Boost Supervised Pretraining for Visual Transfer Learning: Implications of Self-Supervised Contrastive Representation Learning

Jinghan Sun[1,2,*] , **Dong Wei**[2,*] , **Kai Ma**[2] , **Liansheng Wang**[1,†] , **Yefeng Zheng**[2]

[1] Xiamen University, Xiamen, China
[2] Tencent Healthcare (Shenzhen) Co., LTD, Tencent Jarvis Lab, Shenzhen, China
jhsun@stu.xmu.edu.cn, lswang@xmu.edu.cn, {donwei,kylekma,yefengzheng}@tencent.com

A Additional Implementation Details

Common Settings: All experiments are conducted with the PyTorch (Paszke et al. 2019) framework (1.4.0) and four NVIDIA Tesla V100 GPUs.

Few-Shot Recognition: For supervised pretraining, we train 200 epochs with a batchsize of 64. We follow the same augmentation policy as Tian et al. (2020). The initial learning rate is set to 0.05 and decayed twice at 60^{th} and 80^{th} by a factor of 0.1 for all datasets. The SGD optimizer is used with a weight decay of 5×10^{-4} and a momentum of 0.9. For unsupervised pretraining, we train 1,000 epochs for miniImageNet (Vinyals et al. 2016) and CIFAR-FS (Bertinetto et al. 2018), and 500 epochs for tieredImageNet (Ren et al. 2018), respectively, both with an initial learning rate of 0.03. Specifically for MoCo_v2, we use 2,048 negatives for miniImageNet and CIFAR-FS, and 20,480 negatives for tieredImageNet. For the rest, we follow the optimal training strategies as suggested in the original papers. We use the same settings for our proposed CAMtrast as for unsupervised learning. In the pretraining stage, all images are resized to 84×84 pixels. In the testing stage, the images are of size 84×84 pixels for miniImageNet and tieredImageNet, and 32×32 pixels for CIFAR-FS.

PASCAL VOC and Cityscapes: For both datasets, the models are pretrained on the base classes of tieredImageNet as described above. For PASCAL VOC (Everingham et al. 2010) object detection, the input images are resized to 800×800 pixels. We optimize the model with Adam (Kingma and Ba 2014) for 100 epochs with a weight decay of 5×10^{-4}, using a batchsize of 4. The learning rate is initialized to 1×10^{-4} and multiplied by 0.95 per epoch. We fine-tune the pretrained model on the VOC 2007 and VOC 2012 *train* set, and report averaged precision (AP) on the VOC 2007 *test* set. For Cityscapes (Cordts et al. 2016) semantic segmentation, the input images are resized to 768×768 pixels. we fine-tune the models with the SGD optimizer for 30k iterations with a momentum of 0.9, an initial learning rate of 0.01 and a weight decay of 1×10^{-4}, using a batchsize of 8. The standard metric of mean intersection-over-union (mIoU) is reported on the Cityscapes validation set. Online data augmentation including random scaling, cropping, and horizontal flipping is performed during training on both datasets.

B Transfer performance of pretraining on ImageNet

We further evaluate CAMtrast with the ImageNet pretraining, and on VOC07 object detection and COCO object detection and instance segmentation tasks in Table 1. We notice that CAMtrast achieves the best performance on both evaluation datasets and all evaluated metrics, demonstrating generalization on more detection and segmentation tasks and larger databases.

	COCO		VOC07		
Methods	AP_{box}	AP_{seg}	AP	AP_{50}	AP_{75}
Supervised*	38.9	33.9	46.4	77.3	49.0
MoCo_v2*	38.7	34.0	48.5	76.8	52.7
Exemplar_v2*	39.4	34.4	**48.8**	77.2	53.1
CAMtrast (MoCo_v2)	**39.6**	**34.5**	**48.8**	**77.6**	**53.2**

Table 1: Performance on VOC07 object detection and COCO object detection and instance segmentation tasks. The source of non-CAMtrast values is Zhao et al. (2021).

*J. Sun and D. Wei—Contributed equally; J. Sun contributed to this work during an internship at Tencent.

†Corresponding author.

References

Bertinetto, L.; Henriques, J. F.; Torr, P. H.; and Vedaldi, A. 2018. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32: 8026–8037.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of the European Conference on Computer Vision*, 266–282. Springer.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29: 3630–3638.