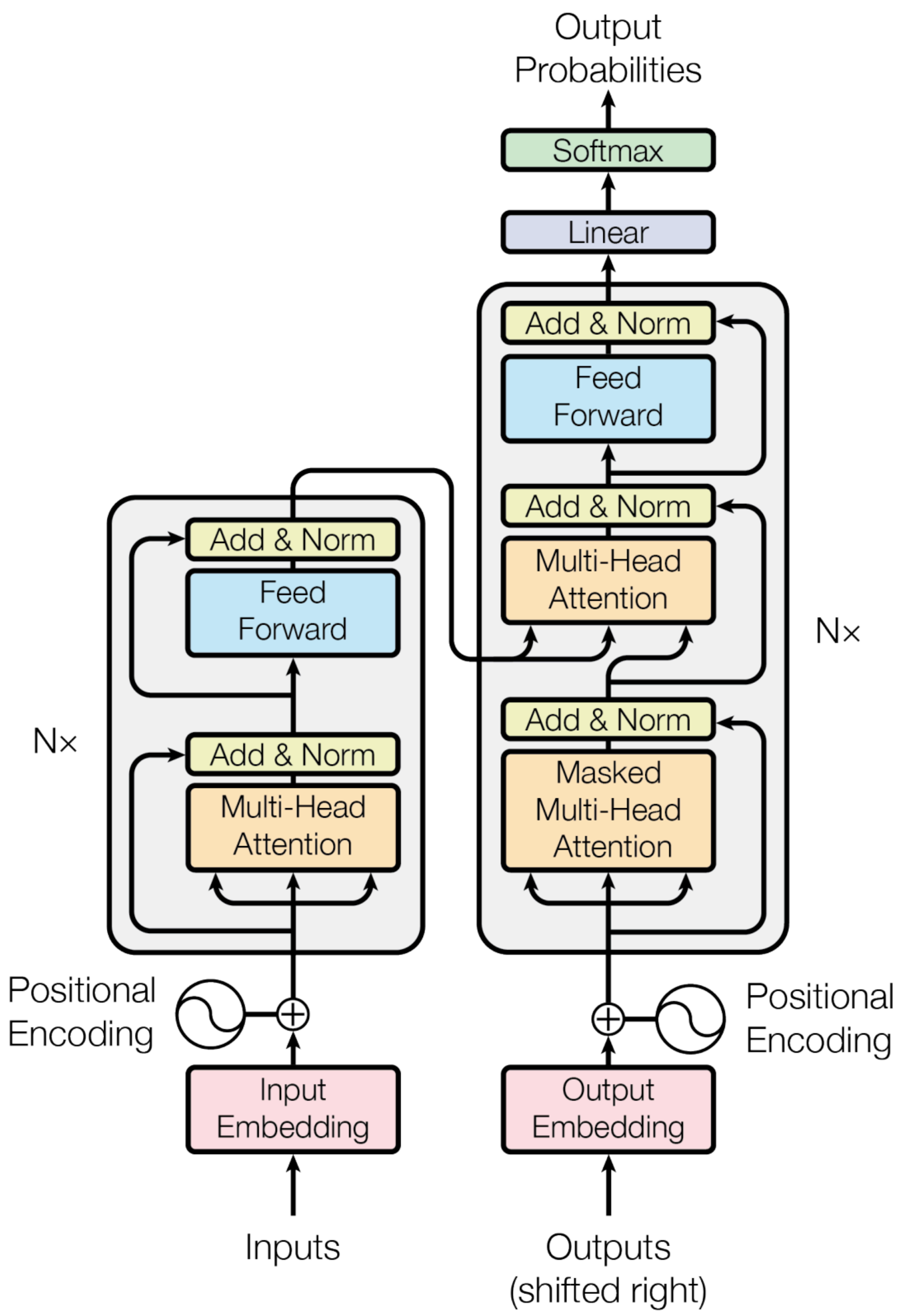
Transformers for MT

Transformers



- Encoder and decoder are both transformers
- Decoder consumes the previous generated token (and attends to input), but has *no recurrent state*

Vaswani et al. (2017)

Transformers

- If we let self attention look at the whole sentence, can access anything in O(1)
- Quadratic in sentence length

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Vaswani et al. (2017)

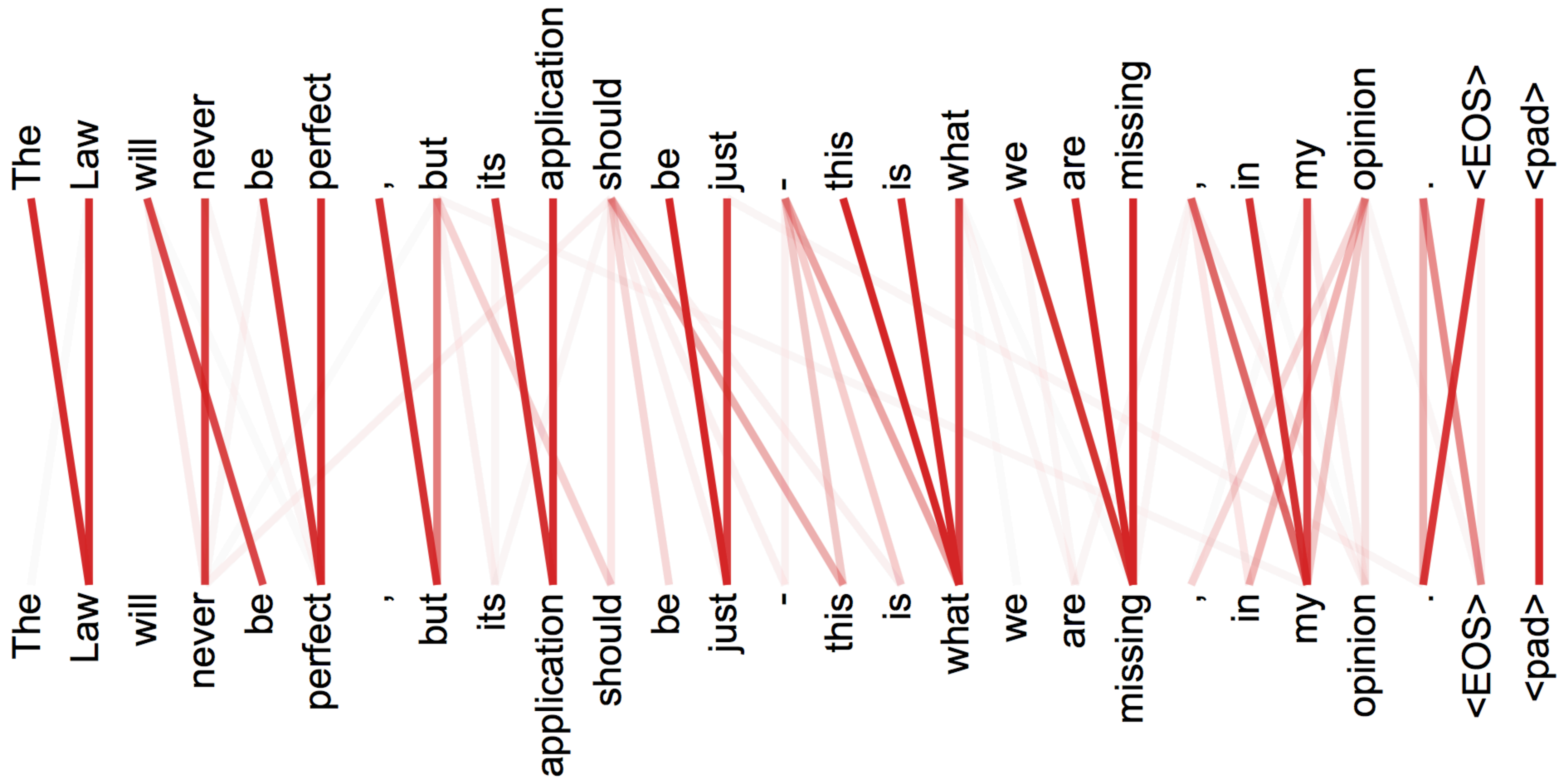
Transformers

Model	BLEU	
	EN-DE	EN-FR
ByteNet [18]	23.75	
Deep-Att + PosUnk [39]		39.2
GNMT + RL [38]	24.6	39.92
ConvS2S [9]	25.16	40.46
MoE [32]	26.03	40.56
Deep-Att + PosUnk Ensemble [39]		40.4
GNMT + RL Ensemble [38]	26.30	41.16
ConvS2S Ensemble [9]	26.36	**41.29**
Transformer (base model)	27.3	38.1
Transformer (big)	**28.4**	**41.8**

- Big = 6 layers, 1000 dim for each token, 16 heads, base = 6 layers + other params halved

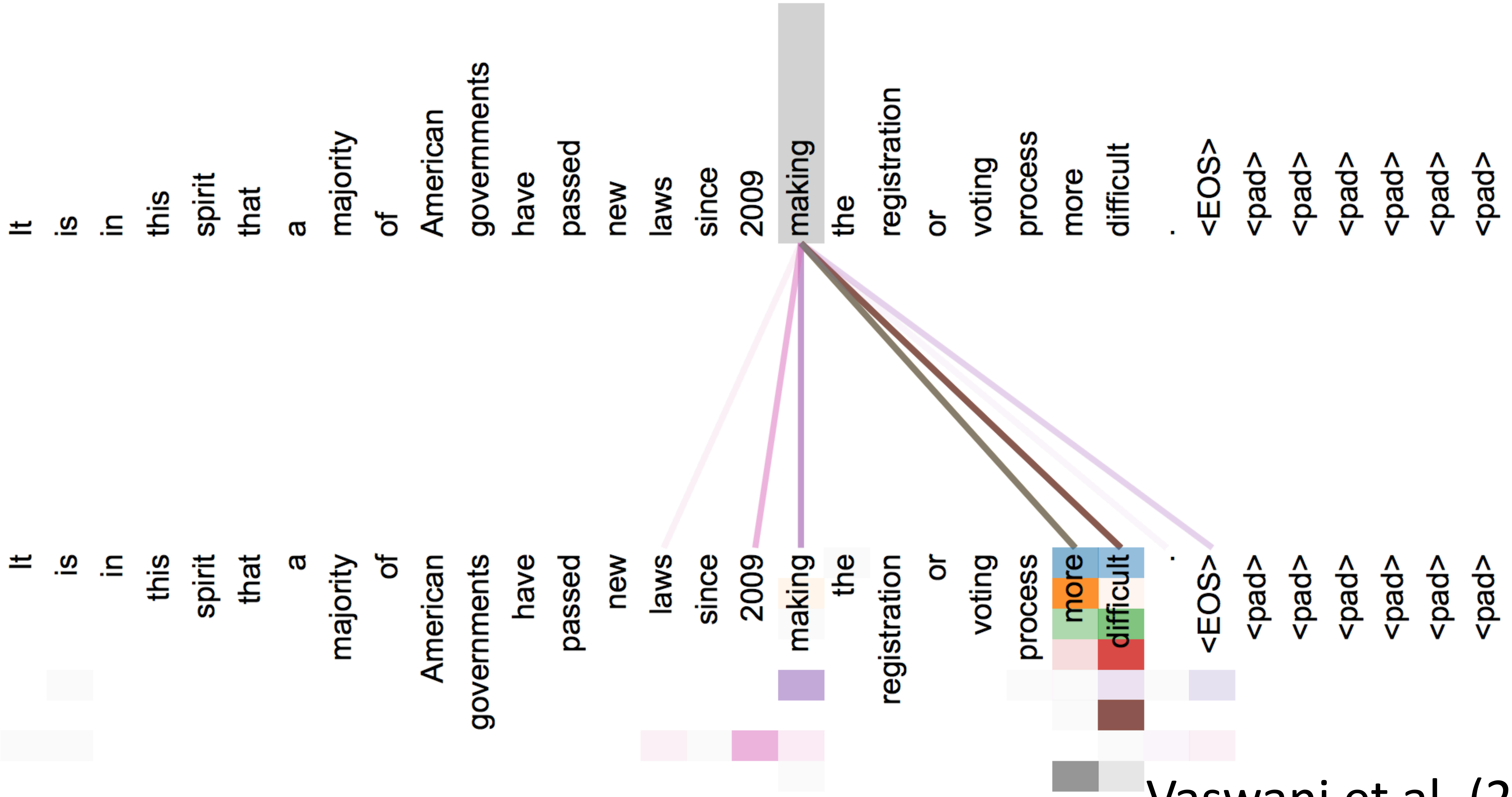
Vaswani et al. (2017)

Visualization: low layer (one head)



Vaswani et al. (2017)

Visualization: high layer (several heads)

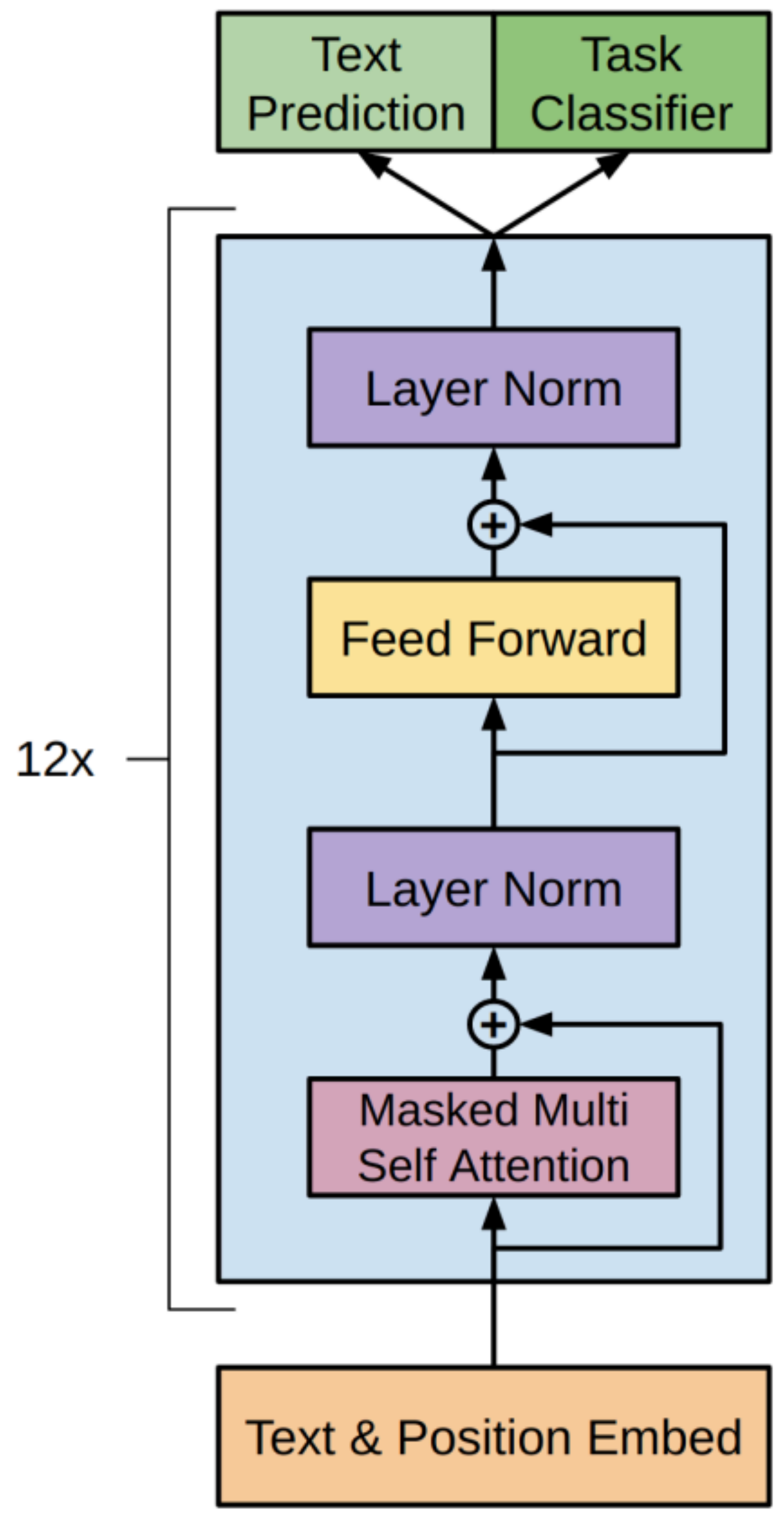


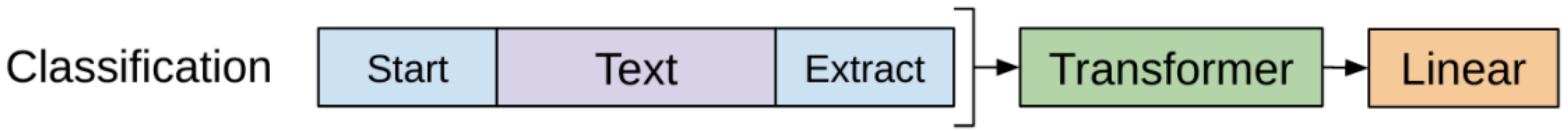
Vaswani et al. (2017)

GPT/BERT

OpenAI GPT

- “ELMo with transformers”
- Fine-tune transformer parameters on the end task



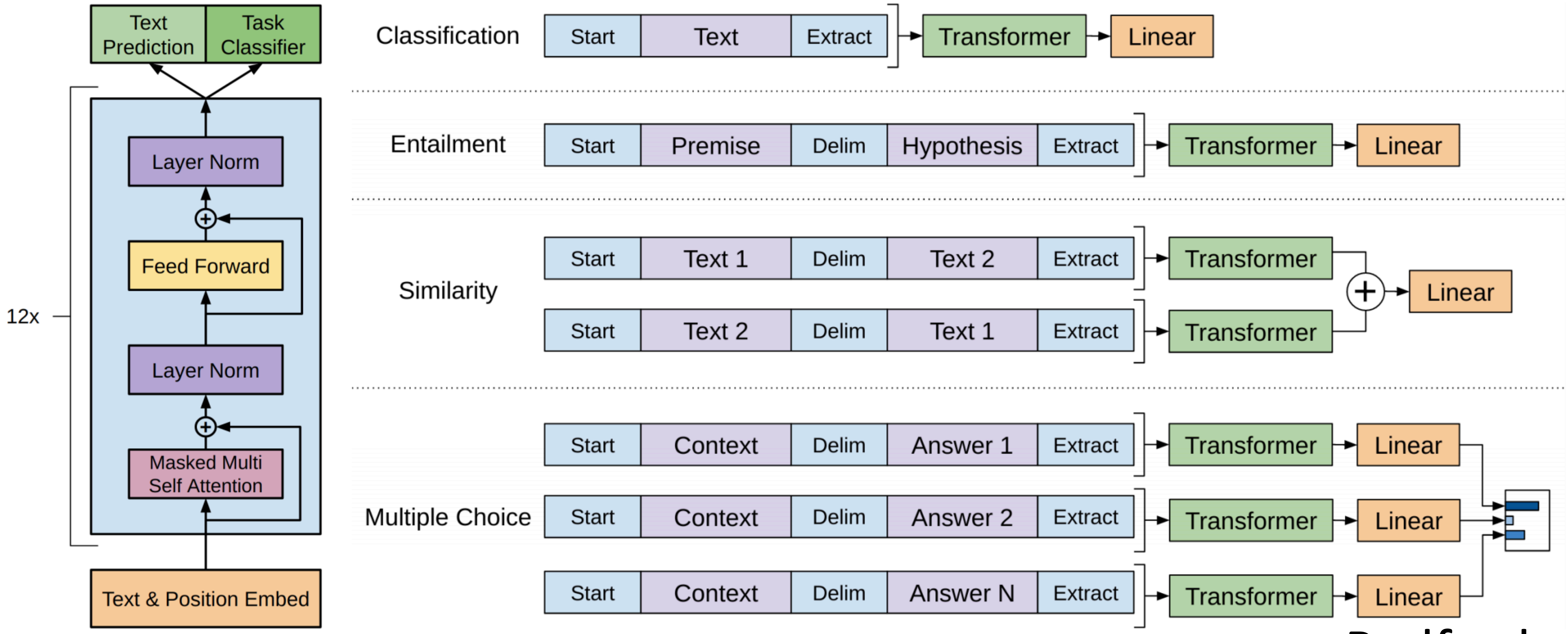


- Assignment 4 architecture but with a pretrained transformer model

Radford et al. (2018)

OpenAI GPT

- "ELMo with transformers"
- Fine-tune transformer parameters on the end task



Radford et al. (2018)

OpenAI GPT

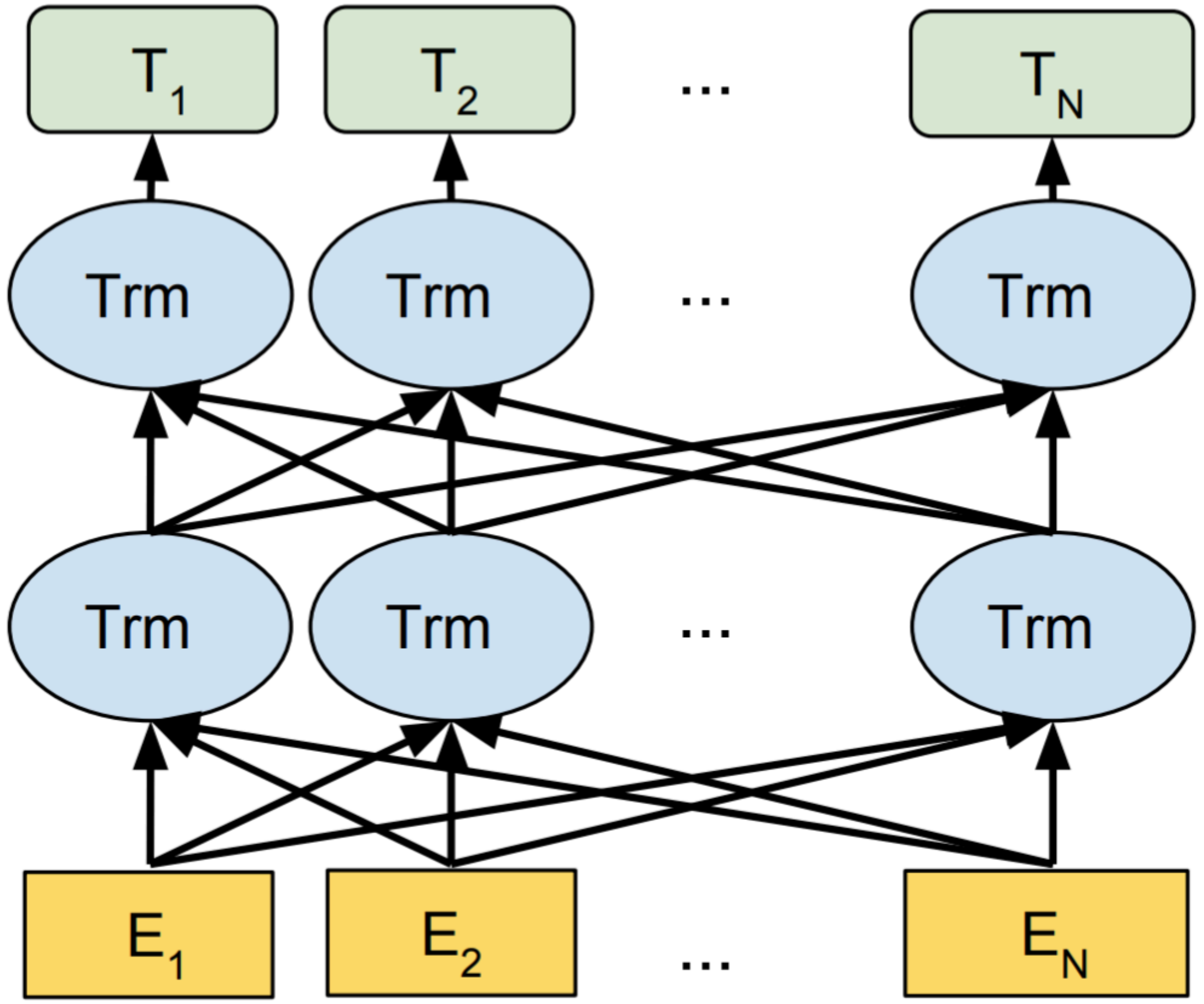
Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	<u>81.0</u>	-	-
Single-task BiLSTM + ELMo + Attn [64]	<u>35.0</u>	90.2	80.2	55.5	<u>66.1</u>	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	<u>68.9</u>
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

- Better than ELMo

Radford et al. (2018)

BERT

- Two-sided Transformer model
- Big model: 24 layers, word dims of 1024, 16 heads
- Small model: 3/4 of this
- Problem: how to do LM when you look at the whole input? Predicting T’s from E’s is trivial



Devlin et al. (2018)

BERT

- Text “infilling” task

I went to the **store** and bought **some milk**

Transformer layers

I went to the [MASK] and bought [MASK] [MASK]

Devlin et al. (2018)

BERT

- Next sentence prediction: predict a true/false label from a [CLS] (classification) input

TRUE I went to the **store** and bought **some milk** || **It** was tasty .

Transformer layers

[CLS] I went to the [MASK] and bought [MASK] [MASK] || [MASK] was tasty .

Devlin et al. (2018)

BERT

- Next sentence prediction: predict a true/false label from a [CLS] (classification) input

FALSE I went to the **store** and bought **some milk** || I flew to Paris

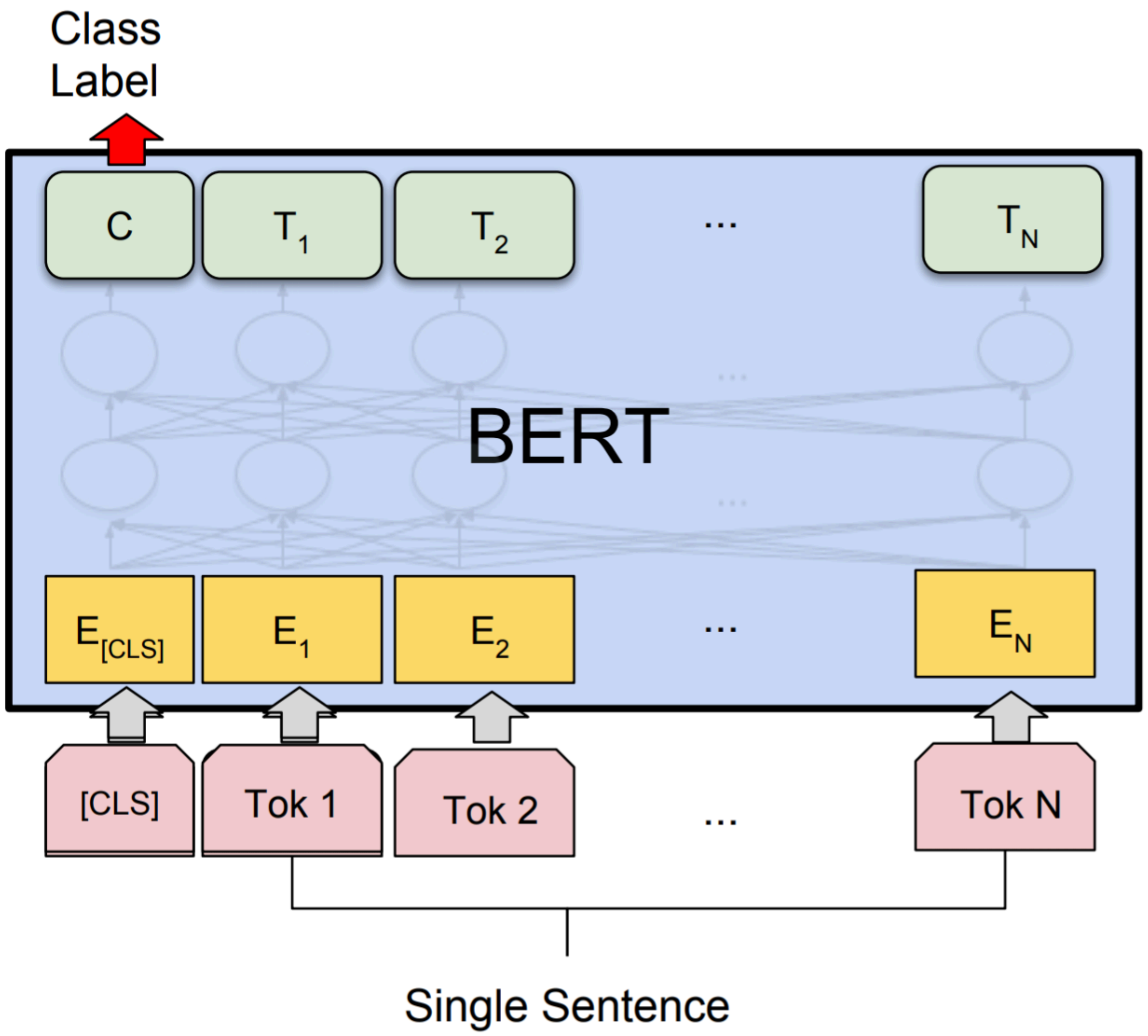
Transformer layers

[CLS] I went to the [MASK] and bought [MASK] [MASK] || [MASK] flew to Paris

Devlin et al. (2018)

BERT

- Can use this like ELMo and just produce embeddings
- Again, works better to fine-tune the whole model on the end task
- "Rewires" the class label output to do the right thing



Devlin et al. (2018)

BERT

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**

- Huge improvements over prior work (even compared to ELMo)
- Effective at "sentence pair" tasks: textual entailment (does sentence A imply sentence B), paraphrase detection

Devlin et al. (2018)

Open AI GPT2

- GPT2: even bigger GPT
- Trained on 40GB of text collected from upvoted links from reddit
- 1.5B parameters — by far the largest of these models trained to date

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Radford et al. (2019)

Open AI GPT2

SYSTEM PROMPT (HUMAN-WRITTEN)

Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

MODEL COMPLETION (MACHINE-WRITTEN, SECOND TRY)

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back.

Scroll down for video

Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today (pictured)

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back

The singer was also wearing a pair of black-rimmed glasses, a black jacket, black jeans and black sandals.

She was carrying a pair of black and white striped gloves and a small black bag.

slide credit: OpenAI