the networks to discard more irrelevant instances without affecting classification accuracy.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[4] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 1, 2, 6

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[6] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, et al. Mobilenetv4-universal models for the mobile ecosystem. *arXiv preprint arXiv:2404.10518*, 2024. 5, 1