

Natural Language Processing



Existing Large Language Models

Kevin Lin – UC Berkeley

March 15, 2023

Existing Large Language Models



Announcements

- HW4 – finetuning LLMs: release today
- HW5 – prompting LLMs: released early April
- Panel Topics Overview
- Today:
 - BERT
 - T5
 - GPT3



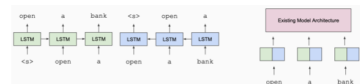
BERT

- Bidirectional Encoder Representations from Transformers (Devlin et al., 2018)
- A working general recipe: pretrain and finetune
- SOTA across token + sentence level tasks
- Deep bidirectional encoder-only model



Previous Work

- ELMO: Deep Contextualized Word Embeddings (Peters et al., 2018)
- Left-to-right, right-to-left unidirectional LSTMs
- Plug in as features
- Single sentences




Previous Work

- GPT: Improving Language Understanding by Generative Pre-training (Radford et al., 2018)
- Finetuning
- Left-to-right
- BooksCorpus (512 length)



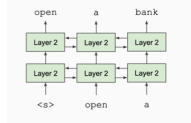
BERT

- Objectives: Masked Language Modeling + Next Sentence Prediction
- Deep encoder-only transformer
- Learn from bidirectional context
 - go to the bank to make a deposit
 - on the river bank




Masked Language Modeling

- Problem: words “see” themselves



Masked Language Modeling

- Solution: masking
- Cloze-style task (Taylor, 1953)
- Denoising-autoencoders
- Select 15%
- No [MASK] during fine-tuning
 - 80% Replace with [MASK]




Next Sentence Prediction

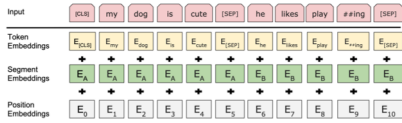
- Learn relationships between sentences
- Predict whether sentence A follows sentence B
- Later shown to be not very helpful

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

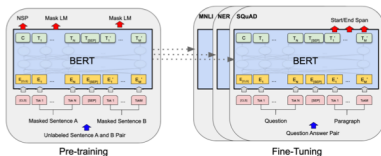
Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence



Input Representation



BERT



Sentence-Level Tasks

- Linear layer on top of [CLS] token

GLUE Results

System	MNLI (tokens)	QQP	QNLI	STS-B	CoLA	STS-B	MRPC	RTE	Average
	70%	86%	90%	87%	57%	75%	73%	72%	
Pre-OpenAI GPT-3	80.4/80.1	66.1	82.3	93.2	55.0	81.0	86.0	61.7	74.0
BE-27/34/41/46/48/49	76.4/76.1	45.4	79.9	89.4	56.0	73.3	84.9	56.8	71.0
OpenAI GPT-3	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT _{Base}	84.6/84.4	71.2	88.1	91.3	52.1	85.8	85.0	66.4	78.9
BERT _{Large}	86.7/86.9	72.1	91.1	94.9	66.5	89.3	89.1	76.1	83.9

MNLI
Premise: Hills and mountains are especially snow-capped in January.
Hypothesis: Arsenium hates nature.
Label: Contradiction

CoLA
Sentence: The wagon rumbled down the road.
Label: Acceptable
Sentence: The car honked down the road.
Label: Unacceptable

Token-Level Tasks

- Extractive QA, NER
- Linear layer on top of token representations

What was another term used for the oil crisis?
 Ground Truth Answers: **oil crisis**, **oil shock**, **oil price shock**, **oil price crisis**
 Prediction: shock

The 1973 **oil crisis** began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC), consisting of the Arab members of OPEC plus Egypt and Syria, proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US\$3 per barrel to nearly \$12 globally. US prices were significantly higher. The embargo caused an **oil crisis**, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "**oil price shock**", followed by the 1979 **oil crisis**, which the "second oil shock".

Training BERT

- Data: Wikipedia (2.5B words) + BooksCorpus (800M words)
- 1M steps
- Batch size: 131,072 words
 - (1024 sequences * 128 length) or (256 sequence * 512 length)
- BERT-large
 - 24 layers, 1024 hidden size, 16 attention heads, 340M parameters
- BERT-base
 - 12 layers, 768 hidden size, 12 attention heads, 110M parameters

BERT Aftermath

- Explosion of variations:
 - RoBERTa: Train longer, remove NSP
 - ALBERT: share weights
 - SpanBERT: mask out contiguous spans
 - Electra: learn from all tokens
- Efficiency:
 - DistillBERT, qBERT, ...
- BERT for X
 - SciBERT: scientific documents
 - ClinicalBERT: clinical documents, ...

BERT Aftermath

- "BERTology"
 - What does an LLM encode about syntax, semantics, knowledge, etc.?
- Generation from BERT
 - Mask-Predict: Parallel Decoding of Conditional Masked Language Models (Ghazvininejad et al., 2019)
 - BERT as Markov Random Field LM (Wang et al., 2019)

T5

T5

- T5: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
- Objectives, architectures, datasets, transfer
- Unified format: text in, text out
- Discriminative and generative tasks

T5 setup

- Start with a basic setup, and get first order effects: objectives, architectures, datasets, transfer

Original text
Thank you for inviting me to your party last week.

Input
Thank you <X> me to your party <Y> week.

Targets
<X> for inviting <Y> last <Z>

T5 setup

- Encoder-decoder
- Each size of BERT-base
- Relative positional embedding
- C4: Colossal Clean Crawled Corpus
 - Filter out javascript, non-English, List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words
 - 750B
 - Note: tokenizer handles French, Romanian, German

T5 setup

Pretrain
BERT_{base} sized encoder-decoder Transformer
Denoising objective
C4 dataset

2nd or 3rd steps
Invert square root learning rate schedule

Finetune
GLUE
CNN/DM
SQuAD
SuperGLUE
WMT14 EnDe
WMT15 EnFr
WMT16 EnRo

2nd or 3rd steps
Invert square root learning rate schedule

Evaluate on validation
step 750000
step 760000
step 770000
step 780000

2nd or 3rd steps
Invert square root learning rate schedule

Evaluate all checkpoints, choose the best

T5 architecture

Fully-visible Causal Causal with prefix

Output
y₅
y₄
y₃
y₂
y₁

Input
x₁ x₂ x₃ x₄ x₅

T5 Architecture

Architecture	Params	Cost	GLUE	CNN/DM	SQuAD	SuperGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	2P	M	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Enc-dec, shared	P	M	82.41	18.78	80.63	70.73	26.72	39.63	27.46
Enc-dec, 6 layers	P	M/2	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39

Decoder
Encoder

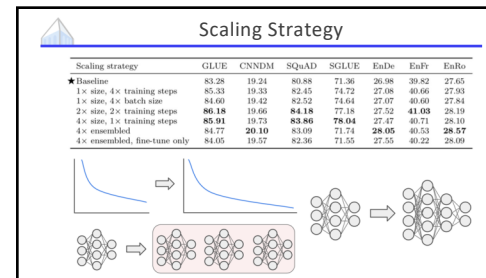
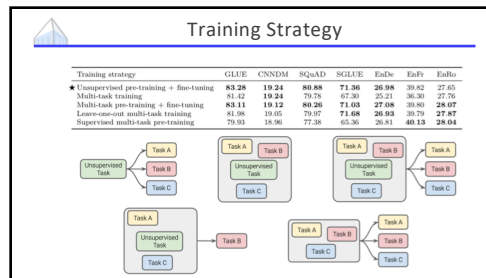
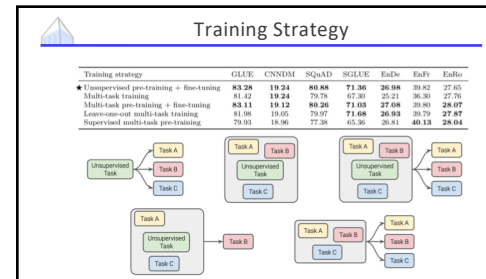
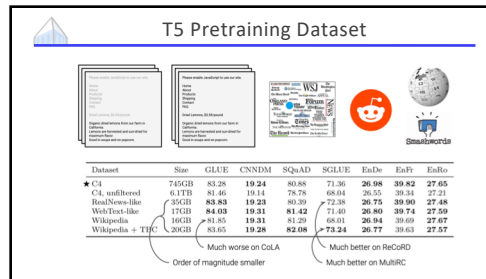
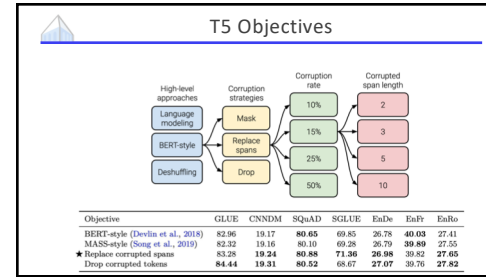
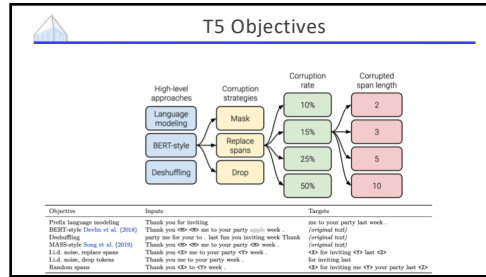
Language model
Prefix LM

T5 Architecture

Architecture	Params	Cost	GLUE	CNN/DM	SQuAD	SuperGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	2P	M	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Enc-dec, shared	P	M	82.41	18.78	80.63	70.73	26.72	39.63	27.46
Enc-dec, 6 layers	P	M/2	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39

Decoder
Encoder

Language model
Prefix LM



T5: Putting It Together

- Encoder-Decoder
- Span Replacement
- C4
- Multi-task pretraining
- Large models, trained longer

T5: Putting It Together

Model	Parameters	# layers	d_{model}	d_q	d_{kv}	# heads
Small	60M	6	512	2048	64	8
Base	220M	12	768	3072	64	12
Large	770M	24	1024	4096	64	16
3B	3B	24	1024	16384	128	32
11B	11B	24	1024	65536	128	128

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4*	69.2*	97.1*	93.0*	91.5*	92.1*	92.3*
T5-Small	77.1	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	90.5	71.6	97.5	92.8	90.4	93.1	92.8

Finetuning Limitations

- Requires large supervised dataset
- Spurious correlations in supervised finetuning dataset
- Poor sample efficiency vs. humans

GPT3

- Language Models are Few-Shot Learners (Brown et al., 2020)
- Decoder-only model

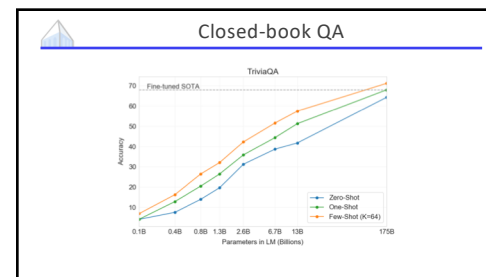
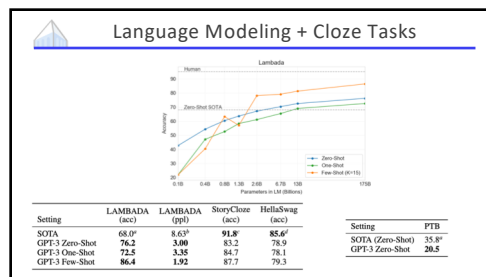
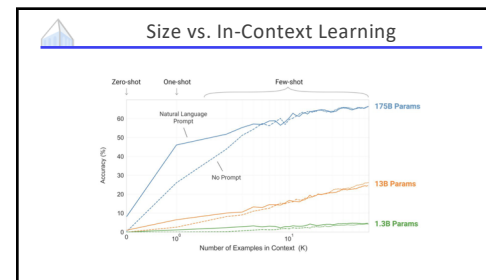
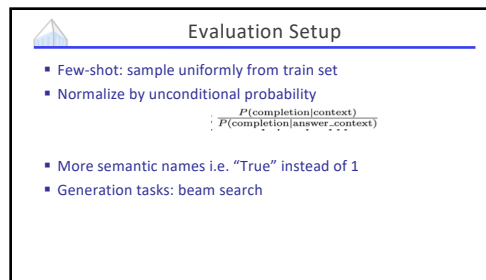
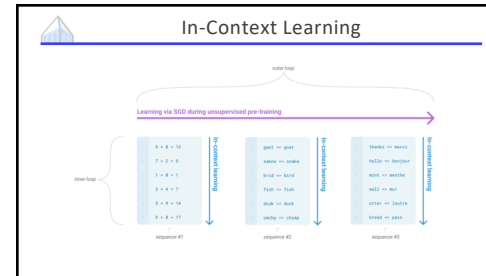
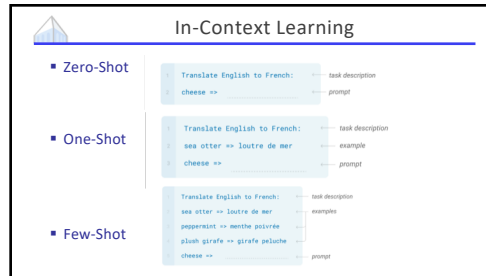
GPT3

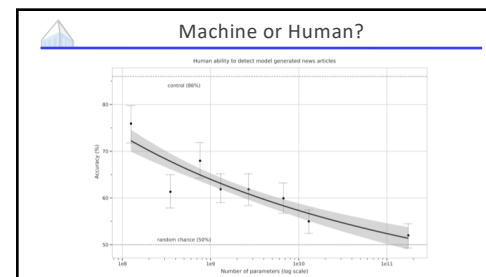
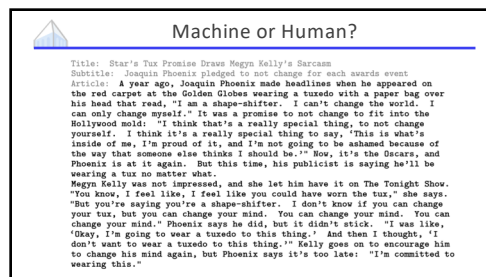
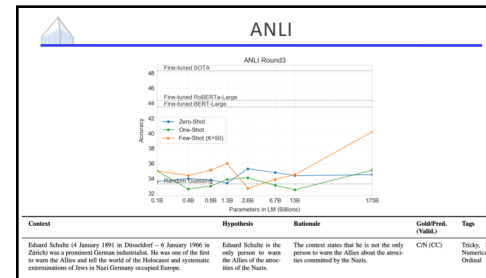
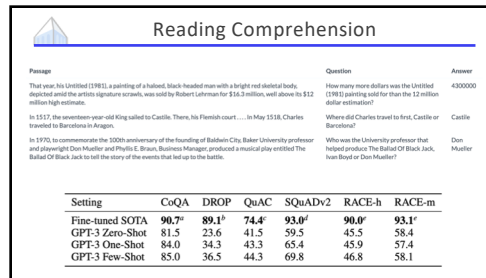
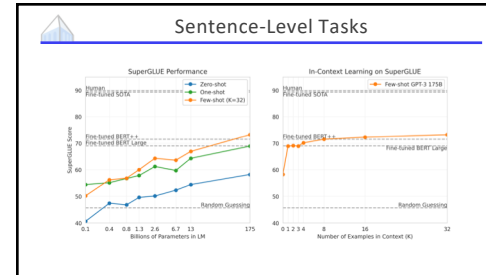
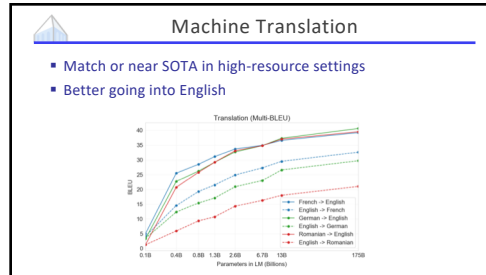
Model Name	n_{tokens}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13.0B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

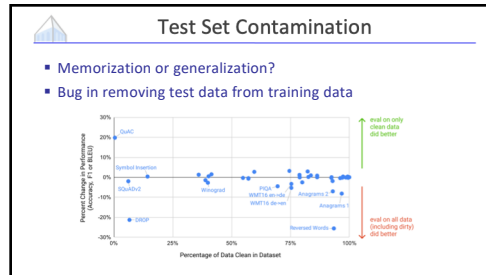
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

GPT3 Training

- Larger models -> larger batch sizes, smaller learning rate
- Model parallelism: across layers
- Adam optimizer
- Gradient clipping: 1
- Linear warmup learning rate, cosine decay
- Weight decay 0.1



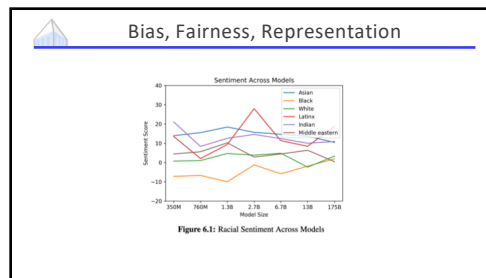




Bias, Fairness, Representation

Table 6.1: Most Biased Descriptive Words in 1718 Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 21.3	Average Number of Co-Occurrences Across All Words: 23.9
Large (146)	Opportunity (132)
Manly (115)	Shallily (121)
Lazy (114)	Naughty (121)
Fortunate (113)	Easy-going (121)
Excellent (113)	Playful (118)
Produce (108)	Right (118)
July (108)	Program (110)
Stable (95)	Gorgeous (100)
Possible (22)	Sacked (95)
Survive (17)	Reinhold (158)



- ### Open Questions
- Scaling
 - Evaluation
 - Misuse, Risks
 - Grounding
 - Controllability
 - Multilingual