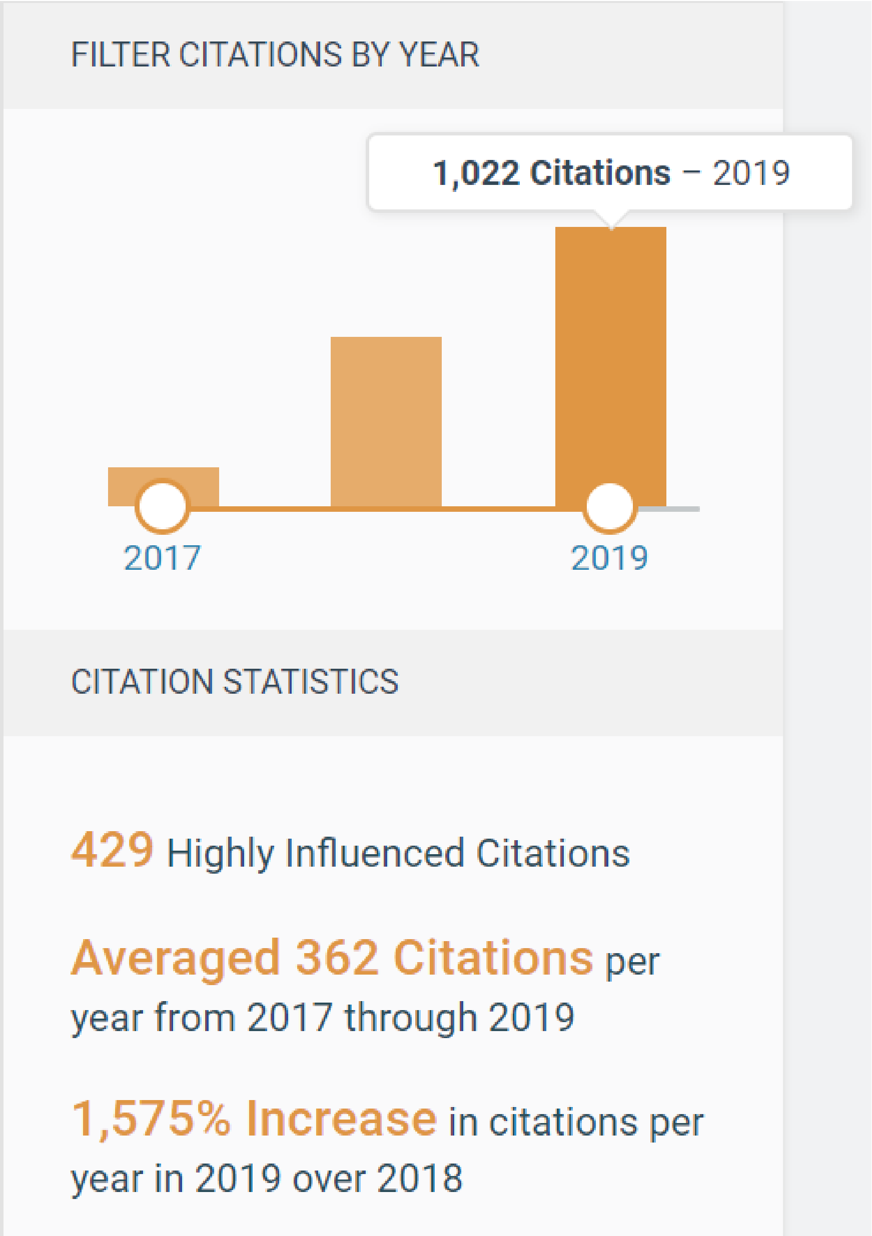
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin; Ming-Wei Chang; Kenton Lee; Kristina Toutanova

Presented by Mitchell Donley & Yinghao Li

Citations

9/3/19



GLUE Benchmark

Rank	Name	Model	Score
1	Facebook AI	RoBERTa	88.5
2	XLNet Team	XLNet-Large (ensemble)	88.4
3	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	87.6
4	GLUE Human Baselines	GLUE Human Baselines	87.1
5	王玮	ALICE large ensemble (Alibaba DAMO NLP)	87.0
6	Stanford Hazy Research	Snorkel MeTaL	83.2
7	XLM Systems	XLM (English only)	83.1
8	张倬胜	SemBERT	82.9
9	Danqi Chen	SpanBERT (single-task training)	82.8
10	Kevin Clark	BERT + BAM	82.3
11	Nitish Shirish Keskar	Span-Extractive BERT on STILTs	82.3
12	Jason Phang	BERT on STILTs	82.0
13	廖亿	RGLM-base (Huawei Noah's Ark Lab)	81.0
14	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden	80.5
15	Neil Houlsby	BERT + Single-task Adapters	80.2
16	Zhuohan Li	Macaron Net-base	79.7
17	蘇大鈞	SESAME-BERT-Base	78.6
18	Linyuan Gong	StackingBERT-Base	78.4
19	GLUE Baselines	BiLSTM+ELMo+Attn	70.0

Containing

General Information & Background

Model Architecture

Performances

General Information

BERT is a language representation model;

Designed to pre-train deep bidirectional representations from unlabeled text;

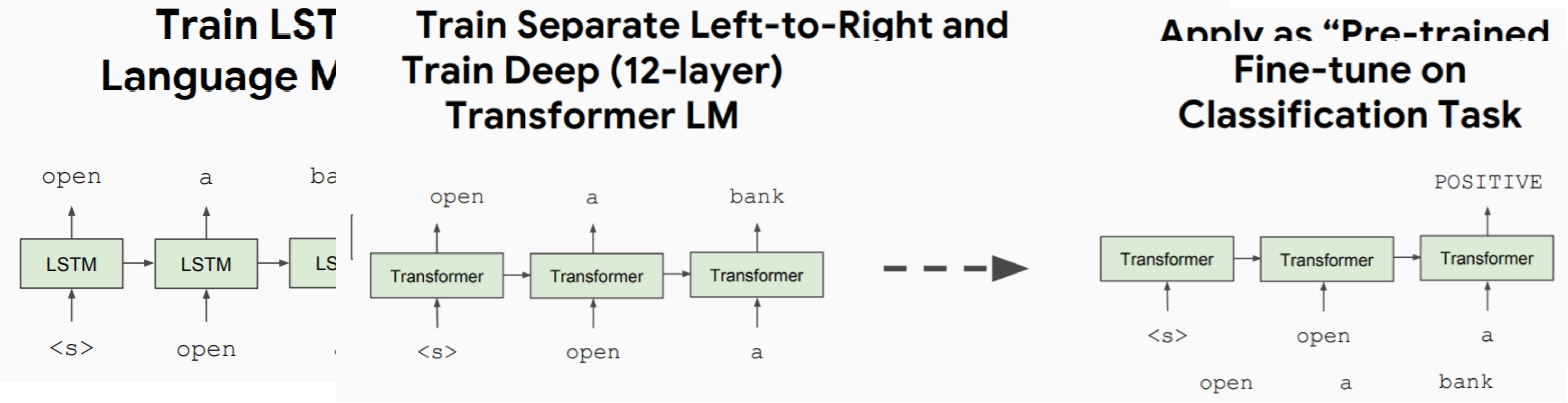
Pre-trained model can be fine-tuned with minimum adjustment to network structure to adapt to different tasks;

Achieves state-of-the-art performances on GLUE Benchmarks;

NOT suitable for text generation (which is the strength of OpenAI GPT).

Language Representation

- **Word Embedding** (Word2Vec, GloVe): irrelevant to the input context;
- Semi-Supervised Sequence Learning:



Heavily borrowed form https://nlp.stanford.edu/seminar/details/jdevlin.pdf

Problem with Previous Contextual Representation Methods

- Training objective: Next word prediction
- Previous language models only use left context or right context, but language understanding is bidirectional.

- Why?
  - Words can "see themselves (from the previous or subsequent inputs)" in a bidirectional encoder.

Masked Language Model (MLM)

- To prevent words from seeing themselves, 15% of the input tokens are randomly selected and
  - 80% of them are substitute by `[MASK]`token;
    - `went to the store -> went to the [MASK]`
  - 10% of them are substitute by random tokens;
    - `went to the store -> went to the covfefe`
  - The rest 10% are unchanged.
    - `went to the store -> went to the store`
- Above procedure is to mitigate the mismatch between pre-training and fine-tuning where `[MASK]` token never appears.

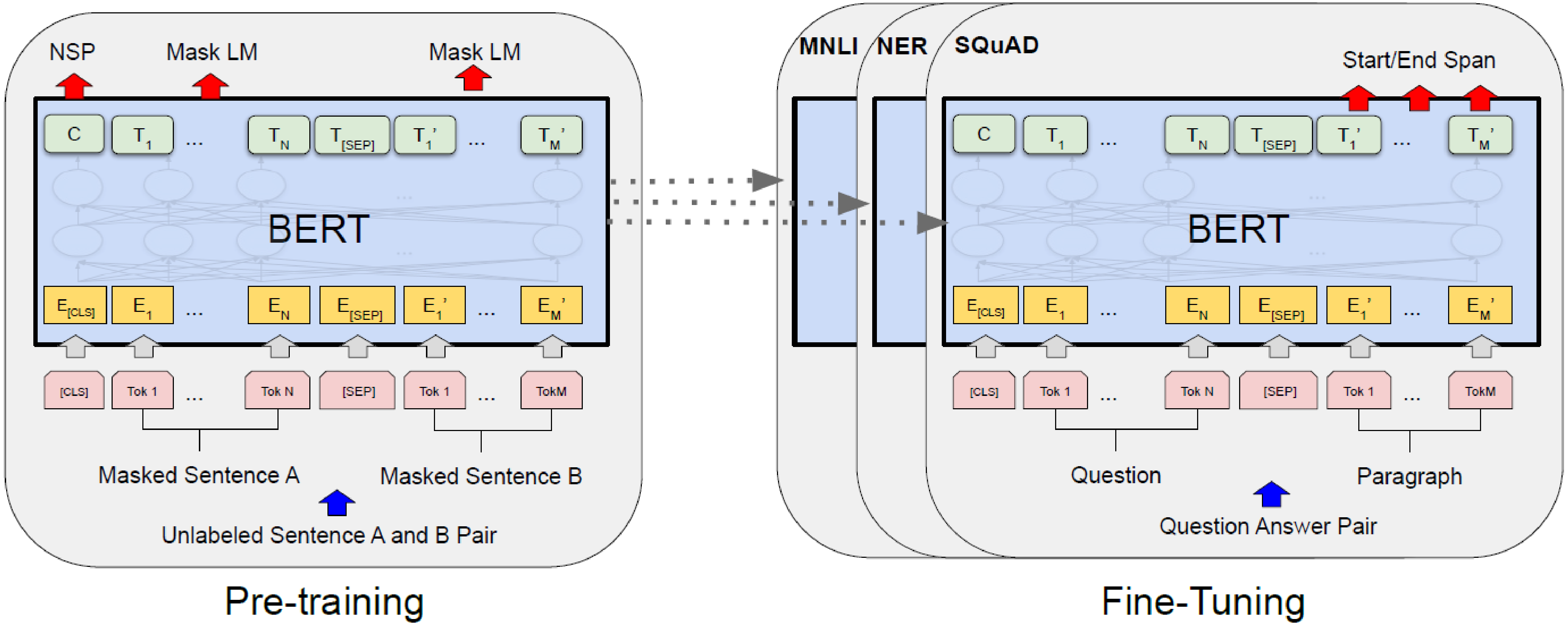
Next Sentence Prediction (NSP)

- To learn relationships between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

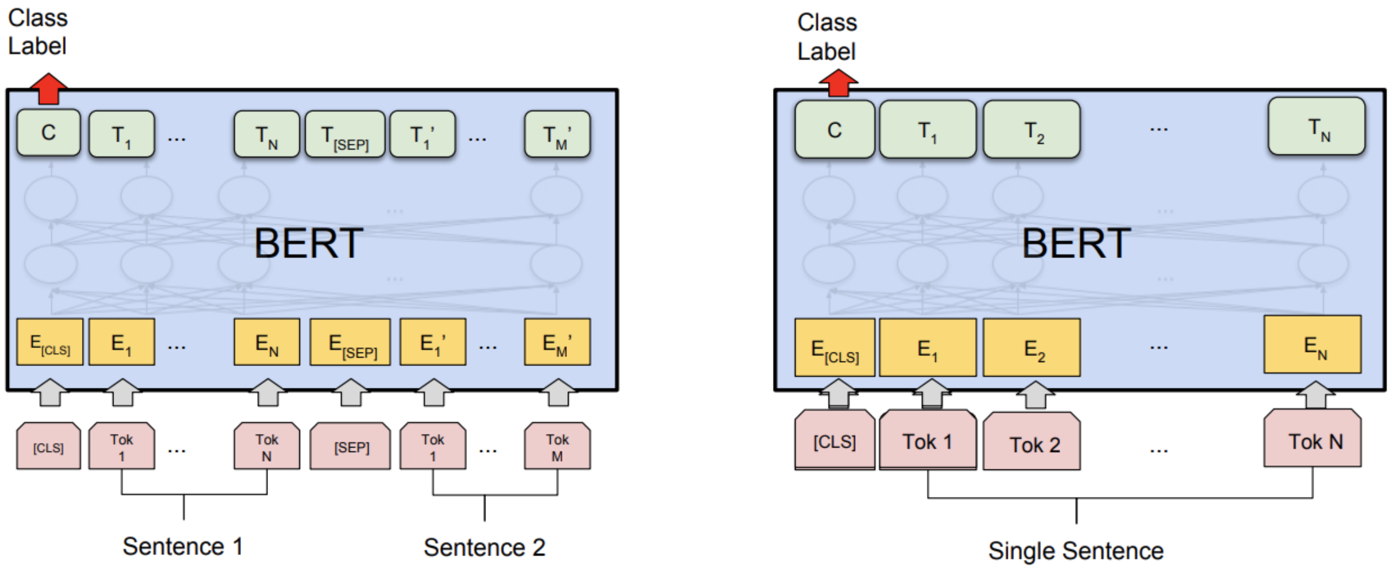
Pre-training and Fine-tuning



Fine-tuning BERT

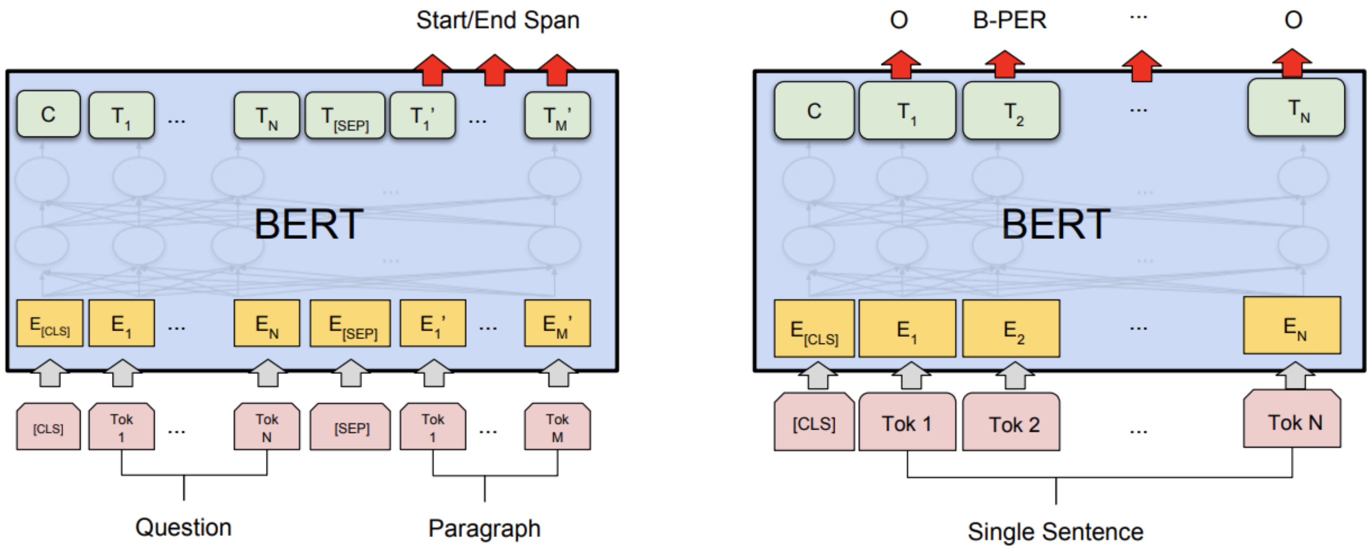
Plug in the task-specific inputs and outputs into BERT and fine-tune all the parameters end-to-end

9/3/19



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

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



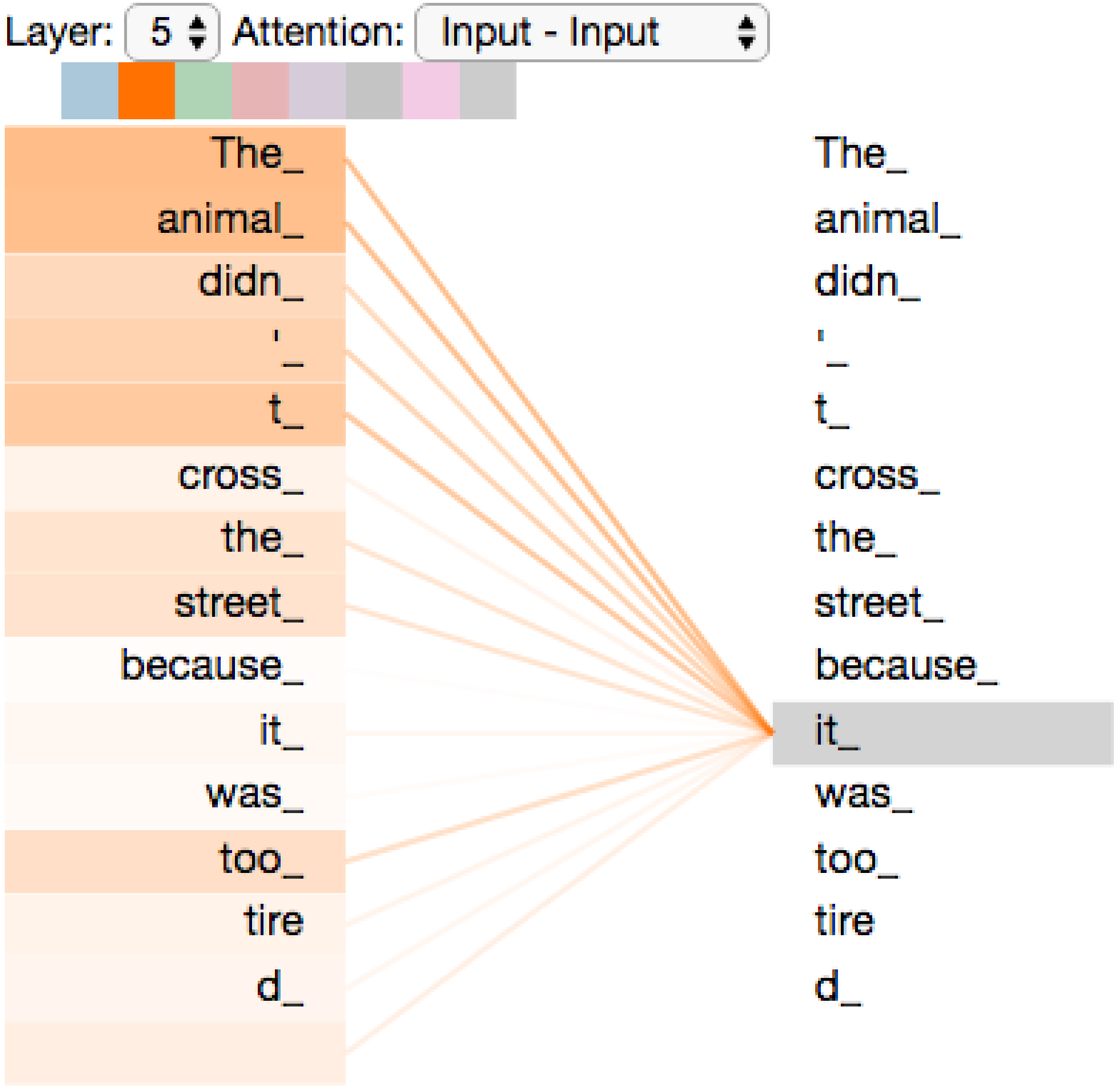
(c) Question Answering Tasks: SQuAD v1.1

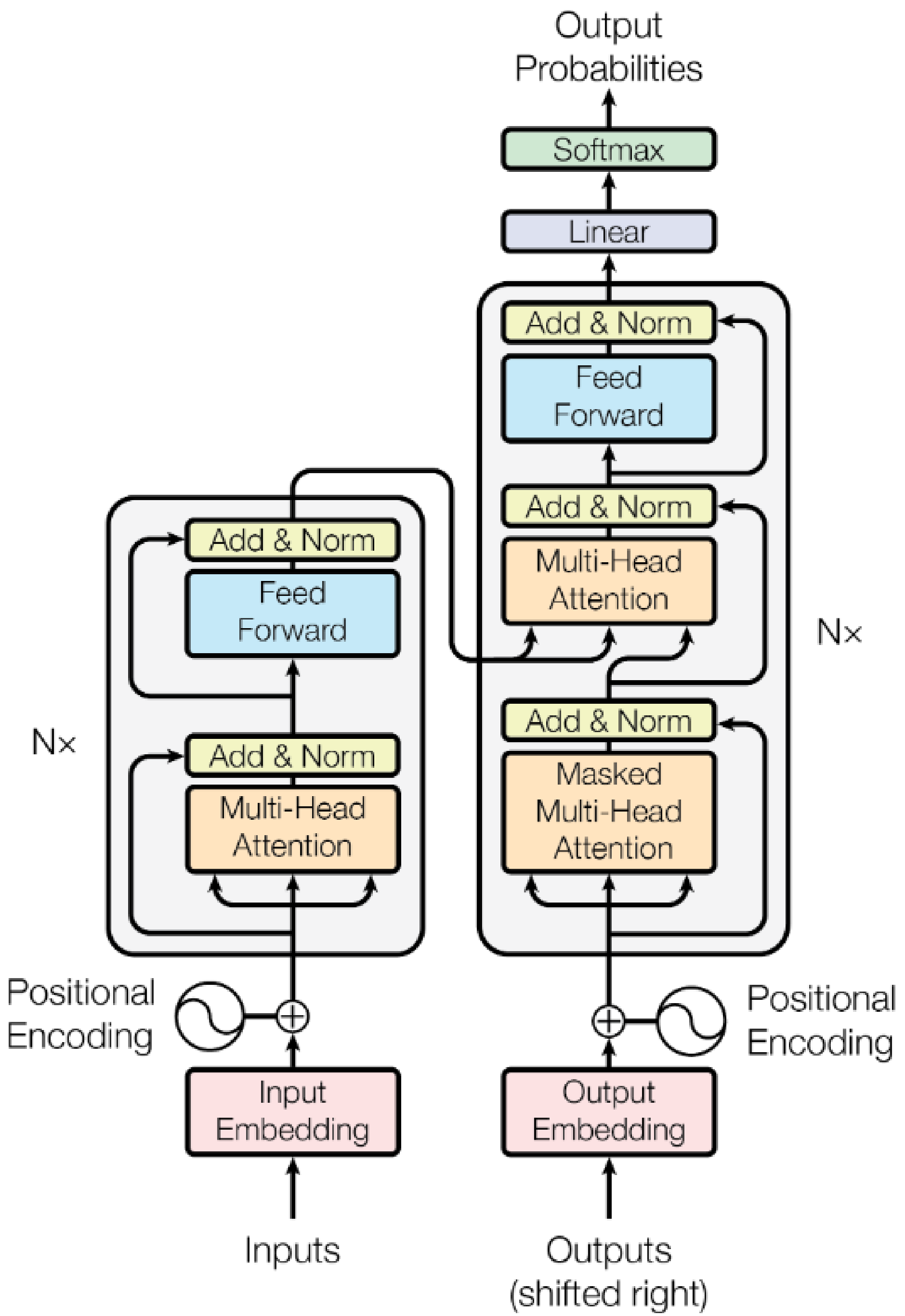
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Model Architecture

Attention

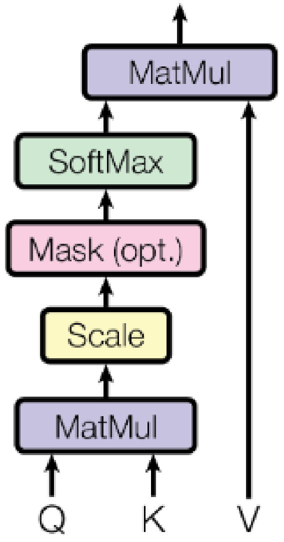
9/3/19





Transformer

Scaled Dot-Product Attention

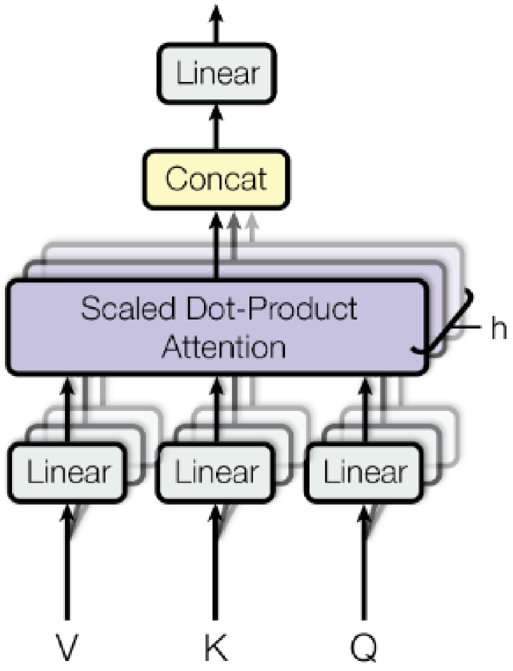


Quer
y

Ke
y

Valu
e

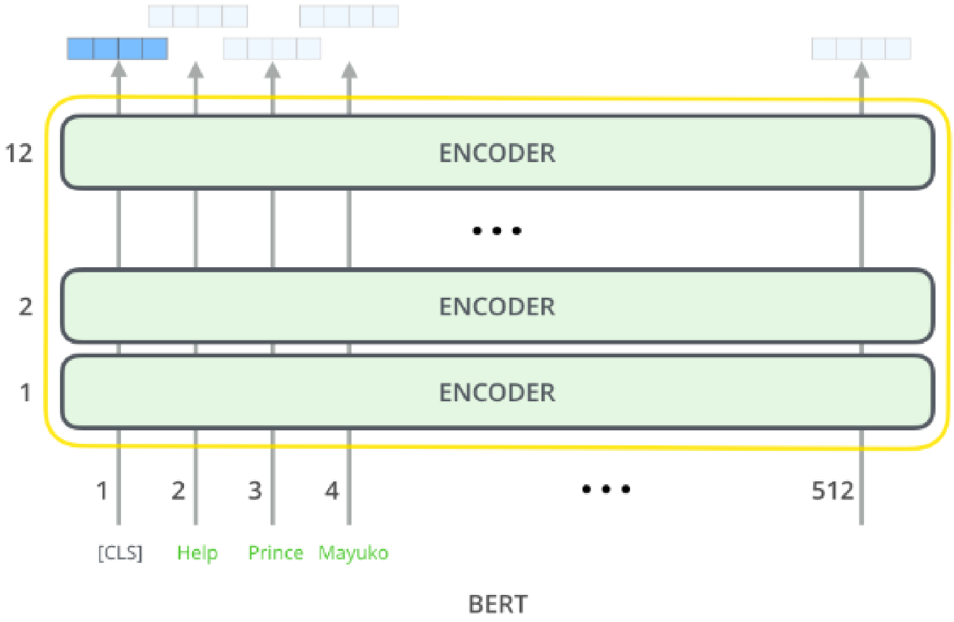
Multi-Head Attention



Bert Architecture

BERT-base

- 12 layers of Transformer encoders
- 768 hidden size
- 12 heads
- 110M parameters

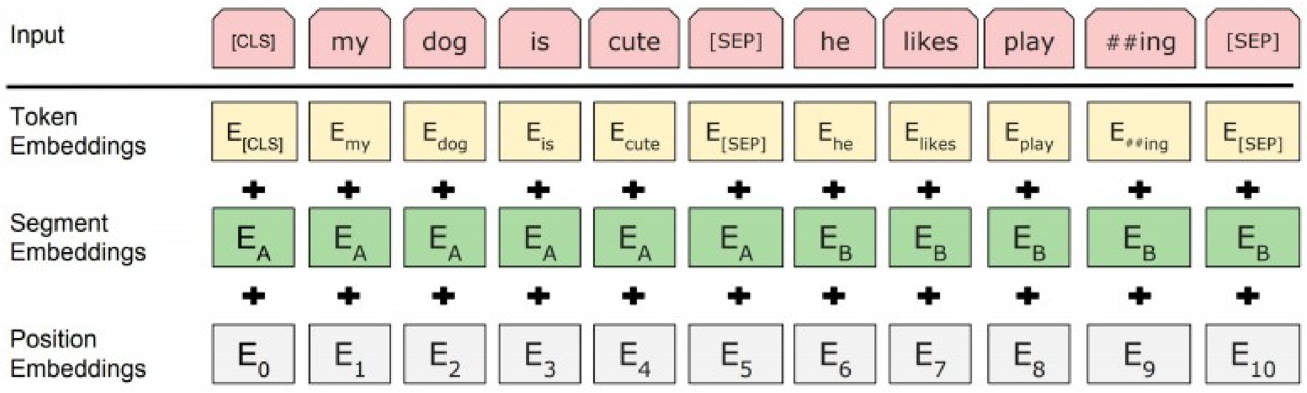


BERT-large

- 24 layers of Transformer encoders
- 1024 hidden size
- 16 heads
- 340M parameters

Input/Output

- 30,000 WordPiece vocabulary
- Token representation: sum of 3 embeddings
- 1 sequence is more efficient



WordPiece tokens

- similar to byte pair encoding (BPE), to solve out-of-vocabulary (OOV) problem
  - Instead "`the boy is playing`", the sentence is tokenized as "`the boy is play ##ing`"
  - "`play ##s`", "`play ##ing`" and "`play ##ed`" are better semantic and syntactic representations than "`plays`", "`playing`" and "`played`"

Experiments

GLUE

- 11 NLP tasks
- Text Entailment
- Sentiment Analysis
- Question Answering
- Question Duplication
- Paraphrase matching
- Pronoun Disambiguity
- Grammatical Acceptability

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	79.9	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**

4	GLUE Human Baselines	GLUE Human Baselines		87.1

MultiNLI

- Premise: At the end of Pennsylvania Ave, people began to line up for a White House tour.
- Hypothesis: People formed a line at the end of Pennsylvania Ave
- Label: {**entailment**, contradiction, neutral}

SQuAD

- Question Answering: Is the answer to a question found in a snippet? If yes where?
- BERT: Introduced "Start" and "End" vectors to determine where to start and stop the answer snippet
- Used the CLS token as the "no answer" token (if start and end had highest probability here)

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \qquad \hat{s_{i,j}} = \texttt{max}_{j \geq i} S \cdot T_i + E \cdot T_j \qquad \hat{s_{i,j}} > s_{\texttt{null}} + \tau$$

SQuAD

BERT Predictions

The atomic number of the periodic table for oxygen?
Ground Truth Answers: 8 8 8 8 8
Prediction: 8

What is the second most abundant element?
Ground Truth Answers: helium helium helium helium helium
Prediction: `<No Answer>`

Which gas makes up 20.8% of the Earth's atmosphere?
Ground Truth Answers: Diatomic oxygen | Diatomic oxygen | Diatomic oxygen gas | Diatomic oxygen | Diatomic oxygen gas
Prediction: Diatomic oxygen

How many atoms combine to form dioxygen?
Ground Truth Answers: two atoms | two | two | two | two
Prediction: two

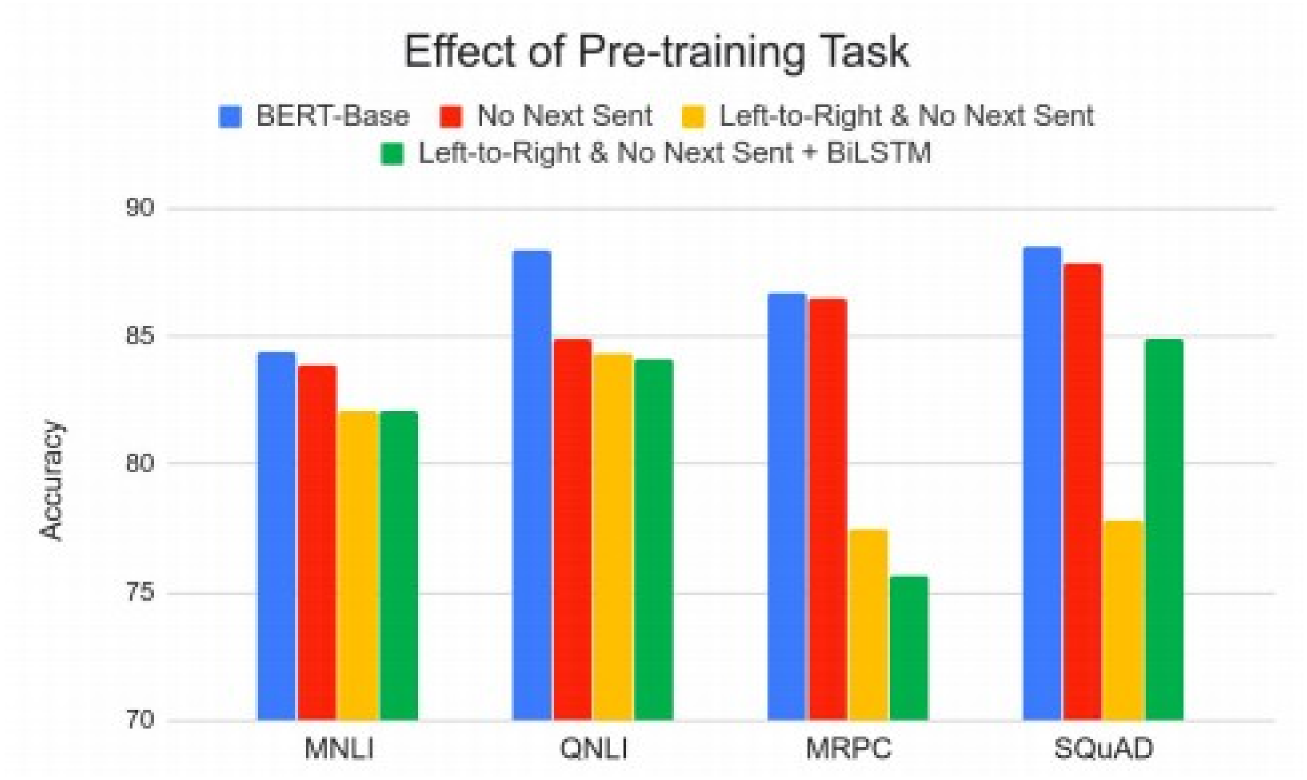
SQuAD Results

9/3/19

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**88.592**	**90.859**
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071
5 Jul 26, 2019	UPM (single model) *Anonymous*	87.193	89.934
6 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
6 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
7 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
8 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
9 May 21, 2019	XLNet (single model) *Google Brain & CMU*	86.346	89.133
10 May 14, 2019	SG-Net (ensemble) *Shanghai Jiao Tong University* https://arxiv.org/abs/1908.05147	86.211	88.848
10 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886

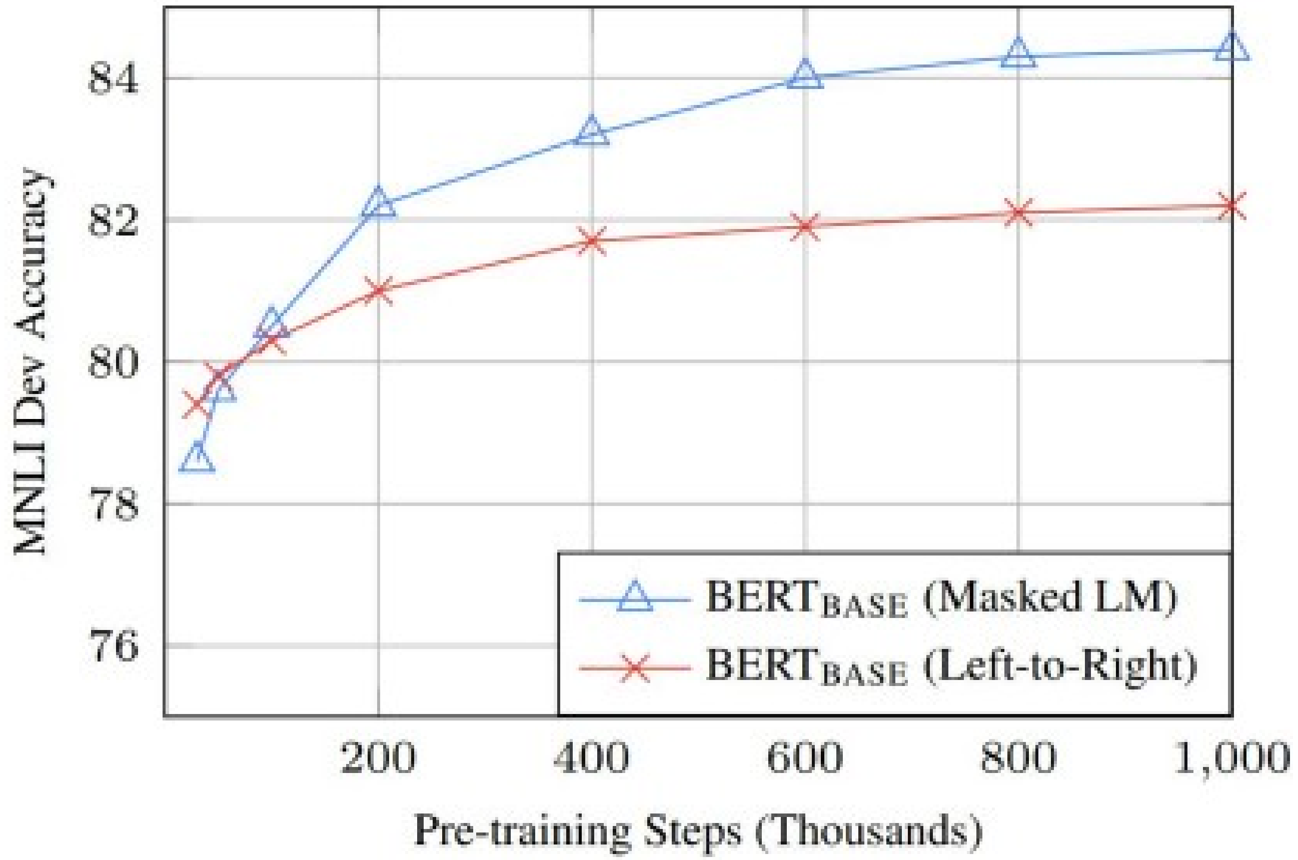
Architecture Effects

Pre-training Effects



- MLM performs much better on sentence semantic similarity (MRPC) and question answering (SQuAD) compared to Left-to-Right

- Only predicting 15% thus takes longer to converge
- Results much better with MLM



Other BERT cases

- Multilingual BERT (text entailment)

System	English	Chinese	Spanish
XNLI Baseline - Translate Train	73.7	67.0	68.8
XNLI Baseline - Translate Test	73.7	68.4	70.7
BERT - Translate Train	81.9	76.6	77.8
BERT - Translate Test	81.9	70.1	74.9
BERT - Zero Shot	81.9	63.8	74.3

  - Zero Shot (no machine translation)
- Domain Specific Text
  - ClinicalBERT

References

- arXiv Page: https://arxiv.org/abs/1810.04805
- GitHub Page: https://github.com/google-research/bert
- Jacob Devlin’s Seminar: https://nlp.stanford.edu/seminar/details/jdevlin.pdf
- Illustrated BERT: http://jalammar.github.io/illustrated-bert/
- Illustrated Transformer: http://jalammar.github.io/illustrated-transformer/
- GLUE Benchmark Leaderboard: https://gluebenchmark.com/leaderboard/
- SQuAD: https://rajpurkar.github.io/SQuAD-explorer/
- MultiLingual BERT: https://github.com/google-research/bert/blob/master/multilingual.md