Debiased Contrastive Learning of Unsupervised Sentence Representations

Kun Zhou[1,3], **Beichen Zhang**[2], **Wayne Xin Zhao**[2,3†] and **Ji-Rong Wen**[2,3]
[1]School of Information, Renmin University of China.
[2]Gaoling School of Artificial Intelligence, Renmin University of China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
francis_kun_zhou@163.com, {zhangbeichen724,batmanfly}@gmail.com, jrwen@ruc.edu.cn

Abstract

Recently, contrastive learning has been shown to be effective in improving pre-trained language models (PLM) to derive high-quality sentence representations. It aims to pull close positive examples to enhance the alignment while push apart irrelevant negatives for the uniformity of the whole representation space. However, previous works mostly adopt in-batch negatives or sample from training data at random. Such a way may cause the sampling bias that improper negatives (*e.g.,* false negatives and anisotropy representations) are used to learn sentence representations, which will hurt the uniformity of the representation space. To address it, we present a new framework **DCLR** (Debiased Contrastive Learning of unsupervised sentence Representations) to alleviate the influence of these improper negatives. In DCLR, we design an instance weighting method to punish false negatives and generate noise-based negatives to guarantee the uniformity of the representation space. Experiments on seven semantic textual similarity tasks show that our approach is more effective than competitive baselines. Our code and data are publicly available at the link: https://github.com/RUCAIBox/DCLR.

1 Introduction

As a fundamental task in the natural language processing (NLP) field, unsupervised sentence representation learning (Kiros et al., 2015; Hill et al., 2016) aims to derive high-quality sentence representations that can benefit various downstream tasks, especially for low-resourced domains or computationally expensive tasks, *e.g.,* zero-shot text semantic matching (Qiao et al., 2016), large-scale semantic similarity comparison (Agirre et al., 2015), and document retrieval (Le and Mikolov, 2014).

Recently, pre-trained language models (PLMs) (Devlin et al., 2019) have become a widely-used semantic representation approach, achieving remarkable performance on various NLP tasks. However, several studies have found that the native sentence representations derived by PLMs are not uniformly distributed with respect to directions, but instead occupy a *narrow cone* in the vector space (Ethayarajh, 2019), which largely limits their expressiveness. To address this issue, contrastive learning (Chen et al., 2020) has been adopted to refine PLM-derived sentence representations. It pulls semantically-close neighbors together to improve the alignment, while pushes apart non-neighbors for the uniformity of the whole representation space. In the learning process, both positive and negative examples are involved in contrast with the original sentence. For positive examples, previous works apply data augmentation strategies (Yan et al., 2021) on the original sentence to generate highly similar variations. While, negative examples are commonly sampled from the batch or training data (*e.g.,* in-batch negatives (Gao et al., 2021)) at random, due to the lack of ground-truth annotations for negatives.

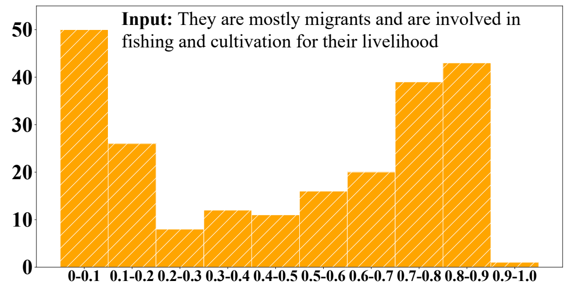


Figure 1: The distribution of cosine similarity between an input sentence and 255 in-batch negatives from the commonly-used Wikipedia Corpus. It is evaluated by the SimCSE model (Gao et al., 2021). Almost half of the negatives have high similarities with the input.

Although such a negative sampling way is simple and convenient, it may cause *sampling bias* and affect the sentence representation learning. First, the sampled negatives are likely to be *false negatives* that are indeed semantically close to the

† Corresponding author

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6120 - 6130
May 22-27, 2022

original sentence. As shown in Figure 1, given an input sentence, about half of in-batch negatives have a cosine similarity above 0.7 with the original sentence based on the SimCSE model (Gao et al., 2021). It is likely to hurt the semantics of the sentence representations by simply pushing apart these sampled negatives. Second, due to the anisotropy problem (Ethayarajh, 2019), the representations of sampled negatives are from the narrow representation cone spanned by PLMs, which cannot fully reflect the overall semantics of the representation space. Hence, it is sub-optimal to only rely on these representations for learning the uniformity objective of sentence representations.

To address the above issues, we aim to develop a better contrastive learning approach with debiased negative sampling strategies.The core idea is to improve the random negative sampling strategy for alleviating the sampling bias problem. First, in our framework, we design an instance weighting method to punish the sampled false negatives during training. We incorporate a complementary model to evaluate the similarity between each negative and the original sentence, then assign lower weights for negatives with higher similarity scores. In this way, we can detect semantically-close false negatives and further reduce their influence. Second, we randomly initialize new negatives based on random Gaussian noises to simulate sampling within the whole semantic space, and devise a gradient-based algorithm to optimize the noise-based negatives towards the most nonuniform points. By learning to contrast with the nonuniform noise-based negatives, we can extend the occupied space of sentence representations and improve the uniformity of the representation space.

To this end, we propose **DCLR**, a general framework towards Debiased Contrastive Learning of unsupervised sentence Representations. In our approach, we first initialize the noise-based negatives from a Gaussian distribution, and leverage a gradient-based algorithm to update the new negatives by considering the uniformity of the representation space. Then, we adopt the complementary model to produce the weights for these noise-based negatives and randomly sampled negatives, where the false negatives will be punished. Finally, we augment the positive examples via dropout (Gao et al., 2021) and combine them with the above weighted negatives for contrastive learning. We demonstrate that our DCLR outperforms a number of competitive baselines on seven semantic textual similarity (STS) tasks using BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019).

Our contributions are summarized as follows:

(1) To our knowledge, our approach is the first attempt to reduce the sampling bias in contrastive learning of unsupervised sentence representations.

(2) We propose DCLR, a debiased contrastive learning framework that incorporates an instance weighting method to punish false negatives and generates noise-based negatives to guarantee the uniformity of the representation space.

(3) Experimental results on seven semantic textual similarity tasks show the effectiveness of our framework.

2 Related Work

In this section, we review the related work from the following three aspects.

Sentence Representation Learning. Learning universal sentence representations (Kiros et al., 2015; Hill et al., 2016) is the key to the success of various downstream tasks. Previous works can be roughly categorized into supervised (Conneau et al., 2017; Cer et al., 2018) and unsupervised approaches (Hill et al., 2016; Li et al., 2020). Supervised approaches rely on annotated datasets (*e.g.,* NLI (Bowman et al., 2015; Williams et al., 2018)) to train the sentence encoder (Cer et al., 2018; Reimers and Gurevych, 2019). Unsupervised approaches consider deriving sentence representations without labeled datasets, *e.g.,* pooling word2vec embeddings (Mikolov et al., 2013). Recently, to leverage the strong potential of PLMs (Devlin et al., 2019), several works propose to alleviate the anisotropy problem (Ethayarajh, 2019; Li et al., 2020) of PLMs via special strategies, *e.g.,* flow-based approach (Li et al., 2020) and whitening method (Huang et al., 2021). Besides, contrastive learning (Wu et al., 2020; Gao et al., 2021) has been used to refine the representations of PLMs.

Contrastive Learning. Contrastive learning has been originated in the computer vision (Hadsell et al., 2006; He et al., 2020) and information retrieval (Bian et al., 2021; Zhou et al., 2022) field with significant performance improvement. Usually, it relies on data augmentation strategies such as random cropping and image rotation (Chen et al., 2020; Yan et al., 2021) to produce a set of semantically related positive examples for learning, and randomly samples negatives from the batch or whole dataset. For sentence representation learning, contrastive learning can achieve a better balance between alignment and uniformity in semantic representation space. Several works further adopt back translation (Fang and Xie, 2020), token shuffling (Yan et al., 2021) and dropout (Gao et al., 2021) to augment positive examples for sentence representation learning. However, the quality of the randomly sampled negatives is seldom studied.

Virtual Adversarial Training. Virtual adversarial training (VAT) (Miyato et al., 2019; Kurakin et al., 2017) perturbs a given input with learnable noises to maximize the divergence of the model's prediction with the original label, then utilizes the perturbed examples to improve the generalization (Miyato et al., 2017; Madry et al., 2018). A class of VAT methods can be formulated into solving a min-max problem, which can be achieved by multiple projected gradient ascent steps (Qin et al., 2019). In the NLP field, several studies incorporate adversarial perturbations in the embedding layer, and show its effectiveness on text classification (Miyato et al., 2017), machine translation (Sun et al., 2020), and natural language understanding (Jiang et al., 2020) tasks.

3 Preliminary

This work aims to make use of unlabeled corpus for learning effective sentence representations that can be directly utilized for downstream tasks, *e.g.,* semantic textual similarity task (Agirre et al., 2015). Given a set of input sentences $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$, our goal is to learn a representation $h_i \in \mathcal{R}^d$ for each sentence x_i in an unsupervised manner. For simplicity, we denote this process with a parameterized function $h_i = f(x_i)$.

In this work, we mainly focus on using BERT-based PLMs (Devlin et al., 2019; Liu et al., 2019) to generate sentence representations. Following existing works (Li et al., 2020; Yan et al., 2021), we fine-tune PLMs on the unlabeled corpus via our proposed unsupervised learning approach. After that, for each sentence x_i, we encode it by the fine-tuned PLMs and take the representation of the `[CLS]` token from the last layer as its sentence representation h_i.

4 Approach

Our proposed framework DCLR focuses on reducing the influence of sampling bias in the contrastive learning of sentence representations. In this framework, we devise a noise-based negatives generation strategy to reduce the bias caused by the anisotropy PLM-derived representations, and an instance weighting method to reduce the bias caused by false negatives. Concretely, we initialize the noise-based negatives based on a Gaussian distribution and iteratively update these negatives towards non-uniformity maximization. Then, we utilize a complementary model to produce weights for all negatives (*i.e.,* randomly sampled and the noise-based ones). Finally, we combine the weighted negatives and augmented positive examples for contrastive learning. The overview of our DCLR is presented in Figure 2.

4.1 Generating Noise-based Negatives

We aim to generate new negatives beyond the sentence representation space of PLMs during the training process, to alleviate the sampling bias derived from the anisotropy problem of PLMs (Ethayarajh, 2019). For each input sentence x_i, we first initialize k noise vectors from a Gaussian distribution as the negative representations:

$$\{\hat{h}_1, \hat{h}_2, \cdots, \hat{h}_k\} \sim \mathcal{N}(0, \sigma^2), \quad (1)$$

where σ is the standard variance. Since these vectors are randomly initialized from such a Gaussian distribution, they are uniformly distributed within the whole semantic space. By learning to contrast with these new negatives, it is beneficial for the uniformity of sentence representations.

To further improve the quality of the new negatives, we consider iteratively updating the negatives to capture the non-uniformity points within the whole semantic space. Inspired by VAT (Miyato et al., 2017; Zhu et al., 2020), we design a non-uniformity loss maximization objective to produce gradients for improving these negatives. The non-uniformity loss is denoted as the contrastive loss between the noise-based negatives $\{\hat{h}_j\}$ and the positive representations of the original sentence (h_i, h_i^+) as:

$$L_U(h_i, h_i^+, \{\hat{h}\}) = -\log \frac{e^{\text{sim}(h_i, h_i^+)/\tau_u}}{\sum_{\hat{h}_j \in \{\hat{h}_j\}} e^{\text{sim}(h_i, \hat{h}_i)/\tau_u}}, \quad (2)$$

where τ_u is a temperature hyper-parameter and $\text{sim}(h_i, h_i^+)$ is the cosine similarity $\frac{h_i^\top h_i^+}{||h_i|| \cdot ||h_i^+||}$. Based on it, for each negative $\hat{h}_j \in \{\hat{h}\}$, we opti-

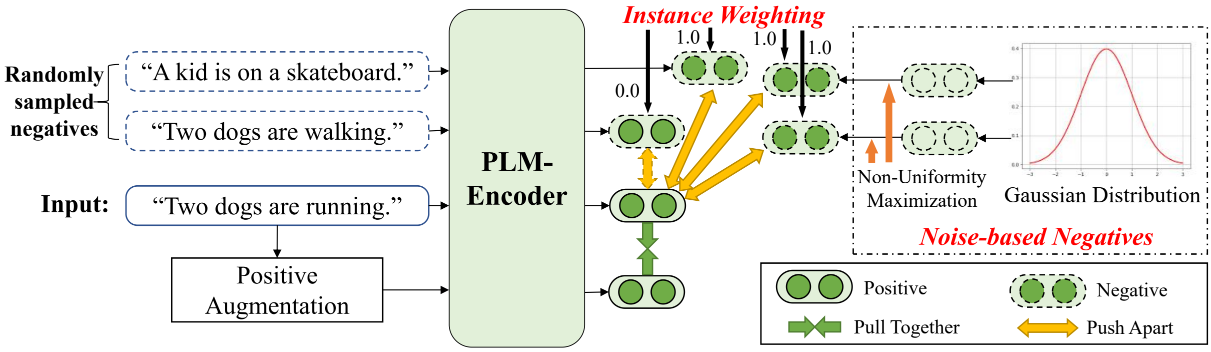


Figure 2: The overview of our DCLR framework with noise-based negatives and the instance weighting strategy. We show the case that a false negative is punished by assigning the weight 0.

mize it by t steps gradient ascent as

$$\hat{h}_j = \hat{h}_j + \beta g(\hat{h}_j)/||g(\hat{h}_j)||_2, \quad (3)$$

$$g(\hat{h}_j) = \nabla_{\hat{h}_j} L_U(h_i, h_i^+, \{\hat{h}\}), \quad (4)$$

where β is the learning rate, $||\cdot||_2$ is the L_2-norm. $g(\hat{h}_j)$ denotes the gradient of $\hat{h}_j$ by maximizing the non-uniformity loss between the positive representations and the noise-based negatives. In this way, the noise-based negatives will be optimized towards the non-uniform points of the sentence representation space. By learning to contrast with these negatives, the uniformity of the representation space can be further improved, which is essential for effective sentence representations.

4.2 Contrastive Learning with Instance Weighting

In addition to the above noise-based negatives, we also follow existing works (Yan et al., 2021; Gao et al., 2021) that adopt other in-batch representations as negatives $\{\tilde{h}^-\}$. However, as discussed before, the sampled negatives may contain examples that have similar semantics with the positive example (*i.e.,* false negatives).

To alleviate this problem, we propose an instance weighting method to punish the false negatives. Since we cannot obtain the true labels or semantic similarities, we utilize a complementary model to produce the weights for each negative. In this paper, we adopt the state-of-the-art SimCSE (Gao et al., 2021) as the complementary model. [1] Given a negative representation h^- from $\{\tilde{h}^-\}$ or $\{\hat{h}\}$ and the representation of the original sentence h_i, we utilize the complementary model to produce the weight as

$$\alpha_{h^-} = \begin{cases} 0, \text{sim}_C(h_i, h^-) \geq \phi \\ 1, \text{sim}_C(h_i, h^-) < \phi \end{cases} \quad (5)$$

where ϕ is a hyper-parameter of the instance weighting threshold, and $\text{sim}_C(h_i, h^-)$ is the cosine similarity score evaluated by the complementary model. In this way, the negative that has a higher semantic similarity with the representation of the original sentence will be regarded as a false negative and will be punished by assigning the weight 0. Based on the weights, we optimize the sentence representations with a debiased cross-entropy contrastive learning loss function as

$$L = -\log \frac{e^{\text{sim}(h_i, h_i^+)/\tau}}{\sum_{h^- \in \{\hat{h}\} \cup \{\tilde{h}^-\}} \alpha_{h^-} \times e^{\text{sim}(h_i, h^-)/\tau}}, \quad (6)$$

where τ is a temperature hyper-parameter. In our framework, we follow SimCSE (Gao et al., 2021) that utilizes dropout to augment positive examples h_i^+. Actually, we can utilize various positive augmentation strategies, and will investigate it in Section 6.1.

4.3 Overview and Discussion

In this part, we present the overview and discussion of our DCLR approach.

4.3.1 Overview of DCLR

Our framework DCLR contains three major steps. In the first step, we generate noise-based negatives to extend in-batch negatives. Concretely, we first initialize a set of new negatives via random Gaussian noises using Eq. 1. Then, we incorporate a gradient-based algorithm to adjust the noise-based negatives by maximizing the non-uniform objective using Eq. 3. After several iterations, we can

[1] For convenience, we utilize SimCSE on BERT-base or RoBERTa-base model as the complementary model.

obtain the noise-based negatives that correspond to the nonuniform points within the whole semantic space, and we mix up them with in-batch negatives to compose the negative set. In the second step, we adopt a complementary model (*i.e.,* SimCSE) to compute the semantic similarity between the original sentence and each example from the negative set, and produce the weights using Eq. 5. Finally, we augment the positive examples via dropout and utilize the negatives with corresponding weights for contrastive learning using Eq. 6.

4.3.2 Discussion

As mentioned above, our approach aims to reduce the influence of the *sampling bias* about the negatives, and is agnostic to various positive data augmentation methods (*e.g.,* token cutoff and dropout). Compared with traditional contrastive learning methods (Yan et al., 2021; Gao et al., 2021), our proposed DCLR expands the negative set by introducing noise-based negatives $\{\hat{h}\}$, and adds a weight term α_{h^-} to punish false negatives. Since the noise-based negatives are initialized from a Gaussian distribution and do not correspond to real sentences, they are highly confident negatives to broaden the representation space. By learning to contrast with them, the learning of the contrastive objective will not be limited by the anisotropy representations derived from PLMs. As a result, the sentence representations can span a broader semantic space, and the uniformity of the representation semantic space can be improved.

Besides, our instance weighting method also alleviates the false negative problem caused by the randomly sampling strategy. With the help of a complementary model, the false negatives with similar semantics as the original sentence will be detected and punished.

5 Experiment - Main Results

5.1 Experiment Setup

Following previous works (Kim et al., 2021; Gao et al., 2021), we conduct experiments on seven standard STS tasks. For all these tasks, we use the SentEval toolkit (Conneau and Kiela, 2018) for evaluation.

Semantic Textual Similarity Task. We evaluate our approach on 7 STS tasks: STS 2012–2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). These datasets contain pairs of two sentences, whose similarity scores are labeled from 0 to 5. The relevance between gold annotations and the scores predicted by sentence representations is measured by the Spearman correlation. Following the suggestions from previous works (Gao et al., 2021; Reimers and Gurevych, 2019), we directly compute the cosine similarity between sentence embeddings for all STS tasks.

Baseline Methods. We compare DCLR with competitive unsupervised sentence representation learning methods, consisting of non-BERT and BERT-based methods:

(1) **GloVe** (Pennington et al., 2014) averages GloVe embeddings of words as the sentence representation.

(2) **USE** (Cer et al., 2018) utilizes a Transformer model that learns the objective of reconstructing the surrounding sentences within a passage.

(3) **CLS**, **Mean** and **First-Last AVG** (Devlin et al., 2019) adopt the `[CLS]` embedding, mean pooling of token representations, average representations of the first and last layers as sentence representations, respectively.

(4) **Flow** (Li et al., 2020) applies mean pooling on the layer representations and maps the outputs to the Gaussian space as sentence representations.

(5) **Whitening** (Su et al., 2021) uses the whitening operation to refine representations and reduce dimensionality.

(6) **Contrastive (BT)** (Fang and Xie, 2020) uses contrastive learning with back-translation for data augmentation to enhance sentence representations.

(7) **ConSERT** (Yan et al., 2021) explores various text augmentation strategies for contrastive learning of sentence representations.

(8) **SG-OPT** (Kim et al., 2021) proposes a contrastive learning method with a self-guidance mechanism for improving the sentence embeddings of PLMs.

(9) **SimCSE** (Gao et al., 2021) proposes a simple contrastive learning framework that utilizes dropout for data augmentation.

Implementation Details. We implement our model based on Huggingface's transformers (Wolf et al., 2020). For BERT-base and RoBERTa-base, we start from the pre-trained checkpoints of their original papers. For BERT-large and RoBERTa-large, we utilize the checkpoints of SimCSE for stabilizing the convergence process. Following Sim-

	Models	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Non-BERT	GloVe (avg.)†	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
	USE†	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
BERT-base	CLS†	21.54	32.11	21.28	37.89	44.24	20.30	42.42	31.40
	Mean†	30.87	59.89	47.73	60.29	63.73	47.29	58.22	52.57
	First-Last AVG‡.	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
	+flow‡	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
	+whitening‡	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
	+Contrastive (BT)†	54.26	64.03	54.28	68.19	67.50	63.27	66.91	62.63
	+ConSERT	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
	+SG-OPT†	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
	+SimCSE	68.40	82.41	74.38	80.91	**78.56**	76.85	**72.23**	76.25
	+DCLR (Ours)	**70.81**	**83.73**	**75.11**	**82.56**	78.44	**78.31**	71.59	**77.22**
BERT-large	CLS†	27.44	30.76	22.59	29.98	42.74	26.75	43.44	31.96
	Mean†	27.67	55.79	44.49	51.67	61.88	47.00	53.85	48.91
	First-Last AVG	57.73	61.17	61.18	68.07	70.25	59.59	60.34	62.62
	+flow†	62.82	71.24	65.39	78.98	73.23	72.72	63.77	70.07
	+whitening	64.34	74.60	69.64	74.68	75.90	72.48	60.80	70.35
	+Contrastive (BT)†	52.04	62.59	54.25	71.07	66.71	63.84	66.53	62.43
	+ConSERT	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
	+SG-OPT†	67.02	79.42	70.38	81.72	76.35	76.16	70.20	74.46
	+SimCSE	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
	+DCLR (Ours)	**71.87**	**84.83**	**77.37**	**84.70**	**79.81**	**79.55**	**74.19**	**78.90**
RoBERTa-base	CLS†	16.67	45.57	30.36	55.08	56.98	45.41	61.89	44.57
	Mean†	32.11	56.33	45.22	61.34	61.98	54.53	62.03	53.36
	First-Last AVG‡	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
	+whitening‡	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
	+Contrastive (BT)†	62.34	78.60	68.65	79.31	77.49	79.93	71.97	74.04
	+SG-OPT†	62.57	78.96	69.24	79.99	77.17	77.60	68.42	73.42
	+SimCSE	**70.16**	81.77	73.24	81.36	80.65	80.22	68.56	76.57
	+DCLR (Ours)	70.01	**83.08**	**75.09**	**83.66**	**81.06**	**81.86**	**70.33**	**77.87**
RoBERTa-large	CLS†	19.25	22.97	14.93	33.41	38.01	12.52	40.63	25.96
	Mean†	33.63	57.22	45.67	63.00	61.18	47.07	58.38	52.31
	First-Last AVG	58.91	58.62	61.44	69.05	65.23	59.38	58.84	61.64
	+whitening	64.17	73.92	71.06	76.40	74.87	71.68	58.49	70.08
	+Contrastive (BT)†	57.60	72.14	62.25	71.49	71.75	77.05	67.83	68.59
	+SG-OPT†	64.29	76.36	68.48	80.10	76.60	78.14	67.97	73.13
	+SimCSE	72.86	83.99	75.62	84.77	81.80	81.98	71.26	78.90
	+DCLR (Ours)	**73.09**	**84.57**	**76.13**	**85.15**	**81.99**	**82.35**	**71.80**	**79.30**

Table 1: Sentence embedding performance on STS tasks (Spearman's correlation). The best performance and the second-best performance methods are denoted in bold and underlined fonts respectively. †: results from Kim et al. (2021); ‡: results from Gao et al. (2021); all other results are reproduced or reevaluated by ourselves.

CSE (Gao et al., 2021), we use 1,000,000 sentences randomly sampled from Wikipedia as the training corpus. During training, we train our models for 3 epoch with temperature $\tau = 0.05$ using an Adam optimizer (Kingma and Ba, 2015). For BERT-base and RoBERTa-base, the batch size is 128, the learning rate is 3e-5. For BERT-large and RoBERTa-large, the batch size is 256, the learning rate is 3e-5 and 1e-5, respectively. For the four backbone models, we set the instance weighting threshold ϕ as 0.9, 0.85, 0.9 and 0.85, respectively. For each batch, we generate $k \times batch\_size$ noise-based negatives as the shared negatives of all instance within it, and k is 1, 2.5, 4 and 5 for BERT-base, RoBERTa-base, BERT-large and RoBERTa-large, respectively. The standard variance of the noise-based negatives is 1, and we update the noise-based negatives four times with the learning rate of 1e-3. We evaluate the model every 150 steps on the development set of STS-B and SICK-R and keep the best checkpoint for evaluation on test sets.

5.2 Main Results

To verify the effectiveness of our framework on PLMs, we select BERT-base and RoBERTa-base as the base model. Table 1 shows the results of different methods on seven STS tasks.

Based on the results, we can find that the non-BERT methods (*i.e.,* GloVe and USE) mostly outperform native PLM representation based baselines (*i.e.,* CLS, Mean and First-Last AVG). The reason is that directly utilizing the PLM native representations is prone to be influenced by the anisotropy issue. Among non-BERT methods, USE outperforms Glove. A potential reason is that USE encodes the sentence using the Transformer model, which is more effective than simply averaging GloVe embeddings.

For other PLM-based approaches, first, we can see that flow and whitening achieve similar results and outperform the native representations based methods by a margin. These two methods adopt specific improvement strategies to refine the representations of PLMs. Second, approaches based on contrastive learning outperform the other baselines in most cases. Contrastive learning can enhance both the alignment between semantically related positive pairs and the uniformity of the representation space using negative samples, resulting in better sentence representations. Furthermore, SimCSE performs the best among all the baselines. It indicates that dropout is a more effective positive augmentation method than others since it rarely hurts the semantics of the sentence.

Finally, DCLR performs better than all the baselines in most settings, including the approaches based on contrastive learning. Since these methods mostly utilize randomly sampled negatives (*e.g.,* in-batch negatives) to learn the uniformity of all sentence representations, it may lead to sampling bias, such as false negatives and anisotropy representations. Different from these methods, our framework adopts an instance weighting method to punish false negatives and a gradient-based algorithm to generate noise-based negatives towards the nonuniform points. In this way, the sampling bias problem can be alleviated, and our model can better learn the uniformity to improve the quality of the sentence representations.

6 Experiment - Analysis and Extension

In this section, we continue to study the effectiveness of our proposed DCLR.

6.1 Debiased Contrastive Learning on Other Methods

Since our proposed DCLR is a general framework that mainly focuses on negative sampling for contrastive learning of unsupervised sentence representations, it can be applied to other methods that rely on different positive data augmentation strategies.

Model	STS-Avg.
BERT-base+Ours	**77.22**
w/o Noise-based Negatives	76.17
w/o Instance Weighting	76.31
BERT-base+Random Noise	75.22
BERT-base+Knowledge Distillation	75.05
BERT-base+Self Instance Weighting	73.93

Table 2: Ablation and variation studies of our approach on the test set of seven STS tasks.

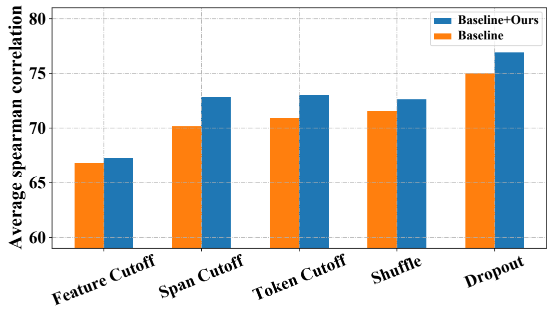


Figure 3: Performance comparison using different positive augmentation strategies on the test set of seven STS tasks.

Thus, in this part, we conduct experiments to examine whether our framework can bring improvements with the following positive data augmentation strategies: (1) *Token Shuffling* that randomly shuffles the order of the tokens in the input sequences; (2) *Feature/Token/Span Cutoff* (Yan et al., 2021) that randomly erases features/tokens/token spans in the input; (3) *Dropout* that is similar to SimCSE (Gao et al., 2021). Note that we only revise the negative sampling strategies to implement these variants of our DCLR.

As shown in Figure 3, our DCLR can boost the performance of all these augmentation strategies, it demonstrates the effectiveness of our framework with various augmentation strategies. Furthermore, the Dropout strategy leads to the best performance among all the variants. It indicates that dropout is a more effective approach to augment high-quality positives, and is also more appropriate for our approach.

6.2 Ablation Study

Our proposed DCLR incorporates an instance weighting method to punish false negatives and also utilizes noise-based negatives to improve the uniformity of the whole sentence representation space. To verify their effectiveness, we conduct an ablation study for each of the two components on seven STS tasks and report the average value

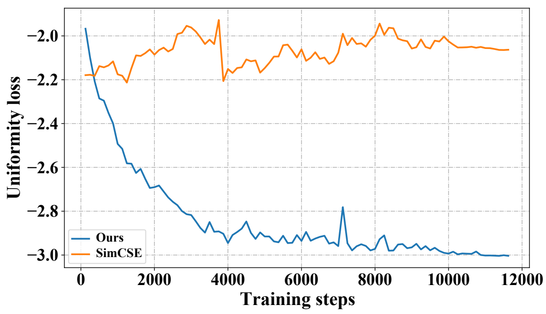


Figure 4: The uniformity loss of DCLR and SimCSE using BERT-base on the validation set of STS-B during training.

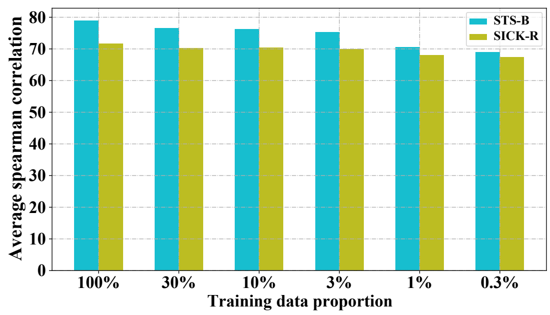
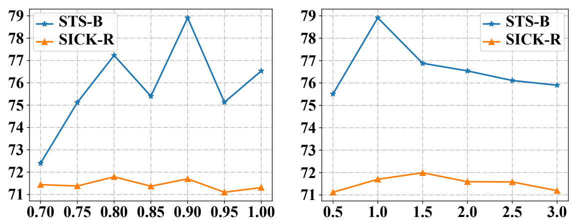


Figure 5: Performance tuning of our DCLR *w.r.t.* different amounts of training data.



(a) Weighting Threshold ϕ (b) Negative Proportion k

Figure 6: Performance tuning *w.r.t.* ϕ and k.

of the Spearman's correlation metric. As shown in Table 2, removing each component would lead to the performance degradation. It indicates that the instance weighting method and the noise-based negatives are both important in our framework. Besides, removing the instance weighting method results in a larger performance drop. The reason may be that the false negatives have a larger effect on sentence representation learning.

Besides, we prepare three variants for further comparison: (1) *Random Noise* directly generates noise-based negatives without the gradient-based optimization; (2) *Knowledge Distillation* (Hinton et al., 2015) utilizes SimCSE as the teacher model to distill knowledge into the student model during training; (3) *Self Instance Weighting* adopts the model itself as the complementary model to generate the weights. From Table 2, we can see that these variations don't perform as well as the original DCLR. These results indicate the proposed designs in Section 4 are more suitable for our DCLR framework.

6.3 Uniformity Analysis

Uniformity is a desirable characteristic for sentence representations, describing how well the representations are uniformly distributed. To validate the improvement of the uniformity of our framework, we compare the uniformity loss curves of DCLR and SimCSE using BERT-base during training.

Following SimCSE (Gao et al., 2021), we utilize the following function to evaluate the uniformity:

$$\ell_{uniform} \triangleq \log \underset{x_i, x_j \overset{i.i.d.}{\sim} p_{data}}{\mathbb{E}} e^{-2\|f(x_i)-f(x_j)\|^2},$$

where p_{data} is the distribution of all sentence representations, and a smaller value of this loss indicates a better uniformity. As shown in Figure 4, the uniformity loss of DCLR is much lower than that of SimCSE in almost the whole training process. Furthermore, we can see that the uniformity loss of DCLR decreases faster as training goes, while the one of SimCSE shows no significant decreasing trend. It might be because our DCLR samples noise-based negatives beyond the representation space, which can better improve the uniformity of sentence representations.

6.4 Performance under Few-shot Settings

To validate the reliability and the robustness of DCLR under the data scarcity scenarios, we conduct few-shot experiments using BERT-base as the backbone model. We train our model via different amounts of available training data from 100% to the extremely small size (*i.e.,* 0.3%). We report the results evaluated on STS-B and SICK-R tasks.

As shown in Figure 5, our approach achieves stable results under different proportions of the training data. Under the most extreme setting with 0.3% data proportion, the performance of our model drops by only 9 and 4 percent on STS-B and SICK-R, respectively. The results reveal the robustness and effectiveness of our approach under the data scarcity scenarios. Such characteristics are important in real-world application.

6.5 Hyper-parameters Analysis

For hyper-parameters analysis, we study the impact of instance weighting threshold ϕ and the proportion of noise-based negatives k. The ϕ is the threshold to punish false negatives, and k is the ratio of the noise-based negatives to the batch size. Both hyper-parameters are important in our framework. Concretely, we evaluate our model with varying values of ϕ and k on the STS-B and SICK-R tasks using the BERT-base model.

Weighting threshold. Figure 6(a) shows the influence of the instance weighting threshold ϕ. For the STS-B tasks, ϕ has a significant effect on the model performance. Too large or too small ϕ may lead to a performance drop. The reason is that a larger threshold cannot achieve effective punishment and a smaller one may cause misjudgment of true negatives. In contrast, the SICK-R is insensitive to the changes of ϕ. The reason may be that the problem of false negatives is not serious in this task.

Negative proportion. As shown in Figure 6(b), our DCLR performs better when the number of noise-based negatives is close to the batch size. Under these circumstances, the noise-based negatives are more capable to enhance the uniformity of the whole semantic space without hurting the alignment, which is key why DCLR works well.

7 Conclusion

In this paper, we proposed DCLR, a debiased contrastive learning framework for unsupervised sentence representation learning. Our core idea is to alleviate the sampling bias caused by the random negative sampling strategy. To achieve it, in our framework, we incorporated an instance weighting method to punish false negatives during training and generated noise-based negatives to alleviate the influence of anisotropy PLM-derived representation. Experimental results on seven STS tasks have shown that our approach outperforms several competitive baselines.

In the future, we will explore other approaches to reducing the bias in contrastive learning of sentence representations (*e.g.,* debiased pre-training). Besides, we will also consider to apply our method for multilingual or multimodal representation learning.

Ethical Consideration

In this section, we discuss the ethical considerations of this work from the following two aspects. First, for intellectual property protection, the code, data and pre-trained models adopted from previous works are granted for research-purpose usage. Second, since PLMs have been shown to capture certain biases from the data they have been pretrained on (Bender et al., 2021), there is a potential problem about biases that are from the use of PLMs in our approach. There are increasing efforts to address this problem in the community (Ross et al., 2020).

Acknowledgement

This work was partially supported by Beijing Natural Science Foundation under Grant No. 4222027, and National Natural Science Foundation of China under Grant No. 61872369, Beijing Outstanding Young Scientist Program under Grant No. BJJWZYJH012019100020098, the Outstanding Innovative Talents Cultivation Funded Programs 2021 and Public Computing Cloud, Renmin University of China. This work is also supported by Beijing Academy of Artificial Intelligence (BAAI). Xin Zhao is the corresponding author.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *NAACL-HLT*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *COLING*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *COLING*, pages 497–511.

Eneko Agirre, Daniel M. Cer, Mona T. Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *NAACL-HLT*, pages 385–393.

Eneko Agirre, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In **SEM*, pages 32–43.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Shuqing Bian, Wayne Xin Zhao, Kun Zhou, Jing Cai, Yancheng He, Cunxiang Yin, and Ji-Rong Wen. 2021. Contrastive curriculum learning for sequential user behavior modeling via data augmentation. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3737–3746. ACM.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *EMNLP*, pages 169–174.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *ACL*, pages 1–14.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *LREC*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. In *EMNLP-IJCNLP*, pages 55–65.

Hongchao Fang and Pengtao Xie. 2020. CERT: contrastive self-supervised learning for language understanding. *CoRR*, abs/2005.12766.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910. Association for Computational Linguistics.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *NAACL-HLT*, pages 1367–1377.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. Whiteningbert: An easy unsupervised sentence embedding approach. *CoRR*, abs/2104.01767.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *ACL*, pages 2177–2190.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *ACL*, pages 2528–2540.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *ICLR*.

Quoc V. Le and Tomás Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196.

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*, pages 9119–9130.

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. *CoRR*, abs/1907.11692.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *ICLR*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton Van Den Hengel. 2016. Less is more: zero-shot learning from online textual documents with noise suppression. In *CVPR*, pages 2249–2257.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. 2019. Adversarial robustness through local linearization. In *NeurIPS*, pages 13824–13833.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3980–3990.

Candace Ross, Boris Katz, and Andrei Barbu. 2020. Measuring social biases in grounded vision and language embeddings. *arXiv preprint arXiv:2002.08911*.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *CoRR*, abs/2103.15316.

Haipeng Sun, Rui Wang, Kehai Chen, Xugang Lu, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Robust unsupervised neural machine translation with adversarial denoising training. In *COLING*, pages 4239–4250.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP - Demos*, pages 38–45.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: contrastive learning for sentence representation. *CoRR*, abs/2012.15466.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *ACL/IJCNLP*, pages 5065–5075.

Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 1488–1496. ACM.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *ICLR*.