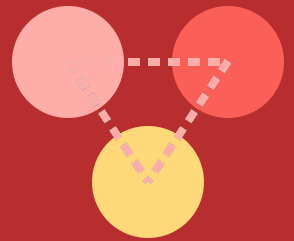
GRADUATE PRESENTATION

Language Models are Unsupervised Multitask Learners

Marzia Tahsin



ARTIFICIAL INTELLIGENCE

Summary

High-level summary that explains the key points of the paper without going into unnecessary detail, but explains concepts

Summary

TRADITIONAL APPROACH

Most natural language processing tasks, such as answering a question or machine translation, are usually done through using supervised learning on task-specific datasets.

THIS PAPER

This paper is focusing on showing that a language model can still learn to do different tasks without any supervised training when it is trained on a large varying dataset.

Summary | WebText Dataset Overview

COLLECTION AND SCALE

8M

Webpages

The dataset created by the authors containined 8 million webpages collected from links shared on Reddit.

CURATION FILTER

≥ 3

Karma Score

Webpages were required to have at least 3 karma, this used Reddit's social ranking scale

QUALITY CONTROL

Filter

Redundant Content

This method filtered out redundant content to make sure only higher quality content was retained.

Performance and results | Zero-Shot Model Scaling

MODEL SCALE

Size ↑

Capacity Expansion

As the language model increases in parameter size its capabilities also expanded

ZERO-SHOT SETTING

0-Shot

No Task training

The model performed well without being specifically trained or fine-tuned for specific tasks

TASK EXECUTION

Better

Multi-Task Results

It shows direct improvement across a broad variety of different downstream tasks

Performance & Benchmarks | Zero-Shot Language Modeling

7/8
SOTA RESULTS

GPT-2 achieved state of the art zero shot results on **7 out of 8** language modeling datasets tested without explicit task-specific fine-tuning

Dataset evaluation matrix

	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

ACC (Accuracy)
Higher score is better

PPL (Perplexity)
Lower score is better

BPB / BPC (Bits Per Byte/Char)
Lower score is better

Strength

A few strengths of the paper compared to other work in the field

No extra training require(Zero-Shot Learning)

In **“Improving Language Understanding by Generative Pre-Training”** (GPT-1), the model required two steps: general pre-training followed by supervised fine-tuning on labeled datasets for each specific task. In the paper I’m reviewing the authors showed that a large language model can perform tasks like translation and questioning right out of the box (zero-shot) using natural language prompts without any fine-tuning or parameter updates.

Reading any text without missing words (Byte-Level BPF)

Operating at the byte level lets GPT-2 handle any string of text without ever hitting an “unknown token”, which expands on the model approach introduced in “**Learning to Generate Reviews and Discovering Sentiment”** into a much larger, general purpose language model

Weaknesses

A few weaknesses of the paper compared to other work in the field

Falls Short on Specialized Tasks

GPT-2 was designed to have a broader purpose which it does well but when it comes to the more specialized tasks it tends to fall through. Models like the machine translation in **“Attention Is All You need”** is much stronger at doing direct translations than GPT-2 due to its limitations.

Data Quality is a proxy, not a guarantee

Filtering WebText by Reddit karma is a weaker substitute for real content than the method used in **“Improving Language Understanding by Generative Pre-Training” (GPT-1),** whose BooksCorpus dataset was drawn from actual published books rather than a social ranking system.

High Karma does not guarantee that a post is accurate, well-written or a good representation of language, it will favor what users find fascinating.

This will favor certain topics , demographic and viewpoints which could introduce bias into the training. This is something that the paper doesn’t go into much detail about.

Extension

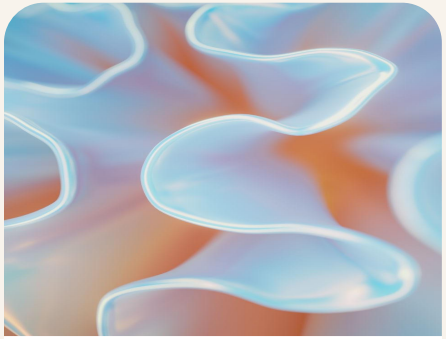
Identify a few examples of how the work can be extended or adapted

Where could this paper go next...

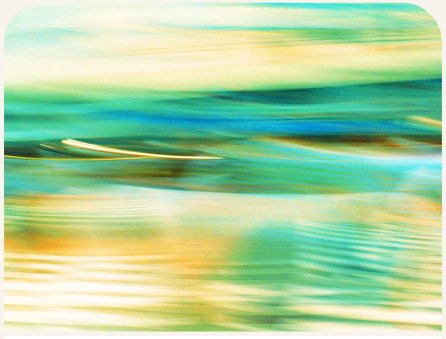
The WebText dataset could be expanded beyond the scope of Reddit and into more sources like books, academic journals and even multilingual text which could help with the bias that was introduced by doing the Karma based filtering, similar to the way GPT-1's BooksCorpus did it. Beyond this, it could also explore even bigger models since it didn't have much improvement at 1542M parameters like what GPT-3 eventually did.

References

Key Publications

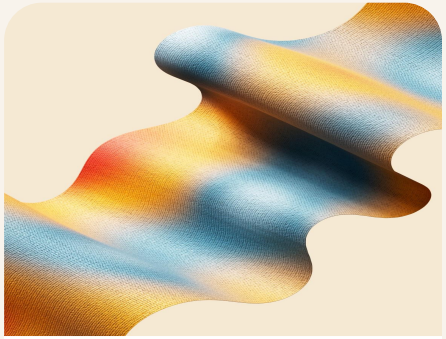


Radford, A., Jozefowicz, R., & Sutskever, I. (2017). **Learning to Generate Reviews and Discovering Sentiment.** *arXiv preprint arXiv:1704.01444.*



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). **Attention Is All You Need**. *Advances in Neural Information Processing Systems (NeurIPS), 30.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). **Language Models are Unsupervised Multitask Learners.** *OpenAI.*



Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). **Improving Language Understanding by Generative Pre-Training.** *OpenAI.*

Thank you!

Questions?