Reading Report 7

Language Models are Unsupervised Multitask Learners

Introduction

- Traditional Approach:

  NLP tasks like question answering and translation use supervised learning on large, annotated datasets
- WebText & Zero-Shot Shift:

  GPT-2, trained on millions of web pages (WebText), performs NLP tasks without explicit supervision
- Key Result:

  Achieves 55 F1 on CoQA without using 127k training examples, surpassing 3 of 4 baselines
- Model Capacity Matters:

  Larger models like GPT-2 (1.5B parameters) improve performance across tasks in a log-linear fashion

Towards Generalized ML

From Narrow Experts to Zero-Shot

- Current ML Systems:

  Perform well on narrow tasks but struggle with generalization and data shifts
- Challenges:

  Erratic performance in tasks like captioning and image classification due to single-task training
- Multitask Learning:

  Benchmarks like GLUE show promise, but multitask training is still nascent
- Zero-Shot Potential:

  Pre-trained language models can perform tasks without fine-tuning, showing promise for generalized ML

Core Approach of Language Modeling

- Language Modeling:
  Framed as unsupervised distribution estimation over sequences of symbols
- Conditional Probabilities:
  Factorize joint probabilities into conditional probabilities for tractable sampling, e.g., $p(x) = \prod p(sn|s1,...,sn-1)$
- Self-Attention & Transformer:
  Modern architectures like the Transformer improve conditional probability computation
- Task Learning:
  General models learn tasks as p(output|input, task), leveraging language to specify tasks, inputs, and outputs
- Multitask Learning:
  Language models can infer and perform multiple tasks without explicit supervision

Experimental Setup

- Training:

  Four Language Models (LMs) were trained, including GPT, BERT, and GPT-2, with varying sizes and learning rates optimized for best perplexity on WebText

- Language Modeling:

  GPT-2 excelled at zero-shot transfer across domains, improving perplexity and achieving state-of-the-art results on 7 of 8 datasets

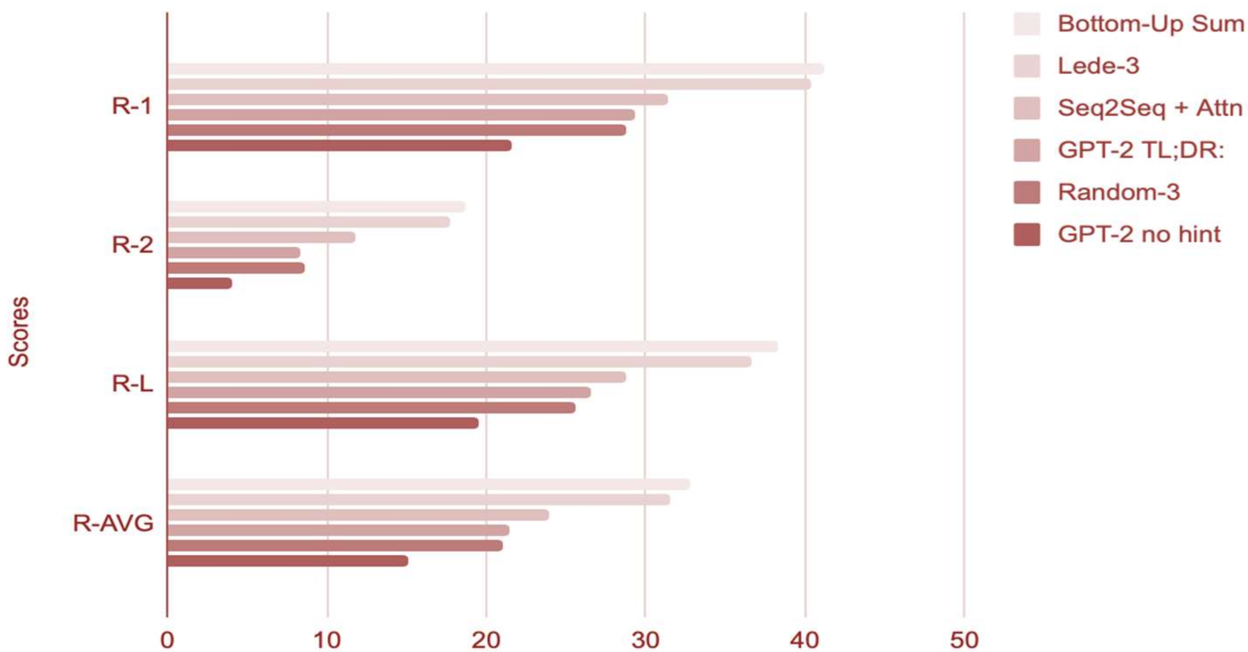
- CBT & LAMBADA:

  Significant performance boosts were seen in Children’s Book Test (93.3% on nouns) and LAMBADA dataset (accuracy from 19% to 52.66%), highlighting GPT-2's ability to handle long-term dependencies

- Challenges:

  While GPT-2 outperformed in many areas, it struggled with datasets like the One Billion Word Benchmark due to heavy preprocessing

Performance Comparision



Generalization v/s Memorization

Aspect	Generalization	Memorization
Definition	Performs well on unseen data	Relies on recalling training data
Dataset Overlap	Low overlap (1-6%)	High overlap (e.g., 13.2% in 1BW)
Impact	Reflects true model capabilities	Inflates reported performance
Example	CoQA (0.5-1.0 F1 gain due to overlap)	GPT-2 shows some memorization, but still underfits WebText
Best Practice	De-duplicate training and test splits	Avoid excessive overlap in datasets

Discussion and Conclusion

- Unsupervised Task Learning:
  Shows potential for models to perform tasks without supervision
- Zero-Shot Performance:
  GPT-2 competes in reading comprehension but underperforms in summarization
- Capacity Matters:
  Performance improves with model capacity; many tasks still fail to exceed random guessing
- Fine-Tuning Exploration:
  Further investigation needed on benchmarks like decaNLP and GLUE
- Overall Insight:
  Large, diverse datasets enable models like GPT-2 to excel across various tasks with minimal supervision

Quiz Question

- **Question:** How do larger language models, trained on vast amounts of data, typically perform in comparison to their smaller counterparts?

A) Smaller models consistently outperform larger ones in all scenarios.

B) The performance of larger models significantly enhances as both the model size and the dataset size increase.

C) Increasing the training data does not provide any benefits for larger models.

D) The performance of larger models is entirely determined by the underlying architecture, not by size or data.

- **Answer:**
    - The Correct Answer is Option B