

Deep Residual Learning

MSRA @ ILSVRC & COCO 2015 competitions

Kaiming He

with Xiangyu Zhang, Shaoqing Ren, Jifeng Dai, & Jian Sun

Microsoft Research Asia (MSRA)

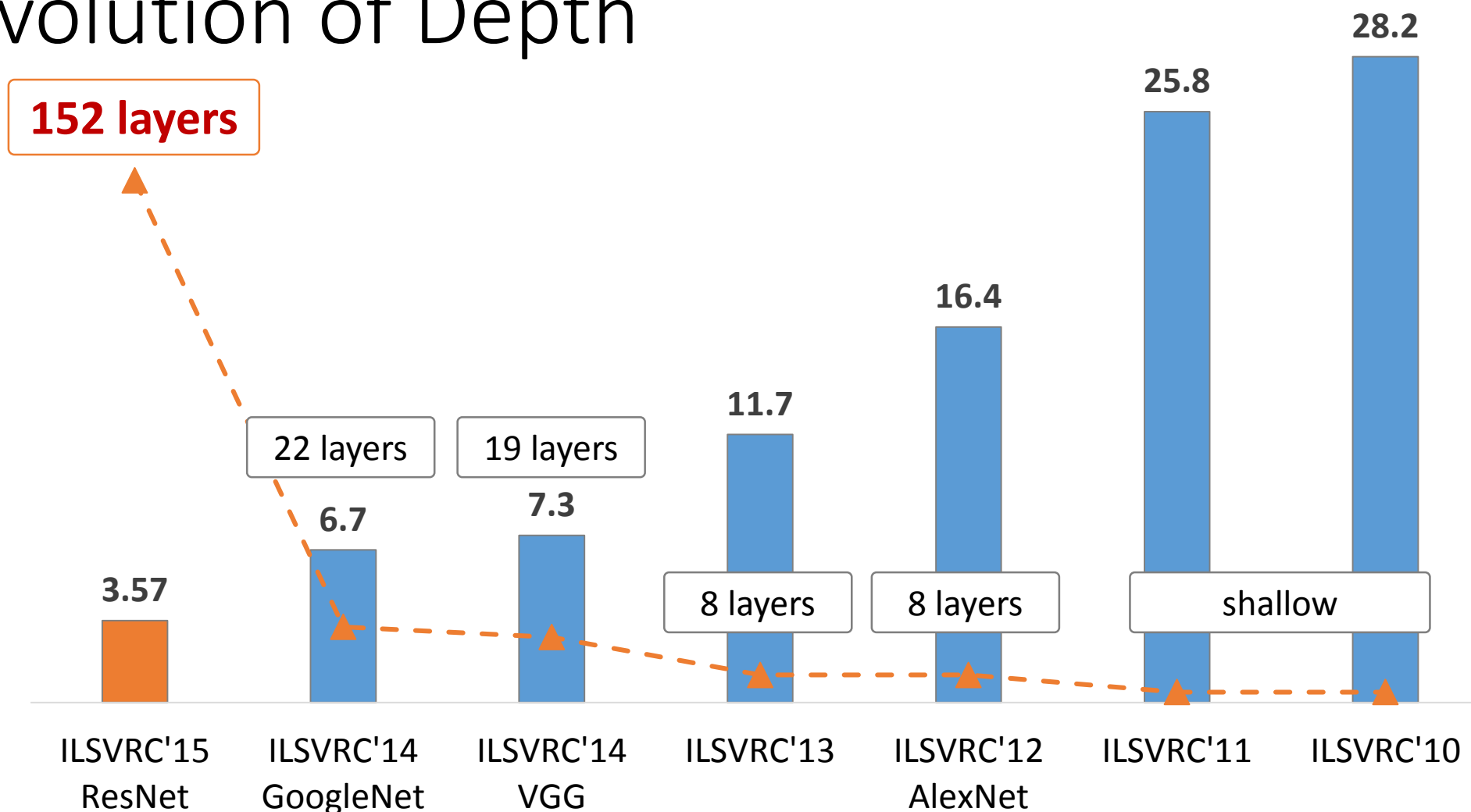
MSRA @ ILSVRC & COCO 2015 Competitions

- **1st places in all five main tracks**

- ImageNet Classification: “*Ultra-deep*” (quote Yann) **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

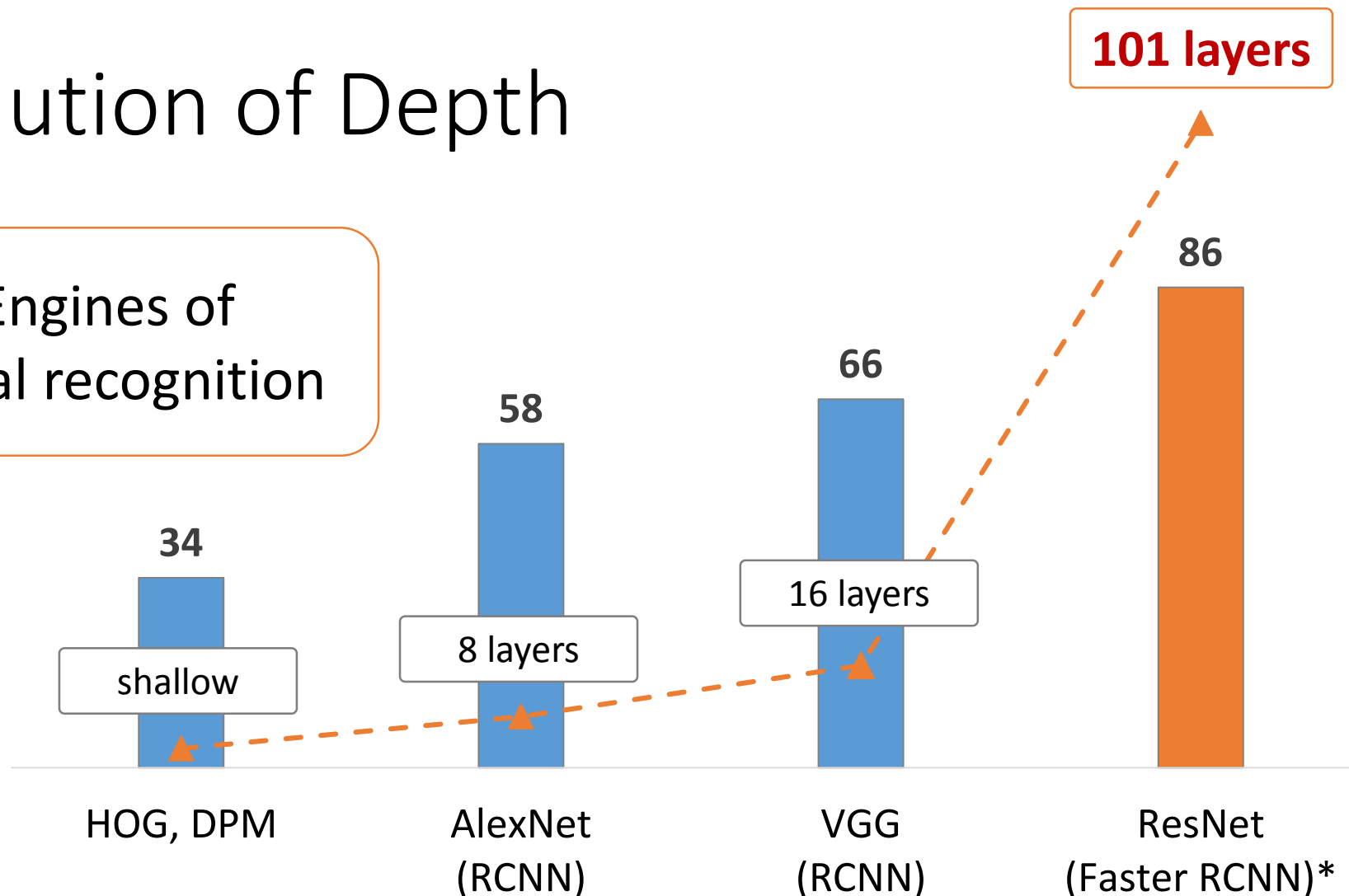


ImageNet Classification top-5 error (%)

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

Revolution of Depth

Engines of
visual recognition



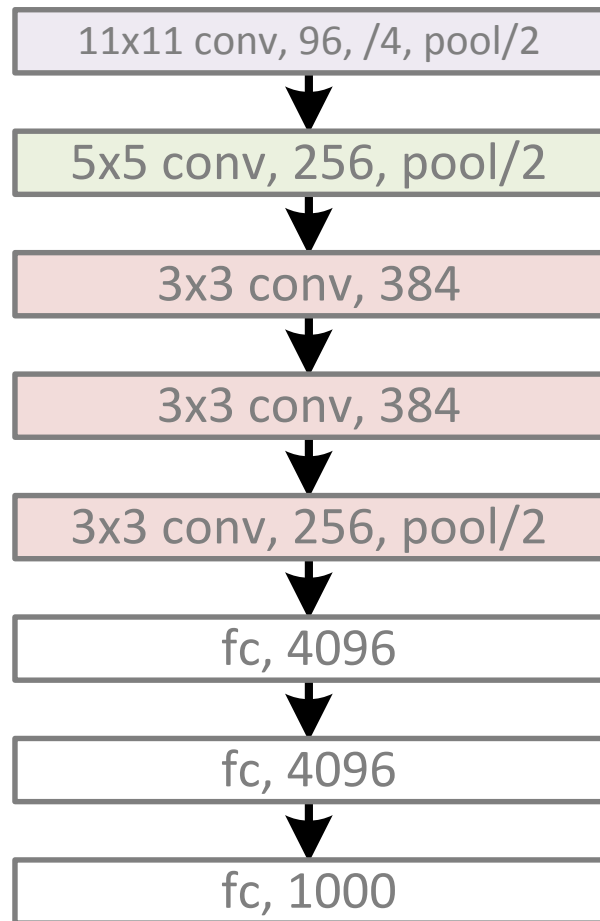
PASCAL VOC 2007 **Object Detection** mAP (%)

*w/ other improvements & more data

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

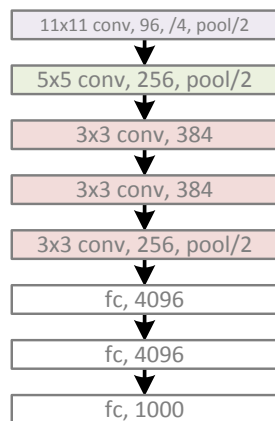
Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)

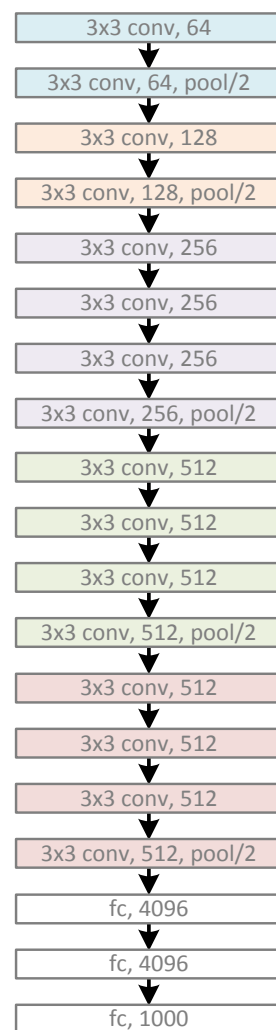


Revolution of Depth

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)

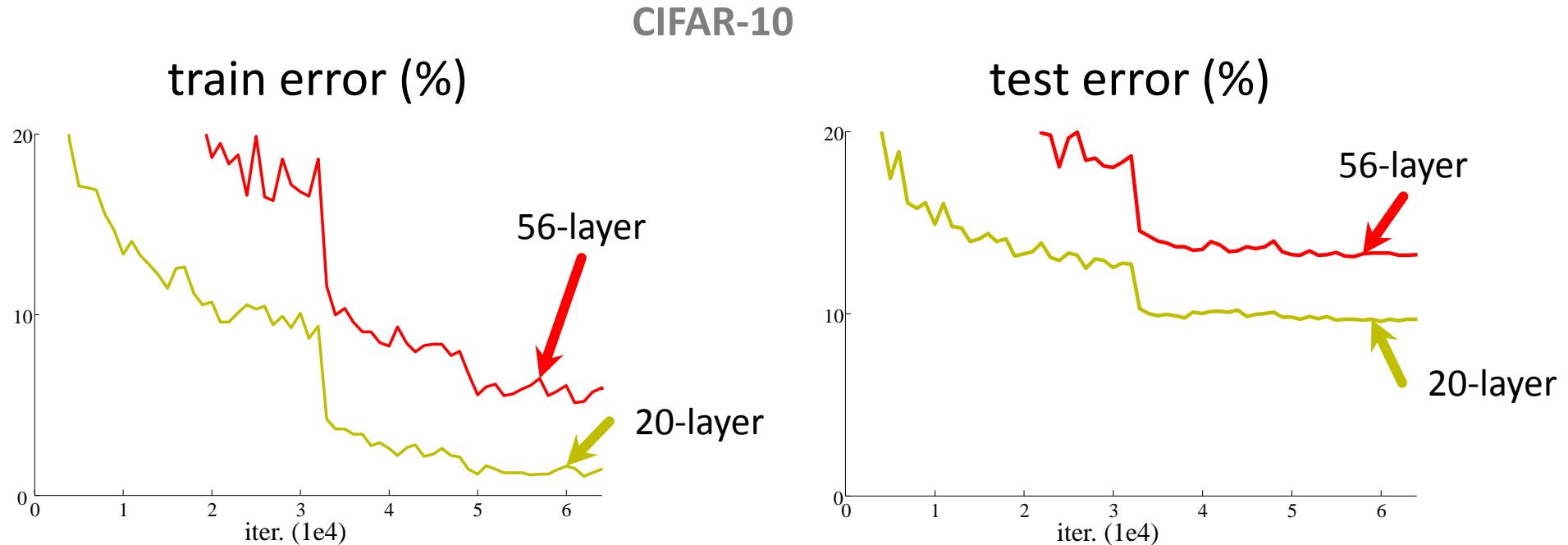


ResNet, **152 layers**
(ILSVRC 2015)



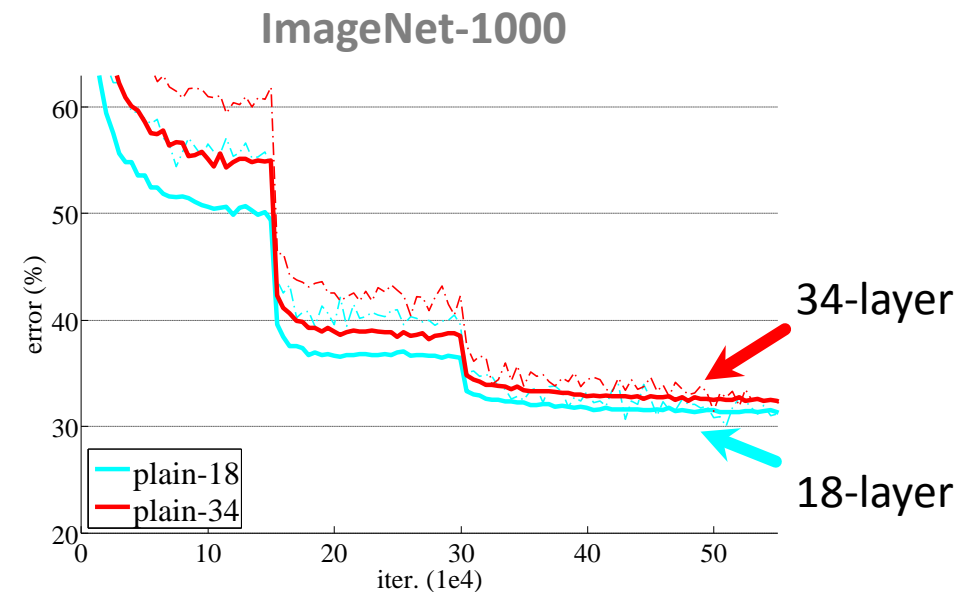
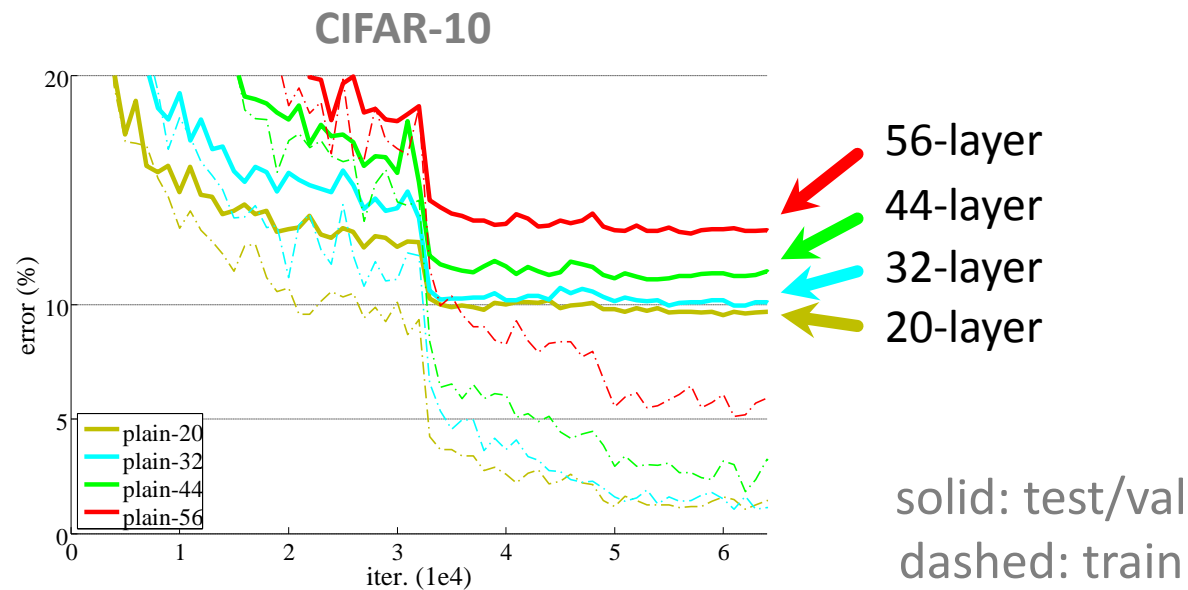
Is learning better networks
as simple as stacking more layers?

Simply stacking layers?



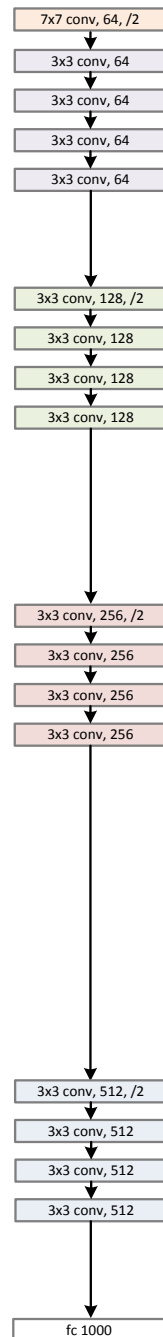
- *Plain* nets: stacking 3x3 conv layers...
- 56-layer net has **higher training error** and test error than 20-layer net

Simply stacking layers?

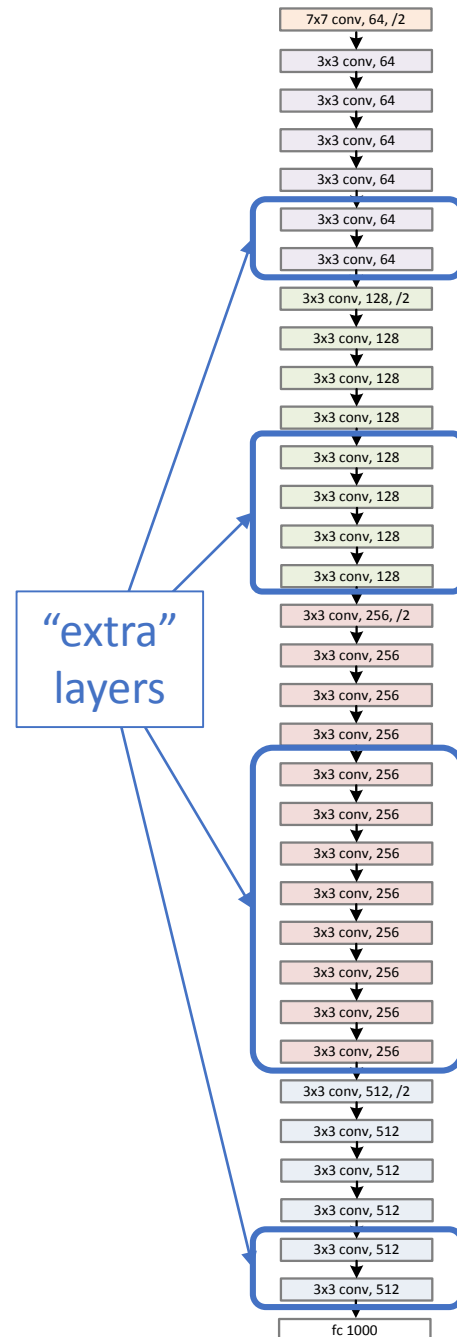


- “Overly deep” plain nets have **higher training error**
- A general phenomenon, observed in many datasets

a shallower model
(18 layers)



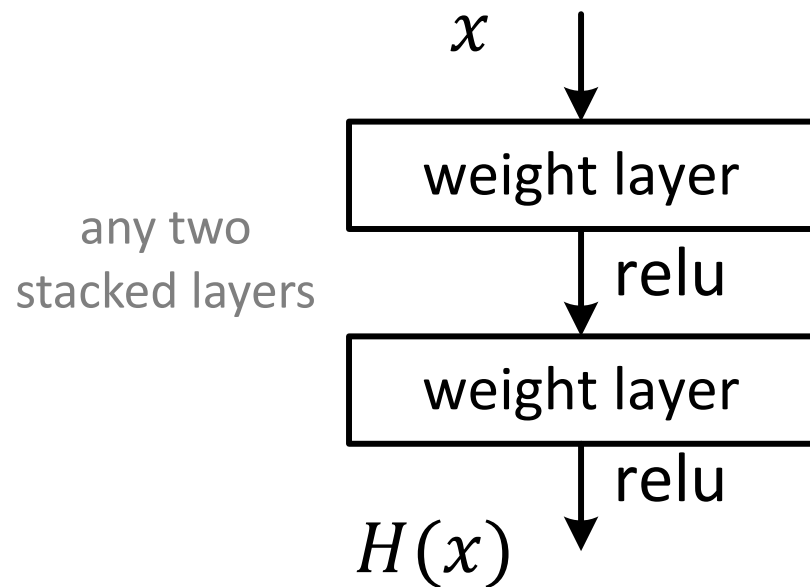
a deeper counterpart
(34 layers)



- 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

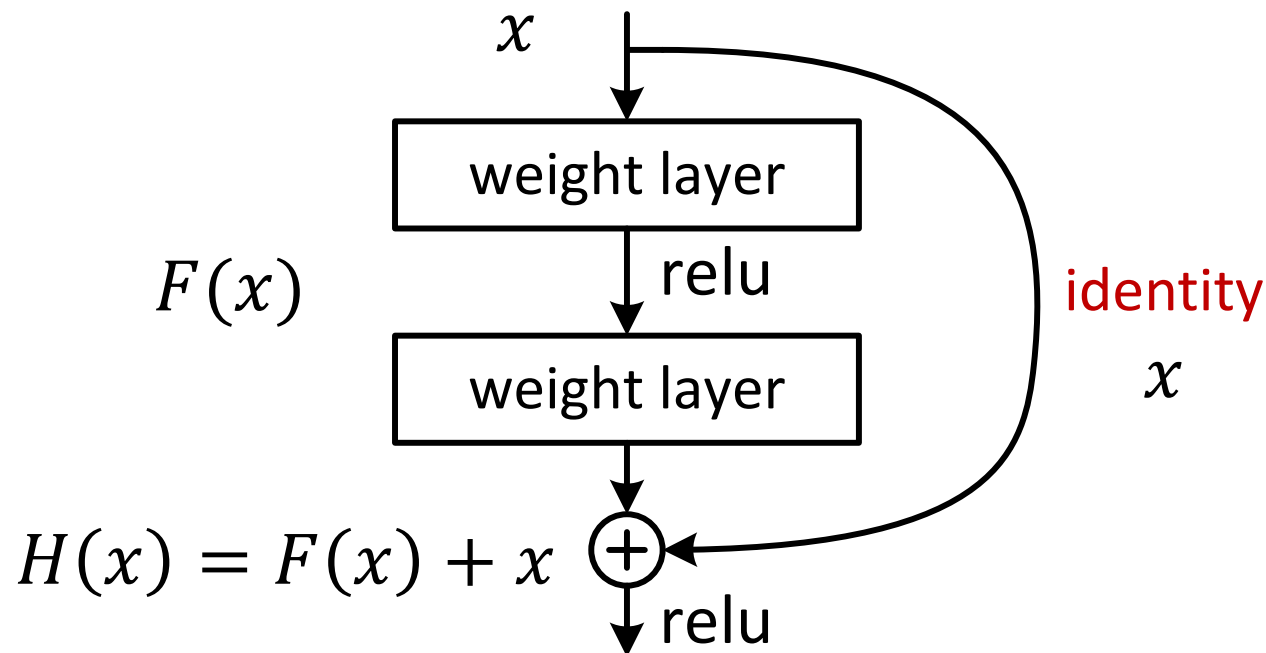
- Plain net



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

Deep Residual Learning

- Residual net



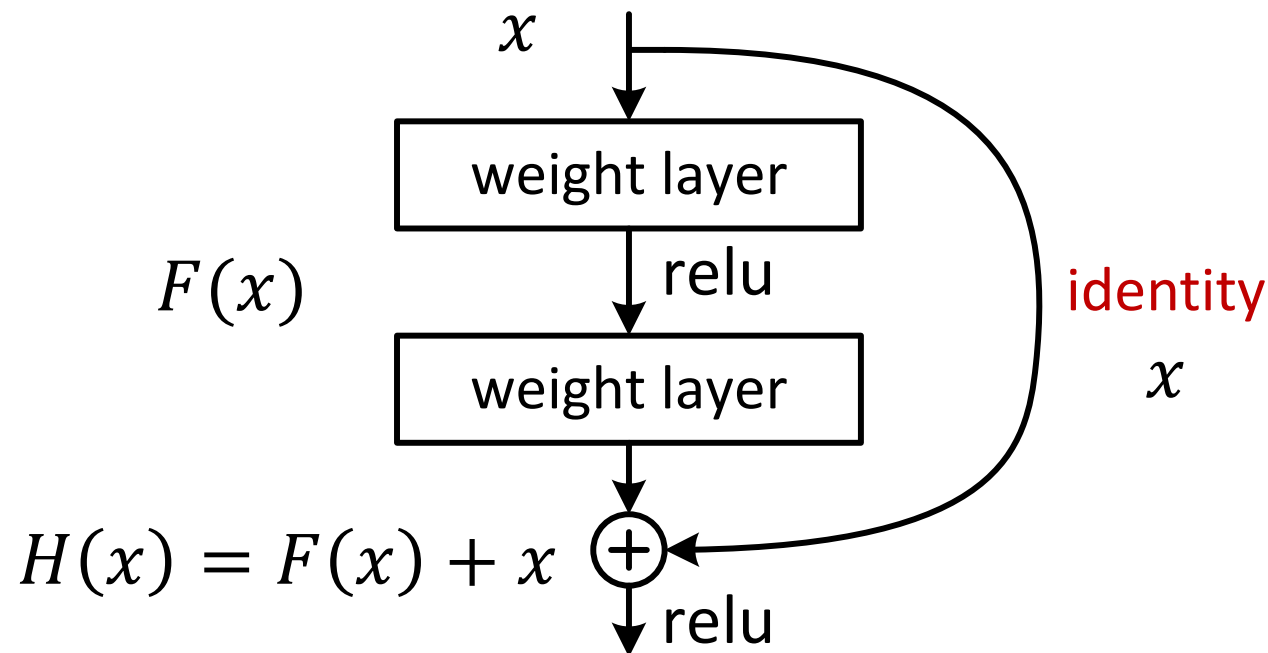
$H(x)$ is any desired mapping,
~~hope the 2 weight layers fit $H(x)$~~

hope the 2 weight layers fit $F(x)$

$$\text{let } H(x) = F(x) + x$$

Deep Residual Learning

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

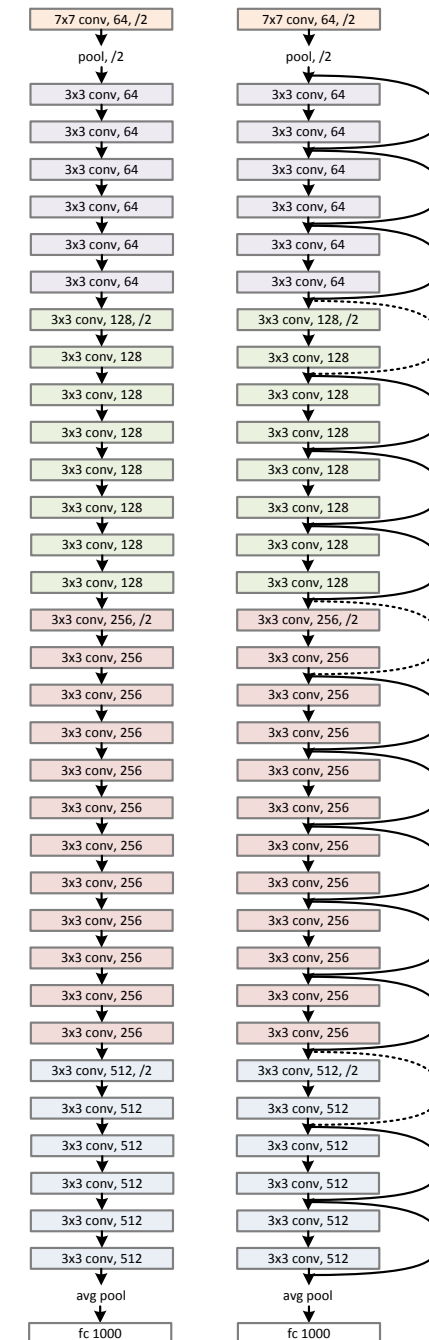


- If identity were optimal, easy to set weights as 0
- If optimal mapping is closer to identity, easier to find small fluctuations

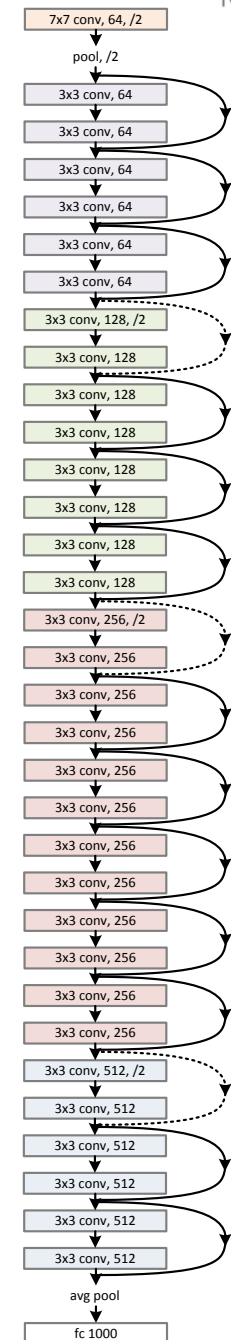
Network “Design”

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

plain net



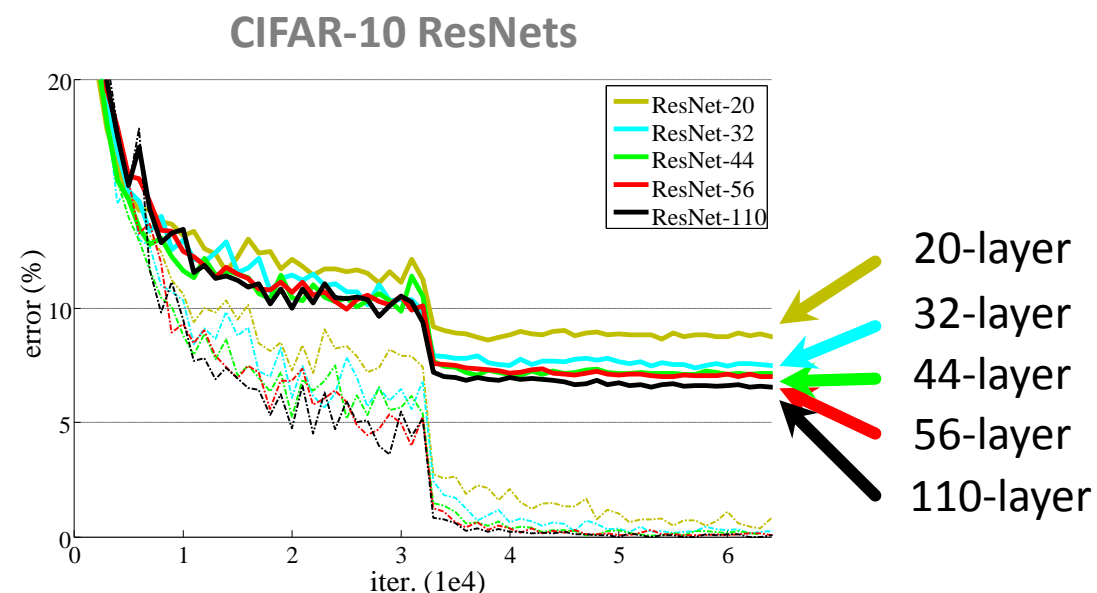
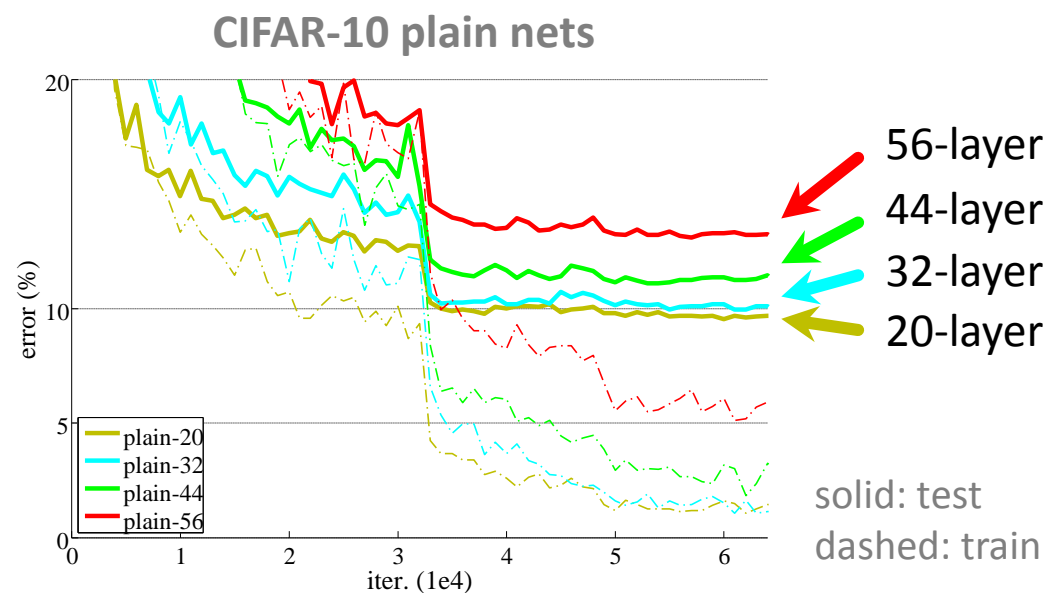
ResNet



Training

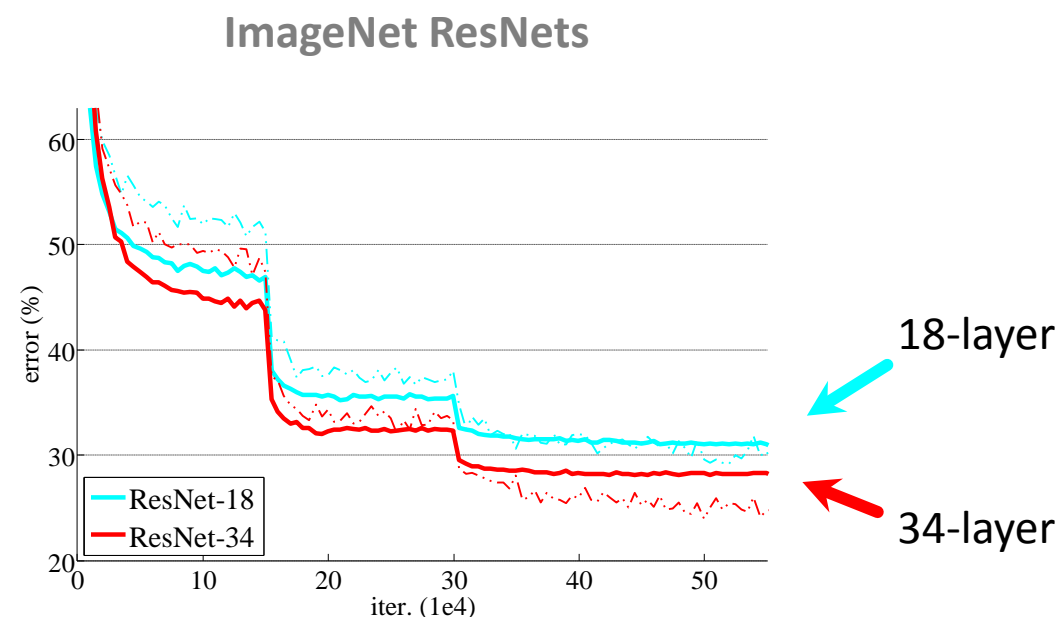
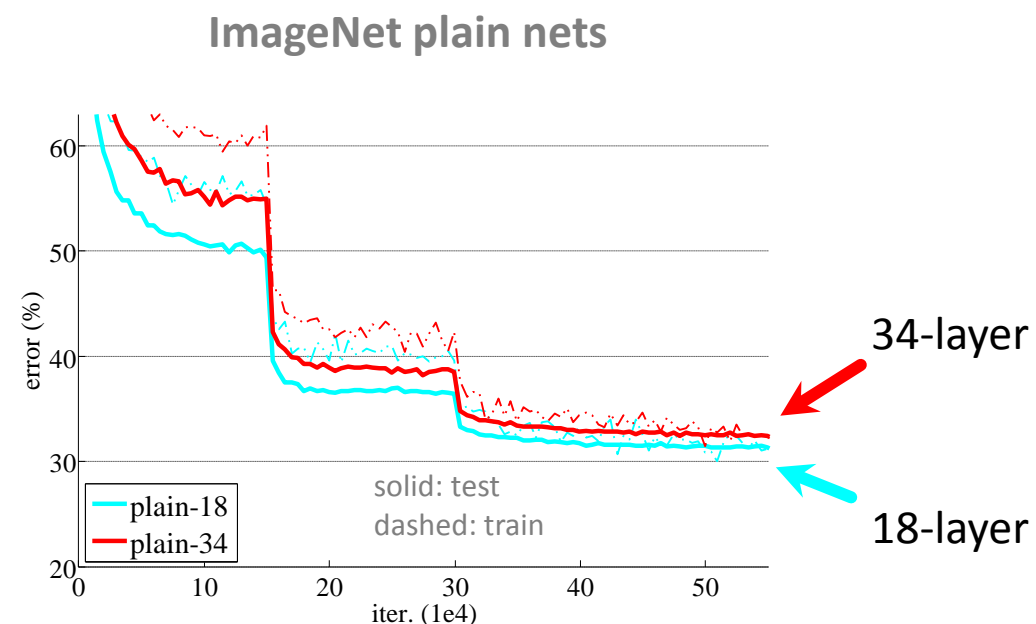
- All plain/residual nets are trained **from scratch**
- All plain/residual nets use Batch Normalization
- Standard hyper-parameters & augmentation

CIFAR-10 experiments



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

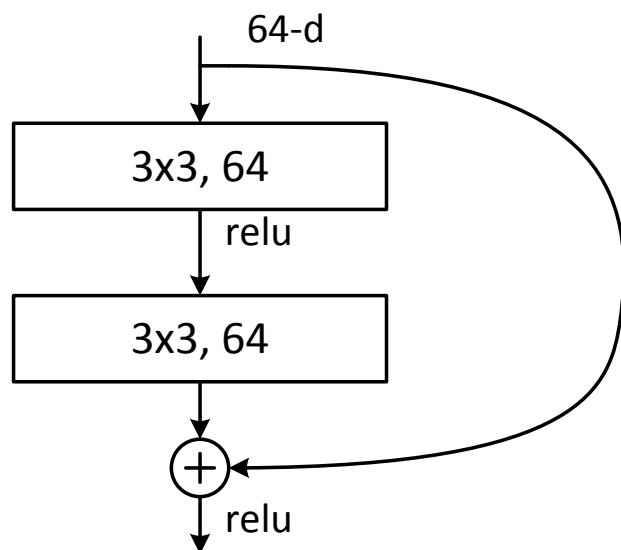
ImageNet experiments



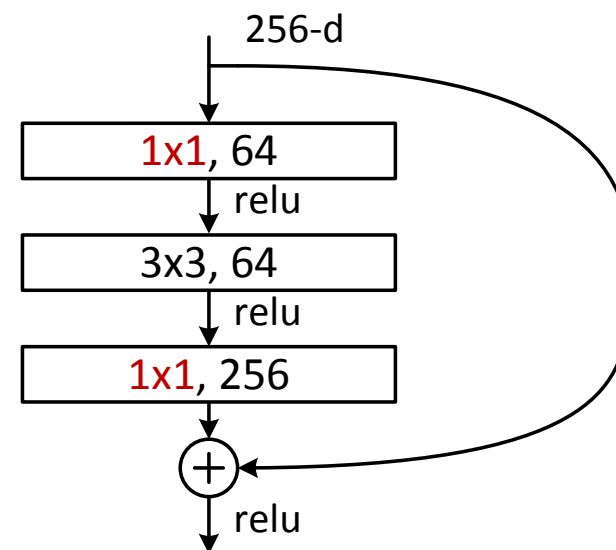
- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

ImageNet experiments

- A practical design of going deeper



all-3x3



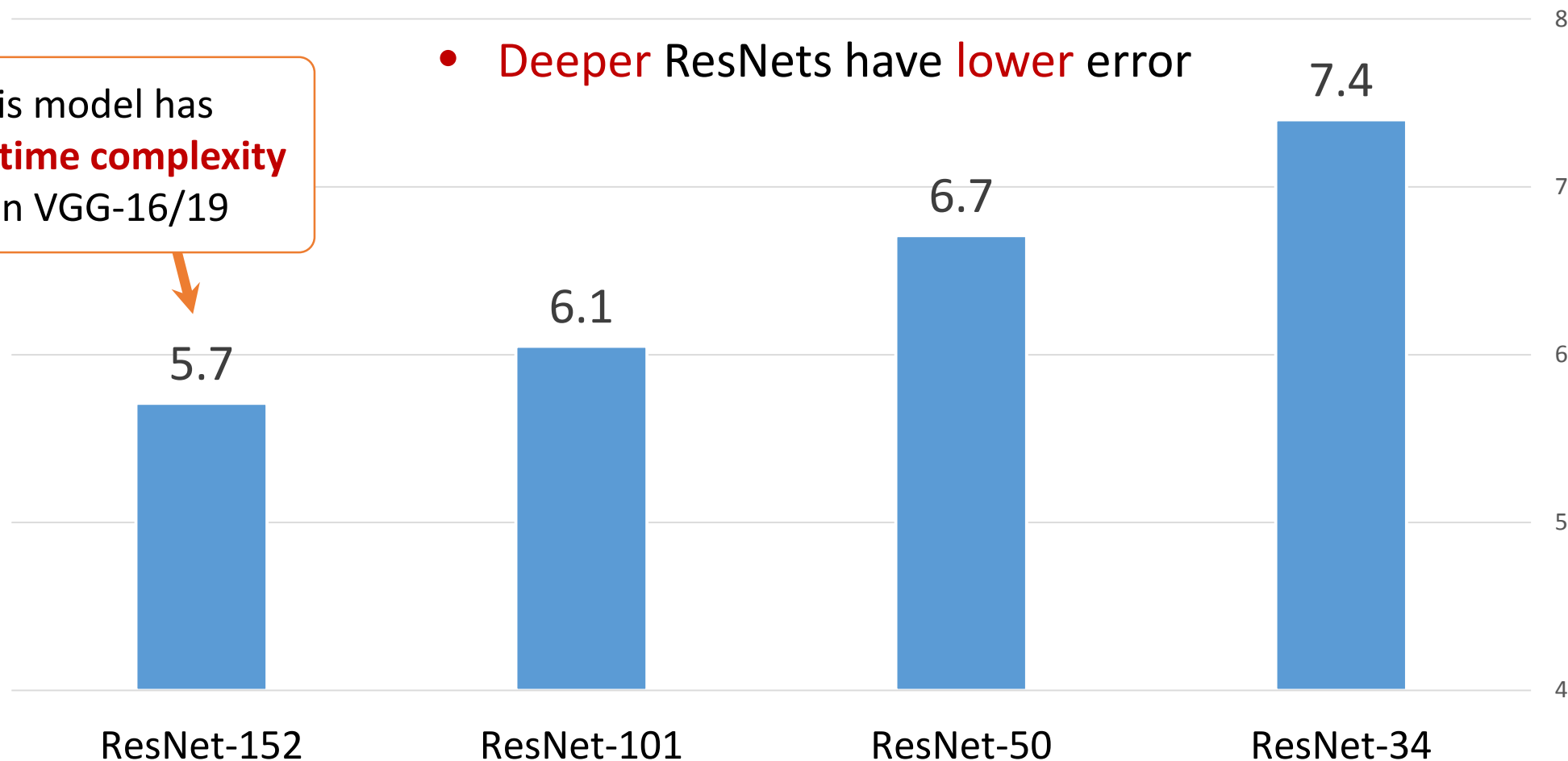
bottleneck

(for ResNet-50/101/152)

ImageNet experiments

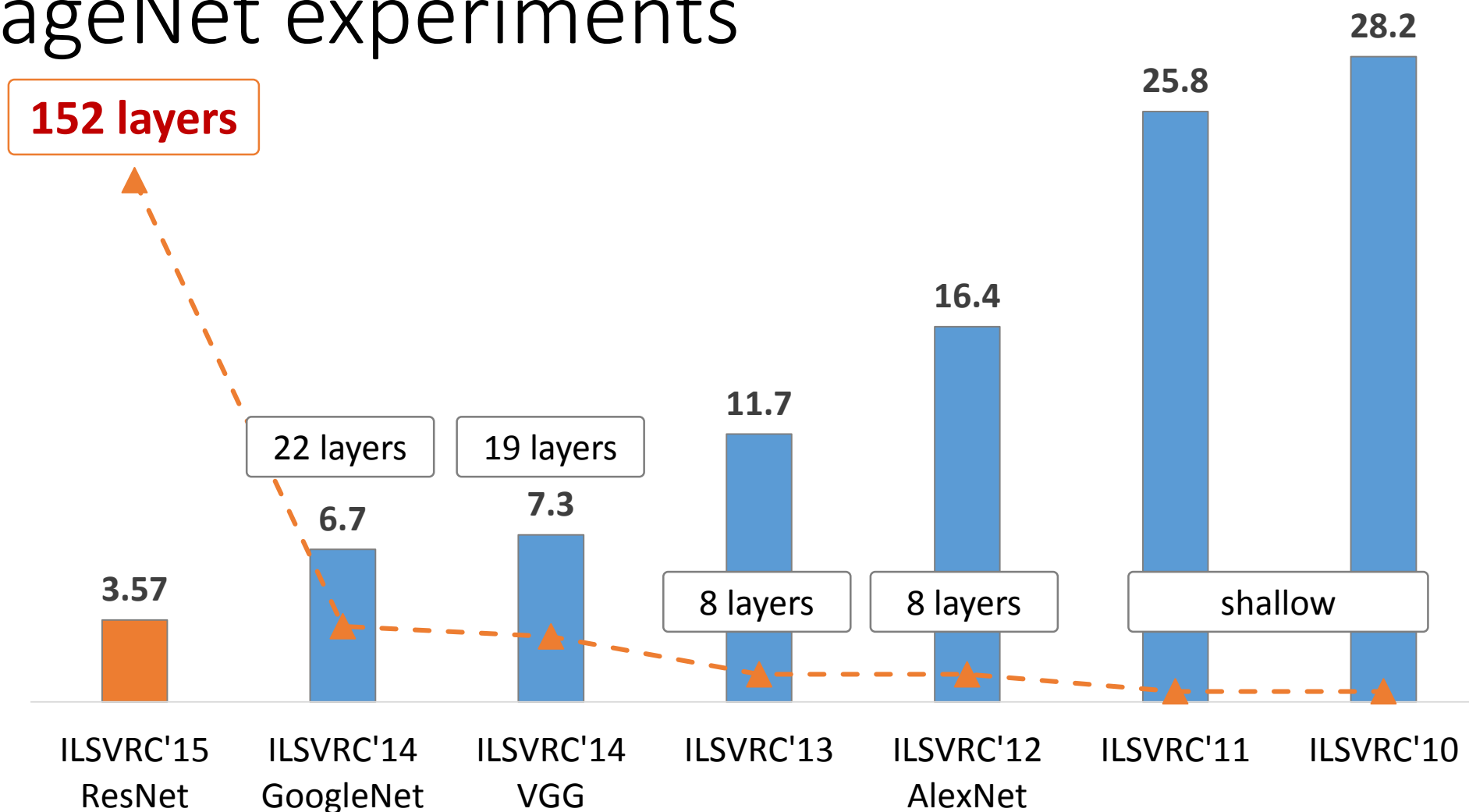
- Deeper ResNets have **lower** error

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



10-crop testing, top-5 val error (%)

ImageNet experiments



ImageNet Classification top-5 error (%)

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