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

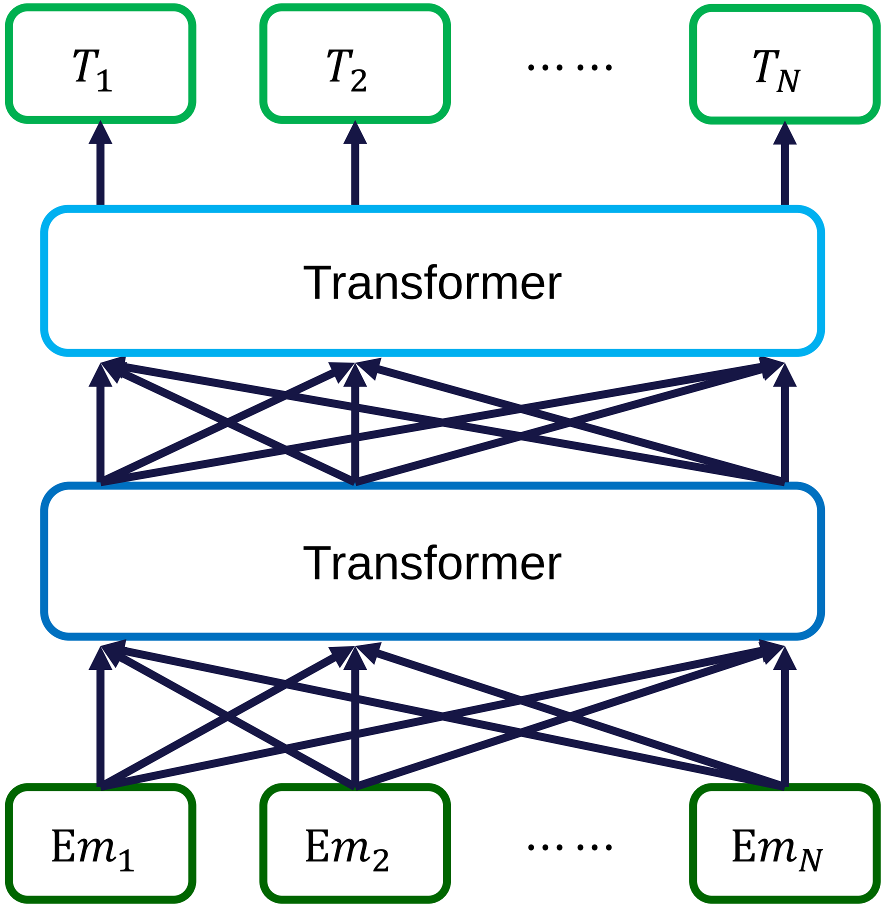
7107053112 沈永平

Abstract

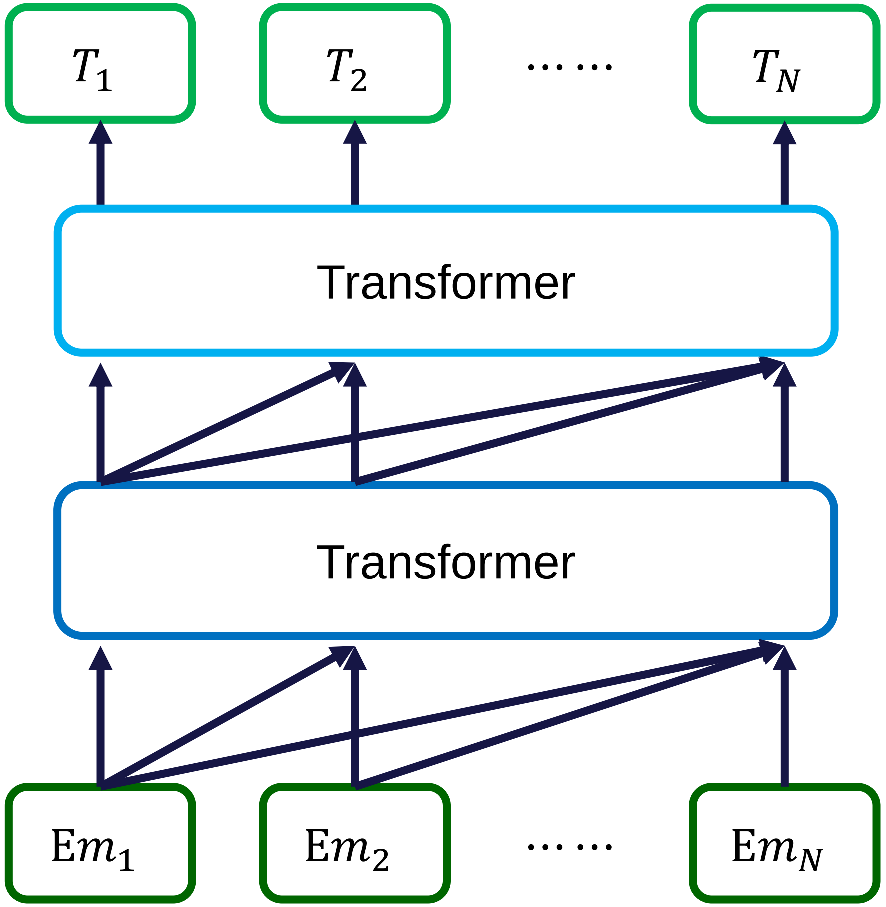
- Purpose : Models can be used for different NLP tasks
  - Single Sentence Classification Tasks
  - Sentence Pair Classification Tasks
  - Question Answering Tasks
  - Single Sentence Tagging Tasks

Model of BERT & OpenAI GPT

BERT

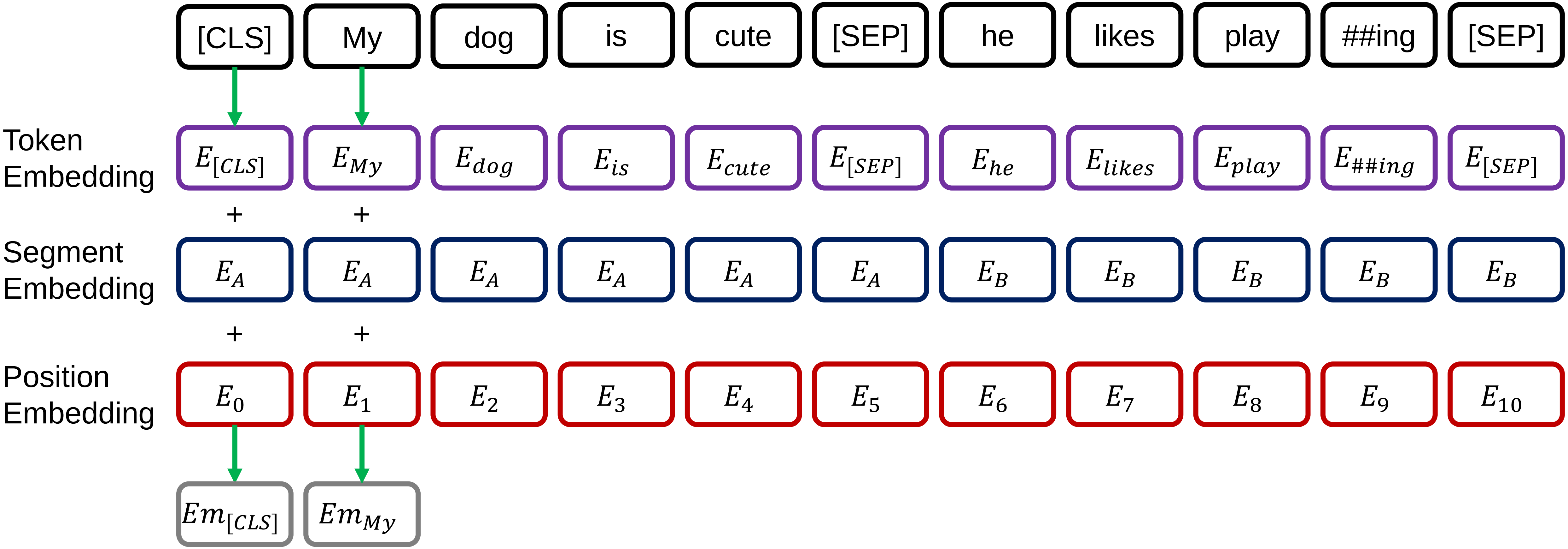


OpenAI GPT



Embedding

Input a sentence : My dog is cute, he likes playing.



	[CLS]	My	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embedding	$E_{[CLS]}$	E_{My}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+									
Segment Embedding	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+									
Position Embedding	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
	$Em_{[CLS]}$	Em_{My}									

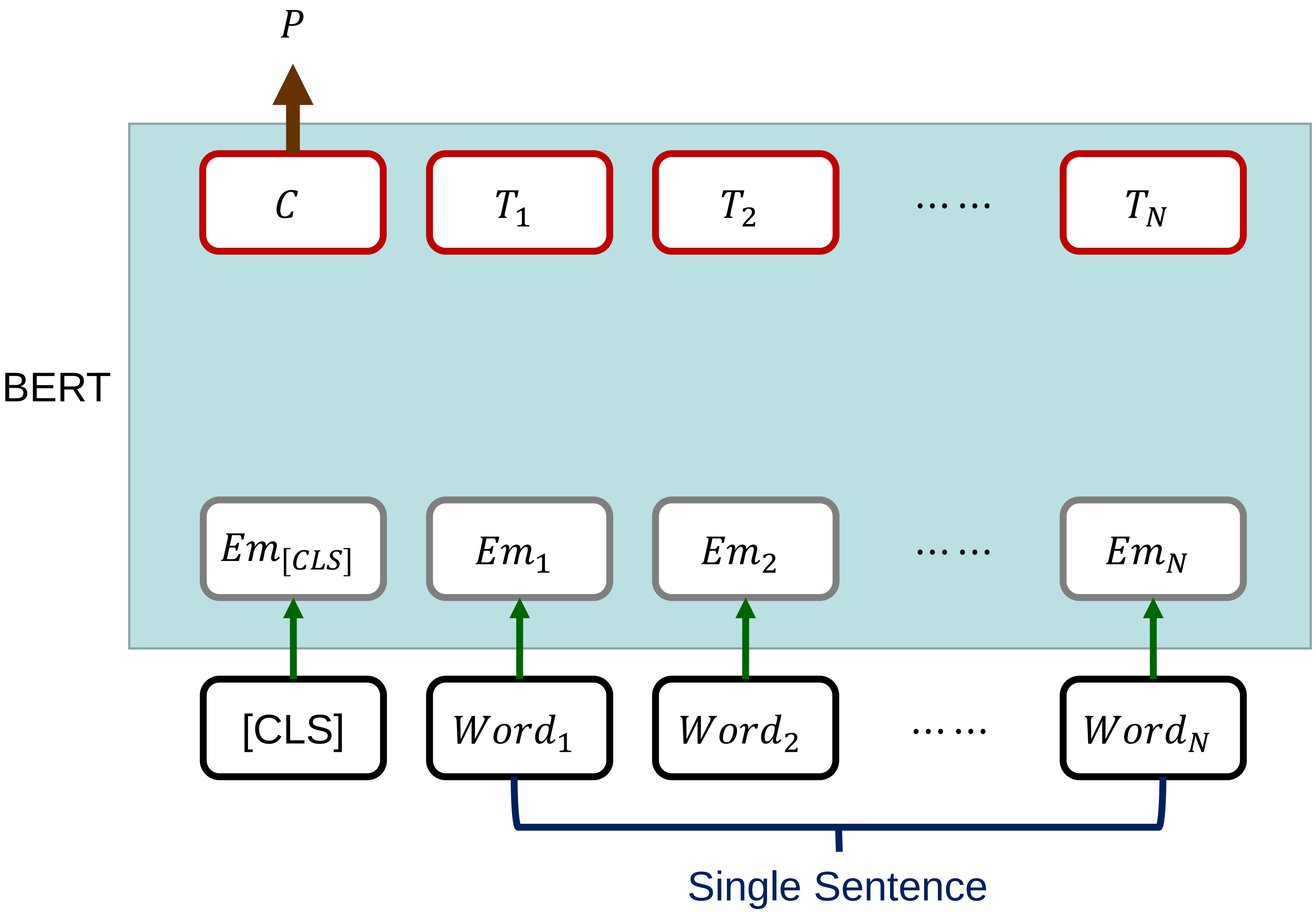
Pre-trained - 1

1. Masked LM : (purpose : predict the masked words)
   A. Chooses 15% of tokens at random in training data and MASK then, but we are not always do that.
   B. 80% of the time: Replace the word with the [MASK] token,
      e.g. My dog is hairy -> My dog is [MASK]
   C. 10% of the time: Replace the word with a random word,
      e.g. My dog is hairy -> My dog is apple
   D. 10% of the time: Keep the word unchanged,
      e.g. My dog is hairy -> My dog is hairy

Pre-trained - 2

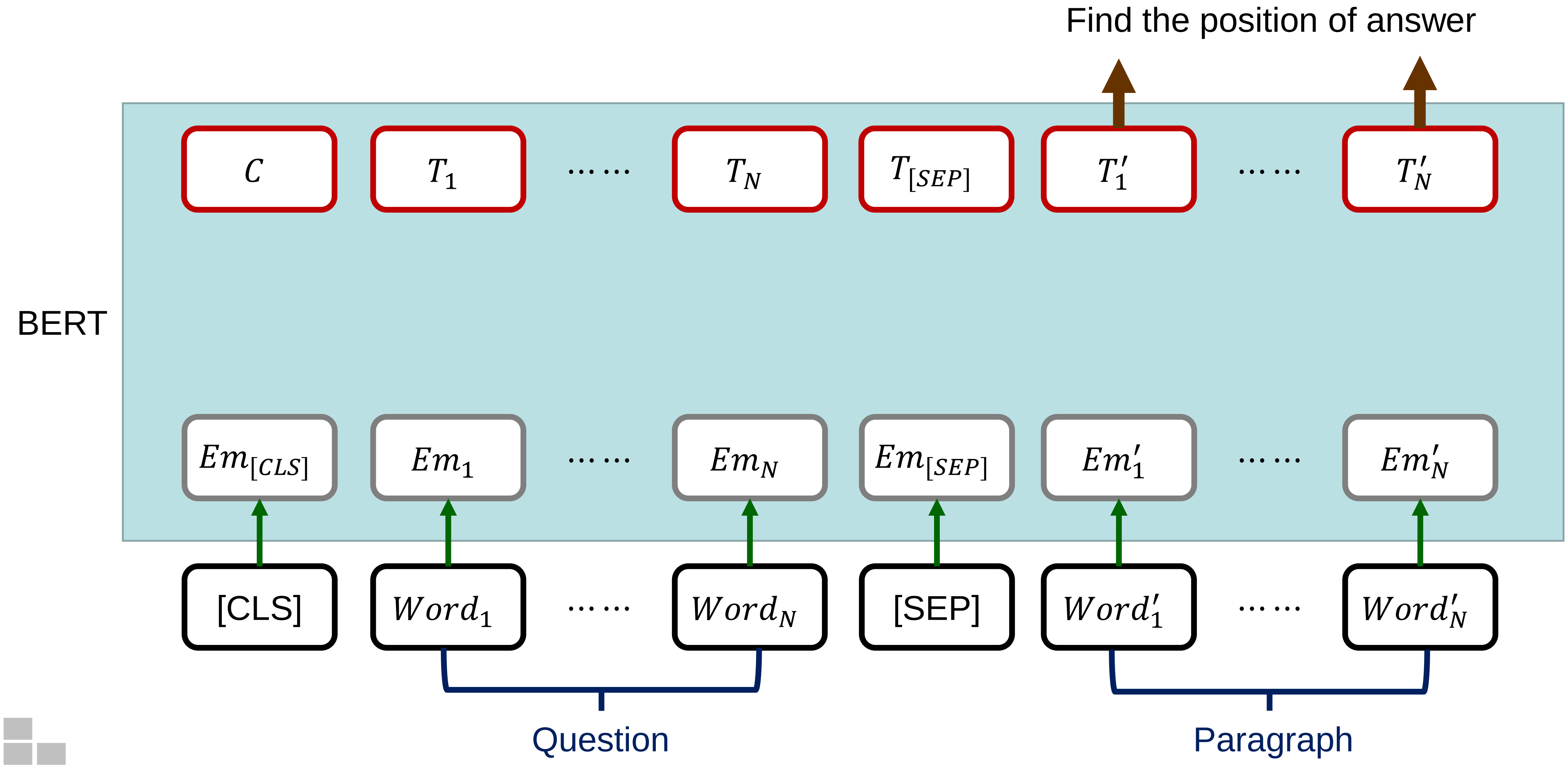
1. Next Sentence Prediction: (purpose : determine if sentence B is the next sentence of sentence A)
   A. choosing the sentences A and B for each pre-training example, 50% of the time B is the actual next sentence that follows A, and 50% of the time it is a random sentence from the corpus.
   B. E.g. :
      1) Input : [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
         label : IsNext
      2) Input : [CLS] the man went to [MASK] store [SEP] penguin [MASK] are flight ##less birds [SEP]
         label : NotNext

Single Sentence Classification Tasks

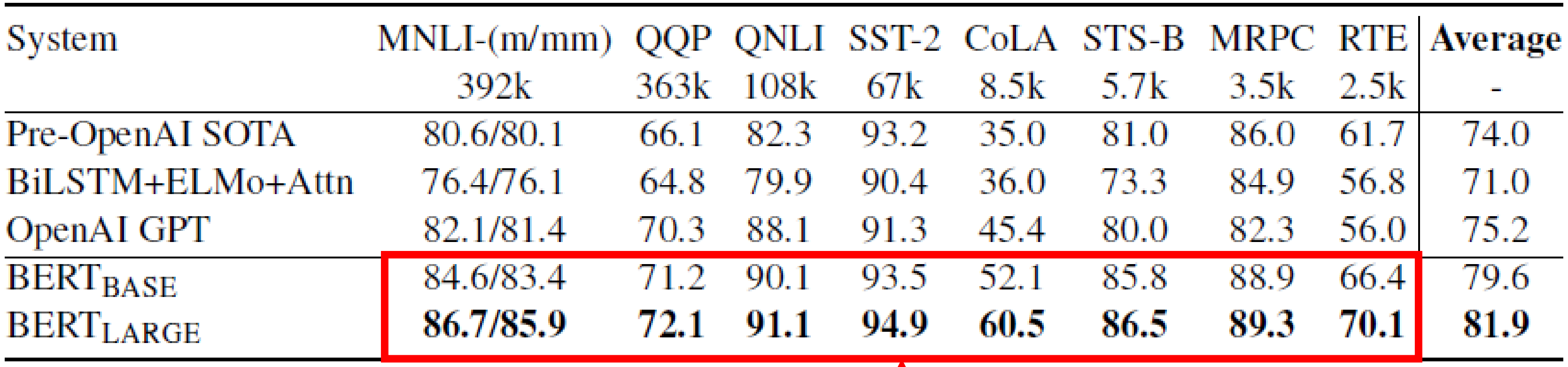


$$p = Softmax(C \cdot W^T)$$

Question Answering Tasks



Experiment result



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	79.9	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$_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT$_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

These results accuracy are trained only 3 epoch