Language Models as Unsupervised Multi-task Learners

Scaling Up Zero-Shot Task Transfer with GPT-2

By Kayvan Shah

Background

Motivation

- LLMs demonstrates strong performance across various natural language processing tasks without task-specific training.
- GPT-2 was trained on WebText, a large dataset of web pages, using unsupervised learning.
- Despite not being trained on specific datasets like CoQA, GPT-2 generates answers that match or surpass supervised baseline systems.
- Performance scales logarithmically with model size, with the 1.5B parameter GPT-2 model achieving state-of-the-art results.
- This showcases the potential for creating more versatile language AI systems that learn and perform tasks from natural data without extensive supervision.
- The key innovation lies in the remarkable zero-shot multitask transfer capabilities of large unsupervised language models, indicating a path towards more flexible and general natural language AI.

Introduction

- Machine learning systems excel at specific tasks they are trained on, but lack generalization ability
- Current systems are narrow experts rather than competent generalists
- Dominant approach is training on task-specific datasets, which limits generalization
- Multitask learning shows promise, but is difficult to scale with current dataset creation methods
- Language models pre-trained on broad data are an alternative approach

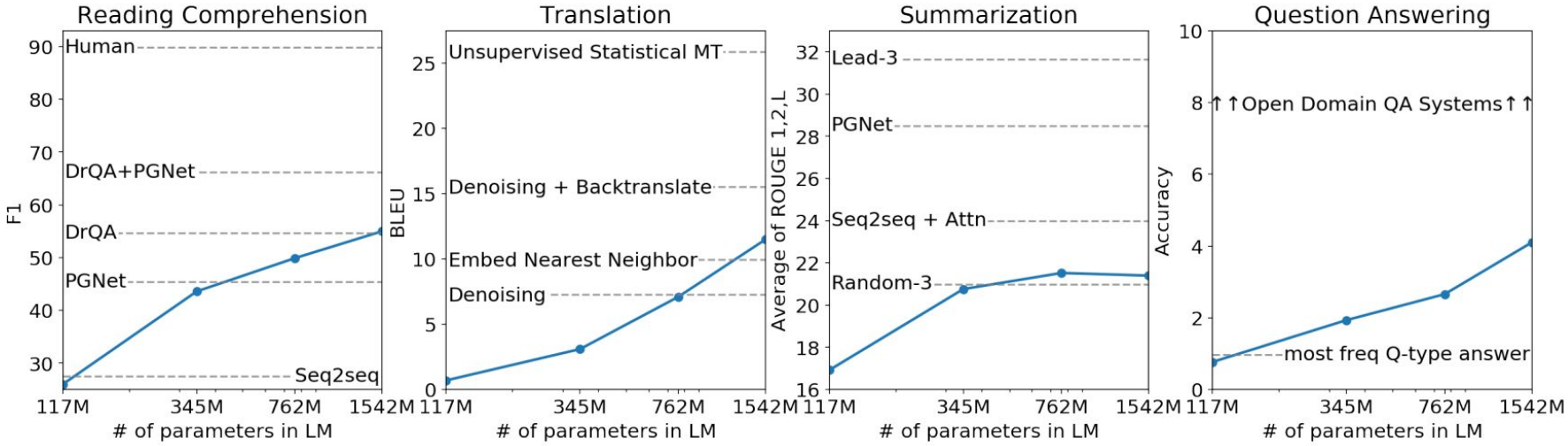


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural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

Introduction (cont...)

- Prior work showed word vectors, contextual reps from RNNs, and self-attention can transfer usefully
- This paper explores language models performing downstream NLP tasks zero-shot without fine-tuning
- Demonstrates potential for more general transfer by achieving competitive results across many tasks
- Key innovation being explored is using LLMs trained on broad data as unsupervised multi-task learners that can generalize to different NLP tasks without specialized supervision or fine-tuning.
- This contrasts with the dominant paradigm of narrow task-specific training.

Approach

About

- The central idea is to use the traditional language modeling objective of predicting the next token in a sequence as a form of unsupervised multitask learning.
- By training on a large corpus of internet data containing many different tasks naturally expressed in text, the hypothesis is that a sufficiently large language model will learn to perform those tasks in order to better model the data.
- This allows evaluating the zero-shot multitask transfer abilities simply by prompting with task descriptions and inputs.

About

- Language modeling estimates probability distribution over sequences

$$P(x) = \Pi\ P(s\_i \mid s\_1..s\_i\text{-}1)$$

- Tasks can be formulated as estimating P(output | input, task specification)
- Prior Work: Demonstrated multitask learning by encoding task, input, and output as sequences.
- Unsupervised Learning: Language modeling objective equivalent to supervised objective on subset.
- Preliminary Experiments: LLMs can perform unsupervised multitask learning, albeit slower.
- Hypothesis: Very LLMs trained on diverse data will learn to infer and perform tasks from natural language sequences.
- Evaluation: Test LLMs zero-shot on various NLP tasks without fine-tuning.

Training Dataset

- Most prior LMs trained on single domain (news, Wikipedia, books)
- This approach motivates using large and diverse dataset across many domains
- Created new WebText dataset by scraping all outbound Reddit links with ≥ 3 karma
- 45M links, extracting text gives 8M docs and 40GB after cleaning
- Excludes Wikipedia to avoid overlap with evaluation datasets

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [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 coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

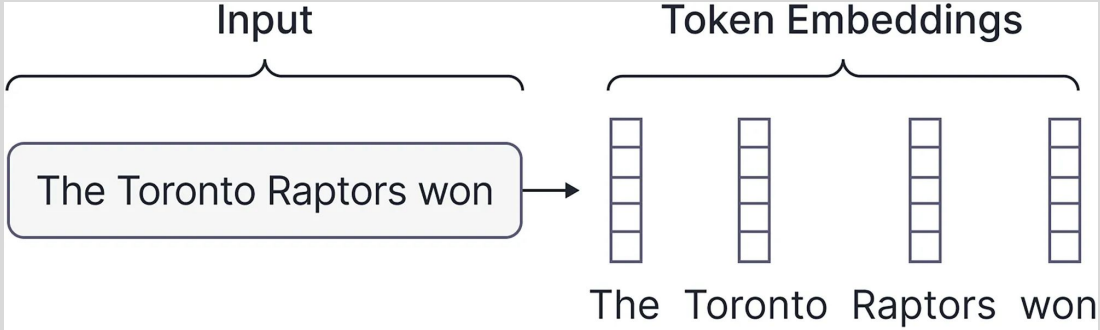
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**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.

Input Representation

- Need generalized input rep to model any Unicode string
- Byte-level too slow, word-level too restrictive
- Use byte-pair encoding (BPE) on bytes with modifications
- Prevents BPE merges across char categories except spaces
- Allows word-level efficiency with byte-level generality
- Can assign probability to any Unicode string regardless of tokenization



Experiments

Trained and evaluated 4 Transformer language models of increasing size (117M to 1.5B params)

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

Table 2. Architecture hyperparameters for the 4 model sizes.

- GPT-2 Model: Over 10x more parameters than original GPT.
- Learning Rate Tuning: Manually adjusted for optimal perplexity on 5% held-out sample of WebText.
- Underfitting: All models still underfit WebText dataset.
- Perplexity Improvement: Held-out perplexity has not improved despite additional training time.

Language Modeling

- Evaluated zero-shot on many language modeling benchmarks without fine-tuning
- Used invertible de-tokenizers to remove dataset artifacts
- Achieved new state-of-the-art on 7 out of 8 datasets
- Large gains on small datasets like PTB, WikiText-2
- Also improved on benchmarks measuring long-range dependencies like LAMBADA, CBT
- Still underperforms on 1 Billion Word Benchmark due to aggressive shuffling

Key Takeaways

- Very large language models trained just on web data transfer remarkably well zero-shot
- Demonstrate strong generalization abilities across diverse domains and tasks
- Model scale is critical, with largest 1.5B model performing best
- Still have room for improvement on extremely shuffled/processed datasets

	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

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Language Modeling Results

Task related Performance

Children's Book Test:

- Tests ability to predict different word types (entities, nouns, verbs, etc.)
- GPT-2 achieved new state-of-the-art, closing gap to human performance
- 93.3% on common nouns, 89.1% on named entities

LAMBADA:

- Tests comprehension requiring long-range context
- GPT-2 improved state-of-the-art perplexity from 99.8 to 8.6
- Improved accuracy from 19% to 63.24% with stop-word filter

Reading Comprehension (CoQA):

- GPT-2 achieved 55 F1, matching 3/4 supervised baselines
- But likely using retrieval heuristics, not full comprehension

Summarization (CNN/DailyMail):

- Generated summaries approach classic neural baselines on ROUGE
- But often focus on recent details, confuse specifics
- Shows some ability to exhibit task behavior with prompting

Task related Performance (cont...)

Winograd Schema Challenge:

- Tests commonsense reasoning to resolve ambiguities
- GPT-2 improved state-of-the-art accuracy by 7% to 70.7%

Reading Comprehension (CoQA):

- GPT-2 achieved 55 F1, matching 3/4 supervised baselines
- But likely using retrieval heuristics, not full comprehension

Translation (WMT English-French, French-English):

- En->Fr achieved 5 BLEU, slightly worse than unsupervised word translation.
- Fr->En achieved 11.5 BLEU, outperforming some prior unsupervised MT baselines.
- Still much worse than current state-of-the-art 33.5 BLEU.
- Surprising given limited French data in WebText.

Question Answering (Natural Questions):

- GPT-2 answered 4.1% exactly correct, 5x better than majority baseline.
- Shows capacity is key limitation of prior neural approaches.
- 63.1% accurate on questions it is most confident about.
- Still much worse than hybrid IR+extractive QA systems.

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%

Table 6. Percentage of test set 8 grams overlapping with training sets.

- Conducted analysis to estimate degree of memorization vs generalization
- Found modest overlap providing small consistent benefit
- But evidence suggests models are still significantly underfitting and generalizing
- Generating fictitious content also demonstrates generalization capabilities

Related Work

- Builds on previous work scaling up language models and leveraging unlabeled data
- Connects to broader line of research on transfer learning and self-supervised pre-training in NLP
- Extends observed trends to larger scale and evaluates emergent abilities

Discussion & Conclusion

Discussion

- Results suggest promising potential for unsupervised task learning.
- Zero-shot performance competitive with supervised baselines in some tasks like reading comprehension.
- Significant gap remains in practical usability for many tasks.
- Fine-tuning may improve performance beyond zero-shot.

Conclusion

- Large LMs perform well across diverse domains and datasets.
- GPT-2 achieved new state-of-the-art on 7/8 language modeling benchmarks.
- Demonstrates emergence of task abilities from broad data.
- Highlights possibility of more general language understanding.

What is the key bottleneck identified for realizing the promise of multi-task learning with current techniques?

A. Lack of large, broad datasets capturing many tasks
B. Each (dataset, objective) pair is a single training example
C. Multi-task interference between disparate tasks
D. Computing infrastructure limitations

Each (dataset, objective) pair is a single training example

THANK YOU