BERT-ATTACK: Adversarial Attack Against BERT Using BERT

1. **What is the primary lesson(s) you took away from this paper (avoid abstract level summary)?**

   The authors in this paper proposed a straightforward, computation-wise, practical attack (BERT-Attack) to generate fluent, effective, and semantically-preserved adversarial attacks/examples that could successfully fool state-of-the-art models in NLP in a black-box manner, such as fine-tuned BERT for various downstream tasks. In this attack, the authors turned BERT against its fine-tuned models and other Deep Neural Networks (DNNs) models in downstream tasks so that they could successfully fool the targeted models into predicting incorrectly.

   The authors in this work evaluated their attack (BERT-Attack) on different types of NLP tasks in the form of Text Classification (they used four datasets for this task) and Natural Language Inference (they used two datasets for this task). As a result, the BERT-Attack attack outperformed state-of-the-art black-box attack approaches in both success rate and perturbed percentage. The average after-attack accuracy was lower than 10%, showing that most examples were successfully perturbed to mislead the state-of-the-art classification models.

   Lastly, the authors used human evaluations to measure the quality of the generated examples in terms of fluency, grammar, and semantic preservation. They asked the human judges to score the grammar correctness of the mixed sentences of generated adversarial samples and original sequences. Then, they asked the human judges to make predictions in a shuffled mix of original and adversarial texts. Consequently, the semantic and grammar scores of the generated adversarial examples were close to the original ones.

2. **Could I have done this work if I had the idea why or why not?**

   I do not think I could have done this work if I had the idea, given my limited knowledge of various model architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and BERT (I am taking CS570 Deep Learning class this semester).

3. **Is there anything I could do to repeat or validate? (Make notes of materials available from the paper- data sets, source code).**

   Yes, the authors open-sourced their attack (BERT-ATTACK) and published it on GitHub here: https://github.com/LinyangLee/BERT-Attack. The authors also used six publicly available datasets: The Yelp Reviews dataset [1], IMDB Movies Reviews dataset [2], AG's News dataset [3], FAKE dataset [4], The Stanford Natural Language Inference (SNLI) Corpus [5], and the Multi-Genre Natural Language Inference (MultiNLI) dataset [6].

   [1] https://www.kaggle.com/code/suzanaiacob/sentiment-analysis-of-the-yelp-reviews-data/data.
   [2] https://ai.stanford.edu/~amaas/data/sentiment/.
   [3] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.
   [4] https://www.kaggle.com/c/fake-news/data.
   [5] https://nlp.stanford.edu/projects/snli/.
   [6] https://cims.nyu.edu/~sbowman/multinli/.

4. **What is my best idea for follow on work that I could personally do?**

   My best idea for the follow-on work that I could personally do is to first repeat the same work using their English datasets to develop a better understanding of how the attack works. Secondly, I could repeat the same work or part of using Arabic text classification datasets, examine the generated results, and test if this kind of adversarial text attack works in different languages like Arabic.

5. **What is my best idea for follow on work that I'd like to see the authors or others do?**

My best idea for the follow-on work that I would like to see the authors or others do is to investigate deeply how the Enhanced Sequential Inference Model (ESIM) model became more robust in the MultiNLI dataset under the BERT-Attack. It would be very interesting to confirm the authors' assumption: "*Under BERT-Attack, the ESIM model is more robust in the MNLI dataset. We assume that encoding two sentences separately gets higher robustness*."—the authors.

6. **Any logistical experimental lessons I learned (this is a little different than #1 here you are looking for techniques)?**

I liked how the authors provided an end-to-end evaluation in their paper, covering almost every evaluation metric discussed in the research literature. I also liked the idea of dividing the evaluations into two categories: automatic evaluation metrics (success rate, perturbed percentage, and query number) and human evaluations (fluency, grammar, and semantic preservation).

Another logistical lesson I learned from his paper is how they beautifully and efficiently included at least 8-9 tables; this large number of tables seems to fit in perfectly because, from my experience, managing such a large number of tables in LaTeX or Overleaf in such a limited number of pages is a challenge and a skill I wish to master soon.

7. **How does this compare to the other papers we read? Most similar? How different? Other comparisons?**

Li et al. (2020) is the third paper I read in this course that utilizes the BERT language model to generate adversarial attacks in a black-box setting. This paper is similar to the previous two papers that propose BERT-based black-box attacks: Garg & Ramakrishnan (2020) and Jin et al. (2020). All these three papers generated black-box attacks, targeting the input space with word-level replacements. Yet, this paper by Li et al. (2020) discussed the sub-word-level replacements in their attacks, which hadn't been discussed by Garg & Ramakrishnan (2020) and Jin et al. (2020).

8. **What is your biggest criticism of the paper?**

There is no big criticism of the paper. Yet, the authors have only proposed black-box attack (BERT-Attack) but have not discussed any suggested defense mechanisms to encounter their attack.

9. **List 3 cited references or terms/concepts that you would be most interested in reading/learning more about.**

I am interested in reading these three references because they are very relevant, and the authors of this paper followed some of their approaches.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. ***Combating adversarial misspellings with robust word recognition.*** *arXiv preprint arXiv:1905.11268.*

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. ***Universal adversarial triggers for attacking and analyzing NLP.*** *Empirical Methods in Natural Language Processing.*

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. ***BERT-based lexical substitution.*** *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.*

I would like to learn more about the BERT model, especially how to fine-tune it on custom corpora. Another thing I would like to learn is the Masked Language Models (MLMs) tasks and how to use such tasks to produce very presentative word embeddings for a given corpus.