Models AI GPT3 and GPT4 Compared

Introduction:

Artificial intelligence (AI) is rapidly becoming an important technological innovation in numerous fields. Among the various applications of this technology, natural language generation is playing a crucial role. In this context, the Generative Pretrained Transformer (GPT) text generation system represents a significant step forward. This system was originally developed by the OpenAI company, and later reproduced by other companies with the same objectives.

The GPT system, based on neural networks, learns the content of language from large volumes of text using various machine learning methods. First introduced in 2018 as GPT-1, later in 2019 as GPT-2, and in 2020 as GPT-3. The latest GPT-3 took considerable attention also from media, for its impressive capacity and versatility in text production. In order to overcome the limits of GPT3, OpenAI has announced the development of an even more advanced system: the GPT4. In this report, we will focus on the technical and practical differences between these two systems.

What is GPT:

The GPT, literally Generative Pretrained Transformer, is an artificial neural network model developed for natural language generation, expanding the family of deep neural network models. Based on machine learning, this system is capable of predicting and generating sequential text based on input.

The basic architecture of a GPT involves an encoder, a decoder, and a separator, allowing the AI to predict and generate natural language not only based on input data, but also on acquired knowledge from training. The system was originally designed to enhance AI performance in many natural language applications, such as automatic translation, speech synthesis, and automatic information retrieval.

GPT-1:

The GPT-1 model, created as a prototype of development, had 117 million parameters. Although it was able to produce coherent texts, its quality was not yet up to expectations.

GPT-2:

GPT-2 was the second iteration of this AI model, launched in 2019. This system demonstrated remarkable capabilities: even with only 1.5 billion parameters, it was able to write coherent texts, generate syntactically and semantically correct text sequences, and distinguish them from human productions. This garnered significant attention even outside the field of computer science. The AI showed remarkable capacity to generate many types of text, including very long and complex ones, but at the same time could (in the context of pure statistical word association) produce false statements and be rooted in gender and racism biases.

GPT-3:

In 2020, the potentials of GPT-3 were presented, proving to be a further big step forward, equipped with approximately 175 billion parameters. This allowed the AI to demonstrate a great capacity for accurate and general text comprehension and production. GPT-3 was able to generate text of very similar quality

to that written by a human being. There was therefore great enthusiasm for the potential use of this AI in numerous sectors, such as automatic translation, automated commercial text production, story creation, and digital anthologies.

GPT-4:

In 2021, the company OpenAI announced the new development in the field of generative pre-trained transformers: the GPT-4 system. With a trillion parameters, the system will be able to process large amounts of data at an unprecedented speed. GPT-4 will present many innovations compared to previous generations, such as advanced contextual comprehension and multitasking processing capabilities. In addition, it will be able to generate high-quality and customized text for a wide range of contexts, including scientific reports and academic documents.

General Context:

AI is a rapidly evolving field that continues to present significant advancements. The use of these devices covers a wide range of sectors, including health, education, banking, marketing, and the automotive industry. Many of these applications are already in use and are constantly being improved. Among the latest innovations in this field, GPT3 has proven to be a major evolution in the field of natural language generation. The GPT3 system, as mentioned before, is able to process human language with a detailed semantic understanding of sentences. Thanks to its extensive vocabulary of words and its expressiveness, it can be used to answer questions, translate languages, write words and sentences in academic and commercial contexts.

However, despite the excellent results and enthusiasm generated by the GPT3 system, based on global feedback, we have not yet reached levels of "human-like" text generation, hence the need to evolve it into GPT4. But what are the technical and practical differences between these two systems? In the following, we will focus on the comparison of these two important AI developments.

Technical differences:

Note: in this section, we will focus on the technical details of GPT3 and GPT4. For non-expert computer audiences, these terms may be complex, but they are very important in understanding the differences between these two projects.

The GPT3 system uses a parallel processing platform on GPU, which allows up to 175 billion parameters to be processed. The number of parameters is the baseline factor for measuring the processing power of AI systems based on neural networks. In this field, GPT3 represents an improvement over the previous iteration, namely the GPT2 system, which had only 1.5 billion parameters.

The fourth version of the system, called GPT4, will have even more parameters than GPT3: in fact, the GPT4 system will use a processing platform on a supercomputer capable of processing up to a trillion parameters. These advances represent a considerable step forward compared to previous generations, and this development is expected to have important commercial and academic applications.

But the innovations do not stop there: in parallel, GPT4 will also present innovations in computer architecture, with the aim of improving performance and response to commands. In particular, we would move from a single linear architecture, in which "tokens" (the smallest textual units with specific

meaning) are entered in a single linear sequence, to a multiple architecture, improving processing capacity.
Another area of progress is contextual comprehension. In the new system, syntactic and semantic comprehension would be integrated with all previous and subsequent content, providing a more complete and profound understanding of the context in which the sentence is used. This would allow the system to produce more complete and coherent responses.
Finally, GPT4 would also be able to process multitasking, meaning it can perform multiple tasks simultaneously, such as answering different questions, simultaneous translations of multiple languages, and generating texts in multiple languages and styles.

Practical differences:
While the technical differences between GPT3 and GPT4 are relevant to programming and data processing, the practical differences between these two systems can influence user experience and the possibilities offered by the AI system.
One of the main advantages of GPT3 is the highly intuitive nature of the interface, which allows the user to make requests and receive quick and accurate responses. Thanks to the system's large data processing capacity, it is possible to obtain a complete and detailed response from GPT3 even to a vaguely formulated question.
However, the ability to generate complete sentences on its own does not indicate the degree of accuracy or usefulness of a system: it may be limited in that it cannot fully understand the conversation context, ask follow-up questions, or recognize irony or sarcasm. In other words, it is important to look beyond the ability to generate complete sentences to evaluate the effectiveness of the answers. For example, it might be difficult for the system to produce an adequate response to a question that implies in-depth specific knowledge, not just general knowledge, about a particular academic topic. In this regard, GPT3 may risk producing answers that, although technically and structurally correct, do not effectively answer the question.
At the same time, the effectiveness of a response depends closely on the context in which it is applied. Its use in commercial contexts could be very useful for creating new marketing content, but its usefulness might be limited for more complex activities that go beyond pure statistical knowledge. In fact, from the point of view of user experience, the use of GPT3 for personal purposes could have some limitations due to the presence of predefined responses and excessive focus on keyword searches. This phenomenon causes an inevitable creative block in understanding and interpreting results, and would limit the exploration of the possibilities offered by the AI system.
However, with GPT4, these restrictions should be overcome thanks to its ability to generate more effective responses to questions and better understanding of linguistic patterns. The use of this system could lead to significant improvements in research and data processing.

Conclusions:
In summary, we can say that the GPT3 system is a significant step forward in terms of natural language generation capacity and is used in various fields. However, despite its great power, the system has some limitations that are reflected in user experience and the ability to produce personalized and effective responses.

With GPT4, OpenAI promises significant innovations in terms of data processing, which should make this new system even more adaptable and versatile. In particular, the greater multitasking capacity and more complete comprehension of the linguistic context could represent a fundamental breakthrough for AI, opening up new avenues for control and customization of this device. However, from a practical point of view, it is still difficult to evaluate the effectiveness of the new features, and we will have to wait for the official presentation of GPT4 to have a complete picture of its actual potential and limitations.

Generatively yours… Mike Yoshi