TinyFoA: Memory Efficient Forward-Only Algorithm for On-Device Learning

Baichuan Huang, Amir Aminifar

Department of Electrical and Information Technology, Lund University, Sweden

AAAI 2025

Motivation

Today, the majority of the state-of-the-art Artificial Neural Networks s (ANNs)/Deep Neural Networks (DNNs) are trained based on the Backpropagation (BP) algorithm. However, **BP is known to be computationally, memory, and energy-hungry**, which is often attributed to its biologically implausible nature [1]. This is particularly relevant in the context of Internet of Things (IoT) and mobile devices, which are limited in terms of resources, namely computing power, memory storage, and battery/energy budget. In this work, we focus mainly on memory efficiency, which is also known to be connected to energy/power consumption.

Method

In this paper, we propose a memory-efficient forward-only algorithm, called TinyFoA, built on joint horizontal and vertical layer-wise training, where it focuses on training only parts of a single layer. As such, TinyFoA only requires storing a single layer at any given time since only one layer updates its parameters per iteration, as shown in Fig. 1 b. By updating only a portion of the full-precision weights with their corresponding gradients, the training memory overhead is even further reduced, as shown in Fig. 1 c.

To further reduce dynamic training memory overheads, we propose to **binarize the weights and activations**, as shown in Fig. 1 d. Binarizing the weights reduces training memory overheads by ~32 times in forward pass; binarizing activations can accelerate the forward pass by up to ~64 times.

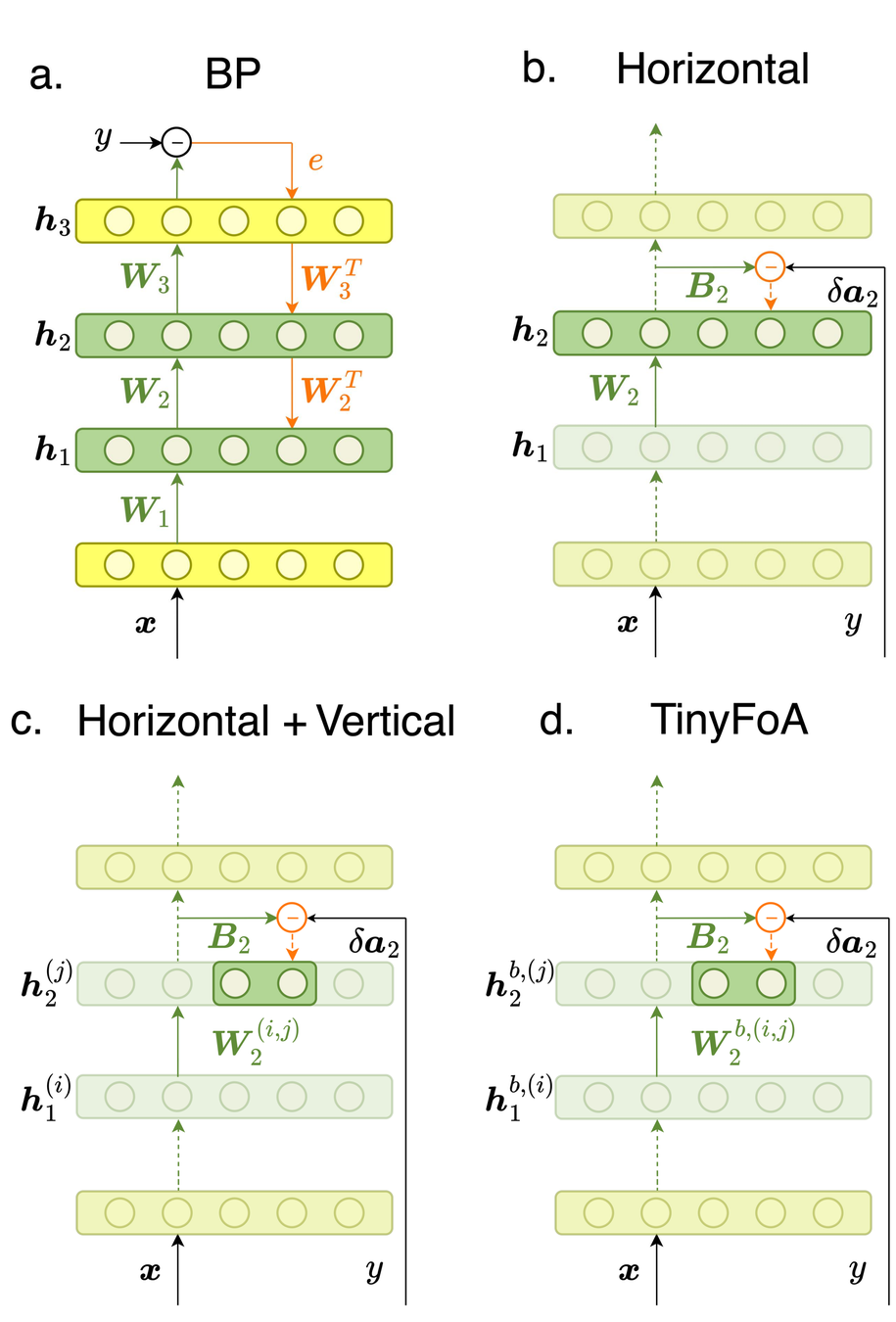

Fig. 1: An overview of TinyFoA

Results

TinyFoA reduces the training memory overheads in the forward pass by **~96×**, the training memory overheads in gradient calculation by **~24×,** and the training memory overheads in parameter update by **~12×,** while maintaining a comparable classification performance to the standard BP training algorithm, as shown in Fig. 2.

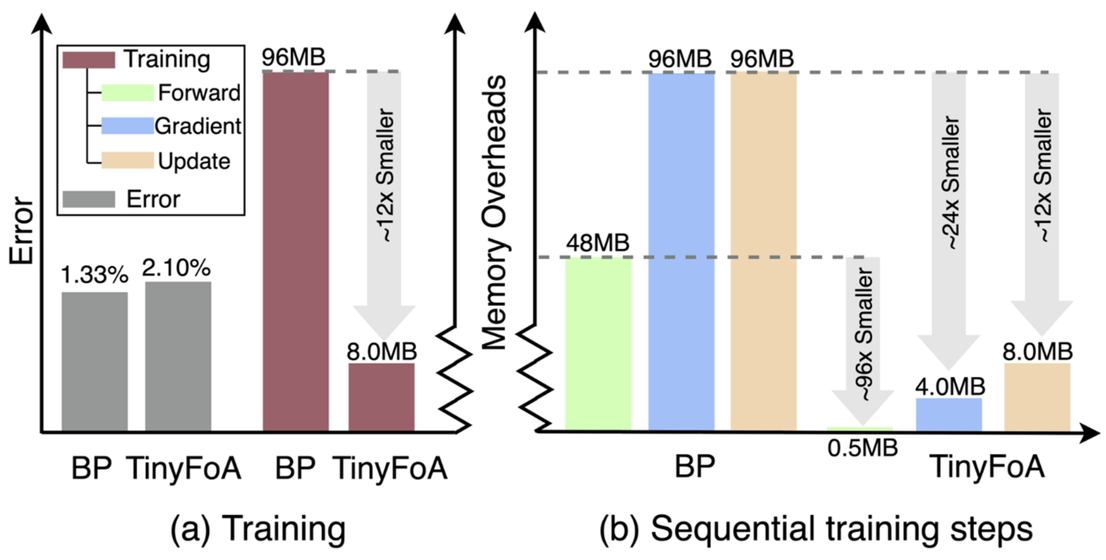

Fig. 2: An overview of memory overheads

TinyFoA outperforms state-of-the-art forward-only algorithms including PEPITA [2] and FF [3], as presented in Table 1.

Setting	MNIST (error)	CIFAR-10 (error)	Forward Pass	Gradient Calculation	Parameter Update
PEPITA [2]	5.02%	52.09%	32.04 MB	16.02 MB	32.01 MB
PEPITA+BW+BA	89.94%	90.00%	1.02 MB	16.00 MB	32.01 MB
FF [3]	1.46%	47.38%	16.04 MB	16.04 MB	32.01 MB
FF+BW+BA	90.47%	90.00%	0.51 MB	16.00 MB	32.01 MB
BP	1.33%	43.62%	48.06 MB	96.06 MB	96.06 MB
BP+BW+BA	3.11%	54.44%	1.53 MB	49.53 MB	96.06 MB
TinyFoA	2.10%	50.38%	**0.50 MB**	**4.01 MB**	**8.00 MB**

Table 1: Error (%) of classification performance and dynamic training memory overheads (BW is binary weights and BA is binary activations)

References

[1] T. P. Lillicrap, et al. Random synaptic feedback weights support error backpropagation for deep learning. Nature communications, 7(1):13276, 2016.

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

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

[4] Srinivasan, R. F., et al. Forward Learning with Top-Down Feedback: Empirical and Analytical Characterization. In The Twelfth International Conference on Learning Representations.