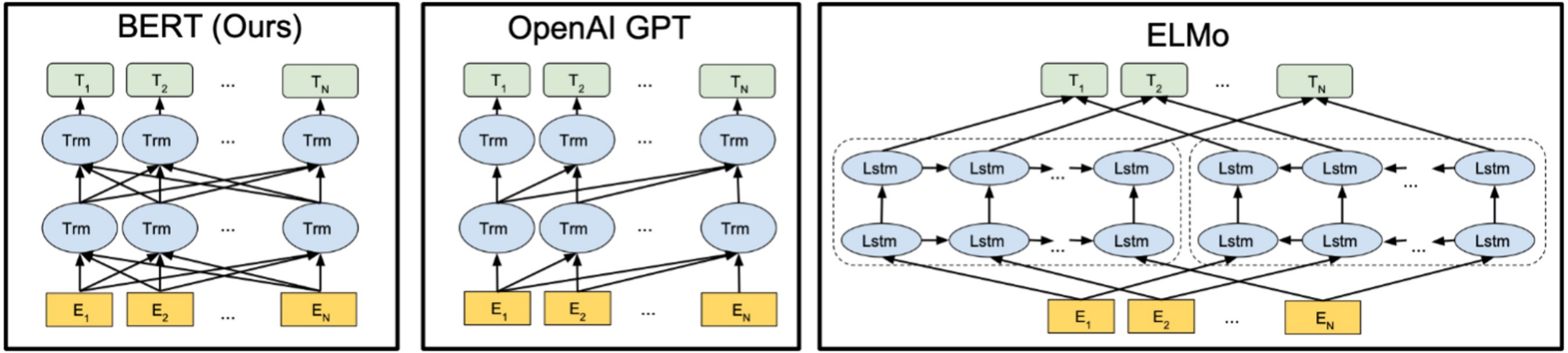
BERT

(BI-DIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

HAPPY 6TH BIRTHDAY BERT!!

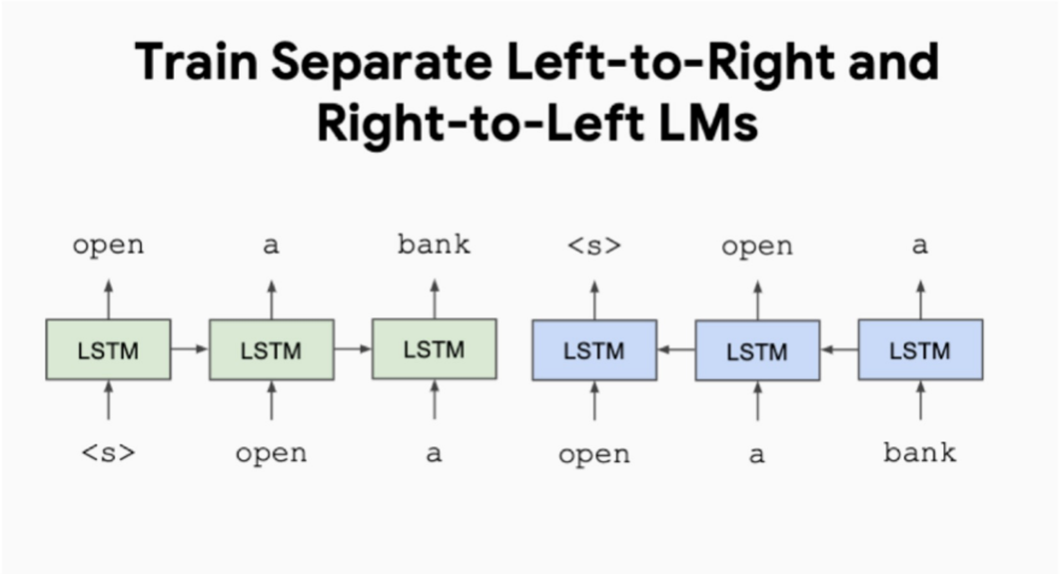
2018 IN MACHINE LEARNING AND NLP

BERT VS OPEN AI GPT VS ELMO



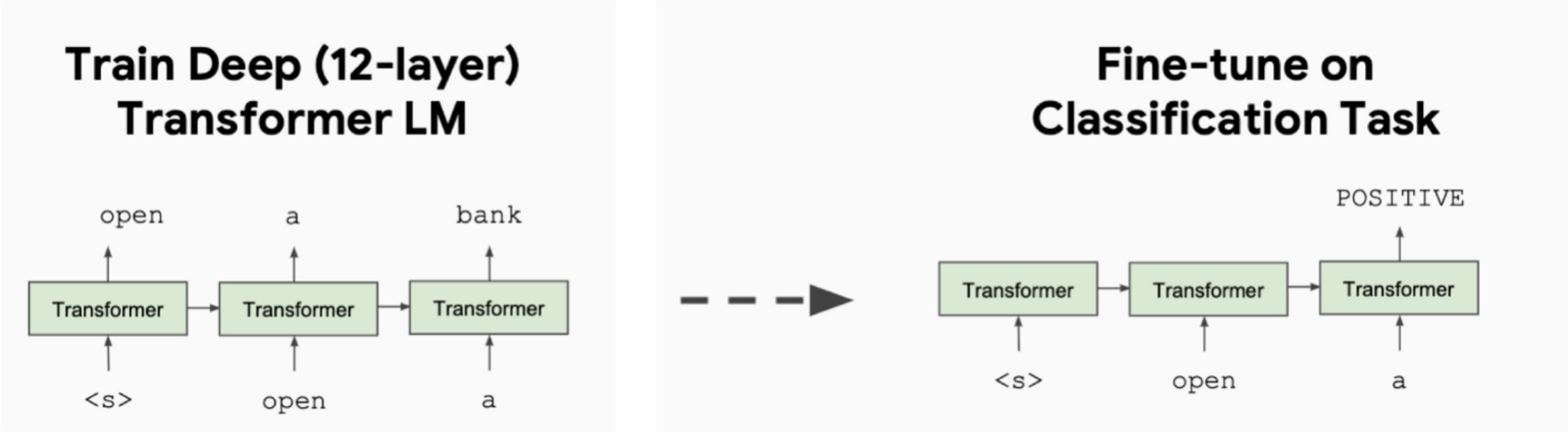
PRE-BERT: ELMO (EMBEDDINGS FROM LANGUAGE MODEL)

Generates contextual word embeddings, meaning the same word can have different meanings depending on the sentence.

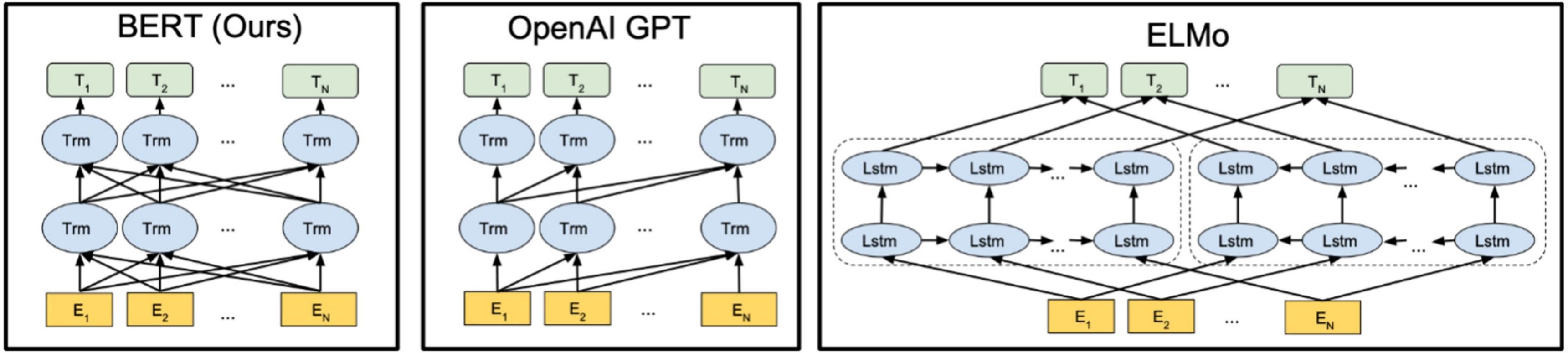


PRE-BERT: OPENAI GPT

A language model that reads text in one direction (left-to-right) using a deep Transformer decoder architecture.



BERT VS OPEN AI GPT VS ELMO



FINE-TUNING APPROACH

BERT uses a deep Transformer encoder and is designed to be fine-tuned for specific tasks.
Key Feature: It learns word representations using **bidirectional context**, meaning it looks at both the words before and after a target word.

- **Why?** Understanding both left and right contexts helps clarify word meaning.
- **Example 1:** "We went to the river bank." (Here, 'bank' refers to the river's edge.)
- **Example 2:** "I need to go to the bank to make a deposit." (Here, 'bank' refers to a financial institution.)

BIDIRECTIONAL CONDITIONING

↓

INDIRECTLY SEE ITSELF IN MULTI-LAYERED CONTEXT?

MASKED LANGUAGE MODELING (MLM)!

Solution: Mask out k% of the input words, and then predict the masked words

store gallon

↑ ↑

the man went to the [MASK] to buy a [MASK] of milk

k: usually 15%

- Too much masking → not enough context
- Too little masking → computationally expensive

MLM (CONTINUED)

Selection of masked tokens:

- 15% are uniformly sampled.

80-10-10 Corruption

- 10% are unchanged.

  Let’s go to the bank’s ATM → Let’s go to the bank’s ATM

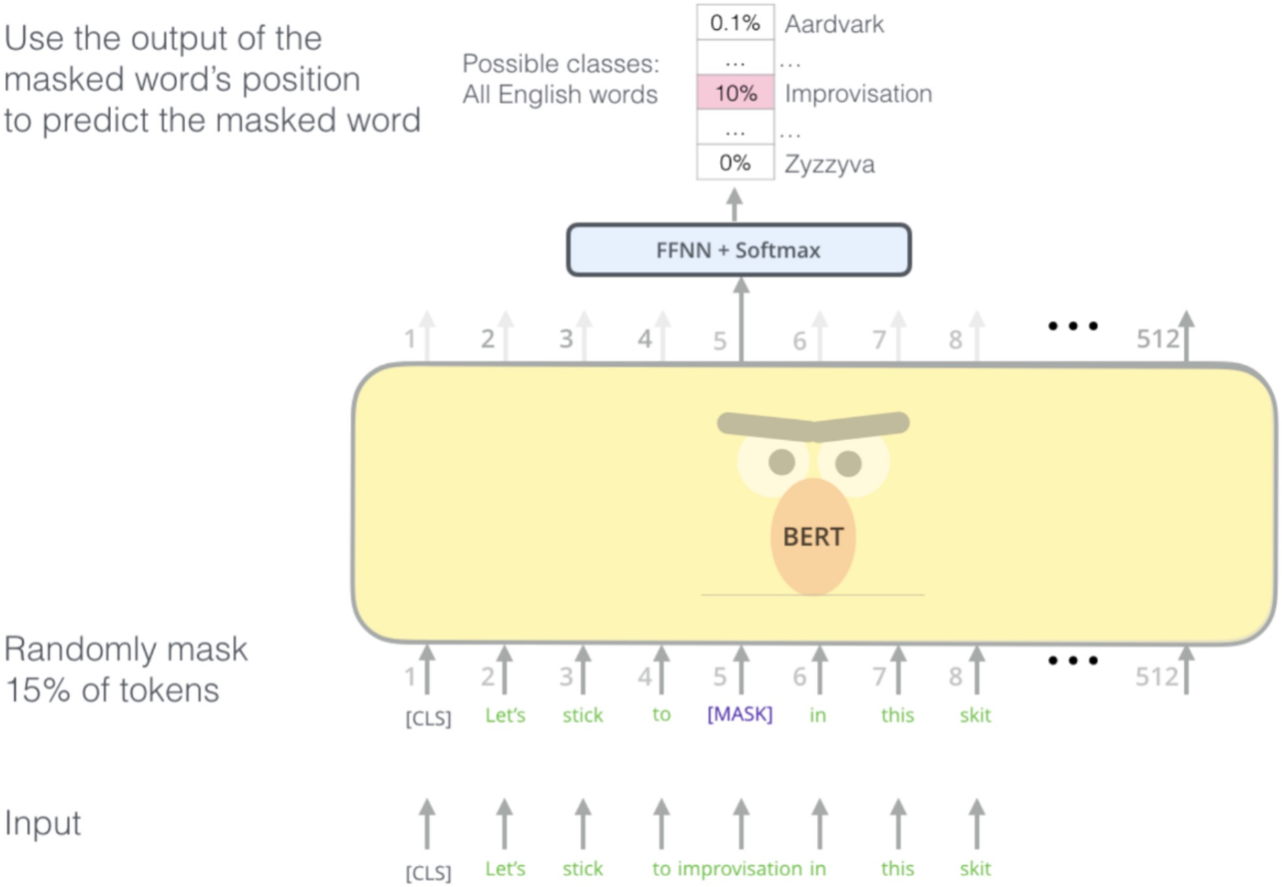
  → Always biased to the correct selection.

- 10% are replaced with a random word in the vocabulary.

  Let’s go to the bank’s ATM → Let’s go to the boo ATM

- 80% of predicted words are replaced with the [MASK] token.

  Let’s go to the bank’s ATM → Let’s go to the [MASK] ATM



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

HANDLING RELATIONSHIPS BETWEEN MULTIPLE SENTENCES:

TWO SENTENCE TASKS

GIVEN TWO SENTENCES A AND B, IS B LIKELY TO BE THE SENTENCE THAT FOLLOWS A OR NOT?

NEXT SENTENCE PREDICTION (NSP)

NSP is designed to reduce the gap between pre-training and fine-tuning

[CLS]: a special token always at the beginning

[SEP]: a special token used to separate two segments

Input = `[CLS] the man went to [MASK] store [SEP]`

`he bought a gallon [MASK] milk [SEP]`

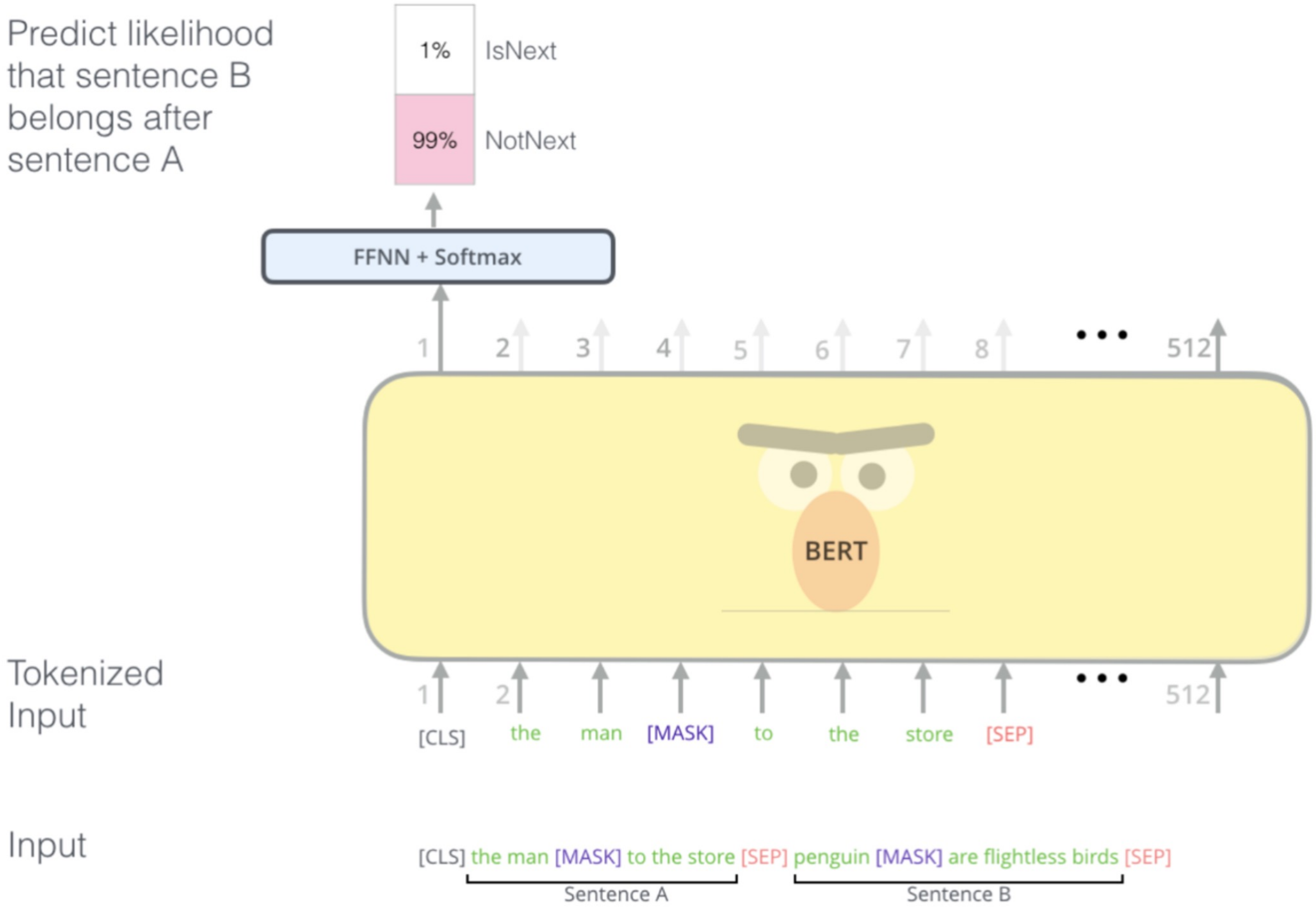
Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP]`

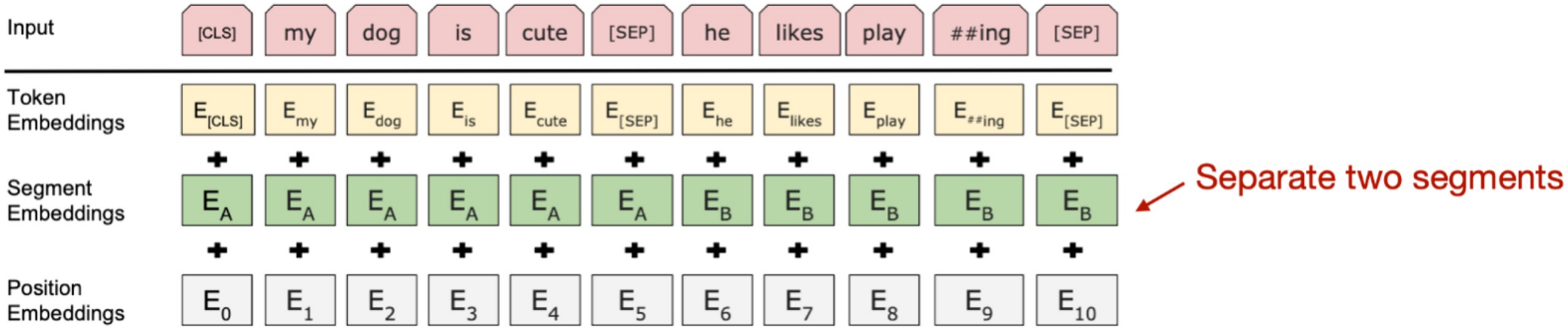
`penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

They sample two contiguous segments for 50% of the time and another random segment from the corpus for 50% of the time



The second task BERT is pre-trained on is a two-sentence classification task. The tokenization is oversimplified in this graphic as BERT actually uses WordPieces as tokens rather than words --- so some words are broken down into smaller chunks.


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}
Separate two segments

BERT BASE AND BERT LARGE

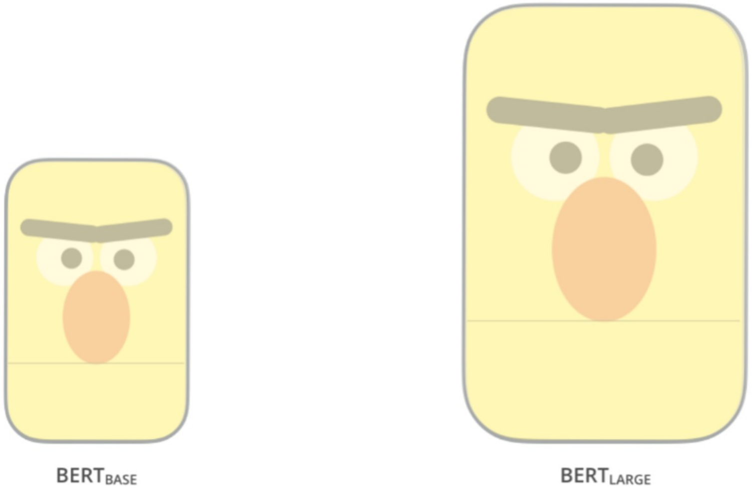
BERT-base: 12 layers, 768 hidden size, 12 attention heads, 110M parameters

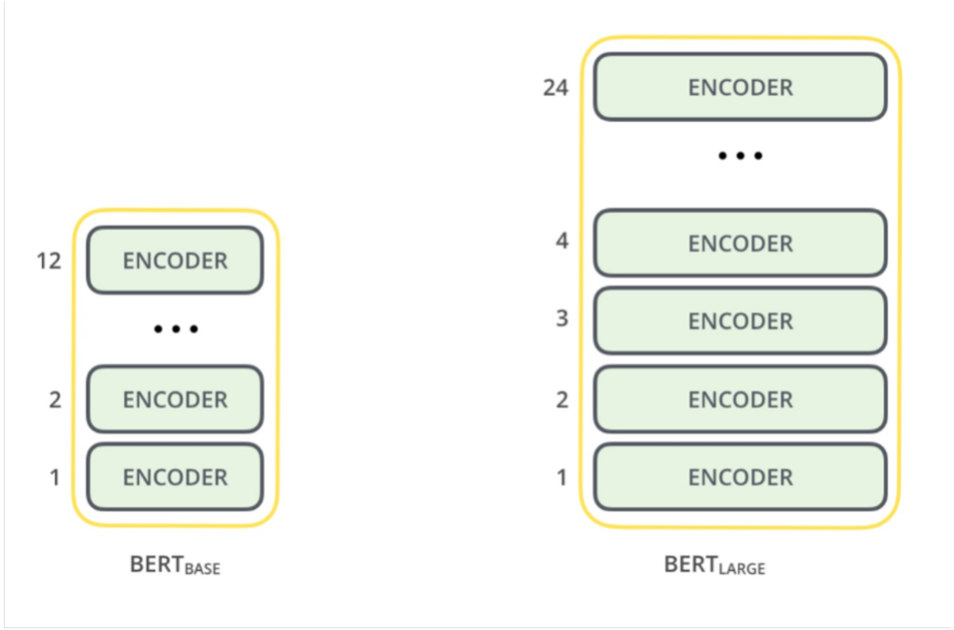
- Same hidden size as OpenAI GPT

• BERT-large: 24 layers, 1024 hidden size, 16 attention heads, 340M parameters

BERT-base: developed for performance comparison with OpenAI GPT

BERT-large: grossly large model for state of the art results


BERT BASE
BERT LARGE


24
ENCODER
4
ENCODER
3
ENCODER
2
ENCODER
1
ENCODER
12
ENCODER
2
ENCODER
1
ENCODER
BERT BASE
BERT LARGE

BERT PRE-TRAINING

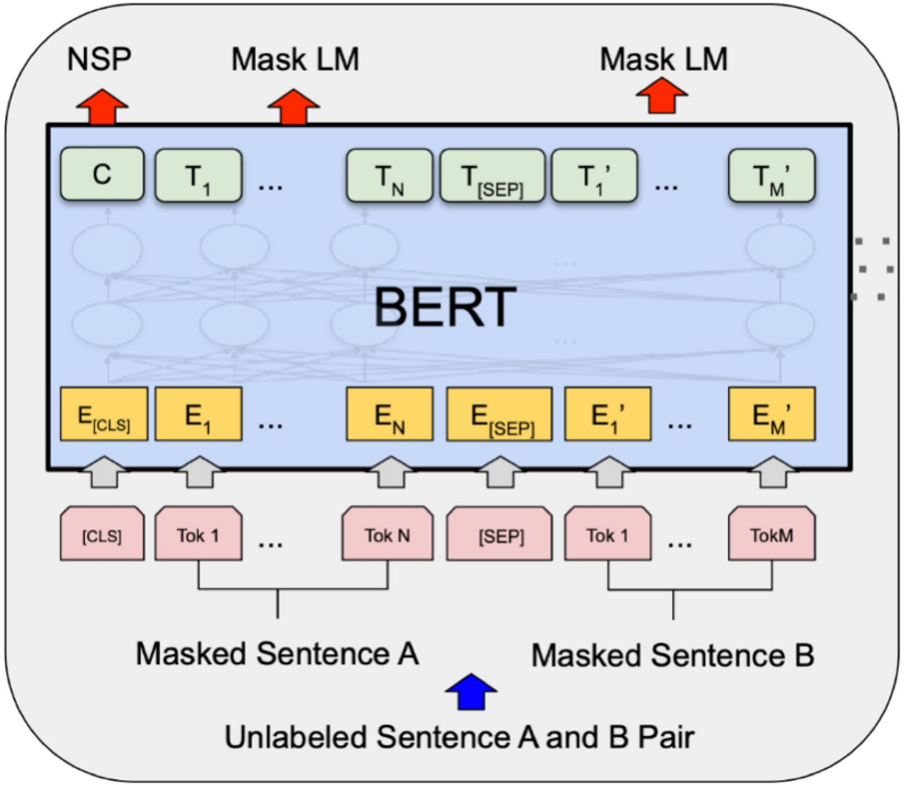
- Training corpus: Wikipedia (2.5B) + BooksCorpus (0.8B)
  - OpenAI GPT was trained on BooksCorpus only.
- Max sequence size: 512 word pieces (roughly 256 and 256

for two non-contiguous sequences)

- Trained for 1M steps, batch size 128k

BERT PRE-TRAINING

- MLM and NSP are trained together
- [CLS] is pre-trained for NSP
- Other token representations are trained for MLM



FINE-TUNING BERT: “PRETRAIN ONCE, FINETUNE MANY TIMES”

Sentence Level Tasks	Token Level Tasks

SENTENCE LEVEL TASKS

- Sentence Pair Classification Tasks

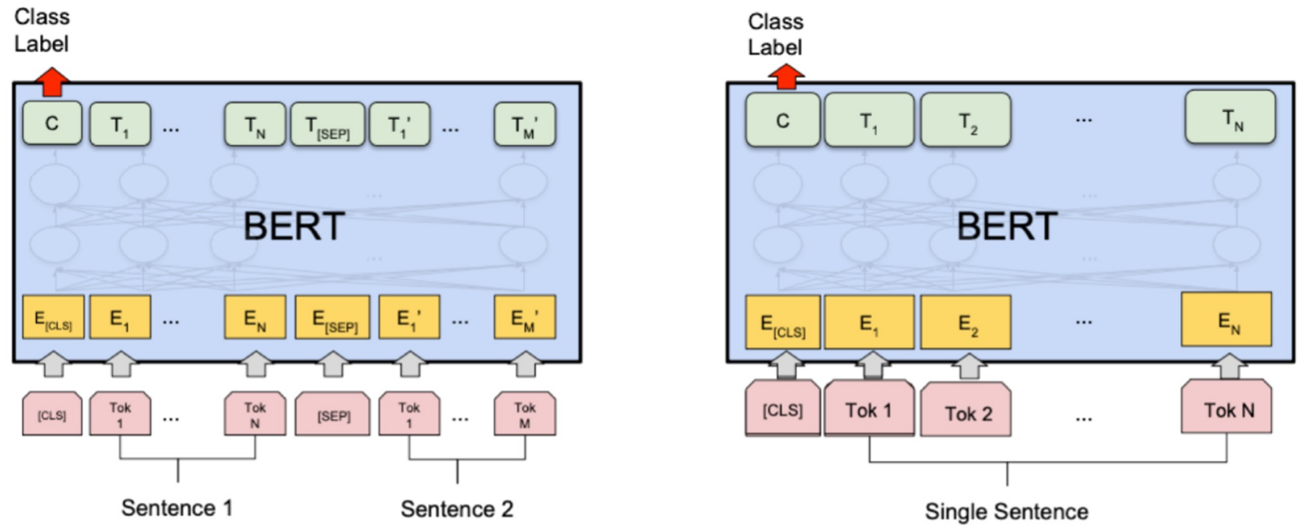
MNLI:
Premise: A soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.
{entailment, contradiction, neutral}

QQP:
Q1: Where can I learn to invest in stocks?
Q2: How can I learn more about stocks?
{duplicate, not duplicate}

- Single Sentence Classification Tasks

SST2:
rich veins of funny stuff in this movie
{positive, negative}

SENTENCE LEVEL TASKS



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

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

- For sentence pair tasks, use [SEP] to separate the two segments with segment embeddings
- Add a linear classifier on top of [CLS] representation and introduce C × h new parameters (C: # of classes, h: hidden size)

TOKEN LEVEL TASKS

- Extractive Question Answering

SQuAD

Question: The New York Giants and the New York Jets play at which stadium in NYC ?

Context: The city is represented in the National Football League by the New York Giants and the New York Jets , although both teams play their home games at MetLife Stadium in nearby East Rutherford , New Jersey , which hosted Super Bowl XLVIII in 2014 .

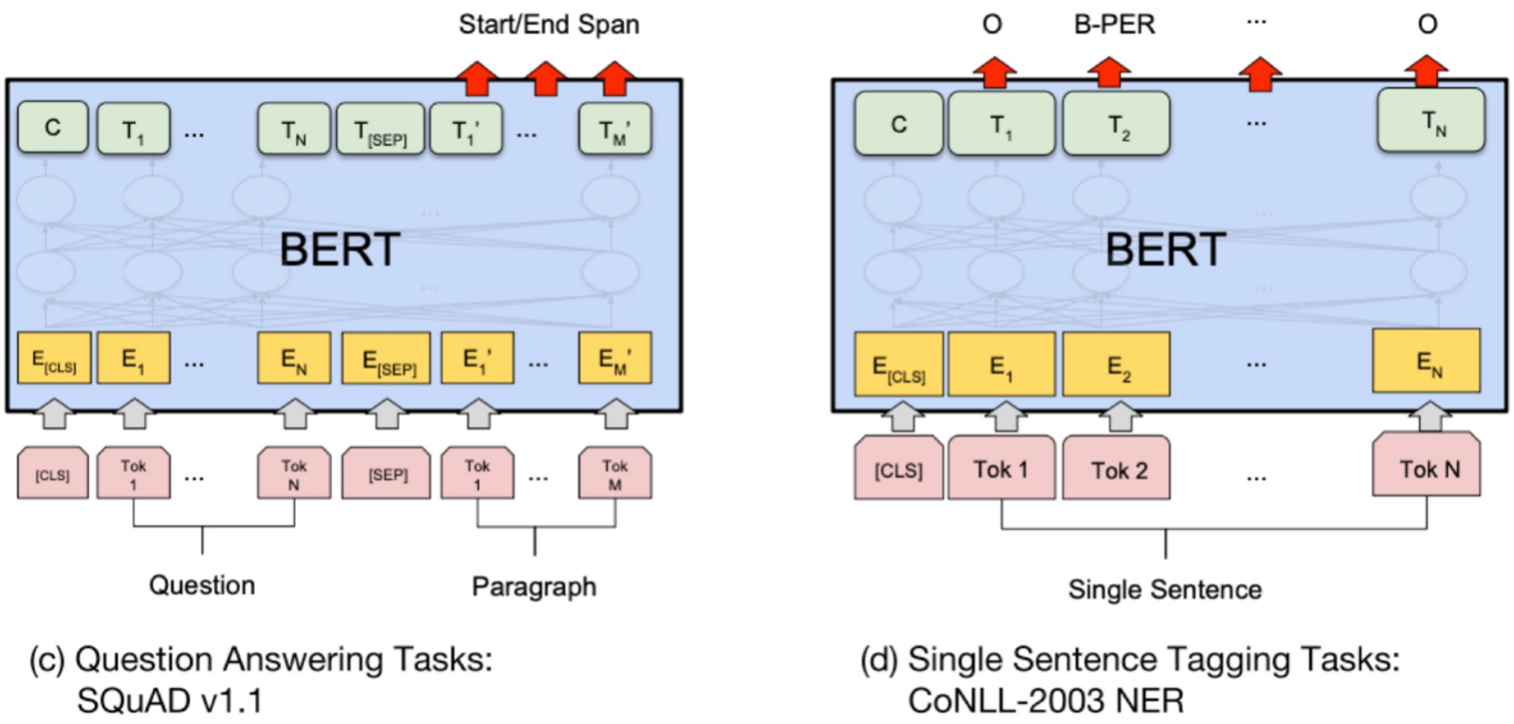
(Training example 29,883)

- Named Entity Recognition

CoNLL 2003 NER

John	Smith	lives	in	New	York
B-PER	I-PER	O	O	B-LOC	I-LOC

TOKEN LEVEL TASKS



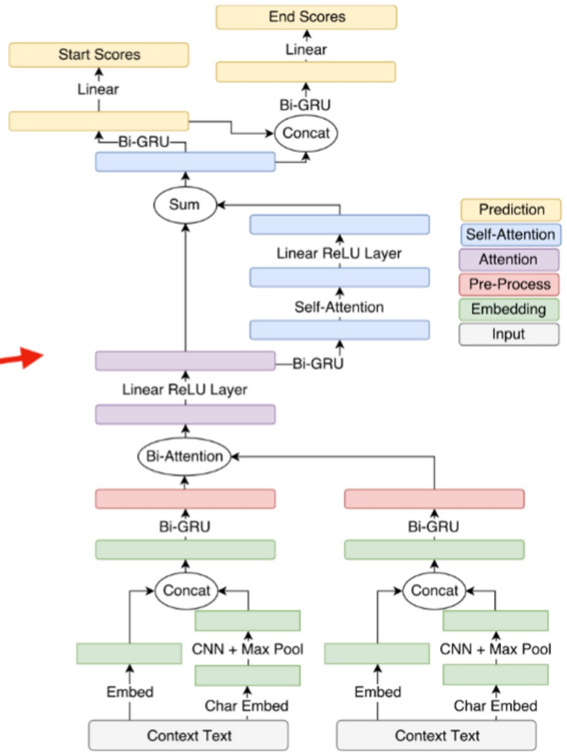
- For token-level prediction tasks, add linear classifier on top of hidden representations

EXPERIMENTAL RESULTS: GLUE

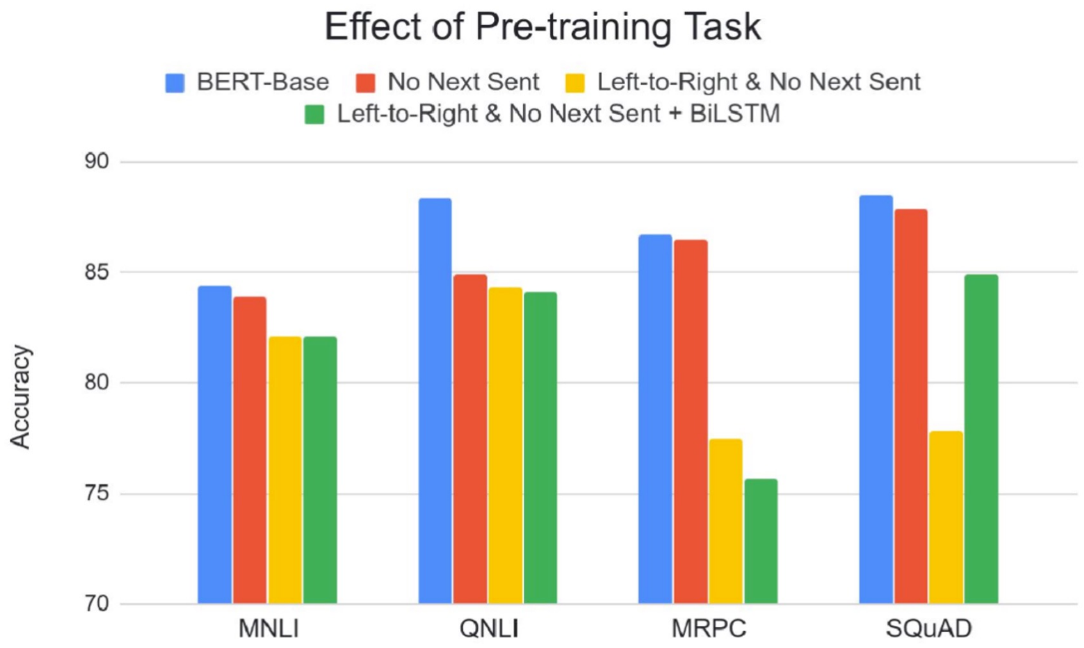
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	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**

EXPERIMENTAL RESULTS: SQUAD

System	Dev EM	Dev F1	Test EM	Test 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**



ABLATION STUDY: PRE-TRAINING TASKS



- MLM>> left-to-right LMs
- NSP improves on some tasks

Later work (Joshi et al. 2020, Liu et al. 2019) argued that NSP is not useful.

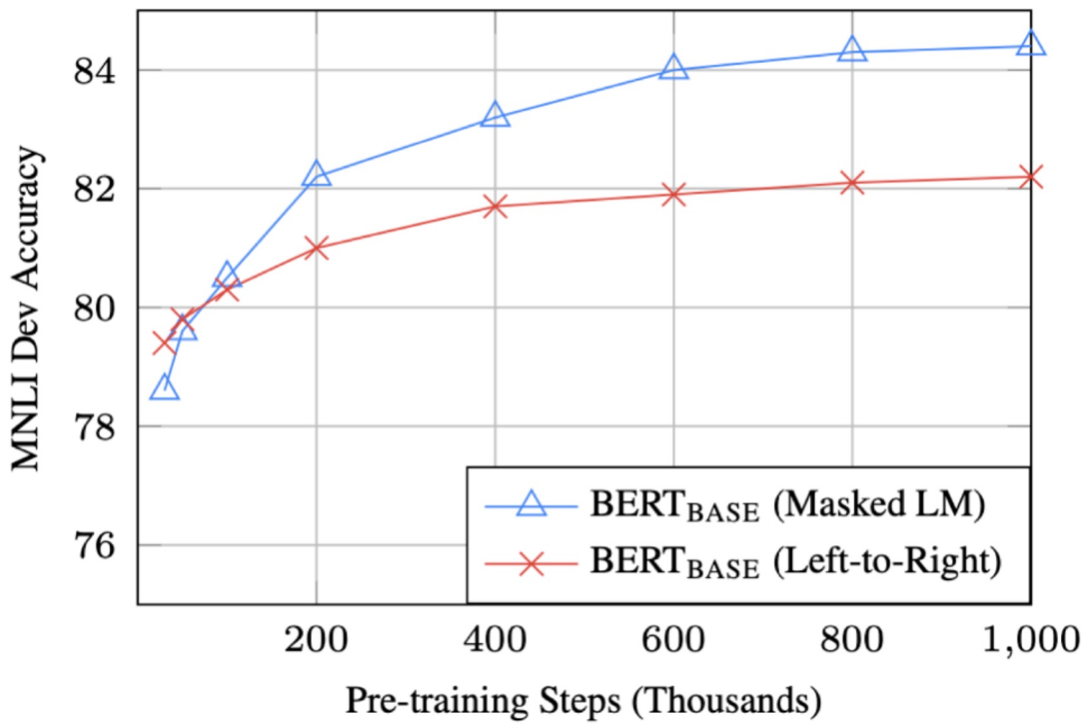
ABLATION STUDY: MODEL SIZE

\# layers — hidden size — # of heads

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

The bigger the better!!!

ABLATION STUDY: TRAINING EFFICIENCY



MLM takes longer to converge because it only predicts 15% of tokens.

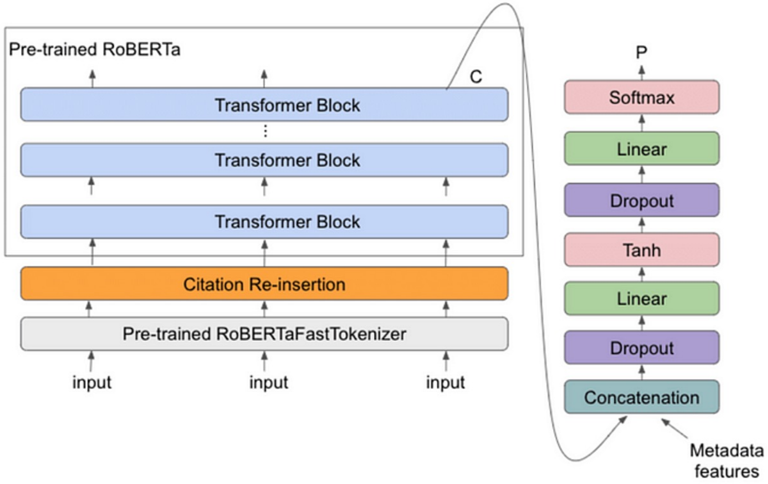
CONCLUSIONS (IN EARLY 2019)

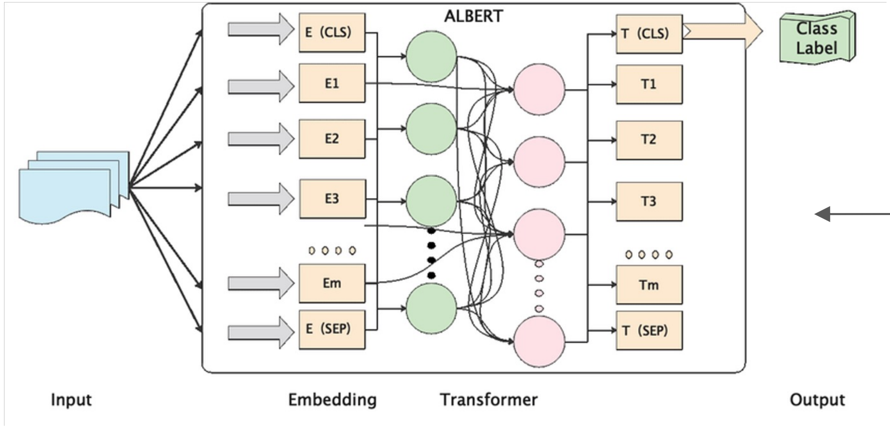
The empirical results from BERT are great, but the biggest impact on the field is:

- With pre-training, bigger == better, without clear limits so far.

AFTER BERT

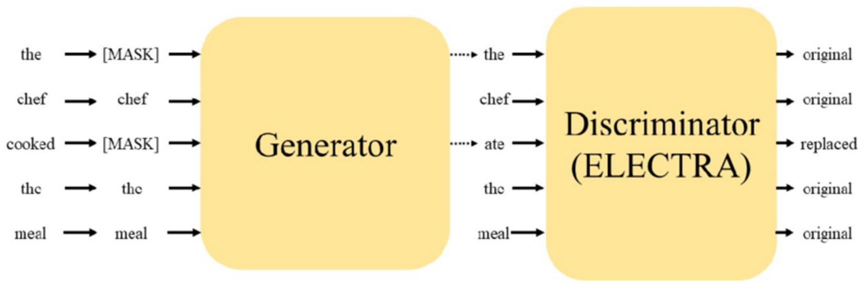
RoBERTa (Liu et al., 2019)





ALBERT (Lan et al., 2020)

ELECTRA (Clark et al., 2020)



AFTER BERT

- Models that handle long contexts ($\gg$ 512 tokens)
  - Longformer, Big Bird (this is really cute), …
- Multilingual BERT
  - Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary
- BERT extended to different domains
  - SciBERT, BioBERT, FinBERT, ClinicalBERT, …
- Making BERT smaller to use
  - DistillBERT, TinyBERT, …

AFTER BERT

Text Generation Using BERT (generally less effective compared to OpenAI's GPT

BERT has a Mouth, and It Must Speak:
BERT as a Markov Random Field Language Model

Alex Wang
New York University
alexwang@nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Mask-Predict: Parallel Decoding of
Conditional Masked Language Models

Marjan Ghazvininejad* **Omer Levy*** **Yinhan Liu*** **Luke Zettlemoyer**
Facebook AI Research
Seattle, WA

Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis--Hastings

Kartik Goyal, Chris Dyer, Taylor Berg-Kirkpatrick

src	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed on November 20th .

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Sascha Rothe, Shashi Narayan, Aliaksei Severyn

NEW RANKINGS! (USING GLUE)

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
8	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
9	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
10	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1

CITATIONS

Devlin, Jacob, et al. “Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *arXiv.Org*, 24 May 2019, arxiv.org/abs/1810.04805.

Fall 2022 Lecture 2: Bert (Encoder-Only Models), www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf. Accessed 23 Oct. 2024.

Alammar, Jay. “The Illustrated Bert, Elmo, and Co. (How NLP Cracked Transfer Learning).” *The Illustrated BERT, ELMo, and Co. (How NLP Cracked Transfer Learning) – Jay Alammar – Visualizing Machine Learning One Concept at a Time.*, jalammar.github.io/illustrated-bert/. Accessed 23 Oct. 2024.

Bert Image from Muppet Wiki: 700 × 1,165

Elmo Image from Muppet Wiki: 800 × 979