CS410 Technology Review

BERT Language Model

Karlin Dye

NetID: karlind2

What is BERT

Language models are key components for applications ranging from speech recognition to information retrieval. In the realm of information retrieval, often a unigram model or bag of words representation is employed. However, there are much more advanced language models that employ deep neural networks that can take word context into account and output embeddings of words in a continuous space. BERT, or Bidirectional Encoder Represantations from Transformers, is one of these deep learning language models that is able to include context from both the left and right sides of words. BERT was developed by researchers at Google AI Language. More, specifically the BERT language model is pre-trained neural network that through transfer learning can be refined for a particular corpus of new documents. Since the output of word embeddings from the BERT language model includes contextual understanding the embeddings perform well in the realm of text classification.

How Does BERT Work

The BERT neural network architecture consists of 24 layers of transformer blocks, 16 attention heads, and 340 million parameters. The BERT model employs a preprocessing step for input text that can takes into account a pair of sequential sentences and includes information about the word positions, sentence positions, and the words themselves. This preprocessing step for text input is an important component of the model training process.

Once input is properly preprocessed the BERT model uses some novel techniques as part of the training process. One challenge involved is how to include both left and right context in a final output embedding. The BERT researchers solved this problem by utilizing a technique called Masked LM (MLM) which masks 15% of input tokens at random and then uses a classification layer to predict the masked token.

An additional strategy that the BERT model uses to help encode context is a technique called Next Sentence Prediction (NSP). The process involves an additional classification layer that is given pairs of sentences and is trained to predict whether one of the sentences in the pair is subsequent from the first. This technique is reliant of on the preprocessing of text input that provides information about sentence positions. During training a random sentence from the corpus is paired with an input sentence 50% of the time and the other 50% of the time is paired with the actual subsequent sentence pair from the input.

BERT has been pre-trained on an extremely large corpus of text and is often fine-tuned for a specific application. BERT can be used in a wide range of applications including sentiment analysis, next sentence classification, question answering tasks, named entity recognition (NER), and many more. The use of BERT for many of these tasks have achieved state-of-the-art results.

BERT for Text Classification

There seems to be many options for using BERT for text classification tasks. One common approach seems to be the use of the Pytorch-Transformers library from HuggingFace. This library contains pre-trained BERT models of different sizes and for different languages and use-cases. The full documentation for how to use the library is available at:

https://huggingface.co/transformers/model_doc/bert.html

Conclusion

BERT has achieved state-of-the-art results on benchmarks for many natural language tasks. With many new open source libraries available to use BERT it will likely become widely used in practice. When faced with a text classification, task practitioners should explore the use of deep learning language models like BERT as part of the model building process.