ROSE: Robust Selective Fine-tuning for Pre-trained Language Models

Lan Jiang[1]*, **Hao Zhou**[2]*, **Yankai Lin**[3,4]†, **Peng Li**[5]†, **Jie Zhou**[2], **Rui Jiang**[1]‡
[1]MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, Department of Automation, BNRist, Tsinghua University, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc., China
[3]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[4]Beijing Key Laboratory of Big Data Management and Analysis Methods , Beijing, China
[5]Institute for AI Industry Research (AIR), Tsinghua University, China
jiangl20@mails.tsinghua.edu.cn

Abstract

Even though the large-scale language models have achieved excellent performances, they suffer from various adversarial attacks. A large body of defense methods has been proposed. However, they are still limited due to redundant attack search spaces and the inability to defend against various types of attacks. In this work, we present a novel fine-tuning approach called **RO**bust **SE**letive fine-tuning (**ROSE**) to address this issue. ROSE conducts selective updates when adapting pre-trained models to downstream tasks, filtering out invaluable and unrobust updates of parameters. Specifically, we propose two strategies: the first-order and second-order ROSE for selecting target robust parameters. The experimental results show that ROSE achieves significant improvements in adversarial robustness on various downstream NLP tasks, and the ensemble method even surpasses both variants above. Furthermore, ROSE can be easily incorporated into existing fine-tuning methods to improve their adversarial robustness further. The empirical analysis confirms that ROSE eliminates unrobust spurious updates during fine-tuning, leading to solutions corresponding to flatter and wider optima than the conventional method. Code is available at https://github.com/jiangllan/ROSE.

1 Introduction

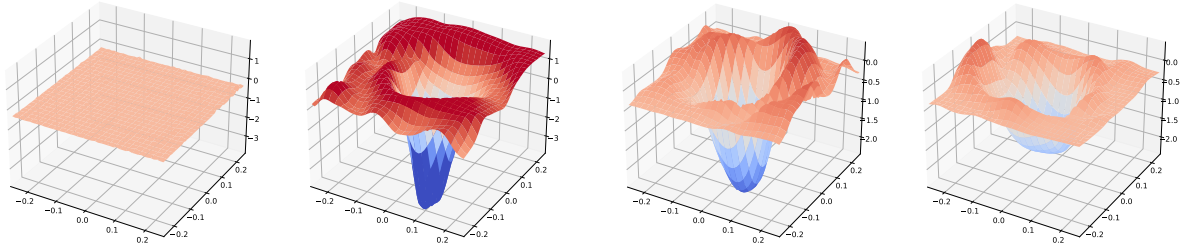
Recently, the discipline of fine-tuning large-scale pre-trained language models has gained prominence, achieving remarkable performances across various natural language processing benchmarks. However, recent studies (Ribeiro et al., 2020; Jin et al., 2020; Nie et al., 2020; Lin et al., 2021; Jiang et al., 2022) have highlighted the lack of adversarial robustness in models fine-tuned on specific downstream tasks, *i.e.*, adapted models are vulnerable to various types of adversarial attacks. A majority of adversarial examples fool models via character or word-level perturbations, either those tokens that aren't found in training sets or superficial cues attached to labels (Li et al., 2021b; Le et al., 2022). The vulnerability of adapted models can be attributed to their tendency to capture shallow and spurious patterns when fine-tuning on downstream tasks, instead of utilizing the general linguistic knowledge they have learned during the pre-training stage (Sagawa et al., 2020; Warstadt et al., 2020; Gouk et al., 2021; Dong et al., 2021).

To address this issue, various defense methods have been proposed, including adversarial training (Goodfellow et al., 2015; Zhu et al., 2020; Ivgi and Berant, 2021), adversarial data augmentation (Zhang et al., 2019; Zheng et al., 2020; Si et al., 2021) and so on. Adversarial training and adversarial data augmentation stand to provide the most promising performance among all the defense methods. They enhance adversarial robustness by re-training models with additional adversarial data generated either via human-crafting or by conducting projected gradient ascent on the benign examples. In essence, these methods prevent models from learning misleading features by covering more diverse training data. However, they are limited to practice as they often require prohibitively large attack search spaces, and are not generally applicable to different types of attacks.

In this work, we present an attack-agnostic and model-agnostic defense method called **RO**bust **SE**lective Fine-Tuning (**ROSE**) to address these challenges from a learning perspective. ROSE is a novel fine-tuning method that conducts robust updates selectively during the fine-tuning stage. The intuition behind our method is straightforward: *only robust and informative updates of parameters should be conducted.* While the improper ones, which make the fine-tuned model capture superfi-

*Equal contribution
†Part of the work was done while Yankai Lin and Peng Li were working at Tencent.
‡Corresponding Author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2886–2897
December 7-11, 2022



(a) w/o Fine-Tuning (b) Overfit (c) Vanilla Fine-Tuning (d) ROSE-Ensemble

Figure 1: The loss surfaces of fine-tuned models with different fine-tuning strategies, initialized from RoBERTa$_{\texttt{BASE}}$. ROSE leads to wider and flatter optima, enhancing the adversarial robustness of adapted models.

cial cues and overfit the training data of the downstream task, should be filtered out.

Specifically, we propose two strategies in response to the above challenges: first-order ROSE and second-order ROSE. Our first-order ROSE employs adversarial perturbations to clarify parameters that are robust against slight perturbation in the hidden space, enabling models to cope with superficial patterns among examples with similar semantics in the current step. Our second-order ROSE allows models to counter superficial patterns across examples with different semantics along the fine-tuning process by smoothing the optimization trajectory. We also propose an ensemble method to aggregate the benefits of the above two strategies. ROSE distinguishes parameters based on robust criteria at each step in the backward process, then performs robust updates while freezing the remaining parameters.

Figure 1 illustrates the loss landscapes of solutions found by different fine-tuning strategies (pre-trained initial, vanilla fine-tuning, overfitting and ROSE-tuning) on specific tasks. ROSE leads to solutions corresponding to broader and flatter optima compared to traditional fine-tuning methods, which implies that it achieves better adversarial robustness as found in Goodfellow and Vinyals (2015). Moreover, our probing experiment illustrates that ROSE prefers deeper linguistic features rather than shallow lexical ones during fine-tuning. The above empirical analysis confirms the inner working of ROSE. ROSE allows for more robust solutions by masking out unreliable and spurious updates when fine-tuning models on downstream tasks.

We conduct extensive experiments to evaluate the effectiveness of ROSE. We compare ROSE to several strong defense methods. The results show that ROSE exhibits superior adversarial robustness on challenging examples, and achieves comparable or even better benign performance on several benchmarks. ROSE is generic and can be incorporated with existing methods to further enhance their adversarial robustness.

2 Methodology

In this section, we will introduce our method in detail. The key motivation of our method is to select parameters that carry stable information for downstream tasks during fine-tuning. Specifically, the vanilla fine-tuning updates all the parameters in the backward process, while ROSE only updates the robust and informative ones. To identify robust parameters, we propose two robustness criteria and corresponding selective fine-tuning strategies: first-order ROSE (Section 2.1) and second-order ROSE (Section 2.2). Furthermore, we explore an ensemble robust selective fine-tuning method (Section 2.3), which aggregates the benefits of the above two strategies. The overall training algorithm of ROSE when applied to AdamW (Loshchilov and Hutter, 2019) is shown in Algorithm 1.

2.1 First-order ROSE

Our first-order ROSE aims to select parameters that are insensitive to first-order perturbation in the hidden space. We employ adversarial inputs to distinguish robust parameters. Different from the conventional virtual adversarial examples generated via PDG-based methods, ROSE adopts dropout to generate adversarial perturbation with little overhead cost. We follow the method used in Gao et al. (2021); Liang et al. (2021), which passes the same input to the model twice in the forward process with different dropout, and obtains two outputs correspondingly. Then in the backward process, ROSE only updates parameters that are insensitive

Algorithm 1 ROSE for AdamW Optimizer

given $\alpha = 0.001, \beta_1, \beta_2 \in [0, 1), \epsilon = 10^{-8}, \lambda \in \mathbb{R}$

initialize time step $t \leftarrow 0$, parameter vector $\boldsymbol{\theta}_{t=0} \in \mathbb{R}^n$, first moment vector $\mathbf{m}_{t=0} \leftarrow \mathbf{0}$, second moment vector $\mathbf{v}_{t=0} \leftarrow \mathbf{0}$, learning rate $\eta \in \mathbb{R}$

repeat

$t \leftarrow t + 1$

$\mathbf{g}_t \leftarrow \nabla \mathcal{L}_t^{\text{SCE}}(\boldsymbol{\theta}_{t-1})$

$\mathbf{r}_t^{\text{F}} \leftarrow \|\nabla \mathcal{L}_t^{\text{KL}}(\boldsymbol{\theta}_{t-1})\|_F$ ▷ calculate first-order risks

$\mathbf{r}_t^{\text{S}} \leftarrow |(1-\beta_1)\|\mathbf{g}_t\|_F / \|\mathbf{m}_{t-1}\|_F - 1|$ ▷ calculate second-order risks

$\mathbf{M}_t \leftarrow \text{CalculateMask}(\mathbf{r}_t^{\text{F}}, \mathbf{r}_t^{\text{S}})$ ▷ calculate mask

$\mathbf{g}_t^{'} \leftarrow \mathbf{M}_t \odot \mathbf{g}_t + (\mathcal{J} - \mathbf{M}_t) \odot \mathbf{m}_{t-1}$ ▷ update gradients

$\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t^{'}$

$\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_{\mathbf{t}}^{'2}$

$\hat{\mathbf{m}}_t \leftarrow \mathbf{m}_t / (1-\beta_1^t)$

$\hat{\mathbf{v}}_t \leftarrow \mathbf{v}_t / (1-\beta_2^t)$

$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \left(\hat{\mathbf{m}}_t / (\sqrt{\hat{\mathbf{v}}_t} + \epsilon) + \lambda \boldsymbol{\theta}_{t-1} \right) \odot \mathbf{M}_t$ ▷ update weights

until *stopping criterion is met*

return optimized parameters $\boldsymbol{\theta}_t$

to the difference between the two outputs.

Formally, we denote an initial pre-trained model as $\boldsymbol{\theta}_0$, the two probability distributions produced with different dropout at the t-step as $\mathcal{P}_t, \mathcal{P}_t^{'}$, and the Kullback-Leibler (KL) divergence between them is defined as follows:

$$\mathcal{L}_t^{\text{KL}} = \mathcal{D}_{KL}(\mathcal{P}_t \| \mathcal{P}_t^{'}) + \mathcal{D}_{KL}(\mathcal{P}_t^{'} \| \mathcal{P}_t). \tag{1}$$

In the backward process, first-order ROSE filters out parameters which incline to learn superficial cues between similar examples. The potential risk $r_t^{\text{F},i}$ of the i-th parameter in model is computed as the ℓ_F norm of the gradient with regard to $\mathcal{L}_t^{\text{KL}}$:

$$r_t^{\text{F},i} = \|(\nabla \mathcal{L}_t^{\text{KL}})_i(\boldsymbol{\theta}_{t-1})\|_F. \tag{2}$$

Then we sort the magnitude of the sensitivity from $r_t^{\text{F},1}, r_t^{\text{F},2}, \cdots, r_t^{\text{F},n}$ into $r_t^{\text{F},k_1}, r_t^{\text{F},k_2}, \cdots, r_t^{\text{F},k_n}$ in ascending order.

Given the upper threshold c_h^F (*e.g.*, 60%) for robust parameters, the mask $\mathbf{M}_t^{\text{F}}$ can be derived as:

$$\mathbf{M}_t^{\text{F},k_i} = \begin{cases} 1 & i/n \leq c_h, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Note that, only the classification loss will be used to update the weights of models, while gradients with regard to $\mathcal{L}_t^{\text{KL}}$ will be discarded after calculating masks.

2.2 Second-order ROSE

Our second-order ROSE smooths the optimization trajectory to prevent models from learning spurious patterns between different groups of data points along the fine-tuning process. More precisely speaking, our second-order ROSE selects and tunes parameters that are less aggressively updated to avoid overfitting on spurious patterns. A straightforward idea is to calculate the second derivatives of the classification loss as the second-order risks. Unfortunately, it requires prohibitive computation and storage costs. Thus we employ a stochastic gradient-based optimizer like AdamW to approximate this solution.

Formally, we denote the softmax cross-entropy loss at the t-step as $\mathcal{L}_t^{\text{SCE}}$, and the first momentum of optimizer as $\mathbf{m}_{t-1}$. Then the second-order risk $r_t^{\text{S},i}$ of the i-th parameter in the model is defined as the relative magnitude between current gradients $\mathbf{g}_t$ and the exponential moving average $\mathbf{m}_{t-1}$, which is computed as:

$$r_t^{\text{S},i} = \left| \frac{\alpha \|\mathbf{g}_t^i\|_F}{\|\mathbf{m}_{t-1}^i\|_F} - 1 \right|, \tag{4}$$

where α is a scaling coefficient. In AdamW, $\alpha = (1-\beta_1)$, and β_1 is the momentum factor.

Similar to our first-order ROSE, we sort the magnitude of the second-order risks $\mathbf{r}_t^{\text{S}}$ in ascending order and calculate the second-order mask $\mathbf{M}_t^{\text{S}}$ with Eq. 3.

2.3 Ensemble ROSE

Since our first-order and second-order ROSE emphasize different kinds of robust parameters, we then propose an ensemble method to aggregate the benefits of the above two mechanisms.

At the t-step, we first calculate both the first-order risks $r_t^{\mathrm{F},1}, \cdots, r_t^{\mathrm{F},n}$ with Eq. 2 and second-order risks $r_t^{\mathrm{S},1}, \cdots, r_t^{\mathrm{S},n}$ with Eq. 4. Then we sort both of them in ascending order. Given upper thresholds c_h^{F} and c_h^{S}, we can compute the first-order and second-order masks as: $\mathbf{M}_t^{\mathrm{F}}$ and $\mathbf{M}_t^{\mathrm{S}}$, respectively. Finally, the ensemble mask $\mathbf{M}_t^{\mathrm{E}}$ at t-step is computed as:

$$\mathbf{M}_t^{\mathrm{E}} = \gamma \mathbf{M}_t^{\mathrm{F}} + (1-\gamma)\mathbf{M}_t^{\mathrm{S}}, \tag{5}$$

where $\gamma \in (0, 1)$ is a scaling coefficient hyper-parameter to control the weight of two masks.

3 Experiment

3.1 Datasets

We demonstrate the effectiveness of our method using four tasks from GLUE (Wang et al., 2019) and AdvGLUE (Wang et al., 2021b) benchmarks. The General Language Understanding Evaluation (GLUE) is a widely-used benchmark, including 9 natural language understanding tasks. The Adversarial GLUE (AdvGLUE) is a robustness benchmark based on GLUE, covering 14 prevalent adversarial textual attack methods. The AdvGLUE adopts careful systematic annotations to curate high-quality and reliable adversarial examples. We do not employ automatic adversarial attack algorithms to evaluate the adversarial robustness, due to their tendency to produce incorrect or puzzling adversarial examples (Li et al., 2021a).

SST-2 (Socher et al., 2013) is a sentiment classification task with single-sentence inputs, which is collected from movie reviews.

RTE (Bentivogli et al., 2009) is a natural language inference task aggregated from a series of textual entailment challenges, originating from news articles and Wikipedia content.

QNLI (Rajpurkar et al., 2016) is also an inference task. Given the context sentence and corresponding question, this task is to determine whether it provides the answer.

QQP[1] (Chen et al., 2018) is a widely used benchmark involving detecting semantic similarity. An annotated binary label indicates whether two texts of each pair are semantically same or not.

3.2 Baselines

We adopt pre-trained checkpoints of RoBERTa$_{\mathrm{BASE}}$ and RoBERTa$_{\mathrm{LARGE}}$ (Liu et al., 2019) as the basis of our experiments. Besides the vanilla fine-tuning method, we select several suitable baselines for comparison including:

R-Drop (Liang et al., 2021) is a generic regularization strategy, which assures consistency between two outputs obtained by different dropout. ROSE borrows the idea of dropout twice, but does not fine-tune all parameters to constrain the divergence between two outputs.

CHILD-TUNING$_D$ (Xu et al., 2021) is a fine-tuning technique, which only updates the most informative subset of parameters of large pre-trained models during the backward process. Although both CHILD-TUNING and ROSE mask out gradients in the backward process, the specific parameters they update are completely different.

SMART (Jiang et al., 2020) is an adversarial training approach, contains a regularization module induced by smoothness and a optimization module inspired from Bregman proximal point method.

FreeLB (Zhu et al., 2020) is an adversarial training approach built on the top of language models, which improves higher invariance in word embedding space and reduces the adversarial risk surrounding examples.

3.3 Experimental Settings

Our implementation of ROSE is based on Huggingface library[2] (Wolf et al., 2020). Batch size for RTE is set to 16, and for other tasks it is set to 32. Dropout rates are all set to 0.1. We carry out grid search of learning rate $\in \{1e-5, 2e-5, \cdots, 1e-4\}$ and upper threshold $\in \{10\%, 20\%, \cdots, 90\%\}$. For ROSE-ensemble, we simply set $\gamma = 0.5$ in Eq. 5. The maximum number of epochs is set to 10. For the replication study, we report the average accuracy over 5 random seeds on the GLUE and AdvGLUE development sets after fine-tuning the pre-trained models on the corresponding GLUE training data.

For all the baselines, we either perform grid search or adopt the parameter combination provided by the official codebase to select the best parameters. Similarly, we report the average results

[1] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

[2] https://github.com/huggingface/transformers

Model	SST-2		RTE		QNLI		QQP		Avg	Avg
	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE+AdvGLUE	$\Delta\downarrow$
RoBERTa$_{\text{BASE}}$										
Vanilla	94.29	24.05	77.91	28.15	92.97	27.43	91.58	19.49	56.98	64.41
R-Drop	95.32	27.84	79.86	31.36	93.30	28.92	91.86	37.44	60.74	58.70
CHILD-TUNING$_D$	94.21	23.82	75.52	16.54	92.36	31.89	91.64	17.95	55.49	65.88
SMART	94.98	35.95	77.54	24.44	93.35	34.29	91.04	46.58	62.27	53.91
FreeLB	94.89	35.81	78.42	32.10	93.12	36.22	92.04	44.10	63.34	52.56
ROSE-First	94.84	37.67	78.34	35.49	92.19	44.19	89.56	44.44	64.59	48.29
ROSE-Second	93.78	36.99	78.16	37.97	92.41	34.63	90.48	45.73	63.77	49.88
ROSE-Ensemble	94.09	39.36	78.63	38.02	92.64	39.59	90.39	47.44	**65.02**	**47.84**
RoBERTa$_{\text{LARGE}}$										
Vanilla	96.08	56.08	85.92	61.73	94.58	63.38	92.09	40.60	73.81	36.72
R-Drop	96.59	53.38	85.56	66.67	95.01	55.95	92.35	44.80	73.79	37.18
CHILD-TUNING$_D$	95.91	51.35	85.92	61.73	94.30	58.11	92.03	43.59	72.87	38.35
SMART	96.67	59.12	85.02	69.14	94.91	61.04	92.12	50.85	76.11	32.14
FreeLB	96.49	59.32	86.76	66.91	94.99	62.30	92.60	48.21	75.95	33.53
ROSE-First	95.58	57.77	85.13	70.62	94.08	64.02	90.67	60.26	77.27	28.20
ROSE-Second	96.29	60.59	85.08	67.49	94.72	63.68	91.68	55.90	76.93	30.03
ROSE-Ensemble	96.10	60.81	85.92	71.11	94.26	64.64	91.46	60.51	**78.10**	**27.67**

Table 1: Model performance on GLUE and AdvGLUE benchmarks. The results are accuracy averaged over 5 random seeds. Values are reported as percentage (%). Besides each task scores, we also report a macro-average. The last column is the drop from GLUE to AdvGLUE, the smaller the better. The **bold** represents ROSE is significantly better (1-tailed t-test, p-value < 0.05) than the baselines.

on two benchmarks over 5 random seeds using the same evaluation schema.

3.4 Main Results

We compare ROSE-First, ROSE-Second, and ROSE-Ensemble to all baselines on SST-2, RTE, QNLI, and QQP tasks from GLUE and AdvGLUE benchmarks. The overall results are summarized in Table 1. We observe that:

(1) Our proposed ROSE substantially improves the robustness of fine-tuned pre-trained language models, while maintaining competitive benign performances to previous methods. Despite the effectiveness of ROSE-First and ROSE-Second varies on tasks and model size, both of them surpass the existing methods. ROSE-Ensemble aggregates the advantages of first-order and second-order strategies, providing the strongest adversarial robustness. In particular, ROSE-Ensemble$_{\text{BASE}}$ outperforms vanilla RoBERTa$_{\text{BASE}}$ model by 8.04% average score. ROSE-Ensemble$_{\text{LARGE}}$ beats RoBERTa$_{\text{LARGE}}$ by 4.29% on average.

(2) ROSE consistently outperforms CHILD-TUNING$_D$ and R-Drop, which both share some similarities with our method. CHILD-TUNING$_D$, which masks out the most inessential parameters in the backward process, shows the worst robustness on most datasets. R-Drop uses dropout to regularize the output of models. Results indicate that R-Drop improves robustness on a number of tasks, but it is not competitive with strong defense methods. We will explore the effectiveness of our robust selection strategies further in Section 3.5.

(3) Our method also surpasses the two strong baselines SMART and FreeLB, which employ the most prevalent adversarial training idea to improve the robustness of pre-trained models. For instance, ROSE-Ensemble$_{\text{BASE}}$ enhances performance by up to 2.75% average score over SMART$_{\text{BASE}}$. ROSE-Ensemble gains 2.15% average score improvement compared to FreeLB with RoBERTa$_{\text{LARGE}}$. Furthermore, SMART and FreeLB are both inefficient and heavily correlated with model structure, while our ROSE does not suffer from these issues.

3.5 Extensions of ROSE to Existing Method

ROSE is a generic method and can be easily incorporated into other well-recognized methods. In this section, we incorporate ROSE into R-Drop and examine whether it is still effective. Since the optimization objective of R-Drop is a weighted sum of softmax cross-entropy loss and KL-divergence, we decouple them from the aggregated loss and use them to perform our first-order and second-order mask calculations, respectively. Noted that, in the backward process we still use gradient calculated with regards to the aggregated loss to update, which is different from the ROSE process.

We primarily adopt the best parameter combinations from the main experiments in Section 3.4,

Model	SST-2		RTE		QNLI		QQP		Avg	Avg
	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE+AdvGLUE	Δ ↓
RoBERTa$_{\text{BASE}}$										
R-Drop	95.32	27.84	79.86	31.36	93.30	28.92	91.86	37.44	60.74	58.70
+ ROSE-First	94.77	32.30	78.34	36.05	92.92	39.46	90.62	42.82	63.41	51.51
+ ROSE-Second	95.05	30.00	73.93	36.79	92.60	32.03	91.96	39.23	61.45	53.87
+ ROSE-Ensemble	92.36	35.68	77.83	36.54	92.50	40.27	90.90	42.56	**63.58**	**49.64**
RoBERTa$_{\text{LARGE}}$										
R-Drop	96.59	53.38	85.56	66.67	95.01	55.95	92.35	44.80	73.79	37.18
+ ROSE-First	96.54	55.68	85.85	68.89	94.38	56.62	91.95	53.21	75.39	33.58
+ ROSE-Second	96.63	55.81	85.27	66.91	94.84	57.16	92.38	50.26	74.91	34.74.
+ ROSE-Ensemble	96.72	56.62	85.20	69.13	94.81	57.97	92.27	54.61	**75.92**	**32.66**

Table 2: Experiment results for R-Drop incorporated with ROSE. ROSE improves the robustness of R-Drop further. The **bold** represents ROSE is significantly better (1-tailed t-test, p-value < 0.05) than R-Drop.

including the learning rates and upper thresholds. We follow the settings from R-Drop for other parameters. We conduct experiments using both RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$.

Results are shown in Table 2. Generally, ROSE improves the adversarial robustness of R-Drop by a large margin, and maintains competitive benign performances at the same time. For example, ROSE-First$_{\text{BASE}}$ promotes the adversarial robustness of R-Drop$_{\text{BASE}}$ on QNLI task from 28.92% to 39.46%. R-Drop patched with ROSE-Second witnesses an improvement on QQP task from 44.80% to 50.26% using RoBERTa$_{\text{LARGE}}$. Notably, our ROSE-Ensemble outperforms R-Drop by roughly 3 points on average for both model sizes. The above results indicate that when incorporated into existing methods, ROSE can enhance their adversarial robustness even further.

3.6 Effect of Scalar γ

Further, we investigate the impact of the scaling coefficient γ in our ROSE-Ensemble. Here we vary the $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and conduct experiments on four tasks, where $\gamma = 0.5$ is the current setting. We adopt the setting from Section 3.4 for other parameters.

γ	RoBERTa$_{\text{BASE}}$		RoBERTa$_{\text{LARGE}}$	
	Avg GLUE+AdvGLUE	Avg Δ ↓	Avg GLUE+AdvGLUE	Avg Δ ↓
0.1	59.35	56.81	75.07	28.99
0.3	58.42	55.13	75.51	30.27
0.5	65.02	47.84	78.10	27.67
0.7	61.09	51.36	75.60	30.52
0.9	59.88	50.82	73.22	32.25

Table 3: Results for ROSE-Ensemble with different γ.

The results are presented in Table 3. It is shown that the best-balanced choice is $\gamma = 0.5$, but ROSE-Ensemble can stably improve the robustness using other γ. Furthermore, ROSE achieves more substantial performance when applied to pre-trained language models of greater complexity.

4 Analysis

In this section, we conduct further analysis to reveal the inner working of ROSE.

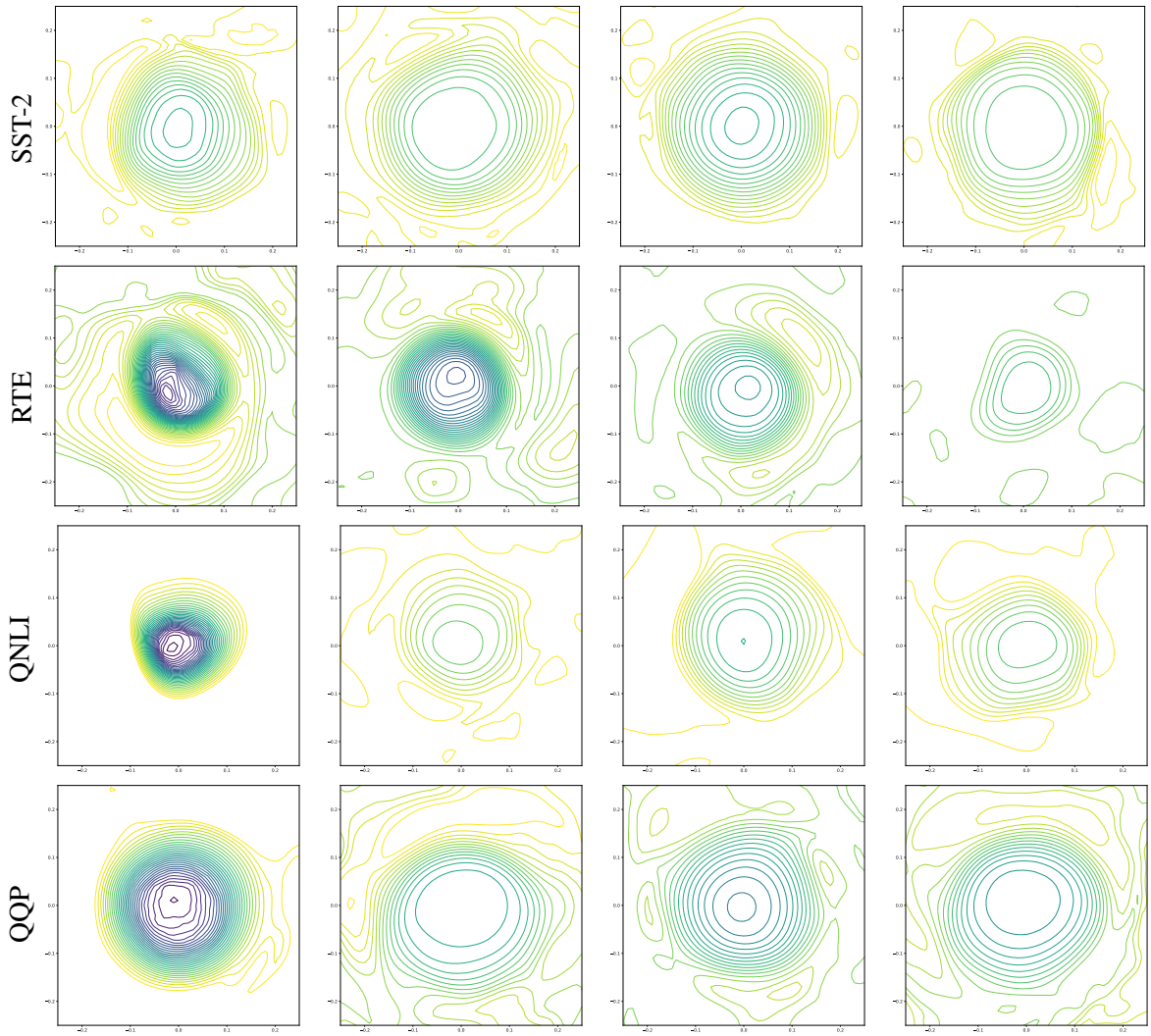
4.1 Two-dimensional Loss Visualization

The loss landscape (Goodfellow and Vinyals, 2015; Li et al., 2018a) is a valid and effective indicator to characterize the property of neural networks. It has been empirically demonstrated that flatter and wider optima correlate well with better robustness. We plot and compare two-dimensional loss landscapes of the solutions found by vanilla fine-tuning and our ROSE. Visualizations on various tasks show that our ROSE generally leads to flatter and wider optima, thus improving the adversarial robustness of adapted models.

Let $\boldsymbol{\theta}$ denote the parameters of a model fine-tuned on downstream tasks. Then the two-dimensional loss curve of model can be plotted with the function:

$$f(\alpha, \beta) = \mathcal{L}(\boldsymbol{\theta} + \alpha\boldsymbol{\delta} + \beta\boldsymbol{\eta}), \tag{6}$$

where $\mathcal{L}$ is the loss function, and α, β are scalar values. $\boldsymbol{\delta}, \boldsymbol{\eta}$ are direction vectors randomly sampled from Gaussian distribution, which denote two direction vectors in the parameter space corresponding to the two axes of the loss surface. To remove the scaling effect of neural nets, we follow the filter-wise normalization in Li et al. (2018a), which scales the $\boldsymbol{\delta}, \boldsymbol{\eta}$ to the same norm as parameters by $\frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\|\boldsymbol{\theta}\|$, $\frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|}\|\boldsymbol{\theta}\|$. We set the range of both α and β to $[-0.25, 0.25]$ and uniformly sample 51 points



(a) Vanilla (b) ROSE-First (c) ROSE-Second (d) ROSE-Ensemble

Figure 2: 2D loss contours of models with vanilla fine-tuning and ROSE on four tasks. The lighter color of the contour lines correlates with a larger loss, and the area with denser contour lines demonstrates steeper loss surfaces.

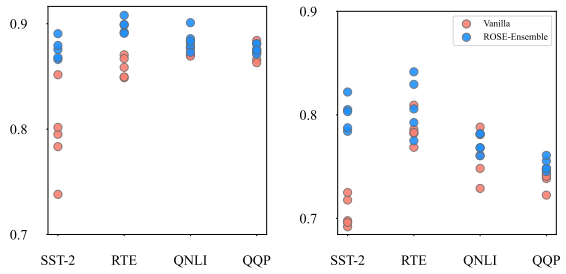
for each axis. Since the parameter space is high-dimensional, experimental results confirm the two directions $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are divergent and orthogonal to each other. We plot and compare the loss surfaces of models with vanilla fine-tuning and ROSE on four tasks.

The visualizations are shown in Figure 2. We can observe that ROSE has a significant influence on the smoothness of the loss landscapes across all datasets. Models fine-tuned with ROSE provide wider and less dense loss contours than vanilla fine-tuning, which shows that they are more robust against noisy perturbations. Specifically, ROSE-First finds solutions with wider bottoms, and ROSE-Second leads to solutions with less dense loss contour. This indicates that ROSE-First and ROSE-Second succeed in defense of local and global perturbations, respectively. Additionally, ROSE-Ensemble is shown to have both of these features, demonstrating that it aggregates the benefits of the two strategies discussed above. Appendix A.1 provides an additional one-dimensional visualization.

4.2 Probing Preference for Different Features

We then employ the probing task from (Warstadt et al., 2020) to test whether models fine-tuned with ROSE prefer linguistic rather than superficial features. In the probing experiment, a model is first trained on ambiguous data which equally supports both linguistic and superficial cues, and then tested on disambiguating data which supports only the linguistic cues. The preference of models for features is measured through Matthews correlation scores between predictions and labels on test sets.

The models are shown a systematic preference for linguistic features if the score is 1, and complete reliance on superficial cues if the score is -1. Therefore a higher score shows a stronger preference for linguistic features. We select two representative experiments gotten by pairing the linguistic feature *Syntactic construction* with two surface features *Lexical content* and *Length*. For each probing task, we report results of adapted models with different fine-tuning strategies on 5 random seeds.



(a) Lexical content (b) Length

Figure 3: Results for probing tasks with different random seeds. Each data point represents one run. Models with a stronger preference for linguistic rather than spurious features achieve higher scores.

Results are plotted in Figure 3. We can observe that, compared to vanilla fine-tuned models, ROSE-tuned models show a stronger preference for linguistic features than any superficial features. This indicates that ROSE successfully enables models to extract deeper linguistic knowledge during fine-tuning, instead of adopting spurious cues from the training data of downstream tasks.

5 Related Work

Adversarial training is the most effective and promising strategy to improve the adversarial robustness of models. Existing adversarial training methods usually employ PDG-based attacks to generate adversarial examples, and force models to maintain the output on them (Zhu et al., 2020; Liang et al., 2018; Wang et al., 2021c). Despite the substantial improvements in robustness, adversarial training often requires significant computational and memory costs and fails to preserve the original labels. Some works focus on constructing reliable adversarial datasets (Gardner et al., 2020; Eger and Benz, 2020), which require huge human annotation and only work for a single task. By contrast, our proposed ROSE is much more efficient and only employs such perturbations to select robust parameters to tune, therefore, there is no need for reliable adversarial examples.

Besides adversarial training methods, our work also relates to a few works of regularization and optimization. In regularization, lots of methods have been proposed, including L^2-penalty (Schwarz et al., 2018; Li et al., 2018b; Chen et al., 2020), weight decay (Kang et al., 2016; Zhang et al., 2021), Mixout regularization (Lee et al., 2020), and so on. The general approach is augmenting the vanilla optimizer with terms that indirectly or directly penalize the aggressive updates. Although these methods are exciting, the regularization is often not scalable, and hard to transfer to another model. Another line of work (Wang et al., 2021a; Dong et al., 2021) attempts to address this issue from an informative-theoretic perspective. In optimization, there has been some work proposed recently to force optimization towards wide valleys (Chaudhari et al., 2017; Jiang et al., 2020). Compared to these works, ROSE uses the simplest idea by selecting parameters with the second-order robustness in fine-tuning stage to smooth the optimization trajectory. ROSE is more efficient and can be incorporated into existing methods to improve their adversarial robustness further.

Note that our method does not fall within the realm of model compression. The target of model compression is to obtain an efficient sub-network with competitive performance, with typical approaches to abandon some parameters when models do inference. While ROSE aims to improve the adversarial robustness of pre-trained language models, which is done via conducting selective parameters updates in the backward process.

6 Conclusion

In this work, we propose an attack-agnostic and model-agnostic defense approach called ROSE, which selectively updates robust parameters during the fine-tuning stage. We present first-order ROSE which selects the parameters robust against slight perturbation in the hidden space, second-order ROSE which filters out aggressive updates, and ensemble ROSE which aggregates the benefits of the above two strategies. Experimental results show that both our ROSE-First and ROSE-Second greatly improve the robust performance on various NLP benchmarks, while ROSE-Ensemble is even more effective. Besides, existing methods achieve better robustness when incorporated with

our ROSE. We also demonstrate empirically that the effectiveness of ROSE can be attributed to the wider and flatter solutions it finds than the conventional fine-tuning methods. We hope ROSE could motivate more defense works for language models.

Limitations

Although ROSE achieves superior adversarial robustness on four datasets, there are still two limitations. First, there are some vital hyper-parameters in ROSE, *e.g.* the scaling coefficient γ, which have a great influence on the performance as shown in Section 3.6. We adopt grid search to select the best parameters, which requires considerable GPU resources. There is still a need for a more automatic method. Once we further understand the inner working mechanism of deep neural networks, such hyper-parameters could be calculated theoretically. Second, due to the limitation of computational resources, we focus on fine-tuning in this work, leaving applying ROSE to pre-training for future work. We hope ROSE could provide a new perspective for general defense work towards more robust language models.

References

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. 2017. Entropy-sgd: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs.

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, and Hanwang Zhang. 2021. How should pre-trained language models be fine-tuned towards adversarial robustness? In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4356–4369.

Steffen Eger and Yannik Benz. 2020. From hero to zéroe: A benchmark of low-level adversarial attacks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 786–803, Suzhou, China. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ian J. Goodfellow and Oriol Vinyals. 2015. Qualitatively characterizing neural network optimization problems. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Henry Gouk, Timothy M. Hospedales, and Massimiliano Pontil. 2021. Distance-based regularisation of deep networks for fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Maor Ivgi and Jonathan Berant. 2021. Achieving model robustness through discrete adversarial training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1529–1544, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled

regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2177–2190. Association for Computational Linguistics.

Lan Jiang, Tianshu Lyu, Yankai Lin, Meng Chong, Xiaoyong Lyu, and Dawei Yin. 2022. On length divergence bias in textual matching models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4187–4193, Dublin, Ireland. 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Guoliang Kang, Jun Li, and Dacheng Tao. 2016. Shakeout: A new regularized deep neural network training scheme. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Thai Le, Noseong Park, and Dongwon Lee. 2022. SHIELD: Defending textual neural networks against multiple black-box adversarial attacks with stochastic multi-expert patcher. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6661–6674, Dublin, Ireland. Association for Computational Linguistics.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2020. Mixout: Effective regularization to finetune large-scale pretrained language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018a. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6391–6401.

Luoqiu Li, Xiang Chen, Zhen Bi, Xin Xie, Shumin Deng, Ningyu Zhang, Chuanqi Tan, Mosha Chen, and Huajun Chen. 2021a. Normal vs. adversarial: Salience-based analysis of adversarial samples for relation extraction. In *The 10th International Joint Conference on Knowledge Graphs*, IJCKG'21, page 115–120, New York, NY, USA. Association for Computing Machinery.

Luoqiu Li, Xiang Chen, Ningyu Zhang, Shumin Deng, Xin Xie, Chuanqi Tan, Mosha Chen, Fei Huang, and Huajun Chen. 2021b. Normal vs. adversarial: Salience-based analysis of adversarial samples for relation extraction. *CoRR*, abs/2104.00312.

Xuhong Li, Yves Grandvalet, and Franck Davoine. 2018b. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2825–2834. PMLR.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4208–4215. International Joint Conferences on Artificial Intelligence Organization.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

Jieyu Lin, Jiajie Zou, and Nai Ding. 2021. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 333–342, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4528–4537. PMLR.

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 Y. 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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021b. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. *CoRR*, abs/2111.02840.

Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. 2021c. Adversarial training with fast gradient projection method against synonym substitution based text attacks. *AAAI Conference on Artificial Intelligence*.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9514–9528. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample BERT fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaoqing Zheng, Jiehang Zeng, Yi Zhou, Cho-Jui Hsieh, Minhao Cheng, and Xuanjing Huang. 2020. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6600–6610, Online. 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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Appendix

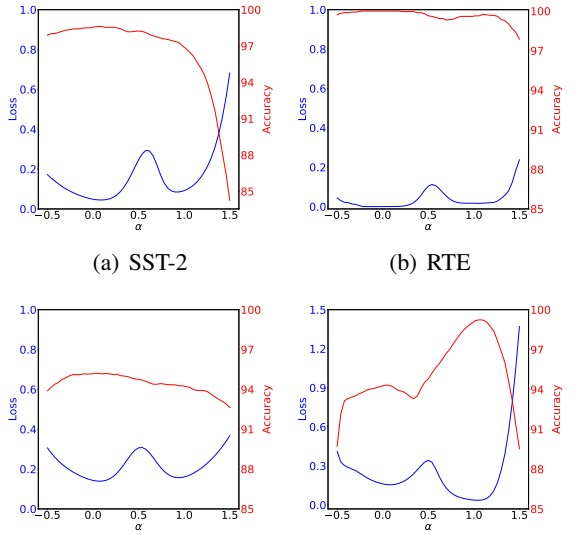
A.1 One-dimensional Linear Interpolation

In order to investigate the robustness of our method, we present parametric one-dimensional visualiza-

tion as described in (Goodfellow and Vinyals, 2015), which plots the value of loss function along the line connecting two different models. Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$ indicate the parameters of these two models respectively. Then we plot the function:

$$f(\alpha) = \mathcal{L}((1-\alpha)\boldsymbol{\theta} + \alpha\boldsymbol{\theta}^{'}), \quad (7)$$

where α is a scalar value. We compare the weights of models obtained by vanilla and ROSE-Ensemble fine-tuning method on four tasks. In particular, for $\alpha \in [-0.5, 1.5]$, we uniformly sample 51 points and plot the function $f(\alpha)$ and superimpose the classification accuracy.



(a) SST-2 (b) RTE

(c) QNLI (d) QQP

Figure 4: 1D loss interpolation (Left axis corresponds to loss, and right to accuracy) between solutions found by vanilla fine-tuning and ROSE-Ensemble on four tasks. Models trained with ROSE ($\alpha = 0$) have flatter and wider curves than vanilla fine-tuning ($\alpha = 1$), which correlates well with robustness.

Figure 4 shows the visualization of fine-tuned models with different strategies. Compared with vanilla fine-tuning, we can observe that our ROSE provides wider and flatter curves. The result indicates that solutions obtained by ROSE tend to be more robust.

A.2 Effectiveness of Dropout

In order to investigate the effectiveness of dropout in ROSE-First, For two outputs produced by different dropout, we inspect the ratio (%) of prediction labels that are not consistent with each other on the first 200, 600, 2000 and 4000 steps.

From Table 4 we can see that, dropout generates adversarial examples as expected with striking low cost. The ratio decreases fast at the beginning and comes to stable finally, which indicates that ROSE-First succeeds to improve the robustness of models against such perturbation over the training process.

Step	200	600	2000	4000
SST-2	12.69	7.80	5.25	4.45
RTE	13.30	9.44	5.12	-
QNLI	18.44	11.61	7.70	6.32
QQP	12.31	9.58	7.35	6.45

Table 4: Results for ROSE-Ensemble with different γ.