Efficient On-Device Machine Learning with a Biologically-Plausible Forward-Only Algorithm

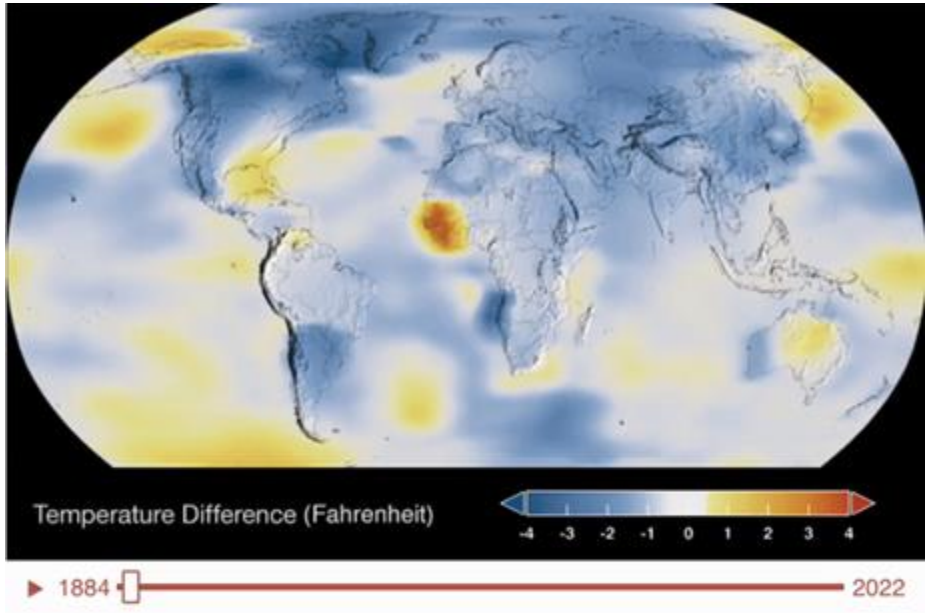
Baichuan Huang, Amir Aminifar

Department of Electrical and Information Technology, Lund University, Sweden

This research has been partially supported by the Swedish Wallenberg AI, Autonomous Systems and Software Program (WASP), the Swedish Research Council (VR), Swedish Foundation for Strategic Research (SSF), the ELLIIT Strategic Research Environment, and the European Union (EU).

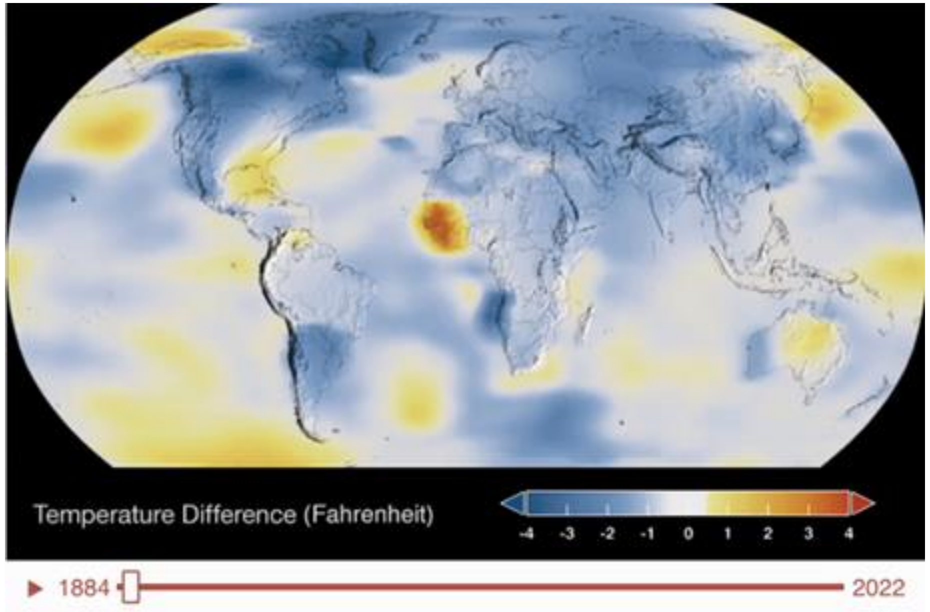
Introduction and Background

Global Warming



NASA; https://en.wikipedia.org/wiki/Climate_change_in_Europe

Global Warming



Europe: an average rise of 2.3°C compared to pre-industrial levels
1°C higher than the global average.

NASA; https://en.wikipedia.org/wiki/Climate_change_in_Europe

Energy Consumption of Training LLMs

GPT-3

GPT-4

D. Patterson, et al. Carbon emissions and large neural network training, 2021.
https://tinyml.substack.com/p/the-carbon-impact-of-large-language
Data sources: U.S. Energy Information Administration, Electric Power Research Institute (EPRI)

Energy Consumption of Training LLMs

	GPT-3		GPT-4
CO_2	1,216,950 lbs	×13	15,238,333 lbs

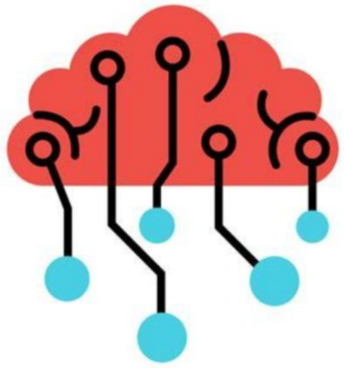
D. Patterson, et al. Carbon emissions and large neural network training, 2021.
https://tinyml.substack.com/p/the-carbon-impact-of-large-language
Data sources: U.S. Energy Information Administration, Electric Power Research Institute (EPRI)

Energy Consumption of Training LLMs

	GPT-3		GPT-4
CO_2	1,216,950 lbs	×13	15,238,333 lbs
	1,287 Megawatt-Hour	× 48	62,318 Megawatt-Hour

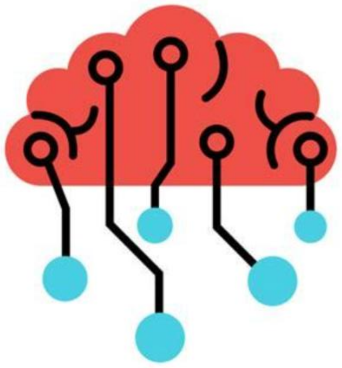
D. Patterson, et al. Carbon emissions and large neural network training, 2021.
https://tinyml.substack.com/p/the-carbon-impact-of-large-language
Data sources: U.S. Energy Information Administration, Electric Power Research Institute (EPRI)

Biologically Plausible Alternatives

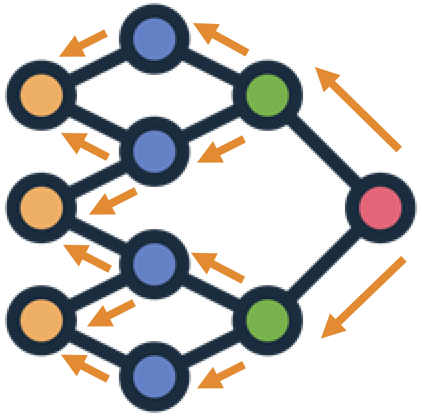


Human Brain
(~20 Watts)

Biologically Plausible Alternatives

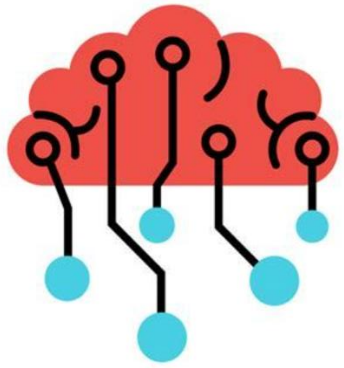


Human Brain
(~20 Watts)

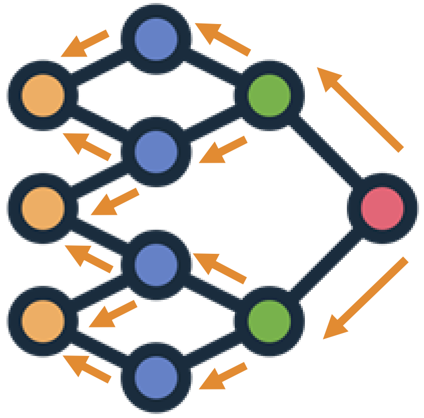


Back-Propagation
(Bio-Implausible)

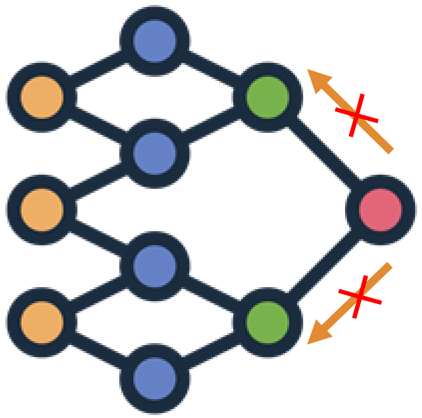
Biologically Plausible Alternatives



Human Brain
(~20 Watts)

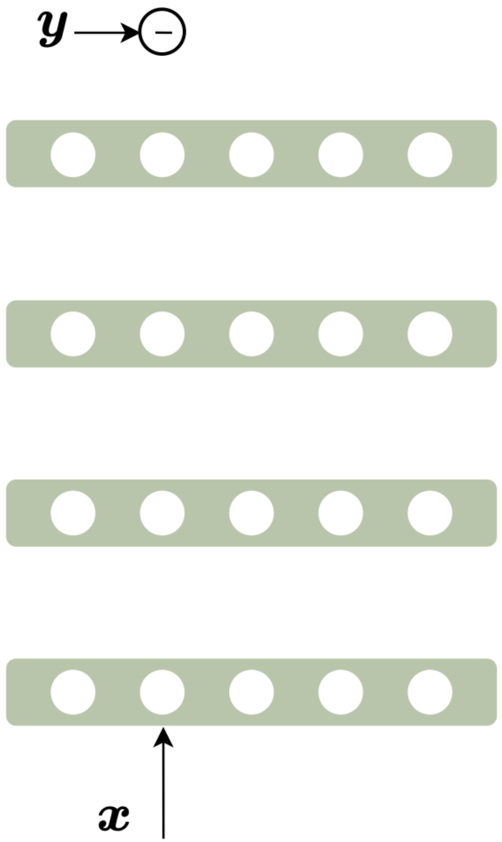


Back-Propagation
(Bio-**Implausible**)



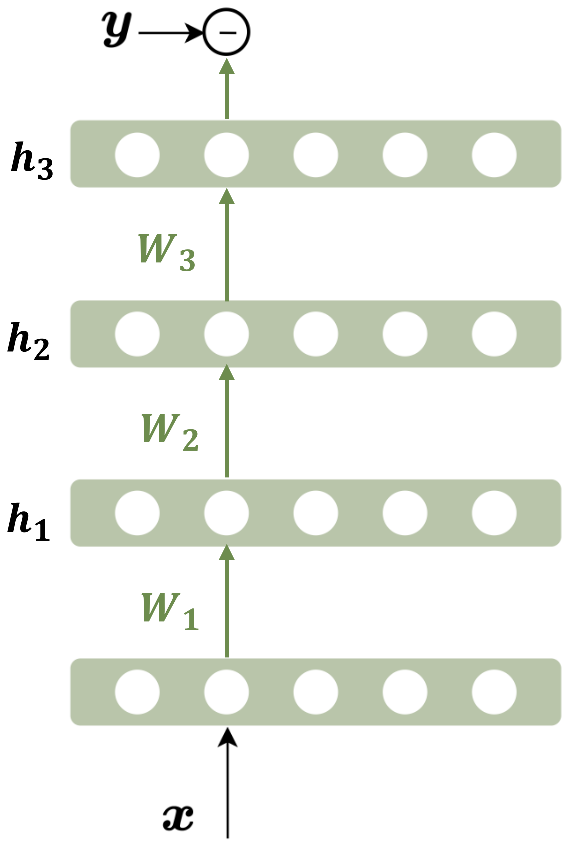
Forward-Only Algorithm
(Bio-**Plausible**)

The Process of Backpropagation



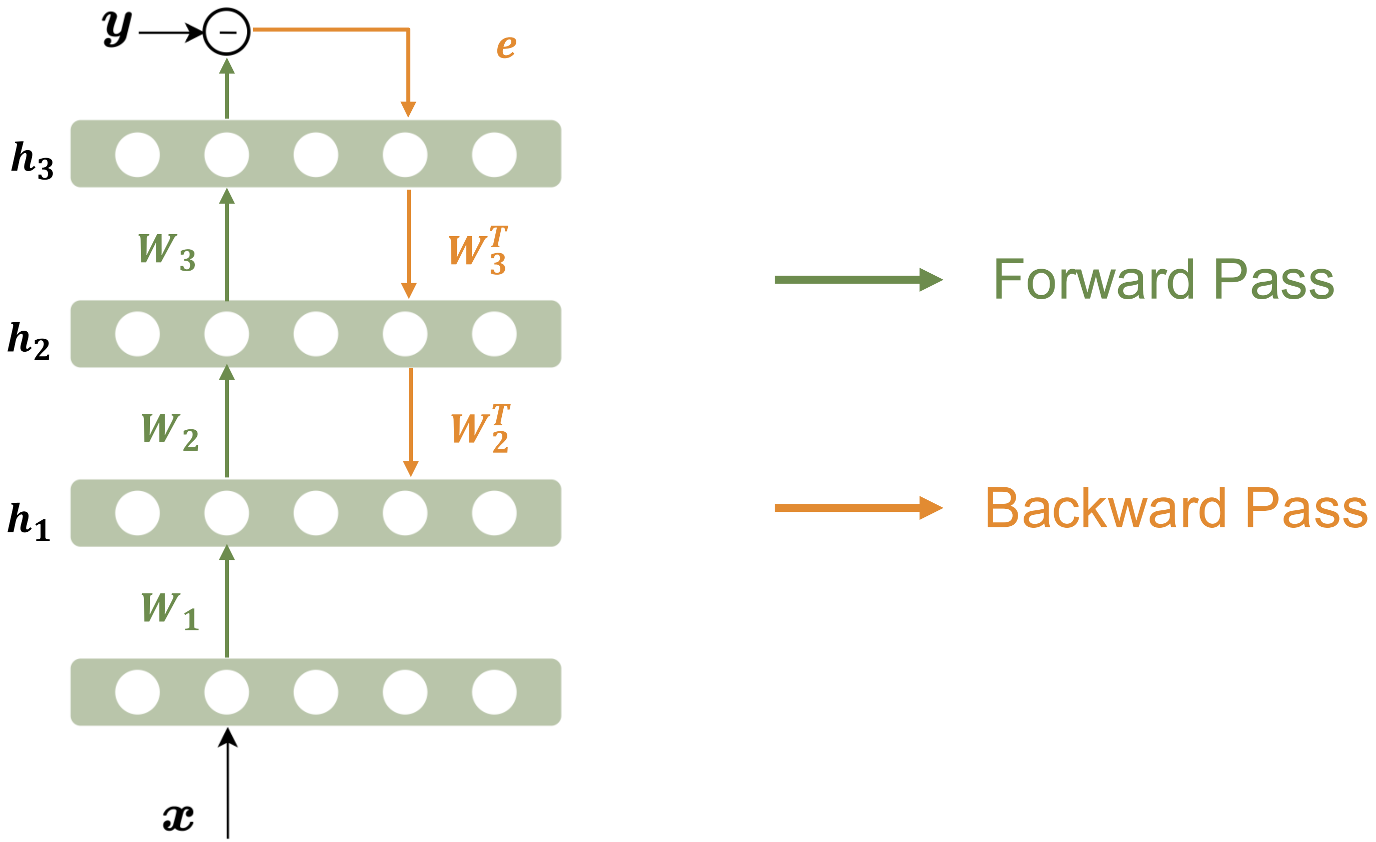
David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Process of Backpropagation



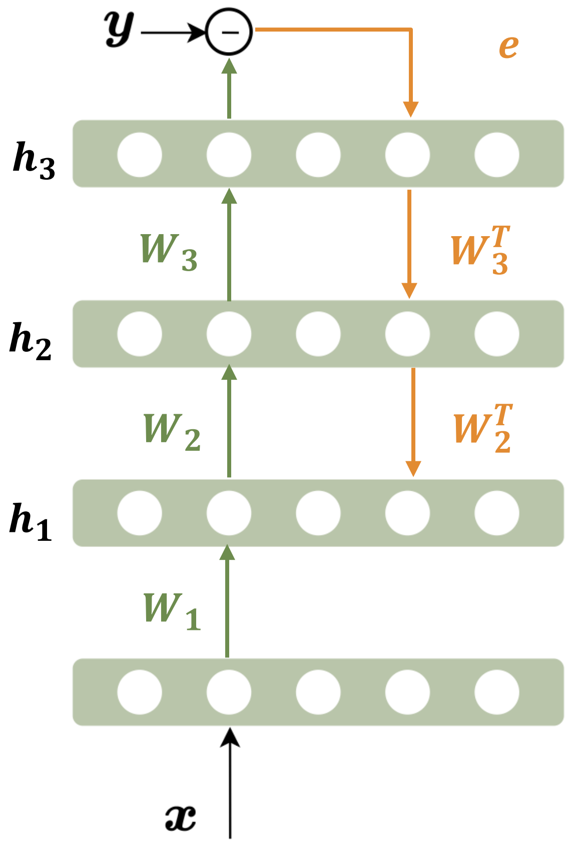
David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Process of Backpropagation



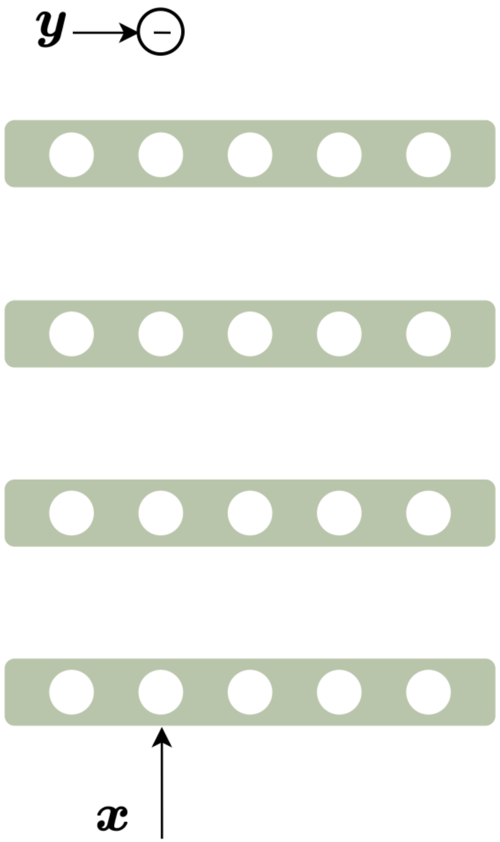
David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

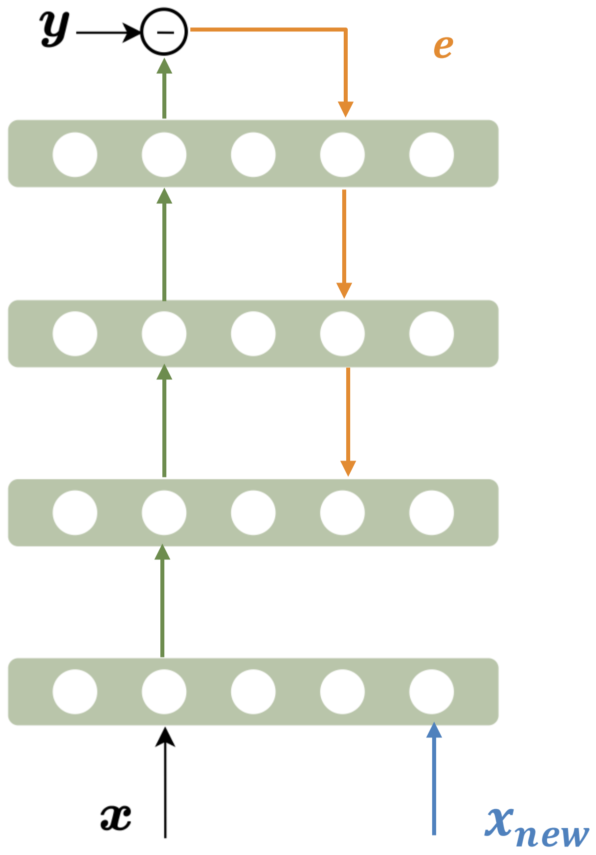
The Biological Implausibility of BP



Locking

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

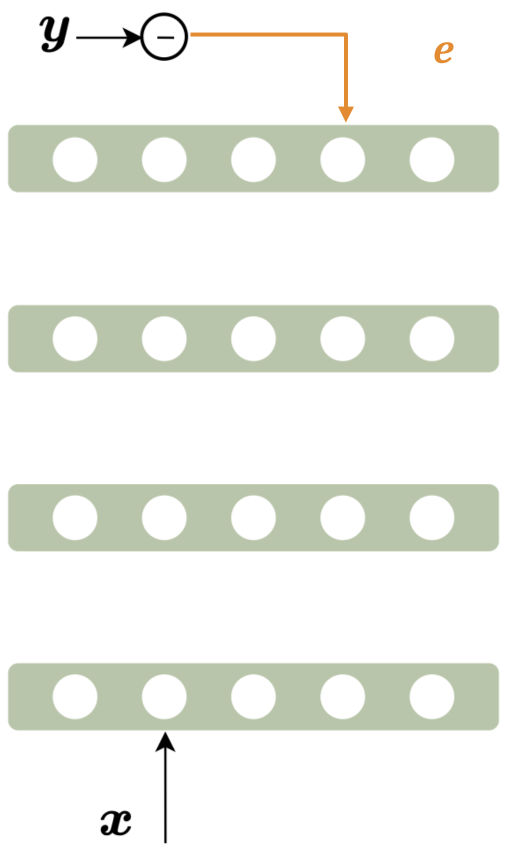
The Biological Implausibility of BP



Locking

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP

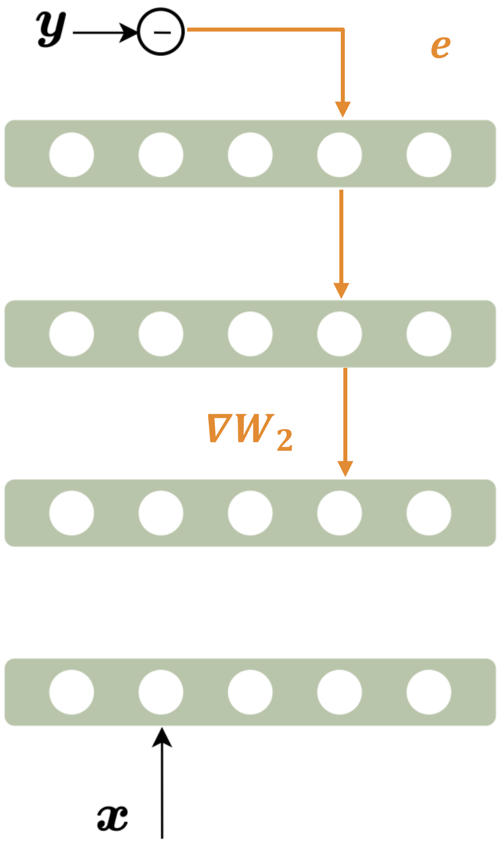


Locking

Non-Locality

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP

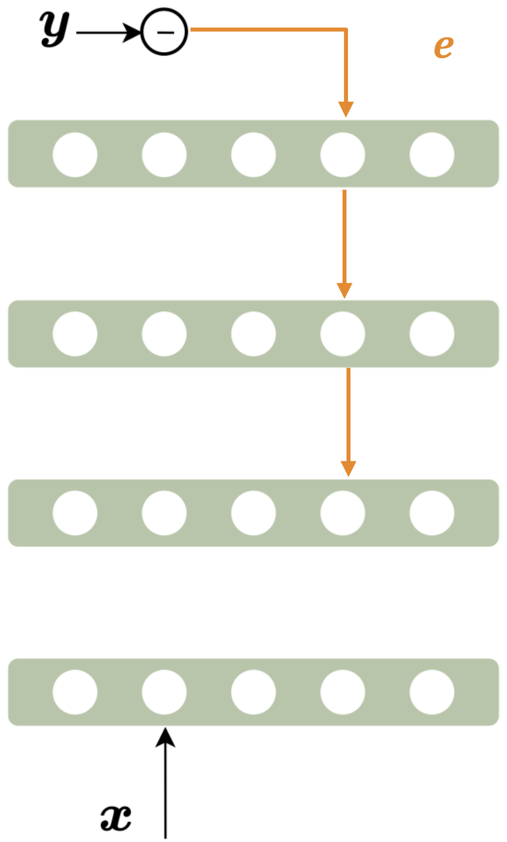


Locking

Non-Locality

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



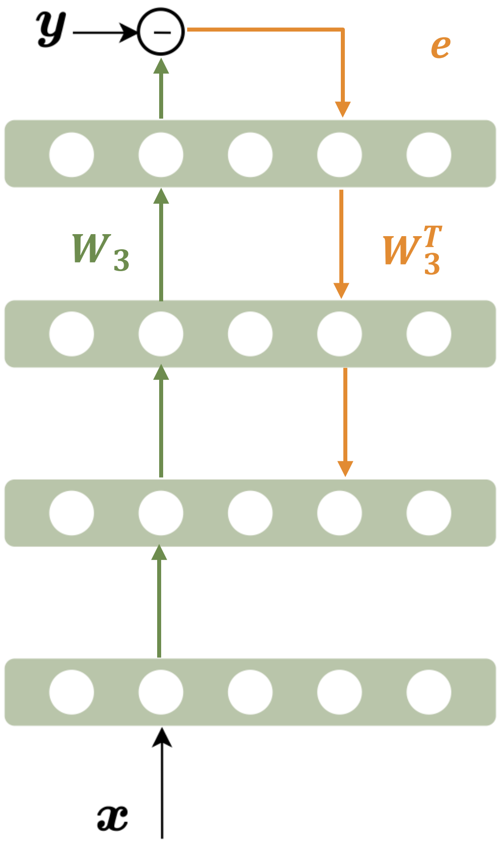
Locking

Non-Locality

Weight Transport

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



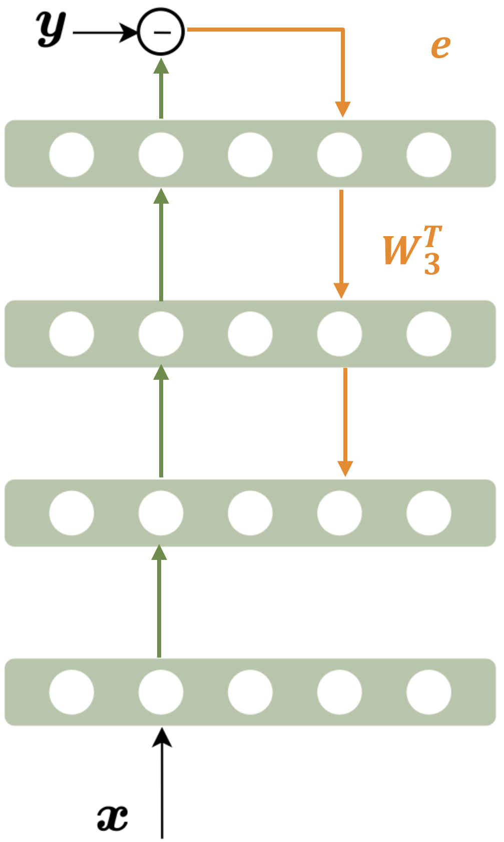
Locking

Non-Locality

Weight Transport

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



Locking

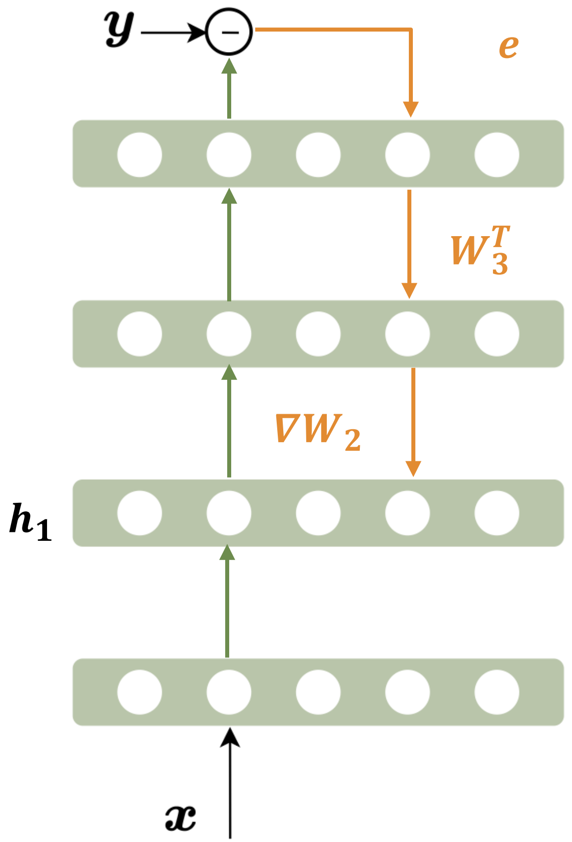
Non-Locality

Weight Transport

Frozen Activities

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



Locking

Non-Locality

Weight Transport

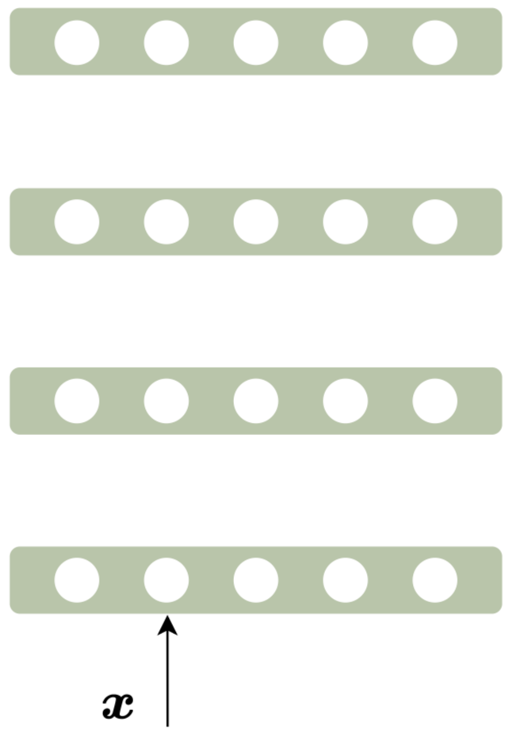
Frozen Activities

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

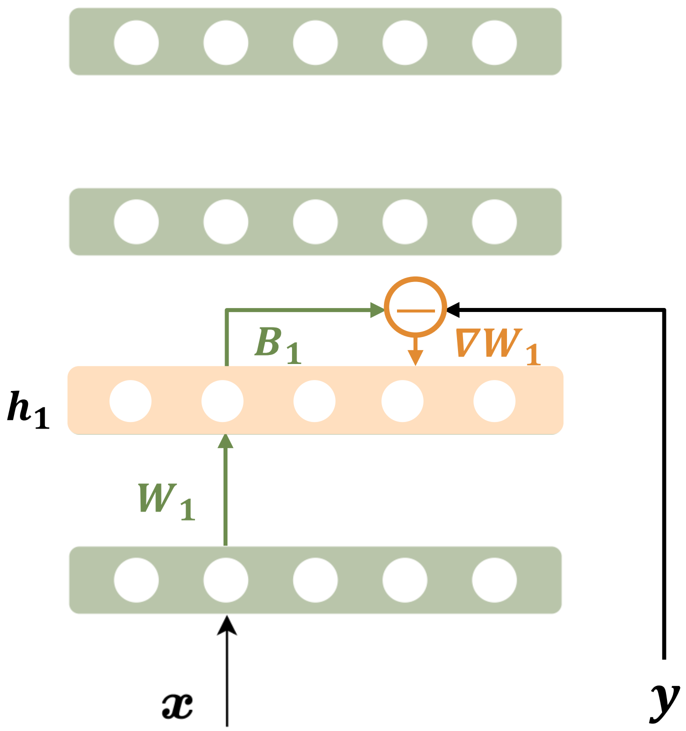
Lund University

Bio-FO: a Biologically-Plausible Forward-Only Algorithm

Our Proposed Bio-FO

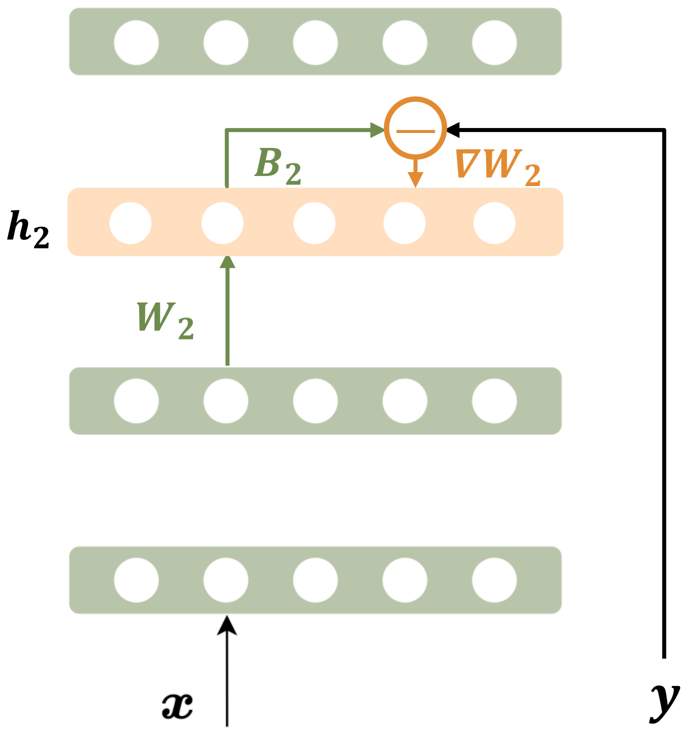


Our Proposed Bio-FO



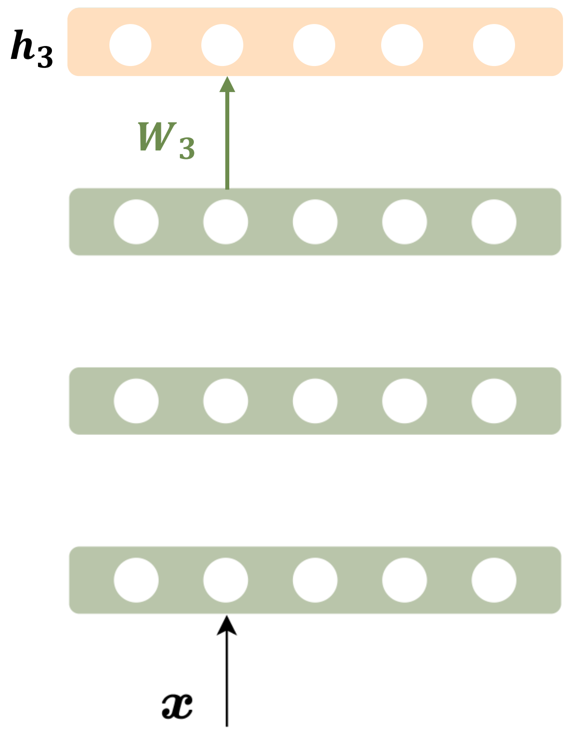
B: Fixed Random Projection

Our Proposed Bio-FO

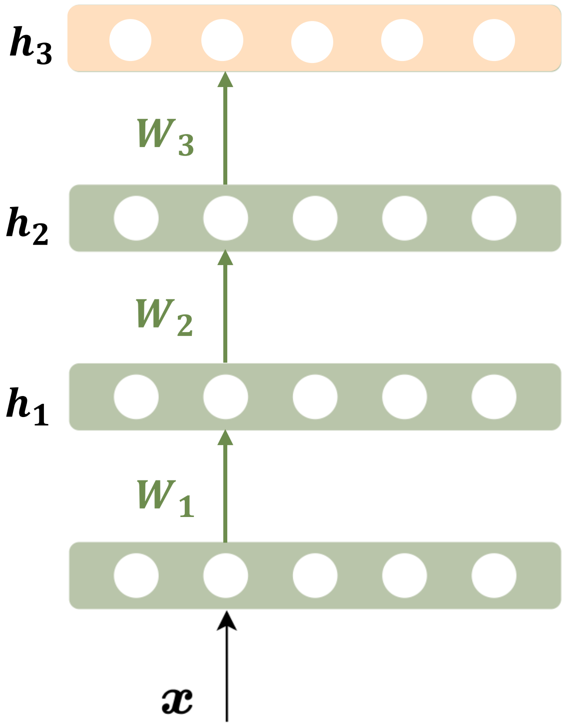


B: Fixed Random Projection

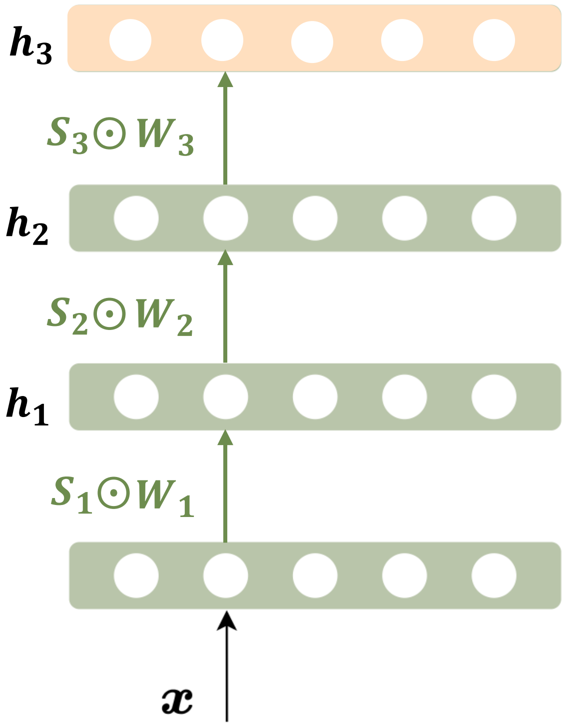
Our Proposed Bio-FO



Our Proposed Bio-FO

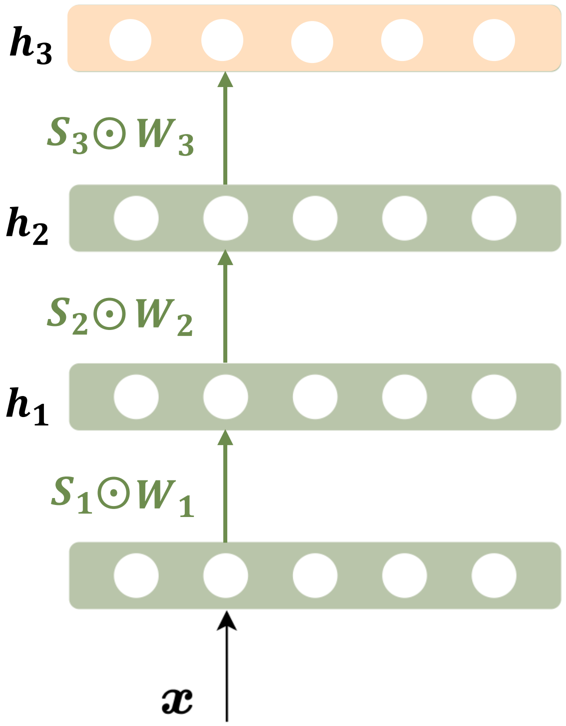


Our Proposed Bio-FO

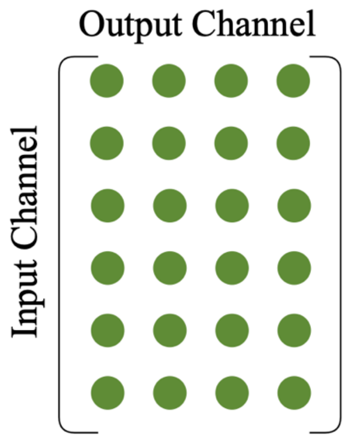


S: Sparsity Mask

Our Proposed Bio-FO

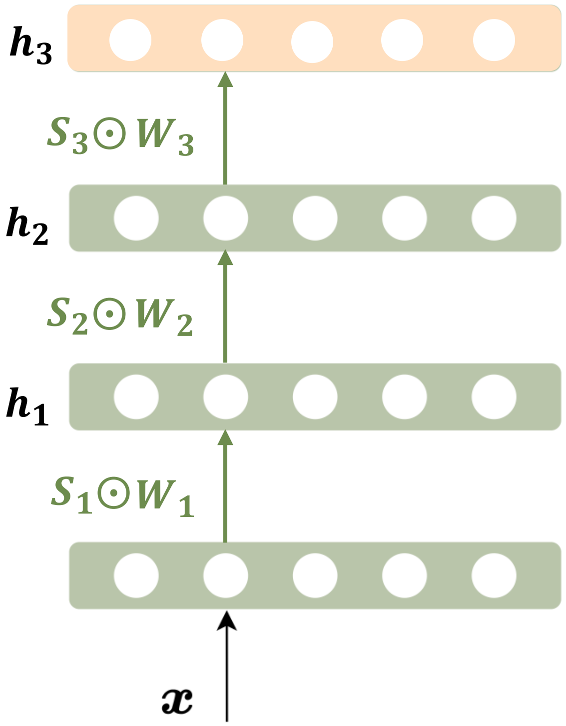


S: Sparsity Mask

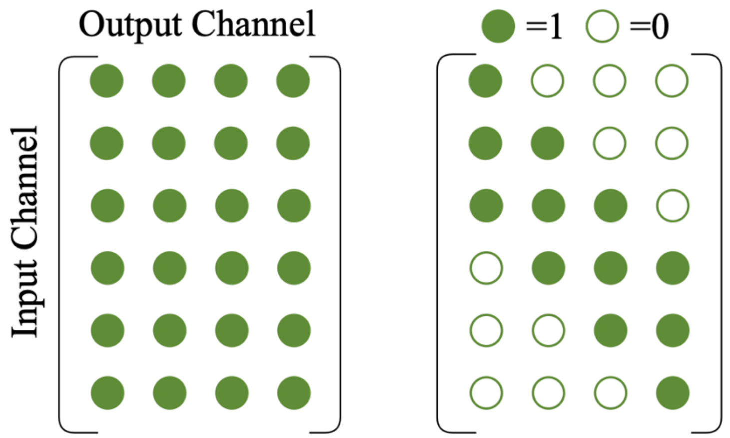


Fully Connected

Our Proposed Bio-FO

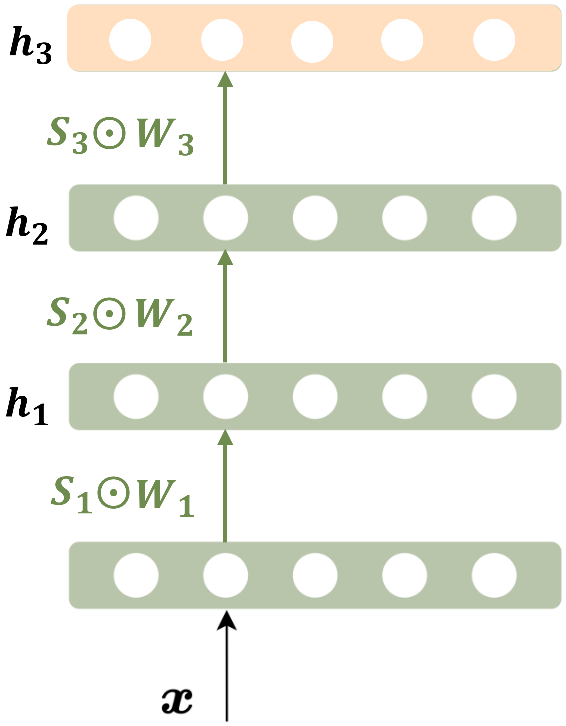


S: Sparsity Mask

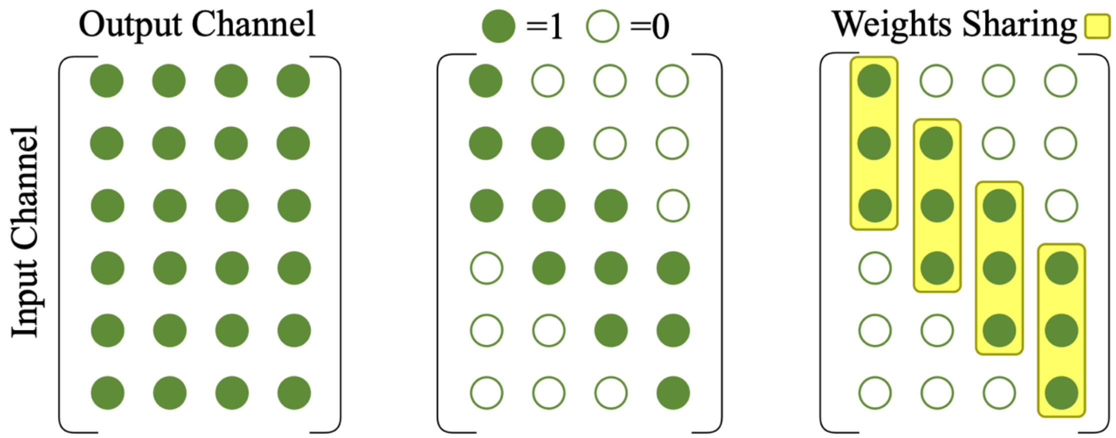


Fully Connected Local Connected

Our Proposed Bio-FO



S: Sparsity Mask



Fully Connected Local Connected CNN

Evaluation and Results

Dataset and Application

MNIST
Grayscale
Image

CIFAR-10(100)
RGB
Images

Mini-ImageNet
Subset of
ImageNet

Vinyals, O., et al. Matching networks for one shot learning. Advances in neural information processing systems, 2016.
A. H. Shoeb. Application of machine learning to epileptic seizure onset detection and treatment. PhD thesis, MIT, 2009.
R. Mark, et al. An annotated ecg database for evaluating arrhythmia detectors. IEEE Transactions on Biomedical Engineering, 1982.

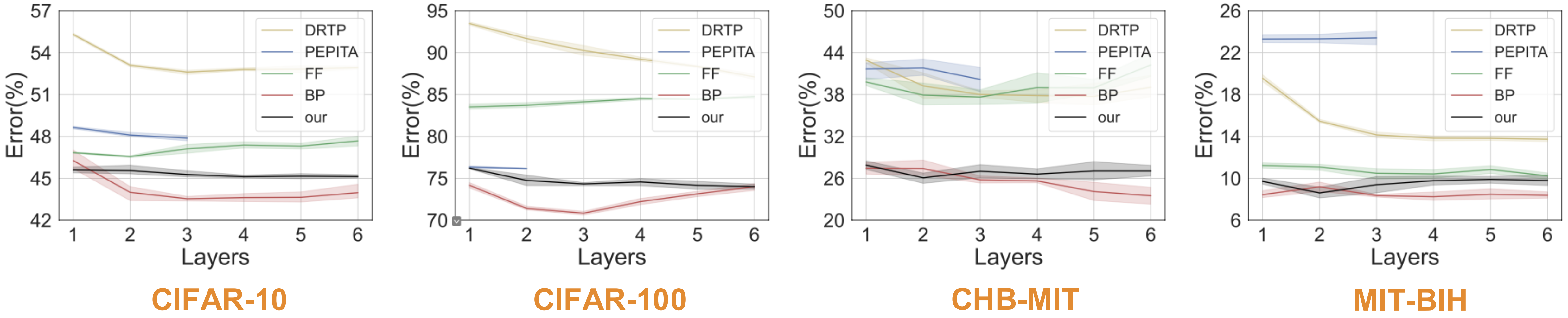
Dataset and Application

MNIST	CIFAR-10(100)	Mini-ImageNet	CHB-MIT	MIT-BIH
Grayscale Image	RGB Images	Subset of ImageNet	Electroencephalogram (EEG)	Electrocardiogram (ECG)

Real-world wearable applications:
Complexity overhead/energy consumption is a major constraint.

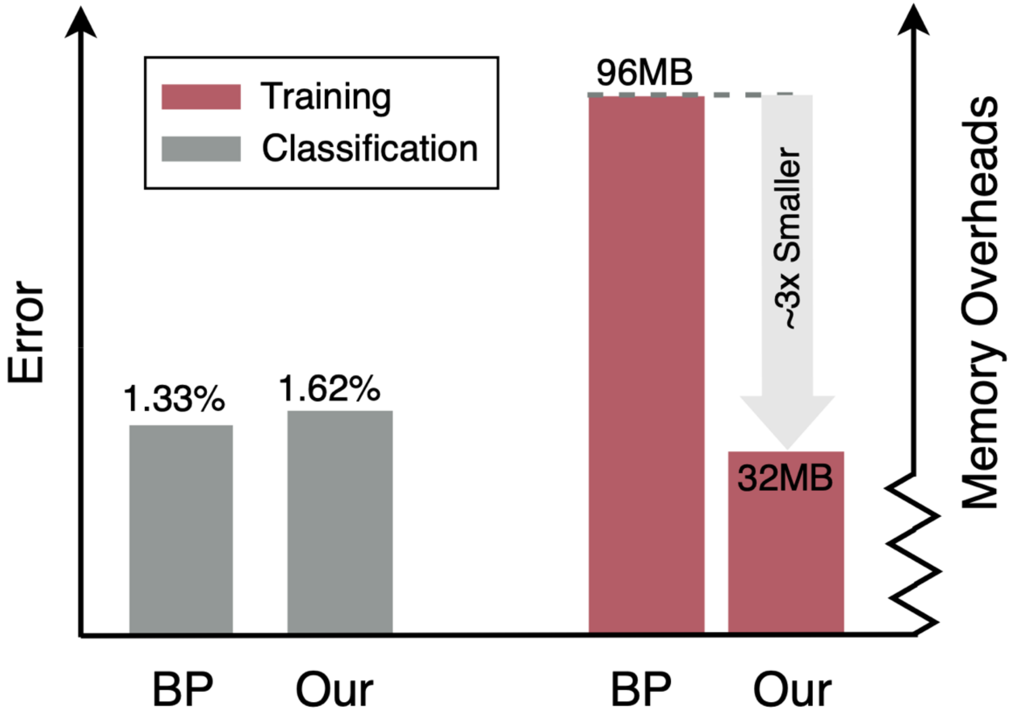
Vinyals, O., et al. Matching networks for one shot learning. Advances in neural information processing systems, 2016.
A. H. Shoeb. Application of machine learning to epileptic seizure onset detection and treatment. PhD thesis, MIT, 2009.
R. Mark, et al. An annotated ecg database for evaluating arrhythmia detectors. IEEE Transactions on Biomedical Engineering, 1982.

Classification Performance



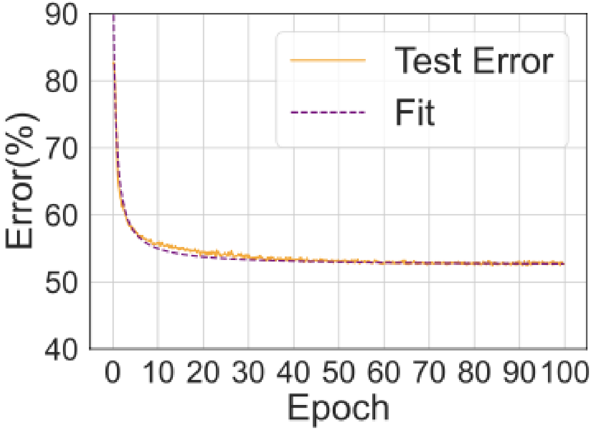
Bio-FO outperforms the state-of-the-art forward-only algorithms, with the potential to achieve comparable performance to BP.

Memory Efficiency

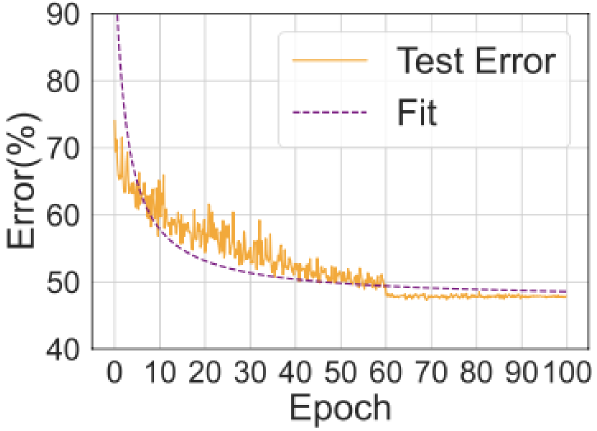


Bio-FO improves the memory efficiency and has approximately 3 times less memory overheads when compared to BP.

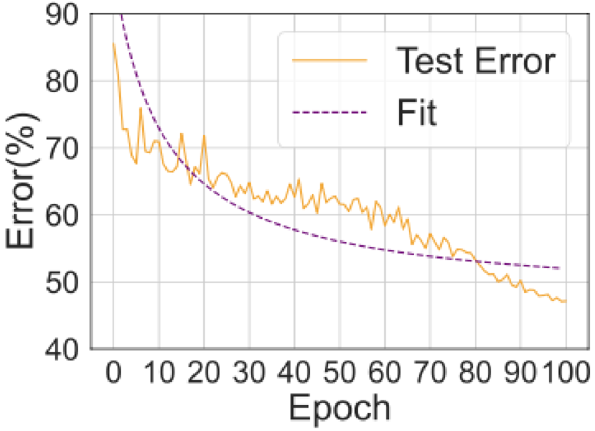
Convergence Rate (CIFAR-10)



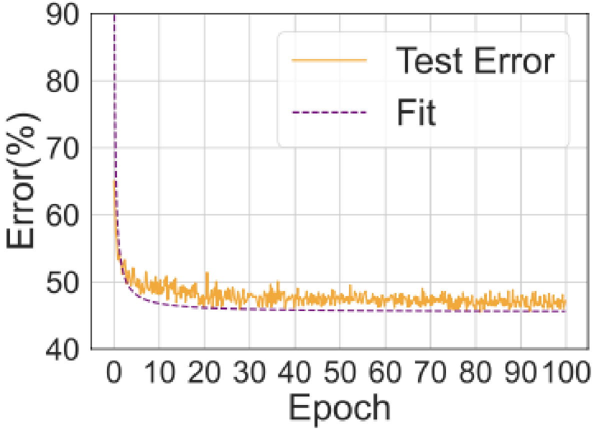
DRTP



PEPITA



FF



Our

Bio-FO enjoys faster convergence than PEPITA, and FF.

Energy Efficiency

LUND UNIVERSITY

Algorithms	Energy Overheads (Wh)		
	CIFAR-100	CHB-MIT	MIT-BIH
DRTP	131.9	6.4	317.7
PEPITA	<u>123.9</u>	5.9	<u>191.0</u>
FF	753.5	<u>4.8</u>	221.9
Our	**37.9**	**3.5**	**121.1**

Bio-FO outperforms the state-of-the-art forward-only algorithms in terms of energy consumption.

Scalability (Architectures)

Datasets	Error (%)		
	Our-FC	Our-LC	Our-CNN
MNIST	1.62	1.36	**0.57**
CIFAR-10	45.12	35.13	**26.08**
CIFAR-100	74.57	64.06	**64.06**

The relevance of Bio-FO with LC and CNN shows the importance of architectures for improving classification performance.

Scalability (mini-ImageNet)

Datasets	Error (%)				
	DRTP	PEPITA	FF	Our	BP
mini-ImageNet	94.20$\pm$0.49	91.23$\pm$0.18	93.64$\pm$0.26	67.39$\pm$0.25	53.49$\pm$0.40

Bio-FO achieves the closest classification performance to BP, on relatively large-scale datasets such as mini-ImageNet.

Conclusion

Challenge

Bio-Implausibility
Incurs
Inefficiency

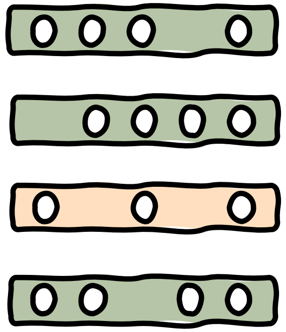
Conclusion

Challenge

Bio-Implausibility
Incurs
Inefficiency

Approach

A Biologically Plausible
Forward-Only
Algorithm



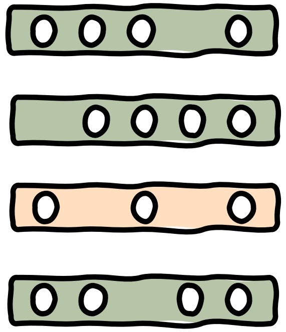
Conclusion

Challenge

Bio-Implausibility
Incurs
Inefficiency

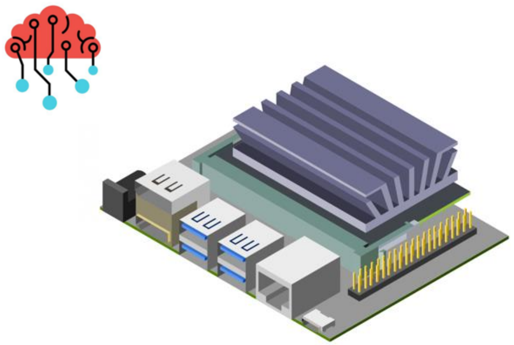
Approach

A Biologically Plausible
Forward-Only
Algorithm



Performance

Memory & Energy
Efficiency
Maintain Performance



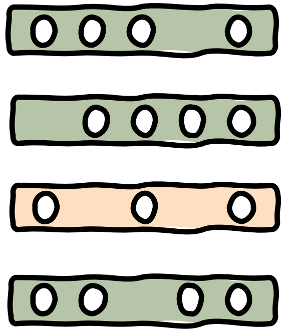
Conclusion

Challenge

Bio-Implausibility
Incurs
Inefficiency

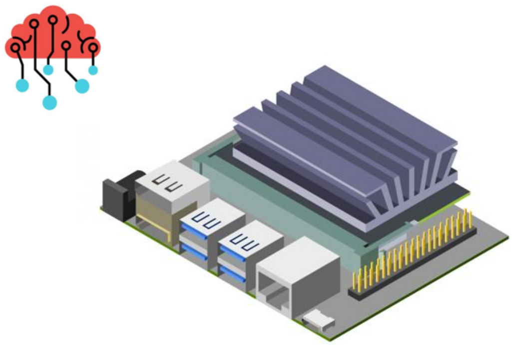
Approach

A Biologically Plausible
Forward-Only
Algorithm



Performance

Memory & Energy
Efficiency
Maintain Performance



Welcome to Our Poster Session

Thank you!