Heidelberg University
Institute of Computer Engineering
Computing Systems Group

Holger Fröning
Hendrik Borras
Christian Simonides

Embedded Machine Learning
Summer Term 2024
Quantization and Pruning

Exercise 5 Quantization and Pruning

- **Return electronically until Wednesday, June 13, 2024, 09:00**
- **Include your names on the top sheet. Hand in only one PDF.**
- **A maximum of four students are allowed to work jointly on the exercises.**
- **When you include plots add a short explanation of what you expected and observed.**
- **Hand in source code if the exercise required programming. You can bundle the source code along with the PDF in a .zip file.**
- **Programming exercises can only be graded if they run on the cluster in the provided conda enviroment. Make sure to document additional steps, which you might have taken to run the exercises.**

In the following exercise you will get first experience with two important optimizations for resource constrained devices: Pruning and Quantization. You only have to do either the pruning or the quantization task, which ever you find more interesting. If you are committed, you are also free to do both.
For both tasks use the VGG11 or ResNet from the previous exercises and train it for 50 epochs. To give you some more leeway in terms of accuracy, you can use the SVHN data set instead of the CIFAR-10. Please make clear in your report, which architecture and dataset you are using.

5.1 Pruning

For the pruning use the native Pytorch functions with L-norm, as randomized pruning likely won't produce good results.

- First use unstructured pruning and vary the pruning rate in with at least five different pruning rates.
- Plot the accuracy during the training and the accuracy lost in comparison to a unpruned network.
- Rerun the same experiment again with structured pruning and also at least five different pruning rates.
- Calculate the number of BOPs for each of your graphs and compare them.
- Shortly discuss: What could be advantages/disadvantages of structured and unstructured pruning?

5.2 Quantization

For the quantization, please use brevitas[1], as it is more flexible that the quantization provided by Pytorch

- First use post training quantization(PTQ) and vary the quantization rate with at least four different values. You can quantize the whole network to the same bit width.
- Plot the accuracy during the training and the accuracy lost in comparison to a unquantized network.
- Rerun the same experiment but use quantization aware training(QAT).
- Calculate the number of BOPs for each of your graphs and compare them.
- Shortly discuss: What could be advantages/disadvantages of PTQ and QAT?

Total: 45 points

[1] https://xilinx.github.io/brevitas/getting_started.html