An Observation of BERT in 2020 Sarcasm Detection Competition Twitter Dataset

Bei Zhao, beizhao3@illinois.edu

Tech Review Assignment

1. Introduction

BERT's full name is Bidirectional Encoder Representations from Transformers. It breaks the traditional fine-tuning pre-trained language model which is instead of using a unidirectional language model, it uses a bidirectional language model (Devlin et al). With the bidirectional language model, we will be able to identify or classify text contents not only based on its past words but also on future words (Ingham, F.) According to Devlin et al, BERT beat its previous version language model, OpenAI GPT. Furthermore, in the General Language Understanding Evaluation, the current top three are all using the improved version of BERT (GLUE Benchmark Leaderboard), which indicates the power of BERT in the text classification area. In this review, it will discuss the observation of BERT in the 2020 Sarcasm Detection Shared Task Competition Twitter Dataset, since the top 7 teams in this competition are using BERT as their main approach.

2. Competition Brief Description

2020 Sarcasm Detection Shared Task Competition is a Classification Competition to identify whether a response is sarcastic or not in Reddit and Twitter dataset supplied by the organizer. The competitors need to train the supplied dataset and submit it into Codalab to do the evaluation and see the rank of their results (Debanjan Ghosh et al). This competition report includes the top 13 teams' papers brief introduction and the interesting thing is that 11 of them are using techniques related to BERT. In the following section, I will mainly discuss the top 5 BERT team's techniques in the Twitter dataset and analyze useful information extracted from them.

3. Top 5 Team's BERT Related Method Analysis in Twitter Dataset

The fifth place in the Twitter dataset is tanvidadu, they were using a fine-tuned RoBERTa-large model (355 Million parameters with over a 50K vocabulary size) on response and its two immediate contexts (Debanjan Ghosh et al). RoBERTa model's full name is Robustly optimized BERT approach. It is trained with dynamic masking, Full-sentences without NSP (Next Sentence Prediction) loss, large mini-batched, and a larger byte-level BPE (Byte-Pair Encoding) (Yinhan Liu et al). According to the evaluation result, RoBERTa's performance is better than BERT itself on GLUE, RACE, SQuAD, SuperGLUE, and XNLI. For the Twitter sarcasm detection, with the advanced pre-trained RoBERTa-large model, tanvidadu team compared the experimental results with three different input methods, Response-only, Context-Response, and Context-Response (Separated). They utilized Context-Response one as their final submission, which indicates the RoBERTa-large model itself has a good separation function implemented and no extra separating needs for Twitter sarcasm classification (Dadu and Pant et al).

The fourth place in the Twitter dataset is ad6398, they were comparing multiple transformer architectures, which are BERT, RoBERTa, and spanBERT(Debanjan Ghosh et al). To improve the experimental results, they tried to add RNN (LSTM) or extra transformers (Siamese type architecture and Dual transformer) to achieve it. Based on their testing result, RoBERTa combined with LSTM has the best score (Kumar and Anand et al).

The third place in the Twitter dataset is andy3223, they were mainly introducing two different models which are target-oriented where only the target is modeled and context-aware where the context is modeled with the target (Debanjan Ghosh et al). It is quite the same procedure as the fifth team, but their model compared all the prior turns while the fifth team only compared the prior two turns, which makes the andy3223 team get a better result than tanvidadu team did. Their best result is also the RoBERTa-large model with a context-aware model (Xiangjue Dong et al).

The second place in the Twitter dataset is nclabj, they were using the majority-voting ensemble of RoBERTa models weight-initialization and different levels of context length to do the experimental testing. Their report shows that the previous 3 turns of dialogues had the best performance in isolation(Debanjan Ghosh et al). For the Twitter dataset, the majority-voting ensemble of RoBERTa with multiple weight-initialization returns the best result (Nikhil Jaiswal).

The first place in the Twitter dataset is miroblog, their score is quite high in this Competition. Their score is at 0.8 level while the other teams are at 0.7 level. Different from the other top 4 teams' RoBERTa method, they were using a novel way in their best result, which implements a BERT classifier followed by BiLSTM and NeXtVLAD, which is a differentiable pooling mechanism then mean and max-pooling usually used (Debanjan Ghosh et al). They also proposed a new data augmentation technique called CRA(Contextual Response Augmentation), which utilizes the conversational context of the unlabeled dataset to generate new training samples (Hankyol Lee et al).

Based on the summaries of the top 5 BERT method teams in the sarcasm classification competition, RoBERTa methods are utilized quite often. Although it doesn't improve any basic concept of BERT, eliminating some BERT limitations in training helps it get a great improvement in text classification. The most impressive team is the winner one, they didn't improve the advanced BERT pre-trained model to obtain a better result as the other teams did. Instead, they mainly focus on optimizing the data feed into the BERT model. With this novel idea, they got much higher scores than the other teams. From this perspective, if the No 1 team optimal data could feed into advanced BERT such as RoBERTa, the F1 score might be higher than the No 1 team currently scores. This needs to be approved by more experiments.

4. Conclusion

This review mainly analyzes the top 5 BERT method teams' results on the Twitter data. It shows that although BERT itself cannot be improved from its basic concept, such as the big changes from unidirectional language model into bidirectional language model, improving BERT through changing the training method for BERT, adding some extra

neural network structures, or improving the data feed into the BERT will always improve the precision, recall and F1 scores for the sarcasm classification. Currently, there are no 0.9 scores in the sarcasm classification, but with more training condition changes, digging into the BERT extended architecture, and raising up new models, a 0.9 level score is not too far to be achieved.

5. References

Amardeep Kumar and Vivek Anand 2020. Transformers on Sarcasm Detection with Context. *Proceedings of the Second Workshop on Figurative Language Processing*

Debanjan Ghosh, Avijit Vajpayee, Smaranda Muresan, and Educational Testing Service 2Data Science Institute, Columbia University {dghosh, avajpayee}@ets.org smara@columbia.edu 2020. A Report on the 2020 Sarcasm Detection Shared Task. *arXiv preprint arXiv:2005.05814.*

GLUE Benchmark Leaderboard. (n.d.). Retrieved November 11, 2020, from https://gluebenchmark.com/leaderboard

Hankyol Lee, Youngjae Yu, and Gunhee Kim. Augmenting Data for Sarcasm Detection with Unlabeled Conversation Context. *arXiv preprint arXiv:2006.06259.*

Ingham, F. (2018, November 27). Dissecting BERT Part 2: BERT Specifics. Retrieved November 11, 2020, from https://medium.com/dissecting-bert/dissecting-bert-part2-335ff2ed9c73

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, and Google AI Language {jacobdevlin,mingweichang,kentonl,kristout}@google.com 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805.*

Nikhil Jaiswal 2020. Neural Sarcasm Detection using Conversation Context. *Proceedings of the Second Workshop on Figurative Language Processing*

Tanvi Dadu and Kartikey Pant 2020. Sarcasm Detection using Context Separators in Online Discourse. *arXiv preprint arXiv:2006.00850.*

Xiangjue Dong, Changmao Li, and Jinho D. Choi 2020.Transformer-based Context-aware Sarcasm Detection in Conversation Threads from Social Media. *arXiv preprint arXiv:2005.11424.*

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Conference paper at ICLR 2020*