Vision-Language Pre-Training with Triple Contrastive Learning

Jinyu Yang[1], Jiali Duan[2], Son Tran[2], Yi Xu[2], Sampath Chanda[2], Liqun Chen[2], Belinda Zeng[2], Trishul Chilimbi[2], and Junzhou Huang[1]
[1]University Of Texas at Arlington, [2]Amazon
jinyu.yang@mavs.uta.edu, jzhuang@uta.edu
{duajiali,sontran,yxaamzn,csampat,liquchen,zengb,trishulc}@amazon.com

1. Supplementary

This supplementary includes additional details that are not included in the main manuscript due to space limits.

1.1. Feature Visualization

We argue that the main reason for the competitive performance achieved by TCL is that it can learn better intra-modal representations which further contribute to cross-modal representation learning. To validate this assumption, we visualize the t-SNE of text features of the current state of the art [2] (left) and TCL (right) as shown in Figure 1. We can clearly see that the feature representations from TCL are more uniformly distributed, which is desirable for intra-modal retrieval tasks (e.g., text-text retrieval), implying that TCL can learn better intra-modal representations.

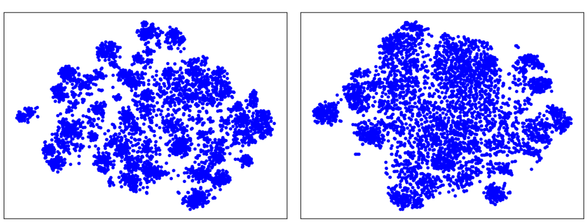

Figure 1. t-SNE visualization of learned features on the COCO dataset.

1.2. Ablation study of the momentum coefficient

To rule out the probability that different experimental settings impact model performance, we set momentum coefficient $m = 0.995$ by following [2]. We retrain our model on COCO [3] with different m to learn the impact of the momentum. Table 1 shows the performance on zero-shot image-text retrieval on Flickr30K [4] and COCO datasets with the evaluation criteria R@1/R@5/R@10. Different from MoCo [1] which claims that a reasonable momentum should be in 0.99∼0.9999, our results suggest that 0.5 performs the best.

m	MSCOCO (5K)						Flickr30K (1K)					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
0.995	60.6	85.9	92.2	46.0	74.1	83.1	67.2	89.3	94.4	52.7	79.0	85.7
0.9	59.7	85.1	92.0	45.5	74.1	83.5	68.0	89.6	94.9	53.3	79.8	86.3
0.5	61.6	85.6	92.2	46.5	74.9	84.0	69.7	89.1	94.3	54.7	79.9	86.9
0.0	61.3	85.8	92.7	46.4	75.2	84.4	70.0	88.6	93.0	53.3	78.5	85.6

Table 1. Ablation study of the momentum coefficient m.

References

[1] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1

[2] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021. 1

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[4] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1