GUIDE: Running Models on Gcloud

CMPT419: PROJECT19

1 Renting a GPU

While there are multiple websites available where you can rent NVIDIA GPUs. It is suggested you use Google Cloud Console since it offers $\approx$ \$400 worth of free credits (this is more than enough for our usage). Also note that before beggining, you should install Google CLI on your local computer. Refer to installation instructions.

1.1 Creating Account and increasing Quota

- Go to GCP's homepage and select- 'Get Started for Free'.

- Follow the signup instructions and you should receive free credits automatically.

- Go to Google Cloud Console and start a project (you might also have to setup billing information to keep your free credits from getting expired).

- Head to Quotas page on GCP and select 'Compute Engine API'

- Under the 'Compute Engine API (Networks)', increase the limit for 'gpus(all region)' and 'L4' (since we are using 'L4' in our case) gpus for your preferred locations. For example, the increase requests may look as given in Figure1 1

1.2 Creating a VM

- Head to Google Cloud Console and select 'Compute Engine' after logging in.

- Click 'Create Instance' from the top menu.

- Name the instance and select a region.

- Under the 'Machine Configuration', select 'GPUs' and choose one ' NVIDIA L4' GPU.

- Leaving other settings as they are, under the 'Boot Disk' menu, select 'Change' and modify the OS to 'Ubuntu' with version 'x86/64 22.04 LTS'. Lastly, change the size of the disk to 100 GB for now. You may increase the it later.

- After all of the above steps, select 'Create'. If the GPU is available in the region, the google cloud platform will create it in a few minutes, ready to use.

Note, if you see an error mentioning 'GPUs unavailable in selected region', redo the whole process with different region selected.
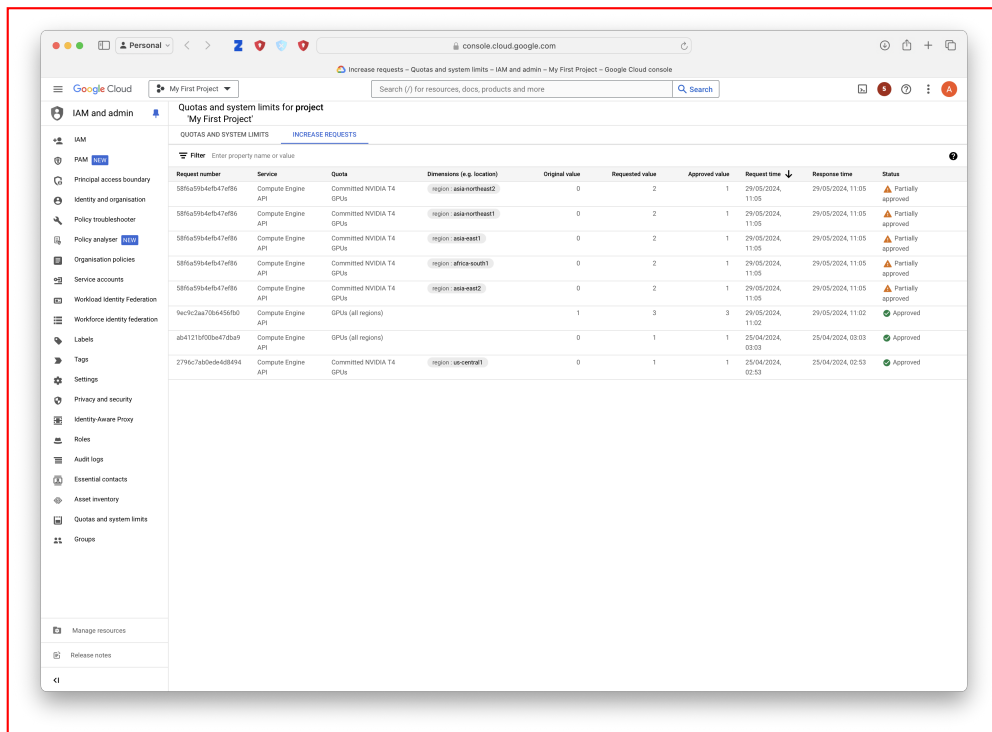
Figure 1: Increase Requests on GCP

1.3 Connecting to VM

There are multiple ways to connect to the created VM. It is recommended you use your own local terminal and ssh into it since connection through other methods expire after inactivity.

For **SSH** connection,

- Click the dropdown arrow under 'Connect' and from the available options, select 'View gcloud command'
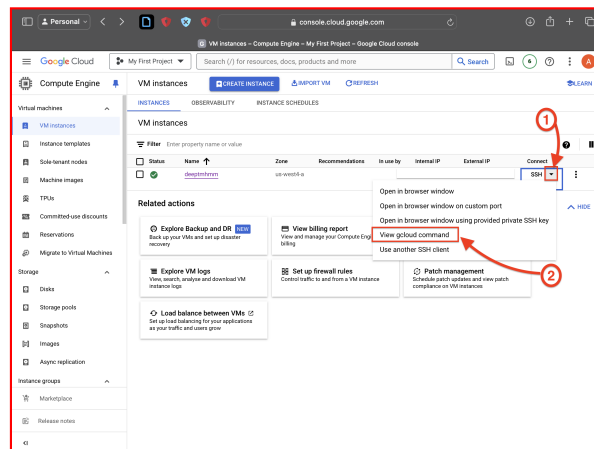

Figure 2: Selecting gcloud Command

- Copy the displayed command and enter it in your **LOCAL** terminal.

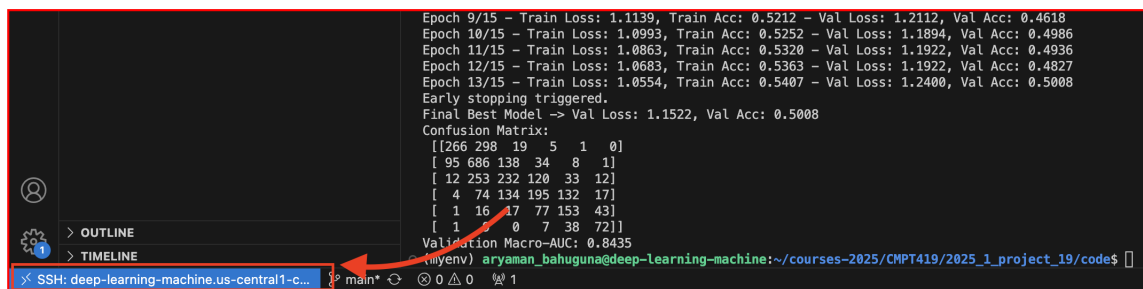You should now have successfully established a connection with the VM through your terminal.

1.4 Using VM w/ Remote SSH on Vscode

If you are using VScode, you can also connect to the VM using the Remote SSH extension. This allows you to use the GUI of VScode to run commands on the VM.

To do so, follow the steps below-

- Install the Remote SSH extension on VScode.

- Open the command palette (Ctrl + Shift + P) and select 'Remote-SSH: Connect to Host...'

- Paste the gcloud command you copied earlier in the input box and hit enter.

- VScode will now connect to the VM and open a new window with the remote file system.

- You can now use the terminal in VScode to run commands on the VM.

See the figure below for verifying the connection-

Figure 3: VScode Connection

2 Installing GPU Drivers

After setting up the VM, GPU drivers must be installed. Refer Compute Engine Documentation-Install GPU Drivers. (Before beggining, ensure you are ssh'd into the VM)

3 TEST: Running Models on Gcloud VM

You may now clone the repository from Github-Repo-Link and run our results on the VM. The datasets link is: Google Drive Link.