University of Nottingham
UK | CHINA | MALAYSIA

Frugal neurons: Restricting plasticity to reduce the energetic costs of learning

Silviu Daniel Ungureanu, Mark C.W. van Rossum

I Introduction

- The brain consumes a significant amount of energy, roughly 20% of the body's total metabolic requirements[1].
- Plasticity is an important component of this cost[2]; fruit flies in conditioning experiments had a 20% shorter lifespan when energy-expensive long-term memory was induced and they were subsequently starved[3]. Similarly, already-starving fruit flies did not form long-term memory, and forcibly inducing it reduced lifespan by 30%[4].

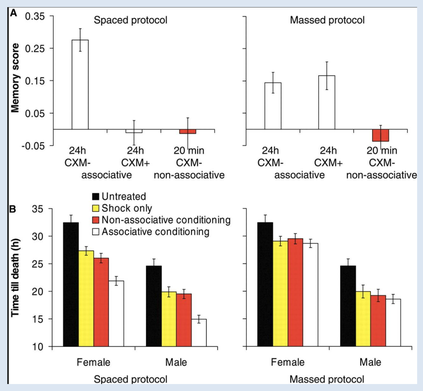

Figure 1: Mery, Kawecki: A cost of long-term learning in Drosophila(2005)

- In traditional machine learning applications, neural networks are trained by updating all synaptic connections at every trial, which is energetically inefficient. We explore the effect of energy constraints in plasticity using artificial neural networks and propose learning algorithms that reduce energy costs while maintaining learning performance.
- Energy costs were modeled as $\mathrm{M} = \sum_{i,t} |\delta w_i(t)|^{\alpha}$, for all weights w_i and timesteps t. $\alpha = 1$ corresponds to the L1 norm (**M1 energy**), while $\alpha \to 0$ gives the L0 norm, equivalent to the total number of individual weight updates (**M0 energy**).

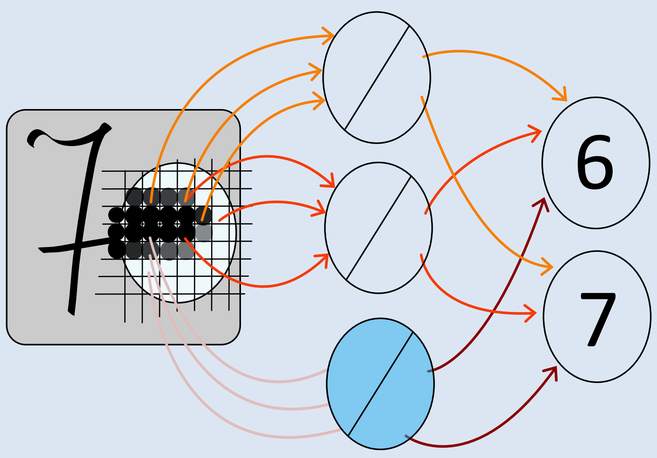

Figure 2: A schematic representation of the network architecture.

II Plastic fraction effects

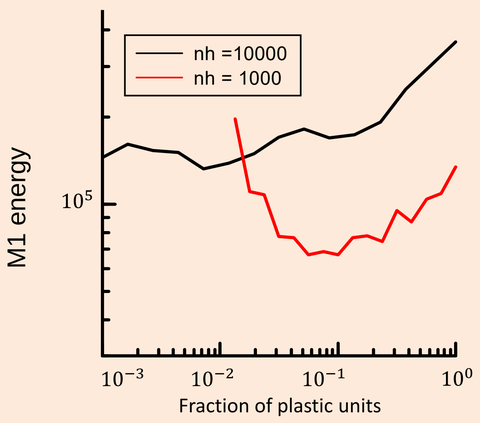

Figure 3: Scaling of M1 energy consumption with the fraction of plastic units. A standard network has a fraction P=1

- We tested plasticity-constraining algorithms that select a subset of the model's weights to be updated. Forward propagation and backpropagation are performed as normal, and only the selected weights are updated. The networks are trained on the standard MNIST dataset to a criterion of 95% accuracy.
- We compare three methods to restrict plasticity: randomly selecting a 'fixed mask', held constant throughout training, selecting the largest updates across all synapses in the network, and selecting the largest updates among each neuron's weights. We find savings with all methods. When the number of selected weights is too small, training time increases significantly, leading to higher energy consumption. Larger plasticity-constrained networks perform better than small networks with the same number of active units, the latter being unable to reach the same accuracy during training.

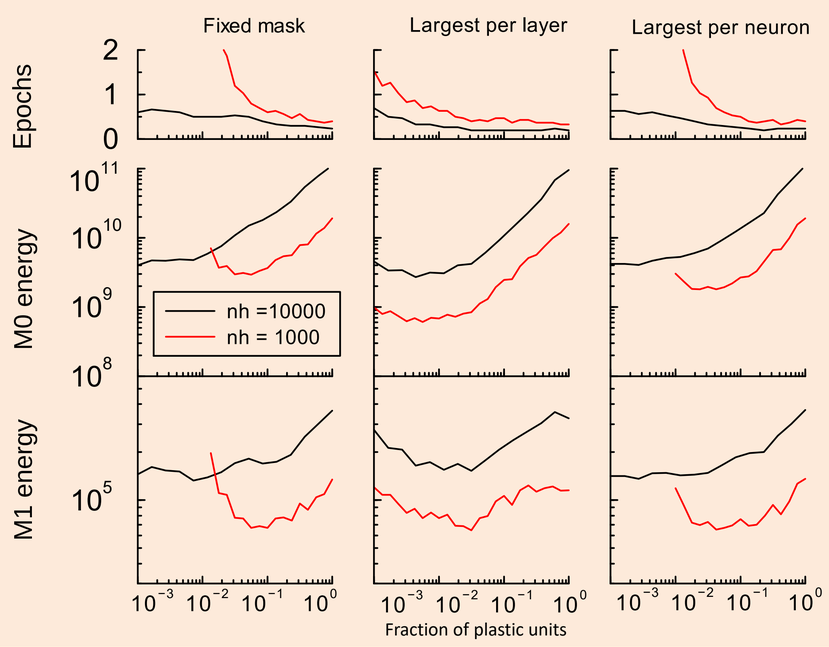

Figure 4: Scaling of energy consumption and training time, for networks with a number of hidden-layer units(nh) of 1000 and 10000, respectively

III Network size effects

- For M1 energy, the algorithms have comparable performance, all of them significantly outperforming the default, unmodified network (with a plasticity fraction P=1). For large networks, a fixed mask consumes slightly less energy than maximum selection.
- When measuring M0 energy, largest update selection outperforms a random fixed mask for medium- and small-sized networks by an order of magnitude, while for large networks, the performance of the three methods appears to be similar.

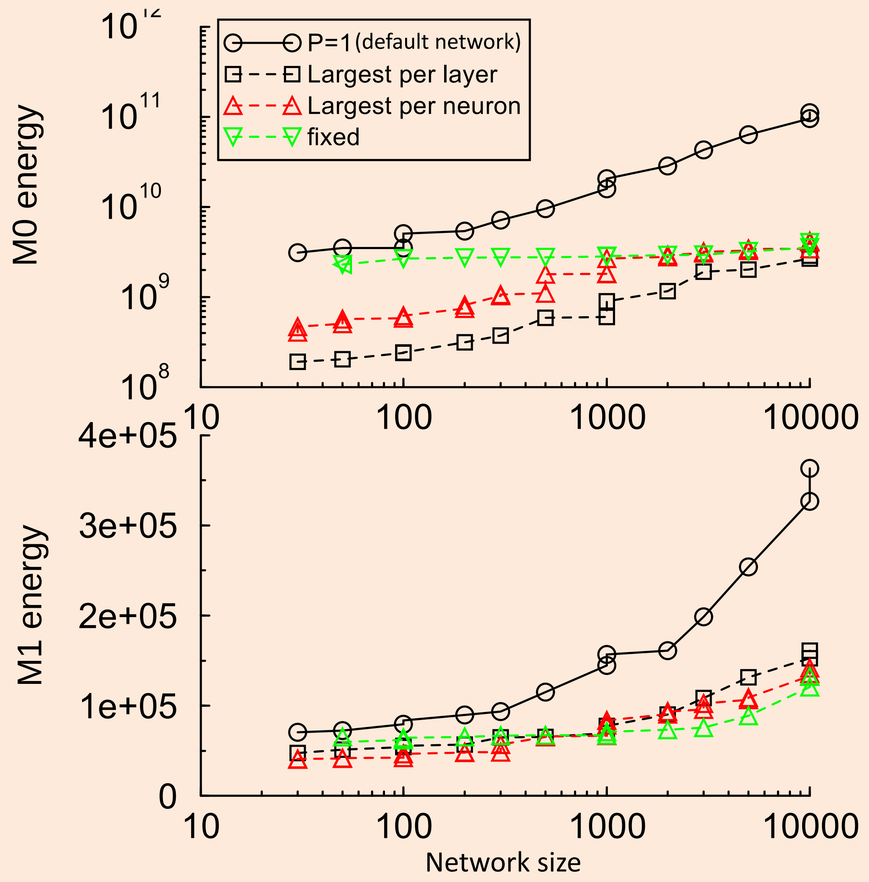

Figure 5: Scaling of optimal energy with network size, for different algorithms and the unmodified networks

IV Energy savings ratio

- The fraction of energy saved increases with the network size.
- For M1 energy, a fixed mask saves less energy for smaller networks, but more energy for large networks, compared to maximum update selection.
- For M0 energy, selecting the largest updates across all synapses gives the best performance for all network sizes, followed by per-neuron selection. However, performance appears to converge for very large networks.

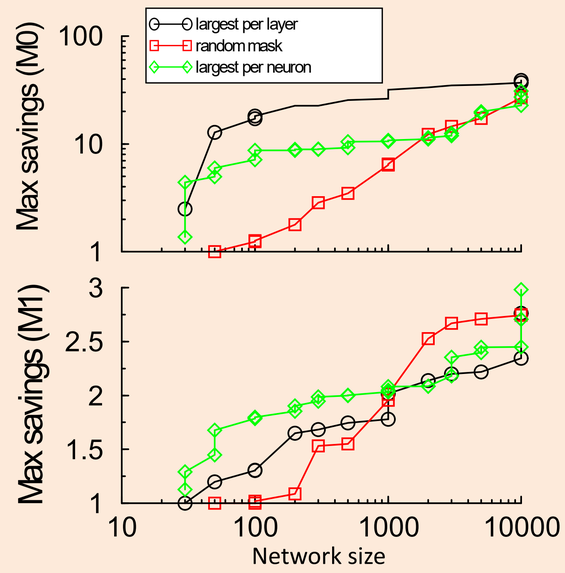

Figure 6: Maximum savings achieved by different algorithms

V Conclusions

- Restricting plasticity can lead to order-of-magnitude savings in energy without impacting accuracy.
- Optimal energy consumption is achieved when the fraction of active neurons is very small.
- The selected energy model has a significant impact, with M1 energy roughly halving as plasticity is constrained, while M0 energy decreases tens of times. However, both paradigms show savings for all three algorithms discussed.
- The largest savings are observed then the networks are much larger than the task requires, which is arguably the case in biology.

References: [1]: D Attwell and S B Laughlin. An energy budget for signaling in the grey matter of the brain. Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism, oct 2001, 21(10):1133–45; [2]: Karbowski, Jan. Metabolic constraints on synaptic learning and memory. Journal of Neurophysiology 122.4 (2019): 1473-1490., 10.1152/jn.00092.2019; [3]: Mery F, Kawecki TJ. A cost of long-term memory in Drosophila. Science. 2005, 308(5725):1148-; [4]: Pierre-Yves Plaçais and Thomas Preat. To favor survival under food shortage, the brain disables costly memory. Science, 339(6118):440–442, 2013.