The many directions of feedback alignment

Brian Cheung (bcheung@berkeley.edu)
Redwood Center for Theoretical Neuroscience, BAIR
UC Berkeley

Daniel Lu Jiang (danieljiang@berkeley.edu)
Redwood Center for Theoretical Neuroscience
UC Berkeley

Abstract

Backpropagation is a key component in training neural network models which have been successfully applied to many perceptual tasks including vision and audio. Despite this success, an analogous learning mechanism has yet to be discovered in biology. The feedback alignment algorithm relaxes the constraints imposed by the backpropagation procedure while still demonstrating successful learning in feedforward neural networks. In this work, we further loosen the constraints of these learning algorithms by removing the directionality constraint of the forward and backward paths. We show these paths can be operated in a *bidirectional* manner to train multiple networks without interference even when the networks are solving different tasks.

Keywords: learning; backpropagation; neural networks

Introduction

Neural networks have been shown to perform remarkably well on a variety of machine learning tasks in various domains. They have made substantial improvements in speech recognition (Hannun et al., 2014), image classification (Krizhevsky, Sutskever, & Hinton, 2012) and machine translation (Sutskever, Vinyals, & Le, 2014). Despite the wide variety of architectures used for these applications, these models all share the same fundamental learning algorithm known as *backpropagation* (Rumelhart, Hinton, & Williams, 1986). The success of this training algorithm has raised the question of whether biological neurons could be employing a similar procedure (Scellier & Bengio, 2017).

One hope of finding an appropriate analogy between artificial neural networks and their biological counterparts is the possibility of discovering more robust and efficient learning mechanisms. Recent work has investigated the properties of learning as departures are taken from the standard backpropagation algorithm (Jaderberg et al., 2016; Liao, Leibo, & Poggio, 2016; Dean et al., 2012). Dean et al. (2012) showed weight updates could be performed asynchronously. Going further, Jaderberg et al. (2016) decoupled the relative ordering of forward and backward passes.

Feedback alignment (Lillicrap, Cownden, Tweed, & Akerman, 2016) has been a significant milestone in the path to finding a learning algorithm that has comparable performance with backpropagation in training multilayer neural networks while also being less restrictive. In standard backpropagation, the error signal is propagated through the feedback path via multiplications by the transpose of the feedforward weights. This requires a precise coupling between the weights in the feedforward and feedback paths of the neural network, where the weights along the feedback path must be exact symmetric copies of the weights along the feedforward path. This coupling is known as the *weight transport problem* and is believed to be biologically implausible (Grossberg, 1987), because it requires that the learned feedforward weights be "transported" to the feedback path in order to generate learning signals for the neural network.

Feedback alignment offers a solution to the weight transport problem in which the weights along the feedback path of a neural network are randomly initialized and static. This simple solution removes the coupling constraint while still demonstrating successful learning.

In this work, we go further by making the backward weights dynamic and having them serve multiple purposes. Specifically, the backward weights serve:

- their original purpose as the backward weights for a neural network
- a secondary purpose as the forward weights for a different neural network

In contrast to the original feedback aligment algorithm, the secondary purpose changes the backward weights over time.

Our hypothesis is that given the additional flexibility of feedback alignment in training a neural network, we can utilize the backward weights to function as the forward weights for another neural network. In our work, we show that despite the multiple uses of the backward weights in a network, each network still trains without interference from the other network.

Overview of Feedback Alignment

Feedback alignment has been empirically shown to train fully-connected neural networks to a similar performance to standard backpropagation. The normal backpropagation formulation for a given weight matrix is:

$$\frac{\partial \mathcal{L}}{\partial W^l} = \frac{\partial \mathcal{L}}{\partial z^l} h^{(l-1)T} \tag{1}$$

which for a standard feedforward neural network with nonlinearity $\sigma()$ is written out as:

$$z^l = W^l h^{l-1} + b^l \tag{2}$$

$$h^l = \sigma(z^l) \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial z^l} = (W^{(l+1)T} \frac{\partial \mathcal{L}}{\partial z^{(l+1)}}) \odot \sigma'(z^l) \tag{4}$$

where $\odot$ is element-wise multiplication.

In feedback alignment, the transposed weight matrices $W^{(l+1)T}$ in equation 4 for each layer are replaced by:

$$\frac{\partial \mathcal{L}}{\partial z^l} = (B^{(l+1)T} \frac{\partial \mathcal{L}}{\partial z^{(l+1)}}) \odot \sigma'(z^l) \tag{5}$$

where $B^{(l+1)T}$ is a fixed randomly initialized weight matrix. This random matrix relieves the constraint of backpropagation where the transpose of the forward weight matrix is normally needed on the feedback path of the network.

Bidirectional Feedback Alignment

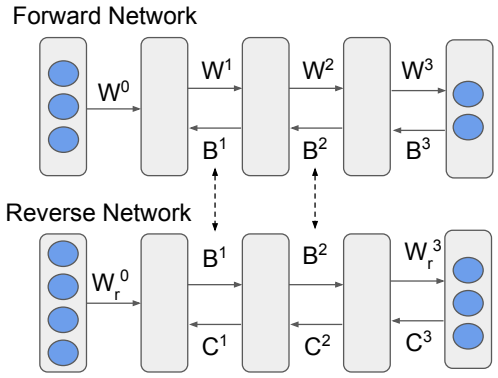

Figure 1: Diagram of our bidirectional feedback method. Only weights of the hidden layers are coupled to match in number of units.

Recent work has shown that a significant fraction of the parameters of a neural network can be removed after training without compromising accuracy (Han, Pool, Tran, & Dally, 2015; Frankle & Carbin, 2018). The work of Frankle and Carbin (2018) suggests that many of the parameters in large neural networks are only utilized for increasing the chances of sampling a good initialization. Otherwise, these weights are left unused during learning.

With this large fraction of unused parameters in mind, we hypothesize that the backward weights of a neural network trained with feedback alignment could be used for multiple purposes without impacting the overall learning process of the forward weights. If only a small random fraction of the backward weights are used for one purpose, there is a low chance that they would interact when co-opted for another purpose. Rather than having the feedback weights carry only gradient information to the lower layers of a neural network, we investigate using these weights as the feedforward weights of a different neural network:

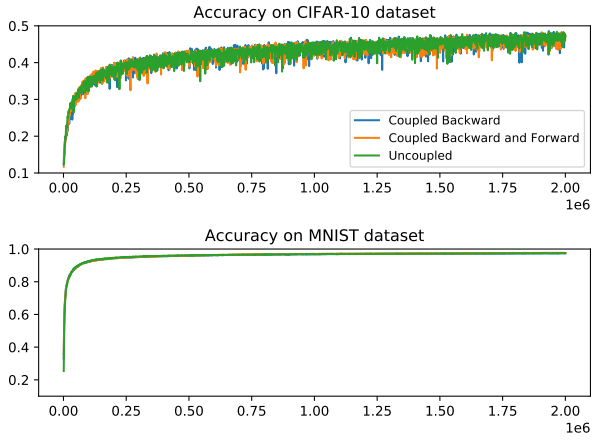

Figure 2: Test accuracy during training of standard Feedback Alignment (green), Coupled Backward (blue), and Coupled Backward and Forward (orange) on the CIFAR-10 (top) and MNIST (bottom) datasets.

$$z_r^l = B^l h_r^{l-1} + b_r^l \tag{6}$$

$$h_r^l = \sigma(z_r^l) \tag{7}$$

In contrast to the original feedback alignment procedure, the backward weights B^l here are not kept fixed throughout learning. These weights are updated by another feedback alignment procedure to train a *reverse network* whose components are denoted by the subscript r.

Backward and Forward Coupling In the previous section, the reverse network possessed its own backward matrices C^l which are randomly initialized and fixed (see Figure 1). We go further to investigate whether coupling the backward matrix of the reverse network to the forward network has any impact on learning using feedback alignment:

$$\frac{\partial \mathcal{L}_r}{\partial z_r^l} = (W^{(l+1)T} \frac{\partial \mathcal{L}_r}{\partial z_r^{(l+1)}}) \odot \sigma'(z_r^l) \tag{8}$$

Experiments

We train two distinct 5 layer fully connected neural networks with 1024 units in each hidden layer. One network is trained on the CIFAR-10 task while the other is trained on the MNIST task. To improve convergence speed, we center the pre-activations across the feature dimension. This centering is equivalent to multiplying by a fixed symmetric matrix in the forward and backward directions. Aside from improving learning speed, we found this centering did not impact our results or conclusions.

Results

We compare the accuracy during learning of standard feedback alignment with bidirectional feedback alignment. In standard feedback alignment, the forward and reverse networks

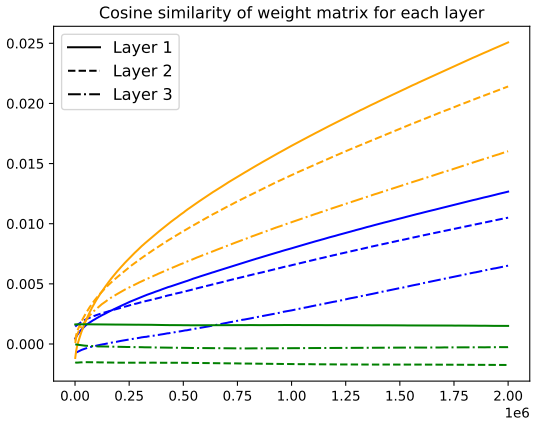

Figure 3: Cosine similarity between weights in the forward and reverse networks during training iterations for standard Feedback Alignment (green), Coupled Backward (blue), and Coupled Backward and Forward (orange).

are independently learning over the CIFAR-10 and MNIST datasets respectively. Figure 2 shows that coupling the backward matrix of the forward network with the reverse network converges at the same rate for both datasets. Furthermore, fully coupling the forward and reverse networks (Backward and Forward coupling) also converges at the same rate as standard feedback alignment.

To determine if there are any interactions while simultaneously training the forward and reverse networks, we compute the cosine similarity of the weight matrices of each coupled layer which is shown in Figure 3. Unsurprisingly, when the networks are uncoupled (standard feedback alignment), the weights in all layers have nearly zero cosine similarity. In contrast, when the forward and reverse networks are coupled by the backward matrix, the alignment between the weights increases as training progresses. When the forward and reverse networks are fully coupled, we see the greatest alignment during training. Interestingly the alignment is larger in the earlier layers when compared to the downstream layers.

Discussion

We show preliminary results demonstrating that the feedforward and feedback pathways in a neural network can travel along the same connections, further relaxing the constraints which were initially thought necessary for training neural networks architectures. Instead of keeping the backward weights fixed during learning, we allow the backward matrix to be updated to solve a different task. We show that this does not impact learning performance of the original feedback alignment algorithm.

We believe that tempering the restrictions of when learning is feasible will lead to algorithms which are more resilient to unforeseen changes and may make it easier to find analogies in biology.

Acknowledgments

We would like to thank the members of the Redwood Center of Theoretical Neuroscience for fruitful discussions and support.

References

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., ... others (2012). Large scale distributed deep networks. In *Advances in neural information processing systems* (pp. 1223–1231).

Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Training pruned neural networks. *arXiv preprint arXiv:1803.03635*.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive science*, *11*(1), 23–63.

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems* (pp. 1135–1143).

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... others (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Jaderberg, M., Czarnecki, W. M., Osindero, S., Vinyals, O., Graves, A., Silver, D., & Kavukcuoglu, K. (2016). Decoupled neural interfaces using synthetic gradients. *arXiv preprint arXiv:1608.05343*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Liao, Q., Leibo, J. Z., & Poggio, T. A. (2016). How important is weight symmetry in backpropagation? In *Aaai* (pp. 1837–1844).

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, *7*, 13276.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533.

Scellier, B., & Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, *11*, 24.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).