DisCo: Remedying Self-supervised Learning on Lightweight Models with Distilled Contrastive Learning

Yuting Gao[1*], Jia-Xin Zhuang[1,2*], Shaohui Lin[3], Hao Cheng[4], Xing Sun[1], Ke Li[1†], and Chunhua Shen[5]

[1]Tencent Youtu Lab, [2]Hong Kong University of Science and Technology
[3]East China Normal University [4]University of California, Santa Cruz
[5]Zhejiang University
yutinggao@tencent.com, lincolnz9511@gmail.com

Abstract. While Self-Supervised Learning (SSL) has received widespread attention from the community, recent researches argue that its performance often suffers a cliff fall when the model size decreases. Since current SSL methods mainly rely on contrastive learning to train the network, we propose a simple yet effective method termed **Dis**tilled **Co**ntrastive Learning (DisCo) to ease this issue. Specifically, we find that the final inherent embedding of the mainstream SSL methods contains the most important information, and propose to distill the final embedding to maximally transmit a teacher's knowledge to a lightweight model by constraining the last embedding of the student to be consistent with that of the teacher. In addition, we find that there exists a phenomenon termed Distilling BottleNeck and propose to enlarge the embedding dimension to alleviate this problem. Since the MLP only exists during the SSL phase, our method does not introduce any extra parameters to lightweight models for the downstream task deployment. Experimental results demonstrate that our method surpasses the state-of-the-art on many lightweight models by a large margin. Particularly, when ResNet-101/ResNet-50 is used respectively as a teacher to teach EfficientNet-B0, the linear result of EfficientNet-B0 on ImageNet is improved by 22.1% and 19.7%, respectively, which is very close to ResNet-101/ResNet-50 with much fewer parameters. Code is available at https://github.com/Yuting-Gao/DisCo-pytorch.

Keywords: Self-supervised Learning, Distillation

1 Introduction

Deep learning has achieved great success in computer vision tasks, including image classification, object detection, and semantic segmentation. Such success

* The first two authors contributed equally. This work was done when Jia-Xin Zhuang was an intern at Tencent Youtu Lab.
† Corresponding author: tristanli@tencent.com

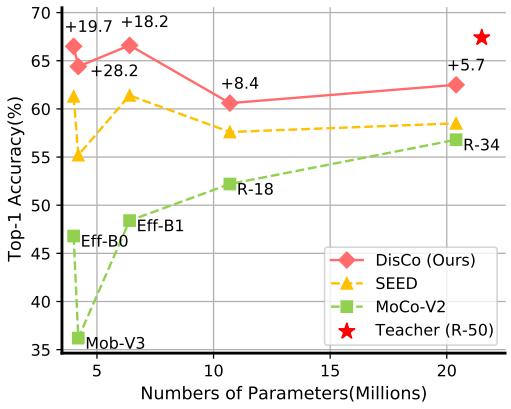

Fig. 1. ImageNet top-1 linear evaluation accuracy on different network architectures. Our method significantly exceeds the result of using MoCo-V2 directly, and also surpasses the state-of-the-art SEED by a large margin. Particularly, the result of EfficientNet-B0 is quite close to the teacher ResNet-50, while the number of parameters of EfficientNet-B0 is only 16.3% of ResNet-50. The improvement brought by DisCo is compared to the MoCo-V2 baseline.

relies heavily on manually labeled datasets, which are time-consuming and expensive to obtain. Therefore, more and more researchers begin to explore how to make better use of off-the-shelf unlabeled data. Among them, SSL is an effective way to explore the information contained in the data itself by using proxy signals as supervision. Usually, after pre-training the network on massive unlabeled data with self-supervised methods and fine-tuning on downstream tasks, the performance of downstream tasks will be significantly improved. Hence, SSL has attracted widespread attention from the community, and many methods have been proposed [6,7,11,14,15,21,25,27]. Among them, methods based on contrastive learning are becoming the mainstream due to their superior results. These methods are constantly refreshing the SOTA results with relatively large networks, but are unsatisfactory on some lightweight models at the same time. For example, the number of parameters of MobileNet-v3-Large/ResNet-152 is 5.2M/57.4M [17,20], and the corresponding linear evaluation top-1 accuracy on ImageNet [30] using MoCo-V2 [8] is 36.2%/74.1%. Compared to their fully supervised counterparts 75.2%/78.57%, the results of MobileNet-v3-Large is far from satisfying. Meanwhile in real scenarios, sometimes only lightweight models can be deployed due to the limited hardware resources. Therefore, improving the ability of self-supervised learning on small models is of great significance.

Knowledge distillation [19] is an effective way to transfer the knowledge learned by a large model (teacher) to a small model (student). Recently, some self-supervised learning methods use knowledge distillation to improve the efficacy of small models. SimCLR-V2 [7] uses logits in the fine-tuning stage to

transfer the knowledge in a task-specific way. CompRess [1] and SEED [13] mimic the similarity score distribution between a teacher and a student model over a dynamically maintained queue. Though distillation is effective, two factors affect the result prominently, *i.e.*, *which* knowledge is most needed by the student, and *how* to deliver it. We propose new insights into these two aspects.

In the current mainstream contrastive learning based SSL methods, a multi-layer perceptron (MLP) is added after the encoder to obtain a low-dimensional embedding. Training loss and the accuracy evaluation are both performed on this embedding. We thus hypothesize that this final embedding contains the most fruitful knowledge and should be regarded as the first choice for knowledge transfer. To achieve this, we propose a simple yet effective DisCo framework to transfer this knowledge from large models to lightweight models in the pre-training stage. Specifically, DisCo takes the MLP embedding obtained by the teacher as the knowledge and injects it into the student by constraining the corresponding embedding of the student to be consistent with that of the teacher using MSE loss. In addition, we find that a budgeted dimension of the hidden layer in the MLP of the student may cause a knowledge transmission bottleneck. We term this phenomenon as *Distilling Bottleneck* and present to enlarge the embedding dimension to alleviate this problem. This simple yet effective operation relates to the capability of model generalization in the setting of self-supervised learning from the Information BottleNeck [33] perspective. It is worth noting that our method only introduces a small number of additional parameters in the pre-training phase, but during the fine-tuning and deployment stage, there is no extra computational burden since the MLP layer is removed.

Experimental results demonstrate that DisCo can effectively transfer the knowledge from the teacher to the student, making the representations extracted by the student more generalized. Our approach is simple and incorporate it into existing contrastive based SSL methods can bring significant gains. Our contributions are summarized as follows:

- We propose a simple yet effective self-supervised distillation method to boost the representation abilities of lightweight models.
- We discover that there exists a phenomenon termed Distilling BottleNeck in the self-supervised distillation stage and propose to enlarge the embedding dimension to alleviate this problem.
- We achieve state-of-the-art SSL results on lightweight models. Particularly, the linear evaluation results of EfficientNet-B0 [32] on ImageNet is quite close to ResNet-101/ResNet-50, while the number of parameters of EfficientNet-B0 is only 9.4%/16.3% of ResNet-101/ResNet-50.

2 Related Work

2.1 Self-supervised Learning

SSL is a generic framework that learns high semantic patterns from data without any tags from human beings. Current methods mainly rely on three paradigms, *i.e.*, pretext tasks based, contrastive learning based, and clustering based.

Pretext tasks based. Approaches based on pretext paradigm first design surrogate tasks, e.g., Rotation [21], Jigsaw [25], and then train the network to solve.
Contrastive learning based. Contrastive learning based approaches have shown impressive performance on self-supervised learning, which enforce different views of the same input to be closer in feature space [9,7,6,18,15,8,14,34,35]. SimCLR-V2 indicates that SSL can be boosted by applying strong data augmentation, training with larger batch size, and adding projection head after the global average pooling. However, SimCLR relies on a very large batch size to achieve comparable performance. MoCo-V2 considers contrastive learning as a look-up dictionary, using a memory bank to maintain consistent representations of negative samples. Thus, MoCo can achieve superior performance without a large batch size, which is more feasible to implement. DINO [5] applies contrastive learning to vision transformers.
Clustering based. Clustering is a kind of promising approach for unsupervised representation learning [3,2]. SwAV [4] maps representations to prototype vectors and is capable to scale to larger datasets.

Mainstream methods from different self-supxervised categories have four things in common: 1) two views for each input image, 2) two encoders for feature extraction, 3) two projection heads to map the representations into a lower dimension space, and 4) the two low-dimensional embeddings are regarded as a pair of positive samples and are pulled closer during training, which can be considered as a contrast process. However, all of these methods suffer a performance cliff fall that is way much more severe than expected on lightweight models, which is what we try to remedy in this work.

2.2 Knowledge Distillation

Knowledge distillation tries to transfer the knowledge from a larger teacher model to a smaller student model. The form of knowledge can be classified into three categories, logits-based, feature-based, and relation-based. Logits-based method KD [19] proposes to make the student mimic the logits of the teacher by minimizing the KL-divergence of the class distribution. Feature-based methods [29,36] directly transfer the knowledge from the intermediate layers of the teacher to the student. AT [36] proposes to use the spatial attention of the teacher as the knowledge and let the student pay attention to the area that the teacher is concerned about. Relation-based approaches explore the relationship between data instead of the output of a single instance. RKD [26] transfers the mutual relationship of the input data within one batch from the teacher to the student. In this work, we use feature-based distillation methods.

2.3 SSL Meets KD

CompRess [1] and SEED [13] try to employ knowledge distillation as a means to improve the representation capability of small models in self-supervised learning, which utilize the negative sample queue in MoCo-V2 to constrain the distribution of positive sample over negative samples of the student to be consistent with that

of the teacher. However, both methods heavily rely on MoCo-V2, which means that a memory bank has to be preserved during the distillation process. Our method also aims to boost the self-supervised learning ability on lightweight models by distilling, however, we do not restrict the self-supervised framework and thus are more flexible. Furthermore, our method surpasses SEED with a large margin on all lightweight models under the same setting.

3 Method

In this section, we introduce the proposed *Distilled Contrastive Learning* (DisCo) framework for lightweight models. We first give some preliminaries on contrastive based SSL and then introduce the overall architecture of DisCo and how DisCo transfers the knowledge from the teacher to the student. Finally, we present how DisCo can be combined with the existing contrastive based SSL methods.

3.1 Preliminary on Contrastive Learning Based SSL

Mainstream contrastive learning-based SSL methods have four commonalities.

Two views: one input image x is transformed into two views v and $v^{'}$ by two drastic data augmentation operations.

Two encoders: two augmented views are input to two encoders of the same structure, one is a learnable base encoder $s(\cdot)$ and the other $m(\cdot)$ is updated according to the base encoder, either shared or momentum updated. The encoder here can use any network architecture, such as the commonly used ResNet. Given an input image, the extracted representation obtained from the last global average pooling of the encoder is denoted as Z, and its dimension is D.

Projection head: both encoders are followed by a small projection head $p(\cdot)$ that maps the representation Z to a low-dimensional embedding E, which contains several linear layers. This procedure can be formulated as $E = p(Z) = W_{(n)} \cdots (\sigma(W_{(1)}Z))$, where W is the weight parameter of the linear layer, n is the number of layers, which is greater than or equal to 1, and σ is the non-linear function ReLU. The importance the of projection head has been addressed in SimCLR-V2 and MoCo-V2. Following MoCo-V2, the default configuration of the projection head is two linear layers, in which the first layer maintains the original feature dimension D, and the second layer reduces the dimension to 128.

Loss function: after obtaining the final embeddings of these two views, they are regarded as a pair of positive samples to calculate the loss.

3.2 Overall Architecture

The framework of DisCo is shown in Figure 2, consisting of three encoders followed by the projection head. The *Student* $s(\cdot)$ in center is the encoder that we want to improve, the *Mean Student* $m(\cdot)$ is updated according to $s(\cdot)$, and *Teacher* $t(\cdot)$ is the self-supervised pre-trained large encoder that is used as teacher in distillation.

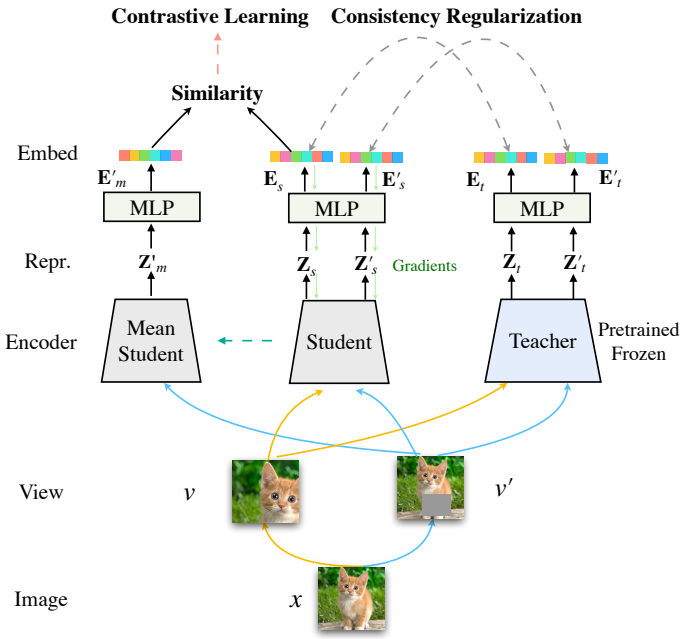

Fig. 2. The framework of the proposed method DisCo. One image is first transformed into two views by two drastic data augmentation operations. In addition to the original contrastive SSL part, a self-supervised pre-trained teacher is introduced, and the final embeddings obtained by the learnable student and the frozen teacher are required to be consistent for each view. Repr. stands for representation.

For each input image x, it is first transformed into two views v and v' by two drastic data augmentation operations. On the one hand, v is input to $s(\cdot)$ and $t(\cdot)$, generating two representations $Z_s = s(v)$, $Z_t = t(v)$, then after the projection head, these two representations are mapped to low-dimensional embeddings, $E_s = p_s(Z_s)$, $E_t = p_t(Z_t)$ respectively. On the other hand, v' is input to $s(\cdot)$, $m(\cdot)$ and $t(\cdot)$ simultaneously, after encoding and projecting, three low-dimensional vectors $E_s^{'} = p_s(s(v'))$, $E_m^{'} = p_m(m(v'))$, and $E_t^{'} = p_t(t(v'))$ are obtained.

$E_m^{'}$ and E_s are the embeddings of two different views, which are regarded as a pair of positive samples and are pulled together in the existing SSL methods. E_s and E_t, $E_s^{'}$ and $E_t^{'}$ are two pairs of embeddings of the student and the teacher of the same view, and each pair is constrained to be consistent during the distilling.

3.3 Distilling Procedure

In most contrastive based SSL methods, the calculation of loss function and the evaluation of accuracy are performed at the final embedding vector E. Therefore, we hypothesize that the last embedding E contains the most fruitful knowledge and should be primarily considered when distilling.

For a self-supervised pre-trained teacher model, we distill the knowledge in the last embedding into the student, that is, for view v and view v', the embedding vector output by the frozen teacher and the learnable student should be consistent. Specifically, we use a consistency regularization term to pull the embedding vector E_s closer to E_t and $E_s^{'}$ closer to $E_t^{'}$. Formally,

$$\mathcal{L}_{dis} = ||E_s - E_t||^2 + ||E_s^{'} - E_t^{'}||^2 \tag{1}$$

To verify that the embedding E contains the most meaningful knowledge, we experiment with several other commonly used distillation schemes in Table 5. The results prove that the knowledge we transmitted and the way it is transferred are indeed the most effective.

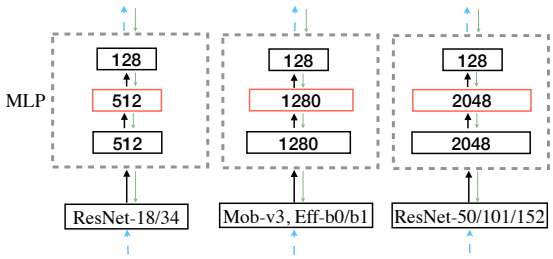

Fig. 3. Default MLP of multiple networks.

Distilling Bottleneck. In our distillation experiment, we found an interesting phenomenon. When the encoder of the student is ResNet-18/34 and the default MLP configuration is adopted, that is, the dimension of embedding output by the encoder is projected from D to D and then to 128, the results of DisCo are not satisfactory. We assume that this degradation is caused by the fact that the dimension of the hidden layer in the MLP is too small, and term this phenomenon as *Distilling Bottleneck*. In Figure 3, we exhibit the default configuration of the projection head of ResNet-18/34, EfficientNet-B0/B1, MobileNet-v3-Large, and ResNet-50/101/152. It can be seen that the dimension of the hidden layer of ResNet-18/34 is too small compared to other networks.

To alleviate the Distilling Bottleneck problem, we expand the dimension of the hidden layer in MLP. It's worth noting that this operation only introduces a small number of parameters at the self-supervised distillation stage, and the MLP will be directly discarded during fine-tuning and deployment, which means no extra computational burden is brought. We experimentally verified that such a simple operation can bring significant gains in Table 4.

This operation can be explained from the Information Bottleneck (IB) [33] perspective. IB is utilized in [31,10] to understand how deep networks work by visualizing mutual information ($I(X;T)$ and $I(T;Y)$) in the information plane, where $I(X;T)$ is the mutual information between input and output, and $I(T;Y)$ is the mutual information between output and label. The training of deep networks can be described by two phases: the first *fitting phase*, where the network

memorizes the information of input, resulting in the growth of $I(X;T)$ and $I(T;Y)$; the subsequent *compression phase*, where the network removes irrelevant information of input for better generalization, resulting in the decrease of $I(X;T)$. Generally, in the *compression phase*, $I(X;T)$ can present the model's capability of generalization while $I(T;Y)$ can present the model's capability of fitting label [10]. We visualize the *compression phase* of our model with different dimensions of the hidden layer in the pre-training distillation stage in the information plane on one downstream transferring classification task. The results in Figure 6 shows two interesting phenomenons:

i. Models with different dimensions of the hidden layer have very similar $I(T;Y)$, suggesting that models have nearly equal capability of fitting the labels.

ii. The Model with a larger dimension in the hidden layer has smaller $I(X;T)$, suggesting a stronger capability of generalization.

These phenomenons show that MLP indeed relates to the capability of model generalization in the setting of self-supervised transfer learning.

3.4 Overall Objective Function

The overall objective function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{dis} + \lambda \mathcal{L}_{co} \tag{2}$$

where $\mathcal{L}_{dis}$ comes from the distillation part, $\mathcal{L}_{co}$ can be the contrastive loss of any SSL method, and λ is a hyper-parameter that controls the weights of the distillation loss and contrastive loss. In our experiments, λ is set to 1. Due to the simplicity of implementation, we use MoCo-V2 as the testbed in the experiments without additional explanation.

4 Experiments

4.1 Settings

Dataset. All the self-supervised pre-training experiments are conducted on ImageNet [30]. For downstream classification tasks, experiments are carried out on Cifar10 and Cifar100 [22]. For downstream detection tasks, experiments are conducted on PASCAL VOC [12] and MS-COCO [23], with train+val/test and train2017/val2017 for training/testing respectively. For downstream segmentation tasks, the proposed method is verified on MS-COCO.

Teacher Encoders. Four large encoders are used as teachers, ResNet-50(22.4M), ResNet-101(40.5M), ResNet-152(55.4M), and ResNet-50*2(55.5M), where X(Y) denotes that the encoder X has Y millions of parameters and the Y does not consider the linear layer.

Student Encoders. Five widely used small yet effective networks are used as student, EfficientNet-B0(4.0M), MobileNet-v3-Large(4.2M), EfficientNet-B1(6.4M), ResNet-18(10.7M) and ResNet-34(20.4M).

Teacher Pre-training Setting. ResNet-50/101/152 are pre-trained using MoCo-V2 with default hyper-parameters. Following SEED, ResNet-50/101 are trained for 200 epochs, and ResNet-152 is trained for 400 epochs. ResNet-50*2 is pre-trained by SwAV, which is an open-source model [1] and trained for 800 epochs.

Self-supervised Distillation Setting. The projection head of all the student networks has two linear layers, with the dimension being 2048 and 128. The configuration of the learning rate and optimizer is set the same as MoCo-V2, and without a specific statement, the model is trained for 200 epochs. During the distillation stage, the teacher is frozen.

Student Fine-tuning Setting. For linear evaluation on ImageNet, the student is fine-tuned for 100 epochs. Initial learning rate is 3 for EfficientNet-B0/EfficientNet-B1/MobileNet-v3-Large, and 30 for ResNet-18/34. For linear evaluation on Cifar10 and Cifar100, the initial learning rate is 3 and all the models are fine-tuned for 100 epochs. SGD is adopted as the optimizer and the learning rate is decreased by 10 at 60 and 80 epochs for linear evaluation. For downstream detection and segmentation tasks, following SEED [13], all parameters are fine-tuned. For the detection task on VOC, the initial learning rate is 0.1 with 200 warm-up iterations and decays by 10 at 18k, 22.2k steps. The detector is trained for 48k steps with a batch size of 32. Following SEED, the scales of images are randomly sampled from [400, 800] during the training and is 800 at the inference. For the detection and instance segmentation on COCO, the model is trained for 180k iterations with the initial learning rate 0.11, and the scales of images are randomly sampled from [600, 800] during the training.

4.2 Linear Evaluation

We conduct linear evaluation on ImageNet to validate the effectiveness of our method. As shown in Table 1, student models distilled by DisCo outperform the counterparts pre-trained by MoCo-V2 (Baseline) with a large margin. Besides, DisCo surpasses the state-of-the-art SEED over various student models with teacher ResNet-50/101/152 under the same setting, especially on MobileNet-v3-Large distilled by ResNet-50 with a difference of 9.2% at top-1 accuracy. When using R50*2 as the teacher, SEED distills 800 epochs while DisCo still distills 200 epochs, but the results of EfficientNet-B0, ResNet-18, and, ResNet-34 using DisCo also exceed that of SEED. The performance on EfficientNet-B1 and MobileNet-v3-Large is closely related to the epochs of distillation. For example, when EfficientNet-B1 is distilled for 290 epochs, the top-1 accuracy becomes 70.4%, which surpasses SEED and when MobileNet-v3-Large is distilled for 340 epochs, the top-1 accuracy becomes 64%. We believe that when DisCo distills 800 epochs, the results will be further improved. Moreover, since CompRess uses a better teacher which trained 600 epochs longer and distills 400 epochs longer than SEED and ours, it's not fair to compare thus we do not report the result in the table.

[1] `https://github.com/facebookresearch/swav`

Table 1. ImageNet test accuracy (%) using linear classification on different student architectures. ◊ denotes the models are pre-trained with MoCo-V2, which is our implementation and †means the teacher is pre-trained by SwAV, which is an open-source model. When using R50*2 as the teacher, SEED distills 800 epochs while DisCo distills 200 epochs. Subscript in green represents the improvement compared to MoCo-V2.

Method	S / T	Eff-b0		Eff-b1		Mob-v3		R-18		R-34	
		T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Supervised		77.1	93.3	79.2	94.4	75.2	-	72.1	-	75.0	-
Self-supervised											
MoCo-V2 (Baseline)◊		46.8	72.2	48.4	73.8	36.2	62.1	52.2	77.6	56.8	81.4
SSL Distillation											
SEED[13]	R-50 (67.4)	61.3	82.7	61.4	83.1	55.2	80.3	57.6	81.8	58.5	82.6
DisCo (ours)	R-50 (67.4)◊	**66.5** (19.7↑)	**87.6** (15.4↑)	**66.6** (18.2↑)	**87.5** (13.7↑)	**64.4** (28.2↑)	**86.2** (24.1↑)	**60.6** (8.4↑)	**83.7** (6.1↑)	**62.5** (5.7↑)	**85.4** (4.0↑)
SEED [13]	R-101 (70.3)	63.0	83.8	63.4	84.6	59.9	83.5	58.9	82.5	61.6	84.9
DisCo (ours)	R-101 (69.1)◊	**68.9** (22.1↑)	**88.9** (16.7↑)	**69.0** (20.6↑)	**89.1** (15.3↑)	**65.7** (29.5↑)	**86.7** (24.6↑)	**62.3** (10.1↑)	**85.1** (7.5↑)	**64.4** (7.6↑)	**86.5** (5.1↑)
SEED [13]	R-152 (74.2)	65.3	86.0	67.3	86.9	61.4	84.6	59.5	83.3	62.7	85.8
DisCo (ours)	R-152 (74.1)◊	**67.8** (21.0↑)	**87.0** (14.8↑)	**73.1** (24.7↑)	**91.2** (17.4↑)	**63.7** (27.5↑)	**84.9** (22.8↑)	**65.5** (13.3↑)	**86.7** (9.1↑)	**68.1** (11.3↑)	**88.6** (7.2↑)
SEED [13]	R50*2 (77.3†)	67.6	87.4	68.0	87.6	68.2	88.2	63.0	84.9	65.7	86.8
DisCo (ours)	R50*2 (77.3)†	**69.1** (22.3↑)	**88.9** (17.7↑)	64.0 (15.6↑)	84.6 (10.8↑)	58.9 (22.7↑)	81.4 (19.3↑)	**65.2** (13↑)	**86.8** (9.2↑)	**67.6** (10.8↑)	**88.6** (7.2↑)

In addition, when DisCo uses a larger model as the teacher, the student will be further improved. For instance, using ResNet-152 instead of ResNet-50 as the teacher, ResNet-34 is improved from 62.5% to 68.1%. It's worth noting, when using ResNet-101/ResNet-50 as the teacher, the linear evaluation result of EfficientNet-B0 is very close to the teacher, while the number of parameters of EfficientNet-B0 is only 9.4%/16.3% of ResNet-101/ResNet-50.

4.3 Semi-supervised Linear Evaluation

Following SEED, we evaluate our method under the semi-supervised setting. Two 1% and 10% sampled subsets of ImageNet training data (∼12.8 and ∼128 images per class respectively) [6] are used for fine-tuning the student models. As is shown in Figure 4, student models distilled by DisCo outperform baseline under any amount of labeled data. Furthermore, DisCo also shows the consistency under different fractions of annotations, that is, students always benefit from larger models as teachers. More labels will be helpful to improve the final performance of the student model, which is expected.

4.4 Transfer to Cifar10/Cifar100

In order to analyze the generalization of representations obtained by DisCo, we further conduct linear evaluation on Cifar10 and Cifar100 with ResNet-18/EfficientNet-B0 as student and ResNet-50/ResNet101/ResNet152 as a teacher. Since the image resolution of the Cifar dataset is 32×32, all the images are resized to 224×224 with bicubic re-sampling before feeding into the model, following

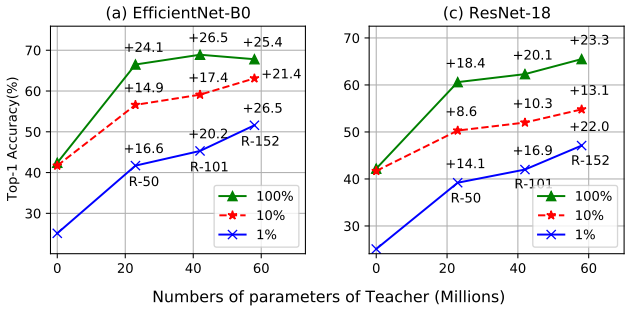

Fig. 4. ImageNet top-1 accuracy (%) of semi-supervised linear evaluation with 1%, 10% and 100% training data. Points where the number of teacher network parameters are 0 are the results of the MoCo-V2 without distillation.

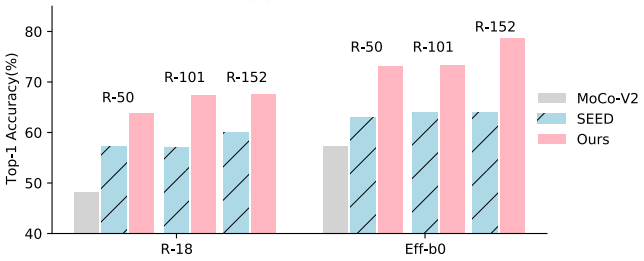

Fig. 5. Top-1 accuracy of students transferred to Cifar100 with and without distillation from different teachers.

[13]. The results are shown in Figure 5, it can be seen that the proposed DisCo surpasses the MoCo-V2 baseline by a large margin with different student and teacher architectures on and Cifar100. In addition, our method also has a significant improvement compared to the-state-of-art method SEED. It is worth noting that as the teacher becomes better, the improvement brought by DisCo is more obvious. The performance trend on Cifar10 is consistent with that on Cifar100, see section 2 in the supplementary material for details.

4.5 Transfer to Detection and Segmentation

We conduct experiments on detection and segmentation tasks for generalization analysis. C4-based Faster R-CNN [28] is used for objection detection on VOC and Mask R-CNN [16] is used for objection detection and instance segmentation on COCO. The results are shown in Table 2. On object detection, our method can bring obvious improvement on both VOC and COCO datasets. Furthermore, as SEED [13] claimed, the improvement on COCO is relatively minor compared to VOC since COCO training dataset has 118k images while VOC has only 16.5k training images, thus, the gain brought by weight initialization is relatively small. On the instance segmentation task, DisCo also shows superiority.

Table 2. Object detection and instance segmentation results on VOC07 test and COCO val2017 with ResNet-34 as backbone. ‡means our implementation. Subscript in green represents the improvement compared to MoCo-V2 baseline.

S	T	Method	Object Detection						Instance Segmentation		
			VOC			COCO			COCO		
			AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
R-34	×	MoCo-V2‡	53.6	79.1	58.7	38.1	56.8	40.7	33.0	53.2	35.3
	R-50	SEED [13]	53.7	79.4	59.2	38.4	57.0	41.0	33.3	53.2	35.3
		DisCo (ours)	**56.5**	**80.6**	**62.5**	**40.0**	**59.1**	**43.4**	**34.9**	**56.3**	**37.1**
			(2.9↑)	(1.5↑)	(3.8↑)	(1.9↑)	(2.3↑)	(2.7↑)	(1.9↑)	(3.1↑)	(1.8↑)
	R-101	SEED [13]	54.1	79.8	59.1	38.5	57.3	41.4	33.6	54.1	35.6
		DisCo (ours)	**56.1**	**80.3**	**61.8**	**40.0**	**59.1**	**43.2**	**34.7**	**55.9**	**37.4**
			(2.5↑)	(1.2↑)	(3.1↑)	(1.9↑)	(2.3↑)	(2.5↑)	(1.9↑)	(2.7↑)	(1.8↑)
	R-152	SEED [13]	54.4	80.1	59.9	38.4	57.0	41.0	33.3	53.7	35.3
		DisCo (ours)	**56.6**	**80.8**	**63.4**	**39.4**	**58.7**	**42.7**	**34.4**	**55.4**	**36.7**
			(3.0↑)	(1.7↑)	(5.7↑)	(1.3↑)	(1.9↑)	(2.0↑)	(1.4↑)	(2.2↑)	(1.4↑)

Table 3. Linear evaluation top-1 accuracy (%) on ImageNet.

Method	Eff-b0	Mob-v3	R-18	R-34
SEED	61.3	55.2	57.6	58.5
DisCo*	65.6	63.8	57.1	58.9
DisCo	**66.5**(0.9↑)	**64.4**(0.6↑)	**60.6**(3.5↑)	**62.5**(3.6↑)

4.6 Distilling BottleNeck Phenomenon

In the self-supervised distillation stage, we first tried to distill small models with default MLP configuration of MoCo-V2 using ResNet-50 as a teacher, and the results are shown in Table 3, denoted by DisCo*. It is worth noting that the dimensions of the hidden layer in DisCo* are exactly as same as SEED. It can be seen that compared to SEED, DisCo* shows superior results on EfficientNet-B0, and MobileNet-v3-Large, and has comparable results on ResNet-18. Then we expand the dimension of the hidden layer in the MLP of the student to be consistent with that of the teacher, that is, $2048D$, it can be seen that the results can be further improved, which is recorded in the third row. This expansion operation brings 3.5% and 3.6% gains for ResNet-18 and ResNet-34 respectively.

Theoretical Analysis from IB perspective. In Figure 6, on the downstream Cifar10 classification task, we visualize the *compression phase* of ResNet-18/34 with different hidden dimensions distilled by the same teacher in the information plane. Following [10], we use binning strategy [24] to estimate mutual information. It can be seen that when we adjust the hidden dimension in the MLP of ResNet-18 and ResNet-34 from $512D$ to $2048D$, the value of $I(X;T)$ becomes smaller while $I(T;Y)$ is basically unchanged, which suggests that enlarging the hidden dimension can make the student model more generalized in the setting of self-supervised transfer learning.

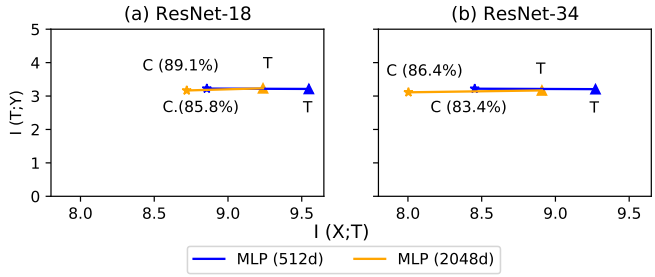

Fig. 6. Mutual information paths from transition points to convergence points in the compression phase of training. T denotes transition points, and C(X%) denotes convergent points with X% top-1 accuracy on Cifar10. Points with similar I(T;Y) but smaller I(X;T) are better generalized.

Table 4. Linear evaluation top-1 accuracy (%) on ImageNet. MLP-d means the hidden dimension of MLP and - denotes the hidden layer of the MLP is directly removed.

Loss		MLP-d	Eff-b0	Mob-v3	R-18
L_{co}	L_{dis}				
Baseline					
✓		1280/1280/512	46.8	36.2	52.2
Effectiveness of loss					
	✓	1280/1280/512	65.6	58.9	54.5
✓	✓	1280/1280/512	65.6	63.7	57.1
Effectiveness of MLP-d					
✓	✓	-/-/-	52.5	60.3	52.5
✓	✓	512/512/512	62.5	62.8	57.1
✓	✓	1024/1024/1024	65.0	63.8	59.2
✓	✓	2048/2048/2048	**66.5**	**64.4**	**60.6**

4.7 Ablation Study

In this section, we testify the effectiveness of two important modules in DisCo, i.e. the distillation loss and the expansion of the hidden dimension of MLP, and the results are shown in Table 4. It can be seen that distillation loss can bring about essential changes, and the result will be greatly improved. Even with only distillation loss, good results can be achieved. Furthermore, as the hidden dimension increases, the top-1 accuracy also increases, but when the dimension is already large, the growth trend will slow down.

4.8 Comparison against other Distillation

We compare with three widely used distillation schemes, namely, 1) *Attention transfer* denoted by AT [36], 2) *Relational knowledge distillation* denoted by RKD [26] 3) *Knowledge distillation* denoted by KD [19]. AT and RKD are feature-based and relation-based respectively, which can be utilized during the

Table 5. Top-1 accuracy (%) on ImageNet compared with various distillation methods.

Method	Eff-b0	Eff-b1	Mob-v3	R-18
Baseline				
MoCo-V2	46.8	48.4	36.2	52.2
Single-Knowledge				
AT	57.1	58.2	51.0	56.2
RKD	48.3	50.3	36.9	56.4
KD	46.5	48.5	37.3	51.5
DisCo (ours)	**66.5**	**66.6**	**64.4**	**60.6**
Multi-Knowledge				
AT + DisCo	66.7	66.3	64.1	60.0
RKD + DisCo	**66.8**	**66.5**	64.4	**60.6**
KD + DisCo	65.8	65.9	**65.2**	**60.6**

self-supervised pre-training stage. KD is a logits-based method, which can only be used at the supervised fine-tuning stage. The comparison results are shown in Table 5. *Singe-Knowledge* means using one of these approaches individually, and it can be seen that all distillation approaches can bring improvement to the baseline but the gain from DisCo is the most significant, which indicates the knowledge that DisCo has chosen to transfer and the way of transmission is indeed more effective. Then, we also try to transfer multi-knowledge from teacher to student by combining DisCo with other schemes. It can be seen that integrating DisCo with AT/RKD/KD can boost the performance a lot, which further proves the effectiveness of DisCo.

4.9 More SSL Methods

We further experiment with two SSL methods that are quite different from the MoCo-V2. i) SwAV is used to testify to the compatibility of the learning paradigm, in which the difference is measured between clusters instead of instances (see supplementary section 3). ii) DINO is used to testify the compatibility towards the backbone type, in which the encoder is a vision transformer instead of CNN, as is shown in Table 3 in the supplemental material. DisCo can bring significant improvement under most of the popular SSL frameworks.

5 Conclusion

In this paper, we propose DisCo to remedy self-supervised learning on lightweight models. The proposed method constraints the final embedding of the lightweight student to be consistent with that of the teacher to maximally transmit the teacher's knowledge. DisCo is not limited to specific contrastive learning methods and can remedy student performance by a large margin.

Acknowledgements This paper is sponsored by the National Natural Science Foundation of China (NO. 62102151), Shanghai Sailing Program (21YF1411200), and CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-031A)

References

1. Abbasi Koohpayegani, S., Tejankar, A., Pirsiavash, H.: Compress: Self-supervised learning by compressing representations. In: NeurIPS. pp. 12980–12992 (2020)
2. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: ICLR (2020)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV. pp. 132–149 (2018)
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS. pp. 9912–9924 (2020)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
7. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS. pp. 22243–22255 (2020)
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. In: CVPR. pp. 9729–9738 (2020)
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: arXiv preprint arXiv:2011.10566 (2020)
10. Cheng, H., Lian, D., Gao, S., Geng, Y.: Evaluating capability of deep neural networks for image classification via information plane. In: ECCV. pp. 168–182 (2018)
11. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. pp. 1422–1430 (2015)
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge. vol. 88, pp. 303–338 (2010)
13. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: Seed: Self-supervised distillation for visual representation. In: ICLR (2021)
14. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS. pp. 21271–21284 (2020)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
18. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: ICML. pp. 4182–4192 (2020)
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPSW (2015)
20. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: ICCV. pp. 1314–1324 (2019)
21. Komodakis, N., Gidaris, S.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
22. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Citeseer (2009)

23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
24. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
25. Noroozi, M., Favaro, P.: Unsupervised learning of vis,ual representations by solving jigsaw puzzles. In: ECCV. pp. 69–84 (2016)
26. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR. pp. 3967–3976 (2019)
27. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR. pp. 2536–2544 (2016)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. vol. 39, pp. 1137–1149 (2015)
29. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: ICLR (2014)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. vol. 115, pp. 211–252 (2015)
31. Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information (2017)
32. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. pp. 6105–6114 (2019)
33. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method (2000)
34. Wang, J., Gao, Y., Li, K., Jiang, X., Guo, X., Ji, R., Sun, X.: Enhancing unsupervised video representation learning by decoupling the scene and the motion (2020)
35. Wang, J., Gao, Y., Li, K., Lin, Y., Ma, A.J., Sun, X.: Removing the background by adding the background: Towards background robust self-supervised video representation learning (2020)
36. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)