Exploring Robust Overfitting for Pre-trained Language Models

Bin Zhu and **Yanghui Rao***
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
zhub35@mail2.sysu.edu.cn, raoyangh@mail.sysu.edu.cn

Abstract

We identify the robust overfitting issue for pre-trained language models by showing that the robust test loss increases as the epoch grows. Through comprehensive exploration of the robust loss on the training set, we attribute robust overfitting to the model's memorization of the adversarial training data. We attempt to mitigate robust overfitting by combining regularization methods with adversarial training. Following the philosophy to prevent the model from memorizing the adversarial data, we find that flooding, a regularization method with loss scaling, can mitigate robust overfitting for pre-trained language models. Eventually, we investigate the effect of flooding levels and evaluate the models' adversarial robustness under textual adversarial attacks. Extensive experiments demonstrate that our method can mitigate robust overfitting upon three top adversarial training methods and further promote adversarial robustness.

1 Introduction

Deep neural networks (DNNs) suffer from adversarial robustness issues (Goodfellow et al., 2015; Szegedy et al., 2014; Papernot et al., 2016a). Recent literature has revealed their vulnerability to crafted adversarial examples on a wide range of natural language processing (NLP) tasks (Papernot et al., 2016b; Ren et al., 2019; Jin et al., 2020; Li et al., 2020). Among the corresponding defensive methods, gradient-based adversarial training (AT) is often considered as the most effective one.

Building upon standard training, AT additionally solves a max-min optimization problem to learn an adversarially robust model (Goodfellow et al., 2015; Kurakin et al., 2017; Madry et al., 2018; Zhang et al., 2020). Surprisingly, a widely observed fact is that AT, which is challenging to optimize, can also converge quickly on pre-trained language models (PrLMs) (Li and Qiu, 2021; Li et al., 2021b). That is, due to their overparameterization, PrLMs can achieve zero robust training error within a few epochs. It is common in practice to achieve zero training error without harming the generalization performance when sufficient data is available, which indicates that overfitting does not occur in the standard training of many modern deep learning tasks (Zhang et al., 2017; Neyshabur et al., 2017; Belkin et al., 2019). Nevertheless, whether PrLMs will overfit when trained to zero robust training error is yet to be explored.

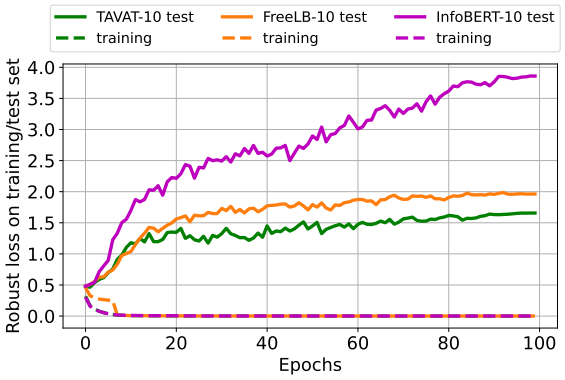


Figure 1: Robust loss against PGD-10 attacks on the SST-2 dataset. The base model is BERT-base (Devlin et al., 2019). "-10" indicates that the number of attack iterations used in AT is 10. The robust test losses of the three top AT methods only increase as the training epochs grow. In contrast, the robust training losses converge to zero within very few epochs.

As revealed by a recent work (Rice et al., 2020), robust overfitting dominates the training procedure of the image classification task, in which the robust test loss increases as the learning rate decays. In contrast, the robust training loss continues to decrease. This motivates us to identify the robust overfitting issue in adversarially robust learning for NLP models. We first visualize the robust test loss of various effective AT methods developed for NLP tasks. We adopt the simple yet effective Projected

*The corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 5506–5522
July 9-14, 2023

Gradient Descent (PGD) attack (Madry et al., 2018) rather than any other textual adversarial attacks to get universal results. This is because textual adversarial attacks integrate too many strategies, and the results under a particular textual adversarial attack may not be generalizable. We can observe from Figure 1 that the robust test loss only increases as the training epochs grow, which is counterintuitive. It also violates the common practice of taking the last checkpoint as an adversarially robust model. In contrast, the robust training loss converges to zero quickly. We refer to the difference between the two robust losses as an adversarial generalization gap. What is worse, the generalization gap appears in the early stage of AT and grows during the whole training phase. This initial finding inspires us to explore the convergence and generalization of AT in-depth and to ask the following question:

- *Why does the robust test loss continue to increase as adversarial training goes?*

We further explore the robust loss and accuracy curves on the training set. More specifically, we re-perform a PGD-10 attack on the training set to check the robust learning curves. We surprisingly observe that on the training set, both the robust loss and robust error under PGD-10 converge to small values. We also evaluate the adversarial robustness under different settings, such as datasets, model architectures, etc., and similar results are observed. With extensive empirical results, we argue that the model overfits the threat model used in AT and loses the adversarial generalization ability. We hypothesise that the model simply memorizes the adversarial data during training and fails to generalize to robust testing. Thus a poor adversarial generalization performance is observed on the test set. We make several attempts to mitigate robust overfitting issues in AT using a series of regularization methods. The underlying philosophy is to prevent the model from memorizing adversarial data. In this way, we prevent the adversarially trained model from robust overfitting. Eventually, we evaluate our methods against textual adversarial attacks and obtain improvements upon the existing AT methods. Our contributions can be summarized as follows:

- We identify the robust overfitting issue in AT for PrLMs. Through in-depth explorations, we attribute the robust overfitting to memorizing the adversarial training data.
- We make empirical attempts to mitigate robust overfitting using a series of regularization methods. We propose calibrating the model's overconfident prediction in AT[1]. Extensive experimental results demonstrate that our methods can mitigate robust overfitting and improve the adversarial robustness of models upon three top AT methods.

2 Related Work

In this section, we briefly review the relevant work on AT and robust overfitting, especially for NLP tasks.

2.1 Adversarial Training

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be the training set, in which $x_i \in \mathcal{X}$ is an input sample with its corresponding true label $y_i \in \mathcal{Y}$. AT aims to learn adversarially robust models by expanding the training set with adversarial data, which can be formulated as the following max-min optimization problem:

$$\min_{\theta} \mathbb{E}_{(\mathcal{X},\mathcal{Y})\sim\mathcal{D}} \left[\max_{\|\delta\|\leq\epsilon} \mathcal{L}(f_\theta(\mathcal{X}+\delta), \mathcal{Y}) \right], \quad (1)$$

where f_θ is a neural network parameterized by θ, $\mathcal{L}(\cdot)$ is the loss function, δ is the adversarial perturbation, and ϵ is the allowed perturbation size.

To tackle the intractable problem, Goodfellow et al. (2015) first proposed to use a one-step gradient-based method to generate adversarial examples, also known as the Fast Gradient Sign Method (FGSM). Madry et al. (2018) extended it to a multi-step method with random starts known as the PGD method.

Unfortunately, AT always leads to a drop in the standard accuracy. Zhang et al. (2019b) theoretically identified a trade-off between robustness and accuracy and proposed TRADES to trade adversarial robustness off against accuracy. Considering that PGD-based AT is time-consuming, there is another line of work focused on accelerating AT (Shafahi et al., 2019; Zhang et al., 2019a; Wong et al., 2020).

For NLP tasks, Miyato et al. (2017) first found that AT could help generalization in a semi-supervised manner. To make AT more reasonable, Sato et al. (2018) proposed to generate interpretable adversarial perturbations in the embedding space to improve standard accuracy. Zhu et al.

[1]Our code is available in public at https://github.com/zedzx1uv/GAT.

(2020) proposed a model named FreeLB to understand natural languages better. To exploit the implicit information in the text, Li and Qiu (2021) crafted fine-grained perturbations for tokens in their model named TAVAT and obtained improvements on both the standard and robust accuracy. Wang et al. (2021a) improved AT from an information theoretic perspective termed InfoBERT. Dong et al. (2021b) proposed RIFT to encourage the model to retain the information from the original pre-trained model. To benchmark the existing defensive methods, Li et al. (2021b) gave a systematic analysis of them under the same attack settings. They also found that removing the norm-bounded projection and increasing adversarial steps could improve adversarial robustness.

To defend against the widely used adversarial word substitutions, Jia et al. (2019) captured the perturbation in a hyper-rectangle and obtained certified robustness. Dong et al. (2021a) further modelled the word substitution attack space as a convex hull to enhance adversarial robustness. Wang et al. (2021c) proposed to project the perturbed word embedding to a valid one so that the crafted adversarial examples are reasonable. By learning a robust word embedding space where synonyms have similar representations, Yang et al. (2022) promoted models' robustness and maintained competitive standard accuracy.

Discrete adversarial data augmentation (Ren et al., 2019; Jin et al., 2020; Zang et al., 2020; Li et al., 2020; Si et al., 2021; Li et al., 2021a) can also significantly improve adversarial robustness by generating valid adversarial examples to expand the training set. However, the adversarially trained model suffers from degraded generalization performance. Another disadvantage is that it only helps defend against the same attacking method with adversarial data augmentation. To this end, Zhu et al. (2022) developed friendly adversarial data augmentation to improve adversarial robustness without hurting standard accuracy.

AT empirically boosts the adversarial robustness of models, but no guarantees can be given for the robustness. Therefore, another series of work devotes to obtaining certified robustness under given adversarial strengths by using randomized smoothing (Ye et al., 2020), interval bound propagation (Jia et al., 2019; Huang et al., 2019; Shi et al., 2020), differential privacy (Wang et al., 2021b), etc.

2.2 Robust Overfitting

Robust overfitting occurs immediately after the learning rate decays in AT across datasets, model architectures, and AT methods in computer vision (Rice et al., 2020; Rebuffi et al., 2021; Dong et al., 2022a). The robust training loss continues to decrease while the robust test loss begins to increase. They also found that only the combination of early stopping and semi-supervised data augmentation works better than early stopping alone. Since it is common in practice to train deep models as long as possible in computer vision, robust overfitting counteracts the gains of robustness by recent variants of AT.

From the perspective of the weight loss landscape, Wu et al. (2020) proposed adversarial weight perturbation to improve robust generalization. Chen et al. (2021) empirically injected learned smoothing into AT to avoid overfitting in AT. Dong et al. (2022a) introduced a new insight into the relationships between noisy labels and robust overfitting. Rebuffi et al. (2021) found that data augmentation with model weight averaging could also mitigate robust overfitting.

Similarly, Dong et al. (2022b) integrated temporal ensemble into AT frameworks, which could be seen as another form of weight averaging. Yu et al. (2022) explored robust overfitting from data loss distributions. They attributed robust overfitting to the small-loss data under a large perturbation size.

In this paper, we mainly focus on the convergence and robust overfitting of gradient-based AT methods, which have rarely been studied in the NLP field.

3 Robust Overfitting for PrLMs

In this section, we explore the robust learning curves for PrLMs. By comparing the data loss distributions between training and testing, we identify the robust overfitting issue and attribute it to the model's memorization of adversarial data.

3.1 Identifying Robust Overfitting

Motivated by the findings from Figure 1, we make a further study on the training set to see whether the model simply memorizes the adversarial training data. We re-perform PGD-10 attacks on the training set. Since the PGD adversary randomly initializes the starting point $x^{(0)}$ at a $\epsilon-$ball centred by the input x, **we expect a decrease in the robustness of the model on the training set**.

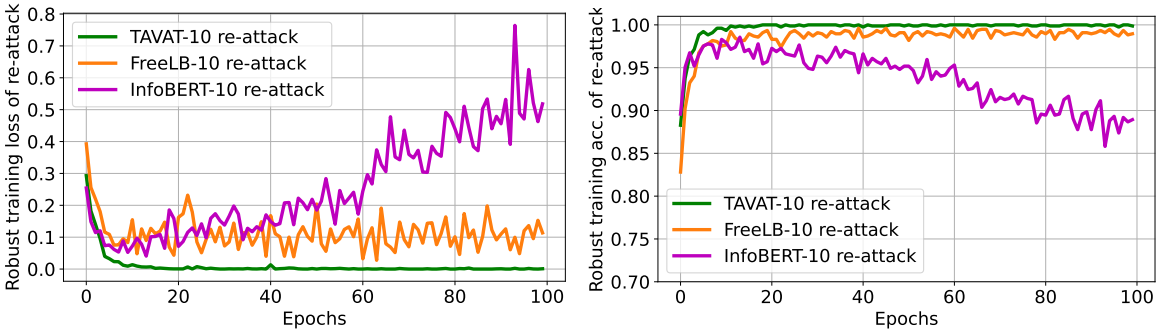


Figure 2: Robust loss and accuracy against re-performed PGD-10 attacks on the SST-2 training set. The base model is BERT-base. The small robust loss and high robust accuracy imply that the model simply memorizes the adversarial data.

We adopt three AT methods, FreeLB (Zhu et al., 2020), TAVAT (Li and Qiu, 2021), and InfoBERT (Wang et al., 2021a), to provide comprehensive results. "-10" refers to the number of attack iterations used in AT is set to 10. As can be seen in Figure 2, the robust losses of the three methods decrease at early epochs, which indicates that the model memorizes the adversarial data quickly.

In subsequent epochs, "TAVAT-10" and "FreeLB-10" maintain small losses. The robust loss of "InfoBERT" gradually increases but is still less than 0.8. For robust accuracy, similar results are observed. It is not surprising that "InfoBERT-10" has a slightly large robust loss and a degraded robust accuracy since we have observed that its robust test loss is abnormally large compared to others in Figure 1.

Comparing the robust loss on the training set with that on the test set, we can conclude that the model can not generalize to the adversarial test set, although it achieves about 100% robust accuracy during training.

We next vary the attack iterations in the re-performed PGD attack to show the model's robustness against unseen attacks with larger perturbations. As shown in Figure 3, when the perturbation size exceeds that used for adversarial training (10 iterations), the robust losses and accuracies begin to sharply increase and decrease, respectively.

Our findings indicate that the model overfits the threat model seen during AT, which has also been shown in (Stutz et al., 2020; Chen et al., 2021). We answer the question raised in Section 1 that due to the overparameterization of PrLMs, they can easily memorize the adversarial data generated during AT, resulting in robust overfitting. Thus the adversarially learned model cannot generalize well on the adversarial test set and the robust test loss continues to increase during robust testing.

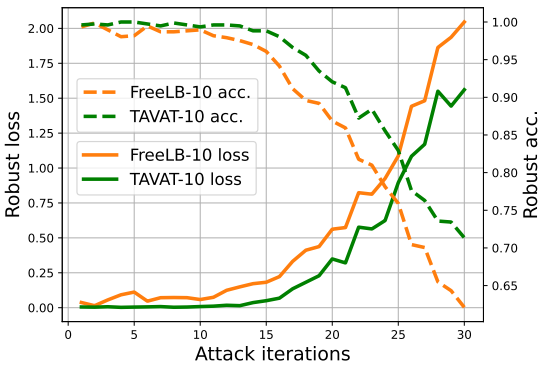


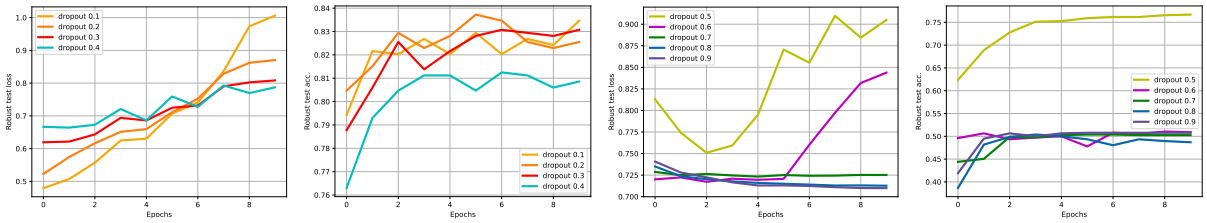
Figure 3: Robust loss and accuracy against re-performed PGD attacks on the SST-2 training set. The base model is BERT-base. The robust loss and accuracy begin to increase and decrease under ∼10 attack iterations, respectively. InfoBERT is excluded due to its abnormal robust loss and accuracy curves, as already shown in Figure 1 and Figure 2.

3.2 More Empirical Evidence

To better support our hypothesis, we provide more empirical evidence across different datasets and model architectures, which can be found in Appendix A.

4 Mitigating Robust Overfitting

In this section, we make several attempts to prevent PrLMs from getting overfitting in AT. In standard training, regularization methods can mitigate overfitting and promote test performance. Thus, it is intuitive to use regularization methods in AT to avoid robust overfitting.



(a) (b) (c) (d)

Figure 4: (a) and (b) Robust test loss and accuracy when the dropout ratio varies in [0.1, 0.4]. (c) and (d) Robust test loss and accuracy when the dropout ratio varies in [0.5, 0.9]. The AT method is FreeLB, and the base model is BERT-base.

4.1 Ensemble Methods

Dropout (Srivastava et al., 2014) randomly drops units from the model during training, which can be recognized as sampling from an exponential number of models. At test time, the model uses all the units to make predictions, which can be seen as an ensemble model.

Dropout is widely used in modern deep learning as a regularizer. We vary the dropout ratio for the attention probabilities and all the fully connected layers in the embeddings, encoder, and pooler for PrLMs. In this way, we aim to see whether dropout can mitigate robust overfitting and whether a large dropout ratio helps.

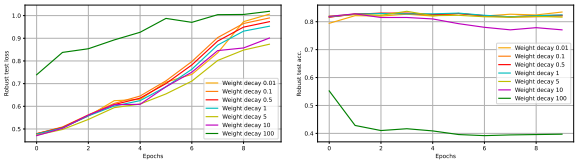
Figure 4(a) and Figure 4(b) show the robust test loss and accuracy when the dropout ratio is in [0.1, 0.4]. For different dropout ratios, the robust loss decreases as the ratio increases. However, the robust loss still increases as the epoch grows. The robust accuracy also decreases as the dropout ratio increases. In Figure 4(c), when the dropout ratio is in [0.7, 0.9], the robust test loss begins to decrease rather than increase. Nevertheless, we can observe in Figure 4(d) that the corresponding robust test accuracy maintains low because the robust loss is still large. It indicates that a large dropout ratio can hurt the robust test performance, though it ostensibly alleviates robust overfitting. This finding also suggests that a proper regularization technique may address robust overfitting.

4.2 Weight Decay

Weight decay (Krogh and Hertz, 1991) aims to adjust the effect of model complexity on the loss function, also known as L_2 regularization. It forces the parameters to converge to smaller values and avoids overfitting. Formally, weight decay adds a regularization term in the loss function as follows:

$$J = J_0 + \frac{\lambda}{2N} \sum_{w} w^2, \tag{2}$$

where J_0 is the original loss function, λ is the coefficient of the regularization term, N is the number of samples in the training set, and w is the set of model parameters.



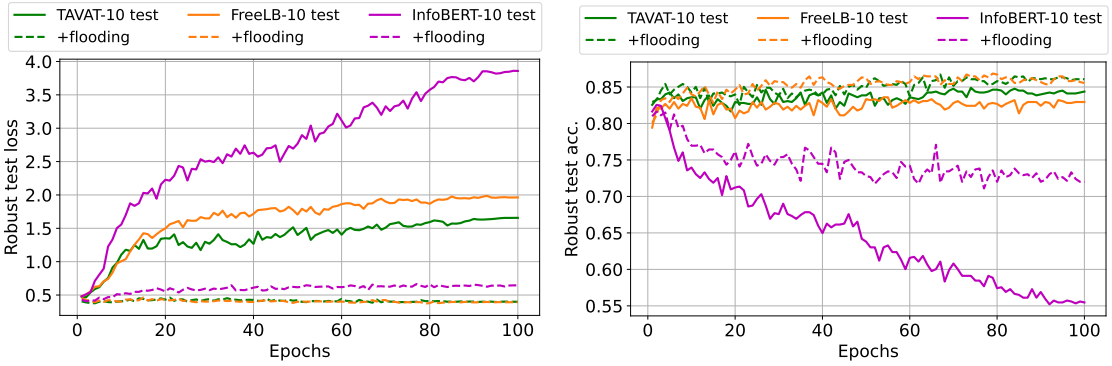
(a) (b)

Figure 5: Robust test loss and accuracy with different coefficient λ for weight decay. The AT method is FreeLB, and the base model is BERT-base.

To assess the effect of weight decay in mitigating robust overfitting, we vary the coefficient λ in a wide range and report the robust loss during testing. From Figure 5(a), we can observe that weight decay can not avoid robust overfitting in AT since the robust test loss continues to increase. Although a slightly larger λ (5 and 10) can make the robust test loss smaller in later epochs, a too-large λ increases the robust test loss overall. Figure 5(b) shows the robust test accuracy of different weight decay coefficients. Similarly, large coefficients hurt the robust accuracy, while small coefficients have little effect on robust accuracy.

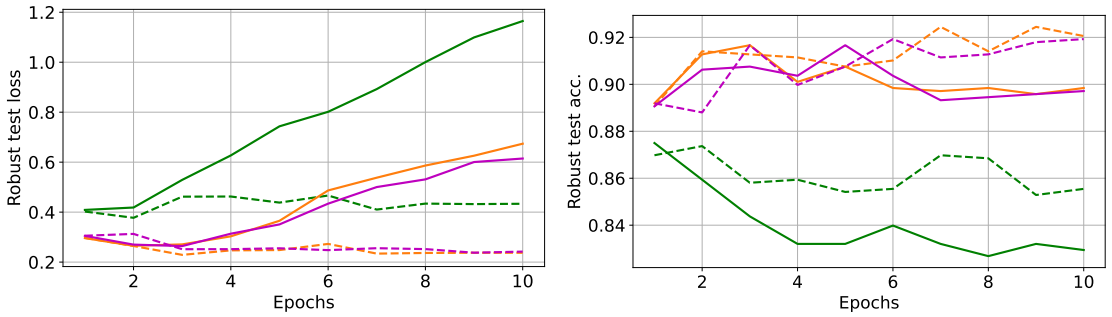
4.3 Flooding

Conventional regularization methods contribute little to alleviating robust overfitting. However, we have shown that proper regularization may help mitigate robust overfitting in Section 4.1. Recall



(a) SST2 & BERT-base

(b) SST2 & BERT-base



(c) AGNEWS & DeBERTa-v3-base

(d) AGNEWS & DeBERTa-v3-base

Figure 6: Robust test loss and accuracy against PGD-10 attacks with and without flooding. The flooding level is 0.1. Dashed lines represent the robust test loss and acc. with flooding, and solid lines represent the robust test loss and acc. without flooding. Due to the large time complexity, we train DeBERTa-v3-base models for 10 epochs.

Methods	Flooding level	Clean %	RA %
BERT-base (Devlin et al., 2019)	0	91.97	7.00
	0.0125	91.82	4.93
	0.025	91.76	7.49
	0.05	91.97	4.13

Table 1: Adversarial robustness of models trained using flooding only (w/o AT). We show that flooding can not boost adversarial robustness by itself. The AT method is FreeLB. "**Clean %**" is the accuracy on the clean test set, and "**RA %**" is the robust accuracy against adversarial attacks. The flooding level follows the original paper (Liu et al., 2022).

our hypothesis that the model memorizes all the adversarial training data and fails to generalize to robust testing. Following the philosophy to prevent the model from memorizing all the adversarial data, it is intuitive and reasonable to calibrate the model's prediction when it gets zero robust training loss. Thus, making the model less confident in some small-loss data is significant.

To this end, we propose to combine "flooding" with AT methods. Ishida et al. (2020) have found that flooding could help generalization. Flooding intentionally prevents further reduction of the training loss when it reaches a reasonably small value b as follows:

$$J = abs(J_0 - b) + b, \tag{3}$$

where J_0 is the original loss function, b is the flooding level, and $abs()$ is the absolute value function. Therefore, we expect that flooding can help mitigate robust overfitting in AT.

It is worth noting that Liu et al. (2022) have claimed that flooding could improve adversarial robustness without AT. However, through empirical experiments, we find that flooding, as a regularizer, can not promote adversarial robustness only by itself, which contradicts their results. We first show the adversarial robustness of models using flooding only. Then we combine flooding with AT methods, exploring its effect in avoiding robust overfitting and improving adversarial robustness.

Table 1 reports the models' adversarial robustness regularized by flooding against TextFooler (Jin et al., 2020). Although it is claimed that flooding can boost adversarial robustness without AT, our

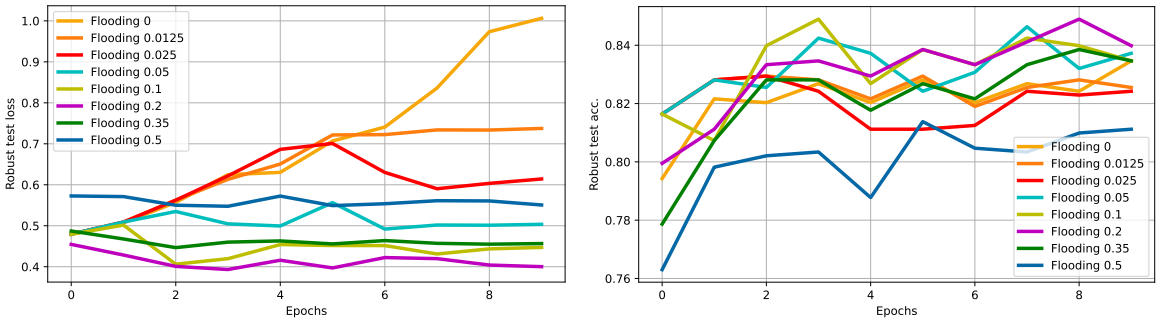


Figure 7: Robust test loss and accuracy of different flooding levels. The AT method is FreeLB, and the base model is BERT-base.

results indicate that flooding contributes little to adversarial robustness. We then combine flooding with AT to exploit its effect in avoiding overfitting (Ishida et al., 2020).

Figure 6 shows the robust test loss and accuracy against PGD attacks across datasets and model architectures. With flooding, the robust test loss of three AT methods maintains a low level and no longer increases as the epoch grows, indicating that flooding can mitigate robust overfitting and bridge the adversarial generalization gap. The robust accuracy also improves, verifying the regularization effect of flooding in AT.

5 Discussion

In this section, we discuss the effect of flooding in AT and investigate why flooding can help adversarial generalization.

5.1 Effect of Flooding Levels

We investigate the effect of flooding levels in mitigating robust overfitting. We vary the flooding level b from 0 to 0.5, and the corresponding robust test loss and accuracy are shown in Figure 7.

When the flooding level is set to 0, the robust test loss continues to increase, as we have shown in previous sections. As the flooding level grows, the overall robust loss decreases and reaches the minimum when the flooding level is 0.2. Larger flooding levels increase the robust loss. However, the robust loss curve no longer rises, which verifies that a reasonable flooding level not only helps alleviate robust overfitting issues but also promotes adversarial robustness. Regarding robust accuracy, similarly, we observe that a proper flooding level can boost the adversarial robustness against PGD attacks.

5.2 Memorization

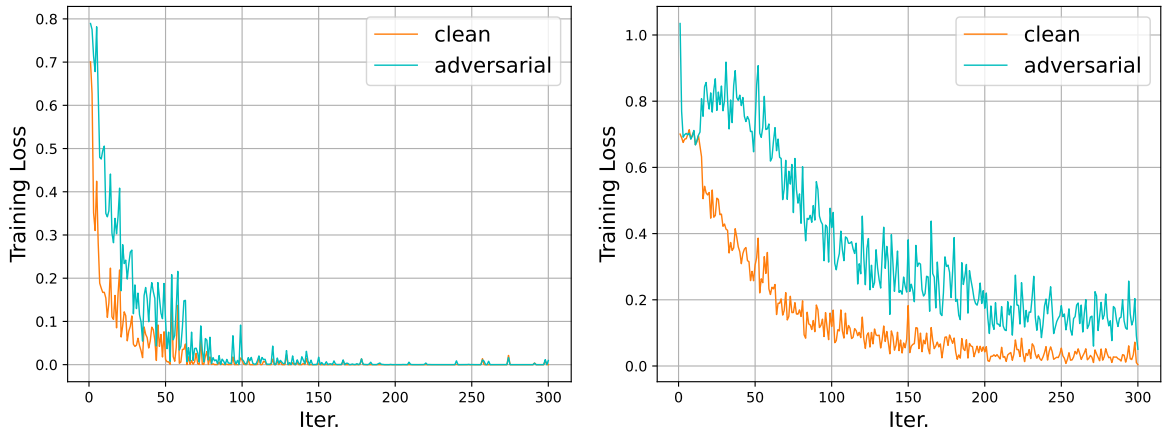
We first give an intuitive explanation of the memorization in AT from the perspective of loss magnitude. Figure 8 demonstrates that the adversarial loss without flooding dominates the training. Therefore the adversarially trained model gets robust overfitting. The learning curves of adversarial loss with flooding is not shown because its value can predictably fluctuate around the flooding level. We vary the adversarial search steps and report the results in Appendix B.

To verify our hypothesis that flooding can prevent the model from memorizing adversarial training data, we investigate if models can achieve zero training error when their training loss are scaling with flooding. We show in Figure 9(a) the learning curves of training accuracy with BERT-base on the SST2 dataset. We conclude that the model gives up on memorizing all the adversarial training data as the flooding level gets higher. In Figure 9(b) we report the learning curves of training accuracy with DeBERTa-v3-base on the AGNEWS dataset. Similarly, the model gives up on memorizing all the adversarial training data.

To conclude, we demonstrate that flooding can mitigate memorization in AT with several model architectures and datasets.

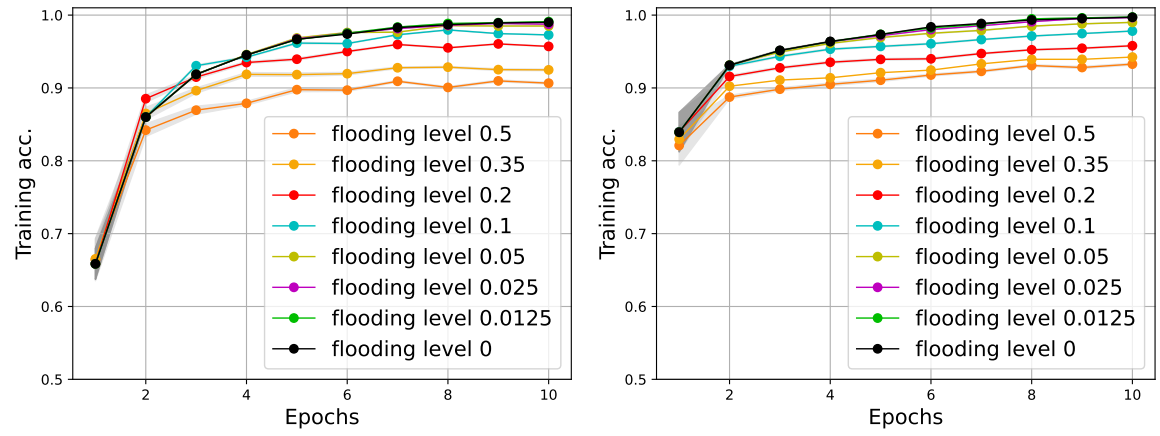
5.3 Robustness against Textual Adversarial Attacks

We have shown that flooding can mitigate robust overfitting against PGD attacks. To provide comprehensive evidence that flooding helps adversarial generalization, we evaluate the model's robustness against textual adversarial attacks.



(a) adv. search steps = 10 (b) adv. search steps = 50

Figure 8: The clean cross-entropy training loss and the adversarial loss with BERT-base on the SST2 dataset.



(a) BERT-base & SST2 (b) DeBERTa-v3-base & AGNEWS

Figure 9: Training accuracy of different flooding levels. The models are trained using FreeLB with BERT-base on the SST2 dataset. Note that a flooding level of zero means that the model is trained without flooding.

5.3.1 Experimental Setup

Datasets We conduct experiments on two widely used text classification datasets, SST2 (Socher et al., 2013)[2] and AGNEWS (Zhang et al., 2015)[3]. SST2 is a sentiment analysis dataset which contains 67349 training samples and 872 validation samples. We use the GLUE (Wang et al., 2019) version of the SST2 dataset. The average text length is 17. AGNEWS is a category classification dataset with four news topics: World, Sports, Business, and Science/Technology. It contains 12000 training samples and 7600 test samples. The average text length is 43. The maximum sentence length kept for the two datasets is 40. For training, we split 10% of the training set as the validation set.

Adversarial Training Methods We adopt three AT methods, FreeLB (Zhu et al., 2020), TAVAT (Li and Qiu, 2021), and InfoBERT (Wang et al., 2021a), as our AT baselines. The three AT methods help boost models' generalization ability and adversarial robustness. The adversarial settings are set consistently. The number of adversarial steps is 10; the step size is 0.01; the adversarial maximum norm is 1; the magnitude of initial adversarial perturbation is 0.02; and all the other settings follow their original papers.

Attacking Methods We adopt TextFooler (Jin et al., 2020), a word-level textual adversarial attacking method, as our attacking baseline. TextFooler is widely used in related literatures on adversarial

[2] https://dl.fbaipublicfiles.com/glue/data/SST-2.zip

[3] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

Methods	Clean %	RA %
BERT-base (Devlin et al., 2019)	91.97	7.00
+FreeLB (Zhu et al., 2020)	92.32	8.94
+FreeLB & flooding	92.66	11.93
+TAVAT (Li and Qiu, 2021)	92.66	14.56
+TAVAT & flooding	93.58	11.24
+InfoBERT (Wang et al., 2021a)	92.32	6.31
+InfoBERT & flooding	93.35	6.77

Table 2: Adversarial robustness against textual adversarial attacks. The dataset is SST-2. The attacking method is TextFooler. The flooding level is set to 0.1.

Methods	Clean %	RA %
DeBERTa-v3-base (He et al., 2021a)	93.10	12.60
+FreeLB (Zhu et al., 2020)	95.20	26.00
+FreeLB & flooding	94.70	26.60
+TAVAT (Li and Qiu, 2021)	91.90	25.60
+TAVAT & flooding	93.58	21.10
+InfoBERT (Wang et al., 2021a)	93.90	27.50
+InfoBERT & flooding	94.90	29.50

Table 3: Adversarial robustness against textual adversarial attacks. The dataset is AGNEWS. The attacking method is TextFooler. The flooding level is set to 0.1.

attacks and robustness. We use TextAttack's (Morris et al., 2020)[4] implementation of TextFooler to provide fair results.

Model Architectures We use BERT-base (Devlin et al., 2019) and DeBERTa-v3-base (He et al., 2021b,a) as our baseline models and load their weights from HuggingFace Transformers[5]. For these models, BERT-base has achieved great performance on NLP tasks as the first pre-trained language model. DeBERTa-v3-base is an advanced variant among the BERT family.

5.3.2 Attacking Results

Table 2 and Table 3 show the standard accuracy (**Clean %**) and robust accuracy (**RA %**) across datasets and model architectures.

On the SST2 dataset, all three AT methods obtain improvements in the standard accuracy. FreeLB and TAVAT boost the adversarial robustness compared with BERT-base, while InfoBERT has degraded robustness. Furthermore, flooding can promote robustness upon FreeLB and InfoBERT, but the combination of TAVAT and flooding has a relatively low robust accuracy compared with TAVAT.

For the AGNEWS dataset, all the combinations can promote standard accuracy except for TAVAT. Like the robust accuracy on the SST2 dataset, flooding can improve adversarial robustness upon FreeLB and InfoBERT while having a degraded robust accuracy compared with TAVAT.

It is an interesting observation that flooding can not boost adversarial robustness upon TAVAT against TextFooler. However, this work mainly focuses on mitigating robust overfitting issues against PGD attacks. This observation indicates that adversarial generalization gaps exist when the model defends against different attacks (e.g., PGD-based attacks and word-level textual adversarial attacks). It may be the generalization gap caused by TAVAT itself. Overall, we leave this question for another promising direction of future work.

It is also surprising that InfoBERT can not promote robustness on the SST2 dataset with the BERT-base architecture. This may be because we fix the attack iterations to 10 instead of using the settings in the original paper.

6 Conclusion

Robust overfitting prevents further improvement of adversarial robustness on PrLMs. While we adopt strong regularizers in AT, weight decay and dropout contribute little to mitigating robust overfitting. To prevent the model from simply memorizing the adversarial training data, we combine flooding with AT. Experimental results on extensive datasets and model architectures demonstrate that a reasonable flooding level helps mitigate robust overfitting.

As a preliminary study, this work identifies the robust overfitting issue for PrLMs. We hope the community can take robust overfitting into account when performing AT to achieve adversarially robust models.

Limitations

In this work, we mainly identify robust overfitting for PrLMs using PGD attacks instead of textual adversarial attacks. The reasons are two folds. First, we aim to check the learning curves during AT. Second, the results of textual adversarial attacks may not be generalizable since they integrate different strategies. In practice, however, it is more inclined to use some textual adversarial attack methods (e.g., TextFooler, TextBugger (Li et al., 2019)) to evaluate the robustness of NLP models. As we have

[4] https://github.com/QData/TextAttack
[5] https://huggingface.co/transformers

clarified in Section 5.3.2, there exists an adversarial generalization gap when the model defends against PGD-based gradient attacks and textual adversarial attacks. While it is difficult to check their robust loss and accuracy curves during AT, it is necessary and promising to explore robust overfitting under textual adversarial attacks and provide helpful insights for promoting the adversarial robustness of PrLMs.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. This work has been supported by the National Natural Science Foundation of China (61972426).

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, pages 15849–15854.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. 2021. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chengyu Dong, Liyuan Liu, and Jingbo Shang. 2022a. Label noise in adversarial training: A novel perspective to study robust overfitting. In *Advances in Neural Information Processing Systems*, pages 17556–17567.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021a. Towards robustness against natural language word substitutions. In *International Conference on Learning Representations*.

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, and Hanwang Zhang. 2021b. How should pre-trained language models be fine-tuned towards adversarial robustness? In *Advances in Neural Information Processing Systems*, pages 4356–4369.

Yinpeng Dong, Ke Xu, Xiao Yang, Tianyu Pang, Zhijie Deng, Hang Su, and Jun Zhu. 2022b. Exploring memorization in adversarial training. In *International Conference on Learning Representations*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4083–4093, Hong Kong, China. Association for Computational Linguistics.

Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. 2020. Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th International Conference on Machine Learning*, pages 4604–4614.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. 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*, pages 8018–8025.

Anders Krogh and John Hertz. 1991. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial machine learning at scale. In *International Conference on Learning Representations*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021a. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5053–5069, Online. Association for Computational Linguistics.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Linyang Li and Xipeng Qiu. 2021. Token-aware virtual adversarial training in natural language understanding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 8410–8418.

Zongyi Li, Jianhan Xu, Jiehang Zeng, Linyang Li, Xiaoqing Zheng, Qi Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021b. Searching for an effective defender: Benchmarking defense against adversarial word substitution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3137–3147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qin Liu, Rui Zheng, Bao Rong, Jingyi Liu, ZhiHua Liu, Zhanzhan Cheng, Liang Qiao, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Flooding-X: Improving BERT's resistance to adversarial attacks via loss-restricted fine-tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5634–5644, Dublin, Ireland. Association for Computational Linguistics.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2017. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *CoRR*, abs/1704.03976.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. 2017. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5947–5956.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016a. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387.

Nicolas Papernot, Patrick D. McDaniel, Ananthram Swami, and Richard E. Harang. 2016b. Crafting adversarial input sequences for recurrent neural networks. *CoRR*, abs/1604.08275.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. 2021. Data augmentation can improve robustness. In *Advances in Neural Information Processing Systems*, pages 29935–29948.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Leslie Rice, Eric Wong, and J. Zico Kolter. 2020. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8093–8104.

Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4323–4330.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pages 3358–3369.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness verification for transformers. In *International Conference on Learning Representations*.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages

1569–1576, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

David Stutz, Matthias Hein, and Bernt Schiele. 2020. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9155–9166.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021a. Info{bert}: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*.

Wenjie Wang, Pengfei Tang, Jian Lou, and Li Xiong. 2021b. Certified robustness to word substitution attack with differential privacy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1102–1112, Online. Association for Computational Linguistics.

Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. 2021c. Adversarial training with fast gradient projection method against synonym substitution based text attacks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13997–14005.

Eric Wong, Leslie Rice, and J. Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, pages 2958–2969.

Yichen Yang, Xiaosen Wang, and Kun He. 2022. Robust textual embedding against word-level adversarial attacks. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, Proceedings of Machine Learning Research, pages 2214–2224.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3465–3475, Online. Association for Computational Linguistics.

Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. 2022. Understanding robust overfitting of adversarial training and beyond. In *Proceedings of the 39th International Conference on Machine Learning*, pages 25595–25610.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. 2019a. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, pages 227–238.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019b. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7472–7482.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan S. Kankanhalli. 2020. Attacks which do not kill training make adversarial learning stronger. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11278–11287.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.

Bin Zhu, Zhaoquan Gu, Le Wang, Jinyin Chen, and Qi Xuan. 2022. Improving robustness of language models from a geometry-aware perspective. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3115–3125, Dublin, Ireland. Association for Computational Linguistics.

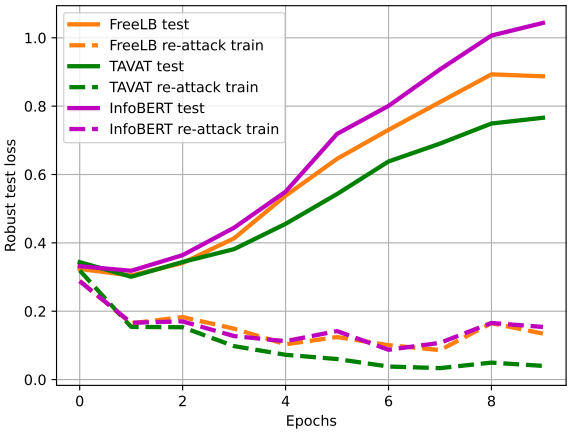
Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

A More Evidence for Robust Overfitting

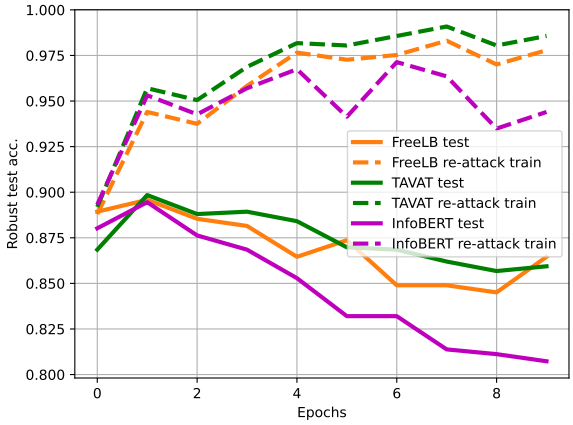
We provide more results across datasets and model architectures to identify robust overfitting for PrLMs in Figure 10, which also empirically verifies our hypothesis that the model's memorization of adversarial training data results in robust overfitting.

B Comparison of the Clean Cross-entropy Loss and the Adversarial Loss

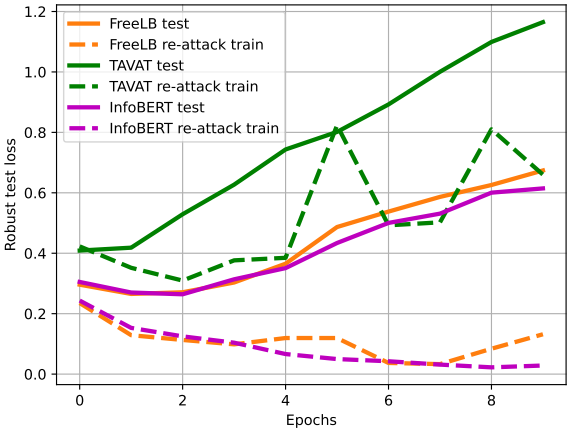
We vary the adversarial search steps and report the clean cross-entropy loss and the adversarial loss during training. In Figure 11, the adversarial loss becomes higher as the number of search steps gets larger, which implies that the adversarial loss dominates the training, leading to robust overfitting.



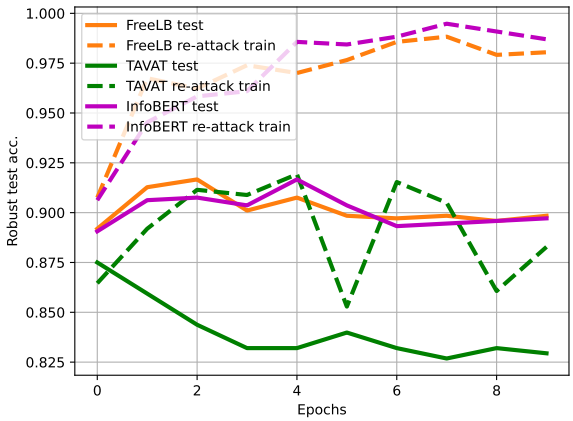
(a) BERT-base & AGNEWS



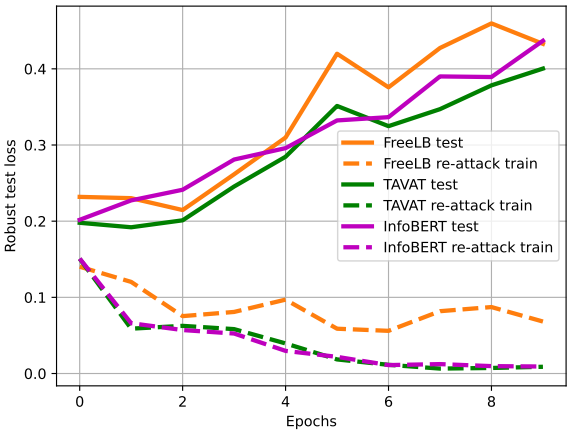
(b) BERT-base & AGNEWS



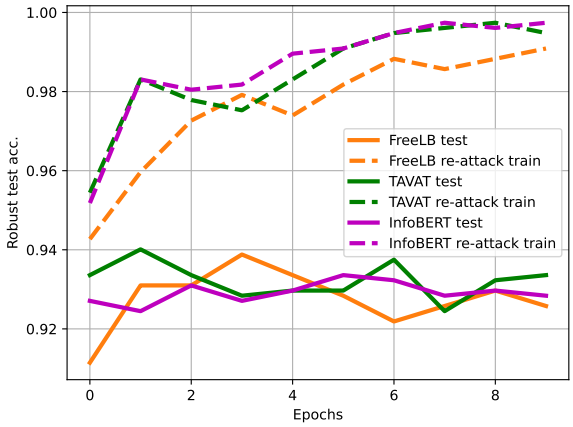
(c) DeBERTa-v3-base & AGNEWS



(d) DeBERTa-v3-base & AGNEWS

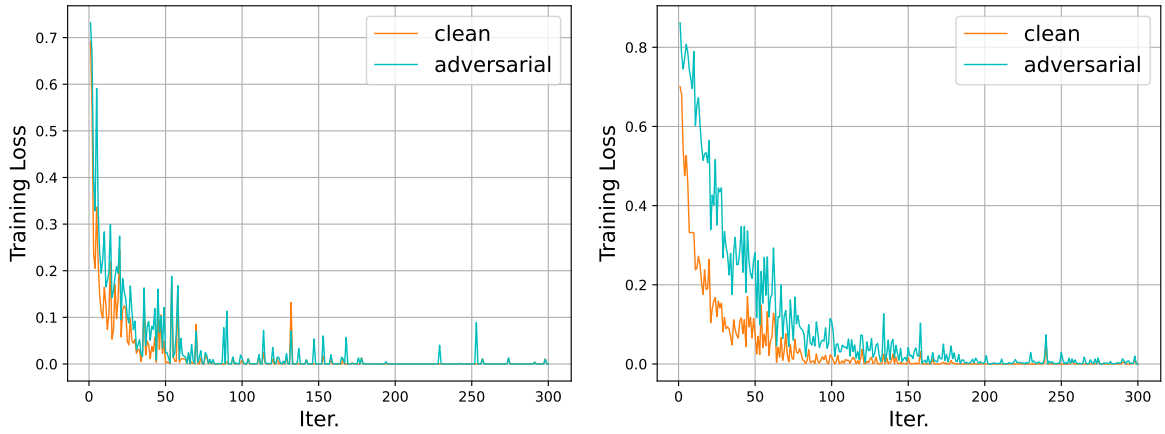


(e) DeBERTa-v3-base & SST2



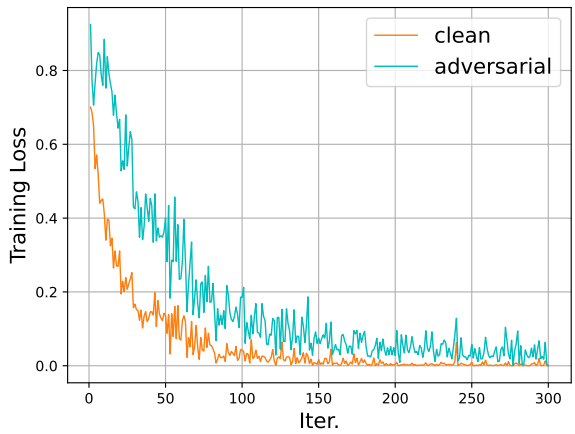
(f) DeBERTa-v3-base & SST2

Figure 10: Robust test loss and accuracy across datasets and model architectures.

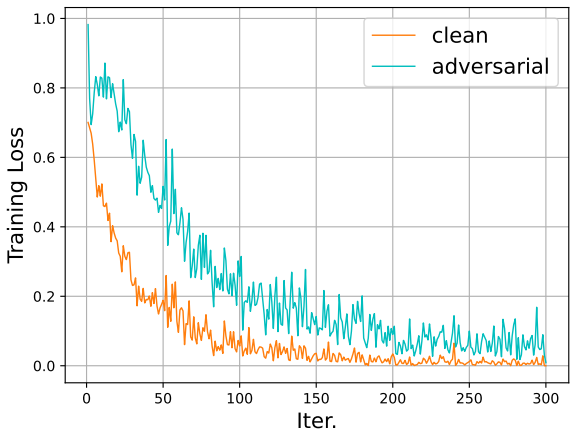


(a) adv. search steps = 3

(b) adv. search steps = 20



(c) adv. search steps = 30



(d) adv. search steps = 40

Figure 11: The clean cross-entropy training loss and the adversarial loss with BERT-base on the SST2 dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7.

☒ A2. Did you discuss any potential risks of your work?
There are no potential risks in this work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1.

☑ A4. Have you used AI writing assistants when working on this paper?
We use Grammarly to check this paper for all the sections.

B ☑ Did you use or create scientific artifacts?

Section 5.

☑ B1. Did you cite the creators of artifacts you used?
Section 5.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We would discuss the license in our codes.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 5.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Not applicable.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Not applicable.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.

C ☑ Did you run computational experiments?

Sections 4 and 5.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In this work, we do not focus on these terms, and we mainly discuss the performance gained.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Following previous work, we provide results with a single run.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.