PROWESS

White Paper

Introduction to Hugging Face

A crucial open-source community member and a valuable AI IP meet in one company.

Although it might sound like something that erupted from an *Aliens* movie, Hugging Face is actually a central hub in the open-source AI ecosystem. The company started its journey to AI powerhouse as a chatbot startup marketing an AI product aimed at teenagers.[1] The company subsequently made the proprietary natural language processing (NLP) model open source and pivoted as an organization to being an AI and machine learning (ML) platform.

Hugging Face's niche in AI focuses on transformer-based models. Google introduced these models in 2017, and they represented a significant leap forward for NLP. Unlike previous NLP models that processed words sequentially, transformers enable NLP models to process words in relation to all other words in a sentence. This revolutionary innovation enables a deeper understanding of context and nuances in language, which in turn can lead to more accurate and sophisticated language models.

Hugging Face's transformers library is unique in its extensive collection of pre-trained models. The library supports a multitude of transformer models like BERT, GPT-2, T5, and RoBERTa. These models cater to a wide range of languages and NLP tasks, including text classification, information extraction, and question answering.

In 2021, Hugging Face branched out into image-based models, introducing onto its platform models for image classification, segmentation, and generation. This move proved to be almost prescient with the 2022 release of Stable Diffusion®, a popular open-source, text-to-image AI model from Stability AI. Beyond Stable Diffusion, Hugging Face houses image-generation models such as VQ-VAE-2, OpenAI's GLIDE, and models for image-to-image translation and image upscaling. As with its transformers library, Hugging Face has also become a major repository for image-based open-source models.

Mainstay of the Open-Source AI Community

Hugging Face has become a central part of the open-source AI community. It has done so through fostering a community that plays a prominent role in model development and sharing. The Hugging Face® platform encourages community members to contribute to and share models, datasets, and training pipelines. This collaborative ecosystem encompasses a diverse range of developers, researchers, and data scientists, which greatly enriches the library of available models and helps ensure a wide variety of tools for various AI tasks. For example, community contributions have expanded the range of language models beyond English to cover several other languages, including Chinese, Spanish, and French. This inclusive approach promotes innovation and democratizes AI research, making advanced models accessible to a broader audience.

The open-source philosophy embraced by Hugging Face significantly improves AI research and development. Hugging Face accelerates innovation in AI by providing open access to state-of-the-art models such as BERT, GPT, and others. This access enables researchers and developers to build upon existing work, fostering a collaborative environment that keeps the field moving forward. Open-source AI tools like those housed on the Hugging Face platform lower the barrier to entry, which allows smaller organizations and independent researchers to contribute to and benefit from cutting-edge AI advancements.

Another core aspect of Hugging Face's commitment to the open-source community is Hugging Face Hub. The Hub serves as a central point for community experimentation and collaboration. Hugging Face Hub provides free tools and datasets for ML development, a Git-based code repository, and Hugging Face Spaces, which are interactive web spaces in which developers and data scientists can build, share, and collaborate on ML applications and demos directly from their repositories.

Partnerships

Hugging Face's relevance in AI extends well beyond the collaborative Hugging Face Hub, however. Hugging Face is involved in a number of partnerships and collaborations with some of the biggest commercial names in the AI space.

AWS

Hugging Face's collaboration with Amazon Web Services (AWS) provides customers with an easier way to train and deploy Hugging Face models on AWS. This collaboration integrates Hugging Face's deep learning (DL) models into the Amazon SageMaker® ML service, simplifying the deployment of AI applications.[2]

Dell Technologies

Dell Technologies and Hugging Face are collaborating to make it easy for enterprises to create, fine-tune, and implement their own open-source generative AI models with the Hugging Face community on industry-leading Dell™ infrastructure–based products and services through a new Dell Technologies portal on the Hugging Face platform to offer simplified on-premises deployment of customized large language models (LLMs) on Dell servers.[3]

IBM

Hugging Face and IBM have collaborated on watsonx.ai®, an enterprise studio for AI builders. This partnership integrates many of Hugging Face's open-source libraries, such as transformers, into the IBM® watsonx® AI and data platform. IBM has also developed its own collection of LLMs, and it plans to make them open source and place them on the Hugging Face Hub.[4]

Intel

Intel and Hugging Face have collaborated to build Optimum Intel, an interface between the Hugging Face Transformers and Diffusers libraries and the different tools and libraries created by Intel to accelerate end-to-end pipelines on Intel® architectures.[5]

Microsoft

Hugging Face and Microsoft have teamed up to bring state-of-the-art NLP models to Microsoft Azure®, making it easier for developers to build applications that understand and interpret human language. This partnership integrates Hugging Face's models directly into Azure Machine Learning Studio and Azure Cognitive Services.[6]

NVIDIA

Hugging Face worked with NVIDIA to expand access to NVIDIA DGX® Cloud AI supercomputing within the Hugging Face platform to train and tune advanced models.[7]

Integration with Other Technologies

In addition to its formal collaboration with other organizations, Hugging Face provides strong compatibility for its platform with major ML frameworks, which facilitates deeper integration with other AI ecosystems. Hugging Face supports popular frameworks such as TensorFlow™ and PyTorch®, which makes it versatile as a development environment. This flexibility also allows for easy adoption in specific projects.

Business Model

Beyond being a key part of the open-source AI community, Hugging Face is also a for-profit company valued around $4.5 billion dollars, and it raised $235 million in its latest round of funding from companies that include Google, Amazon, NVIDIA, Salesforce, AMD, Intel, IBM, and Qualcomm.[8] While the free Hugging Face Hub is a significant focus of Hugging Face, the company also provides a number of paid services (detailed in Table 1).[9]

Table 1 | Paid Hugging Face services

Service	Description
Pro Account	Early access to new features and access to an API for more powerful models, such as Zephyr 7B β, Llama 2 Chat, and Stable Diffusion® XL.
Enterprise Hub	Secure and private collaboration for proprietary models and data along with enterprise tools such as AutoTrain® to automate model training; inferencing can also be done on subscribers' own hardware or handled by Hugging Face via the inference API.
Spaces Hardware	Cloud-based instances billed by the hour with an emphasis on GPUs that build on Hugging Face Spaces.
Inference Endpoints	Dedicated, autoscaling infrastructure for model deployment directly from Hugging Face Hub.

Prospects for the Future

Hugging Face distinguishes itself with its importance to the open-source AI community and its extensive repository of pre-trained models—more than 230,000 models with more than 1,000,000 repositories and 10,000 paying customers.[10] Moreover, Hugging Face also possesses significant expertise with NLP models, as illustrated by its numerous industry partnerships. And while platforms like TensorFlow and PyTorch provide broader ML capabilities, Hugging Face specializes in making NLP models like BERT and GPT easily accessible.

Hugging Face is well aligned with current AI trends like LLMs, ethical AI, and the democratization of AI technologies. Its open-source approach positions it to significantly contribute to and benefit from these trends. The company's recent entry into areas like image generation with Stable Diffusion indicates its readiness to adapt to evolving AI domains. Hugging Face appears to be well positioned to further benefit as AI continues to advance towards more sophisticated, ethical, and accessible models while at the same time nurturing the community that will perpetuate these same trends.

[1] Romain Dillet, TechCruch. "Hugging Face wants to become your artificial BFF." March 2017.
[2] AWS. "AWS and Hugging Face collaborate to make generative AI more accessible and cost efficient." February 2023.
[3] Dell Technologies. "Dell and Hugging Face Simplify On-Premises GenAI Deployment." November 2023.
[4] Hugging Face. "Hugging Face and IBM partner on watsonx.ai, the next-generation enterprise studio for AI builders." May 2023.
[5] Hugging Face. "Scaling Transformer Model Performance with Intel AI." Accessed January 2024. See also: https://github.com/huggingface/optimum-intel.
[6] Hugging Face. "Hugging Face Collaborates with Microsoft to launch Hugging Face Model Catalog on Azure." May 2023.
[7] Brian Caulfield, NVIDIA. "SIGGRAPH Special Address: NVIDIA CEO Brings Generative AI to LA Show." August 2023.
[8] Kif Leswing, CNBC. "Google, Amazon, Nvidia and other tech giants invest in AI startup Hugging Face, sending its valuation to $4.5 billion." August 2023.
[9] Hugging Face. Hugging Face Pricing webpage. Accessed January 2024.
[10] Kyle Wiggers, TechCrunch. "Hugging Face raises $235M from investors, including Salesforce and Nvidia." August 2023.