Boosting Contrastive Learning with Relation Knowledge Distillation

Kai Zheng, Yuanjiang Wang[*]**, Ye Yuan**

Megvii Technology
{zhengkai, wangyuanjiang, yuanye}@megvii.com

Abstract

While self-supervised representation learning (SSL) has proved to be effective in the large model, there is still a huge gap between the SSL and supervised method in the lightweight model when following the same solution. We delve into this problem and find that the lightweight model is prone to collapse in semantic space when simply performing instance-wise contrast. To address this issue, we propose a relation-wise contrastive paradigm with Relation Knowledge Distillation (ReKD). We introduce a heterogeneous teacher to explicitly mine the semantic information and transferring a novel relation knowledge to the student (lightweight model). The theoretical analysis supports our main concern about instance-wise contrast and verify the effectiveness of our relation-wise contrastive learning. Extensive experimental results also demonstrate that our method achieves significant improvements on multiple lightweight models. Particularly, the linear evaluation on AlexNet obviously improves the current state-of-art from 44.7% to 50.1%, which is the first work to get close to the supervised (50.5%). Code will be made available.

Introduction

The rise of Deep Convolutional Neural Networks (DCNN) has led to significant success in computer vision benchmarks. Such success relies heavily on massive labeled datasets, which are prohibitively expensive to obtain. Therefore, self-supervised learning (SSL), an effective way to learn visual representations from unlabeled data, has attracted widespread attention among researchers. A variety of different self-defined pretext tasks (Komodakis and Gidaris 2018; Feng, Xu, and Tao 2019; Zhang, Isola, and Efros 2016; Noroozi and Favaro 2016; Zhang et al. 2019) have been proposed. Recently, instance discrimination (Wu et al. 2018; He et al. 2020; Chen et al. 2020b,a; Grill et al. 2020) has emerged as a dominant pretext task in unsupervised learning. This task considers each image in the dataset as an independent class, which makes the model learn discriminative feature under contrastive objective.

The renewed interest in exploring contrastive learning has derived several awesome works (Chen et al. 2020b; Grill
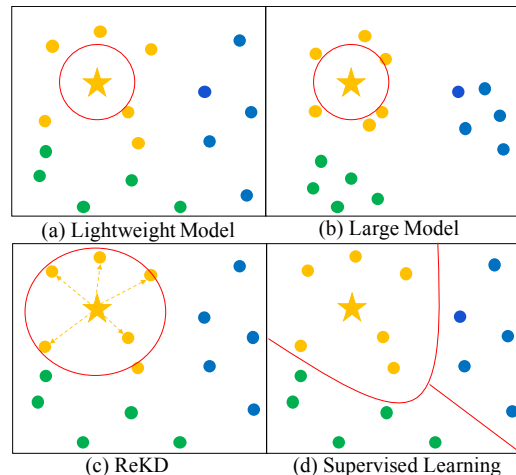
---
[*]Corresponding author(yuanjiang.wang@outlook.com).

Figure 1: Different contrastive learning paradigms. Large model has better semantic feature space (a) than lightweight model (b) under the "one-positive" instance-wise contrastive learning. (c) Our ReKD builds the relation between the instances in semantic space for lightweight model, which roles as semantic label in supervised contrastive learning (d).

et al. 2020; Chen et al. 2020a; Caron et al. 2020), some even close the gap between unsupervised-based method and supervised-based method, which attracts more researches' attention into this field. However, these works focused on how to boost the performance of large models like ResNet-50 and ResNet-50x4, and rare of them paid attention to the lightweight models like MobileNet (Howard et al. 2019) and EfficientNet (Tan and Le 2019). In practice, engineers prefer to select efficient models with low computational complexity and the least storage requirements, which make them easier to deploy in real-time applications, such as video surveillance and autonomous vehicles. Therefore, we pay attention to the SSL on lightweight models' performance. However, the gap between the unsupervised method and supervised method is rather huge in lightweight models. Specifically, we find that MobileNet and EfficientNet are the most typical examples, whose supervised training accuracy is 75.2% and 77.1%, but their unsupervised linear evaluation accuracies using MoCov2 are only 33.3% and 37.9%, which is far from satisfying when compared to ResNet-50's result on supervised (76.5%) and unsupervised (67.6%) training. Mean-

while, the similar conclusion is observed in some recent works (Fang et al. 2020; Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020). Based on the practical usage of the lightweight model but with large margin exists in unsupervised training, this becomes a problem demanding prompt solution for the community.

In this work, we delve into the unsupervised learning and find that those instance-based methods all share a common issue: instances that expect to be close are undesirably pushed apart in the embedding space regardless of the intrinsic semantics in the instance, which might leads to the wrong optimization direction eventually. In our work, we term this phenomenon in SSL as *semantic collapse*. We also observe that this phenomenon varies in different capacity models. The large model that has a superior feature representation, where similar semantic instances are closer than the lightweight model in the embedding space, is less likely to be involved into *semantic collapse* (See Fig. 1(a)/(b)). The tiny accuracy gap between unsupervised and supervised training in the large model proves this. In contrast, the lightweight model with low capacity may easily fall into the semantic trap if using instance-wise contrast (See Fig. 1(a)/(d)), which is detrimental to learn a generalized feature representation.

To solve this issue, we present a **Re**lation **K**nowledge **D**istillation (ReKD) for contrastive learning, which is tailored for lightweight model with junior capacity in feature representation. In ReKD, a *relation knowledge* is proposed to explicitly build the relation between the instances in the semantic space. This knowledge can alleviate the *semantic collapse* existing in instance-based methods, where the semantic information inside the instance is ignored (See Fig. 1(a)/(c)). To acquire the semantic relation knowledge for the lightweight model, we introduce a *heterogeneous teacher* with a *relation miner*. Given the relation knowledge, we optimize the student (lightweight model) by minimizing our proposed *relation contrastive loss*.

In ReKD, we breaks "one-positive" limitation in most instance discriminative methods (Wu et al. 2018; He et al. 2020; Chen et al. 2020a; Grill et al. 2020), and provides informative positives from the semantic level for a better contrastive objective. With this objective, the student obtains fruitful semantic knowledge from a heterogeneous teacher in the feature space, and learns a generalized representation compared with other alternatives. Furthermore, ReKD builds the bridge between clustering-based and contrastive-based method for a better self-supervised visual representation learning. Besides, ReKD is an efficient parallel computing method compared to recent self-supervised knowledge distillation (SSKD) methods (Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020; Fang et al. 2020) that requires a long time for pre-training an *offline teacher*.

Overall, the main contributions of this work include three-fold: (i) We propose a relation knowledge distillation (ReKD) framework tailored for contrastive learning, which also builds the bridge linking cluster-based SSL and contrastive-based SSL. (ii) We provide some insights to demonstrate that our relation knowledge can help mitigate the semantic collapse theoretically. (iii) ReKD achieves a significant boost in multiple lightweight models. Notably, the SSL result on AlexNet almost close the gap with supervised learning. Meanwhile, the improvement against SSKD also verifies our method's effectiveness.

Related Work

Instance Discriminative learning. Instance discriminative based methods (Oord, Li, and Vinyals 2018; Wu et al. 2018) formulate a contrastive learning to learn feature representation, which usually contrasts between one positive and multiple negatives. MoCo (He et al. 2020) and SimCLR (Chen et al. 2020a) obtain the positive from another view generated by the data augmentation on the same image and contrast them against massive negatives. All these methods can be summarized as instance discrimination work that regards an instance as a single class. BYOL (Grill et al. 2020) and SimSiam (Chen and He 2020) come up with a negative-free method, which achieves a competitive result only by constraining the similarity between positives without any negative instances. However, to our knowledge, these instance discrimination works all have an inescapable defect: treating all the instances as independent classes. This deficiency may lead all the instances to be separated apart regardless of whether they belong to the same semantic class or not, which can hurt the semantic-level representation in the model, especially in the lightweight model.

Knowledge Distillation. Knowledge distillation aims to transfer the knowledge learned by a larger model to a smaller one without losing important information. Many forms of knowledge and distillation strategies have been proposed to explore the best way for knowledge distillation. (Hinton, Vinyals, and Dean 2015) proposes using logits with temperature to transfer the category distribution from teacher to student as additional supervision besides the original classification loss. (Romero et al. 2014; Komodakis and Zagoruyko 2017) distill the knowledge via feature/attention map. (Park et al. 2019) involves the mutual relation of data samples as the knowledge. (Chen, Su, and Zhang 2019; Park and Kwak 2020; Shen et al. 2019) propose the multi-teacher scheme to provide diverse knowledge from different teachers to benefit the student. Recently, some works (Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020; Fang et al. 2020) extend the knowledge distillation into self-supervised learning, which formulates the knowledge as the probability distribution. To our knowledge, most of these methods rely heavily on a powerful teacher model that requires a long time pretraining, namely offline teacher. This offline distillation pays little attention to the compatibility with the student and the vast time cost brought by training a teacher. In our work, we look into whether an online teacher can perform as well as an offline teacher or even better. Besides, we formulate an online relation knowledge distillation that is tailored for semantic contrastive objective.

Clustering-based Learning. Early methods (Xie, Girshick, and Farhadi 2016; Yang, Parikh, and Batra 2016; Yang et al. 2017; Chang et al. 2017) aim at integrating clustering into representation learning, which use the semantic clustering results as the supervision to optimize the network.
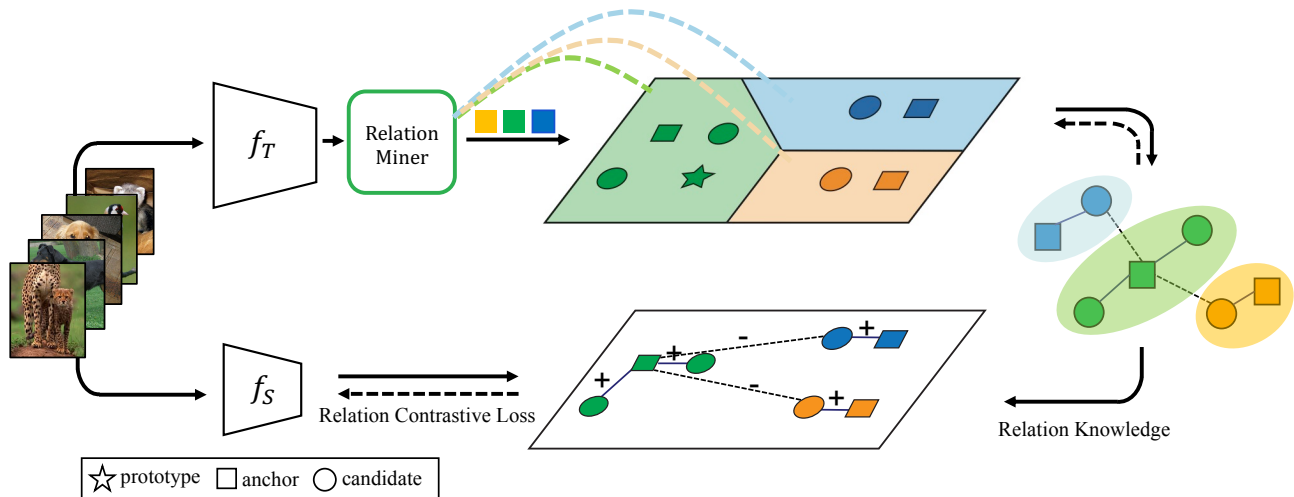
Figure 2: The pipeline of ReKD. A batch of images is fed into the heterogeneous teacher f_T and student f_S simultaneously. The features from heterogeneous teacher go through a relation miner, where the online clustering strategy builds the relation between the candidate and the input anchor through a bank of semantic prototypes. The relation topological structure from the relation miner serves as the relation knowledge to the student for distillation. With the relation contrastive loss, the student and heterogeneous teacher can optimize towards the semantic contrastive objective.

DeepCluster (Caron et al. 2018) uses the clustering labels as the pseudo label to train a classification network. LocalAgg (Zhuang, Zhai, and Yamins 2019) proposes the concept of close neighbor and background neighbor and aims to divide all the samples within the cluster around an anchor into two types. SwAV (Caron et al. 2020) treats the label assignment as an optimal transport problem to get the clustering result and optimize the network. Apart from the conventional clustering, (Huang, Gong, and Zhu 2020) formulates the clustering process as optimization of network in terms of the cluster assignment constraint. Inspired by these works, we find that the cluster-based method can provide extra semantic information in terms of the clustering result. Hence, in our work, we aim to leverage clustering to benefit contrastive learning with semantic information.

Preliminary

Knowledge Distillation. Knowledge distillation (Hinton, Vinyals, and Dean 2015) suggests that the knowledge transferred from an influential teacher model can provide rich information for the student to learn. The objective for this task is to minimize the prediction error between the teacher and the student, which can be summarized as $\mathcal{L}_{distill} = Dist(z^T, z^S)$, where z^T, z^S are the representations (e.g. softmax logit or feature) from teacher and student respectively. $Dist(\cdot)$ is the similarity metric. Although some works (Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020; Fang et al. 2020) have extended the distillation into self-supervised learning with response-based knowledge, it is still worth exploring whether this is the optimal knowledge for SSKD.

Instance Discriminative Learning. Instance discriminative based methods (Wu et al. 2018; He et al. 2020; Chen et al. 2020a) formulate an instance-wise contrastive objec-

tive to learn representation by contrasting the positive with negative. For each image x_i from the training set, the encoder $f(\cdot)$ maps x_i to z_i with $z_i = f(x_i)$. Then the encoder is optimized by an instance-wise contrastive loss function, such as NCE (Oord, Li, and Vinyals 2018). In mean teacher based methods (He et al. 2020; Chen et al. 2020b), $z_i' = f'(x_i)$ is generated as the positive from the mean teacher (a.k.a momentum encoder). Thus, we can rewrite the NCE from the distillation perspective:

$$\mathcal{L}^{NCE} = -\log \frac{\exp(z_i \cdot z_i'/\tau)}{\exp(z_i \cdot z_i'/\tau) + \sum_{n \in D(i)} \exp(z_i \cdot z_n/\tau)}$$

$$= \log(1 + \sum_{n \in D(i)} \exp(z_i \cdot z_n/\tau) \cdot \exp(-z_i \cdot z_i'/\tau)) \quad (1)$$

where $D(i)$ is the negative feature set for instance i, and τ is the temperature parameter. In the Eq. 1, $\exp(-z_i \cdot z_i'/\tau)$ seeks to maximize the similarity for z_i and z_i', which pushes the student's prediction z_i close to mean teacher's *historical* prediction z_i'. For $\sum_{n \in D(i)} \exp(z_i \cdot z_n/\tau)$, it aims to minimize the similarity for z_i and z_n, which separates the negative samples apart. From the distillation perspective, $\sum_{n \in D(i)} \exp(z_i \cdot z_n/\tau) \cdot \exp(-z_i \cdot z_i'/\tau)$ can be regarded as the *historical distillation*, which aims to make the the student mimic mean teacher's historical prediction. In this way, we can unify the objective in these mean-teacher based method as learning the response knowledge from knowledge distillation perspective.

However, the negatives in $D(i)$ are not all correct. NCE regards all the other instances as negatives, which unavoidably involves some positives belonging to the same category. Besides, we argue that only one positive from the historical version can not provide sufficient information for contrastive

objective and thus limits the potential of the student.

Method

The goal of self-supervised learning is to learn rich feature representation at the semantic level. To achieve this, we propose the **Re**lation **K**nowledge **D**istillation (ReKD) in contrastive learning, as illustrated in Fig. 2, which consists of an online *heterogeneous teacher*, a *relation knowledge* and a *relation miner*. In the following subsections, we firstly introduce the ***online heterogeneous teacher*** and its role in the whole framework in section Online Heterogeneous Teacher. Secondly, we customize a novel knowledge named ***relation knowledge*** to capture the semantic information, which indicates the semantic positive/negative relationship mined by the ***relation miner*** in section Relation Knowledge. Thirdly, we formulate a ***relation contrastive loss*** for student's optimization in section Relation Contrastive Loss.

Online Heterogeneous Teacher

Assumes that we have a *heterogeneous teacher* module f_T and a student module f_S. Our objective is to make the student module learn the representation from *heterogeneous teacher* module to escape from the *semantic collapse* phenomenon. The *heterogeneous teacher* and the student use architectures from different families to ensure the superior feature representation in distillation. We maintain two candidate sets $D_t = \{u_1^t, ..., u_L^t\}$ and $D_s = \{u_1^s, ..., u_L^s\}$ to store the feature from the teacher f_T and student f_S respectively. Different from the normal self-supervised knowledge distillation methods (Fang et al. 2020; Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020), we explore the online mechanism for the teacher. In our online *heterogeneous teacher*, the teacher evolves simultaneously in the distillation stage, which is more parallelly efficient.

Relation Knowledge

To alleviate the effect of semantic collapse brought by response knowledge that neglects the relation between instances and limits to historical positive, we try to explicitly model the semantic relation to introduce more diverse positives. To achieve this, we formulate a *relation knowledge*, which captures the semantic positive/negative relationship for each pair of *anchor* z_i from the mini-batch and the *candidate* u_j from the candidate set D_t. The relation is inferred by *relation miner* (see Fig. 3) in the *heterogeneous teacher*'s embedding space. Then, the relation is transferred to student as the guidance for contrastive objective. To achieve this, we maintain a semantic prototype bank $\mathcal{P} = \{p_1, ..., p_M\}$, each prototype in the bank represents an independent semantic category learned by the model. The prototypes are initialized by spherical k-means clusters' centroids at the beginning of the training. After obtaining the prototype bank, the relation miner is then evolved by **connection** step and **update** step alternatively.

Connection. In this step, we take the prototype in $\mathcal{P}$ as the reference to mine the relation between embeddings. Given an *anchor* embedding z_i and a *candidate* embedding u_j, we can calculate the pairwise similarity for both embeddings
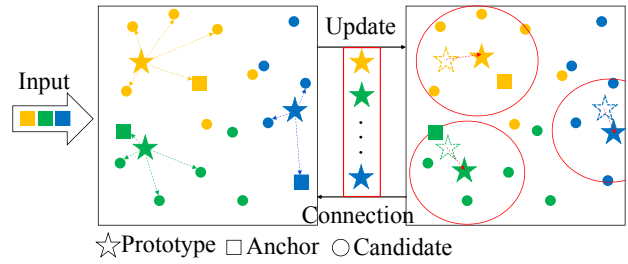

☆Prototype □Anchor ○Candidate

Figure 3: Illustration of Relation Miner. Given a list of input anchors, the *relation miner* connects candidates with anchors by prototype bank and updates prototype bank alternately during the whole training stage.

with prototype p_k. For simplicity, we denote the anchor and candidate embedding uniformly as e, then we assign the embedding e to the prototype assignment $Q(e)$ as follows:

$$Q(e) = \begin{cases} \arg\max_k \mathcal{S}(e, p_k), & \max\{\mathcal{S}(e, p_k)|p_k \in \mathcal{P}\} \geq \theta \\ -1, & \text{otherwise} \end{cases}$$
(2)

where $\mathcal{S}(e, p)$ denotes the similarity for each embedding-prototype pair. Note that if the maximum similarity is lower than θ (a threshold hyperparameter), we attribute to prototype -1, which means this embedding fails in matching any prototypes.

After assigning the *anchor* and *candidate* with the corresponding prototype, we define the relation for each *anchor-candidate* pair. We define it as *positive* pair if the anchor and candidate have the same prototype assignment Q, otherwise we match them as *negative* pair. Thus, for each pair, we obtain a relation (positive or negative). This relation is the core of *relation knowledge* for further distillation, which contains the semantic information based on the semantic prototype retrieval. Then, we can form the positive set $P(i)$, and the negative set $N(i)$ for *anchor* z_i, which is inferred by the relation miner. With this semantic relation, we can introduce diverse semantic positives for student, rather than limits to the low-level historical positive in the optimization.

Update. In this step, we momentum updates the prototypes simultaneously rather than keep frozen or re-initial frequently. In each mini-batch, we update the feature of the *anchor* z_i into assigned prototype p_k in terms of Eq. 3:

$$p_k \leftarrow (1-m)z_i + mp_k \quad m = 1 - (1-\beta) \cdot \mathcal{S}(z_i, p_k) \text{ (3)}$$

where m is the similarity-based coefficient controlling the weight of the *anchor* embedding when updating the prototype. The range of m is $[\beta, 1]$, where β is usually set to a large value to ensure the prototypes to be stable and robust to unexpected noises.

Relation Contrastive Loss

Based on the *relation knowledge* (i.e., $P(i)$ and $N(i)$) produced by the relation miner, we propose our overall relation-wise contrast objective, namely *relation contrastive loss*, which is a more generalized contrastive loss allowing for

multiple positive based on relation. Unlike the historical distillation in Eq. 1, where the student only seeks to maximize the likelihood for historical positive u_i from mean teacher with respect to all negatives u_n in $N(i)$, our relation contrastive loss enforce a *relation* distillation to maximize the likelihood for all semantic positive u_p in $P(i)$ with respect to all negatives u_n in $N(i)$, which proves to be a more reasonable optimization from semantic perspective.

$$\mathcal{L}^{RelCon} = \log(1 + \sum_{n \in N(i)} \exp(z_i \cdot u_n / \tau) \cdot \sum_{p \in P(i)} \exp(-z_i \cdot u_p / \tau)) \tag{4}$$

With this semantic relation knowledge, the self-supervised model can involve the semantic information and learn more generalized representation in contrast to instance-wise contrast methods, which can solve the semantic collapse efficiently. In the following subsection, we theoretically prove why the *heterogeneous teacher* and our *relation knowledge* work.

Theoretical Analysis on ReKD

To prove how the *relation knowledge* benefits in ReKD, we firstly delve into the contribution of positives and negatives in Eq. 4:

$$\mathcal{L}^{RelCon} = \log(1 + \sum_{n \in N(i)} \exp(s_n) \cdot \sum_{p \in P(i)} \exp(-s_p)) \tag{5}$$

where $s_p = z_i \cdot u_p / \tau$, $s_n = z_i \cdot u_n / \tau$ and $N(i), P(i)$ represent the negatives and positives set. Compared with original $\mathcal{L}^{NCE}$ in Eq. 1, we observe that the $\mathcal{L}^{RelCon}$ involves more positives rather than only the historical one. From the positive and negative perspective, we disassemble the $N(i) = TN(i) + FN(i)$ and $P(i) = TP(i) + FP(i)$, where $TN(i), TP(i)$ denote the true negatives and true positives for *anchor i*, $FN(i), FP(i)$ denote the false negatives and false positives. Then, we have the equation:

$$\mathcal{L}^{RelCon} = \log \left\{ 1 + \frac{\sum_{n \in N(i)} \exp(s_n)}{\sum_{p \in P(i)} \exp(s_p)} \right\}$$

$$= \log \left\{ 1 + \frac{\sum_{tn \in TN(i)} \exp(s_{tn}) + \sum_{fn \in FN(i)} \exp(s_{fn})}{\sum_{tp \in TP(i)} \exp(s_{tp}) + \sum_{fp \in FP(i)} \exp(s_{fp})} \right\} \tag{6}$$

Due to workable inequation of $s_{tp} > s_{tn}$ and $s_{fn} > s_{fp}$, we have the inequation:

$$\sum_{tp \in TP(i)} \exp(s_{tp}) \cdot \sum_{fn \in FN(i)} \exp(s_{fn}) > \sum_{tn \in TN(i)} \exp(s_{tn}) \cdot \sum_{fp \in FP(i)} \exp(s_{fp}) \tag{7}$$

Then we apply this inequation in the Eq. 6 and have the formula as follows:

$$\mathcal{L}^{RelCon} = \log \left\{ 1 + \frac{\sum_{tn \in TN(i)} \exp(s_{tn}) + \sum_{fn \in FN(i)} \exp(s_{fn})}{\sum_{tp \in TP(i)} \exp(s_{tp}) + \sum_{fp \in FP(i)} \exp(s_{fp})} \right\}$$

$$> \log \left\{ 1 + \frac{\sum_{tn \in TN(i)} \exp(s_{tn})}{\sum_{tp \in TP(i)} \exp(s_{tp})} \right\} \tag{8}$$

from the Eq. 8, the low bound of $\mathcal{L}^{RelCon}$ can be treated as the $N(i), P(i)$ with pure true negatives $TN(i)$ and true positives $TP(i)$, which indicates that the NCE with incorrect negatives in $N(i)$ can be harmful in optimization. Besides, we conjecture that the purity of positives $TP(i)/P(i)$ and the true positive number $|TP(i)|$ is the point of contrastive based method.

We conclude that the more accurate the *relation knowledge* is, the better performance the student can achieve. This also explains the phenomenon that the supervised-based training surpasses the unsupervised one by a large margin, especially in the lightweight model. With this in mind, our ReKD narrows the gap between supervised and unsupervised for student efficiently, which also mitigates the semantic collapse. Furthermore, the experiments in section *Performance of Relation Knowledge* (in Appendix) support this theoretical analysis.

Experiments

In this section, we demonstrate the performance of ReKD by a standard linear evaluation protocol compared with mainstream SSL and SSKD methods.

Representation Training Setting

In experiments, we validate our algorithm on multiple backbones: AlexNet, MobileNet-V3, ShuffleNet-V2, EfficientNet-b0 and ResNet-18. To enable a fair comparison, we replace the last classifier layer with an MLP layer (two linear layers and one ReLU layer). The dimension of the last linear layer sets to 128. For efficient clustering, we adopt the GPU k-means implementation in faiss (Johnson, Douze, and Jégou 2019). M sets to 1000 as default to model the dataset's semantic distribution (ablation of M refers to appendix).

Representation Evaluation with Self-supervised Method

To evaluate the representation, we freeze the features of the encoder of the self-supervised pre-trained model and train a single classifier layer (a fully connected layer followed by a *softmax*). All the hyperparameters of linear evaluation are strictly kept aligned with the implementations in (Chen et al. 2020b).

To compare with other self-supervised methods on lightweight model, we report the accuracy of AlexNet on ImageNet in Tab.1, where ReKD use ResNet-50 as teacher and AlexNet as student. It is worthy to note that ReKD achieves a significant 7.2% improvement than MoCov2

Method	Conv1	Conv2	Conv3	Conv4	Conv5
Supervised	19.3	36.3	44.2	48.3	50.5
DeepCluster	12.9	29.2	38.2	39.8	36.1
NPID	16.8	26.5	31.8	34.1	35.6
LA	14.9	30.1	35.7	39.4	40.2
ODC	**19.6**	32.8	40.4	41.4	37.3
Rot-Decouple	19.3	**33.3**	40.8	41.8	44.3
MoCov2 †	17.4	27.7	38.1	40.6	42.9
SimCLR †	6.0	30.9	37.7	42.0	40.3
BYOL †	7.4	32.0	39.7	43.9	44.7
SwAV †	11.4	29.4	34.5	40.4	44.2
SimSiam †	17.3	26.8	37.6	39.8	44.5
ReKD (ours)	16.3	33.0	**42.9**(+3.2)	**48.1**(+4.2)	**50.1**(+5.4)

Table 1: ImageNet test accuracy (%) using linear classification on different self-supervised learning methods. † denotes the result reproduced by us on AlexNet backbone.

baseline, which is the first work comparable with the supervised learning. The improvement implies that our ReKD obviously mitigates semantic collapse in lightweight model and that the relation knowledge may role as semantic label in supervised contrastive learning.

To demonstrate that the ReKD can be generalized on different light backbones with types of large models(heterogeneous teachers). We use AlexNet, MobileNet-V3, ShuffleNet-V2, EfficientNet-b0 and ResNet-18 as lightweight models, and use ResNet-50 and ResNet-101 as large model. Tab.3 shows that all lightweight models achieve a consistent and significant improvements, where EfficientNet-b0 increases almost 25% top-1 accuracy. This result validates that ReKD is an effective method, that can also be implemented with various lightweight models and large models (heterogeneous teachers) in a flexible way.

Method	Student	ImageNet Acc.	
		Top-1	Top-5
Supervised	AlexNet	50.5	–
CC	AlexNet	37.3	–
SEED ∗	AlexNet	44.7	69.0
CompRess †	AlexNet	46.8	71.3
ReKD (ours)	AlexNet	**50.1**(+3.3)	**74.4**(+3.1)
SEED	R-18	57.6	81.8
ReKD (ours)	R-18	**59.6**(+2.0)	**83.3**(+1.5)
SEED	Mob-v3	55.2	80.3
ReKD (ours)	Mob-v3	**56.7**(+1.5)	**81.2**(+0.9)
SEED	Eff-b0	61.3	82.7
ReKD (ours)	Eff-b0	**63.4**(+2.1)	**84.3**(+1.6)

Table 2: ImageNet test accuracy (%) using linear classification on different self-supervised knowledge distillation methods. ∗ indicates the result is reproduced by us. † denotes the result reproduced in the same architecture for fair comparison.

Representation Evaluation with Self-supervised Knowledge Distillation Method

To prove the effectiveness of ReKD, we conduct experiments with all the offline SSKD methods (Fang et al.

2020; Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020; Noroozi et al. 2018) on the same backbone (AlexNet) following the linear classification. Note that the original teacher used in (Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020) is MoCov2 with ResNet-50 pre-trained 800 epochs offline. For a fair comparison, we change the teacher of Mo-Cov2 with ResNet-50 of 200 epochs pre-trained model to compare with ReKD. In Tab.2, our ReKD outperforms all the SSKD methods on AlexNet.

Further Analysis

In this section, we analyze ReKD from different perspectives.

Ablation for Components

In Tab.4, we report the impact of applying *heterogeneous teacher* on the selected method (Fang et al. 2020) (Tab.4.b) and our method (Tab.4.c). The baseline (Tab.4.a) is Mo-Cov2 (Chen et al. 2020b) using a mean teacher to guide the student. We see that an extra *heterogeneous teacher* can boost the performance by a significant margin of 2.3% on Top-1 Acc. With the *relation knowledge* we propose, the performance can be further improved by 7.2%. This validates our ReKD with relation knowledge can break the semantic collapse, which also help learn a generalized representation.

Response Knowledge vs. Relation Knowledge

To prove the effectiveness of our *relation knowledge*, we compare it with a response knowledge-based method (Fang et al. 2020) under the same offline teacher setting. We see that in Tab.5, the method using the *relation knowledge* with an offline teacher can improve 2.4% points compared to the response knowledge with an offline teacher. The improvement suggests that the *relation knowledge* would be a better choice when considering knowledge in SSKD. This also implies that the student may benefit from numerous and accurate semantic positive samples connected by the *relation knowledge*, which is ignored by response knowledge (Fang et al. 2020; Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020). This result also supports the theoretical analysis in the previous section *How Relation Knowledge benefits*.

Online Teacher vs. Offline Teacher

We do an ablation study of the online/offline teacher on ReKD. As for the offline teacher case, we first train the teacher with 200 epoch and freeze all the trainable parameters during distillation. For the online teacher case, both teacher and student update simultaneously. In Tab.5, we observe that online teacher cases outperform all the offline teacher cases, which suggests the potential of online teacher in SSKD. The same conclusion is observed in Deep Mutual Learning (Zhang et al. 2018). Intuitively, under the offline paradigm, the huge performance gap between teacher and student may lead to instability and low convergence for the student. Our online teacher can alleviate this and provide a proper curriculum for the student. Besides, this online teacher mechanism can save much time compared with normal offline SSKD methods, since the offline teacher

Method	Teacher	Alex		Mob-v3		Shuff-v2		Eff-b0		R-18	
		T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Supervised	–	50.5		75.2		75.4		77.5		69.8	
Self-supervised MoCov2	–	42.9	66.4	35.3	61.0	52.0	75.8	38.6	65.3	53.3	78.4
ReKD (ours)	R-50 (67.6)	50.1 +7.2	74.4 +8.0	56.7 +21.4	81.2 +20.2	61.9 +9.9	83.8 +8.0	63.4 +24.8	84.3 +19.0	59.6 +6.3	83.3 +4.9
ReKD (ours)	R-101 (69.7)	50.8 +7.9	75.1 +8.7	59.6 +24.3	83.1 +22.1	63.6 +11.6	84.9 +9.1	65.0 +26.4	85.7 +20.4	59.7 +6.4	83.9 +5.5

Table 3: ImageNet test accuracy (%) using linear classification on multiple student architectures. T-1 and T-5 denote Top-1 and Top-5 accuracy using linear evaluation. First column denotes the different methods using for training. Second column indicates Top-1 accuracy of teacher networks in MoCov2 self-supervised learning. First row indicates the student networks, while second row shows the supervised performances of student networks. Third row denotes the self-supervised baseline with MoCov2. Note, all the methods are trained for 200 epochs.

Method	Heterogeneous Teacher	Relation Knowledge	ImageNet Acc.	
			Top-1	Top-5
a			42.9	66.4
b	√		45.2	69.1
c	√	√	**50.1**(+4.9)	**74.4**(+5.3)

Table 4: Ablation of the important components in ReKD: heterogeneous teacher and relation knowledge. The accuracy of Top-1 and Top-5 is evaluated by AlexNet for linear classification on ImageNet.

Distillation Strategy		ImageNet Acc.	
Knowledge	Teacher	Top-1	Top-5
response	offline	44.7	69.0
	online	45.2	69.1
relation	offline	47.1	71.3
	online	**50.1**	**74.4**

Table 5: Ablation of distillation strategies for knowledge type and teacher mechanism. The result is evaluated in AlexNet architecture for linear classification on ImageNet.

consumes an extra pre-trained time. Compared with SSL methods, ReKD achieves much more improvement if allowing the equivalent extra time cost increase. For example, BYOL (Grill et al. 2020) costs almost 100% extra time due to the symmetric structure while only gains $+1.8\%$ accuracy improvement against MoCov2 in Tab.1. In contrast, ReKD achieves a significant $+7.2\%$ increase on accuracy while only takes a $+118\%$ extra time cost increase.

Why Heterogeneous Teacher has better semantic representation

In our relation distillation, it is critically important to choose a well-performed teacher. We analyze the feature representation on the different capacity model, such as AlexNet and ResNet-50. We select MoCov2 as our unsupervised feature extractor and extract all the features with different backbones (AlexNet and ResNet-50) from images in ImageNet,
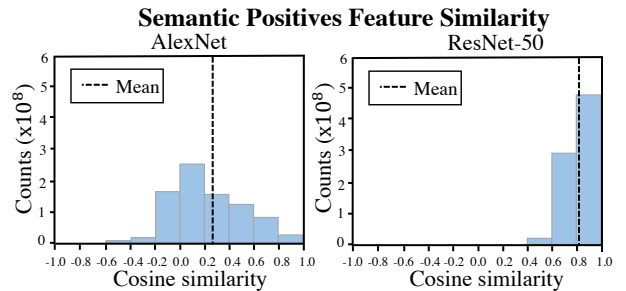

Figure 4: Distribution of similarity between semantic similar pairs from different capacity of backbones.

then we measure the feature distance (i.e. cosine similarity) for each pair of instance with same semantic ground-truth label. Fig. 4 summarizes the resulting similarity distribution, where the features from the large model (ResNet-50) exhibit large similarity, which indicates that the large model can capture more semantic information inherently. Also, the `Mean` in the figure refers to the mean similarity value of all the pairs, which indicates the overall feature extraction capacity. This result also supports the motivation (in Fig. 1(a)/(b)) that the semantic collapse is more severe in the lightweight model under unsupervised training. Therefore, Heterogeneous teacher has better semantic representation.

Conclusion

We propose a Relation Knowledge Distillation (ReKD) to alleviate the semantic collapse in most instances discriminative-based methods. Specifically, the ReKD benefits from relation knowledge, which provides the semantic relation to guide the lightweight model for the semantic contrastive objective. The theoretical analysis supports our main concern about instance-wise contrast and verifys the effectiveness of our relation-wise contrastive learning. Our extensive experiments on SSL and SSKD benchmarks demonstrate the effectiveness of ReKD. Furthermore, we hope our work can raise the community's attention to explore the efficient distillation way for the lightweight model in self-supervised learning.

References

Abbasi Koohpayegani, S.; Tejankar, A.; and Pirsiavash, H. 2020. CompRess: Self-Supervised Learning by Compressing Representations. *Advances in Neural Information Processing Systems*, 33.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, 5879–5887.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2020. Exploring Simple Siamese Representation Learning. *arXiv preprint arXiv:2011.10566*.

Chen, X.; Su, J.; and Zhang, J. 2019. A Two-Teacher Framework for Knowledge Distillation. In *International Symposium on Neural Networks*, 58–66. Springer.

Fang, Z.; Wang, J.; Wang, L.; Zhang, L.; Yang, Y.; and Liu, Z. 2020. SEED: Self-supervised Distillation For Visual Representation. *arXiv preprint arXiv:2101.04731*.

Feng, Z.; Xu, C.; and Tao, D. 2019. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10364–10374.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314–1324.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep Semantic Clustering by Partition Confidence Maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8849–8858.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Komodakis, N.; and Gidaris, S. 2018. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*.

Komodakis, N.; and Zagoruyko, S. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 69–84. Springer.

Noroozi, M.; Vinjimoor, A.; Favaro, P.; and Pirsiavash, H. 2018. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9359–9367.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Park, S.; and Kwak, N. 2020. Feature-level Ensemble Knowledge Distillation for Aggregating Knowledge from Multiple Networks. In *European conference on artificial intelligence*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Shen, C.; Xue, M.; Wang, X.; Song, J.; Sun, L.; and Song, M. 2019. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3504–3513.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, 3861–3870. PMLR.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5147–5156.

Zhang, L.; Qi, G.-J.; Wang, L.; and Luo, J. 2019. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, 2547–2555.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4320–4328.

Zhuang, C.; Zhai, A. L.; and Yamins, D. 2019. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, 6002–6012.