Layer-skipping connections facilitate training of layered networks using equilibrium propagation.

Jimmy Gammell
Sae Woo Nam
Adam N. McCaughan

ABSTRACT

Equilibrium propagation is a learning framework that marks a step forward in the search for a biologically-plausible implementation of deep learning, and is appealing for implementation in neuromorphic analog hardware. However, previous implementations on layered networks encountered a vanishing gradient problem that has not yet been solved in a simple, biologically-plausible way. In this paper, we demonstrate that the vanishing gradient problem can be overcome by replacing some of a layered network's connections with random layer-skipping connections. This approach could be conveniently implemented in neuromorphic analog hardware, and is biologically-plausible.

CCS CONCEPTS

• **Computing methodologies → Bio-inspired approaches**; **Neural networks**.

KEYWORDS

deep learning, neural network, biologically-motivated, equilibrium propagation, vanishing gradient, small-world, neuromorphic hardware

ACM Reference Format:
Jimmy Gammell, Sae Woo Nam, and Adam N. McCaughan. 2020. Layer-skipping connections facilitate training of layered networks using equilibrium propagation.. In *International Conference on Neuromorphic Systems 2020 (ICONS 2020), July 28–30, 2020, Oak Ridge, TN, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3407197.3407618

1 INTRODUCTION

Equilibrium propagation [16] is a learning framework for energy-based networks such as the continuous Hopfield network [8]. It is appealing relative to backpropagation because it is more biologically-plausible, and as a side-effect could be implemented more-easily in neuromorphic analog hardware.

Implementation of equilibrium propagation in [16] was hindered by a vanishing gradient problem whereby networks with as few as 3 hidden layers trained slowly on MNIST [11] - a serious issue given that network depth is critical to performance on difficult datasets [19, 20] and that convergence to a low error rate on MNIST is a low bar to meet. The problem was overcome in [16] by independently tuning a unique learning rate for each layer in the network; however, this approach is unappealing because (1) it introduces additional hyperparameters to tune, (2) it would be inconvenient to implement in neuromorphic analog hardware, and (3) it has not been observed in biological systems.

The purpose of this paper is to demonstrate that in this context the vanishing gradient problem can instead be solved by randomly replacing some of a layered network's connections with layer-skipping connections. [1] Through this modification we have achieved 0% training error (out of 50,000 examples) and $\lesssim$2.5% test error (out of 10,000 examples) on MNIST using a network with three hidden layers and no regularization term in its cost function. These error rates are comparable to those of other biologically-motivated networks [1] and are roughly the same as those of the layered network with unique, manually-tuned learning rates in [16]. Our method could be implemented with relative ease in any system with configurable connectivity. Layer-skipping connections have been observed in biological brains [3], so the approach is biologically-plausible. Similar techniques have seen success in convolutional [7, 21] and multilayer feedforward [10, 23] networks. Our findings outlined in this paper suggest that layer-skipping connections are effective-enough to be appealing in contexts where simplicity and biological plausibility are important.

2 BACKGROUND AND THEORY

2.1 Equilibrium propagation

Similarly to backpropagation, equilibrium propagation [16] trains networks by approximating gradient descent on a cost function. Equilibrium propagation is applicable to any network with dynamics characterized by evolution to a fixed point of an associated energy function; our implementation is a recreation of that in [16], which applies it to a continuous Hopfield network [8]. The mathematical formulation of the framework can be found in [16].

A major reason backpropagation is not biologically-plausible is that to implement it, each neuron would need two distinct mechanisms for information transmission: one to transmit its activation to shallower neurons during the forward-propagation phase, and another to transmit error-correction information to deeper neurons during the backward-propagation phase [2]. While this is easy in a digital computer that can oversee and manipulate an entire network, it would be cumbersome in hardware (biological or otherwise) consisting of many simple, independent computational nodes with limited ability to share information. In contrast, equilibrium propagation consists of a free phase (comparable to forward-propagation) and a weakly-clamped phase (comparable to backward-propagation) during which each neuron only needs to know the activations of

This paper is authored by an employee(s) of the United States Government and is in the public domain. Non-exclusive copying or redistribution is allowed, provided that the article citation is given and the authors and agency are clearly identified as its source.
ICONS 2020, July 28–30, 2020, Oak Ridge, TN, USA
2020. ACM ISBN 978-1-4503-8851-1/20/07.
https://doi.org/10.1145/3407197.3407618

[1] This modification was inspired by small-world topology [22]; however, we have not observed a strong correlation between network performance and common metrics of small-worldness (characteristic path length, clustering coefficient, small-world coefficient).

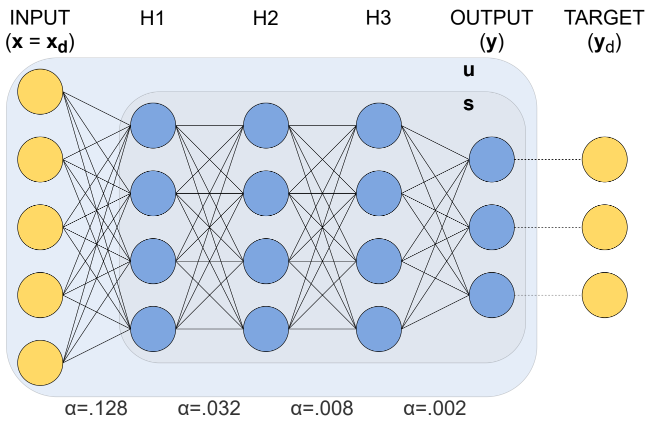

Figure 1: Topology of the layered network tested in [16]. All pairs of neurons in adjacent layers are connected. All connections are bidirectional. To compensate for the vanishing gradient problem, the learning rate is reduced by a factor of 4 each time distance from the output decreases by one layer.

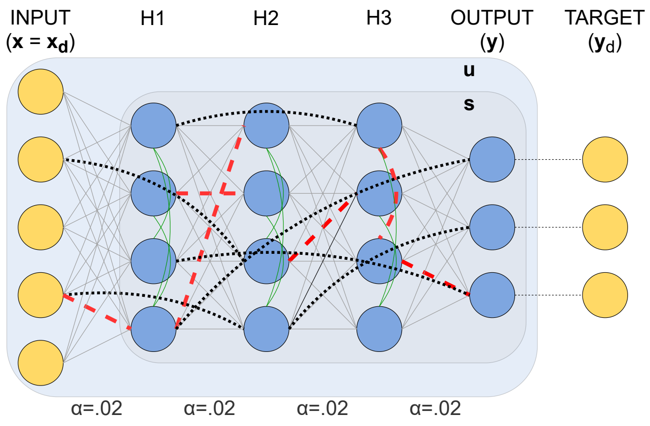

Figure 2: Our modifications to the topology of figure 1 to avoid a vanishing gradient while using a global learning rate. Red dotted lines denote connections that have been removed, black dotted lines denote their replacements, and green solid lines denote added intralayer connections. All connections are bidirectional. This illustration shows a network with $p = 8\%$.

neighboring neurons, so only one mechanism for information transmission is needed [16]. In a similar vein, to implement backpropagation each neuron would need mechanisms to compute both an activation value using activations of deeper neurons and error-correction information using that of shallower neurons. For equilibrium propagation a neuron would only need the ability to compute an activation given those of adjacent neurons [16].

2.2 Vanishing gradient problem

Vanishing gradients are problematic because they reduce a network's rate of training. and could be difficult to represent in neuromorphic analog hardware due to limited bit depth.

The vanishing gradient problem is familiar in the context of conventional feedforward networks, where techniques such as the weight initialization scheme in [6], the use of activation functions with derivatives that do not lead to output saturation [17], and batch normalization [9] have been effective at overcoming it. However, in the context of the networks trained in [16], the vanishing gradient problem persists even when the former two techniques are used. To our knowledge batch normalization has not been used in the context of equilibrium propagation; however, it seems unlikely to be biologically-plausible.

3 IMPLEMENTATION

We recreated the equilibrium propagation implementation in [16] using the Pytorch library. [2] Like the networks in [16], our networks are continuous Hopfield networks with a hardened sigmoid activation function

$$\sigma(x) = \text{Max}\{0, \text{Min}\{x, 1\}\}$$

and squared-error cost function with no regularization term

$$C = ||\boldsymbol{y} - \boldsymbol{y}_d||_2^2,$$

where $\boldsymbol{y}$ is the network's output and $\boldsymbol{y}_d$ is the target output. Tests were run on MNIST [11] grouped into batches of 20 examples, with the 50,000 training examples used for training and the 10,000 validation examples used for computing test errors.

We use two performance-enhancing techniques that were used in [16]: we randomize the sign of β before training on each batch, which has a regularization effect, and we use persistent particles, where the state of the network after training on a given batch during epoch n is used as the initial state for that batch during epoch $n+1$. Persistent particles reduce the computational resources needed to approximate the differential equation governing network evolution, and would be unnecessary in an analog implementation that can approximate the equation efficiently. Note that this technique leads to higher error rates early in training than would be present with a more-thorough approximation of the differential equation.

3.1 Layered topology with per-layer rates

We recreated the 5-layer network evaluated in [16]. It has the standard layered topology shown in 1, and consists of a 784-neuron input layer, 3 500-neuron hidden layers and a 10-neuron output layer. Weights are initialized using the scheme from [6]. As mentioned above, each layer has a unique learning rate; the rates are $\alpha_1 = .128$, $\alpha_2 = .032$, $\alpha_3 = .008$ and $\alpha_4 = .002$ where α_i is the learning rate for the connection weights between layers i and $i+1$ and for the biases in layer i, and the input and output layers are denoted $i = 1$ and $i = 5$, respectively.

[2]https://github.com/jgammell/Equilibrium_Propagation_mobile.git

3.2 Layered topology with global learning rate

To illustrate the vanishing gradient problem and provide a point of reference, we also tested the network in section 3.1 with a single global learning rate of .02.

3.3 Our topology

Algorithm 1: Algorithm to produce our topology

```
Input: Layered network from section 3.2
Input: Integer $n$, giving number of connections to replace
Output: A network with our modified topology
for hidden layer in network do
    Add edge between each pair of neurons in layer
for $i \leftarrow 1$ to $n$ do
    Randomly select pre-existing connection in network;
    Add connection between random unconnected pair of
        neurons in network;
    // Do not allow self connections
    // Do not allow connections between two
       input neurons or between two output neurons
    Remove pre-existing connection;
return modified network
```

To generate a network with our topology, we use algorithm 1. This topology is illustrated in figure 2. The above algorithm is approximately equivalent to the algorithm for generating a small-world network described in [22] with $p = 1 - (\frac{N_o - 1}{N_o})^n$ for $p \lesssim .2$, where N_o is the number of connections in the network; to contextualize the number of replaced connections we will henceforth describe networks with our topology in terms of p instead of n. We have seen good results with $p \approx 8\%$. We have seen similar results when connections are added to the network, rather than randomly replaced (algorithm 1, without removing pre-existing connections).

For these networks we use a global learning rate of .02 and, as in the networks from sections 3.1 and 3.2, initialize connections between neurons in adjacent layers using the scheme from [6]. For all other connections we draw initial weights from the uniform distribution $U[-.05,.05]$ where the value .05 was determined empirically to yield good results.

4 RESULTS

We compared the networks described in section 3 by observing their behavior while training on MNIST [11]. All networks used $\epsilon = .5$, $\beta = 1.0$, 500 free-phase iterations, 8 weakly-clamped-phase iterations, and were trained for 250 epochs.

4.1 Network performance comparison

Figure 3 illustrates that our network significantly outperforms one with a global learning rate, and achieves close to the same training and test error rates as one with unique learning rates, albeit after around 25% more epochs. Both our network and the layered network with unique learning rates achieve approximately a 2.5% test error and 0% training error, whereas the layered network with a global learning rate has test and training error rates around .5% higher than

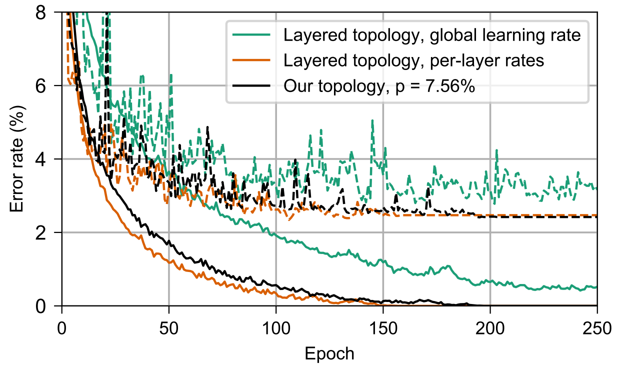

Figure 3: Performance on MNIST of the networks in section 3. Dashed lines show the test error and solid lines show the training error. In green is a layered network with a global learning rate (section 3.2), in orange is a layered network with per-layer rates individually tuned to counter the vanishing gradient problem (section 3.1), and in green is a network with our topology, $p = 7.56\%$ (section 3.3). Observe that our topology is almost as effective as per-layer rates at countering the vanishing gradient problem that impedes training of the layered network with a global learning rate.

the other two networks; it is unclear whether it would converge given enough time, but it is clear that it is inferior.

4.2 Training rates of individual pairs of layers

To observe the extent of the vanishing gradient problem, for each network we tracked the root-mean-square correction to weights in each of its layers during training on MNIST [11]. Figure 4 shows an 11-point centered moving average of these values (without averaging the values are very volatile). It can be seen that for the layered network with a global learning rate, the magnitude of the correction to a typical neuron vanishes with depth relative to the output, with the shallowest weights training around 100 times faster than the deepest weights - this illustrates the vanishing gradient problem. The use of unique learning rates is very effective at making corrections uniform. Our topology with $p = 7.56\%$ is effective at making deeper layers train in a uniform way, but the output layer still trains around 10 times faster than deeper layers; nonetheless, figure 3 suggests that this imperfect solution still yields a significant performance benefit.

The fast training of the output layer in the network with our topology is probably because no layer-skipping connections attach directly to the target output, so for any value of p the shortest path between a deep neuron and the target layer is at least 2 connections long, whereas the path between an output neuron and the target layer is only 1 connection long.

4.3 Effect of p

We evaluated the behavior of our network during its first epoch of training on MNIST for varying p and found that as p increases, the error rate decreases and the root-mean-square corrections to each layer become more-uniform. We anticipate that generalization error

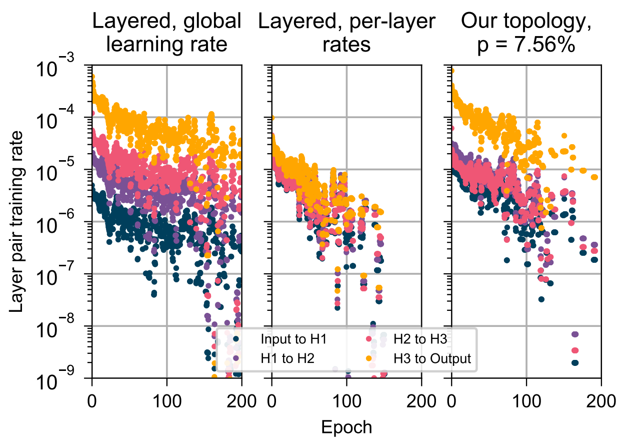

Figure 4: Root-mean-square corrections to weights in different layers while training on MNIST, for the networks in section 3. For clarity, values were subjected to an 11-point centered moving average. (left) A layered network with a single global learning rate (section 3.2). (center) A layered network a unique, individually-tuned learning rate for each layer (section 3.1). (right) A network with our topology, $p = 7.56\%$ (section 3.3). Observe that the layered topology with a global learning rate has a vanishing gradient problem, which is almost completely solved by tuning an individual learning rate for each layer. Our topology improves the situation by making training uniform among the deeper layers, although the shallowest layer still trains more-quickly than the deeper layers.

will suffer for large p as a network loses the benefits associated with a layered topology, but we did not train a large number of networks for long enough to observe such a trend.

We noticed that networks with our topology and $p = 0\%$ perform poorly relative to layered networks with a global learning rate; we believe this is due to a sub-optimal weight initialization. Networks with our topology seem less-sensitive to weight initialization than layered networks.

5 RELATED WORK

References [12, 15, 24] describe other approaches to locally approximating the gradient of a cost function. References [4, 13] explore the use of a random feedback matrix for backwards connections that is more biologically-plausible than identical forwards and backwards connections. Reference [1] explores the present state of biologically-motivated deep learning, and [2] discusses the criteria a biologically-plausible network would need to satisfy. References [5, 14, 18] discuss analog hardware that could potentially implement equilibrium propagation. References [7, 10, 21, 23] use layer-skipping connections for other types of networks and learning frameworks. References [6, 9] give approaches to solving vanishing gradient problems.

6 DIRECTIONS FOR FUTURE RESEARCH

There are several directions in which future research could be taken:

- Evaluating the effectiveness of this approach on hard datasets, such as CIFAR and ImageNet.
- Evaluating the effect of p on a network's test error.
- Exploring the effectiveness of layer-skipping connections on deeper networks.
- Exploring the effectiveness of a network when layer-skipping connections are used during training and removed afterwards.

The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon.

REFERENCES

[1] Sergey Bartunov, Adam Santoro, Blake A. Richards, Geoffrey E. Hinton, and Timothy P. Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *CoRR*, abs/1807.04587, 2018.
[2] Yoshua Bengio, Dong-Hyun Lee, Jörg Bornschein, and Zhouhan Lin. Towards biologically plausible deep learning. *CoRR*, abs/1502.04156, 2015.
[3] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature*, 2009.
[4] Brian Crafton, Abhinav Parihar, Evan Gebhardt, and Arijit Raychowdhury. Direct feedback alignment with sparse connections for local learning. *CoRR*, abs/1903.02083, 2019.
[5] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Prasad Joshi, Andrew Lines, Andreas Wild, and Hong Wang. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, PP:1–1, 01 2018.
[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
[8] John Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 81:3088–92, 06 1984.
[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.
[10] Gokul Krishnan, Xiaocong Du, and Yu Cao. Structural pruning in deep neural networks: A small-world approach, 2019.
[11] Y. LeCun and C. Cortes. The mnist database of handwritten digits, 1998.
[12] Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Y. Bengio. Difference target propagation. pages 498–515, 08 2015.
[13] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random feedback weights support learning in deep neural networks, 2014.
[14] Mitchell Nahmias, Bhavin Shastri, A.N. Tait, and P.R. Prucnal. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. *Selected Topics in Quantum Electronics, IEEE Journal of*, 19:1–12, 09 2013.
[15] Fernando J Pineda. Generalization of back-propagation to recurrent neural networks. *Physical Review Letters*, 59(19):2229–2232, 1987.
[16] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation, 2016.
[17] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan 2015.
[18] Jeffrey M. Shainline, Sonia M. Buckley, Adam N. McCaughan, Jeffrey T. Chiles, Amir Jafari Salim, Manuel Castellanos-Beltran, Christine A. Donnelly, Michael L. Schneider, Richard P. Mirin, and Sae Woo Nam. Superconducting optoelectronic loop neurons. *Journal of Applied Physics*, 126(4):044902, Jul 2019.
[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.
[20] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks, 2015.
[21] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.
[22] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.
[23] L. Xiaohu, L. Xiaoling, Z. Jinhua, Z. Yulin, and L. Maolin. A new multilayer feedforward small-world neural network with its performances on function approximation. In *2011 IEEE International Conference on Computer Science and Automation Engineering*, 2011.
[24] Xiaohui Xie and Hyunjune Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 15:441–54, 03 2003.