



GHDDI

全球健康 药物研发中心
Global Health Drug Discovery Institute

scGPT – user guide

Qi Liu

Data Science Group, GHDDI

qi.liu@ghddi.org

2023.08.29

Get Start

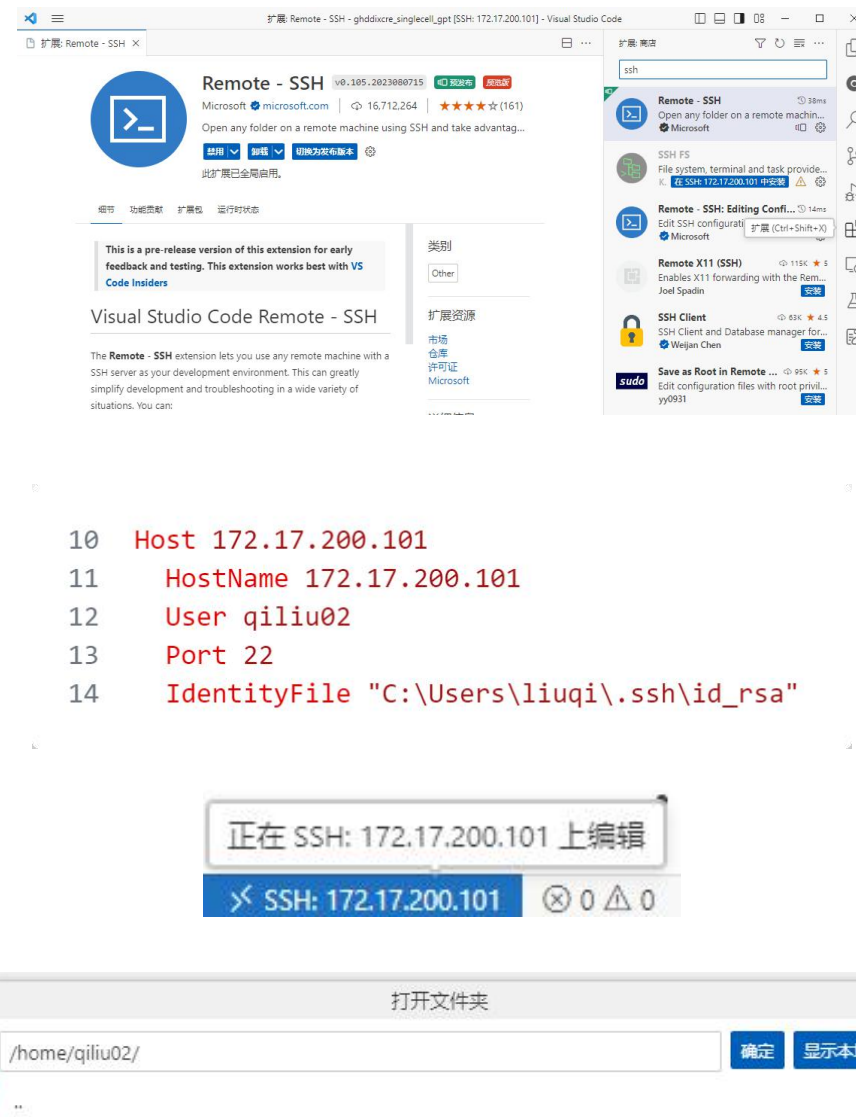
先下载VS Code，远程连接cluster，步骤如下：

1. 下载VS Code: <https://code.visualstudio.com/>
2. 打开VS Code, 点击侧边栏的扩展
 - 搜索框中输入ssh
 - 下载Remote SSH，并配置服务器[2]
 - .ssh/config内容参考右图
3. 在VS Code中连接远程服务器
 - 点击侧边栏的"远程资源管理器"
 - 连接刚才配置好的服务器，选择在当前/新的窗口打开
 - 左下角显示已经连接后，选择侧边栏"打开文件夹"，打开cluster中想要创建项目的路径，或者已经存在的路径后续就可以直接在cluster上进行开发了

本节参考：

[1] [VS Code](<https://code.visualstudio.com/>)

[2] [Remote Development using SSH](https://code.visualstudio.com/docs/remote/ssh#_getting-started)



Conda env

安装好VS Code和远程连接好Cluster后，下面安装conda，并激活scGPT的Conda env。假如当前个人账号没有安装Conda的话，建议先安装[1]，如使用下面命令：

```
wget https://repo.anaconda.com/archive/Anaconda3-2023.07-2-Linux-x86_64.sh
bash ./Anaconda3-2023.07-2-Linux-x86_64.sh
```

101服务器上已经安装好了用于使用scGPT的环境, 可选Chen Liang (梁忱) 提供的环境：

- Python解释器： /home/cliang02/work/bin/cre-python
- Conda环境： /home/cliang02/work/software/common/proglang/mambaforge/envs/cre

或者Qi Liu (刘奇) 提供的环境：

- Conda环境： /home/qiliu02/miniconda3/envs/single_cell_gpt

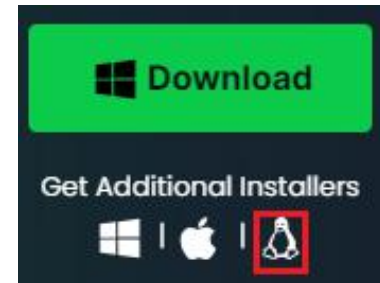
新建Terminal后，可以输入命令激活：

```
conda activate /home/cliang02/work/software/common/proglang/mam
# or
```

```
conda activate /home/qiliu02/miniconda3/envs/single_cell_gpt
```

就切换到了我们运行scGPT的环境了。

注意：conda env中没有安装scgpt package，而是在项目中源码安装scGPT，这是为了方便修改源码。后面的预处理脚本用sys.path.append的方式引用了源码包，若有相关问题，可以联系Qi Liu。



本节参考：

[1] [Anaconda Download](https://www.anaconda.com/download/)

```
(base) [qiliu02@comput101 ~]$ conda activate /home/cliang02/work/software/common/proglang/mambaforge/envs/cre
(cre) [qiliu02@comput101 ~]$ python --version
Python 3.10.11
(cre) [qiliu02@comput101 ~]$ conda activate /home/qiliu02/miniconda3/envs/single_cell_gpt
(single_cell_gpt) [qiliu02@comput101 ~]$ python --version
Python 3.7.13
```

Prepare data

使用我们数据预处理脚本来将.h5ad数据处理成scGPT pretraining需要的input embedding数据集。

首先安装git [1], 终端输入`git --version`测试, 已安装的话跳过. 使用git clone, 将预处理脚本clone到自己的路径:

`git clone https://github.com/qiliu-ghddi/singlecell_gpt`

clone后, 进入到`data`路径下, 将我们的要处理的数据放 (或者软连接) 到`data/raw`,

运行``build_large_scale_data.py``, 给定`.h5ad`文件的路径, 将其用scgpt.scbank处理为便于大规模数据处理的格式;

接着``binning_mask_allcounts.py``, 将上一步的输出, 处理为能够作为pretraining scGPT的embedding数据集。

```
cd data
```

```
conda activate <env> # 参考教程里的激活conda env, 激活运行我们的scgpt环境
```

```
python build_large_scale_data.py --input-dir "raw/" --output-dir "./preprocessed"
```

```
python binning_mask_allcounts.py --data_source "./preprocessed/all_counts/"
```

本节参考:

[1] [Git Download](<https://git-scm.com/downloads>)

[2] [scGPT](<https://github.com/bowang-lab/scGPT>)



Pretrain scGPT

我们conda env环境中没有直接安装scGPT[1]，下面源码安装scGPT，注册wandb[2]，随后预训练scGPT.

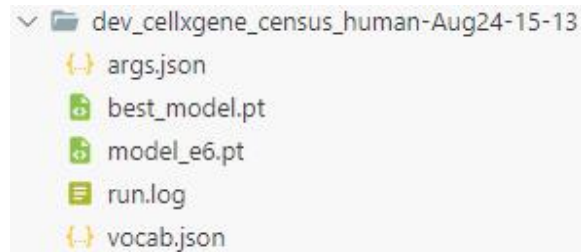
clone scGPT源，将其中的`scGPT/scgpt`源码放在我们的项目根路径下（和code, data平齐）.

由于scGPT使用了wandb作为日志记录的工具，所以建议去wandb上注册一个账号，激活后，新建终端，输入`wandb login`，输入账号和密码即可，只需要配置一次，后续就不用管了.

进入到`code`路径下，运行`python pretrain_scGPT.py`，将`data_source`配置为我们前面处理得到的`binned`数据集的路径，即可开始训练. 更多参数，请输入`python pretrain_scGPT.py --help`查看.

训练完后，在`save/xxx-<datetime>`路径下，会保存`checkpoint`即`best_model`，及参数`args.json`和`vocab.json`文件，可以作为后面`finetune`环节的输入.

```
cd code
python pretrain_scGPT.py --data_source "../data/binned/"
```



dev_cellxgene_census_human-Aug24-15-13

- args.json
- best_model.pt
- model_e6.pt
- run.log
- vocab.json

本节参考：

[1] [scGPT](<https://github.com/bowang-lab/scGPT>)

[2] [Wandb](<https://wandb.ai/>) and [Wandb login](<https://docs.wandb.ai/ref/cli/wandb-login>)

Down-stream tasks and evaluation

scGPT 提供了一些下游finetune任务的jupyter notebooks [1]，将我们训练得到的结果，作为参数，可以运行这些下游任务。

在`examples`中提供了`finetune_integration`作为例子，修改脚本中的`load_model`的值，为我们上面训练得到的`checkpoint`所在的路径，即可运行. 其会下载PBMC 10K`数据，并用我们得到的模型进行`batch integration`的finetuning，最后将结果保存到`wandb`的`run`中。

```
hyperparameter_defaults = dict(
    seed=42,
    dataset_name="PBMC_10K",
    do_train=True,
    load_model="./save/dev_databanks_sample-Aug15-18-16",
    mask_ratio=0.4,
    epochs=30,
    n_bins=51,
    GEPC=True, # Masked value prediction for cell embedding
    ecs_thres=0.8, # Elastic cell similarity objective, 0.0 to 1.0, 0.0 to disable
    dab_weight=1.0,
    lr=1e-4,
    batch_size=64,
    layer_size=128,
    nlayers=4,
    nhead=4,
    # if load model, batch_size, layer_size, nlayers, nhead will be ignored
    dropout=0.2,
    schedule_ratio=0.9, # ratio of epochs for learning rate schedule
    save_eval_interval=5,
    log_interval=100,
    fast_transformer=True,
    pre_norm=False,
    amp=True, # Automatic Mixed Precision
)
run = wandb.init(
    config=hyperparameter_defaults,
    project="finetune-dev_databanks_sample-Aug15-18-16",
    reinit=True,
    settings=wandb.Settings(start_method="fork"),
)
```

本节参考：

[1] [scGPT tutorials](<https://github.com/bowang-lab/scGPT/tree/main/tutorials>)