TOWARD ADVERSARIAL TRAINING ON CONTEXTUALIZED LANGUAGE REPRESENTATION

Hongqiu Wu[1,2] **& Yongxiang Liu**[1,2] **& Hanwen Shi**[1,2] **& Hai Zhao**[1,2,*]**& Min Zhang**[3]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]School of Computer Science and Technology, Soochow University, Suzhou, China
{wuhongqiu,sam.liu,shihanwen}@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn,
minzhang@suda.edu.cn

ABSTRACT

Beyond the success story of adversarial training (AT) in the recent text domain on top of pre-trained language models (PLMs), our empirical study showcases the inconsistent gains from AT on some tasks, e.g. commonsense reasoning, named entity recognition. This paper investigates AT from the perspective of the contextualized language representation outputted by PLM encoders. We find the current AT attacks lean to generate sub-optimal adversarial examples that can fool the decoder part but have a minor effect on the encoder. However, we find it necessary to effectively deviate the latter one to allow AT to gain. Based on the observation, we propose simple yet effective *Contextualized representation-Adversarial Training* (CreAT), in which the attack is explicitly optimized to deviate the contextualized representation of the encoder. It allows a global optimization of adversarial examples that can fool the entire model. We also find CreAT gives rise to a better direction to optimize the adversarial examples, to let them less sensitive to hyperparameters. Compared to AT, CreAT produces consistent performance gains on a wider range of tasks and is proven to be more effective for language pretraining where only the encoder part is kept for downstream tasks. We achieve the new state-of-the-art performances on a series of challenging benchmarks, e.g. AdvGLUE (59.1 → 61.1), HellaSWAG (93.0 → 94.9), ANLI (68.1 → 69.3).

1 INTRODUCTION

Adversarial training (AT) (Goodfellow et al., 2015) is designed to improve network robustness, in which the network is trained to withstand small but malicious perturbations while making correct predictions. In the text domain, recent studies (Zhu et al., 2020; Jiang et al., 2020) show that AT can be well-deployed on pre-trained language models (PLMs) (e.g. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)) and produce impressive performance gains on a number of natural language understanding (NLU) benchmarks (e.g. sentiment analysis, QA).

However, there remains a concern as to whether it is the adversarial examples that facilitate the model training. Some studies (Moyer et al., 2018; Aghajanyan et al., 2021) point out that a similar performance gain can be achieved when imposing random perturbations. To answer the question, we present comprehensive empirical results of AT on wider types of NLP tasks (e.g. reading comprehension, dialogue, commonsense reasoning, NER). It turns out the performances under AT are inconsistent across tasks. On some tasks, AT can appear mediocre or even harmful.

This paper investigates AT from the perspective of the *contextualized language representation*, obtained by the Transformer encoder (Vaswani et al., 2017). The background is that a PLM is typically composed of two parts, a Transformer-based encoder and a decoder. The decoder can vary from tasks (sometimes a linear classifier). Our study showcases that, the AT attack excels at fooling the

[*] Corresponding author; This paper was partially supported by Key Projects of National Natural Science Foundation of China (U1836222 and 61733011); https://github.com/gingasan/CreAT.

decoder, thus generating greater training risk, while may yield a minor impact on the encoder part. When this happens, AT can lead to poor results, or degenerate to random perturbation training (RPT) (Bishop, 1995).

Based on the observations, we are motivated that AT facilitates model training through robust representation learning. Those adversarial examples that effectively deviate the contextualized representation are the necessary driver for its success. To this end, we propose simple yet effective *Contextualized representation-Adversarial Training* (CreAT), to remedy the potential defect of AT. The CreAT attack is explicitly optimized to deviate the contextualized representation. The obtained “global” worst-case adversarial examples can fool the entire model.

CreAT contributes to a consistent fine-tuning improvement on a wider range of downstream tasks. We find that this new optimization direction of the attack causes more converged hyperparameters compared to AT. That means CreAT is less sensitive to the inner ascent steps and step sizes. Additionally, we always re-train the decoder from scratch and keep the PLM encoder weights for fine-tuning. As a direct result of that, CreAT is shown to be more effective for language pre-training. We apply CreAT to MLM-style pre-training and achieve the new state-of-the-art performances on a series of challenging benchmarks (e.g. AdvGLUE, HellaSWAG, ANLI).

2 Preliminaries

This section starts by reviewing the background of adversarial training and contextualized language representation, and then experiments are made to investigate the impact of adversarial perturbations on BERT from two perspectives: (1) output predictions and (2) contextualized language representation. The visualization results lead us to a potential correlation between them and the adversarial training gain.

2.1 Adversarial Training

Adversarial training (AT) (Goodfellow et al., 2015) improves model robustness by pulling close the perturbed model prediction and the target distribution (i.e. ground truth). We denote the output label as y and model parameters as θ, so that AT seeks to minimize the divergence (i.e. Kullback-Leibler divergence):

$$\min_{\theta} \mathcal{D}\left[q(y|\mathbf{x}), p(y|\mathbf{x}+\delta^*, \theta)\right] \tag{1}$$

where $q(y|\mathbf{x})$ refers to the target distribution while $p(y|\mathbf{x}+\delta^*, \theta)$ refers to the output distribution under a particular adversarial perturbation δ^*. The adversarial perturbation is defined as the worst-case perturbation which can be evaluated by maximizing the empirical risk of training:

$$\delta^* = \underset{\delta; \|\delta\|_F \leq \epsilon}{\arg\max} \mathcal{D}\left[q(y|\mathbf{x}), p(y|\mathbf{x}+\delta, \theta)\right] \tag{2}$$

where $\| \delta \|_F \leq \epsilon$ refers to the decision boundary (the Frobenius norm) restricting the magnitude of the perturbation.

In the text domain, a conventional philosophy is to impose adversarial perturbations to word embeddings instead of discrete input text (Miyato et al., 2017). It brings down the adversary from the high-dimensional word space to low-dimensional representation to facilitate optimization. As opposed to the image domain, where AT can greatly impair model performances (Madry et al., 2018; Xie et al., 2019), such bounded embedding perturbations are proven to be positive for text models (Jiang et al., 2020; Zhu et al., 2020; Wang et al., 2021a). However, there is no conclusion as to where such a gain comes from.

2.2 Contextualized Representation and Transformer Encoder

Transformer (Vaswani et al., 2017) has been broadly chosen as the fundamental encoder of PLMs like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XL-Net (Yang et al., 2019), DeBERTa (He et al., 2021). The Transformer encoder is a stacked-up language understanding system with a number of self-attention and feed-forward sub-layers, which continuously place each token into the context and achieve its corresponding contextualized representation as its output.

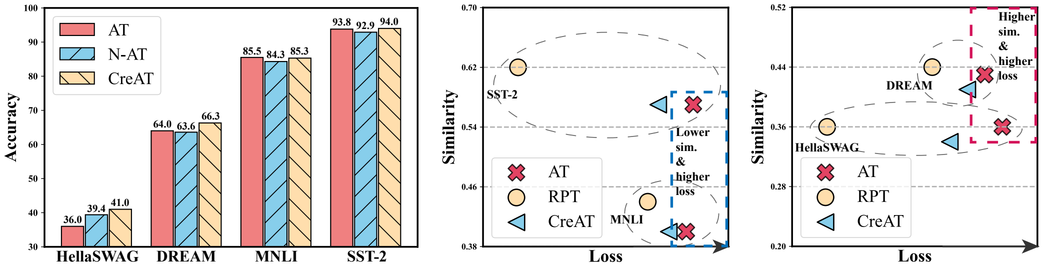

Figure 1: **Left:** Performances across different training methods, where N-AT refers to the regular situation where no adversarial training is applied. **Middle & right:** The relationship between the embedding similarity (contextualized representation) and training loss on different types of tasks. Note that we shift the training loss along the horizontal axis for a better view, the magnitude of which increases along the axis. We also investigate the attack success rate of AT in Appendix B.

One PLM will be deployed in two stages. For pre-training, the encoder will be trained on a large-scale corpus for a long period. For fine-tuning, the pre-trained encoder will be attached with a task-specific classifier as the decoder, which is initialized from scratch (we do not distinguish between the classifier and decoder in what follows). Eventually, the contextualized representation will be re-updated toward the downstream task.

2.3 Impact of Adversarial Training on PLMs

To access the impact of AT, we choose four types of tasks and fine-tune BERT (Devlin et al., 2019) on them respectively: (1) sentiment analysis (SST-2), (2) natural language inference (MNLI), (3) dialogue comprehension (DREAM), (4) commonsense reasoning (HellaSWAG). Among them, the first two tasks are relatively simple, while the last two are much more challenging. We summarize the test accuracies across different training methods in Figure 1 (left).

We calculate two statistical indicators for each training process. The first is the training loss. A higher training loss indicates that the model prediction is more deviated from the ground truth after the perturbation. The second is the cosine similarity of the output embeddings before and after the perturbation. Researchers always take the final output of the encoder (i.e. output embeddings) as the representative of contextual language representation (Li et al., 2020a; Gao et al., 2021). Specifically, we consider the lower bound of all the tokens in the sentence. A lower similarity indicates that the encoder is more disturbed by the perturbation. Note that the indicators here are averaged over the first 20% of the training steps, as the model will become more robust under AT. We depict the above two indicators in Figure 1 (middle and right). The discussion is below. We also depict random perturbation training (RPT) and CreAT (will be presented later) as references here for better comparison with AT. We will mention this figure again in the next section.

• **Observation 1: The gain from AT is inconsistent on the four tasks.** From Figure 1 (left), we can see that AT achieves nice results on two sentence-level classification tasks (0.9 and 1.2 points absolute gain on MNLI and SST-2). On the other two tasks, AT performs badly, even leading to a 3.4 points drop on HellaSWAG.

• **Observation 2: AT contributes the most to deviating model predictions.** From Figure 1 (middle and right), we can see that the training loss caused by AT is always the highest (always on the rightmost side of the graph).

• **Observation 3: AT weakly influences the contextualized representation on DREAM and HellaSWAG.** We can see that the resultant similarities of the output embeddings caused by AT are also both high (equivalent to RPT on HellaSWAG and slightly lower than RPT on DREAM). It suggests that the adversarial examples are gentle for the encoder, although they are malicious enough to fool the classifier. However, we see an opposite situation on SST-2 and MNLI, where the similarities caused by AT are much lower.

The above observations raise a question on AT in the text domain: **whether fooling model predictions is purely correlated with training performances of PLMs?** Grouping observation 1 and observation 2, we see that high training loss does not always lead to better performances (i.e. accuracies). Grouping observation 2 and observation 3, we can see that AT mainly fools the classifier on some of the tasks, but takes little impact on the encoder.

Summing all observations up together, we can derive that **the performance gain from AT on PLMs is necessarily driven by the fact that the model needs to keep its contextualized representation robust from any adversarial perturbation so that its performance is enhanced**. However, this favour can be buried on tasks such as reading comprehension and reasoning, where AT is weakly-effective in deviating the contextualized representation (comparable to RPT). The resultant classifier becomes robust, while the encoder is barely enhanced.

3 Contextualized representation-Adversarial Training

In this section, we propose *Contextualized representation-Adversarial Training* (CreAT) to effectively remedy the potential defect of AT.

3.1 Deviate Contextualized Representation

We let $h(\mathbf{x}, \theta)$ be the contextualized representation (i.e. the final output of the encoder), where θ refers to the model parameters and $\mathbf{x}$ refers to the input embeddings. Similarly, we let $h(\mathbf{x} + \delta, \theta)$ be the contextualized representation after $\mathbf{x}$ is perturbed by δ.

Our desired direction of the perturbation is to push away the two output states, i.e. decreasing their similarity. We leverage the cosine similarity to measure the angle of deviation between two word vectors (Mikolov et al., 2013; Gao et al., 2021). Thus, deviating the contextualized representation is the same as:

$$\min_{\delta} \mathcal{S}\left[h(\mathbf{x}, \theta), h(\mathbf{x} + \delta, \theta)\right] \tag{3}$$

where $\mathcal{S}[a, b] = \frac{a \cdot b}{\|a\| \cdot \|b\|}$.

3.2 CreAT

CreAT seeks to find the worst-case perturbation δ^* which deviates both the output distribution and contextualized representation, which can be formulated as:

$$\delta^* = \arg\max_{\delta; \|\delta\|_F \leq \epsilon} \mathcal{D}\left[q(y|\mathbf{x}), p(y|\mathbf{x} + \delta, \theta)\right] - \tau \mathcal{S}\left[h(\mathbf{x}, \theta), h(\mathbf{x} + \delta, \theta)\right] \tag{4}$$

where τ is the temperature coefficient to control the strength of the attacker on the contextual representation and a larger τ means that the attacker will focus more on fooling the encoder. CreAT is identical to AT when $\tau = 0$.

Why CreAT? In the previous discussion, AT is found to lead to the local worst case that partially fools the decoder part of the model. CreAT can be regarded as the "global" form of AT, which solves the global worst case for the entire model (both the encoder and decoder). The encoder is trained to be robust with its contextualized representation so that the model can perform better.

Training with CreAT Different from commonly-used AT methods (Goodfellow et al., 2015; Zhang et al., 2019b; Wang et al., 2020), where the adversarial risk is treated as a regularizer to enhance the alignment between robustness and generalization, in this paper, we adopt a more straightforward method to combine the benign risk and adversarial risk (Laine & Aila, 2017; Wang & Wang, 2022). Given a task with labels, the training objective is as follows:

$$\min_{\theta} \lambda \mathcal{L}(\mathbf{x}, y, \theta) + (1 - \lambda) \mathcal{L}(\mathbf{x} + \delta^*, y, \theta) \tag{5}$$

where $\mathcal{L}$ is the task-specific loss function and the two terms refer to the training loss under the respective benign example $\mathbf{x}$ and adversarial example $\mathbf{x} + \delta$. We find that $\lambda = 0.5$ performs just well

Algorithm 1 Contextualized representation-Adversarial Training

Input: Model θ, training set T, model step size β, ascent step size α, decision boundary ϵ, number of ascent steps k, temperature coefficient τ

```
 1: while not converged do
 2:     $\{\mathbf{x}, y\} \leftarrow \mathrm{SampleBatch}(T)$
 3:     $\delta_0 \leftarrow \mathrm{Init}()$
 4:     $\mathrm{Forward}(\mathbf{x}, \theta)$                               ▷ Go benign forward pass
 5:     for $j = 1$ to $k$ do
 6:         $\mathrm{Forward}(\mathbf{x} + \delta_{j-1}, \theta)$                    ▷ Go adversarial forward pass
 7:         $\delta_j \leftarrow \mathrm{BackwardUpdate}(\delta_{j-1}, y, \tau, \alpha, \epsilon)$   ▷ Update the perturbation following Eq. 4
 8:     end for
 9:     $\mathrm{Forward}(\mathbf{x} + \delta^*, \theta)$                            ▷ $\delta^* \leftarrow \delta_k$
10:     $\theta \leftarrow \mathrm{BackwardUpdate}(\theta, y, \beta)$            ▷ Update the model parameters following Eq. 5
11: end while
```

Table 1: Results across different tasks over five runs. The average numbers are calculated based on the right side of the table (5 tasks). The variances for all tasks except WNUT and DREAM are low (< 0.3), so we omit them. Our CreAT-trained DeBERTa achieved the state-of-the-art result (94.9) on HellaSWAG (H-SWAG) on May 5, 2022[1].

	MNLI-m (Acc)	QQP (F1)	WNUT (F1)	DREAM (Acc)	H-SWAG (Acc)	AlphaNLI (Acc)	RACE (Acc)	Avg	Gain
$\text{BERT}_{\text{base}}$	84.3	71.6	$48.6_{0.8}$	$63.0_{0.9}$	39.4	65.2	65.3	56.3	-
+ *FreeLB*	85.5	**73.1**	$48.2_{1.3}$	$64.1_{0.9}$	39.8	65.3	62.8	56.4	↑ 0.1
+ *SMART*	**85.6**	72.7	$48.8_{0.8}$	$64.5_{0.6}$	39.2	65.1	63.3	56.2	↓ 0.1
+ *AT*	85.2	72.9	$49.7_{1.2}$	$63.8_{0.6}$	36.0	64.9	63.0	55.5	↓ 0.8
+ *CreAT*	85.3	73.0	$\mathbf{49.9}_{0.9}$	$\mathbf{66.0}_{0.7}$	**40.5**	**67.0**	**68.0**	**58.3**	↑ **2.0**

in our experiments. Eq. 5 is agnostic to both pre-training (e.g. masked language modeling $\mathcal{L}_{\text{mlm}}$) and fine-tuning (e.g. named entity recognition $\mathcal{L}_{\text{ner}}$) of PLMs.

Algorithm 1 summarizes the pseudocode of CreAT. The inner optimization is based on projected gradient descent (PGD) (Madry et al., 2018). At each training step, which corresponds to the outer loop (line 2 $\sim$ line 10), we fetch the training examples and initialize the perturbation δ_0. In the following inner loop (line 5 $\sim$ line 7), we iterate to evaluate δ^* by taking multiple projected gradient steps. At the end of the inner loop, we obtain the adversarial perturbation $\delta^* = \delta_k$. Eventually, we train and optimize the model parameters with the adversarial examples (line 10).

4 Empirical Results

Our empirical results include both general and robust-learning tasks. The implementation is based on *transformers* (Wolf et al., 2020).

4.1 Setup

We conduct fine-tuning experiments on $\text{BERT}_{\text{base}}$ (Devlin et al., 2019). We impose the same bounded perturbation for all adversarial training methods (fix ϵ to 1e-1). We tune the ascent step size α from {1e-1, 1e-2, 1e-3} and the number of ascent steps k from {1, 2} following the settings in previous papers (Jiang et al., 2020; Liu et al., 2020).

We conduct MLM-style continual pre-training and obtain two language models: $\text{BERT}_{\text{base}}^{CreAT}$ based on $\text{BERT}_{\text{base}}$ (Devlin et al., 2019), $\text{DeBERTa}_{\text{large}}^{CreAT}$ based on $\text{DeBERTa}_{\text{large}}$ (He et al., 2021). For the training corpus, we use a subset (nearly 100GB) of C4 (Raffel et al., 2020). Every single model is trained with a batch size of 512 for 100K steps. For adversarial training, we fix α and ϵ to 1e-1,

[1] https://leaderboard.allenai.org/hellaswag/submissions/public

Table 2: Results on AdvGLUE. For dev sets (upper), we report the results over five runs and report the mean and variance for each. For test sets (bottom), the results are taken from the official leaderboard, where CreAT achieved the new state-of-the-art on March 16, 2022[2].

	SST-2	QQP	QNLI	MNLI-m	MNLI-mm	RTE	Avg
	(Acc)	(Acc/F1)	(Acc)	(Acc)	(Acc)	(Acc)	
BERT$_{\text{base}}$	$32.3_{1.4}$	$50.8_{2.1}$/-	$40.1_{0.6}$	$32.6_{1.3}$	$19.3_{0.8}$	$37.0_{2.5}$	35.4
FreeLB-BERT$_{\text{base}}$	$31.6_{0.3}$	$51.0_{2.4}$/-	**45.4**$_{0.3}$	$33.5_{0.4}$	$21.9_{0.9}$	$42.0_{1.5}$	37.6
BERT$_{\text{base}}^{\text{MLM}}$	$32.0_{0.6}$	$48.5_{1.3}$/-	$43.4_{1.4}$	$27.6_{0.4}$	$20.8_{0.5}$	**45.9**$_{2.5}$	36.4
BERT$_{\text{base}}^{CreAT}$	**35.3**$_{1.0}$	**51.5**$_{1.2}$/-	$44.8_{0.6}$	**36.0**$_{0.8}$	**22.0**$_{0.8}$	$45.2_{2.0}$	**39.1** ↑**2.7**
T5$_{\text{large}}$	60.6	63.0/**55.7**	57.6	48.4	39.0	62.8	55.3
SMART-RoBERTa$_{\text{large}}$	50.9	64.2/44.3	52.2	45.6	36.1	70.4	52.0
DeBERTa$_{\text{large}}$	57.9	60.4/48.0	57.9	**58.4**	**52.5**	**78.9**	59.1
DeBERTa$_{\text{large}}^{CreAT}$	**63.5**	**67.5**/54.5	**61.8**	56.7	51.7	72.0	**61.1** ↑**2.0**

and k to 1. Training a base/large-size model takes about 30/100 hours on 16 V100 GPUs with FP16. Readers may refer to Appendix A for hyperparameters details.

To distinguish different training settings, for example, CreAT-BERT$_{\text{base}}$ means we fine-tune the original BERT model with CreAT; BERT$_{\text{base}}^{CreAT}$ means we pre-train the model and then regularly fine-tune it without adversarial training; BERT$_{\text{base}}^{\text{MLM}}$ means we regularly pre-train and fine-tune the model.

We describe a number of adversarial training counterparts below.

• **FreeLB** (Zhu et al., 2020) is a state-of-the-art fast adversarial training algorithm, which incorporates every intermediate example into the backward pass.

• **SMART** (Jiang et al., 2020) is another state-of-the-art adversarial training algorithm, which keeps the local smoothness of model outputs.

• **AT** (i.e. vanilla AT) refers to the special case of CreAT when $\tau = 0$. Both AT and CreAT are traditional PGD attackers (Madry et al., 2018).

4.2 Results on Various Downstream Tasks

We experiment on MNLI (natural language inference), QQP (semantic similarity), two representative tasks in GLUE (Wang et al., 2019a); WNUT-17 (named entity recognition) (Aguilar et al., 2017); DREAM (dialogue comprehension) (Sun et al., 2019); HellaSWAG, AlphaNLI (commonsense reasoning) (Zellers et al., 2019; Bhagavatula et al., 2020); RACE (reading comprehension) (Lai et al., 2017).

Table 1 summarizes the results on BERT$_{\text{base}}$. First, each AT method works well on MNLI and QQP (two sentence classification tasks), and the improvement is quite similar. However, on the right side of the table, we can see AT appears harmful on certain tasks (average drop over BERT: 0.1 for SMART, 0.8 for AT), while CreAT consistently outweighs all the counterparts (absolute gain over BERT: 3.0 points on DREAM, 1.8 on AlphaNLI, 2.7 on RACE) and raises the average score by 2.0 points. Table 1 also verifies our previous results in Figure 1. From Figure 1, we can also see that the embedding similarities under CreAT are the lowest on all four tasks.

AT is sensitive to hyperparameters. For example, FreeLB obtains 36.6 and 39.8 on H-SWAG under different α, 1e-1 and 1e-3. However, we find that CreAT is less sensitive to the ascent step size α, and 1e-1 performs well on all datasets. In other words, the direction of the CreAT attack makes it easier to find the global worst case.

4.3 Adversarial GLUE

Adversarial GLUE (AdvGLUE) (Wang et al., 2021b) is a robustness evaluation benchmark derived from a number of GLUE (Wang et al., 2019a) sub-tasks, SST-2, QQP, MNLI, QNLI, and RTE.

[2] https://adversarialglue.github.io/

Different from GLUE, it incorporates a large number of adversarial samples in its dev and test sets, covering word-level, sentence-level, and human-crafted transformations.

Table 2 summarizes the results on all AdvGLUE sub-tasks. We see that CreAT-trained models consistently outperform fine-tuned ones by a large margin. In addition, we see that directly fine-tuning $\text{BERT}_{\text{base}}^{CreAT}$ works better than fine-tuning with FreeLB. To verify the effectiveness of CreAT-based adversarial pre-training, we pre-train $\text{BERT}_{\text{base}}^{\text{MLM}}$ based on MLM with the same number of steps (100K steps). From Table 2, we see that simply training longer leads to a limited robustness gain, while the CreAT-trained one is much more powerful (over $\text{BERT}_{\text{base}}^{\text{MLM}}$: 3.3 on SST-2, 6.2 on QQP, 8.4 on MNLI-m). Compared with other strong PLMs (Liu et al., 2019; Raffel et al., 2020), the further CreAT-based pre-training let $\text{DeBERTa}_{\text{large}}^{CreAT}$ achieves the new state-of-the-art result.

4.4 Adversarial SQuAD

AdvSQuAD (Jia & Liang, 2017) is a puzzling machine reading comprehension (MRC) dataset derived from SQuAD-1.0 (Rajpurkar et al., 2016), by inserting sentences into the paragraphs, replacing nouns and adjectives in the questions, generating fake answers similar to the original ones.

Table 3: Results on two AdvSQuAD dev sets over three runs.

	AddSent (EM/F1)	AddOneSent (EM/F1)
$\text{BERT}_{\text{base}}$	59.3$_{.4}$/65.7$_{.2}$	67.3$_{.3}$/74.3$_{.4}$
FreeLB-$\text{BERT}_{\text{base}}$	60.9$_{.6}$/67.4$_{.3}$ ↑1.6/1.7	67.9$_{.2}$/74.8$_{.6}$ ↑0.6/0.5
$\text{BERT}_{\text{base}}^{CreAT}$	**62.3**$_{.3}$/**69.5**$_{.2}$ ↑**3.0/3.8**	**69.5**$_{.2}$/**76.5**$_{.3}$ ↑**2.2/2.2**
$\text{DeBERTa}_{\text{large}}$	80.2$_{.2}$/86.5$_{.2}$	83.3$_{.5}$/89.1$_{.1}$
$\text{DeBERTa}_{\text{large}}^{CreAT}$	**81.6**$_{.1}$/**87.3**$_{.3}$ ↑**1.4/0.8**	**84.3**$_{.3}$/**90.2**$_{.2}$ ↑**1.0/1.1**

From Table 3, we see CreAT leads to impressive performance gains by 1 to 3 points on all metrics.

4.5 Adversarial NLI

Adversarial Natural Language Inference (ANLI) (Nie et al., 2020) is an iteratively-strengthened robustness benchmark. In each round, human annotators are asked to craft harder samples to fool the model. In our experiment, each model is trained on the concatenated training samples of ANLI and MNLI.

Table 4: Test results of all rounds on ANLI over five runs.

	Round 1	Round 2	Round 3	Avg
$\text{BERT}_{\text{base}}$	55.1$_{.4}$	44.5$_{.6}$	42.5$_{.7}$	47.4
$\text{BERT}_{\text{large}}$	57.6$_{.4}$	**48.0**$_{.5}$	43.2$_{.9}$	49.6
$\text{BERT}_{\text{base}}^{CreAT}$	**59.2**$_{.8}$ ↑**4.1**	45.9$_{.7}$ ↑**1.4**	**44.0**$_{1.1}$ ↑**1.5**	**49.7** ↑**2.1**
$\text{DeBERTa}_{\text{large}}$	77.8$_{.7}$	66.2$_{.2}$	60.4$_{.8}$	68.1
$\text{DeBERTa}_{\text{large}}^{CreAT}$	**78.7**$_{.7}$ ↑**0.9**	**66.9**$_{.1}$ ↑**0.7**	**62.3**$_{.5}$ ↑**1.9**	**69.3** ↑**1.1**
$\text{InfoBERT}_{\text{large}}$	75.5	51.4	49.8	58.9
$\text{ALUM}_{\text{large}}$	72.3	52.1	48.4	57.6

Table 4 summarizes the test results of all rounds. $\text{BERT}_{\text{base}}^{CreAT}$ is powerful and even surpasses $\text{BERT}_{\text{large}}$, outperforming by 1.6 points and 0.8 point on Round 1 and Round 3. Besides, DeBERTa significantly outperforms the previous state-of-the-art models InfoBERT (Wang et al., 2021a) and ALUM (Liu et al., 2020). However, $\text{DeBERTa}_{\text{large}}^{CreAT}$ further pushes the average score from 68.1 to 69.3, and especially on the most challenging Round 3, it leads to a 1.9 points gain on $\text{DeBERTa}_{\text{large}}$.

5 Analysis

5.1 Ablation Study

To access the gain from CreAT more clearly, in this part, we disentangle the CreAT perturbations level by level. To ensure the fairness of experiments, all the mentioned training methods follow the same optimization objectives in Eq. 4 and Eq. 5. Specifically, we first apply random perturbation training (RPT) on top of BERT fine-tuning. Then we apply AT, where the random perturbations become adversarial. CreAT^{-} refers to the case where the attack is solely optimized to deviate the contextualized representation (i.e. removing the left term in Eq. 4).

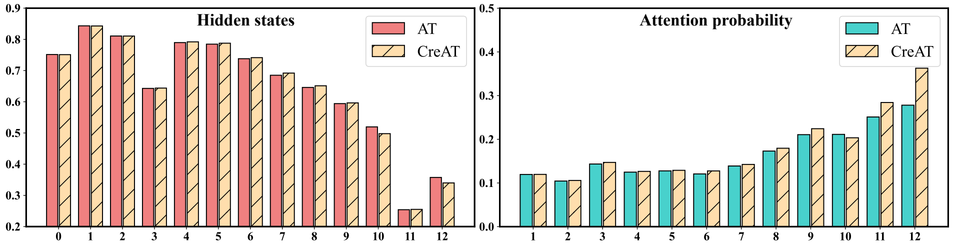

Figure 2: Impact of CreAT attack on the intermediate layers of BERT$_{\text{base}}$ (12 layers).

Table 5: Results of ablation study over three runs.

	MNLI-m	DREAM	AlphaNLI
BERT	84.3	63.0	65.2
+ *RPT*	84.1	63.6	64.9
+ *AT*	85.2	64.0	64.9
+ *CreAT*$^{-}$	83.5	65.0	65.7
+ *CreAT*	**85.3**	**66.4**	**67.1**

From Table 5, we can see that AT is superior to RPT on all three tasks. The improvement from CreAT is greater, outperforming AT by a large margin on DREAM and AlphaNLI. In addition, we can see CreAT^{-} still improves DREAM and AlphaNLI without the adversarial objective. It turns out both parts of the CreAT attack are necessary. CreAT acts as a supplement to AT to find the optimal adversarial examples. The interesting thing is that removing the adversarial objective causes a large drop on MNLI. It implies that the gain from AT sometimes does not come from increasing the training risk, which can vary from tasks or datasets. We will leave this part for the future work.

Table 6: Continual pre-training results on BERT$_{\text{base}}$ in different settings over fine runs. Each model is trained for the same number of steps.

	HellaSWAG	PAWS-QQP	PAWS-Wiki
MLM	42.6	88.5	92.0
+ *FreeLB*	42.7	88.2	92.1
+ *AT*	43.1	89.3	92.8
+ *CreAT*	**43.8**	**90.2**	**93.1**

Table 6 demonstrates a number of adversarial pre-training methods. We can see that CreAT obtains stronger performances than simply training MLM for a longer period on all three tasks. However, the FreeLB-based one is almost comparable to MLM, while the AT-based one is slightly better than it. To explain, we only keep the encoder and drop the MLM decoder when fine-tuning the PLMs on downstream tasks, while CreAT can effectively allow a more robust encoder.

5.2 Impact of CreAT Attack

We take a closer look into the learned contextualized representation of each intermediate BERT layer, including the hidden states and attention probabilities, to compare the impact of CreAT and AT. We calculate the cosine similarity of the hidden states and the KL divergence of the attention probabilities before and after perturbations.

From Figure 2 (left), CreAT and AT look similar on lower layers. For the last few layers ($10 \sim 12$), CreAT causes a stronger deviation to the model (lower similarity). From Figure 2 (right), we observe that CreAT has a strong capability to confuse the attention maps. It turns out AT is not strong enough to deviation the learned attentions of the model. This helps explain why CreAT can fool the PLM encoder more effectively.

6 Related Work

• **Language modeling:** Our work acts as the complement for AT on pre-trained language models (PLMs) (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; Clark et al., 2020; Raffel et al., 2020; Brown et al., 2020; He et al., 2021). From the perspective of language pre-training, it is essentially a common improvement of the conventional pre-training paradigm, e.g. MLM (Devlin et al., 2019; Clark et al., 2020; Yang et al., 2019; Wu et al., 2022). From the perspective of model architecture, it is parallel to strengthening encoder performances (Vaswani et al., 2017).

• **Adversarial attack:** Adversarial training (AT) (Goodfellow et al., 2015; Athalye et al., 2018; Miyato et al., 2019) is a common machine learning approach to create robust neural networks. The technique has been widely used in the image domain (Madry et al., 2018; Xie et al., 2019). In the text domain, the AT attack is different from the black-box text attack (Gao et al., 2018) where the users have no access to the model. Researchers typically substitute and reorganize the composition of the original sentences, while ensuring the semantic similarity, e.g. word substitution (Jin et al., 2020; Dong et al., 2021), scrambling (Ebrahimi et al., 2018; Zhang et al., 2019c), back translation (Iyyer et al., 2018; Belinkov & Bisk, 2018), language model generation (Li et al., 2020b; Garg & Ramakrishnan, 2020; Li et al., 2021). However, a convention philosophy of the AT attack is to impose bounded perturbations to word embeddings (Miyato et al., 2017; Wang et al., 2019b), which is recently proven to be well-deployed on PLMs and facilitate fine-tuning on downstream tasks, e.g. SMART (Jiang et al., 2020), FreeLB (Zhu et al., 2020), InfoBERT (Wang et al., 2021a), ASA (Wu & Zhao, 2022). ALUM (Liu et al., 2020) first demonstrates the promise of AT to language pre-training. Our work practices AT on more types of NLP tasks and re-investigates the source of gain brought by these AT algorithms.

• **Adversarial defense:** Our work act as a novel AT attack to obtain the global worst-case adversarial examples. It is agnostic to adversarial defense for the alignment between accuracy and robustness, e.g. TRADES (Zhang et al., 2019b), MART (Wang et al., 2020), SSL (Carmon et al., 2019).

• **Self-supervising:** AT is closely related to self-supervised learning (Madras et al., 2018; Hendrycks et al., 2019), since it corrupts the model inputs. On the other hand, it is parallel to weight perturbations (Wen et al., 2018; Khan et al., 2018; Wu et al., 2020), and contrastive learning (Hadsell et al., 2006) which relies on the corruption of model structures (Chen et al., 2020; Gao et al., 2021).

AT is still expensive at the moment, though there are works for acceleration (Shafahi et al., 2019; Zhang et al., 2019a; Zhu et al., 2020; Wong et al., 2020). This work is agnostic to these implementations. Another important line is to rationalize the behaviour of the attacker (Sato et al., 2018; Chen & Ji, 2022). In our work, we explain the impact of AT on contextualized language representation.

7 Conclusion

This paper investigates adversarial training (AT) on PLMs and proposes *Contextualized representation-Adversarial Training* (CreAT). This is motivated by the observation that the AT gain necessarily derives from deviating the contextualized language representation of PLM encoders. Comprehensive experiments demonstrate the effectiveness of CreAT on top of PLMs, which obtains the state-of-the-art performances on a series of challenging benchmarks. Our work is limited to Transformer-based PLMs and other architectures, vision models are not studied.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 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, 2021. URL `https://openreview.net/forum?id=OQ08SN70M1V`.

Gustavo Aguilar, Suraj Maharjan, Adrián Pastor López-Monroy, and Thamar Solorio. A multi-task approach for named entity recognition in social media data. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pp. 148–153. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-4419. URL `https://doi.org/10.18653/v1/w17-4419`.

Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of*

Machine Learning Research, pp. 274–283. PMLR, 2018. URL http://proceedings.mlr.press/v80/athalye18a.html.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=BJ8vJebC-.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Byg1v1HKDB.

Christopher M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Comput.*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108. URL https://doi.org/10.1162/neco.1995.7.1.108.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy Liang. Unlabeled data improves adversarial robustness. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11190–11201, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/32e0bd1497aa43e02a42f47d9d6515ad-Abstract.html.

Hanjie Chen and Yangfeng Ji. Adversarial training for improving model robustness? look at both prediction and interpretation. *CoRR*, abs/2203.12709, 2022. doi: 10.48550/arXiv.2203.12709. URL https://doi.org/10.48550/arXiv.2203.12709.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. 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*, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. Towards robustness against natural language word substitutions. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=ks5nebunVn_.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 31–36. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-2006. URL https://aclanthology.org/P18-2006/.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pp. 50–56. IEEE Computer Society, 2018. doi: 10.1109/SPW.2018.00016. URL https://doi.org/10.1109/SPW.2018.00016.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 6894–6910. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.552. URL https://doi.org/10.18653/v1/2021.emnlp-main.552.

Siddhant Garg and Goutham Ramakrishnan. BAE: bert-based adversarial examples for text classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6174–6181. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.498. URL https://doi.org/10.18653/v1/2020.emnlp-main.498.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pp. 1735–1742. IEEE Computer Society, 2006. doi: 10.1109/CVPR.2006.100. URL https://doi.org/10.1109/CVPR.2006.100.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 15637–15648, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a2b15837edac15df90721968986f7f8e-Abstract.html.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1875–1885. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1170. URL https://doi.org/10.18653/v1/n18-1170.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 2021–2031. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1215. URL https://doi.org/10.18653/v1/d17-1215.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 2177–2190. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.197. URL https://doi.org/10.18653/v1/2020.acl-main.197.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8018–8025. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6311.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2616–2625. PMLR, 2018. URL http://proceedings.mlr.press/v80/khan18a.html.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 785–794. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1082. URL https://doi.org/10.18653/v1/d17-1082.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJ6oOfqge.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 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, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 9119–9130. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.emnlp-main.733. URL https://doi.org/10.18653/v1/2020.emnlp-main.733.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized perturbation for textual adversarial attack. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *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*, pp. 5053–5069. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.400. URL https://doi.org/10.18653/v1/2021.naacl-main.400.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: adversarial attack against BERT using BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6193–6202. Association for Computational Linguistics, 2020b. doi: 10.18653/v1/2020.emnlp-main.500. URL `https://doi.org/10.18653/v1/2020.emnlp-main.500`.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models. *CoRR*, abs/2004.08994, 2020. URL `https://arxiv.org/abs/2004.08994`.

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

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3381–3390. PMLR, 2018. URL `http://proceedings.mlr.press/v80/madras18a.html`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun (eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL `http://arxiv.org/abs/1301.3781`.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=r1X3g2_xl`.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993, 2019. doi: 10.1109/TPAMI.2018.2858821. URL `https://doi.org/10.1109/TPAMI.2018.2858821`.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9102–9111, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/415185ea244ea2b2bedeb0449b926802-Abstract.html`.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4885–4901. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.441. URL `https://doi.org/10.18653/v1/2020.acl-main.441`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2383–2392. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1264. URL https://doi.org/10.18653/v1/d16-1264.

Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. Interpretable adversarial perturbation in input embedding space for text. In Jérôme Lang (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 4323–4330. ijcai.org, 2018. doi: 10.24963/ijcai.2018/601. URL https://doi.org/10.24963/ijcai.2018/601.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3353–3364, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/7503cfacd12053d309b6bed5c89de212-Abstract.html.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Trans. Assoc. Comput. Linguistics*, 7:217–231, 2019. URL https://transacl.org/ojs/index.php/tacl/article/view/1534.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 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, 2019a. URL https://openreview.net/forum?id=rJ4km2R5t7.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Infobert: Improving robustness of language models from an information theoretic perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=hpH98mK5Puk.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/335f5352088d7d9bf74191e006d8e24c-Abstract-round2.html.

Dilin Wang, ChengYue Gong, and Qiang Liu. Improving neural language modeling via adversarial training. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6555–6565. PMLR, 2019b. URL http://proceedings.mlr.press/v97/wang19f.html.

Hongjun Wang and Yisen Wang. Self-ensemble adversarial training for improved robustness. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event,* *April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=oU3aTsmeRQV.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rklOg6EFwS.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger B. Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJNpifWAb.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BJx040EFvH.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1ef91c212e30e14bf125e9374262401f-Abstract.html.

Hongqiu Wu and Hai Zhao. Adversarial self-attention for language understanding. *CoRR*, abs/2206.12608, 2022. doi: 10.48550/arXiv.2206.12608. URL https://doi.org/10.48550/arXiv.2206.12608.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Boli Chen, Pengjun Xie, Fei Huang, and Min Zhang. Forging multiple training objectives for pre-trained language models via meta-learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 6454–6466. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-emnlp.482.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 501–509. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00059. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Feature_Denoising_for_Improving_Adversarial_Robustness_CVPR_2019_paper.html.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5754–5764, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1472. URL https://doi.org/10.18653/v1/p19-1472.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 227–238, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/812b4ba287f5ee0bc9d43bbf5bbe87fb-Abstract.html.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 2019b. URL http://proceedings.mlr.press/v97/zhang19p.html.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: paraphrase adversaries from word scrambling. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 1298–1308. Association for Computational Linguistics, 2019c. doi: 10.18653/v1/n19-1131. URL https://doi.org/10.18653/v1/n19-1131.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BygzbyHFvB.

A Training Details

Table 7: Hyperparameters for pre-training.

	$\text{BERT}_{\text{base}}$	$\text{DeBERTa}_{\text{large}}$
Number of hidden layers	12	24
Hidden size	768	1024
Intermediate size	3072	4096
Number of attention heads	12	16
Dropout	0.1	0.1
Batch size	512	512
Learning rate	5e-5	6e-6
Weight Decay	0.01	0.01
Max sequence length	256	256
Warmup proportion	0.06	0.06
Max steps	100K	100K
Gradient clipping	1.0	1.0
Ascent step size	1e-1	1e-1
Decision boundary	1e-1	1e-1
Ascent steps	2	2
FP16	Yes	Yes
Number of GPUs	16	16
Training period	30 hours	100 hours

Table 8: Hyperparameters for fine-tuning BERT. For DeBERTa, we fix all hyperparameters the same, but set the learning rate to 1e-5 for all tasks. (dp: dropout rate, bsz: batch size, lr: learning rate, wd: weight decay, msl: max sequence length, wp: warmup, ep: epochs).

	(A)MNLI	QQP	WNUT	DREAM	H-SWAG	AlphaNLI	RACE	SQuAD
dp	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
bsz	128	128	16	16	16	64	8	16
lr	3e-5	5e-5	5e-5	2e-5	2e-5	5e-5	2e-5	5e-5
wd	0.01	0.01	0	0	0	0	0	0
msl	128	128	64	128	128	128	384	384
wp	0.06	0.06	0.1	0.06	0.06	0.06	0.06	0.06
ep	3	3	5	6	3	2	2	2
α	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1
ϵ	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1
k	1	1	2	1	1	1	1	1