Unifying Visual Contrastive Learning for Object Recognition from a Graph Perspective

Shixiang Tang[1,3], Feng Zhu[3], Lei Bai[2,1,†], Rui Zhao[3,4],
Chenyu Wang[1], and Wanli Ouyang[2,1]

[1] University of Sydney, Australia
[2] Shanghai AI Laboratory, China
[3] Sensetime Research
[4] Qing Yuan Research Institute, Shanghai Jiao Tong University
stan3906@uni.sydney.edu.au, baisanshi@gmail.com [†] corresponding author

1 Transfer to 12 cross-domain classification tasks[5]

In this section, we provide the comparison of our UniVCL with other state-of-the-art methods when we transfer our model to 12 cross-domain classification tasks. Specifically, we follow the setup in BYOL [6]. The datasets for classification task include Food101 [2], CIFAR10 [8], CIFAR100 [8], Birdsnap [1], SUN397 [11], Cars [7], Aircraft [9], VOC2007 [4], DTD [3], Pet2 [10], Caltech-101 [5], and Flowers [6]. Specifically, we freeze the backbone of our pretrained models, and train a classifier on the training set of the datasets mentioned above. We test our models on the testing set of the corresponding dataset. We present these results in Table 1.

Method	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech	Flowers	Avg
BYOL	75.3	91.3	78.4	57.2	62.2	67.8	60.6	82.5	75.5	90.4	**94.2**	96.1	77.6
SimCLR	72.8	90.5	74.4	42.4	60.6	49.3	49.8	81.4	**75.7**	84.6	89.3	92.6	72.0
Sup-IN	72.3	93.6	78.3	53.7	61.9	66.7	61	82.8	74.9	91.5	94.5	94.7	77.2
NNCLR	76.7	**93.7**	79.0	61.4	62.5	67.1	64.1	**83.0**	75.5	91.8	91.3	95.1	78.4
UniVCL	**76.9**	93.1	**79.3**	**64.3**	**62.8**	**67.9**	**64.7**	82.5	75.3	**93.0**	93.5	**96.6**	**79.1**

Table 1. Transfer learning results on fine-grained classification tasks. Specifically, we fix the pretrained backbone, and then train the classifier with the training set of the 12 cross-domain classification datasets. We report the evaluation results by testing the model on the testing set of the dataset.

As shown in Table 1, the transfer results of our method is better than the recent state-of-the-art methods. Specifically, our method is better than other state-of-the-art methods on some fine-grained classfication datasets, *e.g.,* Food101 [2], Birdsnap [1], Cars [7], Aircraft [9], Pets [10] and Flowers [6]. We consider that the reason is that we leverage the neighboring information by the GCN layer, which could exploit the fine-grained information that exist in the ImageNet-1K. For

[5] These experiments are not the improved version of the method UniVCL, just the generalization ability evaluation of the method.

those datasets that have a large domain gap with the ImageNet-1K, *i.e.,* SUN, our method can not help improve the generalization ability. Considering most of the images in ImageNet-1K are object-centric, the scene understanding tasks in SUN can not benefit a lot the self-supervised pretraining from ImageNet-1K.

References

1. Berg, T., Liu, J., Woo Lee, S., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2011–2018 (2014)
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3606–3613 (2014)
4. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
5. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
6. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
7. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
8. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
9. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
10. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
11. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)