Pruning

Quantisation

Data

Teacher model

Transfer Learning

Student model

Knowledge Distillation

≈

Tensor Decomposition

○ FP 32   ● INT 8