Rigre Garciandia
RRG190004
CS 4395.001 - Human Language Technologies
Karen Mazidi
Spring 2023

ACL Paper Summary: "That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks

Rigre Garciandia
RRG190004
CS 4395.001 - Human Language Technologies
Karen Mazidi
Spring 2023

Author List

- Edoardo Mosca (TU Munich, Department of Informatics, Germany edoardo.mosca@tum.de) **[53 citations on google scholar]**
- Shreyash Agarwal (TU Munich, Department of Informatics, Germany shreyash.agarwal@tum.de)
- Javier Rando-Ramirez (ETH Zurich, Department of Computer Science, Switzerland jrando@student.ethz.ch) **[20 citations on google scholar]**
- Georg Groh (TU Munich, Department of Informatics, Germany grohg@in.tum.de) **[1722 citations on google scholar]**

Author with the most citations:

Georg Groh (TU Munich, Department of Informatics, Germany grohg@in.tum.de) **[1722 citations on google scholar]**

Rigre Garciandia
RRG190004
CS 4395.001 - Human Language Technologies
Karen Mazidi
Spring 2023

Summary of the problem addressed by the paper

The problem addressed by the paper is that of **adversarial attacks**. These are attacks which are "perturbed versions of the original text indiscernible by humans which get misclassified by the model" [2]. In other words, these attacks aim to fool the model into misclassifying text data by modifying the original text. Sometimes the changes made by such attacks are identifiable by humans, but increasingly complex attacks such at the BERT-based Adversarial example [2] replace input text in such a way that is not noticeable by humans. These attacks keep evolving and there is thus a need to reliably detect adversarial attacks, especially in safety-critical applications [1]. One example of such an application would be an automatic fraudulent insurance claim detector, which could be fooled through an adversarial attack into not flagging a fraudulent claim. The model-agnostic metric proposed by the authors could be used to train a pre-processing model that detects the likelihood of an adversarial attack.

Summary of prior work

Character-level adversarial attacks, i.e. changing characters in a word to make typos, are relatively easy to detect with a simple spell-checker. The more difficult task is detecting word-level attacks, which typically consist of "deletion, insertion, and replacement by synonyms or paraphrases" [1].

The prior work that the authors compare their results with is described in the paper *Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples* [3]. This technique involves "exploiting the frequency properties of adversarial word substitutions for the detection of adversarial examples" [3]. Another technique for detecting adversarial attacks is in including adversarial examples during the training of a classifier, but this approach "typically require[s] a priori attack knowledge and models to be retrained from scratch to increase their robustness" [3].

Rigre Garciandia
RRG190004
CS 4395.001 - Human Language Technologies
Karen Mazidi
Spring 2023

Techniques to detect adversarial inputs have been explored for image classifiers, which is where the authors took inspiration for their approach. These approaches work by analyzing the model's logits.

Unique Contributions of this paper

The authors come up with a technique that outperforms the state-of-the-art approach described in the paper *Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples* [3]. Their approach is also able to "generalize across multiple datasets, attacks, and target models without needing to retrain" [3].

They achieve this by introducing a new metric called Word-level Differential Reaction (WDR). This metric captures each word's impact on the target class identified by the model. It is used to capture words with a suspiciously high impact on the classifier. They compute the WDR score of each word in the input by running the target model without that word and measuring the reaction. The WDR scores for each word in the text input are then used as the features for an Adversarial Detector model that they have trained to classify inputs as original or adversarial. The model they trained is agnostic to the target model.

The reason that measuring the impact of each word works is because "adversarial attacks based on semantic similarity replace the smallest number of words possible to change the target model's prediction" [1], meaning that we can expect each word that was replaced to have a high impact on the target model's output, which is measured by the WDR score for that word. The WDR score for a word is computed by removing that word from the input and calculating the difference in the logit for the original class and the logit for the highest class excluding the original class. We can expect perturbed words to have a negative WDR score. In a non-adversarial input, removing even the most important word should not change the predicted class.

Rigre Garciandia
RRG190004
CS 4395.001 - Human Language Technologies
Karen Mazidi
Spring 2023

How the authors evaluated their work

The authors tested their adversarial input detection model by using four well-known test classification datasets: *IMDb*, *Rotten Tomatoes Movie Reviews, Yelp Polarity, and AG News.* They used several target models for their classification, such as DistilBERT, LSTM, CNN, and BERT. For each of these models they used four popular word-substitution techniques: *Probability Weighted Word Saliency (PWWS), Improved Genetic Algorithm (IGA), TextFooler, and BERT-based Adversarial Examples (BAE).* They tested each of the target models with a specific dataset and attack, and evaluated their adversarial attack detection model by comparing it to the current state-of-the-art technique of FGWS [3]. They mainly used the F1-score metric to compare their model's performance with the current state-of-the-art, but also looked at recall since false negatives are important when it comes to adversarial detection.

The authors also make sure to mention the limitations of their approach. Since the WDR score needs to be computed for each word in the input text, the authors acknowledge that their method may not be suitable "when input texts are particularly long" [1]. Another limitation that the authors foresee is that new types of attacks could be crafted to circumvent their method. These attacks would avoid relying on only "a few words to substantially affect output logits" [1].

Why I think their work was important

I believe that we may be in the middle of an AI revolution that is a sub-revolution of the larger Information Revolution that has been going on for decades. We will see increased applications of NLP techniques across industries, and as such it will increase the potential rewards of attacking critical systems. Similarly to how firewalls, encryption, and several other techniques are applied to traditional applications, I believe that approaches based on the author's work will become necessary to protect

critical systems, attacks on which could incur large costs, whether that be in human lives, financial losses, or loss of reputation

References

[1] Edoardo Mosca, Shreyash Agarwal, Javier Rando Ramírez, and Georg Groh. 2022. "That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7806–7816, Dublin, Ireland. Association for Computational Linguistics.

[2] Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based Adversarial Examples for Text Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6174–6181, Online. Association for Computational Linguistics.

[3] Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021. Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 171–186, Online. Association for Computational Linguistics.