Adversarial Training with Fast Gradient Projection Method against Synonym Substitution based Text Attacks

Xiaosen Wang[1], Yichen Yang[1], Yihe Deng[2] and Kun He[1]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2] Computer Science Department, University of California, Los Angeles

Introduction

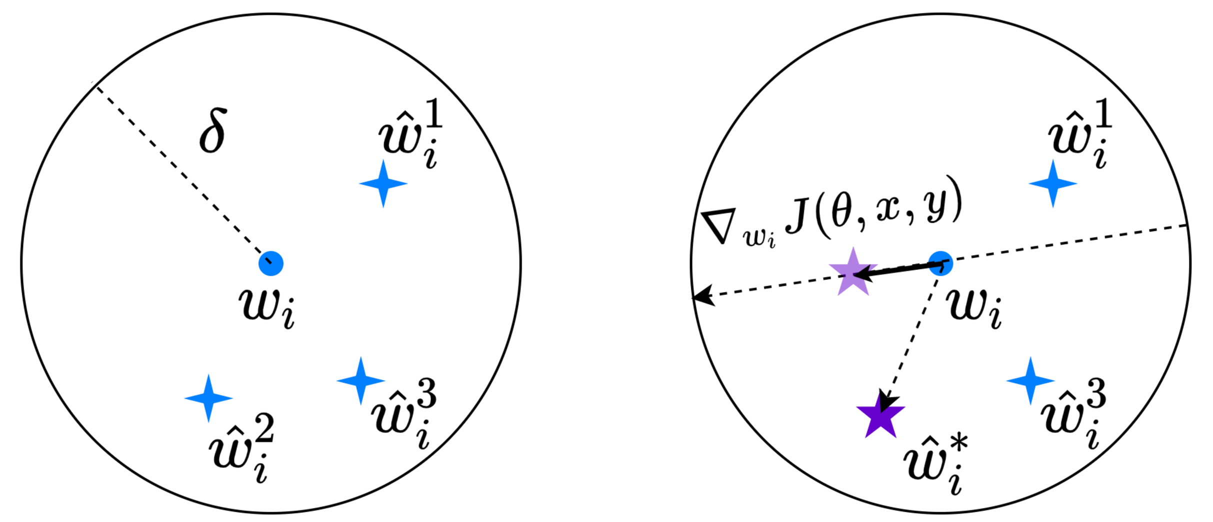
Based on the metrics to evaluate the dissimilarity of two texts, current adversarial attacks can be split into three categories.

- Character Level Attack
- Sentence Level Attack
- Word Level Attack
  - Adding or removing word
  - **Synonym substitution**

Due to the discrete input space and semantic constraints, existing synonym substitution based attacks are **black-box attacks**, that needs thousands of queries on target model and is **time-consuming**.

Goal: **Proposing a synonym substitution based attack with high efficiency and introducing adversarial training as an effective defense against synonym substitution based attack.**

Methods



(a) (b)

Figure 1: Strategies to pick optimal synonym to substitute word w_i.

Fast Gradient Projection Method (FGPM) is a **gradient based** synonym substitution text attack with three steps:

1. Constructing the Synonyms Set

$$S(w_i, \delta) = \{\hat{w}_i \in \mathcal{D} \mid \|\hat{w}_i - w_i\|_2 \leq \delta\}. \quad (1)$$

2. Finding the Optimal Synonym for Each Word

$$\hat{w}_i^* = \mathop{\arg\max}_{\hat{w}_i^j \in S(w_i,\delta)} (\hat{w}_i^j - w_i) \cdot \nabla_{w_i} J(\theta, x, y). \quad (2)$$

3. Determining the Substitution Order

$$\hat{w}_* = \mathop{\arg\max}_{\hat{w}_i^* \in \mathcal{C}_s} (\hat{w}_i^* - w_i) \cdot \nabla_{w_i} J(\theta, x, y). \quad (3)$$

With the high efficiency of FGPM, we further propose **Adversarial Training with FGPM enhanced by Logit pairing (ATFL)**:

$$\tilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1-\alpha) J(\theta, x_{adv}, y) + \lambda \|F(x, \cdot) - F(x_{adv}, \cdot)\|.$$

Algorithm for FGPM

Algorithm 1 The FGPM Algorithm

Input: Benign sample $x = \langle w_1, \cdots, w_i, \cdots, w_n \rangle$
Input: True label y for x
Input: Target classifier ϕ
Input: Upper bound distance for synonyms δ
Input: Maximum number of iterations N
Input: Upper bound for word substitution ratio ϵ
Output: Adversarial example x_{adv}

Initialize $x_{adv}^0 = x$
Calculate $S(w_i, \delta)$ by Eq. (1) for $w_i \in x_{adv}^0$
for $k = 1 \to N$ **do**
  Construct candidate set $\mathcal{C}_s = \{\hat{w}_1^*, \cdots, \hat{w}_i^*, \cdots, \hat{w}_n^*\}$ by Eq. (2)
  Calculate optimal word $\hat{w}_*$ by Eq. (3)
  Substitute $w_* \in x_{adv}^{k-1}$ with $\hat{w}_*$ to obtain x_{adv}^k
  if $\phi(x_{adv}^k) \neq y$ and $R(x_{adv}^k, x) < \epsilon$ **then**
    return x_{adv}^k ▷ Succeed
  end if
end for
return None ▷ Failed

Evaluation on FGPM

	AG's News			*DBPedia*			*Yahoo! Answers*		
	CNN	LSTM	Bi-LSTM	CNN	LSTM	Bi-LSTM	CNN	LSTM	Bi-LSTM
No Attack†	92.3	92.6	92.5	98.7	98.8	99.0	72.3	75.1	74.9
No Attack	87.5	90.5	88.5	99.5	99.0	99.0	71.5	72.5	73.5
Papernot'	72.0	61.5	65.0	80.5	77.0	83.5	38.0	43.0	36.5
GSA	45.5	35.0	40.0	52.0	49.0	53.5	21.5	19.5	19.0
PWWS	37.5	30.0	29.0	55.5	52.5	50.0	5.5	12.5	11.0
IGA	**30.0**	**26.5**	**25.5**	**36.5**	**38.5**	**37.0**	**3.5**	**5.5**	**7.0**
FGPM	37.5	31.0	32.0	40.0	45.5	47.5	6.0	17.0	10.5

Table 1: The classification accuracy (%) of various models under attacks.

	CNN	LSTM	Bi-LSTM	CNN	LSTM	Bi-LSTM	CNN	LSTM	Bi-LSTM
Papernot'	72.0*	80.5	82.5	83.5	61.5*	78.5	79.5	74.5	65.0*
GSA	45.5*	80.0	80.0	84.5	35.0*	73.0	81.5	72.5	40.0*
PWWS	37.5*	**70.5**	**70.0**	83.0	30.0*	**67.5**	80.0	**67.5**	29.0*
IGA	30.0*	74.5	74.5	84.0	26.5*	71.5	79.0	71.0	25.5*
FGPM	37.5*	72.5	74.5	**81.0**	31.0*	73.5	**77.5**	**67.5**	32.0*

Table 2: The classification accuracy (%) of different models for adversaries generated on other models on *AG's News* for transferability evaluation.

	AG's News			*DBPedia*			*Yahoo! Answers*		
	CNN	LSTM	Bi-LSTM	CNN	LSTM	Bi-LSTM	CNN	LSTM	Bi-LSTM
Papernot'	74	1,676	4,401	145	2,119	6,011	120	9,719	19,211
GSA	276	643	713	616	1,006	1,173	1,257	2,234	2,440
PWWS	122	28,203	28,298	204	34,753	35,388	643	98,141	100,314
IGA	965	47,142	91,331	1,369	69,770	74,376	893	132,044	123,976
FGPM	**8**	**29**	**29**	**8**	**34**	**33**	**26**	**193**	**199**

Table 3: The total running time (in seconds) for generating 200 adversarial instances.

Evaluation on ATFL

Dataset	Attack	CNN				LSTM				Bi-LSTM			
		NT	SEM	IBP	ATFL	NT	SEM	IBP	ATFL	NT	SEM	IBP	ATFL
AG's News	No Attack†	**92.3**	89.7	89.4	91.8	**92.6**	90.9	86.3	92.0	**92.5**	91.4	89.1	92.1
	No Attack	87.5	87.5	87.5	**89.0**	90.5	90.5	84.5	**91.5**	88.5	**91.0**	87.0	89.5
	Papernot'	72.0	84.5	87.5	**88.0**	61.5	89.5	81.5	**90.0**	65.0	**90.0**	86.0	89.0
	GSA	45.5	80.0	86.0	**88.0**	35.0	85.5	79.5	**88.0**	40.0	**87.5**	79.0	**87.5**
	PWWS	37.5	80.5	86.0	**88.0**	30.0	86.5	79.5	**88.0**	29.0	**87.5**	75.5	**87.5**
	IGA	30.0	80.0	86.0	**88.0**	26.5	85.5	79.5	**88.0**	25.5	**87.5**	79.0	**87.5**
	FGPM	37.5	78.5	86.5	**88.0**	31.0	85.5	80.0	**88.0**	32.0	84.5	80.0	**87.5**
DBPedia	No Attack†	**98.7**	98.1	97.4	98.4	**98.8**	98.5	93.1	98.7	**99.0**	98.7	94.7	98.6
	No Attack	**99.5**	97.5	97.0	98.0	99.0	**99.5**	95.0	**99.5**	**99.0**	98.0	94.5	**99.0**
	Papernot'	80.5	97.0	97.0	**98.0**	77.0	**99.5**	91.0	**99.5**	83.5	98.0	92.5	**99.0**
	GSA	52.0	96.0	97.0	**98.0**	49.0	**99.0**	84.5	98.5	53.5	98.0	89.5	**99.0**
	PWWS	55.5	95.5	97.0	**98.0**	52.5	**99.5**	84.0	98.5	50.0	95.0	89.5	**99.0**
	IGA	36.5	95.5	97.0	**98.0**	38.5	**99.0**	84.5	98.0	37.0	97.0	90.0	**99.0**
	FGPM	40.0	94.0	97.0	**98.0**	45.5	**99.0**	85.0	98.5	47.5	98.0	89.5	**99.0**
Yahoo! Answers	No Attack†	**72.3**	70.0	64.2	71.0	**75.1**	72.8	51.2	74.2	**74.9**	72.9	59.0	74.3
	No Attack	71.5	67.0	64.5	**72.0**	72.5	69.5	50.5	**74.0**	**73.5**	69.5	56.0	72.0
	Papernot'	38.0	64.0	63.5	**69.0**	43.0	67.0	41.0	**71.0**	36.5	66.5	53.0	**70.5**
	GSA	21.5	59.5	61.0	**63.0**	19.5	63.0	30.0	**69.5**	19.0	62.5	39.5	**64.5**
	PWWS	5.5	59.0	61.0	**62.5**	12.5	63.0	30.0	**68.5**	11.0	62.5	40.0	**65.5**
	IGA	3.5	59.0	61.0	**62.5**	5.5	62.5	31.5	**67.5**	7.0	62.0	40.5	**64.0**
	FGPM	6.0	61.0	63.0	**64.0**	17.0	63.0	35.0	**68.5**	10.5	**64.5**	41.5	63.5

Table 4: The classification accuracy (%) of three defense methods under various attacks.

Attack	CNN				LSTM				Bi-LSTM			
	NT	SEM	IBP	ATFL	NT	SEM	IBP	ATFL	NT	SEM	IBP	ATFL
Papernot'	72.0*	87.0	87.0	**88.5**	80.5	91.0	82.0	**92.0**	82.5	**91.0**	86.0	90.0
GSA	45.5*	87.0	87.0	**88.5**	80.0	90.5	83.0	**91.0**	80.0	**91.0**	87.5	90.0
PWWS	37.5*	87.0	87.0	**88.5**	70.5	**90.5**	83.0	**90.5**	70.0	**90.5**	86.5	90.0
IGA	30.0*	87.0	87.0	**88.5**	74.5	90.5	83.5	**91.0**	74.5	**90.5**	86.5	89.5
FGPM	37.5*	87.0	87.5	**88.5**	72.5	90.5	83.0	**91.5**	74.5	**91.0**	86.5	90.0

Table 5: The classification accuracy (%) of various models for adversaries crafted on CNN model on *AG's News* for evaluating the defense performance against transferability.

Model	Attack	NT	Standard	TRADES	MMA	MART	CLP	ALP
CNN	No Attack†	**92.3**	**92.3**	92.1	91.1	91.2	91.7	91.8
	No Attack	87.5	89.5	89.5	87.5	87.0	**90.5**	89.0
	Papernot'	72.0	85.5	67.0	83.5	83.5	73.0	**88.0**
	GSA	45.5	77.5	36.5	69.0	73.0	42.5	**88.0**
	PWWS	37.5	77.0	33.5	70.5	73.0	38.5	**88.0**
	IGA	30.0	75.0	29.0	67.5	72.0	30.0	**88.0**
	FGPM	37.5	78.0	40.0	73.5	74.5	38.5	**88.0**

Table 6: The classification accuracy (%) of CNN model adversarially trained with different regularization under various adversarial attacks on *AG's News*.

Conclusion

- We propose an efficient gradient based synonym substitution adversarial attack called FGPM, which is at least 20 times faster the existing fastest attack and achieves the similar attack performance and transferability.
- We introduce adversarial training into text domain against synonym substitution adversarial attacks which significantly improves the model robustness.
- We find that recent successful regularizations of adversarial training for image data actually degrade the performance of adversarial training in text domain, suggesting the need for more specialized adversarial training methods for text data.