# TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning

**Anonymous ACL submission**


## Abstract

Masked language models (MLMs) such as BERT have revolutionized the field of Natural Language Understanding in the past few years. However, existing pre-trained MLMs often output an anisotropic distribution of token representations that occupies a narrow subset of the entire representation space. Such token representations are not ideal, especially for tasks that demand discriminative semantic meanings of distinct tokens. In this work, we propose **TaCL** (**T**oken-**a**ware **C**ontrastive **L**earning), a novel continual pre-training approach that encourages BERT to learn an isotropic and discriminative distribution of token representations. TaCL is fully unsupervised and requires no additional data. We extensively test our approach on a wide range of English and Chinese benchmarks. The results show that TaCL brings consistent and notable improvements over the original BERT model. Furthermore, we conduct detailed analysis to reveal the merits and inner-workings of our approach.[1]


## 1 Introduction

Since the rising of BERT (Devlin et al., 2019), masked language models (MLMs) have become the de facto backbone for almost all natural language understanding (NLU) tasks. Despite their clear success, many existing language models pre-trained with MLM objective suffer from the *anisotropic problem* (Ethayarajh, 2019). That is, their token representations reside in a narrow subset of the representation space, therefore being less discriminative and less powerful in capturing the semantic differences of distinct tokens.

Recently, great advancement has been made in continually training MLMs with unsupervised sentence-level contrastive learning, aiming at creating more discriminative sentence-level representations (Giorgi et al., 2021; Carlsson et al., 2021;

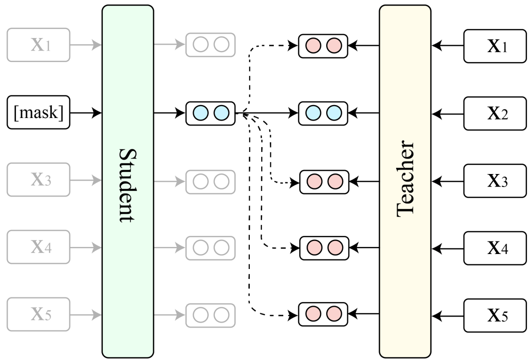


Figure 1: An overview of TaCL. The student learns to make the representation of a masked token closer to its "reference" representation produced by the teacher (solid arrow) and away from the representations of other tokens in the same sequence (dashed arrows).

Yan et al., 2021; Kim et al., 2021; Liu et al., 2021b; Gao et al., 2021). However, such representations are only evaluated as sentence embeddings and there is no evidence that they will benefit other well-established NLU tasks. We show that these approaches hardly bring any benefit to challenging tasks like SQuAD (Rajpurkar et al., 2016, 2018).

In this paper, we argue that the key of obtaining more discriminative and transferrable representations lies in learning contrastive and isotropic token-level representations. To this end, we propose TaCL (**T**oken-**a**ware **C**ontrastive **L**earning), a new continual pre-training approach that encourages BERT to learn discriminative token representations. Specifically, our approach involves two models (a student and a teacher) that are both initialized from the same pre-trained BERT. During the learning stage, we freeze the parameters of the teacher and continually optimize the student model with (1) the original BERT pre-training objectives (masked language modelling and next sentence prediction) and (2) a newly proposed TaCL objective. The TaCL loss is obtained by contrasting the student representations of masked tokens against the "reference" representations produced by the teacher

[1]The code and models will be released upon publication.

without masking the input tokens. In Figure 1, we provide an overview of our approach.

We extensively test our approach on a wide range of English and Chinese benchmarks and illustrate that TaCL brings notable performance improvements on most evaluated datasets (§3.1.1). These results validate that more discriminative and isotropic token representations lead to better model performances. Additionally, we highlight the benefits of using our token-level method compared to current state-of-the-art sentence-level contrastive learning techniques on NLU tasks (§3.2.1). We further analyze the inner workings of TaCL and its impact on the token representation space (§3.2.2).

Our work, to the best of our knowledge, is the first effort on applying contrastive learning to improve token representations of Transformer models. We hope the findings of this work could facilitate further development of methods on the intersection of contrastive learning and representation learning at a more fine-grained granularities.

## 2 Token-aware Contrastive Learning

Our approach contains two models, i.e., a student $S$ and a teacher $T$, both of which are initialized from the same pre-trained BERT. During learning, we freeze $T$ and only optimize the parameters of $S$. Given an input sequence $x = [x_1, ..., x_n]$, we randomly mask $x$ with the same procedure as in Devlin et al. (2019) and feed the masked sequence $\tilde{x}$ into the student model to produce the contextual representation $\tilde{h} = [\tilde{h}_1, ..., \tilde{h}_n]$. Meanwhile, the teacher model takes the original sequence $x$ as input and produces the representation $h = [h_1, ..., h_n]$ (see Figure 1). The proposed token-aware contrastive learning objective $\mathcal{L}_{\text{TaCL}}$ is then defined as

$$-\sum_{i=1}^{n}\mathbb{1}(\tilde{x}_i)\log\frac{\exp(\text{sim}(\tilde{h}_i, h_i)/\tau)}{\sum_{j=1}^{n}\exp(\text{sim}(\tilde{h}_i, h_j)/\tau)}, \quad (1)$$

where $\mathbb{1}(\tilde{x}_i) = 1$ if $\tilde{x}_i$ is a masked token, otherwise $\mathbb{1}(\tilde{x}_i) = 0$. $\tau$ is a temperature hyper-parameter and $\text{sim}(\cdot, \cdot)$ computes the cosine similarity. Intuitively, the student learns to make the representation of a masked token closer to its "reference" representation produced by the teacher and away from other tokens in the same sequence. As a result, the token representations learnt by the student are more discriminative with respect to distinct tokens, therefore better following an isotropic distribution. Similar to Devlin et al. (2019), we also adopt the masked language modelling $\mathcal{L}_{\text{MLM}}$ and next sentence prediction $\mathcal{L}_{\text{NSP}}$ objectives. The overall learning objective $\mathcal{L}$ of the student model during the continual pre-training stage is defined as

$$\mathcal{L} = \mathcal{L}_{\text{TaCL}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{NSP}}. \quad (2)$$

Note that the learning of the student is fully unsupervised and can be realized using the original pre-training corpus. After the learning is completed, we fine-tune the student model on downstream tasks.

## 3 Experiment

We test our approach on a wide range of benchmarks in two languages. For English benchmarks, we evaluate the $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ models. For Chinese benchmarks, we test the $\text{BERT}_{\text{base}}$ model.[2] After initializing the student and teacher, we continually pre-train the student on the same Wikipedia corpus as in Devlin et al. (2019) for 150k steps. The training samples are truncated with a maximum length of 256 and the batch size is set as 256. The temperature $\tau$ in Eq. (1) is set as 0.01. Same as Devlin et al. (2019), we optimize the model with Adam optimizer (Kingma and Ba, 2015) with weighted decay, and an initial learning of 1e-4 (with warm-up ratio of 10%).

### 3.1 Evaluation Benchmarks

For English benchmarks, we use the GLUE dataset (Wang et al., 2019) which contains a variety of sentence-level classification tasks covering textual entailment (RTE and MNLI), question-answer entailment (QNLI), paraphrase (MRPC), question paraphrase (QQP), textual similarity (STS-B), sentiment (SST-2), and linguistic acceptability (CoLA). Our evaluation metrics are Spearman correlation for STS-B, Matthews correlation for CoLA, and accuracy for the other tasks; the macro average score is also reported. Additionally, we conduct experiments on SQuAD 1.1 (Rajpurkar et al., 2016) and 2.0 (Rajpurkar et al., 2018) datasets that evaluate the model's performance on the token-level answer-extraction task. The dev set results of Exact-Match (EM) and F1 scores are reported.

For Chinese benchmarks, we evaluate our model on two token-level labelling tasks, including name entity recognition (NER) and Chinese word segmentation (CWS). For NER, we use the Ontonotes (Weischedel et al., 2011), MSRA (Levow, 2006), Resume (Zhang and Yang, 2018), and Weibo (He and Sun, 2017) datasets. For CWS, we use the

[2] All models are officially released by Devlin et al. (2019).

| | Model | GLUE | | | | | | | | | SQuAD 1.1 | | SQuAD 2.0 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CoLA | SST-2 | MPRC | STS-B | QQP | MNLI | QNLI | RTE | Ave. | EM | F1 | EM | F1 |
| English Benchmark | *Base size models* | | | | | | | | | | | | | |
| | $\text{BERT}_{\text{base}}$ ‖ | 52.1 | 93.5 | 88.9 | 85.8 | 71.2 | 84.6/83.4 | 90.5 | 66.4 | 79.6 | 80.8 | 88.5 | - | - |
| | $\text{BERT}_{\text{base}}$ ‡ | 52.2 | 92.4 | 89.0 | 86.4 | 73.2 | **84.6/84.5** | 90.3 | 63.2 | 79.8 | 80.9 | 88.4 | 73.4 | 76.8 |
| | +*MT*‡ | 51.9 | **92.5** | 89.3 | 87.1 | 75.8 | 84.2/84.0 | 90.6 | **64.1** | 80.0 | 81.0 | 88.5 | 73.2 | 76.3 |
| | $\text{TaCL}_{\text{base}}$ | **52.4** | 92.3 | **90.8** | **89.0** | **80.7** | 84.4/84.3 | **91.1** | 62.8 | **81.2** | **81.6** | **89.0** | **74.4** | **77.5** |
| | *Large size models* | | | | | | | | | | | | | |
| | $\text{BERT}_{\text{large}}$ ‖ | 60.5 | 94.9 | 89.3 | 86.5 | 72.1 | 86.7/85.9 | 92.7 | 70.1 | 82.1 | 84.1 | 90.9 | 78.7 | 81.9 |
| | $\text{BERT}_{\text{large}}$ ‡ | 61.6 | 93.6 | 90.2 | 89.0 | 81.8 | 86.4/86.1 | **92.6** | 67.2 | 83.6 | 84.0 | 90.8 | 77.9 | 81.0 |
| | +*MT*‡ | **62.0** | 93.8 | 90.5 | 89.1 | **82.5** | 86.3/**86.3** | 92.2 | 66.5 | 83.7 | 83.9 | 90.9 | 77.8 | 80.7 |
| | $\text{TaCL}_{\text{large}}$ | 61.1 | **94.1** | **92.0** | **89.7** | **82.5** | **86.5**/85.9 | 92.4 | **70.5** | **84.7** | **84.2** | **91.1** | **78.7** | **81.9** |

| | Model | Ontonotes | | MSRA | | Resume | | Weibo | | PKU | CityU | AS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Dev | Test | Dev | Test | Dev | Test | Dev | Test | Test | Test | Test |
| Chinese Benchmark | ♠ and ♢ published in Li et al. (2020) and Meng et al. (2019) | | | | | | | | | | | |
| | $\text{BERT}_{\text{base}}$ | - | 80.14♠ | - | 94.95♠ | - | 95.53♠ | - | 68.20♠ | 96.50♢ | 97.60♢ | 96.50♢ |
| | $\text{BERT}_{\text{base}}$‡ | 78.29 | 80.23 | 94.13 | 94.97 | 95.37 | 95.70 | 70.63 | 67.98 | 96.51 | 97.83 | 96.58 |
| | +*MT*‡ | 78.42 | 80.36 | 94.20 | 95.01 | 95.29 | 95.62 | 70.81 | 68.02 | 96.53 | 97.79 | 96.54 |
| | $\text{TaCL}_{\text{base}}$ | **79.73** | **82.42** | **94.58** | **95.44** | **96.23** | **96.45** | **72.32** | **69.54** | **96.75** | **98.18** | **96.75** |

Table 1: Benchmark Results. ‖: published in Devlin et al. (2019); and ‡: models from our implementations.

PKU, CityU, and AS datasets from SIGHAN 2005 (Emerson, 2005) for evaluation. The standard F1 score is used for evaluation.

**Baselines:** We compare against two baselines: (1) the original BERT used to initialize the student and teacher; (2) BERT+*MT* (BERT with more training) which is acquired by continually pre-training the original BERT on Wikipedia for 150k steps[3] using the original BERT pre-training objectives.

### 3.1.1 Benchmark Results

Table 1 reports the results on English and Chinese benchmarks.[4] We observe that, on most sequence-level classification tasks in GLUE, TaCL outperforms BERT and BERT+*MT*. Additionally, on all token-level benchmarks (SQuAD, NER, and CWS), TaCL consistently and notably surpasses other baselines. These results indicate that the learning of an isotropic token representation space is beneficial for the model's performance, especially on the token-centric tasks.

## 3.2 Analysis

In this section, we present further comparisons and in-depth analysis of the proposed approach.

### 3.2.1 Sentence-Level vs. Token-Level CL

We compare TaCL against existing sentence-level contrastive learning methods, including DeCLUTR (Giorgi et al., 2021), SimCSE (Gao et al., 2021), and MirrorBERT (Liu et al., 2021b). We also include two ablated models to study the effect of different combinations of pre-training objectives. Specifically, the ablated model-1 is initialized with BERT and trained with the original BERT objectives ($\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{NSP}}$) **plus** the sentence-level contrastive objective as proposed in Liu et al. (2021b). The ablated model-2 is initialized with BERT and trained **only** with the proposed token-aware contrastive objective of Eq. (1). Note that all compared models have the same size as the $\text{BERT}_{\text{base}}$ model.

| Model | $\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{NSP}}$ | CL | SQuAD 1.1 | SQuAD 2.0 |
|---|---|---|---|---|
| BERT | ✓ | × | 80.8/88.5 | 73.4/76.8 |
| *Sentence-Level Contrastive Methods* | | | | |
| DeCLUTR | × | Sen. | 79.9/87.6 | 72.1/75.4 |
| SimCSE | × | Sen. | 80.2/88.0 | 72.5/75.7 |
| MirrorBERT | × | Sen. | 80.3/88.1 | 72.7/75.9 |
| *Ablated Models* | | | | |
| model-1 | ✓ | Sen. | 80.5/88.3 | 73.1/76.5 |
| model-2 | × | Tok. | 81.3/88.7 | 73.8/77.1 |
| TaCL | ✓ | Tok. | **81.6/89.0** | **74.4/77.5** |

Table 2: Comparison of various sentence- and token-level contrastive learning methods. "Sen." or "Tok." denotes training with sentence- or token-level contrastive objectives. Scores of (EM/F1) are reported.

Table 2 shows the performance of different models on SQuAD. We observe decreased performance of existing sentence-level contrastive methods compared with the original BERT. This could be attributed to the fact that such methods only focus on learning sentence-level representations while ignoring the learning of individual tokens. This be-

[3]The number of steps is set the same as our TaCL training.

[4]For all tasks, the average results over five runs are reported.

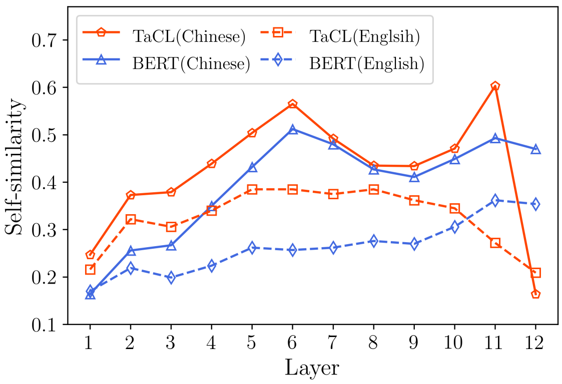


Figure 2: Layer-wise representation self-similarity.

haviour is undesired for tasks like SQuAD that demands informative token representations. Nonetheless, the ablated model-1 shows that the original BERT pre-training objective ($\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{NSP}}$) remedies, to some extent, the performance degeneration caused by the sentence-level contrastive methods. On the other hand, the ablated model-2 demonstrates that our token-aware contrastive objective helps the model to achieve improved results by learning better token representations.

### 3.2.2 Token Representation Self-similarity

To analyze the token representations learnt by TaCL and BERT, we follow Ethayarajh (2019) and define the averaged self-similarity of the token representations within one sequence $x = [x_1, ..., x_n]$ as,

$$s(x) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \text{cosine}(h_i, h_j), \quad (3)$$

where $h_i$ and $h_j$ are the token representations of $x_i$ and $x_j$ produced by the model. Intuitively, a lower $s(x)$ indicates that the representations of tokens within the sequence $x$ are less similar to each other, therefore being more discriminative.

We sample 50k sentences from both Chinese and English Wikipedia and compute the self-similarity of representations over different layers. Figure 2 plots the results of $\text{TaCL}_{\text{base}}$ and $\text{BERT}_{\text{base}}$ averaged over all sentences. We see that, in the intermediate layers, the self-similarity of TaCL is higher than BERT's. In contrast, at the top layer (layer 12), TaCL's self-similarity becomes notably lower than BERT's, demonstrating that the final output token representations of TaCL are more discriminative.

**Qualitative Analysis.** We sample one sentence from Wikipedia and visualize the self-similarity matrix $M$ (where $M_{i,j} = \text{cosine}(h_i, h_j)$) produced by $\text{BERT}_{\text{base}}$ and $\text{TaCL}_{\text{base}}$. The results are shown

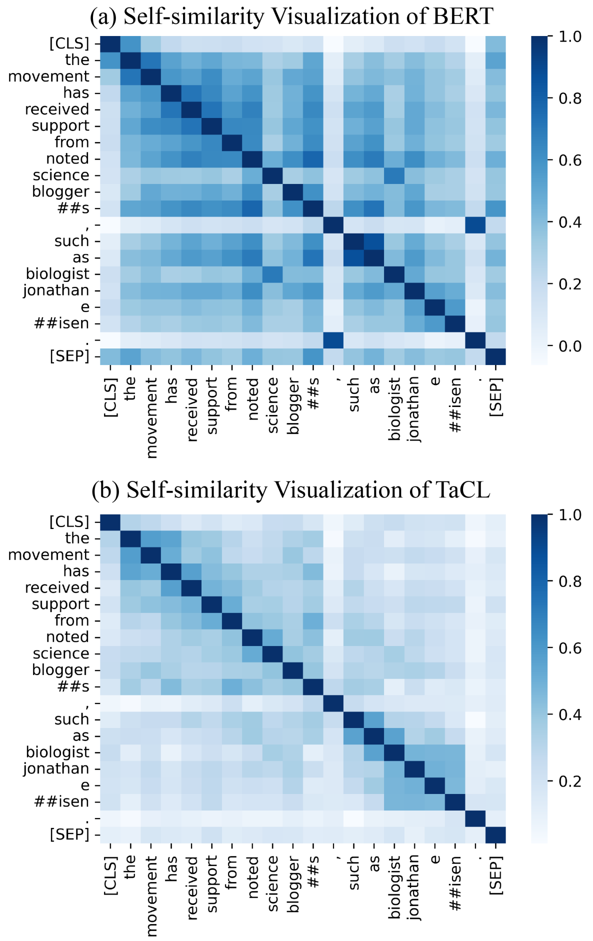


Figure 3: Self-similarity Matrix Visualization: (a) BERT and (b) TaCL. (best viewed in color)

in Figure 3, where a darker color denotes a higher self-similarity score.[5] We see that, as compared with BERT (Fig. 3(a)), the self-similarities of TaCL (Fig. 3(b)) are much lower in the off-diagonal entries. This further highlights that the individual token representations of TaCL are more discriminative, which in return leads to improved model performances as demonstrated (§3.1.1, §3.2.1).

## 4 Conclusion

In this work, we proposed TaCL, a novel approach that applies token-aware contrastive learning for the continual pre-training of BERT. Extensive experiments are conducted on a wide range of English and Chinese benchmarks. The results show that our approach leads to notable performance improvement across all evaluated benchmarks. We then delve into the inner-working of TaCL and demonstrate that our performance gain comes from a more discriminative distribution of token representations.

[5]The entries $M_{i,i}$ in the diagonal have a 1.0 self-similarity by definition, as $\text{cosine}(h_i, h_i) = 1.0$.

## References


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

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.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546. IEEE Computer Society.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. 2021. Container: Few-shot named entity recognition via contrastive learning. *arXiv preprint arXiv:2109.07589*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, SIGHAN@IJCNLP 2005, Jeju Island, Korea, 14-15, 2005*. ACL.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 55–65. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Hangfeng He and Xu Sun. 2017. F-score driven max margin neural network for named entity recognition in chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 713–718. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tian Lan, Deng Cai, Yan Wang, Yixuan Su, Xian-Ling Mao, and Heyan Huang. 2021. Exploring dense retrieval for dialogue response selection. *CoRR*, abs/2110.06612.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia,* *July 22-23, 2006*, pages 108–117. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6836–6842. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

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

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021c. MirrorWiC: On eliciting word-in-context representations from pretrained language models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.

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

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2742–2753.

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

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual bert post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

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

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey

Levine. 2018. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1134–1141. IEEE.

Yixuan Su, Deng Cai, Yan Wang, David Vandyke, Simon Baker, Piji Li, and Nigel Collier. 2021a. Non-autoregressive text generation with pre-trained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 234–243. Association for Computational Linguistics.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2021b. Dialogue response selection with hierarchical curriculum learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1740–1751. Association for Computational Linguistics.

Yixuan Su, Yan Wang, Deng Cai, Simon Baker, Anna Korhonen, and Nigel Collier. 2021c. PROTOTYPE-TO-STYLE: dialogue generation with style-aware editing on retrieval memory. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2152–2161.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021. LexFit: Lexical fine-tuning of pretrained language models. 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 5269–5283, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. *ArXiv*, abs/2109.06304.

Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2794–2802. IEEE Computer Society.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, and Robert Belvin. 2011. Ontonotes release 4.0. ldc2011t03. In *Philadelphia, Penn.: Linguistic Data Consortium* .

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

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1554–1564. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *CoRR*, abs/2109.02492.

# A Statistics of Evaluated Benchmarks

## A.1 English Benchmarks

| Dataset | Train | Test | Evaluation Metric |
|---|---|---|---|
| CoLA | 8.5k | 1k | Matthews correlation |
| SST-2 | 67k | 1.8k | accuracy |
| MRPC | 3.7k | 1.7K | accuracy |
| STS-B | 7k | 1.4k | Spearman correlation |
| QQP | 364k | 391k | accuracy |
| MNLI | 393k | 20k | matched/mismatched accuracy |
| QNLI | 105k | 5.4k | accuracy |
| RTE | 2.5k | 3k | accuracy |

Table 3: GLUE Statistics

| Dataset | Train | Dev | Evaluation Metric |
|---|---|---|---|
| 1.1 | 87.6k | 10.6k | Exact-Match/F1 |
| 2.0 | 130.3k | 11.9k | Exact-Match/F1 |

Table 4: SQuAD Statistics

## A.2 Chinese Benchmarks

| Dataset | Train | Dev | Test | Evaluation Metric |
|---|---|---|---|---|
| Ontonotes | 15.7k | 4.3k | 4.3k | F1 |
| MSRA | 37.0k | 9.3k | 4.4k | F1 |
| Resume | 3.8k | 0.5k | 0.5k | F1 |
| Weibo | 1.4k | 0.3k | 0.3k | F1 |

Table 5: NER Dataset Statistics

| Dataset | Train | Test | Evaluation Metric |
|---|---|---|---|
| PKU | 19.1k | 1.9k | F1 |
| CityU | 53.0k | 1.5k | F1 |
| AS | 708.9k | 14.4k | F1 |

Table 6: CWS Dataset Statistics

# B Related Work

**Pre-trained Language Models.** Since the introduction of BERT (Devlin et al., 2019), the NLP research community has witnessed remarkable progress in the field of language model pre-training on a large amount of free text. Such advancements have led to significant progresses in a wide range of natural language understanding (NLU) tasks (Liu et al., 2019; Yang et al., 2019; Clark et al., 2020; Lan et al., 2021) and text generation tasks (Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Su et al., 2021a,c; Zhong et al., 2021)

**Contrastive Learning.** Generally, contrastive learning methods distinguish observed data points from fictitious negative samples. They have been widely applied to various computer vision areas, including image (Chopra et al., 2005; Oord et al., 2018) and video (Wang and Gupta, 2015; Sermanet et al., 2018). Recently, Chen et al. (2020) proposed a simple framework for contrastive learning of visual representations (SimCLR) based on multi-class N-pair loss. Radford et al. (2021); Jia et al. (2021) applied the contrastive learning approach for language-image pretraining. Xu et al. (2021); Yang et al. (2021) proposed a contrastive pre-training approach for video-text alignment.

In the field of NLP, numerous approaches have been proposed to learn better sentence-level (Reimers and Gurevych, 2019; Wu et al., 2020; Meng et al., 2021; Liu et al., 2021b; Gao et al., 2021; Su et al., 2021b) and lexical-level (Liu et al., 2021a; Vulić et al., 2021; Liu et al., 2021c; Wang et al., 2021) representations using contrastive learning. Different from our work, none of these studies specifically investigates how to utilize contrastive learning for improving general-purpose token-level representations. Beyond representation learning, contrastive learning has also been applied to NLP applications such as NER (Das et al., 2021) and summarisation (Liu and Liu, 2021).

**Continual Pre-training.** Many researchers (Xu et al., 2019; Gururangan et al., 2020; Pan et al., 2021) have investigated how to continually pre-train the model to alleviate the task- and domain-discrepancy between the pre-trained models and the specific target task. In contrast, our proposed approach studies how to apply continual pre-training to directly improve the quality of model representations which is transferable and beneficial to a wide range of benchmark tasks.

# C More Self-similarity Visualizations

In Figure 4, 5, and 6, we provide three more comparisons between the self-similarity matrix produced by TaCL and BERT (the example sentences are randomly sampled from Wikipedia).[6] From the figures, we can draw the same conclusion as in section §3.2.2, that the token representations of BERT follow an anisotropic distribution and are less discriminative. On the other hand, the token representations of TaCL better follow an isotropic distribution, therefore different tokens become more distinguishable with respect to each other.

[6] All results are generated by models with base size.

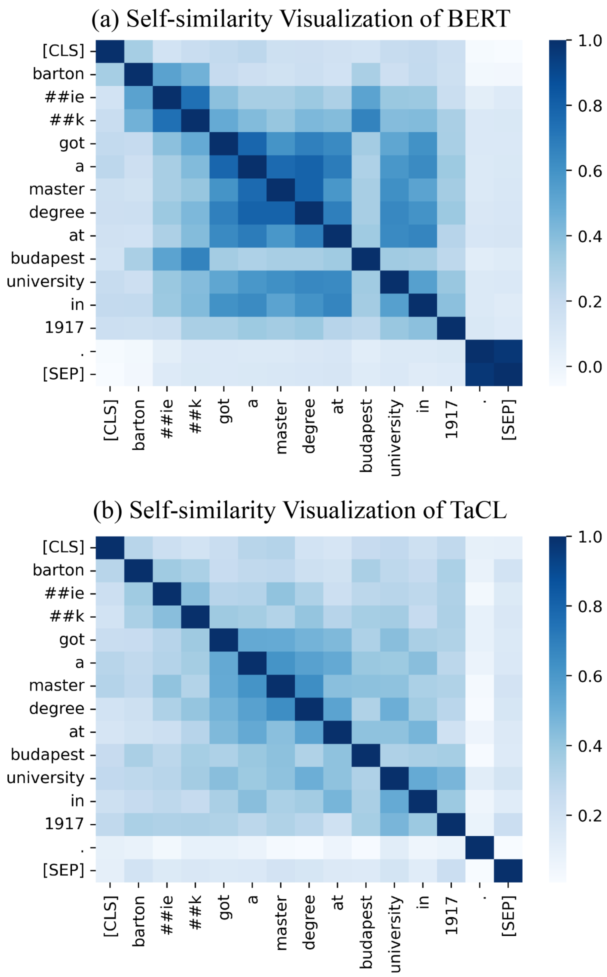


Figure 4: **Example 2**: self-similarity matrix visualization of (a) BERT and (b) TaCL. (best viewed in color)

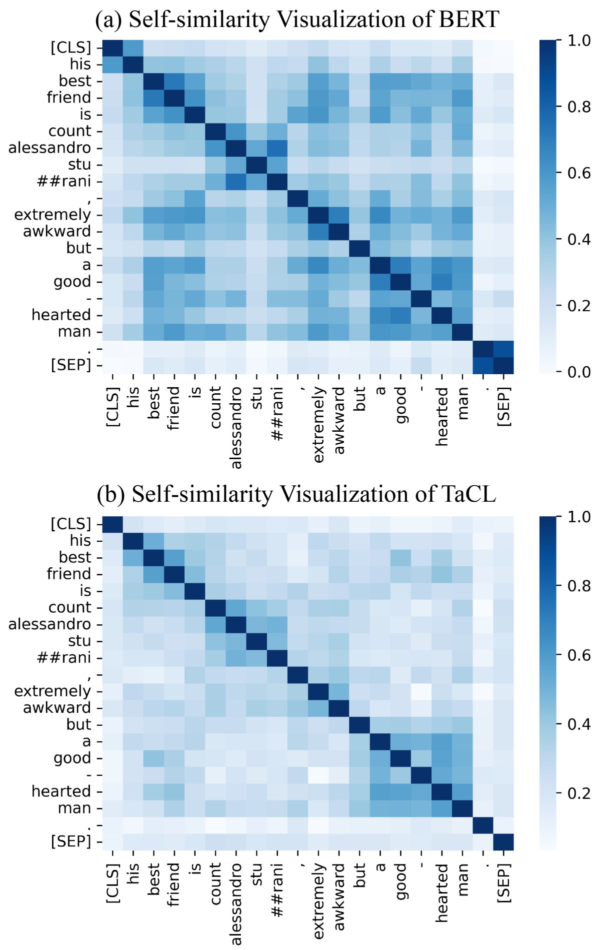


Figure 5: **Example 3**: self-similarity matrix visualization of (a) BERT and (b) TaCL. (best viewed in color)

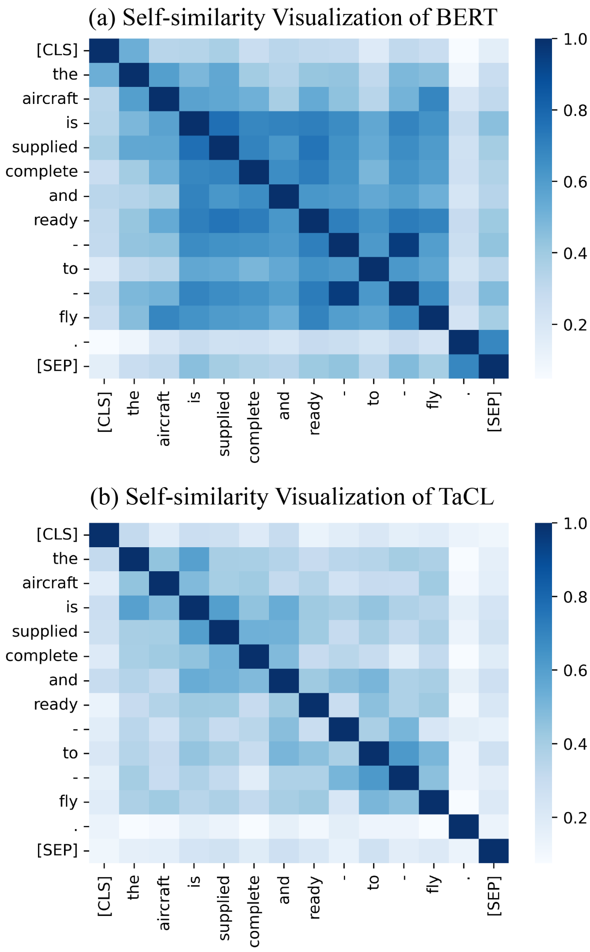


Figure 6: **Example 4**: self-similarity matrix visualization of (a) BERT and (b) TaCL. (best viewed in color)