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 (Token-aware Contrastive Learning), 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; 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 (Token-aware Contrastive Learning), 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 granularity.

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 $L_{TaCL}$ is then defined as

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

(1)

where $\mathbb{I}(\tilde{x}_i) = 1$ if $\tilde{x}_i$ is a masked token, otherwise $\mathbb{I}(\tilde{x}_i) = 0$. $\tau$ 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 $L_{MLM}$ and next sentence prediction $L_{NSP}$ objectives. The overall learning objective $L$ of the student model during the continual pre-training stage is defined as

$$L = L_{TaCL} + L_{MLM} + L_{NSP}.$$  

(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 BERT$_{base}$ and BERT$_{large}$ models. For Chinese benchmarks, we test the BERT$_{base}$ model. 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 weight decay, and an initial learning rate 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).
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 ($L_{MLM}$ and $L_{NSP}$) 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 BERT$_{base}$ model.

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-

Table 1: Benchmark Results. ||: published in Devlin et al. (2019); and ‡: models from our implementations.
Figure 2: Layer-wise representation self-similarity.

 behaviour is undesired for tasks like SQuAD that demands informative token representations. Nonetheless, the ablated model-1 shows that the original BERT pre-training objective ($L_{MLM}$ and $L_{NSP}$) 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,

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

(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 TaCL$_{base}$ and BERT$_{base}$ 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 $M_{i,j} = \cos(h_i, h_j)$) produced by BERT$_{base}$ and TaCL$_{base}$. The results are shown 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 ($\S$3.1.1, $\S$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 $\cos(h_i, h_i) = 1.0$. 
References


Yu Meng, Chenyang 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.


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


A Statistics of Evaluated Benchmarks

A.1 English Benchmarks

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

Table 3: GLUE Statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Evaluation Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>87.6k</td>
<td>10.6k</td>
<td>Exact-Match/F1</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>130.3k</td>
<td>11.9k</td>
<td>Exact-Match/F1</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: SQuAD Statistics

A.2 Chinese Benchmarks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Evaluation Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontonotes</td>
<td>15.7k</td>
<td>4.3k</td>
<td>4.3k</td>
<td>F1</td>
</tr>
<tr>
<td>MSRA</td>
<td>37.0k</td>
<td>9.3k</td>
<td>4.4k</td>
<td>F1</td>
</tr>
<tr>
<td>Resume</td>
<td>3.8k</td>
<td>0.5k</td>
<td>0.5k</td>
<td>F1</td>
</tr>
<tr>
<td>Weibo</td>
<td>1.4k</td>
<td>0.3k</td>
<td>0.3k</td>
<td>F1</td>
</tr>
</tbody>
</table>

Table 5: NER 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). 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)