

Compact Model: Modify the standard operations

used in DNNs.

e.g. (num of parameters) Standard CNN → { dilated conv, separable depthwise conv }

Standard LSTM → { S-LSTM, JANET }

Tensor Decomposition:

$$M = AB$$

$$M \times n \quad M \times r \times n$$

$$(M \times r) + (r \times n) < M \times n$$

hierarchical tensor representation

(HT)

Tensor train decomposition

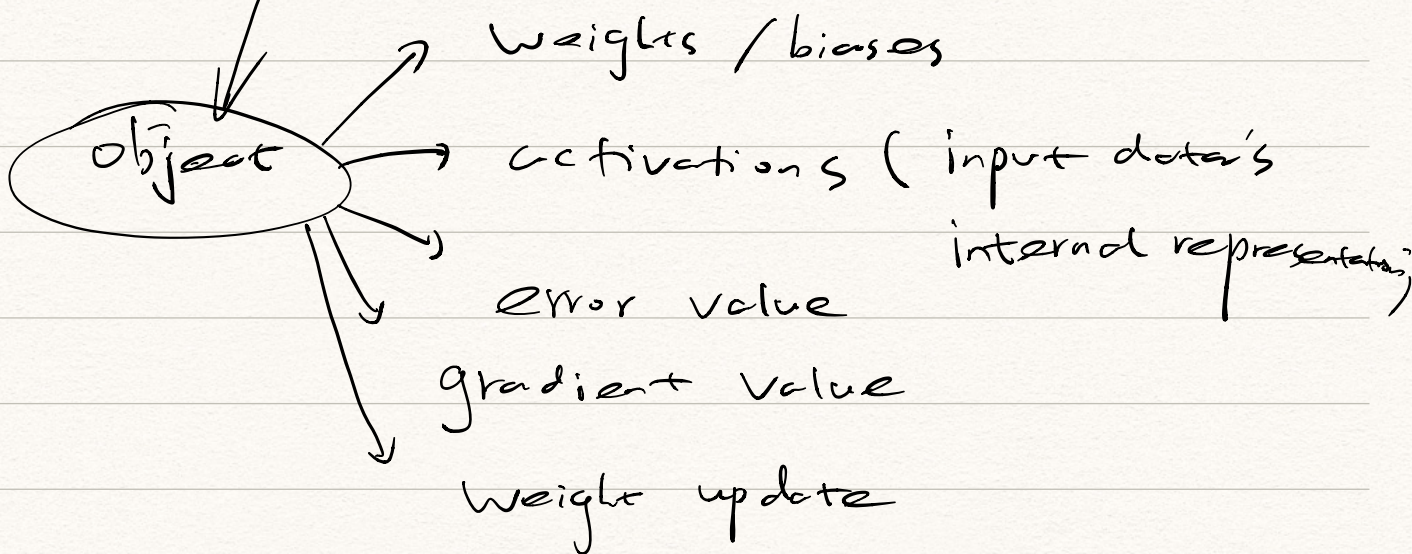
(TT)

Quantization : Convert data object

from 32-floating point to

lower precision or a fixed
point integer or even

binary



Network Sparsification / Pruning :

compress the model by pruning some
weights (edges) or operations (nodes)

"importance" ← weight values
 ← learned via an
 Attention Layer