Deep Residual Learning for Image Recognition

—— Yidan Mei, Yinglei Xu

Deep residual learning for **image recognition**

K He, X Zhang, S Ren, J Sun - … and **pattern recognition**, 2016 - openaccess.thecvf.com

… **Deeper** neural **networks** are more difficult to train. We present a **residual learning** framework to ease the training of **networks** that are substantially **deeper** than those used previously. …

☆ Save 99 Cite Cited by 234341 Related articles All 65 versions »

[PDF] thecvf.com

Revolution of Deep Neural Network

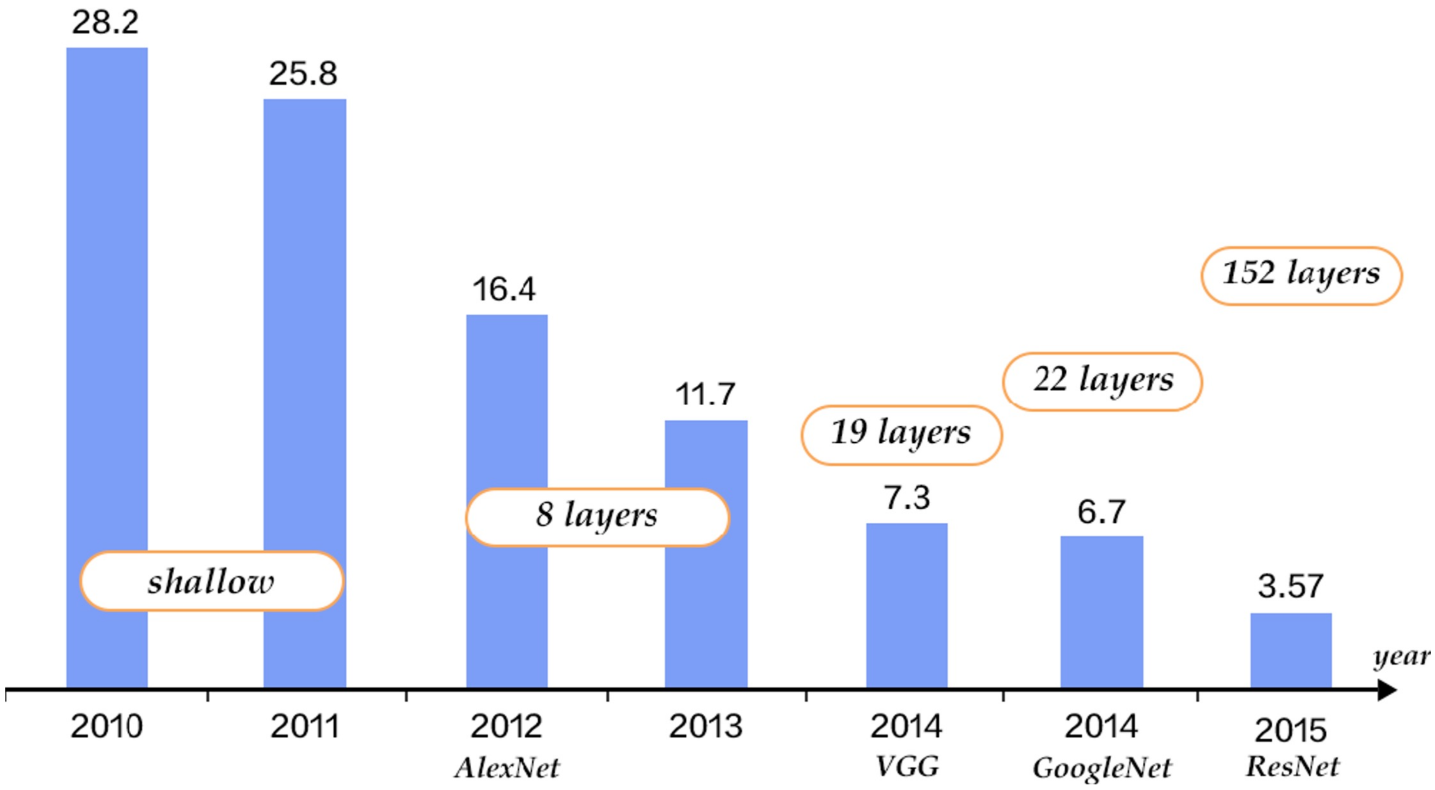
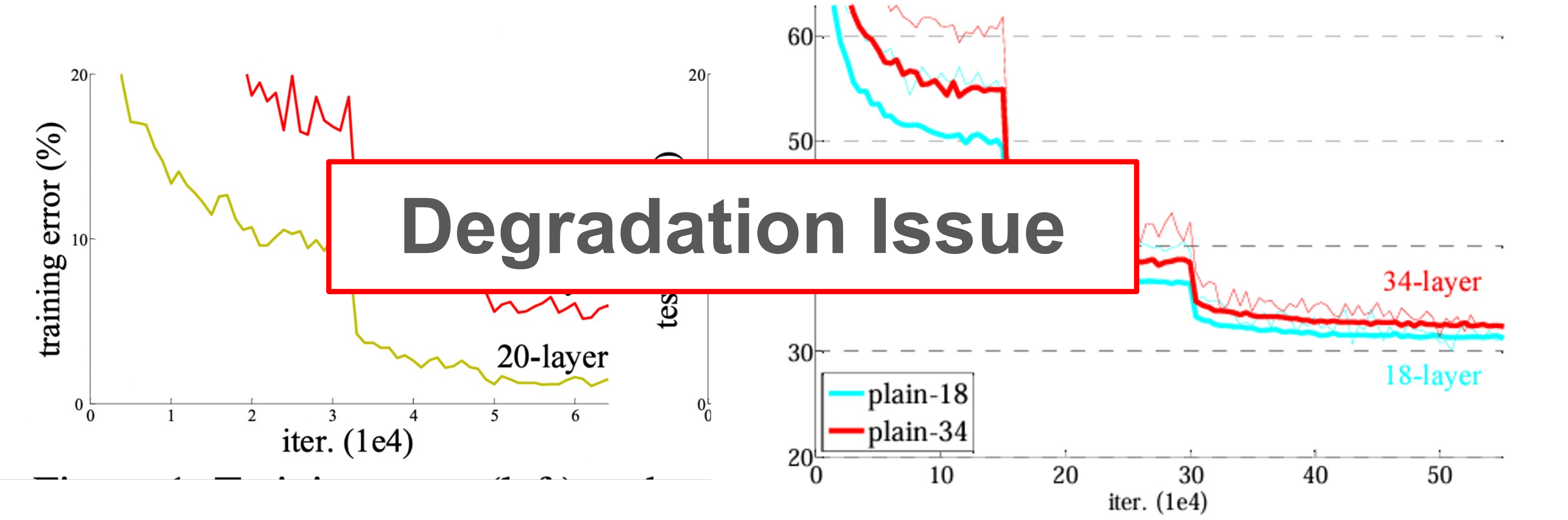


Image Source: http://paddlepaddle.org/

Big Question:
Deeper Networks = Better Performance? NO!



Degradation Issue

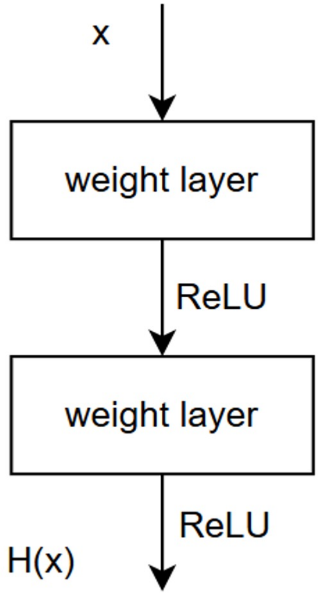
Although solution space of the 18-layer one is a subset of that of the 34-layer one, the deeper network shows higher training error & validation error.

Reasons

- Representational ability? No, deeper networks' solution space include that of shallower networks.
- Overfitting? No, training error also larger.
- Vanishing gradients? No, using BN will prevent it.
- Optimization Difficulty
  - deep plain nets have exponentially low convergence rate → impact the reducing of the training error.

ResNet Architecture: two stacked layers

Plain Network



Fit H(x)

Residual Network

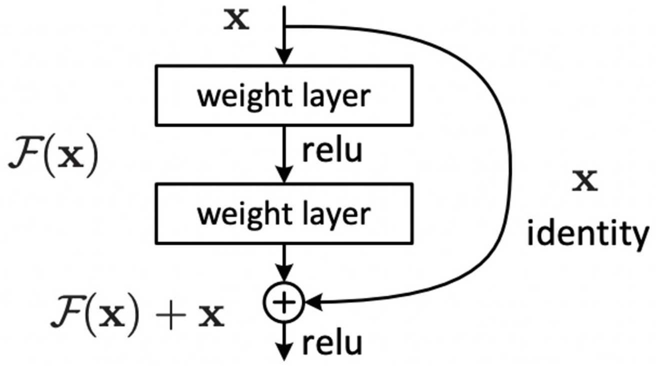


Figure 2. Residual learning: a building block.

Fit residual F(x):= H(x) - x,
then recast by H(x) = F(x) + x

Why ResNet works?

Hypothesize: Easier to optimize the residual mapping than the original H(x).

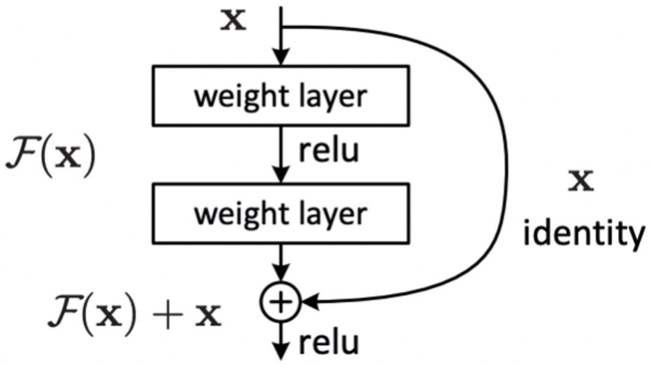
- If optimal mapping is H(x) = x, pushing the residual mapping F(x) to 0 will be easier than using two layers to fit H(x)
- more info can be found in https://arxiv.org/abs/1603.05027

Others also said something about weight initialization using Gaussian distribution → hard to fit identity

Skipping those identity mapping layers → work similarly as a shallower network

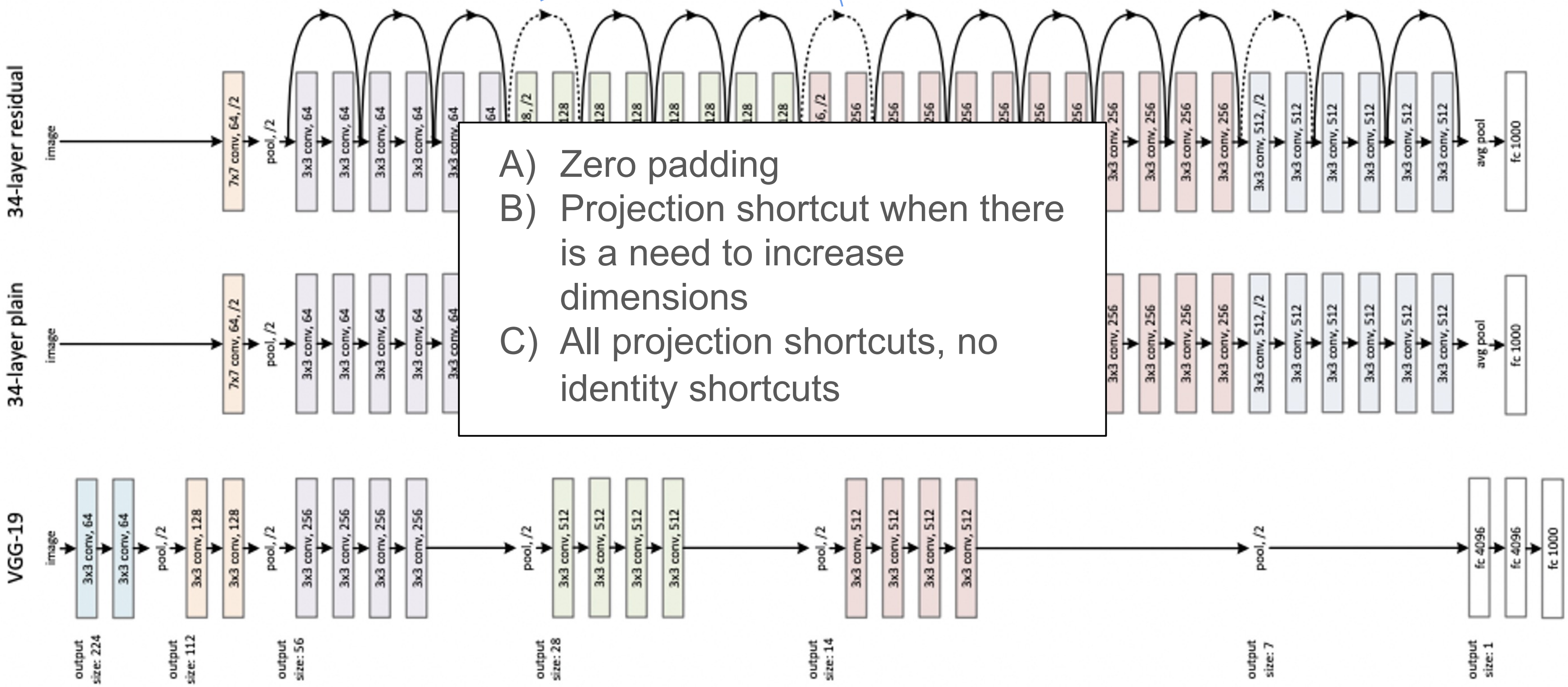
Shortcut Connection

- shortcut connections: are those skipping one or more layers. (e.g., the shortcut connections simply perform identity mapping(X → X) in the picture
  1. Identity shortcut: x, F same dims
     $$\mathbf{y} = \mathcal{F}(x, \{W_i\}) + \mathbf{x}$$
  1. Projection shortcut: x , F different dims
     $$\mathbf{y} = \mathcal{F}(x, \{\mathcal{W}_i\}) + \mathcal{W}_s\mathbf{x}$$
- Will compare different shortcut options in Results
- if F has only 1 layer: similar to linear layer
  y = W_1 x + x, NO advantages
  $$\mathbf{y} = W_i\mathbf{x} + \mathbf{x}$$



Architecture

Increase dimensions by option A/B/C



A) Zero padding
B) Projection shortcut when there is a need to increase dimensions
C) All projection shortcuts, no identity shortcuts

Constructing Deeper Layers: Bottleneck Building Block

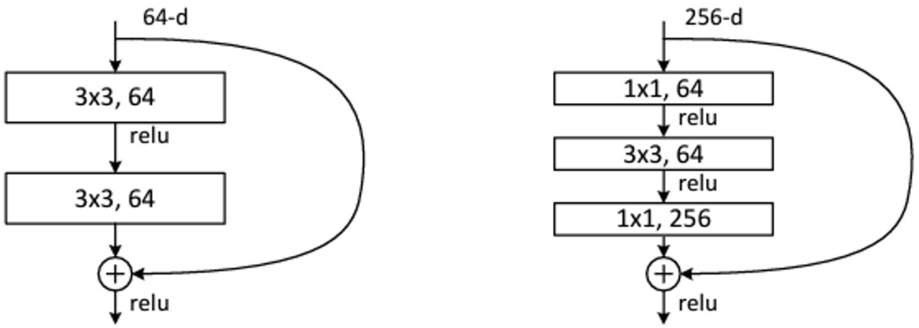


Figure 5. A deeper residual function $\mathcal{F}$ for ImageNet. Left: a building block (on 56×56 feature maps) as in Fig. 3 for ResNet-34. Right: a "bottleneck" building block for ResNet-50/101/152.

- Reason: limited training time authors could afford
- ResNet-50: replace each 2-layer block in ResNet-34 with the 3-layer bottleneck block.
- Parameter-free identity mapping is important in bottleneck

Results

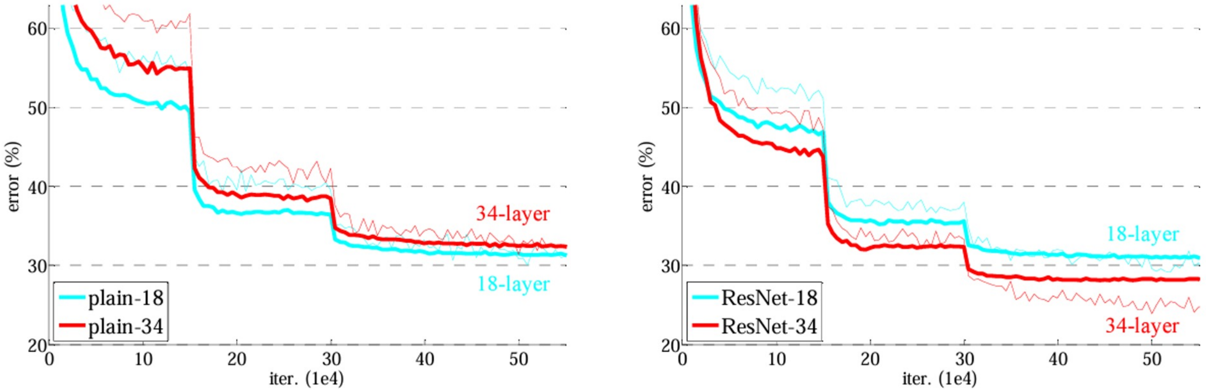


Figure 4. Training on **ImageNet**. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

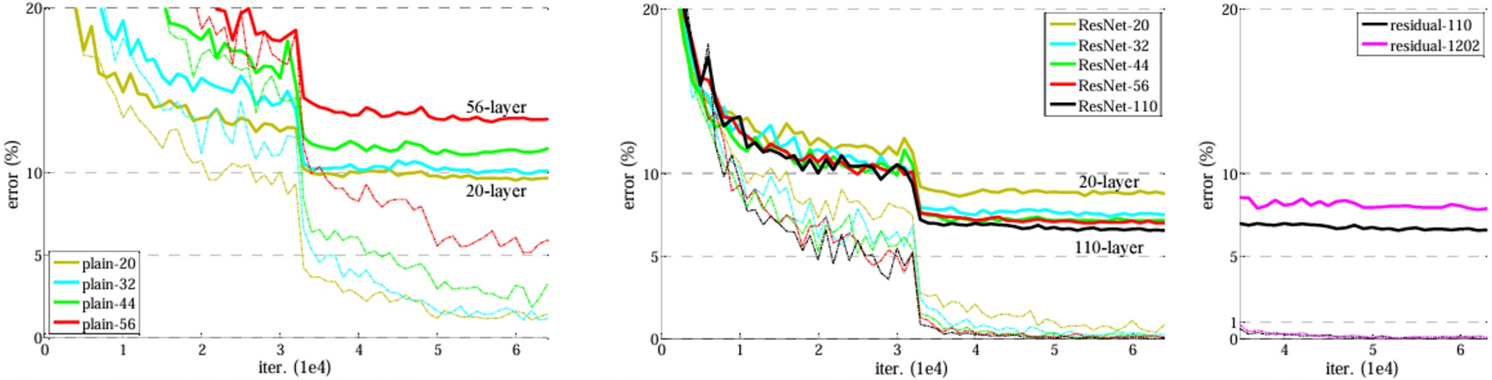


Figure 6. Training on **CIFAR-10**. Dashed lines denote training error, and bold lines denote testing error. **Left**: plain networks. The error of plain-110 is higher than 60% and not displayed. **Middle**: ResNets. **Right**: ResNets with 110 and 1202 layers.

Results Continued

model	top-1 err.
VGG-16 [41]	28.07
GoogLeNet [44]	-
PReLU-net [13]	24.27
plain-34	28.54
ResNet-34 A	25.03
ResNet-34 B	24.52
ResNet-34 C	24.19
ResNet-50	22.85
ResNet-101	21.75
ResNet-152	**21.43**

Table 3. Error rates (%, **10-crop** testing) o
VGG-16 is based on our test. ResNet-50/1
that only uses projections for increasing di

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
	21.99	5.81
	21.84	5.71
	21.53	5.60
	20.74	5.25
	19.87	4.60
	19.38	**4.49**

lel results on the ImageNet
test set).

method	# layers	# params	error (%)
Maxout [10]			9.38
NIN [25]			8.81
DSN [24]			8.22
FitNet [35]	19	2.5M	8.39
Highway [42, 43]	19	2.3M	7.54 (7.72±0.16)
Highway [42, 43]	32	1.25M	8.80
ResNet	20	0.27M	8.75
ResNet	32	0.46M	7.51
ResNet	44	0.66M	7.17
ResNet	56	0.85M	6.97
ResNet	110	1.7M	**6.43** (6.61±0.16)
ResNet	1202	19.4M	7.93

Table 6. Classification error on the **CIFAR-10** test set. All methods are with data augmentation. For ResNet-110, we run it 5 times and show "best (mean±std)" as in [43].

	top-5 err. (**test**)
	7.32
	6.66
	6.8
	4.94
	4.82
	3.57

s. The top-5 error is on the
e test server.

Compare different sho
options

- projection shortcut not essential for degradation

ensemble results (152)

Fewer parameters than other networks

Application of ResNet

- Potential Application
  - Visual Recognition
  - Image Generation (Pixel RNN, Neural Art, etc.
  - Natural Language Processing (Very deep CNN)
  - Speech Recognition
  - Advertising, user prediction

Main takeaways:

- Residual networks consistently outperformed plain networks, especially as depth increased.
- ResNet models effectively addressed the degradation problem, enabling deeper architectures to maintain or improve accuracy.
- ResNet has fewer parameters and lower complexity than other networks
- The ResNet framework generalizes well across various tasks and datasets, including classification, detection, and localization. (Table 7 & 8 if interested)

ResNet @ ILSVRC & COCO 2015 Competitions

1st places in all five main tracks

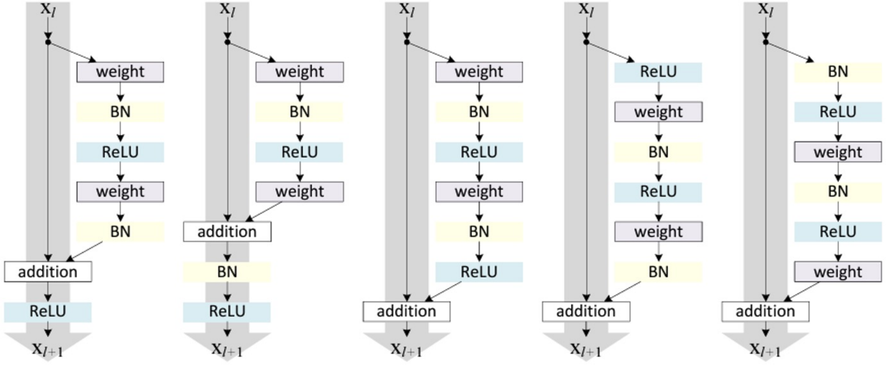
- ImageNet Classification: *"Ultra-deep"* 152-layer nets
- ImageNet Detection: 16% better than 2nd
- ImageNet Localization: 27% better than 2nd
- COCO Detection: 11% better than 2nd
- COCO Segmentation: 12% better than 2nd

Why Identity Mapping? (https://arxiv.org/abs/1603.05027)

- shortcut connections are the most direct paths for the information to propagate.
- Other manipulations (scaling, gating…) on the shortcuts hampers the propagation → optimization problems
- Also proposed a pre-activation design → decrease test error for deeper networks

Table 2. Classification error (%) on the CIFAR-10 test set using different activation functions.

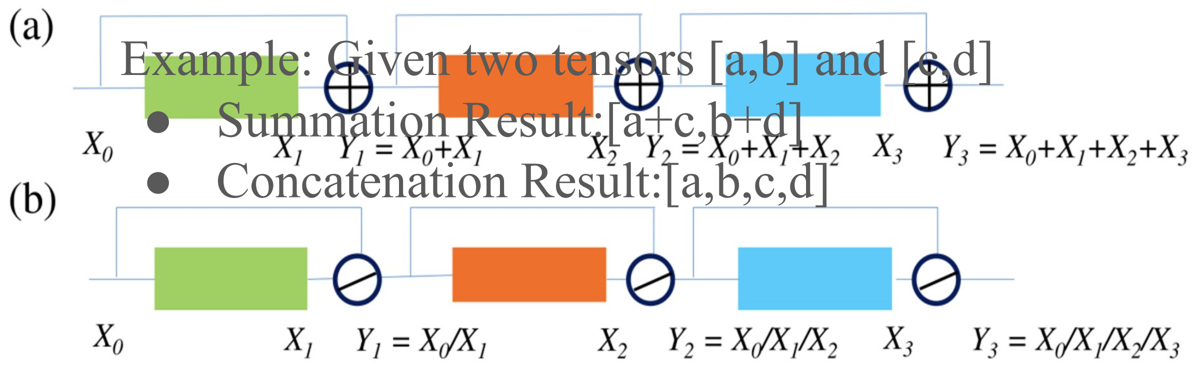
case	Fig.	ResNet-110	ResNet-164
original Residual Unit [1]	Fig. 4(a)	6.61	5.93
BN after addition	Fig. 4(b)	8.17	6.50
ReLU before addition	Fig. 4(c)	7.84	6.14
ReLU-only pre-activation	Fig. 4(d)	6.71	5.91
full pre-activation	Fig. 4(e)	**6.37**	**5.46**



(a) original (b) BN after addition (c) ReLU before addition (d) ReLU-only pre-activation (e) **full pre-activation**

New State-of-Art? ResNet vs. DenseNet

- Drawback of identity shortcuts: limit representation capacity
- ResNet (a) uses summation (identity shortcuts), while DenseNet uses concatenation (dense connections).



Example: Given two tensors [a,b] and [c,d]

- Summation Result:[a+c,b+d]
- Concatenation Result:[a,b,c,d]

Thank You!

Reference

- https://www.youtube.com/watch?v=C6tLw-rPQ2o
- https://github.com/KaimingHe/deep-residual-networks?tab=readme-ov-file
- **https://medium.com/visionwizard/object-segmentation-4fc67077a678**
- https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f
- https://www.researchgate.net/publication/374484296_A_Fruit_Ripeness_Detection_Method_using_Adapted_Deep_Learning-based_Approach#pf3
- https://arxiv.org/abs/1603.05027
- https://openaccess.thecvf.com/content/WACV2021/papers/Zhang_ResNet_or_DenseNet_Introducing_Dense_Shortcuts_to_ResNet_WACV_2021_paper.pdf