NRI

Will future generative AI models be larger or smaller?

Ryoji Kashiwagi

10 June 2024

lakyara vol.387

Nomura Research Institute, Ltd.

Executive Summary

The current wave of generative AI models is founded on the Transformer architecture, heralded by the rise of large language models (LLMs). However, despite their prominence, LLMs exhibit inherent drawbacks and constraints. In response to these issues with LLMs, researchers are ramping up development of small language models that could be a game changer for generative AI.

Ryoji Kashiwagi
Expert Researcher
Financial Market & Digital
Business Research Department

The revolutionary Transformer model

In 2017, eight AI researchers published a paper entitled *Attention Is All You Need*[1] that catalyzed a big AI breakthrough by proposing a novel AI architecture called Transformer. The Transformer architecture opened up new possibilities for large language models (LLMs), dramatically boosting existing deep learning models' performance and unleashing the ongoing wave of progress in generative AI. Transformer models are now used in nearly all of the LLMs offered by major generative AI providers, including OpenAI, Google, Microsoft and Meta.

The Transformer architecture has revolutionized deep learning. Prior to its advent, deep learning models analyzed textual and image data by computing simple word-order and pixel-order relationships (correlations). Transformer models, by contrast, are able to more precisely quantify interrelationships among data being analyzed (position refinement) in addition to bidirectionally expanding the analyzable length of data sequences when analyzing the data's structure (distance expansion).

By virtue of these innovations, Transformer models adhere to scaling laws, meaning that the more data they are trained on (and the more computations they perform), the more accurate their output. Generative AIs that use Transformer models have demonstrated stunning performance generated by ingesting exponentially large training data sets, using huge numbers of GPUs and consuming lots of electricity. Generative AI's rapid progress since Transformer models' advent is well known.

LLMs' limitations

The sustainability of LMMs' rate of progress, however, has been called into

NOTE
[1] The paper (https://arxiv.org/abs/1706.03762) was the first public disclosure of the Transformer architecture.

question. Several potential impediments to LMMs' continued growth have been identified[2], two of which are discussed below.

The first is LMM hallucinations (factually incorrect, nonsensical or contextually incoherent responses). When today's Transformer-based LLMs analyze text and return textual responses, they must do so probabilistically. They consequently have a certain probability of making errors when either interpreting training data or returning responses. Researchers are working on various ways to prevent hallucinations but such efforts have so far only reduced hallucinations' frequency. Another cause of hallucinations is that LLMs are constructed based on information available at a certain point in time and therefore may not be able to accommodate data that postdate their construction. They may even be confounded by information that predated their construction but was not included in their training data. Given this context, it's understandable to harbor concerns that the progression of functionality in Transformer-based LLMs may eventually plateau.

The second potential impediment is scarcity of the resources needed to deploy LLMs, specifically electric power, computing capacity and training data. Scarcity of these physical resources poses a threat to the further advancement of LLMs. Many experts are deeply skeptical of the likelihood of today's LLMs directly paving the way to artificial general intelligence (AGI), the holy grail of AI[3].

New initiatives: RAG, SLMs

Researchers are of course exploring various potential solutions to Transformer-based LLMs' shortcomings. One technology that has proven successful at mitigating hallucinations to some degree is retrieval-augmented generation (RAG). RAG essentially provides AI models with answers to questions up front but instead of literally giving the model correct answers, which would defeat AI's purpose, a human points the model toward information needed to arrive at an answer. Such pointers are provided at the same time the AI model is given a prompt. RAG has demonstrated that it can improve the accuracy of LLMs' output without altering the LLM's internal processes.

Novel technologies have started to emerge to address Transformer models' resource consumption concerns also. A recent case in point is efforts to construct small language models (SLMs) that are trained on smaller datasets and powered by smaller-scale computing resources than LLMs but are still capable of

[2] See *AI roadmap's inherent unpredictability.*

[3] Even Sam Altman, CEO of OpenAI, the developer of ChatGPT, has expressed skepticism about the likelihood of AGI directly ensuing from today's Transformer models (*OpenAI CEO Sam Altman Discusses GPT-5, Sora, and the Road to AGI*).

generating highly accurate output.

Google released RecurrentGemma, its first SLM, on April 16, 2024. Eight days later, Microsoft followed suit, releasing an SLM named Phi-3[4]. Both Google and Microsoft claim their SLMs perform as well as seasoned LLMs such as GPT-4, Llama and Gemini despite being trained on vastly less data than the LLMs. Their SLMs reportedly also consume less energy and have lower computing resource requirements.

Handheld AI era in the offing?

Today's LLMs are generally accessed via a web app or API. The LLMs themselves are housed at large data centers. LLM users merely lease access to the LLM's functionality.

However, if SLM technologies can be locally installed on smartphones or regular PCs, eliminating the need for connection to a large-scale model while delivering promised performance, the landscape of generative AI usage could swiftly transform. SLMs installed in smartphones or PCs could usher in a new era of personal AI in a mobile or desktop form factor, loaded with the user's personal data (use of RAG technology) and optimized to the user's personal use cases.

Existing LLMs are already able to perform a wide variety of tasks and functions fairly accurately, including comprehending text, recognizing speech, composing text, recognizing images and automating tasks. SLMs more or less able to do likewise would be a game changer.

I always wished I had a simultaneous interpretation tool like the magic substance the Japanese cartoon character Doraemon would swallow to be able to speak any foreign language he wanted to. It looks like my wish may actually come true before too long.

[4] See *Tiny but mighty: The Phi-3 small language modes with big potential.*

about NRI

Founded in 1965, Nomura Research Institute (NRI) is a leading global provider of system solutions and consulting services with annual sales above $4.9 billion. NRI offers clients holistic support of all aspects of operations from back- to front-office, with NRI's research expertise and innovative solutions as well as understanding of operational challenges faced by financial services firms. The clients include broker-dealers, asset managers, banks and insurance providers. NRI has its offices globally including New York, London, Tokyo, Hong Kong and Singapore, and over 17,400 employees.

For more information, visit https://www.nri.com/en

The report is provided solely for informational purposes for our UK and USA readers and is not to be construed as providing advice, recommendations, endorsements, representations or warranties of any kind whatsoever.

Whilst every effort has been taken to ensure the accuracy of the information, NRI shall have no liability for any loss or damage arising directly or indirectly from the use of the information contained in this report.

Inquiries to : Financial Market & Digital Business Research Department
Nomura Research Institute, Ltd.
Otemachi Financial City Grand Cube,
1-9-2 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan
E-mail : kyara@nri.co.jp

https://www.nri.com/en/knowledge/publication/fis/lakyara/