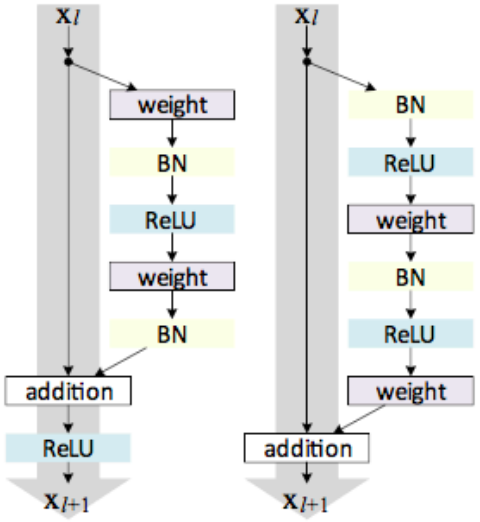
1.1.1 [Resnet-v2] Notes by YK

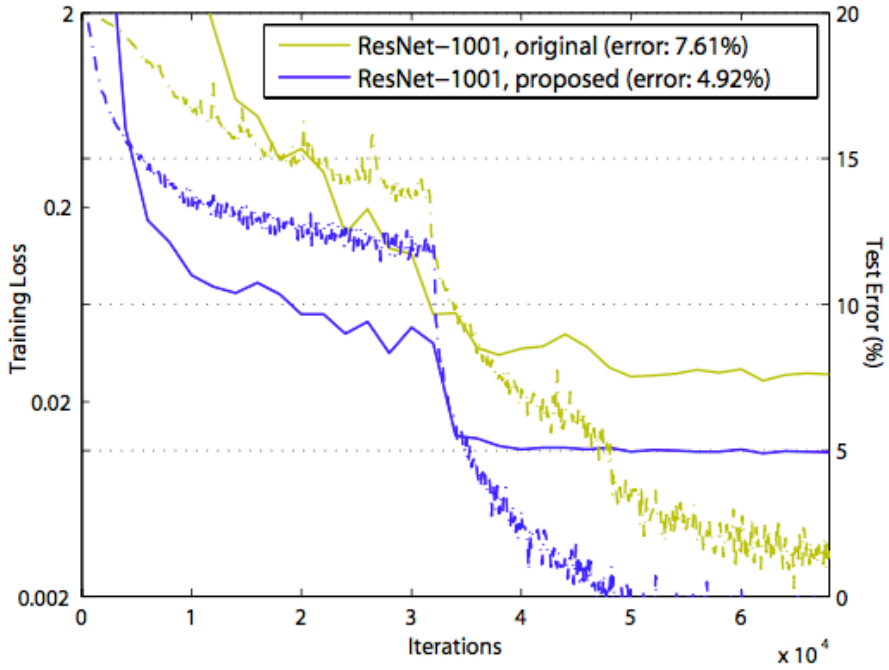
Identity Mappings in Deep Residual Networks

a. Main Contributions

1) analyze the propagation formulation of the residual units
2) show the importance of identity mappings by a series of ablation experiments
3) propose a residual network which propagates information through the entire network, not only within a residual unit.



(a) original (b) proposed



b. Key Points

1) propagation formulation

The original Residual Unit performs the following computation:

$$y_l = h(x_l) + F(x_l, W_l)$$
$$x_{l+1} = f(y_l)$$

where F denotes the residual function and f denotes the function after element-wise addition.
If both h and f are identity mapping, then $x_{l+1} = x_l + F(x_l, W_l)$.
Recursively,$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i)$,where layer L is deeper than l.

This equation leads to nice backward propagation properties:

$$\frac{\partial \epsilon}{\partial x_l} = \frac{\partial \epsilon}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial \epsilon}{\partial x_L}(1 + \frac{\partial}{x_l}\sum_{i=l}^{L-1} F(x_i, W_i)) = \frac{\partial \epsilon}{\partial x_L} + \frac{\partial \epsilon}{\partial x_L}\frac{\partial}{x_l}\sum_{i=l}^{L-1} F(x_i, W_i)$$

This equation presents **two advantages** of indentity mappings as below:

(i)The first term ensures that information of gradients is directly propagated from deeper layers to any shallower unit l.

(ii)The gradient $\frac{\partial \epsilon}{\partial x_l}$ is unlikely to be canceled out for a mini-batch, because $\frac{\partial}{x_l} \sum_{i=l}^{L-1} F$ cannot be always -1 for all samples in a mini-batch.

These two advantages are based on **two conditions**:
(i)the skip connection is identity $h(x_l) = x_l$.
(ii)the function after element-wise addtion is identity $x_{l+1} = y_l$.

2) the importance of indentity mappings

If the shorcut of residual units is not identity, e.g. $h(x_l) = \lambda_l x_l$, the backpropagation equation will be different:

$$\frac{\partial \epsilon}{\partial x_l} = \frac{\partial \epsilon}{\partial x_L} (\prod_{i=l}^{L-1} \lambda_i + \frac{\partial}{x_l} \sum_{i=l}^{L-1} \hat{F(x_i, W_i)})\$$

Therefore, the first term loses its advantages and might impede information propagation.

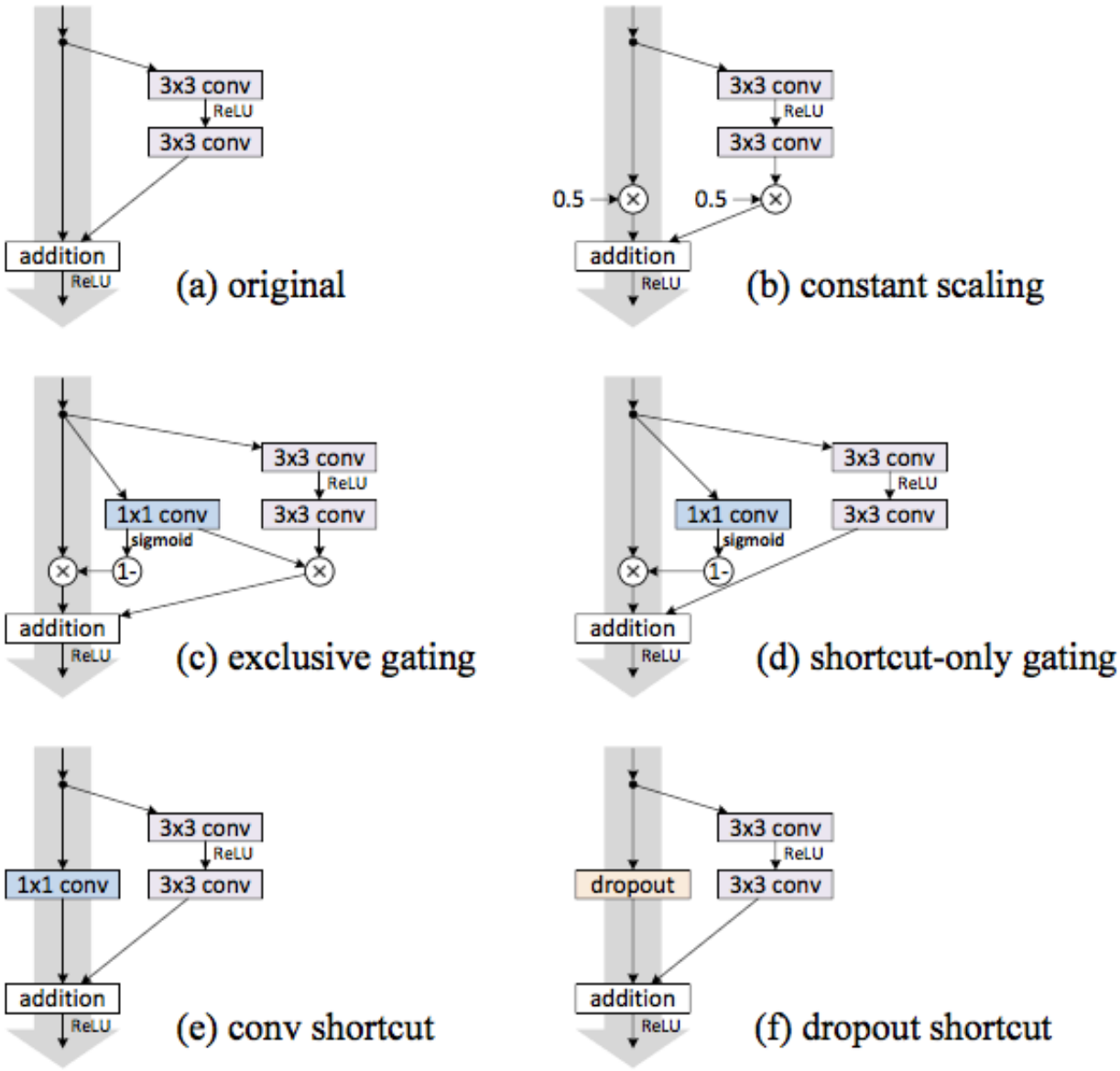
3) the impact of pre-activation

$$pre - activation : BN - Relu - Conv$$
$$post - activation : Conv - Relu - BN$$

The impact of pre-activation is two-fold:
(i)**Ease of optimization** while training(comparing with the baseline ResNet)
(ii)**Reducing overfitting**:
In the original Residual Unit, although the BN normalizes the signal, this is soon added to the shortcut and thus the merged signal is not normalized, which is then used as the input of the next weight layer.

c. Experiments and Results

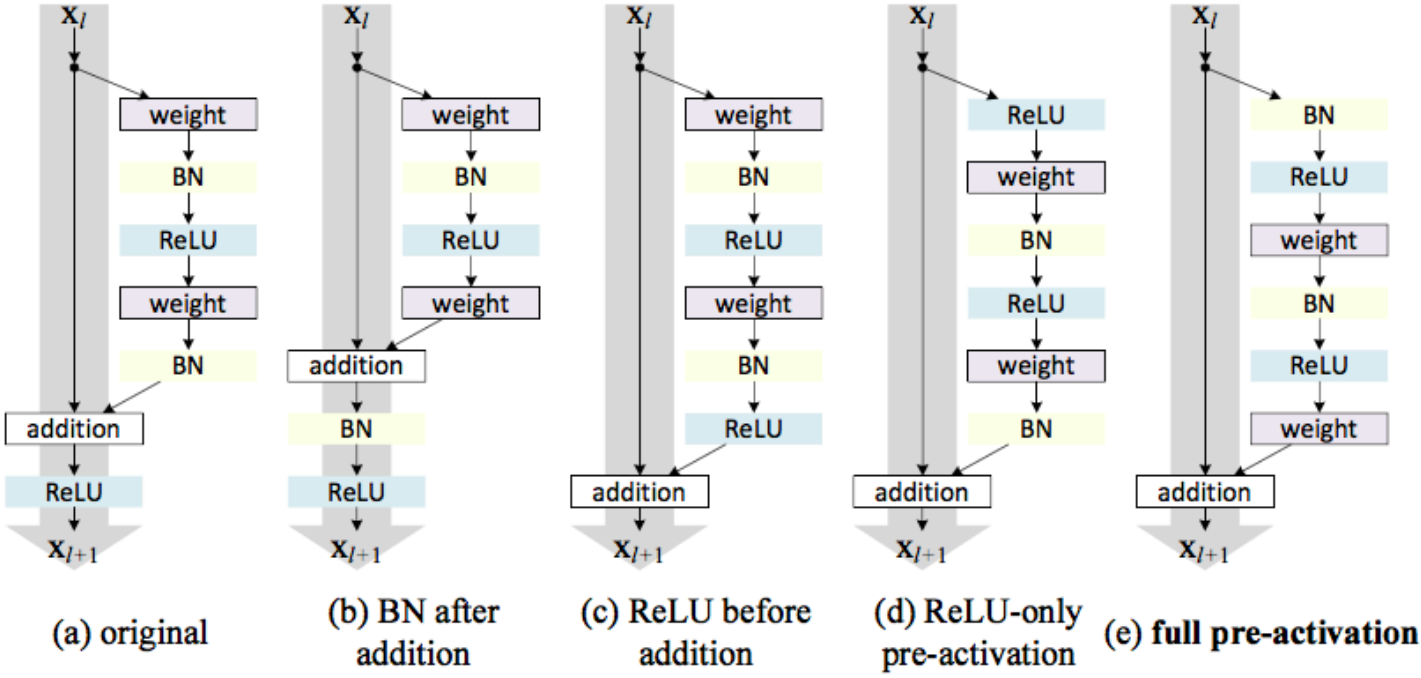
(i) Experiments on Skip Connections
The results show that identity skip connection is important for the accuracy and variation of training.



(ii) Experiments on Activation

The results show that pre-activation is best thanks to its ease for optimization and reducing overfitting.

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**



d. English Writing

(i) Nice phrases
--**exhibits some nice properties**
--**compelling accuracy and nice convergence behaviors**
--**in contrast to conventional wisdom of ...**
--ablation experiments
--propagation formulations
--after-addition activation
--modularized architecture
--truncate
--presumably

(ii)Nice sentences
--This is in contrast to a "plain network" where a feature x_L is **a series of matrix-vector products**.
--This product may also **impede information propagation and hamper the training procedure** as witnessed in the following experiments.
--This **implies** that the gradient of a layer does not **vanish** even when the weights are arbitrarily small.