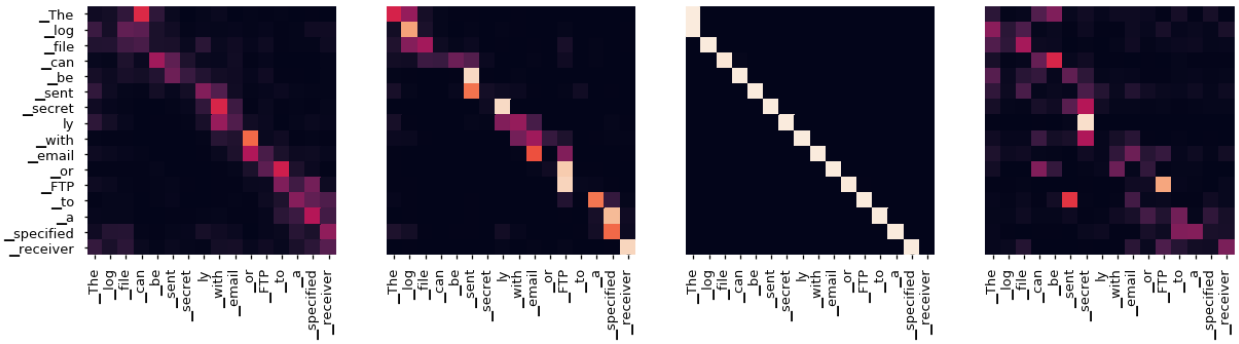
The Legacy of BERT

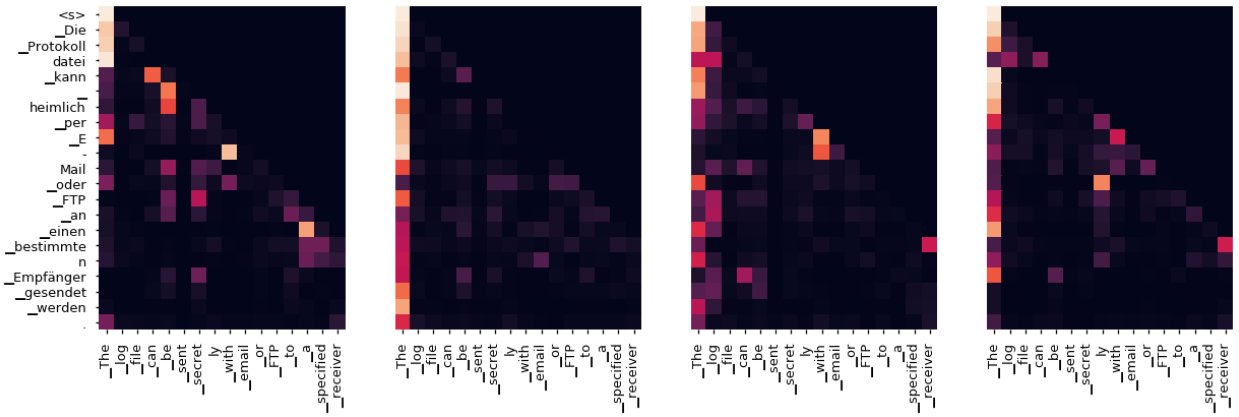
Rodrigo Mendoza
rm36@illinois.edu

BERT, which stands for "Bidirectional Encoder Representations from Transformers" is a very important research paper for the NLP community because it introduced two powerful ideas. Firstly, it demonstrates how it's possible to use a bidirectional training for encoding language into a transformer's encoder and why it's important. Secondly, it provides a very useful pre-trained representation that can be fine-tuned for language-related tasks. Besides those two contributions it also improved the state of the art for some NLP tasks, but despite it's difficulty, it's not as impressive as the legacy it left behind with the key ideas presented in the paper.

The model introduced consists of Transformer blocks, stacked on top of another. Since Transformers (Vaswani et al., 2017) are such an important building block for BERT, here's a short explanation of the intuition behind them. Transformers use the idea of attention (or self-attention) on multiple identical blocks stacked on top of each other. In the case of the original paper, there are 6 layers of encoders which consist of 1. a positional encoding to identify the index or position of each word in the input or output phrase, 2. multi-head attention modules, (the important part of this model) and 3. feed-forward fully connected network and layer normalizations. The attention modules encode the meaning of each word in a sentence relative to the other words, and because the meaning behind words is abstract and intricate, it can be complicated to understand the meaning encoded in these attention modules. For example, in the sentence "The log file can be sent secretly with email or FTP to a specified receiver", the images below show a visualization explained here for the word-to-word weights learned in the second layer. Notice how the third attention head focuses each word with the next word that appears in the sentence.



In the Transformers paper, there are another 6 layers of decoders which are very similar to the encoders, and use it for tasks such as translation. For the same sentence seen above, here's an example learned word-to-word matrix for its German equivalent in the last decoder layer:



In this example, the words "kann" and "can" are deeply associated (shown with a bright pixel), and the whole model learns the association by going through multiple training examples. In the case of BERT, the decoders aren't used directly, but rather the output of the encoders is used directly to train on two tasks: 1. predicting whether the second sentence in the input is indeed the sentence that comes after the first or a random sentence instead, and 2. predicting the words that are masked as part of the input. Thus, the learned word-to-word association matrices can be more difficult to interpret in deep layers.

The task of predicting a masked word is only possible because there is a new word token [MASK] which is randomly replaced on 15% of the words in the training data. This way it's easy to use unsupervised data to train it with: All of Wikipedia pages in English were used, and the training process only had to mask a random subset of words for the model to learn the association between them. The mask allows words in the sentence to have association weights to words both before and after itself, because associating with [MASK] provides no semantic value. This is the first important contribution of BERT, which allows bidirectionality.

Once the model is trained on the two simple tasks mentioned above, it can be trained relatively inexpensively on specific tasks. For comparison, the first unsupervised tasks the model is trained on datasets totalling more than 3 billion words, which lasted 4 full days on 64 TPUs, but all the latter supervised fine-tuning task results in the paper can be replicated with at most 1 hour of training on a single processor. This opened the door to new applications that could use the pre-trained version of BERT and fine-tune it for domain-specific applications.

From those applications which fine-tuned the pre-trained version for an application a couple of examples are BioBERT (Lee et al., 2019) which looked into biomedical text mining, and improved 12.24% on MRR for biomedical question answering, and ClinicalBERT (Huang et al., 2020) which outperformed other NLP techniques to predict from text the 30-day readmission of clinical patients.

There are also models that incorporated different kinds of data other than text along with BERT to get improved performance, for example ERNIE (Zhang et al., 2019) incorporated knowledge graphs into a fused aggregator and videoBERT (Sun et al. 2019) outperformed the

state-of-the-art on video captioning by using a joint video-linguistic model to learn features without explicit supervision.

In conclusion, BERT is a model that pushed the state-of-the-art forward by allowing Transformers to learn bidirectional associations with words, and pre-training the model with a general, unsupervised large dataset. Multiple research papers have used BERT as a starting point (and even more industry applications have done so), proving the importance of this NLP staple.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. In CHIL ’20.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. 2019. ERNIE: Enhanced Representation through Knowledge Integration.