한국코드페어
KOREA CODE FAIR

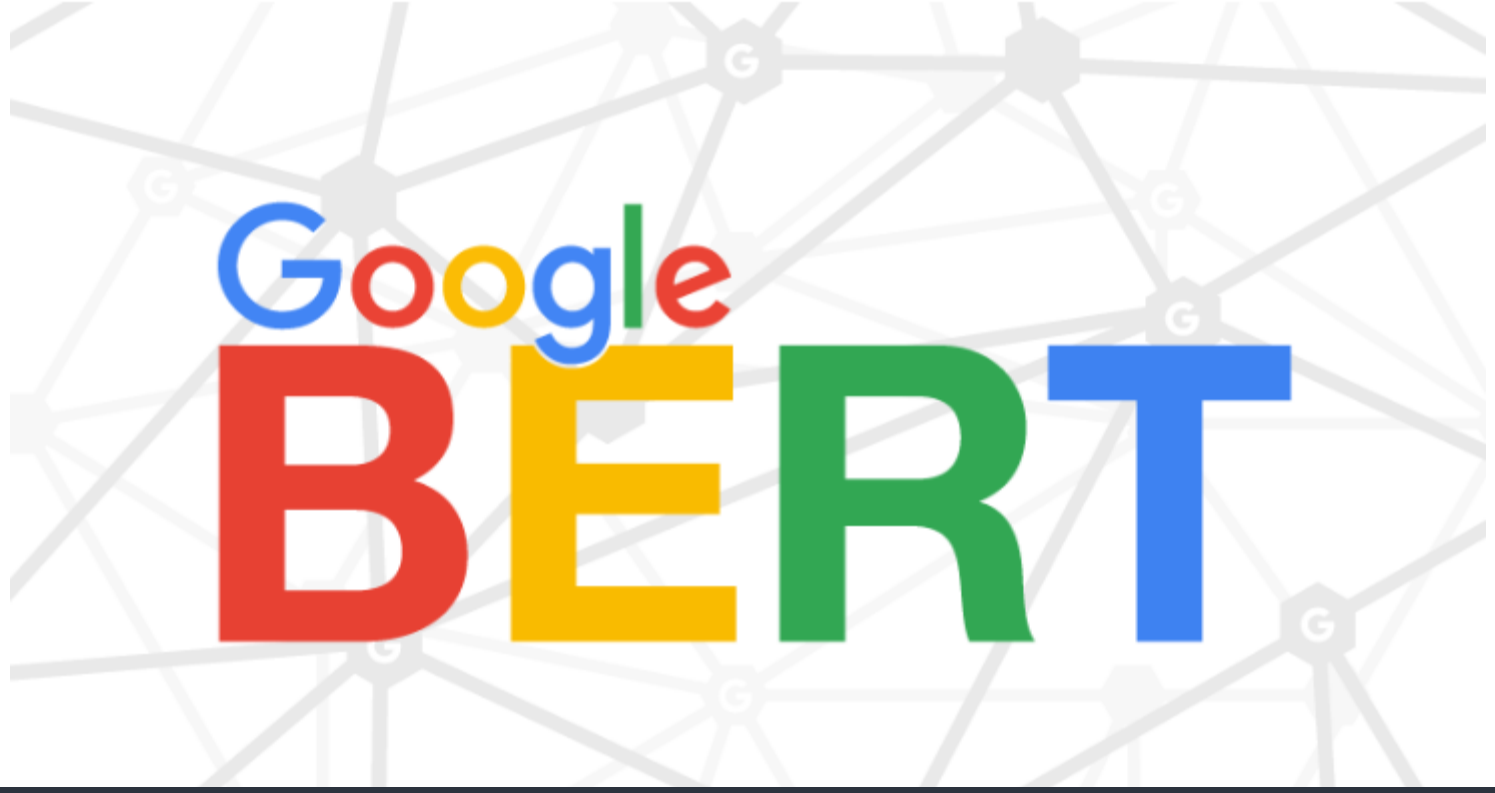
가볍게 시작하는 AI 입문

304: BERT와 실습

BERT 란?

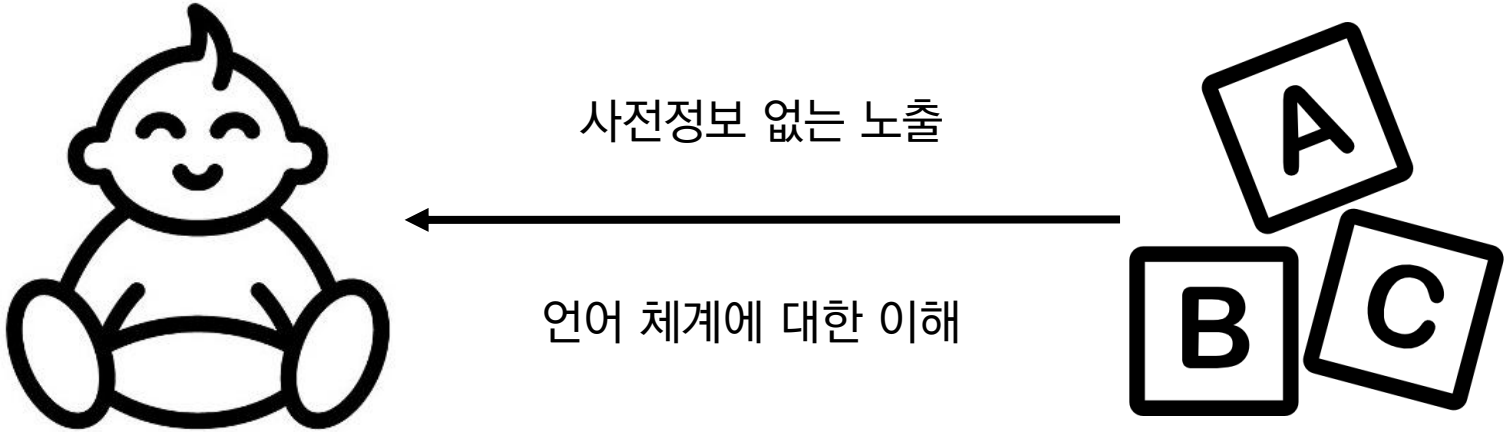
1. BERT: Pre-trained Model
2. BERT: fine-tuning
3. BERT만의 특징

Transformer와 BERT



Attention is All You Need (2017)

BERT: Pre-trained Model



BERT: Pre-trained Model

Input = [CLS] the man went to [MASK] store [SEP]

he bought a gallon [MASK] milk [SEP]

Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]

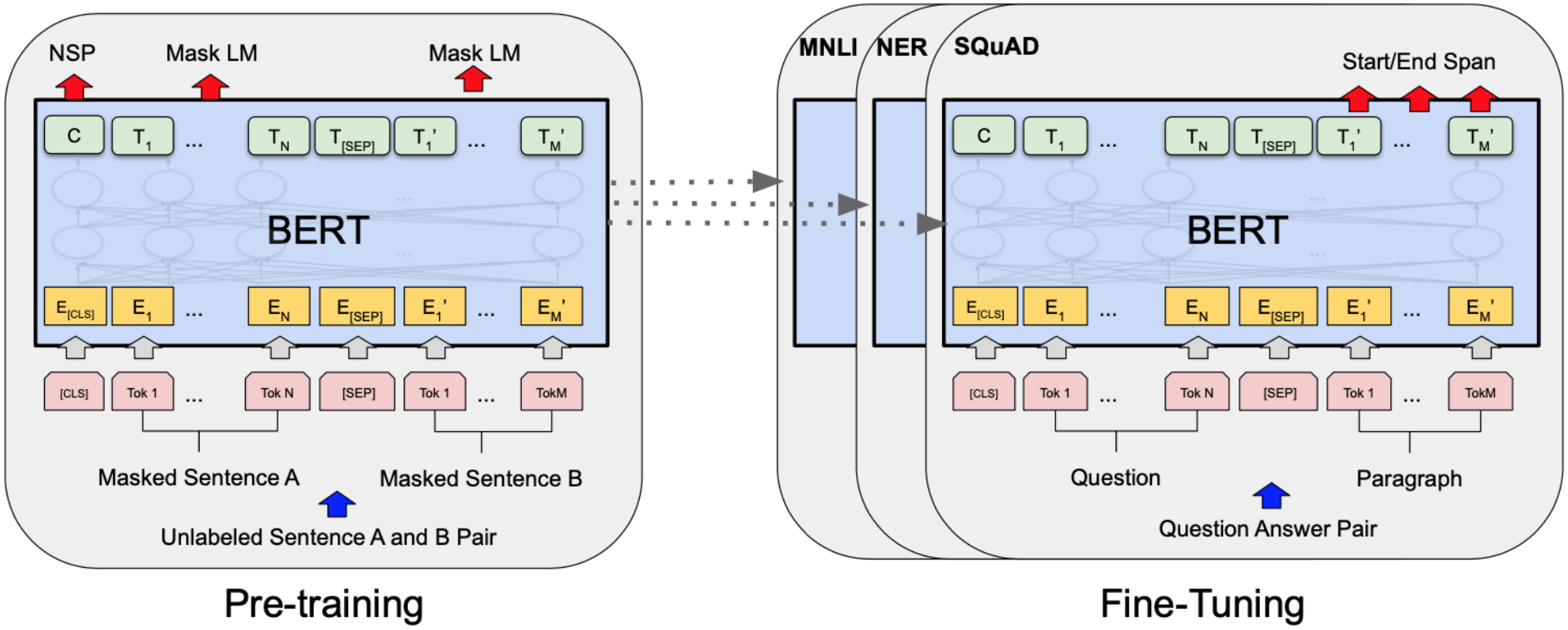
penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

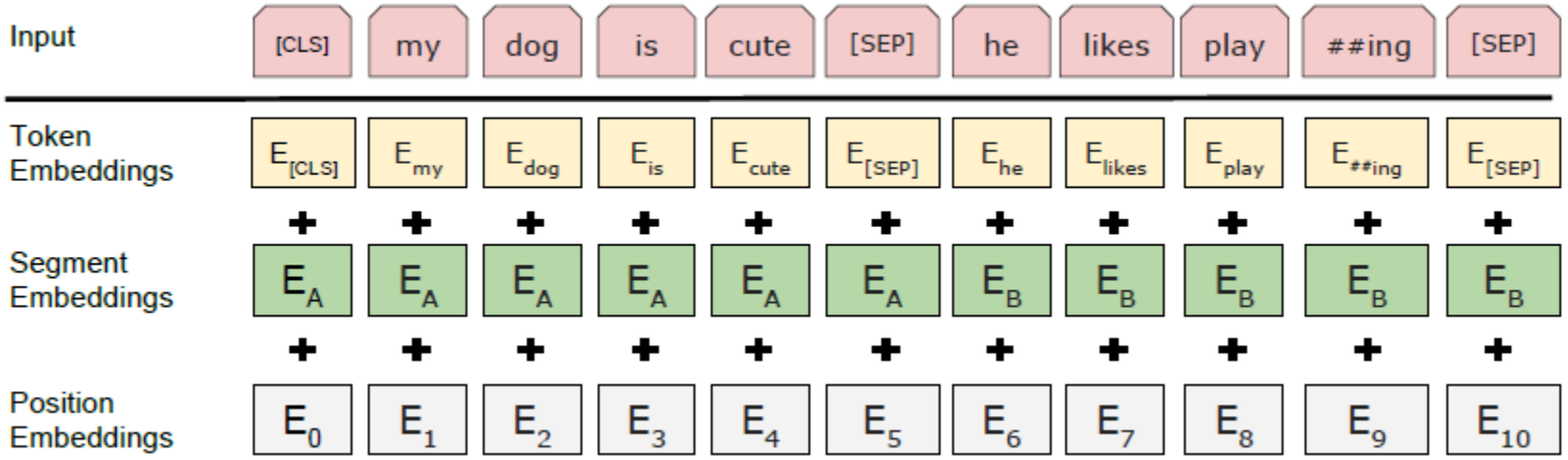
논문의 task 예시 중 MLM

1. Mask Language Model (MLM)
2. Next sentence prediction

BERT: fine tuning



BERT 만의 특징



1. Token Embedding: 특수 토큰을 이용해 단어를 표현
2. Segment Embedding: 두개의 문장을 구분
3. Position Embedding: 위치정보를 학습

BERT 만의 특징

1. [CLS]
   - Special Classification token
   - 시작 토큰으로 삽입, 분류 문제에서 사용
2. [SEP]
   - 첫번째 문장과 두번째 문장을 구분
   - Segment Embedding에서 사용
   - 각 문장의 끝에 삽입

네이버 영화리뷰 감성분석

네이버 영화리뷰 감정분석 with Hugging Face BERT

BERT(Bidirectional Encoder Representations from Transformers)는 구글이 개발한 사전훈련(pre-training) 모델입니다. 위키피디아 같은 텍스트 코퍼스를 사용해서 미리 학습을 하면, 언어의 기본적인 패턴을 이해한 모델이 만들어집니다. 이를 기반으로 새로운 문제에 적용하는 전이학습(transfer learning)을 수행합니다. 좀 더 적은 데이터로 보다 빠르게 학습이 가능하다는 장점이 있습니다. 그래서 최근 자연어처리의 핵심 기법으로 떠오르고 있습니다.

이 예제에서는 한글 NLP의 Hello world라고 할 수 있는 네이버 영화리뷰 감정분석을 구현해보겠습니다. 가장 유명한 모델 중 하나인 Hugging Face의 PyTorch BERT를 사용하였습니다. 아래의 Chris McCormick의 블로그를 참조하여 한글에 맞게 수정하였음을 미리 알려드립니다.

< BERT Fine-Tuning Tutorial with PyTorch >

-> https://mccormickml.com/2019/07/22/BERT-fine-tuning

https://colab.research.google.com/drive/1tIf0Ugdqg4qT7gcxia3tL7und64Rv1dP#scrollTo=WkAHQrj2VjbI