CS 410 Text Information System Tech Review

An Introduction to BERT

Zhenchen Yu (zy23)

October 26th, 2020

Introduction

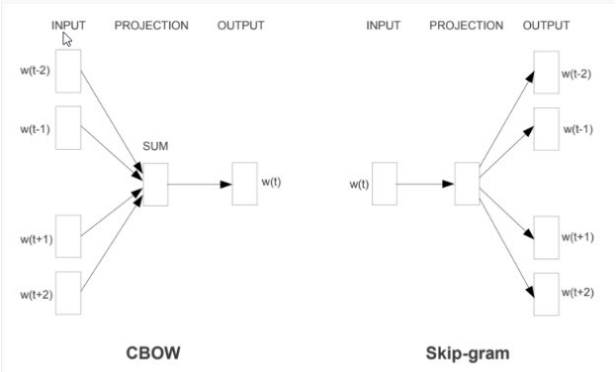
BERT (Bidirectional Encoder Representations from Transformers) is a deep learning language model that was created and introduced by researchers from Google in 2018. When released, it caused an earthquake in the NLP research areas because it beat eleven NLP tasks including GLUE (General Language Understanding Evaluation), SQuAD (Stanford Question Answering Dataset), SWAG (Situations With Adversarial Generations), and others.[1]

BERT is not out of nowhere. It was established based on the previous state of the art models. There are two major strategies that are used in BERT for applying pre-trained language representations to downstream tasks: feature-based and fine-tuning.

Feature-based Approach and Word Embedding

To understand BERT, we need to first understand what is word embedding. In short, word embedding is a from text to text where words that have the same meaning have a similar representation. For example, we have a list of words, [woman, girl, man, boy]. We can map those four words into a two-dimensional vector space [gender, age]. Hence, the procedure of mapping is the word embedding. Of course, it is not an easy task when we have millions of words to map into a lower dimensional space. One solution is to use autoencoder network to complete the mapping. "An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal "noise".[2]

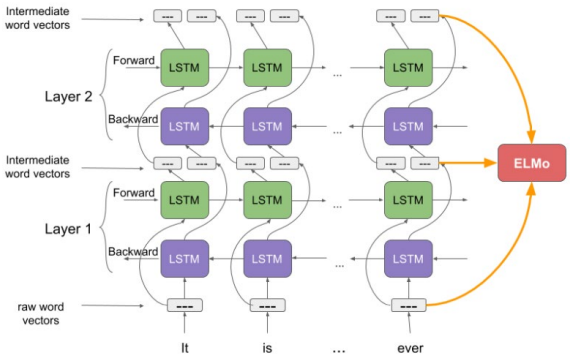
One of the most popular word embedding method is called "Word2Vec". Furthermore, there are two different learning models under word embedding, Continuous Bag-of-Words (CBOW) and Continuous Skip-Gram Model. CBOW uses context to predict the current word while continuous skip-gram uses the current word to predict the surrounding words.



Word2Vec Training Models
Taken from "Efficient Estimation of Word Representations in Vector Space", 2013

[1] https://en.wikipedia.org/wiki/BERT_(language_model)
[2] https://en.wikipedia.org/wiki/Autoencoder

ELMo and its predecessor generalize traditional word embedding research along a different dimension. They extract context-sensitive features from a left-to-right and a right-to-left language model.[3]



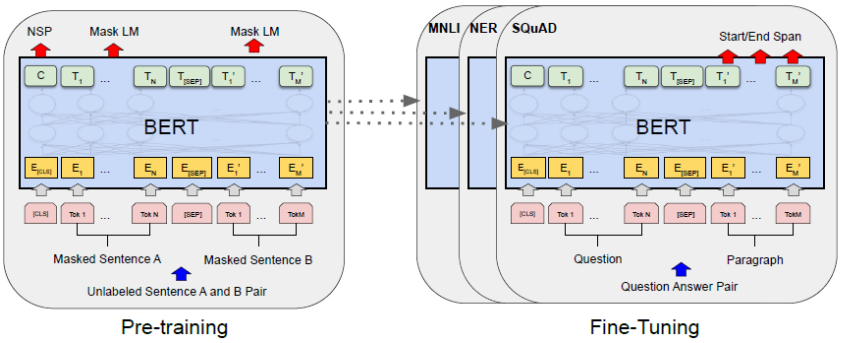
To some extent, BERT's network structure is quite similar to ELMo as they are both bidirectional, but BERT is deeper, and BERT also includes fine-tuning approach.

Fine-tuning Approach

Fine-tuning is a way to leverage the pre-trained model to complete a new task. The advantage of this approach is that it doesn't require a lot of parameters to be trained from the start. One of the state of the art model that used this approach is OpenAI GPT.

BERT

Bert uses feature-based approach for pre-training and use fine-tuning to complete the downstream tasks.



One key innovation in BERT is that it used a novel technique called Masked LM (MLM) which allows bidirectional training in models. Inspired by cloze task, researchers randomly mask some percentage of the input tokens in order to enable bidirectional training. Another task in BERT pre-training is Next Sentence Prediction (NSP), which is used to teach the model the relationship between two sentences. These features are important in tasks such as Question Answering (QA) and Natural Language Inference (NLI).

Fine-tuning in BERT is straight forward. It uses the self-attention mechanism to supervise the learning.

---

[3] Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".