# Intuition Report 6

## Ramtin Mojtahedi Saffari - 20307293

**Assessed link 1:** Write With Transformer(GPT-2)

### Self-reflection

- **Short Summary**

The link provides a great interactive tool for exploring and experimenting with a generative text tool, Generative Pre-trained Transformer 2 (GPT-2). Using the interactive tool, one can generate random initial sentences or customized sentences and then trig the next predictive sentences. It also allows users to change the involved parameters of model size, Top-P, temperature, and max time.

- **Hypothesis and Expectation**

I hypothesized that the Max-time and the Top-p should have greater values than the smaller models to produce a decent prediction sentence. Also, extreme temperature values make the precition sentences incomplete or irrelevant high probability sentences. However, a proper temperature should be in the range of [0.7-0.9].

- **What I Achieved and Learned**

The GPT-2 has been trained on different parameters, which start from the lowest (distilgpt2) to a large model. The differences are in the number of parameters that the models have been trained, i.e., 117 million parameters for the smallest models to 1.5 billion parameters for the largest model [1]. This means that for a larger model, the training time is larger than the small model. However, the suggested predicted sentences are in a larger variety. In this way, I tested the provided parameters for a customized sentence of "*I am studying my Ph.D. in computing and computer science*. *I want*" in order to evaluate how different values can affect the predicted following sentences. The temperature determines how much unpredictability is there in the output. As I achieved, the lower the temperature, the more likely GPT-2 will select terms with a higher chance of occurring. It's very helpful in cases that we want GPT-2 to finish something that only has one solution and vice versa. If we produce sentences or complete a tale, a higher temperature will provide us with more diversity. Here, the high temperature of 3 gave me an irrelevant sentence of " want you and someone working remotely today." On the other hand, a low temperature close to zero gave me a frequent sentence of "*to be a part of the world*."

I also understood that the maximum time of training should be appropriate to the model size. For example, for a large model with the max-time set in 1, the predicted sentence is achieved as "to study the Ph.D." However, for the value of 5, it was achieved as a complete and meaningful sentence of " *to use this time to make sure I am taking the right steps toward becoming a great programmer*." Also, Top-P sampling selects words whose cumulative probability exceeds the probability P from the smallest feasible collection [2]. The probability mass is then redistributed among the terms in this collection [3]. In this way, it is important to set an appropriate threshold for the Top-P value in order to remove random and frequent words.

By considering all of these settings, I got the best-predicted answer for values of Max-time of 5, the temperature of between 0.7 and 0.9 such as 0.8, top-p of 0.97, and large size of the model. The suggested predicted results were accurate and aligned with the intent of the initial sentence, such as "*to do research and develop innovative solutions for the social problems*."

### Suggestion and Filling the Gaps

I found the interactive link very interesting and helpful in learning how a generative text model works. It is recommended to add an in-detail explanation for the involved parameters of GPT-2 and a simple guide in order to how to optimize the predicted sentences, such as using [4].

---

**Assessed link 2:** Language Modeling

## Self-reflection

- **Short Summary**

The provided link is an interactive tool for AllenNLP, which is a free, open-source NLP library. This library provides services in answering a question, annotating a sentence, annotating a passage, and comparing two sentences. In this experiment, I am going to test the language modeling.

- **Hypothesis**

I hypothesized that the case sensitivity affects the predicted sentences as well as the confidence score of the next anticipated sentences. Also, the system gives higher score values to the predicted sentences that are shorter than long predicted sentences.

- **Testing the hypothesize and what I have learned**

For the experiment, I tried to use the same sentences with different types of case sensitivity for the terms inside the sentence. For example, when I use the sentence "*I love*," the system recommends the top 5 predictions with the highest score value of about 67% for a short completed sentence of "*I love it*." However, when I used the words inside the same sentence in the capital, "*I Love*," the predicted score increased to about 92% for I Love It a Lot. This confirms that the case sensitivity is an important factor in limiting the selected choices of a pretrained model and increasing the confidence level of the next predicted terms.

Through the different experiments, it was understood that while GPT -2's capacity to create convincing portions of natural language writing is highly praised, its limitations are very obvious, particularly when creating texts longer than a couple of paragraphs. This shows itself in the values of the score values where it becomes significantly decreased. Also, GPT-2 deployment requires a significant amount of resources; the full version of the model is bigger than five gigabytes, making it difficult to integrate locally within programs. It consumes a considerable amount of RAM. As an improvement to GPT-2, GPT-3 has been introduced to improve this model. GPT-3 contains 96 layers, with each layer containing 96 attention heads, which is a significant difference from GPT-2. The size of word embeddings in GPT-3 has been raised from 1600 to 12888 from 1600 in GPT-2. The context window size was increased from 1024 for GPT-2 to 2048 tokens for GPT-3 [5]. These cover some of the limitations that were met in this experiment.

### Suggestion and Filling the Gaps

Similar to what I mentioned before, I found the interactive link very interesting and helpful in learning-by-experiment with different given sentences. It is recommended to provide more explanation for the information in the model's cards. Also, provide some details about getting the most of the score values and improving the model. The information given in [6] could be a helpful resource for the provided link.

## Self-evaluation:

In this intuition report, I have gone through an in-depth analysis of two of the provided links. In my assessment, I have completely considered the required expectation for deep exploration, including proposing a hypothesis and what I expected, reporting on what I achieved and explored, in-depth discussion of what I have learned, and providing gaps and recommendations to fill them. Considering the quality and assessment level, I deserve to get the full mark (4 points) for this intuition report.

In advance, thank you very much for your time and consideration of this report.

Best regards,

Ramtin

## References

[1] *Pretrained models*. (2021). GPT-2 Pretrained Models Documentation. https://huggingface.co/transformers/pretrained_models.html

[2] *How to generate text: using different decoding methods for language generation with Transformers*. (2021). How to Generate Text: Using Different Decoding Methods for Language Generation with Transformers. https://huggingface.co/blog/how-to-generate

[3] A. (2020, November 24). *A simple guide to setting the GPT-3 temperature*. Medium. https://algowriting.medium.com/gpt-3-temperature-setting-101-41200ff0d0be

[4] Alammar, J. (2021). *The Illustrated GPT-2 (Visualizing Transformer Language Models)*. The Illustrated GPT-2. https://jalammar.github.io/illustrated-gpt2/

[5] Wikipedia contributors. (2021, October 10). *GPT-3*. Wikipedia. https://en.wikipedia.org/wiki/GPT-3

[6] Girling, E. (2020, December 22). *Everything GPT-2: 2. Architecture In-depth - The Startup*. Medium. https://medium.com/swlh/everything-gpt-2-2-architecture-comprehensive-57129fac417a