Multiple Instance Visual-Semantic Embedding

BMVC 2017 Submission # 873

1 Multi-label image annotation on NUS-WIDE

1.1 Diagnostic experiments

In supplementary materials, we provide diagnostic experimental results of the proposed Multiple Instance Visual-Semantic Embedding (MIVSE) model, on the task of multi-label image annotation. This corresponds to Section 5.1.5 of the paper.

As introduced in Section 4 of the paper, MIVSE model establishes the mapping relationship between images and labels by modeling the subregion-to-label correspondence with a rank-weighting scheme. To better justify the contribution of several key components to MIVSE, we conduct diagnostic experiments on the task of multi-label image annotation on NUS-WIDE dataset [1]. We compare our method with its four variants, including:

(a). "**Naive multi-label VSE baseline**" which uses the ranking loss objective as shown in Equation 1 of the paper. It is a naive extension of the VSE model from single-label to multi-label senerio; (b). "**MIVSE w/o rank-weighting**" that encodes subregion-to-label correspondence into the ranking loss objective as in Equation 2, while not including rank-weighting as in Equation 4; (c). "**MIVSE w. manual subregions**": instead of using region-proposal method [6], we manually construct the subregion set by selecting subregions with minimum side length as 2, in the 4×4 rigidly defined image grid (totally 36 subregions), and rank-weighting is not included; (d). "**MIVSE w. hinge loss**" that replaces the ranking loss of MIVSE with hinge loss, while keeps the subregion-to-label correspondence and rank-weighting, whose loss function is defined as $L_{\text{hinge}}(\mathbf{x}_i,\mathbf{y}_i)=\sum_{y_p\in\mathbf{y}_i^+} w(r_p)\cdot\max(0,m+\min_{c\in\mathcal{C}_i}||f(\mathbf{x}_i^c)-s(y_p)||_2^2)$. Finally, our full model is named as "**MIVSE full model**".

We compare our full model with those four variants. The results are shown in Table 1. According to the results, we can evaluate the importance of various components of MIVSE below:

1). Importance of modeling subregion-to-label correspondence: As shown in the Table, "**MIVSE w/o rank-weighting**" outperforms "**Naive multi-label baseline**" by 3.13% averaged over all metrics for $k=3$ and 3.12% for $k=5$, which validates the benefit of modeling subregion-to-label correspondence, as in Equation 2.

2). Importance of rank-weighting: By adding a rank-weighting scheme to the loss in Equation 4, our "**MIVSE full model**" further boost the performance of "**MIVSE w/o rank-weighting**" by 0.99% for $k=3$ and 0.95% for $k=5$, which validates the contribution of rank-weighting in our objective function.

	Approach	Rec_L	$Prec_L$	Rec_A	$Prec_A$	N_+
1. $k=3$	Naive multi-label VSE baseline	31.59	34.75	60.26	49.17	98.77
	MIVSE w/o rank-weighting	38.90	**37.87**	63.12	51.55	98.77
	MIVSE w. manual subregions	34.71	35.92	61.87	50.53	98.77
	MIVSE w. hinge loss	28.51	32.63	57.18	47.09	95.06
	MIVSE full model	**40.15**	37.74	**65.03**	**52.23**	**100.00**
2. $k=5$	Naive multi-label VSE baseline	50.25	26.08	75.62	36.94	98.77
	MIVSE w/o rank-weighting	57.79	28.19	79.16	38.14	**100.00**
	MIVSE w. manual subregions	53.92	26.83	76.81	37.78	**100.00**
	MIVSE w. hinge loss	45.98	21.72	71.86	35.10	96.30
	MIVSE full model	**59.81**	**28.26**	**80.94**	**39.00**	**100.00**

Table 1: Image annotation results of MIVSE and its four variants on NUS-WIDE shown in %, with $k=3$ and $k=5$ annotated labels per image, respectively.

3). Importance of subregion set construction: "**MIVSE w/o rank-weighting**" using automatic region-proposals outperforms "**MIVSE w. manual subregions**" that relies on manually generated subregions by 1.68% for $k=3$ and 1.59% for $k=5$. Thus, the effectiveness of constructing subregions using region-proposal method is validated.

4). Importance of developing based on ranking loss: Hinge loss was widely used in the literature [4, 5] for image classification. However, for the problem of visual-semantic embedding with multiple labels, it tends to be sensitive to the noisy labels [2, 3, 7]. We have validated the superiority of ranking loss by comparing "**MIVSE full model**" with "**MIVSE w. hinge loss**". The performance of this variant decreases significantly by 6.94% for $k=3$ and 7.41% for $k=5$.

Thus, we have validated the contribution of several key components of MIVSE model, including the importance of modeling subregion-to-label correspondence, establishing rank-weighting, constructing subregion set, and developing based on ranking loss, etc.

References

[1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, 2009.

[2] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[3] Yunchao Gong, Yangqing Jia, Thomas K. Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. In *ICLR*, 2014.

[4] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[5] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014.

[6] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *ECCV*, 2014.

[7] Jason Weston, Samy Benjio, and Nicolas Usunier. Wsabie: scaling up to large vocabulary image annotation. In *IJCAI*, 2011.