Homework 2

Zhengxiang Huang

520021910014

Contents

1 Introduction

In this Homework, you shall yourself explore more on how to implement quantization, and deploy LLM on Nano with the multiple online resources.
We sincerely hope you enjoy your journey through this homework as we do. Have fun!

2 Q1: Quantization in PyTorch (85')

2.1 Problem Formulation

We've discussed quantization in class.
Now, try to quantize your own network in Homework 1 Q5 on your own laptop computer.
You shall then be able to answer the following questions:

1. (5') Q: Screenshot the results of "print(qconfig)". Explain the parameters of it. e.g., explain what is "per_channel_symmetric".

2. (10') Q: Implement "fuse_model" function for your model, explain how "Conv2d", "BatchNorm2d", and "ReLU" are fused. Screenshot the results of "print(model)" after you fuse it.

3. (10') Q: Calibrate and quantize your model in PyTorch. Screenshot the results of "print(model)" after you quantize it. Explain why we need calibration phase, and what we are observing during that phase.

4. (20') Q: Measure the size, inference accuracy, and inference time of your model after quantization. Compare the size, time, and accuracy before and after quantization.

5. (40') Q: Submit your quantization code, whether it's runnable or not. (Even it can't work, we attribute some points to you.)

6. (Bonus +15') Q: Quantized model can't be directly deployed on nano, we need onnx workaround. We have a line of "torch.onnx.export" in our "example.py". Search online how to install and run the "quant.onnx" you export, and deploy that stuff onto your Nano board with onnx. If onnx is impossible on nano, write in the report to inform me, and I will attributes bouns to you too. (No hints nor guides, search yourself! hahahaha!)

(**Hint**: It's suggested that you install a CPU version of Pytorch on your own computer if you don't have a powerful Nvidia GPU installed.
"pip install torch==2.0.1 torchvision==0.15.2 –index-url https://download.pytorch.org/whl/cpu" to install a cpu version pytorch on your computer, if you previously have no pytorch installed.)
(**Hint**: Refer to https://pytorch.org/docs/stable/quantization.html#post-training-static-quantization and https://pytorch.org/tutorials/advanced/static_quantization_tutorial.html for more information of how to quantize a model in PyTorch. Also, see our "example.py" and "model.py" provided.)

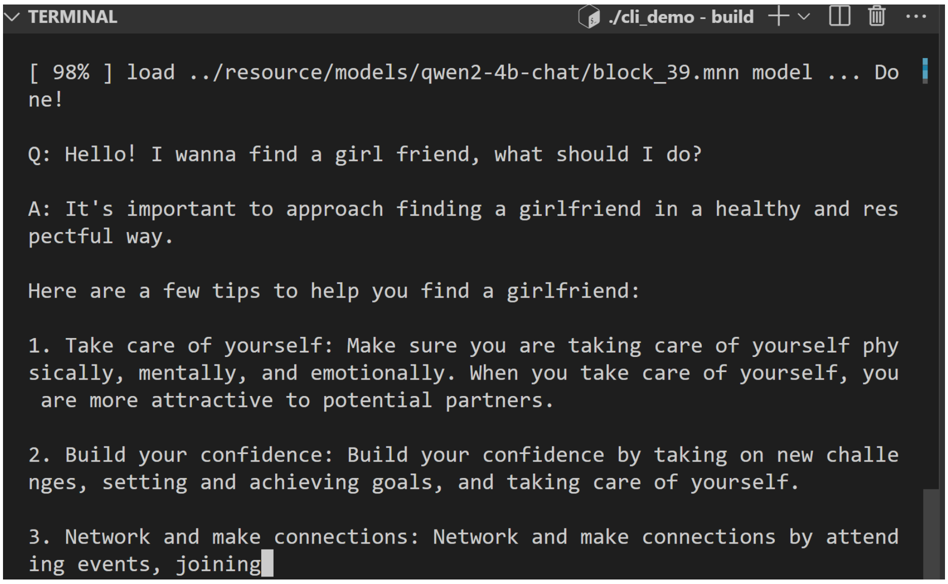
2.2 Solution

(Your answers go here!)

3 Q2: Advanced Topic, LLM on Nano (15')

3.1 Problem Formulation

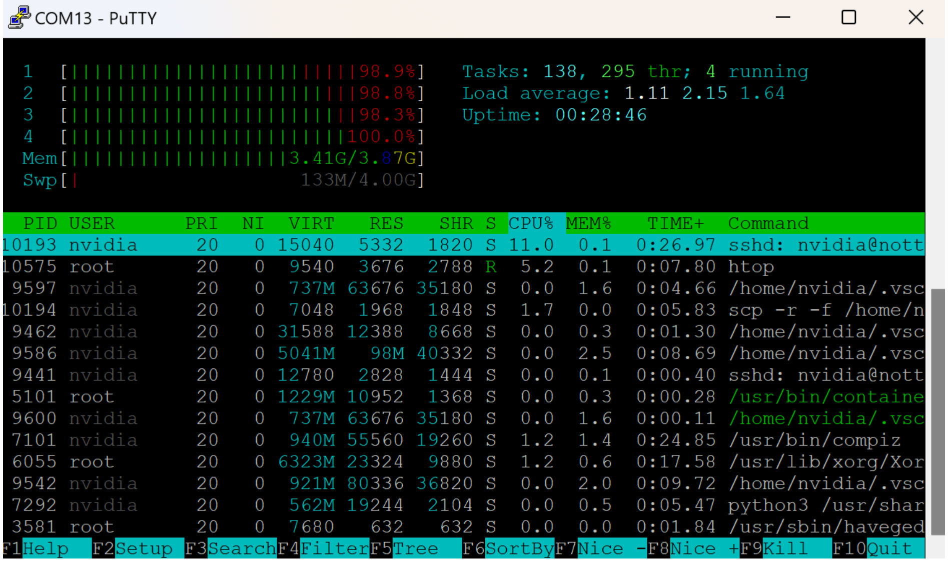
1. (12') Follow our instructions in LLM-README.md to deploy a LLM (Qwen1.5-4B) on Nano. Measure time and space it consumes. Have fun chatting with it! Screenshot your conversions.



```
[ 98% ] load ../resource/models/qwen2-4b-chat/block_39.mnn model ... Do
ne!

Q: Hello! I wanna find a girl friend, what should I do?

A: It's important to approach finding a girlfriend in a healthy and res
pectful way.

Here are a few tips to help you find a girlfriend:

1. Take care of yourself: Make sure you are taking care of yourself phy
sically, mentally, and emotionally. When you take care of yourself, you
 are more attractive to potential partners.

2. Build your confidence: Build your confidence by taking on new challe
nges, setting and achieving goals, and taking care of yourself.

3. Network and make connections: Network and make connections by attend
ing events, joining
```

Figure 1: Nano Chat Bot

The resource LLM takes up.



```
  1  [|||||||||||||||||||||||98.9%]   Tasks: 138, 295 thr; 4 running
  2  [|||||||||||||||||||||||98.8%]   Load average: 1.11 2.15 1.64
  3  [|||||||||||||||||||||||98.3%]   Uptime: 00:28:46
  4  [||||||||||||||||||||||100.0%]
  Mem[|||||||||||||||||3.41G/3.87G]
  Swp[|                133M/4.00G]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
10193 nvidia     20   0 15040  5332  1820 S 11.0  0.1  0:26.97 sshd: nvidia@nott
10575 root       20   0  9540  3676  2788 R  5.2  0.1  0:07.80 htop
 9597 nvidia     20   0  737M 63676 35180 S  0.0  1.6  0:04.66 /home/nvidia/.vsc
10194 nvidia     20   0  7048  1968  1848 S  1.7  0.0  0:05.83 scp -r -f /home/n
 9462 nvidia     20   0 31588 12388  8668 S  0.0  0.3  0:01.30 /home/nvidia/.vsc
 9586 nvidia     20   0 5041M   98M 40332 S  0.0  2.5  0:08.69 /home/nvidia/.vsc
 9441 nvidia     20   0 12780  2828  1444 S  0.0  0.1  0:00.40 sshd: nvidia@nott
 5101 root       20   0 1229M 10952  1368 S  0.0  0.3  0:00.28 /usr/bin/containe
 9600 nvidia     20   0  737M 63676 35180 S  0.0  1.6  0:00.11 /home/nvidia/.vsc
 7101 nvidia     20   0  940M 55560 19260 S  1.2  1.4  0:24.85 /usr/bin/compiz
 6055 root       20   0 6323M 23324  9880 S  1.2  0.6  0:17.58 /usr/lib/xorg/Xor
 9542 nvidia     20   0  921M 80336 36820 S  0.0  2.0  0:09.72 /home/nvidia/.vsc
 7292 nvidia     20   0  562M 19244  2104 S  0.0  0.5  0:05.47 python3 /usr/shar
 3581 root       20   0  7680   632   632 S  0.0  0.0  0:01.84 /usr/sbin/haveged
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

Figure 2: LLM resource

(Hint: If you can't deploy it, screenshot the step you get stuck and we will attribute points to you!)

2. (3') Tell me how many bit are this model's weights quantized into? 4 bits? 8 bits? No quantization? Where did you find that info?

3.2 Solution

(Your answer here!)

4 What have you learned?

4.1 Problem Formulation

(Your answer here!)

4.2 Tell me what you have learned

(Your answer here!)

5 Acknowledgement

(Your answer here!)

References