Generating Natural Language Adversarial Examples Through Probability Weighted-Word-Saliency

1. **What is the primary lesson(s) you took away from this paper (avoid abstract level summary)?**
   The authors in this paper introduced a new word-level/synonym replacement attack (black-box attack) determined by both the word saliency and the classification probability and proposed a greedy algorithm called Probability Weighted Word Saliency (PWWS) for adversarial text attacks. The authors replaced the words in the input texts with synonyms using WordNet (a lexical database for English) and replaced the Named Entities (NEs) with similar NEs to generate adversarial samples. Additionally, the authors experimented on three publicly available datasets using CNNs and LSTMs models to demonstrate the effect of the attacks. The PWWS reduced the accuracy of the DNNs classifiers by up to 84.03%. Lastly, the authors did a human evaluation to show that their perturbations are difficult for humans to perceive and demonstrated that adversarial training using their adversarial examples could help improve the robustness of the text classification models.

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 bidirectional-LSTMs (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 are open-sourcing their attacks (PWWS) and published them on GitHub here: https://github.com/JHL-HUST/PWWS. They further shared with the community their Keras implementations of the PWWS on GitHub here: https://github.com/RenShuhuai-Andy/PWWS. Lastly, the authors used WordNet, a lexical database for English, [4] build a synonym set that contains all synonyms of each selected word in their word substitution strategy.

   The authors in this work used three publicly available datasets: IMDB Movies Reviews dataset [1], AG's dataset of News Articles [2], and Yahoo! Answers dataset [3].

   [1] IMDB dataset: https://ai.stanford.edu/~amaas/data/sentiment/.
   [2] AG's dataset: http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.
   [3] Yahoo! dataset: https://github.com/LC-John/Yahoo-Answers-Topic-Classification-Dataset.
   [4] WordNet: https://wordnet.princeton.edu/.

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 evaluate this method on more datasets and models in terms of effectiveness and efficiency. A comprehensive human study on the similarity between the adversarial examples and their corresponding clean examples would be interesting.

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 performed a human evaluation using six workers on Amazon Mechanical Turk (AMT) to evaluate the adversarial examples generated by their method PWWS in terms of semantic similarity and to show that these examples are hard for humans to perceive. In contrast, other papers in this domain ask the AMT works to classify the generated adversarial examples mixed with clean examples to show that their adversarial examples still deceive a human classifier or an annotator. Yet, I think demonstrating that the adversarial examples could maintain semantic similarity is a deeper analysis than only classifying/labeling these adversarial examples.

Another experimental lesson I learned from this paper, Ren et al. (2019), is that when you have a successful novel idea, you do not have to benchmark it with close pre-existing ideas; make variants/sub-ideas of this big idea and then compare these sub-ideas head-to-head then compare these sub-ideas to the big idea. In this paper, the authors did not compare their methods/attacks (PWWS) to other pre-existing methods/attacks; instead, they created variant attacks like random or gradient attacks to compare to their methods/attacks (PWWS).

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

This paper by Ren et al. (2019) is similar to a few papers I read throughout this course, like Gao et al. (2018), Li et al. (2019), and Alshemali and Kalita (2019), in terms of the attack type (black-box attacks), the targeted space (input space), and the method used (character-level replacements). In contrast, this paper, Ren et al. (2019), differs from a few papers as well in terms of almost everything (attack type: white-box, target: embedding space, and method: word-level replacements) like Ebrahimi et al. (2018) and Kuleshov et al. (2018).

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 the references listed below because they are relevant and have been cited in this paper, and they do not have them on my to-read list.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016a. ***The limitations of deep learning in adversarial settings****. In IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrucken, Germany, March 21-24, 2016, pages 372–387.*

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. ***Deep text classification can be fooled****. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden., pages 4208–4215.*

Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. 2018. ***Interpretable adversarial perturbation in input embedding space for text.*** *In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden., pages 4323–4330.*

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. 2018. ***Greedy attack and gumbel attack: Generating adversarial examples for discrete data****. CoRR, abs/1805.12316.*