BERT

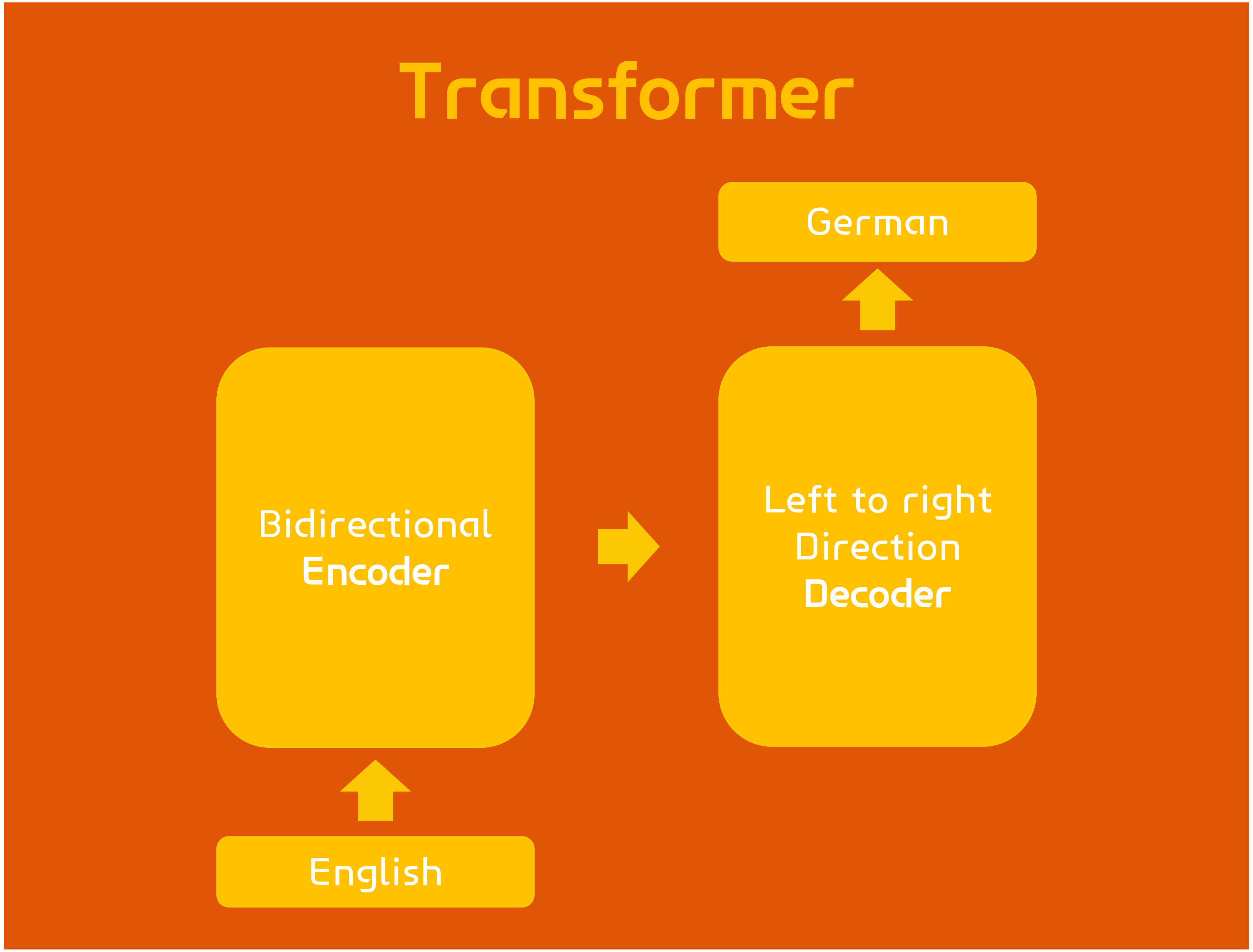
Bidirectional Encoder Representations From Transformers

2021720639 빅데이터학과 이찬우

BERT란?

Bidirectional Encoder Representations from Transformers

BERT란?


Transformer
German
Bidirectional
Encoder
Left to right
Direction
Decoder
English

BERT란?

BERT

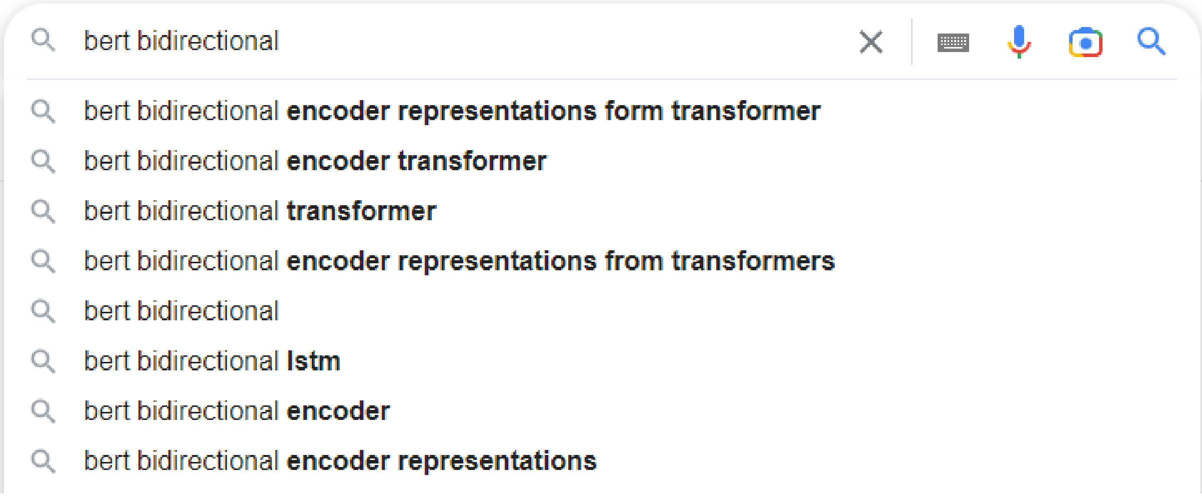
Bidirectional
Encoder

GPT

Left to right
Direction
Decoder

GPT-1

: 단어를 하나씩 읽어 가면서 다음 단어를 예측하는 모델



GPT-1

Left to right
Direction
Decoder

GPT-1

: 단어를 하나씩 읽어 가면서 다음 단어를 예측하는 모델

bert bidirectional encoder representation

Train Data	Label
bert	bidirectional
bert bidirectional	encoder
bert bidirectional encoder	representation

bert → GPT-1 → bidirectional

bert bidirectional → GPT-1 → encoder

bert bidirectional encoder → GPT-1 → representation

GPT-1

: 단어를 하나씩 읽어 가면서 다음 단어를 예측하는 모델

bert bidirectional encoder representation

Train Data	Label
bert	bidirectional
bert bidirectional	encoder
bert bidirectional encoder	representation

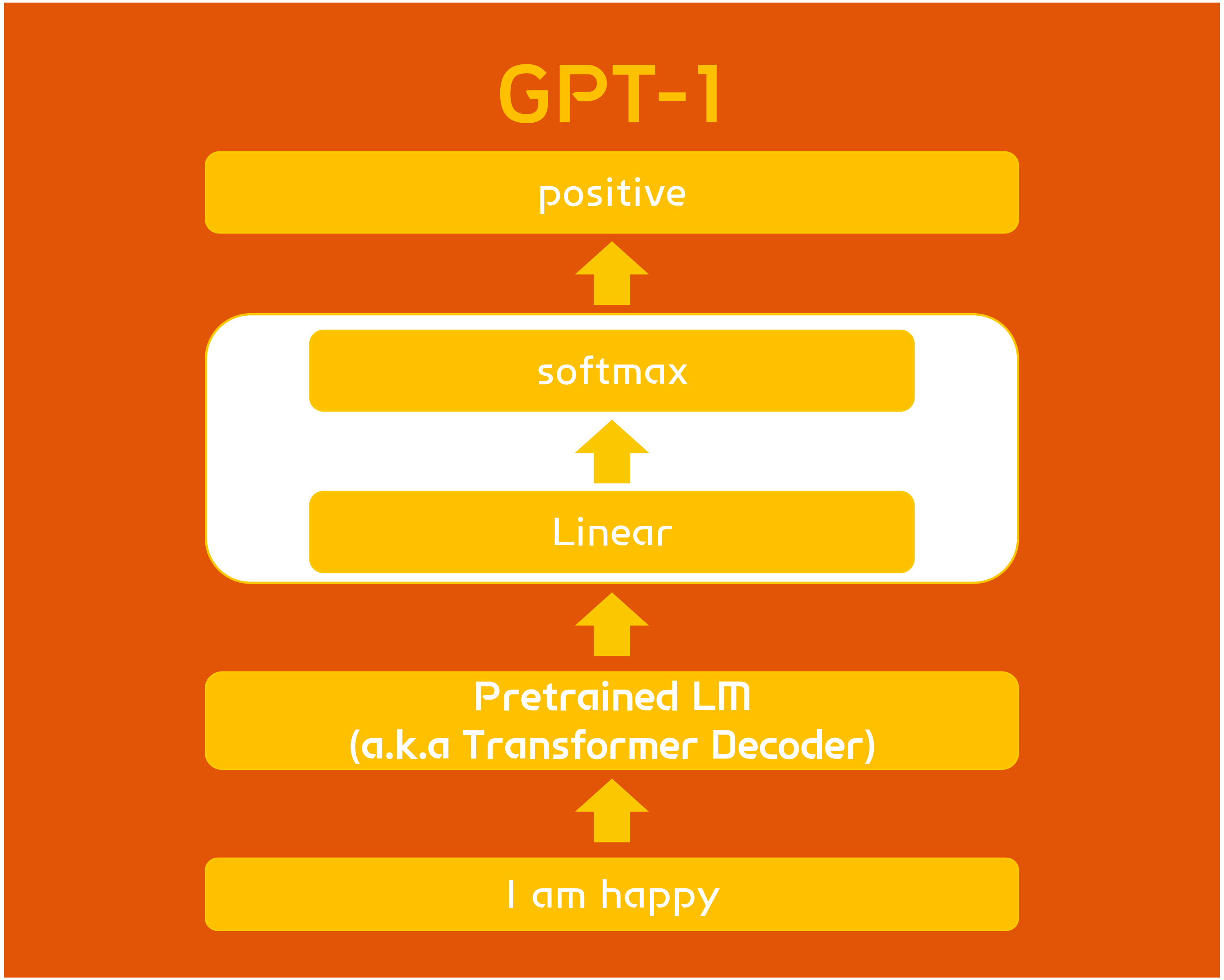
GPT-1

GPT-1

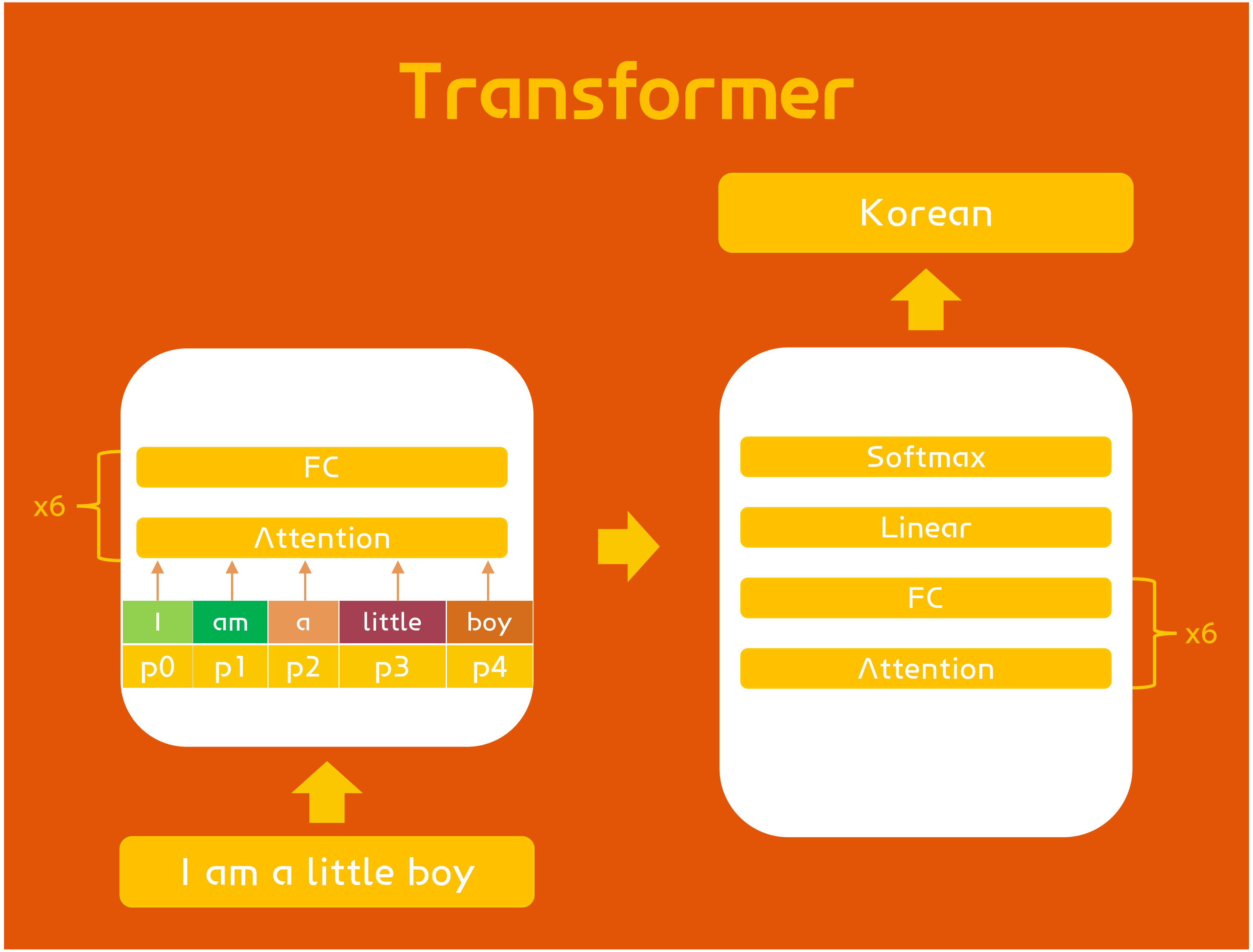
- GPT-1의 트랜스포머의 디코더를 사용한 자연어 처리 능력은 문장을 처리하는 데 부족함이 있을 수 있다.

- 더불어 질의 및 응답 영역은 문맥이해능력이 상당히 중요한데 단순히 왼쪽에서 오른쪽으로 읽어나가는 방식으로는 문맥이해에 약점이 있을 수 있다.

- 이에 단순히 왼쪽에서 오른쪽으로 읽어나가는 디코더보다 양방향으로 문맥을 이해할 수 있는 인코더를 활용한 언어 모델을 BERT라는 이름으로 발표



Transformer



Transformer

text

message

Transformer

Attention!

text

message

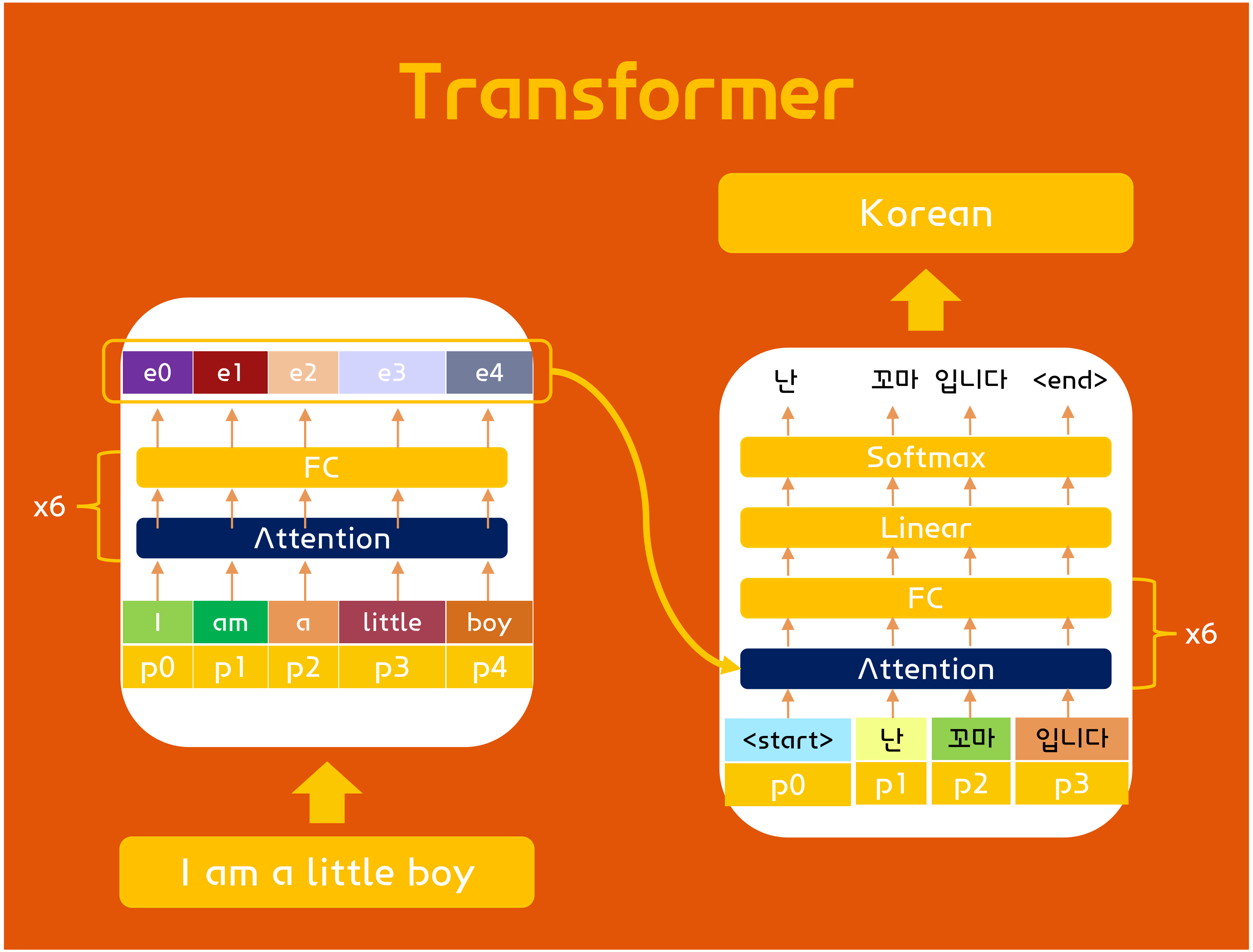
Transformer

- 인코더는 모든 토큰을 한방에 계산한다.

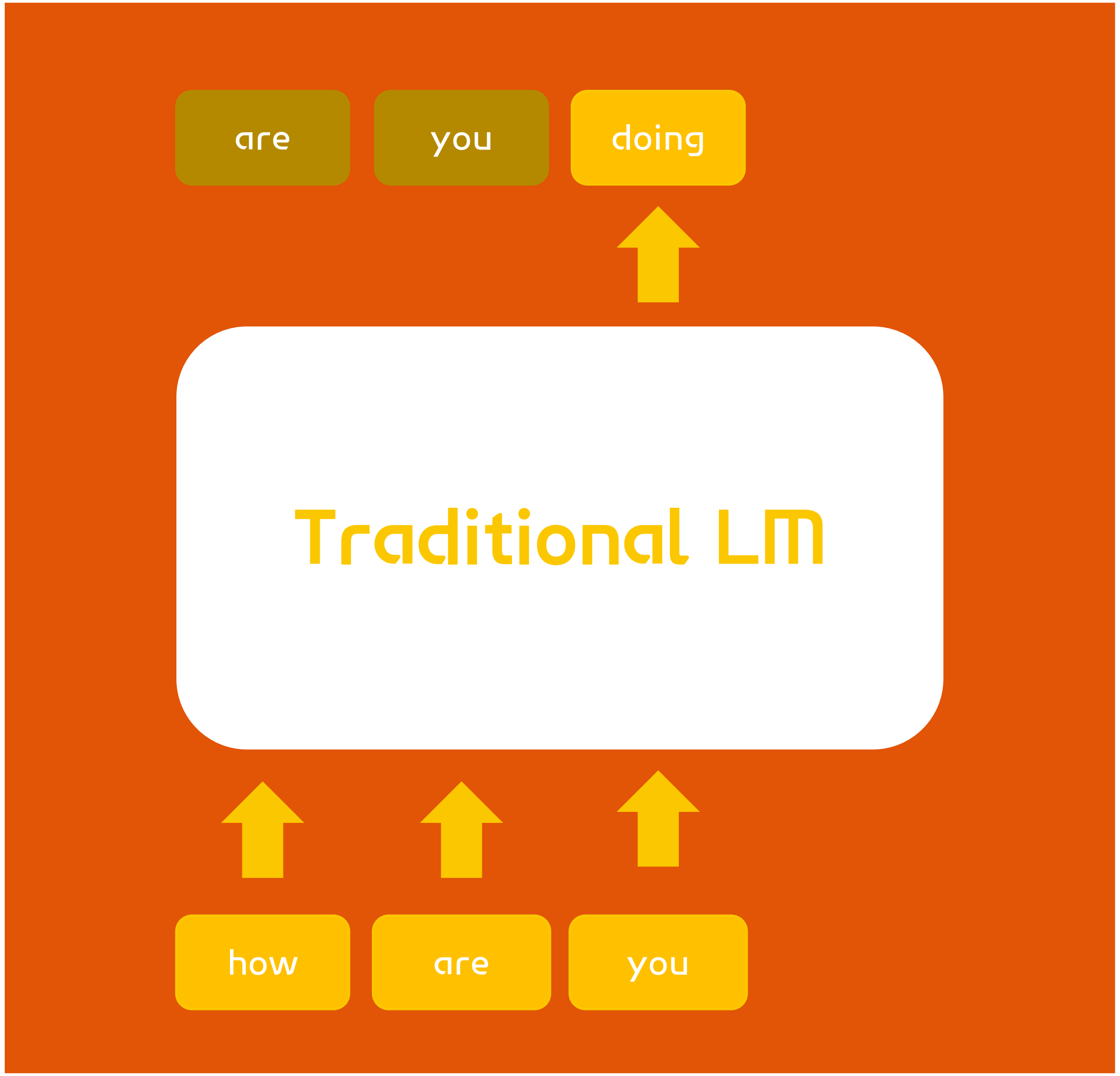
- 왼쪽에서 오른쪽으로 하나씩 읽어가는 과정이 없다.

1. 트랜스포머의 인코더는 양방향으로 문맥을 이해하고

2. 디코더는 왼쪽에서 오른쪽으로 문맥을 이해한다라는게 핵심

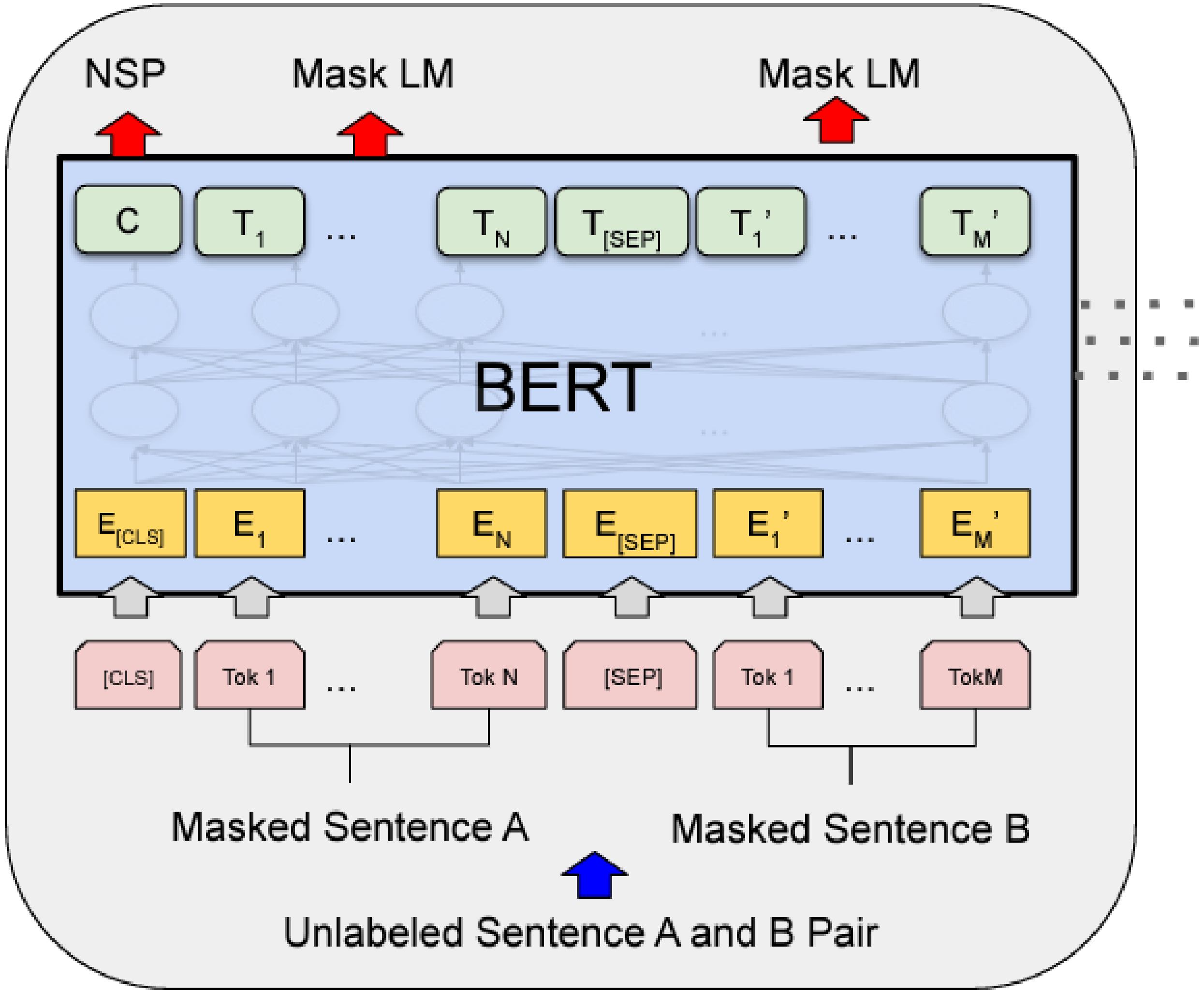


Traditional LM vs. bidirectional LM(BERT)



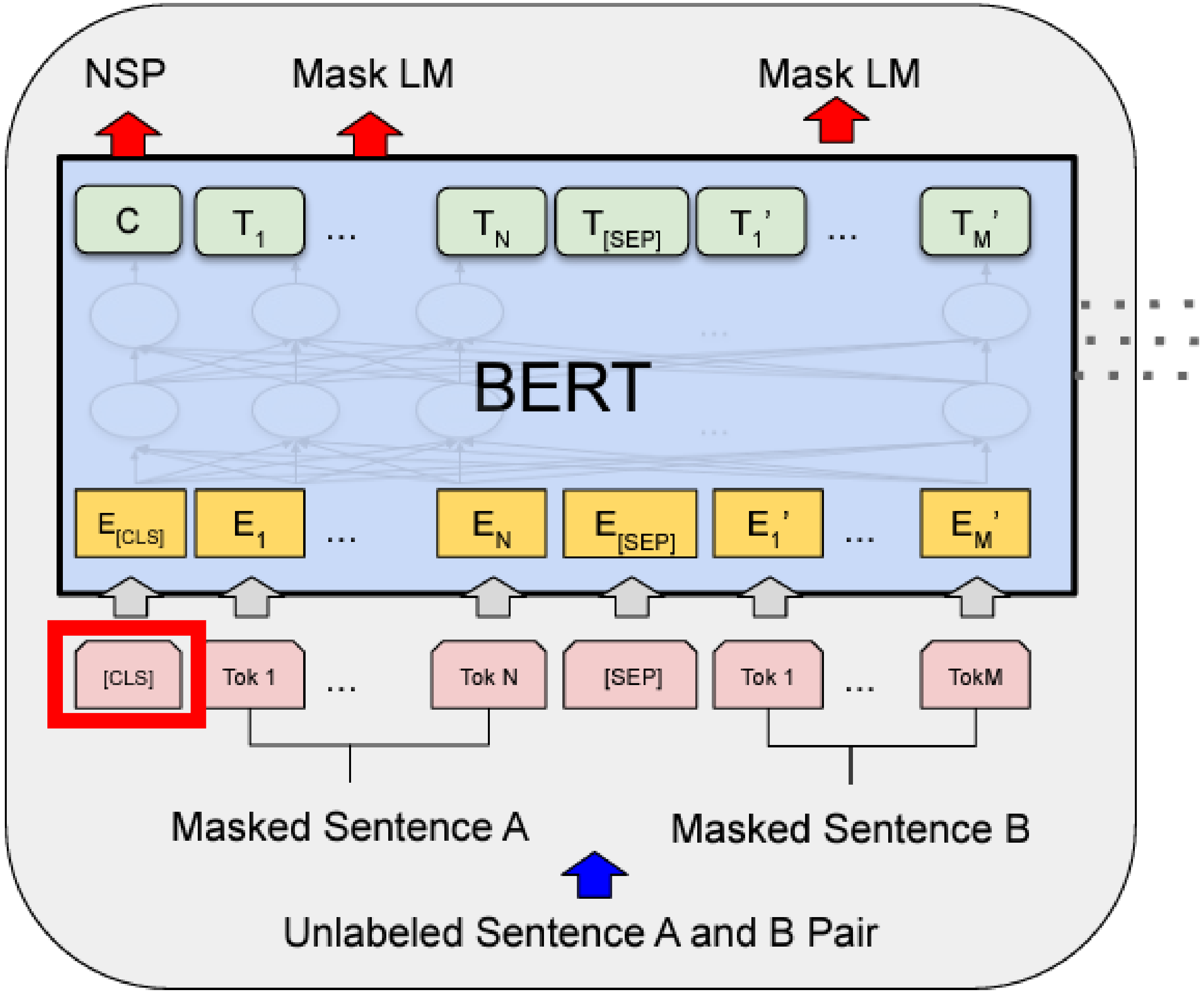
Traditional LM vs. bidirectional LM(BERT)

BERT Pre-training



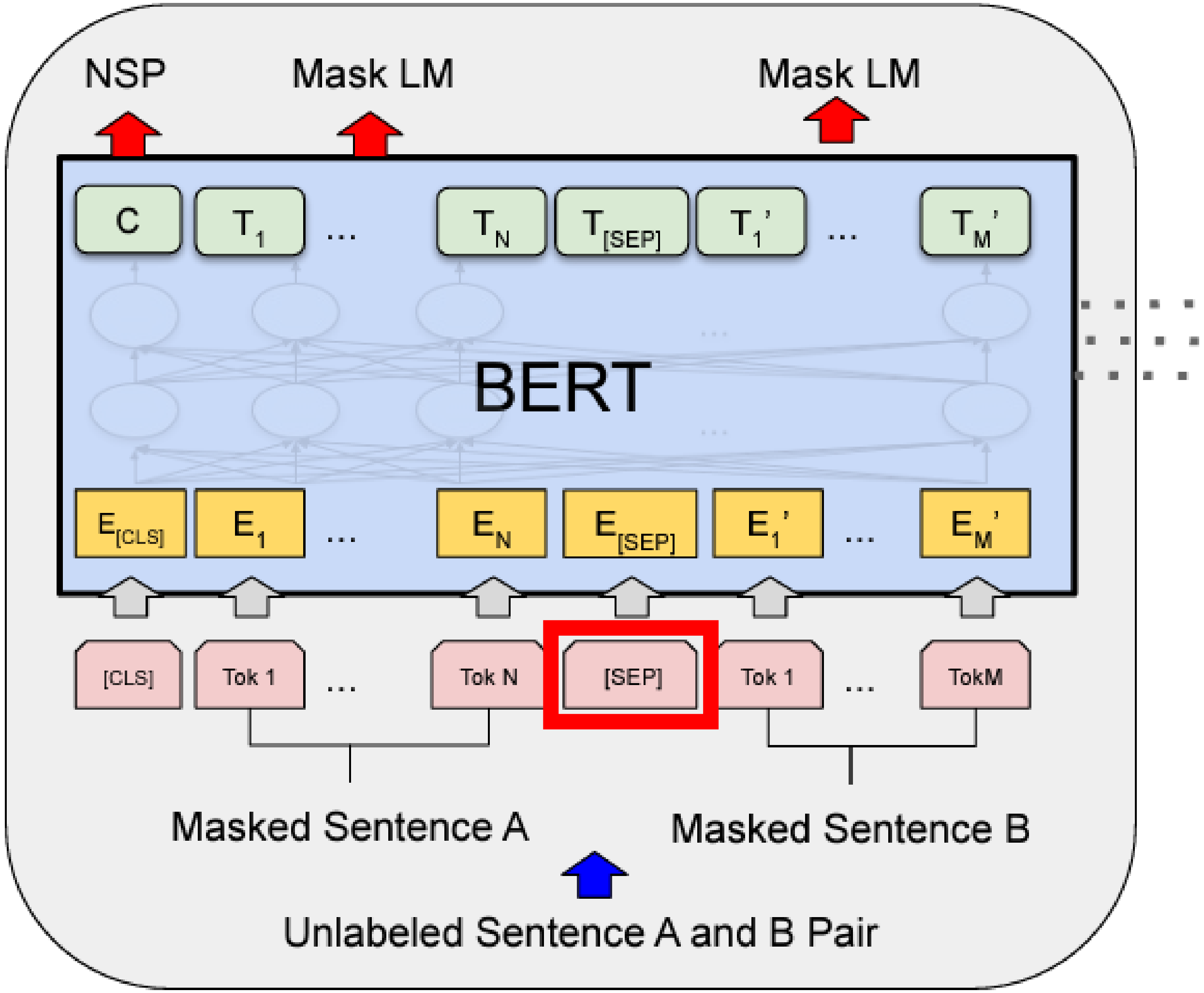
Pre-training

BERT Pre-training



Pre-training

BERT Pre-training



Pre-training

Word Piece Embedding

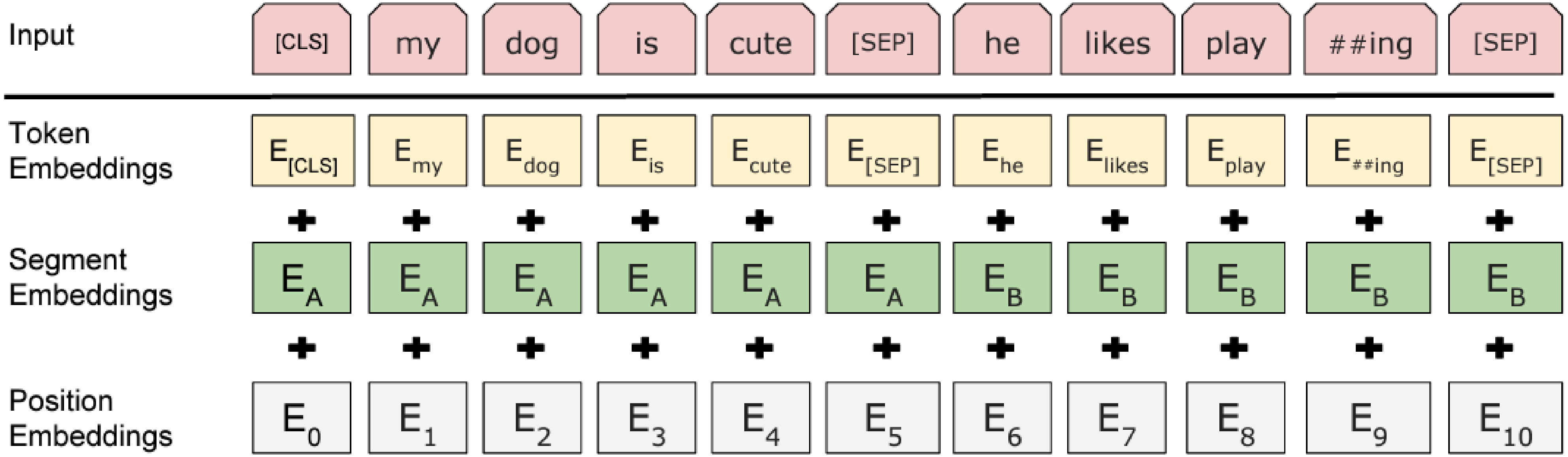


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Word Piece Embedding

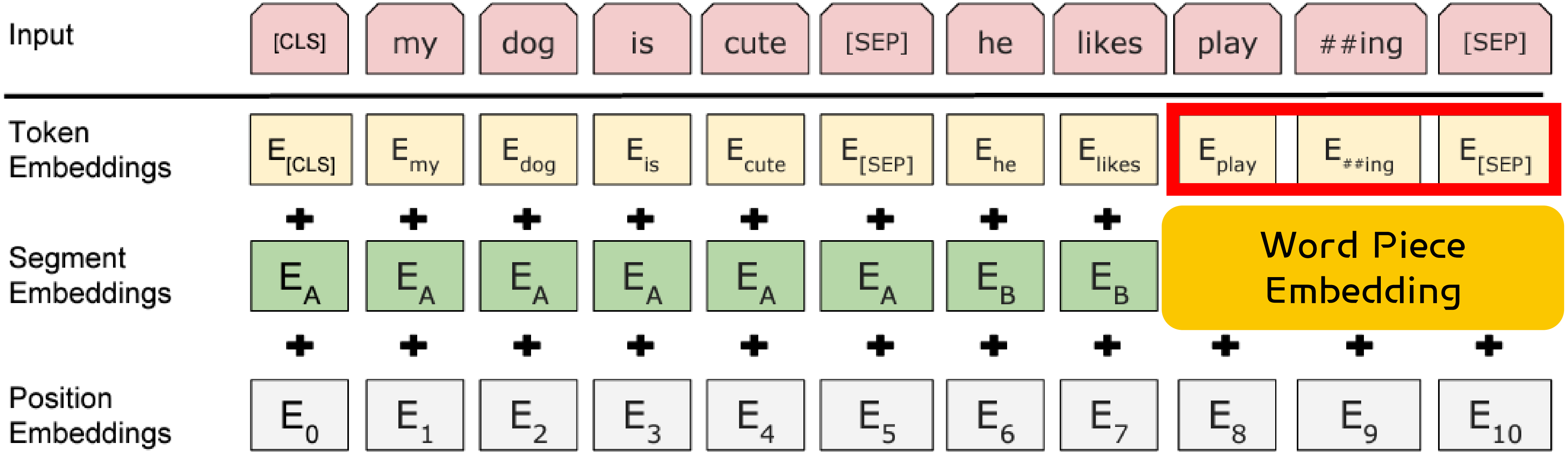
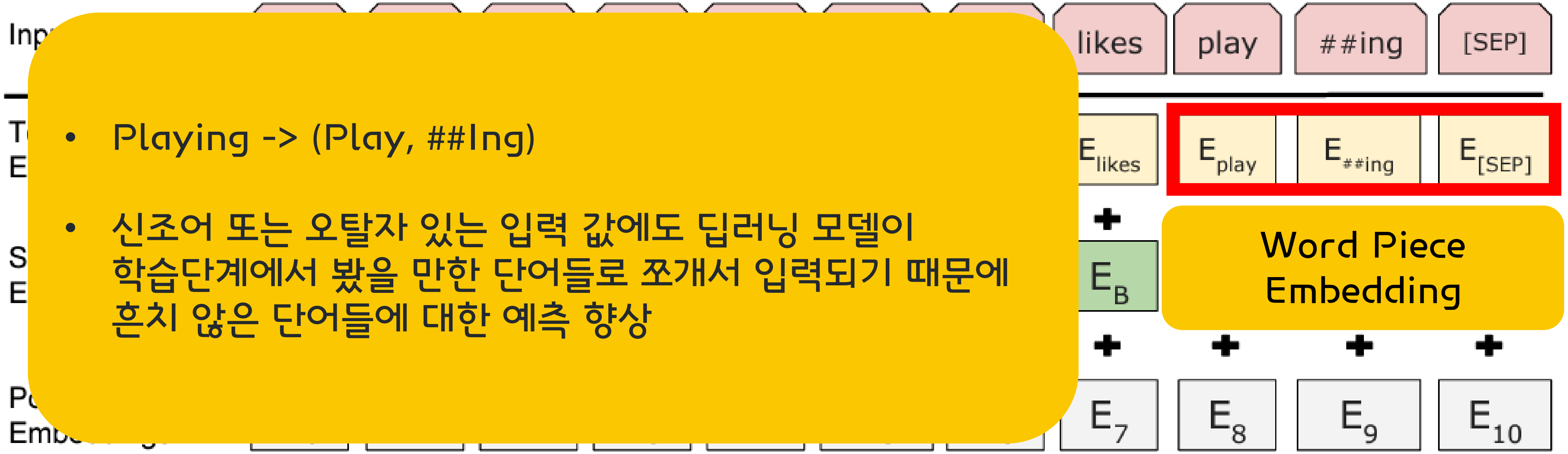


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Word Piece Embedding



- Playing -> (Play, ##Ing)
- 신조어 또는 오탈자 있는 입력 값에도 딥러닝 모델이 학습단계에서 봤을 만한 단어들로 쪼개서 입력되기 때문에 흔치 않은 단어들에 대한 예측 향상

Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Segment Embedding

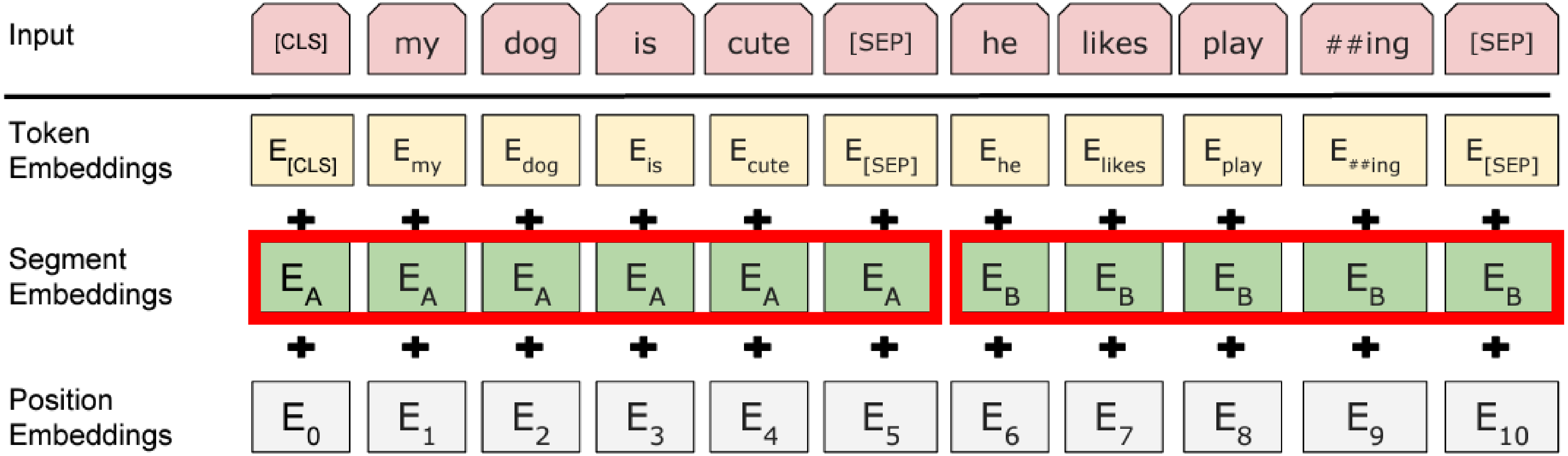


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Positional Embedding

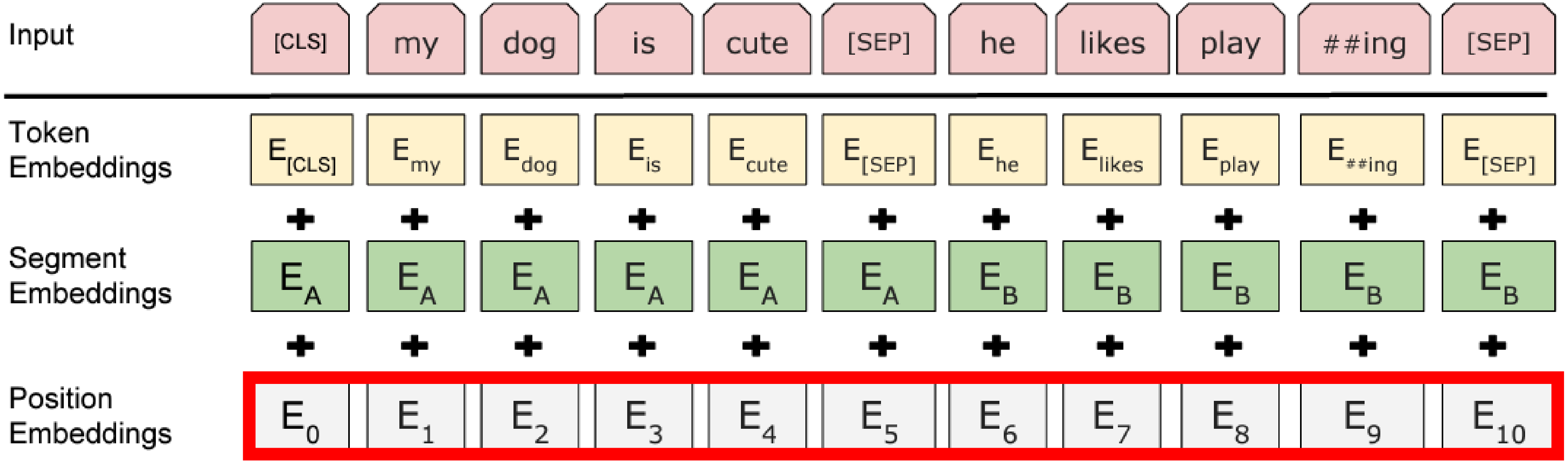


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Positional Embedding

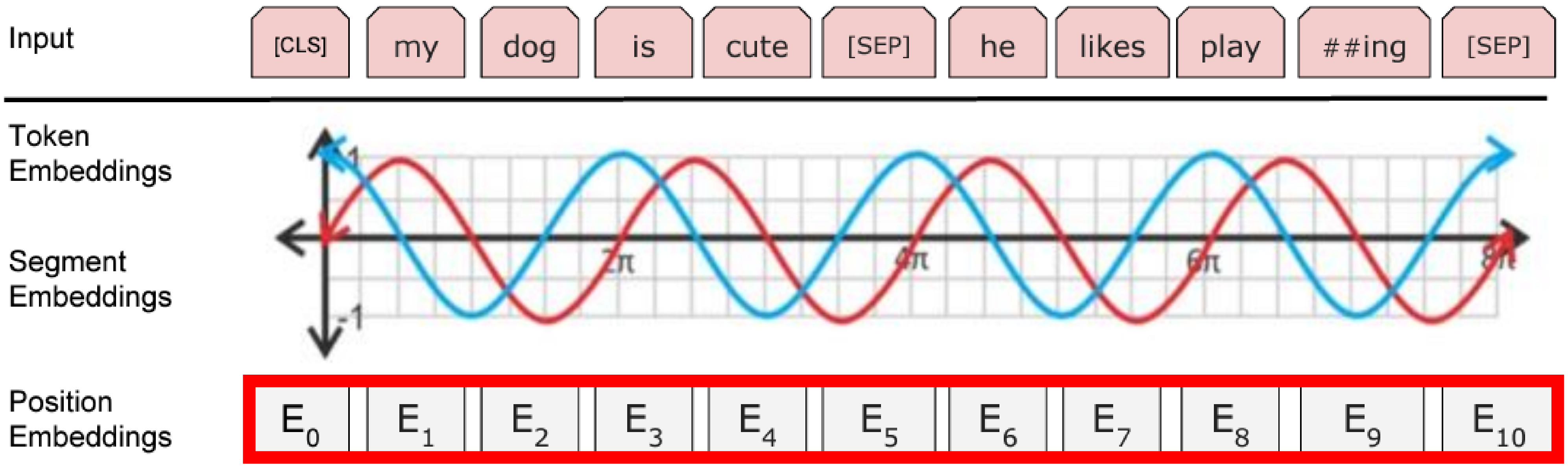


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Positional Embedding

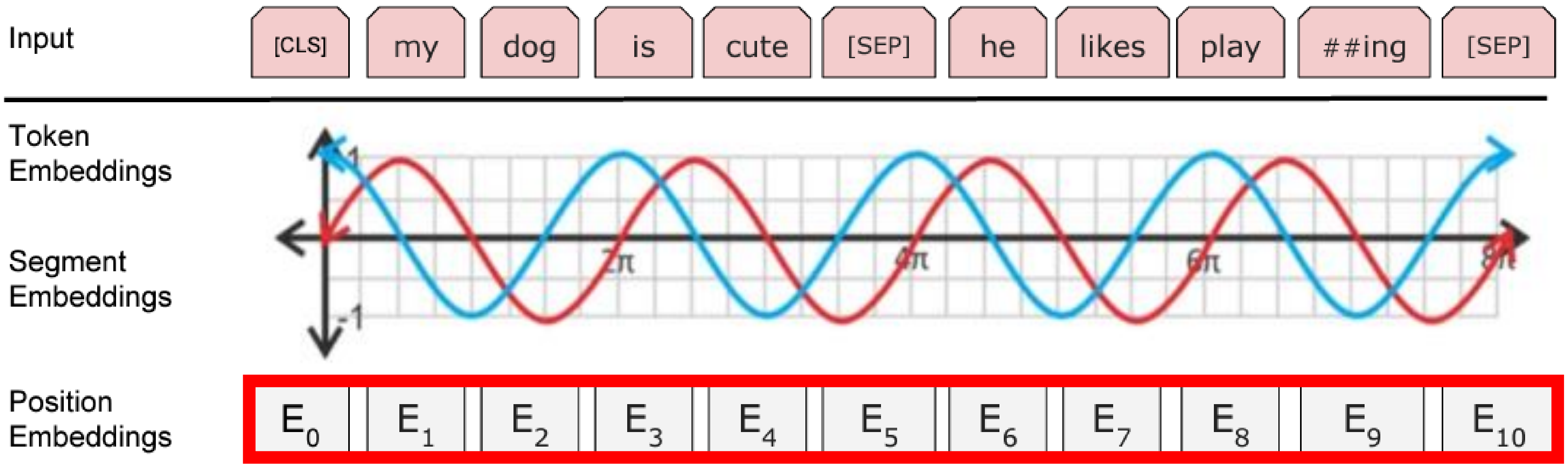


Figure 2: BERT input repre s, the segmenta-
tion embeddings and the po

사인과 코사인의 출력값은 입력값에 따라 달라진다.

Positional Embedding

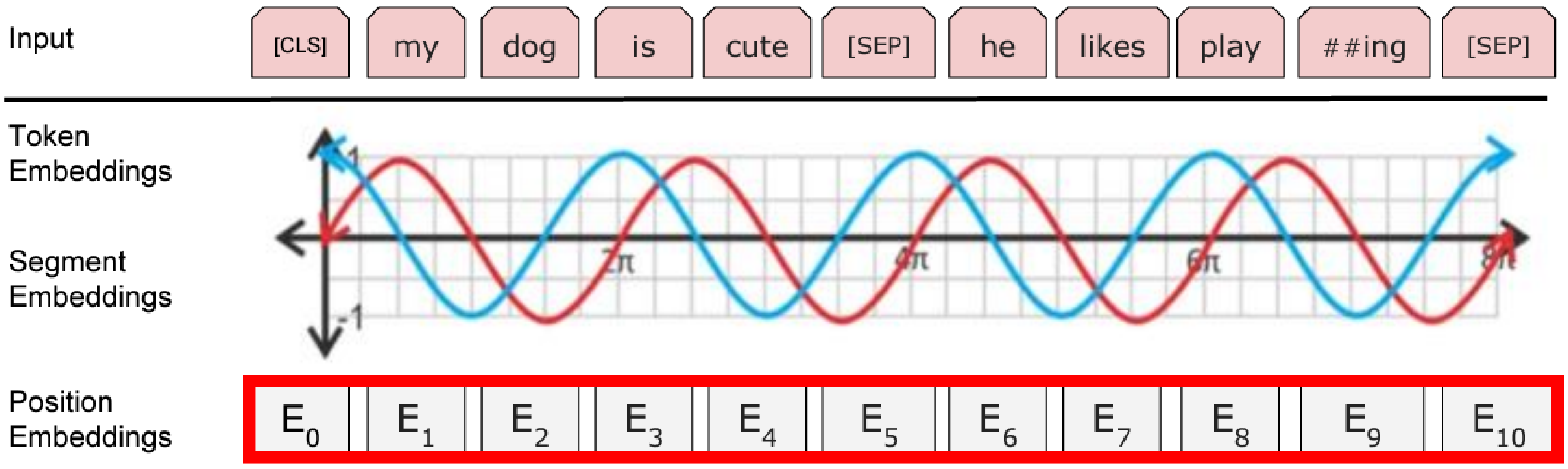


Figure 2: BERT input repre s, the segmenta- tion embeddings and the po

사인과 코사인의 출력값은 규칙적으로 증가 또는 감소한다.

Positional Embedding

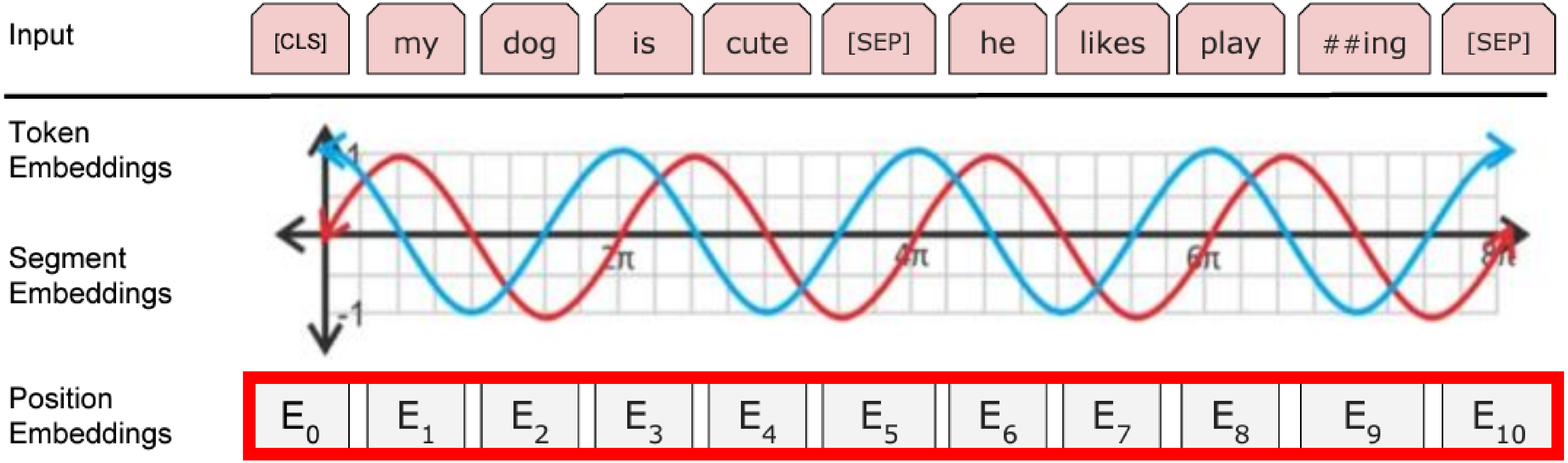


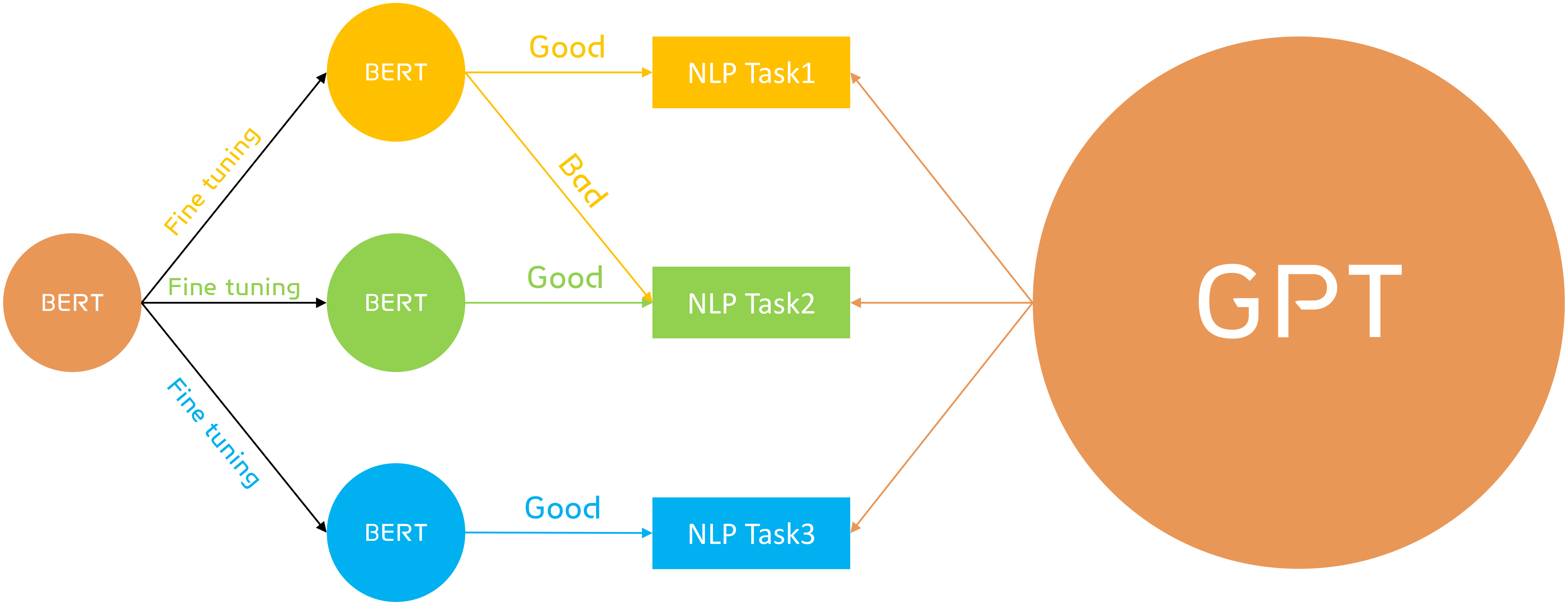
Figure 2: BERT input repre… s, the segmenta-
tion embeddings and the po…

사인과 코사인은 무한대의 길이의 입력값도
상대적인 위치를 출력할 수 있다

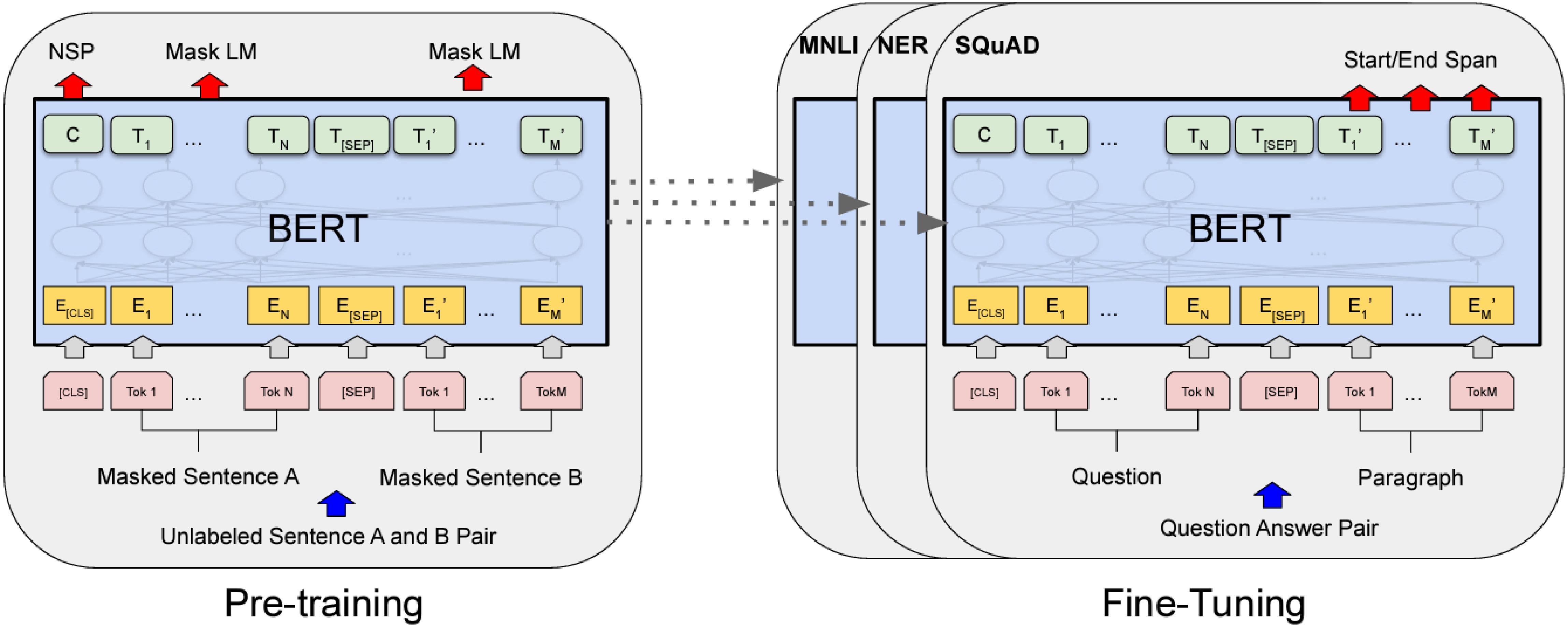
BERT vs GPT

BERT	GPT
Bidirectional LM	Left to Right LM
Loves Fine Tuning	Hates Fine Tuning

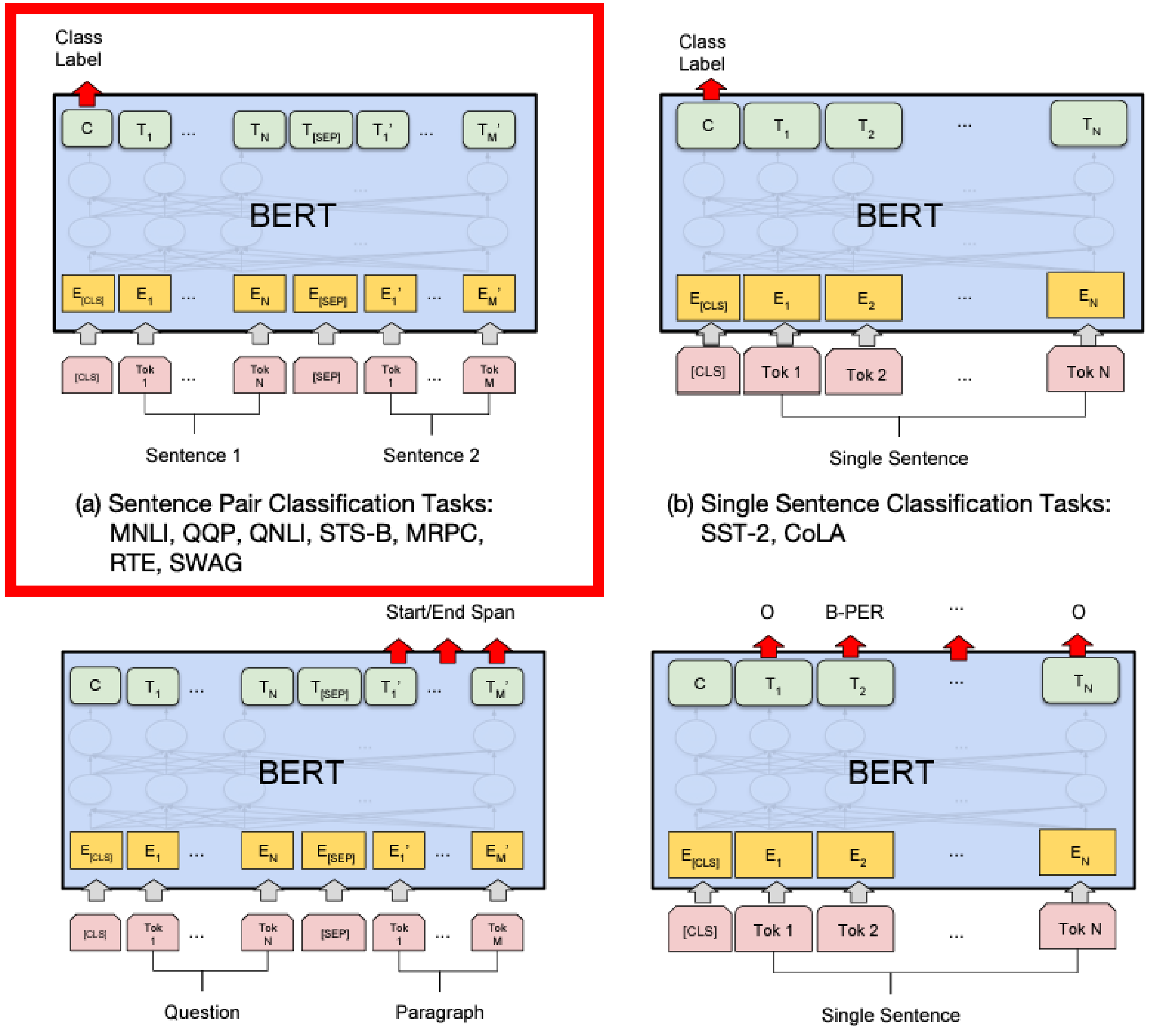
BERT vs GPT


BERT
Fine tuning
Fine tuning
Fine tuning
BERT
BERT
BERT
Good
Bad
Good
Good
NLP Task1
NLP Task2
NLP Task3
GPT

BERT Fine Tuning



BERT Fine Tuning



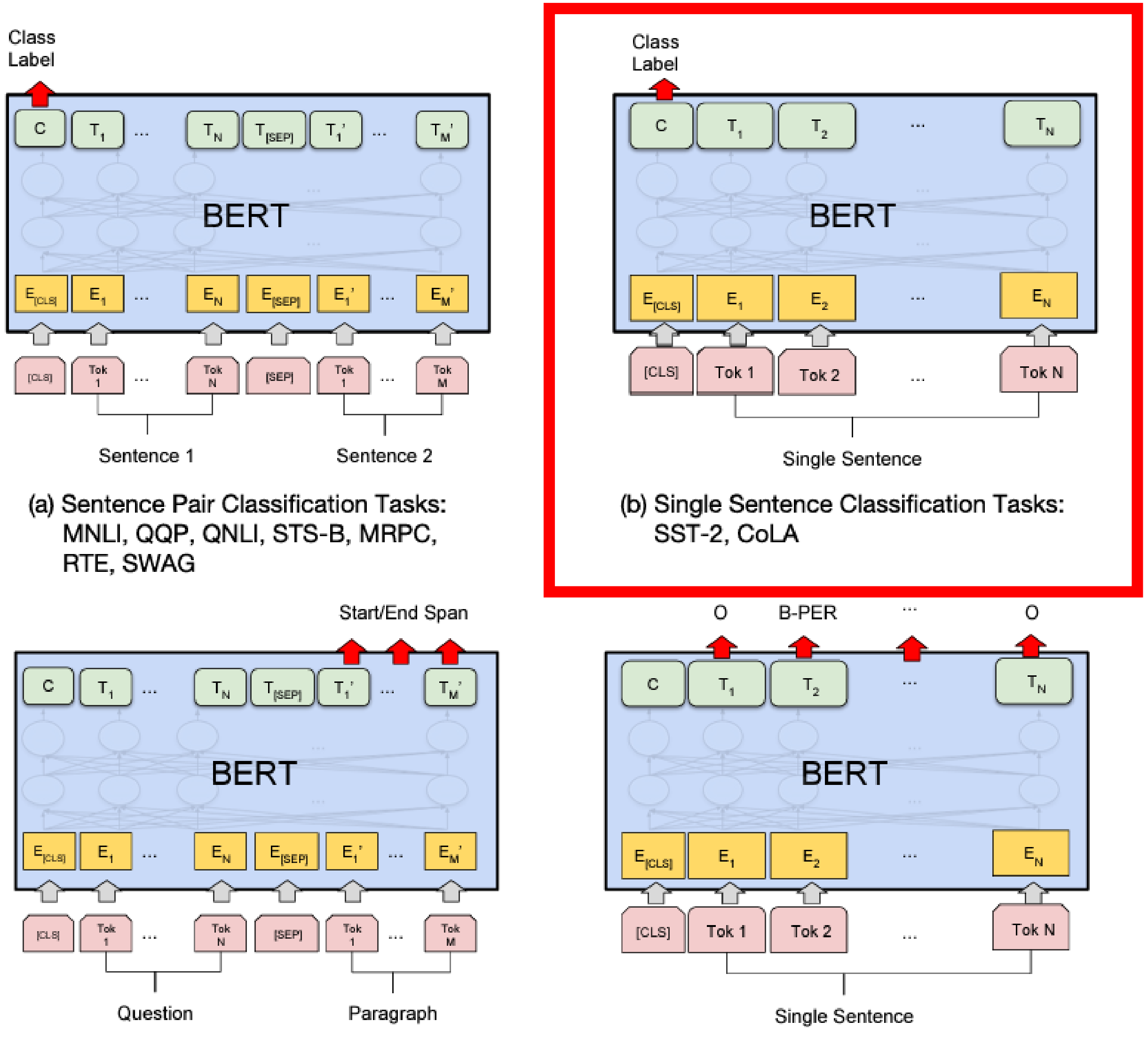
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

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

(c) Question Answering Tasks: SQuAD v1.1

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

BERT Fine Tuning



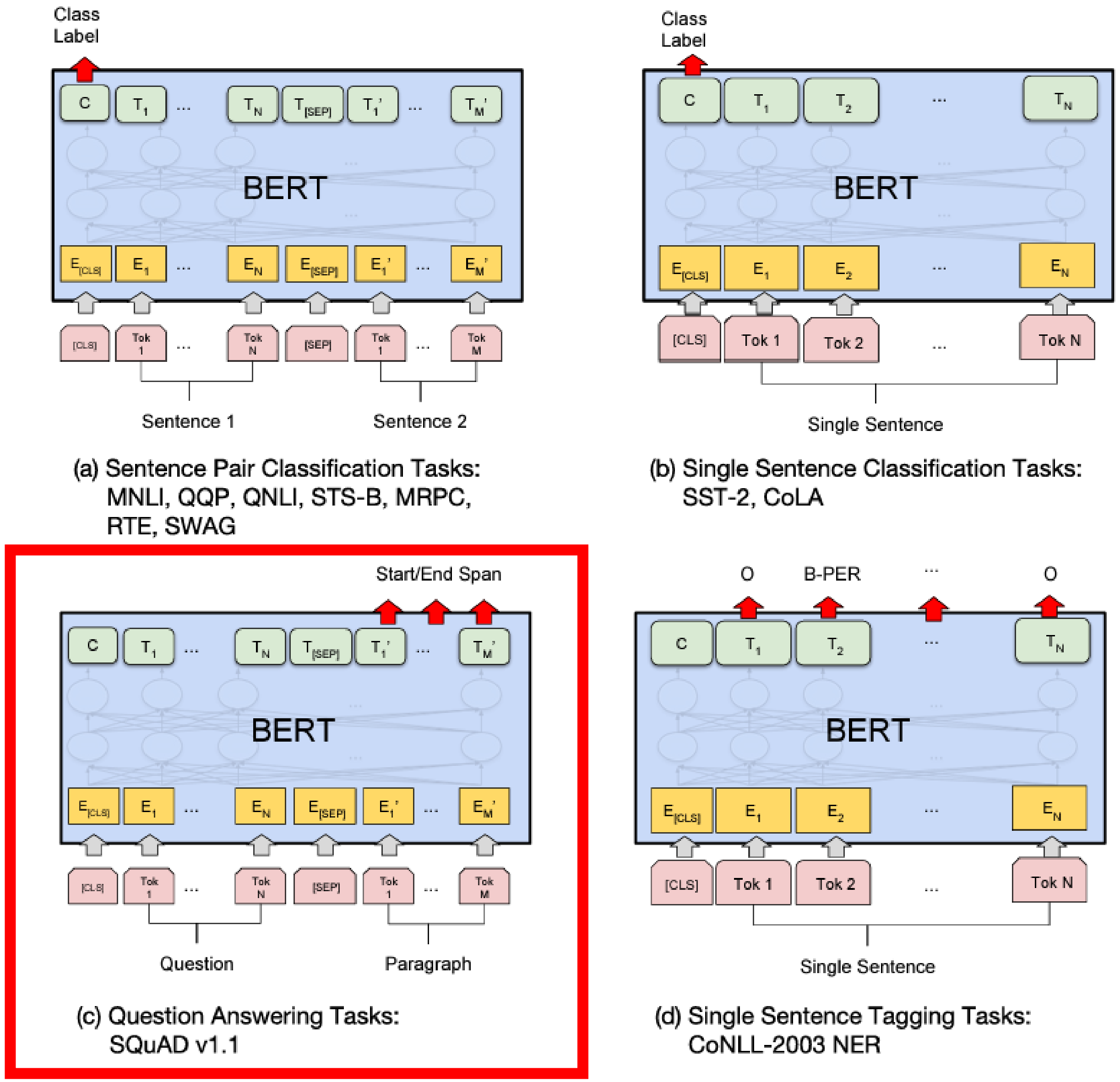
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

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

(c) Question Answering Tasks: SQuAD v1.1

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

BERT Fine Tuning



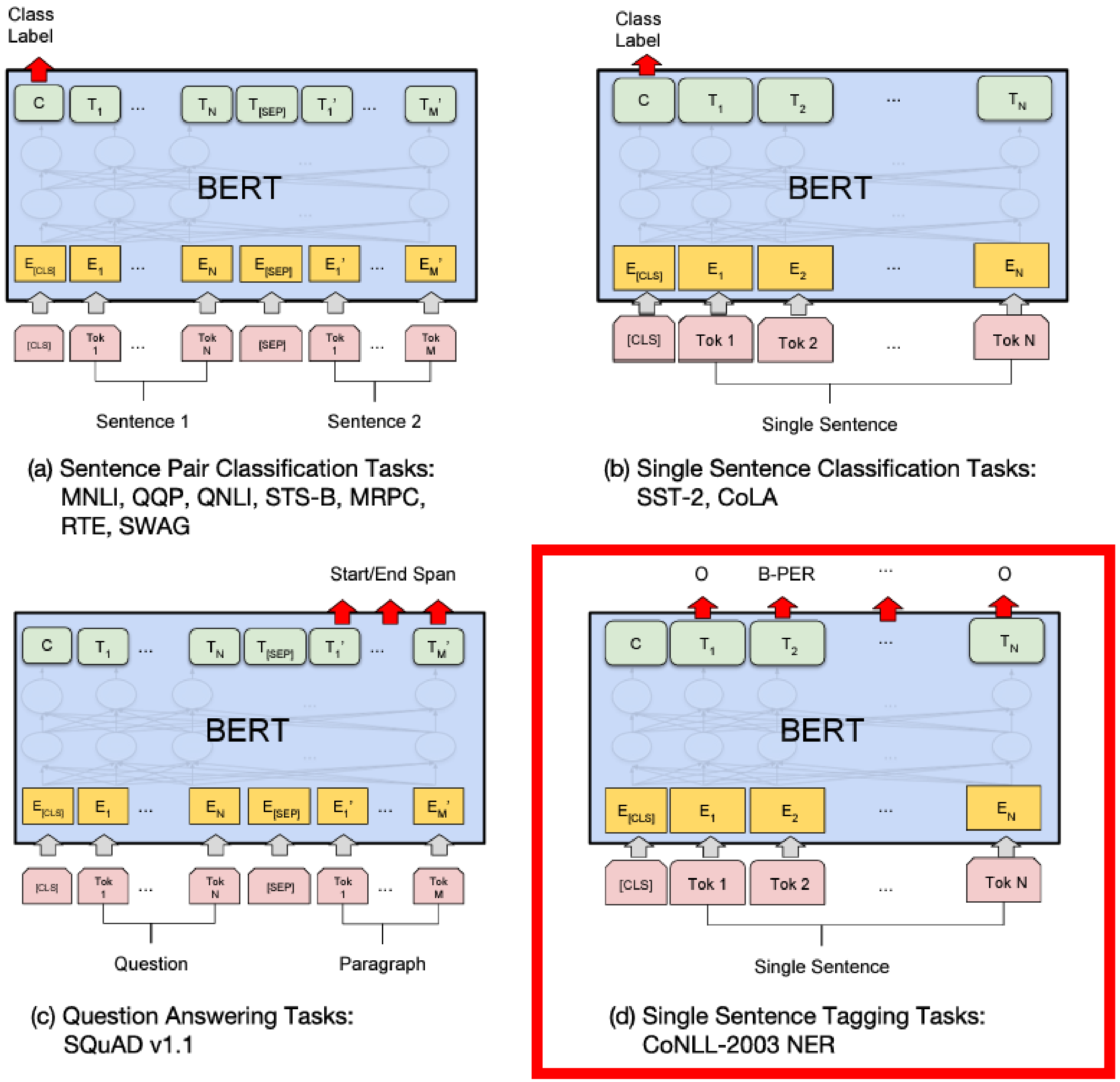
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

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

(c) Question Answering Tasks: SQuAD v1.1

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

BERT Fine Tuning



BERT Performance

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**

Table 1: GLUE Test results, scored by the evaluation server (`https://gluebenchmark.com/leaderboard`). The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.[8] BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

Reference

- BERT: Pre-training of Deep Bidirectional Transformer for Language Understanding
- https://arxiv.org/pdf/1810.04805.pdf