CHAPTER 7

Future of BERT Models

The topics we have covered thus far deal with the architecture and application of the BERT model. The BERT model has not only affected the ML domain, but other fields like content marketing as well. Now let's discuss the development and future possibilities of BERT.

Future Capabilities

Transformer-based ML models like BERT have proven to be successful for state-of-the-art natural processing tasks. BERT, which is a large-scale model, remains one of the most popular language models that delivers state-of-the-art accuracy.

The BERT model has also been used by the Google search team to improve the query understanding capabilities of Google Search. As BERT is a bidirectional model, it is able to understand the context of a word by looking at the surrounding words. BERT is particularly helpful to capture the intent behind search queries.

Ever since its release, the BERT model has influenced the development of various models that are based on BERT. It has to be credited for the introduction of models that not only incorporate its name, but also its core architecture ideas. The variants of BERT are able to successfully beat

N. Sabharwal and A. Agrawal, *Hands-on Question Answering Systems with BERT*,
https://doi.org/10.1007/978-1-4842-6664-9_7

records across a wide array of NLP tasks like sentiment analysis, document classification, question answering, and more.

Here are a few of the models that are based on BERT.

- There are models that pertain to an application or domain-specific corpus. BioBERT is one such model that is trained on biomedical text. Other examples are SciBERT and Clinical BERT. Training on a domain-specific corpus has turned out to be useful and results in better performance when fine-tuning is done on downstream NLP tasks in contrast to fine-tuning BERT, which is trained on BookCorpus and Wikipedia.
- The ERNIE model incorporates knowledge into pretraining and uses a knowledge graph to mask entities and phrases. It is pretrained on a large corpus while taking the knowledge graph into consideration during input.
- The TransBERT model is used for a story ending prediction task that uses a three-stage unsupervised training approach. This is then followed by two supervised steps.
- For making medical recommendations, G-BERT basically combines the power of graph neural networks and BERT. This model is used for medical code recommendations and representations. Encoding of medical codes with hierarchical representations in G-BERT is done with the help of graph neural networks.
- In addition to pretrained models there are also fine-tuned models like DocBERT (document classification) and PatentBERT (patent classification). These models

are fine-tuned for specific tasks. These pretrained BERT-based models can be fine-tuned with the help of NLP tasks, POS, NER, and so on, to achieve better results.

These models are representative of broad classes of BERT- based models. They depict how the BERT model can further be used in different domains with modifications in pretraining or fine-tuning. BERT hence forms a base for the development of other models that are effective in a wide variety of tasks.

One of the developments that relies on the BERT model is RoBERTa, developed by Facebook, which has proven to be highly efficient on GLUE benchmarking. RoBERTa uses the strategy of BERT to mask the text and the machine learns to predict the hidden text. The training is done on a larger number of mini-batches and learning rates, and the hyperparameters are modified to achieve better results. These changes allowed the RoBERTa model to prove its efficiency on MNLI, QNLI, RTE, STS-B, and RACE tasks, and it also shows considerable improvement on the GLUE benchmark.

RoBERTa uses 160 GB of data for pretraining, which includes unannotated NLP datasets and data scrapped from public news articles called the CC-News dataset. These data, along with training of RoBERTa on a 1024 V100 Tesla GPU, takes a day to complete. This results in better performance of RoBERTa over other available models like BERT, XLNet, Alice, and so on.

BERT is incorporated into Google Search, which results in precise and accurate searches. This will affect the content strategy of many users. The content now has to be more precise so that it can be rated better using search engine optimization. The strategies to design the content have to be improvised.

Abstractive Summarization

ML has come a long way in NLP, and one of these applications is in the field of summarization. The most common form of summarization is extractive summarization, which returns the most important sentences out of the content. The other type is abstractive summarization, which uses new sentences, keeping the important ideas or facts intact.

Content selection is integral to any summarization system. In recent approaches, the importance of separating content selection from summary generation is highly emphasized. There are many ongoing studies that attempt to extract content words and sentences that should be the part of summary and use them to guide the generation of an abstract summary.

A brief sentence can be formed by shortening or rewriting a lengthy text. Encoders and decoders are helpful in this context. Comprehensive summaries can be generated in a similar way, by selecting important sentences and dropping the inessential sentence elements, such as prepositional phrases. A summary can be generated through fusing multiple sentences. Selecting important sentences can be done via many approaches, but handling its large cardinality and identifying the sentence relationship to perform fusion has been a tough job. Previously it has been assumed that similar sentences can be fused together because they carry similar information to be processed.

Because abstractive summarization is difficult to perform, there is a lot of development in this area. BERT also has applications in abstractive summarization. The embeddings of multiple sentences can be generated using a BERT model. To perform this task, a [CLS] token can be inserted before the start of the first sentence. The output embeddings have to be processed through multiple layers, which enables the capture of important features. The BERTSUM model is one example.

Natural Language Generation

Natural language generation (NLG) is one of the more active research areas. It is a subgroup of NLP, along with NLU. The basic task of NLG is to convert some text tokens or data into natural language. The basic approach to achieve this is by predefining the templates for a specific domain and filling the empty slots using NLU techniques.

A more complex approach to this is using language modeling. Language modeling is used to model the natural language using the ways of writing, grammar, syntax, and so on, that are required to learn intrinsic features of the source language. We can then use this language in generating language content against some given input data or text.

The applications in terms of language understanding are not limited to NLP, but also extend to NLG. Open-AI's GPT-2 generates text based on the given words and is one of the state-of-the-art models in NLG. The BERT model tries to attain the same feature using HuggingFace transformers.

Recent developments show that the performance of BERT in the field of NLG is not an optial fit. The reason behind it is that the BERT model was trained on MLM rather than being trained autoregressively. Apart from using MLM, the variations such as shuffled input and random words make the BERT model more generalized. Even after all these variations, BERT lags behind GPT-2 because the BERT model is an encoder representation, whereas GPT-2 is a decoder stack, which helps it create context-rich representations.

Machine Translation

Translation is the idea of translating text from one language to another. Automatic or mechanical translation is probably one of the most challenging brain functions given the fluctuations in human language. Recently, pretraining techniques, such as ELMo, GPT and GPT-2, BERT, the cross-language model (XLM), XLNet, and RoBERTa have attracted a lot of attention in the ML and NLP communities.

A Neural Machine Translation (NMT) model usually consists of an encoder to map an input sequence to hidden representations, and a decoder to decode hidden representations and generate a sentence in the target language. BERT has achieved great success in NLU, and incorporating BERT with NMT for performance improvement might be a good research area.

NMT can be improved by fusing the BERT model and NMT, when BERT is drawn by the sensor and decoder using attention models. Research on open supervised NMT (including sentence-level and text-level translation), semisupervised NMT, and unsupervised NMT demonstrates the effectiveness of this approach.

To accurately predict translation quality, a model trained from scratch would theoretically require a large corpus of natural language source text, translations, and their human-labeled quality scores. Creating these datasets at a sufficient scale to train a neural network model is prohibitively expensive. Therefore, researchers have determined that they can transfer learnings from models trained on correctly translated parallel corpora to the task of identifying whether a translation is correct or not. It is far easier to obtain millions of correctly translated sentences to use to pretrain a model in areas where you don't need a quality score.

For future work, there are many interesting directions. First, we have to learn how to speed up the measurement process. Second, we can use such an algorithm in many applications, such as query in response. How to compress the BERT-fused model into a simplified version is another topic. There are other modern functions that include information about distillation to integrate pretrained models with NMT, which is a test method.

Conclusion

This chapter looked at the ongoing research in BERT and in state-of-the-art NLP tasks. With this we have come to the conclusion of this exciting journey into the world of NLP.