EFFICIENT ON-DEVICE MACHINE LEARNING WITH A BIOLOGICALLY-PLAUSIBLE FORWARD-ONLY ALGORITHM

Baichuan Huang [1] **Amir Aminifar** [1]

ABSTRACT

The training of the state-of-the-art Deep Neural Networks (DNNs) consumes massive amounts of energy, while the human brain learns new tasks with remarkable efficiency. Currently, the training of DNNs relies almost exclusively on Backpropagation (BP). However, BP faces criticism due to its biologically implausible nature, underscoring the significant disparity in performance and energy efficiency between DNNs and the human brain. Forward-only algorithms are proposed to be the biologically plausible alternatives to BP, to better mimic the learning process of the human brain and enhance energy efficiency. In this paper, we propose a biologically-plausible forward-only algorithm (Bio-FO), not only targeting the biological-implausibility issues associated with BP, but also outperforming the state-of-the-art forward-only algorithms. We extensively evaluate our proposed Bio-FO against other forward-only algorithms and demonstrate its performance across diverse datasets, including two real-world medical applications on wearable devices with limited resources and relatively large-scale datasets such as mini-ImageNet. At the same time, we implement our proposed on-device learning algorithm on the NVIDIA Jetson Nano and demonstrate its efficiency compared to other state-of-the-art forward-only algorithms. The code is available at https://github.com/whubaichuan/Bio-FO.

1 INTRODUCTION

The state-of-the-art Deep Neural Networks (DNNs) consume massive amounts of energy and pose a threat to the environment (Savazzi et al., 2022). A prime example is GPT-3, a Large Language Model (LLM), that consumes over 1000 megawatt-hour for training alone, which is equivalent to a small town's power consumption for a day (Patterson et al., 2021). In contrast, the human brain learns more efficiently, consuming only around 20 watts (Hsu, 2014; Balasubramanian, 2021; Madhavan, 2024). This is particularly relevant in the context of Internet of Things (IoT) and mobile devices, which are generally extremely limited in terms of resources, namely, computing power, memory storage, and battery/energy budget (Shi et al., 2016; Sopic et al., 2018; Sabry et al., 2022; Huang et al., 2024). Nevertheless, today, DNNs are trained almost exclusively based on the Back-propagation (BP) algorithm (Rumelhart et al., 1986), which is known to lack biological plausibility (Crick, 1989; Lillicrap et al., 2016). As a result, adopting a more biologically plausible approach to training DNNs offers the potential to better mimic the learning processes of the human brain and, in turn, enhance energy efficiency (Hinton, 2022).

The biological implausibility of the BP algorithm stems from several of its inherent requirements/assumptions: *weight transport* (Grossberg, 1987; Burbank & Kreiman, 2012), where BP requires symmetric weights between the forward and backward passes. The error signals are back-propagated by multiplying the *transpose* of the exact same forward weights, which is not known to exist in the biological brain (Lillicrap et al., 2016; Woo et al., 2021); *non-locality* (Whittington & Bogacz, 2019), where the weights update in BP relies on all the nodes from the top to the bottom layers, which means that the error signals span long distances from the output layer in the network, while biological synapses learn from the activations of the neurons they are connected to. Therefore, BP violates the inherent *locality* of biological synaptic plasticity (Hebb, 2005); *update locking* (Jaderberg et al., 2017), where the weights update needs to *wait* for all the dependent layers in the forward pass to complete, which is not consistent with synaptic plasticity (Wang et al., 2016; Tang et al., 2022); and *frozen activities* (Liao et al., 2016), where the intermediate activations are *frozen* to be used for weights update (Cai et al., 2020b),

This research has been partially supported by the Swedish Wallenberg AI, Autonomous Systems and Software Program (WASP), Swedish Research Council, Swedish Foundation for Strategic Research, and ELLIIT Strategic Research Environment. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) partially funded by the Swedish Research Council (VR) through grant agreement no. 2022-06725. [1]Department of Electrical and Information Technology, Lund University, Sweden. Correspondence to: Baichuan Huang <baichuan.huang@eit.lth.se>.

Proceedings of the 8^{th} MLSys Conference, Santa Clara, CA, USA, 2025.

while in the biological brain, the activities are influenced by feedback connections (Lillicrap et al., 2020) and exhibit dynamic adaptation over time (Koban et al., 2019; Blanken et al., 2021; Barabási et al., 2023).

The inherent biological implausibility of BP has led to major criticisms, prompting the exploration of biologically-plausible forward-only algorithms, focusing on training DNNs without resorting to the biologically-implausible back-propagation scheme, to bridge the existing performance–efficiency gap between the DNNs and the cortex (Srinivasan et al., 2023). Biologically plausible algorithms lead to a path towards neurally-inspired deep learning (Miconi, 2017; Richards et al., 2019; Tang et al., 2022), and offer the potential to relieve several challenges in the deep neural networks domain (Gupta et al., 2022), including computational intensity (Zhang et al., 2019; Singhal et al., 2023; Aminifar et al., 2024; Huang et al., 2023), energy/memory intensity (Hinton, 2022; Shervani-Tabar & Rosenbaum, 2023; Huang & Aminifar, 2025b), the demand for massive labeled datasets (Feldmann et al., 2019; Li et al., 2020; Vishwakarma et al., 2024), and vulnerability to perturbations (Büchel et al., 2021; Ma et al., 2023). To date, a range of forward-only techniques, including PEPITA (Dellaferrera et al., 2022b) and the Forward-Forward (FF) algorithm (Hinton, 2022), have emerged. However, these forward-only algorithms only partially address the aforementioned biological implausibility issues, and lack the capacity to resolve these issues associated with BP, i.e., *weight transport*, *non-locality*, *update locking*, and *frozen activities*, as shown in Fig. 1.

In this paper, we propose an efficient on-device learning algorithm, based on the biologically-plausible forward-only algorithm, called Bio-FO. Bio-FO targets the previously-mentioned biological implausibility issues and can be flexibly extended to common networks and relatively large-scale datasets. Firstly, Bio-FO eliminates the symmetric weights in the backward pass by exploiting an auxiliary classifier with the fixed random matrix instead, hence avoiding the issues of *weight transport*. Secondly, the training of Bio-FO is performed locally, without the need for non-local information/global error, hence mitigating the issue of *non-locality*. Thirdly, the weights are updated as soon as the input to the layer (i.e., the activation of the previous layer) is available, addressing the issue of *update locking*. Finally, the activations of the intermediate layers do not need to be frozen and, therefore, the training of each layer could be performed simultaneously, avoiding the issue of *frozen activities*.

A comprehensive evaluation of our proposed Bio-FO is conducted in the context of three widely-used datasets by the forward-only algorithms (Frenkel et al., 2021; Hinton, 2022; Dellaferrera et al., 2022b), including MNIST (LeCun, 1998), CIFAR-10 (Krizhevsky et al., 2010), and CIFAR-100 (Krizhevsky et al., 2010). In addition, to demonstrate the relevance of the proposed forward-only scheme, we also consider two real-world medical applications on wearable devices, namely, seizure detection (Shoeb, 2009) and arrhythmia classification (Mark et al., 1982), for real-time and long-term monitoring in ambulatory settings. Mobile and wearable devices are often extremely limited in terms of computing resources, memory storage, energy, and battery life, and present an excellent case study and motivation for forward-only algorithms because biologically-plausible forward-only algorithms offer more resource-efficient neural network operations (Lin et al., 2016; Hinton, 2022). The results illustrate the relevance of our proposed forward-only algorithm. Our main contributions are summarized below:

- We propose a resource-efficient on-device learning algorithm, namely Bio-FO, based only on forward passes and without the need for BP, targeting the biological implausibility issues of *weight transport*, *non-locality*, *update locking*, and *frozen activities*, as shown in Fig. 1. Our proposed Bio-FO can directly capture structure/patterns and sparsity and can be extended to Locally Connected (LC) Network and Convolutional Neural Network (CNN).
- We extensively evaluate our proposed on-device learning algorithm in terms of prediction performance. Our proposed Bio-FO outperforms the state-of-the-art forward-only algorithms, namely DRTP, PEPITA, and FF, across several datasets, including CIFAR-10, CIFAR-100, CHB-MIT, and MIT-BIH. Overall, Bio-FO demonstrates the closest classification performance to BP. Our evaluation also shows that Bio-FO outperforms other forward-only algorithms on relatively large-scale datasets such as mini-ImageNet.
- We implement our proposed on-device learning algorithm on the NVIDIA Jetson Nano and evaluate it in terms of resource overheads, including computation requirements and energy consumption. Our proposed Bio-FO exhibits faster convergence during the training process, compared to DRTP, PEPITA, and FF. At the same time, our evaluation demonstrates that, overall, Bio-FO is considerably more efficient in terms of resources when evaluated on NVIDIA Jetson Nano.

The remainder of this paper is organized as follows. In Section 2, we review the literature on biologically-plausible forward-only algorithms. Next, in Section 3, we propose our resource-efficient on-device learning algorithm in detail. Then, in Section 4, we describe the experimental setup including datasets, implementation details, and implementation platform. In addition, we experimentally evaluate and compare our proposed on-device learning algorithm against several state-of-the-art algorithms in terms of classification

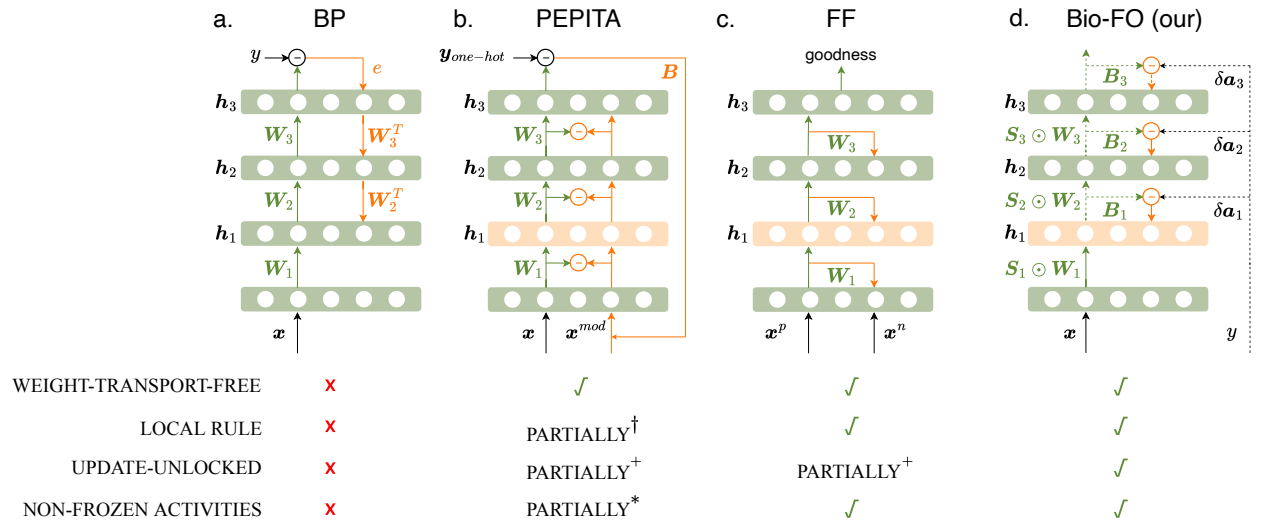


Figure 1: An overview of various training algorithms is presented: a) BP (Rumelhart et al., 1986); b) PEPITA (Dellaferrera et al., 2022b), where $*$ means PEPITA has to store the activations of hidden layers in the first forward pass and † means only space locality, not time locality (Srinivasan et al., 2023). And, + means both PEPITA and FF only partially address the issue of update locking (Srinivasan et al., 2023); c) FF (Hinton, 2022); d) Bio-FO (our), which addresses the *weight transport* issue in the standard pass, *non-locality* issue, the *update locking* issue, and the *frozen activities* issue. The forward paths are illustrated by green arrows, error paths by orange arrows, and input data/label paths by black arrows. Detailed discussion about the biological implausibility of PEPITA and FF can be found in Section 2.

performance and energy consumption. Finally, Section 5 serves as the conclusion of this work.

2 Background and Related Work

As discussed earlier, BP (Rumelhart et al., 1986) has been criticized for its biologically implausible nature, i.e., because of the issues of *weight transport* (Burbank & Kreiman, 2012), *non-locality* (Whittington & Bogacz, 2019), *frozen activities* (Liao et al., 2016), *update locking* (Jaderberg et al., 2017) issues. To solve the issue of *weight transport* (Burbank & Kreiman, 2012), FA (Lillicrap et al., 2016) and DFA (Nøkland, 2016) are proposed to exploit a fixed random matrix, without incurring the need for symmetric weights. However, FA and DFA are still suffering from the issues of *non-locality*, *update locking*, and *frozen activities*. Furthermore, DRTP (Frenkel et al., 2021) introduces the use of a target proxy for the forward pass, not only addressing the issue of *weight transport*, but also eliminating the issue of *update locking*. Despite its great potential, DRTP shows significant performance degradations when compared to BP.

The state-of-the-art forward-only algorithms are proposed to address the challenges inherent in BP, FA, DFA, and DRTP. For instance, PEPITA (Dellaferrera et al., 2022b) adopts the approach of replacing the backward pass with a modulated forward pass. PEPITA addresses the issues of *weight transport*. However, PEPITA still has the issue of *non-locality* because it is only local in space, not in time (Srinivasan et al., 2023). Moreover, PEPITA only partially addresses the *update locking* issue because of the direct global error feedback after the execution of the first forward pass. At the same time, while PEPITA avoids the backward pass, it still needs to store the activations of hidden layers in the standard forward pass for calculating error in the modulated forward pass, hence only partially addressing the *frozen activities* issue. To address the issue of time locality and the constraint of storing the activations in the standard forward pass, PEPITA-TL (Srinivasan et al., 2023) is proposed, which exhibits a major degradation in accuracy compared to the original PEPITA (Dellaferrera et al., 2022b). Besides, PEPITA is exclusively applied to shallow networks, restricted to no more than three hidden layers (Pau & Aymone, 2023); otherwise, PEPITA experiences a decrease in accuracy, e.g., transitioning from three hidden layers to five hidden layers (Srinivasan et al., 2023).

Similarly, the FF algorithm (Hinton, 2022) substitutes the forward and backward passes of backpropagation with two forward passes. While FF effectively resolves the issues of *weight transport*, *non-locality*, and *frozen activities*, it only partially addresses the *update locking* issue (Srinivasan et al., 2023). In addition, FF alters the input data to embed the labels and has the issue of data pollution, because FF generates new labels within the rows of input data. In datasets like CIFAR-100 (Krizhevsky, 2009), the initial four rows of data need to be replaced by the synthetic labels. This

embedding of labels within the input data may significantly impact the performance.

Recently, several extensions (Papachristodoulou et al., 2024; Lorberbom et al., 2024; Niu et al., 2024; Huang & Aminifar, 2025a) have been proposed based on FF and other forward-only algorithms. For instance, in Papachristodoulou et al. (2024), the authors focus on convolutional channel-wise competitive learning for FF. Similarly, in Mostafa et al. (2018); Belilovsky et al. (2019), the authors propose layer-wise training mainly with CNN. Despite these excellent initiatives, the fundamental concept underpinning weight sharing in CNN is not biologically plausible (Bartunov et al., 2018; Tang et al., 2022). In Lorberbom et al. (2024), the authors propose layer collaboration in FF, which has the biological implausibility issues of *non-locality*, *update locking*, and *frozen activities* because layer-collaborative FF requires the sum of goodness values from different layers. In summary, these state-of-the-art forward-only algorithms are only partially biologically plausible.

3 METHOD

3.1 Bio-FO Forward Pass

Let us consider a DNN with L layers. The input $\boldsymbol{x} \in \mathbb{R}^{D_x \times 1}$ and the target $\boldsymbol{y} \in \mathbb{R}^{N_c \times 1}$ are considered for training the DNN, where D_x is the size of input $\boldsymbol{x}$ and N_c is the number of classes. The activations of the hidden layer l of the DNN are denoted as $\boldsymbol{h}_l$, where $\boldsymbol{h}_0 = \boldsymbol{x}$. For the hidden layer l, the activations $\boldsymbol{h}_l \in \mathbb{R}^{D_l \times 1}$ based on $\boldsymbol{h}_{l-1} \in \mathbb{R}^{D_{l-1} \times 1}$, where D_l is the size of output of layer l and D_{l-1} is the size of input to layer l, are calculated as follows:

$$\begin{aligned} \boldsymbol{z}_l &= (\boldsymbol{S}_l \odot \boldsymbol{W}_l)\boldsymbol{h}_{l-1} + \boldsymbol{b}_l \\ \boldsymbol{h}_l &= \sigma_l(\boldsymbol{z}_l) \\ L \geq l &\geq 1, \end{aligned} \tag{1}$$

where $\boldsymbol{z}_l \in \mathbb{R}^{D_l \times 1}$ is logits for layer l and $\odot$ is the Hadamard Product. σ_l is the activation function for the hidden layer l. $\boldsymbol{W}_l \in \mathbb{R}^{D_l \times D_{l-1}}$ and $\boldsymbol{b}_l \in \mathbb{R}^{D_l \times 1}$ are the DNN weights and biases between hidden layers $l-1$ and l, respectively. The sparsity mask $\boldsymbol{S}_l \in \mathbb{R}^{D_l \times D_{l-1}}$ is introduced to allow extensions to common networks, where each element of $\boldsymbol{S}_l$ has a binary value.

3.2 Bio-FO Training Scheme

Here, we introduce our proposed training scheme based on the biologically-plausible forward pass discussed above. Our proposed Bio-FO forward-only training is outlined in Algorithm 1. Let us define the function composition $f_l : \boldsymbol{h}_{l-1} \rightarrow \boldsymbol{h}_l$. Given this function composition, the activations $\boldsymbol{h}_l$ are denoted as follows:

$$f_l(\boldsymbol{h}_{l-1}) : \boldsymbol{h}_l = \sigma_l((\boldsymbol{S}_l \odot \boldsymbol{W}_l)\boldsymbol{h}_{l-1} + \boldsymbol{b}_l).$$

Algorithm 1 Implementation of Bio-FO

Given: Input $(\boldsymbol{x})$ and Target (y), Learning Rate η, and the function composition $f_l(\boldsymbol{h}_{l-1}) : \boldsymbol{h}_l = \sigma_l((\boldsymbol{S}_l \odot \boldsymbol{W}_l)\boldsymbol{h}_{l-1} + \boldsymbol{b}_l)$
$\boldsymbol{h}_0 = \boldsymbol{x}$
for $l = 1, ..., L$ **do**
  # Forward Pass
  $\boldsymbol{h}_l = f_l(\boldsymbol{h}_{l-1})$
  $\boldsymbol{a}_l = \boldsymbol{B}_l \boldsymbol{h}_l$
  # Gradient and Update
  $\delta \boldsymbol{h}_l = \boldsymbol{B}_l^T \delta \boldsymbol{a}_l$
  $\delta \boldsymbol{W}_l = \delta \boldsymbol{h}_l \odot \sigma_l'(\boldsymbol{z}_l) \otimes \boldsymbol{h}_{l-1}^T \odot \boldsymbol{S}_l$
  $\delta \boldsymbol{b}_l = \delta \boldsymbol{h}_l \odot \sigma_l'(\boldsymbol{z}_l)$
  $\boldsymbol{W}_l = \boldsymbol{W}_l - \eta \delta \boldsymbol{W}_l$
  $\boldsymbol{b}_l = \boldsymbol{b}_l - \eta \delta \boldsymbol{b}_l$
end for

Our proposed training scheme is local. This essentially means that, for training the hidden layer l, only the $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ are trainable. That is, the local training of layer l does not affect other layers, i.e., the hidden layers spanning from 1 to $l-1$ are not trainable, which means $\boldsymbol{W}_1$ to $\boldsymbol{W}_{l-1}$, and $\boldsymbol{b}_1$ to $\boldsymbol{b}_{l-1}$ are not affected. To make the distinction, we denote the constant weights of hidden layer $l-1$ as $\overline{\boldsymbol{W}}_{l-1}$ and the constant biases of hidden layer $l-1$ as $\overline{\boldsymbol{b}}_{l-1}$. The function composition with the constant formulation $\overline{f}_{l-1} : \boldsymbol{h}_{l-2} \rightarrow \boldsymbol{h}_{l-1}$ is introduced as follows:

$$\overline{f}_{l-1}(\boldsymbol{h}_{l-2}) : \boldsymbol{h}_{l-1} = \sigma_{l-1}((\boldsymbol{S}_{l-1} \odot \overline{\boldsymbol{W}}_{l-1})\boldsymbol{h}_{l-2} + \overline{\boldsymbol{b}}_{l-1}).$$

Thus, for the hidden layer l, the function compositions $\overline{f}_1$ to $\overline{f}_{l-1}$ are within the constant formulation. The activation $\boldsymbol{h}_l$ is calculated as follows:

$$\boldsymbol{h}_l = f_l \circ \overline{f}_{l-1} \circ \overline{f}_{l-2} ... \circ \overline{f}_1(\boldsymbol{h}_0) = f_l(\boldsymbol{h}_{l-1}),$$

where $\boldsymbol{h}_l$ is derived directly from $\boldsymbol{h}_{l-1}$, not requiring to store other intermediate activations.

For each $\boldsymbol{h}_l$, an auxiliary classifier with the fixed random matrix $\boldsymbol{B}_l \in \mathbb{R}^{N_c \times D_l}$ is employed to project the $\boldsymbol{h}_l \in \mathbb{R}^{D_l \times 1}$ to the output vector $\boldsymbol{a}_l \in \mathbb{R}^{N_c \times 1}$, where $\boldsymbol{B}_l$ is fixed, hence not trainable. Then, the projected vector $\boldsymbol{a}_l$ is used to calculate the error and the gradient, as follows:

$$\begin{aligned} \boldsymbol{a}_l &= \boldsymbol{B}_l \boldsymbol{h}_l, \\ \boldsymbol{p}_l &= \sigma_{\text{output}}(\boldsymbol{a}_l), \\ \delta \boldsymbol{h}_l &= \frac{\partial \text{loss}_l(\boldsymbol{p}_l, y)}{\partial \boldsymbol{a}_l} \frac{\partial \boldsymbol{a}_l}{\partial \boldsymbol{h}_l} = \boldsymbol{B}_l^T \delta \boldsymbol{a}_l = \boldsymbol{B}_l^T (\boldsymbol{p}_l - \boldsymbol{y}), \end{aligned} \tag{2}$$

where $\boldsymbol{a}_l$ is the logits for the auxiliary classifier of layer l. σ_{output} is the output activation function. $\boldsymbol{p}_l \in \mathbb{R}^{N_c \times 1}$ is the probability distribution vector of $\boldsymbol{a}_l$. loss_l is the cross-entropy loss and $\delta \boldsymbol{a}_l = \frac{\partial \text{loss}_l(\boldsymbol{p}_l, y)}{\partial \boldsymbol{a}_l}$ is the gradient of loss_l

with respect to $\boldsymbol{a}_l$. $\delta \boldsymbol{h}_l \in \mathbb{R}^{D_l \times 1}$ is the gradient for updating weights $\boldsymbol{W}_l$ and biases $\boldsymbol{b}_l$. Note that only $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ are trainable.

Next, the updating steps for weights $\boldsymbol{W}_l$ and biases $\boldsymbol{b}_l$ are performed based on the following gradient:

$$\delta \boldsymbol{W}_l = (\boldsymbol{B}_l^T \delta \boldsymbol{a}_l) \odot \sigma_l'(\boldsymbol{z}_l) \otimes \boldsymbol{h}_{l-1}^T \odot \boldsymbol{S}_l,$$
$$\delta \boldsymbol{b}_l = (\boldsymbol{B}_l^T \delta \boldsymbol{a}_l) \odot \sigma_l'(\boldsymbol{z}_l),$$

where $\otimes$ is the Kronecker Product. Note that for the first hidden layer l ($l = 1$), the updating steps for weights $\boldsymbol{W}_1$ and biases $\boldsymbol{b}_1$ are based on $\boldsymbol{h}_0 = \boldsymbol{x}$.

3.3 Biological Plausibility of Bio-FO

In our proposed Bio-FO, for a DNN with L layers, the layers from 1 to L are trained independently. For each layer, the training step does not suffer from the aforementioned biological implausibility issues: (1) To address the *weight transport* problem, our approach exploits an auxiliary classifier with the fixed random matrix ($\boldsymbol{B}_l$), incurring no backpropagation and no symmetric weights in the standard pass, where $\boldsymbol{B}_l$ is fixed and not trainable; (2) Regarding the *non-locality* problem, our proposed approach trains the DNN locally, without the need for the information from the top to the bottom layers. Specifically, for training the hidden layer l, there is no requirement for the information from layers that succeed the layer l, e.g., layer $l+1$. Thus, Bio-FO is *local* in both time and space; (3) With respect to the *update locking* problem, in Bio-FO, the weights update does not need to *wait* for all the dependent layers (succeeding layers) in the forward pass to complete execution. In other words, a new input is able to be fed into the DNN following the previous input; (4) Additionally, for the *frozen activities* problem, as illustrated in the *update locking* part, the activations before $\boldsymbol{h}_{l-1}$ and after $\boldsymbol{h}_{l+1}$ are not frozen for updating the weights between layer $l-1$ and layer l. Our approach does not store the intermediate activations in memory during training. Furthermore, Bio-FO allows us to incorporate the sparsity and locality of the connections via the mask $\boldsymbol{S}$, which are inherent to the cortex (Vision, 2000; Braitenberg & Schüz, 2013; Pulvermüller et al., 2021; Jeon & Kim, 2023; Sarfraz et al., 2023). In summary, Bio-FO targets the biological implausibility issues of *weight transport*, *non-locality*, *update locking*, and *frozen activities*.

4 Evaluation

4.1 Experimental Setup

4.1.1 Dataset

To evaluate our proposed Bio-FO, we consider the MNIST dataset of handwritten digits (LeCun, 1998), the CIFAR-10 dataset of object recognition (Krizhevsky, 2009), and the CIFAR-100 dataset of object recognition (Krizhevsky, 2009). Furthermore, we extend our evaluation to encompass real-world medical applications: epilepsy monitoring and seizure detection based on the CHB-MIT Scalp electroencephalogram (EEG) Dataset (Shoeb, 2009), where this dataset comprises EEG recordings from 22 patients with epilepsy and only two channels, i.e., T7F7 and T8F8 are exploited for real-time seizure monitoring using wearable devices (Sopic et al., 2018); and cardiac arrhythmia classification based on the MIT-BIH Arrhythmia Electrocardiogram (ECG) Dataset (Mark et al., 1982), which includes ECG recordings from 47 patients with cardiovascular problems (Arlington, 1998). Finally, we also consider relatively large-scale datasets such as mini-ImageNet (Vinyals et al., 2016), consisting of 60000 color images of size 84×84 with 100 classes. The mini-ImageNet is a subset of the larger ImageNet dataset (Deng et al., 2009).

4.1.2 Implementation Details

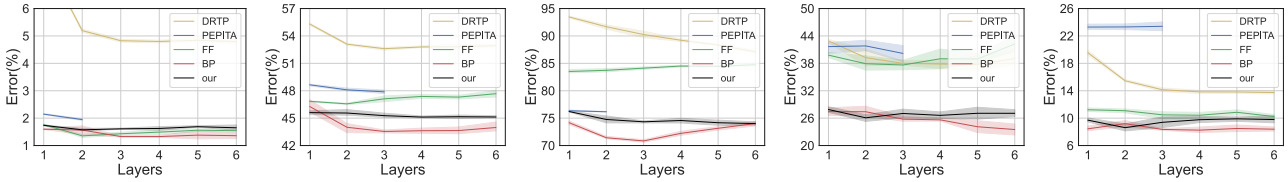
Our proposed Bio-FO is implemented using the PyTorch framework (Paszke et al., 2019). In Bio-FO, we consider the Softmax activation function for σ_{output} and exploit categorical cross-entropy. We consider Kaiming Uniform Initialization (He et al., 2015) for the fixed random matrix $\boldsymbol{B}$. Our experiments utilize balanced datasets, with classification performance assessed using error, i.e., the total number of incorrectly classified inputs divided by the total number of inputs. The mean error with the standard deviation is averaged with five independent runs with different random seeds. We compare and evaluate our proposed Bio-FO against three main state-of-the-art algorithms, i.e., DRTP (Frenkel et al., 2021) (the official implementation (Frenkel, 2021)), PEPITA (Dellaferrera et al., 2022b) (the official implementation (Dellaferrera, 2022)), and FF (Hinton, 2022) (the official implementation (Löwe, 2023)).

4.1.3 Implementation Platform

For classification performance evaluation, all algorithms undergo training on a server equipped with 2×16-core Intel (R) Xeon (R) Gold 6226R (Skylake) Central Processing Units (CPUs) and 1 NVIDIA Tesla T4 Graphics Processing Card (GPU). For resource-usage evaluation, we consider the NVIDIA Jetson Nano (NVIDIA, 2019), with powerful and efficient AI, computer vision, and high-performance computing at just 5 to 10 watts for deploying AI at the edge and embedded IoT applications. The NVIDIA Jetson Nano is equipped with a 128-core NVIDIA Maxwell™ architecture GPU, delivering AI performance up to 472 GFLOPS. The GPU operates at a maximum frequency of 921 MHz, while the Quad-core ARM® Cortex®-A57 MPCore processor has a max frequency of 1.43 GHz. This combination of GPU and CPU capabilities enables the Jetson Nano to handle demanding AI and machine learning tasks efficiently, making

Table 1: Error (%) comparison with the state-of-the-art DRTP (Frenkel et al., 2021), FF (Hinton, 2022), Bio-FO, BP (Rumelhart et al., 1986) with 4 hidden layers, and PEPITA (Dellaferrera et al., 2022b) with 2 or 3 hidden layers. We highlight the **best** and second best results among the forward-only algorithms.

Algorithms	MNIST (LeCun, 1998)	CIFAR-10 (Krizhevsky, 2009)	CIFAR-100 (Krizhevsky, 2009)	CHB-MIT (Shoeb, 2009)	MIT-BIH (Mark et al., 1982)
DRTP	4.79±0.05	52.79±0.12	89.22±0.26	37.85±1.05	13.84±0.27
PEPITA	1.95±0.04	47.85±0.22	76.16±0.04	40.17±2.02	23.40±0.76
FF	**1.46±0.07**	47.38±0.25	85.76±0.18	39.00±2.59	10.86±0.25
Ours (Bio-FO)	1.62±0.08	**45.12±0.12**	**74.57±0.51**	**26.61±0.78**	**9.77±0.49**
BP	1.33±0.04	43.62±0.33	72.22±0.43	25.63±0.40	8.25±0.46



(a) MNIST (b) CIFAR-10 (c) CIFAR-100 (d) CHB-MIT (e) MIT-BIH

Figure 2: Error (%) for DRTP, PEPITA, FF, Bio-FO, and BP, versus the number of layers. The solid line reports the mean over five independent runs, and the shaded area indicates the standard deviation.

it well-suited for edge computing applications.

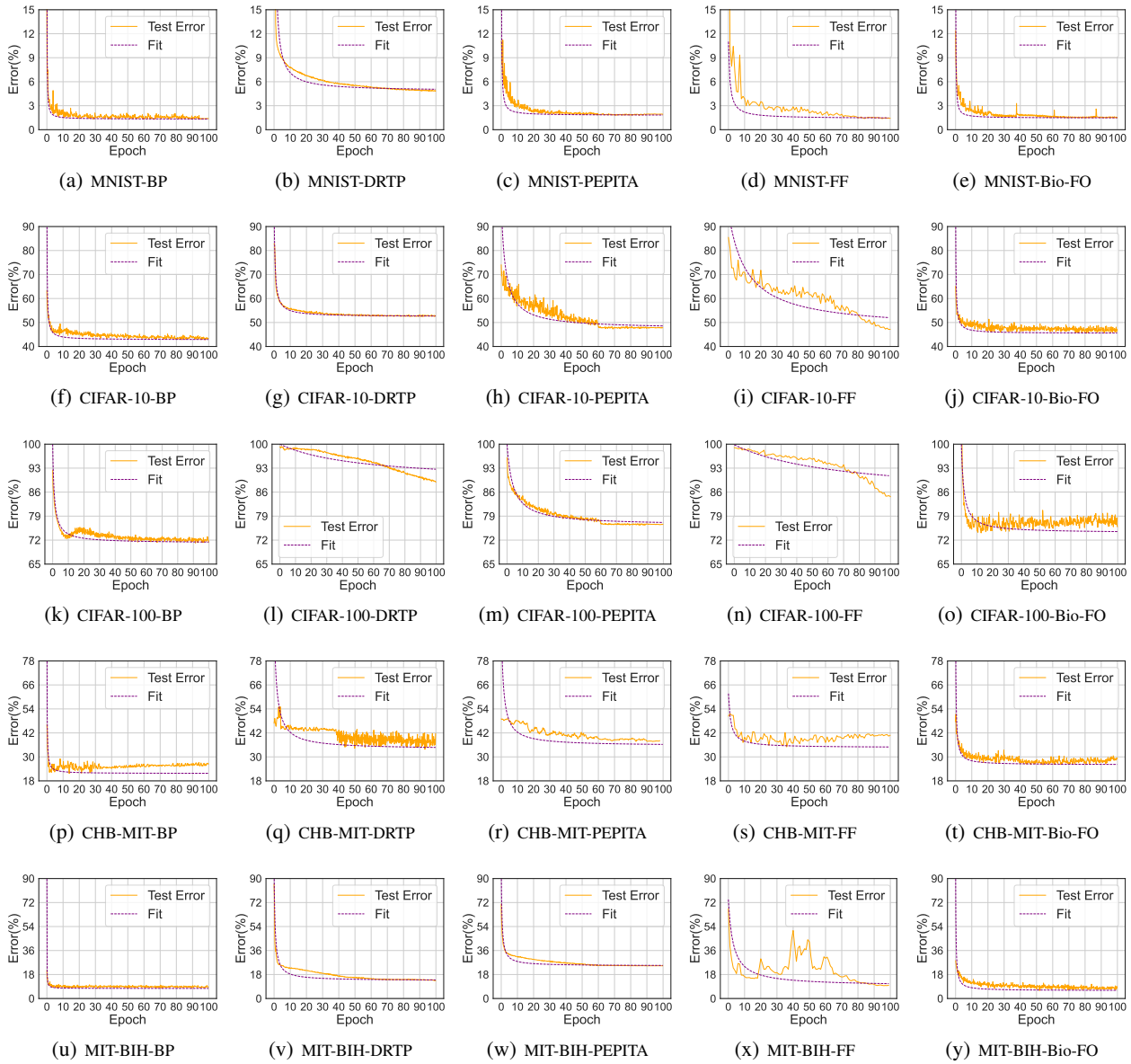
4.2 Experimental Results

In this section, we evaluate Bio-FO in terms of classification performance and convergence rate compared to BP, DRTP, PEPITA, and FF for Fully Connected (FC) networks with the all-one $\boldsymbol{S}$ matrix. Additionally, we also evaluate the energy consumption of Bio-FO compared to the state-of-the-art forward-only algorithm on the NVIDIA Jetson Nano.

4.2.1 Classification Performance

Table 1 presents the error (%) for DRTP, FF, Bio-FO, BP with 4 hidden layers, and PEPITA with 2 or 3 hidden layers. In the case of MNIST, BP achieves the best classification performance, with the lowest mean error of 1.33%. DRTP exhibits the worst classification performance with a mean error of 4.79%, and PEPITA (1.95%) performs better than DRTP. FF (1.46%) attains a similar error as BP and our proposed Bio-FO achieves a comparable error (1.62%) with FF. Additionally, for the other four datasets, namely, CIFAR-10, CIFAR-100, CHB-MIT, and MIT-BIH, our proposed Bio-FO achieves a lower error compared to DRTP, PEPITA, and FF. At the same time, Bio-FO demonstrates only a slightly higher error compared to BP. These results show that, overall, our proposed Bio-FO outperforms the state-of-the-art forward-only algorithms of DRTP, PEPITA, and FF in terms of classification performance.

Next, we vary the number of hidden layers from 1 to 6 for BP, DRTP, FF, and Bio-FO; we vary the number of hidden layers from 1 to 2 or 3 for PEPITA because the official implementation of PEPITA (Dellaferrera, 2022) supports only up to 3 hidden layers (Pau & Aymone, 2023). Fig. 2 illustrates the error (%) for DRTP, PEPITA, FF, Bio-FO, and BP across different layer settings. In the case of MNIST, DRTP consistently demonstrates the highest error across various number of layers, while BP, FF, and Bio-FO have comparable error values across different number of layers. In this case, BP, FF, and Bio-FO also attains a lower error than PEPITA. For CIFAR-10, Bio-FO consistently achieves a lower error compared to DRTP, PEPITA, and FF and performs close to BP. In the context of CIFAR-100 and CHB-MIT, Bio-FO outperforms DRTP and FF significantly, closely approaching the performance of BP. Bio-FO also outperforms PEPITA significantly in CHB-MIT. For MIT-BIH, Bio-FO achieves a slightly lower error than FF, and demonstrates a comparable error with BP, and attains a significantly lower error than PEPITA and DRTP, for different numbers of layers. Furthermore, we investigate a more energy- and memory-efficient Bio-FO by sparsing $\boldsymbol{B}$ in Equation (2) and achieve comparable performance. All the results collectively present that our proposed Bio-FO outperforms the state-of-the-art forward-only algorithms of DRTP, PEPITA, and FF, for different numbers of layers. At the same time, Bio-FO emerges as the forward-only algorithm with the potential to achieve comparable performance to BP.



(a) MNIST-BP (b) MNIST-DRTP (c) MNIST-PEPITA (d) MNIST-FF (e) MNIST-Bio-FO

(f) CIFAR-10-BP (g) CIFAR-10-DRTP (h) CIFAR-10-PEPITA (i) CIFAR-10-FF (j) CIFAR-10-Bio-FO

(k) CIFAR-100-BP (l) CIFAR-100-DRTP (m) CIFAR-100-PEPITA (n) CIFAR-100-FF (o) CIFAR-100-Bio-FO

(p) CHB-MIT-BP (q) CHB-MIT-DRTP (r) CHB-MIT-PEPITA (s) CHB-MIT-FF (t) CHB-MIT-Bio-FO

(u) MIT-BIH-BP (v) MIT-BIH-DRTP (w) MIT-BIH-PEPITA (x) MIT-BIH-FF (y) MIT-BIH-Bio-FO

Figure 3: Test error (%) and fitting the test error by the *plateau equation for learning curves*.

4.2.2 *Convergence Rate*

In this section, we evaluate the convergence rate of our proposed Bio-FO in comparison to BP, DRTP, PEPITA, and FF based on FC networks across five datasets. We also exploit the network with four hidden layers for BP, DRTP, FF, and Bio-FO, and the network with three hidden layers for PEPITA.

To quantify the convergence rate, we adopt the *plateau equation for learning curves* (Dellaferrera et al., 2022a):

$$1 - \text{error} = \frac{(1 - \text{min\_error}) \cdot \text{epochs}}{\text{slowness} + \text{epochs}}. \quad (3)$$

Here, slowness parameter is determined through regression of Equation (3) for test error versus epochs. A lower slowness indicates faster training and a higher convergence rate.

Fig. 3 shows the test error and fitting the test error by the *plateau equation for learning curves* on MNIST, CIFAR-10, CIFAR-100, CHB-MIT, and MIT-BIH. If the test error plateaus over epochs, meaning that there is no further decrease in the test error over a certain number of next epochs, it indicates that the training process has converged. Taking the MNIST dataset as an example, BP converges at the 29th epoch, while DRTP converges around the 100th epoch. In addition, PEPITA converges at the 71st epoch, and FF con-

Table 2: Convergence rate (*slowness*) for test error. A lower slowness indicates faster training and a higher convergence rate. We highlight the **best** and second best results among the forward-only algorithms.

Algorithms	MNIST (LeCun, 1998)	CIFAR-10 (Krizhevsky, 2009)	CIFAR-100 (Krizhevsky, 2009)	CHB-MIT (Shoeb, 2009)	MIT-BIH (Mark et al., 1982)
DRTP	1.494	2.829	256.55	7.401	3.632
PEPITA	0.223	12.517	18.837	6.018	2.746
FF	0.541	51.647	336.99	2.824	12.641
Bio-FO	**0.156**	**0.883**	**4.357**	**1.144**	**1.297**
BP	0.125	1.028	5.072	0.598	0.190

Table 3: Energy consumption (Watt-hour (Wh)) for the state-of-the-art DRTP (Frenkel et al., 2021), PEPITA (Dellaferrera et al., 2022b), FF (Hinton, 2022), and Bio-FO.

Algorithms	MNIST (LeCun, 1998)	CIFAR-10 (Krizhevsky, 2009)	CIFAR-100 (Krizhevsky, 2009)	CHB-MIT (Shoeb, 2009)	MIT-BIH (Mark et al., 1982)
DRTP	121.6	110.8	131.9	6.4	317.7
PEPITA	**89.9**	91.7	123.9	5.9	191.0
FF	174.4	211.1	753.5	4.8	221.9
Bio-FO	99.8	**83.1**	**37.9**	**3.5**	**121.1**

verges at the 83rd epoch. Our proposed Bio-FO converges at the 38th epoch. Overall, these results suggest that Bio-FO enjoys faster convergence than DRTP, PEPITA, and FF.

Table 2 provides the convergence rate for test error, represented by the slowness parameters, across five datasets. The convergence rate (slowness) of Bio-FO is determined by fitting the test error to the *plateau equation for learning curves*, considering the sum of all layers. In the context of MNIST, Bio-FO has a comparable slowness to BP, and a lower slowness than DRTP, PEPITA, and FF (a lower slowness represents faster training). For CIFAR-10 and CIFAR-100, Bio-FO attains a lower slowness than BP. On the other hand, for CHB-MIT and MIT-BIH, Bio-FO demonstrates a higher slowness than BP but a significantly lower slowness than DRTP, PEPITA, and FF. Overall, these results demonstrate that Bio-FO exhibits a faster convergence rate compared to DRTP, PEPITA, and FF, approaching the convergence rate of BP.

4.2.3 *Memory Efficiency of Bio-FO*

We estimate the memory overheads of BP and Bio-FO in a similar way to (Cai et al., 2020a). Bio-FO consumes only 32.01 Megabyte (MB), while BP requires 96.06 MB in training memory with a batch size of 1 for comparison. In theory, BP needs to store/retain the parameters (weights and biases) and activations from all layers (in the forward pass) because calculating the gradients of weights (in the backward pass) requires symmetric weights and activations from top to bottom layers. In contrast, Bio-FO substitutes the forward and backward passes of BP with only forward passes in a layer-wise manner, without the need to store/retain the parameters (weights and biases) and activations from all layers. Therefore, Bio-FO improves the memory efficiency and has approximately 3 times less memory overheads.

4.2.4 *Energy Efficiency of Bio-FO*

In this section, we evaluate the energy consumption of Bio-FO, compared to DRTP, PEPITA, and FF, on the NVIDIA Jetson Nano. We utilize the normalized computation for equal comparison, i.e., considering a four-hidden-layer network for PEPITA computation and considering the number of neurons in each layer is 2000 for PEPITA and DRTP. First, we measure the training time of one epoch for these algorithms individually on the NVIDIA Jetson Nano. Next, the total training time for each algorithm is the product of the training time of one epoch, and the epoch of convergence. Then, the energy overhead is estimated by considering the power consumption of the NVIDIA Jetson Nano, which is assumed to be at 5 watts in our evaluations. Table 3 presents the energy consumption (Wh) on the NVIDIA Jetson Nano for DRTP, PEPITA, FF, and Bio-FO. For MNIST, Bio-FO consumes a comparable energy overhead with PEPITA. For CIFAR-10, CIFAR-100, CHB-MIT, and MIT-BIH datasets, Bio-FO consumes the lowest energy overhead among these forward-only algorithms including DRTP, PEPITA, and FF. Taking CIFAR-100 as an example, Bio-FO consumes 37.9 Wh of energy overhead. In contrast, DRTP consumes 131.9

Table 4: Error (%) for Bio-FO and BP with LC & CNN (DO means Dropout). We highlight the **best** and second best results for the forward-only algorithms.

Algorithms	Bio-FO-FC	Bio-FO-LC	Bio-FO-CNN	BP-FC	BP-LC	BP-LC-DO	BP-CNN
MNIST	1.62±0.08	1.36±0.06	**0.57±0.03**	1.33±0.04	1.50±0.05	1.33±0.05	0.45±0.01
CIFAR-10	45.12±0.12	35.13±0.21	**26.08±0.34**	43.62±0.33	40.21±0.63	35.17±0.09	23.15±0.85
CIFAR-100	74.57±0.51	68.62±0.16	**64.06±0.74**	72.22±0.43	72.55±0.47	65.80±0.14	64.34±0.29

Wh of energy overhead; PEPITA consumes 123.9 Wh of energy overhead; FF consumes 753.5 Wh of energy overhead. Bio-FO improve the energy efficiency up to 19.8 times compared to other state-of-the-art forward-only algorithms for CIFAR-100. Overall, these results demonstrate that our proposed Bio-FO outperforms the state-of-the-art forward-only algorithms of DRTP, PEPITA, and FF in terms of energy consumption.

4.3 Extensions to Other Architectures and Datasets

In this section, we introduce how Bio-FO can capture sparsity and be extended to common networks as well as relatively large-scale datasets. We extend the sparsity $\boldsymbol{S}$ in Equation (1) to LC and CNN.

4.3.1 Extension to LC & CNN

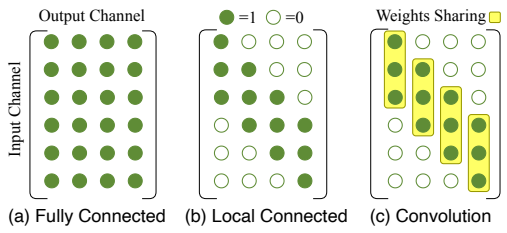


Figure 4: The sparse mask applied to $\boldsymbol{W}$.

We extend Bio-FO to LC, with only $\boldsymbol{S}_{j:j+k,j} = 1$, where j is the column index and k is the kernel size, as shown in Fig. 4 (b). In this case, $\boldsymbol{W} \in \mathbb{R}^{C \times D_l \times D_{l-1}}$, where C is the number of channels, D_{l-1} is the size of input, and D_l is the size of out. LC can be further extended to CNN with weight sharing, as shown in Fig. 4 (c). We compare Bio-FO with BP in terms of LC and CNN.

As shown in Table 4, Bio-FO with LC (Bio-FO-LC) achieves a significantly lower error than Bio-FO with FC (Bio-FO-FC). Moreover, Bio-FO-LC even surpasses BP with LC because Bio-FO-LC is less prone to overfitting. Considering MNIST, Bio-FO-LC (1.36%) exhibits its superiority when compared to DTP-LC (Lee et al., 2015) (1.46%) and its variants, and DFA-LC (Nøkland, 2016) (2.05%), according to (Bartunov et al., 2018); for CIFAR-10, Bio-FO with LC (35.13%) surpasses DTP-LC (Lee et al., 2015) (39.47%) and its variants, FA-LC (Lillicrap et al., 2016) (37.44%), DFA-LC (Nøkland, 2016) (44.41%), and FF-LC (Hinton, 2022) (43.75%).

Although weight sharing in CNN is not biologically plausible (Bartunov et al., 2018; Tang et al., 2022), we further extend Bio-FO to CNN (Bio-FO-CNN) for experimental investigation purposes. As presented in Table 4, Bio-FO-CNN has a lower error compared with Bio-FO-LC; BP-CNN has a lower error compared with BP-LC-DO in the context of three widely used datasets as in the forward-only domain (Frenkel et al., 2021; Hinton, 2022; Dellaferrera et al., 2022b). In addition, taking MNIST as an example, Bio-FO-CNN (0.57%) also demonstrates a lower error than DRTP-CNN (Frenkel et al., 2021) (1.48%), PEPITA-CNN (Dellaferrera et al., 2022b) (1.71%), Collaborative-FF (Lorberbom et al., 2024) (2.10%), CaFo-CNN (Zhao et al., 2025) (1.05%), and CwComp-CNN (Papachristodoulou et al., 2024) (0.58%). For CIFAR-10, the results demonstrate that Bio-FO-CNN decreases the test error from 35.13% to 26.08% compared to Bio-FO-LC. Bio-FO-CNN (26.08%) shows a lower error than DRTP-CNN (Frenkel et al., 2021) (31.04%), PEPITA-CNN (Dellaferrera et al., 2022b) (43.67%), Collaborative-FF (Lorberbom et al., 2024) (51.6%), and CaFo-CNN (Zhao et al., 2025) (30.52%). Moreover, Bio-FO-CNN (26.08%) has a comparable error to CwComp-CNN (Papachristodoulou et al., 2024) (from 21.89% to 27.25% depending on different predictors). In conclusion, our results present the relevance of Bio-FO with LC and CNN.

4.3.2 Extension to mini-ImageNet

We further extend Bio-FO to relatively large-scale datasets such as mini-ImageNet (Vinyals et al., 2016), and compare Bio-FO with state-of-the-art forward-only algorithms. The official implementation for CaFo (Zhao, 2023) is exploited. For the mini-ImageNet dataset, Bio-FO achieves a significantly lower error than DRTP, PEPITA, FF, and CaFo. Bio-FO achieves the closest classification performance to BP as shown in Table 5.

The significant gap between biologically-plausible forward-only algorithms and BP on relatively large-scale datasets is extensively discussed in the machine learning domain (Bartunov et al., 2018). Several algorithms, e.g., Sign-Symmetry (Xiao et al., 2019), Two-Combined Loss (Nøkland & Ei-

Table 5: Error (%) for the state-of-the-art forward-only algorithms on mini-ImageNet.

Algorithms	DRTP	PEPITA	FF	CaFo	Bio-FO	BP
mini-ImageNet	94.20±0.49	91.23±0.18	93.64±0.26	74.58±0.13	**67.39±0.25**	53.49±0.40

dnes, 2019), and DGL (Belilovsky et al., 2020) have been proposed aiming to narrow this gap. However, these algorithms still have the aforementioned biologically plausible issues. Sign-Symmetry (Xiao et al., 2019) has the biological implausibility issues of *non-locality*, *update locking*, and *frozen activities*. Two-combined Loss (Nøkland & Eidnes, 2019) has the biological implausibility issue of *update locking* as layerwise forward and backward passes are required (Frenkel et al., 2021). DGL and its variant (Belilovsky et al., 2020; 2019) suffer from the biological implausibility issue of *weight transport* because of the deeper auxiliary classifiers. In contrast, Bio-FO not only targets the biological implausibility issues of *weight transport*, *non-locality*, *update locking*, and *frozen activities*, but also makes a step forward in improving the classification performance of biologically-plausible forward-only algorithms on relatively large-scale datasets.

5 Conclusions

In this paper, we proposed an efficient on-device learning algorithm, based on the biologically-plausible forward-only algorithm, called Bio-FO, offering the potential to better mimic the learning processes of the human brain and, in turn, enhance energy efficiency. This is particularly relevant in the context of IoT and mobile devices, which are generally extremely limited in terms of resources, namely, computing power, memory storage, and battery/energy budget.

We evaluated our proposed Bio-FO in the context of several widely-used datasets such as MNIST, CIFAR-10, and CIFAR-100. In addition, to demonstrate the relevance of the proposed forward-only algorithm, we also considered two real-world medical applications on wearable devices, with extremely limited amount of resources, namely, seizure detection and arrhythmia classification, for real-time and long-term monitoring in ambulatory settings. The results show that Bio-FO outperforms the state-of-the-art forward-only algorithms, including DRTP, PEPITA, and FF, across datasets such as CIFAR-10, CIFAR-100, CHB-MIT, and MIT-BIH. At the same time, Bio-FO consistently achieves the closest classification performance to BP overall.

Finally, we implemented our proposed on-device learning algorithm on the NVIDIA Jetson Nano and evaluated it in terms of resource overheads, including computation requirements and energy consumption. Our proposed Bio-FO consistently exhibits faster convergence during the training process, compared to DRTP, PEPITA, and FF. At the same time, our evaluation demonstrated that, overall, Bio-FO is considerably more efficient in terms of resource requirements when evaluated on NVIDIA Jetson Nano.

In our future work, we plan to explore the application of the proposed biologically-plausible forward-only algorithm in the context of fine-tuning Large Language Models (LLMs) and the state-of-the-art Transformer-based models. We will investigate our proposed forward-only algorithm's performance and efficiency compared to backpropagation-based techniques and other biologically-plausible forward-only algorithms.

Limitation: Forward-only algorithms designed to enhance energy efficiency in on-device training are still in the early stages of development. As a result, in our current work, we do not evaluate our proposed scheme on hardware designed specifically for forward-only algorithms and do not optimize the forward-only algorithms in terms of efficiency or resource utilization. The extension of our evaluation on hardware designed specifically for forward-only algorithms and the optimization of our algorithm for such hardware will remain as future work. Besides, in this paper, we do not explore forward-only algorithms for advanced model architectures such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Graph Neural Networks (GNNs) (Kipf & Welling, 2017), or Transformers (Vaswani et al., 2017). As such, the extension of the results in this work to advanced model architectures will remain as future work.

Broader Impact: This paper presents the goal to advance the field of forward-only algorithms for on-device machine learning, to better mimic the learning processes of the human brain and enhance energy efficiency. In this paper, we propose an efficient on-device learning algorithm based on forward-only algorithm, which targets the biological-implausibility issues associated with the BP, i.e., *weight transport*, *non-locality*, *update locking*, and *frozen activities*. The research towards more biologically plausible training algorithms will improve our understanding of the underlying learning mechanisms in the human brain. At the same time, by bridging the performance–efficiency gap between the training mechanisms of the artificial neural networks and that of the cortex, our hope is that the new generation of forward-only algorithms also enjoys the remarkable performance and efficiency of the human brain, to reduce the environmental burden of Artificial Intelligence (AI).

REFERENCES

Achlioptas, D. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

Aminifar, A., Huang, B., Abtahi Fahliani, A., and Aminifar, A. Lightff: Lightweight inference for forward-forward algorithm. In *27th European Conference on Artificial Intelligence, ECAI-2024*, volume 392, pp. 1728–1735. IOS Press, 2024.

Arlington, V. Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms. *ANSI-AAMI EC57*, 1998.

Balasubramanian, V. Brain power. *Proceedings of the National Academy of Sciences*, 118(32): e2107022118, 2021. doi: 10.1073/pnas.2107022118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2107022118.

Barabási, D. L., Bianconi, G., Bullmore, E., Burgess, M., Chung, S., Eliassi-Rad, T., George, D., Kovács, I. A., Makse, H., Nichols, T. E., et al. Neuroscience needs network science. *Journal of Neuroscience*, 43(34):5989–5995, 2023.

Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G. E., and Lillicrap, T. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Advances in neural information processing systems*, 31, 2018.

Belilovsky, E., Eickenberg, M., and Oyallon, E. Greedy layerwise learning can scale to imagenet. In *International conference on machine learning*, pp. 583–593. PMLR, 2019.

Belilovsky, E., Eickenberg, M., and Oyallon, E. Decoupled greedy learning of cnns. In *International Conference on Machine Learning*, pp. 736–745. PMLR, 2020.

Blanken, T. F., Bathelt, J., Deserno, M. K., Voge, L., Borsboom, D., and Douw, L. Connecting brain and behavior in clinical neuroscience: A network approach. *Neuroscience & Biobehavioral Reviews*, 130:81–90, 2021.

Braitenberg, V. and Schüz, A. *Cortex: statistics and geometry of neuronal connectivity*. Springer Science & Business Media, 2013.

Büchel, J., Zendrikov, D., Solinas, S., Indiveri, G., and Muir, D. R. Supervised training of spiking neural networks for robust deployment on mixed-signal neuromorphic processors. *Scientific reports*, 11(1):23376, 2021.

Burbank, K. S. and Kreiman, G. Depression-biased reverse plasticity rule is required for stable learning at top-down connections. *PLoS computational biology*, 8(3): e1002393, 2012.

Cai, H., Gan, C., Zhu, L., and Han, S. Tinytl: Reduce memory, not parameters for efficient on-device learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11285–11297. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/81f7acabd411274fcf65ce2070ed568a-Paper.pdf.

Cai, H., Gan, C., Zhu, L., and Han, S. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020b.

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

Dellaferrera, G. Pepita, 2022. https://github.com/GiorgiaD/PEPITA [Accessed: 18-May-2024].

Dellaferrera, G., Woźniak, S., Indiveri, G., Pantazi, A., and Eleftheriou, E. Introducing principles of synaptic integration in the optimization of deep neural networks. *Nature Communications*, 13(1):1885, 2022a.

Dellaferrera, G. et al. Error-driven input modulation: solving the credit assignment problem without a backward pass. In *International Conference on Machine Learning*, pp. 4937–4955. PMLR, 2022b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H., and Pernice, W. H. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature*, 569 (7755):208–214, 2019.

Frenkel, C. Direct random target projection (drtp) - pytorch-based implementation, 2021. https://github.com/ChFrenkel/DirectRandomTargetProjection [Accessed: 18-May-2024].

Frenkel, C., Lefebvre, M., and Bol, D. Learning without feedback: Fixed random learning signals allow for feedforward training of deep neural networks. *Frontiers in neuroscience*, 15:629892, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

Gupta, M., Modi, S. K., Zhang, H., Lee, J. H., and Lim, J. H. Is bio-inspired learning better than backprop? benchmarking bio learning vs. backprop. *arXiv preprint arXiv:2212.04614*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Hebb, D. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

Hinton, G. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

Hsu, J. Ibm's new brain [news]. *IEEE spectrum*, 51(10): 17–19, 2014.

Huang, B. and Aminifar, A. Binary forward-only algorithms. *IEEE Design & Test*, pp. 1–1, 2025a. doi: 10.1109/MDAT.2025.3528366.

Huang, B. and Aminifar, A. Tinyfoa: Memory efficient forward-only algorithm for on-device learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

Huang, B., Abtahi, A., and Aminifar, A. Lightweight machine learning for seizure detection on wearable devices. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2. IEEE, 2023.

Huang, B., Abtahi, A., and Aminifar, A. Energy-aware integrated neural architecture search and partitioning for distributed internet of things (iot). *IEEE Transactions on Circuits and Systems for Artificial Intelligence*, 1(2): 257–271, 2024. doi: 10.1109/TCASAI.2024.3493036.

Jaderberg, M., Czarnecki, W. M., Osindero, S., Vinyals, O., Graves, A., Silver, D., and Kavukcuoglu, K. Decoupled neural interfaces using synthetic gradients. In *International conference on machine learning*, pp. 1627–1635. PMLR, 2017.

Jeon, I. and Kim, T. Distinctive properties of biological neural networks and recent advances in bottom-up approaches toward a better biologically plausible neural network. *Frontiers in Computational Neuroscience*, 17, 2023.

Kane, D. M. and Nelson, J. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Koban, L., Jepma, M., López-Solà, M., and Wager, T. D. Different brain networks mediate the effects of social and conditioned expectations on pain. *Nature communications*, 10(1):1–13, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, Toronto, ON, Canada, 2009.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lee, D.-H., Zhang, S., Fischer, A., and Bengio, Y. Difference target propagation. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, pp. 498–515. Springer, 2015.

Li, J., Xiong, C., and Hoi, S. Mopro: Webly supervised learning with momentum prototypes. In *International Conference on Learning Representations*, 2020.

Liao, Q., Leibo, J., and Poggio, T. How important is weight symmetry in backpropagation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):13276, 2016.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

Lin, Z., Memisevic, R., and Konda, K. How far can we go without convolution: Improving fully-connected networks. *ICLR 2016 workshop: 4th International Conference on Learning Representations*, 2016.

Lorberbom, G., Gat, I., Adi, Y., Schwing, A., and Hazan, T. Layer collaboration in the forward-forward algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14141–14148, 2024.

Löwe, S. Forward-forward, 2023. https://github.com/loeweX/Forward-Forward [Accessed: 18-May-2024].

Ma, G., Yan, R., and Tang, H. Exploiting noise as a resource for computation and learning in spiking neural networks. *Patterns*, 4(10), 2023.

Madhavan, A. Brain-inspired computing can help us create faster, more energy-efficient devices — if we win the race. National Institute of Standards and Technology, 2024. URL [Accessed: 2024-10-29].

Mark, R., Schluter, P., Moody, G., Devlin, P., and Chernoff, D. An annotated ecg database for evaluating arrhythmia detectors. In *IEEE Transactions on Biomedical Engineering*, volume 29, pp. 600–600, 1982.

Miconi, T. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife*, 6:e20899, 2017.

Mostafa, H., Ramesh, V., and Cauwenberghs, G. Deep supervised learning using local errors. *Frontiers in neuroscience*, 12:608, 2018.

Niu, S., Miao, C., Chen, G., Wu, P., and Zhao, P. Test-time model adaptation with only forward passes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Nøkland, A. Direct feedback alignment provides learning in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Nøkland, A. and Eidnes, L. H. Training neural networks with local error signals. In *International conference on machine learning*, pp. 4839–4850. PMLR, 2019.

NVIDIA. Jetson nano developer kit user guide. NVIDIA Corporation, 2019. URL https://developer.download.nvidia.com/embedded/L4T/r32-3-1_Release_v1.0/Jetson_Nano_Developer_Kit_User_Guide.pdf. [Accessed: 2024-10-30].

Papachristodoulou, A., Kyrkou, C., Timotheou, S., and Theocharides, T. Convolutional channel-wise competitive learning for the forward-forward algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14536–14544, 2024.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. Carbon emissions and large neural network training, 2021.

Pau, D. P. and Aymone, F. M. Suitability of forward-forward and pepita learning to mlcommons-tiny benchmarks. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pp. 1–6. IEEE, 2023.

Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., and Wennekers, T. Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, 22(8):488–502, 2021.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11): 1761–1770, 2019.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Sabry, F., Eltaras, T., Labda, W., Alzoubi, K., and Malluhi, Q. Machine learning for healthcare wearable devices: the big picture. *Journal of Healthcare Engineering*, 2022(1): 4653923, 2022.

Sarfraz, F., Arani, E., and Zonooz, B. A study of biologically plausible neural network: The role and interactions of brain-inspired mechanisms in continual learning. *Transactions on Machine Learning Research*, 2023.

Savazzi, S., Rampa, V., Kianoush, S., and Bennis, M. An energy and carbon footprint analysis of distributed and federated learning. *IEEE Transactions on Green Communications and Networking*, 2022.

Shervani-Tabar, N. and Rosenbaum, R. Meta-learning biologically plausible plasticity rules with random feedback pathways. *Nature Communications*, 14(1):1805, 2023.

Shi, W., Cao, J., Zhang, Q., Li, Y., and Xu, L. Edge computing: Vision and challenges. *IEEE internet of things journal*, 3(5):637–646, 2016.

Shoeb, A. H. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Sopic, D., Aminifar, A., and Atienza, D. e-glass: A wearable system for real-time detection of epileptic seizures. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. IEEE, 2018.

Srinivasan, R. F., Mignacco, F., Sorbaro, M., Refinetti, M., Cooper, A., Kreiman, G., and Dellaferrera, G. Forward learning with top-down feedback: Empirical and analytical characterization. In *The Twelfth International Conference on Learning Representations*, 2023.

Tang, M., Yang, Y., and Amit, Y. Biologically plausible training mechanisms for self-supervised learning in deep networks. *Frontiers in Computational Neuroscience*, 16: 789253, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Vishwakarma, H., Lin, H., Sala, F., and Korlakai Vinayak, R. Promises and pitfalls of threshold-based auto-labeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Vision, N. Sparse coding and decorrelation in primary visual cortex during. *science*, 287(1273), 2000.

Wang, T., Yin, L., Zou, X., Shu, Y., Rasch, M. J., and Wu, S. A phenomenological synapse model for asynchronous neurotransmitter release. *Frontiers in computational neuroscience*, 9:153, 2016.

Whittington, J. C. and Bogacz, R. Theories of error backpropagation in the brain. *Trends in cognitive sciences*, 23 (3):235–250, 2019.

Woo, S., Park, J., Hong, J., and Jeon, D. Activation sharing with asymmetric paths solves weight transport problem without bidirectional connection. *Advances in Neural Information Processing Systems*, 34:29697–29709, 2021.

Xiao, W., Chen, H., Liao, Q., and Poggio, T. Biologically-plausible learning algorithms can scale to large datasets. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SygvZ209F7.

Zhang, J., Yu, H.-F., and Dhillon, I. S. Autoassist: A framework to accelerate training of deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/9bd5ee6fe55aaeb673025dbcb8f939c1-Paper.pdf.

Zhao, G. Cafo, 2023. https://github.com/Graph-ZKY/CaFo [Accessed: 18-May-2024].

Zhao, G., Wang, T., Jin, Y., Lang, C., Li, Y., and Ling, H. The cascaded forward algorithm for neural network training. *Pattern Recognition*, 161:111292, 2025.