Title: Hugging Face

URL: https://en.wikipedia.org/wiki/Hugging_Face

PageID: 71431971

Categories: Category:2016 establishments in New York City, Category:American companies established in 2016, Category:Machine learning, Category:Open-source artificial intelligence, Category:Privately held companies based in New York City

-----

Hugging Face, Inc. is an American company based in New York City that develops computation tools for building applications using machine learning . It is most notable for its transformers library built for natural language processing applications and its platform that allows users to share machine learning models and datasets and showcase their work.

History

The company was founded in 2016 by French entrepreneurs Clément Delangue, Julien Chaumond, and Thomas Wolf in New York City , originally as a company that developed a chatbot app targeted at teenagers. [2] The company was named after the U+1F917 ■ HUGGING FACE emoji . [2] After open sourcing the model behind the chatbot, the company pivoted to focus on being a platform for machine learning.

In March 2021, Hugging Face raised US$40 million in a Series B funding round. [3]

On April 28, 2021, the company launched the BigScience Research Workshop in collaboration with several other research groups to release an open large language model . [4] In 2022, the workshop concluded with the announcement of BLOOM , a multilingual large language model with 176 billion parameters. [5] [6]

In December 2022, the company acquired Gradio, an open source library built for developing machine learning applications in Python. [7]

On May 5, 2022, the company announced its Series C funding round led by Coatue and Sequoia . [8] The company received a $2 billion valuation.

On August 3, 2022, the company announced the Private Hub, an enterprise version of its public Hugging Face Hub that supports SaaS or on-premises deployment. [9]

In February 2023, the company announced partnership with Amazon Web Services (AWS) which would allow Hugging Face's products to be available to AWS customers to use them as the building blocks for their custom applications. The company also said the next generation of BLOOM will be run on Trainium, a proprietary machine learning chip created by AWS. [10] [11] [12]

In August 2023, the company announced that it raised $235 million in a Series D funding round, at a $4.5 billion valuation. The funding was led by Salesforce and notable participation came from Google , Amazon , Nvidia , AMD , Intel , IBM , and Qualcomm . [13]

In June 2024, the company announced, along with Meta and Scaleway , their launch of a new AI accelerator program for European startups. This initiative aims to help startups integrate open foundation models into their products, accelerating the EU AI ecosystem. The program, based at STATION F in Paris, will run from September 2024 to February 2025. Selected startups will receive mentoring, access to AI models and tools, and Scaleway’s computing power. [14]

On September 23, 2024, to further the International Decade of Indigenous Languages , Hugging Face teamed up with Meta and UNESCO to launch a new online language translator [15] built on Meta's No Language Left Behind open-source AI model, enabling free text translation across 200 languages, including many low-resource languages. [16]

On April 2025, Hugging Face announced that they acquired a humanoid robotics startup, Pollen Robotics. Pollen Robotics is a France based Robotics Startup founded by Matthieu Lapeyre and

Pierre Rouanet in 2016. [17] [18] In an X tweet, Clément Delangue, CEO of Hugging Face, shared his vision to make Artificial Intelligence robotics Open Source. [19]

Services and technologies

Transformers Library

The Transformers library is a Python package that contains open-source implementations of transformer models for text, image, and audio tasks. It is mainly compatible with the PyTorch library, but previous versions were also compatible with TensorFlow and JAX deep learning libraries. It includes implementations of notable models like BERT and GPT-2 . [20] The library was originally called "pytorch-pretrained-bert" [21] which was then renamed to "pytorch-transformers" and finally "transformers."

A JavaScript version (Transformers.js [22]) has also been developed, allowing models to run directly in the browser through the ONNX runtime.

Hugging Face Hub

The Hugging Face Hub is a platform (centralized web service) for hosting: [23]

Git -based code repositories , including discussions and pull requests for projects;

models, also with Git-based version control;

datasets, mainly in text, images, and audio;

web applications ("spaces" and "widgets"), intended for small-scale demos of machine learning applications.

There are numerous pre-trained models that support common tasks in different modalities, such as:

Natural Language Processing : text classification, named entity recognition, question answering, language modeling, summarization, translation, multiple choice, and text generation.

Computer Vision : image classification, object detection, and segmentation.

Audio: automatic speech recognition and audio classification.

Other libraries

In addition to Transformers and the Hugging Face Hub, the Hugging Face ecosystem contains libraries for other tasks, such as dataset processing ("Datasets"), model evaluation ("Evaluate"), image generation ("Diffusers"), and machine learning demos ("Gradio"). [24]

Safetensors

The safetensors format was developed around 2021 to solve problems with the pickle format in Python. It was designed for saving and loading tensors. Compared to the pickle format, it allows lazy loading and avoids security problems. [25] After a security audit, it became the default format in 2023. [26]

The file format:

size of the header: 8 bytes, an unsigned little-endian 64-bit integer.

header: JSON UTF-8 string, formatted as {"TENSOR_NAME": {“dtype”: “F16”, “shape”: [1, 16, 256], “data_offsets”: [BEGIN, END]}, "NEXT_TENSOR_NAME": {…}, …}.

file: a byte buffer containing the tensors.

See also

OpenAI

Station F

Kaggle

References

External links

Official website

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder

Vibe coding

Vision transformer

Waluigi effect

Word embedding

Character.ai

ChatGPT

DeepSeek

Ernie

Gemini

Grok

Copilot

Claude

Gemini

Gemma

GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5

1

2

3

J

4

4o

4.5

4.1

OSS

5

Llama

o1

o3

o4-mini

Qwen

Base44

Claude Code

Cursor

Devstral

GitHub Copilot

Kimi-Dev

Qwen3-Coder

Replit

Xcode

Aurora

Firefly

Flux

GPT Image 1

Ideogram

Imagen

Midjourney

Qwen-Image

Recraft

Seedream

Stable Diffusion

Dream Machine

Hailuo AI

Kling

Midjourney Video

Runway Gen

Seedance

Sora

Veo

Wan

15.ai

Eleven

MiniMax Speech 2.5

WaveNet

Eleven Music

Endel

Lyria

Riffusion

Suno AI

Udio

Agentforce

AutoGLM

AutoGPT

ChatGPT Agent

Devin AI

Manus

OpenAI Codex

Operator

Replit Agent

01.AI

Aleph Alpha

Anthropic

Baichuan

Canva

Cognition AI

Cohere

Contextual AI

DeepSeek

ElevenLabs

Google DeepMind

HeyGen

Hugging Face

Inflection AI

Krikey AI

Kuaishou

Luma Labs

Meta AI

MiniMax

Mistral AI

Moonshot AI

OpenAI

Perplexity AI

Runway

Safe Superintelligence

Salesforce

Scale AI

SoundHound

Stability AI

Synthesia

Thinking Machines Lab

Upstage

xAI

Z.ai

Category

v

t

e

History timeline

timeline

Companies

Projects

Parameter Hyperparameter

Hyperparameter

Loss functions

Regression Bias–variance tradeoff Double descent Overfitting

Bias–variance tradeoff

Double descent

Overfitting

Clustering

Gradient descent SGD Quasi-Newton method Conjugate gradient method

SGD

Quasi-Newton method

Conjugate gradient method

Backpropagation

Attention

Convolution

Normalization Batchnorm

Batchnorm

Activation Softmax Sigmoid Rectifier

Softmax

Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann

Claude Shannon

Shun'ichi Amari

Kunihiko Fukushima

Takeo Kanade

Marvin Minsky

John McCarthy

Nathaniel Rochester

Allen Newell

Cliff Shaw

Herbert A. Simon

Oliver Selfridge

Frank Rosenblatt

Bernard Widrow

Joseph Weizenbaum

Seymour Papert

Seppo Linnainmaa

Paul Werbos

Geoffrey Hinton

John Hopfield

Jürgen Schmidhuber

Yann LeCun

Yoshua Bengio

Lotfi A. Zadeh

Stephen Grossberg

Alex Graves

James Goodnight

Andrew Ng

Fei-Fei Li

Alex Krizhevsky

Ilya Sutskever

Oriol Vinyals

Quoc V. Le

Ian Goodfellow

Demis Hassabis

David Silver

Andrej Karpathy

Ashish Vaswani

Noam Shazeer

Aidan Gomez

John Schulman

Mustafa Suleyman

Jan Leike

Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category