Paper Critique - 1

Intro of Data Mining - Fall, 2021

Student

Aissata Diallo (Undergraduate) & Sheng Kai Liao (Graduate)

Date:10/6/2021

Paper:

Transformers: “The End of History” for Natural Language Processing?

Critique: (Eddy)

According to the Chernyavskiy et. al, this paper provides an experimental conclusion about the revolutionary NLP (Natural Language Processing) model‘s (BERT) and its variants’ limitations in the NLP domain. The authors hope that by referring to this paper, the NLP model developers could be benefited according to their discussions and opinions about the different limitations in BERT-style models in order to design the next generation of deep NLP architectures. Indeed, based on other research, BERT has several limitations, such as insensitivity in negation and misspellings, and it could be attacked by adversarial contents easily.

According to this paper, they mainly focus on testing the BERT-style model, such as RoBERTa, ALBERT, and XLNet, within two general types of tasks, segmentation and segment labeling, and implement with four different datasets. The major tasks are propaganda detection and keyphrase extraction which both have an identification task for segmentation job and multi-class classification tasks for segment labeling job. In addition, they use hyperparameter fine-tuning to optimize the parameter settings of each model, which can minimize the possibility of parameter settings causing model gaps. Afterward, they discuss the technology-wise issue between the models in order to resolve the issue or have a discussion with it.

This paper tries out several methods into the BERT-style models for improving their performance in the two tasks. For example, in the experiment, they observed that BERT-style models can not explicitly count the length of the character/word/subword since it lacks good sensitivity about the opening/closure mark, such as the quotation marks. Yet, the length features are important for NLP processing in statistic-wise. To solve this issue, the author manually embeds the length of content at the closing of the inputs. Moreover, due to some propaganda techniques often being described as the same word, the author trained a Gazetteer as an external entity list to catch the use frequency of the closure tokens represent the length of span embedding. Furthermore, Conditional Random Fields (CRF) has been used as an extra layer for models in order to observe the relationship between different tags. The purpose is to help the model not only focus on the association between the input tokens but also the relationship between output labels. Also, post-processing techniques are frequently applied in the models to solve certain issues they encountered such as, miss detection of punctuation and quotation marks, over-fitting, repetition of spans and multi-label case in the same span. After the testing, it turned out that RoBERTa has the best performance in propaganda detection tasks and XLNet is the best for keyphrase extraction tasks.

Regarding their experiment result, CRF is helpful for both identification tasks in propaganda detection and keyphase extraction. For sequence classification, basically all methods

benefit the baseline and RoBERTa. Especially the method that improves the expressiveness most is repetition post-processing which at least increases the F1-score 3% for each model case.

For their conclusion, authors not only find out some limitations of BERT-style models but also provide their own solutions or opinions to address those limitations which I think would be helpful for other researchers to have a good understanding of attributes of BERT architecture in order to design next generation NLP architecture. For the future, I think BERT and its variant model can be explored with other NLP problems in order to find out the potential vulnerabilities of the models.

Paper:

Diversity-Based Generalization for Unsupervised Text Classification under Domain Shift

Critique: (Eddy)

According to Krishnan et. al, this paper presents an unsupervised learning model which concentrates on cross-domain adaptability in subjective text classification problems that mainly focus on finding the best balance between the unlabeled target data and labeled source data. The unsupervised learning learns the data without any label, unlike supervised learning which requires each input embedding with a label. However, in this case, unsupervised learning required more samples in order to find the trend of data based on the objective function. So typically, unsupervised learning relies on larger dataset and higher computation power equipment compared to supervised learning. The basic concept of adaptation is that a model be trained in one domain, like street name, but it can be implemented into other domains, for example, address. To achieve that, this paper constructs a novel unsupervised learning model which can be trained without target data, lower computation cost for training and deliver better attendance word for downstream tasks.

In this paper, they compared their models with several strong baseline, such as DANN, DAmSDA, AMN/P-net and HATN. For the dataset, they used “Benchmark Dataset: Amazon Reviews” which contains reviews cross various item domain such as Book, DVD and Electronics on Amazon, and “Crisis Dataset (Tweets)” which created by themselves and contains the tweet posts contents which during the three hurricanes’ crisis.

Their model has three components, BiLSTM (Bidirectional Long Short-term memory) which is similar to traditional LSTM but considers the information from past and future, MHA (Multi-Head Attention) and MHAD (Multi-Head Attention with Diversity) both record the multiple attention heads of the contents in an ATT block but the MHAD added a diversity layer for restraining the attention heads learning respectively, and finally the method call Tri-training which basically train three model at the same time yet the model 1 and the model 2 are diversity constrained and yet the model 3 don’t have such constrained but it is fed with pseudo-label data by the model 1 and model2.

The result reflects that their model scores are close to the strongest baseline (HATN) in “Benchmark Dataset: Amazon Reviews" dataset. And yet, their model was trained without unlabel target data which sometimes may have a huge impact on unsupervised models. At the same time, Their model also got a best score in “Crisis Dataset (Tweets)”. In the computation consuming aspect, the training time of their model is less 2 hours than the best baseline which is pretty impressive.

To conclude, the paper delivers a novel unsupervised model which not only has good adaptability in sufficient diversity but also has lower computation cost compared to the state-of-the-art models for subjective text classification problems. For the future, I think the

model should be tested by other domain shifting tasks so that we can have more comprehensive results for its performance.

Paper:

Scalable Backdoor Detection in Neural Networks

Critique: (Aissata)

The purpose of this research was to find a novel way to protect and optimize deep learning models from Trojan vulnerabilities. Current deep learning models are subject to Trojan attacks because the hackers install a malicious backdoor that causes the resultant model to misidentify samples that have been injected with "a small trigger patch" which in turn puts the deep learning model at risk. In order to keep the deep learning model protected, strong detection performance is necessary, but currently it is expensive and difficult to have good detection performance. In this paper the authors Harikuma et.al were able to create a method that was both able to detect Trojan attacks through the use of reverse engineering of the current model.

In order to secure and optimize the deep learning model Harikuma et.al focused on two main improvements: first they wanted to produce an optimized detection method that has limited computational complexity for the number of growing labels. Second, that requires less information like the scoring for screening when a trigger patch is used by a hacker. Through their observation they were able to find that making prediction vectors for all similar images was a more effective way of stopping detection by Trojan attackers, than classifying all the images to single labels since they do not have that much understanding of the Trojan label. Also, it would have taken longer to do single classification of the labels because they would have to set all the labels as "target labels" and then perform the trigger optimization.

For the results the authors used two testing methods. They classified one group as negative and the other group as positive, assigning each group to 50 trained classes. They tested 50 Pure Models (PM) as the negative group and 50 Trojan models (TM) as the positive group. The Trojan model detection was achieved through a Trojan Scanning procedure which was based on calculating entropy. By computing the entropy score and determining the class distribution of the "recovered trigger" the researchers were able to create a model capable of detecting Trojan in multiple sample labels. Furthermore, Harikuma et.al were able to prove that it is possible to test an upper bound on the scores of the models that were impacted based on the number of class labels and the effectiveness of the Trojan patch they were being tested on.

Overall, Harikuma et.al were able to create a Trojan detection Scanner which is capable of detecting the Trojan model. They were able to achieve all of this through the computation of entropy score, and F1-score, which is inexpensive when compared to other methods that are involved in separation of Trojaned models from pure models. Their research is an effective way to optimize and secure the deep learning model without having to negotiate many resources. The only improvement to this research is that the Trojan Scanner (STS) takes two hours to detect Trojans instead of its counterpart NIC which takes about nine hours. While STS is faster than NIC we still need to work on decreasing the amount of time it takes for the Scanner to detect

Trojans so that in the future as hackers are advancing their methods we can still be faster in detecting Trojans.