“Data techniques that enable tiny computer vision in the real world”

Jelmer Neeven - Plumerai

August 11, 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

dgeimpulse.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
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.
FOTA HUB

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:
- completed AI platform — integrated with a 2.4” LCD Screen, onboard IMU (LIS3DHTR), microphone, buzzer, microSD card slot, light sensor, infrared emitter (IR 940nm)

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

Codecraft:
- no code Programming platform to Get Started With TinyML
- supports Arduino, Python, C or JavaScript etc.

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

Motion / Gesture / Speech / Smell / Sports
Barcode / Face / Image

Artificial Nose
AI Thermal Camera for Safe Camping
Azure IoT Squirrel Feeder

Sense    Train    Inference    Applications
The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

Analytics Toolkit Suite

AutoML

Data Collection

Test & Validation

Data Labeling

Code Generation

Model Output

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
The goal of On Device Learning (ODL) is to make edge devices “smarter” and more efficient by observing changes in the data collected and self-adjusting / reconfiguring the device's operating model. Optionally the “knowledge” gained by the device is shared with other deployed devices.

More sponsorships are available: sponsorships@tinyML.org
tinyML Neuromorphic Engineering Forum
Chair: Prof. Charlotte Frenkel (TU-Delft)

September 27, 2022
https://www.tinyml.org/event/tinyml-neuromorphic-engineering-forum

tinyML is a fast-growing initiative around low-power machine-learning technologies for edge devices. The scope of tinyML naturally aligns with the field of neuromorphic engineering, whose purpose is to replicate and exploit the way biological systems sense and process information within constrained resources.

More sponsorships are available: sponsorships@tinyML.org
tinyML EMEA Innovation Forum 2022
Chair: Prof. Francesco Conti (Univ of Bologna)
Connect, Unify, and Grow the tinyML EMEA Community
October 10-12, 2022

https://www.tinyml.org/event/emea-2022
in person in Cyprus, Grand Resort, Limassol

Registration is open now (late fee after Sept 15)

Keynote speakers:

Massimo BANZI
CTO, Arduino

Alberto L. SANGIOVANNI-VINCENTELLI,
UC-Berkeley, Cadence & Synopsys

More sponsorships: sponsorships@tinyML.org
Join Growing tinyML Communities:

- **Meetup**
  tinyML - Enabling ultra-low Power ML at the Edge

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

10.8k members in 45 Groups in 36 Countries

3k members & 8.4k followers
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)
www.youtube.com/tinyML
### Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday, August 19</td>
<td>Soham Chatterjee, Sleek Tech</td>
<td>ScaleDowStudy Group: Optimisation Techniques: Knowledge Distillation</td>
</tr>
</tbody>
</table>

Webcast start time is 5:30 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

Please use the Q&A window for your questions

tinyml.org/forums  youtube.com/tinyml
Jelmer Neeven

Jelmer Neeven is a deep learning scientist and software engineer with a passion for building smart tools and real-world AI applications. He obtained an MSc in AI from Maastricht University and has been with Plumerai for three years. He spends his time building efficient and robust AI for microcontrollers through researching novel model architectures, data techniques and training methods.
Data Techniques
that enable tiny Computer Vision
in the real World

Jelmer Neeven - jelmer@plumerai.com

TinyML Shenzhen - August 11th, 2022
Plumerai people detection

- Detects each person in view, even if partially occluded.
- Tracks people across time and assigns unique IDs.
- Indoors, outdoors, various lighting conditions.
- 1 MB footprint!

Deploy anywhere. Run on any device, with minimal resources.
Plumerai people detection

Plumerai

1 MB model size on Arm Cortex-A72
55 frames/s at 1.5 GHz, single core
How did we do it?

Highly intertwined, need to cover the entire stack

- Plumerai Data Pipeline
- Plumerai Tiny Models
- Plumerai Inference Engine
- Plumerai Hardware IP Core

Crucial for tinyML!

Custom architectures and training strategies

AI companion to your RISC-V core, easily fits in the smallest FPGAs

World’s fastest inference engine for Arm Cortex-M: 1.6x faster & 2x less RAM than TFLM

Starting point:
Starting point: Visual Wake Words dataset

Subset of Common Objects in Context (COCO)

- Aimed at microcontrollers
- 115k images
- 47% has people
- Bounding boxes available

dataset, Visual Wake Words, that represents a common microcontroller vision use-case of identifying whether a person is present in the image or not, and provides a realistic benchmark for tiny vision models. Within a lim-

1. Chowdhery et al. (2019), "Visual Wake Words Dataset."
2. https://cocodataset.org
Starting point: Visual Wake Words dataset

Theoretical performance

- mAP
- Model size
- Inference time

Amazing! :)

Real-life performance

Terrible :(

Plumerai  Baseline
We need to design a good dataset before we can design a good model!

- Irrelevant context
- Bias: must be worth looking at
- Shortcut: blue background, no people!
- Person-centered!

Why not just use public data?
Plumerai dataset

Not person-centered

Need tooling to make the most of your data!

Nothing to look at...
Plumerai infrastructure
Infrastructure overview

Model architecture

Training strategies
Infrastructure overview

- Data collection
- Labeling
- Data verification & categorization
- Deployment
- Monitoring

- Sample selection
- Model architecture
- Training strategies
- Augmentation
- Sensor devices
- Preprocessing

- Data unit tests
- Hardware
- Data curation
- Failure analysis

- Inference engine
Infrastructure overview

Data collection

Labeling

Sample selection

Model architecture

Training strategies

Data unit tests

Hardware

Augmentation

Sensor devices

Preprocessing

Deployment

Post-processing

Data verification & categorization

Failure analysis

Data curation

Monitoring

Inference engine
Auto-labeling

TinyML: *training* boxes don’t need to fit perfectly!
- Close is good enough
- Focus on **robustness**
- Use large off-the-shelf model for labeling
- Use **tooling** to identify and fix mistakes
Data unit tests

Unit Tests Passed: 73 / 84

- empty_desk
- empty_hallway
- empty_office
- person_desk_face
- person_desk_back
- person_office
- real_life_performance
- person_hallway_within_4m
- person_hallway_4_7m
- person_hallway_beyond_7m
- skin_tone_1
- skin_tone_2
- skin_tone_3

Cannot capture model behavior in a single metric!

Edge cases
Data curation: example issues

- Diversity
- Relevance
- Difficulty
- Mislabeling
- ...
Data curation: example issues

- Diversity
- **Relevance**
- Difficulty
- Mislabeling
- ...

![Fruit bowl](image)
Data curation: example issues

- Diversity
- Relevance
- Difficulty
- Mislabeling
- ...

[Image of a meeting room and a person holding food]

[Logo or icon on the right]
Data curation: example issues

- Diversity
- Relevance
- Difficulty
- Mislabeling
- ...
Data curation cycle

Need **good data tooling** to:

1. Test-driven development
   - Identify failure cases
2. Link problem to training data:
   - Visual similarity
   - Interactive classification
   - Training influence
   - ...
3. Re-train model
4. This is an iterative process!
   - Correct labels
   - Oversampling
   - Targeted augmentation
   - Supervised Contrastive Learning
   - Add / **remove** data
   - ...

**Quality > quantity**
Tooling demo:
Training Influence
Welcome, jelmer! Select an app in the sidebar.

jelmer
Tooling demo: Failure concepts
Welcome, jenner! Select an app in the sidebar.

| jenner |
Demo recap

- Tooling allows us to identify problematic images:
  - Searching for failure concepts
  - Influence on training process
Demo recap

• Tooling allows us to identify problematic images:
  • **Searching for failure concepts**
  • Influence on training process
Demo recap

- Tooling allows us to identify problematic images:
  - Searching for failure concepts
  - **Influence on training process**
Demo recap

• Tooling allows us to identify problematic images:
  • Searching for failure concepts
  • Influence on training process

• **Act on them in real-time!**

• **Custom tooling: quickly add / explore new AI-assisted approaches**
  • Needed for scalable data iteration
  • 30M images in total
  • Terabytes of data!
Other techniques

Oversampling

Targeted Augmentation

Prediction depth

Supervised Contrastive Learning
Conclusion

Production-worthy tinyML requires:

- **Vertical integration!**
- Rapid iteration
- High-quality data

High-quality model!
Run on any device, with minimal resources. It’s just a software update!
Thank you,
Feel free to ask any questions!

Jelmer Neeven - jelmer@plumerai.com

TinyML Shenzhen - August 11th, 2022
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.

www.tinyml.org