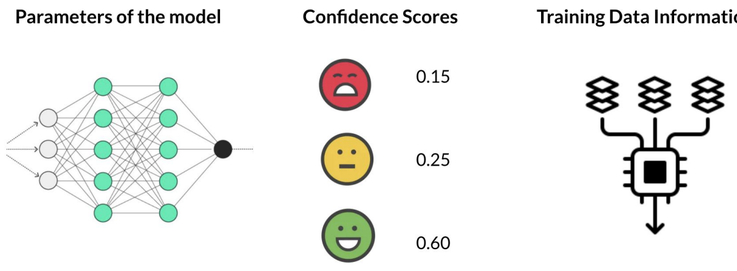
Generating Natural Language Attacks in a Hard Label Black Box Setting

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**. In many real world applications, **the attackers cannot access the details of the target model**.
- Prior Natural language attacks require access to either:



- But in real word applications we rarely have access to the above information.
- We focus on a realistic **decision based** or **hard label black box setting**. In this setting the attacker **does not has access to the parameters, training data or even the confidence score** of the target model.
- Our proposed attack method is able to **craft plausible and semantically similar** adversarial examples using only the **topmost predicted label.**

Problem Formulation

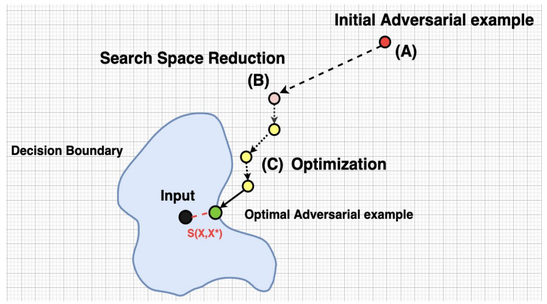
$$\max_{x^*} \; S(X, X^*) \text{ s.t. } C(F(X^*)) = 1$$

- S Semantic Similarity
- X Original text input
- X* Adversarial text input
- C Adversarial Criteria $C(F(X)) = 0$
- F Target model

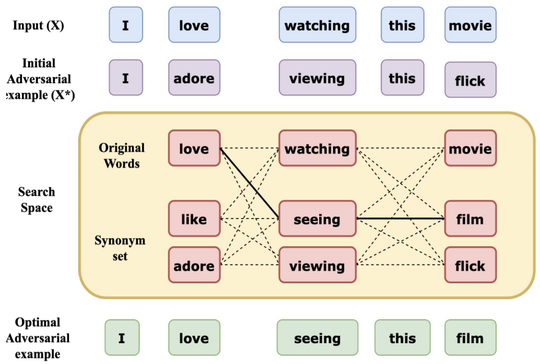
Proposed Approach

In a hard label black box setting, we cannot access the confidence scores of the target model. Therefore, our attack follows the following three steps:

- **Random Initialisation**: Generates an adversarial text by randomly substituting words with synonyms.
- **Search Space Reduction**: Replaces back the substituted words with their original counterparts.
- **Genetic Optimization**: Maximizes the semantic similarity between the adversarial and original input.
- **Overview**



- **Working Example**



Experiments

- We consider **classification and entailment tasks**.
- We attacked **five target models** across **seven benchmark datasets** and used attack success rate, perturbation rate, grammatical correctness and human evaluation to evaluate our proposed attack.
- Our attack achieves more than **90% success rate** across all datasets and target models..
- In comparison to baselines, across all datasets and target models our attack **improves the success rate by atleast 20%, reduces the perturbation and grammatical error rate by atleast 15%**.

Generated Adversarial Examples	Prediction
It's not necessarily [definitely] for kids	Neg → Pos
Will Avast protect my computer [machinery]?	Tech → Music
P: A portion of the nation's income is saved by allowing for investment. H: The nation's income is divided into portions [fractions].	Entail.→ Neutral

Conclusion

- Our novel **decision based attack** does not require access to model parameters, confidence scores, training data information and substitute models.
- Extensive experimentation and ablation studies demonstrate the effectiveness of our attack.

Paper: https://arxiv.org/pdf/2012.14956.pdf
Code: github.com/RishabhMaheshwary/hard-label-attack
Also implemented in: github.com/QData/TextAttack