tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

June 26 - 28, 2023

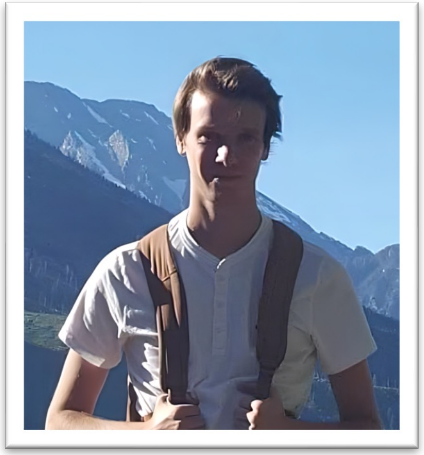
TINY ML

www.tinyML.org

AUDIO-VISUAL ACTIVE SPEAKER DETECTION ON EMBEDDED DEVICES

TINY ML

Baptiste POUTHIER
PhD student



S. DION, L. PILATI
Voice & Audio Team, NXP

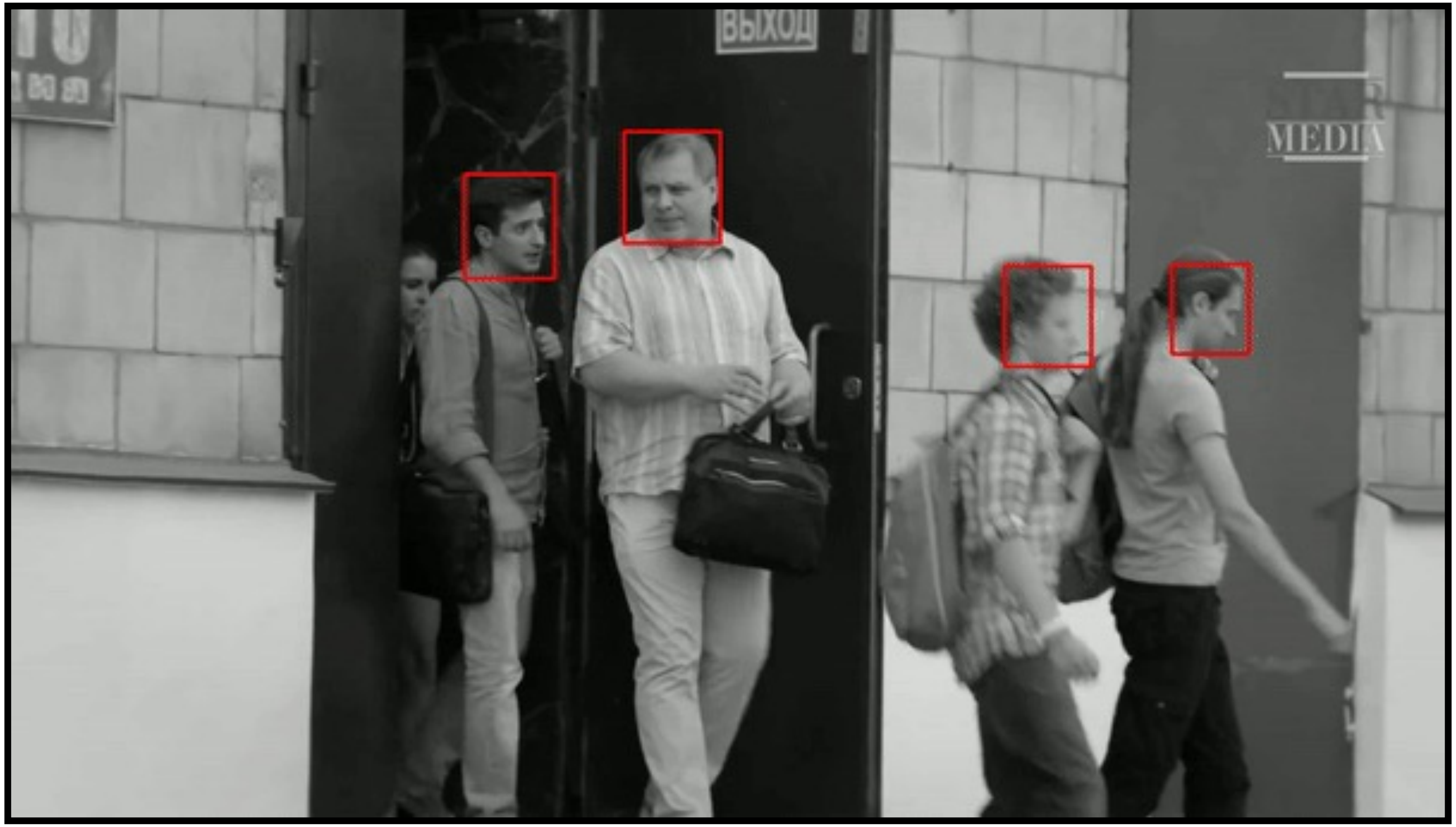
F. PRECIOSO, C. BOUVEYRON
Maasai Team, INRIA

Inria

NXP

SECURE CONNECTIONS
FOR A SMARTER WORLD

Active Speaker Detection – The task

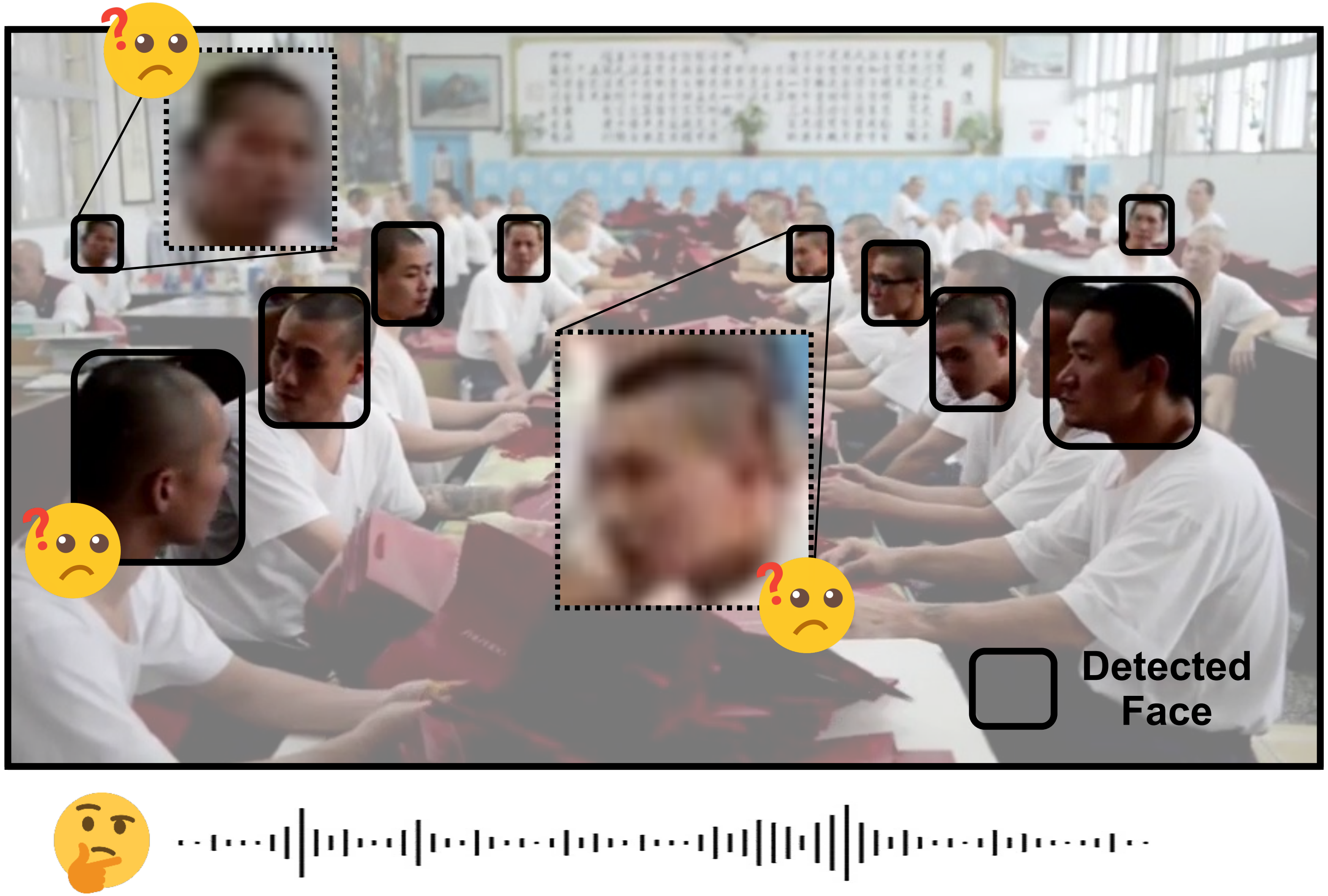


🟩 **Speaking**

🟥 **Not Speaking**

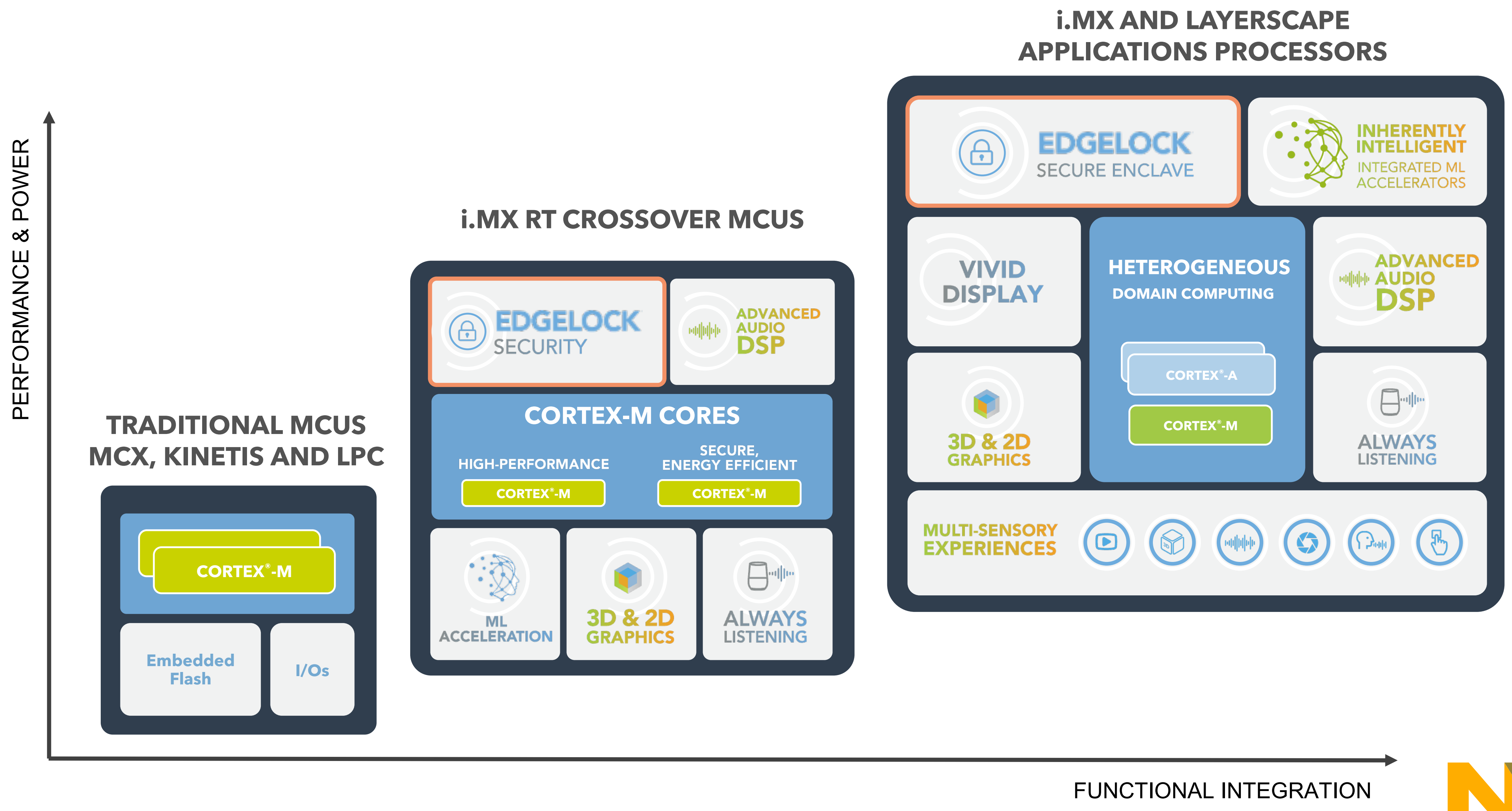
Active Speaker Detection – Use case: video conferencing

Active Speaker Detection – The challenge



- 😕 Low resolution and/or indiscernible faces
- 🤔 Multi-speaker scenario

Integration In Embedded Devices – From MPU to MCU

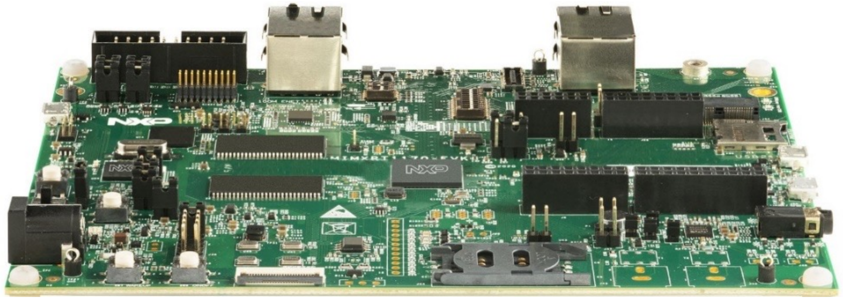


Integration In Embedded Devices – From MPU to MCU

i.MX 8M Plus: High-end NXP MPU

- 4x Arm® Cortex® – A53 (1.8 GHz)

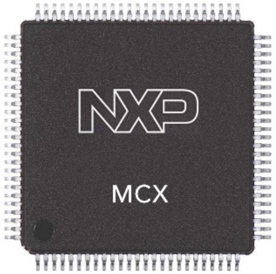
- NPU (2.3 TOPS)



i.MX RT1170: High-end NXP MCU

- Arm® Cortex® – M7 (1 GHz)

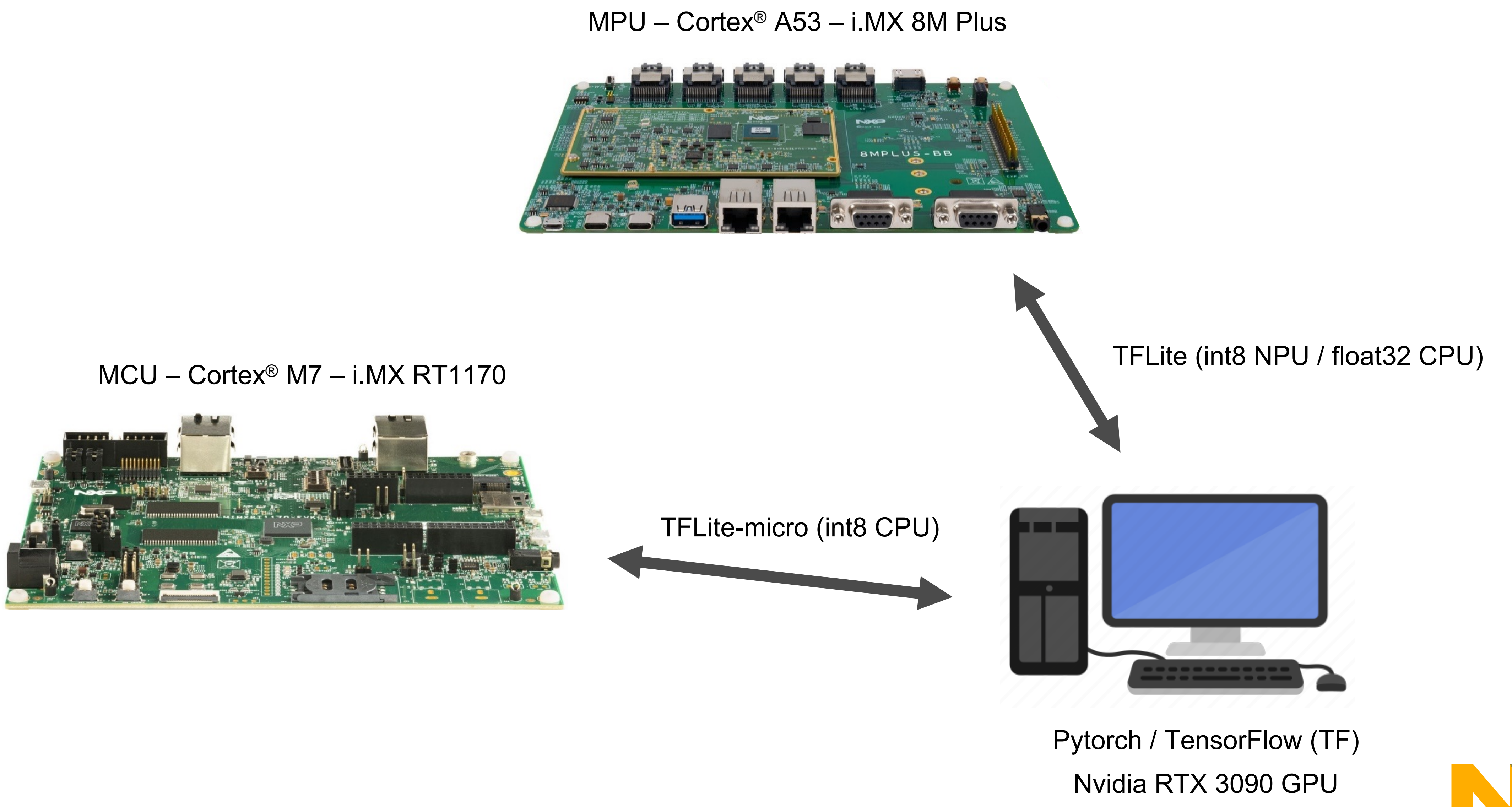
- Arm® Cortex® – M4 (400 MHz)



MCX N Serie: NXP MCU with NPU

- 2x Arm® Cortex® – M33 (150 MHz)

Integration In Embedded Devices – From MPU to MCU



MPU – Cortex® A53 – i.MX 8M Plus

TFLite (int8 NPU / float32 CPU)

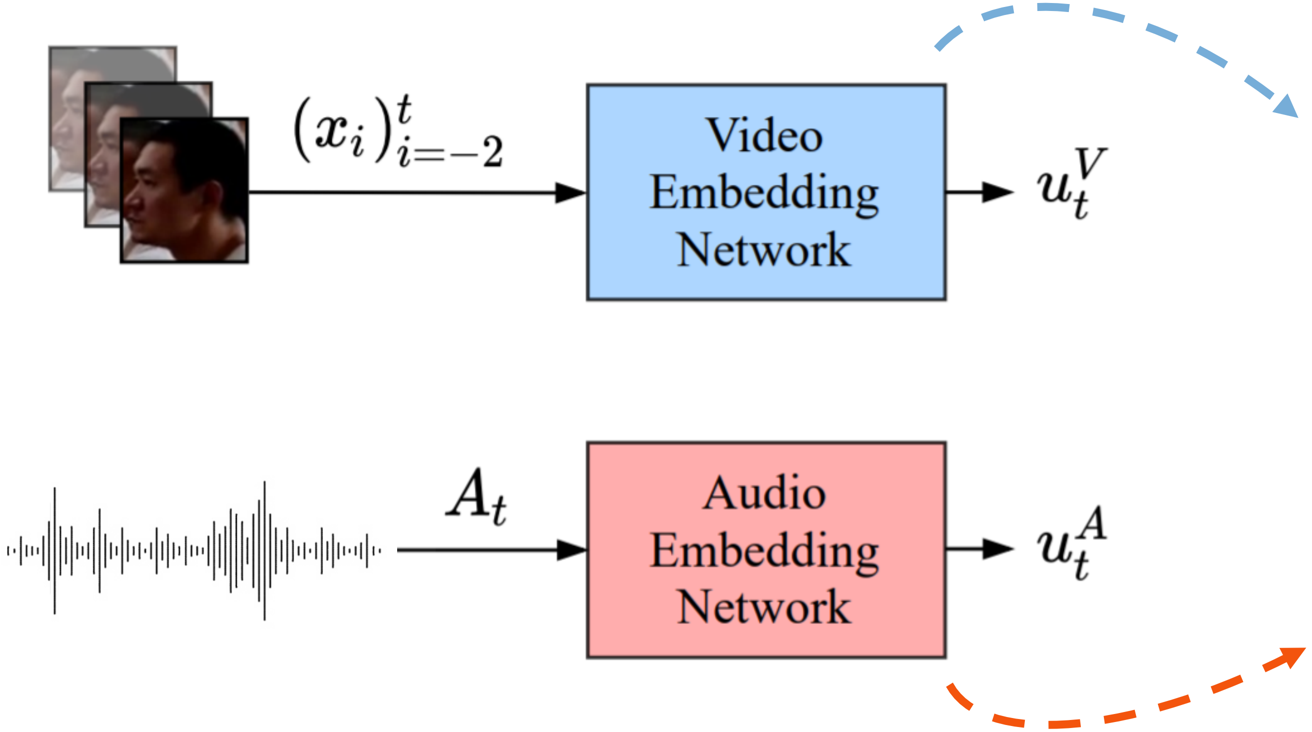
MCU – Cortex® M7 – i.MX RT1170

TFLite-micro (int8 CPU)

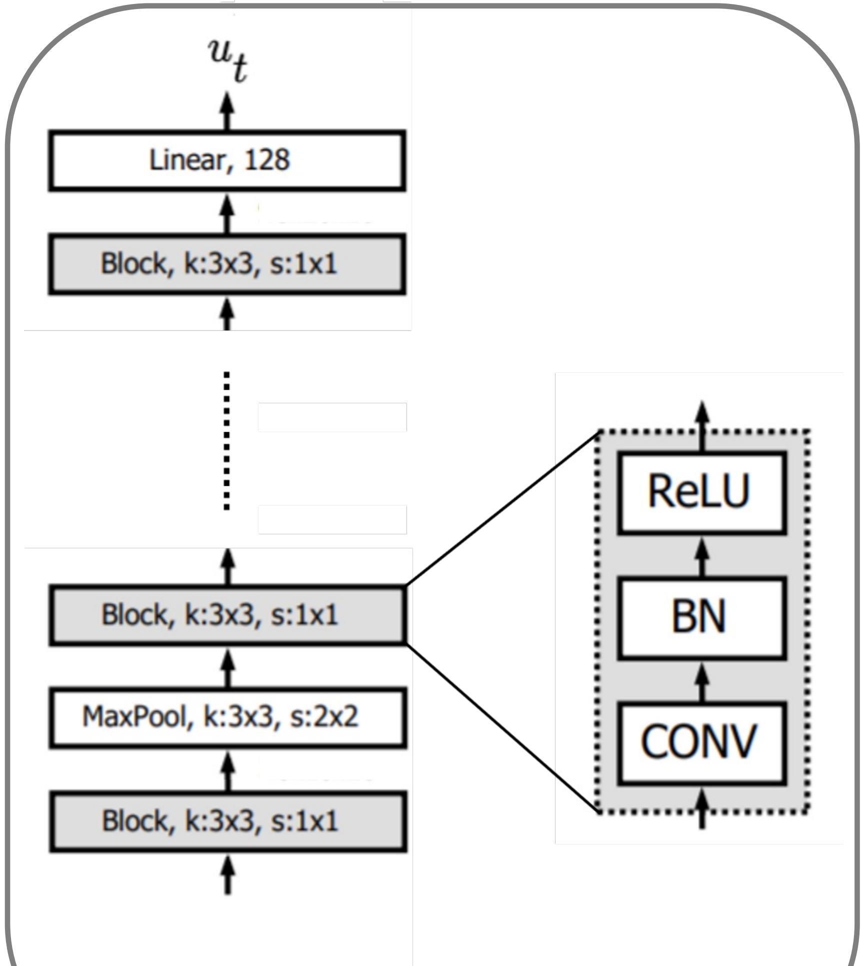
Pytorch / TensorFlow (TF)

Nvidia RTX 3090 GPU

Active Speaker Detection – A two-branch model…

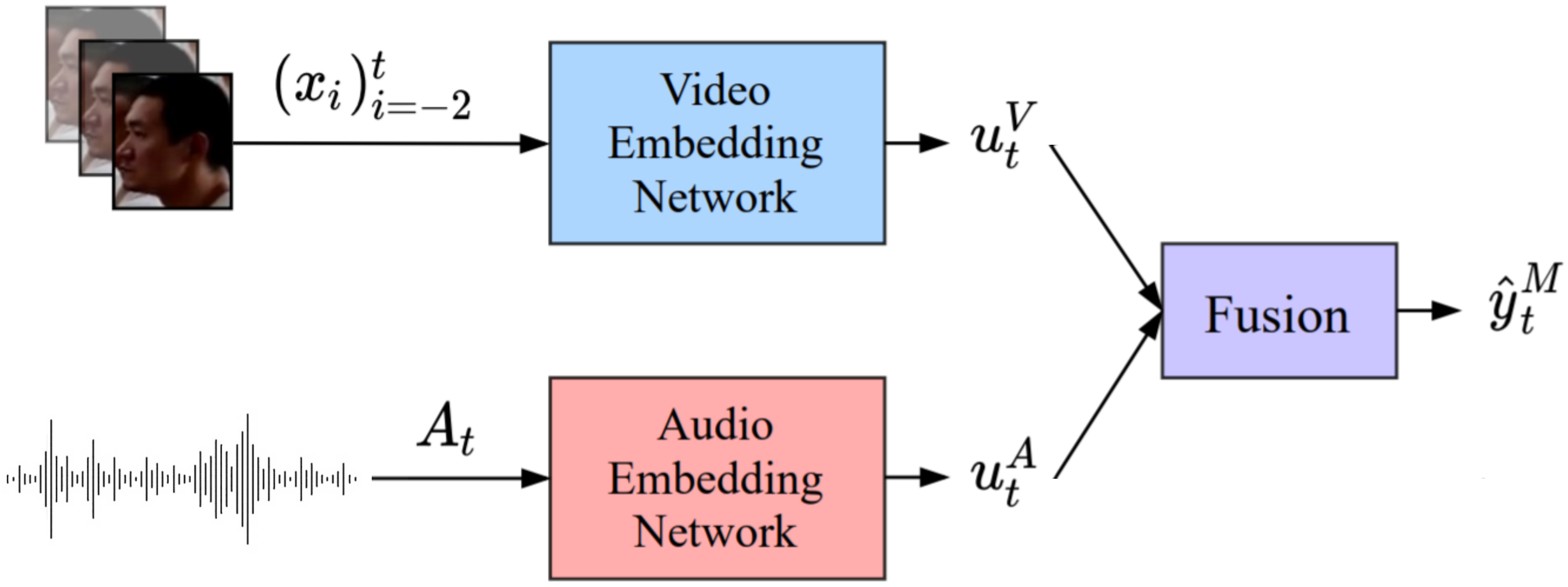


Audiovisual Neural Network

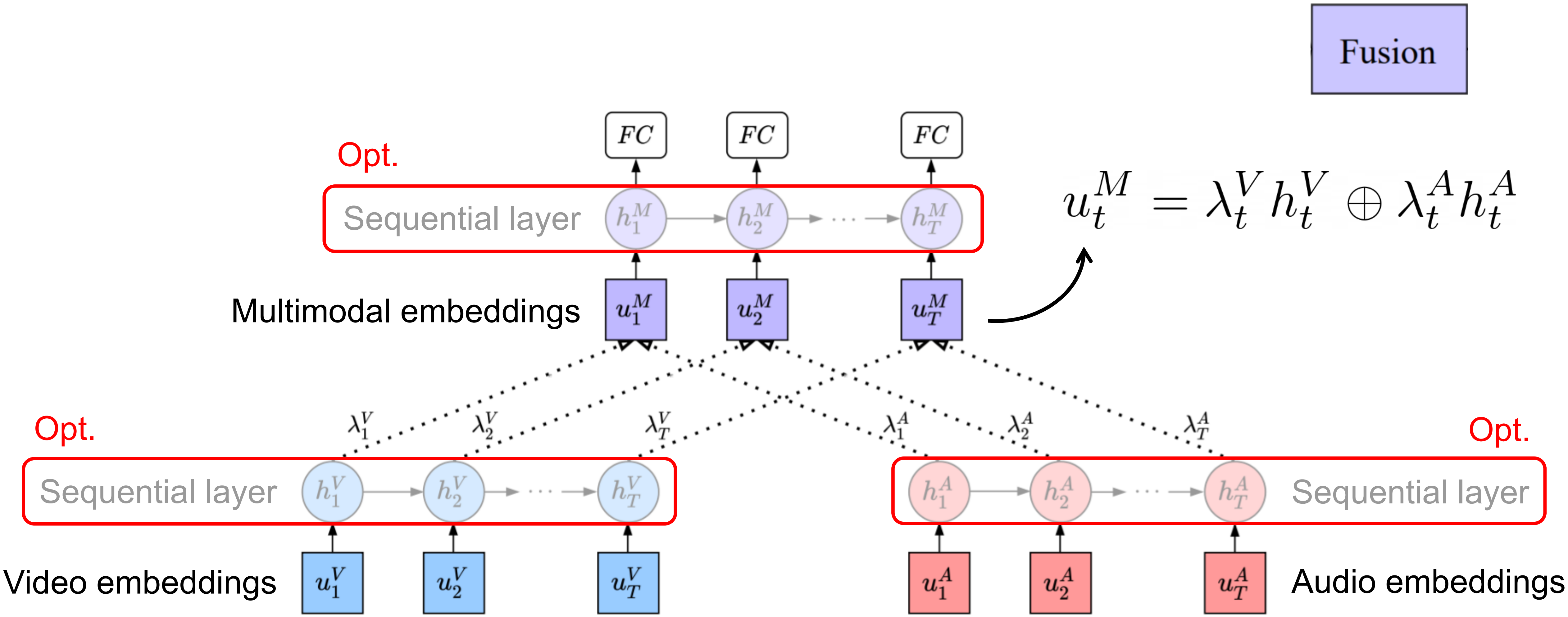


Stacked Convolutions

Active Speaker Detection – … and a fusion

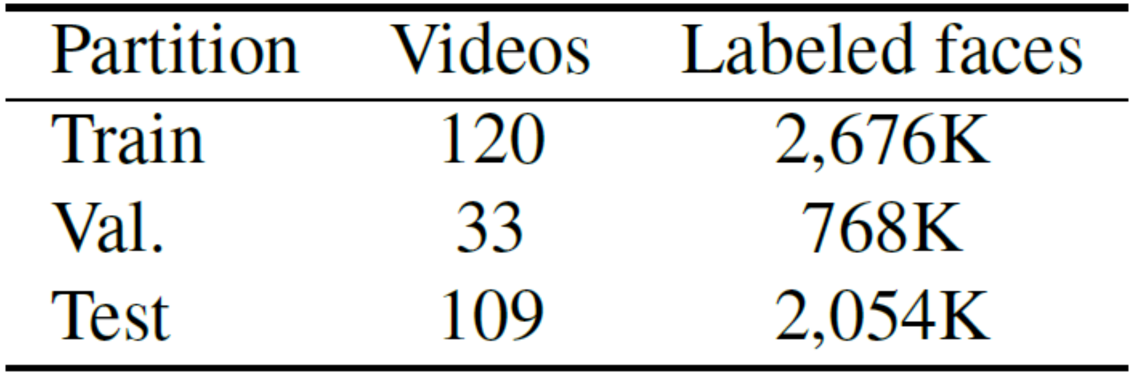


Active Speaker Detection – The fusion mechanism



$$u_t^M = \lambda_t^V h_t^V \oplus \lambda_t^A h_t^A$$

Active Speaker Detection – Training: dataset



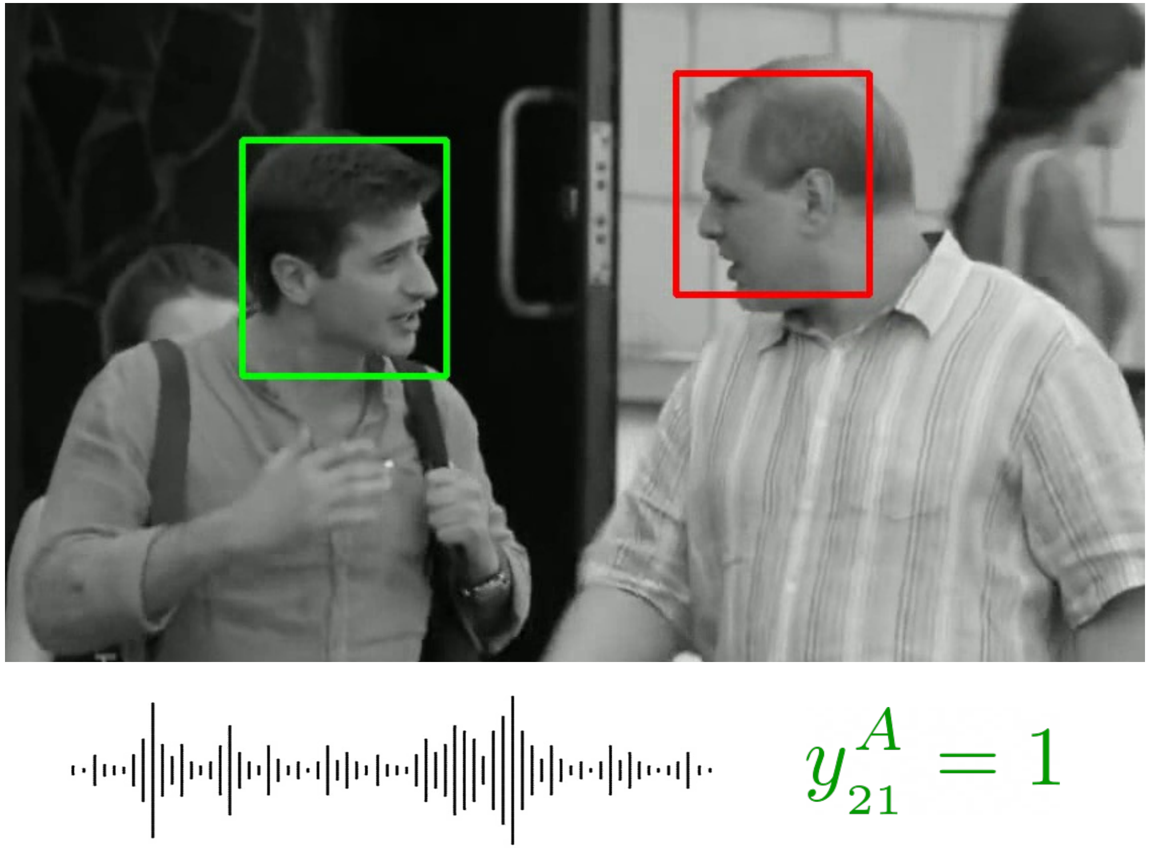
Partition	Videos	Labeled faces
Train	120	2,676K
Val.	33	768K
Test	109	2,054K

AVA-ActiveSpeaker dataset

Faces extracted from videos using available bounding boxes

Active Speaker Detection – Training: labels

$y_{21}^{0\,V} = 1$ $y_{21}^{1\,V} = 0$



$y_{21}^{A} = 1$

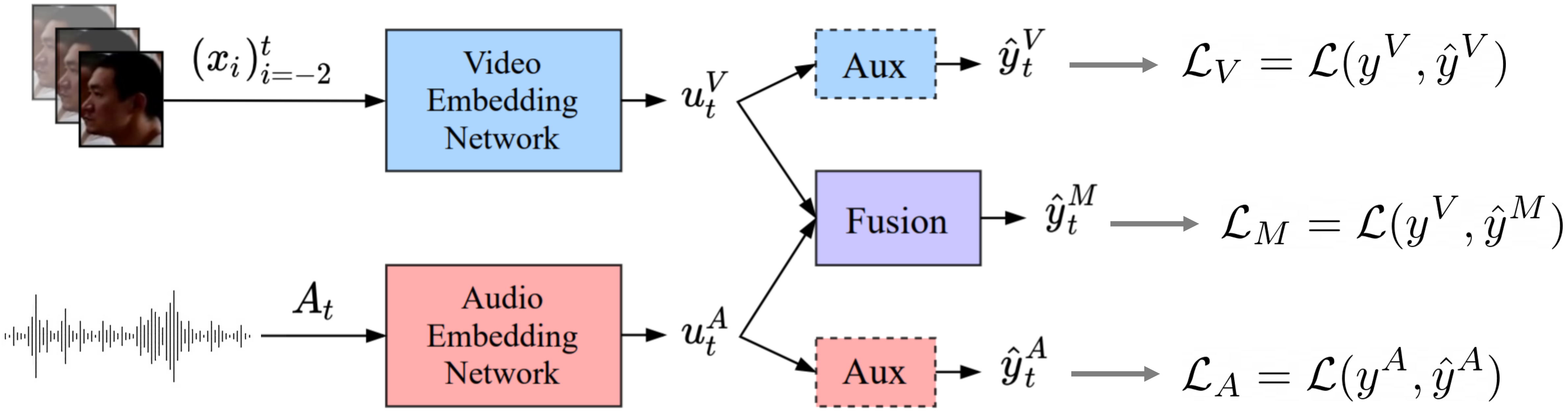
speaker ID

$$y_t^{n\,V} \in \{0, 1\}$$

video frame

$$y_t^A = \begin{cases} 0 & \text{if } \sum_n y_t^{nV} = 0, \\ 1 & \text{otherwise} \end{cases}$$

Active Speaker Detection – Training: multi-objective learning

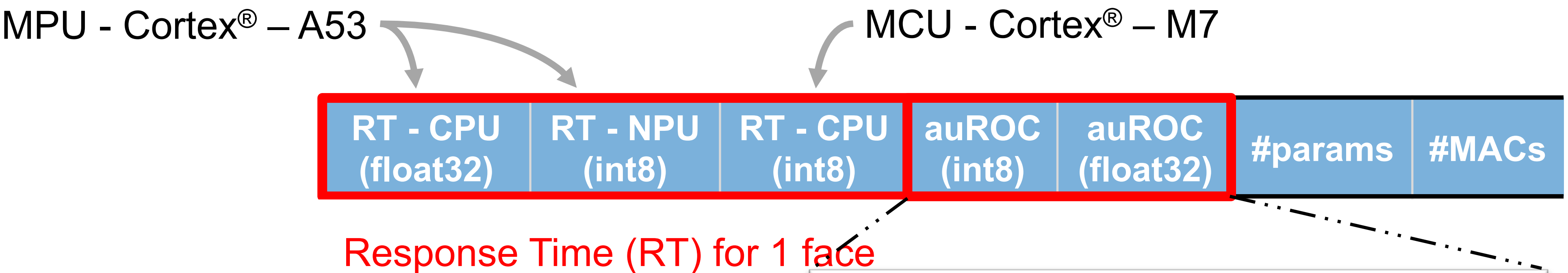


$(x_i)_{i=-2}^{t}$
Video Embedding Network
u_t^V
Aux
$\hat{y}_t^V$
$\mathcal{L}_V = \mathcal{L}(y^V, \hat{y}^V)$
Fusion
$\hat{y}_t^M$
$\mathcal{L}_M = \mathcal{L}(y^V, \hat{y}^M)$
A_t
Audio Embedding Network
u_t^A
Aux
$\hat{y}_t^A$
$\mathcal{L}_A = \mathcal{L}(y^A, \hat{y}^A)$

$$\mathcal{L}_f = \mathcal{L}_M + \mathcal{L}_V + \mathcal{L}_A$$

Neural Architecture Search (NAS) – Search space

- Definition and evaluation of a toolbox for model building

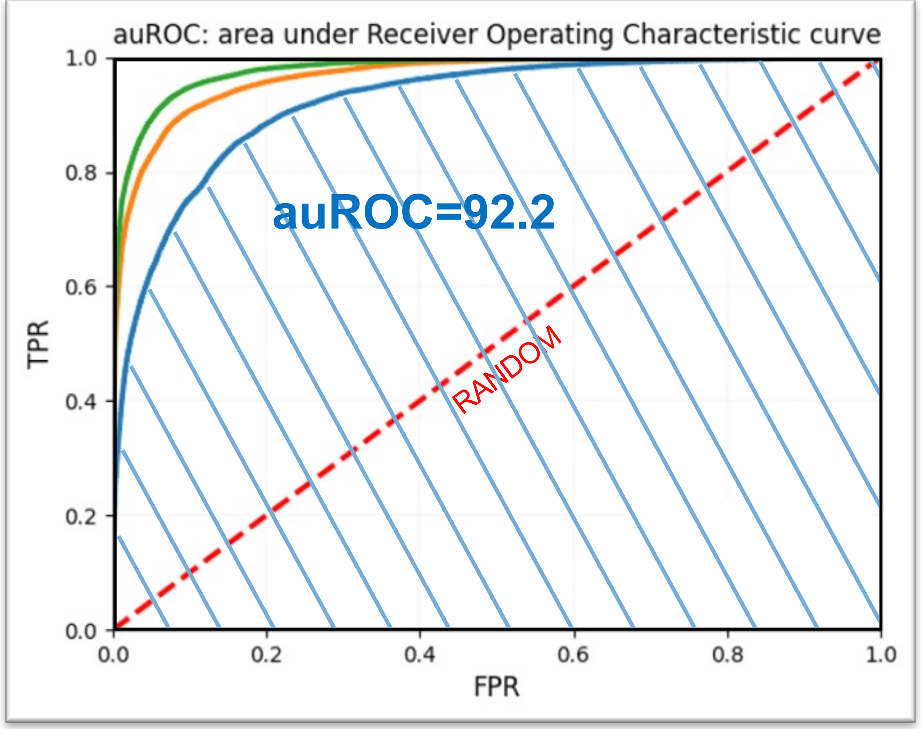


MPU - Cortex® – A53

MCU - Cortex® – M7

RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs

Response Time (RT) for 1 face



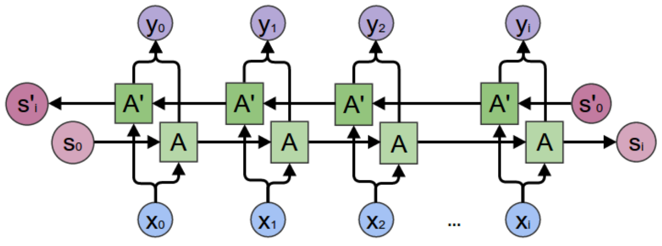
Neural Architecture Search (NAS) – Search space

MPU - Cortex® – A53

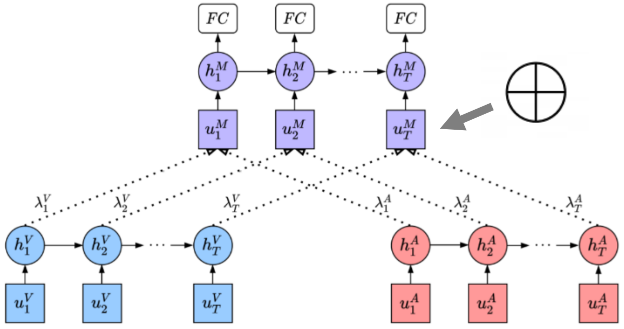
MCU - Cortex® – M7

RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs

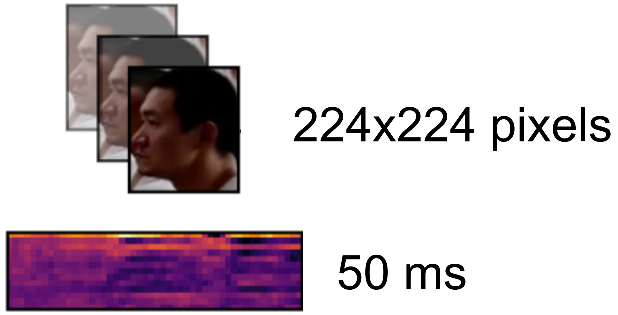
SOTA - *Pouthier et al., INTERSPEECH 2021*



Sequential layers: BiGRU



Fusion: Concatenation



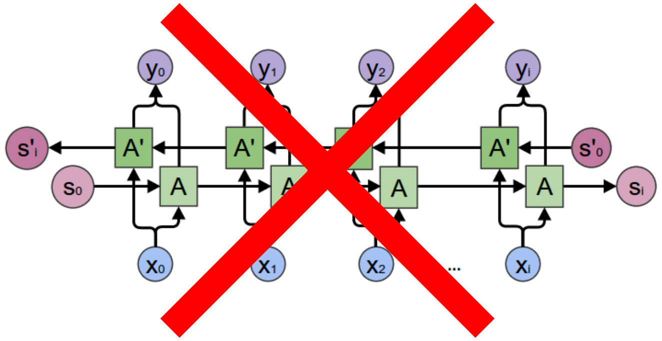
Inputs features: High resolution

Neural Architecture Search (NAS) – Search space

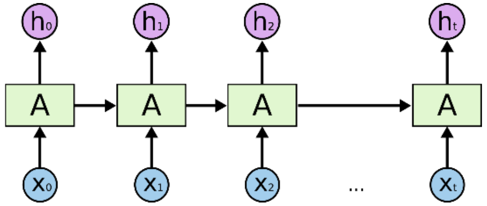
MPU - Cortex® – A53

MCU - Cortex® – M7

	RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs
SOTA - *Pouthier et al., INTERSPEECH 2021*	-	-	-	-	96.3	2.01 M	-
Sequential layers: BiGRU → GRU	39.7 ms	3.2 ms	319 ms	94.7	94.7	1.27 M	104.5 M



Sequential layers: BiGRU



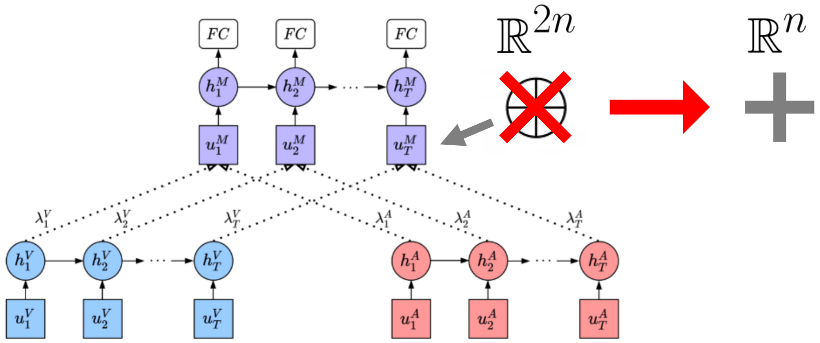
Sequential layers: GRU

Neural Architecture Search (NAS) – Search space

MPU - Cortex® – A53

MCU - Cortex® – M7

	RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs
SOTA - *Pouthier et al., INTERSPEECH 2021*	-	-	-	-	96.3	2.01 M	-
Sequential layers: BiGRU ⟶ GRU	39.7 ms	3.2 ms	319 ms	94.7	94.7	1.27 M	104.5 M
Fusion: Concat. ⟶ ADD	39.5 ms	3.2 ms	319 ms	94.8	94.9	1.22 M	104.5 M



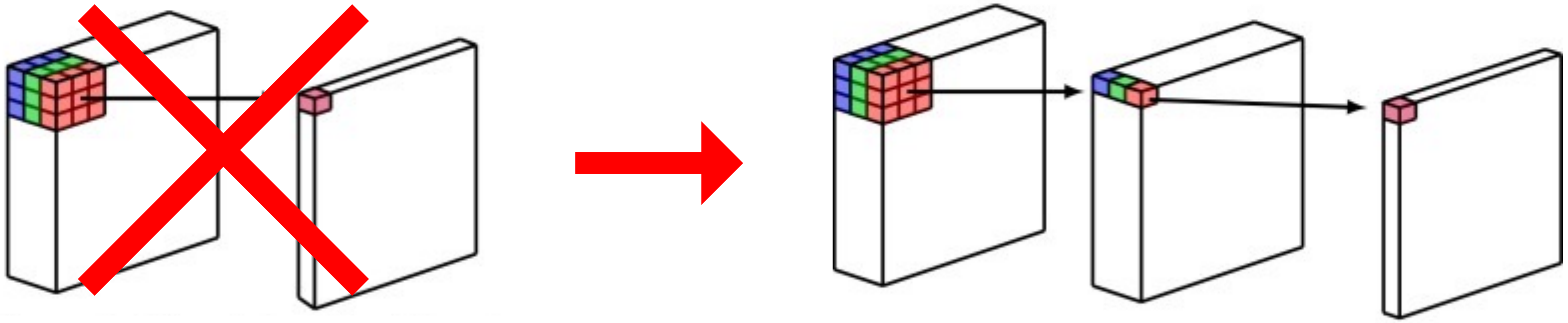
Fusion: ADD

Neural Architecture Search (NAS) – Search space

MPU - Cortex® – A53

MCU - Cortex® – M7

	RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs
SOTA - *Pouthier et al., INTERSPEECH 2021*	-	-	-	-	96.3	2.01 M	-
Sequential layers: BiGRU ⟶ GRU	39.7 ms	3.2 ms	319 ms	94.7	94.7	1.27 M	104.5 M
Fusion: Concat. ⟶ ADD	39.5 ms	3.2 ms	319 ms	94.8	94.9	1.22 M	104.5 M
Convolutions: Standard ⟶ DW-S	18.3 ms	3.2 ms	98 ms	94.1	94.4	0.66 M	8.7 M



Convolutions: Standard

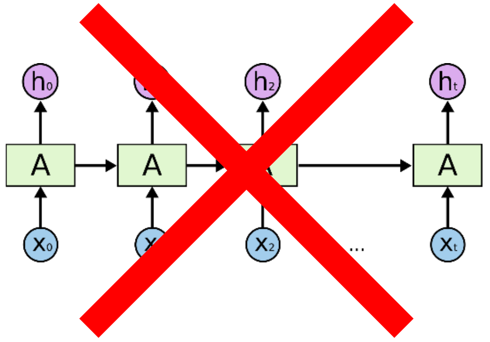
Convolutions: DW-S

Neural Architecture Search (NAS) – Search space

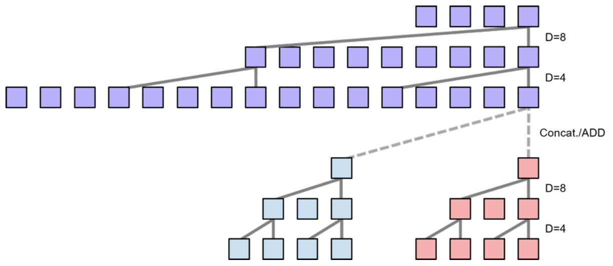
MPU - Cortex® – A53

MCU - Cortex® – M7

	RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs
SOTA - *Pouthier et al., INTERSPEECH 2021*	-	-	-	-	96.3	2.01 M	-
Sequential layers: BiGRU ⟶ GRU	39.7 ms	3.2 ms	319 ms	94.7	94.7	1.27 M	104.5 M
Fusion: Concat. ⟶ ADD	39.5 ms	3.2 ms	319 ms	94.8	94.9	1.22 M	104.5 M
Convolutions: Standard ⟶ DW-S	18.3 ms	3.2 ms	98 ms	94.1	94.4	0.66 M	8.7 M
Sequential layers: GRU ⟶ TCN	17.7 ms	2.7 ms	94 ms	93.2	93.6	0.27 M	8.3 M



Sequential layers: GRU



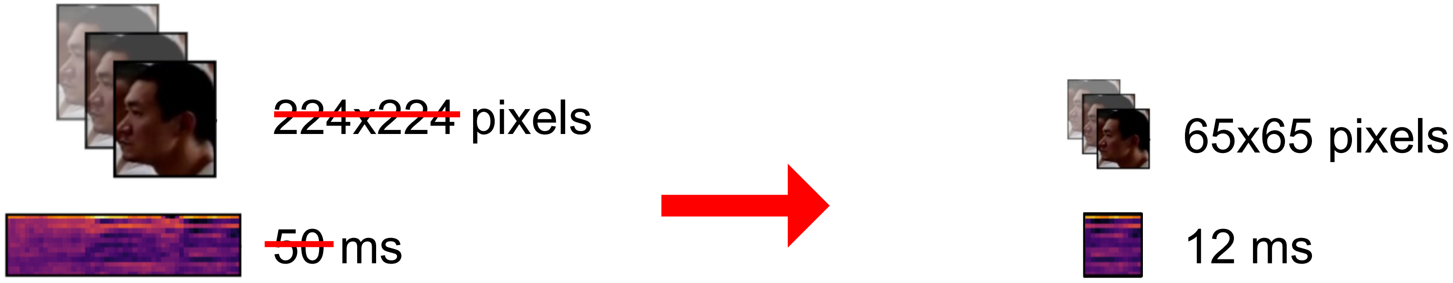
Sequential layers: TCN

Neural Architecture Search (NAS) – Search space

MPU - Cortex® – A53

MCU - Cortex® – M7

	RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs
SOTA - *Pouthier et al., INTERSPEECH 2021*	-	-	-	-	96.3	2.01 M	-
Sequential layers: BiGRU ⟶ GRU	39.7 ms	3.2 ms	319 ms	94.7	94.7	1.27 M	104.5 M
Fusion: Concat. ⟶ ADD	39.5 ms	3.2 ms	319 ms	94.8	94.9	1.22 M	104.5 M
Convolutions: Standard ⟶ DW-S	18.3 ms	3.2 ms	98 ms	94.1	94.4	0.66 M	8.7 M
Sequential layers: GRU ⟶ TCN	17.7 ms	2.7 ms	94 ms	93.2	93.6	0.27 M	8.3 M
Inputs features: High res. → Low res.	2.4 ms	0.7 ms	18 ms	91.4	92.5	0.26 M	3.5 M



~~224x224~~ pixels

~~50~~ ms

65x65 pixels

12 ms

Inputs features: High resolution

Inputs features: Low resolution

Neural Architecture Search (NAS) – Search space

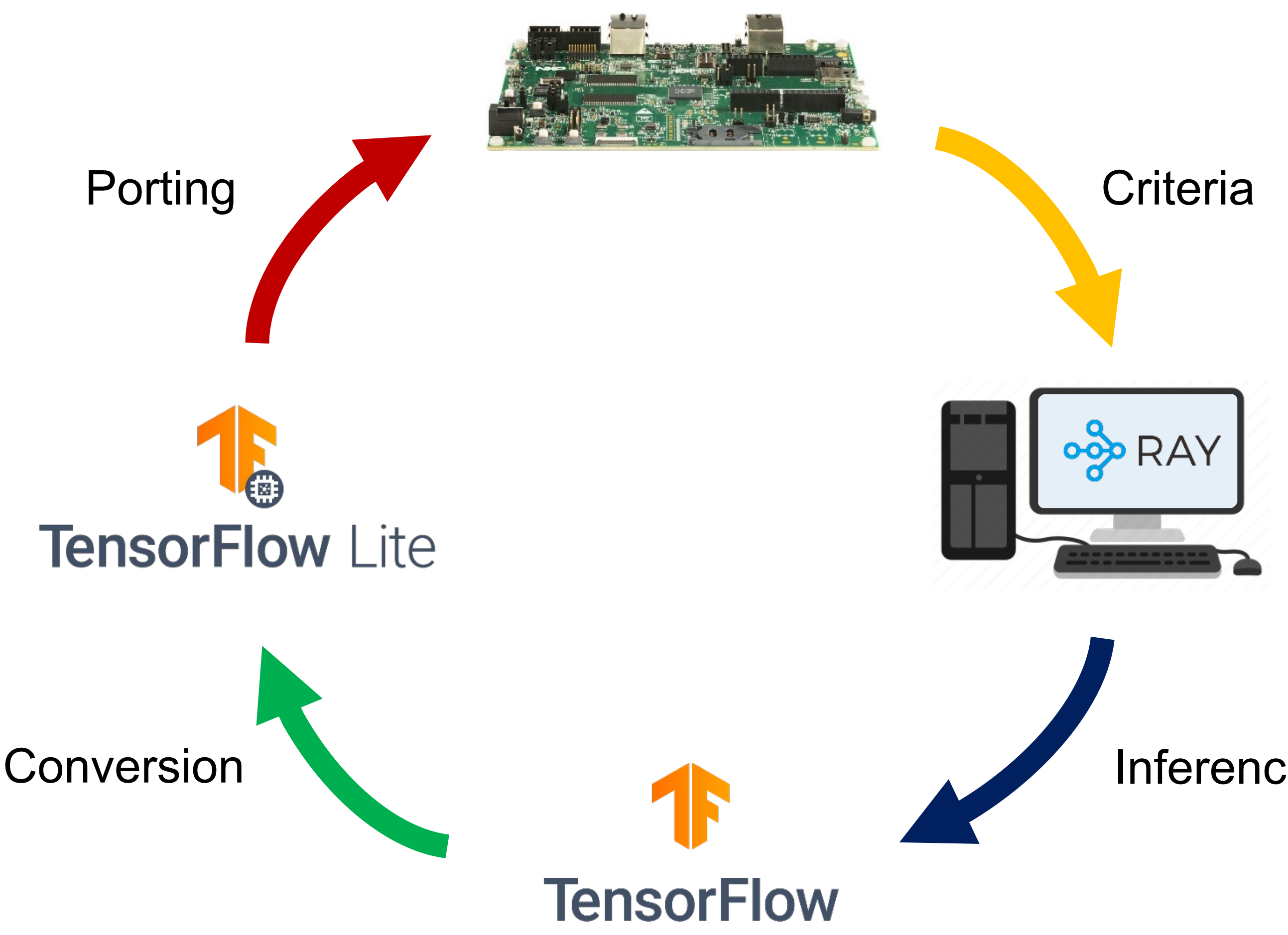
MPU - Cortex® – A53

MCU - Cortex® – M7

	RT - CPU (float32)	RT - NPU (int8)	RT - CPU (int8)	auROC (int8)	auROC (float32)	#params	#MACs
SOTA - *Pouthier et al., INTERSPEECH 2021*	-	-	-	-	96.3	2.01 M	-
Sequential layers: BiGRU ⟶ GRU	39.7 ms	3.2 ms	319 ms	94.7	94.7	1.27 M	104.5 M
Fusion: Concat. ⟶ ADD	39.5 ms	3.2 ms	319 ms	94.8	94.9	1.22 M	104.5 M
Convolutions: Standard ⟶ DW-S	18.3 ms	3.2 ms	98 ms	94.1	94.4	0.66 M	8.7 M
Sequential layers: GRU ⟶ TCN	17.7 ms	2.7 ms	94 ms	93.2	93.6	0.27 M	8.3 M
Inputs features: High res. → Low res.	2.4 ms	0.7 ms	18 ms	91.4	92.5	0.26 M	3.5 M

- ➢ Low resolution inputs features and DW-S convolutions are required
- ➢ What is the best model configuration?

Neural Architecture Search (NAS) – Search algorithm



On-device benchmarking

Criteria

Random search
ASHA scheduler

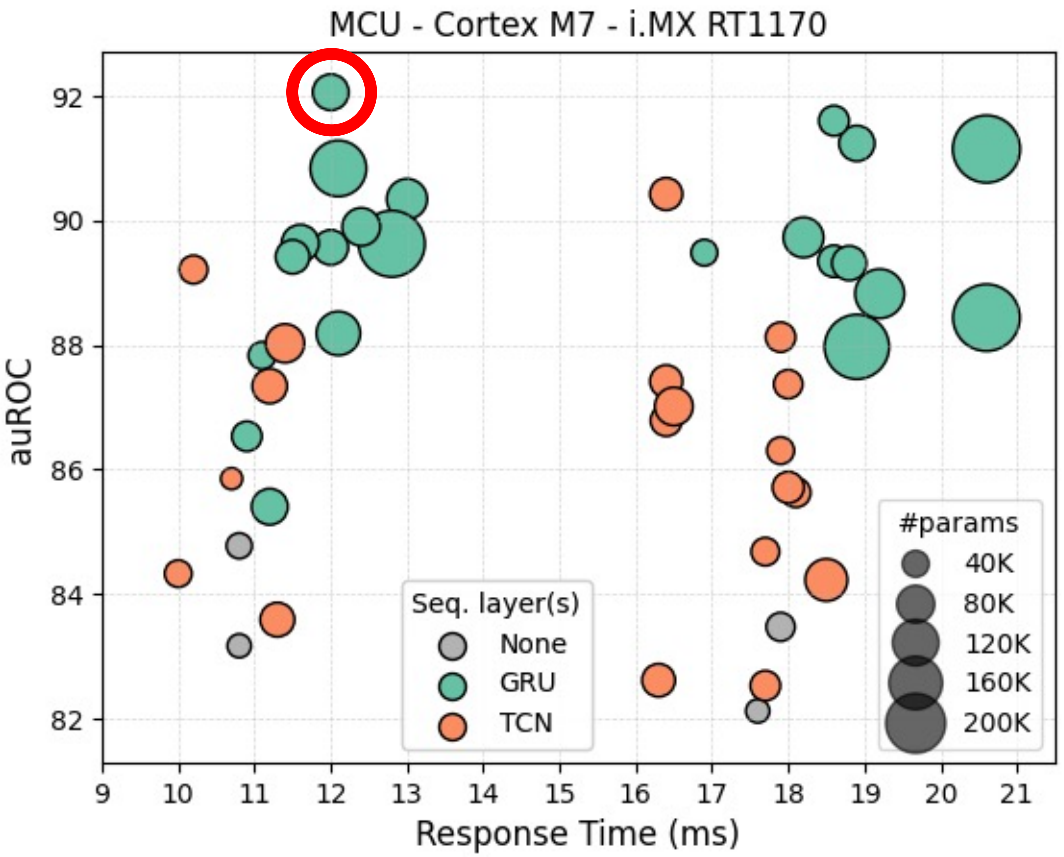
Inference model

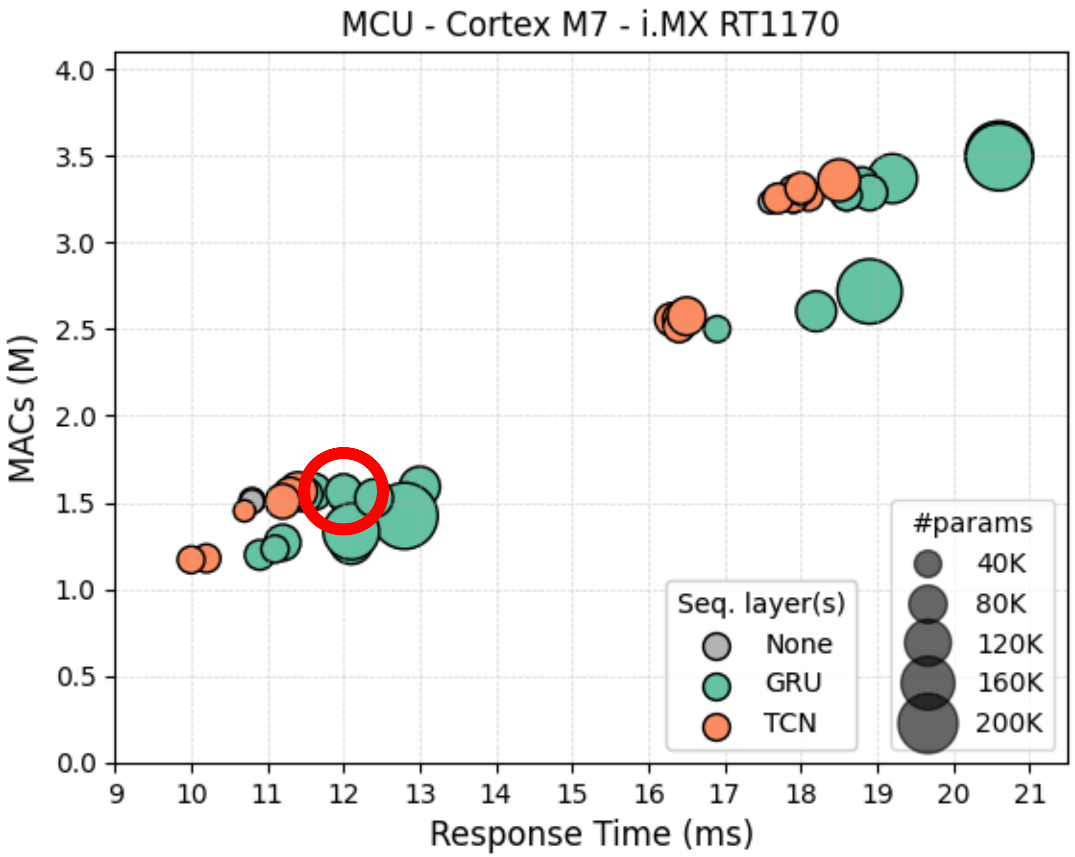
Conversion

Post-Training
Quantization

Porting

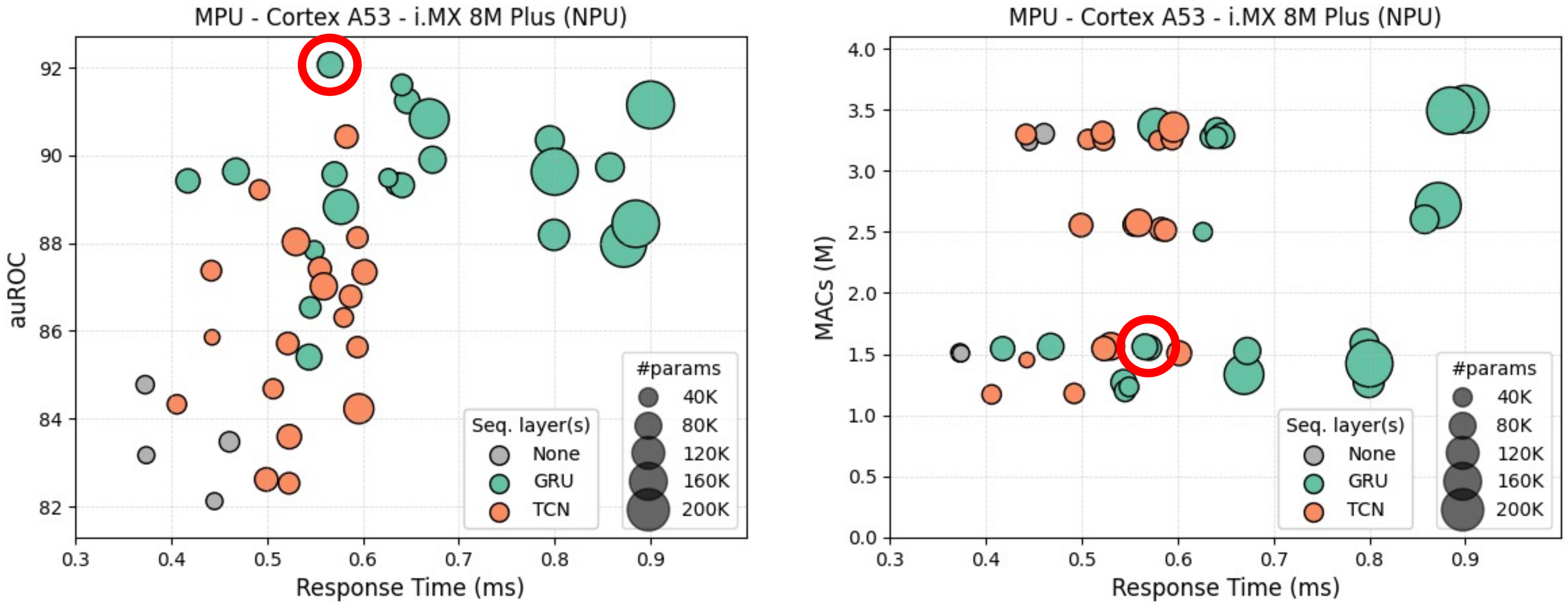
Results – NAS: MCU (int8)





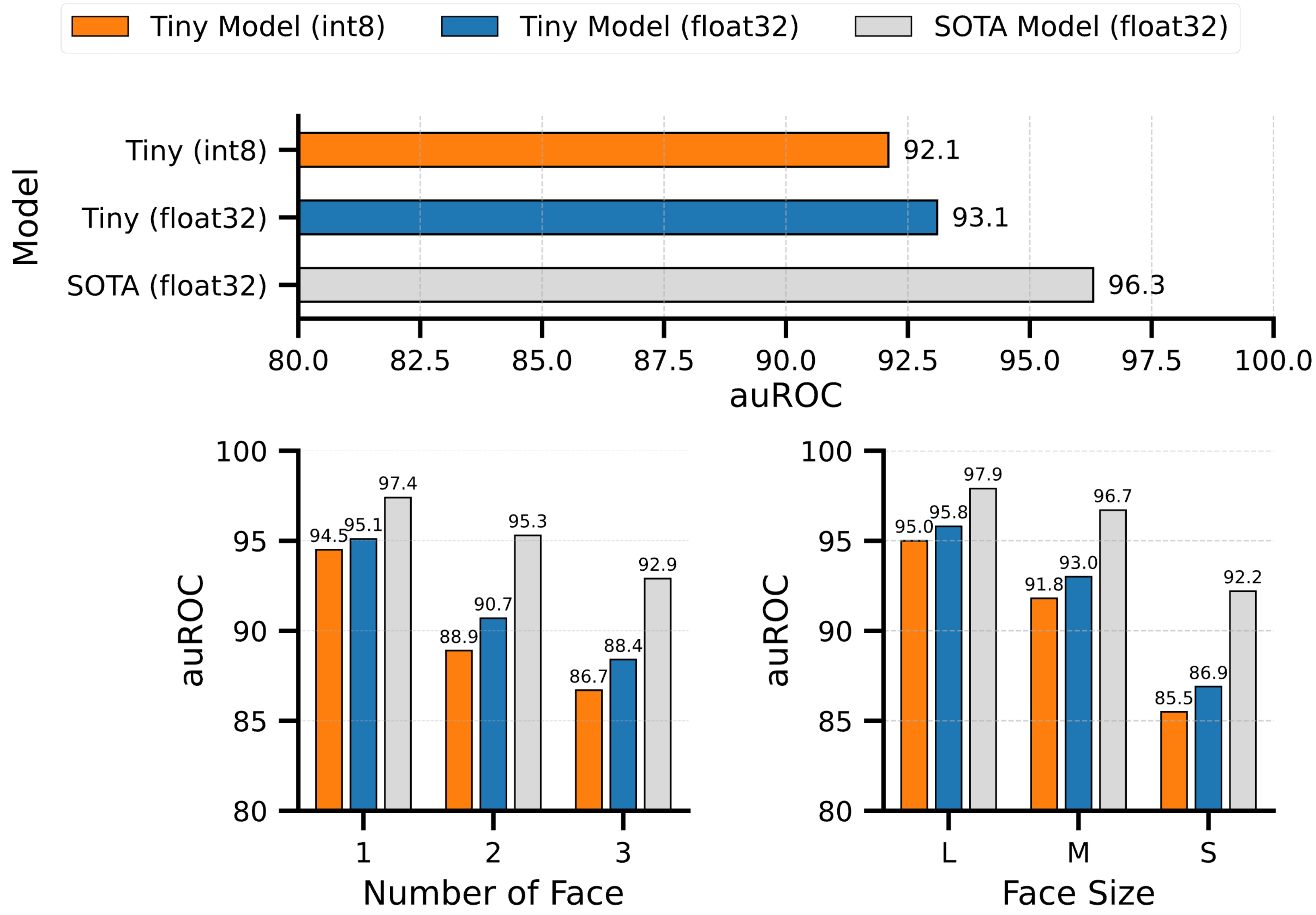
○ Chosen “Tiny” model

Results – NAS: MPU with NPU (int8)

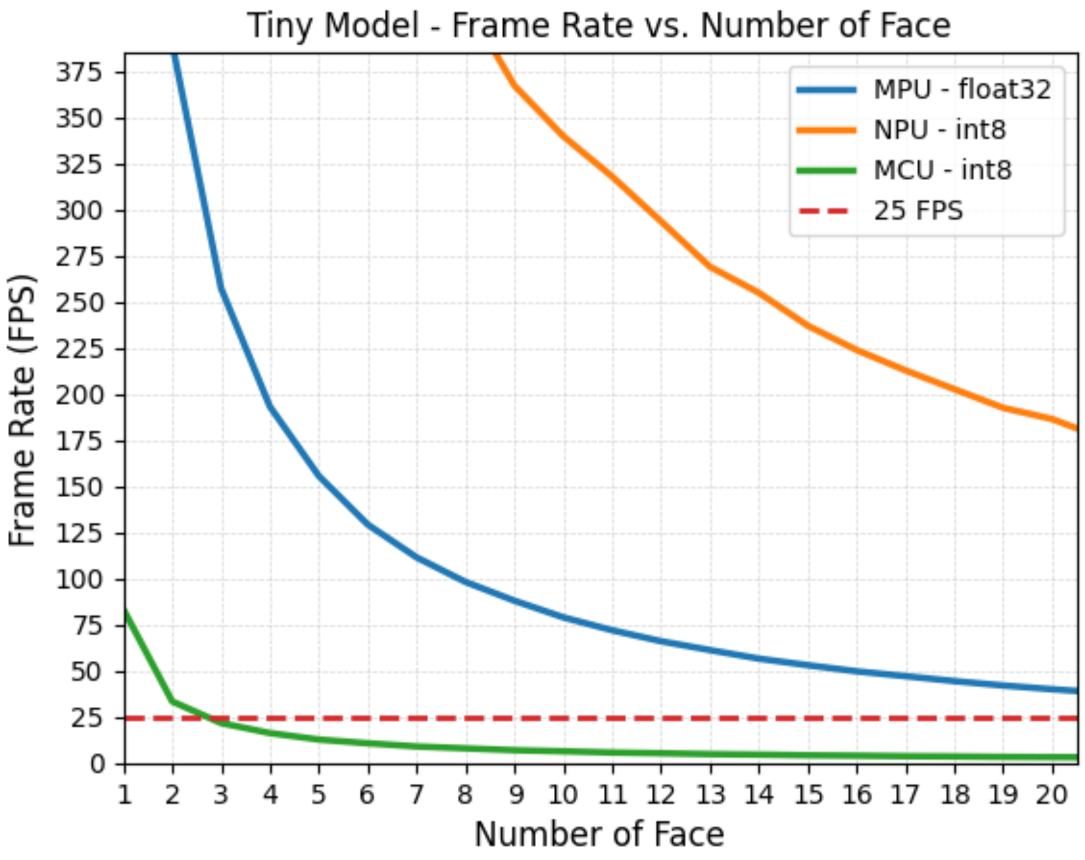


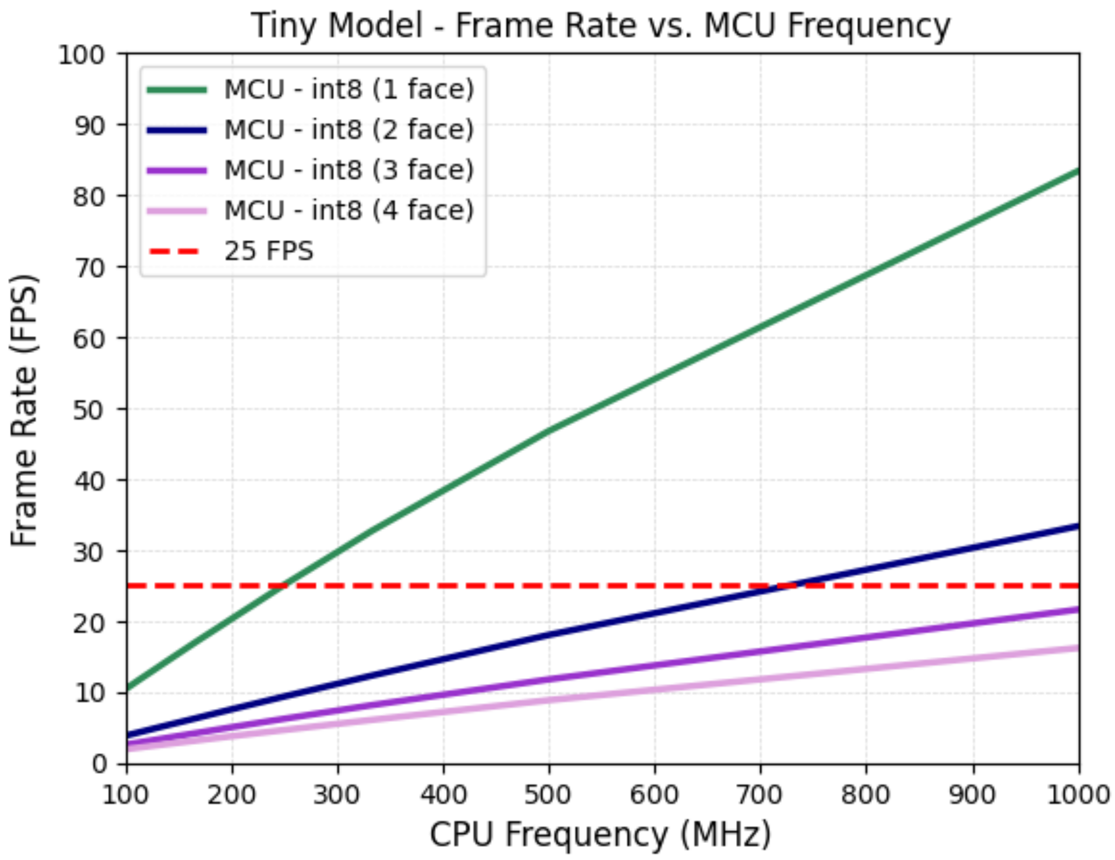
○ Chosen “Tiny” model

Results – Tiny model performance analysis



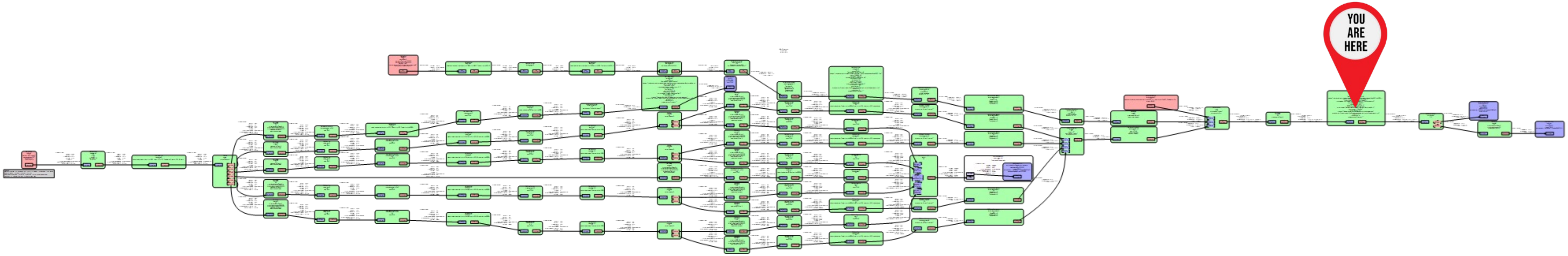
Results – Tiny model performance analysis





Conclusion

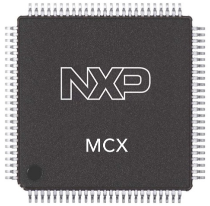
- MPU – Cortex® A53 – i.MX 8M Plus → Fully adapted to the requirements



Integration pipeline on i.MX 8M Plus

- MCU – Cortex® M7 – i.MX RT1170 → Allows limited use of the algorithm

Next: Targeting MCX N Serie – Cortex® M33 (MCU with NPU)



Learn More

- AI and Machine Learning at NXP Semiconductors (www.nxp.com/ai)
- eIQ® ML Software Development Environment (www.nxp.com/eiq)
- eIQ® ML/AI Training Series (www.nxp.com/mltraining)
- eIQ® Neutron Neural Processing Unit (NPU) (www.nxp.com/neutron)

eIQ® ML SW Development Environment

eIQ Toolkit with eIQ Portal GUI to:

- Import/create, convert, optimize , validate and deploy ML models
- Dataset curation tools to create new, augment, label/annotate datasets

eIQ inference with:

- TensorFlow Lite, TensorFlow Lite Micro and DeepViewRT

eIQ Marketplace:

- Add-on wares available from eco-system partners and NXP for ML applications, optimized models, optimization tools, datasets and sensor solutions

Support for i.MX 8M, i.MX 9, i.MX RT, MCX family of devices

Integrated with NXP dev environments (MCUXpresso, Yocto/Linux)

NXP eIQ® Neutron NPU

- Highly scalable ML acceleration cores
- Unified architecture and software support
- Optimized for edge performance and power dissipation

Turnkey Solutions

Smart HMI solution

- i.MX RT117H (kit - SLN-TLHMI-IOT-RD)

Face & emotion recognition solution with Anti-Spoofing

- i.MX RT106F (kit – SLN-VIZN-IOT)

Local voice control solution

- i.MX RT106L (kit – SLN-LOCAL-IOT)

NXP

SECURE CONNECTIONS
FOR A SMARTER WORLD

Copyright Notice

TINY ML EMEA
AMSTERDAM, NL
June 26-28, 2023

www.tinyml.org