Multi-marginal Contrastive Learning for Multi-label Subcellular Protein Localization Supplementary Materials

Ziyi Liu, Zengmao Wang*, Bo Du*
National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
School of Computer Science, Wuhan University
Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan, China
{ziyiliu, wangzengmao, dubo}@whu.edu.cn

1. Datasets

We use two single-label datasets and two multi-label datasets in experiments. These datasets consist of IHC images selected from the HPA. Details of the datasets are as following described.

- HPA-7 dataset [6]: This dataset contains 2413 IHC images from 46 normal human tissues (1540 for training and 873 for testing). Nineteen proteins are located in seven main organelles.
- HPA-8 dataset [3]: The IHC images of 16 proteins located in 1 of the 8 main organelles are selected from HPA. A total of 1971 images from 45 types of tissues are found in the dataset (1238 for training and 733 for testing).
- Multi-HPA dataset [4]: This dataset contains 823 protein images belonging to 2 or 3 locations from seven major subcellular locations.
- HPA-18 dataset [2]: The images of proteins are from four organs: liver, bladder, breast, and prostate. Labels are merged into six categories (i.e. nucleus, mitochondria, vesicles, golgi apparatus, endoplasmic reticulum, and cytoplasm) according to the hierarchical structure of organelles. This dataset only chooses images whose staining intensity level is strong or moderate and quantity is higher than 75%. The images belonging to the same protein are either in the training set (including validation) or the test set. Different from other datasets, we need to predict the labels of protein. 1067 and 119 proteins are selected for training and testing, including 7617 and 238 images respectively.

We ensure that the images from different tissues with the same protein both exist in the training and the testing sets.

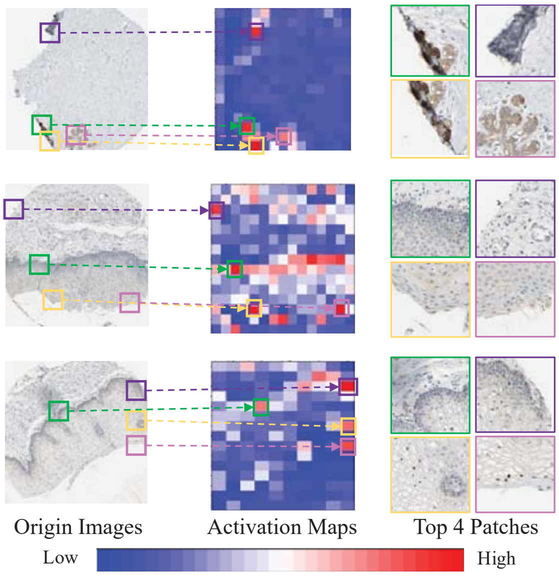


Figure 1. Examples of selection cropped patches from original huge images by activation maps.

2. Patches Selection by Activation Maps

In Figure 1, we show the detailed process of patch selection. Firstly, we generate the activation maps from the channel-wise global average pooling outputs of ResNet features. In Figure 1, we find the top T largest values in the activation maps. Then, we locate the T patches in the original images. Finally, T patches are cropped for further training or test. During the training phase, we set the label of each patch the same as the original images. In the test phase, the probabilities of T patches are averaged to get the final prediction.

3. Evaluation metrics

To evaluate the performance of each method, we choose some popular metrics, such as accuracy, precision, recall,

*Corresponding author.

and F1 score, for the single-label classification task. On the multi-label dataset, label-based metrics(accuracy, precision, recall, F1 score) and example-based metrics(subset accuracy, example-based accuracy, precision, recall, and F1 score) are adopted as evaluation metrics [5,7]. Besides, we select Hamming loss to analyze the sensitivity of parameter β.

3.1. Single-label Metrics

- Accuracy is the percentage of the total number of samples that are correctly classified by the model. The specific formula is shown as:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

- Precision is the percentage of samples with true labels are predicted positive by the classification model. The specific formula is shown as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Recall measures the proportion of the true positive sample that is predicted to be positive by the model.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- F1 score is the summed average of recall and precision. The specific formula is shown as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

3.2. Multi-label Metrics

- The subset accuracy evaluates the fraction of correctly classified examples, i.e. the predicted label set is identical to the ground-truth label set.

$$\text{subset acc}(h) = \frac{1}{p} \sum_{i=1}^{p} I\left[h\left(x_i\right) = Y_i\right] \tag{5}$$

  where p is the total number of samples in the dataset. $h()$ denotes multi-label classifier $h : X \rightarrow 2^Y$, where $h(x)$ returns the set of proper labels for x. I is an indicator function, which returns 0 or 1.

- The example-based metrics includes example-based accuracy, precision, recall and F1 score. The metrics are defined as:

$$Accuracy_{exam}(h) = \frac{1}{p} \sum_{i=1}^{p} \frac{\left|Y_i \cap h\left(x_i\right)\right|}{\left|Y_i \cup h\left(x_i\right)\right|} \tag{6}$$

$$Precision_{exam}(h) = \frac{1}{p} \sum_{i=1}^{p} \frac{\left|Y_i \cap h\left(x_i\right)\right|}{\left|h\left(x_i\right)\right|} \tag{7}$$

$$Recall_{exam}(h) = \frac{1}{p} \sum_{i=1}^{p} \frac{\left|Y_i \cap h\left(x_i\right)\right|}{\left|Y_i\right|} \tag{8}$$

$$F1_{exam}(h) = \frac{2 \times Precsion_{exam}(h) \times Recall_{exam}(h)}{Precision_{exam}(h) + Recall_{exam}(h)} \tag{9}$$

  where || denotes the number of elements in the set.

- The label-based metrics includes label-based accuracy, precision, recall and F1 score. For the jth class label y_j, the following four basic quantities:

$$\begin{aligned} TP_j &= \left|\left\{x_i \mid y_j \in Y_i \wedge y_j \in h\left(x_i\right), 1 \leq i \leq p\right\}\right|, \\ FP_j &= \left|\left\{x_i \mid y_j \notin Y_i \wedge y_j \in h\left(x_i\right), 1 \leq i \leq p\right\}\right|, \\ TN_j &= \left|\left\{x_i \mid y_j \notin Y_i \wedge y_j \notin h\left(x_i\right), 1 \leq i \leq p\right\}\right|, \\ FN_j &= \left|\left\{x_i \mid y_j \in Y_i \wedge y_j \notin h\left(x_i\right), 1 \leq i \leq p\right\}\right|, \end{aligned} \tag{10}$$

  where TP_j, FP_j, TN_j and FN_j represent the number of true positive, false positive, true negative, and false negative samples with respect to y_j, respectively. Label-based accuracy, precision, recall, and F1 score are defined as:

$$Accuracy_{label}(h) = \frac{1}{q} \sum_{j=1}^{q} \frac{TP_j + TN_j}{TP_j + FP_j + TN_j + FN_j} \tag{11}$$

$$Precision_{label}(h) = \frac{1}{q} \sum_{j=1}^{q} \frac{TP_j}{TP_j + FP_j} \tag{12}$$

$$Recall_{label}(h) = \frac{1}{q} \sum_{j=1}^{q} \frac{TP_j}{TP_j + FN_j} \tag{13}$$

$$F1_{label}(h) = \frac{1}{q} \sum_{j=1}^{q} \frac{2 \times TP_j}{2 \times TP_j + FN_j + FP_j} \tag{14}$$

  where q is the number of labels.

3.3. Hamming loss

The Hamming loss counts the number of times the true labels of all samples do not appear in the set of predicted labels [1]. Compared to subset accuracy, Hamming loss can be used to measure the model predicting performance on a single label. The Hamming loss is defined as:

$$hloss(\mathrm{h}) = \frac{1}{N} \sum_{i=1}^{N} \left|h\left(x_i\right) \Delta Y_i\right| \tag{15}$$

where Δ denotes the reciprocal difference between the predicted and target labels. The smaller the calculated value of Hamming's loss, the better the performance of the model.

HPA-7	Acc	Prec	Recall	F1 Score
D	96.18	96.28	96.17	96.19
P	95.36	95.59	95.29	95.37
D+P	**97.95**	**97.98**	**97.96**	**97.96**

HPA-8	Acc	Prec	Recall	F1 Score
D	94.73	95.52	95.48	95.50
P	95.36	95.61	95.29	95.23
D+P	**96.79**	**97.53**	**97.22**	**97.36**

Table 1. Classification results of our method using resample images, cropped patches and both on the HPA-7 and HPA-8 dataset. The bold font indicates the best among compared methods. R + P denotes that both resample images and cropped patches are fed into network.

Multi-HPA	Subset	Example	Example	Example	Example	Label	Label	Label	Label
	acc	acc	prec	recall	F1	acc	prec	recall	F1
D	95.17	95.75	96.03	96.03	96.03	97.59	97.02	95.20	96.10
P	95.17	**97.21**	**97.47**	**98.33**	**97.90**	**98.52**	97.55	**98.13**	**97.84**
D+P	**95.86**	96.98	97.41	97.36	97.39	98.37	**97.65**	96.66	97.15
HPA-18	Subset	Example	Example	Example	Example	Label	Label	Label	Label
	acc	acc	prec	recall	F_1	acc	prec	recall	F_1
D	57.02	63.64	71.07	63.64	67.15	87.33	90.52	28.93	43.85
P	59.50	66.53	73.55	67.36	70.32	88.43	71.55	36.81	48.61
D+P	**61.12**	**68.04**	**74.79**	**68.46**	**71.90**	**88.98**	**89.30**	**37.07**	**52.39**

Table 2. Classification results of our method using downsampled images, cropped patches and both on the Multi-HPA and HPA-18 dataset. The bold font indicates the best among compared methods. D and P denotes that both downsampled images and cropped patches are fed into network.

4. Ablation Study

4.1. Ablation of Global and Local Features

To clarify the advantages of resampled images and cropped patch integration, several experiments have been conducted on the single-label [3,6] and multi-label datasets [2,4]. The whole experimental results are shown in Table 1 and 2.

4.2. Ablation of Patch Selection

In Figure 2, we compare the results of the patch branch with different patch selection strategies, i.e., activation map guided selection and random selection. For different T, networks using activation maps selection yield better performance than random selection. When T is larger than 10, the metrics grow slowly. To balance the calculated consumption and algorithm predicting performance, we set T to 10 in our paper.

4.3. Sensitivity Analysis for Parameter β

We set different values for the hyperparameter β to analyze the effect of the contrastive loss on the DeePSLoc performance and prove that learning the feature similarity between contrastive pairs can help the model extract more discriminative features from IHC images. The details of experimental results are list in Table 3, 4 and 5.

β	Acc	Prec	Recall	F1 Score
0	90.95	92.25	90.95	91.47
0.05	91.98	94.79	91.59	92.87
0.1	93.47	95.38	93.43	94.24
0.15	93.59	95.32	93.41	94.25
0.2	94.29	95.92	93.09	94.42
0.25	**95.57**	**96.53**	**95.50**	**95.96**
0.5	93.58	95.12	93.69	94.32
0.75	92.09	94.69	92.01	93.13

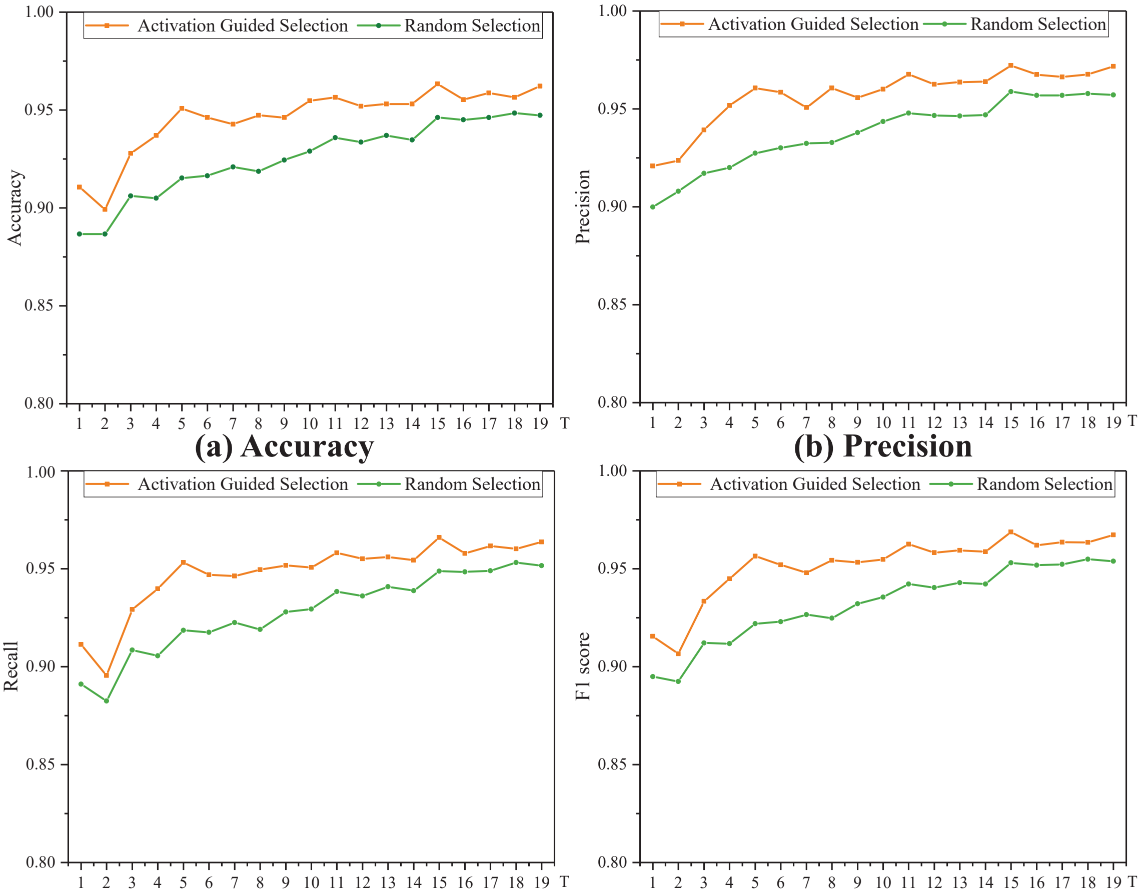
Table 3. Classification results with different values of β in HPA-7.

β	Acc	Prec	Recall	F1 Score
0	89.63	91.43	89.61	90.63
0.05	91.68	92.72	91.68	91.93
0.1	93.45	93.99	93.40	94.01
0.15	93.72	94.32	93.64	93.81
0.2	94.13	94.51	94.03	94.14
0.25	**95.36**	**95.61**	**95.29**	**95.23**
0.5	94.41	94.54	94.33	94.38
0.75	93.59	94.16	93.52	93.66

Table 4. Classification results with different values of β in HPA-8.

β	subset acc	Hamming loss
0	87.93	0.04
0.05	90.69	0.04
0.1	92.07	0.03
0.15	92.76	**0.02**
0.2	94.14	**0.02**
0.25	**95.17**	**0.02**
0.5	94.14	0.03
0.75	85.86	0.07

Table 5. Classification results with different values of β in Multi-HPA.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1 score

Figure 2. Results of DeePSLoc patch branch using activation maps guided and random selection with different T values in HPA-8 dataset. The number of selected patches T is set from 1 to 19.

References

[1] Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012. 2

[2] Wei Long, Yang Yang, and Hong-Bin Shen. Imploc: a multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images. *Bioinformatics*, 36(7):2244–2250, 2020. 1, 3

[3] Justin Y Newberg and Robert F Murphy. A framework for the automated analysis of subcellular patterns in human protein atlas images. *Journal of Proteome Research*, 7(6):2300–2308, 2008. 1, 3

[4] Wei Shao, Mingxia Liu, Yingying Xu, Hongbin Shen, and Daoqiang Zhang. An organelle correlation-guided feature selection approach for classifying multi-label subcellular bioimages. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(3):828–838, 2018. 1, 3

[5] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007. 2

[6] Yingying Xu, Fan Yang, Yang Zhang, and Hongbin Shen. An image-based multi-label human protein subcellular localization predictor (ilocator) reveals protein mislocalizations in cancer tissues. *Bioinformatics*, 29(16):2032–2040, 2013. 1, 3

[7] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013. 2