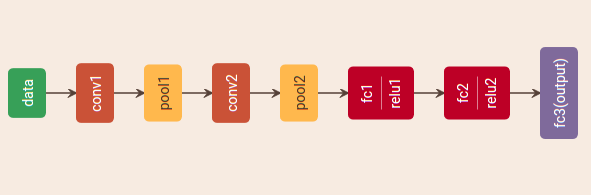
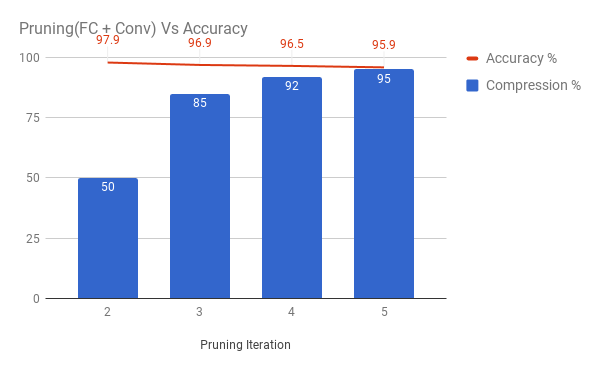
**Deep Compression on Lenet-5 Model:**



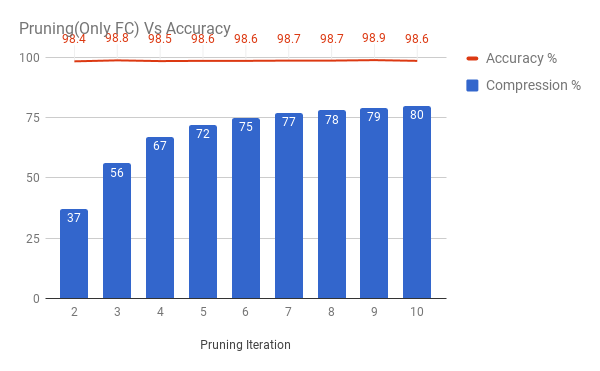
The above image shows the architecture of Lenet-5.

**Pruning on (Conv+FC) layers:**



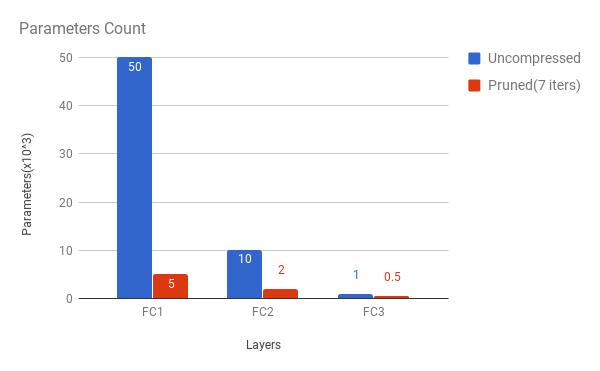
* The plot shows us the effect of pruning(both FC & Conv Layers) on accuracy.
* We observe that the accuracy decreases with the increase in the parameters pruned (parameters pruned pruning iteration)
* This might be as the these conv layers are first two layers and the model may have lost some connections where it learns some important features of the data.
* So,we decided not prune the conv layers for this model.

**Pruning on (Conv+FC) layers:**



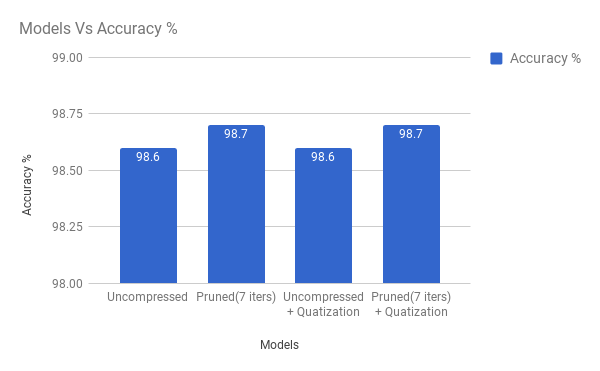
* The plot shows us the effect of pruning(only FC Layers) on accuracy.
* We observe that the accuracy remains constant with the increase in the parameters pruned. (parameters pruned pruning iteration)
* This might be as the these FC layers come later in the model architecture and most of these connections might be redundant.
* We choose to prune for 7 iterations as 90% parameters are pruned.(90% because,it’s mentioned in the deep compression paper that it’s advisable not to prune more than that)

**Parameters Count:**



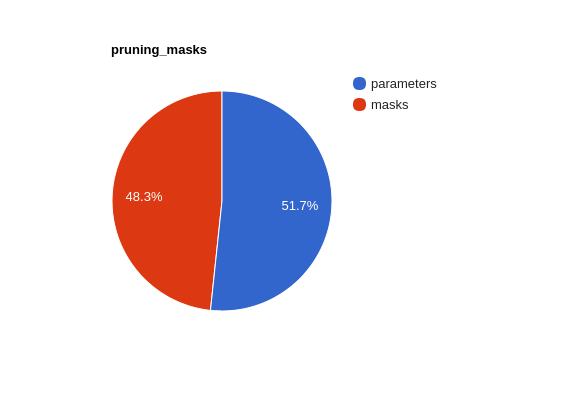
* Parameters Count-- No.of non-zero parameters in the model
* Uncompressed -- ~60K parameters
* Pruned(After 7 iterations) -- ~7.5K parameters and the parameters are reduced by ~8x i.e 87.5%

**Accuracy of the models:**



* There is no loss in the accuracy of the model after pruning is done.

**Prune Masks:**



* The headover with the prune masks in the pb file was about 50% but later on we came up with an idea to not include these masks in the protobuf

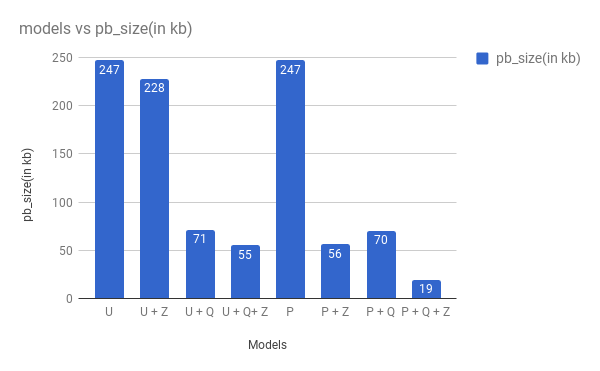
**Model sizes:**

# U--Uncompressed

# P--Pruned(7 iters)

# Q--Quantization

# Z--gzip



**#** PR--Non-zero parameters,PCR--Parameters compression ratio

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **PR** | **PCR** | **Pb Size(kb)** | **Pb Size after Z(Kb)** | **Pb Compression** | **Acc(%)** |
| **U** | ~60K |  | 247 | 228 | ~8% | 98.6 |
| **U+Q** | ~60K |  | 71 | 55 | ~78% | 98.6 |
| **P** | ~7.5K | ~8x | 247 | 56 | ~78% | 98.7 |
| **P+Q** | ~7.5K | ~8x | 70 | 19 | ~93% | 98.7 |

* The protobuf file(file used for deployment) size is reduced by 80% after implementing the pruning and by 93% after quantizing the pruned model

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Download size(kb)** | **In app size(After extraction)(kb)** | **Compression** |
| **Before pruning**  **(Uncompressed+Quantization)** | 55 | 71 |  |
| **After pruning**  **(Compressed+Quantization)** | 19 | 70 | 65% |

* The download bandwidth has been reduced for the compressed quantized model by 65% compared to the uncompressed quantized model and the size after extraction is same in both the models