CS6886W - System Engineering for DL

Assignment 3

Computer Science and Engineering Department
IIT Madras, Chennai, Tamil Nadu

Due Date: Sunday, December 7 , 2025

Instructions:

This assignment focuses on **training MobileNet-v2 on CIFAR-10** and then applying **model compression techniques** to reduce model size while retaining accuracy.

Both accuracy and compression effectiveness will be evaluated.

Repository context: You may start from a template repository that demonstrates quantization based compression:

```
python test.py --weight_quant_bits 8 --activation_quant_bits 8
```

You must adapt it to **MobileNet-v2** and extend it with your **own compression method**.

You are not allowed to use any kind of compression API/library function to do the assignment. You should write your own code.

Please note that there will be relative grading for question no. 3 and 4 respectively. (students with better compression ratio and better accuracy will get higher points)

You can take help from this repository to learn more about **MobileNet-v2**: `https://github.com/pytorch/vision/blob/main/torchvision/models/mobilenetv2.py`

For a basic example of quantization, you may also refer to: `https://github.com/DextroLaev/CS6886-Sys_for_dl-Assignment3`

What to submit: Single PDF with answers, figures, tables and your GitHub repo link. Upload the PDF on **Moodle**.

Your GitHub repository should include clear commands to reproduce results.

The PDF should contain the following:

- Accuracy of the model without any compression.
- storage overheads (e.g., metadata, scaling factors)
- Best compression ratio of the model (when loaded in the memory).
- Best compression ratio of the weights (when loaded in the memory).
- Best compression ratio of the activations (state how you measured the activations).

- Wandb Parallel Coordinates chart.
- Final approximated model size (MB) after compression.

Question 1. Training Baseline (20 points)

(a) Prepare CIFAR-10 with proper normalization and data augmentation; specify transforms. (5)

(b) Describe your MobileNet-v2 configuration (e.g., width multiplier, dropout, BN settings) and training strategy (optimizer, LR schedule, regularization, epochs, batch size). (7)

(c) Report final test top-1 accuracy and include loss/accuracy curves; briefly discuss failure modes. (8)

Question 2. Model Compression Implementation (30 points)

(a) Implement a configurable compression method for reducing both model weights and activations. Clearly explain your design choices. (12)

(b) Show how the compression is applied to MobileNet-v2 (which layers compressed, any exceptions). (6)

(c) Document storage overheads (e.g., metadata, scaling factors) and include them in size estimates. (7)

Question 3. Compression Results

(a) Apply your compression pipeline at different levels of compression (e.g., varying bit-widths or parameters).

(b) Evaluate and compare the accuracy in these settings. Provide the Wandb Parallel Coordinates chart. Sample plot is given below (Minimum number of expected simulation run is 8, if you want to find the best one, you can try running the simulation for more than 8 runs).

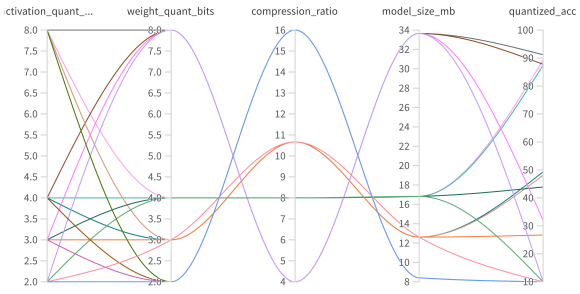

Figure 1: Parallel coordinates plot of VGG16 quantization sweep.

Please not that this is just a sample plot. You will get different values in the plot based on the algorithm you write.

Question 4. Compression Analysis

(a) Compression ratio of the *model*

(b) Compression ratio of the *weights*.

(c) Compression ratio of the *activations* (state how you measured the activations).

(d) Final approximated *model size (MB)* after compression.

Question 5. Reproducibility & Repository (10 points)

(a) Clean, modular,well commented codebase with separation of training, evaluation, and compression. (4)

(b) README with exact commands, environment, and dependency versions; include seed configuration. (4)

(c) Provide the GitHub repository link. (2)