tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"On-Device Domain Adaptation for Noise-Robust Keyword Spotting"

Cristian Cioflan – Doctoral Student, ETH Zürich

November 16, 2023

www.tinyML.org

TINY ML TALKS webcast

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

TINY ML TALKS webcast

Executive Strategic Partners

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

Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partner

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML TALKS webcast

Gold Strategic Partners


Build the
Future of tinyML
on arm

TINY ML TALKS webcast

EDGE IMPULSE

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

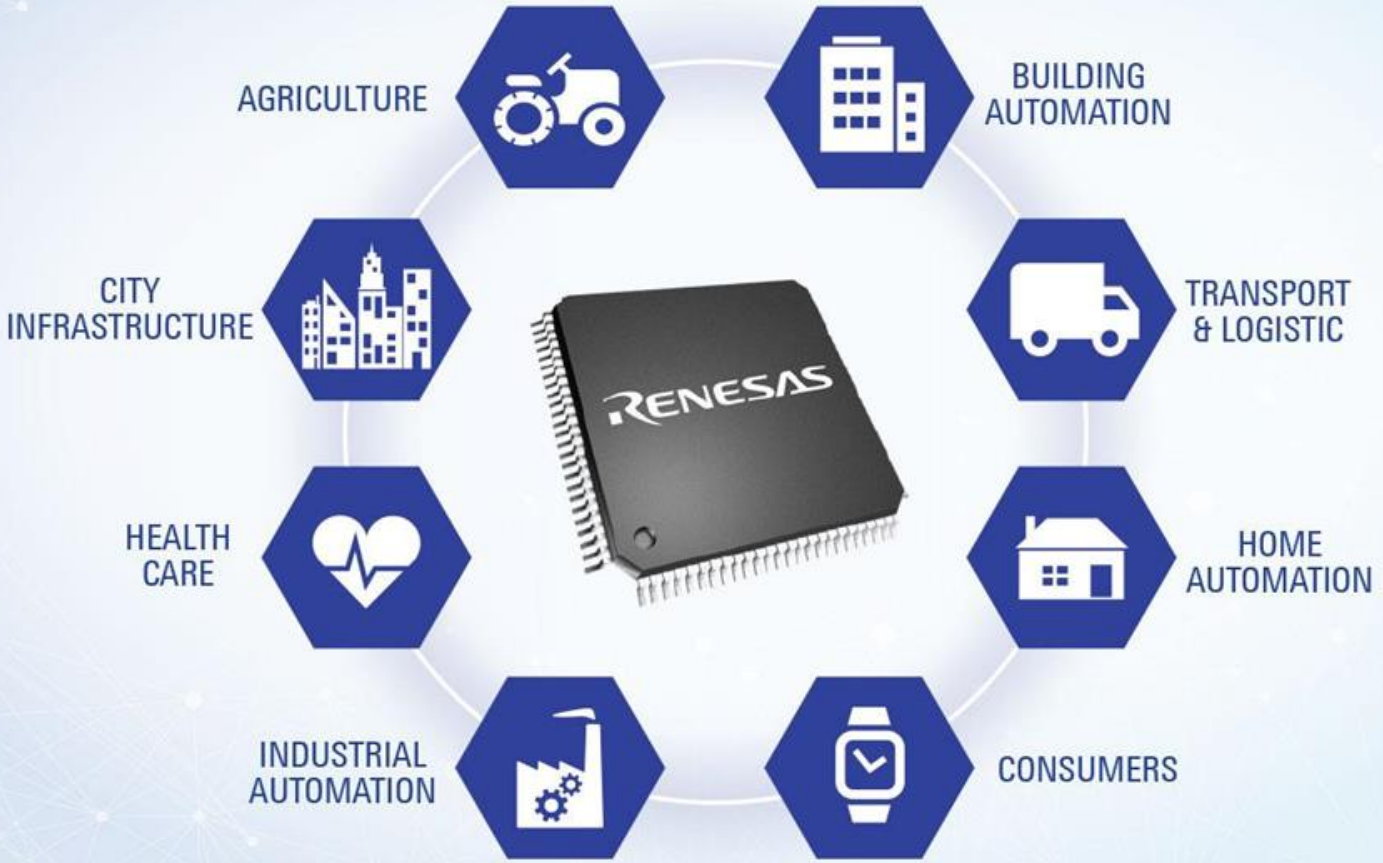
www.infineon.com

innatera

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.



renesas.com

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

www.st.com/ai

synaptics

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

TINY ML TALKS webcast

Silver Strategic Partners

Join Growing tinyML Communities:

17.6k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

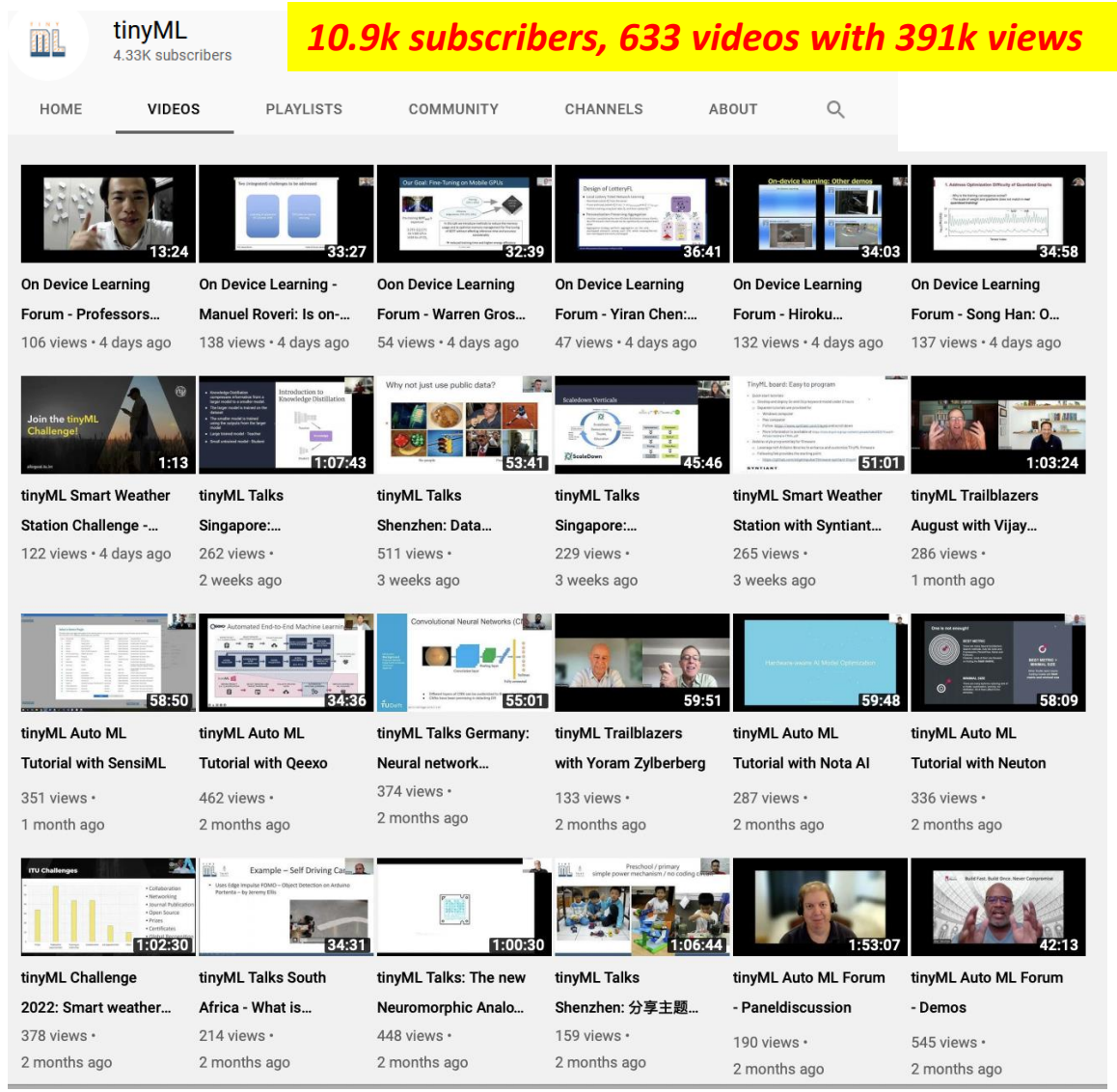
4k members
&
13k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

YouTube



Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

tinyAI Forum on PdM & Anomaly Detection 2023

Interactive live webinar December 5, 2023 at 8AM Pacific Time
Registration is free of charge

tinyML Research Symposium
April 22, 2023
Call for Papers

tinyML Summit April 23-24, 2024
Call for Presentations and Posters

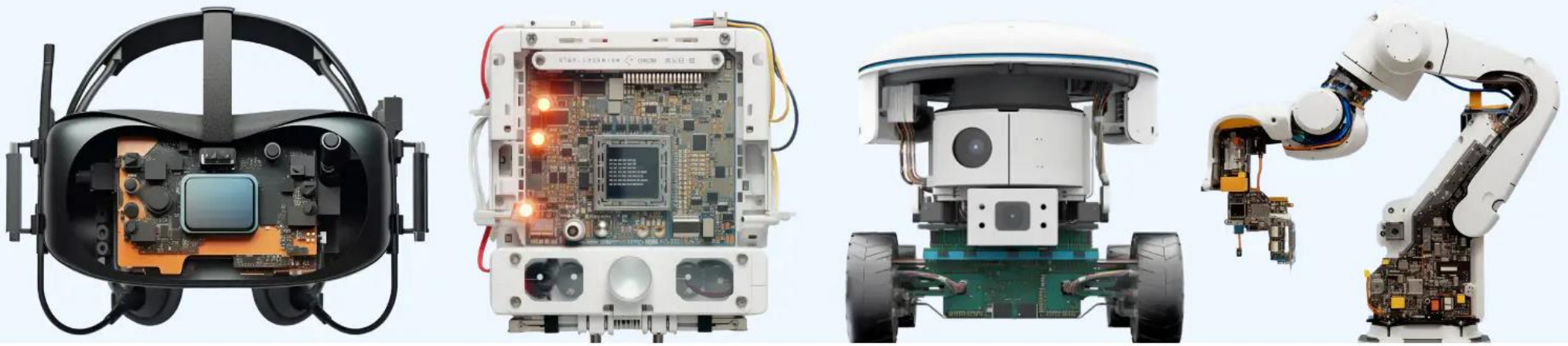
WEVOLVER

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

TINY ML FOUNDATION

2023 Edge AI Technology Report

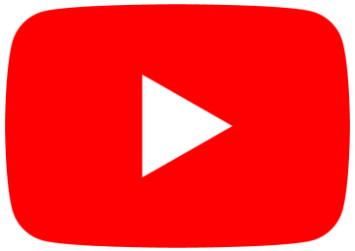
The guide to understanding the state of the art in hardware & software in Edge AI.



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

Cristian Cioflan

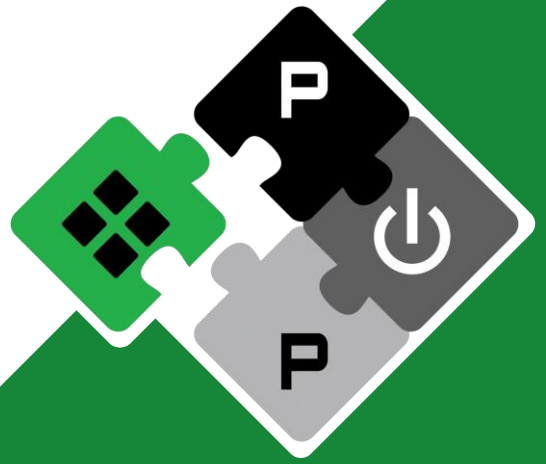
Cristian Cioflan received his B.Sc. degree in Electrical Engineering and Information Technology from University Politehnica of Bucharest in 2018 and his M.Sc. degree from the Swiss Federal Institute of Technology Zürich (ETHZ) in 2020. He is currently pursuing a Ph.D. in the Digital Circuits and Systems group of Prof. Luca Benini. His research interests include audio processing in low-power embedded systems, on-device continual learning, and neural architecture search for energy-efficient learning.

ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

On-Device Domain Adaptation for Noise-Robust Keyword Spotting

Integrated Systems Laboratory (ETH Zürich)



PULP Platform

Open Source Hardware, the way it should be!

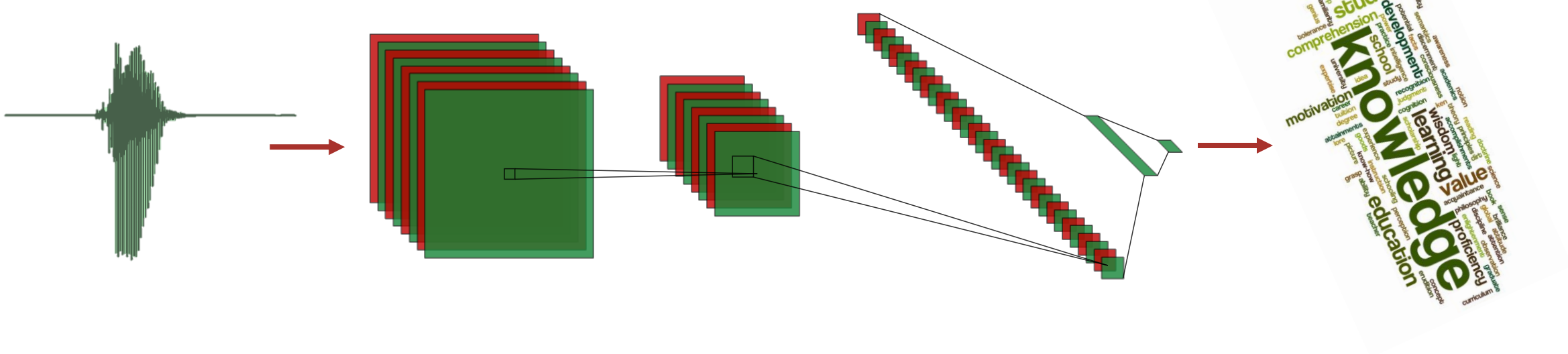
@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

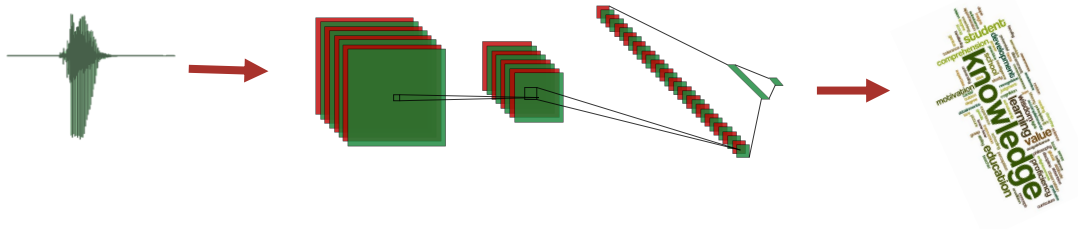
Keyword Spotting – a definition

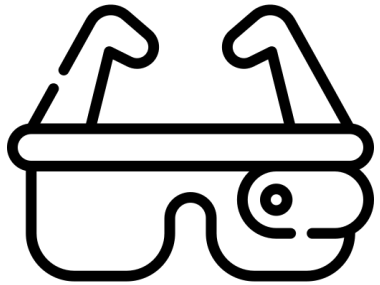
- **Keyword spotting – AI/ML branch**
  - **Process** an audio signal
  - **Recognize** a **target** word from a **predefined set**
  - (to **control** embedded devices)

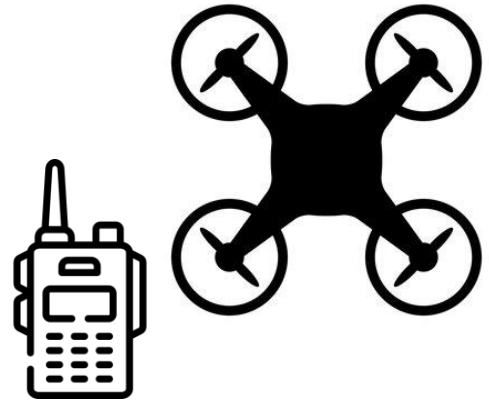


Keyword Spotting – a definition

- **Keyword spotting – AI/ML branch**
  - **Process** an audio signal
  - **Recognize** a **target** word from a **predefined set**
  - (to **control** embedded devices)

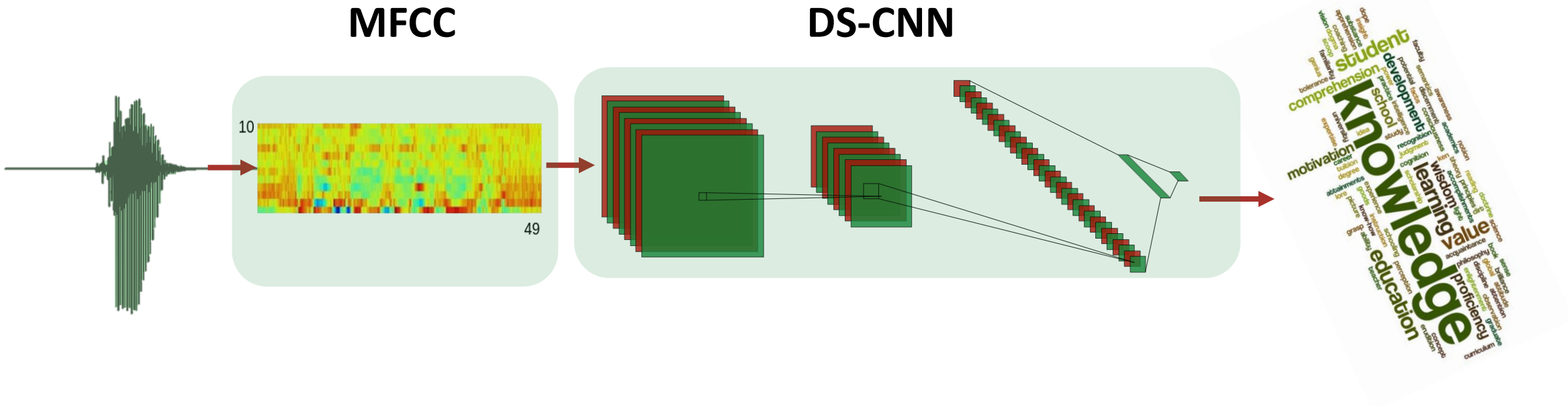






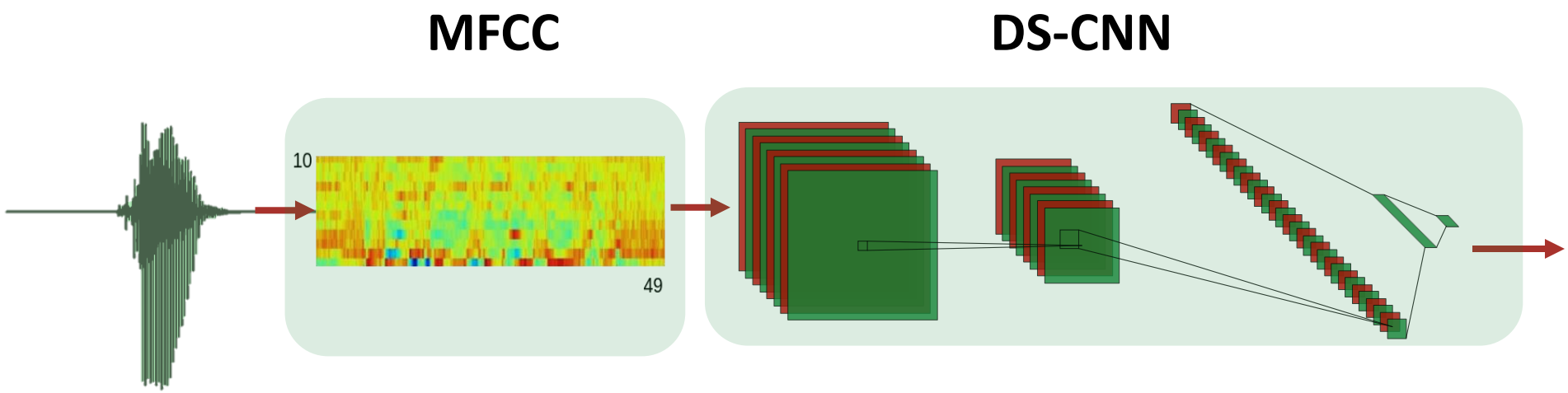
Keyword Spotting – a definition

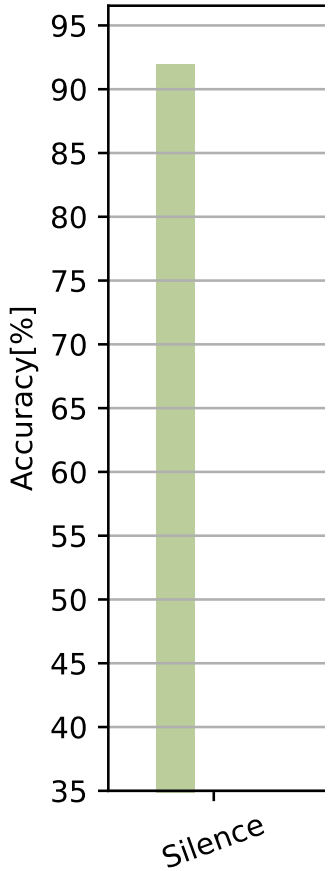
- **Keyword spotting – AI/ML branch**
  - **Process** an audio signal
  - **Recognize** a **target** word from a **predefined set**
  - (to **control** embedded devices)



92% accuracy in clean conditions

- **Keyword spotting – AI/ML branch**
  - **Process** an audio signal
  - **Recognize** a **target** word from a **predefined set**
  - (to **control** embedded devices)





The problem – noisy environments

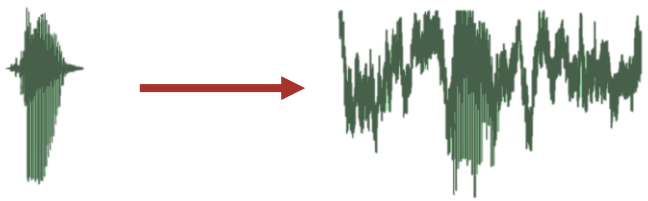
The problem – noisy environments

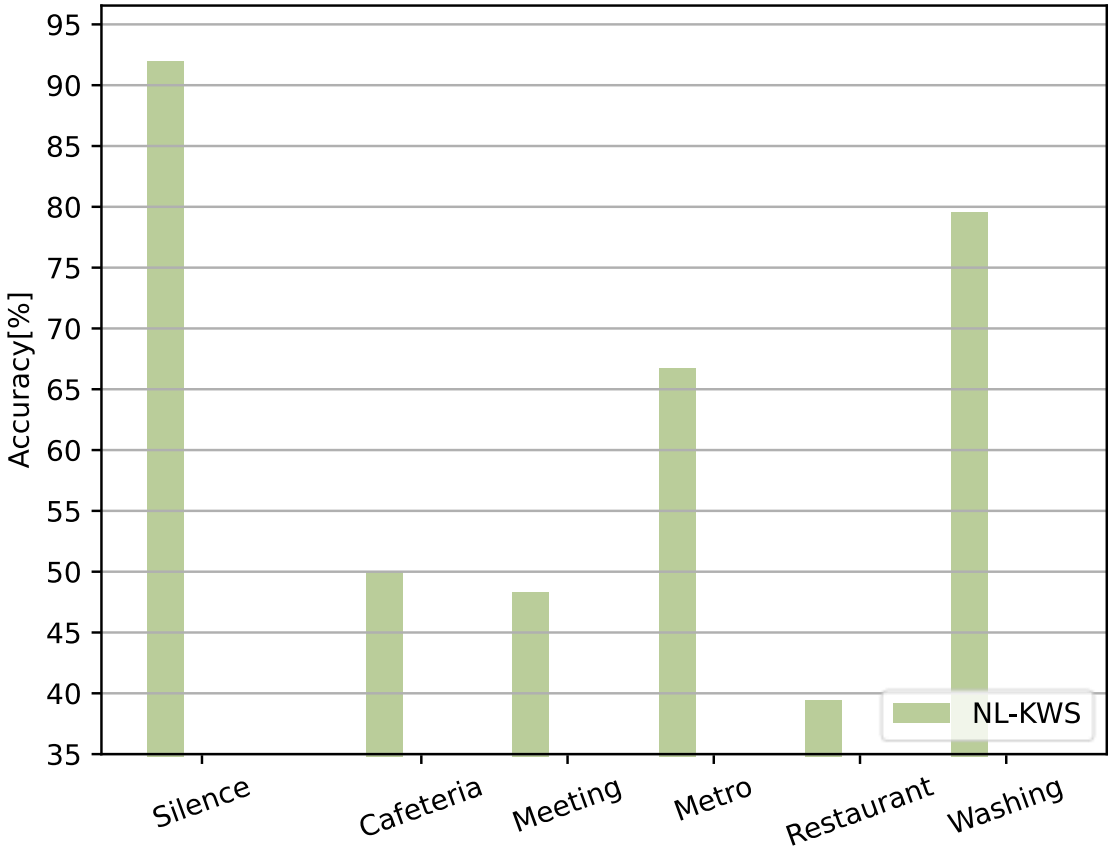
- **Additive background noise pollutes our signal**
- **Negative Signal-to-Noise ratios are common**
  - -10 dB

- **Let us assume a Noiseless Keyword Spotting model**
- **Can a NL-KWS system still recognize the word?**

Can a KWS system still recognize the words?

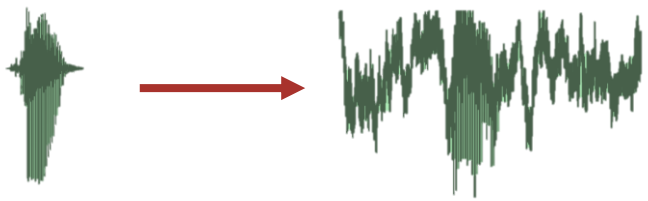
- **Accuracy drops between 12% and 52%**
- **Stationary vs non-stationary noises**
- **Speech noise**
  - @ 0 dB – **43%** accuracy loss
  - Hard to separate target from noise

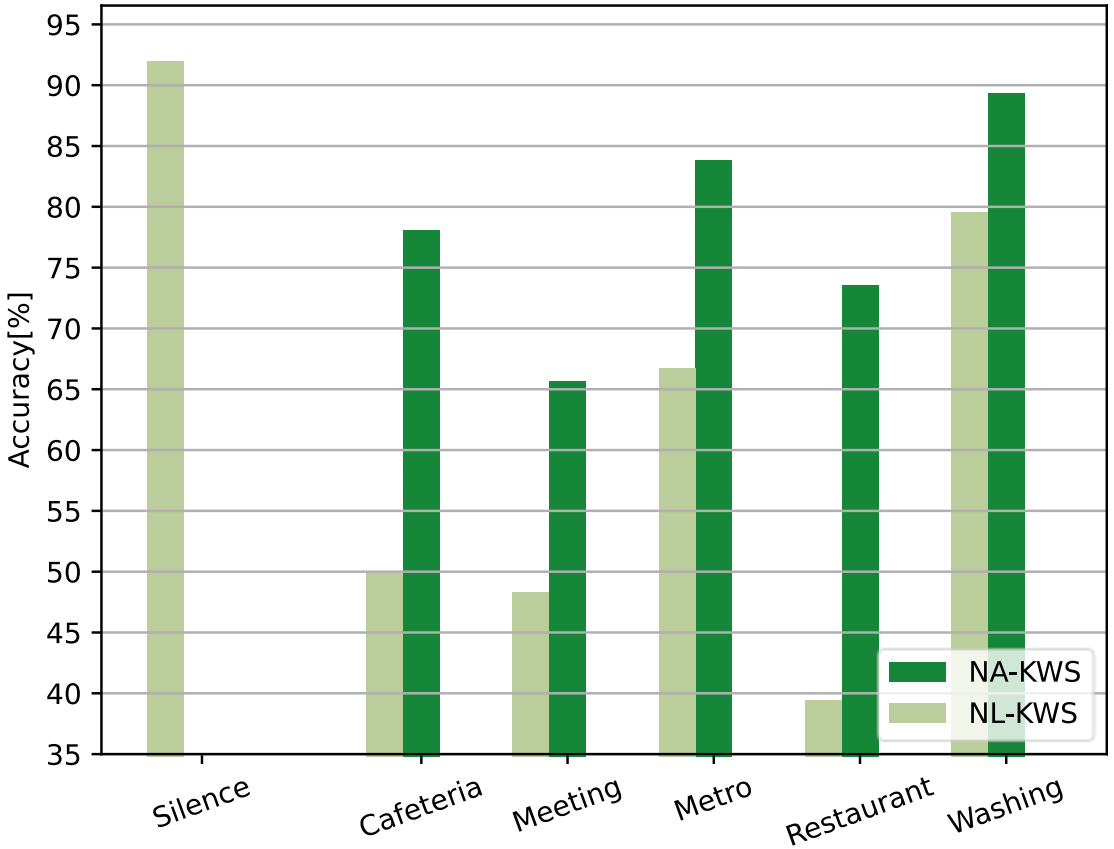




Noise-Aware Keyword Spotting

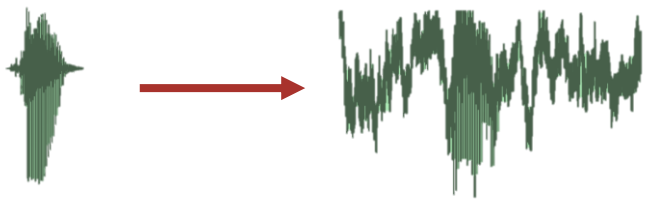
- **NL-KWS vs NA-KWS**
  - Augment training samples with diverse noise types (and SNRs…)
  - Improved results
    - Between **10%** and **34%** over NL-KWS

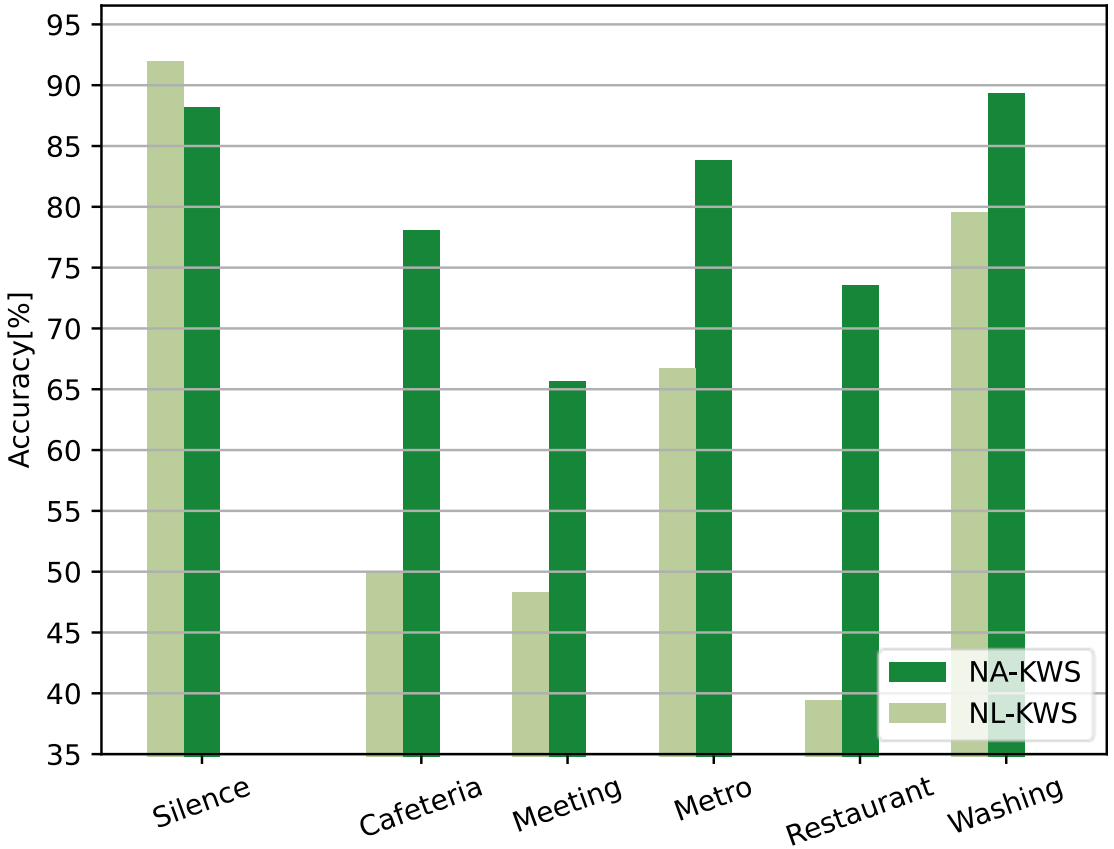




Noise-Aware Keyword Spotting

- **NL-KWS vs NA-KWS**
  - Augment training samples with diverse noise types (and SNRs…)
  - Improved results
    - Between **10%** and **34%** over NL-KWS
    - **-4%** on *Silence*





Can we improve KWS accuracy through direct noise exposure?

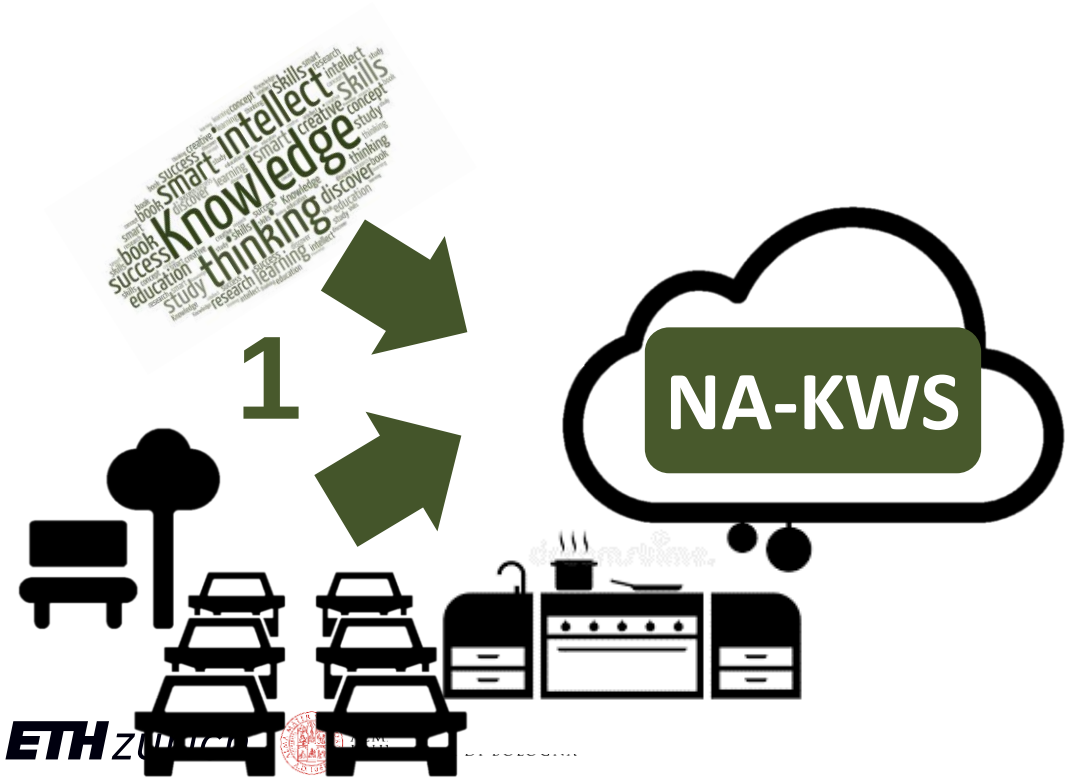
Can we improve KWS accuracy through direct noise exposure?

On-Device Domain Adaptation

On-Device Domain Adaptation – the methodology

1. Noise-Aware Keyword Spotting

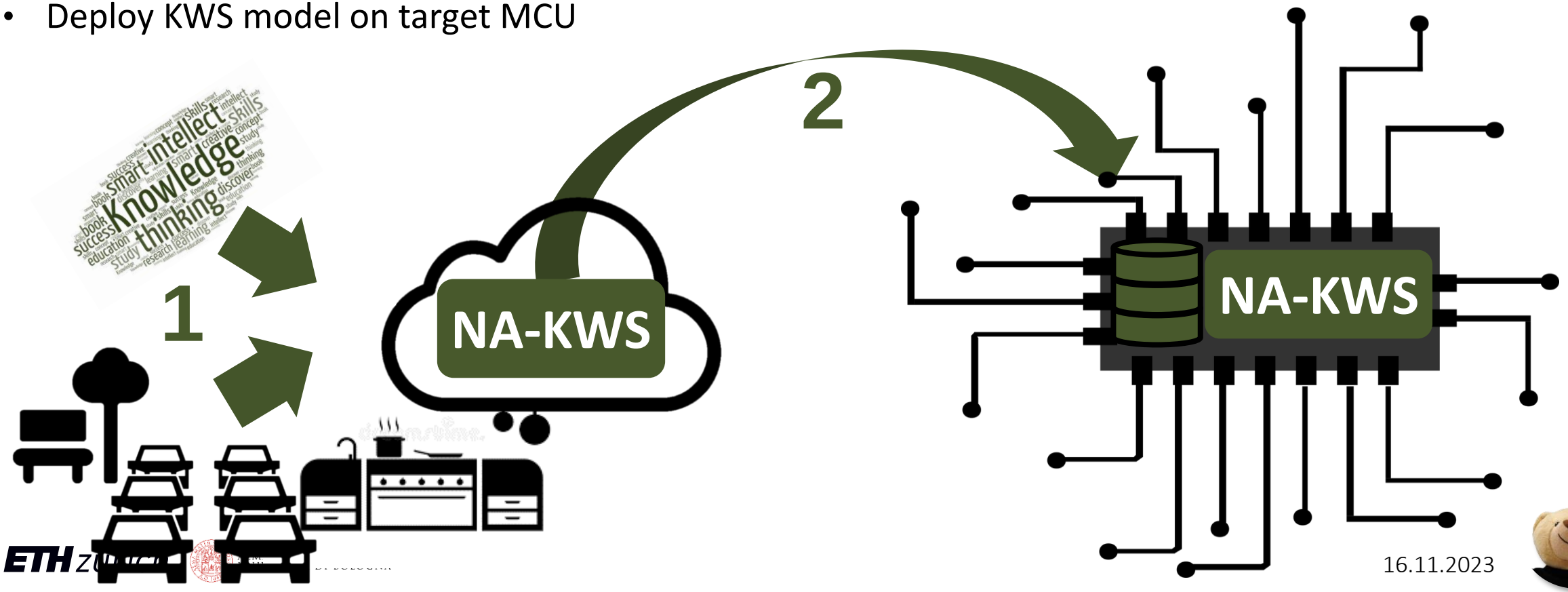
- Data augmentation for improved generalization
- Google Speech Commands v2 x DEMAND



On-Device Domain Adaptation – the methodology

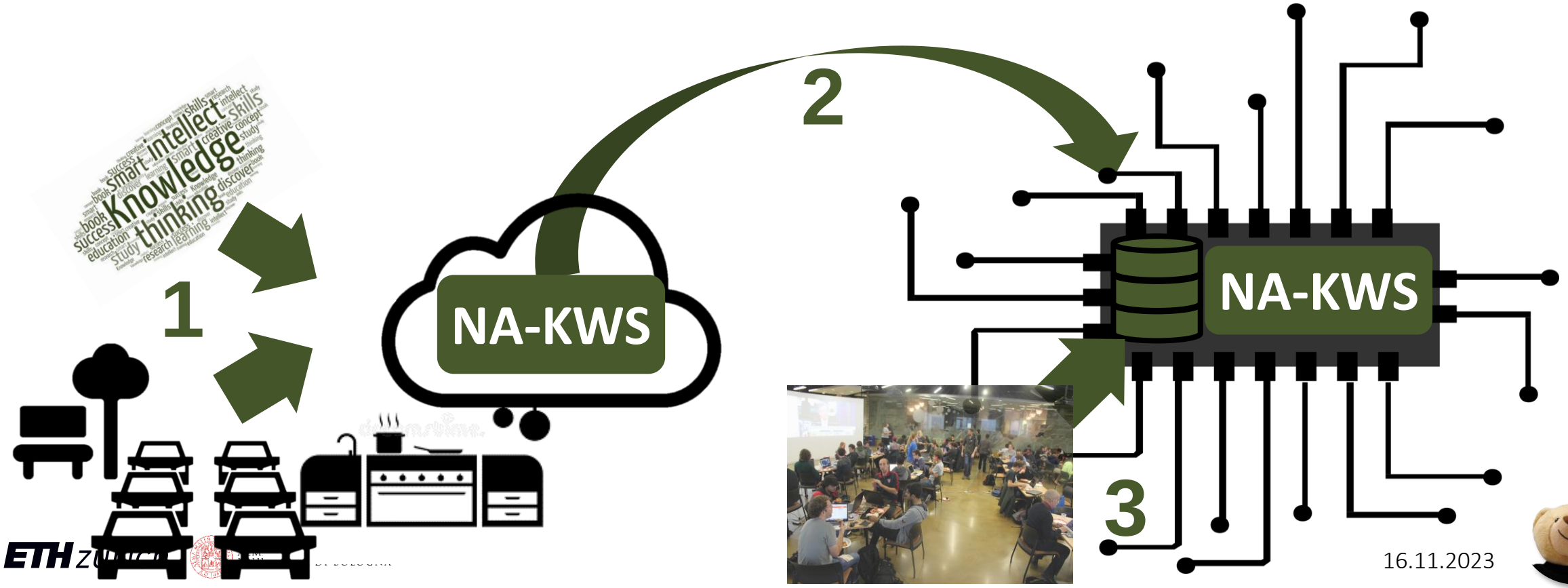
1. **Noise-Aware Keyword Spotting**
2. **Deployment**

- Store pre-recorded utterances
- Deploy KWS model on target MCU



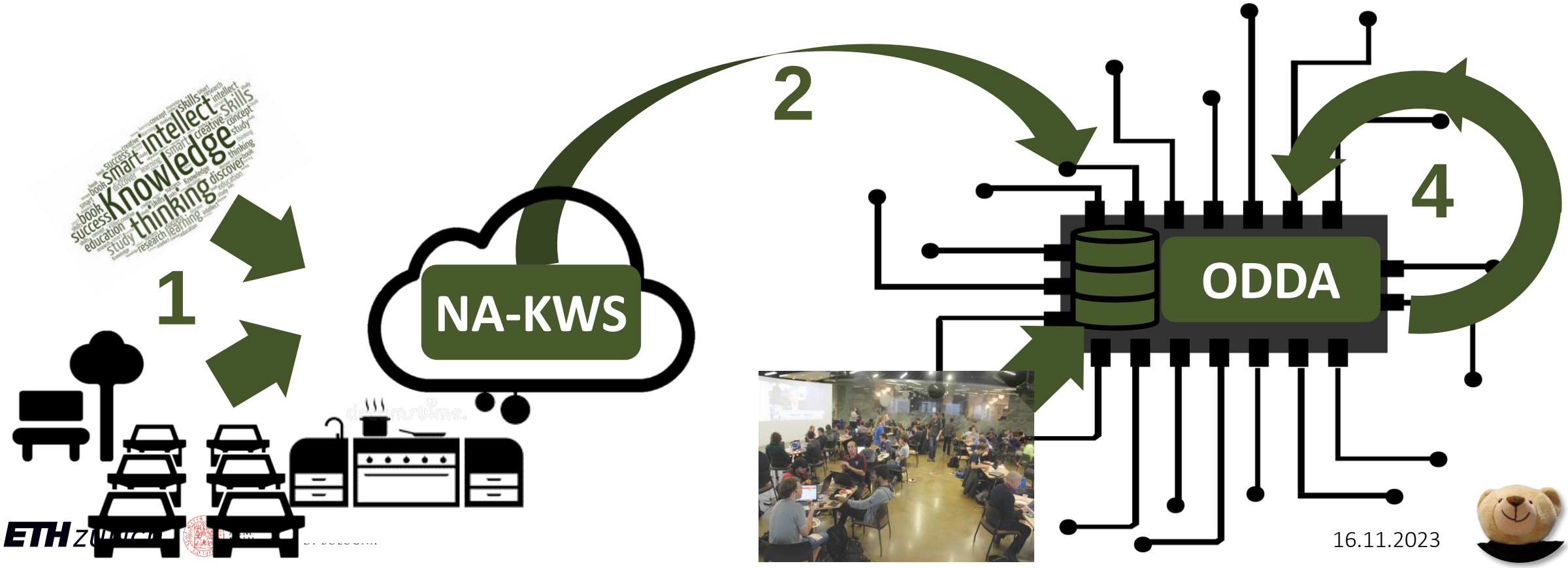
On-Device Domain Adaptation – the methodology

1. **Noise-Aware Keyword Spotting**
2. **Deployment**
3. **Noise exposure**
   - Record on-site noise
   - Augment pre-recorded samples



On-Device Domain Adaptation – the methodology

1. **Noise-Aware Keyword Spotting**
2. **Deployment**
3. **Noise exposure**
4. **ODDA** [1]
   - Refine NA-KWS using augmented utterances

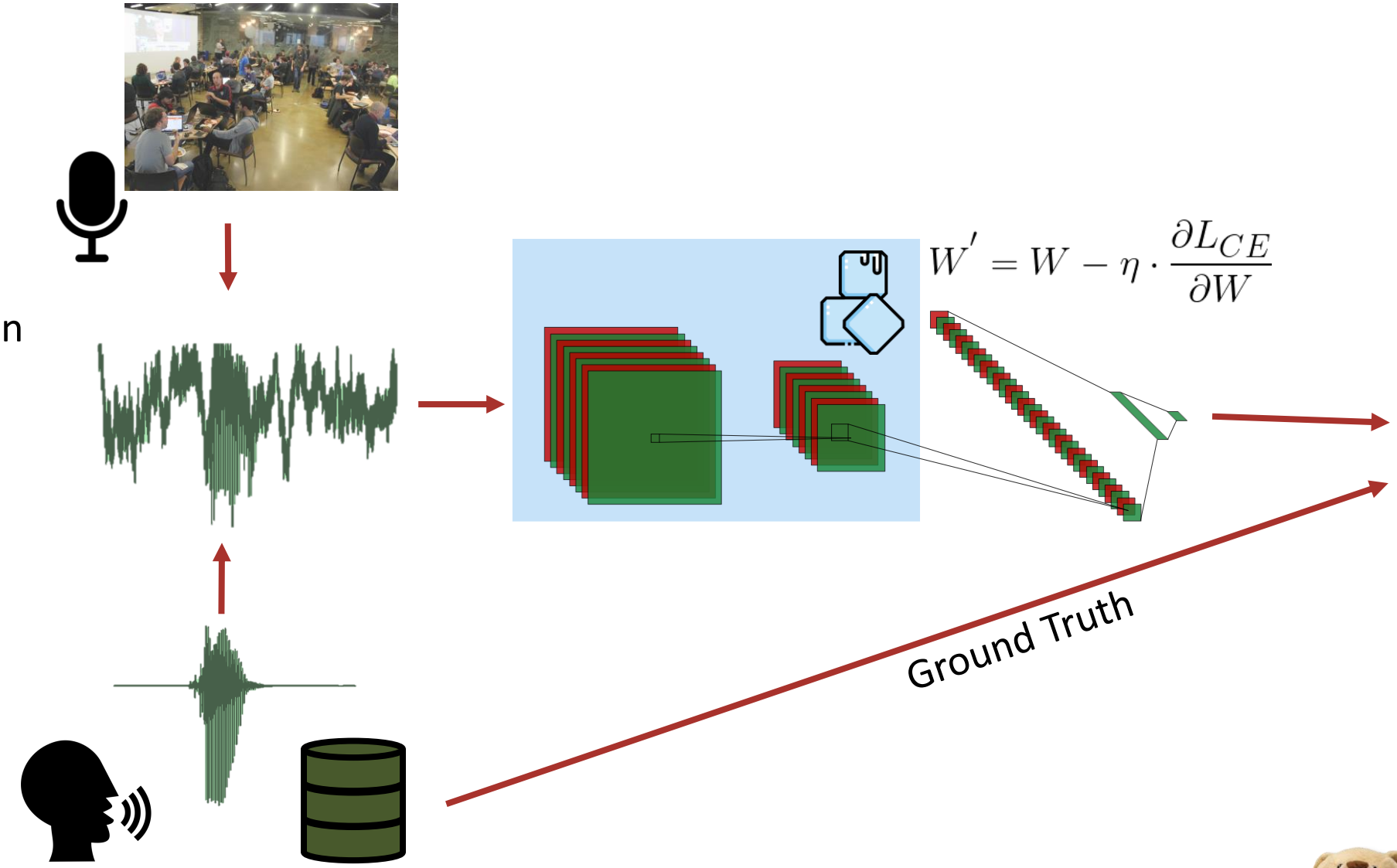


4. ODDA – Refine NA-KWS using augmented utterances

a. Combine additive background noise with pre-recorded utterances

b. Compute CE loss using partially frozen model prediction

c. Selectively update model parameters



Exposure to environmental noise for specialization

On-Device Domain Adaptation

1. **Pretrain** a Noise-Aware Keyword Spotting network
2. **Deploy** NA-KWS (and clean utter.) on embedded device
3. **Augment** utterances with recorded environmental noise
4. **Refine** the network's parameters on the target platform
5. **Recognize** utterances with domain-adapted KWS system

Modelling ODDA

Google Speech Commands v2 (**GSCv2**)[2]

- 1-second audio @ 16 kHz sample rate
- silence & unknown – 10%
- NA-KWS/ODDA:validation:test – 80:10:10

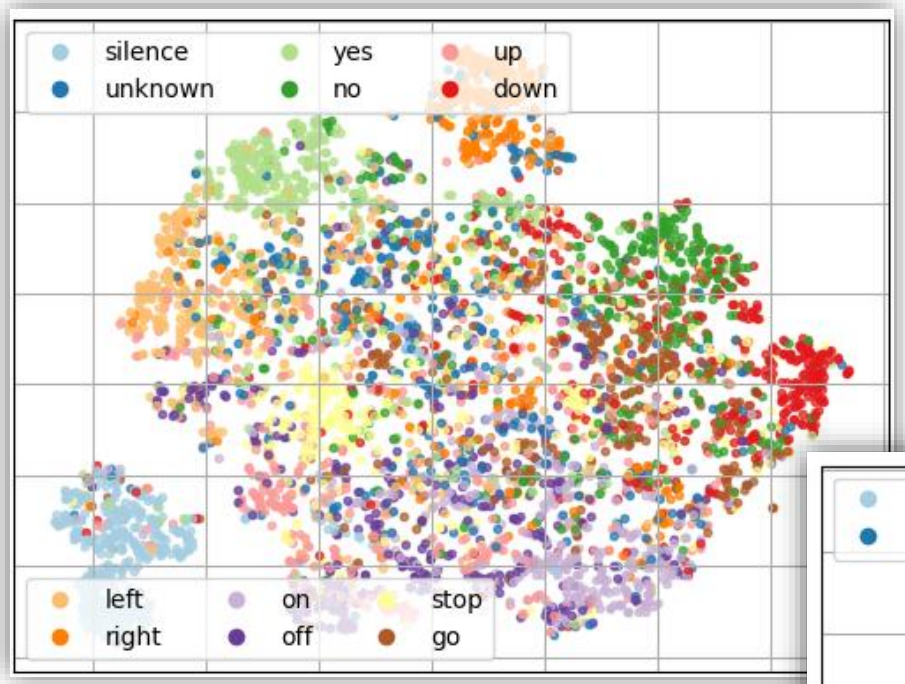
Diverse Environments Multichannel Acoustic Noise Database (**DEMAND**)[3]

- 18 noises in real-world conditions
- 5 LOO settings
  - cafeteria, meeting, metro, restaurant, washing
- SNR: {-10, 10} dB

How well does ODDA perform?

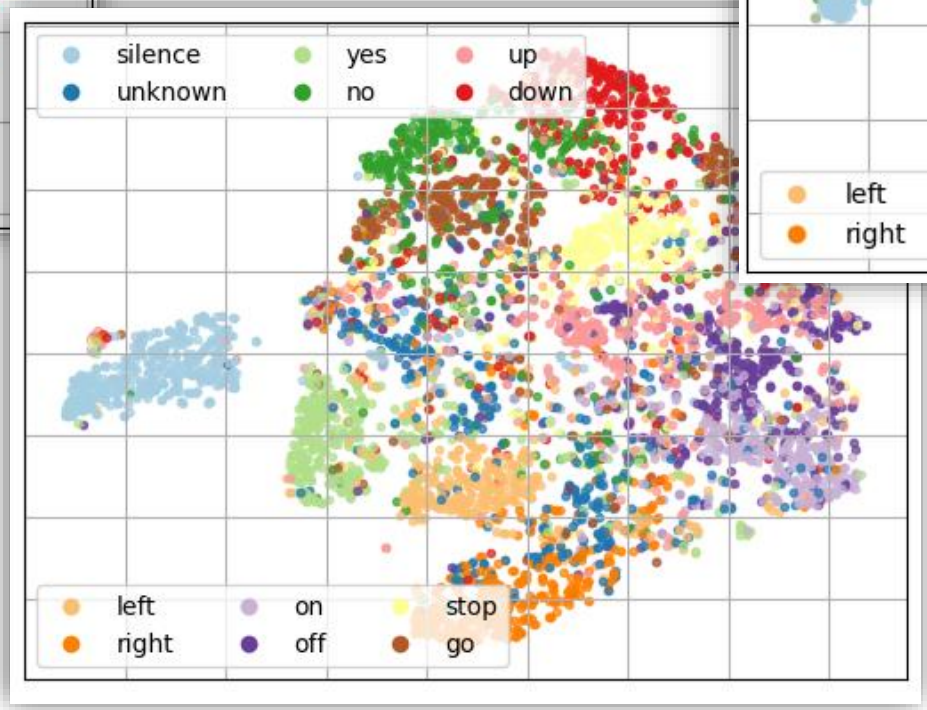
ODDA – Qualitative analysis

NL-KWS

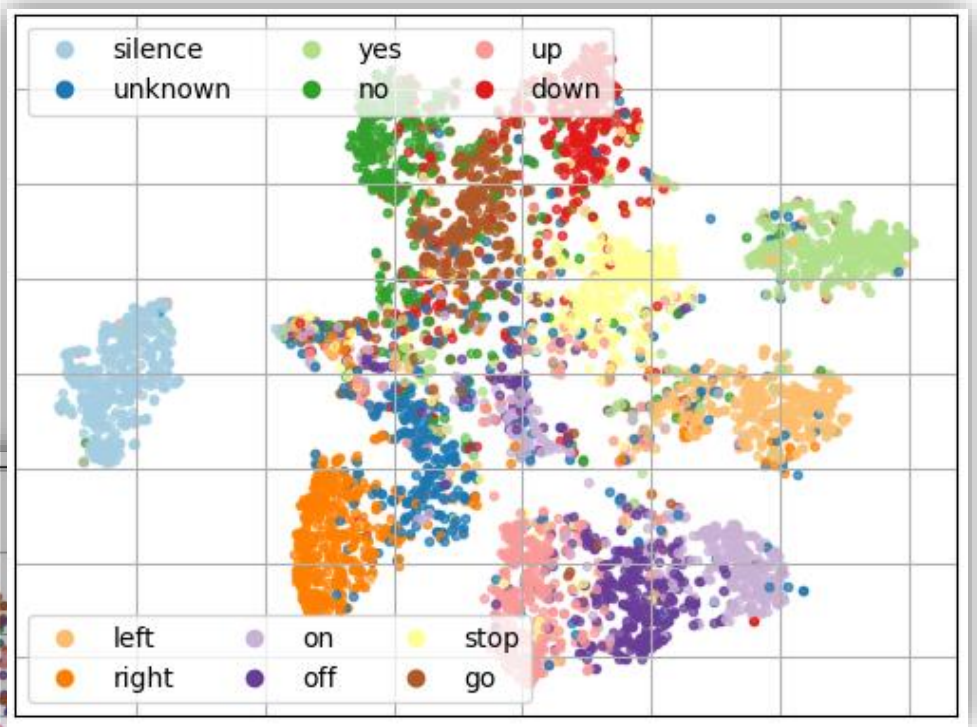


- t-SNE visualization of pre-classifier features
- *meeting* noise

NA-KWS

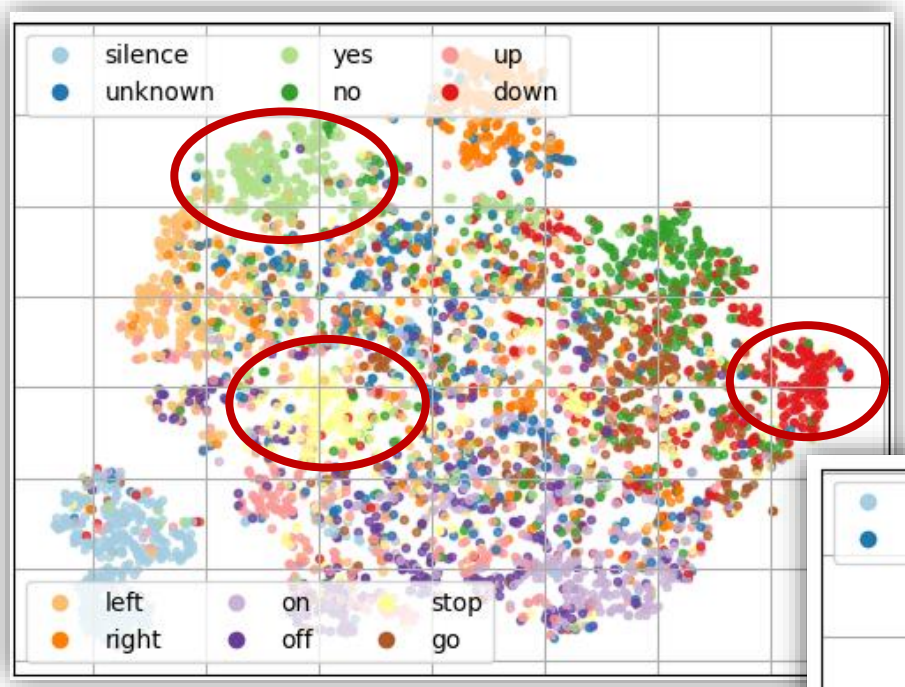


ODDA



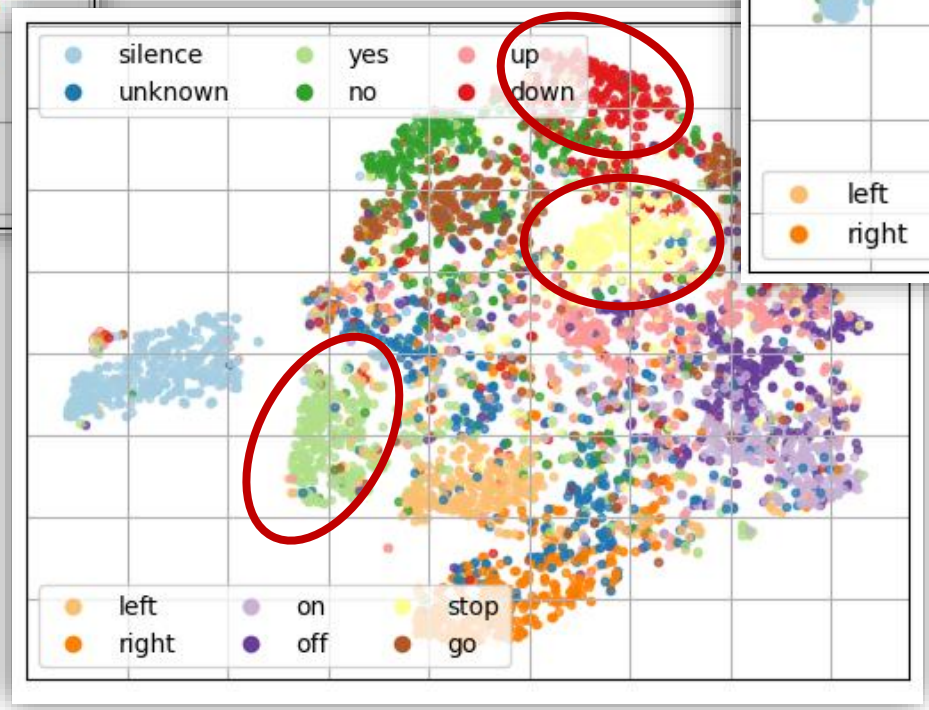
Exposure to env. noise ⟶ feature separation

NL-KWS

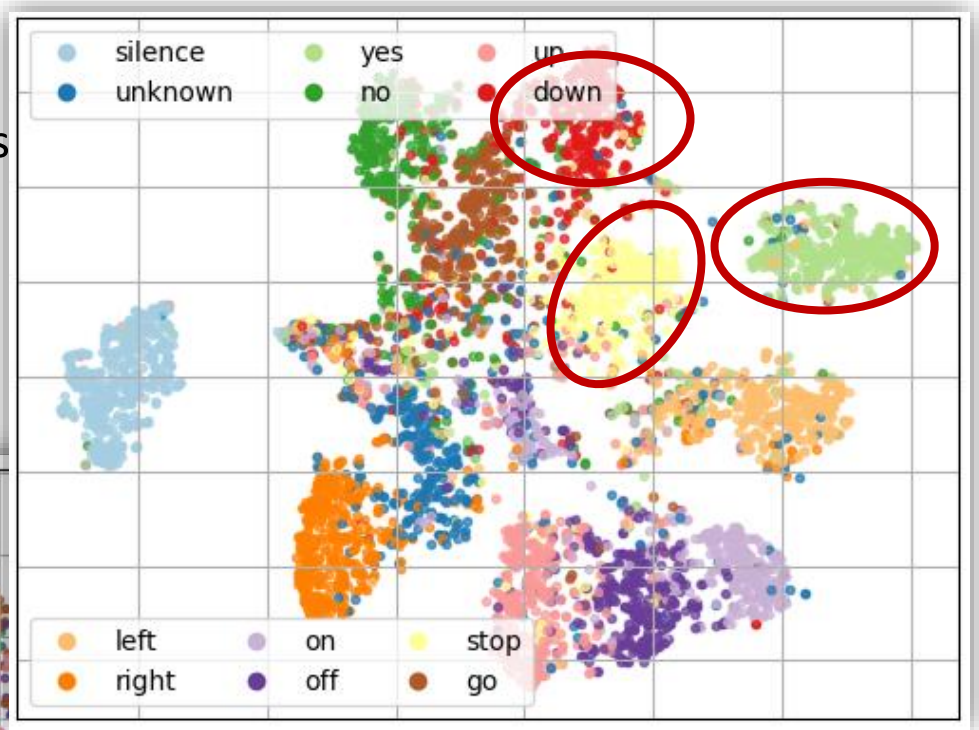


- *Yes, stop, down*
  - **Marginal** improvements

NA-KWS

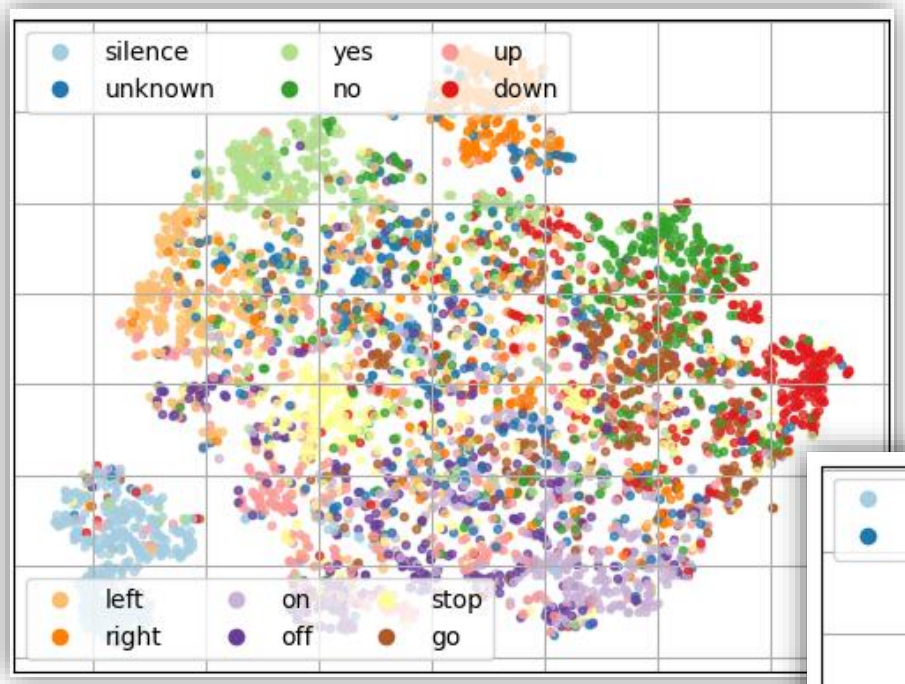


ODDA



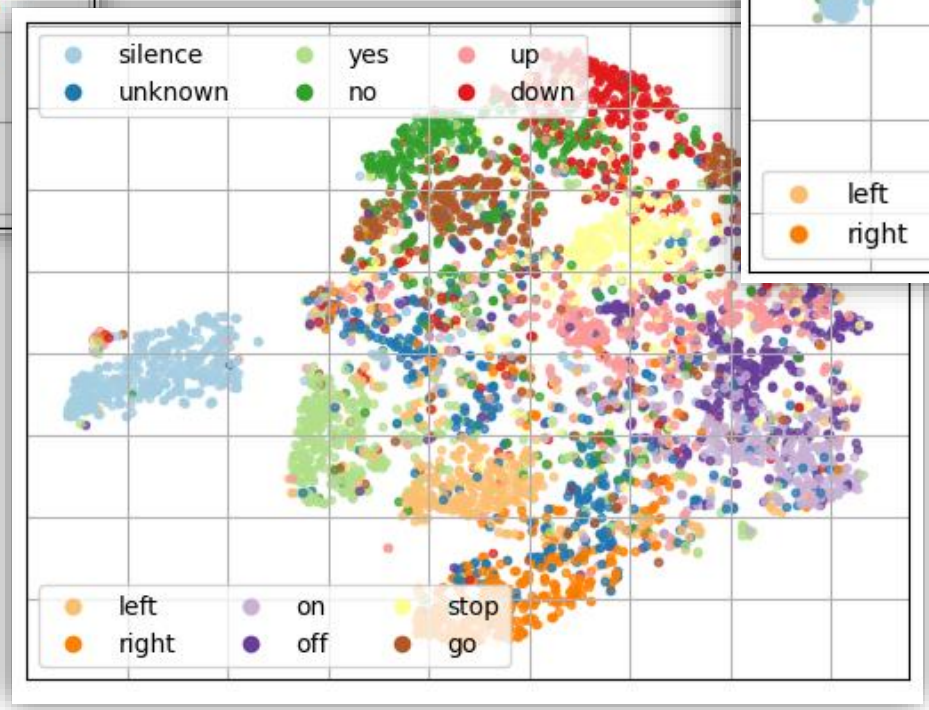
Exposure to env. noise ⟶ feature separation

NL-KWS

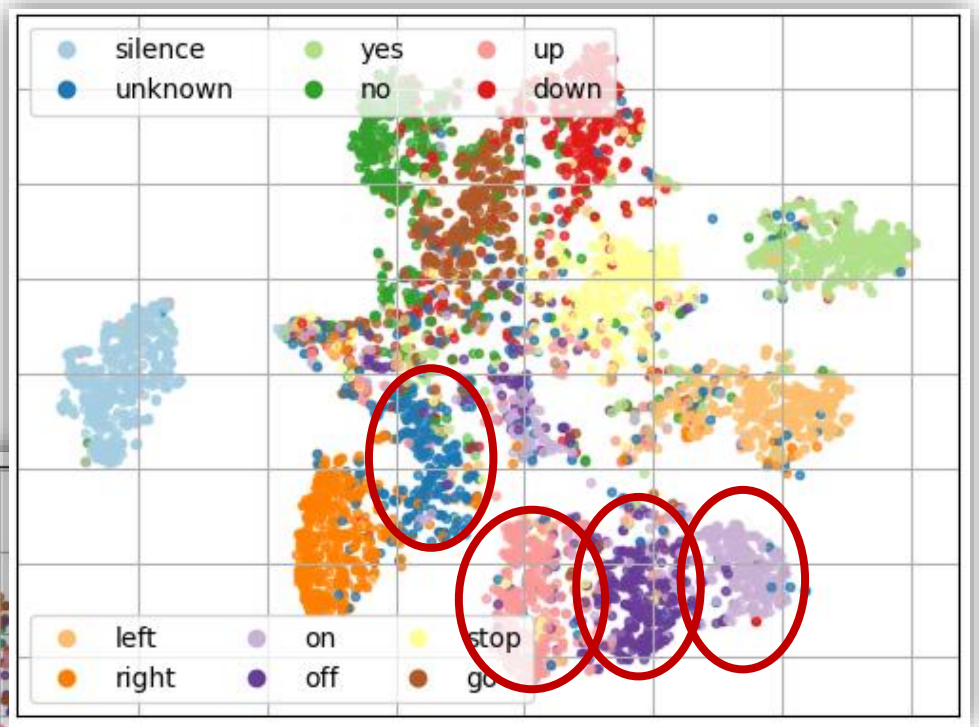


- *Unknown, up, off, on*
  - **Significant** improvements

NA-KWS

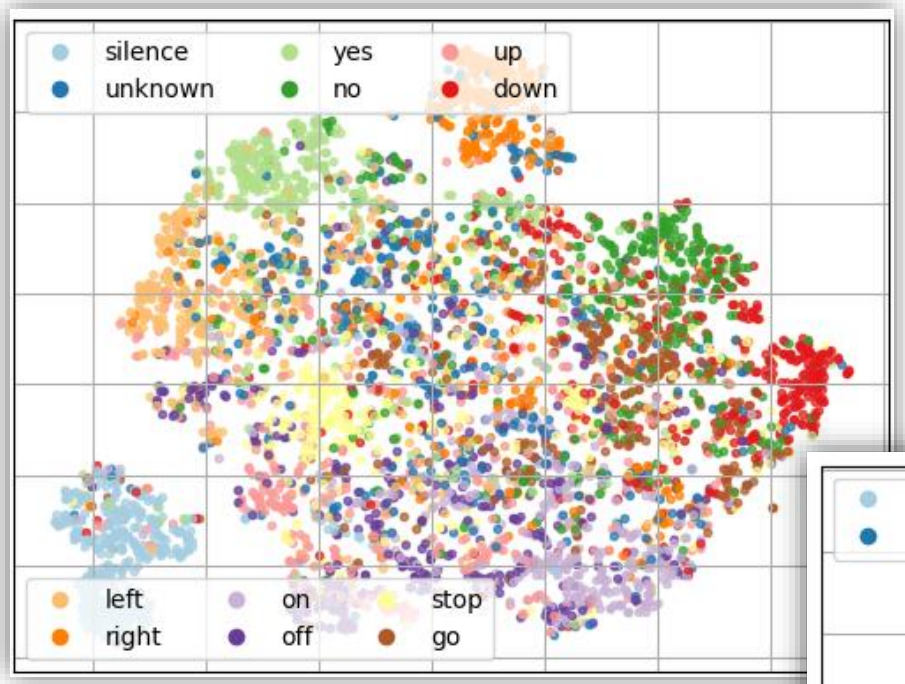


ODDA



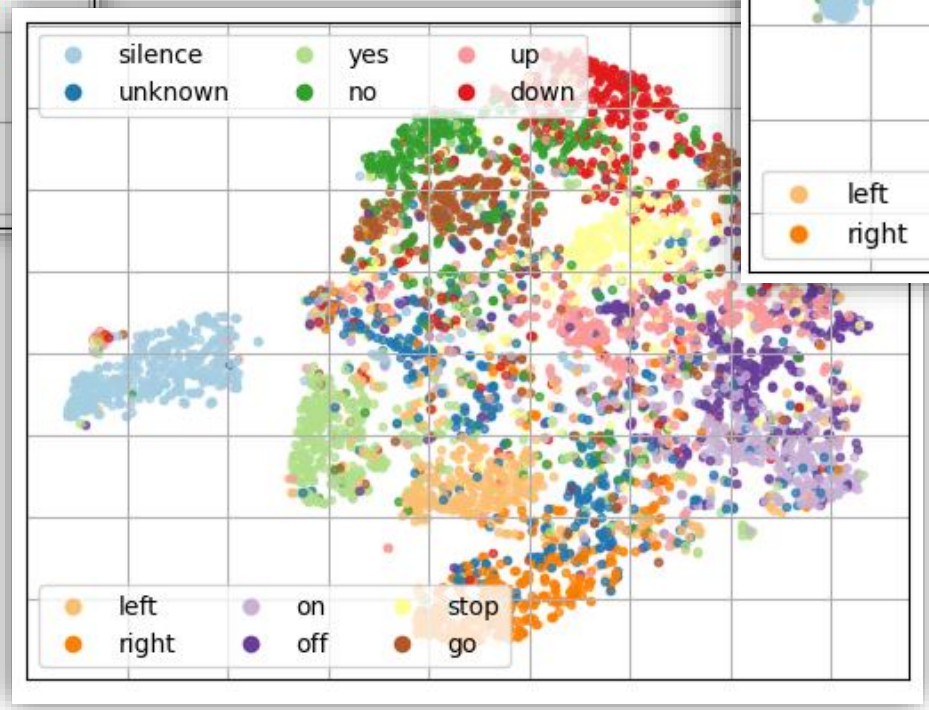
Exposure to env. noise ⟶ feature separation

NL-KWS

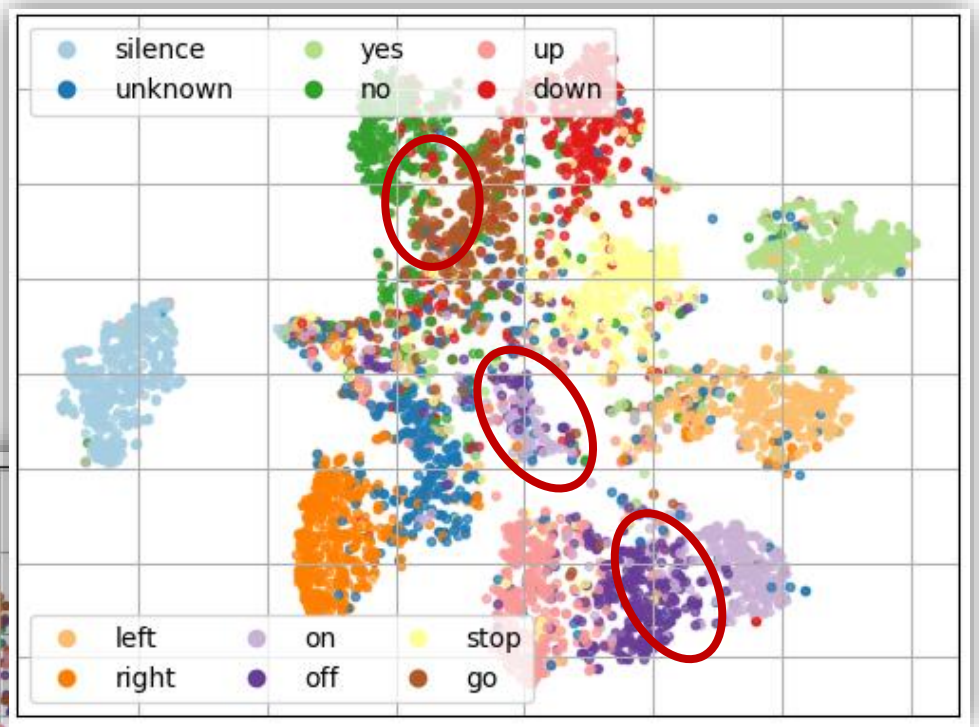


- *On, off*
- *No, go*
  - Short, phonetically similar

NA-KWS

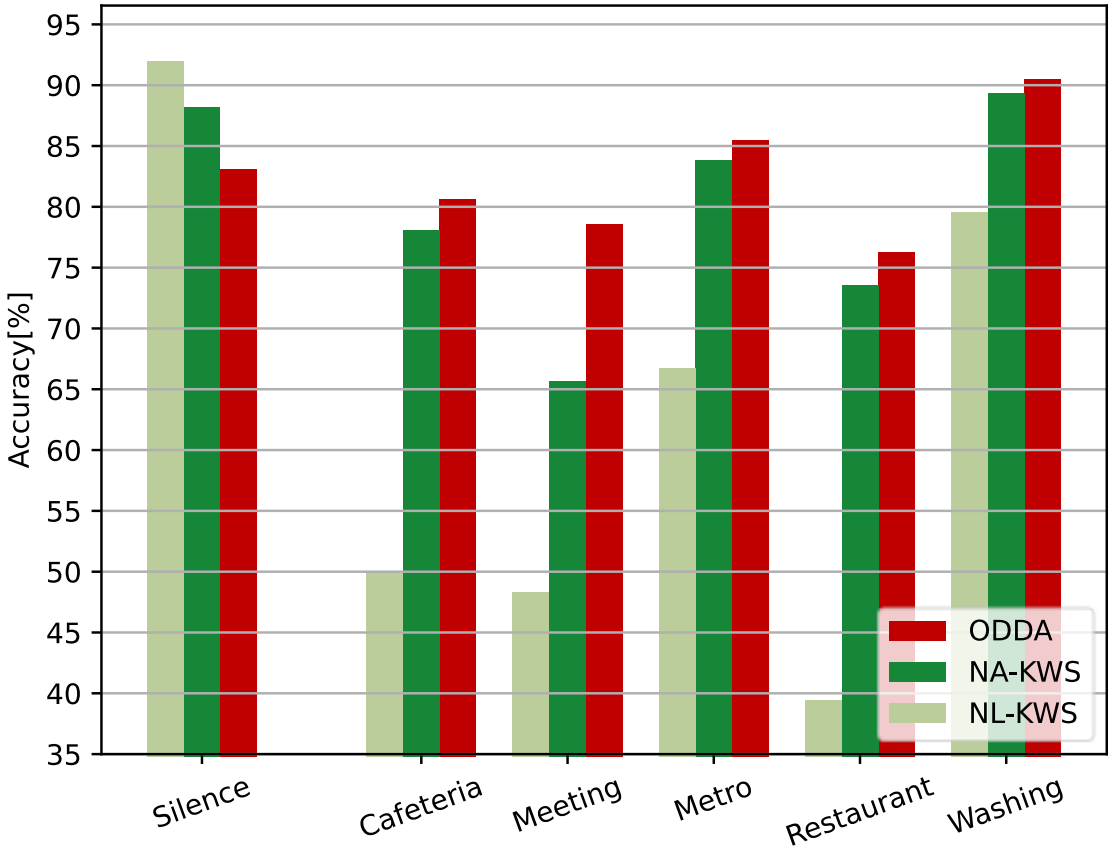


ODDA



ODDA offers accuracy improvements on any noise…

- **For 0 dB SNRs**
  - **12% improvement** on *meeting* over NA-KWS
    - **+30%** over NL-KWS
  - **37% improvement** on *restaurant* over NL-KWS
    - **+3%** over NA-KWS



… and any SNR

- **For 0 dB SNRs**
  - **12% improvement** on *meeting* over NA-KWS
    - **+30%** over NL-KWS
  - **37% improvement** on *restaurant* over NL-KWS
    - **+3%** over NA-KWS
- ODDA's effect **scales with SNR**
  - 5-6% to <1%

		SNR_{test}			
		-10	0	30	AVG
SNR_{ODDA}	**[-10,10]**	59.68	80.96	85.86	79.23
	0	54.34	81.87	87.71	79.29
	[0,10]	51.52	81.06	88.93	76.49
	NA-KWS	49.10	75.85	87.61	74.49

… and any SNR

- **For 0 dB SNRs**
  - **12% improvement** on *meeting* over NA-KWS
    - **+30%** over NL-KWS
  - **37% improvement** on *restaurant* over NL-KWS
    - **+3%** over NA-KWS
- ODDA's effect **scales with SNR**
  - 5-6% to <1%
  - Up to **10%** for known env. SNR

		SNR_{test}			
		-10	0	30	AVG
SNR_{ODDA}	**[-10,10]**	59.68	80.96	85.86	79.23
	0	54.34	81.87	87.71	79.29
	[0,10]	51.52	81.06	88.93	76.49
	NA-KWS	49.10	75.85	87.61	74.49

PULP Platform

Open Source Hardware, the way it should be!

Q&A

Institut für Integrierte Systeme – ETH Zürich
Gloriastrasse 35
Zürich, Switzerland

DEI – Universitá di Bologna
Viale del Risorgimento 2
Bologna, Italy

@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

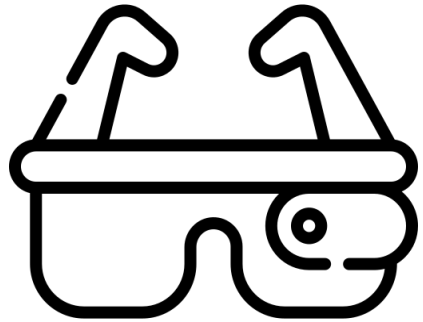
ETH zürich

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

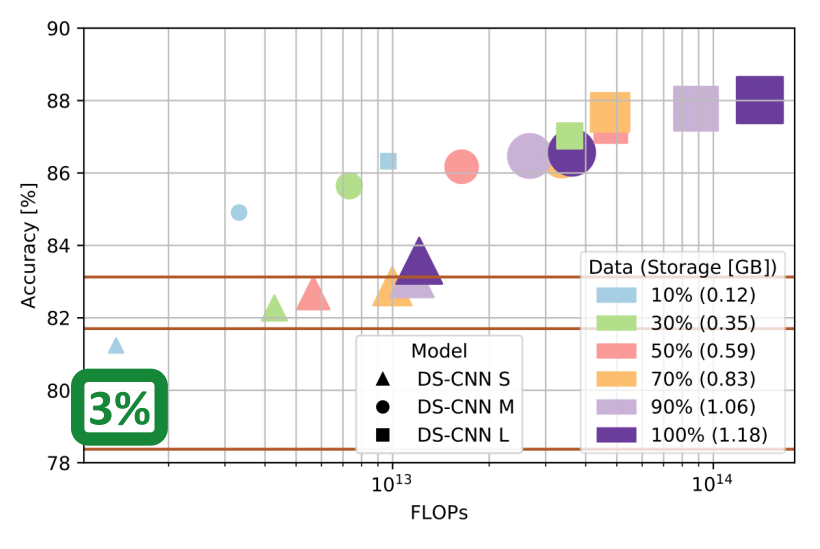
We are improving accuracy,
but at what cost?

TinyML constraints

- **(Extreme) Edge AI/ML**
- **Embedded, miniaturized devices**
  - Limited storage (e.g., model parameters)
  - Limited memory (e.g., activations, gradients)
- **Real-time operation**
  - Limited computational resources
  - Maximize throughput, minimize latency
- **Always-on, battery operated devices**
  - (Ultra-) Low-Power systems



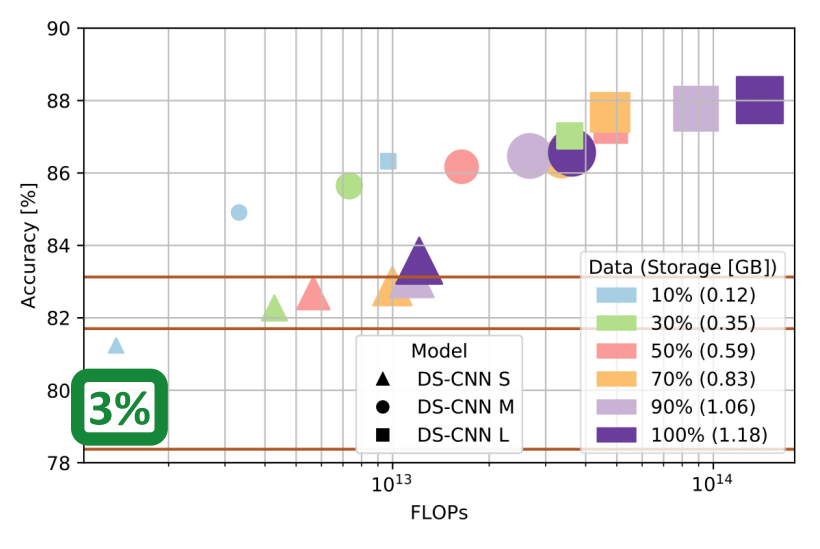
Resource-constrained ODDA



Storage

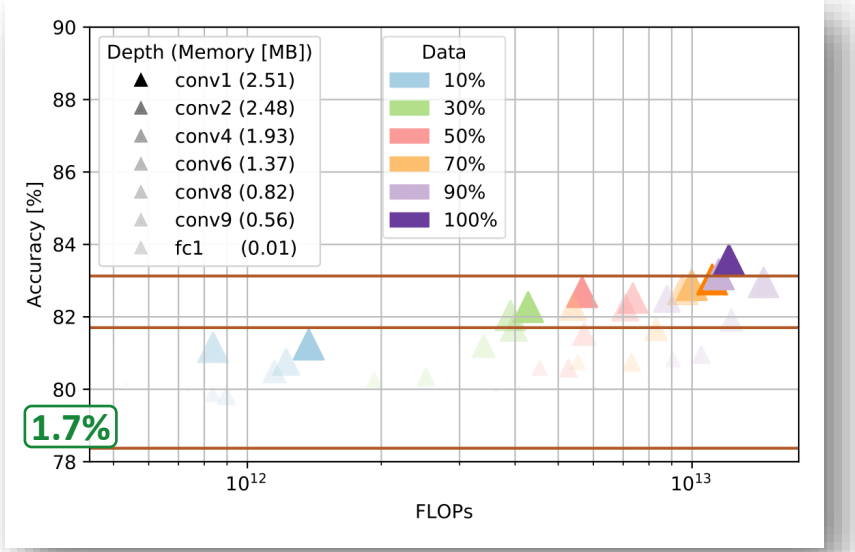
- 120 MB – 310 samples/class
  - +3% over NA-KWS
  - +5% on *meeting* with 3 MB
- Increasing model size
  - +3.7% for M model
  - +5% for L model

Resource-constrained ODDA



Storage

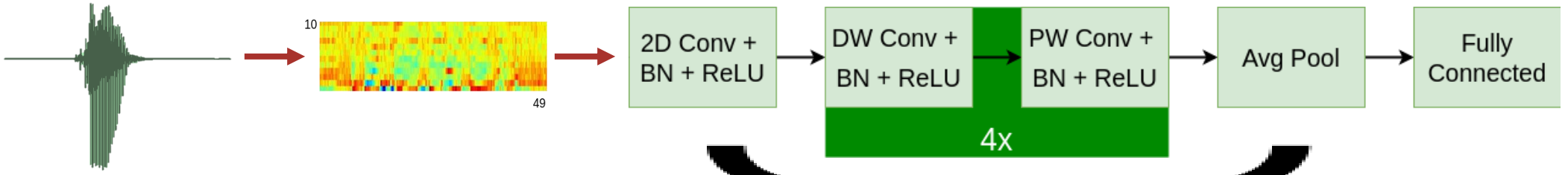
- 120 MB – 310 samples/class
  - +3% over NA-KWS
  - +5% on *meeting* with 3 MB
- Increasing model size
  - +3.7% for M model
  - +5% for L model



Memory

- +1.7% by refining the only classifier
  - 10 kB (i.e., parameters, activations)

Model deployment for On-Device Domain Adaptation



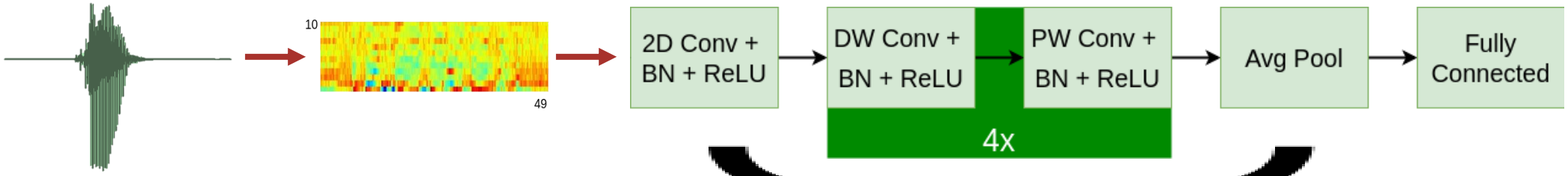
`int16`
1-second @16 kHz

`float16`
window length = 640 ms
window hop = 320 ms
10 MFC coefficients

`int8`
NEMO[4] quantization
DORY[5] deployment

`float32`
PULP-TrainLib[6] deployment

Model deployment for On-Device Domain Adaptation

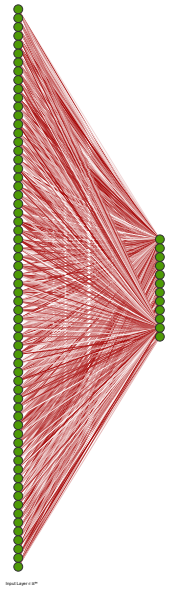


`int16`
1-second @16 kHz

`float16`
window length = 640 ms
window hop = 320 ms
10 MFC coefficients

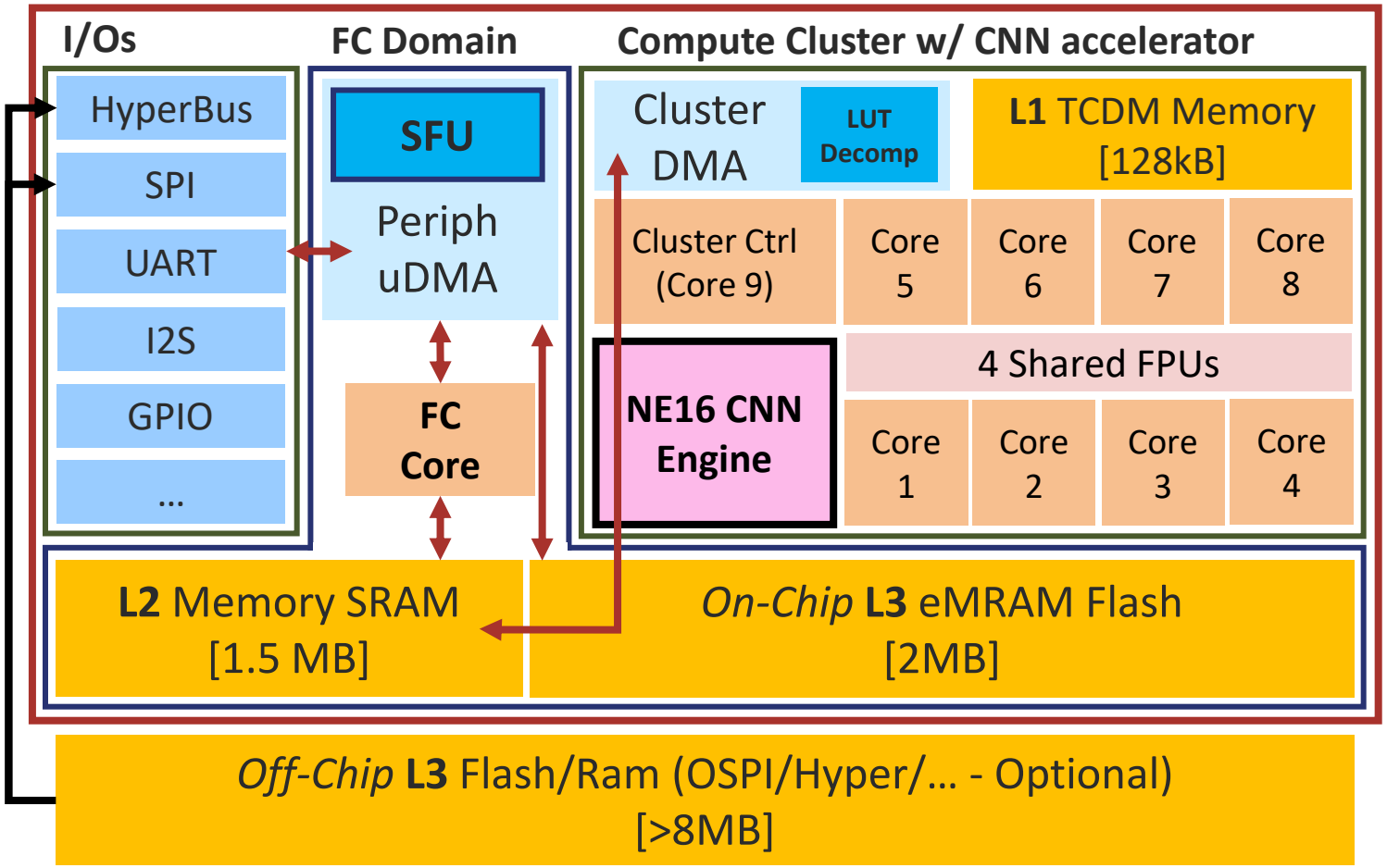
`int8`
NEMO[4] quantization
DORY[5] deployment

`float32`
PULP-TrainLib[6] deployment



- Fully Connected structure
  - 64 input neurons
  - 12 output neurons
  - **768 weights**
- PULP-Trainlib
  - Softmax
  - Cross-Entropy loss

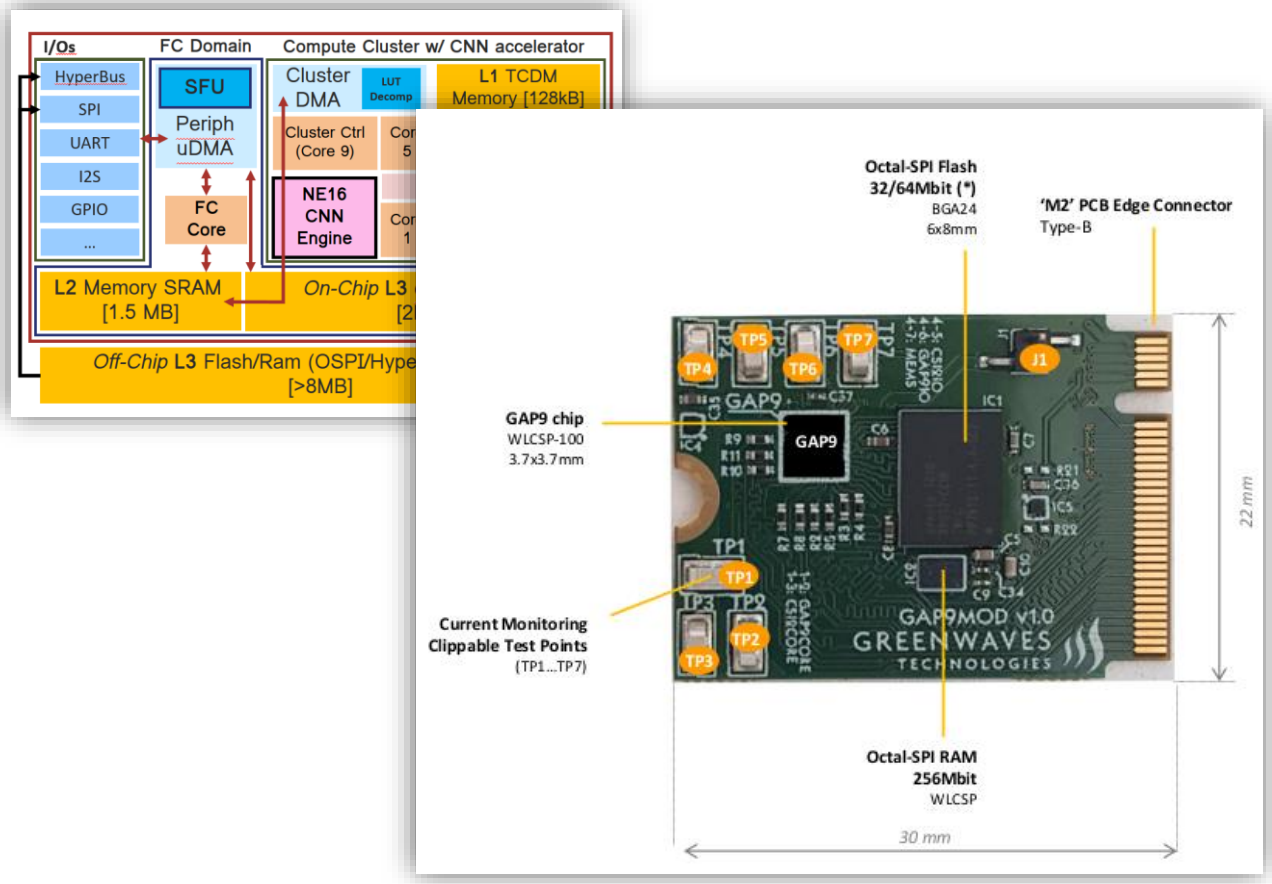
On-Device Domain Adaptation target: GAP9



GAP9 Architecture

- 4 frequency domains
- Hierarchical memory architecture
- Heterogeneous compute units
  - 10 general purpose RISC-V cores
  - 4 shared FPUs

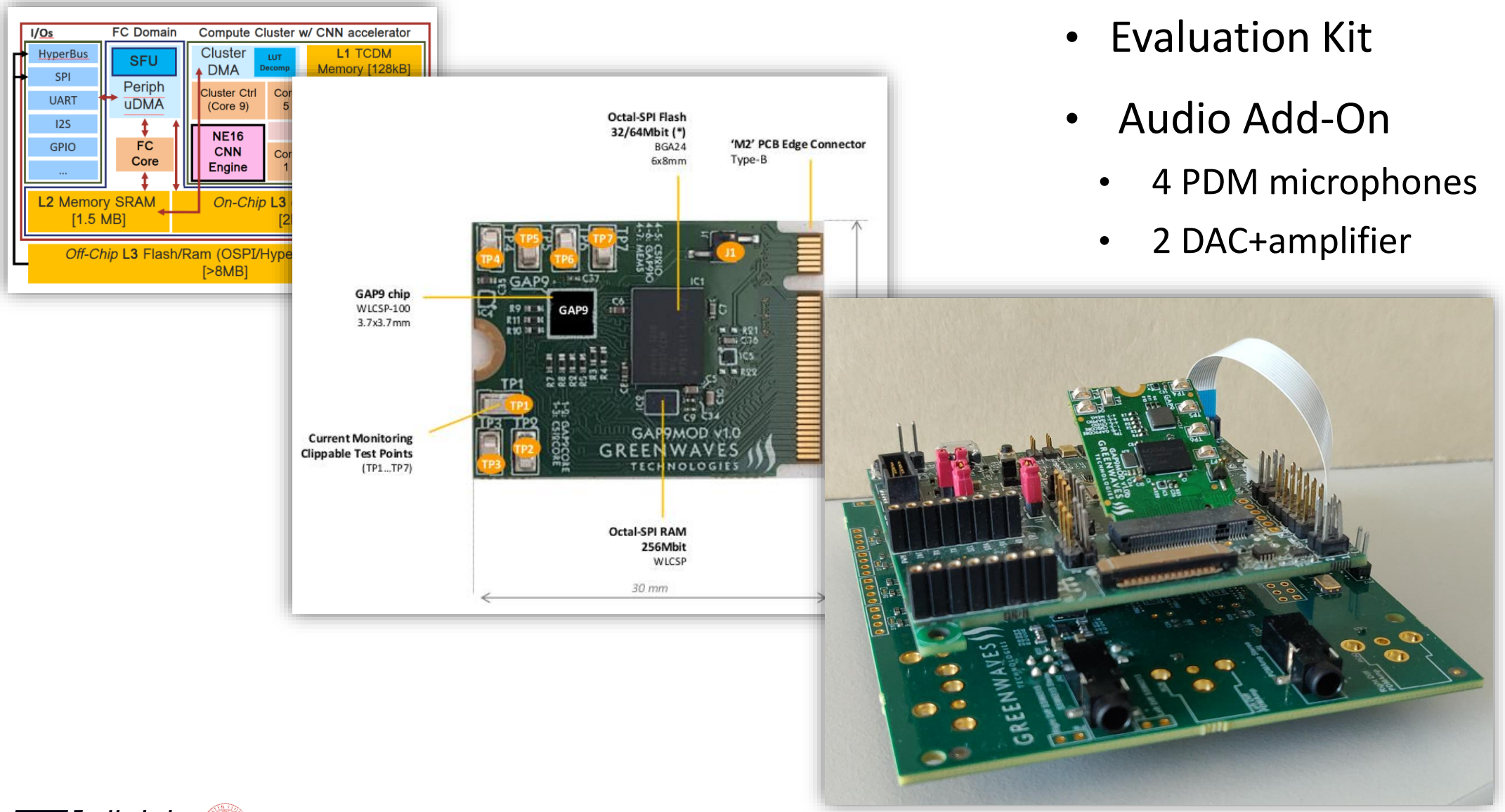
On-Device Domain Adaptation target: GAP9



GAP9Mod

- GAP9 chip
- Octal-SPI RAM (32 Mbit)
- Octal-SPI Flash (256 Mbit)

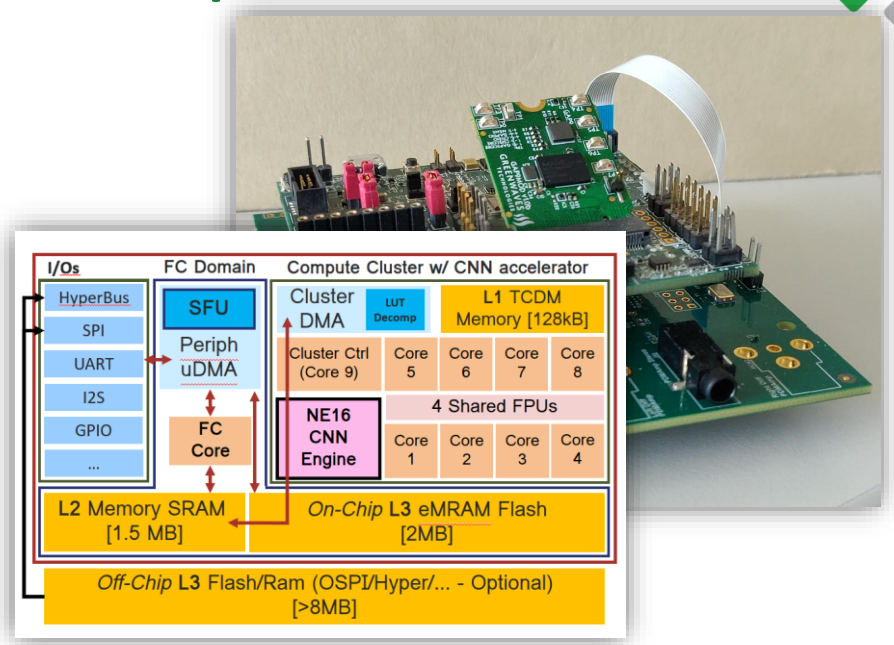
On-Device Domain Adaptation target: GAP9



- Evaluation Kit
- Audio Add-On
  - 4 PDM microphones
  - 2 DAC+amplifier

ODDA pipeline on GAP 9 – a practical example

1. **Flash pre-recorded .WAVs, model parameters**
2. **Streaming inference**
    1. Always-on recording using Audio Add-on mic(s)
    2. MFCC computation using FPUs
    3. Backbone inference on 9-core cluster
    4. Classification using 9-core cluster
3. **On-Device Domain Adaptation**
    1. Record on-site noise, augment pre-recorded samples
    2. MFCC computation using FPUs
    3. Backbone inference on 9-core cluster
    4. Classification using 9-core cluster
    5. Compute loss, compute gradients, update weights



ODDA pipeline on GAP 9 – a practical example

GAP9 in ultra-low-power mode (f_{FC}=50 MHz, f_{CL}=50 MHz, Vdd=650 mV)

Step	Data acq.	MFCC	Backbone	Classifier	Inference
Latency [ms]	50	20.9	26.4	1.6	50
Energy [mJ]	0.54	0.34	0.46	0.02	1.36

- Inference triggered every 50 ms; energy consumption as low as 1.36 mJ

Step	Data prep.	MFCC	Backbone	Classifier	Update	Training
Latency [ms]	2.5	20.9	26.4	1.6	2.3	53.7
Energy [mJ]	0.03	0.34	0.46	0.02	0.03	0.88

- For 1 sample/class for evaluation and 10 samples/class for training:
  - Training process completes in 7.4 s
  - ODDA (incl. evaluation and data acquisition) – energy consumption of 117.5 mJ

Conclusions

- **ODDA improves** over NA-KWS **by specializing** on the target noise
  - Accuracy **gains up to 12%** over NA-KWS, **37%** over NL-KWS at **0 dB** for **DS-CNN S**
  - Average gains of **10% over NA-KWS for – 10 dB**

- **Enables inexpensive on-device learning**
  - **100 samples** (3 MB) for **+5% on *meeting***
  - As little as **10 kB of memory**

- **Demonstrates practical value on GAP9**
  - **0.88 mJ** per training epoch
  - **ODDA completes in 7.4 s**, requiring as low as **118 mJ**

Cristian Cioflan

cioflanc@iis.ee.ethz.ch

`github.com/pulp-platform/odda-for-kws`

Institut für Integrierte Systeme – ETH Zürich
Gloriastrasse 35
Zürich, Switzerland

DEI – Universitá di Bologna
Viale del Risorgimento 2
Bologna, Italy

ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Bibliography

[1] C. Cioflan et al., “Towards On-device Domain Adaptation for Noise-Robust Keyword Spotting”, AICAS, 2022

[2] Pete Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition”, 2018

[3] Joachim Thiemann, Nobutaka Ito & Emmanuel Vincent , “DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments”, ICA, 2013

[4] Francesco Conti, "Technical Report: NEMO DNN Quantization for Deployment Mode", 2020; https://github.com/pulp-platform/nemo

[5] A. Burello et al., "DORY: Automatic End-to-End Deployment of Real-World DNNs on Low-Cost IoT MCUs", IEEE Transactions on Computers, 2021; https://github.com/pulp-platform/dory

[6] D. Nadalini et al., "PULP-TrainLib: Enabling On-Device Training for RISC-V Multi-core MCUs Through Performance-Driven Autotuning", Embedded Computer Systems: Architectures, Modeling, and Simulation, 2022; https://github.com/pulp-platform/pulp-trainlib

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org