Efficient GPU training of SNNs

James Knight & Thomas Nowotny

UNIVERSITY OF SUSSEX

Engineering and Physical Sciences Research Council

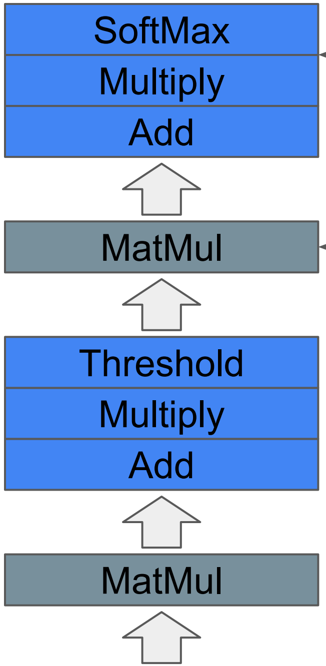
Spiking Neural Network acceleration

- Comp Neuro
  - Long history of simulators for efficiently processing sparse connectivity and activity
  - Focus on simulating single instances of (potentially very) large models
  - Historically, distributed CPU was platform of choice
- Neuromorphic hardware
  - Potential 1000-1000000x energy saving compared to standard hardware [1]
  - On-chip learning still challenging
  - Real-time isn't fast enough for training with current data-hungry algorithms?
- ML using SNNs
  - PyTorch/TensorFlow/JAX used for GPU acceleration by treating SNN as RNNs
  - Auto diff + surrogate gradients/spike times used to train SNNs using BPTT

Frenkel, C., Bol, D., & Indiveri, G. (2021). Bottom-Up and Top-Down Neural Processing Systems Design: Neuromorphic Intelligence as the Convergence of Natural and Artificial Intelligence. 1–25. Retrieved from http://arxiv.org/abs/2106.01288

SNNs as computational graphs

Naive

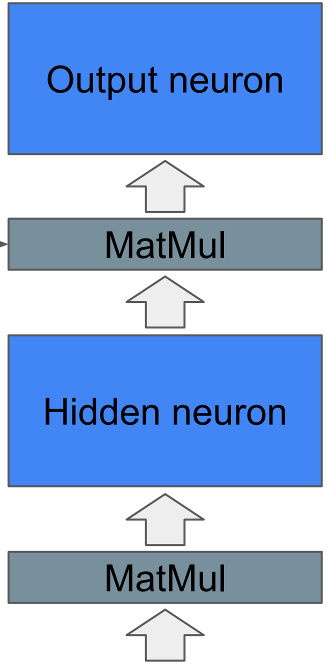


= 8 kernel launches ≈ 40μs latency

Unnecessary global memory accesses

Spike propagation is **not** matrix multiplication — *theoretically* much cheaper computationally

With JIT/custom kernels



= 4 kernel launches ≈ 20μs latency

GeNN

- C++ library for generating SNN simulation code
- Backends to generate CUDA, OpenCL and C++ code
- All features available from Python
- Past focus on Computational Neuroscience and Neurorobotics
- Maximal user control

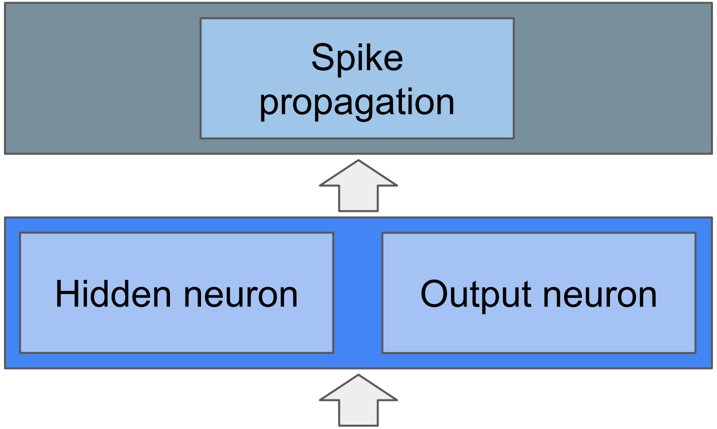
https://genn-team.github.io/genn/

Knight, J. C., & Nowotny, T. (2018). GPUs Outperform Current HPC and Neuromorphic Solutions in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model. Frontiers in Neuroscience, 12(December), 1–19. https://doi.org/10.3389/fnins.2018.00941

Knight, J. C., & Nowotny, T. (2021). Larger GPU-accelerated brain simulations with procedural connectivity. Nature Computational Science, 1(2), 136–142. https://doi.org/10.1038/s43588-020-00022-7

Knight, J. C., Komissarov, A., & Nowotny, T. (2021). PyGeNN: A Python Library for GPU-Enhanced Neural Networks. Frontiers in Neuroinformatics, 15(April). https://doi.org/10.3389/fninf.2021.659005

SNNs in GeNN



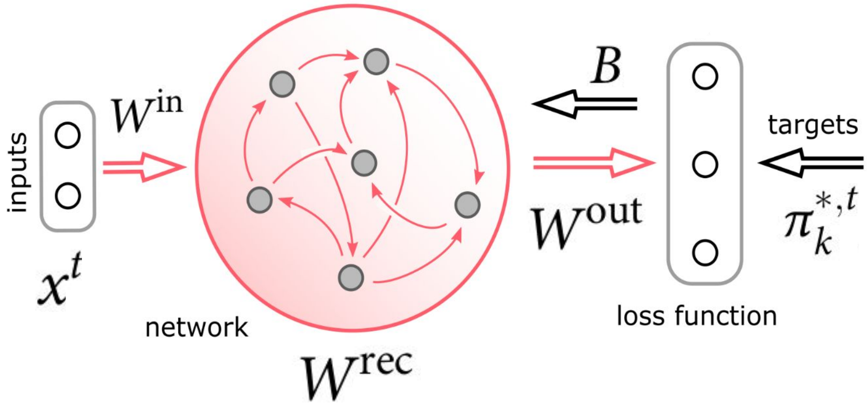
= 2 kernel launches ≈ 10μs latency “pipelined”

Spike transmission isn’t instantaneous

- Breaks dependencies - model doesn’t need to be a directed acyclical graph
- All neuron and synapse updates can be *fused* [1]

1. Knight, J. C., & Nowotny, T. (2021). Larger GPU-accelerated brain simulations with procedural connectivity. Nature Computational Science, 1(2), 136–142. https://doi.org/10.1038/s43588-020-00022-7

Training SNNs with symmetric eProp

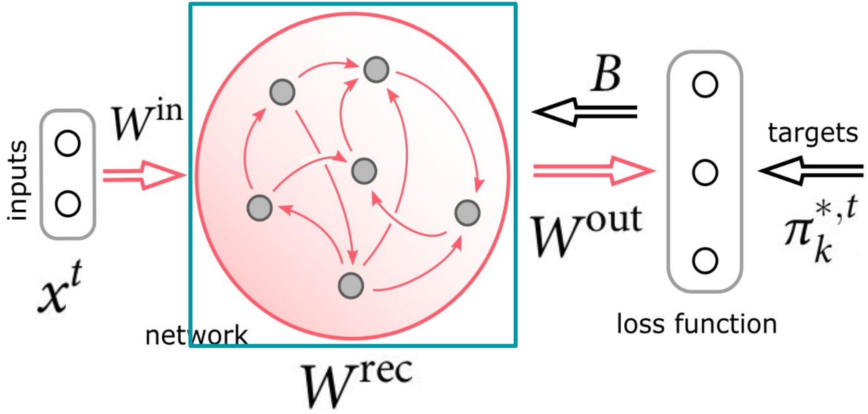


Zenke, F., & Neftci, E. O. (2021). Brain-Inspired Learning on Neuromorphic Substrates. Proceedings of the IEEE, 1–16. https://doi.org/10.1109/JPROC.2020.3045625

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, *11*(1), 3625. https://doi.org/10.1038/s41467-020-17236-y

Training SNNs with symmetric eProp

LIF neuron with adaptation and relative reset

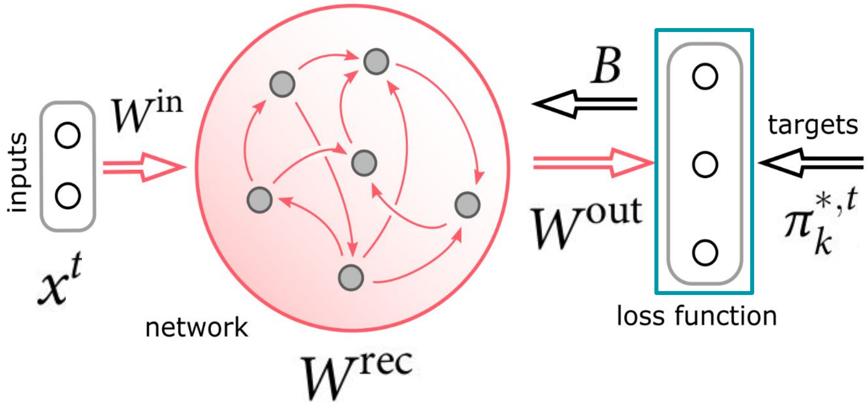


Zenke, F., & Neftci, E. O. (2021). Brain-Inspired Learning on Neuromorphic Substrates. Proceedings of the IEEE, 1–16. https://doi.org/10.1109/JPROC.2020.3045625

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, *11*(1), 3625. https://doi.org/10.1038/s41467-020-17236-y

Training SNNs with symmetric eProp

Non-spiking
output neuron
with trainable bias



Softmax π_k^t calculated
from membrane voltage

Zenke, F., & Neftci, E. O. (2021). Brain-Inspired Learning on Neuromorphic Substrates. Proceedings of the IEEE, 1–16. https://doi.org/10.1109/JPROC.2020.3045625

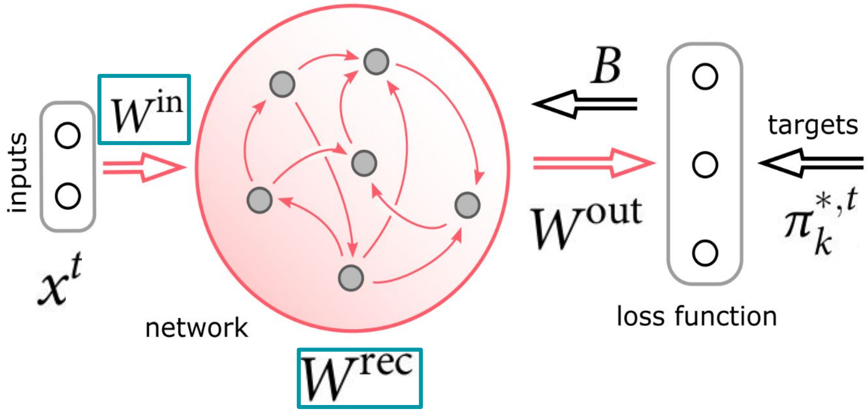
Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, *11*(1), 3625. https://doi.org/10.1038/s41467-020-17236-y

Training SNNs with symmetric eProp

Per-synapse eligibility traces and supervised learning rule

$$\epsilon_{ji,a}^{t+1} = \psi_j^t \bar{z}_i^{t-1} + (\rho - \psi_j^t \beta)\epsilon_{ji,a}^t \quad e_{ji}^t = \psi_j^t \left(\bar{z}_i^{t-1} - \beta \epsilon_{ji,a}^t\right)$$

$$\Delta W_{ji}^{\text{rec}} = -\eta \sum_t \underbrace{\left(\sum_k B_{jk}(\pi_k^t - \pi_k^{*,t})\right)}_{=L_j^t} \bar{e}_{ji}^t.$$



Zenke, F., & Neftci, E. O. (2021). Brain-Inspired Learning on Neuromorphic Substrates. Proceedings of the IEEE, 1–16. https://doi.org/10.1109/JPROC.2020.3045625

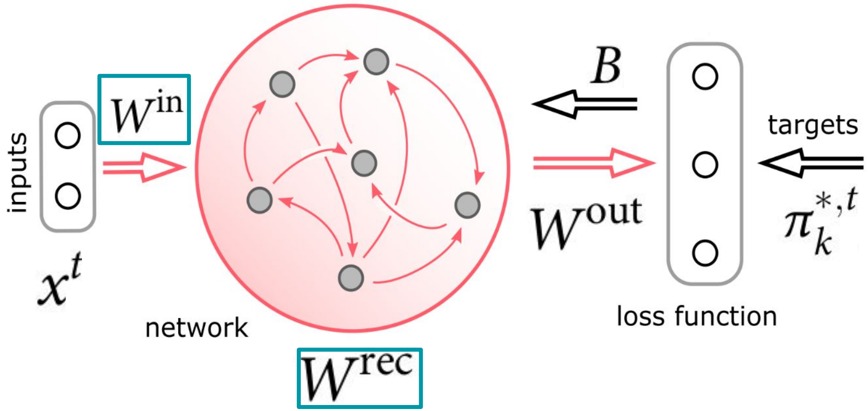
Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, *11*(1), 3625. https://doi.org/10.1038/s41467-020-17236-y

Training SNNs with symmetric eProp

Per-synapse eligibility traces and supervised learning rule

$$\epsilon_{ji,a}^{t+1} = \boxed{\psi_j^t}\bar{z}_i^{t-1} + (\rho - \boxed{\psi_j^t}\beta)\epsilon_{ji,a}^t \qquad e_{ji}^t = \boxed{\psi_j^t}\left(\bar{z}_i^{t-1} - \beta\epsilon_{ji,a}^t\right)$$

$$\Delta W_{ji}^{\mathrm{rec}} = -\eta \sum_t \underbrace{\left(\sum\nolimits_k B_{jk}(\pi_k^t - \pi_k^{*,t})\right)}_{=L_j^t} \bar{e}_{ji}^t.$$



Postsynaptic neuron surrogate gradient

Zenke, F., & Neftci, E. O. (2021). Brain-Inspired Learning on Neuromorphic Substrates. Proceedings of the IEEE, 1–16. https://doi.org/10.1109/JPROC.2020.3045625

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, *11*(1), 3625. https://doi.org/10.1038/s41467-020-17236-y

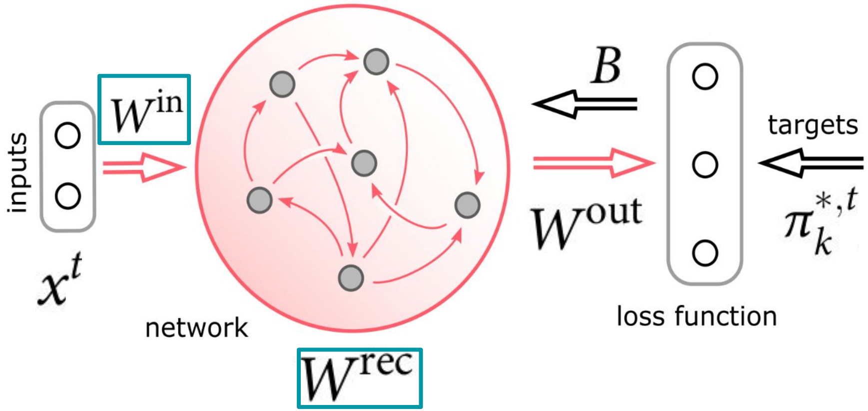
Training SNNs with symmetric eProp

Per-synapse eligibility traces and supervised learning rule

$$\epsilon^{t+1}_{ji,a} = \psi^t_j \boxed{\bar{z}^{t-1}_i} + (\rho - \psi^t_j \beta)\epsilon^t_{ji,a} \qquad e^t_{ji} = \psi^t_j \left(\boxed{\bar{z}^{t-1}_i} - \beta \epsilon^t_{ji,a} \right)$$

$$\Delta W^{\mathrm{rec}}_{ji} = -\eta \sum_t \underbrace{\left(\sum_k B_{jk} (\pi^t_k - \pi^{*,t}_k) \right)}_{=L^t_j} \bar{e}^t_{ji}.$$

Filtered presynaptic activity



Zenke, F., & Neftci, E. O. (2021). Brain-Inspired Learning on Neuromorphic Substrates. Proceedings of the IEEE, 1–16. https://doi.org/10.1109/JPROC.2020.3045625

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, *11*(1), 3625. https://doi.org/10.1038/s41467-020-17236-y

Extensions to GeNN

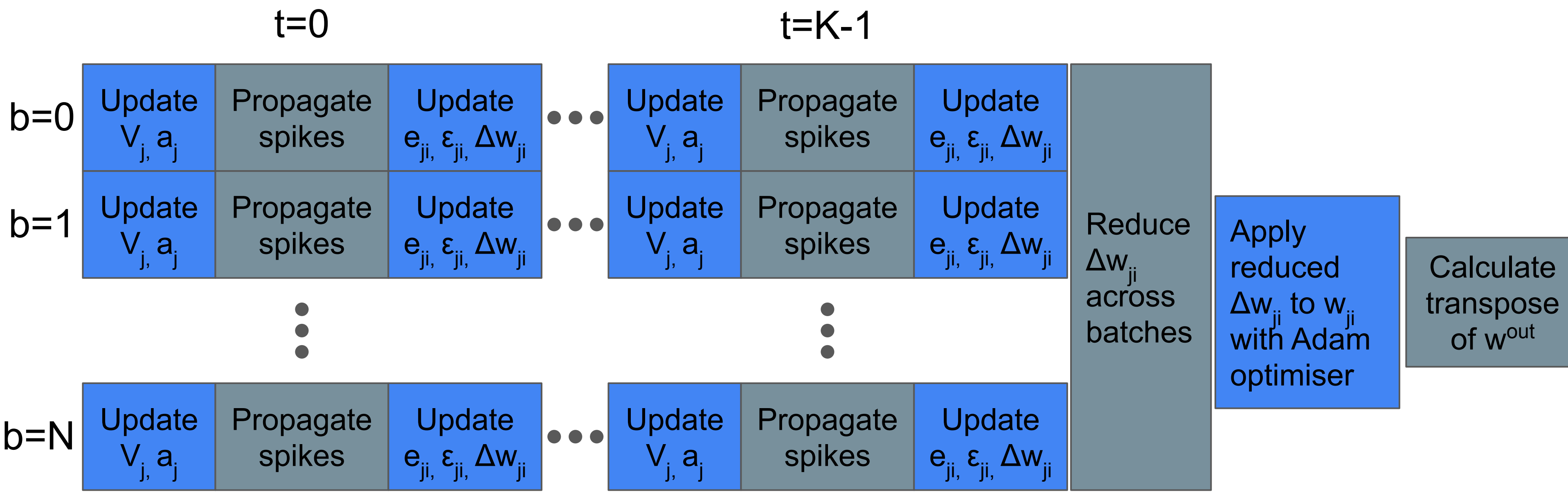
Batching

- Instantiated multiple copies of model simultaneously
- Variables e.g. weights can be shared between instances

Custom updates

- Arbitrary user-defined operations — optimizer
- Efficient matrix transpose — weight transport
- Efficient batch reduction operations (NCCL) — parallel training

Implementing eProp



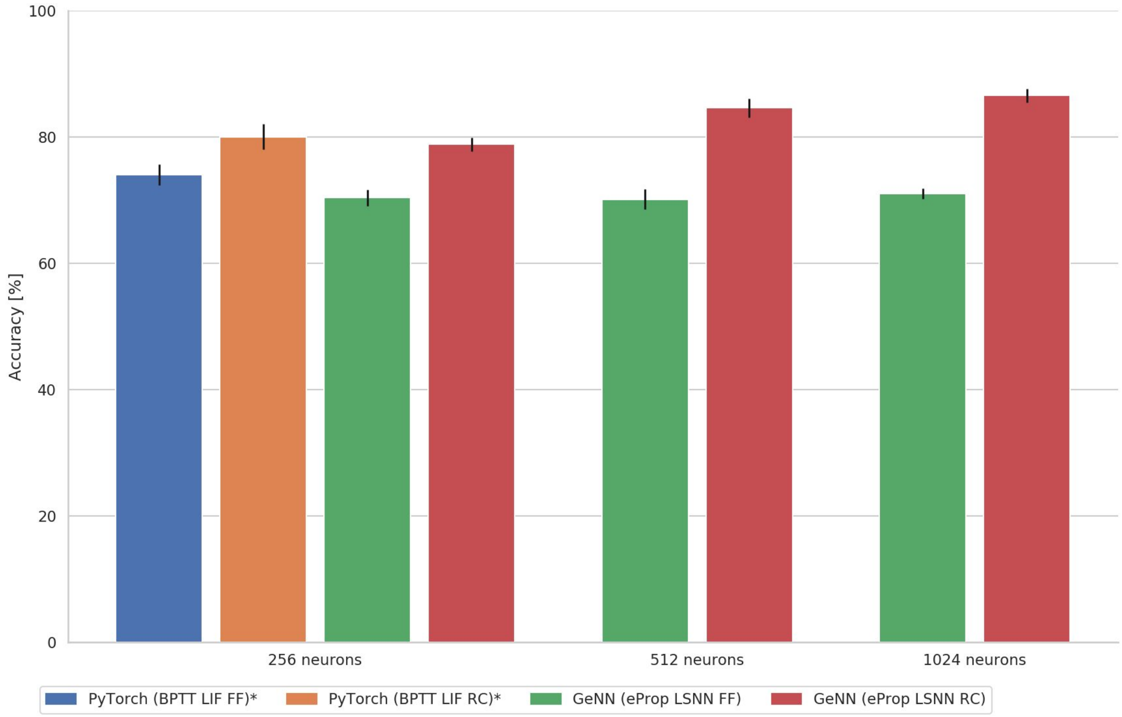
Spiking Heidelberg Digits

"three" "seven"

- 10420 recordings
- 12 speakers
- Digits 0-9 in English and German
- Converted to 700 spike trains using inner ear model

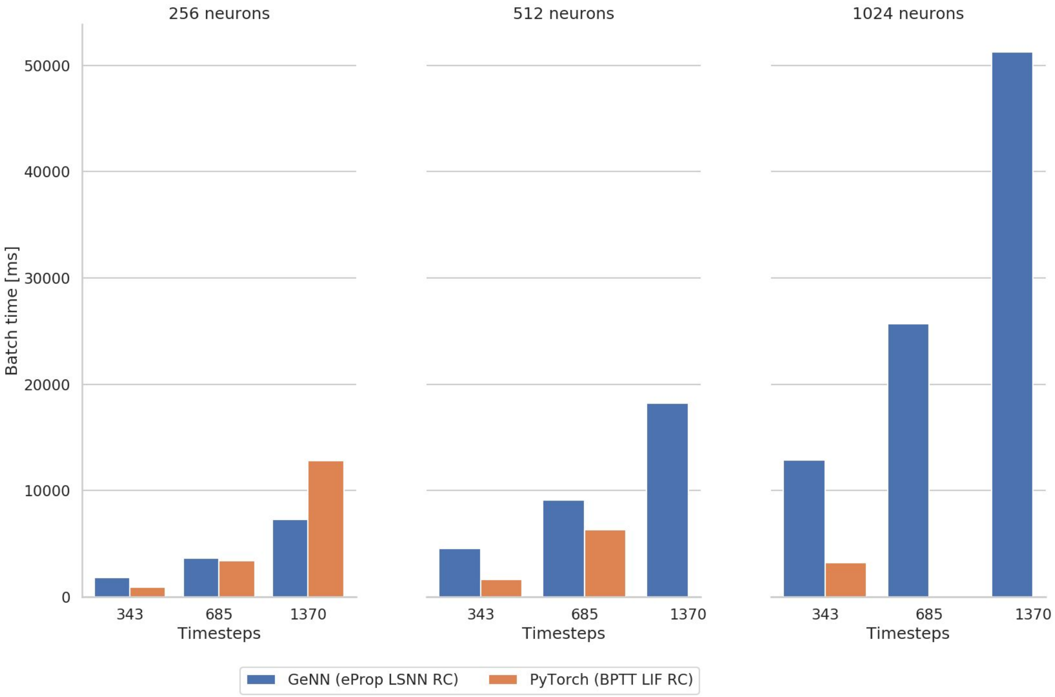
Cramer, B., Stradmann, Y., Schemmel, J., & Zenke, F. (2020). The Heidelberg Spiking Data Sets for the Systematic Evaluation of Spiking Neural Networks. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2020.3044364

Classification accuracy: Spiking Heidelberg Digits



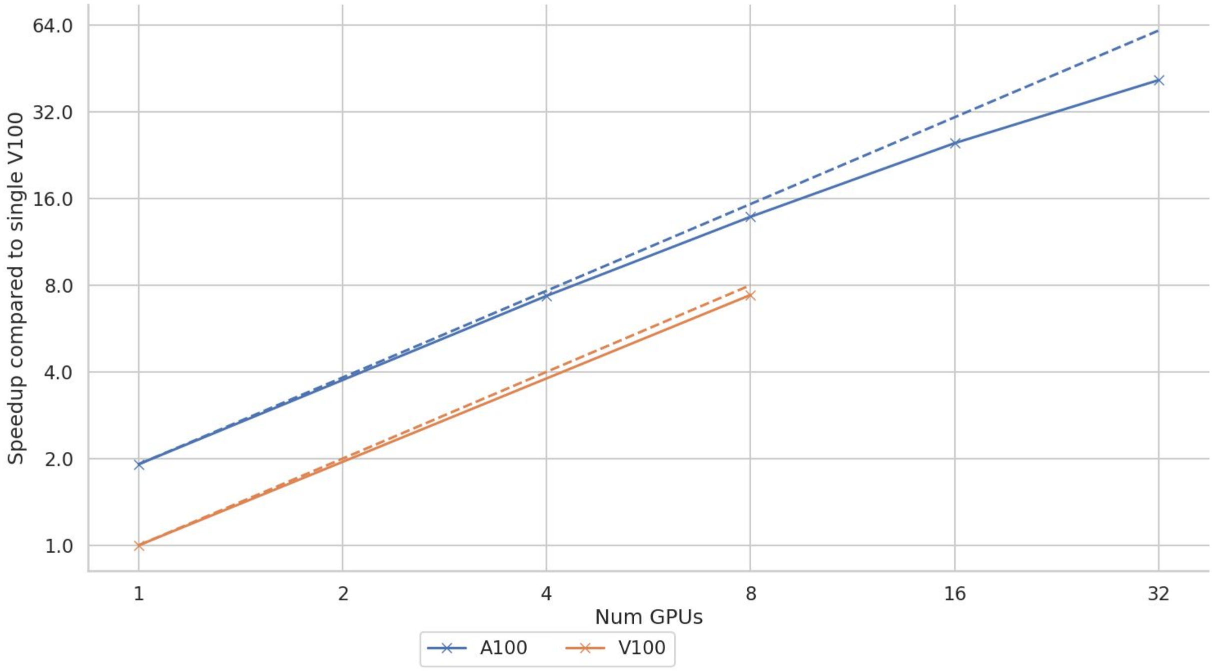
* Zenke, F., & Vogels, T. P. (2020). The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *BioRxiv*, 1–22. https://doi.org/10.1101/2020.06.29.176925

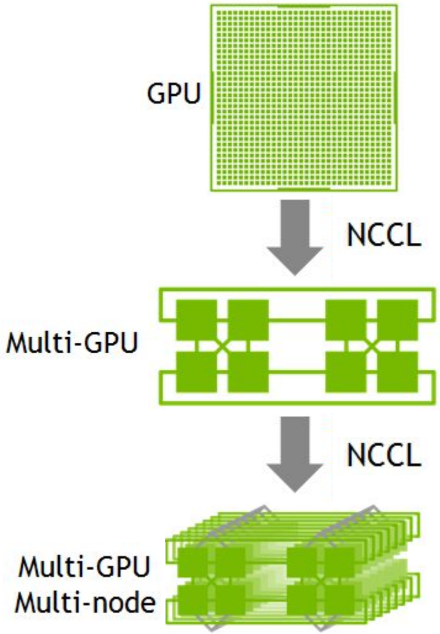
Training performance: Spiking Heidelberg Digits



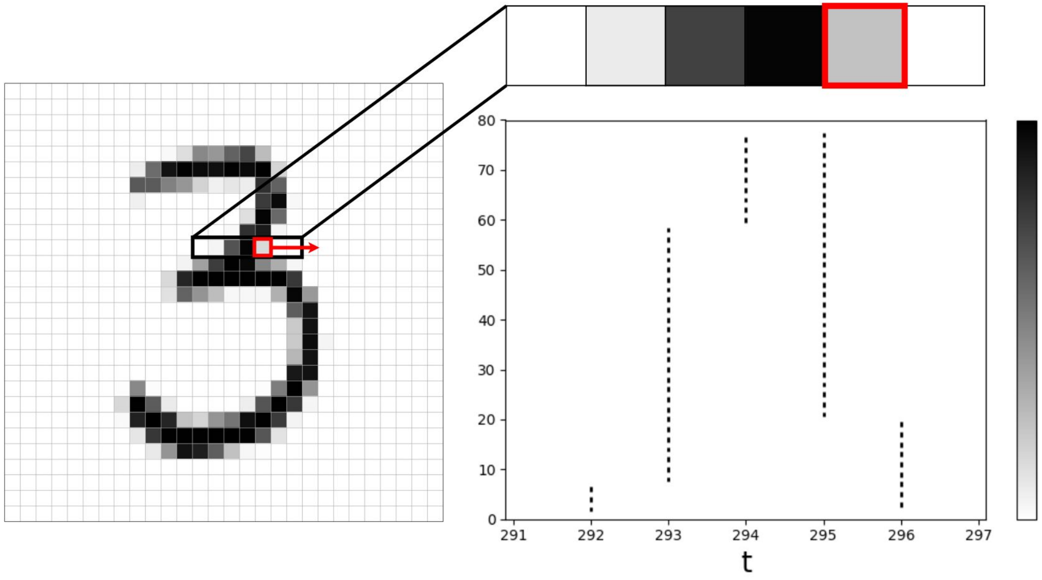
https://github.com/fzenke/spytorch

Multi-GPU training: Spiking Heidelberg Digits



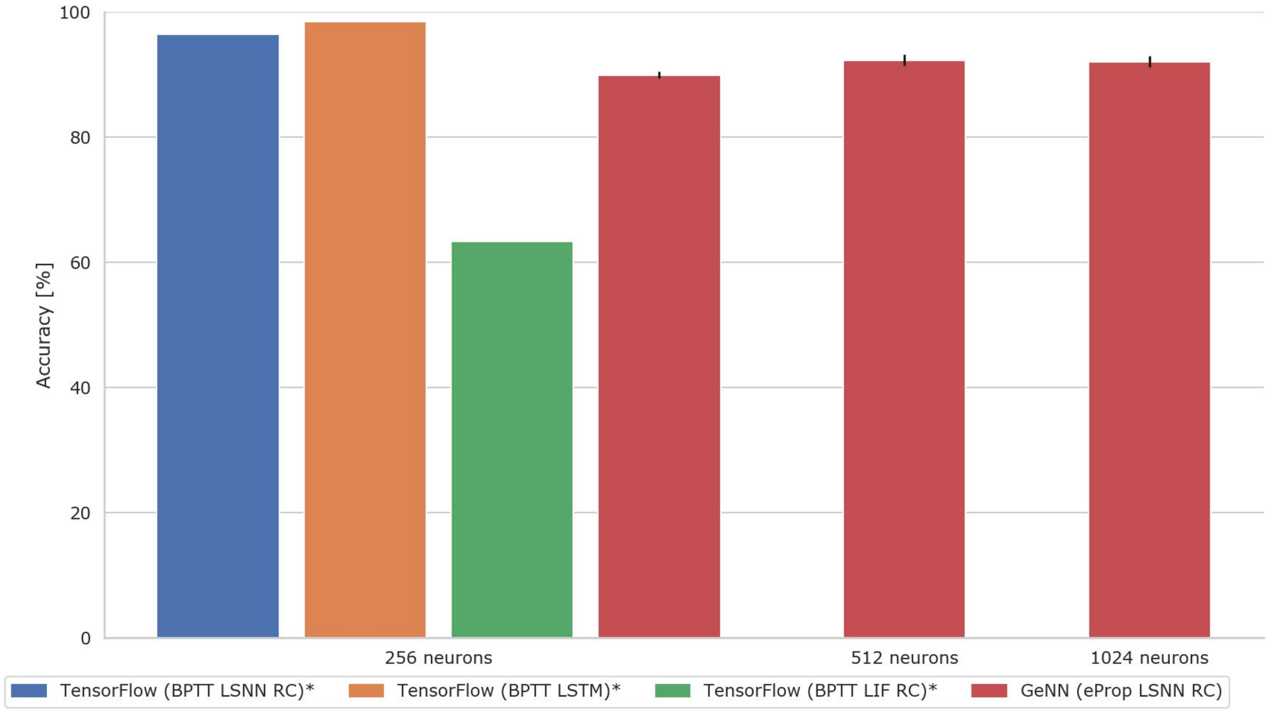


Spiking Sequential MNIST



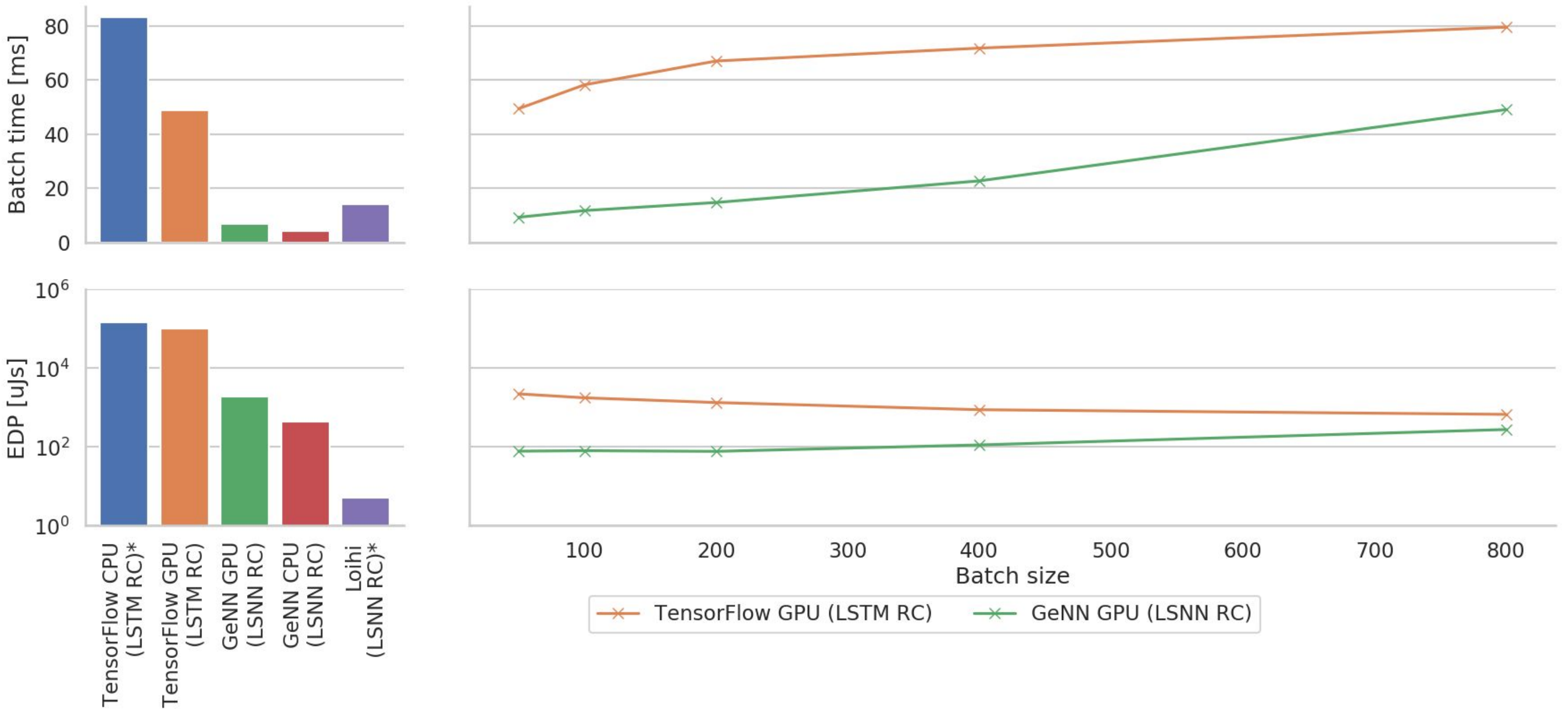
- Pixel values of MNIST digits presented in fixed order
- Each neuron represents a threshold crossing of a gray value

Classification accuracy: Spiking Sequential MNIST



* Plank, P., Rao, A., Wild, A., & Maass, W. (2021). A Long Short-Term Memory for AI Applications in Spike-based Neuromorphic Hardware. Retrieved from http://arxiv.org/abs/2107.03992

Inference performance: Spiking Sequential MNIST



* Plank, P., Rao, A., Wild, A., & Maass, W. (2021). A Long Short-Term Memory for AI Applications in Spike-based Neuromorphic Hardware. Retrieved from http://arxiv.org/abs/2107.03992

EventProp: Fully event-driven learning

Free dynamics	Transition condition	Jumps at transition
$\tau_{\mathrm{mem}} \frac{\mathrm{d}}{\mathrm{d}t} V = -V + I$ $\tau_{\mathrm{syn}} \frac{\mathrm{d}}{\mathrm{d}t} I = -I$	$(V)_n - \vartheta = 0$ $(\dot{V})_n \neq 0$ for any n	$(V^+)_n = 0$ $I^+ = I^- + We_n$

$$\mathcal{L} = l_{\mathrm{p}}(t^{\mathrm{post}}) + \int_0^T l_V(V(t), t)\mathrm{d}t.$$

https://youtu.be/oM7XEsDVcNg

Wunderlich, T. C., & Pehle, C. (2021). Event-based backpropagation can compute exact gradients for spiking neural networks. Scientific Reports, 11(1), 12829. https://doi.org/10.1038/s41598-021-91786-z

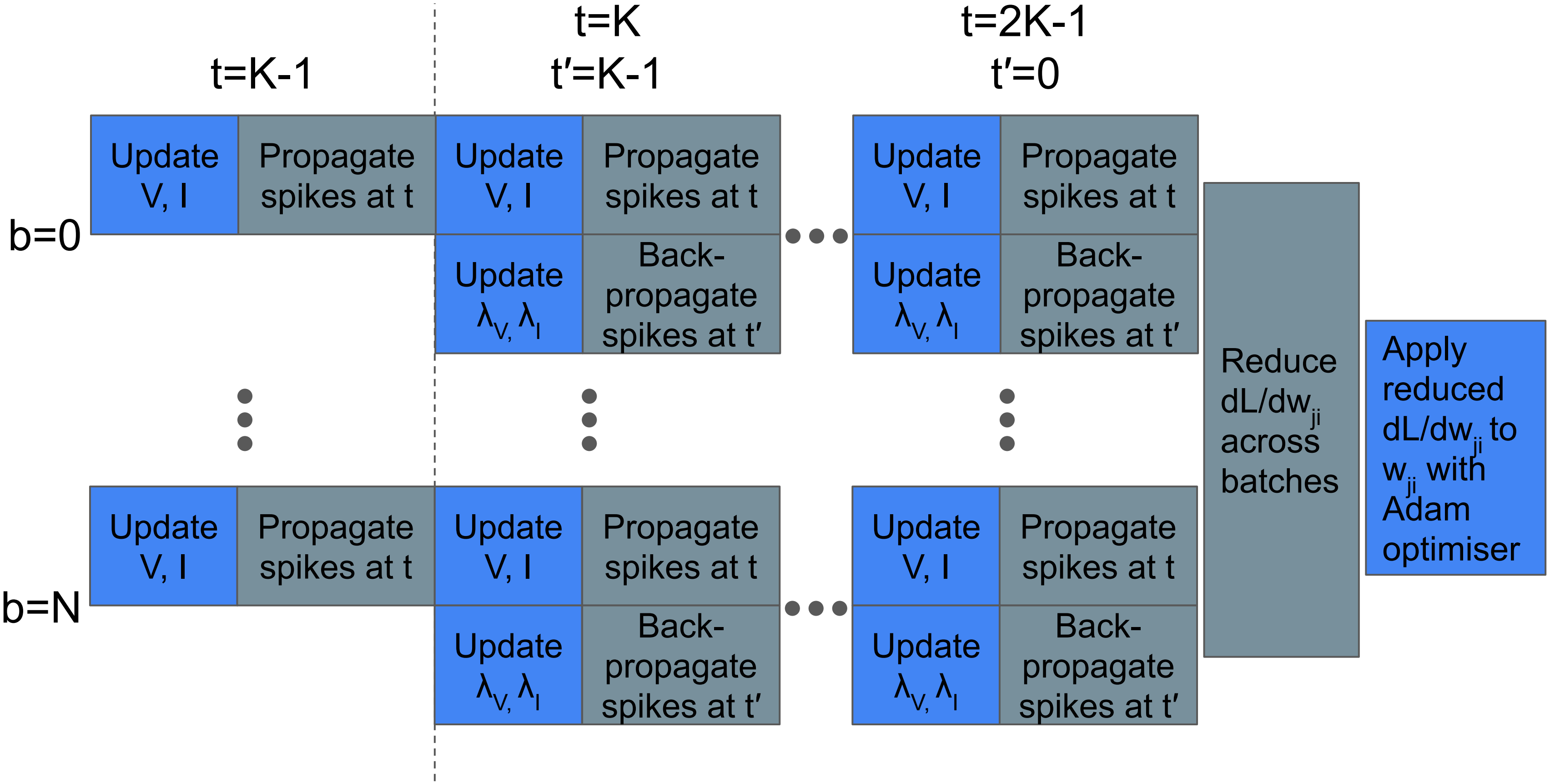
EventProp: Fully event-driven learning

Free dynamics	Transition condition	Jump at transition
$\tau_{\text{mem}}\lambda_V' = -\lambda_V - \frac{\partial l_V}{\partial V}$ $\tau_{\text{syn}}\lambda_I' = -\lambda_I + \lambda_V$	$t - t_k^{\text{post}} = 0$ for any k	$(\lambda_V^-)_{n(k)} = (\lambda_V^+)_{n(k)} + \frac{1}{\tau_{\text{mem}}(\dot{V}^-)_{n(k)}}\Big[\vartheta(\lambda_V^+)_{n(k)} + \left(W^\top(\lambda_V^+ - \lambda_I)\right)_{n(k)} + \frac{\partial l_p}{\partial t_k^{\text{post}}} + l_V^- - l_V^+\Big]$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}w_{ji}} = -\tau_{\text{syn}} \sum_{\text{spikes from } i} (\lambda_I)_j,$$

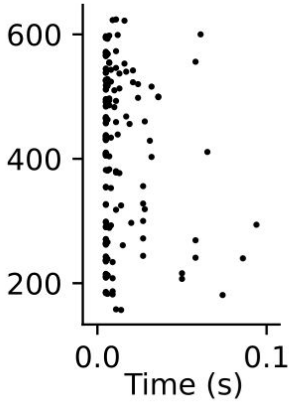
Wunderlich, T. C., & Pehle, C. (2021). Event-based backpropagation can compute exact gradients for spiking neural networks. Scientific Reports, 11(1), 12829. https://doi.org/10.1038/s41598-021-91786-z

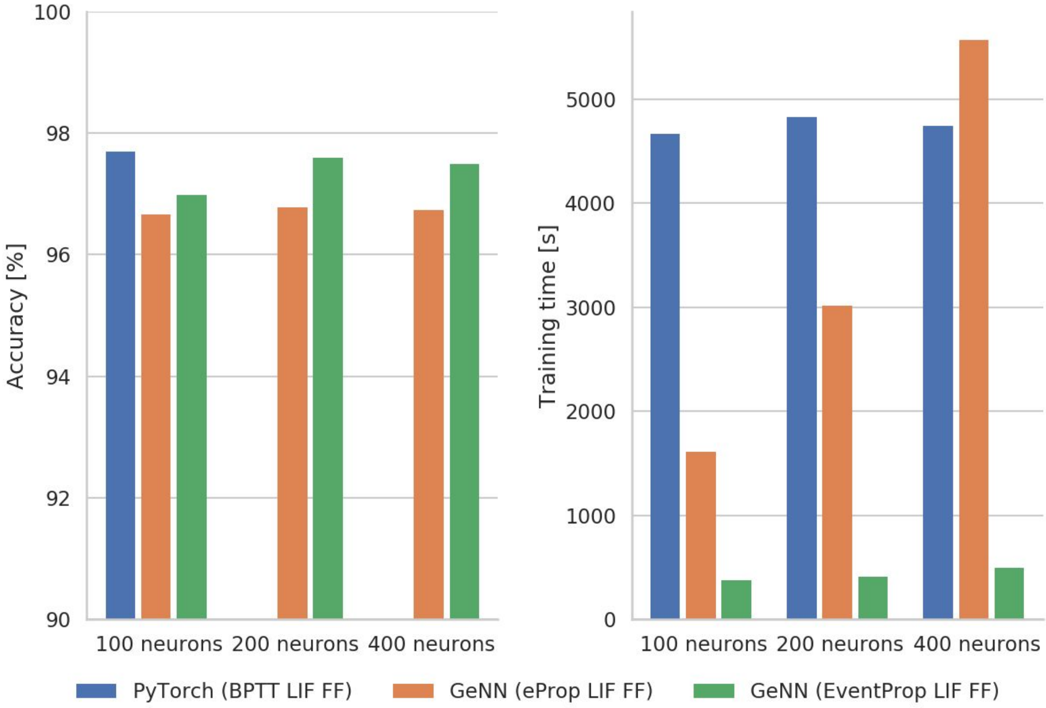
Implementing EventProp



Fully event-driven learning: Latency-encoded MNIST

$$T(x) = \begin{cases} \tau_{\text{eff}} \log\left(\frac{x}{x-\vartheta}\right) & x > \vartheta \\ \infty & \text{otherwise} \end{cases}$$





* Zenke, F., & Vogels, T. P. (2020). The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *BioRxiv*, 1–22. https://doi.org/10.1101/2020.06.29.176925

Future direction

- Simplifications to eProp [1]
- Applying eProp and EventProp to CNNs
- Deep-R [2]
- New higher-level frontend library [3]
- FPGA backend

1. Frenkel, C., & Indiveri, G. (2022). ReckOn: A 28nm Sub-mm2 Task-Agnostic Spiking Recurrent Neural Network Processor Enabling On-Chip Learning over Second-Long Timescales. 2022 IEEE International Solid- State Circuits Conference (ISSCC), 1–3. https://doi.org/10.1109/ISSCC42614.2022.9731734
2. Bellec, G., Kappel, D., Maass, W., & Legenstein, R. (2018). Deep rewiring: Training very sparse deep networks. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 1–24.
3. Turner, J. P., Knight, J. C., Subramanian, A., & Nowotny, T. (2022). mlGeNN: accelerating SNN inference using GPU-enabled neural networks. Neuromorphic Computing and Engineering, 2(2), 024002. https://doi.org/10.1088/2634-4386/ac5ac5

Acknowledgement

Everyone who's supported this work at Sussex, especially:

- Thomas Nowotny
- Andy Philippides
- Garibaldi Pineda Garcia
- Felix Kern (now University of Tokyo)
- James Turner

My funders at the EPSRC

Some very talented students:

- Manvi Agarwal (Basel)
- Ajay Subramanian (NYU)

And finally:

- Gregor Lenz, author of Tonic
- Franz Scherr for his help with eProp

Any questions

J.C.Knight@sussex.ac.uk