# Conditional Semantic Textual Similarity via Conditional Contrastive Learning

**Xinyue Liu[1], Zeyang Qin[1], Zeyu Wang[1], Wenxin Liang[1], Linlin Zong[1], Bo Xu[2]***,
[1]School of Software, Dalian University of Technology, China,
[2]School of Computer Science and Technology, Dalian University of Technology, China
{xyliu, wxliang, llzong, xubo}@dlut.edu.cn
{zyqin, zywang}@mail.dlut.edu.cn

## Abstract

Conditional semantic textual similarity (C-STS) assesses the similarity between pairs of sentence representations under different conditions. The current method encounters the over-estimation issue of positive and negative samples. Specifically, the similarity within positive samples is excessively high, while that within negative samples is excessively low. In this paper, we focus on the C-STS task and develop a conditional contrastive learning framework that constructs positive and negative samples from two perspectives, achieving the following primary objectives: (1) adaptive selection of the optimization direction for positive and negative samples to solve the over-estimation problem, (2) fully balance of the effects of hard and false negative samples. We validate the proposed method with five models based on bi-encoder and tri-encoder architectures, the results show that our proposed method achieves state-of-the-art performance. The code is available at https://github.com/qinzeyang0919/CCL.

## 1 Introduction

The semantic similarity task aims to bring similar texts closer together in the representation space and push dissimilar texts further apart, which serves as a common benchmark for evaluating the performance of general sentence representations (Agirre et al., 2012, 2014, 2015, 2016; Cer et al., 2017). However, a sentence typically encompasses multiple facets. Viewed from varying angles, the sentence should exhibit different degrees of similarity.

Deshpande et al. (2023) firstly extend the semantic similarity task to the Conditional Semantic Textual Similarity (C-STS) task by introducing conditional information, assessing the similarity between two sentences $(s_1, s_2)$ under different conditions. For example, in the context of *"The person's gender"*, the sentences *"A girl standing in a boat resting her arm on an elephant who is passing by"* and *"A woman sits on a brick platform while two elephants are in the distance on the grass"* exhibit a higher degree of similarity. In contrast, the similarity is comparatively lower in terms of *"The person's age"*. Compared to fuzzy semantic similarity tasks, C-STS focuses on fine-grained textual representations, endowing two sentences with varying degrees of similarity under different conditions. Condition-based semantic representations can be utilized in tasks such as fine-grained retrieval (Kamoi et al., 2023; Asai et al., 2023), large language model text attribution (Rashkin et al., 2023; Chen et al., 2024), and other related tasks.

Deshpande et al. (2023) propose the first method for C-STS which we name QuMSE. QuMSE combines Quad loss and MSE loss for joint optimization. Nevertheless, QuMSE has the problem of over-estimation. As shown in Table 1, compared with the labels by human, the prediction results of positive samples (*under positive conditions* $c_{pos}$) are too high, and the prediction results of negative samples (*under negative conditions* $c_{pos}$) are too low.

We analyze the over-estimation problem of QuMSE starting from the definition of Quad loss as follows :

$$\begin{aligned} &\text{Quad}\left(p_1, p_2, n_1, n_2\right) = \\ &max\left(\lambda + \cos\left(n_1, n_2\right) - \cos\left(p_1, p_2\right), 0\right) \end{aligned} \tag{1}$$

where $\cos\left(p_1, p_2\right)$ and $\cos\left(n_1, n_2\right)$ denote the similarity between positive and negative samples, respectively, the value of $\lambda$ determines the degree of optimization for positive and negative samples, that is, the degree of difference between positive and negative samples. In Deshpande et al. (2023), $\lambda = 1$. In this case, Quad loss stops optimizing only when $\cos\left(p_1, p_2\right) - \cos\left(n_1, n_2\right) \geq 1$. For the example depicted in Table 1, when $l_{pos} - l_{neg} < 1$, there exists a misguided effect of Quad loss, leading to an over-estimation issue. It is easy to un-

*Corresponding author.



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 4548–4560
January 19–24, 2025. 

| Sentence 1 | Sentence 2 | Condition | QuMSE | CCL(ours) | Label |
|---|---|---|---|---|---|
| A large room with tile flooring and four pieces of furniture on a Persian style carpet. | A room decorated with multiple rocking chairs and lots of art on the walls. | $c_{pos}$ : The room's function. | 0.99↑↑ | 0.84↑ | 0.75 |
| | | $c_{neg}$ : The art's position. | 0.29↓↓ | 0.45↓ | 0.5 |
| The rear of a bus and a traffic light are pictured in a downtown night scene. | This is a picture of a busy downtown cross walk with several cars in the flow of traffic. | $c_{pos}$ :The location of the scenery. | 0.90↑↑ | 0.68↑ | 0.5 |
| | | $c_{neg}$ :The vehicles in sight. | 0.01↓↓ | 0.36↓ | 0.5 |
| A cart is loaded with luggage at a station while an attendant sits on a bench. | A woman is using a cellphone while sitting at a group of benches with luggage nearby. | $c_{pos}$ :The type of carrier. | 0.86↑↑ | 0.73↑ | 0.75 |
| | | $c_{neg}$ :The person's attention. | 0.21↓↓ | 0.30↓ | 0.5 |

Table 1: Comparative Examples of CCL (ours) and QuMSE. The prediction similarity scores are obtained by fine-tuning the $\text{SimCSE}_{\text{base}}$ model based on the bi-encoder architecture. The similarity scores with underlines represent the human labels. ↑↑ and ↓↓ represent large deviations from human labels. ↑ and ↓ indicate relatively smaller deviations from human labels. The statistical analysis results are in Appendix A.

derstand that the differences in similarity between sentence pairs under varying conditions are not a fixed value. Therefore, how to adaptively select the optimization direction of the samples is a crucial issue.

Contrastive learning can enhance representation performance by constructing positive and negative samples through diverse data augmentation or selection (Chen et al., 2020; Miao et al., 2023). However, constructing effective positive and negative samples in conditional sentence representation learning remains unexplored. Particularly in constructing negative samples, the balance between hard negative samples and false negative samples, to some extent, limits the performance of contrastive learning (Kalantidis et al., 2020; Zhou et al., 2022). Therefore, it is necessary to explore how to utilize limited information to balance the optimization strength of controversial samples.

In this paper, we propose a **C**onditional **C**ontrastive **L**earning (**CCL**) framework to address the issues mentioned above. We first introduce **W**eighted **A**daptive **C**ontrastive **L**oss (**W-ACL**). **W-ACL** utilizes similarity score label information as truncation coefficients, adaptively selecting the optimization directions for positive and negative samples. The adaptive contrastive loss selects the same sentence pairs as positive and negative samples under different conditions, fully leveraging the similarity score label information of positive and negative samples in the C-STS dataset to control the relative difference between positive and negative samples at the same level as the human labels.

We further enhance W-ACL by introducing **B**alanced **C**ontrastive **L**oss (**BCL**), specifically, by considering similarity contrast scores between the same sentences in the same conditions *(positive samples)* and between distinct sentences in the same conditions *(negative samples)*, as well as in different conditions *(negative samples)*. Unsupervised contrastive learning necessitates the construction of entirely similar and dissimilar positive and negative samples. Simply using distinct sentences under the same conditions as negative samples can lead to instances with high label similarity scores becoming false negative samples (Chuang et al., 2020). Such samples can also be regarded as hard negative samples. We achieve a balance in optimizing false negative and hard negative samples by utilizing similarity score labels.

CCL optimizes the aforementioned losses jointly with MSE loss and is compatible with various pretrained models. Table 1 displays the prediction outcomes of the model trained with the CCL framework. It can be seen that the prediction results of CCL are much closer to the human labels than those of QuMSE.

In summary, this paper makes the following contributions:

- We propose W-ACL, to the best of our knowledge, it is the first work that leverages label information to determine the optimization direction for positive and negative sample pairs.

- We introduce BCL into the C-STS task as auxiliary training, employing similarity score labels to achieve a balance in optimizing controversial negative samples.
- We conduct extensive experiments, showing that our proposed CCL method achieves state-of-the-art performance on the C-STS task based on bi-encoder architecture.

## 2 Related work

### 2.1 Conditional Semantic Textual Similarity

The conditional Semantic Textual Similarity(C-STS) task involves two sentences, a condition, and a similarity score label. Conditions can be considered prompt templates for sentences (Asai et al., 2023; Petroni et al., 2019; Jiang et al., 2020). Based on different prompt templates, sentences exhibit distinct representations. Deshpande et al. (2023) have proposed three architectures: cross-encoder, bi-encoder, and tri-encoder. Among these, the bi-encoder architecture inputs a single sentence and condition together into the encoder to obtain representation, enabling the capture of rich contextual information and leading to the best performance among the three architectures. In the tri-encoder architecture, the hypernetwork (Ha et al., 2017) is introduced to project sentences into distinct subspaces according to different conditions (Yoo et al., 2024). Even the current large language models (Achiam et al., 2023) struggle to perform well on C-STS tasks (Deshpande et al., 2023). Therefore, the C-STS task is still worth exploring.

### 2.2 Contrastive Learning

Contrastive learning has been widely applied in various tasks (Liu and Chen, 2024; Wu et al., 2021; Radford et al., 2021), excelling notably in representation learning (Chen et al., 2020; He et al., 2020; Gao et al., 2021). In sentence representation learning, contrastive learning has successfully addressed the issue of anisotropy (Ethayarajh, 2019; Li et al., 2020a) in text representations output by pre-trained models (Gao et al., 2021, 2019; Li et al., 2020b). Existing contrastive learning methods have delved deeply into three key components of contrastive learning: the construction of positive samples (Wu et al., 2022; Kim et al., 2021), the selection of negative samples (Miao et al., 2023), and the design of loss functions (Chuang et al., 2022; Wang and Dou, 2023; Liu et al., 2023). Other studies have addressed issues such as tackling the problem of false negative samples in unsupervised contrastive learning (Zhou et al., 2022; Huynh et al., 2022; Chuang et al., 2020) and devising strategies to construct effective hard negative samples to enhance contrastive learning frameworks (Kalantidis et al., 2020; Robinson et al., 2021).

Our approach differs from the aforementioned methods by employing precise similarity score labels for contrastive losses in the C-STS task, enabling controlled optimization of positive and negative samples while balancing the influence of controversial negatives.

## 3 The Proposed Method

### 3.1 Preliminaries

We investigate the problem of conditional semantic text similarity, considering a batch of samples $D = \{d^i_{pos}, d^i_{neg}\}^N_{i=1}$ , where $N$ denotes the number of samples in the batch. The $d^i_{pos}$ and $d^i_{neg}$ are respectively composed of quadruplets $\langle s^i_1, s^i_2, c^i_{pos}, l^i_{pos}\rangle$ and $\langle s^i_1, s^i_2, c^i_{neg}, l^i_{neg}\rangle$. $s^i_1$ and $s^i_2$ denote two distinct sentences, whereas $c^i_{pos}$ and $c^i_{neg}$ signify two distinct conditions. $l^i_{pos}$ and $l^i_{neg}$ measure the similarity scores for sentences $s^i_1$ and $s^i_2$ under conditions $c^i_{pos}$ and $c^i_{neg}$, respectively, ensuring $l^i_{pos} \geq l^i_{neg}$, as illustrated in Table 1. The objective of the task is to feed sentences and conditions into the model, assessing the similarity between the two sentences when controlled by the conditions.

### 3.2 Conditional Contrastive Learning Framework

The construction of the conditional contrastive learning framework is based on bi-encoder and tri-encoder architectures and is optimized by our proposed loss function. The approach to obtaining conditional similarity involves calculating the cosine similarity between the representations of conditional sentences.

$$similarity = \cos(h_1, h_2), \tag{2}$$

where $h_1$ and $h_2$ represent the conditional representations of the two sentences, respectively.

#### 3.2.1 Model structure

The model structure is based on two architectures: the bi-encoder and the tri-encoder architectures.

The **bi-encoder** architecture is commonly used in semantic similarity tasks (Wu et al., 2022;

Reimers and Gurevych, 2019). According to Deshpande et al. (2023), we concatenate the sentence with the condition and take the [CLS] output of the model's last layer as the conditional sentence representation. The concatenation of sentence $s_1$ with condition $c$ and sentence $s_2$ with condition $c$ are fed into the bi-encoder $f$. The cosine similarity score between the conditional sentence representations $h_1 = f(s_1, c)$ and $h_2 = f(s_2, c)$, output by the bi-encoder, is used as the final similarity score.

The **tri-encoder** architecture encodes sentences and conditions independently, situating the feature fusion process within the embedding space rather than within the encoder. $h_1 = I(f(s_1), f(c))$ and $h_2 = I(f(s_2), f(c))$ represent the calculation process of conditional sentence representations. Here, $I$ denotes the feature fusion function, following Deshpande et al. (2023), we simply employ the Hadamard product.

#### 3.2.2 Optimization objective

The conditional contrastive learning framework utilizes our proposed **W**eighted **A**daptive **C**ontrastive **L**oss(**W-ACL**) and **B**alanced **C**ontrastive **L**oss (**BCL**) for joint optimization with MSE loss.

$$\begin{aligned} l_{CCL} =& \lambda_1 l_{\text{W-ACL}} + \lambda_2 l_{\text{BCL}} \\ &+ \lambda_3 l_{\text{MSE}} + \lambda_4 l_{\text{c-MSE}}, \end{aligned} \tag{3}$$

where, $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are hyperparameters used to balance the training process. $l_{\text{MSE}}$ and $l_{\text{C-MSE}}$ finely constrain similarity scores in the embedding and contrastive learning spaces, respectively.

### 3.3 Weighted Adaptive Contrastive Loss

The W-ACL consists of two components: the Forward Directional Loss (FDL), which alleviates the over-estimation issues associated with contrastive learning, and the Reverse Direction Loss (RDL), which effectively mitigates the waste of sample information.

#### 3.3.1 Forward Directional Loss

Considering the Quad loss (Deshpande et al., 2023),

$$\begin{aligned} &\text{Quad}(h_1^{pos}, h_2^{pos}, h_1^{neg}, h_2^{neg}) = \\ &\max\left(\lambda + \cos(h_1^{neg}, h_2^{neg}) - \cos(h_1^{pos}, h_2^{pos}), 0\right), \end{aligned} \tag{4}$$

where $h_1^{neg} = f(s_1, c_{neg})$, $h_2^{neg} = f(s_2, c_{neg})$, $h_1^{pos} = f(s_1, c_{pos})$, $h_2^{pos} = f(s_2, c_{pos})$.

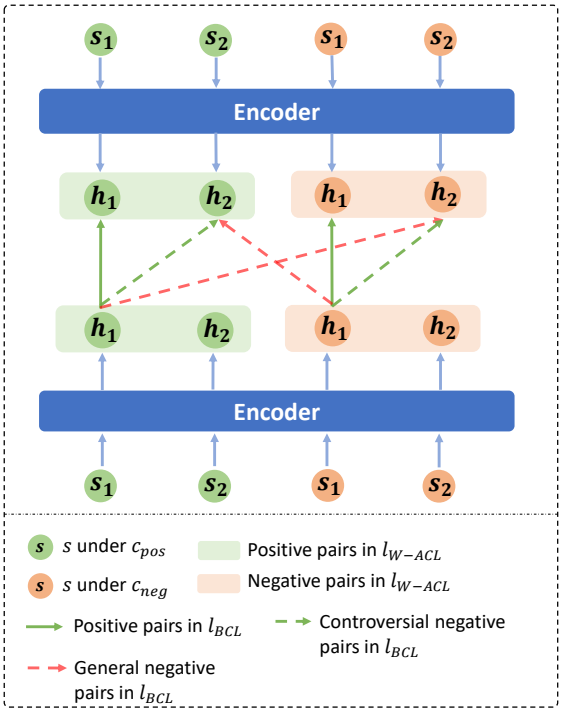


Figure 1: illustrating the construction of positive and negative samples within the CCL framework.

Clearly, optimizing the Quad loss will increase $\cos(h_1^{pos}, h_2^{pos})$ and decrease $\cos(h_1^{neg}, h_2^{neg})$. Given the condition $l_{pos} \geq l_{neg}$ in the previously mentioned dataset, the direction of Quad loss optimization is correct.

When will the optimization stop? This depends on the value of $\lambda$, which is set to 1 according to Deshpande et al. (2023). Thus, optimization of the Quad loss will cease when $\cos(h_1^{pos}, h_2^{pos}) - \cos(h_1^{neg}, h_2^{neg}) \geq 1$. If only based on the condition $l_{pos} \geq l_{neg}$, this is undoubtedly correct. Constructing confident positive and negative samples is sufficient, just like using contrastive loss in datasets without explicit similarity scores (Chen et al., 2020; Gao et al., 2021).

However, considering the semantic similarity dataset, we can ascertain the specific similarity values of $l_{pos}^i$ and $l_{neg}^i$. Therefore, when $\cos(h_1^{pos}, h_2^{pos}) - \cos(h_1^{neg}, h_2^{neg}) > l_{pos} - l_{neg}$, the optimization objective of the loss function becomes inaccurate. According to the above, the Quad loss can be truncated by using the values of $l_{pos}$ and $l_{neg}$ as truncation factors to mask incorrect gradients, resulting in the forward direction loss

$l_{\text{FDL}}$,

$$l_{\text{FDL}}(h_1^{pos}, h_2^{pos}, h_1^{neg}, h_2^{neg}) = \max(l_{pos} - l_{neg} + \cos(h_1^{neg}, h_2^{neg}) - \cos(h_1^{pos}, h_2^{pos}), 0). \tag{5}$$

#### 3.3.2 Reverse Direction Loss

Although the truncation term in the above loss function can ensure the correctness of gradient information, it wastes a portion of sample information.

Considering the condition $\cos(h_1^{pos}, h_2^{pos}) - \cos(h_1^{neg}, h_2^{neg}) > l_{pos} - l_{neg}$, the loss function $l_{FDL}$ takes the value of 0. In this situation, the model does not receive gradient information. We aim to utilize the gradient information of this subset at this juncture; the optimization direction for these samples should decrease the cosine similarity of $(h_1^{pos}, h_2^{pos})$ and increase the cosine similarity of $(h_1^{neg}, h_2^{neg})$. When considering $\cos(h_1^{pos}, h_2^{pos}) - \cos(h_1^{neg}, h_2^{neg}) < l_{pos} - l_{neg}$, we face a problem similar to Quad loss: incorrect gradient information. We add truncation factors to ensure the correctness of the optimization direction for positive and negative samples. Therefore, the reverse direction loss $l_{\text{RDL}}$ is designed as follows,

$$l_{\text{RDL}}(h_1^{pos}, h_2^{pos}, h_1^{neg}, h_2^{neg}) = \max(l_{neg} - l_{pos} - \cos(h_1^{neg}, h_2^{neg}) + \cos(h_1^{pos}, h_2^{pos}), 0). \tag{6}$$

#### 3.3.3 Weighted Adaptive Contrastive Loss

We define $l_{\text{ACL}}$ by combining $l_{\text{FDL}}$ and $l_{\text{RDL}}$,

$$\begin{aligned} l_{\text{ACL}} =& l_{\text{FDL}}(h_1^{pos}, h_2^{pos}, h_1^{neg}, h_2^{neg}) \\ &+ l_{\text{RDL}}(h_1^{pos}, h_2^{pos}, h_1^{neg}, h_2^{neg}). \end{aligned} \tag{7}$$

Considering that the truncation terms in the $l_{\text{FDL}}$ and $l_{\text{RDL}}$ loss functions are both $l_{pos}^i - l_{neg}^i$, the above loss function can be rewritten as,

$$\begin{aligned} l_{\text{ACL}} =& |l_{pos} - l_{neg} + \\ & \cos(h_1^{neg}, h_2^{neg}) - \cos(h_1^{pos}, h_2^{pos})|. \end{aligned} \tag{8}$$

Overall, $l_{\text{ACL}}$ constructs positive and negative samples based on the similarity of the same sentence pair under different conditions. $l_{\text{ACL}}$ can control the optimization direction of positive and negative samples to make $\cos(h_1^{neg}, h_2^{neg}) - \cos(h_1^{pos}, h_2^{pos})$ approach $l_{pos} - l_{neg}$.

Furthermore, we can regard $l_{pos} - l_{neg}$ as the degree of difference between positive and negative samples. A larger value of $l_{pos} - l_{neg}$ indicates a greater confidence in the positive and negative samples. Hence, we should assign greater optimization weight to samples with significant differences. By assigning different weights to samples with different confidence levels based on $l_{pos}^i - l_{neg}^i$,

$$\begin{aligned} l_{\text{W-ACL}} =& (l_{pos} - l_{neg}) \cdot |l_{pos} - l_{neg} + \\ & \cos(h_1^{neg}, h_2^{neg}) - \cos(h_1^{pos}, h_2^{pos})|. \end{aligned} \tag{9}$$

### 3.4 Balanced Contrastive Loss

#### 3.4.1 Definition of BCL

Many studies have demonstrated the efficacy of contrastive learning in sentence representation (Gao et al., 2021; Wu et al., 2022; Wang and Dou, 2023). However, the exploration of contrastive learning in conditional sentence representation remains insufficient. The critical step in contrastive learning is the creation of positive and negative samples. A straightforward construction strategy is to set $h_1^{pos}$ and $h_2^{pos}$ as positive sample pairs. However, this confronts the same issue as previously discussed, wherein the positive samples are not guaranteed to be sufficiently positive samples, that is, $\exists l_{pos}^i < 1$.

Therefore, we utilize augmented samples of the same sentences under the same conditions as positive samples and different sentences under distinct conditions as negative samples to construct positive-negative sample pairs, as illustrated in Figure 1. The loss function is formulated as follows,

$$l_{\text{BCL}} = -log \sum_{i=1}^{N} \frac{e^{cos_{i,i}/\tau}}{e^{cos_{i,i}/\tau} + \sum_{j=1}^{N} \alpha \cdot e^{cos_{i,j}/\tau}}, \tag{10}$$

where $\tau$ represents the temperature coefficient, $cos_{i,i} = cos\left(g\left(f\left(s_1^i, c^i\right)\right), g\left(f'\left(s_1^i, c^i\right)\right)\right)$ and $cos_{i,j} = cos\left(g\left(f\left(s_1^i, c^i\right)\right), g\left(f\left(s_2^j, c^j\right)\right)\right)$ represent the cosine similarity between positive and negative samples, respectively. $f\left(s_1^i, c^i\right)$ and $f'\left(s_1^i, c^i\right)$ denote the representations obtained for the same sentence and condition by the encoder with dropout (Srivastava et al., 2014), and therefore can be regarded as sufficiently confident positive samples (Gao et al., 2021). When $i \neq j$, the representations $f\left(s_1^i, c^i\right)$ and $f\left(s_2^j, c^j\right)$ denote two distinct sentences under different conditions. When $i = j$, there may be hard negative or false negative

| Encoding | Model | lr. | $\tau$ | $\sigma$ |
|---|---|---|---|---|
| Bi-encoder | $\text{RoBERTa}_{\text{base}}$ | 3e-5 | 10 | 0.5 |
| | $\text{RoBERTa}_{\text{large}}$ | 1e-5 | 3 | 0.5 |
| | $\text{DiffCSE}_{\text{base}}$ | 3e-5 | 8 | 1 |
| | $\text{SimCSE}_{\text{base}}$ | 3e-5 | 3 | 0.75 |
| | $\text{SimCSE}_{\text{large}}$ | 1e-5 | 4 | 0.25 |
| Tri-encoder | $\text{RoBERTa}_{\text{base}}$ | 3e-5 | 5 | 0.5 |
| | $\text{RoBERTa}_{\text{large}}$ | 1e-5 | 8 | 1 |
| | $\text{DiffCSE}_{\text{base}}$ | 3e-5 | 9 | 0.75 |
| | $\text{SimCSE}_{\text{base}}$ | 3e-5 | 7 | 1 |
| | $\text{SimCSE}_{\text{large}}$ | 1e-5 | 7 | 0.5 |

Table 2: Optimal hyperparameters of the model on the C-STS validation dataset.

samples. Thus, we weight the negative samples based on the human labels.

$$\alpha = \begin{cases} 0 & i = j, l \geq \sigma \\ 1 & i \neq j \\ 1 - l & i = j, l < \sigma \end{cases} \tag{11}$$

### 3.4.2 Analysis of BCL

We analyze the role of weight terms in BCL from the perspective of gradients. According to Wang and Liu (2021), we analyze the impact of weight similarly. The BCL considering only one pair of positive samples is as follows,

$$l_{\text{BCL}}(x_i) = -log \frac{e^{cos_{i,i}/\tau}}{e^{cos_{i,i}/\tau} + \sum_{j=1}^{N} \alpha \cdot e^{cos_{i,j}/\tau}}. \tag{12}$$

We calculate the gradients of $l_{\text{BCL}}$ for positive sample similarity $\cos_{i,i}$ and negative sample similarity $\cos_{i,j}$, respectively,

$$\frac{\partial l_{\text{BCL}}(x_i)}{\partial cos_{ii}} = -\frac{1}{\tau} \cdot \frac{\sum_{j=1}^{N} \alpha \cdot e^{cos_{i,j}/\tau}}{e^{cos_{i,i}/\tau} + \sum_{j=1}^{N} \alpha \cdot e^{cos_{i,j}/\tau}}, \tag{13}$$

$$\frac{\partial l_{\text{BCL}}(x_i)}{\partial cos_{ij}} = \frac{1}{\tau} \cdot \frac{\alpha \cdot e^{cos_{i,j}/\tau}}{e^{cos_{i,i}/\tau} + \sum_{j=1}^{N} \alpha \cdot e^{cos_{i,j}/\tau}}, \tag{14}$$

After adding weight terms, we can still obtain that the gradient of positive samples is equal to the sum of the gradients of all negative samples (Wang and Liu, 2021), that is,

$$\sum_{j=1}^{N} \left| \frac{\partial l_{\text{BCL}}(x_i)}{\partial cos_{ij}} \right| \Big/ \left| \frac{\partial l_{\text{BCL}}(x_i)}{\partial cos_{ii}} \right| = 1 \tag{15}$$

The gradient can be regarded as the degree of optimization, and $\sigma$ controls the degree of optimization for different negative samples. We divide negative samples into three categories for processing: false negative samples, hard negative samples, and general negative samples, as shown in Eq 11.

For false negative samples (satisfying $i = j, l \geq \sigma$), we exclude them from the optimization objective, that is, $\frac{\partial l_{\text{BCL}}(x_i)}{\partial cos_{ij}} = 0$. Since general negative samples (satisfying $i \neq j$) are different sentences under different conditions, we unify their optimization strength to 1. For different sentences under the same conditions, their similarity should be lower than that of the same sentence under the same conditions, so they are treated as hard negative samples. The more similar the sample pairs, the smaller the optimization strength should be to ensure that the similarity of hard negative samples is between positive samples and general negative samples. Therefore, it is set to a value $1 - l$ that is closely related to the label information.

| Encoding | Model | Spear. ↑ |
|---|---|---|
| LLM† | $\text{FLan-T5}_{\text{XXL}}$ | 30.6 |
| | $\text{Tk-}Instruct_{\text{11B}}$ | 21.9 |
| | GPT-3.5 | 16.6 |
| | GPT-4 | 43.6 |
| Bi-encoder‡ (QuMSE) | $\text{RoBERTa}_{\text{base}}$ | 28.1 |
| | $\text{RoBERTa}_{\text{large}}$ | 27.4 |
| | $\text{DiffCSE}_{\text{base}}$ | 43.4 |
| | $\text{SimCSE}_{\text{base}}$ | 44.8 |
| | $\text{SimCSE}_{\text{large}}$ | 47.5 |
| Bi-encoder‡ (CCL) | $\text{RoBERTa}_{\text{base}}$ | **43.8**(+15.7) |
| | $\text{RoBERTa}_{\text{large}}$ | **46.3**(+18.9) |
| | $\text{DiffCSE}_{\text{base}}$ | **45.3**(+1.9) |
| | $\text{SimCSE}_{\text{base}}$ | **46.8**(+2.0) |
| | $\text{SimCSE}_{\text{large}}$ | **48.1**(+0.6) |
| Tri-encoder‡ (QuMSE) | $\text{RoBERTa}_{\text{base}}$ | 28.0 |
| | $\text{RoBERTa}_{\text{large}}$ | 20.3 |
| | $\text{DiffCSE}_{\text{base}}$ | 28.9 |
| | $\text{SimCSE}_{\text{base}}$ | 31.5 |
| | $\text{SimCSE}_{\text{large}}$ | 35.3 |
| Tri-encoder‡ (CCL) | $\text{RoBERTa}_{\text{base}}$ | **31.4**(+3.4) |
| | $\text{RoBERTa}_{\text{large}}$ | **30.3**(+10.0) |
| | $\text{DiffCSE}_{\text{base}}$ | **33.4**(+4.5) |
| | $\text{SimCSE}_{\text{base}}$ | **34.6**(+3.1) |
| | $\text{SimCSE}_{\text{large}}$ | **35.6**(+0.3) |

Table 3: We report the Spearman correlation of the model on the test split. Models with † indicate that we directly report the scores from Deshpande et al. (2023), while models with ‡ indicate models with conditional contrastive learning framework. Bold font indicates the optimal results.

| Encoding | Model | Spear. ↑ |
|---|---|---|
| Bi-encoder (CCL) | $\text{RoBERTa}_{\text{base}}$ | **45.8** |
| | $w/o\ l_{\text{W-ACL}}$ | 42.9 |
| | $w/o\ l_{\text{BCL}}$ | 43.1 |
| | $w/o\ l_{\text{MSE}}$ | 44.8 |
| | $w/o\ l_{\text{C-MSE}}$ | 43.7 |
| Bi-encoder (CCL) | $\text{SimCSE}_{\text{base}}$ | **46.9** |
| | $w/o\ l_{\text{W-ACL}}$ | 44.3 |
| | $w/o\ l_{\text{BCL}}$ | 46.3 |
| | $w/o\ l_{\text{MSE}}$ | 45.8 |
| | $w/o\ l_{\text{C-MSE}}$ | 46.6 |
| Tri-encoder (CCL) | $\text{RoBERTa}_{\text{base}}$ | **31.4** |
| | $w/o\ l_{\text{W-ACL}}$ | 22.3 |
| | $w/o\ l_{\text{BCL}}$ | 29.3 |
| | $w/o\ l_{\text{MSE}}$ | 21.8 |
| | $w/o\ l_{\text{C-MSE}}$ | 28.0 |
| Tri-encoder (CCL) | $\text{SimCSE}_{\text{base}}$ | 34.8 |
| | $w/o\ l_{\text{W-ACL}}$ | 30.8 |
| | $w/o\ l_{\text{BCL}}$ | 33.9 |
| | $w/o\ l_{\text{MSE}}$ | 32.9 |
| | $w/o\ l_{\text{C-MSE}}$ | **35.4** |

Table 4: Ablation study on loss functions in the conditional contrastive learning framework (CCL), with results obtained on the validation data. We use $w/o$ as the representation of $without$. We display the best results in bold.

| Encoding | Model | $\sigma$ | Spear. ↑ |
|---|---|---|---|
| Bi-encoder (CCL) | $\text{RoBERTa}_{\text{base}}$ | 0 | 44.3 |
| | | 0.25 | 45.4 |
| | | **0.5** | **45.8** |
| | | 0.75 | 44.8 |
| | | 1 | 44.7 |
| Bi-encoder (CCL) | $\text{SimCSE}_{\text{base}}$ | 0 | 46.0 |
| | | 0.25 | 45.8 |
| | | 0.5 | 46.4 |
| | | **0.75** | **46.9** |
| | | 1 | 46.3 |
| Tri-encoder (CCL) | $\text{RoBERTa}_{\text{base}}$ | 0 | 27.9 |
| | | 0.25 | 27.7 |
| | | **0.5** | **31.4** |
| | | 0.75 | 27.1 |
| | | 1 | 26.2 |
| Tri-encoder (CCL) | $\text{SimCSE}_{\text{base}}$ | 0 | 34.5 |
| | | 0.25 | 33.6 |
| | | 0.5 | 34.5 |
| | | 0.75 | 34.1 |
| | | **1** | **34.8** |

Table 5: Analysis of different handling approaches for controversial negative samples. We evaluate the impact of different $\sigma$ values in CCL on C-STS validation data. The results in bold are the best.

## 4 Experiment

### 4.1 Settings

All experiments are conducted on one NVIDIA TITAN RTX GPU. The optimal hyperparameters of the model are presented in Table 2. We simply set the balancing factors in Equation 3, $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, to 1, ensuring that each loss function has an equal infuence. According to Gao et al. (2021), we set the random seeds for all experiments to 42. After each epoch of training is completed, the model's performance is validated on the validation set. We save the model with the best performance on the validation set as the final result. Further training details can be found in Appendix A.

### 4.2 Baselines and Basic Models

We compare our proposed conditional contrastive learning framework CCL with the current method QuMSE (Deshpande et al., 2023) by applying them on five models across bi- and tri-encoder architectures, including $\text{RoBERTa}_{\text{base}}$ (Zhuang et al., 2021), $\text{RoBERTa}_{\text{large}}$ (Zhuang et al., 2021), as well as $\text{SimCSE}_{\text{base}}$ (Gao et al., 2021), $\text{SimCSE}_{\text{large}}$ (Gao et al., 2021) and $\text{DiffCSE}_{\text{base}}$ (Chuang et al., 2022) models. We also conducted comparisons with current large models, including Flan-T5 (Chung et al., 2024), Tk-INSTRUCT (Wang et al., 2022), GPT-3.5 (OpenAI, 2022), and GPT-4 (Achiam et al., 2023).

We follow (Deshpande et al., 2023) by employing Spearman correlation and Pearson correlation as the evaluation metrics, and the larger score means the higher performance. The Spearman results are reported in the main body of the paper and the Pearson results are included in the Appendix.

### 4.3 Main Results

The main results on the C-STS dataset are summarized in Table 3. CCL outperforms QuMSE when applied on all the models, it achieves the SOTA performance on the bi-encoder architecture and demonstrates the broad applicability.

CCL achieves absolute improvements over QuMSE by 15.7 and 18.9 points on $\text{RoBERTa}_{\text{base}}$ and $\text{RoBERTa}_{\text{large}}$ models based on the bi-encoder architecture. On $\text{SimCSE}_{\text{base}}$ and $\text{DiffCSE}_{\text{base}}$ that have been fine-tuned by contrastive learning, CCL achieves improvements of more than one point.

CCL achieves the best result on $\text{SimCSE}_{\text{large}}$ based on the bi-encoder architecture, yielding the Spearman correlation of 48.1. This score surpasses the best large language model, GPT-4, by 4.7 points.

### 4.4 Ablation Study

To investigate the impact of different losses in CCL, we conduct a set of ablation studies by removing $l_{\text{W-ACL}}$, $l_{\text{BCL}}$, $l_{\text{MSE}}$ and $l_{\text{C-MSE}}$ from Equation 3. We conduct ablation experiments on two typical models, $\text{RoBERTa}_{\text{base}}$ and $\text{SimCSE}_{\text{base}}$ across the bi- and tri-encoder architectures.

Table 4 presents the influence of each loss function on conditional semantic representations. Removing $l_{\text{W-ACL}}$ or $l_{\text{BCL}}$ from the conditional contrastive learning framework (CCL) can decrease the model's performance on the C-STS validation set. This is because that $l_{\text{W-ACL}}$ and $l_{\text{BCL}}$ construct positive and negative samples from different perspectives, and can promote the model to generate sentence embeddings that fully consider the similarity difference between samples from a complementary perspective. This explains why the model's performance declines when one of the contrastive losses is removed.

The RoBERTa model can achieve higher benefits from the two MSE losses. However, for the SimCSE model, the benefits obtained from $l_{\text{C-MSE}}$ are minimal or even detrimental. This may be attributed to the fact that SimCSE has already been fine-tuned by contrastive learning and possesses strong representational capabilities, thereby eliminating the necessity for further constraining the model within the contrastive space. We present the ablation results on the Pearson correlation coefficient in Appendix C.2.

### 4.5 Hyperparameter study

#### 4.5.1 Controversial negative samples

Table 5 demonstrates the impact of different treatments for false negative samples and hard negative samples in the CCL framework. We can find that the optimal effect cannot be achieved without controversial negative samples ($\sigma = 0$). All the models achieve optimal performance by eliminating false negative samples and weighting hard negative samples based on the ground truth ($\sigma \neq 0$). The model benefits from an appropriate $\sigma$ to strike a reasonable balance between false and hard negative samples. In Appendix C.3, we illustrate the influence of $\sigma$ on the Pearson correlation of CCL.

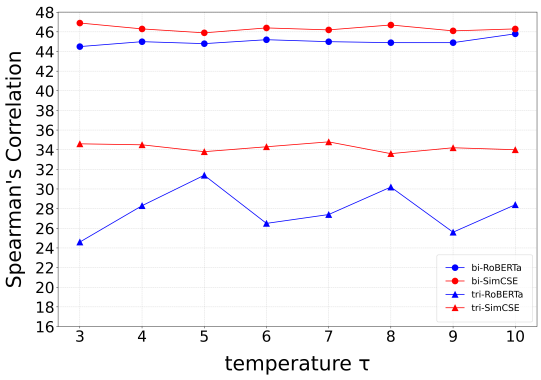


Figure 2: Analysis of temperature coefficient $\tau$ in $l_{\text{BCL}}$.

#### 4.5.2 Temperature Coefficient

We analyze the influence of temperature coefficient in $l_{\text{BCL}}$ and find that the model based on bi-encoder architecture has a certain robustness to temperature coefficient for Spearman correlation, which remains at a high level, as shown in Figure 2. For the Roberta model based on tri-encoder architecture, the model's performance fluctuates significantly with the temperature coefficient. This may be attributed to the inherent limitations of Roberta's representation space and the fact that the Hadamard product is not an effective feature fusion strategy. Pearson's results are detailed in Appendix C.4

## 5 Conclusion

In this paper, we have proposed a conditional contrastive learning framework (CCL) for the C-STS task, which encompasses four loss functions, including two traditional losses MSE and C-MSE, and two novel losses W-ACL and BCL specially designed for the conditional semantic text similarity task. W-ACL and BCL are constructed from two complementary perspectives to define positive and negative samples, enabling adaptive selection of the optimization direction for positive and negative samples while also balancing the influence of controversial negative samples. The proposed framework achieves state-of-the-art performance across five basic models.

## 6 Limitations

The current limitation of our work lies in the absence of more datasets to validate it. We encourage more scholars to join this field to enhance the quality of conditional text representation and further explore fine-grained representation learning within natural language processing.

## Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 62476040) and the Fundamental Research Funds for the Central Universities.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona 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 *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263.

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

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. in* sem 2012: The first joint conference on lexical and computational semantics–volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (semeval 2012). *Association for Computational Linguistics. URL http://www. aclweb. org/anthology/S12-1051*.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587.

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.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Ameet Deshpande, Carlos E Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik R Narasimhan. 2023. C-sts: Conditional semantic textual similarity. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

David Ha, Andrew M Dai, and Quoc V Le. 2017. Hypernetworks. In *International Conference on Learning Representations*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. 2022. Boosting contrastive self-supervised learning with false negative cancellation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 986–996. IEEE.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020b. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. Rankcse: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13785–13802.

Jiexi Liu and Songcan Chen. 2024. Timesurl: Self-supervised contrastive learning for universal time series representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13918–13926.

Pu Miao, Zeyao Du, and Junlin Zhang. 2023. Debcse: Rethinking unsupervised contrastive sentence embedding learning in the debiasing perspective. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1847–1856.

OpenAI. 2022. Introducing chatgpt.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

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.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Hao Wang and Yong Dou. 2023. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples. In *International Conference on Intelligent Computing*, pages 419–431.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative

instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907.

Young Yoo, Jii Cha, Changhyeon Kim, and Taeuk Kim. 2024. Hyper-CL: Conditioning sentence representations with hypernetworks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 700–711.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227.

## A Statistical analysis for over-estimation issue

Due to the relative nature of similarity scores, we define samples with significant over-estimation as follows,

$$\begin{cases} cos\left(h_1^{pos}, h_2^{pos}\right) - l_{pos} > 0.25 \\ l_{neg} - cos\left(h_1^{neg}, h_2^{neg}\right) > 0 \end{cases}, \quad (16)$$

$$\begin{cases} l_{neg} - cos\left(h_1^{pos}, h_2^{pos}\right) > 0.25 \\ cos\left(h_1^{pos}, h_2^{pos}\right) - l_{pos} > 0 \end{cases}. \quad (17)$$

Samples that satisfy either Equation 16 or Equation 17 may be considered to exhibit issues of over-estimation.

Based on the bi-encoder architecture, we conduct a statistical analysis of the results from the $\text{SimCSE}_{\text{base}}$ model trained using QuMSE and CCL separately. The results on the C-STS validation set indicate that the number of over-estimation samples for QuMSE is **134**, while for CCL, it is only **46**. Our method reduces the number of over-estimation samples by over fifty percent, thereby providing evidence for the effectiveness of our approach.

| Model | ep. | bs. | ga. | wd. |
|---|---|---|---|---|
| $\text{RoBERTa}_{\text{base}}$ | 10 | 16 | 4 | 0.1 |
| $\text{RoBERTa}_{\text{large}}$ | 10 | 16 | 4 | 0.1 |
| $\text{DiffCSE}_{\text{base}}$ | 5 | 16 | 4 | 0.1 |
| $\text{SimCSE}_{\text{base}}$ | 5 | 16 | 4 | 0.1 |
| $\text{SimCSE}_{\text{large}}$ | 5 | 16 | 4 | 0.1 |

Table 6: Additional hyperparameters of the CCL. The abbreviations ep., bs., ga., and wd. represent epoch, batch size, gradient accumulation and weight decay, respectively.

| Encoding | Model | Pears. |
|---|---|---|
| Bi-encoder‡ (QuMSE) | $\text{RoBERTa}_{\text{base}}$ | 22.3 |
| | $\text{RoBERTa}_{\text{large}}$ | 21.3 |
| | $\text{DiffCSE}_{\text{base}}$ | 43.5 |
| | $\text{SimCSE}_{\text{base}}$ | 44.9 |
| | $\text{SimCSE}_{\text{large}}$ | 47.6 |
| Bi-encoder‡ (CCL) | $\text{RoBERTa}_{\text{base}}$ | **43.1**(+20.8) |
| | $\text{RoBERTa}_{\text{large}}$ | **45.3**(+24) |
| | $\text{DiffCSE}_{\text{base}}$ | **44.8**(+1.3) |
| | $\text{SimCSE}_{\text{base}}$ | **46.3**(+1.4) |
| | $\text{SimCSE}_{\text{large}}$ | **47.7**(+0.1) |
| Tri-encoder‡ (QuMSE) | $\text{RoBERTa}_{\text{base}}$ | 25.2 |
| | $\text{RoBERTa}_{\text{large}}$ | 18.9 |
| | $\text{DiffCSE}_{\text{base}}$ | 27.8 |
| | $\text{SimCSE}_{\text{base}}$ | 31.0 |
| | $\text{SimCSE}_{\text{large}}$ | 35.6 |
| Tri-encoder‡ (CCL) | $\text{RoBERTa}_{\text{base}}$ | **31.0**(+5.8) |
| | $\text{RoBERTa}_{\text{large}}$ | **29.0**(+10.1) |
| | $\text{DiffCSE}_{\text{base}}$ | **33.6**(+5.8) |
| | $\text{SimCSE}_{\text{base}}$ | **35.0**(+4.0) |
| | $\text{SimCSE}_{\text{large}}$ | **36.1**(+0.5) |

Table 7: We report the Pearson correlation of the model on the C-STS test split. Models with † indicate that we directly report the scores from (Deshpande et al., 2023), while models with ‡ indicate models with conditional contrastive framework. Bold font indicates the optimal results.

## B More training details for CCL

Table 6 describes the additional hyperparameters required in CCL. We utilize the same hyperparameter values for both the bi-encoder and tri-encoder architectures.

| Encoding | Model | Pears. |
|---|---|---|
| Bi-encoder (CCL) | RoBERTa$_{\text{base}}$ | **44.3** |
| | $w/o\ l_{\text{W-ACL}}$ | 41.7 |
| | $w/o\ l_{\text{BCL}}$ | 41.6 |
| | $w/o\ l_{\text{MSE}}$ | 43.7 |
| | $w/o\ l_{\text{C-MSE}}$ | 42.3 |
| Bi-encoder (CCL) | SimCSE$_{\text{base}}$ | **46.3** |
| | $w/o\ l_{\text{W-ACL}}$ | 44.1 |
| | $w/o\ l_{\text{BCL}}$ | 45.9 |
| | $w/o\ l_{\text{MSE}}$ | 45.1 |
| | $w/o\ l_{\text{C-MSE}}$ | 45.8 |
| Tri-encoder (CCL) | RoBERTa$_{\text{base}}$ | **30.7** |
| | $w/o\ l_{\text{W-ACL}}$ | 21.5 |
| | $w/o\ l_{\text{BCL}}$ | 27.7 |
| | $w/o\ l_{\text{MSE}}$ | 14.6 |
| | $w/o\ l_{\text{C-MSE}}$ | 26.5 |
| Tri-encoder (CCL) | SimCSE$_{\text{base}}$ | 34.5 |
| | $w/o\ l_{\text{W-ACL}}$ | 32.1 |
| | $w/o\ l_{\text{BCL}}$ | 33.6 |
| | $w/o\ l_{\text{MSE}}$ | 32.6 |
| | $w/o\ l_{\text{C-MSE}}$ | **35.1** |

Table 8: Ablation study on loss functions in the conditional contrastive learning framework.

| Encoding | Model | $\sigma$ | Pears. |
|---|---|---|---|
| Bi-encoder (CCL) | RoBERTa$_{\text{base}}$ | 0 | 42.9 |
| | | **0.25** | **44.4** |
| | | 0.5 | 44.3 |
| | | 0.75 | 43.3 |
| | | 1 | 43.4 |
| Bi-encoder (CCL) | SimCSE$_{\text{base}}$ | 0 | 45.5 |
| | | 0.25 | 45.2 |
| | | 0.5 | 45.9 |
| | | **0.75** | **46.3** |
| | | 1 | 45.7 |
| Tri-encoder (CCL) | RoBERTa$_{\text{base}}$ | 0 | 26.8 |
| | | 0.25 | 25.8 |
| | | **0.5** | **30.7** |
| | | 0.75 | 24.7 |
| | | 1 | 23.7 |
| Tri-encoder (CCL) | SimCSE$_{\text{base}}$ | 0 | 34.2 |
| | | 0.25 | 32.9 |
| | | **0.5** | **34.6** |
| | | 0.75 | 34.3 |
| | | 1 | 34.5 |

Table 9: Analysis of different handling approaches for controversial negative samples.

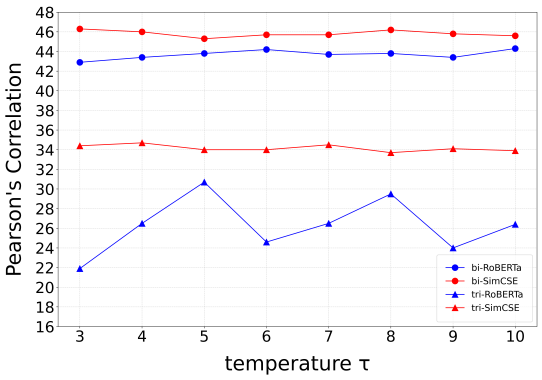


Figure 3: Analysis of temperature coefficient $\tau$ in $l_{\text{BCL}}$.

## C Further experimental analysis.

### C.1 The Pearson correlation for main results

Table 7 shows the main results of the CCL framework for the Pearson correlation on the C-STS test split. Our proposed CCL framework also demonstrates a significant improvement in the Pearson correlation compared to QuMSE (Deshpande et al., 2023), achieving an enhancement of over 20 points on the Roberta model based on the bi-encoder architecture, thereby highlighting the effectiveness of our approach.

### C.2 The Pearson correlation for ablation studies

We show the impact of each component of the CCL framework in Table 8. Consistent with the Spearman correlation, removing any of the contrastive learning losses ($l_{\text{W-ACL}}$ and $l_{\text{BCL}}$) from the CCL framework leads to a decline in model performance. Likewise, the SimCSE model derives minimal benefit from $l_{\text{C-MSE}}$ and may even be adversely affected.

### C.3 The Pearson correlation for controversial negative samples

Table 9 displays the Pearson correlation for various cut-off values of controversial negative samples. $l_{\text{BCL}}$ achieves optimal performance by utilizing a suitably defined parameter $\sigma$, thereby establishing an effective equilibrium between false negative samples and hard negative samples.

### C.4 The Pearson correlation for temperature coefficient

Figure 3 illustrates the Pearson correlation of the model for different temperature coefficients. As

with the Spearman correlation, the models are insensitive to the temperature coefficients, and the Pearson correlations are maintained at a high level, except for the SimCSE model based on the tri-encoder architecture.