Anthony Ghabour
ghabour2@illinois.edu

An Overview of BERT: Bidirectional Encode Representations from Transformers

Recent advancements in computational performance and storage capacity have reinvigorated research and development efforts in the realm of artificial intelligence and machine learning. In the wake of these developments, the sub-domains of Natural Language Processing (NLP) and Natural Language Understanding (NLU) have arguably benefitted most of all.

At their core, machine learning algorithms aim to recognize patterns in the data on which they are trained in the hope that such patterns will generalize to unseen data. By effectively identifying patterns found in sufficiently large amounts of data, and applying those patterns to new data, ML practitioners can often improve outcomes by facilitating the execution of many routine and/or repetitive tasks. Importantly, however, the language data associated with NLP and NLU present unique challenges. In particular, these data are fundamentally unstructured in their representation and involve a great deal variability and ambiguity.

Although humans routinely (and unwittingly) rely on context and a considerable amount of prior knowledge to effectively execute even the most trivial language-based tasks, traditional language models have historically struggled to replicate humans' innate ability to infer context and navigate exceedingly ambiguous syntactic and sematic linguistic features. Consequently, traditional ML models have attempted to tackle NLP-related tasks by developing distinct models uniquely designed to accomplish individual tasks including part-of-speech tagging, question answering, sentiment analysis, language translation, and more. In 2018, however, researchers at Google introduced BERT (short for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers) – a large pre-trained transformer-based model that can subsequently be fine-tuned on smaller datasets to perform a wide range of NLP-related tasks.

Until transformers, most language models were trained to iteratively predict the next word in a sequence conditioned on previous elements. Although somewhat effective, these traditional sequence models (known as Recurrent Neural Networks or RNNs) generally struggled to fully capture the relevant context and long-range dependencies needed for most language-related tasks. In applications where the complete input sequence is known at the time of inference (such as machine translation, part-of-speech tagging, or text summarization), a technique known as bidirectional encoding may be employed to improve performance. In contrast to directional models that read text sequentially, bidirectional models allow predictions to be conditioned on context learned not only from preceding elements in a sequence, but also on information inferred from subsequent elements. By leveraging bidirectional encoding, BERT relies on embeddings (i.e representations) that are far richer than those previously used in other models. In particular, rather

Anthony Ghabour
ghabour2@illinois.edu

than relying on a single embedding for a given token, BERT "recognizes" that a token may have multiple connotations and specifies unique context-dependent embeddings for each token's distinct meaning.

In addition to bidirectional encoding, traditional sequence models have also been further enhanced with the introduction "attention" mechanisms – a mathematical formulation capable of identifying the degree to which different elements in a sequence are relevant when attempting to predict the next element. In the seminal paper entitled, "Attention is all you need" researchers at Google proposed the notion of self-attention applied during the encoding phase of training and introduced transformers, which, in contrast to sequence models, provide direct access to all elements of a sequence in lieu of a compressed representation of the entire input. By leveraging non-sequential processing of sequences as a whole rather than token-by-token, transformers such as BERT have proven to be far more effective at capturing long-range contextual information in sequential data than their predecessors.

Interestingly, the major performance gains introduced by BERT were derived not only from the unique combination/application of pre-existing architectures and design cues (i.e., bidirectional encoding and transformers), but also from a novel training paradigm that involved simultaneously training BERT on two distinct tasks – a Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM component of training enabled BERT to encode the meaning of text within a sequence by hiding (masking) a word at random and requiring BERT to "fill in the blank" by relying on context learned from words on either side of the hidden word. With NSP, BERT also learned about relationships across sequences by being asked to determine whether a given sequence follows the previous sequence.

Importantly, BERT's 340 million parameters were trained using semi-supervised learning with text from Wikipedia (approx. 2.5 billion words) and Google's BooksCorpus of roughly 800 million words – a corpus of approximately 3.3 billion words in total. Given the vast amounts of data and the tremendous computing resources required to develop the pre-trained model, the true value of BERT's advent comes from applying transfer learning, a process that fine-tunes and adapts the model to prescribed use-cases based on smaller task-specific datasets. For example, BERT has been fine-tuned to develop a number of domain-specific models including, but not limited to patentBERT (to perform patent classification), bioBERT (for biomedical text mining), SciBERT (for scientific text), and BEREL (trained on a corpus of Rabbinic Hebrew).

Regardless of the domain in which it's applied, BERT has transformed the way language models are both trained and deployed in industry and its impact in the areas of NLP and NLU are undeniable. Despite being an extremely advanced and highly complex language model, the researchers at Google have reminded us that state-of-the art performance can be achieved not only by introducing

Anthony Ghabour
ghabour2@illinois.edu

completely novel architectures and new mathematical formulations, but also by deftly combining pre-existing design patterns with novel training techniques.

References

J. Devlin, M.-W. Chang, et al. *"BERT: Pretraining of deep bidirectional transformers for language understanding"* in Proceedings of ACL, 2018.

Muller, Britney. "*BERT 101 - State of the Art NLP Model Explained"* March 2, 2022. https://huggingface.co/blog/bert-101

Horev, Rani. "*BERT Explained: State of the art language model for NLP"* November 10, 2018. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Montantes, James. "*BERT Explained: A Complete Guide with Theory and Tutorial" April 12, 2021* https://becominghuman.ai/bert-transformers-how-do-they-work-cd44e8e31359