Fraud's Bargain Attacks to Textual Classifiers via Metropolis-Hasting Sampling (Student Abstract)

Mingze Ni[1], **Zhensu Sun**[2], **and Wei Liu**[1]

[1] University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia
[2] ShanghaiTech University, 393 Middle Huaxia Road, Shanghai, China
Mingze.Ni@student.uts.edu.au, sunzhs@shanghaitech.edu.cn, Wei.Liu@uts.edu.au

Abstract

Recent studies on adversarial examples expose vulnerabilities of natural language processing (NLP) models. Existing techniques for generating adversarial examples are typically driven by deterministic heuristic rules that are agnostic to the optimal adversarial examples, a strategy that often results in attack failures. To this end, this research proposes Fraud's Bargain Attack (FBA), which utilizes a novel randomization mechanism to enlarge the searching space and enables high-quality adversarial examples to be generated with high probabilities. FBA applies the Metropolis-Hasting algorithm to enhance the selection of adversarial examples from all candidates proposed by a customized Word Manipulation Process (WMP). WMP perturbs one word at a time via insertion, removal, or substitution in a contextual-aware manner. Extensive experiments demonstrate that FBA outperforms the baselines in terms of attack success rate and imperceptibility.

Introduction

AI-based models on text data have been broadly applied to many real-world applications, but they are surprisingly fragile to adversarial examples crafted by adding maliciously crafted typos, words, and letters to the input (Yang et al. 2021). Generally, attackers perform adversarial attacks using a two-step process: (1) it heuristically searches preset N essential words to probe the target model; (2) it replaces crucial words with substitutions from lexical databases or masked language models (MLM). Although existing methods can deceive their victim models, they have the following drawbacks: (1) No prior work can use an optimized combination of all word substitution, insertion, and removal strategies in the formation of the attacks. This shortage of study leads to limited search space for adversarial examples and restricts the performance of the attacks. (2) Most algorithms calculate a word importance rank (WIR), and such a rank can be ineffective when attacking more than one word. (3) Launching attacks to a preset static number of perturbed words (NPW) is not adaptive to different target texts.

To address the above problems, in this research, we propose an attacking algorithm, Frauds' Bargain attack (FBA), which utilizes the Metropolis-Hasting (MH) algorithm to improve the quality of adversarial candidates from our proposed stochastic process, Word Manipulation process (WMP). We name our method Frauds' Bargain because WMP makes a malicious "deal" (adversarial candidate) while MH sampler works as an experienced fraud to calculate an accept probability by considering the "deal's" quality measured by the adversarial distribution. Compared with existing attacks, FBA has three outstanding advantages: (1) a much larger searching domain, (2) an adaptive setting of NPW for different target texts, and (3) keeping better semantics by considering semantic preservation for both words and sentences. Specifically, FBA can generate adversarial examples through insertion, substitution, and deletion with MH combinatory optimization, which could enlarge the searching domain. Unlike deterministic WIR, FBA performs stochastic attacks that can probabilistically adapt to a more effective NPW. In addition, we consider synonyms for word perturbations and regulate the FBA with a sentence-level semantic similarity for semantic preservation.

The Proposed Attacking Strategy

Notation Let x and y denote the input text and its corresponding class, respectively. The victim classifier F learns to map the text space to the class space through a categorical distribution, $F(\cdot) : \mathcal{X} \rightarrow (0,1)^K$, where $\mathcal{X}$ represents text space, and K is the number of classes. Given the input text $x = [w_1, \ldots, w_i, \ldots, w_n]$ with n words, we denote an adversarial candidate of x as x', and denote the final chosen adversarial example as x^*.

Word Manipulation Process (WMP) The WMP is defined as an aperiodic Markovian process which means such a process owns discrete time stamps and states. Its aperiodicity guarantees an enlarged searching domain. To perform the WMP, we sequentially propose three dependent manipulations, including actions e, positions $l|e$, and candidate $o|l, e$, for each iterative time. With these three mutually dependent manipulations, we intuitively divide WMP into three steps. The first step is to sample an action e from the set {insert(0), substitute(1), remove(2)} with preset probabilities (P_{add}, P_{sub}, P_{rem}) according to attacks' preference distribution $p(e)$. After determining the action, we determine the position l in the sentence to conduct the chosen manipulation e by drawing l from a customized categorical distribution $p(l|e)$ based on the words' importance. To score the words' importance, this step removes w_i from the input and queries the victim model F to observe the change in

the classification score of the target class. For the last step, if word insertion or substitution is chosen, WMP will provide an insertion or substitution candidate. To find the word candidates, we construct function $p(o|l, e)$ by an MLM and synonyms of the original words (calculated by nearest neighbors using the L2-norm of word embeddings) for parsing fluency and semantic preservation, respectively. By applying the Bayes rule, we can derive the WMP's distribution from iteration t to $t + 1$ as the following equation:

$$\text{WMP}(x_{t+1}|x_t) = p(e|x_t)p(l|e, x_t)p(o|e, l, x_t)$$

Fraud's Bargain Attack (FBA) The FBA utilizes the Metropolis-Hasting (MH) algorithm to enhance the WMP via selecting adversarial candidates evaluated by a customized adversarial distribution. Therefore, we construct the adversarial target distribution as:

$$\pi(x'|x, \lambda) = \frac{R + \lambda \text{Sem}(x', x)}{C},$$

where $\pi(x') : \mathcal{X} \to (0, 1)$ measures the classifier's depravation $1 - F(x)$ with a penalty on semantic similarity, and $\text{Sem}(\cdot) : \mathcal{X}^2 \to [0, 1]$. With such an adversarial distribution, we adaptively apply the MH algorithm to obtain the following accept probability:

$$\alpha(x_{t+1}|x_t) = \min\left(1, \frac{\pi\left(x_{t+1}\right)}{\pi(x_t)} \frac{\text{WMP}\left(x_t \mid x_{t+1}\right)}{\text{WMP}\left(x_{t+1} \mid x_t,\right)}\right)$$

After calculating $\alpha(x_{t+1}|x_t)$, we sample $u \sim Unif(0, 1)$ and compare it with accept probability to decide if we accept x_{t+1} as the new state. In the optimization process, FBA generates a set of adversarial candidates, and we choose the one with the lowest amount of modifications among the successful adversarial candidates.

Experiments

We evaluate the proposed method with two well-performed textual classifiers: BERT Classifier (BERT.C) and TextCNN, on three benchmark datasets: AG News, Emotion, and SST2. We use the successful attack rate (SAR) to measure the attacking performance, while we introduce the ROUGE and Universal Sentence Encoder(USE) to measure the imperceptibility in terms of n-gram and semantic similarities, respectively. For all these three metrics, the higher the value, the better the attack. We compare our work with 3 state-of-art methods: PWWS (Ren et al. 2019), Fast Genetic Attack (Jia et al. 2019), and PSO (Zang et al. 2020).

As shown in Table 2, our method achieved the best performance across all target models and datasets in terms of both attacking successes and imperceptibility.

Conclusion and Future Work

We proposed a novel Fraud's Bargain Attack (FBA) algorithm to generate adversarial examples. The FBA exploits the WMP to generate the adversarial candidates employs the MH sampler to improve their quality. With FBA, we not only make successful attacks but also reserve the semantics. Future research directions include designing defense methods based on FBA.

Attacks	Adversarial examples
PWWS (**Successful attack**. True class score = 21.11%)	i spent wandering around still kinda dazed and not ~~feeling~~ *palpate* particularly sociable but because id been in hiding for a couple for days and it was getting to be a little unhealthy i made myself go down to the cross and hang out with folks
FBA (**Successful attack**. True class score = **0.63%**)	i spent wandering around still kinda dazed and not ~~feeling~~ *sensing* particularly sociable but because id been in hiding for a couple for days and it was getting to be a little unhealthy i made myself go down to the cross and hang out with folks

Table 1: Adversarial example of Emotion for BERT-C.

Model	Attack	SAR%	ROUGE	Sem%
BERT.C (AG News)	PWWS	76.75%	83.56%	73.32%
	FGA	25.17%	77.21%	75.34%
	FBA	**81.90%**	**84.53%**	**80.01%**
TextCNN (AG News)	PWWS	85.31%	83.66%	81.11%
	FBA	**93.12%**	**87.11%**	**82.24%**
BERT.C (Emotion)	PWWS	91.75%	61.41%	90.12%
	FGA	78.21%	59.11%	89.40%
	PSO	94.76%	62.56 %	92.10%
	FBA	**99.15%**	**64.10%**	**92.46%**
TextCNN (Emotion)	PWWS	98.20%	69.94%	89.20%
	FBA	**100%**	**73.04%**	**90.46%**
BERT.C (SST2)	PWWS	93.90%	63.41%	84.22%
	FBA	**99.31%**	**75.40%**	**88.20%**
TextCNN (SST2)	PWWS	98.13%	63.93%	82.10%
	PSO	92.20%	70.01%	81.62%
	FBA	**100%**	**73.40%**	**87.09%**

Table 2: Results comparisons across different models on different datasets. SAR is successful attack rate, ROUGE measures the preservation on the original texts, and Sem represent the semantic similarity.

References

Jia, R.; Raghunathan, A.; Göksel, K.; and Liang, P. 2019. Certified Robustness to Adversarial Word Substitutions. In *EMNLP-IJCNLP*, 4129–4142.

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 ACL*, 1085–1097.

Yang, X.; Liu, W.; Bailey, J.; Tao, D.; and Liu, W. 2021. Bigram and Unigram Based Text Attack via Adaptive Monotonic Heuristic Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 706–714.

Zang, Y.; Qi, F.; Yang, C.; Liu, Z.; Zhang, M.; Liu, Q.; and Sun, M. 2020. Word-level Textual Adversarial Attacking as Combinatorial Optimization. In *Proceedings of the 58th ACL*, 6066–6080.