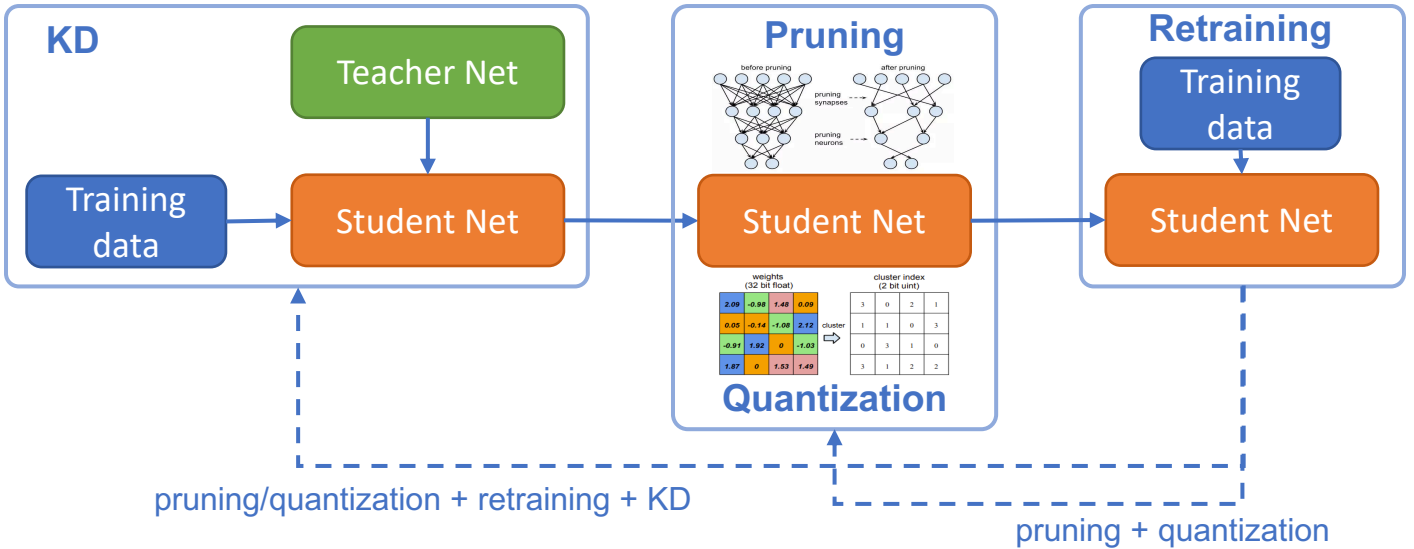
KD
Teacher Net
Training data
Student Net
Pruning
before pruning
after pruning
pruning synapses
pruning neurons
Student Net
weights (32 bit float)
2.09 -0.98 1.48 0.09
0.05 -0.14 -1.08 2.12
-0.91 1.92 0 -1.03
1.87 0 1.53 1.49
cluster
cluster index (2 bit uint)
3 0 2 1
1 1 0 3
0 3 1 0
3 1 2 2
Quantization
Retraining
Training data
Student Net
pruning/quantization + retraining + KD
pruning + quantization