“Minimizing resource usage in microcontrollers for cost effective solutions”

Ilya Gozman – Senior Fellow, Chief AI Architect, Grovety

September 26, 2023
Thank you, tinyML Strategic Partners, for committing to take tinyML to the next Level, together
Executive Strategic Partners
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
Platinum Strategic Partners
DEPLOY VISION AI AT THE EDGE AT SCALE
Gold Strategic Partners
Witness potential made possible at analog.com.

Where what if becomes what is.
Build the Future of tinyML on ARM
The Leading Development Platform for Edge ML

edgeimpulse.com
Driving decarbonization and digitalization. Together.

Infineon serving all target markets as Leader in Power Systems and IoT

www.infineon.com
NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com
Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com
Join Growing tinyML Communities:

tinyML - Enabling ultra-low Power ML at the Edge

The tinyML Community
https://www.linkedin.com/groups/13694488/
Subscribe to tinyML YouTube Channel for updates and notifications *(including this video)*

www.youtube.com/tinyML
tinyML Asia
Technical Forum

November 16, 2023
Seoul, South Korea

Register now
https://www.tinyml.org/event/asia-2023/
2023 Edge AI Technology Report

The guide to understanding the state of the art in hardware & software in Edge AI.

https://www.wevolver.com/article/2023-edge-ai-technology-report
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Ilya is a Senior Fellow and a Chief AI Architect at Grovety, where he worked out his way from a rising talent developer to a veteran expert in AI, a frontline and prospective trend in IT-industry in recent years. He acquired extensive experience in developing general and AI compilers, and chip architectures both in LLVM and TVM backend optimizations; he also led teams working on compiling-related projects, video processing, and protocols support for IP cameras (C/C++). Ilya received Master degree in Applied Mathematics and Computer Science in 2007. Wide range of projects and profound research activity makes Ilya’s experience valuable and demanded.
TinyML - growing interest

Edge AI allows business to improve the AI applications’ overall cost-effectiveness by optimal use of NNs, computing resources and power consumption reduction.

At the same time, the numerous potential benefits of Edge AI face several challenges associated with its implementation and usability. [1]
Fine-Tuning Strategies

**Model modification:**
- Compress off-the-shelf networks by pruning and quantization
- Simplify unsupported operations to primitive blocks
- Transform and merge network layers
- Optimize resource-intensive layers

**Inference time optimization:**
- Use hardware-specific acceleration instructions

**Memory requirements optimization:**
- Optimize schedule of the operation flow
- Store weights on external storage

**Energy Efficiency**
- Throttling MCU/NPU operating frequencies
- Use advantages of heterogenous systems
- Intelligent power management
Approaches to handle Cost Challenges in TinyML

- Minimizing Development costs and time of device
- Reducing Device cost and its power consumption
Fine-Tuning Strategy
Fine tuning platform for CI tests and experiments on target HW

- NN inference on Alif hardware and FVP simulator
- Run on TFLiteMicro and TVM runtimes
- Support of any TVM commit
- Unified API for running NN inferences, various architectures and runtimes
- Actual inference time and power consumption measurements
- Model bottlenecks analysis and numerical mismatches
TVM: ML Compiler Framework

Wide range of ML frameworks and deployment targets

Open-source project, large community

Integrated with ARM® Vela Compiler for acceleration on Ethos™-U55 NPU

Fine-grained control over model compilation, deployment and execution
Why Ethos-U and Alif Ensemble SoC

E7 Processor

- Cortex-M55 160 MHz
- Ethos-U55 128 MAC
- Cortex-A32 800 MHz
- Cortex-M55 400 MHz
- Ethos-U55 256 MAC
- Cortex-A32 800 MHz

Ethos-U55 microNPU

- Configurable MAC Engine
- Elementwise Engine
- System FLASH
- Local Memory
- System SRAM
- Weight Decode
- Control Unit
- DMA

DMA
Our Experience with Alif E5
Pooling with high strides

Benefits:
Inference speed-up: +25%
ARM ML Zoo Models affected: ~17%

Here we already know that IFM.shape == kernel.shape
Padding over channel axis

Benefits:
Inference speed-up: 250% - 400%
ARM ML Zoo Models affected: ~10%

Create a memory area with padding values "before"
Our actual channel data
Create a memory area with padding values "after"
Concatenate everything together over channel axis
Understanding TVM's patterns

Pad
- paddings (4+2)
  - 1x224x224x3

Conv2D
- filter (3x2x2x1)
  - padding = VALID
  - stride_h = 1
  - stride_w = 1
  - 1x223x223x3

Minimum
- input = 127

Optional pad = is_op("nn.pad")

QNN Conv2d = is_op("qnn.conv2d")
  - Optional pad | wildcard()
  - is_constant()
  - is_constant()
  - is_constant()

QNN Bias Add = is_op("qnn.bias_add")
  - qnn.conv2d, is_constant()

QNN Requantize = is_op("qnn.requantize")
  - is_constant(), is_constant(), is_constant(), is_constant(), is_constant()

Clip = is_op("clip")

Optional (is_op("clip"))
Open Challenges

• Per-layer Analysis: computational and memory usage
• The memory scheduling according to the overall network topology [2]
• Transitioning Network Weights to External Storage [5]
• On Device Learning
• Inference of multiple NNs on heterogeneous computing architectures
• Dig in Ethos-U Platforms specific
• More practice and experience on real applications
[1]: 2023 Edge AI Technology Report
https://www.wevolver.com/article/2023-edge-ai-technology-report

[2]: MCUNet: Tiny Deep Learning on IoT Devices
http://tinyml.mit.edu

[3]: Tiny Reservoir Computing for Extreme Learning of Motor Control


https://arxiv.org/abs/2101.08744

[6]: Work With microTVM
https://tvm.apache.org/docs/how_to/work_with_microtvm/index.html

[7]: Arm Ethos-N Processor Series, Product Brief
Thank you for your attention!

gozman@grovety.com
Copyright Notice

This multimedia file is copyright © 2023 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.tinylm.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.

www.tinyml.org