BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Status	Finished
Author	
Publishing/Release Date	
Publisher	
Link	https://arxiv.org/abs/1810.04805
Summary	
Score /5	⭐⭐⭐⭐
Column	
Type	Academic Journal
Original Git Repo	
Git Repo	https://github.com/google-research/bert
Tutorial	http://jalammar.github.io/illustrated-bert/

Abstract

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question

answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

Technique used:

ELMo: make sense of the context

If we're using this GloVe representation, then the word "stick" would be represented by this vector no-matter what the context was. "Wait a minute" said a number of NLP researchers (Peters et. al., 2017 , McCann et. al., 2017, and yet again Peters et. al., 2018 in the ELMo paper), "*stick*" has multiple meanings depending on where it's used. Why not give it an embedding based on the context it's used in – to both capture the word meaning in that context as well as other contextual information?". And so, *contextualized*
word-embeddings were born.

ULM-FiT: Nailing down Transfer Learning in NLP

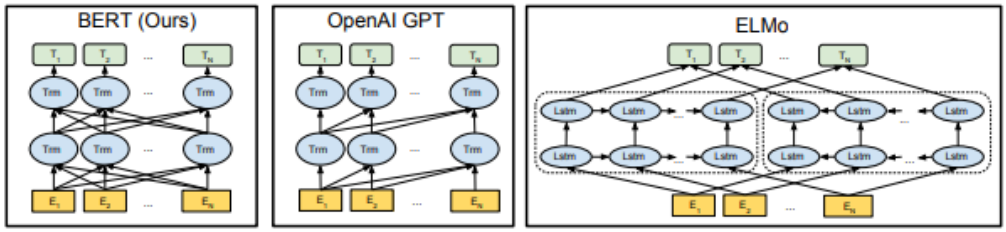
ULM-FiT introduced methods to effectively utilize a lot of what the model learns during pre-training – more than just embeddings, and more than contextualized embeddings. ULM-FiT introduced a language model and a process to effectively fine-tune that language model for various tasks.

The Transformer: Going beyond LSTMs

The release of the Transformer paper and code, and the results it achieved on tasks such as machine translation started to make some in the field think of them as a replacement to LSTMs. This was compounded by the fact that Transformers deal with long-term dependancies better than LSTMs.

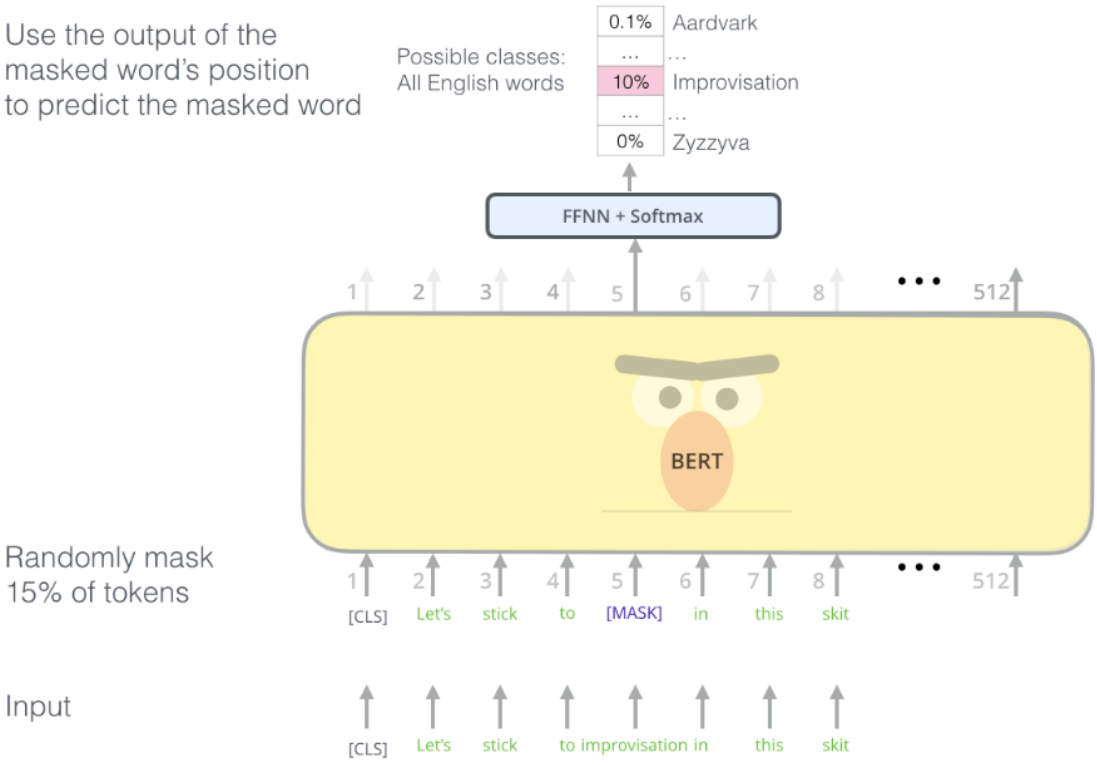
Notes

- As the name suggested, BERT is able to train bi-directional, which considers both the features prior to and after the context to improve the general machine's understanding of data.
- The original paper provides the pre-trained model which allows us to apply the transfer learning method to train our own model.
- This is a momentous development since it enables anyone building a machine learning model involving language processing to use this powerhouse as a readily-available component – saving the time, energy, knowledge, and resources that would have gone to training a language-processing model from scratch.



Two architectures :

- BERT BASE - 12 encoder layers (Transformer blocks). Comparable in size to the OpenAI Transformer in order to compare the performance
- BERT LARGE - 24 encoder layers. A ridiculously huge model which achieved the state-of-the-art results reported in the paper



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.