Attention: *The .pdf version of these slides are sadly not as expressive as the original .pptx file which truly comes to life through the animations. If you wish to look at the .pptx version, send an email to jann.krausse@infineon.com or message me on LinkedIn. Cheers ☺*

Hybrid Spiking Neural Networks for Neural Decoding of Cortical Activity

Jann Krausse (Infineon Technologies, Dresden),
Alexandru Vasilache (FZI Research Center for Information Technology, Karlsruhe)
Co-Authors: Klaus Knobloch, Jürgen Becker

Introduction – Intra-Cortical Brain Machine Interfaces (iBMI)

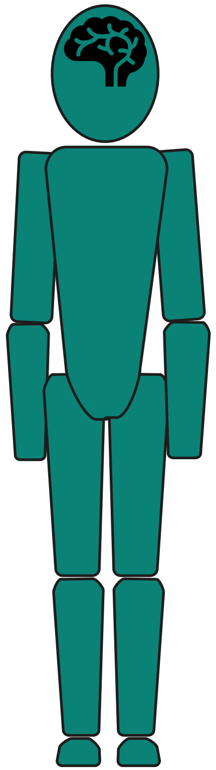
Paralysis affects millions of patiens worldwide

→ *the inability to move some or all of your body*

Introduction – Intra-Cortical Brain Machine Interfaces (iBMI)

Paralysis affects millions of patiens worldwide

→ *the inability to move some or all of your body*

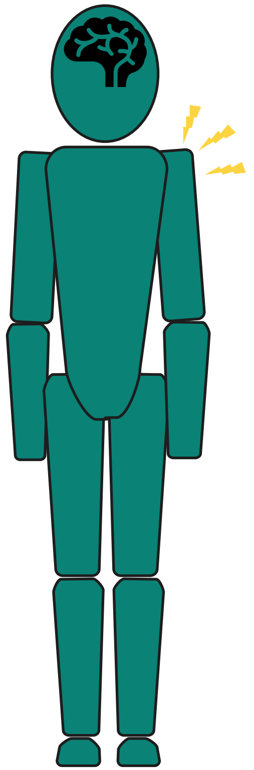


Patient X

Introduction – Intra-Cortical Brain Machine Interfaces (iBMI)

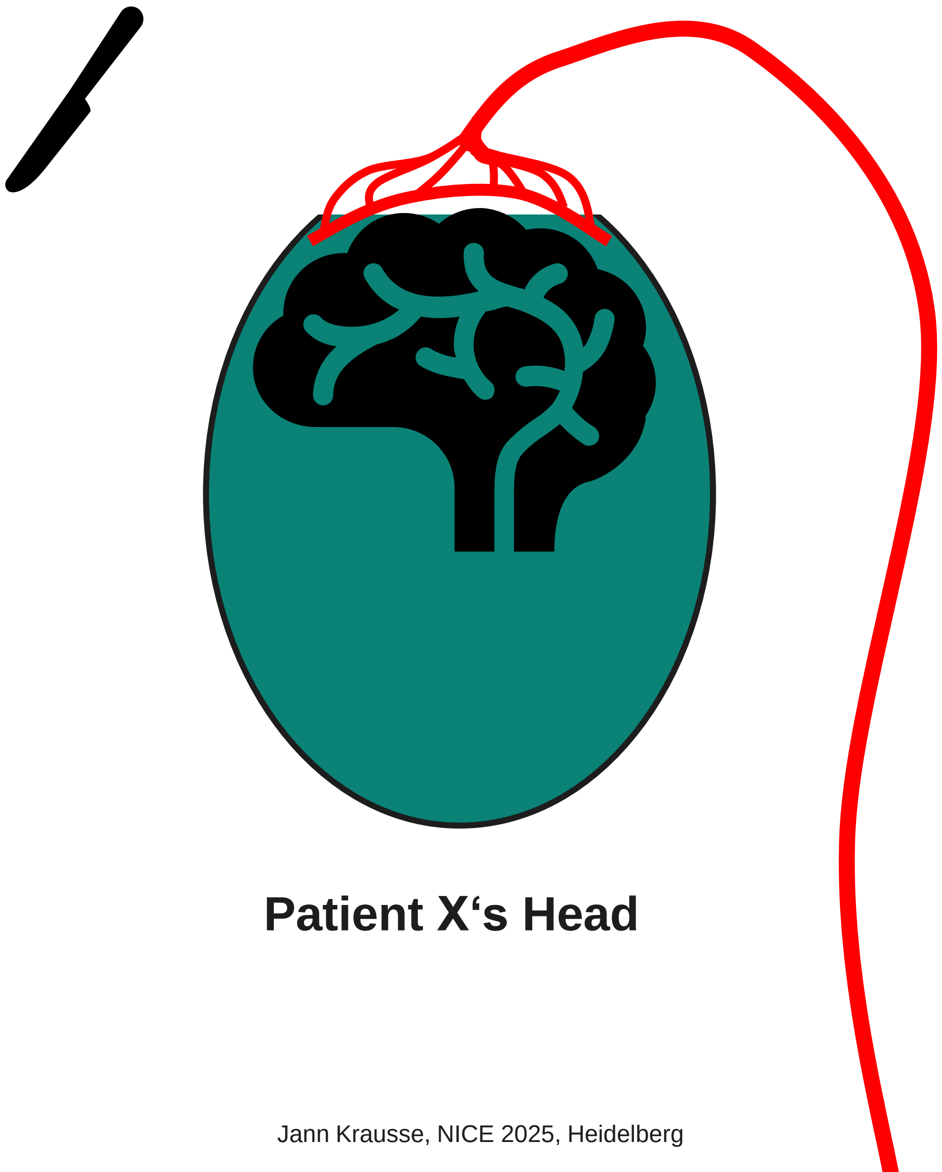
Paralysis affects millions of patiens worldwide

→ *the inability to move some or all of your body*



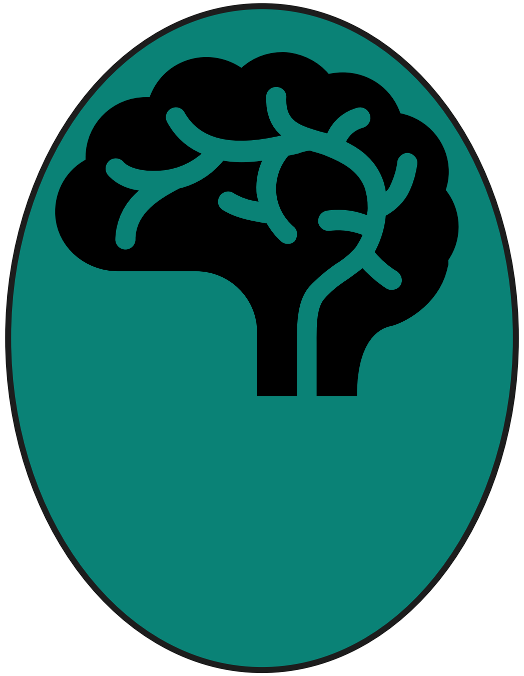
Patient X

Introduction – Intra-Cortical Brain Machine Interfaces (iBMI)



Patient X‘s Head

Introduction – Intra-Cortical Brain Machine Interfaces (iBMI)



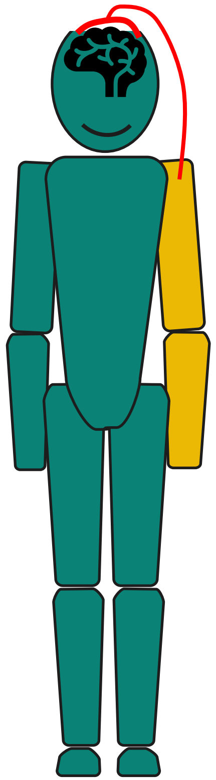
Patient X's Head

Introduction – Intra-Cortical Brain Machine Interfaces (iBMI)

Paralysis affects millions of patiens worldwide

→ *the inability to move some or all of your body*

iBMIs can translate cortical activity and control prostheses

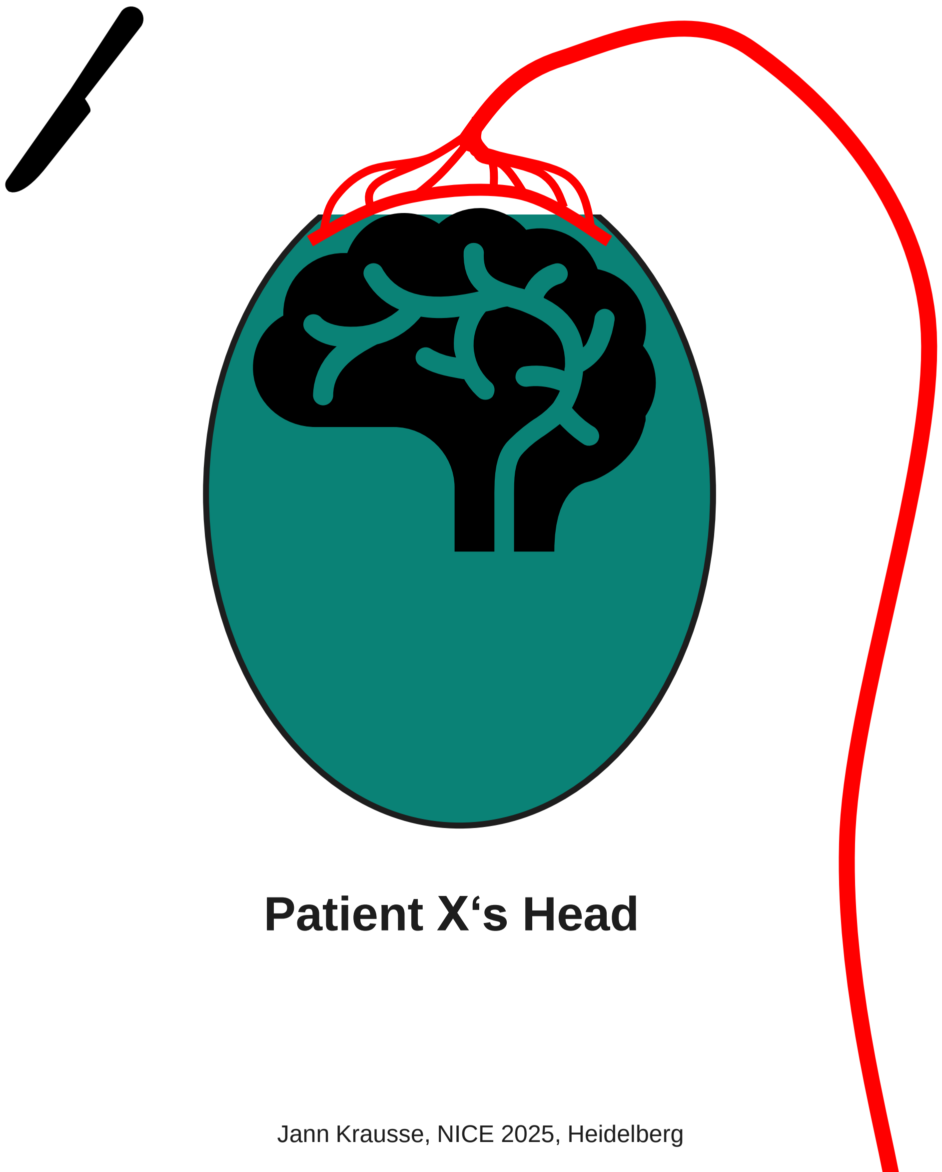


Patient X

Introduction – The Problems with Current iBMIs

Problem 1: Skull Opening

Increases the risk of infection



Problem 2: Bulky Wiring

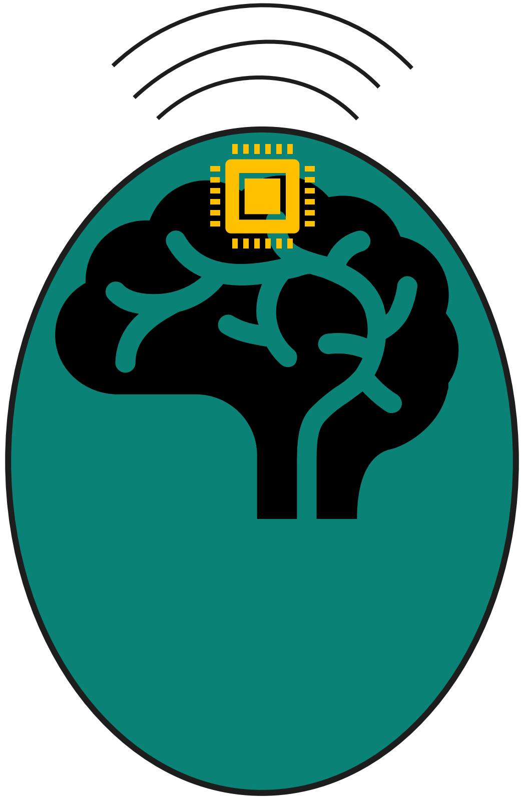
Impairs head movement

Introduction – The Problems with Current iBMIs

Problem 1: Skull Opening
Increases the risk of infection

Problem 2: Bulky Wiring
Impairs head movement

Possible Solution:
Wireless iBMIs



Patient X's Head

Introduction – The Problems with Wireless iBMIs

Possible Solution:

Wireless iBMIs

Demand minimal heat dissipation

Have restricted battery lifetime

Introduction – The Problems with Wireless iBMIs

Possible Solution:
Wireless iBMIs

Demand minimal heat dissipation

Have restricted battery lifetime

↓

Limited bandwidth

↓

Requires **high-quality compression** & **high energy efficiency**

Introduction – The Problems with Wireless iBMIs

Possible Solution:
Wireless iBMIs

Requires **high-quality compression** & **high energy efficiency**

NNs

S

Promising candidate for such neural decoders?

!

Introduction – The Problems with Wireless iBMIs

Possible Solution:
Wireless iBMIs

Requires **high-quality compression** & **high energy efficiency**

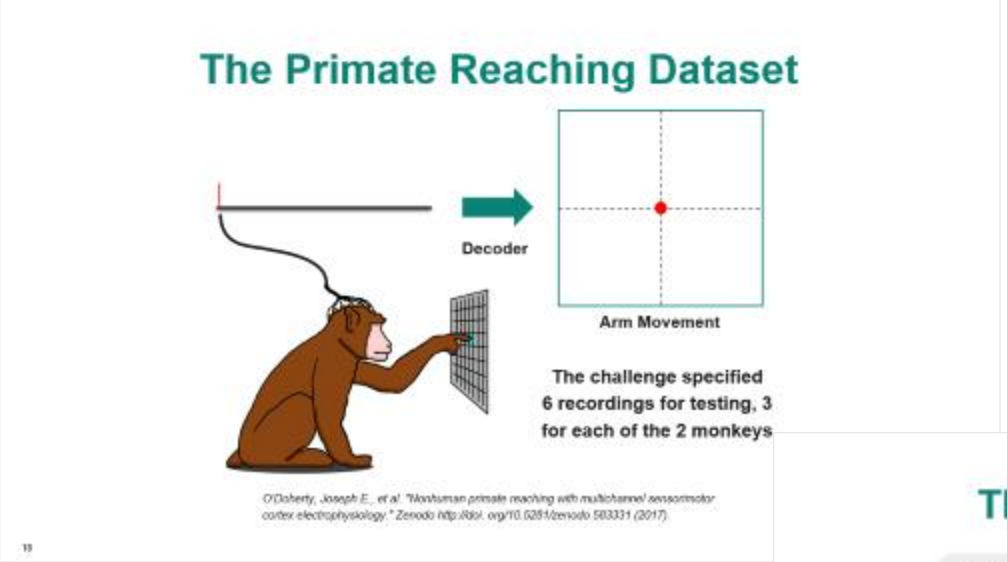
Promising candidate for such neural decoders? ***Neuromorphic Technologies*** **!**

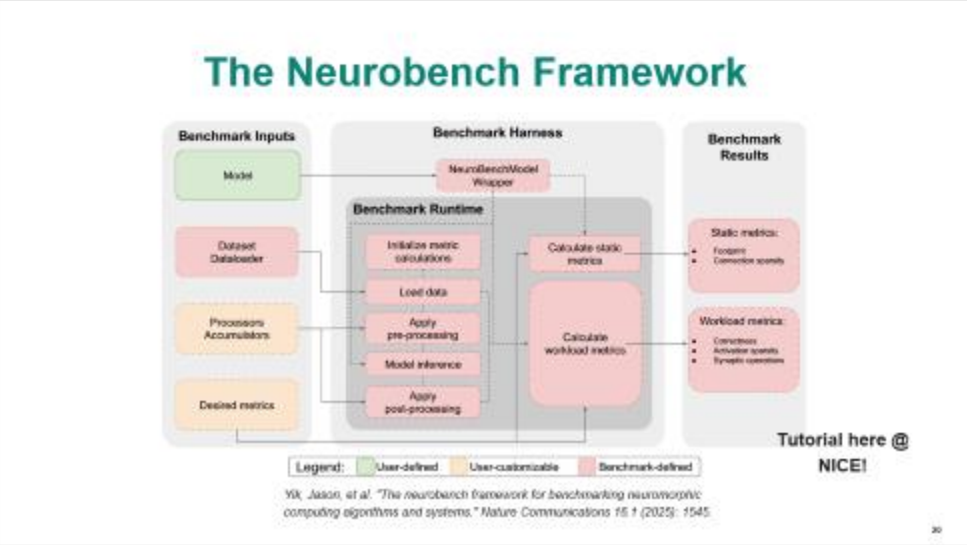
Background – BioCAS'24 Neural Decoding Challenge

– Collaborative effort by *City University of Hong Kong*, *Harvard University*, and *TU Delft*

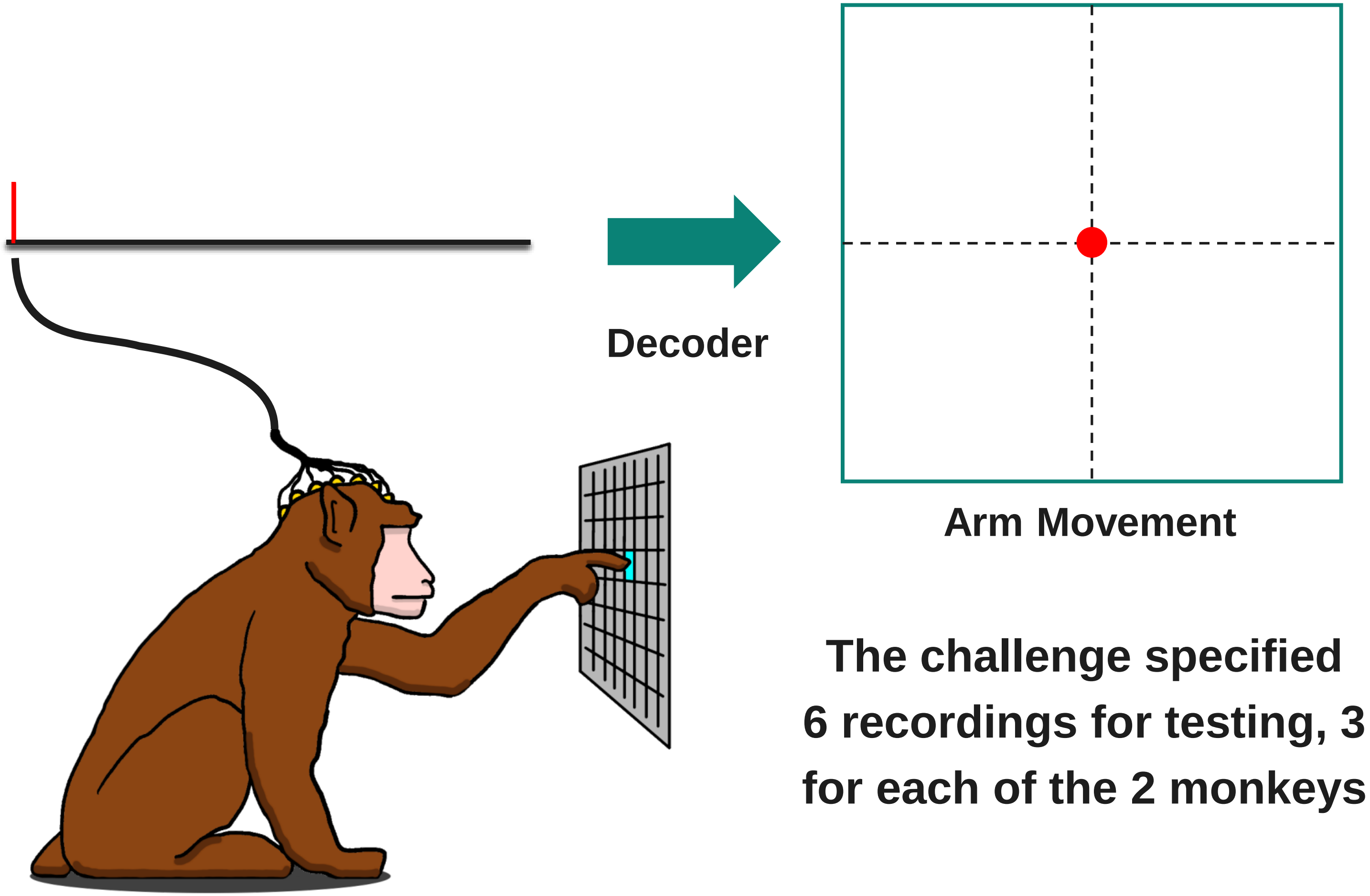
Background – BioCAS’24 Neural Decoding Challenge

– Collaborative effort by *City University of Hong Kong*, *Harvard University*, and *TU Delft*





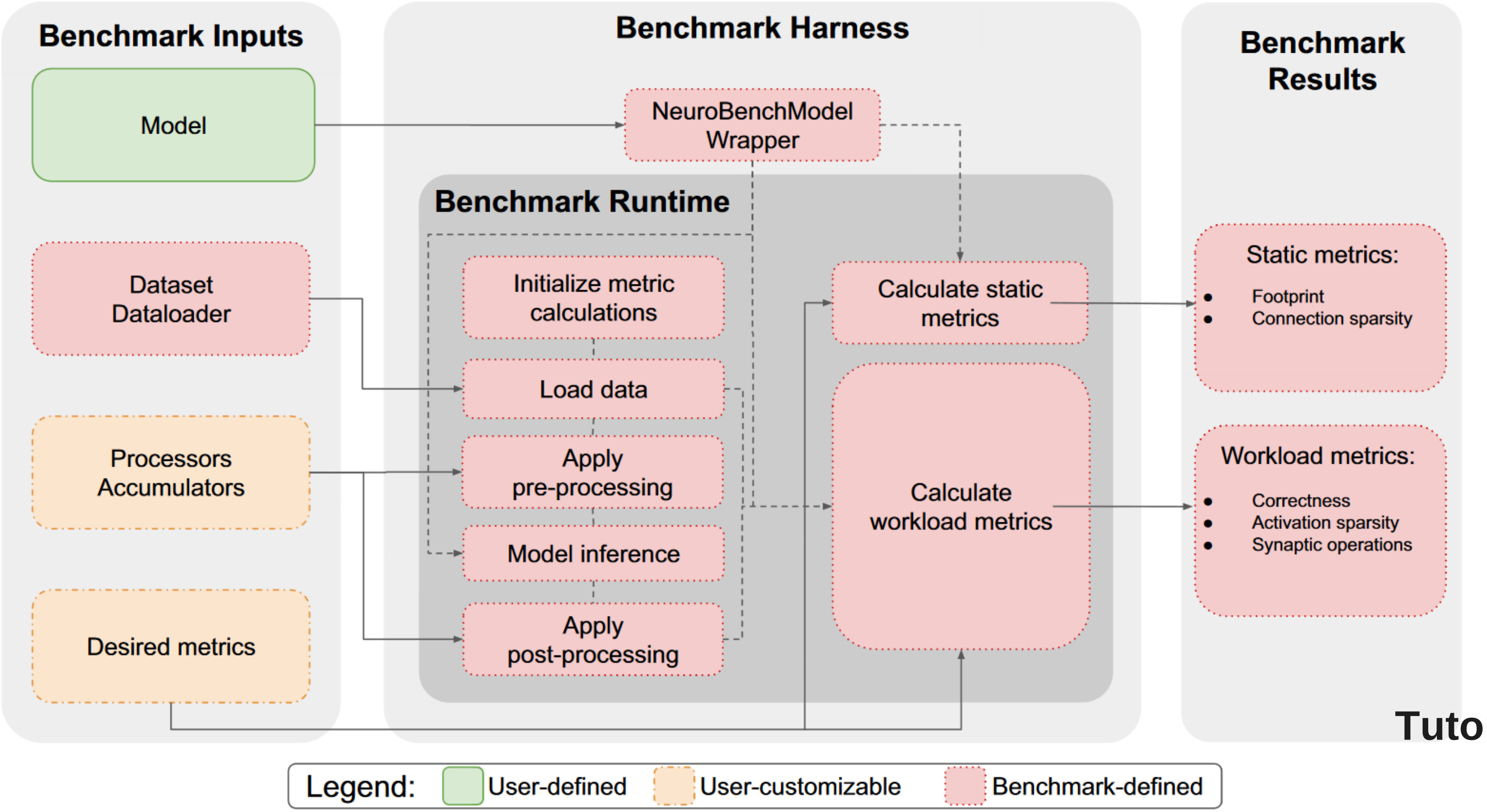
The Primate Reaching Dataset



The challenge specified 6 recordings for testing, 3 for each of the 2 monkeys

O'Doherty, Joseph E., et al. "Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology." Zenodo http://doi. org/10.5281/zenodo 583331 (2017).

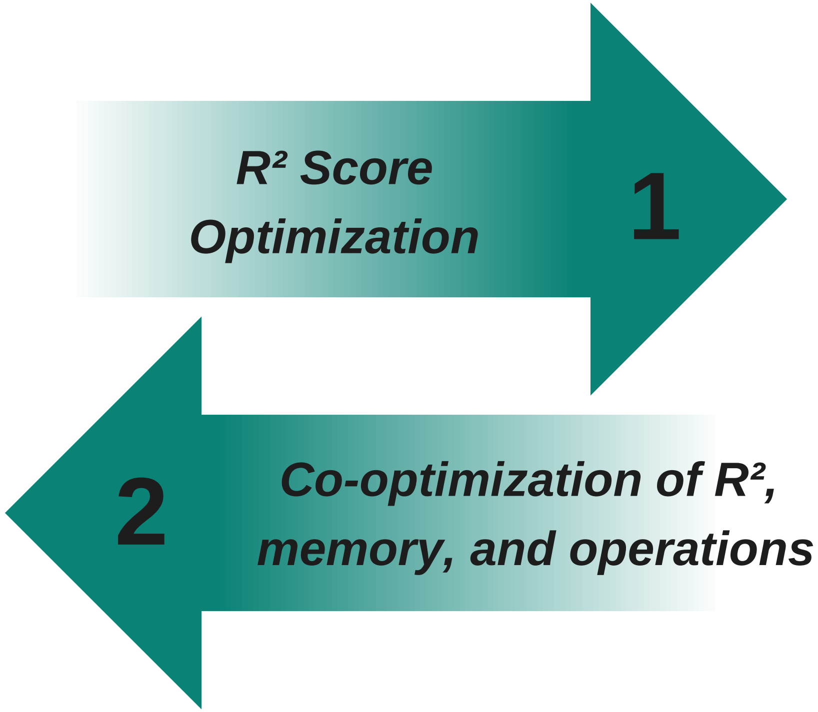
The Neurobench Framework



Tutorial here @ NICE!

Yik, Jason, et al. "The neurobench framework for benchmarking neuromorphic computing algorithms and systems." Nature Communications 16.1 (2025): 1545.

2 Tracks



Both **ANNs and SNNs** were welcomed to compete!

Zhou, Biyan, et al. "Grand Challenge on Neural Decoding for Motor Control of non-Human Primates." 2024 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, 2024.

Methods – Neural Decoding Challenge – Other Approaches

Yik et al., *Neurobench Baseline* [1]

Architecture: Feed-forward ReLU- and LIF-based networks

Demonstration of possible solution

Wang et al., *AEGRU* [2]

Architecture: encoder – GRUs – decoder

During training: additional firing rate reconstruction by auxiliary branch

→ Complex during training, simple during inference

Excellent accuracies

Track 1

Liu et al., *RSNN* [3]

Architecture: LIF-based SNNs with explicit recurrency

Pretraining on all recordings

Iterative pruning and activity regularization

Outstanding compression

Track 2

[1] https://github.com/NeuroBench/neurobench/blob/main/examples/primate_reaching/ANN.py

[2] Liu, Tengjun, et al. "Decoding finger velocity from cortical spike trains with recurrent spiking neural networks." *2024 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2024.

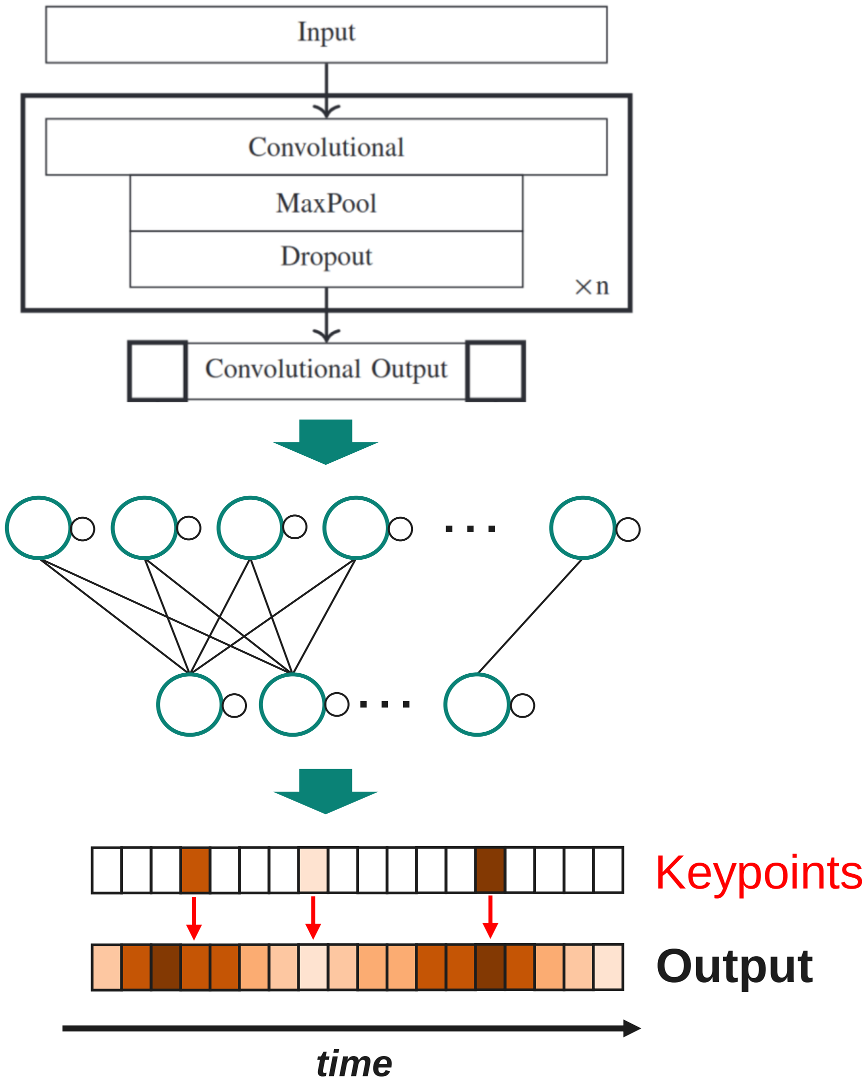
[3] Wang, Yuanxi, Zuowen Wang, and Shih-Chii Liu. "Leveraging recurrent neural networks for predicting motor movements from primate motor cortex neural recordings." *2024 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2024.

Methods – Neural Decoding Challenge – Our Approach

Temporal Convolution

Recurrent Processing

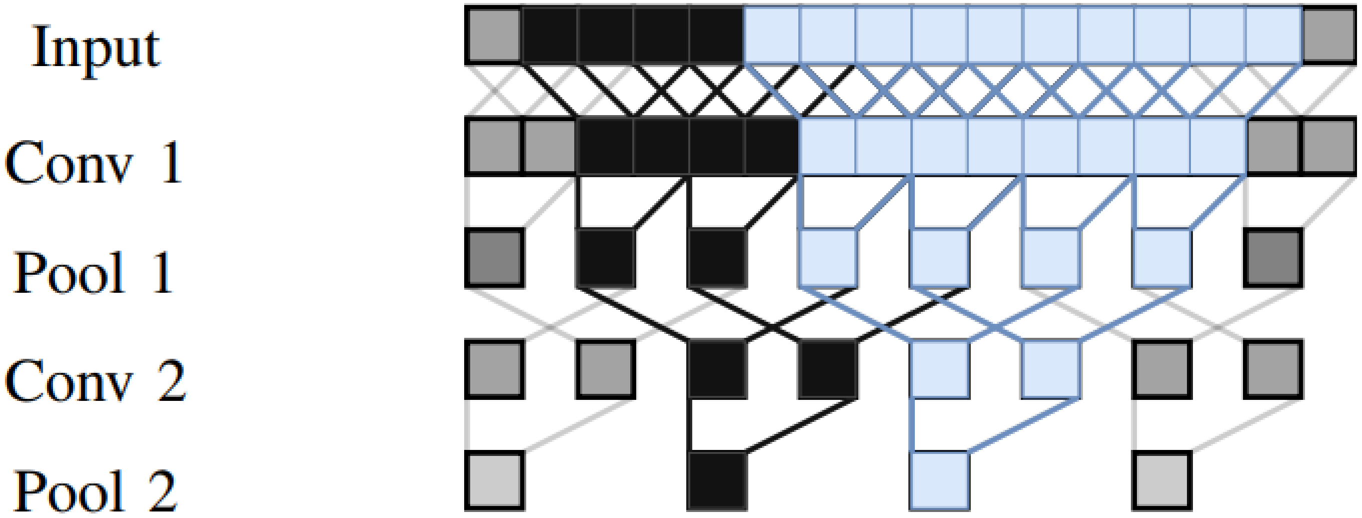
Interpolation



3rd Place in both Tracks!

Methods – Realtime-Capability – Buffering

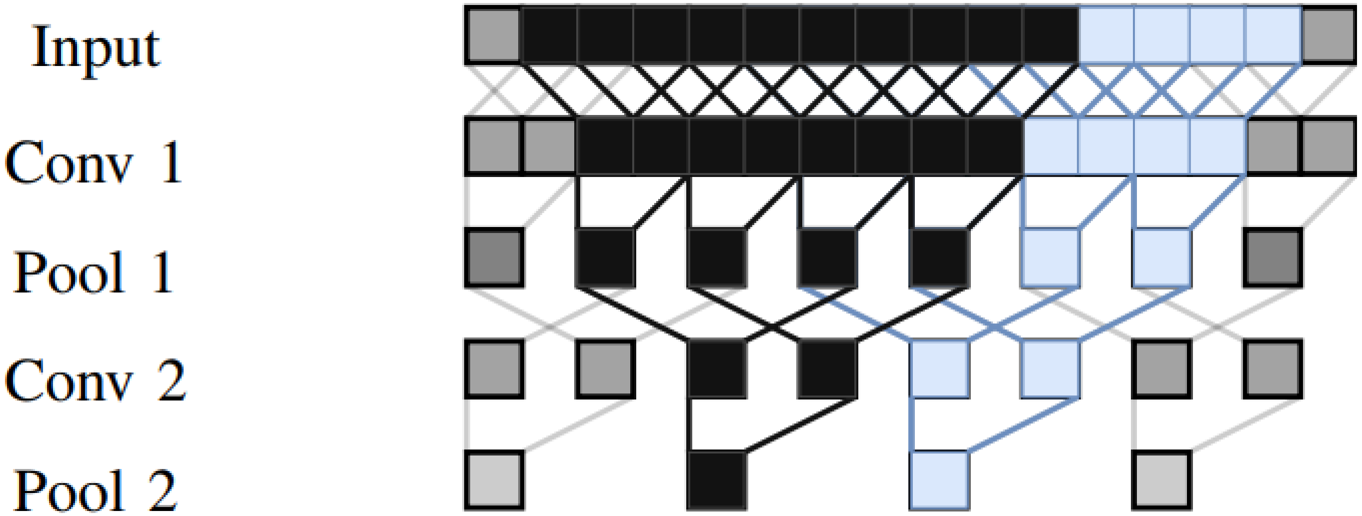
Input data buffering increases latency (and memory)



Buffer size per keypoint

Methods – Realtime-Capability – Buffering

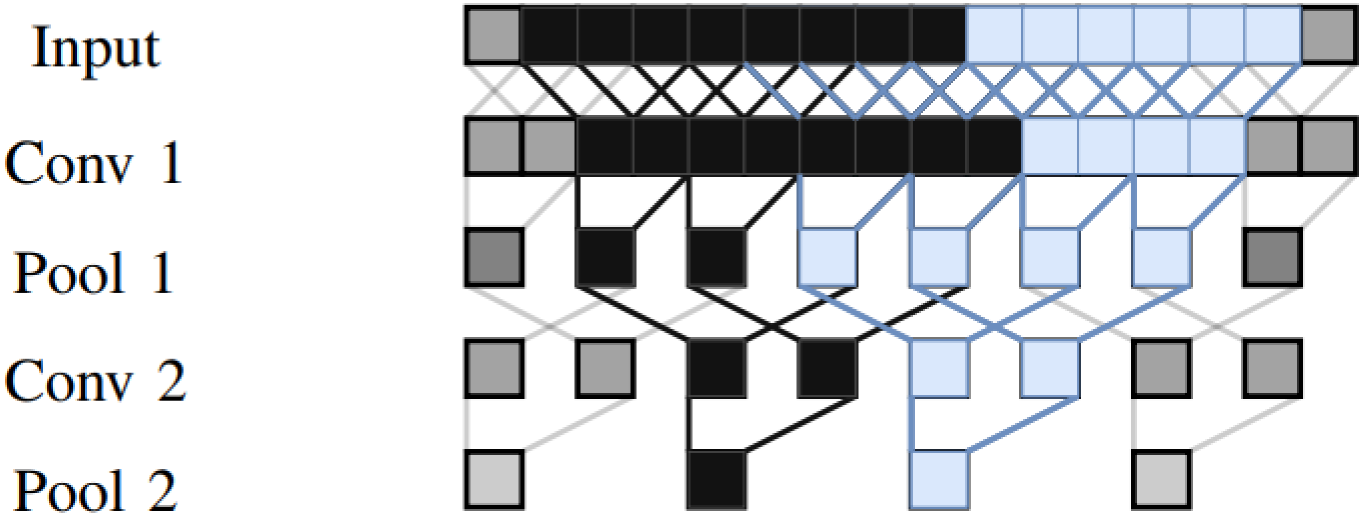
Input data buffering increases latency (and memory)



New input data buffer

Methods – Realtime-Capability – Buffering

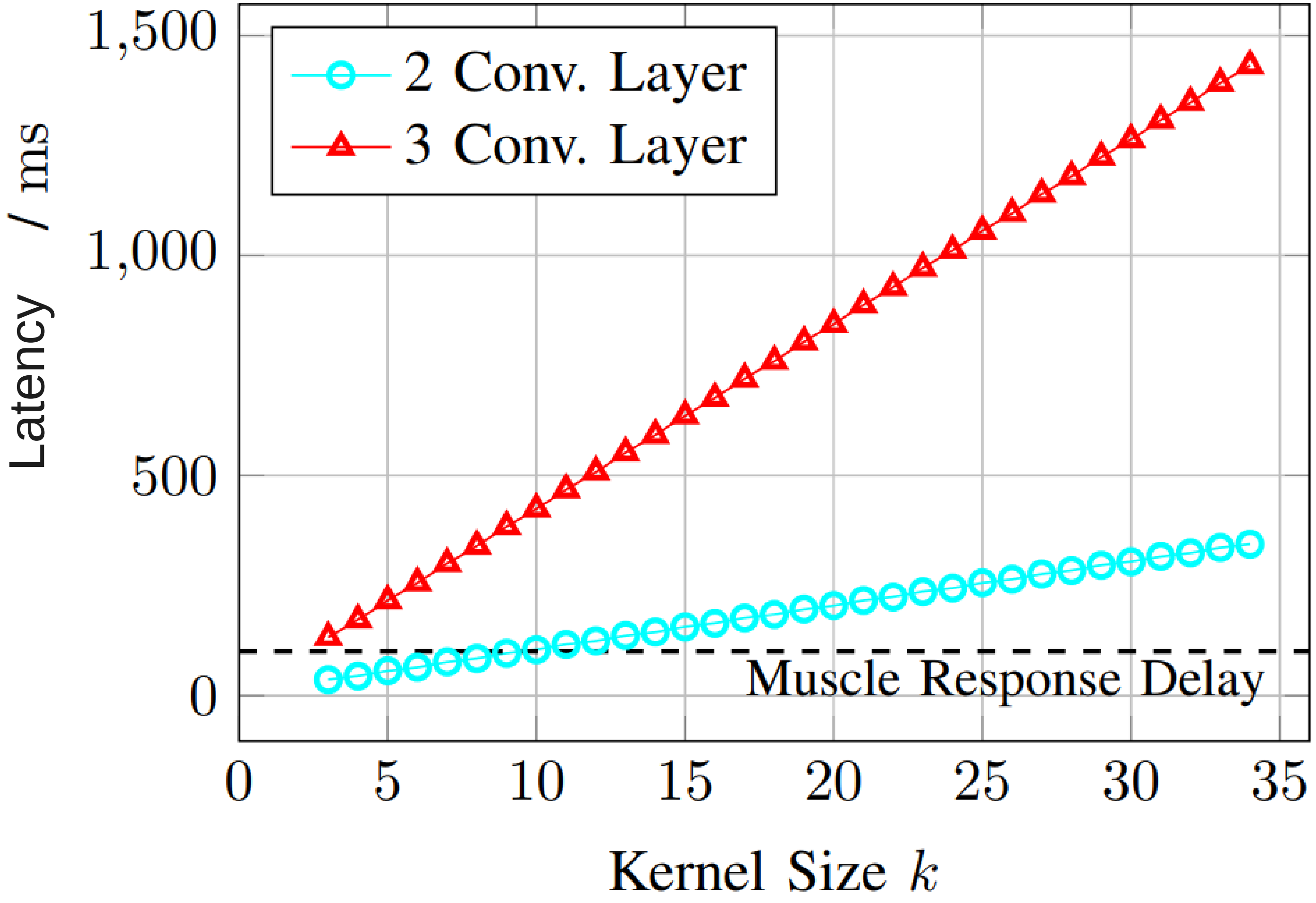
Input data buffering increases latency (and memory)



Required buffer size per update

All models are tested using this buffer architecture

Methods – Realtime-Capability – Architecture



2 Models:

BMnet for max. R^2

RTnet for realtime-capability

Methods – Compression Techniques

1. Spike Regularization

$\mathcal{L}_S = \lambda_S * \text{Spikes}$

2. Weight Regularization

$\mathcal{L}_W = \lambda_W * ||w||_2^2$

3. Fixed Point Quantization

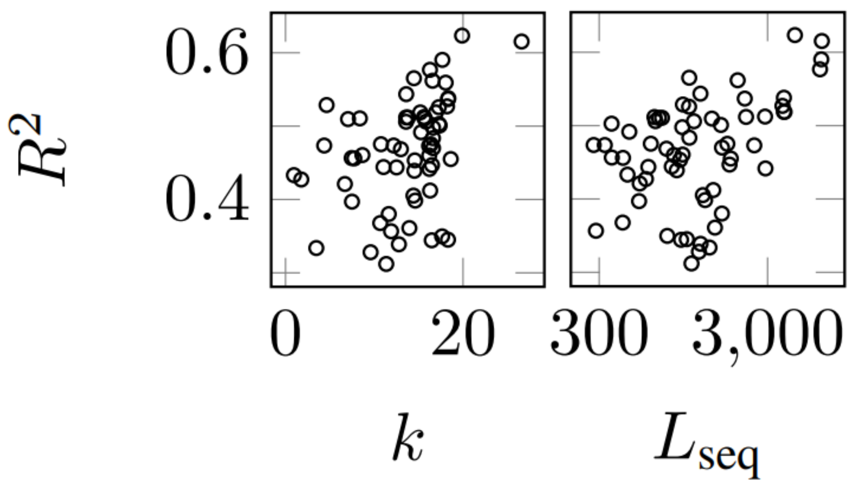
Weights: 1i-7f

Buffers: 1i-4f

~~4. Pruning~~

3rd Model:
sRTnet (compressed RTnet)

Results & Discussion – Hyperparameter Optimization



k..kernel size of temporal convolutions

L_{seq}..length of training data sequences

Kernel size is limited due to memory and computational load

Sequence length is limited due to training time

Results & Discussion – Models Targeting R^2 Optimization

Model	Event-Based	Test R^2
Yik et al. - SNN3 [19]	✓	0.633
Liu et al. - bigRSNN [17]	✓	0.698 ± 0.002
Vasilache et al. - LIF-t1 [16]	✓	0.648 ± 0.022
Wang et al. [18]	✗	0.710 ± 0.050
Yik et al. - ANN3 [19]	✗	0.615
Vasilache et al. - GRU-t1 [16]	✗	0.707 ± 0.012
This Work - BMnet	✓	$\mathbf{0.717 \pm 0.004}$
This Work - RTnet	✓	0.685 ± 0.006

1) Improving our previous results by 7% in R^2

2) Improving the SotA by 1% in R^2

Decreased kernel size of RTnet decreases accuracy

Results & Discussion – Models Targeting Co-Optimization

Model	Event-Based	Test R^2	Memory Footprint / Bytes	MACs	ACs	Realtime-Capable
Yik et al. - SNN2 [19]	✓	0.581	29 248	0	414	✓
Liu et al. - tinyRSNN [17]	✓	0.660 ± 0.002	**27 144**	0	304 ± 12	✓
Vasilache et al. - LIF-t2 [16]	✓	0.566 ± 0.016	168 596	201 ± 0	254.0 ± 0.8	✗
Yik et al. - ANN2 [19]	✗	0.576	27 160	4 970	0	✗
Vasilache et al. - GRU-t2 [16]	✗	0.621 ± 0.014	174 104	627 ± 0	248 ± 0	✗
This Work - sRTnet	✓	$\mathbf{0.675 \pm 0.011}$	105 269	$12\,274 \pm 37$	326 ± 4	✓

1) Improve SotA accuracy of co-optimization models by 1.5% R²

2) Increased memory compared to SotA by factor of 4

Compression techniques improve footprint compared to baseline despite increased kernel size

3) Our hybrid networks are now realtime-capable!

Results & Discussion – Compression – ACs and MACs

Model	Event-Based	Test R^2	Memory Footprint / Bytes	MACs	ACs	Realtime-Capable
Yik et al. - SNN2 [19]	✓	0.581	29 248	0	414	✓
Liu et al. - tinyRSNN [17]	✓	0.660 ± 0.002	**27 144**	0	304 ± 12	✓
Vasilache et al. - LIF-t2 [16]	✓	0.566 ± 0.016	168 596	201 ± 0	254.0 ± 0.8	✗
Yik et al. - ANN2 [19]	✗	0.576	27 160	4 970	0	✗
Vasilache et al. - GRU-t2 [16]	✗	0.621 ± 0.014	174 104	627 ± 0	248 ± 0	✗
This Work - sRTnet	✓	$\mathbf{0.675 \pm 0.011}$	105 269	$12\,274 \pm 37$	326 ± 4	✓

Relative cost of MAC to AC (45nm CMOS) is ~10 [1]

➡ Large kernel size makes sRTnet highly inefficient compared to, e.g., tinyRSNN

Separable convolutional and spiking subnets can be deployed on hybrid platform of specialized CNN and neuromorphic accelerators

➡ This enables truly fair comparison at runtime

[1] Han, Song, et al. "Learning both weights and connections for efficient neural network." *Advances in neural information processing systems* 28 (2015).

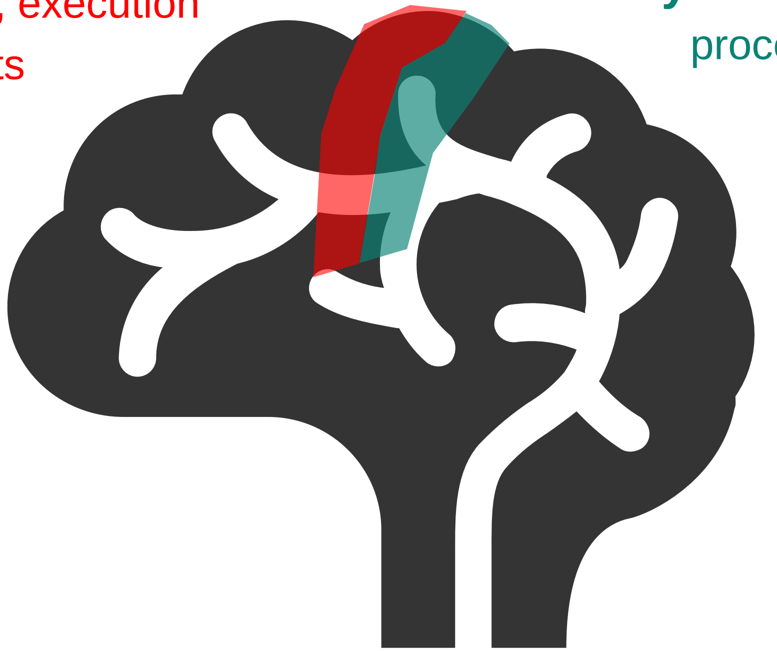
Results & Discussion – Heterogeneous Cortex Data (M1 & S1)

Primary Motor Cortex (M1)

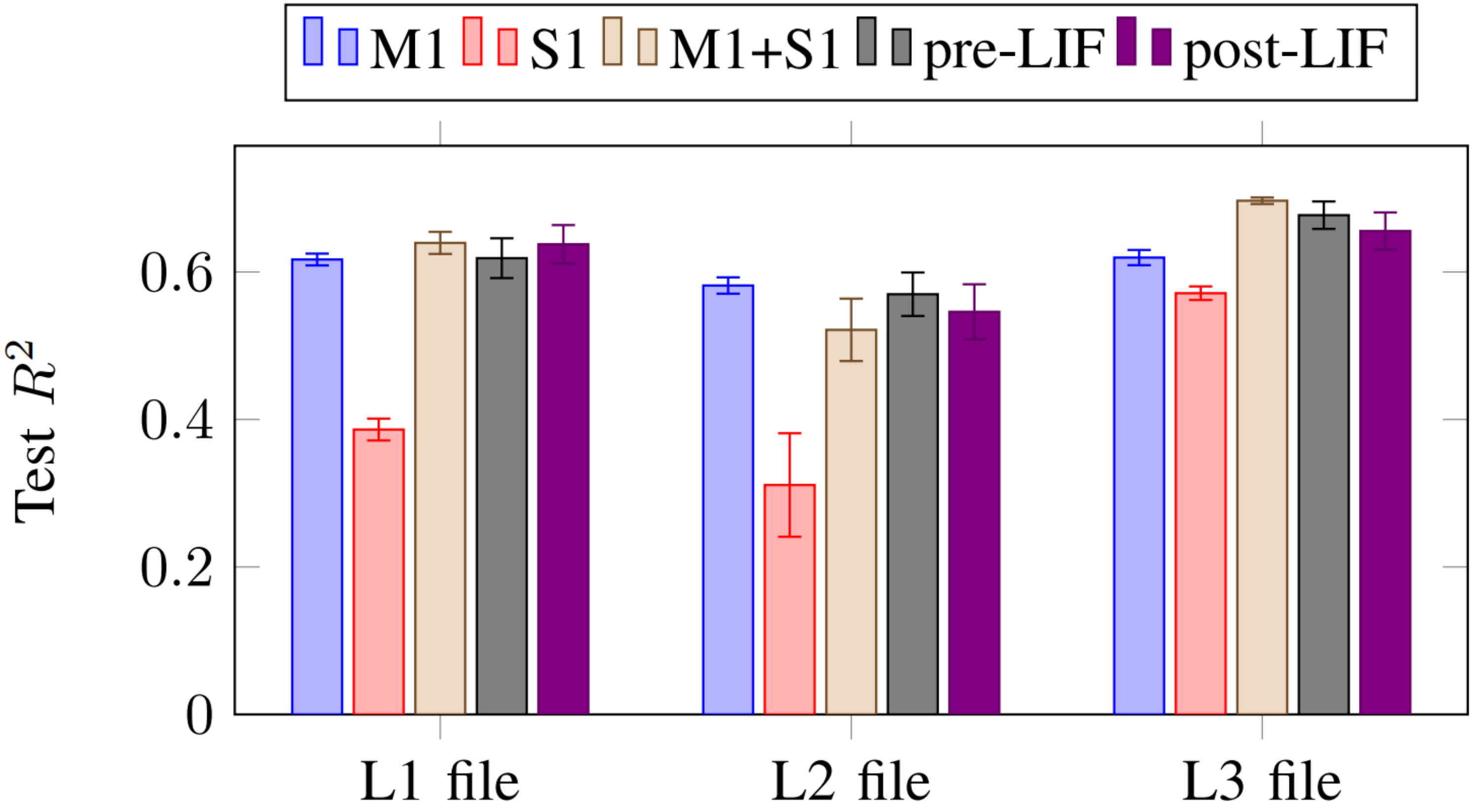
Planning, controlling, execution of muscle movements

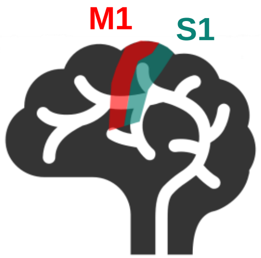
Primary Somatosensory Cortex (S1)

processing tactile sensory information

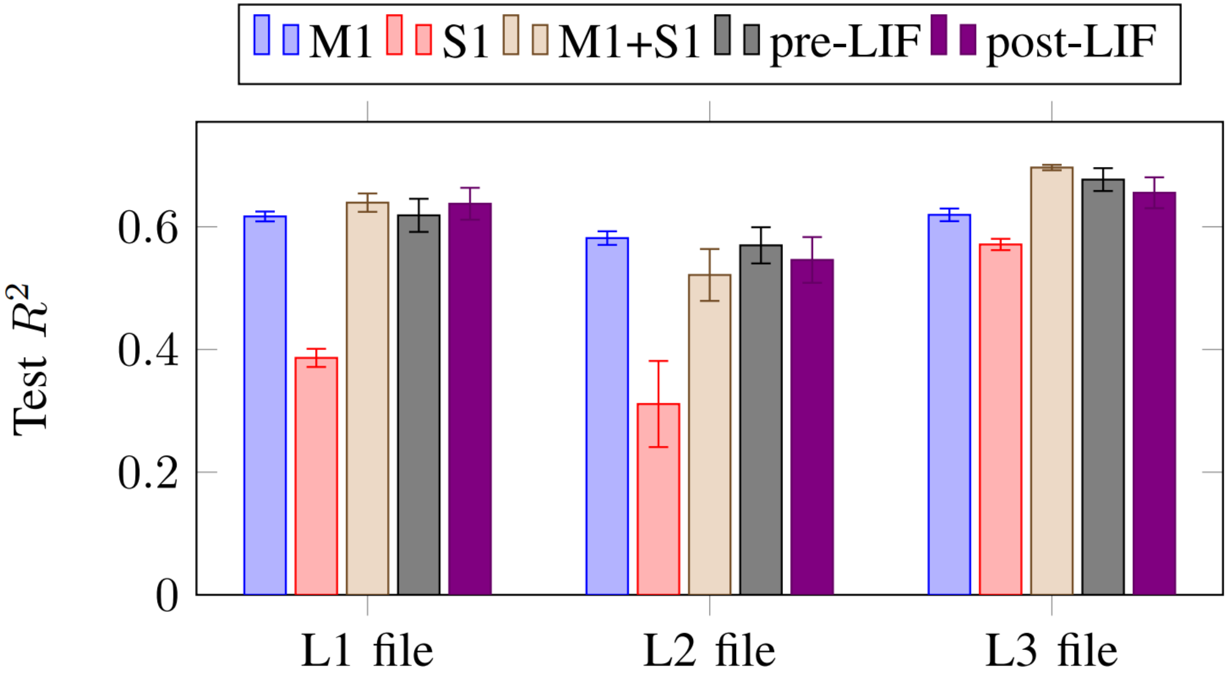


Results & Discussion – Heterogeneous Cortex Data (M1 & S1)

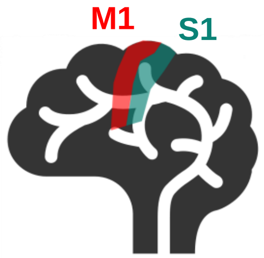




Results & Discussion – Heterogeneous Cortex Data (M1 & S1)



- **pre-LIF**..separate conv. blocks for M1 and S1 data
- **post-LIF**..separate conv. blocks and LIF nets for M1 and S1 data



1) M1 data massively improves the decoding accuracy

2) S1 data complements M1 data to improve decoding accuracy

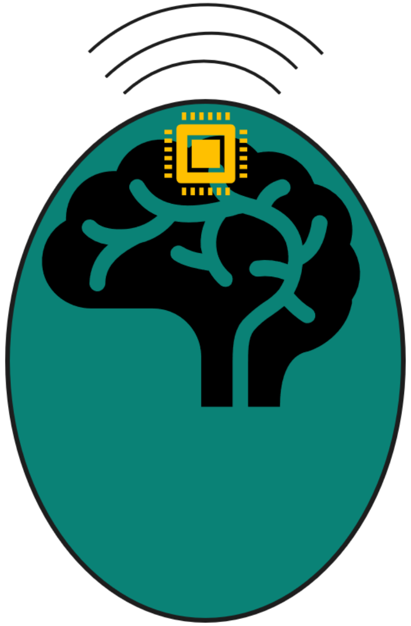
3) Separate conv. blocks or conv. blocks and LIF nets do not improve decoding… **What better methods are there to aid M1 decoding with S1 data?**

Summary

Problem 1: Skull Opening
Increases the risk of infection

Problem 2: Bulky Wiring
Impairs head movement

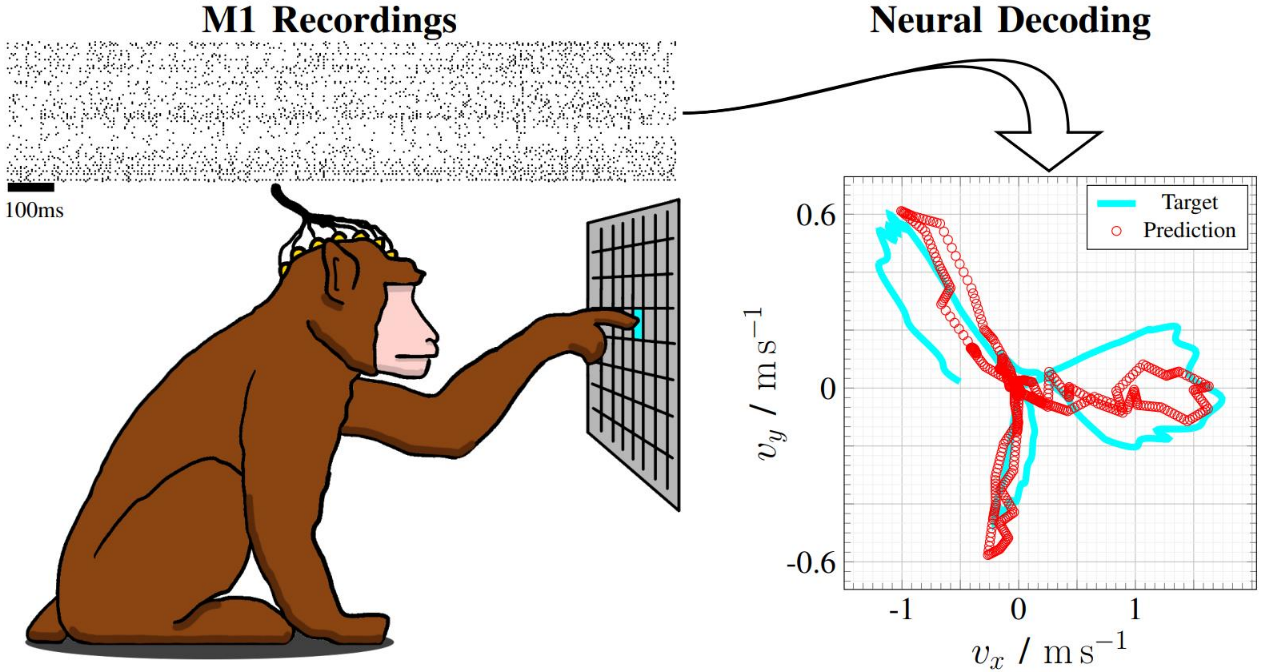
Possible Solution:
Wireless iBMIs



Patient X's Head

neuromorphic computing very promising as a solution to the constraints of wireless iBMIs

Summary



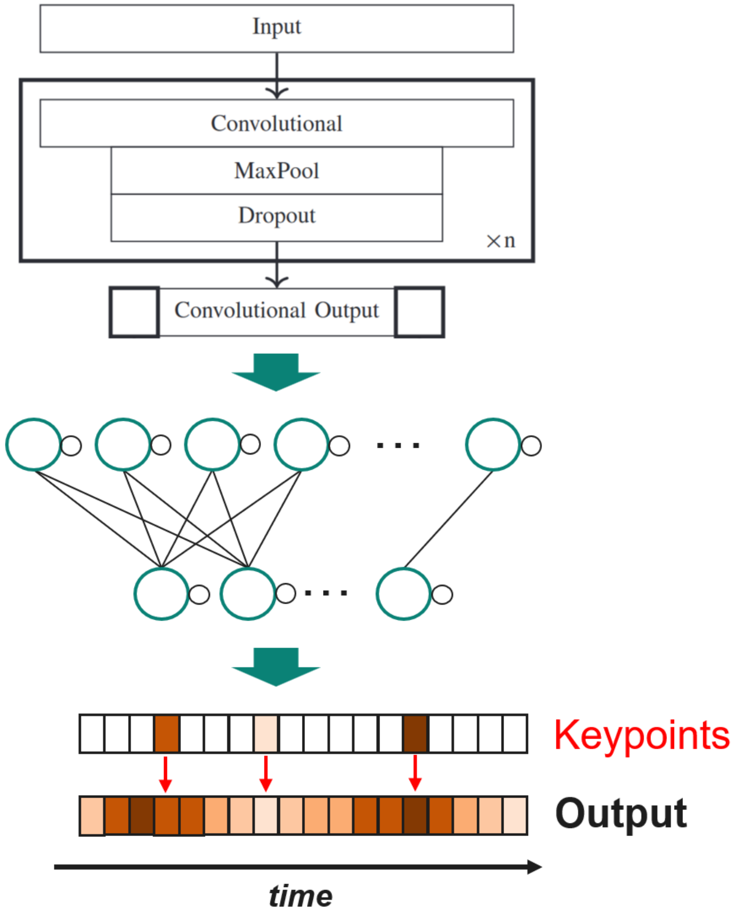
Primate Reaching as new sequence-to-sequence benchmark for efficient AI

Summary

Temporal Convolution

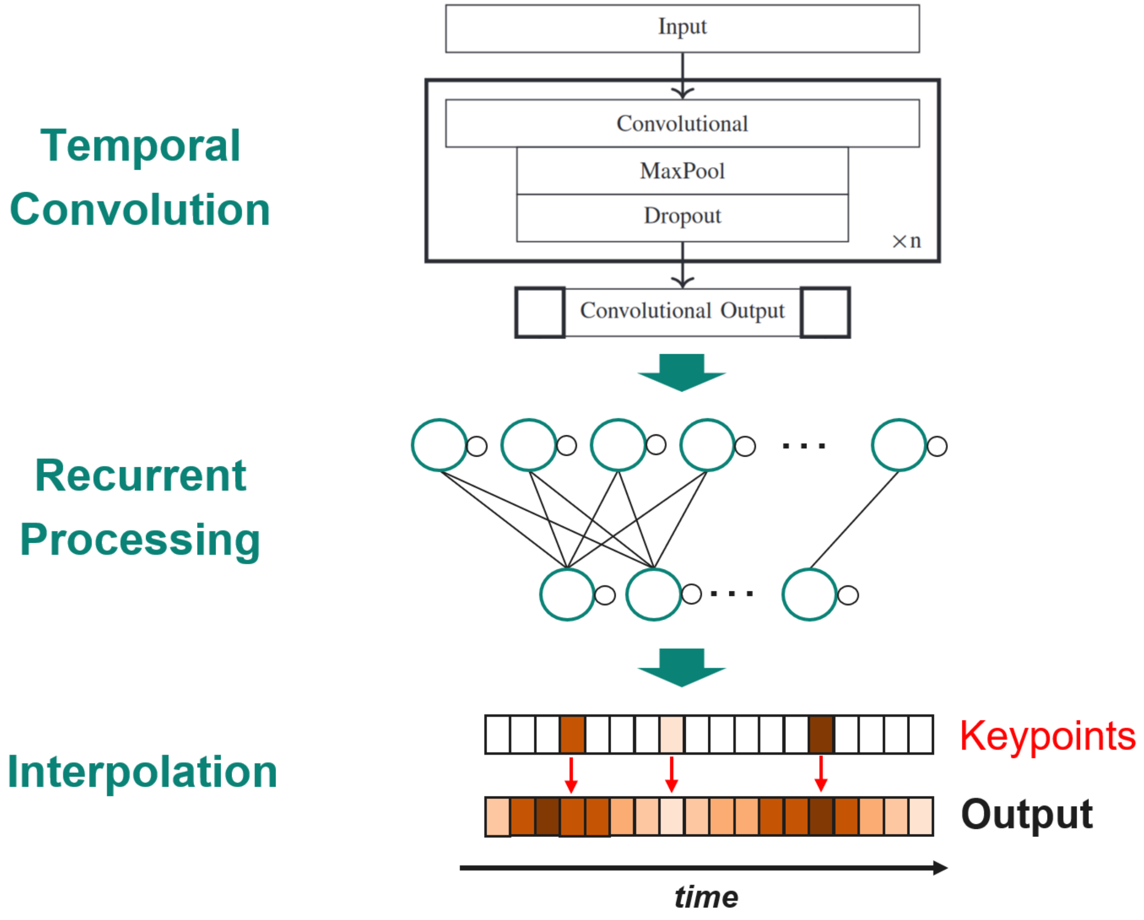
Recurrent Processing

Interpolation



Our approach improves the SotA accuracy by scaling up the context window and training sequence length, and now is realtime-capable!

Summary



Our approach improves the SotA accuracy by scaling up the context window and training sequence length, and now is realtime-capable!

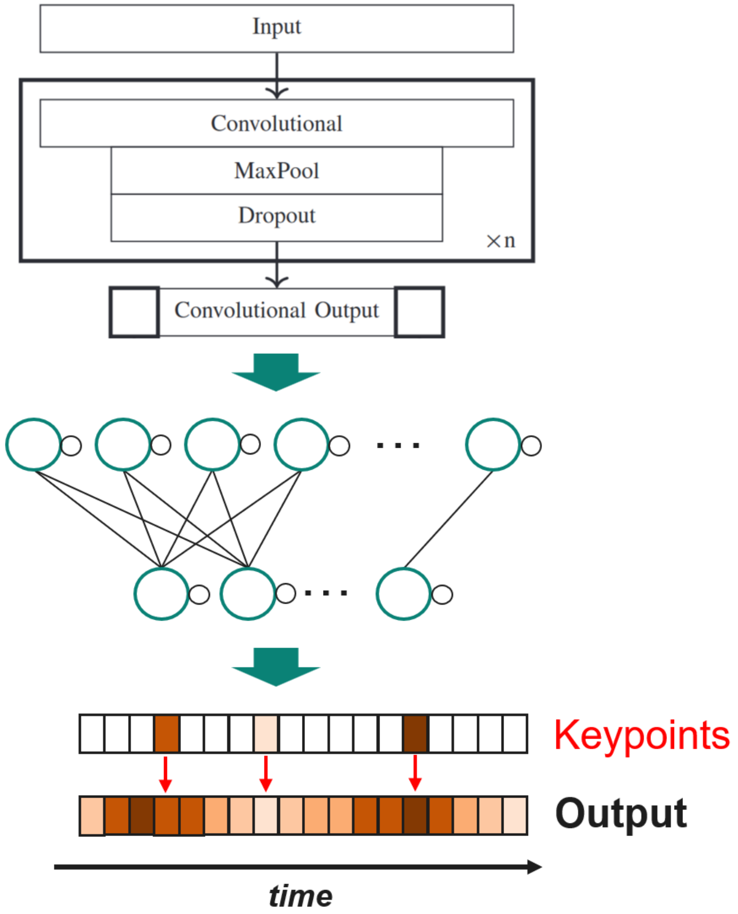
However, the large temporal kernel size makes it hard to be compressed to very small sizes

Summary

Temporal Convolution

Recurrent Processing

Interpolation



Our approach improves the SotA accuracy by scaling up the context window and training sequence length, and now is realtime-capable!

However, the large temporal kernel size makes it hard to be compressed to very small sizes

Final step: deployment on hybrid HW platforms for evaluation at runtime

Acknowledgements

Neuromorphs of Group Becker, NICE 2025

Thank you! ☺