AILAB SEMINAR #19

BERT

220223 한양대학교 인공지능연구실 김민지

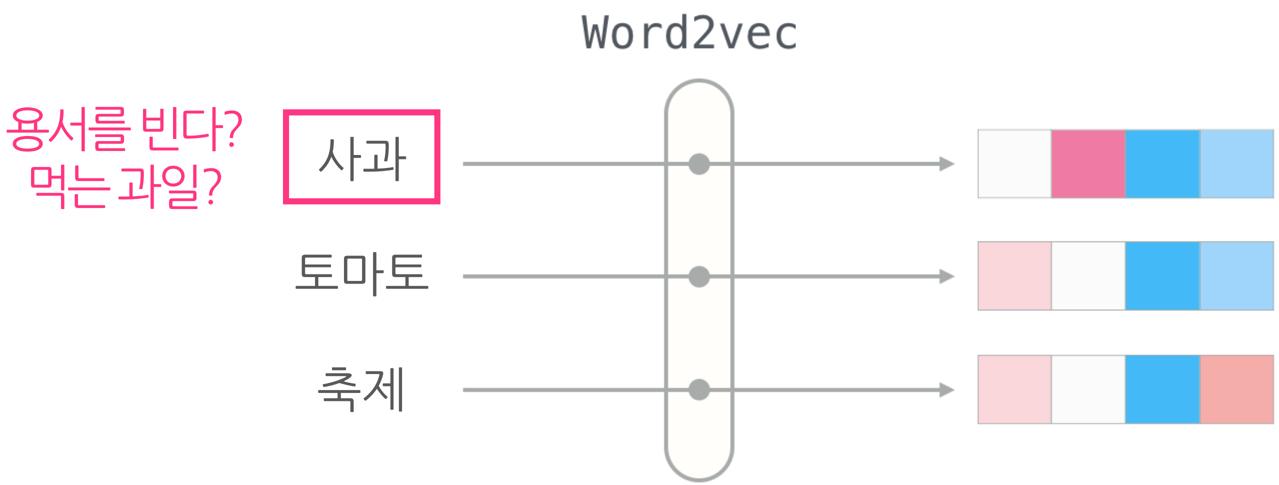
1

PLM

Pre-trained Language Model

(Previous) Word2Vec

- 하나의 단어를 하나의 Vector로 Mapping
- 다의어나 동음이의어를 구분하지 못한다는 한계가 존재

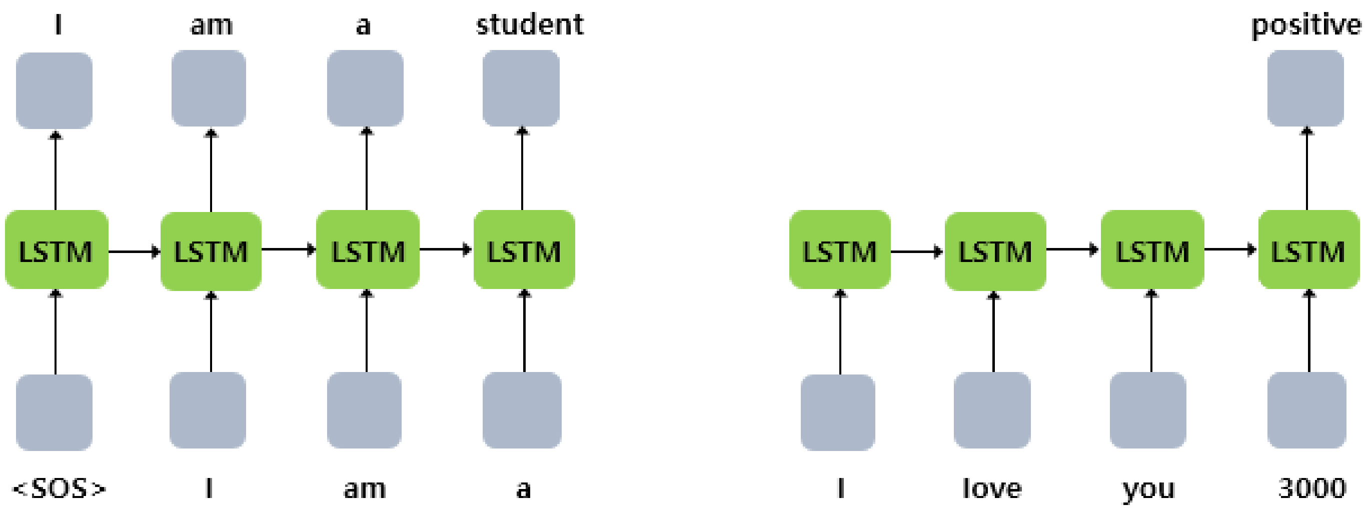


PLM

BERT

Pre-trained Language Model - LSTM

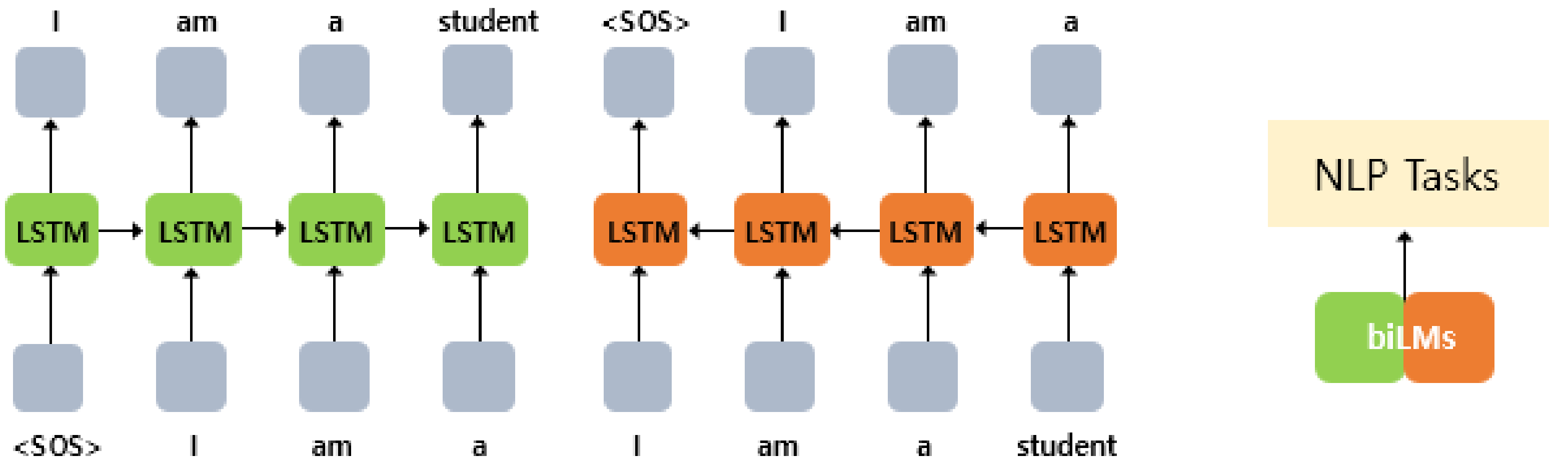
- Semi-supervised Sequence Learning, Google, 2015
- LSTM 언어 모델을 학습(Unlabeled Data)하고 나서 Text 분류 등 Task에 대해 추가 학습



Pre - training　　Fine tuning

Pre-trained Language Model - ELMo

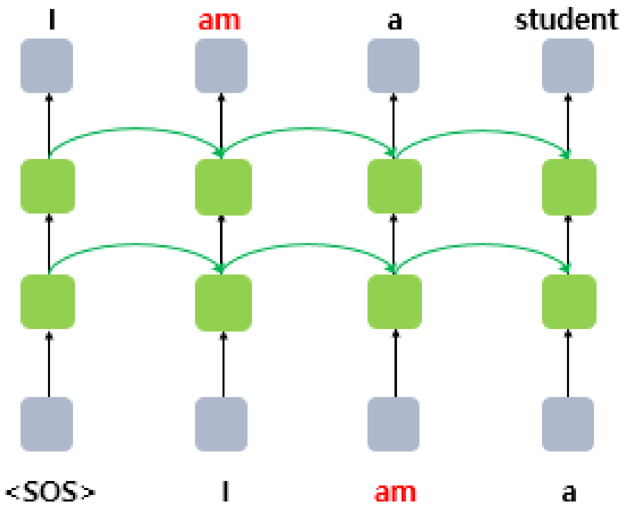
- ELMo: Deep contextualized word representations, AI2 & Univ. of Washington, 2017
- 순방향 언어 모델과 역방향 언어 모델을 각각 Pre-training 후 Input으로 사용



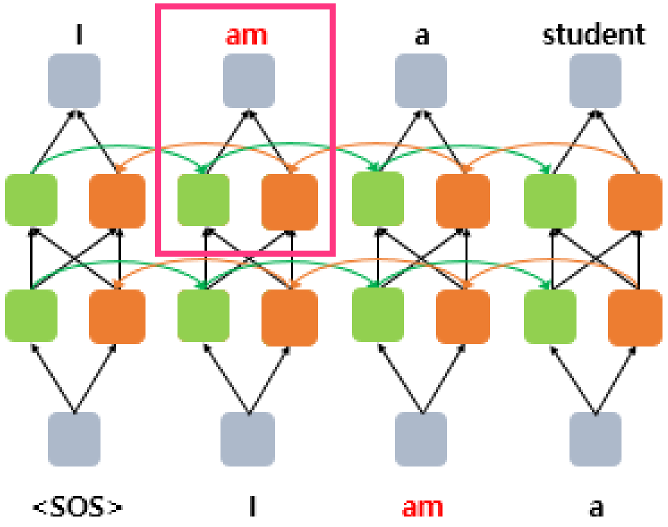
순방향 언어모델 역방향 언어모델 Embedding Vector와 사용

Bidirectional 언어 모델?

- 양방향 Cell로 언어 모델을 구현하면 자기 자신을 보고 예측하는 것과 같음
- 이전 단어를 보고 다음 단어를 예측하는 언어 모델에는 적합하지 않음



Unidirectional LM



Bidirectional LM

2

BERT

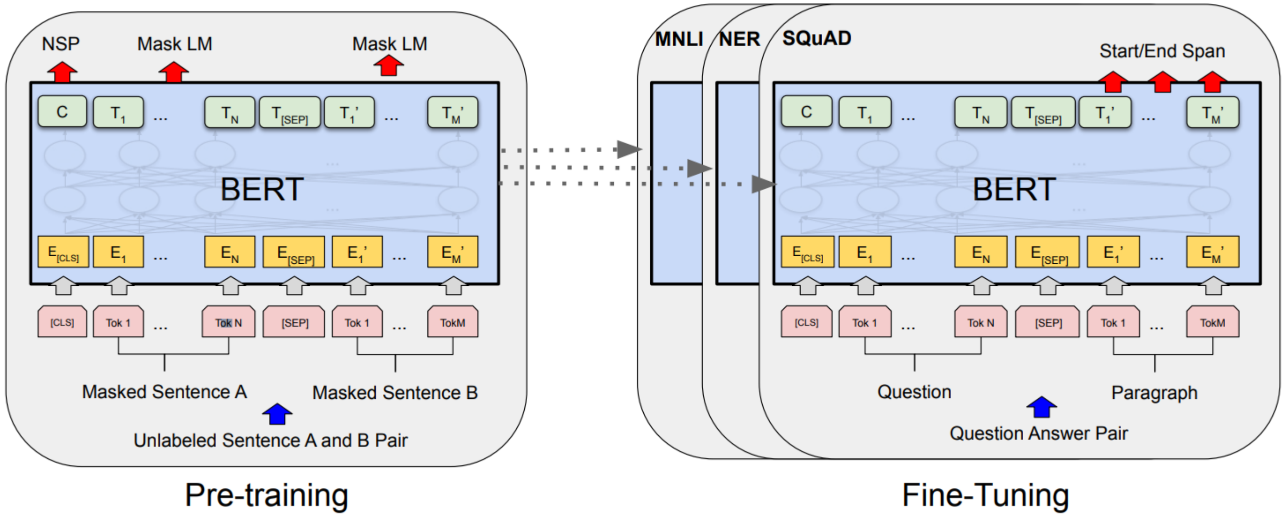
Bidirectional Encoder Representations from Transformers

BERT - Bidirectional Encoder Representations from Transformers

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019
- Pre-train deep bidirectional representation을 위해 제시
- Unsupervised Fine-tuning Approach 기반의 language representation model

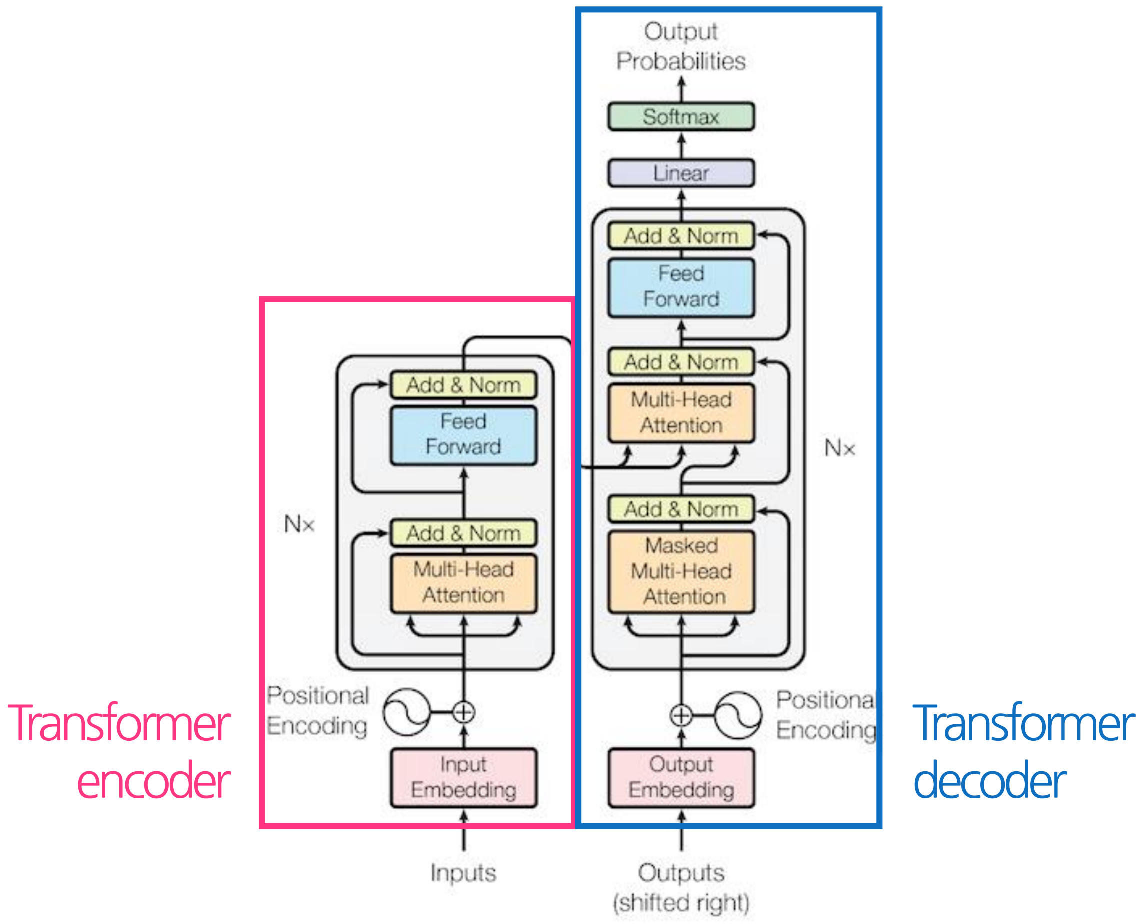
나는 Bert라고 해!

안녕, 난 Elmo야!



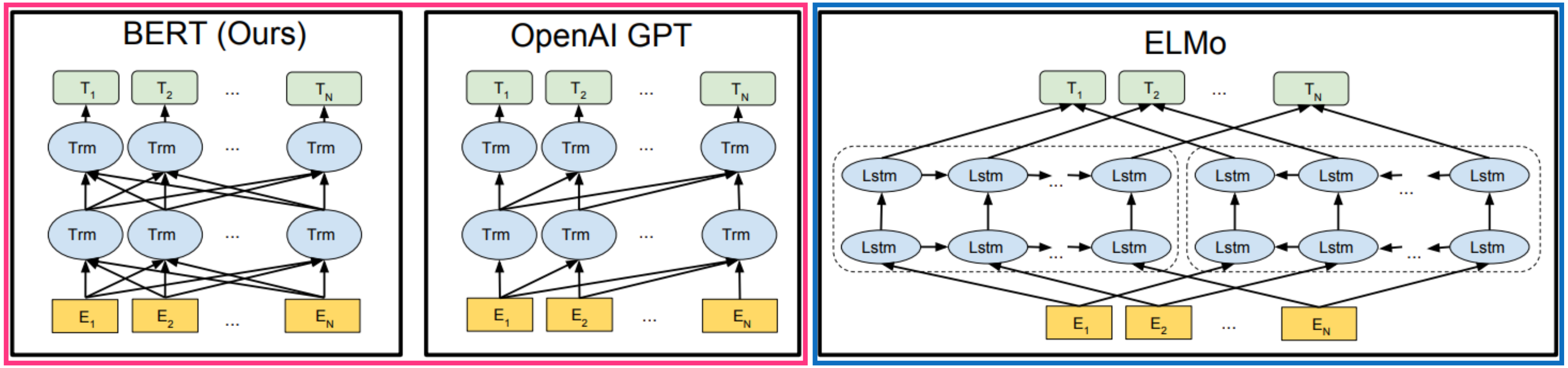
Model Architecture

- *Tensor2tensor*의 Multi-layer bidirectional Transformer encoder를 사용



- ***BERT$_{BASE}$*** (L=12, H=768, A=12, Total Parameters=110M)
- ***BERT$_{LARGE}$*** (L=24, H=1024, A=16, Total Parameters=340M)
- GPT와의 비교를 위한 Base model 제시

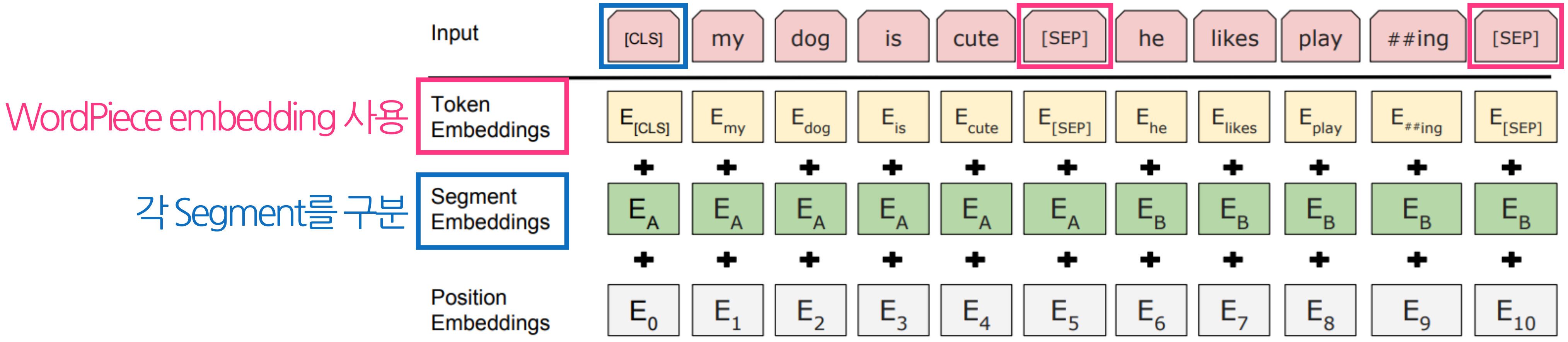
Fine-tuning approach   Feature-based approach



Bidirectional (Encoder)   Left-to-right (Decoder)   Concatenation of left-to-right and right-to-left LSTMs

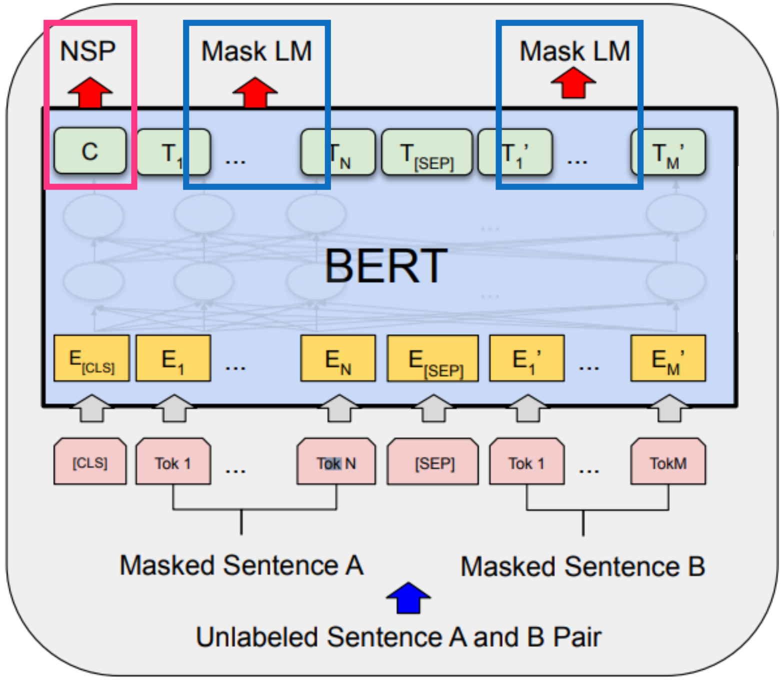
Input / Output Representations

- "Sentence" : 임의의 연속적인 text (문장, 구 등)
- "Sequence" : 1개 또는 2개의 sentence로 구성된 Input token들
- [CLS] : Special classification token (Sequence의 처음에 위치)
- [SEP] : Special separation token (Sentence 구분)



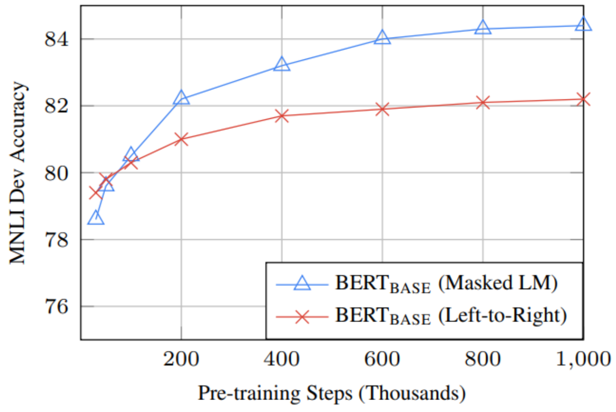
Pre-training BERT

- Deep bidirectional representation을 학습하기 위해 두 가지 Task를 정의
- *BookCorpus* (800M words)와 *English Wikipedia* (2,500M words) 사용
- Task #1: Masked LM (MLM)
- Task #2: Next Sentence Prediction (NSP)



Task #1: Masked LM

- 각 Input token sequence 의 15%를 Random하게 Masking
- Final hidden state에서 Masked Token을 예측



Example My dog is $\boxed{\text{hairy}}$ 15%

80% of the time My dog is [MASK] Replace the word with [MASK] token

10% of the time My dog is apple Replace the word with a random word

10% of the time My dog is hairy Keep the word unchanged

Task #2: Next Sentence Prediction (NSP)

- Question Ansering(QA) & Natural Language Interference(NLI): 문장 간 관계 학습이 필요
- Sentence A와 B를 두 가지 방식으로 구성, Final hidden vector의 C = IsNext / NotNext

Actual Next sentence (50%)

Input = `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]`

Label = `IsNext`

Random Sentence (50%)

Input = `[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]`

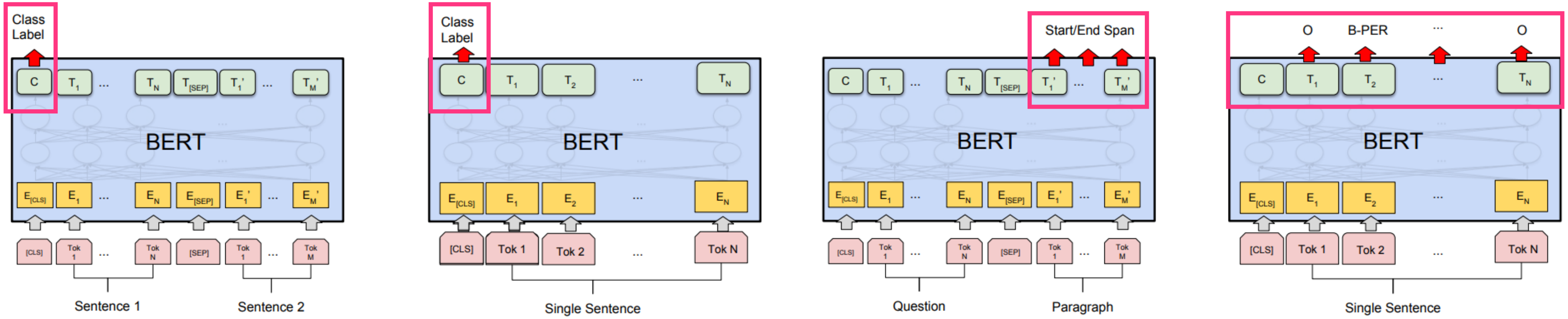
Label = `NotNext`

Fine-tuning BERT

PLM

BERT

- 각 Task에 맞게 Input과 Output을 도입하여 Fine-tuning 후 사용
- Single sentence일 때와 Multiple sentence일 경우 모두를 Self-attention으로 처리
- (Bidirectional Cross Attention 과정을 따로 거칠 필요가 없음)
- Pre-training 에 비해 비교적 Inexpensive



(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

- GLUE Test Results
- https://gluebenchmark.com/leaderboard

F1 Spearman correlation F1

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**

PLM

BERT

Experiments - SQuAD v1.1

- QA Task, The Stanford Question Answering Dataset
- Ablation Study에서 Pre-training의 NSP가 유의미함을 보임

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$BERT_{BASE}$ (Single)	80.8	88.5	-	-
$BERT_{LARGE}$ (Single)	84.1	90.9	-	-
$BERT_{LARGE}$ (Ensemble)	85.8	91.8	-	-
$BERT_{LARGE}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$BERT_{LARGE}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$BERT_{BASE}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Conclusion

- Performance 측면에서 대단한 Model을 제시함

- Unsupervised Pre-training를 수행하고 나면 (Google TPU로 수 일 소요)
- 특정 Task를 위한 Fine-tuning의 비용이 적게 필요하다는 장점

- Bidirectional Architecture를 적합한 Task를 정의하여 제시함

Any Question?

Thank you

들어주셔서 감사합니다

References

- https://jalammar.github.io/illustrated-word2vec/
- https://wikidocs.net/108730
- https://arxiv.org/pdf/1810.04805.pdf
- https://nlp.seas.harvard.edu/2018/04/03/attention.html
- https://github.com/SKTBrain/KoBERT#naver-sentiment-analysis