Language Models are Unsupervised Multitask Learners

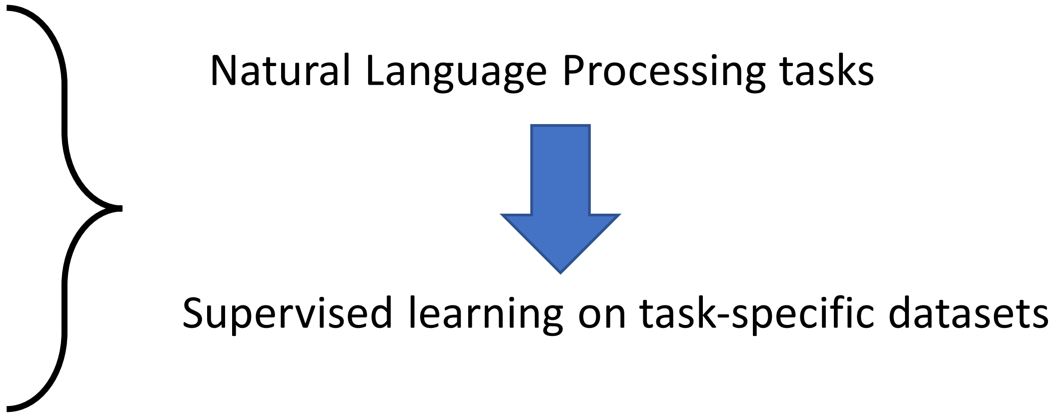
Presentation for Seminar - CS399

M. Maheeth Reddy

1801CS31

Abstract

- Question Answering
- Machine Translation
- Reading Comprehension
- Summarization

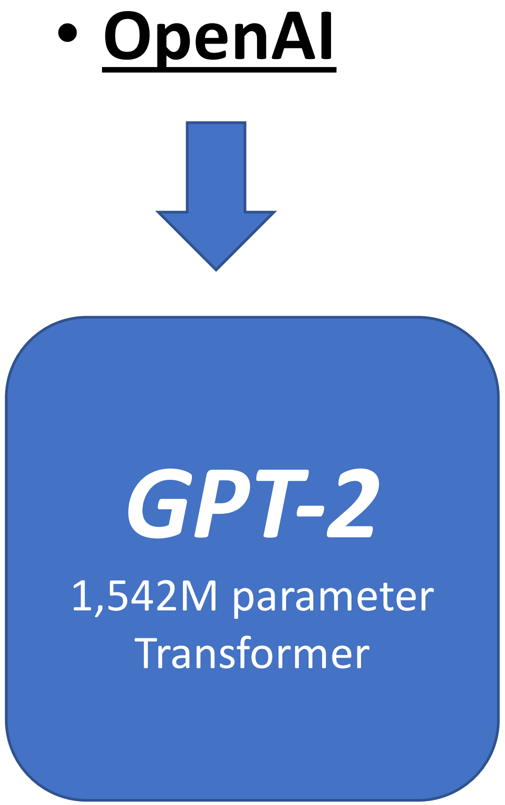


Natural Language Processing tasks

Supervised learning on task-specific datasets

- OpenAI, an AI research lab in San Francisco, have made an attempt to address this problem through unsupervised learning.

OpenAI



- **OpenAI**

GPT-2

1,542M parameter Transformer

- Trained on a dataset of millions of webpages called WebText.
- On CoQA dataset, it achieved 55 F1 score, without any explicit training on the dataset
- It can generate coherent paragraphs of text.
- Findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

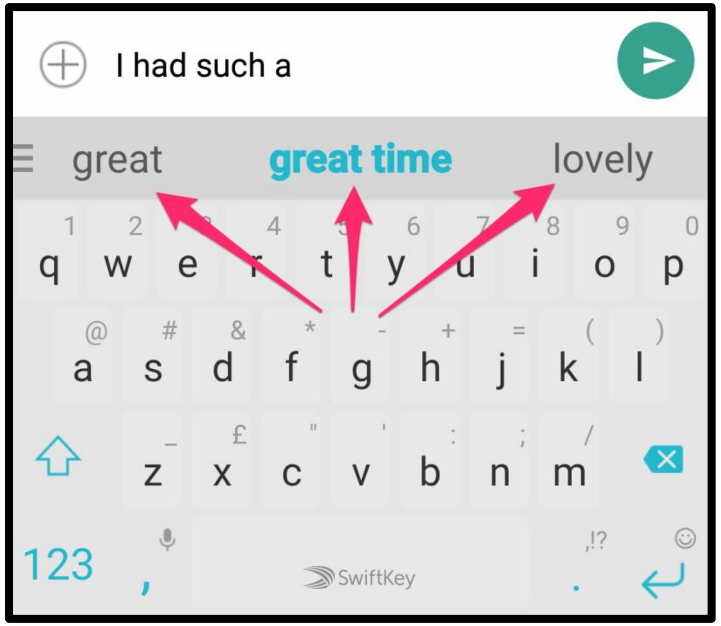
CoQA is a large-scale dataset for building Conversational Question Answering systems.

Introduction

- General procedure to creating ML systems:
  - 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 other examples.
- This has served well to make progress on narrow experts.
- To utilize multitask learning to improve general performance of machine learning models, additional setups are explored.

Approach

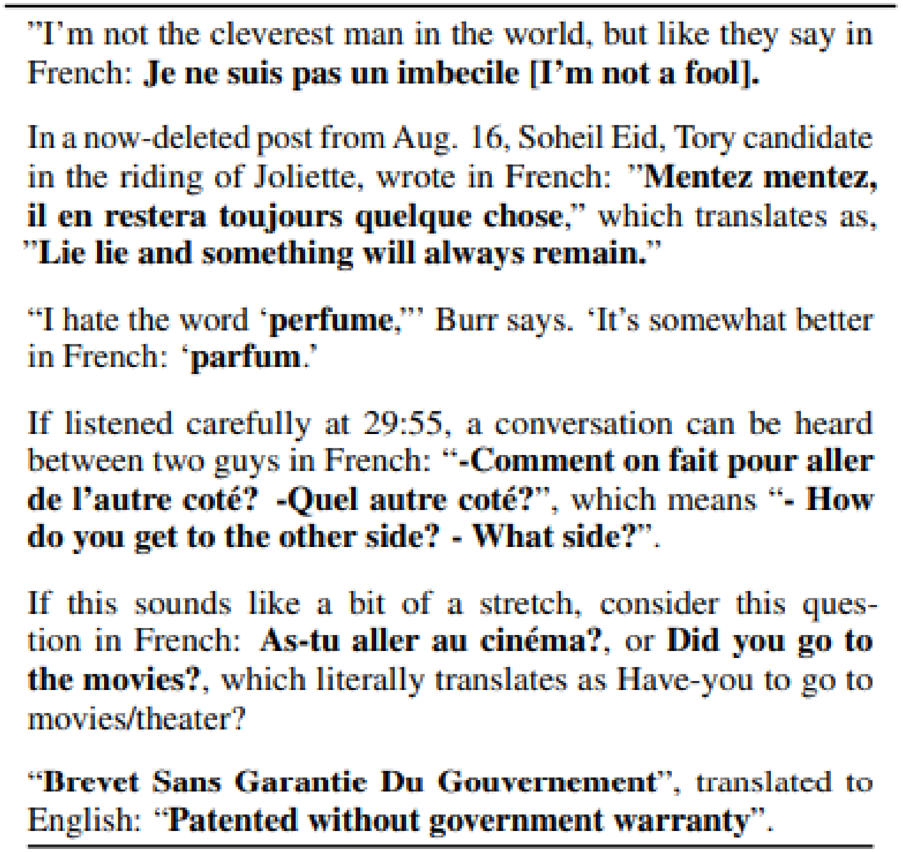
- Language Modeling forms the core of the approach.
- A language model analyzes the pattern of human language for the prediction of words through statistics. It is the core component of modern Natural Language Processing (NLP).
- For GPT-2, Transformer based language model is used. Transformer is used to maintain the wide-range context of words in a paragraph.



I had such a

Training Dataset

- For a multitask learning language model, the training dataset must be large enough so that a wide variety tasks can be performed.
- The dataset made by OpenAI, for GPT-2, called **WebText** is comprised of scrapings of posts from Reddit which have atleast 3 upvotes.
- The number of upvotes is a good but a vague filter of posts which people have found to be interesting.
- **WebText contains 40 GB of text, equivalent to 8 million documents.**



"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.

Details about Model

- Current large scale LMs include pre-processing steps such as lowercasing, tokenization, and out-of-vocabulary tokens which restrict the space of model-able strings.
- GPT-2 utilizes Byte Pair Encoding (BPE) in-order to evaluate the language model on any dataset without the pre-processing.

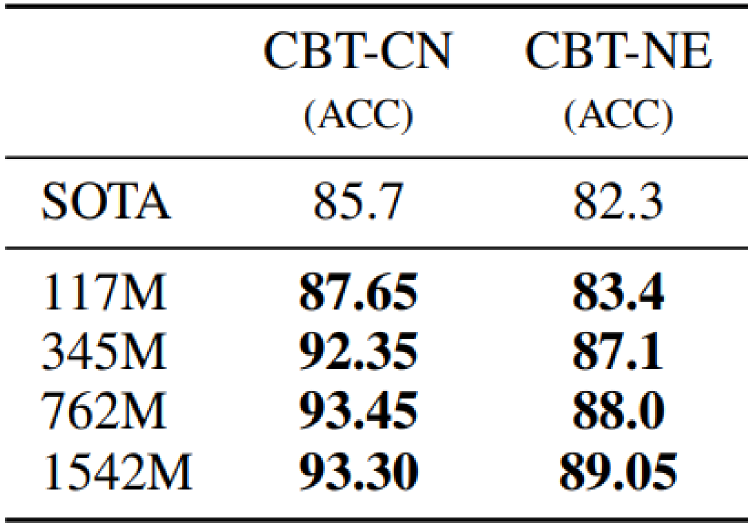
low	**lower**	**lowest**
smart	**smarter**	**smartest**

Experiments on Language Modeling Datasets

- Four Language Models with different no. of parameters were trained as shown here.
- The 117M parameter model is equivalent to original GPT.
- The 345M parameter model is equivalent to largest model from BERT (by Google)
- There is one 762M parameter model.
- The 1542M parameter model is called GPT-2
- All models underfit WebText.

Children’s Book Test (CBT) dataset

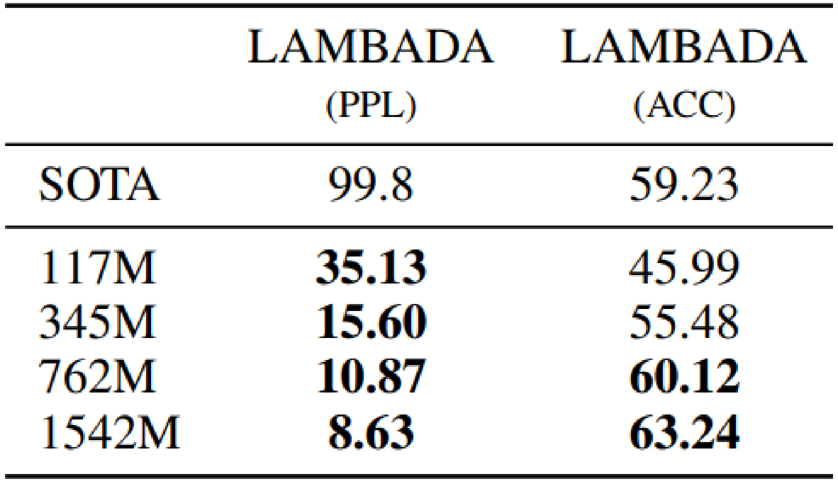
- CBT was designed to test the role of memory and context in language processing and understanding.
- The test requires predictions about different types of missing words in children’s books, given both nearby words and a wider context from the book.
- The prediction accuracy for common nouns and named entities were measured.
- **GPT-2 achieved 93.30% accuracy for common nouns and 89.05% for named entities, hence setting a new benchmark.**



	CBT-CN (ACC)	CBT-NE (ACC)
SOTA	85.7	82.3
117M	**87.65**	**83.4**
345M	**92.35**	**87.1**
762M	**93.45**	**88.0**
1542M	**93.30**	**89.05**

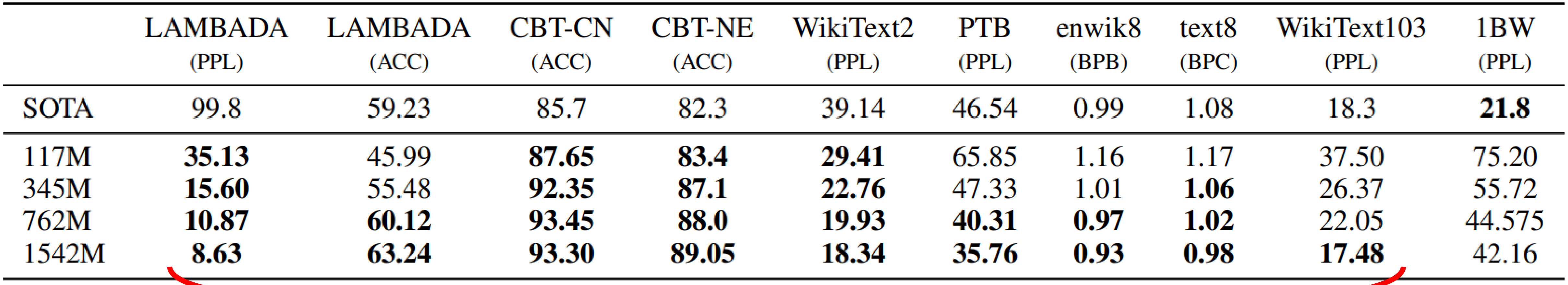
LAMBADA dataset

- It is a collection of narrative passages that requires language models to guess the last word if they are exposed to the whole passage, not just by the last sentence but by a wider context.
- GPT-2 improves the state of the art from 99.8 to 8.63 perplexity and increases the accuracy of LMs on this test from 59.23% to 63.24%.



	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**

Other datasets

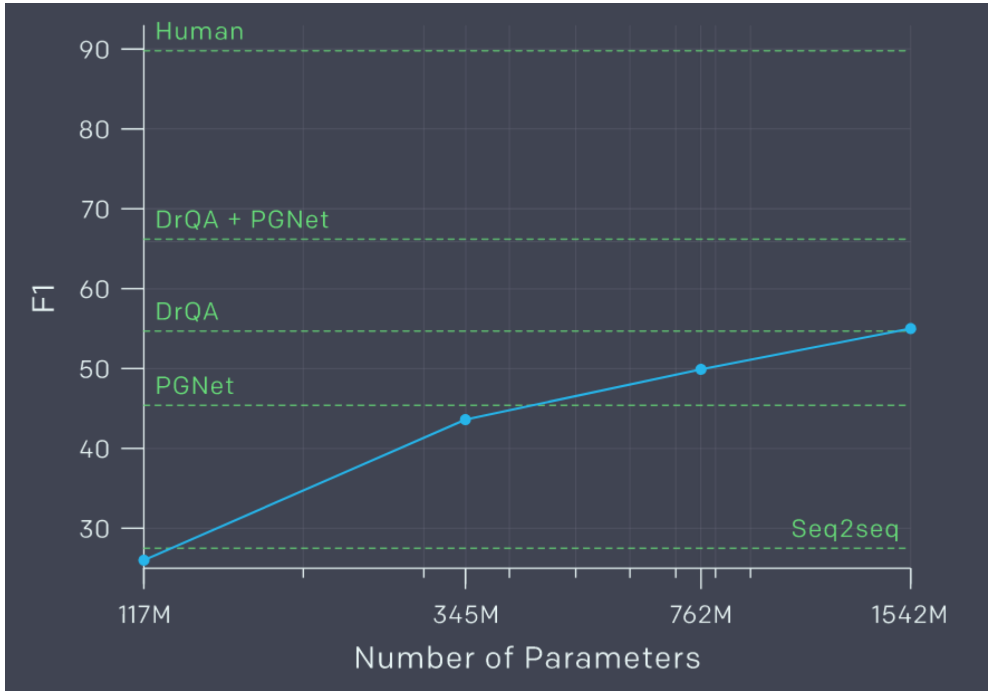


	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**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

Achieved state-of-the-art results on 7 out of 8 language modelling datasets

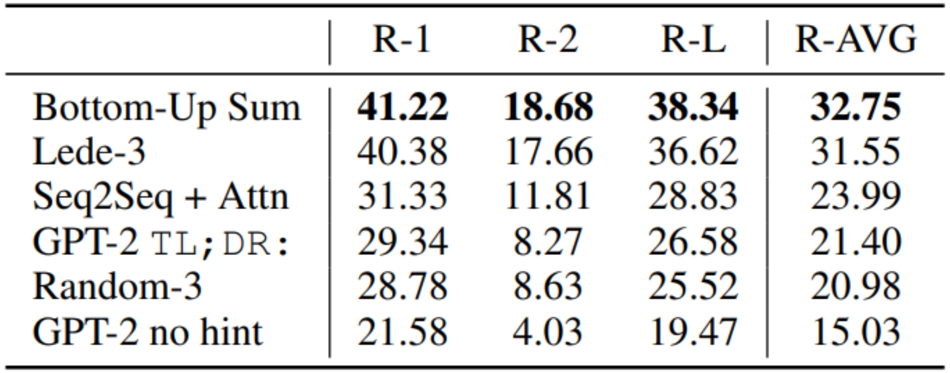
Performance in Reading Comprehension

- The Conversation Question Answering dataset (CoQA) tests reading comprehension capabilities and also the ability of models to answer questions that depend on conversation history.
- GPT-2 achieves 55 F1 score without explicitly training on this dataset.
- This matches or exceeds the performance of 3 out of 4 baseline systems which were trained on CoQA.



Performance in Summarization

- GPT-2's ability to perform summarization was tested on the CNN and Daily Mail dataset, which consists of some unique news articles from the journalists.
- The generated summaries only begin to approach the performance of classic neural baselines.
- The model often focuses on recent content from the article or is confused by specific details such as how many cars were involved in a crash.



	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

Comparison of GPT-2 ROUGE 1,2,L F1 metrics with other models for Text Summarization

Performance in Translation

- GPT-2 doesn’t outperform existing language models in translation.
- On the WMT-14 English-French dataset, GPT-2 performs slightly worse than a word-by-word substitution. (5 BLEU)
- On the WMT-14 French-English dataset, GPT-2 performs worse (11.5 BLEU) than the state-of-the-art benchmark (33.5 BLEU).
- BLEU is a translation performance metric.

Performance in Question Answering

- The Natural Questions dataset was used to evaluate how GPT-2 generates the correct answer to factoid-style questions.
- It was observed that GPT-2 answers 5.3 times more question correctly than the smallest 117M parameter model.
- This suggests that model capacity has been a major factor in the poor performance of neural systems on this kind of task.
- But the performance of GPT-2 is still worse than the 30 to 50% range of open domain question answering systems.

Observations

- Unsupervised learning can be explored for multitask learning language models.
- GPT-2 achieves state of the art performance on 7 out of 8 tested language modeling datasets.
- On reading comprehension the performance of GPT-2 is competitive.
- In case of summarization, GPT-2 is still far from useable for practical applications but is suggestive for research.
- On question answering and translation, language models only begin to outperform trivial baselines when they have sufficient capacity.
- **When a large language model is trained on a sufficiently large and diverse dataset it can perform well across many domains and datasets.**

Language Models are Unsupervised Multitask Learners

Links	
Reference Paper	https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf **https://is.gd/SdhRmT - (Shortened Link)**
GitHub	https://github.com/openai/gpt-2
Web	https://openai.com/blog/better-language-models/

Thank You