BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018) 46k

Bidirectional Encoder Representations from Transformers

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!

ELMo (unsupervised feature-based approach) based on RNN; BERT based on transformer

GPT (unsupervised fine-tuning approach) is unidirectional, left to right; BERT is bidirectional

Two steps: pre-training + fine-tuning

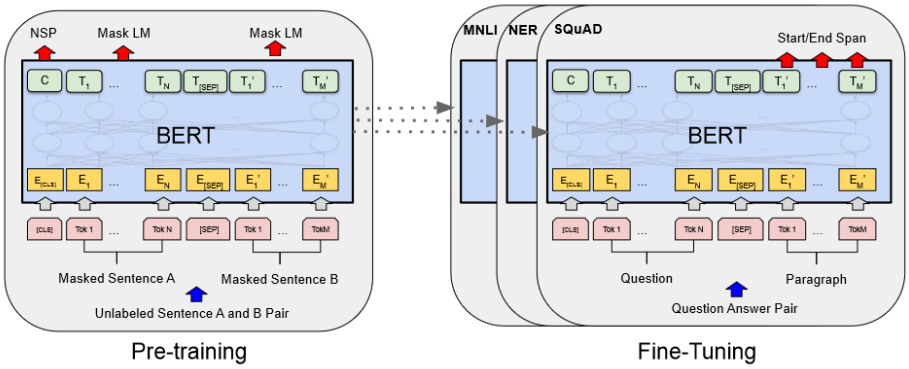

Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are 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).

Parameters: the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A.

Two model sizes:

BERT_BASE (L=12, H=768, A=12, Total Parameters=110M)

BERT_LARGE (L=24, H=1024, A=16, Total Parameters=340M)