
Methods
Sub-Methods
On ASICs
On FPGAs
5.1 AISC & FPGA
5.1.1 Quantization
5.1.2 Sparsity
5.1.3 Operator
5.1.4 Architecture
5.1.1.1 Weight-only
GOBO, FIGNA, MECLA
EdgeLLM
5.1.1.2 Weight-Activation
Mokey, OliVe, OPAL, SPARK, Tender, Li et al. 2024f [1], Yu et al. 2024a [2]
Hur et al. 2023 [3], HLSTransform, SECDA-LLM, Chen et al. 2024a [4], FlightLLM, LNS-LLM
5.1.2.1 Sparse Attention
A3, ELSA, Sanger, SpAtten, FACT, SALO2, ASADI, ALISA, SOFA
5.1.2.2 Sparse Model
TF-MVP, Edge-LLM, BBS, FEASTA
FlightLLM, EdgeLLM
5.1.3.1 Operator Fusion
8bit-Transformer, LPU
FlightLLM
5.1.3.2 Nonlinear Operators
8bit-Transformer, Consmax
5.1.3.3 Storage Optimization
Atalanta
5.1.3.4 Matrix Multiplication
Cambricon-C
5.1.4.1 PE Array
Sanger, SALO2, MECLA, Trapezoid, BBS
LNS-LLM
5.1.4.2 Specific Module
SpAtten, FACT, OPAL, Base-2 Softmax
5.1.4.3 Memory Storage
FACT
5.1.4.4 Multi-device
Cambricon-LLM
DFX