# Simple Temperature Cool-down in Contrastive Framework for Unsupervised Sentence Representation Learning

Anonymous ACL submission

## Abstract

In this paper, we proposes a simple, tricky method to improve sentence representation of unsupervised contrastive learning. Even though contrastive learning has achieved great performances in both visual representation learning (VRL) and sentence representation learning (SRL) fields, we focus on the fact that there is a gap between characteristics and training dynamics of VRL and SRL. We first examine the role of temperature to bridge the gap between VRL and SRL, and find some temperature-dependent elements in SRL; *i.e.*, a higher temperature causes overfitting of the uniformity while improving the alignment in earlier phase of training. Then, we design a *temperature cool-down* technique based on this observation, which helps PLMs to be more suitable for contrastive learning via preparation of uniform representation space. Our experimental results on widely-utilized benchmarks demonstrate the effectiveness and extensiblity of our method.

## 1 Introduction

One of the most important breakthroughs in unsupervised representation learning is the introduction of contrastive learning into the field of deep learning (Chen et al., 2020; He et al., 2020). In the past few years, a number of studies have sought to analyze the success of contrastive learning. For example, optimizing contrastive learning can satisfy two different properties of representations on the hypersphere, which are asymptotically quantified by the uniformity and alignment loss (the former leads to a uniformly distributed representation space and the latter makes a positive instance closer to an anchor (Wang and Isola, 2020)). These approaches have also been widely adopted in the SRL (sentence representation learning) literature, where SimCSE (Gao et al., 2021) successfully implemented the framework for unsupervised contrastive learning by constructing a straightforward dropout-based positive pair.

There has been a steady increase of interest in the role of a temperature ($\tau$) used in NT-Xent loss (normalized temperature cross entropy loss) (Chen et al., 2020). For example, a temperature is inversely proportional to uniformity by controlling the strength of the penalty on negative samples (Wang and Liu, 2021). Also, a higher temperature can lead to a collapse (Zhang et al., 2021a), *i.e.*, degeneration solution of representation learning (Chen et al., 2020; Chen and He, 2021). However, most studies have focused only on VRL (visual representation learning), and little information is known about the role of temperature especially for SRL. In addition, there are several differences between the two fields; *i.e.*, the number of batch size (smaller in SRL), the usage of PLMs (pre-trained language models)), and a temperature value (relatively lower in SRL).

In our study, we first investigate the role of temperature in SimCSE. Interestingly, we find that the higher temperature in the earlier phase of training shows lower alignment and higher uniformity loss, indicating that higher temperature alleviates the excessive repelling of negative instances that are too close to the anchor due to the anisotropic space of PLMs; *i.e.*, feature vectors form a narrow cone-like representation space (Ethayarajh, 2019; Wang et al., 2019; Li et al., 2020). Theoretically, NT-Xent loss with higher temperature will degenerate to the vanilla contrative loss, which repels every negative sample with equal strength (Zhang et al., 2021a). We assume that this can be effective for SRL different from typical VRL works whose models' parameters are initialized by normal distribution[1] and trained from scratch.

Based on the above motivation, we propose *temperature cool-down*, a simple technique specially designed for unsupervised SRL. We set a higher temperature in the first few steps on earlier training,

[1]Thus, their representation spaces are uniformly distributed at the beginning.

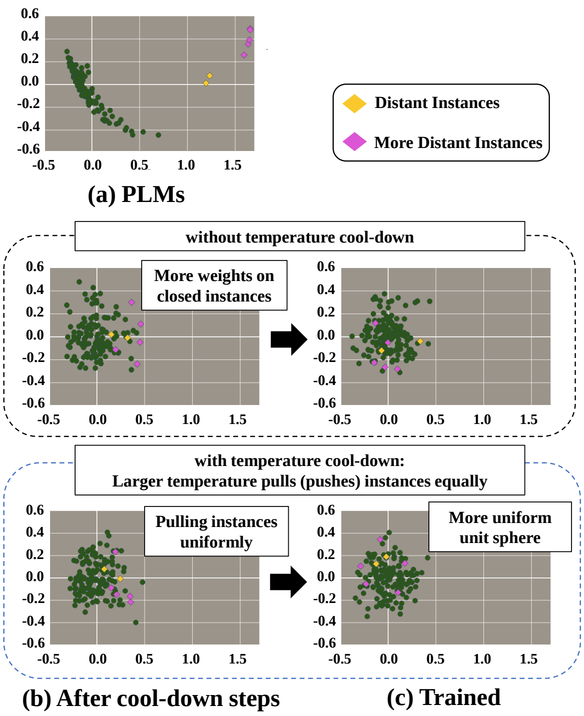


Figure 1: PCA visualization of the representation space during contrastive learning with/without temperature cool-down. (a): Following the literature, BERT-base shows the anisotropic representation space. (b): A model trained with temperature cool-down pulls distant instances (colored pink) more uniformly. (c): A representation space built by temperature cool-down leads to a more uniform unit hypersphere.

and then *cool down* the temperature to the original value. The higher temperature can mitigate the phenomenon where, due to the anisotropic nature of PLMs' representation spaces, a smaller temperature in the early phase of training leads to unintended pulling and pushing of instances because of their excessive proximity to the anchor. In this way, temperature cool-down makes the PLMs' representation spaces better suited for dropout-noise based contrastive learning. Empirically, our temperature cool-down improves SimCSE's performance on the unsupervised sentence representation benchmarks. It also has the extensibility to be used in different SRL methods based on SimCSE.

## 2 Proposed Method

### 2.1 Preliminary and Motivation

**Unsupervised Sentence Representation Learning** Previous studies in the field of SRL have focused on the computation of continuous and static word representations based on the idea of word2vec (Mikolov et al., 2013; Hill et al., 2016; Logeswaran and Lee, 2018). Since the successful introduction of PLMs (Devlin et al., 2018; Liu et al., 2019), several methods using PLMs to generate sentence representations have been reported, but PLMs suffered from some problems such as an anisotropic space (Ethayarajh, 2019).

In line with VRL, previous attempts to apply contrastive learning to SRL have focused on constructing well-crafted pairs to learn a better sentence representation (Sun et al., 2020; Zhang et al., 2020, 2021b; Giorgi et al., 2021; Kim et al., 2021; Yan et al., 2021). Recently, many works have followed the typical SimCSE baseline (Gao et al., 2021), which uses dropout-noise based augmentation. SimCSE utilized NT-Xent loss:

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

where $sim()$, $\mathbf{z}_i$, $\mathbf{z}'_i$, and $\mathbf{z}'_j(i \neq j)$ denote a similarity function, a hidden representation of the anchor, a positive instance, and a negative instance.

**Role of Temperature** According to the gradient of contrastive loss, one of the roles of temperature is to control the distribution of negative gradients (Wang and Liu, 2021). Since the gradients with respect to both positive and negative similarity are proportional to the inverse of the temperature ($\frac{1}{\tau}$), the contrastive loss is the hardness-aware function by which temperature determines the strength of repelling negative samples. For example, a lower temperature boosts the gradient of instances closed to the anchor and thus improves the uniformity (Robinson et al., 2021). In contrast, a higher temperature leads to a balanced weight of gradients and may suffer both performance degradation and collapse of the representation (Zhang et al., 2021a).

We assume that there are *temperature-dependent* factors in SRL due to the nature of PLMs. If there is a strong relationship, a subtle change in the temperature value may lead to an improvement in representational power. This assumption raises the question regarding an inconclusive reason for the lower temperature value used in SimCSE.

### 2.2 Observation

In this section, we examine the effect of temperature in terms of the representation space – *i.e.*, the uniformity and alignment loss –, and the quantitative evaluation results. As shown in Figure 2, the uniformity is proportional to the temperature while

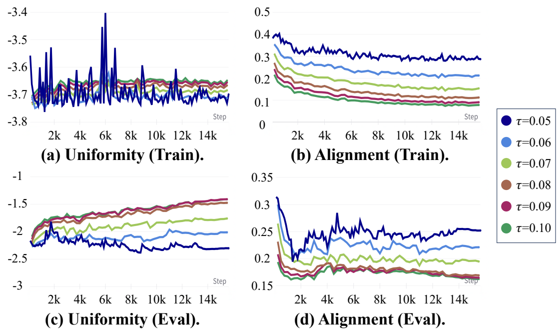


Figure 2: Uniformity and alignment of BERT-base trained by SimCSE with different temperature ($\tau$).

| PLMs | $\tau$ | Avg.STS | PLMs | $\tau$ | Avg.STS |
|---|---|---|---|---|---|
| BERT | <u>0.05</u> | 76.95 | RoBERTa | <u>0.05</u> | 76.64 |
| (base) | 0.06 | 76.96 | (base) | 0.06 | 76.61 |
| | 0.07 | 76.37 | | 0.07 | 75.57 |
| | 0.08 | 75.08 | | 0.08 | 74.86 |
| | 0.09 | 73.26 | | 0.09 | 73.73 |
| | 0.10 | 71.92 | | 0.10 | 72.36 |

Table 1: Results of SimCSE with different temperature on the STS evaluation tasks. An underlined temperature indicates the original SimCSE's hyperparameter.

the alignment is inversely proportional, which is consistent with previous results. Also, a higher temperature leads to worse performance (Table 1), which is similar to the finding of Zhang et al., 2021a. At the same time, we observe that there are unprecedented results; a higher temperature not only leads to *overfitting* of the uniformity (it gets worse[2] in the evaluation datasets), but also improves the alignment. This tendency is more pronounced in the early stages of training.

### 2.3 Temperature Cool-down

Motivated by the previous findings and our observations, we design a simple yet effective technique for contrastive learning in SRL, named *temperature cool-down*. Its logic is similar to the widely-used *warm-up* technique in learning rate schedulers (He et al., 2016, 2019). We start by setting a *initial temperature* ($\tau_i$) value that is larger than the original temperature ($\tau$) in earlier training steps. After a certain ratio of steps ($r_s$), we cool down the temperature to the original one. There are many possible ways to implement an effective cool-down process. In this paper, we explore two candidates: **T**emperature **C**ool-down with **C**onstant (TCC) and with **S**tep function (TCS), each formulated by:

$$\tau_{TCC,t} = \begin{cases} \tau_i, & \text{if } t \in [1, r_s \cdot s) \\ \tau. & \text{otherwise} \end{cases} \quad (2)$$

[2] Both smaller uniformity and alignment are better.

$$\tau_{TCS,t} = \begin{cases} \tau_i, & \text{if } t \in [1, 0.5 \cdot r_s \cdot s) \\ \frac{\tau_i + \tau}{2}, & \text{if } t \in [0.5 \cdot r_s \cdot s, r_s \cdot s) \\ \tau. & \text{otherwise} \end{cases} \quad (3)$$

where $t$, $\tau$, $\tau_i$, $s$, and $r_s$ denote a current training step, original temperature, initial temperature, total training steps, and step ratio, respectively. TCS uses a simple median of the temperature between $\tau_i$ and $\tau$ in the middle of the cool-down steps. We simply divide the TCS steps by $\frac{1}{2}$.

Since the representation spaces of PLMs are anisotropic, lower temperature in the early stages of training can lead to unintended pulling/pushing of instances due to excessive closeness towards the anchor (see Figure 1). This can be mitigated by higher temperature, whose role is to equally pull/push instances regardless of their closeness. In this respect, temperature cool-down prepares the representation spaces of PLMs to be more suitable for dropout-noise-based contrastive learning.

## 3 Experiments

### 3.1 Implementation Details

**Training Setups** We conduct grid search to determine the optimal hyperparameters; initial temperature ($\tau_i$) $\in$ [0.05, 0.014], step ratio ($r_s$) $\in$ [0.01, 0.03], and batch size $\in \{64, 512\}$. We train our models for 1 epoch and evaluate the model every 250 steps on the STS-B development set, following the literature. Also, we train SimCSE based on the paper's hyperparameters configuration.

**Network Implementation** We train SimCSE with temperature cool-down using the pre-trained checkpoints of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) downloaded from huggingface (Wolf et al., 2019). Following SimCSE, we also consider a [CLS] hidden representation as the sentence representation (Gao et al., 2021).

### 3.2 Unsupervised STS Tasks

**Benchmark** We train all models on randomly sampled datasets from English Wikipedia ($10^6$), which is same with the baseline (Gao et al., 2021). We evaluate them on typical sentence representation benchmark: 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 consist of pair of sentences of which similarity score's

| PLMs | Method | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| BERT$_{\text{base}}$ | first-last ♣ | 39.70 | 59.38 | 49.67 | 66.03 | 66.19 | 53.87 | 62.06 | 56.70 |
| | SimCSE | 71.64 | 82.68 | 75.81 | 82.25 | 78.60 | 78.93 | 68.76 | 76.95 |
| | + TCC | **72.52** | **83.83** | 76.60 | 83.29 | **79.60** | **79.60** | **71.26** | **78.10** |
| | + TCS | 72.37 | 83.79 | **76.65** | **83.37** | 79.42 | **79.60** | 71.13 | 78.05 |
| BERT$_{\text{large}}$ | SimCSE | 70.80 | **85.58** | 77.34 | 84.27 | 79.31 | 79.07 | 72.82 | 78.46 |
| | + TCC | **71.50** | 85.25 | 77.09 | **84.43** | 79.12 | 80.21 | **74.45** | **78.86** |
| | + TCS | 71.23 | 85.19 | **77.43** | 84.12 | **79.39** | **80.26** | 73.85 | 78.78 |
| RoBERTa$_{\text{base}}$ | first-last ♣ | 40.88 | 58.74 | 49.07 | 65.63 | 61.48 | 58.55 | 61.63 | 56.57 |
| | SimCSE | 68.65 | 81.70 | 73.44 | 82.30 | 81.09 | 80.51 | 68.76 | 76.64 |
| | + TCC | 69.79 | **82.69** | **74.70** | 82.63 | 81.19 | **82.13** | **69.91** | **77.58** |
| | + TCS | **70.01** | 82.56 | 74.43 | **82.66** | **81.63** | **81.56** | 69.38 | 77.46 |
| RoBERTa$_{\text{large}}$ | SimCSE | 70.85 | 83.67 | 75.83 | 84.24 | 80.27 | 82.42 | 72.41 | 78.53 |
| | + TCC | **71.08** | **84.60** | **76.56** | **84.97** | **80.37** | **83.18** | 71.72 | **78.93** |
| | + TCS | 70.40 | 83.65 | 75.19 | 84.95 | **80.37** | 81.80 | **73.40** | 78.54 |

Table 2: Performance of different unsupervised contrastive learning methods on the STS tasks (Spearman's correlation). Each bold number and underlined number indicates the best and second best performance within the PLMs, respectively. ♣: Results from Gao et al., 2021.

range is from 0 to 5. We utilize SentEval (Conneau and Kiela, 2018) for evaluation.

**Results** Table 2 shows the experimental results. Applying temparture cool-down boosts the performances; both TCC and TCS show better performance in most cases compared with the original SimCSE: nearly 1.5% on BERT-base, 1.4% on RoBERTa-base, 0.5% on BERT-large, and 0.5% on RoBERTa-large.

**Applying to ArcCSE** Here, we applied our temperature cool-down to ArcCSE (Zhang et al., 2022), which is one of the promising baselines extended from SimCSE. It proposed an angular margin contrastive loss (ArcConLoss), which introduces an angular margin term in the similarity function. It also proposed the extra Triplet loss, which requires additional preprocessed data. However, since the data is not accessible, we cannot reproduce the extra Triplet loss. We therefore report the results of ArcCSE without the Triplet loss in Table 3. We follow ArcCSE's default configuration along with our parameters; $\tau_i$ is 0.01 and $r_s \in [0.011, 0.02]$ with a step size of 0.001. We observe that applying temperature cool-down improves the performance, and even shows better performance than the original ArcCSE with the Triplet loss in BERT-base. This result is noteworthy because the extra Triplet loss requires much more computational resources, while our cool-down technique does not.

### 3.3 Uniformity and Alignment

We track the change of uniformity and alignment loss in STS-B development sets. Figure 3 visualizes 3 different methods on BERT-base (more results are in Appendix G), easing the uniformity and improving the alignment in earlier phase by

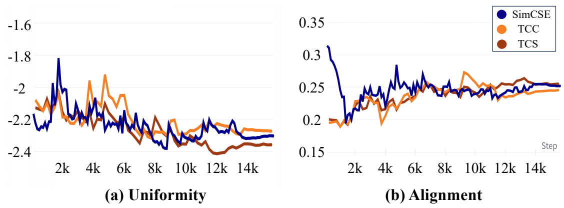


Figure 3: Uniformity and alignment on BERT-base using temperature cool-down.

| PLMs | Method | Avg.STS |
|---|---|---|
| BERT$_{\text{base}}$ | ArcCSE w/o Triplet loss | 77.76 |
| | + TCC | **78.20** |
| | + TCS | 78.09 |
| | ArcCSE ♡ | 78.11 |
| BERT$_{\text{large}}$ | ArcCSE w/o Triplet loss | 78.93 |
| | + TCC | 79.11 |
| | + TCS | 79.23 |
| | ArcCSE ♡ | **79.37** |

Table 3: Results of ArcConLoss with temperature cool-down. ♡: Results from Zhang et al., 2022.

temperature cool-down (steps < 1k) leads to more stable uniformity dynamics (smaller standard deviation). Also, the uniformity and alignment loss for the best checkpoint are better than vanilla SimCSE (see Appendix G).

## 4 Conclusion

We explore a simple, yet tricky, technique to control the temperature value of vanilla contrastive loss, which is widely used in the SRL literature. Motivated by previous studies in VRL and our empirical observations, we design a temperature cool-down that accelerates a higher temperature in earlier training steps and then cools down to the original, lower temperature. It shows performance improvement on various STS tasks, and also has many possibilities for plugging into other contrastive frameworks and designing effective variants.

## 5 Limitation

Although there can be a lot of possibilities of temperature cool-down variants, this paper suggests a few of simple functions. Similar to the learning rate warm-up, there may be effective candidates such as the exponential decay function or cosine function. In addition, there is a lack of mathematical grounding for the proposed approach. Nonetheless, we think that further experiments for gradient analysis can back up the success of our temperature cool-down. We leave exploration towards these researches in the future work.

The results reported in Table 2 may be interpreted as marginal, especially in terms of RoBERTa. As we mentioned before, temperature cool-down is a simple technique for well-preparing PLMs' representation spaces, assuming they initially look like narrow-cone. Thus, we measure the uniformity losses of *untrained* PLMs using in-batch samples (equally 64 for 4 models). Interestingly, we find that the initial uniformity losses of RoBERTa based models (RoBERTa-base:-0.1095, RoBERTa-large:-0.2503) are much smaller than BERT based models (BERT-base : -1.3086, BERT-large : -1.8705). We then visualize the representation spaces of RoBERTa models, which are not included in main paper, and find that they already look similar to cool-down setups (see Figure 1(b)) though those visualizations are limited to 2d manifold representation space. Still uncertain, but we believe this may be the reason of the marginal performance improvement.

More experimental results, which are not included in the main paper due to a limited space, can be found in Appendix. These include the robustness toward different random seeds experiments (Appendix D), evaluation on transfer tasks (Appendix E), and detailed results of the uniformity and alignment (Appendix G).

## 6 Ethical Consideration

We use datasets and pre-trained models in huggingface for only scholar purpose. Following the literature, reported negative biases from training data (English Wikipedia) of PLMs (Bender et al., 2021) can also be found in our works. In addition, there are not 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.

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.

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.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

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

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)*.

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.

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, Online and Punta Cana, Dominican Republic. 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.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 558–567.

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.

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.

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

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

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*.

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

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

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

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*.

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.

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.

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.

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.

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*.

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*.

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. 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.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903.

## A Dataset Details

| 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 4: Detailed configuration of 7 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 5: Detailed configuration of 7 transfer datasets.

We report the statistics of the training, validation, and test sets of the 7 STS evaluation tasks, as well as the 7 transfer tasks which are utilized in Section E: 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). The detailed configuration of the datasets for each evaluation scenario can be found in Table 4 and Table 5, respectively. Following

| **TCC** | batch_size | learning_rate | temp ($\tau$) | init_temp ($\tau_i$) | steps_ratio ($r_s$) |
|---|---|---|---|---|---|
| $\text{BERT}_{\text{base}}$ | 64 | 3e-5 | 0.05 | 0.10 | 0.014 |
| $\text{BERT}_{\text{large}}$ | 64 | 1e-5 | 0.05 | 0.10 | 0.015 |
| $\text{RoBERTa}_{\text{base}}$ | 128 | 1e-5 | 0.05 | 0.07 | 0.013 |
| $\text{RoBERTa}_{\text{large}}$ | 256 | 3e-5 | 0.05 | 0.06 | 0.013 |
| **TCS** | batch_size | learning_rate | temp ($\tau$) | init_temp ($\tau_i$) | steps_ratio ($r_s$) |
| $\text{BERT}_{\text{base}}$ | 64 | 3e-5 | 0.05 | 0.10 | 0.028 |
| $\text{BERT}_{\text{large}}$ | 64 | 1e-5 | 0.05 | 0.10 | 0.018 |
| $\text{RoBERTa}_{\text{base}}$ | 128 | 1e-5 | 0.05 | 0.07 | 0.014 |
| $\text{RoBERTa}_{\text{large}}$ | 256 | 3e-5 | 0.05 | 0.07 | 0.020 |

Table 6: The hyperparameters corresponding to the best results of the STS tasks.

the literature, we use test sets for Table 2 results without using any additional validation sets.

## B Detailed Implementation

Following the literature, 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. We conduct all SimCSE experiments based on the original paper's configuration. We choose learning rate between [1e-5, 3e-5], batch size between [64, 512], and temperature = 0.05. In the case of the initial temperature and cool-down step ratio, we carry out grid-search of the initial temperature between [0.06, 0.12], step ratio between [0.01, 0.03] by increasing each value by 0.01. We do not change the original temperature value ($\tau$=0.05, chosen by SimCSE). Detailed settings of the hyperparameters can be found in Table 6.

| PLMs | Method | uniformity(↓) | alignment(↓) |
|---|---|---|---|
| $\text{BERT}_{\text{base}}$ | SimCSE | -2.101 | 0.2073 |
| | + TCC | **-2.124** | 0.1934 |
| | + TCS | -2.112 | **0.1924** |
| $\text{BERT}_{\text{large}}$ | SimCSE | -2.410 | 0.2493 |
| | + TCC | **-2.586** | 0.2482 |
| | + TCS | -2.518 | **0.2457** |
| $\text{RoBERTa}_{\text{base}}$ | SimCSE | **-2.383** | 0.2413 |
| | + TCC | -2.317 | 0.2196 |
| | + TCS | -2.196 | **0.2087** |
| $\text{RoBERTa}_{\text{large}}$ | SimCSE | -2.868 | 0.2823 |
| | + TCC | -2.817 | **0.2645** |
| | + TCS | **-2.903** | 0.2880 |

Table 7: Uniformity and alignment results. Both losses are better as they become smaller.

| PLMs | Method | Avg.Score |
|---|---|---|
| $\text{BERT}_{\text{base}}$ | SimCSE | 75.83 ± 0.71 |
| | + TCC | **77.42** ± 0.61 |
| | + TCS | 76.46 ± 1.41 |
| $\text{BERT}_{\text{large}}$ | SimCSE | 77.14 ± 1.45 |
| | + TCC | **78.52** ± 0.29 |
| | + TCS | 78.28 ± 0.46 |
| $\text{RoBERTa}_{\text{base}}$ | SimCSE | 76.77 ± 0.06 |
| | + TCC | **77.18** ± 0.78 |
| | + TCS | 77.06 ± 0.65 |
| $\text{RoBERTa}_{\text{large}}$ | SimCSE | 78.04 ± 0.64 |
| | + TCC | **78.47** ± 0.43 |
| | + TCS | 78.04 ± 0.44 |

Table 8: Averaged results of 3 different random seed experiments on the STS evaluation tasks.

## C Detailed Results of ArcConLoss Experiments

In this section, we report detailed results of the ArcConLoss experiments shown in Table 3 of the main paper. As shown in Table 9, the application of our temperature cool-down shows a performance improvement that is comparable to the baseline, without any additional pre-processing or loss function.

## D Robustness of Temperature Cool-down

Since there has been a reported issue of SimCSE's vulnerability to random seeds, we perform additional experiments of temperature cool-down with 3 different random seeds. As shown in Table 8, temperature cool-down improves the performance of SimCSE performance with better robustness.

| PLMs | Method | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| $\text{BERT}_{\text{base}}$ | ArcCSE w/o Triplet loss | 71.76 | 82.77 | **76.81** | **83.56** | 78.87 | 79.36 | 71.16 | 77.76 |
| | + TCC | **72.31** | 83.87 | 76.76 | 83.16 | **79.54** | 79.97 | 71.82 | **78.20** |
| | + TCS | 72.26 | 83.46 | 76.48 | 83.18 | 79.46 | **80.07** | 71.73 | 78.09 |
| | ArcCSE ♡ | 72.08 | **84.27** | 76.25 | 82.32 | **79.54** | 79.92 | **72.39** | 78.11 |
| $\text{BERT}_{\text{large}}$ | ArcCSE w/o Triplet loss | 73.38 | 84.94 | 76.74 | 84.28 | **80.19** | 80.02 | 72.96 | 78.93 |
| | + TCC | **73.92** | 84.53 | 77.24 | 84.72 | 79.66 | 79.96 | 73.76 | 79.11 |
| | + TCS | 72.22 | 85.17 | 77.60 | 84.71 | 79.76 | **80.50** | **74.66** | 79.23 |
| | ArcCSE ♡ | 73.17 | **86.19** | **77.90** | **84.97** | 79.43 | 80.45 | 73.50 | **79.37** |

Table 9: Performance of different unsupervised contrastive learning methods on the STS tasks (Spearman's correlation). Each bold number indicates the best performance within the PLMs. ♡: Results from Gao et al., 2021.

| PLMs | Method | MR | CR | SUBJ | MPQA | SST | TREC | MPRC | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| $\text{BERT}_{\text{base}}$ | SimCSE | **81.37** | **86.49** | **94.46** | 88.66 | <u>84.95</u> | **87.60** | <u>74.32</u> | **85.41** |
| | + TCC | <u>80.77</u> | <u>85.57</u> | 94.24 | **88.86** | **85.28** | <u>87.47</u> | **74.49** | <u>85.21</u> |
| | + TCS | 80.30 | 85.25 | <u>94.31</u> | <u>88.85</u> | 84.35 | 85.80 | 74.14 | 84.71 |
| $\text{BERT}_{\text{large}}$ | SimCSE | 84.30 | 87.98 | <u>94.86</u> | 88.78 | 89.51 | <u>93.00</u> | 74.61 | 87.58 |
| | + TCC | **84.68** | **88.40** | 94.76 | **89.58** | <u>90.39</u> | **93.40** | <u>75.30</u> | <u>88.07</u> |
| | + TCS | <u>84.47</u> | <u>88.37</u> | **95.11** | <u>89.57</u> | **90.72** | 91.80 | **76.58** | **88.09** |
| $\text{RoBERTa}_{\text{base}}$ | SimCSE | <u>81.75</u> | <u>86.97</u> | **93.43** | 87.28 | 86.99 | 84.40 | 75.01 | 85.12 |
| | + TCC | **82.09** | **87.42** | <u>93.15</u> | **88.07** | **87.10** | <u>85.20</u> | **75.42** | **85.49** |
| | + TCS | 81.20 | 86.94 | 92.96 | <u>87.36</u> | <u>87.04</u> | **85.40** | <u>75.19</u> | <u>85.16</u> |
| $\text{RoBERTa}_{\text{large}}$ | SimCSE | **83.17** | **88.40** | **94.08** | **88.57** | **87.53** | **91.20** | <u>72.23</u> | **86.45** |
| | + TCC | 81.85 | 87.47 | <u>93.74</u> | <u>88.54</u> | 86.66 | 90.80 | **73.51** | <u>86.08</u> |
| | + TCS | <u>82.19</u> | <u>88.11</u> | 93.42 | 88.18 | <u>86.99</u> | **91.20** | 71.42 | 85.93 |

Table 10: Performance of different unsupervised contrastive learning methods on the transfer tasks. Each bold number and underlined number indicates the best and the second best performance within the PLMs, respectively.

## E Transfer Tasks

We also perform evaluation on 7 transfer tasks using the SentEval toolkit. As we can see in Table 10, the results of the transfer tasks show slightly lower or comparable performance to the baseline. This is consistent with the intuition that transfer tasks rarely target the sentence representation tasks (Gao et al., 2021).

## F Toward the Possibility of Variant for Temperature Cool-down

In addition to the two methods (TCC and TCS) introduced in the main paper, there will be many different ways to design variants of temperature cool-down, similar to learning rate scheduling. For instance, one of the most commonly used learning rate schedules is *linear warm-up* (Goyal et al., 2017). Following this straightforward mechanism, we introduce a simple approach of *linear temperature cool-down* (called TCL) as below:

$$\tau_{TCL,t} = \begin{cases} \tau_i - \frac{\tau_i - \tau}{r_s \cdot s} \cdot t, & \text{if } t \in [1, r_s \cdot s) \\ \tau. & \text{otherwise} \end{cases} \quad (4)$$

We believe that there may be several other candidates that show effective performance.

## G Additional Results of Uniformity and Alignment

In addition to the results of Section 3.3, we plot the uniformity and alignment of 3 other PLMs during training. As shown in Figure 4, our temperature cool-down methods improve the quality of the representation spaces in terms of both metrics. We also report the uniformity and alignment of the model's best checkpoints in Table 7.

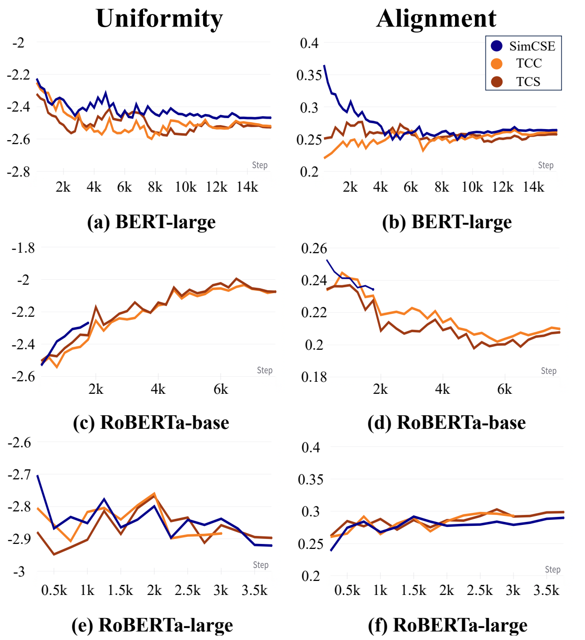


Figure 4: Uniformity and alignment on BERT-large, RoBERTa-base, and RoBERTa-large using temperature cool-down.