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

Dmitry Maslov - Seeed Studio

August 31, 2021
tinyML Talks Sponsors

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
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

- 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

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

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

Computer Vision Algorithms

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
HOTG is building the distributed infrastructure to pave the way for AI enabled edge applications.
Adaptive AI for the Intelligent Edge

LatentAI.com
The new MAX78000 implements AI inferencing 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

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

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
- Automotive
- Smart Home
- Mobile
- Wearables
- IoT
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
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 September 17th, 2021
Winners announced on October 1st, 2021 ($6k value)
Sponsorships available: sponsorships@tinyML.org

https://www.hackster.io/contests/tinymlvision
Successful tinyML EMEA 2021

- Videos are available on [www.youtube.com/tinyML](http://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](http://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>Thursday, September 9</td>
<td><strong>Pete Warden</strong>, Technical lead of the TensorFlow Lite, Google</td>
<td>TinyML and the Developing World</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
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.
Speech-to-intent models on low-power and low-footprint devices
Different types of speech processing

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

Hey Jarvis!
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
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

Can I have matcha latte?

You asked for - “much latte”. Confirm?
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

Intent: washClothes, Slots:
cycle: normal
spin: low
water: default

Normal cycle with low-spin
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)
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.
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).
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)
While it works as is, there are many things that can be improved:
- model pre-training
- seq2seq, LSTM, attention
- trainable filters
- AutoML, synthetic data
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.

arXiv:1904.03670
[ees.s.AS]
Seq2seq, LSTM, attention

All of the model architectures present in a speech-to-text 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 $h = (h_1, ..., h_U)$ with $U \leq T$, while the AttendAndSpell function consumes $h$ and produces a probability distribution over character sequences.

arXiv:1508.01211 [cs.CL]
Trainable filters

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.”

arXiv:1808.00158 [eess.AS]
- 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

- 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 Théodore Bluche, Mäel Primet, Thibault Gisselbrecht
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=PL5efXgSvwk9U CtJ6JKTyWAccSVfTXSIA3
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.
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.
“hardware.ai youtube”
“seeed studio”
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.