A Review of Large Models in Computer Science

1023041002 Shitao Wang
Nanjing University of Posts and Telecommunications
Nanjing, China
1023041002@njupt.edu.cn

***Abstract*—In recent years, large models have achieved significant progress in the field of artificial intelligence, particularly in subfields like natural language processing (NLP) and computer vision (CV). This paper reviews the definition, development history, key technologies, and performance of large models in practical applications. It focuses on the advantages and challenges of large models and explores future development directions.**

***Index Terms*—Large models, natural language processing, computer vision, deep learning, artificial intelligence**

I. Introduction

The scale and complexity of large models, especially deep learning models, have grown rapidly over the past decade, driving breakthroughs in various fields of artificial intelligence. From AlexNet's victory in the ImageNet competition in 2012 to the revolution led by language models like GPT-3, large models have become a critical force in technological advancement. This paper aims to systematically review the development history of large models, analyze their technical characteristics, discuss their applications, and address the challenges they face.

II. Definition and Development History of Large Models

A. Definition

Large models generally refer to deep learning models with a vast number of parameters and high computational complexity. Their typical characteristics include: parameters reaching hundreds of millions or even billions, complex model structures, large-scale training datasets, long training times, and substantial computational resources.

B. Development History

- Early Stage (2012-2014): The success of AlexNet marked the beginning of the era of large models. It utilized deep convolutional neural networks (CNNs) and achieved unprecedented accuracy in image classification tasks.
- Mid Stage (2015-2018): Models like ResNet and Inception further improved performance by introducing residual networks and modular structures. During this period, Generative Adversarial Networks (GANs) and Long Short-Term Memory (LSTM) networks also excelled in generative and sequence tasks.
- Recent Stage (2019-Present): The emergence and application of the Transformer architecture led to breakthroughs in NLP tasks with models like BERT and GPT-3. Large-scale pretraining and fine-tuning techniques have become mainstream.

III. Key Technologies

A. Transformer Architecture

The Transformer, proposed by Vaswani et al. in 2017, is a novel neural network architecture primarily used for NLP tasks. Its core idea is based on the self-attention mechanism, which efficiently handles long-range dependencies. GPT-3 and BERT are representative models based on the Transformer.

B. Pretraining and Fine-tuning

Pretraining and fine-tuning are crucial techniques for training large models. By pretraining on large datasets, models can learn general feature representations and then be fine-tuned to adapt to specific tasks, significantly improving their generalization ability and performance.

C. Model Parallelism and Distributed Training

Given the high computational demands of large models, single-machine training is no longer sufficient. Model parallelism and distributed training techniques divide the model and data across multiple computational nodes to parallelize processing, greatly enhancing training efficiency.

IV. Applications

A. Natural Language Processing

Large models perform exceptionally well in NLP tasks such as machine translation, text generation, and sentiment analysis. GPT-3, with its powerful generation capabilities, has achieved state-of-the-art performance in various NLP tasks.

B. Computer Vision

In the CV field, large models have achieved significant progress in tasks such as image classification, object detection, and image generation. Models like ResNet and EfficientNet are widely used in various vision tasks.

C. Other Fields

Large models are also applied in speech recognition, recommendation systems, game AI, and other fields, demonstrating their broad applicability and powerful performance.

V. Challenges and Future Directions

A. Challenges

- Computational Resource Demand: The training of large models requires extremely high computational resources and energy consumption, limiting their widespread adoption.
- Data Dependency: Large models rely on large-scale, high-quality data, posing challenges in data acquisition and processing.
- Model Interpretability: The complexity of large models makes their internal workings difficult to interpret, affecting their application in critical areas.

B. Future Directions

- Model Compression and Optimization: Techniques like model pruning and quantization reduce the number of parameters and computational demands, improving model efficiency.
- Multimodal Models: Combining multiple data types (such as text, images, and audio) in multimodal models will further enhance AI's comprehensive capabilities.
- Automated Machine Learning (AutoML): AutoML technology can automatically optimize model structures and parameters, reducing the difficulty of developing large models.

Conclusion

Large models in computer science have demonstrated their powerful capabilities and broad application prospects in various fields. Despite numerous challenges, continuous technological advancements and innovations will enable large models to achieve further breakthroughs, injecting new momentum into the development of artificial intelligence.

Acknowledgment

Shitao Wang thanks the Department of Computer Science and Technology at Nanjing University of Posts and Telecommunications for their support in this research. Special thanks to Prof. Wang for his insightful feedback and to my colleagues for their assistance in data collection and analysis. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62076212.

References

[1] Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems.

[2] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems.

[3] Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.