Paper 16 | BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin et. al – published in Oct 2018

Abstract

1. BERT stands for Bidirectional Encoder Representations from Transformers
2. BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers
3. As a result, pre-trained BERT representations can be fine-tuned with just one additional output layer to create SOTA models for many tasks
4. BERT obtains new SOTA results on eleven NLP tasks, including pushing the GLUE benchmark to 80.4% (7.6% absolute improvement) and SQuAD v1.1 QA test F1 to 93.2, outperforming human performance by 2.0

Introduction

1. Language model pre-training has shown to be effective for improving many NLP tasks
2. There are two existing strategies for applying pre-trained language representations to downstream tasks:
    a. Feature-based
        i. Uses tasks-specific architectures that include pre-trained representations as additional features (ELMo model)
    b. Fine-tuning
        i. Introduces minimal task-specific parameters and is trained on the downstream tasks by simply fine-tuning the pretrained parameters (OpenAI GPT)
3. The major limitations of standard language models are unidirectional and this limits the choice of architectures that can be used during pre-training
4. This paper improves the fine-tuning-based approaches. BERT addresses the unidirectional constraints by proposing a new pre-training objective:
    a. The “masked language model” (MLM)
        i. Randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the marked word based only on its context, therefore allowing the representation to fuse the left and the right context
    b. “Next sentence prediction” task
        i. Pre-trains text-pair representations

BERT

1. BERT’s model is a multi-layer bidirectional Transformer encoder based on Vaswani et al. (2017) – “Attention is all you need”
2. Input Representation
    a. Can represent both a single text sentence or a pair of text sentences in one token sequence
    b. For a given token, its input representation is constructed by summing the corresponding token, segment and position embeddings as shown below:

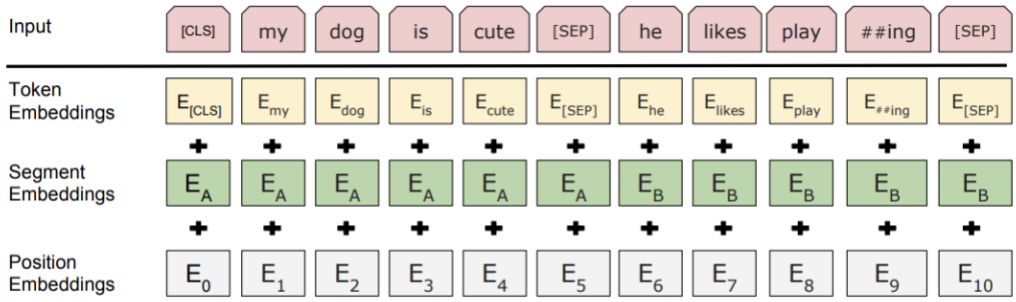

Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

3. Pre-training tasks
    a. Two novel unsupervised prediction tasks:
        i. Masked LM
            - To train a deep bidirectional representation, we mask some percentage of the input tokens at random, and then predicting only those masked tokens
            - In order to mitigate the disadvantage of pre-training and fine-tuning mismatch, we do not always replace "masked" words with the actual token. The training data generator will choose 15% of tokens at random
                a. The transformer encoder does not know which words it will be asked to predict or which have been replaced by random words, forcing it to keep a distributional contextual representation of every input token
            - Second downside of MLM is only 15% of tokens are predicted in each batch → higher training cost (but far better empirical improvements)
        ii. Next Sentence Prediction
            - To train a model to understand sentence relationships, we pre-train a binarized next sentence prediction task that can be generated from any monolingual corpus
            - This is very beneficial for NLP tasks such as QA and NLI
4. Pre-training procedure
    a. Pre-training corpus is the concatenation of BooksCorpus and English Wikipedia (totalling 3300M words). It is important to use document-level corpus rather than a shuffled sentence-level corpus as then we can extract long contiguous sequences
    b. Use gelu (Gaussian Error Linear Units) https://arxiv.org/abs/1606.08415

Conclusion

1. Recent empirical improvements due to transfer learning with language models have demonstrated that rich, unsupervised pre-training is an integral part of many language understanding systems
2. The paper's major contribution is further generalising these findings to deep bidirectional architectures, allowing the same pre-trained model to tackle a broad set of NLP tasks