An Energy-Efficient Multi-DNN Training Processor for GANs with Speculative Dual-Sparsity Exploitation

S. Kang, D. Han, J. Lee, D. Im, S. Kim, S. Kim, J. Ryu, and H-J. Yoo (KAIST)

Motivation
1) Multi-DNN Architectures of GANs

GAN Training Process
Step 1) Fix G & Train D
Step 2) FIX D & TRAIN G

• Different bottlenecks co-exist in multi-DNNs
  (Computation bottleneck & Memory bottleneck)

2) Sparsity Pattern of DNN Training

Generator DNN Sparsity Pattern

• FF stage: Both IA sparsity & OA sparsity (unknown)
• BP stage: Only OA sparsity

System Implementation

Demonstration System

Android Tablet (UI)

System Board

System Architecture

Android Tablet

USB Ctrlr. Flash

Android Tablet

GAPPU (Glu Logic)

FPGA

Flash

DDR3

FPGA

Verification

Face Modification w/ On-Device Training

New Mask

Output Image

On-Device Training

Architectural Efficiency

Overall Architecture

• Total of 32 training cores (DSTC) & Shared memory
  1. Adaptive spatio-temporal workload multiplexing with reconfigurable accumulation network
  2. Dual sparsity (IA & OA) exploitation architecture
  3. Exponent-only ReLU speculation (EORS) to utilize OA sparsity