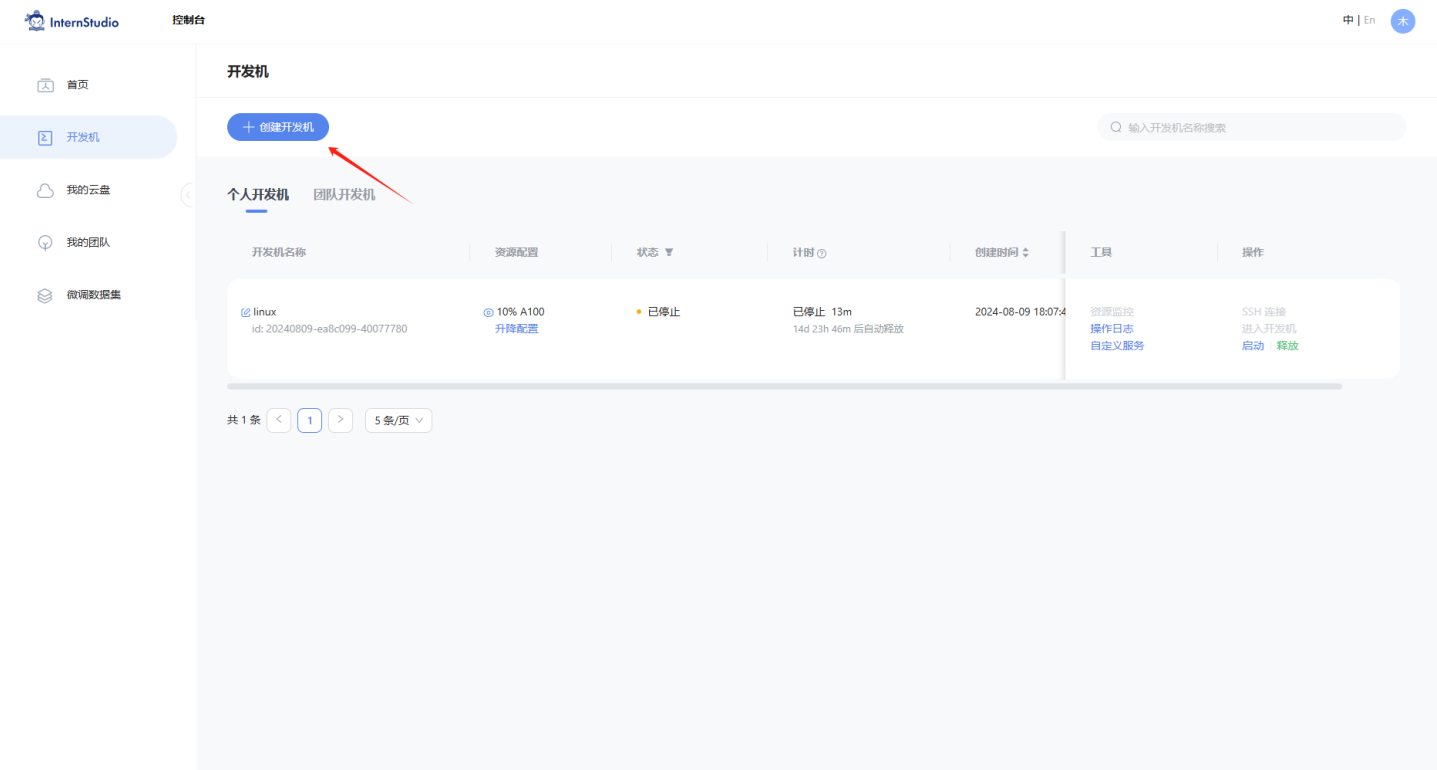


基础岛任务笔记2-8G显存玩转书生大模型Demo

- 1 8G 显存玩转书生大模型 Demo
- 2 记录复现过程并截图
- 3 基础任务（完成此任务即完成闯关）
- 4 使用 Cli Demo 完成 InternLM2-Chat-1.8B 模型的部署，并生成 300 字小故事，记录复现过程并截图。
- 5 进阶任务（闯关不要求完成此任务）
- 6 使用 LMDeploy 完成 InternLM-XComposer2-VL-1.8B 的部署，并完成一次图文理解对话，记录复现过程并截图。
- 7 使用 LMDeploy 完成 InternVL2-2B 的部署，并完成一次图文理解对话，记录复现过程并截图。
- 8 闯关材料提交（完成任务并且提交材料视为闯关成功）
- 9 闯关作业总共分为一个任务，一个任务完成视作闯关成功。
- 10 请将作业发布到知乎、CSDN等任一社交媒体，将作业链接提交到以下问卷，助教老师批改后将获得 100 算力点奖励！！
- 11 提交地址：<https://aicarrier.feishu.cn/share/base/form/shrcnZ4bQ4YmhEtMtnKxZUcf1vd>

创建开发机

在输入开发机名称后，点击创建开发机



选择 10% 的开发机，镜像选择为 Cuda-12.2

首页

开发机

我的云盘

我的团队

微调数据集

← 开发机 / 创建开发机

开发机类型

☒ 个人开发机 ☐ 团队开发机

开发机名称

请输入

镜像

请选择具体镜像

选择镜像 >

资源配置

GPU	显存
<input checked="" type="radio"/> 10% A100 * 1	8192 MiB
<input type="radio"/> 30% A100 * 1	24576 MiB
<input type="radio"/> 50% A100 * 1	40960 MiB
<input type="radio"/> A100 * 1	81920 MiB
<input type="radio"/> A100 * 2	163840 MiB

立即创建

取消

费用 1 算力点 / 小时, 预计消费 8.00 算力点
当前账户余额 166.77 算力点。

基础镜像

返回

Cuda12.2-conda

8 使用量 1

浏览量 1

更新时间 2024-03-28 15:54:29

发布者 wfbt

基础镜像

Ubuntu 20.04 系统环境, 内置 cudatoolkit12.2, 云盘 (/root) 内预装 conda, 默认 conda activate base 环境

版本

详细信息

版本号	版本简介	版本大小	创建时间	操作
gpu-12.2	Ubuntu 20.04, cuda 12.2...	9.59GB	2024-03-28 15:57:04	使用

个人开发机

团队开发机

开发机名称	资源配置	状态 ▼	计时 ⌚	创建时间 ⬆	工具	操作
<div><div>Task</div><div>id: 20240826-8e96ec1-40077780</div></div>	<div><div>10% A100</div><div>升降配置</div></div>	<div><div>排队中</div><div>位于队列第 1</div></div>	-	2024-08-26 21:41:5	<div>资源监控</div> <div>操作日志</div> <div>自定义服务</div>	<div>SSH 连接</div> <div>进入开发机</div> <div>停止 删除</div>

环境配置

我们首先来为 Demo 创建一个可用的环境。

```
1 # 创建环境
2 conda create -n demo python=3.10 -y
3 # 激活环境
4 conda activate demo
5 # 安装 torch
6 conda install pytorch==2.1.2 torchvision==0.16.2 torchaudio==2.1.2 pytorch-
  cuda=12.1 -c pytorch -c nvidia -y
7 # 安装其他依赖
8 pip install transformers==4.38
9 pip install sentencepiece==0.1.99
10 pip install einops==0.8.0
11 pip install protobuf==5.27.2
12 pip install accelerate==0.33.0
13 pip install streamlit==1.37.0
```

创建一个新环境，名字为demo：

```
(base) root@intern-studio-40077780:~# conda create -n demo python=3.10 -y
Retrieving notices: ...working... done
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /root/.conda/envs/demo

  added / updated specs:
    - python=3.10

The following packages will be downloaded:



| package                  | build           |         |          |
|--------------------------|-----------------|---------|----------|
| -----                    | -----           |         |          |
| _libgcc_mutex-0.1        | main            | 3 KB    | defaults |
| _openmp_mutex-5.1        | 1_gnu           | 21 KB   | defaults |
| bzip2-1.0.8              | h5eee18b_6      | 262 KB  | defaults |
| ca-certificates-2024.7.2 | h06a4308_0      | 127 KB  | defaults |
| ld_impl_linux-64-2.38    | h1181459_1      | 654 KB  | defaults |
| libffi-3.4.4             | h6a678d5_1      | 141 KB  | defaults |
| libgcc-ng-11.2.0         | h1234567_1      | 5.3 MB  | defaults |
| libgomp-11.2.0           | h1234567_1      | 474 KB  | defaults |
| libstdcxx-ng-11.2.0      | h1234567_1      | 4.7 MB  | defaults |
| libuuid-1.41.5           | h5eee18b_0      | 27 KB   | defaults |
| ncurses-6.4              | h6a678d5_0      | 914 KB  | defaults |
| openssl-3.0.14           | h5eee18b_0      | 5.2 MB  | defaults |
| pip-24.2                 | py310h06a4308_0 | 2.3 MB  | defaults |
| python-3.10.14           | h955ad1f_1      | 26.8 MB | defaults |
| readline-8.2             | h5eee18b_0      | 357 KB  | defaults |
| setuptools-72.1.0        | py310h06a4308_0 | 2.4 MB  | defaults |
| sqlite-3.45.3            | h5eee18b_0      | 1.2 MB  | defaults |
| tk-8.6.14                | h39e8969_0      | 3.4 MB  | defaults |
| tzdata-2024a             | h04d1e81_0      | 116 KB  | defaults |
| wheel-0.43.0             | py310h06a4308_0 | 110 KB  | defaults |
| xz-5.4.6                 | h5eee18b_1      | 643 KB  | defaults |
| zlib-1.2.13              | h5eee18b_1      | 111 KB  | defaults |
| -----                    | -----           |         |          |
| Total:                   |                 | 55.3 MB |          |



The following NEW packages will be INSTALLED:



|                  |                                                                  |
|------------------|------------------------------------------------------------------|
| _libgcc_mutex    | anaconda/pkgs/main/linux-64::_libgcc_mutex-0.1-main              |
| _openmp_mutex    | anaconda/pkgs/main/linux-64::_openmp_mutex-5.1-1_gnu             |
| bzip2            | anaconda/pkgs/main/linux-64::bzip2-1.0.8-h5eee18b_6              |
| ca-certificates  | anaconda/pkgs/main/linux-64::ca-certificates-2024.7.2-h06a4308_0 |
| ld_impl_linux-64 | anaconda/pkgs/main/linux-64::ld_impl_linux-64-2.38-h1181459_1    |
| libffi           | anaconda/pkgs/main/linux-64::libffi-3.4.4-h6a678d5_1             |
| libgcc-ng        | anaconda/pkgs/main/linux-64::libgcc-ng-11.2.0-h1234567_1         |


```

激活demo环境并且安装torch及其他依赖，时间有点长

CPU 11.83%GPU-1: 10% Nvidia A100 0%内存 0.92 / 24 GB 3.84%显存 0 / 8192 MiB 0%

```
(base) root@intern-studio-40077780: ~# conda activate demo
(demo) root@intern-studio-40077780: ~# conda install pytorch==2.1.2 torchvision==0.16.2 torchaudio==2.1.2 pytorch-cuda=12.1 -c pytorch -c nvidia -y

Collecting package metadata (current_repodata.json): done
Solving environment: unsuccessful initial attempt using frozen solve. Retrying with flexible solve.
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

environment location: /root/.conda/envs/demo

added / updated specs:
- pytorch-cuda=12.1
- pytorch==2.1.2
- torchaudio==2.1.2
- torchvision==0.16.2

The following packages will be downloaded:
```

package	build		
blas-1.0	mk1	6 KB	defaults
brotili-python-1.0.9	py310h6a678d5_8	356 KB	defaults
certifi-2024.7.4	py310h06a4308_0	158 KB	defaults
charset-normalizer-3.3.2	pyhd3eb1b0_0	44 KB	defaults
cuda-cudart-12.1.105	0	189 KB	nvidia
cuda-cupti-12.1.105	0	15.4 MB	nvidia
cuda-libraries-12.1.0	0	2 KB	nvidia
cuda-nvrtc-12.1.105	0	19.7 MB	nvidia
cuda-nvtx-12.1.105	0	57 KB	nvidia
cuda-opencl-12.6.37	0	26 KB	nvidia
cuda-runtime-12.1.0	0	1 KB	nvidia
cuda-version-12.6	3	16 KB	nvidia
ffmpeg-4.3	hf484d3e_0	9.9 MB	pytorch
filelock-3.13.1	py310h06a4308_0	21 KB	defaults
freetype-2.12.1	h4a9f257_0	626 KB	defaults
gmp-6.2.1	h295c915_3	544 KB	defaults
gmpy2-2.1.2	py310heeb90bb_0	517 KB	defaults
gnutls-3.6.15	he1e5248_0	1.0 MB	defaults
idna-3.7	py310h06a4308_0	130 KB	defaults
intel-openmp-2023.1.0	hdb19cb5_46306	17.2 MB	defaults
jinja2-3.1.4	py310h06a4308_0	278 KB	defaults
jpeg-9e	h5eee18b_3	262 KB	defaults
lame-3.100	h7b6447c_0	323 KB	defaults
lcms2-2.12	h3be6417_0	312 KB	defaults
lerc-3.0	h295c915_0	196 KB	defaults
libcublas-12.1.0.26	0	329.0 MB	nvidia
libcufft-11.0.2.4	0	102.9 MB	nvidia
libcufile-1.11.0.15	0	1.0 MB	nvidia
libcurand-10.3.7.37	0	51.8 MB	nvidia

```
CPU 10.2% GPU-1: 10% Nvidia A100 0% 内存 0.55 / 24 GB 2.29% 显存 0 / 8182 MiB 0%
Downloading https://pypi.tuna.tsinghua.edu.cn/packages/65/58/f9c9e6be752e9fcb8b6a0ee9fb87e6e7a1f6bcab2cdc73f02bb7ba
91ada0/tzdata-2024.1-py2.py3-none-any.whl (345 kB)
Requirement already satisfied: charset-normalizer<4,>=2 in ./conda/envs/demo/lib/python3.10/site-packages (from requ
ests<3,>=2.27->streamlit==1.37.0) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in ./conda/envs/demo/lib/python3.10/site-packages (from requests<3,>=2.2
7->streamlit==1.37.0) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in ./conda/envs/demo/lib/python3.10/site-packages (from requests<3
,>=2.27->streamlit==1.37.0) (2.2.2)
Requirement already satisfied: certifi>=2017.4.17 in ./conda/envs/demo/lib/python3.10/site-packages (from requests<3
,>=2.27->streamlit==1.37.0) (2024.7.4)
Collecting markdown-it-py>=2.2.0 (from rich<14,>=10.14.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/42/d7/1ec15b46af6af88f19b8e5ffea08fa375d433c998b8a7639e76935
c14f1f/markdown_it_py-3.0.0-py3-none-any.whl (87 kB)
Collecting pygments<3.0.0,>=2.13.0 (from rich<14,>=10.14.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/f7/3f/01c8b82017c199075f8f788d0d906b9ffbbc5a47dc9918a945e13d
5a2bda/pygments-2.18.0-py3-none-any.whl (1.2 MB)
1.2/1.2 MB 7.8 MB/s eta 0:00:00
Collecting smmap<6,>=3.0.1 (from gitdb<5,>=4.0.1->gitpython!=3.1.19,<4,>=3.0.7->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/a7/a5/10f97f73544edcdef54409f1d839f6049a0d79df68adbclceb24d1
aaca42/smmmap-5.0.1-py3-none-any.whl (24 kB)
Requirement already satisfied: MarkupSafe>=2.0 in ./conda/envs/demo/lib/python3.10/site-packages (from jinja2->altai
r<6,>=4.0->streamlit==1.37.0) (2.1.3)
Collecting attrs>=22.2.0 (from jsonschema>=3.0->altair<6,>=4.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/6a/21/5b6702a7f963e95456c0de2d495f67bf5fd62840ac655dc451586d
23d39a/attrs-24.2.0-py3-none-any.whl (63 kB)
Collecting jsonschema-specifications>=2023.03.6 (from jsonschema>=3.0->altair<6,>=4.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/ee/07/44bd408781594c4d0a027666ef27fab1e441b109dc3b76b4f836f8
fd04fe/jsonschema_specifications-2023.12.1-py3-none-any.whl (18 kB)
Collecting referencing>=0.28.4 (from jsonschema>=3.0->altair<6,>=4.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/b7/59/2056f61236782a2c86b33906c025d4f4a0b17be0161b63b70fd9e8
775d36/referencing-0.35.1-py3-none-any.whl (26 kB)
Collecting rpds-py>=0.7.1 (from jsonschema>=3.0->altair<6,>=4.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/4a/6d/1166a157b227f2333f8e8ae320b6b7ea2a6a38f8e7a3563ad76dff
c8608d/rpds_py-0.20.0-cp310-cp310-manylinux_2_17_x86_64_manylinux2014_x86_64.whl (354 kB)
Collecting mdurl~>=0.1 (from markdown-it-py>=2.2.0->rich<14,>=10.14.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/b3/38/89ba8ad64ae25be8da66a6d463314cf1eb366222074cfda9ee839c
56a4b4/mdurl-0.1.2-py3-none-any.whl (10.0 kB)
Collecting six>=1.5 (from python-dateutil>=2.8.2->pandas<3,>=1.3.0->streamlit==1.37.0)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/d9/5a/e7c31adbe875f2abbb91bd84cf2dc52d792b5a01506781dbcf25c9
1daf11/six-1.16.0-py2.py3-none-any.whl (11 kB)
Installing collected packages: pytz, watchdog, tzdata, tornado, toml, tenacity, smmap, six, rpds-py, pygments, pyarro
w, narwhals, mdurl, click, cachetools, blinker, attrs, referencing, python-dateutil, pydeck, markdown-it-py, gitdb, r
ich, pandas, jsonschema-specifications, gitpython, jsonschema, altair, streamlit
Successfully installed altair-5.4.0 attrs-24.2.0 blinker-1.8.2 cachetools-5.5.0 click-8.1.7 gitdb-4.0.11 gitpython-3.
1.43 jsonschema-4.23.0 jsonschema-specifications-2023.12.1 markdown-it-py-3.0.0 mdurl-0.1.2 narwhals-1.5.5 pandas-2.2
.2 pyarrow-17.0.0 pydeck-0.9.1 pygments-2.18.0 python-dateutil-2.9.0.post0 pytz-2024.1 referencing-0.35.1 rich-13.7.1
rpds-py-0.20.0 six-1.16.0 smmap-5.0.1 streamlit-1.37.0 tenacity-8.5.0 toml-0.10.2 tornado-6.4.1 tzdata-2024.1 watchd
og-4.0.2
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system pa
ckage manager, possibly rendering your system unusable.It is recommended to use a virtual environment instead: https:
//pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppress th
is warning.
(demo) root@intern-studio-40077780: ~ #
```

Cli Demo 部署 InternLM2-Chat-1.8B 模型

第一步：在之前已经创建的demo文件夹中创建一个 `cli_demo.py`。

- 1 `mkdir -p /root/demo`
- 2 `touch /root/demo/cli_demo.py`

第二步：利用vim编辑 `cli_demo.py` 文件,也可将下面的代码直接复制到 `cli_demo.py` 中。

```
1 import torch
2 from transformers import AutoTokenizer, AutoModelForCausalLM
3
4
5 model_name_or_path = "/root/share/new_models/Shanghai_AI_Laboratory/internlm2-
  chat-1_8b"
6
7 tokenizer = AutoTokenizer.from_pretrained(model_name_or_path,
  trust_remote_code=True, device_map='cuda:0')
8 model = AutoModelForCausalLM.from_pretrained(model_name_or_path,
  trust_remote_code=True, torch_dtype=torch.bfloat16, device_map='cuda:0')
9 model = model.eval()
10
11 system_prompt = """You are an AI assistant whose name is InternLM (书生·浦语).
12 - InternLM (书生·浦语) is a conversational language model that is developed by
  Shanghai AI Laboratory (上海人工智能实验室). It is designed to be helpful,
  honest, and harmless.
13 - InternLM (书生·浦语) can understand and communicate fluently in the language
  chosen by the user such as English and 中文.
14 """
15
16 messages = [(system_prompt, '')]
17
18 print("=====Welcome to InternLM chatbot, type 'exit' to
  exit.=====")
19
20 while True:
21     input_text = input("\nUser >>> ")
22     input_text = input_text.replace(' ', '')
23     if input_text == "exit":
24         break
25
26     length = 0
27     for response, _ in model.stream_chat(tokenizer, input_text, messages):
28         if response is not None:
29             print(response[length:], flush=True, end="")
30             length = len(response)
```

```
CPU 7.1% GPU-1: 10% Nvidia A100 0%
内存 0.46 / 24 GB 1.9% 显存 0 / 8182 MiB 0%

5. 将conda环境一键添加到jupyterlab:
lab add {YOUR_CONDA_ENV_NAME}

-----

(base) root@intern-studio-40077780:~# ls
Tutorial demo demo2 file share
(base) root@intern-studio-40077780:~# cd demo
(base) root@intern-studio-40077780:~/demo# ls
hello_world.py linux_task linux_task2 share task.yml task1 test.sh
(base) root@intern-studio-40077780:~/demo# pwd
/root/demo
(base) root@intern-studio-40077780:~/demo# touch cli_demo.py
(base) root@intern-studio-40077780:~/demo# ls
cli_demo.py hello_world.py linux_task linux_task2 share task.yml task1 test.sh
(base) root@intern-studio-40077780:~/demo# vim cli_demo.py
(base) root@intern-studio-40077780:~/demo# cat cli_demo.py
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM

model_name_or_path = "/root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-1_8b"

tokenizer = AutoTokenizer.from_pretrained(model_name_or_path, trust_remote_code=True, device_map='cuda:0')
model = AutoModelForCausalLM.from_pretrained(model_name_or_path, trust_remote_code=True, torch_dtype=torch.bfloat16, device_map='cuda:0')
model = model.eval()

system_prompt = """You are an AI assistant whose name is InternLM (书生 浦语).
- InternLM (书生 浦语) is a conversational language model that is developed by Shanghai AI Laboratory (上海人工智能实验室). It is designed to be helpful, honest, and harmless.
- InternLM (书生 浦语) can understand and communicate fluently in the language chosen by the user such as English and 中文.
"""

messages = [(system_prompt, '')]

print("=====Welcome to InternLM chatbot, type 'exit' to exit.=====")

while True:
    input_text = input("\nUser >>> ")
    input_text = input_text.replace(' ', '')
    if input_text == "exit":
        break

    length = 0
    for response, _ in model.stream_chat(tokenizer, input_text, messages):
        if response is not None:
            print(response[length:], flush=True, end="")
            length = len(response)
(base) root@intern-studio-40077780:~/demo#
```

最后一步：通过 `python /root/demo/cli_demo.py` 启动 Demo（注意环境用“demo”虚拟环境）。问了俩问题，小模型有点不理解问题很正常。

然后，启动一个 Streamlit 服务。

```
1 cd /root/demo
2 streamlit run /root/demo/Tutorial/tools/streamlit_demo.py --server.address
  127.0.0.1 --server.port 6006
```

```
(demo) root@intern-studio-40077780: # ls
tutorial demo demo2 file share
(demo) root@intern-studio-40077780:~# streamlit run Tutorial/tools/streamlit_demo.py --server.address 127.0.0.1

Collecting usage statistics. To deactivate, set browser.gatherUsageStats to false.

You can now view your Streamlit app in your browser.

URL: http://127.0.0.1:8501
```

接下来，在**本地**的 PowerShell 中输入以下命令，将端口映射到本地。

```
ssh -CNg -L 6006:127.0.0.1:6006 root@ssh.intern-ai.org.cn -p 你的 ssh 端口号
```

然后将 SSH 密码复制并粘贴到 PowerShell 中，回车，即可完成端口映射。

在完成端口映射后，通过浏览器访问 <http://localhost:6006> 来启动 Demo。



