ELMo, BERT, GPT

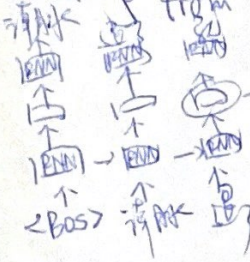
contextualized word embedding.

money bank

river bank

blood bank

Embeddings from Language Model RNN-based language models (trained from lots of sentences 不需要label data)



+ 丢到RNN

its word embedding = (丢到RNN word embedding, 丢到RNN word embedding)

Each layer in deep LSTM can generate a latent representation

final word embedding = a_1 · 第一层 layer word embedding + a_2 · 第二层 layer word embedding

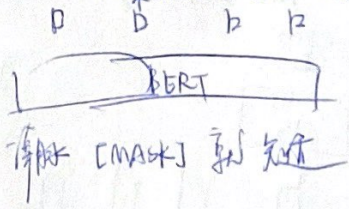
↓ learned with downstream tasks

Bidirectional Encoder Representations from Transformers (BERT): encoder of Transformer

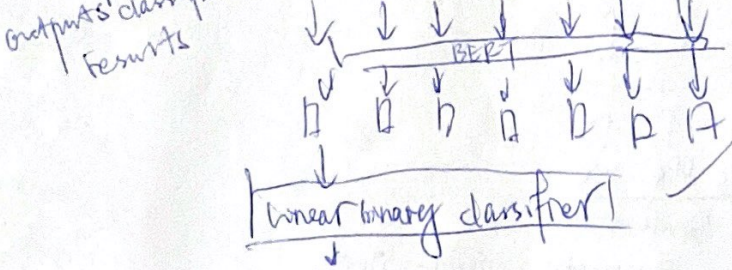
一步 input sequence (one-hot encoding), output 一步 word embedding. learned from a large amount of text without annotation

Although use word as unit here, character may be a better choice for Chinese (词太多 4000多个字)

Training of BERT: ① Masked LM: 15% mask



② next sentence prediction 两个句子是否连接

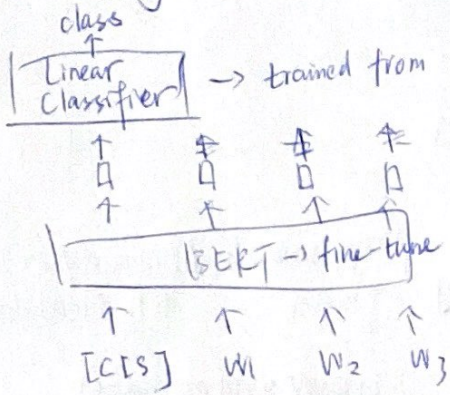


一起训练

还讲了 how to use BERT (与 downstream task 结合在一起).

How to use BERT - Case I

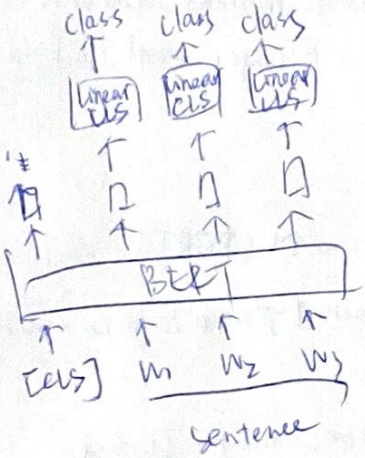
input: single sentence output: class



multilingual BERT (104 languages)

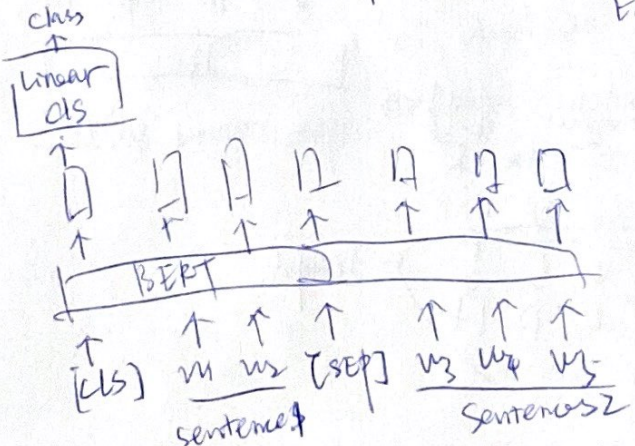
Case II: input: single sentence output: class of each word

Example: slot filling

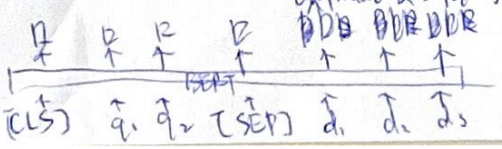


Case III: input: two sentences output: class

Example: NLI



Case IV: Extraction-based question



document
question
→ QA model → start → end

start
end
→ learn from scratch

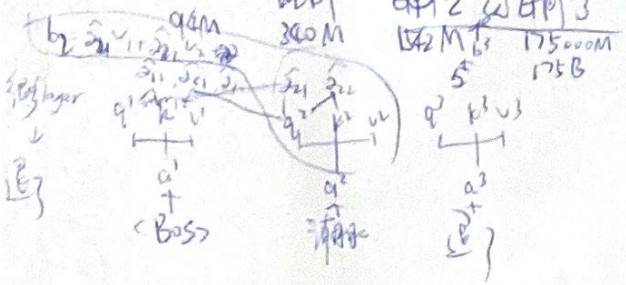
if ecs. 的答案

Generative Pre-Training (GPT)

ELMO BERT GPT2 & GPT3 → transformer decoder

94M 340M 1542M 5배 175000M 175B

GPT-2 는 release 하지 않고 BERT 보다 좋지 않다



Zero-learning?

reading comprehension.

$d_1, d_2, \cdots, d_N$, "Q:", $q_1 \cdots q_n$, "A:"

summarization

GPT-2 ← random $d_1, d_2, \cdots, d_N$, "TL;DR:"

Translation

English sentence = French sentence

_____ = _____ 2

_____ 3 =

In-context

~~Few-shot~~ learning : no gradient descent

closed book QA (open-domain QA)

Trivia QA. 175B with 64 few-shot example. 보다 fine-tuned SOTA

NLI. GPT-3 못함. ANLI Round3