conda create -n llama3 python=3.10

conda activate llama3

conda install pytorch==2.1.2 torchvision==0.16.2 torchaudio==2.1.2 pytorch-cuda=12.1 -c pytorch -c nvidia

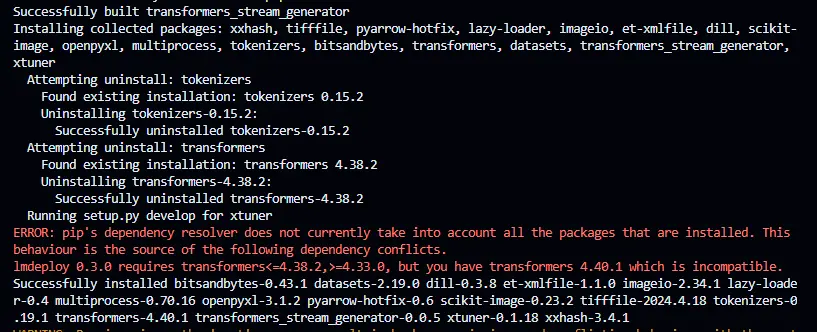
conda activate agent

cd ~/xtuner/

git clone -b v0.1.18 https://github.com/InternLM/XTuner

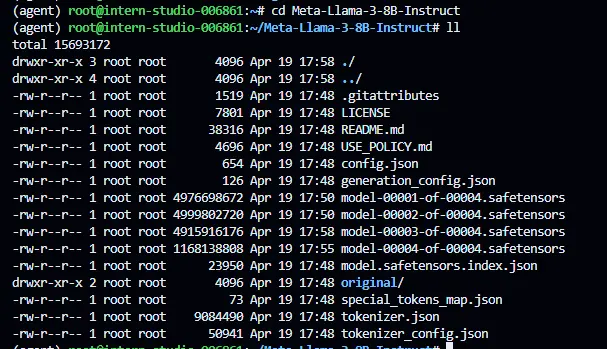
cd XTuner

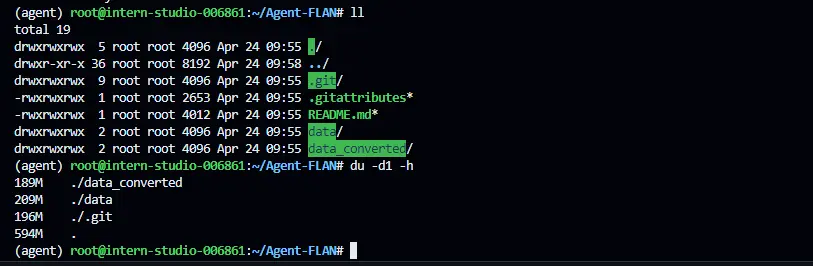
pip install -e .



cd ~

ln -s /root/share/new\_models/meta-llama/Meta-Llama-3-8B-Instruct /root/models/Meta-Llama-3-8B-Instruct





# pretrained\_model\_name\_or\_path = '~/Meta-Llama-3-8B-Instruct'

pretrained\_model\_name\_or\_path = '/root/share/new\_models/meta-llama/Meta-Llama-3-8B-Instruct'

use\_varlen\_attn = True

# Data

agent\_flan\_path = '~/Agent-FLAN/data\_converted'

prompt\_template = PROMPT\_TEMPLATE.llama3\_chat

max\_length = 8192

pack\_to\_max\_length = True

# parallel

sequence\_parallel\_size = 1

# Scheduler & Optimizer

batch\_size = 1 # per\_device

accumulative\_counts = 8

accumulative\_counts \*= sequence\_parallel\_size

dataloader\_num\_workers = 0

max\_epochs = 1

model = dict(

type=SupervisedFinetune,

use\_varlen\_attn=use\_varlen\_attn,

llm=dict(

type=AutoModelForCausalLM.from\_pretrained,

pretrained\_model\_name\_or\_path=pretrained\_model\_name\_or\_path,

trust\_remote\_code=True,

torch\_dtype=torch.float16,

quantization\_config=dict(

type=BitsAndBytesConfig,

load\_in\_4bit=True,

load\_in\_8bit=False,

llm\_int8\_threshold=6.0,

llm\_int8\_has\_fp16\_weight=False,

bnb\_4bit\_compute\_dtype=torch.float16,

bnb\_4bit\_use\_double\_quant=True,

bnb\_4bit\_quant\_type='nf4')),

lora=dict(

type=LoraConfig,

r=32,

lora\_alpha=64,

lora\_dropout=0.1,

bias='none',

task\_type='CAUSAL\_LM'))

MAX\_JOBS=4 pip install flash-attn

pip install deepspeed

export MKL\_SERVICE\_FORCE\_INTEL=1

export MKL\_THREADING\_LAYER=GNU

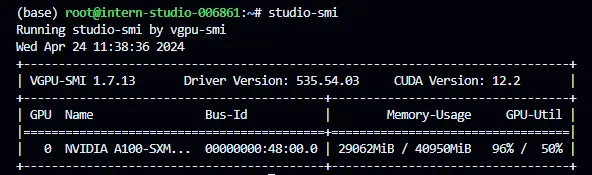
微调指令如下

NPROC\_PER\_NODE=1 xtuner train \

~/Llama3-Tutorial/configs/llama3-agentflan/llama3\_8b\_instruct\_qlora\_agentflan\_3e.py \

--deepspeed deepspeed\_zero2 \

--work-dir ~/project/llama3-ft/agent-flan



# 转换权重

xtuner convert pth\_to\_hf \

~/Llama3-Tutorial/configs/llama3-agentflan/llama3\_8b\_instruct\_qlora\_agentflan\_3e.py \

~/project/llama3-ft/agent-flan/iter\_500.pth \

~/project/llama3-ft/agent-flan/iter\_500\_hf

# 合并权重

export MKL\_THREADING\_LAYER=GNU

export MKL\_SERVICE\_FORCE\_INTEL=1

xtuner convert merge ~/Meta-Llama-3-8B-Instruct \

~/project/llama3-ft/agent-flan/iter\_500\_hf \

~/project/llama3-ft/agent-flan/merged\_500

