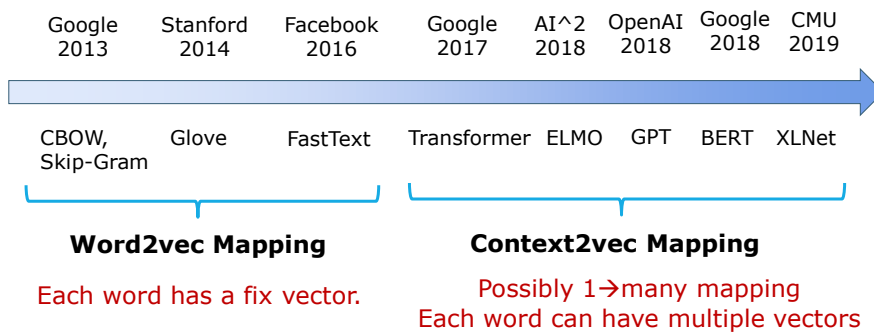


Deep Learning Techniques

DL5. Large Language Models (LLMs)

1

Trend in Language Representation Learning



7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

2

2

Why Contextualized Embedding

□ Fixed Vector

the river bank $\rightarrow (x_1, \dots, x_d)$
 the US bank $\rightarrow (x_1, \dots, x_d)$ **Semantically Ambiguous**

□ Contextualized Mapping

the river bank $\rightarrow (x_1, \dots, x_d)$
 the US bank $\rightarrow (x'_1, \dots, x'_d)$ **Semantically Expressive**

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

3

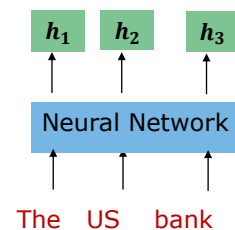
3

Contextualized Embedding

- Mapping input tokens to output embeddings via a function as

$$(h_1, \dots, h_n) := f_{\theta}(t_1, \dots, t_n)$$

- Implementing $f_{\theta}(\cdot)$ by a neural network (RNN or Transformers)
- Training the network on large unlabeled text corpora (**pre-trained language models**)



7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

4

4

Types of Transformer-based Models

□ Encoder only

- BERT, Roberta

□ Decoder only

- GPT1,2,3
- InstructGPT
- ChatGPT

□ Encoder-decoder

- BART
- T5

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

5

5

BERT as a Masked Language Model (MLM)

Target:

pleasant

$$P_{\theta}(x_t | \hat{x}) = \frac{\exp(x_t \cdot h_t)}{\sum_{i=1}^V \exp(x_i \cdot h_t)}$$

Classification over
vocabulary of size V

Contextualized
embedding:

h_1 ... h_t ...

Model:

Transformer Model

masked input:

x_1 x_2 ... x_N

It is a [mask] day

Original input:

It is a pleasant day

Randomly mask
15% of the tokens
Other masking
rate is possible

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

6

6

Should you mask 15% in mask language modeling?

m	Example	Pre-training		Fine-tuning			
		PPL	MNLI	QNLI	SQuAD ³		
15%	We study high [] ing rates [] pre-training language models .	17.7	84.2	90.9	88.0		
40%	We study high [] rates [] pre- [] models .	69.4	84.5 ± 0.3	91.6 ± 0.7	89.8 ± 1.8		
80%	We [] high [] [] models []	1141.4	80.8 ± 3.4	87.9 ± 3.0	86.2 ± 1.8		
Random initialization			61.5 ± 22.7	60.9 ± 30.0	10.8 ± 77.2		

Table 1: Masked examples, validation perplexity (calculated in the same way as Devlin et al., 2019) of different masking rates on the one billion word benchmark (Chelba et al., 2013), and downstream task development performance (SQuAD: F1; accuracy for others). All the pre-trained models have a BERT-large architecture and are trained with the efficient pre-training recipe (§2.2). Full results are provided in Table 7.

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

7

7

Notation and Objective of MLM

- Input sequence $x = (x_1, x_2, \dots, x_N)$.
- Masked subset (targets) (15%) $\bar{x} \subset \{x_1, x_2, \dots, x_N\}$.
- Corrupted sequence \hat{x} (the input with "Mask's")
- Training Objective

$$\max_{\theta} E_{x \in \text{Data}} P_{\theta}(\bar{x} | \hat{x}) = \max_{\theta} E_{x \in \text{Data}} \left(\prod_{x_t \in \bar{x}} P_{\theta}(x_t | \hat{x}) \right)$$

Assuming **conditional independence** among $x_t \in \bar{x}$ given \hat{x} .

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

8

8

Model Optimization in MLM

- Maximizing the likelihood of predicted tokens

$$\max_{\theta} E_{x \in \text{Data}} P_{\theta}(\bar{x} | \hat{x}) = \max_{\theta} E_{x \in \text{Data}} \prod_{x_t \in \bar{x}} P_{\theta}(x_t | \hat{x})$$

- Minimizing the loss function as the **negative log-likelihood**

$$\begin{aligned} \min_{\theta} E_{x \in \text{Data}} \log P_{\theta}(\bar{x} | \hat{x}) \\ &= \min_{\theta} E_{x \in \text{Data}} \left(- \sum_{x_t \in \bar{x}} \log P_{\theta}(x_t | \hat{x}) \right) \\ &= \min_{\theta} E_{x \in \text{Data}} \left(- \sum_{x_t \in \bar{x}} \log \frac{\exp(x_t \cdot h_t)}{\sum_{i=1}^V \exp(x_i \cdot h_i)} \right) \end{aligned}$$

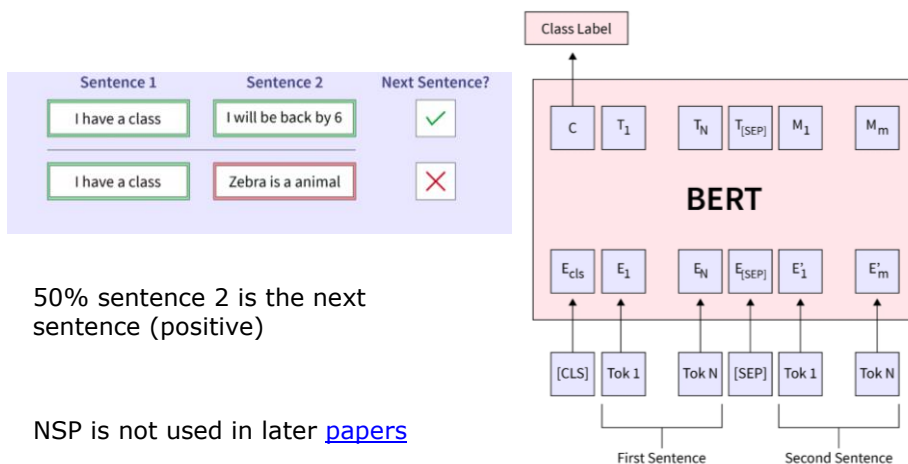
7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

9

9

Next Sentence Prediction (NSP) Objective



7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

10

10

BERT Input Representation (e.g., in the Sentence Entrainment Task)

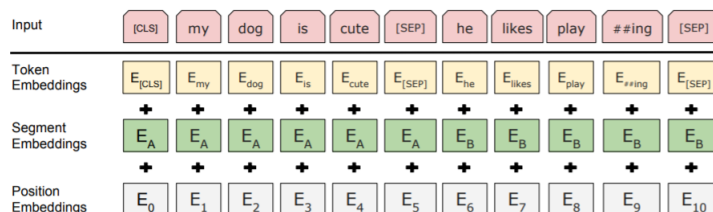


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Without position embeddings, Transformer is invariant to word orders. Segment embeddings are used for next sentence prediction.

11

11

BERT Finetuning for Downstream Tasks

Next sentence prediction

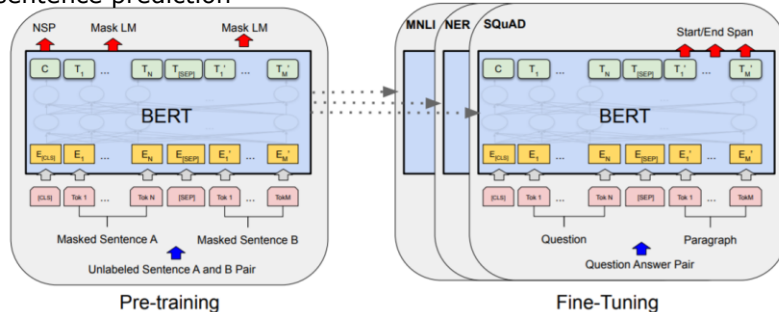
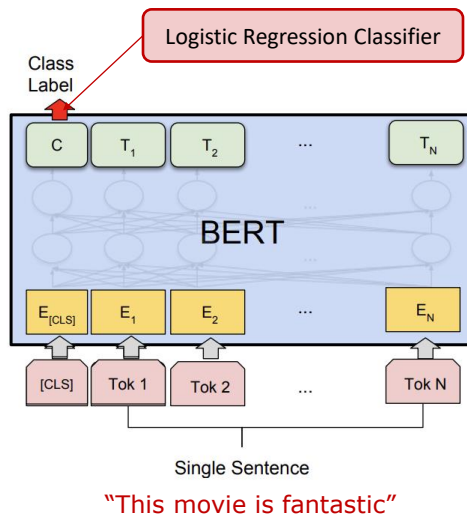


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

12

12

Fine-tuning BERT on classification task

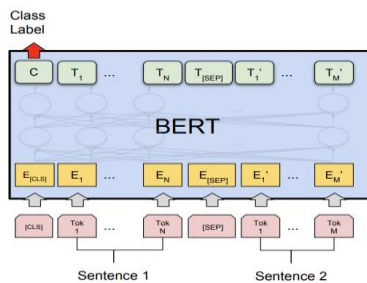


7/1/2020

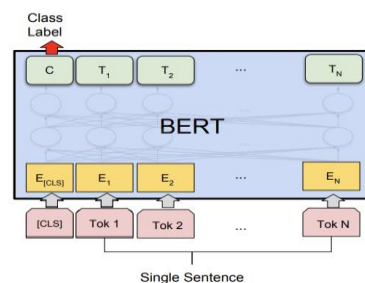
@Yiming Yang, 11-741 lecture on DL Part 5

13

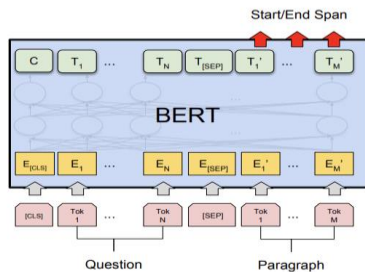
13



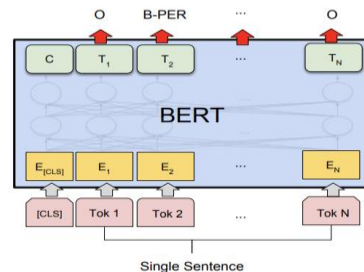
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



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



(c) Question Answering Tasks:
SQuAD v1.1

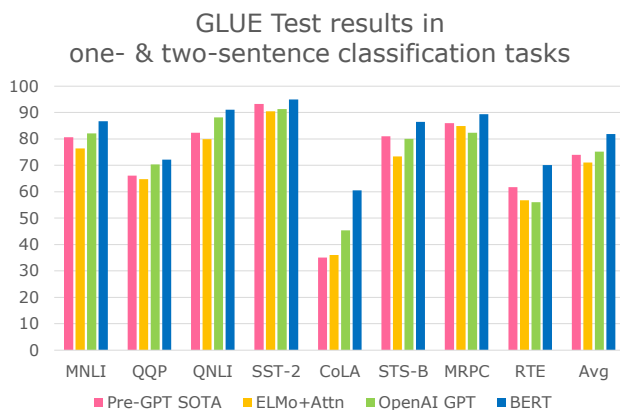


(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

14

14

Performance of ELMo, GPT & BERT (Jacob Devlin et al., 2019)



2/1/2024

@Yiming Yang, 11-741 lecture on DL Part 5

15

15

Types of Transformer Models

□ Encoder only

- BERT, Roberta

□ Decoder only

- GPT1,2,3
- InstructGPT
- ChatGPT

□ Encoder-decoder

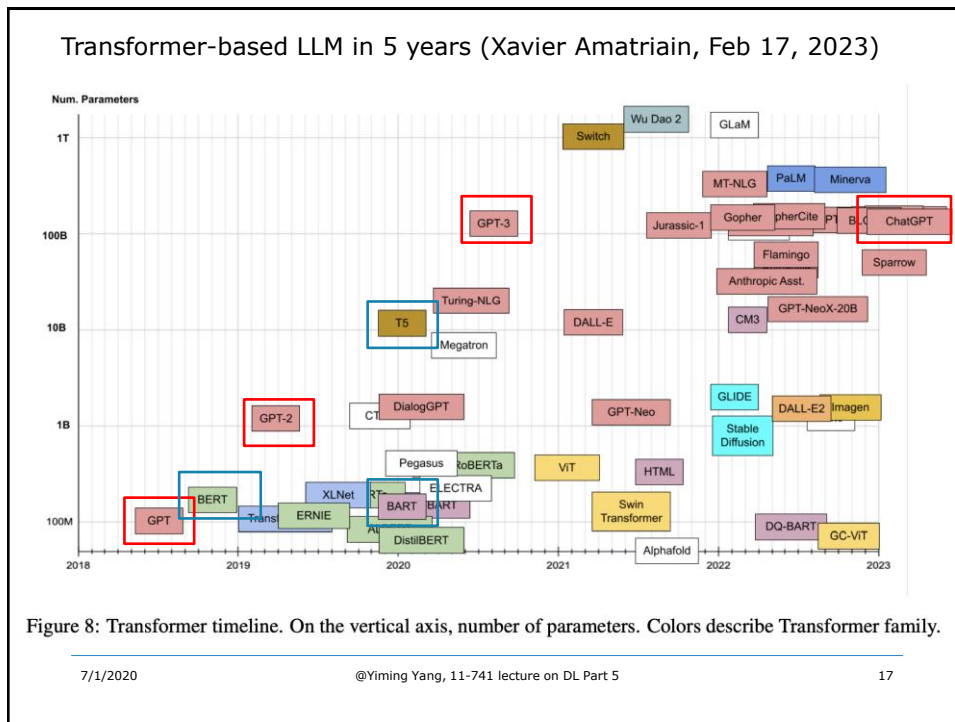
- BART
- T5

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

16

16



17

OpenAI GPT models

Model	time	param	Training data
GPT-1	2018.6	117M	5GB
GPT-2	2019.2	1.5B	40GB
GPT-3	2020.5	175B	570GB
InstructGPT (GPT-3.5)	2022.1	175B	570GB + human written data
ChatGPT (GPT-3.5-turbo)	2022.11	175B	Unreleased
GPT-4	2023.3	16*110B (guessed)	Unreleased

18

18

Autoregressive (AR) LM Objective

- (Left-to-right) Autoregressive Language Modeling Objective

$$\max_{\theta} \sum_{t=1}^T \log P_{\theta}(x_t | x_{<t})$$

- Example: x = The cat sat on the mat

$$\log P_{\theta}(x)$$

$$= \log P_{\theta}(\text{The} \mid [\text{BOS}])$$

$$+ \log P_{\theta}(\text{cat} \mid [\text{BOS}] \text{ the})$$

$$+ \log P_{\theta}(\text{sat} \mid [\text{BOS}] \text{ the cat})$$

...

$$+ \log P_{\theta}(\text{mat} \mid [\text{BOS}] \text{ the cat sat on the})$$

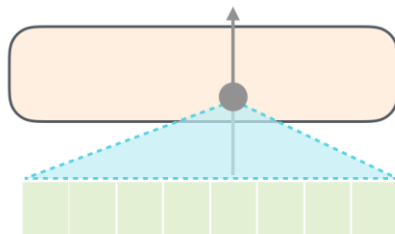
19

19

Masked Self-Attention

Encoder Self-attention

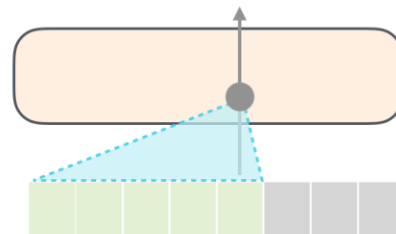
Self-Attention



Attend to the full span

Decoder Self-attention

Masked Self-Attention

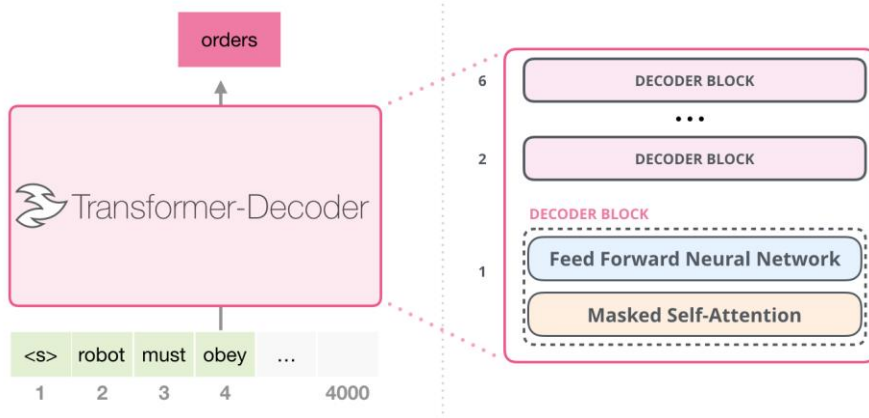


Causal mask: only attend to the left context.

20

20

Autoregressive Transformer



21

21

GPT1 model: pre-training, fine-tuning style

Unsupervised Pre-training

- Use bookcorpus dataset with 5GB of raw text

Supervised Fine-tuning (with additional output layer)

$$P(y|x^1, \dots, x^m) = \text{softmax}(h_l^m W_y)$$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \dots, x^m)$$

Use the last hidden embedding, which attends to the entire context.

(This movie is fantastic, positive)

22

22

GPT-2 from Common Crawl Training Corpus

Gathered from the web dataset of 8 million doc's (40 GB of data)

The texts contain context of English-to-French translations.

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbécile** [I'm not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose.**" which translates as, "Lie lie and something will always remain."

"I hate the word 'perfume,'" Burr says. "It's somewhat better in French: 'parfum.'"

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre côté? -Quel autre côté?**", which means "- How do you get to the other side? - What side?".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

"**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

23

GPT2 for Zero-shot Generation

Autoregressive generation as language model

$$p(x) = \prod_{i=1}^n p(s_i | s_1, \dots, s_{n-1})$$

For specific task using the following prompt:

$$p(\text{output} | \text{input}, \text{task})$$

Example for translation

Task: Translate to French

Input: English text

Output: French text

24

24

GPT2 for Zero-shot Generation

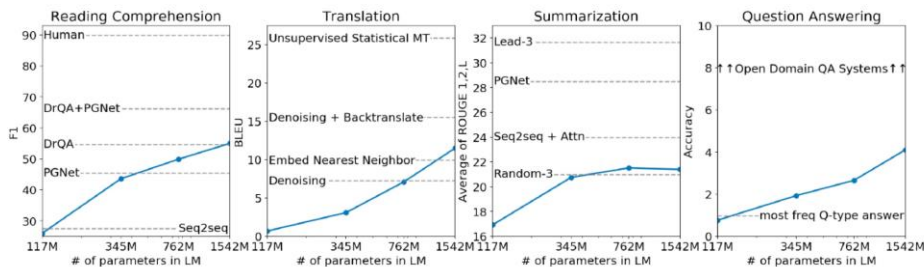


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

25

25

Insights from GPT2

- ✎ Directly use pre-trained language model to generate answers, instead of fine-tuning the model parameters.
- ✎ Zero-shot ability emerges with larger, diverse pre-training corpus.
- ✎ Larger model performs better than smaller model

26

26

GPT3

GPT-3 training data

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

570 GB in total

Model Name	n_{params}
GPT-3 Small	125M
GPT-3 Medium	350M
GPT-3 Large	760M
GPT-3 XL	1.3B
GPT-3 2.7B	2.7B
GPT-3 6.7B	6.7B
GPT-3 13B	13.0B
GPT-3 175B or "GPT-3"	175.0B

Largest GPT2

Llama-2 (07.23) is
6B, 13B, 70B

27

27

GPT3 In-context Learning

Instead of using training data to optimize parameters, use them as inputs:

Sentiment classification

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

LM

Topic classification

Circulation revenue has increased by 5% in Finland. // Finance

They defeated ... in the NFC Championship Game. // Sports

Apple ... development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

LM

No update of model parameter

28

28

GPT3: question-answering benchmark

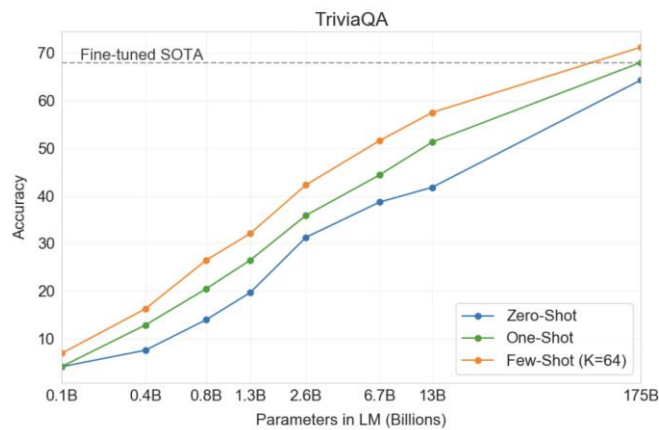
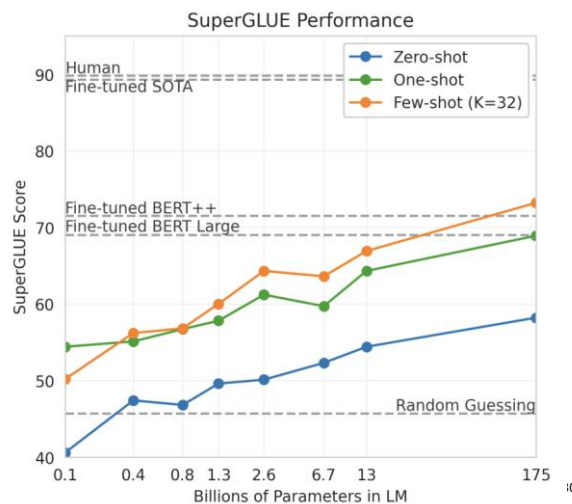


Figure 3.3: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP⁺20]

29

GPT3: Language understanding tasks

superGLUE contains 11 benchmark language understanding tasks.



30

InstructGPT vs. GPT-3 Example Output

Prompt	Explain the moon landing to a 6 year old in a few sentences.
Completion	GPT-3
	Explain the theory of gravity to a 6 year old.
	Explain the theory of relativity to a 6 year old in a few sentences.
	Explain the big bang theory to a 6 year old.
	Explain evolution to a 6 year old.
InstructGPT	People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

InstructGPT has better answers to user's query.
InstructGPT doesn't need in-context examples

31

31

Step1: Supervised Fine-tuning (SFT)

Step 1

Collect demonstration data,
and train a supervised policy.

Model is trained to optimize

$$P_{\theta}(\text{Response}|\text{Instruction})$$

A prompt is
sampled from our
prompt dataset.

Explain the moon
landing to a 6 year old

A labeler
demonstrates the
desired output
behavior.

Some people went
to the moon...

This data is used
to fine-tune GPT-3
with supervised
learning.

SFT

Instruction:

Explain the moon landing to a 6 year old.

Model Output:

People went to moon ...

32

32

Step2,3: Reinforcement Learning from Human Feedback (RLHF)

Step 2

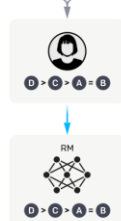
Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



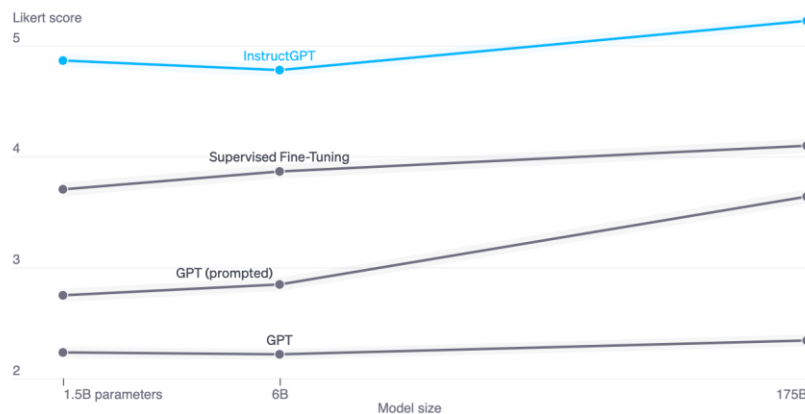
Reward Model:
Assign scores to model generated responses.

RL:
train Policy model (LLM) to generate responses with higher scores.

33

33

Evaluation Results



Quality ratings of model outputs on a 1-7 scale (y-axis), for various model sizes (x-axis), on prompts submitted to InstructGPT models on our API. InstructGPT outputs are given much higher scores by our labelers than outputs from GPT-3 with a few-shot prompt and without, as well as models fine-tuned with supervised learning. We find similar results for prompts submitted to GPT-3 models on the API.

34

34

ChatGPT and GPT-4

- 🔗 Instruction following GPT models developed by OpenAI
- 🔗 Other competitors include
 - ✂ BARD from Google
 - ✂ Claude from Anthropic
 - ✂ Pi from Inflection AI
- 🔗 Open-source alternatives
 - ✂ Llama-2-chat (7B, 13B, 70B)
 - ✂ Vicuna, Alpaca

35

35

Types of Transformer Models

- ❑ **Encoder only**
 - BERT, Roberta
- ❑ **Decoder only**
 - GPT1,2,3
 - InstructGPT
 - ChatGPT
- ❑ **Encoder-decoder**
 - BART
 - T5

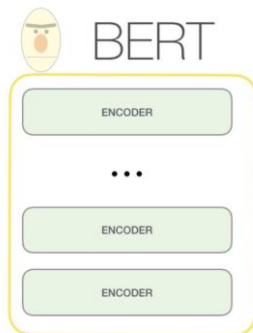
7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

36

36

Encoder and decoder Transformer



Encode text into hidden representation



Autoregressive generation of text (embeddings)

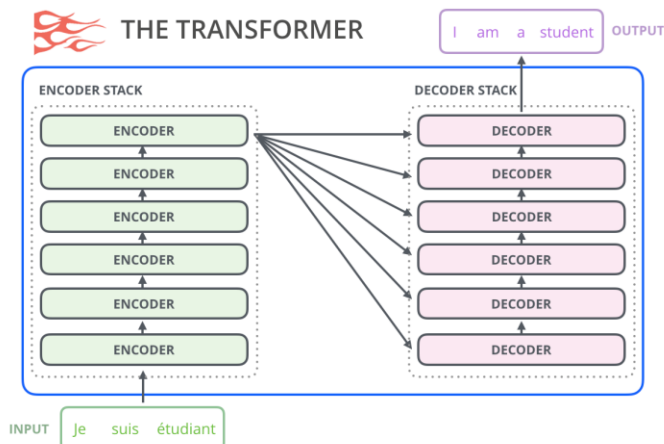
7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

37

37

Encoder-decoder Transformer Overview



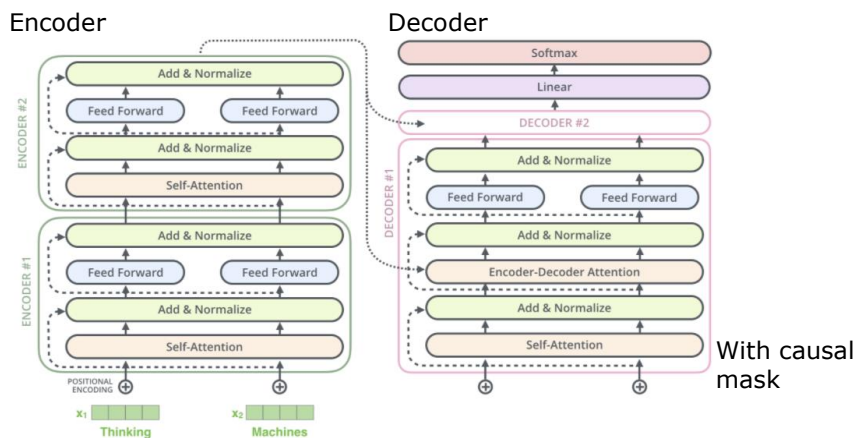
7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

38

38

Encoder-decoder Transformer Design



7/1/2020

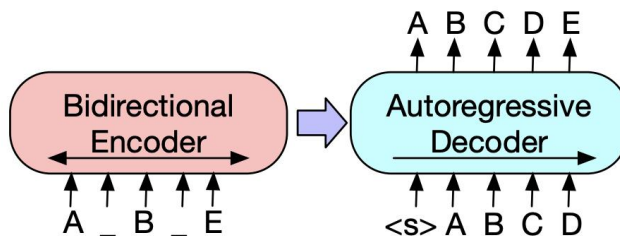
@Yiming Yang, 11-741 lecture on DL Part 5

39

39

BART model: Bidirectional and Autoregressive Transformers.

Example of recovering masked token by generating the original sequence (similar to BERT MLM)



7/1/2020

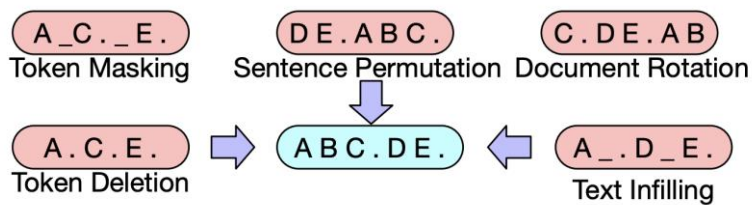
@Yiming Yang, 11-741 lecture on DL Part 5

40

40

BART objective: Denoising Autoencoder

- Encoder-decoder Transformer model can support different types of input corruptions



7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

41

41

BART performance on summarization

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	44.16	21.28	40.90	45.14	22.27	37.25

Table 3: Results on two standard summarization datasets. BART outperforms previous work on summarization on two tasks and all metrics, with gains of roughly 6 points on the more abstractive dataset.

RL (ROUGE-L): the longest common subsequence (LCS) between the generated summary and the reference summary.

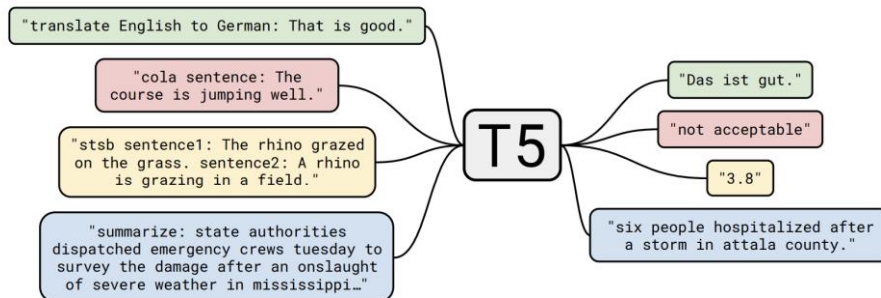
7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

42

42

T5: Text-to-Text Transfer Transformer



7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

43

43

T5 Model: Masked Span Generation

Predicting the masked span of text:



7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

44

44

Performance of T5

Model	GLUE Average	SQuAD F1	SuperGLUE Average
Previous best	89.4 ^a	95.5 ^a	84.6 ^d
T5-Small	77.4	87.24	63.3
T5-Base	82.7	92.08	76.2
T5-Large	86.4	93.79	82.3
T5-3B	88.5	94.95	86.4
T5-11B	90.3	96.22	88.9

GLUE & SuperGLUE: Benchmark datasets evaluate the performance of natural language understanding (NLU) models across a wide range of tasks

SQuAD: Question-answering benchmark datasets

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

45

45

Performance of T5

Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1
Previous best	33.8^e	43.8^e	38.5^f	43.47 ^g
T5-Small	26.7	36.0	26.8	41.12
T5-Base	30.9	41.2	28.0	42.05
T5-Large	32.0	41.5	28.1	42.50
T5-3B	31.8	42.6	28.2	42.72
T5-11B	32.1	43.4	28.1	43.52

WMT: Translation Benchmark datasets

CNN/DM: summarization benchmark datasets

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

46

46

Summary

- LLMs have significant impact on both academia and industries.
- The state-of-the-art LLMs are Transformer-based models, which can be roughly characterized into the categories of encoder-only, decoder-only and encoder-decoder architectures.
- On-going researches focus on how to use LLMs under different scenario: downstream fine-tuning, zero-shot inference, in-context learning, instruction following, etc.

7/1/2020

@Yiming Yang, 11-741 lecture on DL Part 5

47

47

References

- Devlin, Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *ArXiv* abs/1810.04805 (NAACL 2019)
- Matthew E. Peters et al. "Deep contextualized word representations". NAACL 2018.
- Zhilin Yang*, Zihang Dai*, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding" . NeurIPS 2019. (* indicates the co-first authors)
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

48

48

References

- ❑ Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners.
 - ❑ Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.
 - ❑ Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.
-