The spikes must flow: Seeing a joint future of neuroscience and neuromorphic engineering

Friedemann Zenke[1*], Sander M. Bohté[2,3,4], Claudia Clopath[5], Iulia M. Comşa[6], Julian Göltz[7,8], Wolfgang Maass[9], Timothée Masquelier[10], Richard Naud[11], Emre O. Neftci[12,13], Mihai A. Petrovici[7,8], Franz Scherr[9] & Dan F. M. Goodman[14]

[1] Friedrich Miescher Institute for Biomedical Research, Switzerland
[2] CWI, Amsterdam, The Netherlands
[3] Swammerdam Institute for Life Sciences (SILS), University of Amsterdam, The Netherlands
[4] AI Department, Rijksuniversiteit Groningen, Groningen, The Netherlands
[5] Bioengineering Department, Imperial College London, United Kingdom
[6] Google Research, Zürich, Switzerland
[7] Kirchhoff-Institute for Physics, Heidelberg University
[8] Department of Physiology, University of Bern
[9] Institute of Theoretical Computer Science, Graz University of Technology, Austria
[10] CNRS–CERCO UMR 5549, 31300 Toulouse, France
[11] Brain and Mind Research Institute of the University of Ottawa, Department of Cellular Molecular Medicine, University of Ottawa, Canada
[12] Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, United States
[13] Department of Computer Science, University of California, Irvine, Irvine, CA, United States
[14] Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom
* Correspondence: friedemann.zenke@fmi.ch

Abstract

Recent research resolves the challenging problem of building biophysically plausible spiking neural models that are also capable of complex information processing. This advance creates new opportunities in neuroscience and neuromorphic engineering, which we discussed at an online focus meeting.

Introduction

Neurons communicate and compute via discrete sparse events, spikes. This mechanism is radically different from digital computers and the analog activations of deep neural networks underlying modern artificial intelligence. To understand the brain and to mimic its supreme abilities in neuromorphic hardware, we need to understand how networks of spiking neurons learn and exhibit complex, intelligent behavior. The path to this goal has been frustrated by a seeming contradiction. Traditional spiking models closely resemble the mechanisms observed in the brain, but it has proven hard to build models that are capable of learning behaviors with similar complexity and performance as biological circuits. By contrast, deep neural networks are quite unlike biological brains. Yet, for the first time in history, they can solve complex problems at levels that rival the abilities of real brains.

What causes this difference in functional capability? As in deep artificial neural networks, computation in the brain arises from the intricate web of connections that allow large populations of neurons to function in unison as networks capable of complex information processing. As the activity flows through these connections, it undergoes high-dimensional nonlinear transformations. With the appropriate connectivity, this process results in meaningful computation at the network-level. But finding the right connections is problematic because it requires knowledge about how individual neurons deep inside the network affect the output of the whole network. This requirement is known as the credit assignment problem. What distinguishes deep learning is that this problem is solved algorithmically through gradient-based optimization, whereby tuning synaptic connections and neuronal parameters throughout the entire network gradually reduces output errors (Fig 1a). This algorithm relies on gradient information flowing through the network, which is ensured by well-behaved differentiable neuronal activation functions. The existence of such

optimization algorithms is what makes deep learning one of the most promising avenues to understand the brain's inner workings through functional models (Richards et al., 2019).

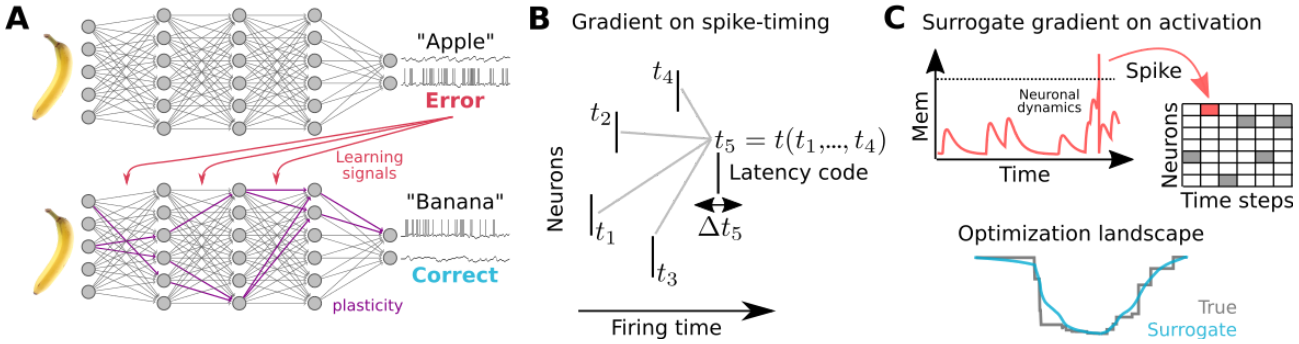

Figure 1: (A) Instilling functions at the network level requires hidden neurons, which are neither connected to the input nor the network's output, to reduce their contribution to errors at the output level. The algorithmic feat of assigning credit or blame to individual hidden neurons and synapses, thereby allowing these neurons to learn, is called the credit assignment problem. (B-C) Schematic view of the two principal schemes underlying the majority of optimization approaches for solving the credit assignment problem in spiking networks. (B) In a spike-timing-based representation, gradient-based updates operate directly on smoothly differentiable spike times. (C) In an activity-based representation, spikes fall onto a time grid whose values are given by thresholding neuronal membrane potentials. Because the spikes' binary functional character precludes computing derivatives, optimizing this representation requires surrogate gradients whereby a smooth surrogate replaces the nondifferentiable step function.

Unfortunately, gradient-based optimization fails in spiking neural networks in which the non-differentiable nature of neuronal spiking dynamics prevents gradients from flowing. Yet, sustained joint efforts by neuromorphic engineers and computational neuroscientists have resulted in several recent developments that allow translating the algorithms underlying the revolution of deep learning to the domain of biologically constrained spiking neural networks.

To provide an interdisciplinary forum for this emerging field, which closes the gap between spiking networks and deep learning, we organized a focus meeting entitled "Spiking neural networks as universal function approximators." Over two days, experts in the field shared recent work in talks, discussed novel ideas, and plotted ways to move forward in open panel discussions. Due to COVID-19, it was an online meeting, which attracted over 700 registered participants from all over the world who actively engaged in vivid discussions.

This meeting report summarizes the key outcomes of this gathering. Central are several innovations that herald a fundamental shift in spiking neural network modeling that combine the best of traditional biologically plausible models and modern performance-optimized artificial neural networks. Importantly, these developments allow building models that:

1. Take advantage of temporal spiking dynamics to efficiently encode and process information.
2. Embrace the computational value of neuronal heterogeneity and multi-time-scale dynamics by jointly optimizing neuronal parameters with the connectivity.
3. Learn through biologically plausible learning rules derived from a normative gradient-based framework, thereby providing new vistas on their mechanistic underpinnings at the micro-circuit-level.

In summary, these advances give us a principled and general new approach to tackle questions about neuronal heterogeneity, specific circuit motifs, and the role of temporal spiking dynamics in the nervous system.

The importance of temporal dynamics in neural processing

Previous work on training spiking neural networks at complex tasks used only the stationary firing rates of neurons, which allowed for a straightforward translation of results from the conventional artificial neural networks used in machine learning. However, as a consequence, these networks

were unable to take advantage of the temporal structure that spikes can carry, a mechanism that the brain exploits extensively for rapid processing and sparse information coding. The work featured in this meeting report overcomes the technical problems of previous studies. Therefore, it allows us, for the first time, to explore the unique temporal coding strategies that spiking networks can employ to solve complex information processing tasks. The technical innovation that made this possible was to find ways of directly translating gradient-based learning to fine-grained temporal spiking while simultaneously keeping the number of emitted spikes minimal (Neftci et al., 2019). This operating setting, with sparse but precisely timed action potentials, is not only reminiscent of cortical processing; it also renders spiking neural network implementations more efficient to run on hardware. We now discuss three major learning paradigms exemplifying this new approach: FORCE training in spiking networks, gradients with respect to single spike times, and surrogate gradients.

Time-continuous processing with instantaneous rates

One of the first studies to showcase the potential of such approaches to build spiking neural networks that solve concrete biological problems adapted the classic FORCE training algorithm for recurrent spiking neural networks (Nicola & Clopath, 2017). The central idea, which sidesteps the problem of having to compute gradients through spikes, is to solve a regression problem at every instant of time over linear combinations of temporally filtered spike trains, while using the postsynaptic potential as the filter kernel. This approach does not require stationary firing rates, readily solves complex sequence generation problems, and is robust to the choice of neuron model.

Efficient low-latency processing with single precisely timed spikes

Another approach assumes that each neuron spikes precisely once in a given time period, and computes gradients with respect to these spike times (Fig. 1b). Kheradpisheh & Masquelier (2020) showed that it yields state-of-the-art accuracy for spike-latency encoded versions of MNIST and fashion MNIST. In similar work, Comşa et al. (2020) not only achieved competitive performance on latency-MNIST, but they also proved that such encoding schemes provide a class of universal approximators. Göltz et al. (2019) demonstrated competitive performance on spike latency-encoded tasks using the accelerated BrainScales-2 analog neuromorphic system. Not only does this lead to vastly reduced latency and power consumption of only 200 mW, allowing processing of more than 10k inputs per second, but the learning scheme is robust to small manufacturing imperfections of the underlying neuromorphic substrate, an essential requirement also for any biological system.

By design, timing-based approaches are well-suited for static stimuli, encoded using a latency code. The method assumes extreme sparseness of spiking since every neuron emits at most one spike. This representation allows efficient event-driven algorithms in which time represents itself, which translates algorithmically into a small memory footprint and low-power computation at the network-level (Kheradpisheh & Masquelier 2020; Göltz et al., 2019). Similar to a binary neural network, all processing occurs as a single volley of spikes propagates through the network. Therefore, the result is ready with low latency. Despite these advantages, only using single spikes in each neuron has its limits and is less suitable to process temporal stimuli such as EEG signals, speech, or videos. This limit, however, can be overcome by training networks with surrogate gradients.

Flexible information processing through surrogate gradient learning

Instead of operating on firing times, surrogate gradients are computed in neuronal simulations with a discrete-time grid, similar to conventional recurrent neural networks in machine learning (Fig. 1c). To capture the essence of spiking dynamics, the approach assumes a binary neuronal output in each time step. Because the binary neuronal activation function is not differentiable, the standard procedures of computing objective function gradients in these networks fail. The trick is to

approximate the non-differentiable step function with a smooth differentiable function, which then yields a surrogate gradient that allows optimizing spiking networks efficiently using standard machine learning software (Neftci et al., 2019). Since the surrogate gradient learning does not impose any strict constraints on the number of spikes emitted by any neuron, it can flexibly handle temporal stimuli in which input neurons spike more than once (Kheradpisheh & Masquelier 2020).

The computational value of coordinated neuronal heterogeneity

Another exciting development in building spiking neural network models is that surrogate gradient techniques can optimize essential neuronal and synaptic parameters, like time constants, jointly with the connectivity. This twist offers exciting new opportunities for modelers to allow for parameter heterogeneity. For instance, Yin et al. (2020) showed that instead of giving each neuron the same adaptation time constant, a common simplifying model assumption, optimizing the time-constants on a per-neuron basis offers decisive computational advantages on several classification benchmarks. Optimizing neuronal parameters is a notable departure from previous modeling standards and opens the way to understanding the functional role of the brain's cellular diversity.

The importance of multi-time scale dynamics

More generally, several studies have shown how individual neurons' dynamical complexity plays a crucial role in shaping computation at the network level. Thus, we now have the essential tools to harness such complexity in spiking network models. Bellec et al. (2020) showed how a slowly moving neuronal firing threshold drastically improved computational performance, allowing spiking networks to solve a plethora of complex computational problems like, for instance, playing Atari games. In a similar vein, Yin et al. (2020) showed that networks with optimal heterogeneous adaptation time scales consistently outperformed networks without such heterogeneity on several time-series classification tasks. In addition to improving overall computational performance, spike frequency adaptation also leads to a significant reduction of spike counts, with the potential of further reducing energy-consumption on neuromorphic implementations.

Linking normative and biologically plausible plasticity models

The work discussed thus far uses gradient-based optimization algorithms, which are not biologically plausible. For instance, the standard algorithm for training recurrent neural networks in machine learning is back-propagation through time (BPTT). It cannot be interpreted as a biologically plausible learning rule because it requires propagating information backward through time. Further, its computation requires knowledge to which individual synapses physically do not have access. This means we can use the algorithm to optimize network models, but it does not provide useful ideas on how neurobiology would achieve a similar optimization. In the context of spiking networks, BPTT has another notable disadvantage. Its memory requirements grow linearly with stimulus duration, creating issues when simulating prolonged stimuli and large networks at high temporal resolution.

Real-time recurrent learning (RTRL) is an alternative algorithm that does not have this issue and only requires propagating information forward in time. It still requires non-local information, thus precluding a direct interpretation as a biologically plausible learning rule. But, approximations of this algorithm can be interpreted as local learning rules (Bellec et al., 2020; Zenke and Neftci, 2020). For example, using a local learning rule “e-Prop” derived in this way, recurrent spiking neural networks with slow spike-triggered adaptation can learn to solve a diversity of challenging tasks, including speech recognition and playing Atari games (Bellec et al., 2020).

Zenke and Neftci (2020) introduced a general mathematical framework that presents a new view on auto-differentiation, allowing flexibly combining elements of BPTT and RTRL with approximations. The framework exposes the fundamental link of local learning rules with approximate forms of RTRL and numerous online learning rules, i.e., e-Prop, OSTL, RFLO, DECOLLE, and SuperSpike, which can all be derived from ignoring specific contributions to the

gradient from recurrent connections. Intriguingly, the notion of synaptic eligibility traces, known to exist in biology, falls out of this normative framework and are tied to synaptic and neuronal dynamics (Bellec et al., 2020; Neftci et al., 2019). Common to these approximations is their improved efficiency, biologically interpretability, and implementability on neuromorphic hardware.

Biologically plausible solutions to the spatial credit assignment problem

While eligibility traces can solve the temporal credit assignment problem (i.e., which past network activity contributed to a specific error or reinforcement signal later in time), solving the spatial credit assignment problem (i.e., which neuron's activity contributes strongly to a particular network-level output), requires dedicated circuits that compute and communicate learning signals between neurons. How the brain accomplishes this feat remains an open question.

Bellec et al. (2020) explored one conceivable way in which a separately trained network module acts as a learning signal generator. Its task is to provide precisely timed and spatially segregated learning signals to a population of neurons as a putative solution to the spatial credit assignment problem. Nevertheless, the precise circuit mechanisms which could exert such control over plasticity are left open in the model.

Payeur et al. (2020) approached this question in a biophysical circuit model using experimentally verified micro-circuit elements and cell-types. The model uses burst-multiplexing, whereby isolated spikes have a different meaning than high-frequency bursts, maintaining two separate information channels through each neuron that allow for the simultaneous flow of feed-forward information and feedback errors. To achieve this, the model relies solely on biologically plausible properties such as dendritic compartments, short-term plasticity, inhibitory microcircuits, and burst-dependent plasticity. Using a reduced-complexity version of their model, the authors demonstrate that it achieves competitive performance on large-scale machine learning benchmarks like ImageNet.

Future challenges and research directions

Although our newly gained ability to build functional spiking neural networks holds the potential to revolutionize how we construct biologically inspired neural network models, there are several notable difficulties ahead. We broadly distinguish between conceptual and technical challenges.

Conceptual challenges. How can we best use functional spiking neural network models to further our understanding of information processing in the brain? Establishing a rapport between artificial and biological spiking networks will be a crucial first step. Doing so will require quantitative ways of comparing network representations across different networks. Initially, it may be viable to adapt and generalize representational similarity analysis currently used to compare neural data with deep neural networks. It is conceivable that the intrinsic temporal structure of neuronal spike trains may require entirely novel analysis techniques.

Another essential step will be to incrementally move toward more plausible architectures by gradually incorporating biological wiring constraints, cell-type diversity, and circuit motifs into our network models. Training such networks on particular tasks will shed light on both the role of such restrictions for efficient information processing and open up new vistas to translate these insights into more efficient generations of neuromorphic hardware.

We should expect different outcomes depending on whether visual inputs use a latency-code, a rate code, or a mixture between the two. Therefore, architecture refinement has to go hand in hand with biologically plausible inputs to provide interpretable results. Hence, detailed knowledge about the brain's input encoding is a prerequisite to making the best of our newfound ability to train spiking neural networks.

While current work focuses on supervised learning, future applications to build better hardware and gain a deeper understanding of the brain require studying unsupervised learning. In doing so, we can hope to answer questions about which objective functions the brain optimizes and how.

Technical challenges. One primary goal is to scale up training of spiking neural networks to larger systems. Although the technical possibilities to simulate large-scale spiking neural networks have existed for years, current training algorithms are not well adapted for these large-scale and often event-based implementations. The current size limitations for functional networks are mainly due to the auto-differentiation libraries used to train spiking networks. Their design has poor support for sparse connectivity and sparse spiking, which renders them inefficient for simulating large network models. Consequently, most models highlighted here consisted of hundreds of neurons, a small number compared to most biological circuits and typical deep neural networks in machine learning. Moving toward neuron numbers comparable to biology and applying these networks to real-world datasets will require the development of novel algorithms, software libraries, and dedicated hardware accelerators that perform well with the specifics of spiking neural networks. Another essential aspect of achieving this goal is developing effective parameter initialization strategies, which are crucial for successful training leading to high task performance.

Finally, to further devise meaningful comparisons between artificial and biological networks, we need to dedicate time and effort to build plausible spike-based datasets that mimic the inputs seen by sensory neurons in the brain. As in deep learning, large datasets are a prerequisite to forming functional networks from optimization principles that generalize well to unseen data. Designing datasets that closely resemble the inputs experienced by sensory neurons will thus be crucial to allow quantitative comparisons between the internal representations of artificial and experimental data from biological neural networks. Finally, an important question remains open: Which tasks do spiking neural networks solve better than their non-spiking relatives?

Conclusion

In summary, while our newly gained ability to instantiate spiking neural networks that perform complex information processing tasks is an exciting advance, demanding technical and conceptual challenges lie ahead to reap its full benefits. In addressing these challenges, we expect a significant shift from the often hand-crafted spiking network models which solve simplistic computational problems toward sophisticated spiking networks that solve demanding computational challenges. This transformation will have a lasting effect on the practical applications in brain-inspired hardware and modeling in computational neuroscience. In particular, it allows building spiking network models that implement the hypothesized function of specific brain circuits and directly compare the model activity to experimental data. Thus far, such comparisons only exist with artificial neural networks whose architecture and dynamics are markedly different from neurobiology. Ultimately, this may well be the beginning of a new era in spiking neural network research, which, when brought to full fruition, may provide us with concrete answers to the long-standing question: Why spikes?

Acknowledgments

The authors thank Walter Senn for his input on the draft and the meeting attendees for lively discussions.

Declaration of Interests

Richard Naud has filed a provisional patent application "Deep learning method with spiking units and spiking neural network system and neuromorphic device".

Iulia M. Comşa is an employee of Google LLC. Portions of the work by Comşa et al. (2019), which is discussed in this submission, are covered by pending PCT Patent Application No. PCT/US2019/055848 ("Temporal Coding in Leaky Spiking Neural Networks"), filed by Google in 2019.
All remaining authors declare no competing interests.

References

- Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., and Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. Nature Communications *11*, 3625.
- Comsa, I.M., Potempa, K., Versari, L., Fischbacher, T., Gesmundo, A., and Alakuijala, J. (2020). Temporal Coding in Spiking Neural Networks with Alpha Synaptic Function: Learning with Backpropagation. ArXiv:1907.13223 [Cs, q-Bio].
- Göltz, J., Kriener, L., Baumbach, A., Billaudelle, S., Breitwieser, O., Cramer, B., Dold, D., Kungl, A.F., Senn, W., Schemmel, J., Meier, K., Petrovici, M.A. (2019). Fast and deep: energy-efficient neuromorphic learning with first-spike times. ArXiv:1912.11443 [Cs, q-Bio, Stat]
- Kheradpisheh, S.R, and Masquelier, T. (2020). Temporal Backpropagation for Spiking Neural Networks with One Spike per Neuron. Int J Neural Syst 30(06):2050027.
- Neftci, E.O., Mostafa, H., and Zenke, F. (2019). Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-based optimization to spiking neural networks. IEEE Signal Process Mag *36*, 51–63.
- Nicola, W., and Clopath, C. (2017). Supervised learning in spiking neural networks with FORCE training. Nat Commun *8*, 2208.
- Payeur, A., Guerguiev, J., Zenke, F., Richards, B.A., and Naud, R. (2020). Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. BioRxiv 2020.03.30.015511.
- Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., Berker, A. de, Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat Neurosci *22*, 1761–1770.
- Yin, B., Corradi, F., and Bohté, S.M. (2020). Effective and Efficient Computation with Multiple-timescale Spiking Recurrent Neural Networks. In International Conference on Neuromorphic Systems 2020, (New York, NY, USA: Association for Computing Machinery), pp. 1–8.
- Zenke, F., and Neftci, E.O. (2020). Brain-Inspired Learning on Neuromorphic Substrates. ArXiv:2010.11931 [Cs].