Literature Review for CS224U, Spring 2020.

Abhishek Goswami
Microsoft
agoswami@microsoft.com

Abstract

This is a literature review for CS224U, Spring 2020 quarter.

1 General Problem

There two general problem we are trying to explore: (1) The role played by attention mechanisms in transformer based models. (2) The Natural Language Inference (NLI) task in the adversarial setting. In recent years attention-based Transformer models have been show to be very effective for the NLI task. But they have some shortcomings, especially in adversarial scenarios.

2 Article Summary

In this section we provide summaries of several papers.

2.1 (Devlin et al., 2019)

The paper introduces several innovations (1) the Masked Language Model (MLM) objective which enables the model to look at both the left and right context during pre-training. This approach stands out, because all prior approaches tried to approach pre-training from a "pure" language modeling objective. As such, the prior approaches were all constrained in sticking to a strictly "left-to-right" or "right-to-left" mindset when doing language modeling. Though the Cloze task is well known in literature, it was mostly used in the context of QA systems or for evaluating models. The BERT authors extended that idea and introduced the MLM objective for pre-training. (2) The paper also promotes an end-to-end fine tuning approach rather than a feature representation approach.To me this bias for end-to-end fine tuning reflects in the way the model is structured, with the use of special tokens [CLS] and [SEP] . Perhaps this also served as the motivation for the next-sentence prediction task the authors introduces. (3) By stacking a large number of transformer layers, and using a unsupervised approach the BERT model is able to capture latent signals, thereby lending itself to do very well on several language modeling tasks.

2.2 (Nie et al., 2019)

The paper relates to a data collection strategy in the adverserial setting. To do this, the authors propose an iterative setting called HAMLET (Human-And-Model-in-the-Loop EnabledTraining). and apply it in the NLI task. The setting treats the human as a hacker, trying to find vulnerabilities in the model. When the human does find a valid vulnerability (as validated by 2 additional validators) that vulnerable example makes it to one of the train/dev/test examples for that round. The train set also includes examples where the model made the right prediction. At the end of a round, a new model is generated using data generated during that round. The authors show that training new models following this paradigm leads to state-of-the-art models.

2.3 (Michel et al., 2019)

The paper primarily studies the role of multi-headed attention at test time for 2 kinds of tasks Machine Translation and NLI. The authors note that at test time, many attention heads can be removed individually without impacting model performance (in terms of BLUE score for MT and Accuracy score for NLI). In fact, in many layers having a single attention head is sufficient without having significant drop in model performance. This opens the door to have a greedy algorithm to identify which heads to remove from the entire network, leading to improved inference time latency. The pruning algorithm relies on computing a head importance score to identify which nodes to prune.

2.4 (Clark et al., 2019)

This paper explores what neural networks learn about language from the perspective of BERT's attention maps. There is a sense of interpretability that we get from attention maps, and the paper tries to extract that using the 144 attention heads in BERT.

2.5 (McCoy et al., 2019)

This paper is related to the adverserial NLI setting. The authors opine that current NLI systems only learn shallow heuristics from the train data. These systems do not generalize well when the test data is drawn from a different distribution. In order to show that current NLI systems learn only shallow heuristics, he authors construct a new test dataset called HANS. The HANS dataset comprises of 3 syntactic heuristics : Lexical, Subsequent, Constituent. The authors show that current deep learning models indeed incorporate such syntactic heuristics quite strongly, and consequently perform quite poorly when test data contradicts those syntactic properties. Interestingly, for test data which contradict syntactic properties, the accuracies are much poorer than random predictions. The authors provide a discussion of whether the poor results on HANS arise from data limitations, model limitations, or both.

3 Compare And Contrast

The papers reviewed in Section 2 fall into two main categories:

Role of multi-headed attention in BERT: These set of papers study the structural properties of a popular deep learning model, BERT (Devlin et al., 2019) from the perspective of what it does (Michel et al., 2019; Rogers et al., 2020; Clark et al., 2019).

Adverserial NLI: These set of papers point out limitations in BERT, particularly in the NLI setting. The limitations seem to be pointing to heuristics inherent in the data (McCoy et al., 2019) and how an iterative process of data collection (Nie et al., 2019) can alleviate some of those limitations.

The two categories have a unifying thread running through them; there are common patterns in BERT's behavior with respect to linguistic and syntactic features that it picks up from the underlying data. This is particularly true for the NLI task, confirmed both heuristically (McCoy et al., 2019) and through an analysis of BERT's attention mechanisms (Michel et al., 2019; Clark et al., 2019). Adversarial NLI exposes these patterns, and provides new data sets to iteratively improve the model by providing richer data.

4 Future Work

We identify some strands of work for future:

1. We want to build Transformer based models for Adverserial NLI and compare them with non-deep learning models. We hope that a comparison of these approaches for the task of Adverserial NLI will give us some grounds to build up more intuition about the Adversarial NLI task.

2. We want to explore whether the role attributed to multi-headed attention in the non-adversarial setting also holds true for adversarial models. The hunch we have is that attention heads will have a markedly different role in the adversarial setting given the nature of the data.

3. We want to study the role of positional embeddings in the adversarial setting.

The author is a SCPD student in a single person team. There are no external collaborators. We are looking forward to a mentor being assigned to us. We are are not sharing this project with any other class.

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.