AI Benchmarking Challenges, Methodology and Progress

Prof. Dr. Jianfeng Zhan

zhanjianfeng@ict.ac.cn or
jianfengzhan.benchcouncil@gmail.com

https://www.benchcouncil.org

ICT, Chinese Academy of Sciences & UCAS & BenchCouncil

AI4S@ Cluster 2021

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

BenchCouncil

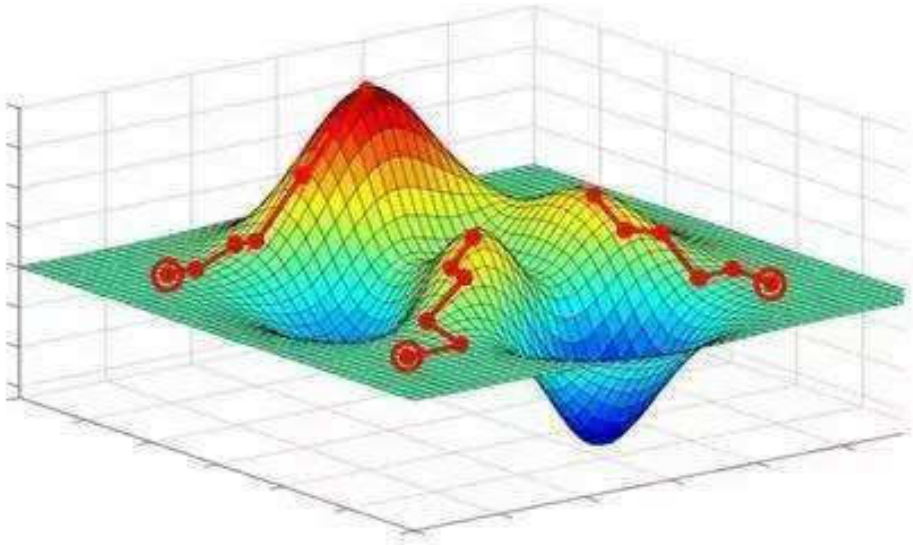
AIBench and ScenarioBench Contributors

ML/DL can do wonders

- : Approximating a high dimensional function
  - Image classification
  - Even traditional scientific computing
- Made possible by our ability to accurately approximate high dimensional functions using finite pieces of data.
- Opens up new possibilities for attacking problems that suffer from the “curse of dimensionality” (CoD)
  - As dimensionality grows, computational cost grows exponentially fast.
- Cited from Machine Learning and Computational Mathematics, Weinan E, 2021 presentation

Learning dynamics are not well understood

- High dimension non-convex optimization problem
  - A slight change leads to a different optimization path
  - Heavily dependent on the experience for parameter tuning



Other AI Benchmarking Challenges

- **Prohibitive cost**
- Metric issue
- Conflicting requirements in different stages
- Short shelf-life
- Scalability
- Repeatability

*[1] **AIBench Training: Balanced Industry-Standard AI Training Benchmarking.**[PDF]*
Fei Tang, Wanling Gao, Jianfeng Zhan, Chuanxin Lan and et al. 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021).

Prohibitive Cost Challenge

- Running an entire training session is mandatory!
  - Some optimizations improve the throughput but finally hurt the model quality.
- Take several weeks to run a complete training session on a small-scale system
  - Simulators with slowdowns 10 to 1,000 times exacerbate the challenge
- A microbenchmark like HPL-AI cannot model the learning dynamics of deep learning

[1] HPL-AI Mixed-Precision Benchmark — HPL-AI 0.0.2 documentation. https://icl.bitbucket.io/hpl-ai/

Other AI Benchmarking Challenges

- Prohibitive cost
- **Metric issue**
- Conflicting requirements in different stages
- Short shelf-life
- Scalability
- Repeatability

*[1] **AIBench Training: Balanced Industry-Standard AI Training Benchmarking.**[PDF]*
Fei Tang, Wanling Gao, Jianfeng Zhan, Chuanxin Lan and et al. 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021).

Metric Issue

- Time to quality
  - It heavily relies upon hyperparameter tuning
  - It is unfair to evaluate hardware and software systems
  - Need to decouple the architecture, system and algorithm evaluation.

- FLOPS
  - Half-Precision
  - Single-precision
  - Double-precision
  - Multi-precision
  - Mixed-precision
  - Not apple-to-apple comparison

Other AI Benchmarking Challenges

- Prohibitive cost
- Metric issue
- **Conflicting requirements in different stages**
- Short shelf-life
- Scalability
- Repeatability

*[1] **AIBench Training: Balanced Industry-Standard AI Training Benchmarking.**[PDF]*
Fei Tang, Wanling Gao, Jianfeng Zhan, Chuanxin Lan and et al. 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021).

Conflicting-requirement Challenge

- Overall
  - component benchmarks can not run on simulators.
  - Micro benchmarks are affordable but can not model learning dynamic.
- Earlier-stage evaluations of a new architecture or system
  - Affordable
  - Portability (Micro benchmarks)
  - Simplicity
- Later-stage evaluations or purchasing off-the-shelf systems
  - Comprehensiveness and representativeness
  - Overall system performance in reality

Other AI Benchmarking Challenges

- Prohibitive cost
- Metric issue
- Conflicting requirements in different stages
- **Short shelf-life**
- Scalability
- Repeatability

*[1] **AIBench Training: Balanced Industry-Standard AI Training Benchmarking.**[PDF]*
Fei Tang, Wanling Gao, Jianfeng Zhan, Chuanxin Lan and et al. 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021).

Short Shelf-life Challenge

- AI model evolutions and changes outpace the AI benchmarks.
  - One year to walk through benchmark design, implementation, community adoption, and large-scale testing.

- Synthetic benchmarks like ParaDNN [1] can traverse many networks, but cannot model learning dynamics.

[1] Wang, Yu Emma, Gu-Yeon Wei, and David Brooks.
"A Systematic Methodology for Analysis of Deep Learning Hardware and Software Platforms,"

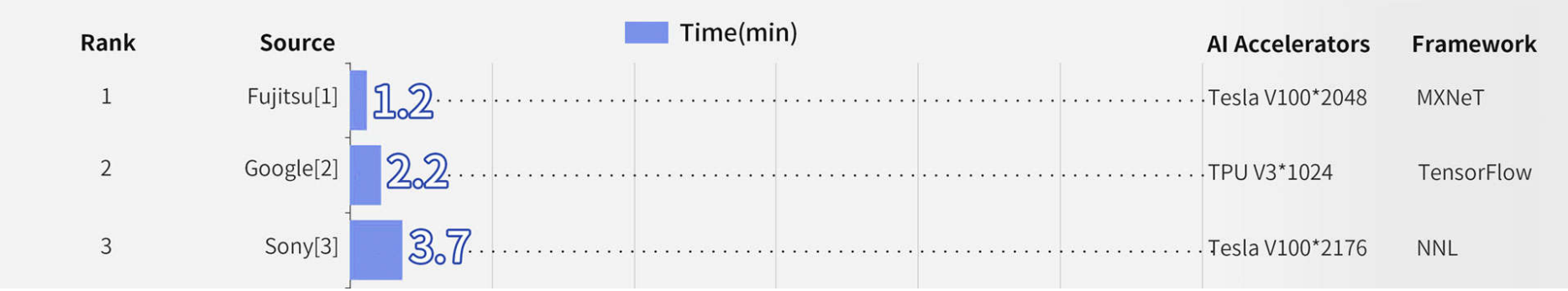
Other AI Benchmarking Challenges

- Prohibitive cost
- Metric issue
- Conflicting requirements in different stages
- Short shelf-life
- **Scalability**
- Repeatability

*[1] **AIBench Training: Balanced Industry-Standard AI Training Benchmarking.**[PDF]*
Fei Tang, Wanling Gao, Jianfeng Zhan, Chuanxin Lan and et al. 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021).

Scalability Challenge

- An AI task's problem scale is fixed, not scalable



Picture from HPC AI500 Ranking, Image Classification

HPL-AI [1] is scalable, but it cannot model the learning dynamics, and fail to consider the model quality. LU decomposition is irrelevant to DL workloads. There are may kernel in AI workloads.

[1] HPL-AI Mixed-Precision Benchmark — HPL-AI 0.0.2 documentation. https://icl.bitbucket.io/hpl-ai/
[2] HPC-AI500 Ranking. *https://www.benchcouncil.org/ranking.html*

Other AI Benchmarking Challenges

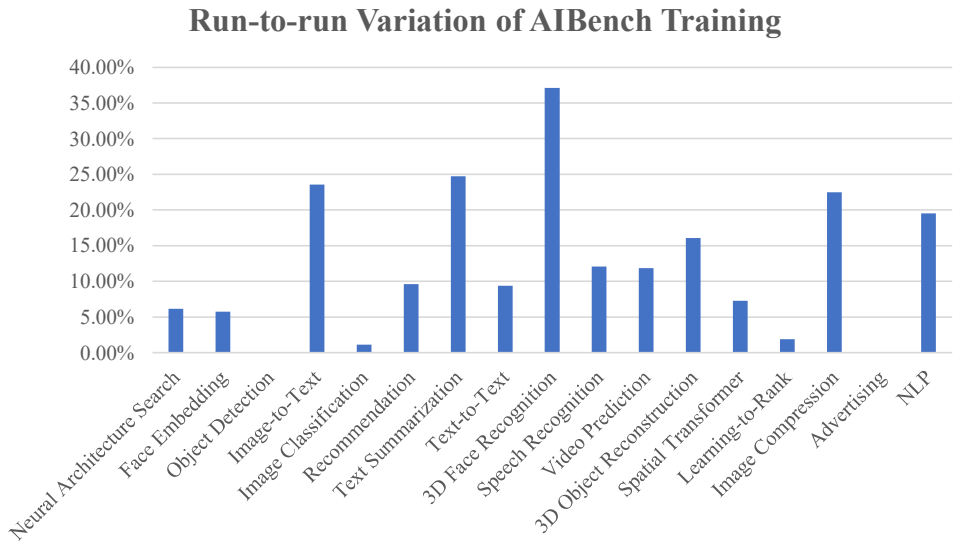
- Prohibitive cost
- Metric issue
- Conflicting requirements in different stages
- Short shelf-life
- Scalability
- **Repeatability**

*[1] **AIBench Training: Balanced Industry-Standard AI Training Benchmarking.**[PDF]*
Fei Tang, Wanling Gao, Jianfeng Zhan, Chuanxin Lan and et al. 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021).

Repeatability Challenge

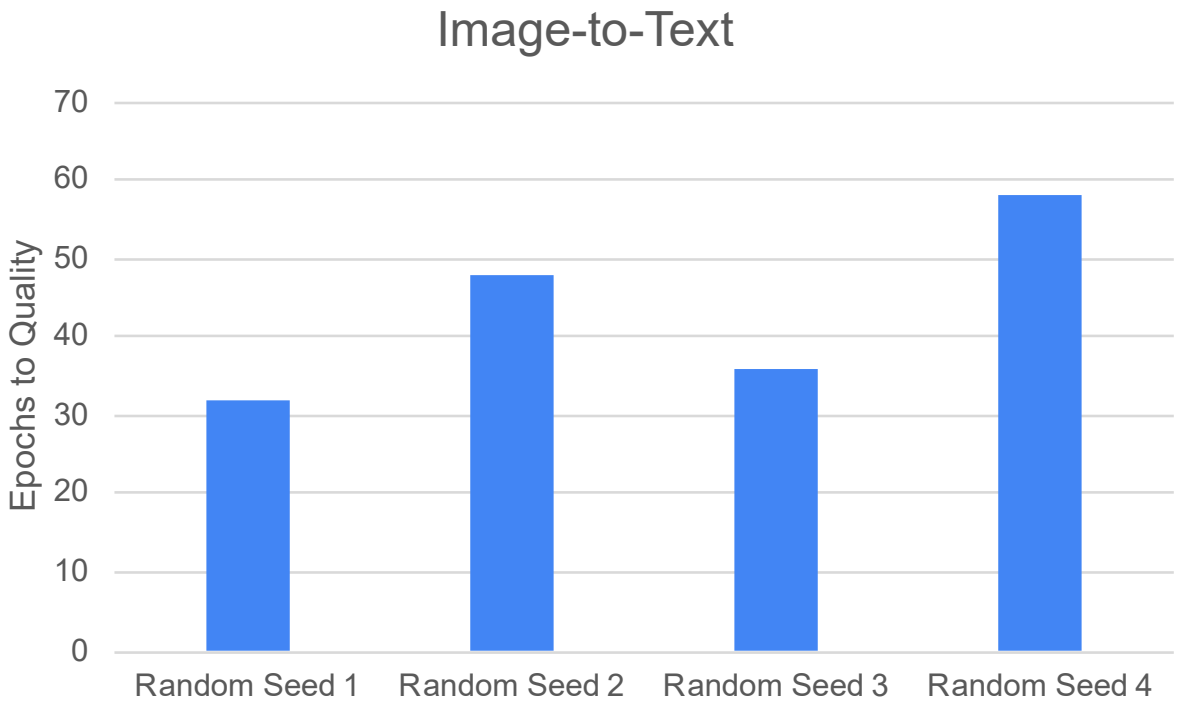
- The benchmark mandates being repeatable, while training deep networks is stochastic

Factors of randomness:
- Model initialization
- Data augment
- Data shuffle
- Dropout
- Etc.



Example: Randomness

- The epochs to achieve target quality vary significantly under different random seeds



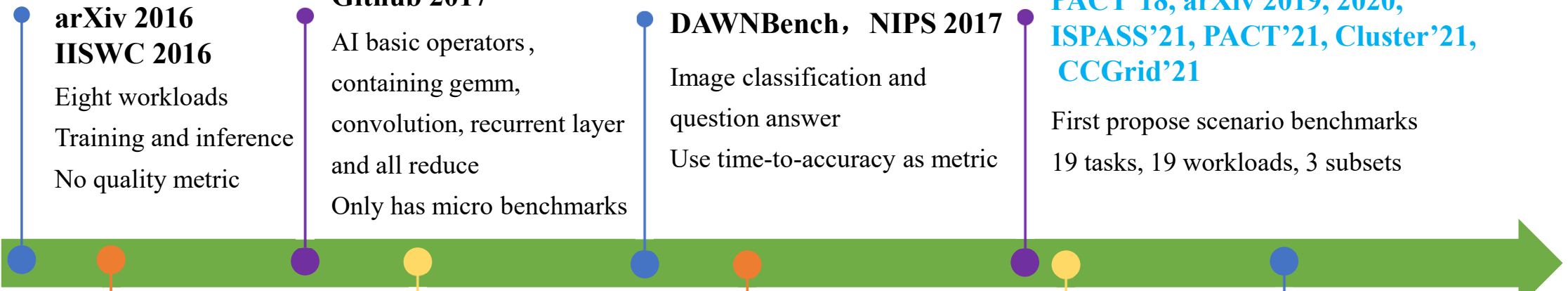
Run Image-to-Text from AIBench Training four times using different random seeds

[1]F. Tang et al., "AIBench Training: Balanced Industry-Standard AI. Training Benchmarking", ISPASS 2021

Overview

- AI Benchmarking Challenges
- **Related Work**
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

AI Benchmarks



Fathom
arXiv 2016
IISWC 2016
Eight workloads
Training and inference
No quality metric

DNNMark
GitHub 2016
GPGPU 2017
Eight micro benchmarks

DeepBench，Baidu
Github 2017
AI basic operators，
containing gemm,
convolution, recurrent layer
and all reduce
Only has micro benchmarks

BenchIP
arXiv 2017
JCST 2018
10 microbenchmarks
11 neural network
models

DAWNBench，NIPS 2017
Image classification and
question answer
Use time-to-accuracy as metric

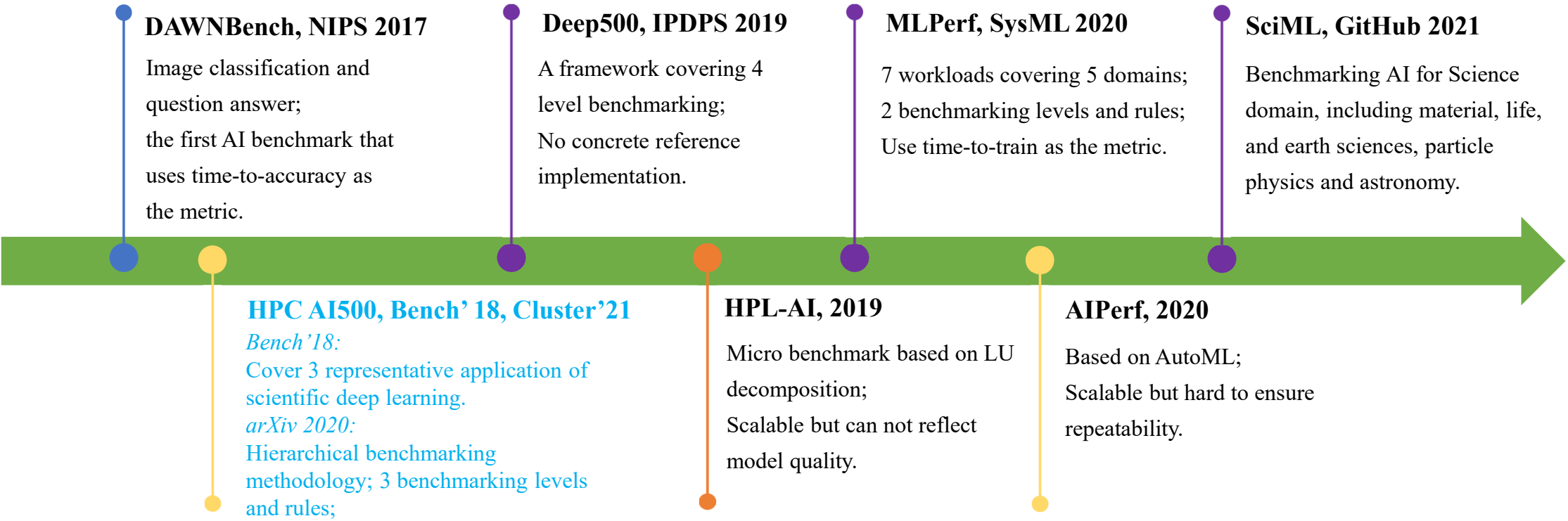
TBD Suite
GitHub 2018
IISWC 2018
Eight workloads，
six domains

AIBench，Bench'18, IISWC'18,
PACT'18, arXiv 2019, 2020,
ISPASS'21, PACT'21, Cluster'21,
CCGrid'21
First propose scenario benchmarks
19 tasks, 19 workloads, 3 subsets

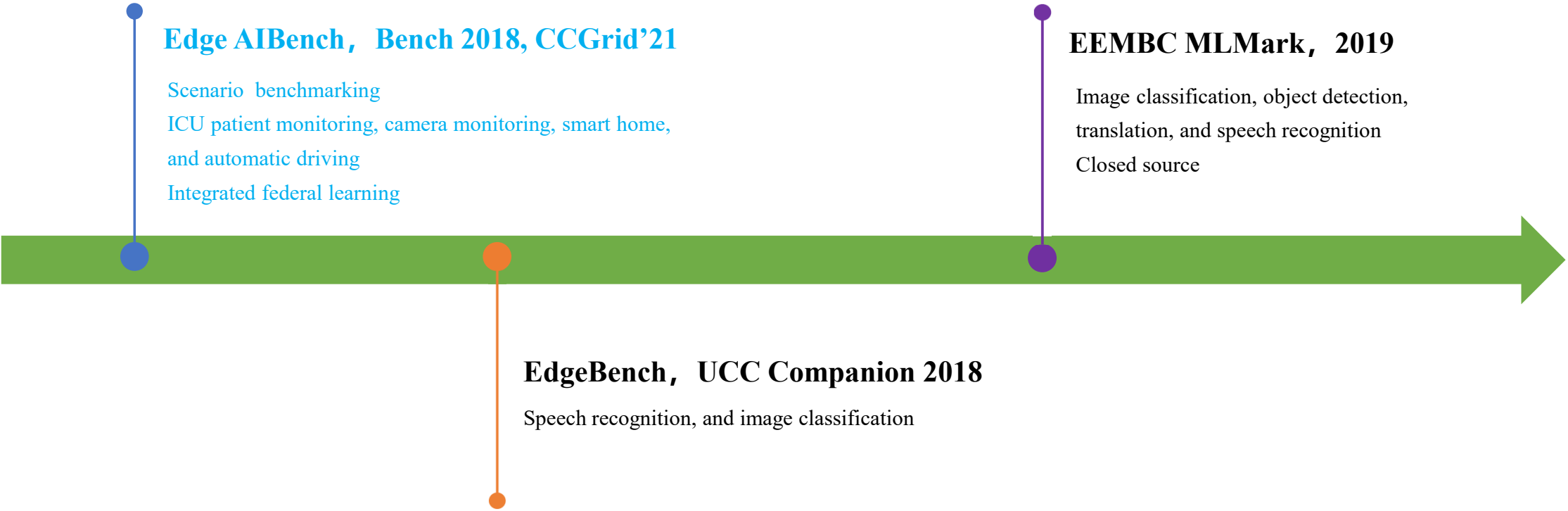
MLPerf, 2018
GitHub 2019
SysML 2020
Five domains
seven workloads

AIIA-DNN
GitHub 2019
Designed to support
training and inference, but
only provides inference
implements now

HPC AI Benchmarking



AI Benchmarks for Edge Computing



AI Benchmarks for IoT

AIoTBench，Bench 2018

Vision, audio, and NLP domain

Supports Android and Raspberry Pie

TensorFlow Lite、Caffe 2

End-to-end、microbenchmarks

ETH Zurich AI Benchmark，ECCV 2018

Only supports vision domain

Only supports Android and TensorFlow Lite

End-to-end

Summary of Related Work

- DawnBench (2017): the first benchmark that proposes time-to-accuracy as the primary metric.
  - I question this metric.
- AIBench Scenario (2018): the first benchmark modeling the critical paths of a real-world application scenario.
- ParaDNN (2020) is the first synthetic AI benchmark.
- HPL-AI (2019) is a micro benchmark that uses mixed-precision LU decomposition to achieve upper bound FLOPS performance.
  - scalable, but LU decomposition not relevant to most of the AI workloads.
- Both ParaDNN and HPL-AI can not model the learning dynamics, and also fail to consider the model quality.

Summary of Related Work

- AIBench and MLPerf are two systematic AI benchmarking projects.
  - They are concurrent and complemental.
- The AIBench suites are by far the most comprehensive AI benchmark suites tackling Challenges #1-5.
  - prohibitive cost
  - conflicting requirements in different stages
  - short shelf-life and fast evolution of AI models
  - scalability challenge due to the fixed problem scale
  - repeatability challenge due to the stochastic nature of AI.
- MLPerf (2019) includes seven benchmarks for training and five benchmarks for inference.
- MLPerf performs the most large-scale testing, but fails to present a benchmarking methodology to justify the choice for and update AI tasks, models, and data sets.
- MLPerf fails to consider the conflicting requirements, shelf-life, scalability challenges (Challenges #2-4).

Overview

- AI Benchmarking Challenges
- Related Work
- **ScenarioBench & AIBench Methodologies**
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

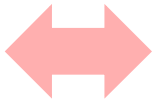
BenchCouncil AIBench & ScenarioBench

- **International Open Benchmark Council (BenchCouncil)**
  - **https://www.benchcouncil.org**
  - a non-profit international organization
    - aims to promote standardizing and incubating Big Data, AI, Chip and other emerging technology.

- **AIBench**
  - **https://www.benchcouncil.org/aibench**

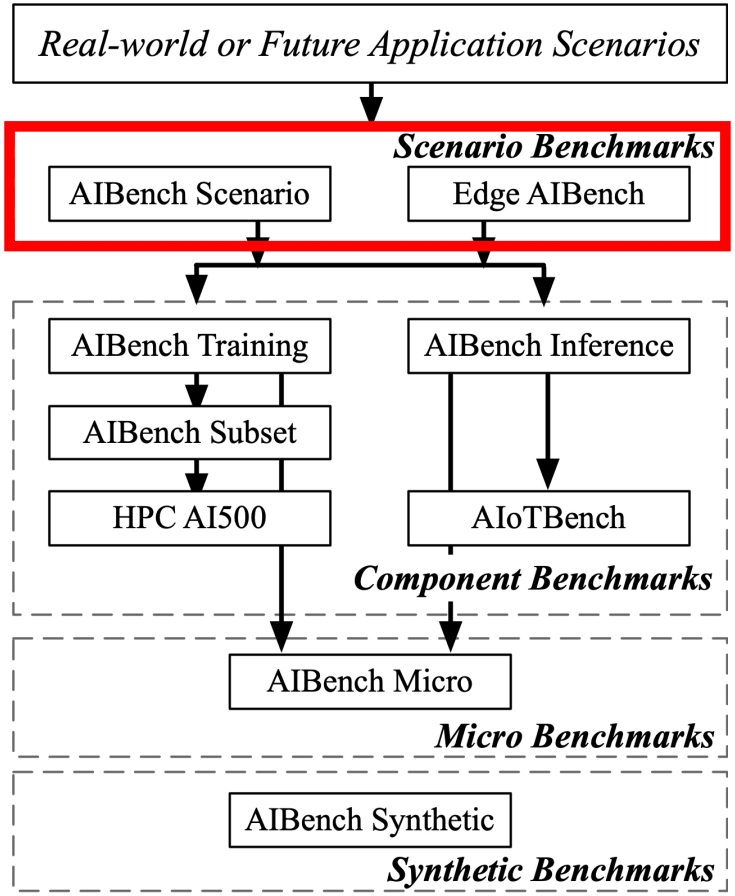
- **ScenarioBench**
  - **Modeling, Characterizing, and Optimizing Ultra-scale Real-world or Future Applications and Systems Using Benchmarks**
  - https://www.benchcouncil.org/scenariobench

ScenarioBench and AIBench Summary

Scenario	 AIBench Scenario   Edge AIBench
Training	 AIBench Training   AIBench Subset  HPC AI500
Inference	 AIBench Inference   AIoTBench
Micro	 AIBench Micro Benchmarks
Synthetic	 AIBench Synthetic

ScenarioBench Methodology

- The goal of ScenarioBench is to propose methodology, tools, and metrics to model, characterize, and optimize ultra-scale real-world or future applications and systems using the benchmarks.
  - AIBench Scenario is a benchmarks suite modeling AI-augmented Internet service scenarios
  - Edge AIBench models end-to-end performance across IoT, edge, and Datacenter.

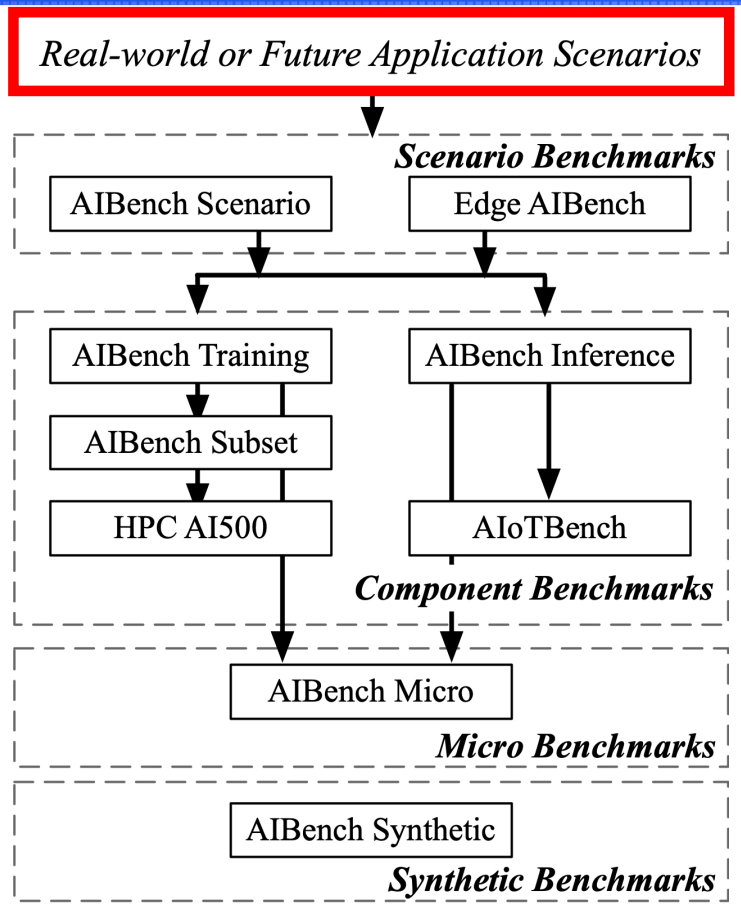


AIBench Methodology

AIBench Methodology

- As a joint work with increasing industry partners, AIBench is a comprehensive AI benchmark project focusing on methodology, frameworks, and continuous improvements.

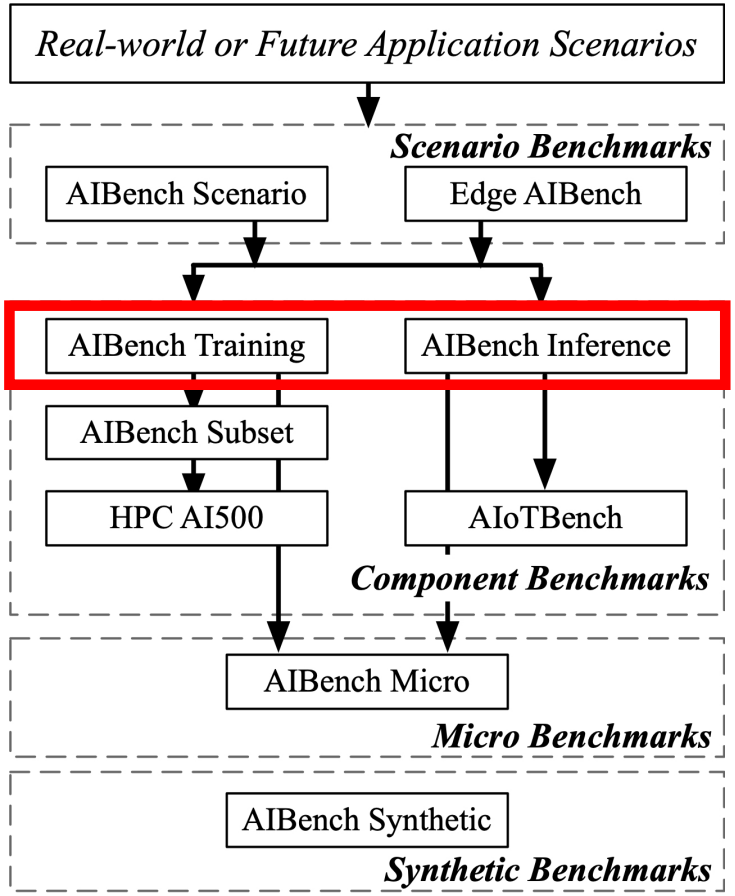
- Tackle the challenges #1-5 mentioned above in a systematic manner.



AIBench Methodology

AIBench Methodology

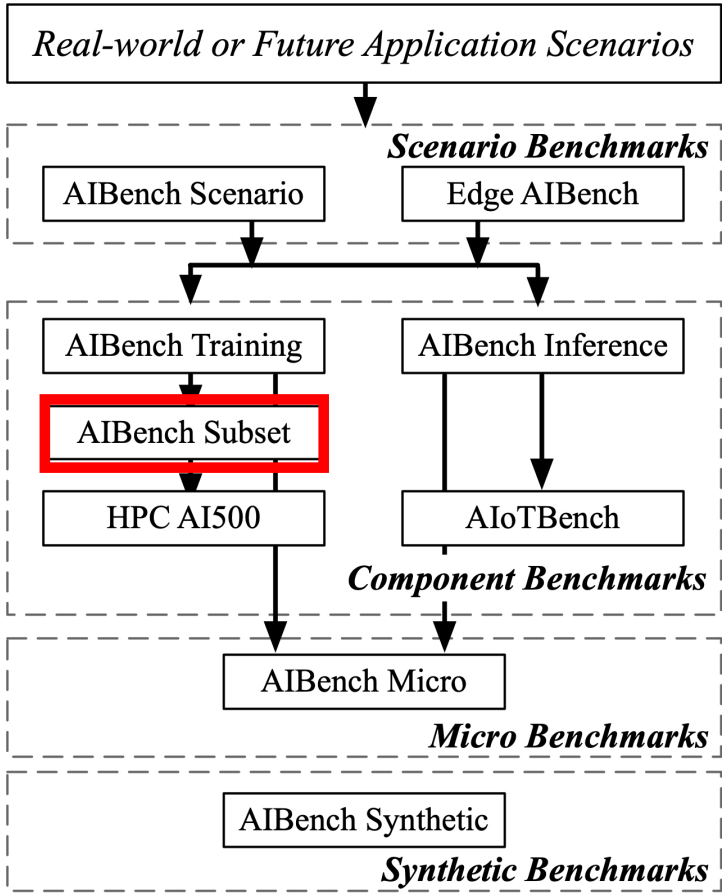
- AIBench Training and AIBench Inference cover nineteen representative AI tasks (will update) with state-of-the-art models to guarantee diversity and representativeness



AIBench Methodology

AIBench Methodology

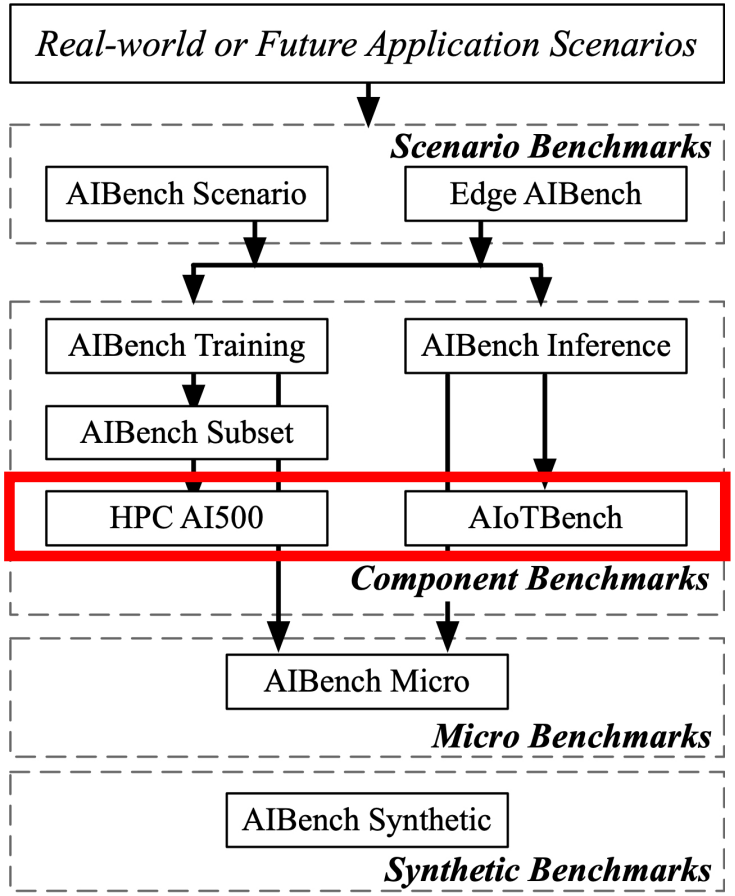
- AIBench Training provides two subsets for repeatable performance ranking (RPR) subset and workload characterization (WC) subset to improve affordability



AIBench Methodology

AIBench Methodology

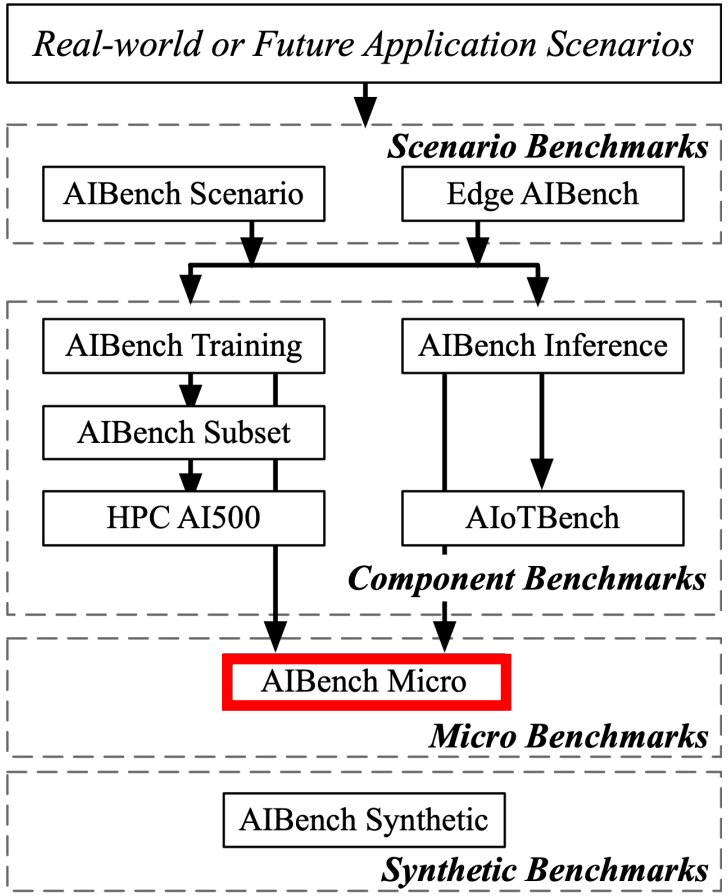
- To evaluate large-scale HPC AI systems, HPC AI500 is derived from the AIBench Training RPR subset
- To evaluate various IoT and embedded devices, AIoTBench is derived from AIBench Inference



AIBench Methodology

AIBench Methodology

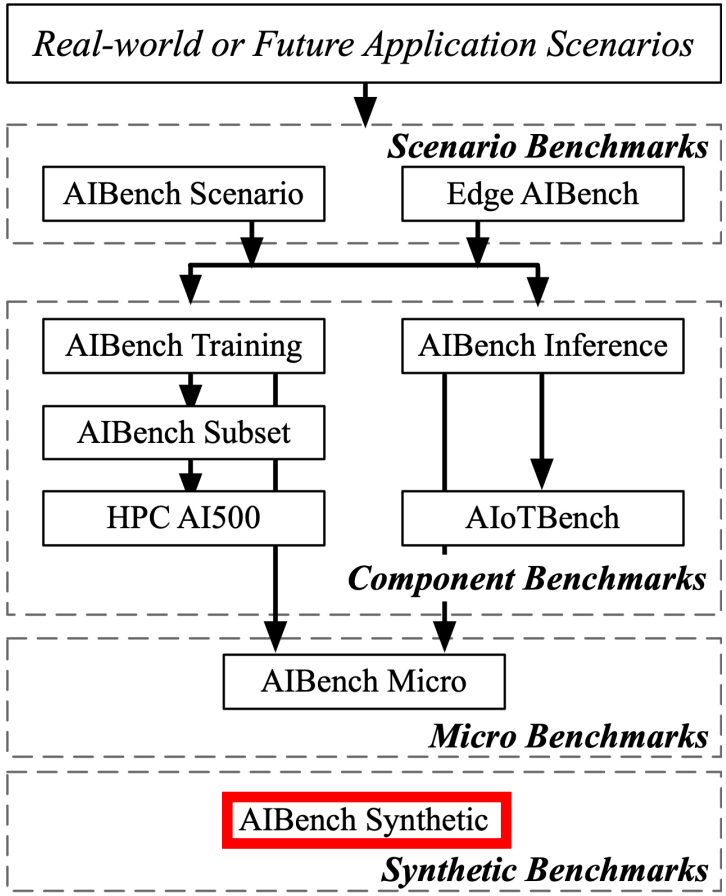
- AIBench Micro provides the intensively-used hotspot functions, profiled from AIBench Training and Inference, for simulation-based architecture researches
  - Different simuation (simulating diverse accelerator architectures) versions



AIBench Methodology

AIBench Methodology

- As complementary to real-world benchmarks, AIBench Synthetic provides several synthetic benchmarks with scalable problem sizes that can model learning dynamics.



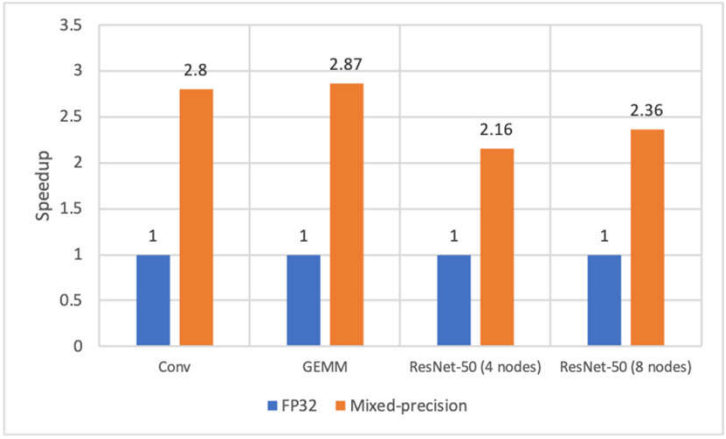
AIBench Methodology

Overview

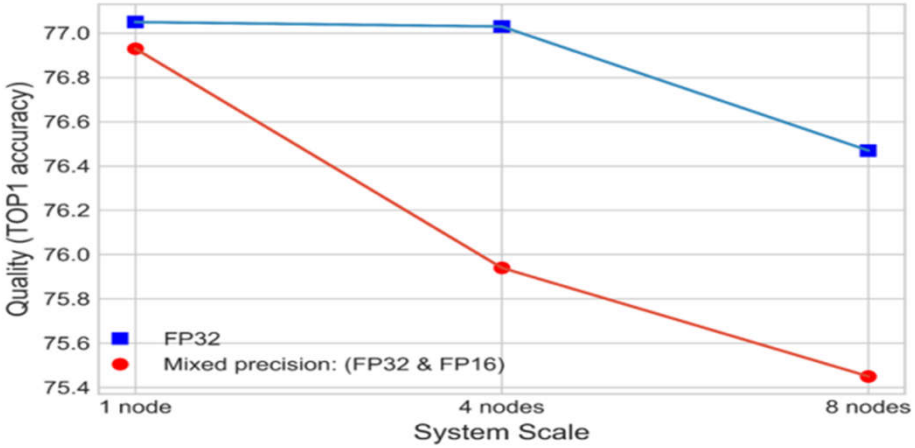
- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- **ScenarioBench**
  - **AIBench Scenario**
  - Edge AIBench
- AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

Is Micro Benchmark Sufficient ?

- AI workloads need to consider both computational efficiency and model quality
  - FLOPS is no longer the only metric
- Mixed-precision training significantly improve FLOPS, however, it may deteriorate the model quality



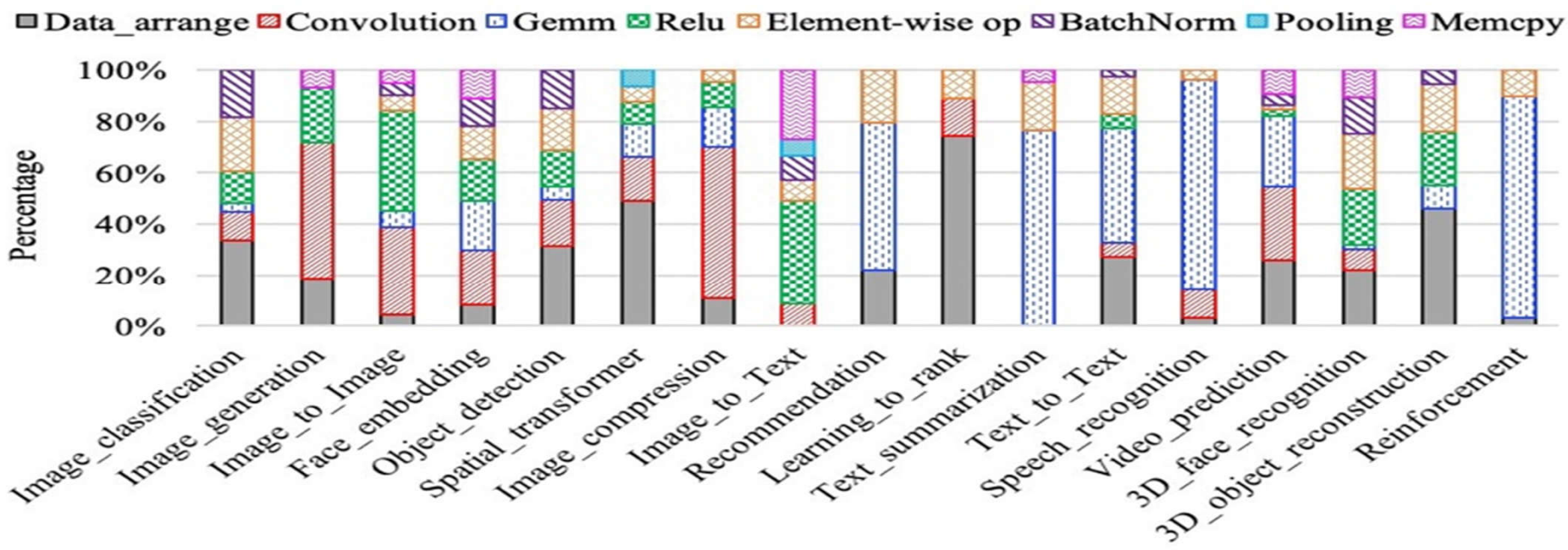
FLOPS comparison of ResNet50 model and operators



The ResNet50 quality comparison

Example: No Single Kernel

- The kernels' runtime breakdown of 17 AI workloads
  - Some micro benchmarks may occupy a little percentage



Is Component Benchmark Sufficient ?

Benchmarking with a Single Component

Image-to-Text
OCR

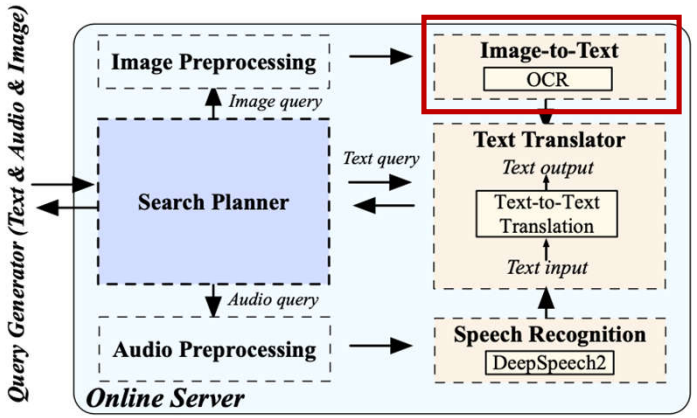
OCR Model Pruning

Tail latency shortens by **51%**

Accuracy decreases **1%**

Optimization Technique

*Putting the Component into a realistic scenario: Online Translation Intelligence (**A scenario benchmark from AIBench**)*



OCR Model Pruning

Tail latency has no improvement

Accuracy decreases **1%**

Inconsistent conclusion

Only a single component may lead to error-prone conclusions

Single Component vs. Realistic Application

- ***E-commerce Search Intelligence*** ***(A scenario benchmark from AIBench)***

- **The overall system tail latency deteriorates even 100X comparing to a single component tail latency**
  - **2.2X** comparing to recommendation component
  - **180X** comparing to text classification component

Benchmarking with a single component cannot reflect the overall system's effects

Model Accuracy vs. QoS

- For *E-commerce Search Intelligence (**a scenario benchmark from AIBench**)*
  - Replace ResNet50 with ResNet152 for image classification
  - Model accuracy improvement **1.5%** => overall system 99th percentile latency deteriorates by **9.7X**
    - Overall system 99th percentile latency
      -

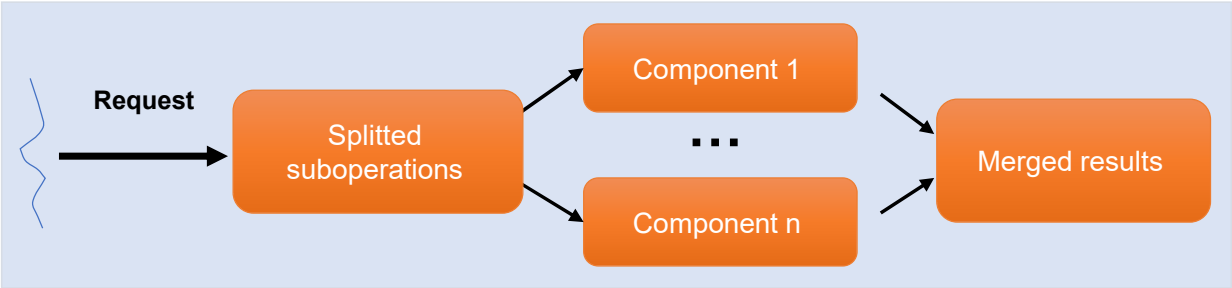
Benchmarking with a single component cannot reflect the tradeoff between model accuracy and QoS

Statistical Model + Component Benchmarks ?

- Does a statistical model predict the overall system tail latency through profiling many components' tail latency performance?
  - NO!
- A simple queueing model
  - E-commerce Search Intelligence **(a scenario benchmark from AIBench)**
    - **8.6X** between the actual average latency and the theoretical one
    - **3.3X** between the actual 99th percentile latency and the theoretical one
- A sophisticated queueing network model
  - E-commerce Search Intelligence **(a scenario benchmark from AIBench)**
    - **4.9X** between the actual average latency and the theoretical one
    - **Difficult** for tail latency predicting: non-superposition property

AIBench Scenario is needed !

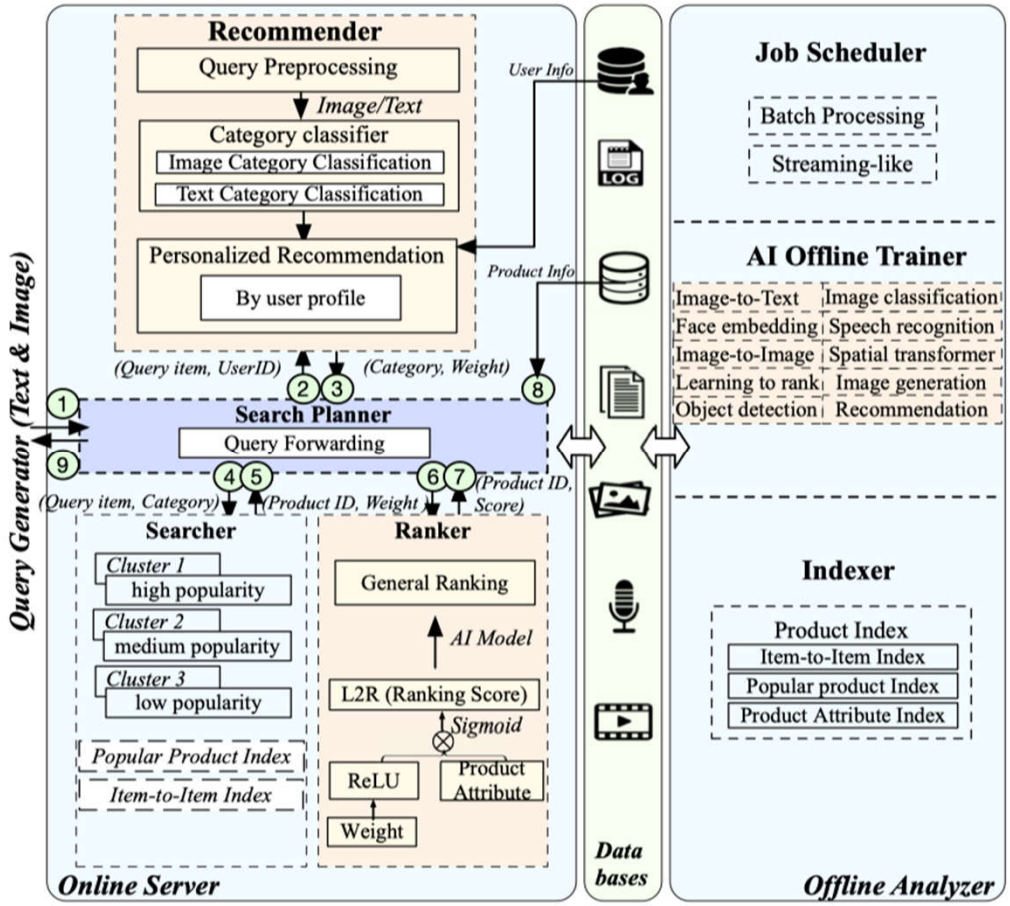
- AIBench Scenario: http://www.benchcouncil.org/aibench/scenario/
  - A proxy of a realistic application scenario
  - The real one is treated as first-class confidential issues
- Capturing the critical path and primary modules
  - The **permutations** of a series of AI and non-AI components



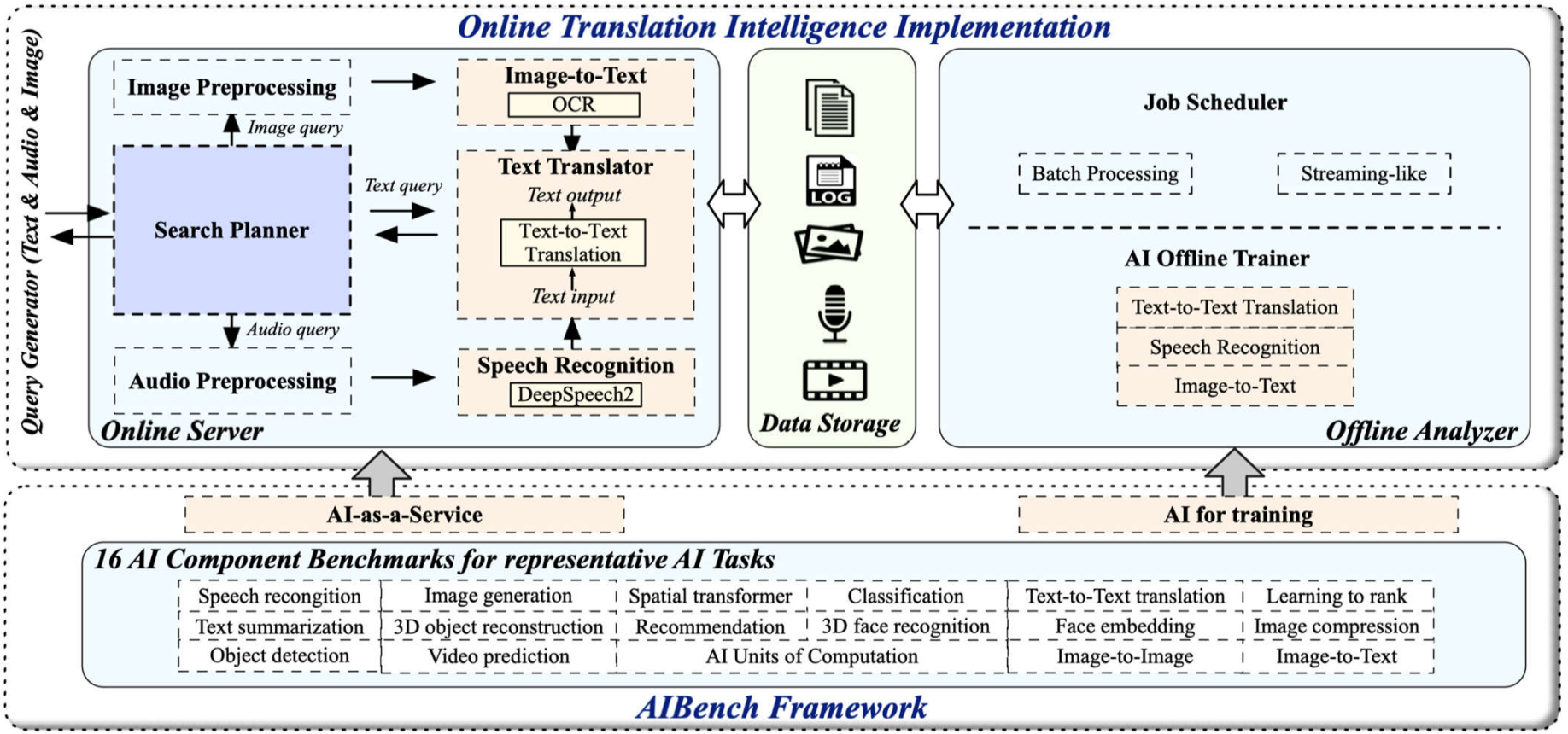
Wanling Gao, Fei Tang, Jianfeng Zhan, Xu Wen, Lei Wang, Zheng Cao, Chuanxin Lan, Chunjie Luo and Zihan Jiang. AIBench Scenario: Scenario-distilling AI Benchmarking. PACT 21

Scenario Benchmark: E-commerce Search Intelligence

- **Query generator**
  - simulate concurrent users and send query requests
- **Online module**
  - personalized searching and recommendations
- **Offline module**
  - a training stage to generate a learning model
- **Data storage module**
  - data storage, e.g., user database, product database



Scenario Benchmark: Online Translation Intelligence



Overview

- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- **ScenarioBench**
  - AIBench Scenario
  - **Edge AIBench**
- AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

Edge AIBench

- Modeling end-to-end performance across IoT, edge, and Datacenter

Edge AIBench

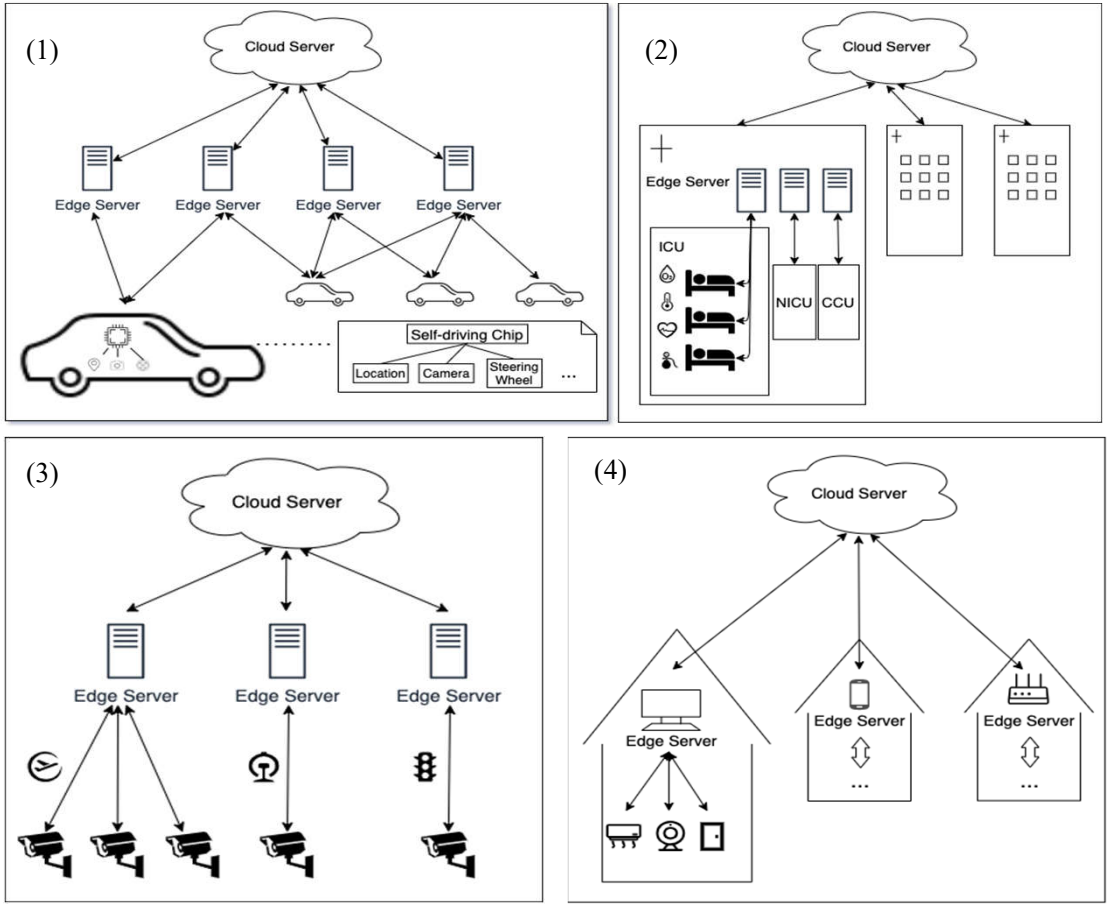
AI Benchmarks for Edge, BenchCouncil

https://www.benchcouncil.org/aibench/edge-aibench/index.html

- Publication
  - *Edge AIBench: towards comprehensive end-to-end edge computing benchmarking.*
    - *Tianshu Hao, Yunyou Huang, Xu Wen, Wanling Gao, Fan Zhang, Chen Zheng, Lei Wang, Hainan Ye, Kai Hwang, Zujie Ren, and Jianfeng Zhan. Bench'18*
    - *https://arxiv.org/pdf/1908.01924.pdf*
  - *AI-oriented Medical Workload Allocation for Hierarchical Cloud/Edge/Device Computing. [PDF]*
    - *Tianshu Hao, Jianfeng Zhan, Kai Hwang, Wanling Gao, Xu Wen. The 21st IEEE/ACM international Symposium on Cluster, Cloud and. Internet Computing (CCGrid 2021).*
  - *Edge AIBench Specification*
    - *https://www.benchcouncil.org/file/EdgeAIBench_Specification.pdf*

Four Typical Edge AI Scenarios

- (1) Autonomous Vehicle
  - High-accuracy, Latency-sensitive
  - Device Mobility
- (2) ICU Patient Monitor
  - Latency-sensitive (msec level)
  - May tolerate some errors
  - Parallel, massive patients
- (3) Surveillance Camera
  - Enormous Data
- (4) Smart Home
  - Heterogenous devices and data



Nine Typical Edge AI Tasks

Task Name	Edge AI Scenarios	Models	Datasets	Implementations
Lane Detection	Autonomous Vehicle	LaneNet	Tusimple/ CULane	Pytorch/Caffe
Traffic Sign Detection	Autonomous Vehicle	Capsule Network	German Traffic Sign Recognition Benchmark	Keras
Heart Failure Prediction	ICU Patient Monitor	LSTM	MIMIC-III	Tensorflow/Keras
Decompensation Prediction	ICU Patient Monitor	LSTM	MIMIC-III	Tensorflow/Keras
Death Prediction	ICU Patient Monitor	LSTM	MIMIC-III	Tensorflow/Keras
Person Re-identification	Surveillance Camera	DG-Net	Market-1501	Pytorch
Action Detection	Surveillance Camera	ResNet18	UCF101	Pytorch/Caffe
Face Recognition	Smart Home	Facenet/Sphere network	LFW/CASIA-Webface	Tensorflow/Caffe
Speech Recognition	Smart Home	DeepSpeech2	LibriSpeech	Tensorflow

Overview

- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- **AIBench**
  - **AIBench Training**
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

Consider Conflicting Benchmarking Requirements

- Benchmarking at different stages
  - Earlier-stage evaluations of a new architecture or system：
    - Portability （Micro benchmarks）
    - Simplicity
  - Later-stage evaluations or purchasing off-the-shelf systems：
    - Comprehensiveness/Representativeness
    - Performance in reality (Component or scenario benchmarks)

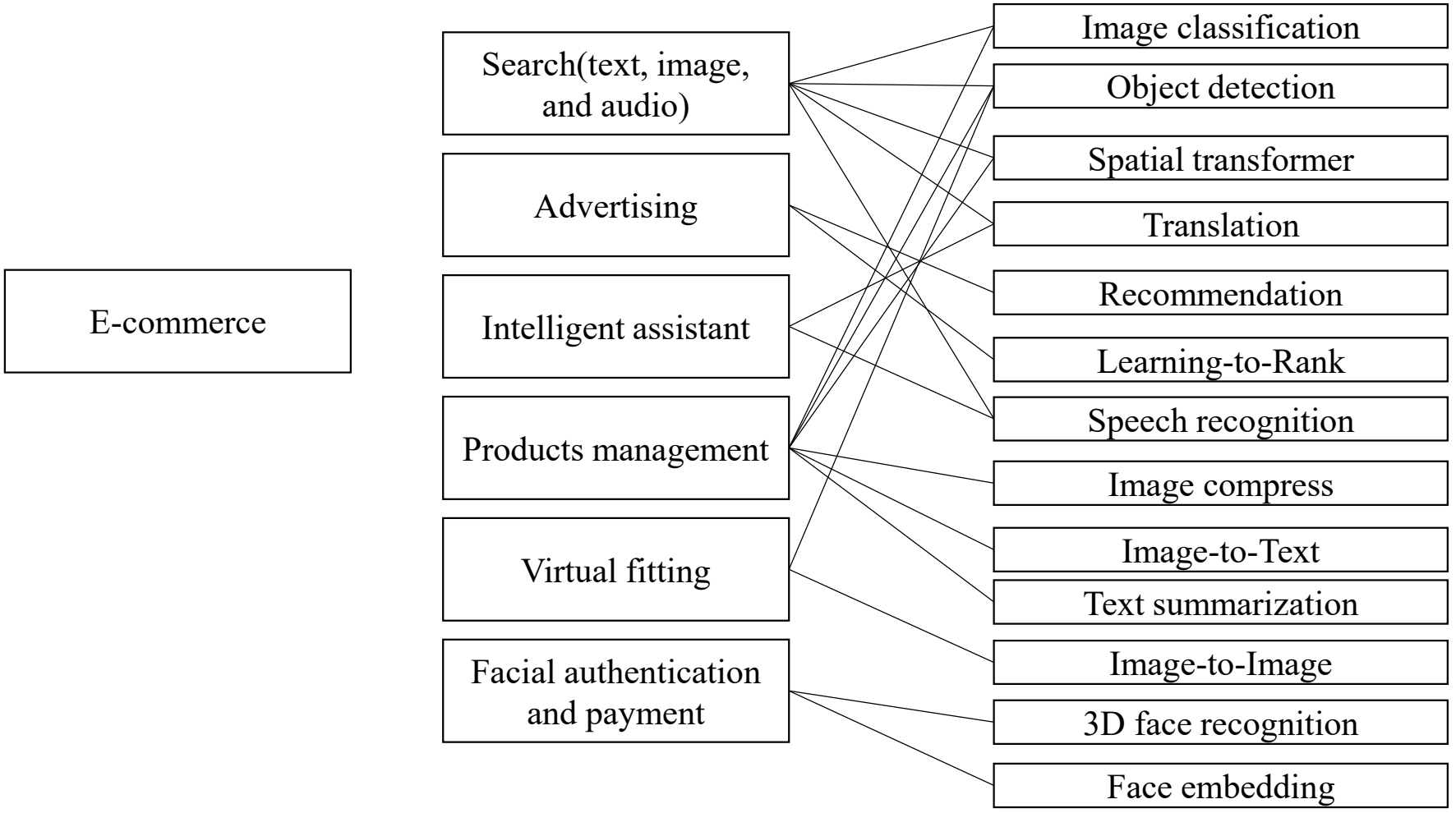
Representativeness and Comprehensiveness

- Diverse behaviors for workload characterization
  - Micro-architecture level
    - FLOPs, memory access pattern, computation pattern, and I/O pattern
  - System level
    - Time to quality, run-to-run variation (the ratio of the standard deviation to the mean of epochs to quality), convergence rate (epochs to quality), and number of hot functions
  - Algorithm level
    - Model architectures and parameters

AIBench Training Workloads

- Coverage of diverse network architectures（CNN、ResNet、LSTM、GRU、Attention, etc.）
  - **Text processing (7)**
    - Text-to-Text, Text summarization, Learning to Rank, Recommendation, Neural Architecture Search, Advertising, Nature Language Processing (NLP)
  - **Image processing (8)**
    - Image Classification, Image Generation, Image-to-Text, Image-to-Image, Face Embedding, Object Detection, Image Compression, Spatial Transformer
  - **Audio processing (1)**
    - Speech Recognition
  - **Video processing (1)**
    - Video Prediction
  - **3D data processing (2)**
    - 3D Face Recognition, 3D Object Reconstruction

Take E-commerce as an Example



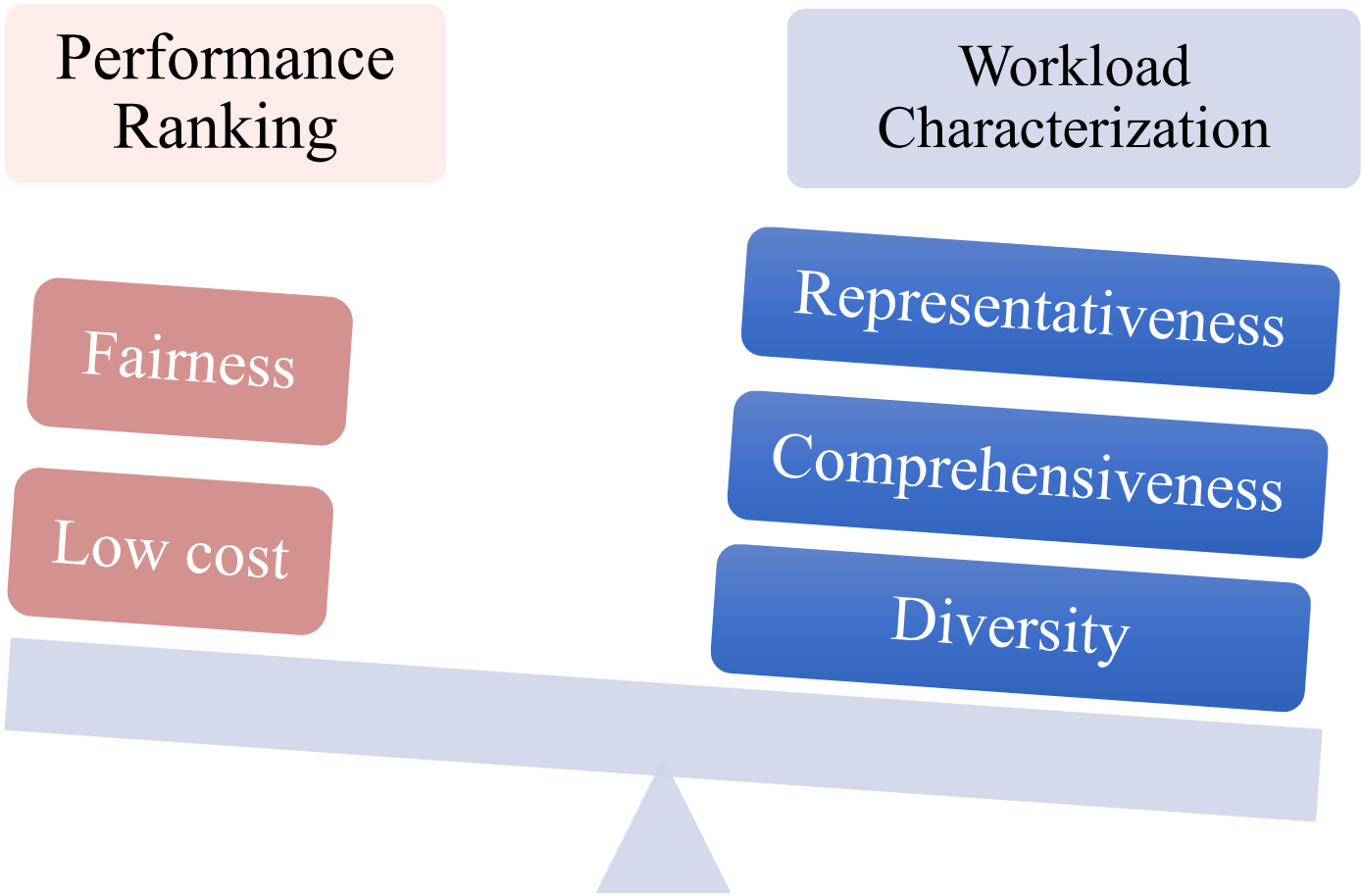
Representativeness

		AIBench Training v1.1	MLPerf Training v0.7
Methodology		Balanced methodology considering conflicting requirements	According to commercial and research relevance
Task		Nineteen tasks and models	six tasks and eight models
Dataset		Text, image, 3D, audio, and video data	Text and image data
Algorithm behavior	Computation	0.09 to 282830 MFLOPs	0.21 to 29000 MFLOPs
	Complexity	0.03 to 110 million parameters	5.2 to 110 million parameters
	Optimizer categories	5	5
	Loss function categories	14	6
System behavior	Hotspot functions	30	9
	Convergence	6 to 96 epochs	3 to 49 epochs
Micro-architecture behavior	Achieved occupancy	0.12 to 0.61	0.12 to 0.54
	IPC efficiency	0.02 to 0.77	0.02 to 0.74
	Gld efficiency	0.28 to 0.94	0.52 to 0.85
	Gst efficiency	0.27 to 0.98	0.75 to 0.98
	DRAM utilization	0.08 to 0.61	0.08 to 0.61

Overview

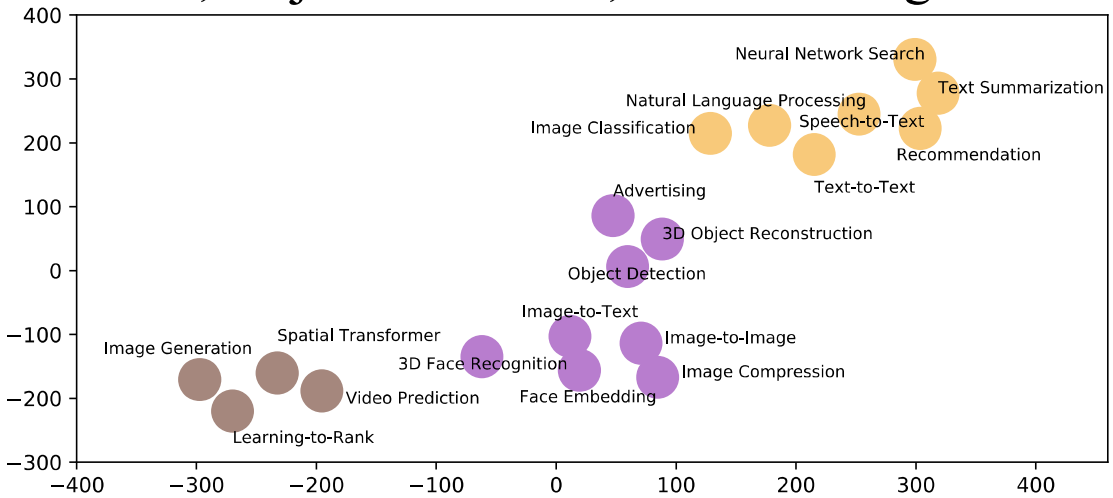
- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- **AIBench**
  - AIBench Training
    - **AIBench Subset**
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

Requirements for Ranking



Two Subsets for Affordability

- Subset for repeatable performance ranking (RPR subset)
  - Image Classification, Object Detection, and Learning to Rank



- Subset for workload characterization (WC subset)
  - Spatial Transformer, Image-to-Text, and Speech-to-Text
    - the nearest to the centroid of three clusters, respectively

Overview

- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- **AIBench**
  - AIBench Training
    - AIBench Subset
    - **HPC AI500**
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

HPC AI500 Benchmarking Methodology

- The criteria for choosing the workloads.
  - Representativeness and Affordability
  - Repeatability
  - Computation complexity
  - Tasks, Models, and datasets
  - Scalability
- AIBench RPR subset satisfies repeatability, representativeness, and affordability.

- HPC AI500 V2.0: The Methodology, Tools, and Metrics for Benchmarking HPC AI Systems. Zihan Jiang, Wanling Gao, Fei Tang, Lei Wang, Xingwang Xiong, Chunjie Luo, Chuanxin Lan, Hongxiao Li, Jianfeng Zhan. **CLUSTER'2021**.
- HPC AI500: A Benchmark Suite for HPC AI Systems. Zihan Jiang, Wanling Gao, Lei Wang, Xingwang Xiong, Yuchen Zhang, Xu Wen, Chunjie Luo, Hainan Ye, Xiaoyi Lu, Yunquan Zhang, Shengzhong Feng, Kenli Li, Weijia Xu, and Jianfeng Zhan. **Bench'18.**

Scalability Requirement

- AIBench subset computation comparison (Single training batch).

Workloads	Computation （FLOPs）
Image Classification	23 G
Object Detection	691 G
Learning to Rank	0.08 M

Image Classification and **Object Detection** meet the computation requirement and are chosen as two typical workloads for HPC AI benchmarking.

Problem Domain， Dataset, and Model of HPC AI500

- Problem Domain
  - Extreme Weather Analysis: detect the patterns of extreme weather, essentially Object Detection. The application that wins Gordon Bell Prize.
  - Image Classification： ResNet50/ImageNet is a de facto benchmark for optimizing HPC AI systems.
- Dataset
  - The extreme weather dataset: 16 channels, 768*1052, 2 TB
  - ImageNet 2012: 3 channels, 256*256, 136 GB
- Model
  - Faster-RCNN
  - ResNet-50 V1.5

Being Consider Other Workloads

- quantum many-body problems
- classical many-body problem, e.g. protein folding
- turbulence
- solid mechanics (plasticity, nonlinear elasticity) control .....
- multi-scale modeling (gas dynamics, combustion, non-Newtonian fluids, etc)

- Can machine learning help?
- Can the success of ML be extended beyond traditional AI ?

- Cited from Machine Learning and Computational Mathematics, Weinan E, 2021 presentation

Overview

- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- **AIBench**
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - **AIBench Inference**
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

Why Comprehensive Inference Workloads?

- Comprehensive workloads are not a burden !
  - Inference time is much shorter
- Diversity of data types, AI models, AI frameworks
- Diversity of workload behaviors
  - Algorithm, System, Architecture

AIBench Inference

- Coverage of diverse network architectures（CNN、ResNet、LSTM、GRU、Attention, etc.）
  - **Text processing (7)**
    - Text-to-Text, Text summarization, Learning to Rank, Recommendation, Neural Architecture Search, Advertising, Nature Language Processing (NLP)
  - **Image processing (8)**
    - Image Classification, Image Generation, Image-to-Text, Image-to-Image, Face Embedding, Object Detection, Image Compression, Spatial Transformer
  - **Audio processing (1)**
    - Speech Recognition
  - **Video processing (1)**
    - Video Prediction
  - **3D data processing (2)**
    - 3D Face Recognition, 3D Object Reconstruction

Overview

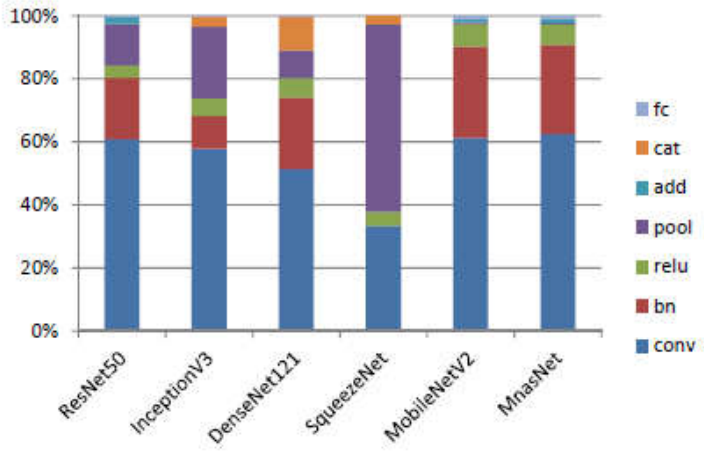
- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- **AIBench**
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - **AIBench Inference**
    - **AIoTBench**
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion

AIoTBench

- AIoTBench aims to evaluate the AI models, frameworks, and hardware on mobile and embedded environments.
  - Out-of-box benchmarks, more affordable
  - Make comprehensive and apple-to-apple comparisons of the algorithm, software, and hardware.
- AIoTBench covers representative and diverse models and frameworks. It has 60 off-the-shelf workload instances in total.
  - Models
    - ResNet50, InceptionV3, DenseNet121, SqueezeNet MobileNetV2 , MnasNet
  - Frameworks
    - Tensorflow Lite, Caffe2, Pytorch Mobile
  - For each model in Tensorflow Lite
    - three quantization versions: dynamic range quantization, full integer quantization, float16 quantization
    - CPU and NNAPI delegate

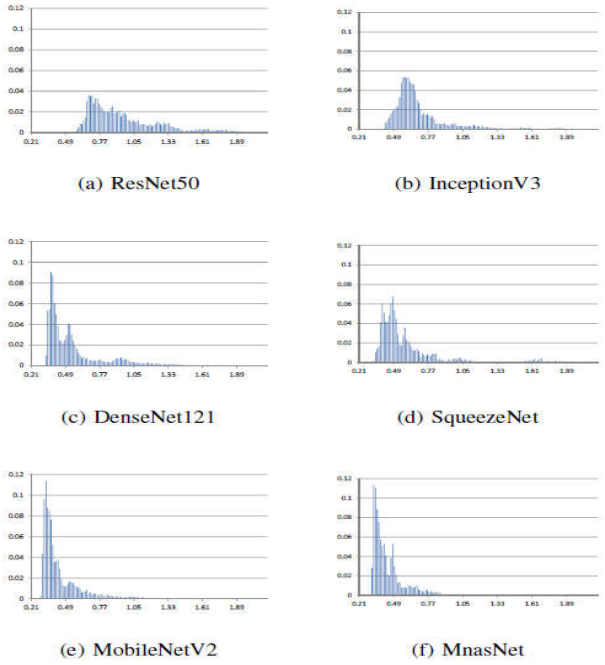
Model diversity

- Different operators have different proportions of processing time in different models



The time breakdown of operators for different models.

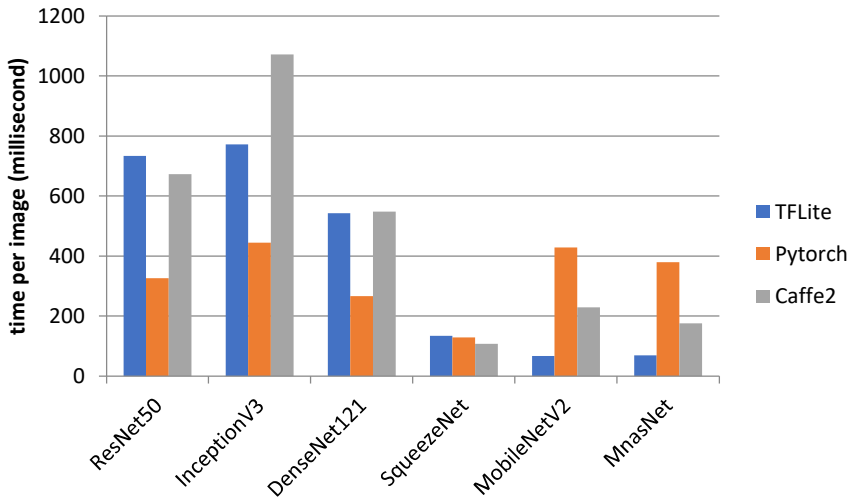
- The convolutions have diverse distributions of processing time in different model



The distribution of convolution processing time. Horizontal axis refers to time (millisecond), vertical axis refers to the frequency.

Framework diversity

- The same model has different performance on different framework



Time per image on Samsung Galaxy S10e

Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology
- **AIBench**
  - AIBench Scenario
    - Edge AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - **Micro Benchmarks**
  - AIBench Synthetic
- Conclusion

Micro Benchmarks

- Tackle the prohibitive cost challenge
- Easily portable across different architectures
- Primitive AI operators and widely-used hotspot functions from AIBench Training and AIBench Inference

Micro Benchmarks

No.	Micro Benchmark
DC-AI-M1	Convolution
DC-AI-M2	Fully Connected
DC-AI-M3	Relu
DC-AI-M4	Sigmoid
DC-AI-M5	Tanh
DC-AI-M6	MaxPooling
DC-AI-M7	AvgPooling
DC-AI-M8	CosineNorm
DC-AI-M9	BatchNorm
DC-AI-M10	Dropout
DC-AI-M11	Element-wise multiply
DC-AI-M12	Softmax

AIBench microbenchmark

GEMM

Convolution

RNN

All Reduce

DeepBench microbenchmark

Other Considerations

- Diverse simulation versions

- Spatially model the workloads
  - Batched kernels.

Overview

- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- **AIBench**
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - **AIBench Synthetic**
- Conclusion

Synthetic Benchmarks

- Tackle the short shelf-life and scalability challenges

- Design Spaces
  - Representative building blocks from AIBench Training
    - Fully connected layer, residual block, etc.
  - Deepth, width, and input sizes
  - Stride and padding sizes

- Train the generated models to convergence to reflect the learning dynamics, which ParaDNN[1] cannot reflect.

[1] Wang, Yu Emma, Gu-Yeon Wei, and David Brooks. "A Systematic Methodology for Analysis of Deep Learning Hardware and Software Platforms," 14.

Overview

- AI Benchmarking Challenges
- Related Work
- ScenarioBench & AIBench Methodologies
- ScenarioBench
  - AIBench Scenario
  - Edge AIBench
- **AIBench**
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- **Conclusion**

Panel questions

- How can the community produce and share scientific data and AI models that are findable, accessible, interoperable, and reusable (FAIR)?
  - GitHub
  - Model zoo: Discover open source deep learning code and pretrained models.
  - The challenge is how to reproduce the results.
- AI algorithms have been used to replace HPC computation based on domain knowledge. Do you see there is any opportunity to automate the process of applying AI to HPC applications?
  - Sure. I just cite some slides from Prof. Weinan E.
- How can available generic models be customized for specific needs and domains?
  - Quite challenging. Most CS guys do not know the domain knowledge well.
  - Benchmarks or libraries my help. We are doing some work.
- How to enable the collaboration between domain scientists and machine learning experts?
  - We tried in the past five years. Definitely failed. We must live together ☺. Now, we can not talk with each other ☺

Conclusion

- ScenarioBench and AIBench distill and abstract real-world application scenarios across Datacenter, HPC, IoT, and Edge into the scenario, training, inference, micro, and synthetic benchmarks.

- http://www.benchcouncil.org/aibench

- **https://www.benchcouncil.org/scenariobench**

Annual BenchCouncil Awards

- https://www.benchcouncil.org/html/awards.html
- BenchCouncil Achievement Award
  - This award recognizes a senior member who has made long-term contributions to benchmarks, data, standards, evaluations, and optimizations. The winner is eligible for BenchCouncil Fellow. ($3000)
- BenchCouncil Rising Star Awards
  - This award recognizes a young researcher who demonstrates outstanding research and practice in benchmarks, data, standards, evaluations, and optimizations. The winner is eligible for BenchCouncil Senior Member. ($1000)
- BenchCouncil Distinguished Doctoral Dissertation Award
  - This award recognizes and encourages superior research and writing by doctoral candidates in the broad field of benchmarks, data, standards, evaluations, and optimizations community. ($1000)

CFPs

- BenchCouncil Transactions on Benchmarks, Standards and Evaluations
  - https://www.editorialmanager.com/tbench/default.aspx
  - http://www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Thank you !