How Should Pre-Trained Language Models Be Fine-Tuned Towards Adversarial Robustness?

Xinhsuai Dong, Luu Anh Tuan, Min Lin, Shuicheng Yan, Hanwang Zhang

NIPS 2021

Motivation

Adversarial Attacks in fine-tuned NLP models:

- Character-level Modification
- Sentence-level manipulation
- Word Substitution

Character-level Modification

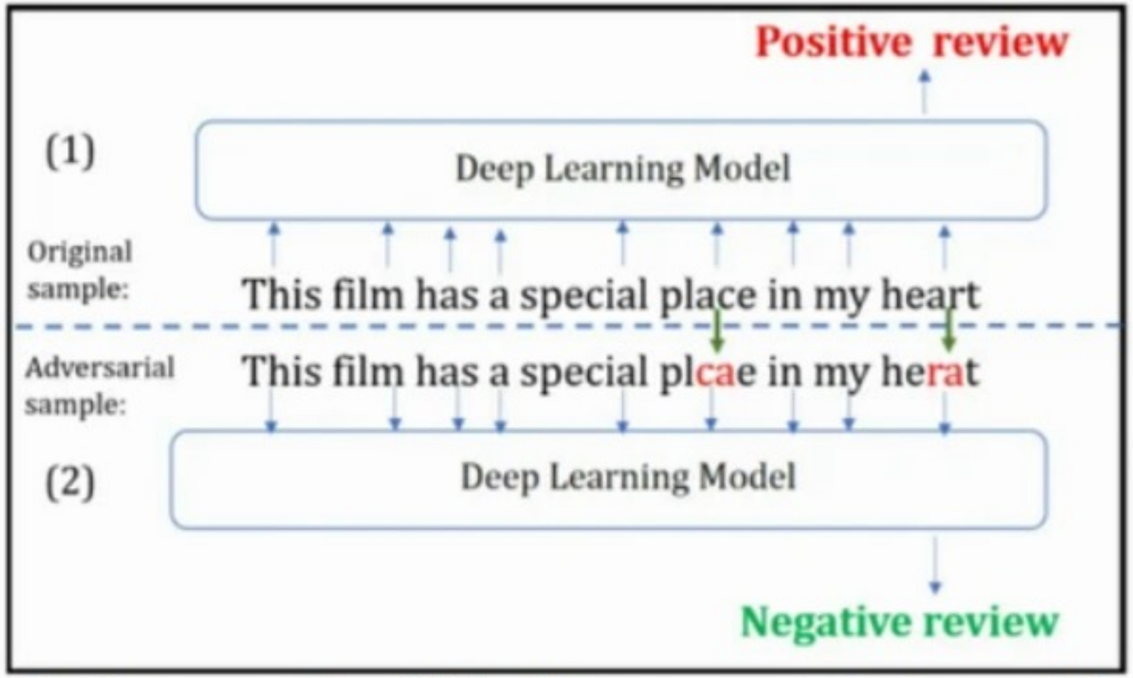


Figure 1: An example of WordBug generated adversarial sequence. Part (1) shows an original text sample and part (2) shows an adversarial sequence generated from the original sample in Part (1). From part (1) to part (2), only a few characters are modified; however this fools the deep classifier to a wrong classification.

Sentence-level manipulation

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Word Substitution (using synonyms)

Original Prediction	*Adversarial* Prediction	Perturbed Texts
Positive Confidence = 96.72%	Negative Confidence = 74.78%	Ah man this movie was ***funny*** (***laughable***) as hell, yet strange. I like how they kept the shakespearian language in this movie, it just felt ironic because of how idiotic the movie really was. this movie has got to be one of troma's best movies. highly recommended for some senseless fun!
Negative Confidence = 72.40%	Positive Confidence = 69.03%	The One and the Only! The only really good description of the punk movement in the LA in the early 80's. Also, the definitive documentary about legendary bands like the Black Flag and the X. Mainstream Americans' repugnant views about this film are absolutely ***hilarious*** (***uproarious***)! How can music be SO diversive in a country of supposed liberty...even 20 years after... find out!

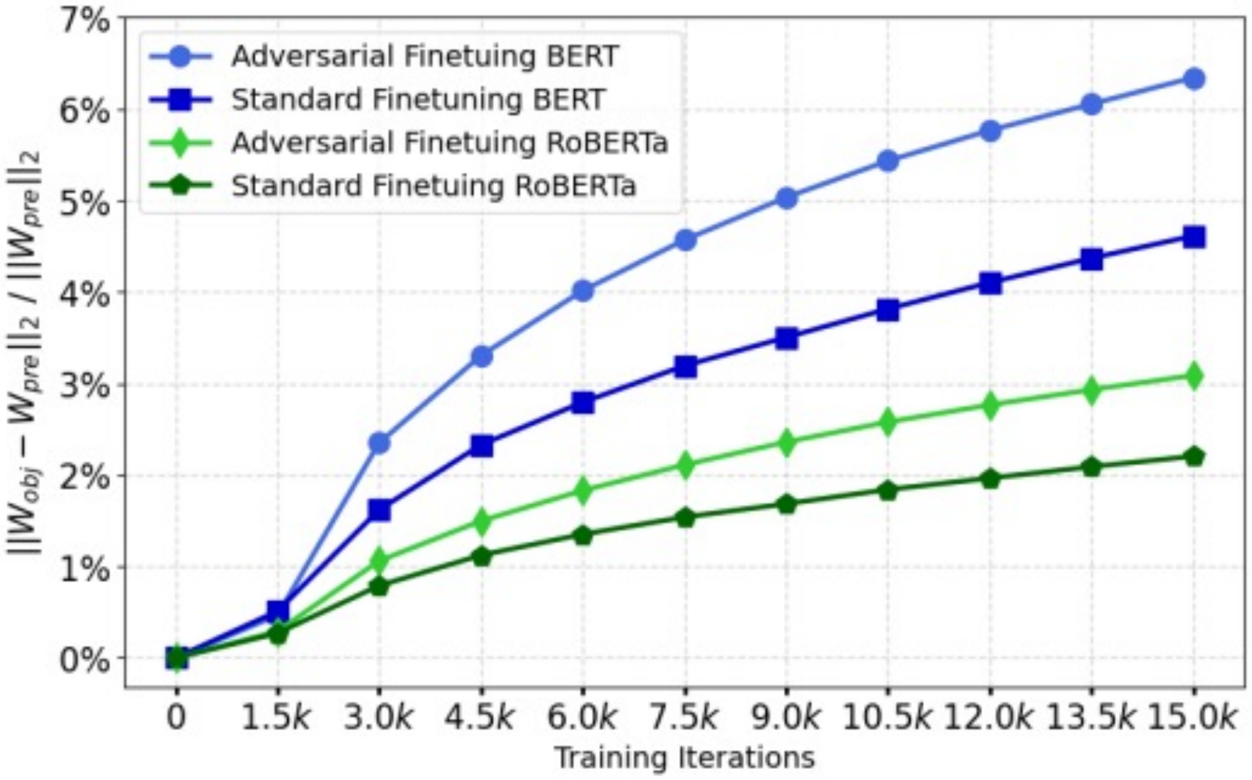
Solution:

Adversarial training:

The training data are augmented by “adversarial” samples generated using an attack algorithm.

$$\min\Big[\mathop{\mathbb{E}}_{x,y\sim p_{\mathcal{D}}}\Big[\max_{\hat{x}\in\mathbb{B}(x)}\mathcal{L}(x,\hat{x},y)\Big]\Big]$$

Problem



(a) In adversarial fine-tuning, the relative L_2 distance continuously grows as the fine-tuning proceeds.

- Adversarial fine-tuning forgets the pre-trained model more than standard fine-tuning.
- Need to retain the generic and robust linguistic features captured by the pre-trained model.

Existing Methods

In the parameter space: add a regularization term in loss function:

$$\lambda \| W_{\text{obj}} - W_{\text{pre}} \|_2$$

- However, change in the model parameter space only serves as an imperfect proxy in function space

- Should use the mutual information between outputs of pre-trained and fine-tuned model

Robust Informative Fine-Tuning (RIFT)

Objective:

Gain better performance on downstream tasks under adversarial attack.

RIFT:

Use mutual information to encourages a fine-tuned model to retain the features learned from the pre-trained model , as these features are benefited to downstream tasks.

Robust Informative Fine-Tuning (RIFT)

$$\max I(S;\, Y, T)$$

$\mathbf{I}()$ is the mutual information

- Here $T = F_t(X)$ $S = F_s(X)$. F_t and F_s are the pre-trained model and the model being fine-tuned respectively.

Robust Informative Fine-Tuning (RIFT)

$$\max I(S;\, Y, T)$$

$\mathbf{I}()$ is the mutual information

$\boldsymbol{H}()$ is the shannon entropy

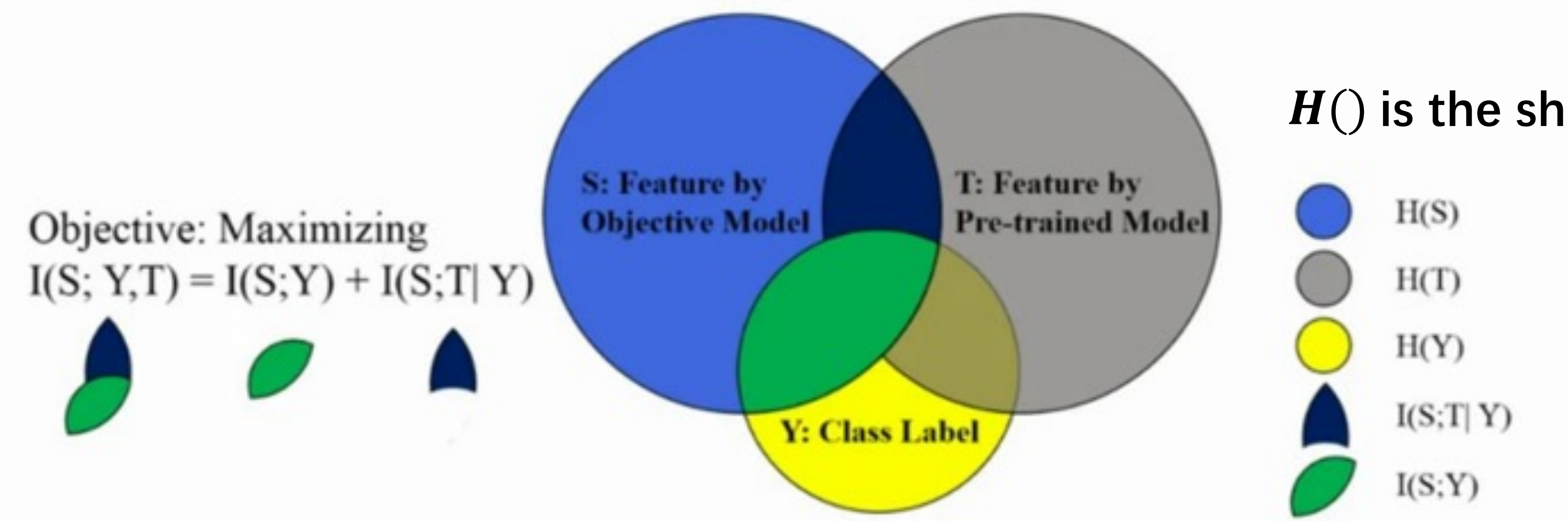
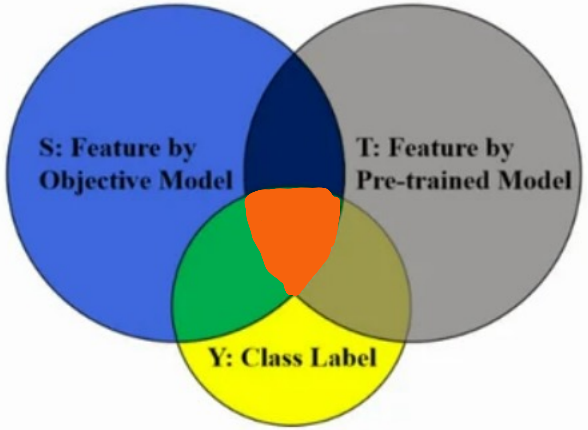


Figure 1: An illustration of the overall objective of RIFT. Maximizing $I(S;Y)$ encourages features of the objective model to be predictive of the class label, while maximizing $I(S;T|\,Y)$ encourages learning robust and generic linguistic information from the pre-trained model. (Random variable S denotes extracted features of X by the objective model and T by the pre-trained language model)

Mutual Information v.s. Conditional Mutual Information

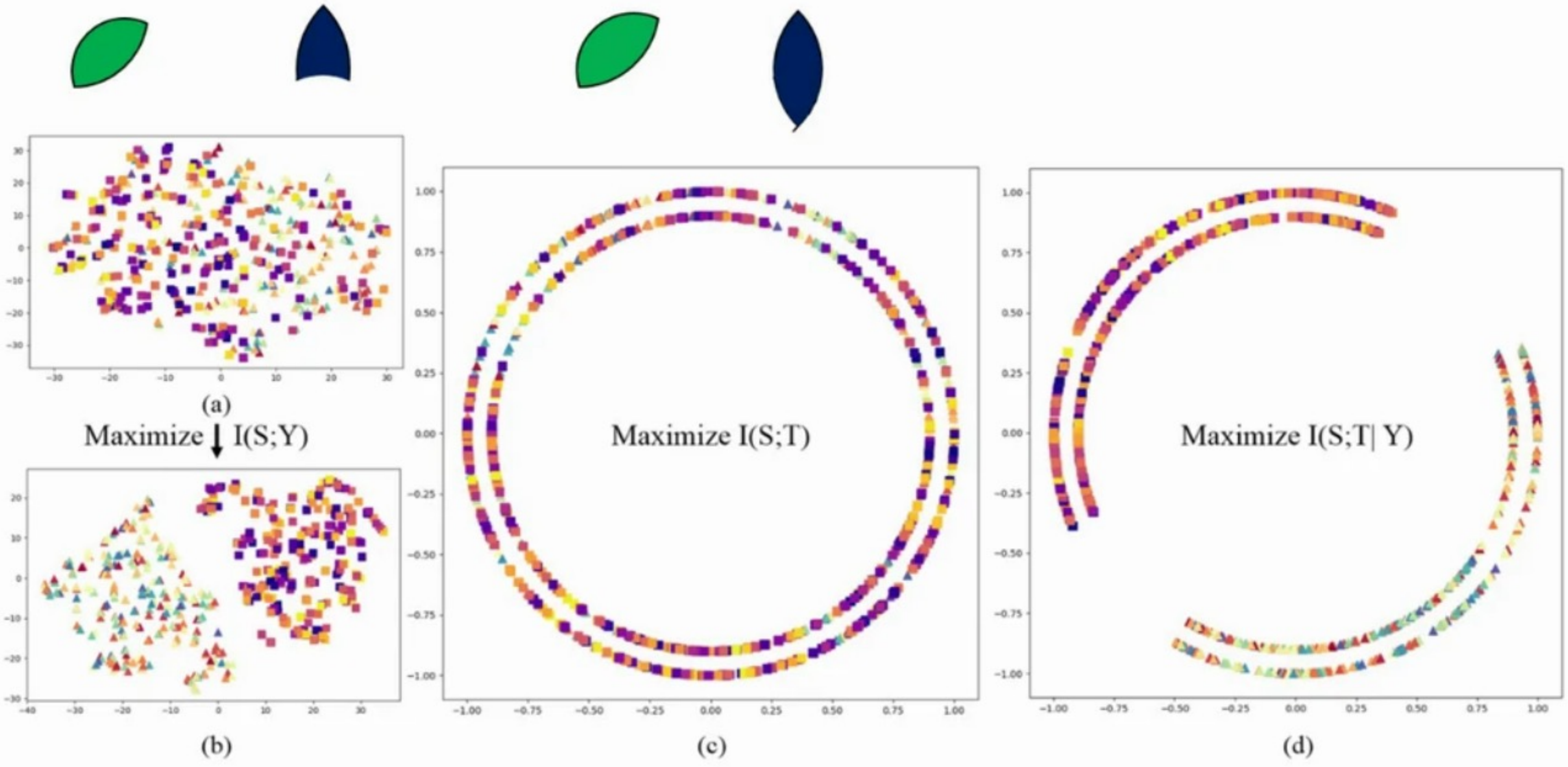
$$I(S;Y) + I(S;T|\,Y) \quad \text{v.s.} \quad I(S;Y) + I(S;T)$$



-  is optimized twice

Mutual Information v.s. Conditional Mutual Information

$I(S;Y) + I(S;T|Y)$ v.s. $I(S;Y) + I(S;T)$



Robust Informative Fine-Tuning (RIFT)

- Overall Objective: $I(S;\,Y,T) = I(S;Y) + I(S;T|\,Y),$

$$
\begin{aligned}
I(S;Y) &= H(Y) - \mathbb{E}_{x,y\sim p_{\mathcal{D}}}[-\log q(y|s)] + \mathrm{KL}\big(p(\cdot|s)\|q(\cdot|s)\big)\\
&\geq H(Y) - \mathbb{E}_{x,y\sim p_{\mathcal{D}}}[-\log q(y|s)],
\end{aligned}
$$

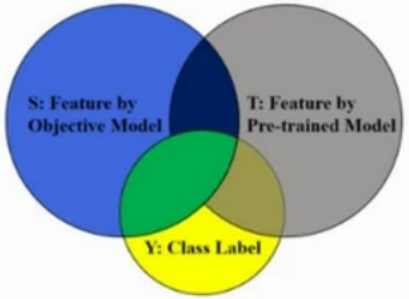
$$
I(S;T|\,Y) = \mathbb{E}_{y\sim p_{\mathcal{D}}(y)}[I(S;T)|\,Y=y] = \mathbb{E}_{y\sim p_{\mathcal{D}}(y)}[\,\mathbb{E}_{x\sim p_{\mathcal{D}}(x|y)}[\log \frac{p(s,t|y)}{p(s|y)p(t|y)}]].
$$

Lemma 1. *Given $\{x_i, y\}_{i=1}^N$ that is sampled i.i.d. from $p_{\mathcal{D}}(x|y)$, $s_i = F_s(x_i)$, and $t_i = F_t(x_i)$, $I(S;T|\,Y)$ is lower bounded by $-\mathcal{L}_{\mathit{info}} = \mathbb{E}_{y\sim p_{\mathcal{D}}(y)}\big[\,\mathbb{E}_{\{x_i,y\}_{i=1}^N}[\frac{1}{N}\sum_{i=1}^N \log \frac{e^{f_y(s_i,t_i)}}{\sum_{j=1}^N e^{f_y(s_i,t_j)}} + \log N]\big]$, and f_y is a score function indexed by y.*

Robust Informative Fine-Tuning (RIFT)

- Overall Objective: $I(S;\,Y,T) = I(S;Y) + I(S;T|\,Y)$

$$\min_{\theta,\phi,\varphi} \mathcal{L}_{\text{r-task}} + \alpha\mathcal{L}_{\text{r-info}},$$



$$\min_{\theta,\phi} \mathcal{L}_{\text{r-task}}, \quad \mathcal{L}_{\text{r-task}} = \mathop{\mathbb{E}}_{x,y\sim p_{\mathcal{D}}}\Big[-\log q(y|F_s(x)) + \beta \mathrm{KL}\big(q(\cdot|F_s(x))\|q(\cdot|F_s(\hat{x}))\big)\Big],$$

$$\min_{\theta,\varphi} \mathcal{L}_{\text{r-info}}, \quad \mathcal{L}_{\text{r-info}} = \mathop{\mathbb{E}}_{y\sim p_{\mathcal{D}}(y)}\Big[\mathop{\mathbb{E}}_{\{x_i,y\}_{i=1}^{N}\sim p_{\mathcal{D}}(x|y)}\big[\frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_y(\hat{s}_i,t_i)}}{\sum_{j=1}^{N} e^{f_y(\hat{s}_i,t_j)}} - \log N\big]\Big],$$

Experimental Results

Table 1: Accuracy(%) of different fine-tuning methods under attacks on IMDB.

Method	Model	Genetic	PWWS
Standard	BERT	38.1±2.5	40.7±1.1
Adv-Base	BERT	74.8±0.4	68.3±0.3
Adv-PTWD	BERT	73.9±0.4	69.1±0.7
Adv-Mixout	BERT	75.4±0.7	68.8±0.6
RIFT	BERT	**77.2**±0.8	**70.1**±0.5

(a) Accuracy (%) based on BERT-base-uncased.

Method	Model	Genetic	PWWS
Standard	RoBERTa	42.1±2.1	45.6±3.1
Adv-Base	RoBERTa	70.3±1.2	63.3±0.7
Adv-PTWD	RoBERTa	69.3±1.4	64.4±0.3
Adv-Mixout	RoBERTa	70.6±1.0	63.9±1.3
RIFT	RoBERTa	**73.5**±0.8	**66.3**±0.7

(b) Accuracy (%) based on RoBERTa-base.

Table 2: Accuracy(%) of different fine-tuning methods under attacks on SNLI.

Method	Model	Genetic	PWWS
Standard	BERT	40.1±0.7	19.4±0.4
Adv-Base	BERT	75.7±0.5	72.9±0.2
Adv-PTWD	BERT	75.2±1.0	72.6±0.5
Adv-Mixout	BERT	76.3±0.8	73.2±1.0
RIFT	BERT	**77.5**±0.9	**74.3**±1.1

(a) Accuracy (%) based on BERT-base-uncased.

Method	Model	Genetic	PWWS
Standard	RoBERTa	43.4±1.2	20.4±1.0
Adv-Base	RoBERTa	82.6±0.6	79.9±0.7
Adv-PTWD	RoBERTa	81.2±0.8	78.9±0.7
Adv-Mixout	RoBERTa	82.6±0.9	80.6±0.3
RIFT	RoBERTa	**83.5**±0.8	**81.1**±0.4

(b) Accuracy (%) based on RoBERTa-base.

Experimental Results

Table 3: Accuracy(%) of RIFT with maximizing $I(S;T|\,Y)$ and $I(S;T)$ respectively.

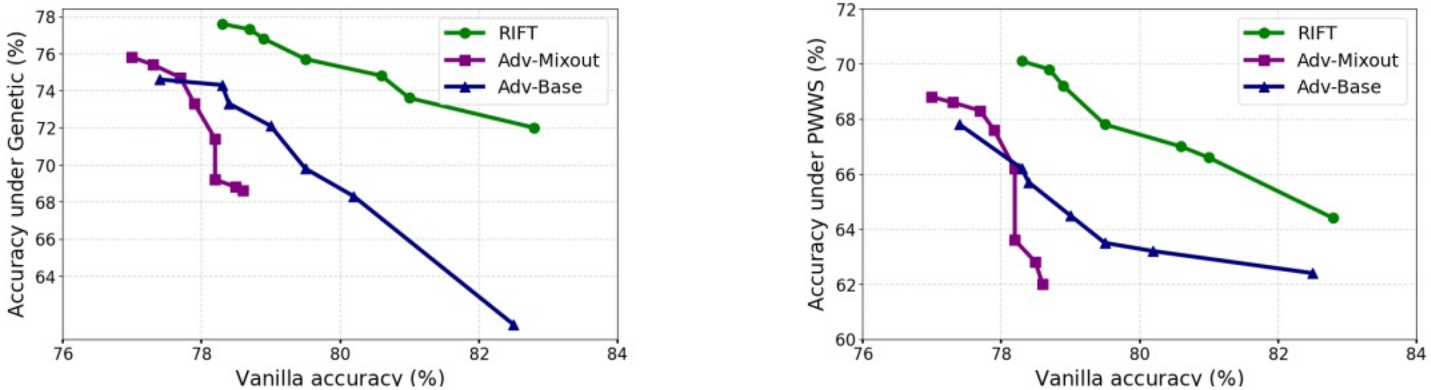
Maximizing	Model	Genetic	PWWS
$\mathbf{I(S;T|\,Y)}$	BERT	**77.2**	**70.1**
$I(S;T)$	BERT	76.1	69.4
$\mathbf{I(S;T|\,Y)}$	RoBERTa	**73.5**	**66.3**
$I(S;T)$	RoBERTa	72.0	65.3

(a) Accuracy (%) under attacks on IMDB.

Maximizing	Model	Genetic	PWWS
$\mathbf{I(S;T|\,Y)}$	BERT	**77.5**	**74.3**
$I(S;T)$	BERT	76.6	72.1
$\mathbf{I(S;T|\,Y)}$	RoBERTa	**83.5**	**81.1**
$I(S;T)$	RoBERTa	82.5	79.4

(b) Accuracy (%) under attacks on SNLI.

Experimental Results



(a) Accuracy (%) under Genetic attacks.

(b) Accuracy (%) under PWWS attacks.

Figure 4: Tradeoff curve between robustness and vanilla accuracy of BERT-based model on IMDB.

Conclusion

- Propose RIFT to fine-tune a pre-trained language model towards robust down-stream performance.
- Only conduct experiments under word substitution attack.