Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Commentary

A commentary of GPT-3 in *MIT Technology Review 2021*

Min Zhang[a,*], Juntao Li[b,*]

[a] *Research Center for Human Language Technology, School of Computer Science and Technology, Soochow University, Suzhou 215006, China*
[b] *Institute of Artificial Intelligence, Soochow University, Suzhou 215006, China*

ABSTRACT

Through the development of large-scale natural language models with writing and dialogue capabilities, artificial intelligence (AI) has taken a significant stride towards better natural language understanding (NLU) and human-computer interaction (HCI). As of today, the GPT-3 model, developed by OpenAI, is the language model with the most parameters, the largest scale, and the strongest capabilities. Using a large amount of Internet text data and thousands of books for model training, GPT-3 can imitate the natural language patterns of humans nearly perfectly. This language model is extremely realistic and is considered the most impressive model as of today.

Despite its powerful modeling and description capabilities, there are significant issues and limitations. First and foremost, the GPT-3 model does not understand writing (natural language generation) well and sometimes generates uncontrollable content. Secondly, training the GPT-3 model requires a large amount of computing power, data, and capital investment, and releases significant carbon dioxide emissions. Developing similar models is only possible in laboratories with adequate resources. Furthermore, as the GPT-3 model is trained with Internet text data rife with error messages and prejudices, it often produces chapters and paragraphs with biased content similar to the training data.[①]

(1) What qualifies it as one of the 10 breakthrough technologies of 2021?

AI has emerged as an essential pillar for socio-economic development and social progress. It is also a strategic technology, which is driving the new era of technological revolution alongside industrial and social changes. A core component of the next generation of AI is NLU, and a breakthrough in its core technology is of great scientific significance and industrial value. Language modeling includes performing abstract and mathematical modeling of natural language using computers, which is one of the principal scientific challenges of NLU. NLU models, in a broad sense, can all be considered as language models from the understanding of mathematically modeling language. However, in a narrow sense, a language model must estimate the probability of a passage, or the probability of occurrence of a certain language segment, or an abstract mathematical representation in a given context. Typically, a narrow language model is referred to when discussing language models. The history of language models can be traced back to the N-Gram model developed in 1948, followed by the distributed theoretical bag-of-words model in 1954, the distributed representation model in 1986, the Word2Vec model in 2013, and finally, the pre-trained model in 2018. In the era of deep learning, pre-trained language models (represented by ELMo, BERT, and GPT) have profoundly impacted the field of natural language processing (NLP), and are regarded as a milestone in NLP.

These deep-learning-based models merely leverage unsupervised language model training objectives to learn and capture myriads of valuable information from massive text data, which can dynamically generate more accurate vector representations and probabilities of words, phrases, sentences, and paragraphs with contextual information. They also have the potential to achieve amazing results on a variety of downstream tasks, including question answering, reading comprehension, text implication, semantic similarity matching, text summarization, code generation, story creation, and more. In addition to their powerful representation learning capabilities and multi-task generalizability, these pre-trained language models also possess powerful few-shot learning capabilities, which allow them to learn a particular task from very few data samples (even under zero-shot setting) and achieve a performance comparable to or superior to supervised learning models. Among the numerous models, the 3rd-generation GPT model (GPT-3) proposed by OpenAI in May 2020 was selected among the “Top 10 Breakthrough Technologies” by MIT Technology Review in 2021, attributing to its extensive parameter scale, exceptional modeling ability, multi-task generalization performance, and few-shot learning ability.

(2) The development process and capability evolution of GPT and other model series

Pre-trained language models are diverse, in which typical milestone models include ELMo, BERT, and GPT. Due to space limitations, only a representative introduction of the GPT series is provided herein.

We will first review the original aim and intermediate development process of pre-training language models before analyzing GPT-3. Traditional language models such as N-Gram use a traditional discrete statistical model based on frequency to calculate the probability of occurrence of a given language segment or to predict the probability of occurrence of the next word given the previous context. There are three

[*] Corresponding authors.
E-mail addresses: minzhang@suda.edu.cn (M. Zhang), ljt@suda.edu.cn (J. Li).

[①] Original source in Chinese: M. Zhang, J. Li, GPT-3, Bulletin of National Natural Science Foundation of China. 35 (3) (2021) 403-406.

https://doi.org/10.1016/j.fmre.2021.11.011

GPT-3 [1].

main limitations of this model. The discrete word representation method has poor description ability, the parameter space expands exponentially, and frequency-based statistical probability models have poor modeling capability, thereby resulting in poor description ability, low robustness, and low accuracy. To overcome these problems, pre-trained language models such as ELMo, BERT, and GPT, have employed large-scale or even networkwide data, evolved from generative language models or masked language models, and used neural networks to train language models. In this manner, the pre-trained language models not only generate the probability outcome that resembles traditional models but also generate a vector representation of the language segment. With neural networks as the specific implementation, powerful mathematical tools (such as derivable and differential methods), as well as extremely large data sets can be utilized. Therefore, pre-trained language models exhibit superior context modeling ability and can deliver a more accurate probability calculation and a dynamic vector representation of language fragments with a strong context correlation.

With the introduction of ELMo, the era of 2nd generation pre-trained language models has begun, i.e., context-sensitive and “pretraining + fine-tuning”. ELMo is a generative model that uses bi-directional LSTM as a feature extractor and performs dynamic modeling based on the context. Compared to the first generation of pre-trained language models represented by Word2Vec, ELMo can handle polysemy more effectively and is adept at generating natural language. BERT is a masked language model that uses a transformer encoder as the feature extractor and is particularly proficient at analyzing and understanding natural language. The GPT is a generative model that also uses a transformer decoder as the feature extractor and exhibits superior performance in natural language generation tasks.

Prior to the above series of models, downstream tasks represented by NLU primarily used supervised learning to train models on corresponding labeled data sets. Each target task required sufficient labeled data, and the trained model could not easily generalize to other tasks. With insufficient data, these types of discriminative models could not deliver satisfactory results. To address this problem, the OpenAI team proposed the 1st generation of the generative pre-trained language model (GPT-1), which is a generative language model based on a transformer decoder. No new features have been added to the model structure, but the level of complexity has increased. For this type of pre-trained language model, only the unsupervised language objective function is needed during training, hence massive unlabeled data can be used to train the model. Additionally, the GPT-1 model unifies the formats of various input data when enhancing downstream tasks to achieve minimum modification to the model structure. Owing to these two properties, GPT-1 only needs simple fine-tuning and supervised training for use in downstream tasks, and it can achieve significant performance improvements, which have demonstrated the powerful generalization ability of the GPT language model. Tests have also shown that in the zero-shot setting, GPT-1 still has some generalization ability. These results have shown the power of GPT models and paved the way for future versions of the model requiring larger parameter scales and additional training data.

Based on GPT-1, GPT-2 included five minor improvements to the model structure, added more training data, further improved the generalization ability, and addressed the need for supervised fine-tuning training when using GPT-1 for downstream tasks. By introducing task information during model training, using more training data than the GPT-1 model (40 GB vs. 5 GB), and building a model with a larger parameter scale (1.5 billion vs. 117 million), GPT-2 has outperformed frontier models in a variety of downstream tasks, including machine translation, reading comprehension, and long-distance dependency modeling. These features of the GPT-2 model have suggested that a larger model capacity and more training data can enhance the ability of the model to generalize and reduce the dependence on supervised training. Furthermore, compared with the training data, the GPT-2 model is still underfitted, hence an increase in the parameter scale of the model is needed.

GPT-3, a successor to GPT-2, further expanded the parameter space (175 billion vs. 1.5 billion) and the data scale (45 TB vs. 40 GB), thus making it the largest language model ever created. The model can perform downstream tasks without fine-tuning and has an excellent performance in zero-shot and few-shot settings. Based on the multi-task generalization abilities of GPT-2, GPT-3 has achieved excellent results on numerous new tasks, including mathematical addition, news article generation, vocabulary interpretation, and code writing. With an increasing number of parameters, this model will become even more powerful.

(3) Discussion of the root causes of success and limitations

Based on a comparison of the original design and development processes of the three generations of GPT models, we can observe that all three generations are based on the transformer decoder. The powerful capabilities of GPT-3 are based on the scale effect, i.e., super generalization ability is only possible by increasing the scale of the model and the training data. Essentially, GPT-3 is still a data-driven model, which uses a large-capacity model to fit massive amounts of data and achieve model convergence. Therefore, the GPT series models have reflected the characteristics of a data-driven model. Hence, the capability of the model is influenced by the coverage, distribution, and quality of the fitted data. Irrespective of whether the data is new, has a different distribution, or has noise, such issues would be catastrophic for the model. According to the latest test results, the GPT-3 model has not met expectations in natural language reasoning, filling in the blanks, long text generation, and reading comprehension tasks, indicating that it is at the data fitting level and does not have a genuine comprehension of natural language. Furthermore, due to the limited quality of Internet data, GPT-3 may produce biased and disturbing content. These inferences suggest that GPT-3 is still at the stage of perceptual intelligence, and a significant amount of development is required before it reaches general intelligence and cognitive intelligence. Hence, GPT-3 is considered to have “memory with a certain ability to generalize,” due to which it easily retains and learns declarative knowledge, rather than understand knowledge, and does not possess true logical reasoning ability or the ability to distinguish right from wrong.

(4) The significance of GPT-3

Although the GPT-3 model does not currently possess the intention or the capability to respond to requests in the real world, it has had an enormous impact on the field of AI. Nearly 10 years have elapsed since deep learning began its explosive foray in various fields in 2012. However, the development of new technologies and algorithms has reached a bottleneck. Data-driven models have reached a plateau with regards to their effects and capabilities. The advent of the GPT-3 model has introduced a stimulant in the field of deep learning and has inspired new thinking. The first question to ask is whether the expansion of capacity that occurs with increasing model scale is predictable and stable. Short-term results indicate that this scale effect will continue to increase the ceiling for deep learning as computing power increases in computer hardware. The other questions to ask are: for deep learning, where is the limit? Eventually, will this model be able to truly understand human language? Finally, can deep learning lead to true AI? Is it possible to develop cognitive abilities and general intelligence?

The GPT-3 model offers many powerful functions that can cope with many practical application scenarios, such as question answering, reading comprehension, summary generation, automatic chat, search matching, code generation, and article generation. As the GPT-3 model presents security and uncontrollability problems, including false content and biased information during content generation, its application value is primarily reflected in intelligent auxiliary tasks, and it cannot directly interface with the end-user. For example, in tasks such as summarizing reports, creating content, and writing, GPT-3 can be used to generate corresponding content according to the task description. Additional manual review and editing are then performed to present the final edited report to the end-user. Furthermore, GPT-3 can be employed for developing games and in other situations that do not have clear task definitions and goals.

(5) Future directions for research and the relevant situation in China

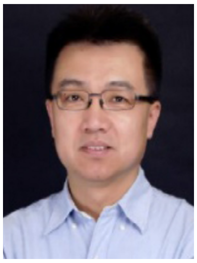
Pre-trained models, such as GPT-3, still have varying engineering applications, ethical, and social problems. In addition, for developing such models, there are additional challenges including interdisciplinary cooperation, open sharing, and resource imbalance. In this regard, China has created a long-term plan and corresponding layout and obtained promising preliminary results. The ultra-large-scale intelligent model system represented by “Wudao” and Pangu” has made breakthroughs in model effects, domain transplantation and generalization, small models, model training efficiency, multi-language, weakly correlated multi-modal pretraining, generalization, controllability, knowledge integration, protein sequence prediction, and other scenarios. It is believed that China will be the world leader in AI fundamental technology innovation, talent and team building, and open source community within the next 10–20 years.

Declaration of Competing Interest

The authors declare that they do not have any conflicts of interest in this work.

Reference

[1] 10 Breakthrough Technologies 2021. https://www.technologyreview.com/2021/02/24/1014369/10-breakthrough-technologies-2021, 2021 (accessed 24 March 2021).



Min Zhang is a Distinguished Professor at Soochow University (China). He is also the Dean of School of Computer Science & Technology and School of Software and the deputy director of the Institute of Artificial Intelligence. He is funded by China National Funds for Distinguished Young Scientists. He received his bachelor’s degree and Ph.D. degree from the Harbin Institute of Technology in 1991 and 1997, respectively. His current research interests include machine translation, natural language processing, information extraction, text generation, dialogue system, large-scale text processing, intelligent computing, and machine learning.

Juntao Li is an associate professor at the Institute of Artificial Intelligence, Soochow University. He received his doctoral degree from Peking University in 2020. His research interests focus on text generation, dialogue systems, and artistic writing (e.g., poetry generation, story generation). He has published over 10 papers as a leading author in TOIS, ACL, EMNLP, AAAI, IJCAI. He has given two tutorials on IJCAI and AAAI. He also serves as a reviewer for Computational Linguistics, TALLIP, TKDE, International Journal of Intelligent Systems, ACL(Area Chair), EMNLP, NAACL, AAAI, IJCAI(Senior PC), etc.