中国科学院软件研究所学术年会'2022
暨计算机科学国家重点实验室开放周

学术论文

Word Level Robustness Enhancement: Fight Perturbation with Perturbation

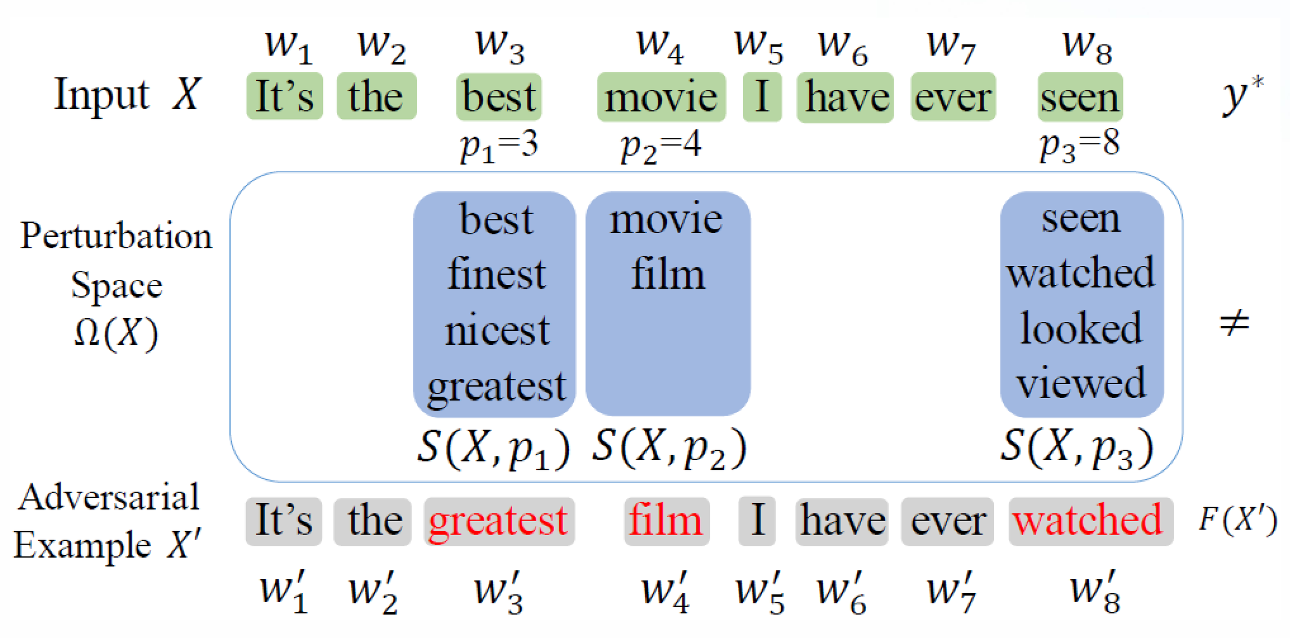
词语级鲁棒性加强：用扰动打败扰动

Pei Huang*, Yuting Yang*, Fuqi Jia, Minghao Liu, Feifei Ma✉, Jian Zhang✉

AAAI 2022

联系人：黄 沛 huangpei@ios.ac.cn 18201101474

Word-level Robustness

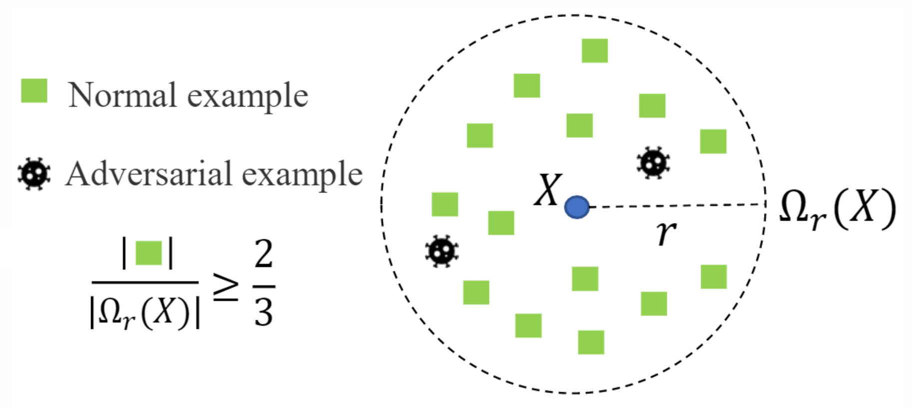


Word substitution alerts NN's prediction, which can be especially dangerous in some applications like privacy detection and fake news detection.

Weak Robustness

Definition 3 (Weak Robustness). *If the value of $PR > 2/3$, f is said to be **weakly robust** on the perturbation space $\Omega_r(X)$, where PR is defined as:*

$$PR := \frac{|\{X' : X' \in \Omega_r(X) \wedge f(X') = y^*\}|}{|\Omega_r(X)|} \qquad (2)$$



Model	Dataset	$PR > 2/3$
BiLSTM	MR	96.84%
	IMDB	96.94%
	SNLI	85.95%
BERT	MR	98.88%
	IMDB	97.61%
	SNLI	96.96%

An interesting phenomenon:
Adversarial examples are everywhere but occupy a small ratio in the perturbation spaces!

Our Defense Method: Fight Perturbation with Perturbation (FPP)

Algorithm 1: Enhancement Classifier F

Input: X
Parameter: A base classifier f
Output: Prediction $\tilde{y}$

for all $p \in P$ **do**
  $s \sim U(0,1)$;
  if $\Delta_{12}(p) > s$ **then**
    $X_{w_p \leftarrow w_{p*}}$ \\ Replace w_p via w_{p*}
  end if
end for
$N \leftarrow -2\ln\epsilon/(2*2/3-1)$
$r \leftarrow \kappa n$
for $i \leftarrow 0$ **to** $N-1$ **do**
  $X_i \sim \Omega_r(X)$;
  $l_i \leftarrow f(X_i)$;
end for
$\tilde{y} \leftarrow \arg\max_{y \in \mathcal{Y}} \sum_{i=0}^{N-1} \mathbb{I}(l_i = y)$
return $\tilde{y}$

Fight Fire with Fire!（以彼之道，还施彼身！）

Step 1: Input perturbation (Destroy the subtle combination of attacker via perturbing.)

Step 2: Random perturbation & voting (Based on weak robustness property, enhance the prediction result via the voting of random perturbations.)

If PR in weak robustness definition is $2/3$ and we want the error rate of enhanced prediction $\epsilon = 10^{-5}$, sample size N in Step 2 needs to satisfy:

$$N > \frac{-2\ln\epsilon}{(2PR-1)^2} > 207$$

Defense Results

Dataset	Method	LSTM Acc	LSTM Textfooler Suc↓	LSTM Textfooler Rob	LSTM SemPSO Suc↓	LSTM SemPSO Rob	BERT Acc	BERT Textfooler Suc↓	BERT Textfooler Rob	BERT SemPSO Suc↓	BERT SemPSO Rob
MR	f	82.47	69.70	25.00	81.82	15.00	89.60	48.35	47.00	73.63	24.00
	Adv	79.85	65.82	27.00	82.27	14.00	**88.00**	35.91	58.00	73.08	24.50
	FGWS	78.73	56.60	34.50	76.73	18.50	83.88	23.98	65.00	56.14	37.50
	SAFER	77.60	22.08	60.00	27.10	56.50	86.32	**7.30**	**82.00**	13.50	67.50
	F	**81.16**	**14.65**	**67.00**	**25.79**	**59.00**	87.72	8.89	**82.00**	**10.87**	**72.50**
IMDB	f	89.94	86.24	13.00	99.45	0.50	93.68	82.63	16.5	92.51	7.00
	Adv	87.64	71.03	25.50	99.95	0.50	91.00	38.95	58.00	58.42	41.00
	FGWS	85.70	77.84	19.50	92.61	6.50	89.60	62.30	34.50	88.52	40.50
	SAFER	86.60	13.97	77.00	25.28	66.50	88.00	7.07	85.50	-	-
	F	**89.30**	**3.70**	**91.00**	**9.89**	**82.00**	**93.40**	**3.11**	**93.50**	-	-
SNLI	f	84.35	72.05	22.50	50.93	39.50	86.77	69.94	26.00	71.10	25.00
	Adv	**84.35**	75.16	20.00	60.87	31.50	82.53	52.98	39.50	54.17	38.50
	FGWS	72.40	38.06	41.50	37.31	42.00	75.60	44.06	40.00	44.76	39.50
	SAFER	56.60	**19.66**	47.00	**17.24**	48.00	67.00	26.90	53.00	27.08	52.50
	F	80.27	22.22	**59.50**	26.53	**54.00**	**83.90**	**15.34**	**69.00**	**24.84**	**59.00**

Table 2: Robustness evaluation results of different defense methods. Acc is the clean accuracy on test set. Suc is the successful attacking ratio. Rob is the robustness accuracy. Only for Suc, the lower the value, the better the defense capability of the model. It is noted with ↓. The numbers in bold denote the best performance for the metric.

- **FPP achieves the highest robustness accuracy (Rob) on all three data sets and two different models.**
- **FPP has a good trade-off between clean accuracy (Acc) and robustness (Rob).**