딥러닝의 통계적이해

14강. 딥러닝 모형을 이용한 자연어 처리 (2)

1. Transformer
2. BERT
3. 최신 자연어 모형 트렌드

SK텔레콤 김기온

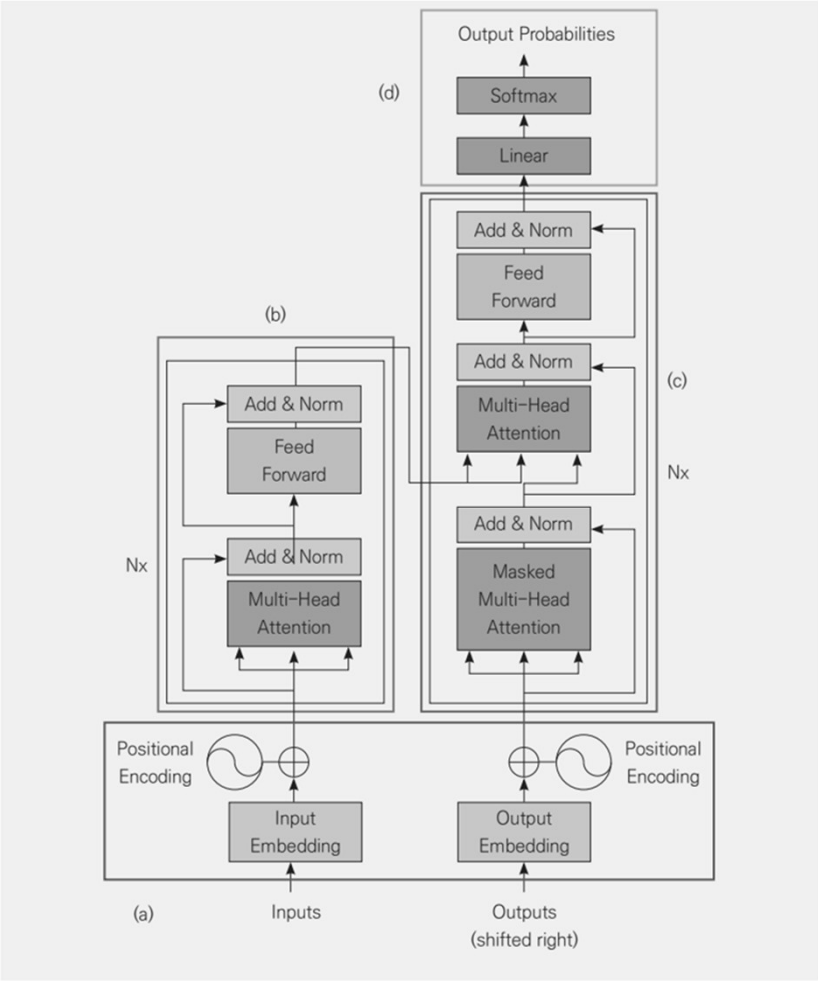
오늘의 학습목표

1. Transformer의 scaled-dot-product와 multi-head attention에 대해 이해한다.
2. BERT 모형에 대해서 이해한다.
3. BERT 이후 자연어 모형의 발전 흐름에 대해 이해한다.

1. Transformer

1. Transformer

Transformer (Overall Architecture)

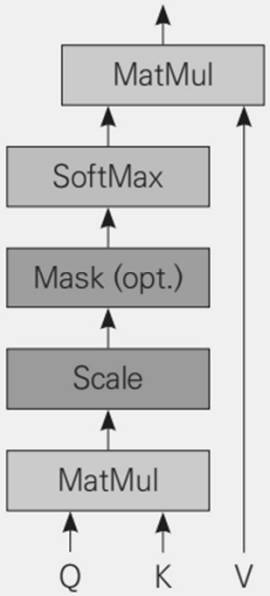


출처 : Vaswani, et. al(2017)

1. Transformer

Scaled dot-product attention

[그림 9.8] 스케일드-닷-프로덕트 어텐션



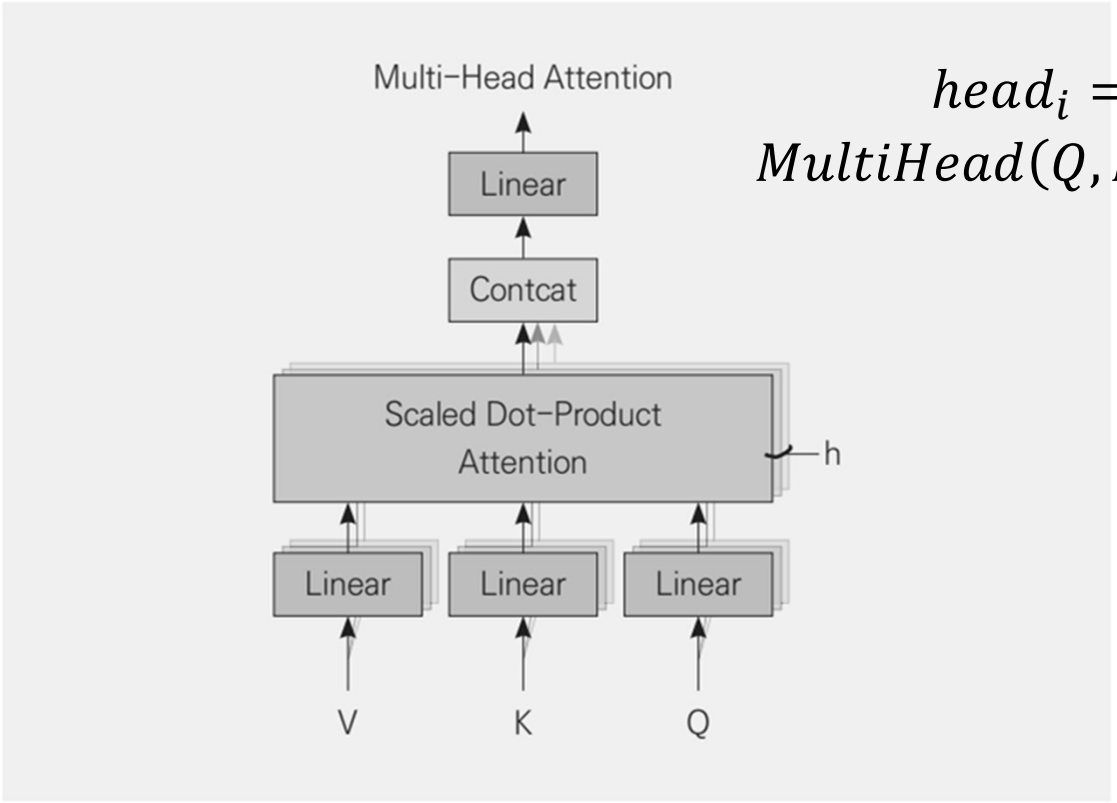
$$Att = Attention(QW^Q, KW^K, VW^V)$$
$$= softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

출처 : Vaswani, et. al(2017)

1. Transformer

Multi-head attention

[그림 9.10] 멀티-헤드 어텐션



$$head_i = Attention(QW_i^Q, KW_i^k, VW_i^V)$$
$$MultiHead(Q,K,V) = concat(head_1, \dots, head_h)W^O$$

출처 : Vaswani, et. al (2017)

2. BERT

2. BERT

BERT

- Bidirectional Encoding Representation from Transformer
- 문장 단위의 embedding을 도출
- Transformer의 Encoder block을 활용

[표 9.4] BERT 모형의 종류

모형의 종류	트랜스포머 블록의 개수	임베딩 차원	멀티 헤드 어텐션 헤드의 수	총 모수의 수
BERT base	12	768	12	110M
BERT large	24	1,024	16	340M

	Training Compute + Time	Usage Compute
$BERT_{BASE}$	4 Cloud TPUs, 4 days	1 GPU
$BERT_{LARGE}$	16 Cloud TPUs, 4 days	1 TPU

https://www.lyrn.ai/2018/11/07/explained-bert-state-of-the-art-language-model-for-nlp/

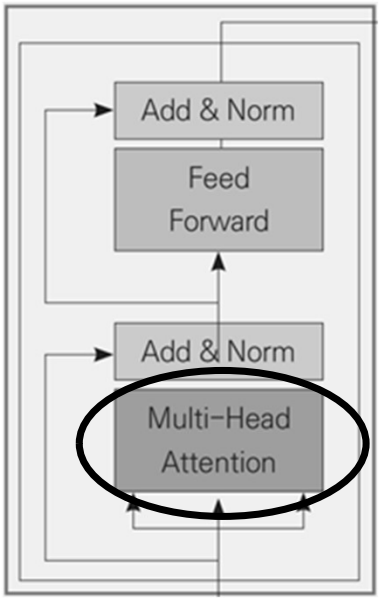
2. BERT

Encoder block

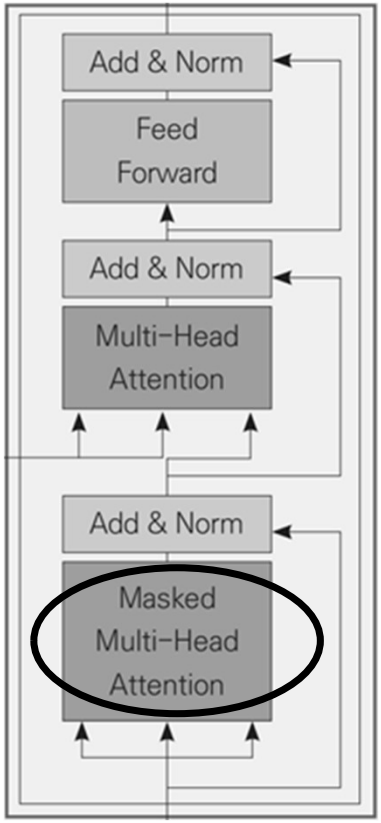
Language Model: A probability distribution over words

$$P(w_1, \ldots, w_n) = P(w_n|w_{n-1}) \cdot \ldots \cdot P(w_2|w_1)$$

Encoder block



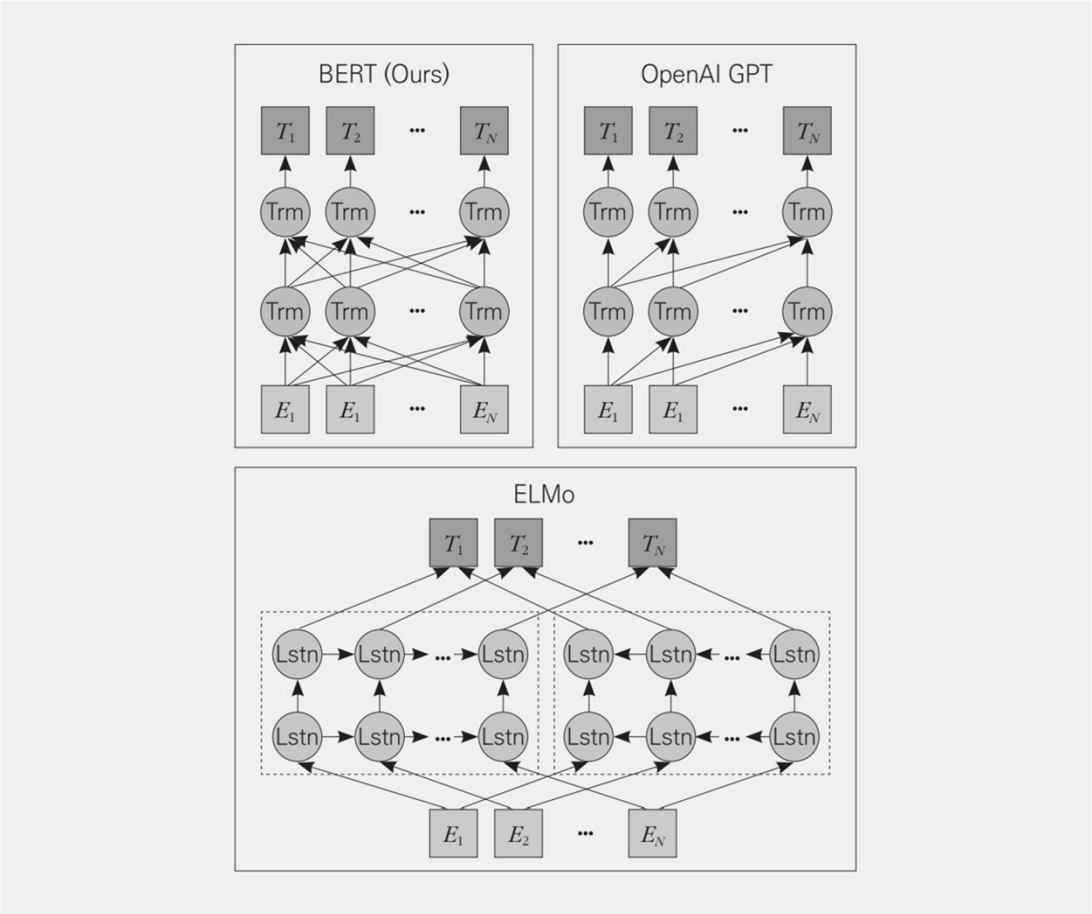
Decoder block



2. BERT

BERT / GPT / ELMO

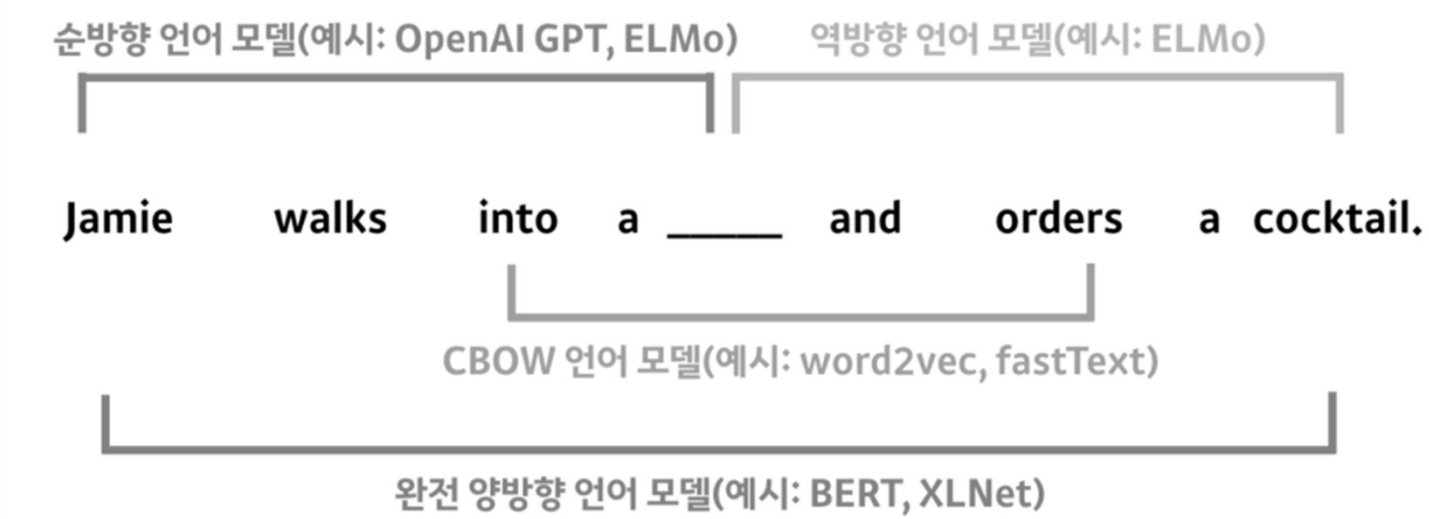
[그림 9.13] BERT, GPT와 ELMo



출처 : Devlin et al. (2018)

2. BERT

BERT

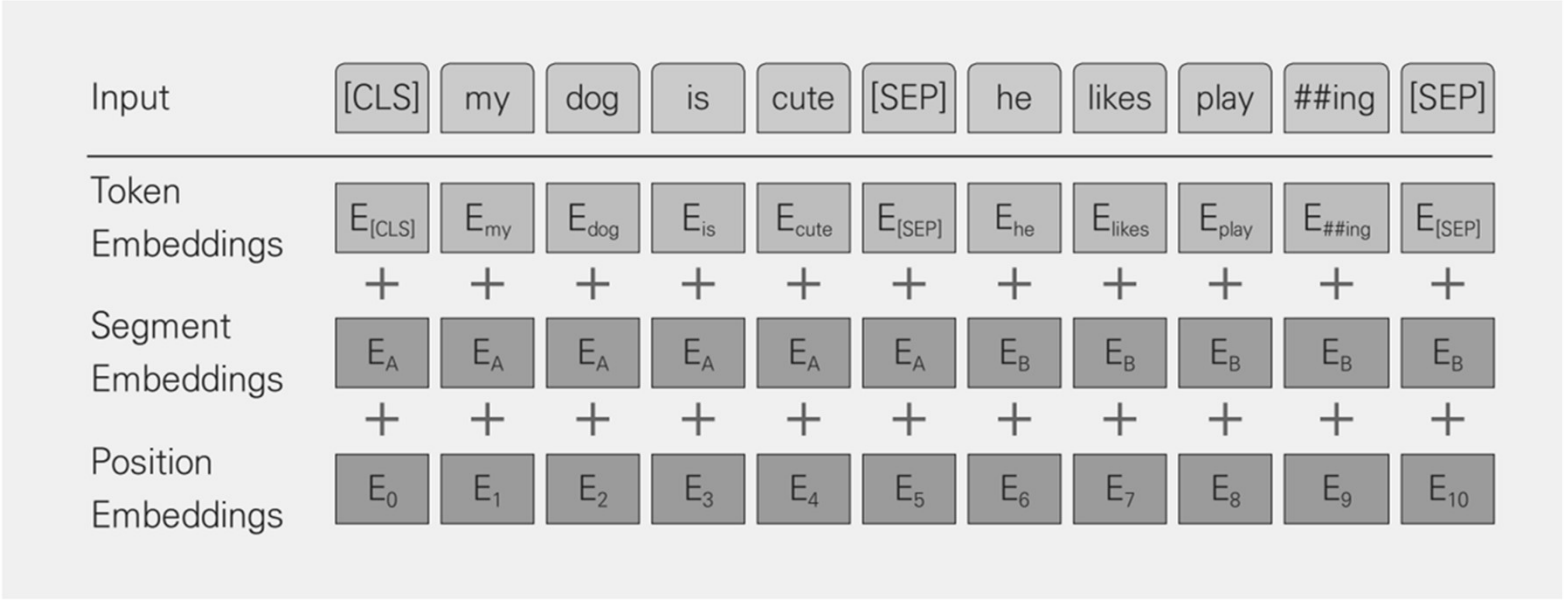


https://www.kakaobrain.com/blog/118

2. BERT

BERT의 입력 표현

[그림 9.12] BERT의 입력 표현

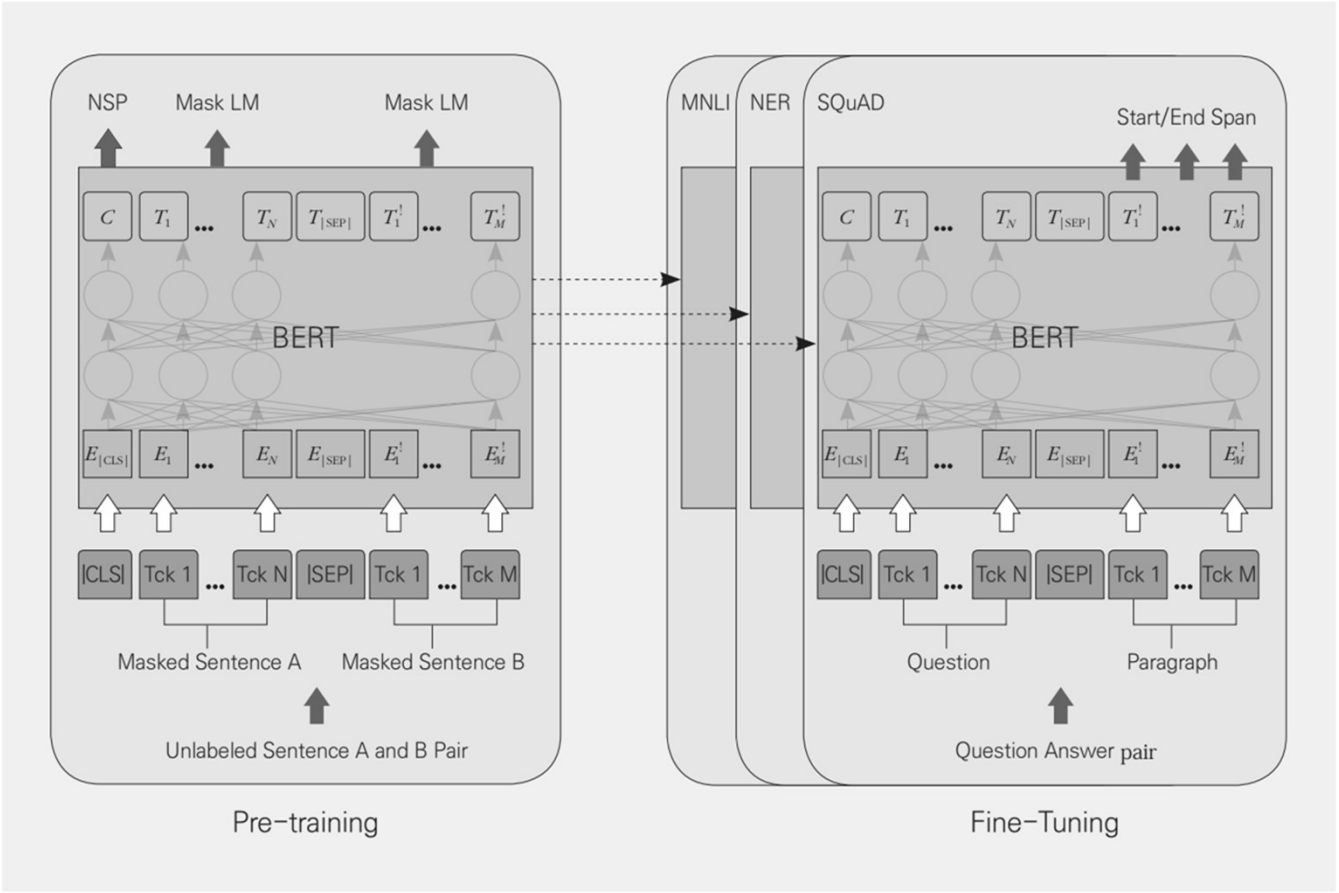


출처 : Devlin et al. (2018)

2. BERT

BERT 학습 과정

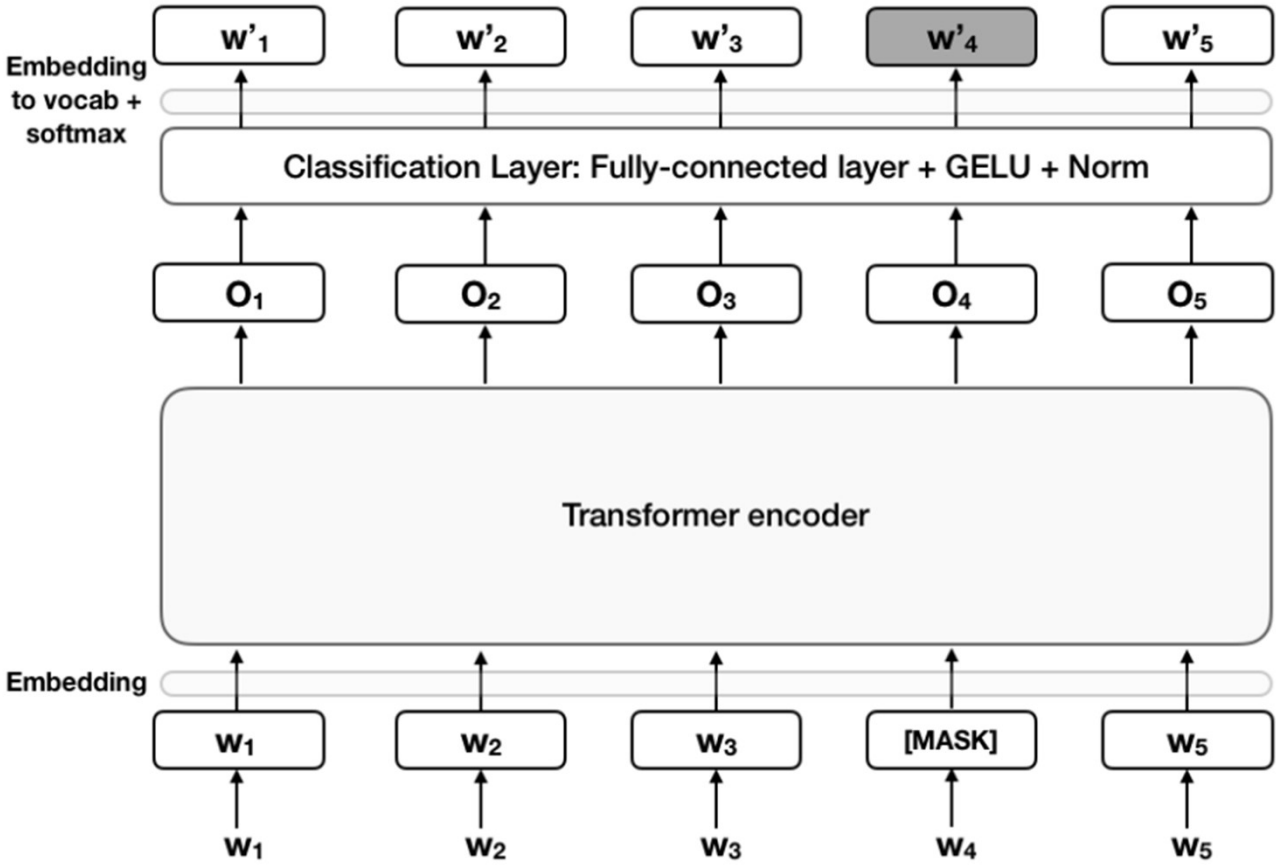
[그림 9.11] 전이학습과 미세조정



출처 : Devlin et al. (2018)

2. BERT

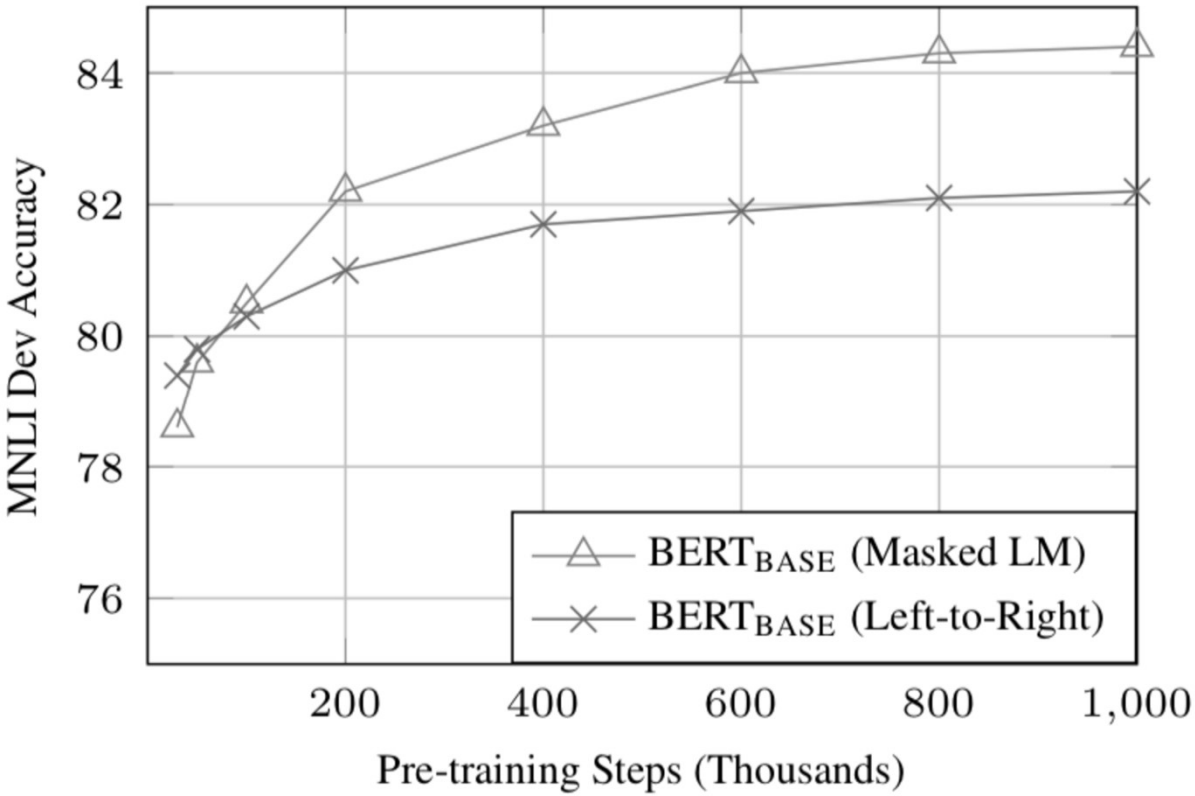
Pretrain – Masked LM



https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

2. BERT

Pretrain – Masked LM



2. BERT

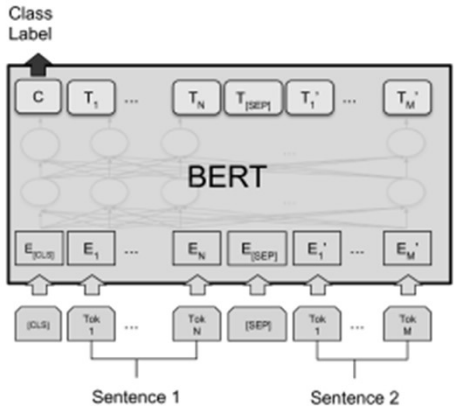
Pretrain – Next Sentence Prediction

[표 9.5] 다음-문장-예측

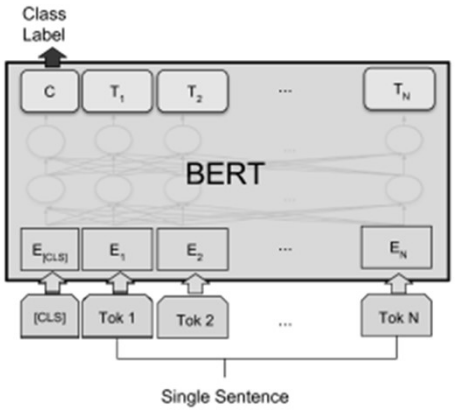
Input = [CLS] the man went to [MASK] store [SEP] he 〉bought a gallon [MASK] milk [SEP] Label = IsNext
Input = [CLS] the man [MASK] to the store [SEP] 〉penguin [MASK] are flight ##less birds [SEP] Label = NotNext

2. BERT

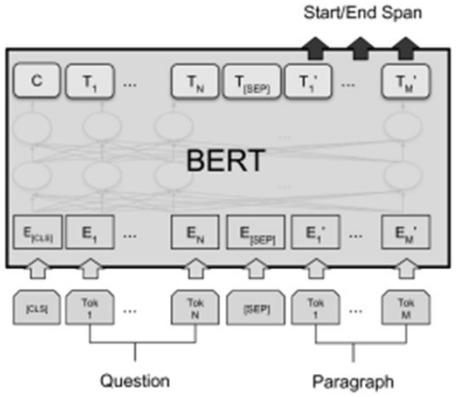
Fine tuning



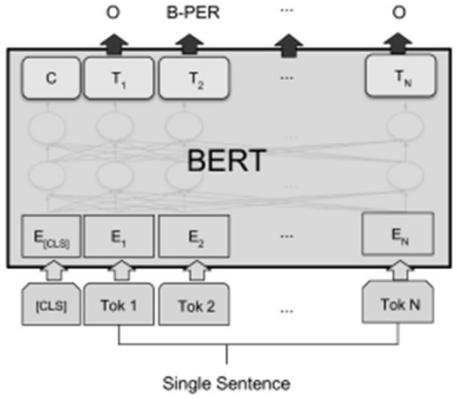
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

2. BERT

BERT 성능비교

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Table 1: GLUE Test results, scored by the evaluation server (`https://gluebenchmark.com/leaderboard`). The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.[8] BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

3. 최신 자연어 모형 트렌드

3. 최신 자연어 모형 트렌드

2018년 이후 모델

년	월	모형	저자
2018년	2월	ELMo	Allen AI, U of Washington
	5월	GPT-1	Open AI
	10월	BERT	Google
	2월	GPT-2	Open AI
	5월	MT-DNN	Microsoft
2019년	7월	XLNet	CMU + Google Brain
	7월	RoBERTa	FAIR
	9월	ALBERT	Google + TTIC
	10월	T5	Google
2020년	3월	ELECTRA	Google
	6월	GPT-3	Open AI

3. 최신 자연어 모형 트렌드

NLP 발전 방향

- ✓ 모형의 크기를 줄이는 방향
  → ALBERT, Trasformer.zip, ELECTRA
- ✓ 정확도를 높이는 방향
  → RoBERTa, XLNet, T5
- ✓ 생성 모형
  → Open GPT3, T5

학습정리

- ✓ Transformer는 scaled-dot-product attention을 여러 개 활용한 multi-head attention을 사용한다.
- ✓ BERT는 대용량 데이터를 대용량 리소스를 투입해 학습시킨 transformer 의 encoder block을 이용한 사전 학습 모형으로 현재의 NLP trend를 대표하는 문장 단위의 embedding 모형이다. Masked LM 기법을 주축으로 학습을 진행한다.
- ✓ BERT 이후의 모형들은 Transformer block을 준용하여, BERT 모형의 크기를 줄이거나, 아주 모형의 크기를 키워서라도 성능을 극대화하려는 두 가지 축으로 발전하고 있다.

딥러닝의 통계적이해

다음시간안내

15강. 딥러닝 실습