Identity Mappings in Deep Residual Networks

Kaiming He, **Xiangyu Zhang**, Shaoqing Ren, and Jian Sun

Microsoft Research

Contributions

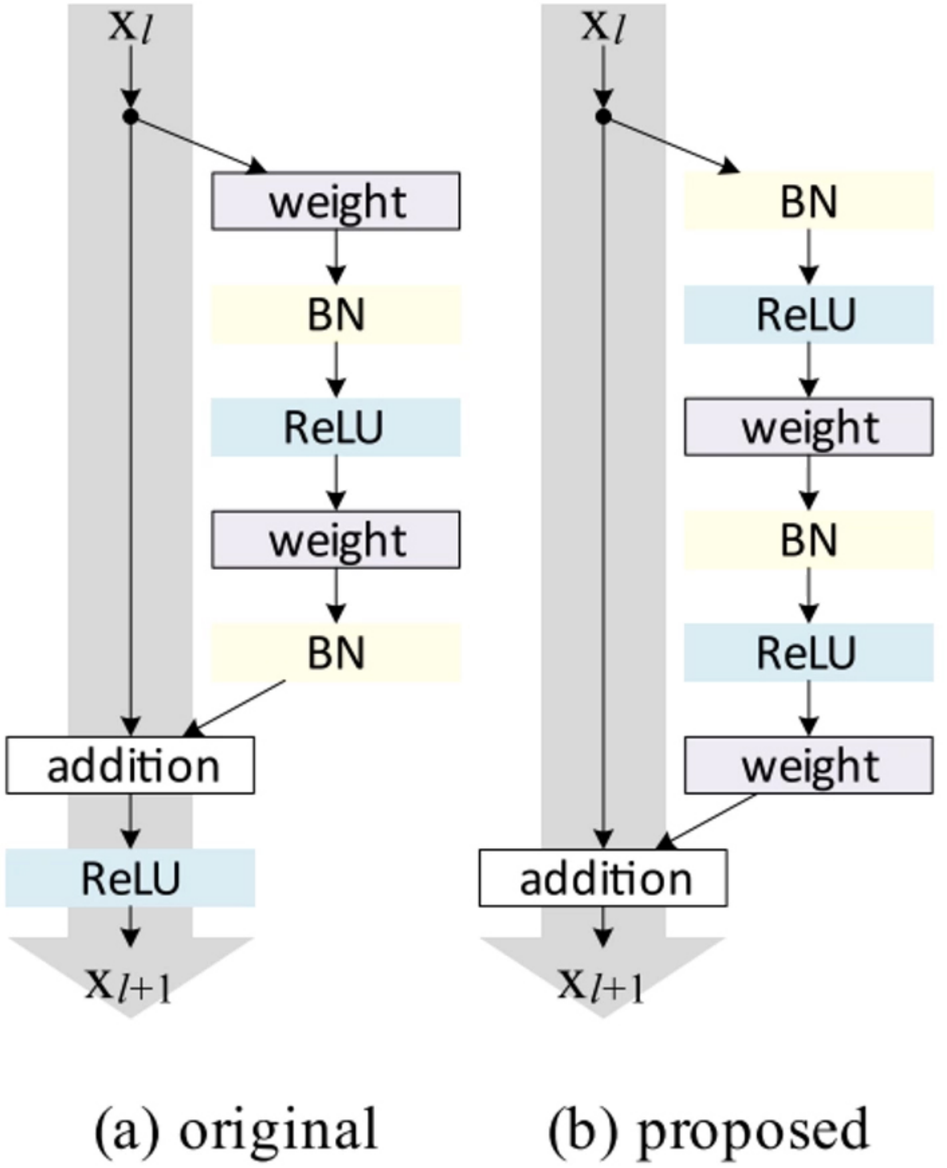
- Importance of “direct” information path
  - -- Identity mapping for shortcut path
  - -- Identity activation function

- Novel “pre-activation” design

Contributions

- Importance of “direct” information path
  - -- Identity mapping for shortcut path
  - -- Identity activation function

- Novel “pre-activation” design

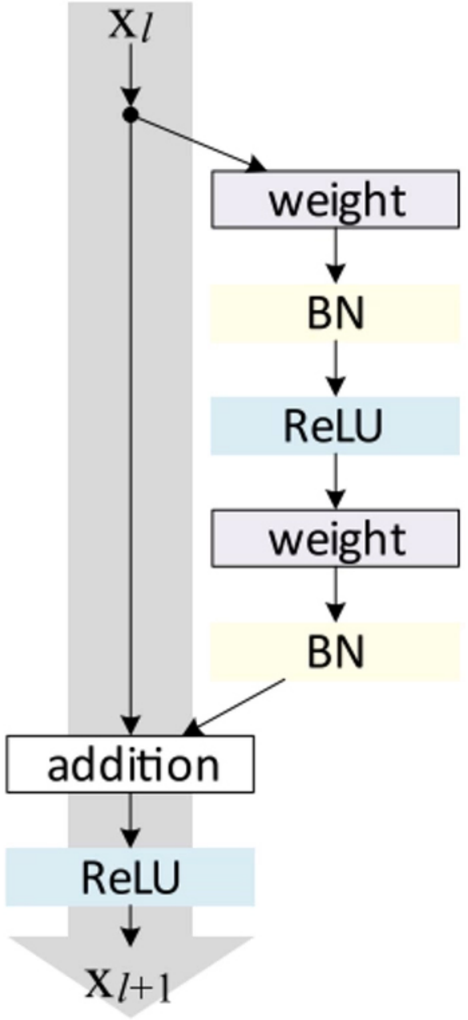


(a) original (b) proposed

Deep Residual Network: a Review

- Residual units in general form

$$\mathbf{y}_l = h(\mathbf{x}_l) + \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l),$$
$$\mathbf{x}_{l+1} = f(\mathbf{y}_l),$$

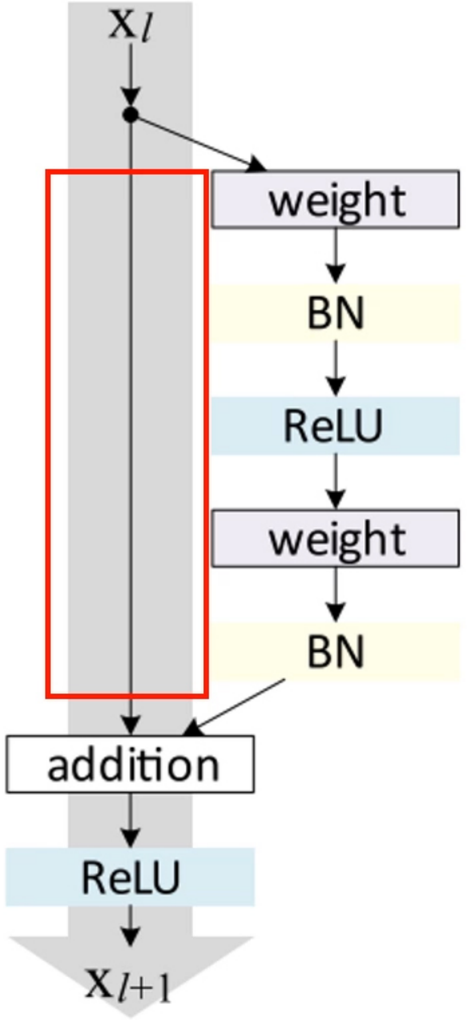


Deep Residual Network: a Review

- Residual units in general form

Skip Connection

$$\mathbf{y}_l = h(\mathbf{x}_l) + \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l),$$
$$\mathbf{x}_{l+1} = f(\mathbf{y}_l),$$

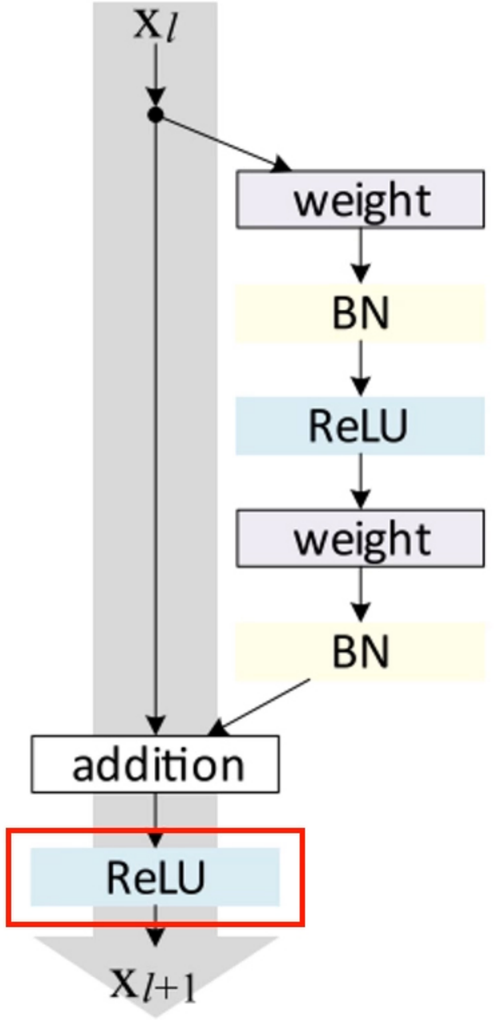


Deep Residual Network: a Review

- Residual units in general form

$$\mathbf{y}_l = h(\mathbf{x}_l) + \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l),$$
$$\mathbf{x}_{l+1} = f(\mathbf{y}_l),$$

Activation Function



Analysis of Deep Residual Networks

- What if both h and f are identity mappings?
- Forward view

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

- Backward view

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

Analysis of Deep Residual Networks

- What if both h and f are identity mappings?
- Forward view

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

- Backward view

"Clean" Information Path

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

On the Importance of Identity Skip Connections

- Let $h(\boldsymbol{x}_l) = \lambda_l \boldsymbol{x}_l$ to break identity shortcut
- Forward view

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

- Backward view

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

On the Importance of Identity Skip Connections

- Let $h(\boldsymbol{x}_l) = \lambda_l \boldsymbol{x}_l$ to break identity shortcut
- Forward view

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

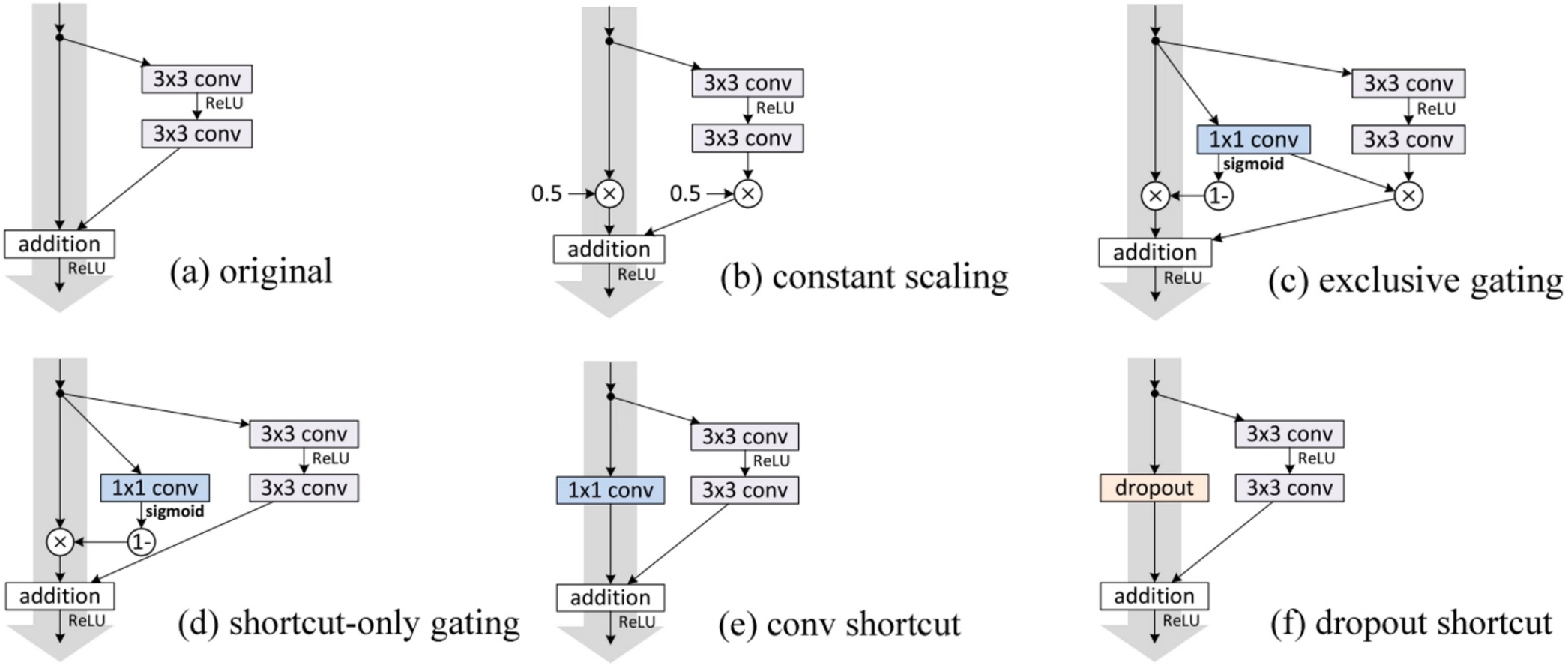
- Backward view

Risk of exponentially explosion or vanishing!

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

Experiments on Skip Connections

- Various types of shortcut connections

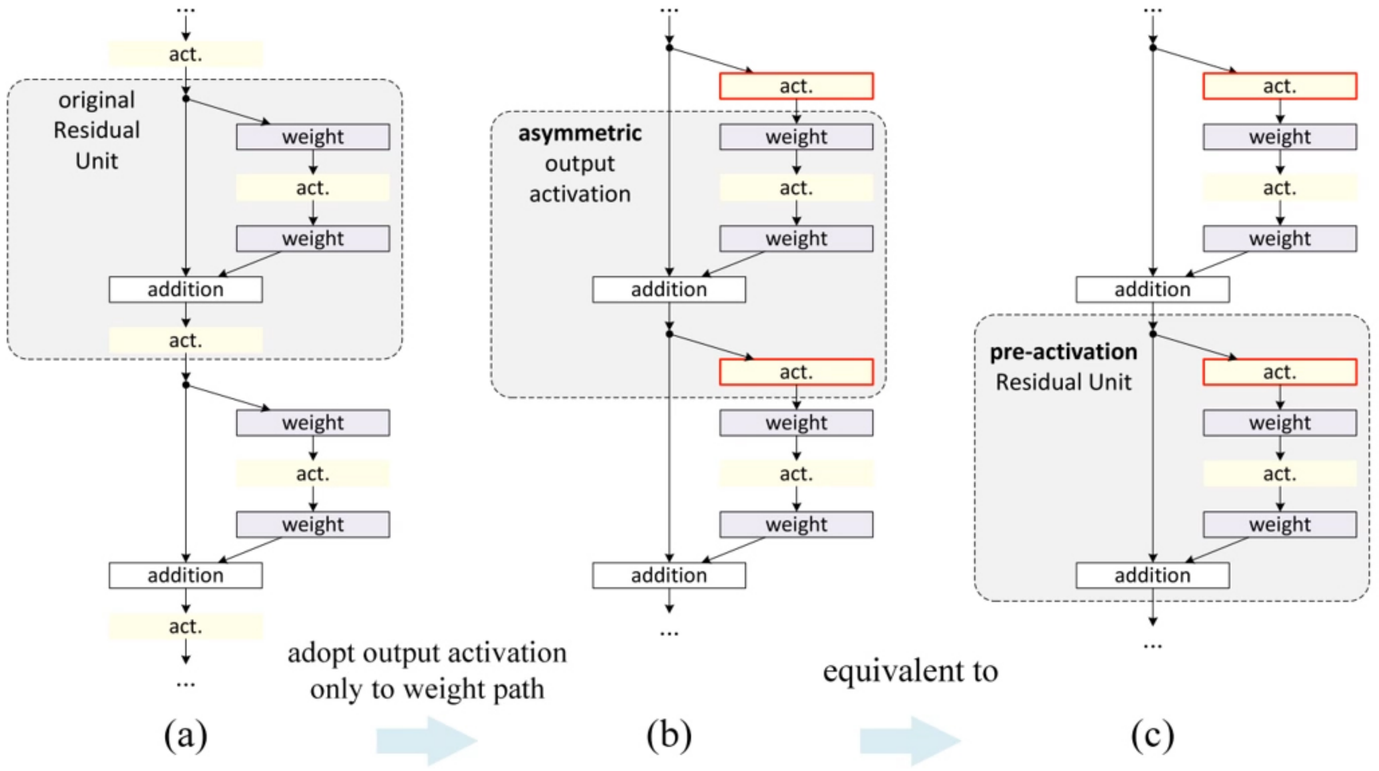


Experiments on Skip Connections

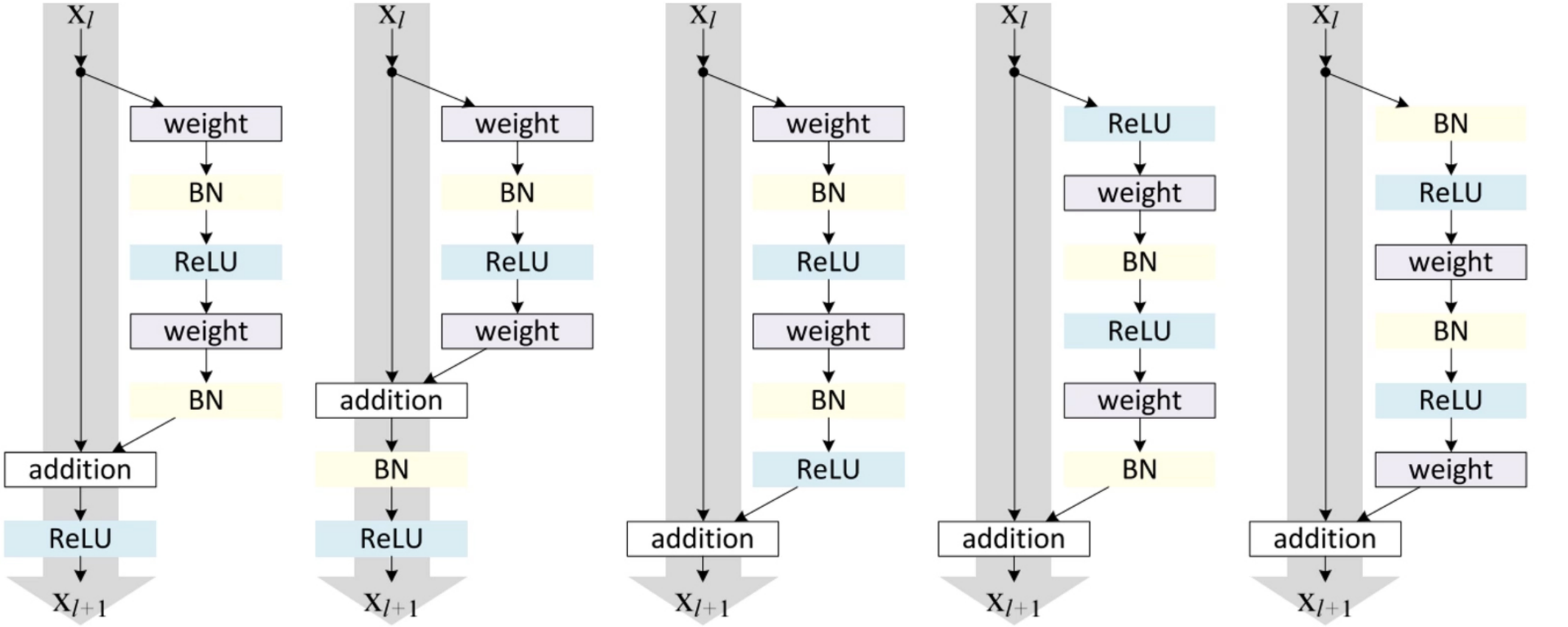
case	Fig.	on shortcut	on $\mathcal{F}$	error (%)	remark
original [1]	Fig. 2(a)	1	1	**6.61**	
constant scaling	Fig. 2(b)	0	1	fail	This is a plain net
		0.5	1	fail	
		0.5	0.5	12.35	
exclusive gating	Fig. 2(c)	$1-g(\mathbf{x})$	$g(\mathbf{x})$	fail	init b_g=0 to −5
		$1-g(\mathbf{x})$	$g(\mathbf{x})$	8.70	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	

On the Usage of Activation Functions

- Not only identity shortcut, but also identity activation function
- Pre-activation design



Experiments on Activation



(a) original

(b) BN after addition

(c) ReLU before addition

(d) ReLU-only pre-activation

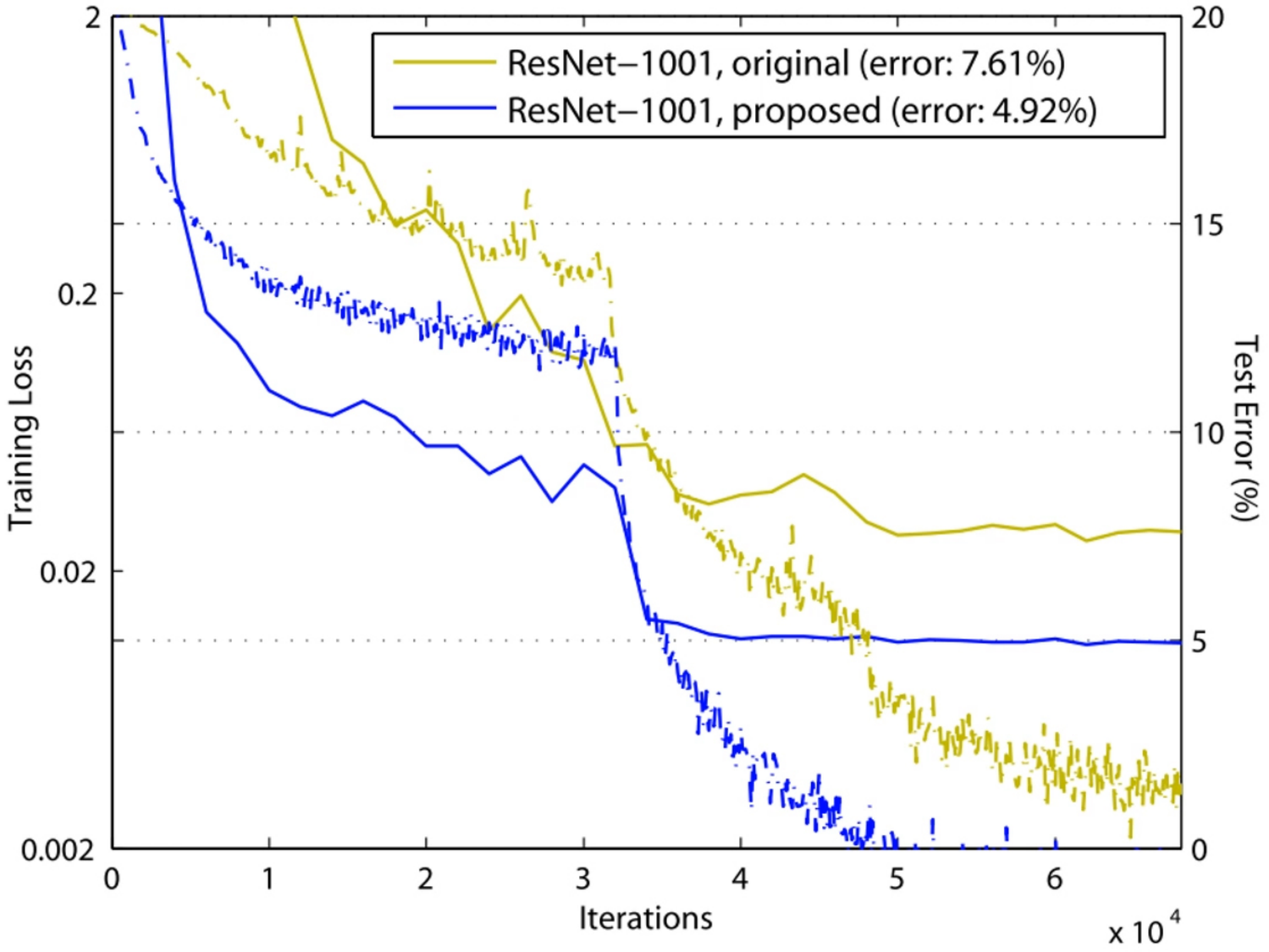
(e) **full pre-activation**

Experiments on Activation

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**

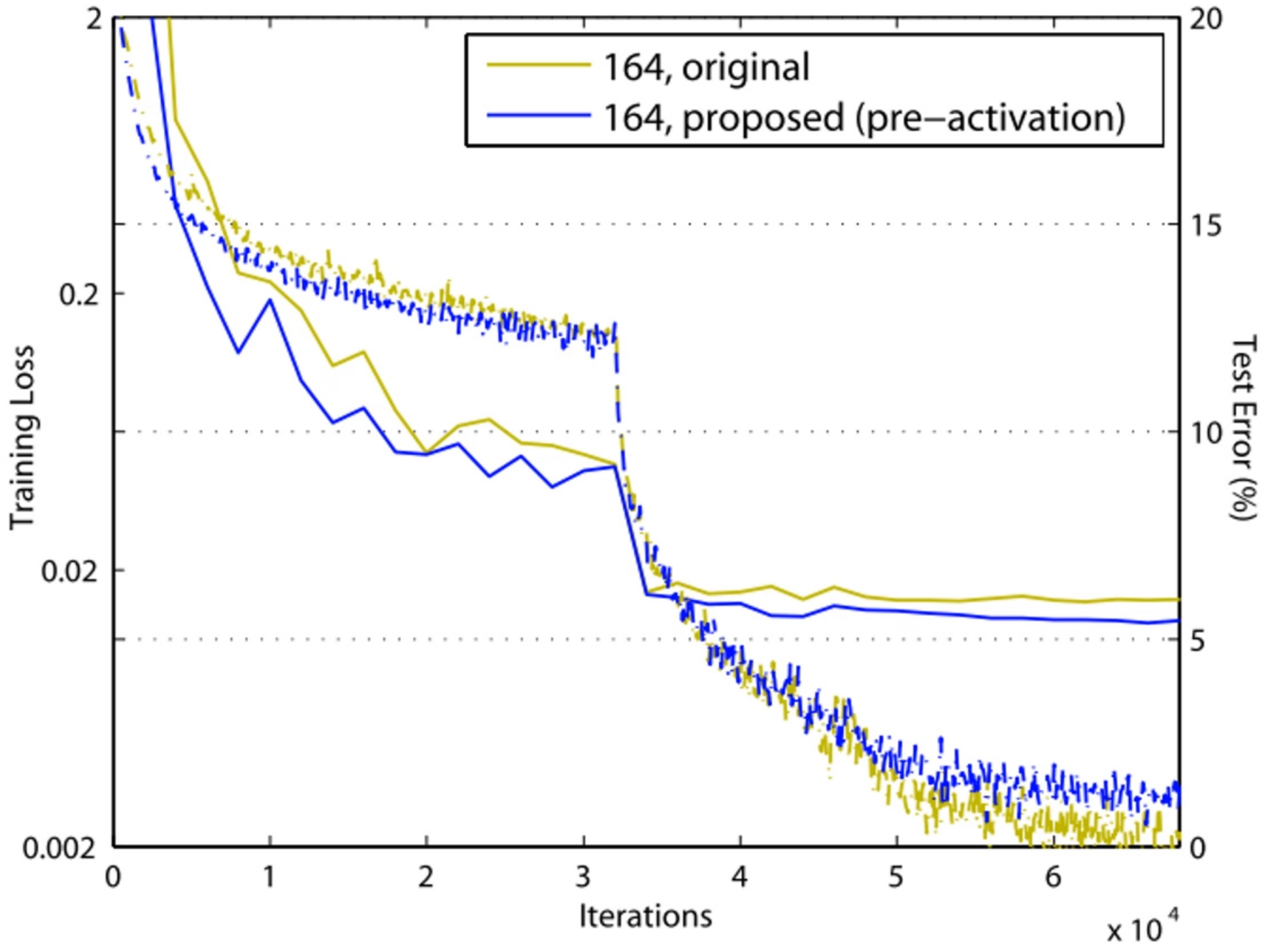
Analysis

- Ease of optimization for very deep networks



Analysis

- Reducing overfitting



Experiments 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

Thank you