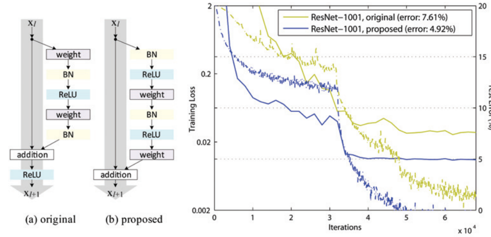
Identity Mappings in Deep Residual Networks

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
Microsoft Research Asia (MSRA)

Highlights

- Novel pre-activation residual structure
- Improved results using **1001**-layer ResNet on CIFAR-10 and **200**-layer on ImageNet



Results on CIFAR

CIFAR-10	error (%)
NIN [13]	8.81
DSN [14]	8.22
FitNet [15]	8.39
Highway [7]	7.72
ELU [12]	6.55
FitResNet, LSUV [16]	5.84
ResNet-110 [1] (1.7M)	6.61
ResNet-1202 [1] (19.4M)	7.93
ResNet-164 [ours] (1.7M)	5.46
ResNet-1001 [ours] (10.2M)	4.92 (4.89±0.14)
ResNet-1001 [ours] (10.2M)†	**4.62** (4.69±0.20)

CIFAR-100	error (%)
NIN [13]	35.68
DSN [14]	34.57
FitNet [15]	35.04
Highway [7]	32.39
ELU [12]	24.28
FitNet, LSUV [16]	27.66
ResNet-164 [1] (1.7M)	25.16
ResNet-1001 [1] (10.2M)	27.82
ResNet-164 [ours] (1.7M)	24.33
ResNet-1001 [ours] (10.2M)	**22.71** (22.68±0.22)

Results on ImageNet

method	train crop size	test crop size	top-1 (%)	top-5 (%)
ResNet-152, original Residual Unit [1]	224×224	224×224	23.0	6.7
ResNet-152, original Residual Unit [1]	224×224	320×320	21.3	5.5
ResNet-152, proposed Residual Unit	224×224	320×320	21.1	5.5
ResNet-200, original Residual Unit [1]	224×224	320×320	21.8	6.0
ResNet-200, proposed Residual Unit	224×224	320×320	**20.7**	**5.3**
Inception v3 [17]	299×299	299×299	21.2	5.6

Importance of Identity Skip Connections

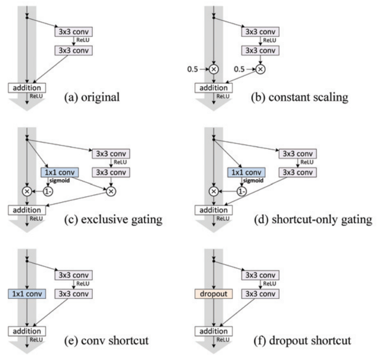
- ResNet with Identity mapping

$$\mathbf{x}_L = \mathbf{x}_l + \sum_{i=l}^{L-1} \mathcal{F}(\mathbf{x}_i, \mathcal{W}_i), \qquad \frac{\partial \mathcal{E}}{\partial \mathbf{x}_l} = \frac{\partial \mathcal{E}}{\partial \mathbf{x}_L}\frac{\partial \mathbf{x}_L}{\partial \mathbf{x}_l} = \frac{\partial \mathcal{E}}{\partial \mathbf{x}_L}\left(1 + \frac{\partial}{\partial \mathbf{x}_l}\sum_{i=l}^{L-1} \mathcal{F}(\mathbf{x}_i, \mathcal{W}_i)\right)$$

- What if we break the identity shortcut?

$$\mathbf{x}_L = \prod_{i=l}^{L-1} \lambda_i \mathbf{x}_l + \sum_{i=l}^{L-1} \hat{\mathcal{F}}(\mathbf{x}_i, \mathcal{W}_i), \qquad \frac{\partial \mathcal{E}}{\partial \mathbf{x}_l} = \frac{\partial \mathcal{E}}{\partial \mathbf{x}_L}\left(\prod_{i=l}^{L-1} \lambda_i + \frac{\partial}{\partial \mathbf{x}_l}\sum_{i=l}^{L-1} \hat{\mathcal{F}}(\mathbf{x}_i, \mathcal{W}_i)\right)$$

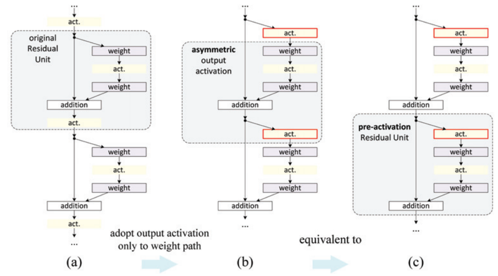
- Various types of shortcut connections



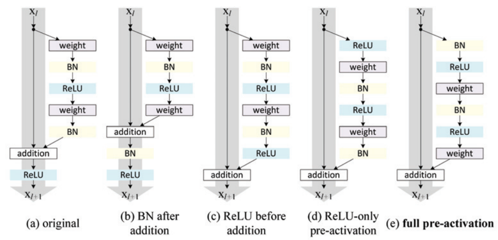
case	Fig.	on shortcut	on $\mathcal{F}$	error (%)	remark
original [1]	Fig. 2(a)	1	1	**6.61**	
constant scaling	Fig. 2(b)	0	1	fail	This is a plain net
		0.5	1	fail	
		0.5	0.5	12.35	
exclusive gating	Fig. 2(c)	$1-g(\mathbf{x})$	$g(\mathbf{x})$	fail	init b_g=0 to −5
		$1-g(\mathbf{x})$	$g(\mathbf{x})$	8.70	init b_g=-6
		$1-g(\mathbf{x})$	$g(\mathbf{x})$	9.81	init b_g=-7
shortcut-only gating	Fig. 2(d)	$1-g(\mathbf{x})$	1	12.86	init b_g=0
		$1-g(\mathbf{x})$	1	6.91	init b_g=-6
1×1 conv shortcut	Fig. 2(e)	1×1 conv	1	12.22	
dropout shortcut	Fig. 2(f)	dropout 0.5	1	fail	

Usage of Activation Function

- Post-activation to pre-activation



- Various usage of activation function



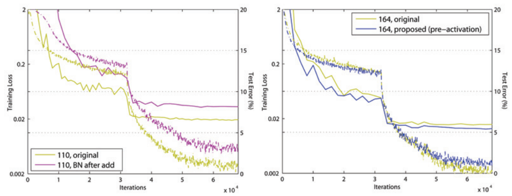
case	Fig.	ResNet-110	ResNet-164
original Residual Unit [1]	Fig. 4(a)	6.61	5.93
BN after addition	Fig. 4(b)	8.17	6.50
ReLU before addition	Fig. 4(c)	7.84	6.14
ReLU-only pre-activation	Fig. 4(d)	6.71	5.91
full pre-activation	Fig. 4(e)	**6.37**	**5.46**

Analysis of Pre-activation Structure

- Ease of optimization

dataset	network	baseline unit	pre-activation unit
CIFAR-10	ResNet-110(1layer)	9.90	8.91
	ResNet-110	6.61	6.37
	ResNet-164	5.93	5.46
	ResNet-1001	7.61	4.92
CIFAR-100	ResNet-164	25.16	24.33
	ResNet-1001	27.82	22.71

- Reducing overfitting



Code available:

- Deep Residual Networks with 1K Layers: https://github.com/KaimingHe/resnet-1k-layers

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016