

基于Qwen-14B-Chat的本地Agent—环境搭建

1. Qwen-14B 模型下载

在云服务器中，使用 ModelScope 平台的 modelscope 包进行模型下载是一种非常高效的方法。

The screenshot shows the ModelScope platform's search interface. The search bar at the top contains the query "qwen". Below the search bar, there are several cards displaying different AI models. The first card is for "通义千问-7B-Chat" (Qwen-7B-Chat), which is described as a 7B parameter model based on Transformer. The second card is for "通义千问-14B-Chat" (Qwen-14B-Chat), which is described as a 14B parameter model based on Transformer. The third card is for "通义千问-VL-Chat" (Qwen-VL-Chat), which is a multimodal model. The fourth card is for "通义千问-14B-Chat-Int4" (Qwen-14B-Chat-Int4), which is a 14B parameter model based on Transformer.

1.1 创建 modelscope_env

```
1 conda create -n modelscope_env python=3.8.5
```

1.2 启动 modelscope_env

```
1 source activate modelscope_env
```

1.3 安装 modelscope 包

```
1 pip install modelscope
```

1.4 修改模型下载文件

将下载文件中的目录修改为你自己服务器的目录：

```
1 from modelscope.hub.snapshot_download import snapshot_download
2 model_dir = snapshot_download('qwen/Qwen-14B-Chat', cache_dir='你自己的目录',
3                                revision='v1.0.8')
```

1.5 执行模型下载文件

```
1 python qwen_14B_chat_download.py
```

```
2023-11-17 21:36:26,317 - modelscope - INFO - Loading ast index from /root/.cache/modelscope/ast_indexer
2023-11-17 21:36:26,317 - modelscope - INFO - No valid ast index found from /root/.cache/modelscope/ast_indexer, generating ast index from prebuilt!
2023-11-17 21:36:26,367 - modelscope - INFO - Loading done! Current index file version is 1.9.5, with md5 7c8956d2efaf4b7a45da16f1198e5 and a total number of 945 components indexed
2023-11-17 21:36:26,654 - modelscope - INFO - Use user-specified model revision: v1.0.8
Downloading: 100%|██████████| 8.21k/8.21k [00:00<00:00, 875kB/s]
Downloading: 100%|██████████| 50.8k/50.8k [00:00<00:00, 1.57MB/s]
Downloading: 100%|██████████| 910/910 [00:00<00:00, 812kB/s]
Downloading: 100%|██████████| 77.0/77.0 [00:00<00:00, 86.3kB/s]
Downloading: 100%|██████████| 2.29k/2.29k [00:00<00:00, 2.37MB/s]
Downloading: 100%|██████████| 1.88k/1.88k [00:00<00:00, 2.26MB/s]
Downloading: 100%|██████████| 249/249 [00:00<00:00, 288kB/s]
Downloading: 100%|██████████| 6.73k/6.73k [00:00<00:00, 628kB/s]
Downloading: 100%|██████████| 1.91G/1.91G [00:21<00:00, 93.7MB/s]
Downloading: 100%|██████████| 1.89G/1.89G [00:21<00:00, 96.0MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:18<00:00, 100MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:17<00:00, 106MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:21<00:00, 89.2MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:17<00:00, 111MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:19<00:00, 96.6MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:18<00:00, 105MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:18<00:00, 102MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:17<00:00, 105MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:17<00:00, 106MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:18<00:00, 102MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:16<00:00, 112MB/s]
Downloading: 100%|██████████| 1.76G/1.76G [00:12<00:00, 146MB/s]
Downloading: 100%|██████████| 1.45G/1.45G [00:14<00:00, 107MB/s]
Downloading: 100%|██████████| 23.8k/23.8k [00:00<00:00, 1.71MB/s]
Downloading: 100%|██████████| 54.5k/54.5k [00:00<00:00, 1.95MB/s]
Downloading: 100%|██████████| 2.64k/2.64k [00:00<00:00, 2.55MB/s]
Downloading: 100%|██████████| 2.44M/2.44M [00:00<00:00, 15.7MB/s]
Downloading: 100%|██████████| 14.3k/14.3k [00:00<00:00, 905kB/s]
Downloading: 100%|██████████| 32.4k/32.4k [00:00<00:00, 2.42MB/s]
Downloading: 100%|██████████| 9.39k/9.39k [00:00<00:00, 7.87MB/s]
Downloading: 100%|██████████| 173/173 [00:00<00:00, 92.4kB/s]
```

2. FastChat环境配置与启动

2.1 创建 fastchat_env

```
1 conda create -n fastchat_env python=3.8.5
```

2.2 启动 fastchat_env

```
1 source activate fastchat_env
```

2.3 安装FastChat包

```
1 bash fastchat_install.sh
```

2.4 安装Qwen依赖

安装01_qwen_server中requirements.txt对应的依赖：

```
1 pip install -r requirements.txt
```

2.5 启动Controller

Controller启动后会占用21001端口：

```
1 nohup bash controller_start.sh > cl_20231118.log &
```

```
(qwen_agent_env) root@autodl-container-faba4ab153-941bc3b5:~/autodl-tmp/jiangxia/01_qwen_server# lsof -i:21001
COMMAND PID USER FD TYPE DEVICE SIZE/OFF NODE NAME
python 2659 root 8u IPv4 6816478      0t0 TCP *:21001 (LISTEN)
```

2.6 修改Worker目录

将worker_start.sh中的模型目录修改为自己的目录：

```
1 python -m fastchat.serve.model_worker --model-path 你的模型目录 --host 0.0.0.0
```

2.7 启动Worker

Controller启动后会占用21002端口：

```
1 nohup bash worker_start.sh > wk_20231118.log &
```

```
(qwen_agent_env) root@autodl-container-faba4ab153-941bc3b5:~/autodl-tmp/jiangxia/01_qwen_server# lsof -i:21002
COMMAND PID USER FD TYPE DEVICE SIZE/OFF NODE NAME
python 3157 root 38u IPv4 6824581      0t0  TCP *:21002 (LISTEN)
```

2.8 启动 OpenAI 接口支持

OpenAI 接口支持服务启动后会占用8000端口：

```
1 nohup bash openai_support.sh > oa_20231118.log &
```

```
(qwen_agent_env) root@autodl-container-faba4ab153-941bc3b5:~/autodl-tmp/jiangxia/01_qwen_server# lsof -i:8000
COMMAND PID USER FD TYPE DEVICE SIZE/OFF NODE NAME
python 3219 root 7u IPv4 6761284      0t0  TCP *:8000 (LISTEN)
```

2.9 在本地机器建立端口映射

```
1 ssh -CNg -L 8000:127.0.0.1:8000 root@region-3.seetacloud.com -p 47819
```