BERT

Bidirectional Encoder Representations from Transformers

https://arxiv.org/pdf/1810.04805.pdf

feature-based & fine-tuning

class	**example**	**Task-specific model**	**method**
Feature-based	ELMo	Need	The representation is provided as feature to tasks
fine-tuning	OpenAI GPT	Don’t need	Fine tune models

Model Architecture

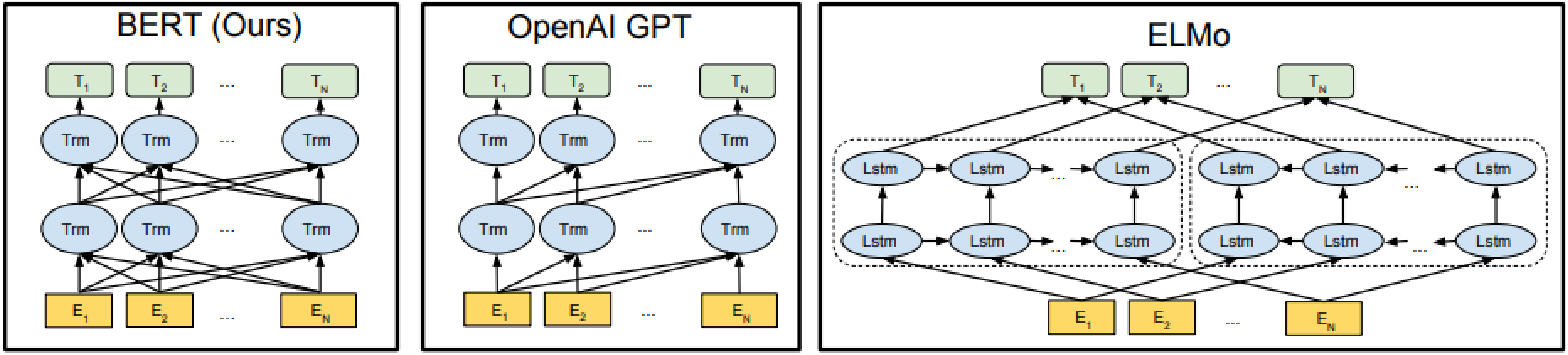


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

Pre-trained models

- **BERT-Base, Uncased:** 12-layer, 768-hidden, 12-heads, 110M parameters
- **BERT-Large, Uncased:** 24-layer, 1024-hidden, 16-heads, 340M parameters
- **BERT-Base, Cased:** 12-layer, 768-hidden, 12-heads , 110M parameters
- **BERT-Large, Cased:** 24-layer, 1024-hidden, 16-heads, 340M parameters (Not available yet. Needs to be re-generated).
- **BERT-Base, Multilingual:** 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- **BERT-Base, Chinese:** Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

Input representation

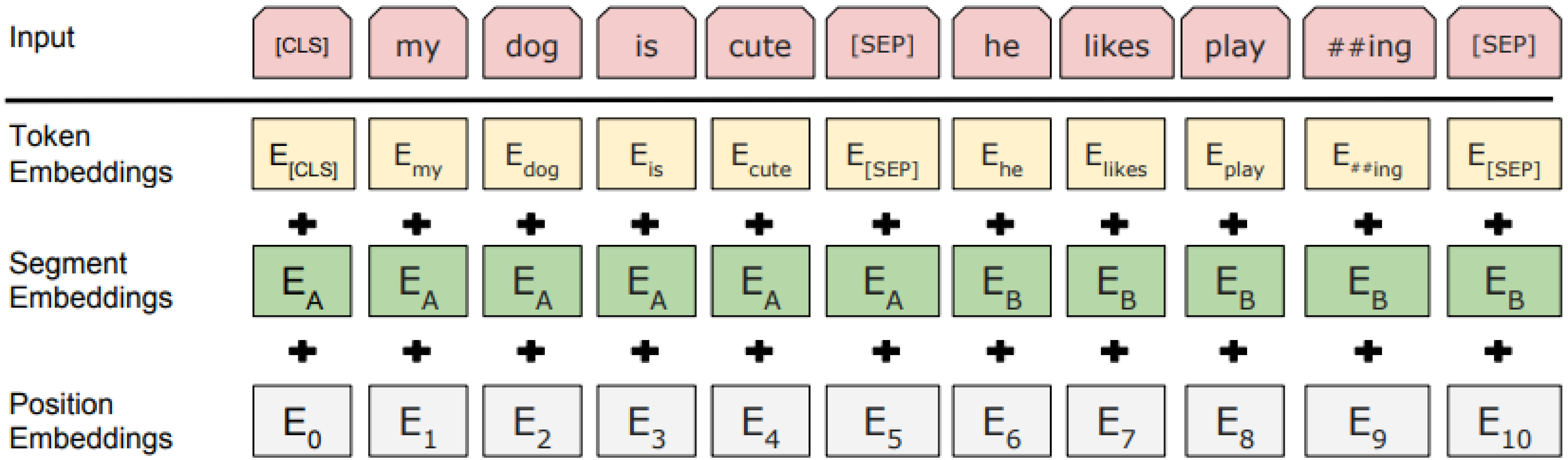
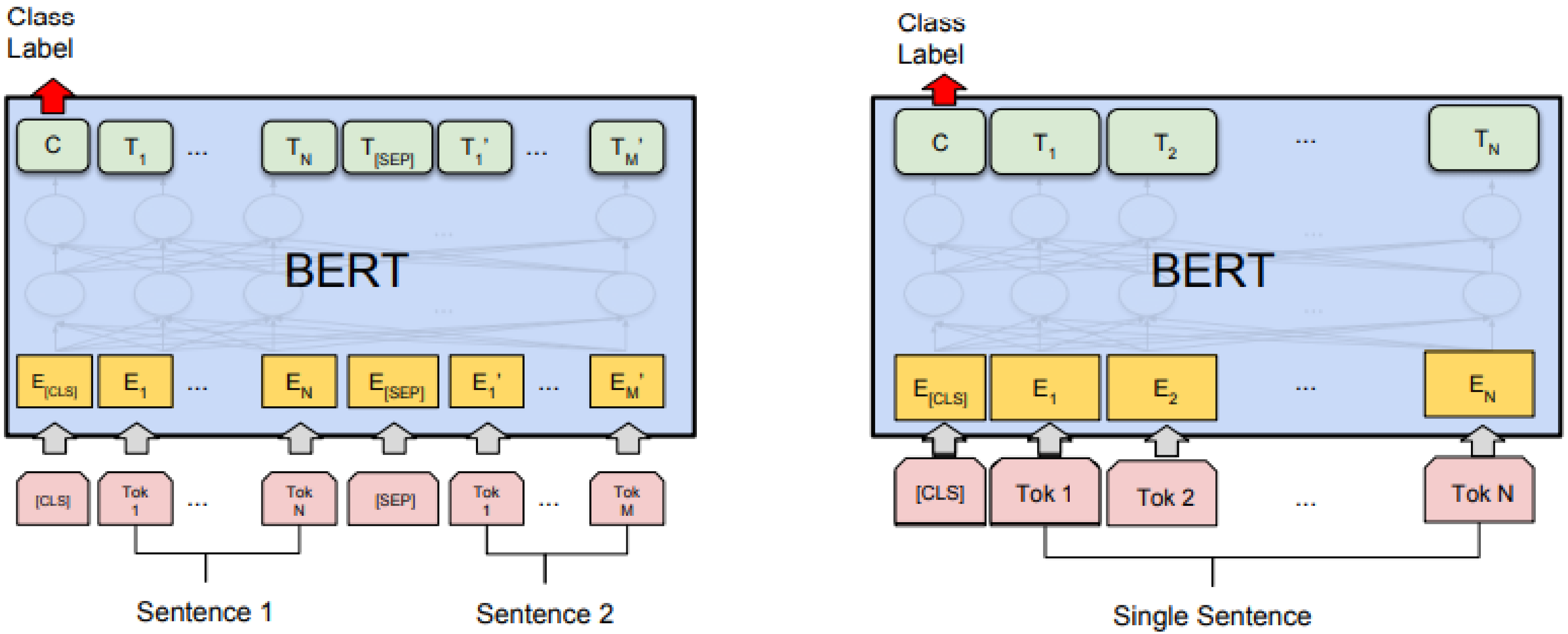


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Pre-training tasks

- **Task #1: Masked LM**
  - 80% of the time: Replace the word with the [MASK] token,
    - e.g., my dog is hairy → my dog is [MASK]
  - 10% of the time: Replace the word with a random word
    - e.g., my dog is hairy → my dog is apple
  - 10% of the time: Keep the word unchanged,
    - e.g., my dog is hairy → my dog is hairy.
- **Task #2: Next Sentence Prediction**
  - Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
    Label = IsNext
  - Input = [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]
    Label = NotNext

Fine-tuning Procedure



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

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

Task — datasets

名称	全名	用途
MNLI	Multi-Genre NLI	蕴含关系推断
QQP	Quora Question Pairs	问题对是否等价
QNLI	Question NLI	句子是否回答问句
SST-2	Stanford Sentiment Treebank	情感分析
CoLA	Corpus of Linguistic Acceptability	句子语言性判断
STS-B	Semantic Textual Similarity	语义相似
MRPC	Microsoft Research Paraphrase Corpus	句子对是否语义等价
RTE	Recognizing Texual Entailment	蕴含关系推断
WNLI	Winograd NLI	蕴含关系推断

Task — MRPC(sentence-pair)

Quality	#1 ID	#2 ID	#1 String	#2 String
1	702876	702977	Amrozi accused his brother , whom he called " the witness " , of deliberately distorting his evidence .	Referrin
0	2108705	2108831	Yucaipa owned Dominick 's before selling the chain to Safeway in 1998 for $ 2.5 billion .	Yucaipa bought Domini
1	1330381	1330521	They had published an advertisement on the Internet on June 10 , offering the cargo for sale , he added .	On Ju
0	3344667	3344648	Around 0335 GMT , Tab shares were up 19 cents , or 4.4 % , at A $ 4.56 , having earlier set a record high of A $	
1	1236820	1236712	The stock rose $ 2.11 , or about 11 percent , to close Friday at $ 21.51 on the New York Stock Exchange .	PG &
1	738533	737951	Revenue in the first quarter of the year dropped 15 percent from the same period a year earlier .	With the sca
0	264589	264502	The Nasdaq had a weekly gain of 17.27 , or 1.2 percent , closing at 1,520.15 on Friday .	The tech-laced Nasdaq
1	579975	579810	The DVD-CCA then appealed to the state Supreme Court .	The DVD CCA appealed that decision to the U.S. Supreme Co
0	3114205	3114194	That compared with $ 35.18 million , or 24 cents per share , in the year-ago period .	Earnings were affected by
0	222621	222514	Shares of Genentech , a much larger company with several products on the market , rose more than 2 percent .	S
0	3131772	3131625	Legislation making it harder for consumers to erase their debts in bankruptcy court won overwhelming House approv	
0	58747	58516	The Nasdaq composite index increased 10.73 , or 0.7 percent , to 1,514.77 .	The Nasdaq Composite index , full of

Training: loading train.tsv——>tokenization——>print 5 train examples——>loading model——>create model_fn(fn+softmax） ——>training & save checkpoints

evaluate: loading dev.tsv——>tokenization——>print 5 dev examples——>loading model——>create model_fn(fn+softmax） ——>loading checkpoints & evaluation

Task — MRPC(sentence-pair)

```
INFO:tensorflow:*** Example ***
INFO:tensorflow:guid: train-1
INFO:tensorflow:tokens: [CLS] am ##ro ##zi accused his brother , whom he called " the witness " , of deliberately di ##stor ##
ting his evidence . [SEP] referring to him as only " the witness " , am ##ro ##zi accused his brother of deliberately di ##sto
r ##ting his evidence . [SEP]
INFO:tensorflow:input_ids: 101 2572 3217 5831 5496 2010 2567 1010 3183 2002 2170 1000 1996 7409 1000 1010 1997 9969 4487 23809
 3436 2010 3350 1012 102 7727 2000 2032 2004 2069 1000 1996 7409 1000 1010 2572 3217 5831 5496 2010 2567 1997 9969 4487 23809
3436 2010 3350 1012 102 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
INFO:tensorflow:input_mask: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
INFO:tensorflow:segment_ids: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
INFO:tensorflow:label: 1 (id = 1)
INFO:tensorflow:*** Example ***
INFO:tensorflow:guid: train-2
INFO:tensorflow:tokens: [CLS] yu ##ca ##ip ##a owned dominic ##k ' s before selling the chain to safe ##way in 1998 for $ 2 .
```

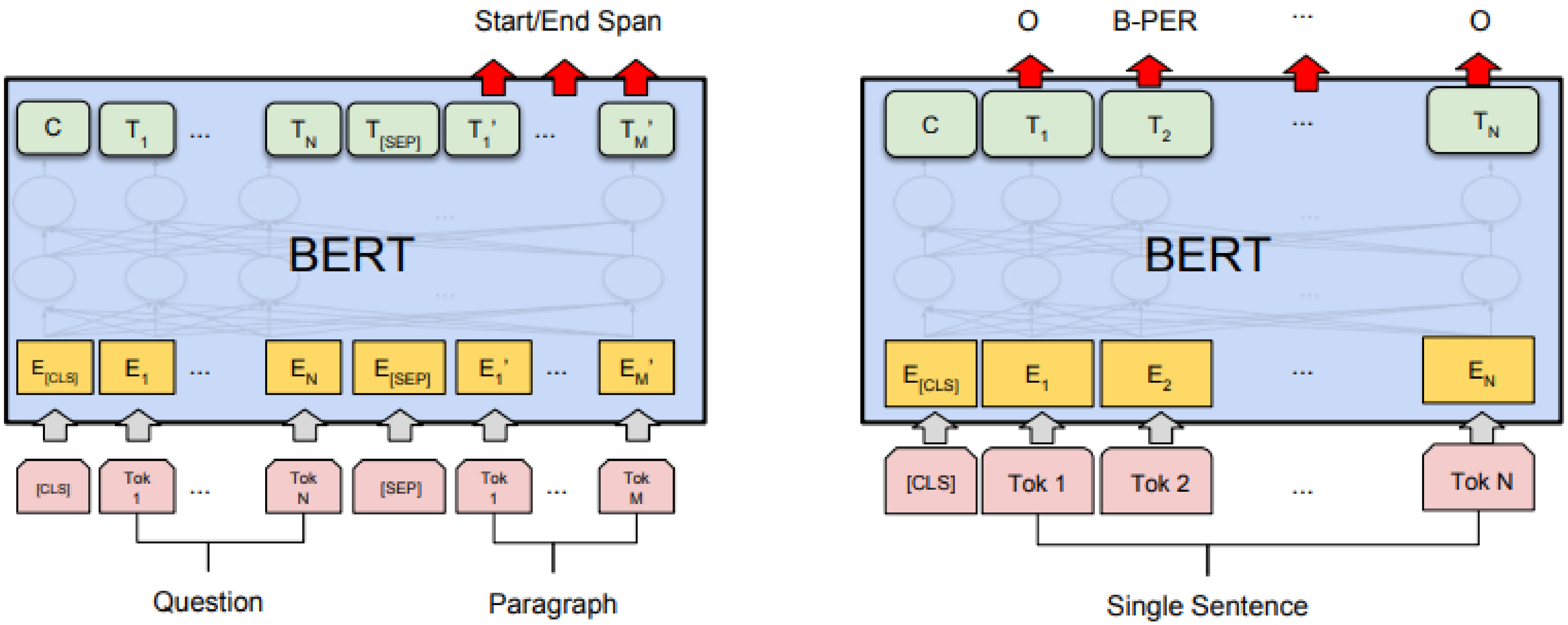
Task — MRPC(sentence-pair)

Steps =len(train_examples)/train_batch_size*num_train_epochs

```
INFO:tensorflow:***** Running training *****
INFO:tensorflow:  Num examples = 3668
INFO:tensorflow:  Batch size = 32
INFO:tensorflow:  Num steps = 343
INFO:tensorflow:Calling model_fn.
INFO:tensorflow:Running train on CPU
INFO:tensorflow:*** Features ***
INFO:tensorflow:  name = input_ids, shape = (32, 128)
INFO:tensorflow:  name = input_mask, shape = (32, 128)
INFO:tensorflow:  name = label_ids, shape = (32,)
INFO:tensorflow:  name = segment_ids, shape = (32, 128)
INFO:tensorflow:**** Trainable Variables ****
INFO:tensorflow:  name = bert/embeddings/word_embeddings:0, shape = (30522, 768), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/embeddings/token_type_embeddings:0, shape = (2, 768), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/embeddings/position_embeddings:0, shape = (512, 768), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/embeddings/LayerNorm/beta:0, shape = (768,), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/embeddings/LayerNorm/gamma:0, shape = (768,), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/encoder/layer_0/attention/self/query/kernel:0, shape = (768, 768), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/encoder/layer_0/attention/self/query/bias:0, shape = (768,), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/encoder/layer_0/attention/self/key/kernel:0, shape = (768, 768), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/encoder/layer_0/attention/self/key/bias:0, shape = (768,), *INIT_FROM_CKPT*
INFO:tensorflow:  name = bert/encoder/layer_0/attention/self/value/kernel:0, shape = (768, 768), *INIT_FROM_CKPT*
```

```
INFO:tensorflow:***** Eval results *****
INFO:tensorflow:  eval_accuracy = 0.85784316
INFO:tensorflow:  eval_loss = 0.45819566
INFO:tensorflow:  global_step = 343
INFO:tensorflow:  loss = 0.45819566
```

Fine-tuning Procedure



(c) Question Answering Tasks:
SQuAD v1.1

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

Task — SQuAD(span)

- **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`

Evaluate:
- EM: the results must exact match the answer
- F1: Cut the phrase of the results into words, and calculate the recall, Precision and F1 together with the answer of the person

out-of-memory: change train_batch_size=6

Task — SQuAD(span)

```
INFO:tensorflow:*** Example ***
INFO:tensorflow:unique_id: 1000000009
INFO:tensorflow:example_index: 9
INFO:tensorflow:doc_span_index: 0
INFO:tensorflow:tokens: [CLS] what race has a very low rate of holding public office in brazil ? [SEP] though brazilian ##s of
 at least partial african heritage make up a large percentage of the population , few blacks have been elected as politicians
. the city of salvador , bahia , for instance , is 80 % people of color , but voters have not elected a mayor of color . journ
alists like to say that us cities with black major ##ities , such as detroit and new orleans , have not elected white mayors s
ince after the civil rights movement , when the voting rights act of 1965 protected the franchise for minorities , and blacks
in the south regained the power to vote for the first time since the turn of the 20th century . new orleans elected its first
black mayor in the 1970s . new orleans elected a white mayor after the wide ##sca ##le disruption and damage of hurricane katr
ina in 2005 . [SEP]
INFO:tensorflow:token_to_orig_map: 16:0 17:1 18:1 19:2 20:3 21:4 22:5 23:6 24:7 25:8 26:9 27:10 28:11 29:12 30:13 31:14 32:15
33:15 34:16 35:17 36:18 37:19 38:20 39:21 40:22 41:22 42:23 43:24 44:25 45:26 46:26 47:27 48:27 49:28 50:29 51:29 52:30 53:31
54:31 55:32 56:33 57:34 58:34 59:35 60:36 61:37 62:38 63:39 64:40 65:41 66:42 67:43 68:43 69:44 70:45 71:46 72:47 73:48 74:49
75:50 76:51 77:52 78:53 79:53 80:53 81:54 82:55 83:56 84:57 85:58 86:59 87:59 88:60 89:61 90:62 91:63 92:64 93:65 94:66 95:67
96:68 97:69 98:70 99:70 100:71 101:72 102:73 103:74 104:75 105:76 106:77 107:78 108:79 109:80 110:81 111:82 112:82 113:83 114:
84 115:85 116:86 117:87 118:88 119:89 120:90 121:91 122:92 123:93 124:94 125:95 126:96 127:97 128:98 129:99 130:100 131:101 13
2:102 133:103 134:103 135:104 136:105 137:106 138:107 139:108 140:109 141:110 142:111 143:112 144:113 145:113 146:114 147:115
148:116 149:117 150:118 151:119 152:120 153:121 154:122 155:122 156:122 157:123 158:124 159:125 160:126 161:127 162:128 163:12
9 164:130 165:130
INFO:tensorflow:token_is_max_context: 16:True 17:True 18:True 19:True 20:True 21:True 22:True 23:True 24:True 25:True 26:True
27:True 28:True 29:True 30:True 31:True 32:True 33:True 34:True 35:True 36:True 37:True 38:True 39:True 40:True 41:True 42:Tru
e 43:True 44:True 45:True 46:True 47:True 48:True 49:True 50:True 51:True 52:True 53:True 54:True 55:True 56:True 57:True 58:T
rue 59:True 60:True 61:True 62:True 63:True 64:True 65:True 66:True 67:True 68:True 69:True 70:True 71:True 72:True 73:True 74
:True 75:True 76:True 77:True 78:True 79:True 80:True 81:True 82:True 83:True 84:True 85:True 86:True 87:True 88:True 89:True
90:True 91:True 92:True 93:True 94:True 95:True 96:True 97:True 98:True 99:True 100:True 101:True 102:True 103:True 104:True 1
05:True 106:True 107:True 108:True 109:True 110:True 111:True 112:True 113:True 114:True 115:True 116:True 117:True 118:True 1
19:True 120:True 121:True 122:True 123:True 124:True 125:True 126:True 127:True 128:True 129:True 130:True 131:True 132:True 1
33:True 134:True 135:True 136:True 137:True 138:True 139:True 140:True 141:True 142:True 143:True 144:True 145:True 146:True 1
```

Todo list

- Transformer : Set the hyperparameter and select the words to connect to each node
- pre_training : add another tasks；
- final hidden state： Like ELMo, learning the linear combination of the representations of each layer;
- Fine-tune： Add a deep network model