Heterogeneity Aware Federated Learning and Demand-Specific Inference

Jianming Tong, Alind Khare, Alexey Tumanov, Tushar Krishna

School of Computer Science [1], School of Electrical and Computer Engineering

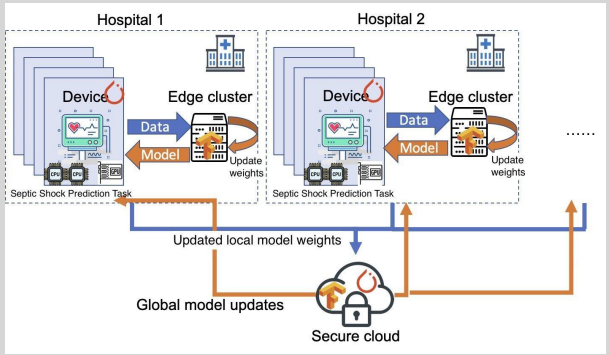
Georgia Tech | Center for Research into Novel Computing Hierarchies

Introduction

Application: Health-care
Demand:
- highly sensitive data with different distributions geographically distributed clients.
- heterogeneous computational power
  - mobile phones may range from GPU (dis)abled or 4-12GB RAM etc.

Problem Statement
- Privacy-aware ML training for highly sensitive heterogeneous data.
- Serving various accuracy and latency demand.

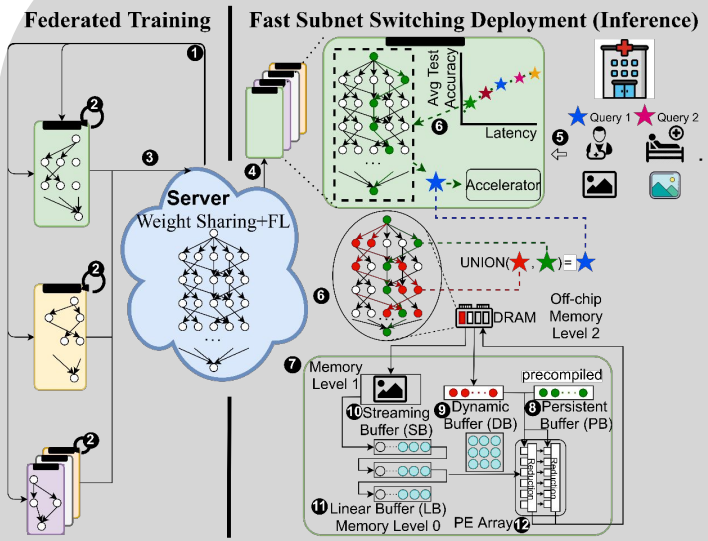
Overall System High-level Overview



State-of-the-Art

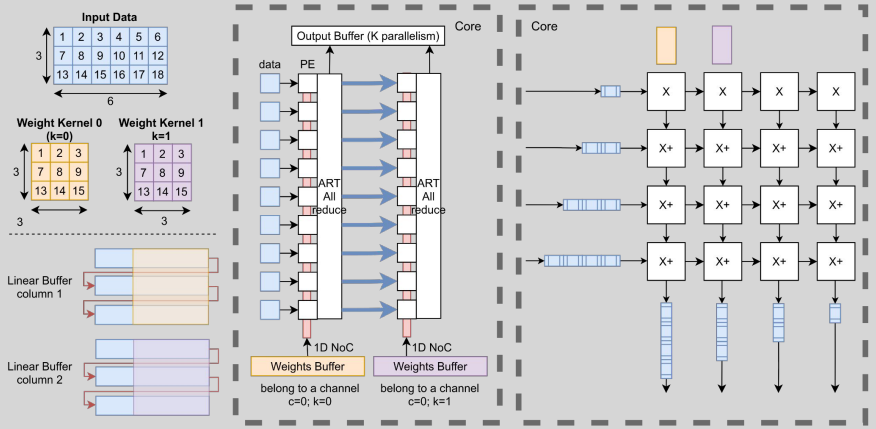
- Federated Learning: Federated Algorithm [1]
  - Homogeneous clients assumption
- Once-For-All [2] requires centralized data
- Xilinx Vitis AI needs minutes to switch models [3]

Proposal System

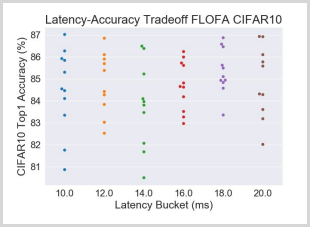


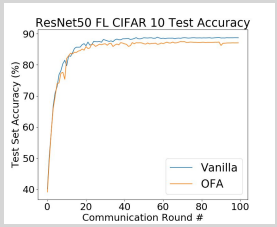
1. Subnet Distribution
2. Local Training
3. Shared-Weight Federated Average
4. Pretrained Weight-sharing Network Deployed on clients
5. Streaming input of quereis
6. NAS on Latency-Accuracy Pareto-Frontier for input query
7. Inference the NAS model on Accel.
8. PB stores subnet's weights (green) unchanged across queries
9. DB stores the weights of model requested by query not stored in PB
10. SB store input feature map
11. Input will shift in LB in the row order

Plug-in-and-play PE Array Design



Federated Learning Evaluation





Results
- Model Family for different Accuracy-Latency
- Same level accuracy as SOTA

Quick Subnet Switching Inference

Xilinx Vitis AI with Xilinx Deep Processing Unit (DPU)

Components	Utilization
CLB LUT	49%
CLB Regs	86%
CLB Logic	92%
DSP	87%
MMCM	33%
BRAM	59%

Results
- 12 FPS for ResNet-50
- Model Switching needs several minutes

Georgia Tech

[1] Communication-Efficient Learning of Deep Networks from Decentralized Data [2] Once-for-All: Train One Network and Specialize it for Efficient Deployment [3] Xilinx Vitis unified software platform