AI Benchmarking Challenges, AIBench Methodology and Summary

Prof. Dr. Jianfeng Zhan

ICT, Chinese Academy of Sciences & BenchCouncil

AIBench Tutorial @ ASPLOS 2021
Overview

- Challenges
- Related Work
- AIBench Methodology
- AIBench
  - AIBench Scenario
    - Edge AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion
AI: Important Modern Computer Workloads

- Datacenter
- Edge Computing
- Supercomputing
- AIoT
Challenges Of Modern Computer Workloads

- **Fragmented**
- **Isolated**
- **Dynamic**
- **Service-based**

Huge Fragmented Application Scenarios

- A marked departure from the past
A Huge Variety of Application Scenarios and Models Scenarios

Picture from:
[1]: http://www.hatdot.com/keji/2552842.html
[2]: https://medium.com/appanion/a-five-minute-guide-to-artificial-intelligence-c4262be85fd3
Challenges Of Modern Computer Workloads

- Fragmented
- Isolated
- Dynamic
- Service-based
Real-world datasets and workloads or even AI models are treated as first-class confidential issues

Isolated between academia and industry, or even among different providers.

Poses a huge obstacle for our communities towards developing an open and mature research field.
Challenges Of Modern Computer Workloads

- Fragmented
- Isolated
- Dynamic
- Service-based
Common requirements are handled collaboratively by datacenters, edge, and IoT devices.

Different distributions of data sets, workloads, ML models may substantially affect the system’s behaviors.

System architectures are undergoing fast evolutions in terms of the interactions among IoT, edge, and datacenters.
Challenges Of Modern Computer Workloads

- Fragmented
- Isolated
- Dynamic
- Service-based
The Side Effect of Service-based Architecture

- SaaS model changes workloads fast
  - workload churn
  - not scalable or even impossible to create a new benchmark or proxy for every possible workload.

- Microservice-based architecture
  - distributed across different datacenters
  - consist of a diversity of various modules with very long and complex execution paths.
  - tail latency matters
The Challenges Of AI Benchmarking

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

AI Bench Training: Balanced Industry-Standard AI Training Benchmarking.[PDF]
Prohibitive Cost Challenge

- Running an entire training session is mandatory!

- 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/
The Challenges Of AI Benchmarking

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

AIBench Training: Balanced Industry-Standard AI Training Benchmarking.[PDF]
Conflicting-requirement Challenge

- Earlier-stage evaluations of a new architecture or system
  - Affordable
  - Portability (Micro benchmarks)
  - Simplicity

- Later-stage evaluations or purchasing off-the-shelf systems
  - Comprehensiveness/Representativeness
  - Reality and overall system performance (Component or scenario benchmarks)
The Challenges Of AI Benchmarking

- Prohibitive cost
- Conflicting requirements
- Short shelf-life
- Scalability
- Repeatability

AIBench Training: Balanced Industry-Standard AI Training Benchmarking.[PDF]
Short Shelf-life Challenge

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

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

The Challenges Of AI Benchmarking

- Prohibitive cost
- Conflicting requirements
- Short shelf-life
- Scalability
- Repeatability

AIBench Training: Balanced Industry-Standard AI Training Benchmarking.[PDF]
Scalability Challenge

- An AI task’s problem scale is often fixed
- HPL-AI [1] is scalable, but it cannot model the learning dynamics, and fail to consider the model quality.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Source</th>
<th>Time(min)</th>
<th>AI Accelerators</th>
<th>Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fujitsu[1]</td>
<td>1.2</td>
<td>Tesla V100*2048</td>
<td>MXNeT</td>
</tr>
<tr>
<td>2</td>
<td>Google[2]</td>
<td>2.2</td>
<td>TPU V3*1024</td>
<td>TensorFlow</td>
</tr>
<tr>
<td>3</td>
<td>Sony[3]</td>
<td>3.7</td>
<td>Tesla V100*2176</td>
<td>NNL</td>
</tr>
</tbody>
</table>

Picture from HPC AI500 Ranking, Image Classification

[1] HPL-AI Mixed-Precision Benchmark — HPL-AI 0.0.2 documentation. [https://icl.bitbucket.io/hpl-ai/](https://icl.bitbucket.io/hpl-ai/)
The Challenges Of AI Benchmarking

- Prohibitive cost
- Conflicting requirements
- Short shelf-life
- Scalability

**Repeatability**

AIBench Training: Balanced Industry-Standard AI Training Benchmarking.[PDF]
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 AlBench Training four times using different random seeds

Overview

- Challenges
- Related Work
- AIBench Methodology
- AIBench
  - AIBench Scenario
    - Edge 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

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

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

**AIBench, Bench 2018**
- arXiv 2019, 2020
- First proposing scenario benchmarks
- 19 tasks, 19 workloads, 3 subsets

**DNNMark**
- GitHub 2016
- GPGPU 2017
- Eight micro benchmarks

**BenchIP**
- arXiv 2017
- JCST 2018
- 10 microbenchmarks
- 11 neural network models

**TBD Suite**
- GitHub 2018
- IISWC 2018
- Eight workloads, six domains

**MLPerf, 2018**
- GitHub 2019
- SysML 2020
- Five domains
- Seven workloads

**AIIA-DNN**
- GitHub 2019
- Designed to support training and inference, but only provides inference
HPC AI Benchmarking

- **DAWNBench, NIPS 2017**
  Image classification and question answer; the first AI benchmark that uses time-to-accuracy as the metric.

- **HPC AI500, Bench 2018, arXiv 2020**
  *Bench ‘18*
  Cover 3 representative application of scientific deep learning.
  *arXiv 2020*
  Hierarchical benchmarking methodology;
  3 benchmarking levels and rules;
  Use Valid FLOPS as the metric;
  Two representative and repeatable AI workloads (Business + Scientific).

- **Deep500, IPDPS 2019**
  A framework covering 4 level benchmarking;
  No concrete reference implementation.

- **HPL-AI, 2019**
  Micro benchmark based on LU decomposition;
  Scalable but can not reflect model quality.

- **MLPerf, arXiv 2019, SysML 2020**
  7 workloads covering 5 domains;
  2 benchmarking levels and rules;
  Use time-to-train as the metric.

- **AIPerf, 2020**
  Based on AutoML;
  Scalable but hard to ensure repeatability.
AI Benchmarks for Edge Computing

Edge AI Bench, Bench 2018, arXiv 2019
Scenario benchmarking
ICU patient monitoring, camera monitoring, smart home, and automatic driving
Integrated federal learning

EdgeBench, UCC Companion 2018
Speech recognition, and image classification

EEMBC MLMark, 2019
Image classification, object detection, translation, and speech recognition
Closed source
AI Benchmarks for IoT

**ETH Zurich AI Benchmark**, ECCV 2018
- Only supports vision domain
- Only supports Android and TensorFlow Lite
- End-to-end

**AIOT, Bench 2018**
- Vision, audio, and NLP domain
- Supports Android and Raspberry Pie
- TensorFlow Lite, Caffe 2
- End-to-end, microbenchmarks
DawnBench (2017) is the first benchmark that proposes time-to-accuracy as the main metric.

AIBench (2018) is the first benchmark that notices the necessity of modeling the critical paths of a real-world application scenario.

ParaDNN (2020) is the first synthetic AI benchmark.

The 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 in AIBench.

Both ParaDNN and HPL-AI can not model the learning dynamics, and also fail to consider the model quality.
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 it 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
- AIBench Methodology
  - AIBench
    - AIBench Scenario
      - Edge AIBench
    - AIBench Training
      - AIBench Subset
      - HPC AI500
    - AIBench Inference
      - AIoTBench
    - Micro Benchmarks
    - AIBench Synthetic
- Conclusion
International Open Benchmark Council (BenchCouncil)

- [http://www.benchcouncil.org](http://www.benchcouncil.org)
- A non-profit international organization
  - Aiming to promote the standardization, benchmarking, evaluation, incubation, and promotion of HPC, Chip, AI, Big Data, Block Chain, and other emerging techniques.

BenchCouncil AIBench

- [http://www.benchcouncil.org/aibench](http://www.benchcouncil.org/aibench)
## AIBench Summary

<table>
<thead>
<tr>
<th>Scenario benchmark</th>
<th>AIBench Scenario</th>
<th>Edge AI Bench</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>AIBench Training</td>
<td>AIBench Subset</td>
</tr>
<tr>
<td>Inference</td>
<td>AIBench Inference</td>
<td>AIoTBench</td>
</tr>
<tr>
<td>Micro</td>
<td>Conv</td>
<td>FC</td>
</tr>
<tr>
<td>Synthetic</td>
<td></td>
<td>AIBench Synthetic</td>
</tr>
</tbody>
</table>
A systematic AI benchmarking project tackling the challenges mentioned above
AIBench Scenario models the critical paths of a real-world application scenario as a permutation of the AI and non-AI modules.

Edge AIBench is an instance of the scenario benchmark suites, modeling end-to-end performance across IoT, edge, and Datacenter.
- AIBench Training and AIBench Inference cover nineteen representative AI tasks (will update) with state-of-the-art models to guarantee diversity and representativeness.
AIBench Training provides two subsets for repeatable performance ranking (RPR) subset and workload characterization (WC) subset to improve affordability.
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 Micro provides the intensively-used hotspot functions, profiled from AIBench Training and Inference, for simulation-based architecture researches.
As complementary to real-world benchmarks, AIBench Synthetic provides several synthetic benchmarks with scalable problem sizes while modeling learning dynamics.
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
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
The kernels’ runtime breakdown of 17 AI workloads

- Some micro benchmarks may occupy a little percentage

*Example: No Single Kernel*
Is Component Benchmark Sufficient?

Benchmarking with a Single Component

Putting the Component into a realistic scenario: Online Translation Intelligence

OCR Model Pruning

Tail latency shortens by 51%

Accuracy decreases 1%

OCR Model Pruning

Optimization Technique

Inconsistent conclusion

Tail latency has no improvement

Accuracy decreases 1%

Only a single component may need to error-prone conclusions

Single Component vs. Realistic Application

- **E-commerce Search Intelligence**

- 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 a single component cannot reflect the overall system’s effects**
Model Accuracy vs. QoS

- For *E-commerce Search Intelligence*
  - 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
      - 1136.79 millisecond => 10985.49 millisecond

Benchmarking a single component cannot reflect the tradeoff between model accuracy and QoS
Does a statistical model predict the overall system tail latency through profiling many components’ tail latency performance?
- NO!

Simple queueing model
- E-commerce Search Intelligence Scenario
  - 8.6X between the actual average latency and the theoretical one
  - 3.3X between the actual 99th percentile latency and the theoretical one

Sophisticated queueing network model
- E-commerce Search Intelligence Scenario
  - 4.9X between the actual average latency and the theoretical one
  - Difficult for tail latency predicting: non-superposition property
AIBench Scenario is needed!


- 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
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

16 AI Component Benchmarks for representative AI Tasks
- Speech recognition
- Text summarization
- Object detection
- Image generation
- 3D object reconstruction
- Video prediction
- Spatial transformer
- Recommendation
- 3D face recognition
- AI Units of Computation
- Text-to-Text translation
- Learning to rank
- Face embedding
- Image compression
- Image-to-Image
- Image-to-Text

AIBench Framework
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
An instance of the scenario benchmark suites
   ◆ modeling end-to-end performance across IoT, edge, and Datacenter

Publication
   ◆ Edge AIBench: towards comprehensive end-to-end edge computing benchmarking.
   ◆ Edge AIBench Specification
Four Typical Edge AI Scenarios

- **(1) Autonomous Vehicle**
  - Latency-sensitive
  - High-accuracy
  - Mobile

- **(2) ICU Patient Monitor**
  - Latency-sensitive
  - Parallel

- **(3) Surveillance Camera**
  - Enormous Data

- **(4) Smart Home**
  - Heterogenous devices and data
## Nine Typical Edge AI Tasks

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Edge AI Scenarios</th>
<th>Models</th>
<th>Datasets</th>
<th>Implementations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane Detection</td>
<td>Autonomous Vehicle</td>
<td>LaneNet</td>
<td>Tusimple/CULane</td>
<td>Pytorch/Caffe</td>
</tr>
<tr>
<td>Traffic Sign Detection</td>
<td>Autonomous Vehicle</td>
<td>Capsule Network</td>
<td>German Traffic Sign Recognition Benchmark</td>
<td>Keras</td>
</tr>
<tr>
<td>Heart Failure Prediction</td>
<td>ICU Patient Monitor</td>
<td>LSTM</td>
<td>MIMIC-III</td>
<td>Tensorflow/Keras</td>
</tr>
<tr>
<td>Decompensation Prediction</td>
<td>ICU Patient Monitor</td>
<td>LSTM</td>
<td>MIMIC-III</td>
<td>Tensorflow/Keras</td>
</tr>
<tr>
<td>Death Prediction</td>
<td>ICU Patient Monitor</td>
<td>LSTM</td>
<td>MIMIC-III</td>
<td>Tensorflow/Keras</td>
</tr>
<tr>
<td>Person Re-identification</td>
<td>Surveillance Camera</td>
<td>DG-Net</td>
<td>Market-1501</td>
<td>Pytorch</td>
</tr>
<tr>
<td>Action Detection</td>
<td>Surveillance Camera</td>
<td>ResNet18</td>
<td>UCF101</td>
<td>Pytorch/Caffe</td>
</tr>
<tr>
<td>Face Recognition</td>
<td>Smart Home</td>
<td>Facenet/Sphere network</td>
<td>LibriSpeech/CASIA-Webface</td>
<td>Tensorflow/Caffe</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Smart Home</td>
<td>DeepSpeech2</td>
<td>LibriSpeech</td>
<td>Tensorflow</td>
</tr>
</tbody>
</table>
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
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
    - Reality and system performance (Component or scenario benchmarks)
Diverse behaviors for workload characterization

- Micro-architecture level
  - FLOPs computation, memory access pattern, computation pattern, I/O pattern

- System level
  - Evaluation time cost, variation, convergence rate, number of hot functions

- Algorithm level
  - Model architectures and parameters
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

E-commerce

- Search (text, image, and audio)
- Advertising
- Intelligent assistant
- Products management
- Virtual fitting
- Facial authentication and payment

Related Technologies:
- Image classification
- Object detection
- Spatial transformer
- Translation
- Recommendation
- Learning-to-Rank
- Speech recognition
- Image compress
- Image-to-Text
- Text summarization
- Image-to-Image
- 3D face recognition
- Face embedding
## Representativeness

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

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology

**AIBench**
- AIBench Scenario
  - Edge AIBench
- AIBench Training
  - AIBench Subset
  - HPC AI500
- AIBench Inference
  - AIoT Bench
- Micro Benchmarks
- AIBench Synthetic

- Conclusion
Requirements for Ranking

Performance Ranking
- Fairness
- Low cost

Workload Characterization
- Representativeness
- Comprehensiveness
- Diversity
Two Subsets: Improve 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
- AIBench Methodology

**AIBench**
- AIBench Scenario
  - Edge 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.
AIBench subset computation comparison (Single training batch).

<table>
<thead>
<tr>
<th>Workloads</th>
<th>Computation (FLOPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Classification</td>
<td>23 G</td>
</tr>
<tr>
<td>Object Detection</td>
<td>691 G</td>
</tr>
<tr>
<td>Learning to Rank</td>
<td>0.08 M</td>
</tr>
</tbody>
</table>

**Image Classification** and **Object Detection** meet the scalability 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
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
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
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
- AIBench Methodology
  - AIBench
    - AIBench Scenario
      - Edge AIBench
    - AIBench Training
      - AIBench Subset
      - HPC AI500
    - AIBench Inference
      - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion
AIoTBench aims to evaluate the AI models, frameworks, and hardwares on mobile and embedded environment.

- Out-of-box benchmarks, more affordable
- to make comprehensive and apple-to-apple comparisons of 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 convolution itself has diverse performance in different models.

![Graph showing model diversity](image)
The same model has different performance on different framework.
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

- To tackle the prohibitive cost challenge
- Easily portable across different architectures
- Basic AI operators and widely-used hotspot kernel functions from AIBench Training and AIBench Inference
## Micro Benchmarks

<table>
<thead>
<tr>
<th>No.</th>
<th>Micro Benchmark</th>
<th>GEMM</th>
<th>Convolution</th>
<th>RNN</th>
<th>All Reduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC-AI-M1</td>
<td>Convolution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M2</td>
<td>Fully Connected</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M3</td>
<td>Relu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M4</td>
<td>Sigmoid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M5</td>
<td>Tanh</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M6</td>
<td>MaxPooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M7</td>
<td>AvgPooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M8</td>
<td>CosineNorm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M9</td>
<td>BatchNorm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M10</td>
<td>Dropout</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M11</td>
<td>Element-wise multiply</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-AI-M12</td>
<td>Softmax</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AIBench microbenchmark

DeepBench microbenchmark
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
To 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

---

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**
AIBench distills and abstracts real-world application scenarios across Datacenter, HPC, IoT, and Edge, into the scenario, training, inference, micro, and synthetic benchmarks

http://www.benchcouncil.org/aibench
Thank you!