A Context Aware Approach For Generating Natural Language Attacks

Rishabh Maheshwary, Saket Maheshwary and Vikram Pudi
IIIT Hyderabad

AAAI Association for the Advancement of Artificial Intelligence
Vancouver, Canada | Feb 2-9, 2021

Introduction

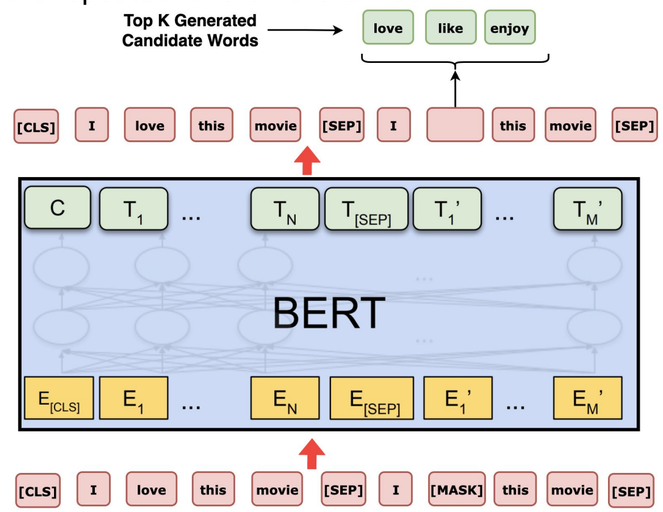
- Deep neural networks are vulnerable to **adversarial attacks**. Existing black box natural language attacks generate adversarial examples by first finding important words and than replacing those with their synonyms.
- Such attacks finds synonyms either using a lexical database such as **WordNet** or from the nearest neighbour in the **counter-fitted embedding space**.
- The major drawbacks of such attacks is that it takes only the word level similarity to find synonyms but **does not consider the overall context surrounding the word to be replaced** thus generating out of context synonyms which degrades the overall semantics.
- Also, It **yields unnatural and complex replacements** which results in non-fluent adversarial examples.
- In this paper, we propose a black box attack that **generates candidate words using the influence of both the original word (to be replaced) as well as its surrounding context.**

- It jointly leverages **masked language modelling** and **next sentence prediction** for context understanding.

Proposed Approach

Our proposed attack consists of two steps:

- **Word Ranking:** This step removes each word from the input and observe the change in the classification score of the target class. This process is repeated for all of the words in the input. All the words are then sorted in descending order based on their score.
- **Word Substitution:** To replace each ranked word, we replace that word with **[MASK]** token and feed both the original input as well as the masked input separated by a **[SEP]** token to **BERT**. **Feeding the sentence pair generates candidate words which not only fits the given context but also retain the meaning of the original word.** The candidate resulting in the maximum change in the confidence score of the target class is selected and the steps 1-2 are repeated for the next ranked word.



Experiments

- We consider text classification and entailment tasks
- We attacked **two target models** across **three benchmark datasets** and used attack success rate, perturbation rate, and grammatical correctness to evaluate our proposed attack.
- Our attack achieves more than **90% success rate** across all datasets and target models..
- In comparison to baselines, on average our attack **improves the success rate and perturbation rate by 15% and 14% respectively** across all target models and datasets.

Generated Adversarial Examples	Prediction
This film is a portrait [sketch] of grace [beauty] in this imperfect world.	Pos → Neg
Vile and tacky are the two best adjectives [words] to describe this movie..	Pos→ Neg

Conclusion

- Our novel attack generates candidate words using the **influence of both the original word (to be replaced) as well as its surrounding context.**
- Extensive experimentation and ablation studies demonstrate the effectiveness of our attack.

Paper: https://arxiv.org/pdf/2012.13339.pdf
Code: github.com/RishabhMaheshwary/contextattack