Language Models are Unsupervised Multitask Learners (GPT-2)

Sung Eun Lee

August 27, 2024

Seoul National University

Contents

❶ Introduction

❷ Structure

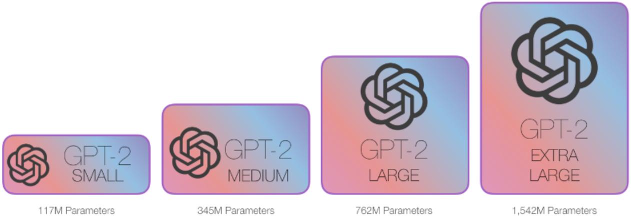
❸ Experiment

❹ Conclusion

Introduction

Introduction

- At the time this paper was written, machine learning systems were highly sensitive to changes in data distribution and the tasks they needed to perform.
- Machine learning systems at that time were specialized for specific tasks they were designed to perform, rather than showing generally good performance across all tasks.
- The paper proposes GPT-2, a language model capable of performing downstream tasks directly in a zero-shot learning setting, without any modification to its parameters or architecture.



Structure

GPT-2 Structure

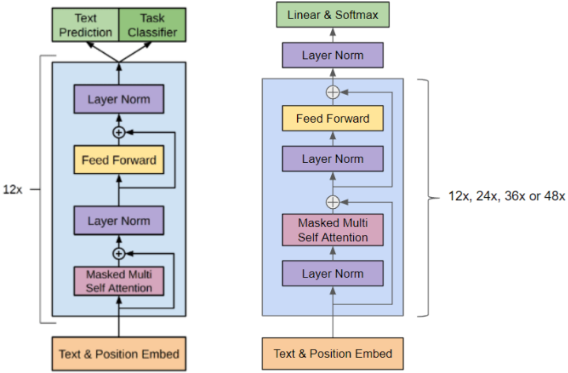
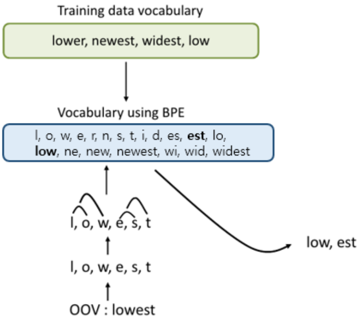


Figure : Comparison of GPT-1 (left) and GPT-2 (right) Architectures.

- Layer normalization has been moved to the input part of the sub-block. and it is applied to the output of the final self-attention block.
- In GPT-2, the vocabulary size increased to 50,257, the context vector size increased to 1024, and the batch size increased to 512.

Token


Training data vocabulary
lower, newest, widest, low
Vocabulary using BPE
l, o, w, e, r, n, s, t, i, d, es, **est**, lo,
low, ne, new, newest, wi, wid, widest
l, o, w, e, s, t
l, o, w, e, s, t
OOV : lowest
low, est

- The paper proposes converting Unicode strings to UTF-8 for processing at the byte level. However, it observes that byte-level language models underperform compared to word-level language models when trained on large-scale datasets.
- language models needed a way to handle **Out-Of-Vocabulary(OOV)** issues, and the paper employed the **Byte Pair Encoding (BPE)** method to address this.

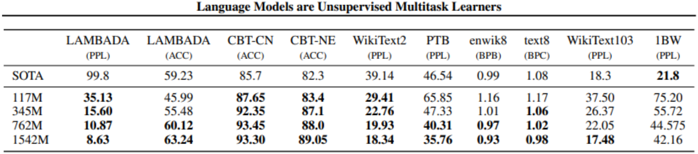
Experiment

Set up

- The smallest model is the same size as **GPT-1**, while the second-largest model is equivalent in size to **BERT-LARGE**.
- The learning rate was manually adjusted using a held-out sample comprising 5% of WebText. The authors claim that all four models **underfit** WebText, indicating that investing more training time could lead to better performance.

Experiment

Language Modeling



Language Models are Unsupervised Multitask Learners

	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

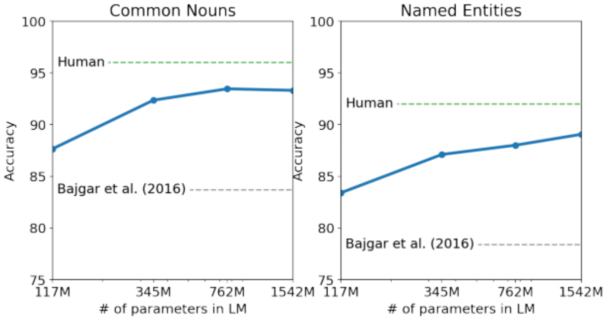
- GPT-2 could be applied to any language model benchmark and was evaluated using the scaled negative log likelihood loss or exponentiated version of the average negative log probability based on the WebText language model.

- GPT-2 also achieved remarkable performance improvements on datasets designed to **measure long-term dependencies** in language models, such as LAMBADA and CBT.

- However, a **performance decline** was observed on the 1BW dataset. this is due to the 1BW dataset being the largest and its destructive pre-processing, which removes long-range structures.

Experiment

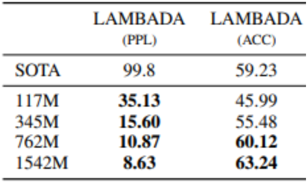
1. Children's Book Test



- ▶ The CBT dataset does **not use perplexity** as an evaluation metric. Instead, it employs **accuracy** as the evaluation metric, where a cloze test is administered by providing a blank and 10 possible choices, and the task is to select the correct word to fill in the blank.
- ▶ As a result, GPT-2 achieved SOTA performance with 93.3% accuracy on common nouns and 89.1% accuracy on entities.

Experiment

2. LAMBADA

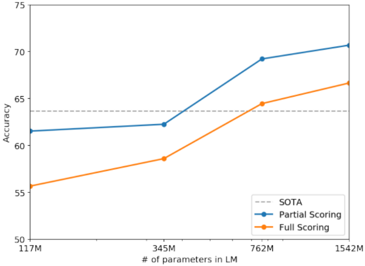


	LAMBADA (PPL)	LAMBADA (ACC)
SOTA	99.8	59.23
117M	**35.13**	45.99
345M	**15.60**	55.48
762M	**10.87**	**60.12**
1542M	**8.63**	**63.24**

- The LAMBADA dataset is designed to measure the long-range dependency of language models.
- It involves predicting the last word of each paragraph, and GPT-2 achieved SOTA performance in terms of **perplexity** and **accuracy** on this task.

Experiment

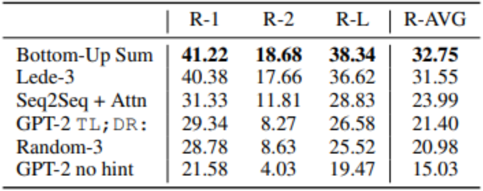
3. Winograd Schema Challenge



- The Winograd Schema Challenge dataset is used to measure a model's ability to resolve ambiguity.
- GPT-2 achieved SOTA performance according to the graph, and the final Extra Large model reached 70.7% which is 7% higher than the previous models.

Experiment

4. Summarization

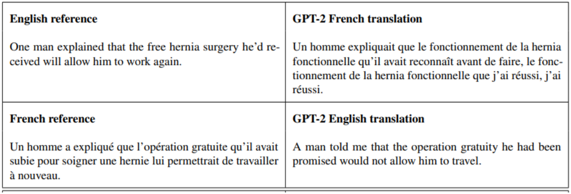


	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

- To evaluate GPT-2's summarization capabilities, the CNN and Daily Mail dataset was used. However, it was observed that GPT-2 does **not perform** well on such summarization tasks.

Experiment

5. Translation

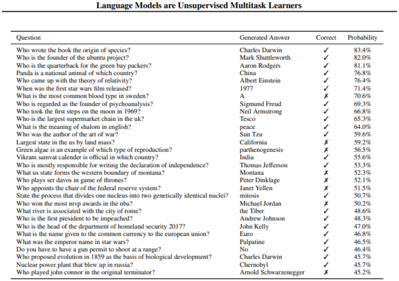


English reference	GPT-2 French translation
One man explained that the free hernia surgery he'd received will allow him to work again.	Un homme expliquait que le fonctionnement de la hernia fonctionnelle qu'il avait reconnaît avant de faire, le fonctionnement de la hernia fonctionnelle que j'ai réussi, j'ai réussi.
French reference	**GPT-2 English translation**
Un homme a expliqué que l'opération gratuite qu'il avait subie pour soigner une hernie lui permettrait de travailler à nouveau.	A man told me that the operation gratuity he had been promised would not allow him to travel.

- On the WMT-14 English-French test set, GPT-2 achieved a BLEU score of 5, which is **lower than the performance** reported in previous unsupervised translation research.
- On the WMT-14 French-English test set, GPT-2 performed better, achieving a BLEU score of 11.5. However, this performance is **still not competitive** compared to other models.

Experiment

6. Question Answering



Language Models are Unsupervised Multitask Learners

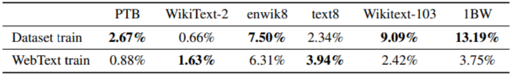
Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	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%

- GPT-2 was evaluated using the exact match metric commonly used in reading comprehension tasks like SQuAD, and it was found to correctly answer 4.1% of the questions. Notably, the smallest model within the WebText LMs did not even surpass 1% accuracy.

- The authors claim that GPT-2 correctly answered 5.3 times more questions, suggesting that the model's capacity was a significant factor contributing to the neural system's poor performance on this type of task.

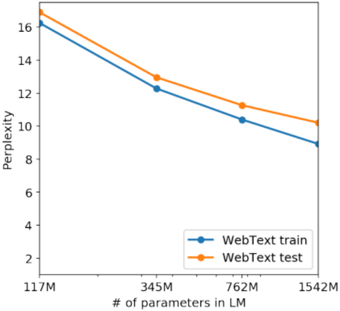
Results

Percentage of test set 8 grams overlapping with training sets.



	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%

The performance of LMs trained on WebText as a function of model size.



- As seen in the graph above, it was confirmed that larger model sizes lead to better performance. but it is speculated that the performance improvement due to data overlap is likely minimal.

Conclusion

Conclusion

1. If a large language model is trained on sufficiently large and diverse datasets, it can achieve strong performance across many domains and datasets.

2. GPT-2 is a model that enhances zero-shot task performance by using a language model that predicts the next token, trained on the vast WebText dataset.

3. This suggests that high-capacity models can learn to perform a variety of tasks without supervision.