Language Models: Unsupervised Multitask Learners

Presented by: Mahsa Sheikhi Karizaki

PennState

Problem Motivation

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models, reading comprehension systems, and image classifiers on the diversity and variety of possible inputs highlights some of the short-comings of this approach.

Objective

- The authors demonstrate that language models begin to learn tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText.

- They demonstrate language models can perform down-stream tasks in a zero-shot setting without any parameter or architecture modification. This approach shows potential by highlighting the ability of language models to perform a wide range of tasks in a zero-shot setting.

Related Work

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples.

- Multitask learning is a promising framework for improving general performance. Recent work reports modest performance improvements and the two most ambitious efforts to date have trained on a total of 10 and 17 (dataset, objective) pairs respectively.

- The current best performing systems on language tasks utilize a combination of pre-training and supervised fine-tuning.

- another line of work has demonstrated the promise of language models to perform specific tasks, such as commonsense reasoning and sentiment analysis

Dataset

Most prior work trained language models on a single domain of text, such as news articles or Wikipedia. This approach motivates building as large and diverse a dataset as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible.

Instead, they created a new web scrape which emphasizes document quality. To do this they only scraped web pages which have been curated/filtered by humans.

Dataset

"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

Model

A Transformer based architecture is used for the LMs. The model largely follows the details of the OpenAI GPT model with a few modifications. The authors trained and benchmarked four LMs.

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

Performance

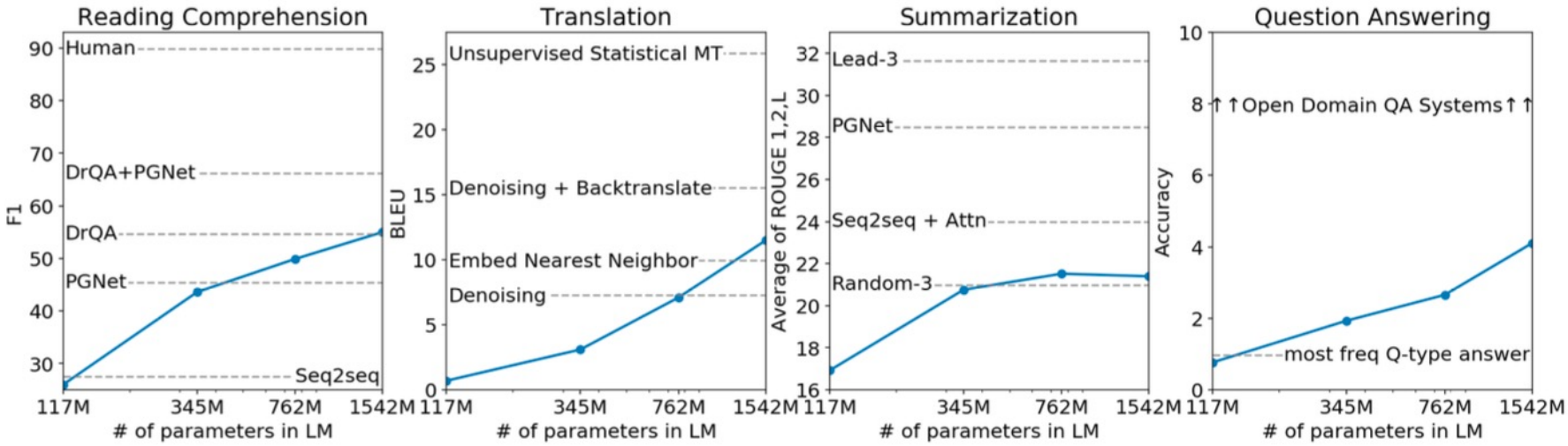

Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA, translation on WMT-14 Fr-En, summarization on CNN and Daily Mail, and Question Answering on Natural Questions.

Results

As an initial step towards zero-shot task transfer, we are interested in understanding how WebText LM's perform at zero-shot domain transfer on the primary task they are trained for: "language modeling".

	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.

Children's Book Test

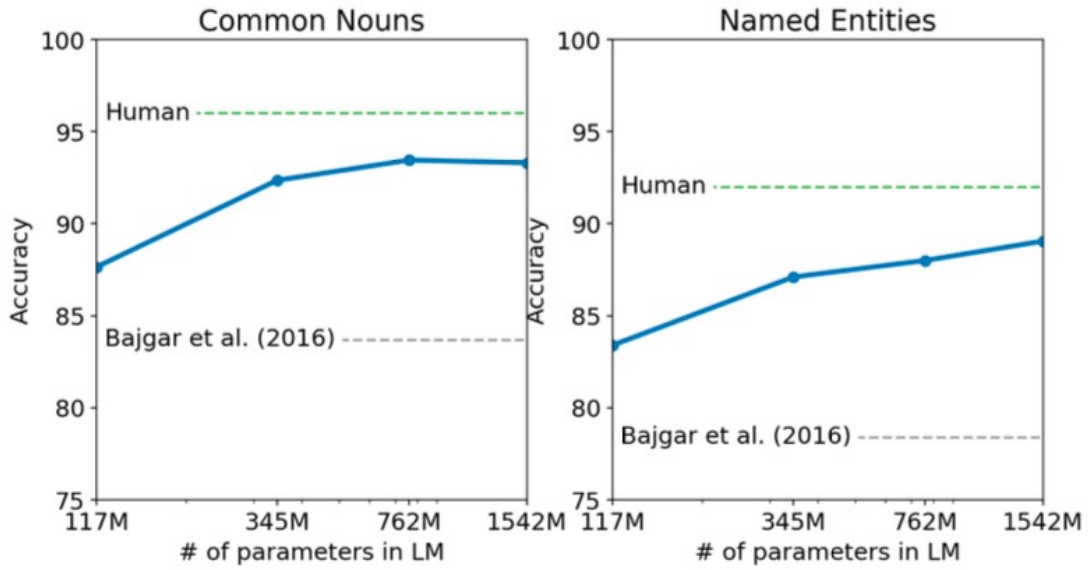

Figure 2. Performance on the Children's Book Test as a function of model capacity.

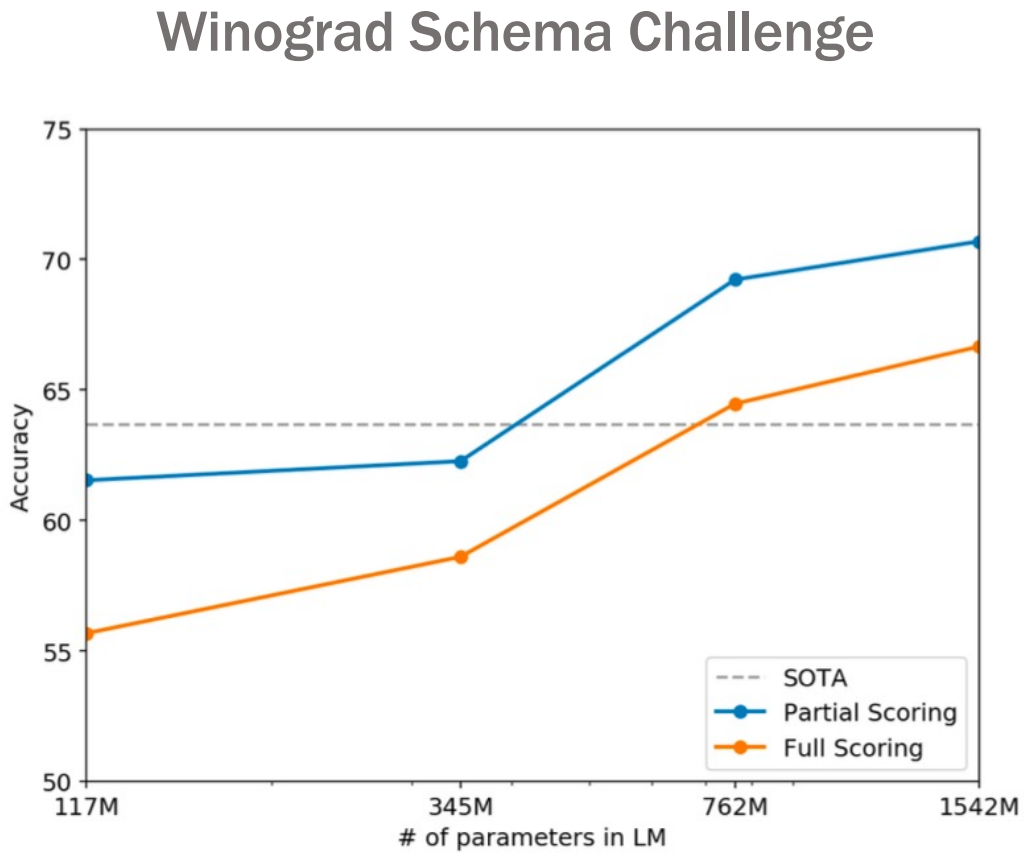

Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

Question Answering

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 4. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2.

Results on Other Tasks

On **reading comprehension**, the performance of GPT-2 is competitive with supervised baselines in a zero-shot setting. However, on other tasks such as **summarization**, while it is qualitatively performing the task, its performance is still only rudimentary according to quantitative metrics.

Generalization vs Memorization

Recent work in computer vision has shown that common image datasets contain a non-trivial amount of near-duplicate images.

	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 5. Percentage of test set 8 grams overlapping with training sets.

Thank you.

Contact Information:
mfs6614@psu.edu

PennState