

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

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

Google AI Language

`{jacobdevlin, mingweichang, kentonl, kristout}@google.com`



Presenter: WENWEI KANG

- Introduction
- Transformer
- Input Representation
- Pre-train
- Fine-tuning
- Evaluation

Introduction

BERT: Bidirectional Encoder Representations from Transformers

- **Pre-train** on both **left and right** context in all layers.
 - Task#1: Masked LM(Language Model)
 - Task#2: NSP (Next Sentence Prediction)
- **Fine-tuning** with one additional output layer.
 - 2 sentence-level tasks
 - 9 token-level tasks

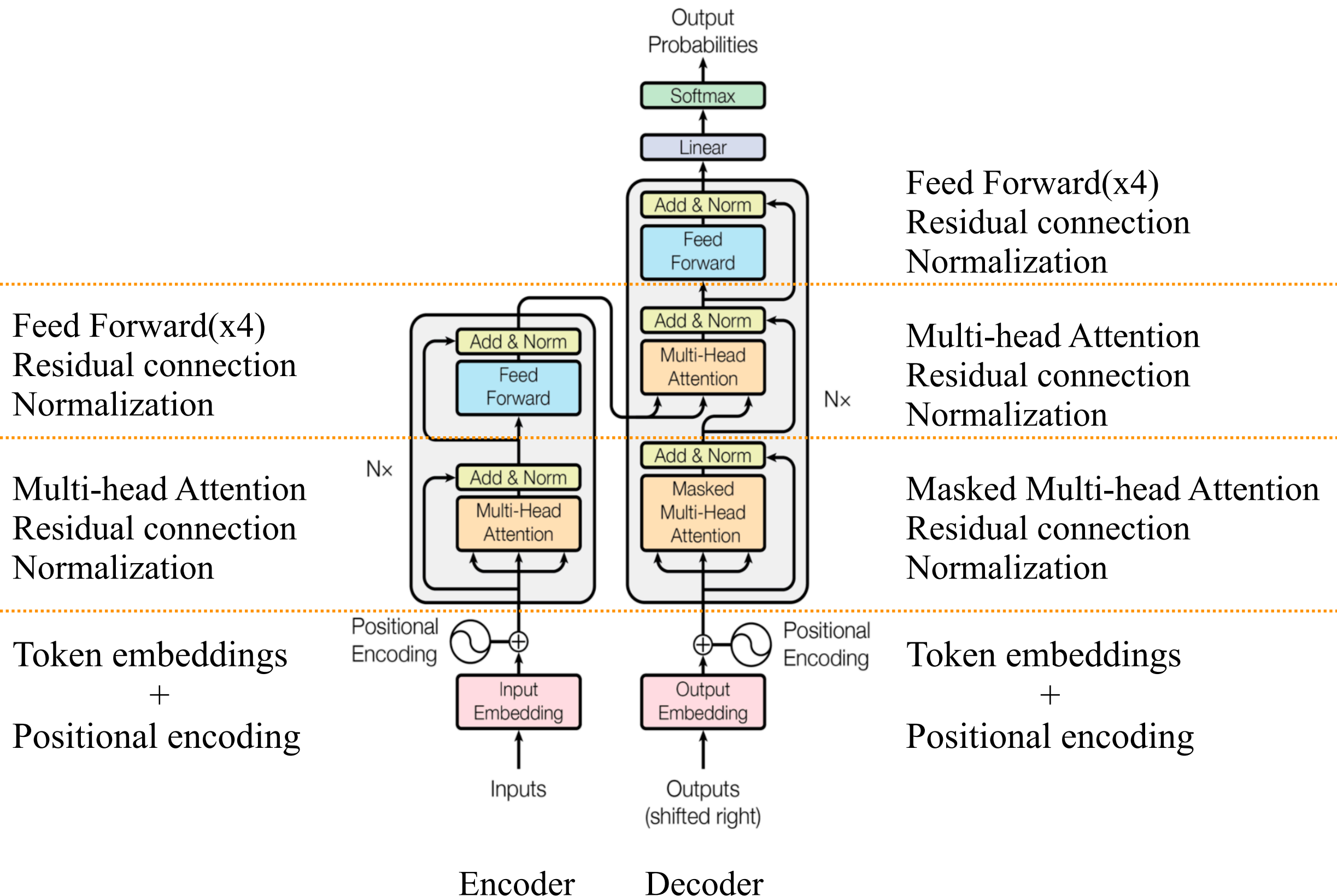
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	Jacob Devlin	BERT: 24-layers, 1024-hidden, 16-heads		80.4	60.5	94.9	85.4/89.3	87.6/86.5	89.3/72.1	86.7	85.9	91.1	70.1	65.1
		BERT: 12-layers, 768-hidden, 12-heads		78.3	52.1	93.5	84.8/88.9	87.1/85.8	89.2/71.2	84.6	83.4	90.1	66.4	65.1

SQuAD1.1 Leaderboard

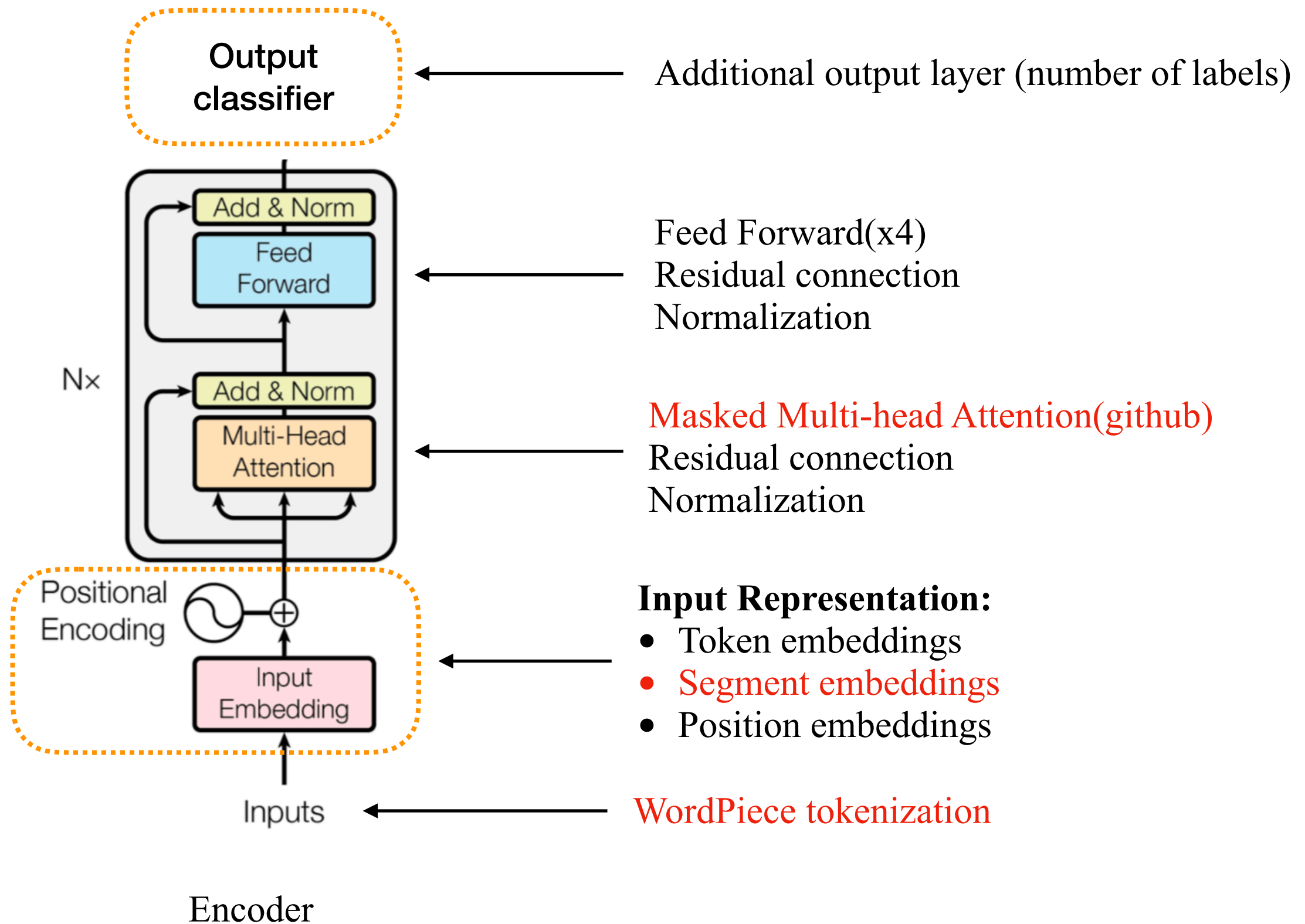
Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language https://arxiv.org/abs/1810.04805	87.433	93.160
2 Oct 05, 2018	BERT (single model) Google AI Language https://arxiv.org/abs/1810.04805	85.083	91.835

Transformer(1/2)



Transformer(2/2)



Input Representation(1/2)

1. Text normalization

John Johanson's, —————> john johanson's,

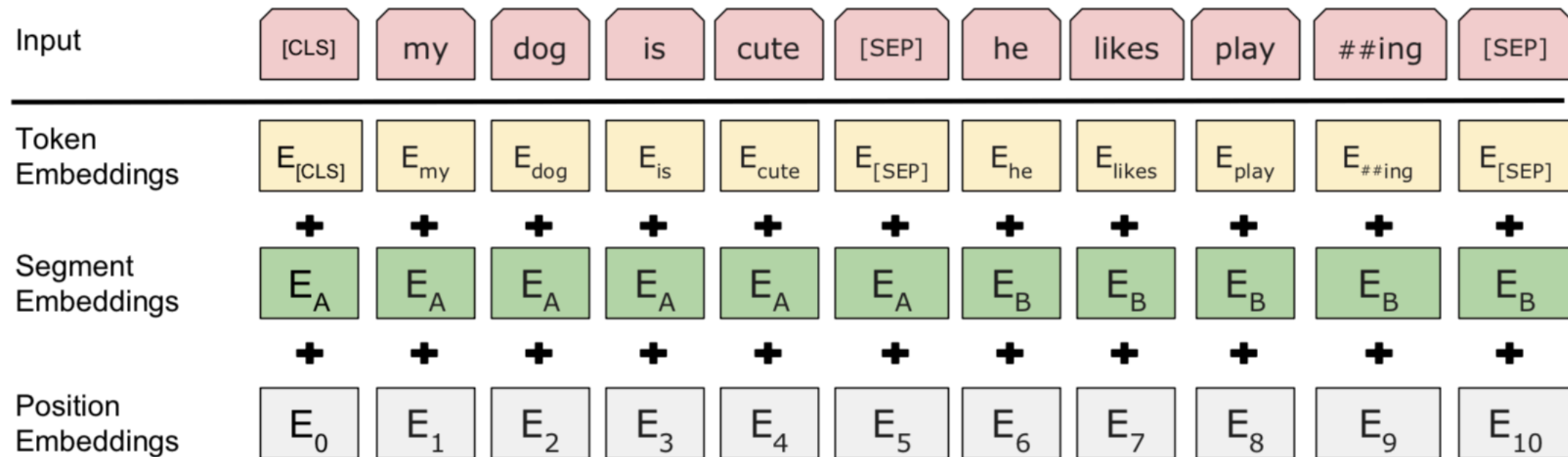
2. Punctuation splitting

john johanson's, —————> john johanson ' s ,

3. WordPiece tokenization

john johanson ' s , —————> john johan ####son ' s ,

Input Representation(2/2)



[CLS]: Output of transformer corresponding to this token is used as the classification task.
[SEP]: Separate sentences.

- **Token Embeddings**: word embeddings (token unit)
- **Segment Embeddings**: A learned sentence A embedding to first sentence and a sentence B embedding to second sentence
- **Position Embeddings**: positional encoding

Pre-train(1/3)

Task#1: Mask LM (Language Model)

Mask **15%** of all WordPiece tokens in each sentence at random, predict the masked words.

Three strategy for masked token:

- 80% of time: Replace the word with the **[Mask]** token.

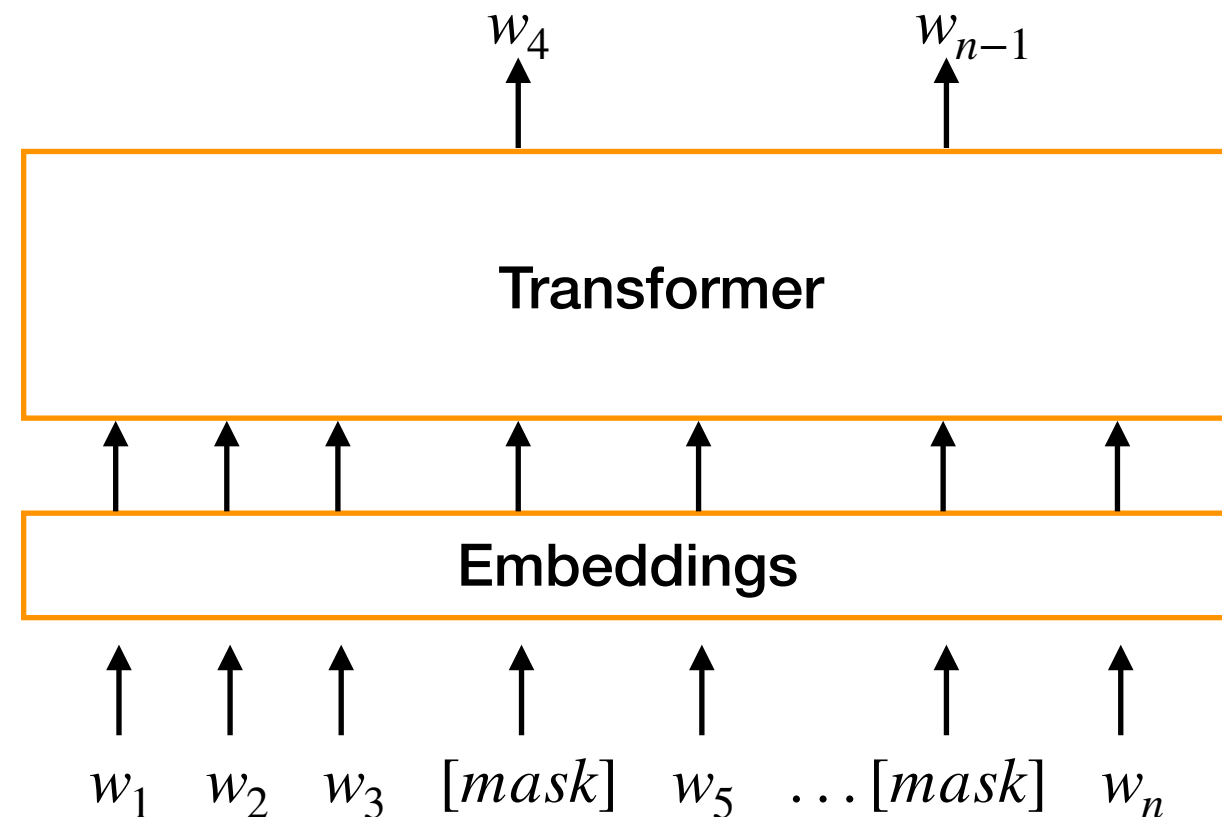
my dog is **hairy** \rightarrow my dog is **[Mask]**

- 10% of time: Replace the word with a random word.

my dog is **hairy** \rightarrow my dog is **apple**

- 10% of time: Keep the word un-changed.

my dog is **hairy** \rightarrow my dog is **hairy**



Pre-train(2/3)

Task#2: NSP (Next Sentence Prediction)

Understanding the relationship between two text sentences.

Pre-train a binarized next sentence prediction task:

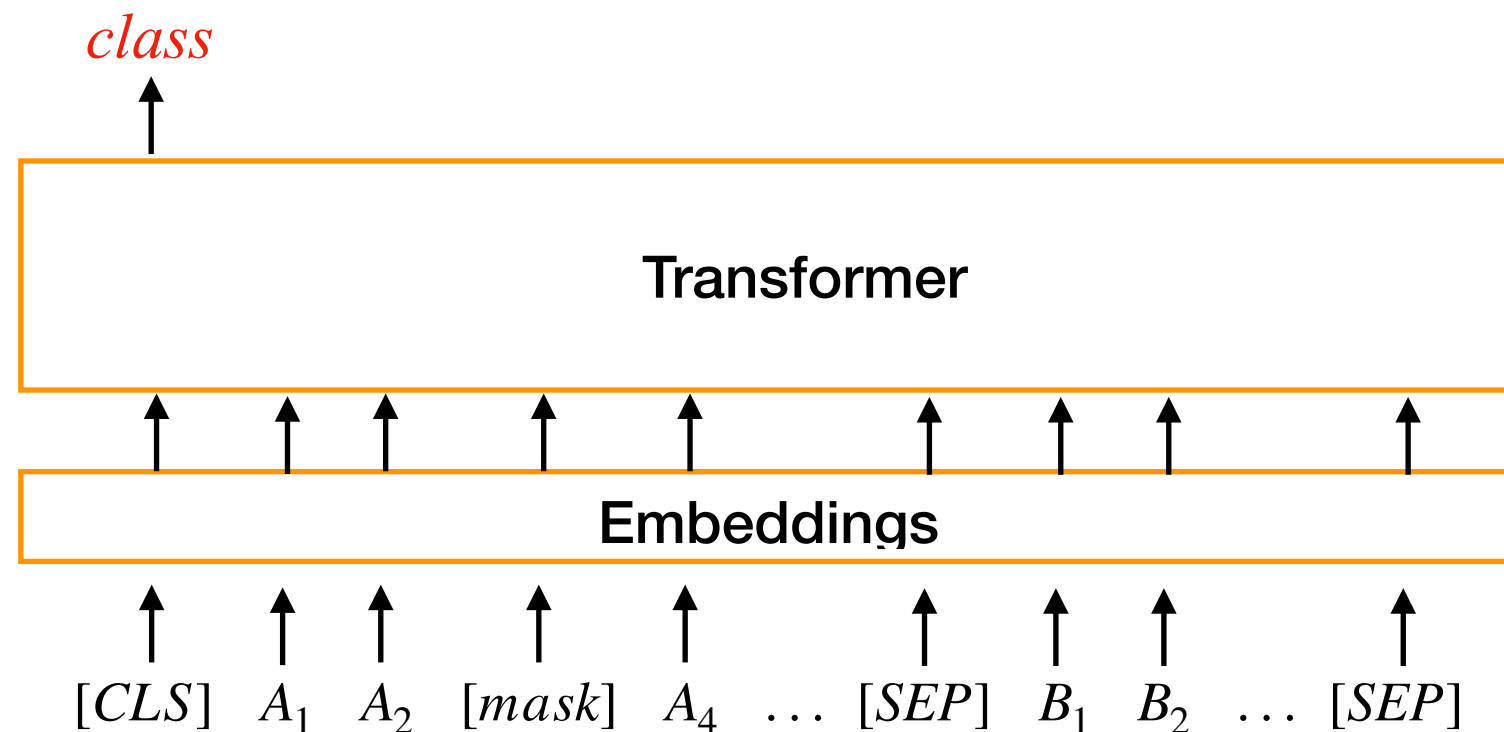
- 50% of time: Actual next sentence.
- 50% of time: Random sentence.

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

Label: IsNext

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

Label: NotNext



Pre-train(3/3)

Pre-training procedure

Data:

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

Two sentences length ≤ 512 tokens

Batch size: 256

Steps: 1,000,000

Optimizer: AdamW, $\beta_1=0.9$, $\beta_2=0.999$, L2 weight decay: 0.01

Learning rate: $1e-4$

Warmup step: 10,000

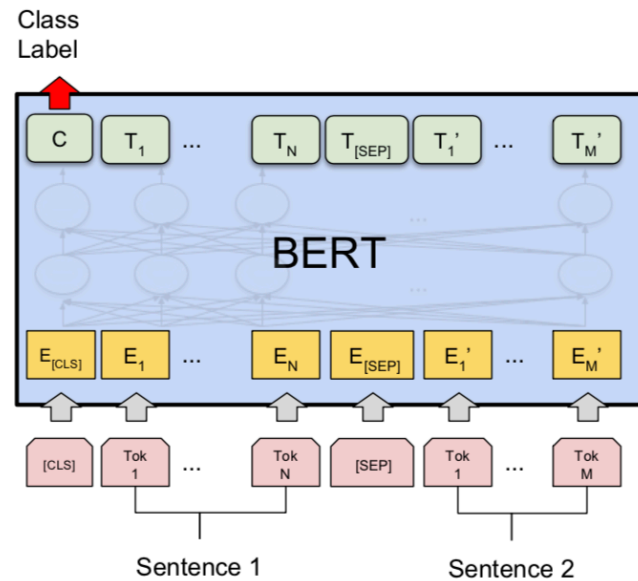
Activation: gelu

BERT_{BASE} : 4 Cloud TPUs in Pod configuration (16 TPU chips total) and pre-train 4 days.

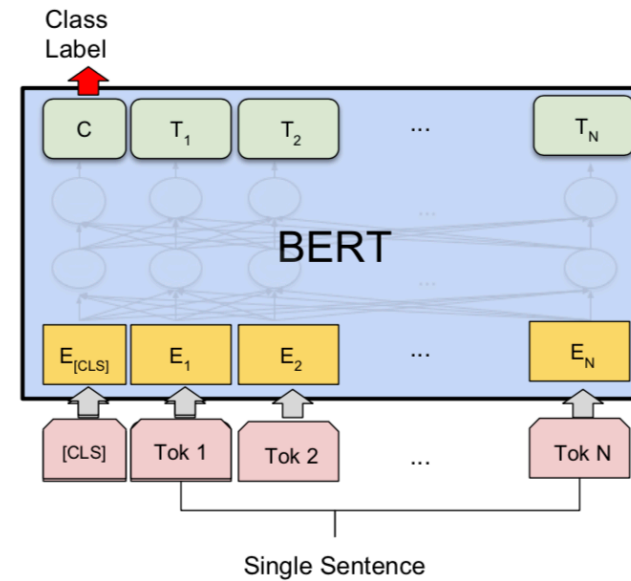
BERT_{LARGE} : 16 Cloud TPUs in Pod configuration (64 TPU chips total) and pre-train 4 days.

Fine-tune(1/9)

9 token-level tasks on GLUE (General Language Understanding Evaluation)

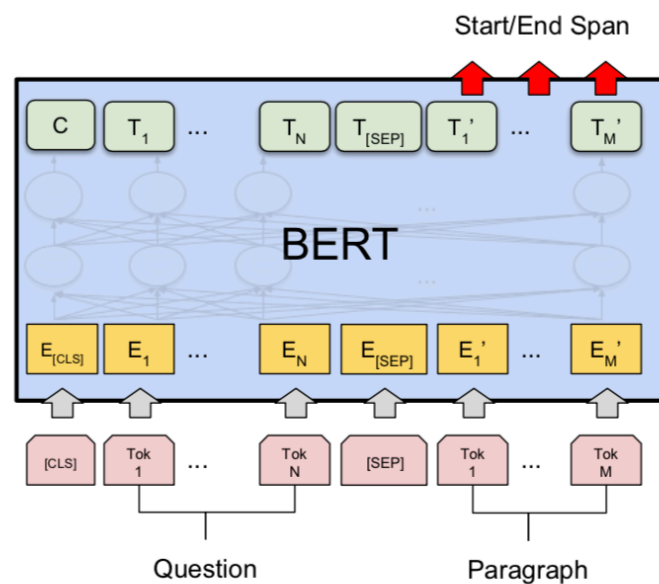


(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

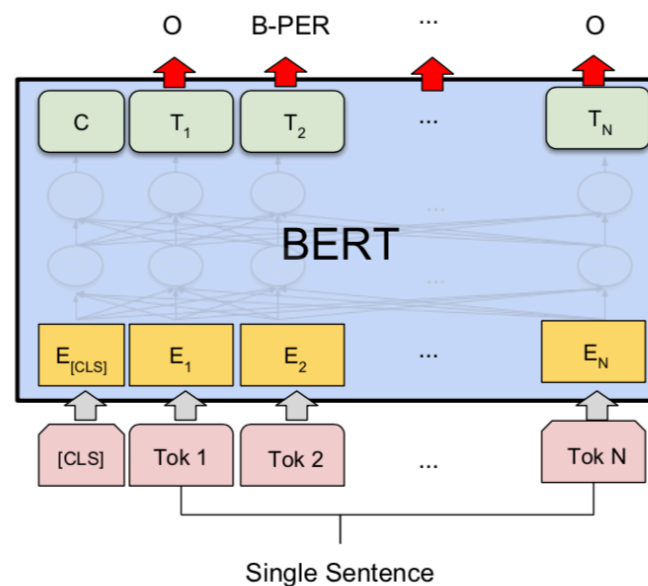


(b) Single Sentence Classification Tasks:
SST-2, CoLA

2 sentence-level tasks



(c) Question Answering Tasks:
SQuAD v1.1



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

Fine-tune(2/9)

Token-level task

- **MNLI**(Multi-Genre Natural Language Inference)

Label:

1. entailment
2. contradiction
3. neutral

- **QQP**(Quora Question Pairs)

Semantical label:

1. yes
2. no

- **QNLI**(Question Natural Language Inference, QA)

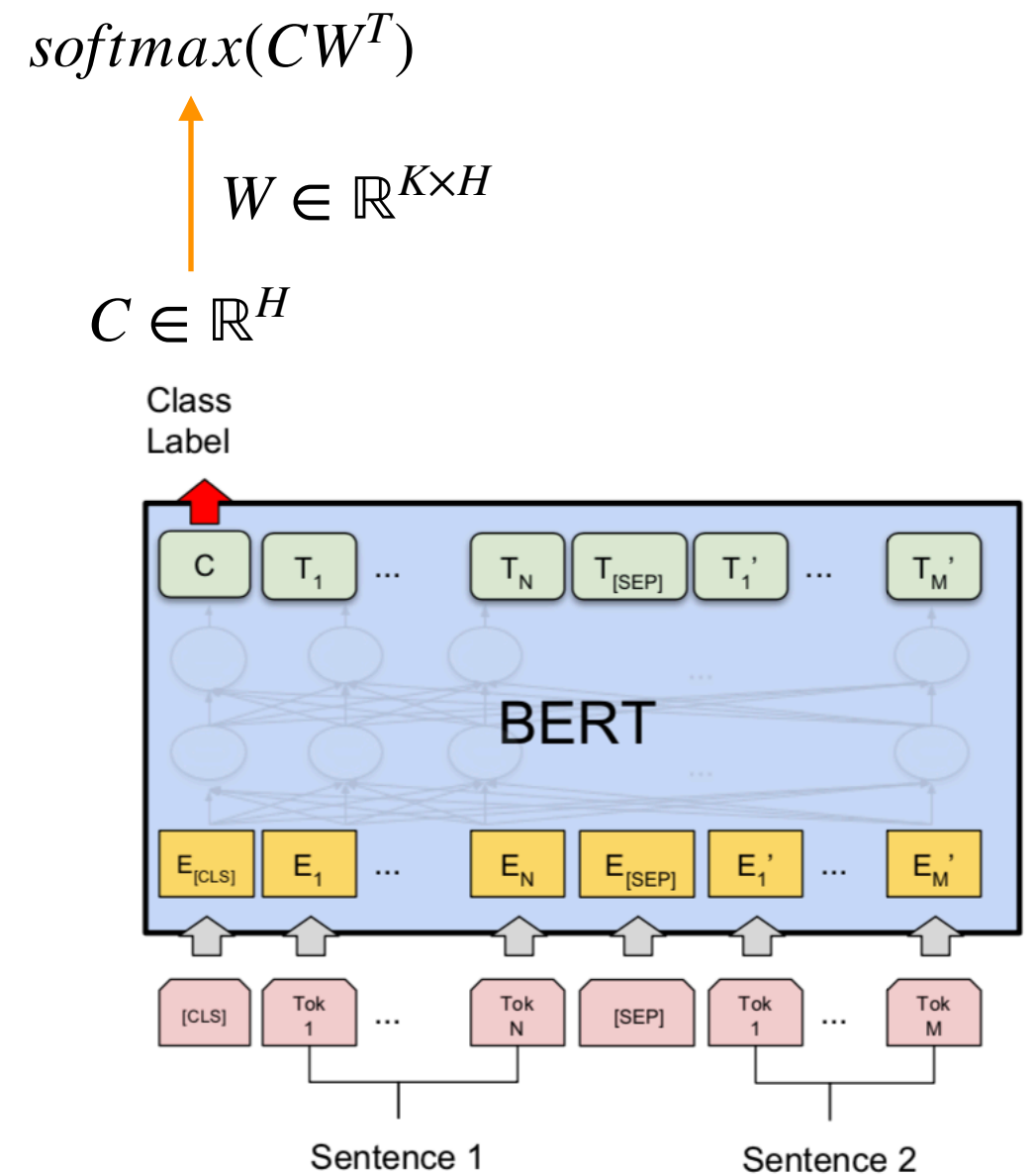
QA label:

1. yes
2. no

- **STS-B**(Semantic Textual Similarity Benchmark)

How similar the two sentence are in terms of semantic meaning?

1. 1~5



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

Fine-tune(3/9)

Token-level task

- **MRPC**(Microsoft Research Paraphrase Corpus)

Semantical label:

1. yes
2. no

- **RTE**(Recognizing Textual Entailment)

Label:

1. entailment
2. contradiction

- **SWAG**(Situations With Adversarial Generations)

Decide among four choices the most plausible sentence.

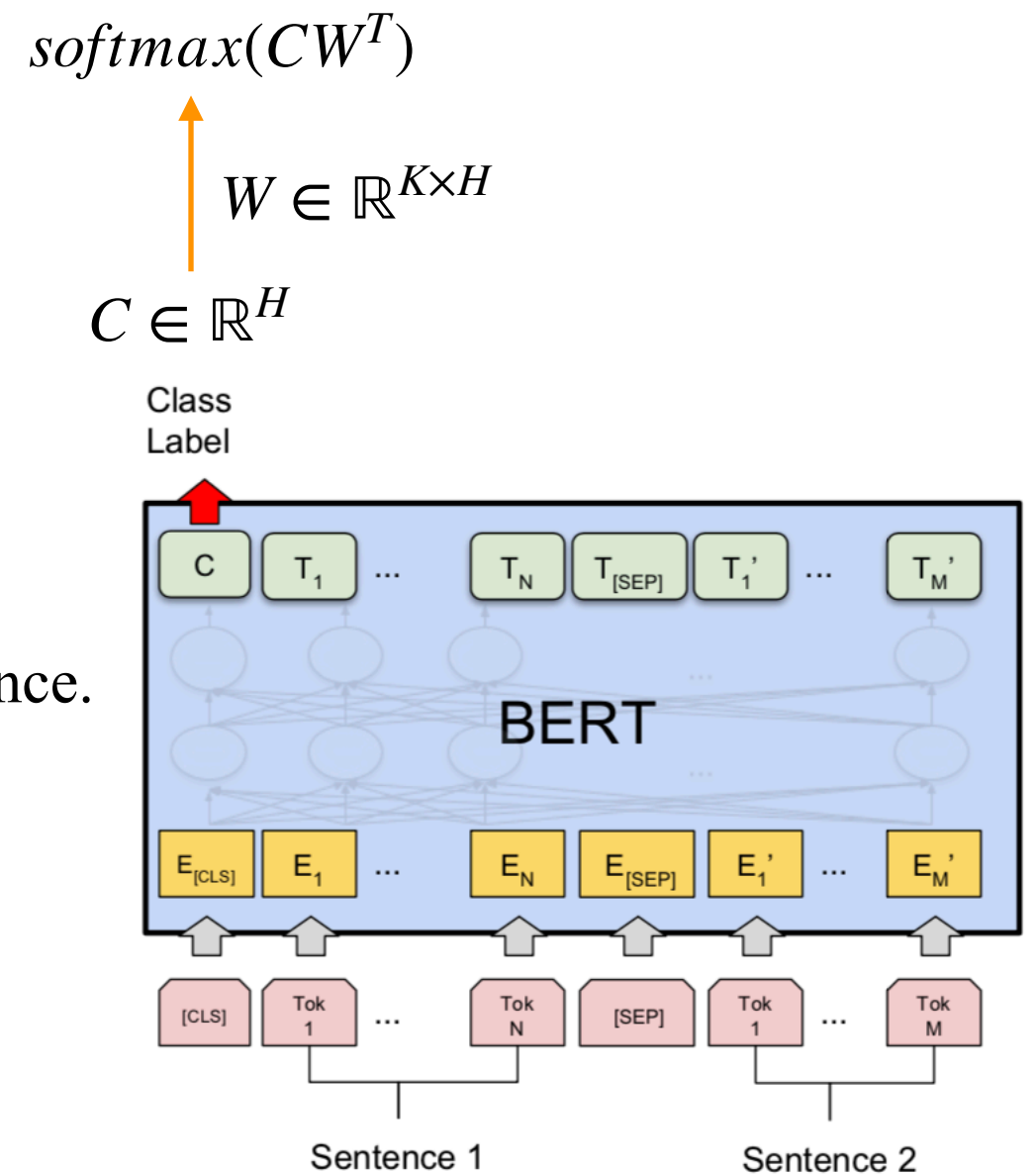
A girl is going across a set of monkey bars. She

(i) jumps up across the monkey bars.

(ii) struggles onto the bars to grab her head.

● (iii) gets to the end and stands on a wooden plank.

(iv) jumps up and does a back flip.



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

Fine-tune(4/9)

Token-level task

- **SST-2**(Stanford Sentiment Treebank)

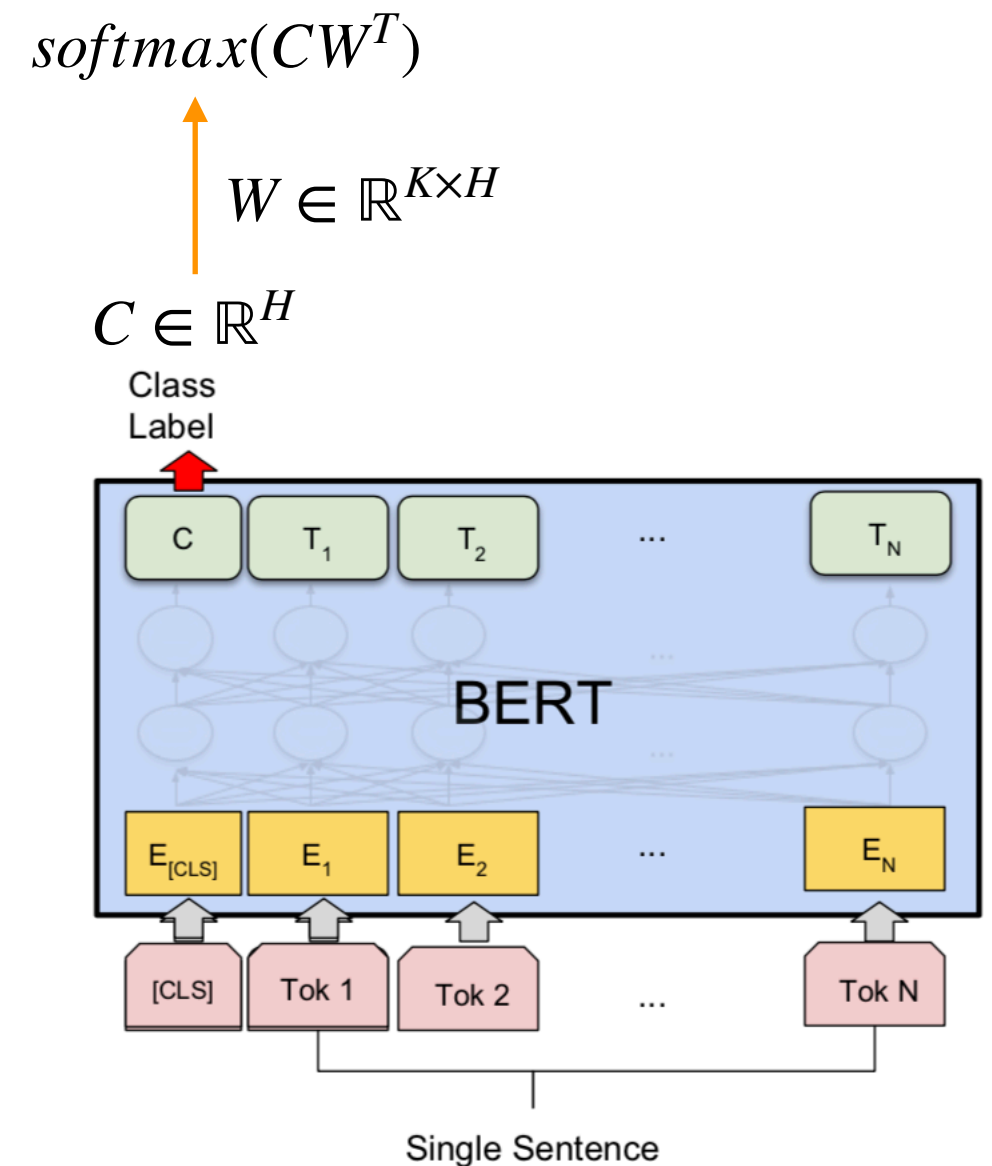
Movie sentiment label:

1. positive
2. negative

- **CoLA**(Corpus of Linguistic Acceptability)

Whether an English sentence is linguistically acceptable:

1. yes
2. no



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

Fine-tune(5/9)

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	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT _{BASE}	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	91.1	94.9	60.5	86.5	89.3	70.1	81.9

OpenAI GPT: L=12, H=768, A=12

BERT_{BASE} : L=12, H=768, A=12

BERT_{LARGE}: L=24, H=1024, A=16

Fine-tune(6/9)

Token-level task

SWAG(Situations With Adversarial Generations)

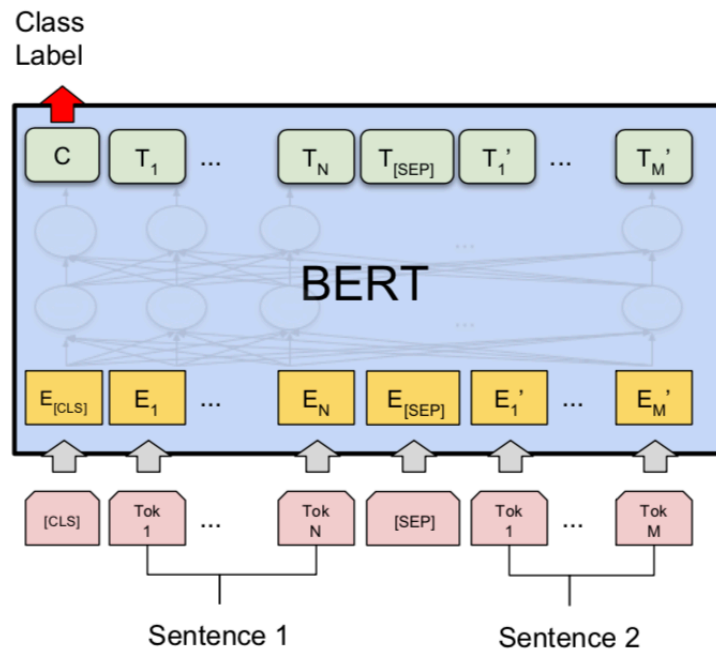
A girl is going across a set of monkey bars. She

- (i) jumps up across the monkey bars.
- (ii) struggles onto the bars to grab her head.
- (iii) gets to the end and stands on a wooden plank.
- (iv) jumps up and does a back flip.

→

Q - A1	N
Q - A2	N
Q - A3	Y
Q - A4	N

$$P_i = \frac{e^{V \cdot C_i}}{\sum_{j=1}^4 e^{V \cdot C_j}} \quad C_i \in \mathbb{R}^H, V \in \mathbb{R}^H$$



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

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
BERT _{BASE}	81.6	-
BERT _{LARGE}	86.6	86.3
Human (expert) [†]	-	85.0
Human (5 annotations) [†]	-	88.0

Fine-tune(7/9)

Sentence-level task

SQuAD v1.1 (Stanford Question Answering Dataset)

- Input Question:

Where do water droplets collide with ice crystals to form precipitation?

- Input Paragraph:

... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...

- Output Answer:

within a cloud

Input Paragraph:

the man went to the store and bought a gallon of milk

Span A: the man went to the

Span B: to the store and bought

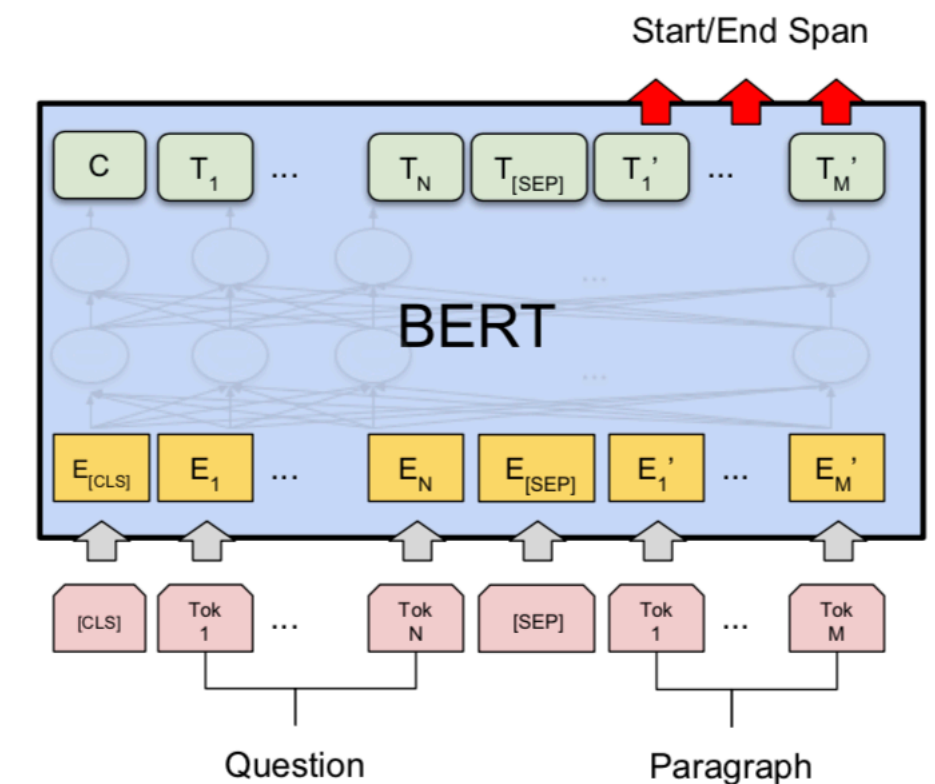
Span C: and bought a gallon of milk

Start:

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \quad S \in \mathbb{R}^H, T_i \in \mathbb{R}^H$$

End:

$$P_i = \frac{e^{E \cdot T_i}}{\sum_j e^{E \cdot T_j}} \quad E \in \mathbb{R}^H, T_i \in \mathbb{R}^H$$



(c) Question Answering Tasks:
SQuAD v1.1

Fine-tune(8/9)

Sentence-level task

SQuAD v1.1 (Stanford Question Answering Dataset)

BERT_{LARGE}(Ensemble)

Use different pre-train checkpoints and fine-tuning seeds.

BERT_{LARGE}(Sgl. + TriviaQA)

Pre-train with SQuAD and TriviaQA.

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT _{BASE} (Single)	80.8	88.5	-	-
BERT _{LARGE} (Single)	84.1	90.9	-	-
BERT _{LARGE} (Ensemble)	85.8	91.8	-	-
BERT _{LARGE} (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERT _{LARGE} (Ens.+TriviaQA)	86.2	92.2	87.4	93.2

Fine-tune(9/9)

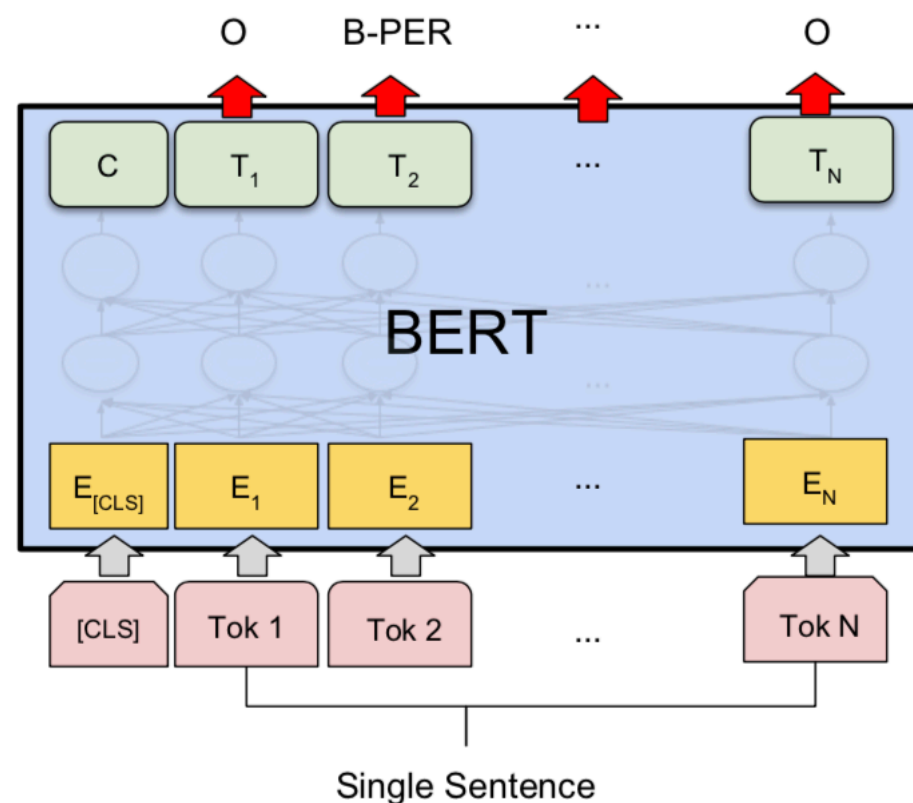
Sentence-level task

NER (Named Entity Recognition)

Sentence words have been annotated as Person, Organization, Location, Miscellaneous or other.

$$T_i \in \mathbb{R}^H \xrightarrow{W \in \mathbb{R}^{K \times H}} P_i \in \mathbb{R}^K$$

where K is NER label set



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



Jim Hen ##son was a puppet ##eer
I-PER I-PER X O O O X

where no prediction is made for X

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
BERT _{BASE}	96.4	92.4
BERT _{LARGE}	96.6	92.8

Evaluation

1. Effect of model size

Fine-tuning task: MRPC (3,600 examples)

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

3. Feature-based Approach with BERT

Fine-tuning task: NER

Layers	Dev F1
Finetune All	96.4
First Layer (Embeddings)	91.0
Second-to-Last Hidden	95.6
Last Hidden	94.9
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1
Sum All 12 Layers	95.5

2. Effect of number of training steps

Fine-tuning task: MNLI

