ConIsI: A Contrastive Framework with Inter-sentence Interaction for Self-supervised Sentence Representation

Meng Sun
Dalian University of Technology
sunmeng20@mail.dlut.edu.cn

Degen Huang*
Dalian University of Technology
huangdg@dlut.edu.cn

Abstract

Learning sentence representation is a fundamental task in natural language processing and has been studied extensively. Recently, many works have obtained high-quality sentence representation based on contrastive learning from pre-trained models. However, these works suffer the inconsistency of input forms between the pre-training and fine-tuning stages. Also, they typically encode a sentence independently and lack feature interaction between sentences. To conquer these issues, we propose a novel **Con**trastive framework with **I**nter-sentence **I**nteraction (ConIsI), which introduces a sentence-level objective to improve sentence representation based on contrastive learning by fine-grained interaction between sentences. The sentence-level objective guides the model to focus on fine-grained semantic information by feature interaction between sentences, and we design three different sentence construction strategies to explore its effect. We conduct experiments on seven Semantic Textual Similarity (STS) tasks. The experimental results show that our ConIsI models based on $\mathrm{BERT}_{\mathrm{base}}$ and $\mathrm{RoBERTa}_{\mathrm{base}}$ achieve state-of-the-art performance, substantially outperforming previous best models SimCSE-$\mathrm{BERT}_{\mathrm{base}}$ and SimCSE-$\mathrm{RoBERTa}_{\mathrm{base}}$ by 2.05% and 0.77% respectively.

1 Introduction

Learning good universal sentence representation is a fundamental task and benefits a wide range of natural language processing tasks such as text classification and machine translation, especially for large-scale semantic similarity computation and information retrieval. With the rise of pre-trained language models (Devlin et al., 2019; Liu et al., 2019), many downstream tasks have achieved remarkable improvements. However, the native sentence representation derived from pre-trained language models without additional supervision are usually low-quality and can not be used directly (Reimers et al., 2019). Recently, contrastive learning has become a popular approach to improve the quality of sentence representation in a self-supervised way.

Contrastive learning is an approach of learning effective feature representation by positive pairs and negative pairs. It generally takes different views as positive or negative pairs for each sentence using various data augmentation ways. And it works by pulling semantically close positive instances together and pushing negative instances away. However, current approaches based on contrastive learning mainly suffer two problems: *train-tuned bias* and *fine-grained interaction deficiency*. Firstly, previous approaches typically input a single sentence to the encoder at a time, which is inconsistent with the pre-training stage of the language models. Most language models concatenate multiple sentences as the input form at the pre-training stage. We argue that the inconsistency of input forms between the pre-training and fine-tuning stages may harm the performance. Secondly, each sentence in a minibatch is encoded independently while training, which lacks fine-grained interaction information between sentences. According to previous works in text matching (Li et al., 2021; Wang et al., 2021; Lu et al., 2022) , modeling a proper interaction between input sentences can improve the performance of semantic feature embedding

* : Corresponding Author

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

for representation-based models, but existing works on sentence representation ignore the importance of this interaction.

Therefore, to conquer these drawbacks of current contrastive learning based methods, we propose ConIsI, a **Con**trastive framework with **I**nter-**s**entences **I**nteraction for self-supervised sentence representation. Firstly, we present to construct a sentence pair as positive instance for each sentence to alleviate the train-tuned bias. By referring to an original sentence and a sentence pair as a positive pair, the model can not only obtain effective representation of a single sentence, but also mitigate the train-tuned bias between the pre-training and fine-tuning stages. Further, to solve the problem of lacking interaction between sentences, we propose a sentence-level objective to perform the inter-sentence interaction during encoding. We pass a pair of sentences as a text sequence into the encoder and the target semantic category of the two sentences is predicted. The sentence pair is sufficiently interacted through the internal interaction mechanism in Transformer-based block (Vaswani et al., 2017) during encoding. Through the inter-sentence interaction, the model can encode fine-grained semantic information and achieve further improvement. Moreover, for a minibatch of n sentences, there are $n \cdot (n-1)/2$ interactive computations. In order to ensure the training efficiency, we do not perform an interactive operation on all data due to too many possible combinations. Instead, we artificially construct a sentence for each original sentence to adjust the difficulty of the interactive objective, which only requires n interactive computations. We propose several models based on three sentence construction strategies, named ConIsI-o1, ConIsI-o2, and ConIsI-s, respectively. The overall model of our proposed ConIsI can be seen in Figure 1.

Our contributions can be summarized as follows:

- We propose to construct each positive pair with an original sentence and a sentence pair based on contrastive learning, which not only learns effective representation by pulling semantically close samples together but also mitigates the train-tuned bias between pre-training and fine-tuning phases.
- We propose a simple but effective sentence-level training objective based on inter-sentence interaction. It alleviates the problem of interaction deficiency among sentences and enriches the semantic information of sentence representation. We also present three sentence construction strategies for interactive sentence pairs and analyze their effects.
- We conduct extensive experiments on seven standard Semantic Textual Similarity (STS) datasets. The results show that our proposed ConIsI-s-$\text{BERT}_{\text{base}}$ and ConIsI-s-$\text{RoBERTa}_{\text{base}}$ achieve 78.30% and 77.34% averaged Spearman's correlation, a 2.05% and 0.77% improvement over SimCSE-$\text{BERT}_{\text{base}}$ and SimCSE-$\text{RoBERTa}_{\text{base}}$ respectively, which substantially outperforms the previous state-of-the-art models.

2 Related Work

Sentence representation built upon the distributional hypothesis has been widely studied and improved considerably. Early works (Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018) inspired by word2vec (Mikolov et al., 2013) lead to strong results by predicting surrounding information of a given sentence. The emergence of pre-trained models such as BERT (Devlin et al., 2019) shows much great potential for sentence representation. Recently, many works have explored how to learn better sentence embeddings from the pre-trained models.

Supervised Methods A common supervised step of learning a model is fine-tuning with labeled data in downstream training sets. Several works build upon the success of using annotated natural language inference (NLI) datasets (including Stanford NLI (Bowman et al., 2015) and Multi-Genre NLI (Williams et al., 2018)) for sentence representation, which projects it as a 3-way classification task (entailment, neutral, and contradiction) to get better sentence embeddings. Conneau et al. (2017) use a BiLSTM-based model as encoder, and they train it on both Stanford NLI and Multi-Genre NLI datasets. Universal Sentence Encoder (Cer et al., 2018) uses the Stanford NLI dataset to enhance the unsupervised training by adopting a Transformer-based model. Sentence-BERT (Reimers et al., 2019) that adopts a Siamese network (Chopra et al., 2005) with a shared BERT encoder is also trained on Stanford NLI and Multi-Genre NLI datasets.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

Unsupervised Methods Some works focus on using the regularization method to improve the quality of raw sentence representation generated by original BERT. Bert-flow (Li et al., 2020) puts forward a flow-based approach to solving the problem that native embeddings of BERT occupy a narrow cone in the vector space. Similarly, Bert-whitening (Su et al., 2021) maps BERT's embeddings to a standard Gaussian latent space by whitening the native embeddings. They all try to alleviate the representation degeneration of pre-trained models and yield substantial improvement.

Self-supervised Methods The sentence-level training objective in language models like BERT inspires a line of work over self-supervised sentence representation learning. BERT includes the next sentence prediction (NSP) task, which predicts whether two sentences are neighboring or not. However, Liu et al. (2019) prove that NSP has minimal effect on the final performance and even does harm to the training model. Therefore, many works have proposed various self-supervised objectives for pre-training sentence encoders. Cross-Thought (Wang et al., 2020) and CMLM (Yang et al., 2021) are two similar approaches that present to predict surrounding tokens of given contextual sentences. And Lee et al. (2020) propose to learn an objective that predicts the correct sentence ordering provided the input of shuffled sentences.

As a self-supervised learning method, contrastive learning with no need for scarce labeled data attracts much attention, and many excellent works have been proposed. Inspired by SimCLR (Chen et al., 2020) which applies data augmentation techniques on the same anchor such as image rotating, scaling, and random cropping to learn image representation in the computer vision community, some works pay attention to getting effective positive pairs by using similar approaches. In the natural language process community, many works apply textual augmentation techniques on the same sentence to obtain different views as positive pairs based on the SimCLR framework. Zhang et al. (2020) extract global feature of a sentence as positive pairs, Wu et al. (2020) and Yan et al. (2021) take some token-level transformation ways such as word or subword deletion or replacement, and Gao et al. (2021) apply dropout mask of Transformer-based encoder to get positive pairs. And Zhang et al. (2021) adopt BYOL (Grill et al., 2020) framework using back-translation data.

3 Methodology

In this section, we present ConIsI, a contrastive framework with inter-sentence interaction for self-supervised sentence representation, which contains two parts: (1) the ConIsI model of joint contrastive learning objective and inter-sentence interactive objective (Section 3.1), and (2) the strategies of sentence construction in the inter-sentence interactive objective (Section 3.2).

3.1 Model

The ConIsI model joints contrastive learning and inter-sentence interactive objectives. The inter-sentence interactive objective is a binary classification task that performs fine-grained interaction between sentences and predicts whether two sentences are in the same semantic category. The overall architecture is shown in Figure 1.

3.1.1 Data Augmentation

To alleviate the train-tuned bias caused by different input forms, we perform sentence-level repetition operation to construct positive instances. For each sentence, our approach proposes to take a sentence pair as positive instance. Specifically, given a tokenized sentence $x = \{t_1, t_2, ..., t_l\}$ (l is the max sequence length), we define the sentence pair as $Y = \{t_1, t_2, ..., t_l, t_1, t_2, ..., t_l\}$, which is the concatenation of two original sentences. For each minibatch of sentences $\mathcal{B} = \{x_i\}_{i=1}^{N}$ (N is the batch size), we perform data augmentation operation on each sentence and then get the positive instances $\mathcal{B}_{\text{Aug}} = \{Y_i\}_{i=1}^{N}$.

3.1.2 Sentence Pair Composition

To perform fine-grained interaction between sentences, we take a pair of sentences as a textual sequence to input into the encoder. The input two sentences can get fine-grained interaction with each other through Transformer-based block. Also, considering the training efficiency, we do not perform interaction on all sentences as there are too many combinations of sentence pairs. Instead, we construct the composed

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

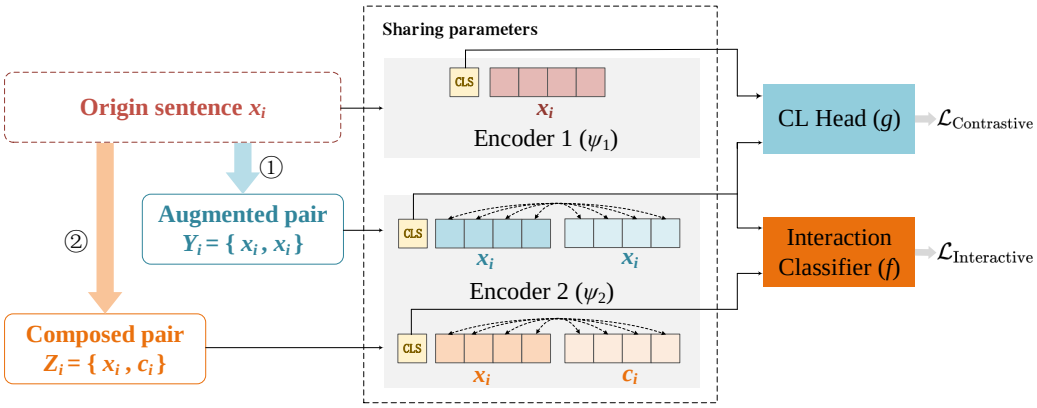

Figure 1: The overall structure of the ConIsI model. It mainly consists of five components: the data augmentation operation (①), the text composition part (②), the encoder $\psi(\cdot)$ mapping the input data to the sentence representation space, the CL Head $g(\cdot)$ and the Interaction Classifier $f(\cdot)$ applying for the contrastive loss and the interactive loss respectively.

sentence pair $Z_i = \{x_i, c_i\}$ for each sentence x_i in $\mathcal{B}$. Specifically, we try to obtain a sentence c_i which belongs to a different semantic category from x_i. Then we concatenate the sentence x_i and the sentence c_i as a composed sentence pair Z_i. We perform the sentence pair composition operation on each sentence in minibatch $\mathcal{B} = \{x_i\}_{i=1}^N$ and then get the composed pairs $\mathcal{B}_{\text{Com}} = \{Z_i\}_{i=1}^N$. We explore three different sentence construction strategies to obtain c_i in section 3.2.

3.1.3 Encoding

We take pre-trained checkpoints of BERT or RoBERTa as the encoder model to obtain sentence representation. For BERT, there are two input forms to fine-tune downstream tasks: one is the single sentence input, and the other is the sentence pair input. Previous works based on contrastive learning input a single sentence to the pre-trained model to learn sentence embeddings, which is inconsistent with the pre-training stage and suffers the train-tuned bias. To alleviate this problem and maintain the model's ability of encoding a single sentence meanwhile, we propose to adopt both two forms. The original sentence x_i is taken as a single sentence and input to the encoder 1. The augmented sentence pair Y_i and the composed sentence pair Z_i are taken as sentence pairs and input to the encoder 2. And to ensure that the augmented sentence pair has the same meaning as the original sentence, the max length of the tokenizer for the former is set double for the latter. The encoder 1 and the encoder 2 share the same parameters.

For RoBERTa whose input forms are a single sentence or several concatenated sentences separated by "</s>" token, we input the original sentence into the encoder 1. And The augmented sequence pair and the composed sentence pair are taken as two concatenated sentences and input to the encoder 2. Similarly, the max length of the tokenizer for encoder 2 is set double for that of encoder 1, and the two encoders share the same parameters.

3.1.4 Contrastive Learning

Contrastive learning aims to learn effective representation by pulling semantically close objects and pushing ones that are dissimilar away. We follow the SimCRL (Chen et al., 2020) contrastive framework and take a cross-entropy objective (Chen et al., 2017) in our approach.

For each minibatch $\mathcal{B} = \{x_i\}_{i=1}^N$, the contrastive loss is defined on $\mathcal{B}$ and the augmented instances $\mathcal{B}_{\text{Aug}} = \{Y_i\}_{i=1}^N$. Let $i \in \{1, .., N\}$ denote the index of an arbitrary instance in augmented set $\mathcal{B}_{\text{Aug}}$, and let $j \in \{1, .., N\}$ be the index of the other instance in $\mathcal{B}_{\text{Aug}}$. We refer to (x_i, Y_i) as a positive pair, while treating the other $N-1$ examples $Y_j (j \neq i)$ in $\mathcal{B}_{\text{Aug}}$ as negative instances for this positive pair. After the positive pair is encoded, we obtain the last hidden state of the special "[CLS]" token as the contextual

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

representation of the corresponding sample, denoted as $h_{[\mathrm{CLS}]}$.

$$\begin{aligned} h^x_{[\mathrm{CLS}]}, h^x_1, ..., h^x_l, h^x_{[\mathrm{SEP}]} &= \psi_1(x) \\ h^Y_{[\mathrm{CLS}]}, h^Y_1, ..., h^Y_l, h^Y_{[\mathrm{SEP}]}, h^{Y'}_1, ..., h^{Y'}_l, h^{Y'}_{[\mathrm{SEP}]} &= \psi_2(Y) \end{aligned} \tag{1}$$

Then we add a predictor layer $g(\cdot)$ to map $h_{[\mathrm{CLS}]}$ to the contrastive embedding space and obtain h , which is given as follows:

$$h = Elu(BN_1(W_1 \cdot h_{[\mathrm{CLS}]} + b_1)) \tag{2}$$

where $W_1 \in R^{d\times d}$ is the weight matrix, $b_1 \in R^{d\times 1}$ is the bias vector, and d is the number of features in hidden layers. Both W_1 and b_1 are trainable parameters. BN_1 is the $\mathrm{BatchNorm1d}$ layer and Elu is the activate function.

Let h^x_i, h^Y_i and h^Y_j be the corresponding outputs of the head $g(\cdot)$. Then for x_i, we try to separate Y_i apart from all negative instances by minimizing the following,

$$\ell^I_i = -log\frac{e^{sim(h^x_i, h^Y_i)/\tau}}{\sum_{j=1}^{N} e^{sim(h^x_i, h^Y_j)/\tau}} \tag{3}$$

where τ denotes the temperature parameter we set as 0.05. We choose cosine similarity $sim(\cdot)$ as the similarity calculation function between a pair of normalized outputs, $sim(h_1, h_2) = \frac{h_1^T h_2}{||h_1||\cdot||h_2||}$.

The contrastive loss is then averaged over all pairs,

$$\mathcal{L}_{\mathrm{Contrastive}} = \sum_{i=1}^{N} \ell^I_i / N \tag{4}$$

3.1.5 Interactive Classification

When applying a training objective after getting sentence embeddings in previous work, each sentence is encoded independently and can not see other sentences while encoding. Therefore, the semantic information contained in each sentence embeddings is insufficient. In contrast, modeling sentence pairs can effectively alleviate this problem. While encoding a sentence pair through the model, the two sentences can obtain fine-grained interaction information from each other. We propose to model an inter-sentence interaction objective between input sentences to enrich semantic information for sentence embeddings.

We encode the sentence pairs into the semantic category space for self-supervised classification. Different from contrastive learning objective, the interactive objective learns fine-grained semantic information through the interaction between sentences. The interactive loss is implemented on the augmented instance Y_i in B_{Aug} and the corresponding composed instance Z_i in $\mathcal{B}_{\mathrm{Com}}$. We refer to the two sentences $\{x_i, x_i\}$ in augmented pair Y_i as being in the same category, and the sentences $\{x_i, c_i\}$ in composed pair Z_i as being in different category. Our model passes Y_i and Z_i to the encoder 2 and obtains the last hidden state of the special "[CLS]" token as their sentence pair embeddings, respectively.

$$\begin{aligned} h^Y_{[\mathrm{CLS}]}, h^Y_1, ..., h^Y_l, h^Y_{[\mathrm{SEP}]}, h^{Y'}_1, ..., h^{Y'}_l, h^{Y'}_{[\mathrm{SEP}]} &= \psi_2(Y) \\ h^Z_{[\mathrm{CLS}]}, h^Z_1, ..., h^Z_l, h^Z_{[\mathrm{SEP}]}, h^{Z'}_1, ..., h^{Z'}_l, h^{Z'}_{[\mathrm{SEP}]} &= \psi_2(Z) \end{aligned} \tag{5}$$

We use a predictor and linear layers to encode $h_{[\mathrm{CLS}]}$ into the semantic category space to obtain r. $r \in R^d$ is the semantic category representation. The formulas are as follows:

$$h = Elu(BN_2(W_2 \cdot h_{[\mathrm{CLS}]} + b_2)) \tag{6}$$

$$r = W_3 \cdot h + b_3 \tag{7}$$

where $W_2, W_3 \in R^{d\times d}$ are the weight matrixs, $b_2, b_3 \in R^{d\times 1}$ are the bias vectors, and d is the number of features in the hidden layers. W_2, W_3 and b_2, b_3 are all learnable parameters, and W_2, b_2 share the same parameters with W_1 and b_1 in $g(\cdot)$ respectively. BN_2 share the same parameters with BN_1 and Elu is the activate function.

Let r^Y_i and r^Z_i denote the corresponding outputs of the head $f(\cdot)$. Then we predict whether each pair

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

is in the same category by optimizing the following objective,

$$\ell_i^{II} = -log\frac{e^{r_i^Y}}{e^{r_i^Y} + e^{r_i^Z}} \tag{8}$$

Then the interactive loss for a mini-batch with N sentence pairs is as follows:

$$\mathcal{L}_{\text{Interactive}} = \sum_{i=1}^{N} \ell_i^{II}/N \tag{9}$$

3.1.6 Overall objective

Finally, our overall objective is,

$$\begin{aligned}\mathcal{L} &= (1-\lambda)\cdot\mathcal{L}_{\text{Contrastive}} + \lambda\cdot\mathcal{L}_{\text{Interactive}} \\ &= (1-\lambda)\cdot\sum_{i=1}^{N}\ell_i^{I}/N + \lambda\cdot\sum_{i=1}^{N}\ell_i^{II}/N\end{aligned} \tag{10}$$

where ℓ_i^I, ℓ_i^{II} are defined in Eq(3) and Eq(8), respectively. λ is the balanced parameter between the contrastive loss and the interactive loss. During training, we jointly optimize a contrastive learning objective and an inter-sentence interactive objective over the original sentences, the augmented sentence pairs and composed sentence pairs. Then we fine-tune all the parameters using the joint objective.

3.2 Sentence Construction Techniques

Intuitively, two semantically opposite sentences are easier for the model to distinguish than two semantically closer sentences. As a self-supervised classification task, the difficulty of the interactive objective can significantly affect the performance of the model. Thus we propose different sentence construction techniques to control the complexity of the inter-sentence interactive objective. We try to construct a sentence c_i that is not in the same semantic category as the original sentence x_i in section 3.1.2. We explore three sentence construction methods, two of which are constructing from the original sentence as shown in section 3.2.1, and one is sampling from other sentences in section 3.2.2.

3.2.1 From Original Sentence

Since the bidirectional language models encode a word based on contextual information, sentences with high textual similarity usually are in high semantic similarity in representation. However, the sentences with high textual similarity may not actually be semantically similar. For example, "this is not a problem." and "this is a big problem." are two sentences with high textual similarity because of similar wording, but they are not semantically similar because of opposite meanings. The models usually fail to distinguish textual similarity and semantic similarity, which has been discussed deeply in the vision field (Robinson et al., 2021; Chen et al., 2021). As a result, a model may overestimate the semantic similarity of any pairs with similar wording regardless of the actual semantic difference between them. Therefore, we propose to construct sentences that are semantically different but are textually similar to the original sentence to improve the fine-grained semantic discrimination ability of the model.

Subword Replacement The subword replacement mechanism randomly substitutes some sub-words in a sentence. Specifically, given a tokenized sub-word sequence $x = \{t_1, t_2, ..., t_l\}$ (l is the max sequence length) after processing by a sub-word tokenizer. Firstly, We mask a certain proportion of the tokenized sequence x at random. If the i-th token is chosen, then we replace the masked token with a random token 80% of the time, leaving the masked token unchanged 20% of the time.

Word Replacement The word replacement mechanism works on full words in a sentence. Different from subword replacement, the word replacement mechanism randomly substitutes some full words with antonyms. If a word is chosen, then we replace the word with its antonym. We use the WordNet (Miller, 1993) to obtain the antonym of a word.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

3.2.2 From Other Sentences

Different from constructing a new sentence from the original sentence, this method selects one other sentence from the training data at random. Specifically, for a given sentence x_i within the minibatch $\mathcal{B} = \{x_i\}_{i=1}^{N}$, we randomly select sentence x_k $(k \in [1, N], k \neq i)$ as c_i for composed pair.

We apply the three sentence construction strategies to our ConIsI model, named ConIsI-o1, ConIsI-o2, and ConIsI-s. Among them, ConIsI-o1 and ConIsI-o2 represent the joint contrastive objective and interactive objective under the subword replacement and word replacement, respectively. ConIsI-s represents the jointing of contrastive learning and the interactive objective under the sampling from other sentences.

4 Experiments

4.1 Data

We train our model on the same one million sentences randomly sampled from English Wikipedia that are provided by SimCSE[0]. All our experiments are fully self-supervised and note that no STS sets are used for training.

We evaluate our approach on multiple Semantic Textual Similarity (STS) datasets: STS12-16 (STS12 - STS16) (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015; Agirre et al., 2016), STS Benchmark (STS-B) (Cer et al., 2017) and SICK-Relatedness (SICK-R) (Marelli et al., 2014), which are seven standard STS benchmark datasets and are extensively used to measure the sentence embeddings and the semantic similarity of sentence pairs. These datasets are composed of pairs of sentences and one golden score between 0 and 5, where a higher score indicates a higher similarity between two sentences in Table 1. The statistics is shown in Table 2.

Sentence1	Sentence2	Golden Score
a plane is taking off .	an air plane is taking off .	5.000
a cat is playing a piano .	a man is playing a guitar .	0.600
a man is playing a guitar .	a man is playing a trumpet .	1.714

Table 1: The sentence samples of STS datasets.

	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Total
Number of train samples	0	0	0	0	0	5479	4500	-
Number of valid samples	0	0	0	0	0	1500	500	-
Number of test samples	3108	1500	3750	3000	1186	1379	4927	-
Number of Unlabeled Texts	6216	3000	7500	17000	18366	17256	19854	89192

Table 2: The statistics of STS datasets.

4.2 Evaluation Setup

Following previous work, we evaluate our method on STS tasks using the SentEval toolkit (Conneau and Kiela, 2018). We take the "[CLS]" embedding generated by the last hidden layer of the encoder 1 in Figure 1 as the sentence representation. To evaluate the sentence representation for a fair comparison, we follow the settings of Sentence-BERT (Reimers et al., 2019) and SimCSE (Gao et al., 2021): (1) we directly take cosine similarities for all STS tasks without training extra linear regressor on top of frozen sentence embeddings for STS-B and SICK-R; (2) we report Spearman's rank correlation coefficients rather than Pearson's; (3) and we take the "all" setting for STS12-STS16 which fuses data from different topics together to make the evaluation closer to real-world scenarios.

[0]https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse/resolve/main/wiki1m for simcse.txt

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

4.3 Training Details

We implement our ConIsI model with Huggingface's transformers package[1] 4.2.1 based on Python 3.8.12 and Pytorch 1.8.0 and run the model on Nvidia 3090 GPU. We start our experiments from pre-trained checkpoints of BERT or RoBERTa. All experiments use the Adam optimizer and the random seed is set as 42. The temperature parameter τ is set as 0.05, and the dropout rate is set as 0.1. Furthermore, the hyper-parameter settings of the models are shown in Table 3. Besides, We train our models for one epoch and evaluate the model every 125 training steps.

Model	Batch size	Max sequence length	Learning rate	Hidden size	λ
ConIsI-s-$\text{BERT}_{\text{base}}$	64	32	3e-5	768	0.8
ConIsI-s-$\text{RoBERTa}_{\text{base}}$	64	32	3e-5	768	0.1
ConIsI-s-$\text{BERT}_{\text{large}}$	64	28	3e-5	1024	0.1

Table 3: Hyper-parameters settings for ConIsI-s models.

4.4 Baselines

We compare our model with previous strong baseline models on STS tasks, including:
(1) Recent state-of-the-art self-supervised models using a contrastive objective: SimCSE (Gao et al., 2021), IS-BERT (Zhang et al., 2020), ConSERT (Yan et al., 2021), Mirror-BERT (Liu et al., 2021), DeCLUTR (Giorgi et al., 2021), CT-BERT (Carlsson et al., 2020), BSL (Zhang et al., 2021), SG-OPT (Kim et al., 2021);
(2) Post-processing methods like BERT-flow (Li et al., 2020) and BERT-whitening (Su et al., 2021);
(3) And naive baselines like averaged GloVe embeddings (Pennington et al., 2014); averaged first and last layer BERT embeddings.

4.5 Main Results

Table 4 shows the evaluation results on seven STS tasks. ConIsI-s-$\text{BERT}_{\text{base}}$ can significantly outperform SimCSE-$\text{BERT}_{\text{base}}$ and raise the averaged Spearman's correlation from 76.25% to 78.30%, which brings a 2.05% average improvement over the SimCSE-$\text{BERT}_{\text{base}}$ model on seven tasks. For the RoBERTa model, ConIsI-s-$\text{RoBERTa}_{\text{base}}$ can also improve upon SimCSE-$\text{RoBERTa}_{\text{base}}$ from 76.57% to 77.34%, a 0.77% increase. And for the ConIsI-s-$\text{BERT}_{\text{large}}$ model, we also achieve better performance, from 78.41% to 79.55%, a 1.14% increase. In general, our method achieves substantial improvement on the seven STS datasets over baseline models.

4.6 Ablation Study

In this section, we discuss the effects of different components. In our model, both the contrastive learning objective and the inter-sentence interactive objective are crucial because they are committed to obtaining the ability of normal semantic encoding and fine-grained semantic information, respectively. If we remove the inter-sentence interactive objective, the model becomes a SimCSE-like model with a different positive instance construction way, causing a drop of 1.30%. If we remove the contrastive learning objective, the performance of **Avg.** drops significantly by more than 10% (see Table 5). This results show that it is important to have common and fine-grained attributes that exist together in the sentence representation space. When compared with SimCSE-$\text{BERT}_{\text{base}}$, our proposed method of taking a sentence pair as positive instance brings an improvement of 0.75%. The result shows that the problem of train-tuned bias is alleviated by the input form of augmented sentence pair.

4.7 Analysis

In this section, we conduct a series of experiments to validate our model better. We use $\text{BERT}_{\text{base}}$ or $\text{RoBERTa}_{\text{base}}$ model and all reported results are evaluated on the seven STS tasks.

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

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe-embeddings(avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$(first-last avg.)◇	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-flow◇	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening◇	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$§	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
BSL-$\text{BERT}_{\text{base}}$†	67.83	71.40	66.88	79.97	73.97	73.74	70.40	72.03
CT-$\text{BERT}_{\text{base}}$◇	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
ConSERT-$\text{BERT}_{\text{base}}$‡	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SG-OPT-$\text{BERT}_{\text{base}}$♭	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
Mirror-$\text{BERT}_{\text{base}}$♮	69.10	81.10	73.00	81.90	75.70	78.00	69.10	75.40
SimCSE-$\text{BERT}_{\text{base}}$◇	68.40	82.41	74.38	80.91	78.56	76.85	**72.23**	76.25
∗ConIsI-s-$\text{BERT}_{\text{base}}$	**70.92**	**84.35**	**76.67**	**83.53**	**78.94**	**82.15**	71.55	**78.30**
$\text{RoBERTa}_{\text{base}}$(first-last avg.)◇	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\text{base}}$whitening◇	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-$\text{RoBERTa}_{\text{base}}$◇	52.41	75.19	65.52	77.12	78.63	72.41	68.62	69.99
SimCSE-$\text{RoBERTa}_{\text{base}}$◇	70.16	81.77	73.24	**81.36**	80.65	80.22	68.56	76.57
∗ ConIsI-s-$\text{RoBERTa}_{\text{base}}$	**71.21**	**83.31**	**75.11**	81.13	**80.73**	**80.50**	**69.39**	**77.34**
SimCSE-$\text{BERT}_{\text{large}}$◇	70.88	84.16	76.43	84.50	**79.76**	79.26	73.88	78.41
∗ ConIsI-s-$\text{BERT}_{\text{large}}$	**72.33**	**86.14**	**77.42**	**84.83**	79.60	**81.76**	**74.78**	**79.55**

Table 4: Sentence embedding performance on STS tasks in terms of Spearman's correlation and "all" setting. ♣: results from (Reimers et al., 2019); §: results from (Zhang et al., 2020); †: results from (Zhang et al., 2021); ‡: results from (Yan et al., 2021); ♭: results from (Kim et al., 2021); ♮: results from (Liu et al., 2021); ◇: results from (Gao et al., 2021); ∗: results from ours.

Model	Avg.
SimCSE-$\text{BERT}_{\text{base}}$	76.25
ConIsI-s-$\text{BERT}_{\text{base}}$	78.30
w/o fine-grained classification loss	77.00 (-1.30)(+0.75)
w/o contrastive loss	67.68 (-10.62)

Table 5: **Avg.** results of seven STS tasks for ConIsI-s-$\text{BERT}_{\text{base}}$ model variants.

4.7.1 Validation of Sentence Construction Strategies

We compare the three models ConIsI-o1, ConIsI-o2, and ConIsI-s to verify the effects of our proposed sentence construction strategies for the inter-sentence interactive objective.

Table 6 shows that our proposed sentence construction techniques for the inter-sentence interactive objective improve the performance of self-supervised sentence representation. Compared with SimCSE-$\text{BERT}_{\text{base}}$ and SimCSE-$\text{RoBERTa}_{\text{base}}$, the Spearman's correlation of ConIsI-o1-$\text{BERT}_{\text{base}}$ and ConIsI-o1-$\text{RoBERTa}_{\text{base}}$ on seven STS tasks have improved by 0.89% and 1.78% respectively, a 1.34% increase on average. The results of ConIsI-o2-$\text{BERT}_{\text{base}}$ and ConIsI-o2-$\text{RoBERTa}_{\text{base}}$ on seven STS tasks have improved by 1.10% and 1.56% respectively, a 1.33% increase on average. The results of ConIsI-s-$\text{BERT}_{\text{base}}$ and ConIsI-s-$\text{RoBERTa}_{\text{base}}$ have improved by 2.05% and 0.77% respectively, a 1.41% increase on average.

As the Table 6 shown, the ConIsI-o1-$\text{RoBERTa}_{\text{base}}$ and ConIsI-o2-$\text{RoBERTa}_{\text{base}}$ implemented by the strategies of "from original sentence" bring more remarkable improvement to the SimCSE-RoBERTa model, exceeding 1.5%. And the ConIsI-s models implemented by the strategy of "from other sentences" gets a lower boost to the SimCSE-RoBERTa model, but a greater improvement to the SimCSE-BERT model. That is, RoBERTa is more capable of encoding fine-grained features and distinguishing textual similarity and semantic similarity than BERT. In contrast, BERT focuses

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

Model	Avg.	Model	Avg.
SimCSE-BERT$_{\text{base}}$	76.25	SimCSE-RoBERTa$_{\text{base}}$	76.57
∗ConIsI-o1-BERT$_{\text{base}}$	77.14	∗ConIsI-o1-RoBERTa$_{\text{base}}$	78.35
∗ConIsI-o2-BERT$_{\text{base}}$	77.35	∗ConIsI-o2-RoBERTa$_{\text{base}}$	78.13
∗ConIsI-s-BERT$_{\text{base}}$	78.30	∗ConIsI-s-RoBERTa$_{\text{base}}$	77.34

Table 6: Validation results of sentence construction strategies.

more on encoding common features in the sentence representation space. We argue that the pre-trained RoBERTa model pays more attention to fine-grained features because of the more refined optimization techniques than BERT in the pre-training phase. So ConIsI-o1-RoBERTa$_{\text{base}}$ and ConIsI-o2-RoBERTa$_{\text{base}}$ achieve better performance than ConIsI-s-RoBERTa$_{\text{base}}$. While ConIsI-s-BERT$_{\text{base}}$ achieves better performance than ConIsI-o1-BERT$_{\text{base}}$ and ConIsI-o2-BERT$_{\text{base}}$.

Overall, our proposed contrastive framework with inter-sentence interaction have improved performance compared with the previous best model SimCSE. The experimental results show that the three sentence construction strategies are effective for the ConIsI model. We take the ConIsI-s model's results as our final ConIsI model's performance.

4.7.2 Effect of Coefficient λ

λ is the weighted hyperparameter for contrastive loss and inter-sentence interactive loss involved in the final joint objective function Eq(10). A smaller λ means a larger contrastive loss weight, indicating that the model pays more attention to common features. And a larger λ means a larger interactive loss weight, indicating that the model focuses more on fine-grained features. Our experiments find that λ plays an essential role in the joint objective, and the experimental results are shown in Table 7. When $\lambda = 0$, the model becomes a SimCSE-like model, and the result shows that our proposed method to take a sentence pair as the positive instance is effective, which brings an improvement over SimCSE-BERT$_{\text{base}}$(Gao et al., 2021) by 0.75%. The results prove that the interactive objective is helpful to enhance the performance of the model under different λ. And when $\lambda = 0.8$, it achieves the best performance on the STS datasets and gets substantial improvement over that when $\lambda = 0$.

λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Avg.	77.00	77.97	77.42	77.78	77.81	78.03	77.76	77.93	**78.30**	77.58

Table 7: **Avg.** results of seven STS tasks under different λ for ConIsI-s-BERT$_{\text{base}}$ model.

5 Conclusion

In this paper, we propose the ConIsI model, which joints contrastive learning and inter-sentence interactive training objective for optimization. We propose to perform a sentence repetition operation on each sentence and then take the augmented pair as a positive instance based on contrastive learning, which alleviates the train-tuned bias of language models. We also propose the inter-sentence interactive objective, which guides the model to focus on fine-grained semantic information by feature interaction between sentences. Moreover, we design three sentence construction strategies in the inter-sentence interactive objective. Experimental results show our proposed ConIsI achieves substantial improvement over the previous state-of-the-art models. In the future, we will further explore more effective inter-sentence interactive way to enrich semantic information in sentence representation, and we hope to apply our approach to other downstream tasks such as machine translation.

References

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–*

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

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

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

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

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

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

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

Ting Chen, Calvin Luo, and Lala Li. 2021. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

A Conneau, D Kiela, H Schwenk, L Barrault, and A Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

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.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual 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 879–895.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284.

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

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.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Haejun Lee, Drew A Hudson, Kangwook Lee, and Christopher D Manning. 2020. Slm: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. 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.

Dan Li, Yang Yang, Hongyin Tang, Jingang Wang, Tong Xu, Wei Wu, and Enhong Chen. 2021. Virt: Improving representation-based models for text matching through virtual interaction. *arXiv preprint arXiv:2112.04195*.

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

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

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

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, et al. 2022. Ernie-search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. *arXiv preprint arXiv:2205.09153*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Lrec*, pages 216–223. Reykjavik.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1993. Wordnet: A lexical database for english. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

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

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 671–688. Association for Computational Linguistics.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in Neural Information Processing Systems*, 34.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shuohang Wang, Yuwei Fang, Siqi Sun, Zhe Gan, Yu Cheng, Jingjing Liu, and Jing Jiang. 2020. Cross-thought for sentence encoder pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–421.

Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. 2021. Distilled dual-encoder model for vision-language understanding. *arXiv preprint arXiv:2112.08723*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

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

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

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6216–6228.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped unsupervised sentence representation learning. 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 5168–5180.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.