Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks

Bojian Yin[1][0000−0002−5074−4337], Federico Corradi[2][0000−0002−5868−8077], and Sander M. Bohté[1,3,4][0000−0002−7866−278X]

[1] CWI, Machine Learning group, Amsterdam, NL
[2] Stichting IMEC Netherlands, Holst Centre, Eindhoven, NL
[3] Univ of Amsterdam, Faculty of Science, Amsterdam, NL
[4] Rijksuniversiteit Groningen, Faculty of Science and Engineering, Groningen, NL

Abstract. Published on "Nature Machine Intelligence 3, 905-913 (2021)".

The success of brain-inspired deep learning in AI is naturally focusing attention back onto those inspirations and abstractions from neuroscience [7]. One such example is the abstraction of the sparse, pulsed and event-based nature of communication between biological neurons into neural units that communicate real values at every iteration or timestep of evaluation, taking the rate of firing of biological spiking neurons as an analog value . Spiking neurons, as more detailed neural abstractions, are theoretically more powerful compared to analog neural units [9] as they allow the relative timing of individual spikes to carry significant information. A real-world example in nature is the efficient sound localization in animals like Barn Owls using precise spike-timing [6]. The sparse and binary nature of communication similarly has the potential to drastically reduce energy consumption in specialized hardware, in the form of neuromorphic computing [3].

Since their introduction, numerous approaches to learning in spiking neural networks have been developed [2, 14, 20, 8, 5]. All such approaches define how input signals are transduced into sequences of spikes, and how output spike-trains are interpreted with respect to goals, learning rules, or loss functions. For supervised learning, approaches that calculate the gradient of the loss function with respect to the weights have to deal with the discontinuous nature of the spiking mechanism inside neurons. Local linearized approximations like SpikeProp [2] can be generalized to approximate "surrogate" gradients [10], or even calculated exactly in special cases [16]. The use of surrogate gradients in particular has recently resulted in rapidly improving performance on select benchmarks, closing the performance gap with conventional deep learning approaches for smaller image recognition tasks like CIFAR10 and (Fashion) MNIST, and demonstrating improved performance on temporal tasks like TIMIT speech recognition [1]. Still, spiking neural networks (SNNs) have struggled to demonstrate a clear advantage compared to classical artificial neural networks (ANNs) [13, 12].

Here, we introduce a novel approach to Spiking Recurrent Neural Networks (SRNNs)[17], networks that include recurrently connected layers of spiking neurons . We demonstrate how these networks can be trained to high performance on hard benchmarks, exceeding existing state-of-the-art in SNNs on all-but-one benchmark, and approaching or exceeding state-of-the-art in classical recurrent artificial neural networks. The high-performance in SRNNs is achieved by applying back-propagation-through-time (BPTT)[15] to spiking neurons using a novel Multi-Gaussian surrogate gradient and using adaptive spiking neurons where the internal time-constant parameters are co-trained with network weights. The Multi-Gaussian surrogate gradient is constructed to include negative slopes, similar to the gradient of the sigmoid-like dSilu activation function [4]: we find that the Multi-Gaussian surrogate gradient consistently outperforms other existing surrogate gradients. Similarly, co-training the internal time-constants of adaptive spiking neurons proved always beneficial. We demonstrate that these ingredients jointly improve performance to a competitive level while maintaining sparse average network activity.

We demonstrate the superior performance of SRNNs for well-known benchmarks that have an inherent temporal dimension, like ECG wave-pattern classification, speech (Google Speech Commands, TIMIT), radar gesture recognition (SoLi), and classical hard benchmarks like sequential MNIST and its permuted variant. We find that the SRNNs need very little communication, with the average spiking neuron emitting a spike once every 3 to 30 timesteps, depending on the task. Calculating the theoretical energy cost of computation, we then show that in SRNNs, cheap Accumulate (AC) operations dominate over more expensive Multiply-Accumulate (MAC) operations. Based on relative MAC vs. AC energy cost [12, 13], we argue that these sparsely spiking SRNNs have an energy advantage ranging from one to three orders of magnitude over RNNs and ANNs with comparable accuracy, depending on network and task complexity.

Using surrogate-gradients, the BPTT-gradient in the SRNNs can be computed using standard deep learning frameworks, where we used PyTorch [11]. With this approach, complicated architectures and spiking neuron models can be trained with state-of-the-art optimizers, regularizers, and visualization tools. At the same time, this approach is costly in terms of memory use and training time, as the computational graph is fully unrolled over all timesteps, precluding online and on-chip learning. Additionally, the abundant spatial and temporal sparsity is not exploited in the frameworks. This also limits the size of the networks to which this approach can be applied: for significantly larger networks, either dedicated hardware and/or sparsity optimized frameworks are needed[19]. Approximations to BPTT like eProp [1] or alternative recurrent learning methods like RTRL[18] may also help alleviate this limitation.

References

1. Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., Maass, W.: A solution to the learning dilemma for recurrent networks of spiking neurons. Nature Communications **11**(1), 1–15 (2020)

2. Bohte, S.M., Kok, J.N., La Poutré, J.A.: Spikeprop: backpropagation for networks of spiking neurons. In: European Symposium on Artificial Neural Networks (ESANN). vol. 48, pp. 17–37 (2000)
3. Davies, M., Srinivasa, N., Lin, T.H., Chinya, G., Cao, Y., Choday, S.H., Dimou, G., Joshi, P., Imam, N., Jain, S., et al.: Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro **38**(1), 82–99 (2018)
4. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks **107**, 3–11 (2018)
5. Falez, P., Tirilly, P., Bilasco, I.M., Devienne, P., Boulet, P.: Multi-layered spiking neural network with target timestamp threshold adaptation and stdp. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2019)
6. Gerstner, W., Kempter, R., Van Hemmen, J.L., Wagner, H.: A neuronal learning rule for sub-millisecond temporal coding. Nature **383**(6595), 76–78 (1996)
7. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron **95**(2), 245–258 (2017)
8. Kheradpisheh, S.R., Ganjtabesh, M., Thorpe, S.J., Masquelier, T.: Stdp-based spiking deep convolutional neural networks for object recognition. Neural Networks **99**, 56–67 (2018)
9. Maass, W.: Networks of spiking neurons: the third generation of neural network models. Neural Networks **10**(9), 1659–1671 (1997)
10. Neftci, E.O., Mostafa, H., Zenke, F.: Surrogate gradient learning in spiking neural networks. IEEE Signal Processing Magazine **36**, 61–63 (2019)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035 (2019)
12. Roy, K., Jaiswal, A., Panda, P.: Towards spike-based machine intelligence with neuromorphic computing. Nature **575**(7784), 607–617 (2019)
13. Sengupta, A., Ye, Y., Wang, R., Liu, C., Roy, K.: Going deeper in spiking neural networks: VGG and residual architectures. Front. Neurosci. **13**, 95 (Mar 2019)
14. Shrestha, S.B., Orchard, G.: Slayer: Spike layer error reassignment in time. In: Advances in Neural Information Processing Systems. vol. 31, pp. 1412–1421 (2018)
15. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proceedings of the IEEE **78**(10), 1550–1560 (1990)
16. Wunderlich, T.C., Pehle, C.: Eventprop: Backpropagation for exact gradients in spiking neural networks. arXiv preprint arXiv:2009.08378 (2020)
17. Yin, B., Corradi, F., Bohté, S.M.: Effective and efficient computation with multiple-timescale spiking recurrent neural networks. In: International Conference on Neuromorphic Systems 2020. pp. 1–8 (2020)
18. Zenke, F., Neftci, E.O.: Brain-inspired learning on neuromorphic substrates. Proceedings of the IEEE pp. 1–16 (2021)
19. Zenke, F., Bohté, S.M., Clopath, C., Iulia M.Comşa, Göltz, J., Maass, W., Masquelier, T., Naud, R., Neftci, E.O., Petrovici, M.A., Scherr, F., Goodman, D.F.M.: Visualizing a joint future of neuroscience and neuromorphic engineering. Neuron **109**(4), 571–575 (Feb 2021)
20. Zenke, F., Ganguli, S.: Superspike: Supervised learning in multilayer spiking neural networks. Neural Computation **30**(6), 1514–1541 (2018)