BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding

Overview:

The paper focuses on the problem of question answering tasks, particularly predicting the answer text span in a passage given a question and an answer-containing passage from Wikipedia. It pioneers the development and evaluation of BERT, a pre-trained Transformer model for natural language processing tasks that successfully addresses a variety of NLP tasks without requiring substantial task-specific architecture modifications. The research utilizes the General Language Understanding Evaluation (GLUE) benchmark, the Stanford Question Answering Dataset (SQuAD v1.1), and the Situations With Adversarial Generations (SWAG) dataset for experiments. Through ablation studies, the paper identifies the importance of the deep bidirectionality of BERT for achieving high performance. BERT has resulted in new state-of-the-art results across eleven natural language processing tasks implying the effectiveness of unsupervised pre-training and the versatility of BERT for both fine-tuning and feature-based methodologies.

Detailed Method:

BERT (Bidirectional Encoder Representations from Transformers) is a model used for natural language processing tasks. It consists of two steps: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data. Then during fine-tuning, the model utilizes the pre-trained parameters on labeled data from specific tasks. Importantly, each task has separate fine-tuned models initiated with the same pre-trained parameters.

BERT's architecture is universal across various tasks and utilizes a multi-layer bidirectional Transformer encoder. Its bidirectional self-attention process allows all input to be related to all output, a good contrast to the GPT Transformer that only utilizes constrained self-attention. BERT uses WordPiece embeddings with a 30000 token vocabulary. Sentences to be run through BERT are represented by summing tokens, segment, and position embeddings, illustrated in Figure 2.

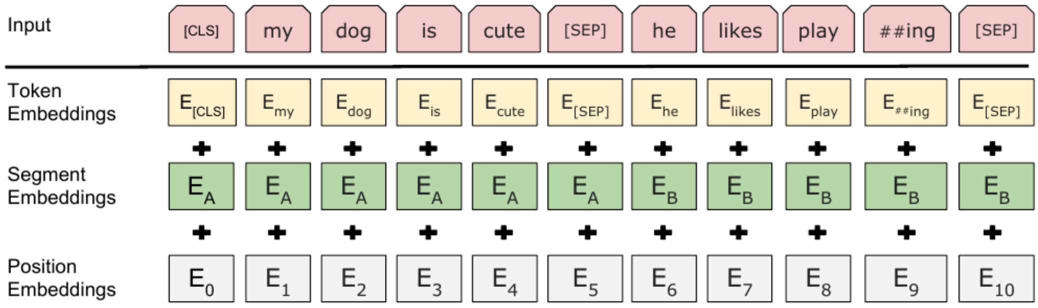

Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmenta- tion embeddings and the position embeddings.

.

During pre-training, BERT uses two unsupervised tasks. First, it masks a fraction of the input tokens at random and predicts those masked tokens, called "masked LM" (MLM). Second, it trains for a next sentence prediction task (NSP), used for understanding the relationship between two sentences. This prediction task is part of the framework you see in Figure 1.

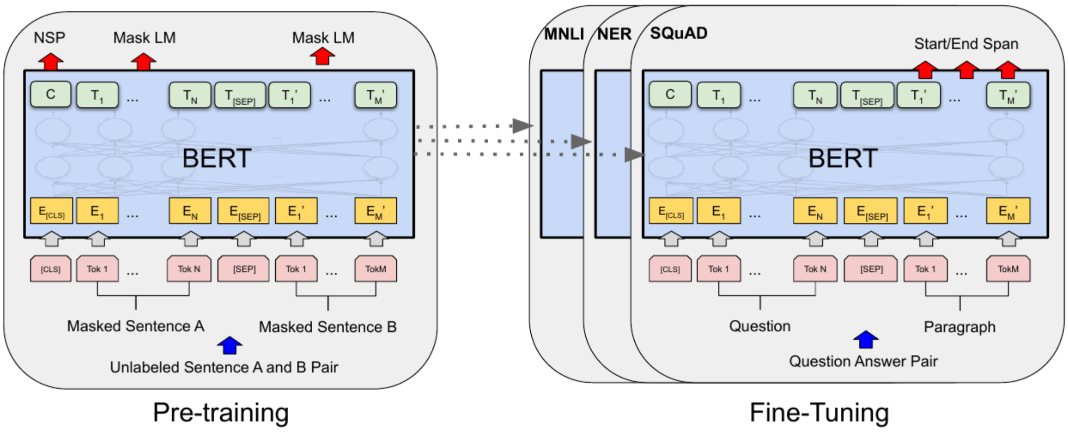

Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architec- tures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special

symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).
.

For fine-tuning, the self-attention mechanism in BERT allows it to model tasks involving single text or text pairs by swapping out the necessary inputs and outputs. The sentence pairs that were used in pre-training can correspond to sentence pairs in downstream tasks like paraphrasing, hypothese-premise pairs in entailment, and question-passage pairs in question answering. This fine-tuning is relatively quick and inexpensive. It can be completed in just hours starting from the same pre-trained model (referenced in Section 4). BERT's design makes it extremely versatile and powerful for a variety of natural language processing applications.