



TaiChi: Improving the Robustness of NLP Models by Seeking Common Ground While Reserving Differences

Huimin Chen[†], Chengyu Wang[‡], Yanhao Wang[†], Cen Chen[†], Yinggui Wang[§]

[†]School of Data Science and Engineering, East China Normal University, Shanghai, China

[‡]Alibaba Group, Hangzhou, China

[§]Ant Group, Hangzhou, China

email: saichen@stu.ecnu.edu.cn

Sentence	Label	Predict
<i>perfect</i> performance by xxx	Positive	Positive
<i>spotless</i> performance by xxx	Positive	Negative

Table 1: Adversarial example generated by PWWS (Ren et al., 2019) for a BERT-based sentiment classification model.

Pre-trained Language Models (PLMs) are vulnerable to adversarial examples

01 Background

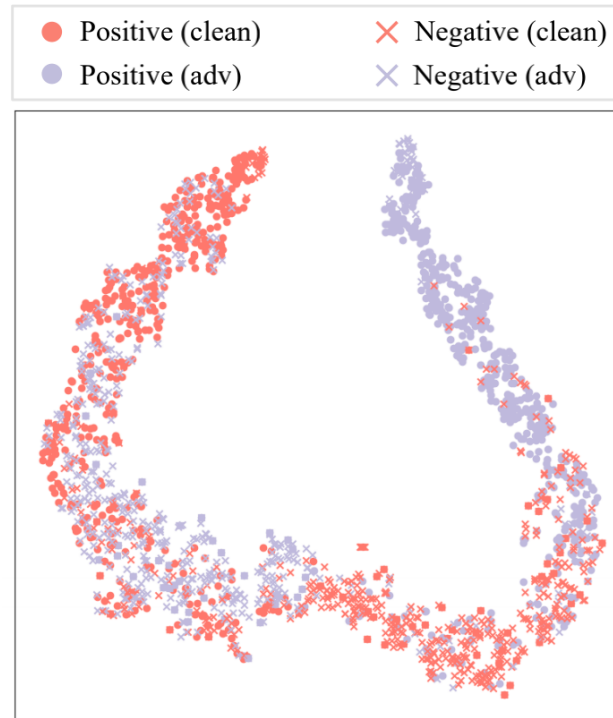


Figure 1: Illustration of BERT sentence representations of clean examples and their adversarial counterparts generated by PWWS on the SST-2 dataset.

We argue that augmenting the dataset with adversarial samples having the same label as their counterparts is not a reasonable approach.

02 Our Approach

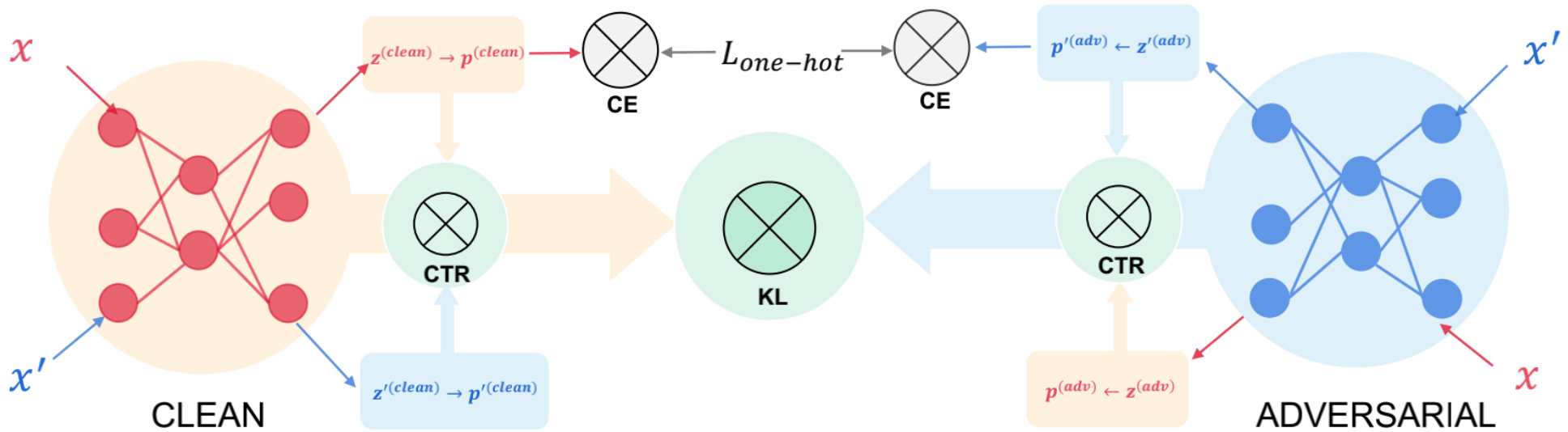


Figure 2: Framework of our method *TaiChi*. The neural network in red (CLEAN) signifies the model trained only on clean samples for classification, whereas the network in blue (ADVERSARIAL) denotes the model designed for adversarial examples. Here, x and x' refer to a clean sample and its adversarial counterpart, respectively. Additionally, z, p and z', p' represent the vector representations and the corresponding logits for x and x' , respectively. Finally, $L_{one-hot}$, **CE**, **CTR**, and **KL** denote the one-hot labeling process, the text classification task, the contrastive data augmentation task, and the model-level fusion task, respectively.

Two models

Three tasks

03 Experiments

Dataset	Method	<i>clean</i> %	<i>deepwordbug</i> %	<i>pwws</i> %	<i>textfooler</i> %	<i>textbugger</i> %
SST-2	Base _{<i>clean</i>}	92.6	21.1	16.0	7.1	37.3
	PGD	92.5	22.8	22.5	10.6	39.9
	FGSM	90.9	26.5	21.9	12.5	36.9
	FreeLB	91.7	21.7	19.0	8.3	35.1
	SMART	92.6	25.3	20.2	12.8	38.8
	ADA	90.8	22.5	28.9	17.3	38.1
	TaiChi	91.7 (↓ 0.9)	34.1 (↑ 7.6)	34.8 (↑ 5.9)	20.3 (↑ 3.0)	51.0 (↑ 11.1)
TREC	Base _{<i>clean</i>}	94.6	42.6	53.4	42.2	64.8
	PGD	95.4	51.0	59.0	46.4	66.6
	FGSM	94.8	48.4	53.2	40.0	62.4
	FreeLB	95.0	52.4	54.0	40.0	62.8
	SMART	94.8	48.6	53.0	40.8	62.8
	ADA	91.9	54.6	64.0	40.4	72.4
	TaiChi	93.9 (↓ 1.5)	61.0 (↑ 6.4)	68.0 (↑ 4.0)	51.8 (↑ 5.4)	75.2 (↑ 2.8)
AG	Base _{<i>clean</i>}	94.4	15.7	25.5	10.0	31.5
	PGD	94.4	41.3	19.6	44.8	19.3
	FGSM	93.4	41.3	58.0	35.6	42.7
	FreeLB	94.0	40.0	55.8	35.2	43.2
	SMART	93.0	18.7	22.7	6.8	29.5
	ADA	92.9	45.4	48.9	35.9	24.1
	TaiChi	93.8 (↓ 0.6)	48.8 (↑ 3.4)	65.7 (↑ 7.7)	45.3 (↑ 0.5)	50.7 (↑ 7.5)

Table 5: Results of different defense methods against four types of adversarial attacks. For ADA and TaiChi, we report their average *clean*% of four types of attacks. For TaiChi, we compare its score on each measure with the best among all remaining methods (in bracket).

Improved robustness and lower generalization costs.

03 Experiments

PLM	Method	<i>clean%</i>	<i>deepwordbug%</i>	<i>pwws%</i>	<i>textfooler%</i>	<i>textbugger%</i>
BERT-Large	Base _{clean}	93.5	15.8	16.6	6.5	36.4
	ADA	90.9	26.8	21.3	11.7	43.5
	TaiChi	92.9 (↓ 0.6)	34.9 (↑ 8.1)	33.2 (↑ 11.9)	15.4 (↑ 3.7)	53.6 (↑ 10.1)
DeBERTa	Base _{clean}	93.0	22.9	20.9	8.8	40.7
	ADA	92.5	29.6	30.5	13.7	49.2
	TaiChi	92.6 (↓ 0.4)	39.5 (↑ 9.9)	37.1 (↑ 6.6)	19.3 (↑ 5.6)	52.7 (↑ 3.5)

Table 6: Additional results of Base_{clean}, ADA, and TaiChi against four types of adversarial attacks on the SST-2 dataset when BERT-Large and DeBERTa are used as base models.

These effects should also be verified across various pretrained models of different sizes or types.

03 Experiments

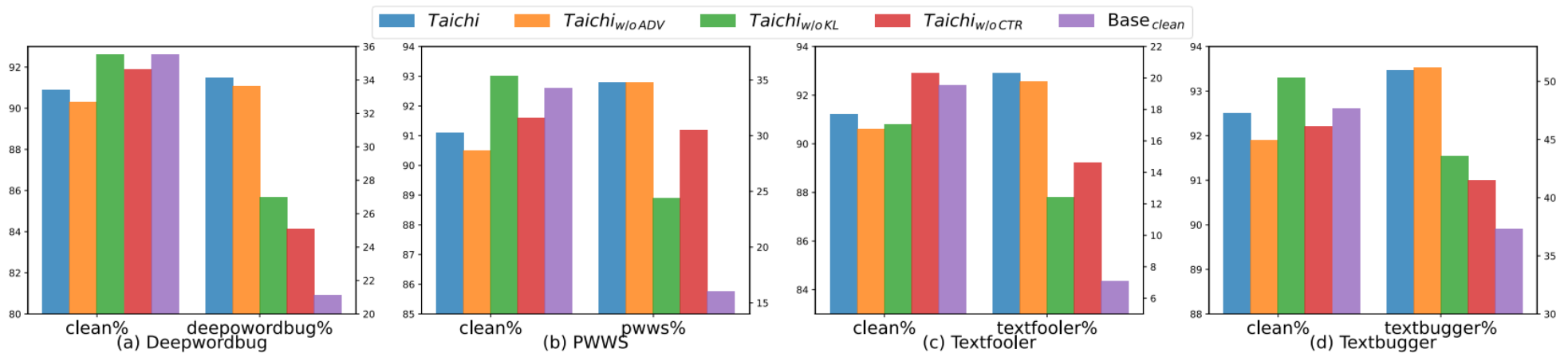


Figure 4: Ablation studies for TaiChi on the SST-2 dataset.

The ablation study confirms that different modules of our method have specific impacts.

03 Experiments

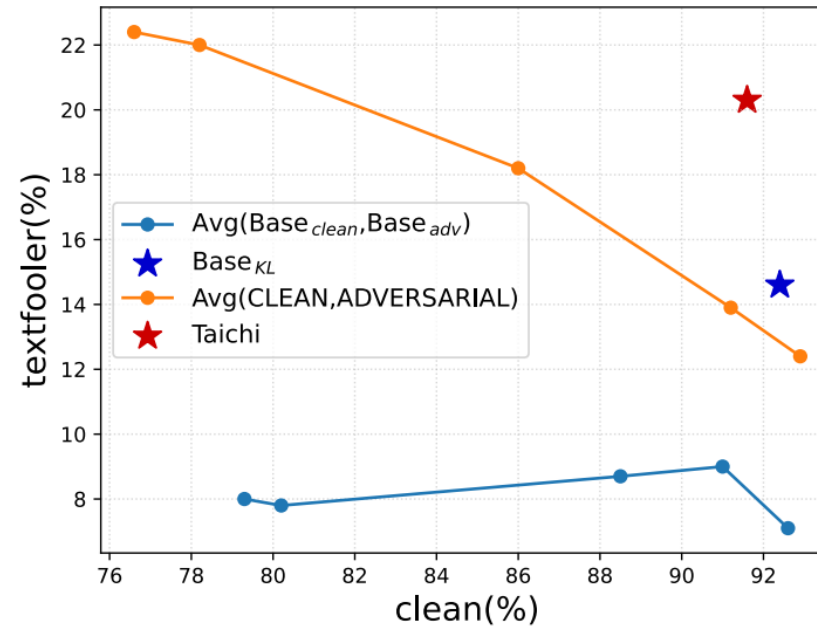


Figure 5: KL divergence loss vs. linear average for model-level fusion on the SST-2 dataset under the TextFooler attack.

This additional experiment demonstrate that our model-level fusion method outperforms the typical linear average method.

03 Experiments

bert-base-uncased	Before Fine-Tune	Base _{clean}	TaiChi _{w/o KL}	TaiChi _{w/o CTR}	TaiChi _{w/o ADV}	TaiChi
MSE (pos, ↓)	0.0393	0.2830	<u>0.0638</u>	0.0966	0.0304	0.0303
MSE (neg, ↑)	0.1069	0.5935	0.7563	0.4713	0.5848	<u>0.6275</u>
KL (pos, ↓)	0.0030	1.3690	0.5222	<u>0.2207</u>	0.1771	0.1206
KL (neg, ↑)	0.0065	<u>4.2273</u>	4.7785	4.0947	2.1123	2.3727
CTR (↓)	6.6380	6.8853	<u>6.0832</u>	6.5214	6.0550	6.0305
clean% (↑)	49.6	<u>92.6</u>	92.9	92.4	90.4	91.6
textfooler% (↑)	1.0	7.1	12.4	<u>14.6</u>	19.8	20.3

Table 7: High-order statistical information exploring the abilities of different models to learn text representations on the SST-2 dataset under the TextFooler attack. Here, the average MSE and KL-divergence losses between *positive* and *negative* sample pairs and the average contrastive learning loss are computed from 1K clean and adversarial sample pairs. Each pair of clean and adversarial samples forms a *positive* sample pair. The *negative* pairs are constructed by assigning a fixed sentence with a negative label to all clean samples with positive labels, and vice versa.

Our method's effectiveness was further validated from the perspective of distribution in vector representation.



Thank you for your time!