Albert J Yeh

Ayeh2@illinois.edu

CS 410

Tech Review of BERT(Bidirectional Encoder Representations from Transformers)

BERT was released by google in October 2018 and it has taken the NLP industry by storm. BERT has achieved impressive benchmark performances which are rarer and rarer in machine learning/artificial intelligence. While BERT is humongous and expensive to train, one of the reasons why it was very appealing to many people in the industry was that BERT was published with a pre trained model and pretty much ready to use out the box so to speak. More specifically it is a specific, large transformer masked language model. The importance of the masked language models are one type of contextual word embedding and can be used as input embeddings. But giving BERT all the credit would be partially misleading as BERT is based off of the even bigger idea which BERT is built on, Transformers. Transformers are a fairly new family of neural network architectures, released in a paper by google in June 2017.

Before one can understand BERT, one must understand that it is build on a mountain of ideas. The first is what we feed into BERT. The input representation and the embedding tool it uses, specifically WordPiece embeddings (similar to Word2Vec/GloVe). The word embeddings are just a feature Vector representation of a word. The model will have a set of pre-learned word embeddings. There is a word embedding lookup table, BERT has a vocabulary of 30,000 tokens and each token has 768 features. For each word in a translation, it would attempt to look up the word in the word embedding find the vectors, so now instead of a group of strings, BERT has a group of numbers it can work with. The features are compared to each other in 768 dimensions and so similar words will have lower scores. In other words the distance reflects word similarity, like I use a “laptop” for work or I use a “computer for work”. They are fed into the neural network, we can expect both outputs to be close to each other because the paradigmatic words have similar embeddings vectors. BERT is pretrained and so it has a technique of breaking down the words to smaller sub words if the word is unknown, so every word can be processed.

With the word embeddings, the Transformers are now able to do the heavy lifting. The original transformer model would take an encoder and take the embeddings of the input sequence and pass it to the decoder and that would output to a sequence. BERT has a slightly different twist, it takes multiple encoders (12 or 24) you take the last layer or last network of the encoder and pass it into a task specific model that you have trained as an input to a new classifier. A common way BERT has been used is to take the raw embeddings that are used as inputs to a new classifier and passed back into the encoders to update the original weights to fine tune it for whatever task is needed. Part of the beauty of BERT is how it can be used, language interpretation has been one of its biggest uses since its introduction, having replaced Google's own translation system built on an older machine learning technique because of BERT's ability to have better "memory" as it really looks at every single word in the sentence at the same time instead of the traditional left to right (again, this goes back to the idea that BERT is bi-directional). This helps a lot with language translation because many languages do not have a similar syntax. The Bidirectional LSTM's go both ways and are able to work with long sentences that previous models did not do very well. Since the introduction of BERT, there has been a lot of newer additions to bert such as RoBERTa, AlBERT, DistilBERT, and for other languages such as AraBERT (Arabic), CamemBERT(French), Ch-RoBERTa (chinese), and BETO(Spanish), to over 100 languages now. Additionally, it can be used as a general-purpose pre-trained model that's powerful for fine tuning specific tasks, such as classification, sentiment analysis, text summarization, question answering, and many more.

What do these all have in common? The need to understand language. So the main idea is to pretrain BERT to understand language, and then fine tune BERT to learn a specific task. The pretraining happens where the model understands what is language and context, and the second step is, how does it solve the problem. Once this is complete, then BERT can understand language, like previously stated, BERT is a language model, and it does it very well.

Lastly, A lite Bert or Albert has been released in September 2019, which has outperformed a lot of the earlier BERT model's, but one unique take is that it has actually reduced parameters, in fact a 89% reduction in parameters compared to the original BERT. However, they have tested that the xlarge or xxlarge versions of BERT have done even better.

The change that Transformers, BERT, AlBERT and its many variations have brought to NLP has shown what is possible as we get closer and closer to computers understanding language.

References:

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/abs/1810.04805

The Illustrated Transformer
http://jalammar.github.io/illustrated-transformer/

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
https://arxiv.org/abs/1909.11942

Attention Is All You Need
https://arxiv.org/abs/1706.03762