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

2018

Panchuk Georgy

HSE Research Seminar

2022

Plan

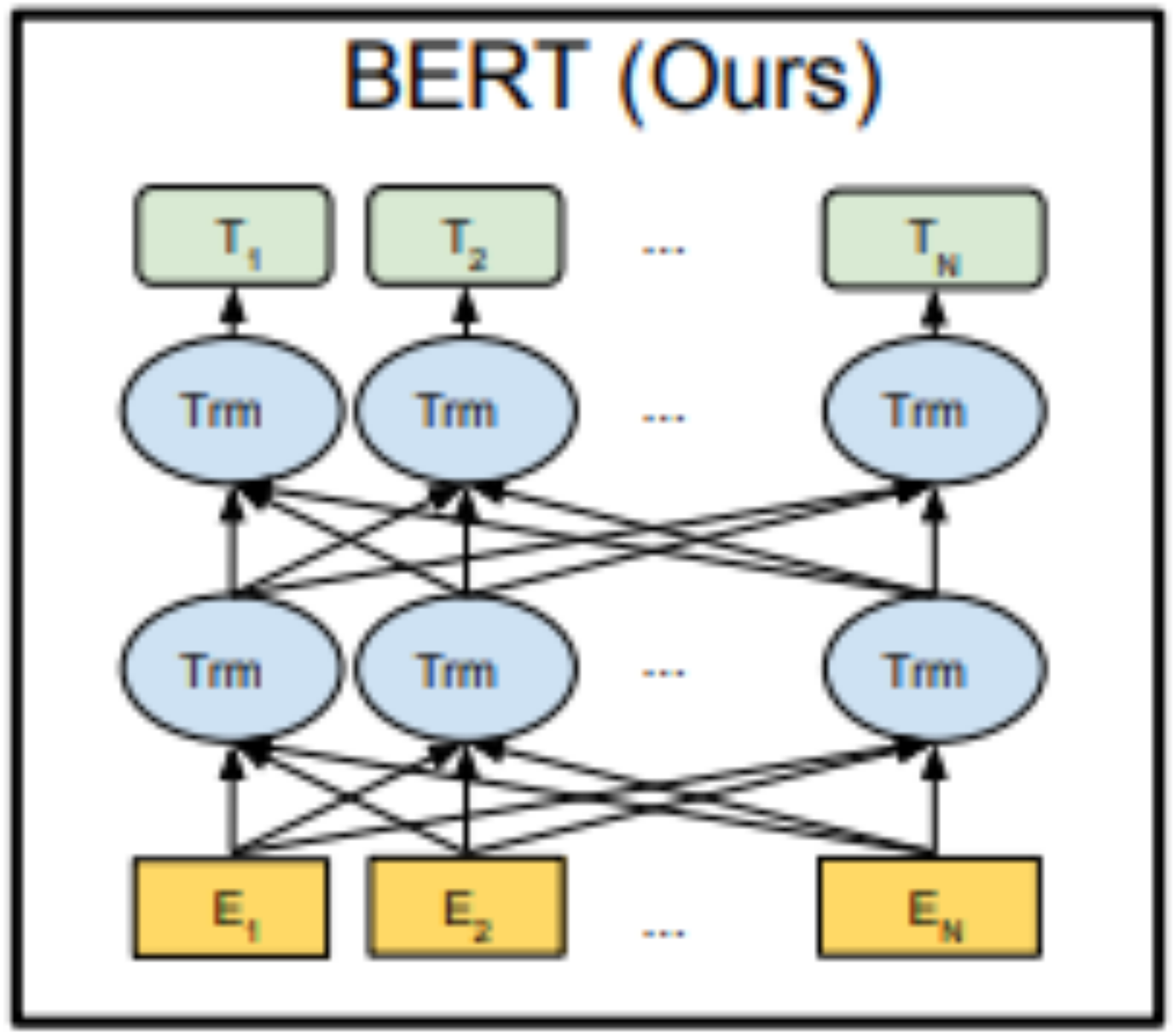
- Introduction
- Recap: NLP approaches
- Architecture
- Pre-training & Fine-tuning
- Ablation studies
- Feature-based
- Questions

NLP approaches

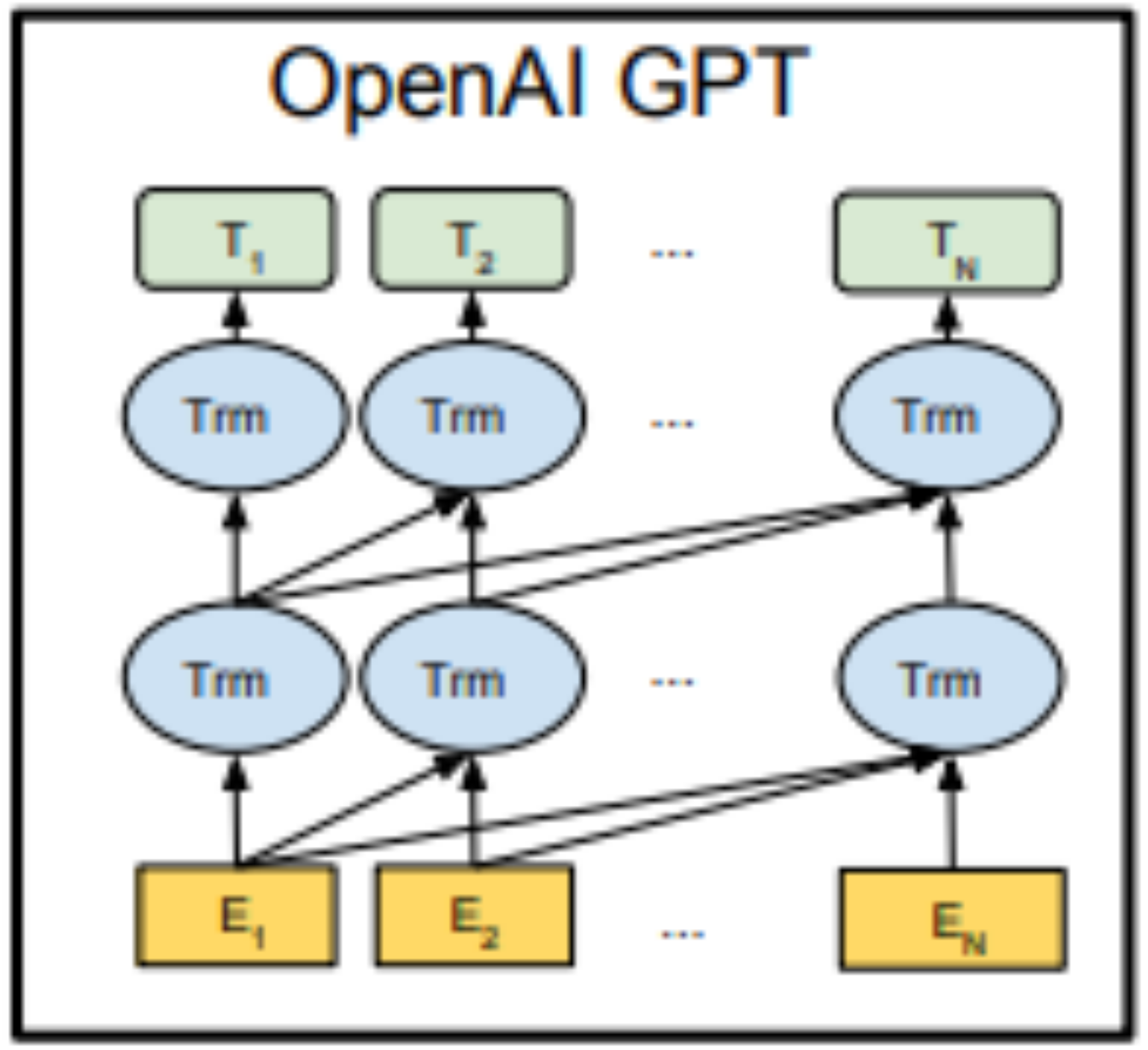
- BoW, TF-IDF
- Word2Vec, GloVe
- CoVe, ELMo
- GPT, BERT

Bidirectional vs unidirectional

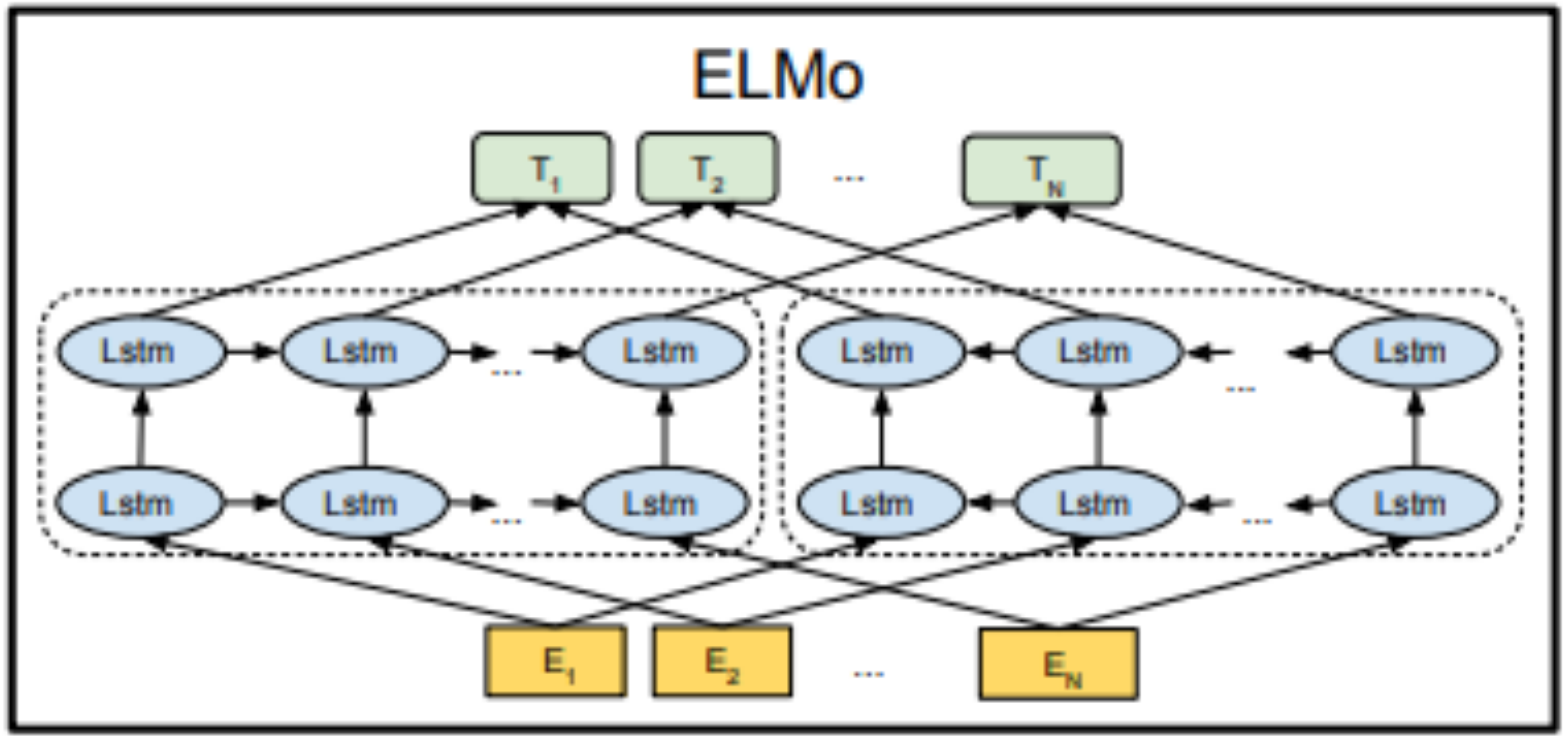
Bidirectional Transformer



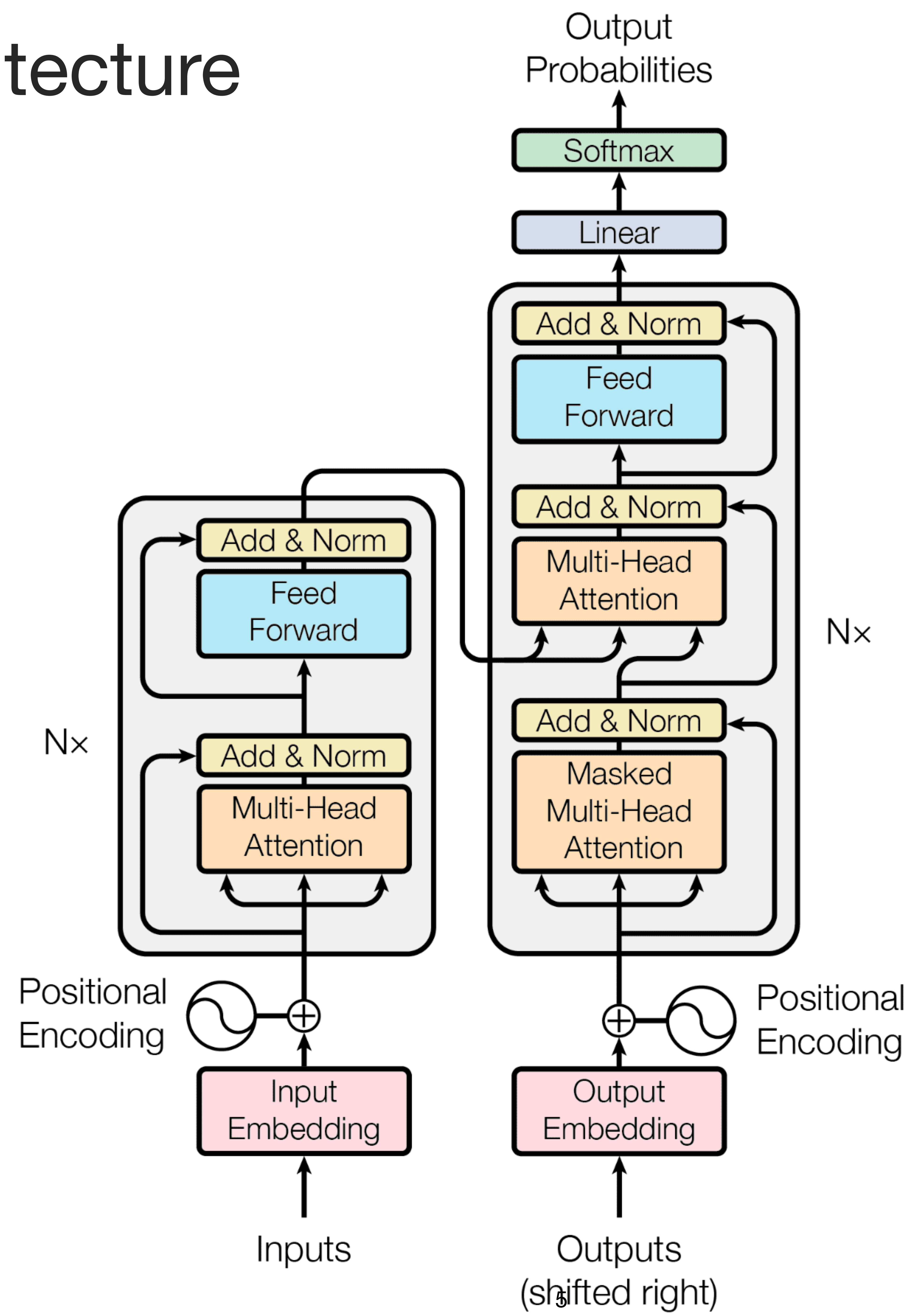
Left-to-right transformer



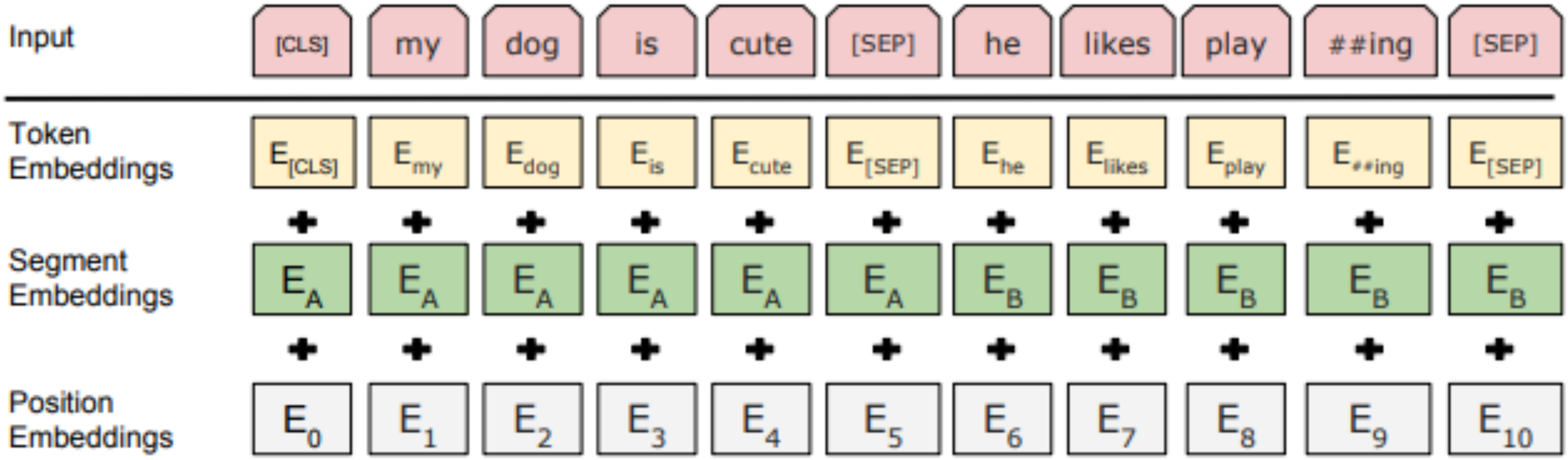
Concatenation of independently trained left-to-right and right-to-left LSTM



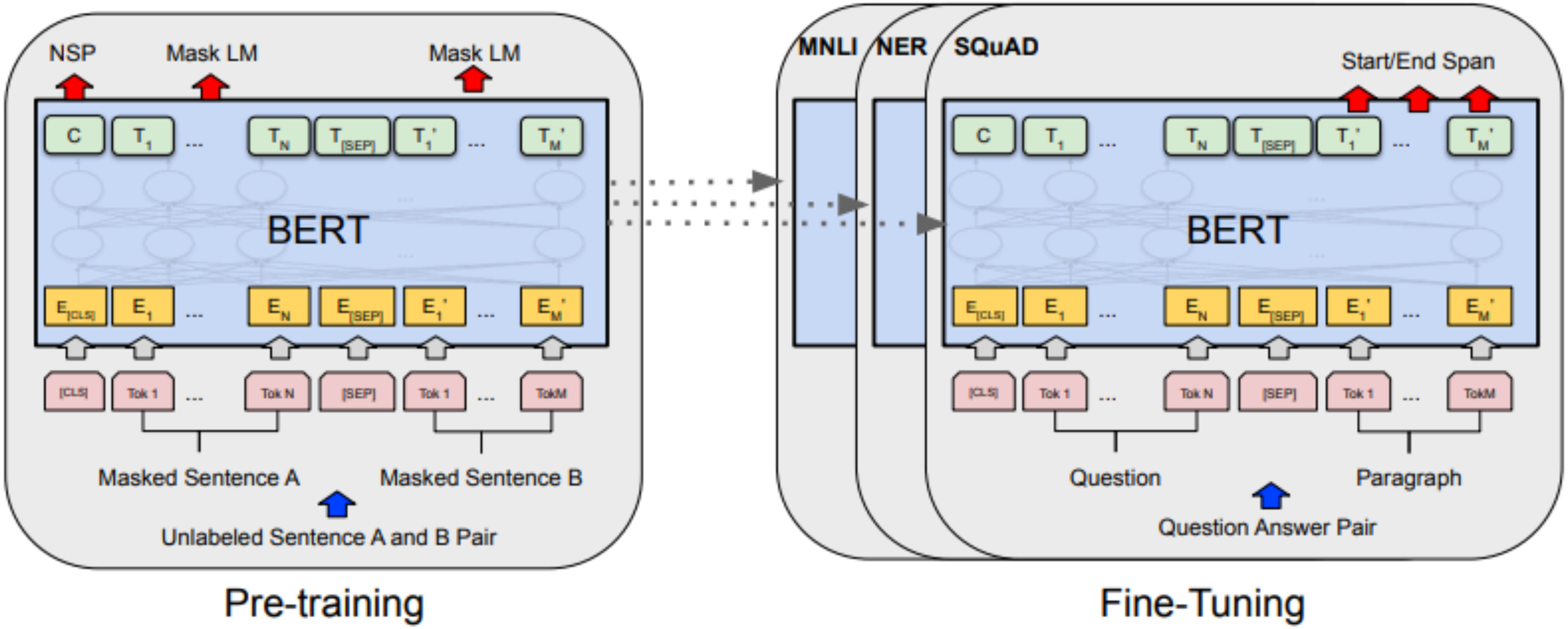
Transformer architecture



Input representation



Pre-training & Fine-tuning



Pre-training. Masked language modelling

1. 80% of the time:
Replace the word with the [MASK] token, e.g.:
George is telling you about BERT -> [MASK] is telling you about BERT

2. 10% of the time: Replace the word with a random word, e.g.:
George is telling you about BERT -> Crocodile is telling you about BERT

3. 10% of the time: Keep the word unchanged, e.g.:
George is telling you about Bert -> George is telling you about BERT
(The purpose of this is to bias the representation towards the actual observed word)

Pre-training. Next sentence prediction

Input:

[CLS] the man wen 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

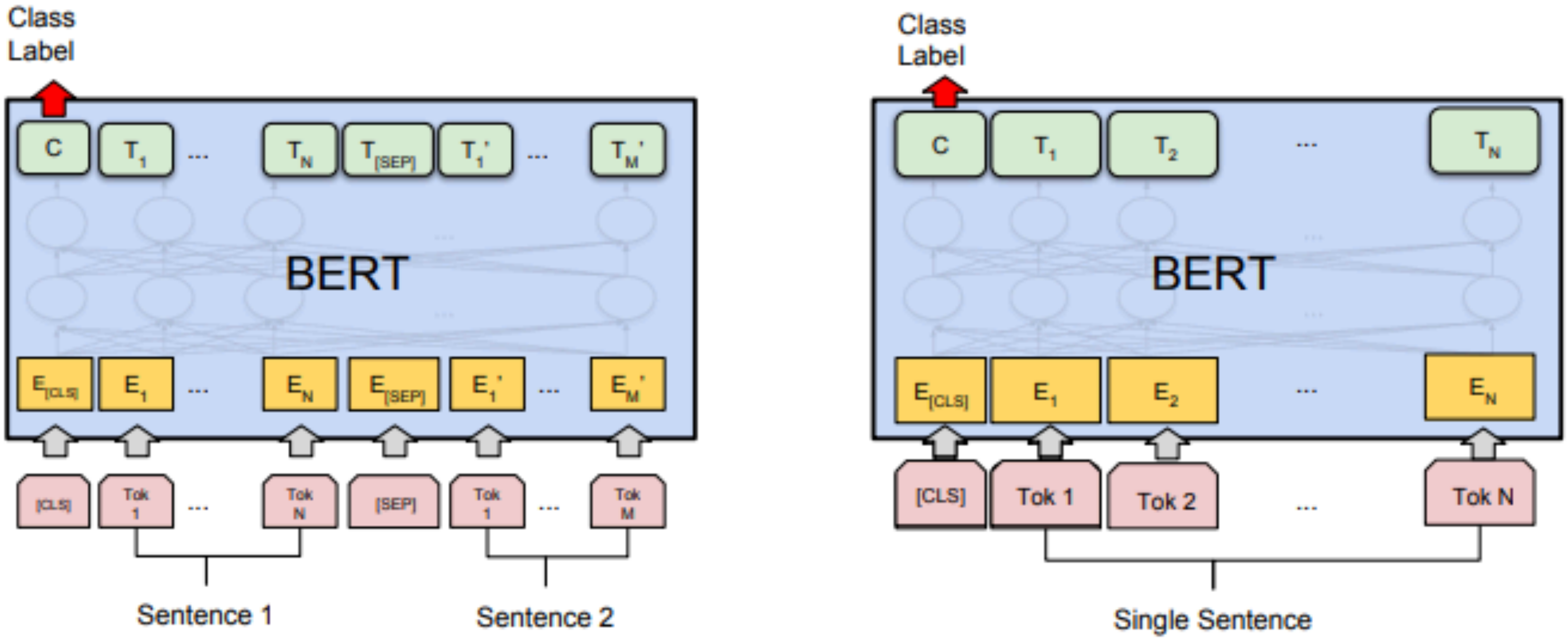
Pre-training data

1. BooksCorpus (800M words)

2. English Wikipedia (2500M words)

It is crucial to use a document-level corpus
rather than a shuffled sentence-level corpus
in order to extract long contiguous sequences.

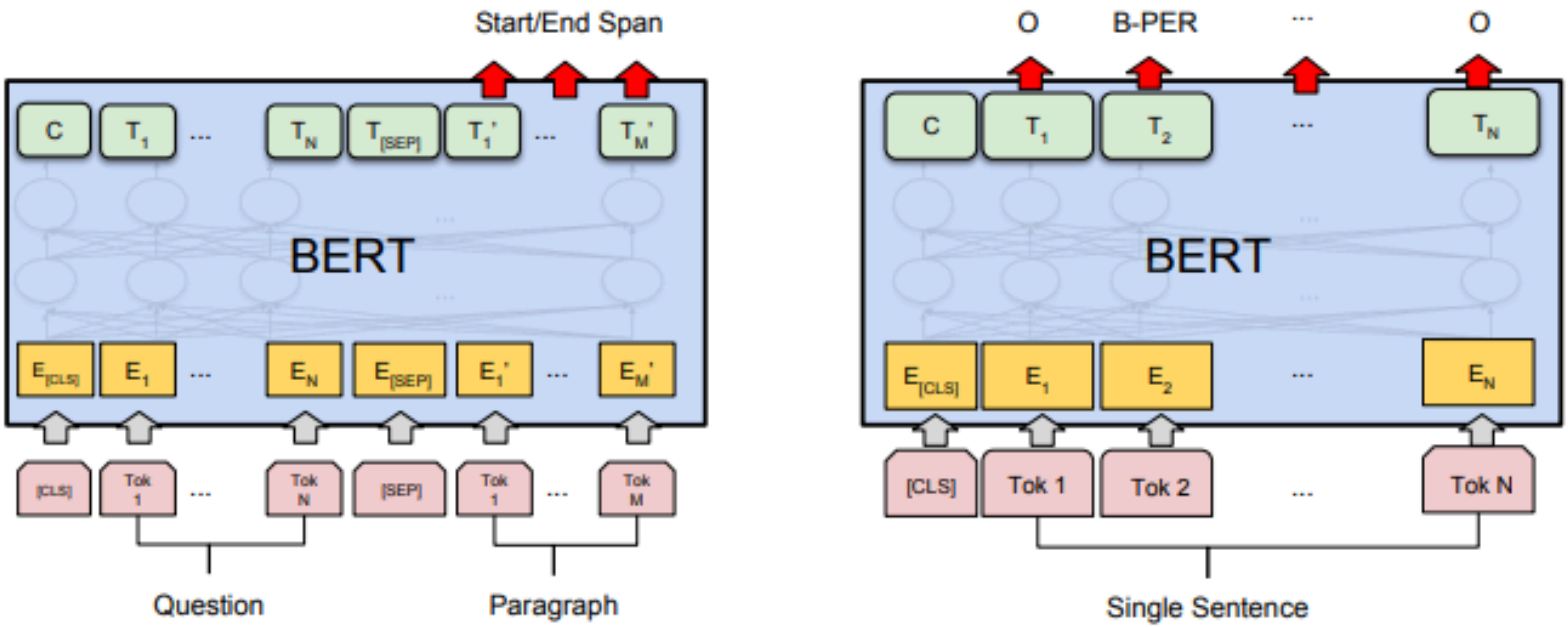
Fine-tuning



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

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

Fine-tuning



(c) Question Answering Tasks:
SQuAD v1.1

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

Benchmarks. 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
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

GLUE Test results, scored by the evaluation server (https://gluebenchmark.com/leaderboard).
The number below each task denotes the number of training examples.

Benchmarks. SQuAD

SQuAD 1.1

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				
BERT$_{\text{BASE}}$ (Single)	80.8	88.5	-	-
BERT$_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
BERT$_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
BERT$_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
BERT$_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

SQuAD 2.0

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				
BERT$_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

Benchmarks. Ablation studies and feature-based approach

Ablation studies

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
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Feature-based approach

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	-

Questions

1. В чем заключается главное архитектурное отличие BERT от GPT?

2. Для каких типов задач особенно важны bidirectionality и использование Next sentence prediction во время pre-training?

3. [MASK] токен используется во время pre-training, но не используется во время fine-tuning, что делается для того, чтобы смягчить последствия этого?

Resources

1. https://arxiv.org/pdf/1810.04805.pdf (original paper)
2. http://peterbloem.nl/blog/transformers (blog post)
3. https://arxiv.org/abs/1706.03762 (Attention is all you need)
4. https://lena-voita.github.io/nlp_course/transfer_learning.html (blog post / NLP textbook)