Biologically Plausible Machine Learning with Privacy

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sequential backpropagation of error is the method by which nearly all deep learning networks are trained today. Yet there are questions regarding its biological plausibility as humans do not learn in this way. There have been a number of recent research papers on algorithms that are motivated by human physiology an davoid known implausibilities like the weight transfer problem. This paper explores how two of these proposed methods compare to standard backpropagation on both unaltered and noisy datasets, the latter of which will be defined using a definition of differential privacy. The first training scheme is direct feedback alignment; this method updates the weight matrices applied to the model's input in parallel rather than sequentially as in backpropagation. The second is target propagation which computes targets rather than the gradient at each layer. With the use of clean and noisy datasets, the relative robusticity of each training method can be compared and it may be found that there are cases where one of the non-traditional schemes outperforms backpropagation.

1 Introduction

The majority of deep neural net architectures use backpropagation[1] (BP) to update the weights or parameters which are used to construct the model. Large neural networks trained using BP have produced models with incredible accuracy for classification tasks in recent years. However, backpropagation does not seem to be biologically plausible for a number of reasons. First, it requires symmetric (the same) weights to be used for both the forward and backward pass calculations; this is the weight transfer problem. Second, the error derivative which is backpropagated sequentially requires exact knowledge of the values for the derivatives of the non-linearities in the network. This would be equivalent to human synapses being capable of bi-directional communication when they are in fact uni-directional. Chemical synapses consist of the axon terminal of one cell transmitting across a synaptic cleft to the dendritic end of another cell. Although unrelated to biology, another limitation to the sequential requirements of BP when updating the model weights during the backward pass is that this process cannot be parallelized because of the dependencies between layers.

The two novel methods of training explored here are direct feedback alignment[2] (DFA) and target propagation[3] (TP). They avoid the weight transfer problem necessitated by BP albeit in different ways. When BP makes its forward and then backwards pass, it is traversing through the same network which is the problem in short. If the body does use such a system, the backwards pass would have to be at the very least through a different network than the forward one. DFA and TP are able to create update values for the forward model's parameters without having direct access to their information during the backwards pass. Direct feedback alignment is capable of updating all layers at the same time which allows for computational parallelization. Target propagation sets target values for the activation at each layer which converge to approximate the gradient direction rather than directly computing the gradient at each layer. These methods are covered in-depth during Section 2 of the paper.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

In addition to comparing these training methods using standard datasets, the paper shows how they perform in a private-setting. Big data companies have had nearly unfettered access to individual's information. However this is changing as seen by the multiple lawsuits brought against Facebook and pieces of legislation such as the General Data Protection Regulation. Differential privacy[4] is a mathematically defined notion of privacy; this allows for ensuring certain levels of privacy as expressed by epsilon in the formula. While BP has been adapted to differential privacy in a number of ways such as using noisy data, gradient clipping or a noisy gradient, it is worth seeing how the two novel training schemes discussed in the paper perform while fulfilling the same level of privacy. Noise tolerance can be an important measure of a model's worth. Some of the best image classifiers will quickly drop to single digit accuracy if the images are injected with small amounts of RGB pixel noise or put through a filter, even though to a human the change is nearly imperceptible. A more biologically plausible training method should be able how to learn these generalization skills more easily while having the advantage of being more noise tolerant in a differentially private setting.

1.1 Biologically Plausible Training

While BP performs well for nearly all tasks, there are questions of where and how machine learning might be improved by using what is known about how humans learn. Such training methods may become useful as neuroscience progresses alongside artificial intelligence and the brain can be modelled. Another apparent drawback of BP is its weak generalizability; most models are tasked with narrow types of classification but humans and general systems AI must perform well on a wide range of tasks which requires greater plasticity in how they can learn. While this may turn out to depend more on the architecture of networks, it is possible that gains will be seen in these novel models when biologically plausible architectures are combined with corresponding learning techniques.

This question of robusticity and plasticity is a facet of biologically plausible learning called "behavioural realism" by Bartunov, et al. The inability of neurons to communicate through the same network bi-directionally is a limit of "physiological realism". Whitting points out that "without local error representation, each synaptic weight update depends on the activity and computations of all downstream neurons. Since biological synapses change their connection strength based solely on local signals (e.g., the activity of the neurons they connect), it appears unclear how the synaptic plasticity afforded by the back-propagation algorithm could be achieved in the brain."[5] A novel algorithm capable of avoiding weight transport and is more robust to noise presents clear advantages.

2 Training Deep Networks

All three of the proposed training methods are alike in that they have two distinct phases of a forward calculation and then backwards error propagation. The reason why the standard BP algorithm is not biologically plausible is that it has direct access to information about the model during the forward pass when calculating how its parameters should update with respect to the error during the backwards pass. The error is a measured difference between a predicted y for a given input x to the model, and the actual output y for that same instance x. As shown below, update values for the forward model are calculated by finding the gradient of the error with respect to those forward weights used.

Gradient

$$\nabla E[\vec{w}] \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots \frac{\partial E}{\partial w_n}\right]$$

Training rule:

$$\Delta\vec{w} = -\eta\nabla E[\vec{w}]$$

i.e.,

$$\Delta w_i = -\eta\frac{\partial E}{\partial w_i}$$

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{2}\sum_d (t_d - o_d)^2$$

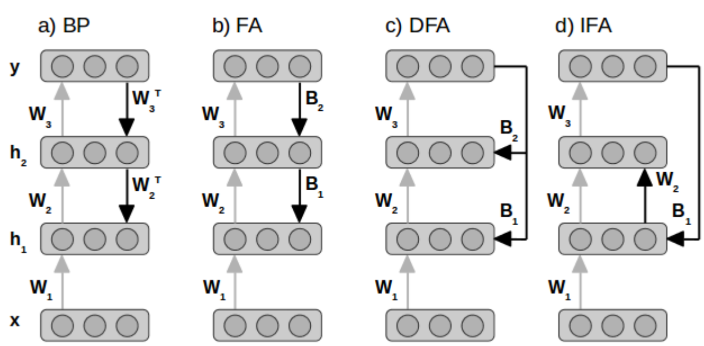
$$
\begin{aligned}
&= \frac{1}{2}\sum_d \frac{\partial}{\partial w_i}(t_d - o_d)^2 \\
&= \frac{1}{2}\sum_d 2(t_d - o_d)\frac{\partial}{\partial w_i}(t_d - o_d) \\
&= \sum_d (t_d - o_d)\frac{\partial}{\partial w_i}(t_d - \vec{w}\cdot\vec{x_d}) \\
\frac{\partial E}{\partial w_i} &= \sum_d (t_d - o_d)(-x_{i,d})
\end{aligned}
$$

2.1 Direct Feedback Alignment

Direct Feedback Alignment avoids weight transport by instead using fixed, random-weight matrices on its forward pass. The equation here describes the model:

$$
\begin{gathered}
a_1 = W_1 x + b_1,\ h_1 = f(a_1) \\
a_2 = W_2 h_1 + b_2,\ h_2 = f(a_2) \\
a_y = W_3 h_2 + b_3,\ \hat{y} = f_y(a_y)
\end{gathered}
$$

where W are the forward model weights, f is some non-linear activation function, and a are net inputs before applying the non-linearity. The update schemes for backpropagation, feedback alignment, direct feedback alignment and indirect feedback alignment are compared in the image below:



Surprisingly, DFA works by updating its weights from the feedback received by random, fixed matrices B. A high level interpretation is that the model learns how to learn from the fixed feedback it receives although this is unclear. The update equations are compared here:

$$
\delta a_2 = (B_2 e) \odot f'(a_2),\ \delta a_1 = (B_1 e) \odot f'(a_1) \tag{8}
$$

where B_i is a fixed random weight matrix with appropriate dimension. If all hidden layers have the same number of neurons, B_i can be chosen identical for all hidden layers. For IFA, the hidden layer update directions are calculated as

$$
\delta a_2 = (W_2 \delta a_1) \odot f'(a_2),\ \delta a_1 = (B_1 e) \odot f'(a_1) \tag{9}
$$

where B_1 is a fixed random weight matrix with appropriate dimension. Ignoring the learning rate, the weight updates for all methods are calculated as

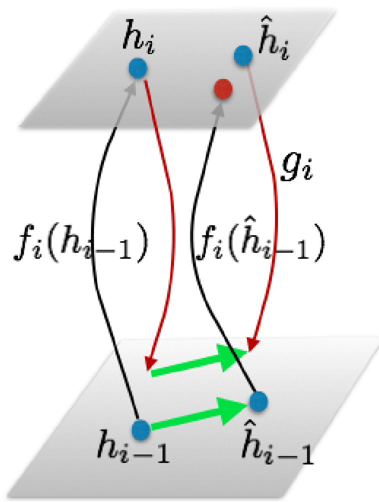
$$
\delta W_1 = -\delta a_1 x^T,\ \delta W_2 = -\delta a_2 h_1^T,\ \delta W_3 = -e h_2^T \tag{10}
$$

2.2 Target Propagation

Target propagation was inspired by the need to find a method which avoided the vanishing gradient problem[6]. Lee, et al. point out that this becomes a greater problem as the number of layers and non-linear functions increase. In the extreme case, there can be a discrete relation between parameters and the cost function which BP would not be able to model accurately. After the error is calculated, updates based on adjusting the target values in the direction of the gradient are backpropagated. These target values are associated with the activation values at each layer. Instead of using symmetric weight, TP uses auto-encoders at each layer which are able to nudge the target values in a direction which lowers loss and thus approximates the gradient used in BP.

2.2.1 Difference Target Propagation

Lee, et al.[7] claim that Bengio's original 2014 formulation of TP is computationally infeasible and update it as shown here:



Difference target propagation (DTP) uses a gradient calculation at the penultimate layer in order to have suitable values when backpropagating through the network. Otherwise, it learns how to update the weights at each layer by computing autoencoder relations between layers for given target values as shown here:

Algorithm 1 Training deep neural networks via difference target propagation

Compute unit values for all layers:
for $i = 1$ to M **do**
  $\mathbf{h}_i \leftarrow f_i(\mathbf{h}_{i-1})$
end for
Making the first target: $\hat{\mathbf{h}}_{M-1} \leftarrow \mathbf{h}_{M-1} - \hat{\eta}\frac{\partial L}{\partial \mathbf{h}_{M-1}}$, ($L$ is the global loss)
Compute targets for lower layers:
for $i = M-1$ to 2 **do**
  $\hat{\mathbf{h}}_{i-1} \leftarrow \mathbf{h}_{i-1} - g_i(\mathbf{h}_i) + g_i(\hat{\mathbf{h}}_i)$
end for
Training feedback (inverse) mapping:
for $i = M-1$ to 2 **do**
  Update parameters for g_i using SGD with following a layer-local loss L_i^{inv}
  $L_i^{inv} = ||g_i(f_i(\mathbf{h}_{i-1}+\epsilon)) - (\mathbf{h}_{i-1}+\epsilon)||_2^2, \quad \epsilon \sim N(0, \sigma)$
end for
Training feedforward mapping:
for $i = 1$ to M **do**
  Update parameters for f_i using SGD with following a layer-local loss L_i
  $L_i = ||f_i(\mathbf{h}_{i-1}) - \hat{\mathbf{h}}_i||_2^2$ if $i < M$, $L_i = L$ (the global loss) if $i = M$.
end for

This is the version used in the remaining proceedings. Thus, although the gradient is computed with respect to the global loss on the first step of the backwards pass, the rest of the network is updated using the autoencoders and avoid having a vanishing gradient due to the chain rule.

3 Differential Privacy

The need to protect people's information has led to the study of how this can be achieved while still allowing the data to have utility. This is an ideal where individuals cannot be identified through sharing their data, but the data is not so noisy or perturbated as to become useless. Differential privacy presents a recent defintion for the amount of information that can be revealed about any given individual. This is measured using the sensitivity of some desired query function (such as calculating the mean or median of a dataset), and then achieving privacy by adding Laplacian noise to the data or function result. Specifically, this definition assumes there exists some adversary who has access to a full neighboring dataset; this is a dataset identical to the one in question but differing by at most one instance x in the input set X. The parameter epsilon is set to allow small or large amounts of privacy leakage given some query function K:

> A randomized algorithm K gives *ε-differential privacy* if for all data sets D and D′ differing on at most one row, and any S ⊆ Range(K),

$$Pr[K(D) \in S] \leq exp(\varepsilon) \times Pr[K(D') \in S]$$

One intuition is that the smaller the dataset, the more likely that the absence or presence of any one individual will influence any statistical measures as opposed to the smaller effect one person would have on a large group. Noise is added most commonly via the Laplace mechanism. Noise is injected proportional to the sensitivity of the function K. It has been proven that (ε,0) differential privacy can be maintained by setting epsilon to the sensitivity of K/λ, and then adding noise drawn from a Laplace(λ) distribution with mean 0.

4 Methods Used

In order to compare the efficacy of these methods, models were trained using standard backpropagation, direct feedback alignment and difference target propagation on the well known MNIST, Fashion-MNIST, CIFAR10 and CIFAR100 datasets. As shown in order of increasing task difficulty, MNIST is a set of 28x28 images showing handwritten digits 0 through 9, and greyscale values for each pixel represented by a value between 0 and 255. Fashion MNIST has the same data attributes as MNIST, but displays images of 10 different articles of clothing from different angles which classification more difficult. CIFAR(Canandian Insitute for Advanced Research)-10 has more complex classes such as airplane, bird, dog, horse, truck, etc. and images are slightly larger at 32x32. The CIFAR sets have three color features RGB(red,green,blue), as opposed to the one greyscale in the MNIST sets. CIFAR-100 is magnitudes more difficult than MNIST with the data instances belonging to 100 possible classes, which are grouped into 20 coarser supergroups. A number of three layer architectures are tested.

Further, the models are tested after being trained on noisy data or by injecting noise to their update or gradient calculations. This experiment allows for insight into the learning abilities of these training algorithms in a private environment, and may show greater behavioural realism as well.

5 Experimental Results

The models used consisted of three layers, with 800 tanh units in the hidden layer and the sigmoid function at the output layer. Unless otherwise noted, learning rate is set to 5*1e-4, and batch size is

Table 1: Test Errors

Error		
Dataset	Training	%
MNIST	BP	7.03%
MNIST	DFA	5.93%
F-MNIST	BP	15.85%
F-MNIST	DFA	14.74%
CIFAR-10	BP	70.35%
CIFAR-10	DFA	63.0%
CIFAR-100	BP	96.58%
CIFAR-100	DFA	87.58%

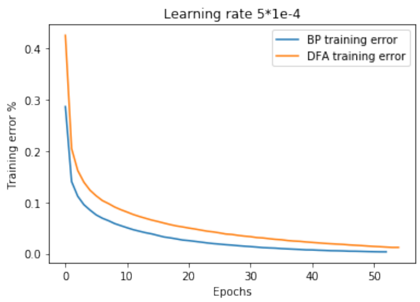


Figure 1: MNIST

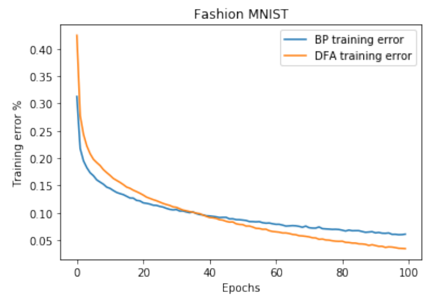


Figure 2: FMNIST

200. The algorithms were set to run for 100 epochs unless stopped when the change in the objective function becomes less than 1e-4.

The results show that DFA is able to perform just as well as BP on relatively simple tasks such as MNIST with a 3-layer network, even surpassing it given enough training time as shown with Fashion-MNIST. However, the model takes around twice as many epochs to converge to a difference in the objective function of 1e-4 if trained using DFA. The model was trained using DFA was able to pass the results of the BP model early on, hinting at the performance benefits DFA might have on low-powered machines if it is only feasible to use one hidden layer.

In order to test the robusticity and peformance of these algorithms in a private setting, a second set of training operations were developed where noise is added during every step of the backwards update pass. This noise is drawn from a Normal distribution with mean of zero and a σ value between .1 and 1. Expressed as such N(0,σ^2). The results shown are from adding noise with σ values equal to .1 and then .4 to the MNIST dataset, and then .2 and .4 to the CIFAR-10 set for both methods. While σ = .4 still allowed for the models to learn accurate features on the MNIST dataset, this amount of noise did not allow for either model to have less than 90% on the CIFAR-10 set. Noise was not added to the CIFAR-100 set as neither model was capable of strong results with even the clean dataset.

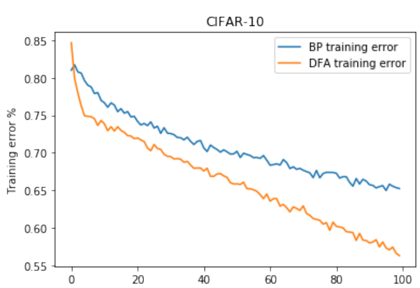


Figure 3: CIFAR10

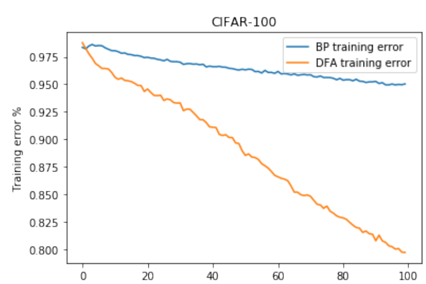


Figure 4: CIFAR100

Table 2: Noisy Errors

Error with Noise			
Dataset	σ	Training	%
MNIST	.1	BP	5.67%
MNIST	.1	DFA	4.95%
MNIST	.4	BP	12.83%
MNIST	.4	DFA	14.22%
CIFAR-10	.1	BP	75.51%
CIFAR-10	.1	DFA	70.78%
CIFAR-10	.4	BP	90.0%
CIFAR-10	.4	DFA	90.0%

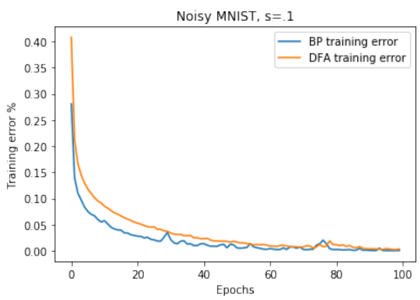


Figure 5: Noisy MNIST, s=.1

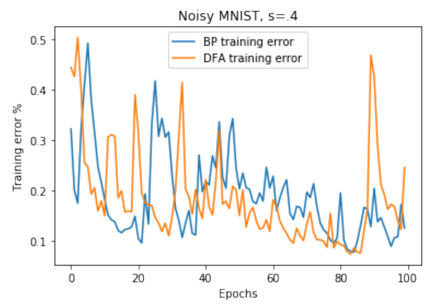


Figure 6: Noisy MNIST, s=.4

Interestingly, the DFA training scheme was able to train the model noticeably quicker than BP for CIFAR-10 when sigma = .1. Not only that, but DFA was able to approximate the same accuracy as given by training without noisy gradients. This result shows that there are settings of light to moderate noise where DFA can train a model more quickly and accurately than BP.

Broader Impact and Further Research

DFA was able to outperform backpropagation for certain tasks given a simple network architecture which shows biologically plausible machine learning is still worth further consideration. The separation of the backwards error propagation from the forward network may be possible as shown by the results of the DFA. Concerns such as differential privacy and the need to learn from noisy data may necessitate the use of more robust training methods in the future. Other papers have shown that while DFA and DTP perform well up to CIFAR level tasks, they fall behind BP gradient descent on jobs such as ImageNet which has more than 20,000 categories. This paper has shown that there are settings where DFA is able to train a given model more effectively than BP for datasets perturbed to light to moderate amounts of noise.

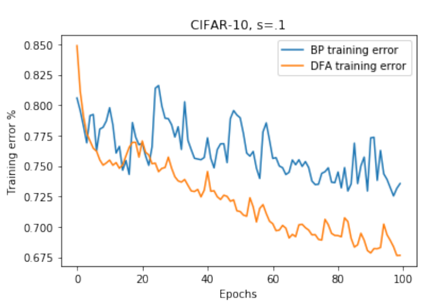


Figure 7: Noisy CIFAR10, s=.1

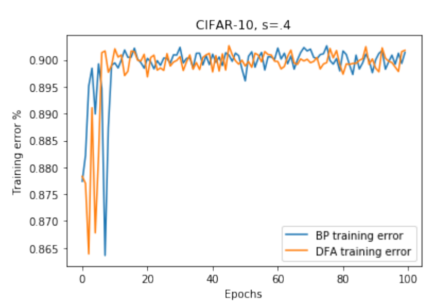


Figure 8: Noisy CIFAR10, s=.4

While tests regarding the biological plausibility of training algorithms were shown here, there is still the question of find more plausible network architectures. There may be little sense in addressing concerns of plausibility if the model architectures DFA/TP are training have nothing to do with biology. One example of a step in this direction are spiking neural networks. According to Bartunov, et al.[8], "the way in which forward and backward pathways in the brain interact is not well-characterized, but we're not aware of existing evidence that straightforwardly supports distinct phases". However, all of these training methods require a forward then backwards pass with the error. Along with computational biology, this is a field where meetings between computer scientists and neurologists can lead to better syncretic technologies. Work on the question of biologically plausible will hopefully improve both areas of science. The brain can be more accurately modelled with more plausible algorithms, but more plausible algorithms would also have the benefit to computer sciences of being more robust to noise or having greater learning plasticity.

References

[1] S.-i. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, 1993.

[2] A. Nøkland, "Direct feedback alignment provides learning in deep neural networks," in *Advances in neural information processing systems*, 2016, pp. 1037–1045.

[3] Y. Bengio, "How auto-encoders could provide credit assignment in deep networks via target propagation," *arXiv preprint arXiv:1407.7906*, 2014.

[4] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[5] J. C. Whittington and R. Bogacz, "Theories of error back-propagation in the brain," *Trends in cognitive sciences*, 2019.

[6] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[7] D.-H. Lee, S. Zhang, A. Fischer, and Y. Bengio, "Difference target propagation," in *Joint european conference on machine learning and knowledge discovery in databases*. Springer, 2015, pp. 498–515.

[8] S. Bartunov, A. Santoro, B. Richards, L. Marris, G. E. Hinton, and T. Lillicrap, "Assessing the scalability of biologically-motivated deep learning algorithms and architectures," in *Advances in Neural Information Processing Systems*, 2018, pp. 9368–9378.