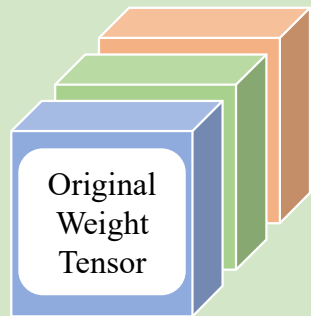


Original Model

Compressed Model



Current Layer



Partition

Weight Partitions

Partition W_1

Partition W_2

Partition W_n

Dictionary Pair-based TD

Dictionary Pair-based TD

Dictionary Pair-based TD

Synthesis Dictionary D_1 $P_1 W_1$

Synthesis Dictionary D_2 $P_2 W_2$

Synthesis Dictionary D_n $P_n W_n$

Decomposed Tensors

Other Layers



Original Output

Other Layers



Output after compression

Compression Process