SimCTC: A Simple Contrast Learning Method of Text Clustering (Student Abstract)

Chen Li,[1] **Xiaoguang Yu,**[2] **Shuangyong Song,**[2] **Jia Wang,**[2] **Bo Zou,**[2] **Xiaodong He**[2]
[1] Sichuan University, No. 24, South Section, First Ring Road, Chengdu, China
[2] JD AI Research, Courtyard 18, Kechuang 11th Street, Daxing District, Beijing, China
lichen19970228@163.com, {cdyuxiaoguang, songshuangyong, cdwangjia5, cdzoubo, hexiaodong}@jd.com

Abstract

This paper presents SimCTC, a simple contrastive learning (CL) method that greatly advances the state-of-the-art text clustering models. In SimCTC, a pre-trained BERT model first maps the input sequence to the representation space, which is then followed by three different loss function heads: Clustering head, Instance-CL head and Cluster-CL head. Experimental results on multiple benchmark datasets demonstrate that SimCTC remarkably outperforms 6 competitive text clustering methods with 1%-6% improvement on Accuracy (ACC) and 1%-4% improvement on Normalized Mutual Information (NMI). Moreover, our results also show that the clustering performance can be further improved by setting an appropriate number of clusters in the cluster-level objective.

Introduction

As one of the most fundamental challenges in unsupervised learning, text clustering has been widely studied for decades and learning high-quality sentence representation is the key point of this task. On the other hand, contrastive learning recently has been proved to be an effective method to learn high-quality sentence representations, and has achieved the state-of-the-art performance in most natural language understanding tasks (Yan et al. 2021). In this paper we try to introduce the contrastive learning to text clustering task.

The main idea of contrastive learning is push the embeddings of two augmented views of the same sequence close to each other and further apart from the embedding of other sequence. Most of the algorithms only focused on instance-level contrastive learning. Actually, as (Li et al. 2021) has indicated that cluster-level contrastive learning can bring improvement in clustering performance of visual representation to some extent, and we assume that this cluster-level contrastive learning is still applicable in textual representation learning. To validate this hypothesis, in this work, we present SimCTC, a simple contrastive learning (CL) method that greatly advances the state-of-the-art text clustering models by simultaneously optimizing the clustering loss, instance-level and cluster-level contrastive loss. We demonstrate that our approach outperforms the state-of-the-art text clustering models on benchmark datasets. Additionally, we also show that the clustering performance can be further improved by setting an appropriate number of clusters in the cluster-level contrastive learning.

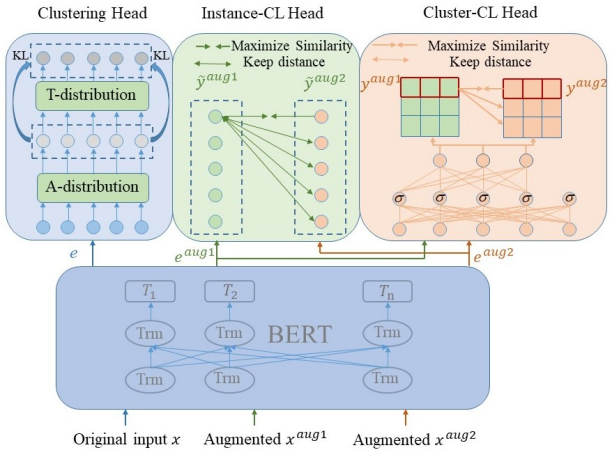

Figure 1: The general framework of our proposed SimCTC.

Method

As illustrated in Figure 1, there are four main components in SimCTC: 1) A BERT encoder that computes sentence representation for each input sequence; 2) A clustering head that tries to bring together sentence representations from the same semantic category; 3) An instance-CL head that applies contrastive learning in the instance level; 4) A cluster-CL head that applies contrastive learning in the cluster level.

For each original input x,we first choose Bertbase and Roberta from the nlpaug library to generate the augmented x^{aug1} and x^{aug2} respectively. Then, x, x^{aug1} and x^{aug2} will be encoded by a BERT-like language model M and produce the sentence representations e, e^{aug1} and e^{aug2} respectively. Finally, we train M by simultaneously optimizing the three loss functions, while the final loss is the sum of them.

Clustering Head: Following in (Zhang et al. 2021), we first apply the student's t-distribution to compute the assigning distribution q_{jk} (A-distribution), which denotes assign-

ing each sentence representation e_j to the k^{th} cluster,

$$q_{jk} = \frac{(1 + \|e_j - \mu_k\|_2^2 \big/ \alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'}^{K} (1 + \|e_j - \mu_{k'}\|_2^2 \big/ \alpha)^{-\frac{\alpha+1}{2}}} \quad (1)$$

where μ_k, α denotes the centroid of the k^{th} cluster and the degree of freedom respectively. $k \in \{1, ..., K\}$ and K denotes the number of clusters. Then, we use a linear layer as the clustering head to approximate the centroids of each cluster and iteratively update the centroids of clusters by leveraging an target dis-tribution p_{jk} (T-distribution),

$$p_{jk} = \frac{q_{jk}^2 \big/ f_k}{\sum_{k'} q_{jk}^2 \big/ f_{k'}} \quad (2)$$

where $f_k = \sum_{j=1}^{M} q_{jk}, k = 1, ..., K$ can be regarded as the soft cluster frequencies and M denotes the batch-size. Finally, we choose the KL divergence to compute the clustering objective, the formula is as follows,

$$L_{Clustering-loss} = \sum_{j=1}^{M} \sum_{k=1}^{K} p_{jk} \log \frac{p_{jk}}{q_{jk}} \quad (3)$$

Instance-CL Head: The Instance-CL loss is defined on the augmented e^{aug1} and e^{aug2}. For a representation e_i^{aug1}, there are $2M - 1$ pairs in total, among which we choose its corresponding augmented representation e_i^{aug2} to form a positive pair $\left\{e_i^{aug1}, e_i^{aug2}\right\}$ and leave other $2M - 2$ pairs to be negative. Let $\tilde{y}_i^{aug1}$ and $\tilde{y}_i^{aug2}$ be the corresponding outputs of the Instance-CL head $g_I()$, i.e., $\tilde{y}_i^{aug1} = g_I(e_i^{aug1}), i = 1, ..., M$. Given a representation e_i^{aug}, the loss is as follows (Li et al. 2021)

$$l_i^{aug1} = -\log \frac{\exp(s(\tilde{y}_i^{aug1}, \tilde{y}_i^{aug2}) \big/ \tau)}{\sum_{j=1}^{N} \exp(s(\tilde{y}_i^{aug1}, \tilde{y}_j^{aug}) \big/ \tau)} \quad (4)$$

where $s(.)$, τ denotes the cosine similarity and the instance-level temperature parameter, respectively. $aug \in \{aug1, aug2\}$ and $g_I()$ consist of a two layer multilayer perceptron (MLP). $\tilde{y}^{aug1} \in \Re^{M \times H}$, $\tilde{y}^{aug2} \in \Re^{M \times H}$ and H denotes the hidden size. The Instance-CL objective is computed over every augmented representation, the formula is as follow,

$$L_{Instance-CL-loss} = \frac{1}{2M} \sum_{i=1}^{M} (l_i^{aug1} + l_i^{aug2}) \quad (5)$$

Cluster-CL Head: Similar to the Instance-CL head, we first use a $g_c()$ to project the e^{aug} into a feature space to obtain the y^{aug1} ($y^{aug1} \in \Re^{C \times M}$) and the y^{aug2} ($y^{aug2} \in \Re^{C \times M}$), where C denotes the number of clusters. Then we can obtain the cluster-level contrastive loss $L_{Cluster-CL-loss}$ by applying contrastive loss on each distribution of clusters.

	SST		SOF		TT		GTS	
Model	Acc	Nmi	Acc	Nmi	Acc	Nmi	Acc	Nmi
BOW	24.3	9.30	18.5	14.0	49.7	73.6	57.5	81.9
TF-IDF	31.5	9.20	58.4	58.7	57.0	80.7	68.0	88.9
STCC	77.0	63.2	51.1	49.0	-	-	-	-
Self-Train	77.1	56.7	59.8	54.8	-	-	-	-
HAC-SD	82.7	63.8	64.8	59.5	**89.6**	83.5	85.8	88.0
SCCL	85.2	71.1	75.5	74.5	78.2	89.2	89.8	94.9
SimCTC	**85.4**	**71.9**	**78.3**	**75.4**	84.7	**93.4**	**90.9**	**96.1**

Table 1: Clustering performance on six datasets

Number of	SST		SOF		TT		GTS	
clusters	Acc	Nmi	Acc	Nmi	Acc	Nmi	Acc	Nmi
C/2	**86.1**	**72.2**	77.3	73.8	84.3	93.1	89.9	95.5
C*1	85.4	71.9	78.3	75.4	**84.7**	**93.4**	**90.9**	96.1
C*2	79.4	64.0	**78.5**	**76.6**	83.9	93.4	90.2	95.5
C*4	85.7	72.2	78.4	74.3	83.9	93.2	90.8	95.8

Table 2: The effect of the number of clusters C.

Experimental Results

We evaluate the proposed method on four benchmark datasets as follows in (Zhang et al. 2021): SST, SOF, TT and GTS. For comparison, we consider the following baselines as follows in (Zhang et al. 2021), including BOW, TF-IDF, STCC, Self-Train, HAC-SD, and SCCL. As can be seen from the Table 1, Our SimCTC model outperforms all baselines on most datasets. In addition, we also explored the impact of C (the number of clusters in the Cluster-CL head). From Table 2, we can see that an appropriate parameter C can further brings improvement on clustering performance.

Conclusion and Future Work

In this paper, we propose a simple method leveraging instance-level and cluster-level contrastive learning for text clustering. Experimental results show that our method can achieve the start-of-the-art results on multiple benchmark datasets. Our future work is mainly focused on: understanding the inner relationship between the number of clusters C in the Cluster-CL head and clustering performance and exploring an optimal C to improve the clustering performance.

References

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021. Contrastive clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 8547–8555.

Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; and Xu, W. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In *59th Annual Meeting of the Association for Computational Linguistics*, 5065–5075.

Zhang, D.; Nan, F.; Wei, X.; Li, S.; Zhu, H.; McKeown, K.; Nallapati, R.; Arnold, A.; and Xiang, B. 2021. Supporting Clustering with Contrastive Learning. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 5419–5430.