A Contrastive Framework for Learning Sentence Representations from Pairwise and Triple-wise Perspective in Angular Space

Yuhao Zhang[1], Hongji Zhu[1], Yongliang Wang[1], Nan Xu[2], Xiaobo Li[1], BinQiang Zhao[1]

[1]Alibaba Group
[2]Institute of Automation, Chinese Academy of Sciences
{zyh277500,zhj283587,wangyongliang.wyl,xiaobo.lixb,binqiang.zhao}@alibaba-inc.com
xunan2015@ia.ac.cn

Abstract

Learning high-quality sentence representations is a fundamental problem of natural language processing which could benefit a wide range of downstream tasks. Though the BERT-like pre-trained language models have achieved great success, using their sentence representations directly often results in poor performance on the semantic textual similarity task. Recently, several contrastive learning methods have been proposed for learning sentence representations and have shown promising results. However, most of them focus on the constitution of positive and negative representation pairs and pay little attention to the training objective like NT-Xent, which is not sufficient enough to acquire the discriminating power and is unable to model the partial order of semantics between sentences. So in this paper, we propose a new method ArcCSE, with training objectives designed to enhance the pairwise discriminative power and model the entailment relation of triplet sentences. We conduct extensive experiments which demonstrate that our approach outperforms the previous state-of-the-art on diverse sentence related tasks, including STS and SentEval.

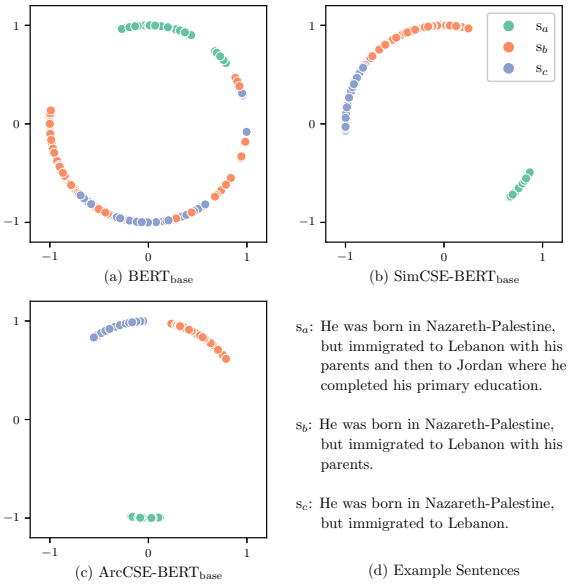


Figure 1: Sentence representations visualization. We generate the representations of three related sentences by passing them to BERT$_{\text{base}}$, SimCSE-BERT$_{\text{base}}$ and ArcCSE-BERT$_{\text{base}}$ multiple times. With different dropout masks, we can generate different representations for each sentence. Then we normalize the embeddings and use t-SNE for dimensionality reduction.

1 Introduction

Learning sentence representations, which encodes sentences into fixed-sized dense vectors such that semantically similar ones stay close, is a fundamental problem of natural language processing. It could benefit a wide range of downstream applications such as information retrieval, semantic similarity comparison, question answering, and so on.

Recently, with the great success of pre-trained Transformer-based language models (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020; Brown et al., 2020; Liu et al., 2019) like BERT, they have been widely adopted for generating sentence representations. A straightforward way is by leveraging the [CLS] embedding (Devlin et al., 2019) or applying mean pooling on the last layers of a BERT-like pre-trained language model (Reimers and Gurevych, 2019). However, the sentence embeddings coming from a pre-trained language model without further fine-tuning could not capture the semantic meaning of sentences very well as shown in Figure 1(a), and sometimes even underperform non-contextualized embeddings like GloVe (Pennington et al., 2014).

To make pre-trained language models more suitable for generating sentence embeddings, supervised methods like SBERT (Reimers and Gurevych, 2019) are proposed, which improve the performance by fine-tuning on a labeled dataset. As labeled data is not available or expensive to annotate in many tasks or domains, it is of great value

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4892 - 4903

for developing unsupervised/self-supervised approaches for learning sentence representations. So recent works like BERT-Flow (Li et al., 2020) and BERT-Whitening (Su et al., 2021) propose post-processing methods to improve the BERT-based sentence representation. They address that the non-smooth anisotropic semantic space of BERT is a bottleneck and alleviate the problem through normalizing flows and whitening operation. To further improve the quality of sentence representations, several works (Kim et al., 2021; Yan et al., 2021; Giorgi et al., 2021; Carlsson et al., 2021; Gao et al., 2021) adopt self-supervised contrastive learning approach, which learns sentence representations by minimizing the distance of positive sentence representation pairs and maximizing the distance of negative pairs. In these works, positive pairs are often constituted through data augmentation or encoders with different structure or parameters, while negative pairs are derived from different sentences within the same batch. Then contrastive learning objective like normalized temperature-scaled cross-entropy loss (NT-Xent) (Chen et al., 2020; Gao et al., 2021) is used for optimizing. A typical example unsup-SimCSE (Gao et al., 2021) achieves state-of-the-art performance with a simple and effective idea of using standard dropout for data augmentation.

Though existing contrastive methods for learning sentence representation have shown promising results, most of them focus on the positive and negative pairs constitution, and the optimization objective itself is not fully exploited. The contrastive learning objective NT-Xent loss used in recent works (Yan et al., 2021; Giorgi et al., 2021; Gao et al., 2021) is a variation of cross-entropy loss with softmax function. Recent studies (Wang et al., 2018; Deng et al., 2019) have shown that the traditional softmax-based loss is insufficient to acquire the discriminating power, as shown in Figure 1(b) in which SimCSE-BERT$_{\text{base}}$ adopts the NT-Xent loss and could not separate s_b and s_c completely. In addition, the current optimization objectives only models sentence relations in a pairwise perspective, which tries to pull sentences with similar semantics closer and push dissimilar ones away from each other. However, there are different degrees of semantic similarity among related sentences. For example in Figure 1(d), s_b is more similar to s_a than s_c is. The current optimization objectives lack the ability to model the partial order of semantics between sentences.

To alleviate these problems, in this paper, we propose a new approach ArcCSE for sentence representation learning. For pairwise sentence relation modeling, we propose Additive Angular Margin Contrastive Loss (ArcCon Loss), which enhances the pairwise discriminative power by maximizing the decision margin in the angular space. Besides, in order to model the partial order of semantics between sentences, we propose a new self-supervised task that captures the entailment relation among triplet sentences. The task is implemented through automatically constituted triplet sentences with entailment relation among them. A visualization example of the generated representations through ArcCSE is shown in Figure 1(c). We evaluate our method on standard semantic textual similarity (STS) tasks and SentEval transfer tasks, and it outperforms the previous state-of-the-art approaches.

2 Related Work

2.1 Unsupervised Sentence Representation Learning

Early works usually learn sentence representations by augmenting the idea of word2vec (Mikolov et al., 2013), such as predicting surrounding sentences (Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018) or summing up n-gram embeddings (Pagliardini et al., 2018). With the rise of pre-trained language models, many works try to generate sentence representations through BERT-like models. A common way is leveraging the [CLS] embedding or applying mean pooling on the last layers of BERT (Reimers and Gurevych, 2019; Li et al., 2020). Instead of using BERT embeddings directly, BERT-Flow (Li et al., 2020) and BERT-Whitening (Su et al., 2021) further improve sentence representation through post-processing.

Recently, several works adopt the contrastive learning framework for sentence representation learning. They propose different strategies to constitute contrastive pairs, either through different data transforming methods (Zhang et al., 2020; Yan et al., 2021; Giorgi et al., 2021), or through encoders with different structures or parameters (Carlsson et al., 2021; Kim et al., 2021; Gao et al., 2021). A typical example SimCSE (Gao et al., 2021) uses dropout as data augmentation strategy and achieves state-of-the-art performance. However, most existing works pay little attention to the training objective and use the traditional contrastive

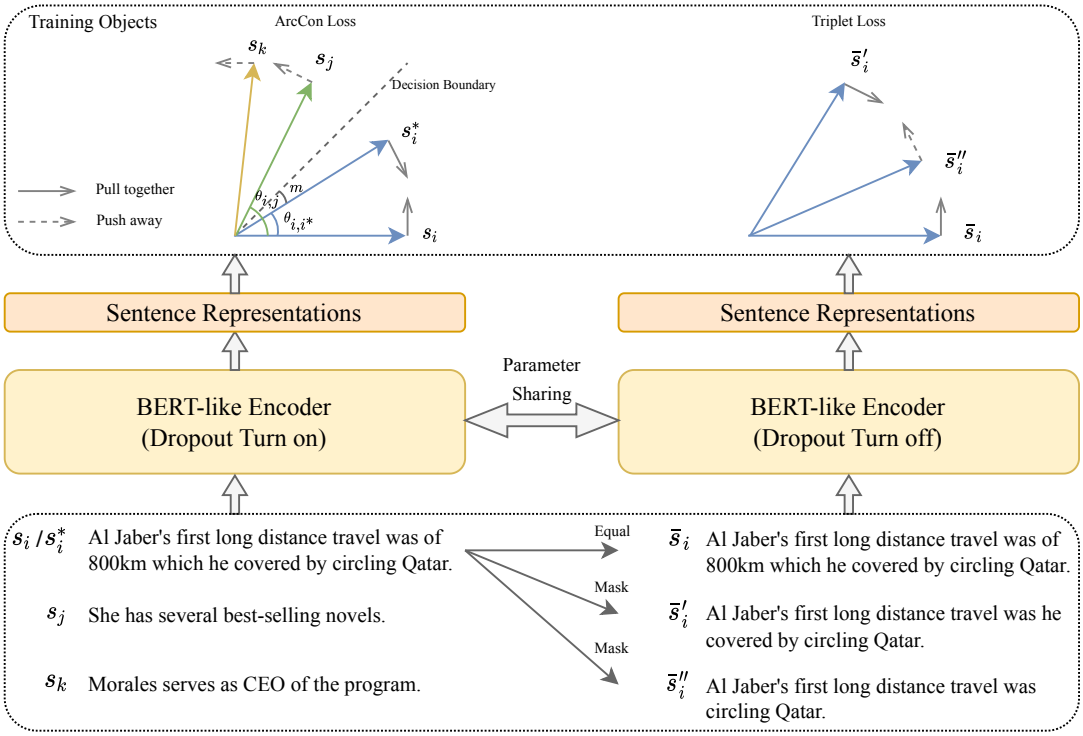


Figure 2: The framework of ArcCSE. ArcCSE models pairwise and triple-wise sentence relations simultaneously. For pairwise sentence relation modeling, we pass sentences to a BERT-like encoder with dropout turn on twice. Then we feed the representations into ArcCon loss which is more discriminative than NT-Xent loss. Triplet sentences are constituted through masking. We pass them to the same BERT-like encoder with dropout turn off and use a triplet loss to model their relations.

loss directly, which is insufficient in discrimination and unable to model the partial order of semantics between sentences. So, in our work, we propose a new approach that jointly models the pairwise and triple-wise sentence relations and further improves the sentence representations' quality.

2.2 Deep Metric Learning Objectives

The goal of Deep Metric Learning (DML) is to learn a function that maps objects into an embedded space, in which similar objects stay close and dissimilar ones are far away. In order to achieve this goal, many approaches have been proposed, and designing appropriate loss functions plays a key role in it. Contrastive training objectives like Contrastive Loss (Chopra et al., 2005), N-Pair Loss (Sohn, 2016), Structured Loss (Song et al., 2016) and Triplet Margin Loss (Ma et al., 2021) apply the definition of metric learning directly. These objectives are among the earliest training objectives used for deep metric learning. Later, softmax-based losses which learn a center for each class and penalize the distances between deep features and their corresponding class centers achieve more promising results in supervised metric learning. Typical examples like Center Loss (Wen et al., 2016), SphereFace (Liu et al., 2017), CosFace (Wang et al., 2018) and ArcFace (Deng et al., 2019) are widely adopted in deep learning applications such as face recognition and sentence classification (Coria et al., 2020). However, these losses need class labels and are not suitable for learning sentence representations. So inspired by ArcFace, we propose a new training objective ArcCon that does not need class labels and can model pairwise sentence relations with more discriminative power than traditional contrastive training objectives.

3 Method

In this section, we present ArcCSE, an angular based contrastive sentence representation learning framework, which could generate superior sentence embeddings from unlabeled data. Given a pre-trained language model $\mathcal{M}$ and an unlabeled text dataset $\mathcal{D}$, the task is fine-tuning $\mathcal{M}$ on $\mathcal{D}$ so that the sentence representations generated through $\mathcal{M}$ could be more semantic discriminative.

Our framework consists of two components that

model pairwise and triple-wise sentence relations simultaneously, as shown in Figure 2. We start with angular margin based contrastive learning in Section 3.1, which models pairwise relations between sentences by pulling semantic similar ones closer while pushing dissimilar ones away. Then we introduce the method which models the partial order of semantics between automatically constituted triplet sentences in Section 3.2.

3.1 Angular Margin based Contrastive Learning

To model the positive/negative pairwise relations between sentences, we first need to generate sentence representations and group them into positive and negative pairs. Then we feed these pairs to a training objective for optimizing.

Given a collection of sentences $\mathcal{D} = \{s_i\}_{i=1}^{N}$, we generate the sentence representations through a BERT-like pre-trained language model $\mathcal{M}$. Following SimCSE, we use dropout as the data augmentation method. For each sentence s_i, we generate two different representations h_i and h_i^* from s_i by passing s_i to $\mathcal{M}$ twice with independently sampled dropout masks. These two representations with the same semantics constitute a positive pair, while the negative pairs are derived from the representations of different sentences within the same batch.

After getting the positive and negative sentence pairs, we put them into a training objective for model fine-tune. The most widely adopted training objective is NT-Xent loss (Chen et al., 2020; Gao et al., 2021), which has been used in previous sentence and image representation learning methods and can be formulated as follows:

$$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{e^{sim\left(h_i, h_i^*\right)/\tau}}{\sum_{j=1}^{n} e^{sim(h_i, h_j)/\tau}} \quad (1)$$

where $sim\left(h_i, h_j\right)$ is the cosine similarity $\frac{h_i^{\top} h_j}{||h_i|| * ||h_j||}$, τ is a temperature hyperparameter and n is the number of sentences within a batch.

Though the training objective tries to pull representations with similar semantics closer and push dissimilar ones away from each other, these representations may still not be sufficiently discriminative and not very robust to noise. Let us denote angular $\theta_{i,j}$ as follows:

$$\theta_{i,j} = \arccos\left(\frac{h_i^{\top} h_j}{||h_i|| * ||h_j||}\right) \quad (2)$$

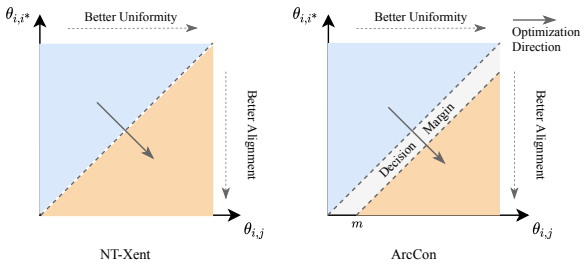


Figure 3: Comparison of NT-Xent loss and ArcCon loss. For sentence representation h_i, we try to make θ_{i,i^*} smaller and $\theta_{i,j}$ larger, so the optimization direction follows the arrow. With an extra margin m, ArcCon is more discriminative and noise-tolerant.

The decision boundary for h_i in NT-Xent is $\theta_{i,i^*} = \theta_{i,j}$, as show in Figure 3. Due to lack of decision margin, a small perturbation around the decision boundary may lead to an incorrect decision.

To overcome the problem, we propose a new training objective for sentence representation learning by adding an additive angular margin m between positive pair h_i and h_i^*. We named it Additive Angular Margin Contrastive Loss (ArcCon Loss), which can be formulated as follows:

$$\mathcal{L}_{\text{arc}} = -\log \frac{e^{\cos\left(\theta_{i,i^*}+m\right)/\tau}}{e^{\cos\left(\theta_{i,i^*}+m\right)/\tau} + \sum_{j \neq i} e^{\cos\left(\theta_{j,i}\right)/\tau}} \quad (3)$$

In this loss, the decision boundary for h_i is $\theta_{i,i^*} + m = \theta_{i,j}$, as show in Figure 3. Compared with NT-Xent, it further pushed h_i towards to the area where θ_{i,i^*} get smaller and $\theta_{i,j}$ get larger, by increasing the compactness of sentence representations with the same semantics and enlarging the discrepancy of different semantic representations. This help enhance the alignment and uniformity properties (Wang and Isola, 2020), which are two key measures of representation quality related to contrastive learning, indicating how close between positive pair embeddings and how well the embeddings are uniformly distributed. The quantitative analysis is illustrated in Section 4.5. Besides, the decision boundary leaves an extra margin m to boundary $\theta_{i,i^*} = \theta_{i,j}$ which is often used during inference, making it more tolerant to noise and more robust. All these properties make ArcCon loss more discriminative than traditional training objectives like NT-Xent. Compared with Arcface (Deng et al., 2019) which is often used in large-scale fine-grained categorization in computer vision community, ArcCon loss does not need clas-

sification labels, and could handle contrastive task properly.

3.2 Modeling Entailment Relation of Triplet Sentences

Previously the training objectives for sentence representation learning like NT-Xent loss only considered pairwise sentence relations, in which sentences are either similar or dissimilar in semantics. But in fact, there are varying degrees of semantic similarity. For example, sentence s_2 could be more similar to sentence s_1 than sentence s_3 to s_1. Existing methods lack the ability to model such partial order of semantics between sentences.

In order to distinguish the slight differences in semantics between different sentences, we propose a new self-supervised task which models the entailment relation of automatically generated triplet sentences. For each sentence s_i in the text dataset $\mathcal{D}$, we first generate an external sentence s_i' by masking contiguous segments of s_i with a masking rate of 20%. Then we enlarge the masking area and get a new sentence s_i'' with a masking rate of 40% to s_i. The masking rates are set up experimentally, and an ablation study about the effect of masking rates is illustrated in Section 4.4. An example of the masking procedure is shown as follows:

s_i Al Jaber's first long distance travel was of 800km which he covered by circling Qatar.

s_i' Al Jaber's first long distance travel was ~~of 800km which~~ he covered by circling Qatar.

s_i'' Al Jaber's first long distance travel was ~~of 800km which he covered by~~ circling Qatar.

We can constitute a triplet (s_i, s_i', s_i'') with entailment relation among them. Though in rare cases, the strategy may generate sentences that do not exhibit the desired relationship and introduce some noise, the entailment relation holds true most of the time. We expect encountering enough data will reinforce the correct ones whereas the impact of incorrect ones will diminish.

Since the s_i, s_i' and s_i'' are similar literally and semantically, generating their representations with dropout noise may obscure their entailment relation and add inaccurate signals to the representation learning process. So we turn off the dropout of the encoder when modeling the triplet relation.

As s_i' is more similar to s_i in semantics than s_i'' is, we could model such relation with a triplet objective:

$$\mathcal{L}_{\text{tri}} = max\left(0, sim(\bar{h}_i, \bar{h}_i'') - sim(\bar{h}_i, \bar{h}_i') + m\right) \tag{4}$$

in which $\bar{h}_i$ is the sentence representation of s_i generated without dropout noise and $sim(i, j)$ is the cosine similarity between i and j. As the semantic difference between s_i' and s_i'' may be subtle depending on the original sentence s_i and the masked words, here we set m to zero.

Combine formula (3) and formula (4), the final form of our training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{arc}} + \lambda \mathcal{L}_{\text{tri}} \tag{5}$$

in which λ is a coefficient.

4 Experiments

4.1 Setups

Evaluation Tasks We evaluate our method on two kinds of sentence related tasks:

- Unsupervised Semantic Textual Similarity (STS): These tasks measure the model's ability to estimate the semantic similarities between sentences.
- SentEval Transfer Tasks: These tasks measure the effectiveness of sentence embeddings used in downstream transfer tasks.

Baselines We compare ArcCSE to several representative methods on STS and SentEval tasks, such as average GloVe embeddings (Pennington et al., 2014), Skip-thought (Kiros et al., 2015), average BERT embeddings from the last layer (Devlin et al., 2019), BERT-Flow (Li et al., 2020), and BERT-Whitening (Su et al., 2021). We also include the recently proposed contrastive learning methods, such as ISBERT (Zhang et al., 2020), CT-BERT (Carlsson et al., 2021), ConSERT (Yan et al., 2021), and the current state-of-the-art method SimCSE (Gao et al., 2021).

Implementation Details We train ArcCSE with the pre-trained checkpoints of $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ (Devlin et al., 2019). We also employ our method to SBERT (Reimers and Gurevych, 2019), which has been trained on NLI datasets, to verify the generalizability of our method.

Following SimCSE (Gao et al., 2021), we use the output of the MLP layer on top of the [CLS] as

Method	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe (avg.)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (last avg.)	30.87	59.89	47.73	60.29	63.73	47.29	58.22	52.57
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-$\text{BERT}_{\text{base}}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
$\text{ConSERT}_{\text{base}}$	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE-$\text{BERT}_{\text{base}}$	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
ArcCSE-$\text{BERT}_{\text{base}}$	**72.08**	**84.27**	**76.25**	82.32	79.54	**79.92**	**72.39**	**78.11**
w/o ArcCon loss	69.94	82.34	75.08	**83.08**	78.97	78.59	71.13	77.02
w/o Triplet loss	69.66	81.92	75.33	82.79	**79.55**	79.56	71.94	77.25
$\text{ConSERT}_{\text{large}}$	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE-$\text{BERT}_{\text{large}}$	70.88	84.16	76.43	84.50	**79.76**	79.26	**73.88**	78.41
ArcCSE-$\text{BERT}_{\text{large}}$	**73.17**	**86.19**	**77.90**	**84.97**	79.43	**80.45**	73.50	**79.37**
$\text{SBERT}_{\text{base}}$	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SimCSE-$\text{SBERT}_{\text{base}}$	69.41	80.76	74.37	82.61	77.64	79.92	**76.62**	77.33
ArcCSE-$\text{SBERT}_{\text{base}}$	**74.29**	**82.95**	**76.63**	**83.90**	**79.08**	**80.95**	75.64	**79.06**
$\text{SBERT}_{\text{large}}$	72.27	78.46	74.90	80.99	76.25	79.23	73.75	76.55
SimCSE-$\text{SBERT}_{\text{large}}$	76.16	83.77	77.27	84.33	79.73	81.67	**77.25**	80.03
ArcCSE-$\text{SBERT}_{\text{large}}$	**76.36**	**85.72**	**78.22**	**85.20**	**80.04**	**82.25**	77.01	**80.69**

Table 1: Sentence representation performance on the STS tasks. We employ our method to BERT and SBERT in both base and large versions and report Spearman's correlation.

the sentence representation during training, and use the [CLS] output without MLP layer for evaluation. The dropout rate is set to 0.1. For ArcCon loss, we set the angular margin m to 10 degrees and the temperature τ to 0.05. When modeling the entailment relation of triplet sentences, we set the masking ratios as 20% and 40% respectively. Since the semantic difference between triplet sentences is more obvious for long sentences, we filter out sentences with less than 25 words and use the left ones for the triplet loss. The loss coefficient λ is set to 0.1 experimentally.

We use one million random sampled sentences from English Wikipedia for training, which has been used in previous work (Gao et al., 2021) [1]. During training, the sentences are sampled by length. We set different maximum sentence lengths for ArcCon loss and triplet loss to save memory. The length is set to 32 for the ArcCon loss in large models, and to the maximum length within a batch for all other cases. We train our model for one epoch and the learning rate is set to 3e-5 for base models and 1e-5 for large models. We search the batch size within {8, 16, 32} and always update the parameters every 64 steps. The model is optimized by the AdamW with Sharpness-Aware Minimization (Foret et al., 2021) and default configurations.

We evaluate our model every 125 training steps on the development set of STS-B, and the best checkpoint is used for the final evaluation on test sets. Our implementation is based on HuggingFace's Transformers (Wolf et al., 2020).

4.2 Unsupervised STS Tasks

We conduct experiments on 7 semantic textual similarity (STS) tasks, including STS tasks 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-Relatedness (Marelli et al., 2014). Within these datasets, each sample contains two sentences and a gold score between 0 and 5 which indicates their semantic similarity. We use SentEval toolkit (Conneau and Kiela, 2018) for evaluation and report the Spearman's correlation following previous works (Reimers and Gurevych, 2019; Gao et al., 2021).

The evaluation results are shown in Table 1,

[1] https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse/resolve/main/wiki1m_for_simcse.txt

Method	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
GloVe (avg.)	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip-thought	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
$\text{BERT}_{\text{base}}$ (last avg.)	78.66	86.25	94.37	88.66	84.40	**92.80**	69.54	84.94
IS-$\text{BERT}_{\text{base}}$	81.09	**87.18**	94.96	88.75	**85.96**	88.64	74.24	85.83
SimCSE-$\text{BERT}_{\text{base}}$	**81.18**	86.46	94.45	88.88	85.50	89.80	74.43	85.81
ArcCSE-$\text{BERT}_{\text{base}}$	79.91	85.25	**99.58**	**89.21**	84.90	89.20	**74.78**	**86.12**
$\text{BERT}_{\text{large}}$ (last avg.)	84.30	89.22	95.60	86.93	89.29	91.40	71.65	86.91
SimCSE-$\text{BERT}_{\text{large}}$	**85.36**	**89.38**	95.39	89.63	90.44	91.80	**76.41**	88.34
ArcCSE-$\text{BERT}_{\text{large}}$	84.34	88.82	**99.58**	**89.79**	**90.50**	**92.00**	74.78	**88.54**

Table 2: Sentence representation performance on SentEval transfer tasks. We report the accuracy results of both $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ level models.

from which we can see that ArcCSE outperforms the previous approaches. Compared with the previous state-of-the-art method SimCSE, ArcCSE-$\text{BERT}_{\text{base}}$ raises the average Spearman's correlation from 76.25% to 78.11%, and ArcCSE-$\text{BERT}_{\text{large}}$ further pushes the results to 79.37%. The performance is even better than strong supervised method SBERT, which has already been trained on NLI datasets. Furthermore, we can also employ our method to SBERT and improve its performance to 79.06% and 80.69% for the base and large models respectively, which is more effective than SimCSE.

We also explore the improvements made by the ArcCon loss and triplet loss independently based on $\text{BERT}_{\text{base}}$. From Table 1 we can see that with ArcCon loss alone, the average Spearman's correlation is 77.25%. When combining the traditional NT-Xent loss with our proposed triplet loss, the average Spearman's correlation is 77.02%. Both of them outperform the previous state-of-the-art method SimCSE, whose average Spearman's correlation is 76.25%. This demonstrates the effectiveness of ArcCon and triplet loss we proposed.

4.3 SentEval Tasks

We evaluate our model with SentEval toolkit on several supervised transfer tasks, including: MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST-2 (Socher et al., 2013), TREC (Voorhees and Tice, 2000) and MRPC (Dolan and Brockett, 2005). For each task, SentEval trains a logistic regression classifier on top of the sentence embeddings and tests the performance on the downstream task. For a fair comparison, we do not include models with auxiliary tasks like masked language modeling.

The results are shown in Table 2. We can see that ArcCSE performs on par or better than baseline methods in both $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ level. This demonstrates the effectiveness of our method in learning domain-specific sentence embeddings.

4.4 Ablation Studies

Effect of Angular Margin The angular margin m in ArcCon loss affects the discriminative power directly. To investigate the effect of m, we conduct an experiment by varying m from 0 degrees to 20 degrees, increased by 2 degrees at each step. We tune the hyper-parameter based on Spearman's correlation on the development set of STS-B following previous works (Kim et al., 2021; Gao et al., 2021). The results are shown in Figure 4.

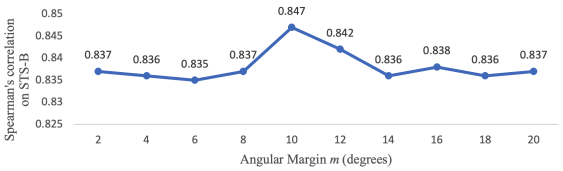


Figure 4: Effect of the angular margin m in ArcCon loss. Results are reported on the development set of STS-B based on the Spearman's correlation.

We can see that the best performance is achieved when $m = 10$, either larger or smaller margin degrade the performance. This matches our intuition since small m may have little effect, and large m may negatively influence the positive pair relation modeling.

Effect of Temperature The temperature τ in ArcCon Loss affects its effectiveness, so we carry out an experiment with τ varying from 0.01 to 0.1, increased by 0.01 at each step. The results are shown in Figure 5. We can see that the model

ArcCSE-BERT$_{\text{base}}$	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
w/ Dropout$_{\text{on/off}}$	**72.08**	**84.27**	**76.25**	**82.32**	**79.54**	**79.92**	**72.39**	**78.11**
w/ Dropout$_{\text{mix/off}}$	70.51	83.59	75.85	82.30	78.87	78.74	71.58	77.35
w/ Dropout$_{\text{on/on}}$	69.62	83.13	74.42	82.15	78.39	78.39	70.89	76.71

Table 3: Effect of on-off Switching of Dropout. We use different dropout settings to generate sentence embeddings used for ArcCon loss and triplet loss respectively. The "on", "off" and "mix" mean turn dropout on, turn dropout off and use different settings for two passes separately.

performs best when $\tau = 0.05$, so we use this value throughout our experiments.

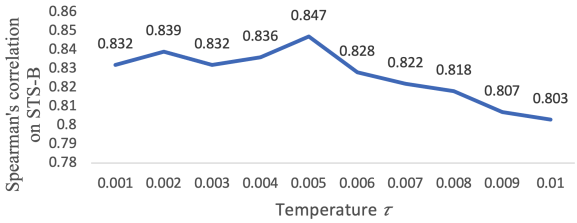


Figure 5: Effect of the temperature τ in ArcCon loss. Results are reported on the development set of STS-B based on the Spearman's correlation.

Effect of Masking Ratios The masking ratios determine the sentences generated for the entailment relation modeling and their differences in semantics, so we conduct an experiment to explore the effect of different masking ratios. The first masking ratio r_1 is varied from 10% to 25%, increased by 5% for each step. The second masking ratio r_2 is derived by adding an extra value r_d to r_1. r_d is varied from 10% to 35%, increased by 5% for each step. The results are shown in Figure 6.

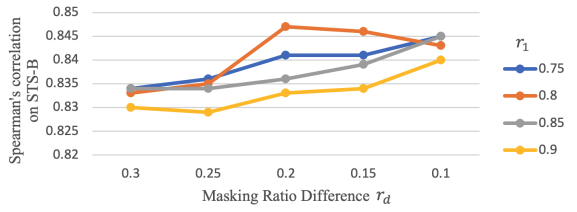


Figure 6: Effect of the masking ratios. Different lines correspond to different values of r_1. The abscissa is r_d, representing the difference between r_1 and r_2. Results are reported on the development set of STS-B based on the Spearman's correlation.

We can see that large differences between the two masking ratios tend to lead lower Spearman's correlation compared to the smaller ones. The reason may be that the larger the semantic difference is, the easier for the model to estimate the entailment relations among the triplet sentences, which makes the triplet loss less helpful. The best performance is achieved when r_1 is 20% and r_2 is 40%, and the corresponding Spearman's correlation is 0.847. We use them as our hyper-parameters.

Effect of on-off Switching of Dropout The on-off switching of dropout in the BERT-like sentence encoder affects the generated sentence representations directly. Since dropout performs a kind of averaging over the ensemble of possible subnetworks, an embedding generated with dropout turned off can be seen as a kind of "averaging" representation, while an embedding generated with dropout turned on can be seen as generated through a subnetwork.

In ArcCSE, we use the embeddings generated with the encoder dropout turned on as input for ArcCon loss, which regularizes the network by making representations generated through different subnetworks similar. When modeling the entailment relation, we generate "averaging" representations with dropout turn-off to avoid inaccurate signals. In order to verify our intuition, we conduct two experiments with different dropout settings. In the first experiment, we feed ArcCon two sentence representations generated with dropout turns on and off respectively. We carry out this experiment with angular margins ranging between 2 degrees to 12 degrees and report the best result. In the second one, we feed the triplet loss representations that are generated with dropout turns on and maintain the other settings. The results are shown in Table 3. We can see that the original settings that turn dropout on for ArcCon and turn dropout off for triplet loss achieve the best performance, which confirms our intuition.

Effect of Coefficient in the Training Objective The coefficient λ in the final optimization objective adjusts the relative weights between ArcCon and the triplet loss, as shown in formula (5). To find the most suitable λ, we conduct an experiment by varying λ from 0 to 1.2 and increased by 0.1 at each step. The results are shown in Figure 7.

We can see that the best performance is achieved

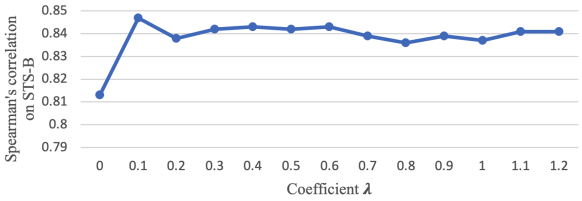


Figure 7: Effect of the coefficient in the training objective. Results are reported on the development set of STS-B based on the Spearman's correlation.

when $\lambda = 0.1$, and the corresponding Spearman's correlation is 0.847. This demonstrates that we can get the best performance by combining ArcCon and the triplet loss with proper λ.

4.5 Alignment and Uniformity Analysis

Alignment and uniformity are two properties closely related to contrastive learning and could be used to measure the quality of representations(Wang and Isola, 2020). Alignment favors encoders that generate similar representations for similar instances. It could be defined with the expected distance between embeddings of the positive paired instances:

$$\ell_{\text{align}} = \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \left\| f(x) - f\left(x^+\right) \right\|^2 \quad (6)$$

where p_{pos} denotes the distribution of positive paired instances. Uniformity prefers uniformly distributed representations, which helps preserve maximal information. It could be defined as:

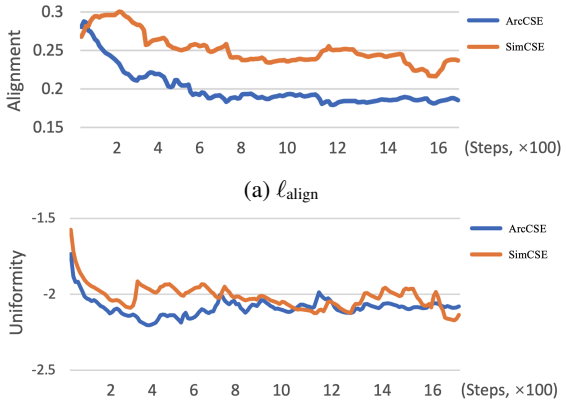
$$\ell_{\text{uniform}} = \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2} \quad (7)$$

where p_{data} denotes whole data distribution.

To justify the inner workings of our approach, we calculate the alignment and uniformity metrics every 10 steps during training on the STS-B development set. We compare our approach with SimCSE and visualize the results in Figure 8. We can see that compared to the original BERT checkpoint, both ArcCSE and SimCSE improve the alignment and uniformity measures during training. ArcCSE performs better on the alignment measure and on par with SimCSE on the uniformity measure. This verifies the intuition of our approach and demonstrates that ArcCSE could help improve the quality of sentence representations.

5 Conclusion

In this work, we propose ArcCSE, a self-supervised contrastive learning framework for learning sentence representation. We propose a new optimizing objective ArcCon loss to model pairwise sentence relations with enhanced discriminating power, and a new self-supervised task to model the partial order of semantics between sentences. Experimental results on semantic textual similarity tasks (STS) and SentEval tasks demonstrate that both techniques bring substantial improvements and our method outperforms previous state-of-the-art method for sentence representation learning.



(a) ℓ_{align}

(b) ℓ_{uniform}

Figure 8: ℓ_{align} and ℓ_{uniform} of ArcCSE and SimCSE, visualized by calculating alignment and uniformity every 10 training steps. For both measures, lower numbers are better.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

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

(SemEval-2016), pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

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, Atlanta, Georgia, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

S. Chopra, R. Hadsell, and Y. 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 vol. 1.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Juan Manuel Coria, Sahar Ghannay, Sophie Rosset, and Hervé Bredin. 2020. A metric learning approach to misogyny categorization. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 89–94, Online. Association for Computational Linguistics.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization.

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 (EMNLP)*. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

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, Online. Association for Computational Linguistics.

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

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, Online. Association for Computational Linguistics.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746.

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

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Xiaofei Ma, Cicero Nogueira dos Santos, and Andrew O. Arnold. 2021. Contrastive fine-tuning improves robustness for neural rankers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 570–582, Online. Association for Computational Linguistics.

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

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*, volume 26. Curran Associates, Inc.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. 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 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1857–1865, Red Hook, NY, USA. Curran Associates Inc.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted

structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012.

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

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. Association for Computing Machinery.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Computer Vision – ECCV 2016*, pages 499–515, Cham. Springer International Publishing.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.