

padeoe的小站

好奇宝宝

输入内容按回车搜索

- [首页](#)
- [关于我](#)

如何快速下载huggingface大模型

2023/09/27

Update: 推荐 **huggingface** 镜像站: <https://hf-mirror.com>

Update: 推荐官方的 **huggingface-cli** 命令行工具、以及本站开发的 **hfd**脚本。

本文已发表至知乎 <https://zhuanlan.zhihu.com/p/663712983>。

Stackoverflow 上有个AI开发入门的最常见问题 [How to download model from huggingface?](#), 回答五花八门, 可见下载 **huggingface** 模型的方法是十分多样的。

其实网络快、稳的话, 随便哪种方法都挺好, 然而结合国内的网络环境, **断点续传**、**多线程下载**等特性还是非常必要的, 否则动辄断掉重来很浪费时间。基于这个考虑, 对各类方法做个总结和排序:

方法类别		推荐程度	优点	缺点
基于URL	浏览器网页下载	☆☆☆	通用性好	手动麻烦/无多线程
	多线程下载器	☆☆☆☆	通用性好	手动麻烦
CLI工具	git clone 命令	☆☆	简单	无断点续传/冗余文件/无多线程
专用CLI工具	huggingface-cli + hf_transfer	☆☆☆	官方下载工具链, 功能最全	无进度条/容错性低
	huggingface-cli	☆☆☆☆☆	官方下载工具	不支持多线程
Python方法	snapshot_download	☆☆☆	官方支持, 功能全	脚本复杂/无多线程
	from_pretrained	☆	官方支持, 简单	不方便存储, 功能不全
	hf_hub_download	☆	官方支持	不支持全量下载/无多线程

另外对于数据集的下载和模型基本相同, 同理参考。

以下对上述方法进行介绍, 并介绍几个常见问题:

- [Q1: 如何下载 Llama 等需要登录的模型?](#)
- [Q2: 如何利用镜像站下载hf模型?](#)

1. 浏览器网页下载

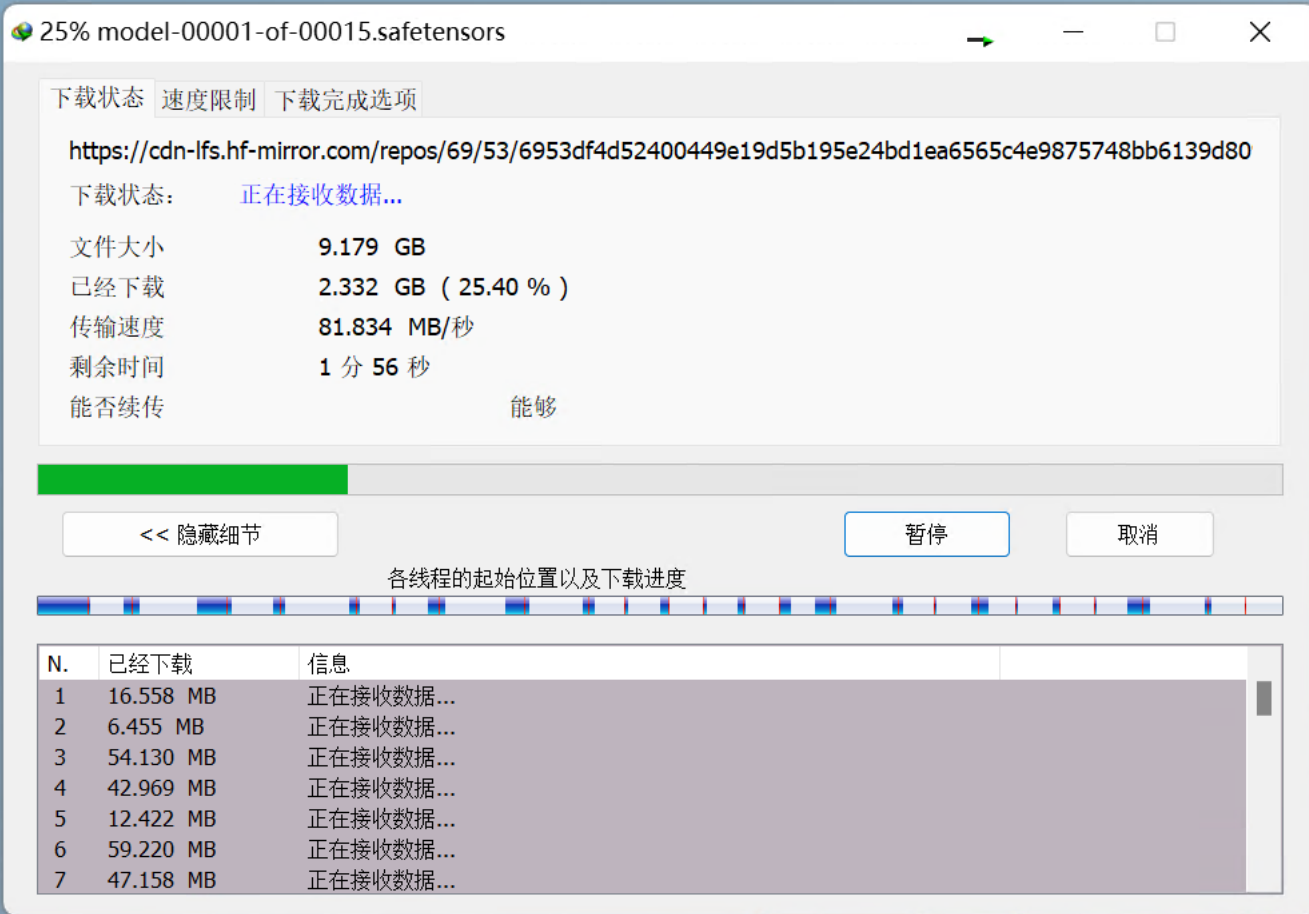
模型项目页的 Files 栏中可以获取文件的下载链接。直接网页复制下载链接，或用其他下载工具下载。

The screenshot shows the Hugging Face model page for `bert-base-chinese`. The 'Files' tab is selected and highlighted with a yellow circle. Below the tabs, the 'main' branch is selected. The file list shows several files, including `model.safetensors` (412 MB). A red circle highlights the download icon for `model.safetensors`, and a context menu is open over it, showing options like 'Open link in new tab', 'Open link in new window', 'Open link in incognito window', 'Save link as...', and 'Copy link address'. The URL bar shows the full path to the file: `https://huggingface.co/bert-base-chinese/resolve/main/model.safetensors`.

2. 多线程下载器

常规工具如浏览器默认采用单线程下载，由于国内网络运营商线路质量、QoS等因素有时候会很慢，多线程加速是一种有效、显著提高下载速度的方法。

经典多线程工具推荐两个：IDM、Aria2。IDM 适用于 Windows、aria2 适用于 Linux。本文头图就是 IDM 工具。因此获取URL后，可以利用这些多线程工具来下载。以我的一次实测为例，单线程700KB/s，IDM 8线程 6MB/s。千兆宽带下，利用IDM能跑到80MB/s+。



当然，手动获取仓库中所有 URL 并导入到多线程下载工具比较麻烦，因此我写了一个命令行脚本 [hfd.sh](#) ([Gist链接](#))，结合自动获取 url 以及 aria2 多线程下载，适合于 Linux。具体原理见下一节。

2.1 hfd 脚本

链接: [hfd.sh](#) ([Gist链接](#))，该工具同样支持设置镜像端点的环境变量:

```
export HF_ENDPOINT="https://hf-mirror.com"
```

基本命令:

```
./hfd.sh bigscience/bloom-560m --tool aria2c -x 4
```

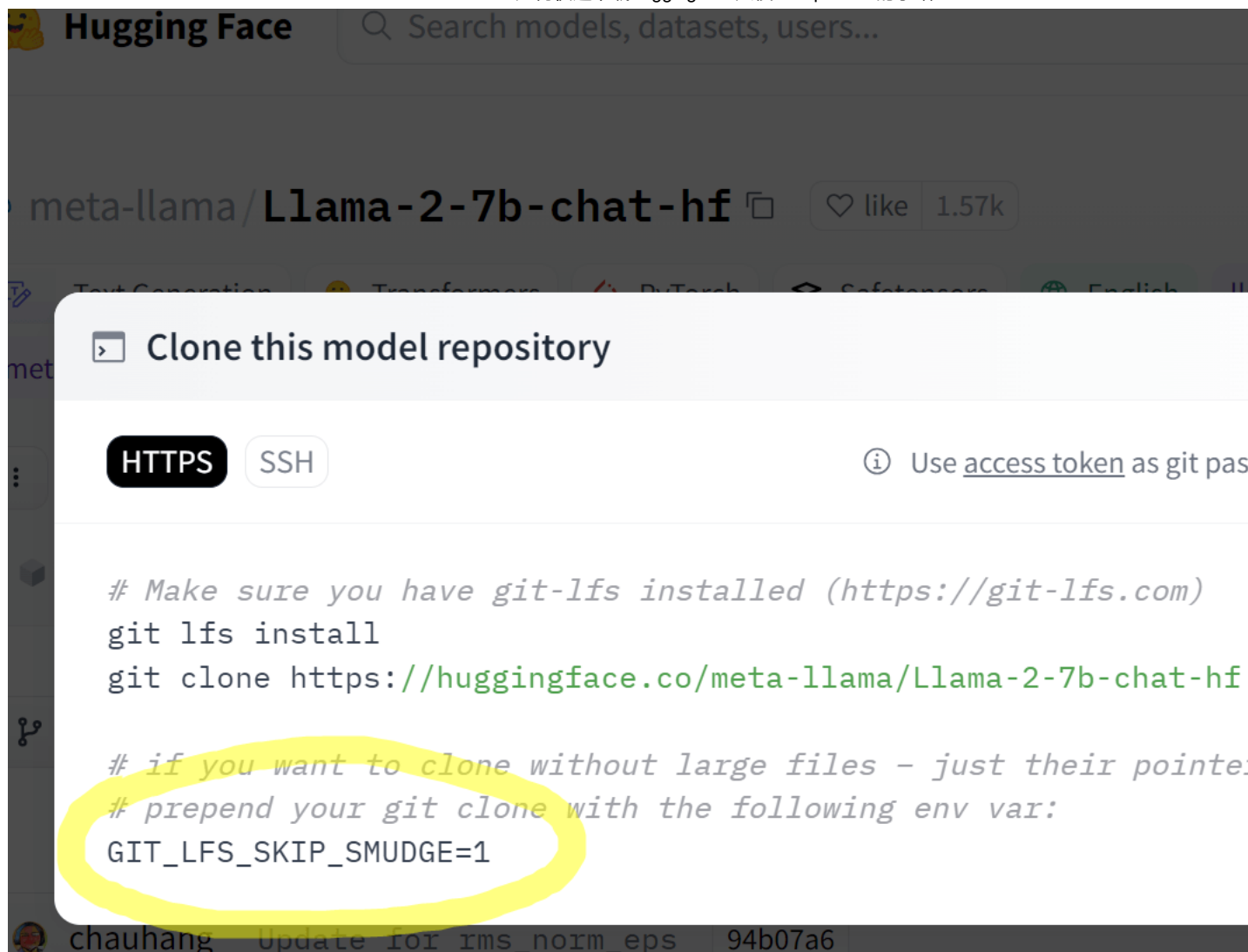
如果没有安装 aria2，则可以默认用 wget:

```
./hfd.sh bigscience/bloom-560m
```

3. Git clone

此外官方还提供了 `git clone repo_url` 的方式下载，这种方法相当简单，然而却是**最不推荐直接用的方法**，缺点有二：

- 1) 不支持断点续传，断了重头再来；
- 2) clone 会下载历史版本占用磁盘空间，即使没有历史版本，`.git` 文件夹大小也会存储一份当前版本模型的拷贝以及元信息，导致整个模型文件夹**磁盘占用两倍以上**，对于有些存在历史版本的模型，**下载时间两倍以上**，对于网络不够稳，磁盘不够大的用户，严重不推荐！



一种比较好的实践是，设置 `GIT_LFS_SKIP_SMUDGE=1` 环境变量（这可能也是为什么官方huggingface页面提到这个参数的原因），再 `git clone`，这样 Git 会先下载仓库中除了大文件之外的文件。然后我们再用一些支持断点续传的工具来下载大文件，这样既支持了断点续传，`.git` 目录也不会太大（一般几百KB）。整个流程，其实就是我上一节提到的 `hfd` 脚本的实现逻辑，感兴趣的可以参考/使用。

4. huggingface-cli+hf_transfer

`huggingface-cli` 和 `hf_transfer` 是 hugging face 官方提供的专门为下载而设计的工具链。前者是一个命令行工具，后者是下载加速模块。

4.1 huggingface-cli

`huggingface-cli` 隶属于 `huggingface_hub` 库，不仅可以下载模型、数据，还可以登录huggingface、上传模型、数据等。

安装依赖

```
pip install -U huggingface_hub
```

注意：`huggingface_hub` 依赖于 `Python>=3.8`，此外需要安装 `0.17.0` 及以上的版本，推荐 `0.19.0+`。

基本用法

```
huggingface-cli download --resume-download bigscience/bloom-560m --local-dir bloom-560m
```

下载数据

```
huggingface-cli download --resume-download --repo-type dataset lavita/medical-qa-shared-task-v1-toy
```

`huggingface-cli` 属于官方工具，其长期支持肯定是最好的。非常推荐。

除了长期支持这个优点，官方工具最大的一个优点，在于可以用模型名直接引用模型。

什么意思呢？我们知道，`from_pretrain` 函数可以接收一个模型的id，也可以接收模型的存储路径。

假如我们用浏览器下载了一个模型，存储到服务器的 `/data/gpt2` 下了，调用的时候你得写模型的绝对路径

```
AutoModelForCausalLM.from_pretrained("/data/gpt2")
```



然而如果你用的 `huggingface-cli download gpt2 --local-dir /data/gpt2` 下载，即使你把模型存储到了自己指定的目录，但是你仍然可以简单的用模型的名字来引用他。即：

```
AutoModelForCausalLM.from_pretrained("gpt2")
```



原理是因为huggingface工具链会在 `.cache/huggingface/` 下维护一份模型的符号链接，无论你是否指定了模型的存储路径，缓存目录下都会链接过去，这样可以避免自己忘了自己曾经下过某个模型，此外调用的时候就很方便。

所以用了官方工具，既可以方便的用模型名引用模型，又可以自己把模型集中存在一个自定义的路径，方便管理。

当然，该工具目前还是有一些缺点的：

一是其**存储逻辑不太直观**，其默认会把模型下载到 `~/.cache/huggingface/hub/` 中，即使设置了 `--local-dir`，也会采用符号链接的形式进行链接，其目的在于防止重复下载。然而我们有时候只想简单的下载到特定目录，其中有一项 `--local-dir-use-symlinks`，设置为 `False` 可以部分解决该问题，虽然仍会临时下载到 `~/.cache/huggingface/hub/`，但下载完成后会移动到 `--local-dir` 指定的目录。

二是由于上述逻辑的问题，主动Ctrl+C中断后，**断点续传**有时存在bug，导致同样的文件无法中断恢复，会重头下载。相信官方后续会改进。

三是**不支持单文件多线程**。目前的行为是多文件并行，一次性会同时下载多个文件。

四是**遇到网络中断会报错退出，不会自动重试**，需要重新手动执行。

4.2 hf_transfer

`hf_transfer` 依附并兼容 `huggingface-cli`，是 hugging face 官方专门为提高下载速度基于 Rust 开发的一个模块，开启后在带宽充足的机器上可以跑到 500MB/s。本人实测了三台不同网络环境的机器，确实有黑科技啊，都把带宽跑满了（千兆）。

然而缺点是：

- 1. **没有进度条**：是真的没有进度条，有进度条说明你没有开启成功。
- 2. **鲁棒性差**，遇到网络不稳定会报错，并提示用户考虑关闭该模块提高容错性。可能这个模块还没有很成熟吧，对国内这种丢包率高的网络还是水土不服。

尽管如此，还是推荐给大家，看各自网络情况吧。

项目地址：https://github.com/huggingface/hf_transfer。

开启方法

(1)安装依赖

```
pip install -U hf-transfer
```



(2)设置 HF_HUB_ENABLE_HF_TRANSFER 环境变量为 1

Linux

```
export HF_HUB_ENABLE_HF_TRANSFER=1
```



Windows Powershell

```
$env:HF_HUB_ENABLE_HF_TRANSFER = 1
```



开启后使用方法同 `huggingface-cli`：

```
huggingface-cli download --resume-download bigscience/bloom-560m --local-dir bloom-560m
```



注意：如果看到进度条，说明 hf_transfer 没开启成功！例如以下情况：

`--resume-download` 参数，指的是从上一次下载的地方继续，一般推荐总是加上该参数，断了方便继续。然而如果你一开始没有开启 `hf_transfer`，下载中途停掉并设置环境变量开启，此时用 `--resume-download` 会由于不兼容导致 `hf_transfer` 开启失败！总之观察是否有进度条就可以知道有没有开启成功，没有进度条就说明开启成功！

5. snapshot_download

huggingface 官方提供了 `snapshot_download` 方法下载完整模型，参数众多、比较完善。**相比下文另两个 python 方法，推荐 `snapshot_download` 方法来下载模型**，支持断点续传、指定路径、配置代理、排除特定文件等功能。然而有两个缺点：

- 1) 该方法依赖于 `transformers` 库，而这个库是个开发用的库，对于自动化运维有点重；
- 2) 该方法调用比较复杂，参数较多，例如默认会检查用户缓存目录下是否已有对应模型，如已有则会创建符号链接，不理解的容易导致问题。外加需要配置代理。最佳实践的参数配置如下：

```
from huggingface_hub import snapshot_download

snapshot_download(
    repo_id="bigscience/bloom-560m",
    local_dir="/data/user/test",
    local_dir_use_symlinks=False,
    proxies={"https": "http://localhost:7890"}
)
```



对于需要登录的模型，还需要两行额外代码：

```
import huggingface_hub
huggingface_hub.login("HF_TOKEN") # token 从 https://huggingface.co/settings/tokens 获取
```



很难记住这么多代码，经常性要下载模型的，不如用上文介绍的官方的命令行工具 `huggingface-cli` 了。

6. from_pretrained

不过多介绍了。常规方法。

7. hf_hub_download

不过多介绍了。常规方法。




Q1:如何下载hf上需要登陆的模型？






由于模型发布者的版权的要求，部分模型无法公开访问下载，需要在 `huggingface` 上申请许可通过后，才可以下载。这类模型称之为 `Gated Model`。基本步骤是：



- 1.申请许可
- 2.获取 `access token`（用于命令行和python方法调用）
- 3.下载




申请许可



此步骤必须在 `huggingface` **官网**注册登录后申请，由于网络安全原因，镜像站一般不支持。


 meta-llama/ **Llama-2-7b-chat-hf**   like 1.57k

 Text Generation  Transformers  PyTorch  Safetensors  English

meta llama-2  text-generation-inference  arxiv:2307.09288

 Train  Deploy  Use in Transformers

 Model card  Files

 **Access Llama 2 on Hugging Face**

This is a form to enable access to Llama 2 on Hugging Face after you have been granted access from Meta. Please visit the [Meta website](#) and accept our license terms and acceptable use policy before submitting this form. Requests will be processed in 1-2 days.

Your Hugging Face account email address **MUST** match the email you provide on the Meta website, or your request will not be approved.


Log in



 or


Sign Up


to review the conditions and access this model content.

Downloads last month
936,388

 **Safetensors**
Model size
Tensor type

 **Hosted**
 Text Generation
Inference API model.

 **Spaces**

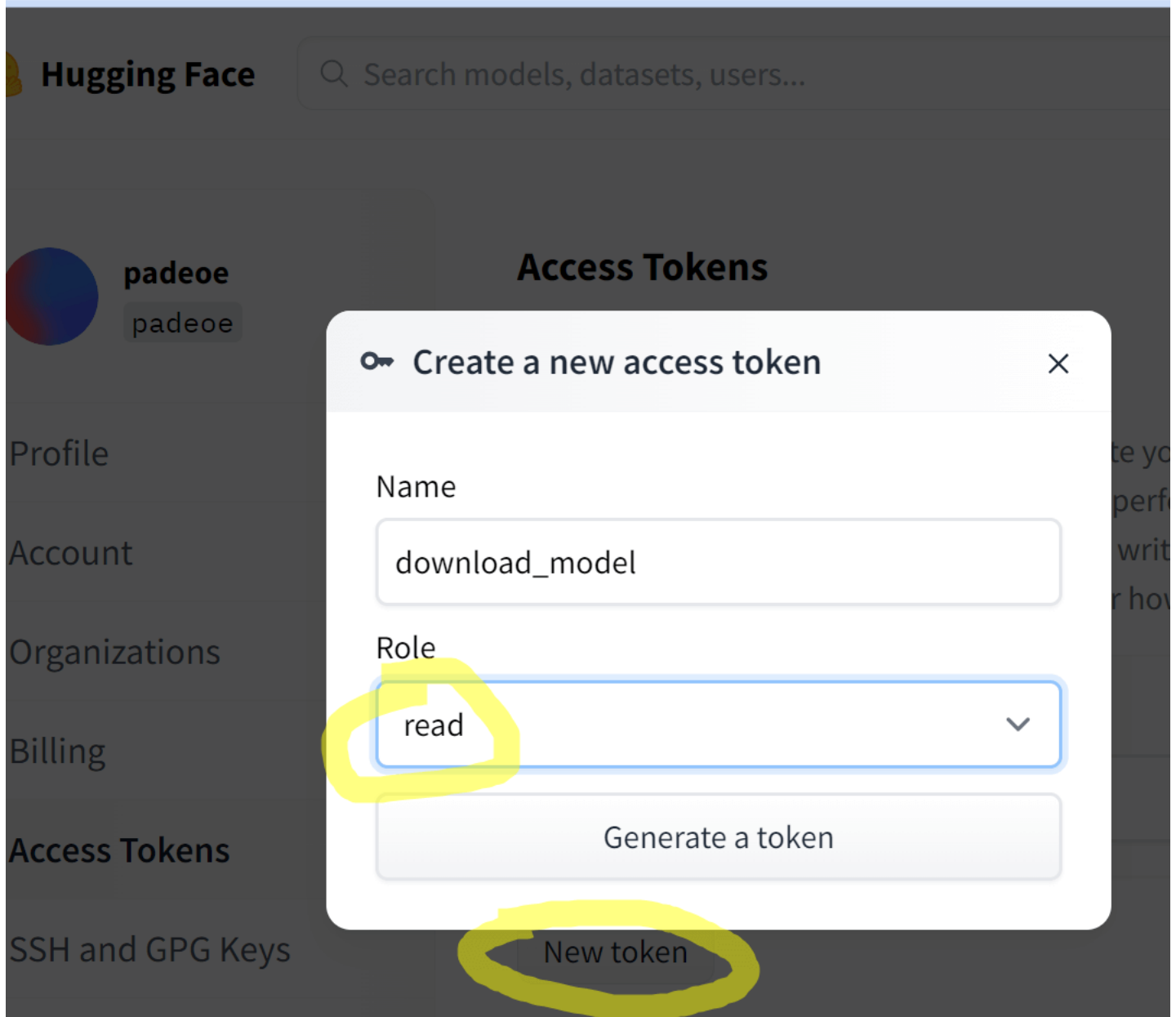
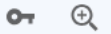
 Hugging Face

申请后一般等待几分钟到几天不等（一般几分钟就行），会发邮件通知你审批结果。

获取 access token

申请通过后，就可以在模型主页的 Files and versions 中看到模型文件了，浏览器的话直接点击下载即可。但是如果想要用工具例如 huggingface-cli 下载，则需要获取 access token。

Access Token 获取地址： <https://huggingface.co/settings/tokens>

huggingface.co/settings/tokens?new_token=true

访问 huggingface 设置页面的 token 管理页，选择 New 一个 token，只需要 Read 权限即可，创建后便可以在工具中调用时使用了。

下载

除了登陆后浏览器直接下载，几种工具的使用方法分别介绍如下：

Git

```
git clone https://hf_username:hf_token@huggingface.co/meta-llama/Llama-2-7b-chat-hf
```



huggingface-cli: 添加 -token 参数

```
huggingface-cli download --token hf_*** --resume-download bigscience/bloom-560m --local-dir bloom-560m
```



curl, wget: 在 header 中添加 token

```
curl -L --header "Authorization: Bearer hf_***" -o model-00001-of-00002.safetensors https://huggingface.co/meta-llama/Llama-2-7b-chat-hf/resolve/main/model-00001-of-00002.safetensors
```



```
wget --header "Authorization: Bearer hf_***" https://huggingface.co/meta-llama/Llama-2-7b-chat-hf/resolve/main/model-00001-of-00002.safetensors
```

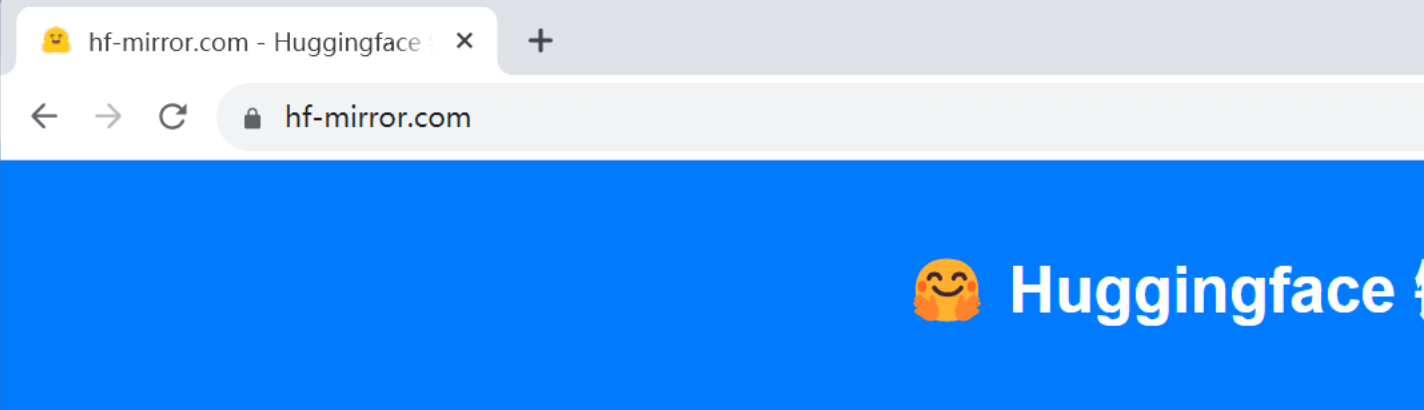


snapshot_download: 调用 login 方法

```
import huggingface_hub
huggingface_hub.login("hf_***")
```

**Q2:如何利用镜像站下载hf模型?**

直接访问镜像站，获取文件URL

镜像站 <https://hf-mirror.com>.

hf-mirror.com - Huggingface

hf-mirror.com

Huggingface

搜索模型...

本站域名 hf-mirror.com，用于镜像 huggingface.co 域名。使用

1. 方法一：我们推荐使用 [huggingface](#) 官方提供的 [huggingface](#)

(1) 安装依赖

```
pip install -U huggingface_hub hf_transfer
```

(2) 基本命令示例：

```
export HF_ENDPOINT=https://hf-mirror.com
```

```
huggingface-cli download --resume-download bigscience/b
```


如需提高下载速度，推荐设置环境变量开启 [hf_transfer](#)，官方

```
export HF_HUB_ENABLE_HF_TRANSFER=1
```

(3) 下载需要登录的模型 (Gated Model)

请添加 `--token hf_***` 参数，其中 `hf_***` 是 *access token*，

```
huggingface-cli download --token hf_*** --resume-downlo
```

GitHub:  padeoe

代码类工具设置 HF_ENDPOINT 环境变量

适用于 huggingface 官方的工具和库，包括：

- huggingface-cli
- snapshot_download
- from_pretrained
- hf_hub_download
- timm.create_model

设置方法

Windows Powershell

```
$env:HF_ENDPOINT = "https://hf-mirror.com"
```



Linux

```
export HF_ENDPOINT="https://hf-mirror.com"
```



Python

```
import os
os.environ['HF_ENDPOINT'] = 'https://hf-mirror.com'
```



注意 os.environ 得在import huggingface库相关语句之前执行。

总结

以上，我们介绍了浏览器、多线程工具、git clone、huggingface-cli、hf_transfer、python方法、hfd脚本等众多方法，各自有其适用场景，大家根据自己的操作系统的支持情况以及个人习惯来选择。

个人推荐：

- Linux/Mac OS/windows **默认推荐使用** huggingface-cli ，对外网连接较好（丢包少）的时候，可尝试 huggingface-cli+hf_transfer（可选）。
- 网络连接不好，考虑用多线程工具，Linux 推荐 aria2 ， Windows 推荐 IDM。
- 偶尔小文件下载，直接访问镜像站，用浏览器下载。
- **不推荐** Git clone （可以被 huggingface-cli 替代），但如确有需要，小模型、小数据集可以 Git clone，建议文件大不要直接 clone，设置环境变量 GIT_LFS_SKIP_SMUDGE=1 再 clone，大文件单独用别的工具下载。

最后，使用问题、建议和技术交流可加群 😊



群聊：群聊



该二维码7天内(12月14日前)有效，重新进入将更新

[cli^{\[1\]}download^{\[1\]}huggingface^{\[1\]}mirror^{\[1\]}下载^{\[1\]}镜像^{\[1\]}镜像站^{\[1\]}](#)
[Previousname.com DDNS](#)
[Next强制使用HTTP/3连接解决SNI阻断的问题](#)



padeoe

共有 21 条评论

1.  **Tao**说道:
[2023-10-31 10:15](#)

服务器连不上huggingface, 找了好久没找到最合适的办法, 最后搜到了hf-mirror无缝下载, 特来感谢 😊

[回复](#)

1.  **padeoe**说道:
[2023-10-31 11:38](#)



[回复](#)

2.  **匿名**说道:
[2023-11-06 12:10](#)

特来感谢

[回复](#)

1.  **padeoe**说道:
[2023-11-06 15:00](#)



[回复](#)

3.  **匿名**说道:
[2023-11-06 16:54](#)

你好, 我是在Win下安装了huggingface_hub, 但是 "export HF_ENDPOINT=" https://hf-mirror.com" " 指令没有用, 显示说HF_ENDPOINT是invalid syntax, 请问这个怎么解决?

[回复](#)

1.  **padeoe**说道:
[2023-11-06 17:47](#)

Windows请使用Powershell执行 `$env:HF_ENDPOINT = "https://hf-mirror.com"` 这个命令来设置环境变量, 本文中也有介绍的

[回复](#)

4.  **17**说道:
[2023-11-07 22:05](#)

设置完毕环境变量后, 原来的from_pretrained函数还是从huggingface上下载, 是我设置的有问题么, 编程小白一个

[回复](#)

1.  **padeoe**说道:
[2023-11-08 09:11](#)


应该是你设置有问题 😊 正确设置后from_pretrained会从镜像站下载, 不过即使正确设置, from_pretrained下载也会有点慢, 不行就先用别的工具下载好, 然后from_pretrained(绝对路径)吧

[回复](#)

5.  **17**说道:
[2023-11-07 22:06](#)

非常感谢!!!!

[回复](#)

6.  **hugo**说道:
[2023-12-02 23:51](#)

非常感谢，微信群超200人了，还能通过什么方式加呀

[回复](#)

1.  **padeoe**说道:
[2023-12-07 13:53](#)

已经更新二维码，建了一个新群，进去后加我微信再拉到旧群

[回复](#)

7.  **匿名**说道:
[2023-12-04 17:23](#)

我在下载baichuan-inc/Baichuan2-13B-Chat-4bits 9.08GB模型文件时经常会断连，下载失败。

```
requests.exceptions.ConnectionError: (MaxRetryError( "HTTPConnectionPool(host=' cdn-lfs.hf-mirror.com' , port=443): Max retries exceeded with url: /repos/d9/9a/d99a1b30a14667579ad3380bc222037f7970e5065e74a9faa784a8a97672f2bd/1d805c460a12cd7be30c7b9a200ad7a14364ab response-content-disposition . . . . . Key-Pair-Id=KVTP0A1DKRTAX (Caused by NewConnectionError( ' : Failed to establish a new connection: [Errno 101] Network is unreachable' ))" ), ' (Request ID: 481a2940-140d-4884-ad8f-6a0564a2ea9e)' )
```

[回复](#)

8.  **wwi**说道:
[2023-12-08 15:16](#)

我对下面这段话理解的不是特别好：

一是其存储逻辑不太直观，其默认会把模型下载到 ~/.cache/huggingface/hub/ 中，即使设置了 --local-dir，也会采用符号链接的形式进行链接，其目的在于防止重复下载。然而我们有时候只想简单的下载到特定目录，其中有一项 --local-dir-use-symlinks，设置为 False 可以部分解决该问题，虽然仍会临时下载到 ~/.cache/huggingface/hub/，但下载完成后会移动到 --local-dir 指定的目录。

请问如果我设置了--local-dir-use-symlinks False 那么我还能使用AutoModelForCausalLM.from_pretrained("gpt2")来加载模型吗？我是不是应该使用AutoModelForCausalLM.from_pretrained("--local-dir")？

[回复](#)

1.  **padeoe**说道:
[2023-12-08 16:18](#)


针对你的两个问题：不能；是的。😄

[回复](#)

9.  **匿名**说道:
[2023-12-15 15:49](#)

我想在内网环境下使用类似hf-mirror镜像的方式下载模型，模型量只有常用的几百个，之前是将这些模型下载到本地（文件结构：机构/模型/版本/），起一个服务，并把huggingface_hub源码中的HF_ENDPOINT替换成服务的url，但是环境中的这个库很容易被替换掉。请问我该怎么在内网中构建一个类似hf-mirror的小规模镜像？

[回复](#)

10.  **匿名**说道:
[2023-12-18 10:29](#)


请问站点模型是同步huggingface版本的吗？

[回复](#)

1.  **padeoe**说道:
[2023-12-19 12:26](#)

完全实时同步的

[回复](#)

11.  **匿名**说道:
[2023-12-18 15:07](#)

请问这个的原理是，使用nginx类似的工具代理么？

[回复](#)

1.  **padeoe**说道:
[2023-12-19 12:26](#)

是的, <https://github.com/padeoe/hf-mirror-site>

[回复](#)

12.  **匿名**说道:
[2024-01-23 03:17](#)

请教下, 最近使用镜像站+官方cli下载后, 会报两个error, 不过我看东西倒是下下来了, 正常吗。
huggingface_hub.utils_errors.FileMetadataError: Distant resource does not seem to be on huggingface.co. It is possible that a configuration issue prevents you from downloading resources from <https://huggingface.co>. Please check your firewall and proxy settings and make sure your SSL certificates are updated.

huggingface_hub.utils_errors.LocalEntryNotFoundError: An error happened while trying to locate the file on the Hub and we cannot find the requested files in the local cache. Please check your connection and try again or make sure your Internet connection is on.

[回复](#)

1.  **padeoe**说道:
[2024-01-24 10:35](#)

第一条错误提示模型元数据获取存在问题, 应该不影响模型文件本身的完整性。第二条错误不确定是否影响。

[回复](#)

发表回复 取消回复

您的电子邮箱地址不会被公开。 必填项已用*标注

显示名称

电子邮箱地址

网站地址

评论 *

[发表评论](#)

Just a [bigfa](#) themeBlog since 2016

