Final Project Report - Natural Language Processing

Individual Submission

Abstract

Question answering task is one of the most important areas in Natural Language Processing (NLP). In the class, we have explored some of the most popular methodologies to perform question answering task including Masked Language Model (Bert), Bart, GPT and some other. The main objective of this final project is to use some of the techniques covered in class to improve upon a pretrained baseline model. The overall structure of the baseline model includes encoding step for both passage and question and then computing logits of starting and ending index by using $p_i^T * W^{start} * q$ and $p_i^T * W^{end} * q$ respectively. The performance of the baseline model is exact match : $48.93, F_1 : 61.29$

1 Introduction

Implementing pretrained model is one of the approaches to improve the baseline model. Training a model from scratch to achieve a good performance on SQUAD dataset can be very computationally expensive, therefore using pretrained model is much preferable. For this project, a pretrained DistilBERT model from Hugging Face repository was used and it has shown a significant improvement over base model on the dev set. DistilBERT uncased model was trained on Squad 1.1 dataset which is identical to what being used to train the baseline model. There are some processes to make this model compatible with what provided in the starter code which involved retokenizing passages and questions, reindexing the logits output and calculating the start and end positions for answers. BERT base model was also explored in this project, however the inference time using BERT base model was very expensive. Therefore, DistilBERT was picked as a final model.

2 Model Selection

When analyzing answers' predictions from the base model, it seemed that the base model usually failed on questions that required a rather deep understanding of a particular subject. Interestingly enough, a good chunk of fail predictions from the baseline model involving questions in Physics and Historical subjects. To answer those questions, it is necessary to capture the complex relationship between contexts and words within a passage and this is where RNN is somewhat limited in how much information it can capture.

Therefore, it makes logical sense to improve the baseline model by using some other model architectures that can better understand these complexities. BERT is one of the options that would potentially lead to a better performance model than the baseline model.

Another observation is that length of a passage doesn't affect predictions of the baseline model.

The table below is a percentile describe of cases when the baseline model's predictions match with gold labels *Length of questions in matched predictions.*

Length of questions in matched predictions is a describe when the baseline model fails to predict gold labels.

	Length of questions in matched predictions
count	13874.00
mean	10.09
std	3.39
min	3.00
25%	8.00
50%	10.00
75%	12.00
max	33.00

	Length of questions in unmatched predictions
count	20639.00
mean	10.35
std	3.70
min	3.00
25%	8.00
50%	10.00
75%	12.00
max	31.00

3 Model Architecture

The main reason to choose DistilBERT over BERT is that DistilBERT can do inference much faster than BERT. From experimentation performed on the dev set, DistilBERT took an average around 15 seconds to do inference using batch size of 64 whereas BERT took around 180 seconds to do inference on the same batch. Therefore, DistilBERT is a more appropriate approach to implement.

DistilBERT is a transformer model developed using transformer architecture instead of RNN or LSTM. Transformer is a deep learning model that utilizes attention layers to calculate the weightings between inputs and outputs and self attention method to attend current word to other contexts. RNN and LSTM are only looking at contexts that are close enough to a word but not contexts that further away. This self-attention mechanism helps the model to only pertain information which closely related to the word so it is able to contextualize embeddings. There can be multiple self-attention heads in transformers.

According to (Sanh et al., 2019) DistilBERT was pretrained on BookCorpus and it had 3 objectives when training which are distillation loss, masked language modeling and cosine embedding loss. Distillation loss is basically the loss between output logits from DistilBERT vs. BERT. Masked language modeling loss is self explanatory. Cosine embedding loss is the loss between hidden states of BERT and hidden states of DistilBERT and this to ensure DistilBERT has similar or close hidden states to BERT base model.

4 Implementation Procedure

There is an additional option, “bert” added in *_select_model*. This option if chosen will get DistilBERT model and tokenizer from Hugging Face using “torch.hub.load”.

Creating batch from dataset in *QADataset* was modified to create batch data from DistilBERT’s tokenizer.

A customized process was created to get DistilBERT working correctly within the starter code and the original vocabulary. Each batch from the dataset was going through DistilBERT’s tokenizer which produced encoded vectors and attention mask vectors. These two were fed into DistilBERT model to get logit outputs. A customized mapping was added to map answer to the corresponding passage index and question id. This process is called within *write_predictions* to get the prediction from DistilBert model.

5 Model Performance and Results

One of the biggest drawbacks of using any pretrained BERT models is that these models are very computationally expensive when doing inference. However, DistilBERT does inference much faster than BERT. From experimental results from this project, DistilBERT is around 10 times faster than large BERT. It took around 45 minutes for DistilBERT to run inference on *squad dev*.

The performance of DistilBERT shows a significant improvement over the baseline model which has $EM : 48.93, F_1 : 61.29$ on *squad dev*. Whereas DistilBERT is able to achieve $EM : 70.56, F_1 : 80.13$ on the same dataset.

It clears that BERT model is very powerful particularly in question answering task since it uses transformer layers in its architecture which has many advantages over RNN/LSTM encoding layer.

6 Bias

Since DistilBERT is a fine tuned model from BERT. It automatically inherits some bias from its parent model. (Sanh et al., 2019)

One example is that DistilBERT tends to infer occupations based on genders (gender bias), therefore it can fail to answer some questions related to the above issue. Racial bias is also a great concern in BERT base model. (Sanh et al., 2019)

7 Potential Improvement

DistilBERT can be fine tuned further to increase its performance in other domains including BioASQ (Tsai et al., 2019) and NewsQA (Trischler et al., 2016) by implementing transfer learning. Data augmentation can also be implemented to increase the robustness of the model when encountering adversarial settings.

8 Conclusions

Based on the results coming out of some research and experimentation done in this project, transformers model shows a significant improvement over the baseline model. This proves that to better perform on question answering task, the model needs to capture the complex relationships between words and contexts within passages in order to accurately determine word spans for answers.

DistilBERT, a fined tuned pretrained transformer model, is able to achieve $EM : 70.56, F_1 : 80.13$ on SQuAD 1.1 dataset which results in a large improvement from the baseline model : $48.93, F_1 : 61.29$

The baseline model has one advantage over DistilBERT which is it's inference time. The baseline model is much faster than DistilBERT model when doing inference process. However considering the boost in the baseline model's performance when using DistilBERT, the tradeoff between speed and performance is very worthwhile. DistilBERT takes around 45 minutes to do inference on the dev set which is very reasonable.

References

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset.

Yi-Ting Tsai, Min-Chu Yang, and Han-Yu Chen. 2019. Adversarial attack on sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 233–240, Florence, Italy. Association for Computational Linguistics.