모두의연구소

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova

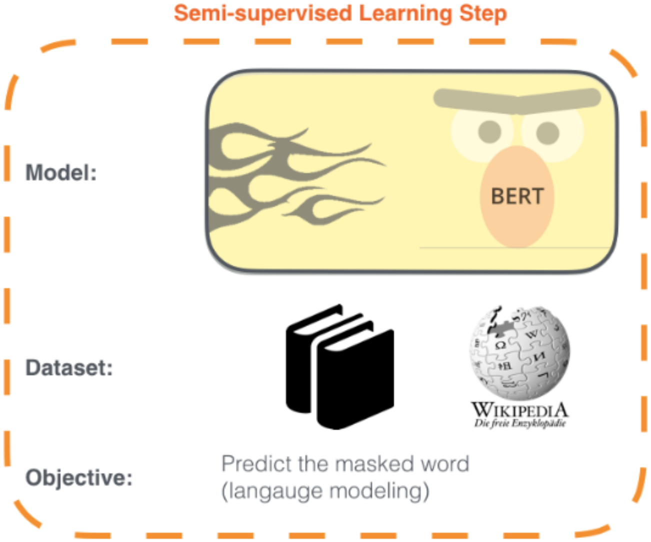
모두의연구소 김승일 연구소장

BERT : Bidirectional Encoder Representation from Transformer

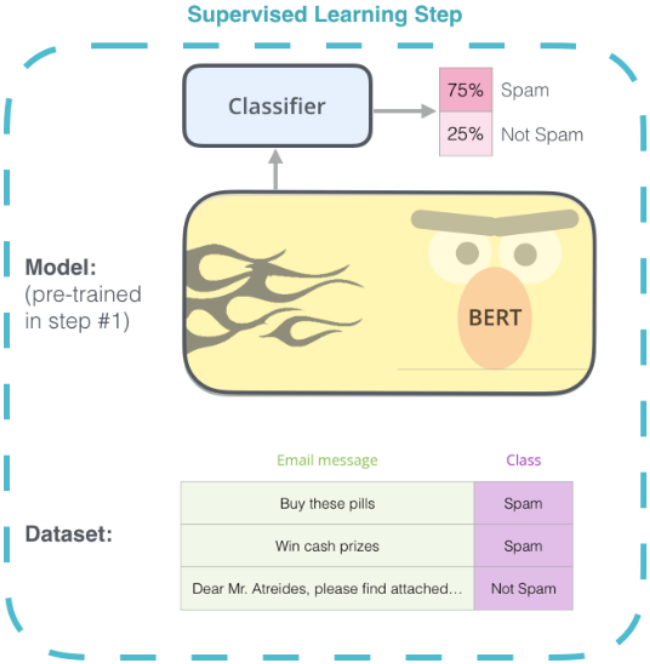
모두의연구소

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



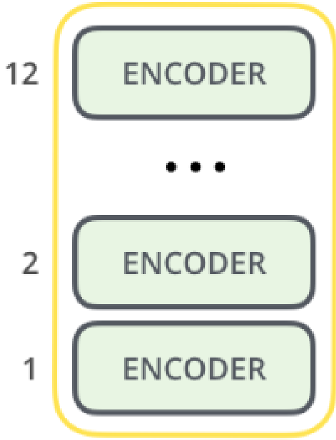
2 - Supervised training on a specific task with a labeled dataset.



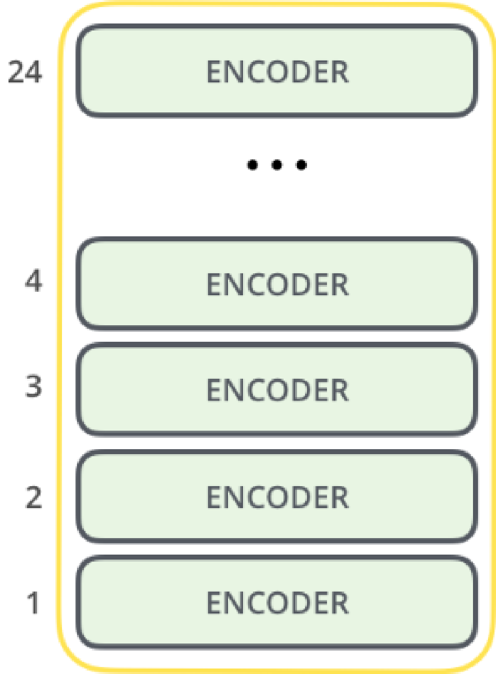
BERT : Bidirectional Encoder Representation from Transformer

- Original transformer : L=6, H=512, A=8
- BERTbase : L=12, H=768, A=12
- BERTlarge : L=24, H=1024, A=16

(L= # of layer, H= hidden size, A= # of self-attention head)

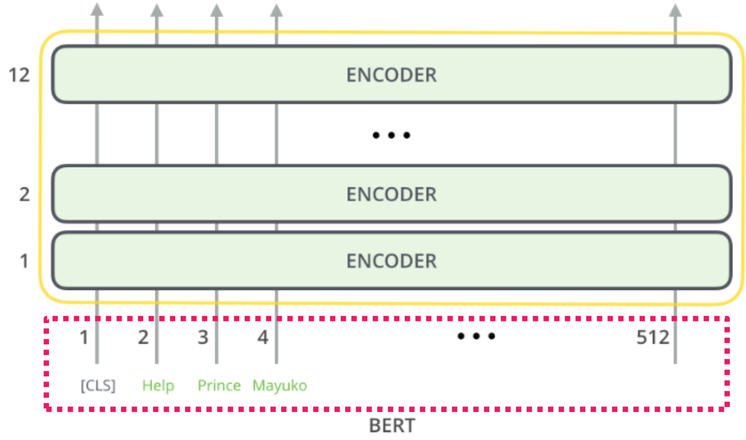


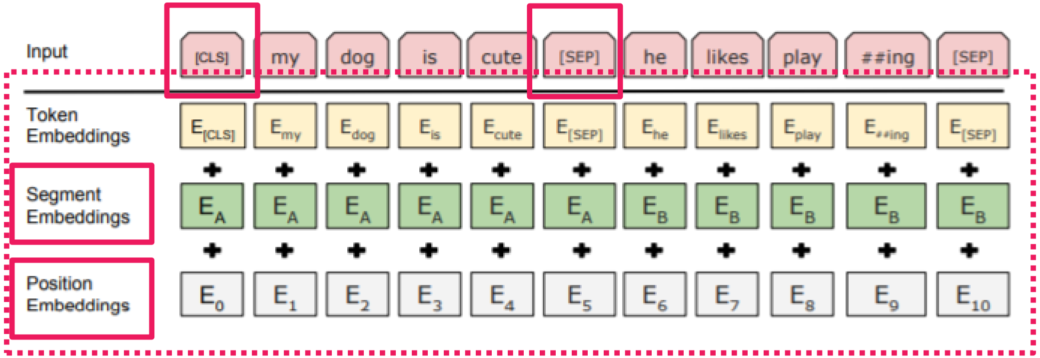
BERTBASE



BERTLARGE

Model Input

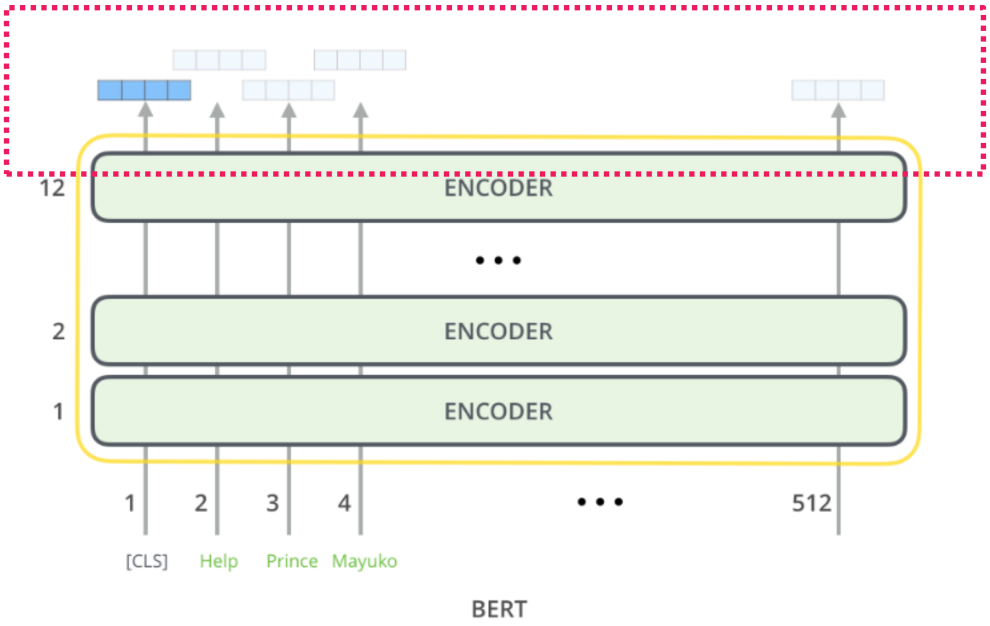




Q1. segment embedding을 사용하는데,
왜 [SEP] token이 필요한가?
Q2. seg/pos embedding 은 어떤 값인지?

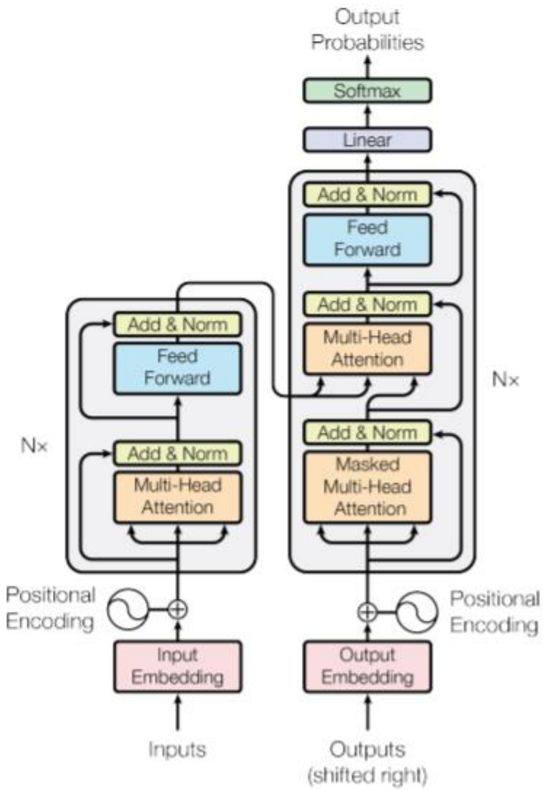
Model Output

각각의 position에 대해 H size의 embedding vector가 출력됨

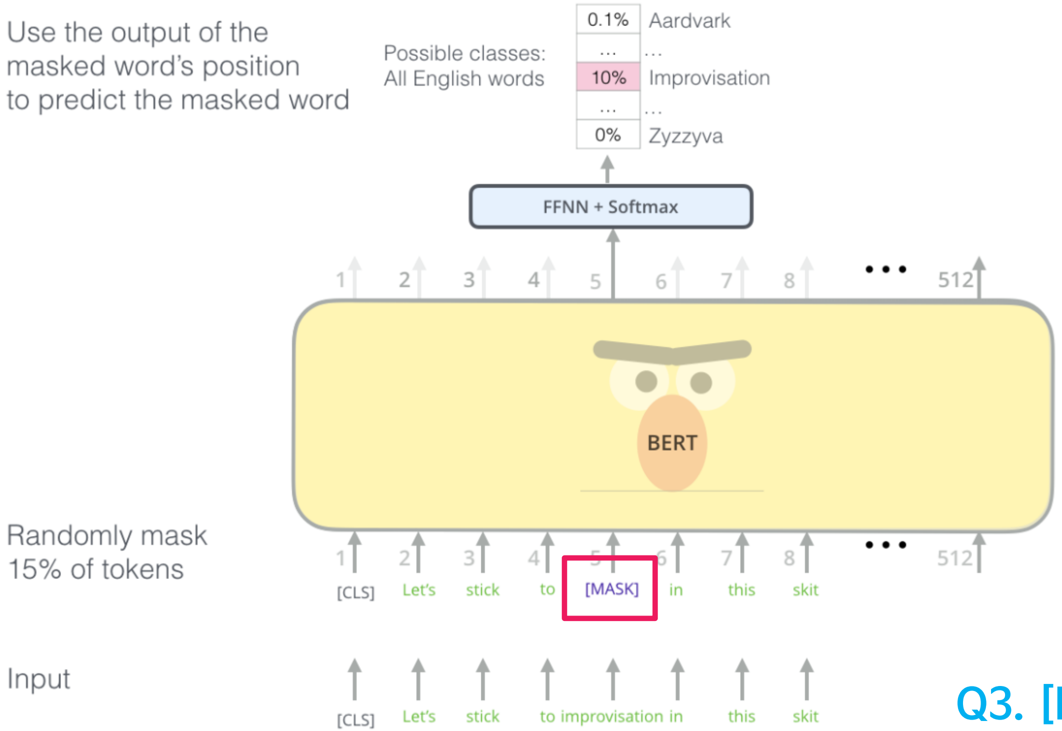


(Pre)Training

- Transformer는 enc-dec 구조로 decoder에서 loss function을 계산해서 training 했음
- BERT는 encoder만 있음.
  (1) Masked Language Model
  (2) Two Sentence Tasks
  2가지 방식을 사용



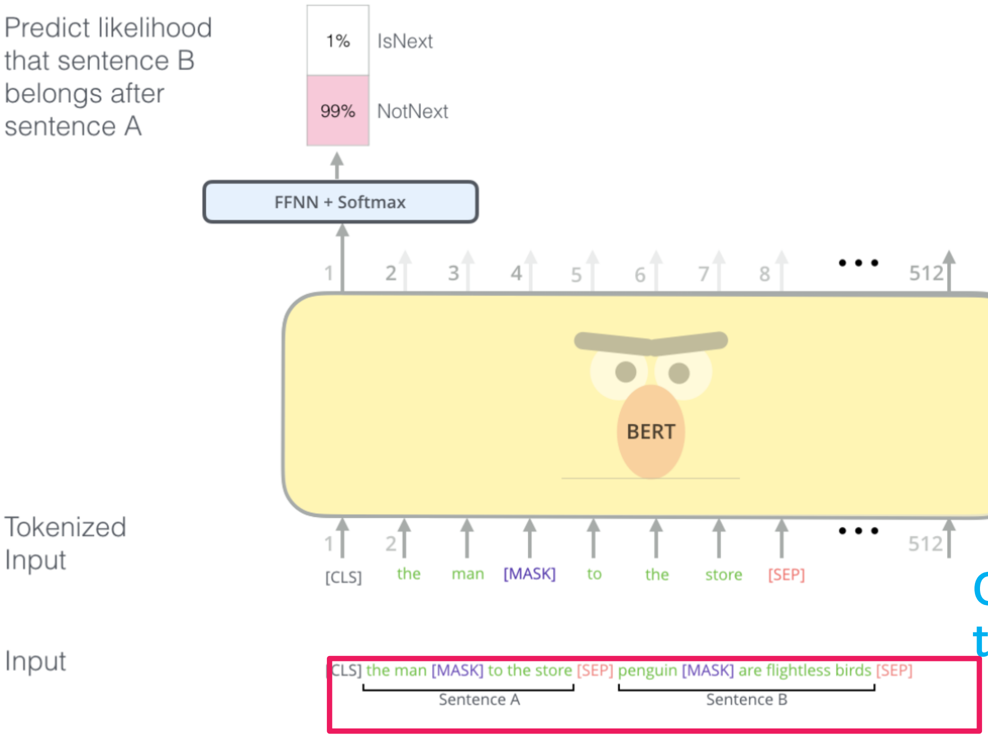
(Pre) Training : Masked LM



- 15%의 token을 random하게 mask로 만듬
- Mask token의
  80%는 mask로
  10%는 random word로
  10%는 unchanged word로
  넣어줌.
  → context를 고려하는 Language model로 학습시키기 위한 것으로 생각됨.
  Mask : 무난한? Context 고려 LM
  random word : 이상한 단어를 넣어도 주변 단어들을 고려하여 정답이 나와야 하니 context를 많이 고려한 LM
  unchanged word : 정답을 넣고 정답이 나오니 context를 덜 고려한 LM
  위 3개의 ensemble 같은 (overfitting방지) 효과가 있지 않을까 추측.

Q3. [Mask] token 을 꼭 넣어주어야 하나?
CBoW 에서는 자기자신을 어떻게 넣어주는가?

(Pre)Training : Two Sentence Tasks



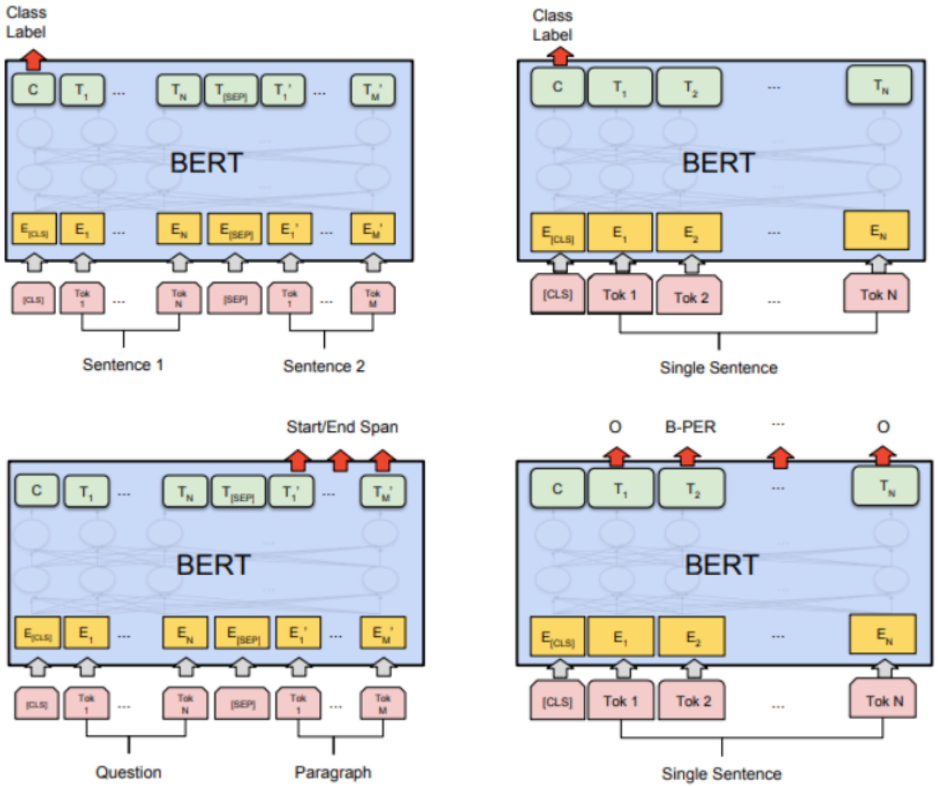
- BERT를 범용적인 NLP 모델로 만들고 싶은데, Mask LM으로 language model은 만들수 있지만, 2개의 문장이 주어지는 task를 해결할 수는 없음. (ex. QA, QQ' similarity 등)
- 입력으로 두 문장을 주고, 이 두 문장이 연속된 의미있는 문장인지를 binary classification

Q4. 512개의 입력 token을 다 못채우는 경우에도, transformer의 오른쪽 노드들은 학습이 잘 되는가?

Fine Tuning : Supervised manner

(a) (b)

(c) (d)



Experiments : 11개의 NLP Tasks

1. MNLI (Multi-Genre Natural Language Inference)
   두 개의 문장을 주고, 두번째 문장이 첫번째 문장과 같은 의미(entailment)인지, 모순(contradiction)이 있는지, 무관한(neutral) 의미인지 판단
2. QQP (Quora Question Pair): 두 질문이 같은 의미의 질문인지 판단
3. QNLI(Question Natural Language Inference): 질문과 문장을 주고, 그 문장에 답이 있으면 pos, 없으면 neg를 판단
4. SST-2 (Stanford Sentiment Treebank): movie review data에서 sentiment 분석(binary)
5. CoLA (Corpus of Linguistic Acceptability) : 문장의 문법이 맞는지 판단
6. STS-B (Semantic Textual Similarity Benchmark) : 두 문장의 의미의 유사도를 1~5 점으로 판단
7. MRPC(Microsoft Research Paraphrase Corpus) : 두 문장의 sentiment가 같은지 판단
8. RTE(Recognizing Textual Entailment) : MNLI 와 비슷
9. SQuAD 1.1 (Stanford Question Answering Database) : 질문을 보고 지문에서 answer text span을 찾아낸다. 즉, 정답의 시작점과 끝점을 찾아냄
10. CoNLL Named Entity Recognition : 단어에 ⟨Person⟩, ⟨Organization⟩, ⟨Location⟩, ⟨Miscellaneous⟩, ⟨Other-Not named entity⟩ 를 annotate
11. SWAG(Situation with Adversarial Generations) : video captioning DB에서 추출한 문장 다음에 올 문장으로 알맞은 것은? (4지선다형)

11개 Tasks에서 SOTA

모두의연구소

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	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD results. The BERT ensemble is 7x

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

Table 3: CoNLL-2003 Named Entity Recognition re-

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Table 4: SWAG Dev and Test accuracies. Test results

김승일 연구소장

E-mail : si.kim@modulabs.co.kr
Blog : www.whyDSP.org
FB: www.facebook.com/lab4all
www.facebook.com/groups/modulabs