Language Models are Few-Shot Learners

Critical evaluation of an AI application

[2] introduces GPT-3 (Generative Pre-trained Transformer), a successor to GPT-2 which is an auto-regressive language model for Natural Language Processing(NLP). GPT-3 is a scaled up version of GPT-2 with no architecture changes with respect to GPT-2. GPT-3 removes the need for fine-tuning a pre-trained NLP model to perform a specific task. Fine tuning is the training of a pre-trained NLP model on labelled dataset specific to the task like translation of sentiment analysis. This effectively updates the weights of the pre-trained model. Fine tuning is time consuming and labelled data for fine tuning is normally scarce. GPT-replaces the need for fine tuning through the use a few-shot mechanism in which a model is given a few demonstrations of the task at inference time as conditioning and does not require updating the weights of the model. GPT-3 was able to achieve state of the art performance on various NLP tasks and benchmarks such as: predicting last word of sentences, picking the best ending to a story or instructions, picking the correct ending sentence for five sentence long stories and closed book question answering. In most of the benchmarks GPT-3 set new records or were on par with fine tuned models. One of the advantages of using few-shot learning was seen on the models performance in translation tasks. The model did not use language specific labelled data outperformed state of the art supervised models in French to English and German to English translations. In addition GPT-3 was able to perform simple arithmetic tasks such as two digit addition and generate news articles. GPT-3 was also efficient at correcting English grammar.

Eight different sized models were trained ranging over three orders of magnitudes from 125 million parameters to 175 billion parameters and the last model named as GPT-3. Common Crawl dataset was used as the training data which consisted of a broad swath of internet data, two internet-based books corpora and English language Wikipedia. Fuzzy de-duplication at document level was performed to prevent redundancy and preserve the integrity of the held-out validation set. A large batch size and small learning rate was used for training and training was done on V100 GPU's of a high-bandwidth cluster provided by Microsoft. Performance was evaluated through traditional language modelling datasets and benchmarks.

A major drawback of the approach used in the paper was the bug reported by the authors related to train data filtering. The bug resulted in only partial removal of all detected overlaps. The bug was not removed due to cost of re-training. The bug raises concern about integrity of the validation dataset as it has been contaminated also known as data leakage. Data leakage causes predictive scores to overestimate the model's utility when run in a production environment [4]. Even though the authors have made a case that dataset contamination is not an issue for GPT-3 they use n-gram overlap detection. Though implementation was done to set N to 5th percentile example length in words, the value of N was constrained to 13 due to performance reasons. This casts doubt on the models ability to approximate the content of the document as well as the effectiveness of their overlap detection methodology [6]. In the GPT-2 paper it was suggested scalable fuzzy matching could be used for better overlap detection but not implemented in GPT-3 [5]

Lack of explain-ability on increased performance in common sense reasoning related tasks is another weakness of the evaluation approach in the paper to the point of suspecting whether the model memorizes the training data. The model was able to outperform SOTA in only one of the common sense reasoning task PIQA (PhysicalQA). PIQA task probes the grounded understanding of the world. While the other two common sense reasoning benchmarks ARC and OpenBookQA are more common sense reasoning intensive bench marks where the GPT-3 did not perform well. Authors were not able to explain this performance but instead tried to make a case by creating synthetic and qualitative tasks involving arithmetic and news generation. In the arithmetic task the GPT-3 performed two digit addition and subtraction but failed to perform arithmetic on numbers with more than two digits. To prove that the model is not memorizing the the specific arithmetic problems they ran searches for three digit arithmetic problems in both train and test set and demonstrated very low percentage of matches. It would have been convincing if they had ran similar search for the two digit arithmetic problems for which model was performing well. This fails to explain the common sense reasoning capabilities of the GPT-3 model for which it has been in the news for.

As mentioned earlier GPT-3 is an autoregressive model which means it predicts future values from past

values, so it has comparatively less context information as compared to a bi-directional model. This was a decision made by the authors for simplicity of compute and sampling. Authors have admitted that their model sees more text during pre-training than a human sees in their lifetime but still needs to improve to achieve same sample efficiency as humans. The authors have also failed to address the environmental impact of their choice of sample efficiency and associated increase in computation. A study conducted by University of Copenhagen estimates that training the GPT-3 would have roughly the same carbon footprint as driving a car the distance to the moon and back, if it had been trained in a data center fully powered by fossil fuels [3]

Finally the due to its unsupervised learning approach the model has retained the bias of the data it was trained on causing to to generate stereotyped and prejudiced content. In addition there has been no mention steps taken or planned future work in the direction of safety implications of the model generated output. Despite OpenAI's disclaimer of unsuitability of the model for healthcare, Doctors and machine learning practitioners at Nabla experimented with GPT-3 and found that it was not ready for the healthcare industry [1] . An example they cited is GPT-3 suggesting committing suicide as a good idea. Yudkowsky (2008) in how to design a Friendly AI mentions that friendliness should be designed from the start and designers should identify the flaws in their designs. Challenge is to define an AI system that evolves along with system checks and balances so that the AI system remains friendly in face of changes [6].

References

[1] MD Anne-Laure Rousseau, Clément Baudelaire, and Kevin Riera. *OPENAI'S LANGUAGE AI WOWED THE PUBLIC WITH ITS APPARENT MASTERY OF ENGLISH–BUT IS IT ALL AN ILLUSION*. URL: https://www.nabla.com/blog/gpt-3/.

[2] Tom Brown et al. "Language Models are Few-Shot Learners." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[3] Will Douglas Heaven. *OPENAI'S LANGUAGE AI WOWED THE PUBLIC WITH ITS APPARENT MASTERY OF ENGLISH–BUT IS IT ALL AN ILLUSION*. URL: https://www.thefreelibrary.com/OPENAI%27S+LANGUAGE+AI+WOWED+THE+PUBLIC+WITH+ITS+APPARENT+MASTERY+OF...-a0654752785.

[4] Shachar Kaufman, Saharon Rosset, and Claudia Perlich. "Leakage in Data Mining: Formulation, Detection, and Avoidance." In: vol. 6. Jan. 2011, pp. 556–563. DOI: 10.1145/2020408.2020496.

[5] Alec Radford et al. "Language Models are Unsupervised Multitask Learners." In: (2018). URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[6] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2002. ISBN: 0137903952. URL: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=ASIN/0137903952.