使用Hugging Face在本地端執行stable-diffusion-3.5–large

WZX

10 min read · Feb 18, 2025

Share ••• More

Hugging face

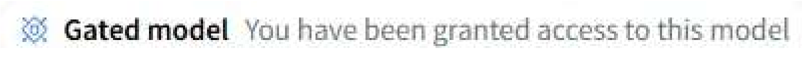
下載 stable-diffusion-3.5–large 模型到本地端

stabilityai/stable-diffusion-3.5-large · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science.

huggingface.co

此模型需要註冊才能擁有訪問權限並使用有權限的token來訪問

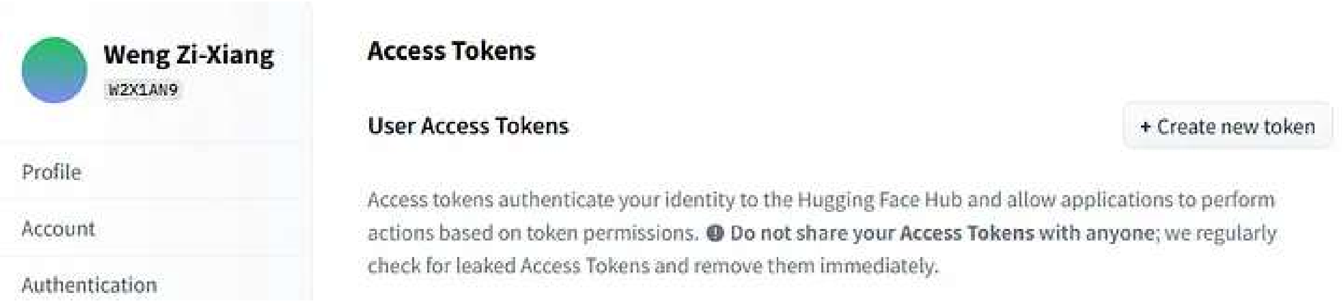


Gated model You have been granted access to this model

訪問權限

創建token

Create new token



Weng Zi-Xiang
W2X1AN9

Profile
Account
Authentication

Access Tokens

User Access Tokens

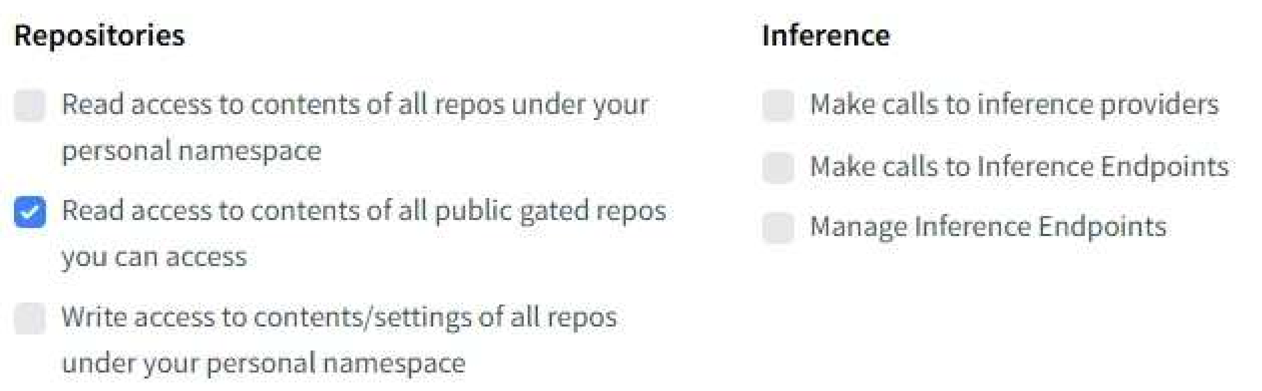
+ Create new token

Access tokens authenticate your identity to the Hugging Face Hub and allow applications to perform actions based on token permissions. ❗ **Do not share your Access Tokens with anyone**; we regularly check for leaked Access Tokens and remove them immediately.

settings

點擊token右邊三個點

Edit permissions將Read access to contents of all public gated repos you can access勾選才能存取



Repositories

☐ Read access to contents of all repos under your personal namespace

☑ Read access to contents of all public gated repos you can access

☐ Write access to contents/settings of all repos under your personal namespace

Inference

☐ Make calls to inference providers

☐ Make calls to Inference Endpoints

☐ Manage Inference Endpoints

執行

首先確保使用最新的 diffusers 庫

```
pip install -U diffusers
```

> *我在執行模型的過程中有遇到兩個警告*

建議安裝 `accelerate` 降低 CPU 記憶體的使用，並加速模型加載

```
pip install accelerate>=0.26.0
```

缺少 `protobuf` ，是 `transformers` 在處理某些分詞器相關功能時所需要

```
pip install protobuf
```

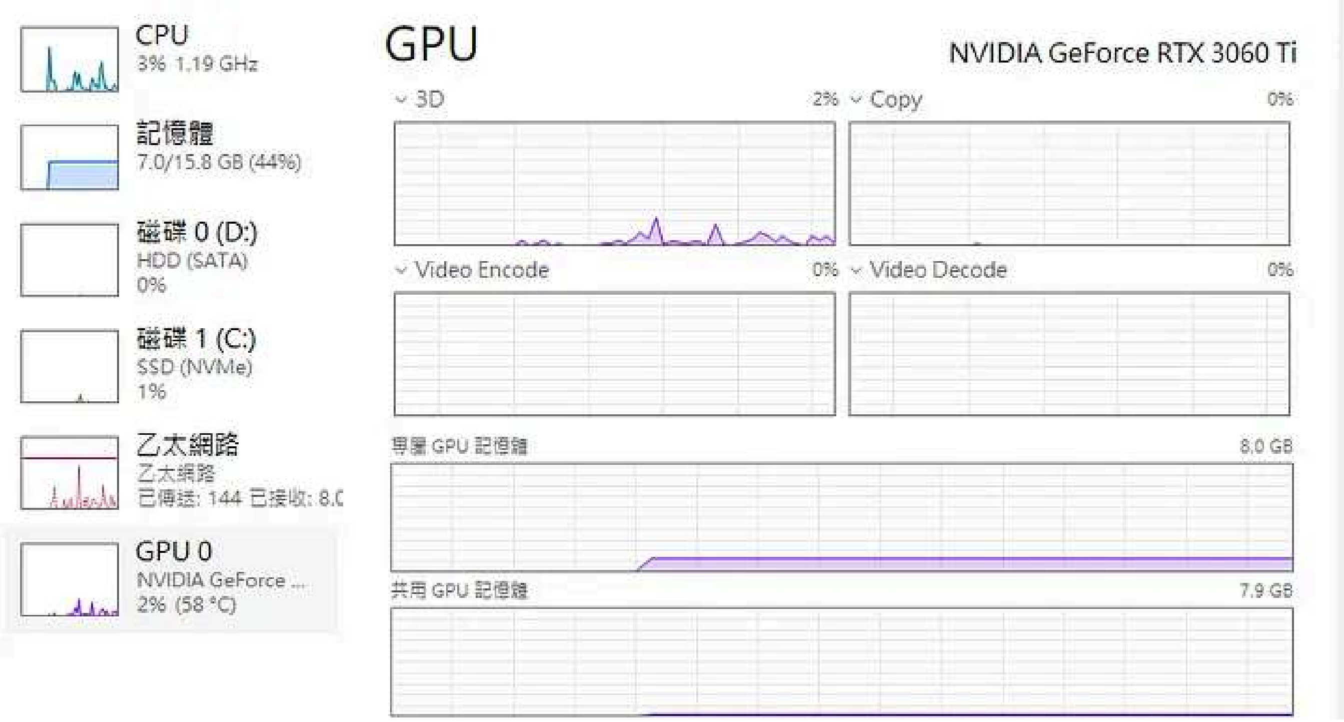
Hugging Face給的範例

```
from diffusers import BitsAndBytesConfig, SD3Transformer2DModel
from diffusers import StableDiffusion3Pipeline
import torch


model_id = "stabilityai/stable-diffusion-3.5-large" # 設定使用的模型
```

```
nf4_config = BitsAndBytesConfig(
    load_in_4bit=True,  # 設定載入4位元量化模型
    bnb_4bit_quant_type="nf4",  # 設定量化類型為 nf4
    bnb_4bit_compute_dtype=torch.bfloat16  # 設定計算使用的資料型態為 bfloat16
)

# 載入預訓練的模型，並使用先前設定的量化配置
model_nf4 = SD3Transformer2DModel.from_pretrained(
    model_id,
    subfolder="transformer",  # 指定子資料夾位置
    quantization_config=nf4_config,  # 使用量化配置
    torch_dtype=torch.bfloat16  # 設定模型的計算資料型態
)

# 載入管道，並使用先前載入的量化模型
pipeline = StableDiffusion3Pipeline.from_pretrained(
    model_id,
    transformer=model_nf4,  # 使用量化的Transformer模型
    torch_dtype=torch.bfloat16  # 設定管道的計算資料型態
)

# 啟用模型的CPU卸載功能，將模型的部分運算從GPU移到CPU上
pipeline.enable_model_cpu_offload()

# 設定提示字
prompt = "A whimsical and creative image depicting a hybrid creature that is a

image = pipeline(
    prompt=prompt,  # 設定提示字
    num_inference_steps=28,  # 設定推理步數
    guidance_scale=4.5,  # 設定指導強度
    max_sequence_length=512,  # 設定最大序列長度
).images[0]

# 儲存生成的圖片
image.save("whimsical.png")
```

由於此模型需要大量的運算資源導致電腦crash

我的硬體：



我的電腦可以執行的版本

```
from diffusers import BitsAndBytesConfig, SD3Transformer2DModel
from diffusers import StableDiffusion3Pipeline
import torch
from huggingface_hub import login

login(token="your_token")

model_id = "stabilityai/stable-diffusion-3.5-large"

nf4_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)
model_nf4 = SD3Transformer2DModel.from_pretrained(
    model_id,
    subfolder="transformer",
    quantization_config=nf4_config,
    torch_dtype=torch.bfloat16
)

pipeline = StableDiffusion3Pipeline.from_pretrained(
    model_id,
    transformer=model_nf4,
    torch_dtype=torch.bfloat16
)
pipeline.enable_model_cpu_offload()

prompt = "A whimsical and creative image depicting a hybrid creature that is a
```

```
image = pipeline(
    prompt=prompt,
    num_inference_steps=16,  # 減少推理步數28->16
    guidance_scale=4.5,
    max_sequence_length=512,
).images[0]

image.save("whimsical.png")
```

幾個降低運算資源的方法

- 降低生成影像的大小： `height = 512 #降低高度, width = 512 #降低寬度`
- 減少推理步數： `num_inference_steps`
- 使用較小的批次： `num_images_per_prompt`
- 清空 GPU 不再需要的記憶體緩存： `torch.cuda.empty_cache()`

Prompt

這是一幅異想天開且富有創意的圖像，描繪了華夫餅和河馬的混合生物，在以早餐為主題的景觀中沐浴在融化的黃油河中。它具有河馬獨特的、笨重的體形。然而，這種生物的身體並不像常見的灰色皮膚，而是像剛從烤盤上拿下來的金黃色酥脆華夫餅。皮具有我們熟悉的華夫餅網格圖案，每格都填充了閃閃發光的糖漿。這個環境將河馬的自然棲息地與早餐桌佈置元素結合在一起，一條溫熱的融化黃油河，背景中茂密的、像煎餅一樣的樹葉間隱約可見超大的餐具或盤子，高聳的胡椒研磨器代替了一棵樹。 當太陽在這個奇幻的世界升起時，它為整個場景投下了溫暖、柔和的光芒。這隻動物對自己的奶油河感到滿意，打了個哈欠。附近，一群鳥飛起來

Output

Hugging Face Pytorch Stable Diffusion Python

Edit profile

Written by WZX

5 Followers · 2 Following