AI Tax: Motivating the Need for End-to-end Performance Analysis of ML Tasks

Vijay Janapa Reddi
Harvard University

Rising Star Award Keynote

BenchCouncil
Nov 15th, 2021
AI Applications
What’s Under the Hood

New Compact SMARC AI-on-Module to Drive Industrial AI at the Edge

By EE Times - March 18, 2021

ADLINK Technology Inc., a global leader in edge computing, has launched the LEC-IMX8MP SMARC module, the first SMARC rev. 2.1 AI-on-Module (AoM) that uses NXP’s next-generation iMX 8M Plus SoC for edge AI applications. The LEC-IMX8MP enables powerful multi-core computing and advanced AI performance at the edge.

Flex Logix raises $55M to design AI chips for edge enterprise applications

AI accelerators are a type of specialized hardware designed to speed up AI applications, particularly neural networks, deep learning, and various...

2 days ago

Chip-Enabled Edge AI Drives Next-Gen IoT

“AI accelerators at the edge can do local processing before sending the...

GPU-accelerated parallel processing in mobile embedded systems.

1 month ago
Benchmarking AI Performance

MLPerf

AI Bench

Measure

Loading content...

7%

Measure your device's performance for:
- Image Classification
- Object Detection
- Image Segmentation
- Language Processing
- Image Classification (offline)
But... are these benchmarks representative of what end-users experience?
ML Execution Pipeline
ML Execution Pipeline
ML Execution Pipeline

Data Capture → Pre-processing → Frameworks

- Crop
- Scale
- NNAPI
- Vendor
ML Execution Pipeline

Data Capture → Pre-processing → Frameworks → Execution

- Crop
- Scale
- NNAPI
- Vendor
- Execution

[Diagram showing data capture, pre-processing, frameworks, and execution stages]
ML Execution Pipeline

Data Capture → Pre-processing → Frameworks → Execution → Post-processing

- Crop
- Scale
- ... 

- NNAPI
- Vendor
- ... 

- CPU Op Op Op
- DSP Op
- GPU Op Op Op
- NPU Op
- ... ... 

- topK

{ people, forest, person, lamps, ... }
ML Execution Pipeline
ML Execution Pipeline

End-to-end performance determines end-user QoE
ML is Complicated
E2E Latency Breakdown
E2E Latency Breakdown
E2E Latency Breakdown

- E2E Performance
- AI Tax
  - Algorithms
    - Data Capture
    - Pre-processing
    - Post-processing
  - Frameworks
    - Drivers
    - Scheduling
  - Hardware
- AI Model
E2E Latency Breakdown

- **E2E Performance**
  - **AI Tax**
    - **Algorithms**
      - Data Capture
      - Pre-processing
      - Post-processing
    - **Frameworks**
      - Drivers
      - Scheduling
    - **Hardware**
      - Offload
  - **AI Model**
What is AI Tax?
What Do We Measure?

Option 1

Option 2

Option 3
On-Device OCR:
A case study
PhotoOCR Normalized Performance (CPU only)
Do we account for pre- and post-processing times in the inference run test?
End-to-end analysis of AI tax: typical Android ML applications versus benchmark utilities

How much do individual stages of the ML execution pipeline influence application performance?

Can we quantify AI tax in real-world systems?
Experimental Setup

- Main evaluation platform
  - Google Pixel (**Snapdragon 845**)
  - Also experiment on Qualcomm **Snapdragon 835, 855, 865**

- Main evaluation framework
  - **NNAPI, TFLite, SNPE**

- Machine learning models
  - **MobileNet-v2, EfficientNet, PoseNet, DeepLab**, etc.

- Measure execution stages (e.g., pre-/post-processing) in TFLite benchmark utility and Android applications
AI Tax: Algorithmic

- How does the performance compare when running ML benchmarks versus running a real end-user mobile application?

- Up to 50% higher latency per inference (on CPU) when doing so through an Android app compared to command line benchmark utility (and its Android wrapper)
AI Tax: Frameworks

- Depending on how the framework is configured, performance may degrade due to limited support for particular models/platforms.

- Case study
  - EfficientNet-Lite0 via NNAPI
  - Observed ~7x slowdown in latency per inference in image processing app when letting NNAPI decide hardware backend.
  - Unsupported TFLite op. -- CPU fallback.
AI Tax: Hardware

- Multi-tenant execution changes the performance profile of applications (which are typically studied in isolation)
- Case study:
  - Single model is offloaded with NNAPI while other models execute inference on the CPU
    - Latency constant, pre-processing progressively worse due to interference
**AI Tax: Hardware**

- Multi-tenant execution changes the performance profile of applications (which are typically studied in isolation)
- Case study:
  - Single model is offloaded with NNAPI while other models execute inference on the CPU
    - Latency constant, pre-processing progressively worse due to interference
  - Multiple models contending for accelerators (via NNAPI) with main application running its pre-processing on CPU
    - Latency scales linearly with number of processes while CPU sees constant activity
    - Why not use some of the CPU resources instead?
E2E Latency Breakdown

- **E2E Performance**
  - **AI Tax**
    - **Algorithms**
      - Data Capture
      - Pre-processing
      - Post-processing
    - **Frameworks**
      - Drivers
      - Scheduling
    - **Hardware**
      - Offload
  - **AI Model**
AI Tax in DataCenters
AI Tax in DataCenters
Increasing Awareness of “AI Tax” in Research

AI Tax in Mobile SoCs: End-to-end Performance Analysis of Machine Learning in Smartphones

Daniel Ricciolo, Brainerd Dagon

ABSTRACT
Artificial intelligence (AI) is becoming increasingly important in our daily lives. With the advent of mobile devices, AI is now integrated into various applications, from virtual assistants to speech recognition. However, the performance and energy efficiency of AI algorithms are critical factors in determining their success.

In this paper, we present an analysis of the performance and energy consumption of AI algorithms on mobile devices. We focus on the implementation of AI models on mobile platforms and the challenges they face, such as limited computational resources and power constraints. Our results highlight the need for efficient AI algorithms that can operate on mobile devices while maintaining high performance.

Jointly Optimizing Preprocessing and Inference for DNN-based Visual Analytics

Daniel Kang, Ankiet Mathur, Teja Veeramachaneni, Peter Bails, Matei Zaharia

ABSTRACT
While deep neural networks (DNNs) are increasingly popular in various fields, the preprocessing and inference stages can significantly impact the overall performance and energy efficiency of the system. In this paper, we propose a joint optimization of preprocessing and inference stages for DNN-based visual analytics.

We develop a unified framework that considers both the preprocessing and inference stages. Our framework allows for the optimization of preprocessing operations, such as data preprocessing and model selection, and the inference stage, which includes the actual prediction. By jointly optimizing both stages, we can achieve significant improvements in both accuracy and efficiency.

This work extends a previous End-to-End AI Application Performance (E2EAP) benchmark [2], which is a collection of benchmarks for evaluating AI applications. Our new benchmark is designed to provide a comprehensive evaluation of the performance and energy efficiency of AI applications on mobile devices.

PUBLISHER Reference Format

1 INTRODUCTION
Machine learning and data analytics are revolutionizing various fields, from healthcare to finance. However, the performance and energy efficiency of AI algorithms are critical factors in determining their success.

In this paper, we present an analysis of the performance and energy consumption of AI algorithms on mobile devices. We focus on the implementation of AI models on mobile platforms and the challenges they face, such as limited computational resources and power constraints. Our results highlight the need for efficient AI algorithms that can operate on mobile devices while maintaining high performance.

Jointly Optimizing Preprocessing and Inference for DNN-based Visual Analytics

Daniel Kang, Ankiet Mathur, Teja Veeramachaneni, Peter Bails, Matei Zaharia

ABSTRACT
While deep neural networks (DNNs) are increasingly popular in various fields, the preprocessing and inference stages can significantly impact the overall performance and energy efficiency of the system. In this paper, we propose a joint optimization of preprocessing and inference stages for DNN-based visual analytics.

We develop a unified framework that considers both the preprocessing and inference stages. Our framework allows for the optimization of preprocessing operations, such as data preprocessing and model selection, and the inference stage, which includes the actual prediction. By jointly optimizing both stages, we can achieve significant improvements in both accuracy and efficiency.

This work extends a previous End-to-End AI Application Performance (E2EAP) benchmark [2], which is a collection of benchmarks for evaluating AI applications. Our new benchmark is designed to provide a comprehensive evaluation of the performance and energy efficiency of AI applications on mobile devices.

PUBLISHER Reference Format

1 INTRODUCTION
Machine learning and data analytics are revolutionizing various fields, from healthcare to finance. However, the performance and energy efficiency of AI algorithms are critical factors in determining their success.

In this paper, we present an analysis of the performance and energy consumption of AI algorithms on mobile devices. We focus on the implementation of AI models on mobile platforms and the challenges they face, such as limited computational resources and power constraints. Our results highlight the need for efficient AI algorithms that can operate on mobile devices while maintaining high performance.

Jointly Optimizing Preprocessing and Inference for DNN-based Visual Analytics

Daniel Kang, Ankiet Mathur, Teja Veeramachaneni, Peter Bails, Matei Zaharia

ABSTRACT
While deep neural networks (DNNs) are increasingly popular in various fields, the preprocessing and inference stages can significantly impact the overall performance and energy efficiency of the system. In this paper, we propose a joint optimization of preprocessing and inference stages for DNN-based visual analytics.

We develop a unified framework that considers both the preprocessing and inference stages. Our framework allows for the optimization of preprocessing operations, such as data preprocessing and model selection, and the inference stage, which includes the actual prediction. By jointly optimizing both stages, we can achieve significant improvements in both accuracy and efficiency.

This work extends a previous End-to-End AI Application Performance (E2EAP) benchmark [2], which is a collection of benchmarks for evaluating AI applications. Our new benchmark is designed to provide a comprehensive evaluation of the performance and energy efficiency of AI applications on mobile devices.
AI Tax in Mobile SoCs: End-to-end Performance Analysis of Machine Learning in Smartphones

Missing the Forest for the Trees: End-to-End AI Application Performance in Edge Data Centers

AI Tax: The Hidden Cost of AI Data Center Applications

Jointly Optimizing Preprocessing and Inference for DNN-based Visual Analytics

AI Bench: Towards Scalable and Comprehensive Datacenter AI Benchmarking

Real-world or Future Application Scenarios

Scenario Benchmarks
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks

Real-world or Future Application Scenarios

End-to-End AI Application Performance

AIBench Scenario
- Edge AIBench
- AIBench Training
- AIBench Inference
- AIBench Subset
- HPC A1500
- AloT Bench

Component Benchmarks
- AI Bench Micro
- AI Bench Synthetic
- Synthetic Benchmarks
Pipeline DNNs

- Keyword Spotting
- Speech Processing

“Hey FB” “Can you do...”

- Back-to-back execution
- Execution dependency

Concurrent DNNs

- Eye Tracking
- Obstacle Detection
- Video Processing

- Concurrent execution
- Execution Deadline

Concurrent + Pipeline

- Obstacle Detection
- Eye Tracking
- Foveated Rendering

- Challenges from both type 1 and 2!
Benchmarks need to expand their scope to drive the industry forward.

End-to-end performance analysis is crucial for representative evaluation.

Expose more information about AI tax in frameworks and benchmarks.