Pre-proposal: Adversarial Analysis of Natural Language Inference Systems

Tiffany Chien
University of California, Berkeley

Jugal Kalita
University of Colorado, Colorado Springs

1. Introduction

In recent years, deep learning models have achieved and continued to improve on state-of-the-art results on pretty much all NLP tasks. However, adversarial examples – small perturbations to a model's inputs that cause its outputs to be wrong – have revealed weaknesses in many models' robustness and generalization abilities. Furthermore, the end-to-end, black-box nature of these models greatly limits the interpretability of their decisions, making it even more difficult to modify them to defend against (known and unknown) adversarial attacks.

One could say that the fundamental shortcoming of all current NLP systems is a lack of a deep, 'real' understanding of language. Therefore, we will focus on the task of natural language inference (NLI), a very direct test of natural language understanding. The setup of the NLI task is to determine whether a *hypothesis* is true given a *premise*, answering *entailment*, *contradiction*, or *neutral* (the hypothesis's truth value can't be determined).

The current top-performing systems for NLI are built on pretrained BERT (Bidirectional Encoder Representations from Transformers) [Devlin *et al.*,] representations of sentences, fine-tuned to specific tasks. The architecture of BERT is built on the Transformer architecture [Vaswani *et al.*,], which is a non-recurrent, purely attention-based neural model architecture. While many analyses of recurrent models (RNNs, LSTMs of various kinds) exist, there is relatively little work on Transformer-based architectures.

Our goal is to pit BERT and other high-performing models with varying architectures against different adversarial attacks, and carefully examine when and why they fail. Those insights will help us interpret what the model is really learning and understanding, and hopefully lead to ideas for defending against those attacks and generally improving the performance of the model. It is also possible that the failures of these models come not from their architecture, but from significant gaps in their training data; we will try to disentangle that blame.

2. Related Work

Various ways of generating adversarial examples (also called challenge datasets) have been shown to break state-of-the-art models:

- McCoy et al. (2019) [McCoy *et al.*,] show that BERT and other models use specific fallible, surface-level syntactic heuristics, by testing them on a challenge dataset (HANS) that violates those heuristics. Their heuristics can be categorised into lexical overlap, subsequence, and constituent.
- Glockner et al. (2018) [Glockner *et al.*,] created a challenge dataset focused on simple lexical inferences and world knowledge by replacing words in the SNLI dataset with related words.
- Naik et al. (2018) [Naik *et al.*,] created examples based on antonyms, numerical reasoning, word overlap (append "and true is true" to hypothesis) negation words (append "false is not true"), length mismatch (append "and true is true" 5 times), and spelling errors.

Liu et al. (2019) [Liu *et al.*,] describe a method of analysing challenge datasets, to figure out whether the model or the original training dataset is the primary cause of weakness of the trained model. Using a metaphor of inoculation, they expose the model to just a few challenge examples, and then re-test it on the original test set and the challenge test set.

1) If the model still fails the challenge dataset, the weakness probably lies in its design/architecture.
2) If the model can now succeed at the challenge dataset (without sacrificing performance on the original dataset), then the original dataset is at fault.
3) If the model does better on the challenge dataset but worse on the original dataset, the model is learning something not generalizable from the extra challenge data. This suggests that the challenge dataset is somehow not representative of the phenomenon it was trying to test, for example being very skewed to a particular label.

One persistent challenge if existing datasets are at fault is that it is impossible to augment the training set with every possible kind of challenge example, meaning that we can only ever effectively defend against known attacks.

3. Proposed Work

The first step will be to reproduce the models' results on MNLI (the dominant NLI dataset).

Then, we will take two approaches to understanding the models' failures on adversarial examples:

1) By testing the models on different kinds of adversarial examples, we can see which models are better and worse at encoding various kinds of information, for example being better at learning syntactic or semantic knowledge.
2) We will examine the models' decision-making process with an existing visualization tool such as [Vig,], which displays the weights learned and activated in different parts of the model by a particular input.

If we gain useful insights about the models from those analyses, we can maybe incorporate them into modifying model design, or adding (e.g. preprocessing) steps to the system pipeline.

4. Conclusion

In this work, we aim to chip away at the problem of model interpretability in modern neural models, by examining successful models' failures on adversarial examples in the task of NLI. Interpretability leads to more useful (and safe) models for practical use, as well as helping to generate ideas for future higher-performing models.

References

[Devlin *et al.*,] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding.

[Glockner *et al.*,] Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655. Association for Computational Linguistics.

[Liu *et al.*,] Nelson F. Liu, Roy Schwartz, and Noah A. Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets.

[McCoy *et al.*,] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.

[Naik *et al.*,] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference.

[Vaswani *et al.*,] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need.

[Vig,] Jesse Vig. Visualizing attention in transformer-based language representation models.