**CSCIGA.3033 HW 3**

Chenmeinian Guo

The experiment tested 3 models’ performance on cloud (NYU cluster) and baremetal (my own PC with GPU/CUDA and with CPU), respectively.

| Model/Environment  ->Time | Baremetal (My own PC using GPU) | Cloud (NYU Cluster) |
| --- | --- | --- |
| Alexnet | 4.985 | 1.15 |
| Resnet50 | 4.362 | 0.447 |
| Vgg16 | 4.041 | 0.588 |

The objective of this project is to conduct a comparative experiment across different environments and with different neural network models. The hypothesis is that NYU cluster environments should have better performance (less training time) in comparison with my own PC environment since the machines in the computing labs are equipped with modern GPUs and are designed for high computing tasks.

The pytorch code I used is based on the github repo mentioned in the lecture PowerPoint (<https://github.com/pytorch/examples/tree/main/imagenet>), with minor tweaks (for example, added code to print out the models’ architecture).

Screenshots (white->my own PC; black->NYU cluster; screenshot of the model’s architecture):

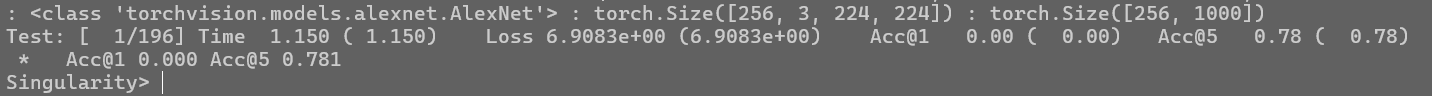
For convenience, I passed the dummy data when testing the models since the accuracy does not matter (as mentioned in the requirements).

The results indicate a significant disparity in performance, favoring the NYU cluster environment over the baremetal setup for all tested models. The NYU cluster's superior performance can be attributed to several factors. First, the machines in NYU cluster are equipped with specialized hardware including more advanced GPU. Second, the cluster also offers optimized software stacks designed to leverage the underlying hardware more efficiently.

See below screenshots for proof and more information:

Alexnet



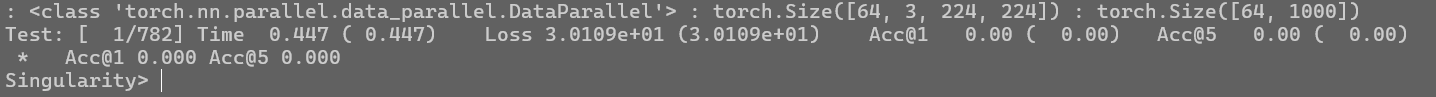


表格

描述已自动生成

Resnet50





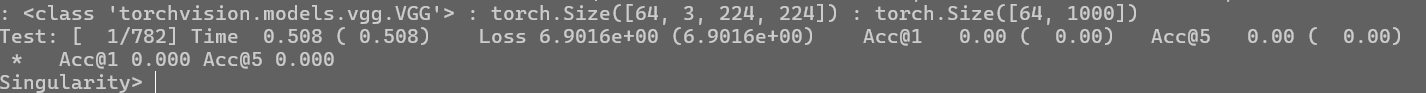
……………..

表格

描述已自动生成

Vgg16





表格

描述已自动生成

Roofline Screenshot (for resnet50)

文本

中度可信度描述已自动生成

文本

描述已自动生成

图片包含 图形用户界面

描述已自动生成

For the Roofline model, I utilized NVIDIA’s ‘nvprof’ tool to profile my training session and capture the actual FLOPS and memory bandwidth utilization.

Here’s the command I used to generate above results shown in the screenshot:

`**nvprof --metrics flops\_sp\_efficiency,achieved\_occupancy,sm\_efficiency, gld\_efficiency, gst\_efficiency python proj\_1.py -a resnet50 –dummy --batch-size 64**`

For simplicity, I’ve only included the screenshot of resnet50 result on NYU cluster as the command/procedures to achieve this on other models/environments are the same. After running the command above, two files were generated (report.nsys-rep and report.sqlite) for the Roofline model analysis.

With the NVIDIA Nsight System installed, we could then open the .nsys-rep file to view the Roofline model report. (see below screenshot)

应用程序, 表格

描述已自动生成

Overall, the study emphasizes the importance of the environment selection in deep learning tasks. The difference of environment (both hardware and software) could drastically affect the efficiency of deep learning model training and experimentation. To enhance the experiment, future works could try to find how closely the actual performance of each model comes to the theoretical limits of the hardware.