Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing

Hailiang Zhao

July 17, 2019

https://hliangzhao.github.io/CV/

Outline

1. Introduction
   - Motivation and Definition
   - Scope and Rating

Outline

1. Introduction
   - Motivation and Definition
   - Scope and Rating
2. A Primer on Artificial Intelligence
   - Deep Learning and Deep Neural Networks
   - Popular Deep Learning Models

Outline

1. Introduction
    - Motivation and Definition
    - Scope and Rating
2. A Primer on Artificial Intelligence
    - Deep Learning and Deep Neural Networks
    - Popular Deep Learning Models
3. Edge Intelligence Model Training and Inference
    - Model Training
    - Model Inference

Outline

1. Introduction
    - Motivation and Definition
    - Scope and Rating
2. A Primer on Artificial Intelligence
    - Deep Learning and Deep Neural Networks
    - Popular Deep Learning Models
3. Edge Intelligence Model Training and Inference
    - Model Training
    - Model Inference
4. Future Research Directions

Outline

1. Introduction
   - Motivation and Definition
   - Scope and Rating
2. A Primer on Artificial Intelligence
   - Deep Learning and Deep Neural Networks
   - Popular Deep Learning Models
3. Edge Intelligence Model Training and Inference
   - Model Training
   - Model Inference
4. Future Research Directions
5. Concluding Remarks

Outline

1 Introduction
- Motivation and Definition
- Scope and Rating

2 A Primer on Artificial Intelligence
- Deep Learning and Deep Neural Networks
- Popular Deep Learning Models

3 Edge Intelligence Model Training and Inference
- Model Training
- Model Inference

4 Future Research Directions

5 Concluding Remarks

About the Slide

This slide is a report on the paper *Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing*, preprinted on arXiv, May 24, 2019. The authors, Zhi Zhou, Xu Chen et al are with the School of Data and Computer Science, Sun Yat-sen University (SYSU).

I agree with insights of this paper *indifinitely*.

Why Edge Intelligence?

1. The edge ecosystem fuels the continuous booming of AI
    1. Big data is a key driver that boosts AI development
    2. Data source: the mega-scale cloud datacenters → the increasingly widespread end devices
    3. Offloading huge end data to cloud is impossible (network conjustion)
    4. Edge computing is a key infrastructure for *AI democratization*
2. Edge computing needs AI to full unlock their potential
    1. AI is functionally necessary for quickly analyzing huge data volumes and extract insights (to realize ubiquitous AI)
    2. AI may provides better mechanisms for *communication on edge*

Definition

Currently, most organizations and presses refer to Edge Intelligence as the paradigm of running AI algorithms locally on the end devices, with data (sensor data or signals) created on the device.

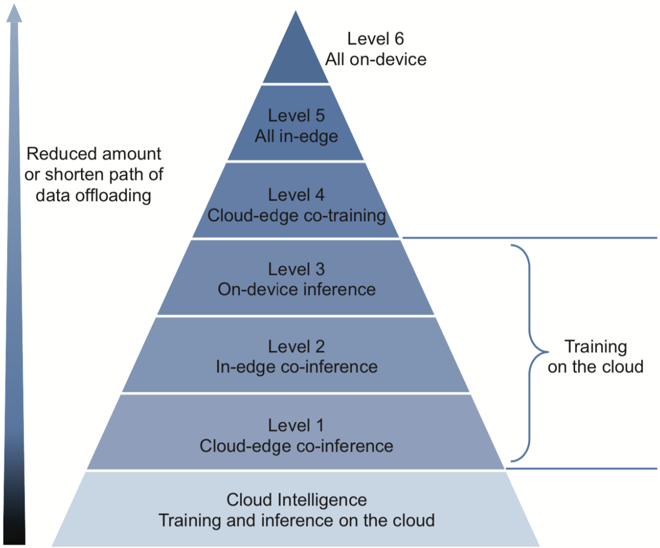
Too narrow!

Edge Intelligence (my definition)

Edge Intelligence is the paradigm of running AI models' training and inference with *device-edge-cloud synergy*, which aims at extracting insights from massive and distributed edge data with the satisfaction of *Quality of Experience (QoE)*.

QoE should be application-dependent and determined by jointly considering multi-criteria such as *AI models' overall performance (training loss and test accuracy)*, *computation latency*, *communication cost*, *energy efficiency*, *privacy*, etc.

A 6-level rating for edge intelligence



Outline

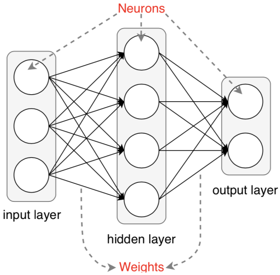
1. Introduction
   - Motivation and Definition
   - Scope and Rating
2. A Primer on Artificial Intelligence
   - Deep Learning and Deep Neural Networks
   - Popular Deep Learning Models
3. Edge Intelligence Model Training and Inference
   - Model Training
   - Model Inference
4. Future Research Directions
5. Concluding Remarks

Deep Learning and Deep Neural Networks

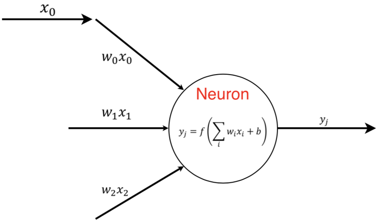
Among the existing machine learning methods, deep learning, by leveraging artificial neural network (ANN) to learn the deep representation of the data, have resulted in an amazing performance in multiple tasks.

Powerful knowledge representation of ANN

An ANN with *single* hidden layer containing *enough* neurons can approximate continuous functions of *any* complexity to *any* accuracy.

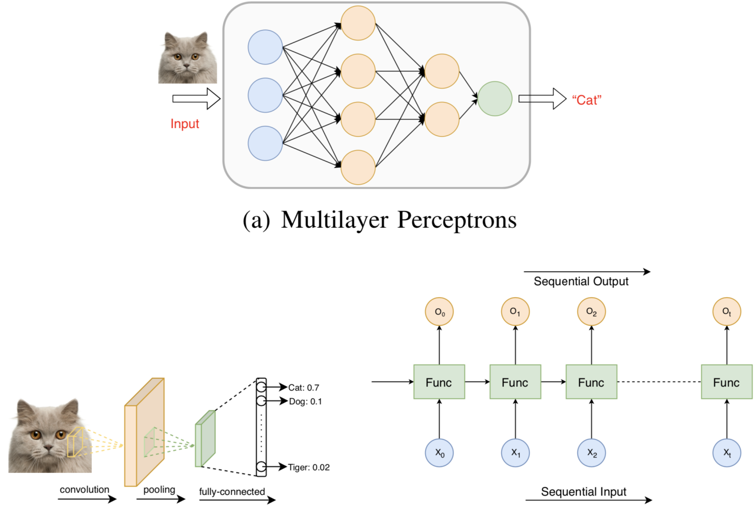


(a) The layers in a DL model



(b) The architecture of a neuron

Threee typical structures of DL models



(a) Multilayer Perceptrons

(b) Convolution Neural Network

(c) Recurrent Neural Network

Threee typical structures of DL models

1. Multilayer Perceptrons (MLP)
   MLP models are the most basic deep neural network, which is composed of a series of fully-connected layers
2. Convolutional Neural Network (CNN)
   CNN models have convolution layers, which can extract the simple features from input by executing convolution operations. Applying various convolutional filters, CNN models can capture the high-level representation of the input data.
3. Recurrent Neural Network (RNN)
   RNN models use sequential data feeding. RNN models are widely used in the task of natural language processing.

Popular Deep Learning Models

1. Convolutional Neural Network (CNN)
   AlexNet → VGG-16 → GoogleNet → ResNet

2. Recurrent Neural Network (RNN)
   The training of RNN is based on Backpropagation Through Time (BPTT). Long Short Term Memory (LSTM) is an extended version of RNNs.

3. Generative Adversarial Network (GAN)
   GAN consists of two main components, namely the generator and discriminator. The generator is responsible for generating new data after it learns the data distribution from a training dataset of real data. The discriminator is in charge of classifying the real data from the fake data generated by the generator.

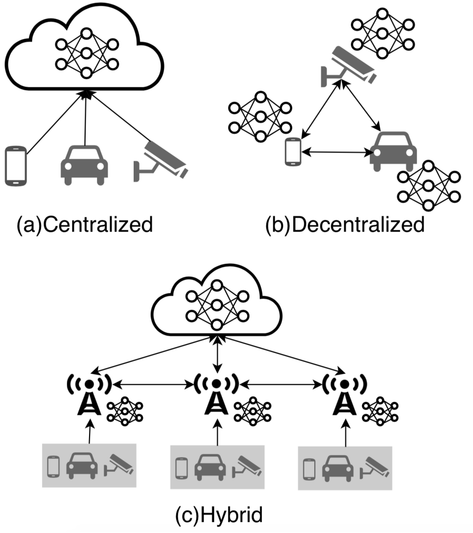
4. Deep Reinforcement Learning (DRL)
   DRL is composed of DNNs and reinforcement learning (RL). In the procedure of value function approximation, DRL chooses CNN (highly non-linear) as the function.

Outline

1. Introduction
   - Motivation and Definition
   - Scope and Rating
2. A Primer on Artificial Intelligence
   - Deep Learning and Deep Neural Networks
   - Popular Deep Learning Models
3. Edge Intelligence Model Training and Inference
   - Model Training
   - Model Inference
4. Future Research Directions
5. Concluding Remarks

Model Training - Architectures

Centralized – Decentralized (√) – Hybrid (√)



(a)Centralized

(b)Decentralized

(c)Hybrid

Model Training - Key Performance Indicators

1. Training loss
   Essentially, the DNN training process solves an optimization problem that seeks to minimize the training loss.

2. Computation latency (for *decentralized* and *hybrid*)
   This indicator is tightly dependent on the capability of the nodes (edge equipment or end device)

3. Communication cost (for *decentralized* and *hybrid*)
   The raw data or intermediate data should be transferred across the nodes. Communication overhead is affected by the size of the original input data, the way of transmission and the available bandwidth.

4. Energy efficiency (for *decentralized* and *hybrid*)
   Edge nodes and end devices are energy-constrained.

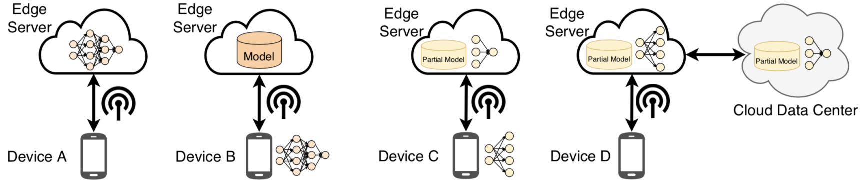
5. Privacy (for *centralized*)
   The raw data or intermediate data should be transferred out of the end devices whatever architecture is chosen. It's inevitable to deal with privacy issues.

Model Training - Enabling Technologies

1. Federated learning
   Decentralized training without aggregating user private data.
2. Aggregation frequency control
   The optimization of communication overhead.
3. Gradient compression
   Use gradient quantization and gradient sparsification to compress the model update.
4. DNN splitting
   A DNN model is splitted inside between two successive layers with two partitions deployed on different locations without losing accuracy.
5. Knowledge transfer learning
   The structure of transfer learning is naturally fit for cloud/edge server (teacher) and edge/end device (student).
6. Gossip training
   Communicate with randomly selected partners.

Model Inference - Architectures

1. Edge-based (send data (features) from device to edge)
2. Device-based (perform the model inference locally)
3. Edge-device (intermediate result on device → edge, final result on edge → device)
4. Edge-cloud (data: device → edge → cloud, result: cloud → edge → device)



(a) Edge-based Mode (b) Device-based Mode (c) Edge-Device Mode (d) Edge-Cloud Mode

Model Inference - Key Performance Indicators

1. Test accuracy
   This is why AI been created.
2. Computation latency
3. Communication conjustion
4. Energy efficiency
5. Privacy
6. Memory footprint
   There is no dedicated high-bandwidth memory for mobile GPUs on mobile devices. Moreover, mobile CPUs and GPUs typically compete for shared and scarce memory bandwidth.

Model Training - Enabling Technologies

1. Model compression
2. Model partition
3. Model early-exit
   Trade-off between accuracy and computation cost.
4. Edge caching
   If the request from mobile devices hit the cached results stored in the edge server, the edge server will return the result directly.
5. Input filtering
   Remove the non-target-object frames of input data to avoid redundant computation.
6. Model selection (train a set of models and choose from it)
7. Support for multi-tenancy (resource allocation and task scheduling for concurrent applications)
8. Application-specific optimization (e.g. hardware acceleration)

Outline

1. Introduction
    - Motivation and Definition
    - Scope and Rating
2. A Primer on Artificial Intelligence
    - Deep Learning and Deep Neural Networks
    - Popular Deep Learning Models
3. Edge Intelligence Model Training and Inference
    - Model Training
    - Model Inference
4. Future Research Directions
5. Concluding Remarks

Future research directions

1. Programming and software platforms
   A common open standard that users can enjoy seamless and smooth services across heterogenegous Edge Intelligence platforms anywhere and anytime.
2. Resource-friendly Edge AI model design
   Resource-efficient edge AI models tailored to the hardware resource constraints of the underlying edge devices and servers.
3. Computation-aware networking technologies
4. Trade-off design with various DNN performance metrics
5. Smart service and resource management
6. Security and privacy issuses
7. Incentive and business models
   Proper incentive mechanism and business model are essential for stimulate effective and efficient cooperation among all members of EI ecosystem.

Outline

1. Introduction
   - Motivation and Definition
   - Scope and Rating
2. A Primer on Artificial Intelligence
   - Deep Learning and Deep Neural Networks
   - Popular Deep Learning Models
3. Edge Intelligence Model Training and Inference
   - Model Training
   - Model Inference
4. Future Research Directions
5. Concluding Remarks

Concluding Remarks

1. *Learning-driven communication* should be classfied into which scope? (future research directions)
2. What about *using AI technologies to solve optimization problems in edge computing*? (future research directions)
3. Where to put *hardware upgrading* (more powerful and customized CPU and GPU cores) of edge devices? (application-specific optimization)
4. What about new AI algorithms? New device-edge-cloud synergy frameworks? New hardware upgrading? Or intersection of them?
5. Only Deep Learning models can be considered as AI?
6. (I do not fully endorsed the classification on Edge Intelligence structure)
7. The division and future research dirctions are ambiguous