Language Models are Unsupervised Multitask Learners

Alec Radford, et al

OpenAI

2019

Abstract

- NLP Tasks were approached with supervised learning
  - Labeling and fine-tuning for a specific task
  - Not enough generalization

- We suggest unsupervised multitask, achieved by
  - Large model GPT-2 (1.5B Parameters)
  - Huge dataset of millions of webpages called WebText

Introduction

- Problem of “Narrow expert”
  - Single task training on single domain dataset
  - Sensitive to slight changes in the data distribution
  - Poor performance on GLUE, decaNLP

- We try to make “Competent generalists”
  - Unsupervised pre-training
  - Zero-shot setting (No fine-tuning)

Approach

- Traditional Language models
  - Estimate conditional distribution $p(output|input)$

- Multitask and meta-learning settings
  - Estimate conditional distribution $p(output|input, task)$
  - Find triplet $(input, output, task)$ in natural language sentence
  - It is achieved by MQAN [1]
  - We expect sufficient data and model size can perform it

---

[1] Bryan McCann et al. *The Natural Language Decathlon: Multitask Learning as Question Answering*. 2018. arXiv: 1806.08730 [cs.CL].

Training dataset

- Prior works used single domain dataset
  - News articles, Wikipedia, Fiction books

- We used WebText, containing 45 million links
  - We scraped links from Reddit, which received at least 3 karma
  - De-duplication and some heuristic based cleaning
  - 8 million documents for a total of 40 GB of text

Input Representation

- Word-level Models
  - Preprocessing is needed
  - Lowercasing, Tokenization, Out-of-vocabulary tokens

- Byte-level Models
  - Treating text as UTF-8 bytes without preprocessing
  - Low performance on large scale datasets
  - 8 million documents for a total of 40 GB of text

Input Representation

- Byte Pair Encoding (BPE) [1]
  - Treating frequent pairs as a token
  - ex) preprocess → ‘pre’, ‘process’

  - If character categories are different, they are divided
  - Categories include letters, numbers, punctuation
  - ex) fair-play → ‘fair’, ‘play’

- Byte-level BPE (BBPE)
  - Despite its name, BPE operates on Unicode code points
  - We changed it to byte-level instead of Unicode symbols

---

[1] Rico Sennrich. “Neural machine translation of rare words with subword units”. In: *arXiv preprint arXiv:1508.07909* (2015).

Model

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

- The 4 models largely follow the details of the OpenAI GPT
- Few modifications of layer normalizations
  - It moved to the input of each sub-block
  - Additional normalization after the final self attention
- A modified initialization
  - Scaled weights of residual layers by $\frac{1}{\sqrt{N}}$
  - N is the number of residual layers

Experiment

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

- Accuracy (ACC)
  - Correctness of word prediction
- Perplexity (PPL)
  - How ”surprised” the model is by the data
- Bits per Byte (BPB)
  - It measures the compression efficiency of a model
- Bits per Character (BPC)
  - It measures the compression efficiency of a model

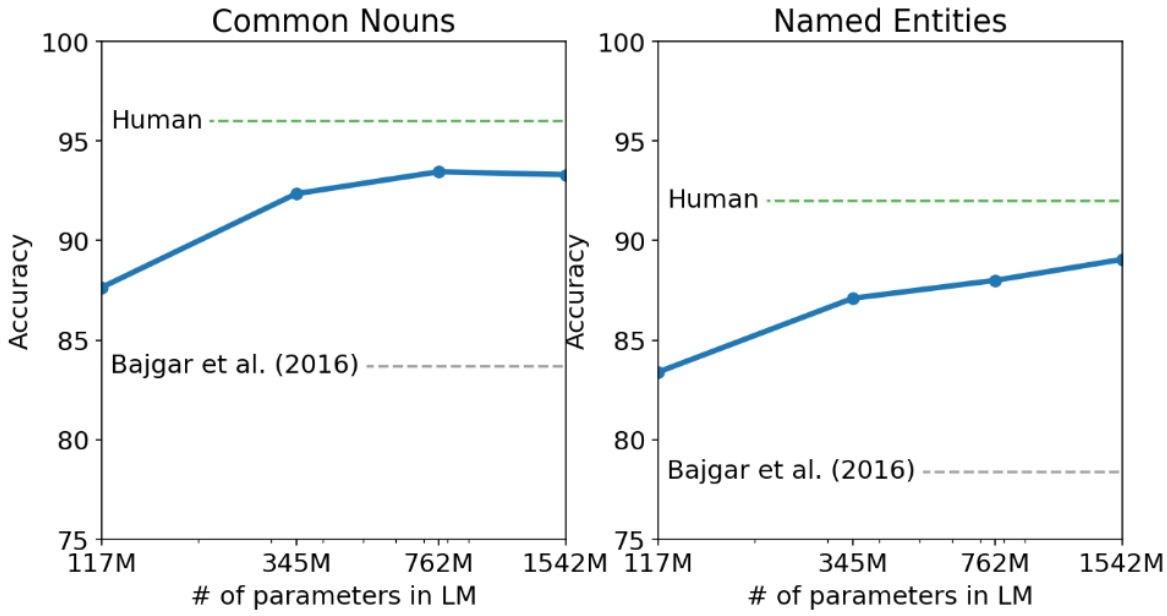
Experiment

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

- SOTA on many datasets (Not most of datasets)
  - Large improvements on small datasets (PTB, WikiText2)
  - Large improvements on long-term dependency (LAMBADA, CBT)
- Worse performance on 1BW [1]
  - It is the biggest dataset
  - Sentence level shuffling to remove long-range structure

[1] Ciprian Chelba et al. “One billion word benchmark for measuring progress in statistical language modeling”. In: *arXiv preprint arXiv:1312.3005* (2013).

Children's Book Test



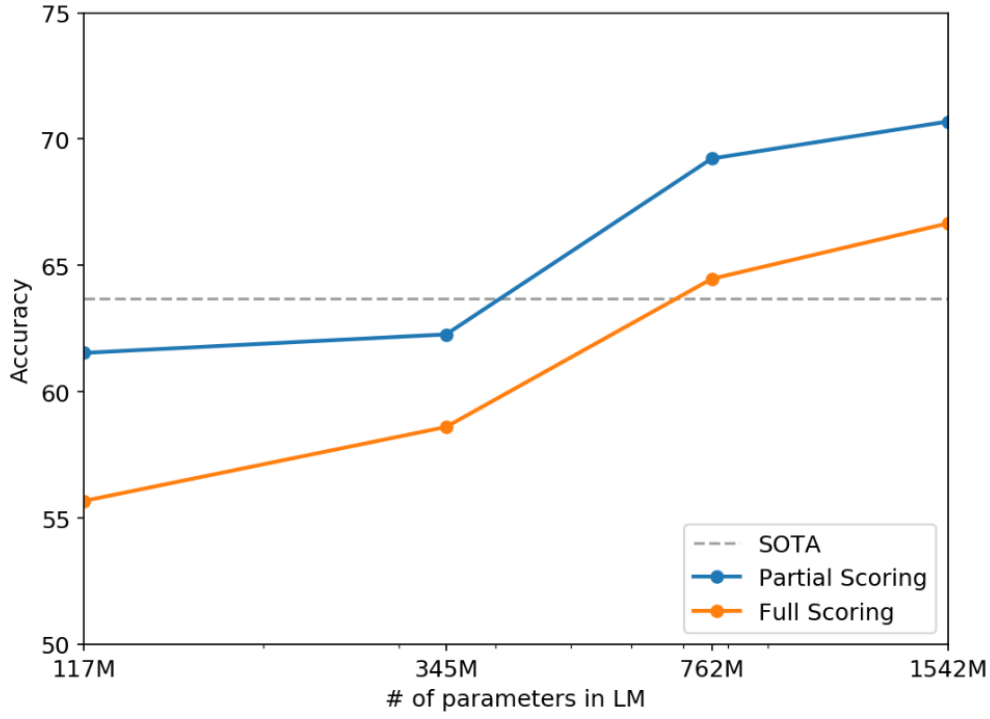
- Predicting missing words in the sentence
- Common Nouns (CN)
  - Dog, Cat, Window, ...
- Name Entities (NE)
  - Alice, Bob, Jack, ...

LAMBADA

- Test ability to handle long-range dependencies
  - Prediction of the final word in a sentence

- Stop-Word filtering
  - The predicted word must logically end the sentence
  - GPT-2 was not leveraging the constraint
  - By excluding words like ‘and’, ‘the’, GPT-2 achieved SOTA

Winograd Schema Challenge



- Commonsense reasoning by resolving textual ambiguities.
  - "The trophy doesn't fit in the suitcase because it is too small."
  - What does 'it' refer to? (Answer: "the suitcase.")

Reading Comprehension

- Conversation Question Answering dataset (CoQA) [1]
  - Dialogues between a question asker and a question answerer
  - Answer questions that depend on conversation history
- Achieved 55 F1 Score
  - Supervised SOTA is nearing the 89 F1
  - GPT-2 used simple retrieval based heuristics

---

[1] Siva Reddy, Danqi Chen, and Christopher D Manning. "Coqa: A conversational question answering challenge". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 249–266.

Summarization

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

- CNN and Daily Mail dataset summarization
  - The text “TL;DR:” is appended as a hint for the task
  - ROUGE 1,2,L metrics (overlap with reference summary)
- Poor performance
  - TL;DR: played important role for task recognition

Translation

- BLEU (Bilingual Evaluation Understudy)
  - Precision about matching n-grams
  - Brevity penalty for overly short translations
- WMT-14 Dataset (English-French)
  - English to French: BLEU score 5
  - French to English: BLEU score 11.5
- Poor performance
  - French data was not enough in WebText

Question Answering

- Natural Questions dataset [1]
  - "Who is the president of the U.S.?"
- GPT-2 Performance
  - 4.1% accuracy when evaluated by exact match metric
  - Five times better than the smallest model
  - Still much worse than supervised systems

---

[1] Tom Kwiatkowski et al. "Natural questions: a benchmark for question answering research". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 453–466.

Generalization vs Memorization

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

- There may be overlap between train-test set
  - Bloom filters containing 8-grams
- Overall Overlap Analysis
  - Overlap provides a small but consistent performance benefit
  - Overlap is not significantly larger than the train-test overlap
- Testing for Memorization
  - Performance improves as model size increases
  - It indicates underfitting rather than over-memorization

Conclusion

- Unsupervised Task Learning
  - Performs well with sufficiently large and diverse dataset
  - Possibility of learning without fine-tuning
- Limitation
  - Far from practically usable for most tasks
  - GPT-2 still struggles with structured reasoning