A Shared Multi-Attention Framework for Multi-Label Zero-Shot Learning

Dat Huynh
Northeastern University
huynh.dat@husky.neu.edu

Ehsan Elhamifar
Northeastern University
eelhami@ccs.neu.edu

1. Evaluation Metrics

1.1. Prediction Performance

To compute the mAP score, we compute the Average Precision of each label as

$$AP_c \triangleq \frac{1}{N_c} \sum_{k=1}^{N} \text{Precision}(k, c) \cdot \text{rel}(k, c), \tag{1}$$

where N_c is the number of images containing label c, Precision(k, c) is the precision for label c when retrieving k best predictions and rel(k, c) is the relevance indicator function that is 1 iff the label c is in the ground-truth of the image at rank k. The mean average precision (mAP) is defined as

$$mAP = 1/|\mathcal{C}| \sum_{c} AP_c, \tag{2}$$

where $|\mathcal{C}|$ is the number of labels.

To compute the F1 score at top K predictions, K labels with highest prediction scores are assigned to each image and are compared with the ground-truth labels. We compute precision and recall for each label independently and report the mean precision and mean recall of all labels as

$$P \triangleq \frac{\sum_c N_c^t}{\sum_c N_c^p}, \quad R \triangleq \frac{\sum_c N_c^t}{\sum_c N_c}, \tag{3}$$

where N_c^t is the number of true positive for label c and N_c^p is the number of positive predictions for label c. F1 score is the harmonic mean between mean Precision and mean Recall and is defined as

$$F1 = \frac{2PR}{P + R}. \tag{4}$$

In the Open Images dataset, due to the large number of classes, each image has unannotated labels. For the F1 score, we treat missing labels as being absent and for the mAP, similar to [1], we evaluate the ranking performance on the labeled data.

1.2. Localization Performance

To provide a quantitative measurement of the localization performance via our attention, we follow [2] to measure whether each label is correctly predicted by paying maximal attention on the ground-truth regions of the label. First, we rescale the attention map to the original size of the image. If the maximal attention region of the attention module chosen to predict a label in an image falls in the ground-truth bounding box of the label, we consider the prediction as true positive, otherwise it is false positive. Finally, we rank all the predictions of each label according to their confidence and compute the Average Precision to measure whether true positive predictions are ranked higher than false positive predictions. Thus, high Average Precision is achieved when the model correctly identifies and localizes present labels.

Notice that this measurement considers all predictions in an image and since only a small fraction of labels is present, most predictions are false positive, resulting in low Average Precision for all methods as reported in the main paper.

2. More Experimental Results

2.1. Multi-Label Learning

In this section, we evaluate the performance of our method for the conventional multi-label learning setting, where all labels have training images, yet the number of samples could be very small or large for different labels.

Baselines: For multi-label learning, we compare with Logistic Regression, WSABIE [3] and WARP [4] (linear embedding methods), Fast0Tag [5] (non-linear embedding), CNN-RNN [6] and One Attention per Label using Bilinear Attention Network [7].

Setting: For NUS-WIDE, we train and test all methods on the set of 81 labels which is annotated by human. Moreover, we follow [6] to remove all test samples without any label in the 81 label set. For Open Images, we train and test on 7,186 seen labels. We set $(\lambda_{div}, \lambda_{rel}, \lambda_{dist})$ to $(1e^{-2}, 1e^{-3}, 1e^{0})$ for NUS-WIDE and $(1e^{-2}, 1e^{-3}, 1e^{-1})$ for Open Images.

Results: Table 1 shows the F1 score at $K \in \{3, 5\}$ of 81 'ground-truth' labels on NUS-WIDE [9], and the F1 score

Method	**NUS-WIDE** multi-label learning							**Open Images** multi-label learning						
	$K = 3$			$K = 5$			mAP	$K = 10$			$K = 20$			mAP
	P	R	F1	P	R	F1		P	R	F1	P	R	F1	
Logistic [8]	46.1	57.3	51.1	34.2	70.8	46.1	21.6	12.2	14.7	13.3	8.4	20.2	11.8	**49.4**
WARP [4]	49.1	61.0	54.4	36.6	75.9	49.4	3.1	7.1	8.5	7.7	5.3	12.6	7.4	46.0
WSABIE [3]	48.5	60.4	53.8	36.5	75.6	49.2	3.1	1.5	3.7	2.2	1.5	3.7	2.2	47.2
Fast0Tag [5]	48.6	60.4	53.8	36.0	74.6	48.6	22.4	14.9	17.9	16.2	9.3	22.3	13.1	45.4
CNN-RNN [6]	49.9	61.7	55.2	37.7	78.1	50.8	28.3	8.7	10.5	9.6	5.4	13.1	10.5	41.0
One Attention per Label [7]	51.3	63.7	56.8	38.0	78.8	51.3	**32.6**	-	-	-	-	-	-	-
One Attention per Cluster ($M = 10$)	51.1	63.5	56.6	37.6	77.9	50.7	31.7	14.9	17.9	16.3	9.2	22.0	13.0	45.1
LESA ($M = 1$)	51.4	63.9	57.0	37.9	78.6	51.2	29.6	15.3	18.4	16.7	9.6	23.2	13.6	45.5
LESA ($M = 10$)	**52.3**	**65.1**	**58.0**	**38.6**	**80.0**	**52.0**	31.5	**16.2**	**19.6**	**17.8**	**10.3**	**24.7**	**14.5**	45.6

Table 1: Performance of **Multi-label** learning methods on NUS-WIDE and Open Images datasets.

Methods	Training (hours)	Inference (sec/img)	Memory (MB)
LESA ($M = 1$)	**2.5**	**0.002**	4150
LESA ($M = 10$)	2.6	**0.002**	4395
One Attention per Label	4.7	0.004	18776
CNN-RNN	7.1	0.009	**2627**

Table 2: Comparison of performance, running time and memory complexity between methods for MLL on NUS-WIDE.

at $K \in \{10, 20\}$ for 7,186 labels on Open Images as well as mAP scores on both datasets. Notice that our method performs on par or better than the state of the art in F1 score across all datasets. We achieve significant improvement with respect to Fast0Tag by 4.2% F1 score at 3 on NUS-WIDE and by 1.6% F1 score at 10 on Open Images. Notice that since Open Images contains a large number of labels, 1.6% translates into $11,497\%$ [1] cumulative improvement over all labels.

For the mAP score, our method gains 9.1% improvement with respect to Fat0Tag on NUS-WIDE. Our method obtains the best F1 score on the NUS-WIDE, and increases the F1 score at 3 by 1.2% with respect to One Attention per Label. However, One Attention per Label achieves the best mAP score performance on NUS-WIDE. This agrees with our observation in the main paper that One Attention per Label performs well when training and testing on seen classes but not in the zero-shot setting in the main paper. Moreover, One Attention per Label cannot scale to thousands of labels in Open Images, due to its large memory requirement. We also observe that the recurrent structure of CNN-RNN has difficulty of capturing correlation between thousands of labels in the Open Images compared to 81 labels in NUS-WIDE. Therefore, it has lower F1 scores than the Logistic baseline.

Notice that all methods except Logistic Regression learn a joint embedding matrix $\boldsymbol{W}_3$ whose rank restricts the set of possible prediction outputs $\{s_i^c\}_{c \in \mathcal{C}}$. On Open Images, the number of labels $\mathcal{C}$ is much larger than the dimension of $\boldsymbol{W}_3$, thus methods based on joint embedding cannot fit data well for the retrieval task. Without this bottleneck, Logistic achieves a high mAP score, yet lower F1 score than ours.

2.2. Multi-Label Zero-Shot Learning

In Table 1 of the main paper, on NUS-WIDE, all methods do better on multi-label zero-shot learning than multi-label generalized zero-shot learning, while on Open Images, the trend is the opposite. We observe on Open Images, the set of seen labels has a significantly large number of positive samples which improve the chance of true positive prediction. On the other hand, labels in the seen set of NUS-WIDE are noisy and sparse, thus decrease the performance.

3. Complexity

Table 2 shows the running time and memory complexity of different methods on the NUS-WIDE dataset. We observe that our method with $M = 10$ shared attention modules has a very similar training and inference time as well as memory complexity compared to using a single attention, and has much better training time and memory complexity compared to One Attention per Label and CNN-RNN. As expected, One Attention per Label has the largest memory requirement, impeding it to be used for classification of a large number of labels. On the other hand, CNN-RNN has significantly larger training time since its sequential structure prevents the training to be parallelized.

References

[1] A. Veit, N. Alldrin, I. K. G. Chechik, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[2] I. L. M. Oquab, L. Bottou and J. Sivic, "Is object localization for free? – weakly-supervised learning with convolutional neural networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[3] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," *IJCAI*, 2011. 1, 2

[1] The mean improvement is multiplied by the total number of labels to get the cumulative improvement.

[4] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, “Deep convolutional ranking for multilabel image annotation,” 2013. 1, 2

[5] Y. Zhang, B. Gong, and M. Shah, “Fast zero-shot image tagging,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[6] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, “Cnn-rnn: A unified framework for multi-label image classification,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[7] J. w. Kim, J. Jun, and B. Zhang, “Bilinear attention networks,” *Neural Information Processing Systems*, 2018. 1, 2

[8] G. Tsoumakas and I. Katakis, “Multi-label classification: An overview,” *Intenational Journal Data Warehousing and Mining*, vol. 3, 2007. 2

[9] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. T. Zheng, “Nus-wide: A real-world web image database from national university of singapore,” *ACM International Conference on Image and Video Retrieval*, 2009. 1