Considerations of deploying AI services in a distributed approach

draft-hong-nmrg-ai-deploy-02

Y-G. Hong (Daejeon Univ.), S-B. Oh (KSA), J-S. Youn (DONG-EUI Univ),
S-J. Lee (Korea University/KT), H-K. Kahng (Korea University)
S-W. Hong (ETRI), H-S. Yoon (ETRI)

nmrg Meeting@IETF 115 – London
November 7. 2022
History and status

– 1\textsuperscript{st} revision : draft-hong-nmrg-ai-deploy-00 (Mar. 2022)

– 2\textsuperscript{nd} revision : draft-hong-nmrg-ai-deploy-01 (Jul. 2022)
  • 1\textsuperscript{st} presentation

– 3\textsuperscript{rd} revision : draft-hong-nmrg-ai-deploy-02 (Oct. 2022)
  • 2\textsuperscript{nd} presentation
Updates after last meeting

– Reconfigure section 4. “Considerations for configuring a system to provide AI services”
  • 4.1. Considerations according to the functional characteristics of the Hardware
  • 4.2. Considerations according to the characteristics of the AI model
  • 4.3. Considerations according to the characteristics of the communication method

– Add a reference
  • ETSI "Mobile Edge Computing; Market Acceleration; MEC Metrics Best Practice and Guidelines" Group Specification ETSI GS MEC-IEG 006 V1.1.1 (2017-01)

– Add two authors
  • S-W. Hong (ETRI)
  • H-S. Yoon (ETRI)
**Motivations**

– Deployment of AI services
  • Focus: training (learning) -> inference (prediction)
  • For inference, not only high-performance servers, but also small hardware, microcontroller, low-performance CPUs, and AI chipsets are optimal target device (due to cost)

– Configuration of the system in terms of AI inference service
  • For training: accuracy of the model
  • For inference:
    • Target device: Local, edge, cloud
    • Objectives: Accuracy, Latency, Network traffic, Resource utilization, etc.
    • Considerations: AI model, Serving framework, Communication method, device capacity, inference data, etc.

– Accelerate the study AI issues and find some possible standardization items in the nmrg
Generic procedure of AI service

![Diagram of AI service workflow]

- Data collection & Store
- Data Analysis & Preprocess
- AI Model Training
- AI Model Deploy & Inference
- Monitor & Maintain Accuracy

Figure 1: AI service workflow
Network configuration structure to provide AI services

Figure 2: AI inference service on Local machine

Figure 3: AI inference service on Cloud server

Figure 4: AI inference service on Edge device
AI inference service on Cloud server and Edge device

Figure 5: AI inference service on Cloud server and Edge device
Considerations according to the functional characteristics of the hardware (1/2)

– (Reference) ETSI Group Specification MEC-IEG 006 V1.1.1 (2017-01) "Mobile Edge Computing; Market Acceleration; MEC Metrics Best Practice and Guidelines"

• It describes various metrics which can potentially be improved through deploying a service on a MEC platform
• It can be identified in order to highlight the benefits of deploying MEC for various services and applications
• Functional metrics
  • latency (both end-to-end, and one-way), energy efficiency, throughput, goodput, loss rate (number of dropped packets), jitter, number of out-of-order delivery packets, QoS, and MOS
• Non-functional metrics
  • service lifecycle (instantiation, service deployment, service provisioning, service update (e.g. service scalability and elasticity), service disposal), service availability and fault tolerance (aka reliability), service processing/computational load, global ME host load, number of API request (more generally number of events) processed/second on ME host, delay to process API request (north and south), number of failed API request
Considerations according to the functional characteristics of the hardware (2/2)

– The performance of AI inference service varies depending on how the hardware such as CPU, RAM, GPU, and network interface is configured for each cloud server and edge device.

– AI inference service can be deployed in the following locations
  • Distant cloud server : High performance and high cost
  • Near edge device : Medium performance and medium cost
  • Local machine : Low performance and low cost

– AI inference service result in (assumption: same AI model)
  • Distant cloud server : High accuracy, short inference time, and long delay to transmit
  • Near edge device : Medium accuracy, medium inference time, and medium delay to transmit
  • Local machine : Low accuracy, long inference time, and short delay to transmit
Considerations according to the characteristics of the AI model (1/2)

– Model size vs. Accuracy vs. Latency

[Source: Google Tensorflow]
Considerations according to the characteristics of the AI model (2/2)

–AI inference service can be deployed in the following locations
  • Distant cloud server : Heavy AI model, high accuracy, Big size, long inference time
  • Near edge device : Medium AI model, medium accuracy, medium size, medium inference time
  • Local machine : Light AI model, low accuracy, small size, short inference time

–AI inference serving framework
  • Traditional web server : ex) FastAPI, Flask, and Django
    • It can be operated on low performance machines
  • Specialized serving framework : ex) Tensorflow serving
    • It can provide high performance.
Considerations according to the characteristics of the communication method

- AI inference service can be utilized
  - Traditional REST method
    - Common and easily deployed
  - Specified communication method (e.g., gRPC)
    - Better performance but need some works

- AI Inference data can be classified
  - Real-time vs. Batch
  - Secure & non-secure
An example of AI system for Object detection services
Latency of object detection services in each device

LATENCY OF OBJECT DETECTION SERVICES IN EACH DEVICE

![Graph showing latency comparisons between local device, edge device, and cloud server]
Relationship to “Challenge document”

– In the NMRG, the “Challenge document” (draft-francois-nmrg-ai-challenges) is the main document for handling AI issues

– This draft is also related to the “Challenge document” and some texts can be added or merged
  • Distributed AI service
  • Lightweight AI service
  • Deployment of AI service

– This draft can be developed as a different document to focus on AI inference (Deployment of AI services)
  • The “Challenge document” includes many items
Thanks!!

Questions & Comments