↓ Model size

↓ number of FLOPs

↑ Compression ratio

Maintain accuracy

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:

$$M = AB$$

$M_{m \times n}$ $A_{m \times r}$ $B_{r \times n}$

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

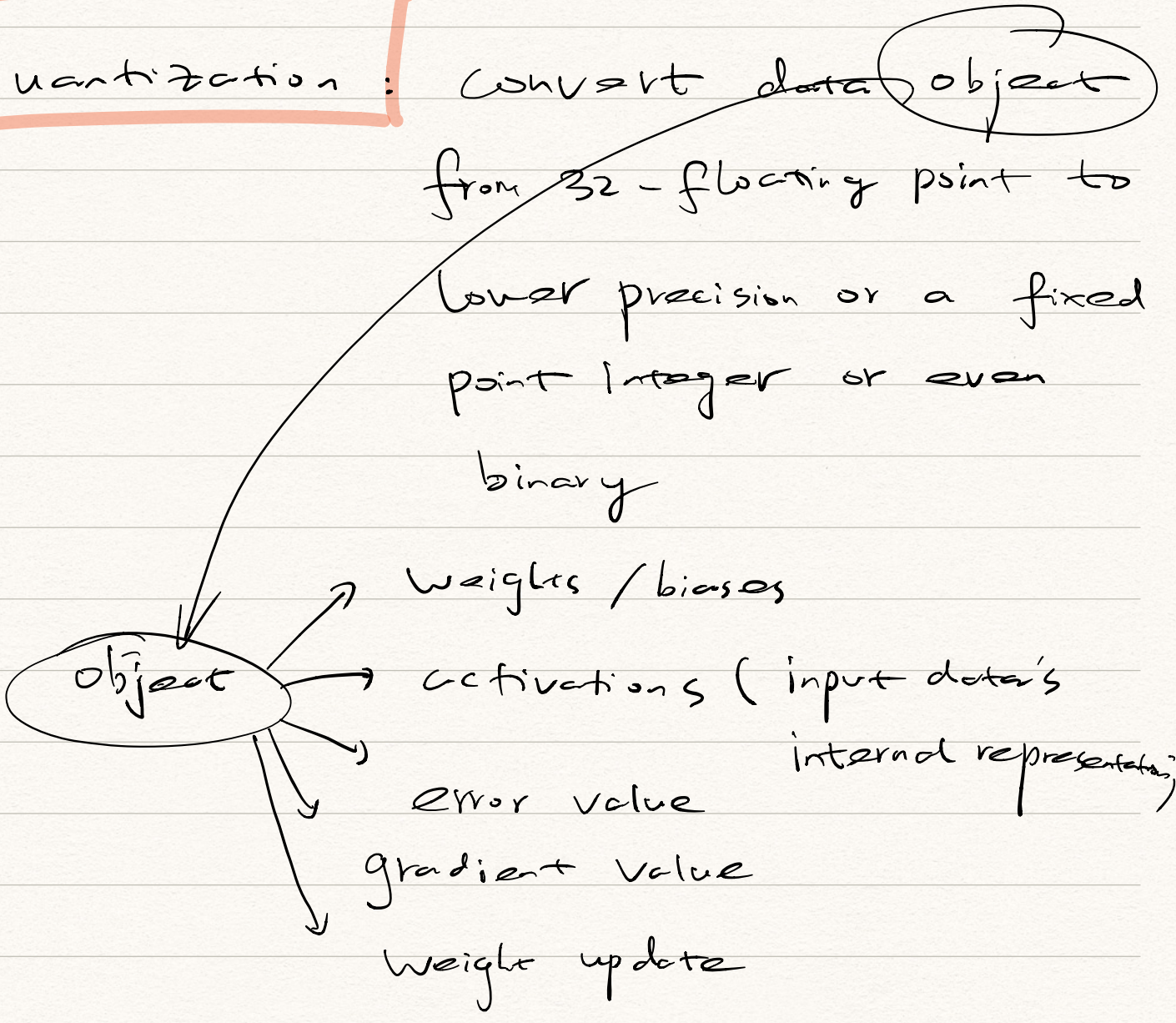
hierarchical tensor representation

(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

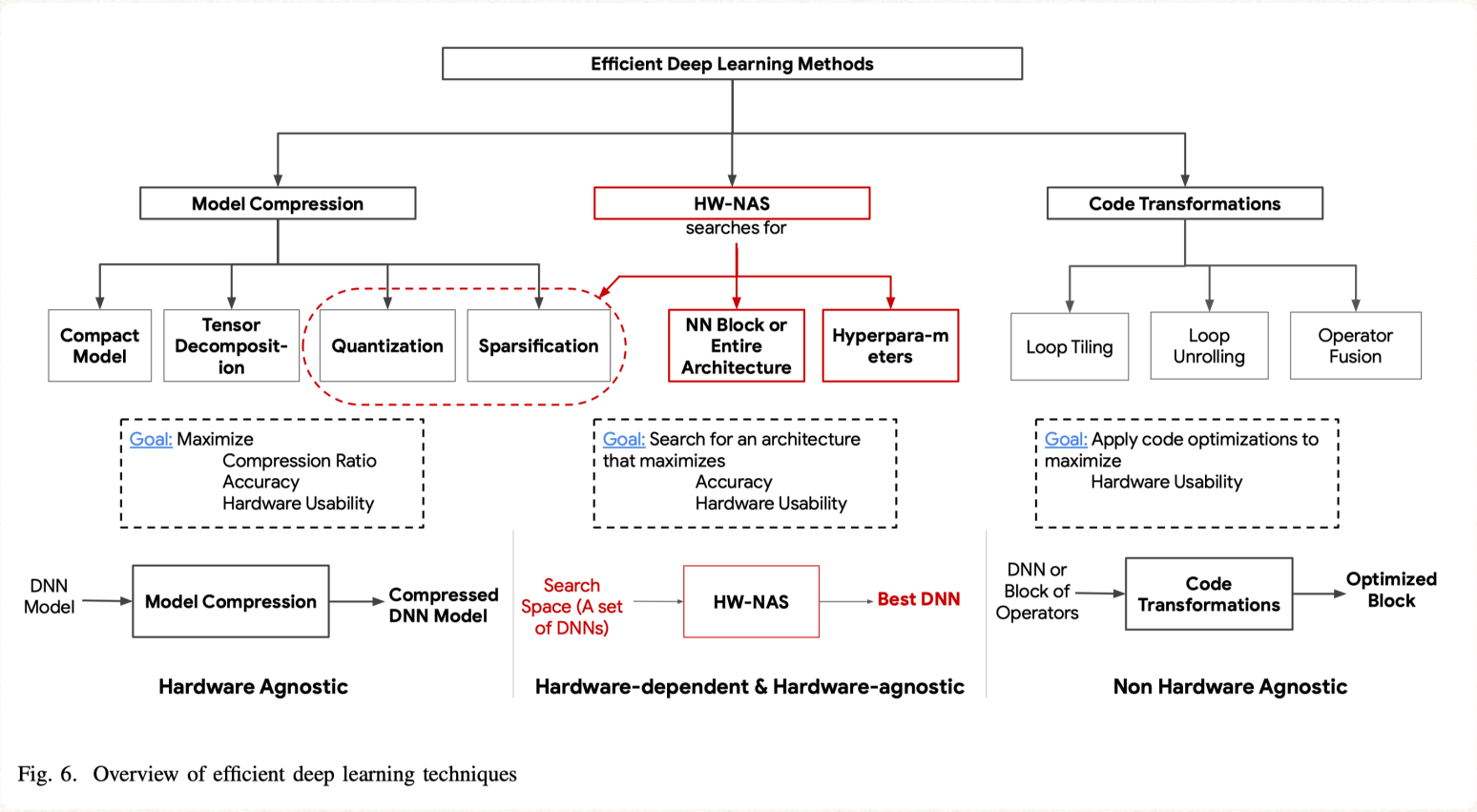


Fig. 6. Overview of efficient deep learning techniques