Hardware-aware Few-shot Learning on a Memristor-based Small-world Architecture

Institute of Neuroinformatics

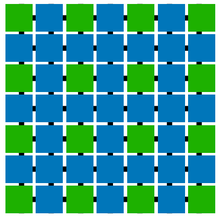
University of Zürich and ETH Zürich

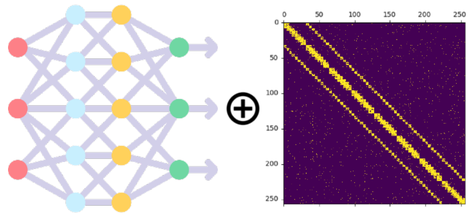
Karthik Charan Raghunathan
Emerging Intelligent Substrates lab

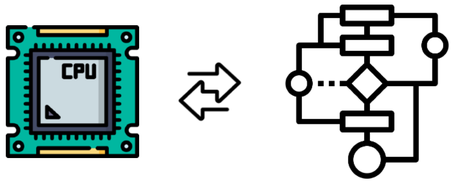
24th April 2024

Take home message

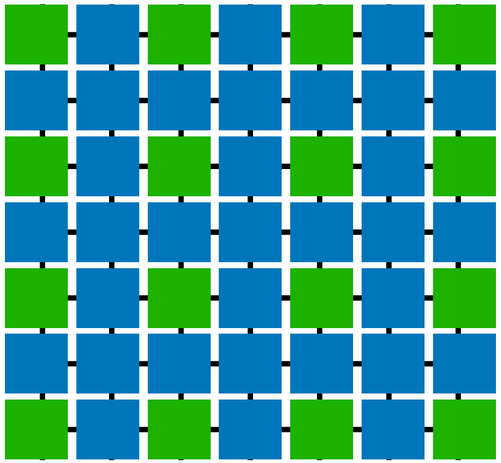
- Demonstrated (using simulations) Online few-shot learning on a novel, energy efficient hardware architecture
- Constraining network architecture does not terminally detriment the performance of the model
- Importance of inner loop dynamics in MAML; Highlighting the need for HW-SW co-design for efficient





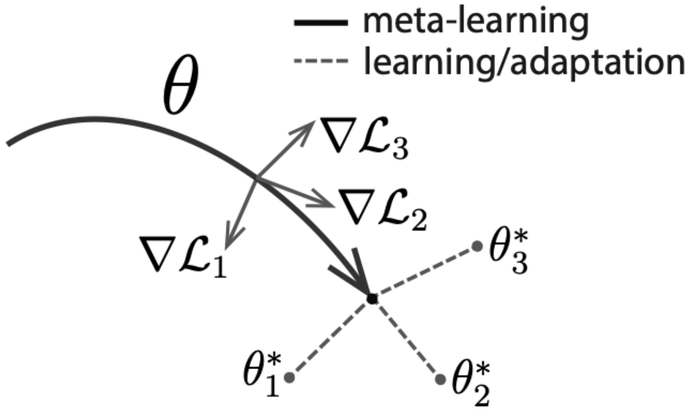


What is this project about?



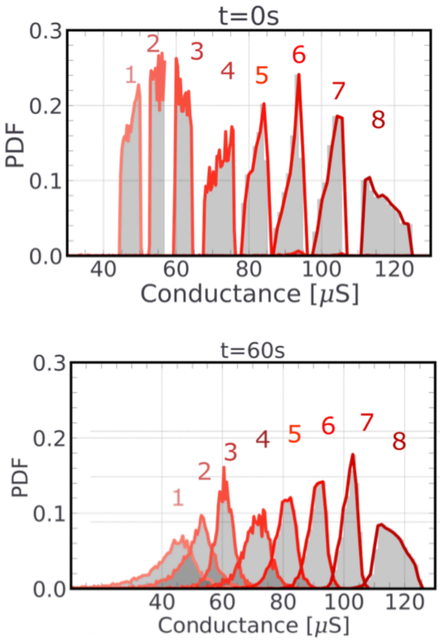
Novel energy efficient architecture

+



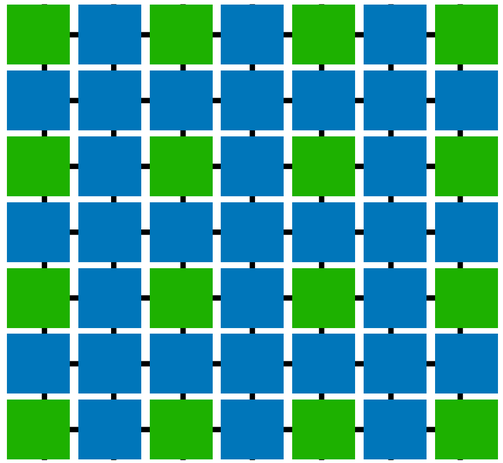
Meta-learning

+



Hardware-aware training

Low-power architecture



+

+

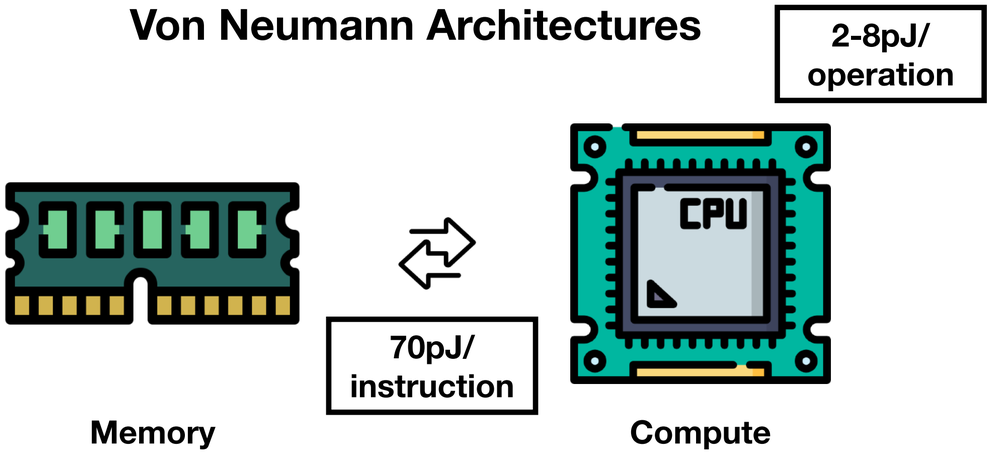
Novel energy efficient architecture

Meta-learning

Hardware-aware training

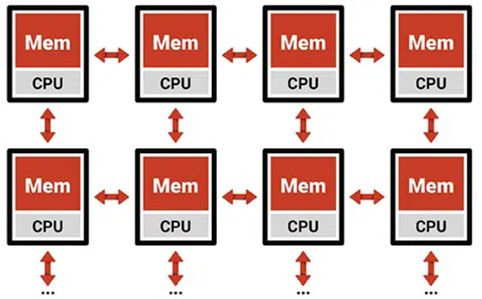
The Memory wall problem

Von Neumann Architectures



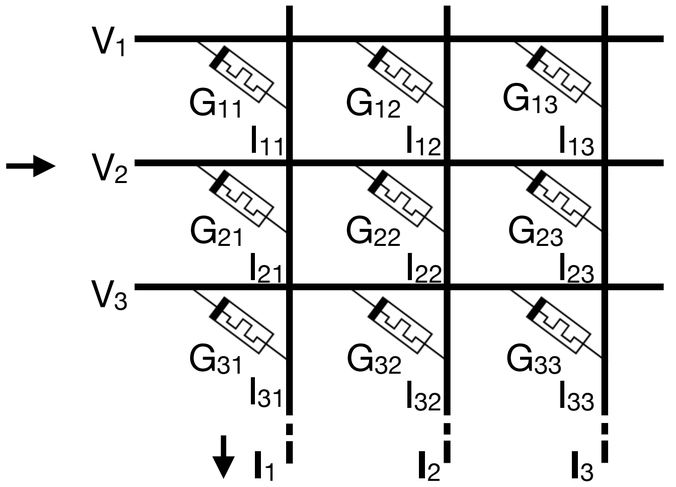
Non-Von Neumann Architectures

Near-memory compute



Memory & Compute

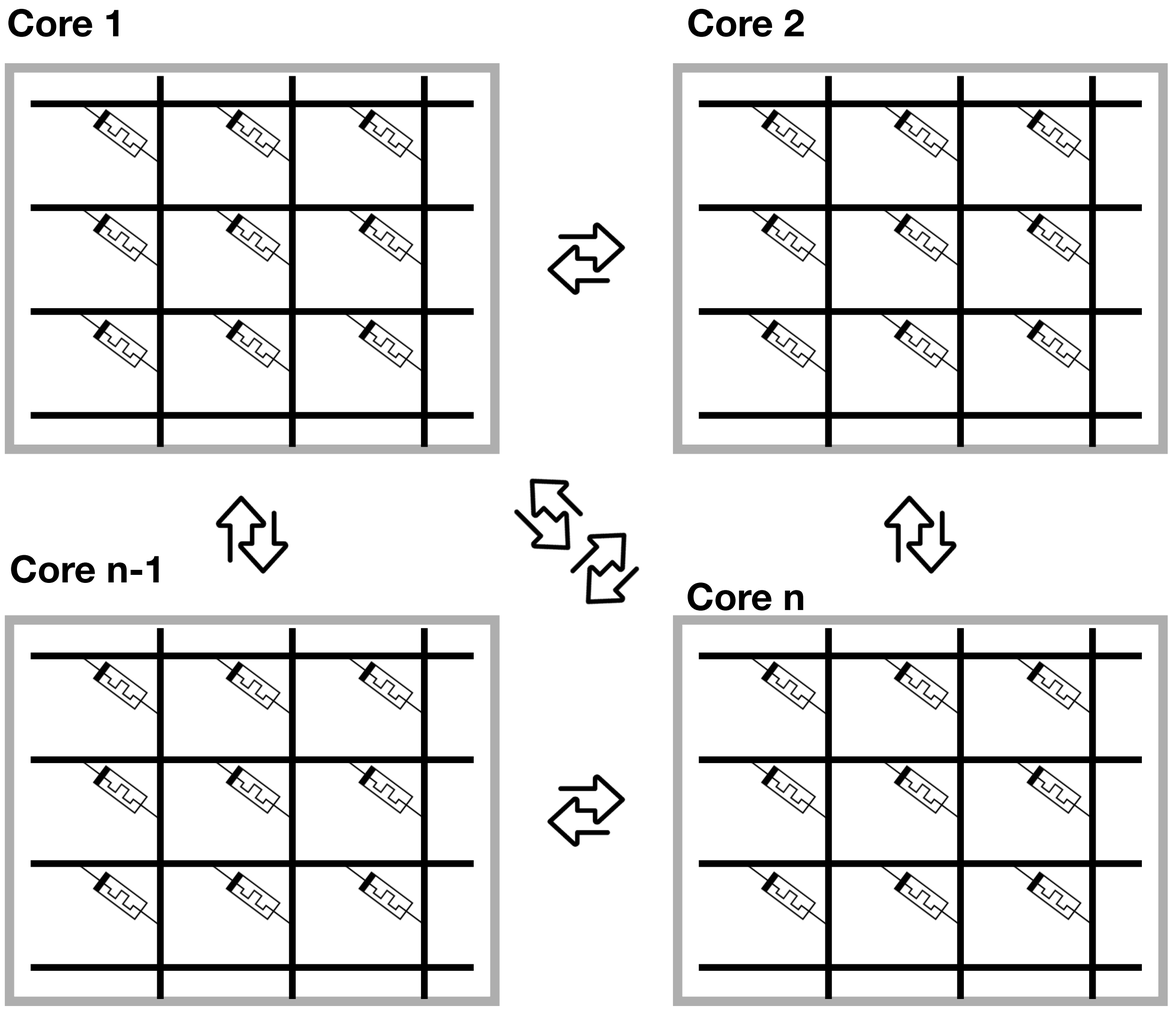
In-memory compute



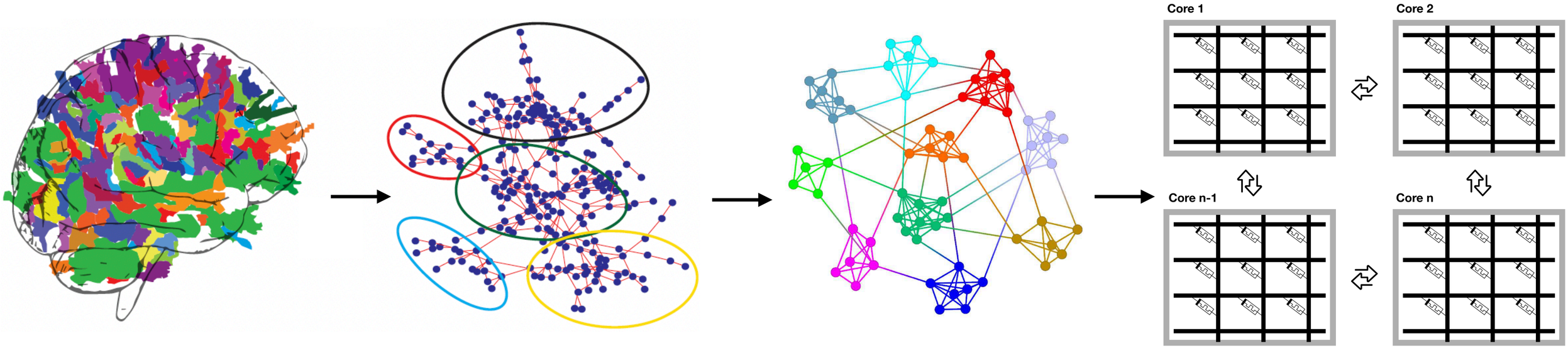
Memory + Compute

Mark, H. (2014, February). Computing's energy problem (and what we can do about it). In *Proceedings of the IEEE International Solid-State Circuits Conference, San Francisco, CA, USA* (pp. 9-13).

Scaling IMC: *Possible solution*



Scaling multi-core IMC

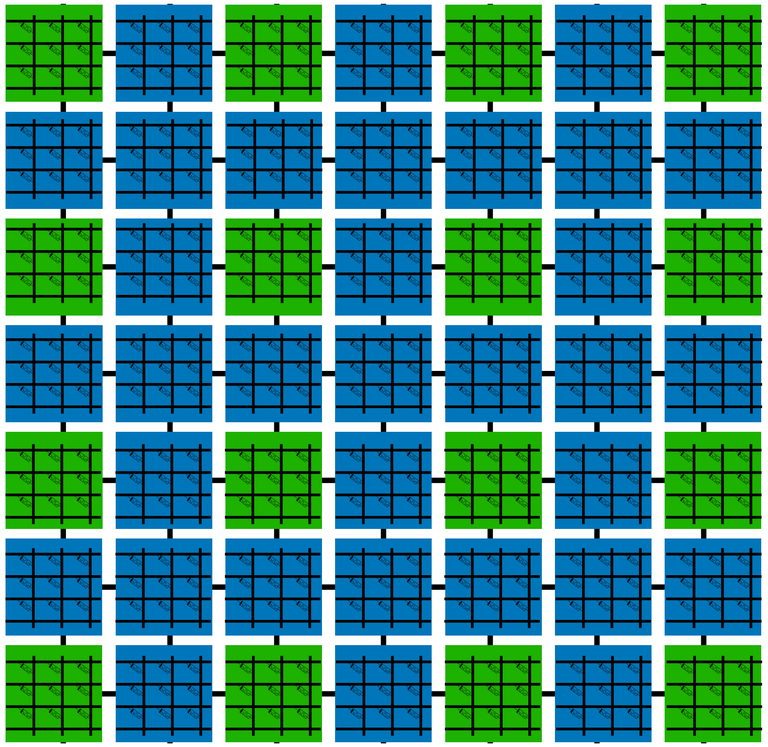


- High clustering coefficient
- Low number of total connections

Faraz Zaidi. Small world networks and clustered small world networks with random connectivity. Social Network Analysis and Mining, 2012. hal-00679660

Dalgaty, T., Moro, F., Demirağ, Y., De Pra, A., Indiveri, G., Vianello, E., & Payvand, M. (2024). Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems. *Nature Communications*, *15*(1), 142.

The Neuromorphic Mosaic



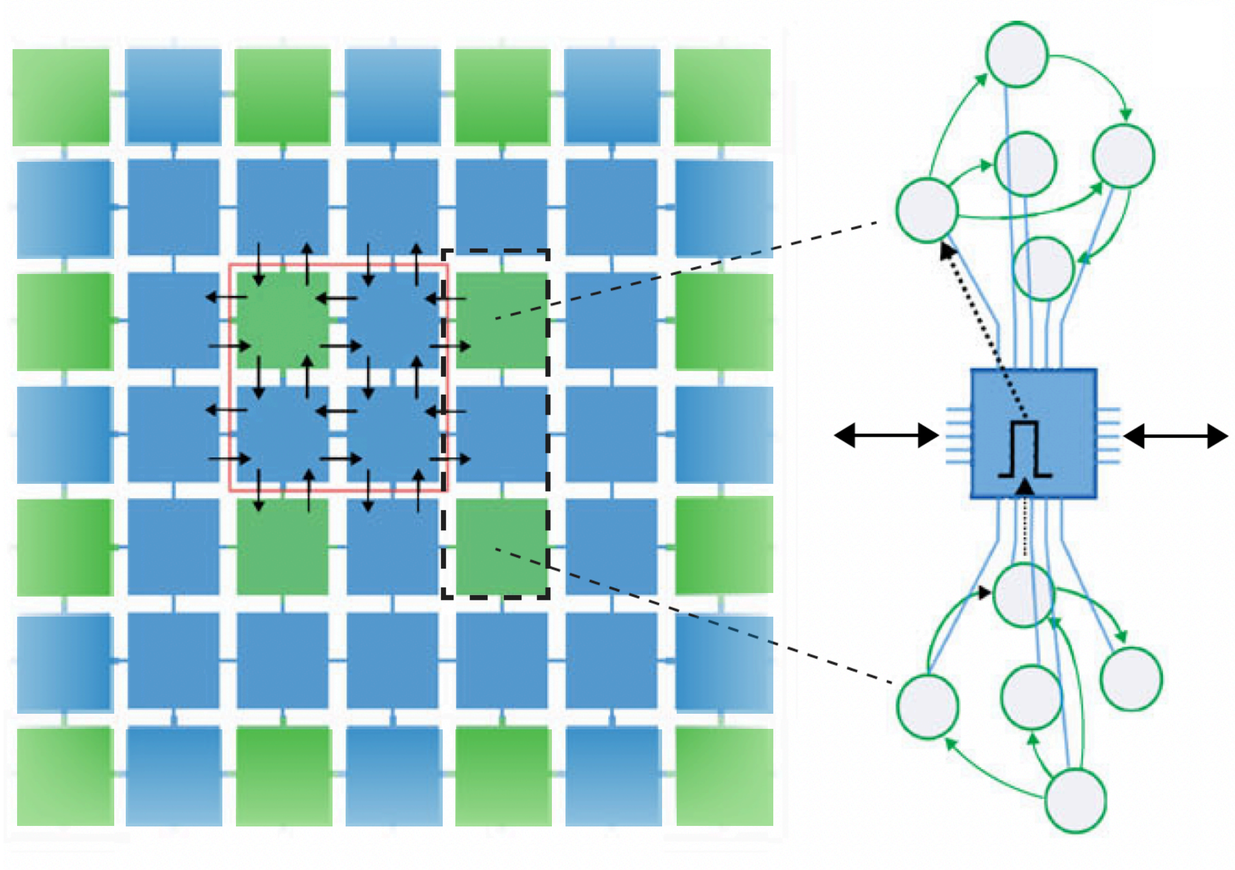
- Non-von Neumann ***systolic*** architecture
- Distributed memristors for ***in-memory computing*** and ***in-memory routing***
- Efficiently implements ***small-world*** graph topologies for Spiking Neural Networks (SNNs)

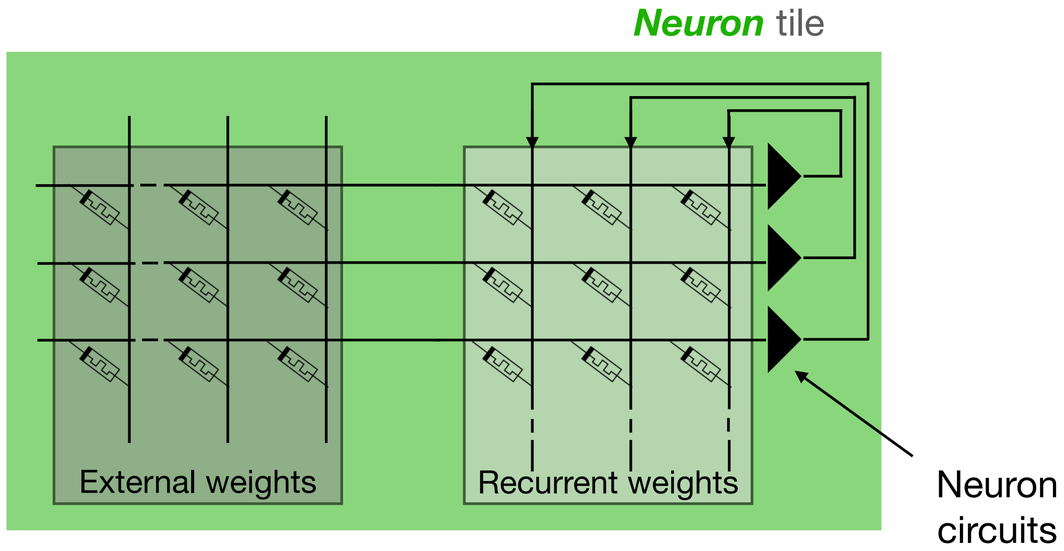
Neuron core

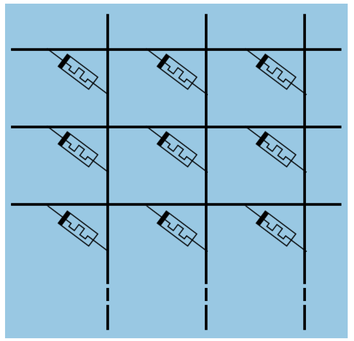
Routing core

Dalgaty, T., Moro, F., Demirağ, Y., De Pra, A., Indiveri, G., Vianello, E., & Payvand, M. (2024). Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems. *Nature Communications*, *15*(1), 142.

The Neuromorphic Mosaic

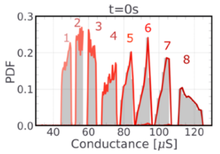




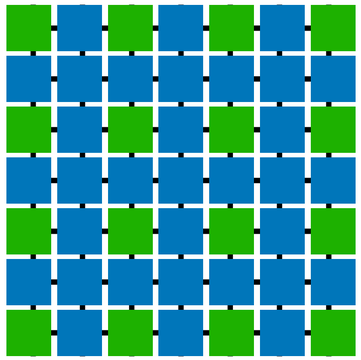


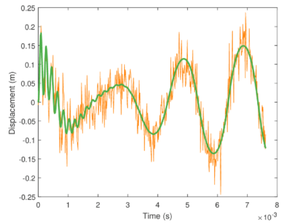
Dalgaty, T., Moro, F., Demirağ, Y., De Pra, A., Indiveri, G., Vianello, E., & Payvand, M. (2024). Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems. *Nature Communications*, *15*(1), 142.

Neuromorphic Mosaic for edge computing

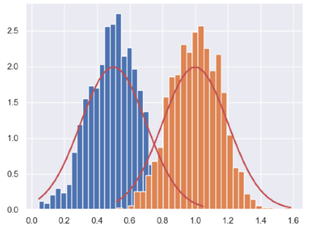


RRAM device non-idealities





Environmental Noise



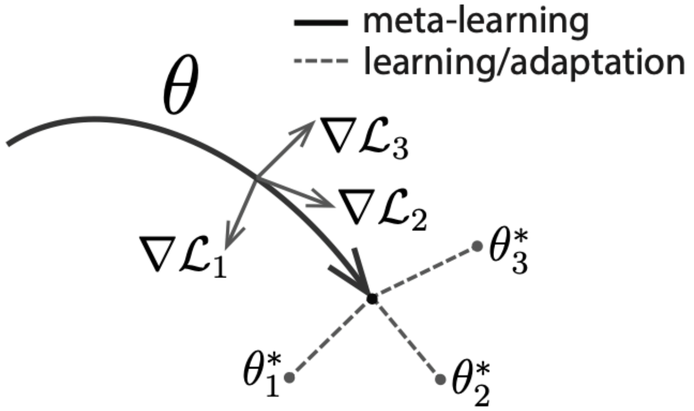
Domain data-shifts

Demirag, Y., Dittmann, R., Indiveri, G., & Neftci, E. O. (2023). Overcoming phase-change material non-idealities by meta-learning for adaptation on the edge. *Proceedings of Neuromorphic Materials, Devices, Circuits and Systems (NeuMatDeCaS)*.

Stewart, K. M., & Neftci, E. O. (2022). Meta-learning spiking neural networks with surrogate gradient descent. *Neuromorphic Computing and Engineering*, *2*(4), 044002.

Meta-learning algorithm

Novel energy efficient architecture

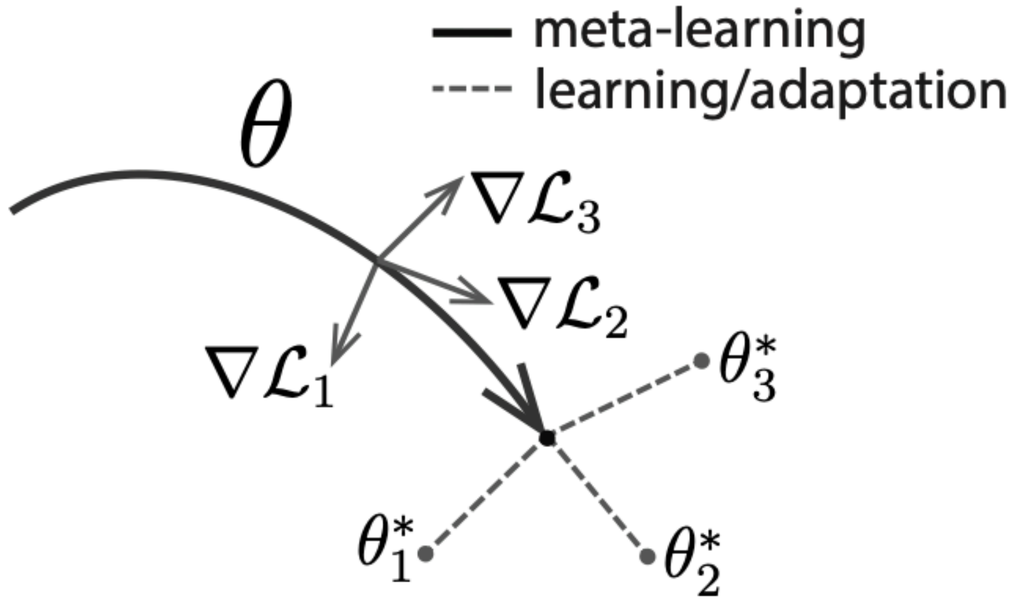


Meta-learning

Hardware-aware training

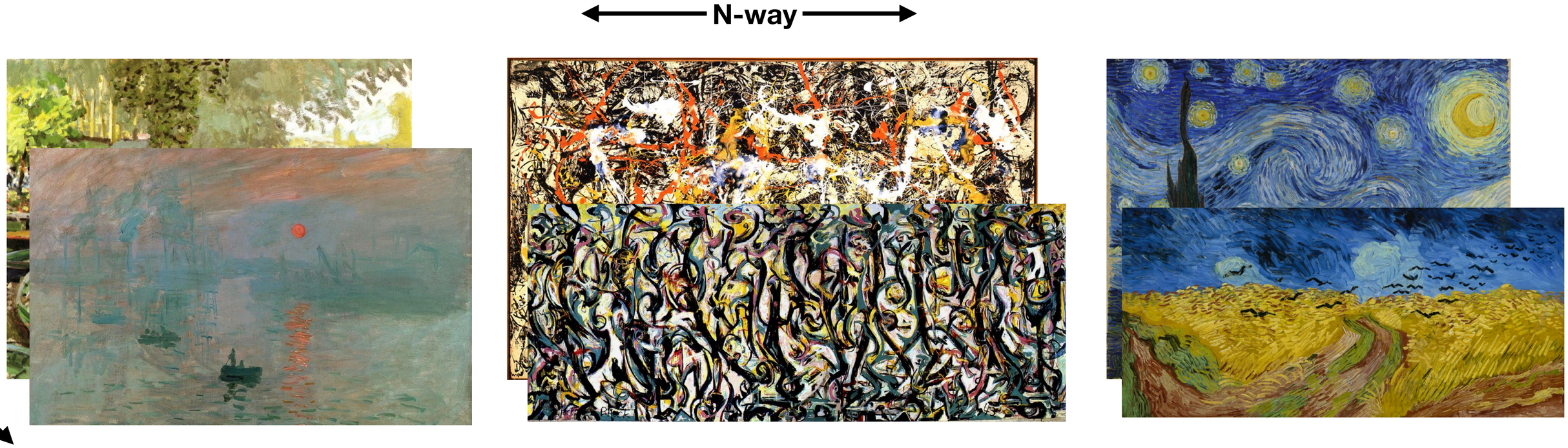
Model Agnostic Meta-learning
(MAML)

- Learning to learn approach
- Model agnostic
- Learn a good initialisation
- Fine-tune to learn new examples (few-shot)
- Gradient-based bi-level optimisation



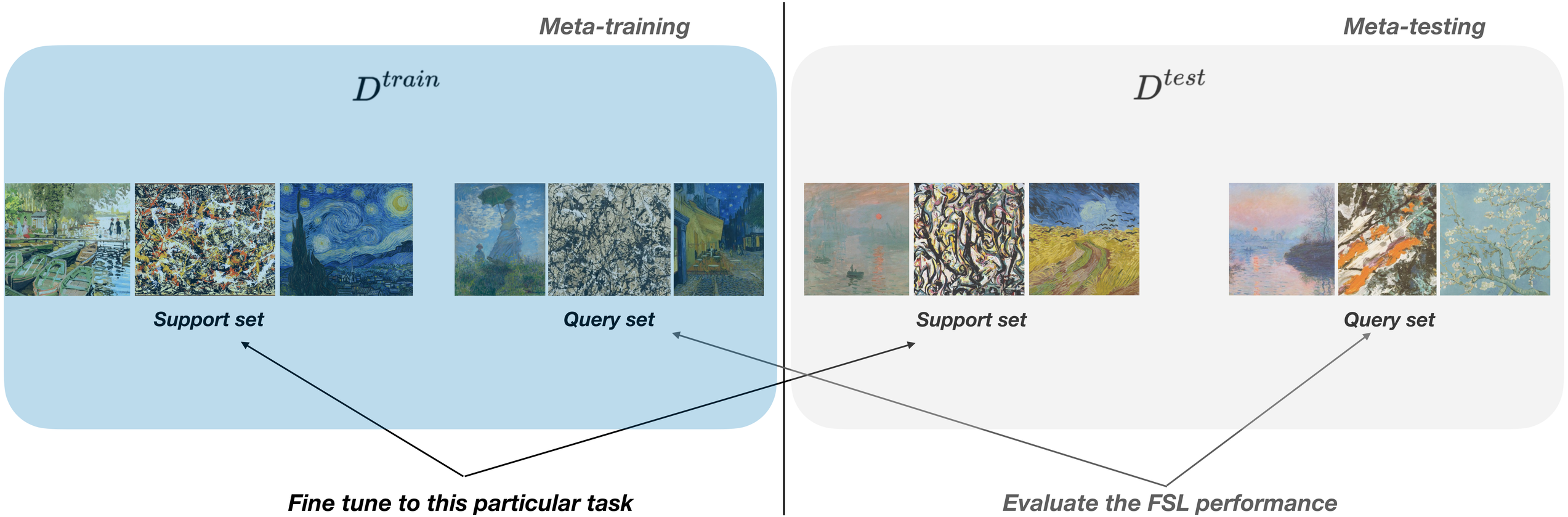
Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126-1135). PMLR.

A few-shot task



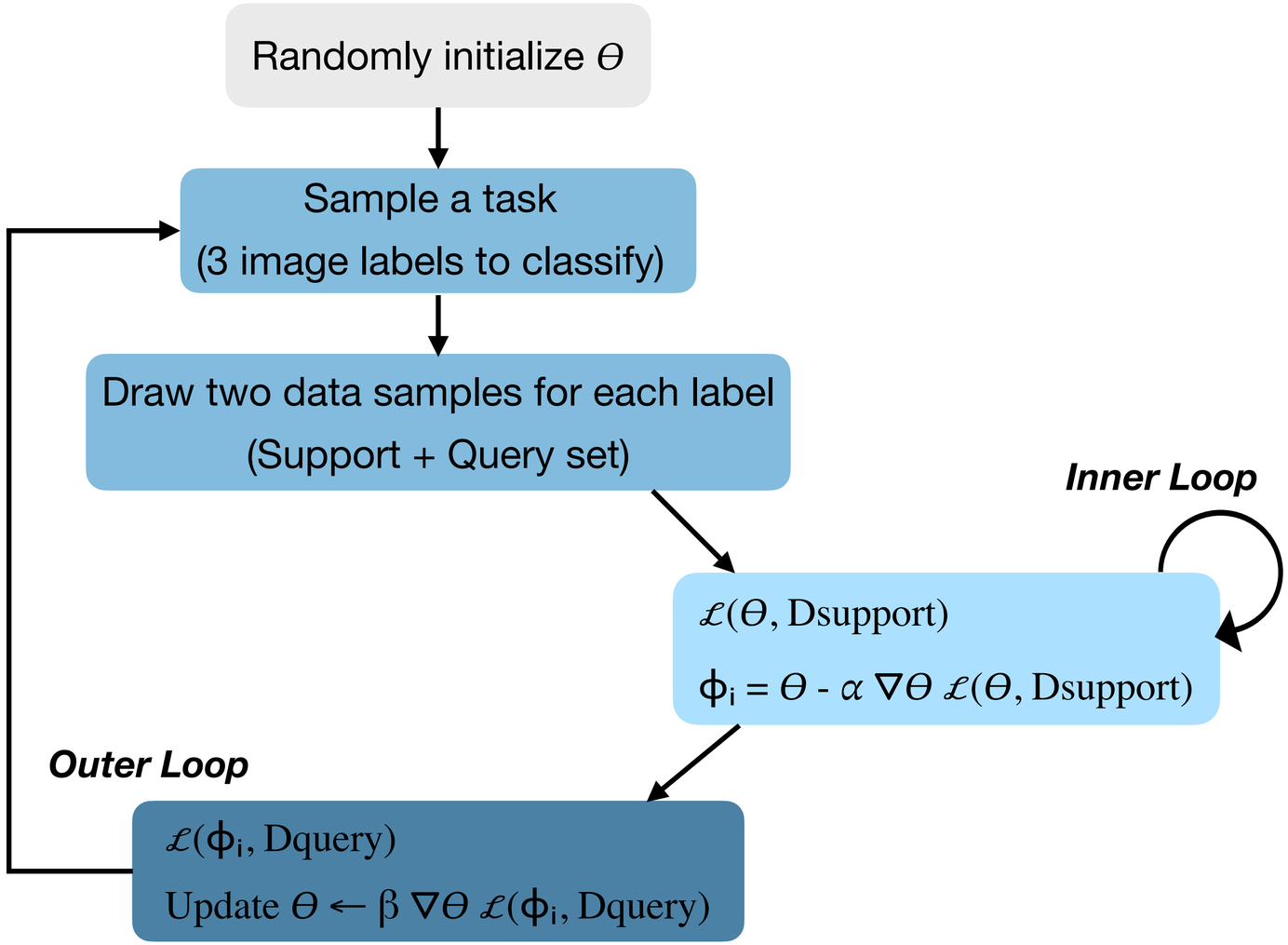
["Claude Monet", "Jackson Pollock", "Vincent Van Gogh"]

A few-shot data loader



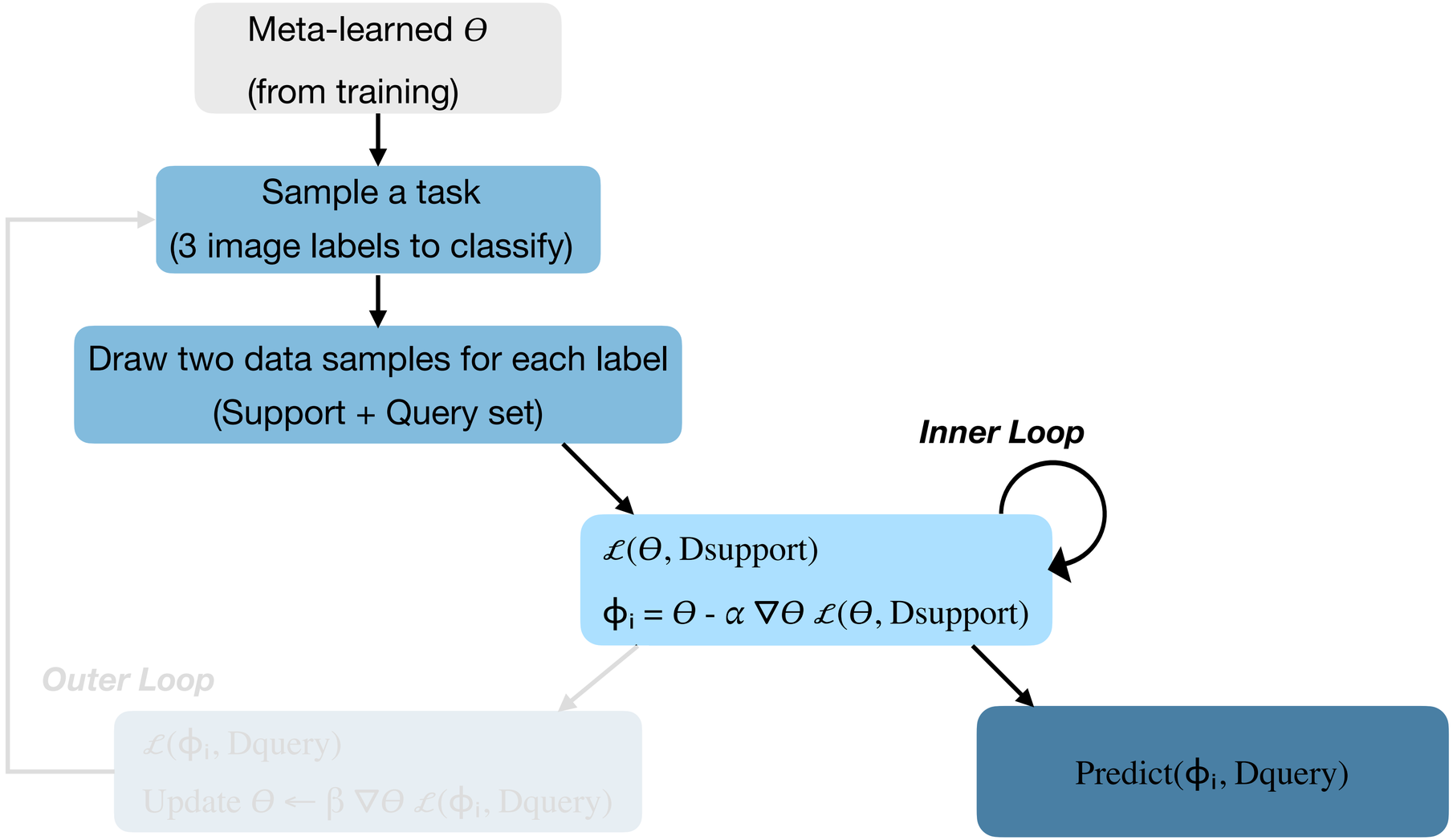
Lifecycle of MAML

Meta-training phase



Lifecycle of MAML

Meta-testing phase

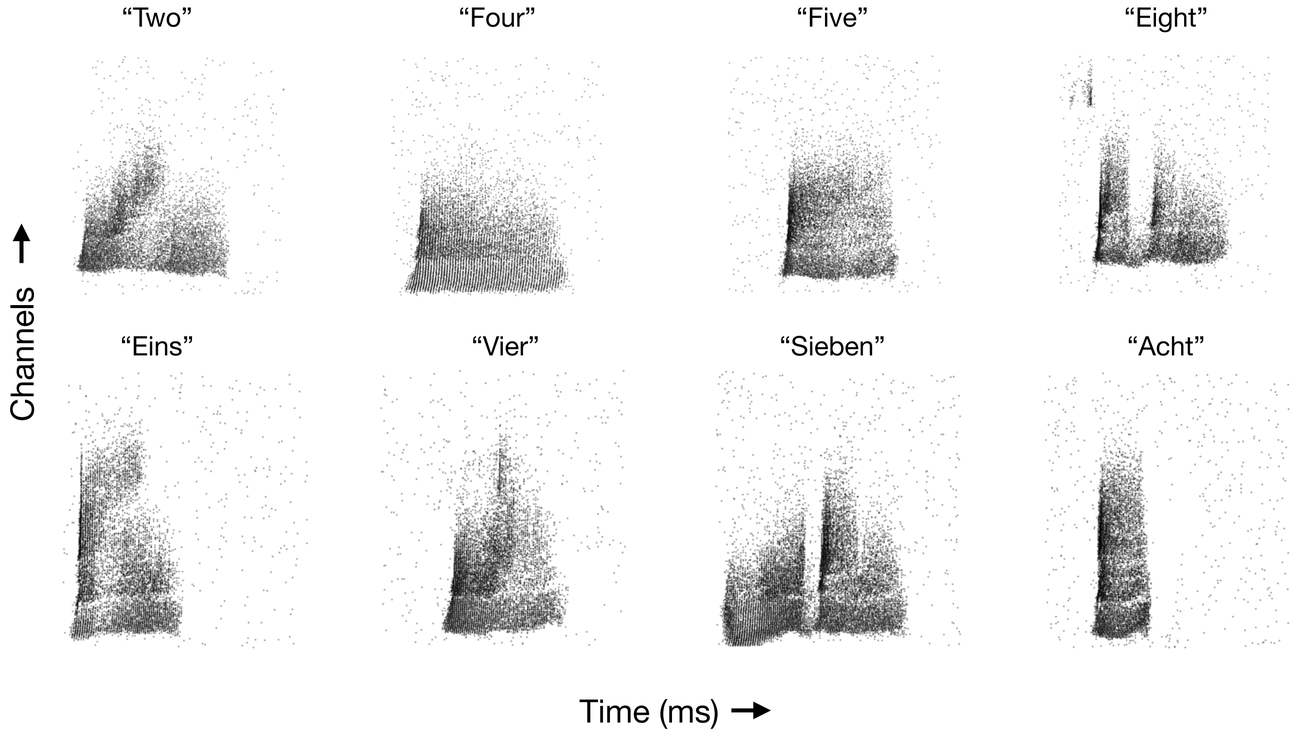


Dataset

Spiking Heidelberg Digits

- Audio-based classification dataset
- 700 frequency channels, 100 time steps
- “Zero”, “One”, “Two”, …. , “Nine”

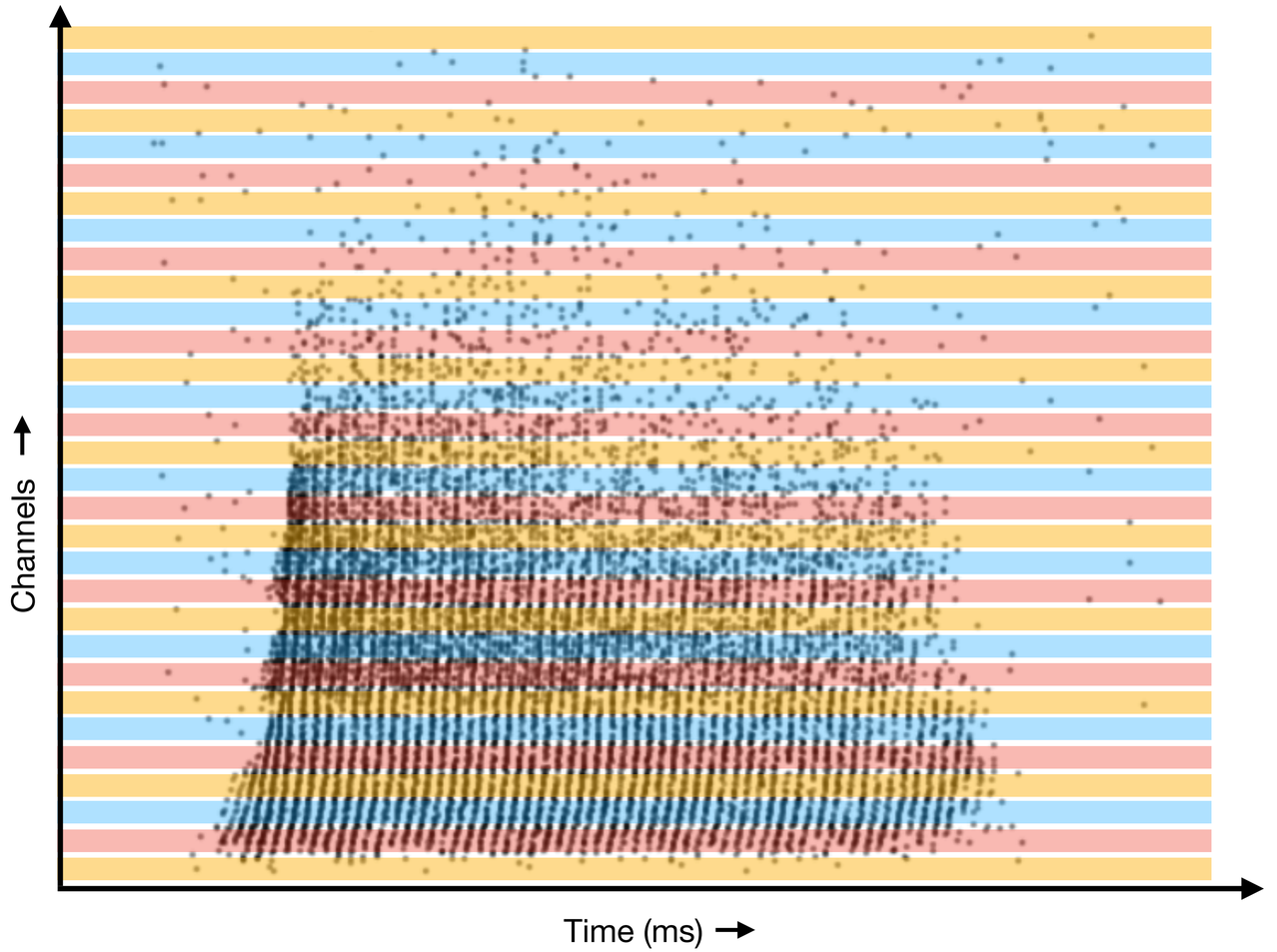
  “Null”, “Eins”, “Zwei”, ….. , “Neun”
- 12 speakers in total (*2 speakers exclusive to the testing set*)
- Class-balanced; 8k training and 2k testing samples



Cramer, B., Stradmann, Y., Schemmel, J., & Zenke, F. (2020). The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), 2744-2757.

Data augmentation

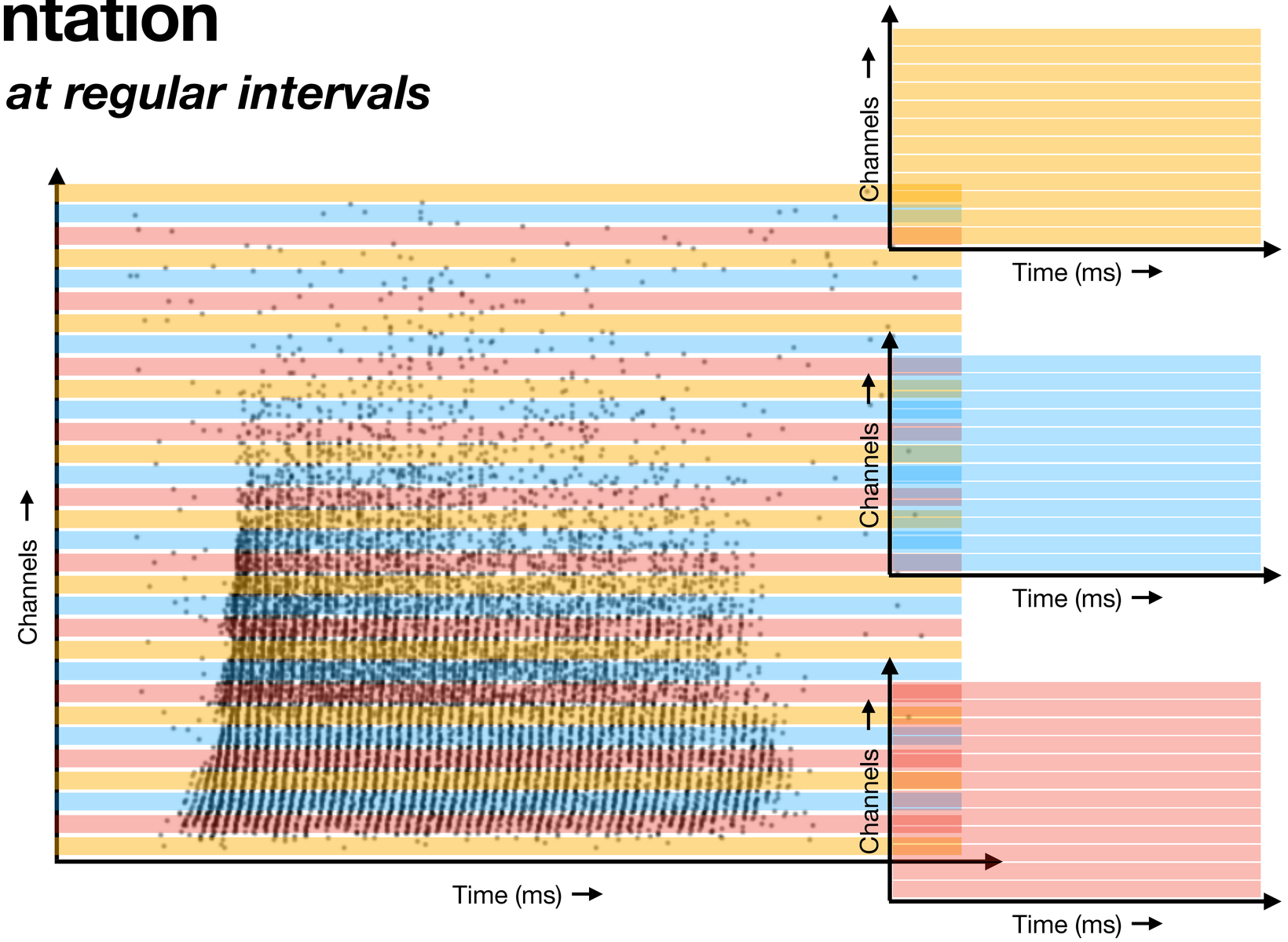
Spatial sampling at regular intervals



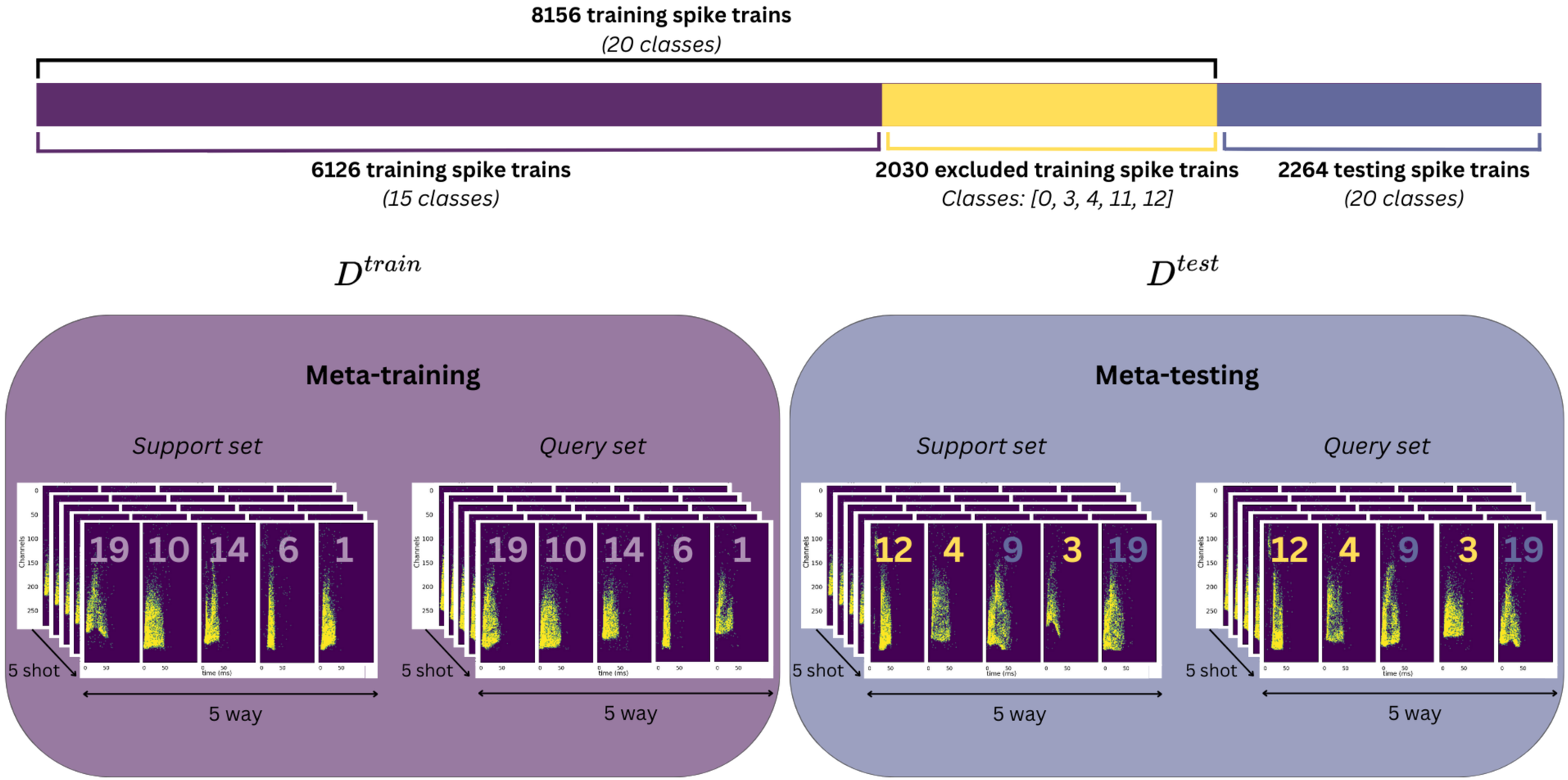
** Not to scale*

Data augmentation

Spatial sampling at regular intervals

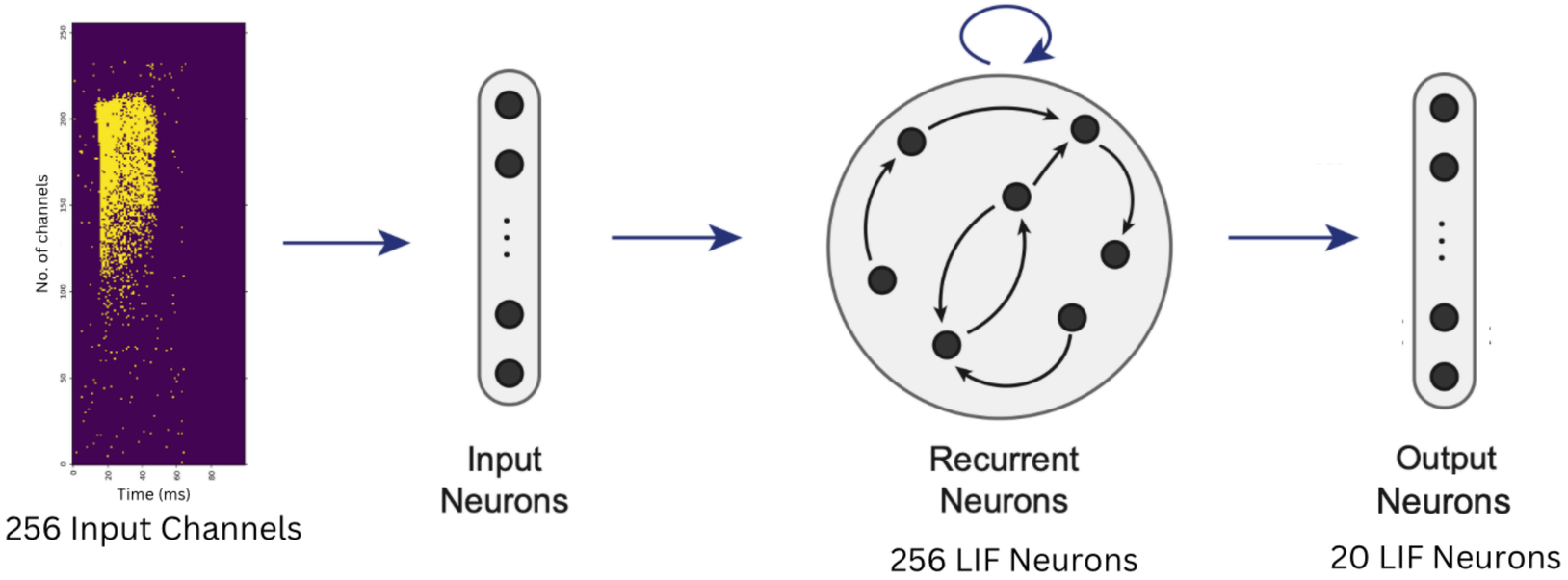


Converting the SHD into a FSL dataset



Model Architecture

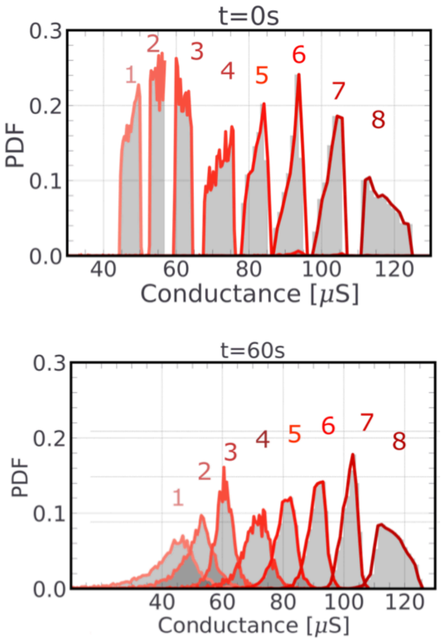
- Layer Seque
- Inner loop: S
- Outer loop: A
- Softmax Cro
- BPTT using s
  compatible w



Hardware-aware training

Novel energy efficient architecture

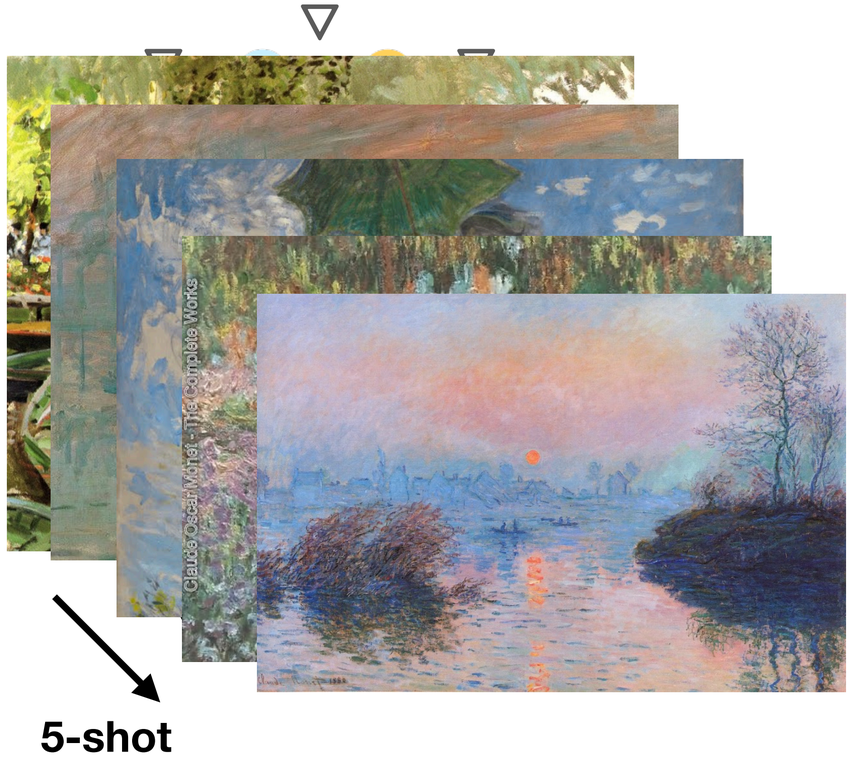
Meta-learning



Hardware-aware training

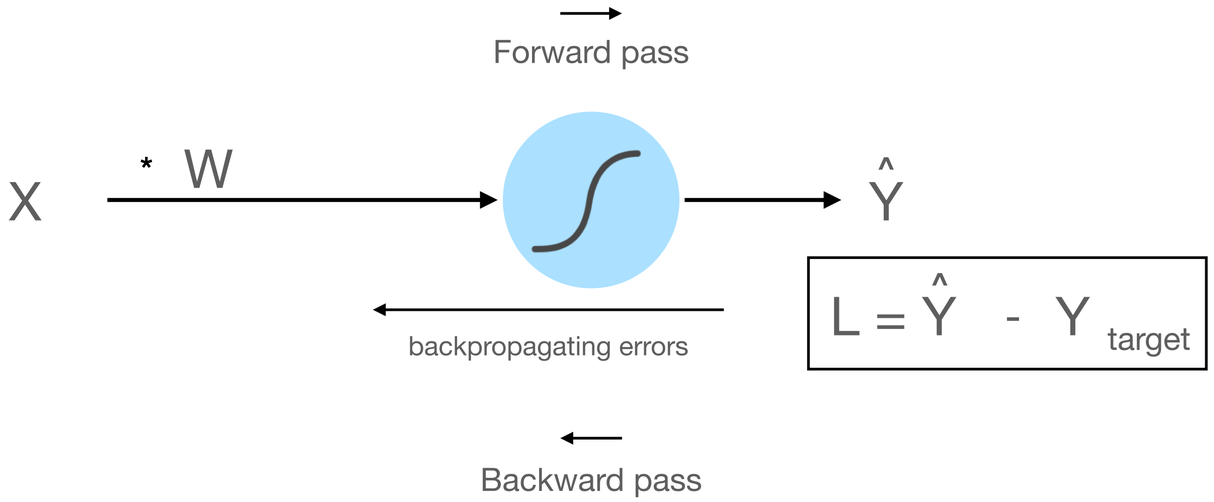
Hardware based constraints

- Update only the output layer weights in the inner loop
- One inner loop gradient step
- 1-shot learning is ideal
- A very high inner loop learning rate

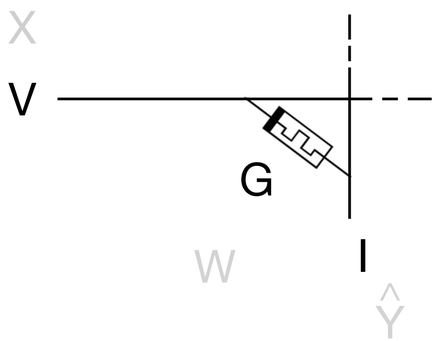


RRAM-aware training

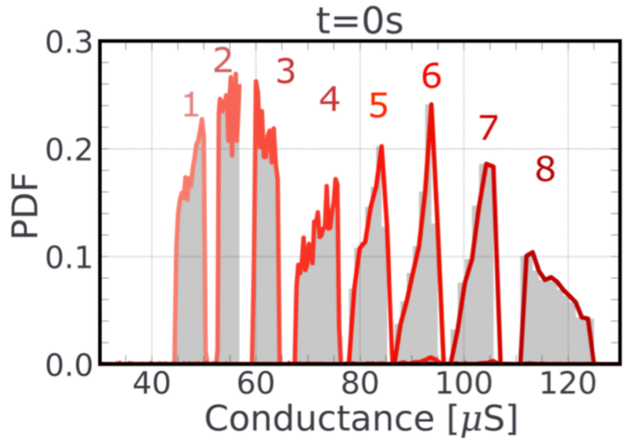
Learning the weights in a Neural network

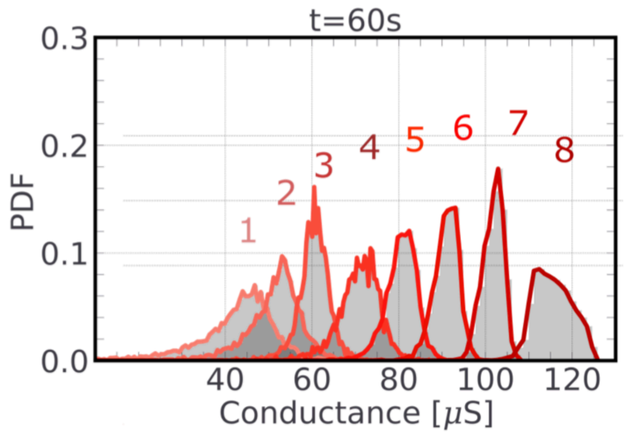


Operation on Hardware



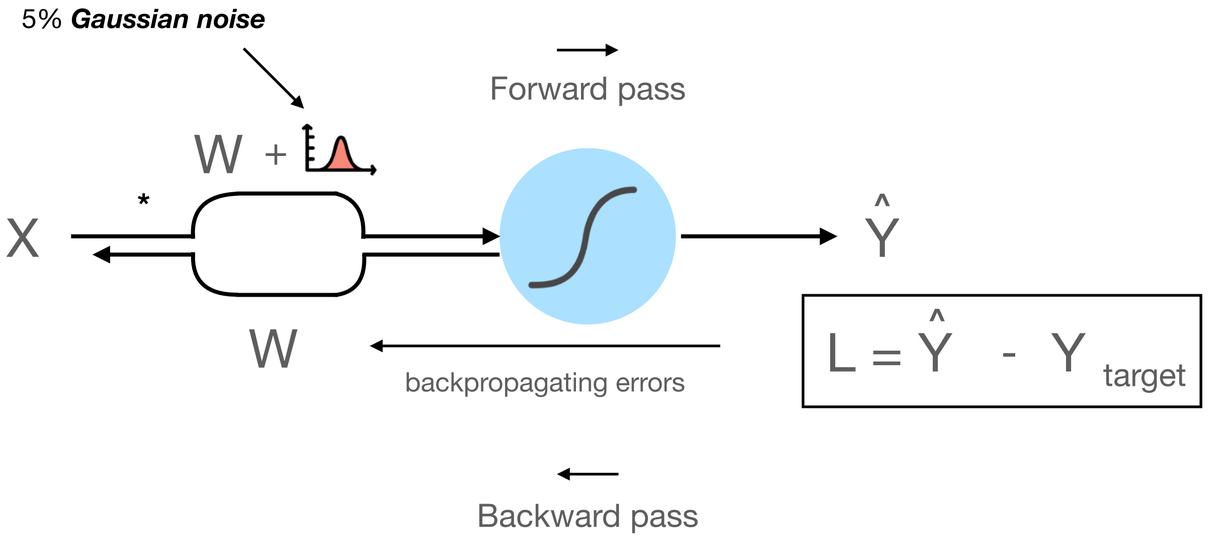
Conductance drift on RRAMs



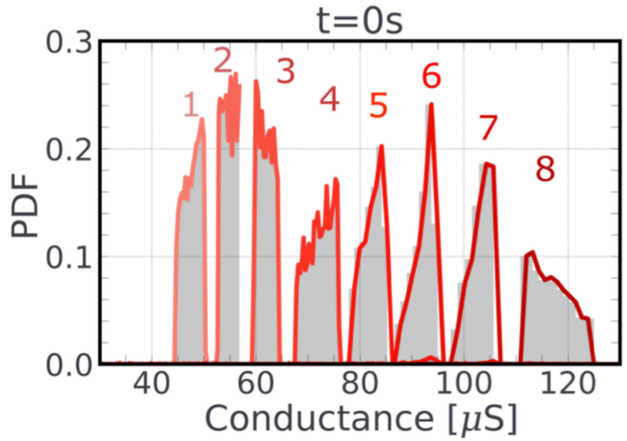


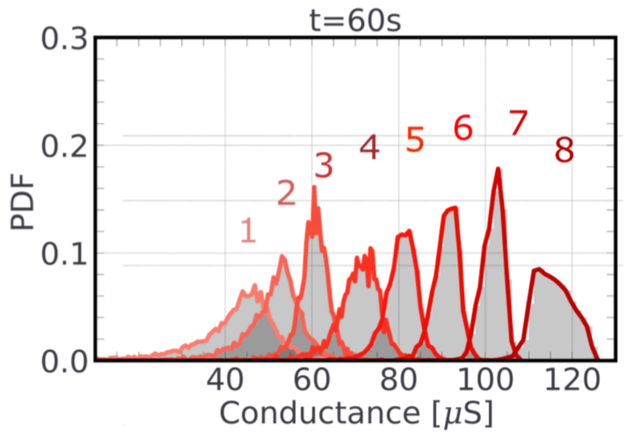
Moro, F., Esmanhotto, E., Hirtzlin, T., Castellani, N., Trabelsi, A., Dalgaty, T., ... & Vianello, E. (2022, May). Hardware calibrated learning to compensate heterogeneity in analog RRAM-based Spiking Neural Networks. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 380-383). IEEE.

RRAM-aware training



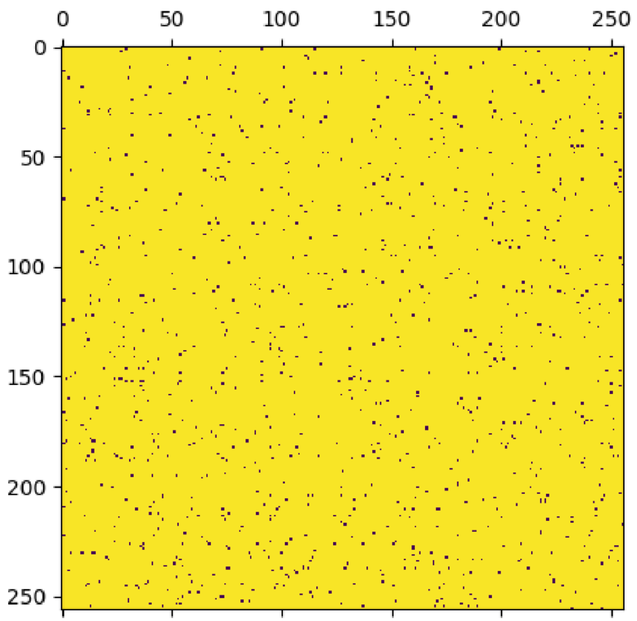
Conductance drift on RRAMs



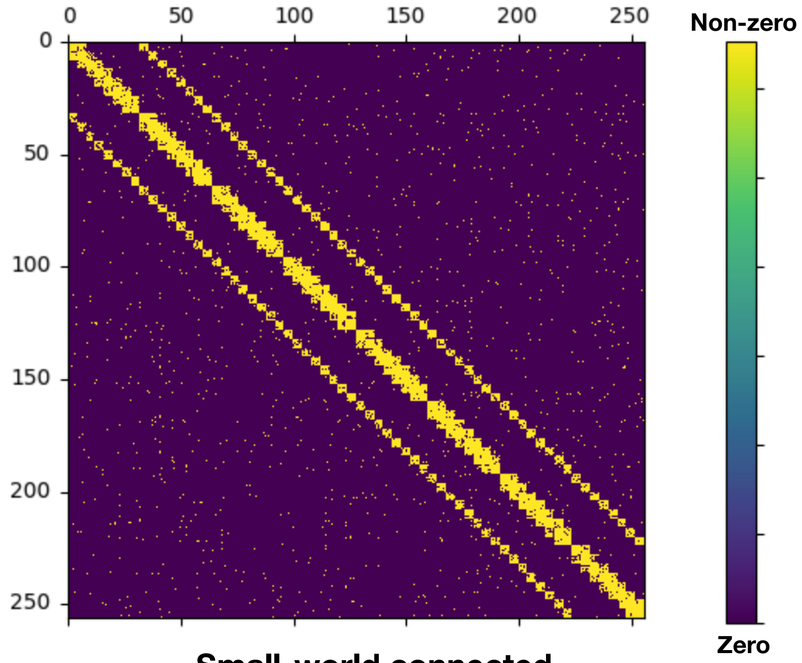


Moro, F., Esmanhotto, E., Hirtzlin, T., Castellani, N., Trabelsi, A., Dalgaty, T., ... & Vianello, E. (2022, May). Hardware calibrated learning to compensate heterogeneity in analog RRAM-based Spiking Neural Networks. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 380-383). IEEE.

Mosaic layout-aware training



Random fully connected

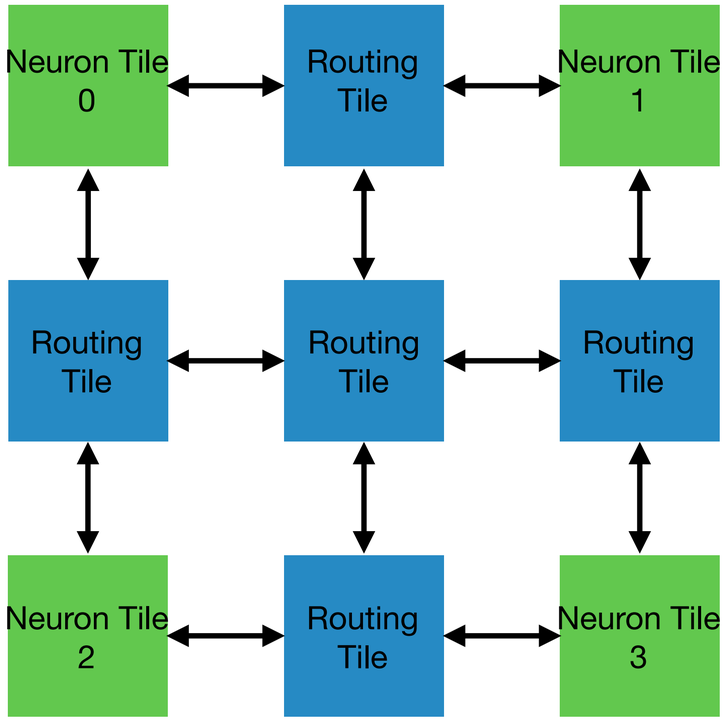


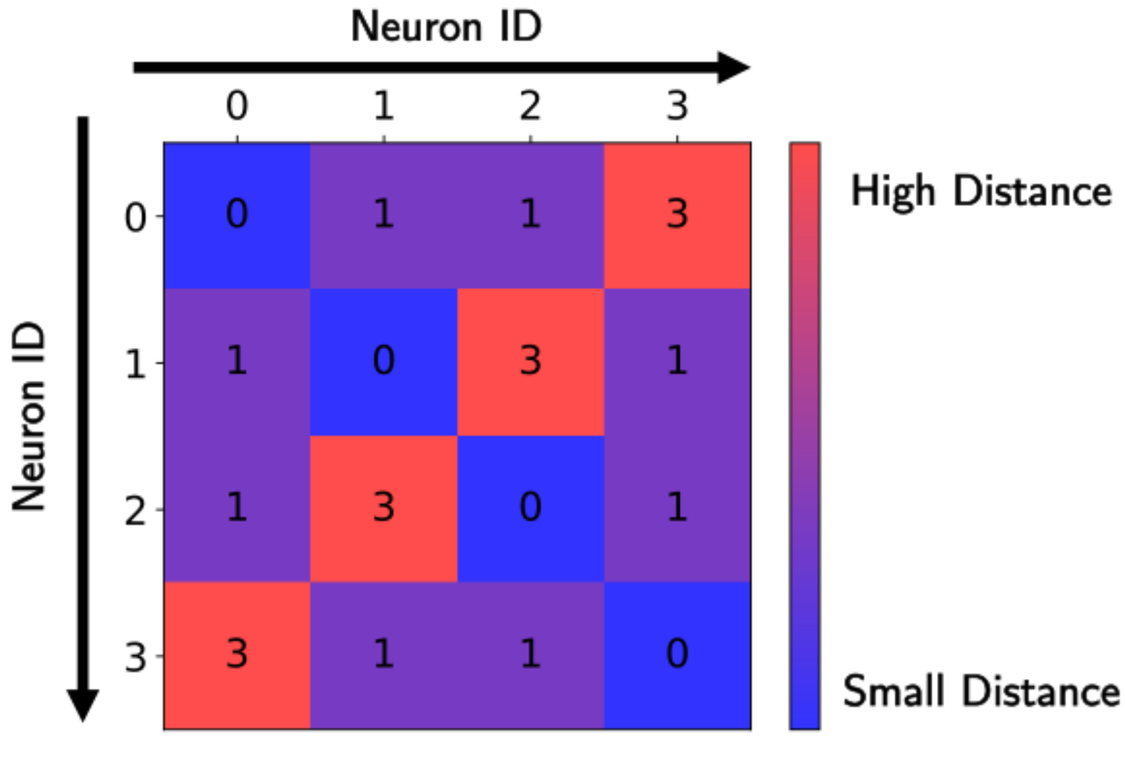
Small-world connected

Dalgaty, T., Moro, F., Demirağ, Y., De Pra, A., Indiveri, G., Vianello, E., & Payvand, M. (2024). Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems. *Nature Communications*, *15*(1), 142.

Calculating the layout-constraint mask

Example of matrix H

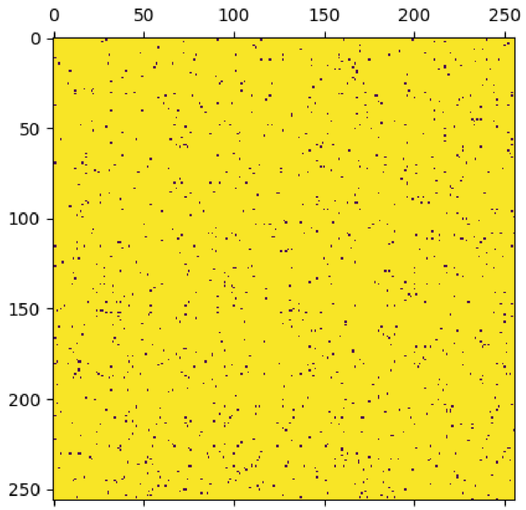




Dalgaty, T., Moro, F., Demirağ, Y., De Pra, A., Indiveri, G., Vianello, E., & Payvand, M. (2024). Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems. *Nature Communications*, *15*(1), 142.

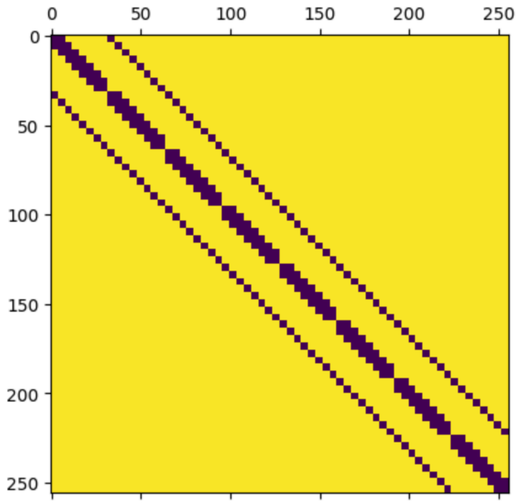
Mosaic layout-aware training

$$S = e^{\beta H} - 1$$



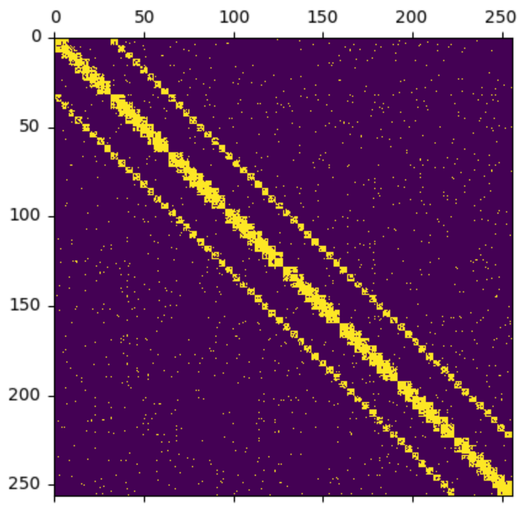
Random fully connected

$\oplus$

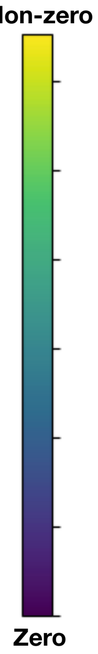


Layout-constraint mask

=



Small-world connected



$$L_M = \sum S \odot W^2$$

Total loss $\longrightarrow$ $L = L_{CE} + \lambda L_M$ $\longleftarrow$ Layout loss

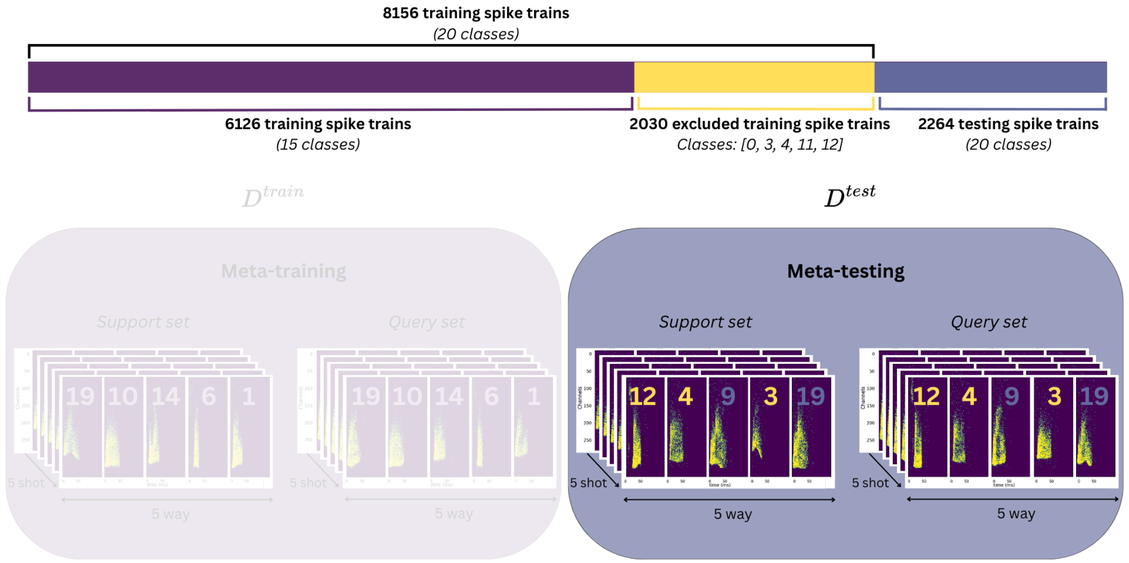
Cross entropy loss

Dalgaty, T., Moro, F., Demirağ, Y., De Pra, A., Indiveri, G., Vianello, E., & Payvand, M. (2024). Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems. *Nature Communications*, *15*(1), 142.

Results

Experimental Setup

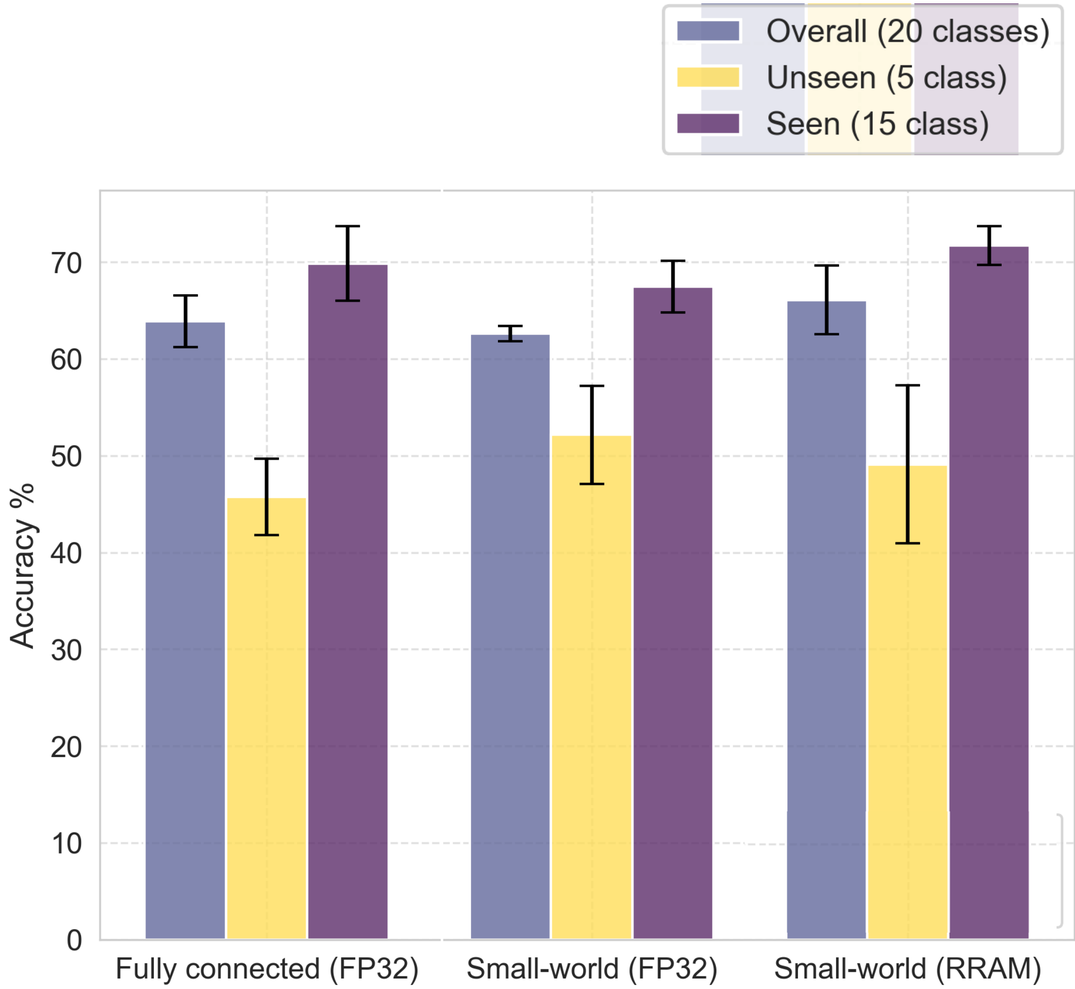
- 1 layer SRNN with 256 LIF neurons (64 neurons for fully-connected baseline)
- Trained on 15 classes (5 excluded classes); Tested on all 20 classes
- 5 way 5 shot tasks
- One inner loop update
- Report the accuracy on seen and unseen classes separately



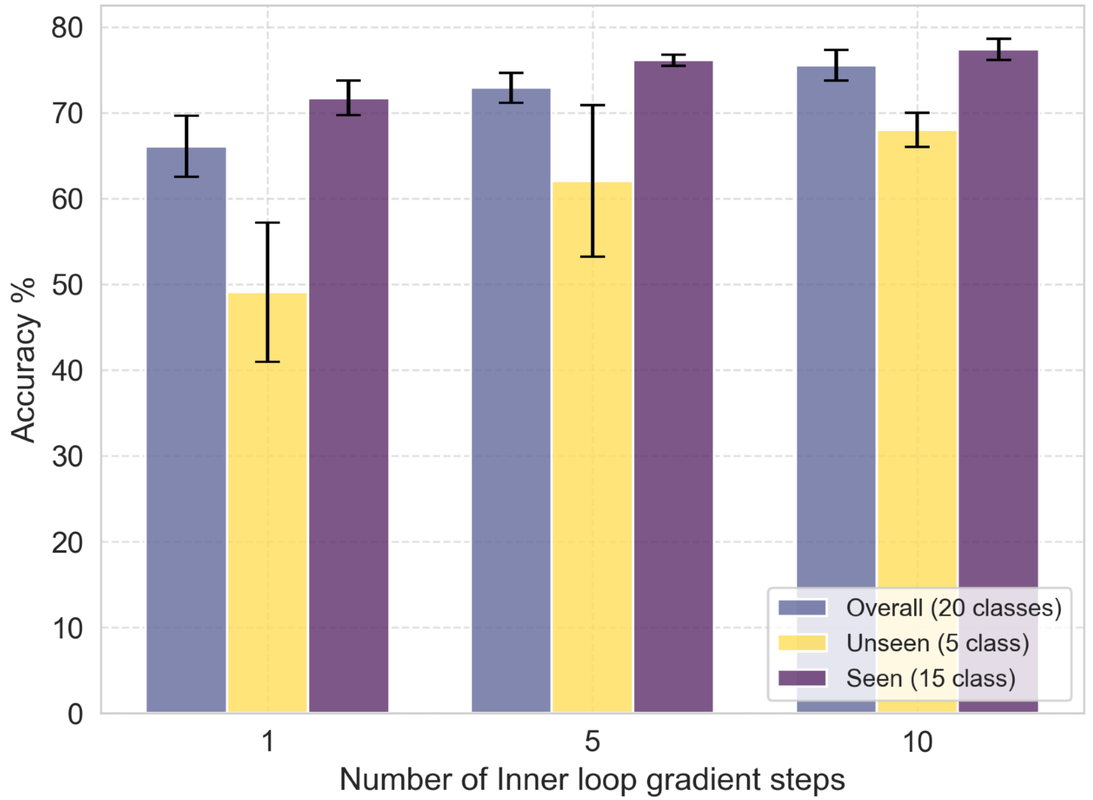
Constrained network architecture does *not* reduce the performance of the model!

(Compared to a fully connected network)

- 70% accuracy, standard SHD benchmark for 1 layer SRNN
- Increase in performance on unseen classes (attributed to sparse dist. of parameters)
- Best performance on seen classes (attributed to noise aware training)



Increasing inner loop gradient steps increases performance!



Discussion

Discussion

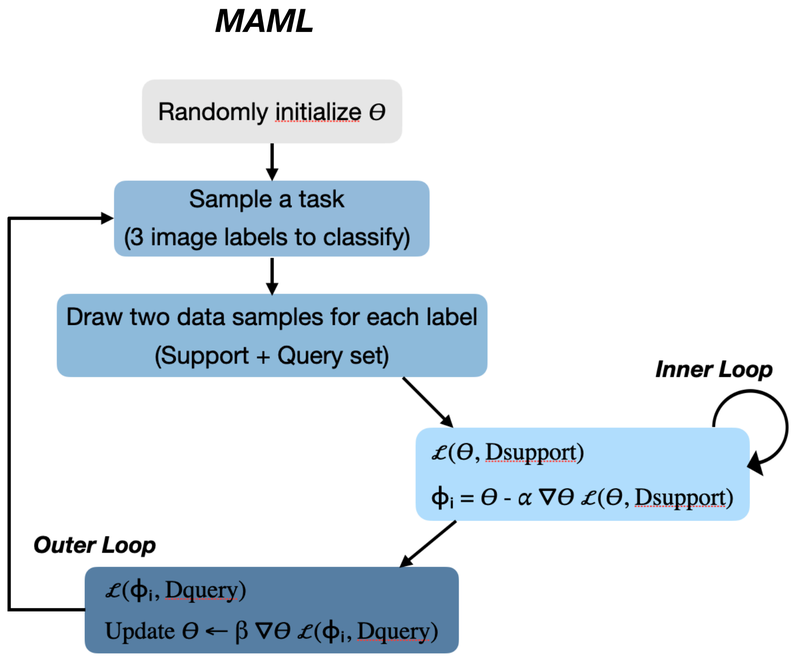
Interesting relationship between learning rates and no. of inner loop updates

➡ 1 inner loop gradient step

Config parameter	Importance
lr_out	
Runtime	
lr_in	

➡ 10 inner loop gradient steps

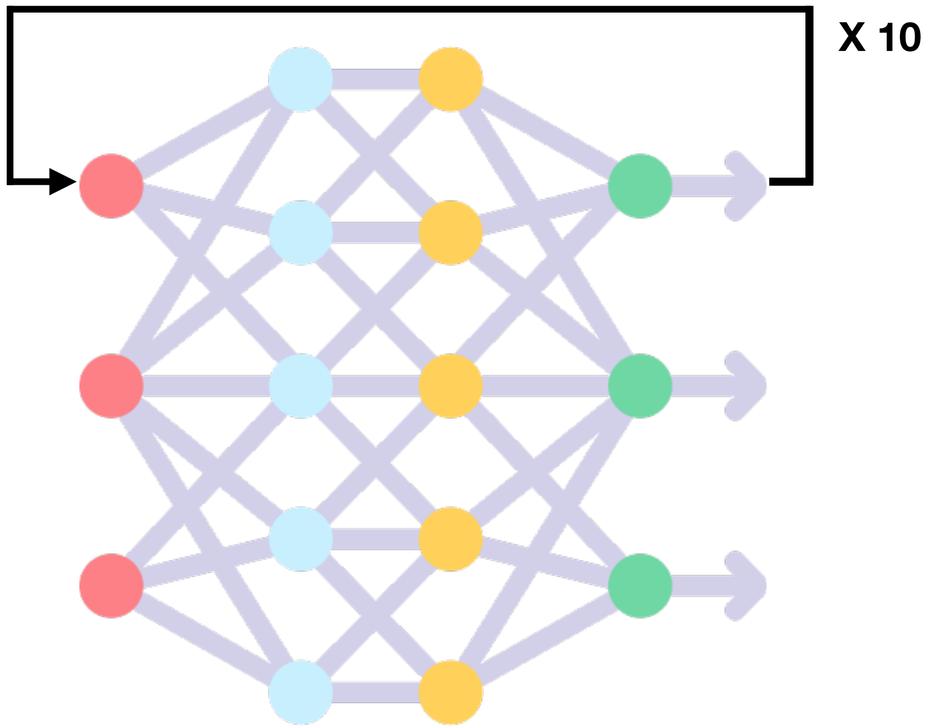
Config parameter	Importance
lr_in	
Runtime	
lr_out	



https://docs.wandb.ai/guides/app/features/panels/parameter-importance.

Discussion

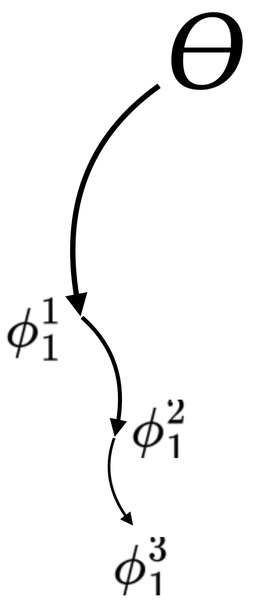
Can re-training compensate for multiple gradient steps?



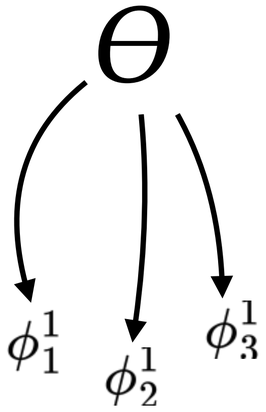
Discussion

Can re-training compensate for multiple gradient steps?

3 inner loop updates

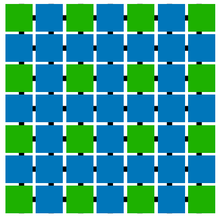


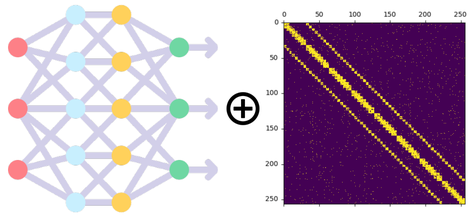
Re-training

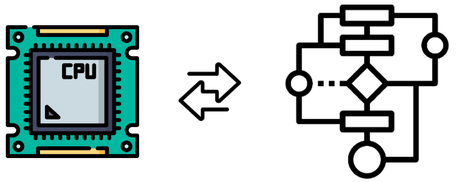


Take home message

- Demonstrated (using simulations) Online few-shot learning on the Mosaic architecture
- Constraining network architecture does not terminally detriment the performance of the model
- Importance of inner loop dynamics in MAML; Highlights the need for HW-SW co-design for efficient







Acknowledgements

Dr. Melika Payvand
Institute of Neuroinformatics

Yigit Demirag
Institute of Neuroinformatics

Prof. Dr. Herbert Jaeger
University of Groningen

Dr. Emre Nefci
Forschungszentrum Jülich

Emerging Intelligent Substrates lab

This work is supported by Horizon Europe **METASPIN** project (grant number 101098651) and the SNSF Starting Grant Project UNITE (TMSGI2- 211461)