WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2025 Volume 3, Issue 1

Computer Simulation in Application

ISSN: 2630-4597 (Online)

Overview of Large Language Models

Jiaxin Li, Fang He, Xiaowei Shen, Yiwei Wei, Haojie Hu, Zhengzhong Cha
Rocket Force University of Engineering, Xi'an 710025, Shaanxi, China

Abstract: This article focuses on typical large language models and conducts an in-depth analysis of their definitions, typical models and the development status of their technologies. As an advanced artificial intelligence technology, large language models are trained based on huge parameters and massive data, and achieve natural language processing with the converter structure as the core. This article elaborates on the developing history and features of various large language models. Meanwhile, it is pointed out that the development of large language model technology faces problems as well as challenges such as non-authentic output and security risks, and in the future, it will develop in the directions of lightweight, multimodal and vertical specialization. The research aims to provide references for the further study and application of large language models and contribute to promoting the healthy development of this technology in various fields.

Keywords: Large language model; Development history; Application scenarios; Technical challenges

Online publication: March 26, 2025

1. Definition of Large Language Models

Large language model technology is an advanced artificial intelligence technology with multiple natural language processing capabilities, including understanding, recognizing, translating, predicting and generating text or other content[1]. These models are called "big" because they use a huge number of parameters and are trained on large-scale datasets[2]. This enables large models to deeply understand and express human language and other complex data types.

Essentially, large language models are a form of deep neural network models. Its computing system is inspired by the human brain and consists of multiple nodes, which are distributed at different levels, similar to neurons. The core of large language models is a special neural network structure called the Transformer. This structure consists of two parts: the encoder and the decoder. The encoder is responsible for converting the input text into a series of vectors, which fully capture the semantic and grammatical information of the text. Subsequently, the decoder will generate output text based on these vectors, such as for answering questions or writing articles.

Large language models are trained on massive amounts of data to help artificial intelligence predict user prompts and generate human-like responses. They power generative artificial intelligence tools such as ChatGPT from OpenAI and Bard from Google.

2. The development history of large models

2.1. Germination stage

In 1956, the concept of "artificial intelligence" was proposed at the Dartmouth Conference. In 1980, the prototype of convolutional neural networks, CNN, was born.

The construction of CNN is as follows:

(1) Input layer: The input layer receives the original image data. Images are usually composed of three color channels (red, green, and blue), forming a two-dimensional matrix that represents the intensity values of pixels.

(2) Convolution and activation: The convolutional layer performs convolution operations on the input image and the convolutional kernel. Then, nonlinearity is introduced by applying activation functions (such as ReLU). This step enables the network to learn complex features.

(3) Pooling layers: The pooling layer reduces the computational complexity by reducing the size of the feature map. It is achieved by selecting the maximum or average value within the pooling window. This is helpful for extracting the most important features.

(4) layers stacked: CNN is usually composed of a stack of multiple convolution and pooling layers to gradually extract higher-level features. Deep-level features can represent more complex patterns.

(5) Full connection and output: Finally, the fully connected layer converts the extracted feature mapping into the final output of the network. This can be a classification label, a regression value or the result of other tasks[3].

In 1998, LeNet-5, the basic structure of modern convolutional neural networks, was born. Machine learning methods changed from early models based on shallow machine learning to models based on deep learning, laying the foundation for the subsequent development of large models.

2.2. Sedimentation stage

In 2013, the natural language processing model Word2Vec was born, proposing the "word vector model". It's found in word embedding, which is designed based on a distributed hypothesis. Words with similar meanings tend to have the same word embedding. [4].

In 2014, Generative Adversarial Network (GAN) was born, marking a new stage in the research of generative models for deep learning. GAN is composed of a generator (G) and a discriminator (D), each instantiated through neural networks[5].

In 2017, Google proposed the Transformer architecture, laying the foundation for the architecture of large model pre-training algorithms.

In 2018, OpenAI and Google respectively released the GPT-1 and BERT large models, and pre-trained large models have become the mainstream in the field of natural language processing. ChatGPT is capable of providing a detailed response, demonstrating powerful performance in multi-language understanding and generative tasks, such as modifying code and writing articles. In 2019, GPT-2 was released, introducing the idea of multi-task learning. Its parameters soared from 110 million of GPT-1 to 1.5 billion, and its text generation capability was significantly enhanced[6].

2.3. Outbreak period

In 2020, OpenAI launched GPT-3, with a model parameter scale of 175 billion and a significant improvement in performance on zero-shot learning tasks. It's the first large language model to exceed a parameter scale of 100 billion. The size of context window has increased from 1024 tokens in GPT-2 to 2048 tokens in GPT-3. GPT-3 combined meta-learning with in-context learning, enhancing the generalization ability.

In January 2021, OpenAI launched DALL-E. DALL-E is an artificial intelligence image generator, which can create images and art forms based on the textual descriptions of natural language. In other words, it is an artificial intelligence system that generates images based on text.

In February 2022, Google published LaMDA. In addition to improving the quality of the generated text, LaMDA has achieved significant enhancements in the key issue of security by further fine-tuning the manually labeled data and enabling the model to learn the ability to retrieve and utilize external knowledge sources.

On November 30, 2022, ChatGPT equipped with GPT-3.5 was launched. Within two months, it surpassed 100 million users, attracting widespread attention. 2022 was also hailed as the first year of large models. In March 2023, GPT-4 was released. GPT-4 is a large multimodal model that accepting image and text inputs and emitting text outputs.

In addition, various large models such as Ernie Bot launched by Baidu, Llama and SAM launched by Meta, have also emerged one after another. Ernie model is based on word feature input, so that it doesn't need to rely on other information in application, depicting stronger versatility and extensibility[7].

Llama supports for local deployment, which means that small companies can also use this open-source model to implement functions that are more suitable for themselves.

SAM focuses on the task of image segmentation. Its technical architecture adopts an encoder-decoder design to achieve efficient segmentation of any object by processing image features with prompt information (such as points, boxes, and text).

3. Problems and development suggestions faced by the Development of large language model technologies

With the vigorous development of large language models, corresponding risks and challenges will inevitably arise. To solve these problems accompanying the development of large language models, the suggestions put forward in this paper are as follows

3.1. Strengthen technological innovation and R&D investment, enhance the performance and stability of large language models, and promote the wide application of large model technology.

Technically, large language models have non-authentic and biased outputs, lack real-time autonomous learning capabilities, require huge computing power support,[8] strongly rely on the quality and quantity of datasets, and present a single language style. In terms of security, there are risks in terms of data privacy, information security, ethical issues and the cost of crime[9]. The future development of large language models will technically shift from "large-scale" to "lightweight" in order to reduce the training cost of large language models. Move from "single-modal" to "multi-modal" to better align with human intentions; From "general" to "vertical", we aim to address industry demands more deeply at a lower cost.

3.2. Strengthen computing resources and technical support, establish a complete big data platform and computing infrastructure to meet the demands of training and reasoning for large language models.

Since large language models have high requirements for computing power, it is necessary to continuously strengthen the construction of computing power. By enhancing computing power, the training and reasoning process of the model can be accelerated, thereby improving the performance and response speed of large language models.

3.3. Strengthen data security and privacy protection, and adopt measures such as encryption, permission control and security monitoring to ensure the security and integrity of data.

Another important issue related to the security of large language models is the timing of regulation. Premature regulatory intervention may hinder the development of AI. Under the premise of ensuring safety, it is necessary to avoid the obstruction of its development. Judging from the current trend, the regulation of AI research and development in our country is relatively lenient. The practice in this regard mainly presents a model of "governance while developing". To achieve this goal, it is necessary to build a standard model, develop specific datasets, and establish basic security standards. In addition, the deep integration of social sciences and AI is also an indispensable key link in the development process of

AI, especially in subjects related to regulation.

4. Conclusion

Large language models (LLMs), with their vast parameters and extensive training data, excel in natural language processing and have broad applications[10]. This paper reviews the architectures, features, performances and applications of several typical large language models, fully demonstrating their application potential.

However, LLM development faces hurdles. Technically, issues such as output inaccuracies, biases, lack of real-time learning, high computational demands, dataset reliance, and monotonous language persist. Security-wise, concerns over data privacy, information security, ethics, and criminal risks remain significant.

Future LLM development will focus on "lightweighting," "multimodality," and "verticalization." Through R&D investment and technological innovation, LLMs aim to boost performance, secure data, and meet computing needs. These trends will foster deeper integration into society, driving economic and social progress, but also necessitate proactive problem-solving for sustainable growth.

Disclosure statement

The author declares no conflict of interest.

References

[1] A. Singh, "Exploring Language Models: A Comprehensive Survey and Analysis," 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-4.

[2] D. H. Anh, D. -T. Do, V. Tran and N. L. Minh, "The Impact of Large Language Modeling on Natural Language Processing in Legal Texts: A Comprehensive Survey," 2023 15th International Conference on Knowledge and Systems Engineering (KSE), Hanoi, Vietnam, 2023, pp. 1-7.

[3] A. Aggarwal, V. Kumar and R. Gupta, "Object Detection Based Approaches in Image Classification: A Brief Overview," 2023 IEEE Guwahati Subsection Conference (GCON), Guwahati, India, 2023, pp. 1-6.

[4] L. N. T. Manalu, M. Arif Bijaksana and A. A. Suryani, "Analysis of the Word2Vec Model for Semantic Similarities in Indonesian Words," 2019 7th International Conference on Information and Communication Technology (ICoICT), Kuala Lumpur, Malaysia, 2019, pp. 1-5.

[5] Q. Lin, T. Li, Y. Zhao, J. Guan, W. Zhang and X. Wang, "Research on Charging Infrastructure Related Detection Technology Based on GAN," 2023 2nd International Conference on Clean Energy Storage and Power Engineering (CESPE), Xi'an, China, 2023, pp. 57-62.

[6] T. Wu et al., "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development," in IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, pp. 1122-1136, May 2023.

[7] J. Li, D. Zhang and A. Wulamu, "Chinese Text Classification Based on ERNIE-RNN," 2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT), Sanya, China, 2021, pp. 368-372.

[8] G. Mani and G. B. Namomsa, "Large Language Models (LLMs): Representation Matters, Low-Resource Languages and Multi-Modal Architecture," 2023 IEEE AFRICON, Nairobi, Kenya, 2023, pp. 1-6.

[9] Q. Wu and Y. Wang, "Research on Intelligent Question-Answering Systems Based on Large Language Models and Knowledge Graphs," 2023 16th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou,

China, 2023, pp. 161-164.

[10] M. Zhou, W. Chen, S. Zhu, T. Cai, J. Yu and G. Dai, "Application of large language models in professional fields," 2023 11th International Conference on Information Systems and Computing Technology (ISCTech), Qingdao, China, 2023, pp. 142-146.

Publisher's note

Whioce Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.