Benchmarks for progress in neuromorphic computing

In order for the neuromorphic research field to advance into the mainstream of computing, it needs to start quantifying gains, standardize on benchmarks and focus on feasible application challenges.

Mike Davies

No one would accuse neuromorphic computing of lacking ambition. With a mission to decipher the multitude of secrets nature deploys to achieve unrivalled efficiency and flexibility in brain-based computing, the field faces one of the most daunting challenges in all of computational science and engineering.

Even with the brain as a guide, reverse engineering such a complex system remains an open-ended and highly unconstrained problem. We must reinvent our understanding of computing from the logic gates up, replacing synchronous sequencing and monolithic memory with millions of parallel dynamical units all communicating with asynchronous spike messages. Familiar programming models are replaced with high-dimensional, temporal and nonlinear computational abstractions yet to be fully comprehended.

In order to deliver on such an ambitious agenda, the neuromorphic field needs to focus more on principles and rigour, less on open-ended exploration and mapping speculative mechanistic features to silicon. Steady progress to real-world value depends on quantitative metrics, discipline and informed prioritization.

Not so fast?

Benchmarking is the traditional methodology for measuring and focusing progress in engineering. In many fields, benchmarks have provided great value, from Dhrystone and SPECint for microprocessor innovation to ImageNet for convolutional deep learning progress more recently. A good benchmark serves to motivate researchers to solve one particular problem chosen as a worthy representative of a broader class of useful problems.

Yet, in a nascent and fragmented field such as neuromorphic computing, some reservation for benchmarking is warranted. An ill-chosen benchmark can focus disproportionate attention on just one piece of the larger puzzle, with the effect of impeding rather than accelerating broader progress.

An example where this is already happening relates to measurements of ‘synaptic ops’ that are commonly reported for neuromorphic designs. Such microscopic metrics are easy to measure but offer little value for assessing the worth of a particular neuromorphic chip. If a chip hasn’t demonstrated even a single meaningful workload, then such an indirect energy measurement should be taken with a grain of salt. On the other hand, if a chip can support meaningful workloads, then it should be measured on those terms. As long as the field has yet to determine the right architectural features and degree of programmability, the emphasis should be on assessing comprehensive workloads, not on microscopic circuit properties. Fixating on readily optimized synaptic op metrics puts the cart before the horse.

This concern about misdirection of focus is also what troubles some who fear ‘benchmarking’ is code for adopting the particular benchmarks that have guided the deep learning community. While neuromorphic researchers universally use the MNIST dataset to test pattern classification algorithms, consensus disappears on whether more advanced vision datasets such as CIFAR-10 and ImageNet should immediately follow.

At root is unease with the idea that the goal of neuromorphic learning should be the ‘training’ paradigm that starts from a tabula rasa network state and learns by ingesting a single dataset, in one computationally intensive leap. Many neuromorphic researchers, with human learning in mind, would prefer to develop incremental, hierarchical and factorization-based approaches. Humans don’t learn new concepts from hundreds of examples presented in all conceivable contexts; humans learn from single examples by leveraging prior learning. From this perspective, it’s unwise to focus on datasets that presuppose a tabula rasa approach.

Another benchmarking challenge arises in the handling of thorny environment modelling issues present in robotics, closed-loop control and active sensing problems. These challenges have stymied standardization even for conventional approaches and can be exacerbated in the neuromorphic domain. Specifically, some neuromorphic systems use analog circuits with real-world time constants and therefore can only run at one particular speed — no slower or faster. These challenges are surmountable, but more thought and care must be devoted to the problem — for example, with a minimal simulation methodology satisfying variability bounds[1].

The need for compelling benchmarks

Despite the challenges and legitimate concerns, the time has nevertheless come for the neuromorphic field to embrace an appropriate set of benchmarks, specifically on two fronts: first, internally oriented, as a way to measure the capabilities of different spiking neuromorphic architectures; and second, in an outward sense, problems that quantify the value of neuromorphic solutions compared to state-of-the-art conventional solutions.

On the first front, the field needs a comprehensive suite of spiking neural network algorithms analogous to SPECint or MLPerf. We might call this SpikeMark. It’s important to bear in mind that different neuromorphic architectures are far more varied than the different von Neumann processors for which SPECint was designed in the 1990s, and support a far more varied set of algorithms than the numerous variants of backpropagation training that MLPerf measures. One further hurdle is that there is no standardized language for neuromorphic programming, such as C for SPECint. Nevertheless, these challenges can be addressed with good written specifications and conventionally coded descriptions of ground truth.

The purpose of such a SpikeMark benchmarking suite (Box 1) would be to evaluate the relative features, flexibility, performance and efficiency of different neuromorphic platforms. It would include both applications suitable for real-world

Box 1 | SpikeMark

A benchmarking suite for spiking neuromorphic systems might include the following workloads:

- Classify spoken keywords using a specified pre-trained deep neural network converted to spiking form[4].
- Classify sequentially presented MNIST digits and TIMIT phonemes using offline-trained long short-term memory spiking neural networks[9].
- Detect hand gestures from the DVS Gesture event-based camera dataset[10], with a convolutional spiking neural network (SNN) trained offline with SNN backpropagation[11] and online with deep continuous local learning[12].
- Solve least absolute shrinkage and selection operator (LASSO) problems with the spiking locally competitive algorithm[13].
- Solve Sudoku and map colouring constraint satisfaction problems using neural sampling[14].
- Identify the shortest path in a variety of graphs using spike-based temporal wavefront propagation[15].
- Perform pattern similarity matching with threshold phasor associative memory[8].
- Control a modelled robotic arm subjected to nonlinear wear using an adaptive controller trained with the neural engineering framework[16].
- Solve simple cognitive problems with the Spaun-embodied brain model[17].
- Simultaneous localization and mapping (SLAM) using Bayesian learning and inference with a robotic head-direction environment model[3].

Today, no single neuromorphic platform has successfully run all of the above workloads. However, several of these examples have been run on multiple platforms, and none depends on exotic platform-specific features beyond typical leaky-integrate-and-fire spiking neural network functionality with local learning rules.

use as well as representative algorithmic primitives with minimal standalone value. Given clear guidelines for standardized cross-platform replication and evaluation, the SpikeMark suite could grow in an open, self-organized way, with popular consensus determining its constituent tasks.

To accommodate a broad diversity of neuromorphic architectures, the suite needs to span a wide range of problem scales and evaluation metrics, including time-to-solution, energy-to-solution, throughput and accuracy. This allows for a fair assessment of each architecture's intended niche. Some systems may excel at small edge workloads while not supporting large-scale workloads. Others, due to functional constraints, may only support a subset of the suite, such as workloads without synaptic plasticity.

For the second imperative of evaluating neuromorphic solutions versus conventional solutions, we can focus on the subset of SpikeMark tasks for which clear and competitive conventional solutions exist. For those particular workloads, including of course any deep learning examples, the range of hardware platforms would be expanded to include central processing units, graphics processing units and relevant specialized devices such as Google's tensor processing unit and Intel's Movidius devices.

Numerous factors and subtleties need to be considered in these evaluations. Radically different programming models across these architectures complicate matters by necessitating different algorithmic solutions for the same problem. Different hardware–software combinations may optimize different metrics, such as latency, throughput, accuracy, energy or resource consumption, so the appropriate combinations need to be selected. Choices of spike-based data encodings, such as temporal versus rate coding, can profoundly affect results. Choices of parameters, such as batch size and data bit widths, can lead to different trade-offs between the evaluation metrics, so these too must be carefully analysed.

The insights gained from exploring this high-dimensional evaluation space will lead to a thorough understanding of neuromorphic architectures and how they relate to conventional solutions. The neuromorphic research community needs to converge sufficiently to support this exploration. As we are already beginning to find, orders of magnitude gains in performance and energy are possible using Intel's Loihi neuromorphic processor[2,3], with fine-grain parallelism and sparse activity providing compelling scaling trends[4]. Continued comprehensive benchmarking of this kind has the potential to move neuromorphic solutions from unsubstantiated promise to mainstream technology.

Grander challenges

The ultimate demonstration of progress and value is to move beyond benchmarks to solve a problem hitherto unsolved by computer. The most memorable milestones in AI are defined by beating human masters at problems associated with human intellect, such as chess, Jeopardy or Go. An increasingly common question in the neuromorphic field is what will be the neuromorphic Go?

As long as neuromorphic researchers are still debating the merits of benchmarking and are just beginning to quantify value at all, it seems early to be dreaming of Go-scale accomplishments. The value of neuromorphic solutions emerging today relates mainly to latency and energy gains when processing temporal data streams in a real-time, non-batched regime. On pure functional terms, neuromorphic solutions have yet to demonstrate practical results that surpass conventional solutions. This won't always be the case, but for now, while the more disruptive algorithms and scaled-up systems are still being developed, value will come from more modest application challenges that leverage the technology's energy and performance benefits.

As such, in the spirit of pitting human against machine, physical games with event-based sensing and closed-loop control provide compelling near-term application challenges for neuromorphic systems.

The game of foosball has the appropriate ingredients for an impressive yet manageable near-term neuromorphic application challenge. Foosball requires quick predictive responses to erratic ball motion, a good match for emerging event-based cameras[5] with neuromorphic processing. Operating with the same sensory input and power budget as a human player, a competitive neuromorphic agent would need to not just sense but anticipate an opponent's moves, with reinforcement-based learning of winning strategies.

Thanks to his efforts at last year's Telluride Neuromorphic Cognition Engineering Workshop, Greg Cohen of Western Sydney University has popularized foosball as a suitable application challenge, and now several groups are pursuing this project, hopefully leading to community-wide sharing of components and competitions staged between neuromorphic players at future workshops.

Longer term, marginally more grown-up applications such as drone racing and RoboCup could benefit from the power and latency advantages of neuromorphic solutions. We can also look forward to eliminating the need for cloud-based computation in mobile AI applications, offering significant privacy and latency gains. One example might be real-time speech translation with locally supervised fine-tuning of the inference model in order to better interpret a particular user's pronunciation.

Admittedly, none of these examples rises to the level of a grand challenge that would mark a new milestone in AI progress. To truly break new ground, we should look to the inherent scalability of neuromorphic systems. Neuromorphic solutions for problems like constraint satisfaction and probabilistic inference may be useful in an edge device, but scaled up, these could prove to be game changers. Combined with hyperdimensional representations[6], rapid analogical reasoning[7] and efficient associative working memory[8] — all possible with sufficient neuromorphic scale — one can see a possible path to systems exhibiting artificial creativity and generative intelligence.

Such systems would be capable of synthesizing new knowledge by recognizing novel relationships between diverse semantic entities, a hallmark of human intelligence. A true grand challenge within reach could then be solving mathematical problems — not just by proving theorems but by positing new ones — or a generalized optimization algorithm to solve difficult engineering problems with creative insights.

Fulfilling this vision of neuromorphic scaling is likely to require significant process technology innovations, particularly related to denser synaptic storage, three-dimensional integration, and post-CMOS devices. This is a large and costly space of possibility to explore. Efficient progress critically depends on embracing the right benchmarks and measuring progress in order to provide the clearest possible view of the road ahead. ❐

Mike Davies
Intel Labs, Intel Corporation, Hillsboro, OR, USA.
e-mail: mike.davies@intel.com

Published online: 11 September 2019
https://doi.org/10.1038/s42256-019-0097-1

References

1. Stewart, T. C., DeWolf, T., Kleinhans, A. & Eliasmith, C. *Front. Neurosci.* **9**, 464 (2015).
2. Davies, M. et al. *IEEE Micro* **38**, 82–99 (2018).
3. Tang, G., Shah, A. & Michmizos, K. P. Preprint available at https://arxiv.org/abs/1903.02504 (2019).
4. Blouw, P., Choo, X., Hunsberger, E. & Eliasmith, C. Preprint available at https://arxiv.org/abs/1812.01739 (2018).
5. Gallego, G. et al. Preprint available at https://arxiv.org/abs/1904.08405 (2019).
6. Kanerva, P. *Cogn. Comput.* **1**, 139–159 (2009).
7. Levy, S. D. & Gayler, R. in *Proc. First Artifical General Intelligence Conference* (AGI, 2008).
8. Frady, E. P. & Sommer, F. T. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1902653116 (2019).
9. Bellec, G., Salaj, D., Subramoney, A., Legenstein, R. & Maass, W. in *32nd Conference on Neural Information Processing Systems* (NIPS, 2018).
10. Amir, A. et al. in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017).
11. Shrestha, S. B. & Orchard, G. in *32nd Conference on Neural Information Processing Systems* (NIPS, 2018).
12. Neftci, E. O., Mostafa, H. & Zenke, F. Preprint available at https://arxiv.org/abs/1901.09948 (2019).
13. Tang, P. T. P., Lin, T.-H. & Davies, M. Preprint available at https://arxiv.org/abs/1705.05475 (2017).
14. Fonseca Guerra, G. A. & Furber, S. B. *Front. Neurosci.* **11**, 714 (2017).
15. Ponulak, F. & Hopfield, J. J. *Front. Comput. Neurosci.* **7**, 98 (2013).
16. DeWolf, T., Stewart, T. C., Slotine, J.-J. & Eliasmith, C. *Proc. R. Soc. B* **283**, 20162134 (2016).
17. Eliasmith, C. et al. *Science* **338**, 1202–1205 (2012).