TOWARD ADVERSARIAL TRAINING ON CONTEXTUALIZED LANGUAGE REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Beyond the success story of adversarial training (AT) in the recent text domain on top of pre-trained language models (PrLMs), our empirical results showcase that current AT can appear mediocre or even harmful on certain tasks, e.g. reading comprehension and commonsense reasoning. This paper investigates AT from the perspective of contextualized language representation. We find that the gain from AT does not derive from increasing the training risk, but from deviating the language representation. The fact is that the current AT attack is better at fooling the decoder (i.e. the classifier), but can be trivial to the encoder. Based on the observations, we propose simple yet effective *Contextualized representation-Adversarial Training* (CreAT), in which the attack is explicitly optimized to deviate the contextualized representation and obtains the global worst-case adversarial examples. CreAT is proven to be all-powerful compared to AT, with performance gain covering a wider range of downstream tasks. We apply CreAT to language pre-training. Our CreAT-empowered DeBERTa outperforms naive DeBERTa by a large margin, achieving the new state-of-the-art performances on a wide range of challenging benchmarks, e.g. AdvGLUE (59.1 $\rightarrow$ 61.1), HellaSWAG (93.0 $\rightarrow$ 94.9), ANLI (68.1 $\rightarrow$ 69.3), PAWS (50.3 $\rightarrow$ 54.5).

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 recent years, the success of AT in the text domain is attractive. There are studies (Zhu et al., 2020; Jiang et al., 2020) demonstrating that AT can be well deployed on pre-trained language models (PrLMs) (e.g. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)) and produces impressive performance gain on a number of benchmarks (e.g. GLUE, SNLI).

However, there are still some concerns about AT, for example, whether it is the adversarial examples that facilitate model training. Some studies (Moyer et al., 2018; Aghajanyan et al., 2021) point out that a similar effect can be achieved under random perturbations. On the other hand, there is insufficient practice of AT in the text domain. Most of the results reported in the current papers are based on natural language understanding tasks (e.g. sentiment analysis, language inference). It is not known for sure whether AT is effective on all types of NLP tasks.

Motivated by these issues, this paper practices AT on more types of NLP tasks (e.g. reading comprehension, dialogue, commonsense reasoning, named entity recognition), but yields inconsistent results with the previous papers. It shows that current AT can appear mediocre or even harmful on those tasks. We investigate AT from the perspective of *contextualized language representation*, which is obtained by Transformer encoder (Vaswani et al., 2017). Loosely speaking, a PrLM is composed of two parts, an encoder and a decoder. The encoder is broadly based on Transformer, while the decoder varies from tasks (e.g. a linear classifier). It is expected that the AT attack will similarly impact both parts. Our empirical results showcase that, on those tasks where AT fails, the AT attack is better at fooling the decoder, while the deviation caused to the encoder is trivial, comparable to random perturbation training (RPT) (Bishop, 1995).

Based on the observations, we are motivated that those adversarial examples that effectively deviate the contextualized representation are the driver for the success of AT. That is, the performance gain from AT on PrLMs derives from robust representation learning. 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 and obtains the global worst-case adversarial examples which fool the entire model. We find that CreAT is all-powerful across different types of tasks with no additional overhead compared to AT.

The Transformer architecture is the current paradigm for PrLMs. Our work facilitates the training of its encoder part, which benefits the entire model. This paper is organized as follows. § 2 presents the preliminary experiments, where we investigate and discuss the impact of AT based on contextualized language representation. § 3 presents our proposed CreAT. § 4 presents the experiments on two stages, pre-training and fine-tuning. On fine-tuning, CreAT is shown to be superior to the current state-of-the-art AT algorithms (e.g. FreeLB, SMART). We then apply CreAT to MLM-style language pre-training. Our CreAT-empowered pre-trained model achieves the new state-of-the-art performances on a wide range of challenging benchmarks (e.g. AdvGLUE, HellaSWAG, ANLI, PAWS). § 5 presents our further analysis, where we take a closer look at how CreAT works on the intermediate language representation (hidden states, attention maps). § 6 discusses the related work.

2 PRELIMINARIES

In this section, we start by reviewing the background of adversarial training and contextualized language representation, and then experiments are made to investigate the impact of adversarial perturbations on PrLMs (e.g. BERT) from two perspectives: (1) output predictions and (2) contextualized language representation. The visualization results lead us to a subtle 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:

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

where $\| \delta \| \leq \epsilon$ refers to the decision boundary restricting the magnitude of the perturbation. This paper defaults to the Frobenius norm.

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. In the image domain, AT can greatly impair model performances (Madry et al., 2018; Xie et al., 2019). Contrarily on PrLMs, such bounded embedding perturbations are proven to enhance performances (Jiang et al., 2020; Zhu et al., 2020; Wang et al., 2021a). However, there is no current evidence to uncover where such gain comes from.

2.2 CONTEXTUALIZED REPRESENTATION AND TRANSFORMER ENCODER

Transformer (Vaswani et al., 2017) has been broadly chosen as the fundamental encoder of PrLMs like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XL-Net (Yang et al., 2019), DeBERTa (He et al., 2021). 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.

A PrLM 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 usually initialized from scratch (we do not distinguish between

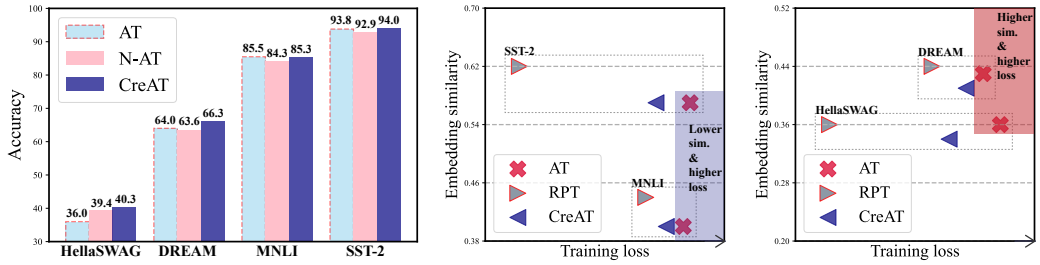


Figure 1: **Left:** Performances across different training methods, where N-AT refers to the regular situation where no adversarial training is applied. **Middle & right:** The relationship between the embedding similarity (contextualized representation) and training loss on different types of tasks. 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.

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 PrLMs

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 for BERT, 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 value 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. Note that 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, however, AT performs poorly, even leading to a 3.4 points drop on HellaSWAG.

• **Observation 2: AT contributes the most to deviate model predictions.** From Figure 1 (middle and right), we can see that the training loss under AT is always the highest (always on the rightmost side of the graph). It suggests that AT works well to fool the model as it is meant to.

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

The above observations raise a question of interest on AT in the text domain: **whether fooling model predictions is positively correlated with the training performances of PrLMs?** 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 PrLMs is driven by robust language representation learning**. 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 in this paper, which is widely used to measure the distance between two word vectors (Mikolov et al., 2013; Gao et al., 2021). Thus, deviating the contextualized representation is as follows:

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

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:

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

where τ is the temperature coefficient to control the strength of the attacker on the contextual representation and 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, we find that AT can lead to the local worst case, which mainly fools the classifier (e.g. a single-layer perceptron) and ignores the impact on the encoder. However, our empirical results demonstrate that a robust encoder allows the model to perform better in aspects like comprehension and reasoning. CreAT can be regarded as a "global" form of AT, which solves the global worst case for the entire model (both the encoder and classifier). The resultant model is encouraged to resist those attacks and keep its contextualized representation robust.

Model 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 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 PrLMs.

Algorithm 1 summarizes the pseudocode of CreAT. Our inner optimization is based on projected gradient descent (PGD) (Madry et al., 2018). At each training step, which corresponds to the outer

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: $\{\mathbf{x}, y\} \leftarrow \text{SampleBatch}(T)$
3: $\delta_0 \sim \mathcal{N}(0, 1)$
4: **for** $j = 1$ to k **do**
5: $\text{ForwardNoGrad}(\mathbf{x}, \theta)$ ▷ Go benign forward pass
6: $\text{Forward}(\mathbf{x} + \delta_{j-1}, \theta)$ ▷ Go adversarial forward pass
7: $\delta_j \leftarrow \text{BackwardUpdate}(\delta_{j-1}, y, \tau, \alpha, \epsilon)$ ▷ Update the perturbation following Eq. 4
8: **end for**
9: $\text{Forward}(\mathbf{x}, \theta), \text{Forward}(\mathbf{x} + \delta^*, \theta)$ ▷ $\delta^* \leftarrow \delta_k$
10: $\theta \leftarrow \text{BackwardUpdate}(\theta, y, \beta)$ ▷ Update the model 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 SOTA (94.9) on HellaSWAG (H-SWAG) on May 5, 2022[1] and we do not report it here for space limitation.

	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**	$48.6_{1.1}$	$64.9_{1.0}$	36.7	65.3	62.8	55.7	$\downarrow$ 0.6
+ *SMART*	**85.6**	72.7	$48.0_{0.6}$	$63.2_{0.4}$	38.2	64.8	63.3	55.5	$\downarrow$ 0.8
+ *AT*	85.2	72.9	$49.7_{1.2}$	$63.8_{0.6}$	36.0	64.9	63.0	55.5	$\downarrow$ 0.8
+ *CreAT*	85.3	73.0	$\mathbf{50.1}_{0.9}$	$\mathbf{66.0}_{0.7}$	**40.3**	**67.1**	**68.0**	**58.3**	$\uparrow$ **2.0**

loop (line 2 $\sim$ line 10), we fetch the training samples and initialize the perturbation. Then we go to the inner loop (line 5 $\sim$ line 7), where we iterate to evaluate δ^* by taking multiple projected gradient steps. Note that the benign forward pass (line 5) does not require gradients, which may accelerate the training process. 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

This section reports the empirical results of fine-tuning and pre-training with CreAT on various benchmarks. Our implementation is based on *transformers* (Wolf et al., 2020).

4.1 Experimental Setup

We conduct fine-tuning on $\text{BERT}_{\text{base}}$ (Devlin et al., 2019). For adversarial training, we fix the ascent step size α and decision boundary ϵ to 1e-1, temperature τ to 1, and choose the number of ascent steps k in $\{1, 2\}$ 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: $\text{BERT}^{CreAT}_{\text{base}}$ based on $\text{BERT}_{\text{base}}$ (Devlin et al., 2019), $\text{DeBERTa}^{CreAT}_{\text{large}}$ 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, 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 D for more hyperparameters.

We describe a number of adversarial training counterparts below (details in Appendix A).

• **FreeLB** (Zhu et al., 2020) is a state-of-the-art fast adversarial training algorithm, which incorporates every intermediate examples into model learning (α =1e-1, ϵ =1e-1, k =2).

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

Table 2: Results on AdvGLUE benchmark. 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)	
$\text{BERT}_{\text{base}}$	$32.3_{1.4}$	$50.8_{2.1}$/-	$40.1_{0.6}$	$32.6_{1.3}$	$19.3_{0.8}$	$37.0_{2.5}$	35.4
FreeLB-$\text{BERT}_{\text{base}}$	$31.6_{0.3}$	$51.0_{2.4}$/-	$\mathbf{45.4}_{0.3}$	$33.5_{0.4}$	$21.9_{0.9}$	$42.0_{1.5}$	37.6
$\text{BERT}^{\text{MLM}}_{\text{base}}$	$32.0_{0.6}$	$48.5_{1.3}$/-	$43.4_{1.4}$	$27.6_{0.4}$	$20.8_{0.5}$	$\mathbf{45.9}_{2.5}$	36.4
$\text{BERT}^{CreAT}_{\text{base}}$	$\mathbf{35.3}_{1.0}$	$\mathbf{51.5}_{1.2}$/-	$44.8_{0.6}$	$\mathbf{36.0}_{0.8}$	$\mathbf{22.0}_{0.8}$	$45.2_{2.0}$	**39.1** ↑**2.7**
$\text{DeBERTa}^{\text{MLM}}_{\text{large}}$	$\mathbf{65.4}_{0.3}$	$72.3_{1.8}$/-	$65.8_{1.1}$	$68.4_{0.7}$	$43.7_{0.6}$	$75.4_{1.5}$	65.2
$\text{DeBERTa}^{CreAT}_{\text{large}}$	$65.3_{0.8}$	$\mathbf{78.5}_{1.8}$/-	$\mathbf{70.3}_{1.1}$	$\mathbf{70.0}_{0.4}$	$\mathbf{44.6}_{0.9}$	$\mathbf{76.5}_{1.0}$	**67.5** ↑**2.3**
SMART-$\text{RoBERTa}_{\text{large}}$	50.9	64.2/44.3	52.2	45.6	36.1	70.4	52.0
T5_{large}	60.6	63.0/**55.7**	57.6	48.4	39.0	62.8	55.3
$\text{DeBERTa}_{\text{large}}$	57.9	60.4/48.0	57.9	**58.4**	**52.5**	**78.9**	59.1
$\text{DeBERTa}^{CreAT}_{\text{large}}$	**63.5**	**67.5**/54.5	**61.8**	56.7	51.7	72.0	**61.1** ↑**2.0**

• **SMART** (Jiang et al., 2020) is another state-of-the-art adversarial training algorithm, whose idea is similar to TRADES (Zhang et al., 2019b) (α =1e-3, ϵ =1e-5, k =1, μ =0.02).

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

CreAT can be adapted to both implementations of FreeLB and SMART. This part is in Appendix A.

4.2 Results on Fine-tuning

We experiment on a wide range of tasks, including MNLI (natural language inference) and 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 and AlphaNLI (commonsense reasoning) (Zellers et al., 2019; Bhagavatula et al., 2020); RACE (reading comprehension) (Lai et al., 2017).

Table 1 summarizes the results based on BERT. What we can see first is that each of the adversarial training methods can achieve remarkable improvement on MNLI and QQP (two sentence classification tasks), and the improvement is quite similar. What is totally different is that, on the right side of the table, we can see adversarial training only brings performance drops in most cases (average drop over BERT: 0.6 point for FreeLB, 0.8 for SMART, 0.8 for AT), suggesting that adversarial training can be harmful on certain tasks. However, CreAT appears all-powerful, outweighing all the counterparts (absolute gain over BERT: 3.0 points on DREAM, 1.9 on AlphaNLI, 2.7 on RACE), and raising the average score by 1.9 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 the four mentioned tasks and it is a trade-off between the training loss and deviating the representation.

Starting from the next section, we start to report the pre-training results. To distinguish different training settings, for example, FreeLB-$\text{BERT}_{\text{base}}$ means we fine-tune the original BERT model with FreeLB; $\text{BERT}^{CreAT}_{\text{base}}$ means we continually pre-train the model and then regularly fine-tune it without adversarial training; $\text{BERT}^{\text{MLM}}_{\text{base}}$ means we regularly pre-train and fine-tune the model.

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. 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.

We summarize the results on all AdvGLUE sub-tasks in Table 2. For dev sets, two CreAT-empowered models consistently outperform naive $\text{BERT}_{\text{base}}$ and $\text{DeBERTa}_{\text{large}}$ by a large margin.

[2] `Anonymous` (since the platform does not allow anonymous submission).

In addition, we can see that directly fine-tuning $\text{BERT}_{\text{base}}^{CreAT}$ works better than fine-tuning with FreeLB, which also accelerates the training cycle. To verify the effectiveness of CreAT-based adversarial pre-training, we pre-train another two models, $\text{BERT}_{\text{base}}^{\text{MLM}}$ and $\text{DeBERTa}_{\text{large}}^{\text{MLM}}$ based on MLM with the same number of steps (i.e. 100K steps). We see that simply training longer leads to limited performance gain on robustness tasks, while $\text{BERT}_{\text{base}}^{CreAT}$ 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). For test sets, where $\text{DeBERTa}_{\text{large}}^{CreAT}$ is compared with a number of strong PrLMs (Liu et al., 2019; Raffel et al., 2020), it outperforms T5 and DeBERTa, raising the average score from 59.1 to 61.1 and achieving the new SOTA performance.

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. From Table 3, we see that two CreAT-empowered models consistently outperform naive models by 1 to 3 points in all metrics.

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

	AddSent (EM/F1)	AddOneSent (EM/F1)
$\text{BERT}_{\text{base}}$	$59.3_{.4}/65.7_{.2}$	$67.3_{.3}/74.3_{.4}$
$\text{FreeLB-BERT}_{\text{base}}$	$60.9_{.6}/67.4_{.3}$ ↑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**

4.5 ADVERSARIAL NLI

Adversarial Natural Language Inference (ANLI) (Nie et al., 2020) is an iteratively-strengthened challenging benchmark. In each round, human annotators are asked to craft harder samples to fool the model. In our experiment, we concatenate the training samples of ANLI and MNLI for training. We summarize the test results of all rounds in Table 4. We can first see that $\text{BERT}_{\text{base}}^{CreAT}$ is so powerful that it even surpasses $\text{BERT}_{\text{large}}$, outperforming by 1.6 points and 0.8 point on Round 1 and Round 3. Besides, we see that DeBERTa significantly outperforms the previous SOTA 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, especially on the most challenging Round 3, outperforming $\text{DeBERTa}_{\text{large}}$ by 1.9 points.

Table 4: Test results of all rounds on ANLI over five runs. We report the previous SOTA results at the bottom[3], InfoBERT and ALUM.

	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

4.6 PAWS

Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019c) generates challenging sentence pairs from Quora and Wikipedia by word swapping and back translation, thus amplifying the problem of the existing language models failing to capture word order and syntactic structure. For PAWS-QQP, we evaluate the model in three training settings: zero-shot setting (we train the model with only QQP training samples), normal setting (train the model with only PAWS samples), and full setting (train the model with both QQP and PAWS samples).

From Table 5, we can see that two CreAT-empowered models perform consistently better on both PAWS-QQP and PAWS-Wiki. Especially in the zero-shot setting on PAWS-QQP, where the model is not exposed to the reordered training samples, the advantages of CreAT are particularly notable, outperforming BERT and DeBERTa by 2.6 points and 4.2 points respectively.

[3] `https://github.com/facebookresearch/anli/checkpoint_results`.

Table 5: Results on PAWS over five runs, where PAWS-QQP$_-$, PAWS-QQP$_+$ and PAWS-QQP correspond to the zero-shot, full, and normal training settings respectively.

	PAWS-QQP$_-$	PAWS-QQP$_+$	PAWS-QQP	PAWS-Wiki	Avg
BERT$_{\text{base}}$	$32.3_{.6}$	$82.0_{2.0}$	$87.0_{1.2}$	$91.4_{.1}$	73.2
BERT$_{\text{base}}^{CreAT}$	$\mathbf{34.9}_{.9}$	$\mathbf{86.7}_{1.3}$	$\mathbf{90.2}_{.7}$	$\mathbf{93.1}_{.2}$	**76.2 ↑3.0**
DeBERTa$_{\text{large}}$	$50.3_{.7}$	$94.1_{.5}$	$94.8_{.2}$	$94.9_{.1}$	83.5
DeBERTa$_{\text{large}}^{CreAT}$	$\mathbf{54.5}_{.8}$	$\mathbf{95.6}_{.1}$	$\mathbf{95.7}_{.2}$	$\mathbf{95.5}_{.0}$	**85.3 ↑1.8**

5 ANALYSIS

5.1 ABLATION STUDY

Previous experiments have shown the advantages of CreAT over AT. 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 the regularly trained BERT model. The next level is to apply AT, where the random perturbations become adversarial ones. Here we introduce a new strong baseline, where we additionally impose the adversarial perturbations to the output embeddings of the encoder (contextualized representation). We denote this as AT$^+$. Eventually, we progress AT$^+$ to CreAT. The difference between them is that the deviation to the contextual representation for CreAT relies on its optimization, whereas it is bounded for AT$^+$.

Table 6: 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
+ *AT*$^+$	85.2	62.5	65.8
+ *CreAT*	**85.3**	**66.4**	**67.1**

From Table 6, we can see that AT is superior to RPT which achieves limited results on top of BERT on all three datasets. However, the improvement from CreAT is much more impressive, outperforming AT by a large margin on DREAM and AlphaNLI. In addition, we observe that AT$^+$ still obtains the local worst case, which confirms the need to allow the attacker to optimize in the direction of deviating the contextualized representation.

Table 7 demonstrates a number of adversarial pre-training methods. We can see that CreAT obtains stronger performances than regularly training MLM for a longer period on all three tasks. However, the FreeLB-based one is almost comparable to regular MLM, while the AT-based one is slightly better than it. It implies that language modeling is basically to learn language representation, and therefore facilitating language representation can lead to promising language model performances.

Table 7: 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**

5.2 IMPACT OF CREAT ATTACK

Previous visualization (Figure 1) has shown us CreAT can deviate the contextualized representation more effectively. In this part, we will take a closer look at its impact. What we do actually is to observe the learned representation of each intermediate layer of BERT, including the hidden states and attention probabilities. 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), we can see that CreAT and AT perform similarly on lower layers. For the last few layers, especially the last layer, CreAT causes a stronger deviation to the model. From Figure 2 (right), we can see that CreAT has a strong capability to confuse attention maps. In contrast, we also find that the current AT algorithms (e.g. PGD, FreeLB) can only cause limited attention deviation to the model. This helps explain why CreAT can fool the PrLM encoder more effectively.

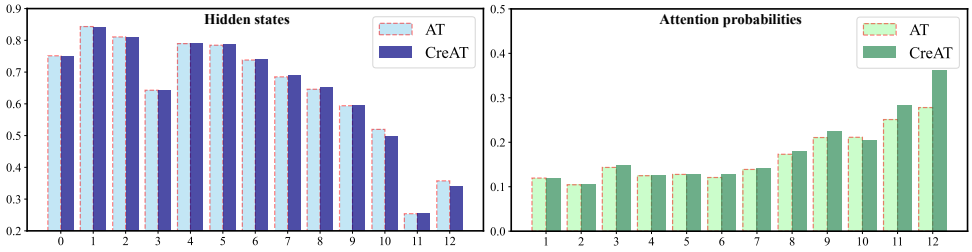


Figure 2: Impact of CreAT attack on the intermediate layers of $\text{BERT}_{\text{base}}$ (12 layers).

6 Related Work

• **Language modeling:** Our work acts as the complement for AT on pre-train language models (PrLMs) (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), RTD (Clark et al., 2020), PLM (Yang et al., 2019). From the perspective of model architecture, it is parallel to strengthening encoder performances (Vaswani et al., 2017), or Transformer variants (Zaheer et al., 2020; Shi et al., 2021; Wu et al., 2021; Tay et al., 2021).

• **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 PrLMs and facilitate fine-tuning on downstream tasks, e.g. SMART (Jiang et al., 2020), FreeLB (Zhu et al., 2020), InfoBERT (Wang et al., 2021a). 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 and Limitation

This paper re-investigates current AT on PrLMs and proposes *Contextualized representation-Adversarial Training* (CreAT). This is motivated by our observation that the AT gain does not derive from increasing the training risk, but from deviating the language representation. Comprehensive experiments show that our CreAT-empowered model achieves the new SOTA performances on a series of challenging benchmarks. Our work is limited to Transformer-based PrLMs. We will follow up with more attempts to explore the impact of our approach on traditional RNN-based models.

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.

Han Shi, Jiahui Gao, Xiaozhe Ren, Hang Xu, Xiaodan Liang, Zhenguo Li, and James Tin-Yau Kwok. Sparsebert: Rethinking the importance analysis in self-attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9547–9557. PMLR, 2021. URL http://proceedings.mlr.press/v139/shi21a.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.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10183–10192. PMLR, 2021. URL http://proceedings.mlr.press/v139/tay21a.html.

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, Hai Zhao, and Min Zhang. Code summarization with structure-induced transformer. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pp. 1078–1090. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.93. URL https://doi.org/10.18653/v1/2021.findings-acl.93.

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.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. 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/c8512d142a2d849725f31a9a7a361ab9-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 CreAT Adaption

FreeLB Considering the convenience that the attacker and the model are simply optimizing in opposite directions, FreeLB merges two backward passes into once using the gradient accumulation. Thus, similar to CreAT, FreeLB optimizes the ensemble of the benign risk and adversarial risk:

$$\min_{\theta} \lambda \mathcal{L}(\mathbf{x}, y, \theta) + (1-\lambda)\mathcal{L}(\mathbf{x}+\delta^*, y, \theta) \tag{6}$$

where $\lambda = 0.5$. Note that the equation only holds when $k = 2$ and we omit the other cases.

SMART Different from CreAT, SMART acts as a regularizer on top of the benign risk:

$$\min_{\theta} \mathcal{L}(\mathbf{x}, y, \theta) + \mu \mathcal{D}\left[p(y|\mathbf{x}, \theta), p(y|\mathbf{x}+\delta^*, \theta)\right] \tag{7}$$

where $\mu = 0.02$. The second term keeps the model prediction smooth around the input points.

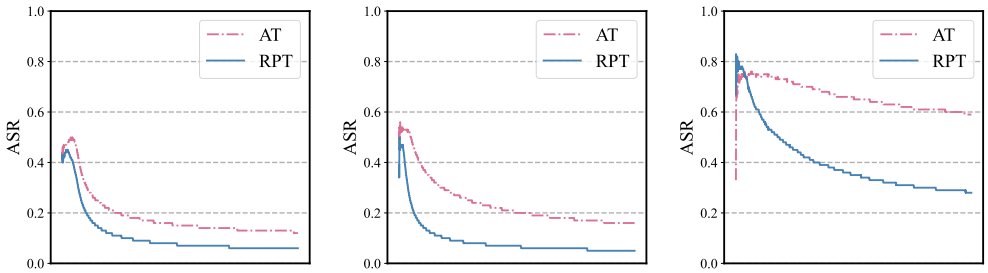
For CreAT-FreeLB, we need to modify the training objective, since it should be the same for both the attacker and the model. To this end, we utilize the objective in Eq. 4. As the result of the regularization term $\tau\mathcal{S}\left[h(\mathbf{x}, \theta), h(\mathbf{x}+\delta, \theta)\right]$, the model will additionally learn to keep its contextualized representation smoothed. That is why we need to go through the benign pass first (line 4).

For CreAT-SMART, the situation is much easier. The perturbation can be solved by $\delta^* = \arg\max_{\delta;\|\delta\|\leq\epsilon} \mathcal{D}\left[p(y|\mathbf{x}, \theta), p(y|\mathbf{x}+\delta, \theta)\right] - \tau\mathcal{S}\left[h(\mathbf{x}, \theta), h(\mathbf{x}+\delta, \theta)\right]$. Thus, we do not need the ground truth y when updating it (line 7).

Algorithm 2 CreAT-FreeLB

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: $\{\mathbf{x}, y\} \leftarrow \text{SampleBatch}(T)$
3: $\delta_0 \sim \mathcal{N}(0, 1)$
4: $\text{ForwardNoGrad}(\mathbf{x}, \theta)$
5: **for** $j = 1$ to k **do**
6: $\text{Forward}(\mathbf{x}+\delta_{j-1}, \theta)$
7: $\nabla\theta, \nabla\delta_{j-1} \leftarrow \text{Backward}(\delta_{j-1}, y, \theta)$
8: $\nabla\theta \leftarrow \text{Accumulate}(\nabla\theta)$
9: $\delta_j \leftarrow \text{Update}(\delta_{j-1}, -\nabla\delta_{j-1}, \alpha)$
10: **end for**
11: $\theta \leftarrow \text{Update}(\theta, \nabla\theta/k, \beta)$
12: **end while**



(a) MNLI (b) AlphaNLI (c) HellaSWAG

Figure 3: The attack success rate with training steps. The implementation here is FreeLB.

Algorithm 3 CreAT-SMART

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

while not converged **do**
  $\{\mathbf{x}, y\} \leftarrow \text{SampleBatch}(T)$
  $\delta_0 \sim \mathcal{N}(0, 1)$
  $\text{Forward}(\mathbf{x}, \theta)$
  for $j = 1$ to k **do**
    $\text{Forward}(\mathbf{x} + \delta_{j-1}, \theta)$
    $\delta_j \leftarrow \text{BackwardUpdate}(\delta_{j-1}, \tau, \alpha, \epsilon)$
  end for
  $\delta^* \leftarrow \delta_k$
  $\text{Forward}(\mathbf{x} + \delta^*, \theta)$
  $\theta \leftarrow \text{BackwardUpdate}(\theta, y, \beta)$
end while

B Attack Success Rate

The attack success rate (ASR) is defined as the proportion of attacks that cause the model to flip its predictions on the examples. A higher ASR means a more effective attack. A typical black-box adversarial text attack algorithm can yield an ASR between 0.7 and 0.9, varying from datasets.

Different from the text attack, however, we find that the AT attack on word embeddings only yields a much lower ASR. From Figure 3, we can see that the ASR decays very quickly as the training proceeds. For the AT attack, it finally leads to an ASR of around 0.1 and 0.2 on MNLI and AlphaNLI. Only on more challenging HellaSWAG, AT maintains a relatively high ASR of around 0.7.

This reflects that the AT attack in the text domain, although effective in deviating the output distribution of the classifier, can hardly result in the prediction flipping. An obvious explanation is that we typically use a small number of steps (e.g. $k = 2$) to obtain the adversarial examples. For the text attack, however, the search space for the adversarial examples is very large.

More importantly, we can derive that those adversarial examples which flip the model predictions are also not the driver for the performance gain from AT.

C AT on Transformer

We also find that applying AT or CreAT on scratch Transformer (e.g. random initialized BERT) can be harmful. It suggests that the pre-trained weights are the key to the model resisting attacks and learning better as a result. It also leaves the possibility that the failure of the AT attack to effectively deviate the contextual representation on some tasks is also related to the pre-trained weights. We will leave this part for the future work.

D TRAINING DETAILS

Table 8: 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 9: 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	PAWS	WNUT	DREAM	H-SWAG	AlphaNLI	RACE	SQuAD
dp	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
bsz	128	128	32	16	16	16	64	8	16
lr	3e-5	5e-5	5e-5	5e-5	2e-5	2e-5	5e-5	2e-5	5e-5
wd	0.01	0.01	0.01	0	0	0	0	0	0
msl	128	128	128	64	128	128	128	384	384
wp	0.06	0.06	0.06	0.1	0.06	0.06	0.06	0.06	0.06
ep	3	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	1e-1
ϵ	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1	1e-1
k	1	1	1	2	1	1	1	1	1