**DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter**

Many big tech companies and other major companies are using BERT for a long time and they created and developed other things such as T5 and GPT3 that were based on BERT. Also, the merit of BERT is that we can customize the BERT for our own purposes. I think this is the reason why many companies and organizations are using BERT. Although BERT is a great model, it required a huge amount of good quality data and time consumption. From the BERT, time consumption is very important to increase the prediction. Based on my research, I found that developers or engineers spend more than a day operating BERT. Because of these issues, the concept and model of distilBERT were developed and this is the content that I read about.

Until this report, most studies were focused on using distillation for specific task models. This report used the knowledge distillation that is during the pre-training phase and the result was kept the 97% of language understanding capabilities when reduced 40% of BERT’s size.

In the field of NLP, large-scale pre-trained language models were the trends and became the basic tool for many tasks. Although the performance was getting improved, sometimes need to have a huge number of parameters to have a better and higher quality of prediction or performance. However, these brought the problem of increased cost since the size of the model is getting bigger. And, distilBERT brought the better size of model with similar quality of performance.

There will be a teacher network and student network. Basically, the teacher network is the well-prepared educated knowledge, and the student network is the network that wants to receive this information. On top of that, there is a compression technique that is the student network is performing like the Teacher network and we called this as Knowledge distillation.

The system and procedure of distilBERT will be divided into student and teacher networks. As student network in distilBERT has the same general architecture with BERT which means that token-type embedding and pooler are removed and the number of layers is reduced by double. Once the optimization is important in the pre-training process in finding initialization, the teacher network’s one of the layer in two will be initialized.

DistilBERT only had 3% of performance decreased compared with BERT performance. They started to test the model in experiments with General Language Understanding.

From the Downstream task, they tested the ability of performance for distilBERT under the environment of classification test (IMBd sentiment classification), and question answering test (SQuAD). From the task, distilBERT’s performance is only 0.6% lower than BERT even though distilBERT is 40% smaller size than BERT. On top of that, in the question-answering task, distilBERT were only 3.9 fall behind of BERT.

Once they found the result in performance rate, and size, they wanted to test the inference speed. They tested the speed of each model’s number of parameters that do a full pass and compared between the number of parameters and inference time. As the result of the speed test, even though, distilBERT had 40% fewer parameters than BERT, had 60% faster speed than BERT.

As the title of the report, distilBERT had 40% smaller, 60% faster, and kept 97% of the performance compared to BERT. As I learned from the class, the size of models and data and training volume are getting bigger and bigger nowadays which means that it will take a longer time to do it. From this situation, I think distilBERT is a great model that could be the option for people or organizations that did not have enough environment that training big size of data, model and volume.