tinyML® Research Symposium

Enabling Ultra-low Power Machine Learning at the Edge

March 27, 2023

TINY ML

www.tinyML.org

Symposium
March 27th, 2023
San Francisco, USA

TINY ML

life.augmented

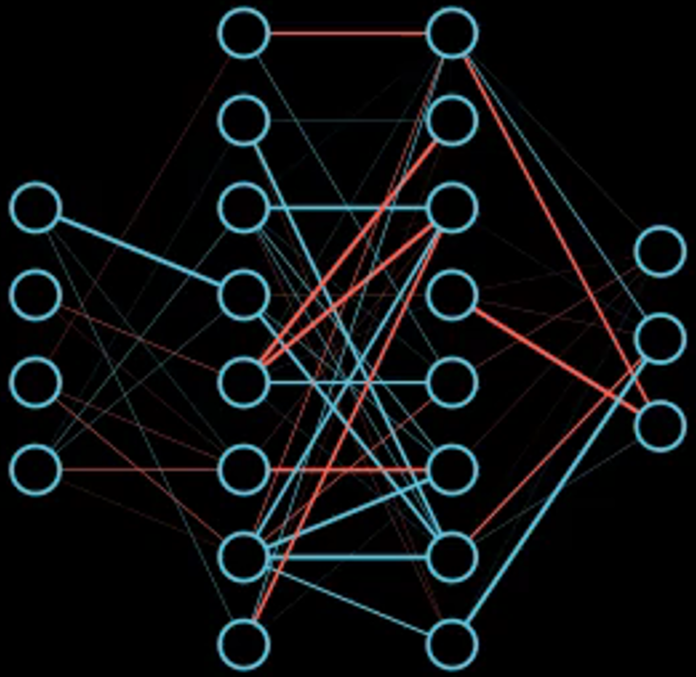
TinyRCE: Forward Learning Under Tiny Constraints

Danilo Pau, Prem Kumar Ambrose

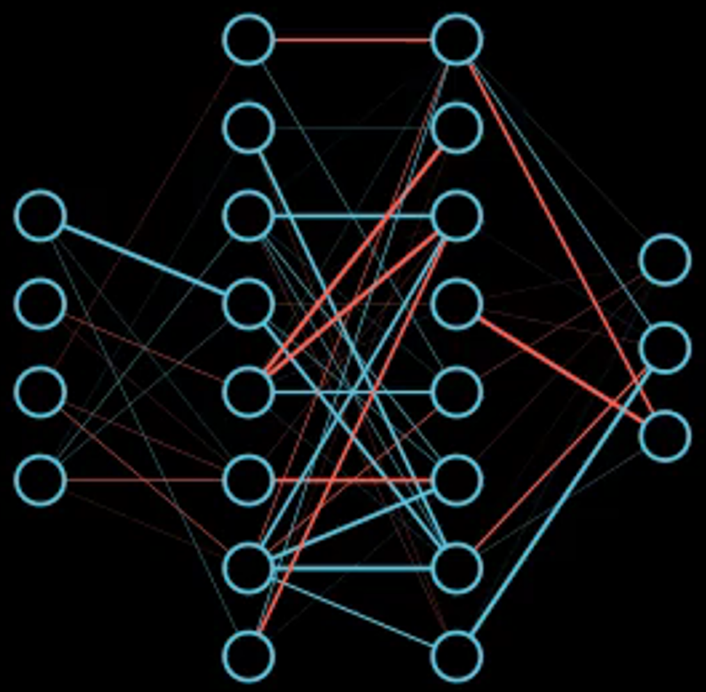
System Research and Applications

STMicroelectronics, Agrate Brianza, Italy

Backpropagation

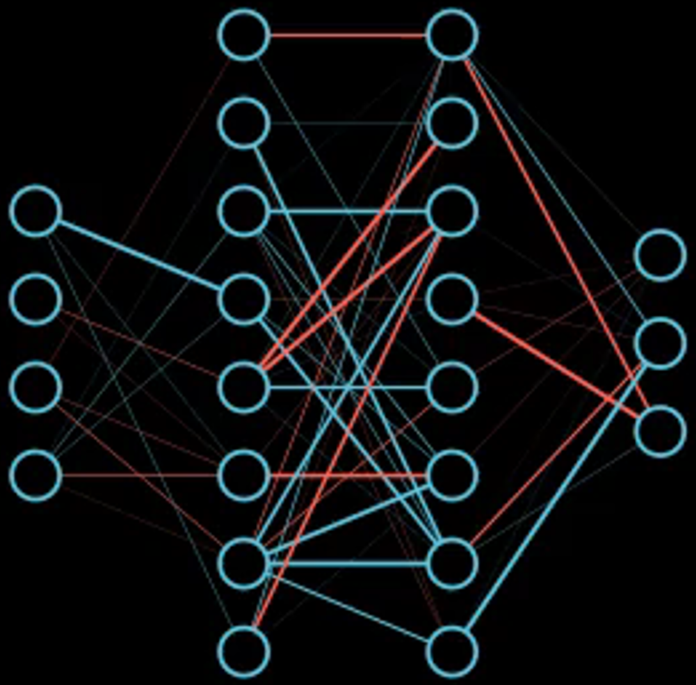


Backpropagation



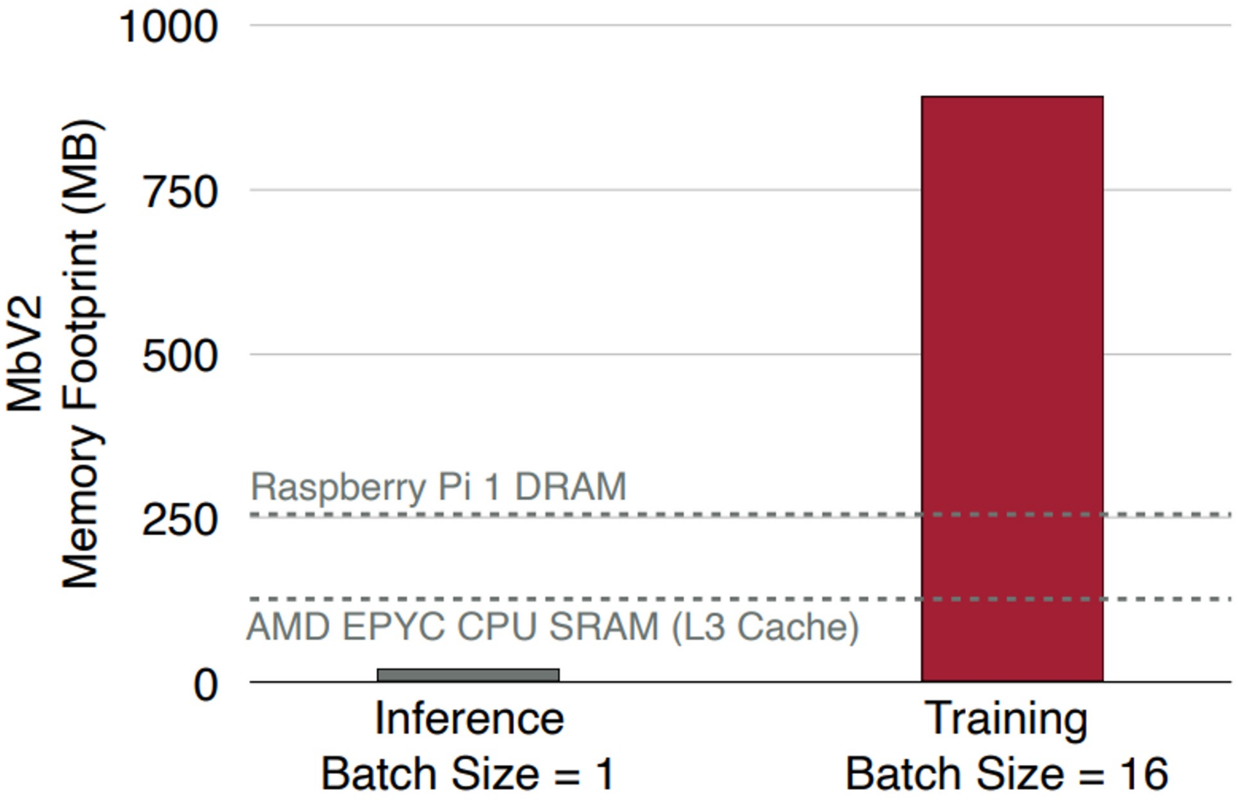
life.augmented

Backpropagation



life.augmented

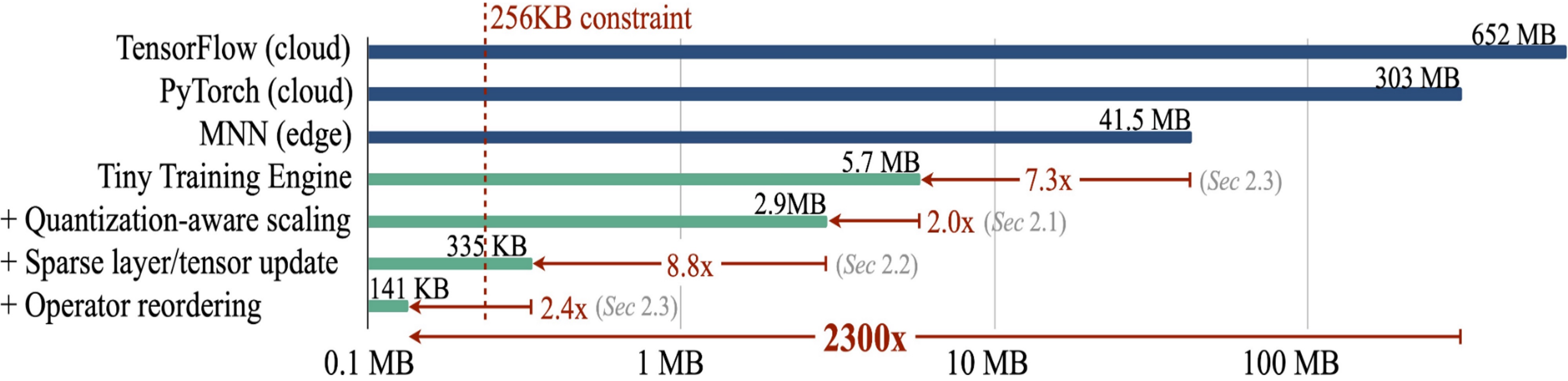
Reduce activations, not trainable parameters for efficient on-device learning[1]



[1]H. Cai, C. Gan, L. Zhu, and S. Han, "Tinytl: Reduce memory, not parameters for efficient on-device learning," in Advances in Neural Information Processing Systems, vol. 33, 2020

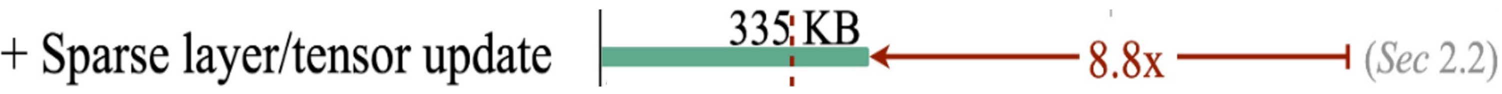
life.augmented

MCUNetV3[2]



[2]J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, C. Gan, and S. Han, “On-device training under 256kb memory,” 2022. [Online]. Available: https://arxiv.org/abs/2206.15472

MCUNetV3[2]



[2]J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, C. Gan, and S. Han, “On-device training under 256kb memory,” 2022. [Online]. Available: https://arxiv.org/abs/2206.15472

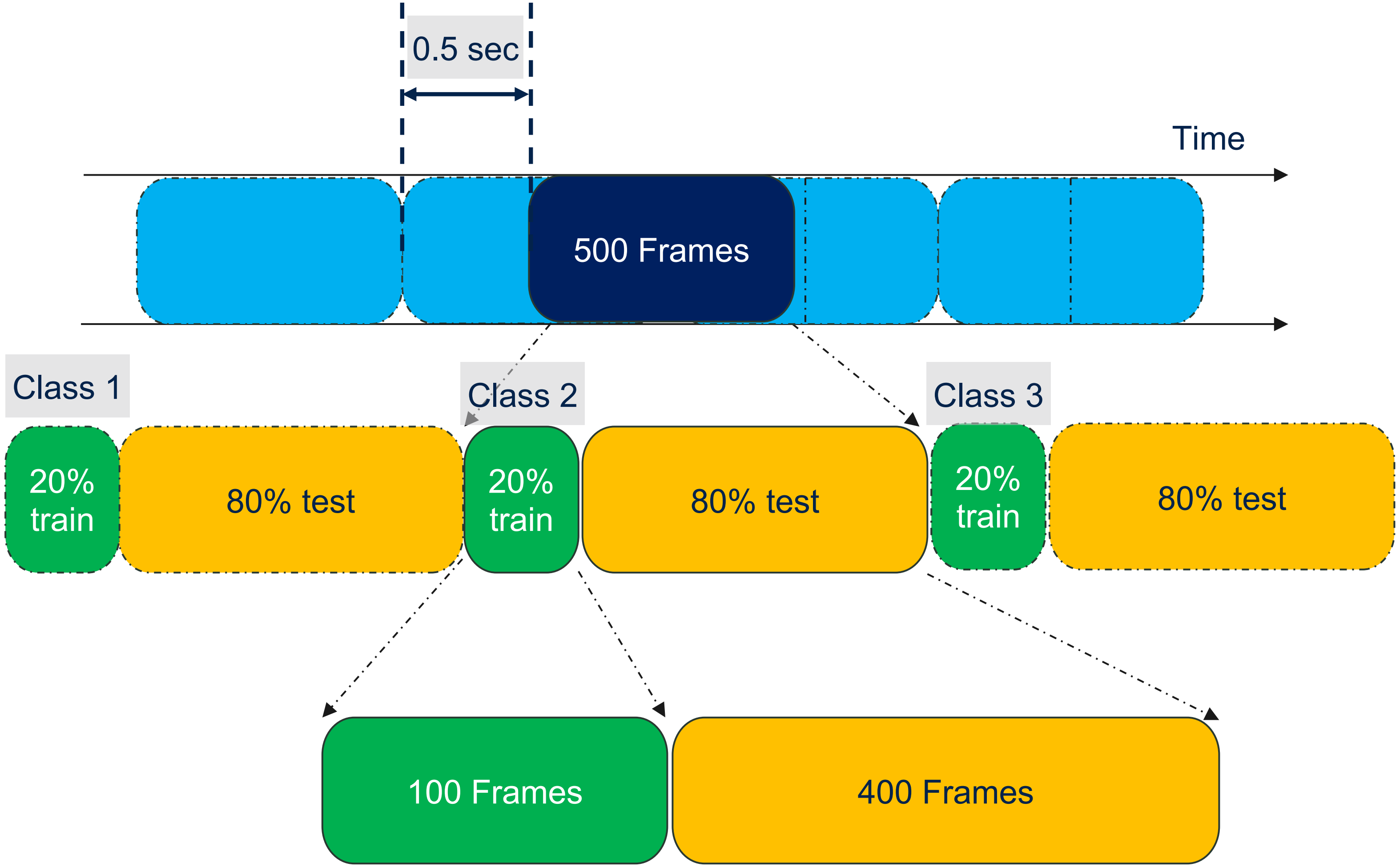
Question

Can **incremental learning** of **multiple** categories happens, totally **on-line** w/out catastrophic forgetting, **w/out K-fold back-prop** and be deployable on **memory constrained** devices?

Case studies

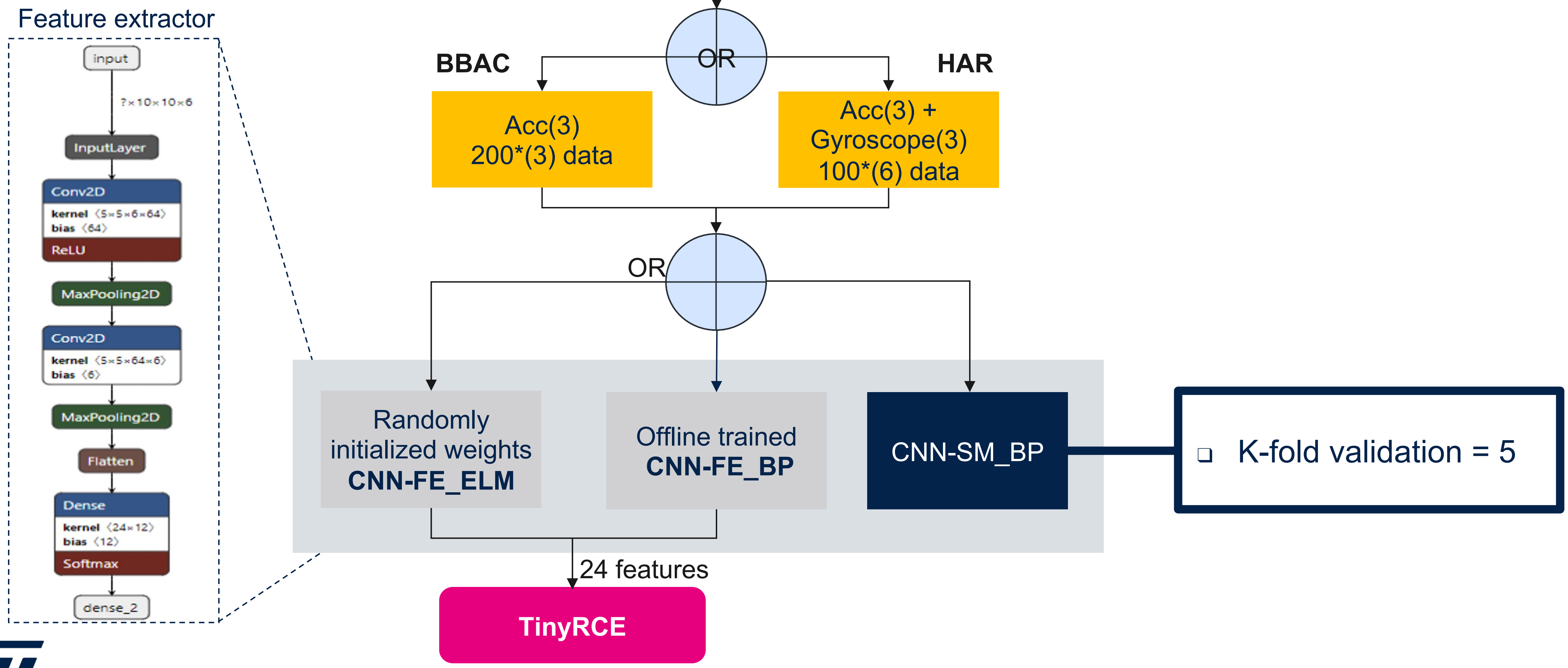
	SHL	PAMAP2	CWRU
Sampling freq [Hz]	100	100	12000
Type	HAR	HAR	BBAC
Sensor data	Accelerometer (3 axis) + gyroscope (3 axis)	Accelerometer (3 axis) + gyroscope (3 axis)	Accelerometer (3 axis)
Number of classes	8	12	10
Number of users	3	8	1
Frame size	1 sec (6 * 100)	1 sec (6 *100)	16.6ms (3*200)
References	H. Gjoreski et al., "The University of Sussex-Huawei Locomotion and Transportation Dataset for Multimodal Analytics With Mobile Devices," in IEEE Access, vol. 6, pp. 42592-42604, 2018, doi: 10.1109/ACCESS.2018.2858933.	A. Reiss and D. Stricker, "Introducing a New Benchmarked Dataset for Activity Monitoring," 2012 16th International Symposium on Wearable Computers, Newcastle, UK, 2012, pp. 108-109, doi: 10.1109/ISWC.2012.13.	https://engineering.case.edu/bearingdatacenter/download-data-file

Interleaved incremental learning and testing



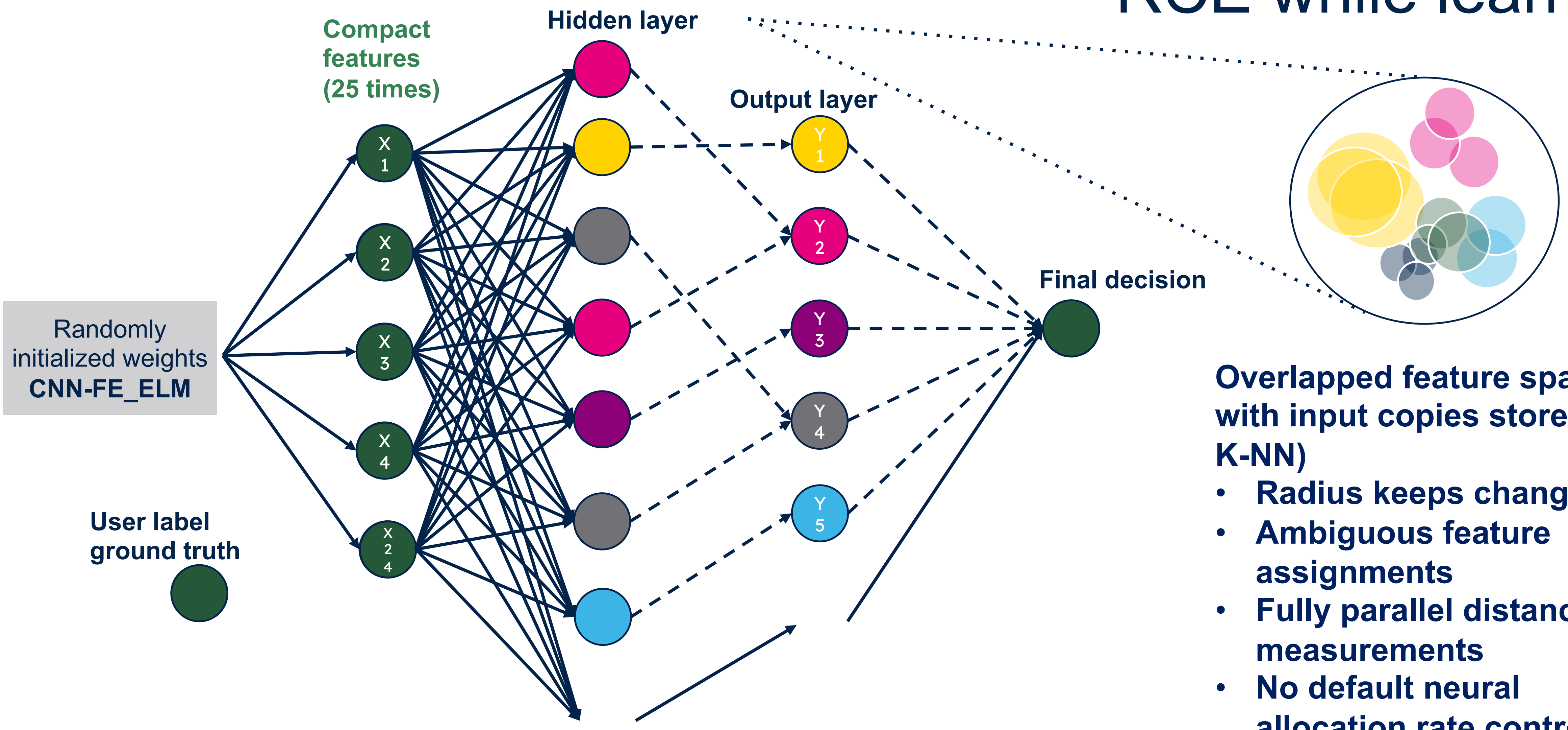
life.augmented

Proposed framework



life.augmented

RCE while learning



Overlapped feature space with input copies stored (like K-NN)

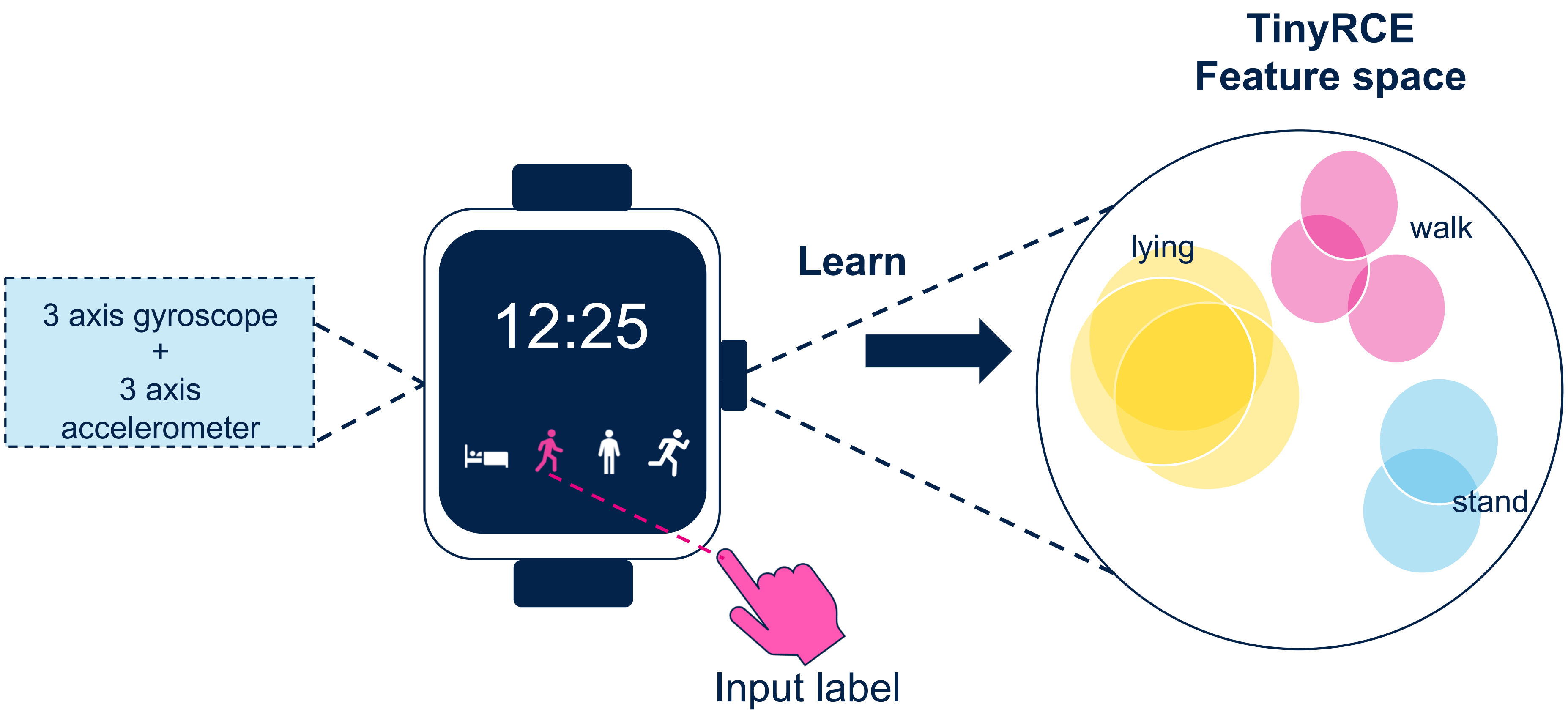
- **Radius keeps changing**
- **Ambiguous feature assignments**
- **Fully parallel distance measurements**
- **No default neural allocation rate control**

[35] M. J. Hudak, "RCE networks: an experimental investigation," IJCNN-91- Seattle International Joint Conference on Neural Networks, Seattle, WA, USA, 1991, pp. 849-854 vol.1, doi: 10.1109/IJCNN.1991.155290.

[36] MICHAEL J. HUDAK (1992) RCE CLASSIFIERS: THEORY AND PRACTICE, Cybernetics and Systems, 23:5, 483-515, DOI: 10.1080/01969729208927478

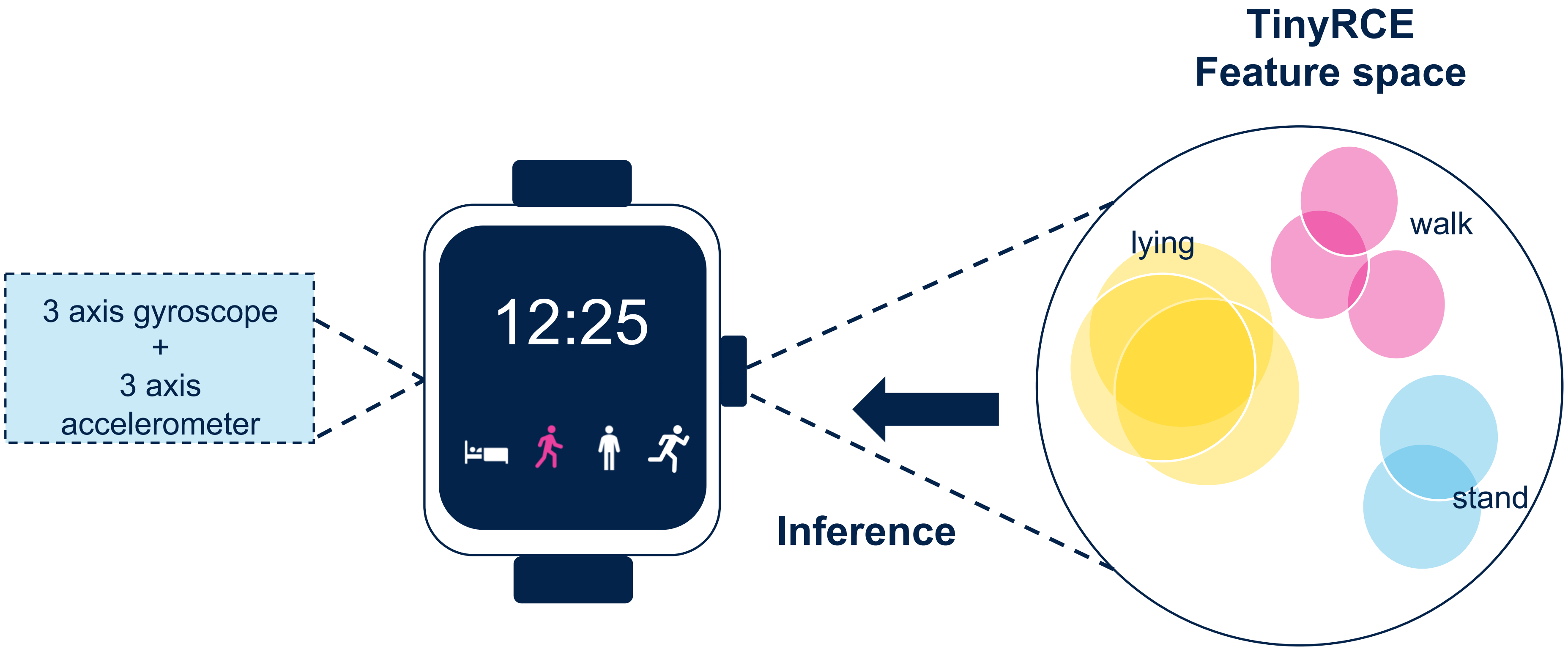
User annotation process

TinyRCE was meant to detect unknown input patterns, thus triggering the user to provide the corresponding label.



User annotation process (e.g. HAR)

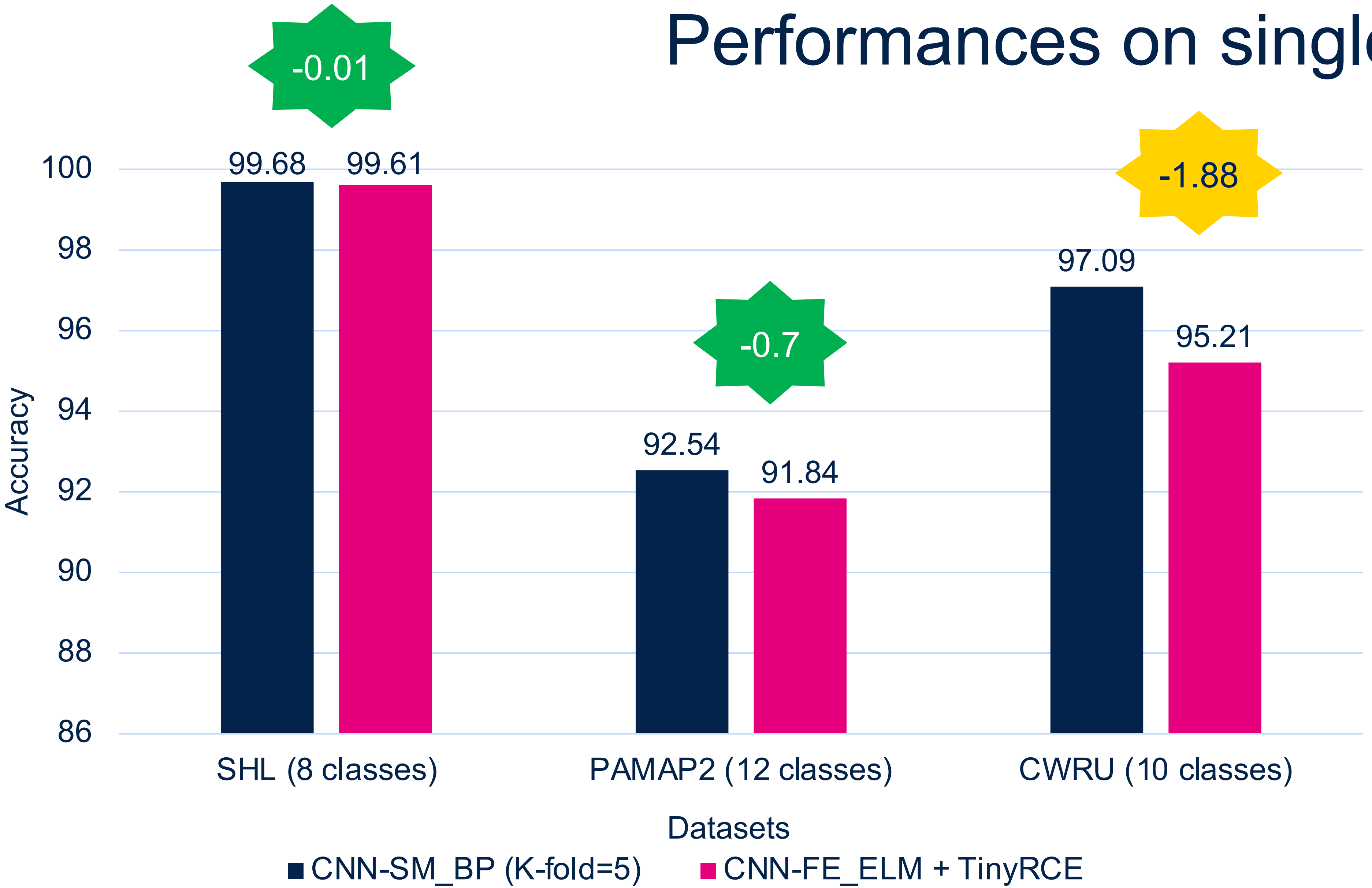
TinyRCE was meant to detect unknown input patterns, thus triggering the user to provide the corresponding label.



TinyRCE vs baseline RCE

✓	Learn within memory caps	The maximum number of hidden neurons the MCU shall store is depending on the budgeted memory.
	Aging	Increased/decreased depending on $\frac{dist(h_j, x_t)}{R_j}$
	Pruning	Prevent uncontrolled instantiation of hidden neurons (causing overflowing available RAM) during the learning phase
?	Unambiguous assignment	When input falls into multiple hidden neurons regions of influence, the output of the one to which the input falls closer (Hamming or Euclidian distance) is considered

Performances on single user



life.augmented

Deployability on MCU

* random weights

Metrics	CNN-SM_BP (single inference) FP32	CNN-SM_BP (learning) FP32 (K fold=1)	CNN-FE_ELM + TinyRCE (single inference) FP32	CNN-FE_ELM + TinyRCE (learning) FP32
MACC	1.213 M	**8 G**	1.251 M	**23.4 M**
FLASH (KiB)	76.45	-	76.45 (*)	
RAM (KiB)	13.59	**954.57**	40.2	
Latency STM32L4 (ms) @80MHz	123.4	**813.25 sec**	127.2	**2.38 sec**
Latency STM32H7 (ms) @480 MHz	11.21	**73.86 sec**	11.55	**216**

$$MACC\_Inference = h * [(n * 5) + 10]$$

$$MACC\_Learning = (h * [(n * 5) + 10]) * (F * E)$$

life.augmented

Conclusions

TinyRCE

- Capable of personalized, incremental learning
- w.r.t. backpropagation
  - **-0.01% to -1.88% single user**
  - **-2% to -3.52% avg 8 users, 8 commons classes**
- Deployable on MCUs, w.r.t. backpropagation
  - **learning 342x less complex**
  - **inference same complexity**
- Memory
- **same FLASH with random weights**
- **24x less RAM**

Tiny learning

- Open new path to applications
- Save dramatically complexity on the cloud
- Scale faster!
- More reactive Federated Learning

Future works

- Extend to **MLCommons/Tiny** tasks
- **Deep quantization** of the CNN_FE to save further memory
- **Add concept drift detection** to trigger learning phase autonomosly

life.augmented

Thanks for listening.

Q&A

danilo.pau@st.com

Copyright Notice

This presentation in this publication was presented at the tinyML® Research Symposium (March 27,2023). 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