BERT

Bidirectional Encoder Representations from Transformer

Team 7

March 13, 2020

Outline

1. Transformer
    1.1. Encoder
    1.2. Self-Attention

2. BERT
    2.1. Archtecture & Input
    2.2. 2 stage Training
    2.3. Pre-Training
    2.4. self-supervised learning
    2.5. FineTune
    2.6. Conclusion 2-stage training

3. BERT-Family
    3.1. SpanBERT, Roberta, ALBERT

Transformer

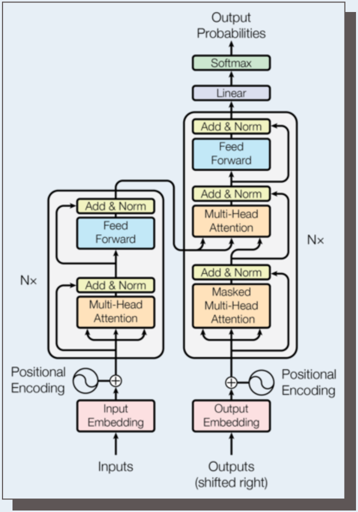


Figure: **Transformer**: Attention is all you need

Overview

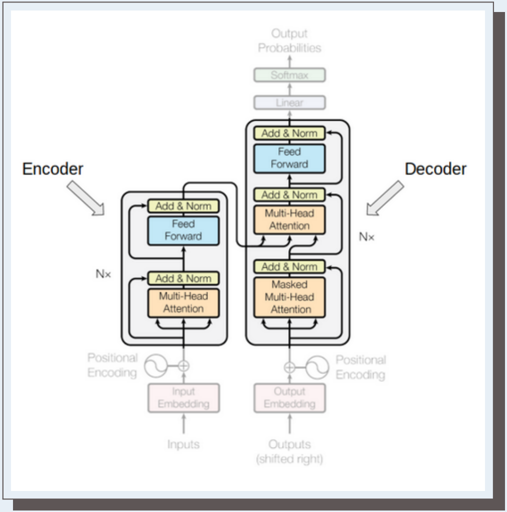


Figure: **Encoder-Decoder**

Encoder

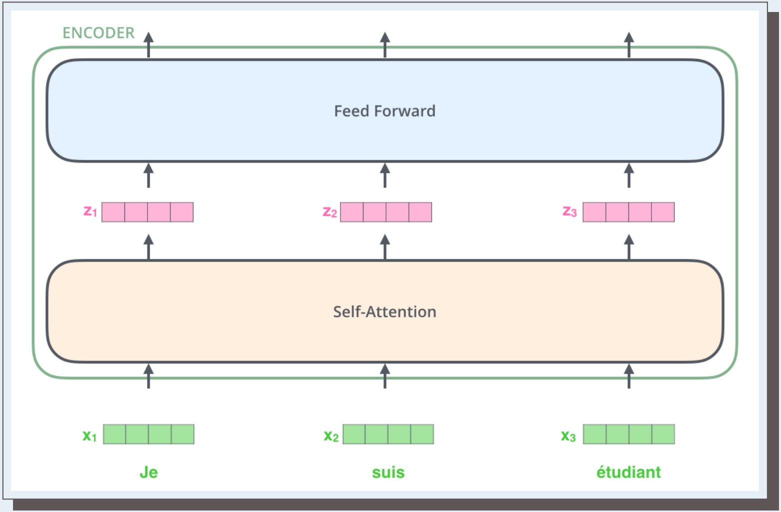


Figure: **1 Encoder-layer**

Self-Attention

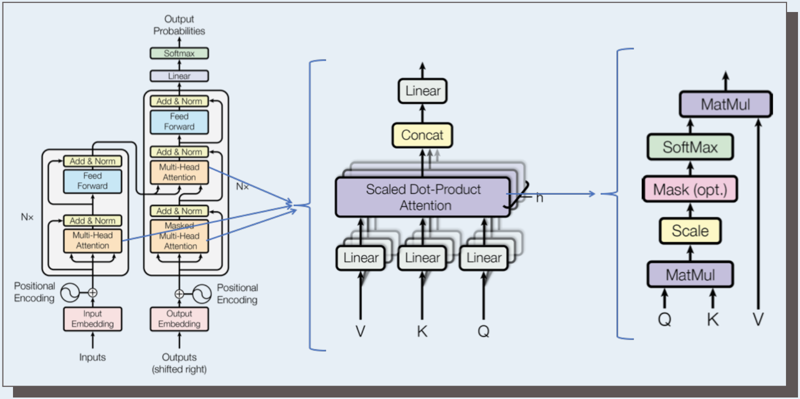


Figure: **Self-Attention**

BERT

Figure: **BERT**

Archtecture

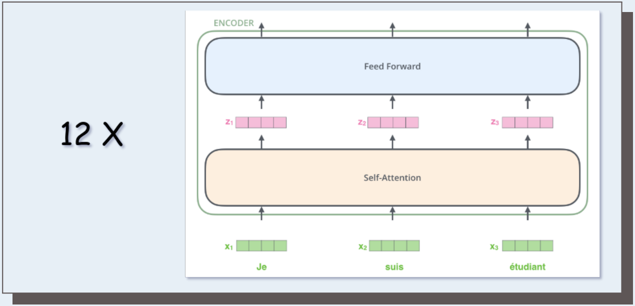
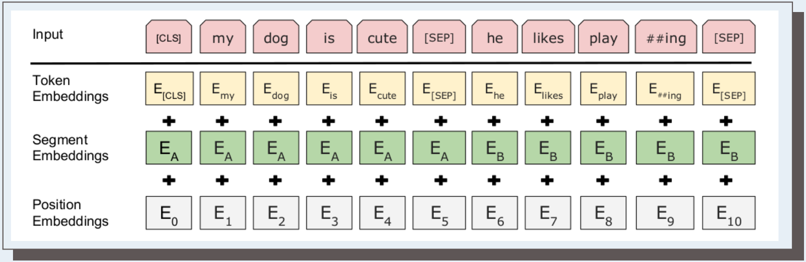


Figure: **BERT-Architecture**

Input



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

Figure: **BERT-Input**

2-stage-training

Pre-training

Self-supervised tasks.

Finetune

3 finetune tasks.

Pre-Training

Pretraining tasks

2 pretraining tasks

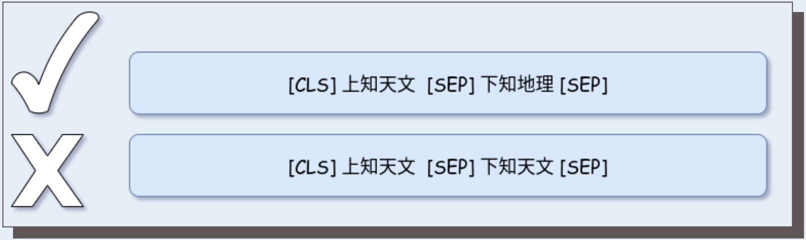
NSP

Next Sentence Prediction

MLM

Masked Language Model - self supervised learning

NSP



✓ [CLS] 上知天文 [SEP] 下知地理 [SEP]

✗ [CLS] 上知天文 [SEP] 下知天文 [SEP]

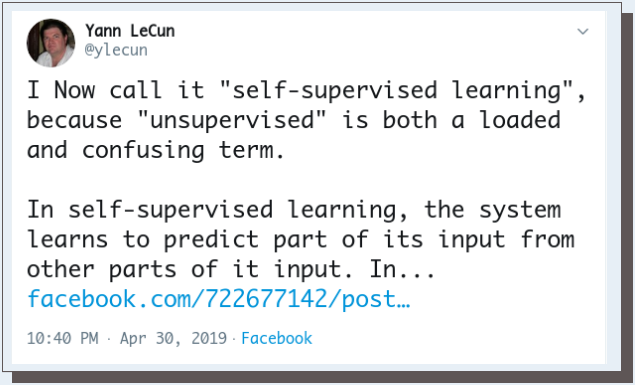
Figure: **Next Sentence Prediction**

CLS Classifier

Figure: **[CLS] token classifier**

SSL

What is self-supervised learning ?



Yann LeCun
@ylecun

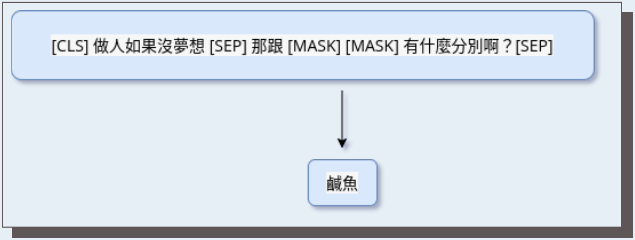
I Now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of it input. In... facebook.com/722677142/post…

10:40 PM · Apr 30, 2019 · Facebook

Figure: **Self-supervised learning definition**

MLM



[CLS] 做人如果沒夢想 [SEP] 那跟 [MASK] [MASK] 有什麼分別啊？[SEP]

鹹魚

Figure: **Masked LM**

MLM Classifier

Figure: **MLM classifier**

Fine-Tune

Downstream tasks:

NER

Name Entity Recognition

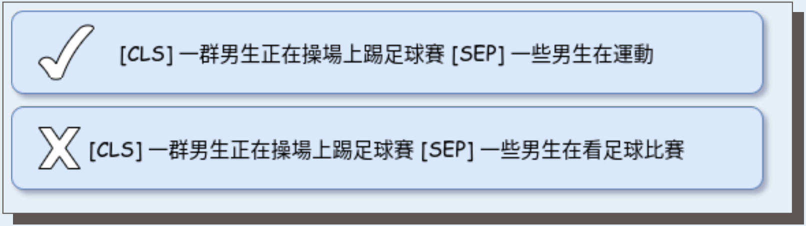
NLI

Natural Language Inference

QA

Question Answering

NLI



✓ [CLS] 一群男生正在操場上踢足球賽 [SEP] 一些男生在運動

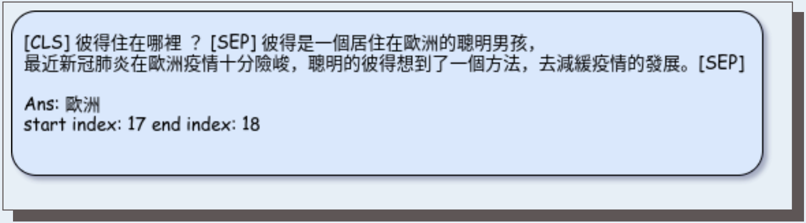
✗ [CLS] 一群男生正在操場上踢足球賽 [SEP] 一些男生在看足球比賽

Figure: **Natural Language Inference**

CLS Classifier

Figure: **[CLS] token classifier**

QA



[CLS] 彼得住在哪裡 ？ [SEP] 彼得是一個居住在歐洲的聰明男孩，
最近新冠肺炎在歐洲疫情十分險峻，聰明的彼得想到了一個方法，去減緩疫情的發展。[SEP]

Ans: 歐洲
start index: 17 end index: 18

Figure: **Question Answering Example**

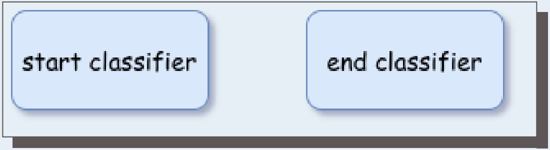


Figure: **Random initialize two classifiers**

Conclusion 2-stage training

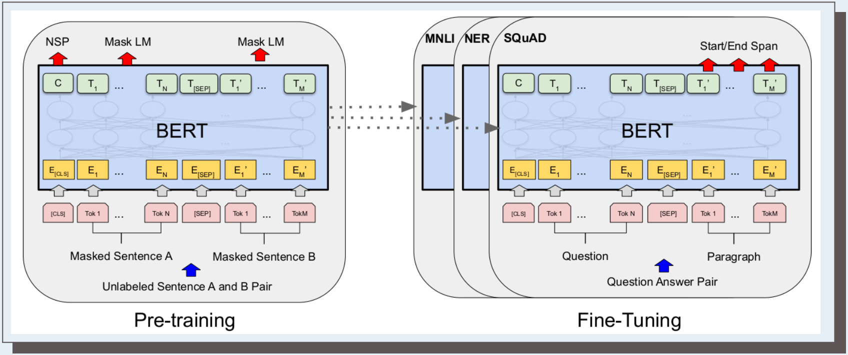


Figure: **2-stage training**

BERT-Family (1)

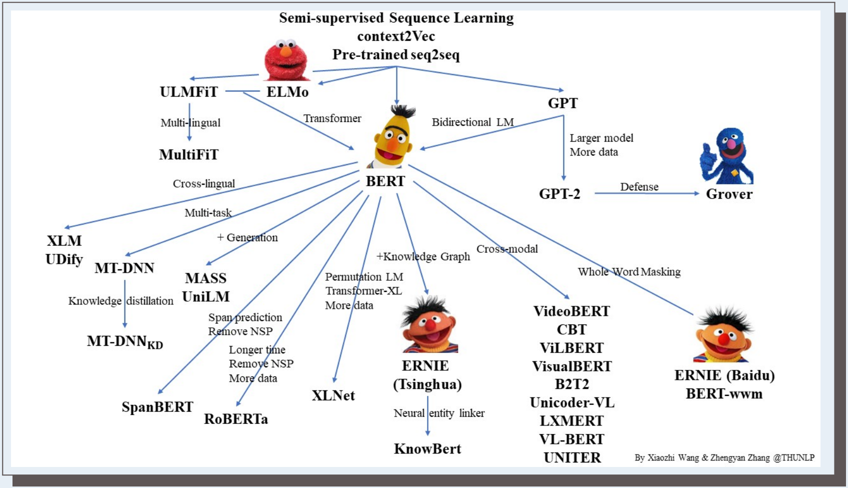


Figure: **BERT Family**

BERT-Family (2)

SpanBERT - EMNLP2019

Improving Pre-training by Representing and Predicting Spans

Roberta - ICLR2020 Reject

A Robustly Optimized BERT Pretraining Approach

ALBERT - ICLR2020 Spotlight

A Lite BERT for Self-supervised Learning of Language Representations

SpanBERT, Roberta, ALBERT

Difference with BERT - SpanBERT

1. change NSP task to SBO (span boundary objective) task.
2. mask random consecutive spans.

Difference with BERT - Roberta

1. More big batch size when training bert.
2. More optimal hyperparameters.
3. More data.
4. Remove NSP task.

Difference with BERT - ALBERT

1. change NSP task to SOP (Sentence Order Prediction) task.
2. One layer recall 12 times to build 12 layers. (share parameters)
3. factorize embedding table. initize small and project to big dimension.