Layer-skipping connections facilitate training of layered networks using equilibrium propagation.

Jimmy Gammell Sae Woo Nam Adam N. McCaughan

July 28, 2020

- Equilibrium propagation:[1] biologically-motivated learning framework

[1] Benjamin Scellier and Yoshua Bengio. *Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation*. 2016. arXiv: 1602.05179 [cs.LG].

- Equilibrium propagation:[1] biologically-motivated learning framework
    - Simple neurons relative to backpropagation

[1] Scellier and Bengio, *Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation*.

- Equilibrium propagation:[1] biologically-motivated learning framework
  - Simple neurons relative to backpropagation
  - Potential application of neuromorphic analog hardware

[1] Scellier and Bengio, *Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation*.

Introduction

- Equilibrium propagation is a biologically-motivated learning framework

Introduction

- Equilibrium propagation is a biologically-motivated learning framework
- Appealing as potential application for neuromorphic analog hardware

Introduction

- Equilibrium propagation is a biologically-motivated learning framework
- Appealing as potential application for neuromorphic analog hardware
- Vanishing gradient problem with conventional layered network topology

Introduction

- Equilibrium propagation is a biologically-motivated learning framework
- Appealing as potential application for neuromorphic analog hardware
- Vanishing gradient problem with conventional layered network topology
- We will show that adding layer-skipping connections can help solve the problem

Introduction

- Equilibrium propagation is a biologically-motivated learning framework
- Appealing as potential application for neuromorphic analog hardware
- Vanishing gradient problem with conventional layered network topology
- We will show that adding layer-skipping connections can help solve the problem
  - Still biologically-plausible

Introduction

- Equilibrium propagation is a biologically-motivated learning framework
- Appealing as potential application for neuromorphic analog hardware
- Vanishing gradient problem with conventional layered network topology
- We will show that adding layer-skipping connections can help solve the problem
  - Still biologically-plausible
  - Simple to implement in hardware with configurable connectivity

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio
- Biologically-motivated learning framework

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio
- Biologically-motivated learning framework
  - For energy-based networks, e.g. continuous Hopfield network

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio
- Biologically-motivated learning framework
  - For energy-based networks, e.g. continuous Hopfield network
  - Gradient descent on cost function

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio
- Biologically-motivated learning framework
  - For energy-based networks, e.g. continuous Hopfield network
  - Gradient descent on cost function
- More biologically-plausible than backpropagation

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio
- Biologically-motivated learning framework
  - For energy-based networks, e.g. continuous Hopfield network
  - Gradient descent on cost function
- More biologically-plausible than backpropagation
  - Neurons only need to communicate activation values, not error-correction information

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio
- Biologically-motivated learning framework
  - For energy-based networks, e.g. continuous Hopfield network
  - Gradient descent on cost function
- More biologically-plausible than backpropagation
  - Neurons only need to communicate activation values, not error-correction information
  - Prediction (free) and correction (weakly-clamped) phases use same type of computation

Background: equilibrium propagation

- Seminal paper: Equilibrium Propagation: Bridging the Gap Between Energy-Based Models and Backpropagation by Benjamin Scellier and Yoshua Bengio
- Biologically-motivated learning framework
  - For energy-based networks, e.g. continuous Hopfield network
  - Gradient descent on cost function
- More biologically-plausible than backpropagation
  - Neurons only need to communicate activation values, not error-correction information
  - Prediction (free) and correction (weakly-clamped) phases use same type of computation
- For same reasons, appealing for implementation on neuromorphic analog hardware

Background: vanishing gradient problem

- Vanishing gradient problem in conventional layered networks

Background: vanishing gradient problem

- Vanishing gradient problem in conventional layered networks
  - Leads to slow training

Background: vanishing gradient problem

- Vanishing gradient problem in conventional layered networks
  - Leads to slow training
  - Potential issues in systems with limited bit depth

Background: vanishing gradient problem

- Vanishing gradient problem in conventional layered networks
  - Leads to slow training
  - Potential issues in systems with limited bit depth
- Original paper solved by manually tuning an independent learning rate for each layer - unappealing solution

Background: vanishing gradient problem

- Vanishing gradient problem in conventional layered networks
  - Leads to slow training
  - Potential issues in systems with limited bit depth
- Original paper solved by manually tuning an independent learning rate for each layer - unappealing solution
  - More hyperparameters to tune

Background: vanishing gradient problem

- Vanishing gradient problem in conventional layered networks
  - Leads to slow training
  - Potential issues in systems with limited bit depth
- Original paper solved by manually tuning an independent learning rate for each layer - unappealing solution
  - More hyperparameters to tune
  - Inconvenient to implement in neuromorphic analog hardware

Background: vanishing gradient problem

- Vanishing gradient problem in conventional layered networks
  - Leads to slow training
  - Potential issues in systems with limited bit depth
- Original paper solved by manually tuning an independent learning rate for each layer - unappealing solution
  - More hyperparameters to tune
  - Inconvenient to implement in neuromorphic analog hardware
  - Seems unlikely to happen in biological systems

Our solution: layer-skipping connections

- Our solution: modify layered topology by adding random layer-skipping connections

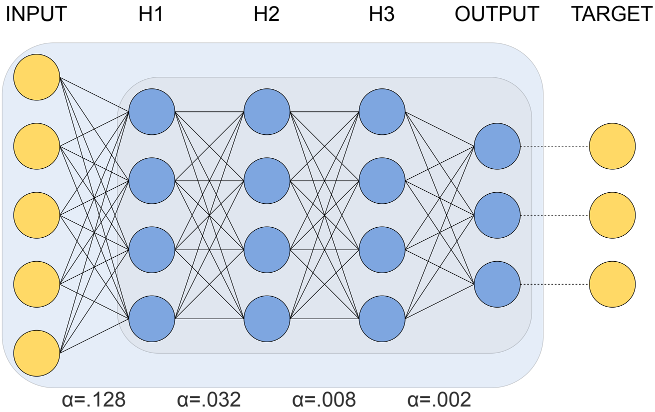
Our solution: layer-skipping connections

- Our solution: modify layered topology by adding random layer-skipping connections
  - Inspired by small-world topology, but little correlation with common small-worldness metrics (e.g. characteristic path length, clustering coefficient, small-world coefficient)

Our solution: layer-skipping connections

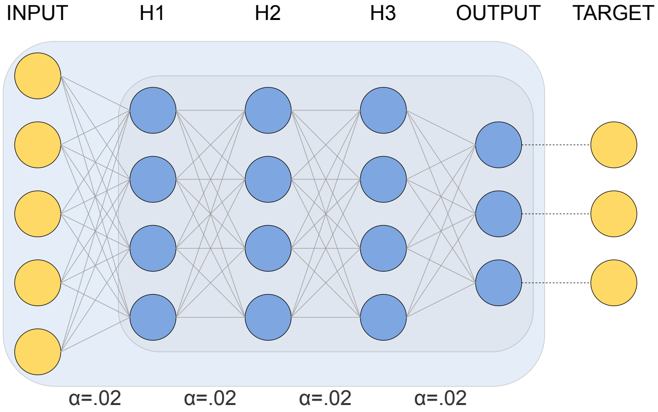
- Our solution: modify layered topology by adding random layer-skipping connections
  - Inspired by small-world topology, but little correlation with common small-worldness metrics (e.g. characteristic path length, clustering coefficient, small-world coefficient)
  - Could be conveniently implemented in neuromorphic systems with configurable connectivity

Conventional layered topology



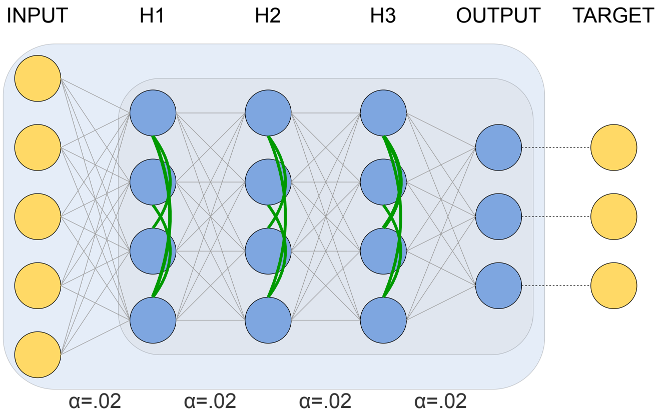
- Topology used in original paper
- Independently-tuned learning rates for each layer

Our topological modifications



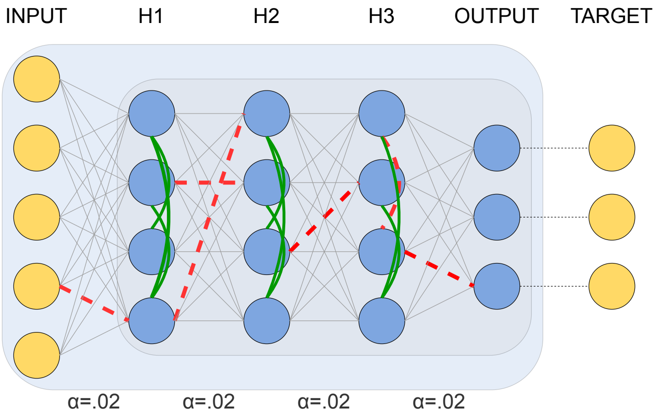
- Use original topology as starting point
- Single global learning rate across all layers

Our topological modifications



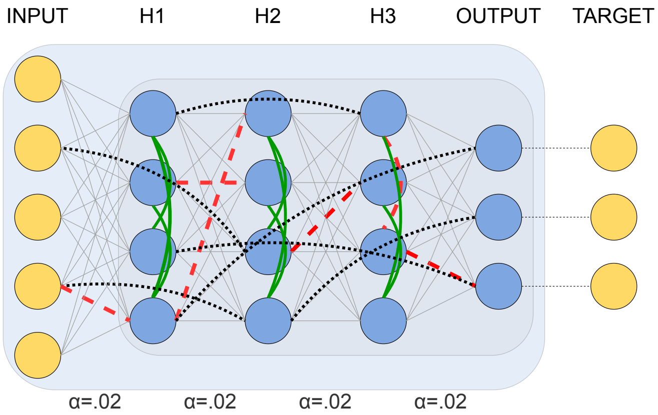
- Make hidden layers fully-connected

Our topological modifications



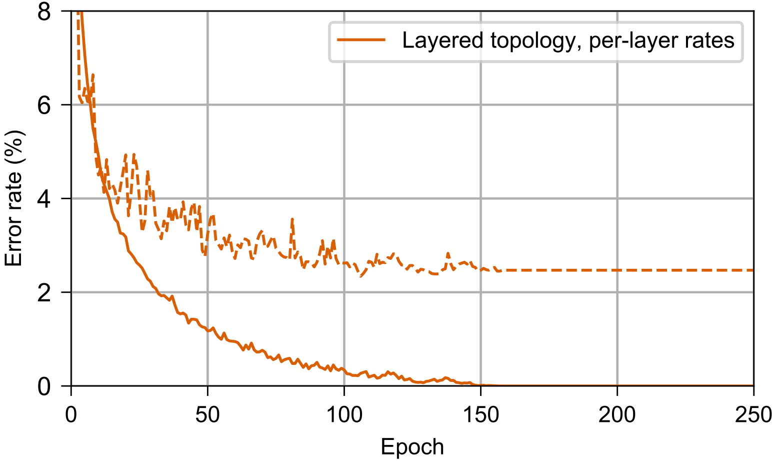
- Consider each connection and remove with probability p

Our topological modifications

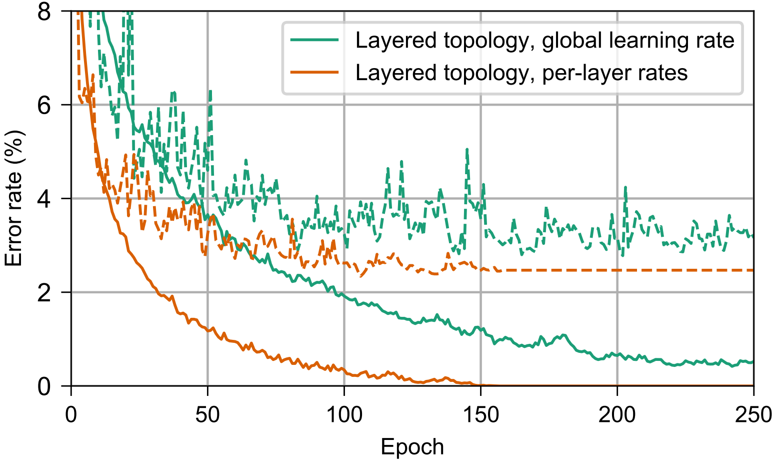


- For each removed connection, randomly connect two separated neurons
  - Only 1 connection per pair
  - No connections within input or output layers

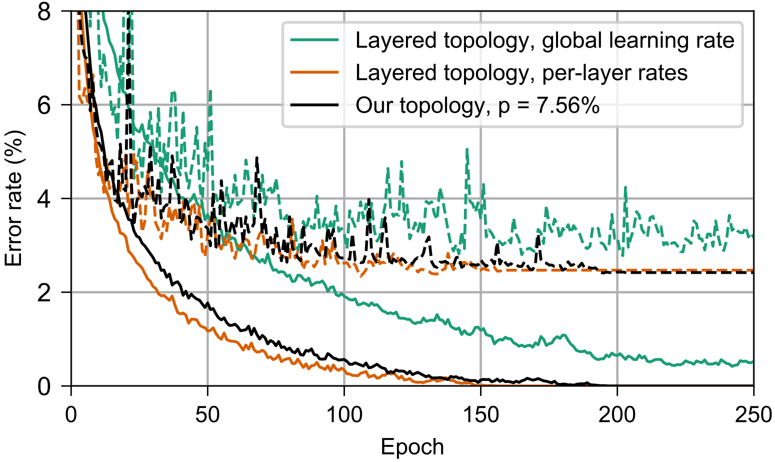
Results: performance of network with layer-skipping connections



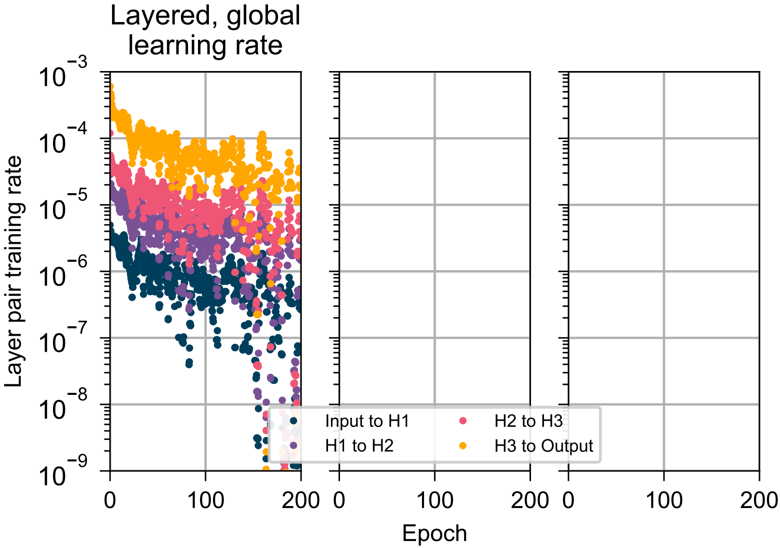
Results: performance of network with layer-skipping connections



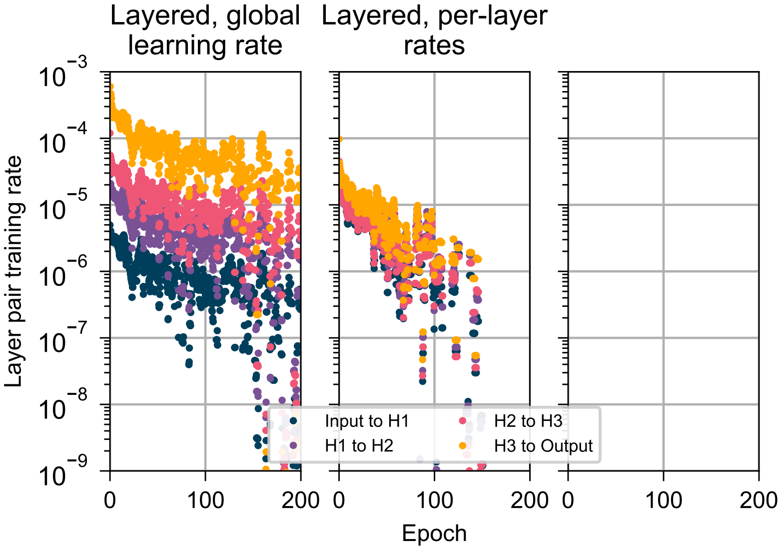
Results: performance of network with layer-skipping connections



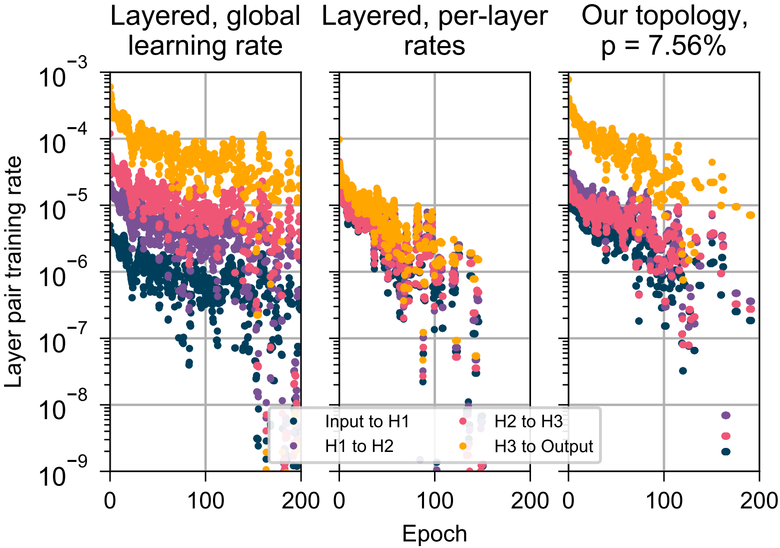
Results: effect on training rates of layers



Results: effect on training rates of layers



Results: effect on training rates of layers



Results: takeaways

Directions for future research

- Evaluate effectiveness on harder datasets, e.g. CIFAR, ImageNet, where network depth is very important

Directions for future research

- Evaluate effectiveness on harder datasets, e.g. CIFAR, ImageNet, where network depth is very important
- Evaluate effect of p on network test error

Directions for future research

- Evaluate effectiveness on harder datasets, e.g. CIFAR, ImageNet, where network depth is very important
- Evaluate effect of p on network test error
- Evaluate effectiveness on deeper networks

Directions for future research

- Evaluate effectiveness on harder datasets, e.g. CIFAR, ImageNet, where network depth is very important
- Evaluate effect of p on network test error
- Evaluate effectiveness on deeper networks
- Try training a network with added layer-skipping connections, then removing them afterwards