Regularization Aided Pruning

Search for L1/L2 Regularizers

→

Train Parameters

→

Prune

→

Splice

→

Retrain

Parameter Profiling

Inspect Weights Distributions

→

Determine Central Values
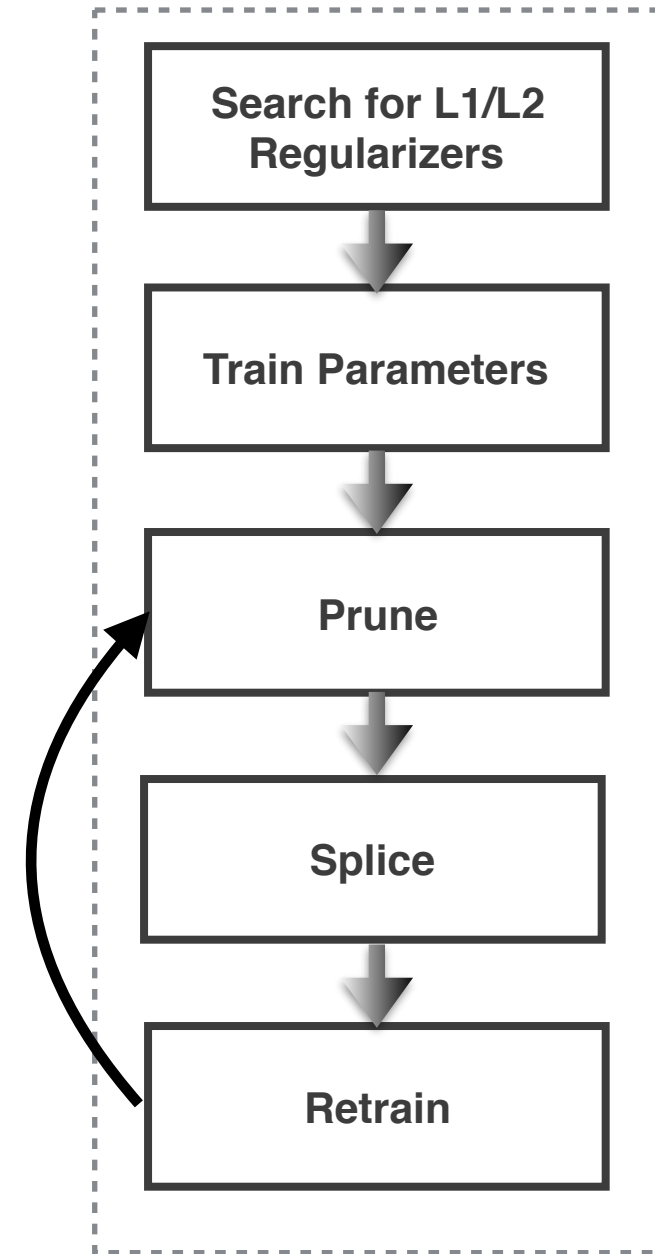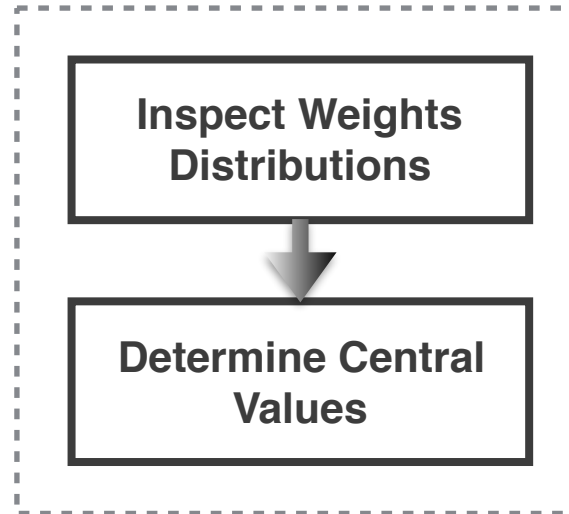
Dynamic fixed-point quantisation

Determine layer-wise dynamic range

→

Quantize

→

Retrain