HW2: Model Inference of the QAT LeNet

We have trained a CNN model similar to LeNet with QAT.

In this homework, we are going to implement the functional inference model in a high-level programming language, Python, and adjust the bit-width of the partial sums in each layer.

Action Items:

- □ Implement a high-level functional model for each layer of the CNN, including convolution, pooling, and fully-connected layers with 8-bit quantization of the input activations, output activations, and weights accordingly.
- □ Pass all unit tests of `OpTestCase`.
- □ Fill in all TODOs in `homework2.ipynb`.
- □ Answer all questions in `homework2.ipynb`.

How to launch Jupyter Notebook?

Choose either Option 1 or Option 2. If you are familiar with Jupyter Notebook, simply launch `homework2.ipynb` and start writing your homework.

Option 1: with Google Colaboratory on the Cloud

1. Open your Colab
2. Upload homework2.ipynb to Colab.
3. Run `!pip install numba==0.55.1` in Colab before using it.

- We don't want to install `Numba` again or upload any files when checking your homework. Comment out all comments you use in Step 3 before submitting your homework.

Option 2: with Conda on your computer

1. Install miniconda
2. Create a Conda virtual environment

```
conda create --name vlsi
conda activate vlsi
```

3. Install the following packages for this homework

```
conda install -c conda-forge matplotlib
conda install -c anaconda jupyter
conda install -c numba numba
```

4. Type `jupyter notebook` and launch Jupyter Notebook!

What do I need to submit?

1. Make sure you have done everything in `homework2.ipynb` and save `qat_prepare.pt`.

2. We don't want to install `Numba` again, upload/download any files to Colab, or retrain any models again when checking your homework. Comment out those lines of code for those processes!

```
# from google.colab import files
# uploaded = files.upload()

# for fn in uploaded.keys():
#     print('User uploaded file "{name}" with length {length} bytes'.format(
#         name=fn, length=len(uploaded[fn])))
...

# files.download(...)
...

# !pip install numba
...
model.load_state_dict(torch.load('qat_prepare.pt'))
#train(model, trainloader, 1)
```

3. Click `Kernel` and then click `Restart Kernel & Run All` on the Jupyter Notebook of `homework2.ipynb`.

    - Make sure everything goes smoothly without any warining or error messages while running your `homework2.ipynb`!

4. Upload `parameters.zip`, `homework2.ipynb`, and `qat_prepare.pt` to EECLASS. Do not zip these files or put them in a folder! Simply upload these four separate files.