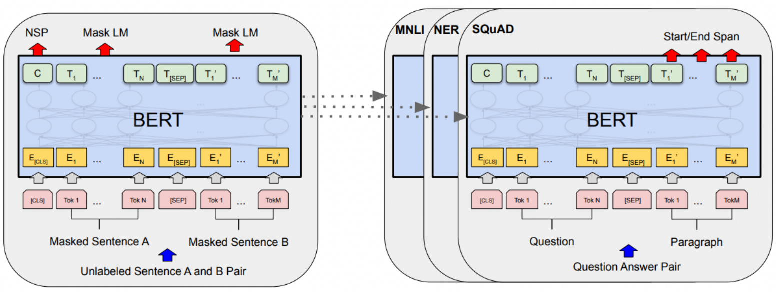
Review On Bert

CS410 Technology Review : Fall 2021

Minsoo Kim
minsoo4

Abstract
BERT is frequently used language model in variety of NLP fields. This article will review the paper "BERT:Pretraining of Deep Bidrectional Transformers for Language Modeling"

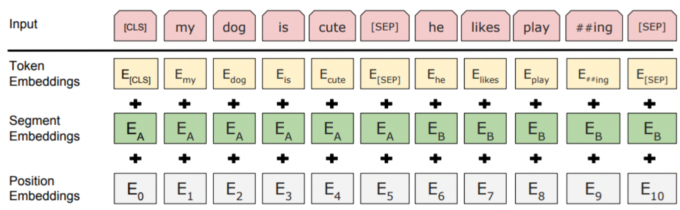
Introduction
There are two existing strategies for applying pre-trained language model, one is feature-based and another is fine-tuning. Both of them uses unidirectional language models during pre-training. This apporach has limitation because the every token can only attend to previous tokens. So some tasks like Question and Answering which needs to incorporate context have a short comings.
The important improvements of BERT is fine-tuning based on Bidirectional Encoder Representations from Transformers. Especially, bidirectional encoding makes the model to see the current word token which makes overfit to the model. However BERT used Maksed Language Model, which masks some of the tokens and predicting it. Another improvement of BERT is 'next sentence prediction'



Pre-training Fine-Tuning

Model Architecture
BERT has unified architecture across different tasks. It is a multi-layer bidirectional Transformer Encoder based on original tensor2tensor library. Two model sizes are as follows.

- $BERT_{BASE}$ (L=12, H=768, A=12, Total Parameters=110M)
- $BERT_{LARGE}$ (L=24, H=1024, A=16, Total Parameters=340M)



Input/Output Representations
BERT takes input in both a single sentence and a pair of sentences in one token sequence. This input representation makes the model fit to both of NLP tasks like language inference

and question and answering. It contains the first token [CLS]. If it is pair of sentences(<Question, Answer>), it has special token [SEP] between them.

Pretraining BERT
#1. Masked LM
Most conditional language model approaches are focused on unidirectional conditioning although learning bidirectional is conceptually better. Bidrectional approach has limitation of seeing it self. So hidden layers will be affected by this 'seeing'. BERT introduces Masked Language Model. Researchers have masked some of the tokens and the model predicted it. This masking has downside, which is mismatch between pre-traning and fine-tuning since there is [MASK] token does not appear during fine-tuning. So this paper changes chosen token into one of three. 1) masked toen 2) random token 3) unchanged.

#2. Next Sentence Prediction
question and answering and natural language interfence are based on understanding the relationship between two sentences. Additional approach is needed for it because language modeling can not capture the relationship between the sentences. So the BERT has pretrained for binarized next sentence prediction task. Half of pretraining data was actual next sentence and half was randomly sentence from corpus. Thie pretraining improved score on both of question and anwering and natural language inference.

*** Pretraining Data**
For pretraining, BooksCorpus(800M words) and English wikipedia(2,500M words) were used.

Fine-tuning BERT
As BERT takes pair of sentences, common pattern would be independently encoding two sentence before cross attention. However, BERT used self-attention mechanism to unify two stages. It simply encodes a concatenated sequence(text pair). So the bidirectional cross attention between two sentences were made.
For each tasks, it plugs in task specific inputs and outputs. pair of sentences were sentences for paraphrasing, hypothesis-premise of question-passage pairs fo question and answering and degenerate text-0 in text classification.
Fine-tuning is less expensive than pre-training.

Experiments
Paper explains that BERT has achieved state-of-the-art results on 11 results including following results. This is part of all scores from paper focused on question and answering which is the reason I am reviewing BERT for.

[GLUE score to 80.5%]

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

[Squad v1.1 question answering Test F1 to 93.2]

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

[Squad v2.0 question answering Test F1 to 83.1]

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

References

- **J Devlin, MW Chang, K Lee, K Toutanova, “BERT : Pre-training of deep bidirectional transformers for language understanding,” https://arxiv.org/abs/1810.04805**