Optimizations

Software Optimizations

Compression

Quantization

Weight Pruning

Knowledge Distillation

Hardware Optimizations

GPUs

TPUs

NPUs