Large-Scale Artificial Intelligence Models

Hsiao-Ying Lin, Huawei International

As artificial intelligence models rapidly grow in size, their potential is stimulated.

What can large-scale artificial intelligence (AI) models do? Large-scale AI language models can generate articles indistinguishable from ones written by humans and help software developers complete codes. Despite an individual being unfamiliar with ancient Chinese languages, that individual may be able to view AI model-generated images that depict topics of ancient Chinese poems from the Tang dynasty. Moreover, AI models can now generate series of pictures from a sequence of text descriptions to tell stories. Large-scale AI models present new opportunities for such undertakings. In this article, the growth of large-scale AI models, the potential they offer, and their requirements are summarized.

Digital Object Identifier 10.1109/MC.2022.3151419
Date of current version: 6 May 2022

PROLOGUE

Let us reflect on a key milestone in the prestigious history of deep-learning-based language models. Generative Pretrained Transformer 3 (GPT-3), developed by OpenAI in 2020, is a natural language processing (NLP) model. GPT-3 achieved a stunning breakthrough when it produced articles, generated summaries of entire books, autocompleted codes, and so on. Although GPT-3's predecessor, GPT-2, has only approximately 1.5 billion parameters and is trained on nearly 40 GB of text data, GPT-3 has roughly 175 billion parameters and is trained on nearly 45 TB of text data. GPT-3 is thus impressive not only for its capability but also its size. The considerable increase in model size from GPT-2 to GPT-3 reflects a trend toward large-scale AI models.

LARGE-SCALE UNIMODAL MODELS

Since GPT-3's release, creating larger models to achieve superior performance or handle more complex tasks has become a trend. The resulting models are often considered large-scale AI models, although there is no actual rigorous definition. For instance, DeepMind released an NLP model, Gopher, which has 280 billion parameters. Microsoft launched the Megatron-Turing Natural Language Generation model,

0018-9162/22©2022IEEE

EDITOR **HSIAO-YING LIN**
IEEE Member; hsiaoying.lin@gmail.com

which has 530 billion parameters. More large-scale language models are currently being developed, including the Switch Transformer, which results in an NLP model with 1.6 trillion parameters. A core component of the Switch Transformer's structure is the mixture of experts (MoE), which combines several expert networks in parallel and plays a central role in increasing the model's size. Relative to the complexity of the human brain which has about 100 trillion parameters, there seems to be still room for model size to grow.

Similarly, the sizes of Chinese NLP models are also growing rapidly. Although the Chinese language differs greatly from English, particularly when you consider its different word segmentations, Chinese NLP models also obtain superior performance via hyperscale model sizes. Wu Dao-Wen Yuan, which was developed by the Beijing Academy of Artificial Intelligence (BAAI) as a component of Wu Dao 1.0 (a suite of superscale Chinese-oriented models), is among the first Chinese NLP models and has 2.6 billion parameters. Other major Chinese companies have continued announcing their own hyperscale Chinese NLP models. In April 2021, Alibaba's Discovery, Adventure, Momentum and Outlook (DAMO) research team introduced a Chinese NLP model, Pretraining for Language Understanding and Generation (PLUG), which has 27 billion parameters. The model size of PLUG is 10 times more than that of Wu Dao-Wen Yuan. A few weeks later, another Chinese NLP model, Pangu-α, was launched with 200 billion parameters. These Chinese NLP models can handle language tasks such as summarizing text, answering questions, and generating dialogue. Moreover, generating poems and couplets in an ancient Chinese style is doable. In late 2021, Peng Cheng Laboratory and Baidu announced the Enhanced Representation Through Knowledge Integration (ERNIE) NLP model, which has 260 billion parameters. ERNIE has been deployed as a part of Baidu's cloud services and can process Chinese-language documents, such as financial contracts.

Although deep learning techniques were originally borrowed from computer vision to solve NLP problems, researchers have begun to apply NLP techniques to implement computer vision tasks. The trend of large-scale models in an NLP domain is propagated to the computer vision domain. The Vision Transformer (ViT)-H model, which utilizes the transformer architecture for computer vision tasks, initially had only 632 million parameters. An improved version, Scaling ViTs, increased the model size to 2 billion parameters. Researchers have also investigated translating the success of self-supervised learning in NLP to computer vision. The large-scale vision models Self-Supervised (SEER) and Image GPT, with 1.3 and 6 billion parameters, respectively, were the result. Following the ViT, vision architecture based on a sparse MoE (V-MoE) is a new type of architecture suitable for parallel processing. An instance model using the V-MoE architecture with 15 billion parameters achieved a 90.35% test accuracy on ImageNet, close to the present state of the art (the optimal top-one accuracy is 90.88% at the time of writing[1]).

Universal approximation theorems provide some clues as to why the size of AI models is crucial, suggesting that a neural network can approximate any continuous function. First, an arbitrary-width version of universal approximation theorems states that a neural network with at least one hidden layer can approximate any continuous function.[2] One of the initial versions of universal approximation theorems was proved for neural networks with sigmoid activation functions and later extended to include a broader class of activation functions. Second, an arbitrary-depth version of universal approximation theorems, where neural networks are with limited width, was successfully proved for convolutional neural networks.[3] These theorems assert that, as long as a model is sufficiently deep or wide, it can be trained to complete a designated task by approximating a continuous function. However, little is known about designing effective and efficient training algorithms to find a neural network for a designated task.

LARGE-SCALE MULTIMODAL MODELS

Combining natural language and image modalities is a new strategy used to improve the ability of AI models in a manner akin to how humans learn by reading and seeing. Models employing such a strategy are classified as multimodal machine learning models; they can use multiple modalities such as images, video, text, and audio. An example of a multimodal model is the Contrastive Language-Image Pretraining network, developed by OpenAI for the mapping of pairs of images and text in image classification tasks. Multimodal models aim at handling multimodal tasks such as image captioning and cross-modal searches. Multimodal machine learning seeks to map multimodal information into a unified representation space, approximating the designated task as a function in that space and mapping the result from that space back to the target modality. It is still an active area to find a proper representation space or coordinated ones due to the differences among multiple modalities such as data distributions and noise levels. In addition to a supervised learning approach, an unsupervised learning approach plays a key role for multimodal tasks. For example, generative models, which are a kind of unsupervised learning techniques and capture the joint probability distribution of observed and target variables,

can approximate the joint probability distribution of texts and images for generating images from texts.[4,5] Because of developments in multimodal machine learning, some recent work has started to theoretically discuss when and why multimodal learning can outperform unimodal learning.[6]

Expanding hyperscale model sizes offers an exciting opportunity for multimodal models to grow. Figure 1 presents an example of images generated from an English text description. Soon after the introduction of GPT-3, DALL·E was introduced, which is a GPT-3-based multimodal model used for generating images from a given text caption and named after a portmanteau of the artist Salvador Dalí and Pixar's WALL·E. DALL·E uses a GPT-3 version with 12 billion parameters and is capable of multimodal tasks, like image-to-image translation, fashion, and interior design. Alibaba DAMO announced its multimodal model Multi-Modality to Multi-Modality Multitask Megatransformer (M6), which has a size of 100 billion parameters. M6 has been used for providing multiple services, such as cross-modal searches and AI-assisted fashion design. BAAI introduced Wu Dao 2.0 (with its 1.75 trillion parameters)—the successor to Wu Dao 1.0—as a multimodal model that can handle both English and Chinese text and images (namely, bilingual and bimodal models). Wu Dao 2.0 can generate poems written in ancient Chinese styles, answer questions, write essays, and provide captions for images. Going beyond the mere combination of text and image modalities, Zi Dong Tai Chu (ZDTC) is the first model combining the three modalities of natural language, images, and audio. The ZDTC model was launched by the Chinese Academy of Sciences (CAS). With 100 billion parameters, ZDTC can perform multimodal tasks such as generating images from audio and producing audio from images.

A Table That Has a Train Model On It With Other Cars and Things

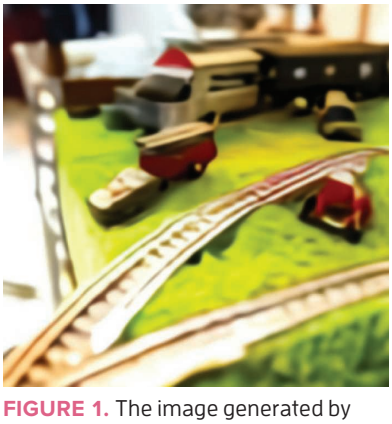

FIGURE 1. The image generated by DALL·E from a text description, adapted from Ramesh et al.[5]

Table 1 summarizes some recent large-scale models focused on such topics as NLP, Chinese NLP, vision, and multimodal models. Due to their hyperscale model sizes, these models are most likely deployed in cloud environments and provided as cloud services.

LARGE-SCALE TRAINING IS A CORNERSTONE OF LARGE-SCALE AI MODELS

Senior Fellow and Senior Vice President of Google Research Jeff Dean predicted that larger and more capable machine learning models are one of five impactful trends[7] that have already been demonstrated by large-scale models. Another expected trend is improvements in the efficiency of large-scale machine learning. As training large-scale AI models is extremely compute intensive, researchers and engineers from various domains are collaborating to innovate efficiency enhancements. The following are some proposed strategies:

- *Hardware acceleration*: In addition to CPUs and GPUs, dedicated hardware accelerators such as tensor and neural network processing units for AI training and inference at the chip level, have been designed and yielded enhanced efficiency.
- *Large-scale parallelism*: Large-scale parallelism includes data, model, and pipeline types. Heterogeneous computing nodes are integrated and coordinated such that all nodes seamlessly collaborate on a given training task, reducing intranode and internode communication delays and properly dispatching and merging computations in distinct model architectures.
- *Integration of designs*: Employing innovative parallel designs in machine learning frameworks is crucial to facilitating large-scale model training with specialized distributed computing platforms. For example, the Switch Transformer was trained using a specific software stack called *Mesh Tensorflow*, a distributed deep learning framework, and Wu Dao 2.0 was trained using FastMoEs,[8] a distributed training system implemented based on PyTorch (a common deep learning framework).
- *Efficient model architecture*: Training-efficient model architectures can be designed to increase model size for more complex or generalized tasks. For example, the Switch Transformer was designed using the MoEs architecture, whose computational overhead remains nearly unchanged, even as the number of parameters increases.[9]

LARGE-SCALE MODELS AS PRETRAINED MODELS

An aim of training large-scale models is obtaining a generalized model that can complete multiple similar tasks. By doing so, developers hope to move closer to developing artificial general intelligence, which can learn and solve problems like humans do. Additionally, because the cost of training a large-scale model is considerable, supporting multiple similar tasks substantially reduces overhead. The concept of employing models for multiple similar tasks is alluded to by related terms, such as pretrained models, foundation models, and transfer learning. Figure 2 abstractly visualizes this concept.

TABLE 1. A summary of selected large-scale models for NLP, Chinese NLP, vision, and multimodal models.

NLP				Chinese NLP			
Creator	Model Name	Size in B	Time	Creator	Model Name	Size in B	Time
Open AI	GPT3	175	2020.05	BAAI	Wu Dao-Wenyuan	2.6	2021.03
Microsoft	Megatron-Turing NLG	530	2021.01	Alibaba DAMO	PLUG	27	2021.04
Google Brain	Switch Transformers	1600	2021.01	PCL-Mindspore	Pangu	200	2021.04
DeepMind	Gopher	280	2021.12	PCL-Baidu	ERNIE	260	2021.12
Vision				**Multimodal**			
Creator	Model Name	Size in B	Time	Creator	Model Name	Size in B	Time
Open AI	ImageGPT	6	2020.06	Open AI	DALL·E	12	2020.05
Google Brain	ViT-H	0.632	2020.09	BAAI	Wu Dao 2.0	1750	2021.05
Facebook	SEER	1.3	2021.03	Alibaba DAMO	M6	100	2021.05
Google Brain	ViT-G/14	2	2021.06	CAS	ZDTC	100	2021.09
Google Brain	V-MoEs	15	2022.01	Baidu	ERNIE-ViLG	10	2021.12

PCL: Peng Cheng Laboratory; ViT: Vision Transformer; CAS: the Chinese Academy of Sciences; NLG: natural language generation; ViLG: vision-language generation.

A model is first trained for a general task using an initial training data set through a process called *pretraining*. The resultant model is later retrained for another specific task by using another task-specific training data set, which is usually much smaller than the original training data set. These specific tasks are called *downstream tasks*, and the retraining process is referred to as *fine-tuning*. The original model's knowledge is then utilized for a more specific task. Therefore, this process is called *transfer learning*. It is inspired by the idea that previously learned knowledge can be employed to solve new problems. Thus, the original model is called a *foundation* or *pretrained model*. Adopting pretrained models as core components (rather than developing wholly new models) for downstream tasks is the norm.

The aforementioned large-scale models can function as foundation models. For example, GPT-3 not only generates natural language and produces written articles but also performs downstream tasks, such as generating programming code and inferring structured data.

Hyperscale size is effective for improving model performance in unimodal and multimodal tasks; it is also a step toward developing artificial general intelligence. Large-scale models present extremely high computational requirements. Advanced hardware–software co-design techniques are continually developed to power next-generation, large-scale models. The trend of increasing AI model size does not appear to be ceasing. Nevertheless, only a few major companies and resourceful institutes can keep pace with this trend

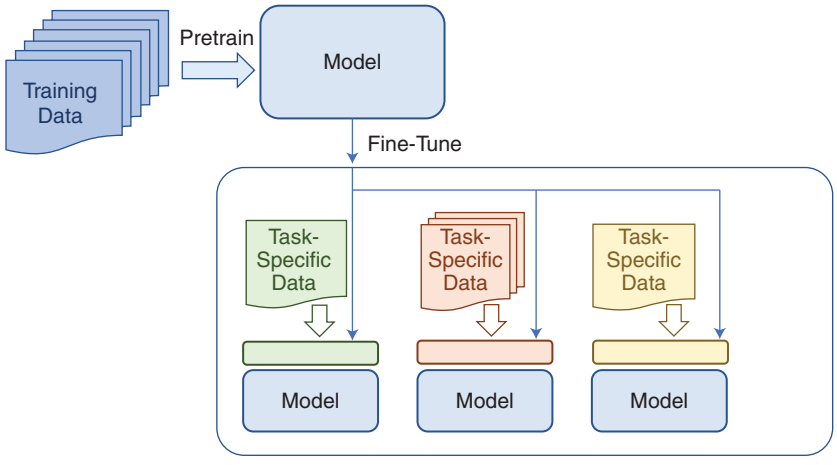

FIGURE 2. A high-level representation of pretraining and fine-tuning processes.

because the barriers to entry are considerable. Therefore, alternative methods for improving model performance are desirable. Some research has explored techniques other than increasing model size for achieving superior performance in specific and generalized tasks. Increasing model size is not the ultimate goal; it is only one of various means for improving model performance.

REFERENCES

1. "Image classification on ImageNet: Leaderboard," paperswithcode.com. https://paperswithcode.com/sota/image-classification-on-imagenet (Accessed: Jan. 26, 2022).
2. G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics Control, Signals, Syst.*, vol. 2, no. 4, pp. 303–314, 1989, doi: 10.1007/BF02551274.
3. Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6231–6239.
4. H. Zhang *et al.*, "ERNIE-ViLG: Unified generative pre-training for bidirectional vision-language generation," 2021, *arXiv:2112.15283*.
5. A. Ramesh *et al.*, "Zero-shot text-to-image generation," 2021, *arXiv: 2102.12092*.
6. Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multimodal learning better than single (Provably)," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2021, arXiv: 106.04538.
7. J. Dean, "Google Research: Themes from 2021 and beyond," Google AI Blog, 2022. https://ai.googleblog.com/2022/01/google-research-themes-from-2021-and.html (accessed Jan. 20, 2022).
8. J. He, J. Qiu, A. Zeng, Z. Yang, J. Zhai, and J. Tang, "FastMoE: A fast mixture-of-expert training system," 2021, arXiv: 2103.13262.
9. W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," 2021, *arXiv:2101.03961*.

DISCLAIMER

This article contains the views of the author. The opinions expressed here are hers alone.

HSIAO-YING LIN is a principal researcher at Huawei International, 138588, Singapore, and a Member of IEEE. Contact her at hsiaoying.lin@gmail.com.

Digital Object Identifier 10.1109/MC.2022.3166645