What Is a Pretrained AI Model?

A pretrained AI model is a deep learning model that's trained on large datasets to accomplish a specific task, and it can be used as is or customized to suit application requirements across multiple industries.

Author: Angie Lee

Imagine trying to teach a toddler what a unicorn is. A good place to start might be by showing the child images of the creature and describing its unique features.

Now imagine trying to teach an artificially intelligent machine what a unicorn is. Where would one even begin?

Pretrained AI models offer a solution.

A pretrained AI model is a deep learning model — an expression of a brain-like neural algorithm that finds patterns or makes predictions based on data — that's trained on large datasets to accomplish a specific task. It can be used as is or further fine-tuned to fit an application's specific needs.

Instead of building an AI model from scratch, developers can use pretrained models and customize them to meet their requirements.

To build an AI application, developers first need an AI model that can accomplish a particular task, whether that's identifying a mythical horse, detecting a safety hazard for an autonomous vehicle or diagnosing a cancer based on medical imaging. That model needs a lot of representative data to learn from.

This learning process entails going through several layers of incoming data and emphasizing goals-relevant characteristics at each layer.

To create a model that can recognize a unicorn, for example, one might first feed it images of unicorns, horses, cats, tigers and other animals. This is the incoming data.

Then, layers of representative data traits are constructed, beginning with the simple — like lines and colors — and advancing to complex structural features. These characteristics are assigned varying degrees of relevance by calculating probabilities.

As opposed to a cat or tiger, for example, the more like a horse a creature appears, the greater the likelihood that it is a unicorn. Such probabilistic values are stored at each neural network layer in the AI model, and as layers are added, its understanding of the representation improves.

To create such a model from scratch, developers require enormous datasets, often with billions of rows of data. These can be pricey and challenging to obtain, but compromising on data can lead to poor performance of the model.

Precomputed probabilistic representations — known as weights — save time, money and effort. A pretrained model is already built and trained with these weights.

Using a high-quality pretrained model with a large number of accurate representative weights leads to higher chances of success for AI deployment. Weights can be modified, and more data can be added to the model to further customize or fine-tune it.

Developers building on pretrained models can create AI applications faster, without having to worry about handling mountains of input data or computing probabilities for dense layers.

In other words, using a pretrained AI model is like getting a dress or a shirt and then tailoring it to fit your needs, rather than starting with fabric, thread and needle.

Pretrained AI models are often used for transfer learning and can be based on several model architecture types. One popular architecture type is the transformer model , a neural network that learns context and meaning by tracking relationships in sequential data.

According to Alfredo Ramos, senior vice president of platform at AI company Clarifai — a Premier partner in the NVIDIA Inception program for startups — pretrained models can cut AI application development time by up to a year and lead to cost savings of hundreds of thousands of dollars.

Since pretrained models simplify and quicken AI development, many developers and companies use them to accelerate various AI use cases.

Top areas in which pretrained models are advancing AI include:

Natural language processing. Pretrained models are used for translation , chatbots and other natural language processing applications. Large language models , often based on the transformer model architecture, are an extension of pretrained models. One example of a pretrained LLM is NVIDIA NeMo Megatron , one of the world's largest AI models.

Speech AI. Pretrained models can help speech AI applications plug and play across different languages. Use cases include call center automation , AI assistants and voice-recognition technologies .

Computer vision. Like in the unicorn example above, pretrained models can help AI quickly recognize creatures — or objects, places and people. In this way, pretrained models accelerate computer vision , giving applications human-like vision capabilities across sports , smart cities and more.

Healthcare. For healthcare applications, pretrained AI models like MegaMolBART — part of the NVIDIA BioNeMo service and framework — can understand the language of chemistry and learn the relationships between atoms in real-world molecules, giving the scientific community a powerful tool for faster drug discovery .

Cybersecurity. Pretrained models provide a starting point to implement AI-based cybersecurity solutions and extend the capabilities of human security analysts to detect threats faster. Examples include digital fingerprinting of humans and machines, and detection of anomalies, sensitive information and phishing .

Art and creative workflows. Bolstering the recent wave of AI art , pretrained models can help accelerate creative workflows through tools like GauGAN and NVIDIA Canvas .

Pretrained AI models can be applied across industries beyond these, as their customization and fine-tuning can lead to infinite possibilities for use cases.

Companies like Google, Meta, Microsoft and NVIDIA are inventing cutting-edge model architectures and frameworks to build AI models.

These are sometimes released on model hubs or as open source, enabling developers to fine-tune pretrained AI models, improve their accuracy and expand model repositories.

NVIDIA NGC — a hub for GPU-optimized AI software, models and Jupyter Notebook examples — includes pretrained models as well as AI benchmarks and training recipes optimized for use with the NVIDIA AI platform.

NVIDIA AI Enterprise , a fully managed, secure, cloud-native suite of AI and data analytics software, includes pretrained models without encryption. This allows developers and enterprises looking to integrate NVIDIA pretrained models into their custom AI applications to view model weights and biases, improve explainability and debug easily.

Thousands of open-source models are also available on hubs like GitHub , Hugging Face and others.

It’s important that pretrained models are trained using ethical data that’s transparent and explainable, privacy compliant, and obtained with consent and without bias .

To help more developers move AI from prototype to production, NVIDIA offers several pretrained models that can be deployed out of the box, including:

NVIDIA SegFormer , a transformer model for simple, efficient, powerful semantic segmentation — available on GitHub.

NVIDIA’s purpose-built computer vision models , trained on millions of images for smart cities, parking management and other applications.

NVIDIA NeMo Megatron , the world’s largest customizable language model, as part of NVIDIA NeMo , an open-source framework for building high-performance and flexible applications for conversational AI , speech AI and biology.

NVIDIA StyleGAN , a style-based generator architecture for generative adversarial networks, or GANs. It uses transfer learning to generate infinite paintings in a variety of styles.

In addition, NVIDIA Riva , a GPU-accelerated software development kit for building and deploying speech AI applications, includes pretrained models in ten languages.

And MONAI , an open-source AI framework for healthcare research developed by NVIDIA and King’s College London, includes pretrained models for medical imaging.

Learn more about NVIDIA pretrained AI models .

Original URL: https://blogs.nvidia.com/blog/2022/12/08/what-is-a-pretrained-ai-model/