TinyML Benchmarking panel

Applications driven benchmarking for TinyML

Moderator: Petruț Bogdan

Overview

Introductions

Goals of the TinyML Datasets and Benchmarking Working Group

Visual Wake Words v2

Discussion

Join the discussion in person or online!

Martin Croome

Thomas Basikolo

Alf Kuchenbuch

Petruț Bogdan

Martin Croome

VP Marketing

Greenwaves Technologies

What we do

GREENWAVES TECHNOLOGIES

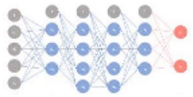
=

Battery

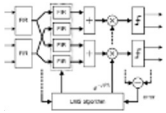
AI

Digital Signal Processing

+



+



Why are benchmarks relevant?

- Benchmarks are relevant to different customers at different levels:
  - Application level: Benchmark these denoising solutions from different suppliers
  - General level: What is the expected latency of your chip/toolchain
  - Specific level: What will be the performance of my network
- GOPS/W gives very little information
  - Generally ignores data movement which is at least 50 percent of the problem
  - Generally ignores getting ready to compute
  - Generally ignores possible network optimizations and how easy they are to implement on your chip - i.e. compression techniques to reduce computation/data movement
  - Some of what we do just doesn't fit at all - i.e. updating uSec latency audio signal processing with a neural network.
- What we use so far
  - "Well known networks" i.e. MNv123 etc. OK for image. Not great for audio
  - ML Perf Tiny - Submitted in V1 round. Little diffusion. Networks are all too easy and not representative of the problems our customers give us. E.g. Audio source separation is WAY more difficult than KWS.

Thomas Basikolo

Programme Officer

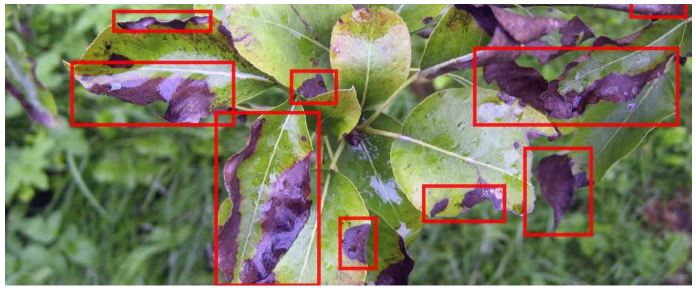
ITU

A new perspective on Benchmarking

- Well-defined/standardized datasets that represent real-world scenarios
- Real-world deployment challenges that tinyML systems may face
- Open collaboration
- Consider ***for Good*** applications or scenarios in Benchmarking

Next-Gen tinyML Smart Weather Station

Wildlife Monitoring



Plant Disease Detection

Alf Kuchenbuch

VP Sales

BrainChip

BrainChip At A Glance

- **First to commercialize** neuromorphic IP platform and reference chip.
- **Competitive advantage** extended with launch of second-gen Akida IP platform
- **World-class board & executive team** focused on commercialization and R&D.
- **Growing commercial ecosystem and partnerships** with leading AI tech companies
- **Strong patent portfolio** that protects the business

Trusted By:

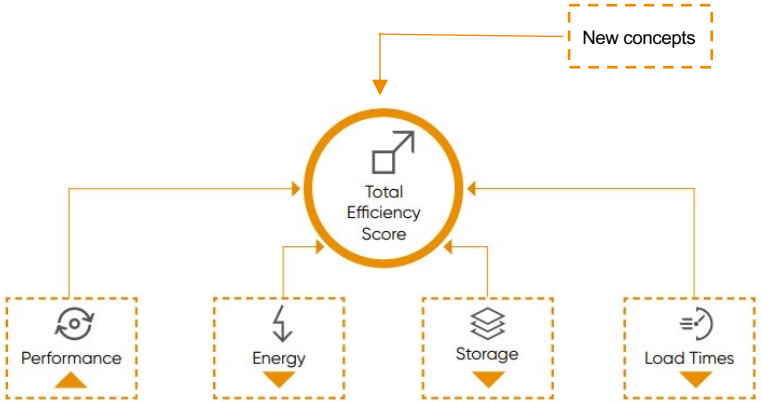
MegaChips, Valeo, RENESAS, VORAGO TECHNOLOGIES, NASA, Mercedes-Benz

Partnered with:

arm, intel foundry services, EDGE IMPULSE, TEKSUN CULTIVATING TECHNOLOGY, PROPHESEE, Ai Labs, EMOTION3D, NVISO, SiFive

Future of Edge AI Benchmarking

- Considerations are different in constrained devices
  - TOPS isn't the defining metric
- MLPerf's TinyML benchmarks take a great step forward
  - Representative Edge AI work loads
  - Performance and energy metrics
  - Notion of model size – that is not yet incorporated into a value metric.
  - Areas of potential extensions: load times, system offload
- Next generation benchmarks could be improved further
  - Overall value metric that combines performance and energy improvement versus a standard
  - Incorporation of load times and system load
  - Expanding use cases for "Tiny"?



Holistically assessing performance and efficiency in a constrained device

Petruț Bogdan

Neuromorphic Architect

Innatera Nanosystems

Scope of the benchmarking WG

- We want to help academia/industry **improve algorithms with realistic/practical data**, this will help in driving appropriate benchmarks
- We want to identify the **challenging applications** to be able to drive the collection/reuse of specific datasets
  - Identify **applications**
  - Identify the **data sets**
  - Identify the current shortcomings/what is missing/what needs work on (How to handle augmentation/negative cases)
  - Can we increase datasets to make them a more realistic **production quality level** (rather than just a small toy example)? Note the intent here is to have at least a full specific example (to reflect the complexity, not necessarily to make it possible for someone to make an actual product out of it).
  - Can we **break down a large dataset** to make easy/medium/hard subsets of data or subsets of output classes?
  - What are **relevant benchmark topologies/criteria for the application**?
  - **Share best practices** for community and industry to move forward (not to be stuck in out-of-date benchmarks)
  - Need to identify the **appropriate cadence of changing benchmarks** (to support above, but not make it impossible to compare vendors to one another because benchmarks change).
  - Identify standards for re-use between different NNs (e.g. standardize input resolution/output classes, so that you can plug and play different NNs for running on same data)
  - How do we deal with **Network Architecture Search** optimization possibilities? **How to compare** what different users can do? (e.g. open category in MLPerf)

Visual Wake Words V2

(a) ‘Person’

(b) ‘Not-person’

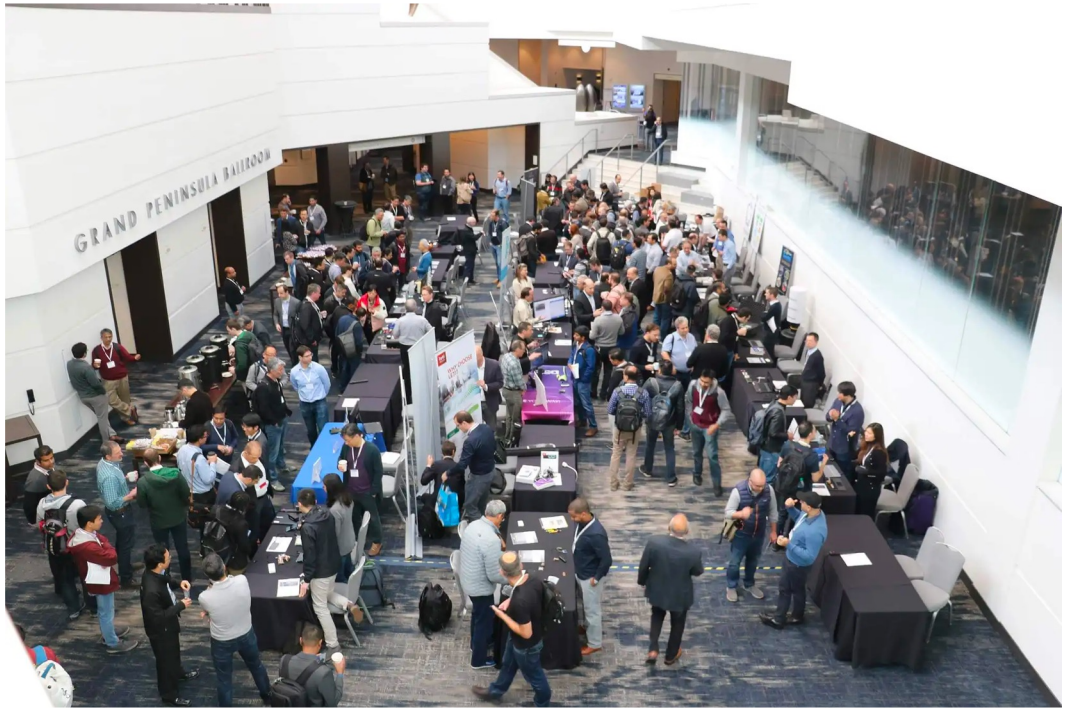
VWW Flaws

- Small
  - ~200k Training Images
- Label Errors
  - ~7% label errors on the Validation set
- Hard to Use
  - Manually generated via deprecated TF code (TF Slim)
- Non-standard Evaluation Methodology

VWW Label:
No Person

Why tinyML Foundation?

- Expertise
- Influence
- Virtuous Cycle



Thank you!

Summary & Discussion

Join the discussion
in person or online!

Copyright Notice

www.tinyml.org