Contextualized Perturbation for Textual Adversarial Attack

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

   The authors of this paper proposed, CLARE, a contextualized adversarial example generation model that generates fluent and grammatical outputs through a mask-then-infill procedure using Masked Language Models (MLMs) and modifies the inputs in a context-aware manner. CLEAR introduced three contextualized perturbations: *Replace*, *Insert* and *Merge*, which permits generating outputs of varying lengths, where it could flexibly integrate these perturbations and apply them at any position in the inputs and then use them to attack the victim model more effectively with fewer edits as possible.

   The authors in this work performed comprehensive experiments and human evaluation to demonstrate that CLARE outperformed the baselines regarding attack success rate, textual similarity, fluency, and grammaticality. For instance, CLARE outperformed BERT-Attack, the strongest baseline, by a more than 5.4% attack success rate with fewer average modifications to the text. CLARE also slightly underperformed TextFooler at 68% with a 95% confidence interval (66%—70%) versus 70% with a 95% confidence interval (68%—73%). Additionally, the authors studied the performance of these three perturbations when applied individually. Interestingly, they found that *Merge* **only** underperformed the other two (*Replace* and *Insert*), partly because the attacks are restricted to bigram noun phrases. Yet, by combining all these three perturbations, CLARE achieved the best performance with the least modifications

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 Large Language Models (LLMs) architectures and their Masked-Language Modeling tasks, like the BERT and RoBERTa models (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 (CLARE) and published them on GitHub here: https://github.com/cookielee77/CLARE. The authors in this work used seven publicly available datasets: Yelp Restaurant Reviews [1], AG's News Corpus [2], Multi-Genre NLI (MultiNLI) dataset [3], and Question-answering NLI (QNLI) dataset [4].

   [1] https://www.tensorflow.org/datasets/catalog/yelp_polarity_reviews.
   [2] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.
   [3] https://cims.nyu.edu/~sbowman/multinli/.
   [4] https://rajpurkar.github.io/SQuAD-explorer/.

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 (text classification) using Arabic text classification 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?**

   I do *not* have an idea for the follow-on work I would like to see the authors or others do.

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

   I do *not* have any logistical experimental lessons I learned from this paper.

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

   This paper by Li et al. (2021) is the fifth paper that utilized Google's BERT language model or its variants. In this paper, the authors used the RoBERTa model to generate the masked tokens in their adversarial examples as the first paper among the previous papers. Yet, it is similar to those papers utilizing the Large Language Models (LLMs) and their Masked-Language Modeling tasks.

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

   There is *no* big criticism of the paper.

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. The authors of this paper discussed these concepts/models, and I think having a background about them is important.

   Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. ***Language models are unsupervised multitask learners.*** *OpenAI Blog, 1(8):9.*

   Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. ***DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.*** *In The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS.*

   Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. ***Sentence-level fluency evaluation: References help, but can be spared!*** *In Proc. of CNLP, pages 313–323.*