“Introduction to ScaleDown: An Open Source Initiative”

Archana Vaidheeswaran - Women Who Code

July 21, 2022
Thank you, tinyML Strategic Partners*, for committing to take tinyML to the next Level, together
Executive Strategic Partners
Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more: www.arm.com/techtalks
The Leading Development Platform for Edge ML

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

- **Perception**
  - Object detection, speech recognition, contextual fusion

- **Reasoning**
  - Scene understanding, language understanding, behavior prediction

- **Action**
  - Reinforcement learning for decision making

A platform to scale AI across the industry

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com
Platinum Strategic Partners
Fastest Video Analytics Solutions on Arm CPUs

www.deeplite.ai
High-Value or Safety-Critical Use Cases?

For your most important projects, use Reality AI Tools®

TinyML software that covers the full engineering lifecycle:

- AutoML for non-visual sensing based on advanced signal processing math
- Hardware design analytics
- Explanation of TinyML models in terms of underlying physics
- Automated Data Readiness assessment

https://reality.ai/  @SensorAI  info@reality.ai  Reality AI
Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.
Sony
Semiconductor Solutions Corporation
Gold Strategic Partners
Witness potential made possible at analog.com.

Where what if becomes what is.
FOTAHU

Making Over-the-Air Firmware and ML models Updates Simple and Accessible!

- Securely update your IoT devices regardless of their Hardware Platform (Silicon) Provider and physical location.
- Unlock TinyML business value through OTA Firmware and ML models update.
- Pay-as-you-go

www.fotahub.com
contact@fotahub.com
The Latent AI Efficient Inference Platform™ (LEIP) combined with application templates makes optimizing neural networks for production inference a cinch.

THE EASY BUTTON FOR DEVELOPERS resulting in faster time to market.

Reach out to info@latentai.com
TOGETHER, WE ACCELERATE THE BREAKTHROUGHS THAT ADVANCE OUR WORLD

www.nxp.com/ai
Deploy TinyML into the Real World - Plug and Play ML

**Sensors:**
- Modulated and ready-to-use sensors to simplify the setup process
- Support 500+ Grove modules

**Wio Terminal:**
- Complete AI platform integrated with a 2.4'' LCD Screen, onboard IMU (LIS3DHTR), microphone, buzzer, microSD card slot, light sensor, infrared emitter (IR 940nm)

**Codecraft:**
- No code Programming platform to Get Started With TinyML
- Supports Arduino, Python, C or JavaScript etc.

**Edge Impulse:**
- To optimize data utilization and enable deploy a machine learning model faster than ever

**TensorFlow Lite:**
- To easily train low memory usage machine learning models

**Applications:**
- Artificial Nose
- AI Thermal Camera for Safe Camping
- Azure IoT Squirrel Feeder
The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

Analytics Toolkit Suite

AutoML

Data Collection

Data Labeling

Model Evaluation

Tool & Validation

Code Generation

Team Collaboration

Version Control and Model Management

sensiml.com/tinyML
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
ENGINEERING
EXCEPTIONAL
EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com
SynSense builds sensing and inference hardware for ultra-low-power (sub-mW) embedded, mobile and edge devices. We design systems for real-time always-on smart sensing, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai
tinyML EMEA Innovation Forum 2022

Connect, Unify, and Grow the tinyML EMEA Community
October 10-12, 2022

https://www.tinyml.org/event/emea-2022

Event will be held in person in Cyprus.

EMEA 2022 Call for Presentations is open till August 1st, 2022.

More sponsorships are available: sponsorships@tinyML.org
Our next tinyML Trailblazers Series
Success Stories with Vijay Janapa Reddi
(Associate Professor, Harvard University)

LIVE ONLINE August 3rd, 2022 at 8 am PDT

Register now!
Join Growing tinyML Communities:

tinyML - Enabling ultra-low Power ML at the Edge

The tinyML Community
https://www.linkedin.com/groups/13694488/

10.6k members in 45 Groups in 36 Countries

2.9k members & 8.1k followers
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)
www.youtube.com/tinyML
Reminders

Slides & Videos will be posted tomorrow

Please use the Q&A window for your questions

tinyml.org/forums    youtube.com/tinyml
Hi I am Archana
Data Product Manager
Introduction to ScaleDown

- Doing TinyML is hard
  - Less educational resources
  - Need for hardware
  - Fragmented hardware and software ecosystem
- ScaleDown aims to democratize TinyML Education
- How?
  - By providing open-source educational materials
  - By giving users access to hardware via a free library
  - By abstracting away hardware and software difficulties through our framework
  - Quick support for new algorithms
Last year saw lead times for chip orders double to more than 20 weeks.

Moreover, there was an estimated chip shortage of 10M units last year, with a shortage of more than 2.6M units in the last quarter alone.
Scaledown Verticals

ScaleDown: Democratizing TinyML Education

Hardware Library
- Books
- Blogs
- Study Group
- Research

Projects

Framework
- Optimisation
- Benchmarking
- TinyML Ops

Optimisations
- Quantization
- Pruning
- Knowledge Distillation

Framework
- PyTorch
- Tensorflow
- OpenVino
Education & Hardware Library

1. Blogs
2. Articles
3. Documentation
4. Website
5. Study Group with TinyML Foundation
6. Research Papers
7. Social Media Content
8. Conferences
9. Courses with LinkedIn Learning
10. Publishing Books

1. Start Your Own Hardware Library
2. Logistics
3. New Hardware
4. Integration of Framework and Support
Framework & Projects

1. Algorithms
2. Support New Hardware
3. Support New Toolkits
4. TinyMLOps
5. Benchmarking
6. AutoML

Projects

1. Use the framework and library to build projects
2. Publish them to our blog or hackster
3. Help grow the community and educate other people
Quantization

- Neural Networks generally use 32 bit floating point weights
- These require more memory to save and custom logic to execute efficiently (which many micro-controllers do not have)
- By converting weights to INT8 (8 bits), we can save memory and compute faster
- However, it leads to a loss in accuracy since we lose precision

source: heartbeat.fritz.ai
Model Pruning

- Weights can be Pruned
- Neurons can be Pruned
- Executing Pruned Networks Efficiently is Difficult
- Structured Sparsity can help: Essentially pruning blocks of weights
Knowledge Distillation

- Works by transferring the knowledge learned by a large teacher model to a smaller student model.
- The student model is easier to execute.
- The student model can be trained with unlabelled data.

Student models can often achieve similar or more accuracy than the teacher.
Edge Computing Frameworks

- For Optimizing Models
  - TensorFlow Lite
    - Quantization
    - Pruning
    - Weight Clustering
  - OpenVINO
    - Quantization
  - PyTorch
    - Quantization
Edge Computing Frameworks

- Model Representations
  - TensorFlow Lite
    - TFLite format
  - OpenVINO
    - Support for NCS and other Intel Hardware
  - ONNX
- Model Deployment
  - TensorFlow Lite
  - OpenVINO
  - Deepstream and TensorRT
How can you get involved?
How can you get Involved?

- Request for Hardware
- Write Content and Resources
- Research
- Build the Package
How can you get Involved?

Request for Hardware
Write Content and Resources
Research
Build the Package
How can you get Involved?

Request for Hardware

Write Content and Resources

Research

Build the Package
Creating Content

How can you started?

- Get Acclimated to User-Guide and Website Development
- Write a blog on TinyML
- Build a project and publish that as a blog
- Create Study Group Content
- Present at meetup and conferences like those by the TinyML Foundation
How can you get Involved?

- Request for Hardware
- Write Content and Resources
- Research
- Build the Package
Research Paper Ideas

Development:
1. Support New Research Algorithms like Early Exits
2. Provide Wrapper via ScaleDown
3. Increase Model Security via wrapper

Community
1. Paper Reading via TinyML SG

Research Papers
1. TinyMLOps
2. White Papers
   a. TinyMLOPs with ScaleDown
   b. Optimisation Techniques with ScaleDown
   c. Hardware Library Ecosystem with ScaleDown
   d. Implementing New Research Algorithms
How can you get Involved?

- Request for Hardware
- Write Content and Resources
- Research
- Build the Package
Contributing to ScaleDown: Code

- Repo: github/scaledown-team/scaledown
- Create an issue or pick up existing issues and action items
- Contributing
  - Each type of algorithm has its own folder:
  - Add your algorithm implementation
  - Extras:
    - Add documentation and examples
    - Add tests
    - Make sure the API can support multiple frameworks
  - Create PR!
Contributing to ScaleDown: Documentation

- Repo: github/scaledown-team/scaledown-team.github.io
- Made with mkdocs
- Contributing
  - Add examples or projects and publish to our blog
  - Create user guides and recipes for each algorithm
  - Add learning materials for each algorithm
Roadmap

Q2:2022:
- ScaleDown support for Knowledge Distillation,
- Support for TFLite and PyTorch,

Q3:2022:
- Starting Hardware Library to 1 city
- Publish 1 research paper
- Support Quantization and Pruning
- Support for OpenVino and ONNX

Q4:2022:
- ScaleDown support for TinyMLOps, monitoring, security and model updates
- Complete 1 cycles of Hardware Library
Join Us!
What areas do you want to focus on?

Check out the slack link (ScaleDown)
What’s Next?

Study Group: Optimization Techniques: Knowledge Distillation

Study Group: Optimization Techniques: Knowledge Distillation Paper Implementation

Study Group: Apply for your Hardware: Welcome to the Hardware Library

https://www.meetup.com/tinyml-enabling-ultra-low-power-ml-at-the-edge-singapore/
Knowledge Distillation

1. Introduction to Knowledge Distillation
2. Hands-On Implementation of Knowledge Distillation
3. Research Paper Reading
4. Research Paper Implementation
Copyright Notice

This multimedia file is copyright © 2022 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinyml.org
Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. 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.