38th Conference on Uncertainty in Artificial Intelligence
Eindhoven, Netherlands
August 1-5, 2022
uai2022

Detecting Textual Adversarial Examples through Randomized Substitution and Vote

Xiaosen Wang, Yifeng Xiong, Kun He
School of Computer Science and Technology, Huazhong University of Science and Technology

Introduction

Natural language processing models are known to be vulnerable to adversarial examples and **synonym substitution based attacks** are widely adopted for generating textual adversarial examples. Various defense methods have been proposed to mitigate the threat of textual adversarial examples, *e.g.* adversarial training, input transformations, detection, *etc.*

In this work, we propose a **simple yet effective detection method called RS&V** to resist adversarial attacks.

Motivation

- **Replacement sequence.** We regard the optimization process of synonym-based attacks as searching a specific sequence for word replacement, in which **the words mutually influence each other and contribute together to mislead the target classifier.**
- **Hypothesis.** We can eliminate the perturbation if we **break the mutual interaction of the words in the replacement sequence**.
- **Observation.** Randomly substituting words with its synonyms could consistently and significantly **improve the robust accuracy against adversarial examples while maintaining the high clean accuracy** under various substitution rates.
- **RS&V.** We detect adversarial examples to **vote** the prediction label by accumulating the logits of k samples generated by **randomly substituting the words in the input text with synonyms**.

Algorithm

Algorithm 1 The RS&V Algorithm

Input: Input text $x = \{w_1, w_2, ..., w_n\}$, target classifier f, substitution rate p, number of votes k, stopword selection portion s.
Output: Detection result and restored label

Calculate the stopword set $\mathcal{W}$ containing the top s high frequency words in the training set
Initialize converted text set $\mathcal{X} = \emptyset$
for $i = 1 \to k$ **do** ▷ Randomized Substitution
  Initialize a new text $x_i = x$
  Randomly sample $n \cdot p$ words for $\mathcal{P}$ from $x_i / \mathcal{W}$
  for each word $w_t \in \mathcal{P}$ **do**
    Randomly select a synonym $\hat{w}_t^j \in \mathcal{S}(w_t)$
    Substitute $w_t \in x_i$ with $\hat{w}_t^j$
  end for
  $\mathcal{X} = \mathcal{X} \cup x_i$
end for
Calculate the prediction label for input text x: $\bar{y} = \arg\max f(x)$
▷ Vote & Detection
Calculate the voted label: $\bar{y}_v = \arg\max \sum_{i=1}^{k} f(x_i)$
if $\bar{y} = \bar{y}_v$ **then**
  return False, $\bar{y}$ ▷ Benign sample
end if
return True, $\bar{y}_v$ ▷ Adversarial example

The framework of RS&V

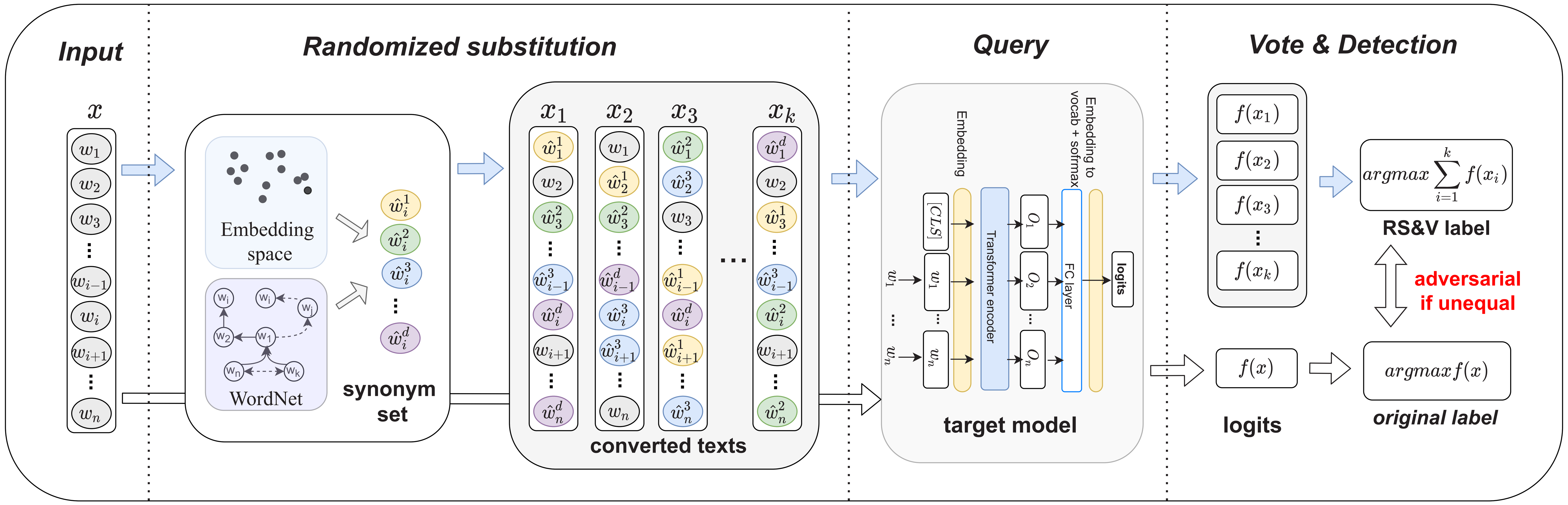

Figure 1: The overall framework of the proposed RS&V method.

Experimental Results

Table 1: The classification accuracy (%) and F1 score (%) of various detection methods for Word-CNN and BERT on three datasets. N/A denotes the normally trained model without the detection module.

Dataset	Model	Method	Clean	GA		PWWS		PSO		Textfooler		HLA		Average	
				Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
AG's News	CNN	N/A	92.1	43.6	–	37.1	–	36.4	–	24.8	–	41.9	–	36.8	–
		DISP	91.6	77.3	83.7	76.5	85.0	78.0	85.8	69.0	80.8	79.2	85.9	76.0	84.2
		FGWS	91.3	76.2	80.3	75.8	83.4	76.2	84.0	77.5	88.3	80.0	85.8	77.1	84.4
		RS&V	91.3	**84.1**	**90.2**	**85.1**	**92.2**	**86.8**	**94.0**	**86.0**	**94.8**	**88.1**	**94.8**	**86.0**	**93.2**
	BERT	N/A	94.9	68.5	–	74.9	–	59.2	–	61.3	–	62.0	–	65.2	–
		DISP	94.5	85.3	77.8	85.2	70.3	83.8	80.9	83.0	81.1	85.6	84.2	84.6	78.9
		FGWS	94.6	87.2	82.9	88.8	82.5	87.7	88.0	89.2	91.5	89.1	90.5	88.4	87.1
		RS&V	94.6	**90.5**	**90.4**	**91.3**	**88.8**	**91.7**	**94.2**	**92.5**	**96.2**	**92.3**	**90.8**	**92.8**	**92.1**
	RoBERTa	N/A	93.5	71.9	–	78.7	–	66.8	–	74.6	–	68.8	–	72.2	–
		DISP	93.4	84.8	75.8	84.8	65.3	85.9	84.9	82.8	67.3	86.8	85.3	85.0	75.7
		FGWS	93.1	86.2	80.0	87.2	74.6	87.0	86.6	88.5	85.2	88.3	88.9	87.4	83.1
		RS&V	93.4	**89.6**	**88.9**	**90.3**	**87.0**	**91.5**	**94.7**	**91.2**	**92.6**	**91.5**	**94.7**	**90.8**	**91.6**
IMDB	CNN	N/A	87.2	6.2	–	1.5	–	2.7	–	0.6	–	17.4	–	5.7	–
		DISP	87.2	48.8	68.3	43.1	64.8	53.3	74.4	39.3	61.2	62.0	77.4	49.3	69.2
		FGWS	86.5	64.7	82.8	64.7	84.0	69.7	87.5	72.6	90.0	72.8	87.4	68.9	86.3
		RS&V	86.3	**79.6**	**94.0**	**80.2**	**94.8**	**80.9**	**95.0**	**79.2**	**94.1**	**81.7**	**94.5**	**80.3**	**94.5**
	BERT	N/A	91.9	15.4	–	26.7	–	5.6	–	9.5	–	15.7	–	14.6	–
		DISP	91.8	64.3	77.5	63.7	72.2	68.7	84.1	62.0	77.6	74.5	86.7	66.6	79.6
		FGWS	92.5	80.6	90.9	79.5	88.2	82.0	92.9	83.0	93.2	84.8	94.0	82.0	91.8
		RS&V	92.1	**87.8**	**96.0**	**88.2**	**95.4**	**88.5**	**96.9**	**89.1**	**97.2**	**89.9**	**97.2**	**88.7**	**96.5**
	RoBERTa	N/A	94.2	18.3	–	29.9	–	7.0	–	34.3	–	21.8	–	22.3	–
		DISP	93.9	66.3	77.4	64.1	70.0	67.4	81.7	68.7	73.3	76.7	86.1	68.6	77.7
		FGWS	94.4	81.0	90.1	82.0	89.1	83.2	92.9	86.6	92.7	85.7	93.4	83.7	91.6
		RS&V	94.6	**89.4**	**95.9**	**88.8**	**94.7**	**91.0**	**97.4**	**91.4**	**96.5**	**90.8**	**96.9**	**90.3**	**96.3**
Yahoo! Answers	CNN	N/A	69.5	4.7	–	5.6	–	2.6	–	3.9	–	4.3	–	4.2	–
		DISP	69.8	37.4	67.1	35.6	63.8	39.3	70.6	35.9	66.5	45.0	76.3	38.6	68.9
		FGWS	68.0	49.7	82.6	48.2	80.9	49.4	82.2	40.6	72.1	39.9	75.0	45.6	78.6
		RS&V	69.3	**63.0**	**92.6**	**62.1**	**92.8**	**63.2**	**93.3**	**61.6**	**93.2**	**62.6**	**91.9**	**62.5**	**92.8**
	BERT	N/A	76.7	13.8	–	25.6	–	8.9	–	17.9	–	11.5	–	15.5	–
		DISP	76.7	50.0	74.5	50.5	68.6	53.8	80.4	51.7	74.8	56.0	81.9	52.4	76
		FGWS	75.7	62.2	88.0	62.7	85.9	62.4	88.7	66.0	90.4	65.5	91.1	63.8	88.8
		RS&V	75.8	**67.4**	**92.3**	**68.7**	**91.0**	**69.7**	**93.7**	**71.6**	**94.1**	**70.0**	**93.9**	**69.5**	**93.0**
	RoBERTa	N/A	74.7	19.8	–	33.7	–	15.2	–	41.7	–	19.6	–	26.0	–
		DISP	74.7	48.0	68.7	50.4	61.3	50.9	75.1	53.7	57.6	55.2	78.6	51.7	68.3
		FGWS	74.8	62.3	87.5	64.9	85.9	63.3	88.5	67.2	86.3	65.7	90.1	64.7	87.7
		RS&V	76.0	**66.4**	**90.3**	**66.8**	**86.7**	**68.1**	**92.2**	**68.3**	**86.8**	**68.7**	**92.8**	**67.7**	**89.8**

Conclusion

We propose a novel detection method RS&V against synonym substitution based adversarial attacks for text classification.

- **Novelty.** We identify that randomized synonym substitution could destroy the mutual interaction among words in the replacement sequence for adversarial attacks. Based on this observation, we propose RS&V to effectively detect adversarial examples.
- **Effectiveness.** Empirical evaluations demonstrate that RS&V could achieve better detection performance than existing baselines while maintaining a high performance on benign samples.
- **Generality.** RS&V is generally applicable to all existing deep neural networks without any additional training or modification on the model architectures.
- **Impact.** RS&V identifies the fragility of textual adversarial examples, which might inspire more defense and detection methods by pre-processing the text without degrading clean accuracy.