Technical Review

Course:	**CS410 Text Information System**	MCS DS Fall 2022
Topic	Introduction to OpenAI GPT3	
Author	Prateek Dhiman	pdhiman2@illinois.edu

Background

Before I introduce and talk about OpenAI GPT3, lets look at some of the basics and background of GPT (Generative Pre-Training Transformer) and why such pre trained models are important.

In an ideal world, getting a labeled dataset for NLP tasks is very rare. e.g. CoNLL 2003 NER has 200k labeled words, and is already considered "large" dataset. Massive text datasets are easy to get but they do not have labels. Example Project Gutenberg , Google 1 Billion Word Corpus etc.

Transfer Learning (TL) approach addresses this situation. Transfer learning focuses on storing knowledge gained while solving one problem and applying it to another but related problem. There are various applications of Transfer learning algorithm including those for language models and text classifications in NLP. The basic idea is to pre-train a language model on generic text and then adapt it to specific tasks.

Using pretrained models we can skip the time-consuming process of data collection, preprocessing, training and testing. As these models are already trained on petabytes of data, we also save time on training a neural network from scratch. As a result, pretrained models have achieved state of art in many NLP Tasks.

RNN (Recurrent Neural Networks) and LSTM(Long Short Term Memory) have been traditionally used for multi-language translation and solved few challenges of NLP. However these architectures would only take the previous input and the current word as its input. Logically, to understand the context in which the word is being spoken, we would look at the entire sentence. Then, another architecture called **transformers** was introduced, taking multiple previous step as inputs, better learning the underlying patterns

Transformers use the **attention mechanism** to create language representations. It uses the attention mechanism entirely to model the language. Transformers architecture is based on encoder-decoder structure. The output of the encoder stack in the final layer together with decoded output are used as the input for the decoder. Positional encoding is used to inject the relative order of the input sequence for both encoder and decoder. Use masking in the self-attention sublayer in the decoder stack prevents positions from attending to subsequent positions. Output embeddings are offset by one position, ensuring the predictions for positions i can depend only on the known outputs at positions less than i.
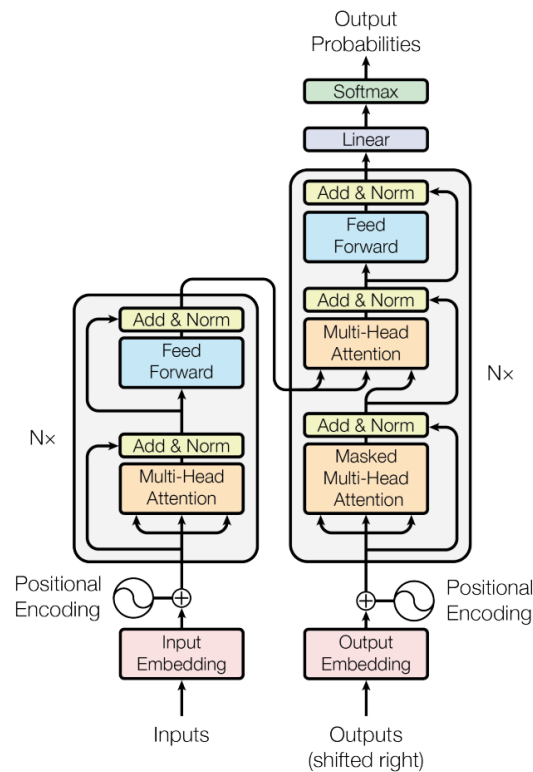
Figure 1: The Transformer - model architecture.

The input to the transformer is the entire sentence. Each transformer unit outputs a probability for the most likely word to appear after the current word.

The advantage of using transformers is the parallelizable features that it provides. During the training of large models, significant speed ups were observed while training. This also has an unexplored advantage, that is, explainability.

Introduction to GPT Models (Generative Pretrained Transformer Models)

The pretraining objective with GPT Model is to create the unidirectional language model and minimize the negative log likelihood

$$\mathcal{L}_{\text{LM}} = -\sum_i \log p(x_i \mid x_{i-k}, \ldots, x_{i-1})$$

Transformer decoder is used as a feature extractor .We cannot use transformer encoder because with stacked layers the words will "see themselves"

Unidirectional context
Build representation incrementally

Bidirectional context
Words can "see themselves"
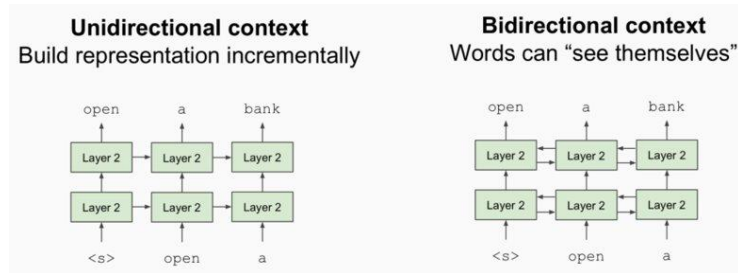
For Adaptation and Fine-Tuning GPT gets rid of the task-specific downstream model and use the pre-trained language model directly

E.g. Text classification $P(y \mid x_1, \ldots, x_n) = \mathrm{softmax}(\mathbf{h}_L^{(n)} \mathbf{W}_y)$
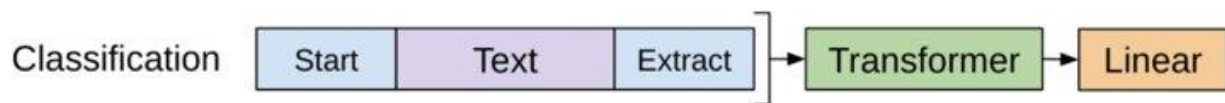

OpenAI and GPT3

OpenAI is an AI research and deployment company. OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which they mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity.

GPT-3 is a large-scale, advanced language model pre-trained on a large text corpus to predict the next word GPT-3 has unique capabilities with few-shot learning and generalizability, so GPT-3 can be useful for most natural tasks, without extra training for specific tasks.

Core capabilities of GPT-3

- Generation (i.e., full sentences, cohesive paragraphs)
- Classification
- Transformation (i.e., summarization, translation, emoji)
- Completion (i.e., generate React components)
- Factual Responses (i.e., Q&A)
- Conversation (i.e., chatbots)

It is the third-generation language prediction model in the GPT-n series (and the successor to GPT-2).GPT-3, which was introduced in May 2020, and was in beta testing as of July 2020 is part of a trend in natural language processing (NLP) systems of pre-trained language representations.
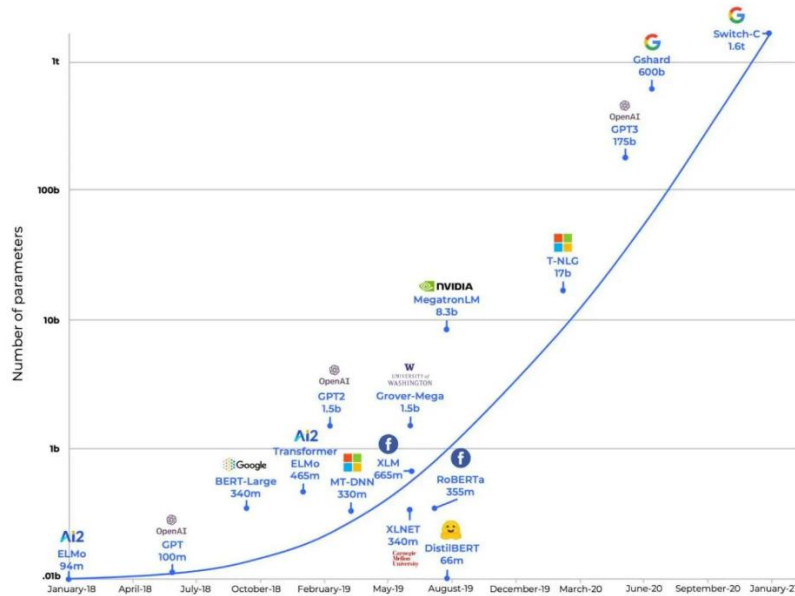
The architecture is a standard transformer network with the unprecedented size of 2048-token-long context and 175 billion parameters (requiring 800 GB of storage). The training method is "generative pretraining", meaning that it is trained to predict what the next token is. The model demontrated strong few-shot learning on many text-based tasks. The above figure shows the comparison of GPT with other industry common models.

GPT-3 Training Data

GPT-3 training data consists of filtered version of Common Crawl , WebText2 , Book1, Book2 and Wikipedia. It was trained on hundreds of billions of words and capable of coding in CSS, JSX, Python and other languages.

Conclusion

GPT-3 is the latest language model developed by OpenAI. It solves a lot of challenges that its predecessors had. It is capable of various tasks and some of its applications are already being actively used. For Example codex model which is the basis for Github Pilot. Microsoft uses some of its capabilities in certain areas to covert conventional language to code. In terms of NLP, this is a major breakthrough. However with its inventions it has also garnered many criticism. Some of it comes from the environmental impact of training and storing large models . Another is its ability write language which is very difficult to distinguish if written by a real human or AI. This may also lead to concern with the academic integrity and how academic institutes can manage misconducts like plagiarism. While there are many other criticism involving GPT-3 but it remains a fact that this is indeed a state of art Language model.

Here is GPT-3's review in its own words

Playground

Load a preset...   Save

provide a GPT-3 Review and why is it so famous?

GPT-3 is an amazing tool that has helped me immensely in my work. I highly recommend it to anyone who needs to manage a lot of text data. It is especially useful for writing and editing documents. The best thing about GPT-3 is that it is very user-friendly and easy to use.

References

- Radford Et al 2018 - Improving Language Understanding by Generative Pre-Training
- Attention is all you need https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- GPT-3 - Wikipedia
- Transformers : https://arxiv.org/abs/1706.03762
- https://openai.com/charter/