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

\[ y_l = h(x_l) + F(x_l, W_l), \]
\[ x_{l+1} = f(y_l), \]
Deep Residual Network: a Review

- Residual units in general form

\[ y_l = h(x_l) + F(x_l, W_l), \]
\[ x_{l+1} = f(y_l), \]
Deep Residual Network: a Review

• Residual units in general form

\[
y_l = h(x_l) + F(x_l, \mathcal{W}_l), \\
x_{l+1} = f(y_l),
\]

Activation Function
Analysis of Deep Residual Networks

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

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

• Backward view

$$\frac{\partial \mathcal{E}}{\partial x_l} = \frac{\partial \mathcal{E}}{\partial x_L} \frac{\partial x_L}{\partial x_l} = \frac{\partial \mathcal{E}}{\partial x_L} \left( 1 + \frac{\partial}{\partial x_l} \sum_{i=l}^{L-1} F(x_i, W_i) \right)$$
Analysis of Deep Residual Networks

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

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, \mathcal{W}_i),$$

• Backward view

$$\frac{\partial \mathcal{E}}{\partial x_l} = \frac{\partial \mathcal{E}}{\partial x_L} \frac{\partial x_L}{\partial x_l} = \frac{\partial \mathcal{E}}{\partial x_L} \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, \mathcal{W}_i) \right)$$
On the Importance of Identity Skip Connections

• Let \( h(x_i) = \lambda_i x_i \) to break identity shortcut

• Forward view

\[
x_L = \prod_{i=l}^{L-1} \lambda_i x_i + \sum_{i=l}^{L-1} \hat{F}(x_i, \mathcal{W}_i),
\]

• Backward view

\[
\frac{\partial \mathcal{E}}{\partial x_l} = \frac{\partial \mathcal{E}}{\partial x_L} \left( \prod_{i=l}^{L-1} \lambda_i + \frac{\partial}{\partial x_l} \sum_{i=l}^{L-1} \hat{F}(x_i, \mathcal{W}_i) \right).
\]
On the Importance of Identity Skip Connections

• Let $h(x_i) = \lambda_i x_i$ to break identity shortcut

• Forward view

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

• Backward view

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

Risk of exponentially explosion or vanishing!
Experiments on Skip Connections

- Various types of shortcut connections

(a) original

(b) constant scaling

(c) exclusive gating

(d) shortcut-only gating

(e) conv shortcut

(f) dropout shortcut
# Experiments on Skip Connections

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

<table>
<thead>
<tr>
<th>case</th>
<th>Fig.</th>
<th>ResNet-110</th>
<th>ResNet-164</th>
</tr>
</thead>
<tbody>
<tr>
<td>original Residual Unit [1]</td>
<td>Fig. 4(a)</td>
<td>6.61</td>
<td>5.93</td>
</tr>
<tr>
<td>BN after addition</td>
<td>Fig. 4(b)</td>
<td>8.17</td>
<td>6.50</td>
</tr>
<tr>
<td>ReLU before addition</td>
<td>Fig. 4(c)</td>
<td>7.84</td>
<td>6.14</td>
</tr>
<tr>
<td>ReLU-only pre-activation</td>
<td>Fig. 4(d)</td>
<td>6.71</td>
<td>5.91</td>
</tr>
<tr>
<td><strong>full pre-activation</strong></td>
<td>Fig. 4(e)</td>
<td><strong>6.37</strong></td>
<td><strong>5.46</strong></td>
</tr>
</tbody>
</table>
Analysis

• Ease of optimization for very deep networks
Analysis

- Reducing overfitting
# Experiments on ImageNet

<table>
<thead>
<tr>
<th>method</th>
<th>train crop size</th>
<th>test crop size</th>
<th>top-1 (%)</th>
<th>top-5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-152, original Residual Unit [1]</td>
<td>$224 \times 224$</td>
<td>$224 \times 224$</td>
<td>23.0</td>
<td>6.7</td>
</tr>
<tr>
<td>ResNet-152, original Residual Unit [1]</td>
<td>$224 \times 224$</td>
<td>$320 \times 320$</td>
<td>21.3</td>
<td>5.5</td>
</tr>
<tr>
<td>ResNet-152, proposed Residual Unit</td>
<td>$224 \times 224$</td>
<td>$320 \times 320$</td>
<td>21.1</td>
<td>5.5</td>
</tr>
<tr>
<td>ResNet-200, original Residual Unit [1]</td>
<td>$224 \times 224$</td>
<td>$320 \times 320$</td>
<td>21.8</td>
<td>6.0</td>
</tr>
<tr>
<td>ResNet-200, proposed Residual Unit</td>
<td>$224 \times 224$</td>
<td>$320 \times 320$</td>
<td>$\mathbf{20.7}$</td>
<td>$\mathbf{5.3}$</td>
</tr>
<tr>
<td>Inception v3 [17]</td>
<td>$299 \times 299$</td>
<td>$299 \times 299$</td>
<td>21.2</td>
<td>5.6</td>
</tr>
</tbody>
</table>
Thank you