BAE: BERT-based Adversarial Examples for Text Classification

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

The authors in this paper proposed black-box attacks, BAE, an adversarial example generation approach using contextual perturbations from BERT and its Masked Language Model (MLM) task for word replacements. The BAE attacks replace or/and insert tokens in the original text/input by masking a token of the text and leveraging the BERT-MLM to produce alternatives for the masked tokens. Generally, the BAE attacks are almost always (only one exception) more effective than the baseline model/attack (TextFooler: a model for natural language attack on text classification and inference), achieving significant drops of 40% to 80% in test accuracies, with higher average semantic similarities. Specifically, the BAE-Replace+Insert attack was the strongest attack of BAE since it allows both replacement and insertion at the same token position.

The authors used three popular text classification models in this work: the word-LSTM model, the word-CNN model, and a fine-tuned BERT model (base-uncased classifier). They also used seven publicly available datasets for various tasks such as sentiment classification, opinion polarity detection, subject classification, and question type classification. These datasets are Amazon Reviews, Yelp Reviews, IMDB Movies Reviews, Movie Review Data, MPQA Opinion Corpus, Subjectivity Corpus, and TREC 10.

Lastly, the authors performed a human evaluation to evaluate the *naturalness* of the adversarial examples and show that BAE attacks yield adversarial examples with improved grammaticality and semantic coherence. The authors' human evaluation revealed that the improved grammaticality of the adversarial examples generated by BAE attacks outperformed the baseline TextFooler, which can be attributed to the BERT-MLM.

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 CNNs (word-based CNNs and character-based CNNs) and LSTMs (word-based LSTMs), 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 have open-sourced their attacks (BAE) and published them on GitHub here: https://github.com/QData/TextAttack/blob/master/textattack/attack_recipes/bae_garg_2019.py. The authors integrated their attacks (BAE) in the TextAttack, a Python framework for adversarial attacks, adversarial training, and data augmentation in NLP: https://github.com/QData/TextAttack.

The authors in this work used seven publicly available datasets: Amazon Reviews [1], Yelp Reviews [2], IMDB Movies Reviews [3], Movie Review Data [4], MPQA Opinion Corpus [5], Subjectivity Corpus [6], and TREC 10 [7].

[1] https://www.kaggle.com/datasets/bittlingmayer/amazonreviews.
[2] https://www.kaggle.com/code/suzanaiacob/sentiment-analysis-of-the-yelp-reviews-data/data.
[3] https://ai.stanford.edu/~amaas/data/sentiment/.
[4] https://www.cs.cornell.edu/people/pabo/movie-review-data/.
[5] https://mpqa.cs.pitt.edu/corpora/mpqa_corpus/.
[6] https://www.cs.cornell.edu/people/pabo/movie-review-data/.
[7] https://cogcomp.seas.upenn.edu/Data/QA/QC/.

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 it (sentiment analysis) using Arabic sentiment analysis datasets, examine the generated results, and test if this kind of adversarial text attack works in different languages like the Arabic language.

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 study the transferability of the adversarial attacks/examples generated by BAE attacks to other Deep Neural Networks (DNNs) models and study to which degree these attacks will affect the performance of those models.

Another idea I would like to see the authors or others do is to investigate the effectiveness and efficiency of using mitigation or defense mechanisms such the adversarial training and data augmentation. The authors demonstrated that BERT is very susceptible to these attacks, and showing that BERT could recover significantly would be an interesting investigation.

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 utilized the BERT model and its Masked Language Model (MLM) task to predict masked tokens to replace a token or be inserted as a token in the input sentence. Yet, this method does not guarantee the semantic coherence to the original input; therefore, the authors used a Universal Sentence Encoder (USE) based sentence similarity scorer to rank the predicted tokens by BERT-MLM.

Another logistical lesson I learned from this paper is the simple, straightforward mathematical problem definition of their adversarial examples/attacks. Their mathematical definition was easy to understand and applicable to other adversarial attacks' definitions.

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

This paper by Garg & Ramakrishnan (2020) is the first paper I read that uses Large Language Models (LLMs) like BERT to craft adversarial examples/attacks. This paper is similar to many papers I have read in this course in terms of the attack type (black-box attacks), the targeted space (input space), and the method used (word-level replacements or insertions) such as Gao et al. (2018) and Li et al. (2019). In contrast, this paper of Garg & Ramakrishnan (2020) differs from a few papers as well in terms of attack type (white-box attacks), the targeted space (embedding space), and the method used (character-level replacements or insertions) such as Kuleshov et al. (2018).

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

There is no big criticism of the paper. Yet, the authors have only proposed novel black-box attacks (BAE) but have not discussed any suggested defense mechanisms to counter their attacks.

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

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.