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

(**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)

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

Google AI Language

Outline

- **Background & Motivation**
- BERT Architecture
- Pre-Training
- Experiments
- Summary & Conclusion
- Strengths & Weaknesses

- Questions
- Related Work

Unsupervised Pre-training

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Deep contextualized word representations

Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

Universal Language Model Fine-tuning for Text Classification

Jeremy Howard[*] fast.ai University of San Francisco j@fast.ai
Sebastian Ruder[*] Insight Centre, NUI Galway Aylien Ltd., Dublin sebastian@ruder.io

Semi-supervised Sequence Learning

Andrew M. Dai Google Inc. adai@google.com
Quoc V. Le Google Inc. qvl@google.com

Unsupervised Pre-training

Feature-based Approach

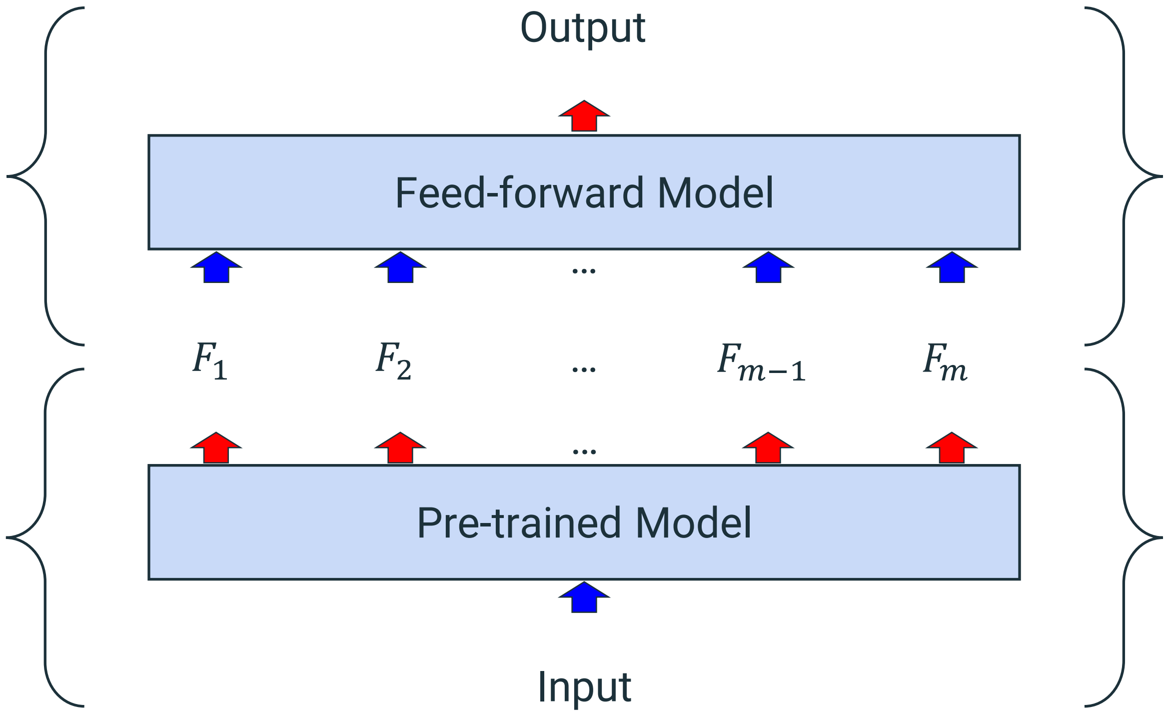
training **complex**
task-specific model

fixed
pre-trained model

Fine-tuning Approach

training **minimal**
task-specific model

fine-tuning
pre-trained model



Unidirectional vs. Bidirectional LM

$$\overbrace{\texttt{We went to the }\textbf{river}}^{\text{left context}}\ \underline{\textbf{bank}}.$$

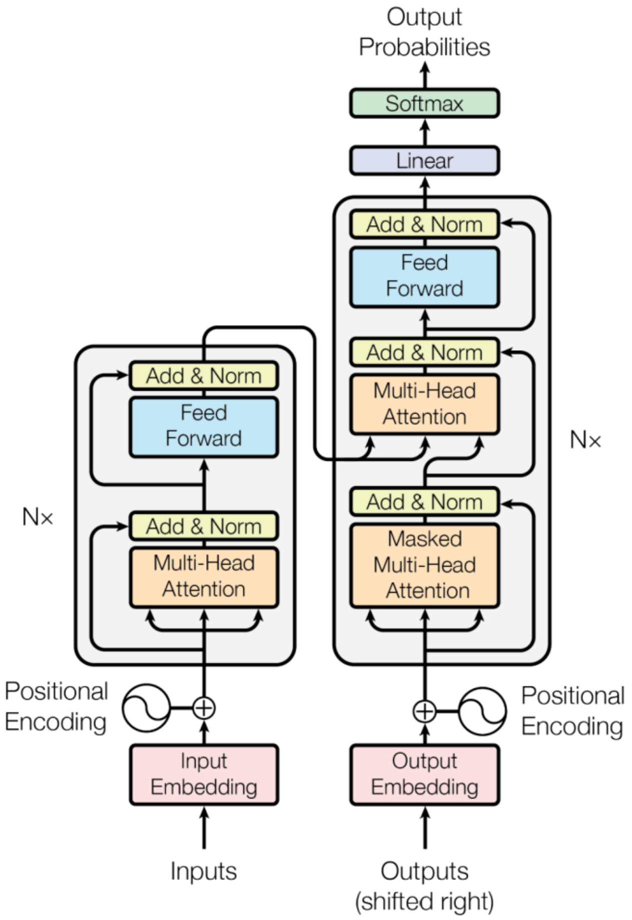
$$\texttt{We went to the }\underline{\textbf{bank}}\ \underbrace{\texttt{to make a }\textbf{deposit}.}_{\text{right context}}$$

Outline

- Background & Motivation
- **BERT Architecture**
- Pre-Training
- Experiments
- Summary & Conclusion
- Strengths & Weaknesses

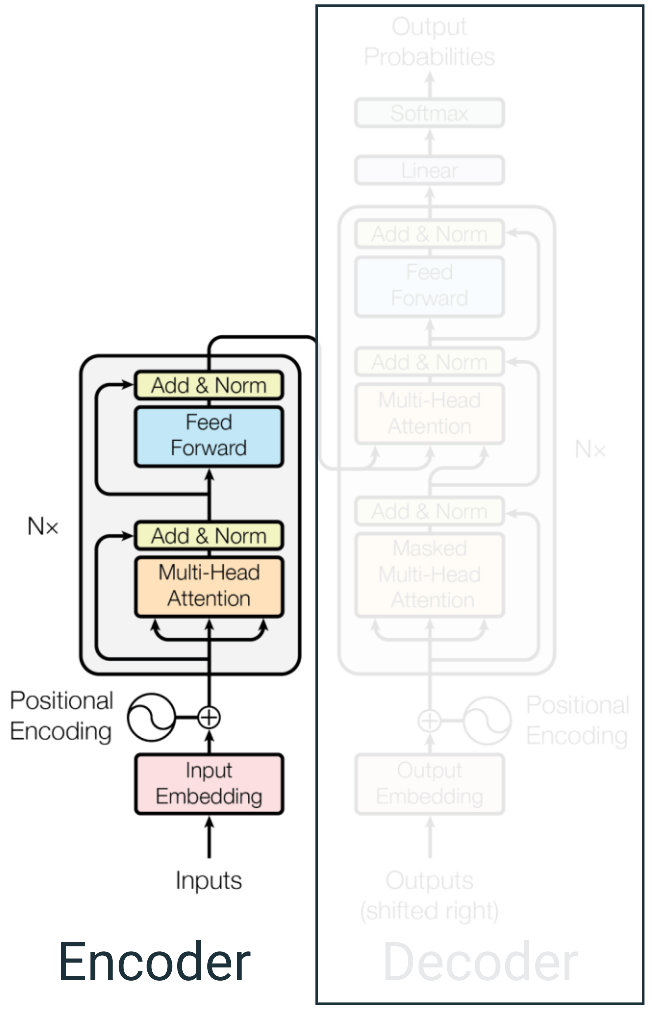
- Questions
- Related Work

Architecture

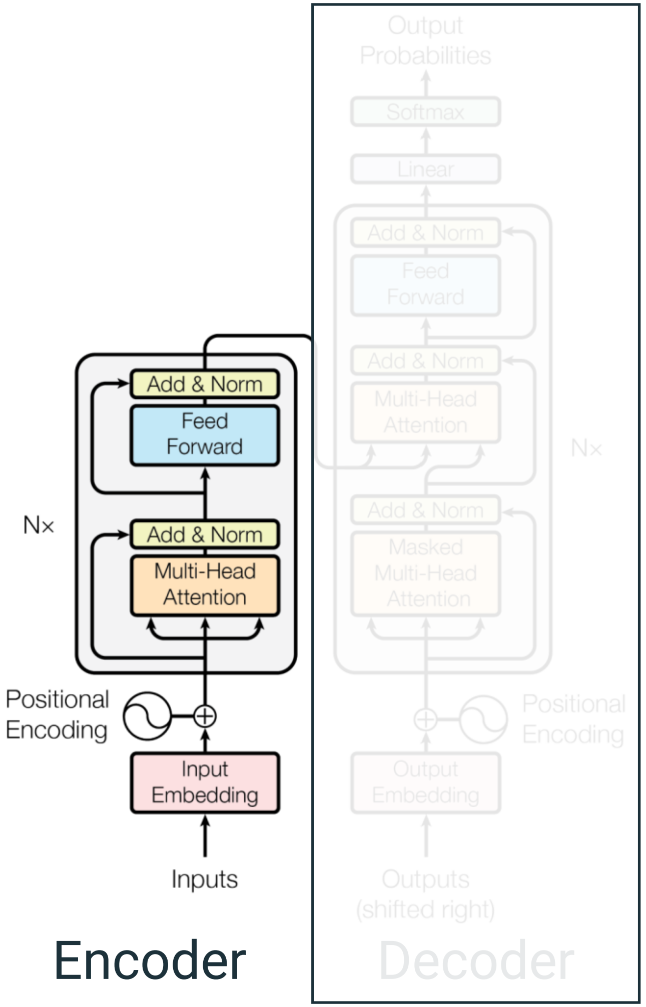


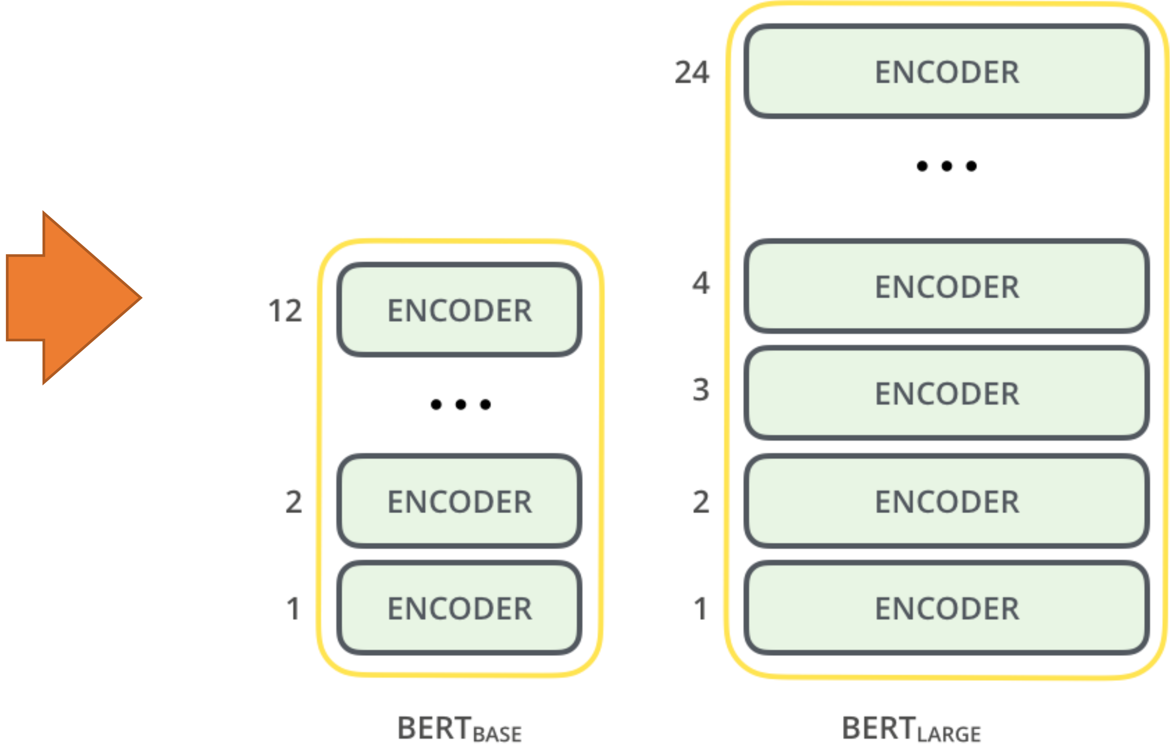
Encoder Decoder

Architecture



Architecture

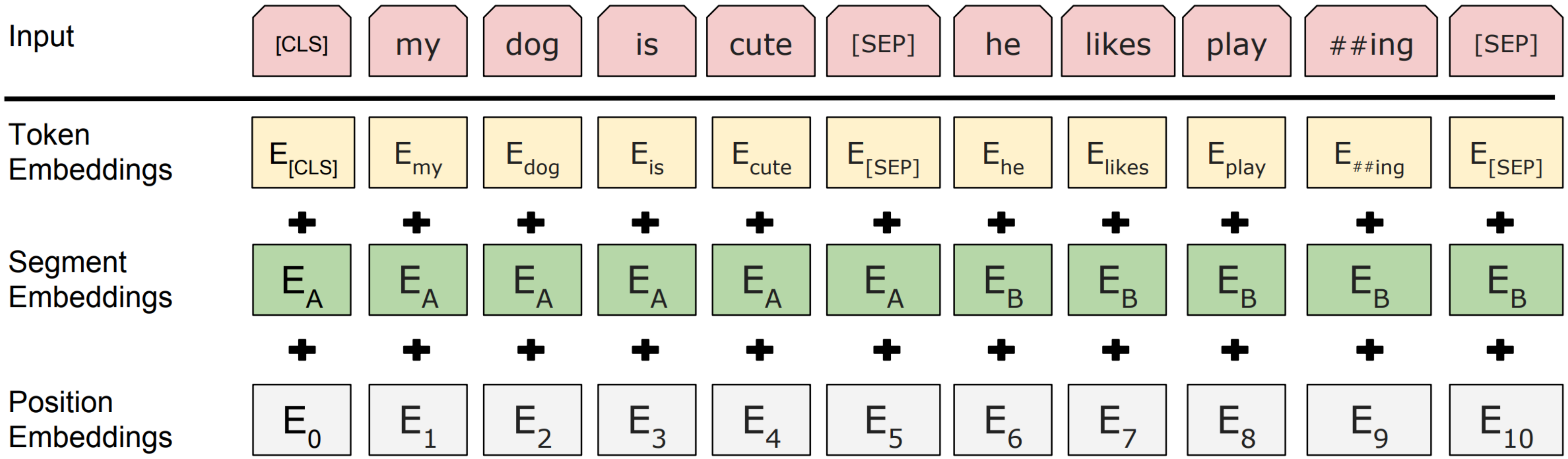




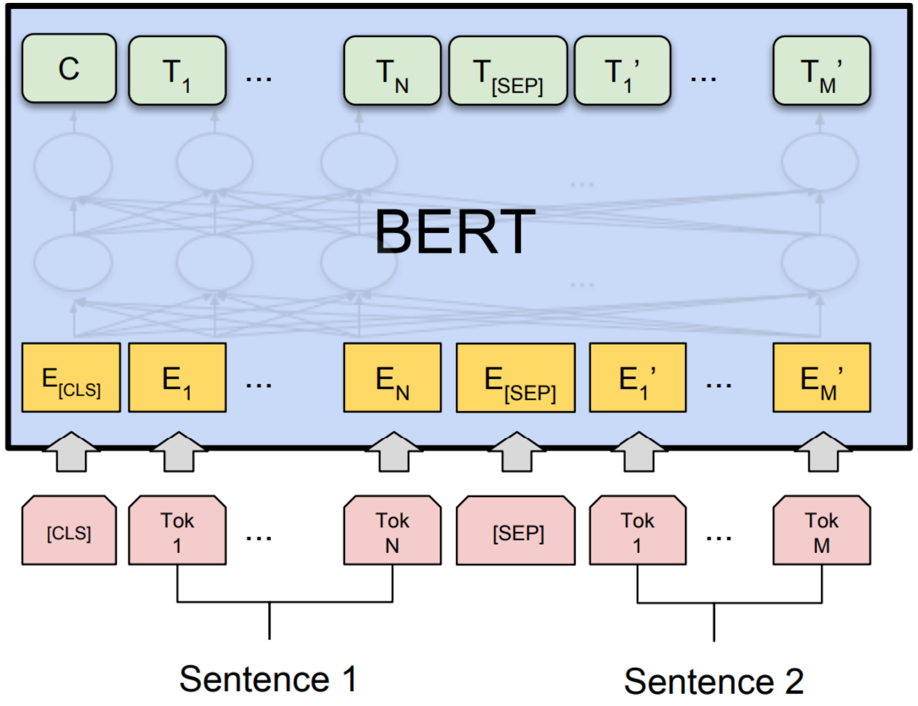
110M Parameters

340M Parameters

Input Representation



BERT Model


C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
BERT
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
Sentence 1
Sentence 2

Outline

- Background & Motivation
- BERT Architecture
- **Pre-Training**
- Experiments
- Summary & Conclusion
- Strengths & Weaknesses

- Questions
- Related Work

Left-to-Right Language Model (LTR LM)

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$\underbrace{\texttt{Alaska is about twelve time}}_{\text{left context}}\ \underline{\mathbf{\texttt{bigger}}}\ \underbrace{\texttt{than New York}}_{\text{masked}}$$

Masked Language Model (MLM)

$\overbrace{\texttt{Alaska is about twelve time}}^{\text{left context}}$ **bigger** $\overbrace{\texttt{than New York}}^{\text{right context}}$

80%: `Alaska is about twelve time` **[MASK]** `than New York`

10%: `Alaska is about twelve time` **apple** `than New York`

10%: `Alaska is about twelve time` **bigger** `than New York`

Next Sequence Prediction (NSP)

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]`

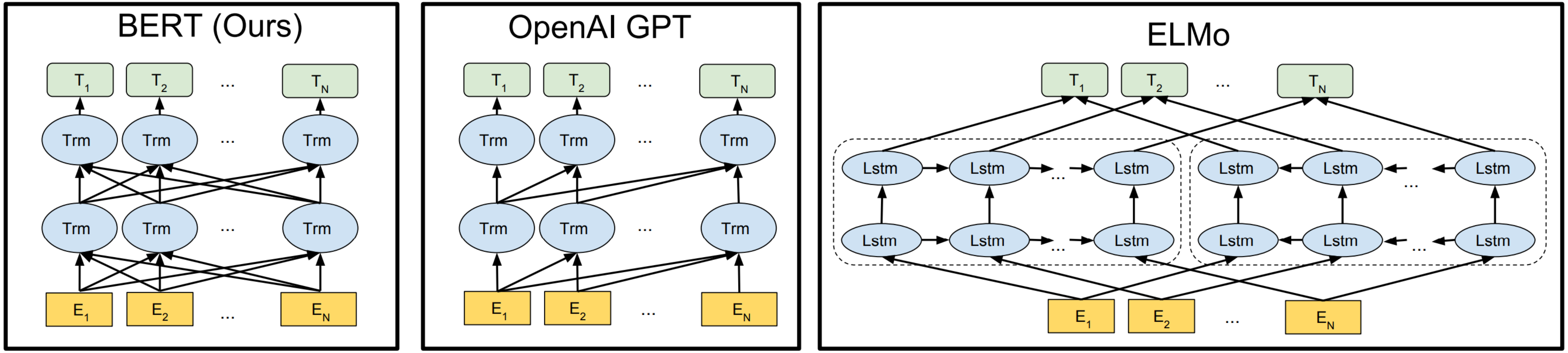
Label: **NotNext**

Outline

- Background & Motivation
- BERT Architecture
- Pre-Training
- **Experiments**
- Summary & Conclusion
- Strengths & Weaknesses

- Questions
- Related Work

BERT vs. GPT vs. ELMo



Multi-Genre Natural Language Inference (MNLI)

Sentence 1:

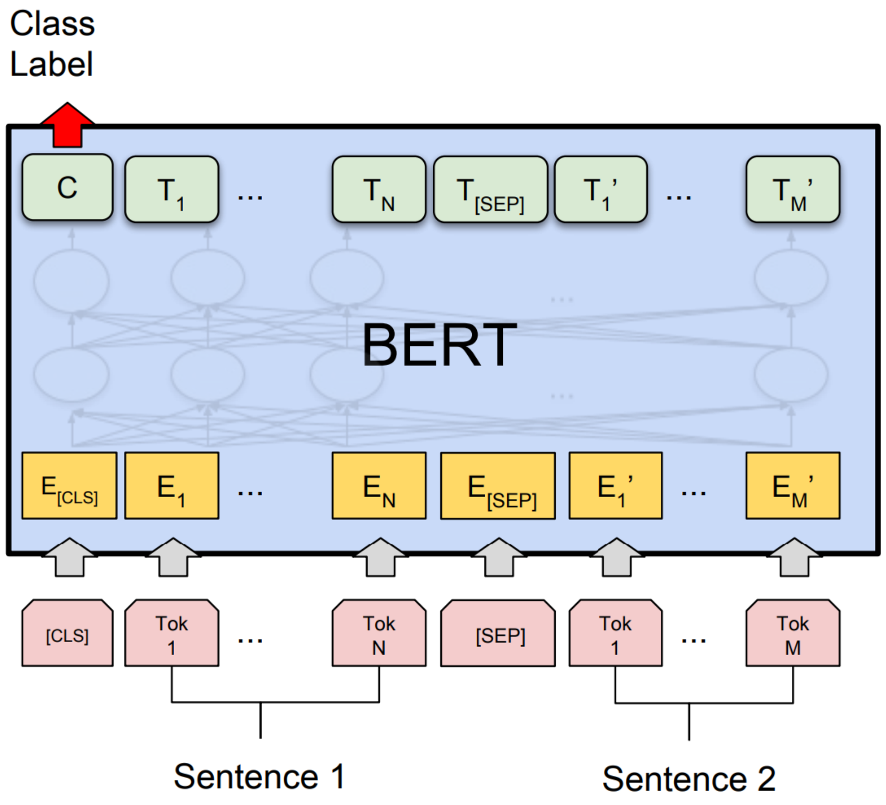
`At the other end of Pennsylvania Avenue, people began to line up for a White House tour.`

Sentence 2:

`People formed a line at the end of Pennsylvania Avenue.`

Label:

`contradiction/neutral/`**entailment**



Stanford Sentiment Treebank (SST-2)

Sentence:

`It's probably not easy to make such a worthless film …`

Label:

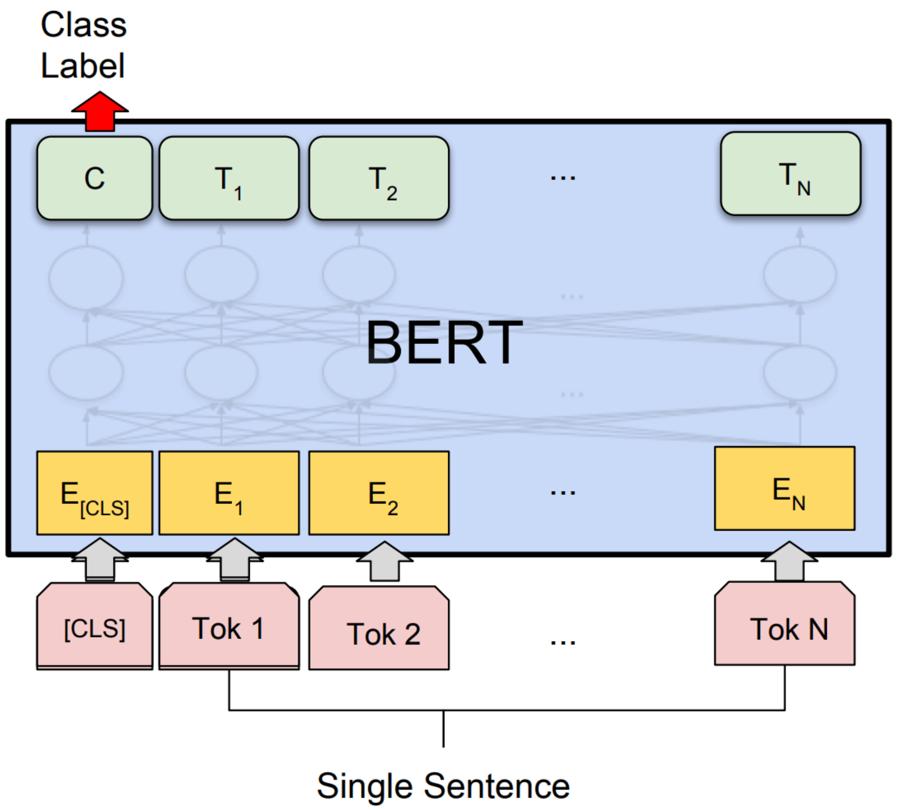
positive/**negative**

Sentence:

`Steven Spielberg brings us another masterpiece`

Label:

positive/negative



General Langauge Understanding Evaluation (GLUE)

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	79.8	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**

Stanford Question Answering Dataset (SQuAD v1.1)

Paragraph:

`… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals` **`within a cloud.`** `…`

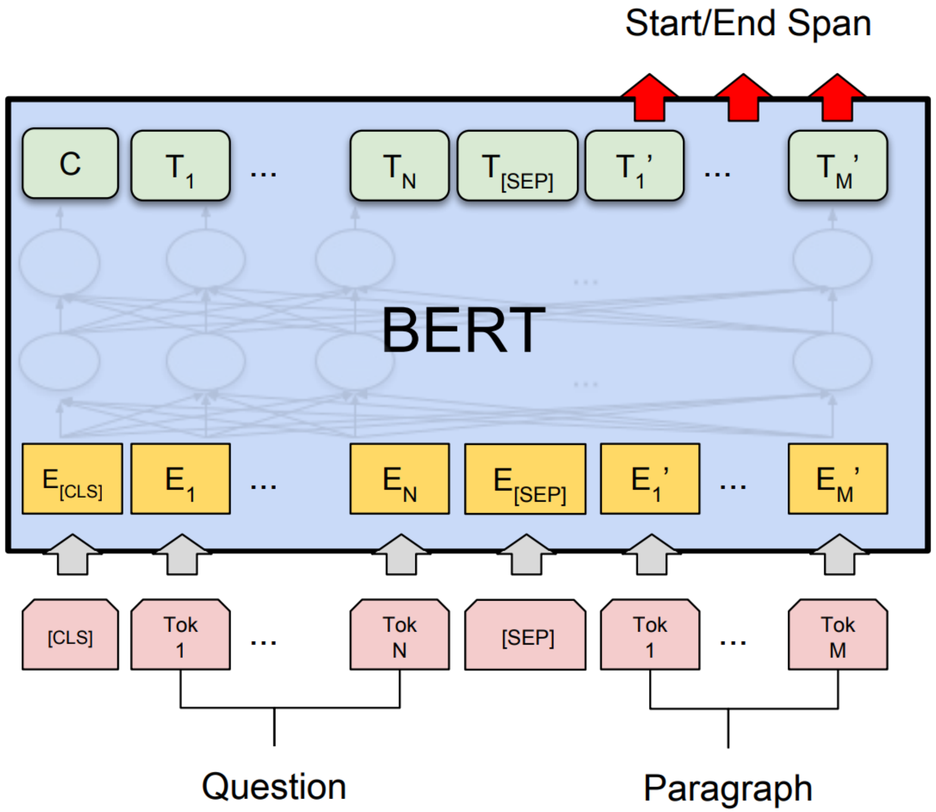
Question:

`Where do water droplets collide with ice crystals to form precipitation?`

Answer:

`within a cloud`

$$S \cdot T_i + E \cdot T_j$$



Stanford Question Answering Dataset (SQuAD v1.1)

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**

Stanford Question Answering Dataset (SQuAD v1.1)

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**

Stanford Question Answering Dataset (SQuAD v1.1)

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**

Stanford Question Answering Dataset (SQuAD v1.1)

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**

Situations With Adversarial Generations (SWAG)

Startphrase:

`On stage, a woman takes a seat at the piano. She …`

Endings:

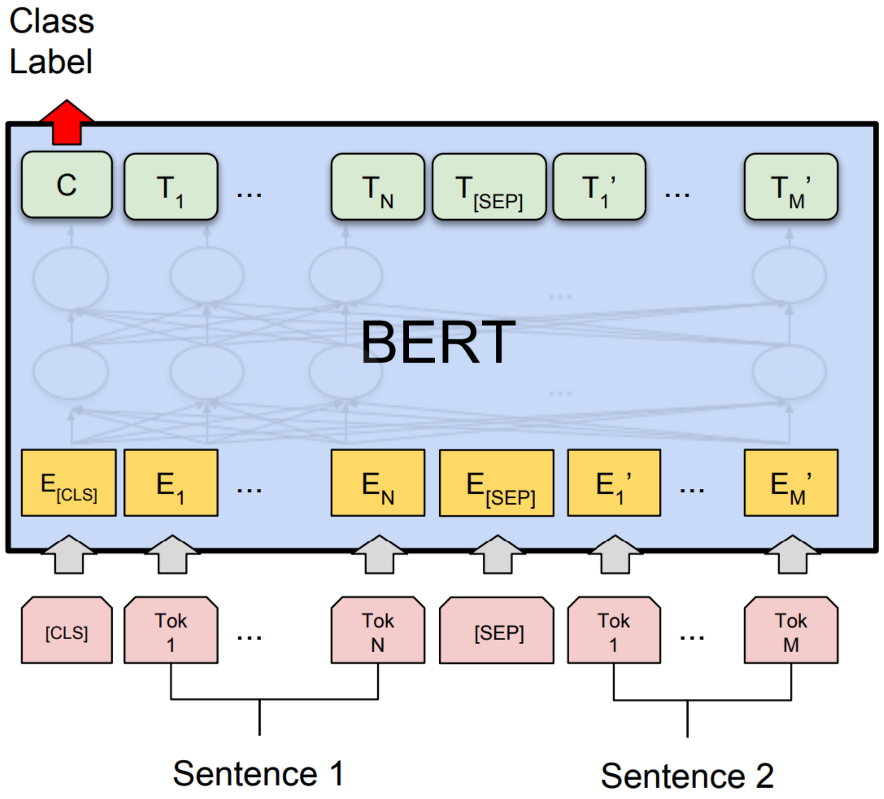
a) `sits on a bench as her sister plays with the doll.`

b) `smiles with someone as the music plays.`

c) `is in the crowd, watching the dancers.`

d) **<u>`nervously sets her fingers on the keys.`</u>**

4 x



Situations With Adversarial Generations (SWAG)

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
BERT$_{\text{BASE}}$	81.6	-
BERT$_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Ablation Study: Effect of Pre-Training

	Dev Set				
Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8

Ablation Study: Effect of Pre-Training

Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
	Dev Set				
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8

Ablation Study: Effect of Pre-Training

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8

Ablation Study: Model Size

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

#L: Number of Encoders

#H: Hidden Vector Size

#A: Number of Attention Heads

$\text{BERT}_{\text{BASE}}$

$\text{BERT}_{\text{LARGE}}$

Ablation Study: Feature-based

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
BERT$_{\text{LARGE}}$	96.6	92.8
BERT$_{\text{BASE}}$	96.4	92.4
Feature-based approach (BERT$_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Ablation Study: Feature-based

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Outline

- Background & Motivation
- BERT Architecture
- Pre-Training
- Experiments
- **Summary & Conclusion**
- Strengths & Weaknesses

- Questions
- Related Work

Summary & Conclusion

- **BERT** is proposed to overcome the limitation of unidirectional LMs
- **Masked LM** is introduced for bidirectional pre-training
- **NSP** is introduced to enable BERT to understand the relationship between sentences

- BERT advances the state-of-the-art for eleven NLP tasks
- Bidirectional LMs are more powerful than left-to-right LMs
- Task-specific models can benefit from larger more expressive pre-trained representation
- BERT can also be used in a feature-based approach

Outline

- Background & Motivation
- BERT Architecture
- Pre-Training
- Experiments
- Summary & Conclusion
- **Strengths & Weaknesses**

- Questions
- Related Work

Experiments

Strengths

- BERT was evaluated on many different NLP tasks
- $\text{BERT}_{\text{BASE}}$ has the same model size as GPT
- Evaluated the effects of their pre-training methods
- Clear description of the NLP tasks and the task-specific models

Weaknesses

- Often only the results of the dev set instead of the test set were used
- No comparison with a transformer-based model using left-to-right and right-to-left LMs.

BERT

Strengths

- Achieves better results than previous state-of-the-art methods
- Parallelizable architecture
- Fast fine-tuning (2-4 epochs)
- Minimal additional task-specific parameters are required
- Suitable for many different NLP tasks

Weaknesses

- Resource and time intensive pre-training (slower convergence than left-to-right pre-training)
- For small datasets sometimes fine-tuning is unstable
- Lack of ability to handle long text sequences (max. 512 tokens)

Questions?

Outline

- Background & Motivation
- BERT Architecture
- Pre-Training
- Experiments
- Summary & Conclusion
- Strengths & Weaknesses

- Questions
- **Related Work**

Domain Specific Pre-training

BioBERT: a pre-trained biomedical language representation model for biomedical text mining

Jinhyuk Lee [1,†], Wonjin Yoon [1,†], Sungdong Kim [2], Donghyeon Kim [1], Sunkyu Kim [1], Chan Ho So [3] and Jaewoo Kang [1,3,*]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea, [2]Clova AI Research, Naver Corp, Seong-Nam 13561, Korea and [3]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, Korea

ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission

Kexin Huang
Health Data Science, Harvard T.H. Chan School of Public Health

Jaan Altosaar
Department of Physics, Princeton University

Rajesh Ranganath
Courant Institute of Mathematical Science, New York University

SciBERT: A Pretrained Language Model for Scientific Text

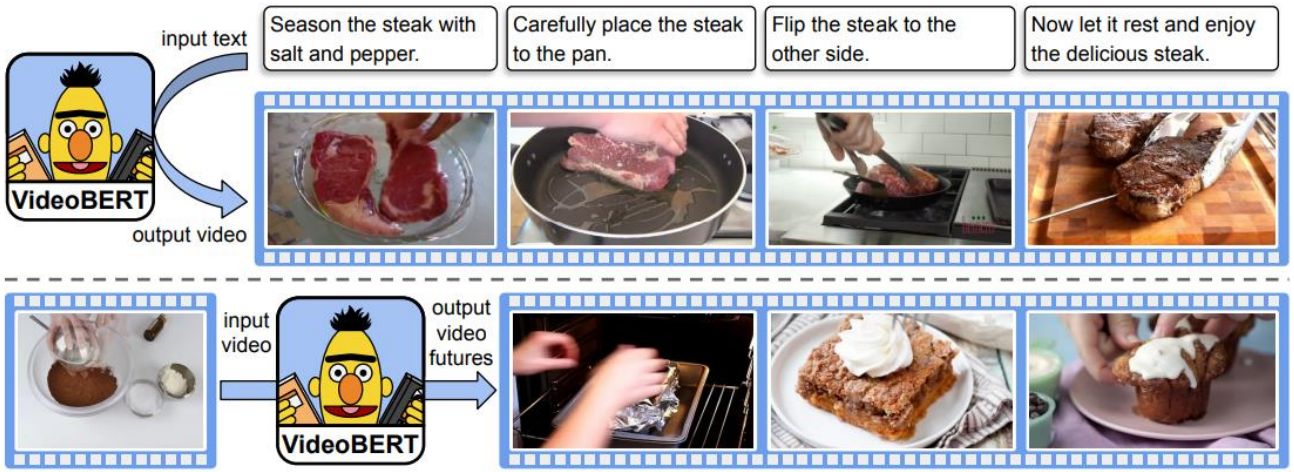
Iz Beltagy   Kyle Lo   Arman Cohan
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,kylel,armanc}@allenai.org

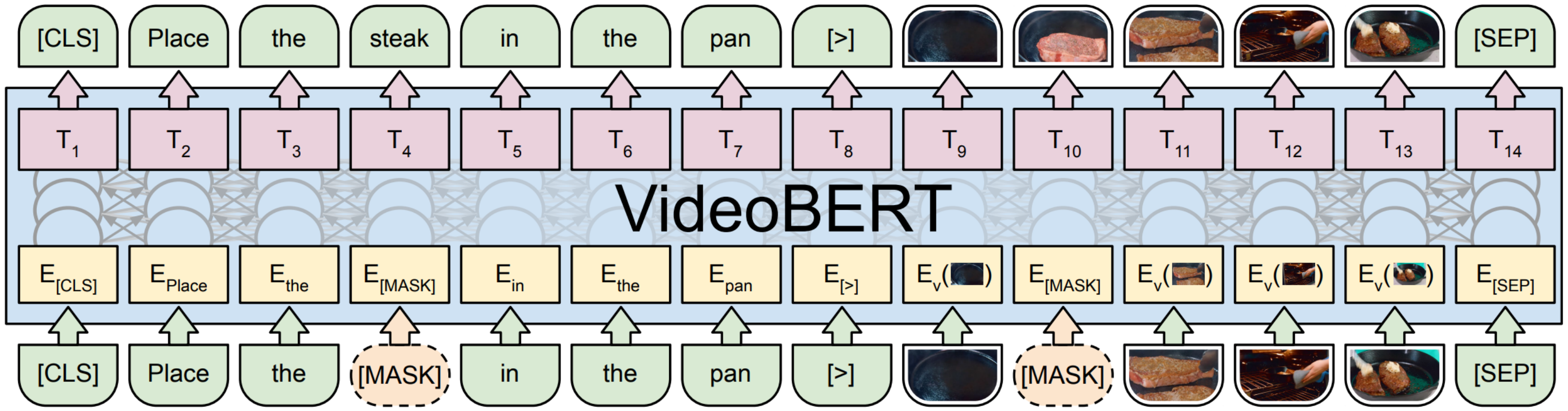
Multilingual

How multilingual is Multilingual BERT?

Telmo Pires* Eva Schlinger Dan Garrette
Google Research
{telmop,eschling,dhgarrette}@google.com

VideoBERT


input text
Season the steak with salt and pepper.
Carefully place the steak to the pan.
Flip the steak to the other side.
Now let it rest and enjoy the delicious steak.
VideoBERT
output video
input video
VideoBERT
output video futures


[CLS] Place the steak in the pan [>] [SEP]
T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10} T_{11} T_{12} T_{13} T_{14}
VideoBERT
$E_{[CLS]}$ E_{Place} E_{the} $E_{[MASK]}$ E_{in} E_{the} E_{pan} $E_{[>]}$ E_v( ) $E_{[MASK]}$ E_v( ) E_v( ) E_v( ) $E_{[SEP]}$
[CLS] Place the [MASK] in the pan [>] [MASK] [SEP]

Distilliation

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF
Hugging Face
{victor,lysandre,julien,thomas}@huggingface.co

RoBERTa

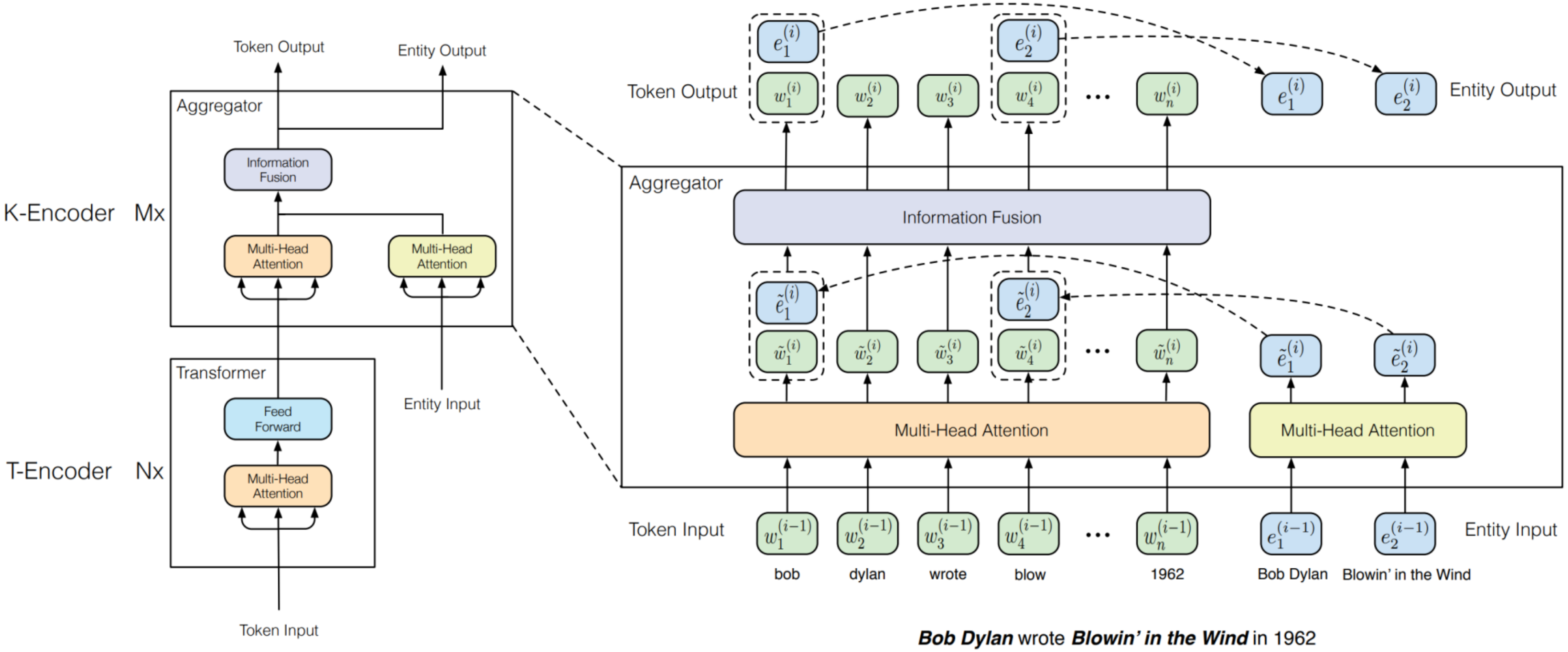
RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu[*§] **Myle Ott**[*§] **Naman Goyal**[*§] **Jingfei Du**[*§] **Mandar Joshi**[†]
Danqi Chen[§] **Omer Levy**[§] **Mike Lewis**[§] **Luke Zettlemoyer**[†§] **Veselin Stoyanov**[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

[§] Facebook AI
{yinhanliu,myleott,naman,jingfeidu,
danqi,omerlevy,mikelewis,lsz,ves}@fb.com

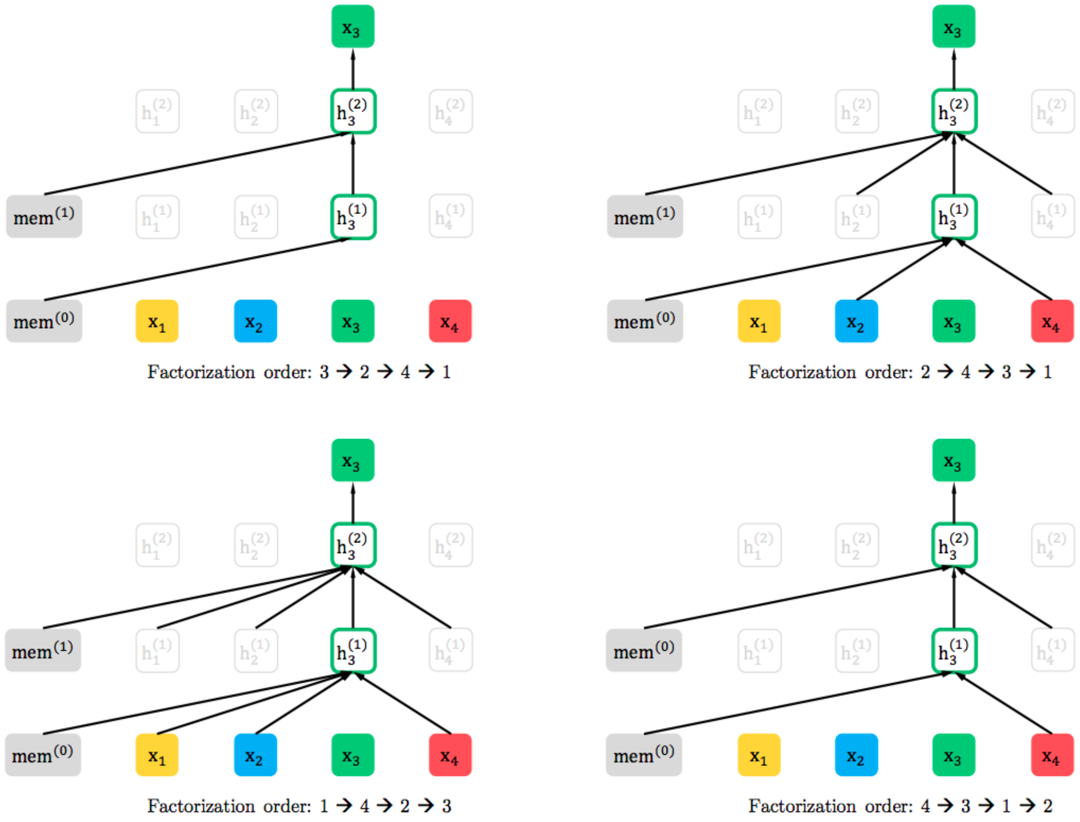
ERNIE



(a) Model Achitecture

(b) Aggregator

XLNet



XLNet: Generalized Autoregressive Pretraining for Language Understanding

Zhilin Yang[*1], **Zihang Dai**[*12], **Yiming Yang**[1], **Jaime Carbonell**[1],
Ruslan Salakhutdinov[1], **Quoc V. Le**[2]
[1]Carnegie Mellon University, [2]Google AI Brain Team
{zhiliny,dzihang,yiming,jgc,rsalakhu}@cs.cmu.edu, qvl@google.com

Additional Slides

Stanford Question Answering Dataset (SQuAD v2.0)

Paragraph:

`… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals` **`within a cloud.`** `…`

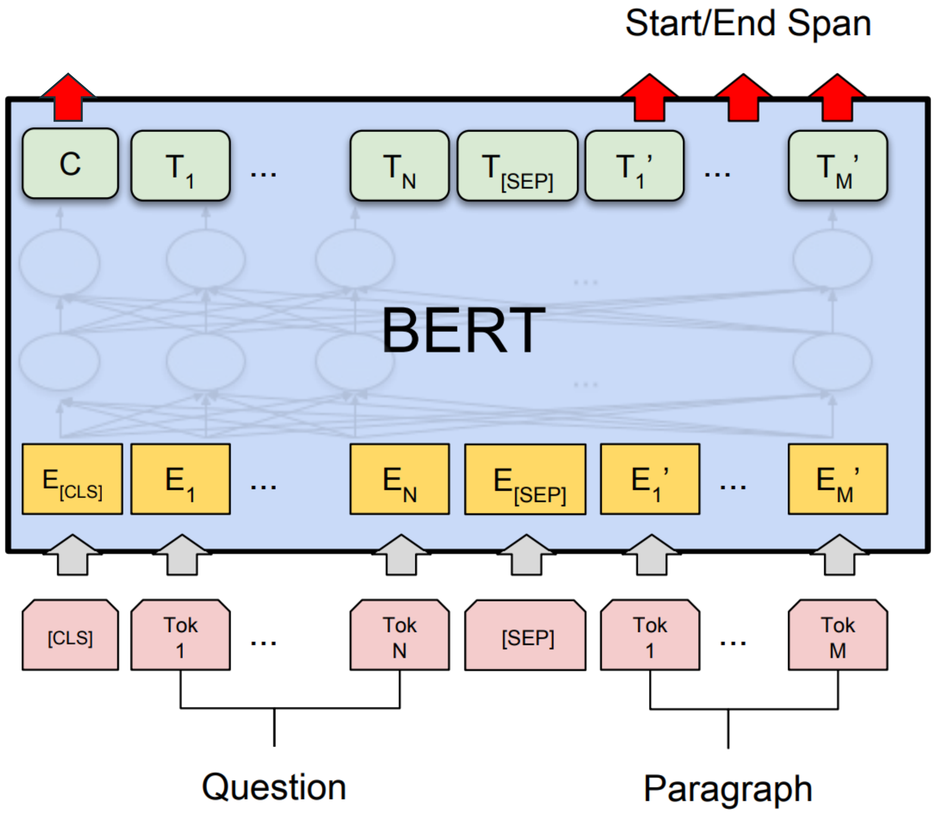
Question:

`Where do water droplets collide with ice crystals to form precipitation?`

Answer:

`within a cloud`

$$S \cdot C + E \cdot C < max_{i \leq j} S \cdot T_i + E \cdot T_j$$



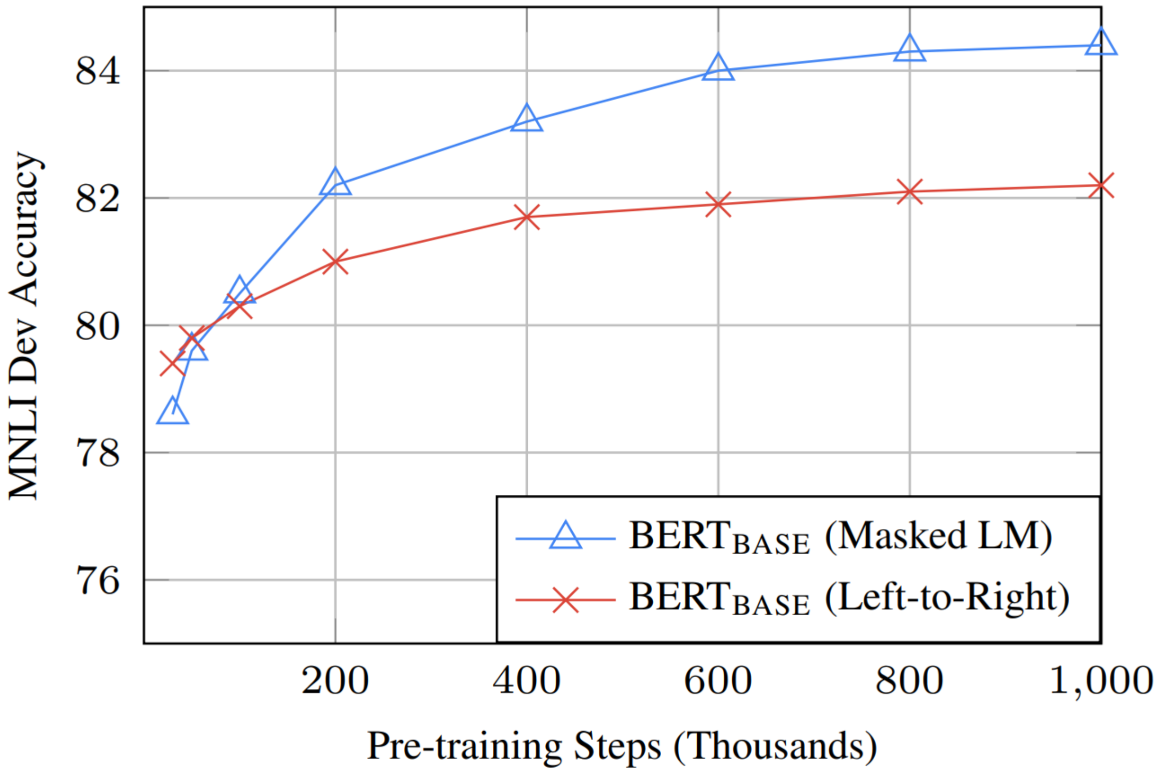
Stanford Question Answering Dataset (SQuAD v2.0)

System	Dev EM	Dev F1	Test EM	Test 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

Ablation Study: Masking Strategies

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

Ablation Study: MLM vs. LTR LM



BERT Limitations

Commonsense Reasoning

Sentence:

`The trophy doesn't fit in the suitcase because `**`it`**` is too small.`

Answer:

`the trophy / `**`the suitcase`**

Attention Is (not) All You Need for Commonsense Reasoning

Tassilo Klein[1], Moin Nabi[1]
[1]SAP Machine Learning Research, Berlin, Germany
{tassilo.klein, m.nabi}@sap.com

***HellaSwag*: Can a Machine *Really* Finish Your Sentence?**

Rowan Zellers[♠] Ari Holtzman[♠] Yonatan Bisk[♠] Ali Farhadi[♠♡] Yejin Choi[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
https://rowanzellers.com/hellaswag

Long Texts

CogLTX: Applying BERT to Long Texts

Ming Ding
Tsinghua University
dm18@mails.tsinghua.edu.cn

Chang Zhou
Alibaba Group
ericzhou.zc@alibaba-inc.com

Hongxia Yang
Alibaba Group
yang.yhx@alibaba-inc.com

Jie Tang
Tsinghua University
jietang@tsinghua.edu.cn

What BERT is not

Context	$\text{BERT}_{\text{LARGE}}$ predictions
Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her ____	*car, house, room, truck, apartment*
The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a ____	*note, letter, gun, blanket, newspaper*
At the zoo, my sister asked if they painted the black and white stripes on the animal. I explained to her that they were natural features of a ____	*cat, person, human, bird, species*