Investigating Learning in Deep Neural Networks using Layer-Wise Weight Change

Ayush M. Agrawal[*, 1, 6, 7], Atharva A. Tendle[*, 2, 6, 7], Harshvardhan D. Sikka[3, 6, 7], Sahib Singh [4, 6, 7], Amr Kayid[5, 6, 7]
*Equal Contribution, [1]University of Nebraska, [2]University of Nebraska-Lincoln, [3]Georgia Institute of Technology, [4]Ford R&A, [5]German University in Cairo, [6]Manifold Computing Group, [7]Openmined

Focus

We attempt to understand the per-layer learning dynamics of deep neural networks as this might provide useful insights into how neural networks learn.

We believe that understanding these trends can help researchers evaluate higher level trends which can potentially yield efficient training regimes and better performance.

Preliminaries

Relative Weight Change:

We utilize this metric to understand layer-wise weight change. Intuitively this is the average of the absolute value of the percentage change in the magnitude of a given layer's weight.

$$RWC_L = \frac{\|w_t - w_{t-1}\|_1}{\|w_{t-1}\|_1}$$

L represents a single layer in the neural network.

$\mathbf{w_t}$ represents the vector of weights associated with **L** at a given training step t.

We use the $\mathbf{L_1}$ **norm** to characterize the difference in magnitude of the weights and normalize the difference by dividing by the magnitude of the layer's weights during the previous training step.

Datasets & Architectures

We use 4 benchmark datasets and 3 architectures for our experiments:

- MNIST
- Fashion MNIST
- CIFAR-10
- CIFAR-100
- Resnet18
- VGG-19_bn
- Alexnet

Experiments

We attempt to make our experiments reproducible and comparable by using a standard set of hyperparameters.

All experiments are averaged over 5 seed values to account for variance.

Table 1. Detailed hyperparameters used for training

Architecture	Datasets	LR	Momentum	Weight Decay
ResNet18	CIFAR-10	0.1	0.9	0.0001
	CIFAR-100	0.1	0.9	0.0001
	MNIST	0.1	0.9	0.0001
	FMNIST	0.1	0.9	0.0001
VGG19_bn	CIFAR-10	0.05	0.9	0.0005
	CIFAR-100	0.05	0.9	0.0005
	MNIST	0.05	0.9	0.0005
	FMNIST	0.05	0.9	0.0005
AlexNet	CIFAR-10	0.001	0.9	0.0001
	CIFAR-100	0.01	0.9	0.0001
	MNIST	0.1	0.9	0.0001
	FMNIST	0.1	0.9	0.0001

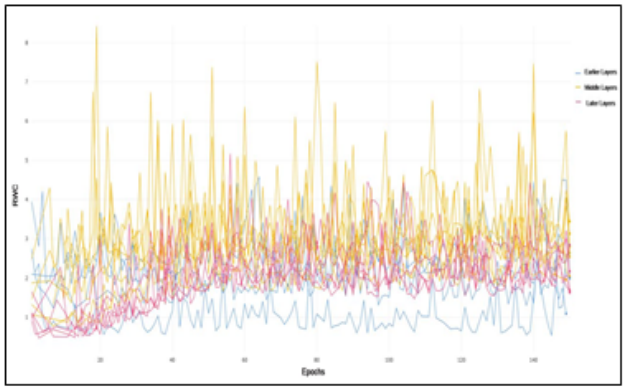
Results

Using RWC as a metric, we noticed that in general, later layers have a higher relative weight change in comparison to the earlier layers through training.

The earlier layers tend to stop learning after a point.

We also noticed a relationship between task complexity and the relative weight change that occurs through the layers.

This difference was evident between the experiments conducted on CIFAR-10 and CIFAR-100 since the latter is essentially a more complex version of the prior.



Correspondence: aagrawal@nebraska.edu
Full paper: https://arxiv.org/pdf/2011.06735.pdf