TECHNOLOGY REVIEW - BERT

Erik Larsen – erikl2@illinois.edu

BERT, an acronym for Bidirectional Encoder Representations from Transformers [1], a statistical language model, has been one of the most influential advances in Natural Language Processing (NLP) in recent years. First conceived in 2018, it was applied to Google search in 2019, and by October 2020, BERT was used for nearly all English-language queries [2]. BERT and its derivatives have achieved state-of-the art results in many NLP tasks including question answering, and continue to be highly influential [3].

The key elements that contribute to BERT's advances include pretrained contextual embeddings in combination with a deep bidirectional neural network. In statistical NLP, words and documents must be represented numerically. Early attempts used a simple one-hot encoding to indicate the presence of words in a document without regard to their order or context [4]. Word embeddings, such as those obtained by GloVE [5] were able to map words based on their semantic similarity, improving results. Contextual embeddings such as BERT and ELMo [6], in contrast with models such as Word2Vec, can distinguish between different meanings of the same word in different contexts. These contextual embeddings are generated by "pre-training" on large corpii without labels. BERT's initial contextual embeddings were generated with unlabeled data from the 2.5 billion word English Wikipedia corpus along with the 800 million word BookCorpus.

BERT's neural network architecture was the result of several advances. The sequential nature of text data led to the use of Recurrent Neural Networks for language models, which process data in a series of time steps [7] and the use of gating mechanisms [8], which allow the network to "remember" longer sequences, and bidirectional models [9], which learn from the "past" and "future" of sequences. The seminal paper "Attention is All You Need" introduced the Transformer model, which are able to direct attention to important language patterns without relying on sequential processing [10]. BERT's transformer architecture was thus able to be trained on a very large dataset efficiently because processing could be parallelized.

Upon its publication, BERT ACHIEVED state-of-the-art performance on several natural language understanding benchmarks. One of BERT's strengths is its ability to be leveraged through transfer learning and "fine-tuned" to specific domains such as financial services [11] and biomedical research [12].

Further improvements of BERT include RoBERTa [13], which optimized hyperparameters and training data to exceed BERT's performance and set a new state-of-the-art.

OpenAI's GPT-3 language model, using a similar transformer architecture, became the largest language model ever, trained on hundreds of billions of words and achieving state-of-the-art results in generating text [14].

Google's T5, an extension of BERT's transformer model, demonstrated the ability to achieve multiple tasks with single model [15]. In the next paradigm of

Google search, its Multitask Unified Model [16] will be able to answer even more complex questions by applying the transformer model to multiple aspects of a query.

References:

1. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
2. https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193
3. Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. "A primer in bertology: What we know about how bert works." *Transactions of the Association for Computational Linguistics* 8 (2020): 842-866.
4. One-hot
5. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
6. Peters, Matthew E., et al. "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365* (2018).
7. Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *nature* 323.6088 (1986): 533-536.
8. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
9. Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45.11 (1997): 2673-2681.
10. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
11. Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." *arXiv preprint arXiv:1908.10063* (2019).
12. Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.
13. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).
14. Brown, Tom B., et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).
15. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019).
16. Metzler, Donald, et al. "Rethinking search: making domain experts out of dilettantes." *ACM SIGIR Forum*. Vol. 55. No. 1. New York, NY, USA: ACM, 2021.