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

Xiaosen Wang[1*], **Yichen Yang**[1*], **Yihe Deng**[2*], **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
{xiaosen, yangyc}@hust.edu.cn, yihedeng@g.ucla.edu, brooklet60@hust.edu.cn

Abstract

Adversarial training is the most empirically successful approach in improving the robustness of deep neural networks for image classification. For text classification, however, existing synonym substitution based adversarial attacks are effective but not very efficient to be incorporated into practical text adversarial training. Gradient-based attacks, which are very efficient for images, are hard to be implemented for synonym substitution based text attacks due to the lexical, grammatical and semantic constraints and the discrete text input space. Thereby, we propose a fast text adversarial attack method called *Fast Gradient Projection Method (FGPM)* based on synonym substitution, which is about 20 times faster than existing text attack methods and could achieve similar attack performance. We then incorporate FGPM with adversarial training and propose a text defense method called *Adversarial Training with FGPM enhanced by Logit pairing* (ATFL). Experiments show that ATFL could significantly improve the model robustness and block the transferability of adversarial examples.

Introduction

Deep Neural Networks (DNNs) have garnered tremendous success over recent years (Krizhevsky, Sutskever, and Hinton 2012; Kim 2014; Devlin et al. 2019). However, researchers also find that DNNs are often vulnerable to *adversarial examples* for image data (Szegedy et al. 2014) as well as text data (Papernot et al. 2016). For image classification, numerous methods have been proposed with regard to adversarial attack (Goodfellow, Shlens, and Szegedy 2015; Wang et al. 2019) and defense (Goodfellow, Shlens, and Szegedy 2015; Guo et al. 2018). Among which, adversarial training that adopts perturbed examples in the training stage so as to promote the model robustness has become very popular and effective (Athalye, Carlini, and Wagner 2018).

For natural language processing tasks, however, the lexical, grammatical and semantic constraints and the discrete input space make it much harder to craft text adversarial examples. Current attack methods include character-level attack (Liang et al. 2018; Li et al. 2019; Ebrahimi et al. 2018), word-level attack (Papernot et al. 2016; Samanta and Mehta 2017; Gong et al. 2018; Cheng et al. 2018; Kuleshov et al. 2018; Neekhara et al. 2019; Ren et al. 2019; Wang, Jin, and He 2019), and sentence-level attack (Iyyer et al. 2018; Ribeiro, Singh, and Guestrin 2018). For character-level attack, recent works (Pruthi, Dhingra, and Lipton 2019) have shown that a spell checker can easily fix the perturbations. Sentence-level attacks, usually based on paraphrasing, demand longer time in adversary generation. For word-level attack, replacing word based on embedding perturbation or adding/removing word is often vulnerable to the problem of hurting semantic consistency and grammatical correctness. Synonym substitution based attacks could better cope with the above issues and produce adversarial examples that are harder to be detected by humans. Unfortunately, synonym substitution based attacks exhibit lower efficiency compared with existing image attack methods.

As text adversarial attack has attracted increasing interests very recently since 2018, its counterpart, text adversarial defense, is much less studied in the literature. Some research (Jia et al. 2019; Huang et al. 2019) is based on interval bound propagation (IBP), originally proposed for images (Gowal et al. 2019), to ensure certified text defense. Zhou et al. (2019) learn to discriminate perturbations (DISP) and restore the embedding of the original word for defense without altering the training process or the model structure. Wang, Jin, and He (2019) propose a Synonym Encoding Method (SEM), which inserts an encoder before the input layer to defend synonym substitution based attacks.

To our knowledge, adversarial training, one of the most efficacious defense methods for image data (Athalye, Carlini, and Wagner 2018), has not been implemented as an effective defense method against synonym substitution based attacks due to the inefficiency of current adversary generation methods. On one hand, existing synonym substitution based attack methods are usually much less efficient to be incorporated into adversarial training. On the other hand, although gradient-based image attacks often have much higher efficiency, it is challenging to adapt such methods directly in the text embedding space to generate meaningful adversarial examples without changing the original semantics, due to

[*]The first three authors contribute equally.
[†]Corresponding author.

the discreteness of the text input space.

To this end, we propose a gradient-based adversarial attack, called *Fast Gradient Projection Method* (FGPM), for efficient synonym substitution based text adversary generation. Specifically, we approximate the classification confidence change caused by synonym substitution by the product of gradient magnitude and projected distance between the original word and the candidate word in the gradient direction. At each iteration, we substitute a word with its synonym that leads to the highest product value. Compared with existing query-based attack methods, FGPM only needs to calculate the back-propagation once to obtain the gradient so as to find the best synonym for each word. Extensive experiments show that FGPM is about 20 times faster than the current fastest text adversarial attack, and it can achieve similar attack performance and transferability compared with state-of-the-art synonym substitution based adversarial attacks.

With such high efficiency of FGPM, we propose *Adversarial Training with FGPM enhanced by Logit pairing (ATFL)* as an efficient and effective text defense method. Experiments show that ATFL promotes the model robustness against white-box as well as black-box attacks, effectively blocks the transferability of adversarial examples and achieves better generalization on benign data than other defense methods. Besides, we also find some recent proposed variants of adversarial training for images, such as TRADES (Zhang et al. 2019), MMA (Ding et al. 2020) that exhibit great effectiveness for image data, cannot improve the performance of adversarial training for text data, indicating the intrinsic difference between text defense and image defense.

Related Work

This section provides a brief overview on word-level text adversarial attacks and defenses.

Adversarial Attack

Adversarial attacks fall in two settings: (a) *white-box attack* allows full access to the target model, including model outputs, (hyper-)parameters, gradients and architectures, etc. (b) *black-box attack* only allows access to the model outputs.

Methods based on word embedding usually fall in the white-box setting. Papernot et al. (2016) find a word in dictionary such that the sign of the difference between the found word and the original word is closest to the sign of the gradient. However, such word does not necessarily preserve the semantic as well as syntactic correctness and consistency. Gong et al. (2018) further employ the Word Mover's Distance (WMD) in an attempt to preserve semantics. Cheng et al. (2018) also propose an attack based on the embedding space with additional constraints targeting seq2seq models.

In black-box setting, Kuleshov et al. (2018) propose a *Greedy Search Attack (GSA)* that perturbs the input by synonym substitution. Specifically, GSA greedily finds a synonym for replacement that minimizes the classification confidence. Ren et al. (2019) propose a *Probability Weighted Word Saliency (PWWS)* that greedily substitutes each target word with a synonym determined by the combination of classification confidence change and word saliency. Alzantot et al. (2018) also use synonym substitution and propose a population-based algorithm called *Genetic Algorithm (GA)*. Wang, Jin, and He (2019) further propose an *Improved Genetic Algorithm (IGA)* that allows to substitute words in the same position more than once and outperforms GA.

Our work produces efficient gradient based white-box attacks, while guaranteeing the quality of adversarial examples by restricting the perturbation to synonym substitution, which only appears in black-box attacks.

Adversarial Defense

There are a series of works (Miyato, Dai, and Goodfellow 2016; Sato et al. 2018; Barham and Feizi 2019) that perturb the word embeddings and utilize the perturbations for adversarial training as a regularization strategy. These works aim to improve the model performance on the original dataset, but do not intend to defend adversarial attacks. Thus, we do not take such works into consideration.

A stream of recent popular defense methods (Jia et al. 2019; Huang et al. 2019) focuses on verifiable robustness. They use IBP to train models that are provably robust to all possible perturbations within the constraints. Such endeavor, however, is currently time-consuming in the training stage as the authors have noted (Jia et al. 2019) and hard to be scaled to relatively complex models or large datasets. Zhou et al. (2019) train a perturbation discriminator that validates how likely a token in the text is perturbed and an embedding estimator that restores the embedding of the original word to block adversarial attacks.

Alzantot et al. (2018) and Ren et al. (2019) adopt the adversarial examples generated by their attack methods for adversarial training and achieve some robustness improvement. Unfortunately, due to the relatively low efficiency of adversary generation, they are unable to craft plenty of perturbations during the training to ensure significant robustness improvement. To our knowledge, word-level adversarial training has not been practically applied for text classification as an efficient and effective defense method.

Besides, Wang, Jin, and He (2019) propose *Synonym Encoding Method (SEM)* that uses a synonym encoder to map all the synonyms to the same code in the embedding space and force the classification to be smoother. Trained with the encoder, their models obtain significant improvement on the robustness with a little decay on the model generalization.

Different from current defenses, our work focuses on fast adversary generation and easy-to-apply defense method for complex neural networks and large datasets.

Fast Gradient Projection Method

In this section, we formalize the definition of adversarial examples for text classification and describe in detail the proposed adversarial attack method, *Fast Gradient Projection Method (FGPM)*.

Text Adversarial Examples

Let $\mathcal{X}$ denote the input space containing all the possible input texts, $\mathcal{Y} = \{y_1, \cdots, y_m\}$ the output space and $\mathcal{D}$ the dictionary containing all the possible words in the input texts. Let $x = \langle w_1, \cdots, w_i, \cdots, w_n \rangle \in \mathcal{X}$ where $w_i \in \mathcal{D}$ denote

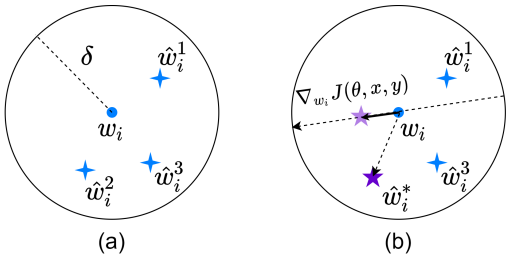

Figure 1: Strategies to pick optimal synonym to substitute word w_i. (a) Pick synonym $\hat{w}_i^*$ that minimizes the classification confidence among all the synonyms $\hat{w}_i^j \in S(w_i, \delta)$. (b) Pick synonym $\hat{w}_i^*$ that maximizes the product of the magnitude of gradient and the projected distance between $\hat{w}_i^*$ and w_i in the gradient direction.

an input sample consisting of n words. A classifier ϕ is expected to learn a mapping $\mathcal{X} \rightarrow \mathcal{Y}$ so that for any sample x, the predicted label $\phi(x)$ equals its true label y with high probability. Let $F(x, y)$ denote the logit output of classifier ϕ on category y. The adversary adds an imperceptible perturbation Δx on x to craft an adversarial example x_{adv} that misleads classifier ϕ:

$$\phi(x_{adv}) \neq \phi(x) = y, \; x_{adv} = x + \Delta x \quad s.t. \, \|\Delta x\|_p \leq \epsilon,$$

where ϵ is a hyper-parameter for the perturbation upper bound, and $\|\cdot\|_p$ is the L_p-norm distance metric, which often denotes the word substitution ratio $R(x, x_{adv})$ as the measure for the perturbation caused by synonym substitution:

$$R(x, x_{adv}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{w_i \neq w_i'}(w_i, w_i').$$

Here $\mathbf{1}_{w_i \neq w_i'}$ is indicator function, $w_i \in x$ and $w_i' \in x_{adv}$.

Generating Adversarial Examples

Mrksic et al. (2016) have shown that counter-fitting can help remove antonyms which are needlessly considered as "similar words" in the original GloVe vector space to improve the capability of indicating semantic similarity. Thus, we post-process the GloVe vectors by counter-fitting and define a synonym set for each word $w_i \in x$ in the embedding space as follows:

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

where δ is a hyper-parameter that constrains the maximum Euclidean distance for synonyms in the embedding space and we set $\delta = 0.5$ as in Wang, Jin, and He (2019).

Once we have the synonym set $S(w_i, \delta)$ for each word w_i, the next steps are for the optimal synonym selection and substitution order determination.

Word Substitution. As shown in Figure 1 (a), for each word w_i, we expect to pick a word $\hat{w}_i^* \in S(w_i, \delta)$ that earns the most benefit to the overall substitution process of adversary generation, which we call optimal synonym. Due to the high complexity of finding optimal synonym, previous works (Kuleshov et al. 2018; Wang, Jin, and He 2019)

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, δ) by Eq. (1) for w_i ∈ x_adv^0
for k = 1 → N do
    Construct candidate set C_s = {ŵ_1^*, ···, ŵ_i^*, ···, ŵ_n^*} by Eq. (2)
    Calculate optimal word ŵ_* by Eq. (3)
    Substitute w_* ∈ x_adv^{k-1} with ŵ_* to obtain x_adv^k
    if φ(x_adv^k) ≠ y and R(x_adv^k, x) < ε then
        return x_adv^k                                   ▷ Succeed
    end if
end for
return None                                              ▷ Failed
```

greedily pick a synonym $\hat{w}_i^* \in S(w_i, \delta)$ that minimizes the classification confidence:

$$\hat{w}_i^* = \underset{\hat{w}_i^j \in S(w_i, \delta)}{\arg\max} \; (F(x, y) - F(\hat{x}_i^j, y)),$$

where $\hat{x}_i^j = \langle w_1, \cdots, w_{i-1}, \hat{w}_i^j, w_{i+1}, \cdots, w_n \rangle$. However, the selection process is time consuming as picking such a $\hat{w}_i^*$ needs $|S(w_i, \delta)|$ queries on the model. To reduce the calculation complexity, based on the local linearity of deep models, we use the product of gradient magnitude and projected distance between the original word and its synonym candidate in the gradient direction in the word embedding space to estimate the amount of change for the classification confidence. Specifically, as illustrated in Figure 1 (b), we first calculate the gradient $\nabla_{w_i} J(\theta, x, y)$ for each word w_i where $J(\theta, x, y)$ is the loss function used for training. Then, we estimate the change by calculating $(\hat{w}_i^j - w_i) \cdot \nabla_{w_i} J(\theta, x, y)$. To determine the optimal synonym $\hat{w}_i^*$, we choose a synonym with the maximum product value:

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

Substitution Order. For each word w_i in text $x = \langle w_1, \cdots, w_i, \cdots, w_n \rangle$, we use the above word substitution strategy to choose its optimal substitution synonym and obtain a candidate set $\mathcal{C}_s = \{\hat{w}_1^*, \cdots, \hat{w}_i^*, \cdots, \hat{w}_n^*\}$. Then, we need to determine which word in x should be substituted. Similar to the word substitution strategy, we pick a word $\hat{w}_i^* \in \mathcal{C}_s$, that has the biggest product of the gradient magnitude and the perturbation value projected in the gradient direction, to substitute $w_i \in x$:

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

In summary, to generate an adversarial example, we adopt the above word replacement and substitution order strategies

for synonym substitution iteratively till the classifier makes a wrong prediction. The overall FGPM algorithm is shown in Algorithm 1.

To avoid the semantic drift caused by multiple substitutions at the same position of the text, we construct a candidate synonym set for the original sentence ahead of synonym substitution process and constrain all the substitutions with word $w_i \in x$ to the set, as shown at line 2 of Algorithm 1. We also set the upper bound for word substitution ratio $\epsilon = 0.25$ in our experiments. Note that at each iteration, previous query-based adversarial attacks need $\sum_{i=1}^{n}|S(w_i, \delta)|$ times of model queries (Kuleshov et al. 2018; Ren et al. 2019), while FGPM just calculates the gradient by back-propagation once, leading to much higher efficiency.

Adversarial Training with FGPM

For image classification, Goodfellow, Shlens, and Szegedy (2015) first propose adversarial training using the following objective function:

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

In recent years, numerous variants of adversarial training (Kannan, Kurakin, and Goodfellow 2018; Zhang et al. 2019; Song et al. 2019; Ding et al. 2020) have been proposed to further enhance the robustness of models.

For text classification, previous works (Alzantot et al. 2018; Ren et al. 2019) have shown that incorporating their attack methods into standard adversarial training can improve the model robustness. Nevertheless, the improvement is limited. We argue that adversarial training requires plenty of adversarial examples generated based on instant model parameters in the training stage for better robustness enhancement. Due to the inefficiency of text adversary generation, existing text attack methods based on synonym substitution could not provide sufficient adversaries for adversarial training. With the high efficiency of FGPM, we propose a new text defense method called *Adversarial Training with FGPM enhanced by Logit pairing (ATFL)* to effectively improve the model robustness for text classification.

Specifically, we modify the objective function as follows:

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

where x_{adv} is the adversarial example of each x generated by FGPM based on the instant model parameters θ during training. In all our experiments, we set $\alpha = 0.5$ and $\lambda = 0.5$, and provide ablation study for α and λ in Appendix. As in Kannan, Kurakin, and Goodfellow (2018), we train the model on adversarial examples and treat the logit similarity of benign examples and their adversarial counterparts as an regularizer to improve the model robustness rather than just adding a portion of adversarial examples of the already trained model into the training set and retrain the model.

Experimental Results

We evaluate FGPM with four attack baselines, and ATFL with two defense baselines, IBP and SEM, on three popular benchmark datasets involving CNN and RNN models. Code is available at https://github.com/JHL-HUST/FGPM.

Experimental Setup

We first introduce the experimental setup, including baselines, datasets and models used in experiments.

Baselines. For fair comparison, we restrict the perturbations of Papernot et al. (2016) within synonyms and denote this baseline as *Papernot'*. To evaluate the attack effectiveness of FGPM, we compare it with four adversarial attacks, *Papernot'*, GSA (Kuleshov et al. 2018), PWWS (Ren et al. 2019), and IGA (Wang, Jin, and He 2019). Furthermore, to validate the defense performance of our ATFL, we take two competitive text defense methods, SEM (Wang, Jin, and He 2019) and IBP (Jia et al. 2019), against the above word-level attacks. Due to the low efficiency of attack baselines, we randomly sample 200 examples on each dataset, and generate adversarial examples on various models.

Datasets. We compare the proposed methods with baselines on three widely used benchmark datasets including *AG's News*, *DBPedia ontology* and *Yahoo! Answers* (Zhang, Zhao, and LeCun 2015). *AG's News* consists of news articles pertaining four classes: World, Sports, Business and Sci/Tech. Each class includes 30,000 training examples and 1,900 testing examples. *DBPedia ontology* is constructed by picking 14 non-overlapping classes from *DBPedia 2014*, which is a crowd-sourced community effort to extract structured information from Wikipedia. For each of the 14 ontology classes, there are 40,000 training samples and 5,000 testing samples. *Yahoo! Answers* is a topic classification dataset with 10 classes, and each class contains 140,000 training samples and 5,000 testing samples.

Models. We adopt several deep learning models that can achieve state-of-the-art performance on text classification tasks, including Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The embedding dimension of all models is 300 (Mikolov et al. 2013). Specifically, we replicate a CNN model from Kim (2014), which consists of three convolutional layers with filter size of 5, 4, and 3 respectively, a dropout layer and a final fully connected layer. We also use a Long Short-Term Memory (LSTM) model which replaces the three convolutional layers of the CNN with three LSTM layers, each with 128 cells (Liu, Qiu, and Huang 2016). Lastly, we implement a Bi-directional Long Short-Term Memory (Bi-LSTM) model that replaces the three LSTM layers of the LSTM with a bi-directional LSTM layer having 128 forward direction cells and 128 backward direction cells.

Evaluation on Attack Effectiveness

To evaluate the attack effectiveness, we compare FGPM with the baseline attacks in two aspects, namely model classification accuracy under attacks and transferability.

Classification Accuracy under Attacks. In Table 1, we provide the classification accuracy under FGPM and the competitive baseline attacks on three standard datasets. The more effective the attack method is, the more the classification accuracy of the target model drops. We observe that IGA, adopting the genetic algorithm, can always achieve the best attack performance among all attacks. Compared with other attacks, FGPM could either achieve the best attack performance or on par with the best one. Especially, *Papernot'*,

	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 different models under various competitive adversarial attacks. The first two rows of *No Attack*† and *No Attack* show the model accuracy on the entire original test set and the sampled examples respectively. The lowest classification accuracy among the attacks is highlighted in bold to indicate the best attack effectiveness. The second lowest classification accuracy is highlighted in underline.

	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 adversarial examples generated on other models on *AG's News* for the transferability evaluation. * indicates that the adversarial examples are generated based on this model.

the only gradient-based attack among the baselines, is inferior to FGPM, indicating that the proposed gradient projection technique significantly improves the effectiveness of white-box word-level attacks. Besides, we also display some adversarial examples generated by FGPM in Appendix.

Transferability. The transferability of adversarial attack refers to the ability to reduce the classification accuracy of different models with adversarial examples generated on a specific model (Goodfellow, Shlens, and Szegedy 2015), which is another serious threat in real-world applications. To illustrate the transferability of FGPM, we generate adversarial examples on each model by different attack methods and evaluate the classification accuracy of other models on these adversarial examples. Here, we evaluate the transferability of different attacks on *AG's News*. As depicted in Table 2, the adversarial examples crafted by FGPM is on par with the best transferability performance among the baselines.

Evaluation on Attack Efficiency

The attack efficiency is important for evaluating attack methods, especially if we would like to incorporate the attacks into adversarial training as a defense method. Adversarial training needs highly efficient adversary generation so as to effectively promote the model robustness. Thus, we evaluate the total time (in seconds) of generating 200 adversarial examples on the three datasets by various attacks. As shown in Table 3, the average time of generating 200 adversarial examples by FGPM is nearly 20 times faster than GSA, the second fastest synonym substitution based attack but with weaker attack performance and lower transferability than FGPM. Moreover, FGPM is on average 970 times faster than IGA, which produces the maximum degradation of the classification accuracy among the baselines. Though *Papernot'* crafts adversarial examples based on gradient, which makes each iteration faster, it needs much more iterations to obtain adversarial examples due to low attack effectiveness. On average, FGPM is about 78 times faster than *Papernot'*.

Evaluation on Adversarial Training

From the above analysis, we see that compared with the competitive attack baselines, FGPM can achieve much higher efficiency with good attack performance and transferability. Such performance enables us to implement effective adversarial training and scale to large neural networks and datasets. In this subsection, we evaluate the defence performance of ATFL and conduct comparison with SEM and IBP against adversarial examples generated by the above attacks. Here we focus on two factors, defense against adversarial attacks and defense against transferability.

Defense against Adversarial Attacks. We use the above attacks on models trained by various defense methods to evaluate the defense performance. The results are shown in Table 4. For normal training (NT), the classification accuracy on all datasets drops dramatically under different adversarial attacks. In contrast, both SEM and ATFL can promote the model robustness stably and effectively among all models and datasets. IBP, originally proposed for CNN to defend the adversarial attacks in image domain, can improve the robustness of CNN on three datasets but with much higher computation cost. More importantly, with many restrictions added on the architectures, the model hardly converges when trained on LSTM and Bi-LSTM, resulting in both weakened generalization and adversarial robustness instead. Compared with SEM, moreover, ATFL can obtain higher classification accuracy on benign data, and is very competitive under almost all adversarial attacks.

	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: Comparison on the total running time (in seconds) for generating 200 adversarial instances.

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 competitive defense methods under various adversarial attacks on the same set of 200 randomly selected samples for the three standard datasets.

Defense against Transferability. As transferability poses a serious concern in real-world applications, a good defense method should not only defend the adversarial attack but also resist the transferability of adversarial examples. To evaluate the ability of blocking transferability, we evaluate each model's classification accuracy on adversarial examples generated by different attack methods under normal training on *AG's News*. As shown in Table 5, ATFL is much more successful in blocking the transferability of adversarial examples than the defense baselines on CNN and LSTM and achieve similar accuracy to SEM on Bi-LSTM.

In summary, ATFL can significantly promote the model robustness, block the transferability of adversarial examples successfully and achieve better generalization on benign data compared with other defenses. Moreover, when applied to complex models and large datasets, ATFL maintains stable and effective performance.

Evaluation on Adversarial Training Variants

Many variants of adversarial training, such as CLP and ALP (Kannan, Kurakin, and Goodfellow 2018), TRADES (Zhang et al. 2019), MMA (Ding et al. 2020), MART (Wang et al. 2020), have tried to adopt different regularizations to improve the effectiveness of adversarial training for image data. The loss functions for these variants are depicted in Appendix. Here we try to answer the following question: can these variants also bring improvement for texts?

To validate the effectiveness of these variants, we run the above methods on *AG's News* with three models. As shown in Table 6, standard adversarial training can improve both generalization and robustness of the models. Among the variants, however, only ALP can further improve the performance of adversarial training. Some recent variants (e.g. TRADES, CLP) that work very well for images significantly degrade the performance of standard adversarial training for texts, indicating that we need more specialized adversarial training methods for texts.

Conclusion

In this work, we propose an efficient gradient based synonym substitution adversarial attack method, called *Fast Gradient Projection Method* (FGPM). Empirical evaluations

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
Papernot'	83.5	87.5	87.5	**88.0**	61.5*	**91.0**	82.0	**91.0**	78.5	**91.0**	86.5	89.5
GSA	84.5	87.0	87.5	**88.5**	35.0*	90.5	83.5	**91.0**	73.0	**91.0**	86.5	89.5
PWWS	83.0	87.0	87.5	**89.0**	30.0*	**90.5**	85.0	**90.5**	67.5	**90.5**	86.5	90.0
IGA	84.0	87.0	87.5	**88.5**	26.5*	90.5	83.5	**91.5**	71.5	**91.0**	87.0	90.0
FGPM	81.0	87.5	87.5	**89.0**	31.0*	90.5	83.5	**91.5**	73.5	**91.0**	87.0	89.5
Papernot'	79.5	88.0	87.0	**88.5**	74.5	**91.0**	82.5	**91.0**	65.0*	**91.0**	86.5	89.0
GSA	81.5	87.0	87.5	**88.5**	72.5	90.5	84.0	**91.0**	40.0*	**91.0**	87.5	90.0
PWWS	80.0	86.5	87.0	**89.0**	67.5	90.5	83.5	**91.5**	29.0*	**90.5**	87.0	90.0
IGA	79.0	87.0	87.0	**88.5**	71.0	90.5	83.5	**91.0**	25.5*	**91.0**	86.5	89.5
FGPM	77.5	87.5	87.5	**89.0**	67.5	90.5	83.5	**91.0**	32.0*	**91.0**	87.0	89.5

Table 5: The classification accuracy (%) of various models under competitive defenses for adversarial examples generated on other models on *AG's News* for evaluating the defense performance against transferability. * indicates that the adversarial examples are generated based on this model.

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**
LSTM	No Attack†	**92.6**	**92.6**	91.9	91.3	90.8	92.1	92.0
	No Attack	90.5	**92.0**	90.5	89.0	90.0	91.0	91.5
	Papernot'	61.5	88.0	66.0	86.0	86.0	69.0	**90.0**
	GSA	35.0	83.0	37.5	78.0	79.0	40.5	**88.0**
	PWWS	30.0	84.0	32.0	78.0	79.5	46.5	**88.0**
	IGA	26.5	83.0	24.0	77.5	79.5	34.0	**88.0**
	FGPM	31.0	83.0	32.5	81.5	80.5	41.0	**88.0**
Bi-LSTM	No Attack†	92.5	**92.8**	92.4	91.4	92.3	92.4	92.1
	No Attack	88.5	89.5	**90.5**	88.5	90.0	**90.5**	89.5
	Papernot'	65.0	**89.5**	65.5	85.5	86.0	66.0	89.0
	GSA	40.0	86.0	35.5	81.0	80.5	38.5	**87.5**
	PWWS	29.0	86.5	30.0	80.0	80.5	52.0	**87.5**
	IGA	25.5	86.0	29.0	78.5	80.0	34.5	**87.5**
	FGPM	32.0	86.5	32.0	82.0	80.5	46.0	**87.5**

Table 6: The classification accuracy (%) of different classification models adversarially trained with different regularization under various adversarial attacks on the same set of 200 randomly selected samples for the *AG's News* dataset.

on three widely used benchmark datasets demonstrate that FGPM is about 20 times faster than the current fastest synonym substitution based adversarial attack method, and FGPM can achieve similar attack performance and transferability. With such high efficiency, we introduce an effective defense method called *Adversarial Training with FGPM enhanced by Logit pairing* (ATFL) for text classification. Extensive experiments demonstrate that ATFL can significantly promote the model robustness, block the transferability of adversarial examples effectively, and achieve better generalization on benign data than text defense baselines. Besides, 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.

Our work offers a way to adopt gradient for adversarial attack in discrete space, making it possible to adapt successful gradient based image attacks for text adversarial attacks. Besides, considering the prosperity of adversarial training for image data and high efficiency of gradient based methods, we hope our work could inspire more research of adversarial training in text domain.

Acknowledgements

This work is supported by National Natural Science Foundation (62076105) and Microsft Research Asia Collaborative Research Fund (99245180). We thank Kai-Wei Chang for helpful suggestions on our work.

References

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.; Srivastava, M. B.; and Chang, K. 2018. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2890–2896.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 274–283.

Barham, S.; and Feizi, S. 2019. Interpretable adversarial training for text. *arXiv preprint arXiv:1905.12864* .

Cheng, M.; Yi, J.; Chen, P.-Y.; Zhang, H.; and Hsieh, C.-J. 2018. Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples. *arXiv preprint arXiv:1803.01128* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)*, 4171–4186.

Ding, G. W.; Sharma, Y.; Lui, K. Y. C.; and Huang, R. 2020. MMA Training: Direct Input Space Margin Maximization through Adversarial Training. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Ebrahimi, J.; Rao, A.; Lowd, D.; and Dou, D. 2018. HotFlip: White-Box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 31–36.

Gong, Z.; Wang, W.; Li, B.; Song, D.; and Ku, W.-S. 2018. Adversarial Texts with Gradient Methods. *arXiv preprint arXiv:1801.07175* .

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T. A.; and Kohli, P. 2019. Scalable Verified Training for Provably Robust Image Classification. *International Conference on Computer Vision (ICCV)* .

Guo, C.; Rana, M.; Cisse, M.; and van der Maaten, L. 2018. Countering adversarial images using input transformations. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Huang, P.-S.; Stanforth, R.; Welbl, J.; Dyer, C.; Yogatama, D.; Gowal, S.; Dvijotham, K.; and Kohli, P. 2019. Achieving Verified Robustness to Symbol Substitutions via Interval Bound Propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4083–4093.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long Papers)*, 1875–1885.

Jia, R.; Raghunathan, A.; Göksel, K.; and Liang, P. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4120–4133.

Kannan, H.; Kurakin, A.; and Goodfellow, I. J. 2018. Adversarial Logit Pairing. *arXiv Preprint arXiv:1803.06373* .

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 26th Annual Conference on Neural Information Processing System (NIPS)*, 1097–1105.

Kuleshov, V.; Thakoor, S.; Lau, T.; and Ermon, S. 2018. Adversarial examples for natural language classification problems. *OpenReview submission OpenReview:r1QZ3zbAZ* .

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2019. TextBugger: Generating adversarial text against real-world applications. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*.

Liang, B.; Li, H.; Su, M.; Bian, P.; Li, X.; and Shi, W. 2018. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 4208–4215.

Liu, P.; Qiu, X.; and Huang, X. 2016. Recurrent Neural Network for Text Classification with Multi-Task Learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2873–2879.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR)*.

Miyato, T.; Dai, A. M.; and Goodfellow, I. J. 2016. Adversarial training methods for semi-supervised text classification. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Mrksic, N.; Séaghdha, D. Ó.; Thomson, B.; Gasic, M.; Rojas-Barahona, L. M.; Su, P.; Vandyke, D.; Wen, T.; and

Young, S. J. 2016. Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 142–148.

Neekhara, P.; Hussain, S.; Dubnov, S.; and Koushanfar, F. 2019. Adversarial Reprogramming of Text Classification Neural Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5216–5225.

Papernot, N.; McDaniel, P. D.; Swami, A.; and Harang, R. E. 2016. Crafting adversarial input sequences for recurrent neural networks. In *Proceedings of IEEE Military Communications Conference (MILCOM)*, 49–54.

Pruthi, D.; Dhingra, B.; and Lipton, Z. C. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 5582–5591.

Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1085–1097.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 856–865.

Samanta, S.; and Mehta, S. 2017. Towards Crafting Text Adversarial Samples. *arXiv preprint arXiv:1707.02812* .

Sato, M.; Suzuki, J.; Shindo, H.; and Matsumoto, Y. 2018. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 4323–4330.

Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2019. Improving the generalization of adversarial training with domain adaptation. *Proceedings of the 7th International Conference on Learning Representations (ICLR)* .

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Wang, X.; He, K.; Song, C.; Wang, L.; and Hopcroft, J. E. 2019. AT-GAN: An Adversarial Generator Model for Non-constrained Adversarial Examples. *arXiv preprint arXiv:1904.07793* .

Wang, X.; Jin, H.; and He, K. 2019. Natural language adversarial attacks and defenses in word level. *arXiv Preprint arXiv:1909.06723* .

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, 7472–7482.

Zhang, X.; Zhao, J. J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 649–657.

Zhou, Y.; Jiang, J.-Y.; Chang, K.-W.; and Wang, W. 2019. Learning to discriminate perturbations for blocking adversarial attacks in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4906–4915.