

1.1.1 [Resnet-v2] Identity Mappings in Deep Residual Networks

a. What to Solve

- 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.



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:

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

This equation presents **two advantages** of identity 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 $\frac{\partial}{\partial x_l} \sum_{i=l}^{L-1} F$ 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 addition is identity $x_{l+1} = y_l$.

2) the importance of identity mappings

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

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

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. Results



d. English Writing