Beyond Words: Large Language Models Expand AI’s Horizon

The powerful models making waves in natural language processing are rippling across fields from healthcare to robotics and beyond.

Author: Rick Merritt

Back in 2018, BERT got people talking about how machine learning models were learning to read and speak. Today, large language models , or LLMs, are growing up fast, showing dexterity in all sorts of applications.

They’re, for one, speeding drug discovery, thanks to research from the Rostlab at Technical University of Munich, as well as work by a team from Harvard, Yale and New York University and others . In separate efforts, they applied LLMs to interpret the strings of amino acids that make up proteins, advancing our understanding of these building blocks of biology.

It’s one of many inroads LLMs are making in healthcare, robotics and other fields.

Transformer models — neural networks, defined in 2017, that can learn context in sequential data — got LLMs started.

Researchers behind BERT and other transformer models made 2018 “a watershed moment” for natural language processing, a report on AI said at the end of that year. “Quite a few experts have claimed that the release of BERT marks a new era in NLP,” it added.

Developed by Google, BERT (aka Bidirectional Encoder Representations from Transformers) delivered state-of-the-art scores on benchmarks for NLP. In 2019, it announced BERT powers the company’s search engine.

Google released BERT as open-source software , spawning a family of follow-ons and setting off a race to build ever larger, more powerful LLMs.

For instance, Meta created an enhanced version called RoBERTa , released as open-source code in July 2017. For training, it used “an order of magnitude more data than BERT,” the paper said, and leapt ahead on NLP leaderboards. A scrum followed.

For convenience, score is often kept by the number of an LLM’s parameters or weights, measures of the strength of a connection between two nodes in a neural network. BERT had 110 million, RoBERTa had 123 million, then BERT-Large weighed in at 354 million, setting a new record, but not for long.

In 2020, researchers at OpenAI and Johns Hopkins University announced GPT-3 , with a whopping 175 billion parameters, trained on a dataset with nearly a trillion words. It scored well on a slew of language tasks and even ciphered three-digit arithmetic.

“Language models have a wide range of beneficial applications for society,” the researchers wrote.

Within weeks, people were using GPT-3 to create poems, programs, songs, websites and more. Recently, GPT-3 even wrote an academic paper about itself .

“I just remember being kind of blown away by the things that it could do, for being just a language model,” said Percy Liang, a Stanford associate professor of computer science, speaking in a podcast .

GPT-3 helped motivate Stanford to create a center Liang now leads, exploring the implications of what it calls foundational models that can handle a wide variety of tasks well.

Last year, NVIDIA announced the Megatron 530B LLM that can be trained for new domains and languages. It debuted with tools and services for training language models with trillions of parameters.

“Large language models have proven to be flexible and capable … able to answer deep domain questions without specialized training or supervision,” Bryan Catanzaro, vice president of applied deep learning research at NVIDIA, said at that time.

Making it even easier for users to adopt the powerful models, the NVIDIA Nemo LLM service debuted in September at GTC. It’s an NVIDIA-managed cloud service to adapt pretrained LLMs to perform specific tasks.

The advances LLMs are making with proteins and chemical structures are also being applied to DNA.

Researchers aim to scale their work with NVIDIA BioNeMo , a software framework and cloud service to generate, predict and understand biomolecular data. Part of the NVIDIA Clara Discovery collection of frameworks, applications and AI models for drug discovery, it supports work in widely used protein, DNA and chemistry data formats.

NVIDIA BioNeMo features multiple pretrained AI models , including the MegaMolBART model, developed by NVIDIA and AstraZeneca.

Transformers are also reshaping computer vision as powerful LLMs replace traditional convolutional AI models. For example, researchers at Meta AI and Dartmouth designed TimeSformer , an AI model that uses transformers to analyze video with state-of-the-art results.

Experts predict such models could spawn all sorts of new applications in computational photography, education and interactive experiences for mobile users.

In related work earlier this year, two companies released powerful AI models to generate images from text.

OpenAI announced DALL-E 2 , a transformer model with 3.5 billion parameters designed to create realistic images from text descriptions. And recently, Stability AI, based in London, launched Stability Diffusion ,

LLMs also help developers write software. Tabnine — a member of NVIDIA Inception , a program that nurtures cutting-edge startups — claims it’s automating up to 30% of the code generated by a million developers.

Taking the next step, researchers are using transformer-based models to teach robots used in manufacturing, construction, autonomous driving and personal assistants.

For example, DeepMind developed Gato , an LLM that taught a robotic arm how to stack blocks. The 1.2-billion parameter model was trained on more than 600 distinct tasks so it could be useful in a variety of modes and environments, whether playing games or animating chatbots.

“By scaling up and iterating on this same basic approach, we can build a useful general-purpose agent,” researchers said in a paper posted in May.

It’s another example of what the Stanford center in a July paper called a paradigm shift in AI. “Foundation models have only just begun to transform the way AI systems are built and deployed in the world,” it said.

Learn how companies around the world are implementing LLMs with NVIDIA Triton for many use cases.

Original URL: https://blogs.nvidia.com/blog/2022/10/10/llms-ai-horizon/