Are skip connections necessary for biologically plausible learning rules?

Daniel Jiwoong Im, Rutuja Patil, Kristin Branson
Janelia Research Campus, HHMI
{imd,patilr,bransonk}@janelia.hhmi.org

Abstract

Recognizing that backpropagation has been the workhorse of deep learning, it is time to explore other alternative learning methods. Several biologically motivated learning rules have been introduced, such as random feedback alignment and difference target propagation. Nevertheless, none of these methods have produced competitive performance against backpropagation. In this paper, we show that biologically motivated learning rules with skip connections between intermediate layers can perform as well as backpropagation on the MNIST dataset and is robust to various sets of hyper-parameters.

1 Introduction

The backpropagation (BP) of global error [11] has delivered tremendous success to solving hard AI tasks using deep learning. All deep learning models, such as deep neural networks, recurrent neural networks, and deep reinforcement learning, use BP as the main credit assignment tool to update their weights (model parameters) [6]. Nevertheless, the implausibility of BP for the brain is widely shared in the community. BP requires symmetric backward connectivity patterns (weight transpose) and it does not deliver the error signals through a distinct pathway. Such concerns have inspired researchers to develop biologically motivated learning rules[1] while trying to maintain the performance of artificial neural networks. The two famous learning rules are random feedback alignment (FA) algorithm [8] and difference target propagation (DTP) [7]. Instead of having symmetric weight connections, FA uses random feedback weights as a backward pathway to propagate the error information. DTP trains separate feedback neural network that produces the target activities and then minimizes the error between target activities and forward propagated activities. Several other methods were stemmed from FA and DTP, such as directed feedback alignment [9] and simplified difference target propagation (SDTP) [1]. However, *Bartunov et. al.* demonstrate not only that these methods are not scalable to problems like the ImageNet dataset, but also that the performance decays as they add more biologically plausible constraints [1]. Additionally, we find that the performance is not robust with respect to the hyper-parameters for variants of DTP methods.

The brain consists of many skip and dense connections. For example, the brain has a similar structure as residual nets, where cortical layer VI neurons get input from layer I, and skips intermediary layers [12]. Similar skip connection structure exist in multiple other layers [3]. *Oh et. al.* [10] laid out the adult mouse brain connectome and showed skip connections between inter-regions. The whole-brain and corticocortical connections can be fit by one-component lognormal distributions. In general, the log-normal distribution connectivity implies that the sparse long-range connections exist in the brain, which are the skip connections in our context [2, 10].

The performance of computer vision and natural language processing problems increased over the last five years with an increase in depth of deep neural networks. Especially with the introduction of

[1] We refer to credit assignment methods as learning rules.

Preprint. Under review.

	Sigmoid			Relu		
	Learning Rate	Early Stop	Depth	Learning Rate	Early Stop	Depth
BP	(0.1-0.001)	(200k-800k)	(3-7)	(0.001-0.00001)	(100k-300k)	(5-10)
FA	(0.01-0.0001)	(200k-800k)	(3-7)	(0.001-0.00001)	(600k-1000k)	(5-10)
DTP	(0.01-0.0001)	(200k-800k)	(3-7)	(0.001-0.00001)	(50k-150k)[2]	(5-10)

Table 1: The table provides a range of hyper-parameters explored for various learning rules. Throughout the hyper-parameter search, we explored three sets of learning rates and five sets of the early stopping starting points. We tried 0.001, 1e-3, and 1e-4 learning rates and 0.01, 0.001, and 1e-03 learning rates for the architectures with ReLU and sigmoid activations respectively. We tried 200k,400k,600k,800k, 1,000k early stop starting point for NN and DN. We tried three different range of early stop starting points that are uniformly spaced out for ConvNet and DenseConvNet.

skip and dense connection from residual neural networks [4] and densely connected convolutional networks [5], training with hundreds and thousands of layers became possible. The core idea behind the performance gain with deep and dense connections is that shorter path from earlier layers to later layers is created, and information (gradients) gets propagated more efficiently through more layers. In our experiments, we demonstrate that such type of skip connections help even more for biologically motivated methods.

Biologically motivated learning rules are yet to be combined with biologically inspired architectures. Taking inspiration from the connectivity in the brain and taking the performance advancement in deep skip and dense networks as exemplars, we claim that skip and dense connection are necessary for biologically plausible learning rules to perform as par as backpropagation. We show that FA and DTP with densely connected deep neural networks perform comparable to BP even with an increase in depth, and show that they are much more robust against different hyper-parameters compare to non-densely connected networks.

2 Methods

We use fully connected neural network and convolutional network architectures with dense connectivity. The dense connectivity refers to direct connections from any layer to all subsequent layers. More formally, we define densely connected neural network to be $\mathbf{h}_l = f_l([h_1; \cdots; h_{l-1}]; \theta_l) = \sigma(W_{l,1}h_1 + \cdots W_{l,l-1}h_{l-1} + b_l)$, where $\theta_l = \{W_{l,1}, \cdots, W_{l,l-1}, b_l\}$ are the parameters of neural network with weights $W_{l,l-1}$ connecting from layer $l-1$ to l, and $h_0 = x$. $[\cdot;\cdot]$ refers to the concatenation between vectors.

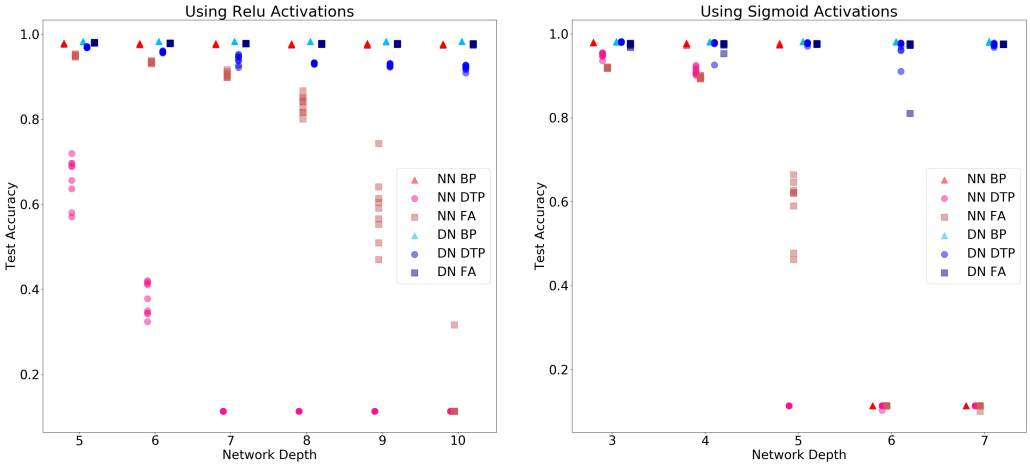
We can apply standard BP on dense network. The gradient of hidden layer l and parameter θ_l can be derived using chain rule: $\frac{d\mathcal{L}}{dh_i} = \sum_{i=1}^{L} \left(\frac{dh_i}{dh_l}\right)^T \frac{d\mathcal{L}}{dh_i}$ and $\frac{d\mathcal{L}}{d\theta_l} = \left(\frac{dh_l}{d\theta_l}\right)^T \frac{d\mathcal{L}}{dh_l}$. Similarly for FA, we can replace the transpose weight matrices with fixed random connections. For DTP, the decoder network is defined as $\hat{h}_l = g(h_{l+1}; \lambda_{l+1})$ which is learned to act as an inverse transformation $f^{-1}(h_{l+1}; \theta_{l+1})$. Then, the target activation l, $\hat{h}_l$, becomes $h_l \leftarrow h_l - \sum_{j=l+1}^{L}(g_j(h_j) - g_j(\hat{h}_j))$. We can minimize the standard difference target loss and reconstruction loss for θ and λ [7].

Note that all the above can be easily extended to convolutional neural network.

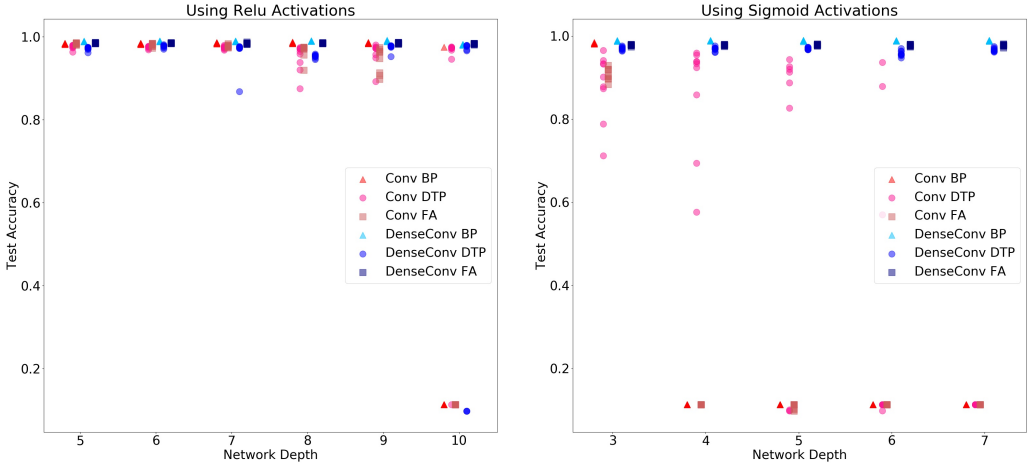
3 Experiment

We conducted our experiments on different learning methods with different network architectures on the MNIST dataset. We compared the performance of feedfoward neural network (NN) against dense neural network (DN) and convolution network (ConvNet) against dense convolutional network (DenseConvNet) for BP, FA and DTP. It is well known that BP without batch normalization suffers from vanishing gradients as the feedfoward neural network depth increases, especially with sigmoid activation function. This holds for FA and DTP as well. Therefore, we want to evaluate whether the performance of FA and DTP remains consistent with increase in network depth for densely

[2]Early stopping range between 200k-600k explored for multi-lyer perceptron (NN and DN) and 50k-150k explored for convolutional neural networks (ConNet and DenseConvNet)



(a) NN vs DN



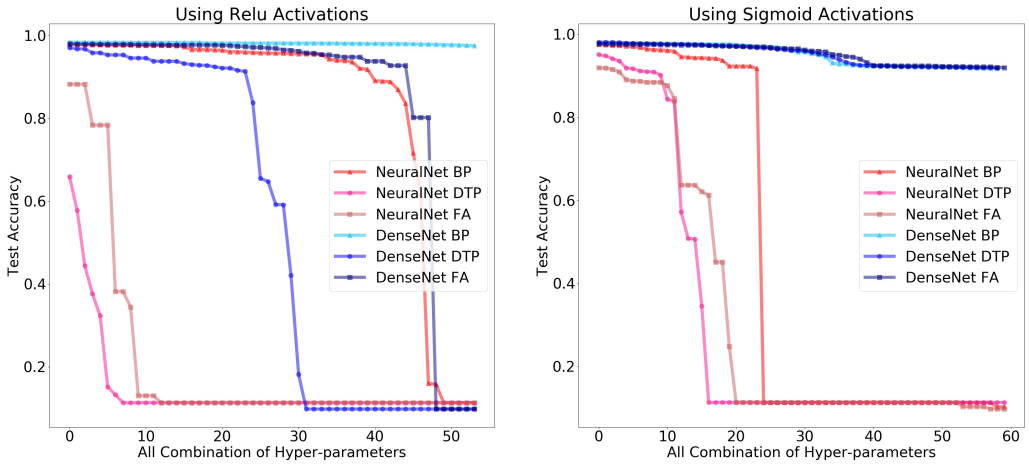
(b) ConvNet vs DenseConv

Figure 1: The performance over BP, FA, DTP with respect to different network depths.

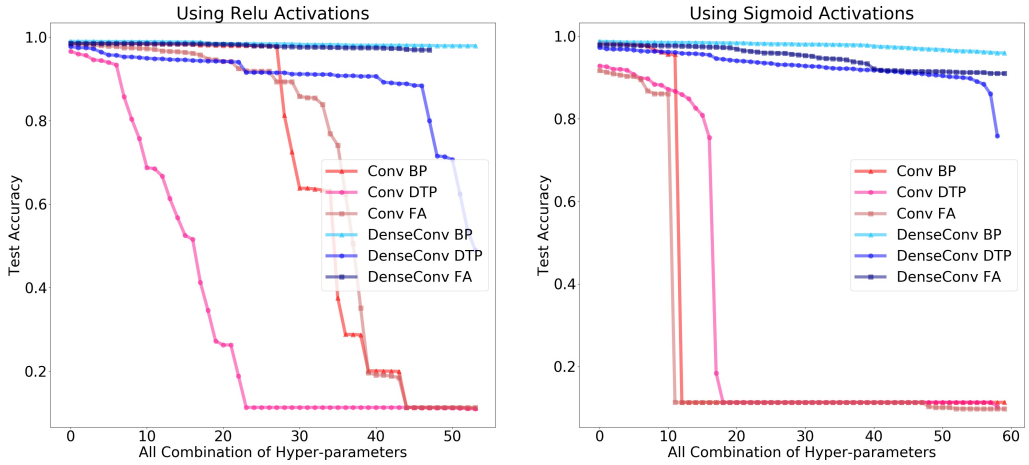
connected network. Furthermore, we want to know whether they are more robust to different sets of hyper-parameters.

In our experiments, the dataset was divided into 50,000 training, 10,000 validation, and 10,000 test dataset. We trained each network with a batch size of 128 and 0.00001 L2 weight decay coefficient. Table 1 presents the set of hyper-parameters we tested. We explored learning rates between (0.00001, 0.1) and explored early stopping criterion starting points between (20,000, 1,000,000). We used 128 hidden units for each fully connected hidden layer. We used a convolutional filter size of three and channel size is depth times 16 for convolutional neural networks. We explored both sigmoid and ReLu activation functions for multi-layer perceptron and convolutional neural networks. We measured the performance of the model with network depth from three to seven layers for sigmoid activation and five to ten layers for ReLu activation.

Figure 1 presents the test accuracy over BP, FA, and DTP with respect to network depth. The results of BP, FA, DTP are paired with NN and DN in Figure 1a, and paired with ConvNet, and DenseConvNet in Figure 1b. The best hyper-parameters for each model is chosen across 10 folds. We observe that test accuracy for all three methods drop for NN and ConvNet with network depths, whereas the test accuracy maintains for DN and DenseConvNet. This illustrates that the network did not suffer from propagating error signals all the way to bottom layers when having dense connections. It is well-known that BP suffers from vanishing gradients with deep neural networks, and yet the dense connectivity allows short pathway from the top layers to bottom layers. We suspect that FA



(a) NN vs DN



(b) ConvNet vs DenseConv

Figure 2: The sensitivity of test accuracy over multiple hyper-parameter for BP, FA, DTP. X-axis are all combination of hyper-parameter settings for learning rates, epochs, and network depth (sorted based on accuracy rate).

benefits from the dense connectivity the same way as BP. Even though DTP is a local learning rule, the activities in the top layers need to be informative in order for the local weight updates to make sense, otherwise, the local updates are based on random signals from the adjacent layers for the fully connected neural network. However, the dense connectivity will enable transfer the information from the top layers to bottom, since every hidden layer are adjacent layers from each other.

Figure 2 presents the sensitivity of test accuracy over multiple hyper-parameters. We explored combinations of learning rates, epochs, and depth, which are ordered based on sorted test accuracy. We can see that the performance of FA and DTP for NN and ConvNet (red lines) varies across a wide range. In fact, there are big accuracy discrepancy for depth 3, 4, and 5 for NN and ConvNet. However, the performance of all three learning rules for DN and DenseConvNet (blue lines) remain nearly constant lines. This illustrates that having dense connections makes the model more robust to different hyper-parameters.

4 Discussion

Building an intelligence system may require an appropriate objective/reward function, credit assignment method, specific architecture types, or even combination of all. Brain is a good archetype that illustrates what the necessity of building an intelligent system is by showing existence of dopamine rewards, error propagation, and neural connection structures. Currently, it is difficult for existing credit assignment methods to produce competitive performance under the following constraints: asymmetric backward connectivity and distinct error signal pathway. For example, the error rates for all biologically motivated learning rules were in the range of 93$\sim$98% whereas BP error rate is 63.9% on ImageNet classification challenge using AlexNet [1]. This is because we study the biologically motivated learning methods on its own rather than considering them with others together. Thus, re-examining the other key biological conditions that induce better learning performance is required. In this paper, we posit that the skip connections in our brain could be one of the key architectural conditions that will enhance existing and still unexplored methods. Through this mini experiment, we hint that biologically motivated learning rules like FA and DTP need to be composed with dense and skip connections. Furthermore, it is possible that having lognormal distributed skip connections just like the brain could be the computationally efficient way to propagate information. We leave this to future work.

References

[1] Sergey Bartunov, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and Timothy Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Neural Information Processing Systems*, 2018.

[2] György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. In *Nature Reviews Neuroscience*, pages 264—-278. 2014.

[3] David Fitzpatrick. The Functional Organization of Local Circuits in Visual Cortex: Insights from the Study of Tree Shrew Striate Cortex. *Cerebral Cortex*, 1996.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey E Hinton. Deep learning. In *Nature*, pages 436—-444. 2015.

[7] Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propagation. *European conference of machine learning*, 2015.

[8] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. In *Nature Communication*, pages 207—-214. 2016.

[9] Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. *Neural Information Processing Systems*, 2016.

[10] Seung Wook Oh, Julie A. Harris, and Hongkui Zeng. A mesoscale connectome of the mouse brain. In *Nature*, pages 207—-214. 2014.

[11] David E Rumelhard, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagation errors. *Nature*, 1986.

[12] Alex M. Thomson. Neocortical layer 6, a review. *Front. Neuroanat*, 2010.