MI2RL Medical Imaging & Intelligent Reality Lab.

www.mi2rl.co

공학 논문 발표 : BERT

2021.11.24

김인환

Convergence Medicine/Radiology,
University of Ulsan College of Medicine
Asan Medical Center

울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

01

목차

내용

1)Motivation

2) 기존 RNN의 문제와 Transformer의 등장

3) BERT

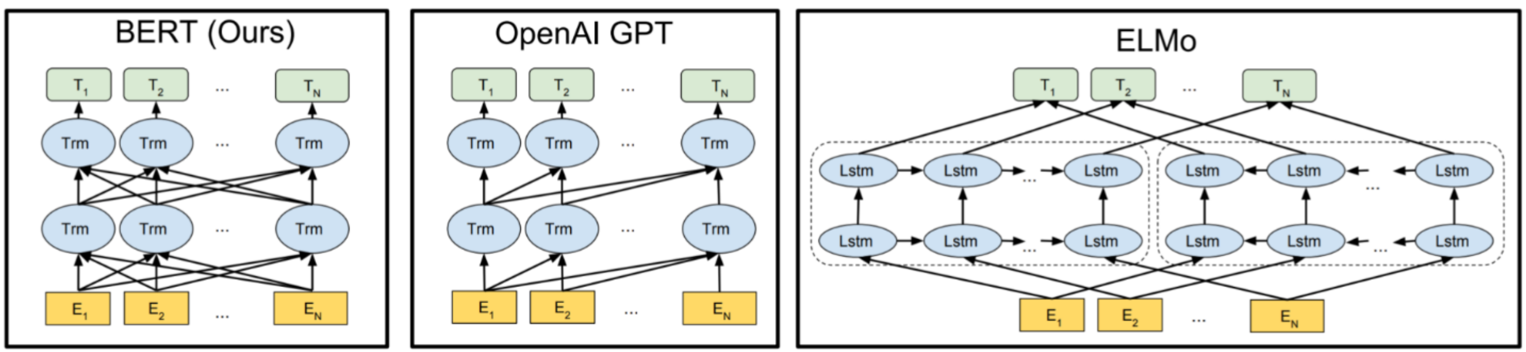
01

Motivation

Big Model 의 시초

대용량의 데이터를 사전 학습 후 이를 이용해 사용자가 원하는 모델로 Fine turning 시 높은 정확도의 모델을 제작할수 있게 됨.

- Differences in pre-training model architectures



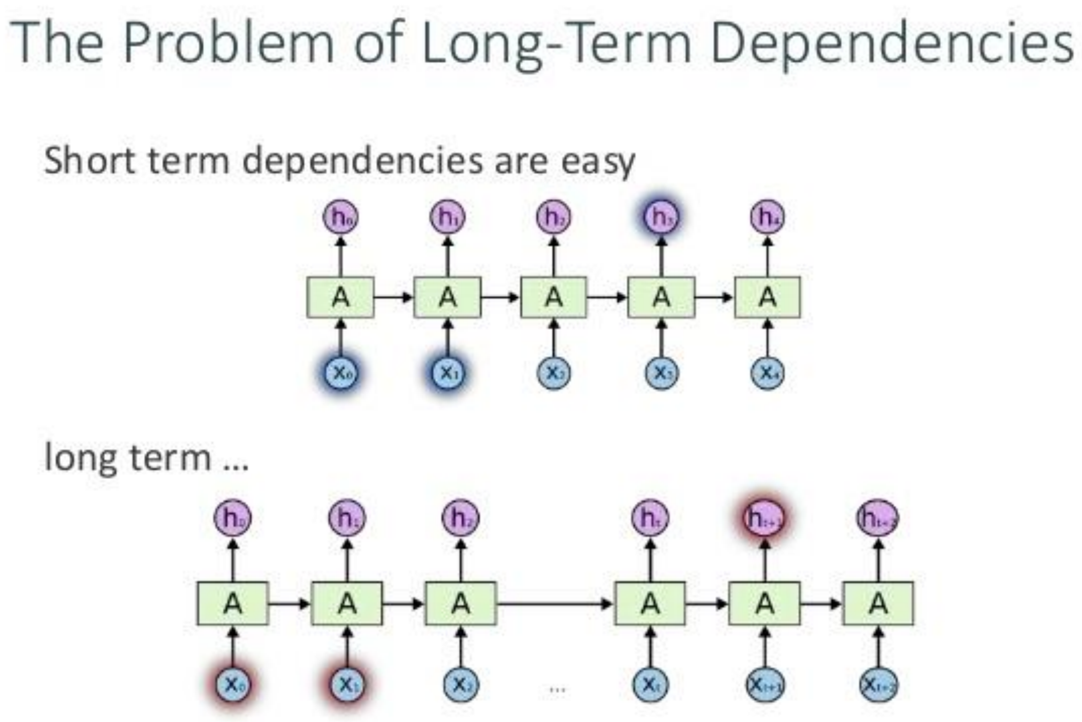
기존에 존재하던 pre training model 들은 단방향 학습으로 진행 되었지만 BERT는 양방향 학습을 이용하여 이전 연구된 모델보다 높은 정확도를 도출함.

기존 RNN의 문제와 Transformer 등장

01

기존 RNN의 문제와 Transformer 등장

RNN 의 Long term Dependency 문제



01

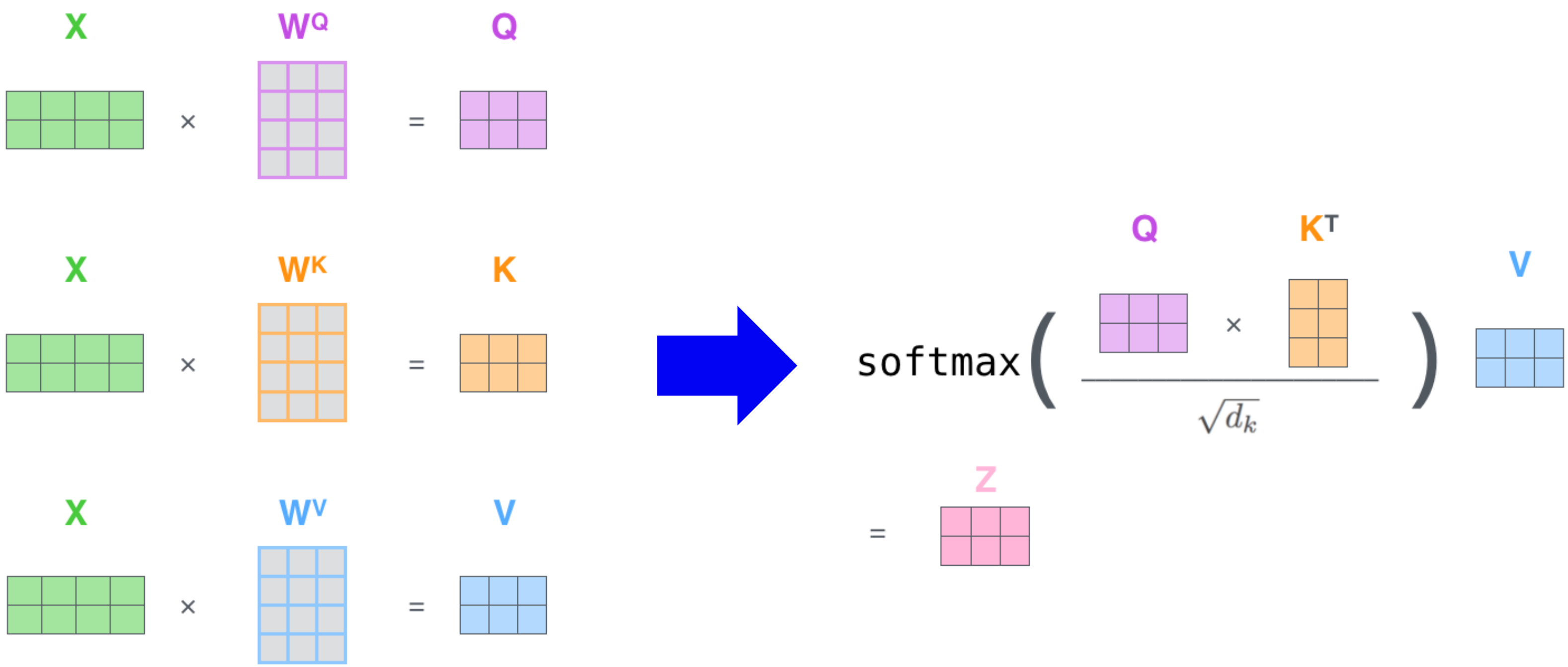
기존 RNN의 문제와 Transformer 등장

RNN 의 Long term Dependency 문제만 있는것은 아닙니다;;;

울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

기존 RNN의 문제와 Transformer 등장

Attention 등장!



가장 많이 받는 질문!

Attention 메커니즘의 장점은 무엇인가요???

정확히 어떤 구조로 동작되나요???

자세하게 알아봅시다.

01

기존 RNN의 문제와 Transformer 등장

문장에서 단어 간의 의미 및 연결성 찾기!

A dog ate the food because it was hungry.

이때 it 이 의미하는게 과연 무엇일까???

기존 RNN의 문제와 Transformer 등장

문장에서 단어 간의 의미 및 연결성 찾기!

A $\boxed{\text{dog}}$ ate the food because $\boxed{\text{it}}$ was hungry.

이때 it 이 의미하는게 과연 무엇일까???

dog!

01

기존 RNN의 문제와 Transformer 등장

문장에서 단어 간의 의미 및 연결성 찾기!

A dog ate the food because it was hungry.

이때 it 이 의미하는게 과연 무엇일까???

dog!

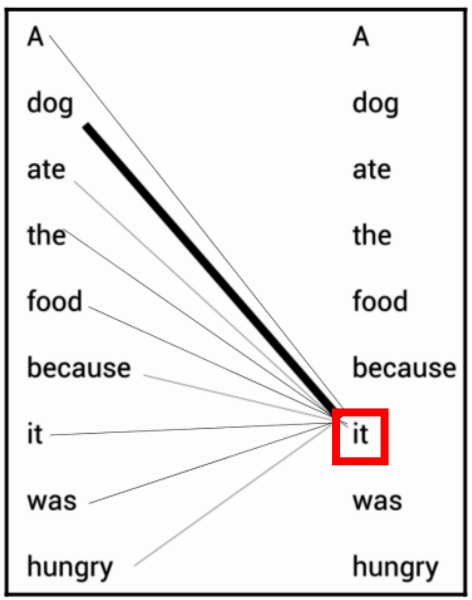
이걸 딥러닝 모델이 어떻게 찾도록 할까???

01

기존 RNN의 문제와 Transformer 등장

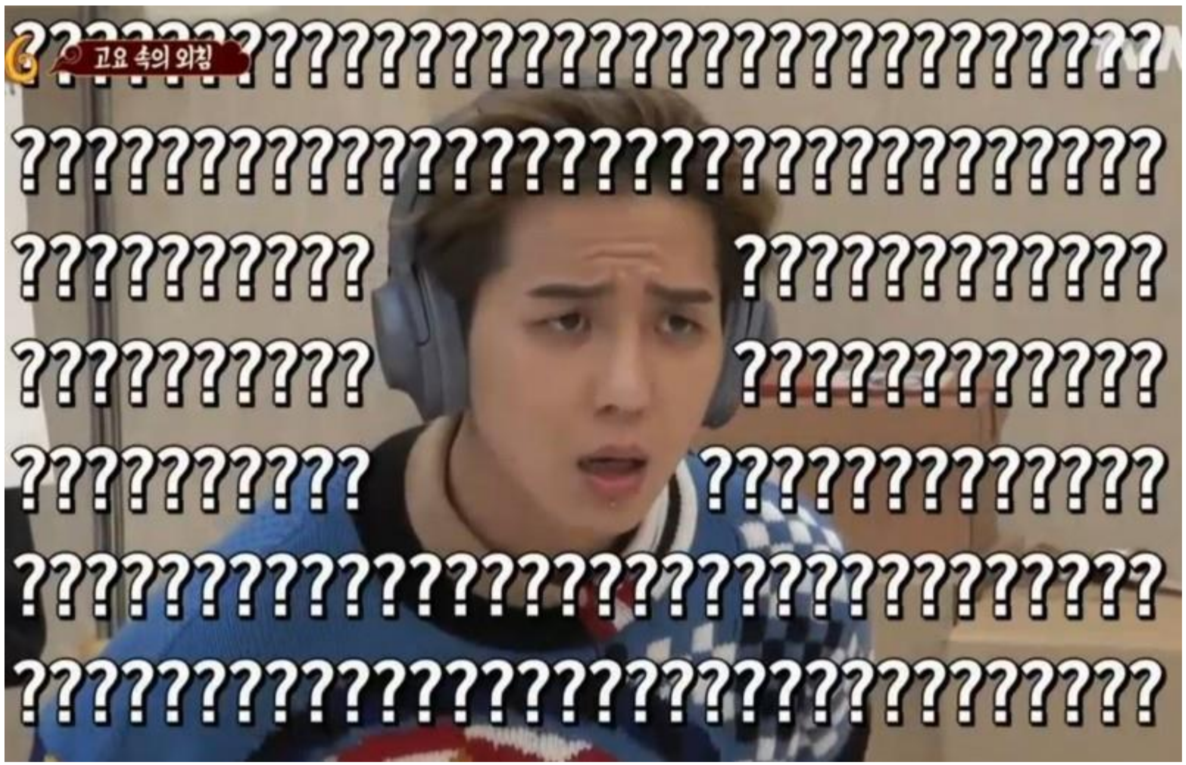
모든 단어와 연결성 및 유사도를 계산하여 가장 의미 있는 단어와 연결!

A dog ate the food because it was hungry.



단어 it 과 문장내의 모든 단어들 간의 관련성을 판단!

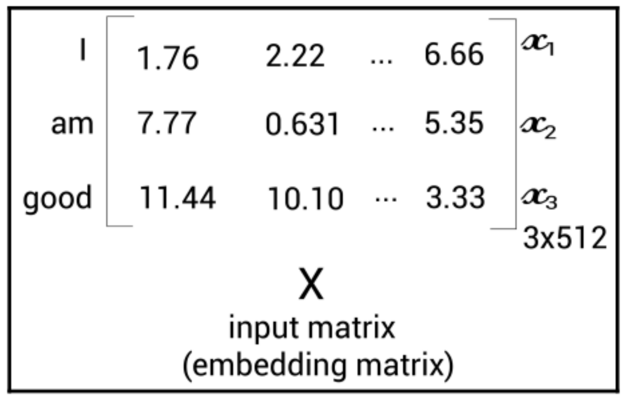
어떻게요???



01

기존 RNN의 문제와 Transformer 등장

첫번째로 단어를 Vector로 변경해 줍니다!



이미지와는 다르게 단어는 별도의 숫자로 되어 있지않음!

고로 단어를 Vector로 변경해 줘야 합니다!

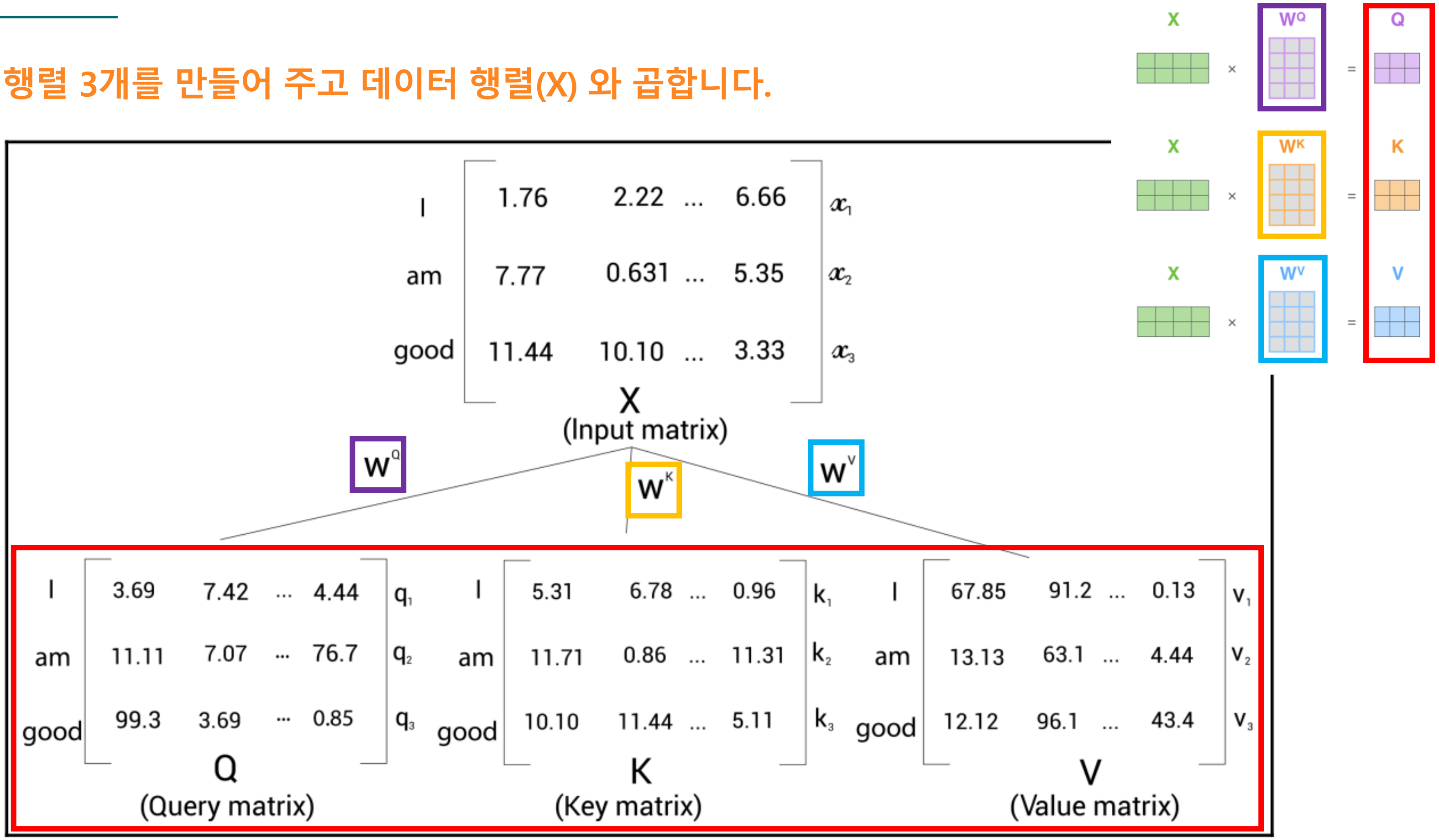
참고자료

Efficient Estimation of Word Representations in Vector Space
GloVe: Global Vectors for Word Representation

01

기존 RNN의 문제와 Transformer 등장

Weight 행렬 3개를 만들어 주고 데이터 행렬(X) 와 곱합니다.

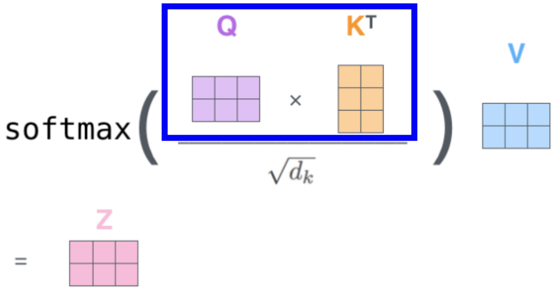


W(Q,K,V) 를 학습하는 것이 목적입니다!
일단 우리는 W가 학습이완료된 것으로 가정하고 설명 진행하겠습니다@@

01

기존 RNN의 문제와 Transformer 등장

각 단어 별로 유사도를 구합니다.



$$
\begin{array}{c|ccccc|c}
 & & & & & \\
\text{I} & 3.69 & 7.42 & \cdots & 4.44 & & q_1 \\
\text{am} & 11.11 & 7.07 & \cdots & 76.7 & & q_2 \\
\text{good} & 99.3 & 3.69 & \cdots & 0.85 & & q_3
\end{array}
\cdot
\begin{array}{ccc}
\text{I} & \text{am} & \text{good} \\
\hline
5.31 & 11.71 & 10.10 \\
6.78 & 0.86 & 11.44 \\
\vdots & \vdots & \vdots \\
0.96 & 11.31 & 5.11 \\
\hline
k_1 & k_2 & k_3
\end{array}
=
\begin{array}{c|ccc|}
 & \text{I} & \text{am} & \text{good} \\
\text{I} & q_1.k_1 & q_1.k_2 & q_1.k_3 \\
\text{am} & q_2.k_1 & q_2.k_2 & q_2.k_3 \\
\text{good} & q_3.k_1 & q_3.k_2 & q_3.k_3
\end{array}
$$

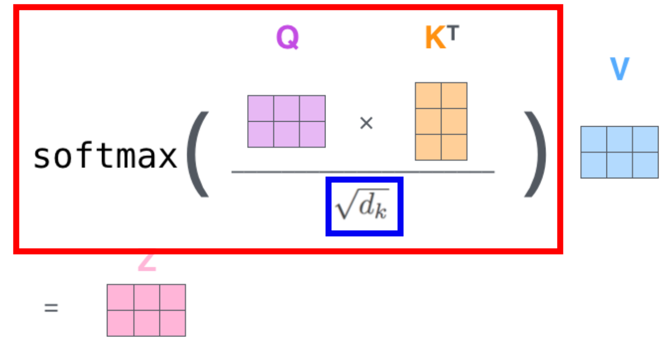
$$Q \qquad K^T$$

$$
QK^T =
\begin{array}{c|ccc|}
 & \text{I} & \text{am} & \text{good} \\
\text{I} & 110 & 90 & 80 \\
\text{am} & 70 & 99 & 70 \\
\text{good} & 90 & 70 & 100
\end{array}
$$

01

기존 RNN의 문제와 Transformer 등장

유사도를 정규화 한 후 attention score를 구합니다.



$$\frac{QK^T}{\sqrt{d_K}} = \frac{QK^T}{8} = \begin{array}{c|ccc} & \text{I} & \text{am} & \text{good} \\ \hline \text{I} & 13.75 & 11.25 & 10 \\ \text{am} & 8.75 & 12.375 & 8.75 \\ \text{good} & 11.25 & 8.75 & 12.5 \end{array}$$

$$\text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right) = \begin{array}{c|ccc} & \text{I} & \text{am} & \text{good} \\ \hline \text{I} & 0.90 & 0.07 & 0.03 \\ \text{am} & 0.025 & 0.95 & 0.025 \\ \text{good} & 0.21 & 0.03 & 0.76 \end{array}$$

01

기존 RNN의 문제와 Transformer 등장

Attention score를 이용한 weighted L C을 계산합니다!

$$\text{softmax}\left(\frac{Q \times K^{T}}{\sqrt{d_k}}\right) V = Z$$

$$Z = \underbrace{\begin{array}{c|ccc} & \text{I} & \text{am} & \text{good} \\ \hline \text{I} & 0.90 & 0.07 & 0.03 \\ \text{am} & 0.025 & 0.95 & 0.025 \\ \text{good} & 0.21 & 0.03 & 0.76 \end{array}}_{\text{softmax}\left(\frac{QK^{T}}{\sqrt{d_K}}\right)} \; \underbrace{\begin{array}{c|cccc|c} & \text{I} & \text{am} & & \text{good} & \\ \hline \text{I} & 67.85 & 91.2 & \ldots & 0.13 & v_1 \\ \text{am} & 13.13 & 63.1 & \ldots & 4.44 & v_2 \\ \text{good} & 12.12 & 96.1 & \ldots & 43.4 & v_3 \end{array}}_{V}$$

$$Z = \begin{array}{c|c} z_1 & \text{I} \\ z_2 & \text{am} \\ z_3 & \text{good} \end{array}$$

가중결합

$$z_1 = 0.90\underbrace{\boxed{67.85 \mid 91.2 \mid \ldots}}_{v_1(\text{I})} + 0.07\underbrace{\boxed{13.13 \mid 63.1 \mid \ldots}}_{v_2(\text{am})} + 0.03\underbrace{\boxed{12.12 \mid 96.1 \mid \ldots}}_{v_3(\text{good})}$$

기존 RNN의 문제와 Transformer 등장

1개로 믿을만 한가??

$$Z_1 = \mathrm{softmax}\left(\frac{Q_1 K_1^T}{\sqrt{d_k}}\right) V_1$$

$$Z_2 = \mathrm{softmax}\left(\frac{Q_2 K_2^T}{\sqrt{d_k}}\right) V_2$$

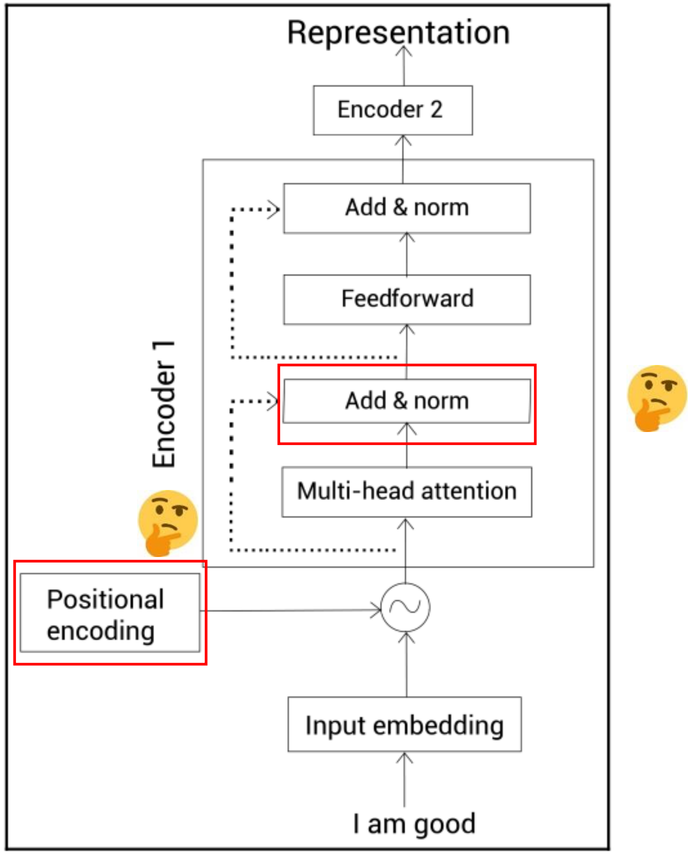
$$\vdots$$

$$\text{Multi-head attention} = \mathrm{Concatenate}(Z_1, Z_2, \ldots Z_i, \ldots Z_8) W_0$$

01

기존 RNN의 문제와 Transformer 등장

문장에서 단어의 순서 또한 중요합니다! -> 문장 내 단어의 순서도 학습하자!



BERT 준비완료

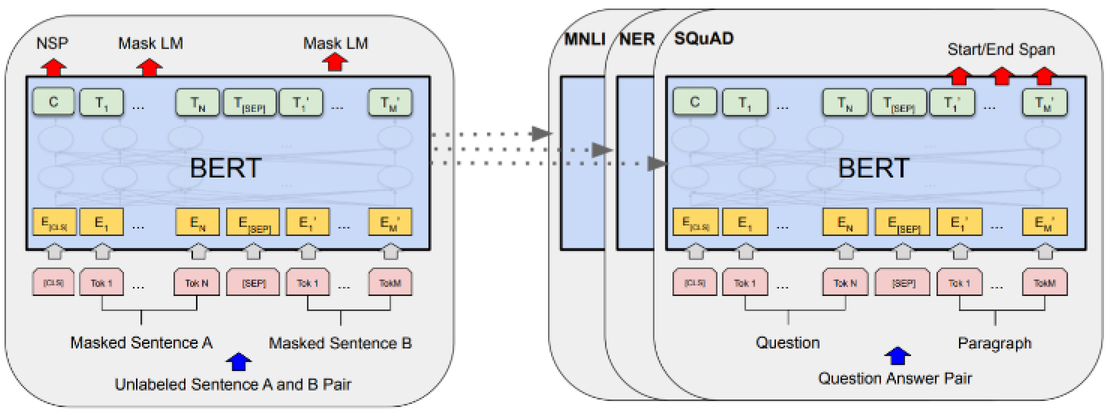
지금까지 Transformer 의 Encorder 부분을 봤습니다.

BERT에선 Transformer 의 Decorder 부분은 사용하지 않으므로
이번 발표에선 Encorder 에 대해서만 논하겠습니다@@

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- BERT
  - ✓ Designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers
    - Masked language model (MLM): bidirectional pre-training for language representations
    - Next sentence prediction (NSP)



Pre-training Fine-Tuning

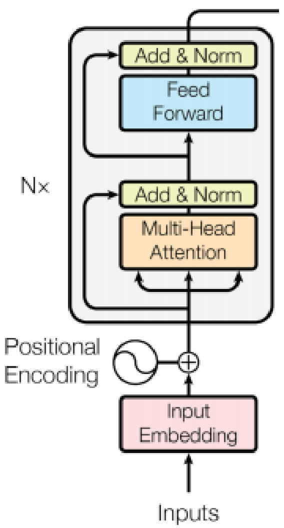
    - Pre-trained BERT model can be fine-tunes with just one additional output layer to create SOTA models for a wide range of NLP tasks (QA, NER, Sentiment Analysis, etc.)

MLM, NSP 학습방식은 데이터 획득에 유리.

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- BERT: Model Architecture
  - ✓ Multi-layer bidirectional Transformer encoder
    - L: number of layers (Transformer block)
    - H: hidden size
    - A: number of self attention heads
  - ✓ $\text{BERT}_{\text{BASE}}$
    - L = 12, H=768, A = 12
    - Total parameters = 110M
    - Same model size as OpenAI GPT
  - ✓ $\text{BERT}_{\text{LARGE}}$
    - L = 24, H=1,024, A = 16
    - Total parameters = 340M



BERT: Bidirectional Encoder Representations from Transformer

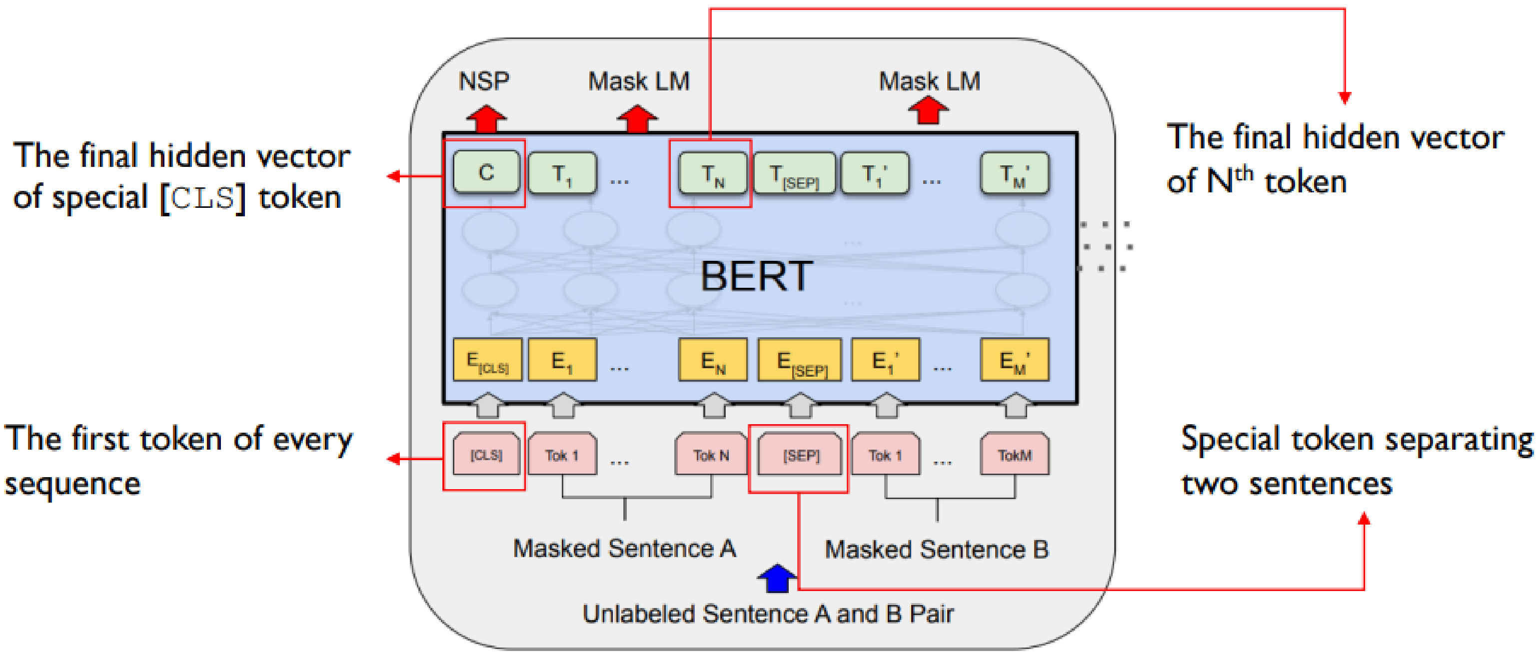
Devlin et. al (2018)

- BERT: Input/Output Representations
  - ✓ To make BERT handle a variety of down-stream tasks, the input representation is able to unambiguously represent both a single sentence and a pair of sentences (ex: Question-Answer)
    - **Sentence**: an arbitrary span of contiguous text, rather than an actual linguistic sentence
    - **Sequence**: the input token sequences to BERT, which may be a single sentence or two sentences packed together

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- BERT: Input/Output Representations

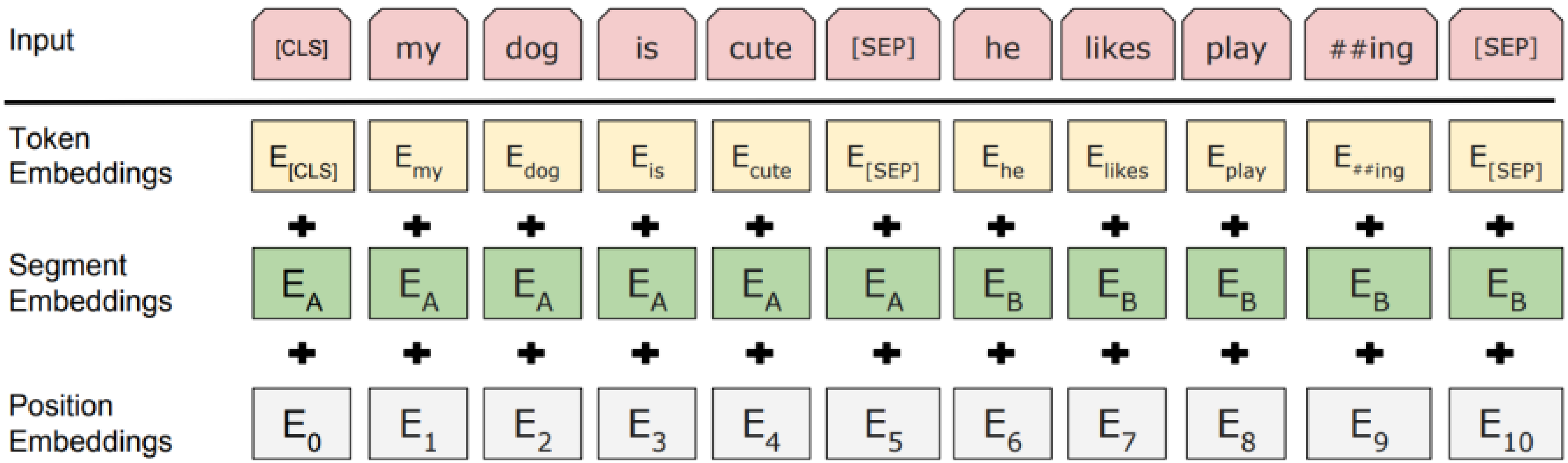


다음페이지에 상세하게....

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

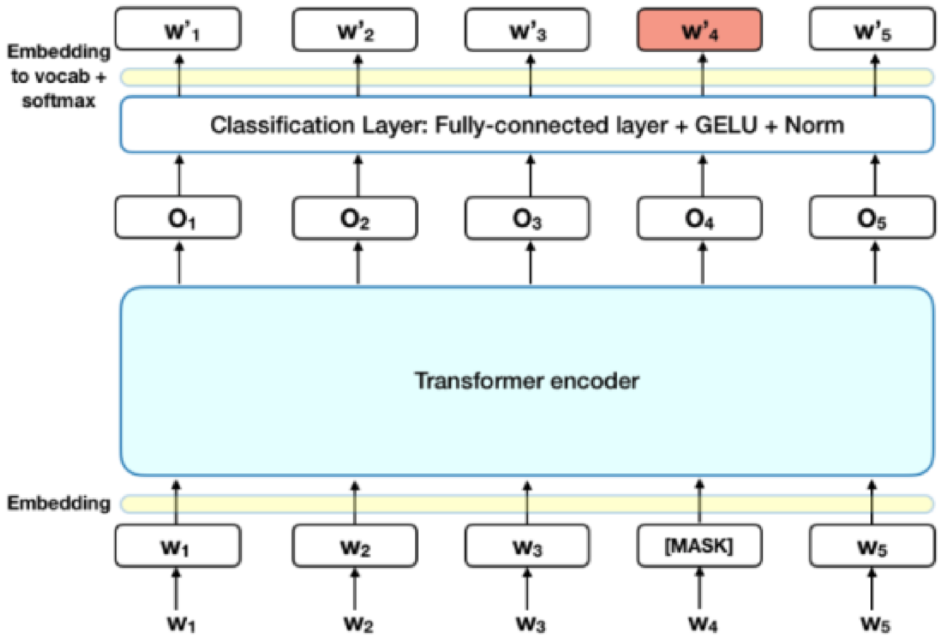
- BERT: Input/Output Representations
  - ✓ Input representation is the sum of
    - (1) Token embedding: WordPiece embeddings with a 30,000 token vocabulary
    - (2) Segment embedding
    - (3) Position embedding: same as in the Transformer



BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Pre-training BERT
  - ✓ Task 1: Masked Language Model (MLM)
    - ▪ 15% of each sequence are replaced with a [MASK] token
    - ▪ Predict the masked words rather tan reconstructing the entire input in denoising encoder



https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Pre-training BERT
  - ✓ Task 1: Masked Language Model (MLM)
    - (Caution!) A mismatch occurs between pre-training and fine-tuning, since the [MASK] token does not appear during fine-tuning
    - (Solution) If the i-th token is chosen to be masked, it is replaced by the [MASK] token 80% of the time, a random toke 10% of the time, and unchanged 10% of the time
      - (80%) `my dog is hairy → my dog is [MASK]`
      - (10%) `my dog is hairy → my dog is apple`
      - (10%) `my dog is hairy → my dog is hairy`

BERT: Bidirectional Encoder Representations from Transformer

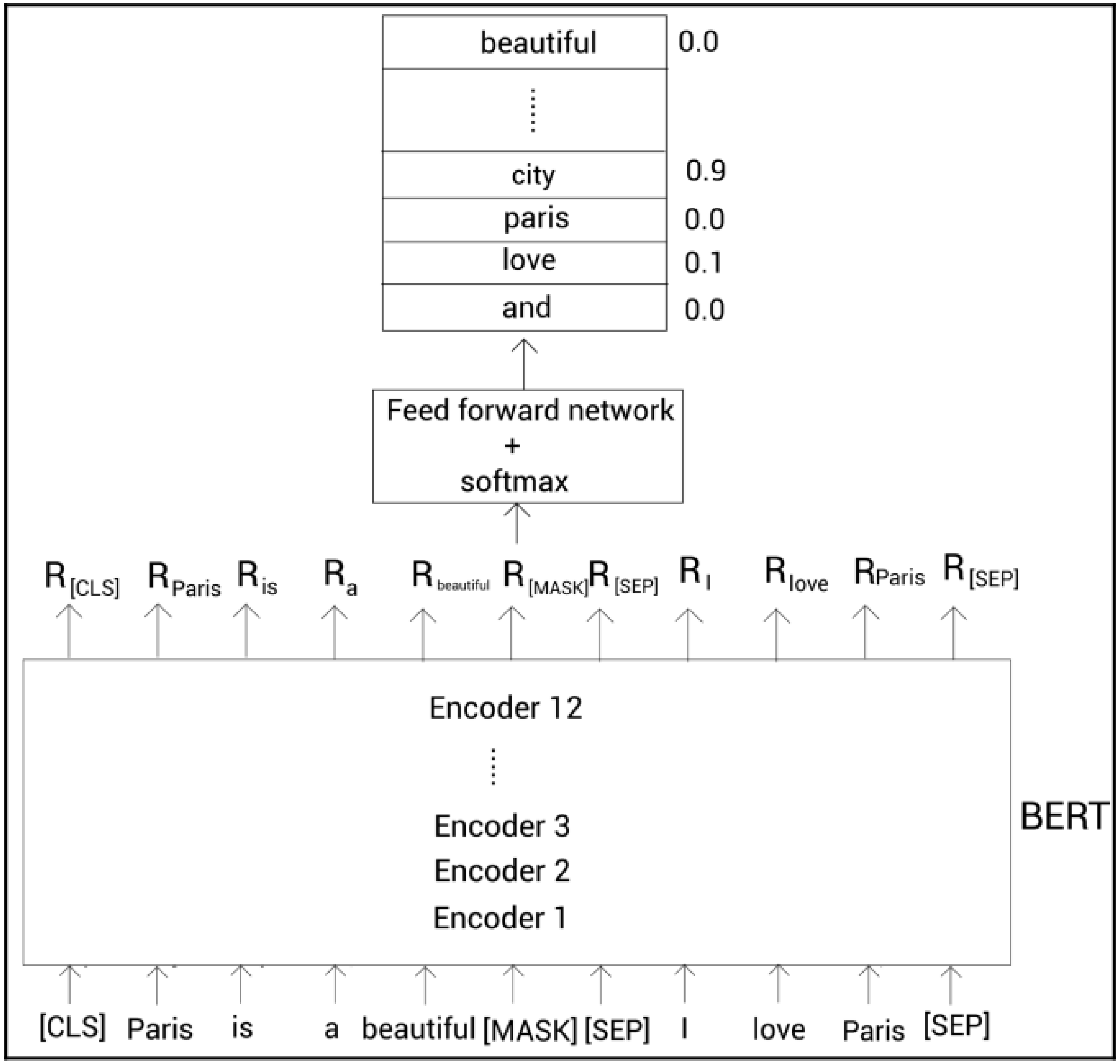
Devlin et. al (2018)

- Pre-training BERT
  - ✓ Task 1: Masked Language Model (MLM)
    - (Caution!) A mismatch occurs between pre-training and fine-tuning, since the [MASK] token does not appear during fine-tuning
    - (Solution) If the i-th token is chosen to be masked, it is replaced by the [MASK] token 80% of the time, a random toke 10% of the time, and unchanged 10% of the time

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)



BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Pre-training BERT
  - ✓ Task 2: Next Sentence Prediction (NSP)
    - Many important downstream tasks such as QA and NLI are based on understanding the relationship between two sentences, which is not directly captured by language modeling
    - A Binarized next sentence prediction task that can be trivially generated from any monolingual corpus is trained
      - 50% of the time B is the actual next sentence that follows A (IsNext)
      - 50% of the time it is a random sentence from the corpus (NotNext)
      - C is used for next sentence prediction
    - Despite its simplicity, pre-training towards this task is very beneficial both QA and NLI

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Pre-training BERT
  - ✓ Task 2: Next Sentence Prediction (NSP)

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

Chandler: Oh, it is gonna 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?

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Pre-training BERT
  - ✓ Task 2: Next Sentence Prediction (NSP)

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

Chandler: Oh, it is gonna be okay.

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

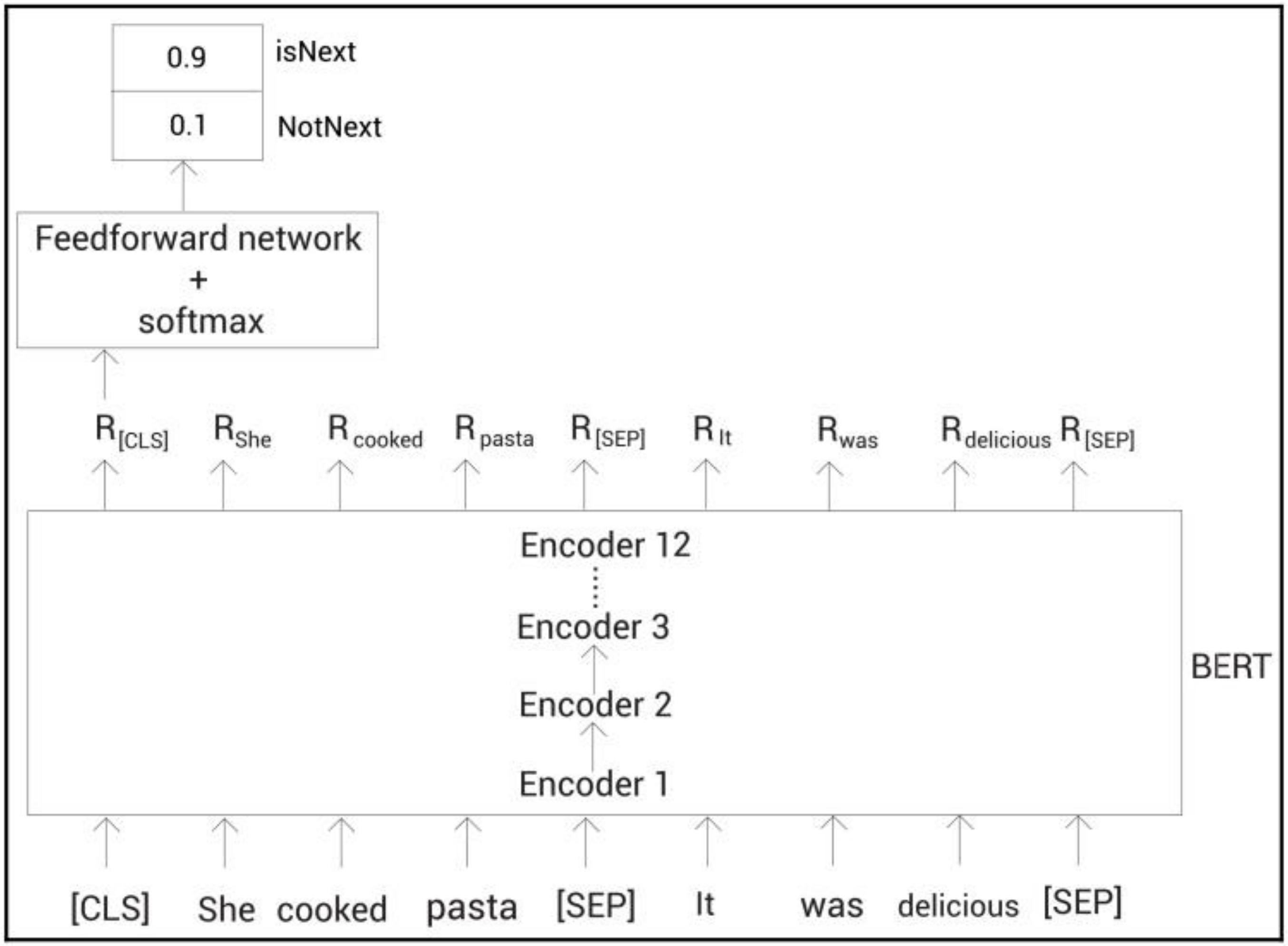
Monica: We got some time.

Rachel: Okay, should we get some coffee?

Chandler: Sure. Where?

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)



BERT: Bidirectional Encoder Representations from Transformer

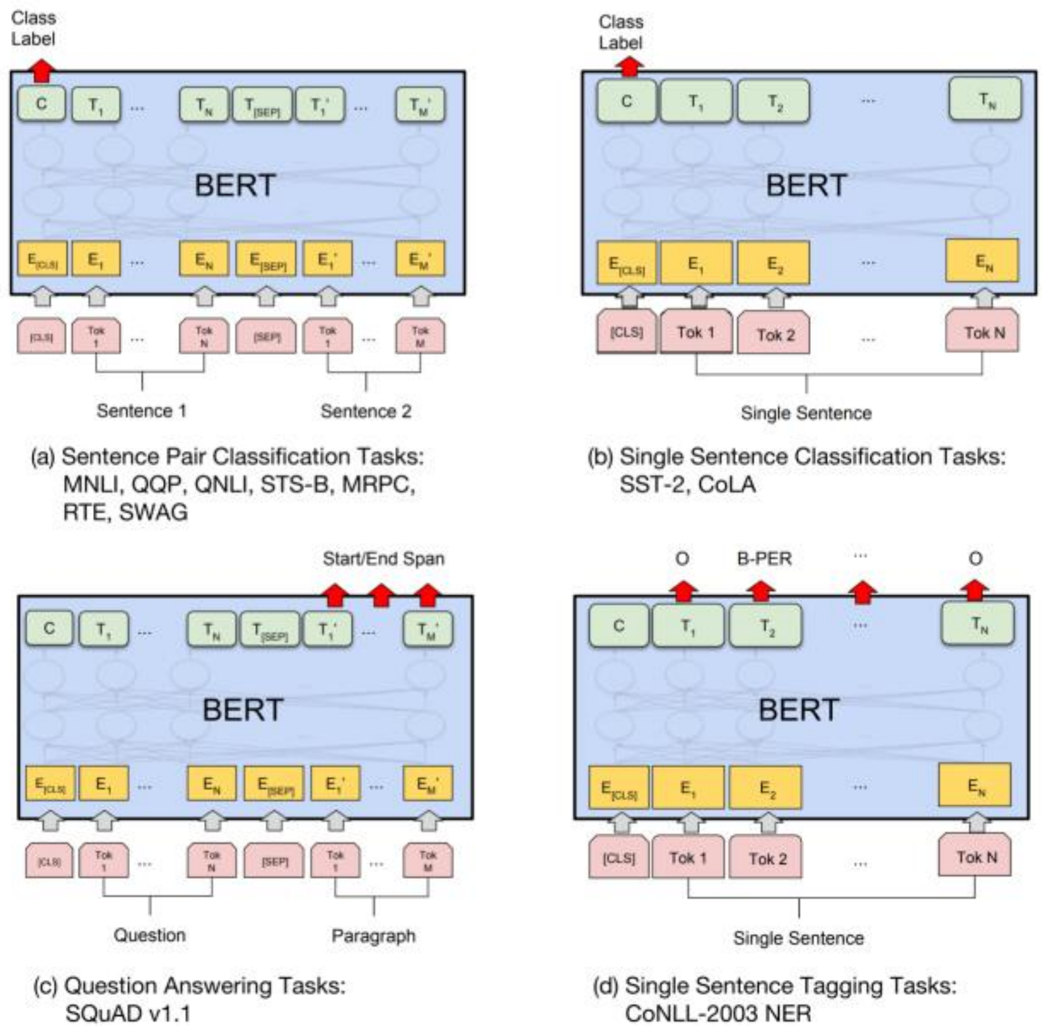
Devlin et. al (2018)

- Pre-training BERT
  - ✓ Hyper-parameter settings
    - Maximum token length: 512
    - Batch size: 256
    - Adam with learning rate of 1e-4, beta1 = 0.9 beta2 = 0.999
    - L2 weight decay of 0.01
    - Learning rate warmup over the first 10,000 steps, linear decay of the learning rate
    - Dropout probability of 0.1 on all layers
    - GeLU activation function rather than standard ReLU
    - $\text{BERT}_{\text{BASE}}$ took 4 days with 16 TPUs and $\text{BERT}_{\text{LARGE}}$ took 4 days with 64 TPUs
    - Pre-train the model with sequence length of 128 for 90% of the steps
    - The rest 10% of the steps are trained with sequence length of 512

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Fine-tuning BERT



(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: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Experiments
  - ✓ A collection of diverse NLU tasks

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	ERNIE Team - Baidu	ERNIE		90.2	72.2	97.5	93.0/90.7	92.9/92.5	75.2/90.8	91.2	90.6	98.0	90.9	94.5	49.4
2	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		90.1	73.2	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.8	90.6	99.2	87.4	94.5	48.7
3	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
4	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	53.1
5	XLNet Team	XLNet (ensemble)		89.5	70.2	97.1	92.9/90.5	93.0/92.6	74.7/90.4	90.9	90.9	99.0	88.5	92.5	48.4
6	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
7	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
8	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
9	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
10	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8

https://gluebenchmark.com/leaderboard

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
$\mathrm{BERT}_{\mathrm{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\mathrm{BERT}_{\mathrm{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Experiments
  - ✓ Ablation study 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

  - ✓ Ablation study 2: Effect of Model Size

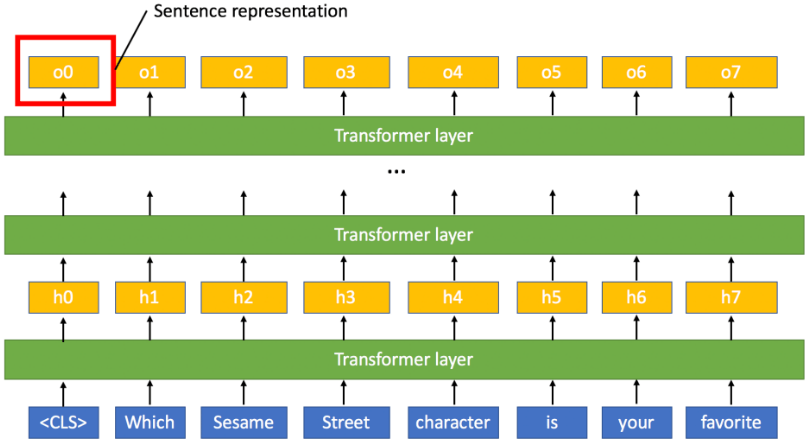
Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

BERT: Bidirectional Encoder Representations from Transformer

Devlin et. al (2018)

- Experiments
  - ✓ Ablation study 3: Feature-based Approach with BERT
    - CoNLL-2003 NER task

Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-



Collaborators

Radiology
Joon Beom Seo, SangMin Lee[A,B], Dong Hyun, Yang, Hyung Jin Won, Ho Sung Kim, Seung Chai Jung, Ji Eun Park, So Jung Lee, Jeong Hyun Lee, Gilsun Hong

Pathology
Hyunjeong Go, Gyuheon Choi, Gyungyub Gong, Dong Eun Song

Cardiology
Jaekwan Song, Jongmin Song, Young-Hak Kim

Anesthesiology
Sung-Hoon Kim, Eun Ho Lee

Neurology
Dong-Wha Kang, Chongsik Lee, Jaehong Lee, Sangbeom Jun, Misun Kwon, Beomjun Kim

Surgery
Beom Seok Ko, JongHun Jeong, Songchuk Kim, Tae-Yon Sung

Internal Medicine
Jeongsik Byeon, Kang Mo Kim

Emergency Medicine
Dong-Woo Seo

LG · CGBio · OSSTEM IMPLANT · kakaobrain · VUNO · ANYMEDI · CORELINE SOFT · MI2L · 울산대학교 의과대학 · 서울아산병원 Asan Medical Center