offCampusInstructions

February 7, 2024

1 Using Llama2 or Mixtral off campus through UNCC's VPN

We have installed a Llama-70B quantized model on 6 V100 NVIDIA GPUs, which are part of a CCI server. The SIS department has also installed a Mixtral model on a different HPC server. There are two ways of accessing these models using the OpenAI chat completion API: 1. Use a computer on campus that is connected to internet through Eduroam (not NinerWifi-Guest). 2. Use a computer off campus running the "Cisco Secure Client" VPN for UNCC.

If you are on campus (option 1), you do not need to do anything extra.

If you are off campus (option 2), you will need to use the Educational Cluster through VPN, as explained below.

1.1 Setting up a Jupyter notebook server on Centaurus

1.1.1 1. Connect to the Centaurus server through `ssh`

Open a Terminal (command line) window and type the command below, using your UNCC user name:

```
ssh <username>@hpc-student.uncc.edu
```

This will ask for your password, and dual factor authentication.

1.1.2 2. Copy the `examples` folder in your home directory

```
cp -r /projects/class/itcs4111_001/hw03 ~/
```

1.1.3 3. Change current directory to the `hw03` folder

```
cd ~/hw03
```

1.1.4 4. Submit the batch job that will start the Jupyter notebook server

```
sbatch jn-server.slurm
```

This will print a message showing the job number, such as "Submitted batch job xxxxxx"

1.1.5 5. Find the `ssh` command in the log file

`grep ssh 4111-jn-server-xxxxxx.log` (replace xxxxxx with the actual job number you saw at the previous step)

This will output a complete ssh command, such as **`ssh -N -L 8234:gal-c1:8234 <username>@hpc-student.uncc.edu`**

1.1.6 6. Find the `http` address for the notebook server in the log file

`grep 127 4111-jn-server-100748.log` (replace 100748 with the actual job number you see at the previous step)

This will output a web address, such as **`http://127.0.0.1:8234/?token=18446ac637ff267e18c54dd34558bffa0`**

1.2 Connecting to the notebook server and opening the notebook

Open another Terminal (command line) window and connect to the notebook server using the complet ssh command found above, for example:

```
ssh -N -L 8234:gal-c1:8234 <username>@hpc-student.uncc.edu
```

Open a new window or tab in your favorite browser and paste the http address found above, for example:

```
http://127.0.0.1:8234/?token=18446ac637ff267e18c54dd34558bffa09aff29dc3f0e62c
```

This will show the contents of the `llm` folder in your account on the educational cluster. Click on `llama` or `mixtral`, then click on the Jupyter notebook file. This will open a new tab showing the code in the file, that you can edit.

1.3 Edit code in the notebook, then download a local copy

Once the notebook is opened in the browser, edit and evaluate the code acordingly, saving from time to time. This will update the notebook in your account on the HPC server.

Note that the notebook server is set to automatically close 2 hours after it was started. If that happens and you need more time, you will have to redo the steps above starting from step 4 (your code edits will be preserved in the notebook file). Note that if you navigated out of the `~/hw03` directory on Centaurus, you would have to navigate back into the directory before starting again. If you disconnected from Centaurus, you would repeat all instructions except for step 2 (you already copied these files).

When you are done, download a local copy by clicking File -> Download as => Notebook (.ipynb). It is this local copy that you will submit on Canvas. You can follow similar steps to download a PDF version that you submit on Canvas.

2 When you are finished working

What happens if you finish working before your Jupyter Notebook server automatically closes? The server will remain open until the 2 hours are up, unless the job is cancelled. However, unlike a regular Jupyter Notebook server, even shutting down the server will not end the job and free the resources. Other students are trying to use these resources as well, and courtesy is important when utilizing shared computing resources.

The command to cancel a job, and save your classmates some stress of waiting on resources, is fairly simple. For example:

`scancel xxxxxx` (replace xxxxxx with the job number you saw in the previous steps)

If you forget your job number, you can use

`squeue -u <username>` (replace "<username>" with your own username)

to find the job number of any jobs that you have open.

Thank you in advance for your courtesy.