# All You Need is Beyond a Good Init: Exploring Better Solution for Training Extremely Deep Convolutional Neural Networks with Orthonormality and Modulation

Di Xie (xiedi@hikvision.com), Jiang Xiong and Shiliang Pu
Hikvision Research Institute

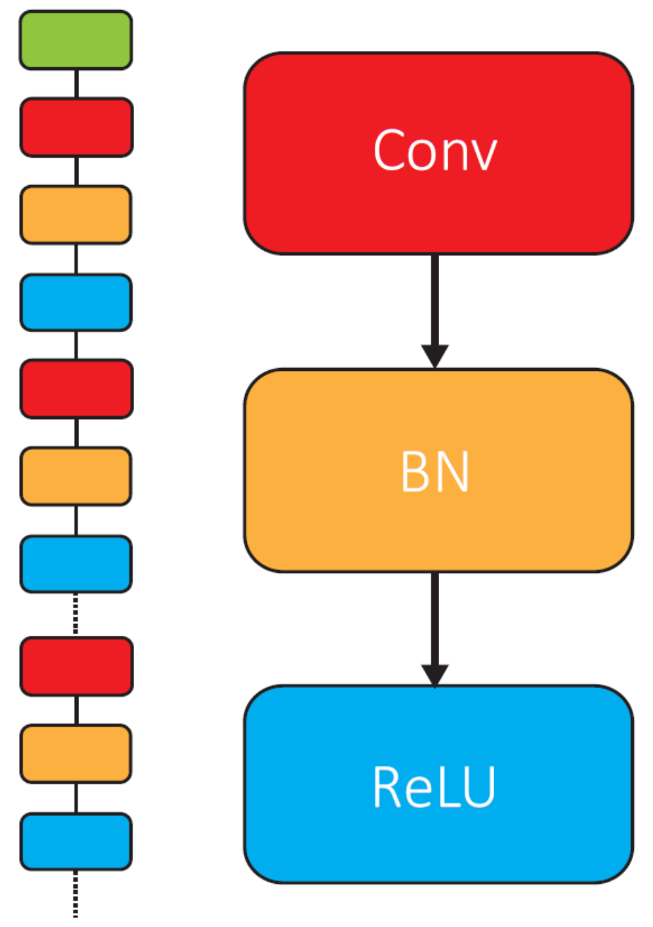
IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

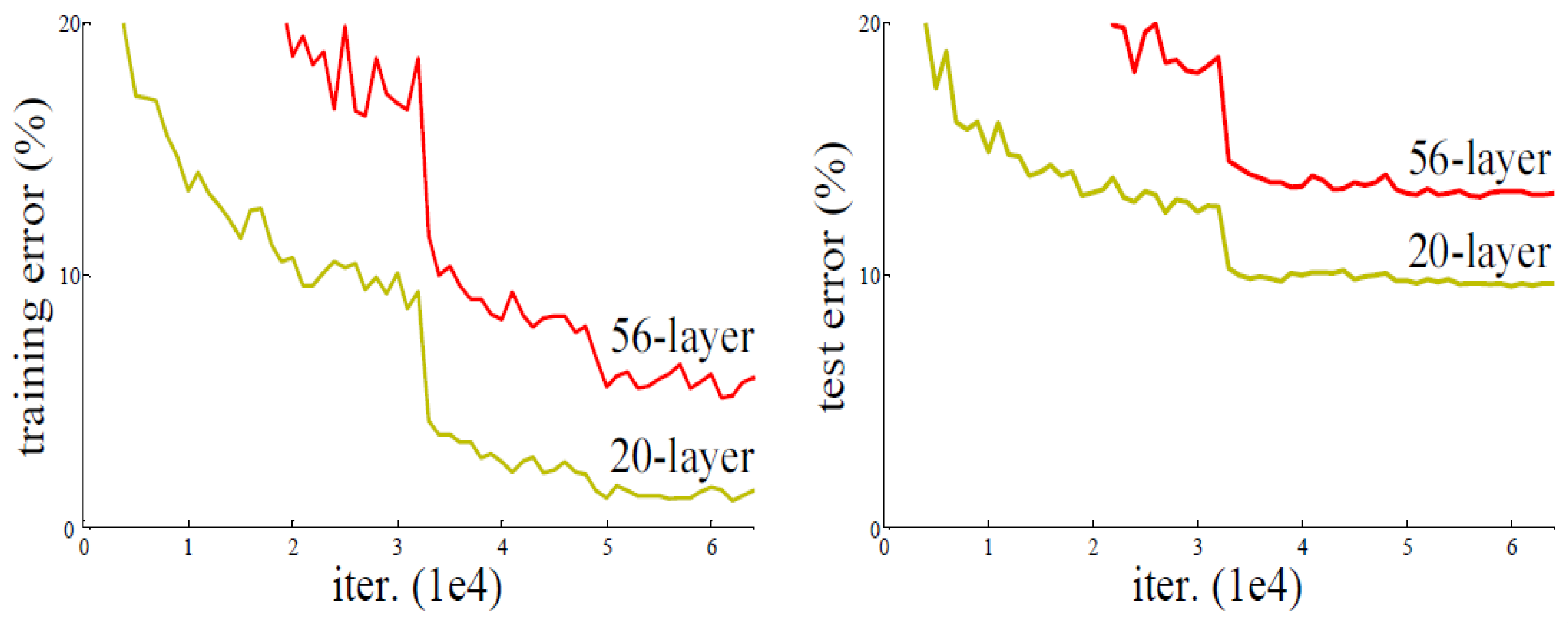
## Abstract

Deep neural network is difficult to train and this predicament becomes worse as the depth increases. The essence of this problem exists in the magnitude of backpropagated errors that will result in gradient vanishing or exploding phenomenon. We show that a variant of regularizer which utilizes orthonormality among different filter banks can alleviate this problem. Moreover, we design a backward error modulation mechanism based on the quasi-isometry assumption between two consecutive parametric layers. Equipped with these two ingredients, we propose several novel optimization solutions that can be utilized for training a specific-structured (repetitively triple modules of Conv-BN-ReLU) extremely deep convolutional neural network (CNN) WITHOUT any shortcuts/ identity mappings from scratch. Experiments show that our proposed solutions can achieve distinct improvements for a 44-layer and a 110-layer plain networks on both the CIFAR-10 and ImageNet datasets. Moreover, we can successfully train plain CNNs to match the performance of the residual counterparts.

Besides, we propose new principles for designing network structure from the insights evoked by orthonormality. Combined with residual structure, we achieve comparative performance on the ImageNet dataset.

## Motivation





The performance of plain neural networks degrades as its depth increases beyond a certain layer numbers (usually 18~20 layers). Figure is referred from [1]. What are the potential factors and how to overcome this problem?

## Problems Reside in Backprop

1. Batch Normalization (BN) perfectly stabilizes forward signal but biases the distribution of backward signal a bit after one pseudo-normalization transformation:

$$\frac{\partial\,\ell}{\partial\,x_i}=\frac{1}{\sqrt{\sigma_B^2+\epsilon}}(\delta_i-\mu_\delta-\frac{\hat{x}_i}{m}\sum_{j=1}^{m}\delta_j\hat{x}_j)$$

2. The Jacobian matrix of BN is rank-deficient that violate the perfect dynamic isometry:

$$\mathbf{J}=\mathbf{P}^T\rho\begin{bmatrix}1-\frac{\lambda_1}{m} & 0 & 0 & \cdots & 0\\ 0 & 1-\frac{\lambda_2}{m} & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix}_{m\times m}\mathbf{P}$$

The entries of Jacobian is $\frac{\partial y_j}{\partial x_i}=\rho\left[\Delta(i=j)-\frac{1+\hat{x}_i\hat{x}_j}{m}\right]$ and $U_{ij}=1+\hat{x}_i\hat{x}_j$, $\rho=\frac{\gamma}{\sqrt{\sigma_B^2+\epsilon}}$

$\lambda_1$ and $\lambda_2$ are eigenvalues of $\mathbf{U}$ and $1-\frac{\lambda_1}{m}\approx 0, 1-\frac{\lambda_2}{m}\approx 0$ given that sufficient batch size.

## Solutions

1. Introducing orthonormality (norm-preserving) constraints of weights to alleviate vanishing or exploding phenomenon and provides more probabilities by limiting set of parameters in an orthogonal space instead of inside a hypersphere:

$$\frac{\lambda}{2}\sum_{i=1}^{D}\|\mathbf{W}_l^T\mathbf{W}_l-\mathbf{I}\|_F^2$$

2. Modulation with a layer-wise mechanism dynamically (if necessary) since $\mathbf{J}\mathbf{J}^T\approx\rho^2\mathbf{I}$

* Recently advance found by [2] supports our work convincingly, which shows that orthogonality exists in primate's brain!

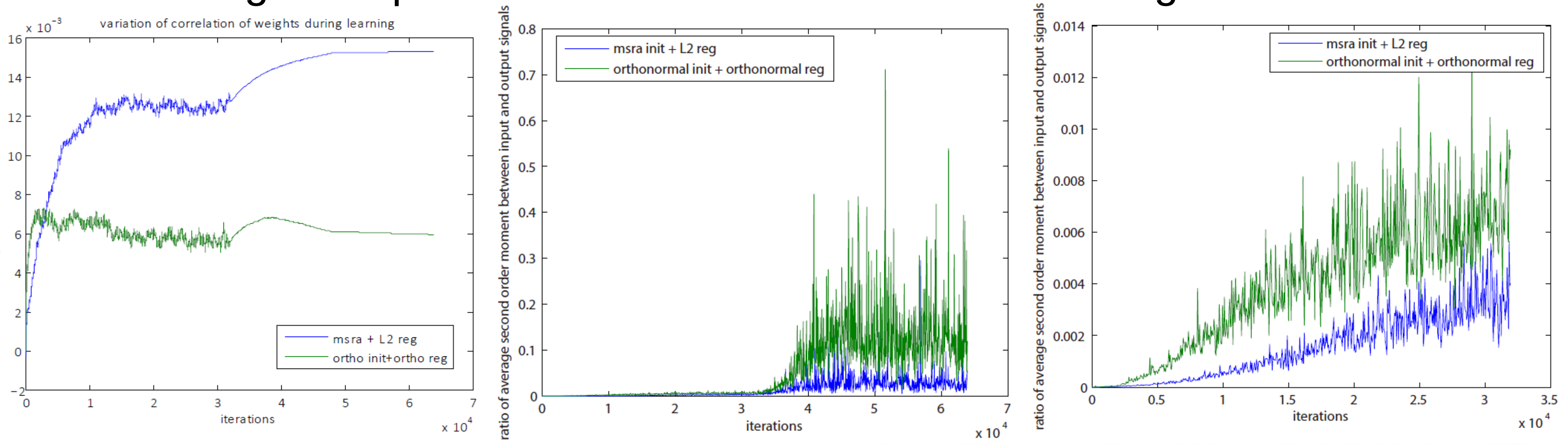
## References

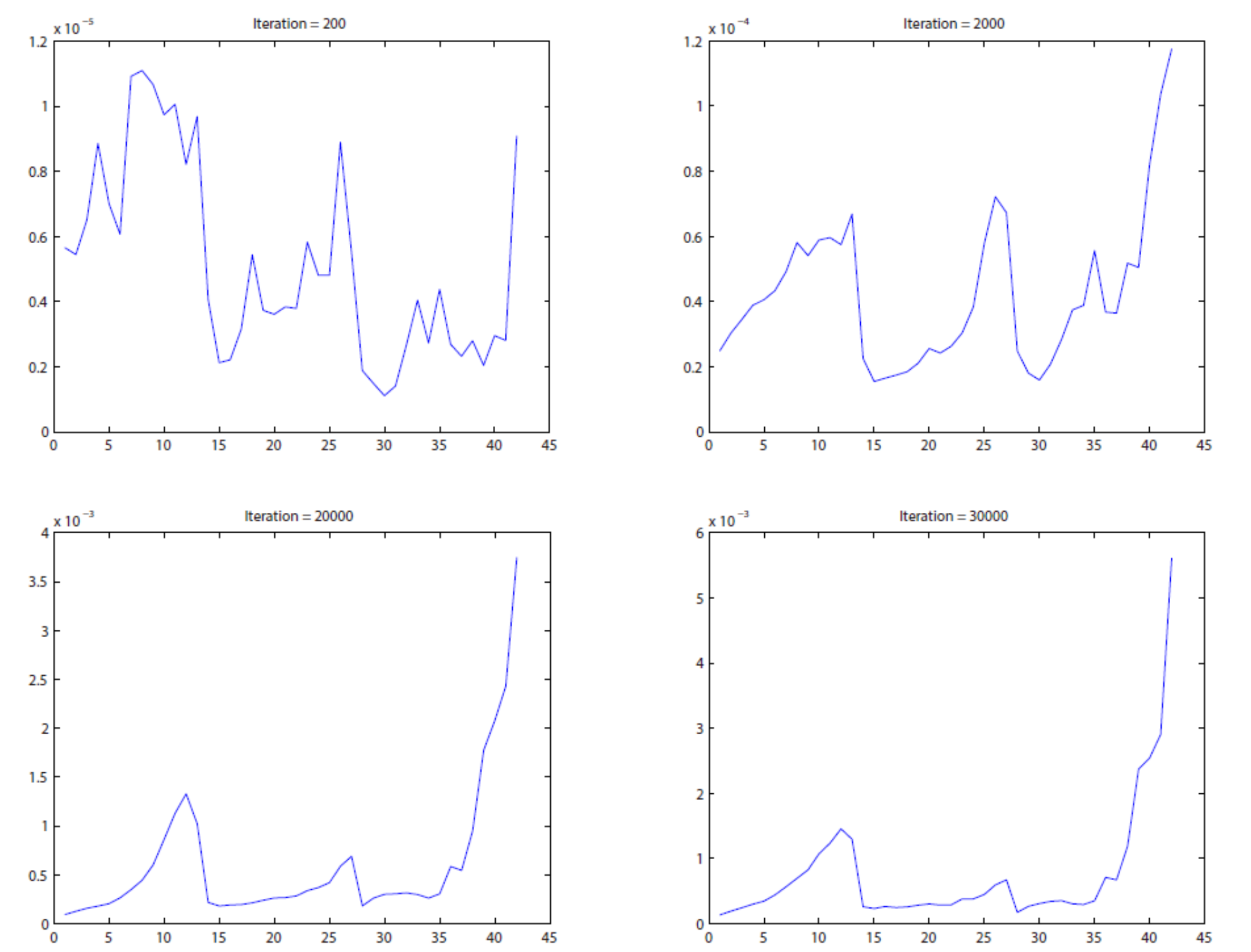
[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv:1512.03385, 2015.

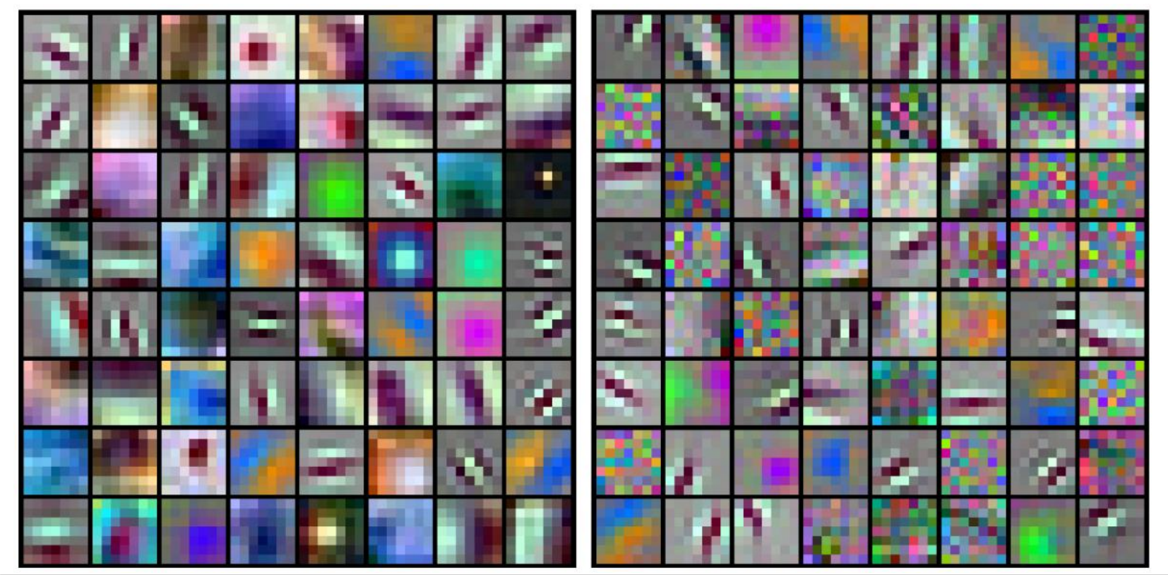
[2] Le C, Tsao D Y. The Code for Facial Identity in the Primate Brain[J]. Cell, 2017, 169(6):1013.

## Proofs & Results & Conclusions

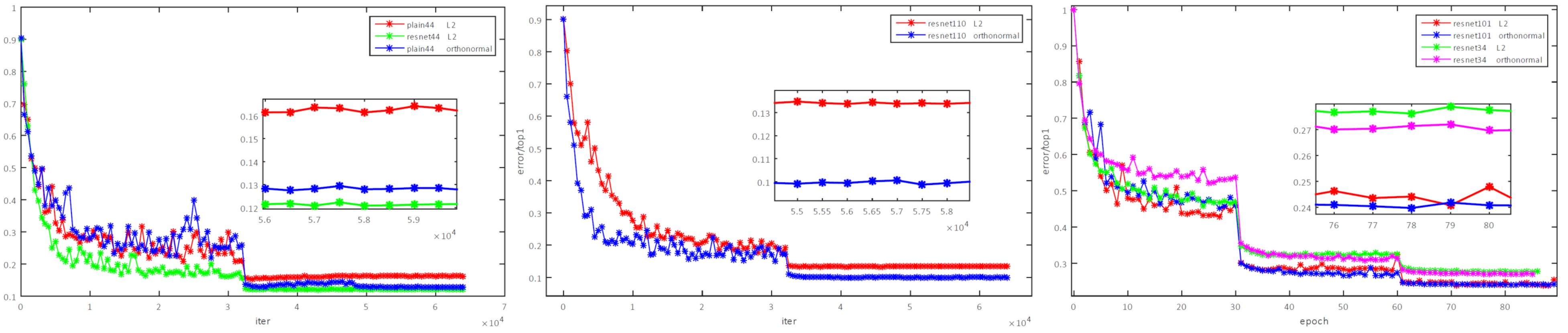
1. Orthonormality regularization enhances the magnitude of backward signals in extremely deep networks.
2. It probably exists a potential evolution pattern in training deep networks.
3. Orthonormality constraints helps reduce redundancies between filters, which inspires filter pruning to optimize the architecture of networks.
4. Regularization methods other than weight decay help explore different manifold space that may approach better local minima.
5. How to keep fidelity of signals instead of learning rate tuning is a key factor for making a deep network to have a reasonable learning behavior.







| Method | Top-1 Accuracy (%) | | |
|---|---|---|---|
| | 44-layer | 110-layer | 44-layer* |
| Nesterov | 85.0 | 10.18 | 61.9 |
| AdaGrad | 77.86 | 30.3 | 36.1 |
| AdaDelta | 70.56 | 66.48 | 52.6 |
| Adam | 39.85 | 10.0 | N/A |
| RmsProp | 10.0 | 10.0 | N/A |
| SGD | 84.14 | 11.83 | 65.2 |
| Ours | **88.42** | **81.6** | **70.0** |


resnet110 L2
resnet110 orthonormal
resnet101 L2
resnet101 orthonormal
resnet34 L2
resnet34 orthonormal
iter
epoch