Robust Tuning of Pre-trained Language Models: a Parameter-efficient Approach

Anonymous ACL submission

Abstract

Fine-tuning pre-trained language models (PLMs) has demonstrated remarkable performance in downstream tasks. These models, however, are vulnerable to adversarial attacks. Defenses based on adversarial fine-tuning, i.e., fine-tuning PLMs with adversarial examples, have been proposed to counter this vulnerability. However, such defenses suffer from unsatisfactory performance due to catastrophic forgetting, meaning they fail to retain the robust features learned during pre-training. In this paper, we propose a novel parameter-efficient adversarial fine-tuning method that tunes only a small subset of the model's parameters, leaving the majority intact. Our method involves training a defense soft prompt prepended to inputs, which leads to robust predictions by PLMs. Our extensive experiments demonstrate the effectiveness of our proposed defenses across various benchmarks and PLMs.

1 Introduction

Pre-trained Language Models (PLMs) have revolutionized natural language processing by shifting the paradigm from traditional supervised learning, which involves training task-specific models from scratch, to adapting general-purpose PLMs for specific downstream tasks through fine-tuning. Despite their remarkable performance, fine-tuned PLMs are vulnerable to adversarial examples; carefully crafted sentences with changes that are imperceptible to humans and cause misclassifications by classifiers (Zhang et al., 2020; Jin et al., 2020; Moraffah and Liu, 2024). Such attacks compromise the trustworthiness of these models, particularly in high-stakes applications, highlighting the urgent need for developing defense methods.

Adversarial training, which involves training models on adversarial examples, is a defense strategy designed to enhance the robustness of classifiers against such attacks and has demonstrated optimal robust performance (Lin et al., 2024).

In the context of PLMs, since training from scratch is infeasible, adversarial training is implemented through fine-tuning on these examples, commonly referred to as adversarial fine-tuning (Dong et al., 2021a; Jiang et al., 2022a). However, unlike traditional adversarial training, adversarial fine-tuning often leads to subpar performance, primarily due to catastrophic forgetting, i.e., the loss of robust features learned during pre-training. This issue arises from the inherent nature of fine-tuning and the requirement for adversarial fine-tuning to be conducted over several epochs (Dong et al., 2021a).

We propose a novel parameter-efficient adversarial fine-tuning method that freezes the pre-trained model parameters and only tunes a much smaller set of parameters. By altering only a small subset of the model's parameters, our proposed defense ensures that the core features learned by the pre-trained weights are largely preserved, thus alleviating catastrophic forgetting. This parameter-efficient tuning helps the model maintain the robust features learned during the pre-training while learning the necessary information from adversarial examples and adapting to the downstream task simultaneously. In particular, we propose a defense soft prompt that limits learned parameters to a set of virtual tokens prepended to the text input. Our soft prompt is trained with a min-max adversarial objective, which ensures when combined with the input, the soft prompt effectively guides the PLM to select the robust path and make robust decisions to adversarial attacks. While parameter-efficient fine-tuning has been extensively explored for low-resource scenarios (Han et al., 2024), to the best of our knowledge, our method is the first to explore its role in adversarial defense. Our experiments validate superior performance of our defense compared to state-of-the-art adversarial fine-tuning methods on several benchmarks.

2 Related Work

Several types of adversarial defenses for text, including adversarial purification (Moraffah et al., 2024), certified robustness (Wang et al., 2021b), manifold-based defenses (Minh and Luu, 2022), and adversarial training (Zhu et al., 2020; Jiang et al., 2022a) have been developed. Among all defense methods, adversarial training is known to be the most effective and promising strategy to improve the adversarial robustness of models (Jiang et al., 2022a). In the context of PLMs, adversarial training appears as adversarial fine-tuning, which fine-tunes the pre-trained models on adversarial examples. Existing adversarial fine-tuning methods, either greedily fine-tune PLMs on adversarial attacks (Zhu et al., 2020), or selectively fine-tune the models on samples that carry robust information (Jiang et al., 2022b; Dong et al., 2021b). These methods overfit the adversarial attack they are trained on, resulting in catastrophic forgetting of robust and informative features learned during the pre-training and thus low robust accuracy.

3 Methodology

To address catastrophic forgetting of adversarial fine-tuning on PLM, we propose a **P**arameter-**E**fficient **A**dversarial **F**ine-**T**uning (PEAFT) defense, which learns a defense soft prompt that, when prepended to PLM, results in robust predictions (cf. Figure 1). Since our method only learns a few parameters while freezing the pre-trained weights, it alleviates catastrophic forgetting.

3.1 Preliminaries

Soft Prompting Tuning is a parameter-efficient tuning technique that integrates k virtual tokens $\{p_1, p_2, \ldots, p_k\}$ as learnable embedding vectors to adapt the PLM to the downstream task (Lester et al., 2021). These tokens are prepended to the embedding representations of the input tokens. During the fine-tuning, instead of updating all model parameters, only the k virtual token embedding vectors are updated. Formally, for an input sequence $X = \{x_1, x_2, \ldots, x_q\}$, the embeddings are derived by prepending k randomly initialized soft prompts to the input sequence. Let $E(x)$ denote the embedding function. The initial embeddings, $\mathbf{E}_{\text{init}}$, are thus defined as: $\mathbf{E}_{\text{init}} = [p_1, p_2, \ldots, p_k, E(x_1), E(x_2), \ldots, E(x_q)]$ where each p_i is a vector in $\mathbb{R}^d$, and d is the embedding dimension. Soft prompt tuning of the PLMs is

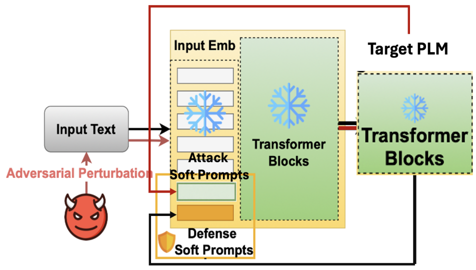


Figure 1: An overview of the proposed PEAFT. The defense soft prompt learned by PEAFT prepends to PLM and guides the model to correct predictions for adversarial examples.

then achieved by $\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f_{p,\theta}(x_i), y_i)$, where $\mathcal{L}$ is the Cross-Entropy loss, and $f_{p,\theta}$ represents the PLM parameterized by the soft prompts p, and the original parameters θ that are frozen.

Adversarial Training is a defense mechanism that aims to minimize the worst-case training loss for the adversarial examples (Madry et al., 2018). This is formulated via a min-max objective defined as $\min_\theta \max_{\|\delta\|\le\epsilon}\mathcal{L}(f_\theta(x_{adv}), y)$, where the inner maximization is responsible for generating adversarial examples that are crafted by learning and adding a small perturbation δ to the original input $\mathbf{x}$: $\mathbf{x}_{\text{adv}} = \mathbf{x} + \delta$. The model's parameters are then learned to minimize the training loss over these adversarial examples. In particular, each training epoch of adversarial training consists of two steps: (1) Generation of adversarial examples through solving $\delta^* = \arg\max_{\|\delta\|\le\epsilon}\mathcal{L}(f_\theta(\mathbf{x}+\delta), y)$; and (2) Updating the model parameters via $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}(f_\theta(\mathbf{x}+\delta^*), y)$, where η is the learning rate.

3.2 Proposed Defense

To mitigate the catastrophic forgetting caused by adversarial fine-tuning, we propose a parameter-efficient adversarial fine-tuning method that freezes the pre-trained model parameters and only tunes a much smaller set of parameters. Our proposed method consists of a defense soft prompt that limits its learned parameters to a set of virtual tokens prepended to the text input. Our soft prompt is trained with a min-max adversarial objective which ensures when combined with the input $\min_p \max_{\|\delta\|\le\epsilon}\mathcal{L}(f_{p,\theta}(x_{adv}), y)$. In this objective, only p (the soft prompt) is learned, where p is significantly smaller than θ. By preserving the pre-trained model's parameters θ, we effectively retain the robust and informative features learned during

Method	Target	MRPC			QNLI			RTE			SST2		
		AUA	ACC	AVG	AUA	ACC	AVG	AUA	ACC	AVG	AUA	ACC	AVG
PEAFT	roberta-base	**70.00**	73.20	**71.60**	32.50	**92.30**	**62.40**	**49.50**	68.50	**59.00**	**48.00**	**93.80**	**70.90**
	deberta-v3-base	**64.41**	**86.27**	**75.34**	32.22	**93.18**	**62.70**	**55.24**	**80.14**	**67.69**	**34.17**	94.61	**64.39**
	electra-base	**67.43**	**88.97**	**78.20**	**28.26**	91.61	**59.94**	**57.36**	**73.29**	**65.33**	**36.76**	**94.27**	**65.52**
	roberta-large	**66.23**	**87.74**	**76.99**	**36.52**	**93.68**	**65.10**	**55.78**	**85.19**	**70.49**	**41.50**	94.08	**67.79**
	deberta-v3-large	**68.41**	**90.44**	**79.43**	**34.72**	**94.40**	**64.56**	**58.42**	**88.81**	**73.62**	**42.32**	**95.30**	**68.81**
	electra-large	**71.22**	**89.48**	**80.35**	**33.60**	**94.10**	**63.85**	**62.23**	**88.09**	**75.16**	**50.91**	95.15	**73.03**
	Llama	**52.89**	**95.11**	**74.00**	**28.90**	**89.16**	**59.03**	**66.82**	**89.32**	**78.07**	**37.51**	**98.81**	**68.18**
FreeLB	roberta-base	12.23	**74.34**	43.29	29.00	84.10	56.55	16.41	**69.11**	42.76	12.30	82.50	47.40
	deberta-v3-base	14.11	81.21	47.66	14.12	88.10	51.11	8.12	71.50	39.81	13.76	**94.95**	54.36
	electra-base	8.90	78.91	43.91	19.11	**91.73**	55.42	12.10	69.01	40.56	11.58	93.81	52.70
	roberta-large	11.76	87.50	49.63	12.63	**93.68**	53.16	1.44	80.14	40.79	11.47	**94.84**	53.16
	deberta-v3-large	7.35	87.50	47.43	OOM	OOM	OOM	OOM	OOM	OOM	21.90	94.15	58.03
	electra-large	29.95	63.20	46.58	10.71	48.26	29.49	7.58	81.23	44.41	8.60	**95.41**	52.01
	Llama	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
ROSE	roberta-base	28.12	71.14	49.63	**37.12**	85.10	61.11	18.87	66.81	42.84	31.64	84.12	57.88
	deberta-v3-base	31.60	80.91	56.26	**33.12**	85.31	59.22	19.10	73.00	46.05	15.81	84.17	49.99
	electra-base	25.70	71.61	48.66	17.81	88.40	53.11	31.28	61.17	46.23	26.79	88.00	57.40
	roberta-large	26.94	83.78	55.36	18.39	87.13	52.76	15.10	71.97	43.54	19.75	86.39	53.07
	deberta-v3-large	22.00	70.26	46.13	12.96	76.94	44.95	22.00	72.98	47.49	19.65	78.46	49.06
	electra-large	27.81	60.10	43.96	22.90	64.67	43.79	24.64	78.35	51.50	15.47	84.43	49.95
	Llama	11.26	90.20	50.73	25.50	80.69	53.09	32.87	82.26	57.56	9.94	89.66	49.80

Table 1: Comparison of the proposed PEAFT with SOTA defense on the GLUE dataset attacked by Textfooler.

the pre-training. As mentioned earlier, solving the min-max objective consists of two steps, i.e., generation of adversarial examples and updating PLM parameters. Therefore, we split our soft prompt into two sets of tokens, the first part called the defense soft prompt is in charge of defense (learned by the outer minimization) and the second part, called the attack soft prompt is responsible for the attack generation (learned via inner maximization).

Learning the Attack Soft Prompt. The objective is to learn a soft prompt that when prepended to the input generates its corresponding adversarial example. This is achieved by solving the inner maximization of the adversarial training objective. The solution to the optimization is provided by the Projected Gradient Descent (PGD) (Madry et al., 2017). The perturbation δ is calculated based on the input mask M derived from the attention mask of the inputs. For l_2 norm initialization, the perturbation is initialized with random values scaled by the input mask. The magnitude of the perturbation is then adjusted based on the dimensions of the embeddings $\delta_0 = (\mathrm{U}(-1,1) \cdot M) \cdot \frac{\epsilon_{\text{init}}}{\sqrt{q \cdot d}}$, where $\mathrm{U}(-1,1)$ denotes the uniform distribution between -1 and 1, M is the input mask, q is the input length, d is the embedding dimension, and ϵ_{init} is the initial perturbation magnitude, which is a hyperparameter. In the k-th adversarial iteration, the embeddings are updated using $\mathbf{E}_{adv}^{(k)} = \mathbf{E}_{\text{init}} + \delta_k$, and $\delta_k = \epsilon \cdot \text{sign}(\nabla_{\mathbf{E}} \mathcal{L}_{\text{adv}})$.

Learning the Defense Soft Prompt. Once the adversarial soft prompt is learned, it is prepended to the input which acts as an adversarial example. The defense soft prompt is then learned by minimizing the training loss over these examples. To further ensure the performance over benign examples, we utilize a weighted combination of losses over benign ($\mathcal{L}_{\text{bgn}}$) and adversarial examples ($\mathcal{L}_{\text{adv}}$). The final objective of our framework is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bgn}} + \lambda \cdot \mathcal{L}_{\text{adv}}$, where λ is a hyperparameter controlling impact of adversarial loss.

4 Experiments

4.1 Experimental Setup

Datasets, Targets, and Baselines. Following previous research (Zhu et al., 2020; Jiang et al., 2022b), we utilize *six* widely-used datasets for our experiments. We use four tasks from the GLUE (Wang et al., 2018): MRPC, QNLI, RTE, and SST2, and attack them with TextFooler (Jin et al., 2020), one of the strongest adversarial attacks. We also evaluate our defense on the AdvGLUE (Wang et al., 2021a), which is designed to evaluate the vulnerabilities of modern LLMs under various types of adversarial attacks. We test our defense across various target model sizes and model backbones: RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), and Electra (Clark, 2020). For all models, we adopt base and large variants. For the sake of comprehensiveness, we also use Llama (using a sequence classification head) (Touvron et al., 2023). Our baselines are: *(1) FreeLB* (Zhu et al., 2020): an adversarial fine-tuning method, which finetunes on adversarial examples generated by adding perturbations to the embeddings; and *(2) ROSE* (Jiang et al., 2022b): a fine-tuning method that selectively fine-tunes the PLMs on robust samples[1].

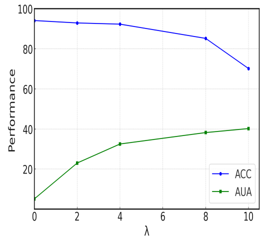
Method	Target	RTE AUA	RTE ACC	RTE AVG	SST2 AUA	SST2 ACC	SST2 AVG
PEAFT	roberta-base	48.20	**68.501**	58.35	**66.14**	93.86	**80.00**
	deberta-v3-base	59.82	80.14	69.98	**64.58**	94.61	**79.60**
	electra-base	**62.27**	73.29	**67.78**	**66.02**	**94.27**	**80.15**
	roberta-large	**71.59**	**85.19**	**78.39**	**60.92**	94.08	**77.50**
	deberta-v3-large	**72.71**	**88.81**	**80.76**	**68.64**	95.30	**81.97**
	electra-large	70.37	**88.09**	79.23	61.35	95.15	78.25
FreeLB	roberta-base	**64.19**	58.12	**61.15**	39.18	93.11	66.15
	deberta-v3-base	**65.43**	**81.58**	**73.51**	52.7	**94.95**	73.83
	electra-base	43.21	74.0	58.61	44.59	93.81	69.20
	roberta-large	56.79	80.14	68.47	50.67	94.83	72.75
	deberta-v3-large	OOM	OOM	OOM	47.97	94.15	71.06
	electra-large	69.13	81.22	75.18	**63.51**	**95.41**	**79.46**
ROSE	roberta-base	35.49	78.34	56.92	37.67	**94.84**	66.26
	deberta-v3-base	32.09	78.26	55.18	39.5	90.76	65.13
	electra-base	31.85	**75.74**	53.80	42.72	90.37	66.55
	roberta-large	70.62	85.13	77.88	57.77	**95.58**	76.68
	deberta-v3-large	70.5	83.71	77.11	52.61	**95.50**	74.10
	electra-large	**77.82**	85.71	**81.77**	59.64	93.2	76.42

Table 2: Comparison of the proposed PEAFT with SOTA defenses on the AdvGLUE dataset.

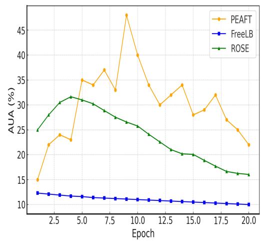
Evaluation Metrics. We report the accuracy under the attack (AUA), which measures the model's accuracy on adversarial examples, and the accuracy of benign samples from the test set (ACC). To assess the final performance on adversarial and begin examples, we report the average accuracy (AVG).

4.2 Experimental Results

Comparison with State-of-the-art Defenses. We compare our proposed PEAFT with SOTA adversarial fine-tuning defenses and report the results in Table 1 and 2. We observe that PEAFT consistently outperforms the SOTA defenses in terms of both accuracy under the attack (AUA) and benign accuracy (ACC) by a large margin. We can also observe that the average accuracy on both benign and adversarial examples obtained by PEAFT is significantly higher than the baselines. In the following, we elaborate on our in-depth observations: (1) FreeLB exhibits poor performance in all cases. This is due to fine-tuning the model on any adversarial examples, resulting in catastrophic forgetting of the robust features learned during the pre-training; (2) due to its selective fine-tuning strategy, ROSE obtains higher AUA compared to FreeLB. However, due to the low occurrence of updates for robust samples, it overfits the adversarial perturbations, resulting in lower ACC compared to PEAFT; and (3) the proposed PEAFT achieves over 30% higher AUA compared to the best-performing baseline. Note that the FreeLB's poor performance which uses the same training objective as ours but to fine-tune the entire model, further emphasizes the role of soft prompt in alleviating the catastrophic forgetting and obtaining higher AUA and AVG.



(a) λ vs. performance on the QNLIdataset



(b) AUA vs. Epochs on SST2 dataset

Figure 2: PEAFT's Behavior Analysis.

Hyperparameter Analysis. We demonstrate effect of the adversarial loss and trade-off between the original accuracy and accuracy under the attack by varying the hyperparameter λ in the attack objective. $\lambda = 0$ indicates normal training on benign examples. As shown in Figure 2(a), as λ increases the AUA and ACC increases and decreases, respectively. This shows trade-off between accuracy on benign samples and accuracy under the attack, while emphasizing that incorporating the adversarial loss indeed leads to learning robust features.

Analysis of the Model Performance over Epochs. To demonstrate the effectiveness of defense soft prompt on alleviating catastrophic forgetting, we plot the AUA for PEAFT and the baselines trained over different number of epochs. As shown in the Figure 2(b), both FreeLB and ROSE exhibit mostly decreasing trend over different epochs, indicating that their training results in forgetting robust information thus a decrease in AUA. Our proposed defense, on the other hand, learns robust features over epochs, resulting in increasing AUA.

5 Conclusion

We propose a parameter-efficient adversarial fine-tuning method that addresses catastrophic forgetting while improving robustness against adversarial examples. Our approach, based on defense soft prompting, enhances PLM robustness without compromising pre-trained knowledge. Experiments show significant improvements across benchmarks.

[1]Implementation will be made public upon acceptance.

6 Limitations

Building upon the foundational studies of adversarial training for text (e.g., (Zhu et al., 2020)), the defense mechanism proposed in this paper also entails the generation of adversarial attacks within the continuous embedding space. However, this approach may not represent the most optimal strategy to generate worst-case adversarial attacks. The primary focus of this research is to tackle the issue of catastrophic forgetting, with the exploration of more optimal adversarial attacks being earmarked for future work. Moreover, the defense method proposed in this paper is specifically tailored for classification tasks that utilize discriminative Pre-trained Language Models (PLMs). For tasks that involve the use of generative LLMs, there is a distinct necessity to devise alternative defensive strategies tailored to those models.

References

K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, and Hanwang Zhang. 2021a. How should pre-trained language models be fine-tuned towards adversarial robustness? *Advances in Neural Information Processing Systems*, 34:4356–4369.

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*, volume 34, pages 4356–4369. Curran Associates, Inc.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Lan Jiang, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, and Rui Jiang. 2022a. ROSE: Robust selective fine-tuning for pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2897, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lan Jiang, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, and Rui Jiang. 2022b. Rose: Robust selective fine-tuning for pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2897.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Guang Lin, Chao Li, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. 2024. Adversarial training on purification (ATop): Advancing both robustness and generalization. In *The Twelfth International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks.

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*.

Dang Nguyen Minh and Anh Tuan Luu. 2022. Textual manifold-based defense against natural language adversarial examples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6612–6625.

Raha Moraffah, Shubh Khandelwal, Amrita Bhattacharjee, and Huan Liu. 2024. Adversarial text purification: A large language model approach for defense. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 65–77. Springer.

Raha Moraffah and Huan Liu. 2024. Exploiting class probabilities for black-box sentence-level attacks. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1557–1568, St. Julian's, Malta. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021a. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

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*.