CSE 256 Paper 2 Review

Review the following paper according the the prompts.
When you are done, submit the form. You will get a copy of your answers.

Make sure to print out the form in PDF format and submit on Gradescope.

Paper: Language Models are Few-Shot Learners
https://papers.nips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

* Required

Full name *

Le

UCSD student ID *

A59011332

Official UCSD email *

ley004@ucsd.edu

What are the key contributions (< 10 sentences) *

1. They demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, and sometimes it can be even better than prior state-of- the-art fine-tuning approaches.
2. They train GPT-3, an autoregressive language model with 175 billion parameters and test its performance in the few-shot setting without any gradient updates or fine-tuning.
3. They find that GPT-3 gets strong performance on tasks including translation, question-answering, and cloze tasks
4. They find that there are also some tasks where GPT-3 still struggles. Also GPT-3 faces methodological issues related to training on large web corpora on some datasets.
5. They find that GPT-3 can generate news articles that are indistinguishable to humans, and they discuss the social impact of that ability.

Discuss three(3) strong points of the paper (< 3 sentences each) *

1. They use the same model and architecture as GPT-2, but they use alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer.

2. For the training data, they filter CommonCrawl data based on similarity to a high-quality reference corpora; They perform fuzzy deduplication of CommonCrawl to prevent overfitting of the validation set; They add known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

3. In the training process, they use the scale of the gradient noise to select the appropriate batch size; They use a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network to avoid running out of memory

Discuss three weak point of the paper (< 3 sentences each) *

1. On text synthesis task, GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs.

2. The paper's experiments do not include any bidirectional architectures or other training objectives such as denoising. Also the paper's experiments do not include the results from fine-tuned GPT-3.

3. They paper's objective weights every token equally and lacks a notion of what is most important to predict and what is less important.

Rate the paper's originality and creativity *

	1	2	3	4	5	
Nothing new	○	○	○	◉	○	Very novel

Rate the paper's technical depth *

	1	2	3	4	5	
Not very technical	○	○	○	◉	○	Highly technical

Rate the paper's readability *

	1	2	3	4	5	
Well-written and easy to read	○	○	○	◉	○	Difficult to follow

What are the main questions you are left with after reading this paper? These can be technical questions or motivation related *

The news article generate from GPT-3 may contain inaccurate information or geographical and sexist content, how can we improve the model to overcome this issue?

Anything else you want to say about this paper

The paper is awesome and it find the limitation of former architecture is that they need finetuning on a dataset of training examples. Therefore the authors propose the GPT-3 model to remove this process and get competitive results.

A copy of your responses will be emailed to ley004@ucsd.edu.

Submit

Page 1 of 1

Clear form