[Review] (GPT-3) Language Models are Few-Shot Learners

NeurIPS 2020

Hankyo Jeong

August 27, 2024

Seoul National University

Introduction

- Recent years have featured a trend towards **pre-trained language representations** in NLP systems, applied in increasingly flexible and **task-agnostic** ways for downstream transfer.
  - transfomer architecture

- However, a major limitation to this approach is that while the architecture is task-agnostic, there is still a need for **task-specific datasets and task-specific fine-tuning**
  - need for large dataset of labeled examples for every new tasks
  - the potential to exploit spurious correlations in training data grows
  - humans do not require large supervised datasets to learn most language tasks

Introduction

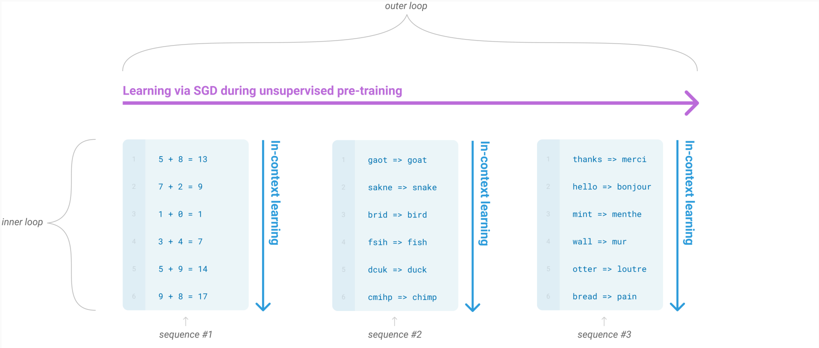


Figure 1: Language model meta-learning: During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "in-context learning" to describe the inner loop of this process, which occurs within the forward-pass upon each sequence.

Introduction

- Another recent trend in LM may offer a way forward: **scaling law**
  - GPT-1: 100 million
  - BERT: 300 million
  - GPT-2: 1.5 billion
  - Megatron-lm: 8 billion
  - T5(google): 11 billion
  - Turing-NLG(MS): 17 billion

- Each increase has brought improvements in text synthesis and/or downstream NLP tasks

- Since in-context learning involves absorbing many skills and tasks within the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains with scale.

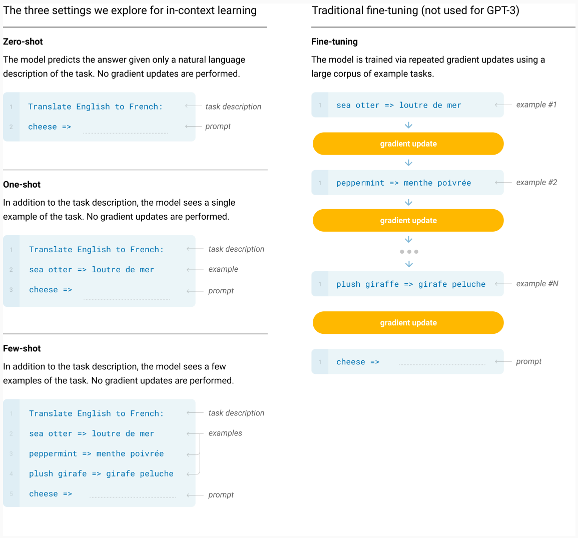
Introduction

- Another recent trend in LM may offer a way forward: **scaling law**
  - GPT-1: 100 million
  - BERT: 300 million
  - GPT-2: 1.5 billion
  - Megatron-lm: 8 billion
  - T5(google): 11 billion
  - Turing-NLG(MS): 17 billion

- Each increase has brought improvements in text synthesis and/or downstream NLP tasks

- Since in-context learning involves absorbing many skills and tasks within the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains with scale.

Approach


The three settings we explore for in-context learning
Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.
1 Translate English to French:
2 cheese =>
task description
prompt
One-shot
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.
1 Translate English to French:
2 sea otter => loutre de mer
3 cheese =>
task description
example
prompt
Few-shot
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.
1 Translate English to French:
2 sea otter => loutre de mer
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese =>
task description
examples
prompt
Traditional fine-tuning (not used for GPT-3)
Fine-tuning
The model is trained via repeated gradient updates using a large corpus of example tasks.
1 sea otter => loutre de mer
example #1
gradient update
1 peppermint => menthe poivrée
example #2
gradient update
1 plush giraffe => girafe peluche
example #N
gradient update
1 cheese =>
prompt

Approach

Context →	Please unscramble the letters into a word, and write that word: skicts =
Target Completion →	sticks

Context →	anli 3: anli 3: We shut the loophole which has American workers actually subsidizing the loss of their own job. They just passed an expansion of that loophole in the last few days: $43 billion of giveaways, including favors to the oil and gas industry and the people importing ceiling fans from China. Question: The loophole is now gone True, False, or Neither?
Correct Answer →	False
Incorrect Answer →	True
Incorrect Answer →	Neither

Context →	Q: What is 556 plus 497? A:
Target Completion →	1053

Approach

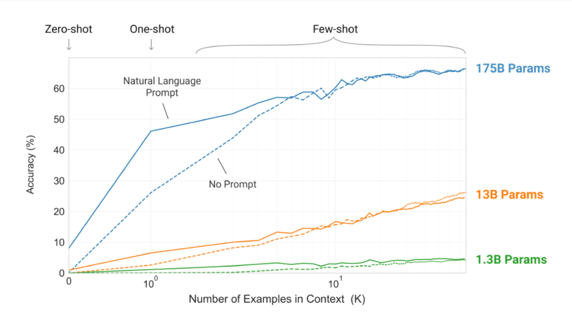
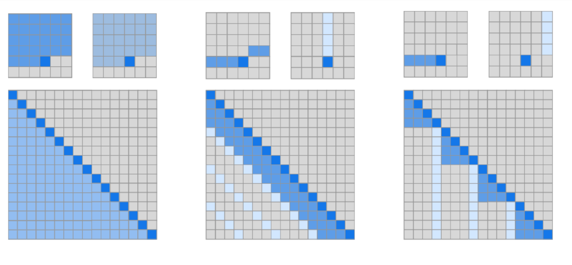


Figure 2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description. The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information.

Approach: Model and Architectures

- **Model and Architectures**
  - Same as GPT-2 with the exception that we use alternating dense and locally banded **sparse attention patterns** in the layers of the transformer, similar to the Sparse Transformer



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

Approach: Datasets

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Figure 3: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset.

- To reduce contamination, we searched for and attempted to remove any overlaps with the development and test sets of all benchmarks studied in this paper. Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model.

Approach: Training Process

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Figure 4: All models were trained for a total of 300 billion tokens.

- larger models can typically use a larger batch size, but require a smaller learning rate. We measure the gradient noise scale during training and use it to guide our choice of batch size.

cf)
Scaling laws for neural language models, 2020.,
An empirical model of large-batch training, 2018.

Results

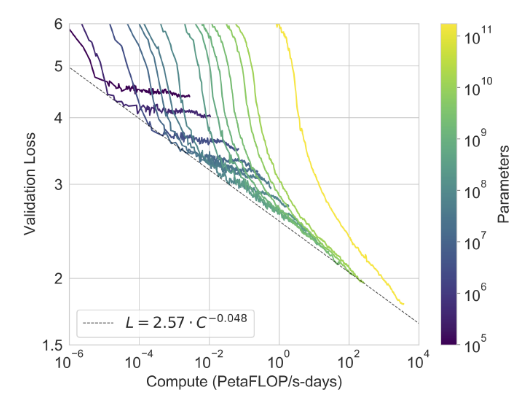


Figure 5: Smooth scaling of performance with compute.
Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training

Results: Tasks

1. Language modeling
2. Closed Book Question Answering
3. Translation
4. Winograd-Style Tasks
5. Common Sense Reasoning
6. Reading Comprehension
7. SuperGLUE
8. Natural language Inference
9. Synthetic and Qualitative Tasks

Measuring and Preventing Memorization Of Benchmarks

Since training dataset is sourced from the internet, it is possible that the model was trained on some of benchmark test sets.

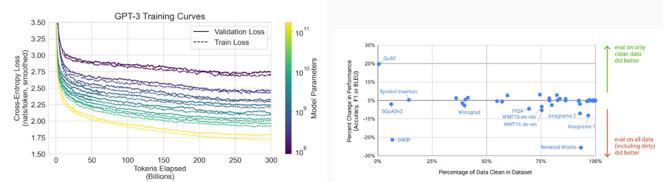


Figure 6: Benchmark contamination analysis We constructed cleaned versions of each of benchmarks to check for potential contamination in training set. The x-axis is a how much of the dataset is known with high confidence to be clean, and the y-axis shows the difference in performance when evaluating only on the verified clean subset.

Limitations: 1.Performance

It still has notable weaknesses in text synthesis and several NLP tasks.

- On text synthesis, GPT-3 samples still sometimes repeat themselves semantically at the document level
- Within the domain of discrete language tasks like "common sense physics": "If I put cheese into the fridge, will it melt?"
- NLI
- reading comprehension

Limitations: 2.structural and algorithmic

- since GPT-3 is auto-regressive language model, our experiments do not include any bidirectional architectures or other training objectives such as denoising.
- Low performance for following tasks:
  - fill-in-the-blank
  - comparing two pieces of content
  - considering a long passage and then generating a very short answer.

Limitations: 3.fundamental problem

- Current pre-training objective weights every token equally and lacks a nothin of what is most important to predict and what is less important.
- large pretrained language models are not grounded in other domains of experience, such as video or real-world physical interaction, and thus lack a large amount of context about the world
- Promising future directions in this vein might include learning the objective function from humans, fine-tuning with reinforcement learning, or adding additional modalities such as images to provide grounding and a better model of the world