Nepal Engineers' Association, Gandaki Province, Technical Journal

Revolutionizing Natural Language Processing with GPT-based Chatbots: A Review

Smita Adhikari [1], Bhawana Dhakal [2]

[1] *Department of Electronics and Computer, Institute of Engineering Paschimanchal Campus*

[2] *Officer, Gandaki University*

*(*Corresponding email: smita.adhikari@pasc.tu.edu.np)*

Received: January 17, 2023; Revised Day: April 03, 2023; Accepted: July 07, 2023

Abstract

OpenAI introduced a language model called GPT (Generative Pre-trained Transformer model. The algorithm learns to predict the following word in a phrase based on the context of the preceding words after being trained on a large text dataset. GPT employs a transformer architecture, a class of neural networks that have been demonstrated to perform exceptionally well on tasks involving natural language understanding. Pre-training was one of the fundamental advances of GPT, allowing the model to learn a variety of broad language representations from a vast amount of text input before being fine-tuned on tasks. Through this pre-training phase, the model can recognize pertinent information from the data that could be used for various upcoming tasks, such as information extraction, language translation, summarization, etc. Since its initial release, GPT has undergone several iterations since its initial release, including GPT-1, GPT-2, and GPT-3. GPT-1 was the model's first version, followed by GPT-2 and GPT-3. With a significantly larger model performance and a broader dataset, GPT-2, launched in 2019, performed better on various tasks. GPT-3, the most recent version of the model, was released in 2020. With 175 billion parameters, it is the largest model to date and could generalize across various domains. The GPT language model has a variation called ChatGPT that was created exclusively for conversational AI challenges. It is pre-trained on conversational data and tailored to tasks such as generating conversational responses and comprehending conversations. In general, ChatGPT is a powerful conversational AI technology that has demonstrated promising outcomes in various use cases and industries, including customer service, e-commerce, and entertainment.

Keywords: AI (Artificial Intelligence); Neural Network; Transformer architecture; GPT (Generative Pre-training Transformer); NLP (Natural Language Processing)

1. Introduction

OpenAI created the GPT (Generative Pre-trained Transformer), a particular class of language model. It can produce text comparable to what a human writer would write because it was trained on a sizable text dataset. GPT can be used for various activities involving the processing of natural languages, including text production, language translation, and summarization. It has been particularly successful in generating coherent and fluent text and has been used to create news articles, poems, and even computer code. GPT technology has the potential to revolutionize the way we interact with computers and communicate with each other, making it simple for humans to communicate with machines and for machines to understand and generate human-like text. The model is trained using a vast corpus of text data, including texts from books, papers, and websites. A few applications in which training is used to generate additional content that is like the original data and can produce human-like writing include language translation, language modeling, and chatbot creation.GPT-1 is a language generation model developed by Open AI, was trained on a huge dataset of internet text, and can produce human-like text in various styles and formats. This language generation model uses transformer architecture, a sort of neural network designed for handling sequentially (Vaswani et al.,2017). It has been used for many things, such as question-and-answer sessions and translation. It cannot carry on conversations as a chatbot could. The GPT-2 model mainly focused on utilizing a larger dataset and increasing the model's parameters to learn an even more powerful language model.

Capability transfer is available in GPT 2. The implementation of supervised learning on task-specific data sets is often applied to various natural language processing tasks, such as information extraction, machine translation, comprehension, and reading and summarization. The language model begins to understand such tasks without explicit instructions while training on Web Text. After training on a well enough large and varied dataset, a big language model can function well across several domains and datasets. The reality that GPT-2, having a 1.5B parameter Transformer, produces state-of-the-art results on 7 of the 8 tested language tasks indicates that high-capacity designs trained to maximize the probability of a well enough diverse text corpus beginning to understand how to perform an incredible number of objectives without the necessity of explicit supervision. These improvements are reflected in the model samples, which also have coherent text paragraphs but still underfit WebText (OpenAI et al., n.d.).

OpenAI created the GPT-3 model, a neural network-based language model with 175 billion parameters that produce text that resembles human speech through unsupervised learning. GPT-3 can produce text on various themes and in diverse styles because it has been pre-trained using a huge volume of text data and information. GPT-3 has been trained using a substantial dataset and several different parameters. It works effectively in zero-shot and few-shot settings for downstream NLP tasks. It can write articles that are difficult to differentiate from those authored by people. It can also perform various activities, such as adding and subtracting numbers, writing SQL queries and codes, decoding sentences made up of words, constructing React and JavaScript codes, etc. Training GPT on a smaller dataset specific to that task can improve it on certain tasks like question answering or text completion, etc. ChatGPT is a fine-tuned version of the GPT model specifically designed for conversational AI. It can be used for tasks such as language understanding and generating appropriate responses in a conversation. However, ChatGPT can still generate natural-sounding text and can be fine-tuned on specific tasks using much fewer data than GPT-3, making it more accessible and suitable for developers who want to integrate GPT-like functionality into their applications.

1.1 Objective

To introduce the GPT language model, its variation ChatGPT and their applications in conversational AI.

2. Previous Transformer based Natural Language Processing model for Conversational AI applications

Intelligent conversational computer programs known as chatbots are created to mimic human speech to provide automated online support and guidance. Chatbots have been widely used by numerous businesses to give clients virtual support due to their growing benefits. Chatbots use two branches of artificial intelligence: machine learning and natural language processing. However, there are numerous obstacles and restrictions to their use (Caldarini et al., 2022). The transformer architecture in NLP models uses self-attention processes to understand the context and relationships between words in a sentence or document. This allows the model to handle long-term dependencies and generate more coherent, context-aware language output. These transformer-based NLP models have achieved state-of-the-art results in various NLP tasks, including language modeling, text generation, and translation.

2.1 DialoGPT

DialoGPT is a conversational language model developed by Hugging Face. The GPT-2 model, a transformer-based neural network architecture, is its foundation. The model may be fine-tuned for conversational tasks like question answering, dialogue production, and text completion because it has already been pre-trained on a sizable amount of text data.

The GPT model used in DialoGPT is trained on a wide variety of internet content, enabling it to produce human-like responses with an elevated level of fluency and coherence. The model employs a transformer architecture comprising numerous feed-forward and self-attentional neural network layers. The feed-forward neural network and the model's self-attention technique allow it to process the input and forecast outcomes while weighing the importance of various terms in the input sentence. During fine-tuning, the model has been trained in a specific conversational task, such as posing and answering questions or developing a discourse. During the fine-tuning procedure, the weights of the pre-trained model are modified to suit the specific task better. Once dialoGPT has been improved, it can be utilized in various conversational applications, such as chatbots, virtual assistants, and interactive fiction. Smaller datasets can be used to fine-tune it, and it can produce human-like responses to the given environment. A substantial, fully customizable neural conversational answer-generating model is the DialoGPT model (dialogue-generative pre-trained transformer). DialoGPT enhances the Hugging Face PyTorch transformer that achieves performance equivalent to that achieved by people with both automatic and human evaluation in single-turn discussion scenarios. DialoGPT was trained using 147 million conversation-like interactions from Reddit comment chains.

Compared to strong baseline systems, DialoGPT-based conversational systems deliver more appropriate, instructive, and relevant responses to the context. Toward advanced development in neural response generation and the creation of much more intelligent open-domain conversation platforms, this pre-trained model and the training process are available to the public. A sizable real-world Reddit dataset was used to train the DialoGPT open-domain pre-trained model. A distributed training pipeline and a set of pre-trained models are included in the program. They may be modified to quickly build a conversation model on a decently big bespoke dataset. Due to DialoGPT's total open-sourcing and straightforward implementation, users can extend the pre-trained dialogue platform to bootstrap training by applying various datasets. It acts as a foundation for innovative applications and processes. Future research will put a lot of emphasis on detecting and controlling harmful production. Reinforcement learning is used to improve the relevance of the replies that are generated and stop the model from producing inappropriate responses (Zhang et al., 2020).

2.2 Natural Language Explanation-GPT (NLX-GPT)

NLX-GPT is a GPT language model version designed to explain natural language. It is qualified for tasks like responding to inquiries about complicated ideas and producing human-like explanations because it was trained on a collection of text that comprises explanations and descriptions of technical terms. The model has been developed, Hugging Face. The purpose of using natural language explanation (NLE) models is to offer a fine-grained, high-level explanation of a black box system's decision-making process that is human-friendly. Current NLE models use a language model widely recognized as an explanatory model, like the GPT, to describe the decision-making activities of a vision and a vision-language model known as a task model. Due to the complete independence between the task and explanations models, it is possible to distinguish between the justifications and the method of forecasting the answer. The task model does, however, demand extra memory resources and inference time. A generic, condensed, and realistic language model called NLX-GPT can predict and explain a response. Pre-training on many image-caption pairings is recommended for a broad understanding of images. Then, the solution is developed as a text prediction job together with the justification. It's not a bot itself, but it could make one. The resulting overall framework outperforms the current SoA (Service-Oriented Architecture) model in terms of assessment scores, significantly reduces parameters, and is 15 times faster. However, it lacks regional proposals and a task model (Sammani et al., 2022).

2.3 Meena

Meena employs many neural network designs, including the transformer-based encode and decode model and the sequence-to-sequence model. The transformer-based model is used to generate context-aware responses, while the sequence-to-sequence model is used to produce more general responses. Meena also uses a variant of the Transformer architecture called the Transformer-XL, which allows it to handle long-term dependencies in the input better. Meena uses unsupervised pre-training, where a model is initially learned on a sizable text dataset before being refined on a smaller, more focused dataset. This enables the model to acquire general language knowledge and patterns before concentrating on a particular task.

The public-domain social media interactions are mined and filtered to provide the dataset needed to train Meena. These source data effectively message trees with numerous speakers, with the initial message serving as the root and any subsequent messages serving as its child nodes. Any path through the tree results in a dialogue in which each message serves as a conversational turn. A training example in the form of a (context, response) pair is produced by considering each step in a conversation track as a response and all the prior turns (up to 7) as the context. Filtering is used on data to improve the generation quality depending on various conditions. The finalized Meena dataset contains 341GB of text (40B words). GPT-2, in contrast, has already been learned using 40GB of Internet text (Radford et al., 2019). (8 million web pages). Meena is a 2.6B parameter Evolved Transformer (ET) seq2seq model that consists of 1 ET encoder unit and 13 ET decoder units. The evolutionary NAS (Neural Architecture Search) architecture based on the transformer is called the Evolved Transformer (Real et al., 2017).

Diverse beam search is a modification of standard beam search that attempts to produce a more diverse set of responses by considering the variety of the produced responses and their likelihood. This method helps generate more human-like responses and avoid repetitive or generic ones. The decoding technique used by Meena is a re-ranking step. This post-processing step re-scores the generated responses based on a set of attributes designed to measure the quality and diversity of the responses. This step helps to further improve the quality and human-like nature of the generated responses. A multi-turn open-domain chatbot named Meena was taught from beginning to end using information gathered and extracted via public-domain social media interactions. Sensibleness and Specificity Average (SSA), a human evaluation metric that captures crucial characteristics of a human-like multi-turn discussion, is simply trained to reduce the perplexity of the following token using a neural network with 2.6B parameters. There is a significant link between confusion and SSA. Provided that the finest perplexity end-to-end trained Meena achieves a high SSA (72% on multi-turn evaluation), it is possible to get a perplexity level SSA of 86% if perplexity can be improved. Additionally, Meena's full version (which includes a filtering mechanism as well as customized decoding) performs 23% better than other chatbots in terms of absolute SSA, scoring 79% SSA (Adiwardana et al., 2020).

2.4 ChatGPT

ChatGPT, developed by OpenAI, is a variant of the GPT-2 model, an unsupervised machine learning model that can generate natural language text. Adapting a transformer architecture, a type of neural network best suited for tasks requiring natural-language processing is used to pre-train the model on a sizable amount of text. For applications like language generation, the model's ability to parse variable-length sequences and manage a lot of information, thanks to the transformer design, is helpful. ChatGPT is fine-tuned on conversational data like dialogue and question-answering; therefore, it is more efficient in that domain. The model can understand the context and respond appropriately to prompts in a conversation. It uses sequence-to-sequence generation with a decoder part to generate the most likely output sequence. The model employs a technique known as the masked language model, in which some phrase input components are hidden, and the decoder portion is trained to anticipate which words are absent. Natural language processing activities, including text generation, text completing, and question answering, can be performed with ChatGPT. It is used in conversational AI and chatbot applications to generate human-like responses. Reinforcement learning is a kind of machine learning in which an agent learns to interact with its environment to maximize a reward. It is based on the idea of an agent taking actions in an environment and receiving a reward or punishment based on those actions. The objective of reinforcement learning is to discover a strategy that, over time, maximizes the total reward (GPT-3.5 + ChatGPT: An Illustrated Overview, n.d.).

ChatGPT's architecture is based on transformer architecture. Self-attention layers in the transformer design enable the model to evaluate the relative weight of several phrases or words in each input. As an outcome, the model can understand the context and importance of the input more clearly and produce more compelling and coherent responses. In addition to the self-attention levels, a transformer architecture also features feed-forward layers with residual connections. These components allow the model to identify more intricate patterns within the data and, more precisely, represent the relationships between different words or sentences. It can recognize a discussion's context and provide appropriate responses in line with that context. It can then connect with consumers more fascinatingly and organically. As a result, the model is provided a big textual dataset during the pre-training phase and taught to predict the subsequent word for every sequence. As a result, the model may learn the relationships between various words and phrases and the language's patterns and structure. (Majid, 2022c).

The model used to train ChatGPT, which ended training in early 2022, is from the GPT-3.5 series. The model uses the same techniques as InstructGPT, Reinforcement Learning from Human Feedback (RLHF), with minor variations in the data-gathering configuration. The initial model was trained through supervised fine-tuning, in which human AI instructors gave interactions in which they took on the roles of both the user and the AI assistant. Trainers have access to sample written recommendations to use as a guide while composing their replies. The InstructGPT dataset, converted into a conversation format, is combined with this new dialogue dataset. It was necessary to gather comparison data, which consisted of two or more model replies graded by quality, to develop a reward model for reinforcement learning. Conversations between AI trainers and the chatbot are recorded to get this data. A model-written message is chosen randomly from a pool of potential completions after being ranked by AI trainers. Using these reward models, proximal policy optimization is used to fine-tune the model. After being pre-trained, ChatGPT can be tuned on specific tasks using much fewer data than GPT-3, which makes it more accessible and suitable for developers who want to integrate GPT-like functionality into their applications. It can be used for many NLP applications, including chatbots, text-to-speech, and language translation, and is especially helpful for tasks like question-answering dialogue systems and text completion. The model can produce text close to human-written material because it can grasp the context and produce cohesive, coherent prose. ChatGPT is an effective tool for natural language production since it can generate text in a variety of styles and on certain subjects, as well as adjust to the input given (Ouyang et al., 2022).

Artificial intelligence (AI) has greatly enhanced natural language understanding (NLU) along with human-computer interaction (HCI) through the development of large natural language models, including writing and dialogue capabilities. Currently, OpenAI's GPT-3 model is the language that provides the greatest capability, scope, and number of features. After being trained on a huge amount of material from the Internet and hundreds of books, GPT-3 can surprisingly accurately emulate human language patterns. Given its incredible realism, this language model is possibly the most remarkable one currently in use. Despite possessing outstanding descriptive and modeling abilities, there are significant obstacles and constraints. GPT-3 model has trouble producing manageable text (natural language production), and it does so occasionally. The high processing requirements, data needs, and capital expenses associated with constructing the GPT-3 model also result in significant carbon dioxide emissions (Zhang & Li, 2021).

3. Advantages and Limitations of using GPT models in the field of AI

3.1 Advantages

Transfer learning: GPT models are suitable for transfer learning, which means that the model can apply the skills it acquires in one job to perform better in another.

Human-like responses: GPT models can generate human-like responses, making them well-suited for chatbots and virtual assistants.3.2 Limitations of chatbots designed using GPT models:

High performance: GPT models can produce high-quality text and comprehend inputs in natural language because they have been trained on big datasets. It has been demonstrated that they excel at tasks requiring natural language understanding, including answering questions, text production, and language translation.

Pre-training: GPT models can be customized for applications with only minimal data because they have already been trained. As a result, they are more effective and economical to use than models trained from the start.

Continuous learning: GPT models are made to keep picking up the latest information from new data, letting them adapt and enhance their performance over time.

3.2 Limitations

Lack of Domain Knowledge: GPT models are trained on various texts. These models may need more specific domain knowledge in some fields or businesses. As a result, they may need to improve in use situations where domain expertise is necessary.

Lack of Explain ability: The neural networks that form the basis of GPT models can be challenging to analyze and comprehend. Due to this, it may be difficult to comprehend how the model generates its results or to find any flaws or biases in the model.

Costs of computation: GPT models are computationally expensive and demand a lot of data and processing power to run and train. They may become less reachable because of specific groups or people.

Language Model bias: GPT model can generate biased or offensive responses, as it is pre-trained on the internet data, which can contain biases and stereotypes. Therefore, it's crucial to fine-tune the model using data from the relevant area.

4. Diverse Applications of GPT Models

GPT can improve the conversational ability of chatbots by generating more natural and human-like responses. This can be measured by evaluating the generated text's coherence, relevance, and overall quality.

Fine-tuning GPT models on task-specific datasets can improve the model's performance on that task. The improvement can be measured using accuracy, precision, recall, and F1-score metrics.

GPT is useful for content generation in journalism and creative writing, as it can produce new text based on a given prompt with high fluency and coherence. The quality of the generated content can be evaluated based on metrics such as readability, coherence, and novelty.

Using GPT for text summarization, important phrases, and sentences can be identified and condensed into a shorter, more manageable format. The effectiveness of the summarization can be evaluated by comparing the summary with the original text and measuring the level of information loss and coherence.

GPT can be fine-tuned on different datasets for various NLP tasks, including sentiment analysis, question-answering, machine translation, text completion, and named entity recognition. Measuring parameters for these tasks include accuracy, precision, recall, F1-score, and BLEU score for machine translation.

Additionally, the GPT model can be fine-tuned on specific use cases by training on a dataset that is specific to that use case.

5. Conclusion

The field of natural language processing has undergone a revolution thanks to the Generative Pre-trained Transformer (GPT), a new generation of language models (NLP). The GPT technology has also inspired the development of other models, which have achieved state-of-the-art performance in NLP tasks. As GPT technology continues to evolve, we expect to see even more impressive results and new applications soon. This research has evaluated current developments in creating and using chatbots and models that employ GPT models. These models have been used for several purposes, such as text creation, language translation, customer service, and support, and they have shown success in providing human-like responses. The usage of GPT models in chatbot generation is anticipated to increase in response to the developments in natural language processing and the growing accessibility of vast data. The use of GPT-based chatbots or models is anticipated to increase in the next years due to continued developments in natural language processing and the growing accessibility of massive datasets. A GPT model can be fine-tuned on a particular dataset to help it understand that subject or topic's language, terminology, and context. This can enhance the model's comprehension and production of material in that area or domain. However, additional research is necessary to ascertain how to improve these models' performance in particular areas to handle better scenarios found in the actual world.

References

Adiwardana, D., Luong, M., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., & Le, Q. V. (2020). Towards a Human-like Open-Domain Chatbot. *ArXiv: Computation and Language*. https://arxiv.org/pdf/2001.09977.pdf

Caldarini, G., Jaf, S., & McGarry, K. (2022). A Literature Survey of Recent Advances in Chatbots. *Information*, *13*(1), 41. https://doi.org/10.3390/info13010041

GPT-3.5 + ChatGPT: An illustrated overview. (n.d.). Dr Alan D. Thompson – Life Architect. https://lifearchitect.ai/chatgpt/

Majid, U. (2022c, December 29). *How Chat GPT utilizes the advancements in Artificial Intelligence to create a revolutionary language model*. Pegasus One. https://www.pegasusone.com/how-chat-gpt-utilizes-the-advancements-in-artificial-intelligence-to-create-a-revolutionary-language-model/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Cornell University - ArXiv*. https://doi.org/10.48550/arxiv.2203.02155

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). Improving Language Understanding by Generative Pre-Training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y. L., Tan, J., Le, Q. V., & Kurakin, A. (2017c). Large-scale evolution of image classifiers. *International Conference on Machine Learning*, 2902–2911. http://proceedings.mlr.press/v70/real17a/real17a.pdf

Sammani, F., Mukherjee, T., & Deligiannis, N. (2022b). NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks. *Cornell University - ArXiv*. https://doi.org/10.48550/arxiv.2203.05081

So, D. R., Liang, C., & Le, Q. V. (2019). The Evolved Transformer. *ArXiv: Learning*. https://arxiv.org/pdf/1901.11117

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020). DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://doi.org/10.18653/v1/2020.acl-demos.30

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf