Lab04
Model Compression: Pruning & Quantization

Deep Learning, 2025 Fall

Overview

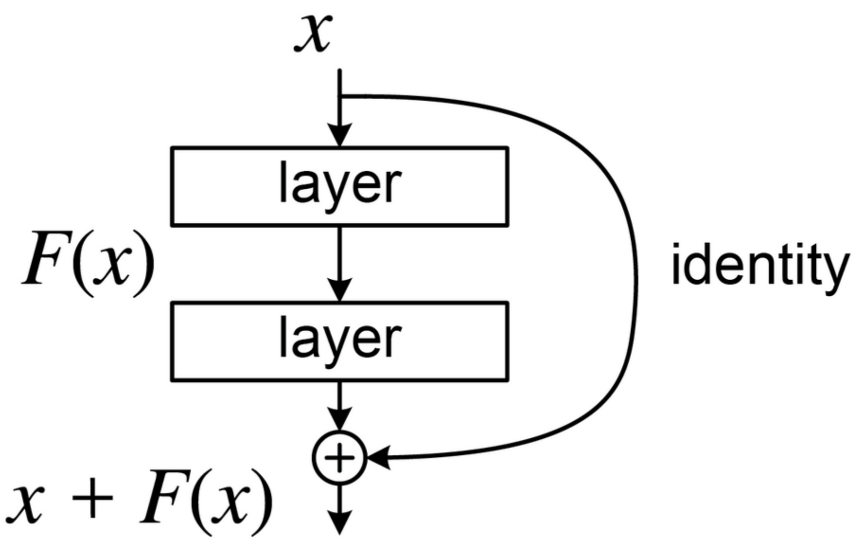
- Dataset: CIFAR10
- Model: ResNet20 (pretrained)
- Tasks:
  - Task 1: unstructured pruning
  - Task 2-1: post-training static quantization using Pytorch's FX graph mode
  - Task 2-2: manual post-training static quantization

Dataset: CIFAR10

- CIFAR-10 is a widely used benchmark dataset in the fields of machine learning and computer vision.
- It contains 60,000 color images of size 32x32 across 10 classes. The dataset is pre-divided into a training set of 50,000 images and a test set of 10,000 images to ensure standardized evaluation.

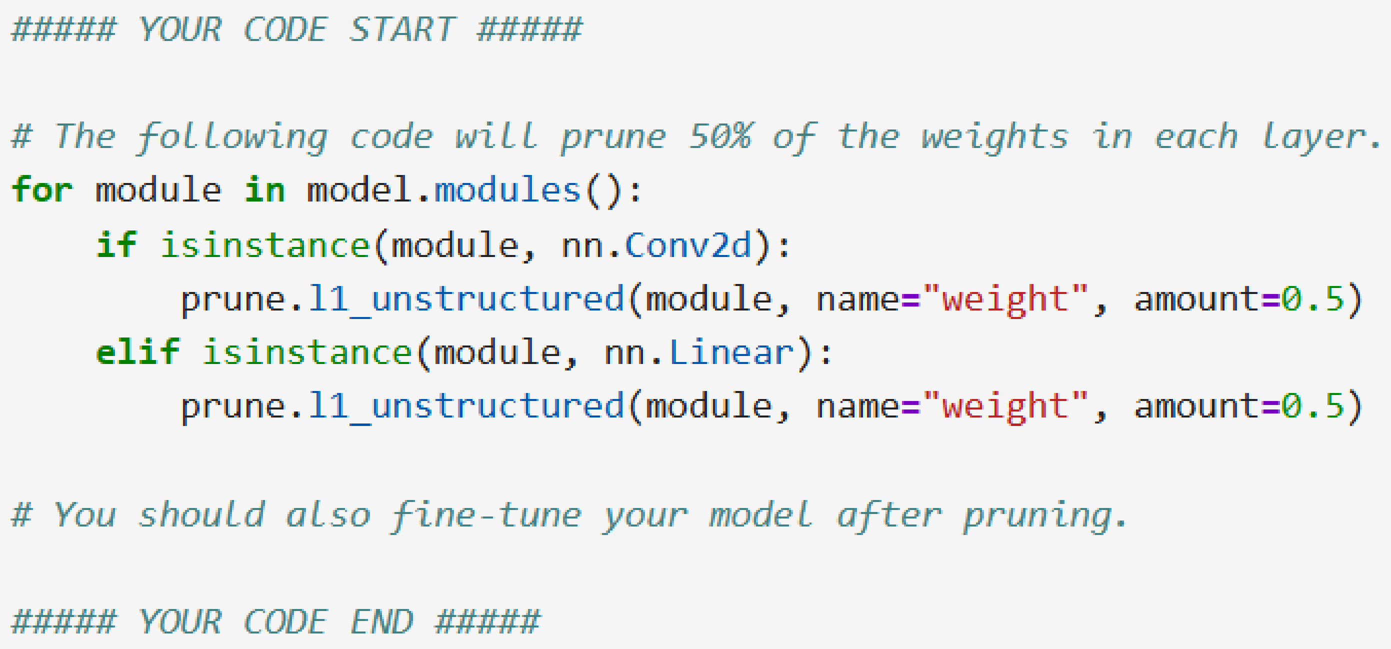
Model: ResNet20

- ResNet is a popular deep learning model for image classification. Its key feature is the use of skip (residual) connections, which make training deep networks easier and more stable.
- Pretrained model from chenyaofo/pytorch-cifar-models with test accuracy of **92.60%**



Task 1: Unstructured Pruning

- task1_pruning.ipynb
- Use unstructured pruning to prune the pretrained model to achieve **>=50% sparsity** while keeping **test accuracy >=90%**.
- Performance ranking based on sparsity.



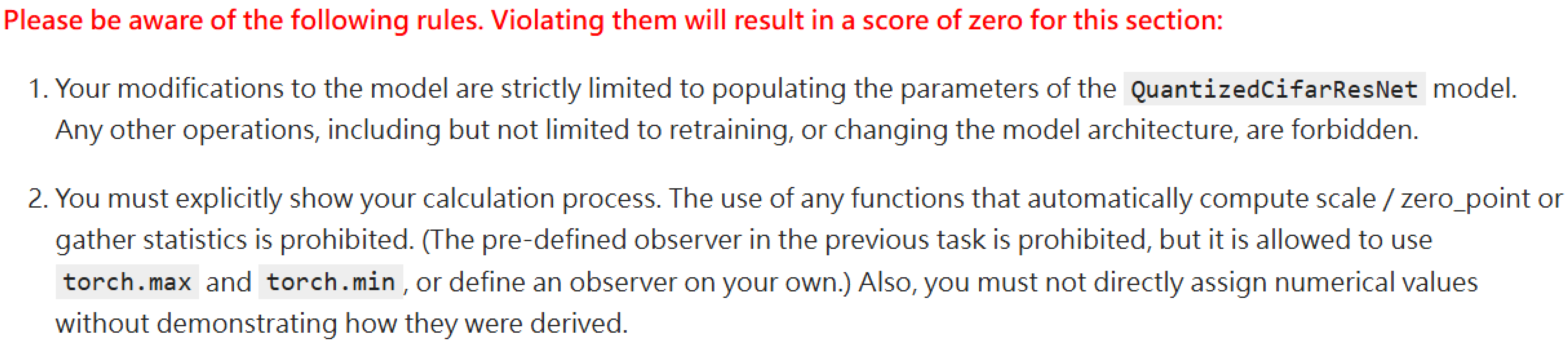
```
##### YOUR CODE START #####

# The following code will prune 50% of the weights in each layer.
for module in model.modules():
    if isinstance(module, nn.Conv2d):
        prune.l1_unstructured(module, name="weight", amount=0.5)
    elif isinstance(module, nn.Linear):
        prune.l1_unstructured(module, name="weight", amount=0.5)

# You should also fine-tune your model after pruning.

##### YOUR CODE END #####
```

Task 2-1: Post-Training Static Quantization Using Pytorch's FX Graph Mode

- task2_1_quantization_api.ipynb
- Pytorch FX graph mode
  - Prepare, calibrate, convert
  - See notebook for more details
- Use at least five different amount of data for calibration and compare the changes in accuracy of the quantized models on the test set, and document your observations in your report.

Task 2-2: Manually Quantizing the Model

- task2_2_quantization_manual.ipynb
- Manually quantize the model.
- The quantized model should have **test accuracy >= 90.0%**
- Performance ranking based on test accuracy.



Please be aware of the following rules. Violating them will result in a score of zero for this section:

1. Your modifications to the model are strictly limited to populating the parameters of the `QuantizedCifarResNet` model. Any other operations, including but not limited to retraining, or changing the model architecture, are forbidden.

2. You must explicitly show your calculation process. The use of any functions that automatically compute scale / zero_point or gather statistics is prohibited. (The pre-defined observer in the previous task is prohibited, but it is allowed to use `torch.max` and `torch.min`, or define an observer on your own.) Also, you must not directly assign numerical values without demonstrating how they were derived.

Report

- Please answer these questions below in your report.
  - Task 1: Please describe the approaches you took to increase the model's sparsity. (10%)
  - Task 2-1: Please describe the different data sizes you used for model calibration and the corresponding changes in test accuracy (a chart would be helpful). Based on these results, how would you recommend determining the appropriate data size? (15%)
  - Task 2-2: Please explain how you calculated the scale, zero-point, and the quantized weights. (10%)
  - Feedback about this assignment. (5%)

Grading

Task	Code	Report
Task 1	Function correct & Sparsity > 50% & test accuracy > 90.0% (15%) Performance ranking (10%) **(Only qualifiers will be ranked.)**	10%
Task 2-1	Function correct (10%)	15%
Task 2-2	Function correct & test accuracy > 90.0% (15%) Performance ranking (10%) **(Only qualifiers will be ranked.)**	10%

Feedback (report): 5%

File List

- 2025_DL_Lab04.zip
  - task1_pruning.ipynb (task 1)
  - task2_1_quantization_api.ipynb (task 2-1)
  - task2_2_quantization_manual.ipynb (task 2-2)
  - resnet20.py (Required in task 1)
  - resnet20_int8.py (Required in task 2-2)
  - check_pruning.py (Check your model in task 1)
  - check_quantization.py (Check your model in task 2-2)

Submission

Replace [student_id] with your student id.
Upload your [student_id].zip file to new E3.
Naming error will result in 5 point deduction.

[student_id].zip

- [student_id]_task1_pruning.ipynb (task 1)
- [student_id]_ task2_1_quantization_api.ipynb (task 2-1)
- [student_id]_ task2_2_quantization_manual.ipynb (task 2-2)
- [student_id]_pruning.pt (Generated in task 1)
- [student_id]_quantization.pt (Generated in task 2-2)
- [student_id]_report.pdf (Your report)

Reminder

- Submission deadline: 2 weeks **(2025/11/10 23:59 p.m.)**
- Use check_pruning.py and check_quantization.py to **check your model before submission**.
- If you encounter any problems with this assignment, please post your questions in the Facebook group. **TAs may not respond to private messages**.
- Plagiarism is strictly prohibited. Offenders will receive a score of zero.
- If you can't get access to a GPU, you can use **Google Colab** to complete your assignment as well.

Good Luck!