Masked Language Modeling

Writer: 이미지처리팀-류채은

Ref:

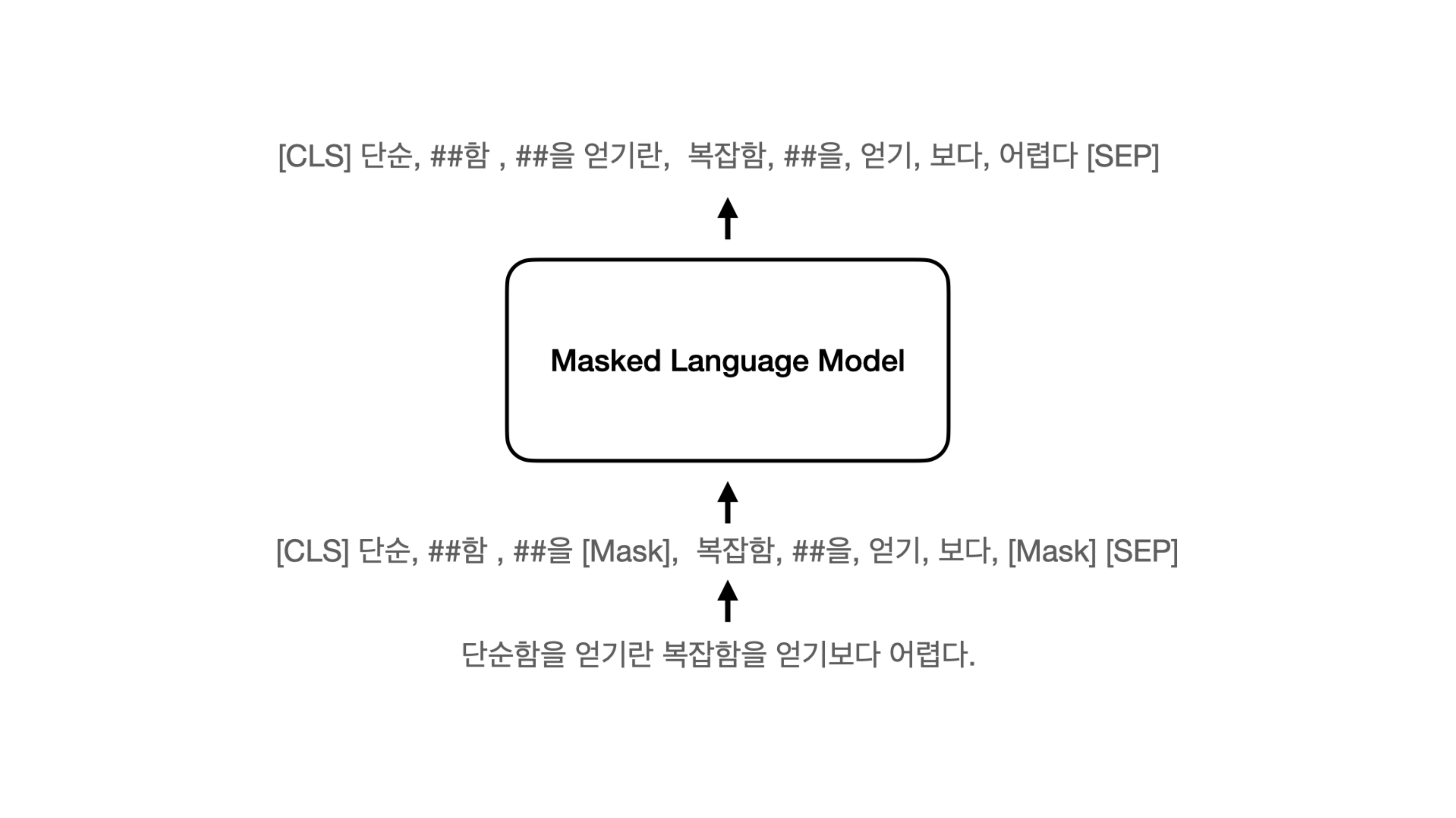
- https://towardsdatascience.com/masked-language-modelling-with-bert-7d49793e5d2c
- https://velog.io/@nawnoes/나만의-언어모델-만들기-Masked-Language-Model-학습
- https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[Background: What is Masked Language Modeling?]

Masked Language Modeling = **“Fill in the Gaps”**

마스킹된 언어 모델링으로, 입력으로 사용하는 문장의 토큰 중 특정 확률(15%)로 선택된 토큰을 **[MASK]** 토큰으로 변환시키고, 언어 모델을 통해 변환되기 전 **[MASK]** 토큰을 예측하는 언어 모델입니다.

```
In Autumn the ______ fall from the trees.
```


[CLS] 단순, ##함 , ##을 얻기란, 복잡함, ##을, 얻기, 보다, 어렵다 [SEP]
Masked Language Model
[CLS] 단순, ##함 , ##을 [Mask], 복잡함, ##을, 얻기, 보다, [Mask] [SEP]
단순함을 얻기란 복잡함을 얻기보다 어렵다.

For BERT, this guess will come from reading *a lot*
— and learning linguistic patterns incredibly well.

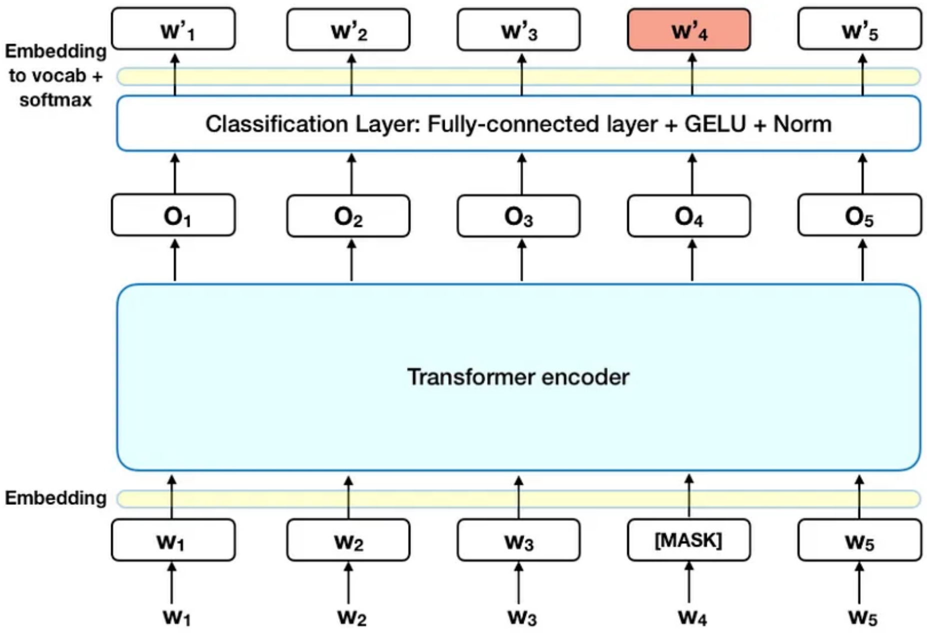
How Bert Works
As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), **the Transformer encoder reads the entire sequence of words at once**. Therefore it is considered bidirectional, though it would be more accurate to say that it’s non-directional. **This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word)**.

Training Process of BERT

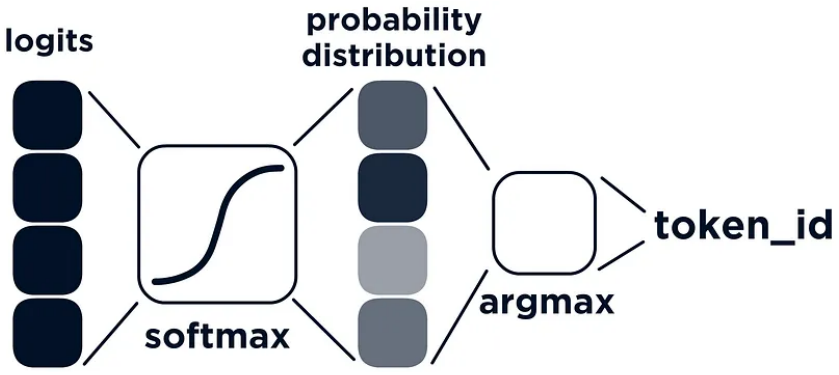
When training language models, there is a challenge of defining a prediction goal. Many models **predict the next word** in a sequence (e.g. “The child came home from ___”), a **directional approach that inherently limits context learning.** To overcome this challenge, BERT uses **two training strategies**:

1. **Masked LM (MLM)** ← ⭐
2. **Next Sentence Prediction (NSP)**

[MLM]



MLM in Summary

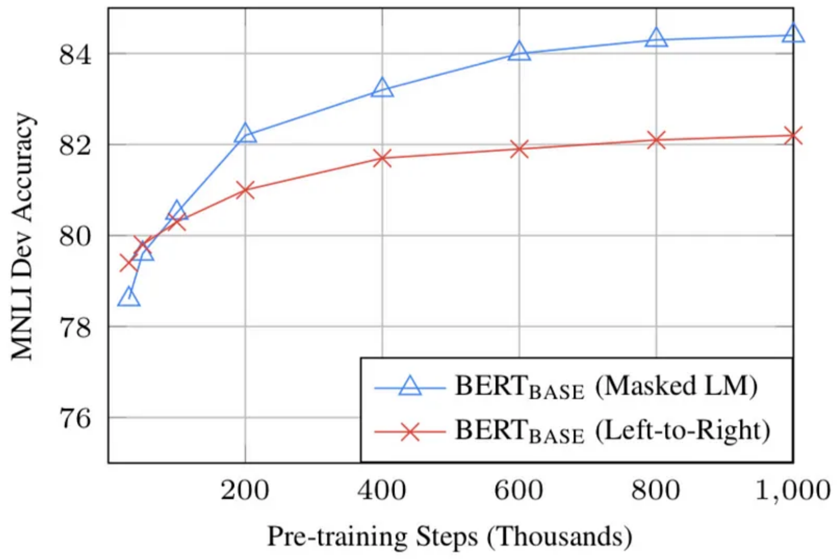


How mask is predicted: the **logits** — which has a vector length equal to the model vocab size. The predicted **token_id** is extracted from this logit using a softmax and argmax transformation.

The goal of predicting word at the masked region requires the three following stages:

1. Adding a classification layer on top of the encoder output.
2. Multiplying the output vectors by the embedding matrix, transforming them into the vocabulary dimension.
3. Calculating the probability of each word in the vocabulary with softmax.

   (The BERT loss function takes into consideration only the prediction of the masked values and ignores the prediction of the non-masked words.)



BERT's bidirectional approach (MLM) ***converges slower*** *than left-to-right approaches* but bidirectional training still outperforms left-to-right training after a small number of pre-training steps.