PACT: Pretraining with Adversarial Contrastive Learning for Text Classification

Md Tawkat Islam Khondaker[♠] Muhammad Abdul-Mageed[♠]♦ Laks V.S. Lakshmanan[♠]

[♠]The University of British Columbia, [◊]MBZUAI

{tawkat@cs.,muhammad.mageed@,laks@cs.}ubc.ca

Abstract

We present PACT (Pretraining with Adversarial Contrastive Learning for Text Classification), a novel self-supervised framework for text classification. Instead of contrasting against inbatch negatives, a popular approach in the literature, PACT mines negatives closer to the anchor representation. PACT operates by endowing the standard pretraining mechanisms of BERT with adversarial contrastive learning objectives, allowing for effective joint optimization of token- and sentence-level pretraining of the BERT model. Our experiments on 13 diverse datasets including token-level, singlesentence, and sentence-pair text classification tasks show that PACT achieves consistent improvements over SOTA baselines. We further show that PACT regularizes both token-level and sentence-level embedding spaces into more uniform representations, thereby alleviating the undesirable anisotropic phenomenon of language models.¹

1 Introduction

Pretrained language models (PLM) like BERT (Devlin et al., 2019) revolutionized several NLP tasks such as text classification, question answering, etc. With the success of PLMs, different pretraining objectives were proposed to further improve model performance (Liu et al., 2019; Lan et al., 2020; Clark et al., 2020). PLMs can also be finetuned on downstream task data (Howard and Ruder, 2018). One of the exciting lines of work aimed at sharpening PLM representations is related to contrastive learning (CL) (Hadsell et al., 2006). These works are motivated by recent success in computer vision (Chen et al., 2020a; Dosovitskiy et al., 2014; Chen et al., 2020b, 2017). The basic idea behind CL is to pull positive samples close to each other while pushing apart negative samples in the embedding space. While these positive and negative sam-



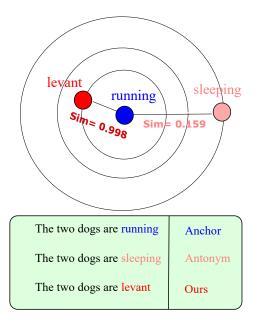


Figure 1: Visual comparison between our adversarial approach and existing approach on the token-level negative. Our approach adversarially generated token (*levant*) is closer to the anchor token (*running*) than the antonym token (*sleeping*) in the embedding space.

ples are already labeled and hence can be used for *supervised finetuning* (Khosla et al., 2020; Gunel et al., 2021), it is still challenging to mine positive and especially negative samples for *self-supervised pretraining*.

In NLP, CL has been used both for language model pretraining (Gao et al., 2021; Fang et al., 2020; Yan et al., 2021; Qu et al., 2020) and finetuning (Suresh and Ong, 2021; Zhang et al., 2022c). Prior works on CL-based pretraining have applied different data augmentation methods such as dropout (Gao et al., 2021), backtranslation (Fang et al., 2020), adversarial attack, and token-shuffling (Yan et al., 2021) for mining positive samples. However, these methods rely mostly on sampling independently from the training batch (in-batch negatives) to collect negative samples, regardless of how uninformative these negative samples may be for the learned representation. Effectively pretraining a language model requires informative negative examples that are mapped nearby the positive samples but should be far apart from one another (*hard negatives*) (Robinson et al., 2021). Existing works (Wang et al., 2021) attempt to synthetically generate semantic negative examples by replacing some tokens with antonyms. However, such approaches require human cognitive and external knowledge resources (e.g., dictionary), and there is no guarantee that representations obtained by such replacements are close to the actual representation in the embedding space (i.e., there is no guarantee they are actually hard negatives).

Another issue with PLMs is that they suffer from anisotropy (Ethayarajh, 2019; Li et al., 2020) in the embedding space. That is, representations obtained by PLMs tend to occupy a narrow cone in the hyperspace, making them less informative. This makes it harder for classifiers to push apart samples belonging to different classes. Although prior works attempt to address this issue using CL separately for token-level (Su et al., 2022) embedding and sentence-level (Gao et al., 2021), the sentence-level work focused mainly on acquiring representations that practically work fine for semantic similarity tasks but not for sentence-level classification (e.g., sentiment analysis). In addition, it is yet to be explored how to jointly optimize both token-level and sentence-level representations to achieve better uniformity (i.e., to alleviate anisotropy).

To address issues above, we present PACT, our self-supervised pretraining method for text classification. First, we introduce adversarial masked language modeling (MLM) to mine negative tokens by adding a small perturbation to the masked token representation in order to reduce the maximum likelihood of the correct token. Since we regulate the perturbation within a small margin, it guarantees to produce adversarial tokens within the vicinity of the masked tokens (Figure 1). Next, we adversarially perturb the next sentence prediction (NSP) objective of BERT to minimize the maximum likelihood of correct prediction. For the contrastive learning objective, we treat the obtained token-level and sentence-level representations as negative pairs. Since both of these are acquired with only small perturbations, their representations stand as hard negatives. Our proposed method is completely selfsupervised and simple in that it aligns with the original pretraining objective of BERT. We further

show that the joint token-level and sentence-level pretraining ensures uniformity in acquired representations and thus alleviates anisotropy.

Our contributions are as follows:

- 1. We propose PACT, a novel pretraining framework for BERT which jointly optimizes tokenlevel and sentence-level representations using CL.
- 2. We introduce adversarial MLM and Sequence objectives to mine adversarial hard negative samples in a self-supervised fashion.
- 3. Our experiments on 13 different token and sentence classification tasks show that PACT achieves consistent improvement over the SOTA baselines.
- 4. We show that PACT demonstrates better sentence-level (Section 6.1) and token-level (Section 6.2) uniformity than other baselines that alleviate the problem of anisotrophy.

2 Related Works

Contrastive Learning. CL aims to learn effective embeddings by pulling semantically close neighbors together while pushing apart nonneighbors (Hadsell et al., 2006). CL employs a similarity objective to learn the embedding representation in the hyperspace (Chen et al., 2017; Henderson et al., 2017). In computer vision, Chen et al. (2020a) propose a framework (SimCLR) for CL of visual representations without specialized architectures or a memory bank. Chen et al. (2020b) dynamically build a queue of in-batch negative samples. The authors use a moving-averaged encoder with the dynamic queue to facilitate unsupervised CL. Hu et al. (2021) argue that such queue may not be able to track the change of the learned representations. Hence, the authors propose an adversarial contrast (AdCo) model consisting of two adversarial networks. One is a backbone representation network that encodes the representation of input samples. The other is a collection of negative adversaries that are used to discriminate against positive queries over a minibatch. By this way, AdCo updates negative samples as a whole by making them sufficiently challenging to train the representation network. In NLP, several memory-based methods have been explored in the context of sentence representation learning (Karpukhin et al., 2020; Gillick et al., 2019; Logeswaran and Lee, 2018).

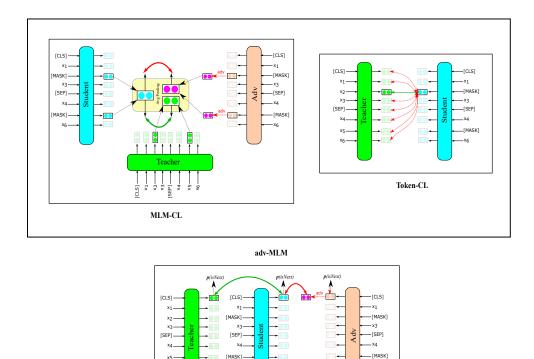




Figure 2: Overview of our proposed framework. *PACT* consists of two principal CL objectives: (i) adv-MLM (*top*) and (ii) adv-Sequence (*bottom*). adv-MLM further consists of two losses: (a) \mathcal{L}_{MLM-CL} (*top-left*) and (b) \mathcal{L}_T (*top-right*). The core idea behind adv-MLM and adv-Sequence objectives is to *pull* the representations of *teacher*-BERT and *student*-BERT together and *push* the representations of *student*-BERT and *adv*-BERT apart.

Self-Supervised CL. CL approaches in NLP can be broadly categorized into two types: (i) selfsupervised and (ii) supervised. For self-supervised CL, one of the most notable works is SimCSE (Gao et al., 2021), which augments an input sentence with another view of the same sentence after applying dropout. Meng et al. (2021) introduce an auxiliary model to train the student model using corrupted text sequence through ELECTRAstyle (Clark et al., 2020) pretraining. Fang et al. (2020) pretrain BERT with back-translation and show improvement in natural language understanding (NLU) tasks. DeCLUTR (Giorgi et al., 2021) adopts the architecture of SimCLR and jointly trains two encoders to maximize the agreement between a span of a sequence. Wu et al. (2020) advances DeCLUTR with both word- and span-level data augmentation strategies.

Supervised CL. For the supervised setting, Khosla et al. (2020); Gunel et al. (2021) propose to directly use the representations of the same class as positive pairs and different classes as negative pairs. Pan et al. (2022) propose adversarial perturbation of the word embedding layer of BERT during finetuning.

Similarly, Lee et al. (2021) propose to generate adversarial tokens for text generation tasks. Suresh and Ong (2021) introduce an additional weighting network to capture inter-label relationships for fine-grained classification, while Zhang et al. (2022c) initialize additional label embeddings to match the representations of instances and corresponding labels.

Adversarial Learning. In the literature (e.g., (Miyato et al., 2017; Pan et al., 2022; Qiu et al., 2021)), adversarial perturbation was used for data augmentation as a way to improve model robustness (i.e., by making the model invariant to adversarial samples). Diverging from the literature, we employ adversarial perturbation in a completely different way: instead of using adversarial samples as positive pairs to enhance robustness, we use the adversarial samples generated through selfsupervised learning as negative pairs. Therefore, our model learns to *differentiate* between the anchor and the adversarial representations instead of making them invariant in the embedding space. Thus, PACT is pretrained to be discriminative of negative samples closely located near the anchor

representation in the embedding space.

3 Proposed Framework

PACT introduces two novel CL objectives for BERT (Devlin et al., 2019), namely adversarial MLM and adversarial Sequence. For this purpose, we design a self-supervised framework consisting of three BERT-base models: one teacher model (teacher-BERT), one adversarial model (adv-BERT), and the main model (student-BERT). The purpose of teacher-BERT is to provide positive examples for student-BERT. We pass the same examples to teacher-BERT and student-BERT to generate both token- and sentence-level representations. Since these representations are obtained without further manipulation, we can consider them positive pairs for the contrastive learning (CL) objective. On the other hand, the purpose of adv-BERT is to provide negative examples that are closely located with the anchor examples (examples obtained from student-BERT) in the embedding space. After obtaining the representation from the adv-BERT, we add a small adversarial perturbation to minimize the likelihood of the correct representation (which means the likelihood of both the correctly predicted token and the next sentence prediction will be minimized). In this way, the manipulated representation coming from the adv-BERT is considered a negative pair for the anchor representation in CL objective.

Now we describe the proposed two CL (adversarial MLM and adversarial Sequence) objectives. The overall framework is shown in Figure 2.

Adversarial MLM: BERT uses masked language modeling (MLM) objective, which takes an input sequence, $\mathbf{X} = \{x_1, x_2, ..., x_i, ..., x_n\}$, masks out a random token (e.g. *i*-th token), and attempts to predict the original token given a contextualized representation of the sequence:

$$p_{MLM}(\hat{x}_i \,|\, \mathbf{h}(i)) = \frac{exp\left(\psi(\hat{x}_i)^T \,\mathbf{h}(i)\right)}{\sum_{x_t \in V} exp\left(\psi(x_t)^T \,\mathbf{h}(i)\right)}$$

where, $\psi(.)$ is the token embedding matrix and $H = \{\mathbf{h}(i)\}_{i=1}^{n}$ is the contextualized vector representation generated by BERT. The pretraining objective is to minimize MLM loss and maximize the likelihood of the correct tokens at a set of masked positions \mathcal{M} :

$$\mathcal{L}_{MLM} = \mathbf{E}\left(-\sum_{i \in \mathcal{M}} \log p_{MLM}(x_i \,|\, \mathbf{h}(i))\right)$$

We introduce adversarial learning to this MLM objective of adv-BERT to perturb the contextualized representation of the masked token by a small margin so that the maximum likelihood of the correct token is minimized. For this purpose, we use the same MLM objective as BERT. BERT optimizes the masked representation by going along the direction of the gradient to predict the correct token. However, instead of taking the original masked embedding to predict the correct token, we manipulate it by adding a small perturbation. Unlike BERT, we take the opposite direction of the gradient and add it with the original masked representation. As a result of this perturbation, we get an adversarial token representation where the probability of predicting the correct token is minimized:

$$\delta = \arg \min_{\hat{\delta}} p_{MLM}(\hat{x}_i | \mathbf{h}(i) + \hat{\delta}) \ s.t. \ ||\hat{\delta}|| < \epsilon, \ \epsilon > 0$$

$$\mathbf{h_{adv}}(i) = \mathbf{h}(i) + \delta$$

We use *fast gradient sign method* (FGSM) (Good-fellow et al., 2015) to approximate the perturbation δ with a linear approximation around $\mathbf{h}(i)$ and an *L2* norm constraint:

$$\mathbf{h_{adv}}(i) = \mathbf{h}(i) - g / ||g||_2$$

where $g = \nabla_{\mathbf{h}(i)} \log p_{MLM}(\hat{x}_i | \mathbf{h}(i))$. We normalize the gradient g by $||g||_2$, to keep the adversarial representation $\mathbf{h}_{adv}(i)$ close to $\mathbf{h}(i)$ in the embedding space.

We pass the contextualized representations $(\mathbf{h}(i), \mathbf{h_T}(i), \mathbf{h_{adv}}(i))$ of the masked tokens of all three models through non-linear projection layers and take the average of the representations to obtain \mathbf{Z} , $\mathbf{Z_T}$, and $\mathbf{Z_{adv}}$:

$$z(i) = \phi(\mathbf{W}.h(i) + b)$$
$$\mathbf{Z} = \frac{\sum z(i)}{|z|}$$

We apply CL loss on the obtained \mathbf{Z} , where for the *i*-th sample in the batch, the model learns to increase the similarity between the representation of the student model ($\mathbf{Z}^{(i)}$) and the teacher model ($\mathbf{Z}_{\mathbf{T}}^{(i)}$), while decreasing the similarity between the representation of the student model $(\mathbf{Z}^{(i)})$ and the adversarial model $(\mathbf{Z}_{adv}^{(i)})$.

$$\mathcal{L}_{MLM-CL} = -\sum_{i=1}^{N} \log \frac{\exp(sim(\mathbf{Z}^{(i)}, \mathbf{Z_{T}}^{(i)})/\tau)}{\sum_{\{\hat{\mathbf{Z}} = \{\mathbf{Z_{T}}^{(i)}\} \cup \mathbf{Z_{adv}}^{(i)}\}} \exp(sim(\mathbf{Z}^{(i)}, \hat{\mathbf{Z}}^{(k)})/\tau)}$$

Following Su et al. (2022), we further apply token-CL among the masked tokens of student-BERT and teacher-BERT:

$$\mathcal{L}_T = -\sum_{i=1}^N \log \frac{exp(sim(\mathbf{h}(i), \mathbf{h_T}(i))/\tau)}{\sum_{k=1}^N exp(sim(\mathbf{h}(i), \mathbf{h_T}(k))/\tau)}$$

Our final token-level CL is the summation of the above two losses:

$$\mathcal{L}_{adv-MLM} = \mathcal{L}_{MLM-CL} + \mathcal{L}_{T}$$

Adversarial Sequence: Additionally, we propose to adversarially modify the next sentence prediction (NSP) objective of BERT. Given two sequences X_1 and X_2 , the NSP loss is based on the prediction of whether the two sequences are next to each other:

$$\mathcal{L}_{NSP} = \mathbf{E} \left(-\log p_{NSP}(isNext \mid \mathbf{C}) \right)$$

where C is the contextualized representation of the [CLS] token, h([CLS]). Similar to adversarial MLM, we apply FGSM to the NSP² objective of adv-BERT to obtain C_{adv} and use CL loss to push it apart from the [CLS] representation of student-BERT, C:

$$\mathbf{C_{adv}} = \mathbf{C} - g / ||g||_2$$

where $g = \nabla_{\mathbf{C}} \log p_{NSP}(isNext \mid \mathbf{C})$

$$\mathcal{L}_{adv-Seq} = -\sum_{i=1}^{N} \log \frac{exp(sim(\mathbf{C}^{(i)}, \mathbf{C}_{\mathbf{T}}^{(i)})/\tau)}{\sum_{\{\hat{\mathbf{C}} = \{\mathbf{C}_{\mathbf{T}}^{(i)}\} \cup \mathbf{C}_{adv}^{(i)}\}} exp(sim(\mathbf{C}^{(i)}, \hat{\mathbf{C}}^{(k)})/\tau)}$$

To avoid catastrophic forgetting (McCloskey and Cohen, 1989; Sun et al., 2019), we continue pretraining student-BERT and adv-BERT with \mathcal{L}_{MLM} , \mathcal{L}_{NSP} objectives. The final loss function for student-BERT is the linear combination of the various loss terms:

$$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{NSP} + \mathcal{L}_{adv-MLM} + \mathcal{L}_{adv-Seq}$$
(1)

Dataset	Task	Classification Type	Source
CoLA	Linguistic acceptability	Single-sentence	GLUE
SST-2	Sentiment Analysis	Single-sentence	GLUE
NC	News Classification	Single-sentence	XGLUE
MRPC	Paraphrase Identification	Sentence-pair	GLUE
QQP	Question Paraphrase	Sentence-pair	GLUE
MNLI	Natural language inference	Sentence-pair	GLUE
QNLI	Question answer entailment	Sentence-pair	GLUE
RTE	Textual entailment	Sentence-pair	XGLUE
QAM	Question passage entailment	Sentence-pair	XGLUE
QADSM	Query-Advertisement Matching	Sentence-pair	XGLUE
PAWSX	Paraphrase Identification	Sentence-pair	XGLUE
NER	Named-entity recognition	Token-level	XGLUE
POS	Part-of-speech tagging	Token-level	XGLUE

Table 1: Summary of the datasets used in this paper.

We initialize all three models with pretrained BERT-base weights at the start of training. We freeze teacher-BERT weights and update adv-BERT with \mathcal{L}_{MLM} , \mathcal{L}_{NSP} losses. We now present our experiments.

4 **Experiments**

4.1 Datasets

To evaluate the efficacy of PACT, we conduct experiments on 13 diverse datasets from GLUE (Wang et al., 2019) and XGLUE (Liang et al., 2020) benchmarks. We cover both single-sentence and sentence-pair text classification tasks. We further experiment on named entity recognition (NER) and part-of-speech (POS) tagging datasets to evaluate the model's performance on the token-level classification tasks. A summary of the datasets that we experiment on is presented in Table 1. We describe the detailed experimental setup in Appendix A.

4.2 Baselines

We compare PACT with the following state-of-theart BERT-base models pretrained with contrastive learning objective:

- **TaCL (Su et al., 2022)** propose token-level contrastive learning to produce diverse token representation from BERT.
- SimCSE (Gao et al., 2021) propose dropoutbased data augmentation as positive pairs and in-batch examples as negative pairs.
- Mirror-BERT (Liu et al., 2021) construct positive pairs by random span masking as well as different dropout masks.
- SCD (Klein and Nabi, 2022) optimize a joint self-contrastive and decorrelation objective by leveraging the instantiation of standard dropout at different rates.

²Although we demonstrate the adversarial Sequence objective with NSP, it can be implemented on any transformer model with sequence-level pretraining. For example, AL-BERT (Lan et al., 2020) uses sentence-order prediction (SOP) instead of NSP. We can similarly apply FSGM on the SOP loss to compute adversarial Sequence. To be able to directly compare with existing CL-pretrained SOTA in the literature, we focus on the BERT-based architecture in this work.

- DiffCSE (Chuang et al., 2022) introduce an unsupervised contrastive learning framework that is sensitive to the difference between the original sentence and an edited sentence. The edited sentence is obtained by stochastically masking out the original sentence and then sampling from a masked language model.
- **BERT-PT.** In addition to comparing with BERT-base, we further pretrain it (i.e., BERT-base) with MLM and NSP objectives for an equal number of training steps as PACT to facilitate fair comparison.

5 Results

Sentence-level Classification: We report performance of the models for both single-sentence and sentence-pair classification tasks in Table 2. Especially, PACT either outperforms other models (CoLA, NC) or maintains comparable performance (SST-2) with the published results on the singlesentence classification tasks. For sentence-pair classification tasks, PACT achieves the best score (QNLI, RTE, QADSM, and PAWSX) or the joint best score (QQP and QAM) except for MNLI. Overall, PACT improves performance for both singlesentence classification and sentence-pair classification tasks. The consistent improvement across 11 different sentence-level classification tasks shows the efficacy of the proposed sentence-level (adv-Sequence) contrastive objective.

Token-level Classification: Table 3 shows our evaluation results on token-level classification tasks. As observed, PACT outperforms other models in both NER and POS tagging tasks. This result highlights that PACT can also improve over token-level classification tasks by pretraining on token-level contrastive learning. Although TaCL is also pretrained on token-level CL, our adversarial MLM-based CL objective helps the model differentiate very similar tokens that may belong to different classes during the finetuning stage, which helps the model perform better on the downstream tasks.

Overall, PACT is pretrained to contrast both token-level and sentence-level adversarial hard negatives. Therefore, the joint optimization helps the model produce discriminative representations to improve on the downstream tasks.

6 Analysis

6.1 Uniformity-Tolerance Dilemma

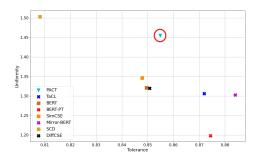


Figure 3: Uniformity vs tolerance (higher is better). *Uniformity* indicates how uniformly the representations are distributed and *tolerance* indicates how closely the representations from the same class are located in the embedding space. PACT (*red*-circled) produces higher uniformity while maintaining impressive tolerance.

In Wang and Liu (2021), the authors find that both uniformity and tolerance are the significant properties in contrastive learning. Wang and Isola (2020) show that the contrastive loss can be disentangled into two parts, which encourages the positive features to be aligned and the representations to match a uniform distribution in a hypersphere. Therefore, we employ the uniformity metric with gaussian potential kernel proposed by Wang and Isola (2020); Wang and Liu (2021),

$$L_{uniformity} = \log \mathop{\mathbf{E}}_{x_i, x_j} \mathop{\mathbf{E}}_{\sim p_{data}} \left[e^{-t || f(x_i) - f(x_j) ||_2^2} \right]$$

Where, x_i and x_j are two different examples and f(.) is the model encoder. Following Gao et al. (2021); Zhang et al. (2022b), we set t = 2.

On the contrary, we measure the tolerance using the mean similarities of samples belonging to the same class formulated as,

$$L_{tolerance} = \mathop{\mathbf{E}}_{x_i, x_j \sim p_{data}} \left[(f(x_i)^T f(x_j)) \cdot I_{l(x_i) = l(x_j)} \right]$$

Where $l(x_i)$ is the class of example x_i and I is a binary indicator function.

Ideally, models are expected to project the representations uniformly distributed in the embedding space and at the same time representations of the same class as closely as possible. We compute $L_{uniformity}$ and $L_{tolerance}$ of the models taking samples from two single sentence classification tasks (*SST-2* and *NC*) and plot them in Figure 3. We observe that PACT achieves high uniformity while maintaining a good tolerance compared to *all*

	Sin	gle-Sent	ence				Sente	nce-Pai	ir		
	Cola	SST2	NC	MRPC	QQP	MNLI	QNLI	RTE	QAM	QADSM	PAWSX
BERT	52.1	<u>93.5</u>	-	88.9	89.2	<u>84.6</u>	90.5	66.4	-	-	-
BERT	52.1	92.7	92.8	88.2	89.2	84.4	90.6	66.9	69.0	71.5	93.3
BERT-PT	51.3	91.9	92.6	87.6	89.1	83.4	90.4	65.6	67.8	71.5	93.7
TaCL ‡	52.4	92.3	-	<u>90.8</u>	-	84.4	91.1	62.8	-	-	-
TaCL	52.4	91.8	92.7	87.9	89.1	84.1	91.2	65.9	68.7	70.9	93.4
SimCSE	50.9	92.9	92.7	87.8	89.2	84.2	90.4	64.2	68.8	71.1	93.2
Mirror-BERT	52.8	92.3	92.8	86.8	89.2	84.2	90.9	66.5	68.9	72.2	93.5
SCD	52.0	92.2	92.8	86.9	89.1	84.3	90.2	65.3	68.8	71.4	92.7
DiffCSE	50.3	92.7	92.7	88.1	89.1	84.2	91.1	64.4	69.0	71.7	92.9
PACT	53.1	93.2*	93.1	89.1	89.2 *	84.2*	91.4 *	67.1	69.0	72.5	93.8 *

Table 2: Performance of the models on single-sentence and sentence-pair classification tasks. We evaluate *MRPC* with F₁-score, *CoLA* with Matthew's correlation and the others with accuracy. ||: published in Devlin et al. (2019) and \ddagger : published in Su et al. (2022). Best performance in our experiments that outperforms the results in published literature is highlighted in *bold*. Best performance in our experiments that does not outperform the published results is highlighted in *red*. Best performance in published works that also outperform our own experiments is highlighted in *underline*. \clubsuit indicates statistically significant result in *t-test* with *p* < 0.05.

	NER	POS
BERT	90.80	96.76
BERT-PT	90.64	96.64
TaCL	91.18	96.70
SimCSE	91.18	96.67
Mirror-BERT	90.49	96.71
SCD	90.51	96.41
DiffCSE	90.54	96.61
PACT	91.24 *	96.85

Table 3: Performance of the models on token-level classification tasks. Following Liang et al. (2020), we evaluate *NER* with F₁-score and *POS* with accuracy. Best performance is highlighted in *bold*. \clubsuit indicates statistically significant result in *t-test* with *p* < 0.05.

other models.³ The uniformity achieved by PACT potentially stems from contrasting with the hard adversarial sentence-level representations during the pretraining. As a result, PACT achieves more discriminative capability and reduces anisotropy in the embedding space.

6.2 Token-level Uniformity

Following Su et al. (2022), we conduct a qualitative experiment by visualizing the similarity among the tokens (Figure 4). We pass an example sentence to BERT (Figure 4a), TaCL (Figure 4b), and PACT (Figure 4c) to compute the cosine similarity between every two tokens. We observe that similarity along the diagonal is the highest for all the models because of the self-similarity. However, TaCL and PACT produce lower similarity scores along the off-diagonal compared to BERT. In fact, it is

noticeable that PACT even produces better distinguishable token representations than TaCL in some areas (red rectangular portion). This indicates that PACT produces token-level discriminative representation, like TaCL, resulting in an isotropic distribution in the embedding space.

6.3 Data Imbalance

Since real-world datasets are usually imbalanced (Cao et al., 2019; Bao et al., 2020), we study how PACT performs on imbalanced scenarios. Following Cao et al. (2019) and Zhang et al. (2022c), we construct imbalanced classification training datasets with different imbalance degrees, $\rho = |class_{max}| / |class_{min}|$, where $|class_{max}|$, $|class_{min}|$ denotes the number of samples in the maximum and the minimum class respectively. We conduct this experiment on three binary classification datasets where we sample $|class_{max}| / \rho$ number of the minimum class for $\rho = 2, 3, 4, 5, 10$. As can be seen from Table 4, performance decreases almost monotonically for all the models as ρ increases. However, PACT generally maintains a preferable performance compared to other methods. We conjecture that the higher sequence-level uniformity helps PACT to generate more discriminative representations, which make it easier to draw a boundary between two classes even with fewer examples from one class, resulting in an enhanced capability of the model to differentiate the classes.

7 Label-Wise Similarity Distribution

We conduct an experiment to analyze the similarity distribution for different labels. Specifically, we

³While SCD produces better uniformity than PACT, SCD performs worst in terms of tolerance.

	PAWSX					QADSM					QAM				
ρ	2	3	4	5	10	2	3	4	5	10	2	3	4	5	10
BERT	91.9	89.6	89.3	88.4	82.9	68.2	63.5	63.5	52.8	53.0	67.	1 64.7	61.8	60.6	55.3
BERT-PT	92.1	91.6	91.1	88.4	83.7	66.4	62.7	53.9	52.4	52.3	64.	9 62.1	60.5	57.5	53.6
TaCL	88.6	86.6	86.2	83.2	76.9	66.9	63.5	54.4	52.9	52.8	65.	5 64.5	62.4	61.3	56.1
SimCSE	92.1	91.9	91.9	87.3	81.3	68.	61.4	55.3	54.1	52.8	66.	6 63.4	60.8	59.9	54.9
Mirror-BERT	91.6	91.4	91.4	86.5	82.6	67.9	63.6	56.9	54.1	53.1	66.	1 64.8	62.6	60.8	54.4
SCD	92.2	90.6	88.9	88.4	79.7	66.9	61.9	61.9	53.1	52.8	67.	2 65.3	62.2	60.8	57.2
DiffCSE	92.3	90.6	87.3	86.7	78.9	68.4	64.6	59.1	54.1	53.2	67.	2 65.1	62.8	60.6	56.9
PACT	92.5	90.1	89.9	88.6	83.7	68.4	64.0	55.5	54.2	53.2	67.	7 65.4	63.5	61.3	57.8

Table 4: Performance of the models in data imbalance settings. Best performances are highlighted in *bold*.

4e+0 2e+0

8e+0

6e+0

4e+0 2e+0

1c+0

8e+0

6e+0 4e+0

1e+0+ 1e+0+ 8e+0

6e+0 4e+0 2e+0

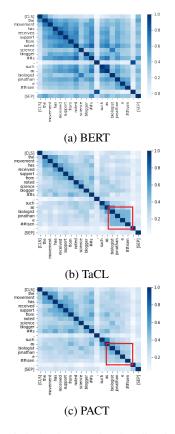
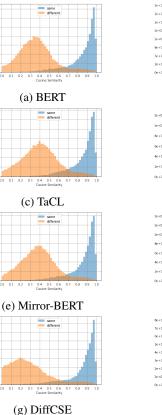


Figure 4: Self-similarity matrix visualization for token representations. *Red*-rectangle indicates the area where PACT produces more discriminative representation than TaCL.

analyze how the similarity distribution of representations coming from the same class differ from the similarity distribution coming from different classes. For this purpose, we take the models finetuned on *NC* and plot the cosine similarity score for every pair of examples in the evaluation set. Figure 5 shows the similarity distribution for different models. We denote the similarity score for representations of the same class in *blue* and representations of the different classes in *orange*. As the Figure shows, PACT pushes the distributions to the opposite poles better than other models.



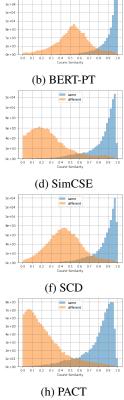


Figure 5: Label-wise similarity distribution of the models. *blue* distribution indicates cosine similarity for representations of the same label and *orange* distribution indicates cosine similarity for representations of the different labels.

We further quantify the distributions by computing Earth Mover's Distance (EMD) (Rubner et al., 2000; Ramdas et al., 2017) score. We present the result in Table 5. Ideally, the distribution of same class and the distribution of different classes (second column) should be well-apart (higher EMD). Moreover, the distribution of the same class (third column) should be close to one (lower EMD) while the distribution of the different classes (fourth column) should be close to zero (lower EMD), respec-

	<i>Same ∼ Diff.</i> ↑	Same \sim 1.0 \downarrow	<i>Diff.</i> \sim 0.0 \downarrow
BERT	0.488	0.145	0.367
BERT-PT	0.383	0.095	0.522
TaCL	0.447	0.143	0.409
SimCSE	0.556	0.158	0.280
Mirror-BERT	0.468	0.141	0.389
SCD	0.429	0.128	0.442
DiffCSE	0.555	0.145	0.296
PACT	0.557	0.206	0.237

Table 5: EMD scores of the models for Figure 5. Second column indicates EMD between the two distributions. Third column indicates EMD between the distribution of *same* class and 1.0. Fourth column indicates EMD between the distribution of *different* class and 0.0. \uparrow indicates higher is better and \downarrow indicates lower is better.

tively. We observe that PACT differentiates the two distributions with higher EMD, by pushing them to opposite directions, corroborating Figure 5. Although EMD score of the second column is higher for PACT compared to some other models, PACT achieves the lowest EMD in the third column. Overall, as PACT achieves higher sentence-level uniformity as a result of being pretrained on adversarial hard negatives, it has more discriminative representations for the different classes. This results in a lower similarity and EMD score across the different classes.

We further conduct experiments to analyze the representation transferability of PACT, which we present in Appendix B.

8 Ablation Study

	NC		PAWSX		NER	
PACT	92.93		93.27		94.93	
- $\mathcal{L}_{adv-Sequence}$	92.72	(-0.21)	92.83	(-0.44)	94.88	(-0.05)
- $\mathcal{L}_{adv-MLM}$	92.82	(-0.11)	93.14	(-0.13)	94.70	(-0.23)
$-\mathcal{L}_{MLM-CL}$	92.83	(-0.10)	93.21	(-0.06)	94.79	(-0.14)
$-\mathcal{L}_T$	92.89	(-0.04)	93.18	(-0.09)	94.81	(-0.12)

Table 6: Ablation study on the contribution of proposed $\mathcal{L}_{adv-MLM}$ and $\mathcal{L}_{adv-Sequence}$ losses.

We conduct ablation studies to analyze the efficacy of our proposed two losses, $\mathcal{L}_{adv-MLM}$ and $\mathcal{L}_{adv-Sequence}$. For this purpose, we experiment on the validation sets of one single-sentence (*NC*), one pair-sentence (*PAWSX*), and one token-level (*NER*) datasets and report performance in Table 6.

As we observe, removing $\mathcal{L}_{adv-MLM}$ loss hurts performance on the *NER* dataset more than the other two datasets. This shows that $\mathcal{L}_{adv-MLM}$ contributes mostly for the token-level tasks. On the other hand, if we remove $\mathcal{L}_{adv-Sequence}$ loss, performance drops mostly on *NC* and *PAWSX* datasets, indicating its contribution to sentence-level tasks. Overall, performance degrades on all the datasets, if we remove any of these two losses, which highlights the positive contribution of each of the two losses.

9 Conclusion

We propose PACT, a contrastive learning selfsupervised framework for jointly optimizing tokenand sentence-level representations. We introduced adversarial MLM and Sequence objectives to mine adversarial hard negative samples close to the anchor representations in the embedding space. Our evaluation over 13 different tasks show that PACT achieves consistent improvements over the SOTA baselines. We further show that PACT exhibits better token- and sentence-level uniformity that alleviate the issue of anisotropy in PLMs.

10 Limitations

Although PACT improves over SOTA baselines on token-level and sentence-level classification tasks, we empirically find that it exhibits subpar performance on semantic text similarity (STS) tasks (Table D.1). We hypothesize the reason is that PACT does not explicitly attempt to align positive representations. This is in contrast to other self-supervised methods such as those based on exploiting dropout and back-translation that intrinsically learn another view of each data point (as these would function as the positive pairs), hence benefitting STS tasks. It is to be noted that our work is focused on text classification, while STS tasks are focused on generating similarity scores between two sentences (as opposed to text classification). Therefore, STS tasks are out of the scope of this work. Another limitation is related to the pretraining time and resources: pretraining PACT requires three BERT models, which costs additional GPU resources. However, we only pretrain the student-BERT and the adv-BERT, while keeping the parameters of the teacher-BERT fixed. To put this in perspective, the additional pretraining steps of the student-BERT and the adv-BERT (150K steps each) are still significantly lower than the original BERT (1M steps). Moreover, this pretraining is a one-time execution and after that we only use the student-BERT for the downstream tasks. Finally, we outline a series of negative results in Appendix D. We hope these negative results will spur further research in this area.

References

- 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.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural networkbased collaborative filtering. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, page 767–776, New York, NY, USA. Association for Computing Machinery.
- Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pretraining 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.
- 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.
- 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.
- Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 528–537, Hong Kong, China. 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.
- Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.
- 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.
- Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735–1742. IEEE.

- Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. 2021. Adco: Adversarial contrast for efficient learning of unsupervised representations from selftrained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083.
- Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. 2020. Robust pre-training by adversarial contrastive learning. *Advances in neural information processing systems*, 33:16199–16210.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for opendomain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.
- 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*, volume 33, pages 18661–18673. Curran Associates, Inc.
- Tassilo Klein and Moin Nabi. 2022. SCD: Selfcontrastive decorrelation of sentence embeddings. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 394–400, Dublin, Ireland. Association for Computational Linguistics.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.
- Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.
- 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.

- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6008–6018, Online. Association for Computational Linguistics.
- Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. 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.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, volume abs/1803.02893.
- Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114.
- Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semisupervised text classification. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.
- Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.
- Yao Qiu, Jinchao Zhang, and Jie Zhou. 2021. Improving gradient-based adversarial training for text classification by contrastive learning and auto-encoder. 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 1698–1707. Association for Computational Linguistics.

- Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. 2020. Coda: Contrastenhanced and diversity-promoting data augmentation for natural language understanding. *arXiv preprint arXiv:2010.08670*.
- Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. 2017. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47.
- Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.
- Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121.
- Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2022. TaCL: Improving BERT pre-training with token-aware contrastive learning. In *Findings of the Association* for Computational Linguistics: NAACL 2022, pages 2497–2507, Seattle, United States. Association for Computational Linguistics.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.
- Varsha Suresh and Desmond 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, pages 4381–4394, Online and Punta Cana, Dominican Republic. 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.
- 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 (Volume 1: Long Papers), pages 2332–2342, Online. Association for Computational Linguistics.

- Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *IEEE Conference* on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 2495–2504. Computer Vision Foundation / IEEE.
- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.
- Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: contrastive learning for sentence representation. *CoRR*, abs/2012.15466.
- 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.
- Qiying Yu, Jieming Lou, Xianyuan Zhan, Qizhang Li, Wangmeng Zuo, Yang Liu, and Jingjing Liu. 2022. Adversarial contrastive learning via asymmetric infonce. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V, pages 53–69. Springer.
- Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Axi Niu, Jiu Feng, Chang D Yoo, and In So Kweon. 2022a. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX, pages 725–742. Springer.
- Chiyu Zhang, Muhammad Abdul-Mageed, and Ganesh Jawahar. 2022b. Infodcl: A distantly supervised contrastive learning framework for social meaning. *arXiv preprint arXiv:2203.07648*.
- Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022c. Label anchored contrastive learning for language understanding. In *Proceedings of*

the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1437–1449, Seattle, United States. Association for Computational Linguistics.

Appendices

A Implementation Details

We use the pretrained BERT-base from Huggingface (Wolf et al., 2020) as the backbone architecture. Following Su et al. (2022), we pretrain PACT on Wikipedia 150k steps with 10% of the total optimization steps for warm-up. During the pretraining, we set the learning rate to 1e-4 with a batch size of 256 on 4 Nvidia 40GB GPUs. As evident from Table 2 and Table 3, PACT improves the performance over the baselines without incorporating additional individual weights for each of the loss terms in the final objective function (Eqn. 1). Incorporating such weights could have further improved the performance of PACT on individual downstream tasks. However, we opt out from including such weighting hyperparameters for three reasons. First, we focus on the practical scenarios where searching for the optimal pretrained hyperparameters for each task is not a feasible option due to the computational cost of pretraining. Second, our main goal is to offer a method that is easy to deploy in the real world in that it can work well on a diverse range of downstream tasks. Third, we wanted to have fair comparisons to our baseline methods as not all of these search for the optimal values of the pretraining hyperparameters (Su et al., 2022; Klein and Nabi, 2022). Nevertheless, search for best pretraining hyperparamters can be investigated in the future.

During the finetuning on downstream tasks, we run CoLA, SST-2, and MRPC for 20 epochs and the others for 10 epochs. We set the batch size to 32, maximum sequence length to 256, and use the AdamW optimizer with initial learning rate as {5e-6, 1e-5, 2e-5, 3e-5, 4e-5, 5e-5} with linear learning scheduler. We choose the best model on the Dev set for reporting on the test set. Following the standard protocol, we use Matthew's correlation for CoLA, F1-score for MRPC and NER, and accuracy for other datasets as the evaluation metrics. For each task, we run the experiments three times with different random seeds and report the average score. We further conduct statistical significant test for PACT using *t-test* against finetuned BERT with *p*-value < 0.05.

B Finetuned Representation Transferability

Although we usually finetune a model on the same task the model is evaluated on, we were inquisitive about the transferability of our finetuned representations across tasks . To test this transferability, we select different sentence-pair classification datasets and study how the models perform when finetuned on one dataset and evaluated on another. Although tasks are different across these sentencepair datasets, the core idea behind all these tasks is to measure sentence-pair relevance. Hence, we hypothesize a good model should be able to generalize across tasks by performing favorably in the zero-shot setting.

Table B.1 shows performance of the models finetuned on one dataset (first part of the pair) when evaluated on another (second part of the pair). We see that PACT outperforms other models for most of the pair combinations. We observe that PACT produces uniform representations in the embedding space, which allows the representations to be more informative. As a result, the learned representations increase the generalization capability of PACT and help the model perform better across the tasks.

C Difference Between Adversarial Sample Generation in CV and NLP

In computer vision, Jiang et al. (2020) propose to create two different views: one with standard augmentation and another with adversarial perturbation to train with contrastive loss. Similarly, Zhang et al. (2022a) add adversarial perturbation to the images under l_{∞} to maximize robustness. Furthermore, Yu et al. (2022) consider generating positive and negative views of an original image by directly adding weighted perturbation.

The major difference of adversarial sample generation in computer vision (CV) works and NLP works discussed in Section 2 is that while CV works focus on directly perturbating images in the continuous space, due to the discrete nature of text, NLP works primarily perturb on token-level. Perturbing a whole image can create an additional view of the anchor image, however, NLP works perturb each token separately. To get a sentence-level adversarial representation, some works (Pan et al., 2022) perturb the representational token ([CLS]) to get augmented view of the anchor sentence. Finally, although adding a small perturbation in an anchor image can produce an adversarial view, while ad-

	QADSM-QNLI	QNLI-QADSM	QADSM-MRPC	MRPC-QADSM	MRPC-PAWSX	PAWSX-MRPC	QNLI-QAM	QAM-QNLI
BERT	63.1	52.8	69.3	49.7	45.4	69.1	61.6	78
BERT-PT	63.9	53.9	69.1	49.6	45.3	67.1	61.5	77.5
TaCL	61.8	55.6	69.6	48.8	45.4	68.6	60.9	78.2
SimCSE	63.4	53.1	69.6	49.8	45.4	68.3	61.8	77.6
Mirror-BERT	64.7	55.3	69.4	49.9	45.5	65.2	60.9	77.3
SCD	65.5	52.1	66.7	49.5	45.3	63.7	61.5	77.3
DiffCSE	63.4	52.6	69.9	50.1	45.5	67.6	62.5	78.0
PACT	65.5	52.8	71.3	49.8	45.6	71.6	61.8	78.5

Table B.1: Performance of the models for finetuned representation transferability. For *GLUE* datasets, we evaluate on the validation sets. Best performances are highlighted in *bold*.

versaries can be obtained in NLP by changing the words(e.g., with antonyms) in a sequence (Wang et al., 2021) instead of continuous perturbation.

D Negative Results

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
SimCSE	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
DiffCSE	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
SCD	66.94	78.03	69.89	78.73	76.23	76.30	73.18	74.19
PACT	38.63	56.76	42.74	59.28	60.88	51.34	61.52	53.02

Table D.1: Performance on STS tasks (Spearman's correlation) for different models. PACT exhibits subpar performance on STS tasks.

-	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
SimCSE	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
PACT	38.63	56.76	42.74	59.28	60.88	51.34	61.52	53.02
PACT-dropout	65.23	77.31	68.09	78.57	75.17	74.58	69.34	72.61

Table D.2:Performance comparison of PACT andPACT-dropout on STS tasks.

In this section, we outline a series of experiments that did not exhibit promising results:

- 1. To improve tolerance, we added another probabilistic pretraining objective that teaches the model whether two segments of a sequence are the same. Motivated by SimCSE (Gao et al., 2021), we therefore, incorporated a dropout-based augmentation to align the positive examples. Although this objective indeed improves performance for some tasks such as paraphrase detection (*PAWSX*) and semantic text similarity (Table D.2), it results in inferior performance in other classification tasks.
- 2. We attempted to add phrase-level CL on top of token- and sentence-level CL in a selfsupervised manner. For this purpose, we computed point-wise mutual information (PMI) to collect frequent bigram, trigram, and quadgram from Wikipedia instead of masking random spans. We considered sentences containing the same phrases as positive pairs, however, this new phrase-level objective did not

improve performance. We conjecture that this objective is contradicting the token-level CL objective. That is, at the phrase-level, CL pulls tokens belonging to the same phrase together, while at the token-level, CL pushes non-identical tokens apart.

3. We further experimented on PACT's efficacy on low-resource data setting. Particularly, we sampled 10%, 25%, and 50% data from each class for multiple datasets to evaluate PACT, but it exhibited inferior performance. This can be potentially attributed to high uniformity of PACT. In low resource setting, we need to pull representations from the same class close together with limited data. Since PACT already distributes the representations uniformly in the embedding space, it makes it harder for PACT to pull them together with fewer training data.

E Ethics Statement

E.1 Data Collection and Release.

We collect pretraining data from Wikipedia for academic research purpose. The code to collect the data is publicly available. We will also share the dataset we used for pretraining upon request. For the downstream tasks, we use 13 benchmark datasets from GLUE and XGLUE (Table 1). To ensure proper credit assignment, we refer users to the original publications. We use the same train, dev, and test splits provided by the benchmark datasets.

E.2 Intended Use.

The intended use of PACT is for the text classification tasks. We aim to help researchers to pretrain the models with adversarially hard negative examples in the self-supervised setting. PACT can also be used for achieving better tokenand sentence-level uniformity, thus alleviating the anisotropy in PLMs.

E.3 Potential Misuse and Bias.

Some of the pretraining data may contain potential harmful and biased contents. For these reasons, we recommend that PACT not be used for research or in applications without careful prior consideration of potential misuse and bias.