0.0.1 Pruning

Network pruning is the process of removing unimportant connections, leaving only the most informative connections. There has been a substantial amount of research into how pruning can be used to reduce overfitting and network complexity [1]–[4], but more recent research shows that some pruning methodologies can produce pruned networks with no loss of accuracy [5].

0.0.2 Quantization and Weight Sharing

Quantization is the process of limiting the number of bits used to represent each weight within a network, this process results in many weights using identical or very close weight values. These repeated weight values creates an ideal situation to use weight sharing techniques.

The paper Deep Compression by Han et al [6] weight sharing is taken a step further and clustering is employed to gather the quantized weights into bins (whose value is denoted by the centroid of that bin) then an index is assigned to each weight that points to the weights corresponding bin, the bins value is the centroid of that cluster, which is further fine-tuned by subtracting the sum of the gradients for each weight in the bin their respective centroid see Fig. 1.

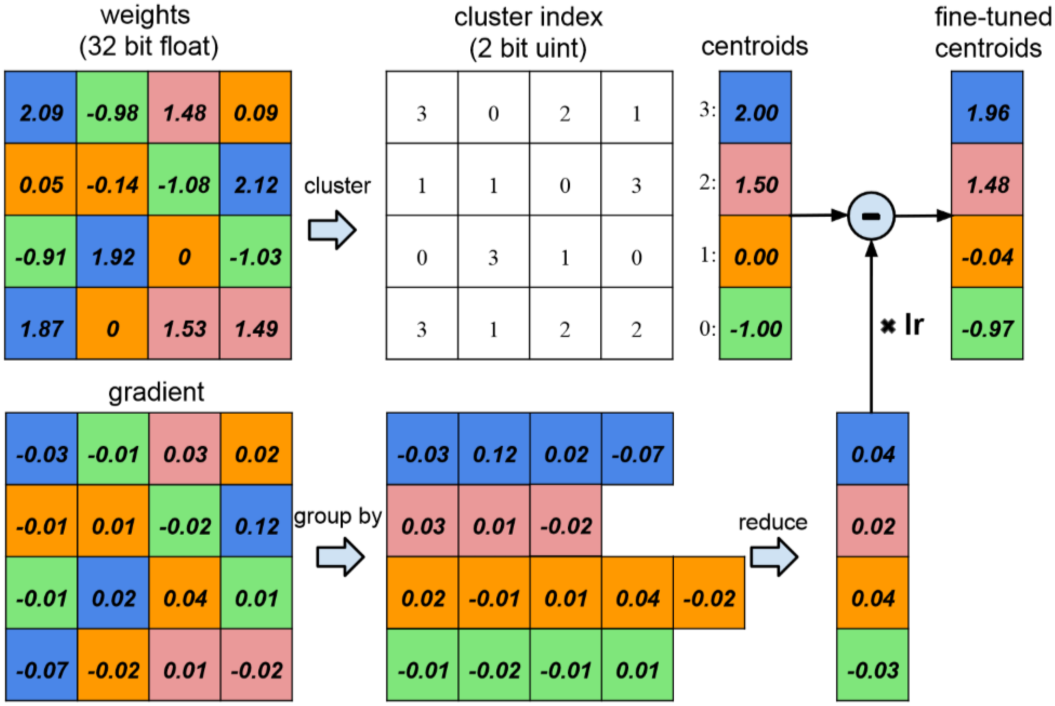

Figure 1: Weight sharing by quantization with centroid fine-tuning using gradients **(Adopted figure from [6])**

0.0.3 Distillation

0.0.4 Low-rank Factorization

0.0.5 Network Design Strategies