Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

Presented by Anita Tadakamalla
February 21, 2020
CS 747, Spring 2020
Hierarchical Features and Role

• Low, mid, and high-level features
• More layers enrich the “levels” of the features
• Previous ImageNet models have depths of 16 and 30 layers
Is learning better networks as easy as stacking more layers?
Why is it not ok to add more layers?

- It's not ok because it introduces problems during training such as:
  - Vanishing/Exploding gradients
    - Can be addressed by normalized initialization and intermediate normalization
  - Degradation problem
    - How can this be solved?
Degradation Problem

Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.
Degradation Problem
Degradation Problem
Degradation Problem
Construction Insight

- Consider a shallow architecture and its deeper counterpart
- The deeper the model would just need to copy the shallower model with identity mappings
- Constructed solution suggests that a deeper model should produce no higher training error than its shallow counterpart
Residual Functions

• We explicitly reformulate the layers as learned residual functions with reference to the layer input, instead of learning un-referenced functions

• $H(x) = F(x) + x$

Figure 2. Residual learning: a building block.
ResNet when the dimensions of $x \neq F$

1. The shortcut still performs the identity mapping, with extra zero entries padded for increasing dimensions
   - No extra parameters in this method
   - Very quick

2. Projection matches the dimensions with a 1x1 convolution

$$y = \mathcal{F}(x, \{W_i\}) + W_s x.$$
ResNet Architecture

Linear projections
For dimension matching

\[ Y = F(x, \{W_i\}) + W_s x \]
Experiments

• 152 layers on ImageNet
  • 8x deeper than VGG nets
  • Less Parameters
    • ResNet-152 (11.3 billion FLOPs) vs VGG-16/19 (15.3/19.6 billion FLOPs)

• Ensemble of ResNets achieve 3.57% error on ImageNet test
  • 1st Place in ILVRC 2015

• CIFAR-10 with 100 and 1000 layers

• COCO object detection
  • +28% improvement
  • 1st place on COCO detection and segmentation
Experiments on ImageNet dataset

- ImageNet dataset has 1000 classes
- 1.28M images were used for training
- 50K images were used for validation
- 100K images were used for final testing
- Batch normalization
- Mini-batch size of 256
- Learning rate of 0.1 and divide by 10 when error plateaus
- Weight decay of 0.0001
- Momentum of 0.9
- Trained for 60 x 10^4 iterations
- He initializer
Training on ImageNet

Plain networks of 18 and 34 layers

Residual networks of 18 and 34 layers

* Thin curve denotes training error, and thick curve denotes validation error or the center crops
** The residual networks have no extra parameters compare to their plain counterparts
Identity vs Projection Shortcuts

A. Zero-padding shortcuts
B. Projection shortcuts are used for increasing dimensions only
C. All shortcuts are Projections
   This introduces significantly more parameters

<table>
<thead>
<tr>
<th>Options</th>
<th>Increasing dimension</th>
<th>Same dimension</th>
<th>Top-1 err</th>
<th>Top-5 err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain</td>
<td>no shortcut</td>
<td>no shortcut</td>
<td>28.54</td>
<td>10.02</td>
</tr>
<tr>
<td>A</td>
<td>Zero-pad (parameter-free)</td>
<td>Identity</td>
<td>25.03</td>
<td>7.76</td>
</tr>
<tr>
<td>B</td>
<td>Projection shortcuts</td>
<td>Identity</td>
<td>24.52</td>
<td>7.46</td>
</tr>
<tr>
<td>C</td>
<td>Projection shortcuts</td>
<td>Projection shortcuts</td>
<td>24.19</td>
<td>7.40</td>
</tr>
</tbody>
</table>
Deeper Bottleneck Architectures (ResNet-50/101/152)

* Done to accommodate for less training time
Results on ImageNet dataset

- Error rates (%, 10-crop testing) on ImageNet validation. ResNet-50/101/152 are of option B
- Error rates (%) of ensembles
- The top-5 error is on the test set of ImageNet and reported by the test server
Experiments on CIFAR-10 dataset

- CIFAR-10 dataset has 10 classes
- 45K images were used for training
- 5K images were used for validation
- 10K images were used for testing
- Batch normalization
- Mini-batch size of 128
- Learning rate of 0.1 and divide by 10 at step 32K and 48K
- Weight decay of 0.0001
- Momentum of 0.9
- Termination of training at step 64K

<table>
<thead>
<tr>
<th>output map size</th>
<th>32×32</th>
<th>16×16</th>
<th>8×8</th>
</tr>
</thead>
<tbody>
<tr>
<td># layers</td>
<td>1+2n</td>
<td>2n</td>
<td>2n</td>
</tr>
<tr>
<td># filters</td>
<td>16</td>
<td>32</td>
<td>64</td>
</tr>
</tbody>
</table>
Results on CIFAR-10 dataset

Classification error on the CIFAR-10 test set. All methods are with data augmentation. For ResNet-110, it was run 5 times and show “best (mean ± std)”
Effect of number of layers on the CIFAR-10 dataset

Plain networks

Residual networks

* Dashed lines denote training error; bold lines denote testing error
Main Contributions

• Residual learning eases optimization.
• Solved the degradation problem.
• Faster training of deep neural networks.
• Decreases the error rate for deeper networks.
ResNets @ ILSVRC & COCO 2015 Competitions

• 1st places in all five main tracks
  • ImageNet Classification: “Ultra-deep” 152-layer nets
  • ImageNet Detection: 16% better than 2nd
  • ImageNet Localization: 27% better than 2nd
  • COCO Detection: 11% better than 2nd
  • COCO Segmentation: 12% better than 2nd
152 layers

ILSVRC'15 ResNet: 3.57%
ILSVRC'14 GoogleNet: 6.7%
ILSVRC'14 VGG: 7.3%
ILSVRC'13: 11.7%
ILSVRC'12 AlexNet: 16.4%
ILSVRC'11: 25.8%
ILSVRC'10: 28.2%

ImageNet Classification top-5 error (%)