↓ { 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:**

$$\underset{m\times n}{M} = \underset{m\times r}{A}\,\underset{r\times n}{B}$$

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

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

object

- weights / biases
- activations ( input data's internal representation )
- error value
- gradient value
- weight update

**Network Sparsification / Pruning** :

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

" importance " ← weight values

↖ learned via an Attention Layer