Deep Residual Networks: ResNet

CS8004: Deep Learning and Applications

Source: Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun.
"Deep Residual Learning for Image Recognition". arXiv 2015.

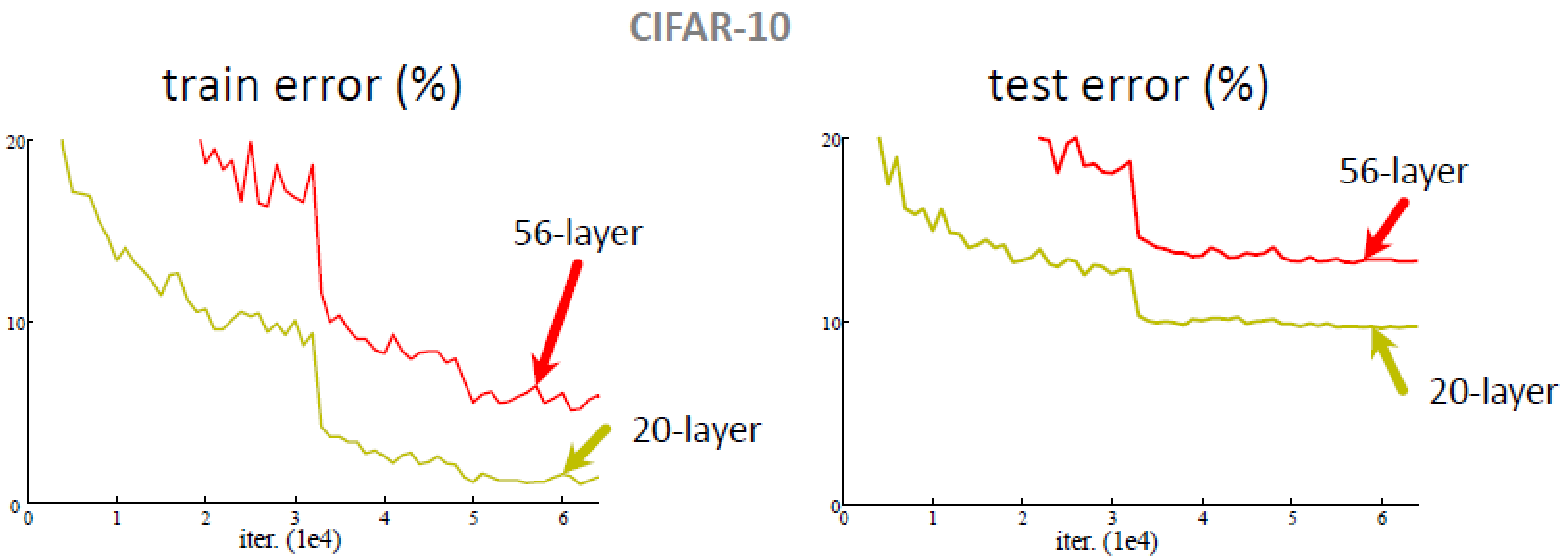
Deeper Networks

- A deeper network can provide solution to more complex problems.
  - With increasing feature map size.
  - Involving more nonlinearity
- But training a deeper network is much more difficult and a complex problem.
- Why ?
  - Due to exploding /vanishing gradients ?
  - Or due to overfitting ?

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Plain CNN

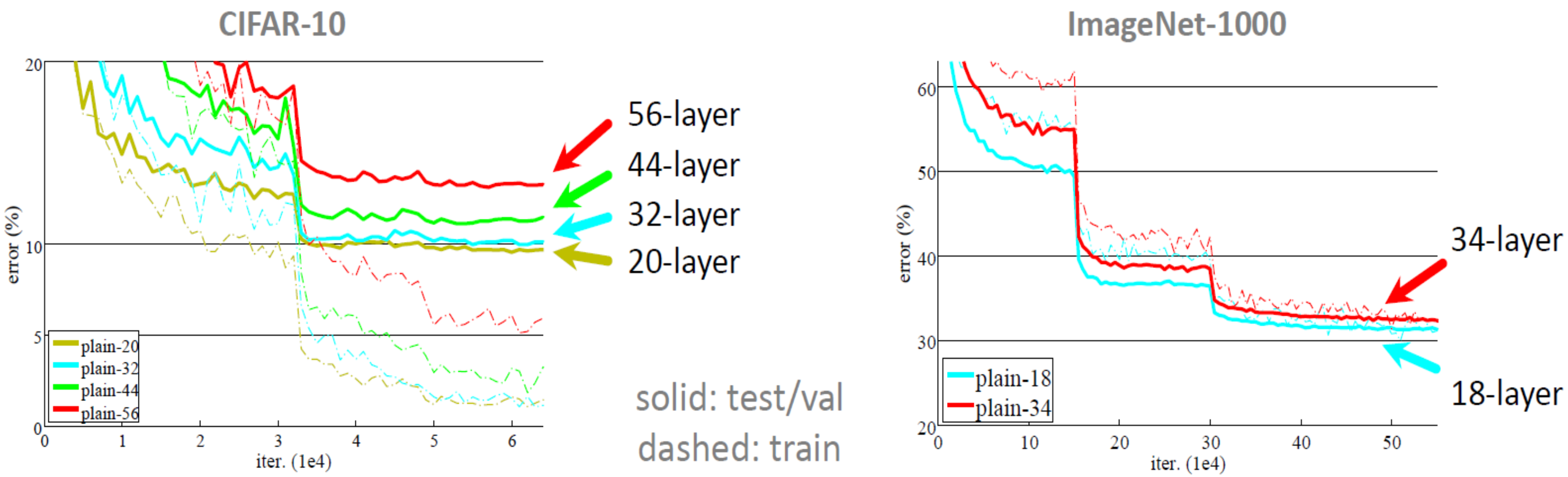
- Plain nets: A CNN with 3x3 Conv layers
- 56-layer net has higher training error and test error than 20-layers net.



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Plain CNN

- Very deep plain CNN have higher training error.
- A general phenomenon, observed in many datasets

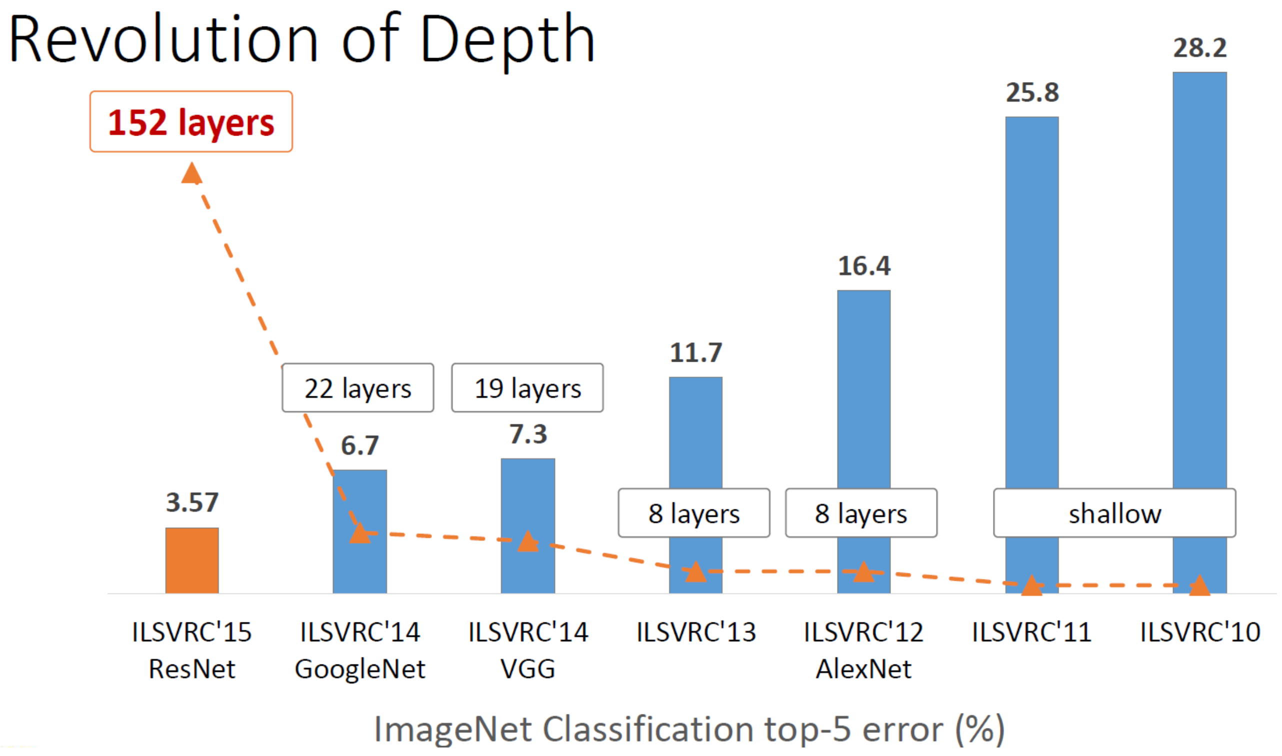


So, it is not a problem of overfitting !
Then it is vanishing \exploding gradients ?

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Deep CNN

- Solutions: Training relatively less number of layers
  - ReLU for solving gradient vanishing problem
  - Dropout …

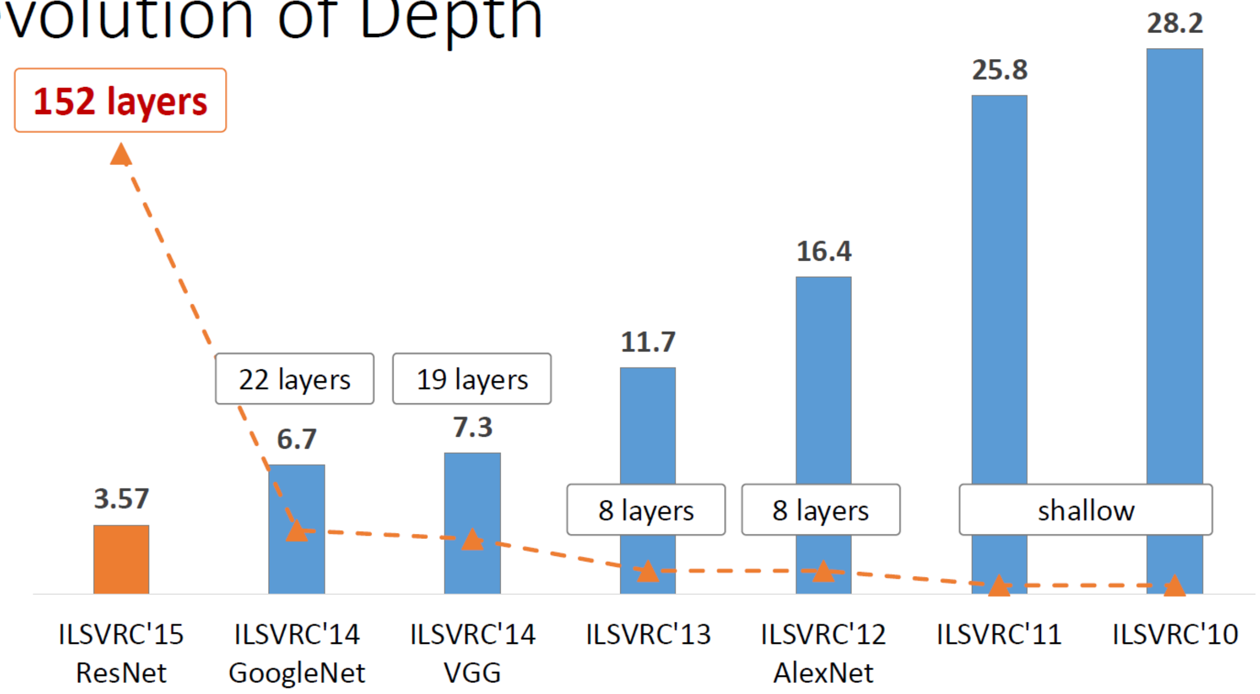


Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Deep CNN

- Solution: ~ 10 layers
  - Normalized initialization.
  - Intermediate normalization layers.

Revolution of Depth

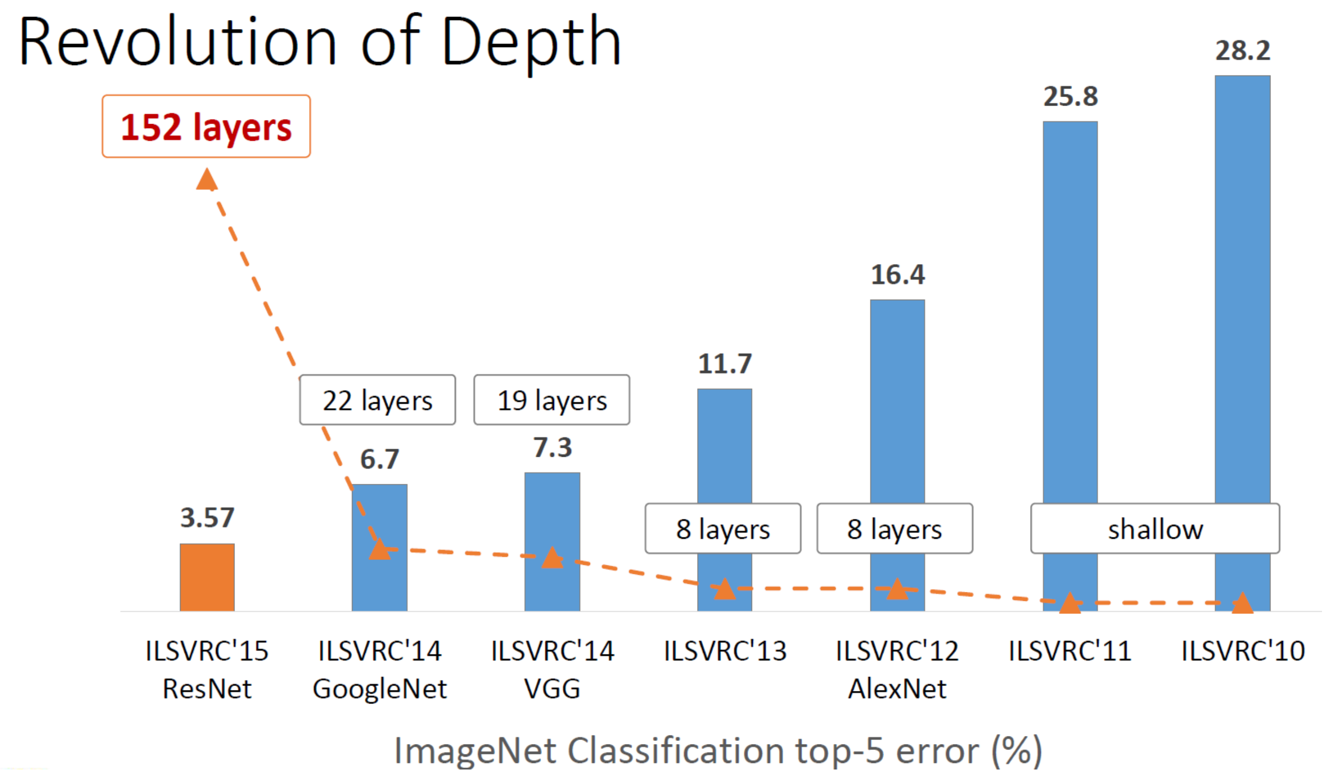


ImageNet Classification top-5 error (%)

Residual Learning for Image Recognition". arXiv 2015.

Deep CNN

- Solution beyond 100 layers
  - Residual network



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Kaiming He

- Research Scientist
- Facebook AI Research (FAIR)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

A Deep Residual Learning Framework

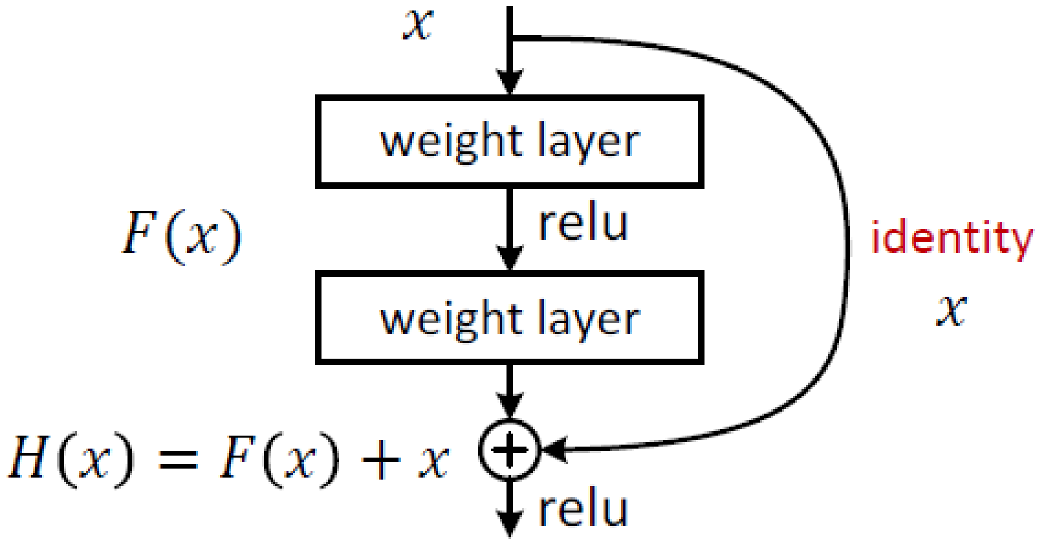
- Fit a residual map in place of directly fitting a desired underlying mapping H(x).
- Let the stacked nonlinear layers fit another mapping of the form F(x) := H(x)−x.
- The original mapping is recast into F(x)+x.
- Hypothesis: It is easier to optimize the residual mapping than to optimize the original, unreferenced mapping.
- If an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.

Source: Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

A Deep Residual Learning Framework

- Think of a shortcut connection.

- Identity mappings neither increase the computational complexity nor add extra parameters



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

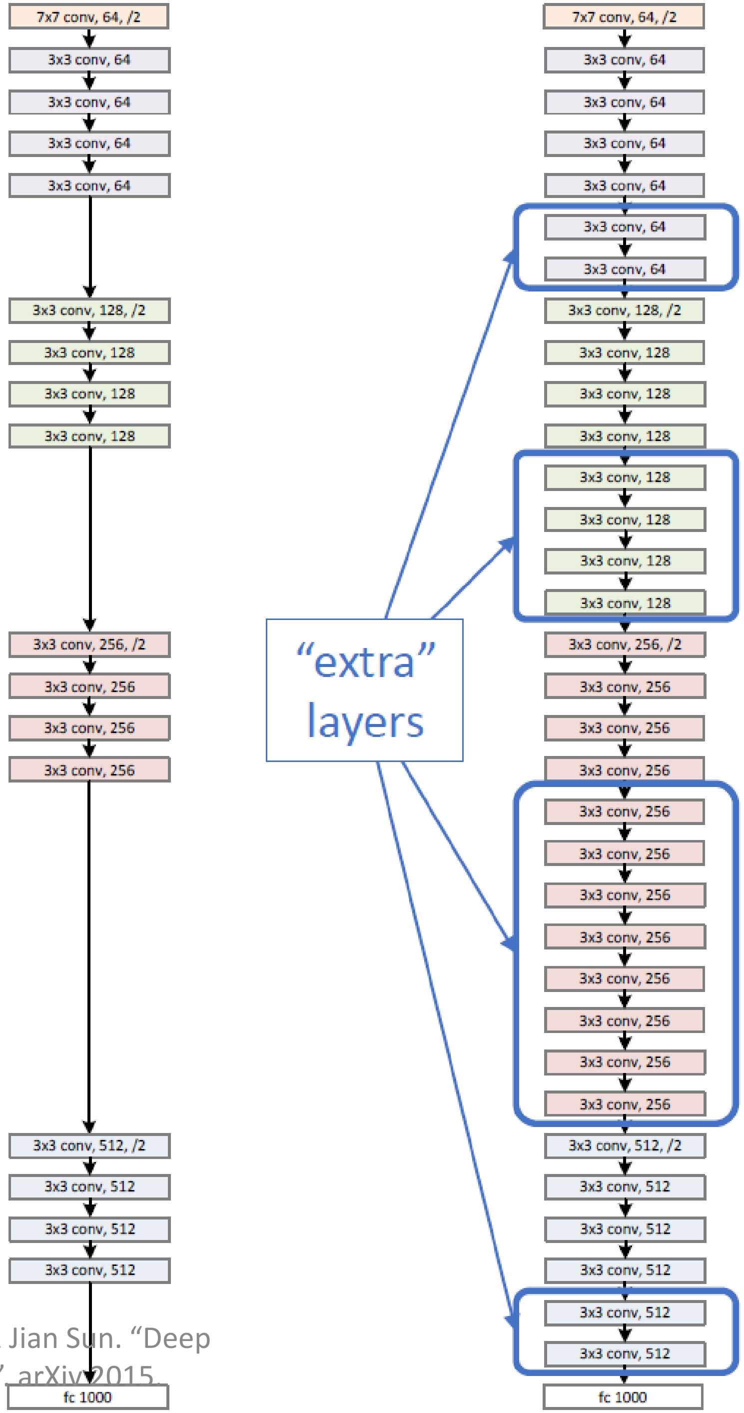
A Deep Residual Learning Framework

- The entire network can still be trained end-to-end using SGD with backpropagation.
- Can be easily implemented using common libraries without modifying the solvers.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Residual Network

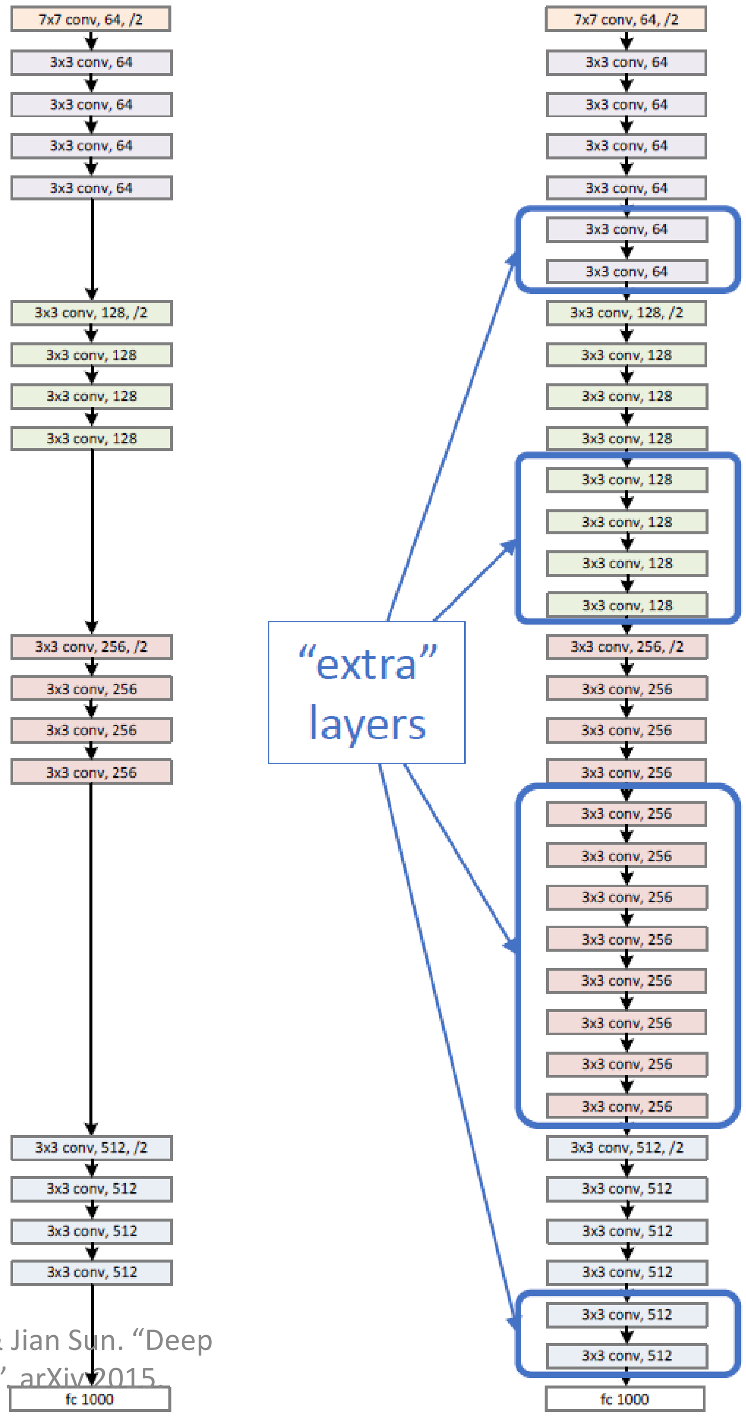
- Naïve solution
  - If extra layers are an identity mapping, then training error does not increase.



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition", arXiv 2015.

Residual Network...

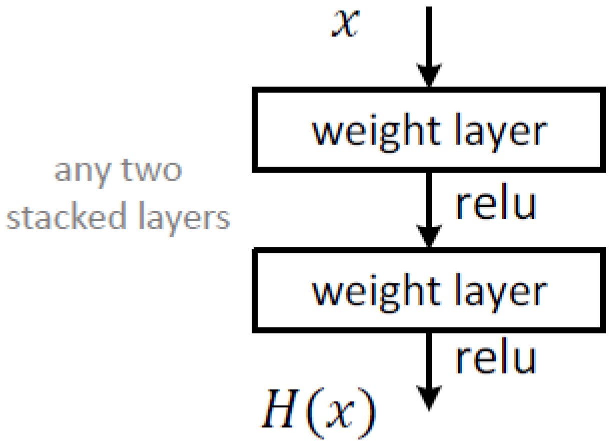
- Deeper networks also maintain the tendency of results.
  - Features in same level will be almost same.
  - Adding layers makes smaller differences.
  - Optimal mappings are closer to an identity map.



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Residual Network…

- Plain block
  - Difficult to make identity mapping because of multiple non-linear layers

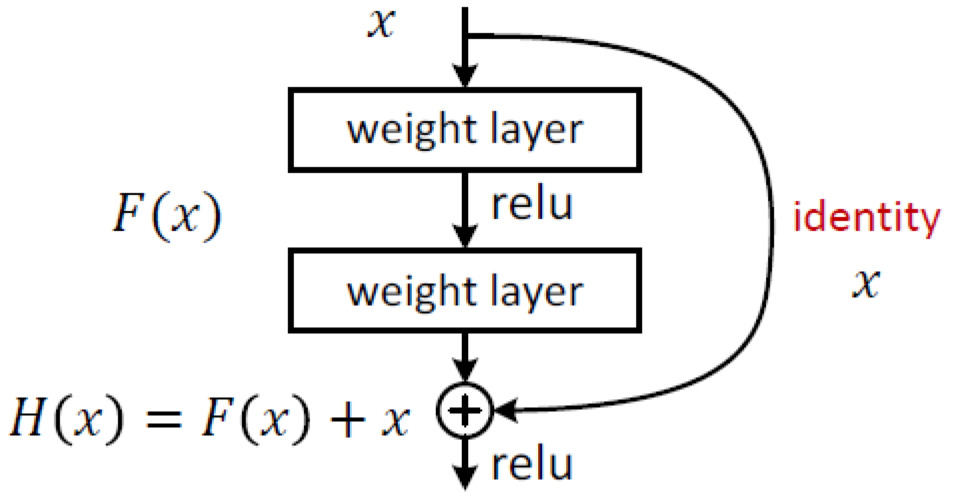


Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Residual Network…

- Residual block
  - If identity were optimal, easy to set weights as 0.
  - If optimal mapping is closer to identity, easier to find small fluctuations.

-> Appropriate for treating perturbation as keeping a base information



Residual Network…

- Let us see how does the Residual Network (ResNet) work.
- Suppose a^l is the input of the $l-th$ layer.
- Recall the output form of the $l-th$ layer.
  - $z^{l+1} = {W^{l+1}}^T a^l + b^{l+1}$
- Activation :
  - $a^{l+1} = g(z^{l+1}) = g({W^{l+1}}^T a^l + b^{l+1})$
  - $a^{l+2} = g(z^{l+2}) = g({W^{l+2}}^T a^{l+1} + b^{l+2})$
- In ResNet it is modified as
  - $a^{l+2} = g(z^{l+2} + a^l) = g({W^{l+2}}^T a^{l+1} + b^{l+2} + a^l)$.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Effect of a Skip Connection

- For the sake of simplicity let us assume that there is identity activation function $g(x) = x$ involved between the layers. Also let us take $b^l = 0$ and use W^l in place of ${W^l}^T$ for the sake of simplicity.
- In ResNet

$$a^{l+2} = z^{l+2} + a^l = W^{l+2}\, a^{l+1} + a^l$$
$$a^{l+4} = z^{l+4} + a^{l+2} = W^{l+4}\, a^{l+3} + W^{l+2}\, a^{l+1} + a^l$$
$$\cdots$$
$$\cdots$$
$$a^L = W^L\, a^{L-1} + W^{L-2}\, a^{L-3} + \cdots + W^{l+2}\, a^{l+1} + a^l$$
$$a^L = a^0 + \sum_{j=1}^{L/2} W^{2j} a^{2j-1}$$

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Effect of a Skip Connection…

- Comparison with Plain Network expression.

$$a^{l+1} = W^{l+1}\, a^{l}$$

$$a^{l+2} = W^{l+2}\, a^{l+1} = W^{l+2}\, W^{l+1}\, a^{l}$$

$$\cdots$$

$$\cdots$$

$$a^{L} = W^{L} \cdots W^{l+2}\, W^{l+1}\, a^{l}$$

$$a^{L} = \prod_{j=1}^{L} (W^{j}) a^{0}$$

ResNet Recurrence

$$a^{L} = a^{0} + \sum_{j=1}^{L/2} W^{2j} a^{2j-1}$$

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Effect of a Skip Connection…

- Two main advantages: The feature a^L of a deeper layer can be expressed as the feature a^l of a shallower layer plus a residual term.

$$a^L = a^l + \sum_{j=l}^{L/2} W^{2j} a^{2j-1}$$

- The feature a^L of a deeper layer is the summation of residual terms plus the initial input term a^0.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Effect of a Skip Connection…

- Imagine a situation when all the weights are zero in the following equation.

$$a^L = a^l + \sum_{j=l}^{L/2} W^{2j} a^{2j-1}$$

- That means the network has perfectly reconstructed the identity mapping.

- This is the basis of considering residual network to improve training time and performance.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Effect of a Skip Connection…

- Backward propagation in ResNet:

  If L is the loss function.

$$\frac{\partial L}{\partial a^l} = \frac{\partial L}{\partial a^L} \cdot \frac{\partial a^L}{\partial a^l}$$

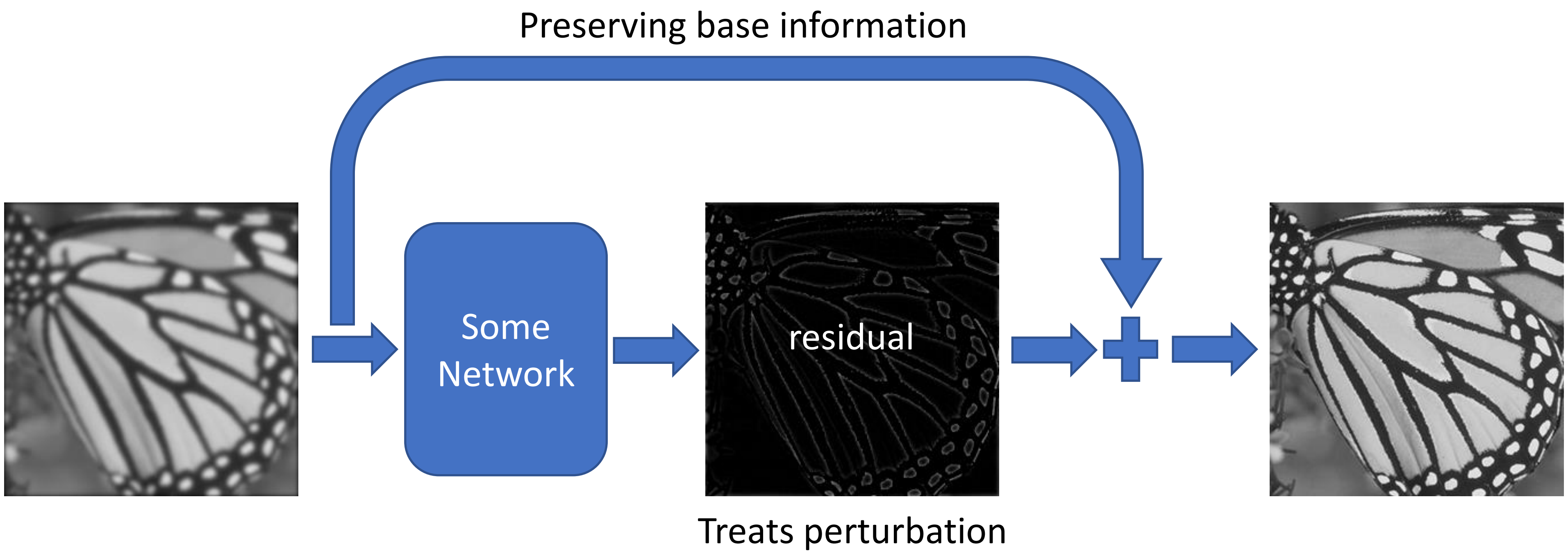
$$\frac{\partial L}{\partial a^L} = \frac{\partial L}{\partial a^L}(1 + \frac{\partial}{\partial a^l} \sum_{j=l}^{L/2} W^{2j} a^{2j-1})$$

- This efficiently eliminates the gradient descent problem .
  - Any backpropagation ensures that the information $\frac{\partial L}{\partial a^L}$ directly back propagates to the $l-th$ layer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

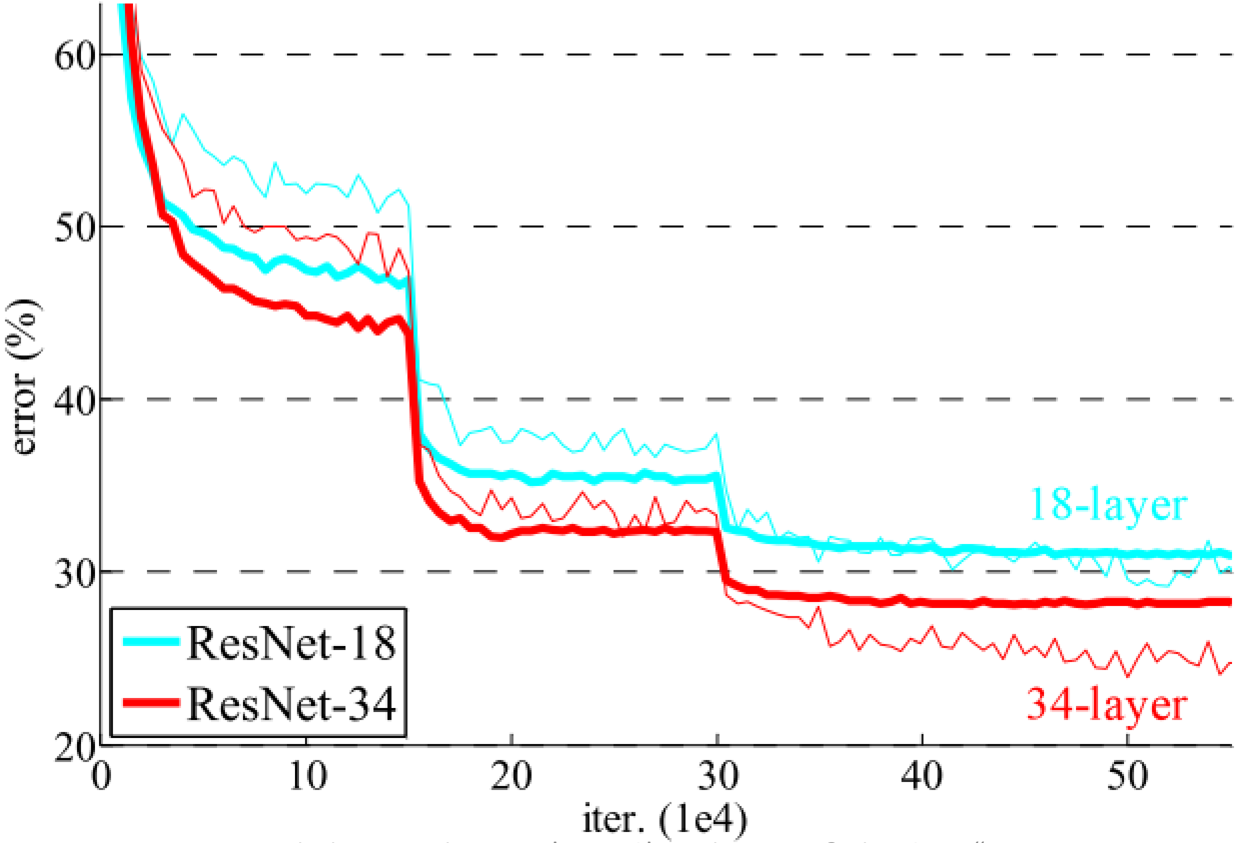
Residual Network…

- Difference between an original image and a changed image.



Residual Network…

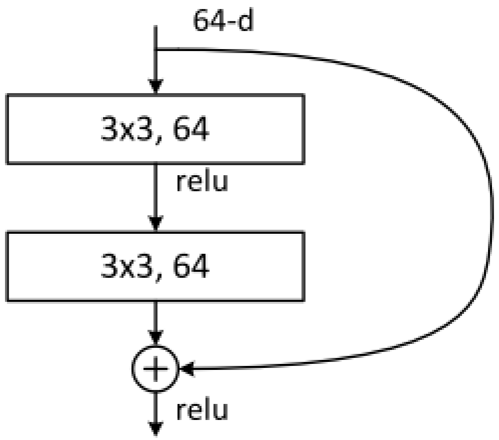
- Deeper ResNets have lower training error



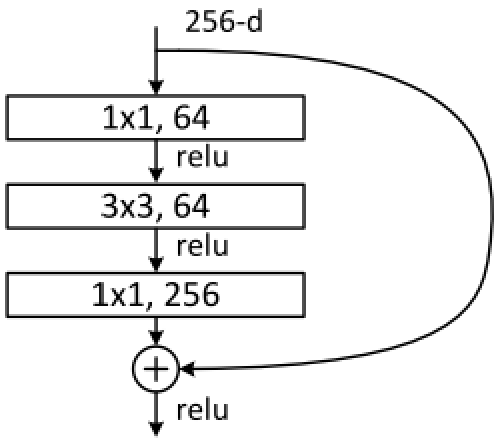
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Residual Network…

- Residual block
  - Very simple
  - Parameter-free



A naïve residual block



"bottleneck" residual block
(for ResNet-50/101/152)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Residual Network…

- Shortcuts connections
  - Identity shortcuts : $x^{l+1} = F\left(x^l, \{W_{l+1}\}\right) + x^l$
  - Projection shortcuts : $x^{l+1} = F\left(x^l, \{W_{l+1}\}\right) + W_s x^l$
- Why projection shortcuts ?
  - To match the dimension of the output $F(x, \{W_{l+1}\})$ from the residual weight layers.
  - For example if $F(x, \{W_{l+1}\})$ is of dimension 256 and initial x is of dimension 128. Then W_s projects x to make it a vector of dimension 256.
  - Hence W_s is of dimension 256×128.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

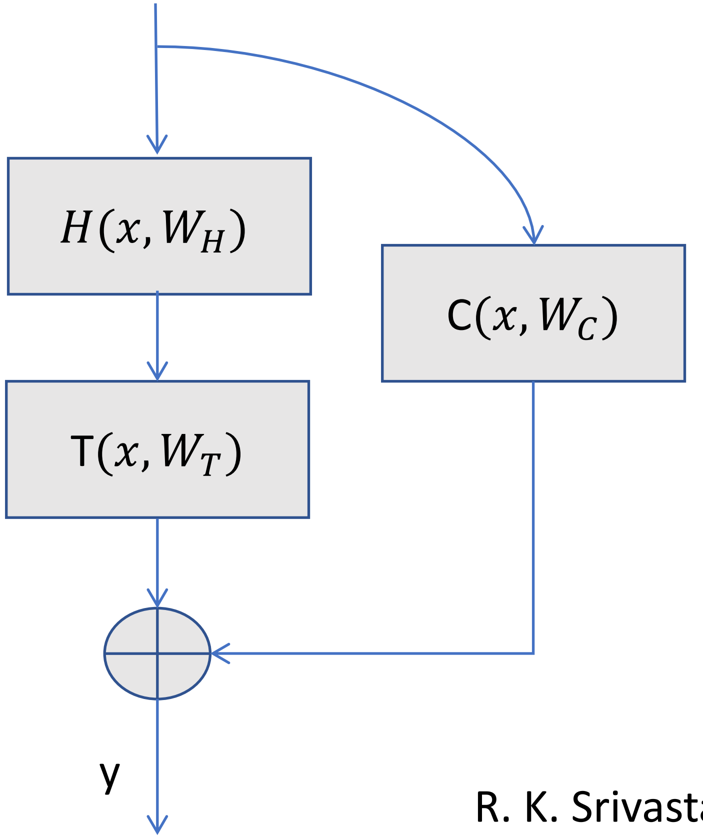
Identity Mapping in ResNet

- Identity shortcuts are defined by
  - $x^{l+1} = F\left(x^l, \{W_{l+1}\}\right) + x^l$
- Performance of ResNet with other shortcuts is also studied by He et al. (2016) where x is replaced by $h(x)$.
- Such shortcut connections are defined by
  - $x^{l+1} = F\left(x^l, \{W_{l+1}\}\right) + h(x^l)$
  - $y^{l+1} = f(x^{l+1})$
- It is reported that identity skip connections (shortcuts) achieves the fastest error reduction and lowest training loss among all variants (scaling, gating) convolutions.
- Other choices lead to higher training loss and error.
- These experiments suggest that keeping a “clean" information path is helpful for easing optimization.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Identity Mappings in Deep Residual Networks”. arXiv 2016.

Similar Architecture – Highway Net

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot C(\mathbf{x}, \mathbf{W_C}).$$



R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. arXiv:1505.00387, 2015.

Highway Net vs. ResNet

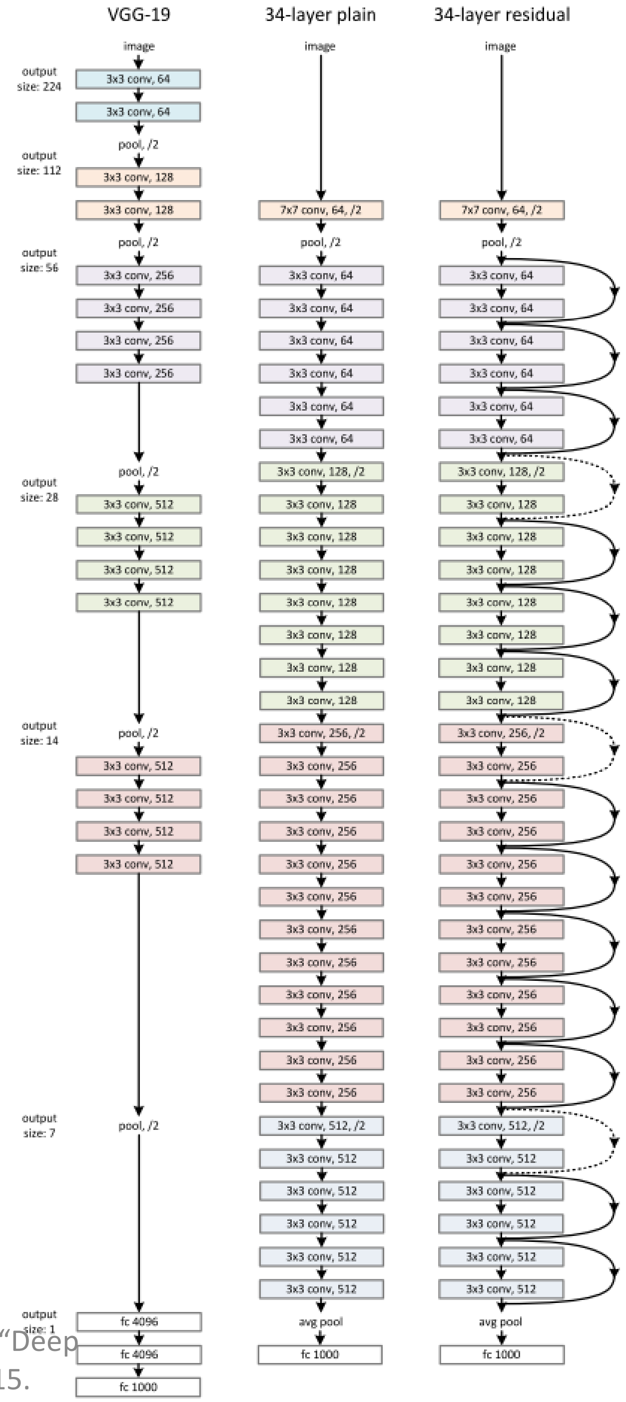
- C and T are data dependent
- Both the gates have parameters.
- When a gated shortcut is “closed” the layers in highway networks represent non-residual functions.
- High-2 way networks have not demonstrated accuracy gains with depth of over 100 layers.

R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. arXiv:1505.00387, 2015.

ResNet Design

- Basic design (VGG-style)
  - All 3x3 conv (almost)
  - Spatial size/2 => #filters x2
  - Batch normalization
  - Simple design, just deep

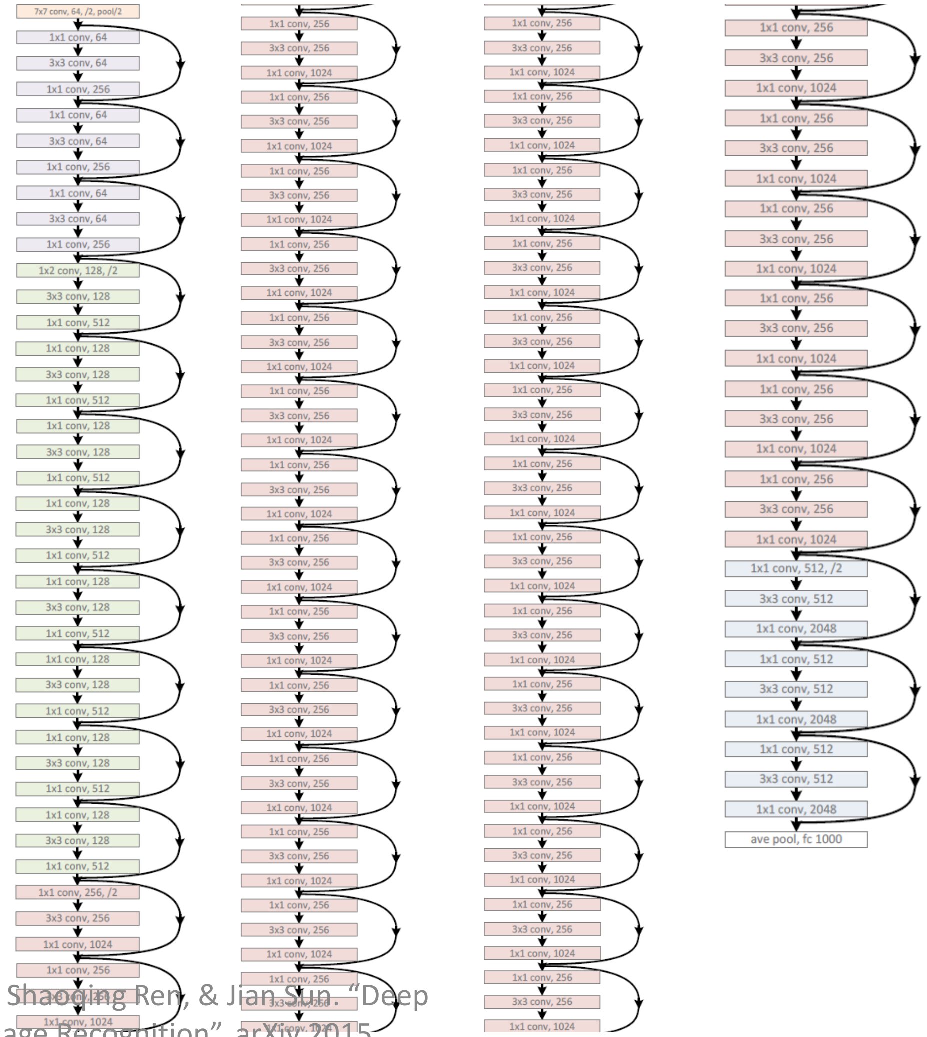
- Other remarks
  - No max pooling (almost).
  - No hidden fully connected layers.
  - No dropout.



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Network Design

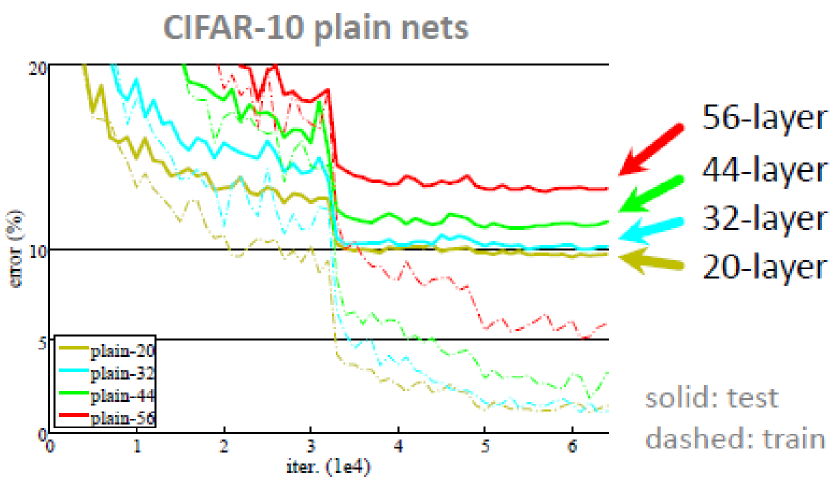
- ResNet-152
  - Use bottlenecks.
  - ResNet-152(11.3 billion FLOPs) has lower complexity than VGG-16/19 nets (15.3/19.6 billion FLOPs).

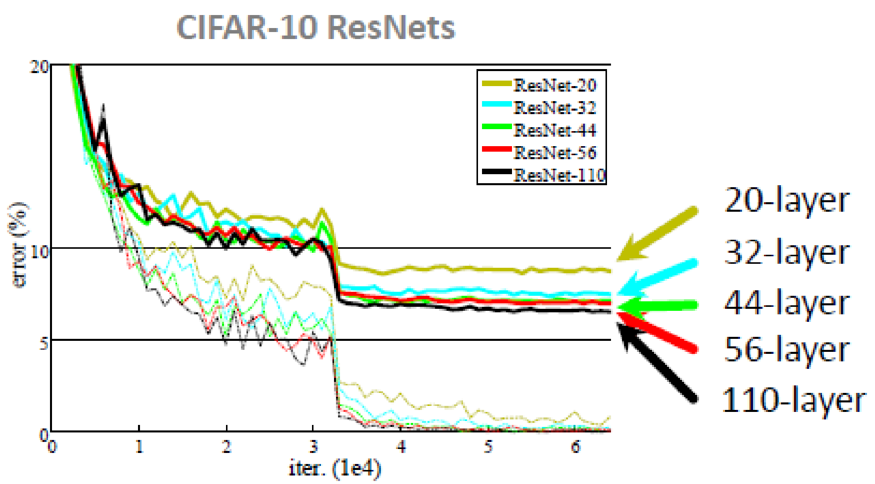


Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Results

- Deep Resnets can be trained without difficulties.
- Deeper ResNets have lower training error, and also lower test error.





Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

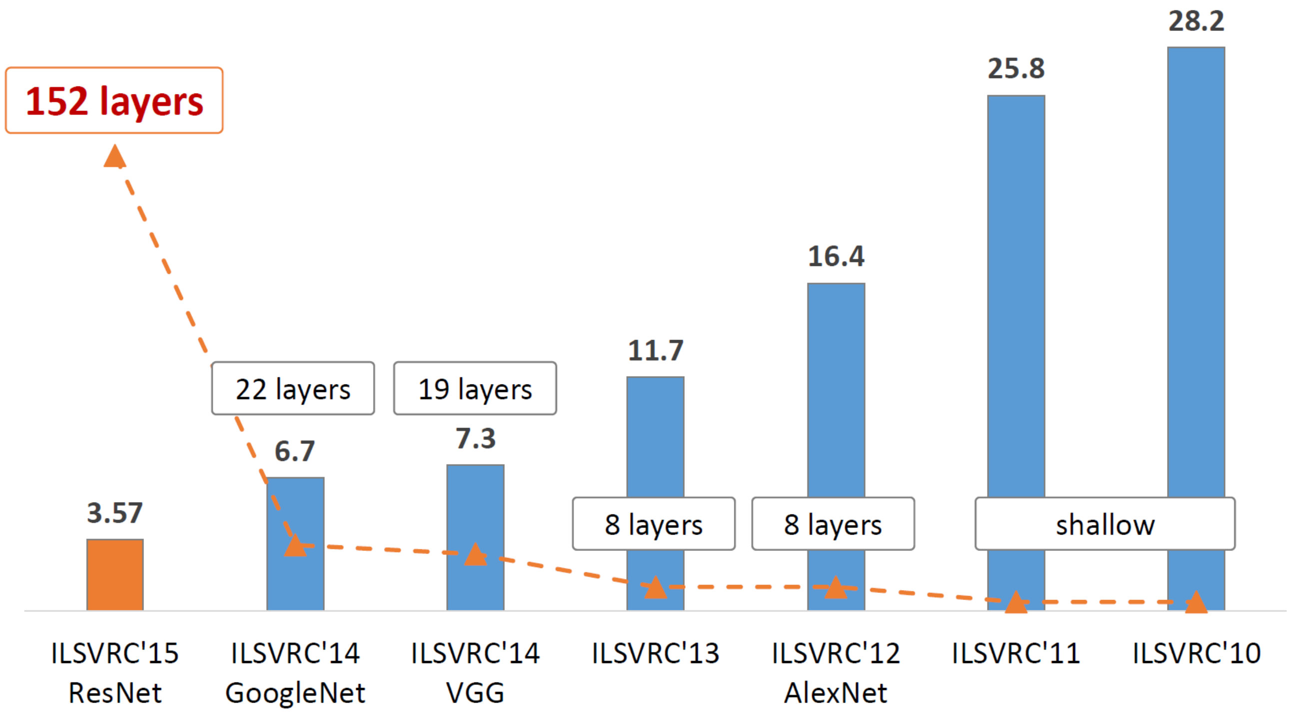
Results

- 1st places in all five main tracks in “ILSVRC & COCO 2015 Competitions”
  - ImageNet Classification
  - ImageNet Detection
  - ImageNet Localization
  - COCO Detection
  - COCO Segmentation

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Quantitative Results

- ImageNet Classification



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Result

- Performances increase absolutely.

task	2nd-place winner	MSRA	margin (relative)
ImageNet Localization (top-5 error)	12.0	9.0	**27%**
ImageNet Detection (mAP@.5)	53.6	62.1	**16%**
COCO Detection (mAP@.5:.95)	33.5	37.3	**11%**
COCO Segmentation (mAP@.5:.95)	25.1	28.2	**12%**

absolute 8.5% better!

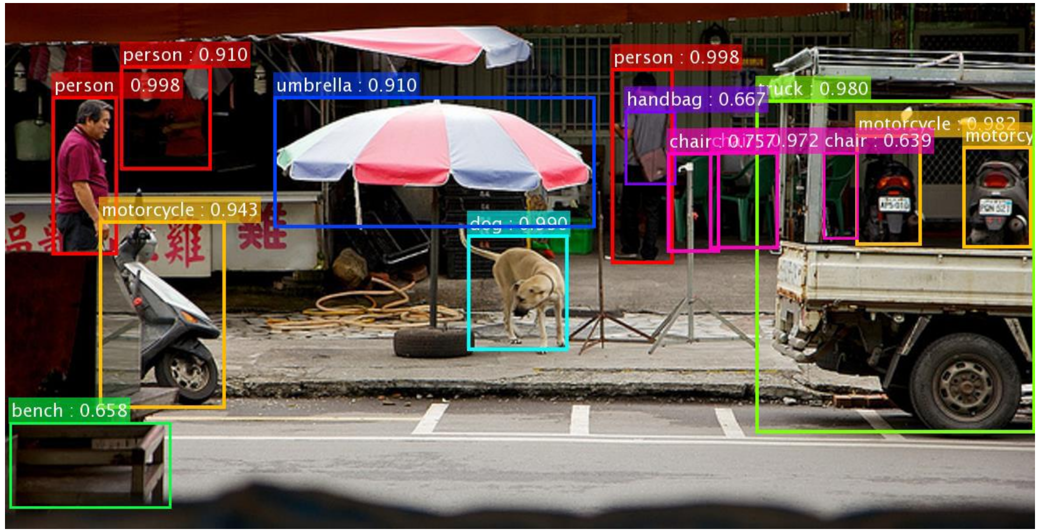
- Based on ResNet-101
- Existing techniques can use residual networks or features from it

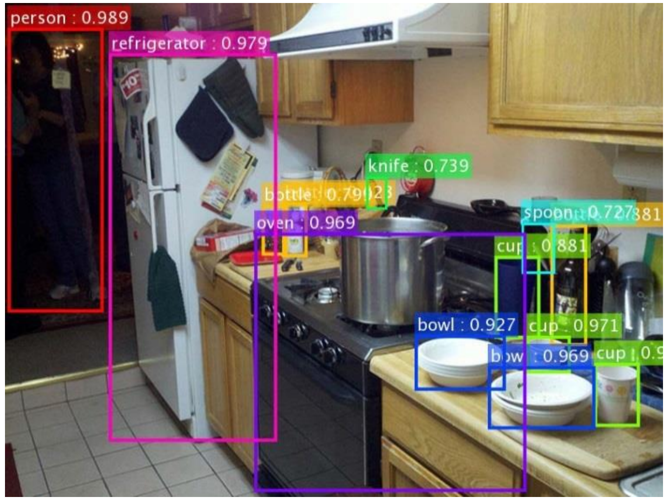
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Qualitative Result

- Object detection
  - Faster R-CNN + ResNet

80 Object Categories
More than 300,000 images

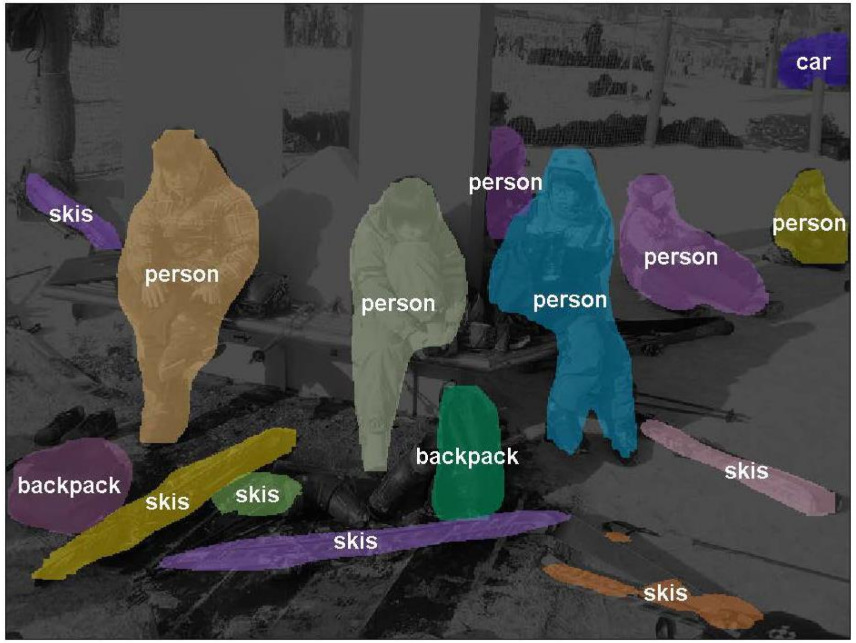




Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Jifeng Dai, Kaiming He, & Jian Sun. “Instance-aware Semantic Segmentation via Multi-task Network Cascades”. arXiv 2015.

Qualitative Results

MS COCO –Instance Segmentation



Qualitative Results

ImageNet - Classification,
Localization & Detection

200 object classes **456,567 images** **DET**
1000 object classes **1,431,167 images** **CLS-LOC**

Flute

Strawberry

Traffic light

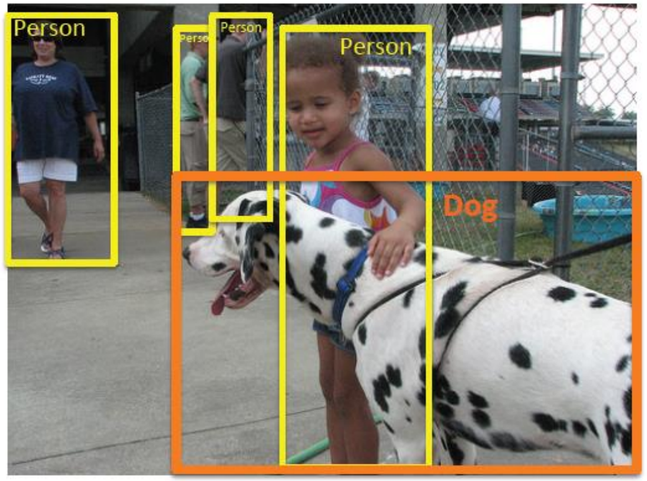
Backpack

Matchstick

Sea lion

Bathing cap

Racket



http://image-net.org/challenges/LSVRC/

Further Deep Residual Network

- 1202 layer network was also explored, but with the dataset mentioned in previous slides, its test set error was showing increase in error as compared to 110 layer network.

- Probably due to overfitting.

- Data set size was not sufficient to train such a high sized network.

Conclusion

- ResNet training is simple and computationally convenient.
- Accuracy improved.

- New versions
  - Wide ResNet
  - ResNext

Reference

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
- Slides of Deep Residual Learning @ ILSVRC & COCO 2015 competitions
- Slides of Hyeonwoo Noh, Pohang University of Science and Technology on ResNet.