tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

“Speech-to-intent model deployment to low-power low-footprint devices”

Dmitry Maslov - Seeed Studio

August 31, 2021

www.tinyML.org

tinyML Talks Sponsors

TINY ML TALKS webcast

arm
tinyML Strategic Partner

Deeplite

EDGE IMPULSE
tinyML Strategic Partner

emza visual sense
tinyML Strategic Partner

GREENWAVES TECHNOLOGIES
tinyML Strategic Partner

LatentAI
Adaptive AI for a Smarter Edge
tinyML Strategic Partner

HOTG
tinyML Strategic Partner

maxim integrated™

Qeexo

Qualcomm
tinyML Strategic Partner

RealityAI®
tinyML Strategic Partner

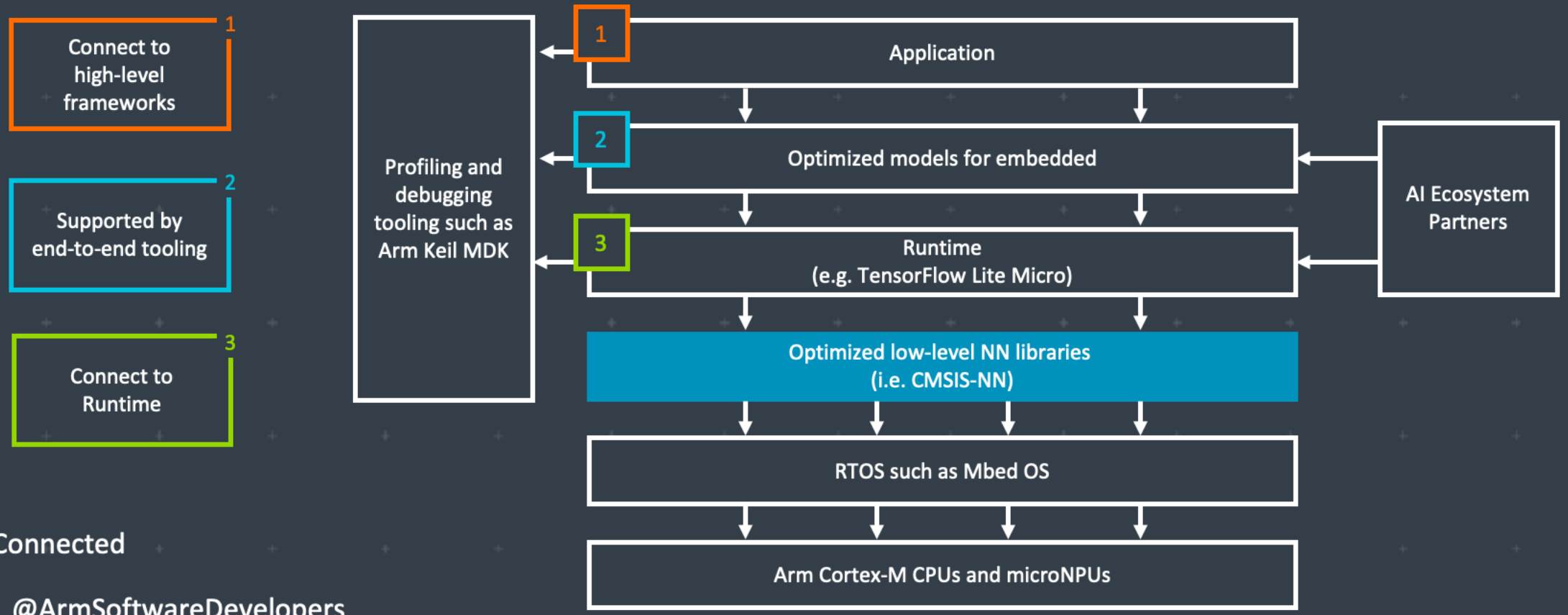
SensiML™
tinyML Strategic Partner

SynSense

SYNTIANT
tinyML Strategic Partner

Additional Sponsorships available – contact Olga@tinyML.org for info

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

arm

Deeplite

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**

BECOME BETA USER **bit.ly/testdeeplite**

mobilityXlab arm AI 100 CBINSIGHTS

TinyML for all developers

C++ library

Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

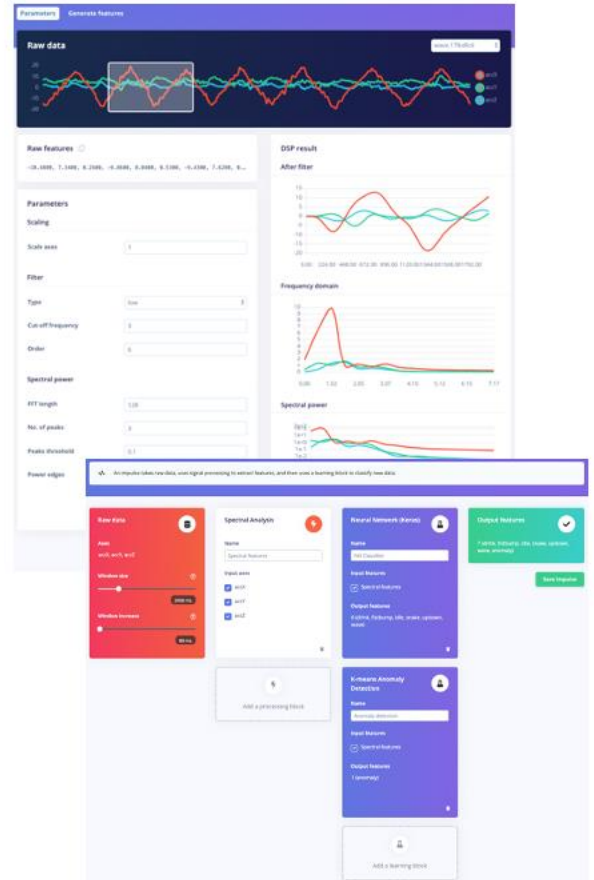
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test

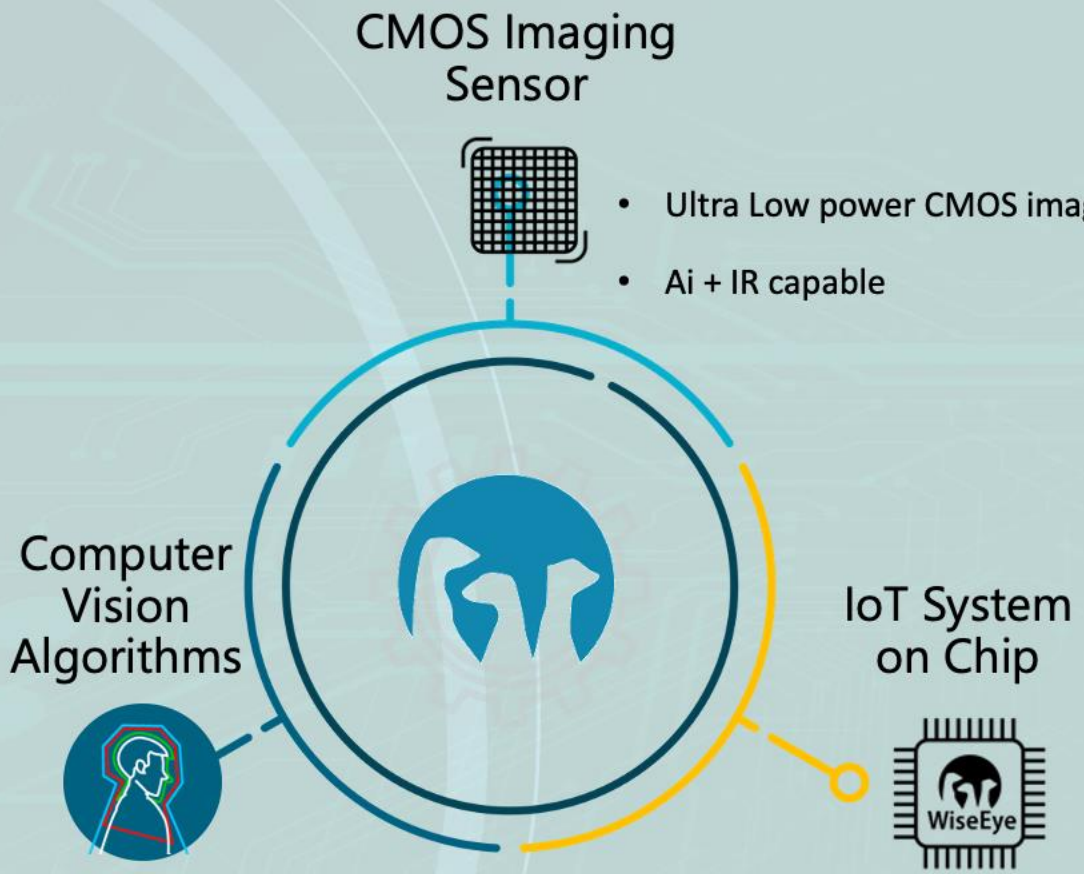


www.edgeimpulse.com

emza
visual sense

The Eye in IoT

Edge AI Visual Sensors



- Ultra Low power CMOS imager
- Ai + IR capable

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

- 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

Visible Image

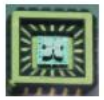
Sound

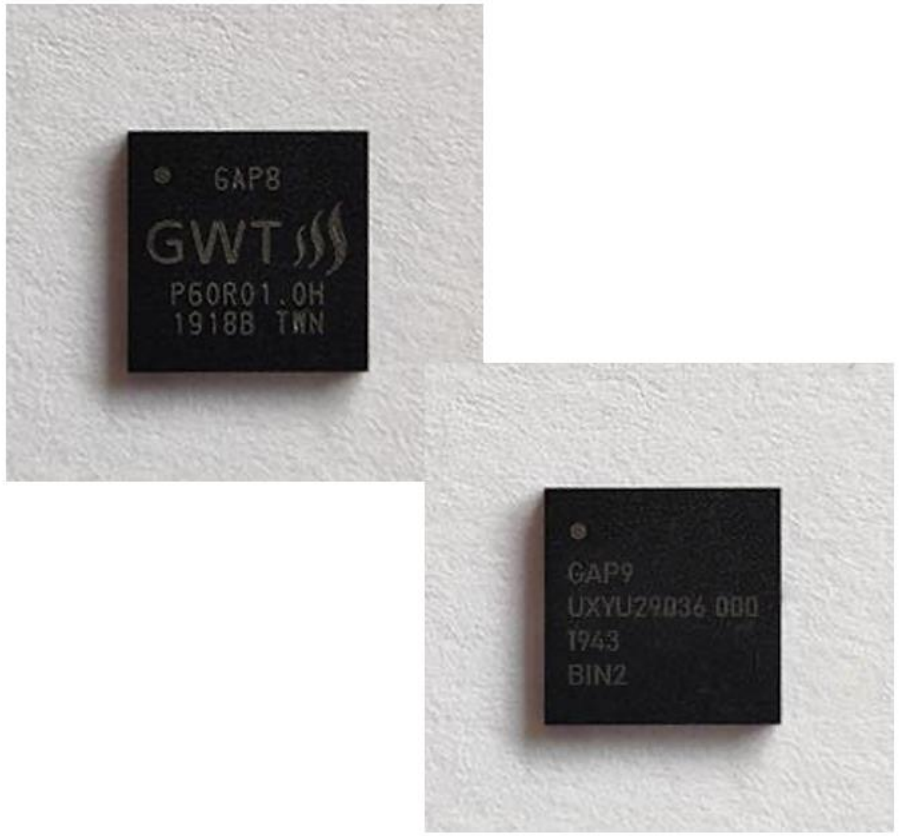
IR Image

Radar

Bio-sensor

Gyro/Accel

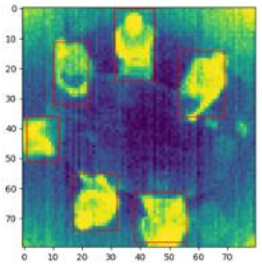




Wearables / Hearables

Battery-powered consumer electronics

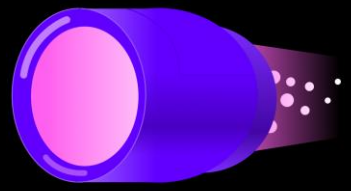
IoT Sensors

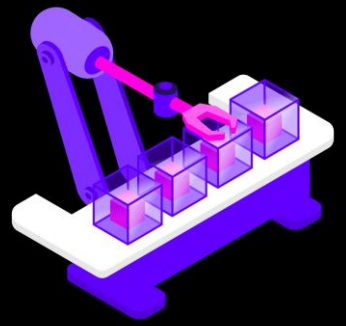


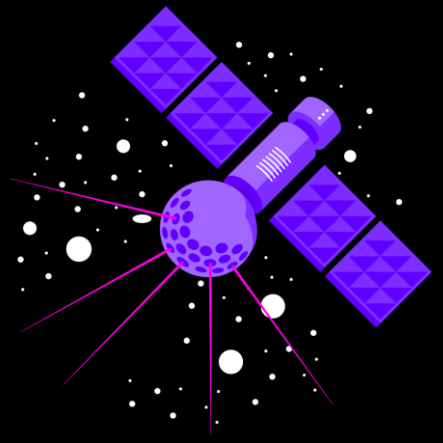
GREENWAVES TECHNOLOGIES

Distributed infrastructure for TinyML apps

HOTG
Decoupling intelligence







Develop at warp speed

Automate deployments

Device orchestration

HOTG is building the distributed infrastructure to pave the way for AI enabled edge applications

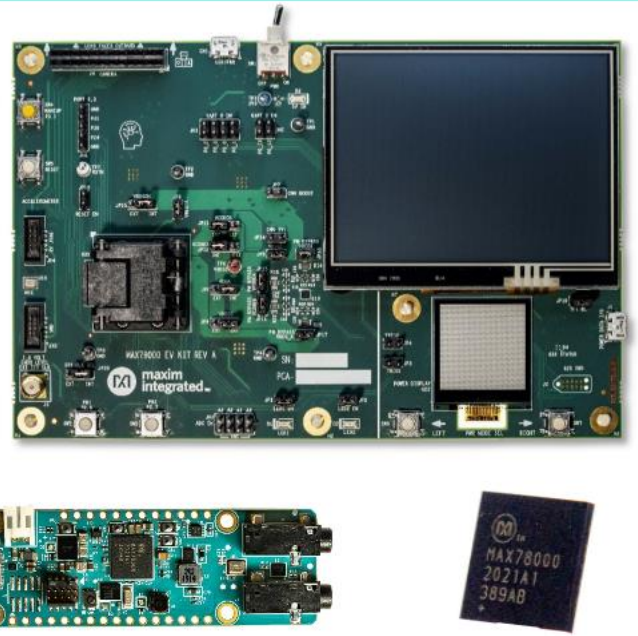
LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

Maxim Integrated: Enabling Edge Intelligence

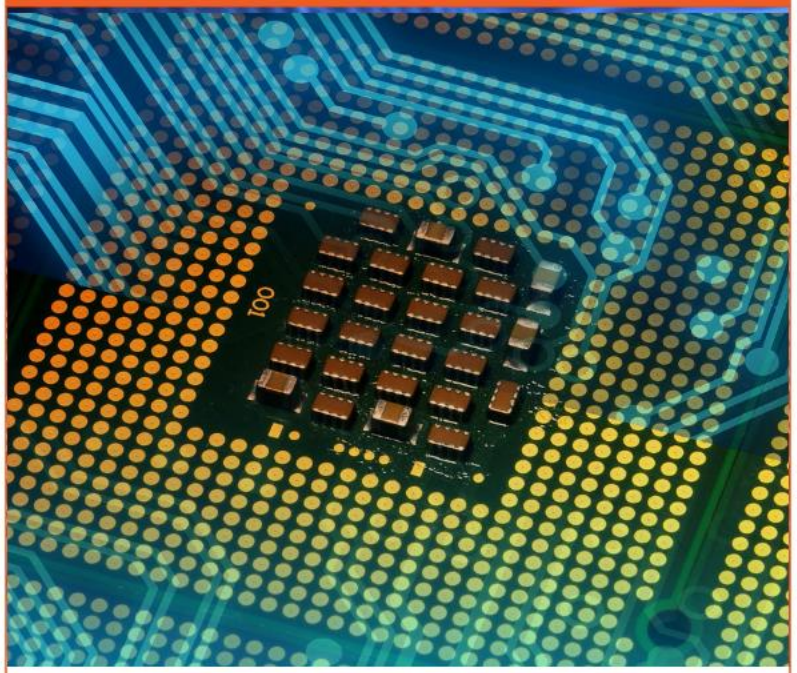
Advanced AI Acceleration IC



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

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

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

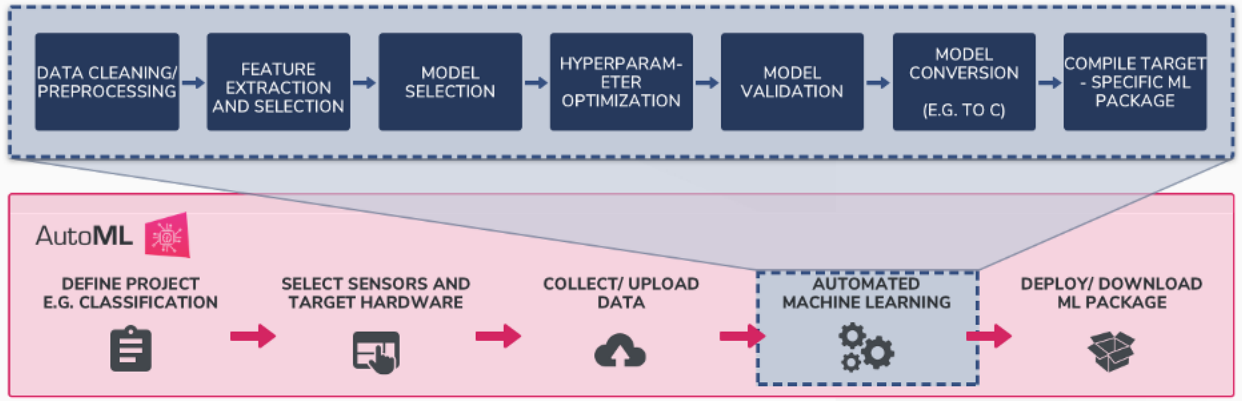
Qeexo AutoML

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



For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

Qualcomm
AI research

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

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai info@reality.ai @SensorAI Reality AI

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

SensiML™

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

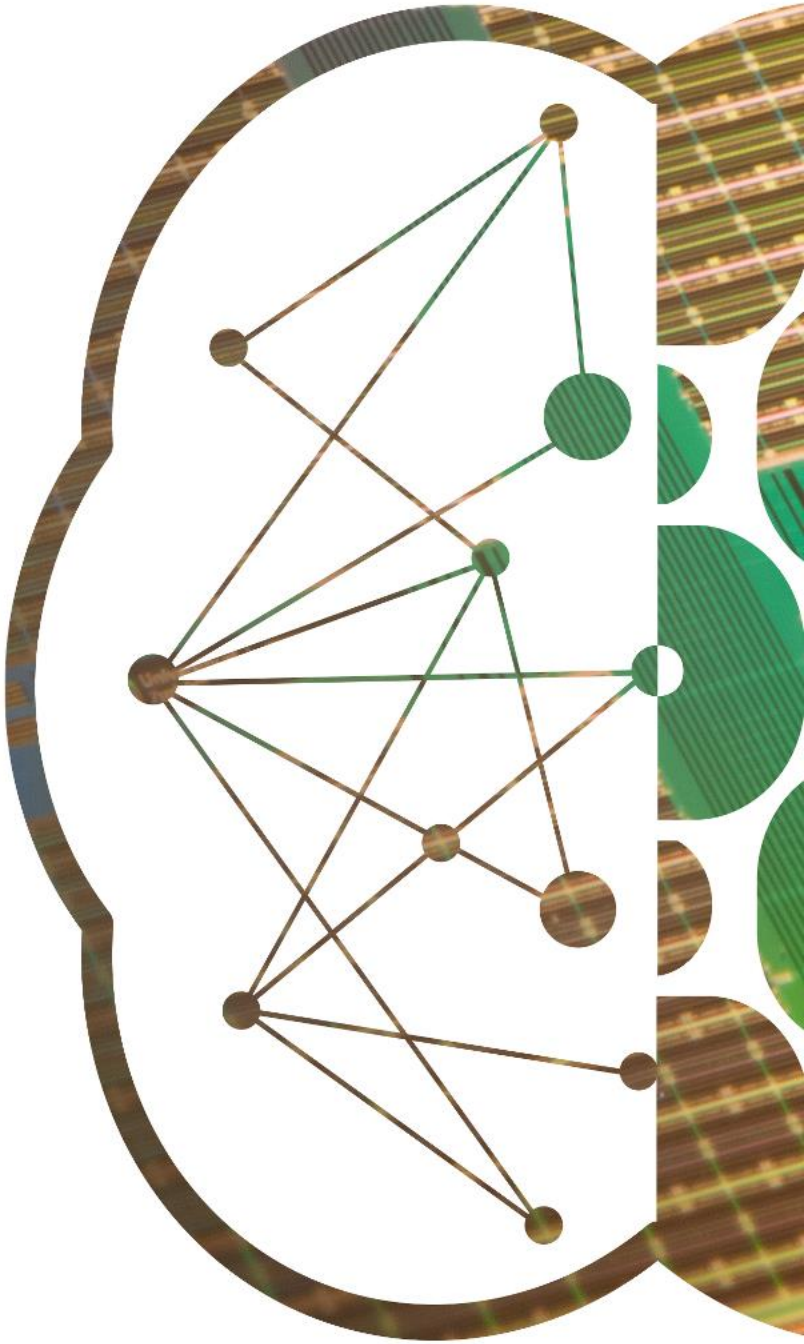
We enable the creation of production-grade smart sensor devices.

sensiml.com

SynSense

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

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

collaboration with hackster.io

AN AVNET COMMUNITY

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

Submissions accepted until September 17th, 2021
Winners announced on October 1st, 2021 ($6k value)
Sponsorships available: *sponsorships@tinyML.org*

https://www.hackster.io/contests/tinyml-vision

Successful tinyML EMEA 2021

TINY ML EMEA Virtual JUNE 7-10,2021

- Videos are available on www.youtube.com/tinyML

- **4** days of tinyML excitement
  - **2** tutorials
  - **5** keynotes
  - **15** tinyTalks
  - **7** lightning talks
  - **3** panel discussions & networking
  - **16** papers in the Student Forum
  - **4** partner sessions
  - **16** sponsoring companies

- **58** speakers, **1687** registered attendees!

250+ videos with 133k views

YouTube

Next tinyML Talks

Date	Presenter	Topic / Title
Thursday, September 9	**Pete Warden,** Technical lead of the TensorFlow Lite, Google	TinyML and the Developing World

Webcast start time is 8 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting

TINY ML TALKS webcast

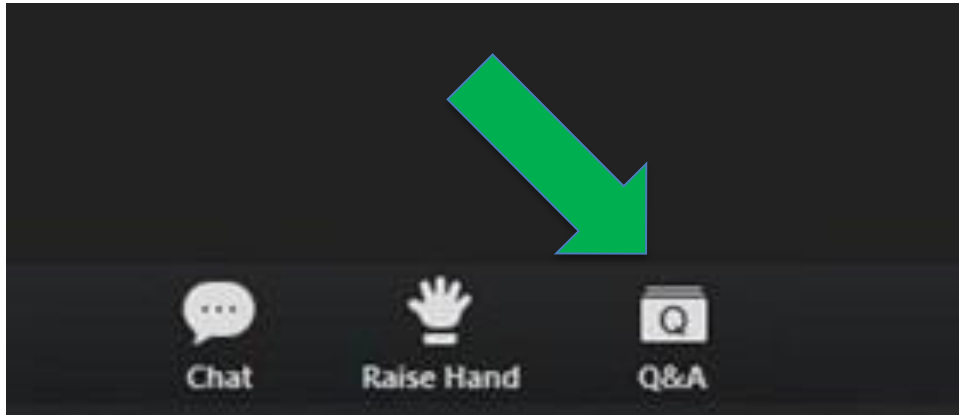
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions



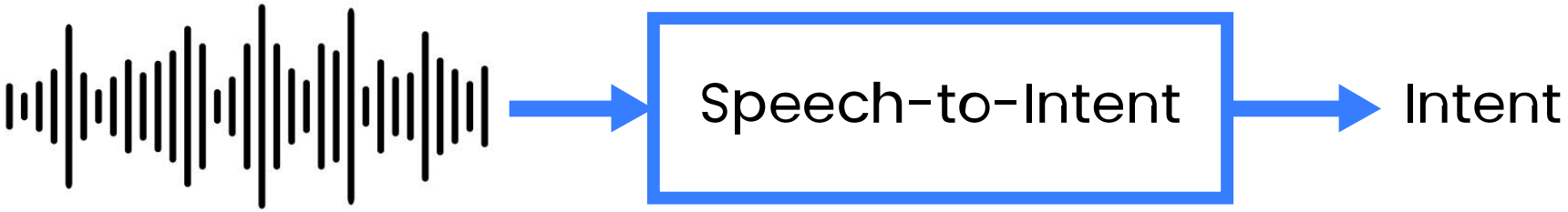
TINY ML TALKS webcast

Dmitry Maslov

Dmitry Maslov is a Machine Learning engineer, working at Seeed Studio on applications for embedded devices, both MCUs and SBCs. Recently he published a series of TinyML projects combined into a course, where he utilizes Edge Impulse/Tensorflow Lite for Microcontrollers to tackle challenging sensor data analysis tasks using Seeed Studio's Wio Terminal as a reference hardware. He also runs the Hardware.ai YouTube channel that is focused on embedded ML and robotics.

TINY ML TALKS webcast

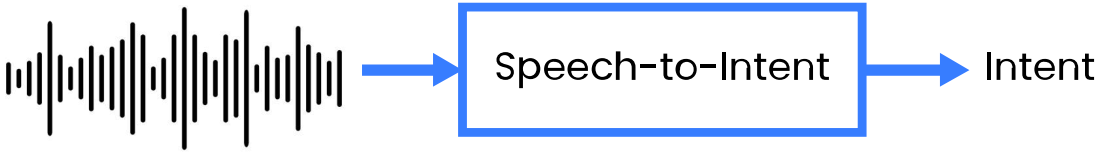
Speech-to-intent models on low-power and low-footprint devices



TensorFlow Lite

Different types of speech processing

- Open domain speech transcription, aka STT
- Keyword spotting
- Speech-to-Intent



TINY ML TALKS webcast

Keyword spotting

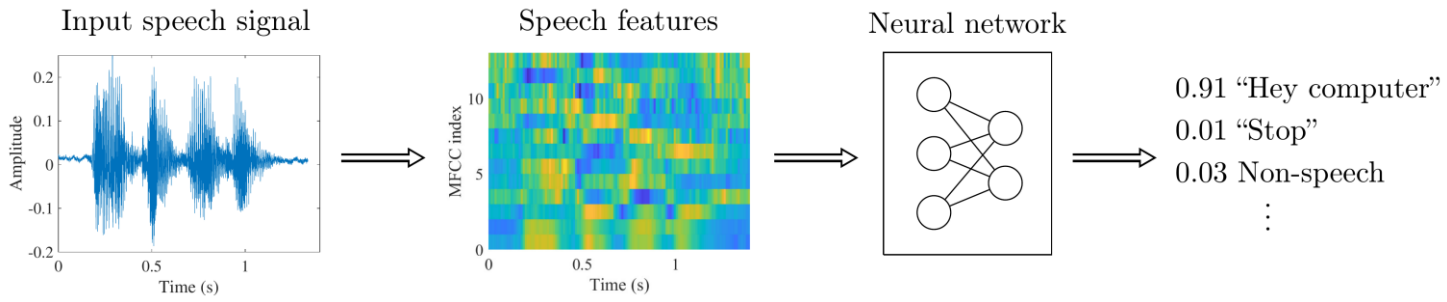
Keyword spotting is a task of detecting a key phrase within a stream of audio.

Benefits:

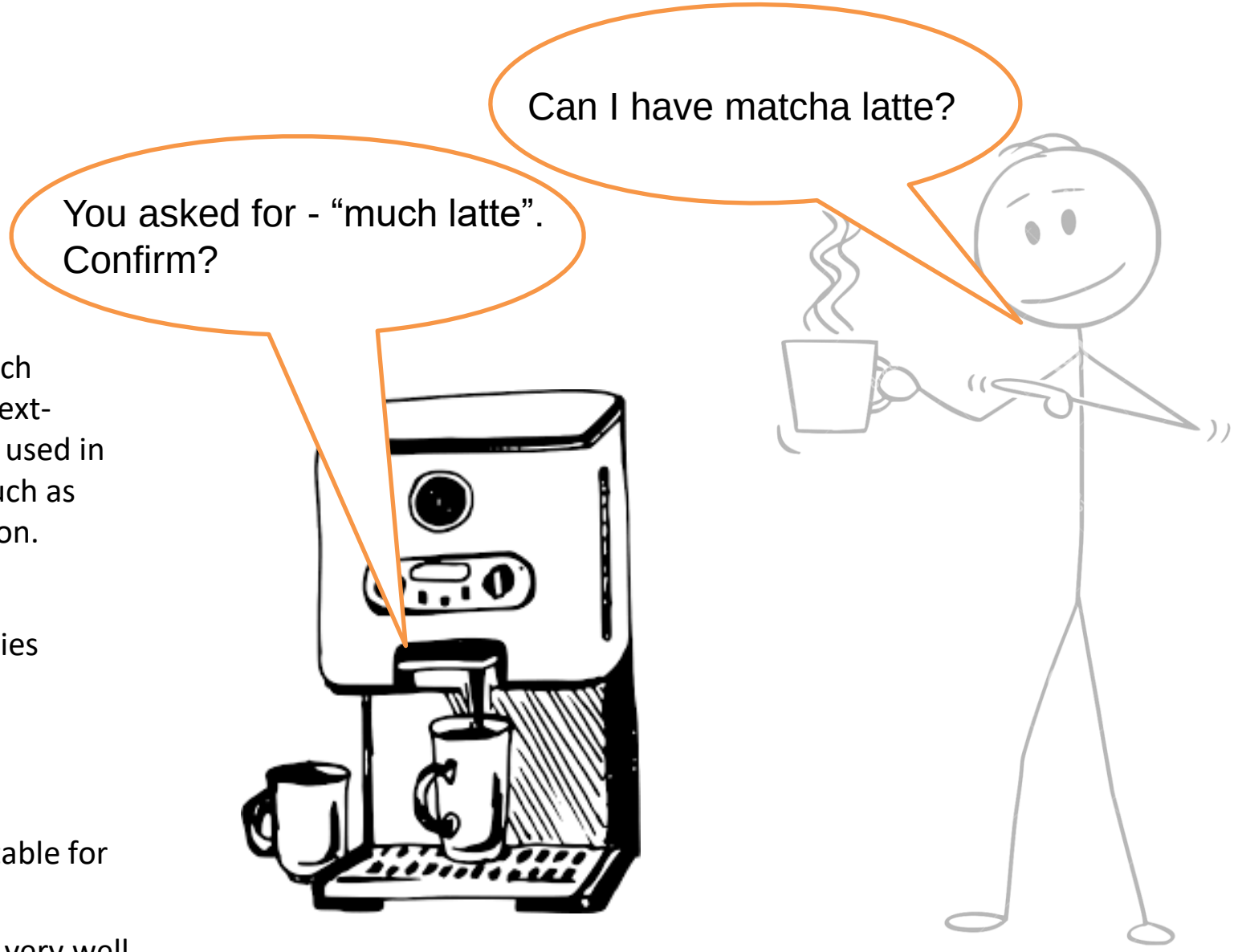
- very small models, fit even to very constrained resources devices
- normally faster than real time

Drawbacks:

- Recognizing single separate words becoming more difficult as vocabulary size increases
- Needs complete retraining for a new task



TINY ML TALKS webcast



Large-vocabulary continuous speech recognition (LVCSR) models + Text-based Natural language parser, used in mostly Cloud-based systems, such as Alexa, Google Assistant and so on.

Benefits:

- Can react to wide variety of queries
- Robust to environmental noises

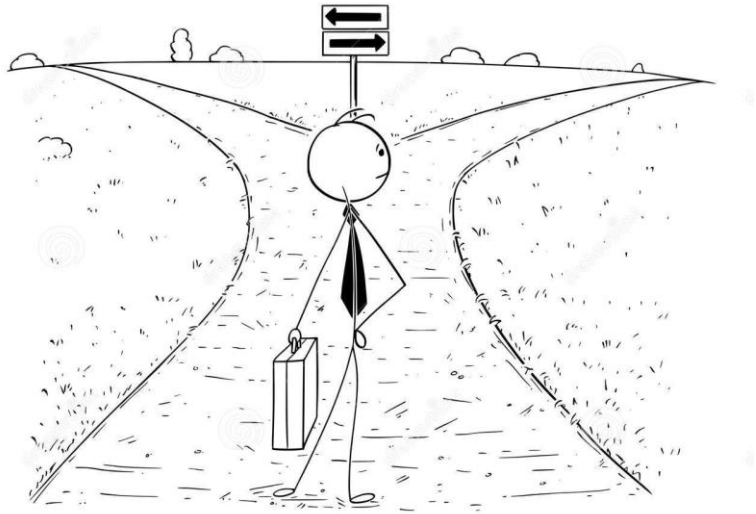
Drawbacks:

- demanding a lot of resource and computing power, not very suitable for the Edge computing
- Struggle with specific words, not very well represented in common use-case scenarios

Open-domain transcription + NLU

Comparison

Keyword spotting works well on microcontrollers, fairly easy to train with variety of no-code open-source tools available, e.g. Edge Impulse, but cannot handle large(er) vocabularies well



LVCSR models + Text-based Natural language parser approach is robust and somewhat easier to implement, given abundance of publicly available ASR engines, but is not suitable for running even on SBCs, let alone microcontrollers

Enter the better way

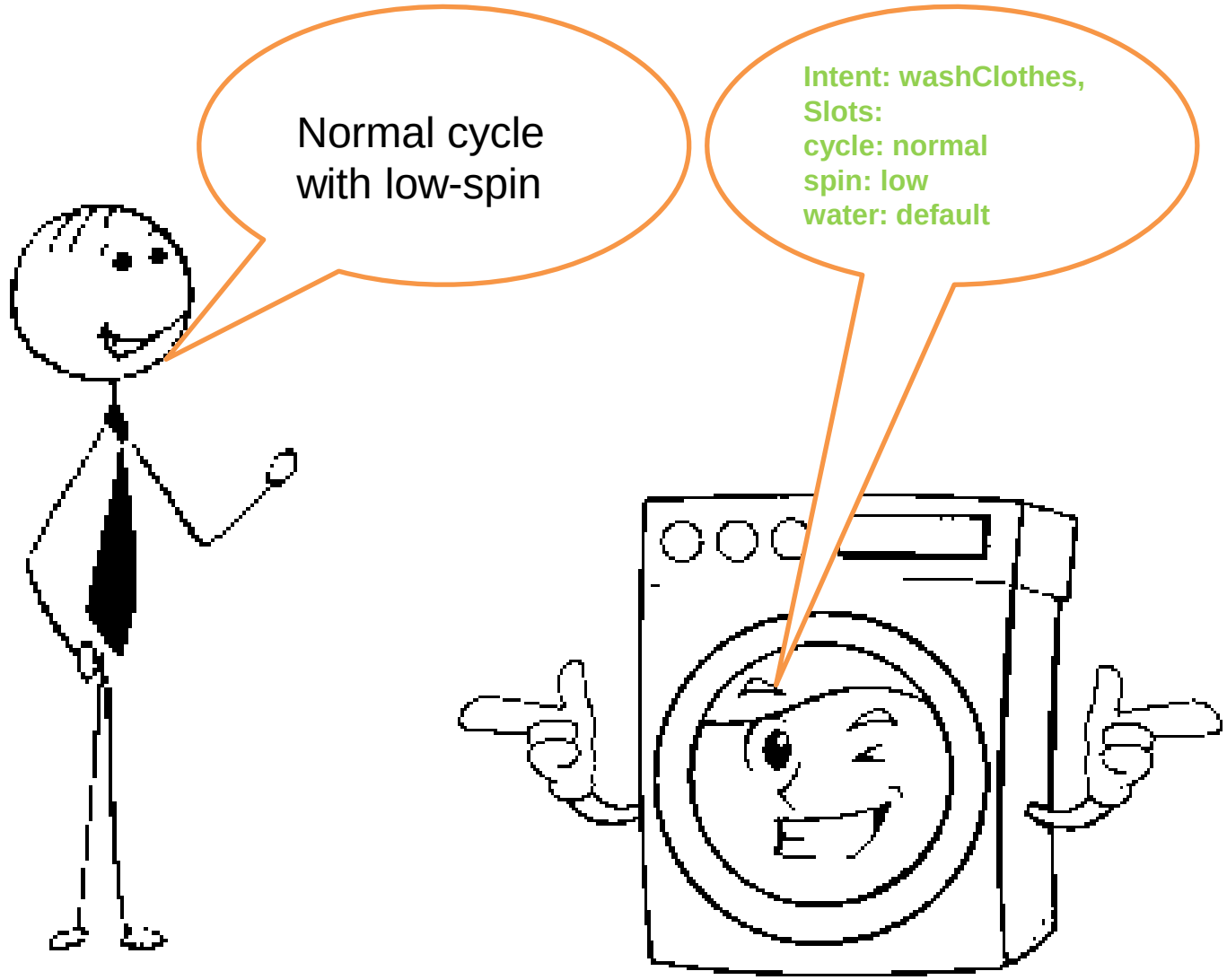
Well represented in research, but lacking widely available open-source implementations suitable for microcontrollers.

Production-ready, not open-source:

- Picovoice
- Fluent.ai

Production-ready, FOSS, not suitable for microcontrollers:

- Speechbrain.io



TINY ML TALKS webcast

The device

ARM Cortex M4F-based 120 Mhz, overclock to 200 MHz

320x240 LCD screen

Built-in sensors: microphone, light sensor, accelerometer

Wi-Fi + Bluetooth

Highly compatible with variety of ML training and deployment platforms: easy to set up and use TFLite Micro and Edge Impulse (code and no-code solutions available)

TINY ML TALKS webcast

TINY ML TALKS webcast

The dataset

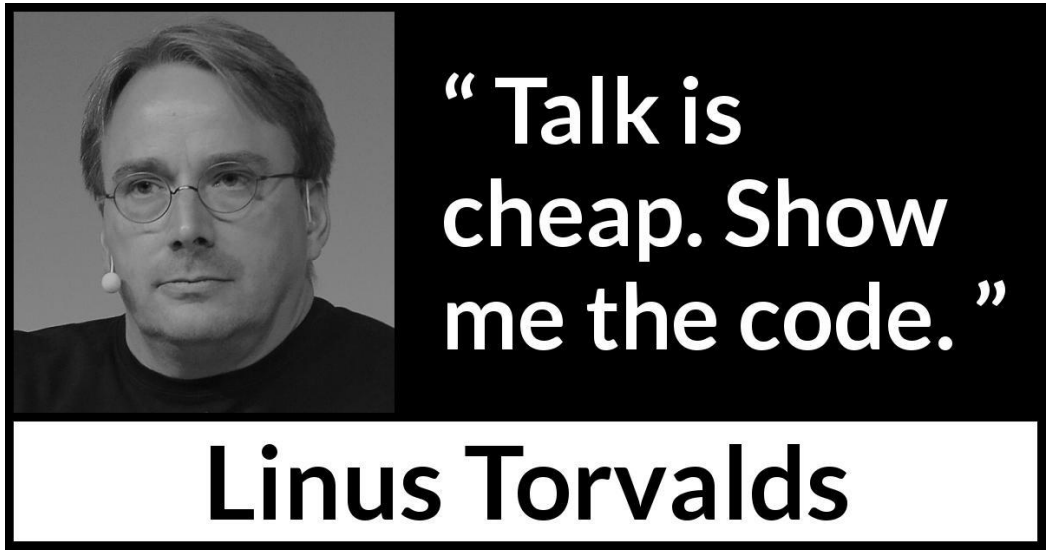
FLUENT SPEECH COMMANDS DATASET

Dataset contains 97 speakers saying 248 different phrases. The 248 utterances map to 31 unique intents, that are divided into three slots: action, object, and location. The goal in preparing this dataset was to provide a benchmark for end-to-end spoken language understanding models.

Trained on Fluent Speech Commands dataset, tested on 160 phrases recorded with Wio Terminal microphone.

No samples from Wio Terminal were used in training, although the training samples were augmented with audiomentations library, since Fluent Speech Commands dataset recordings all have very clean, noise-free sound.

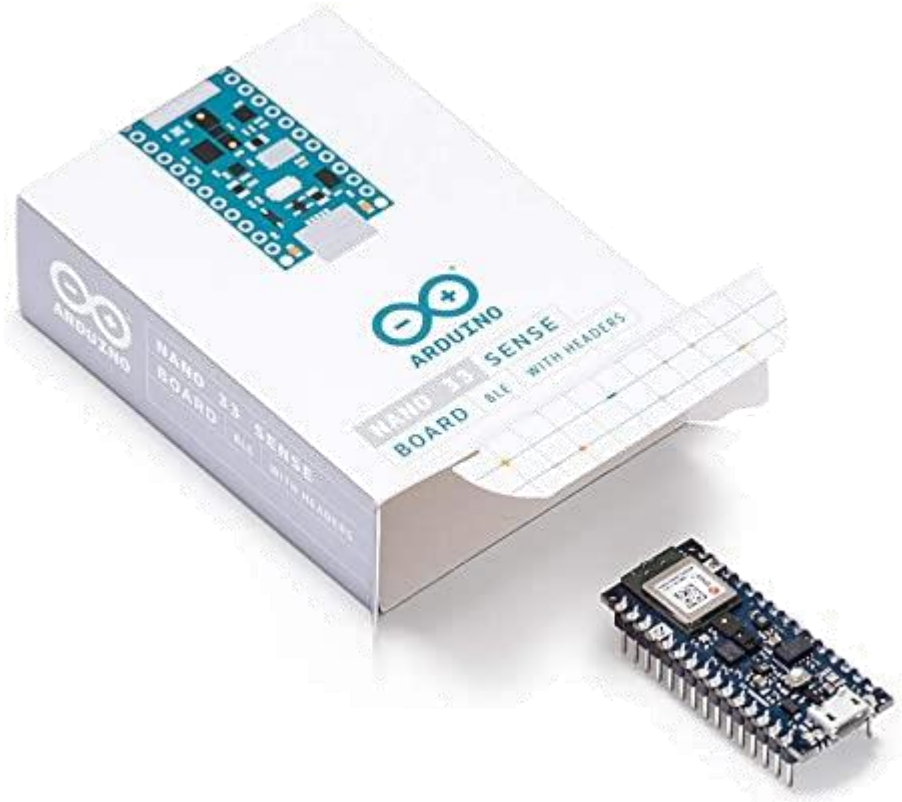
TINY ML TALKS webcast



One of the main motivations behind my work on this project was to create an open-source easily accessible package for training and deploying Speech-to-Intent models on Microcontrollers and SBCs. While it may not be fancy or SOTA at this point, it can (and will) be extended with more advanced model architectures and training techniques (more on that in improvement section).

TINY ML TALKS webcast

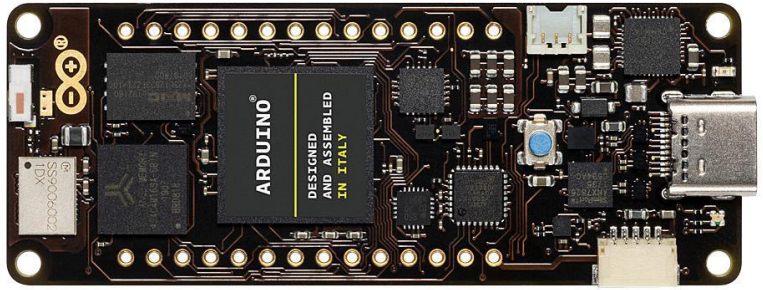
Other devices?

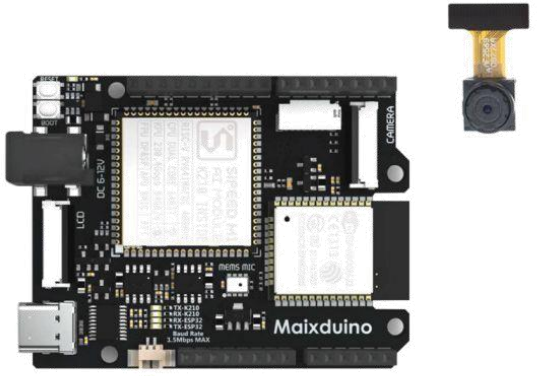


Can easily be run on:

- other Cortex M4F microcontrollers
- other ARM Cortex MCUs
- SBCs with Tensorflow Lite Interpreter

- other architecture MCUs (ESP32, K210)





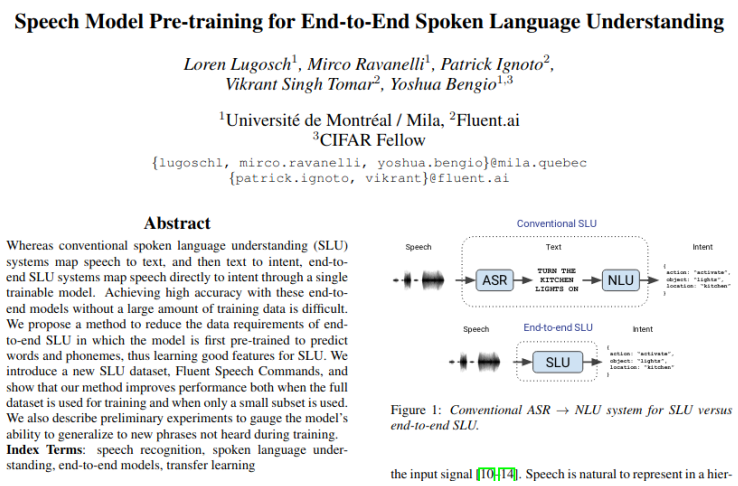
Improvement

While it works as is, there are many things that can be improved:

- model pre-training
- seq2seq, LSTM, attention
- trainable filters
- AutoML, synthetic data

TINY ML TALKS webcast

Model pre-training



Speech Model Pre-training for End-to-End Spoken Language Understanding

Loren Lugosch[1], *Mirco Ravanelli*[1], *Patrick Ignoto*[2], *Vikrant Singh Tomar*[2], *Yoshua Bengio*[1,3]

[1]Université de Montréal / Mila, [2]Fluent.ai
[3]CIFAR Fellow
{lugoschl, mirco.ravanelli, yoshua.bengio}@mila.quebec
{patrick.ignoto, vikrant}@fluent.ai

Abstract

Whereas conventional spoken language understanding (SLU) systems map speech to text, and then text to intent, end-to-end SLU systems map speech directly to intent through a single trainable model. Achieving high accuracy with these end-to-end models without a large amount of training data is difficult. We propose a method to reduce the data requirements of end-to-end SLU in which the model is first pre-trained to predict words and phonemes, thus learning good features for SLU. We introduce a new SLU dataset, Fluent Speech Commands, and show that our method improves performance both when the full dataset is used for training and when only a small subset is used. We also describe preliminary experiments to gauge the model's ability to generalize to new phrases not heard during training.

Index Terms: speech recognition, spoken language understanding, end-to-end models, transfer learning

Figure 1: *Conventional ASR → NLU system for SLU versus end-to-end SLU.*

the input signal [T0,T4]. Speech is natural to represent in a hier-

arXiv:1904.03670 [eess.AS]

It is intuitive, that as long as language is the same, transfer learning can be leveraged to improve model accuracy and decrease training time.

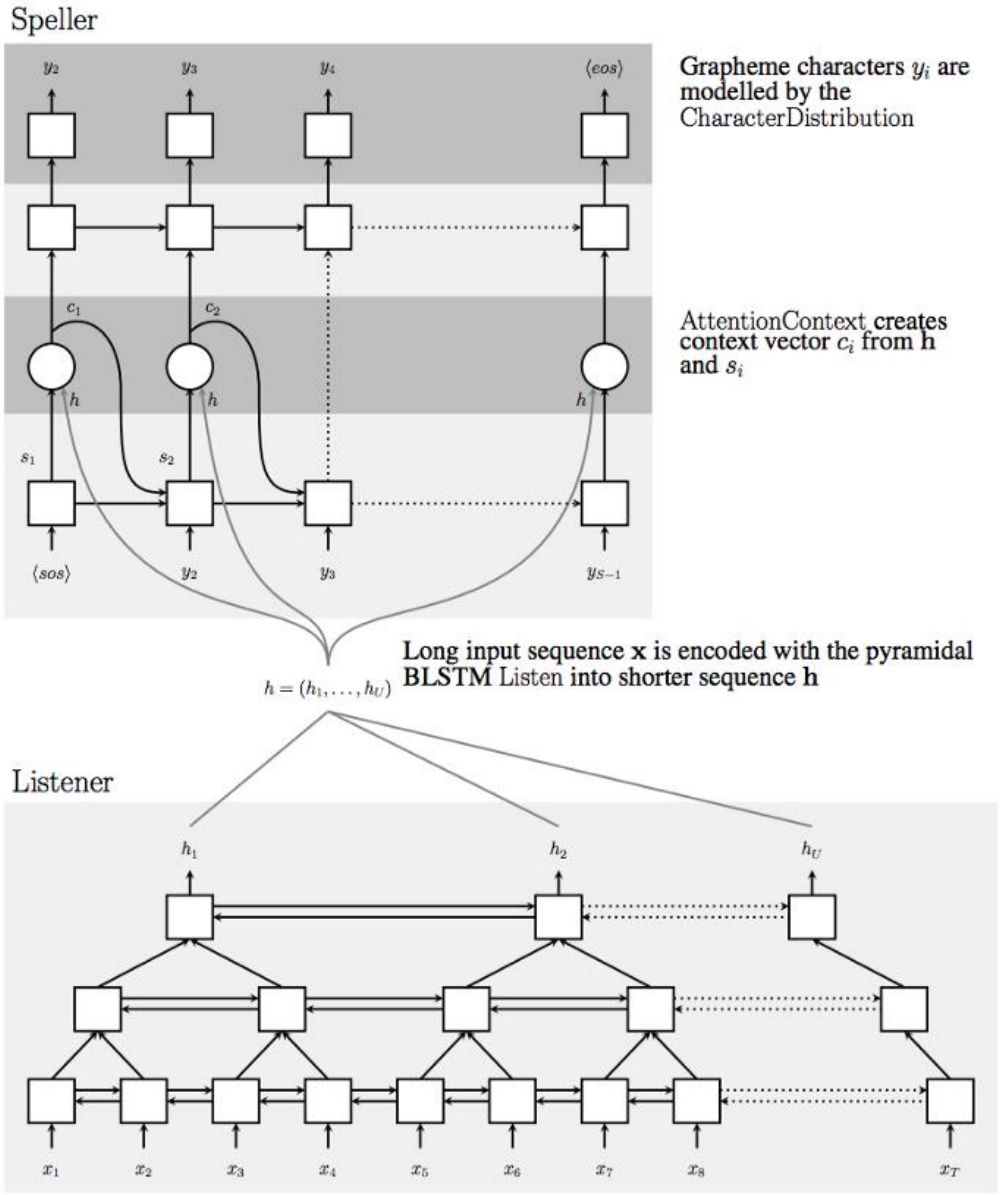
General idea is to pre-train the model on phoneme classification task and then transfer this knowledge to classifying intents and slot, which are words consisting of phonemes.

TINY ML TALKS webcast

Seq2seq, LSTM, attention

All of the model architectures present in Speech-to-Intent package at the moment only utilize Convolutional and Fully Connected layers with Global Max Pooling after feature extractor. That hinders model ability to process time-domain information and also limits it to fixed-length inputs and outputs.

Listen, Attend and Spell (LAS) model consists of two sub-modules: the listener and the speller. The listener is an acoustic model encoder, whose key operation is Listen. The speller is an attention-based character decoder, whose key operation is AttendAndSpell. The Listen function transforms the original signal x into a high level representation h = (h1,...,hU) with U ≤ T, while the AttendAndSpell function consumes h and produces a probability distribution over character sequences.



arXiv:1508.01211 [cs.CL]

TINY ML TALKS webcast

Trainable filters

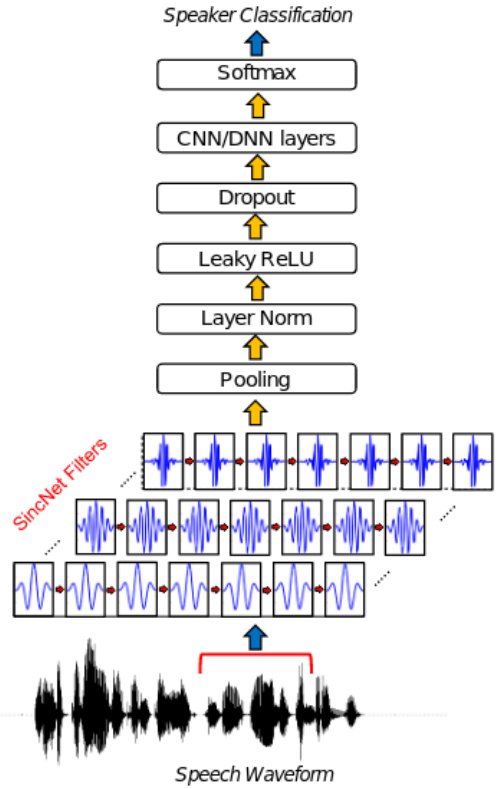


Fig. 1: Architecture of SincNet.

arXiv:1808.00158 [eess.AS]

While a lot of the times we rely on computing input features for speech recognition model with hand-written algorithms (FFT, MFCC, MFE) it is possible to allow the network to learn the discriminatory filters during training process.

"SincNet is based on parametrized sinc functions, which implement band-pass filters. In contrast to standard CNNs, that learn all elements of each filter, only low and high cutoff frequencies are directly learned from data with the proposed method. This offers a very compact and efficient way to derive a customized filter bank specifically tuned for the desired application."

TINY ML TALKS webcast

- Listen, Attend and Spell William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals https://arxiv.org/abs/1508.01211
- Speech Model Pre-training for End-to-End Spoken Language Understanding Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, Yoshua Bengio https://arxiv.org/abs/1904.03670
- Speaker Recognition from Raw Waveform with SincNet Mirco Ravanelli, Yoshua Bengio https://arxiv.org/abs/1808.00158
- FLUENT SPEECH COMMANDS DATASET https://www.kaggle.com/tommyngx/fluent-speech-corpus
- Small-Footprint Open-Vocabulary Keyword Spotting

with Quantized LSTM Networks Th´eodore Bluche, Ma¨el Primet, Thibault Gisselbrecht https://arxiv.org/pdf/2002.10851.pdf

TINY ML TALKS webcast

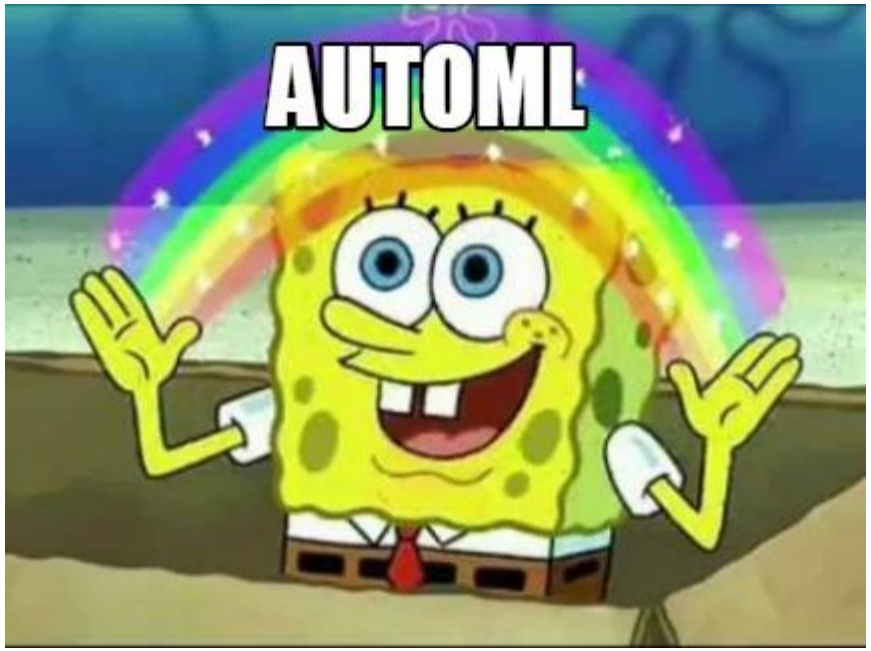
Project repository https://github.com/AIWintermuteAI/Speech-to-Intent-Micro

TinyML Course with Wio Terminal https://wiki.seeedstudio.com/Wio-Terminal-TinyML/

TinyML Course Playlist https://www.youtube.com/playlist?list=PL5efXgSvwk9UCtJ6JKTyWAccSVfTXSlA3

It might not give you SOTA, but at least you can say “Well, the computer tried everything and this is the best it could come up with.”

Specifically it can be very useful for hyperparameter search and finding the right number of filters in Convolutional layers, that can produce the best model for device’s RAM and FLASH memory constraints.



TINY ML TALKS webcast

And the answer is - absolutely!

We encourage you to fork the code repository, try training on your own dataset and perhaps try implementing more advanced architectures or model training techniques.

TINY ML TALKS webcast

"hardware.ai youtube"

"seeed studio"

Hardware.ai

Workshop

Graphical Programming for TinyML, the Easiest way to Start with Embedded ML

9:30 AM - 10:30 AM

Add to Calendar

Huiying Lai
Application Engineer, Seeed

Benjamin Cabé
Principal Program Manager, Azure IoT

Panel

Inspiring the Next Generation of ML Developers

8:30 AM - 9:00 AM

Add to Calendar

Moderator: Jen Fox
Sr. Program Manager, Microsoft

Mike Senese
Executive Editor, Make Magazine

Eric Pan
Founder and CEO, Seeed Studio

seeed

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyML.org