Edge Intelligence: the Confluence of Edge Computing and Artificial Intelligence

Hailiang Zhao
hliangzhao@zju.edu.cn

College of Computer Science and Technology, Zhejiang University

November 17, 2019
Outline

1 Introduction
- 5G, edge, and AI
- Relations between Edge Computing and AI
- Birth of Edge Intelligence
Outline

1 Introduction
   - 5G, edge, and AI
   - Relations between Edge Computing and AI
   - Birth of Edge Intelligence

2 Research Roadmap of Edge Intelligence
   - Roadmap overview
   - Quality of Experience
   - Intelligence-enabled Edge Computing
   - Artificial Intelligence on Edge
Outline

1 Introduction
   - 5G, edge, and AI
   - Relations between Edge Computing and AI
   - Birth of Edge Intelligence

2 Research Roadmap of Edge Intelligence
   - Roadmap overview
   - Quality of Experience
   - Intelligence-enabled Edge Computing
   - Artificial Intelligence on Edge

3 AI for Edge
   - State of the Art
   - Grand Challenges
Outline

1 Introduction
   - 5G, edge, and AI
   - Relations between Edge Computing and AI
   - Birth of Edge Intelligence

2 Research Roadmap of Edge Intelligence
   - Roadmap overview
   - Quality of Experience
   - Intelligence-enabled Edge Computing
   - Artificial Intelligence on Edge

3 AI for Edge
   - State of the Art
   - Grand Challenges

4 AI on Edge
   - State of the Art
   - Grand Challenges
Outline

1 Introduction
   - 5G, edge, and AI
   - Relations between Edge Computing and AI
   - Birth of Edge Intelligence

2 Research Roadmap of Edge Intelligence
   - Roadmap overview
   - Quality of Experience
   - Intelligence-enabled Edge Computing
   - Artificial Intelligence on Edge

3 AI for Edge
   - State of the Art
   - Grand Challenges

4 AI on Edge
   - State of the Art
   - Grand Challenges
5G is coming!

What 5G brings to us

1. enhanced Mobile BroadBand (eMBB)
2. Ultra-Reliable Low Latency Communications (URLLC)
3. massive Machine Type Communications (mMTC)
Processing data nearby

Why edge?
1. explosion of data generated by mobile and IoT devices
2. oppressive network congestion in backbone
3. ...

Multi-access Edge Computing (MEC)
1. communication/computation/caching/control at the edge directly
2. provide services
3. perform computations
4. manage resources

MEC avoids unnecessary communication latency and enabling faster responses for end users.

---

A typical pre-5G HetNet
What about Artificial Intelligence?

1. powerful in **big data processing & insights extracting**
2. DNNs: powerful **knowledge representation**
3. Typical structures of DNNs
   1. Multilayer Perceptrons (MLP)
   2. Convolutional Neural Network (CNN) (AlexNet → VGG-16 → GoogleNet → ResNet)
   3. Recurrent Neural Network (RNN) (RNN → LSTM)
4. Popular DNN models
   1. Generative Adversarial Network (GAN)
   2. Deep Reinforcement Learning (DRL)
Can they integrate with each other?

1. **AI provides Edge Computing with methods and technologies**
   - Complicated resource allocation problems need to solve
   - Huge volumes of data need to analysis
   - AI can help in **model formulation & optimization**

2. **Edge Computing provides AI with scenarios and platforms**
   - More and more data is created by widespread and **geographically distributed** mobile and IoT devices
   - Many more application scenarios (intelligent networked vehicles, autonomous driving, smart home, smart city, ...)
   - **Hardware acceleration** on resource-limited IoT devices

Their integration leads to the birth of

**Edge Intelligence (a.k.a. Edge AI)**
Edge Intelligence: our definition

We divide it into **AI for edge** and **AI on edge**.

1. **AI for edge**
   1. provide a better solution to the constrained optimization problems
   2. AI is used for energizing edge with more intelligence and optimality
   3. **Intelligence-enabled Edge Computing (IEC)**

2. **AI on edge**
   1. carry out the entire process of AI models on edge
   2. run model training and inference with device-edge-cloud synergy
   3. **Artificial Intelligence on Edge (AIE)**
Outline

1 Introduction
   - 5G, edge, and AI
   - Relations between Edge Computing and AI
   - Birth of Edge Intelligence

2 Research Roadmap of Edge Intelligence
   - Roadmap overview
   - Quality of Experience
   - Intelligence-enabled Edge Computing
   - Artificial Intelligence on Edge

3 AI for Edge
   - State of the Art
   - Grand Challenges

4 AI on Edge
   - State of the Art
   - Grand Challenges
Roadmap Overview

Quality of Experience (QoE)
- Performance
  - Problem-based Indicators
  - Training loss + Test Accuracy
- Cost
  - Computation Resource (delay)
  - Communication Resource (latency)
  - Energy Consumption
- Privacy (Security)
- Efficiency
- Reliability

AI for Edge
- Service
  - Computation Offloading
  - User Profile Migration
  - Mobility Management
- Content
  - Data Provisioning
    - Provisioning
    - Placement
    - Composition
    - Caching
- Topology
  - Edge Site Orchestration
    - Data Acquisition
    - Network Planning

AI on Edge
- Model Adaptation
  - Model Training
    - Federated Learning
    - Knowledge Distillation
  - Algorithm Asynchronization
  - Thoroughly Decentralization
- Framework Design
  - Model Inference
    - Partitioning
    - Splitting
- Processor Acceleration
  - Instruction Set Design
  - Parallel Computation
  - Near-data Processing
QoE: indicators

1. **performance**
   - AI for edge: problem-dependent
   - AI on edge: training loss, inference loss

2. **cost**
   - computation cost (CPU time, CPU frequency)
   - communication cost (transmit power, frequency band, access time)
   - energy consumption (battery capacity)

3. **privacy (security)**
   - leads to the birth of Federated Learning

4. **efficiency**
   - excellent performance with low overhead

5. **reliability**
   - robustness
   - handle with failure
AI for edge: a recapitulation

1. **Service**
   1. optimize computation offloading via DQN\textsuperscript{23}

2. **Content**
   1. service placement via MAB\textsuperscript{4}
   2. service deployment via DRL\textsuperscript{5}

3. **Topology**
   1. optimize UAVs via Multi-agent Learning\textsuperscript{6}
   2. learning-driven communication\textsuperscript{7}


AI on edge: a recapitulation

1. **model adaptation** (too many of them)
   - model compression, conditional computation, algorithm asynchronization, thoroughly decentralization, ...

2. **framework design**
   - model training: Federated Learning on edge\(^8\), knowledge distillation-based methods\(^9\)
   - model inference: model splitting/partitioning (Edgent)\(^10\)

3. **processor acceleration**\(^11\)
   - design special instruction sets
   - design high parallel computing paradigms
   - move computation closer to memory

---


Outline

1 Introduction
- 5G, edge, and AI
- Relations between Edge Computing and AI
- Birth of Edge Intelligence

2 Research Roadmap of Edge Intelligence
- Roadmap overview
- Quality of Experience
- Intelligence-enabled Edge Computing
- Artificial Intelligence on Edge

3 AI for Edge
- State of the Art
- Grand Challenges

4 AI on Edge
- State of the Art
- Grand Challenges
Utilize DQN for performance optimization

- **Problem Definition**
  - Performance Optimization in MEC

- **Model Construction**
  - Deep Q-Network (DQN)

- **Algorithm Design**

**Goal**
- Need-to-be-minimized delay
  - execution delay
  - handover delay
  - task dropping cost

**Decision Variables**
- Binary Task
  - energy allocation
  - edge server selection
- Partial Task
  - offloading or not
  - partition point

**Observe States**
- energy state
- task request state
- resource usage

**Observe Actions** (Discretization)
- energy allocation
- edge server selection
- offloading decision

**Remove Constraints**
- add penalty
- transfer to goal
- add assumption

**Constraints**
- battery energy level
- task execution deadline
- radio frequency bandwidth
- computing resources

**Memory Pool** (Database of Samples)
- Mini-batch
- DNN
- Weight Updating
- (alternative)
- Action
- Policy
- Environment
- State

**Action**
- Cost
**Grand challenges**

1. **model establishment**
   - unrestrained searching space
   - state/action set cannot be infinite

2. **algorithm deployment**
   - cannot obtain analytic (approximate) optimal solution
   - too many iterations → hard to deploy in an online manner
   - who undertake the responsibility?

3. **balance between optimality and efficiency**
Outline

1. Introduction
   - 5G, edge, and AI
   - Relations between Edge Computing and AI
   - Birth of Edge Intelligence

2. Research Roadmap of Edge Intelligence
   - Roadmap overview
   - Quality of Experience
   - Intelligence-enabled Edge Computing
   - Artificial Intelligence on Edge

3. AI for Edge
   - State of the Art
   - Grand Challenges

4. AI on Edge
   - State of the Art
   - Grand Challenges
Model Adaptation: a classification

- Model Compression
  - Exploit the inherent sparsity structure of gradients and weights
- Conditional Computation
  - Selectively turn off some unimportant calculations
- Algorithm Asynchronization
  - Aggregate local models in an asynchronous way
- Thoroughly Decentralization
  - Remove the central aggregator to avoid any possible leakage

Methods

- Quantizing
- Dimensional Reduction
- Pruning
- Precision Downgrading
- Components Sharing
- Components Shutoff
- Input Filtering
- Early Exit
- Results Caching

Approaches

- Singular Value Decomposition
- Huffman Coding
- Block-wise Dropout
- Participator Selection
- Random Gossip Communication
- Smart Contract
- Game Theory
Grand challenges

1. **data availability**
   - 1. where to find *usable* data?
   - 2. incentive mechanisms
   - 3. obvious bias from distributed end users (non i.i.d.)

2. **model selection**
   - 1. select befitting threshold of learning accuracy & scale of models
   - 2. select probe training frameworks and accelerator architectures

3. **coordination mechanism**
   - 1. same method achieves different results
   - 2. compatibility and coordination (cloud-edge-device synergy)
   - 3. establish a unified API interface?