Homework Assignment #2

Submission Due: 2021/04/06 23:59

Objective

1. In this homework assignment, you will learn to model inference of the CNN pretrained in the last assignment.
2. You should implement the functional model of the inference in high-level programming language, either Python or C/C++. However, we strongly recommend using Python, which can be executed on Google Colaboratory too.

Action Items

1. Implement a high-level functional model for each layer of the CNN, including convolution, pooling, and fully-connected layer with 8-bit quantization of the input activations, output activations, and weights accordingly.
2. Use 32-bit signed integers for the partial sums. That is, the accumulation of activations is limited to 32-bit precision in convolution and fully-connected layers. Saturate the result if the computation exceeds min/max values of 32-bit singed precision.
3. Verify your implementation by checking the output activations of conv1, pool1 layer, and fc3 layer with the following input images. Ensure that the fc3 results are the same as the CNN outputs in the previous assignment. This is the baseline implementation to start with.
   - ✓ Use the two test set images, index 1300 and index 3108, to verify whether your quantized inference matches the previous homework results.
4. Justify the model accuracy with the CIFAR10 **test dataset** as in the previous homework. Check if it matches the result in the previous homework.
   - ✓ You may encounter the speed issue for program execution. Refer to Numba official website to check how to accelerate your calculation: https://numba.readthedocs.io/en/stable/user/jit.html
   - ✓ Also, there is an example of matrix multiplication with Numba for your reference: https://drive.google.com/file/d/1PMFyA72UTy02AcyCH4Y-0Pw4ZrGdIyQT/view?usp=sharing
   - ✓ Other online reference: https://medium.com/jacky-life/%E9%AD%AF%E8%9B%87%E8%AE%8A%E8%9F%92%E8%9B%87%E8%A8%98-41e9c047e8e5
5. **Approach A:**
   Plot the distribution of partial sums of all quantized layers in the CNN with the CIFAR10 test dataset. Observe the dynamic range of partial sums. Based on the observation, can you reduce the bit-width of partial sums while still maintaining the accuracy? If yes, modify your inference and verify its accuracy accordingly.
   - ✓ In addition to the previous two test set images, use another two training set images, index 21280 and index 30702, to verify whether your quantized inference matches the previous homework results. Do you encounter overflow with any images?
6. **Approach B:**
   If you keep reducing the bit-width of partial sums, you can expect the accuracy will become lower and

lower. What is the minimum bit-width for the accuracy drop within 1%? Modify your inference and verify its accuracy accordingly.

7. **Approach C:**

   Again, plot the distribution of partial sums of **each quantized layer** in the CNN with the CIFAR10 test dataset. Determine the minimum bit-width of partial sums in each layer without hurting the accuracy. Modify your inference and verify its accuracy accordingly.

   ✓ Observe the distribution of partial sums of each quantized layer with the CIFAR10 **training dataset**. Compare with the observation for each quantized layer with test dataset in Approach C.

8. **Discussion:**

   Evaluate the three approaches based on the following energy model:

$$E_W = s_{mul} \times N_{mul} + s_{add} \times N_{add},$$

$$s_{mul} = 64 \times \left(\frac{B_{mul}}{8}\right)^2, s_{add} = B_{add},$$

   where N_{mul} and N_{add} are the number of multiplications and additions, respectively. B_{mul} and B_{add} are the bit-widths of multiplier and adder, respectively. s_{mul} denotes the power scaling factor of multiplication; s_{add} denotes the power scaling factor of addition. We use the energy weighting, E_W, to evaluate the total energy of different approaches.

   ✓ You need to calculate each layer's energy weighting and sum them all to obtain the overall E_W, if each layer has different B_{mul} or B_{add}.

   ✓ We only consider convolution and fully-connected operations, ignoring pooling and ReLU operations in this energy model.

   ✓ Disclaimer: Note that this energy model is artificial and oversimplified. DO NOT apply it to your research work.

9. Submission:

   a. Submit the PDF report according to the questions. Include the plots in your writeup. Use the following file name:

      `hw2_YourStudentID.pdf`

      Use the report template, which consists of Design Concept, Simulation and Discussion, and Summary. Note that for this homework assignment, you may describe the coding structure specific for the CNN architecture in the section of Design Concept.

   b. Also, hand in your source code with the following filenames:

      `hw2a_YourStudentID.py`, **`hw2b_YourStudentID.py`**, and **`hw2c_YourStudentID.py`** for each approach (Use proper file extension if you use C/C++ instead (and good luck…).)

      Note: you should detail how to execute the programs in the report explicitly. Also, put proper header/comments in the source codes.