Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

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

   The authors in this paper proposed adversarial examples for Sequence-to-Sequence (Seq2Seq) models that input discrete text strings and output an almost infinite number of possibilities. Seq2Seq is an optimization-based framework that aims to learn an input sequence close enough to the original sequence in terms of distance metrics in word embedding spaces or sentiment classification. The authors faced many hurdles, such as the challenges caused by the discrete input space and the almost infinite output space, and to overcome these hurdles, the authors respectively proposed a projected gradient method combined with group lasso and gradient regularization and designed novel loss functions to conduct a non-overlapping attack and targeted keyword attack.

   The authors applied their Seq2Seq algorithm to Machine Translation and Text Summarization tasks. They also verified the effectiveness of their proposed algorithm, where they could make the Seq2Seq model produce desired outputs (adversarial examples) with high success. The authors also showed that their algorithm only needs to change 2 or 3 words on average and can generate entirely different outputs for more than of sentences 80% and showed also that only 2.2% of adversarial examples have semantic meaning differ from the original sentences. Their results proved the proposed framework (Seq2Sick) was powerful and effective, where it could achieve high success rates in both *non-overlapping* and *targeted keywords* attacks with relatively small distortions and preserve similar sentiment classification results for most of the generated adversarial examples.

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 LSTMs models (word-based LSTMs), Encoder-Decoders models with attention, and Sequence-to-Sequence (Seq2Seq) 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 (Seq2Sick) and published them on GitHub here: https://github.com/cmhcbb/Seq2Sick. The authors in this work used four publicly available datasets: Documents, Tasks, and Measures (DUC2003) [1], Documents, Tasks, and Measures (DUC2004) [2], English Gigaword [3], and WMT'16 Multimodal Translation task [4].

   [1] DUC2003: https://duc.nist.gov/duc2003/tasks.html.
   [2] DUC2004: https://duc.nist.gov/duc2004/.
   [3] Gigaword: https://www.tensorflow.org/datasets/catalog/gigaword.
   [4] WMT'16 Task: https://www.statmt.org/wmt16/translation-task.html#download.

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 (machine translation) using parallel machine translation 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 Seq2Sick attacks to other Deep Neural Networks (DNNs) models such as RNNs and CNNs and study to which degree these attacks will affect the performance of those models. The authors claimed that the seq2seq models are more robust since they have discrete input space and exponentially large output space, whereas CNN-based classifiers are highly susceptible to these attacks.

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 summarized and compared ten existing works designed to attack RNN models in terms of gradient-based, word-level RNN, sequential output, and targeted attack. I also like how the paper was short and delivered the idea of the Seq2Sick attacks clearly.

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

This paper by Cheng et al. (2020) is the first-ever paper that discusses the adversarial example/attacks for Sequence-to-Sequence (Seq2Seq) models to fool machine translation and text summarization tasks. The authors proposed "*a projected gradient method to address the issue of discrete input space, adopt group lasso to enforce the sparsity of the distortion, and develop a regularization technique to further improve the success rate."* – the authors.

In general, this work of Cheng et al. (2020) falls in the white-box attacks that consider word-level replacements in the input space. Even though most of the papers I have read so far are about Recurrent Neural Networks (RNNs), this paper is similar to a few papers I read earlier, like Ebrahimi et al. (2018) and Li et al. (2019).

8. What is your biggest criticism of the paper?

There is no big criticism of the paper. Yet, the authors have only proposed an optimization-based white-box attacks (Seq2Sick) but have not discussed any suggested defense mechanisms to encounter their attacks.

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

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. ***Neural machine translation by jointly learning to align and translate.*** *arXiv preprint arXiv:1409.0473.*

Michel, P.; Li, X.; Neubig, G.; and Pino, J. M. 2019. ***On evaluation of adversarial perturbations for sequence-to-sequence models.*** *arXiv preprint arXiv:1903.06620.*

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. ***Sequence to sequence learning with neural networks****. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13, 2014, Montreal, Quebec, Canada.*

I would like to learn more about the Sequence-to-Sequence (Seq2Seq) models and their applications in NLP, especially using Arabic. The Seq2Seq models are used in various NLP applications such as Machine Translation, Question Answering, Chatbots, Text Summarization, etc.