A Context Aware Approach for Generating Natural Language Attacks

Rishabh Maheshwary, Saket Maheshwary, Vikram Pudi
Data Sciences and Analytics Center, Kohli Center on Intelligent Systems
International Institute of Information Technology, Hyderabad, India
{rishabh.maheshwary, saket.maheshwary}@research.iiit.ac.in, vikram@iiit.ac.in

Abstract
We study an important task of attacking natural language processing models in a black box setting. We propose an attack strategy that crafts semantically similar adversarial examples on text classification and entailment tasks. For each word to be replaced, our proposed attack generates candidate words using the influence of both the original word and its surrounding context. The generated candidates (1) fit well within the sentence thus retaining the overall semantics of the sentence, (2) have the same meaning as that of the original word and (3) are grammatically correct and fluent. We use BERT to generate candidates for each word to be replaced in the input. BERT is a masked language model trained on masked language modeling (MLM) — predicts the [MASK] word using both left and right context and next sentence prediction (NSP) — predicts the next sentence given previous sentences separated by the [SEP] token. For each word to be replaced, we substitute that word with a [MASK] token to get a masked input. Then, we leverage the NSP and feed both the original input as well as the masked input separately by the [SEP] to BERT.

Proposed Approach
Given a target model $F : \mathcal{X} \rightarrow \mathcal{Y}$, that classifies an input text $\mathcal{X}$ to a set of class labels $\mathcal{Y}$. Our goal is to generate an adversarial text sequence $\mathcal{X}_{ADV}$ that is misclassified by $F$ i.e. $F(\mathcal{X}) \neq F(\mathcal{X}_{ADV})$ and is semantically similar to $\mathcal{X}$. We solve this problem using the following two steps:

**Word Ranking:** Given an input text $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, this step assigns high scores to words which significantly impact the final prediction of $F$. To score a word $x_i$, this step removes $x_i$ from the input and queries $F$ to observe the change in the classification score of the target class. This process is repeated for all of the words in $\mathcal{X}$. All the words are then sorted in descending order based on their score.

**Word Substitution:** Given the word $x_i$ and $M$ the BERT masked language model, this step finds a word replacement for $x_i$ using the information of both the original word $x_i$ and its surrounding context. It consists of the following steps:

1. **Candidate Generation:** It replaces $x_i$ with a [MASK] token to get a masked input $\mathcal{X}_i = \{x_1, \ldots, x_{i-1}, \text{[MASK]}, x_{i+1}, \ldots, x_n\}$. But, simply masking a word and using $M$ to predict the masked word might generate candidate words that are not synonyms of $x_i$. For example, consider the sentence "I love this movie", if the word "love" is masked then $M$ might predict words like

Code: [https://github.com/RishabhMaheshwary/contextattack](https://github.com/RishabhMaheshwary/contextattack)

---

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
This film is really robust? natural language attack on text classification and entailment. As our attack queries $M$ for each word to be substituted, the attack generation process slows down. We leave these improvements for future work.

References

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating natural language adversarial examples. arXiv preprint arXiv:1804.07998.

Jin, D.; Jin, Z.; Zhou, J. T.; and Szolovits, P. 2019. Is bert really robust? natural language attack on text classification and entailment. arXiv preprint arXiv:1907.11932.

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, 1085–1097.

| Model       | Orig% | Attack | Acc% | Pert% | I%  |
|-------------|-------|--------|------|-------|-----|
| BERT (IMDB) | 90.9  | TF     | 13.6 | 6.1   | 0.5 |
|             |       | Ours   | 9.5  | 4.2   | 0.38|
| LSTM (IMDB) | 89.8  | GA    | 3.0  | 14.7  | 0.78|
|             |       | PWWS  | 2.0  | 3.38  | 0.46|
|             |       | TF    | 0.3  | 5.1   | 0.53|
|             |       | Ours  | 0.3  | 3.2   | 0.32|
| BERT (MR)   | 86.5  | TF    | 11.5 | 16.7  | 1.26|
|             |       | Ours  | 10.7 | 16.3  | 0.7 |
| LSTM (MR)   | 80.0  | TF    | 3.1  | 14.9  | 1.04|
|             |       | PWWS  | 3.7  | 14.38 | 0.86|
|             |       | Ours  | 2.1  | 14.0  | 0.7 |
| BERT (SNLI) | 89.1  | TF    | 4.0  | 18.5  | 9.7 |
|             |       | PWWS  | 3.6  | 18.9  | 1.9 |
|             |       | Ours  | 4.0  | 18.5  | 9.7 |
| LSTM (SNLI)| 84.0  | GA    | 3.0  | 23.3  | 9.9 |
|             |       | TF    | 3.5  | 18.0  | 7.7 |
|             |       | Ours  | 3.0  | 17.0  | 2.3 |

Table 1: Result comparison. Orig% is the original accuracy, Acc% is the after attack accuracy, Pert% is the average perturbation rate, I% is the grammatical error increase rate.

| Type       | Examples                                                |
|------------|---------------------------------------------------------|
| Orig.      | Vile and tacky are best adjectives to describe it.     |
| TF         | Vile and tacky are best **qualifier** to describe it.  |
| Ours       | Vile and tacky are best **words** to describe it.      |
| Orig.      | This film is a portrait of grace in this world.         |
| TF         | This film is a **spitting** of grace in this universe.  |
| Ours       | This film is a **sketch** of beauty in this world.      |

Table 2: Demonstrates adversarial examples generated on BERT by TF and our attack. Substituted words are in bold.