Paper Review #3

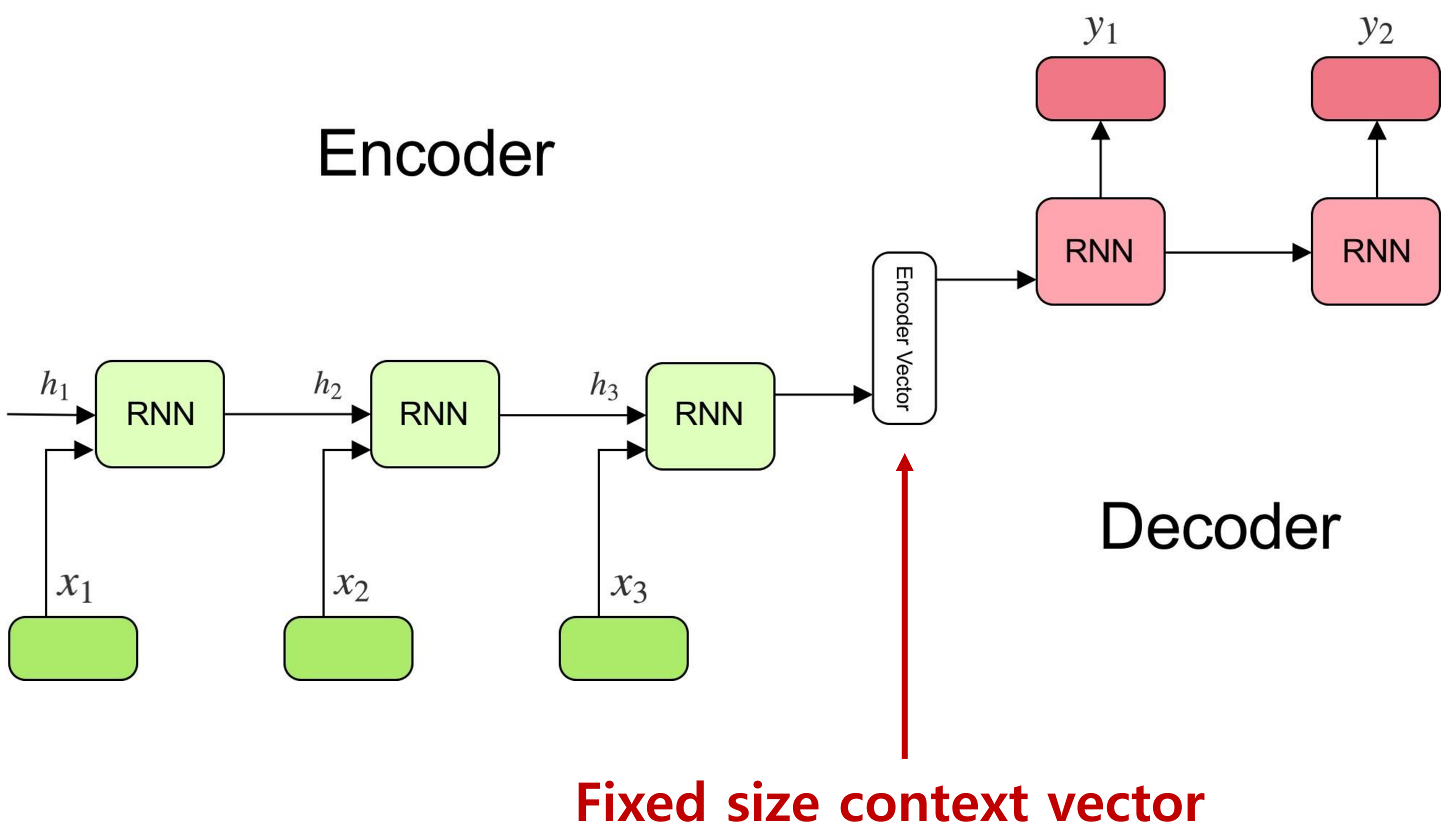
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Dept. of Computer Science & Engineering
201502755 Meeyun Kim

Contents

1. Introduction

RNN based Encoder & Decoder

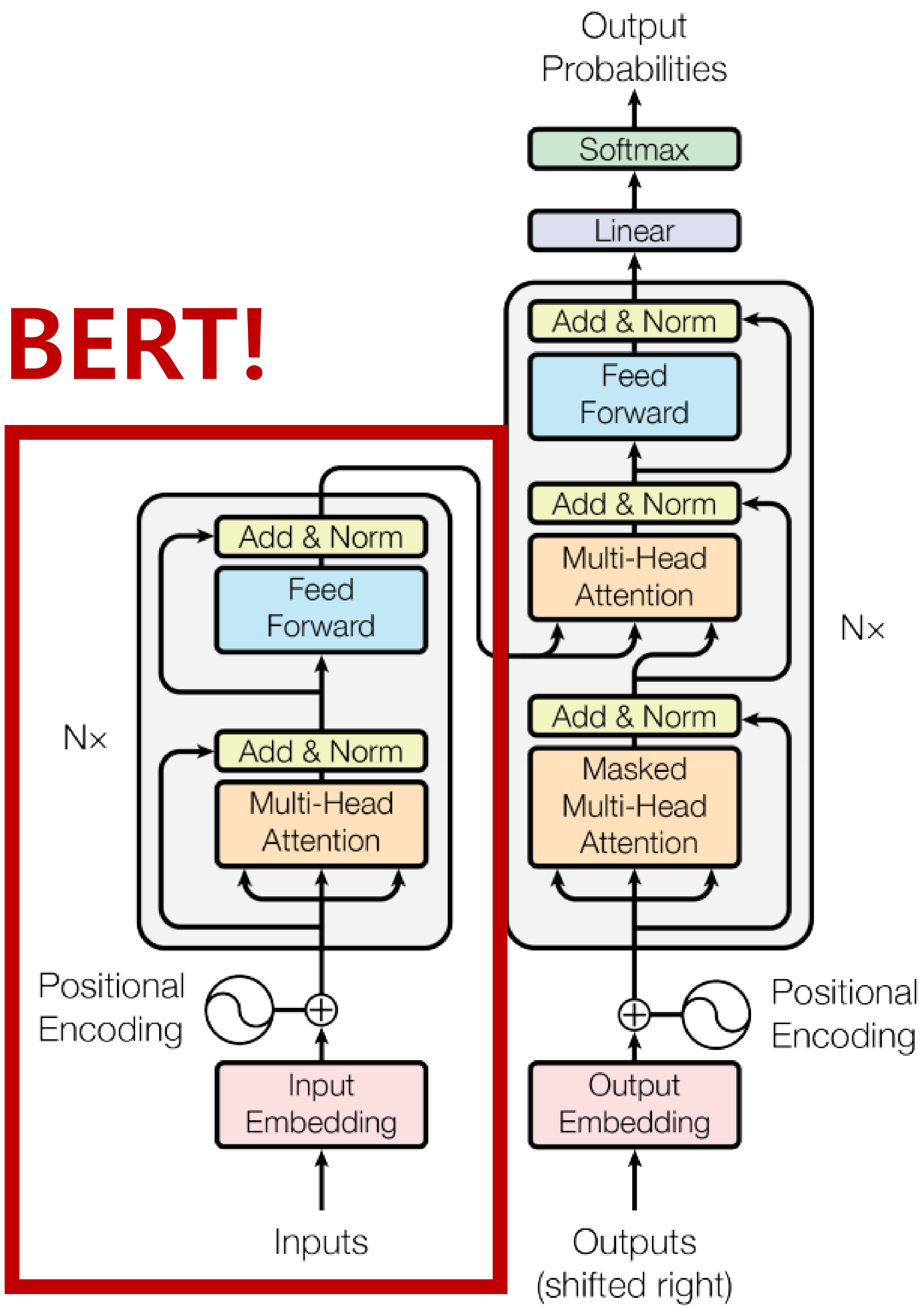


Transformer

- Encoder + Decoder

- Without **RNN**, Only **Attention!**

- Adopt **Parallelization**
  → **Dot Production**



2. About BERT

1) What is BERT?

Bidirectional **E**ncoder **R**epresentations from **Transformers**

cf> **OpenAI GPT →** Decoder **(**left-to-right, **unidirectional)**

2) Model Architecture

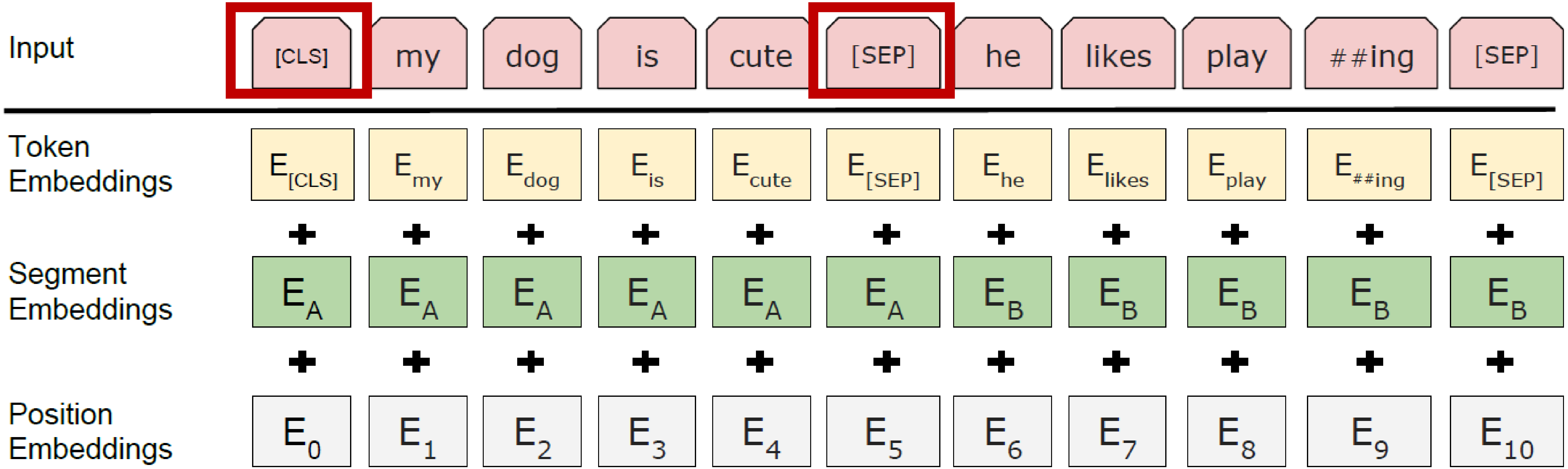
- **$\text{BERT}_{\text{BASE}}$: L=12, H=768, A=12, Total Parameters=110M**

- **$\text{BERT}_{\text{LARGE}}$: L=24, H=1024, A=16, Total Parameters=340M**

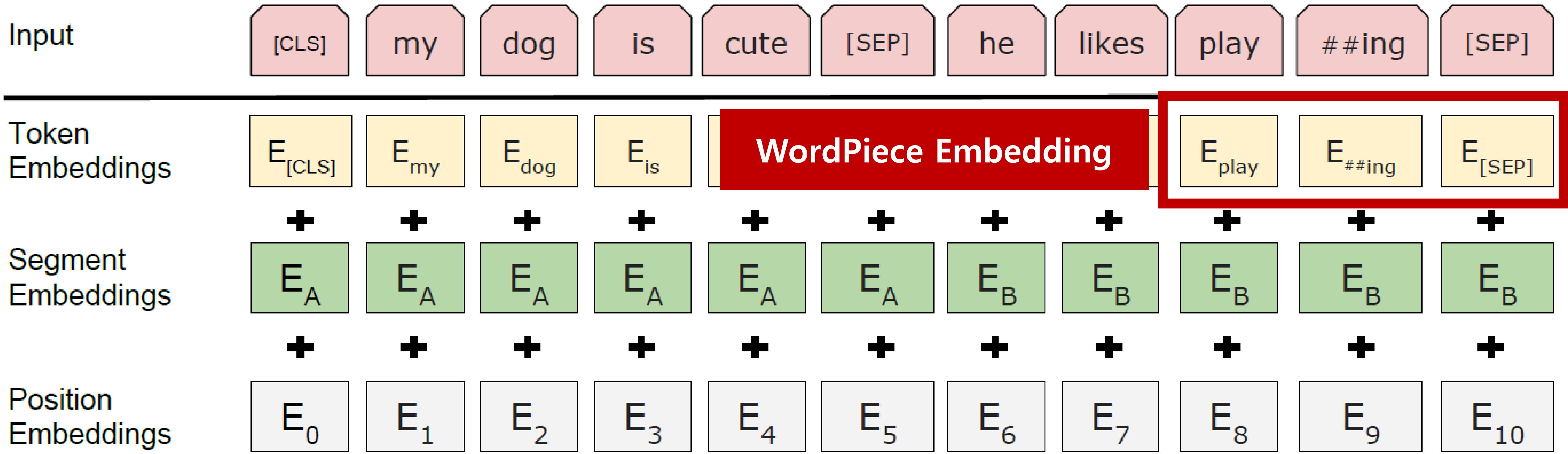
(number of layers - **L**, the hidden size - **H**, the number of self-attention heads - A)

- **BookCorpus**(800M words) + **English Wikipedia**(2,500M words)

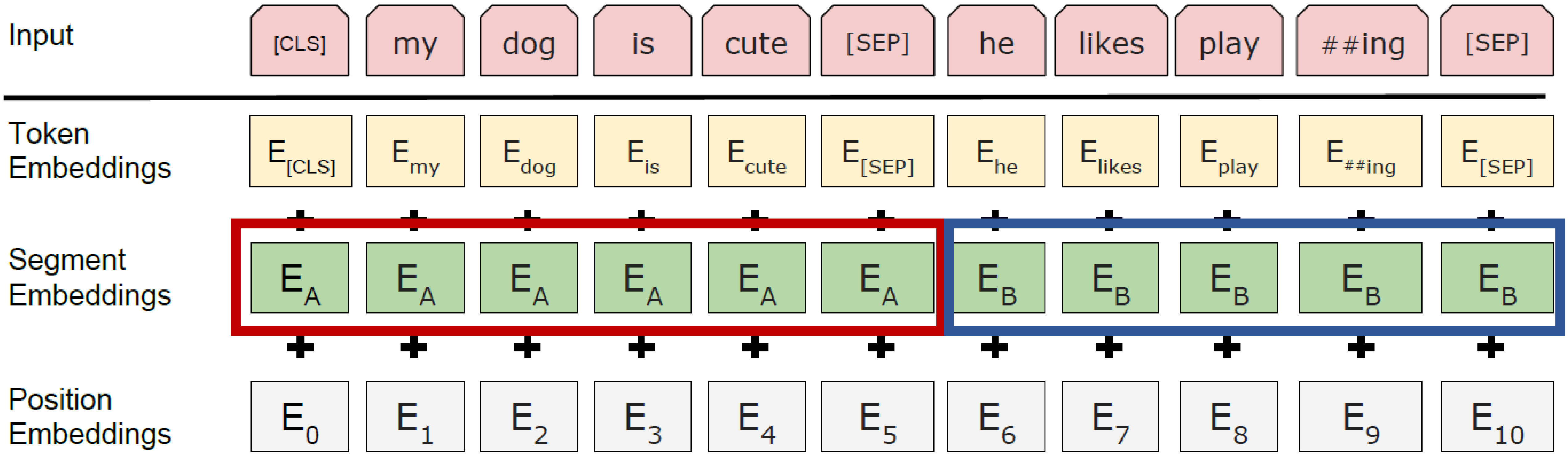
2) Input Representation



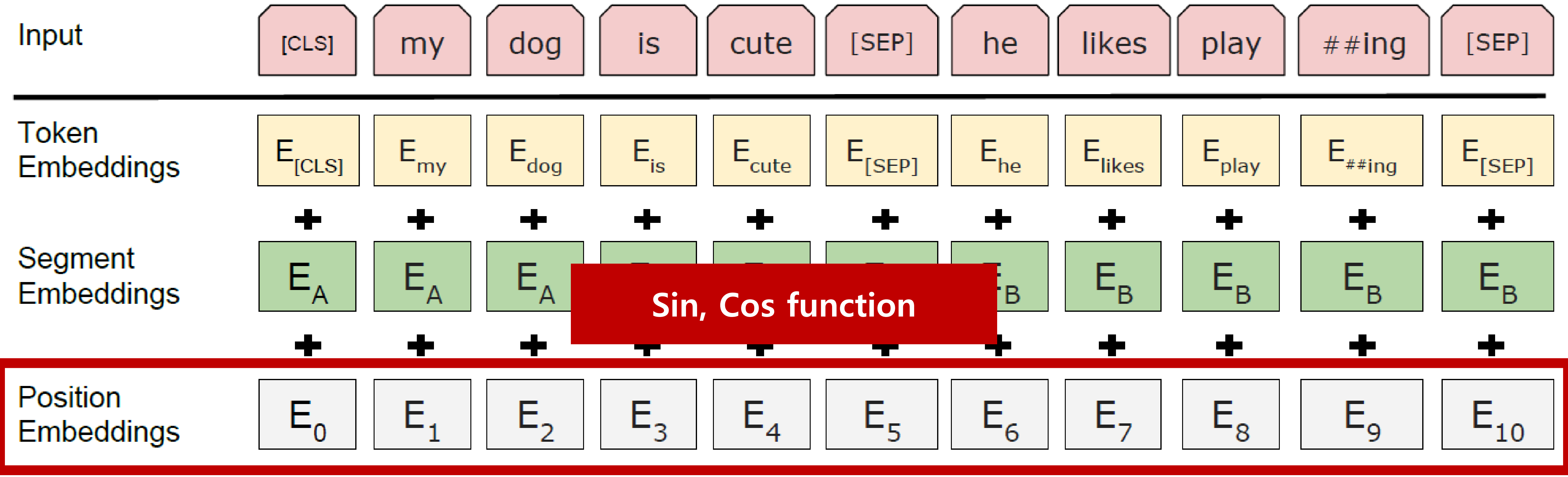
2) Input Representation



2) Input Representation



2) Input Representation



1) Outputs different value for each position

2) No limit on input

2) Pre-training Tasks

BERT = $\boxed{\text{Pre-training}}$ + Fine-tuning

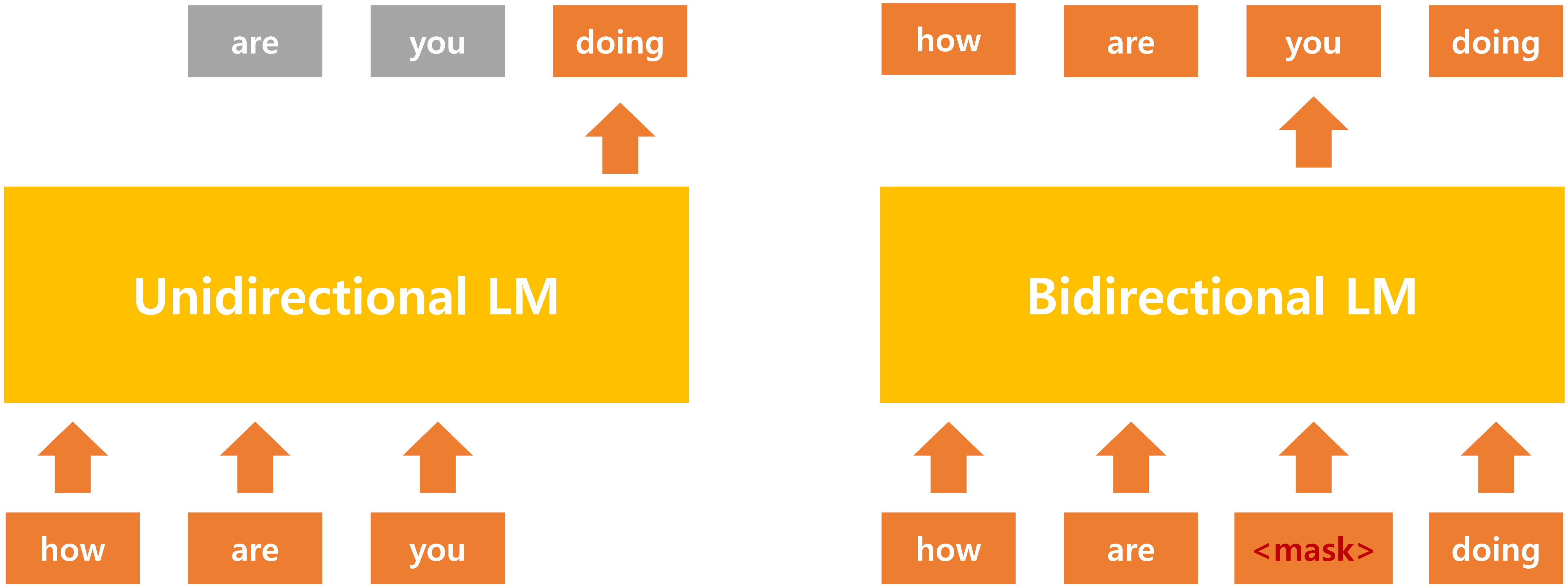
(1) Masked Language Model(MLM)

(2) Next sentence prediction

2) Pre-training Tasks

(1) Masked Language Model(MLM)

Unidirectional(Traditional) LM vs Bidirectional(Masked) LM(BERT)



2) Pre-training Tasks

(1) Masked Language Model(MLM) (Cont’d)

- **Generator** chooses **15%** of the token positions at **random** for prediction:

→ 80% : **[MASK]** token

→ 10% : **Random** token

→ 10% : **Unchanged** token

2) Pre-training Tasks

(2) Next Sentence prediction

- for task understanding **relationship between two sentences**(like **QA**)

- **Binarized next sentence prediction task**

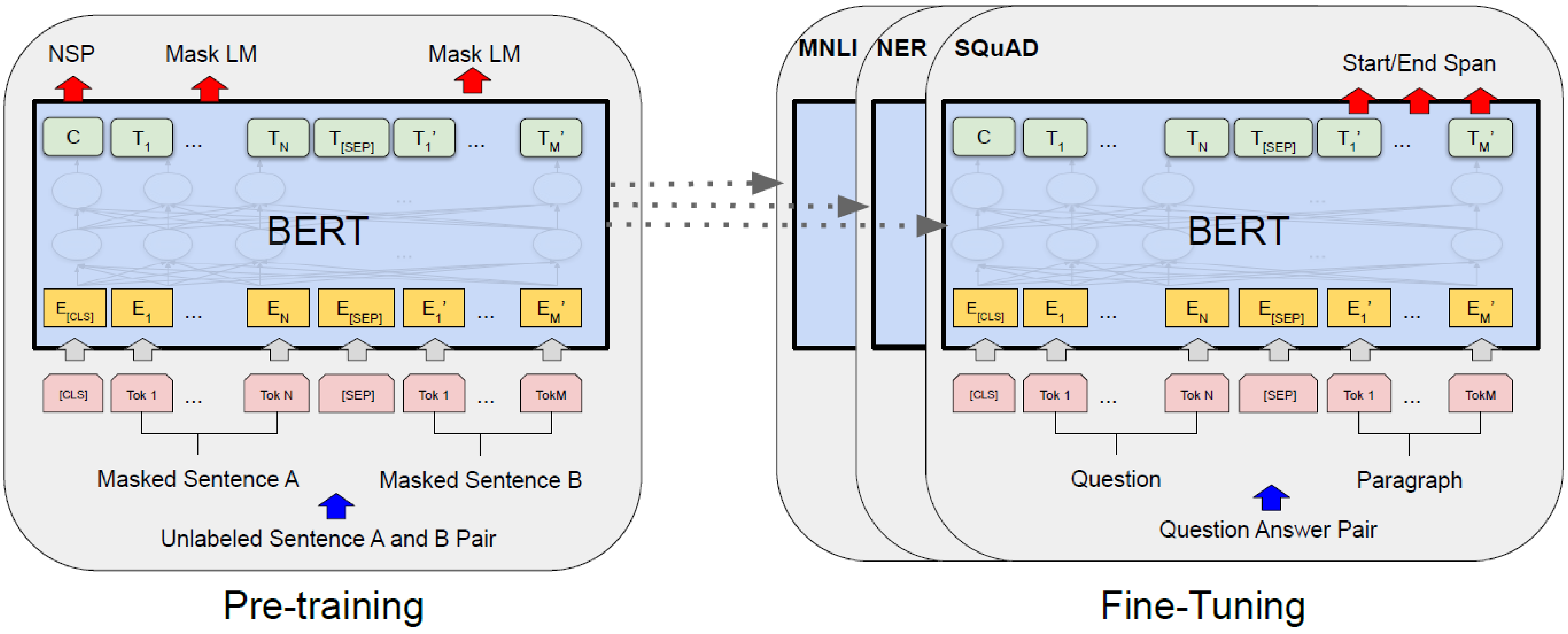
→ **50% : B** is the **actual next sentence**(labeled as **IsNext)**

→ **Remaining 50% : B** is a **random sentence**(labeled as **NotNext)**

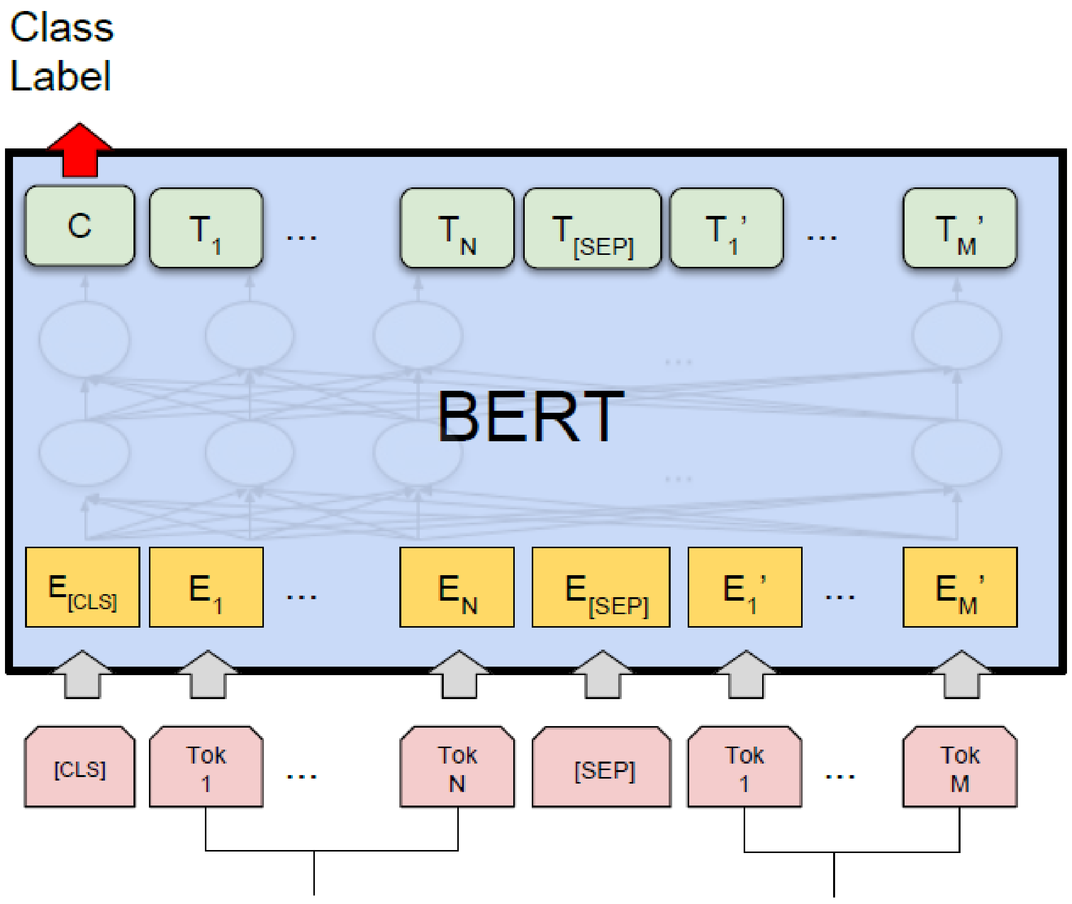
3) Fine-tuning Tasks

BERT = Pre-training + $\boxed{\text{Fine-tuning}}$

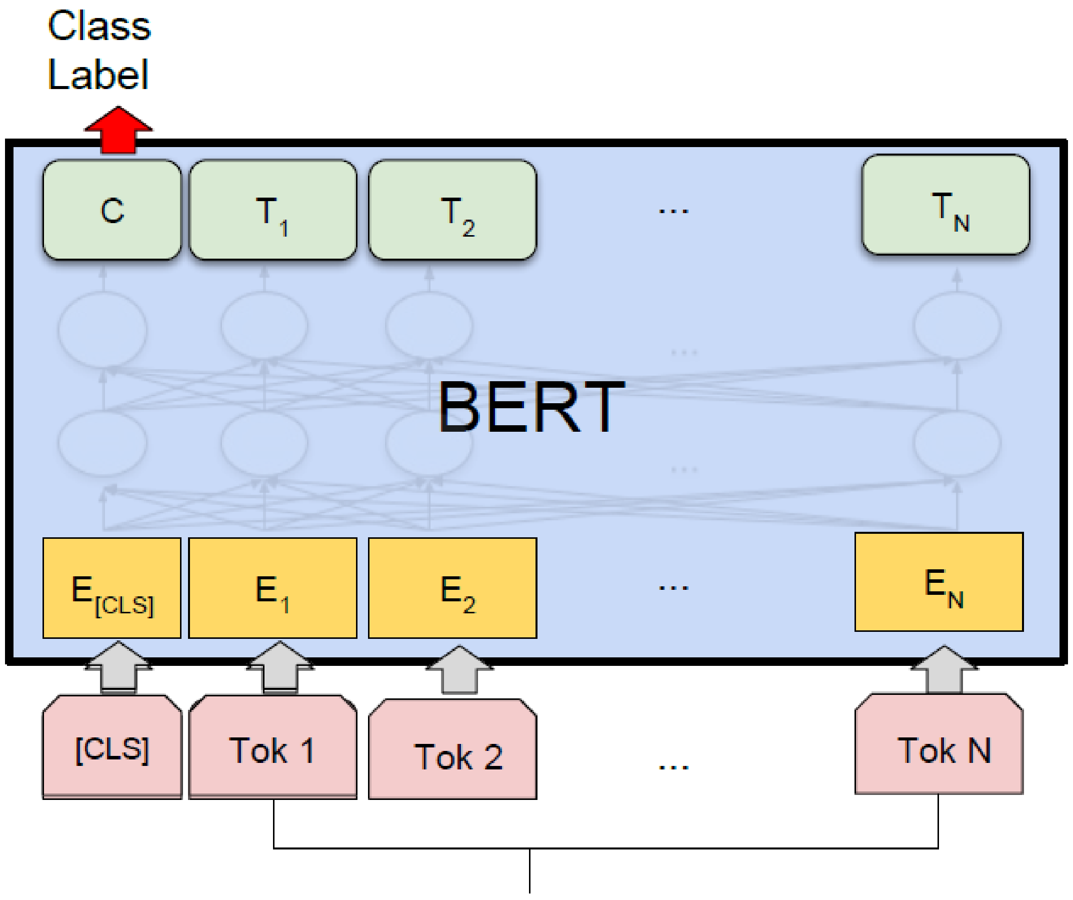
3) Fine-tuning Tasks



3) Fine-tuning Tasks

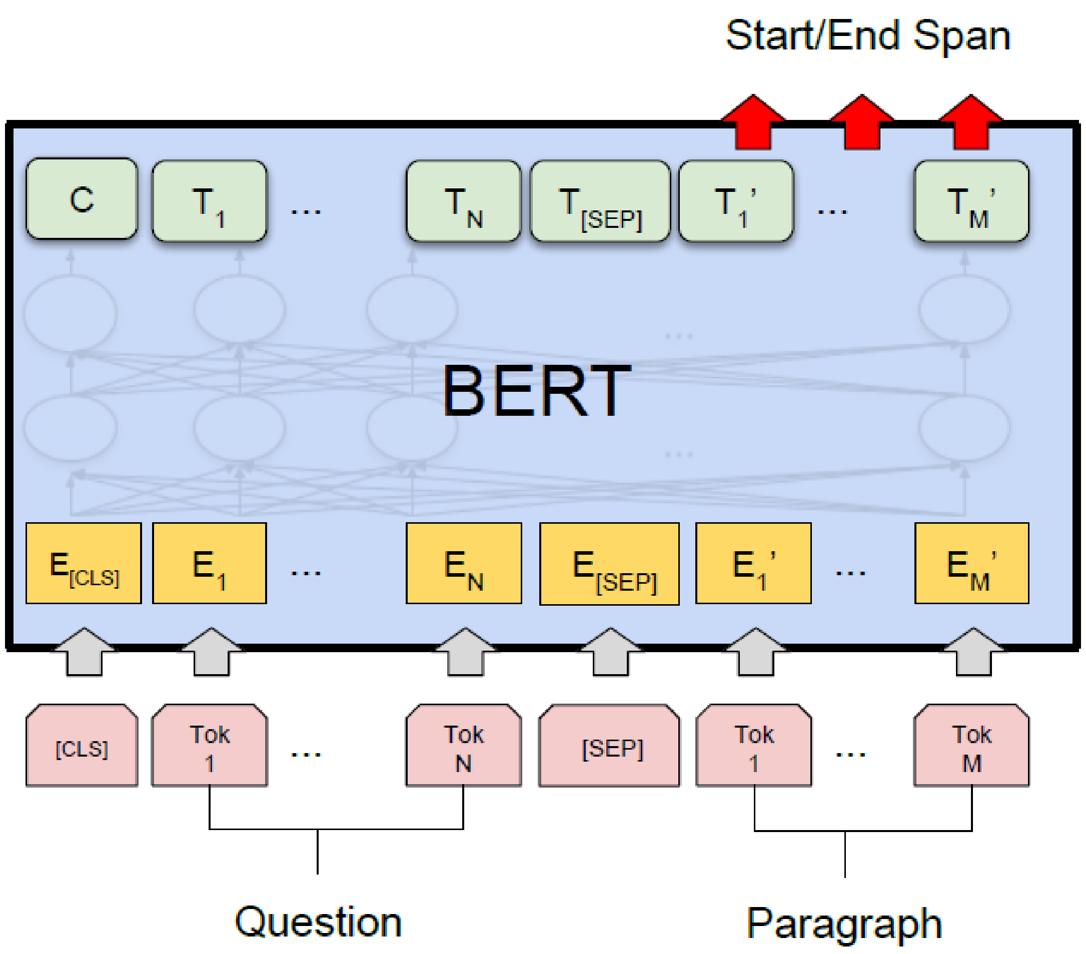


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

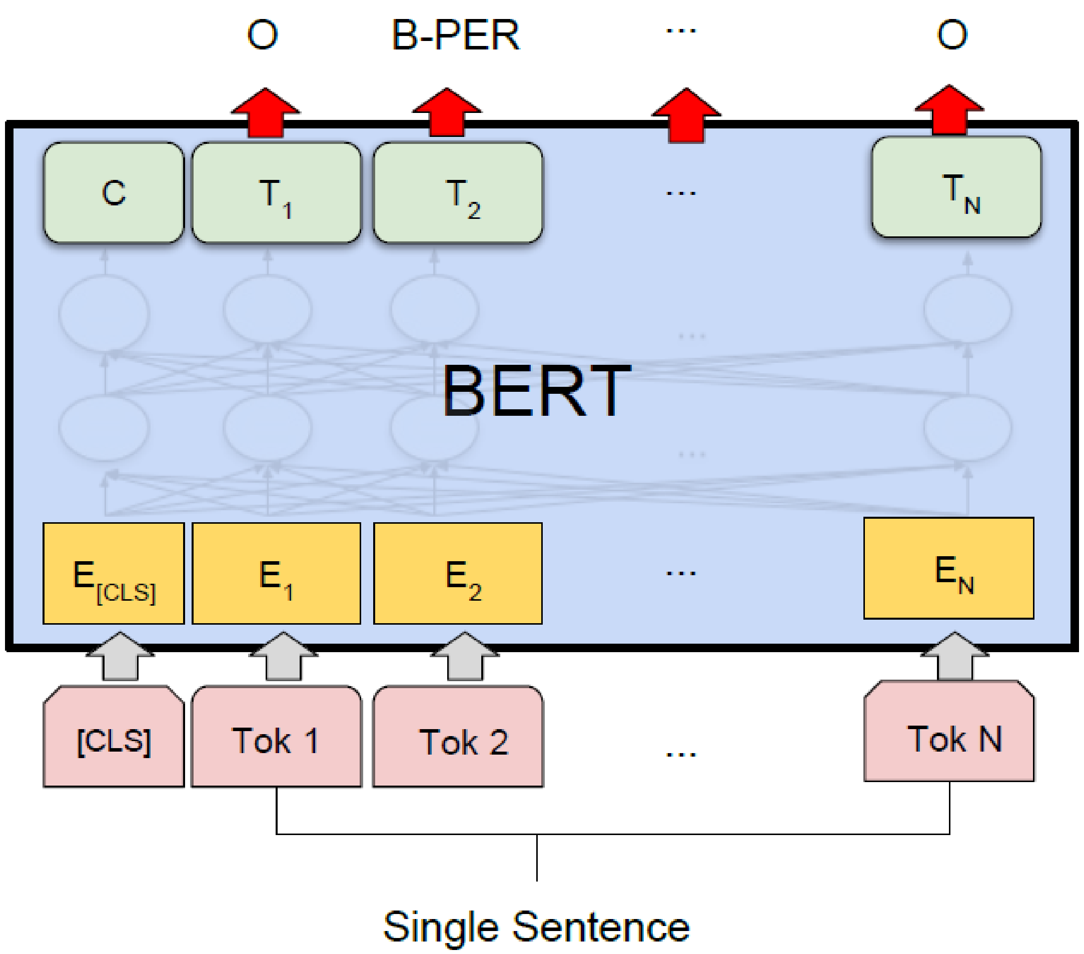


(b) Single Sentence Classification Tasks: SST-2, CoLA

3) Fine-tuning Tasks



(c) Question Answering Tasks:
SQuAD v1.1



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

3. Result

1) GLUE Results

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

1) GLUE Results

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**

- **GLUE → sequence classification task**

- **BERT** achieves **SOTA**

2) SQuAD v1.1 Results

- SQuAD → Question Answering task

- **BERT$_{\text{LARGE}}$** achieves **SOTA** (with wide margin)

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$_{\text{BASE}}$ (Single)	80.8	88.5	-	-
BERT$_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
BERT$_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
BERT$_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
BERT$_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

3) SWAG Results

- SWAG → grounded common-sense inference task

- **$\text{BERT}_{\text{LARGE}}$** achieves **SOTA** (with wide margin)

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

4) Ablation Studies

(1) Effect of Pre-training Tasks

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	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

4) Ablation Studies

(2) Effect of Model Size

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