Revisiting the Critical Factors of Augmentation-Invariant Representation Learning (Supplementary Materials)

Junqiang Huang[†], Xiangwen Kong[†], and Xiangyu Zhang

MEGVII Technology, Beijing, China
{huangjunqiang,kongxiangwen,zhangxiangyu}@megvii.com

1 Setup

1.1 Details for Data Augmentations

In Sec. 4.3, we ablate the data augmentations in self-supervised learning. The combinations of data augmentations we used are the same as the symmetric version proposed in [3], including random cropping, resizing, horizontal flipping, color jittering, gray scale converting, Gaussian blurring, and solarization. The probability and parameters of the data augmentations are detailed in Tab. 1.

Table 1: The probabilities and parameters of the data augmentations

Data Augmentations	Probability	Parameters
Random Crop	1.0	$(0.08, 1.0)$
Resize	1.0	224
Horizontal Flip	0.5	-
Color Jitter	0.8	(0.4, 0.4, 0.2, 0.1)
Gray Scale	0.2	-
Gaussian Blur	0.5	(0.1, 2.0)
Solarization	0.2	128

2 More Experiments

We provide other interesting experiments in Supplementary Materials.

2.1 Longer Pre-training

Here, we benchmark the MoCo v2+ and BYOL with longer training. We also introduce the supervised pre-training as a baseline. We start by investigating

[†]Equal Contribution

12 pre-trained models that vary along two dimensions, training time (i.e., 100, 200, 300, 500 epochs) and pre-training methods (i.e., supervised, MoCo v2+, BYOL). Furthermore, we plot the results of downstream tasks in Fig. 1. For most downstream tasks, it can be seen that all pre-training methods tend to saturation or even degradation with longer training time (e.g., 500 epochs). On the contrary, the linear classification accuracy consistently benefits from longer training. This contradiction unveils the fact that longer training in self-supervised learning helps models better adapt to the pre-training dataset, and does not automatically improve the quality of learned representations.

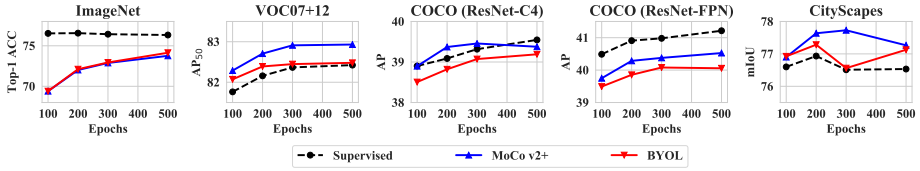

Fig. 1: The linear evaluation and transfer learning results of different pre-training methods with various training time

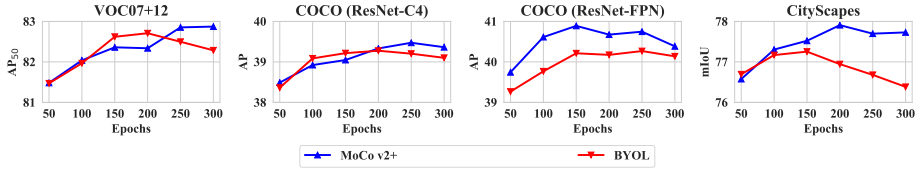

Fig. 2: The transfer performances of intermediate checkpoints

The invalidity of longer training invokes us to examine the overfitting problem in self-supervised learning. We train MoCo v2+ and BYOL on ImageNet for 300 epochs. During training, we fetch the intermediate checkpoints for every 50 epoch, and evaluate them on downstream tasks to see how representations evolve as the optimization proceeds.

The results are plotted in Fig. 2. Without exception, the performances on all downstream tasks meet saturation or even decline after a period of training time, suggesting that overfitting to pretext tasks does happen in self-supervised learning. According to [5], lower layers of convolutional neural networks converge rapidly during training. As for self-supervised pre-training, low-level and mid-level representations that are of more importance in transferring to downstream tasks [6] may stop evolving when the learning rate decays to a small value. We, therefore, hypothesize that currently used optimization strategies (e.g., learning

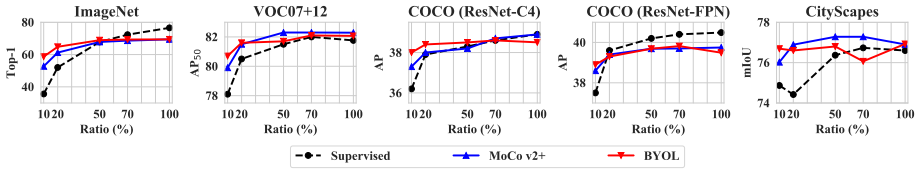

Fig. 3: The results of linear evaluation and downstream tasks under different subsampling ratio

Table 2: The performances of linear evaluation and transfer learning. All models use ResNet-50 as backbone and are trained for 100 epochs. Supervised-X stands for supervised learning on ImageNet-X, where X is the number of classes. Note that $2^\flat$ means class taxonomy with artifact and non-artifact, and that $2^\diamond$ means the class taxonomy with imbalanced data distribution

	ImageNet	VOC07	VOC07+12	COCO				CityScapes
	Acc	AP_{50}	AP_{50}	AP^{C4}_{box}	AP^{C4}_{seg}	AP^{FPN}_{box}	AP^{FPN}_{seg}	mIoU
Supervised-$2^\flat$	9.2	70.2	79.1	36.8	32.3	37.6	34.0	75.7
Supervised-$2^\diamond$	0.7	29.2	46.8	20.9	19.2	26.3	24.4	63.6
Supervised-10	10.6	66.5	77.1	35.8	31.6	36.4	33.0	73.0
Supervised-79	45.4	74.4	80.9	38.3	33.6	39.8	36.0	76.1
Supervised-127	53.3	74.9	81.0	38.5	33.4	39.7	36.0	75.2
Supervised-1000	77.1	76.4	81.8	38.9	33.9	40.5	36.4	76.0
MoCo v2+	69.1	77.0	82.3	38.9	34.1	39.7	35.8	76.9
S-MoCo v2+	69.3	76.0	82.3	38.4	33.6	39.7	36.1	77.1
BYOL	69.4	76.3	82.1	38.5	33.8	39.5	35.7	77.6

rate decays to zero) are the reason for overfitting. A better optimization strategy is waiting to be developed.

2.2 Data Variation

Variation of pre-training data. Inspired by [2], we explore the relationship between self-supervised learning methods and the size of pre-training data. We uniformly sample some classes in ImageNet. For any sampled classes, all its training images will be added to the training set. There are five ratios of sampling: 10%, 20%, 50%, 70%, 100%. To keep training iterations constant, we extend the training epochs adaptively.

As shown in Fig. 3, the linear classification accuracy is monotonically related to the subsampling ratio in both supervised and self-supervised pre-training. The transfer performances reach a plateau with a relatively large subsampling ratio ($\geq$50%). For a low-data regime ($\leq$20%), self-supervised pre-training has a distinct advantage over supervised pre-training in transferring to downstream tasks.

Semantic information of labels. Supervised pre-training with different class taxonomies in ImageNet has been discussed in [4]. In this part, we study the comparison between supervised and self-supervised pre-training given different class taxonomies. The total number of training samples remains the same regardless of the change of taxonomy. We use the WordNet tree [1] and follow the practice of bottom-up clustering in [4], where leaf nodes belonging to the same ancestor are iteratively clustered together. According to this rule, we present four taxonomies that contain 2, 10, 79, and 127 classes respectively. We notice that the sample proportion of 2-class taxonomy is extremely imbalanced (about 1:327). To exclude the effect caused by imbalanced data distribution, we introduce another merging rule, which divides all classes into artifact and non-artifact. The data distribution is well-balanced (an approximate 1:1 ratio). Tab. 2 shows that when the semantic information of labels is inadequate (the number of classes is less than 79) or the data is highly imbalanced (the taxonomy with 2 imbalanced classes), self-supervised learning methods seem to be a better choice for pre-training.

2.3 More Network Architectures

In this subsection, we attempt to make sure that our conclusions still apply to other architectures. We adopt ResNet-18 and ResNet-101 as the backbone. We pre-train MoCo v2, MoCo v2+, and BYOL-SGD for 100 epochs. Tab. 3 shows the linear accuracy of MoCo v2 receives a large promotion with sophisticated model configurations (51.9% vs. 57.3% for ResNet-18 and 65.0% vs. 70.8% for ResNet-101), which is comparable to BYOL-SGD (57.8% for ResNet-18 and 71.7% on ResNet-101). Both MoCo v2+ and BYOL achieve competitive results on VOC07 and VOC07+12.

Archs	Models	ImageNet Acc	VOC07			VOC07+12		
			AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
ResNet-18	MoCo v2	51.9	70.3	40.9	41.0	78.3	50.2	54.2
	MoCo v2+	57.3	71.2	41.2	41.3	78.6	50.7	55.9
	BYOL-SGD	57.8	71.1	42.0	43.3	78.7	51.1	55.4
ResNet-101	MoCo v2	65.0	76.7	50.2	55.1	82.5	59.3	65.8
	MoCo v2+	70.8	76.9	50.3	55.3	82.8	59.1	65.5
	BYOL-SGD	71.7	76.9	50.2	55.1	82.6	59.3	65.6

Table 3: Linear evaluation on ImageNet and detection results on VOC07 and VOC07+12 with ResNet-18 and ResNet-101.

2.4 Other Anchors for NormRescale

Here, we explore other anchor choices for NormRescale. We use the released supervised[1] and self-supervised (MoCo v2[2]) pre-trained model as the anchor to rescale the LARS-trained BYOL.

$\mathbf{w}_S$	VOC07			VOC07+12		
	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
w/o rescale	71.7	38.8	37.0	79.1	48.7	51.7
BYOL-SGD	76.6	48.1	51.6	82.1	56.7	62.9
Constant	76.1	47.9	51.8	82.3	56.8	62.7
MoCo v2	76.4	48.1	52.0	82.2	56.5	62.8
Supervised	76.2	48.1	52.2	82.4	56.7	63.1

Table 4: The fine-tuning results on VOC07 and VOC07+12 of LARS-trained BYOL re-scaled by different SGD-trained models or constant.

As we can see in Tab. 4, Using the released model like supervised or MoCo v2 also bring about good results. Besides, we find the norm values of LARS-trained weight are roughly 10 times to that of SGD-trained weight (also shown in Fig. 3b in our paper), so we simply rescale the weight norm by a factor of 0.1 (abbreviated as "Constant"). Tab. 4 shows that even rescale the norm with a constant, the model does not experience significant performance degradation. NormRescale is rather robust to the choice of anchor model.

[1] https://download.pytorch.org/models/resnet50-0676ba61.pth
[2] https://github.com/facebookresearch/moco

References

1. Fellbaum, C.: Wordnet: An electronic lexical database: Bradford book (1998)
2. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 6390–6399. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00649
3. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent - A new approach to self-supervised learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
4. Huh, M., Agrawal, P., Efros, A.A.: What makes imagenet good for transfer learning? arXiv preprint arXiv:1608.08614 (2016)
5. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer (2014)
6. Zhao, N., Wu, Z., Lau, R.W., Lin, S.: What makes instance discrimination good for transfer learning? arXiv preprint arXiv:2006.06606 (2020)