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

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language

NACCL ’19

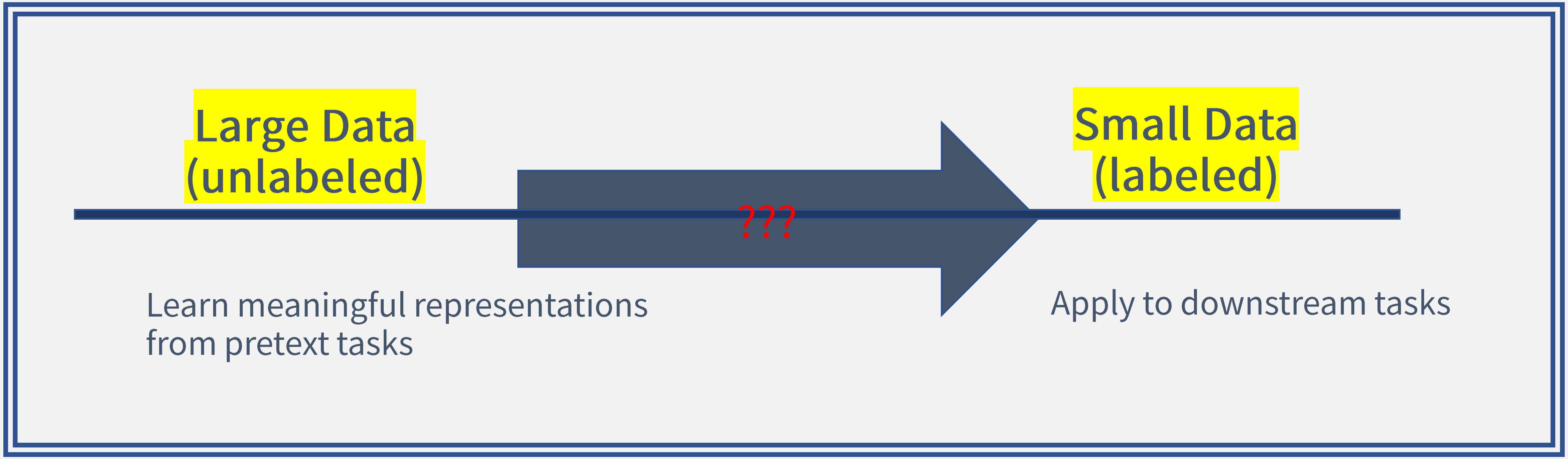
Noah Lee (noahlee357@korea.ac.kr)
2021.09.23

01 Problem Statement

How can we further develop current pre-training methods for language representations?

02 Motivation

Semi-Supervised Learning



02 Motivation

Self-Supervised Learning

Large Data (unlabeled)

???

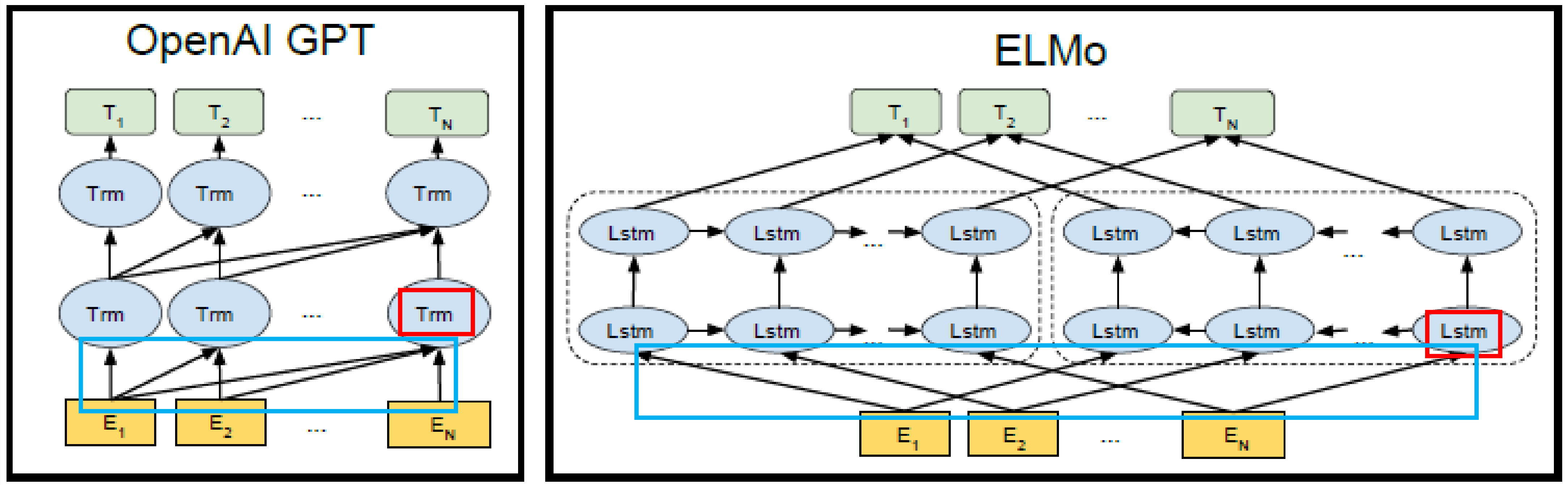
Small Data (labeled or unlabeled)

Learn meaningful representations from **pretext tasks**…
without human supervision!

Apply to **downstream tasks**

02 Motivation

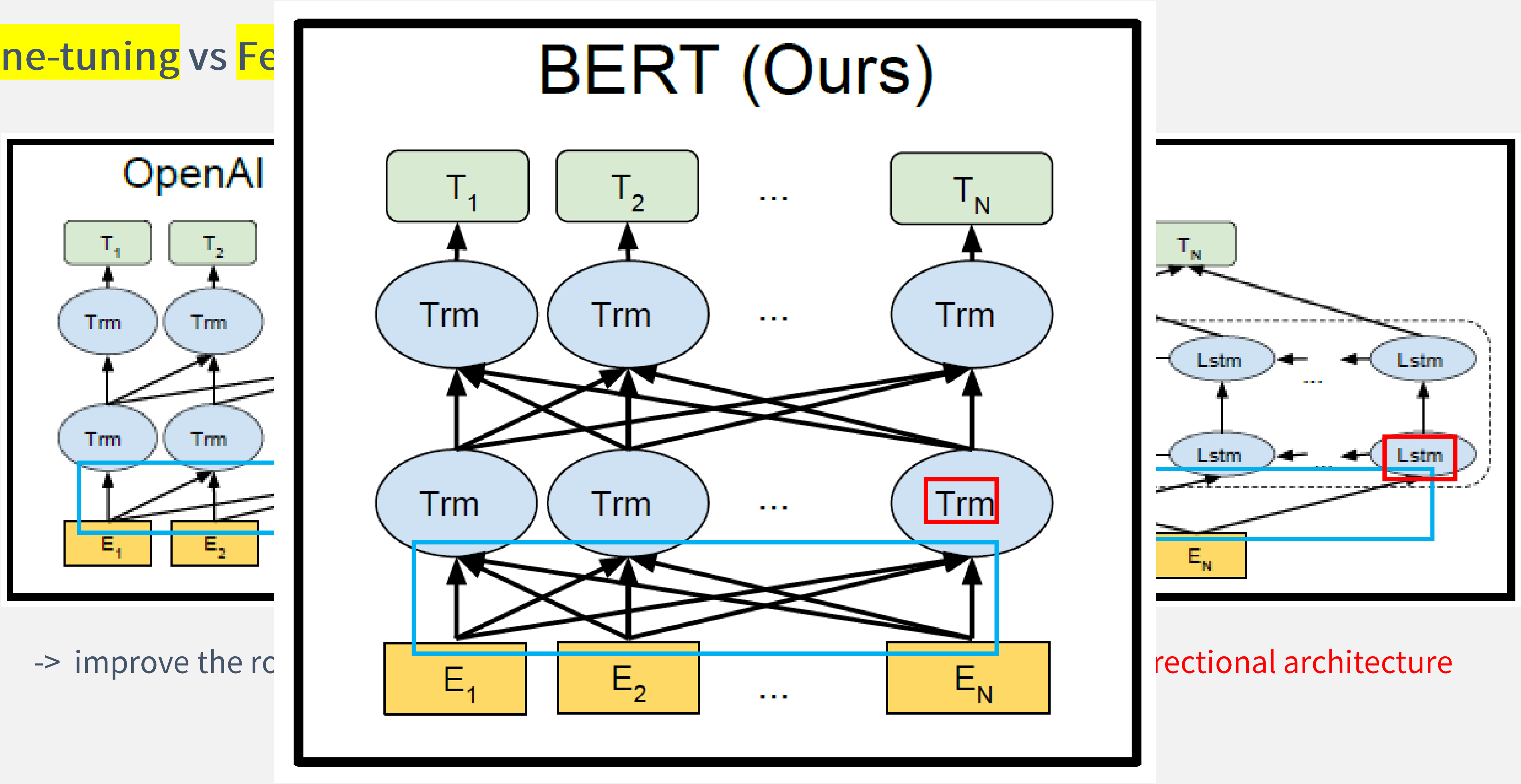
Fine-tuning vs Feature-Based



-> improve the robustness of text generation models by fine-tuning + deeply bidirectional architecture

02 Motivation

Fine-tuning vs Fe



-> improve the ro

rectional architecture

03 Contribution

1) Presents BERT to demonstrate the importance of **bidirectional pre-training** for language representations

2) Shows how pre-trained representations **reduce the need for task-specific architectur**e engineering

3) Advances the state of the art(SOTA) performances for **11 NLP tasks**

04 Datasets

Sentence Pair CLS
Single Sentence CLS
Question Answer
Single Sentence Tagging

[Pre-training Datasets]

- BookCorpus (800M words)
- English Wikipedia (2500M words)

[Fine-tuning Datasets 1/2]

- GLUE(General Language Understanding Evaluation): except WNLI*
  - MNLI : entailment classification
  - QQP : binary CLS if semantically equivalent
  - QNLI : SQuAD -> binary CLS, whether correct answer
  - SST-2 : binary single-sentence CLS for sentiment
  - CoLA : binary single- sentence whether sentence linguistically acceptable
  - STS-B : how two sentences are semantically alike from 1 to 5
  - MRPC : whether sentences in the pair are semantically equivalent
  - RTE: binary entailment task but with less training data
  - WNLI: small inference dataset…*

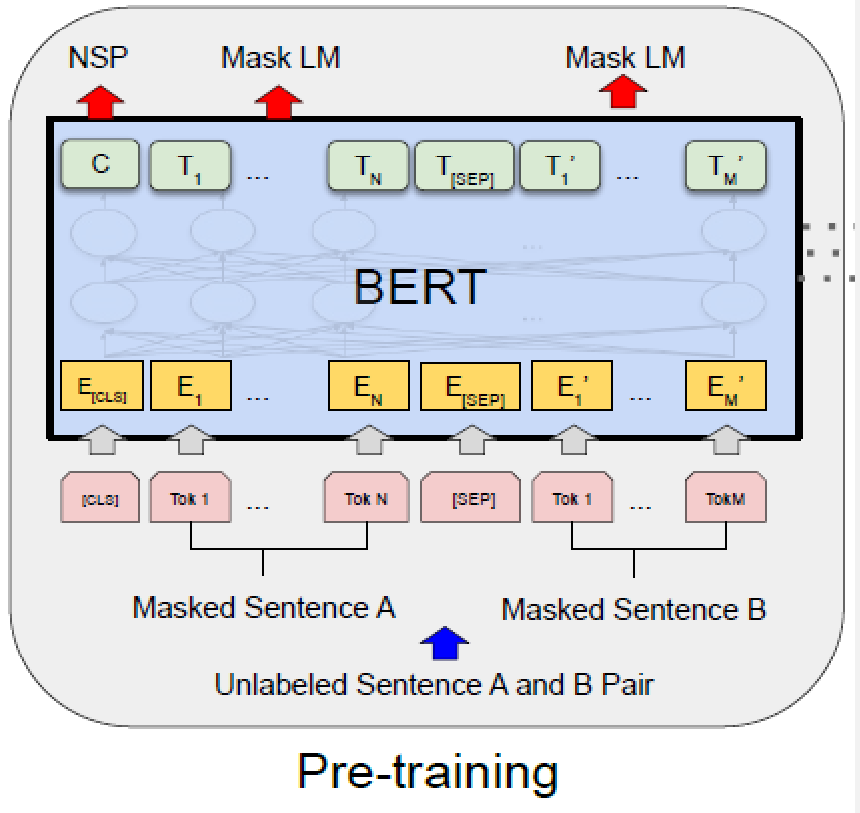
* Excluded due to poor performance (likely by the adversarial examples with shared sentences)

[Fine-tuning Datasets 2/2]

- SQuAD v1.1
  - QA dataset from 100k crowd-sourced QA pairs
  - Predict the answer text span
- SQuAD v2.0
  - Extension of v1.1 : more realistic as to allowing “no answer”
- SWAG (Situations With Adversarial Generations)
  - 113k sentence-pair completion
  - 1 given sentence : four possible continuation

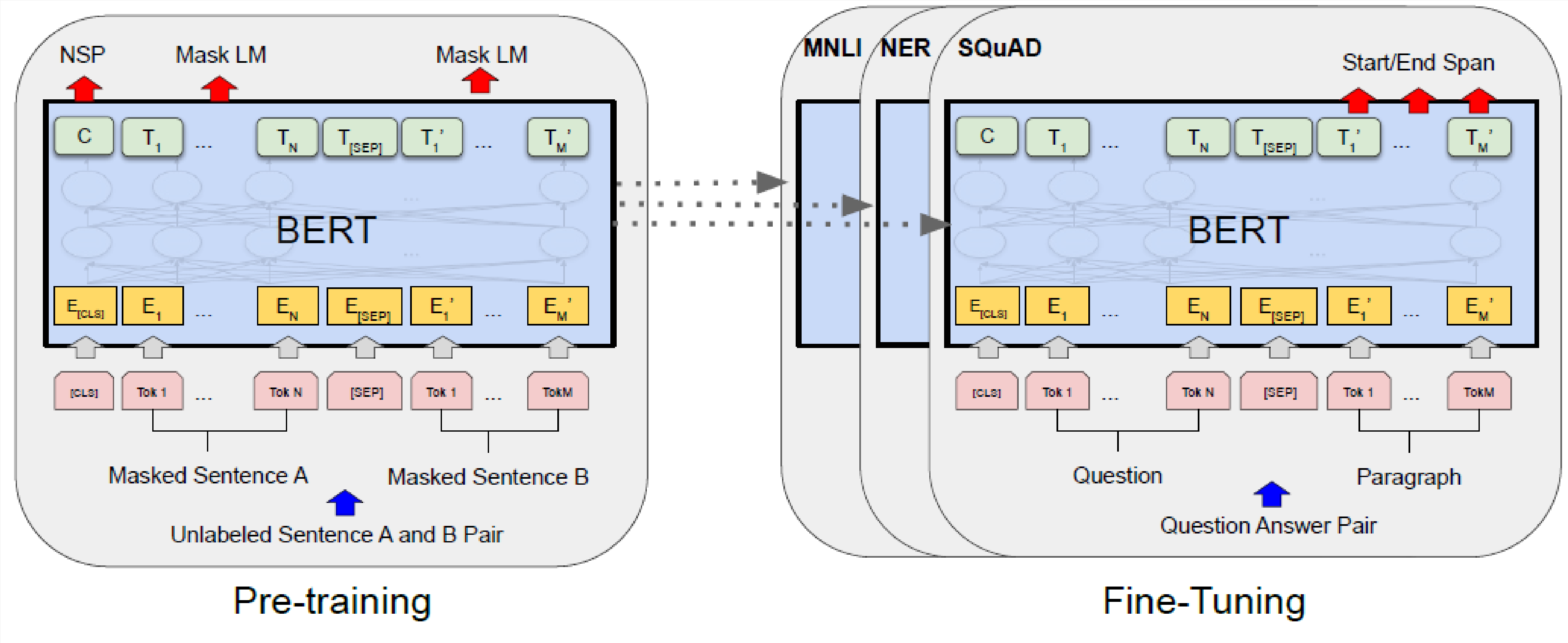
BERT Overview

- Framework : pre-training -> fine-tuning
- For fine-tuning, utilize unified architecture over most tasks (minimal changes)
- Multi-layer bidirectional Transformer encoder based
- BERT_BASE (L=12, H=768, A=12, Total Parameters=110M)
- BERT_LARGE (L=24, H=1024, A=16, Total Parameters=340M)



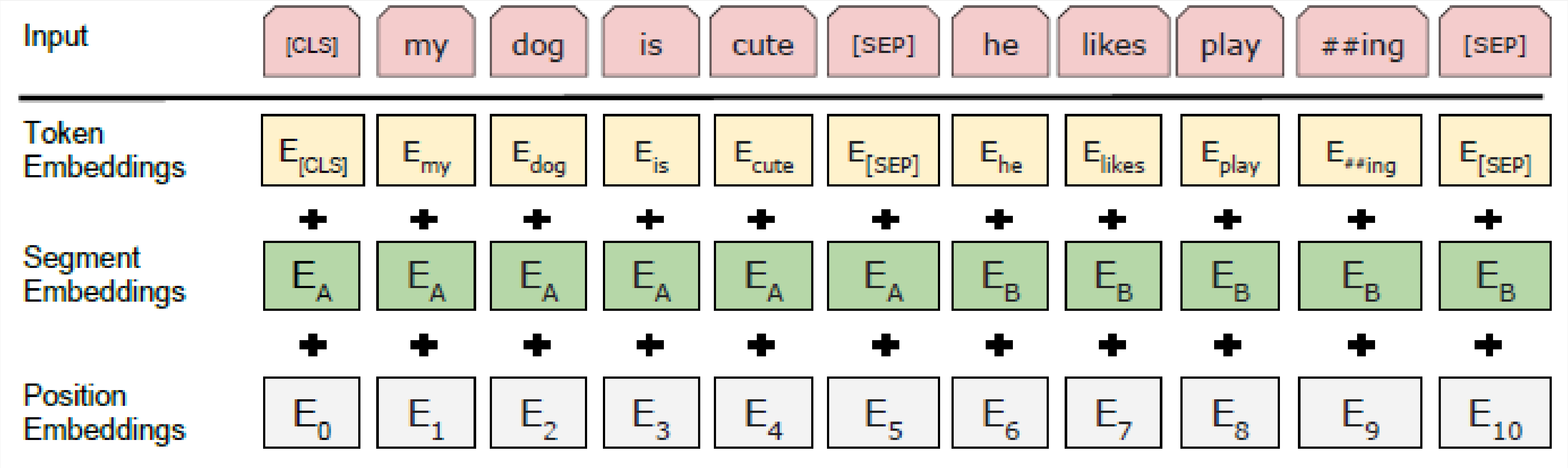
05 Model

BERT Overview



05 Model

Input Embedding



Sum of emb_token, emb_segment, emb_position

05 Model

Pre-training

1) Masked LM

80% of the time: Replace the word with the [MASK] token, e.g., my dog is hairy → my dog is [MASK]

10% of the time: Replace the word with a random word, e.g., my dog is hairy → my dog is apple

10% of the time: Keep the word unchanged, e.g., my dog is hairy → my dog is hairy. The purpose of this is to bias the representation towards the actual observed word.

2) Next Sentence Prediction (NSP)

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

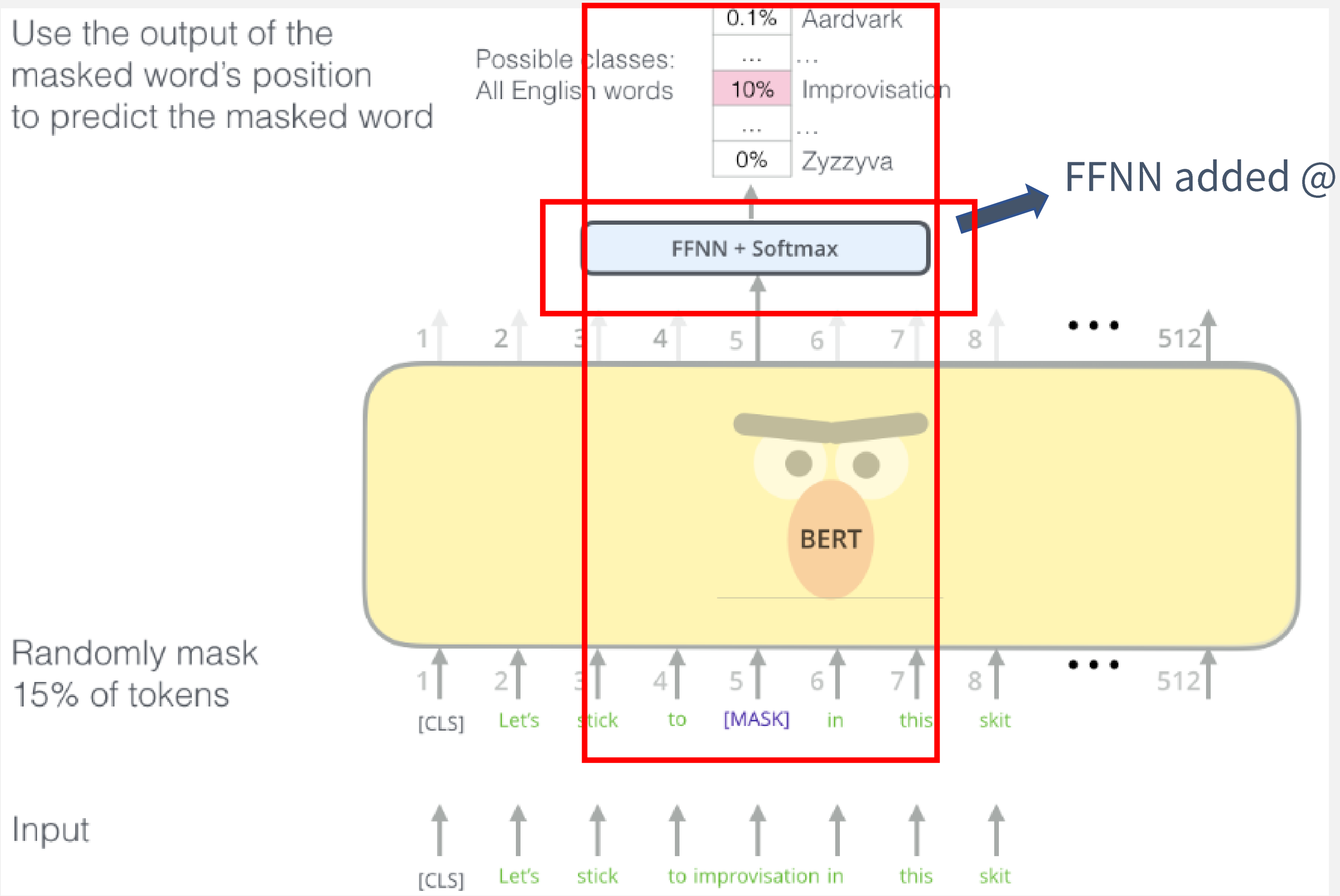
Label = IsNext

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

Label = NotNext

05 Model

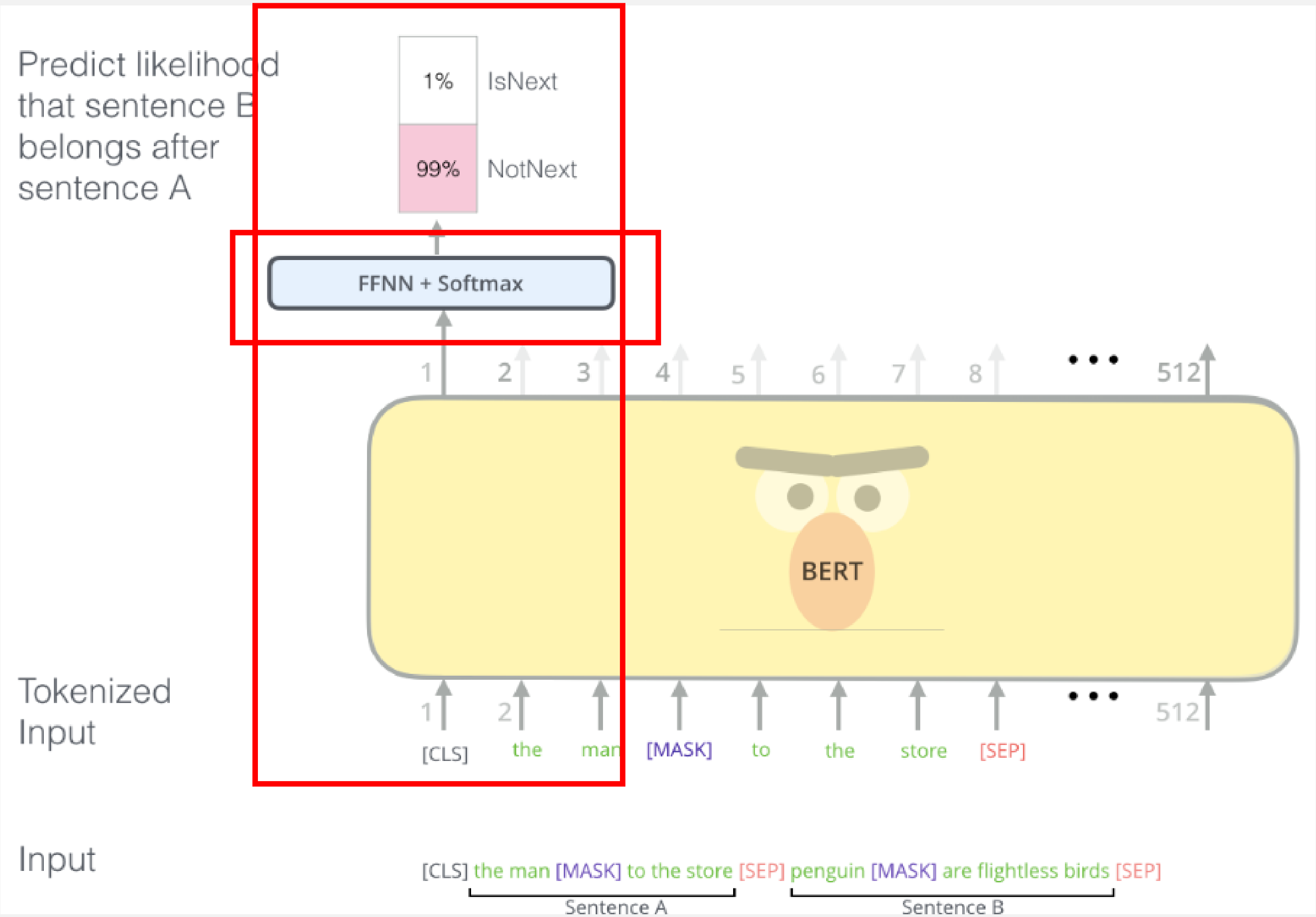
Pre-training Outlook: MLM



FFNN added @ specific position

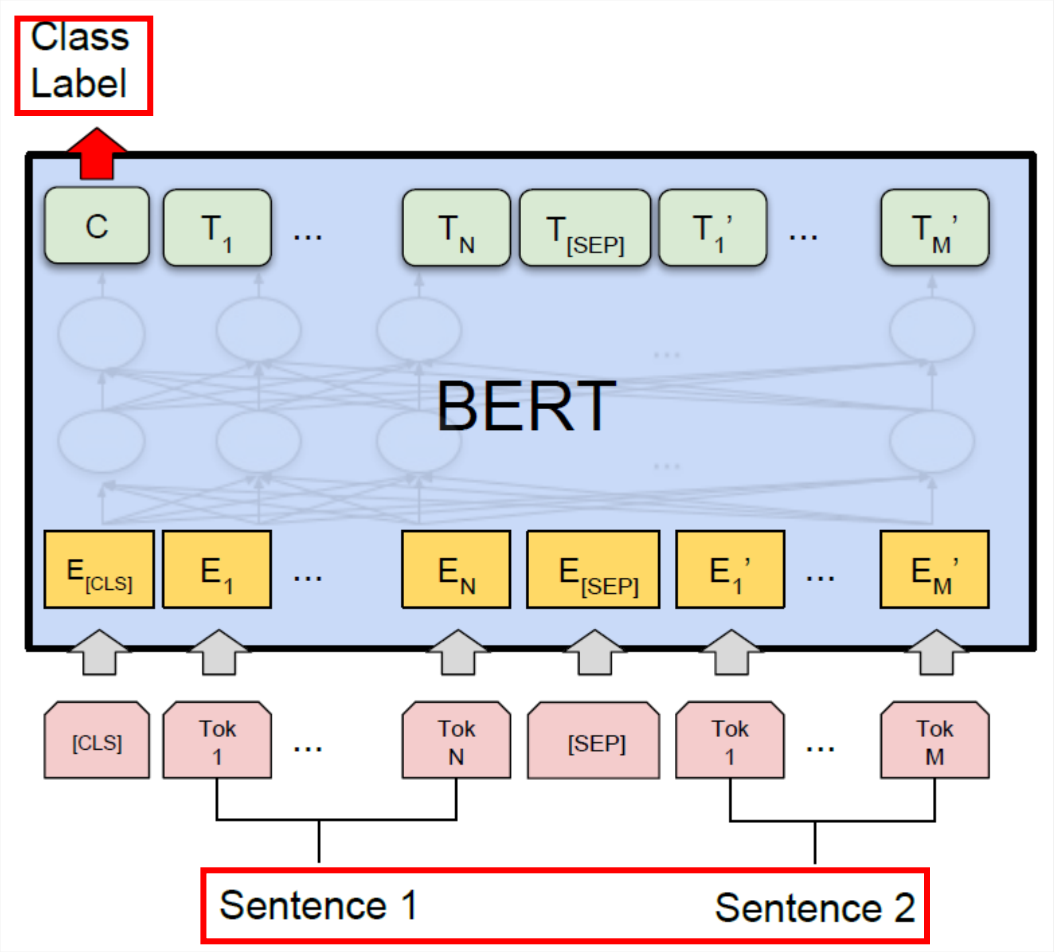
05 Model

Pre-training Outlook: NSP



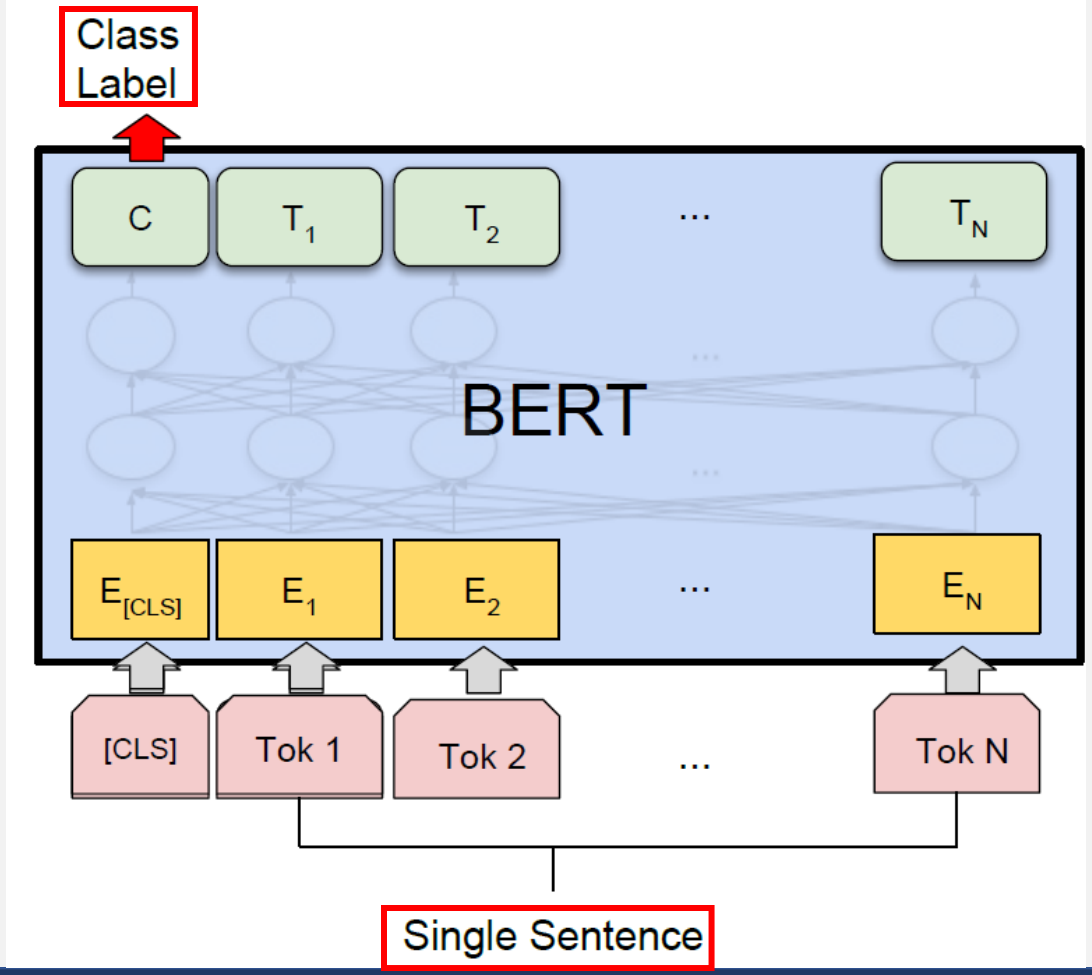
05 Model

Fine-tuning: Sentence Pair CLS



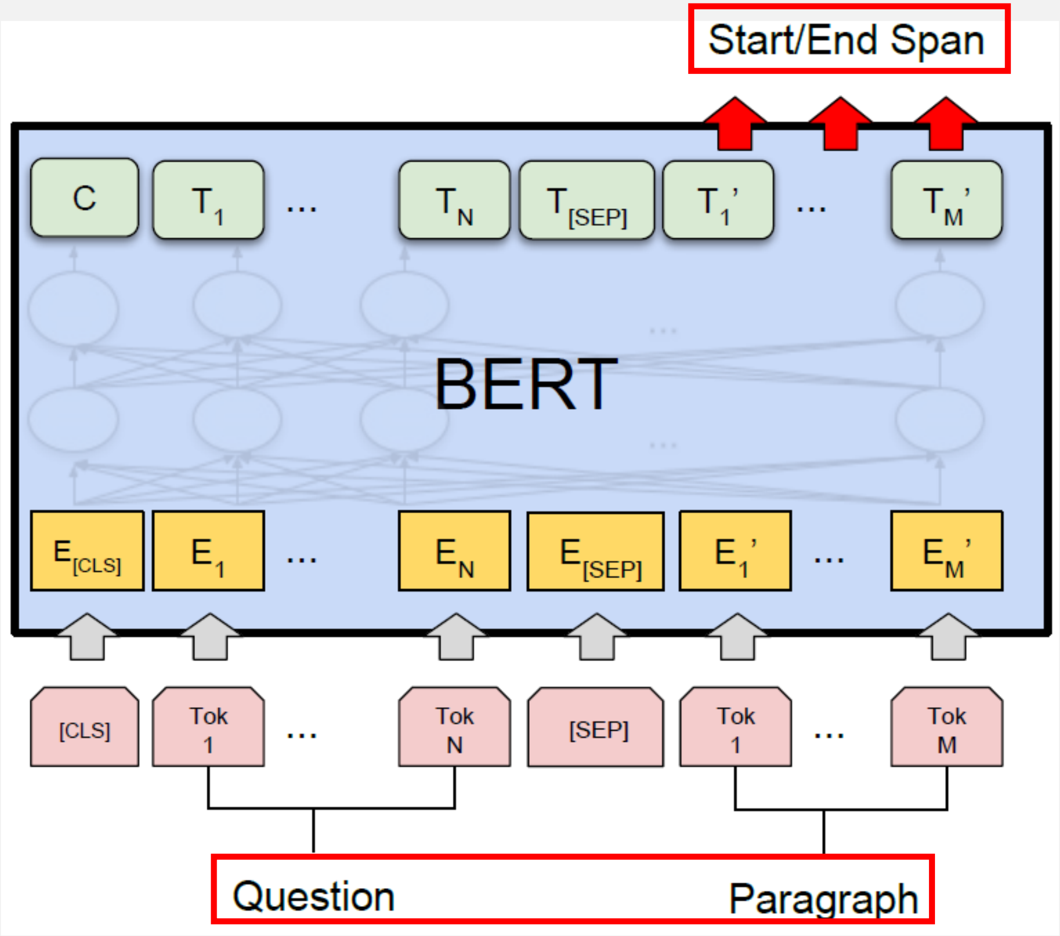
05 Model

Fine-tuning: Single Sentence CLS



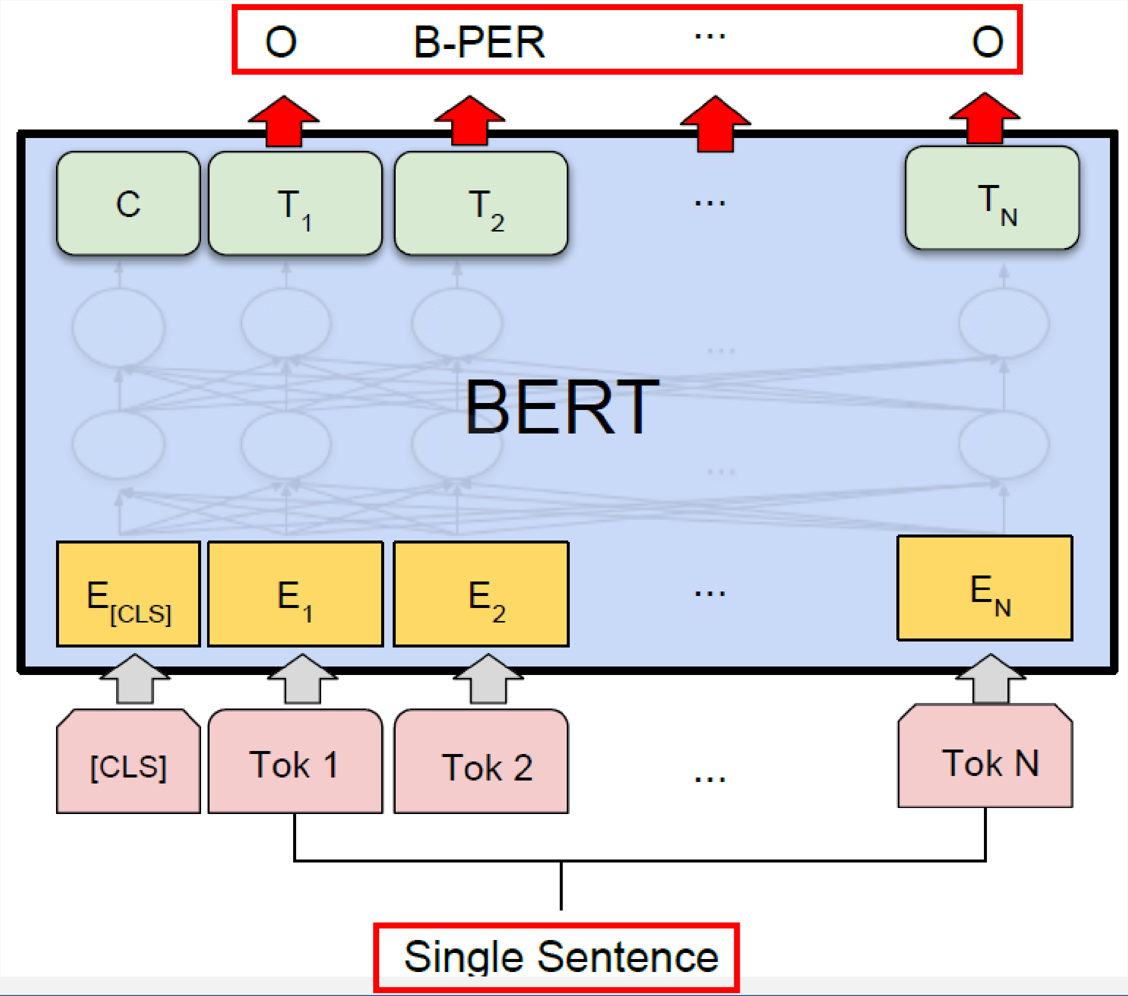
05 Model

Fine-tuning: QA Task



05 Model

Fine-tuning: Single Sentence Tagging



06 Experiment

Downstream Performance: GLUE

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**

06 Experiment

Downstream Performance: SQuAD 1.1, 2.0

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				
$\text{BERT}_{\text{BASE}}$ (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**

SQuAD 1.1

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

SQuAD 2.0

06 Experiment

Downstream Performance: SWAG

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Considerable performance…

07 Conclusion

1) Presents BERT to demonstrate the importance of **bidirectional pre-training** for language representations

2) Shows how pre-trained representations **reduce the need for task-specific architectur**e engineering

3) Advances the state of the art(SOTA) performances for **11 NLP tasks** (8 GLUE tasks, SQuAD 1.1&2.0, SWAG)

Q&A