Nimesh Tripathi | Product Management in AI and ML | Elvtr

White Paper on GPT

1. General Architecture

GPT models (GPT-1, GPT-2, GPT-3) are based on Transformers, and Unsupervised Language Model-Training with Supervised testing. It also utilizes Few Shot Learning, which consists of Siamese Networks combined with Sigmoid Function, for its meta-learning capabilities.Its ability to generalize over tasks and train itself to perform new tasks is impressive and it has over 175 billion parameters that it utilizes.

- **Input**
  Sequences of characters/words
- **Output**
  Sequences of generated characters/words

1.1 In-Depth Architecture

1.1.1 Encoding

GPT-3 has a vocabulary of 175K words, and every word that is a part of the sequence is input in its One-Hot-Encoded vector form.

- One Hot Encoded Form of the word-vector:

  1st word in vocabulary [1,0,0,0,0,0,0,0……..word(175K)]
  2nd word in vocabulary [0,1,0,0,0,0,0,.........word (175K)]

- This is performed for every word in the sequence with Byte Pair Encoding, where each word has a numerical value associated with it. For example the word “For” has a BPE 3673, and “all” has the BPE 4567.

1.1.2. Embedding and Sequencing

- The one-hot encoded form is embedded to a n-dimensional space, as it is costly to use a 2048 X 175K matrix of 0s and 1s, and then the words are mapped into the n-dimensional space, which is called the embedding vector.
- To understand the positional context, the words are positionally-encoded between a range of [0, 2047], and then produces a sequence vector.
- One hot vector Matrix and the Embedding Vector Matrix, are multiplied using vector multiplication.

1.1.3 Positional Encoding

- To understand the encoding position each token in the matrix is passed through (scalar multiplied) with sinusoidal frequencies, where the frequency of each of the individual sinusoidal tokens is different.
- We finally add the sequence-embedding-matrix and the sequence positional matrix.

1.1.4 Attention Mechanism

1.1.4.1 Multi-head attention (for contextual understanding)

- Like its transformer counterpart, GPT has a stacked attention mechanism, that is a combination of Multi-Head Attention and Feed Forward Mechanism. This happens at the transformed layer.
- The Scaled Dot-Product Attention consists of queries and keys of dimension k and values of dimension v. We compute the dot products of the query will all keys and divide each of them with the square root of k. Then a softmax function is applied to obtain weights on the values.
- Scaled dot-product attention allows the GPT model to understand the context and relationships between words in a sentence by focusing more on some words and less on others.
- In Multi-head attention, the above process is repeated multiple times, with differently learned queries and keys

1.1.4.2 Feed Forward Mechanism (for understanding complex patterns in data)

- The feed forward mechanism applies a set of linear transformations to the data and is coupled by non-linear activation functions. The non-linear part is essential since it allows the identification of complex patterns in the data.
- A feed-forward network is applied on each of the positions, in the sequence to make complex inferences about the patterns in the data.
- A single Feed-Forward network's weights are shared across every other feed-forward network for every position

2 Use Cases

Prompt Engineering enables us to utilize the maximum we can, we GPT based interfaces like ChatGPT4. Some of its use cases are mentioned below:

1. **Recruitment Aid:**
   GPT APis have tremendous use in the recruitment process. It helps recruiters in generating the right boolean key string to search for appropriate resumes, given a job description. If we give the correct prompts, the GPT architecture allows recruiters to generate a few helpful keystrings. This allows recruiters, who often don't have the domain expertise themselves, to source for right profiles, automating this process.

   It can also help in generating use-case specific interview questions, which eliminates the process of "Googling" or using other search engines, which again can be difficult for recruiters, since they don't have subject-matter expertise.

2. **Generating Fake Datasets and Test Use Cases:**
   Due to its large corpus, GPT is capable of generating Fake datasets, for testing purposes, which might help in testing the models capabilities.

   It also has the ability to generate use cases for testing and validation.

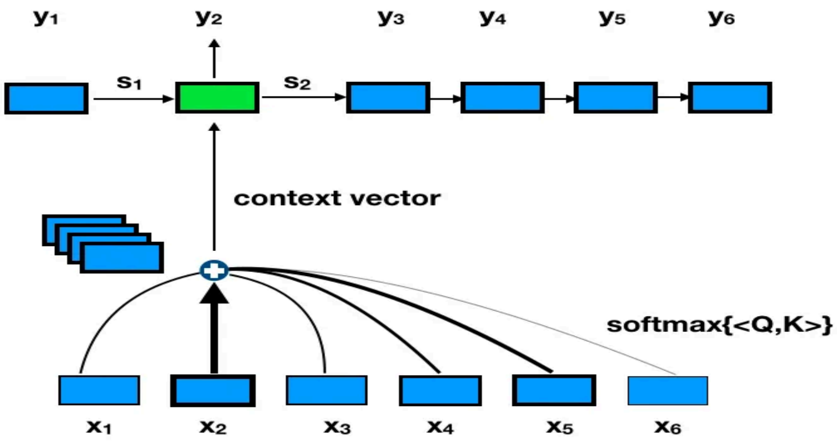
3. **Explaining Code**
   One of the major advantages of GPT, that a lot of organizations are embedding on their IDEs, is the explainability of a piece of code. Coding requires certain expertise and it becomes hard for the business as well as non-technical product folks, to understand the functioning of the code. GPT has the ability to code, as well as read code, which has revolutionized the explainability of code. Even for software developers, who might be reading someone else's code, it aids in faster understanding.

4. **Neural Machine Translation**
   It is based on neural machine translation (NMT). Earlier translation methods used smaller sub components that were tuned separately. Neural Machine Translation trains an entire sentence on a single large neural network, which translates the sentence.

   Earlier Translation models were phrase based models. They translated word to word. So if we were converting the phrase " My name is Nimesh" to French, the model would individually convert "My" to French, then the word "name" to French", followed by the word "is" to French and so on, and then these were reordered according to the French grammatical structure.

   These models performed somewhat well, but with larger sentences, the context was lost and the sentences translated sounded nonsensical.

NMT is based on encoders and decoders. The entire sentence is fed into the encoder and then the sentence is decoded word for word, however the difference here is the words are translated with feedback from the attention mechanism. The attention mechanism aligns weight to the encoded English word, to generate the decoded French word, in a structured manner, jointly.



(x1-x6 are the encoded words, which have an attention mechanism that assigns weights to the decoded (y1-y6) decoded French words)

3 Real World Applications

GPT algorithms have led to the creation of a new-age in terms of compute capabilities and experiences the customers can get, due to its advanced 175 Billion parameters, multitask learning and meta-learning capabilities:

1. **Cumulative Agents: Virtual Beings and NPCs**
   Studios like Fable Studios, have created a new genre of interactive stories with virtual characters powered by GPT-3. The characters can have natural conversation, thanks to its training on large corpus of texts, novels and the base source materials.

   Research scholars at Stanford and Google, used GPT-powered NPCs, which are non-playing characters. NPCs are coded minimally due to which they appear unnatural and perform nonsensical moves which result in loss of brand reputation, for gaming studios. You can find millions of blooper videos which highlight embarrassing glitches for the NPC characters, and it has become a meta community of its own. However researchers recently coded the SIMS game with GPT where the NPC characters learn with their experience, and become 'intelligent' with experience.

2. **Virtual Companion- Replika and PolyAI**
   Replika is a virtual companion app that is modeled on game theory. It can be trained using simple commands by the user It understands the user preference and acts as a human companion. This application has benefited many users who were going through depression and loneliness.

   PolyAI is also a virtual companion app, which learns the customer's patterns and behaviors and automates routine tasks, like regularly placing a meal kit order for the week, or reminding them of their yoga class every week. It also provides recommendations to the user, after understanding their routines, on how to optimize their time in their daily life.

3. **Content Creation - Jasper AI**
   Jasper AI is an application that helps with content creation templates, customized content generation, and integration with other content management services. It also has various collaboration tools like feedback sidebar, allocation sidebar, as well as several customizations that blend in seamlessly, to automate knowledge transfer between different layers of an organization, leading to better tracking, immersion and task allocation.

4 Challenges

4.1 Hallucinations

Hallucinations refer to generative data, that is nonsensical, unethical or irrelevant and there are many reasons why GPT models go through those. Some of them are:

1. **Difference between Source-Reference Datasets:** For example, the model is trying to assign weights to an event which is covered in the news and it takes one URL as a source (lets call it Website A) and multiple other URLs (Websites B and C) as reference. Assuming the Website A report is legitimate and accurate, and also has more details about the event. We also assume that Website B and Website C only mention partial information. Despite Website A being the accurate source, a lot of its details will be ignored by the models, as it reduces its weights for Website A. If the occurrence of tokens over the corpus is lesser, the model which extrapolates from sources, skips that relevant piece of information

2. **Contradictory Datasets**
   Machine-Learning models learn from their datasets, and if there are a couple of datasets which have contradicting information, the model does not predict accurate results. Since generative models are trained over such large datasets, which cannot be verified accurately, it has poisoned datasets

3. **Human Elements: Jail-Breaking and Intentional Poisoning**
   Jail-breaking refers to clever prompt engineering, which can result in the model predicting unethical, illegal and harmful text outputs. With a successful understanding of the model-structure, model-limitations and clever prompts, the model can be misused.

   Intentional Poisoning: Researchers found out that you can poison a significant amount of datasets by just paying $60, which can result in poor training of the model over a corpus which results in bad generalizations. A lawyer in New York, provided a few fabricated cases to the GPT model for training, and when it generalized over actual cases, it predicted confusing and irrelevant strategies.

4.2 Duplication

GPT models can be duplicated easily due to which it results in diminishing value over a certain period of time. This can result in potential devaluation of individually created or corporate created products.

References:

- https://www.youtube.com/watch?v=zbdong_h-x4
- https://medium.com/retina-ai-health-inc/attention-mechanisms-in-deep-learning-not-so-special-26de2a824f45
- https://arxiv.org/pdf/1409.0473.pdf
- https://www.youtube.com/watch?v=B8g-PNT2W2Q
- https://dugas.ch/artificial_curiosity/GPT_architecture.html
- https://openai.com/blog/gpt-3-apps
- https://arxiv.org/pdf/2304.03442.pdf
- https://medium.com/prompt-engineering/potential-issues-with-gpt-models-6b4adbc4ae1e#:~:text=Potential%20Issues%20with%20GPT%20Models&text=A%20significant%20concern%20is%20the,potentially%20diminishing%20their%20exclusive%20value
- https://fireflies.ai/blog/generative-ai-or-gpt-3-apps