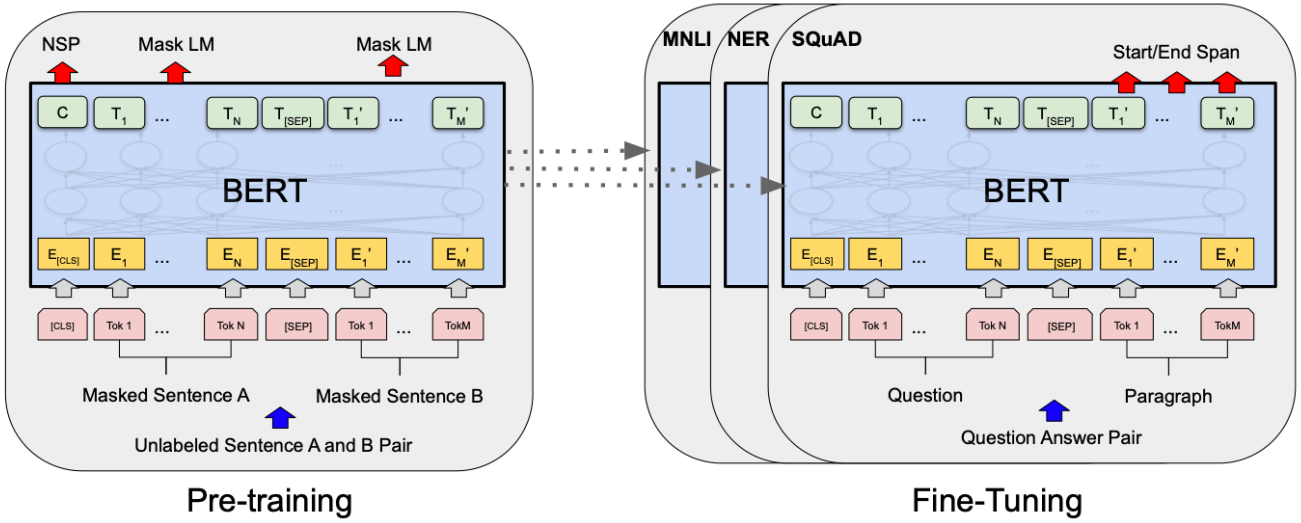
BERT:
Pre-training of Deep Bidirectional Transformers for Language Understanding

ESC 조민주

BERT



Classification layer for Fine-tuning

Pre-trained BERT

BERT 사용 x : 분류할 데이터 -> LSTM, CNN 등의 모델 -> 분류

BERT 사용 o : **관련 corpus 많이 -> BERT** -> 분류할 데이터 -> LSTM, CNN 등의 모델 -> 분류

ESC 조민주

BERT

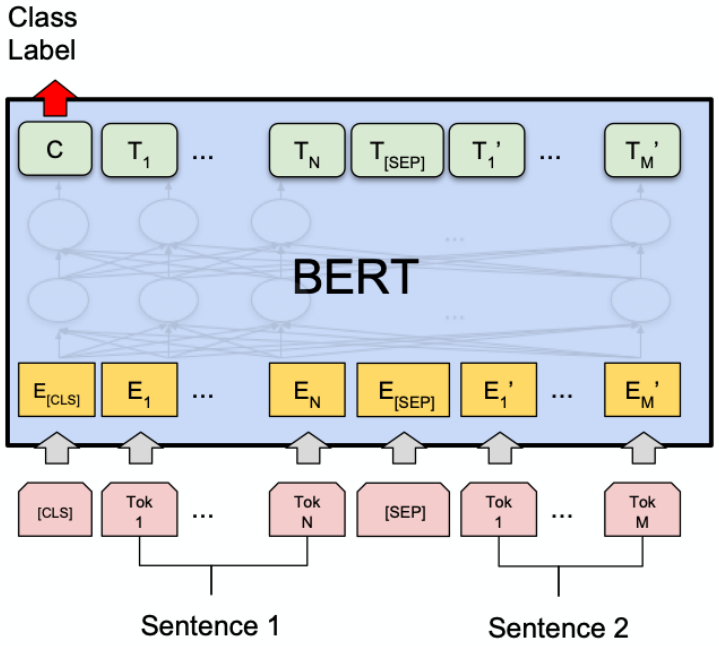
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	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**

구성(Transformer layer, Hidden-size, Self-attention head, Total Parameter 순서)

BERT Base : 12, 768, 12, 110m

BERT Large : 24, 1024, 16, 340m

BERT : 구조 및 방법



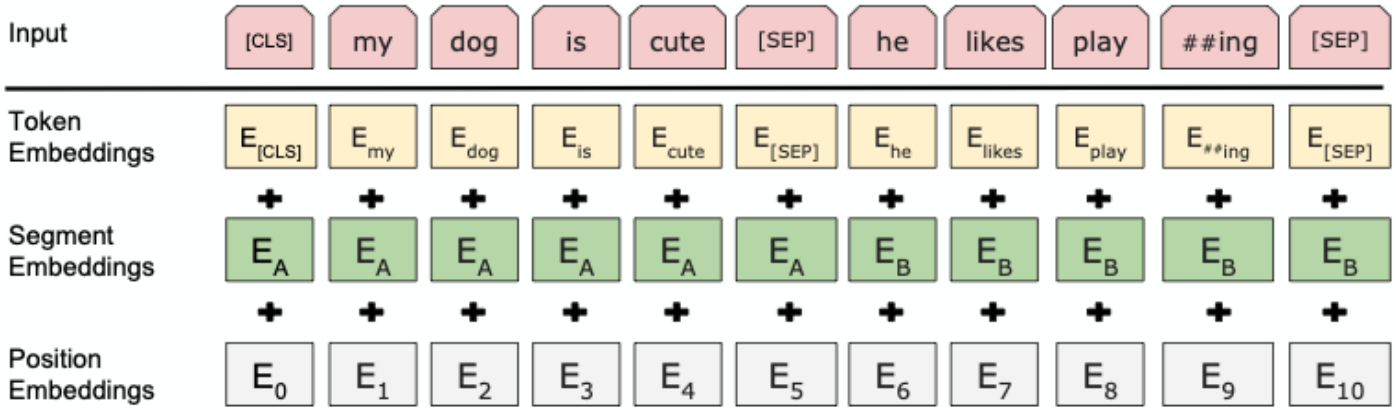
Contextual Representation of token

Transformer layer

Input embedding layer

BERT : 구조 및 방법

Input



BERT : 구조 및 방법

Pre-training

학습한 corpus

Pre-training data The pre-training procedure largely follows the existing literature on language model pre-training. For the pre-training corpus we use the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). For Wikipedia we extract only the text passages and ignore lists, tables, and headers. It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the Billion Word Benchmark (Chelba et al., 2013) in order to extract long contiguous sequences.

Tokenize

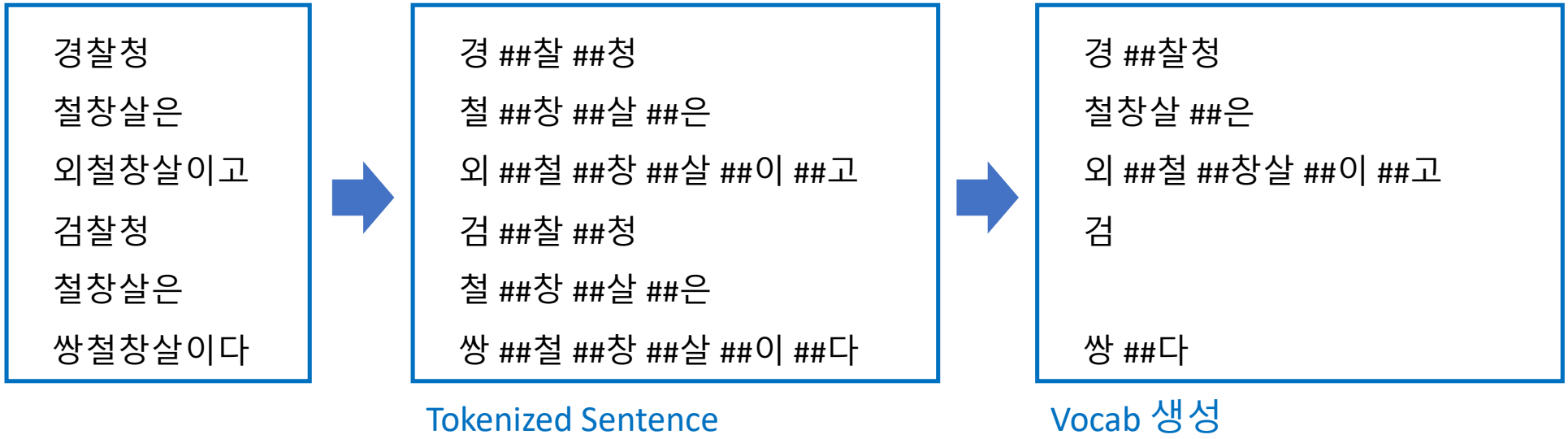
WordPiece tokenizing 방법 사용

1) input 문장을 위의 방법으로 tokenize
2) 1)의 token으로 token sequence를 학습
3) 두 개의 token sequence가 학습에 사용됨

BERT : 구조 및 방법

WordPiece Tokenizing

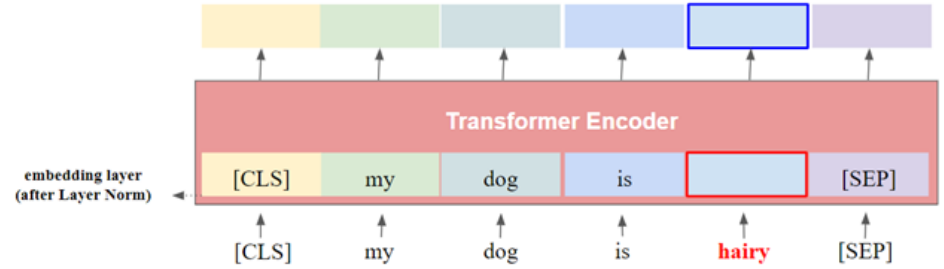
경찰청 철창살은 외철창살이고 검찰청 철창살은 쌍철창살이다



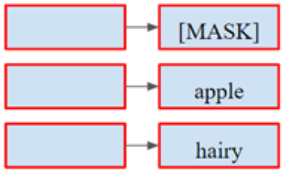
ESC 조민주

BERT : 구조 및 방법

MLM(Masked Language Model)



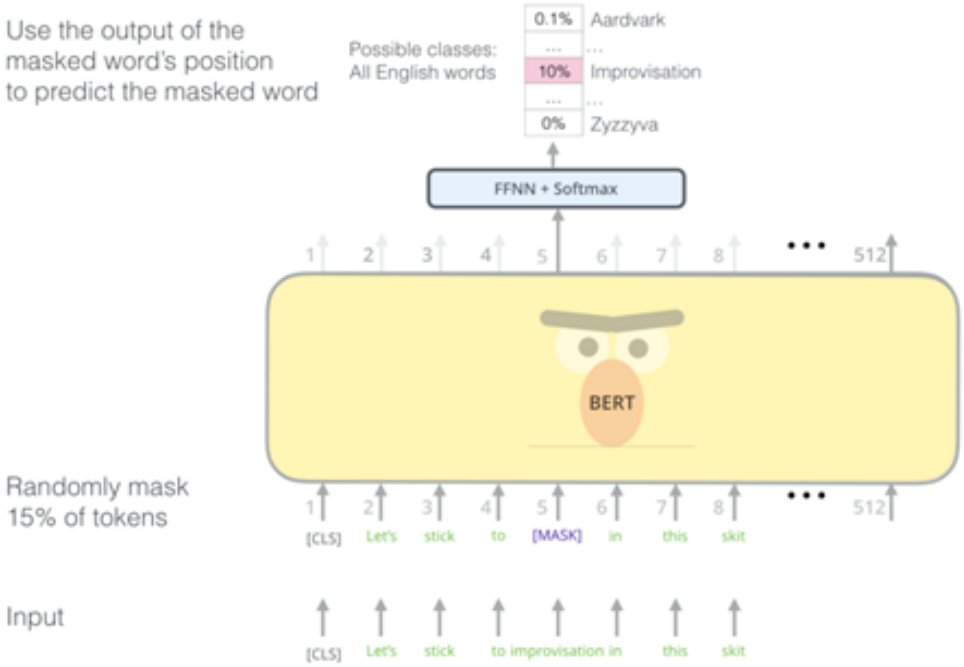
Mask **15%** of all WordPiece tokens in each sequence at **random**. (e.g., **hairy**)



80% of the time : Replace **[MASK]** token.

10% of the time : Replace the word with a **random** word

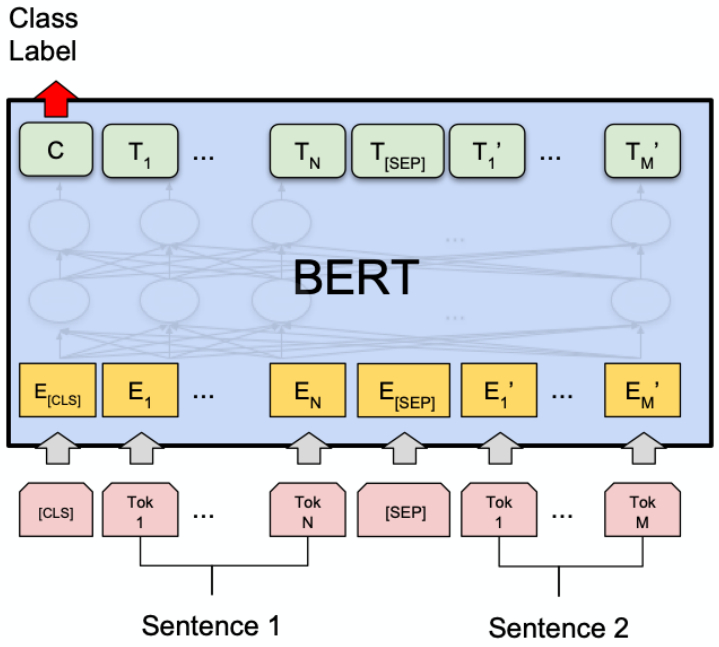
10% of the time : Keep the word **unchanged**.



ESC 조민주

BERT : 구조 및 방법

NSP(Next Sentence Prediction) – isNext? NotNext?



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

he bought a gallon [MASK] milk [SEP]

Label = IsNext

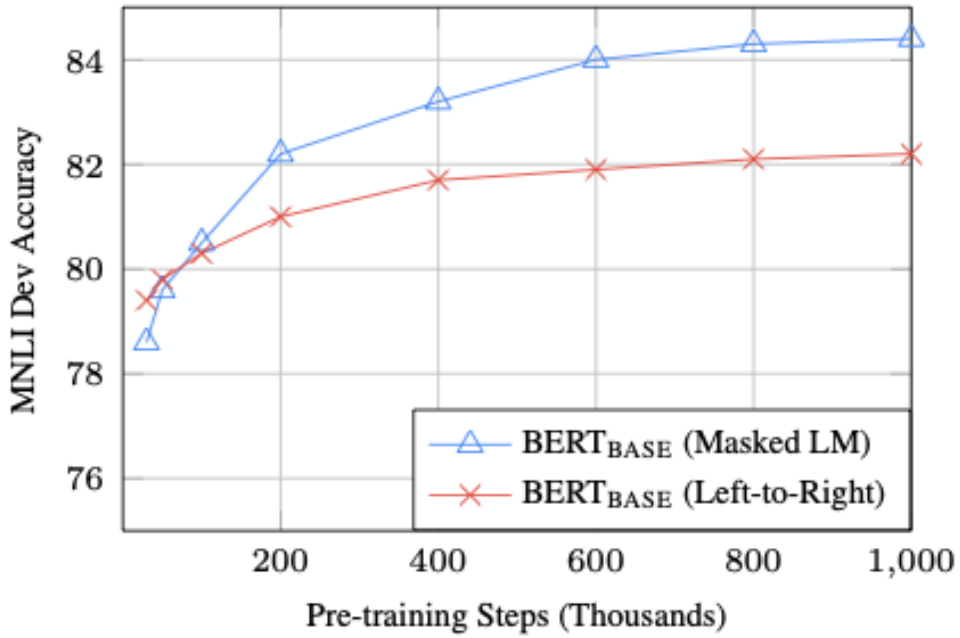
Input = [CLS] the man [MASK] to the store [SEP]

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

Label = NotNext

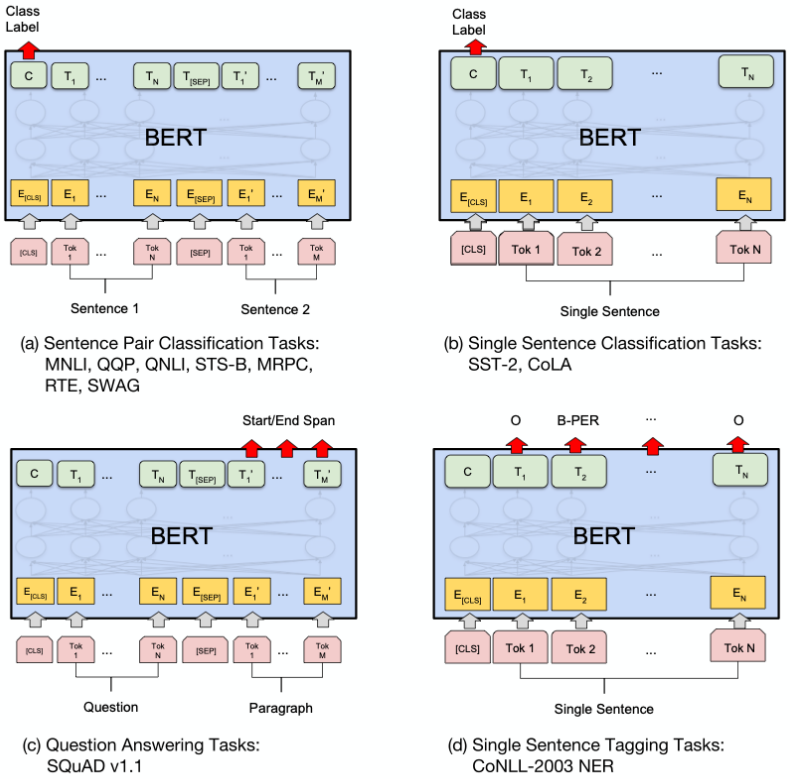
BERT : 구조 및 방법

MLM(Masked Language Model)



BERT : 구조 및 방법

Transfer Learning



(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

ESC 조민주

BERT : 성능

Hyperparams			Dev Set Accuracy			
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

끝

https://arxiv.org/pdf/1810.04805.pdf
https://www.youtube.com/watch?v=riGc8z3YIgQ
https://ebbnflow.tistory.com/151
https://keep-steady.tistory.com/19