Revisiting the Training of Very Deep Neural Networks without Skip Connections

Oyebade K. Oyedotun*, Abd El Rahman Shabayek *, Djamila Aouada *, Björn Ottersten *

* Interdisciplinary Centre for Security, Reliability and Trust - University of Luxembourg, L-1855 Luxembourg

Abstract

We investigate two scenarios that plague the training of very deep PlainNets (models without skip connections): (1) the relatively popular challenge of 'vanishing and exploding units' activations', and (2) the less investigated 'singularity' problem, which is studied in the literature. In contrast to earlier works that study only the saturation and explosion of units' activations in isolation, this paper harmonizes the inconspicuous coexistence of the aforementioned problems for very deep PlainNets. We argue that the aforementioned problems would have to be tackled simultaneously for the successful training of very deep PlainNets. Finally, different techniques that can be employed for tackling the optimization problem are discussed.

Introduction

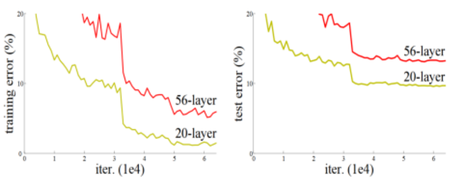
- Very deep models → Deep Neural Networks (DNNs) with over 15 layers.
- Generalization performance of DNNs generally increase with depth increase.

Motivation

- Simple architecture, since there is one information path from the input to the output of the model.
- Hierarchical representations are easier to interpret.

Problem Statement

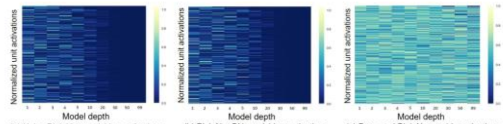
- Very deep PlainNets are difficult to optimize [1, 2, 3].



10 dataset [1]

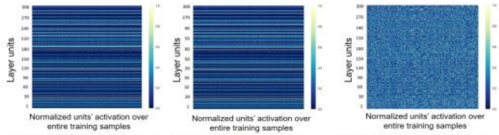
Proposed investigation

- **Units' activation evolution**: Units' activation stability



(a) Naive PlainNet: vanishing activations (b) PlainNet-BN: vanishing activations (c) Proposed PlainNet: stable activations

- **Singularity**: Hidden representation condition

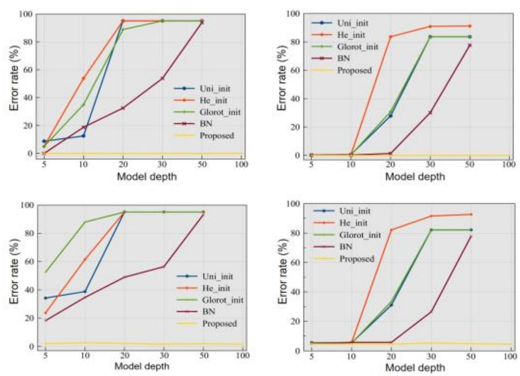


(a) Naive PlainNet: near singularity (b) PlainNet-BN: near singularity (c) Proposed PlainNet: no singularity

Alleviating the training problem of very deep PlainNet

- **Batch normalization (BN)**
- **Leaky rectified linear units (LReLU)**
- **Max-norm constraint for model weights**

❑ Proposed solution → BN + LReLU + Maxnorm



(a) COIL-20 (b) USPS

Experiments

Table 1: Ablation studies for the different components of the proposed solution.

Model component	Train error	Test error
Batch normalization (BN)	84.56%	83.21%
LReLU	92.37%	92.03%
Max-norm	86.22%	86.85%
BN + LReLU	78.38%	79.52%
BN + max-norm	82.90%	81.86%
LReLU + max-norm	83.62%	82.11%
Proposed: BN + LReLU + max-norm	**0.11%**	**5.48%**

Table 2: Results on CIFAR-10 dataset

Model	Skip conn.	Layers	Parameters	Test error
Highway network [2]	Yes	19	2.30M	7.54%
ResNet [3]	Yes	56	0.85M	6.97%
ResNet [3]	Yes	110	1.7M	6.43%
All CNN [30]	No	8	1.30M	7.25%
NiN [31]	No	10	1.30M	8.81%
Delta init. [15]	No	32	17.80M	18.00%
PlainNet-BN [3]	No	56	0.85M	15.00%
Proposed PlainNet	**No**	**50**	**0.72M**	**6.65%**

Conclusion

It is common to observe poor generalization when the depth of DNNs without skip connections (i.e. very deep PlainNets) exceeds 15 layers. In this paper, our investigation results reveal that the successful training of very deep PlainNets would rely on simultaneously alleviating vanishing/exploding units' activations and singularity of units' activations. Lastly, we demonstrate an approach for alleviating the training problems.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR (pp. 770-778).

[2] Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. NIPS (pp. 2377-2385).

[3] Oyedotun, O. K., Aouada, D., & Ottersten, B. (2017, November). Training very deep networks via residual learning with stochastic input shortcut connections. ICONIP (pp. 23-33).

This work was funded by:

The National Research Fund (**FNR**), Luxembourg, under the project reference R-AGR- 0424-05-D/Bjorn Ottersten.

Fonds National de la Recherche Luxembourg