Certified Robustness to Adversarial Word Substitutions

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

   The authors in this paper studied the possibility of guaranteeing that a model is robust against all adversarial perturbations/examples of a given input and the possibility of robust training models. The authors obtained such guarantees by leveraging Interval Bound Propagation (IBP), a technique previously applied to feedforward networks and CNNs in computer vision. This IBP efficiently computes a tractable *upper bound* on the loss of the worst-case perturbation. When this upper bound on the worst-case loss is small, the model is guaranteed robust to all adversarial perturbations/examples, providing a *certificate* of robustness.

   Additionally, the authors used robust training to train models to optimize the IBP upper bound by certifiably evaluating robust training on two tasks — sentiment analysis on the IMDB dataset and natural language inference on the SNLI dataset. Across various model architectures, such as bag-of-words, CNN, LSTM, and attention-based, certifiably robust training consistently yields robust models to all perturbations on many test examples.

   Lastly, the authors interestingly found that a normally-trained model has only 8% and 41% accuracy on IMDB and SNLI, respectively, when evaluated on adversarially perturbed test examples. With certifiably robust training, they achieved 75% adversarial accuracy for both IMDB and SNLI. Data augmentation fares much worse than certifiably robust training, with adversarial accuracies falling to 35% and 71%, respectively.

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 bag-of-words, CNNs, LSTMs, and attention-based models (I am taking CS570 Deep Learning class this semester).

   I also do not think I could mathematically define the Interval Bound Propagation (IBP) even if I had the idea. The authors of this work introduced nearly eight mathematical definitions.

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 data, experiments, and models, and they publicly published their source code on CodaLab here: https://bit.ly/2KVxIFN.

   The authors also used two free datasets: the IMDB Movie Reviews – Sentiment Analysis dataset [https://ai.stanford.edu/~amaas/data/sentiment/] and the Stanford Natural Language Inference (SNLI) Corpus [https://nlp.stanford.edu/projects/snli/].

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

   I do not have an idea for the follow-on work that I could personally do.

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 use the Interval Bound Propagation (IBP) to handle other adversarial examples/attacks such as character-level typos, word insertions, and deletions; rather than just word substitutions as in this work.

   Another idea I would like to see the authors do is "*to train models that get state-of-the-art clean accuracy while also being provably robust; achieving this remains an open problem*" – 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 mathematically defined the Interval Bound Propagation (IBP) to minimize the upper bound on the worst-case loss that any combination of word substitutions can induce in training procedures.

I also liked the supplemental material section and how much valuable information could be used there as authors to explain the fine-grained details of our works profoundly and illustrate our ideas more clearly.

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

In this paper, Jia et al. (2019) used the approach of Alzantot et al. (2018), which used word substitution perturbations, where the attacker could replace every word in the input with a similar word (that must not change the label). Yet, the authors made three modifications to this approach: allowing substitutions relative to the original sentence and disallowing repeated substitutions, using a faster language model that allows them to query more extended contexts, and using the language model constraint only at test time.

Jia et al. (2019) is similar to Alzantot et al. (2018) and Kuleshov et al. (2018) in terms of the targeted space (embedding space) and the used method (word-level replacements) and similar as well to Gao et al. (2018) in terms of the attack type (black-box attacks). In contrast, Jia et al. (2019) differs from Ebrahimi et al. (2018) in terms of the attack type (white-box attacks), the targeted space (input space), and the used method (character-level replacements) and differs also from Kuleshov et al. (2018) in terms of the attack type (white-box attacks).

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 (Jia and Liang, 2017; Iyyer et al., 2018; Ribeiro et al., 2018) because they discuss adversarial examples/attacks that I have not explored yet, which are based insertion of irrelevant text and paraphrasing, respectively.

R. Jia and P. Liang. 2017. ***Adversarial examples for evaluating reading comprehension systems.*** *In Empirical Methods in Natural Language Processing (EMNLP).*

M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. 2018. ***Adversarial example generation with syntactically controlled paraphrase networks.*** *In North American Association for Computational Linguistics (NAACL).*

M. T. Ribeiro, S. Singh, and C. Guestrin. 2018. ***Semantically equivalent adversarial rules for debugging NLP models****. In Association for Computational Linguistics (ACL).*

I would like to learn/improve the skill of mathematically defining natural language processing tasks, machine learning, and deep learning architectures through equations. Writing codes for these tasks and architectures is much easier than defining them mathematically, at least for me.