how to use a model from hugging face on my local machine

All   Videos   Short videos   Images   Forums   Shopping   Web   More ▾   Tools ▾

✦ AI Overview   Learn more   ⋮

To use a model from Hugging Face on your local machine, you'll typically
download it using the Hugging Face Transformers library, then load it into
your Python code, and finally use it for inference or fine-tuning. ⊘

Here's a more detailed breakdown:

1. Prerequisites:

- **Install necessary libraries:** You'll need the `transformers` library, and potentially a
  framework like PyTorch or TensorFlow, depending on the model's architecture.

  - Install transformers: `pip install transformers`
  - Install PyTorch: `pip install torch`
  - Install TensorFlow: `pip install tensorflow`

- **Create a Hugging Face account:** This allows you to access the Hub and download
  models.

- **Get a user access token:** This token is used to authenticate your downloads. ⊘

2. Downloading the Model:

- **Find the model on the Hugging Face Hub:** Browse the [Hugging Face Hub](#) and
  find the model you want to use. ⊘

- Use the `from_pretrained()` function: In your Python code, use the `transformers`
  library's `AutoModel.from_pretrained()` function to download and load the
  model. ⊘

  - **Example:**

```python
from transformers import AutoModel

model = AutoModel.from_pretrained("google/flan-t5-large")  # Replace with the model ID
```

Specify the model ID:

The `from_pretrained()` function takes the model ID as an argument, which you can find
on the model's page on the Hugging Face Hub. ⊘

Optional: Specify a revision:

You can also specify a specific version (revision) of the model using the `revision`
parameter. ⊘

- **Example:**

```python
model = AutoModel.from_pretrained("google/flan-t5-large", revision="4d33b01d79672f27f001f6d
```

3. Using the Model:

- **Load the tokenizer:** You'll also need the tokenizer associated with the model to
  convert text into a format that the model can understand.

  - **Example:**

```python
```

```python
tokenizer = AutoTokenizer.from_pretrained("google/flan-t5-large")
```

- **Prepare your input:** Use the tokenizer to encode your input text.
- **Make predictions:** Use the model to make predictions on your input.
  - **Example:**

```python
Python                                                          ⧉
inputs = tokenizer("Hello, how are you?", return_tensors="pt")
outputs = model(**inputs)
```

- **Fine-tune the model (optional):** If you want to improve the model's performance on a specific task, you can fine-tune it on your own data.

4. Local Storage:

Hugging Face Cache:
Models downloaded from the Hugging Face Hub are stored in a cache directory on your local machine, which is typically located at `~/.cache/huggingface/hub`.

Load from local directory:
If you have a model downloaded to a specific directory, you can load it using the `local_files_only=True` parameter in `from_pretrained()`.

- **Example:**

```python
Python                                                          ⧉
model = AutoModel.from_pretrained("./my_model_directory", local_files_only=True)
```

5. Additional Tips:

- **Check the model's license:** Make sure the model's license allows your intended use.

Choosing and Implementing Hugging Face Models - Medium
1 Nov 2024 — Incorporating into your code. Once you've found a model you'd like to try, it's easy to get going- click the "Use this M...

Ⓜ Medium · Stephanie Kirmer   ⋮

Using hugging face models with private company data? - Beginners
26 Sept 2023 — TLDR: You can install the transformers library (🤗 Transformers) plus e.g. tensorflow or pytorch, download an appropria...

🤗 Hugging Face Forums   ⋮

How to Use a Hugging Face Model Without Internet Access? - Medium
2 Feb 2025 — To use Hugging Face models offline, you need to: * Download the model beforehand in an internet-enabled environment. *

Ⓜ Medium   ⋮

Show all

Generative AI is experimental.   👍   👎

Videos   ⋮

▶ Running a Hugging Face LLM on your laptop
4:35
YouTube · Learn Data with Mark
4 Aug 2023

DeepSeek AI   Ask ChatGPT

G ·········   ✕   🎤   📷   🔍   M

From 00:12	From 00:52	From 02:27	From 03:00
Downloading a Model	Choose a Model To Download	Initialize the Model	Create Our Pipeline

▶
14:55

How to Run HuggingFace Models Locally (Without Ollama ...

YouTube · Skilled Engg
6 Dec 2024

▭ 15 key moments in this video   ⌄

▶
12:10

Locally Run Huggingface LLMs like Llama on Your Laptop or ...

YouTube · Cameron McKenzie
28 Jan 2025

4 key moments in this video   ⌄

▶
22:59

HuggingFace + Langchain | Run 1,000s of FREE AI Models ...

YouTube · Tech With Tim
1 month ago

▭ 10 key moments in this video   ⌄

View all ⟩

🙂 Hugging Face Forums
https://discuss.huggingface.co › run-models-on-a-deskt...

Run models on a desktop computer? - Beginners

5 Feb 2023 — Yes, you can **run all models from the hub locally**. Maybe you can start by here:
Installation Setting up a local python environment, and installing the required ...
Top answer: hello @antcodes , Yes, you can run all models from the hub locally. Maybe you can start ...

People also ask   ⋮

How to import a model from Hugging Face?   ⌄

How to use Hugging Face models offline?   ⌄

How do you run a Hugging Face space locally?   ⌄

Can Hugging Face models be used commercially?   ⌄

                                                                    Feedback

🔴 Reddit · r/huggingface
9 comments · 1 year ago

How do I download and run a model? : r/huggingface

I highly recommend **using** this after downloading a **model** and **using** LM studio to **host** a **local** server to
connect **the model** with **the** web app.
9 answers · Top answer: I prefer in this order. 1. Ollama with Ollama Web UI (yes runs text to image), 2...
How do I **use HuggingFace**? : r/SillyTavernAI - Reddit   6 answers   16 May 2024
What's **the** simplest way to download **models** from ...   25 answers   8 Jun 2024
More results from www.reddit.com

🙂 Hugging Face Forums
https://discuss.huggingface.co › download-models-for-l...

Download models for local loading - Beginners

10 Nov 2020 — **how can I download the models from huggingface directly in my specified local
machine** directroy rather it downloads automatically into cached ...

G   ✕   🎤   📷   🔍   M

6 Ways to Run LLMs Locally (also how to use HuggingFace)

18 Jun 2024 — Hugging Face also provides transformers, a Python library that streamlines running a LLM locally. The following example uses the library to run ...

Stack Overflow
6 answers · 4 years ago

Load a pre-trained model from disk with Huggingface ...

Where is **the** file located relative to your **model** folder? I believe it has to be a relative PATH rather than an absolute one.

How to download a **model** from **huggingface** ...   16 answers   19 May 2021

python - how to **use huggingface models** without ...   1 answer   20 Oct 2023

More results from stackoverflow.com

YouTube · Sam Witteveen
120K+ views · 2 years ago

LangChain - Using Hugging Face Models locally (code ...

▶   This video shows you how to use the end points, how to load the models
            locally (and access model that don't work in the end points) and load ...
10:22

▢   9 key moments in this video   ⌄

🤗   Hugging Face
     https://huggingface.co › docs › hub › models-downloadi...

Downloading models   ⋮

Downloading **models**. Integrated libraries. If a **model** on **the** Hub is tied to a supported library, loading **the model** can be done in just a few lines.

Ⓜ   Medium · Stephanie Kirmer
     620+ likes · 5 months ago

Choosing and Implementing Hugging Face Models

Once you've found a model you'd like to try, it's easy to get going- **click the "Use this Model" button on the top right of the Model Card page**, ...

Images   ⋮

HuggingFace + Langchain | R...   Downloading models   LangChain - Using Hugging F...
▶ YouTube   ⋮   🤗 Hugging Face   ⋮   ▶ YouTube   ⋮

Show more images   ⌄

People also search for   ⋮

Hugging Face   Kaggle   GitHub   OpenAI   Groq   Civitai

Feedback

🔍 How to use a model from hugging   🔍 **Huggingface** local model
    face on my local machine **android**

G   ✕   🎤   📷   🔍   M

How to **run** Hugging Face **models**   How to use Hugging Face **models**
online   🔍   in Python   🔍

Huggingface download model   **Huggingface-cli download** model
locally   🔍   🔍

1 2 3 4 5 6 7 8 9 10   Next

Results are personalised - Try without personalisation

United Kingdom   **DD1, Dundee** - From your IP address - Update location

Help   Send feedback   Privacy   Terms