Improving Random Feedback Alignment Using Transfer Learning

Joshua T. Loong[*]

[*] Faculty of Science, University of Waterloo

Abstract:

There has been increasing interest in how biological principles exhibited by the human brain could help inform development of more robust artificial neural networks. This work introduces a new method to try and improve random feedback alignment, a biologically plausible alternative to backpropagation, and its performance in new domains by using a feedback matrix generated through random initialization over the latent space representation of weights transferred from another domain.

Introduction & Background:

Deep learning and artificial neural networks have garnered a significant amount of attention for their successful application in a variety of domains with high dimensional data such as computer vision and natural language processing (Liu et al., 2017). Several advancements, of which the increased availability of detailed training data and computing power are most prominent, have contributed to this explosion in interest in, and the ability of, these methods since the methods were first developed in the eighties (Goodfellow, Bengio & Courville, 2016).

Although inspired by the workings of human brain neurons, backpropagation, one of the key components of modern artificial neural networks, has been demonstrated to be biologically implausible. There are a several reasons for this, but the one primarily addressed by this study has been described as the weight transport problem (Lillicrap et al., 2016). Backpropagation has found incredible success because it allows the upstream weights in the network to be updated using the copies of downstream errors. This is relatively trivial to do exactly and precisely in modern computers, but it as a pattern of behavior that is not observed in the brain. As early as the late eighties, neuroscientists spoke of the fact that the existence of backpropagation in the brain would require that information be rapidly transmitted backwards through their axons, and that this appeared quite unlikely to occur (Crick, 1989). Lillicrap et al. remarked in their 2016 paper:

> *…whilst the brain does exhibit widespread reciprocal connectivity that would be consistent with the transfer of error information across layers, it is not believed to exhibit such precise patterns of reciprocal connectivity.*

It was due to this disconnect that researchers began proposing alternatives to backpropagation that could be biologically plausible and that avoided this weight transport problem. In 2016, Lillicrap et al. published their work on random feedback alignment (RFA), a method where random, static weights were used to initialize the network along the feedback path. Specifically, whereas backpropagation networks learn via the exact transpose of the error matrix (δ_{BP}=W^Te), RFA conducts learning through a random projection of the error matrix (δ_{FA}=Be, where B is a random fixed matrix). This method was remarkable given its comparable performance on the MNIST data set to backpropagation and its avoidance of the weight transport problem. Since then a number of feedback alignment variants have been proposed expanding on progress towards more human-like machine learning (Nøkland, 2016).

In fact, due to a human's remarkable learning capabilities it is believed that further incorporating mechanisms observed in biological neural networks may lead to the increased robustness, generalizability, and/or efficiency of artificial neural networks (Cheung

& Jiang, 2018; Kruger et al., 2013; Lillicrap et al., 2016). Conversely, it has been expressed that furthering our understanding of deep learning may lead to better insight into the workings of our brains (Tripp, 2018). It was for these reasons that this study wanted to expand on random feedback alignment by incorporating transfer learning, another biologically inspired method used commonly in machine learning.

Transfer learning, or transfer of learning, is a concept that initially originated in psychology. It is intuitive and has been observed that, at a higher-level, humans can use learnings from past experiences to help make learning new tasks easier (Helfenstein, 2005). At a much lower-level, it has also been observed that organisms, such as honeybees, have brain mechanisms to help them recognize novel stimuli using transferred representations of stimuli it had seen before (Giurfa, 2008). This remarkable ability of biological brains sparked interest in the ability to apply this concept to machine learning. One of the largest issues in machine learning is that models are always trained to fit their specific domain and moving to new data becomes problematic for a variety of reasons. Due to this gap, this transfer learning has become remarkably popular in recent years to address this issue. The process involves training a network on one domain with lots of data, and transfer aspects of the model weights to improve learning on a new domain; whether in respect to training time or efficiency, or model robustness (Hendrycks, Lee & Mazeika, 2019; Pan & Yang, 2010).

It was the belief of this author that one could combine these two ideas: feedback alignment and transfer learning. The original feedback alignment method proposed a completely random instantiation of the feedback matrix. However, it is unlikely and probably inefficient, that the brain starts learning from a completely random point. In this work, a network was trained on one domain and a feedback matrix for the new domain was created randomly on a representation of these weights that attempted to retain information about the previous domain. It was this semi-random matrix that was then transferred to a feedback alignment network to train on a new domain. With this transfer learning variant of feedback alignment, this paper contrasts it to the original feedback alignment method and traditional transfer learning with backpropagation to understand any potential performance gains.

Methods:

This project made use of two data sets: the MNIST hand written digits and the EMNIST hand written letters (Cohen et al., 2017). Feedback alignment has shown to be very comparable in performance to backpropagation on the handwritten digit data set as was outlined in the initially Lillicrap et al. paper (2016). It was thought that there would be definite potential that components of this data set would be transferrable to performance on the letter data set.

A three layer fully connected network was trained on the MNIST data set using regular backpropagation with 784 hidden units in the first layer and 10 in the output layer.The learning rate was set to 0.5 and was trained for 300 epochs. The final weights and error values were exported into binary files. It was not key that this network achieve high performance as the purpose was just to have a starting point that one could transfer from.

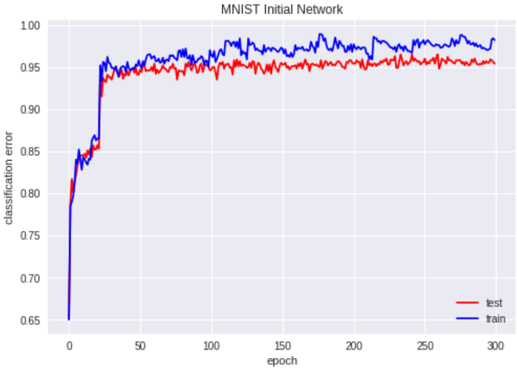
To iterate on potential transfer methods, an experimental simulator was created to measure their effectiveness. In Lillicrap et al., the authors proposed that RFA achieves similar performance to backpropagation when $e^TWBe > 0$ on average (2016). Although this applies more strictly in a vector sense, what this formula implied was that the random matrix was pushing the teaching signal to be within approximately 90º of the signal that would have been used in backpropagation. With an implementation of this formula, one could simulate how effective new methods were at approximating the teaching signal given the weight matrix and error information exported from the initial network.

Using this simulator, a method was derived that pushed the teaching signal in a stronger direction than what would have happened using the original RFA. Using principal component analysis (PCA), a method of dimensionality reduction invented by Pearson, the final layer weights of the network were decomposed into a lower dimensional space that attempted to capture the variance and information encoded within those weights (1901). A random uniform distribution was created on this the latent space by utilizing the mean as the lower bound and the mean in addition to one standard deviation as the upper bound. The new feedback matrix was created by transforming this random space back into the original weight dimensions.

Several final networks were trained to compare performance. The original RFA, where the feedback matrix was drawn on a random uniform distribution between -0.5 and 0.5, and the new transfer RFA method proposed above were both trained back on the original MNIST digit data to gauge performance. Finally, on the EMNIST letter data the original RFA, the new transfer RFA, and regular backpropagations with traditional transfer learning methods were all trained to compare performance. These were all three layer fully connected networks with 784 units in the first layer and 10 in the output layer with learning rates of 0.5 The RFA methods were trained for 150 epochs, while the backpropagation networks were trained for 300 epochs.

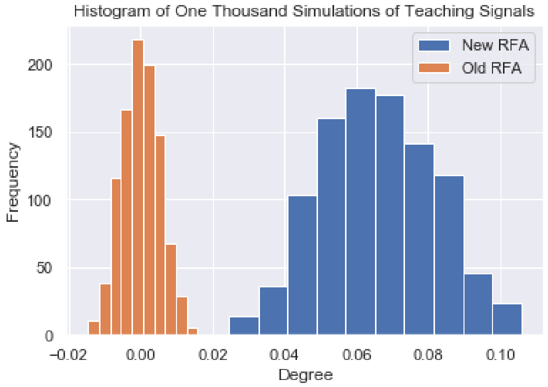
Results:

The performance of the initial MNSIT network trained with regular backpropagation can be seen below:



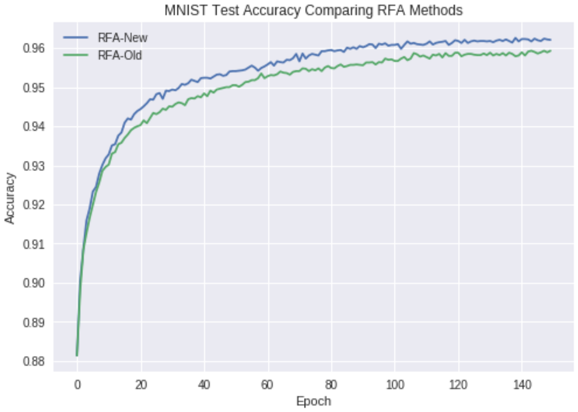
This network achieved acceptable performance, and the weight and error information were exported to be used in transfer learning.

As part of experimentation, this exported information was used to compare the methods based on their teaching signals. Under one thousand simulations, the new transfer RFA method appeared to push signals in a more favorable direction than the old RFA method that used a random uniform distribution between 0.5 and -0.5.



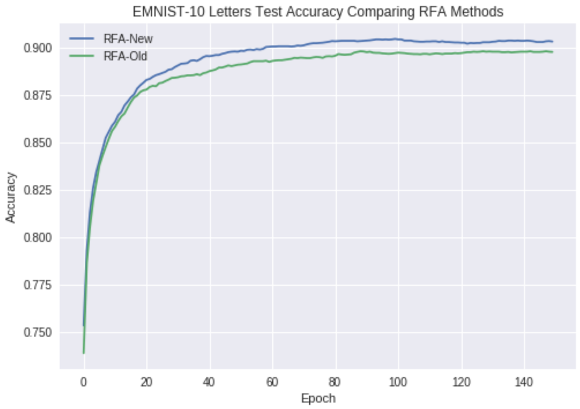
The new method appeared to push the teaching signal in a direction that was greater than zero significantly more than the old RFA.

Although this was promising, it was imperative that the method be further tested before proceeding. It was for this reason that the new transfer RFA method was trained on MNIST against the old RFA method. The results can be seen below:

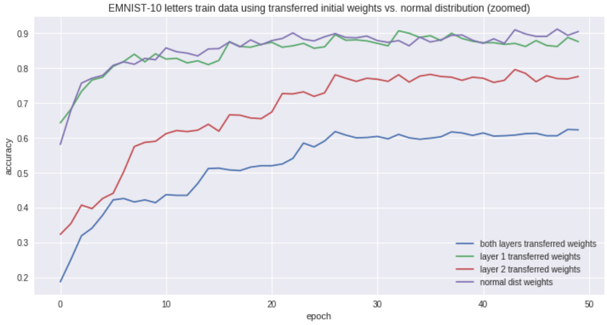


Although somewhat minor, one can see an increased slope and converged test accuracy after using the proposed transfer method. With this promising result, the final networks were trained.

The two RFA methods were trained on the first ten letters of EMNIST letter data sets. With results below:



The proposed transfer RFA method continued to provide a lift in final test accuracy as compared to the original RFA method on the new domain. The ability to transfer information encoded in the previous network appeared to be successful in some respect.



For further comparison, again using the first ten letters in the EMNIST letter data set, the following networks were trained: regular backpropagation, backpropagation with layer one transferred, backpropagation with layer two transferred, and backpropagation with both layers transferred. Although the network was trained for 300 epochs, the zoomed in results of the first 50 can be seen below:

From this one can see that the first layer transferring was more successful than the other methods and much more comparable to regular backpropagation.

When comparing these traditional methods with the new transfer RFA, it was observed that transfer RFA has the advantage of starting from a higher test accuracy in the first epoch. Furthermore, by the fiftieth epoch the transfer RFA had reached comparable performance to the best of the backpropagation techniques with a much higher slope. However, it should be noted that all of these improvements were only marginal and further experimentation is definitely required to understand how these findings generalize to other data.

Discussion:

The results shown here present an interesting iteration of the feedback alignment method. Creating the feedback matrix based on a random instantiation of a latent representation of weights transferred from a similar domain appeared to yield performance gains, albeit minor, in the new domain. The transferred matrix was appeared to be able to retain some information that was helpful in the learning process when compared to the original RFA method. This was further promising when compared to traditional backpropagation and backpropagation with transfer learning, as the new method was able to perform just as comparably in the long run but start from a higher accuracy and improve at a higher rate.

From a biological perspective, it is obvious that there are no mechanisms in the brain that are conducting principle component analysis. However, part of the basic principles of the weight transport problem are that there are no mechanisms for the brain to transfer exact error or weight information given the current understanding of biological neural networks (Crick, 1989; Lillicrap et al., 2016). However, if it is observed at a higher and lower level that brains can transfer some type of learned representations of information to help with new tasks and domains, then it must be conducting this in a way that is effective yet consistent with the weight transport problem.

It was from this interplay of ideas that the proposed method arose. PCA was merely a means to create a lower dimensional representation of previously learned information from which a feedback matrix could be created that satisfies both: the inherent idea of preserving important information for learning new tasks, and not relying on the precise transfer of weights. The implantation of this idea showed that the method was able to be relatively successful within the two domains that were chosen for this study.

However, it is quite possible that PCA was not the most accurate or efficient method for dimensionality reduction. Further experimentation in this direction may be worthwhile, especially drawing from other work that also use latent representations of weight parameters for learning such as work done by Praider et al. (2018). This would be in addition to experimenting with performance improvements in deeper networks, higher dimensional data, and different ways to draw random distributions in the latent space. If consistently improved results can be reported empirically, it would also be of use in the future to more concretely formulate mathematically the behaviour being observed. On the other side of the spectrum, it would be useful to do more neurological studies trying to understand the exact mechanisms at work happening when humans conduct transfer learning at the lower and higher level. Better insight into this process may translate into more meaningful directions on which to implement this in artificial neural networks. Hopefully the methods described here can provide further validation that incorporating neuroscience concepts into machine learning methods can better the understanding of both fields.

References:

Cheung, B., & Jiang, D. L. (2018). The many directions of feedback alignment. *Conference on Cognitive Computational Neuroscience*

Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373.*

Crick, F. (1989). The recent excitement about neural networks. *Nature, 337(6203),* 129-132.

Giurfa, M. (2008). Behavioral and Neural Analysis of Associate Learning in the Honeybee. *Learning and Memory: A Comprehensive Reference, 561–585.* doi:10.1016/b978-012370509-9.00067-x

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Helfenstein, S. (2005). General Discussion. *Transfer: Review, reconstruction, and resolution* (No. 59). University of Jyväskylä.

Hendrycks, D., Lee, K., & Mazeika, M. (2019). Using Pre-Training Can Improve Model Robustness and Uncertainty. *arXiv preprint arXiv:1901.09960*.

Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., ... & Wiskott, L. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision?. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1847-1871.

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.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11-26.

Nøkland, A. (2016). Direct feedback alignment provides learning in deep neural networks. In *Advances in neural information processing systems* (pp. 1037-1045).

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (2018). Latent Projection BNNs: Avoiding weight-space pathologies by learning latent representations of neural network weights. *arXiv preprint arXiv:1811.07006.*

Tripp, B. (2018). A deeper understanding of the brain. *NeuroImage*, *180*, 114-116.