Technology Review: Pre-Trained Bert Models for Language Understanding and Question Answering

Chris Kabat

University of Illinois

CS 410 Text Information Systems

Technology Review: Pre-Trained Bert Models for Language Understanding and Question Answering

Introduction

This document's purpose to review the use of pre-trained BERT models for language understanding and question answering. BERT was introduced in the paper BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin, Chang, Lee, & Toutanova, 1999). BERT stands for Bidirectional Encoder Representations from Transformers and is one of the most popular language models in use today. It has some unique features that we will describe in this documentation along with the applications of BERT and some alternatives.

What is BERT?

BERT is a deep learning, transformer language model created by researchers at Google. Pre-trained language models like BERT are trained on a very large corpus of text. Specifically, BERT is trained on the BooksCorpus (800M words) and English Wikipedia (2,500M words) (Devlin, Chang, Lee, & Toutanova, 1999). The training allows the model to "learn" the language from that corpus of text and be able to predict the relationships between words and sentences. When BERT was being researched, most pre-trained models were unidirectional. This means in the pre-training process, only the previous token can be accounted for during the attention layer of the transformer model. The BERT model uses two mechanisms during pretraining. One is a masked language model (MLM) that allows the model to be bidirectional (i.e. getting context from tokens in both directions). The other is a next sentence prediction task which determines the probability of a second sentence given the first. Using transfer learning techniques and fine-tuning, the BERT models can be applied to many different NLP tasks. It is important to note that the BERT requires the training of as many of 345 million parameters takes a lot of computing

horsepower. Advances in both cloud computing and CPU/GPU processors have allowed models like this to be advanced and were not as practical previously.

Applications of BERT

BERT models can be fined tuned to handle many different tasks. Some examples (Jagtap, 2020) specific to NLP are (but not limited to):

- Sequence classification: Predicting the class of a given sequence.
- Named entity recognition: Identify entities in text and classify them (e.g. Famous People)
- Natural language inference: Determining if a hypothesis is true using NLP.
- Ground common sense inference: Finding the most probable continuation of a sentence.
- Text summarization: Summarizing a document into a few sentences.
- Question answering: Using text within a document to answer a question.

Specific to question answering, the BERT model can take the question and the passage of text containing the answer as single sequence. It then can train a start and end vector that represent the most likely start and end of the answer in that sequence of words (i.e. the start and end vector with the highest probability of being the answer).

Benchmarks and Testing

A key benchmark for measuring the success of these types of models is the Standford Question Answering Dataset (SQuAD) (Rajpurkar, Zhang, Lopyrev, & Liang, 2016). This dataset is a number of question/answer pairs that can be used to test the effectiveness of the models. Below is a figure from the SQuAD 2.0 results published in the BERT : Pre-training of Deep

Bidirectional Transformers for Language Understanding paper (Devlin, Chang, Lee, & Toutanova, 1999) :

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

Figure 1- SQuAD 2.0 Results

As you can see, even back when this paper was published, the model was very close to human parity. If you look at the leaderboards today (The Stanford Question Answering Dataset (rajpurkar.github.io), you can see many models, some based on the BERT model have surpassed human parity.

Alternatives to BERT

There are a number of additional models that use similar approaches to BERT. Some examples are:

- ELMo: Feature based NLP model deep contextualized word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (Peters, et al., 2018).
- OpenAI GPT3: Generative Pre-trained Transformer 3 – it's the third version of the tool to be released. In short, this means that it generates text using algorithms that are pre-trained (Marr, 2020).

- Turing Natural Language Generation (T-NLG) is a 17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks (Corby Rosset, 2020).

All of these models have different techniques for pretraining, and fine tuning for specific applications. Many of these models are more effective in certain applications than others.

Conclusion

The growing field of Natural Language Processing has been heavily influenced by the success of the BERT model. It has influenced a lot of additional research and competing models while also spawning off a number of domain language models as well. As computing power becomes more powerful and these models are able to be trained with a larger and larger corpus of data, there will be even more advancement in this space. These models are being trained with millions of parameters to attempt to learn words, word structures, and word relations like a human does. Some of these applications have not only reached but surpassed human parity.

References

Corby Rosset. (2020, Feb 13). *Turing-NLG: A 17-billion-parameter language model by Microsoft*. Retrieved from Microsoft Research Blog: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (1999). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*

Jagtap, R. (2020, Jun 28). *BERT: Pre-Training of Transformers for Language Understanding - Understanding Transformer-Based Self-Supervised Architectures*. Retrieved from medium.com: https://medium.com/swlh/bert-pre-training-of-transformers-for-language-understanding-5214fba4a9af

Marr, B. (2020, Oct 5). *What Is GPT-3 And Why Is It Revolutionizing Artificial Intelligence?* Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2020/10/05/what-is-gpt-3-and-why-is-it-revolutionizing-artificial-intelligence/?sh=74d789d7481a

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of textt. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383-2392. Retrieved from https://rajpurkar.github.io/SQuAD-explorer/