Label Anchored Contrastive Learning for Language Understanding

Zhenyu Zhang, Yuming Zhao, Meng Chen, Xiaodong He
JD AI, Beijing, China
{zhangzhenyu47,zhaoyuming3,chenmeng20,xiaodong.he}@jd.com

Abstract

Contrastive learning (CL) has achieved astonishing progress in computer vision, speech, and natural language processing fields recently with self-supervised learning. However, CL approach to the supervised setting is not fully explored, especially for the natural language understanding classification task. Intuitively, the class label itself has the intrinsic ability to perform hard positive/negative mining, which is crucial for CL. Motivated by this, we propose a novel label anchored contrastive learning approach (denoted as LaCon) for language understanding. Specifically, three contrastive objectives are devised, including a multi-head instance-centered contrastive loss (ICL), a label-centered contrastive loss (LCL), and a label embedding regularizer (LER). Our approach does not require any specialized network architecture or any extra data augmentation, thus it can be easily plugged into existing powerful pre-trained language models. Compared to the state-of-the-art baselines, LaCon obtains up to 4.1% improvement on the popular datasets of GLUE and CLUE benchmarks. Besides, LaCon also demonstrates significant advantages under the few-shot and data imbalance settings, which obtains up to 9.4% improvement on the FewGLUE and FewCLUE benchmarking tasks.

1 Introduction

In recent years, contrastive learning (CL) has been widely applied to self-supervised representation learning and led to major advances across computer vision (CV) (He et al., 2019; Chen et al., 2020b), speech (Saeed et al., 2021; Chen et al., 2021), and natural language processing (NLP) (Fang and Xie, 2020; Gao et al., 2021; Yan et al., 2021). The basic idea behind these works is to pull together an anchor and a "positive" sample in the embedding space, and to push apart the anchor from many "negative" samples. Since no labels are available, a positive pair often consists of data augmentations of the sample (a.k.a "views"), and negative pairs are formed by the anchor and randomly chosen samples from the mini-batch. In visual representations, an effective solution to generate data augmentations is to take two random transformations of the same image (e.g., cropping, flipping, distortion and rotation) (Chen et al., 2020b; Grill et al., 2020; Chen et al., 2020c). For natural language, similar approaches are adopted such as word deletion, reordering, substitution, and back-translation etc. (Fang and Xie, 2020; Wang et al., 2021) However, data augmentation in NLP is inherently difficult because of its discrete nature. Therefore, some previous works (Gao et al., 2021; Yan et al., 2021) also use dropout technique (Srivastava et al., 2014) to obtain sentence augmentations.

Unlike self-supervised setting, some researchers propose supervised contrastive learning (SCL) (Khosla et al., 2020; Gunel et al., 2021; Suresh and Ong, 2021) which can construct positive pairs by leveraging label information. Examples from the same class are pulled closer than the examples from different classes, leveraging the semantics of labels to construct negatives and positives rather than shallow lexical information via data augmentation. Despite the aforementioned advantages brought by SCL, we argue that CL under supervised learning is not fully explored because the label information can be better utilized. On the one hand, labels are usually not merely categorical indices in the label vocabulary, but also contain specific semantic meanings, especially in the language understanding tasks. Thus labels can be used as positive/negative samples or anchors when calculating contrastive loss. On the other hand, label embedding enjoys a built-in ability to leverage alternative sources of information related to labels, such as class hierarchies or textual descriptions. Once we obtain representative label embeddings, they can be utilized to enhance the image/text representations, and finally facilitate the classification task. Previous la-

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1437 - 1449

bel embedding based classification models (Wang et al., 2018; Xiao et al., 2019) have demonstrated the effectiveness of leveraging label information.

Motivated by above analysis, we propose a novel label anchored supervised contrastive learning approach (denoted as LaCon), which combines the advantages of both contrastive learning and label embedding techniques. Specifically, we have the following three novel designs: 1) Instance-centered contrastive loss (ICL), which uses the InfoNCE (van den Oord et al., 2018) to encourage each text representation and its corresponding label representation to be closer while pushing far away mismatched instance-label pairs. We further apply a multi-head mechanism to catch different aspects of text semantics. 2) Label-centered contrastive loss (LCL), which takes label as anchor, and encourages the label representation to be more similar to the corresponding instances belonging to the same class in a mini-batch than the instances with different labels. 3) Label embedding regularizer (LER), which keeps the inter-label similarity as low as possible thus the feature space of each class is more dispersed to prevent representation degeneration. By combining above three losses, LaCon can learn good semantic representations within the same space for both input instances and labels. It's also well aligned with the two key properties related to CL: alignment and uniformity (Wang and Isola, 2020), where alignment favors encoders that assign similar features to similar samples. Uniformity prefers a feature distribution that preserves maximal information, i.e., the uniform distribution on the unit hypersphere.

To validate the effectiveness of LaCon, we perform extensive experiments on eight language understanding tasks. We take the popular pre-trained language model BERT-base (Devlin et al., 2019) as text encoder without loss of generality. For simplicity, we predict the classification label by matching the instance representation with label embeddings directly. Since our approach does not require any specialized network architecture or any extra data augmentation, LaCon can be easily plugged into other pre-trained language models. Additionally, we also explore the capability of LaCon under more difficult task settings, including few-shot learning and data imbalance situations.

To summarize, our contributions are as follows:

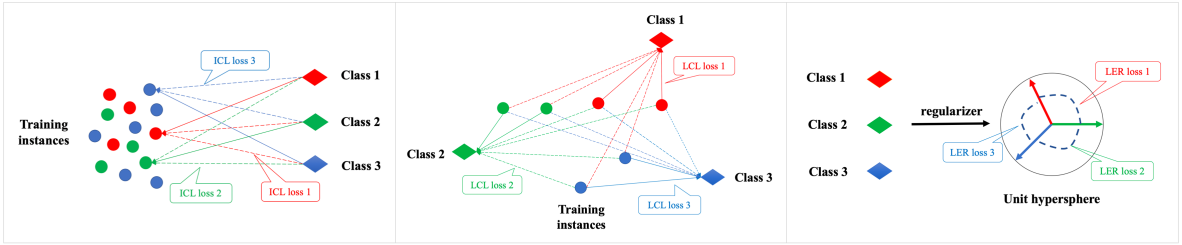
- We propose a novel label anchored contrastive learning approach for language understanding, which is equipped with a multi-head instance-centered contrastive loss, a label-centered contrastive loss, and a label embedding regularizer. All three contrastive objectives help the model learn the joint semantic representations for both input instances and labels.
- We conduct extensive experiments on eight public language understanding tasks from GLUE (Wang et al., 2019) and CLUE (Xu et al., 2020) benchmarks, and experimental results show the competitiveness of LaCon. Additionally, we also experiment on more difficult settings including few-shot learning and data imbalance situations. LaCon experimentally obtains up to 9.4% improvement over BERT-base on FewGLUE (Schick and Schütze, 2021) and FewCLUE (Xu et al., 2021) benchmark tasks.
- We analyze the contribution of each ingredient of LaCon, and also visualize the learned instance and label representations, showing the necessity of each loss component and the advantage of LaCon on representation learning over BERT fine-tuned with cross entropy.

2 Model

In this section, we introduce the details of LaCon. We focus on the language understanding classification tasks. For a multi-class classification problem with C classes, we work with a batch of training examples $\{x_i, y_i\}$, where $1 \leq i \leq N$ and $1 \leq y_i \leq C$. Our target is to learn discriminative representations for both instances and class labels. As Figure 1 shows, we propose three supervised CL based objectives, including the instance-centered contrastive loss, the label-centered contrastive loss, and the label embedding regularizer loss.

2.1 The Input Encoder

The input of LaCon contains two parts consisting of the text and all the labels for the task. Since SOTA language understanding classification models follow the "pre-training then fine-tuning" two-stage paradigm, here we take the prevalent pre-trained language model (PtLM) as input encoder. In this paper, we select BERT-base (Devlin et al., 2019) as the backbone for PtLM (denoted as f) without loss of generality. Given a text $\mathbf{x} = \{w_1, w_2, ..., w_M\}$ containing M tokens, the output of PtLM (i.e. BERT) is $\mathbf{E} = f_{PtLM}([CLS], w_1, ..., w_M)$ where



(a) Instance-centered Contrastive Loss (b) Label-centered Contrastive Loss (c) Label Embedding Regularizer Loss

Figure 1: Overview of LaCon. The full line is the similarity between a instance and corresponding label, and the dash line is the similarity between the mismatched instance and label. The lines with the same color denote the per-instance or per-label loss.

the $[CLS]$ token is the inserted sentence representation token and $\mathbf{E} \in R^{(M+1)\times d}$. d is the dimension of the model. We use the first token output $E_{[CLS]} \in R^d$ to represent the whole input text. We then apply a projection head (denoted as g) that is a 3-layer MLP with ReLU activation function for each hidden layer to the $E_{[CLS]}$, where the dimension of the g is also d. For a mini-batch X with N training samples, the text representations can be obtained as Equation 1.

$$H = g \circ f_{PtLM}(X), \text{ where}$$
$$X = x_1, x_2, ..., x_N \text{ and } H \in R^{N\times d} \quad (1)$$

Along with each mini-batch, LaCon also maps the C classes into C label embeddings. For simplicity, we look up in a learnable weight matrix W_{emb} and map the k^{th} label into the k^{th} row of W_{emb}, where the W_{emb} is randomly initialized. The dimension d is the same as PtLM. We then normalize the vectors for both L and H using the l_2 norm.

$$L = lookup([1, 2, ..., C], W_{emb})$$
$$\text{where } L \in R^{C\times d} \text{ and } W_{emb} \in R^{C\times d} \quad (2)$$

2.2 Instance-centered Contrastive Loss

Given a mini-batch of input text and corresponding label (x_i, y_i), as shown in Figure 1 (a), the instance-centered contrastive loss (ICL) takes each instance x_i as anchor, and mines positive and negative samples from class labels. ICL aims to encourage each text representation and corresponding label representation to be closer while pushing far away mismatched instance-label pairs. As shown in Equation 3, we modify the InfoNCE (van den Oord et al., 2018) to calculate the loss. Here we leverage the cosine similarity function as the distance metric sim. τ is the temperature hyper-parameter, which can be tuned to improve the performance. Lower temperature increases the influence of examples that are harder to separate, effectively creating harder negatives. Similar to the self-supervised CL, ICL also takes only one positive and many negatives. Differently, the positive is not generated from data augmentation, and the negatives are not randomly sampled from the same mini-batch. By treating the class labels as data samples, ICL can mine better positive and negatives with the supervision signal. By minimizing the ICL, the instance representation is aligned to its label representation in the same semantic space, which encourages the model to learn a more representative embedding for each class label.

$$\mathcal{L}_{ICL} = -\frac{1}{N}\sum_{x_i,y_i} log\frac{exp(sim(H_{x_i}, L_{y_i})/\tau)}{\sum_{1\le p\le C} exp(sim(H_{x_i}, L_p)/\tau)} \quad (3)$$

Inspired by the image-augmented views proposed in CV (He et al., 2019; Chen et al., 2020b,c; Grill et al., 2020), we also leverage the multi-head mechanism proposed in Transformer (Vaswani et al., 2017) to compute the ICL for each head representation with smaller representation dimension. Each head can be regarded as a clipped local view of the instance or label representation. Suppose we have m heads for both instance representation and label representation, then for the k_{th} head, the corresponding representations for training sample (x_i, y_i) are $h^k_{x_i}$ and $l^k_{y_i}$, and the dimension of each vector becomes $d' = d/m$. Then, we apply the contrastive loss for each head by following Equation 4. Compared with InfoNCE, $\mathcal{L}_{ICL}$ and $\mathcal{L}'_{ICL}$ do not suffer from small batch size issue (He et al., 2019; Chen et al., 2020c) because we only need to contrast the instance representation with corresponding label representation for per example loss.

$$\mathcal{L}'_{ICL} = -\frac{1}{N}\sum_{k=1}^{m}\sum_{x_i,y_i} log\frac{exp(sim(h^k_{x_i}, l^k_{y_i})/\tau)}{\sum_{1\le p\le C} exp(sim(h^k_{x_i}, l^k_p)/\tau)} \quad (4)$$

2.3 Label-centered Contrastive Loss

As shown in Figure 1 (b), we can take the class label in a mini-batch as anchor, and mine positive/negative samples from corresponding instances. Suppose there are $|P|$ classes in the batch, where $P = \{p|1 \leq p \leq C \wedge |A(p)| > 0\}$. We define that $A(p)$ denotes the set of indices of all positive instances whose label is p, i.e. $A(p) = \{x_i|y_i = p\}$. And $B(p)$ represents the set of negative instances whose label is not p, i.e. $B(p) = \{x_j|y_j \neq p\}$. Then we can calculate the label-centered contrastive loss (LCL) as Equation 5, which promotes the instances of a specific label to be more similar than the others for each label. Similar to the previous SCL (Khosla et al., 2020; Gunel et al., 2021), LCL also contains many positives per anchor and many negatives. Different from SCL which sums up all the softmax scores among all pairs of instances of the same class in a batch, LCL is based on comparing a specific label representation with corresponding instances (i.e. $A(p)$). LCL is more stable as the label representation serves as the anchor which can be stably updated.

$$\mathcal{L}_{LCL} = -\frac{1}{|P|}\sum_{p\in P}\sum_{a\in A(p)} log\frac{exp(sim(L_p, H_a)/\tau)}{\sum_{b\in B(p)} exp(sim(L_p, H_b)/\tau)} \quad (5)$$

ICL and LCL are complementary to each other and more computationally efficient than previous SCL. We conduct the detailed theoretical analysis in Appendix A.3 due to space limitation.

2.4 Label Embedding Regularizer

Recent researches (Wang and Isola, 2020) demonstrate that it is common and useful to add a regularization term during training to eliminate the anisotropy problem. Inspired by this, We devise a label embedding regularizer as shown in Equation 6 to promote the uniformity of our model and prevent model degeneration. As illustrated in Figure 1 (c), the label embedding regularizer (LER) encourages the label representations to be dispersed in the unit hypersphere uniformly. The LER loss is the exponential mean of the cosine similarity for all pairs of label representations. As $-1 \leq sim(L_i, L_j) \leq 1$, it is quite sensitive to the loss change as the gradient is larger than 1 for $exp(x)$ w.r.t $x \geq 0$. Thus, we add 1.0 to the cosine similarity so that the value of LER varies from 0 to e^2 - 1.

$$\mathcal{L}_{LER} = avg(\sum_{i\neq j}(exp(1.0 + sim(L_i, L_j)) - 1.0) \quad (6)$$

Finally, the overall loss function of LaCon is summarized as follows:

$$\mathcal{L} = \mathcal{L}'_{ICL} + \mathcal{L}_{LCL} + \lambda * \mathcal{L}_{LER} \quad (7)$$

where λ is a hyper-parameter to balance the influence of our regularization term.

2.5 Matching based Class Prediction

Since LaCon is capable of learning instance and label representations jointly, we can predict the class by matching the instance representation to all label representations directly during inference, just as shown in Equation 8. We denote this simple and direct approach as **LaCon-vanilla**. H_x is the instance representation and L_j is the label representation of Class j. sim denotes the cosine similarity and $1 \leq j \leq C$ denotes the corresponding label. The advantage of LaCon-vanilla is that it does not require any complicated network architecture and can be easily plugged into the mainstream PtLMs. As a result, our inference-time model contains exactly the same number of parameters as the model using the same encoder but trained with cross entropy loss.

$$pred = \arg_{max}(j)\{score_j|sim(H_x, L_j)\} \quad (8)$$

2.6 LaCon with Label Fusion

Previous researches (Akata et al., 2016; Wang et al., 2018; Xiao et al., 2019; Pappas and Henderson, 2019; Miyazaki et al., 2020) have proved that incorporating the label semantics into the sentence representation can improve the model performance because the label information can highlight the alignment of input tokens and label information via carefully designed fusion mechanism. Inspired by LEAM (Wang et al., 2018), here we design a fusion block to enhance the instance representations by utilizing the learnt discriminative label embeddings. We firstly calculate the cosine similarity interaction matrix G between words and labels, and then apply a convolution then max-pooling layer ($conv_{max}$) to measure the attention score (β_i) for each word attending the instance representation. The fusion process is illustrated as Equation 9. Then the fused vector z is fed into the projection head g to get the enhanced instance representation.

$$\begin{aligned} m = conv_{max}(G), \text{ where } G_{ij} = \frac{< L_i, E_j >}{||L_i|| \cdot ||E_j||} \\ z = \sum_i \beta_i E_i, \text{ where } \beta = softmax(m) \end{aligned} \quad (9)$$

To distinguish with the vanilla model above, we name this approach as **LaCon-fusion**. Please note that the fusion block is just applied between the text encoder and projection head, so the class prediction keeps the same as the LaCon-vanilla. Since the fusion block is not the main focus of this paper, we leave exploring more advanced fusion networks to future work.

Datasets	type	class	metric	train	dev
DBPedia	genre	14	ACC	14K	70K
Tnews	genre	15	F1	14.2K	10K
QQP	PI	2	ACC	10K	40.4K
MRPC	PI	2	ACC	4.07K	1.7K
QNLI	NLI	2	ACC	10K	5.4K
RTE	NLI	2	ACC	2.5K	278
CoLA	LA	2	M' corr	8.5K	1K
YelpRev	senti	2	ACC	10K	10K

Table 1: The statistics of datasets that are from GLUE (Wang et al., 2019) and CLUE (Xu et al., 2020)

3 Experiments

3.1 Experimental Setup

Datasets. We experiment on 8 public datasets listed in Table 1. which are from GLUE (Wang et al., 2019), CLUE (Xu et al., 2020), DBpedia, and Yelp Dataset Challenge 2015. They cover five representative tasks including sentiment analysis (senti), genre classification (genre), paragraph identification (PI), natural language inference (NLI) and linguistic acceptability (LA). To improve the comparability and experiment confidence of the models, we follow the experimental setup in (Chen et al., 2020a) and use part of the training sets via sampling and the full original test sets for evaluation. We randomly sample without replacement at most 5K (binary-class) / 1K (multi-class) training instances per class from the whole datasets except for MRPC, RTE, and CoLA. We use the wilcoxon rank test (Wilcoxon, 1945) to check the statistic significance. The results of 10 runs are reported for each dataset in the format as "avg$\pm std.dev$".

Training & Evaluation. During training, we run experiments for MRPC, RTE and CoLA with 10 random seeds on the whole training datasets and run the sampling strategy with 10 repeats for the remaining datasets. The average evaluation metrics are reported to avoid the noise and unstable randomness of a single run. We use the AdamW optimizer with initial learning rate as {1e-5, 2e-5, 3e-5} with linear learning scheduler, 6% of warm-up steps of total optimization steps, and batch size as {8,16,32,64,96}, where the hyper-parameters are tuned for different datasets. For evaluation, we leverage accuracy (ACC), Macro-F1 score and Matthew's corr (M' corr) metrics to evaluate the performance. We run 10 epochs for all the datasets[1] and then evaluate the models on dev set. Our implementation is based on Huggingface Transformers[2].

3.2 Baselines

Since LaCon is based on CL and label embedding technique, we compare with several SOTA models in language understanding including BERT-base fine-tuned with cross-entropy (CE) loss, label embedding based models, and self-supervised CL and supervised CL models.

- **CE**: we directly follow the instructions of original paper (Devlin et al., 2019) to finetune BERT for both English and Chinese langauge understanding tasks.
- **LEAM**: Wang et al. (2018) apply cosine similarity to get matching scores between words and labels and use CNN on the matching matrix to get the label-aware attention weighted text representation for classification.
- **LSAN**: Xiao et al. (2019) propose a label specific attention network that leverages label-attention and self-attention mechanism with an adaptive attention fusion strategy for multi-label classification. We use *softmax* instead of *sigmoid* for the model output due to our multi-class classification setting.
- **CE+CL**: Yan et al. (2021) propose to learn sentence representations by joint fine-tuning PtLM with InfoNCE and cross-entropy based on feature augmentation. Here we leverage the framework of ConSERT (Yan et al., 2021) and the feature augmentation in SimCSE (Gao et al., 2021) to finetune and evaluate the PtLM on classification datasets.
- **CE+SCL**: Gunel et al. (2021) propose to boost sentence representation learning by fine-tuning PtLM with both supervised contrastive learning loss (Khosla et al., 2020) and cross entropy loss. We follow the instructions in (Gunel et al., 2021) to set hyper-parameters.

[1] We split 5% of training set as validation for early stop.

[2] https://github.com/huggingface/transformers

Methods	YelpRev	DBPedia	Tnews	QNLI	RTE	QQP	MRPC	CoLA
CE	82.0±0.5	98.7±0.3	54.5±0.3	87.1±0.2	67.3±1.9	82.2±0.5	85.6±1.6	60.9±0.8
LEAM	82.1±0.6	98.7±0.5	54.1±0.3	87.2±0.7	67.3±1.3	81.9±0.5	85.6±1.3	60.9±1.0
LSAN	82.2±0.6	98.7±0.7	54.9±0.8	87.1±0.3	69.7±1.0	81.2±0.5	86.1±0.7	61.6±0.9
CE+CL	82.2±0.6	98.5±0.5	53.9±0.5	87.3±0.3	67.8±1.5	82.4±0.3	83.1±0.7	61.1±0.7
CE+SCL	81.4±0.8	98.5±0.6	54.6±0.2	87.7±0.1	69.1±2.2	82.5±0.6	88.1±0.9	62.3±0.6
LaCon-vanilla	82.3±0.5	98.9±0.5	**56.8±0.6**	88.1±0.2	71.4±0.7	82.8±0.5	87.5±0.9	62.4±1.0
LaCon-fusion	**83.1±0.8**	**99.5±0.2**	56.7±0.3	**88.4±0.3**	**72.2±0.9**	**83.7±0.5**	**88.6±0.7**	**62.8±0.5**

Table 2: The experimental results for the Language Understanding Tasks. Best scores for each dataset are highlight in **bold** (all with significance value $p < 0.05$).

LEAM and **LSAN** are label embedding based methods while **CE+CL** and **CE+SCL** are contrastive learning based methods. To compare all models fairly, we use BERT-base encoder for all the baselines and our proposed model.

3.3 Main Results

We report the experimental results of eight language understanding tasks in Table 2. It's observed that, LaCon-vanilla outperforms all the baselines in 7 datasets except MRPC, and LaCon-fusion achieves the best performance across all datasets. Specifically, 1) LaCon-vanilla outperforms BERT fine-tuned with CE by 4.1%, 2.3%, 1.9%, and 1.5% on RTE, Tnews, MRPC, and CoLA respectively, which indicates our proposed novel CL approach can facilitate the representation learning; 2) Compared with previous supervised contrastive learning method (CE+SCL), LaCon-fusion can still obtain very exciting improvements of 3.1%, 2.1%, 1.7%, 1.2% points on RTE, Tnews, YelpRev, QQP, which demonstrates the label fusion block can enhance the instance representations effectively; 3) Compared to previous label embedding methods (LEAM and LSAN) which are also equipped with the label fusion block, LaCon-fusion outperforms them with a large margin, which proves that LaCon can learn more discriminative joint representations for both labels and instances.

3.4 Ablation Study

In this section, we conduct three groups of ablation studies to investigate the contribution of each component in LaCon. We only conduct experiments on MRPC, RTE, and CoLA datasets due to space limitation. The experimental results are shown in Table 3. First, we replace the multi-head ICL with single head version (LaCon w/ $\mathcal{L}_{ICL}$). Table 3 shows the performance drops on all three datasets. We conjecture that the multi-head version can learn different parts of the local features of the representation, which can catch the text semantics in more fine-grained granularity. Second, we remove each of our proposed CL loss separately, and the results in the second part of Table 3 demonstrate that ICL plays a more important role while LCL and LER are complementary to further improve the performance. We also try to add each CL loss in an accumulative way, please refer to Appendix A.1 for more details. Finally, we try to remove the projection head g from LaCon and the performance degrades significantly, which indicates g is critical in CL. Previous researches (Chen et al., 2020c) also find the projector head can eliminate the non-task relevant features of the encoder in CL and benefit the downstream tasks. Meanwhile, Table 3 shows that it is basically useless by adding g to BERT directly (BERT w/ g), indicating that the projector head needs to be used with CL.

Methods	MRPC	RTE	CoLA
LaCon-vanilla	**87.5±0.8**	**71.4±0.7**	**62.4±1.1**
LaCon w/ $\mathcal{L}_{ICL}$	87.0±1.2	69.2±1.4	61.5±0.9
$-\mathcal{L}'_{ICL}$	86.6±1.3	68.1±0.9	61.3±0.7
$-\mathcal{L}_{LCL}$	87.1±0.6	70.5±0.8	61.2±1.1
$-\mathcal{L}_{LER}$	87.3±1.1	70.2±1.3	62.2±1.6
$-g$	86.8±0.6	69.6±0.6	62.2±0.9
BERT w/ g	84.9±1.7	66.5±2.1	61.0±1.2

Table 3: Ablation study. Best scores for each dataset are highlight in **bold** (all with significance test $p < 0.05$).

4 Discussion

In this section, we conduct further experiments under more challenging few-shot and data imbalance settings. We also discuss the hyper-parameter tuning and the impact of class number on LaCon.

4.1 LaCon for Few-shot Learning

Few-shot learning is critical for applications of language understanding models because the high-quality human annotated datasets are usually costly and limited. Previous researches (Liang et al., 2021; Gunel et al., 2021; Aghajanyan et al., 2021) find that fine-tuning PtLM with cross entropy loss

in NLP tends to be unstable across different runs especially when supervised data is limited. This limitation can result in model degeneration and model shift. Besides, some researches (Müller et al., 2019) also demonstrate that the cross-entropy optimization goal is not reachable due to the bounding of the gradient, which can also easily result in overfitting. Since LaCon is equipped by CL, it's interesting to validate if LaCon can overcome the shortcomings of CE under few-shot learning settings.

Model	YelpRev	Tnews	EPRSTMT	BUSTM
CE	59.0	52.5	84.4	65.6
LaCon	**65.0**	**55.8**	**90.6**	**75.0**
Model	**QNLI**	**RTE**	**MRPC**	**QQP**
CE	73.0	54.0	65.8	64.0
LaCon	**76.0**	**60.0**	**69.0**	**71.0**

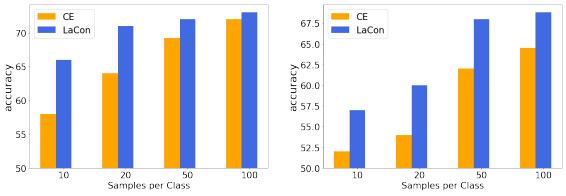
Table 4: Performance under few-shot learning settings.

We conduct further experiments with vanilla LaCon on 5 public English datasets from FewGLUE (Schick and Schütze, 2021) and 3 public Chinese datasets (Tnews, EPRSTMT and BUSTM) from FewCLUE (Xu et al., 2021). We build all the few-shot learning datasets by sampling **20 samples** for each class to form training set. We also held out the same amount of samples for validation set but keep the whole test set unchanged. We train the model for 20 epochs and select the best model based on validation set. Table 4 shows that LaCon significantly outperforms the BERT-base fine-tuned with CE loss with a huge margin. Specifically, we observe 9.4%, 7%, and 6.2% absolute improvement on BUSTM, QQP, and EPRSTMT.

Additionally, we also conduct more strict experiments by changing the number of samples per class from {10, 20, 50, 100}. Figure 2 demonstrates that the smaller the sample size per class is, the larger gain the model obtains. All above results indicate that the similarity-based CL losses in LaCon are able to hone in on the important dimensions of the multidimensional hidden representations hence lead to better and more stable few-shot learning results when fine-tuning PtLM.

4.2 LaCon for Data Imbalance Setting

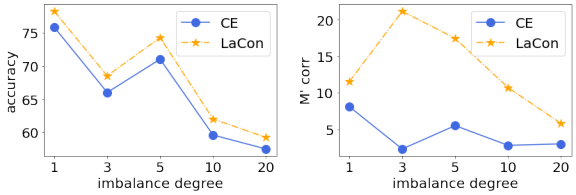
The real-world datasets are usually imbalanced for different classes (Cao et al., 2019; Bao et al., 2020), where several dominant classes contain most of the samples while the rest minority classes only hold a handful of samples. In this section, we conduct experiments to validate the capacity of LaCon under data imbalance setting. We follow the previous



(a) QQP (b) RTE

Figure 2: Few-shot learning with different number of training samples.

research (Cao et al., 2019) to construct imbalanced classification training datasets with different imbalance degree ($\rho = |class_{max}|/|class_{min}|$, where $|class_{max}|$ / $|class_{min}|$ denotes number of samples in maximum / minimum class). For space limitation, we conduct experiments with vanilla LaCon on QNLI and CoLA. The minority class contains 32 samples and the majority class contains $32 \times \rho$ in our experiments. As shown in Figure 3, we vary the imbalance degree (ρ) from {1, 3, 5, 10, 20} and observe that LaCon outperforms BERT with CE consistently, demonstrating that LaCon also has advantage on the data imbalance setting.



(a) QNLI (b) CoLA

Figure 3: LaCon with Different Imbalance Degree (ρ).

	QNLI		CoLA	
Methods	**minor**	**major**	**minor**	**major**
CE	43.4	71.8	2.3	81.3
LaCon	56.1	74.0	12.9	82.2

Table 5: F1 score for both majority and minority classes. Due to space limitation, we show results of $\rho = 10$.

We argue that LaCon may alleviate data imbalance issue on two aspects: 1) For the infrequent classes, treating labels as anchor or positive/negative may mitigate the data insufficient issue to some extent. 2) Label representations are shared across the whole dataset during training, which may transfer the knowledge from frequent classes to infrequent classes. To validate above conjecture, we present the performance on the test sets for majority and minority classes separately.

Table 5 shows that LaCon outperforms the baseline on both majority and minority classes and the gain on minority class is much larger.

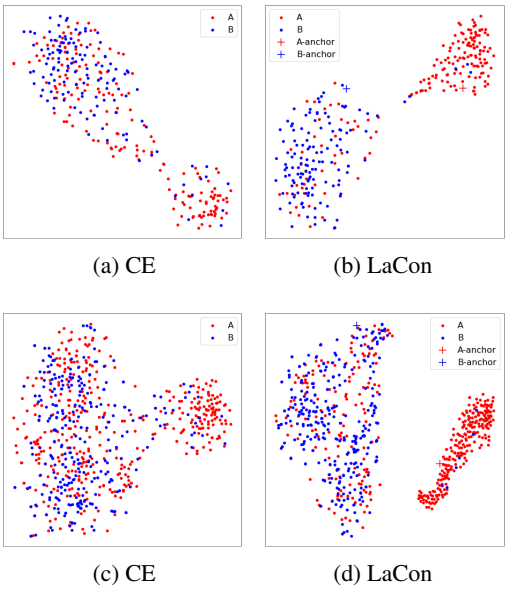
4.3 Visualization

To demonstrate the effectiveness of LaCon on representation learning, we visualize the learned instance representations of LaCon and CE on the MRPC and CoLA dataset. In Figure 4, we use t-SNE (Van der Maaten and Hinton, 2008) to visualize both the high dimensional representations of the instances and labels on a 2D map. Different classes are depicted by different colors. As shown in Figure 4 (a), the instances of class A and class B are sparsely located and overlapped in a large area, making it hard to find a hyper-plane to separate them. However, in Figure 4 (b), the instances gather into two compact clusters and the instances stay close to the corresponding class. For CoLA, Figure 4 (c) and (d) show the similar trends. It indicates that LaCon can learn more discriminative instance representations than CE. Besides, in Figure 4 (b), the instances are near the corresponding label anchor, proving that LaCon can also learn a representative label embedding for each class.

4.4 Hyper-parameter Tuning

In this section, we take the RTE dataset as an example for illustrating the hyper-parameter tuning process. The similar hyper-parameter tuning strategy is applied for other datasets. The tuning scripts will be released in our source code. Figure 5 shows the influence of different hyper-parameters.

For each experiment, we conduct a grid-based hyper-parameter sweep for τ between 0.05 and 0.5 with step 0.05, λ between 0.1 and 1.0 with step 0.1, and select the best hyper-parameter for the given dataset. The τ is the most influential hyper-parameter that needs to be tuned carefully with minimum step 0.05. Larger τ results in lower accuracy in LaCon and the recommended value is around 0.1 and 0.2. Figure 5 (b) illustrates that the optimal number of heads in Equation 4 is 6 and both the most and fewest heads result in low accuracy while heads with middle sizes get relatively better accuracy scores. Small number of head shows little diversity in feature clipping while larger one results in very short vectors with poor representation capacity. The label embedding regularizer weight λ in Figure 5(c) can be set in a wide range, where either without $\mathcal{L}_{LER}$ or large λ will result in poor performance.



(a) CE (b) LaCon

(c) CE (d) LaCon

Figure 4: Visualization of label and instance representations for MRPC (a&b) and CoLA (c&d) using T-SNE.

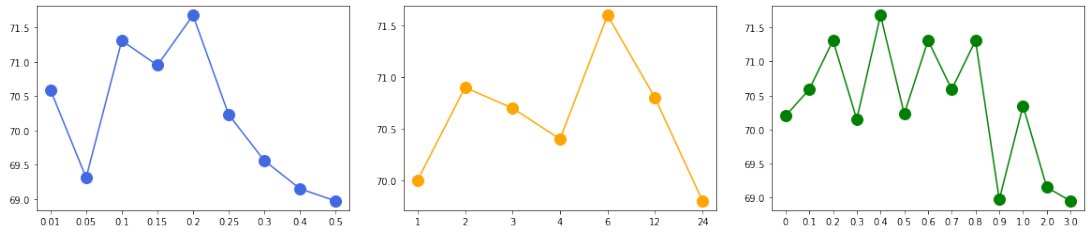
4.5 The Impact of Class Number

As the number of classes influence the difficulty of classification task directly, in this section, we discuss the impact of class number on our proposed model LaCon. We pick the DBPedia dataset for experiment. The original DBPedia dataset includes 14 labels. We gradually increase the label number from 2 to 14 and randomly select 1000 samples for each label in our experiment as training set. Meanwhile, we keep the whole samples for the chosen labels in the evaluation set unchanged. Figure 6 demonstrates that with the increase of the labels, the performance of all models degrades as the task becomes more difficult. However, LaConfusion outperforms CE+SCL consistently on different number of labels, which shows the advantage of leveraging labels as anchors or positive/negative samples during contrastive learning.

5 Related Work

5.1 Contrastive Learning

Contrastive Learning has become a rising domain and achieved significant success in various CV, speech and NLP tasks (He et al., 2019; Chen et al., 2020b; Fang and Xie, 2020; Han et al., 2021; Saeed et al., 2021; Gunel et al., 2021; Gao et al., 2021; Yan et al., 2021). There are two kinds of CL approaches, which are self-supervised CL and supervised CL. The self-supervised CL contrasts a



(a) τ (b) *heads* (c) λ

Figure 5: Illustration of hyper-parameters tuning (RTE is taken for example and other datasets are similar).

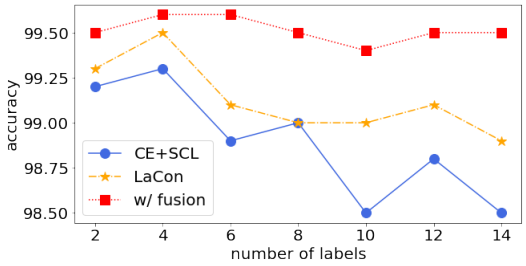


Figure 6: The impact of class number on LaCon. Experiments were conducted on DBPedia.

single positive for each anchor (i.e., an augmented version of the same image) against a set of negatives consisting of the entire remainder of the batch. However, due to the intrinsic discrete nature of natural language, data augmentations are less effective than that in CV. Recently, researchers (Khosla et al., 2020; Gunel et al., 2021) propose supervised CL, which contrasts the set of all samples from the same class as positives against the negatives from the remainder of the batch. Suresh and Ong (2021) propose label-aware SCL method via assigning weights to instances of different labels, which treats the negative samples differently.

LaCon belongs to the scope of supervised CL. Different from (Khosla et al., 2020; Gunel et al., 2021), LaCon can take the labels as anchors or mine negative/positive from labels, which does not need to construct positive pairs from the data augmentation. Meanwhile, Gunel et al. (2021) combine CL and CE losses at the same time, but LaCon is purely equipped with three CL objectives, including the instance-centered contrastive loss, the label-centered contrastive loss and the label embedding regularizer.

5.2 Label Representation Learning

Label representation learning aims to learn the embeddings of labels in classification tasks and has been proven to be effective in various CV (Frome et al., 2013; Akata et al., 2016) and NLP tasks (Tang et al., 2015; Pappas and Henderson, 2019; Nam et al., 2016; Zhang et al., 2018; Wang et al., 2018; Xiao et al., 2019; Miyazaki et al., 2020). In this work, we compare with two representative label embedding based models, which are LEAM (Wang et al., 2018) and LSAN (Xiao et al., 2019). Both learn label embeddings and sentence representations in a joint space based on attention mechanism and fuse them to improve the classification. Differently, LaCon learns the label and instance representations jointly via purely supervised contrastive learning. Besides, our experiments also verify that after obtaining the discriminative label and instance representations, even simple fusion block can facilitate the language understanding tasks.

6 Conclusions

In this paper, we proposed a novel supervised contrastive learning approach for language understanding. To utilize the class labels sufficiently, we devise three novel contrastive objectives, including a multi-head instance-centered contrastive loss, a label-centered contrastive loss, and a label embedding regularizer. Extensive experiments were conducted on eight public datasets from GLUE and CLUE benchmarks, showing the competitiveness of LaCon against various strong baselines. Besides, we also demonstrate the strong capacity of LaCon on more challenging few-shot and data imbalance settings, which leads up to 9.4% improvement on the FewGLUE and FewCLUE benchmarks. LaCon does not require any complicated network architecture or any extra data augmentation, and can be easily plugged into mainstream pre-trained language models. In the future, we will explore more advanced representation fusion approaches to enhance the capability of LaCon and plan to extend LaCon to the computer vision and speech fields.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2016. Label-Embedding for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. 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 1565–1576.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020a. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2147–2157. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020c. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

Yulong Chen, Jianping Zhao, Weiqi Wang, Ming Fang, Haimei Kang, Lu Wang, Tao Wei, Jun Ma, Shaojun Wang, and Jing Xiao. 2021. SEQ-CPC : Sequential contrastive predictive coding for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 3880–3884. IEEE.

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*, page 4171–4186.

Hongchao Fang and Pengtao Xie. 2020. CERT: contrastive self-supervised learning for language understanding. *CoRR*, abs/2005.12766.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2121–2129.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2214–2225. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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, Online. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *CoRR*, abs/2106.14448.

Jorma Kaarlo Merikoski. 1984. On the trace and the sum of elements of a matrix. *Linear Algebra and its Applications*, 60:177–185.

Taro Miyazaki, Kiminobu Makino, Yuka Takei, Hiroki Okamoto, and Jun Goto. 2020. Label embedding using hierarchical structure of labels for twitter classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 6317–6322.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? 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 4696–4705.

Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. All-in text: Learning document, label, and word representations jointly. In *Proceedings of the thirtieth AAAI conference on artificial intelligence*, pages 1948–1954.

Nikolaos Pappas and James Henderson. 2019. GILE: A Generalized Input-Label Embedding for Text Classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.

Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 3875–3879. IEEE.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2339–2352. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Varsha Suresh and Desmond C. Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4381–4394. Association for Computational Linguistics.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. PTE: Predictive text embedding through large-scale heterogeneous text networks. *In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015-August:1165–1174.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: contrastive learning with semantic negative examples for natural language understanding. 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 2332–2342. Association for Computational Linguistics.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020. Improving neural language generation with spectrum control. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. 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 466–475.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Hu Yuan, Huilin Xu, Guoao Wei, Xiang Pan, and Hai Hu. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *CoRR*, abs/2107.07498.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5065–5075. Association for Computational Linguistics.

Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2018. Multi-task label embedding for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4545–4553.

A Appendix

A.1 Accumulative Ablation Study

In this section, we supplement more ablation results by adding each proposed CL loss cumulatively. We conduct experiments on MRPC, RTE, and CoLA datasets and keep the setting consistent with Section 3.4. Table 6 demonstrates that the contribution of each component in more details.

Methods	MRPC	RTE	CoLA
$\mathcal{L}_{ICL}$	86.2±1.5	67.8±1.1	61.2±0.9
$\mathcal{L}^{'}_{ICL}$	86.9±1.1	68.2±0.7	61.2±1.5
$\mathcal{L}_{LCL}$	87.0±1.7	68.3±1.0	61.1±1.3
$\mathcal{L}^{'}_{ICL} + \mathcal{L}_{LCL}$	87.3±1.1	70.2±1.3	62.2±1.6
$\mathcal{L}^{'}_{ICL} + \mathcal{L}_{LER}$	87.1±0.6	70.5±0.8	61.2±1.1
LaCon-vanilla	**87.5±0.8**	**71.4±0.7**	**62.4±1.1**

Table 6: Ablation study via adding losses cumulatively.

A.2 Experimental Results of More Datasets

We supplement more experimental results on the remaining datasets of GLUE and CLUE benchmarks. We follow the same experimental setup with Section 3.1. Please note that SST-B is a regression task that is beyond the capacity of the proposed LaCon. The official CLUE benchmark has replaced CMNLI with OCNLI dataset, and the CSL dataset is a keyword recognition task, which is not suitable for our proposed model. Thus, we omit the experiments on above three datasets and report the performance on the remaining language understanding tasks including SST-2, MNLI, AFQMC, OCNLI and IFLYTEK.

Table 7 shows that, LaCon-vanilla consistently outperforms BERT fine-tuned with CE, and LaCon-fusion still beats the baselines among all datasets, which further demonstrates the superiority of our proposed method.

A.3 Theoretical Analysis

In this section, we conduct the theoretical analysis to prove the rationality and necessity of our proposed ICL and LCL losses. We also explain why these two losses are complementary. Finally, we analyze the computational efficiency of ICL and LCL compared to InfoNCE (van den Oord et al., 2018) and SCL (Khosla et al., 2020).

The recent researches (Li et al., 2020; Gao et al., 2019) reveal that the anisotropy problem of pre-trained language models, which shows that the learnt embeddings occupy a narrow cone in the dense vector space, harming the uniformity of the models and limiting the representation capacity. The singular values of the contextual embeddings decay drastically with most of them nearly zeros (Wang and Isola, 2020). CL is proposed to eliminate the long-tail distribution problem of singular values, aiming to enhance the representation capacity (Gao et al., 2021; Yan et al., 2021; Gunel et al., 2021). From the spectrum perspective (Wang et al., 2020; Wang and Isola, 2020) that analyzes the distribution and uniformity of the learned embedding space, CL flattens singular values of the embeddings thus improves the capacity of language

Methods	SST-2	MNLI	AFQMC	OCNLI	IFLYTEK
Baseline*	91.4±0.3	73.1±0.3	74.5±0.3	68.3±0.7	61.5±0.5
CE	91.2±0.1	72.5±0.3	70.9±0.5	66.8±0.4	60.6±0.2
LaCon-vanilla	91.4±0.3	73.3±0.4	72.5±0.3	67.4±0.3	60.9±0.2
LaCon-fusion	**92.5±0.2**	**73.9±0.3**	**74.8±0.5**	**69.1±0.6**	**63.5±0.7**

Table 7: Performance on the remaining datasets of GLUE and CLUE. Baseline* means the best performance of our baselines. The evaluation metrics are the same as the official GLUE (Wang et al., 2019) and CLUE (Xu et al., 2020) benchmarks (all with significance value $p < 0.05$).

models.

$$\begin{aligned}\mathcal{L}_{ICL} &= -\frac{1}{\tau}E_{(x,y)\sim A(y)}(H_x L_y) \\ &+E_{x\sim A(y)}[logE_{y^-\notin A(y)}(e^{H_x L_{y^-}/\tau})]\end{aligned} \tag{10}$$

$$\begin{aligned}&E_{x\sim A(y)}[logE_{y^-\notin A(y)}(e^{H_x L_{y^-}/\tau})] \\ &= \frac{1}{N}\sum_{i=1}^{i=N} log(\frac{1}{C-1}\sum_{j\neq i}^{1\leq j\leq C} e^{H_{x_i} L_{y_j}/\tau}) \\ &\geq \frac{1}{N(C-1)\tau}\sum_{i=1}^{i=N}\sum_{j=1,j\neq i}^{j=C} H_x L_{y^-}\end{aligned} \tag{11}$$

Therefore, we can form an asymptotic equivalent objective of the $\mathcal{L}_{ICL}$ (Equation 3) as Equation 10. $(x, y) \sim A(y)$ denotes instances (i.e. x) with corresponding label (i.e. y) and y^- denotes the label that is different from y. The first item keeps instances and corresponding labels similar and the second item pushes the mismatched instances and labels apart. We can further derive Equation 11 using Jensen's inequality because $e(.)$ is convex. Therefore, minimizing the $\mathcal{L}_{ICL}$ equals to minimization of summation of all elements in $HL^T \in R^{N\times C}$. Because both H and L are normalized, $tr(H^T L)$ is a constant due to all diagonal elements are ones. $sum(HL^T)$ is an upper bound of the largest singular value (Merikoski, 1984) and minimization of the $sum(HL^T)$ will flatten the singular values distribution of HL^T. As the HL^T is a non-squared matrix, we need to optimize both the left and right singular values using HL^T and LH^T in order to effectively eliminate the anisotropy and promote the uniformity of pre-trained language models in classification tasks to enhance the model capacity. Thus, we also need to optimize the label-centered contrastive loss $\mathcal{L}_{LCL}$ at the same time. From above analysis, we can see that $\mathcal{L}_{ICL}$ and $\mathcal{L}_{LCL}$ are complementary to each other. Similarly, we can derive that minimizing $\mathcal{L}_{LCL}$ results in the minimization of $sum(LH^T) \in R^{C\times N}$.

Although both ICL and LCL calculate the $N \times C$ similarity scores for a mini-batch, they are different. The ICL is the average of the instance-level per sample loss while the LCL is the per label loss. The ICL intends to align each instance to corresponding label correctly. The LCL makes the instances of different labels far away from each other and instances of the same label more compact. They consider different aspects of instance and label representation through operating the $N \times C$ similarity scores differently according to Equation 3 and 5.

Compared to InfoNCE, ICL improves the computational efficiency from $O(N^2)$ to $O(NC)$ because we only need to contrast the instance representation with corresponding label representations for per example loss, which is extremely useful for language understanding tasks (Wang et al., 2019; Xu et al., 2020) that commonly consist of 2 or 3 labels. Similarly, LCL is also more computationally efficient as it only contrasts one label representation to several instances rather than computes all pairs of instances belonging to a given label in the mini-batch. Thus it improves the complexity from $O(N^2)$ to $O(CN)$ compared with SCL too.