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

Jimmy Gammell   Sae Woo Nam   Adam N. McCaughan

July 28, 2020

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a 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].

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a biologically-motivated learning framework
  - Gradient descent on cost function (alternative to backpropagation)

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

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a biologically-motivated learning framework
  - Gradient descent on cost function (alternative to backpropagation)
  - Energy-based networks, e.g. continuous Hopfield network

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

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a biologically-motivated learning framework
  - Gradient descent on cost function (alternative to backpropagation)
  - Energy-based networks, e.g. continuous Hopfield network
- Advantageous due to simplicity of neurons

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

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a biologically-motivated learning framework
  - Gradient descent on cost function (alternative to backpropagation)
  - Energy-based networks, e.g. continuous Hopfield network
- Advantageous due to simplicity of neurons
  - One computation in both phases of training

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

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a biologically-motivated learning framework
  - Gradient descent on cost function (alternative to backpropagation)
  - Energy-based networks, e.g. continuous Hopfield network
- Advantageous due to simplicity of neurons
  - One computation in both phases of training
  - One type of information to transmit in both phases of training

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

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a biologically-motivated learning framework
  - Gradient descent on cost function (alternative to backpropagation)
  - Energy-based networks, e.g. continuous Hopfield network
- Advantageous due to simplicity of neurons
  - One computation in both phases of training
  - One type of information to transmit in both phases of training
  - Biologically plausible (relative to backpropagation)

---

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

Motivation: appeal of equilibrium propagation

- Equilibrium propagation:[1] a biologically-motivated learning framework
  - Gradient descent on cost function (alternative to backpropagation)
  - Energy-based networks, e.g. continuous Hopfield network
- Advantageous due to simplicity of neurons
  - One computation in both phases of training
  - One type of information to transmit in both phases of training
  - Biologically plausible (relative to backpropagation)
  - Implementable in neuromorphic analog hardware

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

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
    - Leads to slow training

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues
  - State-of-the-art performance by deep networks

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues
  - State-of-the-art performance by deep networks
- Not yet solved in simple, biologically-plausible manner

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues
  - State-of-the-art performance by deep networks
- Not yet solved in simple, biologically-plausible manner
- Original paper: manually tune independent learning rate for each layer

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues
  - State-of-the-art performance by deep networks
- Not yet solved in simple, biologically-plausible manner
- Original paper: manually tune independent learning rate for each layer
- Per-layer rates unappealing for following reasons:

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues
  - State-of-the-art performance by deep networks
- Not yet solved in simple, biologically-plausible manner
- Original paper: manually tune independent learning rate for each layer
- Per-layer rates unappealing for following reasons:
  1. More hyperparameters to tune

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues
  - State-of-the-art performance by deep networks
- Not yet solved in simple, biologically-plausible manner
- Original paper: manually tune independent learning rate for each layer
- Per-layer rates unappealing for following reasons:
  1. More hyperparameters to tune
  2. Inconvenient in neuromorphic hardware

Motivation: vanishing gradient problem

- Problem: vanishing gradients in layered networks
  - Leads to slow training
  - Potential bit-depth issues
  - State-of-the-art performance by deep networks
- Not yet solved in simple, biologically-plausible manner
- Original paper: manually tune independent learning rate for each layer
- Per-layer rates unappealing for following reasons:
  1. More hyperparameters to tune
  2. Inconvenient in neuromorphic hardware
  3. Seems unlikely in biological systems

Introduction

- Our solution: modified topology including layer-skipping connections

Introduction

- Our solution: modified topology including layer-skipping connections
- Addresses above issues with per-layer rates

Introduction

- Our solution: modified topology including layer-skipping connections
- Addresses above issues with per-layer rates
    1. Two new hyperparameters; constant with depth

Introduction

- Our solution: modified topology including layer-skipping connections
- Addresses above issues with per-layer rates
  1. Two new hyperparameters; constant with depth
  2. Inspired by small-world topology - observed in biological brains

Introduction

- Our solution: modified topology including layer-skipping connections
- Addresses above issues with per-layer rates
  1. Two new hyperparameters; constant with depth
  2. Inspired by small-world topology - observed in biological brains
  3. Easy to implement in networks with configurable connectivity

Introduction

- Our solution: modified topology including layer-skipping connections
- Addresses above issues with per-layer rates
  1. Two new hyperparameters; constant with depth
  2. Inspired by small-world topology - observed in biological brains
  3. Easy to implement in networks with configurable connectivity
- Improves network training speed

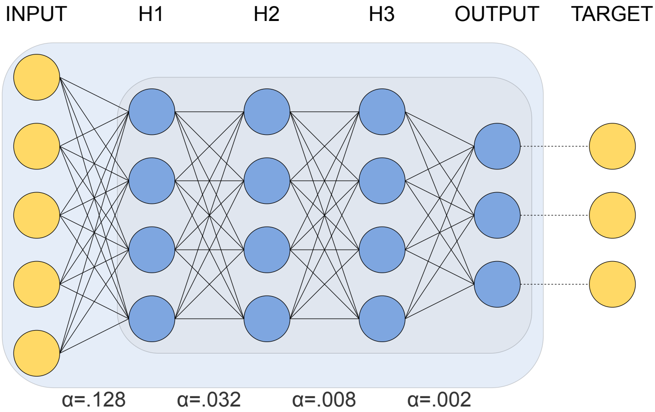
Introduction

- Our solution: modified topology including layer-skipping connections
- Addresses above issues with per-layer rates
  1. Two new hyperparameters; constant with depth
  2. Inspired by small-world topology - observed in biological brains
  3. Easy to implement in networks with configurable connectivity
- Improves network training speed
- Increases uniformity with which each layer trains

Introduction

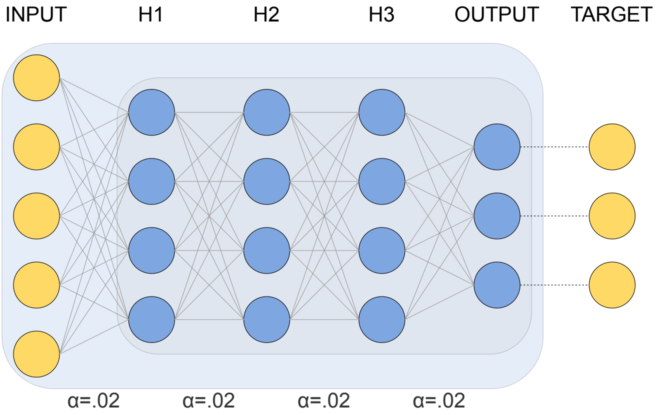
- Our solution: modified topology including layer-skipping connections
- Addresses above issues with per-layer rates
  1. Two new hyperparameters; constant with depth
  2. Inspired by small-world topology - observed in biological brains
  3. Easy to implement in networks with configurable connectivity
- Improves network training speed
- Increases uniformity with which each layer trains
- Performance similar to that of per-layer rates

Original topology



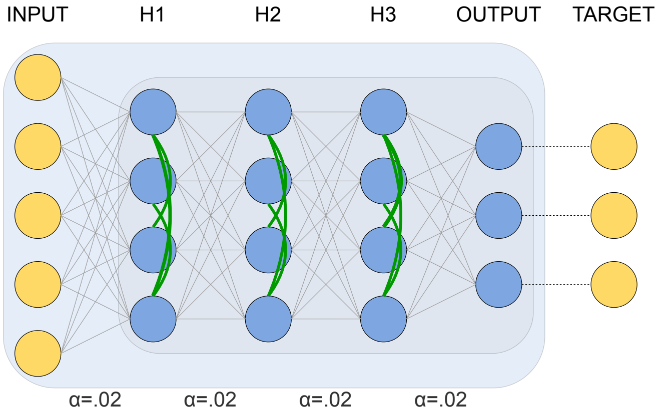
- From original paper
- Per-layer learning rates

Our topological modifications



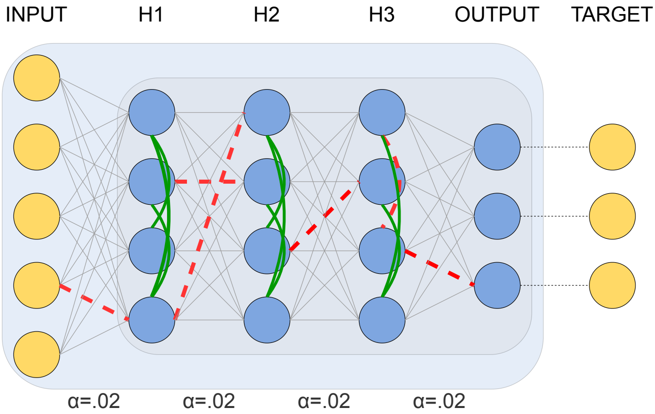
- Starting point: original topology
- One learning rate for all layers

Our topological modifications



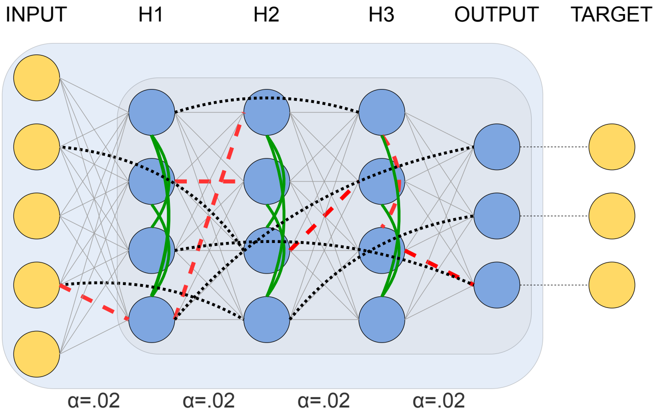
- Hidden layers fully-connected

Our topological modifications



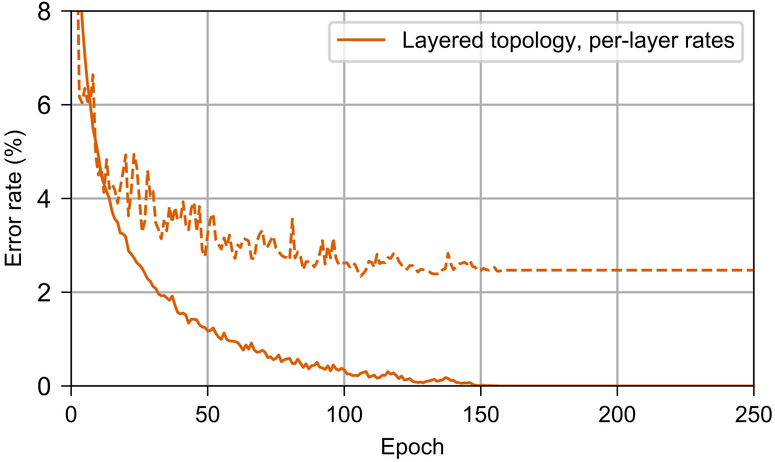
- Consider each connection
- Remove with probability p

Our topological modifications

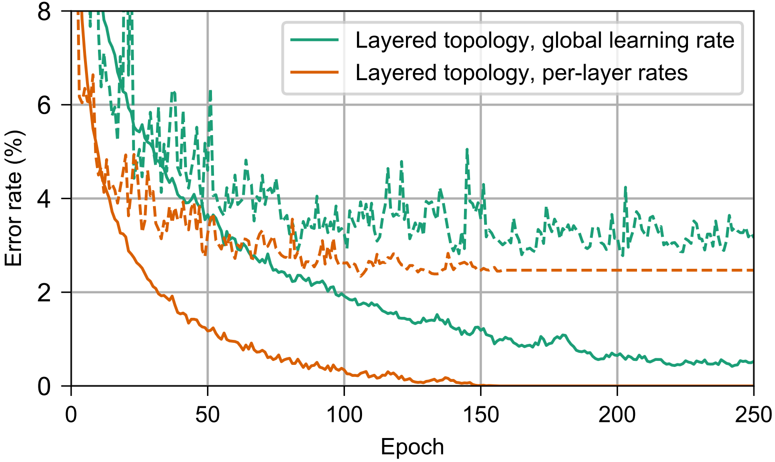


- For each removed connection, randomly connect different pair
- No connections in input or output layers

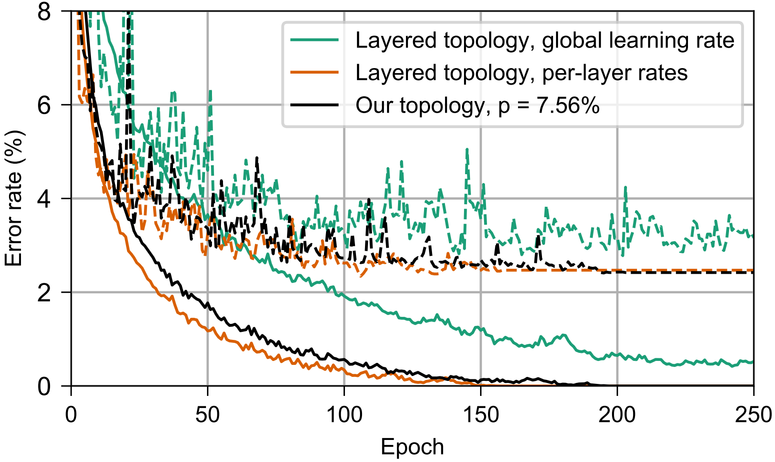
Results: performance of network with layer-skipping connections



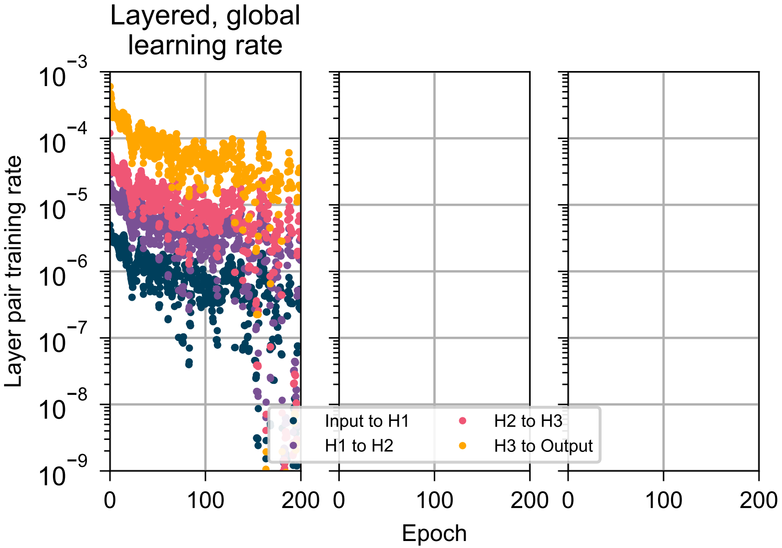
Results: performance of network with layer-skipping connections



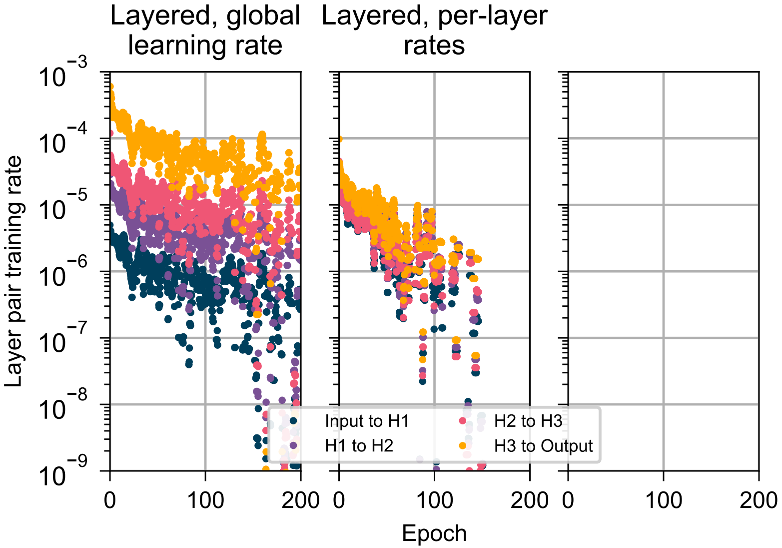
Results: performance of network with layer-skipping connections



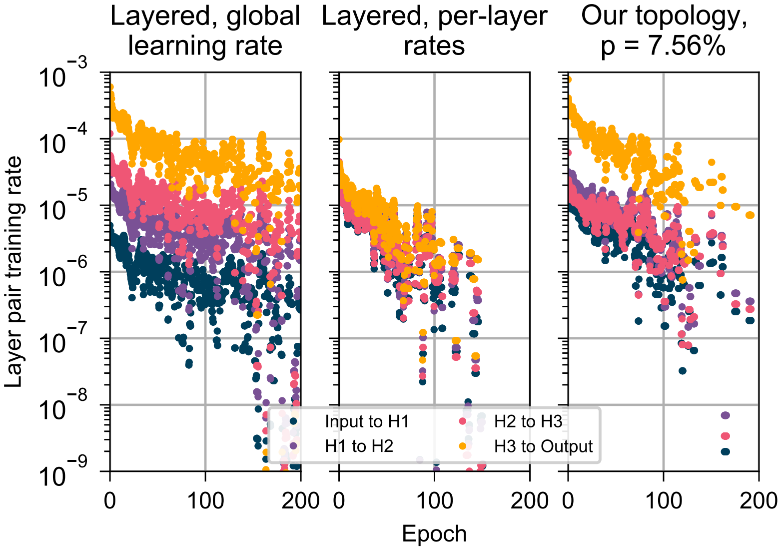
Results: effect on training rates of layers



Results: effect on training rates of layers



Results: effect on training rates of layers



Results: takeaways

- Our topology increases training speed

Results: takeaways

- Our topology increases training speed
  - Much faster than original topology, one learning rate

Results: takeaways

- Our topology increases training speed
    - Much faster than original topology, one learning rate
    - Slightly slower than per-layer rates

Results: takeaways

- Our topology increases training speed
  - Much faster than original topology, one learning rate
  - Slightly slower than per-layer rates
- Layers train in more-uniform manner

Results: takeaways

- Our topology increases training speed
  - Much faster than original topology, one learning rate
  - Slightly slower than per-layer rates
- Layers train in more-uniform manner
  - Output layer faster - no added connections to target layer

Results: takeaways

- Our topology increases training speed
  - Much faster than original topology, one learning rate
  - Slightly slower than per-layer rates
- Layers train in more-uniform manner
  - Output layer faster - no added connections to target layer
- Good solution where simplicity, biological plausibility are important

Directions for future research

- Try on harder datasets (e.g. CIFAR, ImageNet) where depth is very important

Directions for future research

- Try on harder datasets (e.g. CIFAR, ImageNet) where depth is very important
- Effect of p on test error

Directions for future research

- Try on harder datasets (e.g. CIFAR, ImageNet) where depth is very important
- Effect of p on test error
- Effectiveness on deeper networks

Directions for future research

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