NATURAL LANGUAGE PROCESSING

LECTURE 12: GPTs

goorm

KAIST AI
Graduate School of AI

DAVIAN

Improving language understanding by generative pre-training

GPT1

- It introduces special tokens, such as <S> /<E>/ $, to achieve effective transfer learning during fine-tuning
- It does not need to use additional task-specific architectures on top of transferred representations (e.g., ELMO)

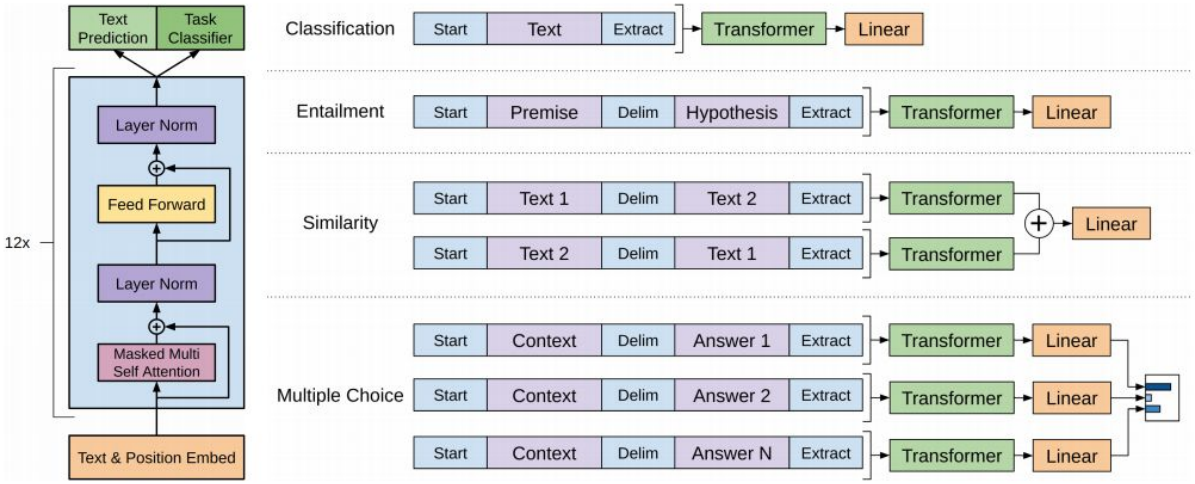


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

- 12-layer decoder-only transformer
- 12 head / 768 dimensional states
- GELU activation unit

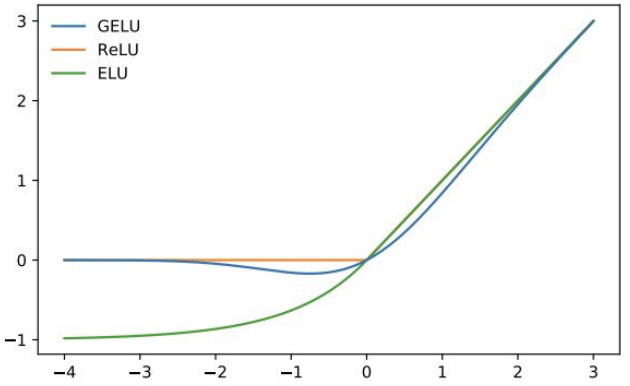


Figure 1: The GELU ($\mu = 0, \sigma = 1$), ReLU, and ELU ($\alpha = 1$).

https://blog.openai.com/language-unsupervised/

Improving language understanding by generative pre-training

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

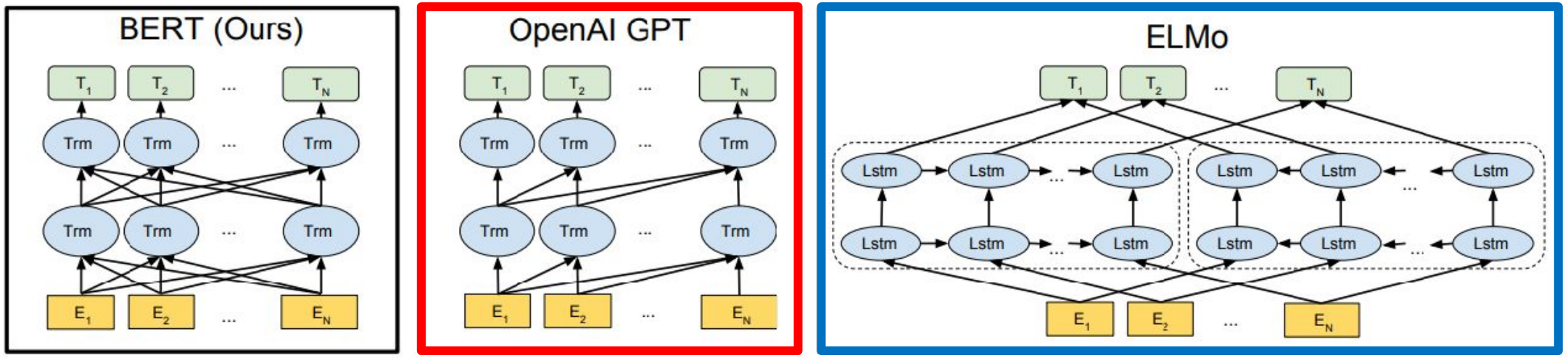
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

https://blog.openai.com/language-unsupervised/

BERT

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

- Learn through masked language modeling tasks
- Use large-scale data and large-scale model



BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL'19

GPT2: Language models are unsupervised multi-task learners

Just a really big transformer LM

- Trained on 40GB of text
  - Quite a bit of effort going into making sure the dataset is good quality
  - Take webpages from reddit links with high karma
- Language model can perform down-stream tasks in a zero-shot setting – without any parameter or architecture modification

https://openai.com/blog/better-language-models/