# DenoSent: A Denoising Objective for Self-Supervised Sentence Representation Learning

**Xinghao Wang, Junliang He, Pengyu Wang, Yunhua Zhou, Tianxiang Sun, Xipeng Qiu***

School of Computer Science, Fudan University
{xinghaowang22, jlhe22, pywang22}@m.fudan.edu.cn, {zhouyh20, txsun19, xpqiu}@fudan.edu.cn

## Abstract

Contrastive-learning-based methods have dominated sentence representation learning. These methods regularize the representation space by pulling similar sentence representations closer and pushing away the dissimilar ones and have been proven effective in various NLP tasks, e.g., semantic textual similarity (STS) tasks. However, it is challenging for these methods to learn fine-grained semantics as they only learn from the *inter-sentence* perspective, i.e., their supervision signal comes from the relationship between data samples. In this work, we propose a novel denoising objective that inherits from another perspective, i.e., the *intra-sentence* perspective. By introducing both discrete and continuous noise, we generate noisy sentences and then train our model to restore them to their original form. Our empirical evaluations demonstrate that this approach delivers competitive results on both semantic textual similarity (STS) and a wide range of transfer tasks, standing up well in comparison to contrastive-learning-based methods. Notably, the proposed intra-sentence denoising objective complements existing inter-sentence contrastive methodologies and can be integrated with them to further enhance performance. Our code is available at https://github.com/xinghaow99/DenoSent.

## Introduction

Sentence representation learning is a fundamental task for natural language processing, which aims to embed sentence-level semantics into vectors of a fixed-sized $d$. High-quality sentence representations are expected to form a uniform space where similar semantics stay close, which is proven beneficial to various downstream tasks such as semantic textual similarity and information retrieval.

Transformer (Vaswani et al. 2017)-based pre-trained language models (PLMs) like BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019) have shown remarkably superior performance on token-level tasks and can be adapted to various downstream tasks through fine-tuning, but they perform poorly in encoding sentence-level semantics due to the well-known anisotropy phenomenon in their representation space. Therefore, further training these PLMs for sentence-level representation learning remains a challenge.

Recently, contrastive methods have been adopted to sentence representation learning (Gao, Yao, and Chen 2021; Yan et al. 2021; Giorgi et al. 2021) and brought substantial improvement in both STS tasks and transfer tasks like sentiment analysis. These methods regularize the pre-trained language models (PLMs) representation space to be less anisotropic (Ethayarajh 2019; Wang and Isola 2020), yielding competitive performance in downstream tasks.

However, one limitation of contrastive-learning-based methods is that their performance is highly dependent on the strategies of constructing positive pairs and selecting negative pairs. For instance, previous works adopted standard dropout (Gao, Yao, and Chen 2021), different data augmentation strategies (Yan et al. 2021) and different prompts (Jiang et al. 2022) to construct positive pairs and may include a true-negative sample selection (Zhou et al. 2022) to alleviate the above problem. Nevertheless, contrastive methods solely learn the representation from the inter-sentence perspective, i.e., their supervision signal comes from the relationship between data samples, making it challenging to capture fine-grained semantics.

To address the above issue, we start from another perspective, i.e., the intra-sentence perspective, to learn sentence representations. In this work, we propose a novel denoising objective for sentence representation learning, which corresponds to another main branch of self-supervised learning (Liu et al. 2021) other than contrastive, the generative branch, to provide intra-sentence supervision signals. Specifically, we adopt an encoder-decoder model structure that is identical to the original Transformer, except we only keep the encoded sentence representation to do cross-attention with a noisy version of the original sentence input. The training objective is to recover the noisy input to its original. Furthermore, the structure of our training framework has been designed to enable self-supervised integration of both intra-sentence and inter-sentence objectives.

Our main contributions can be summarized as follows:

1. We propose a novel training objective to learn high-quality sentence representations from an intra-sentence perspective, i.e., utilize an auto-encoder structure and learn sentence representations by reconstructing the input sentence.
2. We incorporate both discrete noises and continuous noises into our training framework, which facilitates our

*Corresponding author

proposed denoising objective.

3. We demonstrate that the proposed denoising objective is complementary to the contrastive objective, thereby proposing a promising sentence representation learning framework that combines both the intra-sentence and inter-sentence supervision signals.

# Preliminaries

## Sentence Representation Learning

Sentence representations strive to encapsulate the underlying semantics and are adaptable for diverse applications. Each dense vector that represents a sentence enables direct measurement of semantic similarities, facilitates information retrieval, and supports training of classifiers tailored to diverse downstream tasks. There are two paradigms for generating sentence representations: frequency-based methods such as Bag-of-Words-based and TF-IDF-based and neural network-based methods like variants of Word2Vec (Mikolov et al. 2013; Hill, Cho, and Korhonen 2016; Kiros et al. 2015; Logeswaran and Lee 2018) and variants of Transformer (Reimers and Gurevych 2019; Li et al. 2020a; Su et al. 2021; Jiang et al. 2022). Contrastive sentence representation learning (Zhang et al. 2020; Kim, Yoo, and Lee 2021; Meng et al. 2021; Giorgi et al. 2021; Yan et al. 2021; Gao, Yao, and Chen 2021; Janson et al. 2021; Zhou et al. 2022; Zhang et al. 2022) has become the main trend in this field for its effectiveness. On the other hand, generative methods of learning high-quality sentence representations (Wang, Reimers, and Gurevych 2021; Wu and Zhao 2022) are less investigated.

## Self-Supervised Learning

Self-supervised learning is an ideal method for learning representations, owing to its intrinsic nature of not requiring any manual labels. It has been demonstrated to be effective across various modalities. (Devlin et al. 2018; Chen et al. 2020a; Schneider et al. 2019). There are principally two main branches of methods in self-supervised learning: **Contrastive learning** and **Generative learning** (Liu et al. 2021; Balestriero et al. 2023).

Contrastive learning (Sung et al. 2018) has been proven a promising approach in the field of sentence representation learning. The goal of contrastive learning is to pull semantically similar sentences closer together, while pushing dissimilar ones further apart in the representation space. For self-supervised contrastive learning, certain data augmentation strategies are necessary to form positive pairs, adhering to the principle of not using any labels. In the vision modality, methods such as cropping, resizing, rotation, and cutout are adopted to generate a positive sample from the input image. For contrastive sentence representation learning, ConSERT (Yan et al. 2021) employs strategies such as adversarial attacks, token shuffling, cutoff, and dropout on the token embedding matrix to create positive samples. Meanwhile, SimCSE enhances performance by passing the same sentence into the pre-trained language model twice, thereby forming positive pairs. Contrastive learning has also been adopted as a pre-training objective for sentence representation learning. (Wang et al. 2022b; Su et al. 2022)

Compared to contrastive learning, generative learning approaches are less investigated in the field of self-supervised sentence representation learning. Generative sentence representation learning attempts to generate original sentences from their corrupted or masked version (Yang et al. 2020; Wang, Reimers, and Gurevych 2021). Recently, PaSeR(Wu and Zhao 2022) was introduced, which auto-regressively generates important phrases from the original sentences; however, it necessitates the identification of these phrases beforehand.

## AutoEncoder

AutoEncoder (Kingma and Welling 2013) is a neural network architecture that is designed to learn a compressed and efficient representation of the input data and it consists of two main components: an encoder and a decoder. The encoder maps the input data to a lower-dimensional representation, known as the bottleneck or latent representation. The decoder then reconstructs the bottleneck representation back to the original input space. Same to contrastive learning, autoencoders can also be trained in a self-supervised manner.

# Methodology

Figure. 1 illustrates the overview of our proposed training framework, DenoSent. In this work, we aim to utilize intra-sentence supervision signals, using the original sentence as a guide. We achieve this by training an auto-encoder to reconstruct the original sentence from its noisy version. In our proposed training framework, the auto-encoder closely mirrors the architecture of the original sequence-to-sequence Transformer. However, in our implementation, the length of the encoded source sequence is constrained to 1 through pooling (detailed in the implementation section), serving as the sentence representation. The decoder component is utilized exclusively during training and is subsequently discarded for evaluation. We introduce perturbations to the sentences in both the discrete and continuous space, and train our model to restore them to their original form from the perturbed sentences and their corresponding representations. We empirically demonstrate that our proposed denoising objective operates orthogonally to the contrastive objective, allowing both objectives to be seamlessly integrated into our framework. Experimental results reveal that the amalgamation of both intra-sentence and inter-sentence supervision signals yields competitive results on a broad range of tasks.

## Turn Vanilla Transformer into a Sentence Representation Learner

The proposed denoising objective is both straightforward and efficacious. The following three modifications are made to the original Transformer (Vaswani et al. 2017) model to turn it into a sentence representation learner:

- Apply pooling strategies to reduce the length of the encoder output to 1, serving as the sentence representation, and seamlessly execute sequence-to-sequence learning.

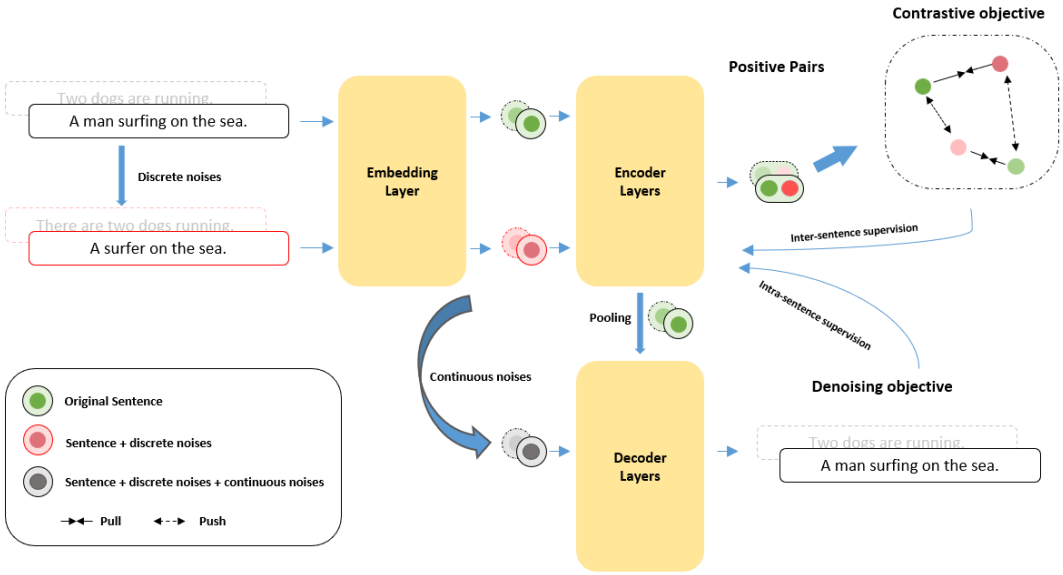



Figure 1: Overview of DenoSent. The proposed sentence representation learning framework is a combination of two objectives, providing both inter-sentence and intra-sentence supervision signals. Note that we use pooling strategies to downsize the encoder outputs from [n_tokens, hidden_dim] to [1, hidden_dim].

- Discard the multi-head attention technique in the decoder and use single-head attention instead.
- In the prediction stage, use a denoising strategy to predict the original sentence rather than the standard auto-regressive technique.

As a sequence-to-sequence model, the vanilla Transformer first encodes an input sequence of symbol representations $\{x_1, ..., x_{n_1}\}$ to a sequence of continuous representations $z_x = \{z_1, ..., z_{n_1}\}$ through self-attention layers and feed-forward layers, where $n_1$ denotes the input sequence length. The Transformer decoder accepts a shifted-right target sequence of symbol representations $\{\langle s\rangle, y_1, ..., y_{n_2-1}\}$, where $\langle s\rangle$ denotes a start token for a sequence and $n_2$ for target sequence length, then transforms it to continuous representations $z_y$, and predict the target sequence $\{y_1, ..., y_{n_2}\}$. In the Transformer decoder, there is an additional attention layer other than the self-attention layer and the feed-forward layer in each block, which performs cross-attention operations across $z_x$ and $z_y$:

$$CrossAttention(z_x, z_y) = softmax(\frac{z_y z_x^T}{\sqrt{d}})z_x \quad (1)$$

Denote $d$ as the number of the hidden dimensions thus $z_x \in \mathbb{R}^{n_1 \times d}$, $z_y \in \mathbb{R}^{n_2 \times d}$ in the original Transformer. In the context of prediction, the Transformer model utilizes an auto-regressive approach to generate each token, where each generated token is dependent on preceding tokens:

$$p(y) = \prod_{i=1}^{n} p(y_i|y_0, ..., y_{i-1}) \quad (2)$$

where $y_0$ denotes the start token $\langle s\rangle$.

In DenoSent, we employ pooling strategies on the encoder outputs to compress each sentence into a vector of a fixed-sized $d$. This can be alternatively viewed as reducing the input sequence length to 1, i.e., $z_x \in \mathbb{R}^{1 \times d}$ here. After introducing certain perturbations to the input sentence, we feed the perturbed sentence into the Transformer decoder. Our model is then trained to reconstruct the original input sentence using solely the encoded sentence representation. In the training process, $z_x$ obtains intra-sentence supervision signals in cross-attention operations(Eq. 1) in the decoder and is forced to capture more semantic information to help recover the original sentence from its noisy version. Unlike the vanilla Transformer, which applies a causal mask on the attention matrix to facilitate auto-regressive training, DenoSent aims to predict each input sequence token based on the entire noisy sentence:

$$p(x) = \prod_{i=1}^{n} p(x_i|\tilde{x}_1, ..., \tilde{x}_n) \quad (3)$$

where $\tilde{x}_i$ denotes the noisy version of token $x_i$.

Hence, let $\mathbf{S}$ be a corpus of sentences, the self-supervised denoising loss can be formed as:

$$\ell_{denoising} = -\sum_{s_i}^{S} \sum_{j=1}^{s_i} logP(t_j|\tilde{t}_1, ..., \tilde{t}_{s_i}; \Theta) \quad (4)$$

Here, $t_j$ represents the original token, while $\tilde{t}_j$ denotes its noisy counterpart; $\Theta$ symbolizes the parameters of the model. The introduction of noise is detailed in the following subsection.


Figure 2: The two-stage perturbation process wherein both discrete and continuous noises are sequentially incorporated into the original sentences. The discrete perturbation is achieved through back-translation or the use of a large language model (LLM), while the continuous perturbation is implemented by applying substantial dropout on the embedded sentences.

## The Perturbed Sentences: Discrete Noises and Continuous Noises

Previous contrastive sentence representation learning techniques have employed a variety of data augmentation methods to construct positive pairs for contrastive learning. In the process, such operations introduce both discrete (e.g., token shuffling, token cutoff, *inter alia*) and continuous (e.g., adversarial attack, dropout, *inter alia*) noises to the original sentences, which enhance the generalization and alignment capabilities of the sentence encoder (Yan et al. 2021; Gao, Yao, and Chen 2021). In this work, we propose a two-stage perturbation strategy that integrates discrete noises and continuous noises sequentially(Figure. 2). These perturbations facilitate the generation of noisy input sentences, enabling us to train our sentence representation learner using the proposed denoising objective.

**Discrete Noises** Discrete noises are introduced directly at the token level, resulting in a sequence of tokens $\{\tilde{x}_1, ..., \tilde{x}_{n'}\}$ derived from the original sequence $\{x_1, ..., x_n\}$. Simple token manipulations, such as deletion, swapping, or shuffling, have been shown to adversely affect performance, as they can disrupt the original semantics of the sentence. (Yan et al. 2021; Gao, Yao, and Chen 2021) Here we propose to use two off-the-shelf data augmentation strategies to provide discrete noises, without compromising the inherent semantics of the original sentence. Specifically, we achieve this by leveraging the back-translation technique or a large language model (LLM) to rewrite the sentences. Machine translation aims to preserve original semantics in another language. By translating and back-translating sentences, we can obtain augmented sentences with similar semantics but varied syntax and expression. LLMs, on the other hand, can generate text based on the user's input and instructions after instruction fine-tuning (Ouyang et al. 2022; Wei et al. 2022). Cheng et al. have demonstrated that it is possible to generate sentence similarity labels for use in contrastive learning training, highlighting the ability of LLMs to capture sentence semantics. In our work, we exclusively use LLMs to rewrite the original sentences, introducing noise while preserving the underlying semantics. In practice, we utilize the pre-trained translation models for translation purposes, and OpenAI `gpt-3.5-turbo` for the instruction-following LLM. In our experiments, we discovered that employing the back-translation strategy results in marginally superior performance compared to using an LLM(See Table. 3). Consequently, we adopt back-translation as the default strategy for incorporating discrete noises in the rest of the literature.

**Continuous Noises** The introduction of continuous noises plays a crucial role in our proposed denoising objective, as it offers much greater control over the level of introduced noises within the continuous space. In our training framework, we employ dropout (Srivastava et al. 2014) at a substantial rate on the embedded sentences, setting most of the elements of the decoder input to zero. We subsequently train our model to reconstruct the sentence from the heavily corrupted input, drawing upon the output from the encoder, which serves as the sentence representation. This approach compels the model to retain sufficient semantic information in the encoded representation to facilitate the restoration of the original sentence. The level of noise introduced can be controlled by the dropout rate, which determines the difficulty of the learning task.

## Combine with Contrastive Learning

As the main trend in self-supervised sentence representation learning, contrastive learning (Chen et al. 2020b) has been proven effective in previous works (Gao, Yao, and Chen 2021). The contrastive objective provides inter-sentence supervision signals by learning one sentence's representation from other sentences. Specifically, given a sentence $s$, a semantic-related positive example $s^+$ and a set of negative examples $s^-$ are needed to perform contrastive learning. Formally, denote $z$, $z^+$ and $z^-$ as the representation of $s$, $s^+$ and $s^-$, respectively, contrastive-learning-based methods utilize the InfoNCE (Oord, Li, and Vinyals 2018) loss:

$$\ell_{contrastive} = -log\frac{e^{sim(z,z^+)/\tau}}{\sum_{i=1}^{N} e^{sim(z,z_i^-)/\tau}} \quad (5)$$

where $\tau$ denotes the temperature hyperparameter, $N$ is the number of negative samples for each training sample, and $sim$ for cosine similarity.

Unlike contrastive learning, the proposed denoising objective (as described in Eq. 4) offers intra-sentence supervision signals by learning the representation directly from the sentence. Therefore, the denoising objective works independently from previous contrastive methods. Both objectives can be readily integrated:

$$\ell = \ell_{contrastive} + \ell_{denoising} \quad (6)$$

For the contrastive objective, we add discrete perturbations to construct $s^+$ and in-batch negative samples $s^-$ for training. We reach our final results by optimizing Eq. 6.

# Experiment

In our study, we evaluate the effectiveness of DenoSent on a variety of sentence-level tasks, including semantic textual similarity (STS), reranking, retrieval, and classification. To assess performance on STS tasks, we employed the SentEval toolkit (Conneau and Kiela 2018), in line with previous research. The remaining tasks were evaluated using the Massive Text Embedding Benchmark (MTEB) toolkit (Muennighoff et al. 2022).

## Datasets

**Semantic textual similarity tasks.** STS tasks assess sentence similarity. Given a sentence pair, the similarity score is calculated based on the model-generated sentence representations, which is then compared to human-annotated similarity. We evaluate DenoSent on 7 STS tasks: **STS 2012–2016**(Agirre et al. 2012, 2013, 2014, 2015, 2016), **STS Benchmark**(Cer et al. 2017) and **SICK-Relatedness** (Marelli et al. 2014) using the SentEval toolkit(Conneau and Kiela 2018), following previous research. Spearman correlation based on cosine similarity is reported as the main metric (Reimers, Beyer, and Gurevych 2016).

**Reranking & Retrieval tasks.** For reranking tasks, the model generates sentence representations for a given query and a set of reference sentences (relevant and irrelevant), and ranks the references based on their similarity to the query representation. Retrieval tasks, similar to reranking tasks, involve the model embedding a query and documents in a corpus, and ranking the documents by similarity. We evaluate DenoSent on 4 reranking tasks: **AskUbuntuDupQuestions** (Lei et al. 2015), **MindSmallReranking** (Wu et al. 2020), **SciDocsRR** (Cohan et al. 2020), and **StackOverflowDupQuestions** (Liu et al. 2018), and a retrieval task: **QuoraRetrieval** (Thakur et al. 2021). We report the mean MRR@1 and MAP@1 as the main results.

**Classification tasks.** For classification tasks, each sentence in the datasets has a corresponding label. Sentence representations are obtained by the provided model and an extra logistic regression classifier is trained on these representations and their corresponding label. We evaluate DenoSent on 10 classification tasks: **AmazonCounterfactual** (O'Neill et al. 2021), **AmazonReviews** (McAuley and Leskovec 2013), **Banking77** (Casanueva et al. 2020), **Emotion** (Saravia et al. 2018), **MassiveIntent** (FitzGerald et al. 2022), **MassiveScenario** (FitzGerald et al. 2022), **MTOPDomain** (Li et al. 2020b), **MTOPIntent** (Li et al. 2020b), **ToxicConversations** (Kaggle 2019), **TweetSentimentExtraction** (Kaggle 2020). We report the classification accuracy as the main metric.

## Baselines

In this study, the proposed method was evaluated and compared to the following established methods.

**Glove** (Pennington, Socher, and Manning 2014) takes the Glove embedding of each word in the sentence as the sentence's representation. **InferSent** (Conneau et al. 2017) uses Glove with some signal enhancement and is trained further on the NLI dataset. **Universal Sentence Encoder** (Cer et al. 2018) uses the Transformer model and learns the objective of reconstructing surrounding sentences in a paragraph. **BERT(CLS, Mean, First-Last Avg.)** (Devlin et al. 2018) directly utilizes BERT's outputs as sentence representations, using different pooling strategies. **BERT-Flow** (Li et al. 2020a) reversibly maps the BERT output space from a cone to the standard Gaussian distribution space. **BERT-Whitening** (Su et al. 2021) improves the quality of sentence representation by simple vector whitening. **ConSERT** (Yan et al. 2021) and **SimCSE**(Gao, Yao, and Chen 2021) is based on contrastive learning and uses different data augmentation strategies to construct positive sentence pairs. **DCLR** (Zhou et al. 2022) uses an instance weighting strategy to alleviate the false-negative problem in contrastive learning. **DiffCSE** (Chuang et al. 2022) is optimized on SimCSE to improve the effectiveness of the sentence representation model using forged samples. **PromptBERT** (Jiang et al. 2022) uses prompts to generate sentence representations. **SNCSE** (Wang et al. 2022a) is a contrastive learning method based on soft negative examples. **CMLM** (Yang et al. 2021) incorporates the learning of sentence representation into MLM training. **PaSeR** (Wu and Zhao 2022) proposed an intra-sentence objective that learns sentence representation by utilizing the encoded sentence representation to predict masked phrases in the input sentence.

## Implementation Details

For the implementation of the proposed method, we use pretrained *bert-base-uncased* as the encoder and randomly initialized transformer layers as the decoder for all our experiments. We use the unsupervised Wiki dataset used in SimCSE (Gao, Yao, and Chen 2021) as our self-supervised training dataset. For back translation data augmentation, we use pre-trained machine translation models (Tiedemann and Thottingal 2020) to translate the training sentences to Chinese and then translate them back to English. We use a learning rate of 5e-5 and AdamW (Loshchilov and Hutter 2017) as the optimizer. For the input sequence length, we use a value of 32. For the denoising objective, we use $\{0.8, 0.825, 0.85, 0.875, 0.9\}$ as the dropout rates for continuous perturbations, $\{12, 14, 16\}$ as the number of decoder transformer layers and perform a sweep on these parameters then select the checkpoint that has the highest spearman correlation on the STS-Benchmark development set for evaluation. We use 0.825 as the dropout rate and 16 transformer layers for reported results. For the contrastive objective, we use a temperature $\tau = 0.03$. For the pooling strategy, we fit every sentence with the same template "[X] means [MASK]." and use the encoder output vector of the [MASK] token as the sentence representation through all our experiments. We conduct all the experiments on a machine with 8 NVIDIA GeForce RTX 3090 GPUs.

| Model | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg. |
|---|---|---|---|---|---|---|---|---|
| *Non-BERT Models* | | | | | | | | |
| GloVe embeddings (avg.)♣ | 55.14 | 70.66 | 59.73 | 68.25 | 63.66 | 58.02 | 53.76 | 61.32 |
| InferSent-GloVe♣ | 52.86 | 66.75 | 62.15 | 72.77 | 66.87 | 68.03 | 65.65 | 65.01 |
| Universal Sentence Encoder♣ | 64.49 | 67.80 | 64.61 | 76.83 | 73.18 | 74.92 | 76.69 | 71.22 |
| *BERT&Post-Processing Models* | | | | | | | | |
| BERTbase (CLS)■ | 21.54 | 32.11 | 21.28 | 37.89 | 44.24 | 20.30 | 42.42 | 31.40 |
| BERTbase (Mean)■ | 30.87 | 59.89 | 47.73 | 60.29 | 63.73 | 47.29 | 58.22 | 52.57 |
| BERTbase (first-last avg.)■ | 39.70 | 59.38 | 49.67 | 66.03 | 66.19 | 53.87 | 62.06 | 56.70 |
| BERTbase-flow♣ | 58.40 | 67.10 | 60.85 | 75.16 | 71.22 | 68.66 | 64.47 | 66.55 |
| BERTbase-whitening♣ | 57.83 | 66.90 | 60.90 | 75.08 | 71.31 | 68.24 | 63.73 | 66.28 |
| *Contrastive-based Models* | | | | | | | | |
| ConSERT-BERTbase♡ | 64.64 | 78.49 | 69.07 | 79.72 | 75.95 | 73.97 | 67.31 | 72.74 |
| SimCSE-BERTbase♣ | 68.40 | 82.41 | 74.38 | 80.91 | 78.56 | 76.85 | 72.23 | 76.25 |
| DCLR-BERTbase■ | 70.81 | 83.73 | 75.11 | 82.56 | 78.44 | 78.31 | 71.59 | 77.22 |
| DiffCSE-BERTbase♢ | 72.28 | 84.43 | 76.47 | 83.9 | **80.54** | 80.59 | 71.23 | 78.49 |
| PromptBERT-BERTbase♡ | 71.56 | <u>84.58</u> | 76.98 | **84.47** | 80.6 | **81.6** | 69.87 | 78.54 |
| SNCSE-BERTbase♠ | 70.67 | **84.79** | <u>76.99</u> | 83.69 | <u>80.51</u> | <u>81.35</u> | **74.77** | <u>78.97</u> |
| DenoSent-BERTbase(contrastive only) | <u>73.09</u> | 82.19 | 75.56 | 83.51 | 79.38 | 80.10 | 71.86 | 77.96 |
| *Generative-based Models* | | | | | | | | |
| CMLM-BERTbase♦ | 58.20 | 61.07 | 61.67 | 73.32 | 74.88 | 76.60 | 64.80 | 67.22 |
| PaSeR-BERTbase♦ | 70.21 | 83.88 | 73.06 | 83.87 | 77.60 | 79.19 | 65.31 | 76.16 |
| DenoSent-BERTbase(generative only) | 69.50 | 83.83 | 75.09 | 82.78 | 77.75 | 77.59 | 66.78 | 76.19 |
| *Generative+Contrastive Models* | | | | | | | | |
| **DenoSent-BERTbase** | **75.57** | 83.77 | **77.24** | <u>84.30</u> | 79.51 | 80.81 | <u>74.09</u> | **79.33** |

Table 1: Evaluation performance on 7 STS tasks. The reported metric is spearman correlation($\times 100$) based on cosine similarity following previous works. Bolded results and underlined results correspond to the best and second-best results in the same dataset, respectively. ♣: results from Gao, Yao, and Chen 2021. ■: results from Zhou et al. 2022. ♢: results from Chuang et al. 2022. ♡: results from Jiang et al. 2022. ♠: results from Wang et al. 2022a. ♦: results from Wu and Zhao 2022.

## Main Results

Table 1 illustrates the performance of DenoSent on 7 STS tasks compared to previous methods. All experiments are conducted under a self-supervised/unsupervised setting except for non-BERT models. The results reveal that methods that either do not use a PLM or rely solely on post-processing are less effective than those that apply contrastive and generative approaches on a PLM. In the context of single-objective learning, the contrastive objective proves to be more effective for semantic textual similarity tasks than the generative objective since it directly optimizes representation similarities. The proposed denoising objective shows competitive performance compared to contrastive methods despite the fact that it is completely complementary to them. The utilization of the contrastive objective alone in the DenoSent model resulted in a 1.71% absolute improvement in performance compared to the SimCSE model. This demonstrates the effectiveness of incorporating discrete noises and the [MASK] token pooling strategy, as the contrastive DenoSent model is identical to the SimCSE model in all other aspects. The proposed framework effectively integrates both inter-sentence and intra-sentence objectives, resulting in superior performance on STS tasks.

| Model | Avg. Classification Accuracy |
|---|---|
| Glove | 56.42 |
| BERT(CLS) | 60.32 |
| SimCSE | 62.73 |
| PaSeR | 63,23 |
| PromptBERT | 63.78 |
| SNCSE | 62.82 |
| **DenoSent** | **64.46** |

Table 2: Evaluation performance on classification tasks.

In order to assess the generalizability of DenoSent, a comprehensive set of experiments was conducted on reranking, retrieval and classification tasks. The results, as illustrated in Figure 3, demonstrate that DenoSent consistently outperforms SimCSE on reranking and retrieval tasks, and exhibits a higher degree of robustness across various tasks and domains compared to other baselines. Table 2 presents the evaluation results for the average accuracy across 10 sentence-level classification tasks. The results indicate that DenoSent exhibits the highest performance on classification tasks, demonstrating its strong capability for generalization.


Figure 3: Absolute performance difference on reranking and retrieval tasks compared to SimCSE. AUD, MS, SciD, SODQ and QR denotes AskUbuntuDupQuestions, MindSmallReranking, SciDocsRR, StackOverflowDupQuestions and QuoraRetrieval, respectively.

The results of these tasks indicate that utilizing both intra-sentence and inter-sentence objectives not only improves performance on STS tasks, but also leads to enhancements in the overall generalizability.

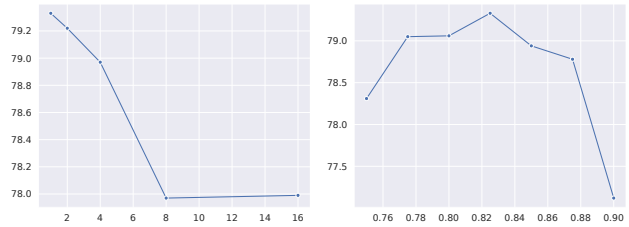
| Model | Avg. STS |
|---|---|
| **DenoSent** | **79.33** |
| DenoSent$^{\diamondsuit}$ | 78.98 |
| w/o *denoising* | 77.96 |
| w/o *contrastive* | 76.19 |
| w/o *discrete noise* | 77.99 |
| w/o *denoising* +*discrete noise* | 76.17 |
| w/o *contrastive* +*discrete noise* | 74.54 |
| w/ *CLS Pooling* | 78.66 |

Table 3: Ablation on the components in DenoSent. $\diamondsuit$ denotes using an LLM to introduce discrete noise.

## Ablation Study

**Effects of proposed components.** In Table 3, we investigate the impacts of different proposed components in the DenoSent framework. The utilization of both contrastive and denoising objectives has been demonstrated to be crucial for achieving high performance. The combination of these objectives results in a significant improvement in performance. Additionally, the incorporation of discrete noises has been found to enhance performance for both objectives consistently. The utilization of [MASK] token pooling, instead of [CLS] pooling, has also been shown to provide a slight boost in performance, as previously reported in Jiang et al. 2022.

**Effects of different number of attention heads in the decoder.** For the denoising objective, we use single-head attention instead multi-attention in our experiments. The results, depicted in Figure 4a, indicate that an increase in the number of attention heads results in a decrease in performance. This may be due to the fact that the multi-head attention technique enhances transformer models by offering multiple perspectives on attention. However, in the case of the DenoSent decoder, the memory input sequence length for the transformer layers is fixed at 1, rendering the utilization of multiple attention heads redundant. On the other hand, an increasing number of attention heads results in a reduction of the dimensionality of the sentence representation during computation. This decrease in dimensionality impairs the representation capabilities of the model, thereby leading to a decline in performance.



(a) Impact of attention heads. (b) Impact of dropout rates.

Figure 4: Average STS performance using different numbers of attention heads and dropout rates.

**Effects of the continuous noise level.** In the proposed method DenoSent, we employ dropout as a technique for introducing controlled corruption to sentences in the continuous space. The dropout rate is used to define the level of corruption added to the sentence. It is crucial that the injected noise is substantial enough to render the learning task sufficiently challenging, thus enabling our model to learn meaningful semantic information in sentence representations. As illustrated in Figure 4b, the performance of the model is sensitive to the choice of dropout rate, with optimal results observed for moderate values. If the value is set too high, the input becomes excessively corrupted, rendering the task overly challenging and impeding the model's learning capability. Conversely, if the level of corruption is too low, the denoising task becomes overly simplistic, preventing the model from effectively leveraging the semantic information embedded in the sentence representation.

## Conclusion

In this work, we introduce DenoSent, a self-supervised sentence representation learning framework that incorporates both intra-sentence and inter-sentence objectives. We propose a novel denoising objective that uses sentence representation to recover a noisy sentence input to its original. We introduce both discrete and continuous noises to perturb the input sentence to facilitate our denoising objective. Furthermore, we combine the denoising objective with the contrastive objective, allowing representations to benefit from both intra-sentence and inter-sentence supervision. We evaluate our model on numerous tasks ranging from semantic textual similarity, reranking, retrieval and classification, showing superior performance and generalization ability.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62236004 and No. 62022027). The authors would like to thank the anonymous reviewers for their comprehensive and insightful reviews and suggestions.

## References

Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W.; Lopez-Gazpio, I.; Maritxalar, M.; Mihalcea, R.; Rigau, G.; Uria, L.; and Wiebe, J. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability.

Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W.; Mihalcea, R.; Rigau, G.; and Wiebe, J. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity.

Agirre, E.; Banea, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Mihalcea, R.; Rigau, G.; and Wiebe, J. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation.

Agirre, E.; Cer, D.; Diab, M.; and Gonzalez-Agirre, A. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity.

Agirre, E.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; and Guo, W. 2013. *SEM 2013 shared task: Semantic Textual Similarity.

Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; Schwarzschild, A.; Wilson, A. G.; Geiping, J.; Garrido, Q.; Fernandez, P.; Bar, A.; Pirsiavash, H.; LeCun, Y.; and Goldblum, M. 2023. A Cookbook of Self-Supervised Learning. arXiv:2304.12210.

Casanueva, I.; Temčinas, T.; Gerz, D.; Henderson, M.; and Vulić, I. 2020. Efficient Intent Detection with Dual Sentence Encoders.

Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; St. John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; Strope, B.; and Kurzweil, R. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A Simple Framework for Contrastive Learning of Visual Representations.

Cheng, Q.; Yang, X.; Sun, T.; Li, L.; and Qiu, X. 2023. Improving Contrastive Learning of Sentence Embeddings from AI Feedback. arXiv:2305.01918.

Chuang, Y.-S.; Dangovski, R.; Luo, H.; Zhang, Y.; Chang, S.; Soljačić, M.; Li, S.-W.; tau Yih, W.; Kim, Y.; and Glass, J. 2022. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. arXiv:2204.10298.

Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; and Weld, D. S. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers.

Conneau, A.; and Kiela, D. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. *arXiv preprint arXiv:1803.05449*.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. arXiv:1705.02364.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ethayarajh, K. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

FitzGerald, J.; Hench, C.; Peris, C.; Mackie, S.; Rottmann, K.; Sanchez, A.; Nash, A.; Urbach, L.; Kakarala, V.; Singh, R.; Ranganath, S.; Crist, L.; Britan, M.; Leeuwis, W.; Tur, G.; and Natarajan, P. 2022. MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Giorgi, J.; Nitski, O.; Wang, B.; and Bader, G. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. In *Meeting of the Association for Computational Linguistics*.

Hill, F.; Cho, K.; and Korhonen, A. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

Janson, S.; Gogoulou, E.; Ylipää, E.; Cuba Gyllensten, A.; and Sahlgren, M. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations, 2021*.

Jiang, T.; Huang, S.; Zhang, Z.; Wang, D.; Zhuang, F.; Wei, F.; Huang, H.; Zhang, L.; and Zhang, Q. 2022. PromptBERT: Improving BERT Sentence Embeddings with Prompts. *arXiv preprint arXiv:2201.04337*.

Kaggle. 2019. ToxicConversations.

Kaggle. 2020. TweetSentimentExtraction.

Kim, T.; Yoo, K. M.; and Lee, S.-g. 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)*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. *Advances in neural information processing systems*.

Lei, T.; Joshi, H.; Barzilay, R.; Jaakkola, T.; Tymoshenko, K.; Moschitti, A.; and Marquez, L. 2015. Semi-supervised Question Retrieval with Gated Convolutions.

Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; and Li, L. 2020a. On the Sentence Embeddings from Pre-trained Language Models. arXiv:2011.05864.

Li, H.; Arora, A.; Chen, S.; Gupta, A.; Gupta, S.; and Mehdad, Y. 2020b. MTOP: A Comprehensive Multilingual Task-Oriented Semantic Parsing Benchmark.

Liu, X.; Wang, C.; Leng, Y.; and Zhai, C. 2018. LinkSO: A Dataset for Learning to Retrieve Similar Question Answer Pairs on Software Development Forums. In *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering*.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2021. Self-supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Logeswaran, L.; and Lee, H. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization.

Marelli, M.; Menini, S.; Baroni, M.; Bentivogli, L.; Bernardi, R.; and Zamparelli, R. 2014. A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

McAuley, J.; and Leskovec, J. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems*.

Meng, Y.; Xiong, C.; Bajaj, P.; Tiwary, S.; Bennett, P.; Han, J.; and Song, X. 2021. COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining. arXiv:2102.08473.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2022. MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316*.

O'Neill, J.; Rozenshtein, P.; Kiryo, R.; Kubota, M.; and Bollegala, D. 2021. I Wish I Would Have Loved This One, But I Didn't – A Multilingual Dataset for Counterfactual Detection in Product Reviews.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Reimers, N.; Beyer, P.; and Gurevych, I. 2016. Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Saravia, E.; Liu, H.-C. T.; Huang, Y.-H.; Wu, J.; and Chen, Y.-S. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised Pre-training for Speech Recognition. arXiv:1904.05862.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*.

Su, H.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.-t.; Smith, N. A.; Zettlemoyer, L.; Yu, T.; et al. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Su, J.; Cao, J.; Liu, W.; and Ou, Y. 2021. Whitening Sentence Representations for Better Semantics and Faster Retrieval. arXiv:2103.15316.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; and Gurevych, I. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Tiedemann, J.; and Thottingal, S. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Li, Y.; Huang, Z.; Dou, Y.; Kong, L.; and Shao, J. 2022a. SNCSE: Contrastive Learning for Unsupervised Sentence Embedding with Soft Negative Samples. arXiv:2201.05979.

Wang, K.; Reimers, N.; and Gurevych, I. 2021. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. arXiv:2104.06979.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652.

Wu, B.; and Zhao, H. 2022. Sentence Representation Learning with Generative Objective rather than Contrastive Objective. arXiv:2210.08474.

Wu, F.; Qiao, Y.; Chen, J.-H.; Wu, C.; Qi, T.; Lian, J.; Liu, D.; Xie, X.; Gao, J.; Wu, W.; and Zhou, M. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; and Xu, W. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Yang, Z.; Yang, Y.; Cer, D.; Law, J.; and Darve, E. 2020. Universal sentence representation learning with conditional masked language model. *arXiv preprint arXiv:2012.14388*.

Yang, Z.; Yang, Y.; Cer, D.; Law, J.; and Darve, E. 2021. Universal Sentence Representation Learning with Conditional Masked Language Model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Zhang, Y.; He, R.; Liu, Z.; Lim, K. H.; and Bing, L. 2020. An Unsupervised Sentence Embedding Method by Mutual Information Maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhang, Y.; Zhu, H.; Wang, Y.; Xu, N.; Li, X.; and Zhao, B. 2022. A Contrastive Framework for Learning Sentence Representations from Pairwise and Triple-wise Perspective in Angular Space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Zhou, K.; Zhang, B.; Zhao, W. X.; and Wen, J.-R. 2022. Debiased Contrastive Learning of Unsupervised Sentence Representations. *arXiv preprint arXiv:2205.00656*.