PCL: Peer-Contrastive Learning with Diverse Augmentations for Unsupervised Sentence Embeddings

Qiyu Wu[1]*, **Chongyang Tao**[2], **Tao Shen**[2], **Can Xu**[2], **Xiubo Geng**[2], **Daxin Jiang**[2]†
[1]The University of Tokyo, Tokyo, Japan
[2]Microsoft Corporation
[1]qiyuw@g.ecc.u-tokyo.ac.jp
[2]{chotao,shentao,caxu,xigeng,djiang}@microsoft.com

Abstract

Learning sentence embeddings in an unsupervised manner is fundamental in natural language processing. Recent common practice is to couple pre-trained language models with unsupervised contrastive learning, whose success relies on augmenting a sentence with a semantically-close positive instance to construct contrastive pairs. Nonetheless, existing approaches usually depend on a mono-augmenting strategy, which causes learning shortcuts towards the augmenting biases and thus corrupts the quality of sentence embeddings. A straightforward solution is resorting to more diverse positives from a multi-augmenting strategy, while an open question remains about how to unsupervisedly learn from the diverse positives but with uneven augmenting qualities in the text field. As one answer, we propose a novel Peer-Contrastive Learning (PCL) with diverse augmentations. PCL constructs diverse contrastive positives and negatives at the group level for unsupervised sentence embeddings. PCL performs peer-positive contrast as well as peer-network cooperation, which offers an inherent anti-bias ability and an effective way to learn from diverse augmentations. Experiments on STS benchmarks verify the effectiveness of PCL against its competitors in unsupervised sentence embeddings.[1]

1 Introduction

Sentence embedding learning, which aims at deriving semantically meaningful fixed-sized vectors for sentences, is a natural language processing (NLP) technique of great significance, especially for time-sensitive downstream tasks (Reimers and Gurevych, 2019). Recently, contrastive learning (CL) is proven effective to learn representation (Wu et al., 2018; Tian et al., 2020; He et al., 2020) and substantially improve its performance (Yan et al., 2021; Gao et al., 2021) when coupling with pre-trained language models (PLMs). The main idea of contrastive learning for sentence embedding is pulling semantic neighbors together and pushing semantic non-neighbors apart (Hadsell et al., 2006), which naturally requires effective contrastive pairs. As effective contrastive pairs are usually scarce and require much human effort to collect, how to learn sentence embeddings in an fully unsupervised manner has become a challenging yet attractive research area (Wang et al., 2021; Giorgi et al., 2021).

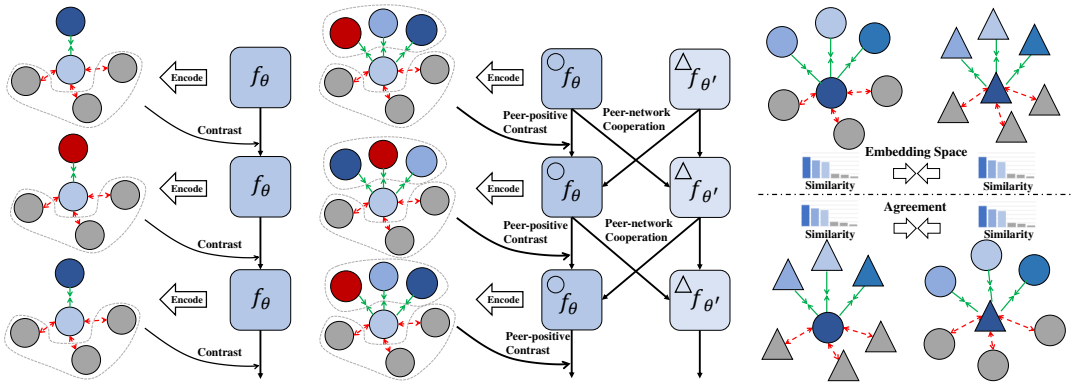
Augmenting	Order	N-gram	Bag-of-words
Shuffled Sentence	×	×	✓
Inversed Sentence	×	✓	✓
Word Repetition	✓	×	✓
Word Deletion	✓	×	×

Table 1: Text augmentation strategies change semantics in the sentence but still has shortcuts to learn. Employing limited number of strategies causes learning shortcuts towards the augmenting bias.

The key to unsupervised contrastive learning for sentence embedding is to augment a given anchor sentence with an effective positive instance to construct the pairs. Hence, many efforts have been made to design augmentation methods by adding noises or using heuristics, which mainly fall into two categories in terms of augmentation format – *discrete* and *continuous*. The former operates directly on words or n-grams in the sentence, e.g., synonym substitution (Su et al., 2021b), shuffling and word deletion (Yan et al., 2021). The latter operates on latent embeddings derived by neural encoder(s), e.g., SimCSE with twice dropouts (Gao et al., 2021). However, these existing approaches usually depend on a mono-augmenting format (i.e., either discrete or continuous) with a limited number of augmenting strategies, which suffer from learning shortcuts (Ilyas et al., 2019; Du et al., 2021)

*Work done during the internship at Microsoft.
† Corresponding author.
[1]Our implementation is available at https://github.com/qiyuw/PeerCL.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 12052 - 12066
December 7-11, 2022



(a) Conventional CL (b) PCL with diverse augmentations (c) Cooperative learning objective

Figure 1: The main idea of PCL. Circles and triangles denote contrastive instances encoded by f_θ and $f_{\theta'}$, respectively. The red ones denote ineffective positives, blue ones denote effective positives, and gray ones denote negatives. **(a)** Conventional CL on ineffective positive instances causes shortcut learning towards mono-augmenting biases. **(b)** PCL improves the probability of 'at-least-one' effective positives and performs contrasts among peer positives and cooperation between peer networks. PCL maintains two networks, and each network learns from its peer network to achieve a common agreement. **(c)** Cooperative learning objective considers peer-positive contrasts to achieve PCL with diverse augmentations. The top panel illustrates the consistency of embedding space, while the bottom panel illustrates the agreement between the peer networks.

towards the augmenting biases and thus corrupt the quality of learned embeddings. For example, learning shortcuts caused by discrete augmenting biases are shown in Table 1, and SimCSE based solely on dropout in continuous format is biased towards the sentence length (Wu et al., 2021b).

To prevent the learning shortcuts caused by the potential biases from the mono-augmenting strategy, one straightforward solution coming into our mind is that we can consider more diverse augmentations for a given sentence in both continuous and discrete formats. Besides learning from diverse instances for inherent anti-bias ability, it can also bring a great opportunity for more effective learning. In particular, controlling the qualities of the noisy and heuristic augmentations for different sentences is almost impossible[2]. As illustrated in Figure 1(a), the resulting contrastive instances may become ineffective and even poisonous for conventional CL. Nonetheless, diverse instances from various augmentation strategies can notably improve the possibility of *at-least-one* effective positives in the contrastive instances, so how to leverage the rich relations among the diverse augmentations for more effective CL is worth exploiting.

To this end, we propose a brand-new Peer-Contrastive Learning (PCL) with diverse augmentations, and an illustration of its overall framework is shown in Figure 1(b). Firstly, PCL not only performs the vanilla positive-negative contrasts but also takes the opportunity to learn the rich structured relations among the diverse positives (i.e., *peer-positive*) to highlight the more possibly effective ones. Then, to learn the structured relations in a fully unsupervised manner, we propose a cooperative learning framework consisting of two peer embedding networks (i.e., *peer-network*). The two networks learn from each other to prevent error reinforcement in sole-network and achieve a common agreement from different views (as shown in Figure 1(c)). Consequently, the sentence embedding network is equipped with (i) anti-bias abilities by CL on the diverse augmentations and (ii) improved effectiveness by the unsupervised PCL, leading to a high quality of sentence embeddings.

We conduct experiments on 7 standard semantic textual similarity (STS) tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017; Marelli et al., 2014) to evaluate PCL. Results demonstrate that PCL significantly outperforms state of the art on 7 STS tasks. Typically, PCL achieves a 2.85% improvement over the previous best approach in the averaged Spearman's correlation of 7 STS tasks in the $\text{BERT}_{\text{base}}$ setting. PCL also outperforms previous approaches across different PLMs initialization and model sizes. Moreover, ablation study and analysis show that the two proposed components, i.e., peer-positive contrasts and peer-network cooperation, are both capable of improving unsupervised sentence embedding learning.

[2]It may cause poisonous positive. For example, given a sentence *"A dog is chasing a cat."*, a possible shuffled sentence is *"A cat is chasing a dog."*, which is semantically different.

2 Peer-Contrastive Learning (PCL)

This section begins with a formal definition of unsupervised sentence embeddings, followed by detailed formulations of our PCL with diverse augmentations (§2.1 and §2.2). Lastly, we will elaborate on our training and inference procedure for unsupervised sentence embeddings (§2.3).

Unsupervised Sentence Embedding. Given a sentence $x_i \sim X$, the target of this task is to learn a neural network f_θ (parameterized by θ) without any human-labeled data. Then, the network can be applied to x and derive a dense real-valued vector representation, i.e., $\boldsymbol{h}_i = f_\theta(x_i) \in \mathbb{R}^d$. Consequently, $\boldsymbol{h}_i$ can be used to represent the semantics of x_i and fulfill downstream sentence-related tasks, e.g., semantic textual similarity. Thereby, this task depends on the designs of unsupervised (a.k.a self-supervised) objectives based on X to learn f_θ effectively.

2.1 Contrastive Representation Learning

The recent common practice of representation learning in an unsupervised manner is contrastive learning (Zhang et al., 2020; Yan et al., 2021; Gao et al., 2021), which aims to learn effective representations to pull similar instances together and push apart the other instances. Thereby, compared to supervised contrastive learning that has already offered contrastive pairs, how to augment the given anchor (e.g., an image and sentence) with effective positive and negative instances to construct the contrastive pairs is critical in the unsupervised scenario. More recently, a simple contrastive learning framework (SimCLR) is proposed in visual representation learning (Chen et al., 2020), which constructs positive instances by different views (e.g., chop) of an image then learn to minimize the following InfoNCE loss.

$$L_\theta^{(\mathrm{c})}(X, \delta; \theta) = \tag{1}$$
$$-\mathbb{E}_{x_i \sim X}\left[\log \frac{e^{s[f_\theta(x_i), f_\theta(\delta(x_i))]/\tau}}{\sum_{x_j \sim X \wedge j \neq i \vee \delta(x_i)} e^{s[f_\theta(x_i), f_\theta(x_j)]/\tau}}\right],$$

where $\delta(\cdot)$ denotes using a different view of the image x_i as the positive instance during visual contrastive learning, x_j denotes negative instances against x_i to construct contrastive pairs with $\delta(\cdot)$, and $s[\cdot, \cdot]$ denotes a similarity metric between two dense vectors. And $L_\theta^{(\mathrm{c})}$ denotes this loss function is optimized w.r.t the subscript θ. It is also noteworthy that $x_j \sim X$ is usually implemented by using other in-batch instances during mini-batch SGD (a.k.a in-batch negatives).

However, when switching to unsupervised sentence embedding, augmenting an input sentence by a fully random chop or permutation may become very intractable. This is because these operations are most likely to destroy the original sentence in both semantics and syntax and cause trivial positive augmentations. Hence, many research efforts have been made to design δ for effective positive augmentations in the NLP community. These efforts mainly fall into two categories in terms of augmentation format – 'discrete' and 'continuous'. Discrete augmentation format denotes operating directly on the inputted sentence, where δ is defined as word deletion, shuffling (Yan et al., 2021), back translation (Xie et al., 2020), etc. In contrast, continuous ones operate on hidden states or network parameters, where δ is defined as network twice dropout (Gao et al., 2021), etc.

Nonetheless, compared to visual contrastive learning that barely introduces new data distribution for the positive instances, such heuristic augmentation methods in the text field cause shortcut learning (Ilyas et al., 2019; Du et al., 2021) – each method exposes the learning procedure to potential biases towards the augmented instances and thus corrupts the quality of learned embeddings. Therefore, existing unsupervised contrastive sentence embedding works usually depend on the limited- or even mono-augmenting methods for their positive instances and inevitably suffer from the biases in the positive instances.

2.2 Contrast-Cooperation with Peers

To prevent learning shortcuts caused by the potential biases from mono-augmenting strategy and exploit rich relations among diverse augmentations for more effective positives, we propose a brand-new contrastive learning method, called peer-contrastive learning (PCL). Besides the vanilla contrastive objective, it contains a novel 'contrast-cooperation' learning mechanism, which we will detail in the following.

2.2.1 Multi-Augmenting Strategy

First, we adopt a multi-augmenting strategy for extensively diverse augmentations. Given a sentence $x_i \sim X$, it considers extensive augmentation methods from both continuous and discrete perspectives.

This can be formally written as

$$\Delta = \{\delta_k | \delta_k \in \Delta^{(c)} \cup \Delta^{(d)}\}, \tag{2}$$

where Δ denotes a set of multiple augmentation methods from both the continuous $\Delta^{(c)}$ set and discrete $\Delta^{(d)}$ set, and $|\Delta| = K$. Then, we can obtain diverse augmented positives by applying Δ to a sentence $x_i \sim X$, i.e.,

$$\hat{X}^i = \{x_k^i = \delta_k(x_i) | \delta_k \in \Delta\}. \tag{3}$$

The contrastive sentence embedding based on diverse positives can mitigate the biases towards mono-augmenting strategy, but it comes with a double-edged sword. That is, the $\hat{x}_k^i \sim \hat{X}^i$ varies a lot with many factors (e.g., input sentence x_i and augmentation method δ_k), making it hard to control the quality of each x_k^i. To one extreme, one augmentation can become ineffective and even poisonous if its semantics is largely changed and thus corrupt the model.

2.2.2 Contrast among Peer Positives

To effectively learn from the uneven qualities of the augmented positives, we propose a brand-new peer-contrastive learning framework that not only performs the vanilla positive-negative contrast but a positive-positive contrast. This is because our diverse augmentations provide a great opportunity to model rich structured relations among the positives and improve the probability of 'at-least-one' effective positive in $\hat{X}^i$. And the positive-positive contrast can mimic 'peer-competition' to highlight more likely effective positives but weaken the others' effects by suppressing them.

Formally, we first derive a group-wise probability distribution by contrasting the anchor x_i with both diverse positives $\hat{X}^i$ and in-batch negatives $x_j \sim X \wedge j \neq i$. That is,

$$\begin{aligned} \boldsymbol{p}^{\text{P-Cf}}_{\theta^1,\theta^2}(x_i) &:= \text{P-Cf}(x_i, \Delta^{(d)}; \theta^1, \theta^2) = \\ &\text{softmax}(\{s[f_{\theta^1}(x_i), f_{\theta^2}(\hat{x}_k^i)/\tau]\}_{\hat{x}_k^i \sim \hat{X}^i} + \\ &\qquad \{s[f_{\theta^1}(x_i), f_{\theta^2}(x_j)/\tau]\}_{x_j \sim X \wedge j \neq i}), \end{aligned} \tag{4}$$

where '+' here denotes a union of two sets. Identical to the vanilla contrastive sentence embedding (Gao et al., 2021), we also leverage a softmax normalization to fulfill peer-contrast among augmented positives. Please note we introduce θ^1 and θ^2 for clear deliveries in the remaining sections, and the two parameters here can be either tied (i.e., $\theta^1 = \theta^2 = \theta$) or not. Although using the augmented positives to 'compete' each peer sounds attractive for contrastive learning, one critical question remains about how to learn merely from effective positives and guide the positive-peer contrasts $\boldsymbol{p}^{\text{P-Cf}}_{\theta^1,\theta^2}(x_i)$ in a fully unsupervised way.

2.2.3 Cooperation across Peer Networks

We propose a cooperative learning framework to learn contrasts among the augmented positives. It contains two peer embedding networks, and the two networks learn from each other to prevent error reinforcement in sole-network and achieve a common agreement from different views. Specifically, we first build a peer network θ' which acts like a momentum encoder (He et al., 2020) to cooperatively learn with θ. Here, θ and θ' can be untied or even heterogeneous. Then, we present the loss of the momentum-like cooperative learning, which is a combination of two Kullback–Leibler divergence losses. That is

$$\begin{aligned} L^{(p)}_{\theta,\theta'}(X, \Delta; \theta, \theta') = \\ \mathbb{E}_{x_i}[\text{KL}[\boldsymbol{p}^{\text{P-Cf}}_{\theta,\theta}(x_i), \boldsymbol{p}^{\text{P-Cf}}_{\theta',\theta'}(x_i)] + \\ \text{KL}[\boldsymbol{p}^{\text{P-Cf}}_{\theta,\theta'}(x_i), \boldsymbol{p}^{\text{P-Cf}}_{\theta',\theta}(x_i)]]. \end{aligned} \tag{5}$$

We call this 'momentum-like' since θ' is not strictly a history of θ for more different views but depends on the second KL term to prevent significant divergence from θ. Meanwhile, the first KL term is to reach an agreement between the main network θ and its peer network θ'. This 'learning-from-agreement' paradigm, including mutual-distillation (Zhang et al., 2018) and denosing-by-agreement (Wei et al., 2020), is proven effective in improving performance and learning from label noises by prior supervised works. In contrast, we hold a distinct motivation that the implicit relations in a group of diverse positives and in-batch negatives are expected to unsupervisedly match each peer embedding network with another view (e.g., structures and parameters).

Remark. The meaning of '*peer*' has two folds: (i) It denotes that we want to learn the rich structured relations among the diverse augmented positives to highlight the effective ones; (ii) It involves a cooperative learning framework based on peer networks for modeling positive-positive contrasts to achieve PCL with diverse augmentations.

2.3 Training and Inference

Training Objective. We write the loss as a combination of (i) our proposed contrast-cooperation learning for both θ and θ' simultaneously to highlight more effective positives and (ii) vanilla contrastive learning that is applied to θ and θ' separately and based on our diverse augmentations Δ for their strong anti-bias initializations. That is,

$$L^{(\text{PCL})} = L^{(p)}_{\theta,\theta'}(X, \Delta; \theta, \theta') + \beta \sum_{\delta_k \in \Delta} \left[L^{(\text{c})}_{\theta}(X, \delta_k; \theta) + L^{(\text{c})}_{\theta}(X, \delta_k; \theta') \right], \quad (6)$$

where β is a hyperparameter to control if the training inclines to vanilla CL for the anti-bias purpose. Hence, β could be annealing to provide strong unbiased initializations for different views at the beginning and then focus on contrast-cooperation with peers for more effective learning.

Inference. Due to the symmetrical learning paradigm, we empirically found θ and θ' achieve comparable performance in our pilot experiments. Nonetheless, we only use the main embedding network θ rather than ensembles them to encode each sentence for fair comparisons with its competitors.

3 Experiments

3.1 Unsupervised Corpus and Benchmark

We train and evaluate our model in a fully unsupervised manner. Following Gao et al. (2021), we train our model on 10^6 randomly sampled sentences from Wikipedia English. We evaluate our model on the semantic textual similarity (STS) tasks without using any STS training data. We report results on 7 datasets, namely the STS benchmark (STSb) (Cer et al., 2017) the SICK-Relatedness (SICK-R) dataset (Marelli et al., 2014) and the STS tasks 2012 - 2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016) (STS12-STS16). These datasets provide a gold standard semantic similarity between 0 and 5 for each sentence pair, which include texts from various domains, and we obtain them from the SentEval toolkit (Conneau and Kiela, 2018).

3.2 Implementation of PCL

Augmentation Strategies In this paper we utilize five unsupervised augmentation strategies that are commonly adopted in previous works (Wei and Zou, 2019; Yan et al., 2021; Gao et al., 2021). Augmentations from discrete perspectives $\Delta^{(d)}$ includes: 1) Shuffled Sentence (SS) shuffles the position of words in the sentence; 2) Inverted Sentence (IS) inverts the original sentence as the augmented sample; 3) Words Repetition (WR) duplicates part of words and randomly insert them into the original sentences; 4) Words Deletion (WD) deletes part of words in the sentences. The augmentations from the continuous perspective $\Delta^{(c)}$ include Dropout (DP). It generates augmentation instances in the embedding level by passing the original sentence again into the encoder with different dropout masks. More implementation details about augmentation are presented in § A due to the page limit.

Network Implementation We initialize the networks θ and θ' with the PLMs checkpoint downloaded from Huggingface's `Transformers`[3] of BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). The encoder consists of 12 and 24 Transformer layers for the base and large model, respectively. The hidden size is set to 768 and 1024, and the number of attention heads is set to 12 and 16 for base and large models, respectively. We choose the representation of the `[CLS]` token as the embedding of the input sentence. The hyperparameter β is set to 1 for training simplicity without tuning, and the number of augmentations K is 9 for base models. Due to the computation resource limitation, particularly for large models, we set K to 4, and the two networks θ and θ' are tied, in which the cooperative learning is performed between the two passes through the network.

Training Setups. We follow common practices and carry out preliminary grid search on the development set of STSb to decide the hyper-parameter configuration. The learning rate is set to 3e-5 for base models and 1e-5 for large models, respectively. Except for learning rate, We use the same training hyper-parameters for all experiments with the batch size of 64 and the maximum length of 32. The temperature parameter τ is set to 0.05, and the dropout probability is set to 0.1. We train our model for 1 epoch and evaluate the model on the STSb development set every 125 steps, and keep the best checkpoint by following Gao et al. (2021).

Evaluation Setups. We evaluate PCL on 7 STS tasks, including STS12-STS16, STSb, and SICK-R as introduced in § 3.1. No training data of STS tasks are used during training and evaluation. Gao

[3] https://github.com/huggingface/transformers

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
GloVe embeddings (avg.)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-$\text{BERT}_{\text{base}}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
$\text{ConSERT}_{\text{base}}$	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SG-OPT_{base}	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
$\text{BERT}_{\text{base}}$-Mirror	69.10	81.10	73.00	81.90	75.70	78.00	69.10	75.50
SimCSE-$\text{BERT}_{\text{base}}$	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
PCL-$\text{BERT}^{\dagger}_{\text{base}}$	**72.84**	**83.81**	**76.52**	83.06	**79.32**	**80.01**	**73.38**	**78.42**
– Avg. of seeds$^{\dagger *}$	72.74	83.36	76.05	**83.07**	79.26	79.72	72.75	78.14
$\text{RoBERTa}_{\text{base}}$ (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\text{base}}$-whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-$\text{RoBERTa}_{\text{base}}$	52.41	75.19	65.52	77.12	78.63	72.41	68.62	69.99
$\text{RoBERTa}_{\text{base}}$-Mirror	66.60	**82.70**	74.00	82.40	79.70	79.60	69.70	76.40
SimCSE-$\text{RoBERTa}_{\text{base}}$	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
PCL-$\text{RoBERTa}^{\dagger}_{\text{base}}$	71.13	82.38	**75.40**	83.07	**81.98**	**81.63**	**69.72**	77.90
– Avg. of seeds$^{\dagger *}$	**71.54**	**82.70**	75.38	**83.31**	81.64	81.61	69.19	**77.91**
$\text{BERT}_{\text{large}}$-flow	65.20	73.39	69.42	74.92	77.63	72.26	62.50	70.76
SG-$\text{OPT}_{\text{large}}$	67.02	79.42	70.38	81.72	76.35	76.16	70.20	74.46
$\text{ConSERT}_{\text{large}}$	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE-$\text{RoBERTa}_{\text{large}}$	72.86	83.99	75.62	84.77	**81.80**	81.98	71.26	78.90
PCL-$\text{RoBERTa}^{\dagger}_{\text{large}}$	74.08	84.36	76.42	85.49	81.76	82.79	71.51	79.49
– Avg. of seeds$^{\dagger *}$	73.76	84.59	76.81	85.37	81.66	**82.89**	70.33	79.34
PCL-$\text{BERT}^{\dagger}_{\text{large}}$	74.87	**86.11**	78.29	**85.65**	80.52	81.62	**73.94**	**80.14**
– Avg. of seeds$^{\dagger *}$	**74.89**	85.88	**78.33**	85.30	80.13	81.39	73.66	79.94

Table 2: The models' performance comparison on STS tasks. We report the Spearman's correlation ρ (%) on 7 STS datasets. We highlight the highest numbers among models with the same pre-trained encoder. †: Our models. *: We also run our models five times with different random seeds and report the *average* of these five results on each column as the final number.

et al. (2021) has studied the evaluation settings for sentence embedding. We adopt their suggestions and follow the standard settings in Sentence-BERT (Reimers and Gurevych, 2019). Specifically, we do not train an additional regressor for STSb and SICK-R, use Spearman's correlation as the metric, concatenate all tasks and report the overall Spearman's correlation. To fairly compare with previous approaches, we use the evaluation scripts released by Gao et al. (2021)[4]. Moreover, we train our model for five times with different random seeds and report the *average* of these five results. We also evaluate PCL on 7 transfer tasks (Conneau and Kiela, 2018). PCL achieves competitive performance and the detailed results are presented in Appendix C.2. As mentioned in previous works (Reimers and Gurevych, 2019; Gao et al., 2021), the main goal of sentence embeddings is to cluster semantically similar sentences. Hence we only take STS as the main results in this paper.

[4]https://github.com/princeton-nlp/SimCSE

3.3 Competitive Baselines

We compare our model with previous state-of-the-art unsupervised sentence embedding approaches, including basic embedding approaches (e.g. average of GloVe (Pennington et al., 2014), BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) embeddings) and contemporary contrastive learning approaches (e.g. IS-BERT (Zhang et al., 2020), ConSERT (Yan et al., 2021), SG-OPT (Kim et al., 2021), Contrastive Tension (Carlsson et al., 2021), DeCLUTR (Giorgi et al., 2021), Mirror-BERT (Liu et al., 2021b) and SimCSE (Gao et al., 2021)). SGPT (Muennighoff, 2022) and Sentence-T5 (Ni et al., 2021) are proposed with new paradigm and far larger models, which underperform with comparable model size. Trans-Encoder (Liu et al., 2021a) proposes a cooperative method with in-domain pairwise data for mutual benefits of bi- and cross-encoder, making the results incomparable. Please refer to § B for more details.

3.4 Main Quantitative Results

Experimental results on STS tasks are shown in Table 2. We can find that our PCL significantly outperforms the previous best result on all seven tasks as well as the average STS score with a large margin compared to the baseline methods based on $\text{BERT}_{\text{base}}$ or $\text{RoBERTa}_{\text{base}}$ PLMs. Specifically, PCL improves the previous best result on average STS score from 76.25 to 78.42 for $\text{BERT}_{\text{base}}$ and 76.57 to 77.91 for $\text{RoBERTa}_{\text{base}}$, respectively. As SimCSE (Gao et al., 2021) did not report their performance on $\text{BERT}_{\text{large}}$, we compare all large models together, and the results are shown in the last rows of the table. We can observe that PCL outperforms the best result on all tasks apart from the STS16. Despite this, our PCL still obtain an improvement from 78.90 to 80.14 on the average STS score. Our PCL achieves more significant improvement over the base models than the large models, And even so, PCL still outperforms SimCSE on almost all tasks in large models and all tasks in base models, which shows that PCL is effective across different model sizes and different types of PLMs.

3.5 Analysis of Diverse Augmentations

Diversity and the number of augmentations are two crucial factors of PCL. In this section, we test the performance of PCL with varying K and diversity. For PCL and all variants of PCL in this section, we train them for 5 times with different random seeds, and take the average as the final results.

The number of augmentations. To mitigate the model bias towards the mono-augmenting strategy, we propose to augment the input sentence with a group of positive instances. The number of augmentations K is a crucial hyper-parameter in this framework. To check if the performance of PCL is sensitive to K, we conduct experiments on $\text{PCL-BERT}_{\text{base}}$ with varying K on 7 STS tasks and report the average STS score. We keep the diversity of augmentations Δ as much as possible when $K > 1$. As shown in Figure 2, the performance of PCL maintains an upward trend with increasing K. This indicates that multiple augmentation strategy improves unsupervised sentence embeddings compared with learning with mono-augmenting strategy. This supports our motivation that contrastive learning with mono-augmenting strategy causes learning shortcuts. More detailed results on all 7 STS tasks are presented in § C.3.

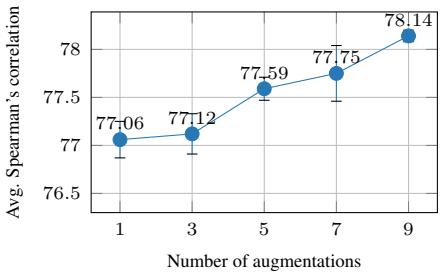


Figure 2: Effect of the number of augmentations.

The diversity of augmentations. Another critical factor is the diversity of augmentations. We fix $K = 9$ and reduce the diversity of augmentations Δ to check if PCL is sensitive to the diversity. We conduct experiments on $\text{PCL-BERT}_{\text{base}}$ with $K = 9$ but only use *one* type of augmentation strategy from discrete and continuous perspective, respectively. In other words, we keep at least one DP augmentation for all variants. We compare PCL with five mono-augmentation variants that are denoted as PCL_{DP}, PCL_{SS}, PCL_{IS}, PCL_{WR} and PCL_{WD}, respectively. The details of augmentation strategies are introduced in § A. Average Spearman's correlation scores of 7 STS tasks are shown in the Figure 3. Experimental results show that PCL significantly outperforms its mono-augmenting variants, even keeping the K constant, indicating a better generalization. This supports our motivation that PCL with diverse augmentations can mitigate the shortcut learning biased towards mono-augmenting strategy. Particularly, SimCSE utilizes dual-dropout to construct the contrastive pairs, hence the PCL_{DP} variant (9 positive instances generated by the dropout) can be regarded as *SimCSE w/ 9 augmented positive samples*. Our proposed contrasts and cooperation among peers improve SimCSE from 76.25 to 77.14, but the score is still lower than PCL with a large margin. This is another piece of evidence that shows the advantage of the diversity of augmentations. The detailed results on all 7 STS tasks are presented in § C.3.

3.6 Ablation Study

We first check the impact of the proposed two components of PCL, peer-network cooperation and peer-positive contrast, i.e., the two terms in Equation 6 respectively. We designed two variants of PCL on $\text{PCL-BERT}_{\text{base}}$ by removing the cooperation loss and contrast loss, which are denoted as PCL_{noP} and PCL_{noC} respectively. To ensure the networks have the essential ability to learn sentence embeddings, we keep the contrastive loss with a

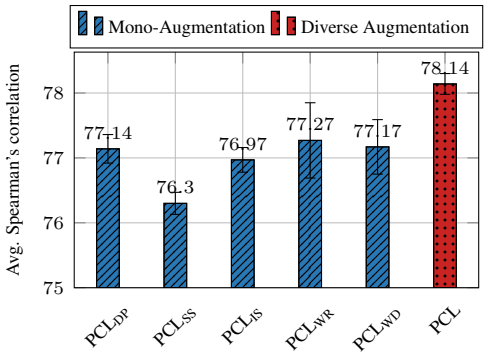


Figure 3: Effect of the diversity of augmentations.

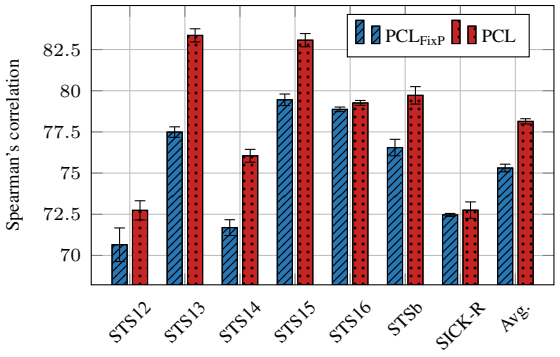


Figure 4: Ablation study on whether updating the extra peer encoder. $\text{PCL}_{\text{FixedP}}$ denotes a variant of PCL that cooperatively learning with a fixed peer network.

Tasks	PCL	PCL_{noP}	PCL_{noC}	SimCSE
STS12	**72.74**	71.15	72.58	68.40
STS13	**83.36**	83.07	80.62	82.41
STS14	**76.05**	75.72	74.15	74.38
STS15	**83.07**	82.93	82.31	80.91
STS16	**79.26**	78.37	79.23	78.56
STSb	**79.72**	78.67	78.47	76.85
SICK-R	**72.75**	70.37	72.06	72.23
Avg.	**78.14**	77.12	77.06	76.25

Table 3: Ablation study on $\text{PCL-BERT}_{\text{base}}$. PCL_{noP} denotes PCL w/o peer-network cooperation. PCL_{noC} denotes PCL w/o peer-positive contrast.

single DP augmentation for PCL_{noC}, which is equal to the setting in Figure 2 when $K = 1$. We train each variant for five times with different random seeds and take the average of these seeds as the final results. As in Table 3, the average scores of PCL drop by 1.02 and 1.08 when removing the cooperation loss and contrast loss, respectively. This indicates that our proposed peer cooperation and peer contrast are both beneficial to unsupervised learning of sentence embeddings. Among the two components, the peer cooperation loss plays a more important role as it incorporates contrasts among peer positives and peer networks, enabling an inherent anti-bias ability and an effective way to learn from diverse augmentations.

Fixed peer encoder vs. trainable peer encoder. Particularly, We are also curious about the impact of 'learning from agreement' (i.e., the second term in Equation 5) in the cooperative learning objective. Therefore, we further test additional variants of PCL with a fixed peer encoder, denoted as $\text{PCL}_{\text{FixedP}}$. Specifically, we download a checkpoint of SimCSE as the peer encoder but fix its parameters while training. Experimental results show that the performance of $\text{PCL}_{\text{FixedP}}$ drops with a large margin on STS12, STS13, STS14, STS15, STSb, and the average STS. The reason may be that although SimCSE is by far the best practice of sentence embeddings, it is still biased towards a mono-augmenting strategy. Hence, cooperative learning with a biased peer network can be harmful to the network with diverse augmentations. This also indicates that it is necessary to simultaneously update the two peer networks and learn the agreement between them in PCL. Furthermore, there can be a considerable discrepancy between the embedding spaces produced by two methods, which hinders the cooperative training of two networks.

4 Related Works

Unsupervised Sentence Embedding. Common practice of unsupervised sentence embedding is to take the average of pre-trained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) PLMs, like BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), Wu et al. (2021a) takes the average of word embeddings as context embedding to enhance the language pre-training. Other works also take the `[CLS]` embedding from the last layer of PLMs with post-processing (Li et al., 2020; Su et al., 2021a). Some works (Kiros et al., 2015; Logeswaran and Lee, 2018; Hill et al., 2016) directly train a deep model for sentence embeddings using co-occurrence information. Recent approaches couple PLMs with CL (Zhang et al., 2020; Yan et al., 2021; Kim et al., 2021; Carlsson et al., 2021; Giorgi et al., 2021; Gao et al., 2021; Xie et al., 2022) with a particular single strategy to construct contrastive pairs. It is straightforward to extend mono-augmentation into multi-augmentation to learn expressive representations. For example, CLEAR (Wu et al., 2020) uses various token/span manipulations for noise-invariant representations while Mirror-BERT (Liu et al., 2021b) employs several fast augmentation strategies. However, these methods usually take the augmented positives equally, regardless of the **uncontrollable qualities**.

Thereby, we take a step further to consider the contrasts among the augmented positives to figure out which augmentation is relatively reasonable. This is achieved by our novel cooperative learning method with peer networks. More related to our work, ESimCSE (Wu et al., 2021b) found learning on dual-dropout causes sentence length bias so it employs another augmentation strategy, i.e., word repetition, to prevent the length bias. However, word repetition introduces **learning shortcut** by itself, not to mention it makes the sentence unnatural and even **semantics-wrong**. To circumvent this dilemma, we propose exhaustive augmentations to ensure "*at-least-one*" true positive and reduce learning shortcuts by complementary augmenting strategies. By doing so, PCL achieve a better performance on STS tasks. Please refer to § D.1 for more discussion details.

Contrastive Learning. The main idea of CL is to pull semantic close neighbors close and push non-neighbors apart (Hadsell et al., 2006; Zbontar et al., 2021). It is shown to be a successful way to learn representation. Approaches in computer vision (CV) (Chen et al., 2017; Wu et al., 2018; Tian et al., 2020; He et al., 2020; Zbontar et al., 2021) try to make an image to be invariant to transformations on itself, while remaining discriminative to other images. More references in CV are discussed in the recent survey (Jaiswal et al., 2021). CL is also coupled with PLMs to learn sentence embeddings. But, recent works (Xiao et al., 2021) argue that learning invariance to particular transformations may be harmful to the robustness of the model. This also supports our idea of leveraging diverse augmentations to improve unsupervised sentence embeddings from another angle.

Learning from Agreement. Another line of work close to ours is learning from agreement, e.g., Decoupling (Malach and Shalev-Shwartz, 2017), Co-teaching (Han et al., 2018; Yu et al., 2019), and mutual CL Yang et al. (2021). This paradigm has been proven effective in improving model performance and learning with label noises by prior fully-supervised works. As text data is discrete and compositional, qualities of multiple augmentations can be uneven, which may corrupt the generalization of sentence embeddings. Besides widely used regularization like dropout (Srivastava et al., 2014) and weight decay (Krogh and Hertz, 1991), we consider learning from agreement paradigm to offer a robust way to learn from our diverse positives.

5 Conclusion

In this paper, we propose a brand-new contrastive learning framework, dubbed as peer-contrastive learning (PCL), to capture rich relations among diverse positive peers and highlight effective positives, which are learned by cooperative learning by peer networks. Besides inherent anti-bias ability by diverse augmentations, it can learn from unsupervised corpus more effectively than vanilla contrastive in the text field. Experiments show that the number and diversity of augmentations are crucial to PCL. Ablation study also shows that the two components of PCL, i.e., peer-positive contrast and peer-network cooperation, are both beneficial to unsupervised CL for sentence embeddings.

Limitation. We also recognize that our PCL framework has its certain limitations: (i) Due to peer positives and encoders, our framework needs higher (~3×, i.e., 2.5 GPU-hours for base models) computation overheads compared to vanilla CL. Nonetheless, the acceptable extra overhead and same inference makes our framework still practical and scalable. (ii) As a work addressing the general shortcut learning problem in a fundamental task, the proposed PCL is only evaluated on resources in English. It can be further extend to more applications such as in low-resource or other languages. (iii) The performance of our framework relies on the choice of augmentation methods, and it is hard to strictly claim which combination of the methods is optimal except experimental verification. Although we have analysed the effect of varying combinations of augmentations with extensive experiments, we can only select several widely-adopted augmentations to evaluate the general effectiveness of our framework.

Ethics Statement

This paper investigates unsupervised contrastive learning for sentence embedding. There will not be any ethical problems or negative social consequences from the research. The data in this paper are all publicly available and are widely adopted by researchers. The proposed method does not introduce ethical bias in the data.

Acknowledgement

We thank Yoshimasa Tsuruoka, Ryokan Ri and Jing Zhou for valuable discussions. Qiyu Wu was supported by JST SPRING, Grant Number JPMJSP2108.

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 SemEval*.

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 SemEval*.

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 *Proceedings of SemEval*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of SemEval*.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * SEM 2013 shared task: Semantic textual similarity. In *Proceedings of SemEval*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *ICLR*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Mengnan Du, Varun Manjunatha, R. Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu models. In *NAACL*.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for bert's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

John Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. *ArXiv*, abs/2006.03659.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. *ArXiv*, abs/1905.02175.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. *ArXiv*, abs/2106.07345.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Anders Krogh and John A Hertz. 1991. A simple weight decay can improve generalization. In *NIPS*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *EMNLP*.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2021a. Transencoder: Unsupervised sentence-pair modelling through self-and mutual-distillations. *arXiv preprint arXiv:2109.13059*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459.

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*, abs/1907.11692.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Eran Malach and Shai Shalev-Shwartz. 2017. Decoupling" when to update" from" how to update". In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 961–971.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021a. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Peng Su, Yifan Peng, and K. Vijay-Shanker. 2021b. Improving BERT model using contrastive learning for biomedical relation extraction. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 1–10.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *ECCV*.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In *EMNLP*, pages 671–688.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13723–13732.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *ArXiv*, abs/1901.11196.

Qiyu Wu, Chen Xing, Yatao Li, Guolin Ke, Di He, and Tie-Yan Liu. 2021a. Taking notes on the fly helps language pre-training. In *International Conference on Learning Representations*.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021b. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *ArXiv*, abs/2109.04380.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

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*.

Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. 2021. What should not be contrastive in contrastive learning. *ArXiv*, abs/2008.05659.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. *arXiv: Learning*.

Yutao Xie, Qiyu Wu, Wei Chen, and Tengjiao Wang. 2022. Stable contrastive learning for self-supervised sentence embeddings with pseudo-siamese mutual learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3046–3059.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. 2021. Mutual contrastive learning for visual representation learning. *ArXiv*, abs/2104.12565.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *EMNLP*, pages 1601–1610.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328.

A Augmentation Strategies

We propose diverse augmentation strategies for each sentence. In this paper we utilize five unsupervised augmentation strategies that are commonly adopted in previous works (Wei and Zou, 2019; Yan et al., 2021; Gao et al., 2021). Augmentations from discrete perspectives $\Delta^{(d)}$ includes: 1) Shuffled Sentence (SS) shuffles the position of words in the sentence. SS corrupts the order of the original sentence but preserves the semantic information of words; 2) Inverted Sentence (IS) inverts the original sentence as the augmented sample. Apart from the reading order, IS preserves all language properties even including n-gram statistics (Dufter and Schütze, 2020); 3) Words Repetition (WR) duplicates part of words and randomly insert them into the original sentences; 4) Words Deletion (WD) deletes part of words in the sentences. WD and WR change the length and words of the original sentence but roughly preserve the reading order. The deletion and repetition ratio are empirically set to 0.2. The augmentations from the continuous perspective $\Delta^{(c)}$ include Dropout (DP). It generates augmentation instances in the embedding level by passing the original sentence again into the encoder with different dropout masks. The above five strategies can be repeatedly applied in practice. As there is randomness in the processes of augmentation and encoding, repeatedly generated instances with the same strategy can be regarded as diverse positives. But the diversity may accordingly decline. Note that the primary goal of this paper is to verify the effectiveness of our PCL framework, hence all of the chosen augmentations are common and simple. We speculate that our PCL can be further improved with more fine-tuned augmentation strategies.

B Baselines

We compare PCL with previous state-of-the-art unsupervised sentence embedding approaches. Basic approaches include taking the average of GloVe, BERT, or RoBERTa embeddings. Besides, BERT-whitening and BERT-flow post-process the embeddings distribution of BERT. We also compare PCL with recent approaches using contrastive learning, including IS-BERT, ConSERT, SG-OPT, Contrastive Tension, DeCLUTR, and SimCSE. The following are the details of these baselines,

- GloVe (Pennington et al., 2014) maps words into a meaningful space where the distance between words is related to semantic similarity. The results of the average of GloVe embeddings are from Reimers and Gurevych (2019).
- Su et al. (2021a) takes the average of the first and last layers of BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) embeddings. We report the results from Gao et al. (2021).
- BERT-whitening (Su et al., 2021a) and BERT-flow (Li et al., 2020) post-process the embeddings distribution of BERT. We report the results from Gao et al. (2021) for a fair comparison.
- IS-BERT (Zhang et al., 2020) encourages the representation of a specific sentence to encode all aspects of its local context information, using local contexts derived from other input sentences as negative examples for contrastive learning. We report the results from the original paper.
- ConSERT (Yan et al., 2021) contrasts a pair of sentences augmented by different augmentation methods. We report the results in the original paper.
- SG-OPT (Kim et al., 2021) is a contrastive learning method using self-guidance. The results are from the original paper.
- Contrastive Tension (CT) (Carlsson et al., 2021) propose a training objective that aligns the embeddings of the same sentence encoded by two different encoders. We report the results from Gao et al. (2021).
- DeCLUTR (Giorgi et al., 2021) is a contrastive approach that takes different spans from the same document as contrastive pairs. The results are from Gao et al. (2021)
- Mirror-BERT (Liu et al., 2021b) employs several fast augmentation strategies for effective representations. The results are from the original paper.
- SimCSE (Gao et al., 2021) contrasts a pair of embeddings of one sentence encoded with different dropout masks. The results are from the original paper. We had re-run SimCSE with same setups and it performs worse than the numbers reported in the original paper (e.g., 75.36 averaged over 5 seeds on $\text{BERT}_{\text{base}}$).

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Models w/o PLMs								
GloVe embeddings (avg.)	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip-thought	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
Base Models								
Avg. BERT embeddings	78.66	86.25	94.37	88.66	84.40	92.80	69.54	84.94
BERT-`[CLS]` embedding	78.58	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS-BERT$_{\text{base}}$	81.09	87.18	94.96	88.75	85.96	88.64	74.24	85.83
SimCSE-BERT$_{\text{base}}$	81.18	86.46	94.45	88.88	85.50	89.80	74.43	85.81
SimCSE-RoBERTa$_{\text{base}}$	81.04	87.74	93.28	86.94	86.60	84.60	73.68	84.84
Ours-BERT$_{\text{base}}$	80.11	85.25	94.22	89.15	85.12	87.40	76.12	85.34
Ours-RoBERTa$_{\text{base}}$	81.83	87.55	92.92	87.21	87.26	85.20	76.46	85.49
Large Models								
SimCSE-RoBERTa$_{\text{large}}$	82.74	87.87	93.66	88.22	88.58	92.00	69.68	86.11
Ours-BERT$_{\text{large}}$	82.47	87.87	95.04	89.59	87.75	93.00	76.00	87.39
Ours-RoBERTa$_{\text{large}}$	84.47	89.06	94.60	89.26	89.02	94.20	74.96	87.94

Table 4: Transfer task results (measured as accuracy).

We reported higher numbers for a fair comparison.

Due to the surge of this topic, many concurrent works emerge with two trends: *new model structure* and *more in-domain data*. SGPT (Muennighoff, 2022) and Sentence-T5 (Ni et al., 2021) are proposed with new paradigm and far larger models, which underperform with comparable model size. Trans-Encoder (Liu et al., 2021a) proposes a cooperative method with in-domain pairwise data for mutual benefits of bi- and cross-encoder, making the results incomparable.

C Additional Experimental results

C.1 Comparison of controlled setups

We can also find some variants of SimCSE in our controlled experiments. For example, PCL$_{noP}$ in Table 3 is regarded as *SimCSE w/ multi-augmentations*. PCL$_{K=1}$ in Figure 2 can be regarded as *SimCSE w/ peer-network cooperation*, and PCL$_{DP}$ in Figure 3 is regarded as *SimCSE w/ 9 dropout augmented positive samples*. As we discussed in the § 3.5 and § 3.6, the comparison between the variants of SimCSE and PCL show the advantages and importance of our proposed peer-contrast and peer-cooperation.

C.2 Transfer tasks

We also evaluate PCL on 7 transfer tasks (Conneau and Kiela, 2018). As the Table 4 shows, PCL achieves competitive performance compared with baselines. Note that as mentioned in previous works (Reimers and Gurevych, 2019; Gao et al., 2021), the main goal of sentence embeddings is to cluster semantically similar sentences. Hence we only take STS as the main results in this paper.

C.3 Detailed experimental results on analysis of diverse augmentations

In this section, we present detailed results of experiments of diverse augmentations on all 7 STS tasks. We test the performance of PCL with varying K and diversity. Experimental results are shown in Table 5. As the results and our analysis in § 3.5 show, the performance of PCL maintains an upward trend with increasing K. Besides, it is also shown that PCL significantly outperforms its mono-augmenting variants, even keeping the K of them constant, which indicates a better generalization. As a result, PCL with more diverse augmentations performs better. We also speculate that our PCL can be further improved with larger K and more fine-tuned augmentation strategies.

D Discussion

D.1 Distinction between PCL and other contemporary methods.

Unsupervised sentence embedding w/ multiple positive augmentations. It's straightforward to extend mono-augmentation into multi-augmentation to learn expressive representations. For example, CLEAR (Wu et al., 2020) uses various token/span manipulations for noise-invariant representations while Mirror-BERT (Liu et al.,

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Effect of the number of augmentations								
$\text{PCL}_{\text{K=1}}$	72.58	80.62	74.15	82.31	79.23	78.47	72.06	77.06
$\text{PCL}_{\text{K=3}}$	72.66	82.96	74.44	81.94	78.38	77.93	71.55	77.12
$\text{PCL}_{\text{K=5}}$	73.44	81.80	74.59	82.63	79.40	79.05	72.25	77.59
$\text{PCL}_{\text{K=7}}$	73.49	81.93	74.84	82.24	79.75	79.37	72.62	77.75
$\text{PCL}_{\text{K=9}}$	72.74	83.36	76.05	83.07	79.26	79.72	72.75	78.14
Effect of the diversity of augmentations								
PCL_{DP}	71.20	82.53	74.66	82.67	78.92	78.06	71.94	77.14
PCL_{SS}	70.60	80.73	74.11	82.18	78.90	77.91	69.69	76.30
PCL_{IS}	70.95	81.31	74.51	82.24	79.23	78.44	72.09	76.97
PCL_{WR}	71.82	82.56	74.75	82.34	78.85	78.72	71.88	77.27
PCL_{WD}	73.08	81.84	74.17	82.50	78.81	78.52	71.23	77.17
PCL	72.74	83.36	76.05	83.07	79.26	79.72	72.75	78.14

Table 5: Effect of the number and diversity of augmentations. We report the Spearman's correlation ρ (%) on 7 STS datasets. All variants are run for five times with different random seeds and the *average* of these five results on each column is reported as the final number.

2021b) employs several fast augmentation strategies for effective representations. However, these methods usually take the augmented positives equally, regardless of the **uncontrolable qualities**. For example, given "*Two men are wrestling on the floor*", we get "*Two men are squirming on the floor*" and "*Two persons are wrestling on the floor*" by word replacement, but only the 2nd is reasonable. Thereby, we take a step further to consider the contrasts among the augmented positives to figure out which augmentation is relatively reasonable. This is achieved by our novel cooperative learning method with peer networks.

Unsupervised sentence embedding for anti-bias. More related to our work, ESimCSE (Wu et al., 2021b) found learning on dual-dropout causes sentence length bias so it employs another augmentation strategy, i.e., word repetition, to prevent the length bias. However, word repetition introduces **learning shortcut** (i.e., *order* and *BoW* as in Table 1) by itself, not to mention it makes the sentence unnatural and even **semantics-wrong** (e.g., repeating "*no*"). To circumvent this dilemma, we propose exhaustive augmentations to ensure "*at-least-one*" true positive and reduce learning shortcuts by complementary augmenting strategies (See Figure 1 and 3: if we employ every strategy, the shortcuts can be blocked). Nonetheless, PCL still has a better performance (78.42 vs. 78.27) compared with ESimCSE on STS tasks.

Cooperative learning has more parameters, is it the reason leading better performance? One of the advantages of our cooperative learning is to effectively learn from the positives with uncontrolled qualities, or it can be also interpreted as 'noisy labels'. In the noisy circumstance, more parameters not necessarily lead to better performance. And it can be anticipated that it possibly leads to worse performance because the over fitting in the noisy positives.

D.2 Efficiency & Impact of augmentations.

Efficiency Due to peer positives and encoders, our framework needs higher (~3×, i.e., 2.5 GPU-hours for base models) computation overheads compared to vanilla CL (Gao et al., 2021). Nonetheless, the same inference makes our framework still practical and scalable. Since PCL contains more loss items, we do not see any significant difference in convergence time.

Impact of augmentations The performance of our framework relies on the augmentation methods, and it is hard to claim which combination of the methods is optimal except experimental verification. In this work, we only intuitively select several methods without extensive trials. We have illustrated the performance of mono-augmentation in Figure 3.