“Cracking a 600 million year old secret to fit computer vision on the edge”
Shivy Yohanandan - Xailient

July 6, 2021
tinyML Talks Sponsors

Additional Sponsorships available – contact Olga@tinyML.org for info
Arm: The Software and Hardware Foundation for tinyML

1. Connect to high-level frameworks
   - Profiling and debugging tooling such as Arm Keil MDK

2. Supported by end-to-end tooling
   - Optimized models for embedded
   - Runtime (e.g. TensorFlow Lite Micro)

3. Connect to Runtime
   - 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)
WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

- **Automatically compress** SOTA models like MobileNet to <200KB with little to no drop in accuracy for inference on resource-limited MCUs.

- **Reduce** model optimization trial & error from weeks to days using Deeplite’s **design space exploration**.

- **Deploy more** models to your device without sacrificing performance or battery life with our **easy-to-use software**.

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

Test
- Test impulse with real-time device data flows

www.edgeimpulse.com
The Eye in IoT
Edge AI Visual Sensors

- CMOS Imaging Sensor
  - Ultra Low power CMOS imager
  - AI + IR capable

- Computer Vision Algorithms
  - Machine Learning algorithm
  - <1MB memory footprint
  - Microcontrollers computing power
  - Trained algorithm
  - Processing of low-res images
  - Human detection and other classifiers

- IoT System on Chip
  - Machine Learning edge computing silicon
  - <1mW always-on power consumption
  - Computer Vision hardware accelerators

info@emza-vs.com
Enabling the next generation of **Sensor and Hearable products** to process rich data with energy efficiency
Maxim Integrated: Enabling Edge Intelligence

**Advanced AI Acceleration IC**

The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

[www.maximintegrated.com/MAX78000](http://www.maximintegrated.com/MAX78000)

**Low Power Cortex M4 Micros**

Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

[www.maximintegrated.com/microcontrollers](http://www.maximintegrated.com/microcontrollers)

**Sensors and Signal Conditioning**

Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

[www.maximintegrated.com/sensors](http://www.maximintegrated.com/sensors)
Qeexo AutoML

Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

End-to-End Machine Learning Platform

Data Cleaning & Preprocessing → Feature Extraction & Selection → Model Selection → Hyperparameter Optimization → Model Validation → Model Conversion (e.g., to C) → Compile Target + Specific ML Package

For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Automotive
- Smart Home
- Mobile
- Wearables
- IoT
Advancing AI research to make efficient AI ubiquitous

A platform to scale AI across the industry
Add Advanced Sensing to your Product with Edge AI / TinyML

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
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
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
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
Focus on:
(i) developing new use cases/apps for tinyML vision; and (ii) promoting tinyML tech & companies in the developer community

Submissions accepted until August 20th, 2021
Winners announced on September 1, 2021 ($6k value)
Sponsorships available: sponsorships@tinyML.org

https://www.hackster.io/contests/tinyml-vision
## Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday, July 7</td>
<td>Tobias Peikenkamp, Hahn-Schickard</td>
<td>Analysis of ECG Data by Energy Efficient Decision Trees on a Reconfigurable ASIC</td>
</tr>
</tbody>
</table>

Webcast start time is 8 am Pacific time

Please contact [talks@tinyml.org](mailto:talks@tinyml.org) if you are interested in presenting
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Shivy Yohanandan

Dr. Shivy Yohanandan is the co-founder and Chief Technology Officer at Xailient – the computer vision platform that is revolutionizing AI by teaching algorithms how to process images and video like humans! He holds a PhD in AI and Computer Science but started his career as a Neuroscientist and Bioengineer from the University of Melbourne. Passionate about vision, Shivy spent 4 years bringing vision to the blind by helping build Australia’s first bionic eye as a Research Engineer. Previously, Shivy worked as a research scientist for 3 years at IBM Research in AI for healthcare including computer vision in medical imaging and building a brain-machine interface to decode brainwaves for controlling a robotic arm.
Cracking a 600-Million-Year-Old Secret to Fit Computer Vision on the Edge

Cambrian period (ca 541 myo)
The bionic eye - how it works

First prototype: Wide-view neurostimulator

1. Camera captures image and transmits data to an external, body worn processing unit.

2. Data processed and sent to implanted system via external wire.

3. Implanted receiver passes signals onto retinal implant.

4. Implanted electrode array stimulates retina.

5. Electrical signals sent from retina via visual pathway to vision processing centres in the brain.

Image courtesy of Bionic Vision Australia
Worked with human eyes and brains for a decade...

...now ready to build machine eyes and brains.
My motivation to do a PhD was a meme lol

2015

AlphaGO
1920 CPUs, 280 GPUs,
1 MW

Human Brain
1 coffee,
20 W

1,000,000 W vs 20 W = 50,000x 😱
Training a single AI model can emit as much carbon as five cars in their lifetimes.

Deep learning has a terrible carbon footprint.
“Tesla is using a 1.8-exaFLOP AI supercomputer packed with 5,760 GPUs that train neural networks it hopes one day will power autonomous vehicles”
Are we really getting smaller?
Should you care?
US cities set up 'cooling centres' as historic heatwave bakes Pacific north-west

Temperatures in Canada and north-west US reached record highs on 29 June

Western Canada burns, deaths mount after world’s most extreme heat wave in modern history

It's Official: UN Just Confirmed Antarctica's New Heat Record at 18.3 Celsius

2021 WTF?!
• More people want AI on the **tiny edge** device, but AI can’t fit 😞

• Others solving this
  
  • **Model compression** (HW agnostic = you have more HW choice)
  
  • **Proprietary AI chips** (HW dependent = you have less choice)

• **Problem** with these solutions: accuracy, robustness and speed not good enough because most use e.g. YOLO, R-CNN, SSD
YOLO

You Only Look Once: Unified, Real-Time Object Detection

Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi
Classifies ~100,000 random regions (anchors)!
What YOLO’s classifier sees

What you and I see
IT'S NOT YOU
IT'S YOLO
YOLO Flaws

• Localization and classification on shrunken low-res image

• Throws away >95% of pixels = no wonder accuracy is a problem

• Cisco: “66% cameras will be 4k in coming years”
  • but YOLO will still have to shrink image to 416x416
  • that's like 1% of original image -- ouch!!

• YOLO is not scalable in Industry 4.0 😞
What about YOLOv123456?
It's not YOU, it's YOLO

What about tiling?
Bigger models overfit which is why you struggle with robustness.
Scaling Accuracy Croc Plot
Real-world Accuracy 😞

Hype Cycle for Artificial Intelligence, 2020
Research >>>GAP<<< Real-World

**Research:**
- 0 to 1
- If achieving 70% on a benchmark, move on to the next one

**Real-World:**
- Even 90% is not enough!
- Exacerbated by long tail of scenarios
• YOLO did good job for 6 years, but tiny edge constraints is new problem which requires new solution

• YOLO inspired from someone’s imagination

• Can we get inspiration from natural intelligence?
A solution?

OBJECT DETECTOR
13,800,000,000 BC

C O S M O S

A SPACETIME ODYSSEY
Stage 1

~600 million years of evolution!
The problem with YOLOs, SSDs & RCNNs
Face Recognition – Three Phases

1. Detecting a face: Is there a face in the input image?
2. Locating a face: Where is the face in the input image?
3. Identifying a face: Extract features and match with a database

Complexity (MACs) of each phase normalized to Phase 1

- Phase 1: 1x
- Phase 2: 59000x
- Phase 3: 2500x

Each step requires a different amount and type of computation.
Not just 🙄👂👌
Xailient Detectum™
(single class object detector)
>1,200 Detectum™ users!

Plenty of single-class applications in real-world 😊

console.xailient.com
“Face-off” on Raspberry Pi 3B+
Sony IMX500 AI chip

powered by a coin cell battery for a year
Advanced Computer Vision with MAX78000 & Xailient

https://www.youtube.com/watch?v=lDCcwBaKJDg
In the Australian Outback on Solar Power

Hive you heard the news? B honey’s Purple Hive Project has won a Webby Award in the Apps and Software category for Machine Learning and Bots. We are buzzing with excitement! P.S. Happy World Bee Day 🐝 #Bhoney #PurpleHiveProject #AI #techinnovations #tech
ANPR: License plate detection and reading

in a basement
with no internet
Creator: “toddler don't have patience so your vision algo better be fast!”
"we couldn't go faster because the camera wasn't fast enough!"
Models & Tools Tested

Single-Core Inference Time (Millisecond)

- Pelee (NCNN)
  - 5476 ms
- MobileNetV2-YOLO (NCNN)
  - 10x speed improvement
- NanoDet (NCNN)
  - 50x speed improvement
- QuickYOLO (LCE)
  - 54x speed improvement
- Xailient
  - 509x speed improvement

Speed improvement compared to YOLOv5-s on PyTorch

Platforms:
- Seeed NPI i.MX6ULL Cortex A7 800 MHz
- Raspberry Pi 2 B v1.1 Cortex A7 900 MHz
- ASUS TinkerBoard v1 Cortex A17 1.8 GHz
More under NDA!
You asked, we listened...

Xailient Detectum™
(multi class object detector)
White paper: https://www.xailient.com/technical-details
join Detectum™ waitlist @
detectum.xailient.com
Adaptive radiation: a single species rapidly adapts to fill available niches in an environment.
Adaptive radiation: a single species rapidly adapts to fill available niches in an environment.
Xailient AdRad™

• Continuous accuracy monitoring and fine-tuning loop!

• Detection & localization is the **biggest accuracy bottleneck** in current deployments because the **camera field of view has the highest variance**
Face Recognition – Three Phases

1. Detecting a face: Is there a face in the input image?
2. Locating a face: Where is the face in the input image?
3. Identifying a face: Extract features and match with a database

Each step requires a different amount and type of computation

Complexity (MACs) of each phase normalized to Phase 1

- Phase 1: 1x
- Phase 2: 59000x
- Phase 3: 2500x
Summary

6 years old

218,112 KB

600,000,000 years old

44 KB
Summary

• Unlike YOLO, Detectum™ came from cracking a 600 myo biology secret which revealed nature’s object detector NN

• Nature does clever localization technique (stage 1), then classification on high-res ROIs (stage 2), which is how it’s smaller, faster and more accurate!

• Smaller model (44 KB) means more robust (accurate)

• Detectum™ is software (HW agnostic) so you have more HW choice
What’s in it for you?

• I spent 9 years reverse engineering a 600 myo biology secret, and I’m sharing that with you so you don’t have to spend years reinventing that wheel.

• You spend less time fine-tuning models like YOLO, and more time on the actual use-case you’re trying to solve.

• **Object detection is just the tip of the iceberg**, you still have much more complex AI (stage 2) to do after that!

• *Imagine what you can now do with richer pixel info available more easily for your own AI?*
“understand the underlying principles and forces of flight”

“In 1799, he set forth the concept of the modern aeroplane as a fixed-wing flying machine”

Fundamental breakthroughs are made to last!
Not enough people seeking to understand fundamentals of natural intelligence...

...in order to build artificial intelligence
Copyright Notice

This multimedia file is copyright © 2021 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.

www.tinyML.org