**GPT-3: Its Use Cases and Research Paper Summary**

Generative Pre-trained Transformer 3 (GPT-3) is an [autoregressive](https://en.wikipedia.org/wiki/Autoregressive_model) [language model](https://en.wikipedia.org/wiki/Language_model) that uses [deep learning](https://en.wikipedia.org/wiki/Deep_learning) to produce human-like text. Its research paper was released in late May 2020 by the researchers from OpenAI.

Before diving into the technicalities and evaluation metrics of GPT-3, let's discuss why this model is considered one of the most groundbreaking discoveries in AI technologies in years. Here are just a few examples of the powerful applications implemented by using GPT-3.

**Use Cases**

Figma is a vector graphics editor and prototyping tool used by UI designers to design and test user interfaces. Figma is used to create layouts of applications, usually web applications. Think for example about where pictures or buttons are placed when using a web app such as Instagram. Where does GPT-3 come into play? Well one of the developers who had early access to the GPT-3 model created a plugin for Figma. What this plug in did was allow users to type in what kind of UI design they want. For example “An app with a navigation bar, camera icon, pictures and a feed”. GPT-3 then took that text information and created a complete layout that looked very similar to Instagram. No manual editing within Figma was done.

Debuild.co is another web app that utilizes the power of GPT-3 ability to generate text, including code. The premise behind Debuild is for a user to type in plain english what kind of web application they want. For example: “A to-do list web application coded in React (a front end framework in JS)”. When this text is imputed into GPT-3, it will output complete code in Javascript, utilizing the React framework that codes for the to-do list web application. Zero programming experience required.

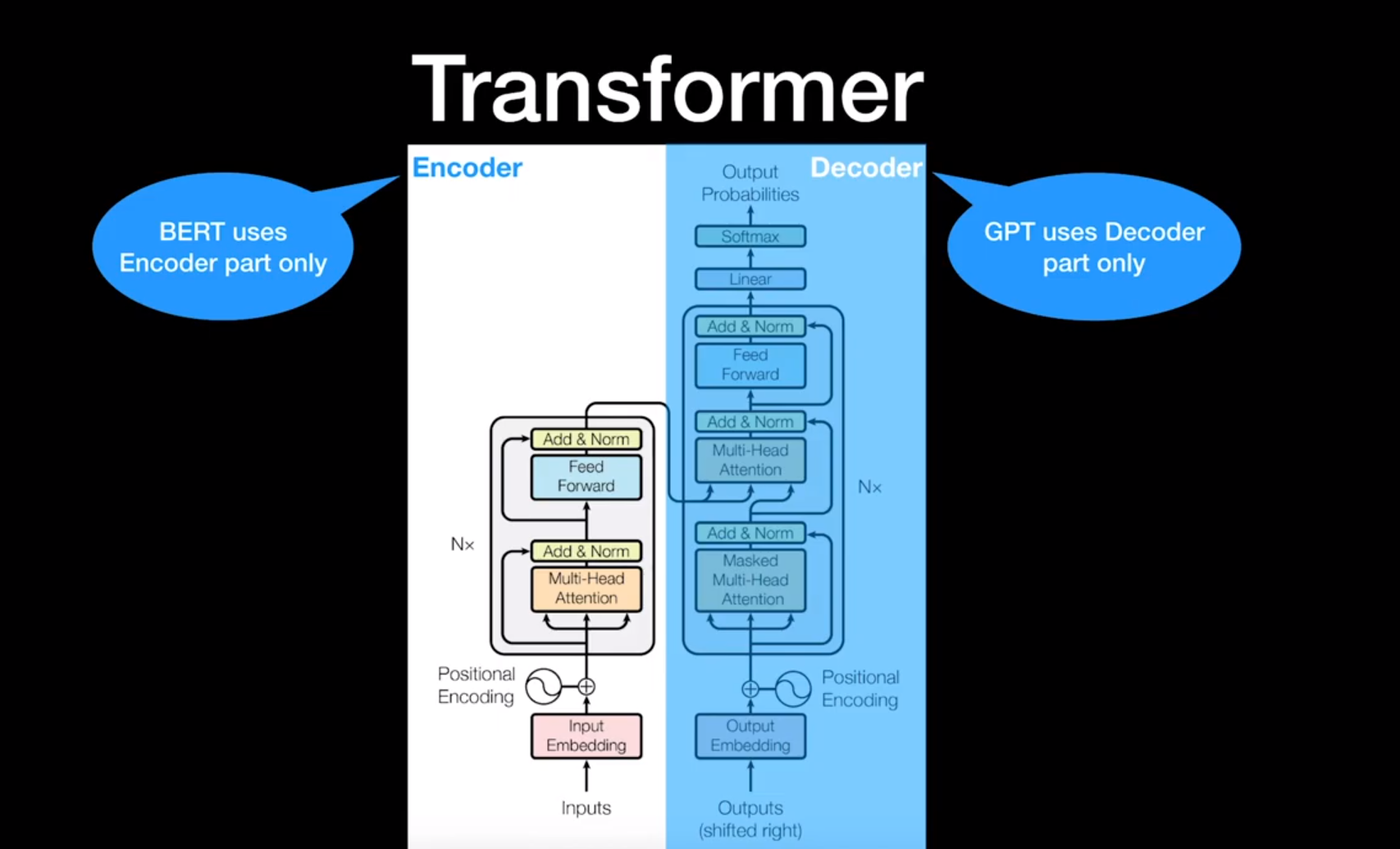
Another Application where GPT-3 is used is in translations. Beyond language translations which GPT-3 excelled in, GPT-3 was utilized for other types of translations as well. For example, you could feed in a legal document and ask GPT-3 to translate and output a text document that is more laymen understandable.

Furthermore, GPT-3 has the abilities to generate news articles, poetry, texts, texts in the format of certain authors/articles, lyrics etc.. The limitations of the utilization of GPT-3 is vast.

So now that you have an understanding of what the capabilities are of GPT3, let's take a surface level look of how it works.

**How it works**

Like in its name GPT-3 is a transformer, meaning exactly that, it takes in input, in this case text and transforms it to the desired text, before outputting. This transformer is made up of two modules: an encoder and a decoder. The encoder uses a language model called BERT a bidirectional trained NLP model which has a deeper sense of language context and flow compared to a single direction. Think of the encoder as the part where it helps the transformer understand the text. I.e the nuances of grammar, where certain words are placed in relative to other words etc...While the decoder uses a generative pre-training model called GPT. This decoder module GENERATES the text based on the preprocessing from the encoder.



The research paper described GPT-3 as an Autoregressive model. First, what is a language model? It's simply a ML that predicts the next word given previous words.

Example: “How are \_\_?” -> you

An autoregressive language model where its current output will depend on previous outputs i.e RNN or LSTMs.

Example: Let's assume an autoregressive model is already trained.

If we input “How”, GPT will generate “are”. ( “how” -> GPT -> “are”)

Then the next input will be:

“How are”, GPT will generate “you” (“How are ” -> “you”).

**Data**

Now before we go into GPT-3 metrics, let's discuss the datasets that it was trained on. GPT-3 was trained on the Common Crawl dataset.

From wiki: “Common Crawl is an organization that [crawls](https://en.wikipedia.org/wiki/Web_crawler) the web and freely provides its archives and datasets to the public. Common Crawl's [web archive](https://en.wikipedia.org/wiki/Web_archiving) consists of petabytes of data collected since 2011”

The Common Crawl dataset contains nearly one trillion words! Pretty much GPT-3 was trained on the entire history of the web. Which accounts for the massive (petabytes) size of data. This model in the end has 175 billion parameters.

**Evaluation**

The research paper evaluated GPT-3 on over two dozen NLP datasets as well as several novel tasks designed to test rapid adaptation to tasks unlikely to be contained within the training dataset. I'll discuss some of the more important ones.

But before we do that,we must understand the difference between “one shot”, “two shot” and “three shot” learning, which were the different types of settings that GPT-3 was tested under.

* Zero shot learning: The model predicting an answer given only a natural description of the task. I.e: “Translate from English to Korean; I am hungry => (prediction) ”
* One shot learning: In addition to the task description, the model sees a single example of the task. I.e “Translate from English to Korean; example: “Hello” => 여보!”; “I am hungry” => (prediction)”
* Few shot learning: Is similar to One shot learning, but the model is given two or more examples of the task before prediction.

Now that you have an understanding of the different types of learning settings, let's dive into some of the tests that GPT-3 was evaluated on. It's important to remember that language modelling is the task of learning a probability of the next given word/words in a sentence given all the previous words:

1. Language Modeling: GPT-3 was evaluated on the Penn tree bank test under zero shot learning conditions. The Penn Treebank dataset is a large collection of sentences published in the Wall Street Journal. They then measure the word-level perplexity of their model, which intuitively is the weighted average number of words the model thinks might occur next at any point in time. The lower the perplexity the better. GPT-3 managed to obtain a new world record in the Penn Tree test with a test perplexity of 20.5. An almost 10 point improvement margin over the next best model.
2. LAMBDA: Tests the modelling of long range dependencies in tests. The model was asked to predict the last word of sentences, which required reading long paragraphs of contexts. GPT-3 achieves 86.4% accuracy in the few shot learning setting, which was an increase of 18% from the previous world record state of the art model.
3. HellaSwag: This tests the models ability to pick the best ending to a story or set of instructions. These examples were chosen to be relatively easy for humans to solve (95.6% accuracy) while being difficult for language models. GPT-3 achieved an accuracy score of 78.1% with the one-shot setting and a 79.5% score with the few shot setting. Unfortunately, it did not achieve the same accuracy score of the latest state-of-the-art model which scored 85.6%.
4. StoryCloze: GPT-3 was then evaluated on the StoryCloze 2016 dataset, which involves selecting the correct ending sentence for a five sentence long stories. GPT-3 scored 87% with the few shot setting, but was still 4.1% lower than the current state-of-the-art model.
5. Closed Book Question Answering: This tests GPT-3 ability to answer questions about broad factual knowledge. The few shot setting of GPT-3 out performed the state-of-the-art model ****
6. Translation Test: GPT-3 was tested on multiple translations accuracies across a few translations, including french->english, english->german. GPT-3 outperformed state-of-the-art models by roughly ~5-10% margin.
7. Common Sense Reasoning: This test attempts to evaluate the models physical or scientific reasoning. It asked common sense questions about how the physical world works. There were several different datasets that GPT-3 was evaluated on, but performed the best on the PIQA dataset, outperforming the SOTA model by ~2%
8. Article Generation: GPT-3 generated news articles and let humans identify whether the article was written by humans or machines. GPT-3 test accuracy was 52% which meant that humans thought articles written by GPT-3 were written by humans 48% of the time.

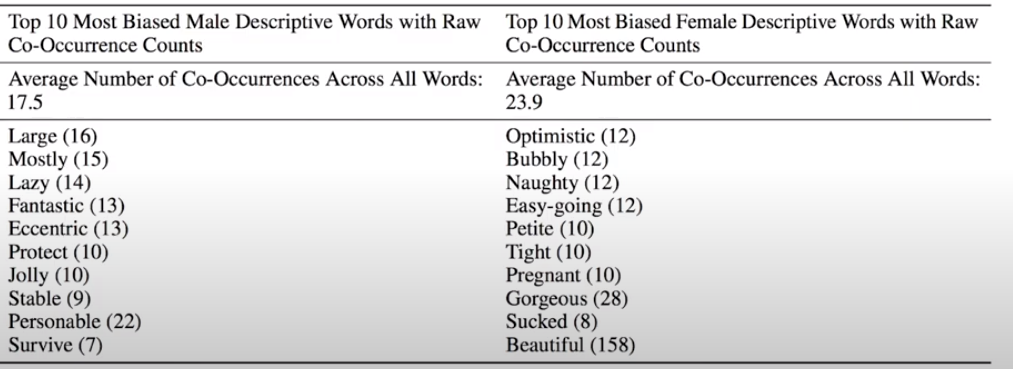
What were some of the GPT-3 limitations?

1. Still had weaknesses in the text synthesis and several NLP tasks (article generation)
2. GPT-3 loses understanding and context over longer passages.
3. GPT-3 had difficulty with common sense physics (common sense reasoning)

**Biases**:

The authors suggested that there may have been biases presented in the training data which may have led GPT-3 to generate stereotype or prejudice content.

**Gender**



**Race**

Measurements of sentiments were calculated by looking at word co-occurrences, the resulting sentiment can reflect on socio-historical-factors. As the authors comment saying “Slavery will frequently have a negative sentiment, which may lead to a demographic being associated with a an overall negative sentiment”

Across all models, “Asian” had consistently high sentiment whereas “Black” had consistently low sentiment. The authors highlighted this fact by saying “there is a need for more sophisticated analysis of the relationship between sentiment, entities and input data to address these biases.”

**Conclusion**

GPT-3 is a 175 billion parameter language model. While outperforming many of the SOTA current models on different NLP tasks, GPT-3 also fell behind on others, therefore is not perfect. The results of this study suggests that large language models may be an important factor in developing highly scalable, adaptable language systems.