DisCo: Effective Knowledge Distillation For Contrastive Learning of Sentence Embeddings

Xing Wu\textsuperscript{1,2,4}, Chaochen Gao\textsuperscript{1,2}, Jue Wang\textsuperscript{3}, Liangjun Zang\textsuperscript{1}, Zhongyuan Wang\textsuperscript{4}, Songlin Hu\textsuperscript{1,2}

\textsuperscript{1}Institute of Information Engineering, Chinese Academy of Sciences
\textsuperscript{2}School of Cyber Security, University of Chinese Academy of Sciences
\textsuperscript{3}Zhongnan University of Economics and Law, \textsuperscript{4}Kuaishou Technology

\{wuxing,gaochaochen,zangliangjun,husonglin\}@iie.ac.cn
201821090281@stu.zuel.edu.cn, wangzhongyuan@kuaishou.com

Abstract

Contrastive learning has been proven suitable for learning sentence embeddings and can significantly improve the semantic textual similarity (STS) tasks. Recently, large contrastive learning models, e.g., Sentence-T5, tend to be proposed to learn more powerful sentence embeddings. Though effective, such large models are hard to serve online due to computational resources or time cost limits. To tackle that, knowledge distillation (KD) is commonly adopted, which can compress a large “teacher” model into a small “student” model but generally suffer from some performance loss. Here we propose an enhanced KD framework termed Distill-Contrast (DisCo). The proposed DisCo framework firstly utilizes KD to transfer the capability of a large sentence embedding model to a small student model on large unlabelled data, and then finetunes the student model with contrastive learning on labelled training data. For the KD process in DisCo, we further propose Contrastive Knowledge Distillation (CKD) to enhance the consistencies among teacher model training, KD, and student model finetuning, which can probably improve performance like prompt learning. Extensive experiments on 7 STS benchmarks show that student models trained with the proposed DisCo and CKD suffer from little or even no performance loss and consistently outperform the corresponding counterparts of the same parameter size. Amazingly, our 110M student model can even outperform the latest state-of-the-art (SOTA) model, i.e., Sentence-T5(11B), with only 1% parameters.

1 Introduction

Pretrained Language Models (PLMs) built on Transformer blocks (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Lan et al., 2019; Raffel et al., 2019; Clark et al., 2020) have shown strong transferability and achieved great success in many natural language processing (NLP) tasks (Wang et al., 2018), like classification (Sun et al., 2019a), sequence labelling (Tsai et al., 2019), etc. For the specific task of semantic textual similarity (STS), which measures the semantic similarity between two sentences and is widely applied in many real-world applications (like text retrieval, text deduplication, etc.), the current state-of-the-art (SOTA) methods (Gao et al., 2021; Wang et al., 2021; Yan et al., 2021) that achieves remarkable performance are also all based on PLMs. Specifically, those methods commonly train an encoder to derive an embedding vector for each sentence, using the popular scheme of contrastive learning.

Recently, sentence embedding methods tend to use larger model sizes and training data scales for better performance. For instance, the latest SOTA model, i.e., Sentence-T5 (Ni et al., 2021), is built with 11 billion parameters and trained on 2 billion question-answer pairs. Though effective, such large models are hard to be applied in real-world applications with limited computational resources or time cost for model inference.

Model compression is a feasible way to tackle the problem mentioned above and Knowledge Distillation (KD) (Romero et al., 2014; Kim and Rush, 2016; Hu et al., 2018; Sanh et al., 2019; Sun et al., 2020b; Wang et al., 2020; Jiao et al., 2019) is commonly used. KD is to transfer the knowledge learned in a large “teacher” model to a smaller “student” model, and thus expects to reduce the computational overhead and model storage while retaining the performances. Generally, KD is conducted on the same training data that the teacher model is built on (Sun et al., 2020b).

Here we conduct a probe experiment to explore KD on sentence embedding models trained with contrastive learning. Specifically, we distill the...
Figure 1: The proposed DisCo framework consists of two stages: knowledge distillation (stage 1) and further finetuning with contrastive learning (stage 2). In stage 1, Knowledge Distillation uses the MSE objective function, while Contrastive Knowledge Distillation uses the InfoNCE objective function.

well-trained SimCSE-RoBERTa-Large (Gao et al., 2021) with 330 million parameters to a smaller student model with 110 million parameters. As the teacher model is trained on the Natural Language Inference (NLI) (Bowman et al., 2015) dataset, we minimize the difference between the embedding learned by the teacher model and that of the student model for each sentence in NLI. However, as shown in Appendix A, the performance of the student model is substantially lower (-1.81%) than the teacher model in the STS task evaluation. Thus, it is challenging to use only a single KD process to transfer the capability of the large teacher model adequately to the small student model, especially on the limited NLI training data.

To alleviate the performance decrease, in this paper we propose an enhanced KD framework termed Distill-Contrast (DisCo) for contrastive learning of sentence embeddings, which is simple but effective. As shown in Figure 1, the proposed DisCo framework consists of two stages. In the first stage, we conduct KD to transfer the capability of a well-trained large sentence embedding model to a small student model, using a large set of unlabelled data instead of the same training data for adequate knowledge transfer. Then in the second stage, the student model is further finetuned with supervised contrastive learning on the teacher model’s training data. It alleviates the domain bias brought by the unlabelled data and enables the student model to better fit the STS measurement in the original training data domain.

For the KD process in the proposed DisCo framework, inspired by the performance improvement brought by the consistency of training objectives between prompt learning and autoregressive PLMs (e.g., GPT-3), we further propose a novel KD method termed Contrastive Knowledge Distillation (CKD). As illustrated in the (b) subfigure of Figure 1, the proposed CKD shares an identical knowledge transfer process as KD methods, but uses an identical loss function of contrastive learning, i.e., InfoNCE loss (Hjelm et al., 2018), for knowledge transfer, rather than MSE loss in KD methods. For KD methods, given a sentence, they would force the sentence embedding derived by the student model to be close to that learned by the teacher. Differently, for the proposed CKD, given a sentence, the InfoNCE loss not only encourages the sentence embedding derived by the student model to be close to that learned by the teacher, but also encourages the former to be far away from sentence embeddings of other sentences learned by the teacher model. Moreover, the proposed CKD can bring two consistencies in the DisCo framework. (1) The objective functions of the distillation process and the teacher model’s training process are consistent. (2) The objective functions of the first stage (i.e., the distillation process) and the second stage in the proposed DisCo framework are consistent. That may probably bring further performance improvement.

In our experiments, we use the well-trained sentence embedding model SimCSE-RoBERTa-large
(330M) as the teacher, and small Transformer-based models as the students, using different sizes of parameters (110M/52M/14M). Evaluations on 7 STS benchmark datasets show that, the student models trained through the proposed DisCo framework can well compress the large sentence embedding model with little or even no performance decrease, and significantly outperforms a single KD process. In that way, the student models are more parameter efficient than those trained directly through contrastive learning on the NLI dataset. For instance, our 110M student model can even slightly outperform the latest SOTA model, i.e., Sentence-T5(11B), with only 1% parameter. Moreover, DisCo with the newly proposed CKD method also consistently outperforms that with KD method, well demonstrating the effectiveness and reasonableness of CKD.

We summarize our contributions as follows:

1. (1) We propose a simple but effective framework termed DisCo, which utilizes KD together with an extra contrastive learning process to better compress large sentence embedding models.

2. (2) We propose a novel KD method termed Contrastive Knowledge Distillation (CKD), for the KD process in the DisCo framework. CKD brings better consistencies among teacher model training, KD and student model finetuning.

3. (3) We conduct extensive experiments on 7 STS benchmarks and well demonstrate the effectiveness of the proposed DisCo and CKD for model compression. Student models trained with DisCo and CKD can even slightly outperform the latest SOTA model with only 1% parameters.

2 Related Work

Knowledge Distillation Knowledge Distillation (Hinton et al., 2015) is a commonly used model compression technique and the knowledge distillation methods for pre-trained models have been extensively studied. (Sanh et al., 2019) proposes to distill the predicted logits related to the task from the teacher model into a student model. (Sanh et al., 2019) proposes to distill a teacher model to a student model with 6 Transformer blocks with the corresponding predicted logits. (Sun et al., 2019b) proposes to distill both the predicted logits and the [CLS] representation in the intermediate layers. (Jiao et al., 2019) proposes effective frameworks to distill both the intermediate layers and the prediction layers for the Transformer-based teacher and student models. (Aguilar et al., 2020) formulates two ways to distill the internal knowledge to improve the student model’s generalization capabilities. (Sun et al., 2020b) trains a specially designed teacher model, and transfers to a task-agnostic student model. The most related work to our work is (Sun et al., 2020a), which proposes to distill knowledge through intermediate layers of the teacher model via a contrastive objective function. However, it is designed for only classification tasks and excludes the STS tasks. Different from (Sun et al., 2020a), here we focus on STS tasks by combing knowledge distillation with contrastive learning. In addition, we do not explore the influence of the intermediate layer of the teacher model on knowledge distillation. Instead, we propose an enhanced KD framework and a novel KD method based on contrastive learning by replacing the objective function.

Contrastive Learning Contrastive learning has been explored in learning sentence embeddings and has become a promising trend. (Fang et al., 2020) pre-trains language representation models using contrastive self-supervised learning at the sentence level. (Giorgi et al., 2020) designs a self-supervised contrastive learning objective function for learning universal sentence embeddings, which does not require labelled training data. (Wu et al., 2020) proposes contrastive learning for sentence embeddings by employing multiple sentence-level augmentation strategies to learn a noise-invariant sentence embeddings. (Yan et al., 2021) fine-tunes BERT (Devlin et al., 2019) through contrastive learning to solve the collapse issue of BERT-derived sentence embeddings. One of the most related works is (Gao et al., 2021), which incorporates labelled NLI sentence pairs in contrastive learning and achieves remarkable performance in learning sentence embeddings. Different from (Gao et al., 2021) which improves sentence embeddings by modifying the model structure, we propose to achieve improvement by distilling knowledge from a larger powerful teacher model. The other most related one is (Ni et al., 2021), which introduces a multi-stage contrastive learning recipe involving fine-tuning firstly on 2 Billion question-answers pairs from commu-
In this paper, we adopt two knowledge distillation methods. One is the commonly used KD method that enforces $h_i^S$ to be close to $h_i^T$ for each sentence $x_i$, using MSE loss in general. The other is our newly proposed contrastive knowledge distillation (CKD) method, which uses InfoNCE loss to enforce $h_i^S$ to be close to $h_i^T$ and meanwhile enforce $h_i^S$ to be away from $h_j^T$ corresponding to any other sentence $x_j$.

Knowledge Distillation

MSE loss is a commonly used loss function in knowledge distillation (Jiao et al., 2019), which measures the difference between $h_i^S$ and $h_i^T$ with L2-norm for each $x_i$ in a mini-batch with $N$ sentences as follows.

$$\mathcal{L}_{KD} = \sum_{i=1}^{N} MSE(h_i^S, h_i^T)$$ (2)

In cases where the dimension of $h_i^T$ is different from that of $h_i^S$, a learnable linear projection matrix $M$ is needed to adjust the dimension of $h_i^S$ to be the same as $h_i^T$. Then the loss function above can be redefined as follows.

$$\mathcal{L}_{KD} = \sum_{i=1}^{N} MSE(Mh_i^S, h_i^T)$$ (3)

Proposed Contrastive Knowledge Distillation

For the proposed CKD, given a sentence $x_i$ in a mini-batch, $h_i^S$ and $h_i^T$ form a positive pair, and meanwhile $h_i^S$ and $h_j^T$ forms a negative pair, where $h_j^T$ is the sentence embedding of any other sentence $x_j$ within the mini-batch. Then we leverage the widely-used contrastive learning loss, i.e., InfoNCE loss, to encourage $h_i^S$ to be close to $h_i^T$ and meanwhile away from $h_j^T$:

$$\mathcal{L}_{CKD} = -\log \frac{e^{f(h_i^S, h_i^T)/\tau}}{\sum_{j=1}^{N} e^{f(h_i^S, h_j^T)/\tau}}$$ (4)

where $f(u, v)$ is the cosine similarity between $u$ and $v$, $\tau$ is a temperature hyperparameter.

The memory bank mechanism is widely adopted in contrastive learning (He et al., 2020; Chen et al., 2020). It allows reusing the encoded sentence embeddings from the immediate preceding mini-batches by maintaining a queue of a fixed size, which can enlarge the size of negative pairs for contrastive learning and thus can bring performance improvement. Here, we also incorporate the memory bank mechanism in the proposed CKD to allow the output embeddings of the student model to be compared with more output embeddings of the teacher model, without increasing the batch size. Specifically, we construct a memory bank queue for the output embeddings of the teacher model from community QA websites and then on the contrastive version of the NLI dataset. (Ni et al., 2021) achieves state-of-the-art performance in STS tasks with as high as 11B model parameters. Although our proposed DisCo is also a multi-stage training framework and uses the NLI data set in the second stage, different from (Ni et al., 2021), we focus on compressing large sentence embedding models with little or even no performance loss to enable them to be served online.

3 Distill-Contrast Framework

As illustrated in Figure 1, the proposed DisCo framework consists of two stages: knowledge distillation (KD) on large unlabelled data and student model finetuning with contrastive learning on labelled training data.

3.1 Knowledge Distillation on Unlabelled Data

In the proposed DisCo framework, the knowledge distillation stage mainly follows the well-known teacher-student structure and aims to transfer the capability of the large teacher model to the small student model. Note that here it is conducted on large unlabelled data, rather than the original labelled training data, to make the knowledge transfer more adequate. The teacher model is a parameter-fixed encoder with $T$ layers of transformer blocks trained with contrastive learning on labelled training data, e.g., NLI. The student model is a to-be-learned encoder with $S$ ($< T$) layers transformer blocks, whose parameters can be initialized with pre-trained models like BERT. For KD, the student model is trained to imitate the behavior of the teacher model. Specifically, given a mini-batch with $N$ sentences $X = \{x_1, x_2, ..., x_N\}$, for each sentence $x_i \in X$, the teacher model and the student model would encode it into $h_i^T$ and $h_i^S$, respectively, as follows.

$$h_i^S = \text{Student}(x_i)$$

$$h_i^T = \text{Teacher}(x_i)$$

(1)

where $*$ means the parameters of the teacher model is fixed.

In this paper, we adopt two knowledge distillation methods. One is the commonly used KD method that enforces $h_i^S$ to be close to $h_i^T$ for each sentence $x_i$, using MSE loss in general. The other is our newly proposed contrastive knowledge distillation (CKD) method, which uses InfoNCE loss...
consequent mini-batches. And the embeddings in the memory bank queue will be progressively replaced. Namely, when the sentence embeddings of the teacher model for the current mini-batch are enqueued, the “oldest” ones in the queue are removed if the queue is full. With the memory bank queue, the InfoNCE loss is further modified as follows. 

\[
\mathcal{L}_{\text{CKD}} = - \log \frac{e^{f(h^q_i,h^q_j)/\tau}}{\sum_{j=1}^{N} \left(e^{f(h^q_i,h^q_j)/\tau}\right) + \sum_{q=1}^{Q} \left(e^{f(h^q_i,h^q_j)/\tau}\right)}
\]

where \(h^q_i\) denotes a sentence embedding of the teacher model in the memory bank queue with a size of \(Q\). Similarly, in cases where the dimension of \(h^T_q\) is different from that of \(h^S_i\), Equation 5 is redefined as follows.

\[
\mathcal{L}_{\text{CKD}} = - \log \frac{e^{f(Mh^S_i,h^T_j)/\tau}}{\sum_{j=1}^{N} \left(e^{f(Mh^S_i,h^T_j)/\tau}\right) + \sum_{q=1}^{Q} \left(e^{f(Mh^S_i,h^T_q)/\tau}\right)}
\] (6)

### 3.2 Student Model Finetuning with Contrastive Learning on Labelled Data

To alleviate the potential domain bias brought by the large unlabelled data, we further conduct student model finetuning with contrastive learning on the labelled training data. The finetuning process enables the student model to better fit the STS measurement in the original domain.

Suppose that the original training data consists of tuples \((x_i, x_i^+, x_i^-)\), where \(x_i\) is a sentence, \(x_i^+\) is a similar sentence to \(x_i\), and \(x_i^-\) is a dissimilar one. We conduct contrastive learning to finetune the student model on the training data, using the InfoNCE loss as follows,

\[
\mathcal{L}_{\text{CL}} = - \log \frac{e^{f(h^S_i,h^{S+}/\tau)}}{\sum_{j=1}^{N} \left(e^{f(h^S_i,h^{S+}/\tau)} + e^{f(h^S_i,h^{S-}/\tau)}\right)}
\] (7)

where \(N\) is the size of a mini-batch of sentences, \(h^S_i\) and \(h^{S+}\) denote the sentence embeddings of \(x_i\) and \(x_i^+\) output by the student model, respectively.

After being finetuned with contrastive learning, the small student model can then be applied to real-world applications with generally much lower computational costs and little or even no performance loss, as demonstrated by experiments below.

### 4 Experiment

#### 4.1 Experiment Setup

**Datasets** We construct an unlabelled dataset with 5M high-quality English sentences from open-source news \(^1\), termed News-5m, for the KD stage of the proposed DisCo framework. The preprocessed data can be downloaded from the link\(^2\). And for the student model finetuning stage, we directly leverage the labelled NLI dataset, which is also the dataset for building the large teacher model. Specifically, the NLI dataset consists of 275K sentence pairs, each being either an entailment hypothesis or a contradiction hypothesis for a premise (i.e., sentence). Following \(\text{Gao et al., 2021}\), we use the entailment pairs as positives and contradiction pairs as negatives to build the needed tuples for Eq. 7.

**Baselines** We compare with the latest SOTA models, i.e., 110M/330M/3B/11B Sentence-T5 \(\text{Ni et al., 2021}\), and 330M/110M/52M/14M SimCSE \(\text{Gao et al., 2021}\). Note that for each size of parameters, Sentence-T5 explores a variety of experimental settings, and here we choose the best results for comparison. The results of baseline models are reported from the corresponding papers, except for 52M and 14M SimCSE, which have no reported results or published models, and thus we train and evaluate both by ourselves.

**Evaluation** We evaluate all methods on 7 widely used STS benchmarks, i.e., STS 2012–2016 \(\text{Agirre et al., 2012, 2013, 2014, 2015, 2016}\) and STS-B \(\text{Cer et al., 2017}\) to measure the semantic similarity of any two sentences with the cosine similarity between the corresponding sentence embeddings. After deriving the semantic similarities of all sentence pairs in the test set, we follow \(\text{Gao et al., 2021}\) to use Spearman correlation\(^3\) to measure the correlation between the ranks of predicted similarities and that of the ground-truth similarities. Specially, we utilize the public SentEval toolkit\(^4\) to evaluate the models on the dev set of STS-B to

---

\(^1\)http://data.statmt.org/news-commentary/v16/  
\(^2\)https://github.com/chaochen99/DisCo  
\(^3\)https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient  
\(^4\)https://github.com/facebookresearch/SentEval
search for better settings of the hyper-parameters. Then the best-performing checkpoint is evaluated on the STS test sets.

### 4.2 Training Details

| Model          | #layers | embed size | #params  |
|----------------|---------|------------|----------|
| Teacher        | 24      | 1024       | 330M     |
| *-BERT-base    | 12      | 768        | 110M     |
| *-Tiny-L6      | 6       | 768        | 52M      |
| *-Tiny-L4      | 4       | 312        | 14M      |

Table 1: Model sizes of the teacher model and different student models. * represents KD, CKD, DisCo-KD or DisCo-CKD.

**Model Settings** For the teacher model, we choose the 330M pre-trained checkpoints of SimCSE-RoBERTa-large, which composes of 24 layers of transformer blocks. An MLP layer is added on top of the [CLS] representation to get the sentence embedding, and the dimension of sentence embeddings is 1024. During training, the parameters of the teacher model are fixed and will not be updated. For the student model, we have three different settings of the parameter sizes, i.e., 110M/52M/14M. They are in similar network structures to the teacher model, except that the number of layers and the dimensions of sentence embeddings are smaller. The 110M models are initialized from the pre-trained BERT-base, and the 52M and 42M models are initialized from the pre-trained TinyBERT. We list the model information of the teacher model and the different student models in Table 1. Note that the dimension of sentence embeddings output by each student model is different from that of the teacher model, and thus for the knowledge distillation state in the proposed DisCo framework, we need to add a layer of linear projection upon the embeddings output by the student model to map them to 1024-dimensional ones, as denoted by Eq. 3 and Eq. 6.

**Optimization Settings** In the knowledge distillation stage on large unlabelled data, we train our student models for 20 epochs, using the Adam optimizer with a batch size of 512. Learning rate is set as $2e^{-4}$ for 110M model, and $3e^{-4}$ for 52M and 14M models. Particularly for the proposed CKD, we follow (Li et al., 2021) to use a memory bank queue with the size being 65536. Following (Gao et al., 2021), we evaluate the performance of each student model every 125 training steps on the dev set of STS-B. The training process uses the early-stop strategy with the patience being 3. Namely, the KD or CKD process will stop if the performance of a student model on the dev set is not updated for 3 consecutive epochs.

In the student model finetuning stage, we load each student model from the best performing checkpoint in the knowledge distillation stage and train it for 5 epochs using the Adam optimizer, with a batch size of 128. Learning rate is set as $1e^{-5}$ for 52M model, and $5e^{-5}$ for 110M and 14M models. Each student model is still evaluated every 125 training steps on the dev set of STS-B, and the best checkpoint is used for the final evaluation on test sets.

### 4.3 Experiment Results

In Table 2, we report the performance of the student models under three settings of parameter sizes, i.e., DisCo-*.BERT-base (110M), DisCo-*.Tiny-L6 (52M), and DisCo-*.Tiny-L4 (14M), with * being KD or CKD. We also report the performance of student models derived via just a single KD process for comparison, i.e., KD/CKD-BERT-base (110M), KD/CKD-Tiny-L6 (52M), KD/CKD-Tiny-L4 (14M), which are also trained on the same unlabelled data.

It can be seen that, in all settings of parameter sizes, student models trained through the proposed DisCo framework outperform their corresponding counterparts trained through just a single KD process. Moreover, using the proposed CKD instead of KD brings consistent further improvements. That also verifies the consistencies brought by CKD among teacher model training, KD, and student model finetuning are beneficial to improving the performance of the student model. In that way, the 52M DisCo-CKD-Tiny-L6 can even achieve comparable results to the 110M DisCo-KD-BERT-base.

**110M Student Models:** Both DisCo-KD-BERT-base and DisCo-CKD-BERT-base achieve better performance than the teacher model, i.e., SimCSE-RoBERTa-Large. And amazingly, DisCo-CKD-BERT-base can slightly outperform the latest SOTA 11B Sentence-T5, with only 1% parameters. Mean-
| #Params | Model                  | STS12 | STS13 | STS14 | SICK15 | STS16 | STS-B    | SICK-R | Avg.  |
|---------|------------------------|-------|-------|-------|--------|-------|----------|--------|-------|
| 11B     | Sentence-T5♣           | 80.11 | 88.78 | 84.33 | 88.36  | 85.55 | 86.82    | 80.60  | 84.94 |
| 3B      | Sentence-T5♣           | 79.02 | 88.80 | 84.33 | 85.31  | 86.25 | 79.51    | 83.76  | 84.59 |
| 330M    | Sentence-T5♣           | 79.10 | 87.32 | 83.17 | 84.36  | 86.73 | 79.84    | 84.11  | 83.44 |
|         | SimCSE-RoBERTa-Large♣  | 77.46 | 87.27 | 82.36 | 83.93  | 86.70 | 81.95    | 83.76  | 81.57 |
| 110M    | Sentence-T5♣           | 78.05 | 85.84 | 82.19 | 84.03  | 86.04 | 79.75    | 83.34  | 82.52 |
|         | SimCSE-RoBERTa-base♣   | 76.53 | 85.21 | 80.95 | 85.03  | 85.83 | 80.50    | 82.52  | 81.29 |
|         | SimCSE-BERT-base♣      | 75.30 | 84.67 | 80.19 | 84.50  | 84.25 | 80.39    | 81.57  | 81.09 |
|         | KD-BERT-base           | 75.6  | 86.75 | 81.31 | 83.63  | 86.01 | 81.56    | 83.05  |       |
|         | CKD-BERT-base          | 76.48 | 86.94 | 82.42 | 83.65  | 86.27 | 81.03    | 83.45  |       |
|         | DisCo-KD-BERT-base     | 78.57 | 88.32 | 83.52 | 84.56  | 87.6  | 81.55    | 84.57  |       |
|         | DisCo-CKD-BERT-base♣   | 79.51 | 88.85 | 84.10 | 85.06  | 87.97 | 81.34    | 85.04  |       |
| 52M     | SimCSE-Tiny-L6♠        | 75.66 | 83.49 | 79.82 | 85.14  | 80.41 | 83.08    | 81.09  |       |
|         | KD-Tiny-L6             | 75.19 | 86.26 | 80.64 | 82.52  | 84.81 | 80.04    | 82.29  |       |
|         | CKD-Tiny-L6            | 75.83 | 86.88 | 82.12 | 83.51  | 85.97 | 80.26    | 83.17  |       |
|         | DisCo-KD-Tiny-L6       | 77.97 | 87.32 | 82.89 | 83.85  | 86.46 | 80.89    | 83.85  |       |
|         | DisCo-CKD-Tiny-L6      | 78.20 | 88.21 | 83.75 | 84.61  | 87.53 | 81.29    | 84.58  |       |
|         | SimCSE-Tiny-L4♠        | 74.90 | 78.07 | 73.56 | 81.51  | 77.24 | 77.78    | 77.30  | 77.19 |
| 14M     | KD-Tiny-L4             | 74.41 | 83.84 | 78.89 | 84.71  | 80.45 | 82.78    | 77.93  | 80.43 |
|         | CKD-Tiny-L4            | 74.27 | 84.71 | 80.19 | 81.94  | 83.63 | 79.94    | 81.16  |       |
|         | DisCo-KD-Tiny-L4       | 76.58 | 85.40 | 81.76 | 86.72  | 82.71 | 84.87    | 79.89  | 82.56 |
|         | DisCo-CKD-Tiny-L4      | 77.24 | 85.50 | 81.94 | 87.10  | 85.16 | 80.00    | 82.84  |       |

Table 2: Sentence embedding performance on 7 semantic textual similarity (STS) test sets, in terms of Spearman’s correlation. ♣: results from (Reimers and Gurevych, 2019; Gao et al., 2021). ♠: Small SimCSE models trained by ourselves with the code and data from (Gao et al., 2021).

while, DisCo-KD-BERT-base can achieve comparable performance to the 3B Sentence-T5.

**52M Student Models:** Both DisCo-KD-Tiny-L6 and DisCo-CKD-Tiny-L6 can still outperform the teacher model, i.e., SimCSE-RoBERTa-Large. And DisCo-CKD-Tiny-L6 achieves comparable performance to the 3B Sentence-T5, with only 1/60 parameters.

**14M Student Models:** Both DisCo-KD-Tiny-L4 and DisCo-CKD-Tiny-L4 suffer from some performance loss, but the loss is much less than that brought by their corresponding counterpart with just a single KD process, i.e., KD-Tiny-L4 and CKD-Tiny-L4. That also demonstrates the superiority of our proposed DisCo framework. Moreover, both of them can still outperform the SimCSE-BERT-base model, with only 1/8 parameters.

The analyses above show that the small student models derived from the proposed DisCo framework can even outperform the large teacher model. We argue that it can be attributed to both following potential reasons. Firstly, the KD stage on large unlabelled data can not only adequately transfer the model capability of the teacher model to the student model, but also make the student model see more data, which is somehow like a semi-supervised setting. Secondly, after the KD stage, the student model is initialized as a local optimum near to the one corresponding to the teacher model, and the further fine-tuning process with contrastive learning can probably enable it to reach another better local optimum and thus gain performance improvements.

**5 Further Analyses**

In this section, we conduct some analyses on the DisCo framework. Following (Gao et al., 2021), all results are evaluated on the development set of STS-B unless otherwise specified.

**5.1 Uniformity and Alignment**

We investigate the alignment and uniformity of models of different parameter sizes and knowledge distillation methods in the DisCo framework. Following (Wang and Isola, 2020), we compute the alignment loss and uniformity loss to measure the quality of the learned sentence embeddings, which
are defined as follows.

\[ L_{\text{align}} = -E_{v,v^+ \sim p_{\text{pos}}} \| f(v) - f(v^+) \| \]
\[ L_{\text{uniform}} = \log E_{v,w_{1,i,d} \sim p_{\text{data}}} e^{-2\|f(v) - f(w)\|} \]  

where \( p_{\text{pos}} \) denotes all positive pairs of similar sentences, and \( p_{\text{data}} \) is the data distribution. \( L_{\text{align}} \) is the expected distance between the embeddings of the two sentences in a positive pair, and \( L_{\text{uniform}} \) denotes the uniformity of the embedding distribution. For both \( L_{\text{align}} \) and \( L_{\text{uniform}} \), a lower value indicates better performance.

As shown in Figure 2, DisCo-CKD models can achieve better alignment, while DisCo-KD models can have better uniformity. Meanwhile, among student models with different parameter sizes, the 6-layer ones can achieve better uniformity, the 4-layer ones can achieve better alignment, and the 12-layer ones can reach a balance in good uniformity and alignment.

5.2 Effects of the Scales of Unlabelled Data for KD

We investigate the effects of different scales of the unlabelled data in the KD stage on DisCo-CKD-BERT-base model. We increase the dataset scale from 1M to 20M gradually and plot the best results on STS-B development set in Figure 3. As the scale of unlabelled data for the CKD stage increases, the performance of DisCo-CKD-BERT-base will firstly increase then tend to converge. Specifically, when the amount of unlabelled data increases from 1M to 5M, Spearman’s correlation on the STS-B development set tends to converge and varies between 88.9 and 89.0. Compared with the 2 billion question-answer pairs used in Sentence-T5 (11B), the proposed DisCo framework is kind of data-efficient during the KD stage and thus very large unlabelled data is not needed.

5.3 Initialize Student Model with SimCSE

We further explore whether initializing the student models with the parameters of a well-trained SimCSE models can bring further improvement, as SimCSE can yield significantly superior performance than BERT for STS.

In table 4, we report the evaluation results of student models initialized from either the well-trained SimCSE or the general pre-trained models (i.e., BERT-base, and TinyBERT) in different settings of parameter sizes, i.e., 110M/52M/14M. It can be seen that initialization with SimCSE achieves comparable performance to that of initialization with general pre-trained models, which reflects the robustness of the proposed DisCo framework in some sense.

5.4 Performance on Transfer Tasks

Following (Gao et al., 2021), we further evaluate the performance of the proposed DisCo framework on transfer tasks, to see the transferability of the sentence embeddings output by the student models learned through DisCo. The transfer tasks include: MR (movie review) (Pang and Lee, 2005), CR (product review) (Hu and Liu, 2004), SUBJ (subjectivity status) (Pang and Lee, 2004), MPQA (opinion-polarity) (Wiebe et al., 2005), SST-2 (binary sentiment analysis) (Socher
| #Params | Model | MR | CR | SUBJ | MPQA | SST | TREC | MRPC | Avg. |
|---------|-------|----|----|------|------|-----|------|------|------|
| 330M    | SimCSE-RoBERTa-Large | 88.12 | 92.37 | 95.11 | 90.49 | 92.75 | 91.80 | 76.64 | 89.61 |
| 110M    | Sentence-T5 ♣ | 86.56 | 91.31 | 96.01 | 90.57 | 90.77 | 94.60 | 72.93 | 88.96 |
|         | SimCSE-RoBERTa-base ♣ | 84.92 | 92.00 | 94.11 | 89.82 | 91.27 | 88.80 | 75.65 | 88.08 |
|         | SimCSE-BERT-base ♣ | 82.68 | 88.88 | 94.52 | 89.82 | 88.41 | 87.60 | 76.12 | 86.86 |
|         | DisCo-KD-BERT-base | 86.51 | 91.55 | 95.15 | 91.02 | 91.10 | 93.20 | 76.64 | 89.31 |
|         | DisCo-CCKD-BERT-base | 86.50 | 91.34 | 95.16 | 91.21 | 91.76 | 90.60 | 76.93 | 89.07 |
| 52M     | SimCSE-Tiny-L6 | 81.96 | 88.93 | 94.30 | 89.84 | 86.66 | 89.20 | 75.25 | 86.45 |
|         | DisCo-KD-Tiny-L6 | 84.48 | 90.44 | 94.75 | 91.19 | 90.73 | 91.60 | 76.52 | 88.39 |
|         | DisCo-CCKD-Tiny-L6 | 85.61 | 90.97 | 94.78 | 91.43 | 90.99 | 90.60 | 76.70 | 88.73 |
| 14M     | SimCSE-Tiny-L4 | 81.46 | 89.88 | 92.01 | 90.48 | 87.64 | 86.20 | 74.67 | 86.05 |
|         | DisCo-KD-Tiny-L4 | 81.60 | 90.25 | 92.52 | 90.36 | 87.20 | 86.50 | 75.07 | 86.14 |

Table 3: Results on transfer tasks of different sentence embedding models, in terms of accuracy. ♣: results from (Reimers and Gurevych, 2019; Gao et al., 2021; Ni et al., 2021).

| #Params | Initialization | dev_stsb |
|---------|----------------|----------|
| 110M    | BERT-base      | 88.96    |
|         | SimCSE         | 89.03    |
| 52M     | TinyBERT       | 88.74    |
|         | SimCSE         | 88.70    |
| 14M     | TinyBERT       | 87.10    |
|         | SimCSE         | 87.21    |

Table 4: Evaluation results of student models initialized from the well-trained SimCSE or general pre-trained models (i.e., BERT-base and TinyBERT) in different settings of parameter sizes.

...to transfer the capability of a large sentence embedding model to a small student model on large unlabelled data, and a student model finetuning stage to make it better fit the STS measurement in the original data domain. For the KD stage of DisCo, we further propose a novel contrastive learning based KD method termed Contrastive Knowledge Distillation (CKD), which can bring performance improvement with better consistencies among teacher model training, KD, and student model finetuning. Experimental results on 7 STS benchmarks show that, the proposed DisCo and CKD are effective, and the learned student model can even slightly outperform the latest SOTA model Sentence-T5 (11B) with only 1% parameters.

6 Conclusion

In this paper, we propose an enhanced KD framework termed Distill-Contrast (DisCo), to compress large sentence embedding models with little or even no performance loss. DisCo consists of a KD stage to transfer the capability of a large sentence embedding model to a small student model on large unlabelled data, and a student model finetuning stage to make it better fit the STS measurement in the original data domain. Following (Gao et al., 2021), we train a logistic regression classifier on the frozen sentence embeddings generated by different methods, and use the default configuration of SentEval for evaluation. The evaluation results on the transfer tasks are reported in Table 3. It can be seen that, though the small student models underperform the large teacher model (i.e., SimCSE-RoBERTa-Large) on transfer tasks, they still consistently outperform the corresponding counterparts of the same parameter size. That also validates the effectiveness of the proposed DisCo framework.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In...
Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 385–393.

Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. 2020. Knowledge distillation from internal representations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7350–7357.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training deep text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. arXiv preprint arXiv:2005.12766.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821.

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. arXiv preprint arXiv:2006.03659.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670.

Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. 2018. Attention-guided answer distillation for machine reading comprehension. arXiv preprint arXiv:1808.07644.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. arXiv preprint arXiv:1606.07947.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In NeurIPS.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Jianmo Ni, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, Yinfei Yang, et al. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. arXiv preprint arXiv:2108.08877.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. arXiv preprint cs/0409058.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint cs/0506075.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Khour, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019a. How to fine-tune bert for text classification? In China National Conference on Chinese Computational Linguistics, pages 194–206. Springer.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019b. Patient knowledge distillation for bert model compression. arXiv preprint arXiv:1908.09355.

Siqi Sun, Zhe Gan, Yu Cheng, Yuwei Fang, Shuohang Wang, and Jingjing Liu. 2020a. Contrastive distillation on intermediate representations for language model compression. arXiv preprint arXiv:2009.14167.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020b. Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Ariavzahagam, Xin Li, and Amelia Archer. 2019. Small and practical bert models for sequence labeling. arXiv preprint arXiv:1909.00100.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 200–207.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Liang Wang, Wei Zhao, and Jingming Liu. 2021. Aligning cross-lingual sentence representations with dual momentum contrast. arXiv preprint arXiv:2109.00253.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In International Conference on Machine Learning, pages 9929–9939. PMLR.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. arXiv preprint arXiv:2002.10957.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39(2):165–210.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. arXiv preprint arXiv:2012.15466.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Yin, and Steven Hoi. 2021. Align before fuse: a contrastive framework for self-supervised sentence representation transfer. arXiv preprint arXiv:2105.11741.

Zhilin Yang, Zhang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.
### Table 5: The KD probe experiment results on the NLI Dataset

| #Params | Model                        | STS12 | STS13 | STS14 | SICK15 | STS16 | STS-B | SICK-R | Avg. |
|---------|------------------------------|-------|-------|-------|--------|-------|-------|--------|------|
| 330M    | SimCSE-RoBERTa-Large♣        | 77.46 | 87.27 | 82.36 | 86.66  | 83.93 | 86.70 | 81.95  | 83.76|
| 110M    | KD-BERT-base                 | 74.92 | 85.50 | 80.04 | 85.10  | 82.5  | 84.92 | 80.70  | 81.95(-1.81%)|
| 52M     | KD-Tiny-L6                   | 74.86 | 84.97 | 79.81 | 85.56  | 81.79 | 84.56 | 80.70  | 81.75(-2.01%)|
| 14M     | KD-Tiny-L4                   | 72.07 | 81.26 | 76.81 | 83.95  | 79.37 | 81.64 | 79.30  | 79.20(-4.56%)|

♣: results from (Gao et al., 2021).

## A The KD Probe Experiment on the NLI Dataset

We conduct a probe experiment by applying the KD process on the NLI data set under different parameter scales. Each student model is still evaluated every 125 training steps on the dev set of STS-B, and the best checkpoint is used for the final evaluation on test sets. The performances of student models on 7 semantic textual similarity (STS) test sets are shown in Table 5. The performance of the student models is substantially lower than the teacher model in the STS task evaluation.