Loss of Plasticity in Deep Continual Learning

Aksenenko Veronika

Definitions

Train-once setting – training occurs once on a large dataset and then never again

Continual learning – continually learns from new data

Problems

Catastrophic forgetting – deep learning networks, when exposed to new data, tend to forget most of what they have previously learned

Plasticity loss – ability to keep learning from a new data

Loss of Plasticity in ImageNet

ImageNet: 1000 classes (700 images for each class)

Continuall ImageNet:

1) 2000 tasks of binary classification
2) 700 images for class were divided by 600 (for train) and 100 (for test)

ImageNet Results

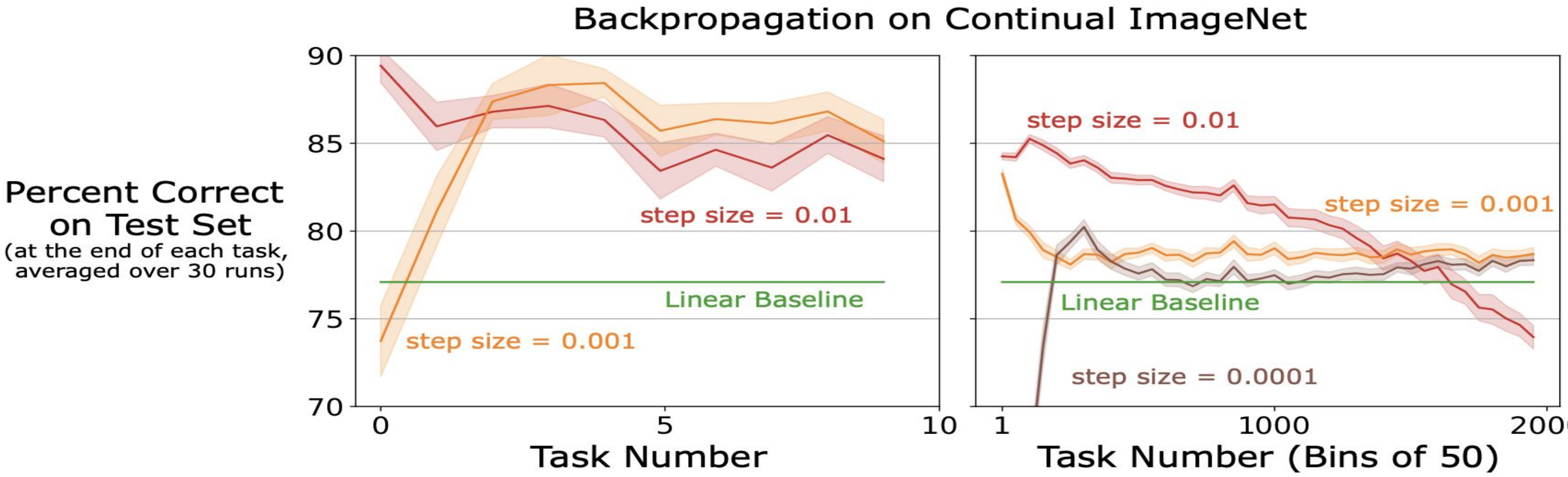


Figure 1: **Loss of plasticity on a sequence of ImageNet binary classification tasks.** The first plot shows performance over the first ten tasks, which sometimes improved initially before declining. The second plot shows performance over 2000 tasks, over which the loss of plasticity was extensive. The learning algorithm was backpropagation applied in the conventional deep-learning way.

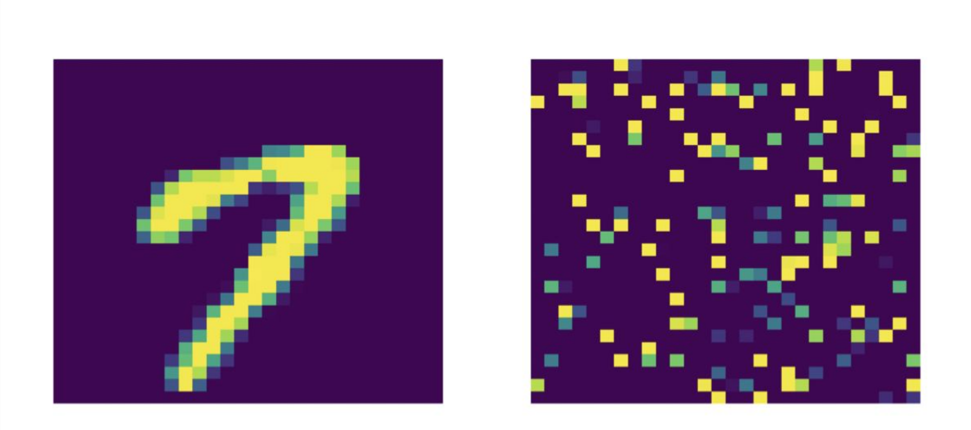
Loss of Plasticity in MNIST

MNIST: 60.000 images of handwritten digits from 0 to 9

Permuted MNIST: Permuting pixels with chosen permutation

Online Permuted MNIST:

Created 800 permuted MNIST



MNIST Results

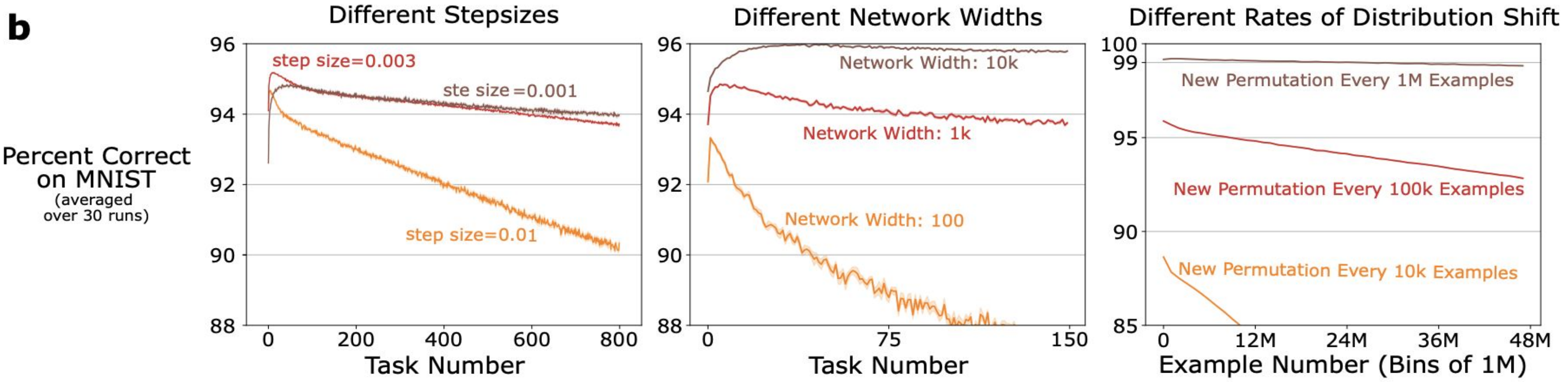
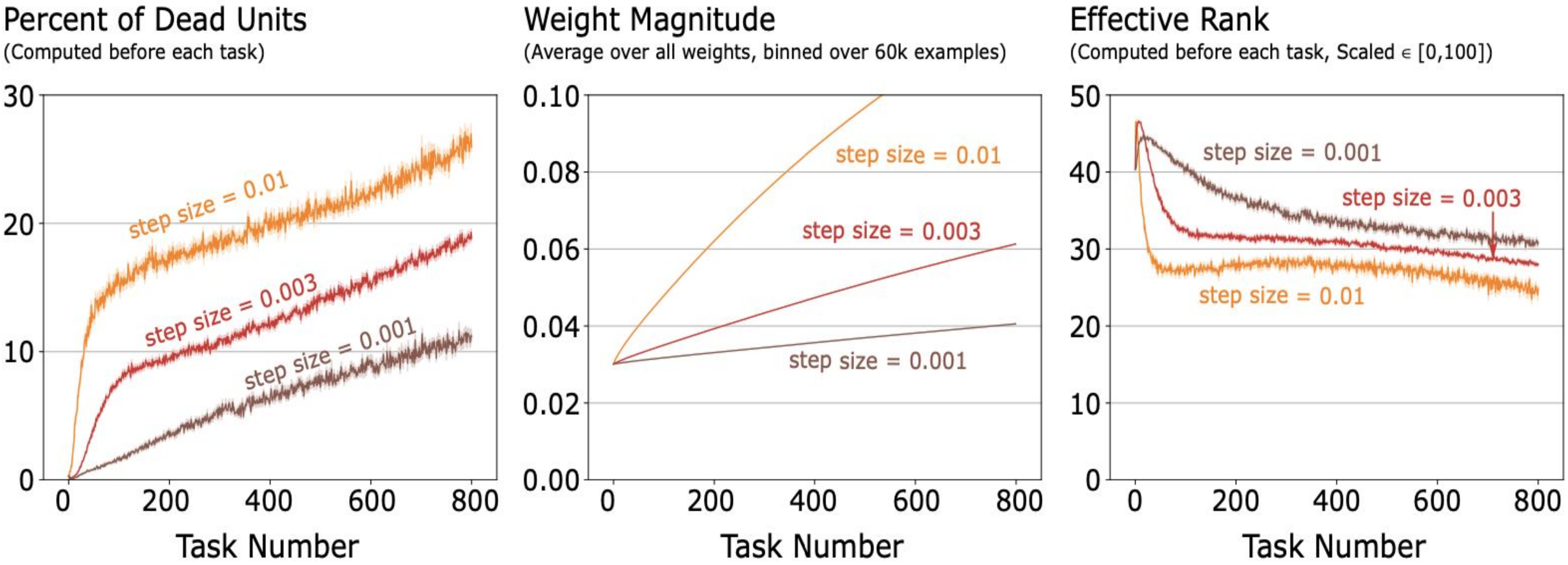


Figure 2: **a:** Left: An MNIST image with the label '7'; Right: A corresponding permuted image. **b:** Loss of plasticity in Online Permuted MNIST is robust over step sizes, network sizes, and rates of change.

Results

- Loss of plasticity is a general phenomenon, and it can be catastrophic in some cases
- Backpropagation does not work

Reasons of Plasticity Loss



Methods for Mitigation Loss Of Plasticity

L2-regularization – decrease weight magnitudes

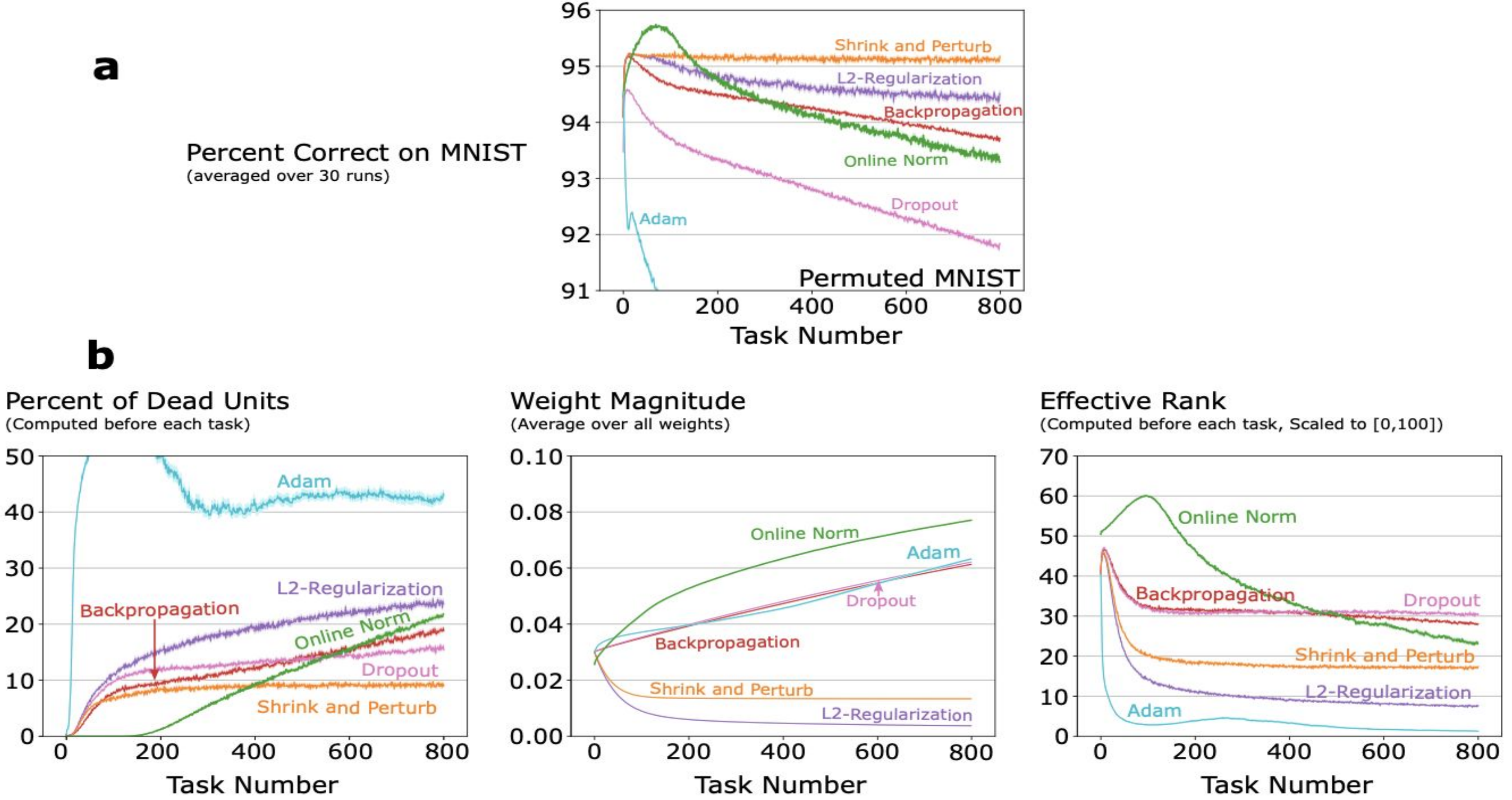
Dropout – increase in the effective rank

Batch normalization – decrease number of dead units

Shrink-and-perturb – decrease weight magnitudes, number of dead units and increase the effective rank

Adam

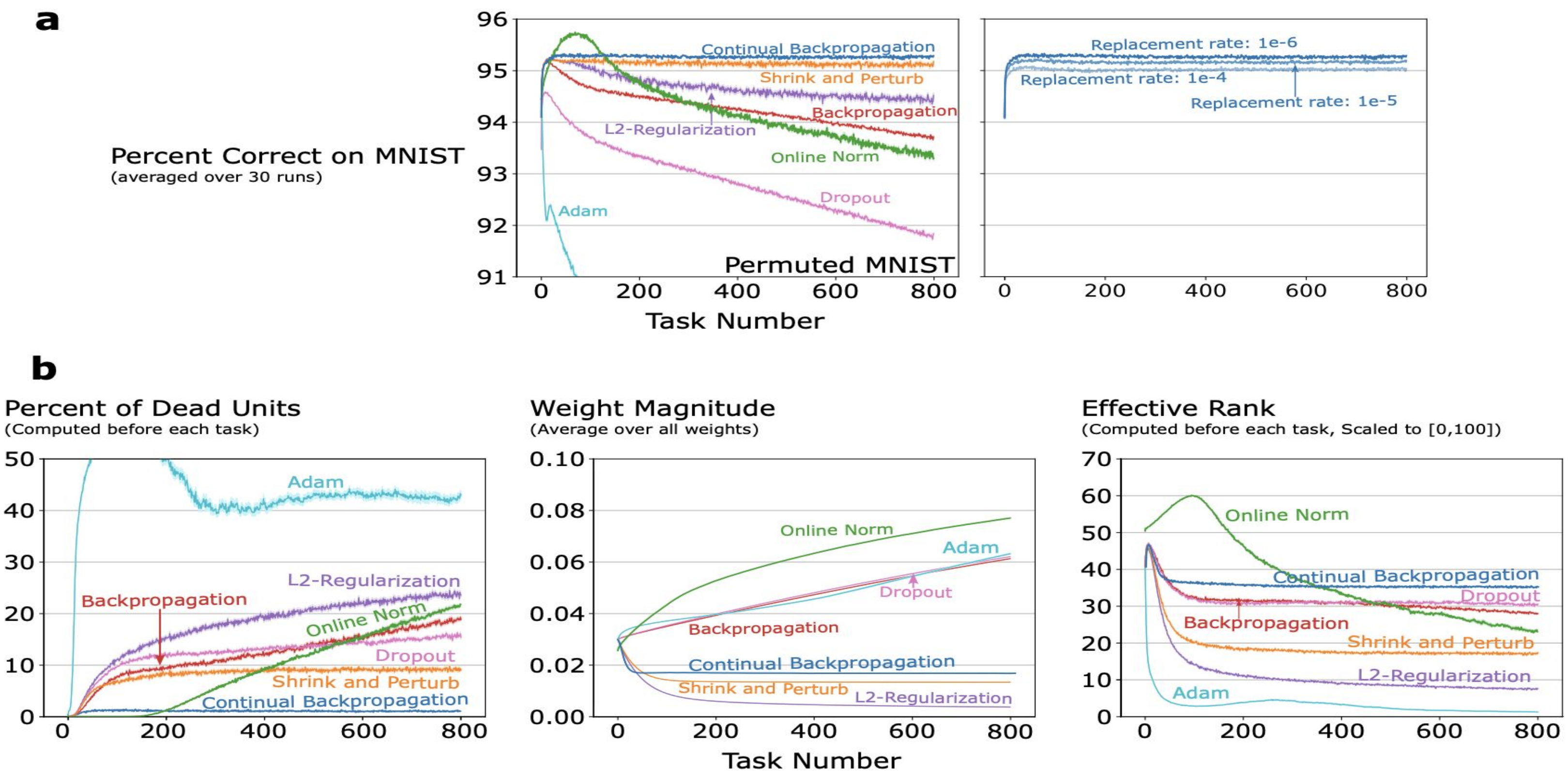
Results


a
Percent Correct on MNIST
(averaged over 30 runs)
Shrink and Perturb
L2-Regularization
Backpropagation
Online Norm
Dropout
Adam
Permuted MNIST
Task Number
b
Percent of Dead Units
(Computed before each task)
Adam
L2-Regularization
Backpropagation
Online Norm
Dropout
Shrink and Perturb
Task Number
Weight Magnitude
(Average over all weights)
Online Norm
Adam
Dropout
Backpropagation
Shrink and Perturb
L2-Regularization
Task Number
Effective Rank
(Computed before each task, Scaled to [0,100])
Online Norm
Backpropagation
Dropout
Shrink and Perturb
L2-Regularization
Adam
Task Number

Continual Backpropagation

1) Extension of conventional backpropagation
2) Selectively reinitialize low-utility units
    a) Find low-utility units
    b) Reinitialize them

Results



a
Percent Correct on MNIST
(averaged over 30 runs)
96
95
94
93
92
91
Continual Backpropagation
Shrink and Perturb
Backpropagation
L2-Regularization
Online Norm
Dropout
Adam
Permuted MNIST
0
200
400
600
800
Task Number
Replacement rate: 1e-6
Replacement rate: 1e-4
Replacement rate: 1e-5
b
Percent of Dead Units
(Computed before each task)
50
40
30
20
10
0
Adam
L2-Regularization
Backpropagation
Online Norm
Dropout
Shrink and Perturb
Continual Backpropagation
Weight Magnitude
(Average over all weights)
0.10
0.08
0.06
0.04
0.02
0.00
Online Norm
Adam
Dropout
Backpropagation
Continual Backpropagation
Shrink and Perturb
L2-Regularization
Task Number
Effective Rank
(Computed before each task, Scaled to [0,100])
70
60
50
40
30
20
10
0
Online Norm
Continual Backpropagation
Dropout
Backpropagation
Shrink and Perturb
L2-Regularization
Adam

Summary

- Demonstrated loss of plasticity in deep continual learning
- Proposed potential causes and suggested methods to correct
- Introduces a new algorithm which solves the problem