**Presentation details:**

**Topic: [8] Language Models are Few-Shot Learners.**

**https://crossminds.ai/video/gpt-3-language-models-are-few-shot-learners-paper-explained-6025d6a8fc2dc9d9ab36b587/?playlist\_id=6011f07becbeebc970a2ef20**

**Paper Reference:** [2005.14165.pdf (arxiv.org)](https://arxiv.org/pdf/2005.14165.pdf)

**1)What is the methodology of the paper .**

**1.1)Why this methodology is used compared to other methodologies?**

**1.2)Functionality and Working of the Methadology ?**

**1.3)What are the result produced by the Methadology ?**

**2)What are the current limitations of the methodology ?**

**3)How to improvise on the current approach ?**

**3.1)Ideas wrt to the algorithm .**

**3.2)Implementation wrt to the demo ?**

Notes:

Introduction:

1)This paper is written from authors of OpenAI.This paper discusses about language model GPT-3 which is a succession of Language models of Open AI platform.This language model has order of magnitude which is larger than any language model. **Voice assistants such as Siri and Alexa** are examples of how language models help machines in processing speech audio. If we train the language model with huge data it can solve the tasks that it has never seen before.For example to a language mode if we input a sentence say , ‘How r ’ it will be you .It solves the problem of which word is most likely to come next. Language model can generate a language or sentence in a probabilistic way . In this paper first they train the language model with some sort of text data . The datasets used in this paper are , ‘Common Crawl’ dataset for the quantity . It is a crawl of entire internet . It also consists of WebText2 ,Book1,Book2,Wikipedia datasets. These datasets have weight in training mix as 60%,22%,8%,8% and 3% respectively. They got these datasets from website and they trained the language model on that. They trained the GPT-3 model with various sizes. They then compared various language models like ‘BERT’ , ‘T5’ and found how many days they required to be trained in with above dataset.

If we consider GPT-3 it had about 175Billion parameters and the magnitude thus is very large. Basically in the paper , they trained various sizes of GPT-3 model to under the number of attenuation layers , each attenuation head and no of dimensions in each head and also the batch size. As we know that GPT-3 is a transformer model, which is a several layers of attenuation mechanisms . In order to transform the transformer model to attenuation model the different attenuation layers have information that is routed around the model it makes inferences and comes up with next word. Its not like BERT bi-directional its auto-regressive and always goes from left to right. Different GPT-3 sizes are based on more layers and wider layers.

\*How do they train the models in the paper?

1.In bert we first pre-train (teach with data), then fine tuning(sentimental classification is the task) with the database of labeled instances we train the model for the specific task using supervised methodology. Find tuning is performed on model with repeated gradient updates using a large corpus.

Eg: translating English to French:

Every word is gradient updated to form the final sentence.

Now with question-answering task , we take the pertained model and we fine tune it with Q-A dataset. Problem is with we always dont have a huge training dataset.

So, the main motive of the paper is to , reduce the step of fine tuning for every task and use a zero-short fashion.

\*What are the 3 settings we explore for in-context learning ?

1)Zero-shot: In this approach the model predicts the answer given only natural language description of the task. No gradient updates are performed. We basically input the task description and the prompt and ask the language model to predict the next word. The entire prediction is like in the training data the model as seen the similar task description and the prompt already.

Task des: Translate English to French

Prompt: Cheese

2)One –shot: In addition to the task description, the model sees a single example of the task. Here the example is pulled from the training dataset of the database . This example is not used for training the model , it doesn’t do any gradient updating it is just used to give better clarity to the model about what task it needs to do.

Task des: Translate English to French

Eg: sea otter=>loutre de mer

Prompt: Cheese

3)Few-short learners: It is similar to one-shot learners but the difference here is multiple examples are provided before predicting the word.

Task des: Translate English to French

Eg1: sea otter=>loutre de mer

Eg2: peppermint=>menthe poivree

Eg3: plush girafe=>girafe peluche

Prompt: Cheese=>

This paper talks about the fine tuning techniques , were instead of taking entire dataset from the database , it learns as a few short learners.

\*Experimental Results :

1.As the Compute increases ,validation loss decreases and goes down and down. As we scale up with lot of parameters the model gets better and better.

2.If we consider , the Q-A task and plot various Short learners techniques with accuracy , we can observe that few shot learners had better accuracy.

3.From the Results of Open-Domain QA tasks . Few short clearly out-performs open domain . But there is some cases which out-performs the few shot learners due to the NaturalQS(factual wikipedia).

Summary:

According to the paper , once the model is trained the data is stored in the weights of the function . Next when we give the model a set of task descriptions , ‘k’ examples and a prompt , if will go to weights of the function do some sort of regex and pulls out the data . There is no much reasoning happening here .

4.Translation-As the model goes up in parameters generally the performance increases. We can also observe that performance is very good when the language goes to English(target). Doesn’t matter much from the source language to English. This shows it is trained with lot of English dataset.

5.Some methods of Unsupervised out-performs the supervised for translation.

6.In ‘Winograd’ themes. A Winograd refers to a classifical task in NLP that involves determining which word a pronoun refers to , when the pronoun is grammatically ambiguous but semantically unambiguous to humans. This model out-performs fine tuned bert –large when compared to Robert a-large. But from the above we can see that based on the 0 or 1 or more examples the accuracy of the model veries.

7.In ‘PhysicalQA’,where it is a common sense reasoning(MCQS science Questions).GPT-3 achieves 81 accuracy ,zero shot has 80.5% , 82.8% for few shot.

Eg: If I drop a ball where will it fall?

Here there no much difference between 0 or 1 or more short learners and it is capable of out-performing Fine tuned SOTA. This may due to contamination in the datasets or there would be some significant difference.

8.In Reading Comprehension,the task was to use datasets including abstractive ,mcq and span based answer formats in both dialog and single question settings.Now GPT-3 performs best for COQA (conversational dataset). The measurement is done on a ‘SuperGlue’ which is a nlp benchmark. But GPT-3 doesn’t out-perform SOTA but it out-performs the fine tuned BERT Model.

9.NLI , which is the ability to understand the relationship between 2 sentences.The model here classifies if the second sentence logically follows the first, contradicts the first or is neutral. This is performed by Arithmetic addition.

Eg:

Task: “Arithmetic Addition”

Eg1: what is 3+2 =5

Prompt: What is 1+1=

From the graph we can see that ,the lower parameter models perform ok , bt as we go up the higher parameter models it performs really well for 2d addition and 2d subtraction. When we get to 4 digit multiplication the performance drastically dropped even for higher parameter models. This is because multiplication is computational harder. Another reason could be longer digit numbers are less frequent in the training data.

Explain various GPT-3 zero , one, and few shots .So , we can see that with few shot the 2 digit multiplication had almost 100% accuracy. Especially for GPT-3 One shot learner had a higher accuracy when compare to 0-shot learners.Almost from 76.9 to 99%.

10. Now the task is ‘SAT Analogies’, this is a test of MCQS taken for college entrance exam. On this task GPT-3 achieves 65.2% in the few-shot setting, 59.1% in the one-shot setting, and 53.7% in the zero-shot setting, whereas the average score among college applicants was 57% . The largest model achieves 65% accuracy in the few-shot setting. (explain the example.)

\*Current Limitations of GPT-3 tasks:

1) GPT-3 has Notable weaknesses in text synthesis and several NLP tasks. On text synthesis, although the overall quality is high, GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves.

2) GPT-3 has several structural and algorithmic limitations it doesn’t do very well with finding if 2 sentences are semantically correct and reading comprehension.(Was the news article written by human or computer).

3)In case of forming a short answer by going forward and backward to a sentence , large bidirectional model would be stronger at fine-tuning than GPT-3.

4)Large pre-trained 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. So even if GPT-3 is used with few shot learners it may not perform better than the humans.

5) A limitation, or at least uncertainty, associated with few-shot learning in GPT-3 is ambiguity about whether few-shot learning actually learns new tasks “from scratch” at inference time, or if it simply recognizes and identifies tasks that it has learned during training. Like does it just find patterns in the train datatset and predict. Synthetic tasks such as wordscrambling or defining nonsense words seem especially likely to be learned, whereas translation clearly must be learned during pretraining, although possibly from data that is very different in organization and style than the test data. But this is a ambuiguity and we cannot trust it entirely .

6) A limitation associated with models at the scale of GPT-3, regardless of objective function or algorithm, is that they are both expensive and inconvenient to perform inference on, which may present a challenge for practical applicability of models . Large models such as GPT-3 contain a very wide range of skills,most of which are not needed for a specific task.

7) Finally, GPT-3 shares some limitations common to most deep learning systems like decisions are not interpretable . It has higher variance in performance than humans and is highly biased. Bias is the model leads it to provide pre-judged content .

8) Any socially harmful activity that relies on generating text could be augmented by powerful language models. Examples include misinformation, spam, phishing, abuse of legal and governmental processes. It would be difficult to distinguish the human written text and one generated by these models and avoid the harmful activities.

9)As the models are highly biased it may produce text that are pre-judgmental and would thus hurt the sentiments of the humans.(Like news content generation against politics or individuals.)

10)High Energy useage. Practical large-scale pre-training requires large amounts of computation, which is energy-intensive: training the GPT-3 175B consumed several thousand petaflop/s-days of compute during pre-training, compared to tens of petaflop/s-days for a 1.5B parameter GPT-2 model.

\*Future Work:

1) For providing short answers for read comprehension , Bi-directional models perform better than GPT-3. I feel , when we make a bidirectional model at the scale of GPT-3(by increasing the parameters), and/or trying to make bidirectional models work with few- or zero-shot learning, is a promising direction for future research, and could help achieve the “best of both worlds”.

2) We know that for real world physical interaction the dataset is not sufficient and pre-trained models are not very robust. In order to address this we can perform fine tuning with re-enforcement learning and also adding additional modalities such that images to provide grounding and better model.

3) Another limitation broadly shared by language models is poor sample efficiency during pre-training. While GPT-3 takes a step towards test-time sample efficiency closer to that of humans (one-shot or zero-shot), it still sees much more text during pre-training than a human sees in the their lifetime . Improving pre-training sample efficiency is an important direction for future work, and might come from grounding in the physical world to provide additional information, or from algorithmic improvements

###################################################################

1)existing NLP method requires lot of fine tuning with datasets even after pre-training with huge corpus .(humans learn the language with fewer examples)

2)Therefore we scale up the language models to improve task agnostic .(agnostic refers to self learning techniques.)

3)Pros of GPT-3:

3.1)Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting.

3.2) For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model.

3.3)GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic.

3.4)GPT-3 **can create anything with a text structure**, and not just human language text. It can also automatically generate text summarizations and even programming code.

4)Cons of GPT-3 model:

4.1)At the same time, we also identify some datasets where GPT-3’s few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

4.2)Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. 4.3)GPT-3 **lacks the ability to reason beyond what is statistically sound text**. 4.4)Additionally, if the model comes across content that may not be adequately represented on the internet, it cannot generate any meaningful and coherent output text.