

# CS5242 Lecture 6

Sep 27th, 2022

YOU, Yang

# Outline

- Part 1: Deep CNN Architectures
  - We focus on ResNet
- Part 2: A Real-World Application of Deep Learning
  - Crafting Better Contrastive Views for Siamese Representation Learning
    - Self-Supervised Learning
    - Suggestions for your future research directions or engineering solutions



Yann LeCun  
@ylecun

...

1. Self-Supervised Learning
2. ResNets (not intellectually deep, but useful)
3. Gating -> Attention -> Dynamic connection graphs.
4. Differentiable memory.
5. Permutation-equivariant modules, e.g. multihead self-attention -> Transformers.



David Chalmers @davidchalmers42 · Sep 13

what are the most important intellectual breakthroughs (new ideas) in AI in the last ten years?

[Show this thread](#)

11:14 PM · Sep 15, 2022 · Twitter for Android

---

295 Retweets 12 Quote Tweets 1,748 Likes

## Why should we study ResNet and Self-Supervised Learning?

Slides Credits: from Dr. Kaiming He

# Overview

- Introduction
- Background
  - From shallow to deep
- Deep Residual Networks
  - From 10 layers to 100 layers
  - From 100 layers to 1000 layers
- Applications

# Introduction

# Introduction

## Deep Residual Networks (ResNets)

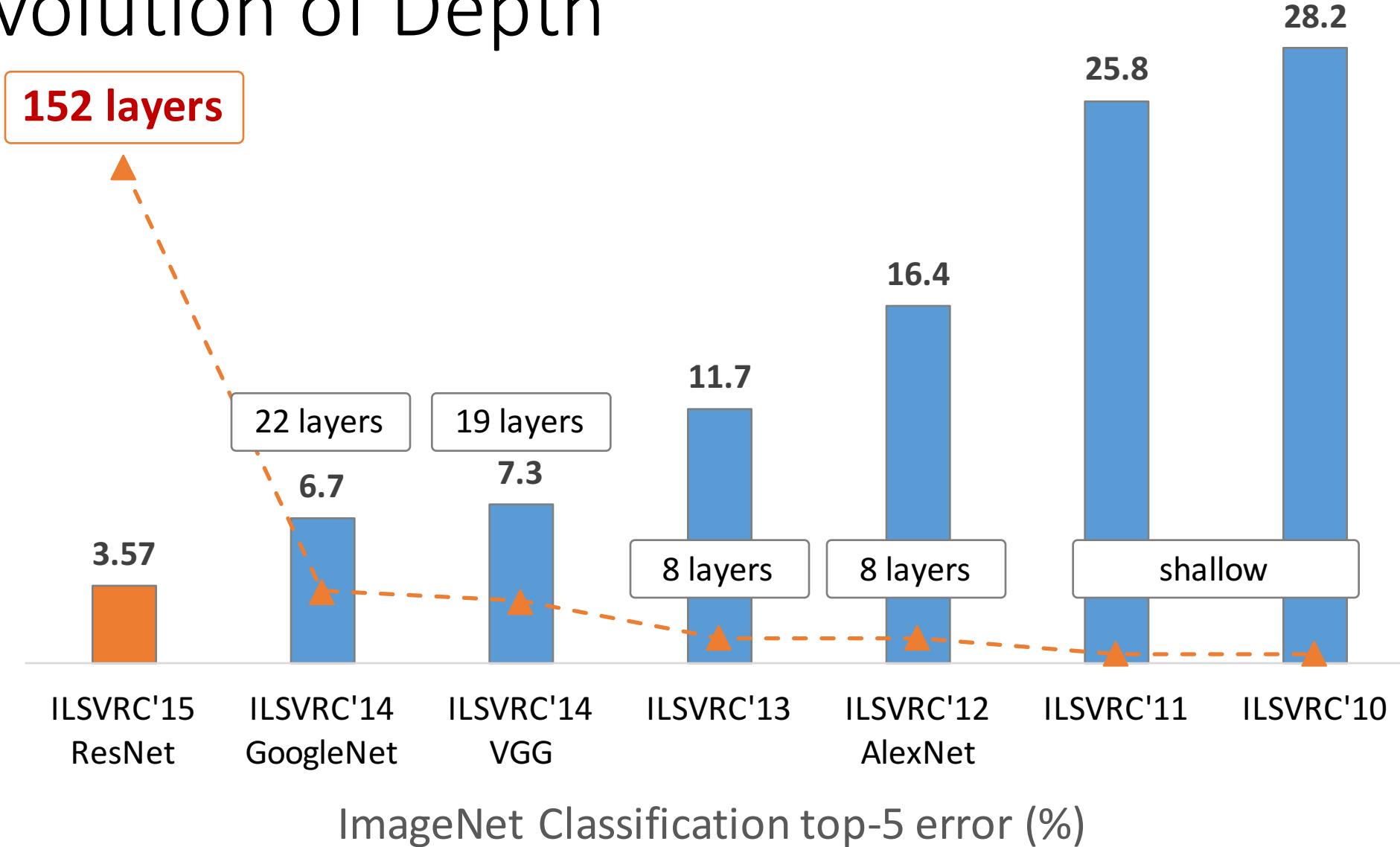
- “Deep Residual Learning for Image Recognition”. CVPR 2016
- A simple and clean framework of training “very” deep nets
- State-of-the-art performance for
  - Image classification
  - Object detection
  - Semantic segmentation
  - and more...

# 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

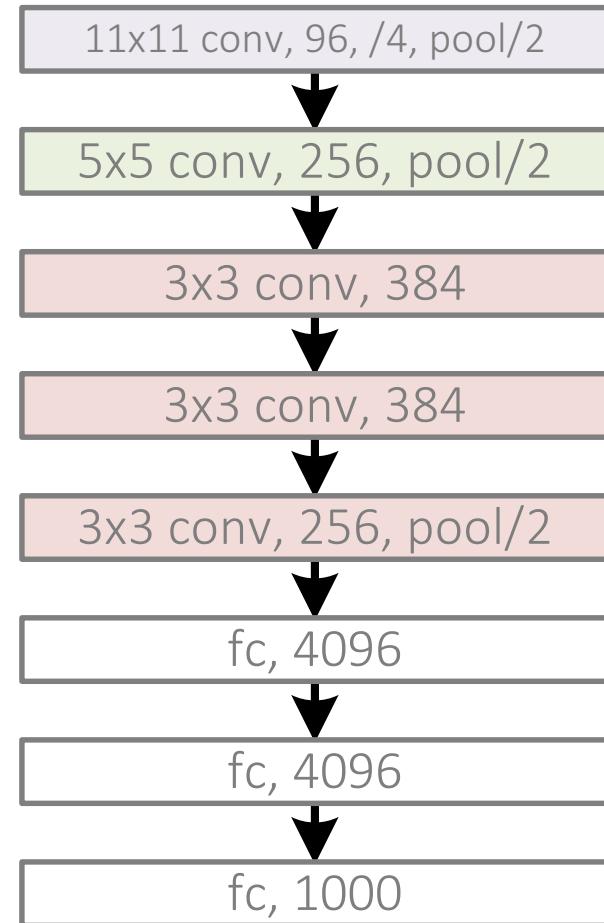
\*improvements are relative numbers

# Revolution of Depth



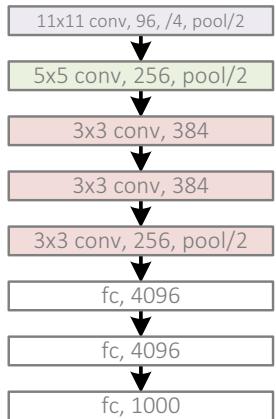
# Revolution of Depth

AlexNet, 8 layers  
(ILSVRC 2012)

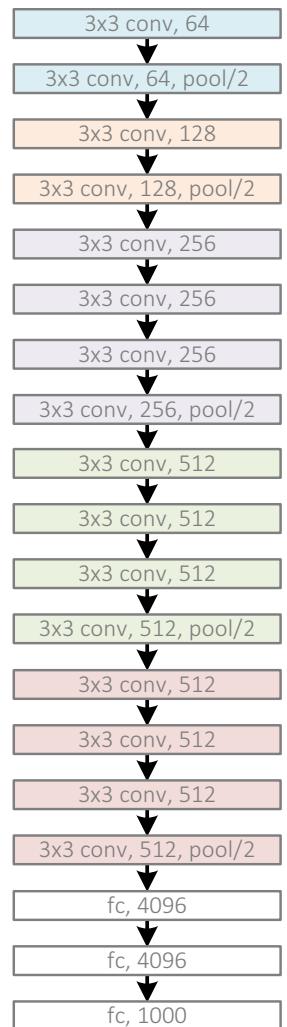


# Revolution of Depth

AlexNet, 8 layers  
(ILSVRC 2012)



VGG, 19 layers  
(ILSVRC 2014)



GoogleNet, 22 layers  
(ILSVRC 2014)



# Revolution of Depth

AlexNet, 8 layers  
(ILSVRC 2012)



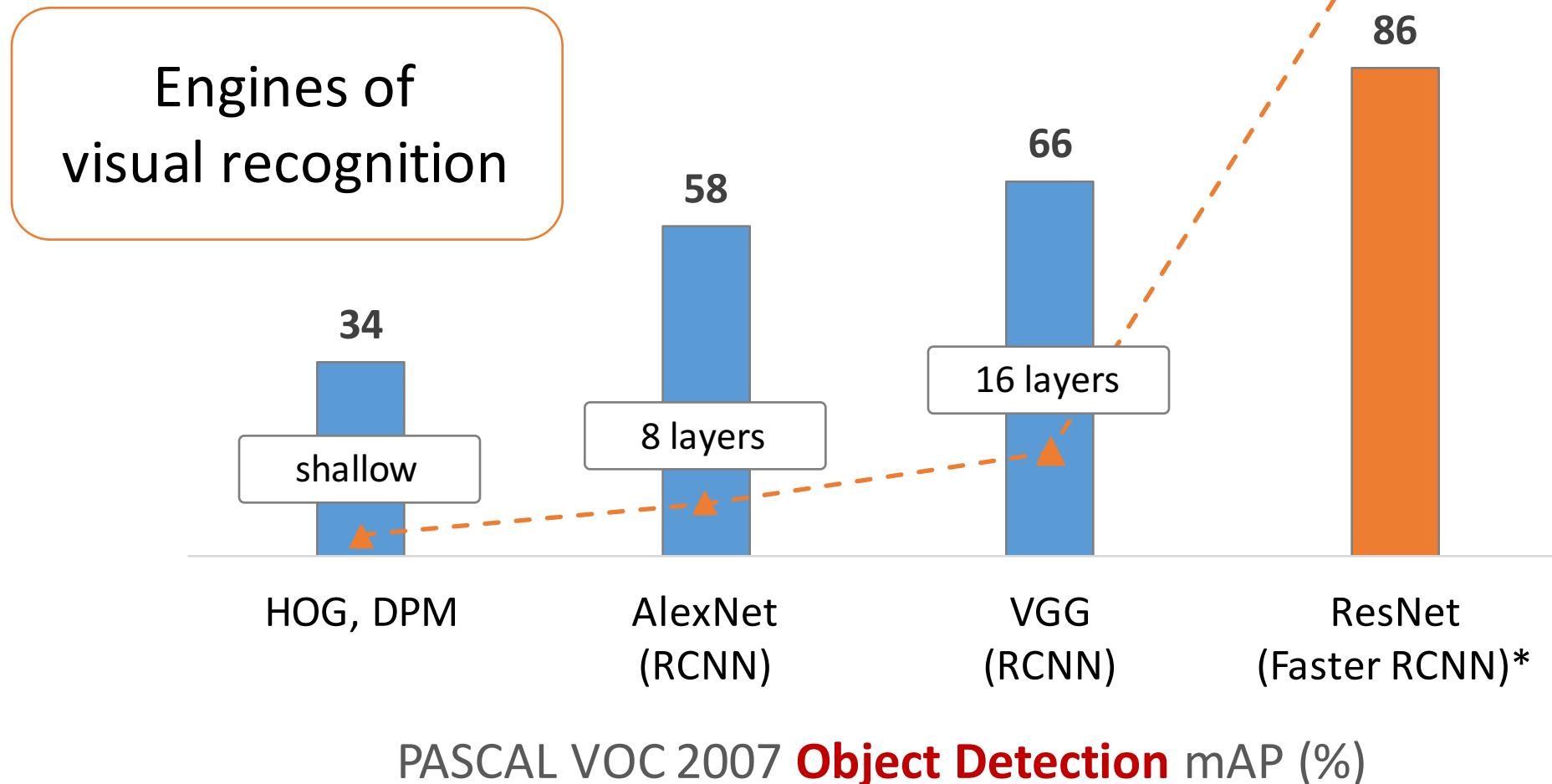
VGG, 19 layers  
(ILSVRC 2014)



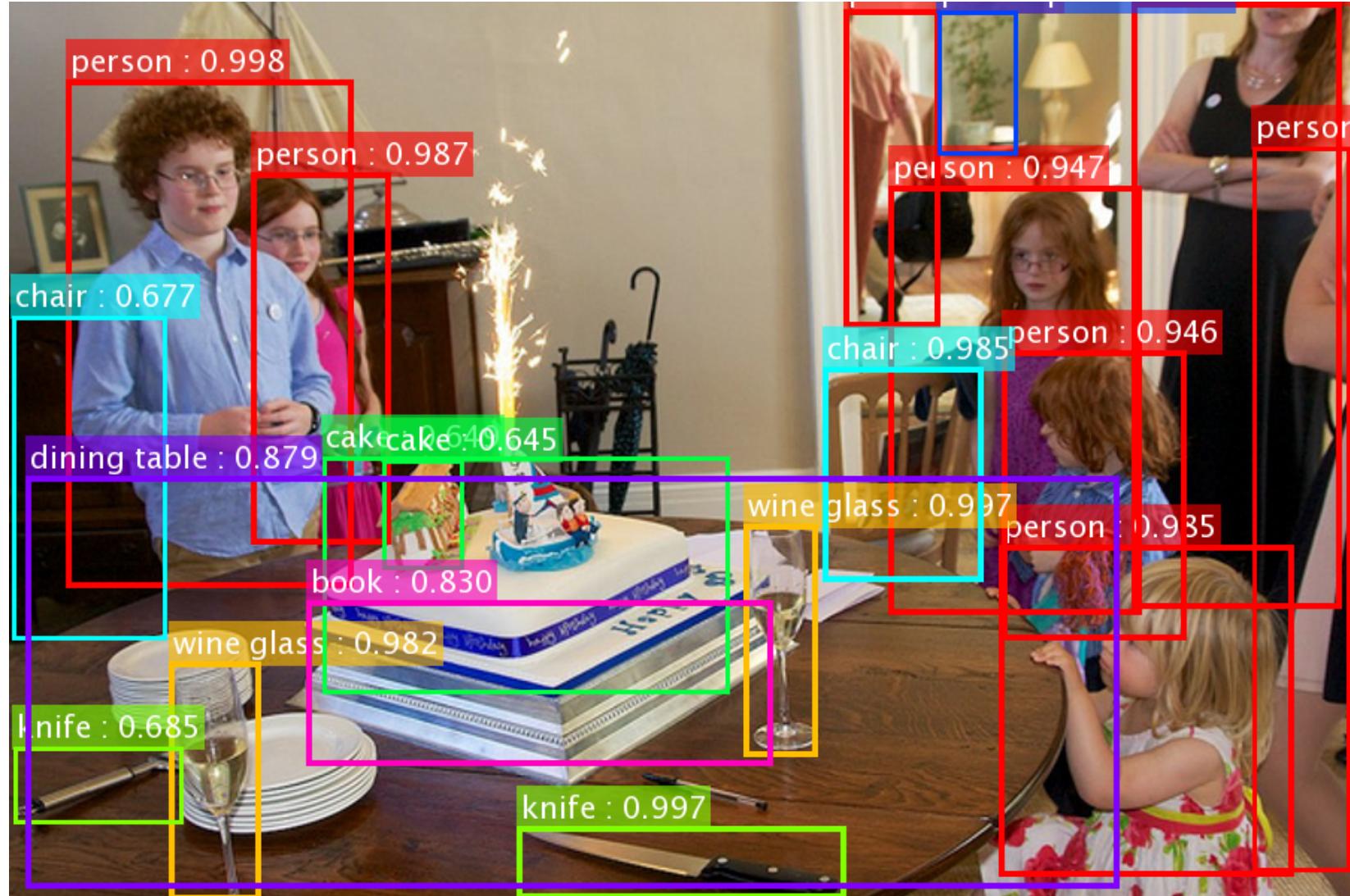
ResNet, **152 layers**  
(ILSVRC 2015)



# Revolution of Depth



\*w/ other improvements & more data



## ResNet's object detection result on COCO

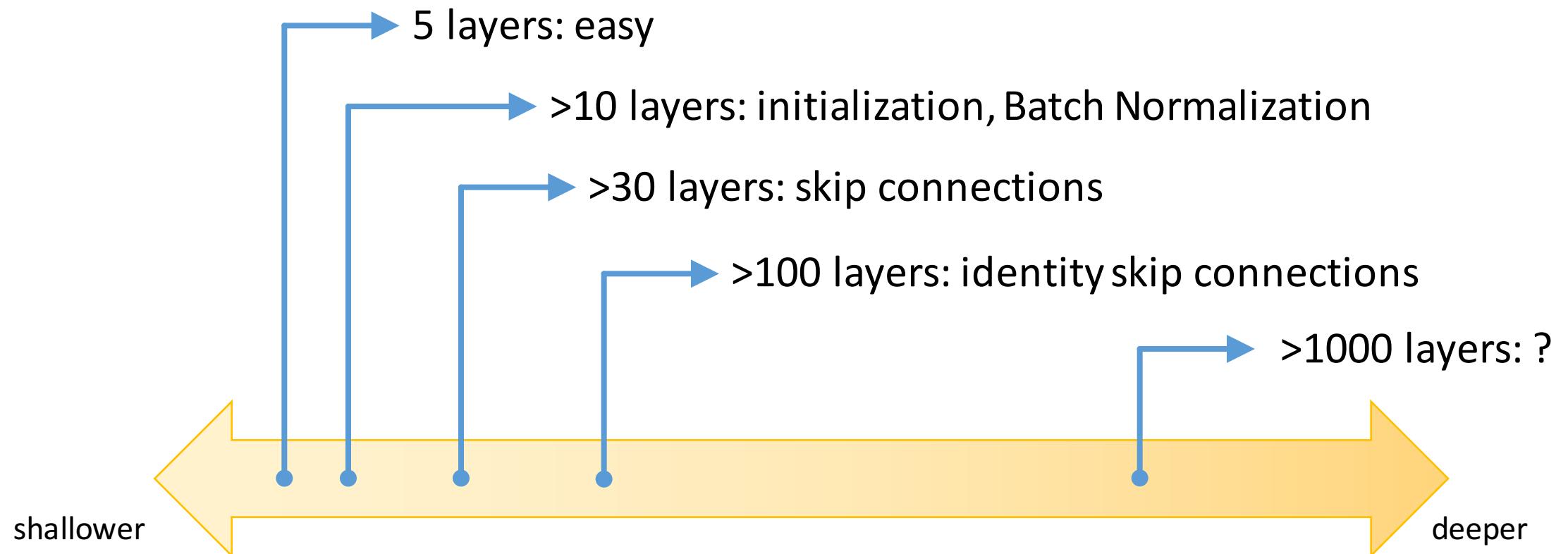
\*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.

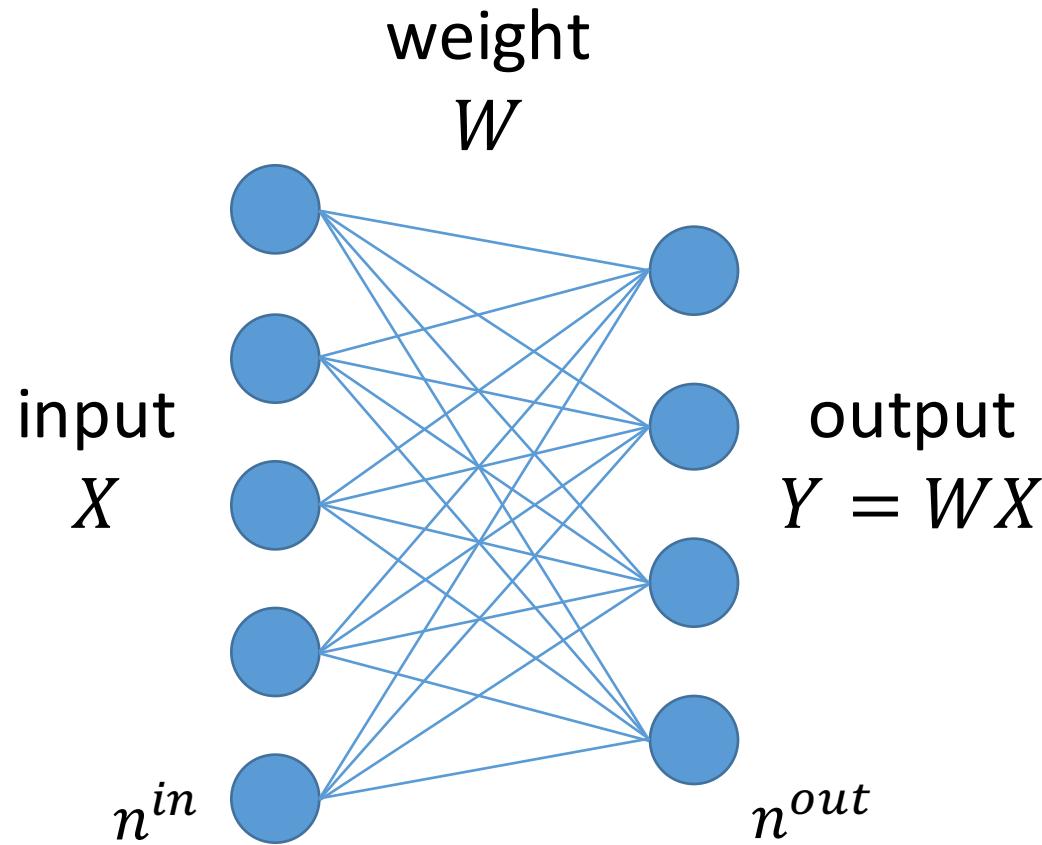
# Background

From shallow to deep

# Spectrum of Depth



# Initialization



If:

- Linear activation
- $x, y, w$ : independent

Then:

1-layer:

$$Var[y] = (n^{in} Var[w]) Var[x]$$

Multi-layer:

$$Var[y] = \left( \prod_d n_d^{in} Var[w_d] \right) Var[x]$$

# Initialization

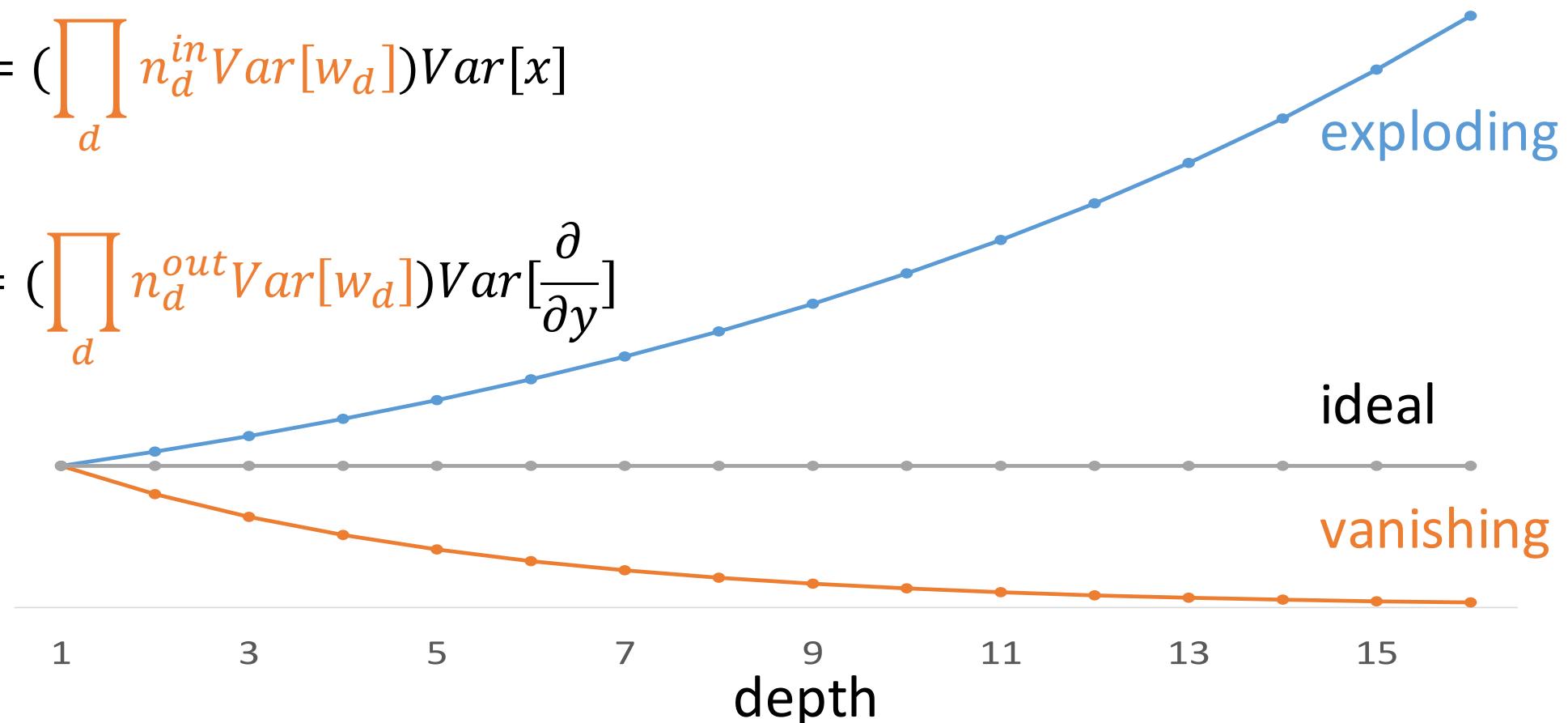
Both forward (response) and backward (gradient) signal can vanish/explode

Forward:

$$Var[y] = \left( \prod_d n_d^{in} Var[w_d] \right) Var[x]$$

Backward:

$$Var\left[\frac{\partial}{\partial x}\right] = \left( \prod_d n_d^{out} Var[w_d] \right) Var\left[\frac{\partial}{\partial y}\right]$$



LeCun et al 1998 "Efficient Backprop"

Glorot & Bengio 2010 "Understanding the difficulty of training deep feedforward neural networks"

# Initialization

- Initialization under **linear** assumption

$$\prod_d n_d^{in} \text{Var}[w_d] = \text{const}_{fw} \text{ (healthy forward)}$$

and

$$\prod_d n_d^{out} \text{Var}[w_d] = \text{const}_{bw} \text{ (healthy backward)}$$



$$n_d^{in} \text{Var}[w_d] = 1$$

or\*

$$n_d^{out} \text{Var}[w_d] = 1$$

\*:  $n_d^{out} = n_{d+1}^{in}$ , so  $\frac{\text{const}_{bw}}{\text{const}_{fw}} = \frac{n_{last}^{out}}{n_{first}^{in}} < \infty$ .

It is sufficient to use either form.

“Xavier” init in Caffe

LeCun et al 1998 “Efficient Backprop”

Glorot & Bengio 2010 “Understanding the difficulty of training deep feedforward neural networks”

# Initialization

- Initialization under **ReLU** activation

$$\prod_d \frac{1}{2} n_d^{in} Var[w_d] = const_{fw} \text{ (healthy forward)}$$

and

$$\prod_d \frac{1}{2} n_d^{out} Var[w_d] = const_{bw} \text{ (healthy backward)}$$

$$\frac{1}{2} n_d^{in} Var[w_d] = 1$$

or

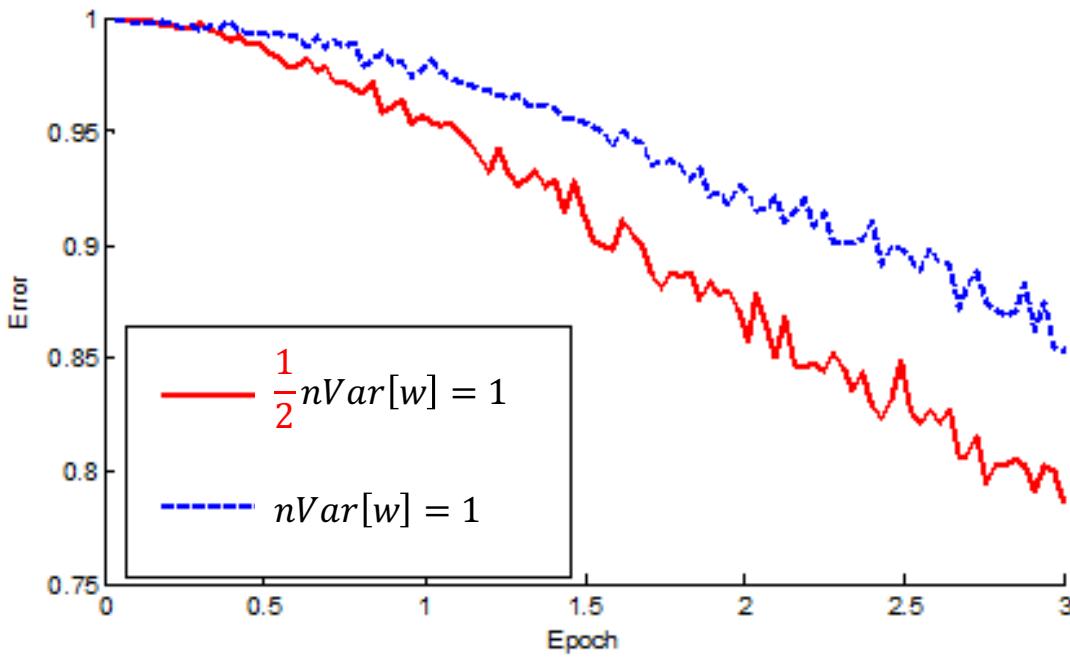
$$\frac{1}{2} n_d^{out} Var[w_d] = 1$$

With  $D$  layers, a factor of 2 per layer has exponential impact of  $2^D$

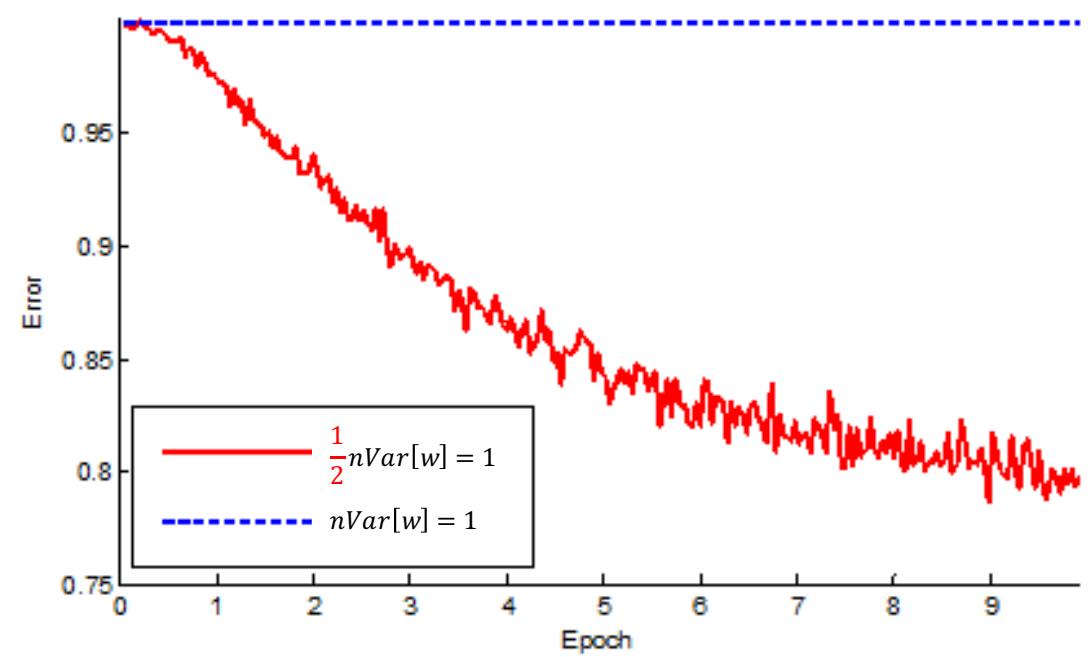
“MSRA” init in Caffe

# Initialization

22-layer ReLU net:  
good init converges faster



30-layer ReLU net:  
good init is able to converge

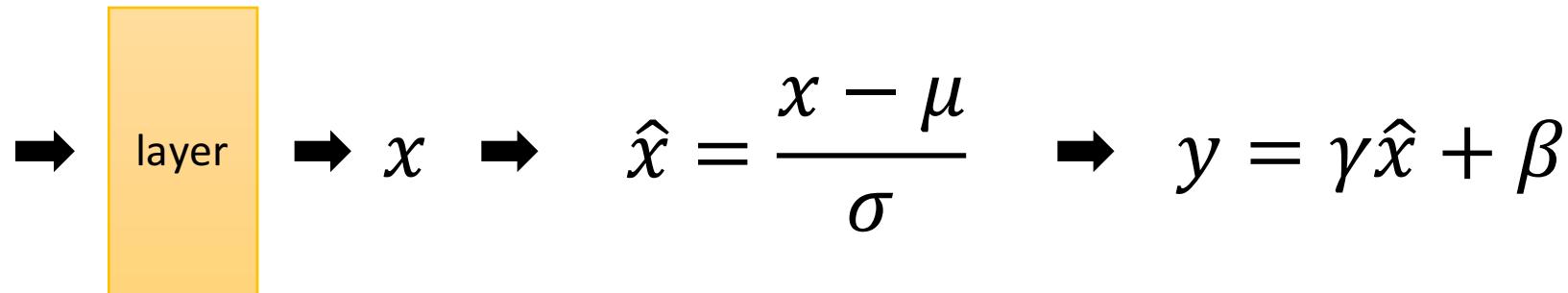


\*Figures show the beginning of training

# Batch Normalization (BN)

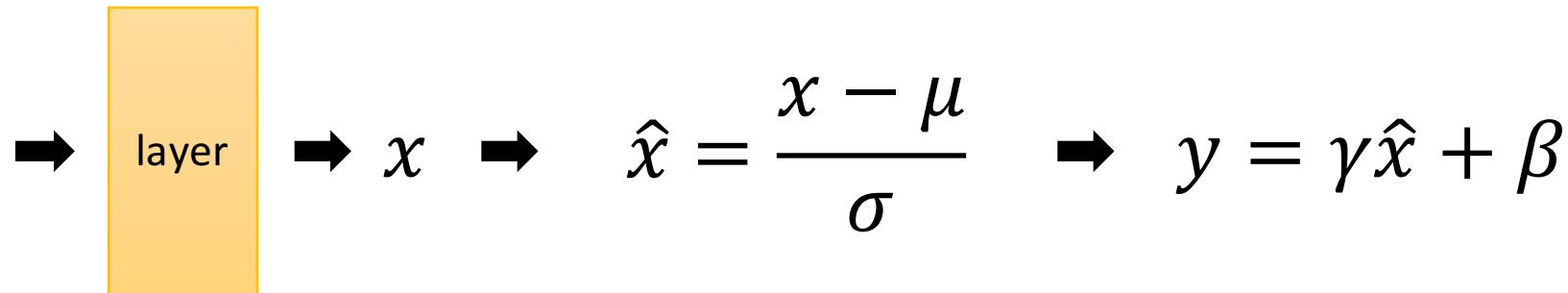
- Normalizing input (LeCun et al 1998 “Efficient Backprop”)
- BN: normalizing **each layer**, for **each mini-batch**
- Greatly accelerate training
- Less sensitive to initialization
- Improve regularization

# Batch Normalization (BN)



- $\mu$ : mean of  $x$  in mini-batch
- $\sigma$ : std of  $x$  in mini-batch
- $\gamma$ : scale
- $\beta$ : shift
- $\mu, \sigma$ : functions of  $x$ ,  
analogous to responses
- $\gamma, \beta$ : parameters to be learned,  
analogous to weights

# Batch Normalization (BN)



2 modes of BN:

- Train mode:
  - $\mu, \sigma$  are functions of  $x$ ; backprop gradients
- Test mode:
  - $\mu, \sigma$  are pre-computed\* on training set

**Caution:** make sure your BN  
is in a correct mode

```
running_mean = momentum * running_mean + (1 - momentum) * sample_mean  
running_var = momentum * running_var + (1 - momentum) * sample_var
```

\*: by running average, or post-processing after training

# Batch Normalization (BN)

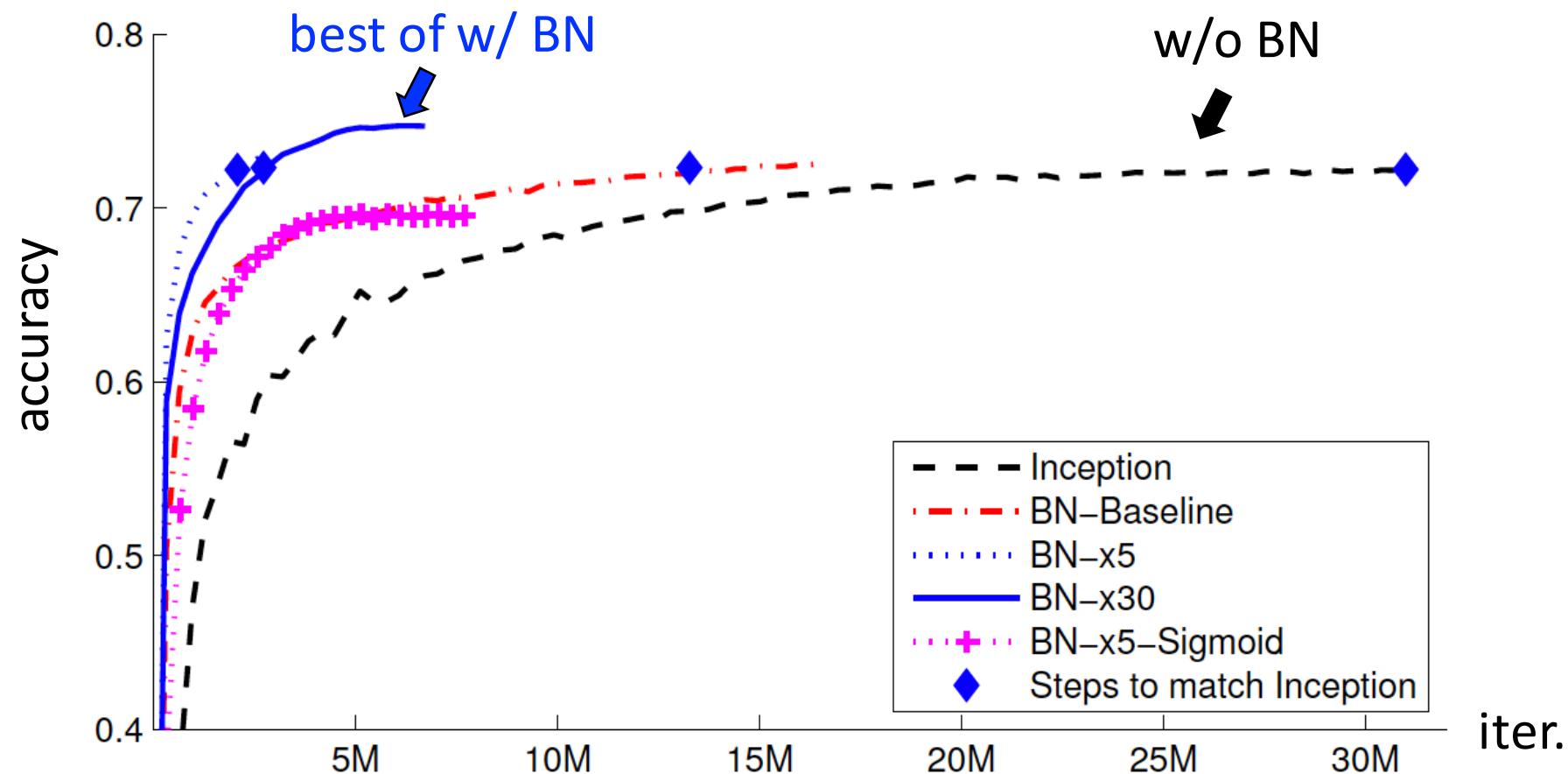


Figure taken from [S. Ioffe & C. Szegedy]

# Deep Residual Networks

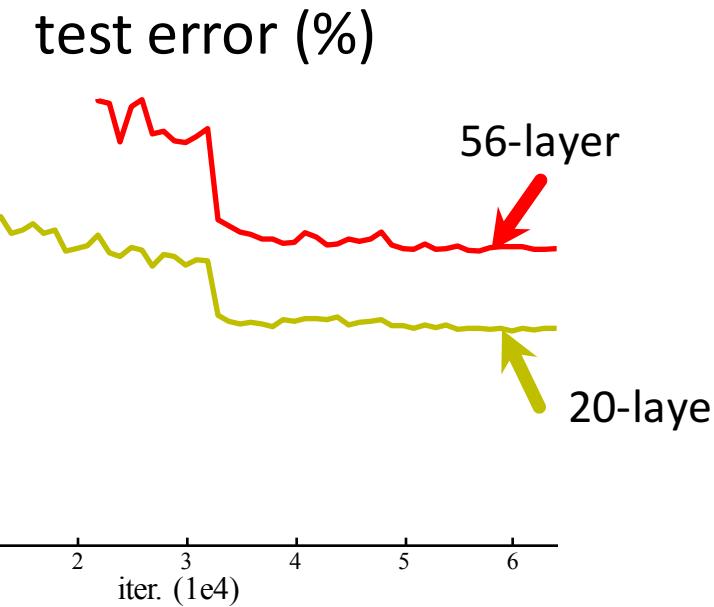
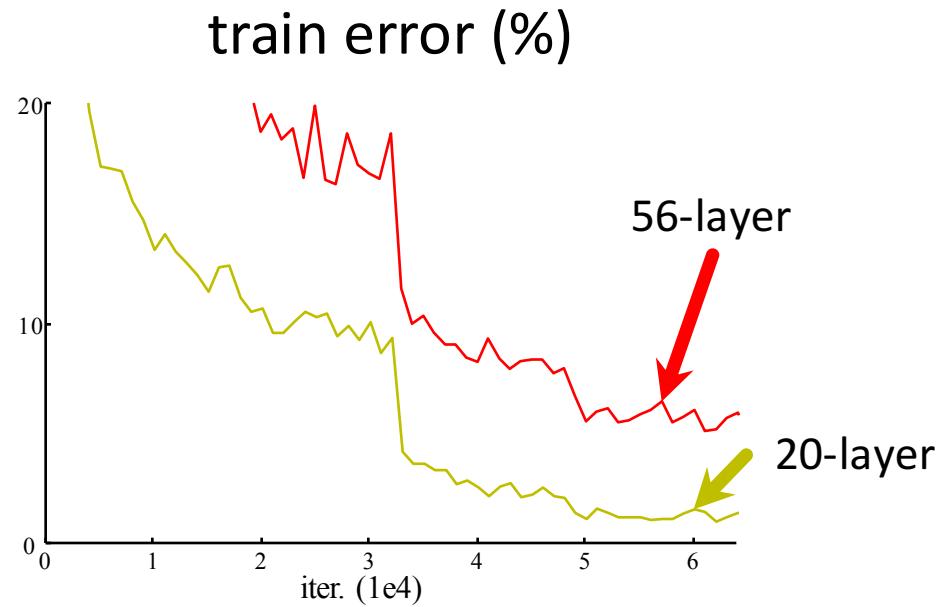
From 10 layers to 100 layers

# Going Deeper

- Initialization algorithms ✓
- Batch Normalization ✓
- **Is learning better networks as simple as stacking more layers?**

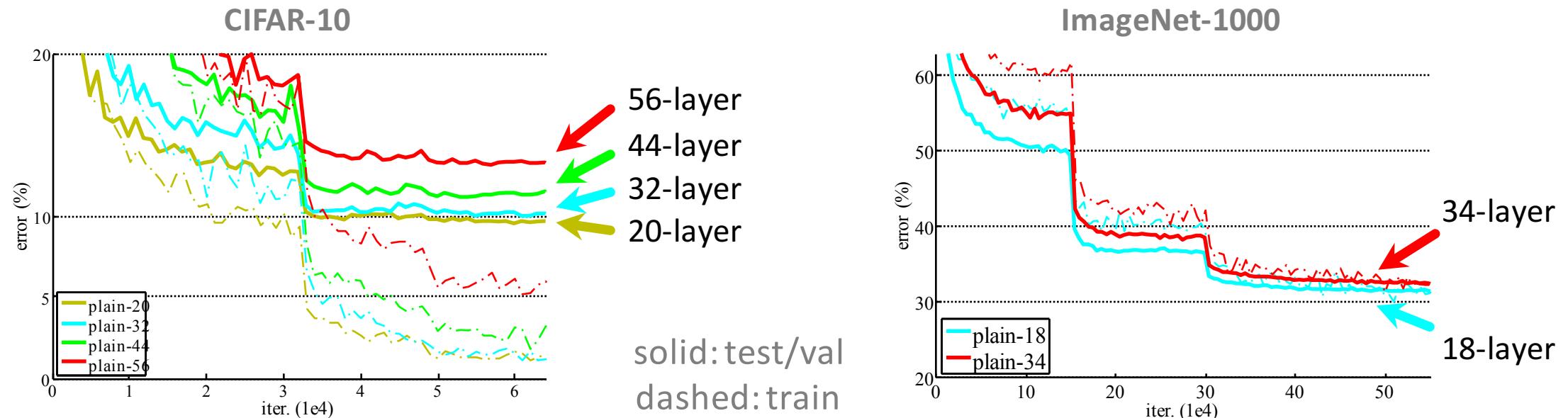
# Simply stacking layers?

CIFAR-10



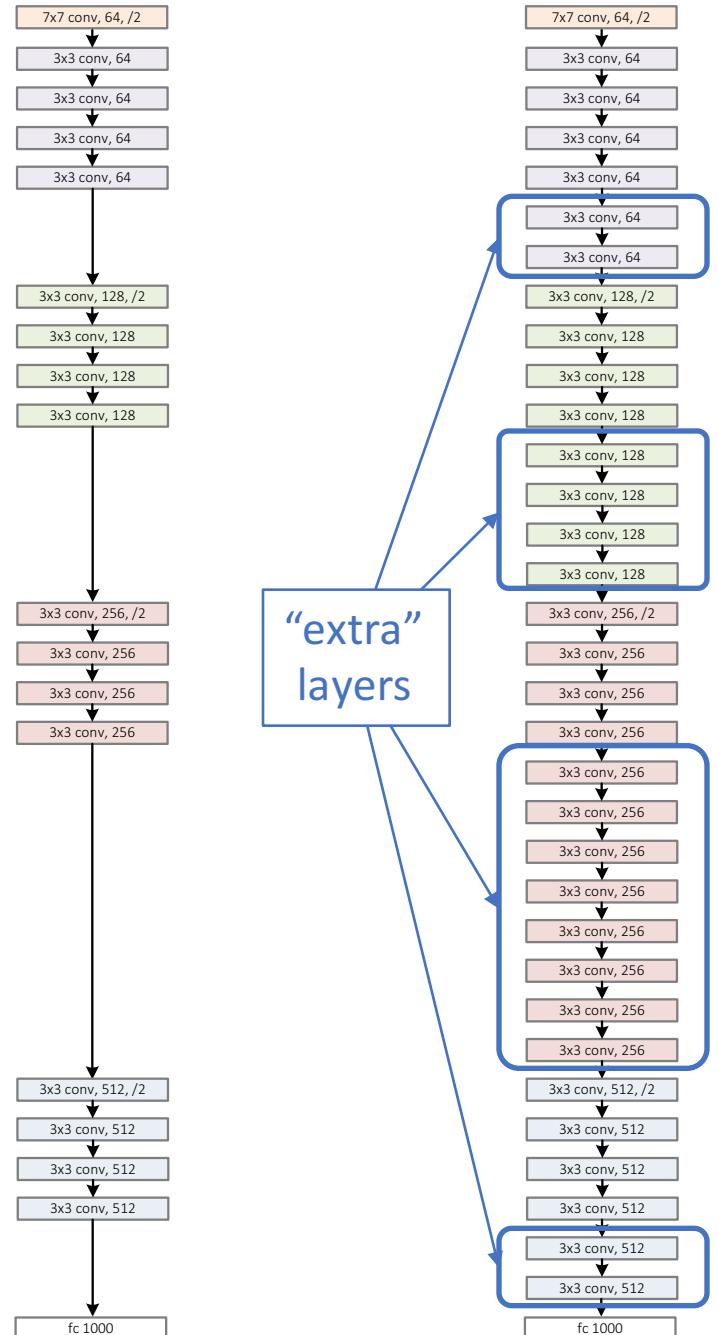
- *Plain* nets: stacking  $3 \times 3$  conv layers...
- 56-layer net has **higher training error** and test error than 20-layer net

# Simply stacking layers?



- “Overly deep” plain nets have **higher training error**
- A general phenomenon, observed in many datasets

a shallower  
model  
(18 layers)

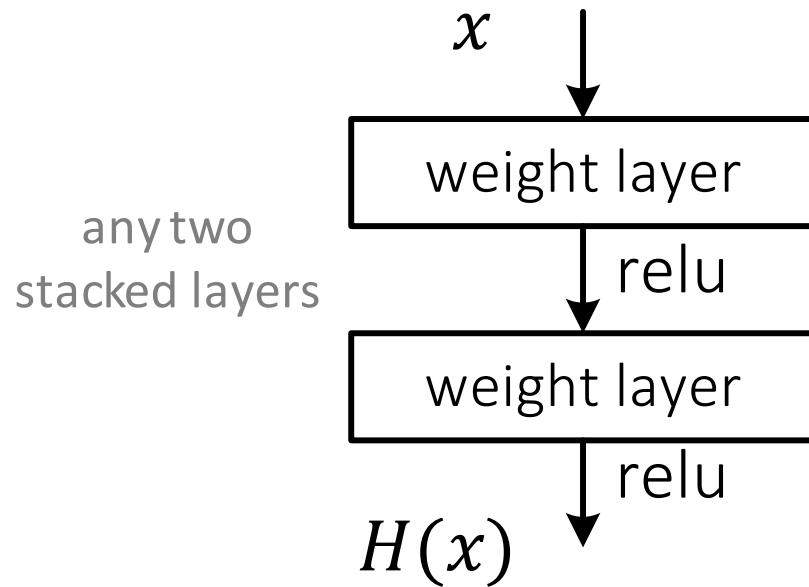


a deeper  
counterpart  
(34 layers)

- Richer solution space
- A deeper model should not have **higher training error**
- A solution *by construction*:
  - original layers: copied from a learned shallower model
  - extra layers: set as **identity**
  - at least the same training error
- **Optimization difficulties**: solvers cannot find the solution when going deeper...

# Deep Residual Learning

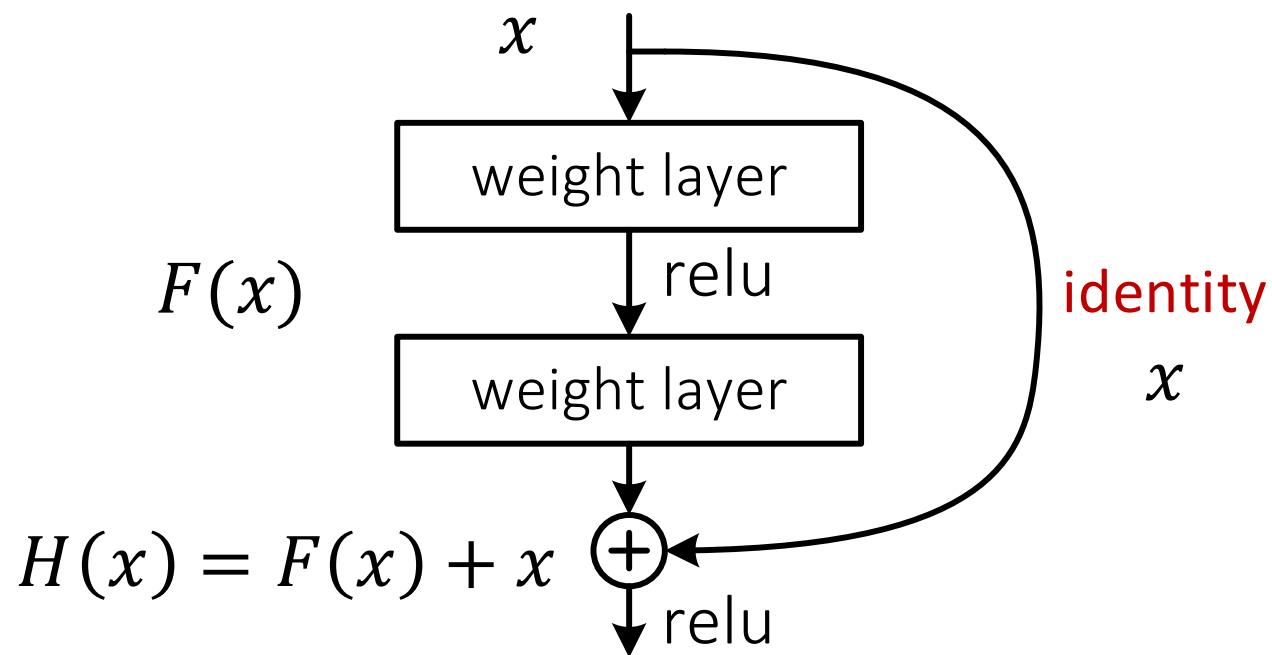
- Plain net



$H(x)$  is any desired mapping,  
hope the 2 weight layers fit  $H(x)$

# Deep Residual Learning

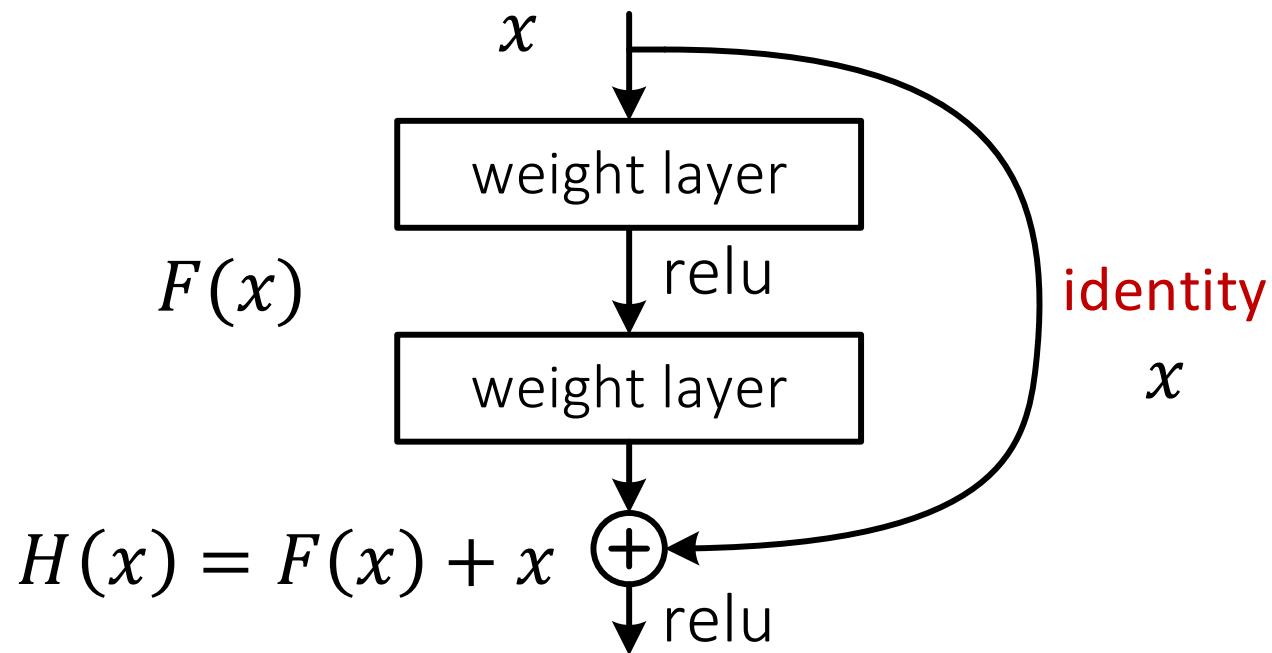
- **Residual net**



$H(x)$  is any desired mapping,  
hope the 2 weight layers fit  $H(x)$   
hope the 2 weight layers fit  $F(x)$   
let  $H(x) = F(x) + x$

# Deep Residual Learning

- $F(x)$  is a **residual mapping w.r.t. identity**



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

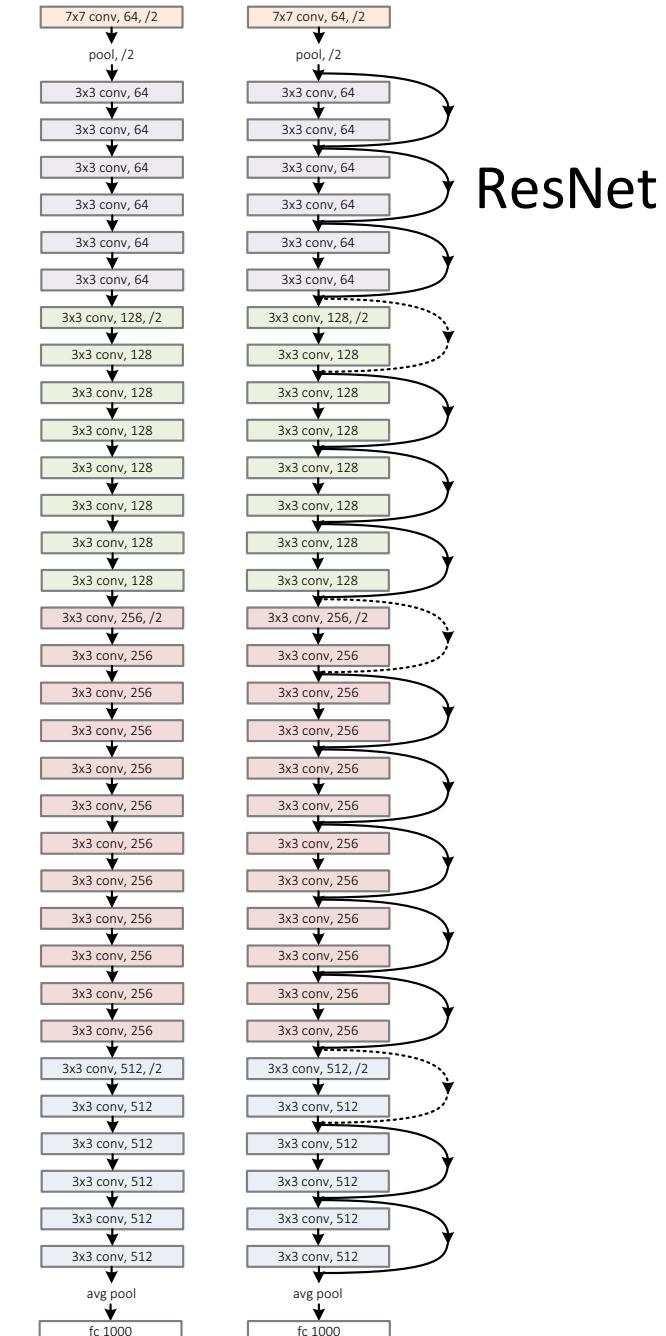
# Related Works – Residual Representations

- VLAD & Fisher Vector [Jegou et al 2010], [Perronnin et al 2007]
  - Encoding **residual** vectors; powerful shallower representations.
- Product Quantization (IVF-ADC) [Jegou et al 2011]
  - Quantizing **residual** vectors; efficient nearest-neighbor search.
- MultiGrid & Hierarchical Precondition [Briggs, et al 2000], [Szeliski 1990, 2006]
  - Solving **residual** sub-problems; efficient PDE solvers.

# Network “Design”

- Keep it simple
- Our basic design (VGG-style)
  - all 3x3 conv (almost)
  - spatial size /2 => # filters x2 (~same complexity per layer)
  - **Simple design; just deep!**
- Other remarks:
  - no hidden fc
  - no dropout

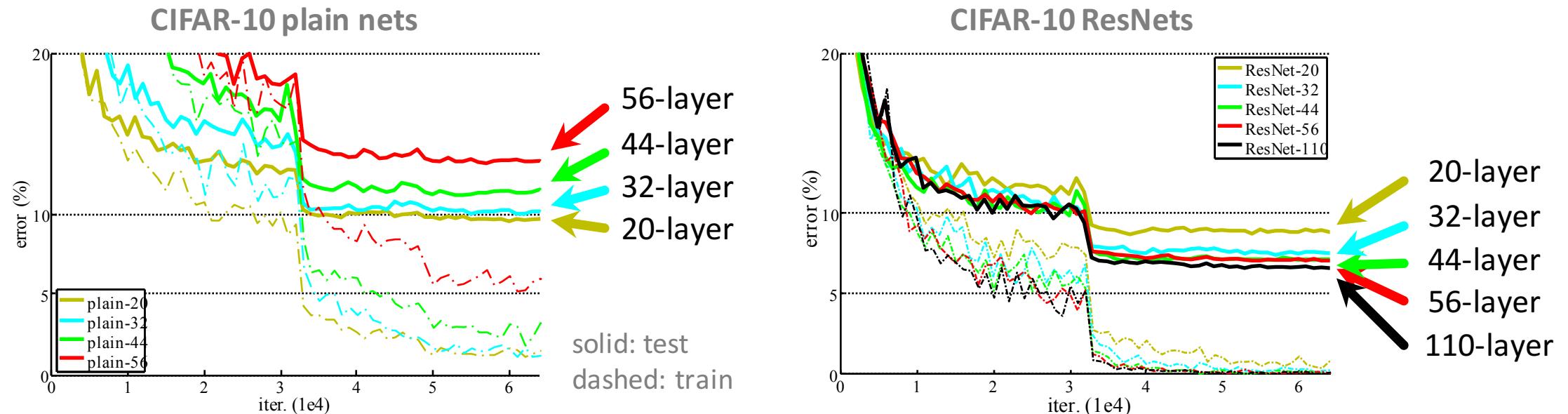
plain net



# Training

- All plain/residual nets are trained **from scratch**
- All plain/residual nets use Batch Normalization
- Standard hyper-parameters & augmentation

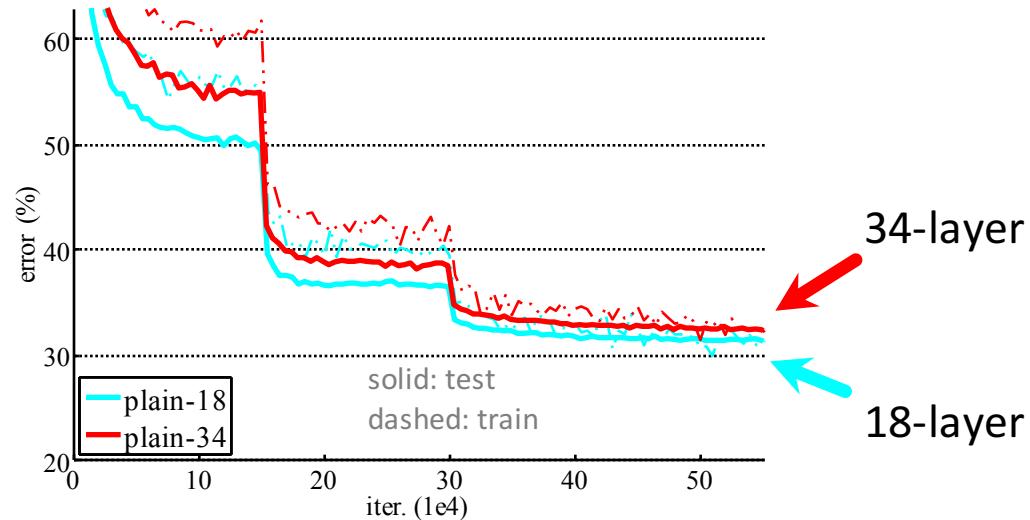
# CIFAR-10 experiments



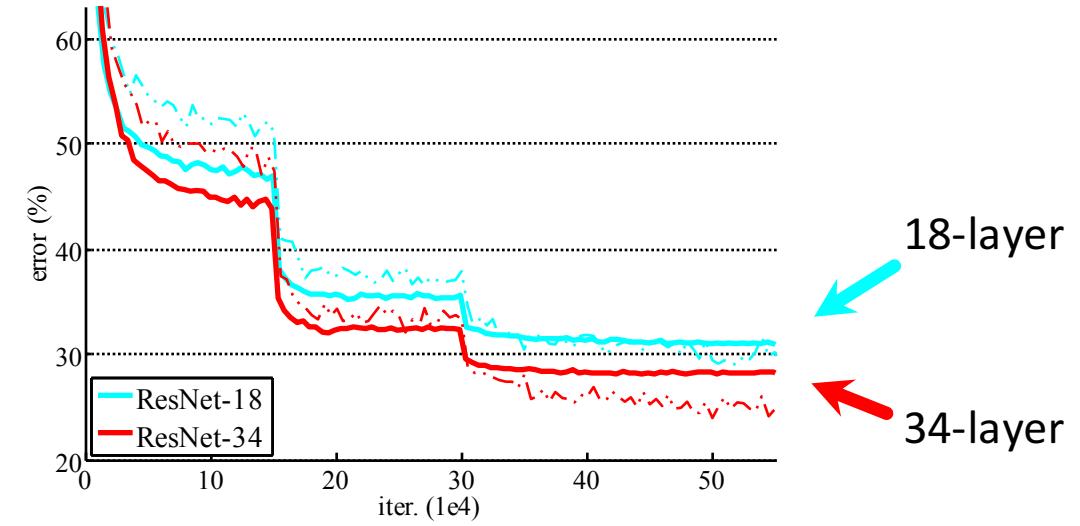
- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

# ImageNet experiments

ImageNet plain nets



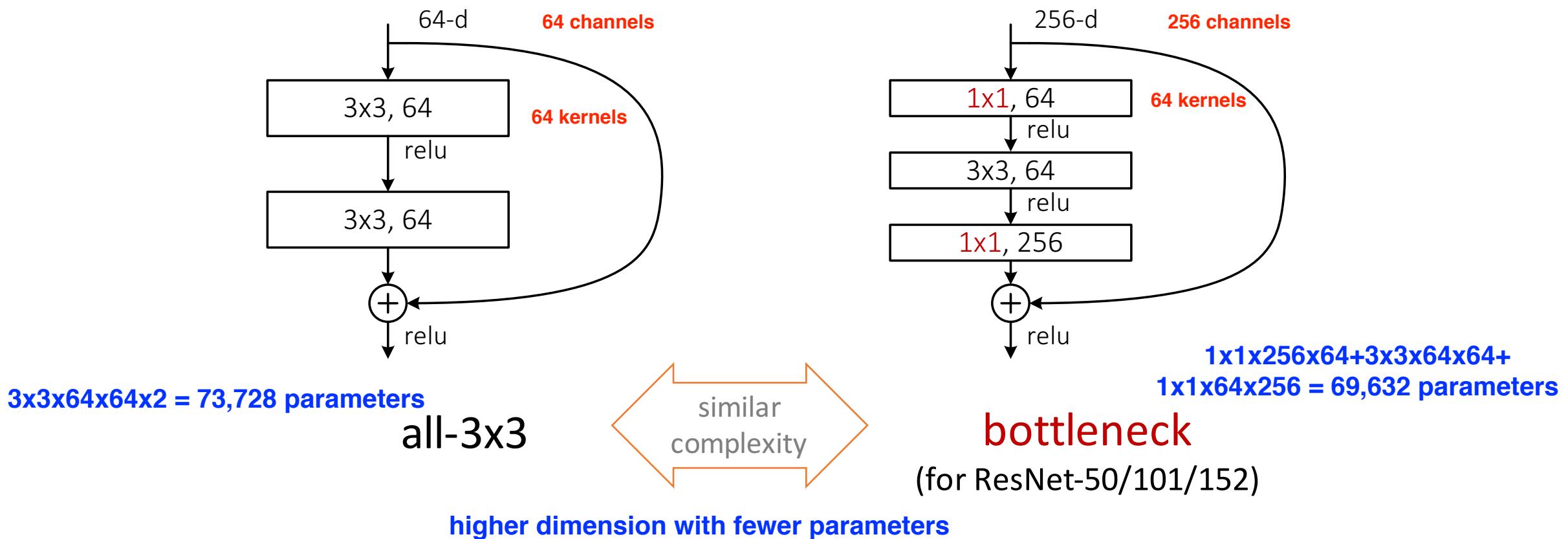
ImageNet ResNets



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

# ImageNet experiments

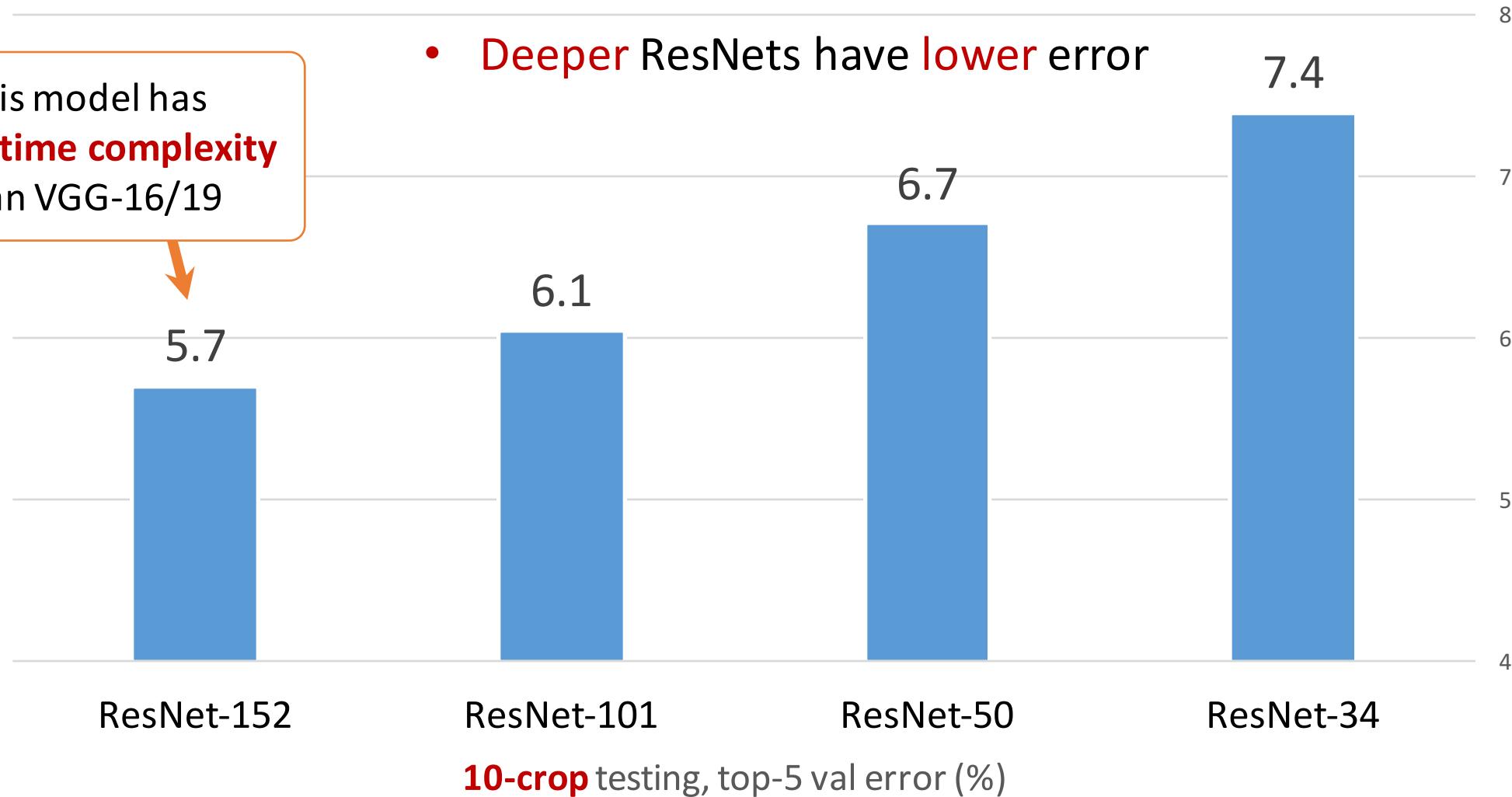
- A practical design of going deeper



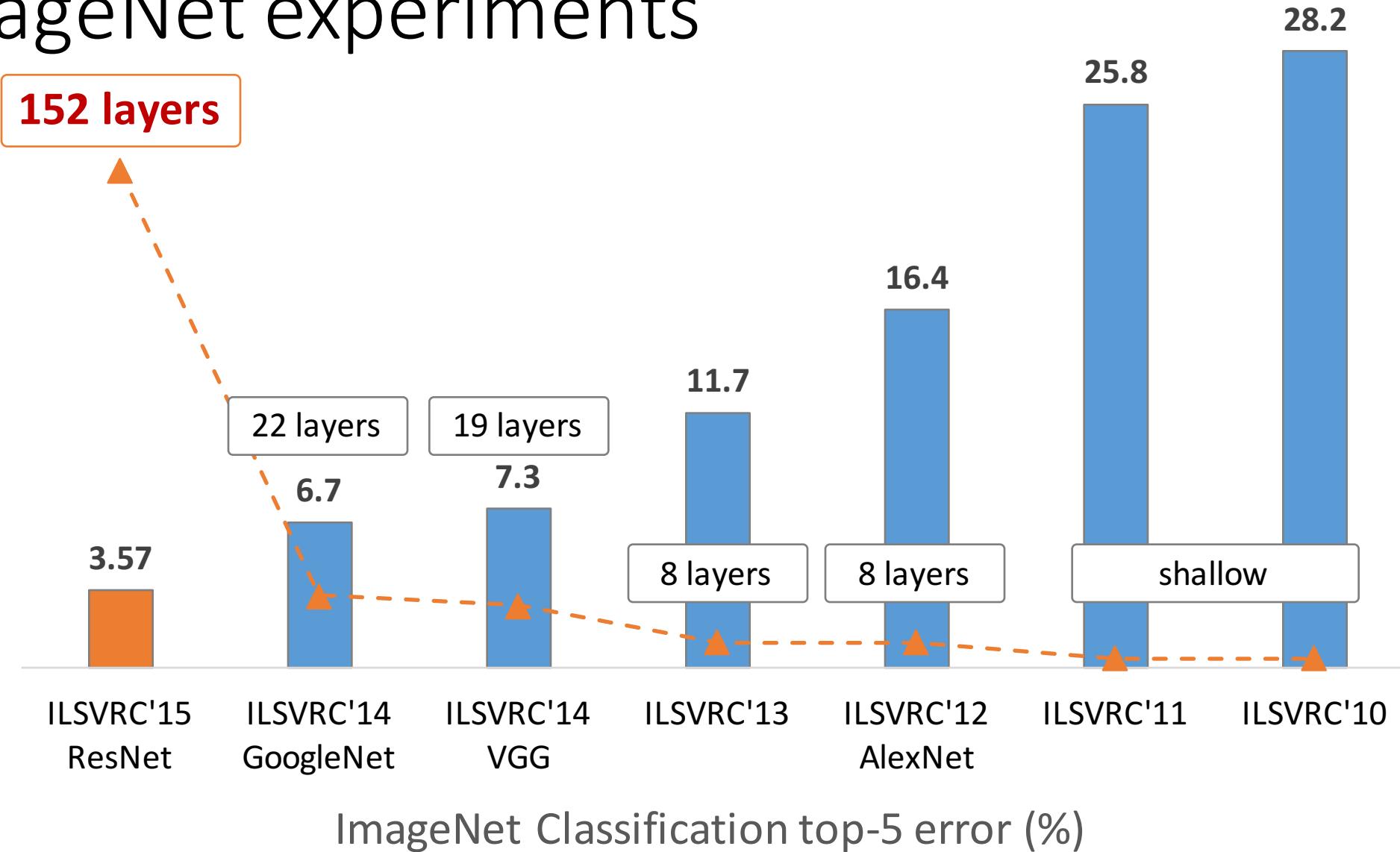
# ImageNet experiments

this model has  
**lower time complexity**  
than VGG-16/19

- Deeper ResNets have lower error



# ImageNet experiments



# Discussions

## Representation, Optimization, Generalization

# Issues on learning deep models

- **Representation** ability

- Ability of model to fit training data, if optimum could be found
- If model A's solution space is a superset of B's, A should be better.

- **Optimization** ability

- Feasibility of finding an optimum
- Not all models are equally easy to optimize

- **Generalization** ability

- Once training data is fit, how good is the test performance

# How do ResNets address these issues?

- **Representation** ability

- No explicit advantage on representation (only re-parameterization), but
- Allow models to go **deeper**

- **Optimization** ability

- Enable very smooth forward/backward prop
- Greatly ease optimizing **deeper** models

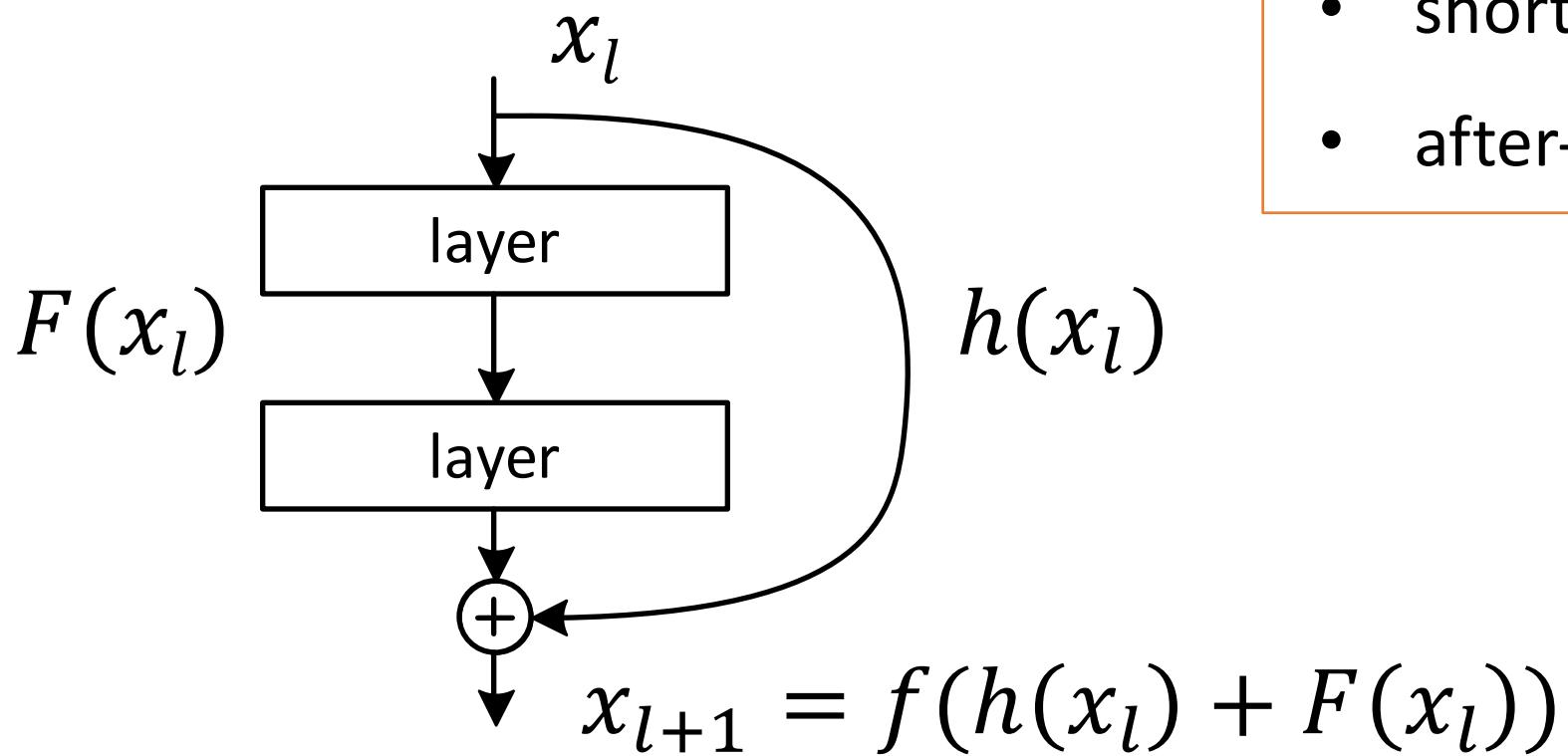
- **Generalization** ability

- Not explicitly address generalization, but
- **Deeper**+thinner is good generalization

# On the Importance of Identity Mapping

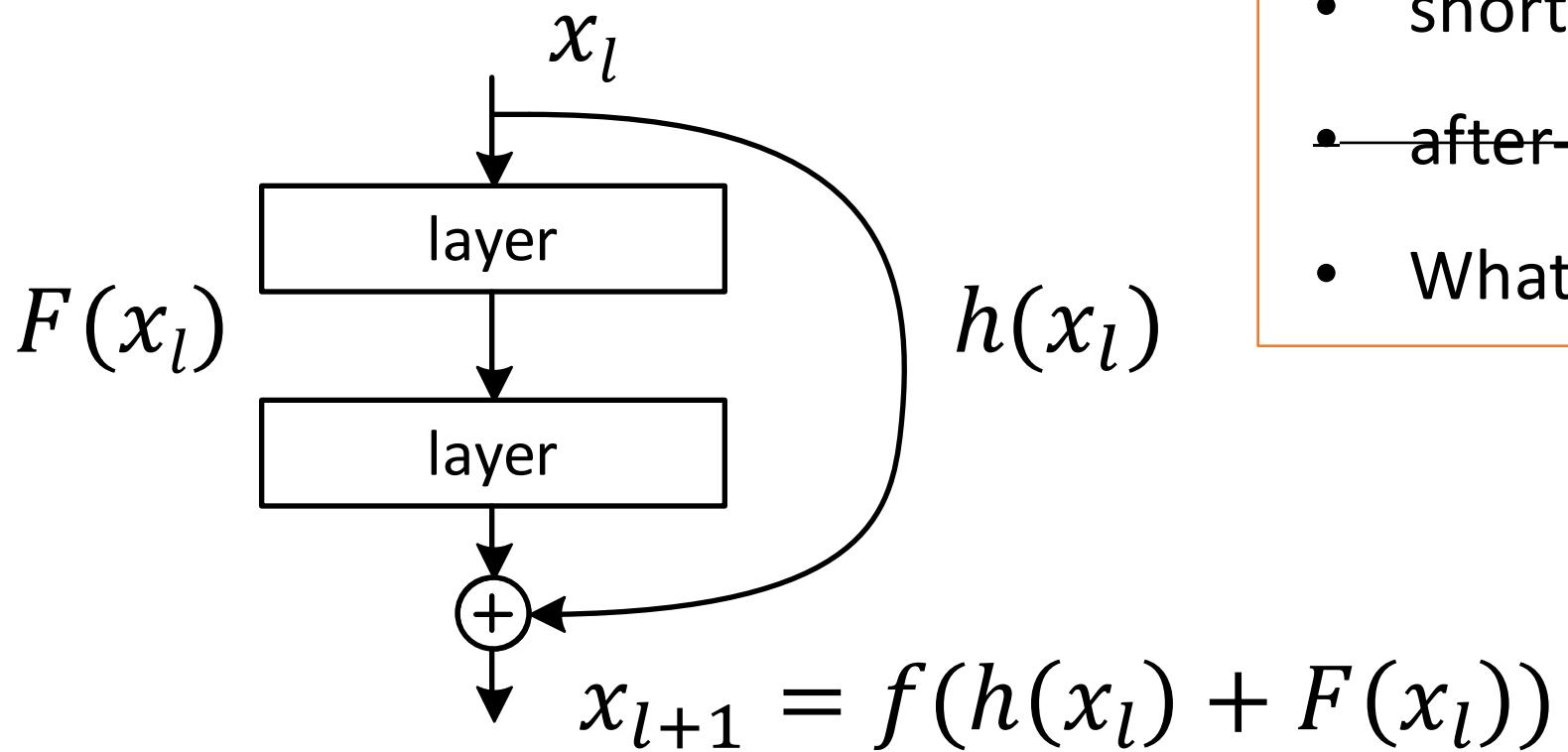
From 100 layers to 1000 layers

# On identity mappings for optimization



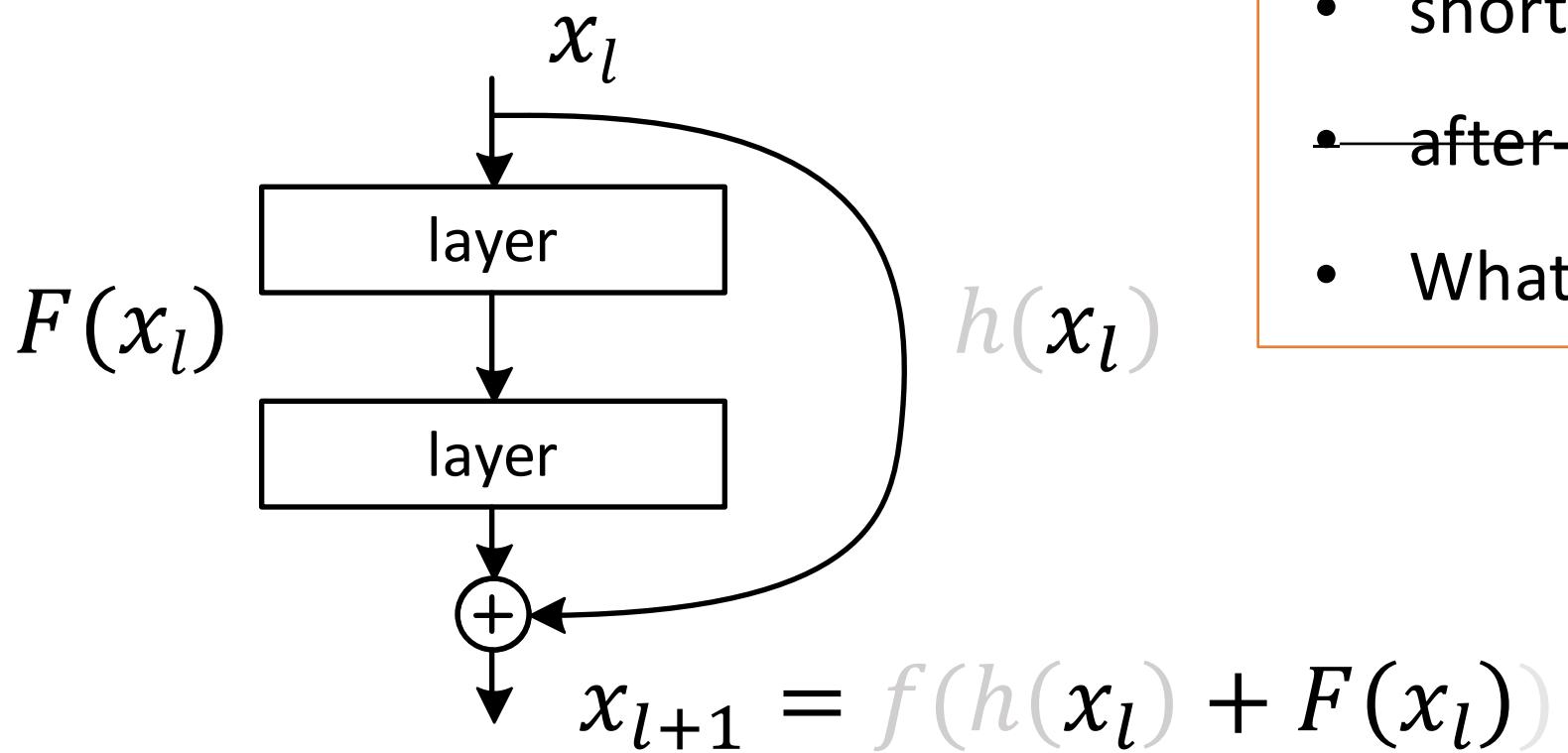
- shortcut mapping:  $h = \text{identity}$
- after-add mapping:  $f = \text{ReLU}$

# On identity mappings for optimization



- shortcut mapping:  $h = \text{identity}$
- after-add mapping:  $f = \text{ReLU}$
- What if  $f = \text{identity}$ ?

# On identity mappings for optimization



- shortcut mapping:  $h = \text{identity}$
- after-add mapping:  $f = \text{ReLU}$
- What if  $f = \text{identity}$ ?

# Very smooth forward propagation

$$x_{l+1} = x_l + F(x_l)$$



$$x_{l+2} = x_{l+1} + F(x_{l+1})$$

# Very smooth forward propagation

$$x_{l+1} = x_l + F(x_l)$$



$$x_{l+2} = x_{l+1} + F(x_{l+1})$$

$$x_{l+2} = x_l + F(x_l) + F(x_{l+1})$$

# Very smooth forward propagation

$$x_{l+1} = x_l + F(x_l)$$



$$x_{l+2} = x_{l+1} + F(x_{l+1})$$

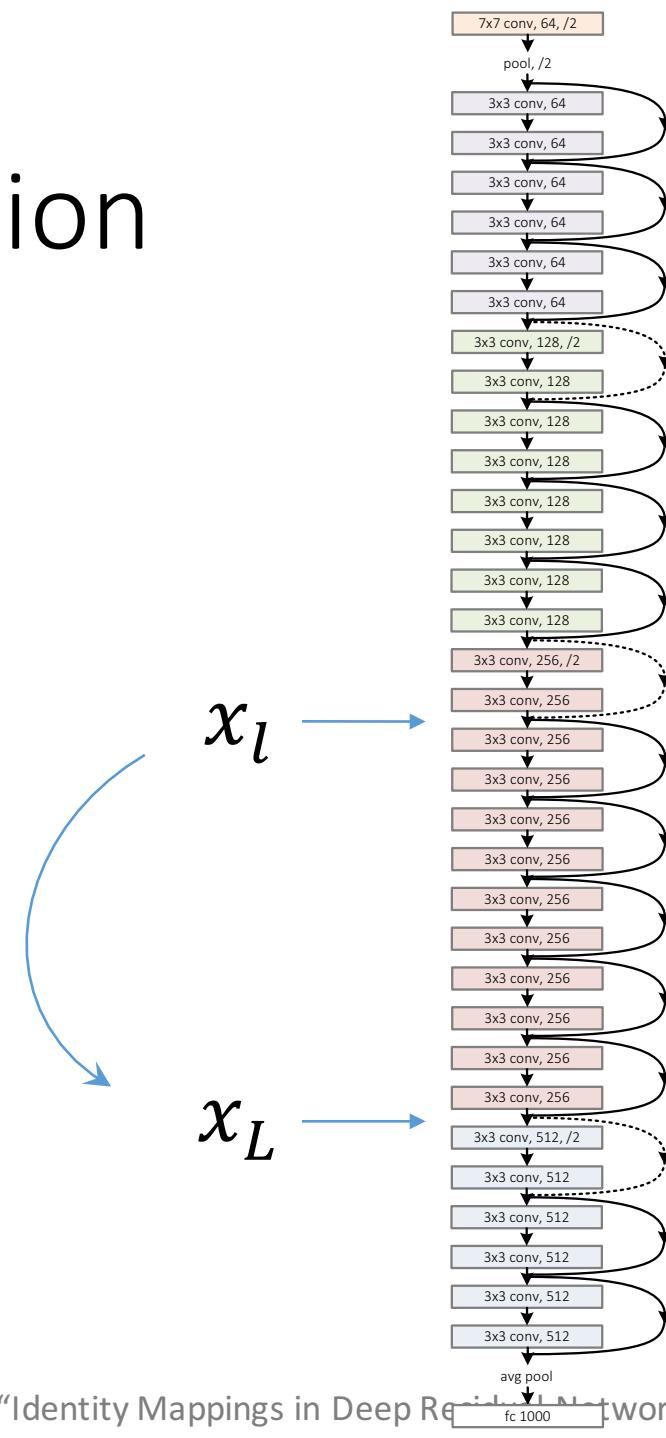
$$x_{l+2} = x_l + F(x_l) + F(x_{l+1})$$

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

# Very smooth forward propagation

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

- Any  $x_l$  is **directly** forward-prop to any  $x_L$ , plus **residual**.
- Any  $x_L$  is an **additive** outcome.
  - in contrast to **multiplicative**:  $x_L = \prod_{i=l}^{L-1} W_i x_l$



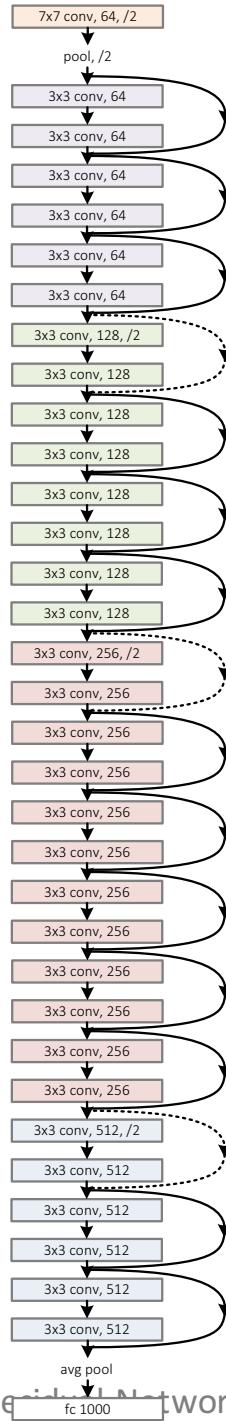
# Very smooth backward propagation

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



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

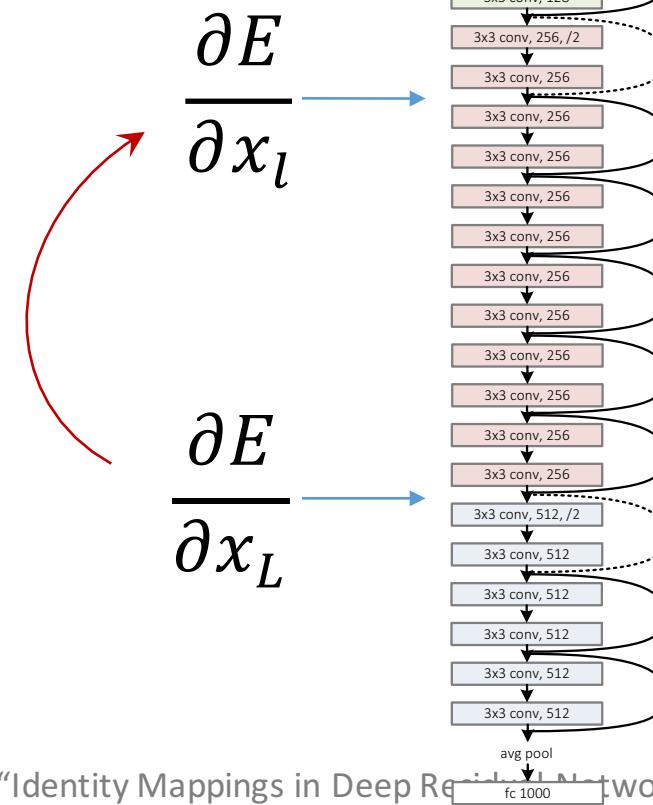
$$\frac{\partial E}{\partial x_l} \quad \frac{\partial E}{\partial x_L}$$



# Very smooth backward propagation

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

- Any  $\frac{\partial E}{\partial x_L}$  is **directly** back-prop to any  $\frac{\partial E}{\partial x_l}$ , plus **residual**.
- Any  $\frac{\partial E}{\partial x_l}$  is **additive**; unlikely to vanish
  - in contrast to **multiplicative**:  $\frac{\partial E}{\partial x_l} = \prod_{i=l}^{L-1} W_i \frac{\partial E}{\partial x_L}$



# Residual for every layer

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

Enabled by:

- shortcut mapping:  $h = \text{identity}$
- after-add mapping:  $f = \text{identity}$

backward:  $\frac{\partial E}{\partial x_l} = \frac{\partial E}{\partial x_L} \left( 1 + \frac{\partial}{\partial x_l} \sum_{i=1}^{L-1} F(x_i) \right)$

# Experiments

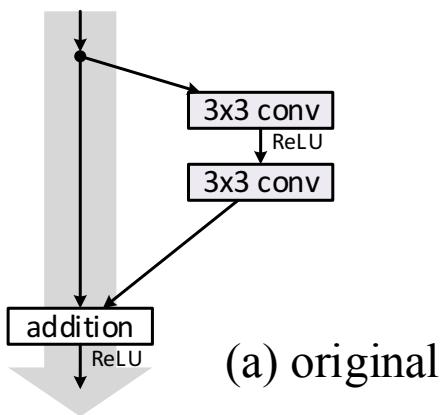
- Set 1: what if shortcut mapping  $h \neq$  identity
- Set 2: what if after-add mapping  $f$  is identity
- Experiments on ResNets with more than 100 layers
  - deeper models suffer more from optimization difficulty

Experiment Set 1:  
what if shortcut mapping  $h \neq$  identity?

\* ResNet-110 on CIFAR-10

$$h(x) = x$$

error: 6.6%

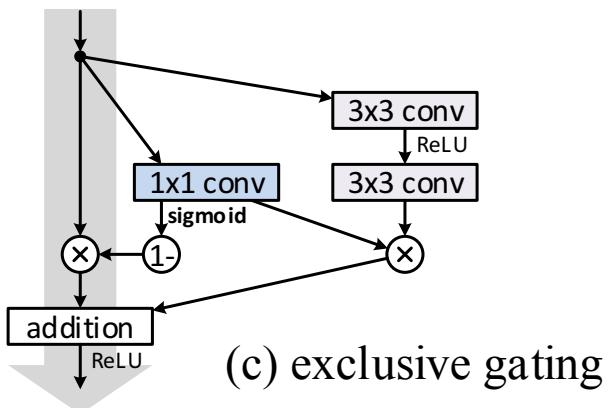


(a) original

$$h(x) = \text{gate} \cdot x$$

error: 8.7%

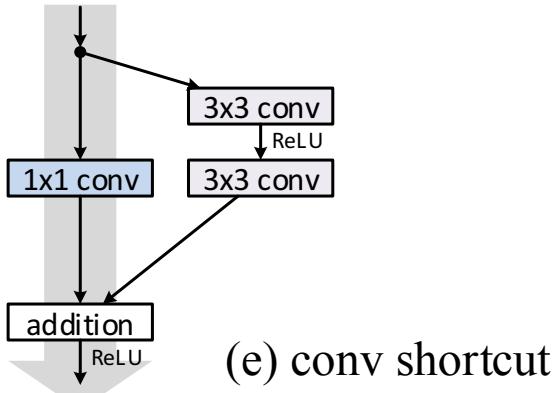
\*similar to "Highway Network"



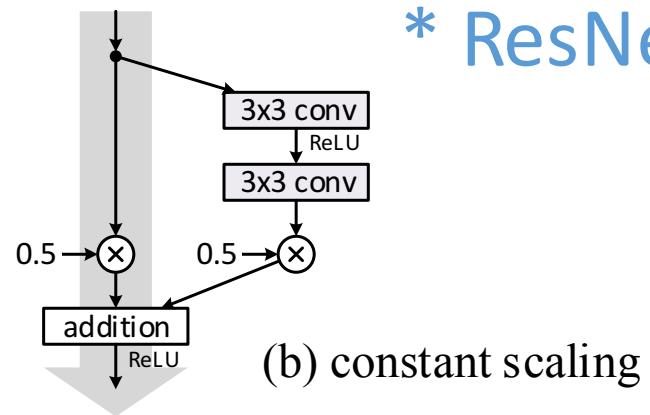
(c) exclusive gating

$$h(x) = \text{conv}(x)$$

error: 12.2%



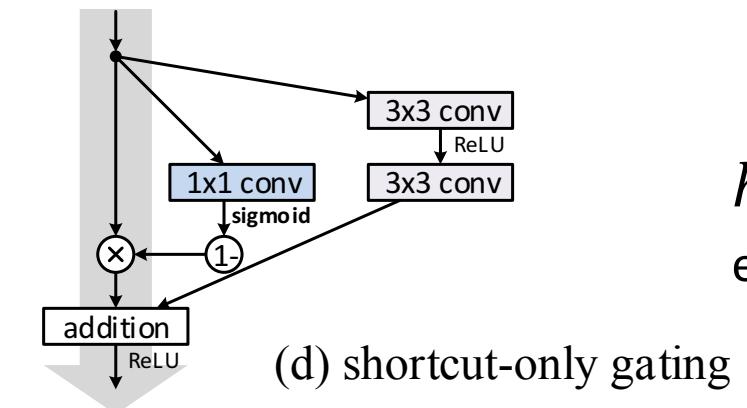
(e) conv shortcut



(b) constant scaling

$$h(x) = 0.5x$$

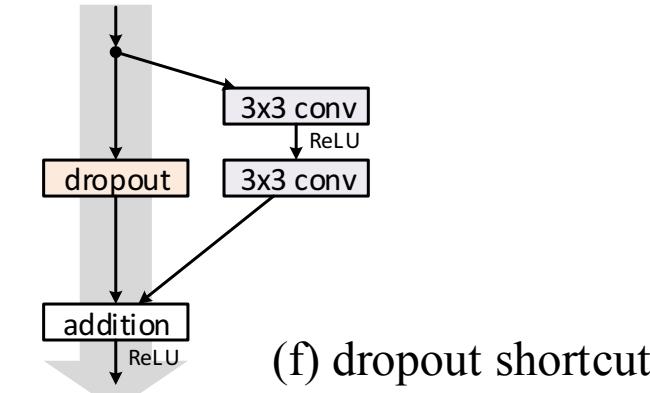
error: 12.4%



(d) shortcut-only gating

$$h(x) = \text{gate} \cdot x$$

error: 12.9%



(f) dropout shortcut

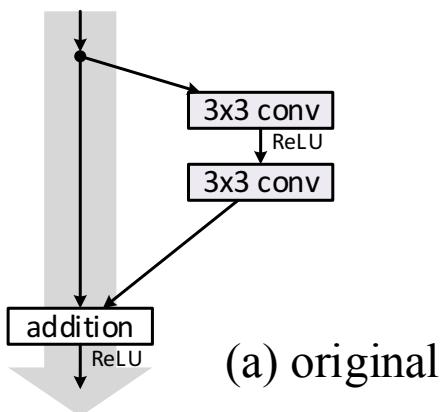
$$h(x) = \text{dropout}(x)$$

error: > 20%

\* ResNet-110 on CIFAR-10

$$h(x) = x$$

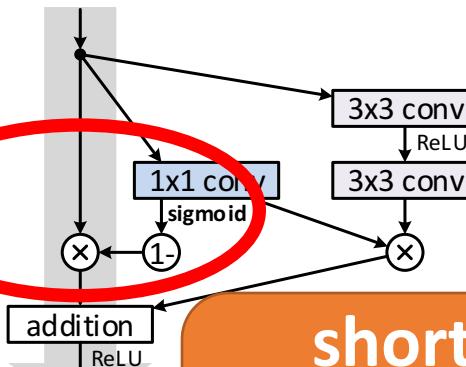
error: 6.6%



(a) original

$$h(x) = \text{gate} \cdot x$$

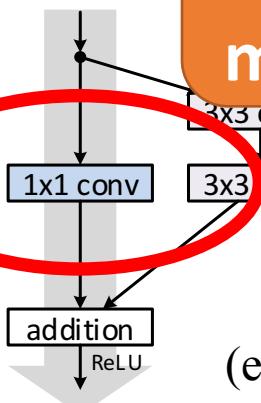
error: 8.7%



(b) constant scaling

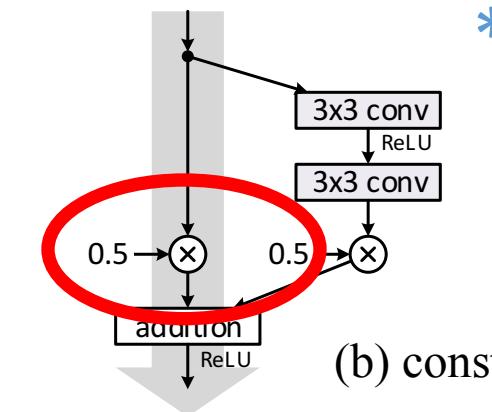
$$h(x) = \text{conv}(x)$$

error: 12.2%



(c) convolutional shortcut

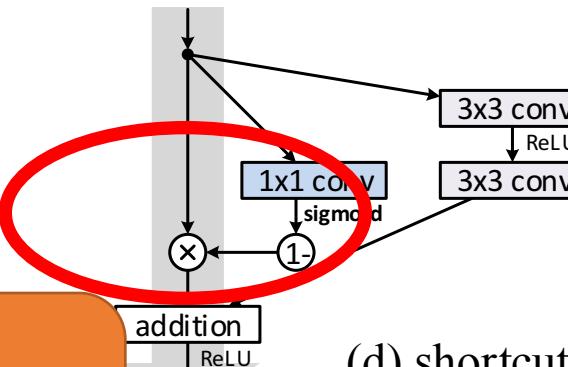
shortcuts  
blocked by  
multiplications



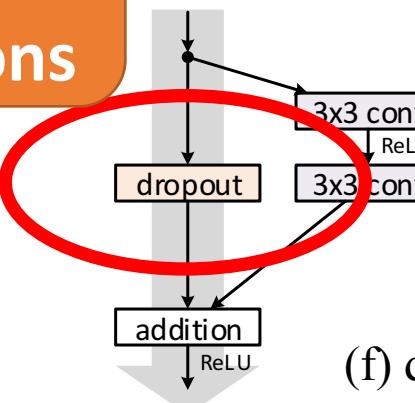
(d) shortcut-only gating

$$h(x) = 0.5x$$

error: 12.4%



(e) convolutional shortcut



$$h(x) = \text{dropout}(x)$$

error: 12.9%

$$h(x) = \text{dropout}(x)$$

error: > 20%

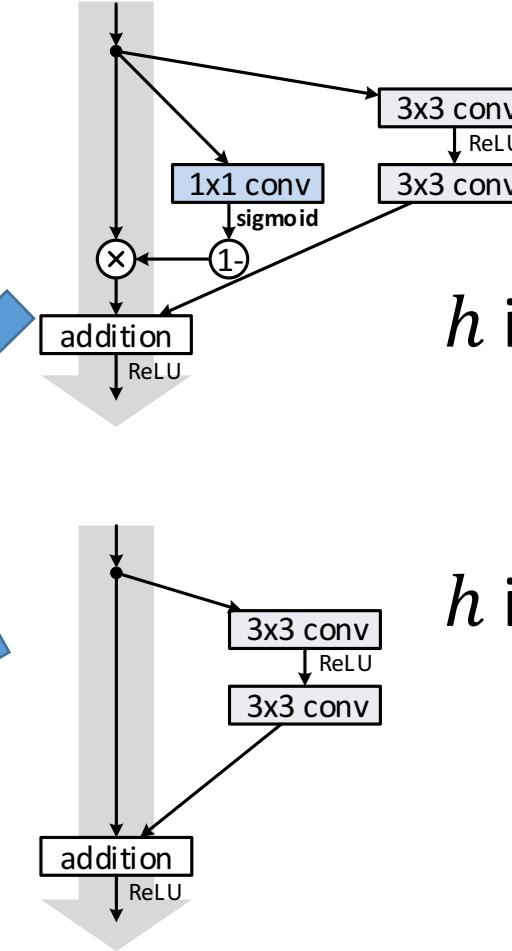
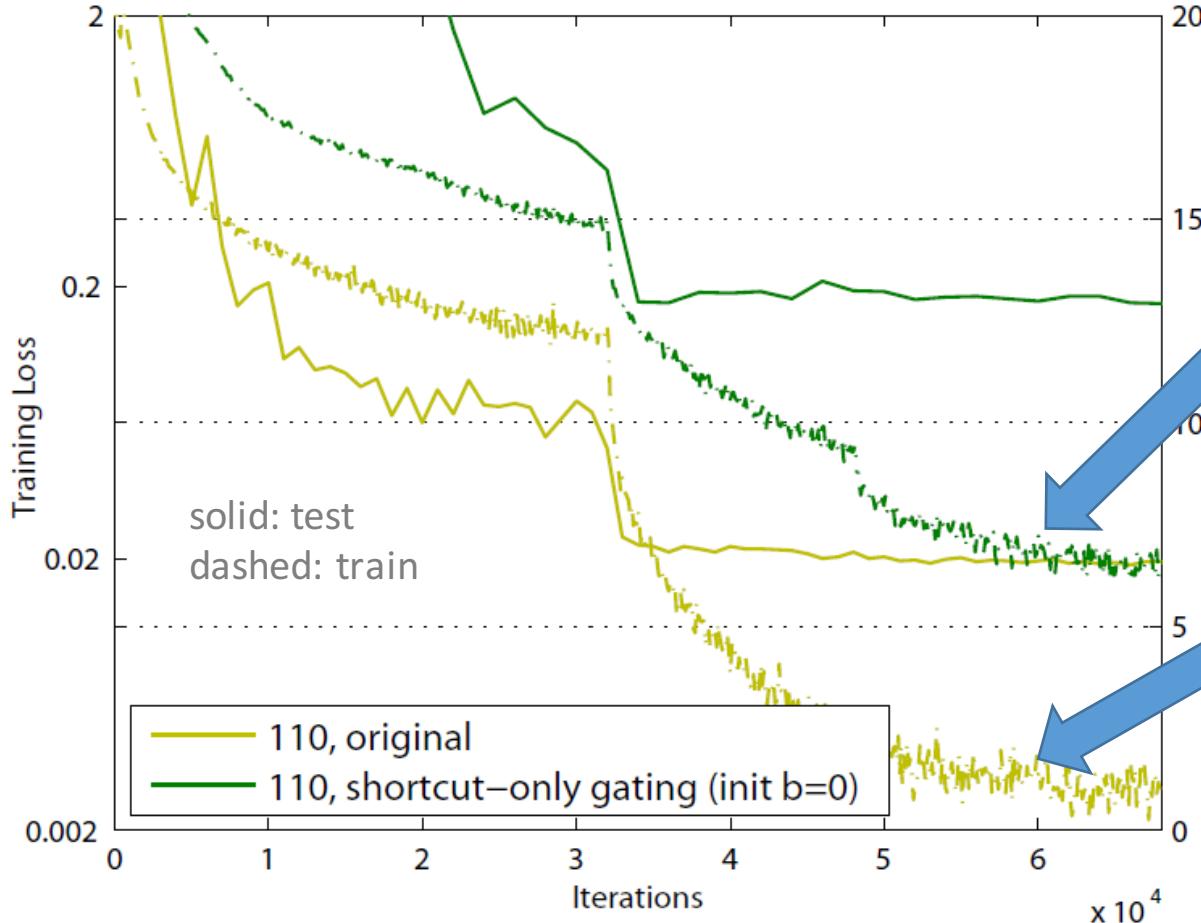
If  $h$  is multiplicative, e.g.  $h(x) = \lambda x$

forward:  $x_L = \lambda^{L-l} x_l + \sum_{i=l}^{L-1} \hat{F}(x_i)$

- if  $h$  is multiplicative, shortcuts are blocked
- direct propagation is decayed

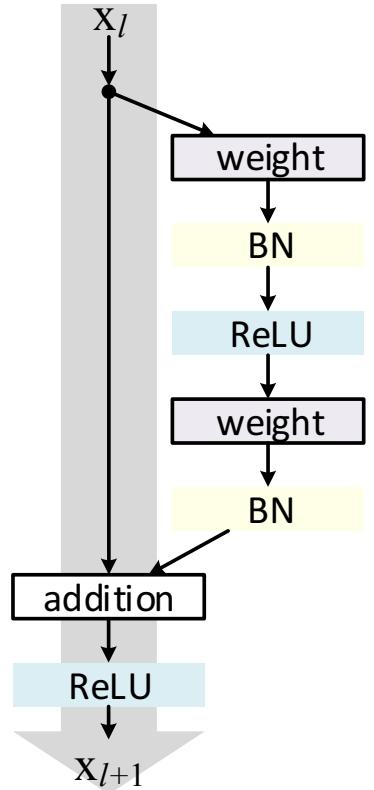
backward:  $\frac{\partial E}{\partial x_l} = \frac{\partial E}{\partial x_L} (\lambda^{L-l} + \frac{\partial}{\partial x_l} \sum_{i=1}^{L-1} \hat{F}(x_i))$

\*assuming  $f$  = identity

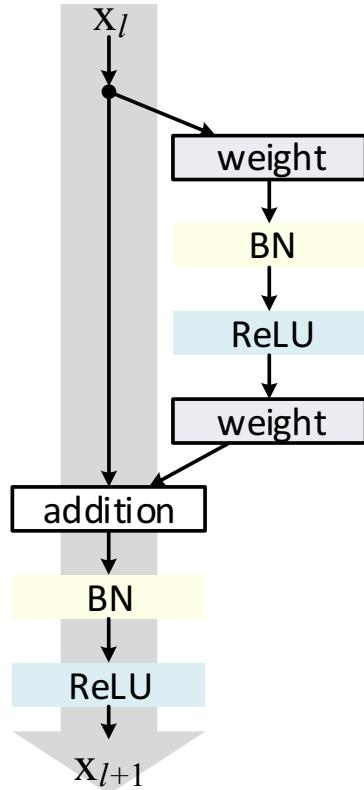


- gating should have better representation ability (identity is a special case), but
- optimization difficulty dominates results

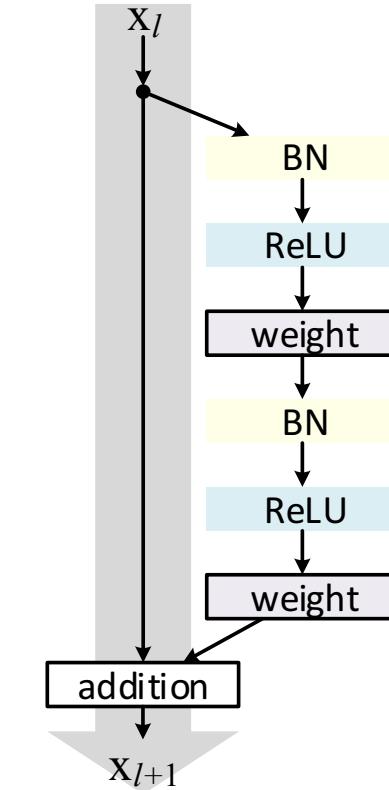
Experiment Set 2:  
what if after-add mapping  $f$  is identity



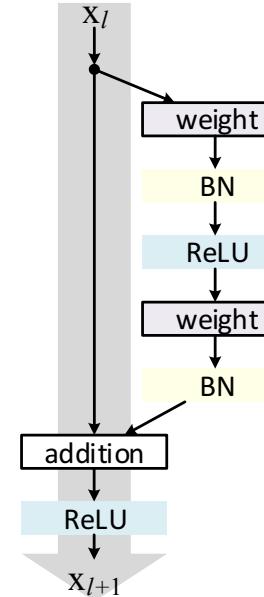
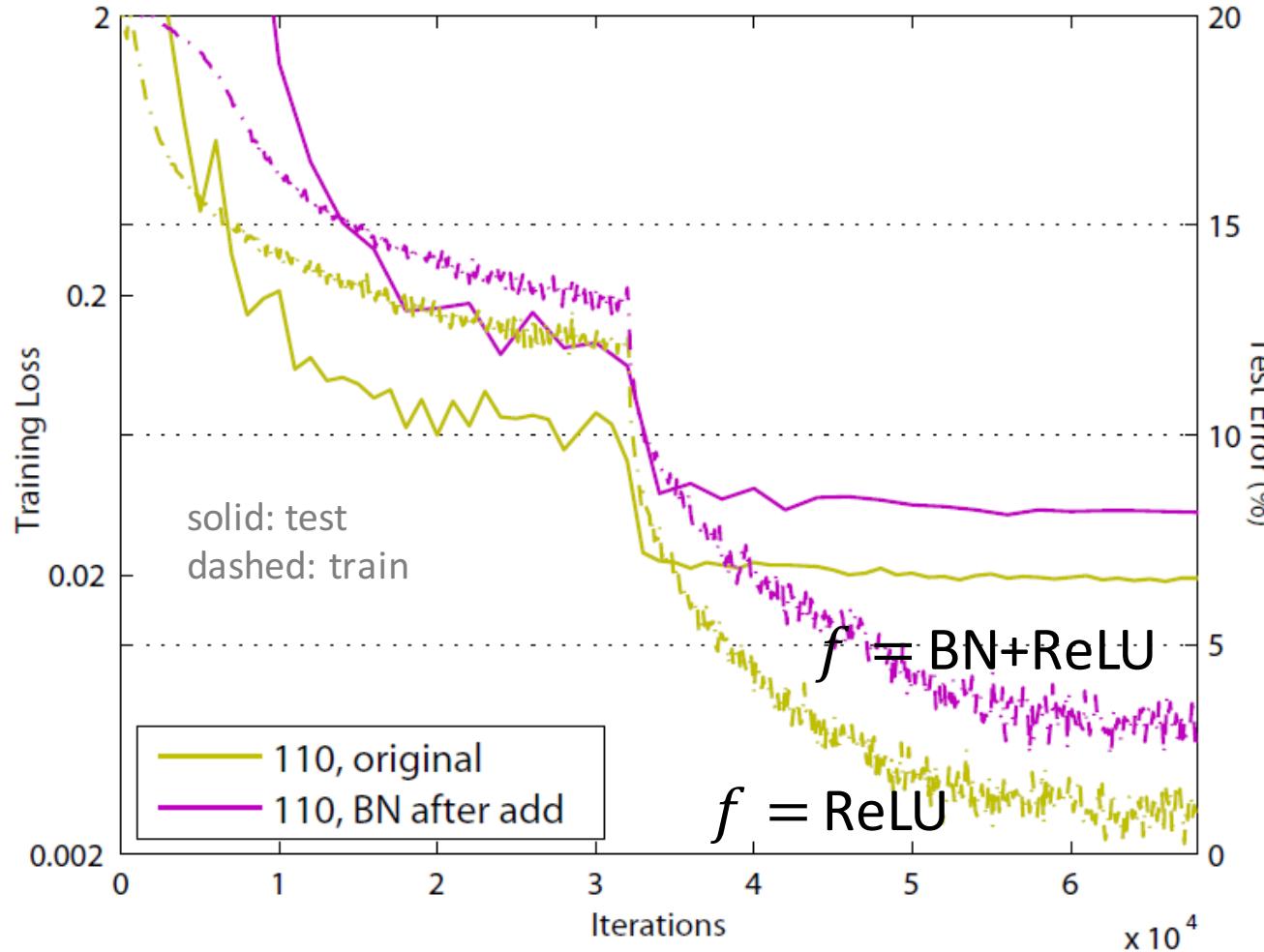
$f$  is ReLU  
(original ResNet)



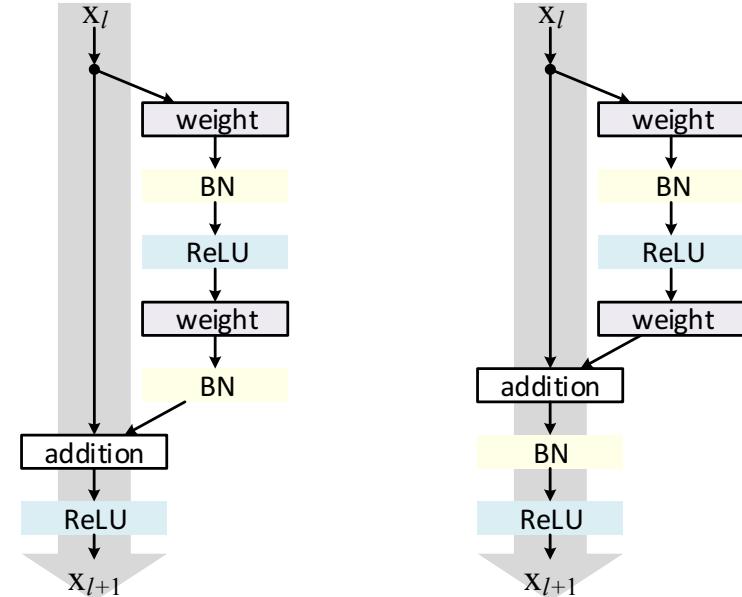
$f$  is BN+ReLU



$f$  is identity  
**(pre-activation** ResNet)

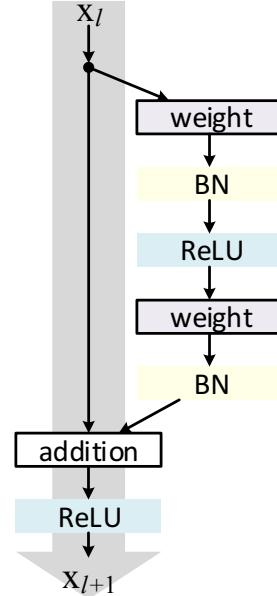
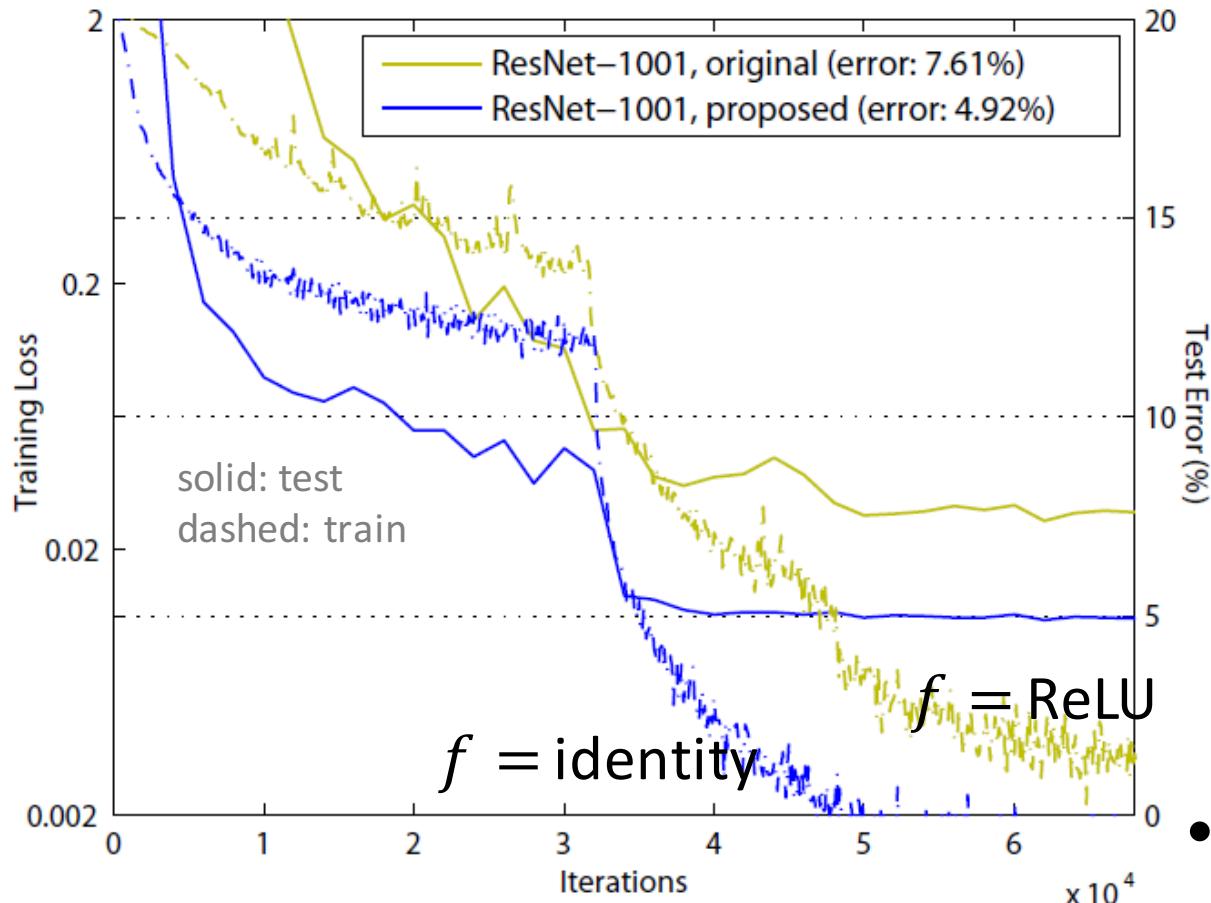


$$f = \text{ReLU} \quad f = \text{BN} + \text{ReLU}$$



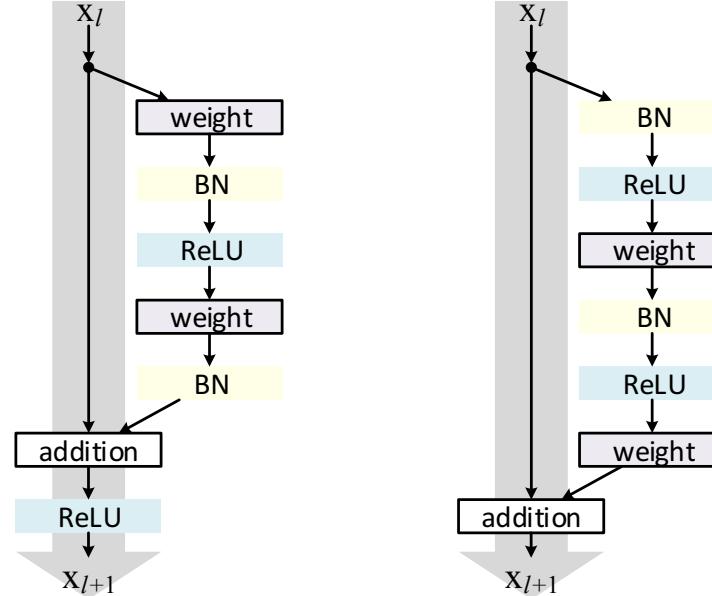
- BN could block prop
- Keep the shortest pass as smooth as possible

# 1001-layer ResNets on CIFAR-10



$$f = \text{ReLU}$$

$$f = \text{identity}$$



- ReLU could block prop when there are 1000 layers
- pre-activation design eases optimization (and improves generalization; see paper)

# Comparisons on CIFAR-10/100

CIFAR-10

method	error (%)
NIN	8.81
DSN	8.22
FitNet	8.39
Highway	7.72
ResNet-110 (1.7M)	6.61
ResNet-1202 (19.4M)	7.93
ResNet-164, pre-activation (1.7M)	5.46
<b>ResNet-1001</b> , pre-activation (10.2M)	<b>4.92</b> ( $4.89 \pm 0.14$ )

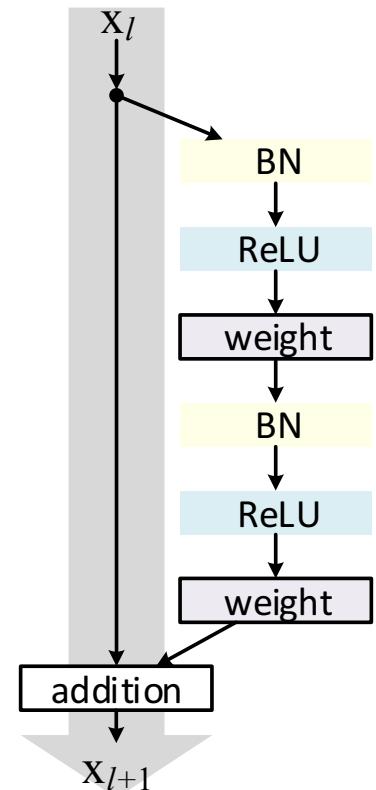
CIFAR-100

method	error (%)
NIN	35.68
DSN	34.57
FitNet	35.04
Highway	32.39
ResNet-164 (1.7M)	25.16
ResNet-1001 (10.2M)	27.82
ResNet-164, pre-activation (1.7M)	24.33
<b>ResNet-1001</b> , pre-activation (10.2M)	<b>22.71</b> ( $22.68 \pm 0.22$ )

\*all based on moderate augmentation

# Summary of observations

- Keep the shortest path as smooth as possible
  - by making  $h$  and  $f$  identity
  - forward/backward signals directly flow through this path
- Features of any layers are additive outcomes
- **1000-layer** ResNets can be easily trained and have better accuracy



# Applications

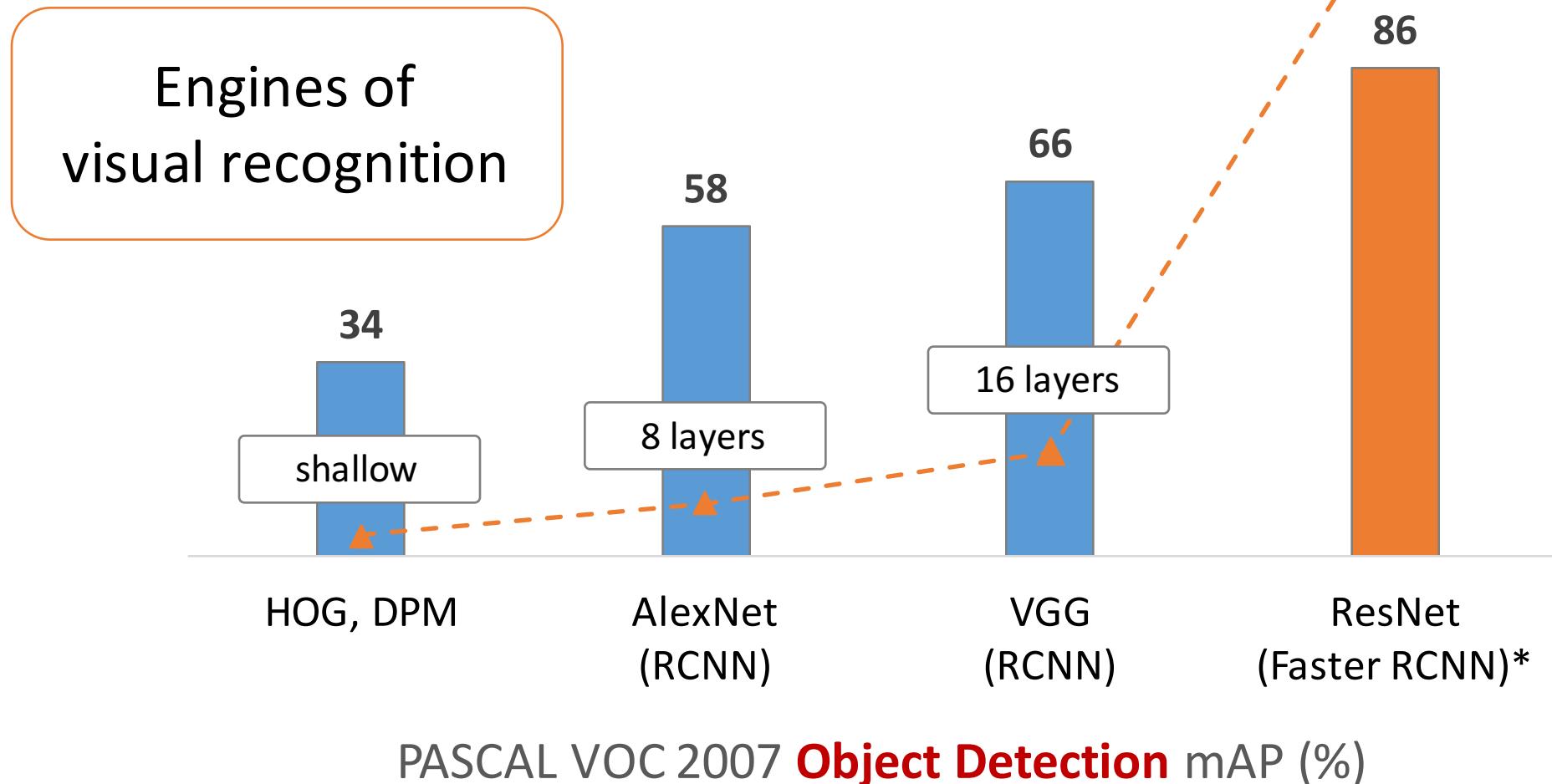
*“Features matter”*

*“Features matter.”* (quote [Girshick et al. 2014], the R-CNN paper)

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

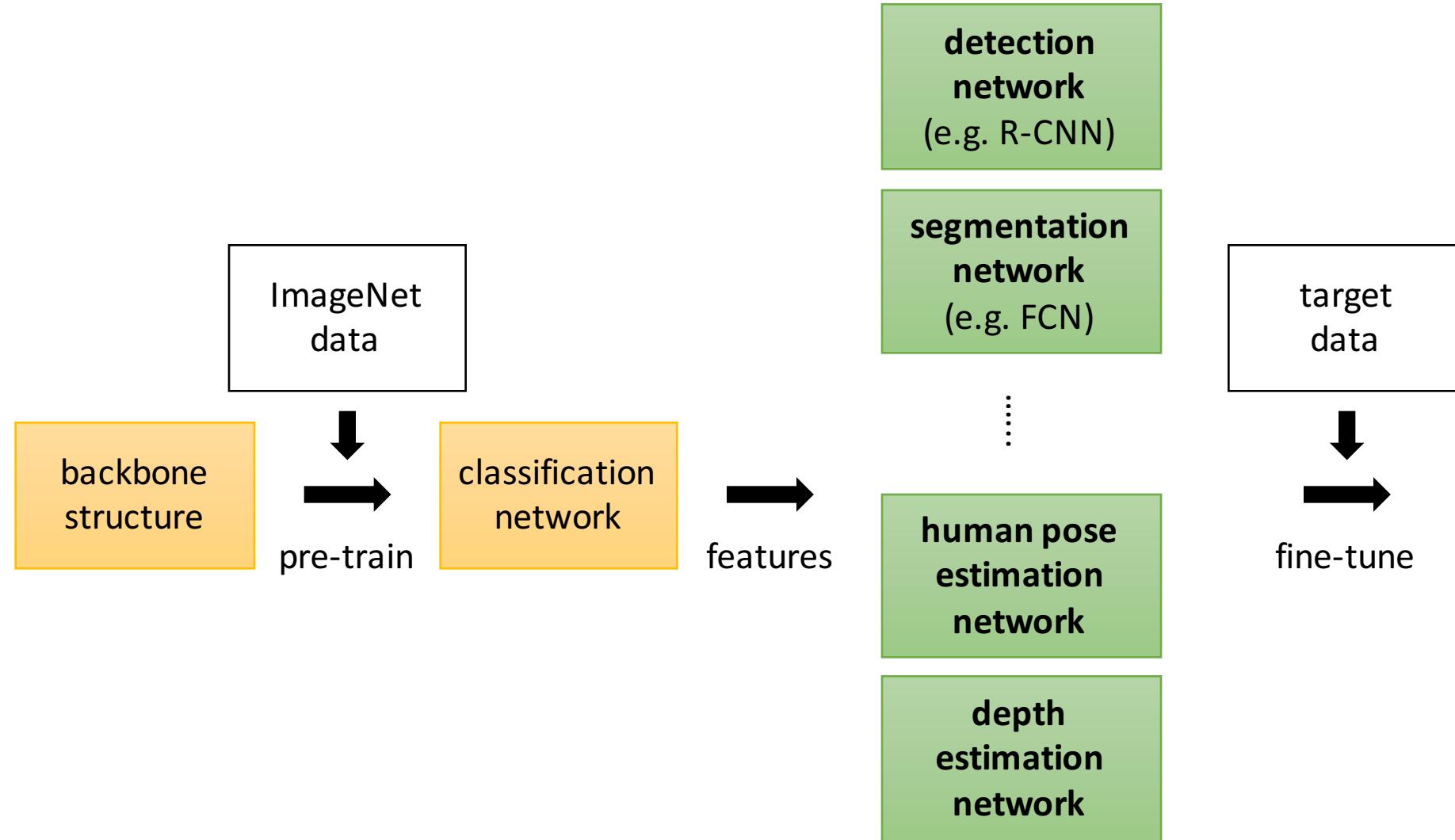
- Our results are all based on **ResNet-101**
- Deeper features are **well transferrable**

# Revolution of Depth



\*w/ other improvements & more data

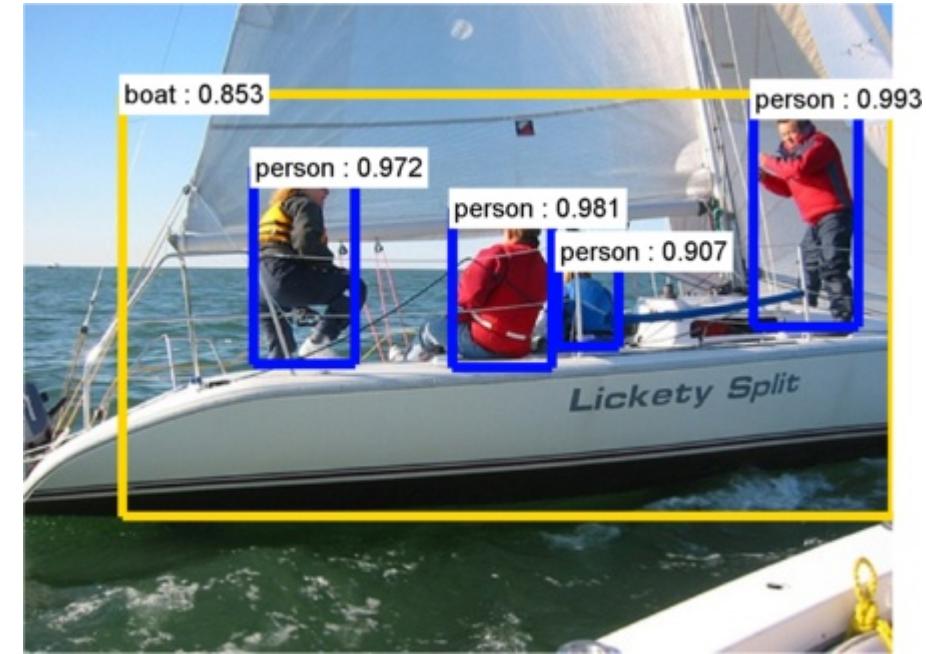
# Deep Learning for Computer Vision



# Example: Object Detection

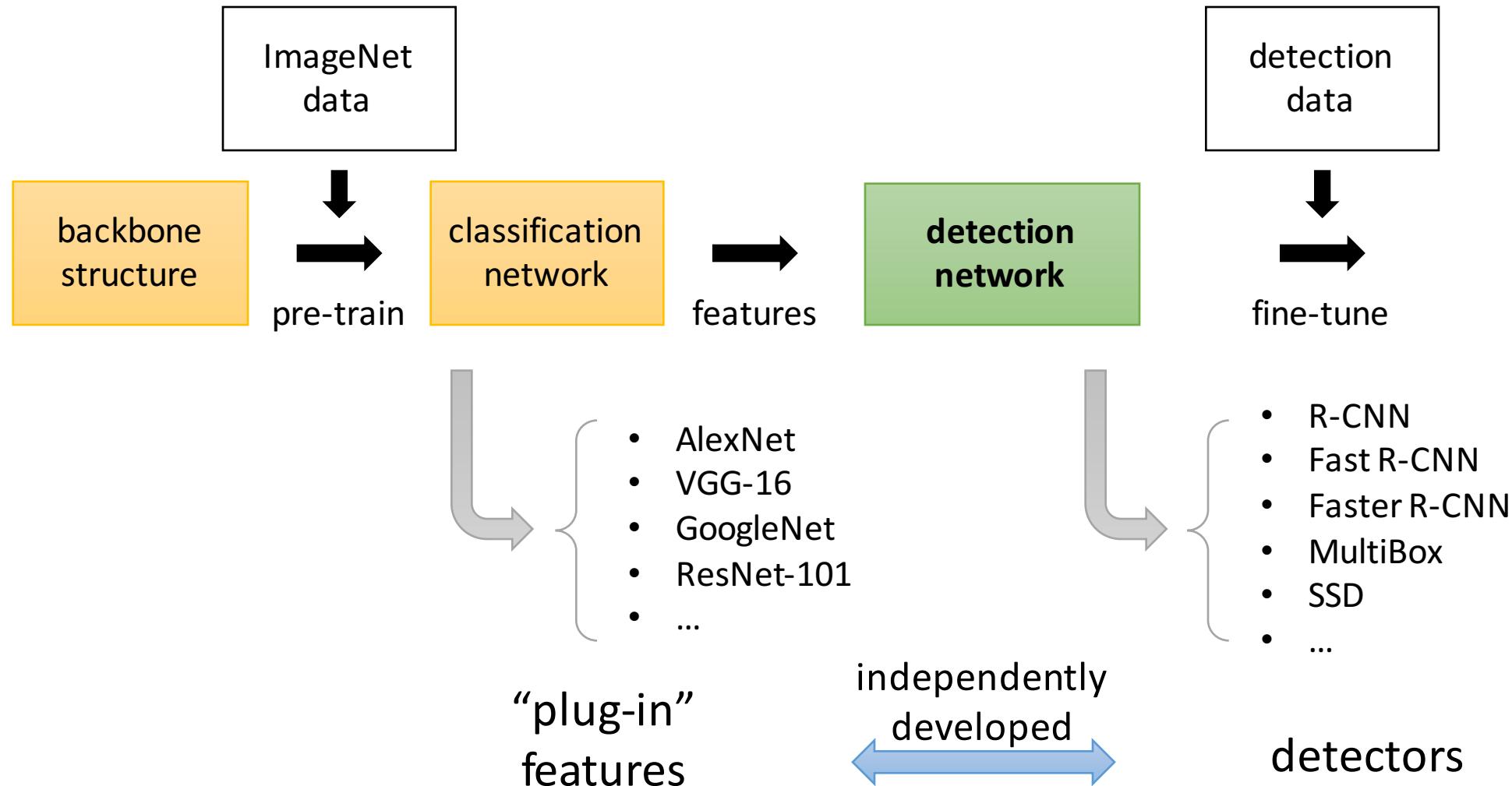


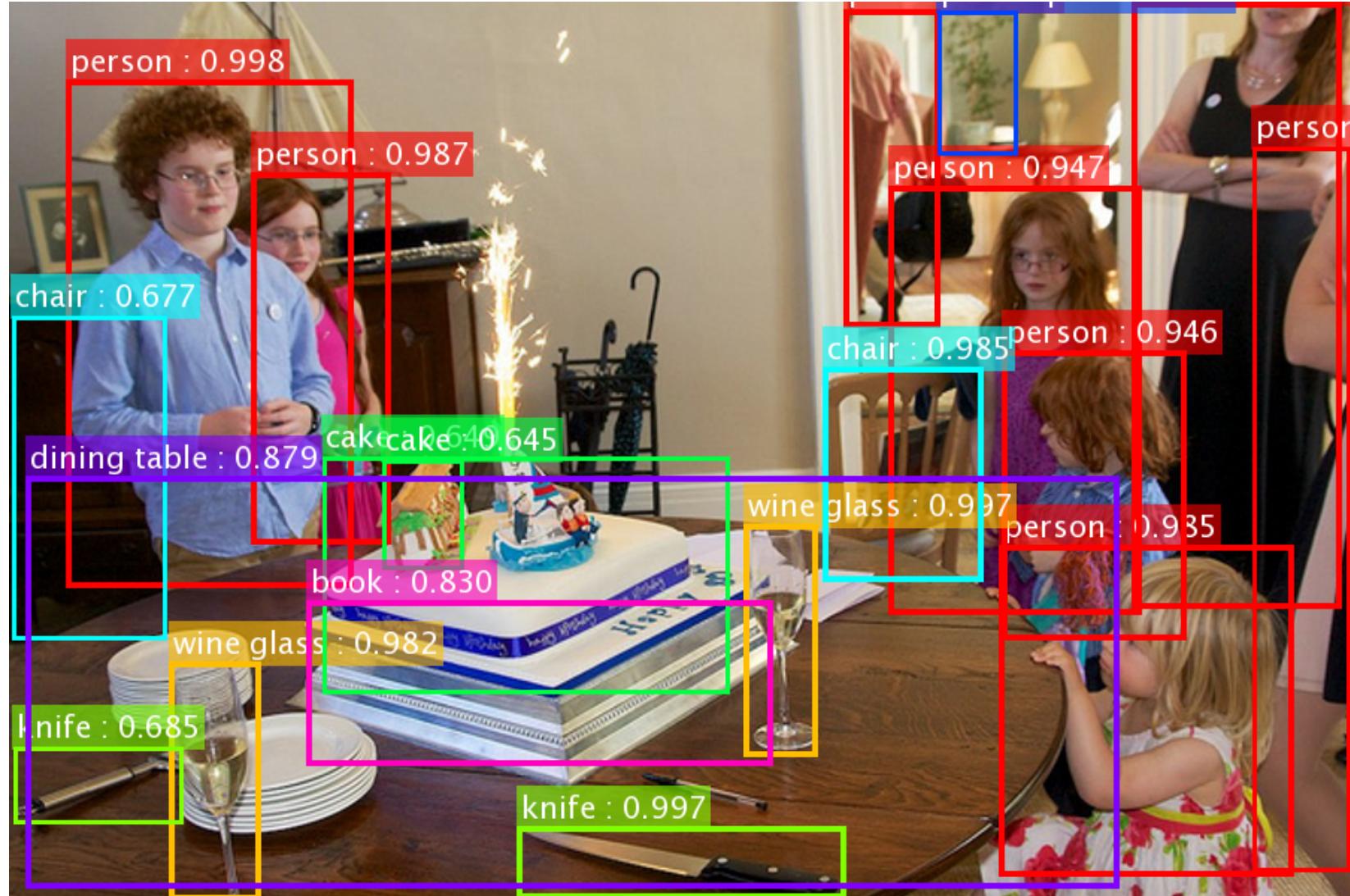
Image Classification  
(what?)



Object Detection  
(what + where?)

# Object Detection

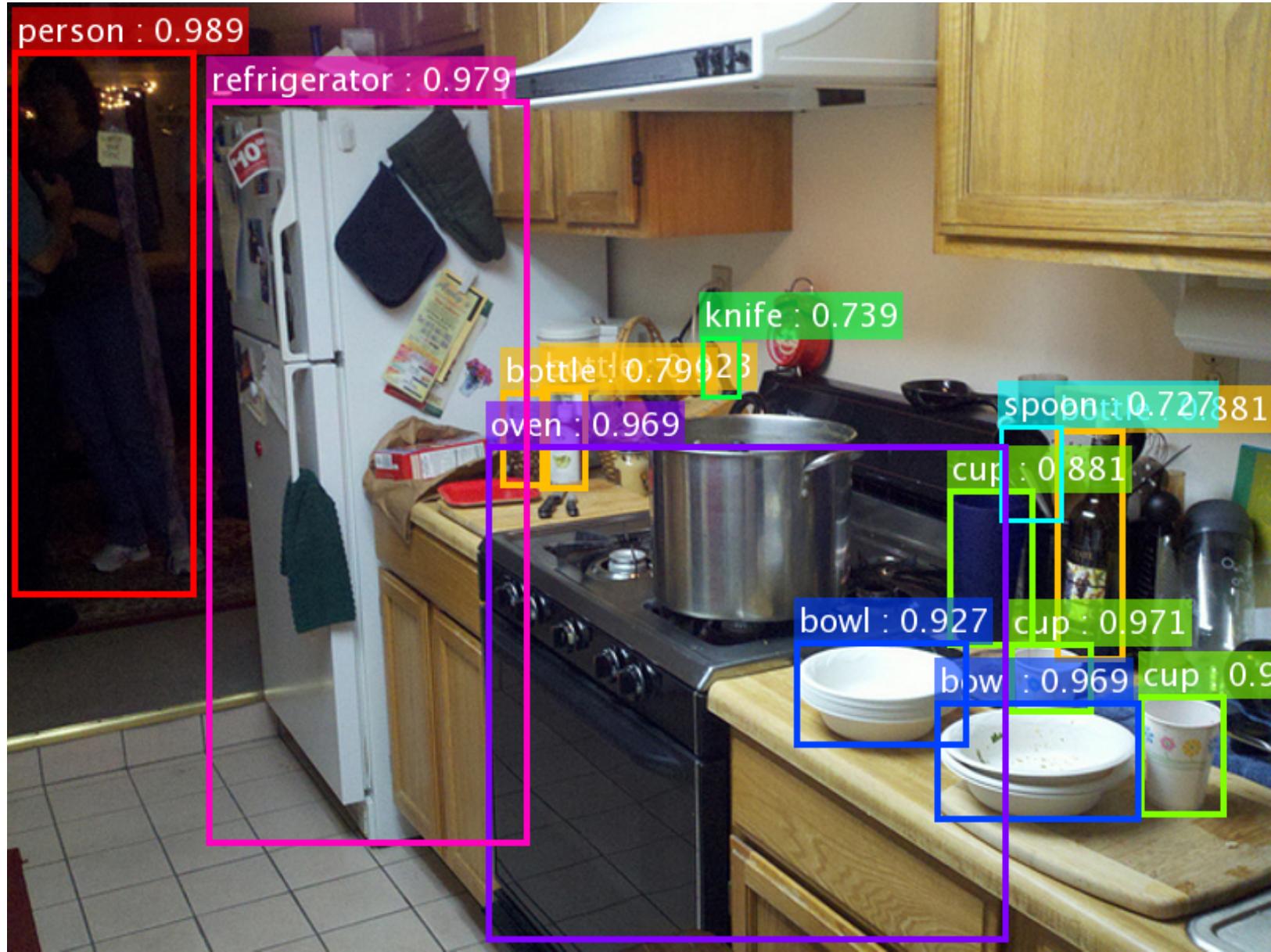




## ResNet's object detection result on COCO

\*the original image is from the COCO dataset

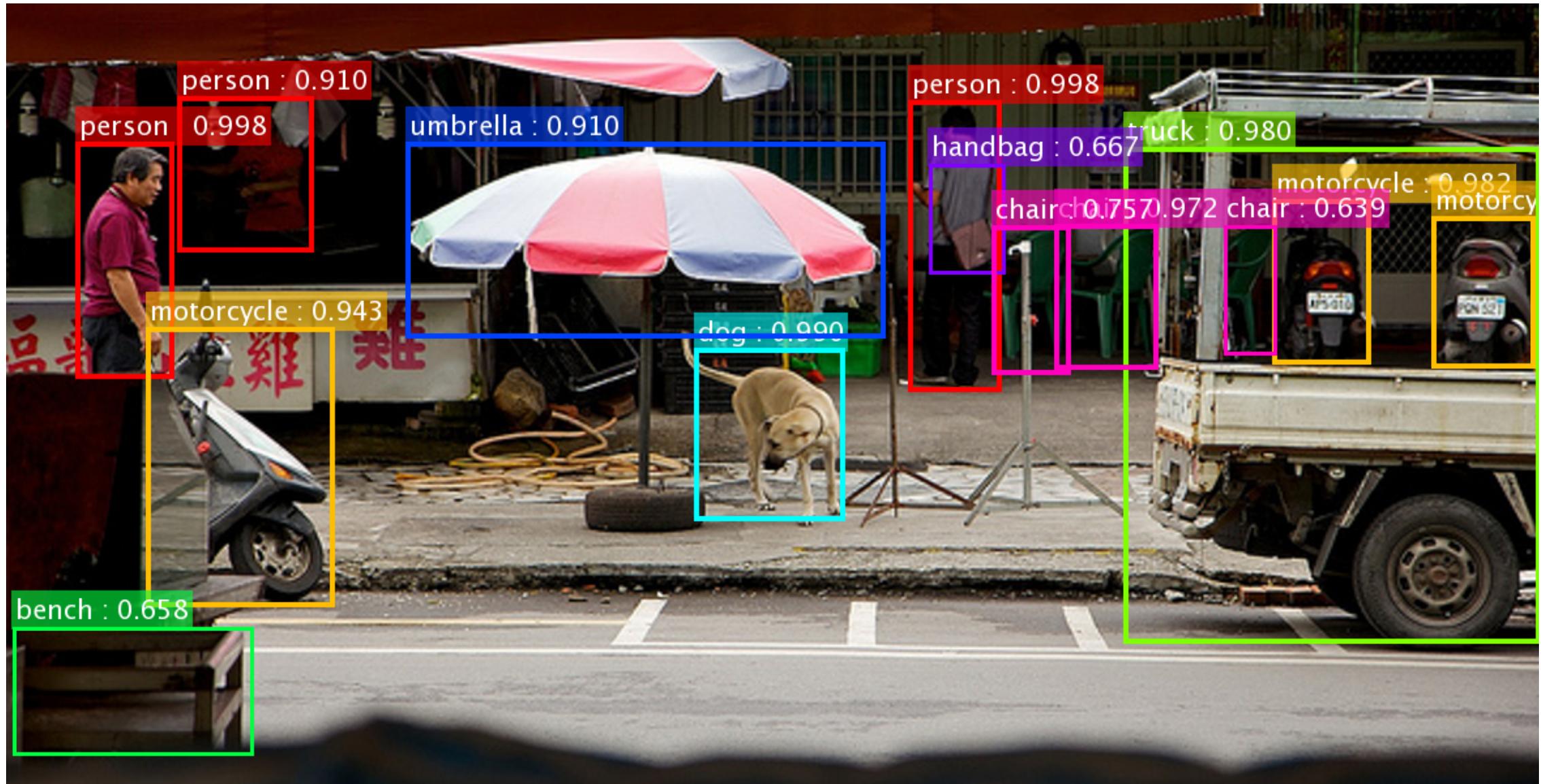
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.  
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.



\*the original image is from the COCO dataset

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

Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.



\*the original image is from the COCO dataset

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

Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.



this video is available online: <https://youtu.be/WZmSMkK9VuA>

Results on real video. Models trained on MS COCO (80 categories).  
(frame-by-frame; no temporal processing)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.  
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.

# Potential Applications

ResNets have shown outstanding or promising results on:

