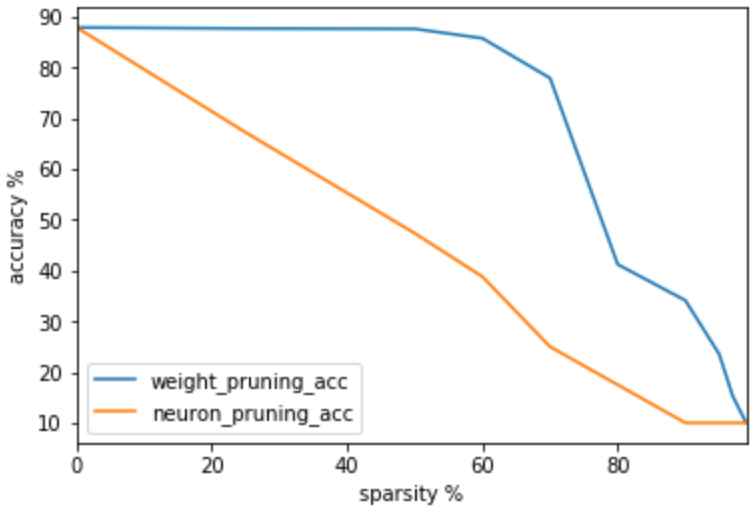
Sparsity v/s Accuracy for Pruned Models



Yes, there is considerable difference in the Accuracy curve for different sparsity levels in case of both techniques (weight & neuron pruning).

- In case of weight pruning, there is No/Minimal loss of accuracy till the sparsity level around 40% and starting from 50% sparsity level, the validation accuracy started to decrease.
- In case of neuron pruning, the loss accuray is almost negatively correlated...as sparsity increases, accuracy is decreasing linearly.

Reasons:

- In case of weight pruning, we are eliminating least k% weights in the magnitude. Obviously, weight which are least in magnitude contribute less share (as their co-efficient is very small) in the final prediction, hence replacing those with zeros is having minimal effect on the accuracy till certain level. At around 50% sparsity level, we are trying to eliminate more contributing weights decreasing the accuracy.

- In case of neuron pruning, we are eliminating least k% columns according to L2 norm. Here the loss of accuracy is more as we are eliminating entire columns of which few weight coefficients may be high (although total norm of column is low) & we are replacing those with zeros.