Pretrained Language Models 1

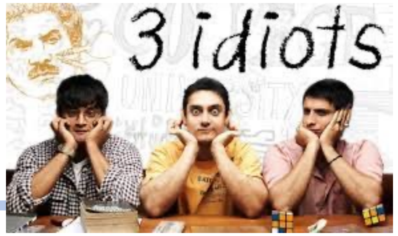
BERT

Woohwan Jung

한양대학교
HANYANG UNIVERSITY

DSLAB
데이터사이언스 연구실

Attention Is All You Need

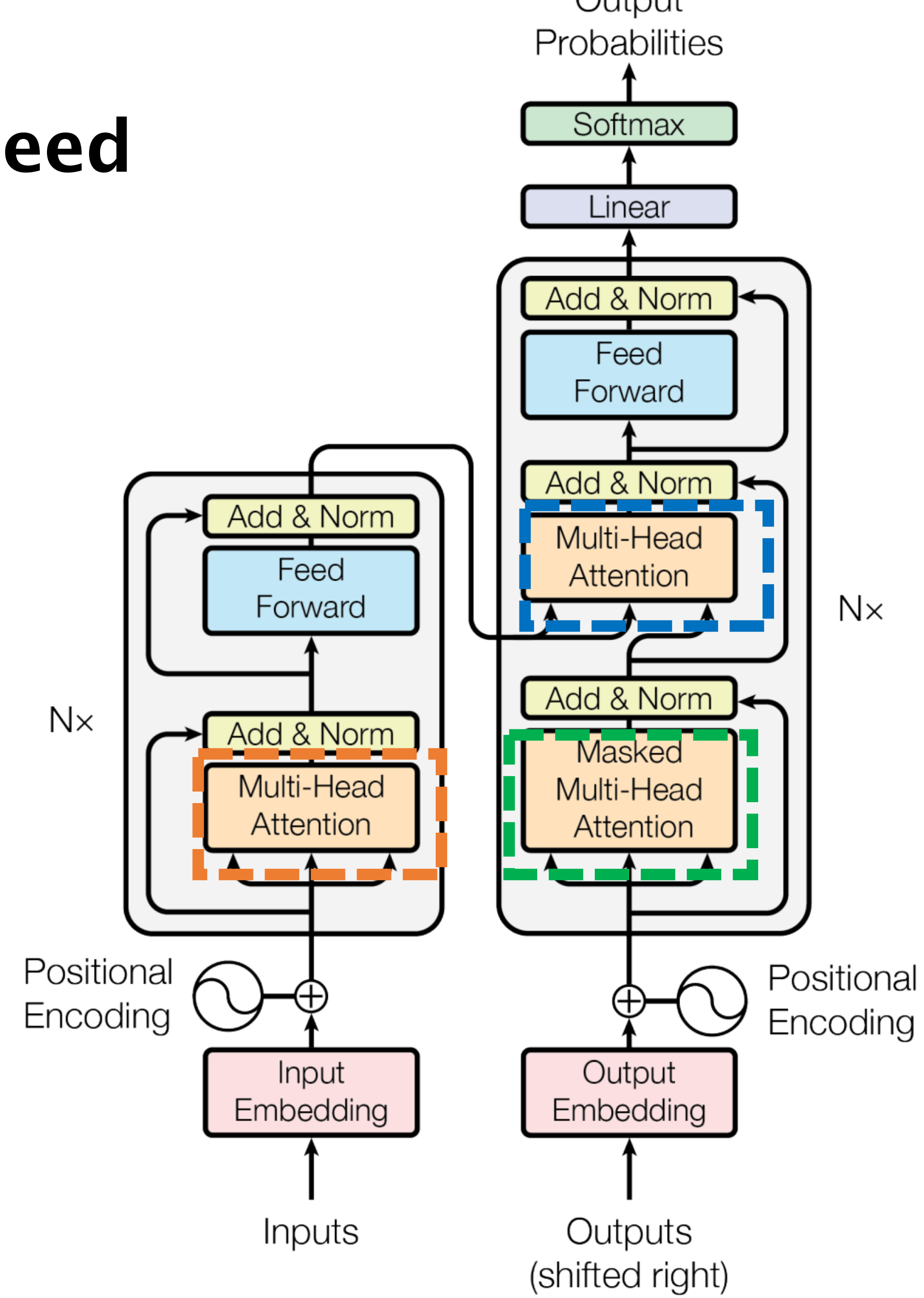


- Three attentions

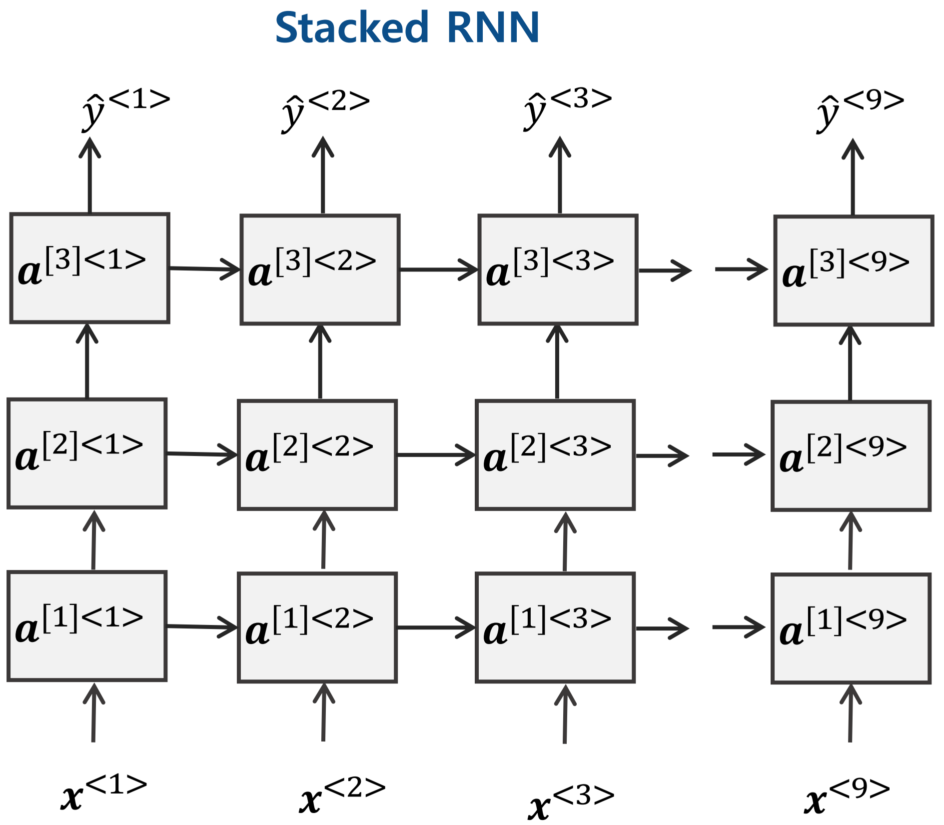
  Self-attention (encoder)

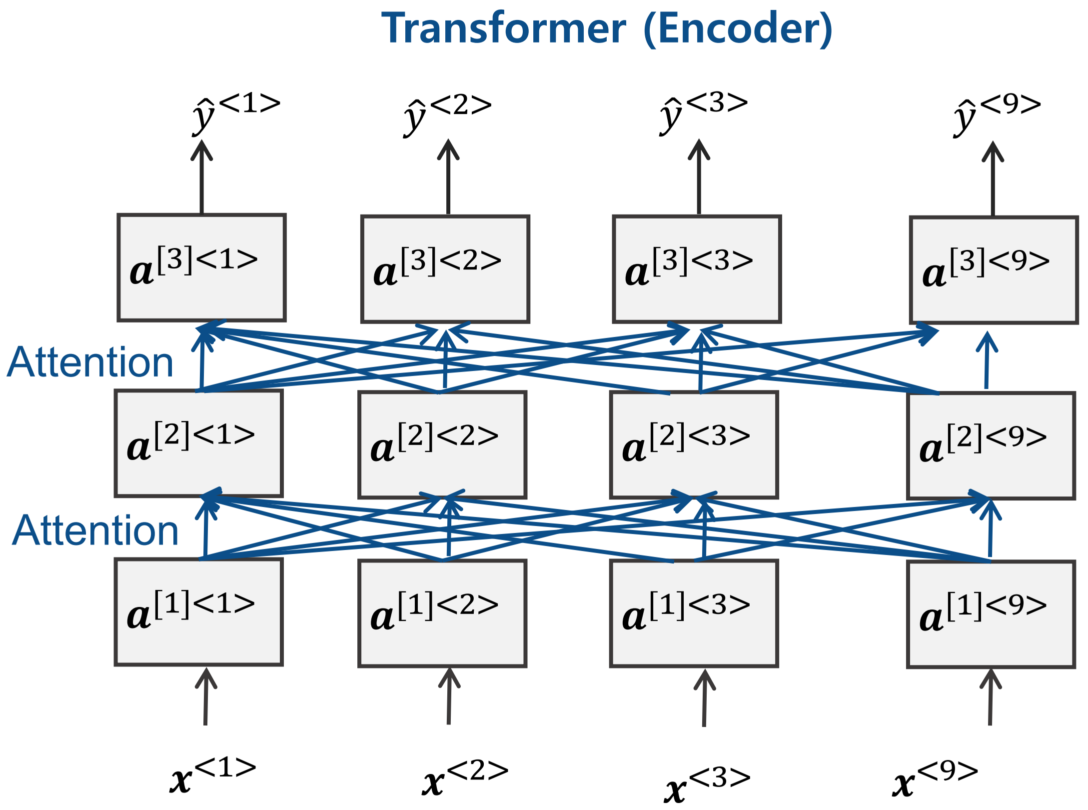
  Self-attention (decoder)

  Encoder-decoder attention



RNN vs Transformer





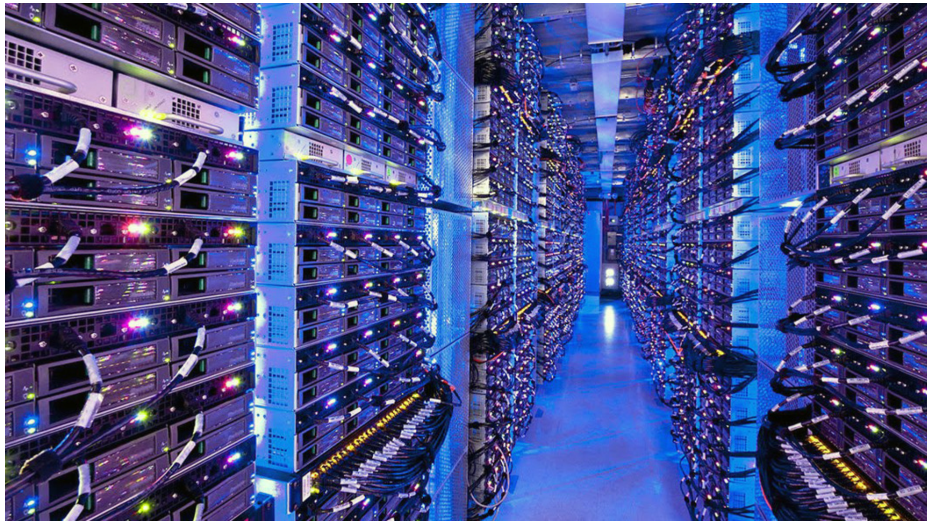
Highly parallelizable

Large model capacity & more parameters

Require more training data

Transformer is a very powerful model

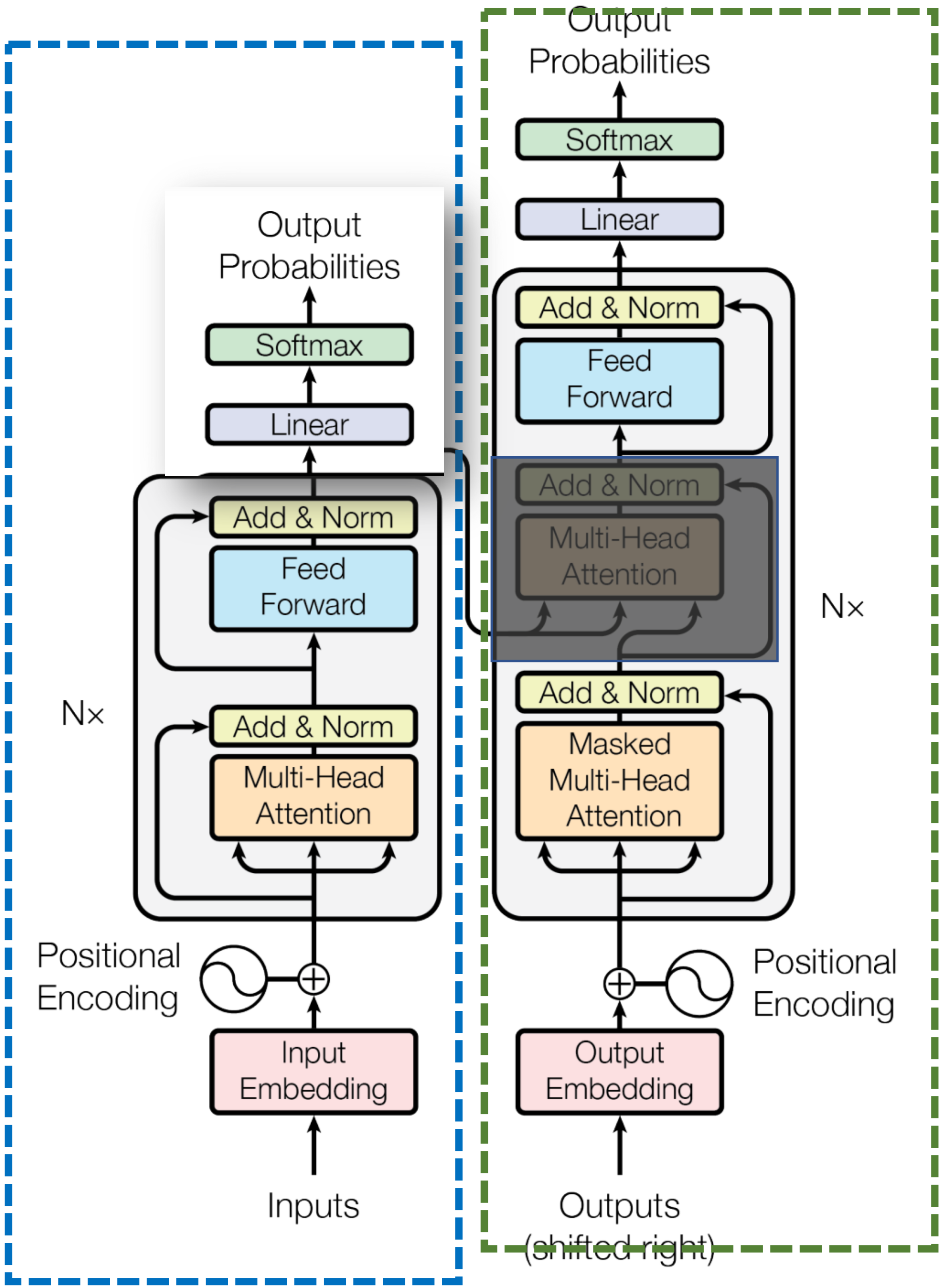
- Especially when there is a large amount of training data and resources



Encoder + Big data

auto-encoder

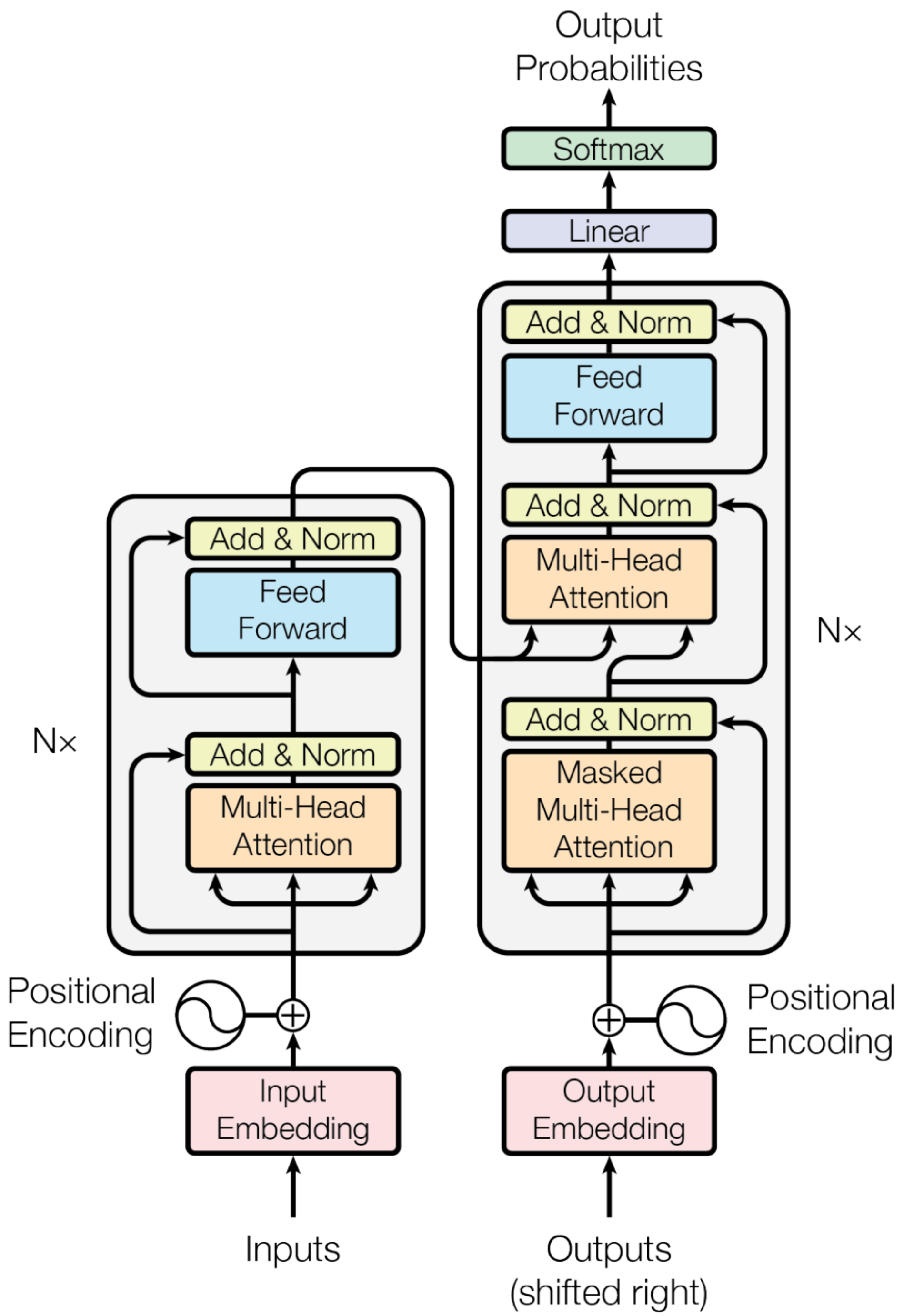
BERT (2018)
RoBERTa (2019)
ALBERT (2019)
DistilBERT (2019)
Reformer (2020)
Electra (2020)
…



Decoder + Big data

auto-regressive

GPT (2018)
GPT-2 (2019)
GPT-3 (2020)
GPT-4 (2023)
XLNet (2019)
…



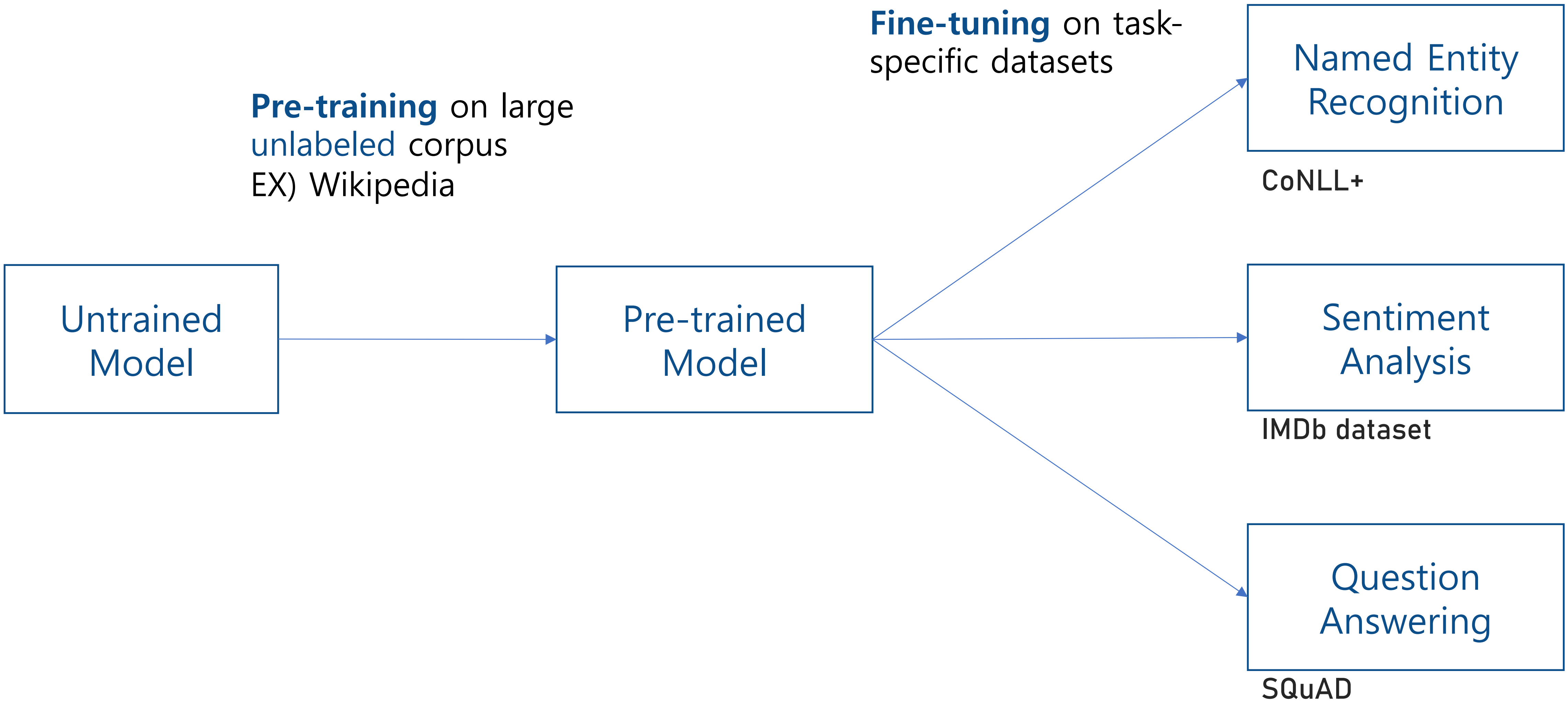
Encoder + Decoder + Big data

BART (2019)
T5 (2020)
mBART (2020)
...

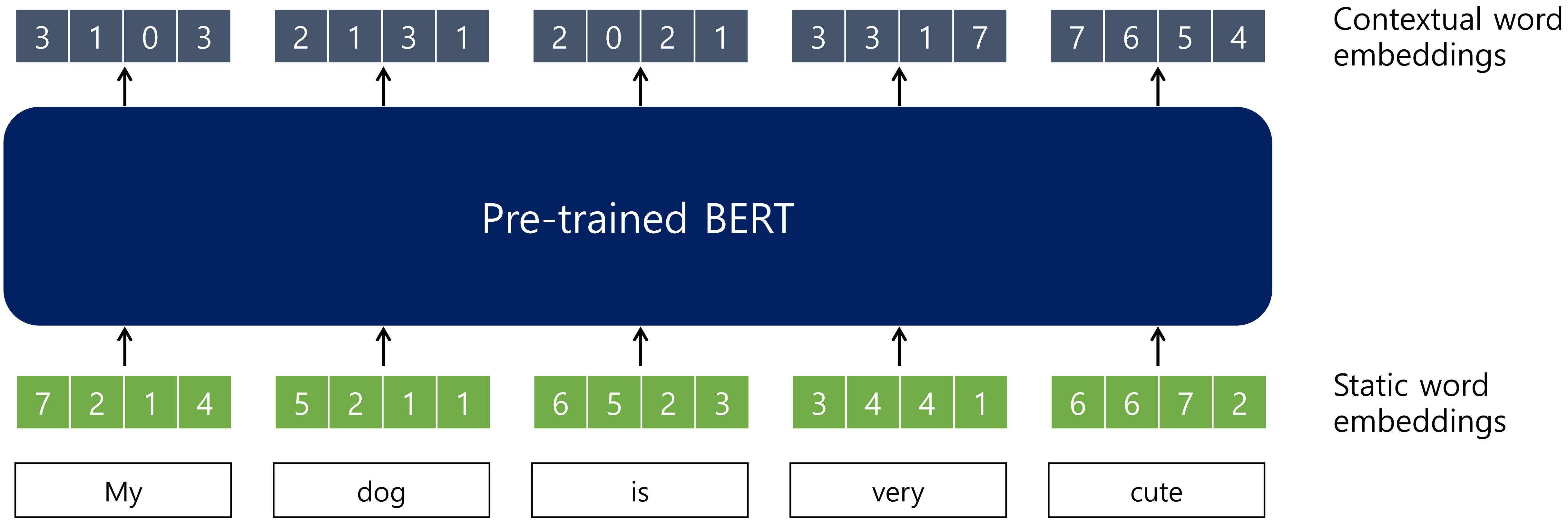
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

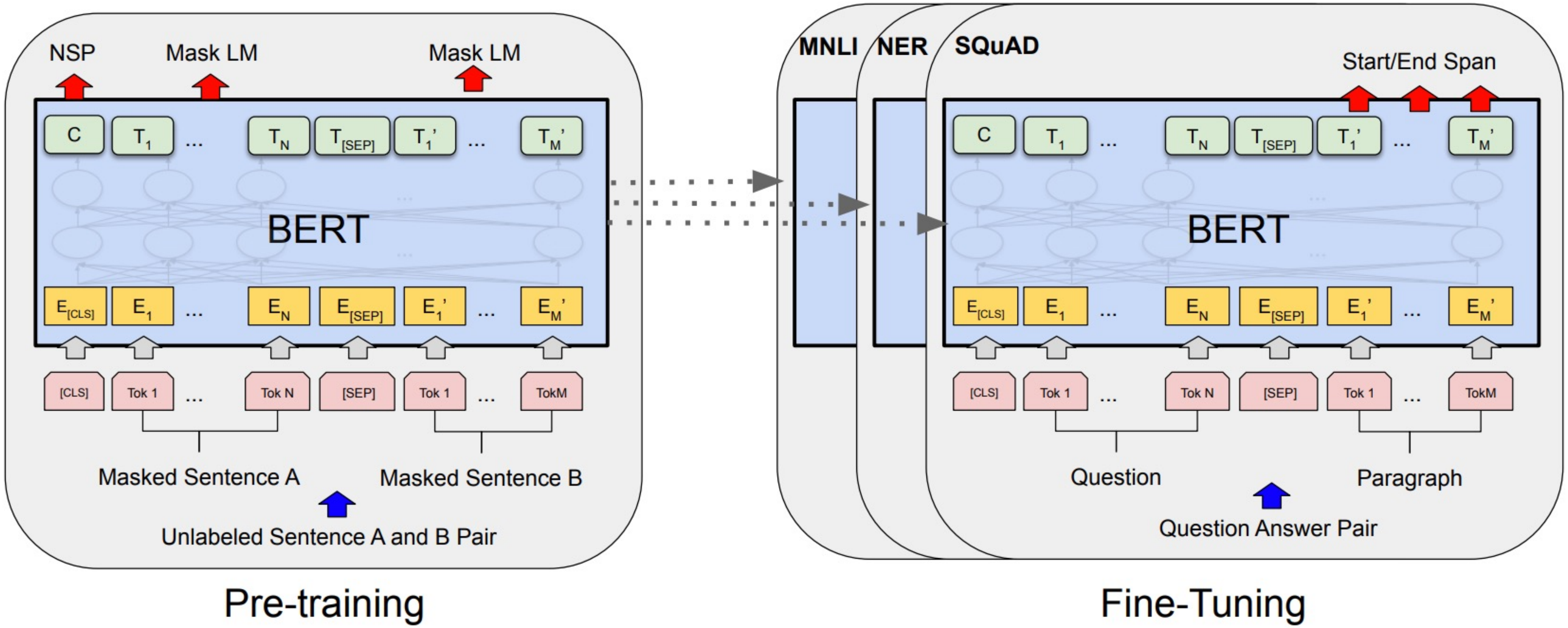
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

2018


Pre-training on large
unlabeled corpus
EX) Wikipedia
Fine-tuning on task-
specific datasets
Untrained
Model
Pre-trained
Model
Named Entity
Recognition
CoNLL+
Sentiment
Analysis
IMDb dataset
Question
Answering
SQuAD

BERT (Bidirectional Encoder Representations from Transformers)



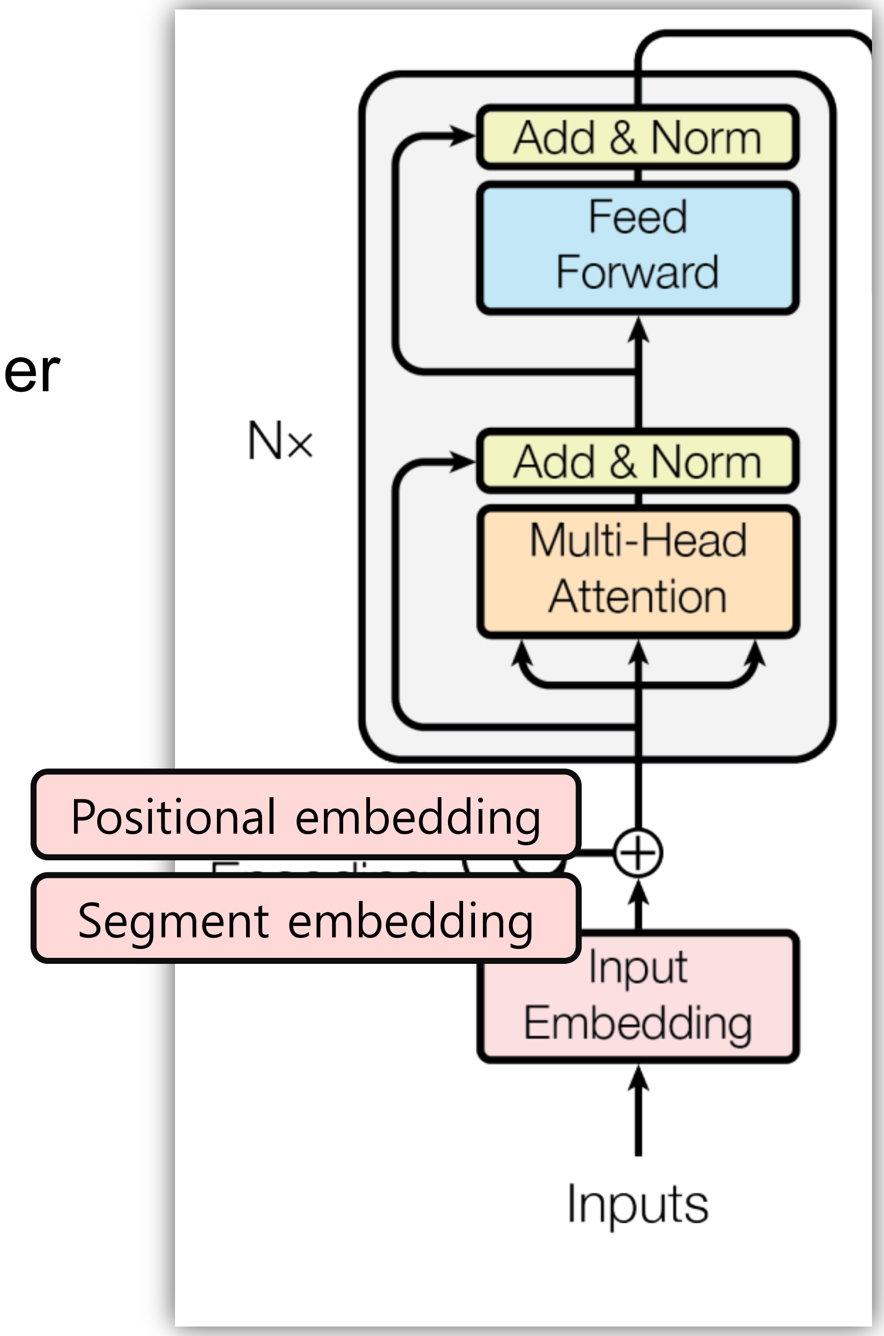


- The architecture
- Pretraining
- Fine-tuning

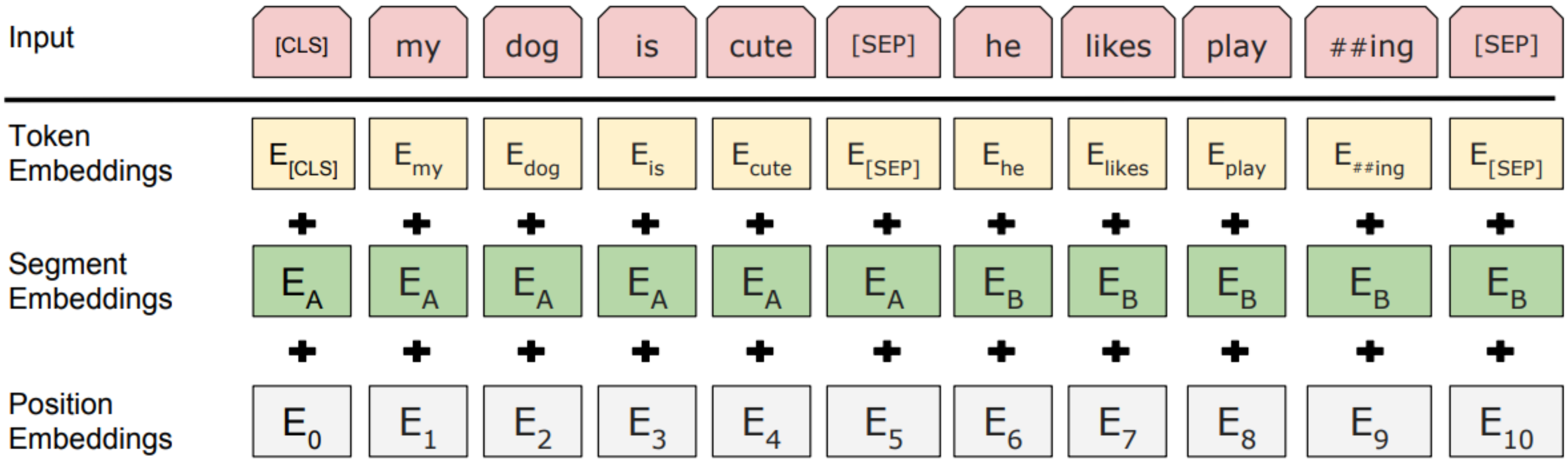
Model Architecture

- Multi-layer bidirectional Transformer encoder
- Two variations

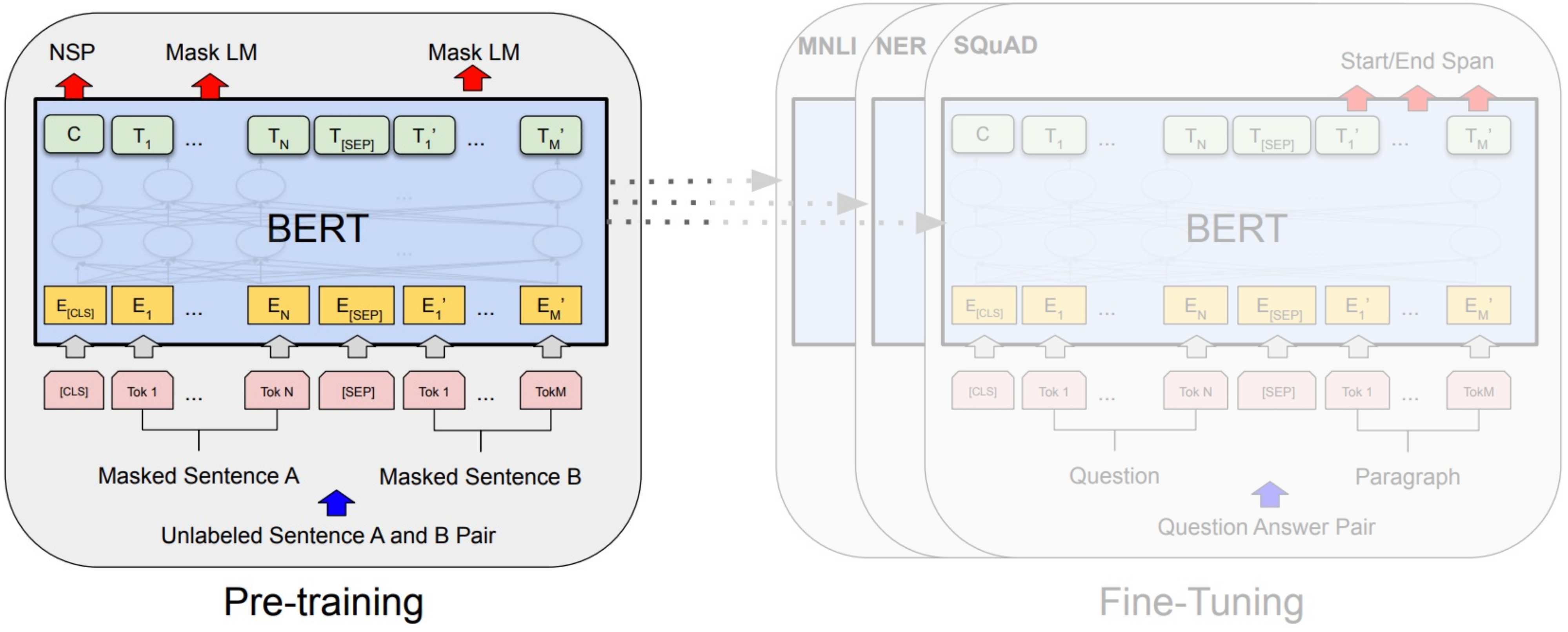
	BERT_{BASE}	BERT_{LARGE}
N: # layers	12	24
H: hidden size	768	1024
A: # attention heads	12	16
Total # params	110M	340M



Input representation

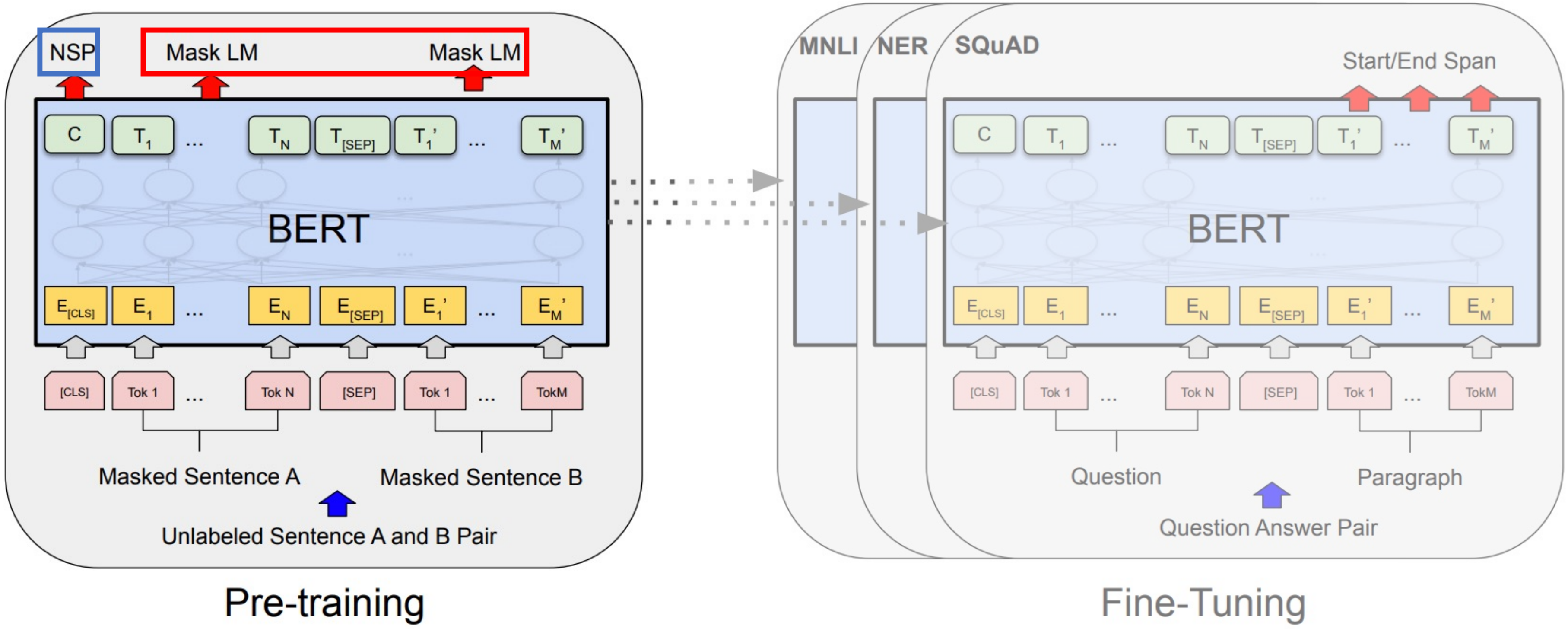


Pretraining

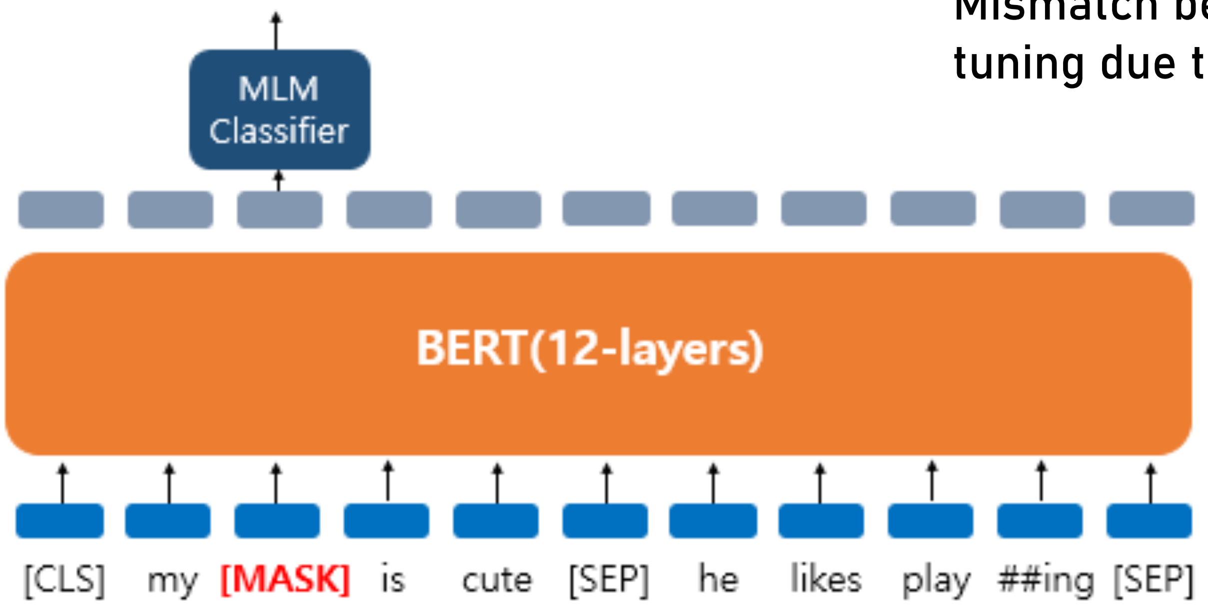


Pre-training tasks

- Masked language model (MLM)
  - Train a deep bidirectional representation
- Next sentence prediction (NSP)
  - Train a model that understands sentence relationships



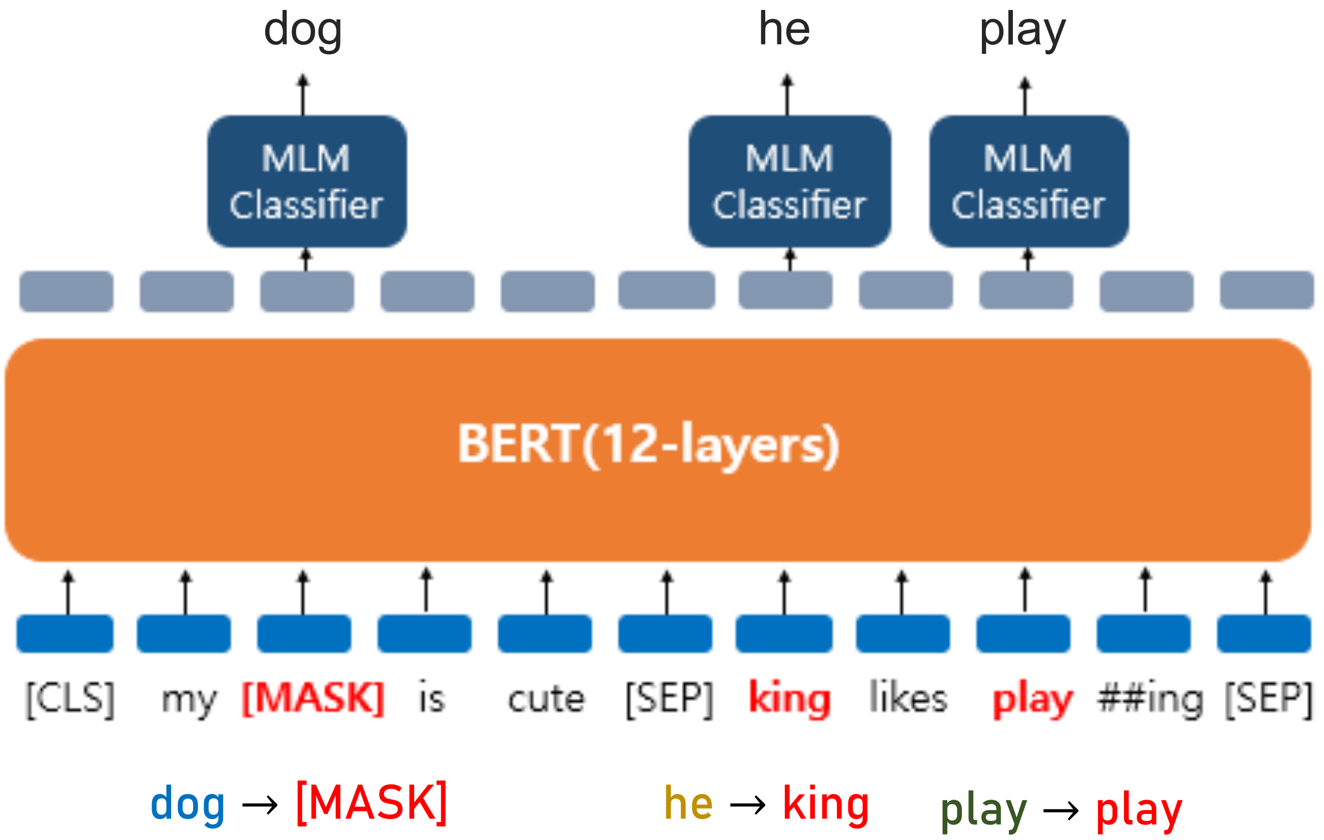
Masked Language Model (MLM)



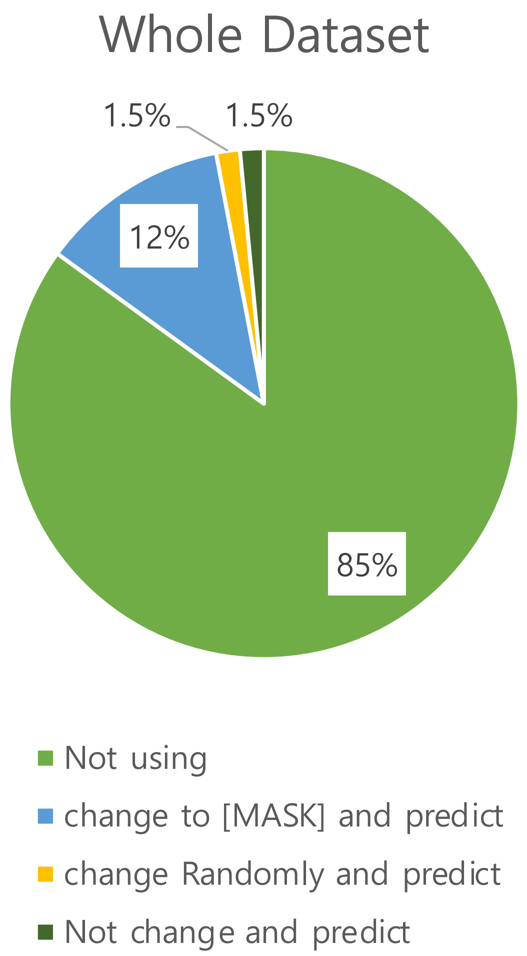
Mismatch between pre-training and fine-tuning due to the use of [MASK] tokens

- Train a deep bidirectional representation
- Mask 15% of input tokens at random
- Predict the masked tokens

Masked Language Model (MLM)
: Mitigating the mismatch

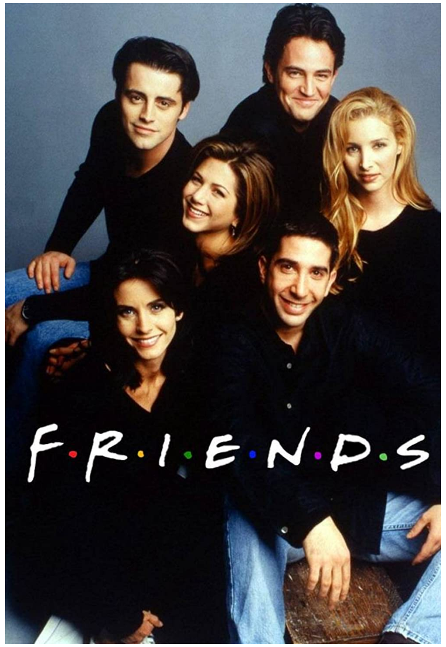


My **dog** is cute. **he** likes **play**ing



Next sentence prediction(NSP)

- Train a model that understands sentence relationships
- Example)



Monica: This is harder than I thought it would b
e.

Chandler: Oh, it is going to be okay.

Rachel: Do you guys have to go to the new hous
e right away, or do you have some time?

Monica: We got some time.

Rachel: Okay, should we get some coffee?

Chandler: Sure. Where?

Next sentence prediction(NSP)

Monica: This is harder than I thought it would be.

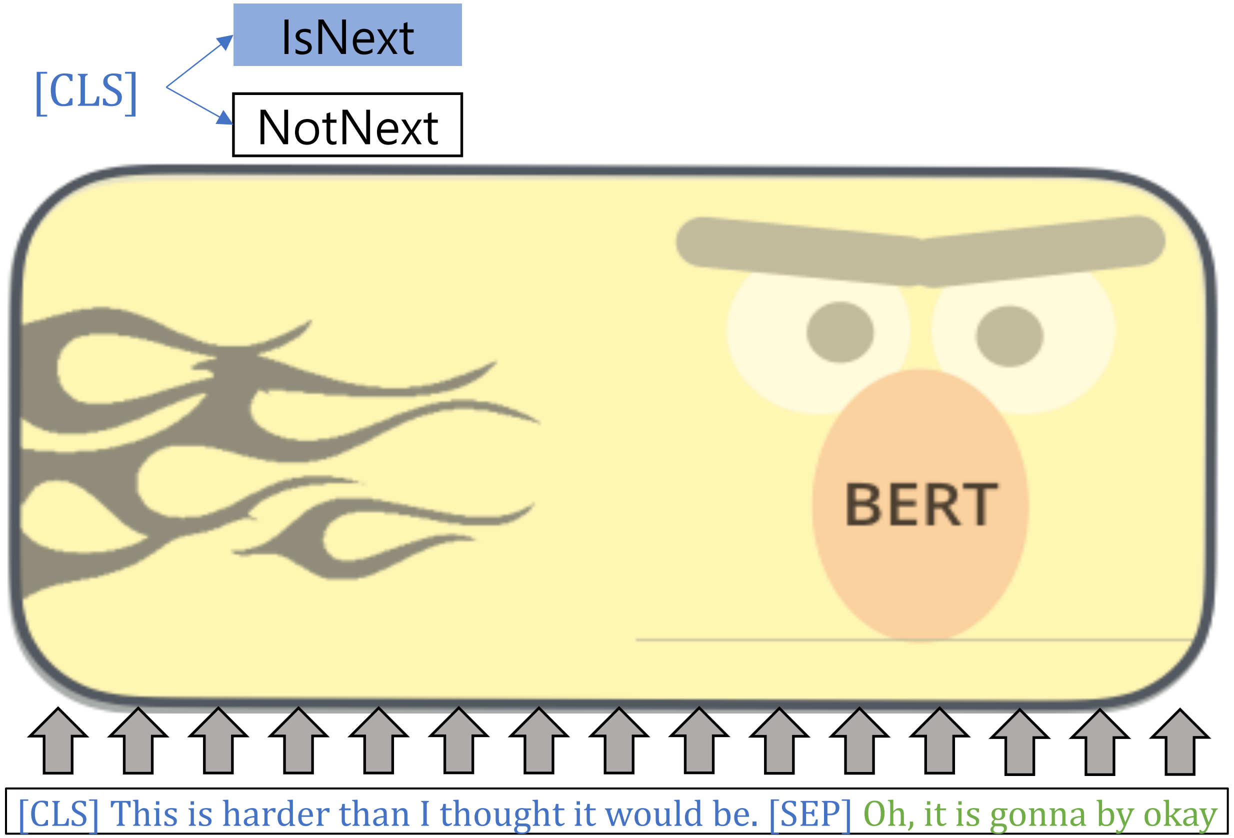
Chandler: Oh, it is going to be okay.

Rachel: Do you guys have to go to the new house right away, or do you have some time?

Monica: We got some time.

Rachel: Okay, should we get some coffee?

Chandler: Sure. Where?



Next sentence prediction(NSP)

Monica: This is harder than I thought it would be.

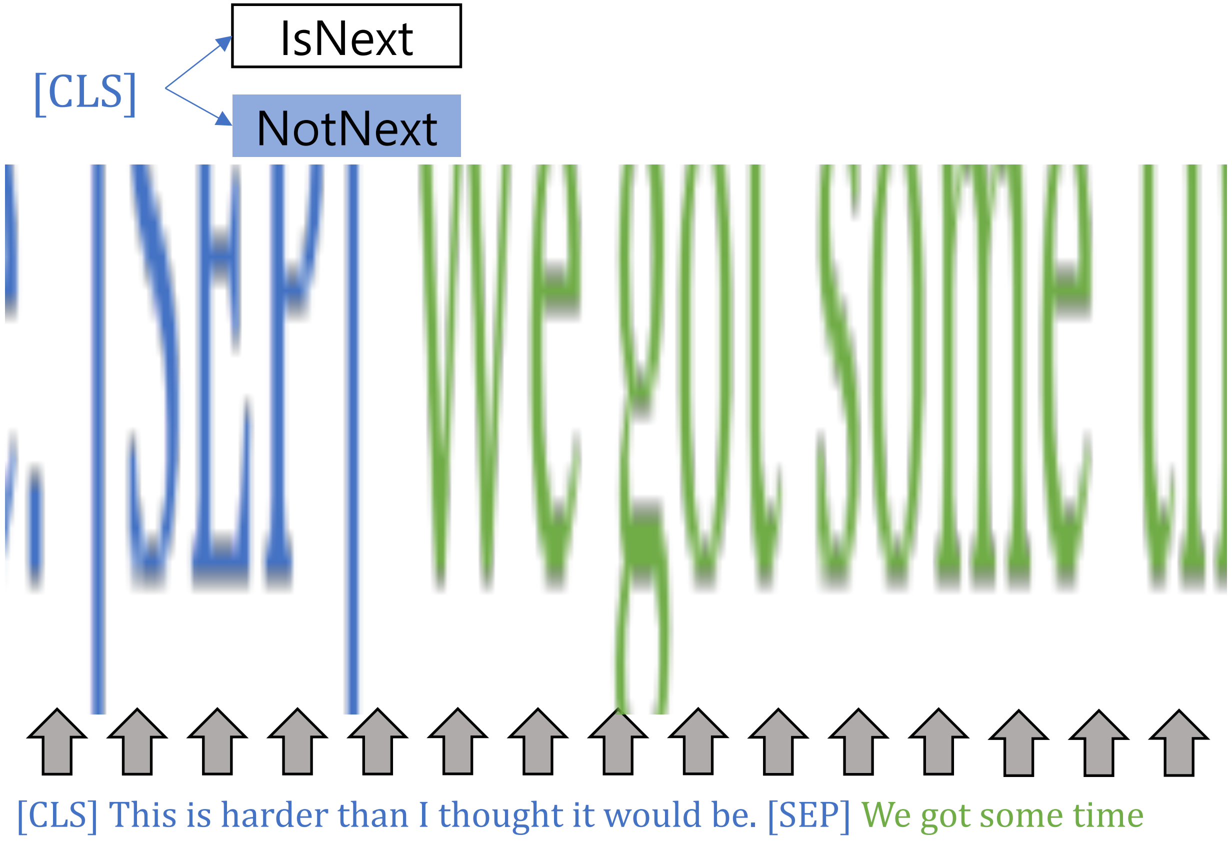
Chandler: Oh, it is going to be okay.

Rachel: Do you guys have to go to the new house right away, or do you have some time?

Monica: We got some time.

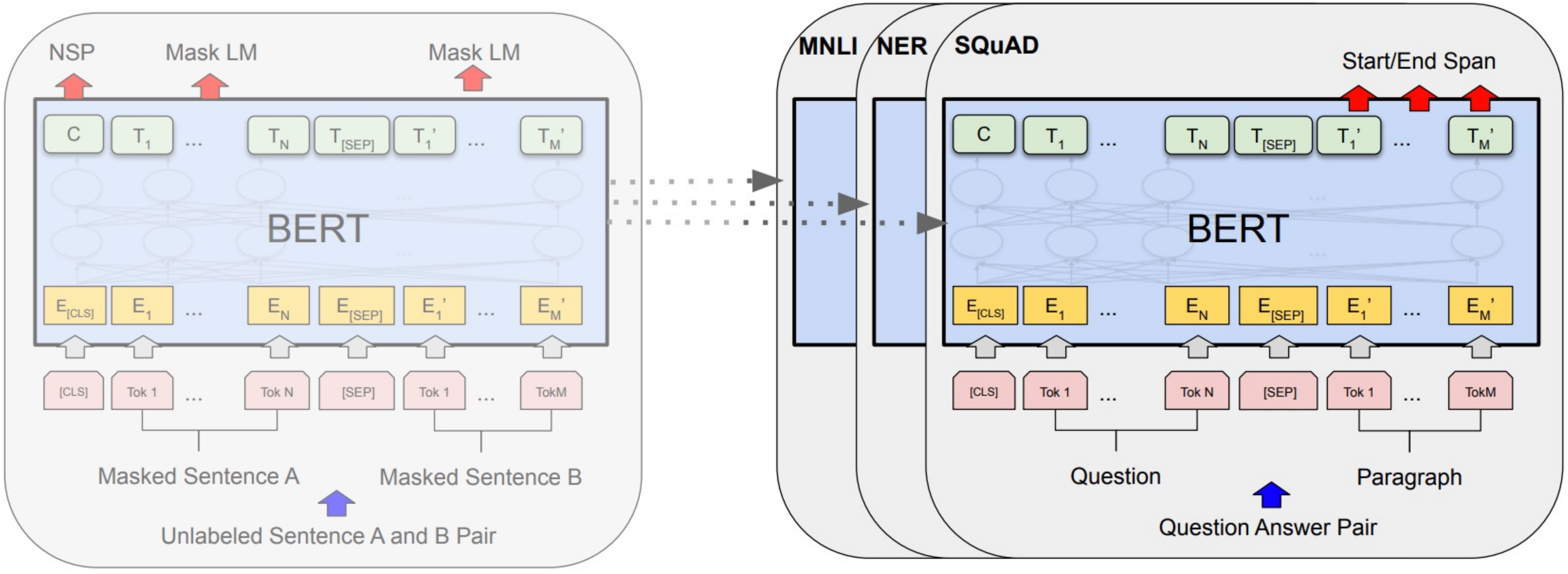
Rachel: Okay, should we get some coffee?

Chandler: Sure. Where?



Youtube 고려대학교 산업경영공학부 DSBA 연구실 08-5 : BERT

Fine-Tuning



Pre-training
Fine-Tuning

Fine-tuning
: Single-text classification

- Examples
  - Sentiment analysis
  - News classification
- Use [CLS] token for classification

Similar to the many-to-1 topology in RNN!

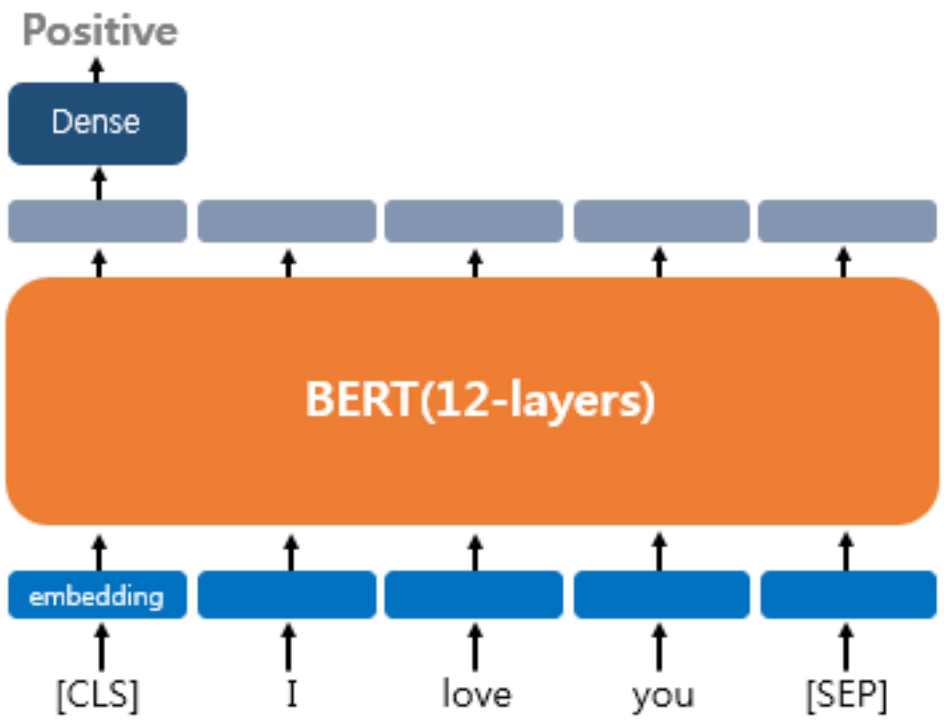


Image reference : wikidocs 딥 러닝을 이용한 자연어 처리 입문 – B

Fine-tuning
: Tagging

- Examples
  - Part-of-speech (POS) Tagging
  - Named Entity Recognition (NER)
- Use [CLS] token for classification

Similar to the many-to-many topology in RNN!

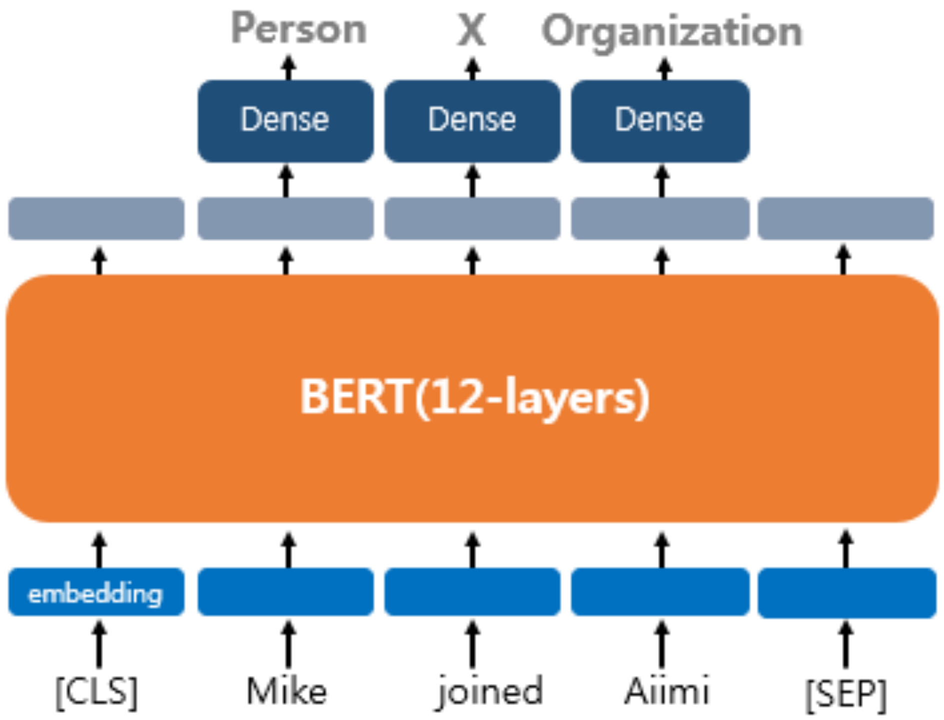
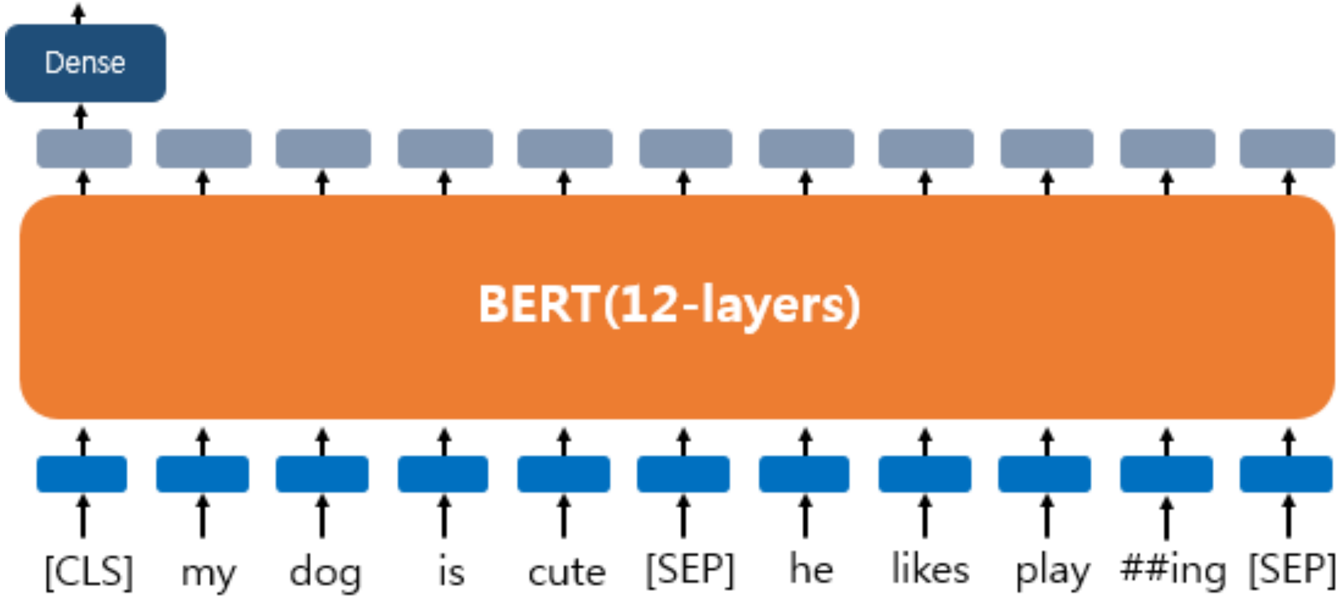


Image reference : wikidocs 딥 러닝을 이용한 자연어 처리 입문 – B

Fine-tuning
: Text-pair classification



Example)
Natural Language Inference (NLI)

Infer the relationship between two sentences
3 classes (contradiction, entailment, neutral)

Image reference : wikidocs 딥 러닝을 이용한 자연어 처리 입문 – B

Fine-tuning
: Question answering

Passage (P) + Question (Q) → Answer (A)

Document, Context

Extractive QA , sub-sequence

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.......

Q Why did Alyssa go to Miami?

A To visit some friends

Fine-tuning
: Question answering

Passage (*P*) + Question (*Q*) → Answer (*A*)

Document, Context

Extractive QA , sub-sequence

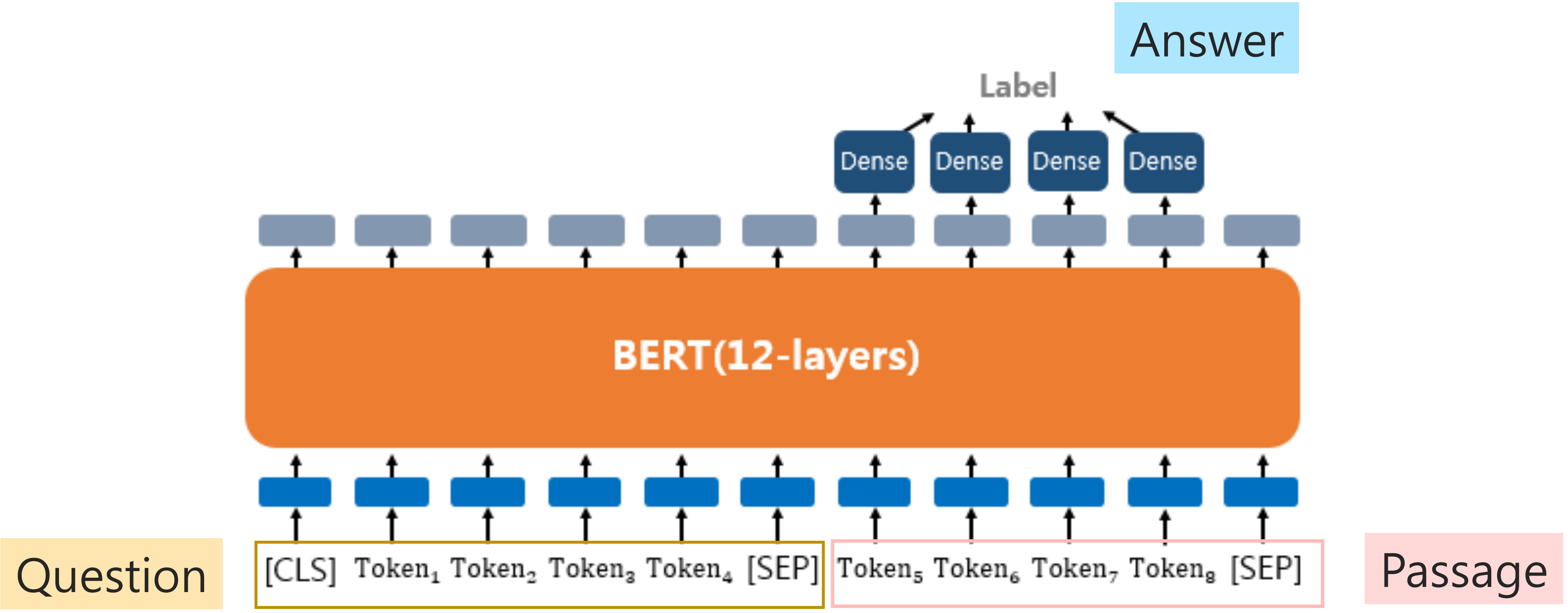


Image reference : wikidocs 딥 러닝을 이용한 자연어 처리 입문 – B

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.