A Forward-backward Learning strategy for CNNs via Separation Index Maximizing at the First Convolutional Layer

Ali Karimi [1] Ahmad Kalhor [1] Mona Ahmadian [2]

[1]University of Tehran [2]University of Surrey

BMVC 2023

Introduction

Although error backpropagation algorithms can achieve very high accuracy in image classification problems, they are constrained by the vanishing gradient issue. It is not possible for the first layer to be properly learned. The vanishing gradient occurs more when we have more layers and use deeper neural networks. With the development of deeper networks, this problem will become more visible. Due to this issue, a variety of approaches have been proposed, including activation functions like Rectified linear unit (ReLU) [4] that avoid saturation in the early layers, use normalization techniques such as batch normalization or layer normalization to reduce the dependence of the gradients on the scale of the activation values [3]. But the issue still exists.

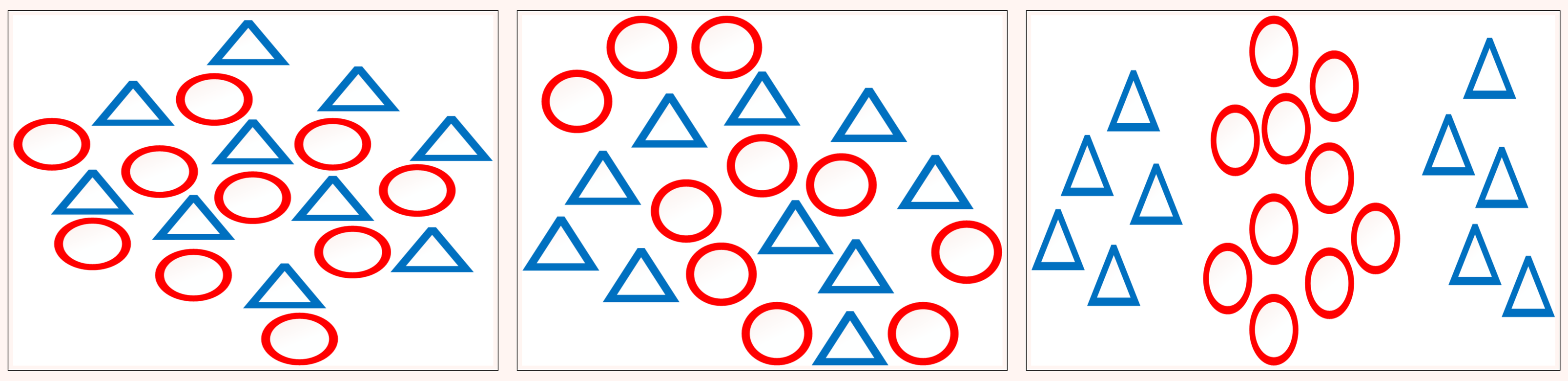
It is possible to solve the vanshing gradient problem using forward learning, but error backpropagation is still more accurate and powerful [2] . We have proposed a method that combines forward learning and error backpropagation with a slight modification to improve accuracy.

Related Works

- A greedy layer-wise learning approach was introduced by Bengio et al [1] . The main purpose of greedy layer-wise pretraining is to initialize the weights of a deep neural network layer by layer, beginning with the first layer. In each layer, a separate autoencoder is trained, using the previous layer's output as an input. After each layer has been trained, the entire network is fine-tuned using backpropagation.
- Hinton [2] has introduced a novel learning strategy for neural networks named the Forward-Forward Algorithm. By using this algorithm, forward and backward backpropagation passes are replaced by two forward passe.

Separation Index (SI)

Separation Index (SI) [5] indicates how much the data points with different labels are separated from each other. In addition, it has been explained that while the SI increases layer by layer in a deep neural network (DNN), the margin among different classes will increase and the generalization of the network.



(a) (b) (c)

Figure 1. An illustrative example where data points of two classes (with the circle and triangle indicators) have (a) low ($SI \approx 0$) , (b) medium ($SI \approx 0.5$), or (c) high ($SI \approx 1$) separation indices.

Method

Our novel learning strategy is developed in two phases:

- At the first phase, as the forward learning part, After a batch normalization layer is applied to the data, the patches from the input data are extracted and principal component analysis (PCA) is carried out on the data based on these patches. The eigenvalues derived by PCA in the convolution layer are used as the values for the filter in the convolution layer. The output of the convolution layer is then received, and QLS is performed on them to calculate the updated values of the convolution layer filters. This continues until we reach maximum classification accuracy in the first layer.
- At the second phase, as the backward learning part, the output of the first layer is given as input to the second layer and the network is trained by an error backprogration algorithm.

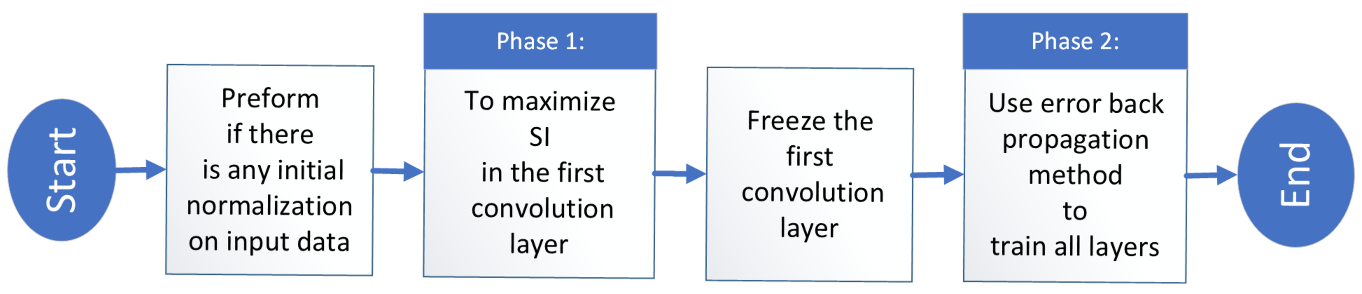


Figure 2. This flowchart indicates the learning method in which the quasi-LS is utilized to learn the first convolution layer (phase 1), and other layers are learned by an error backpropagation algorithm (phase 2).

Experiments

In all architectures and datasets the accuracy is increased by our proposed learning strategy. This improvement is more significant in datasets with fewer classes, as demonstrated in Table 1, The CIFAR10 dataset has the same number of features as the CIFAR100 dataset, but the CIFAR10 dataset has fewer classes. The accuracy of our learning strategy in CIFAR10 shows a greater improvement ratio to backpropagation than in CIFAR100. Furthermore, the improvement ratio in datasets with more features is much higher.

Dataset	Learning Method	AlexNet	VGG16	ResNet50	InceptionV3
CIFAR10	Backpropagation algorithm	84.61	92.95	93.17	94.20
	Our proposed learning strategy	**85.84**	**94.81**	**95.02**	**95.43**
	Percentage of improvement	1.23	1.86	1.85	1.23
CIFAR10 - (plane,truck)	Backpropagation algorithm	96.45	97.18	97.64	97.09
	Our proposed learning strategy	**97.09**	**98.36**	**98.49**	**97.77**
	Percentage of improvement	0.64	1.18	0.85	0.68
CIFAR10 - (plane,cat,bird)	Backpropagation algorithm	96.21	96.80	96.88	96.81
	Our proposed learning strategy	**96.62**	**97.63**	**97.16**	**97.14**
	Percentage of improvement	0.41	0.83	0.28	0.33
CIFAR100	Backpropagation algorithm	62.22	70.98	75.30	76.31
	Our proposed learning strategy	**62.45**	**71.74**	**75.58**	**76.72**
	Percentage of improvement	0.23	0.76	0.28	0.41
Fashion-MNIST	Backpropagation algorithm	92.53	94.17	95.24	95.78
	Our proposed learning strategy	**92.65**	**94.73**	**95.38**	**95.91**
	Percentage of improvement	0.12	0.56	0.14	0.13

Table 1. Comparing the accuracy of our learning strategy in different architectures and datasets with error backpropagation algorithm

The Impact of Number of Layers In Forward Learning

We have applied the forward learning method to further layers and Table 2 summarizes the accuracy results. As it is seen, the best accuracy is for when we apply our forward learning method to only the first layer.

No. of Forward layers	Run Time(m:s)	ACC
0 (Backpropagation Method)	144 : 16	92.95
1 (Our Method)	127 : 44	94.81
2	131 : 18	94.76
3	125 : 45	94.32
4	124 : 58	93.64
5	120 : 06	93.73
6	114 : 05	93.66
...	...	...
11	87 : 03	93.29
12	85 : 51	93.25
13	81 : 14	93.13

Table 2. Comparison of Time Complexity and Accuracy of the Proposed Method in Different Layers with Backpropagation Learning Strategy for VGG16 on CIFAR10.

Conclusion

The comparison results clearly demonstrated our learning strategy improved the learning of CNNs in all cases. For example in the classification of CIFAR10 images, the accuracy of VGG and ResNet50 increased near two percentages compared to the error backpropagation. Furthermore, in comparison with greedy layerwise learning, the proposed method was superior.

References

[1] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

[2] Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[4] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[5] Mohsen Saffar and Ahmad Kalhor. Evaluation of dataflow through layers of convolutional neural networks in classification problems. *Expert Systems with Applications*, 224:119944, 2023.

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.