Pathologies of Neural Models Make Interpretations Difficult

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

   The authors in this paper studied how a model's prediction can be influenced by *unimportant* words using input reduction. This process removes unimportant words from the input while maintaining the model's prediction, called "leave-one-out". The words remaining after input reduction should be important for prediction. Yet, the reduced input is meaningless to humans but retains the same model prediction with high confidence.

   The authors also constructed more of these counterintuitive examples, *rubbish examples*, by augmenting input reduction with beam search and experimenting with three tasks: SQUAD for reading comprehension, SNLI for textual entailment, and VQA for visual question answering. The input reduction with beam search consistently reduces the input sentence to very short lengths — often only one or two words — without lowering model confidence on its original prediction.

   The authors drew connections to adversarial examples and confidence calibration and explained why the observed pathologies are a consequence of the overconfidence of neural models. They also encouraged high model uncertainty on reduced examples with entropy regularization. The pathological model behavior under input reduction is mitigated, leading to better-reduced examples. Lastly, crowdsourced experiments confirmed that reduced examples appear nonsensical to humans, highlighting that input reduction uncovers pathological model behaviors.

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 interpretation of AI systems or Artificial Neural Networks (ANNs) predictions. In my past 600-level class, CS 675 Fairness, Accountability and Transparency in AI and Automated Systems, I read a few papers (3-4 papers) about interpreting AI systems using frameworks like LIME and SHAP. I never had the chance to replicate any of these papers or part of them. Yet, this is a very interesting field that uses adversarial examples/attacks to advance the interpretation of AI systems.

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 input reduction method (rubbish examples), and they publicly published their source code on GitHub here: https://github.com/ihsgnef/pathologies.

   The authors also used three free datasets in their experiments: the Stanford Question Answering Dataset (SQuAD) [https://rajpurkar.github.io/SQuAD-explorer], the Stanford Natural Language Inference (SNLI) corpus [https://nlp.stanford.edu/projects/snli], and Visual Question Answering (VQA) dataset [https://visualqa.org]. They used three published models: DRQA Document Reader [https://github.com/facebookresearch/DrQA] with SQuAD; Bilateral Multi-Perspective Matching (BIMPM) [https://github.com/zhiguowang/BiMPM] with SNLI; and Show, Ask, Attend, and Answer model [https://github.com/adamcasson/show_ask_attend_answer] with VQA.

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 input reduction method works. Secondly, I could repeat the same work or part of it (question answering) using Arabic question answering datasets, examine the generated results, and test if these rubbish examples work in different languages like Arabic.

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 that 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 liked how the authors mathematically defined the importance metrics in natural language contexts, introduced the efficient gradient-based approximation, and used this approximation in all their experiments.

   I also liked the idea of having a human study as an evaluation method for their experiments' results, and at the same time, they generated new results that could be a contribution by comparing them to the experiments' results. The authors recruited two groups of crowd workers and tasked them to compare the human accuracy on original and reduced examples. They showed one group the original inputs and the other the reduced. The results were that humans could no longer give the correct answer, showing a significant accuracy loss on all three tasks of reading comprehension, textual entailment, and visual question answering.

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

   This paper, Feng et al. (2018), is the first paper I read that does not discuss any attacks/adversarial examples. Feng et al. (2018) discuss only the rubbish examples produced by removing the least important word of an input sentence that does not change the model accuracy/confidence. Feng et al. (2018) is similar to both Gao et al. (2018) and Ebrahimi et al. (2018) in terms of the input space but differs from all papers in terms of the attack type and the methods used. The significant difference is that Feng et al. (2018) aims to preserve the success rate of a model, whereas all other papers aim to reduce the success rate of a model.

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 this reference (Li et al., 2018). The authors built their work in this paper based on this reference method of identifying the importance of words called "leave-one-out".

   Jiwei Li, Will Monroe, and Daniel Jurafsky. 2016b. ***Understanding neural networks through representation erasure.*** *arXiv preprint arXiv: 1612.08220.*

   I am also interested in reading this reference (Simonyan et al., 2014). Using the input gradient, the authors used this reference to approximate a word's removal, which computes the importance of all words in the backward pass.

   Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. ***Deep inside convolutional networks: Visualising image classification models and saliency maps.*** *In Proceedings of the International Conference on Learning Representations.*

   I would like to learn about the beam search algorithm mainly used in most adversarial examples/attacks papers I have read so far. Yet, a few papers used the greedy algorithm, but the beam search algorithm results are more accurate.