NATURAL LANGUAGE PROCESSING

LECTURE 10: BERT

goorm

KAIST AI
Graduate School of AI

DAVIAN

BERTology: Self-Supervised Pre-training via Language Modeling

Pre-training overview

Initialize part of the model with networks trained using unsupervised learning

- Works Great !

But requires training a separate (usually extremely large) model

- Example: ELMo uses a 2-layer BiLSTM with 4096 units in each layer,
  also incorporated a size 2048 character-level CNN,
  pre-trained with 10 passes over a 1 billion word corpus

Improving language understanding by generative pre-training

GPT1

- It introduces special tokens, such as <S> /<E>/ $, to achieve effective transfer learning during fine-tuning
- It does not need to use additional task-specific architectures on top of transferred representations (e.g., ELMO)

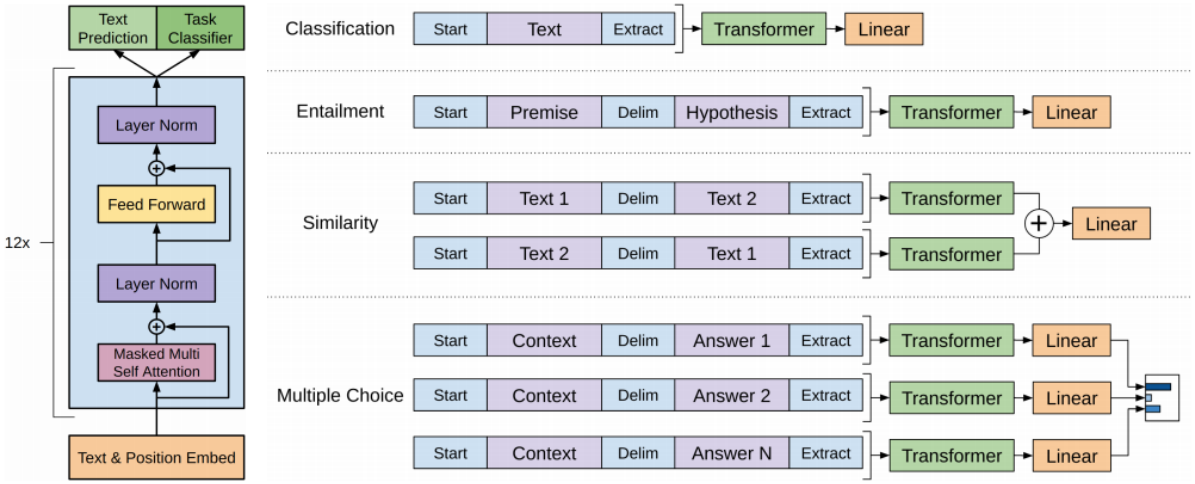


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

- 12-layer decoder-only transformer
- 12 head / 768 dimensional states
- GELU activation unit

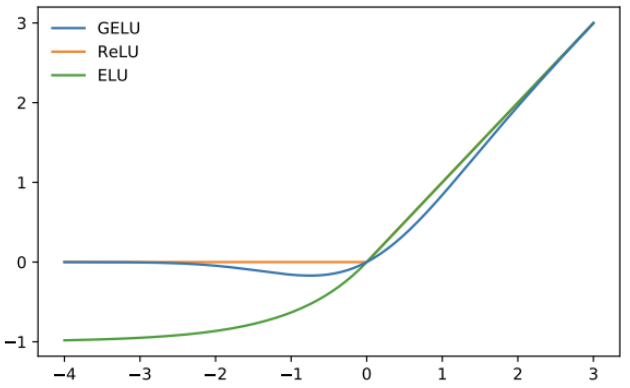


Figure 1: The GELU ($\mu = 0, \sigma = 1$), ReLU, and ELU ($\alpha = 1$).

https://blog.openai.com/language-unsupervised/

Improving language understanding by generative pre-training

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

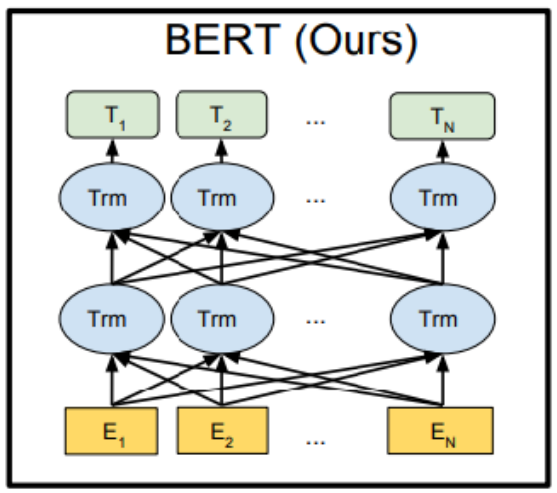
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

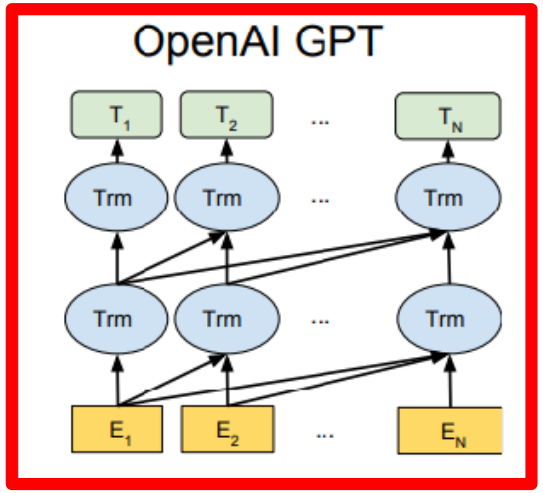
https://blog.openai.com/language-unsupervised/

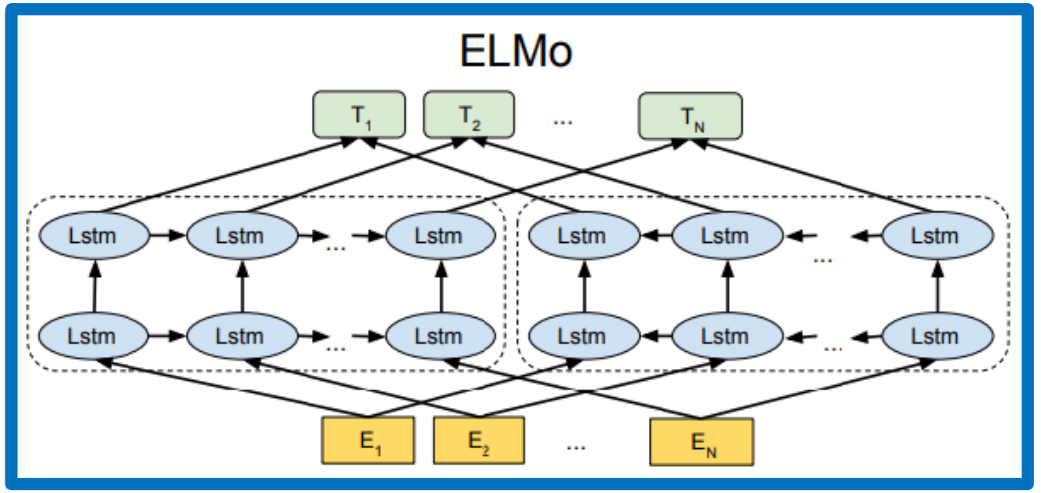
BERT

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

- Learn through masked language modeling tasks
- Use large-scale data and large-scale model







Unidirectional

LSTM

Masked LM: Problem with previous methods

Problem

- Language models only use left context or right context, but language understanding is bi-directional.

If we use the bidirectional language model?

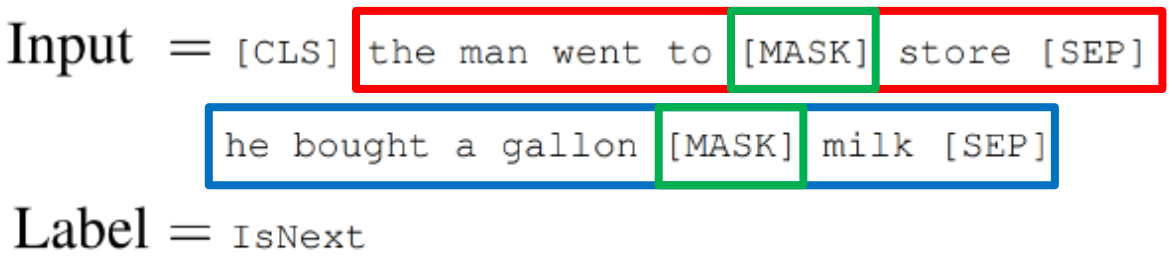
- Problem: Words can "see themselves"(cheating) in a bidirectional encoder

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

Pretraining tasks in BERT

- Masked Language Model (MLM)
  - Mask some percentage of the input tokens at random, and then predict those masked tokens.
  - 15% of the words to predict
    - 80% of the time, replace with [MASK]
    - 10% of the time, replace with a random word
    - 10% of the time, keep the sentence as same

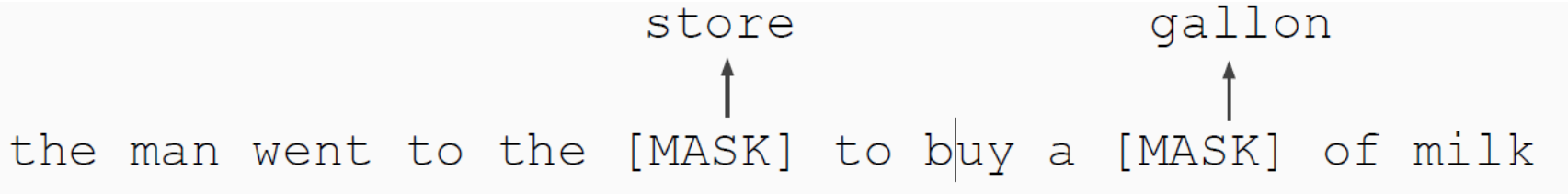
- Next Sentence Prediction (NSP)
  - Predict whether Sentence B is an actual sentence that proceeds Sentence A, or a random sentence



Masked LM

Solution

- Mask out *k*% of the input words, and then predict the masked words
  - We always use *k*= 15%



- Too little masking : Too expensive to train

- Too much masking : Not enough to capture context

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

Masked LM

Problem

- Mask token never seen during fine-tuning

Solution

- 15% of the words to predict, but don't replace with [MASK] 100% of the time. Instead:
  - 80% of the time, replace with [MASK]

    went to the store → went to the [MASK]
  - 10% of the time, replace with a random word

    went to the store → went to the running
  - 10% of the time, keep the same sentence

    went to the store → went to the store

Next sentence prediction

To learn the relationships among sentences, predict whether Sentence B is an actual sentence that proceeds Sentence A, or a random sentence

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`

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

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

1. Model Architecture
   - BERT BASE: L = 12, H = 768, A = 12
   - BERT LARGE: L = 24, H = 1024, A = 16
2. Input Representation
   - WordPiece embeddings (30,000 WordPiece)
   - Learned positional embedding
   - [CLS] – Classification embedding
   - Packed sentence embedding [SEP]
   - Segment Embedding
3. Pre-training Tasks
   - Masked LM
   - Next Sentence Prediction

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

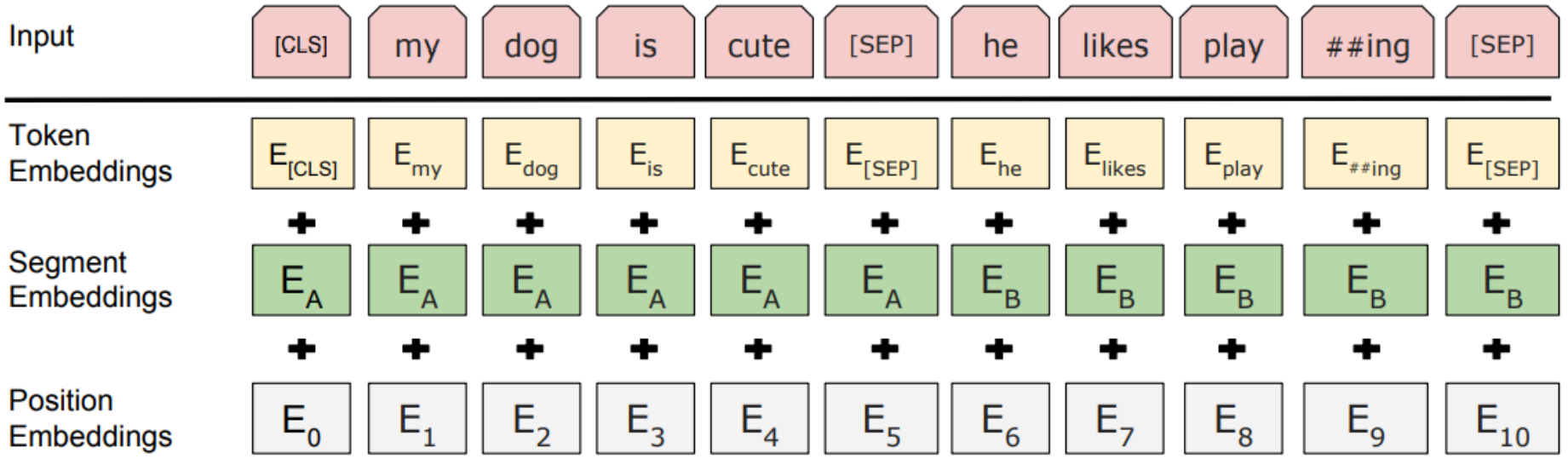
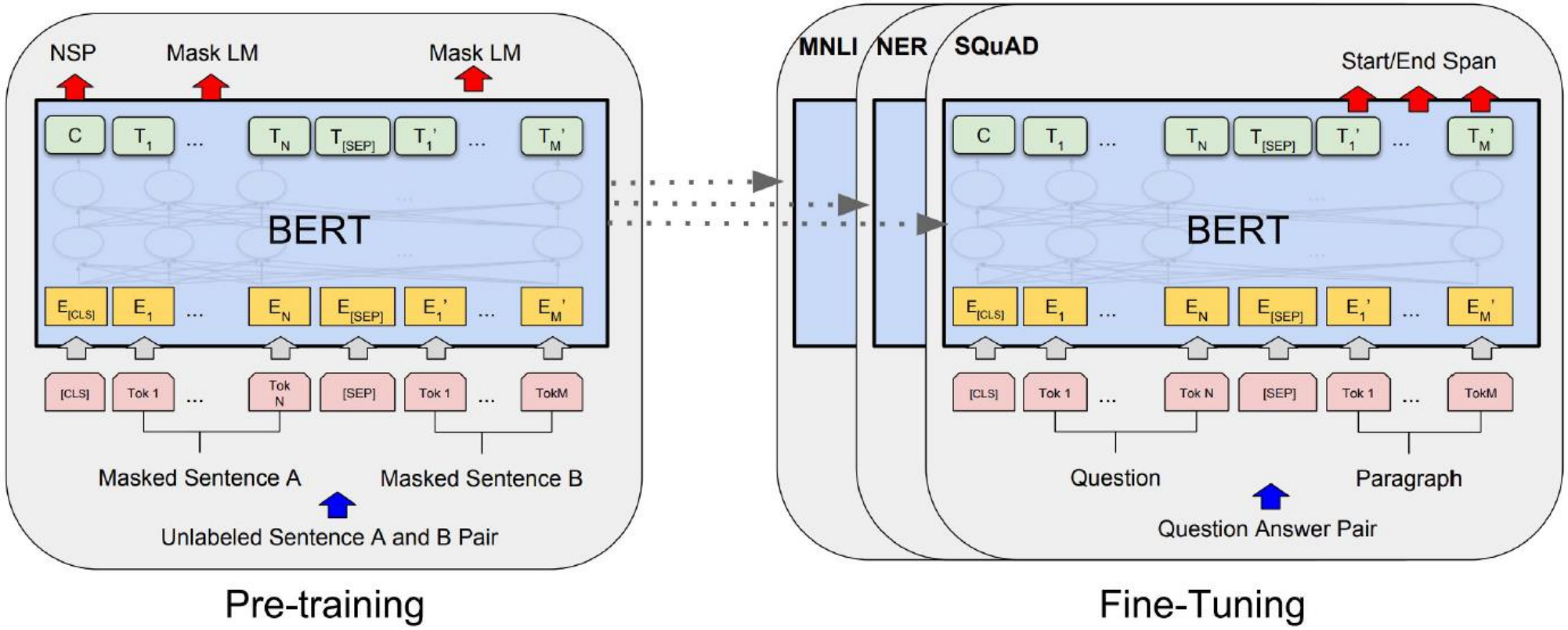


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

BERT: Fine-tuning process

Transfer Learning



BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

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

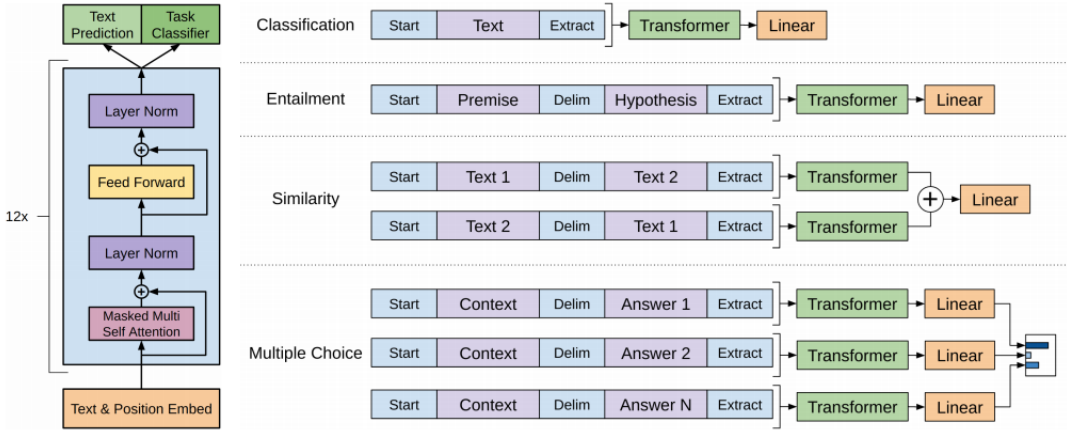
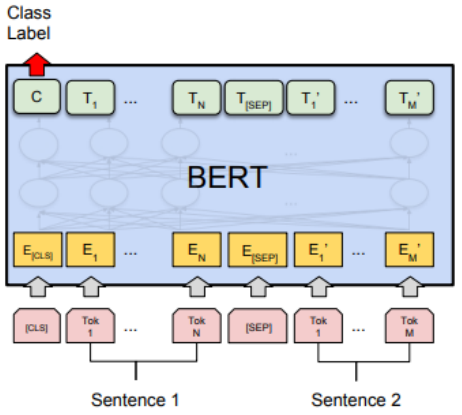
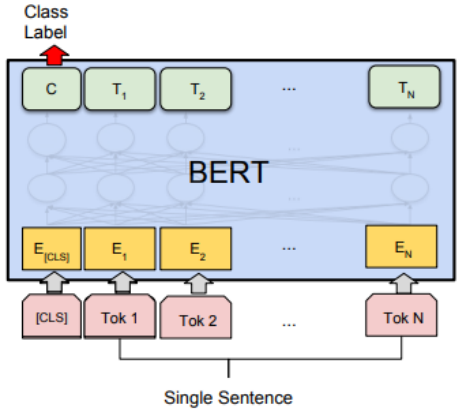


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

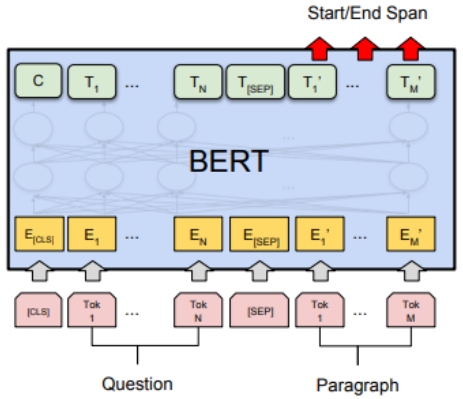
https://blog.openai.com/language-unsupervised/



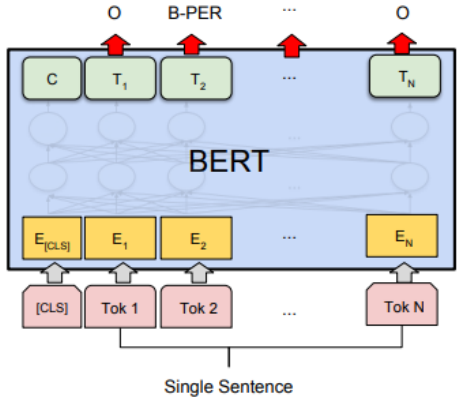
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



(c) Question Answering Tasks: SQuAD v1.1



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

BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

Comparison of BERT and GPT

- Training-data size
  - GPT is trained on BookCorpus(800M words) ; BERT is trained on the BookCorpus and Wikipedia (2,500M words)
- Training special tokens during training
  - BERT learns [SEP], [CLS], and sentence A/B embedding during pre-training
- Batch size
  - BERT – 128,000 words ; GPT – 32,000 words
- Task-specific fine-tuning
  - GPT uses the same learning rate of 5e-5 for all fine-tuning experiments; BERT chooses a task-specific fine-tuning learning rate.

BERT: GLUE benchmark results

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

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.
Hypothesis: Jainism hates nature.
Label: Contradiction

CoLa

Sentence: The wagon rumbled down the road.
Label: Acceptable

Sentence: The car honked down the road.
Label: Unacceptable

BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

Machine Reading Comprehension (MRC) Question Answering

Given

Document

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.

Question

Where is Daniel?

Answer

A: garden

Reading Comprehension

BERT: On SQuAD 1.1

What was another term used for the oil crisis?
Ground Truth Answers: first oil shock | shock | shock | first oil shock | shock
Prediction: shock

The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."

Only new parameters: Start vector and end vector

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

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
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

BERT: On SQuAD 2.0

- Use token 0 ([CLS]) to emit logit for "no answer"
- "No answer" directly competes with answer span
- Threshold is optimized on dev set

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

What action did the US begin that started the second oil shock?
Ground Truth Answers: <No Answer>
Prediction: <No Answer>

The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
12 Nov 08, 2018	BERT (single model) *Google AI Language*	80.005	83.061
20 Sep 13, 2018	nlnet (single model) *Microsoft Research Asia*	74.272	77.052

References

- https://nlp.stanford.edu/seminar/details/jdevlin.pdf
- http://jalammar.github.io/illustrated-bert/