A Simple Angle-based Approach for Contrastive Learning of Unsupervised Sentence Representation

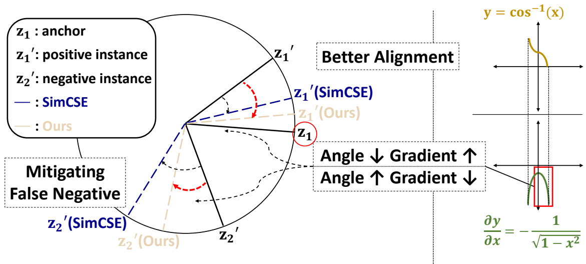
Anonymous ACL submission

Abstract

Contrastive learning has been successfully adopted in VRL (visual representation learning) by constructing effective contrastive pairs. A promising baseline SimCSE has made notable breakthroughs in unsupervised SRL (sentence representation learning) following the success of contrastive learning. However, considering the difference between VRL and SRL, there is still room for designing a novel contrastive framework specially targeted for SRL. We propose a novel angle-based similarity function for contrastive objective. By examining the gradient of our contrastive objective, we show that an angle-based similarity function incites better training dynamics on SRL than the off-the-shelf cosine similarity: (1) effectively pulling a positive instance toward an anchor instance in the early stage of training and (2) not excessively repelling a false negative instance during the middle of training. Our experimental results on widely-utilized benchmarks demonstrate the effectiveness and extensibility of our novel angle-based approach. Subsequent analyses establish its improved sentence representation power.

1 Introduction

Contrastive learning has achieved promising results in VRL (visual representation learning) (Hadsell et al., 2006; Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019; He et al., 2020; Chen et al., 2020). However, the adoption of contrastive learning in SRL (sentence representation learning) has suffered from several limitations such as inherently difficult data augmentations due to a discrete nature of NLP (natural language processing) (Li et al., 2022) and a limited property of PLMs' (pre-trained language models) representation spaces (Gao et al., 2018; Ethayarajh, 2019; Wang et al., 2019; Li et al., 2020a). Unlike earlier attempts to construct positive pairs (Zhang et al., 2017; Wei and Zou, 2019; Xie et al., 2020; Sun et al., 2020; Zhang et al., 2020b, 2021b; Giorgi et al., 2021; Kim et al., 2021; Yan et al., 2021), which are similar to the works in VRL, SimCSE (Gao et al., 2021) found using the independently sampled dropout (Srivastava et al., 2014) mask is simple but effective for augmentations for unsupervised contrastive learning and can alleviate the problem of anisotropy – a narrow cone-like representation space leads to a lack of expressiveness (Ethayarajh, 2019; Li et al., 2020a; Gao et al., 2021). A number of studies based on SimCSE reported a successful utilization of contrastive learning in SRL (Zhou et al., 2022; Zhang et al., 2022a; Chuang et al., 2022; Zhang et al., 2022b; Wu et al., 2022; Liu et al., 2023).



(a) 2D Hypersphere Manifold of Representation Space (b) Derivative of $\cos^{-1}$

Figure 1: Difference between contrastive learning for unsupervised sentence representation using different similarity functions. Compared to the widely-utilized cosine similarity function (SimCSE), our novel angle-based similarity function shows different training dynamics, which lead to a better alignment and mitigate a sampling bias by not repelling the negative instance strongly. We infer that this phenomenon is due to the gradient property of the angle-based similarity function as seen in (b).

However, indeed there are differences between SRL and VRL (Nie et al., 2022), which suggests that consideration of the nature of SRL should precede a blind adoption of VRL's success. Among several points that differentiate SRL, we focus on two important points: (1) the number of in-batch negative instances; (2) the property of training dynamics as SRL usually uses pre-trained models.

More specifically, several works utilize a smaller number of negative instances (e.g., 64 ~ 512 (Gao et al., 2021; Zhou et al., 2022; Zhang et al., 2022a; Chuang et al., 2022; Wu et al., 2022; Liu et al., 2023)), while the larger number of negative instances (e.g., 4096 ~ 65536 (He et al., 2020; Chen et al., 2020)) is used in VRL. Also, the number of training epochs is relatively smaller (e.g., 1 ~ 4 (Gao et al., 2021; Zhou et al., 2022; Zhang et al., 2022a; Wu et al., 2022; Liu et al., 2023)) to train pre-trained language models (PLMs). Considering the differences, we aim to design a novel contrastive objective with better properties for SRL.

Towards this end, we first investigate which component of the contrastive objective is effective for SRL. By analyzing a gradient of the contrastive objective, we find that a temperature value of normalized temperature-scaled cross entropy (NT-Xent) loss (Chen et al., 2020) and a derivative of the similarity function has a correlation with a magnitude of gradient. This indicates that both of them affect training dynamics. Conforming to previous works that have reported the role of temperature (Wang and Liu, 2021; Zhang et al., 2021a), we focus more on exploring better similarity functions that take into account the nature of PLMs and SRL.

In this regard, we design a novel angle-based similarity function for contrastive learning of unsupervised sentence representation. Comparing the derivatives of the naive cosine similarity function used in SimCSE and the proposed angle-based function, we find an interesting property from the derivative of our angle-based function – it exponentially increases (absolute value) from 90 to 0 degrees. We expect that this property could lead to following positive impacts: (1) the angle-based approach improves the *alignment* during the early stages of training due to the anisotropic space of PLMs with smaller angles; (2) the angle-based approach mitigates the problem of inappropriate in-batch negative sampling (*i.e.*, false negative (Chuang et al., 2020; Robinson et al., 2020; Zhou et al., 2022)) during the middle of training as it does not strongly repel the negative instances with higher angle differences (see Figure 1).

Under the assumption that the angle-based approach can solve some issues, we propose a *simple angle-based approach for contrastive sentence embedding framework* (SimACE), which equips with the aforementioned angle-based function. The proposed method is straightforward. We change the vanilla cosine similarity function to the angle-based function by applying an inverse function of cosine (arccosine) and adjusting its range suitable for softmax logits of contrative objective. SimACE outperforms the baseline SimCSE on off-the-shelves unsupervised sentence representation benchmark, with relatively small in-batch negative instances. Also, SimACE shows more robust performance and even outperforms the baseline in a multi-task benchmark for sentence representation. In addition, we apply our novel design to recent state-of-the-art methods based on SimCSE and show that simply replacing the original cosine similarity function with our angle-based similarity function can improve the performance. These results demonstrate the extensibility of our work. To verify the difference between SimCSE and SimACE and the reason for improved performance, we conduct several experimental analyses, including semantic space visualization, reporting uniformity and alignment, and training dynamics in terms of angle. We note that the reason for SimACE's success is that the angle-based approach is appropriate especially for unsupervised SRL, though it shows unprecedented results and tendencies that are not in line with prior works in VRL (Wang and Isola, 2020; Wang and Liu, 2021; Zhang et al., 2021a).

2 Related Works and Preliminary

Unsupervised SRL In SRL, high-quality representation greatly correlated with human evaluations on similarities and has been proven to be effective when transferred to downstream tasks. Despite the success of transformer-based PLMs on transfer tasks (Devlin et al., 2018; Liu et al., 2019), PLMs-based representations underperformed conventional static word embeddings, such as Word2Vec (Mikolov et al., 2013) and its augmented version (Pennington et al., 2014), particularly in sentence representation benchmark (STS tasks (Cer et al., 2017)). As PLMs turned out to have high-dimensional conical space (Ethayarajh, 2019), post-processing methods (Li et al., 2020b; Su et al., 2021) instantly tried to mitigate the problem in PLMs, but were limited to improving the performance. Contrastive learning-based methods aim at smoothing the bottleneck of its anisotropic property, by constructing finely tailored contrastive pairs (Yan et al., 2021; Gao et al., 2021; Zhou et al., 2022; Zhang et al., 2022a; Wu et al., 2022; Chuang et al., 2022; Zhang et al., 2022b; Liu et al., 2023) or designing an apt contrastive objective (Gao et al., 2021; Zhang et al., 2022b). In unsupervised contrastive learning, it mainly falls into two components in terms of achieving these goals: 1) constructing the well-crafted pairs; 2) designing an appropriate contrastive objective. Most efforts have focused on constructing the former (Zhang et al., 2022a; Zhou et al., 2022) or adding auxiliary objective on contrastive loss (Chuang et al., 2022; Zhang et al., 2022b; Wu et al., 2022; Liu et al., 2023).

Preliminary In unsupervised SRL, SimCSE systematically proposed the major components for learning sentence representations, and many recent works (Zhou et al., 2022; Zhang et al., 2022a; Chuang et al., 2022; Zhang et al., 2022b; Wu et al., 2022; Liu et al., 2023) are originated from the following framework. First, given a collection of sentences $D = \{x_i\}_{i=1}^m$, positive views are derived from independently passing x_i to encoder twice (*i.e.,* dropout augmentation), while negative pairs through in-batch negative sampling (Chen et al., 2017). Secondly, they use NT-Xent loss, which is based on similarity function $sim(\mathbf{z}_i, \mathbf{z}_j)$:

$$l_i = -log \frac{e^{sim(\mathbf{z}_i, \mathbf{z}_i^{'})/\tau}}{\sum_{j=1}^{N} e^{sim(\mathbf{z}_i, \mathbf{z}_j^{'})/\tau}}, \tag{1}$$

where $\mathbf{z}_i$, $\mathbf{z}_i^{'}$, and $\mathbf{z}_j^{'}(i \neq j)$ denotes the hidden representation of the anchor, positive instance, and negative instance. The hidden representation with $'$ means the augmented view of instance, which is a dropout-based one in SimCSE, and τ dictates temperature. Although there have been several works dealing with understanding the contrastive learning (Wang and Liu, 2021; Zhang et al., 2021a) in the field of VRL, little is known about the unique property of contrastive learning for SRL. Regardless of the progress in the area of SRL, the major problem of grounding based on deeper analysis, such as the role of temperature or the possibility of different similarity functions, persists.

3 Angle-based Contrastive Learning

3.1 Motivation

In this section, we first investigate the gradient of contrastive loss to find which factors affect the training dynamics in SRL. For simplicity, we consider $\mathbf{z}$ as input hidden representation like Equation 1, which then can be reformulated using the softmax probability. Treating the $sim(\mathbf{z}_i, \mathbf{z}_i^{'})/\tau$ in Equation 1 as the logit of a vanilla Cross-Entropy loss, we can define the probability (λ_i) of each negative sample as below:

$$\begin{aligned} k_{i,j} &= sim(\mathbf{z}_i, \mathbf{z}_j^{'})/\tau, \quad \forall i = 1, ..., N, \quad \forall j = 1, ..., N, \\ \lambda_i &= \frac{e^{k_{i,i}}}{\sum_{j=1}^{N} e^{k_{i,j}}}, \quad \forall i = 1, ..., N, \quad \ni \sum_{j=1}^{N} e^{\lambda_j} = 1. \end{aligned} \tag{2}$$

We can simply calculate the gradient according to the derivative of the softmax function as follows:

$$\begin{aligned} l_i &= -log(\lambda_i), \\ \frac{\partial l_i}{\partial k_{i,j}} &= -\frac{1}{\lambda_i} \frac{\partial \lambda_i}{\partial k_{i,j}}, \\ where \quad \frac{\partial \lambda_i}{\partial k_{i,j}} &= \lambda_i \frac{\partial log(\lambda_i)}{\partial k_{i,j}} = \lambda_i (1\{i = j\} - \lambda_j). \end{aligned} \tag{3}$$

Using the chain rule, we can compute the gradient for $\mathbf{z}_i$ as follows:

$$\begin{aligned} \frac{\partial k_{i,j}}{\partial z_i} &= \frac{1}{\tau} \frac{\partial sim(\mathbf{z}_i, \mathbf{z}_j^{'})}{\partial z_i}, \\ \frac{\partial l_i}{\partial z_i} = \frac{\partial l_i}{\partial k_{i,j}} \cdot \frac{\partial k_{i,j}}{\partial z_i} &= \frac{1}{\tau}(\lambda_j - 1\{i = j\}) \frac{\partial sim(\mathbf{z}_i, \mathbf{z}_j^{'})}{\partial z_i}. \end{aligned} \tag{4}$$

In Equation 4, we can find that both the derivative of the similarity function and the value of temperature influence the gradient of loss. The role of the temperature has been covered in the asymptotic analysis of several previous studies (Wang and Liu, 2021; Zhang et al., 2021a), most notably finding that it is strongly related to entropy, determining the gradient weight for negative instances.

In contrast, we focus on the influence of the similarity function and assume that a change in the similarity function will also lead to a significant change in the training dynamics.

3.2 Angle-based Similarity Function

Most of the works, including SimCSE, use a naive cosine similarity ($cossim$) for similarity function (sim). Nevertheless, there have been several attempts to deal with other candidates of the similarity function; e.g., RBF (radial basis function) (Zhang et al., 2020a), angular distance (Zhang et al., 2022b), or hyperbolic distance (Ge et al., 2023). Among them, we focus on an angular relation between different sentence representations, where the previous work has raised the issue of gradient dissipation with regard to angle in SRL (Nie et al., 2022). To model the angular similarity between hidden representations, we apply arccosine (cos^{-1}) to the dot product of two normalized representations[1]. Given a mini-batch $\{s_i\}_{i=1}^n$, we denote $cossim(z_i, z_j)$ as the cosine similarity function of

[1] A ℓ_2 normalized dot product is analogous of cosine similarity function.

two hidden representations for two samples s_i, s_j. Then the straightforward angle similarity (θ) can be described as:

$$\theta_{i,j} = cos^{-1}(cossim(\mathbf{z}_i, \mathbf{z}_j)), \quad (5)$$

where $\theta_{i,j}$ represents the angular distance between $\mathbf{h}_i$ and $\mathbf{h}_j$. Note that this vanilla form of angle relation is not appropriate for contrastive learning, since it is not an increasing function. The modified version of the angle-based similarity function will be introduced in Section 3.3.

We now compare the derivative of the cosine similarity ($cossim$) and the newly designed angle-based one (θ). The derivative of each similarity function can be derived as follows:

$$\frac{\partial cossim(\mathbf{z}_i, \mathbf{z}_j^{'})}{\partial z_i} = \frac{\mathbf{z}_j^{'}}{||\mathbf{z}_i||||\mathbf{z}_i||} - cossim(\mathbf{z}_i, \mathbf{z}_j^{'})\frac{\mathbf{z}_i}{||\mathbf{z}_i||^2},$$
$$\frac{\theta_{i,j}}{\partial z_i} = -\frac{1}{\sqrt{1 - cossim(\mathbf{z}_i, \mathbf{z}_j^{'})^2}} \cdot \frac{\partial cossim(\mathbf{z}_i, \mathbf{z}_j^{'})}{\partial z_i}. \quad (6)$$

The derivative of arccosine ($cos^{-1}(x)$) is $-\frac{1}{\sqrt{1-x^2}}$ for -1 < x < 1. The range of values for this function is negative infinity and -1 for 0 and 90 degrees respectively, and the function is concave (see Figure 1(b)). So, if we use the angle-based similarity function for InfoNCE loss, we can infer that the strength of both pulling positive instance and repelling negative instance is stronger for small angles, while the strength of pulling and repelling becomes weaker as the angle gets larger since the magnitude of the gradient decreases accordingly. Based on this intuition, we expect that the gradient property of the angle-based function can be effective especially for contrastive learning in SRL for the following two reasons. First, since the embedding spaces of several PLMs are anisotropic such that sentence representations are converged into narrow cone (Gao et al., 2018; Ethayarajh, 2019; Wang et al., 2019; Li et al., 2020a), we believe that strongly repelling negative instances while pulling positive instances will be effective in improving the alignment of the semantic space during the early stages of training. Secondly, since the repelling power of negative instance is exponentially decreased as the angle gets larger in the middle of training, angle-based contrastive learning can mitigate the problem of false negative instance[2] (Chuang et al., 2020; Robinson et al., 2020; Zhou et al., 2022). In this regard, we believe that different instances will not be separated by more than a certain threshold angle, and assume that the embedding space of the model after angle-based contrastive learning is narrower than that of the model trained by cosine similarity-based contrastive loss.

[2] An in-batch negative sampling of unsupervised contrastive learning may lead to repelling the semantically-closed instance, unintentionally.

Our methodology may appear similar to method used in Zhang et al., 2022b due to the use of angular space. However, the motivation behind the previous work is entirely derived from VRL method, named ArcFace Loss (Deng et al., 2019). In contrast, the foundation for our proposed SimACE is a comprehensive understanding and consideration of SRL characteristics, coupled with mathematical reasoning and subsequent analyses to validate it. Detailed analyses of the angle-based function's characteristics which can back up our assumptions are covered in Section 5.

3.3 SimACE

Now, we propose SimACE: *simple angle-based approach for contrastive sentence embedding framework*. It adopts the angle-based similarity function suitable for unsupervised contrastive learning. Before directly leveraging the angle-based function (θ) defined in Equation 5, we modify the range of θ by subtracting a value from $\frac{\pi}{2}$. This is because of the nature of contrastive learning with the cross-entropy objective, which involves increasing the similarity of a positive pair and decreasing that of a negative pair. This adjustment shifts the similarity range from [-1, 1] to [$\frac{\pi}{2}$ - π, $\frac{\pi}{2}$ - 0] = [-$\frac{\pi}{2}$, $\frac{\pi}{2}$]:

$$\theta_{i,j} = \frac{\pi}{2} - cos^{-1}(cossim(\mathbf{z}_i, \mathbf{z}_j^{'})), \quad (7)$$

Then, the new loss function based on our angle-based similarity function is defined as follows:

$$L_{ang} = -log\frac{e^{\theta_{i,i'}/\tau}}{\sum_{j=1}^{N} e^{\theta_{i,j'}/\tau}}. \quad (8)$$

In addition, to mitigate the issue of the relatively narrower space (mentioned in Section 3.1), we apply a margin penalty to the angle between the anchor and the positive sample, leveraging its inherent property of angle-based similarity. We simply subtract the angular margin (m) between the anchor ($\mathbf{z}_i$) and the positive pair ($\mathbf{z}_i^{'}$). Subtracting the margin term to the hidden representation of the positive instance is in line with the adversarial perturbation, an effective scheme for semantic space interpolation (Hadsell et al., 2006; Chen et al., 2021; Robinson et al., 2021). We expect this negative pertur-

PLMs	Method	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
$\text{BERT}_{\text{base}}$	first-last ♣	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
	SimCSE ♣	68.40	82.41	74.38	80.91	78.56	76.85	**72.23**	76.25
	ArcCon*	**71.76**	82.77	**76.81**	83.56	78.87	79.36	71.16	77.76
	SimACE*	71.63	**83.44**	76.65	**83.85**	**79.95**	**79.99**	71.86	**78.20**
$\text{BERT}_{\text{large}}$	SimCSE ♣	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
	ArcCon*	73.38	84.94	76.74	84.28	**80.19**	**80.02**	72.96	78.93
	SimACE*	**73.89**	**85.07**	**77.67**	**84.87**	79.18	79.96	**74.61**	**79.32**
$\text{RoBERTa}_{\text{base}}$	first-last ♣	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
	SimCSE ♣	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
	ArcCon*	69.01	81.30	73.02	81.47	81.54	80.43	68.94	76.53
	SimACE*	**70.50**	**84.16**	**76.33**	**83.38**	**82.45**	**82.24**	**69.69**	**78.39**
$\text{RoBERTa}_{\text{large}}$	SimCSE ♣	**72.86**	83.99	75.62	84.77	81.80	81.98	71.26	78.90
	ArcCon*	70.03	83.15	75.26	83.76	81.43	80.64	70.22	77.78
	SimACE*	72.12	**84.41**	**77.25**	**85.05**	**81.92**	**83.35**	**71.37**	**79.35**

Table 1: Performance of several unsupervised contrastive learning methods using different similarity functions on STS tasks (Spearman's correlation). Each bold number and underlined number indicates the best and second-best performance within the PLMs, respectively. We reproduce the results of ArcConLoss (proposed by ArcCSE (Zhang et al., 2022b)), following configurations with a grid search for their hyper-parameters. ♣: Results from Gao et al., 2021. ∗: Results of our experiments.

PLMs	SimCSE	ArcCon	SimACE
$\text{BERT}_{\text{base}}$	$75.97_{\pm 0.69}$	$76.76_{\pm 0.76}$	$77.46_{\pm 0.47}$
$\text{BERT}_{\text{large}}$	$77.62_{\pm 0.58}$	$78.66_{\pm 0.21}$	$79.02_{\pm 0.26}$
$\text{RoBERTa}_{\text{base}}$	$76.77_{\pm 0.18}$	$76.27_{\pm 0.75}$	$77.87_{\pm 0.44}$
$\text{RoBERTa}_{\text{large}}$	$78.29_{\pm 0.32}$	N/A	$79.14_{\pm 0.15}$

Table 2: Mean and standard deviation across 5 different runs of different methods with random seeds. Unfortunately, since RoBERTa-large models trained by ArcConLoss with different random seeds show a gradient explosion, we report these results as N/A (Not Applicable or Not Available). We report p-values for each baseline in the Appendix (Table 9), which are highly statistically significant (p < 0.001).

PLMs	SimCSE	SimACE
$\text{BERT}_{\text{base}}$	$46.16_{\pm 0.36}$	$48.19_{\pm 0.27}$
$\text{BERT}_{\text{large}}$	$50.35_{\pm 0.22}$	$51.62_{\pm 0.13}$
$\text{RoBERTa}_{\text{base}}$	$47.33_{\pm 0.09}$	$49.46_{\pm 0.24}$
$\text{RoBERTa}_{\text{large}}$	$50.43_{\pm 0.17}$	$51.66_{\pm 0.08}$

Table 3: Performance of averaged results on MTEB benchmark (total 56 datasets). Results are highly statistically significant (p < 0.001). Detailed results can be found in Appendix (Table 12).

bation can lead to a discrimination of the positive pair's feature space and enhance the alignment.

Consequently, the final form of our SimACE's training objective is:

$$L_{ang} = -log \frac{e^{(\theta_{i,i'}-m)/\tau}}{e^{(\theta_{i,i'}-m)/\tau} + \sum_{j\neq i}^{N} e^{\theta_{i,j'}/\tau}}. \tag{9}$$

4 Experiments

4.1 Unsupervised Corpus and Benchmark

Following the literature, we train SimACE on datasets randomly sampled from English Wikipedia (10^6) same with the baseline SimCSE (Gao et al., 2021). Then, we evaluate SimACE on 7 STS tasks: STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (STS-B) (Cer et al., 2017) and SICK Relatedness (SICK-R) (Marelli et al., 2014). These datasets contain pairs of two sentences along with a gold score ranging from 0 to 5 whose scores represent their semantic similarity. We obtain these datasets from the SentEval (Conneau and Kiela, 2018) toolkit.

4.2 Implementation Details

Training Setups We follow standard practices and conduct a preliminary grid search using the STS-B development dataset to determine the hyper-parameter configuration. We carry out a grid search of learning rate ∈ {1e-5, 3e-5}, temperature (τ) ∈ [0.06, 0.07], and batch size ∈ {32, 128}. Then, we set the same training hyper-parameters for all experiments with 10 (radians) for the margin. We train our models for 1 epoch and evaluate the model every 125 steps on the development set. A server equipped with a single GPU (NVIDIA A100, 40GB memory) was used for all our experiments. Detailed hyperparameter settings can be found in Table 7.

Evaluation Setups We evaluate SimACE on 7 STS tasks as introduced in Section 4.1. For the need of reproducibility, we update the baselines' scores which are different from those reported in the original paper. In addition, we also report the averaged results of different random seeds to ensure a fair comparison to the baseline, considering a reported problem that the performance of unsupervised SimCSE is unstable depending on random seeds (Jiang et al., 2022).

Network Implementation We train SimACE with the pre-trained checkpoints of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) downloaded from Huggingface's `Transformers` (Wolf et al., 2019). Each encoder consists of 12 and 24 Transformer layers for the base and large sizes, respectively. The base model has a hidden size of 768 and 12 attention heads, while the large model has a hidden size of 1024 and 16 attention heads. Following the literature (Gao et al., 2021), we choose the representation of the [CLS] token as the sentence representation during training, and use the [CLS] output without the pooler for evaluation.

4.3 Comparative Results

We aim to compare our angular similarity function with other candidates: we employed the original cosine similarity function from SimCSE, and ArcConLoss from Zhang et al., 2022b of which loss functions are based on cosine similarity and the modified cosine similarity inspired by ArcFace (Deng et al., 2019), respectively. Experimental results on STS tasks are shown in Table 1. Despite the fewer in-batch negative instances than SimCSE, SimACE improves the average score on STS from **76.95 to 78.20** for BERT-base and from **78.46 to 79.32** for BERT-large, respectively. Interestingly, SimACE shows more powerful performance on RoBERTa-base and RoBERTa-large, which further pushes the results from **76.64 to 78.39** and **78.53 to 79.35**, respectively. These results imply that training dynamics can be differentiated depending on PLMs. We will do a deep dive into the grounding of SimACE's capability in Section 5.

4.4 Robustness of Angular-based Approach

To ensure the robustness with regard to different random seeds, we conduct 5 runs of model training with the configurations outlined in Appendix (Table 7), each initialized with distinct random seeds. Subsequently, we calculate the mean and standard deviation values. The results provided in Table 2 show both the superior performance and the robustness of our method compared to the baselines using different similarity functions.

4.5 Results on MTEB benchmark

To validate a generalization ability of SimACE, we evaluate our method in the additional sentence embedding benchmark, named Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022). This benchmark consists of total 56 tasks: 10 semantic textual similarity (STS) tasks, 12 classification tasks, 11 clustering tasks, 3 pair classification tasks, 4 reranking tasks, 15 retrieval tasks, and 1 summarization tasks. As seen in Table 3, SimACE shows better performance compared to the baseline SimCSE within all PLMs.

4.6 Extension to SOTAs

In the previous section, we reported the comparative results to confirm the superiority of our method. From now on, we aim to confirm the effectiveness of our angle-based similarity function from a different perspective. We employ several recent state-of-the-arts and replace their cosine similarity function with our angle-based one. Specifically, we utilize PCL (Wu et al., 2022) and RankCSE (Liu et al., 2023). A detailed explanation of each method can be found in Section D. Concretely, we use 3 versions of modified SimCSE objectives: group-wise relations (P-Cf) loss (Eq. 12), and two different ranking distillation losses (Eq. 14). As a result, we replace $sim(\cdot,\cdot)$ of PCL and RankCSE with our $\theta(\cdot,\cdot)$ (Eq. 5). Furthermore, other loss terms and training details including hyperparameter settings are the same as in the original papers.

Comparative Results Table 4 reports the results. We can observe that our angle-based versions of PCL and RankCSE outperform their original cosine similarity version in terms of the average STS score. Interestingly, we can observe that RankCSE-listMLE with our angle-based similarity function shows the best result on all PLMs. These results show that our angle-based similarity function is adaptable across different SRL methods on different PLMs. As before, we report the robustness of random seeds in the Appendix (Table 10).

5 Analysis

5.1 Difference of Semantic Space between PLMs

From Table 1, we can see that our angle-based similarity function (SimACE) encourages the PLMs more suitable for computing correct similarities between two sentence representations, regardless of

PLMs	Method	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
BERT$_{\text{base}}$	PCL	72.44	82.16	**74.69**	82.09	**79.13**	**79.30**	71.95	77.39
	+ angle	**73.29**	**82.39**	74.48	**82.22**	78.77	79.24	**72.24**	**77.52**
	RankCSE$_{\text{listNet}}$	69.02	82.88	73.54	80.18	77.65	77.73	73.22	76.32
	+ angle	71.06	84.46	75.49	82.60	78.91	79.53	74.06	78.02
	RankCSE$_{\text{listMLE}}$	74.47	**85.77**	78.09	84.71	**81.48**	81.76	**74.40**	80.06
	+ angle	**75.83**	85.48	**78.46**	**85.19**	81.02	**81.94**	73.60	**80.22**
BERT$_{\text{large}}$	RankCSE$_{\text{listNet}}$	72.78	85.38	77.15	83.89	79.46	80.46	74.31	79.06
	+ angle	73.10	85.89	77.78	84.67	80.39	80.80	74.70	79.62
	RankCSE$_{\text{listMLE}}$	73.97	**86.18**	**78.73**	85.15	**80.91**	81.24	74.68	80.11
	+ angle	**74.35**	85.97	78.41	**85.18**	80.77	**81.38**	**74.83**	**80.13**
RoBERTa$_{\text{base}}$	PCL	68.20	81.05	72.68	**81.23**	80.02	**79.58**	**69.82**	76.08
	+ angle	**70.30**	**81.48**	**72.78**	81.18	**80.07**	79.37	68.41	**76.23**
	RankCSE$_{\text{listNet}}$	72.45	83.79	74.36	82.92	81.12	81.81	69.88	78.05
	+ angle	73.26	83.81	75.38	84.27	81.78	82.33	70.53	78.77
	RankCSE$_{\text{listMLE}}$	73.52	84.35	75.76	83.91	82.65	**82.88**	**70.88**	79.14
	+ angle	**74.24**	**84.54**	**76.07**	**84.41**	**82.67**	82.86	70.74	**79.36**
RoBERTa$_{\text{large}}$	RankCSE$_{\text{listNet}}$	71.80	82.09	73.76	81.96	79.03	80.41	70.57	77.09
	+ angle	73.19	84.01	75.91	84.81	81.11	82.76	70.82	78.94
	RankCSE$_{\text{listMLE}}$	73.86	84.14	76.41	85.25	**81.99**	**83.11**	71.65	79.49
	+ angle	**74.60**	**84.86**	**77.15**	**85.42**	81.67	82.99	**71.81**	**79.79**

Table 4: Performance of original PCL and RankCSE, and their angle-based version (denoted as '+angle'). We conduct each experiment using 5 different random seeds and report the average of the results, whose mean and standard deviation are reported in the Appendix (Table 10). We cannot run PCL based on the large models due to a shortage of our GPU memory (40GB). We report p-values for each baseline in the Appendix (Table 9), most of which are highly statistically significant (p < 0.001) except PCL and RankCSE-ListMLE on BERT-large.

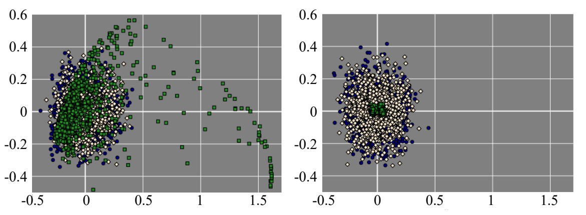


Figure 2: Visualization of 2D manifold representation space of (a) BERT-base and (b) RoBERTa-base, with different methods (PLMs: ■, SimCSE: ●, SimACE: ◆). We use 1000 random samples from the train dataset (Wiki), and apply PCA (Pearson, 1901) to approximate sentence embeddings. (b): RoBERTa-base model shows relatively narrower space, which may lead to high-performance gain of our angle-based approach.

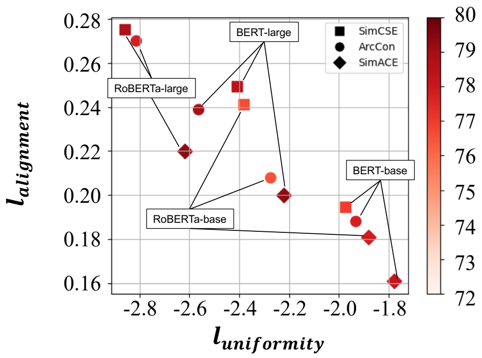


Figure 3: $\ell_{uniform} - \ell_{alignment}$ plot for contrastive methods with different similarity functions measured on the STS-B dev set. The colors of the points represent the average Spearman score on 7 STS tasks.

their size. Interestingly, SimACE is more effective in RoBERTa, which motivates us to explore the geometrical difference of semantic space between PLMs, as shown in Figure 2. From the visualization of two base models (BERT-base and RoBERTa-base), we suggest the following two intuitions.

Firstly, although the vanilla RoBERTa-base has a more anisotropic space than the vanilla BERT-base, the performance improvement for RoBERTa-base with SimACE is much larger than the performance improvement for BERT-base with SimACE. It seems likely that SimACE may be more discriminative in a narrow semantic space than SimCSE, as it densely aligns positive pairs to a greater extent. Secondly, we can observe that the semantic space optimized by SimACE is narrower than that of cosine similarity-based contrastive loss (SimCSE), which supports our intuitions that different instances will not be separated than a certain angular threshold. This also implies that there are meaningful factors rather than the wider size of the semantic space (*i.e.,* uniformity), and we will discuss these factors in the aspect of training dynamics in Sections 5.2 and 5.3.

5.2 Uniformity and Alignment Analysis

Firstly introduced into SRL by SimCSE (Gao et al., 2021), uniformity and alignment are the widely

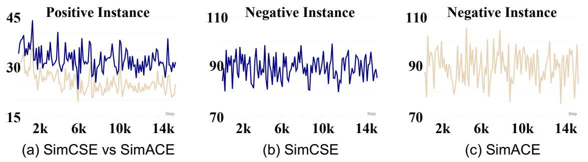


Figure 4: Change of angle (y-axis) between anchors and positive (a) and negative ((b)&(c)) instances during training on BERT-base. We average the angle values of all in-batch negative instances. We compare SimCSE (●) and SimACE (●). (a): SimCSE shows larger angle of positive instance (mean for SimCSE = 32.22 / mean for SimACE = 25.43) than SimACE. (b)&(c): SimCSE also shows a smaller change in the angle of negative instances (standard deviation for SimCSE: 4.67 / standard deviation for SimACE: 6.40).

utilized quantitative evaluation metrics that measure the quality of sentence representation after contrastive learning. Optimizing these two losses turned out to be equivalent to optimizing the contrastive loss under the assumption of infinite negative instances (Wang and Isola, 2020), where the former indicates how well the representation vectors are uniformly distributed, while the latter computes the distance between the anchor and the positive instance given the distribution of positive pairs. For both uniformity and alignment, the lower value indicates well-trained by contrastive learning. Each loss can be formulated as:

$$l_{uniform} \triangleq log \underset{x_i,x_j \sim P_{data}}{\mathbb{E}} e^{-t\|f(x_i)-f(x_j)\|_2^2}. \quad (10)$$

$$l_{alignment} \triangleq log \underset{x_i,x_j \sim P_{pos}}{\mathbb{E}} \|f(x_i) - f(x_j)\|_2^{\alpha}. \quad (11)$$

Figure 3 shows the uniformity-alignment plot for the baselines employed in Section 4.3. Aligned with our intuitions, SimACE enhances alignment in all PLMs by giving more attention to positive pairs. Notably, SimACE consistently exhibits a higher uniformity loss compared to the cosine similarity-based approaches. This occurs because SimACE non-aggressively pushes away negative instances with higher angle differences during the middle of training. These findings diverge from the aforementioned research which suggests that better uniformity leads to superior sentence representations, based on cosine similarity function (Gao et al., 2021; Chuang et al., 2022; Zhou et al., 2022; Zhang et al., 2022b). As a result, this prompts us further to explore the training dynamics of the gradient property.

5.3 Effect of Our Angle-based Approach

Among the several components that determine the training dynamics of contrastive learning, our study aims at developing a simple but more effective similarity function than the off-the-shelf cosine similarity. Although both SimACE and SimCSE achieve the goal of contrastive learning, there exists a visible difference in a gradient property during optimizing the loss function, as mentioned in Eq. 6. Figure 4 visualizes the difference by plotting the change of angle between representations to explore the difference in training dynamics.

In line with the contrastive objective, SimACE is also well-optimized toward the right direction ($\theta_{i,j} > \theta_{i,i'}$). Specifically, the results show that the hidden representation $\mathbf{z}_i$ derived from SimACE is strongly pushed toward the area where $\theta_{i,i'}$ is much smaller (around 25 degrees) than that of SimCSE (around 32 degrees). It confirms our intuition that the angle-based similarity function has a strong gradient signal at relatively small angles, which tends to pull similar sentences more strongly, as shown in Figure 1. Meanwhile, we can observe that SimACE has a more diverse similarity distribution for negative instances, as shown in Figure 4 (b) and (c). At the points where the angle gets larger, the strength of pulling and repelling becomes weaker since the magnitude of the gradient decreases. It aligns with the findings of Nie et al., 2022 that weak gradient signals at the area ($\theta_{i,j} > \theta_{i,i'}$) play a key role in contrastive learning for SRL.

6 Conclusion

We have proposed a novel angle-based similarity function for unsupervised contrastive learning of sentence representation, whose property delivers a more positive impact on training dynamics in SRL. Through extensive experiments, we have demonstrated that angle-based similarity can be a promising alternative to the traditional cosine similarity function. After finding different aspects of uniformity and alignment, we have also performed additional experiments dealing with training dynamics and visualization of semantic space to gain a deeper understanding. Furthermore, we have found that our idea can be effectively plugged into the recent state-of-the-art in SRL, boosting their performances. We hope that our work will be an important milestone for future research.

7 Limitation

While our proposal focuses on leveraging an angle-based distance between instances as a function for calculating a similarity between two different instances, it is important to note that there exist other alternatives that can be utilized to achieve the same objective, as shown in Appendix F.

We argue our main contribution lies in the fact that we introduce the framework of using an angle-based similarity function for predicting similarity between different sentences. In addition, we show that the utilization of the angle-based similarity function serves as a notable example of enhancing off-the-shelves methodologies. Therefore, we expect that researchers within the community can collaborate to improve the contrastive learning framework shortly by exploring several similarity functions in contrastive learning for unsupervised sentence representation learning. Moreover, there is abundant space for further progress in improving our angular-based contrastive learning. Further studies of analyzing the property of contrastive learning, such as gradient analysis, need to be undertaken for a deeper understanding of the framework.

On top of that, we believe it is feasible since our method builds on the foundational literature of the SimCSE baseline, which is extendable to multilingual settings (Wang et al., 2022), although we have not performed a multilingual scenario with our method. There is also scope for further analysis of contrastive learning and BERT-based models from both mathematical and theoretical perspectives.

8 Ethical Consideration

Considering intellectual property, we utilize sampled data and pre-trained models in huggingface for only scholar purpose. Like the previous study, there can be reported negative biases from training data (Wiki) of PLMs (Bender et al., 2021) used in our works. Besides them, we do not see any other ethical problems.

References

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

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

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

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.

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.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

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

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

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.

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

Ting Chen, Calvin Luo, and Lala Li. 2021. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845.

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

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

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

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

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*, pages 4690–4699.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27.

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

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2018. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

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

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. 2023. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6840–6849.

Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. 2021. Tweac: transformer with extendable qa agent classifiers. *arXiv preprint arXiv:2104.07081*.

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.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

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

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837.

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.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. *arXiv preprint arXiv:1708.00391*.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3:71–90.

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

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020b. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020c. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv:2008.09335*.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. Rankcse: Unsupervised sentence representations learning via learning to rank. *arXiv preprint arXiv:2305.16726*.

Xueqing Liu, Chi Wang, Yue Leng, and ChengXiang Zhai. 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*, pages 2–5.

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

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

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

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Zhijie Nie, Richong Zhang, and Yongyi Mao. 2022. On the inadequacy of optimizing alignment and uniformity in contrastive learning of sentence representations. In *The Eleventh International Conference on Learning Representations*.

James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish i would have loved this one, but i didn't–a multilingual dataset for counterfactual detection in product reviews. *arXiv preprint arXiv:2104.06893*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

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

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697.

Darsh J Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. *arXiv preprint arXiv:1809.02255*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y 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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, S Yu Philip, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for nlp tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

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*, pages 200–207.

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

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2019. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*.

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

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

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, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3597–3606.

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022. Pcl: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Wei Xu, Chris Callison-Burch, and William B Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

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.

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X Pham, Chang D Yoo, and In So Kweon. 2021a. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *International Conference on Learning Representations*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Xiao Zhang, Rui Zhao, Yu Qiao, and Hongsheng Li. 2020a. Rbf-softmax: Learning deep representative prototypes with radial basis function softmax. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 296–311. Springer.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021b. Bootstrapped unsupervised sentence representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5168–5180.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020b. 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.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022a. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11730–11738.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022b. 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)*, pages 4892–4903.

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

A Detailed Explanation of Datasets

We report the statistics of train, validation, test datasets of STS and 7 transfer tasks which are utilized in Section J: 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). Each detailed configuration can be found in Table 5 and Table 6, respectively. Following the literature, we use test sets for Table 1 results without using any additional validation sets.

Dataset	train	valid	test
STS12	-	-	3108
STS13	-	-	1500
STS14	-	-	3750
STS15	-	-	3000
STS16	-	-	1186
STS-B	5749	1500	1379
SICK-R	4500	500	4927

Table 5: Detailed configuration of STS datasets.

Dataset	train	valid	test
MR	10662	-	-
CR	3775	-	-
SUBJ	10000	-	-
MPQA	10606	-	-
SST-2	67349	872	1821
TREC	5452	-	500
MPRC	4076	-	1725

Table 6: Detailed configuration of 7 transfer datasets from SentEval.

B Implementation Details

Following SimCSE, which is a widely used baseline for unsupervised settings, we train SimACE using the two representative PLMs, BERT$_{\text{base}}$ & BERT$_{\text{large}}$ and RoBERTa$_{\text{base}}$ & RoBERTa$_{\text{large}}$. We use the [CLS] token as the sentence representation for training and save the best model checkpoint by using the validation score on the development set of STS-B.

Unsupervised STS tasks We conduct all SimCSE experiments based on the original paper's configuration. We choose a learning rate between [1e-5, 3e-5], batch size between [64, 512], and temperature = 0.05. In the case of ArcConLoss, We carry out grid-search of batch size between [16, 32, 64], learning rate between [1e-5, 3e-5], and temperature = 0.05. Detailed settings of SimACE's hyperparameters can be seen in Table 7.

Connection to Off-the-shelves For these experiments, we follow all settings of hyperparameters in the original paper: PCL and RankCSE. Since the introduction of the angle-based similarity function requires an additional margin term, we follow the

	batch_size	learning_rate	max_seq	eval_steps
$\text{BERT}_{\text{base}}$	64	3e-5	32	125
$\text{BERT}_{\text{large}}$	32	1e-5	32	125
$\text{RoBERTa}_{\text{base}}$	128	1e-5	32	125
$\text{RoBERTa}_{\text{large}}$	128	1e-5	32	125
	temperature	margin	eval_metric	pooler
$\text{BERT}_{\text{base}}$	0.06	10°	stsb	cls
$\text{BERT}_{\text{large}}$	0.06	10°	stsb	cls
$\text{RoBERTa}_{\text{base}}$	0.05	10°	stsb	cls
$\text{RoBERTa}_{\text{large}}$	0.05	10°	stsb	cls

Table 7: The hyperparameters that correspond to the best results of the STS tasks. stsb : Saving the best checkpoint of the model based on validation on STS-B dataset. The unit of margin value is degree ($^{\circ}$). cls : Using the representation of the [CLS] token, consisting of a linear layer and the following activation function.

same margin (m=10) as the vanilla SimACE implementation. Furthermore, there is no other grid-search for hyperparameter tuning.

Method	Similarity	Batch size	Epoch	Time
SimCSE	Cosine	64	1	64min
	ArcCon	64	1	76min
	Angular	64	1	68min
PCL	Cosine	64	1	134min
	Angular	64	1	130min
ListNet	Cosine	64	4	374min
	Angular	64	4	372min
ListMLE	Cosine	64	4	369min
	Angular	64	4	372min

Table 8: Comparison of training time between original cosine similarity-based method and angular similarity function in several baselines. We report the results of BERT-base model. Cosine : SimCSE-variants. ArcCon: ArcConLoss-based method. Angular : SimACE-variants. min: elapsed minutes.

C Training Efficiency

There may be concern about computational efficiency when using the arccosine function for our proposed angular similarity function. Dealing with this issue, we report the training time of SimCSE and SimACE on several baseline methods using in the main paper's experiments. We measure the required time for training when using a single NVIDIA Tesla A100 GPU (40GB memory). For a fair comparison, we use the same experimental settings, including batch size, epoch, and others, although their training configurations are different with each other. As seen in Table 8, we do not find any meaningful difference between the angular-based function and other baselines.

D Training Objective of Baseline Methods

We briefly introduce each method in Section 4.6, focusing on each one's loss function which is based on the cosine similarity. We simply replace the original similarity function with our angular-based one:

- **PCL** contrasts the anchor (x_i) with augmented positives (X^i) from a different discrete augmentation set ($\Delta^{(d)}$) and in-batch negatives, which models a group-wise relation (P-Cf) for cooperation across two peer networks ($f(\cdot)$ and $f'(\cdot)$):

$$p^{\text{P-Cf}}_{f,f'}(x_i) := \text{P-Cf}(x_i, \Delta^{(d)}; f, f')$$
$$= \text{softmax}(\{sim(f(x_i), f'(\hat{x}^i_k)/\tau)\}_{\hat{x}^i_k \sim \hat{X}^i} + \{sim(f(x_i), f'(x_j)/\tau)\}_{x_j \sim X \wedge j \neq i}), \quad (12)$$

where $sim(\cdot,\cdot)$ denotes cosine similarity between two different representations.

- **RankCSE** proposed cosine similarity-based loss terms for ranking consistency and ranking distillation. The ranking consistency loss aims to minimize Jensen-Shannon (JS) divergence:

$$L_{\text{consistency}} = \sum_{i=1}^{N} JS(P_i || Q_i), \quad (13)$$

where P_i and Q_i denote the probability distribution (λ) of similarity score lists ($S(x_i)$, $S(x_i)'$) obtained from independent networks $f(\cdot)$ and $f(\cdot)'$, respectively. In addition, this work explores two list-wise ranking methods, ListNet (Cao et al., 2007) and ListMLE (Xia et al., 2008), for ranking distillation:

$$L_{\text{rank}} = \sum_{i=1}^{N} rank(S(x_i), S^T(x_i)), \quad (14)$$

where $rank(\cdot,\cdot)$ indicates the list-wise method. $S(x_i)$ and $S^T(x_i)$ denote similarity

PLMs	SimCSE	ArcCon	PCL	RankCSE$_{\mathrm{listNet}}$	RankCSE$_{\mathrm{listMLE}}$
BERT$_{\mathrm{base}}$	0.001	0.05	0.12	0.001	0.001
BERT$_{\mathrm{large}}$	0.001	0.01	N/A	0.001	0.85
RoBERTa$_{\mathrm{base}}$	0.001	0.001	0.57	0.001	0.04
RoBERTa$_{\mathrm{large}}$	0.001	0.001	N/A	0.001	0.05

Table 9: Statistical significance of experimental results (p-value) across different random seeds. Most cases show statistically highly significant in terms of performance improvement.

PLMs	PCL		RankCSE$_{\mathrm{listNet}}$		RankCSE$_{\mathrm{listMLE}}$	
	Original	Ours	Original	Ours	Original	Ours
BERT$_{\mathrm{base}}$	$77.39_{\pm 0.22}$	$77.52_{\pm 0.39}$	$76.32_{\pm 0.12}$	$78.02_{\pm 0.26}$	$80.06_{\pm 0.08}$	$80.22_{\pm 0.06}$
BERT$_{\mathrm{large}}$	N/A	N/A	$79.06_{\pm 0.17}$	$79.62_{\pm 0.26}$	$80.11_{\pm 0.15}$	$80.13_{\pm 0.11}$
RoBERTa$_{\mathrm{base}}$	$76.08_{\pm 0.63}$	$76.23_{\pm 0.24}$	$78.05_{\pm 0.04}$	$78.77_{\pm 0.14}$	$79.14_{\pm 0.18}$	$79.36_{\pm 0.21}$
RoBERTa$_{\mathrm{large}}$	N/A	N/A	$77.09_{\pm 0.28}$	$78.94_{\pm 0.20}$	$79.49_{\pm 0.35}$	$79.79_{\pm 0.18}$

Table 10: Mean and standard deviation across 5 different runs of different methods with random seeds. Unfortunately, since large-size models trained by PCL with different random seeds show a gradient explosion, we report these results as N/A (Not Applicable or Not Available). We report p-values for each baseline in the Appendix (Table 9), which are highly statistically significant (p < 0.001).

score lists obtained from a student model and a teacher model. All the aforementioned similarity score lists are based on cosine similarity $sim(\cdot,\cdot)$ between two different inputs x_i and $x_i^{'}$.

The ranking consistency loss refers to maintaining consistency between two sentence representations obtained using different dropout masks by optimizing the Jensen-Shannon(JS) divergence between two similar sentence representations. RankCSE tries to guide the student model to learn better sentence representations by distilling the listwise ranking knowledge through ListNet (Cao et al., 2007) and ListMLE (Xia et al., 2008) algorithms, which minimize the cross entropy between the top one probability distribution and maximizing the likelihood of the ground truth permutation, respectively.

E Statistical Results of Experiments

In addition to Section 4.5, we report the full statistical information of our experimental results. These statistics include the statistical significance (p-value) and the standard deviation of performance on STS correlation. As seen in Table 9, most results, except two PCL results and a RankCSE-listMLE on BERT-large, show statistically highly significant. The calculated standard deviation of results for Table 4 is reported in Table 10. In line with the results of the main paper, plugging the angular-based method shows better performance and robustness compared to the original method using the cosine similarity function.

F Experiments of Different Objectives

We compare several candidates of different contrastive objectives with regard to sentence representation learning. These objectives include replacing the cosine similarity function with RBF, and 4 different losses proposed in Nie et al., 2022. RBF can be defined as below:

$$\phi(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{c}\|^2}{2\sigma^2}\right). \tag{15}$$

Considering the contrastive pairs, we set $\mathbf{c}$ as the anchor instance and calculate the similarity logits with all in-batch negative instances ($\mathbf{x}$). We also properly tuned the hyperparameter value σ by conducting grid-search. We report the overall results in Table 11. As seen in the table, our proposed method mostly shows better performance compared to other methods, except for the case of BERT-base. We think that the angular property may play a more important role in the larger models in terms of both model size and inductive bias (in general, RoBERTa is better than BERT).

G Detailed Results on MTEB benchmark

We evaluate several PLMs trained by SimACE on MTEB benchmark (Muennighoff et al., 2022). MTEB benchmark is designed to provide better evaluation for sentence embedding quality. The benchmark consists of several datasets including prior works and newly introduced by the paper.

Method	$BERT_{base}$	$BERT_{large}$	$RoBERTa_{base}$	$RoBERTa_{large}$
Ours(SimACE)	77.46	**79.02**	**77.87**	**79.14**
RBF	76.04	77.58	76.58	78.32
DCL$^{\heartsuit}$	71.13	72.73	73.18	72.43
MPT$^{\heartsuit}$	77.25	77.35	76.42	78.84
MET$^{\heartsuit}$	**78.38**	78.38	77.38	78.71
MAT$^{\heartsuit}$	77.76	77.76	76.95	78.82

Table 11: Comparative results of different optimization objectives, including different similarity functions and modified contrastive objectives. We report the averaged performance of different random seeds same with the Table 2. Each bold number and underlined number indicates the best performance within PLMs. DCL: Debiased contrastive objective. MPT: Minimum Dot Product Triplet Loss. MET: Minimum Euclidean Distance Triplet Loss. MAT: Minimum Angle Triplet Loss. ♡: Results from Nie et al., 2022.

PLMs	**Method**	Clas	Clus	Pair	Rank	Retr	STS	Sum	Avg.
$BERT_{base}$	original	61.66	30.12	56.33	43.44	10.59	54.36	29.82	38.33
	SimCSE	62.28	29.04	74.65	53.96	20.29	74.33	**30.10**	46.16
	SimACE	**63.56**	**33.87**	**75.25**	**54.92**	**22.09**	**75.70**	29.51	**48.19**
$BERT_{large}$	SimCSE	64.50	35.62	76.15	**55.96**	28.08	74.94	**31.00**	50.35
	SimACE	**64.83**	**38.09**	**77.26**	54.95	**30.15**	**75.97**	30.14	**51.62**
$RoBERTa_{base}$	SimCSE	64.00	34.32	74.65	53.96	19.82	73.96	28.43	47.33
	SimACE	**64.51**	**37.79**	**75.25**	**54.92**	**23.12**	**75.78**	**29.68**	**49.46**
$RoBERTa_{large}$	SimCSE	**65.28**	36.55	76.93	**55.44**	25.42	77.42	**30.84**	50.43
	SimACE	64.98	**38.92**	**77.33**	54.82	**28.44**	**77.79**	29.21	**51.66**

Table 12: Performance of SimACE on MTEB benchmark. A bold face number indicates the best performance within the PLMs. We report averaged results of different random seeds. Considering the space, we use abbreviation for a task name: Clas: 12 classification tasks, Clus: 11 clustering tasks, Pair: 3 pair classification tasks, Rank: 4 reranking tasks, Retr: 15 retrieval tasks, STS: 10 sts tasks, Sum: 1 summarization tasks.

There are all 56 datasets: 12 classification datasets are AmazonCounterfactual (O'Neill et al., 2021), AmazonPolarity (McAuley and Leskovec, 2013), AmazonReviews (McAuley and Leskovec, 2013), Banking77 (Casanueva et al., 2020), Emotion (Saravia et al., 2018), Imdb (Maas et al., 2011), MassiveIntent (FitzGerald et al., 2022), MassiveScenario (FitzGerald et al., 2022), MTOPDomain (Li et al., 2020c), MTOPIntent (Li et al., 2020c), ToxicConversations[3], and TweetSentimentExtraction[4], 11 cluster datasets are ArxivClusteringS2S, BiorxivClusteringS2S, BiorxivClusteringP2P, MedrxivClusteringP2P, MedrxivClusteringS2S [56], RedditClustering (Geigle et al., 2021), RedditClusteringP2P, StackExchangeClusteringP2P (Muennighoff et al., 2022), StackExchangeClustering (Geigle et al., 2021), and TwentyNewsgroupsClustering[7], 3 pair classification datsets are SprintDuplicateQuestions (Shah et al., 2018), TwitterSemEval2015 (Xu et al., 2015), and TwitterURLCorpus (Lan et al., 2017), 4 reranking tasks are AskUbuntuDupQuestions[8], MindSmall (Wu et al., 2020), SciDocsRR (Cohan et al., 2020), and StackOverflowDupQuestion (Liu et al., 2018), 15 retrieval datasets are from Thakur et al., 2021, 10 STS datasets are 8 from STS benchmark, STS22[9], and BIOSSES[10], and 1 summarization dataset is SummEval (Fabbri et al., 2021).

We report the averaged results within tasks in Table 12. As seen in Table, models trained by SimACE show considerable performance compared to SimCSE. Specifically, 2 base PLMs trained by SimACE show better performance on all tasks, while 2 large PLMs trained by SimACE show better performance on most tasks except classification, reranking, and summarization task. Nonetheless, SimACE outperforms SimACE on STS, along the lines with results of main experiment (Table 1).

[3]https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification

[4]https://www.kaggle.com/competitions/tweet-sentiment-extraction

[5]https://arxiv.org/help/api/

[6]https://api.biorxiv.org/

[7]https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

[8]https://github.com/taolei87/askubuntu

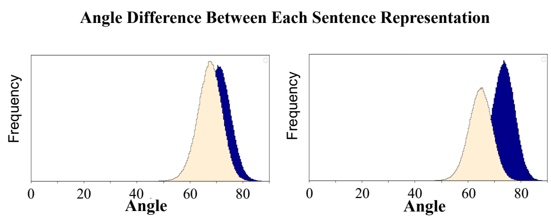
[9]https://competitions.codalab.org/competitions/33835

[10]urlhttps://tabilab.cmpe.boun.edu.tr/BIOSSES/DataSet.html

PLMs	Method	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
BERT$_{\text{base}}$	SimCSE	81.37	86.49	94.46	88.66	84.95	87.60	74.32	85.41
	with MLM	81.64	86.81	**95.76**	88.32	85.94	89.40	73.74	85.94
	ArcCon	81.31	85.80	94.44	88.96	**88.56**	87.40	74.43	85.41
	with MLM	82.26	87.74	95.57	88.45	85.72	**91.60**	74.84	86.60
	SimACE*	81.19	85.22	94.42	**89.14**	86.05	86.60	75.71	85.48
	with MLM*	**82.63**	**87.92**	95.68	88.91	86.33	91.00	**76.41**	**86.98**
BERT$_{\text{large}}$	SimCSE	84.30	87.98	94.86	88.78	89.51	93.00	74.61	87.58
	with MLM	85.78	89.72	95.83	87.94	90.83	93.00	72.87	88.00
	ArcCon	85.34	88.98	95.32	89.58	**91.27**	89.40	75.71	87.94
	with MLM	85.77	90.04	**95.98**	89.01	91.05	**93.40**	75.36	**88.66**
	SimACE*	84.34	89.51	95.24	**89.88**	90.61	92.40	**76.00**	88.28
	with MLM*	**86.15**	**90.33**	95.81	88.89	91.16	92.60	75.54	88.64
RoBERTa$_{\text{base}}$	SimCSE	81.75	86.97	93.43	87.28	86.99	84.40	75.01	85.12
	with MLM	84.14	89.04	94.49	88.07	89.24	87.20	74.38	86.65
	ArcCon	81.61	87.36	93.22	87.65	87.86	85.60	76.00	85.61
	with MLM	83.36	88.90	94.42	87.54	89.40	**89.80**	**76.81**	87.18
	SimACE*	81.87	87.36	92.87	87.54	86.93	87.00	74.61	85.45
	with MLM*	**84.35**	**89.57**	**94.65**	**88.28**	**90.28**	**89.80**	75.19	**87.45**
RoBERTa$_{\text{large}}$	SimCSE	83.17	88.40	94.08	88.57	87.53	91.20	72.23	86.45
	with MLM	83.00	87.87	94.64	87.38	87.92	90.80	**75.07**	86.67
	ArcCon	83.30	**89.38**	93.59	88.59	88.63	92.40	74.03	87.13
	with MLM	76.56	64.69	90.41	70.25	84.84	40.60	66.38	70.53
	SimACE*	82.90	88.90	93.60	**88.91**	87.64	91.60	73.04	86.66
	with MLM*	**84.56**	88.50	**94.85**	88.68	**89.07**	**93.00**	74.09	**87.54**

Table 13: Performance of different unsupervised contrastive learning methods on transfer tasks. Each bold number and underlined number indicates the best and second performance best within the PLMs, respectively. *: Our method.

H Deeper Analysis of Uniformity and Alignment

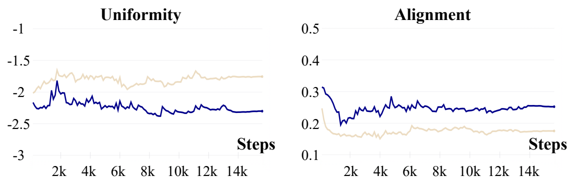


(a) BERT-Base (b) RoBERTa-Base

Figure 5: Histogram of the angle between each sentence representation. We use the BERT-base model trained by SimCSE (●) and SimACE (●).

To intuitively understand the characteristic of SimACE, we visualize the histogram of the angle between representations, as shown in Figure 5. SimCSE plots a higher average on angles than SimACE. From the results, we interpret that the lower angular average results in better alignment than SimCSE because it pulls the positive sample at the beginning of training and doesn't push the negative far enough when past the middle of training.

Following the literature, we also plot the change of uniformity and alignment during contrastive learning. We observe that SimACE improves alignment more than SimCSE, while its uniformity is getting worse during training. In the early stages of training, Figure 6 shows that SimACE's alignment drops below 0.2, which verifies our intuitions that the property of gradient and the training dynamics of SimACE can lead to better alignment, as we have discussed in Section 5.2. Moreover, as depicted in the figure, a higher value of uniformity than SimCSE also backs up our assumption of an angle-based approach.



(a) Uniformity (b) Alignment

Figure 6: Uniformity and Alignment of BERT-base trained by SimCSE (●) and SimACE (●).

I Training Dynamics of Angle with Different Temperatures

Motivated by Section 3.1, we further analyze the role of temperature in terms of training dynamics. In particular, we conduct additional experiments similar to Section 5.3, by using BERT-base trained by SimACE with 3 different temperature

PLMs	Method	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
$\text{BERT}_{\text{base}}$	SimACE	**71.63**	**83.44**	**76.65**	**83.85**	**79.95**	**79.99**	**71.86**	**78.20**
	$+m = 0$	70.20	81.76	75.56	82.44	79.52	78.94	71.09	77.08
	$+m = -10$	64.73	78.83	70.47	79.60	74.67	74.92	70.98	73.46
$\text{BERT}_{\text{large}}$	SimACE	**73.89**	**85.07**	**77.67**	**84.87**	**79.18**	**79.96**	**74.61**	**79.32**
	$+m = 0$	72.39	84.12	76.92	83.88	79.13	79.53	73.99	78.57
	$+m = -10$	69.68	83.32	74.35	81.00	78.62	78.42	74.04	77.06
$\text{RoBERTa}_{\text{base}}$	SimACE	**70.50**	**84.16**	**76.33**	**83.38**	**82.45**	**82.24**	**69.69**	**78.39**
	$+m = 0$	70.38	83.19	74.85	82.86	80.74	80.65	69.04	77.39
	$+m = -10$	67.35	80.29	71.90	81.56	79.73	79.52	69.12	75.64
$\text{RoBERTa}_{\text{large}}$	SimACE	**72.12**	**84.41**	**77.25**	**85.05**	**81.92**	**83.35**	**71.37**	**79.35**
	$+m = 0$	71.92	84.12	76.95	84.76	80.99	82.98	71.14	78.98
	$+m = -10$	67.68	80.44	72.47	81.68	78.66	79.27	71.07	75.90

Table 14: Performance of SimACE with subtracting margin values on STS tasks. A bold face number indicates the best performance within the PLMs. All results are based on default random seed (42) same with Table 1. $+m$: A different margin value is applied to SimACE. -10 indicates the additive margin (see margin term in Equation 9).

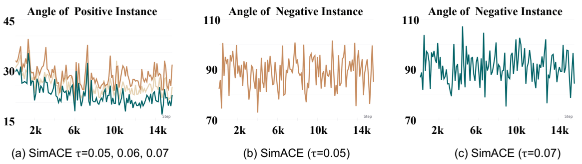


Figure 7: Change of angle between anchor and positive, negative instance during training on BERT-base. We average the angle values of all in-batch negative instances. We compare SimACE with different temperatures (0.05, 0.06, 0.07). **(a), (b), (c)**: A smaller temperature (0.05, ●) leads to a narrower range of angles (larger positive angle (mean = 28.22), smaller negative angle (mean = 88.75)), while a larger temperature (0.07, ●) leads to the wider range of angles (smaller positive angle (mean = 22.65), larger negative angle (mean = 90.90)).

values. For a fair comparison, we choose $\tau = 0.05$, which is the same as SimCSE, $\tau = 0.06$ (original SimACE's hyperparameter as seen in Table 7), and a larger value $\tau = 0.07$.

As we mentioned before, the temperature is related to the entropy of sentence embedding since it plays a role in altering gradient weight for negative instances. Concretely, the temperature value is proportional to the entropy of the distribution. It indicates that higher temperature leads to higher entropy so that embedding space becomes more tolerant of similar samples and thus improves the alignment, while lower temperature leads to lower entropy which improves uniformity.

Similar to findings of the role of temperature, we may assume two premises: (1) InfoNCE loss with high temperature will repulse every negative sample equally; (2) InfoNCE loss with low temperature will give more gradient weight to the negative instance which is more similar to anchor. These assumptions also align with our intuition from Equation 4. We can infer that the inverse of temperature value shows a similar pattern with the derivative of the similarity function, which we find some notable points in Section 5. Still, there is a major difference between the temperature and the similarity function: the temperature is a constant value for all instances.

As seen in Figure 7, the results partially satisfy our assumptions. First, higher temperature leads to improving alignment (Figure 7(a)). In contrast, it is interesting to see that a lower temperature value does not lead to an improvement in uniformity (Figure 7(a) and (b)). This result is an unanticipated finding since it violates both previous studies in the field of VRL and our intuition based on gradient analysis. We think that the anisotropic space of PLMs and the smaller number of negative instances may be problematic since degeneration to a simple contrastive loss due to lower temperature does not have enough power to equally push all negative instances.

J Results of Transfer Tasks

Following the literature, we also compare different contrastive methods on the off-the-shelves transfer tasks. We first freeze the feature extractor of sentence embeddings and then train a classifier. We conduct experiments using a standard configuration from SentEval(Conneau and Kiela, 2018), which uses 10-fold evaluation protocols to report the final test results. For fair comparison to the baseline SimCSE, we also train AngConLoss and SimACE with MLM (Masked Language Modeling) (Devlin et al., 2018), which is a typical pre-trained method for a BERT-like model, and report these results.

As seen in Table 13, SimACE shows a performance improvement compared to the baseline SimCSE. Moreover, similar to the SimCSE, we find that adding the MLM also improves the performance of vanilla SimACE. This backs up experimental results about the extensibility of SimACE, which was mentioned before in Section 4.6.

K Ablation of Angular Margin

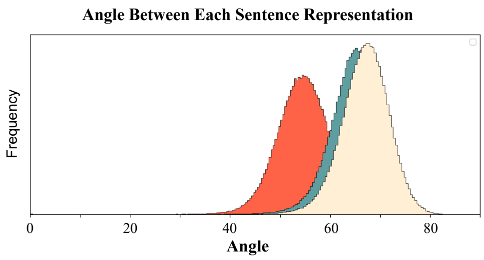
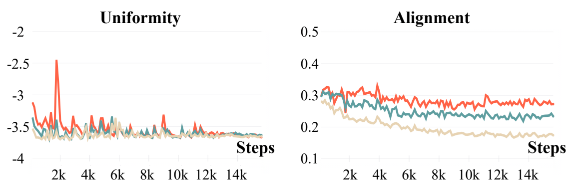


Figure 8: Histogram of the angle between each sentence representation. We use BERT-base model trained by SimACE with different margins: ● is m=10 (original), ● is m=0 (no margin), and ● is m=-10 (additive margin).

In addition to Figure 8, we also evaluate several SimACE with different margins on STS benchmarks within PLMs. Specifically, we compare 3 cases: our proposed subtractive margin, additive margin ($m = -10$) similar to ArcCSE (Zhang et al., 2022b), and no margin ($m = 0$). As seen in Table 14, SimACE method with the original subtractive margin shows the best averaged performance on STS tasks. While a vanilla SimACE with no margin shows comparable performance to the baseline, the method with an additive margin suffers severe performance degradation.



(a) Uniformity (b) Alignment

Figure 9: Uniformity and Alignment of the BERT-base model trained by SimACE with different margin (●: $m = 10$ (original), ●: $m = 0$ (no margin), and ●: $m = -10$ (additive margin)). Averaged STS correlation scores for the original SimACE, SimACE with no margin, and with additive margin are 78.20, 76.69, and 73.46, respectively.

In addition, we drag the observation into the angular margin to further understand the relationship between angular distribution and alignment. Therefore, we conduct supplementary experiments to plot uniformity and alignment of SimACE with varying margin $m \in$ {-10, 0, 10}. As shown in Figure 9 (a), the angular margin leads the inductive bias against alignment, showing that margin penalty for negative perturbations encourages the representations to well-align due to the property of large gradient magnitude at the beginning of training.