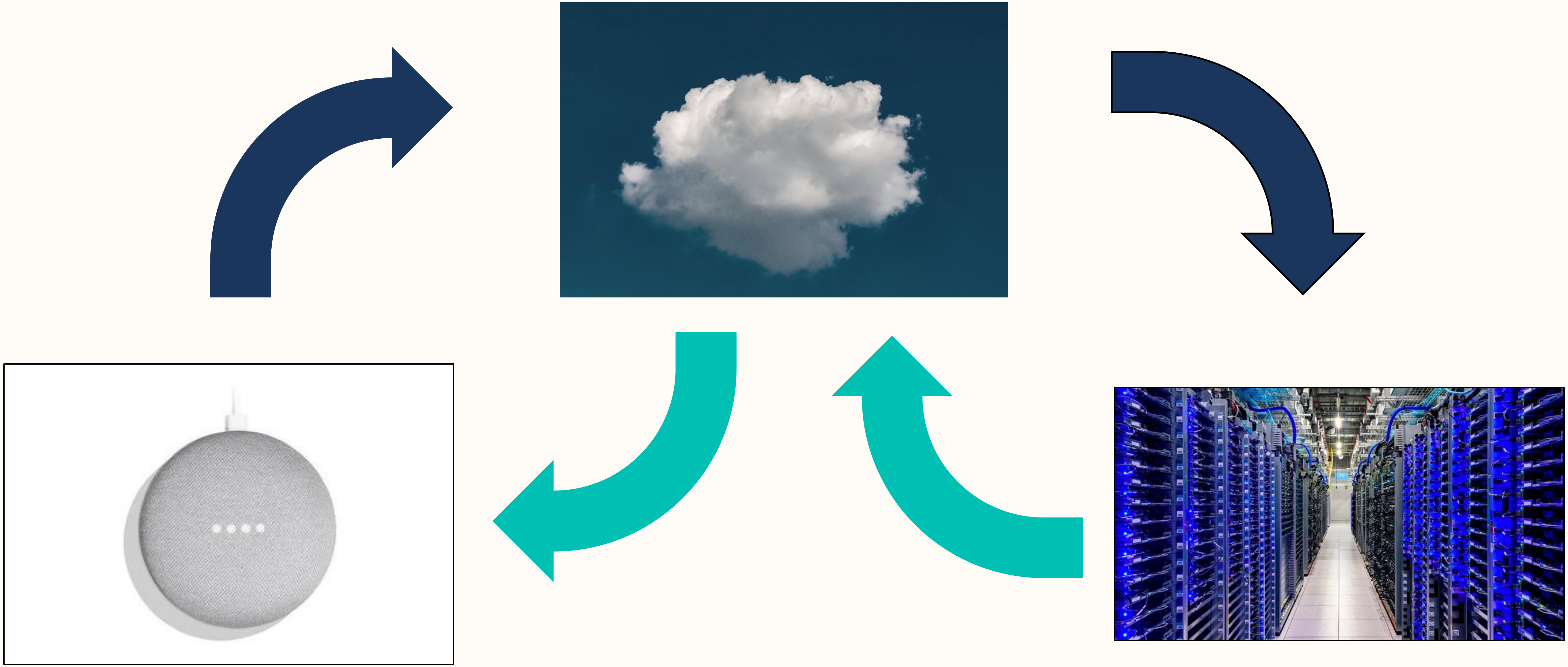
EventProp training for efficient neuromorphic applications

Thomas Shoesmith, James C. Knight, Balazs Meszaros,
Jonathan Timcheck, and Thomas Nowotny

Key Word Spotting (KWS)



GeNN & EventProp

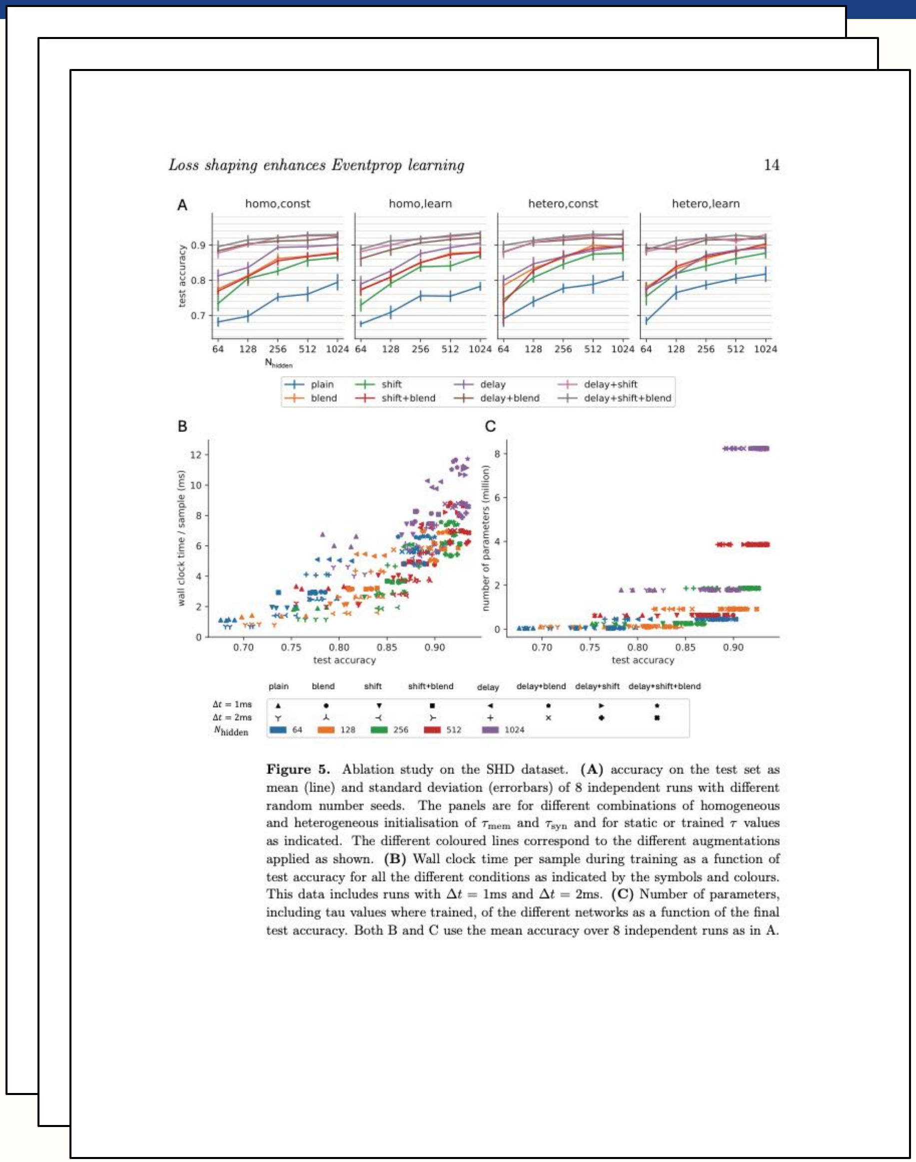
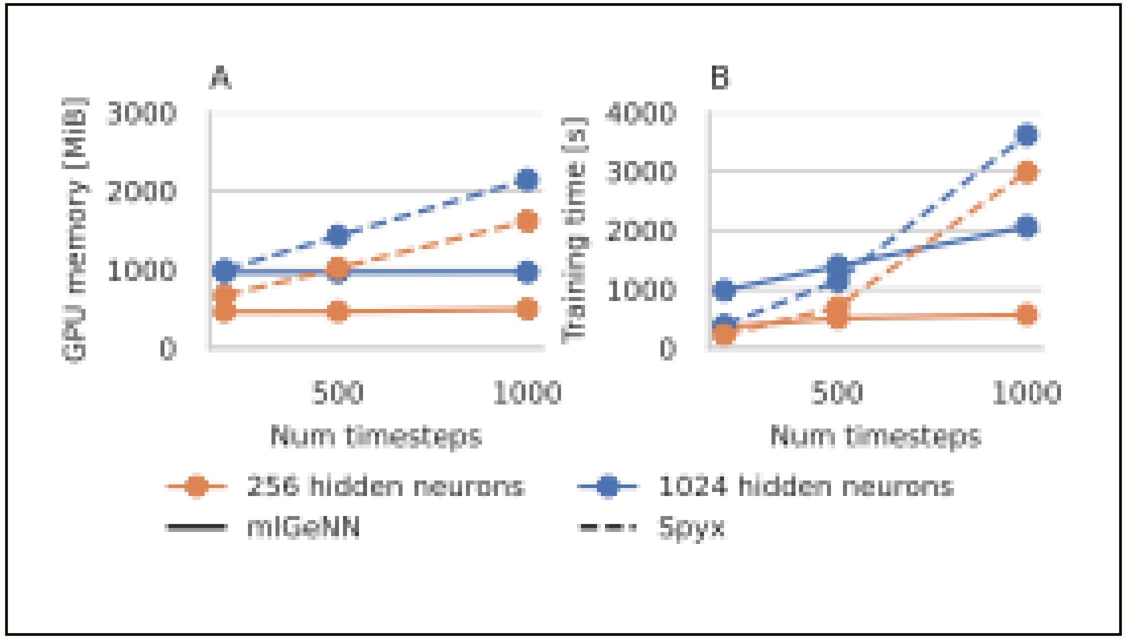


Figure 5. Ablation study on the SHD dataset. **(A)** accuracy on the test set as mean (line) and standard deviation (errorbars) of 8 independent runs with different random number seeds. The panels are for different combinations of homogeneous and heterogeneous initialisation of τ_{mem} and τ_{syn} and for static or trained τ values as indicated. The different coloured lines correspond to the different augmentations applied as shown. **(B)** Wall clock time per sample during training as a function of test accuracy for all the different conditions as indicated by the symbols and colours. This data includes runs with Δt = 1ms and Δt = 2ms. **(C)** Number of parameters, including tau values where trained, of the different networks as a function of the final test accuracy. Both B and C use the mean accuracy over 8 independent runs as in A.

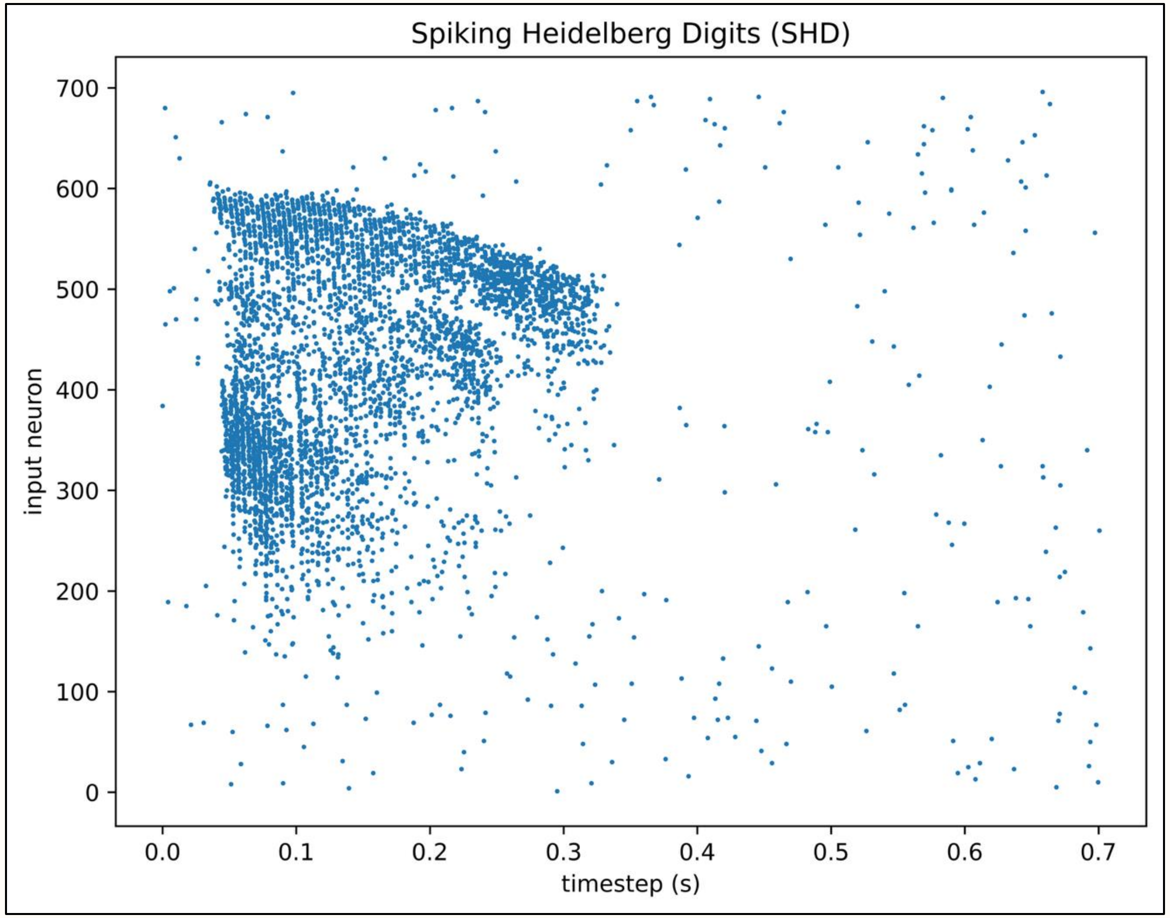
mlGeNN

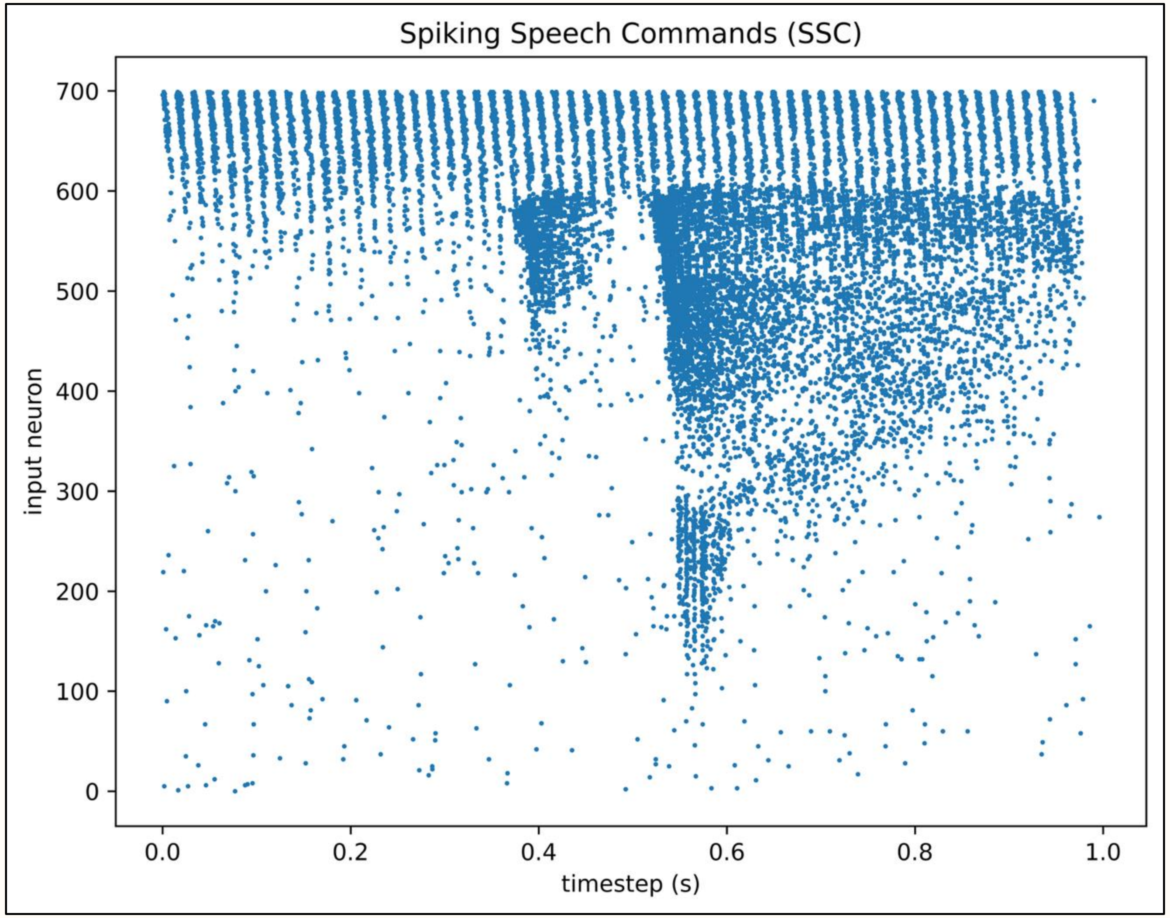
- A SNN Simulator which can be run on conventional hardware.
- Can achieve state of the art performance
- EventProp Implementation, which takes advantage of sparse activities
- Optimised memory usage



T. Nowotny, J. P. Turner, and J. C. Knight, “Loss shaping enhances exact gradient learning with EventProp in Spiking Neural Networks,” en, Jun. 2024.
T. C. Wunderlich and C. Pehle, “Event-based backpropagation can compute exact gradients for spiking neural networks,” Scientific Reports, vol. 11, no. 1, 2021

Datasets





B. Cramer, Y. Stradmann, J. Schemmel, and F. Zenke, "The Heidelberg Spiking Data Sets for the Systematic Evaluation of Spiking Neural Networks," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–14, 2022. DOI: 10.1109/TNNLS.2020.3044364.

Building the pipeline

GeNN

NetX
(HDF5)

NxKernel

https://github.com/genn-team/ml_genn_netx

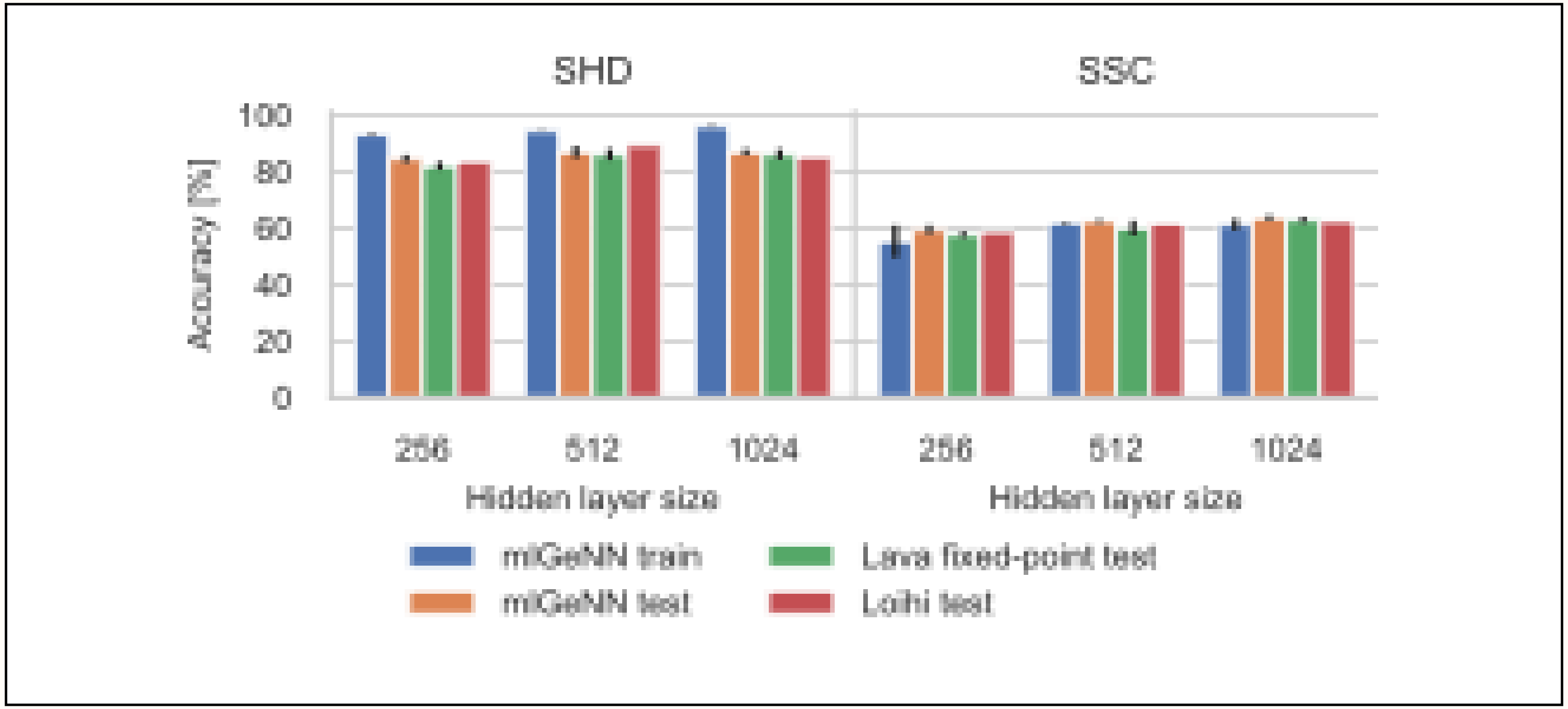
GeNN vs Lava / NxKernel

<u>LIF Neuron Characteristics</u>

GeNN	Lava / NxKernel
$I(t+\Delta t) = \alpha_s I(t) + \sum_j w_{ij}\, S_i(t)$	$I(t+\Delta t) = \alpha_s I(t) + \sum_j w_{ij}\, S_i(t)$
$V(t+\Delta t) = \alpha_m V(t) + \boxed{(1 - \alpha_m)} I(t+1)$	$V(t+\Delta t) = \alpha_m V(t) + I(t+1)$

- Weights to be scaled by $1 - \alpha$
- LIF voltage thresholds to be scaled
- Post Training Quantisation of weights (32bit to 8bit)

Results - Accuracy



Results - Energy

SPIKING HEIDELBERG DIGITS CLASSIFICATION PERFORMANCE COMPARISON.

Hidden size	Weight precision	Hardware	Hardware inference cost per sample			
			Energy		Latency	EDP
			Total [mJ]	Dynamic [mJ]	[ms]	[µJ × s]
256	fp32	Jetson Orin Nano GPU (batch=1)‡	82.3	23.8	25.5	2102.0
512	fp32	Jetson Orin Nano GPU (batch=1)‡	85.0	25.3	26.0	2214.8
1024	fp32	Jetson Orin Nano GPU (batch=1)‡	92.0	29.1	27.4	2522.5
256	fp32	Jetson Orin Nano GPU (batch=128)‡	3.7	1.9	103.3	384.6
512	fp32	Jetson Orin Nano GPU (batch=128)‡	6.5	3.7	156.4	1018.2
1024	fp32	Jetson Orin Nano GPU (batch=64)‡	15.5	9.7	161.2	2506.6
256	int8	Loihi 2†	0.19	0.11	2.33	0.44
512	int8	Loihi 2†	0.27	0.15	2.37	0.63
1024	int8	Loihi 2†	0.50	0.31	2.56	1.29

Loihi 2 vs Jetson (batch size of 128)
~24x lower total energy
~1,616x lower Energy Delayed Product

Loihi 2 vs Jetson (batch size of 1)
~315x lower total energy
~3,516x lower Energy Delayed Product

Conclusion

- Maintaining similar performance after conversion.
- A successful pipeline for training rSNN using EventProp in ml-GeNN to be deployed on Intel’s Loihi 2
- Future work intended to be being done on delays.
- Exploring more advanced quantised approaches.

A. Patino-Saucedo et al., “Hardware-aware training of models with synaptic delays for digital event-driven neuromorphic processors,” arXiv preprint arXiv:2404.10597, 2024.

Acknowledgments

James C. Knight, Balazs Meszaros, Jonathan Timcheck, and Thomas Nowotny

This work was funded by the EPSRC, grants EP/V052241/1 and EP/S030964/1; and Intel Corporation. TS is also funded by a PhD studentship from the University of Sussex.

Compute time was provided through Gauss Centre for Supercomputing application number 21018 and EPSRC (grant number EP/T022205/1) and local GPU hardware was provided by an NVIDIA hardware grant award

Thank You!

&

Any Questions?