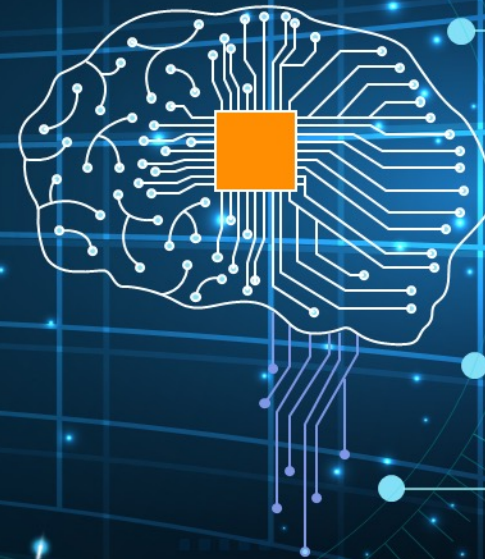


The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)

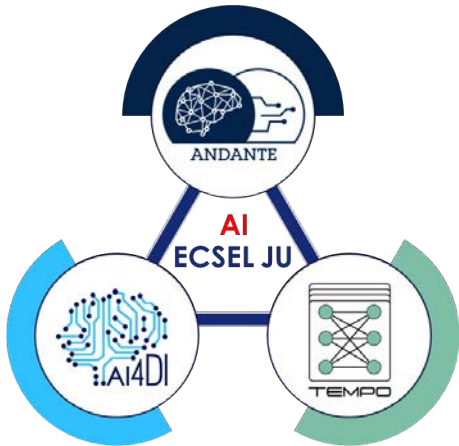


AI



Vienna, Austria
25-26 July 2022

The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)



Deploying a Convolutional Neural Network on Edge MCU and Neuromorphic Hardware Platforms

Simon Narduzzi, CSEM, Switzerland

Dorvan Favre, CSEM, Switzerland

Nuria Pazos Escudero, Haute—Ecole Arc, Switzerland

L. Andrea Dunbar, CSEM, Switzerland

Vienna, Austria 25-26 July 2022

Edge computing

Edge processing is a significant target in Machine Learning

- Embedded ML: autonomous cars, drones, IoT, wearables...
- Reduced workload on servers
- Carbon footprint

Efforts from research and industry:

- Growth of interest: TinyML Community
- ANDANTE, TEMPO, AI4DI, ...

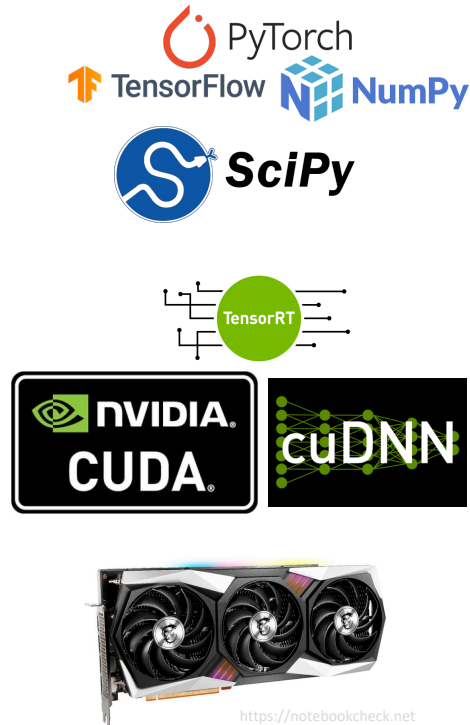
Edge ML is still challenging...

- What are the barriers?



Standardization of Edge Devices

Massive acceptance of Deep Learning
due to standardization of tools



Edge?

~~Standardization~~

~~Select the best~~

~~Comparison~~

Edge Devices

Model

Manual!

Optimization

Dedicated hardware

Reduce energy

Challenges

- The standardization of devices requires selection of the “best” one
- Selection = Comparison of devices
 - Lack of standard deployment tools
 - Lack of optimization standards
 - Lack of automated tools
 - Variability of device memory, ISA, compute capability...
 - Lack of baseline performance (absolute? relative?)

What has been done in the field yet?

Comparison

- TinyMLPerf : Benchmarks algorithms for (a few) Tiny devices
- Ostrau et al., 2020: Benchmarking for event-based processors

Deployment

- Heim et al. 2021, Falbo et al. 2019, Sanchez-Iborra et al. 2020:
 - Intra-processor comparison
 - Cortex-M processors
 - STM32 processors
- Osman et al., 2022:
 - CNN on Arduino Nano BLE and STM32 NUCLEO-F401RE

Our work

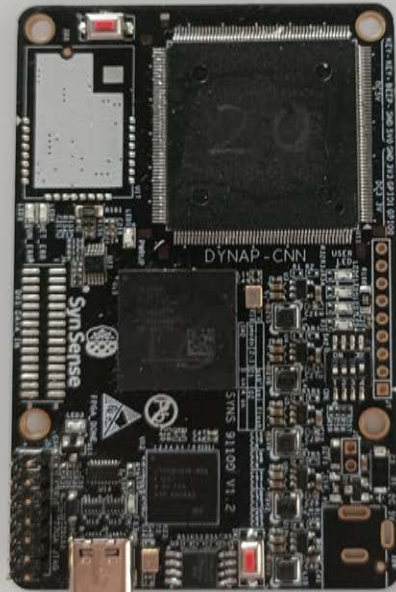
Tomorrow: “Benchmarking Neuromorphic Computing for Inference” by Loreto Mateu

Our work

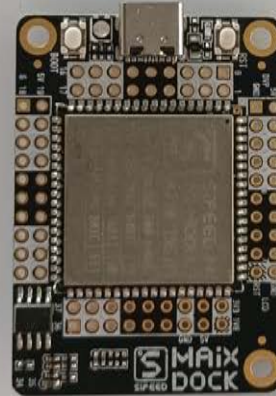
- Inter-processor comparison
- Automated deployment pipeline
- Open-source tool

Experimental setup: Board

DynapCNN



Kendryte K210



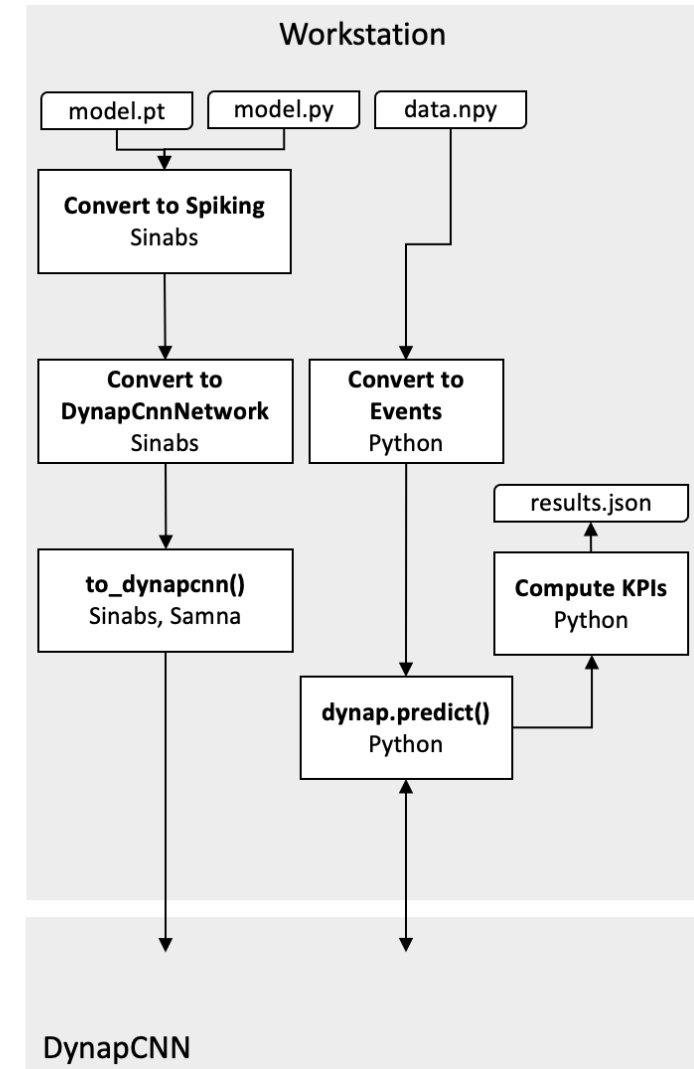
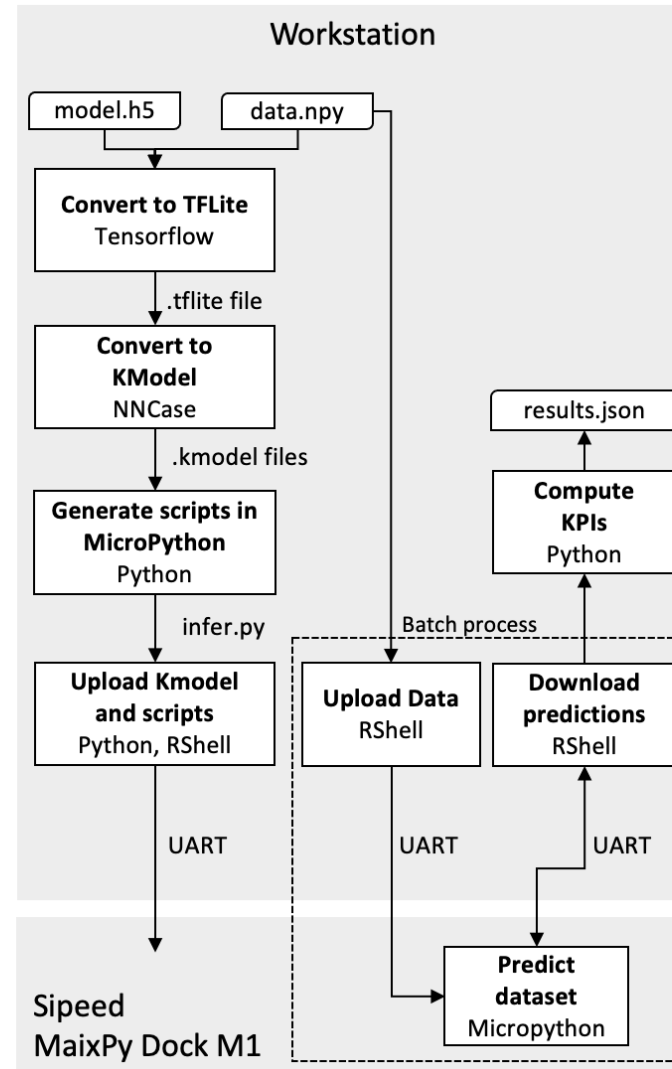
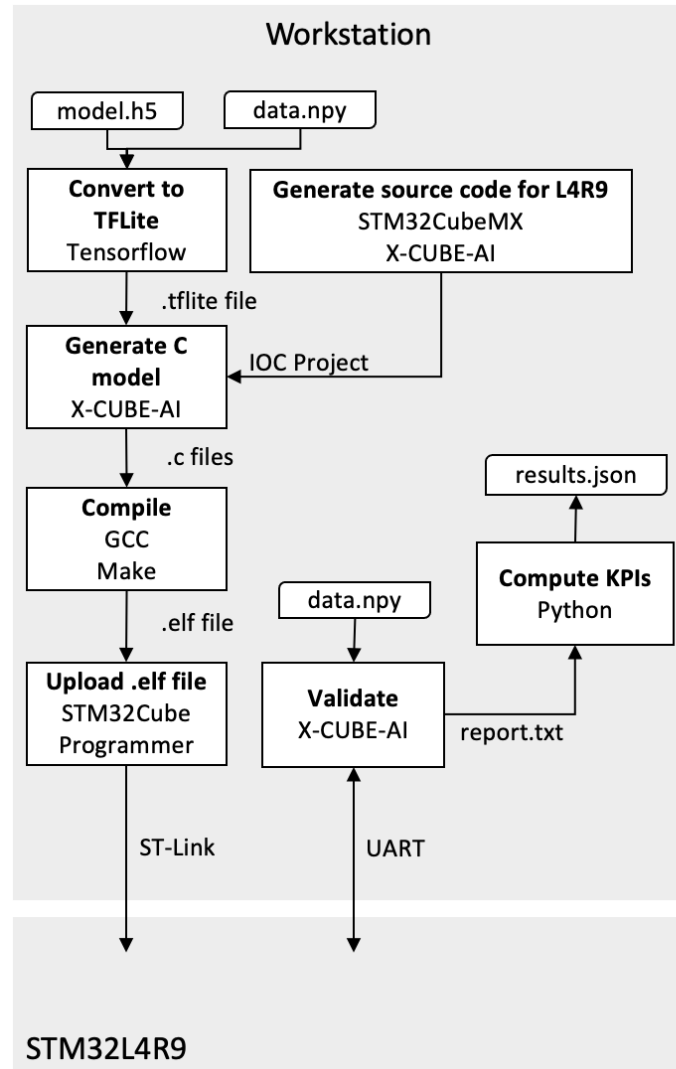
STM32L4R9



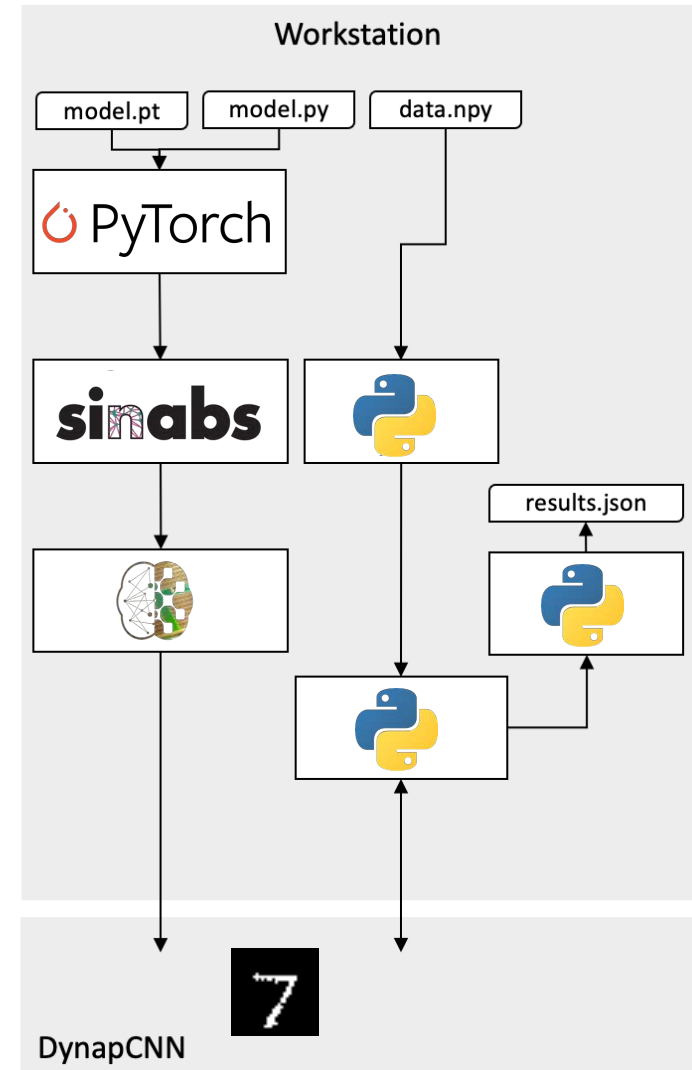
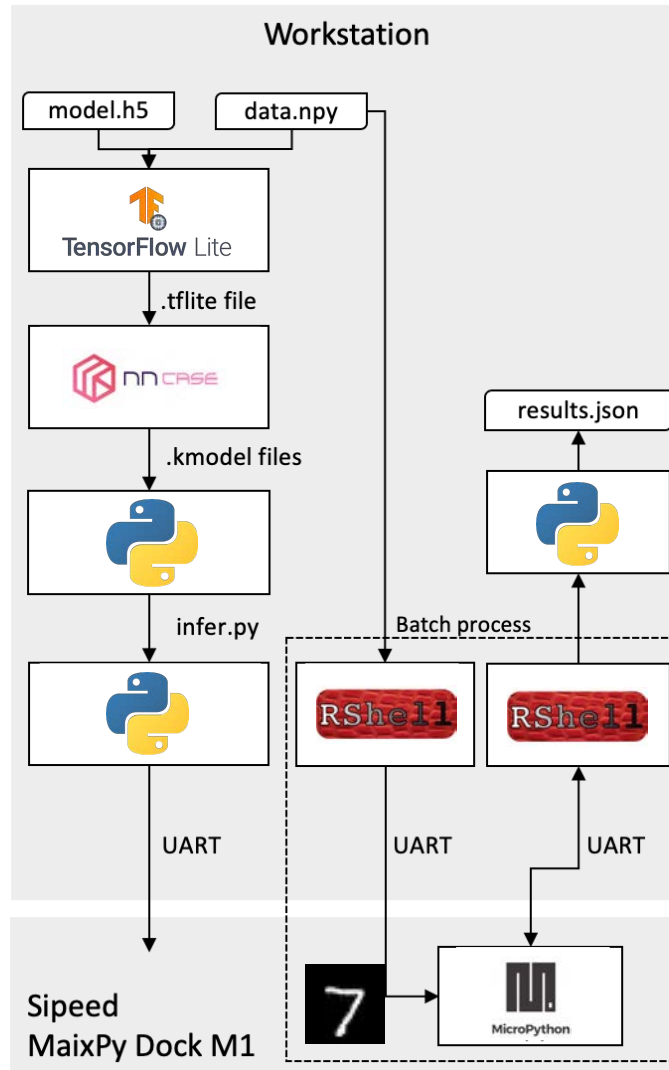
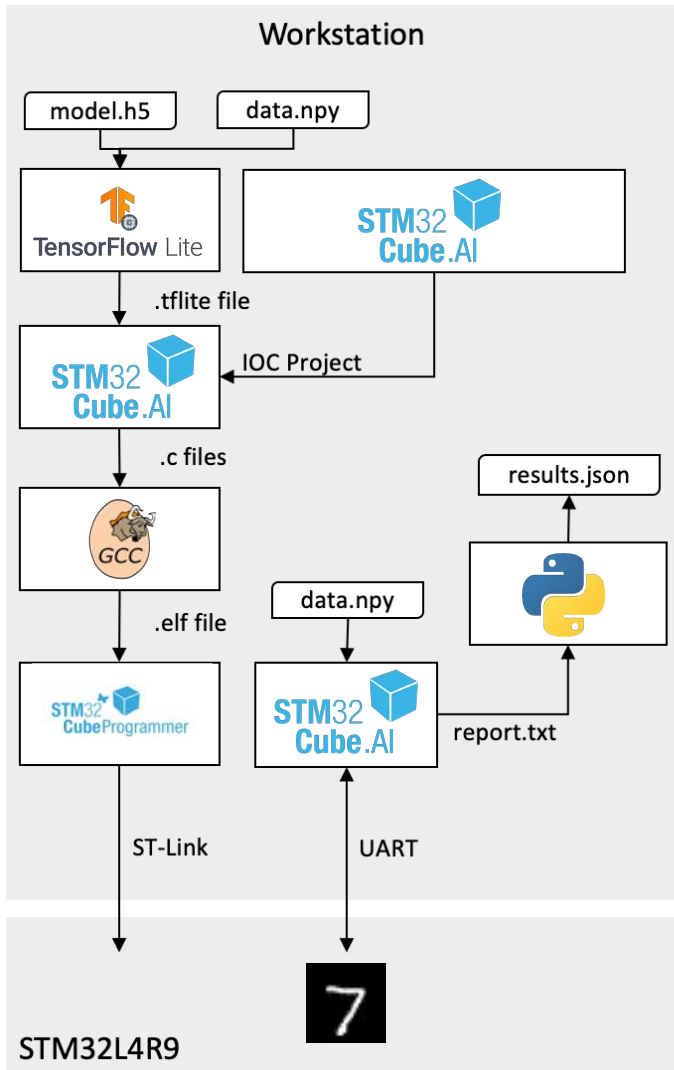
Experimental setup: Board

	DynapCNN	Kendryte K210	STM32L4R9
ISA	Event-based	Dual-core RISC-V 64b	ARM Cortex-M4
Power	1mW	300mW	66mW
Max Freq	-	900 MHz	120 MHz
TOPS/W	-	3.3	-
Standalone	X	✓	✓
Event-Based	✓	X	X
Language	Python	MicroPython	C

Deployment Pipelines



Deployment Pipelines (illustrated)



Implementation: challenges

Documentation quality

Tools

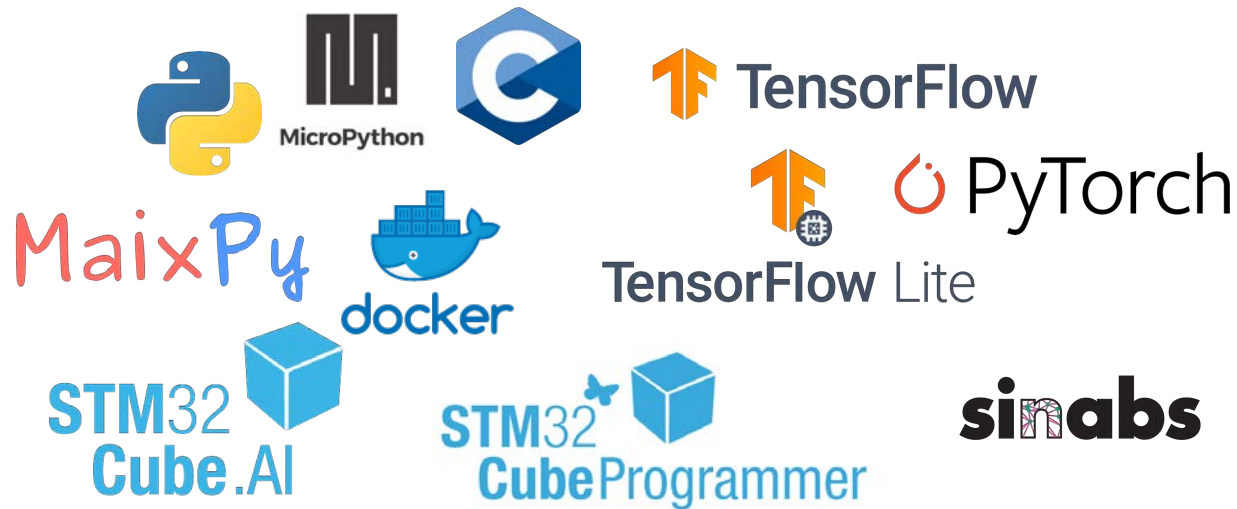
Instruction Set

Connectivity (UART, USB)

Speed

Portability

Memory



Data representation

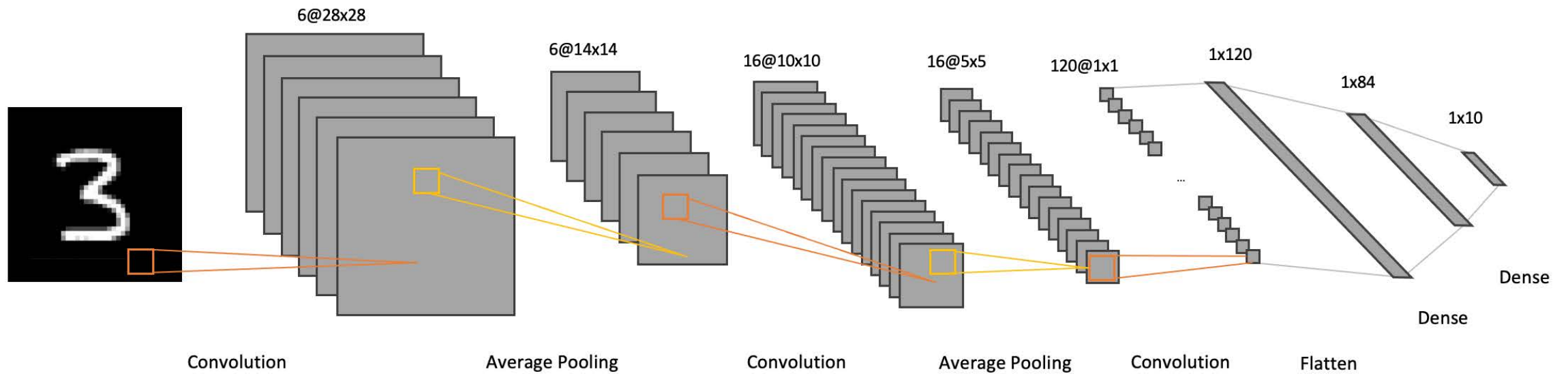
Frame



Events



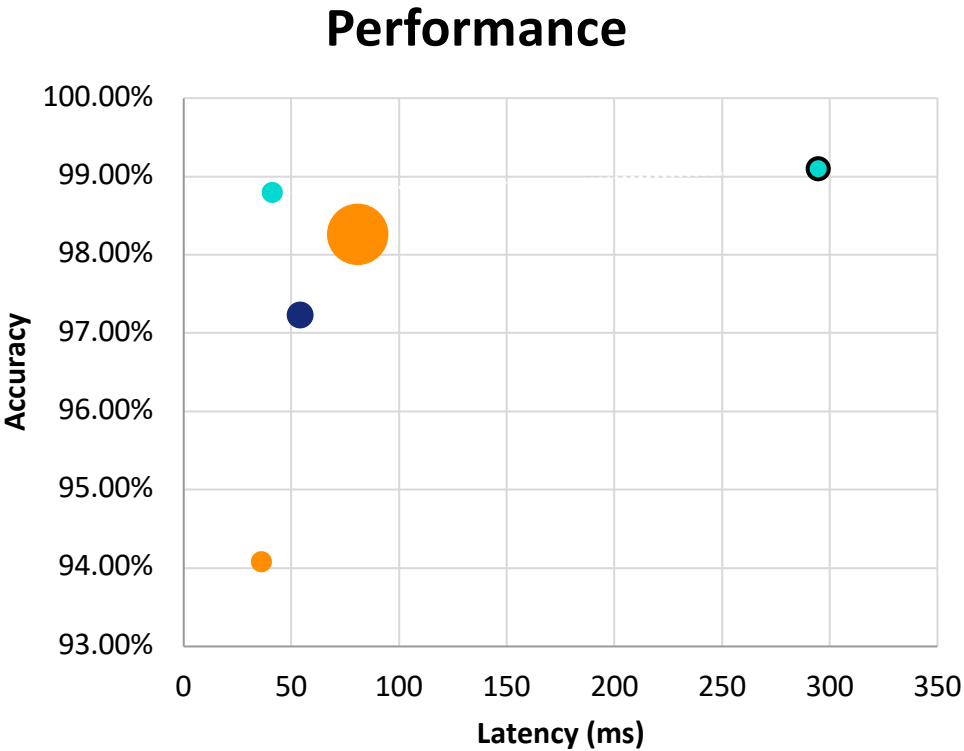
Network Architecture








LeNet-5 Architecture


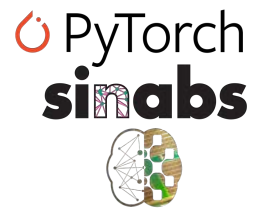
"Gradient-based learning applied to document recognition", LeCun et al., 1998

Deployment Results



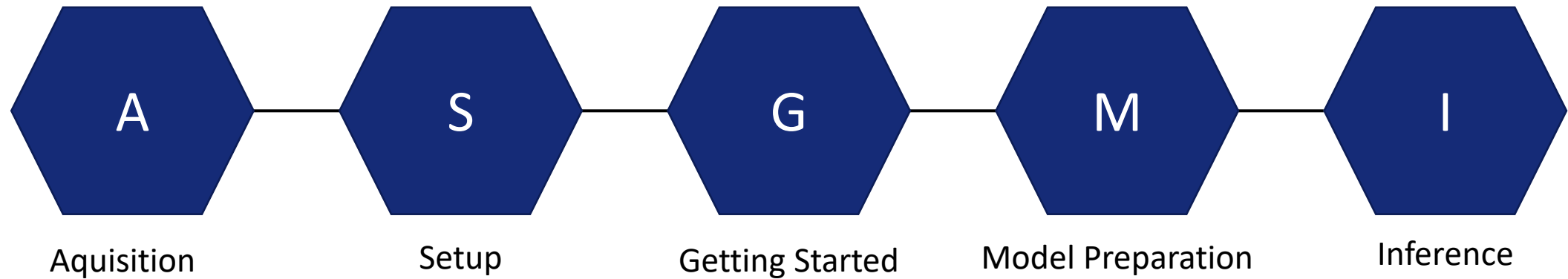






Platform	Kendryte K210	STM32L4R9		DynapCNN
Bit precision	float-32	float-32	int-8	float-32
Size (KB)	94.2	359.2	90.5	-
Accuracy	97.23%	98.26%	94.07%	98.79% / 99.09%
Latency (ms)	54.17	80.82	36.23	41.3 / 294.9
Energy (μJ)	-	-	-	144.5

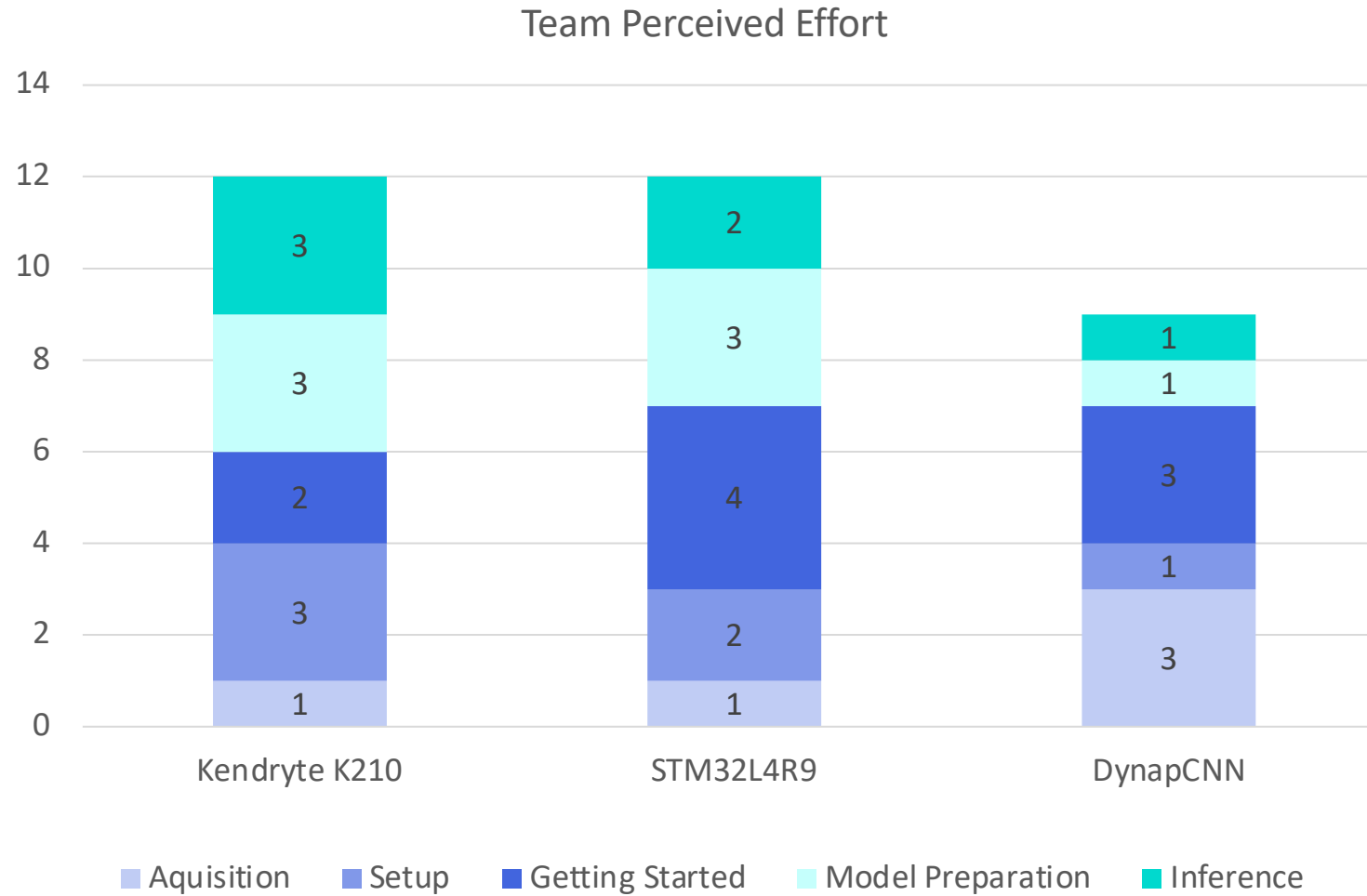
Ease of Deployment



Scale:

- 1: Straight forward (script, command...)
- 5: Need constructor support

Deployment Results



Conclusion and Outlook

Conclusion

- Automated the deployment on 3 constructors
- Provided a framework to evaluate the effort
- Comparison is still a challenge

Outlook

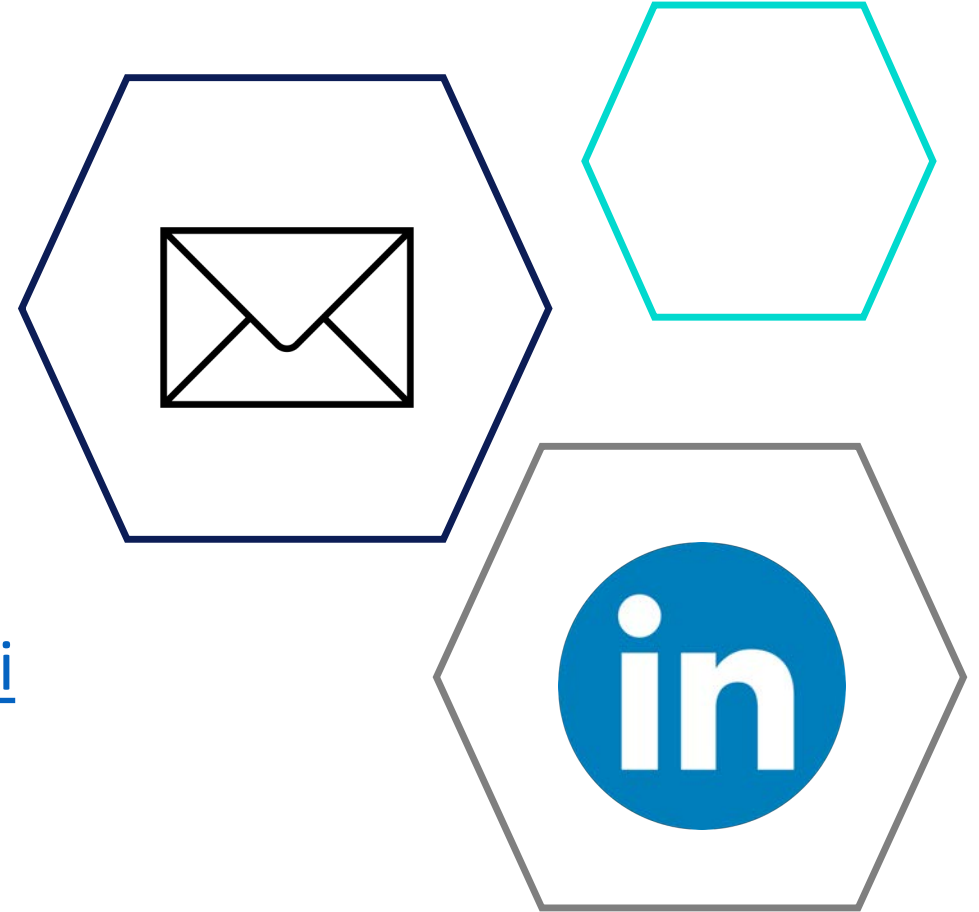


- Extend the support of devices
- Release the code as library
- Develop testbenches
- Define the KPIs (Hard and Soft)
- Validate / Improve the usability measure

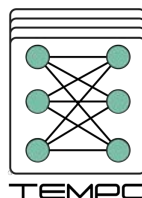
Please get in touch!

Simon Narduzzi

- simon.narduzzi@csem.ch
- <https://linkedin.com/in/narduzzi>



Event Organisers



The Key Digital Technologies Joint Undertaking - the Public-Private Partnership for research, development and innovation – funds projects for assuring world-class expertise in these key enabling technologies, essential for Europe's competitive leadership in the era of the digital economy. KDT JU is the successor to the ECSEL JU programme. www.kdt-ju.europa.eu

The AI4DI project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826060. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the national authorities. www.ai4di.eu

The TEMPO project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Switzerland. www.tempo-ecsel.eu

The ANDANTE project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Portugal, Spain, Switzerland. www.andante-ai.eu



Thank You
For your attention

@ simon.narduzzi@csem.ch