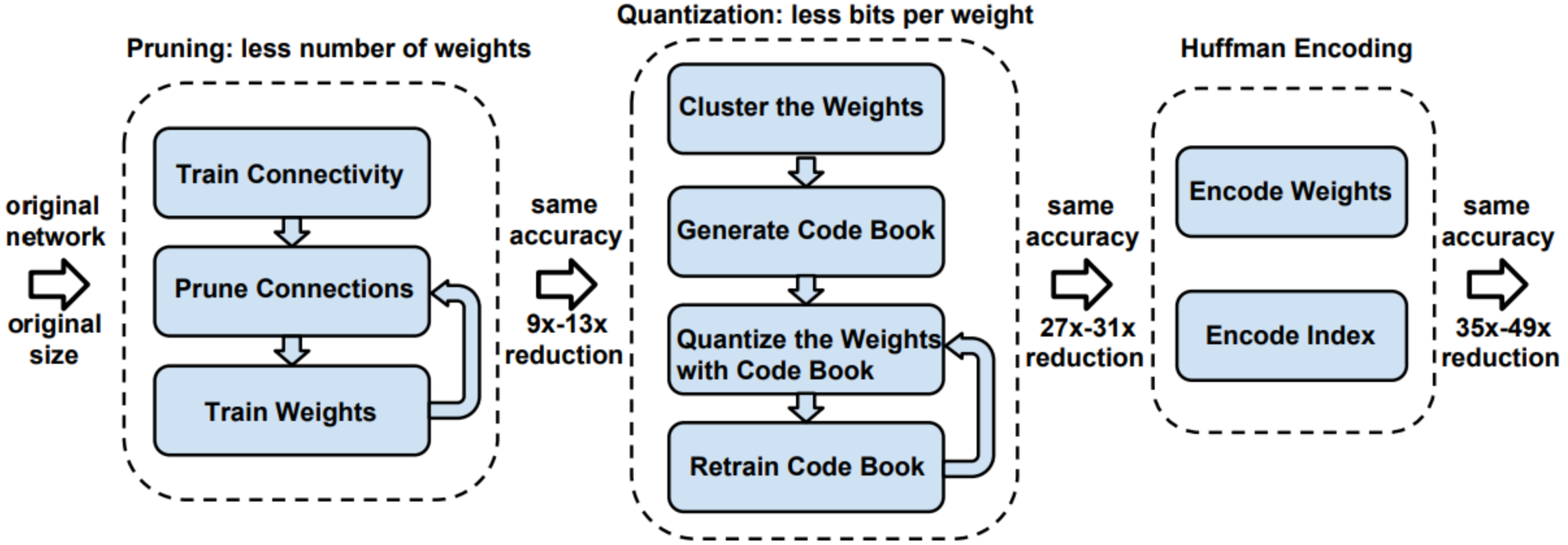
Pruning: less number of weights
Train Connectivity
Prune Connections
Train Weights
original network
original size
same accuracy
9x-13x reduction
Quantization: less bits per weight
Cluster the Weights
Generate Code Book
Quantize the Weights with Code Book
Retrain Code Book
same accuracy
27x-31x reduction
Huffman Encoding
Encode Weights
Encode Index
same accuracy
35x-49x reduction