A LOW POWER AND HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE INFERENCE APPROACH FOR ONBOARD DATA PROCESSING

mandar.harshe@klepsydra.com
www.klepsydra.com
Technical Specifications:
• AI model: Multibranch DNN based on AlexNet for 224x224x3 images
• Processor: Intel 3.3GHz Dual core
• Data rate: 10 Frames per second (FPS)
• Implementations: Klepsydra AI, Tensor Flow Lite, Open CV
• OS: Ubuntu 20.04

https://klepsydra.com/klepsydra-ai-online-demo/
Edge computing challenges

- Cloud interaction dependency
- Over dimensioned hardware and therefore excessive costs
- Safety issues:
  - Data loss
  - Long latencies
  - Unreliable systems
- Constraints can vary:
  - Low CPU
  - Low Latency
  - High Throughput
KLEPSYDRA AI APPROACH

**Trading Software**

- Bigger computer did not solve the problem
- Use revolutionary data processing software
- Top investment banks make billions using these techniques.

**Parallelisation**  
**Pipeline**
Threading model consists of:
- Number of cores assigned to event loops
- Number of event loops per core
- Number of parallelisation threads for each layer

Layers can be parallelised and are vectorized.
KLEPSYDRA AI PERFORMANCE

AlexNet Power Consumption comparison for Raspberry Pi4

AlexNet Processed data volume comparison for Raspberry Pi4

- Klepsydra
- OpenCV
- TensorFlow

Power Consumption (%)

Processed data (Hz)

Data rate (Hz)
Smaller networks:
- Image size 80 x 80
- Tuning for throughput

At best latency, limited by FPU

Raspberry Pi 4 : Ubuntu 20.04
“Heavy” convolutions
- Image size 224 x 224
- Padding same

KLEPSYDRA AI PERFORMANCE

Intel : Ubuntu 20.04
Conclusions:

• More deterministic.
• Zero data loss.
• Klepsydra CPU usage is 50% less than TensorFlow’s.
• Klepsydra Throughput is 2x more than TensorFlow’s and even more with respect to OpenCV!

https://klepsydra.com/klepsydra-ai-online-demo/
Premier Sponsor
Automated TinyML

Zero-code SaaS solution

Create tiny models, ready for embedding, in just a few clicks!

Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Executive Sponsors
Arm: The Software and Hardware Foundation for tinyML

1. Connect to high-level frameworks
2. Supported by end-to-end tooling
3. Connect to Runtime

Profiling and debugging tooling such as Arm Keil MDK

Application

Optimized models for embedded

Runtime (e.g. TensorFlow Lite Micro)

Optimized low-level NN libraries (i.e. CMSIS-NN)

RTOS such as Mbed OS

Arm Cortex-M CPUs and microNPUs

AI Ecosystem Partners

Stay Connected

@ArmSoftwareDevelopers
@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

© 2020 Arm Limited (or its affiliates)
TinyML for all developers

- **Dataset**
  - Acquire valuable training data securely
  - Enrich data and train ML algorithms

- **Edge Device**
  - Real sensors in real time
  - Open source SDK
  - Embedded and edge compute deployment options

- **Impulse**
  - Test impulse with real-time device data flows

- **Test**

---

www.edgeimpulse.com
Advancing AI research to make efficient AI ubiquitous

- **Power efficiency**: Model design, compression, quantization, algorithms, efficient hardware, software tool
- **Personalization**: Continuous learning, contextual, always-on, privacy-preserved, distributed learning
- **Efficient learning**: Robust learning through minimal data, unsupervised learning, on-device learning

**A platform to scale AI across the industry**

- **Perception**: Object detection, speech recognition, contextual fusion
- **Reasoning**: Scene understanding, language understanding, behavior prediction
- **Action**: Reinforcement learning for decision making

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant’s chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com @Syntiantcorp
Platinum Sponsors
Add Advanced Sensing to your Product with Edge AI / TinyML

Reality AI®

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions

Prebuilt sound recognition models for indoor and outdoor use cases

Solution for industrial anomaly detection

Pre-built automotive solution that lets cars "see with sound"

Reality AI Tools® software

Build prototypes, then turn them into real products

Explain ML models and relate the function to the physics

Optimize the hardware, including sensor selection and placement

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
Gold Sponsors
Adaptive AI for the Intelligent Edge

LatentAI.com
Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.

sensiml.com
Silver Sponsors
Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® EMEA Technical Forum 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at tinyML EMEA. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org