Residual Networks Behave Like Ensembles of Relatively Shallow Networks

Andreas Veit Michael Wilber Serge Belongie



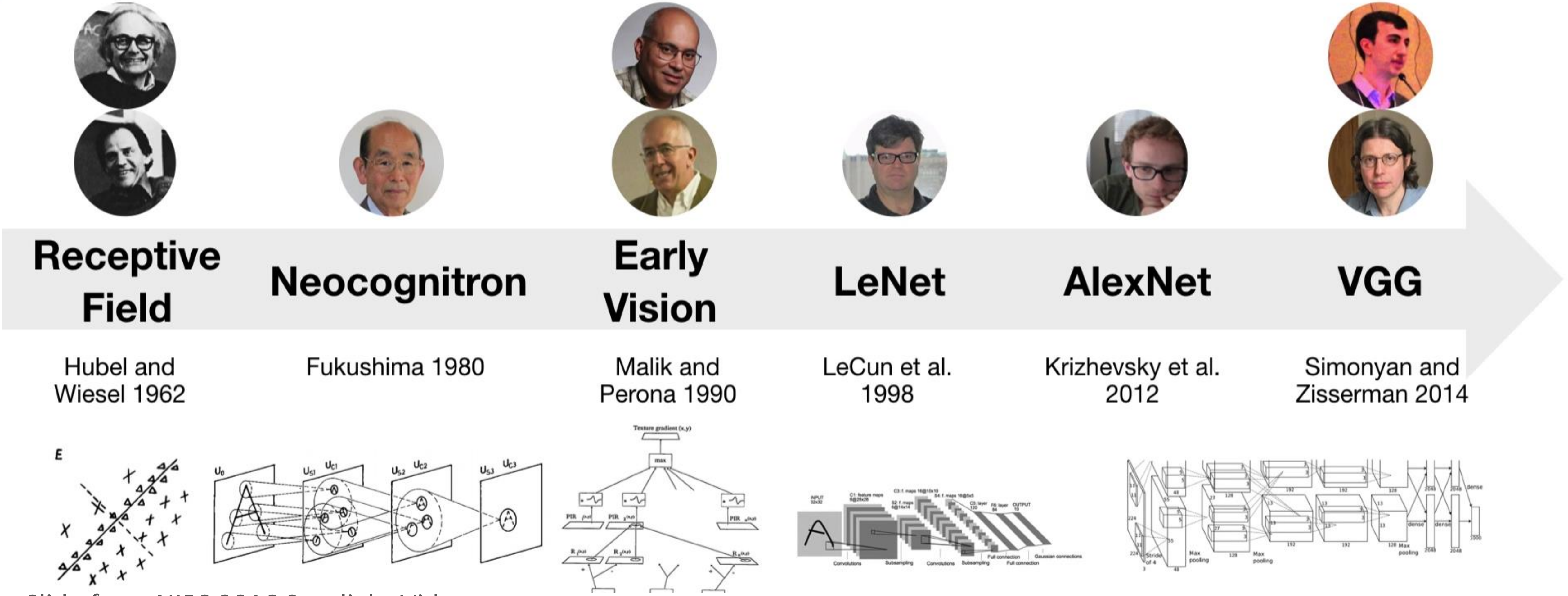
05/2016 ARXIV, Advances in Neural Information Processing Systems 29 (NIPS 2016)

Overview

- Introduction
- Background
  - Previous investigations on neural networks
  - Deep Residual Networks (ResNets): 10 to 100 layers
  - Importance of Identity Mapping: 100 to 1000 layers
- Key takeaway 1
  - Existing systems are feed-forward, with only one path.
  - ResNets contain many paths instead, shown by the «unraveled view».
- Key takeaway 2
  - Path lengths are binomially distributed.
  - |Gradient| decreases exponentially with increasing path length.
  - Only short paths contribute gradient during training.
- Q & A

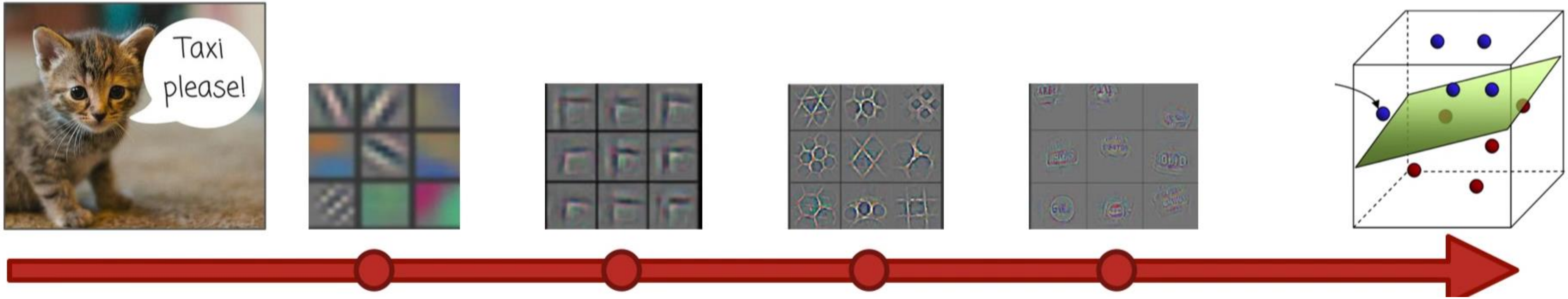
Introduction

Sequential vision pipelines **influence our thinking.**



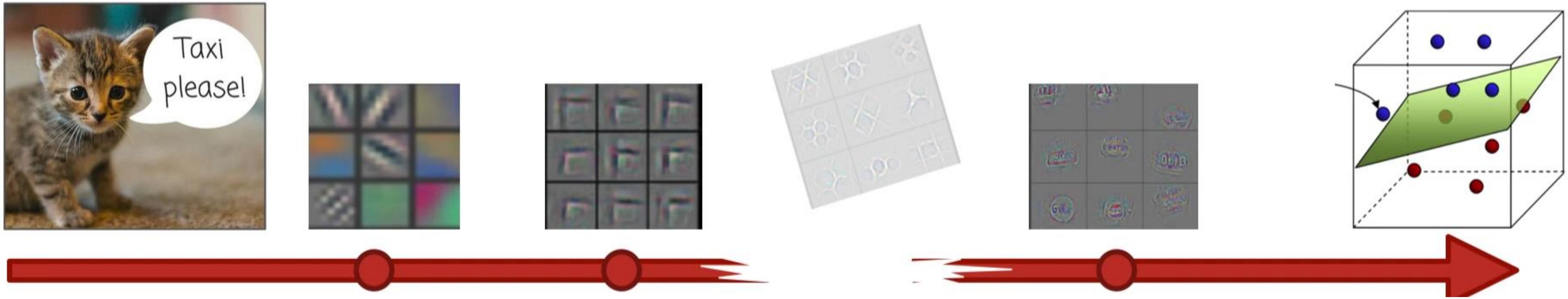
Slide from NIPS 2016 Spotlight Video

Existing systems are feed-forward, with only one path



Slide from NIPS 2016 Spotlight Video

Existing systems are feed-forward, with only one path



What happens when we delete a step?

Slide from NIPS 2016 Spotlight Video

Existing systems are feed-forward, with only one path

Slide from NIPS 2016 Spotlight Video

Existing systems are feed-forward, with only one path

Slide from NIPS 2016 Spotlight Video

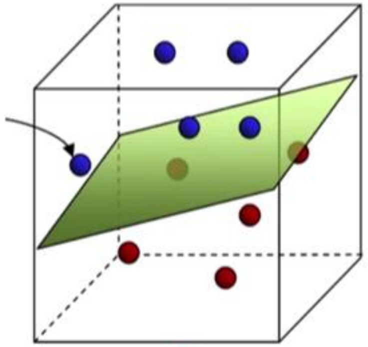
Existing systems are feed-forward, with only one path

Any alternatives?

Slide adapted from NIPS 2016 Spotlight Video

Existing systems are feed-forward, with only one path

Any alternatives?
ResNets!



Slide adapted from NIPS 2016 Spotlight Video

What is the reason behind ResNets' increased performance?

Hypothesis by He et al. 2016[†]:

«via a simple but essential concept – going deeper.»

Veit et al. 2016:

A complementary explanation…

[†] He et al. 2016, "Identity Mappings in Deep Residual Networks"

Background

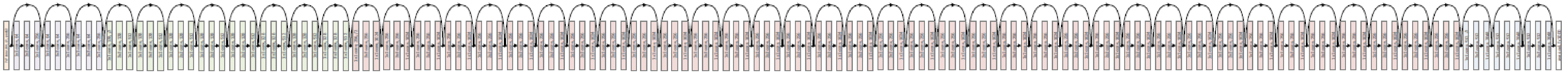
Previous investigations:
What do we know about neural networks?

- Shown by Bengio et al. 1994 and Hochreiter 1991:
  - Length of paths affect magnitude of the gradient during backpropagation.

- Lesion studies on AlexNet by Yosinski et al. 2014:
  - Early layers little co-adaptation: General, applicable to many datasets and tasks
  - Later layers have more co-adaptation: Specific

generality -> specificity

Deep Residual Networks (ResNets)



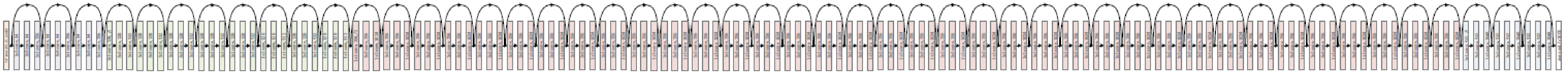
- «Deep Residual Learning for Image Recognition». CVPR 2016
  by Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun

- A Simple and clean framework of training very deep nets

- State-of-the-art performance for
  - Image classification
  - Object detection
  - Semantic segmentation
  - and more…

Slide adapted from ICML 2016 Tutorial by Kaiming He

Deep Residual Networks (ResNets)



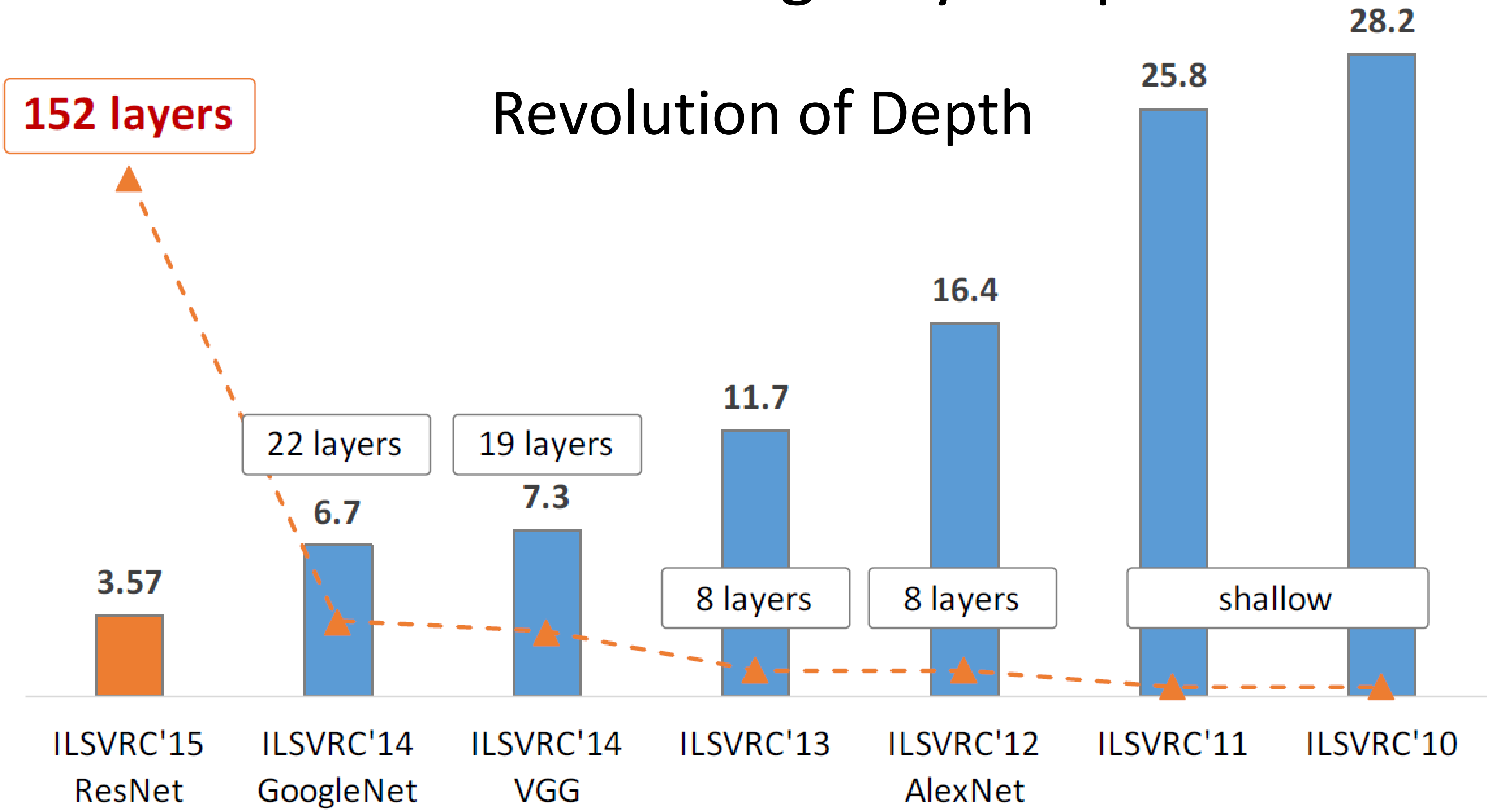
- «Deep Residual Learning for Image Recognition». CVPR 2016
  by Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun

- A Simple and clean framework of training very deep nets

- State-of-the-art performance for
  - Image classification
  - Object detection
  - Semantic segmentation
  - and more…

Slide adapted from ICML 2016 Tutorial by Kaiming He

ResNets for «training very deep nets»



Slide from ICML 2016 Tutorial by Kaiming He

ImageNet Classification top-5 error (%)

ResNets for achieving «state-of-the-art performance»

ResNets @ ILSVRC & COCO 2015 Competitions

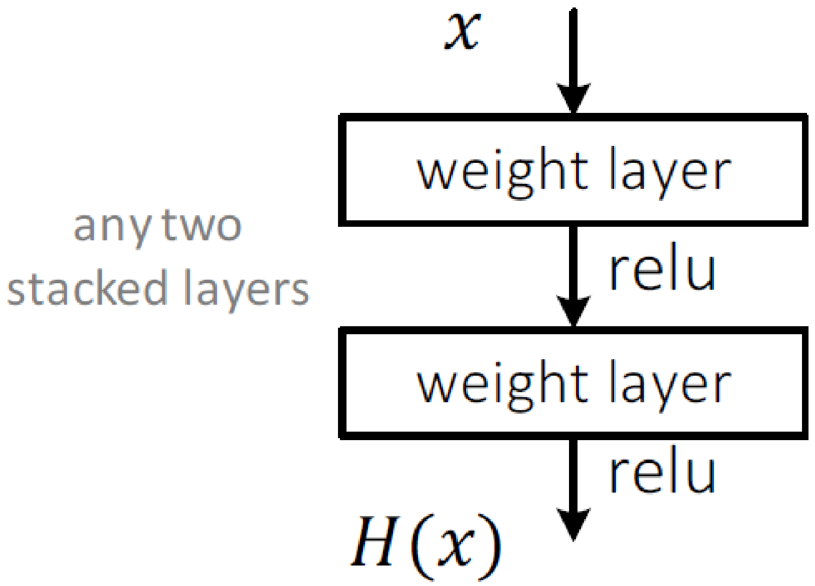
- **1st places in all five main tracks**
    - ImageNet Classification: *"Ultra-deep"* 152-layer nets
    - ImageNet Detection: 16% better than 2nd
    - ImageNet Localization: 27% better than 2nd
    - COCO Detection: 11% better than 2nd
    - COCO Segmentation: 12% better than 2nd

Slide from ICML 2016 Tutorial by Kaiming He

*improvements are relative numbers

Deep Residual Learning

Plain Net



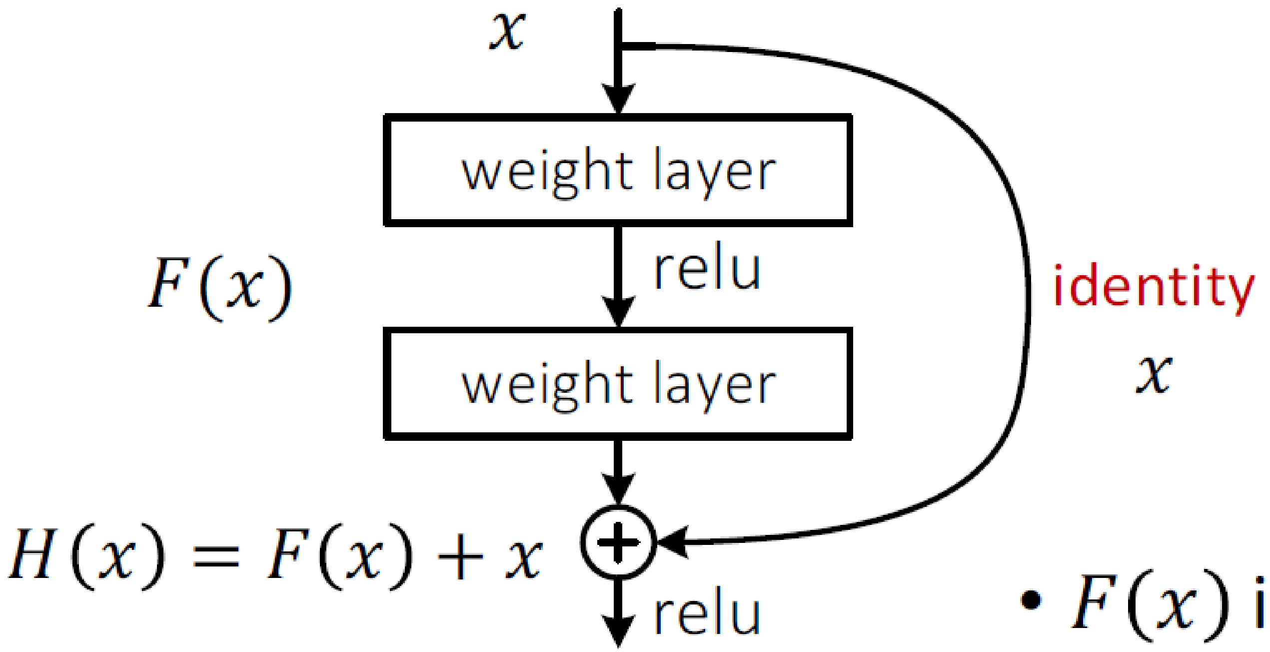
$H(x)$ is any desired mapping,

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

Slide from ICML 2016 Tutorial by Kaiming He

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 $H(x) = F(x) + x$

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

Slide from ICML 2016 Tutorial by Kaiming He

An issue on learning deep models

- **Optimization** ability

…(other issues)

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

How do ResNets address this issue?

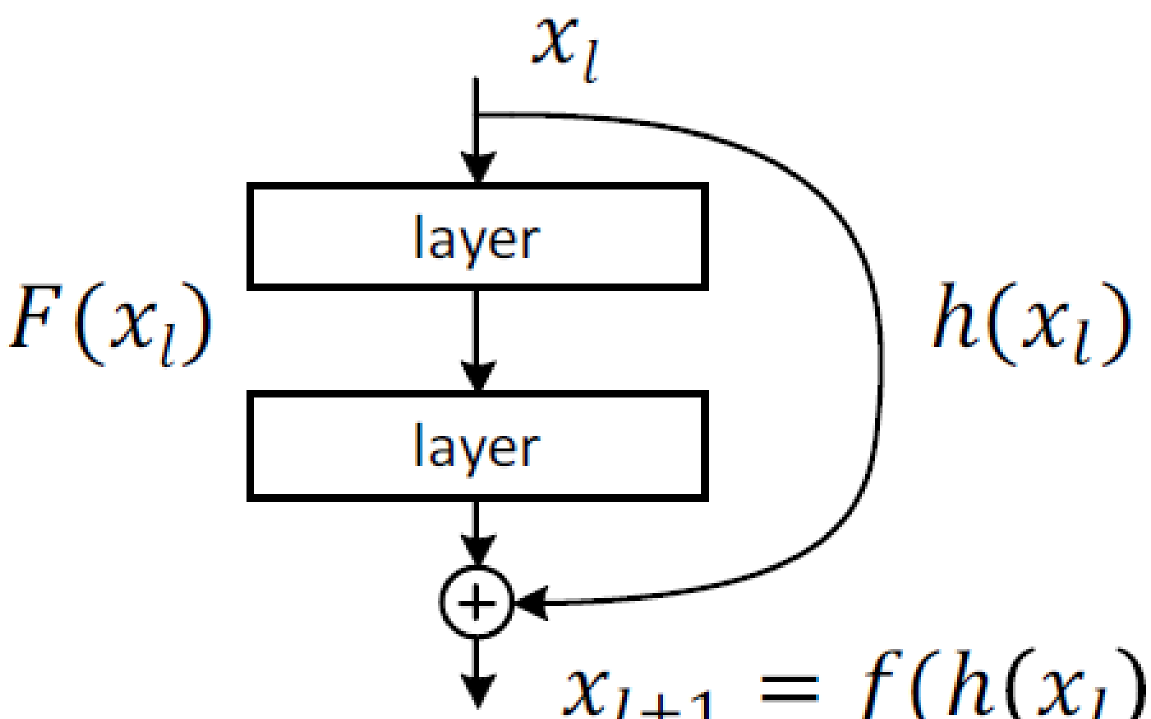
- **Optimization** ability

…

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

Slide adapted from ICML 2016 Tutorial by Kaiming He

On identity mappings for **optimization**

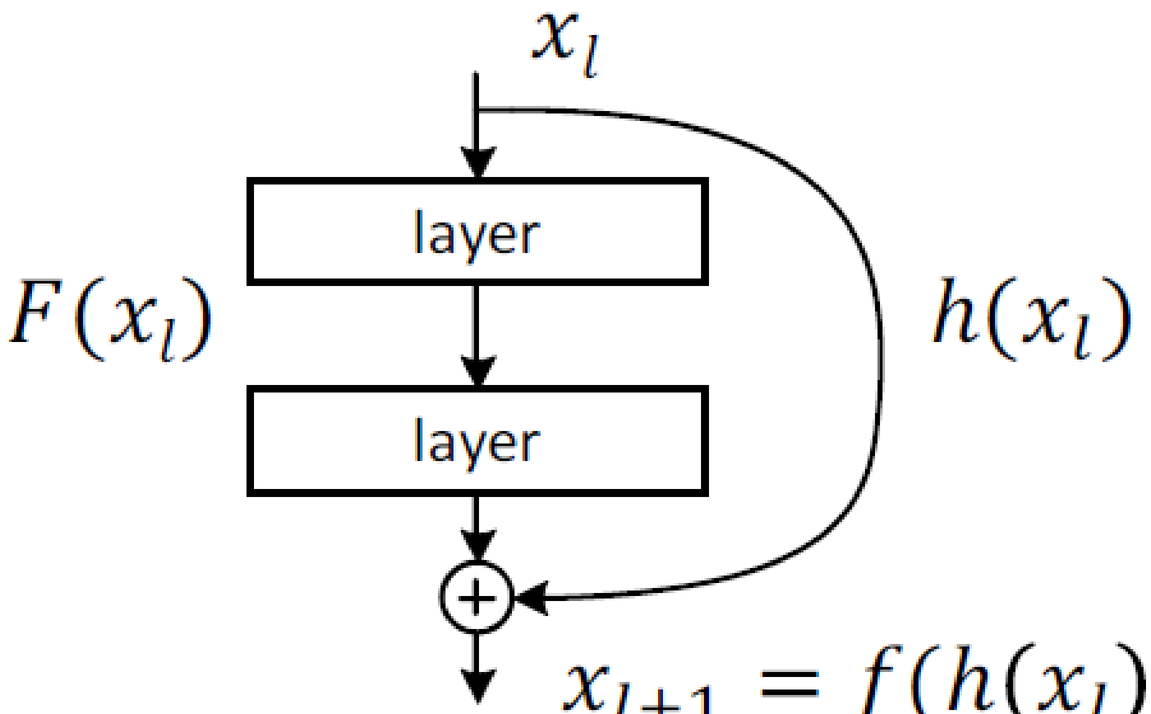


$$x_{l+1} = f(h(x_l) + F(x_l))$$

- shortcut mapping: h = identity
- after-add mapping: f = ReLU

Slide from ICML 2016 Tutorial by Kaiming He

On identity mappings for **optimization**

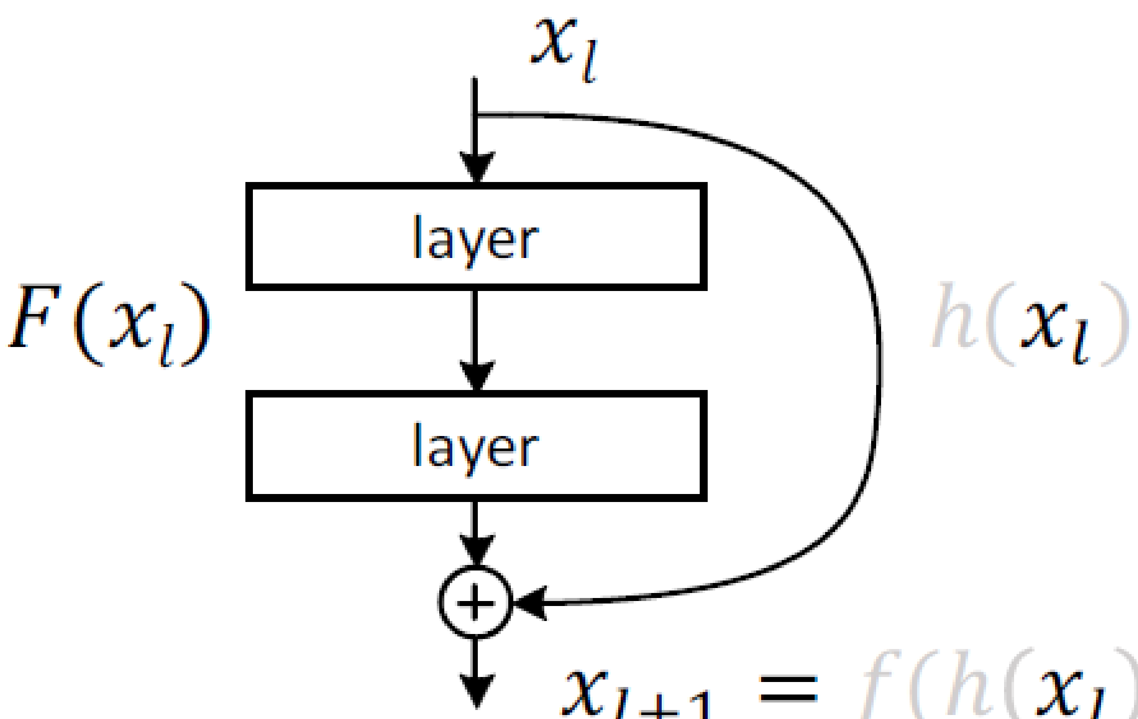


$$x_{l+1} = f(h(x_l) + F(x_l))$$

- shortcut mapping: h = identity
- ~~after-add mapping: f = ReLU~~
- What if f = identity?

Slide from ICML 2016 Tutorial by Kaiming He

On identity mappings for **optimization**



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

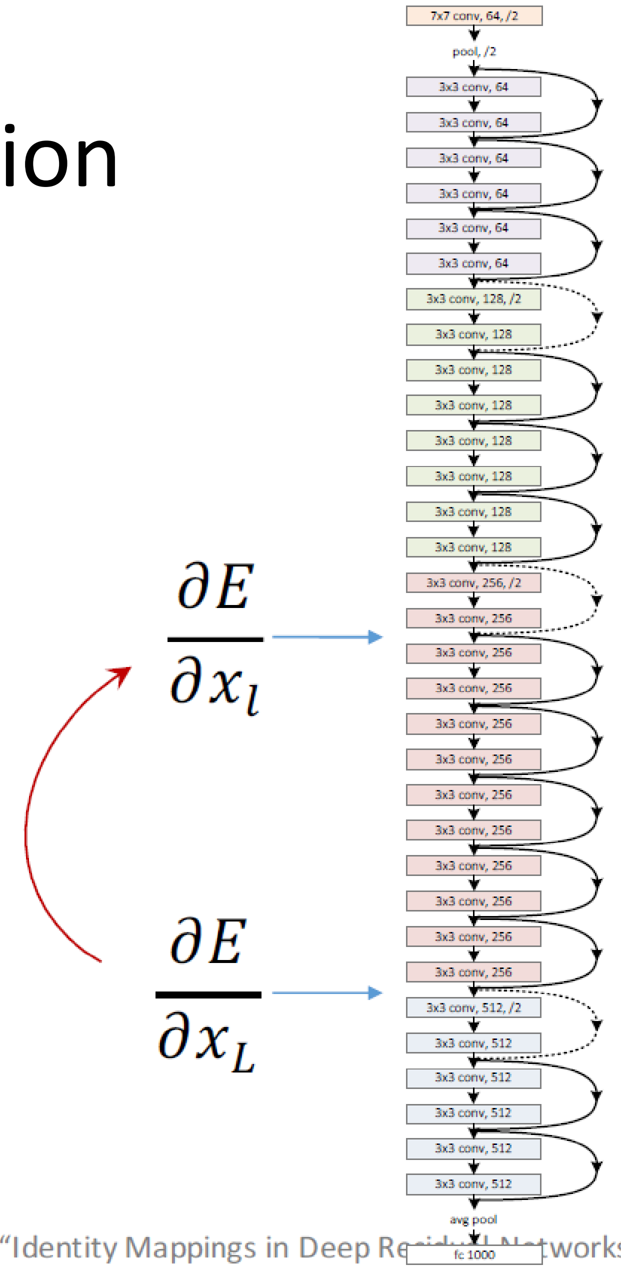
- shortcut mapping: h = identity
- ~~after-add mapping: f = ReLU~~
- What if f = identity?

Slide from ICML 2016 Tutorial by Kaiming He

Very smooth backward propagation

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

- 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: $\frac{\partial E}{\partial x_l} = \prod_{i=l}^{L-1} W_i \frac{\partial E}{\partial x_L}$



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

Slide from ICML 2016 Tutorial by Kaiming He

Key takeaways

Key takeaways

Residual networks contain many paths.

Previous networks have a single path.

Only short paths contribute gradient during training.

Vanishing gradient suppresses gradient from long paths.

Slide from NIPS 2016 Spotlight Video

Key takeaways

Residual networks contain many paths.

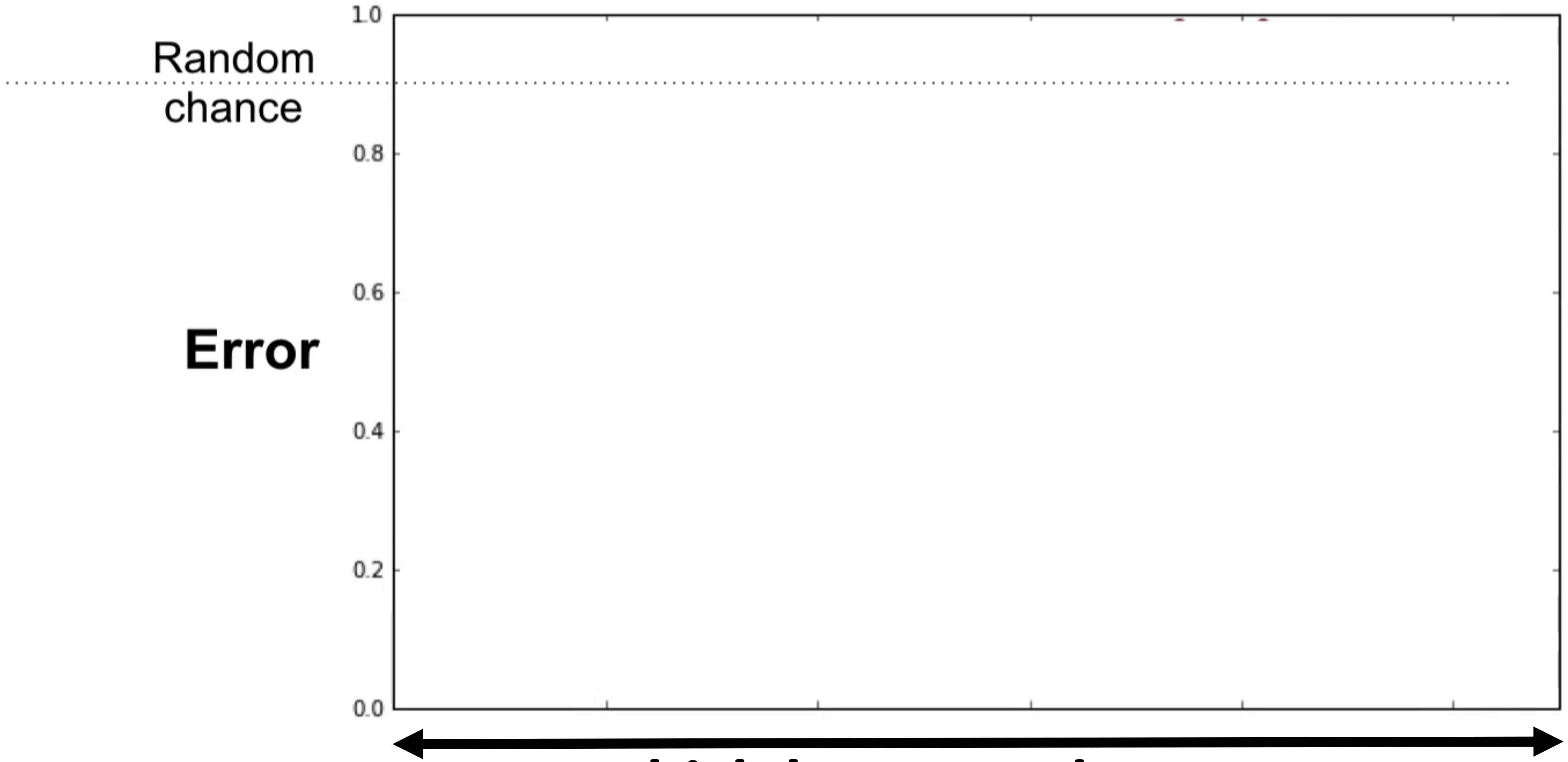
Previous networks have a single path.

Only short paths contribute gradient during training.

Vanishing gradient suppresses gradient from long paths.

Slide from NIPS 2016 Spotlight Video

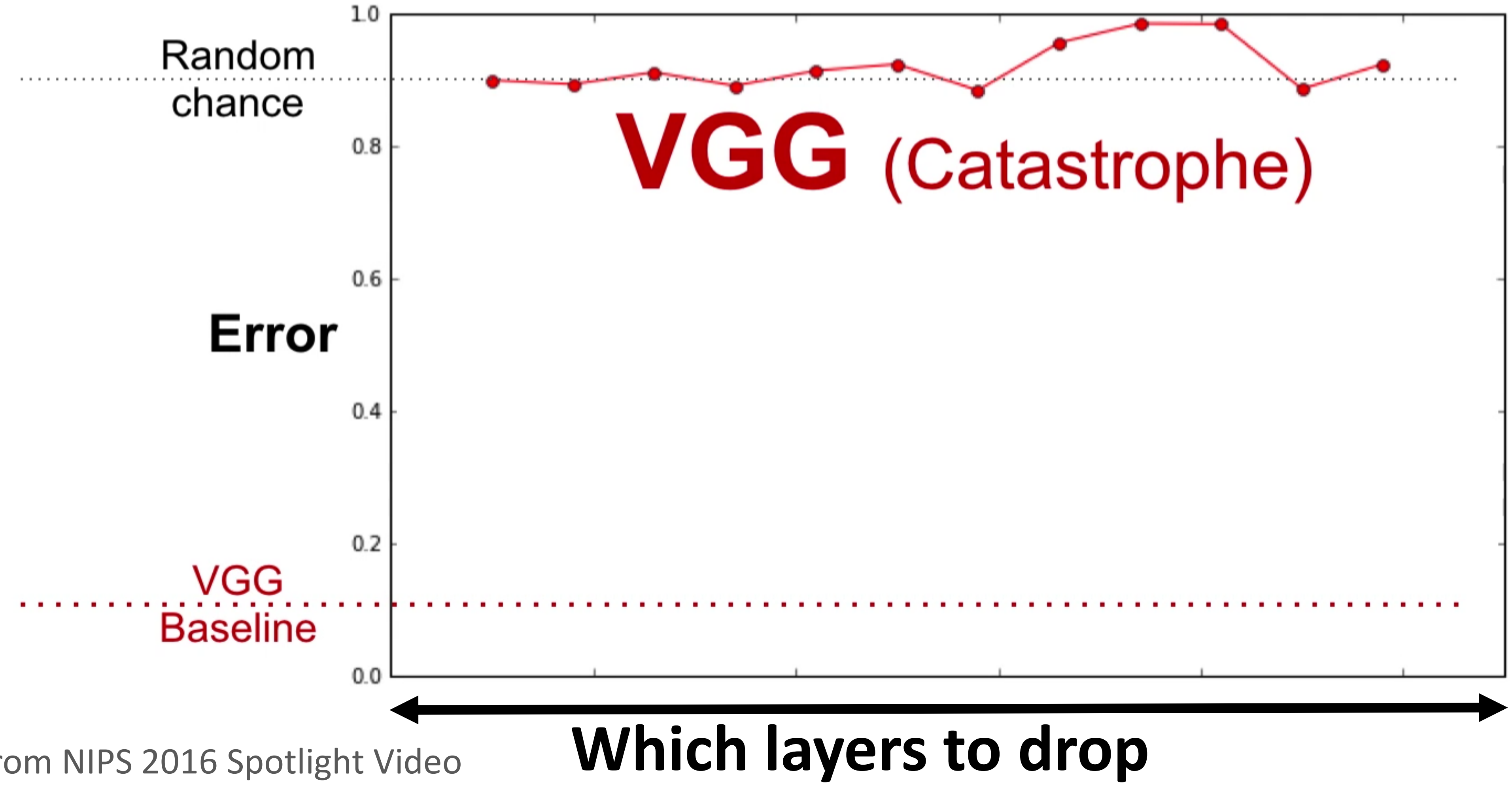
For example, **what happens when we delete layers at test time?**



Which layers to drop

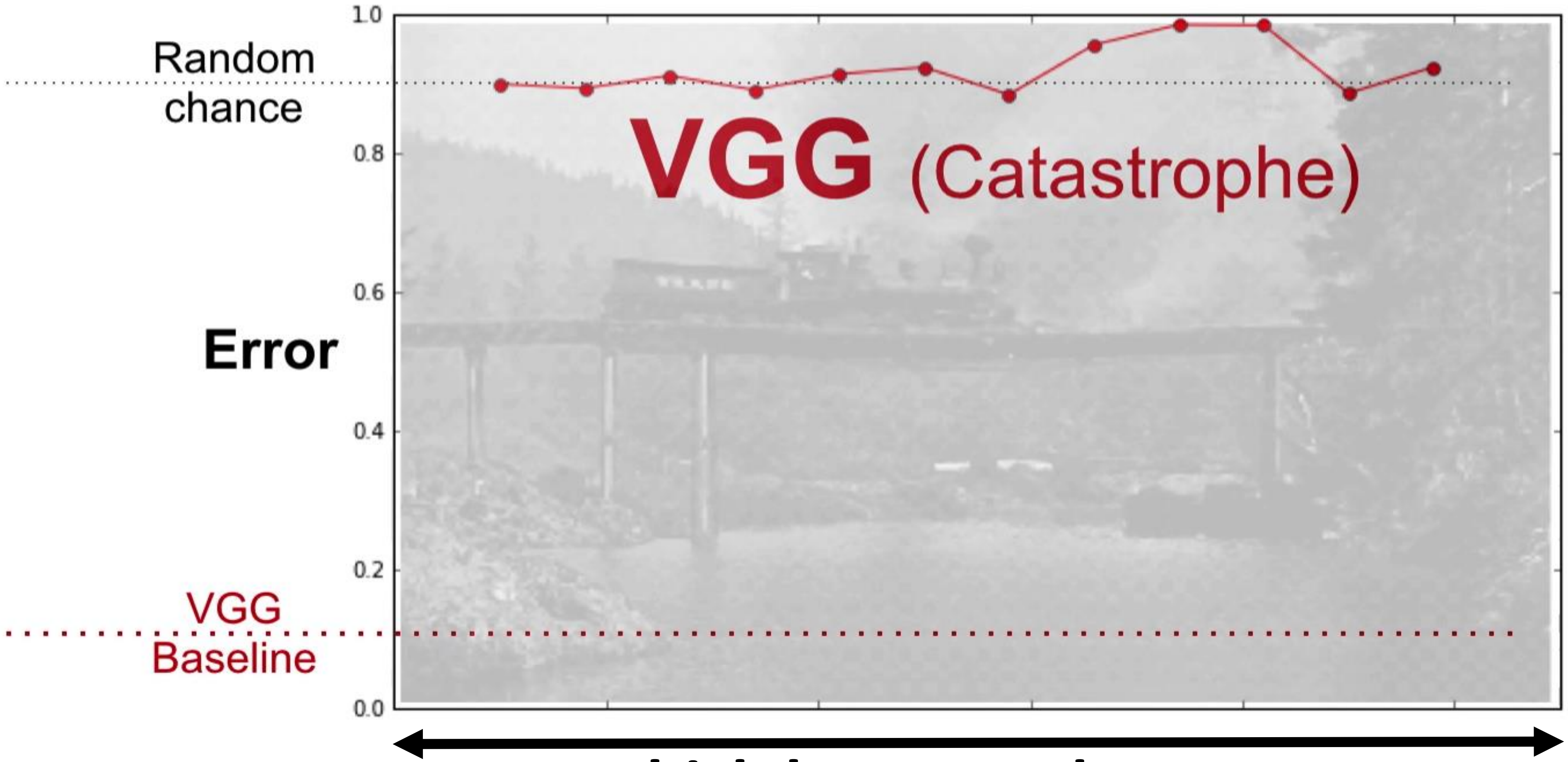
Slide from NIPS 2016 Spotlight Video

For example, **what happens when we delete layers at test time?**



Slide from NIPS 2016 Spotlight Video

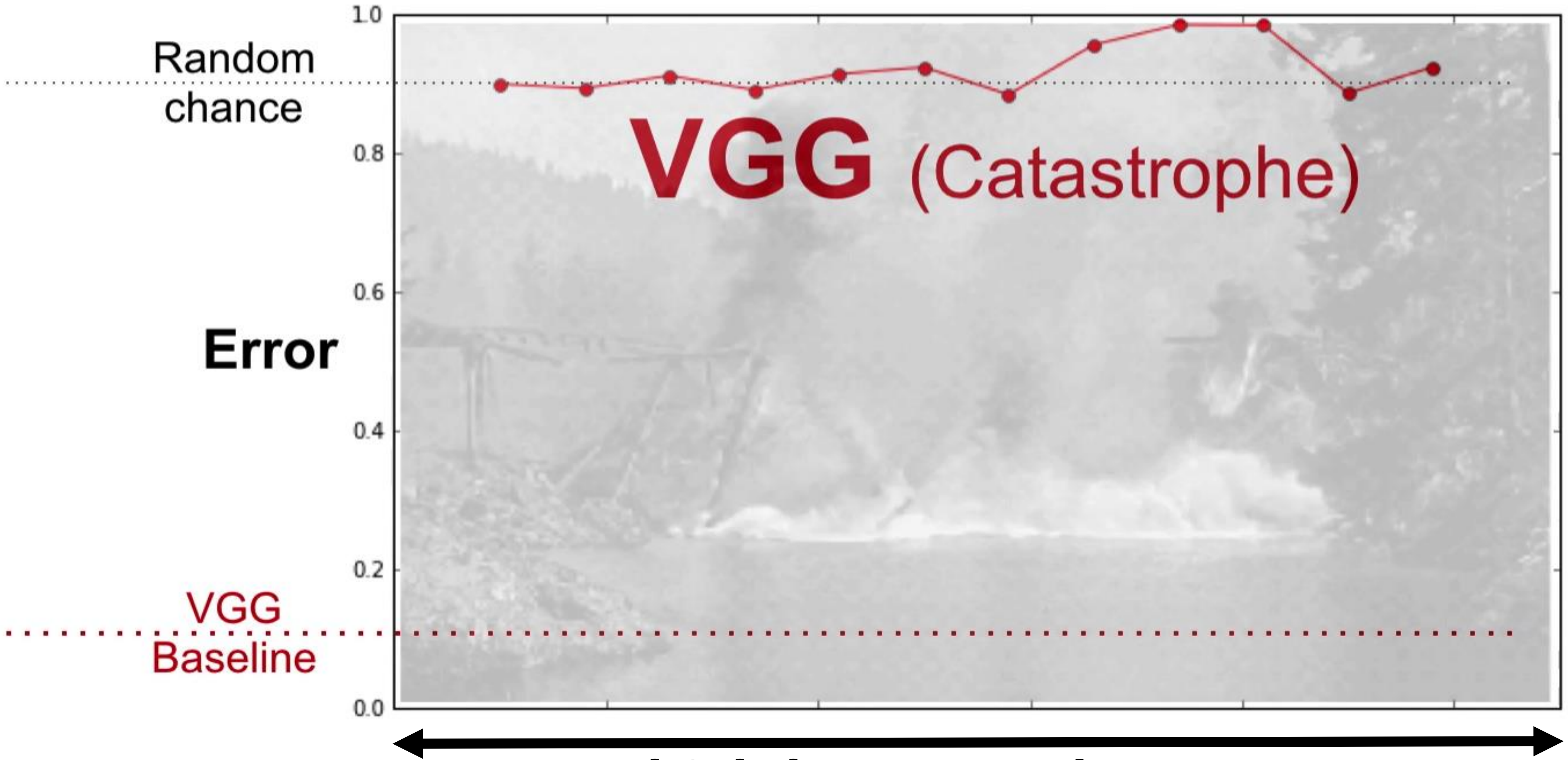
For example, **what happens when we delete layers at test time?**



Which layers to drop

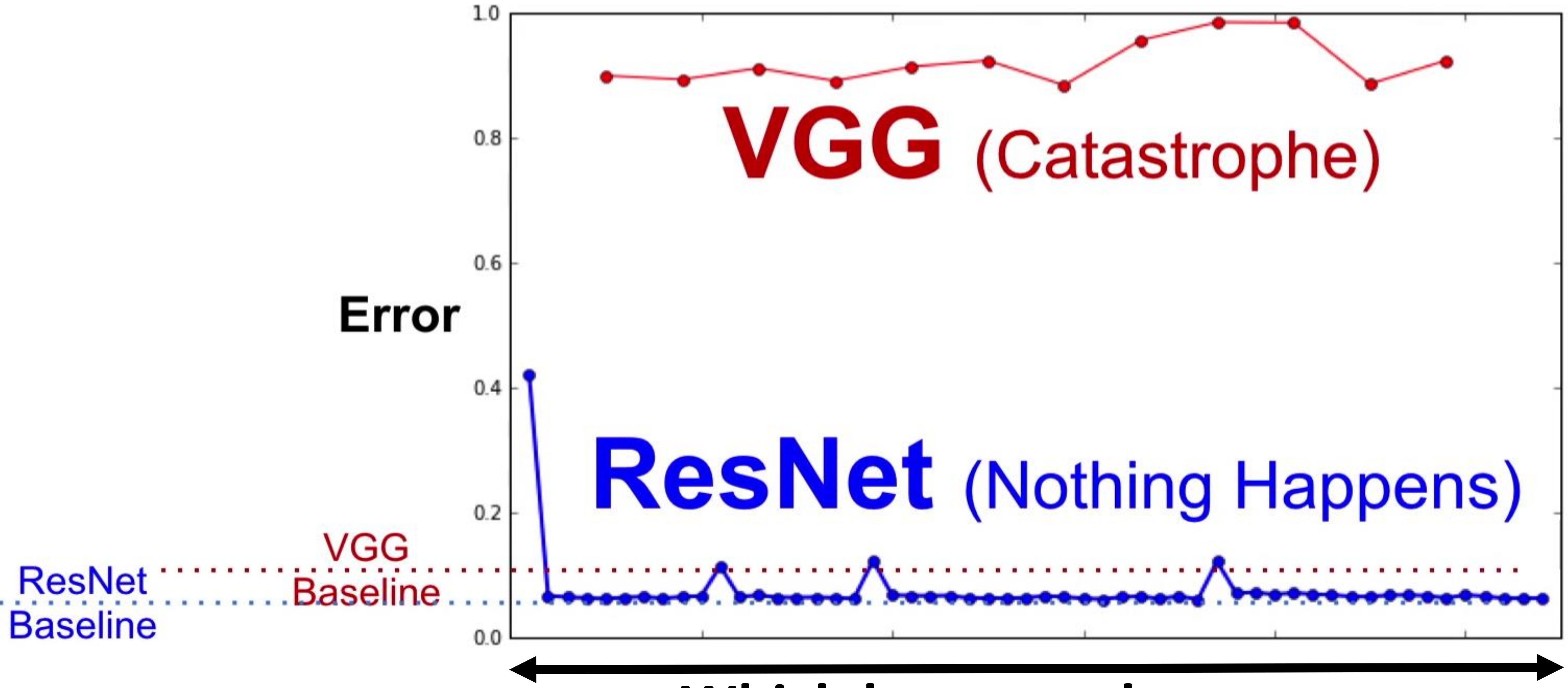
Slide from NIPS 2016 Spotlight Video

For example, **what happens when we delete layers at test time?**



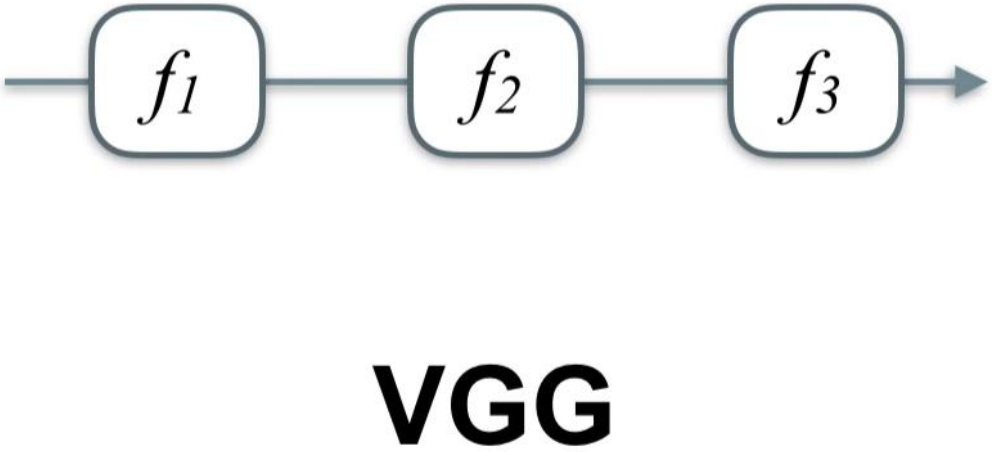
Slide from NIPS 2016 Spotlight Video

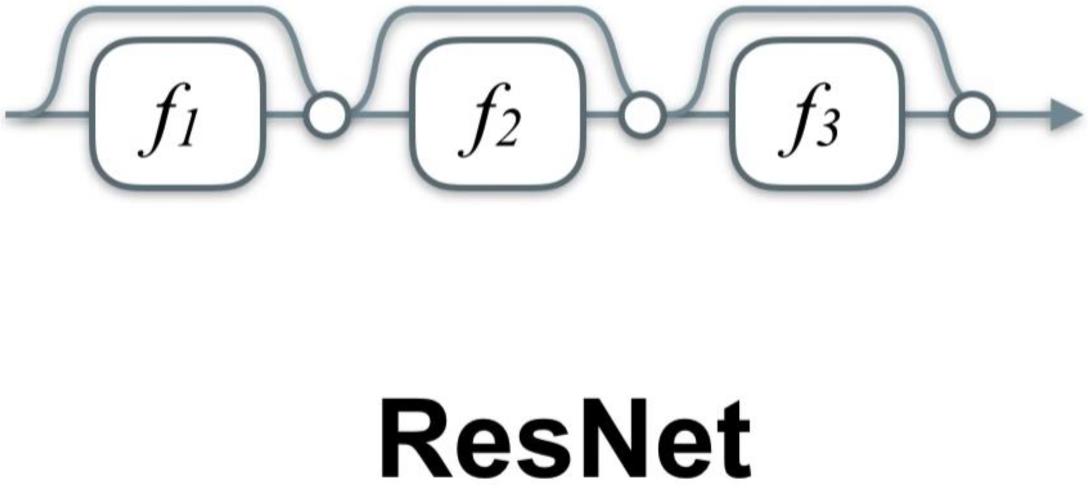
For example, **what happens when we delete layers at test time?**



Slide from NIPS 2016 Spotlight Video

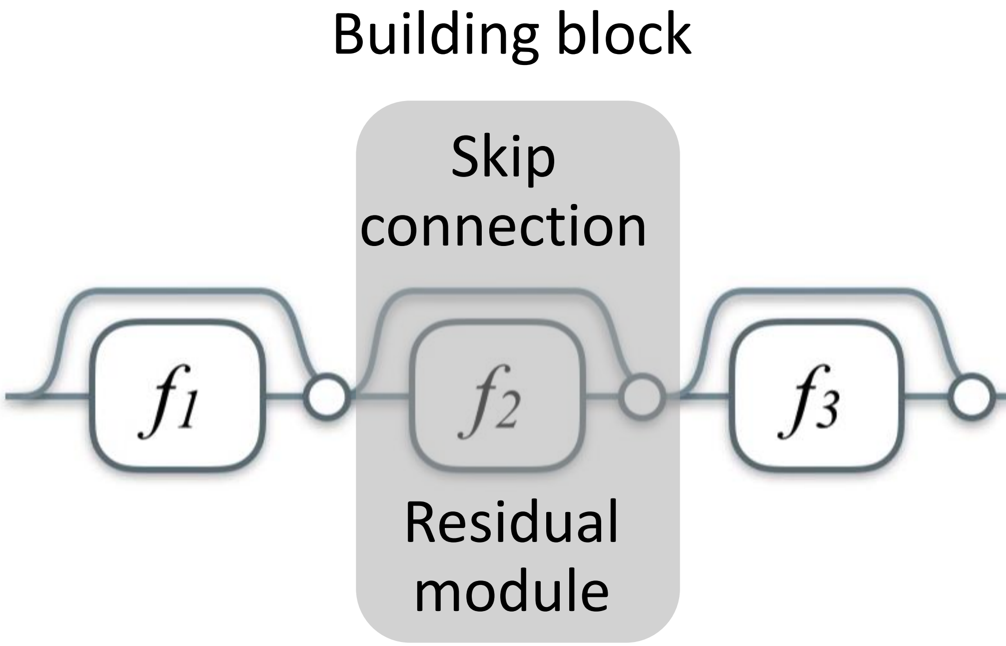
Why does this happen? **The «unraveled view»**



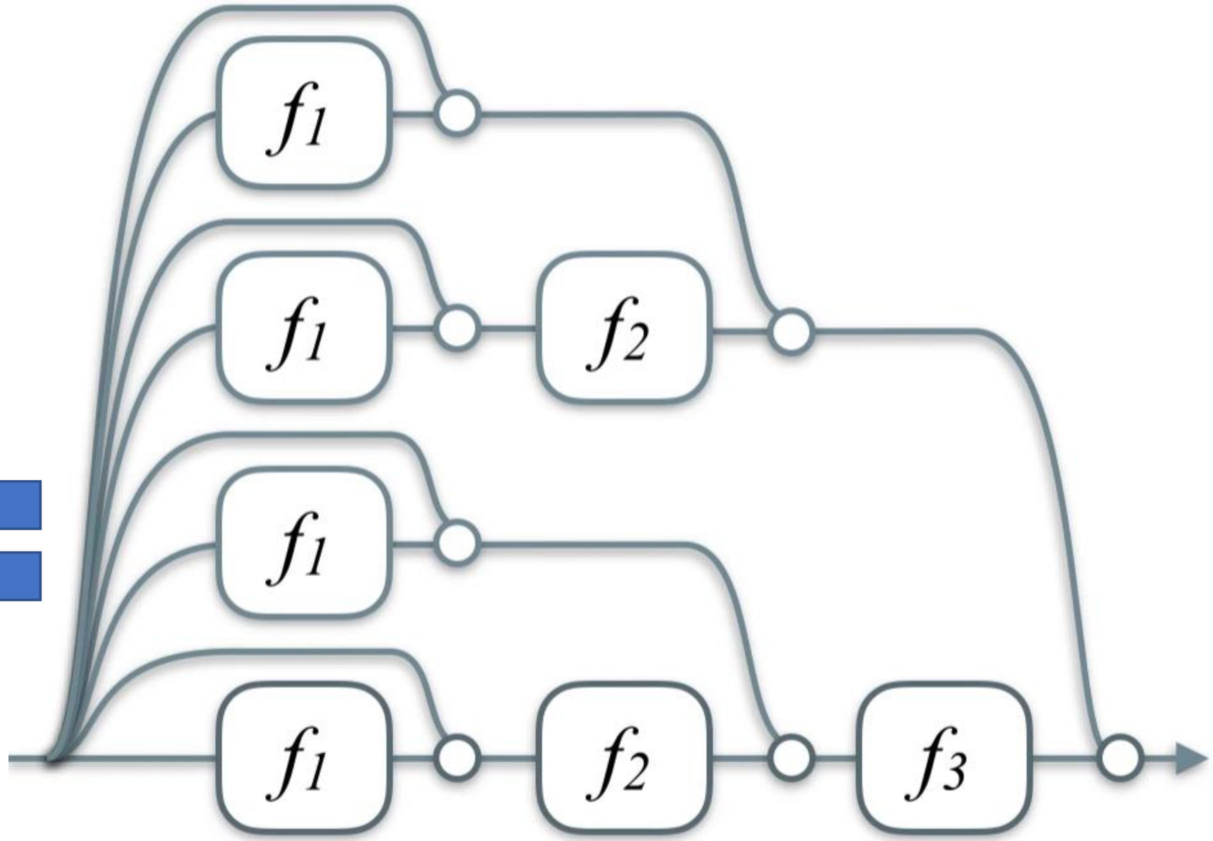


Slide from NIPS 2016 Spotlight Video

Why does this happen? **The «unraveled view»**



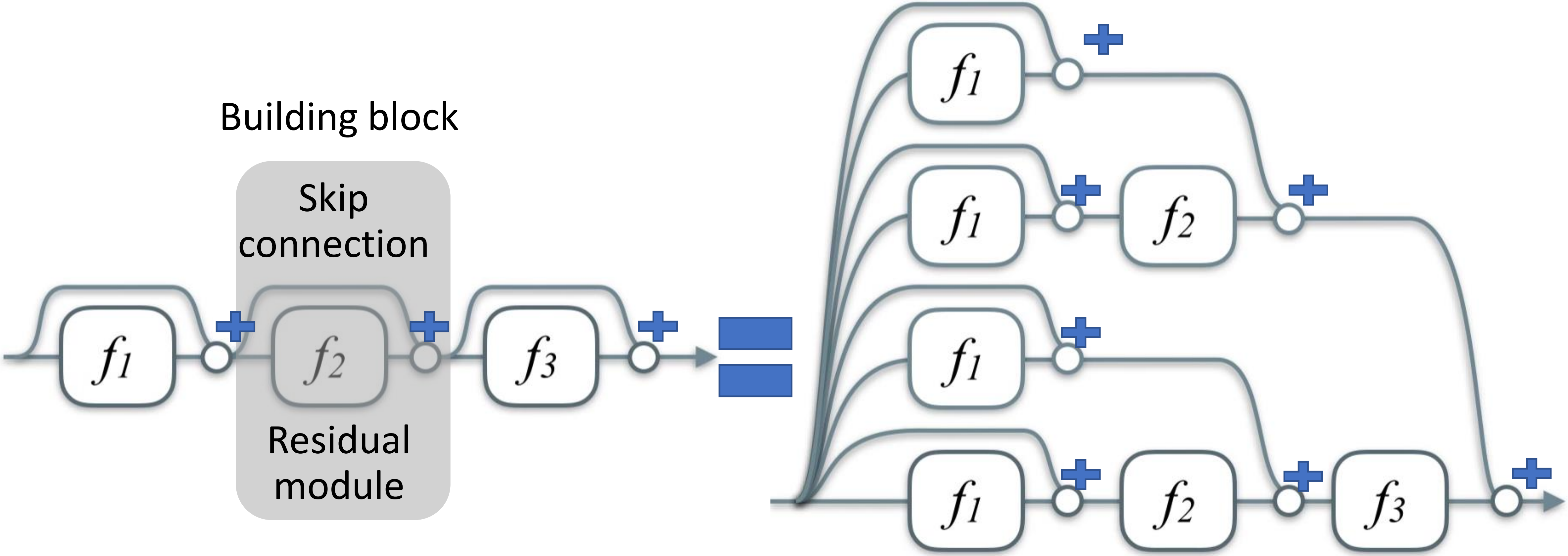
(a) Conventional 3-block residual network



Unraveled view of (a)

Slide adapted from NIPS 2016 Spotlight Video

Why does this happen? **The «unraveled view»**



(a) Conventional 3-block residual network

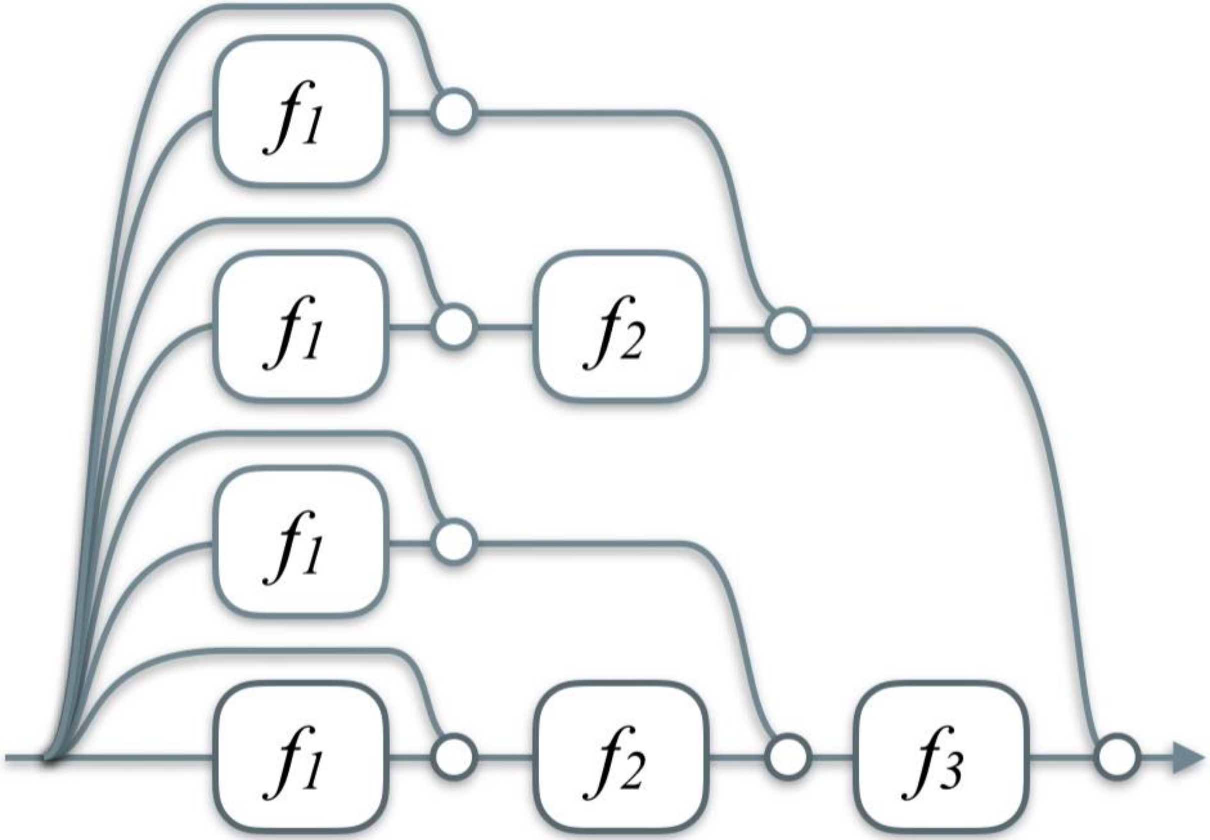
Unraveled view of (a)

Slide adapted from NIPS 2016 Spotlight Video

Why does this happen? **The «unraveled view»**

The unraveled view is equivalent and showcases the many paths in ResNet.

VGG



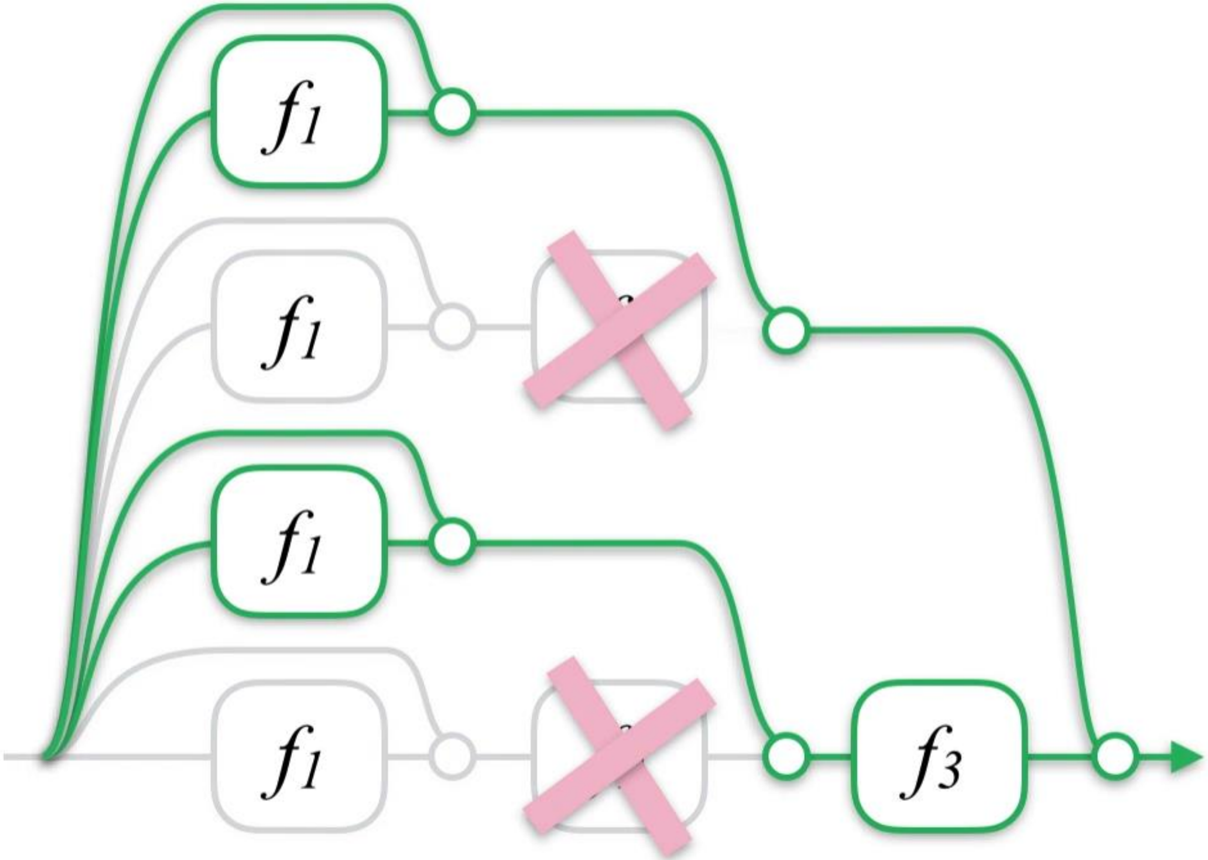
ResNet

Slide from NIPS 2016 Spotlight Video

Deletion of one layer

VGG

All paths are affected

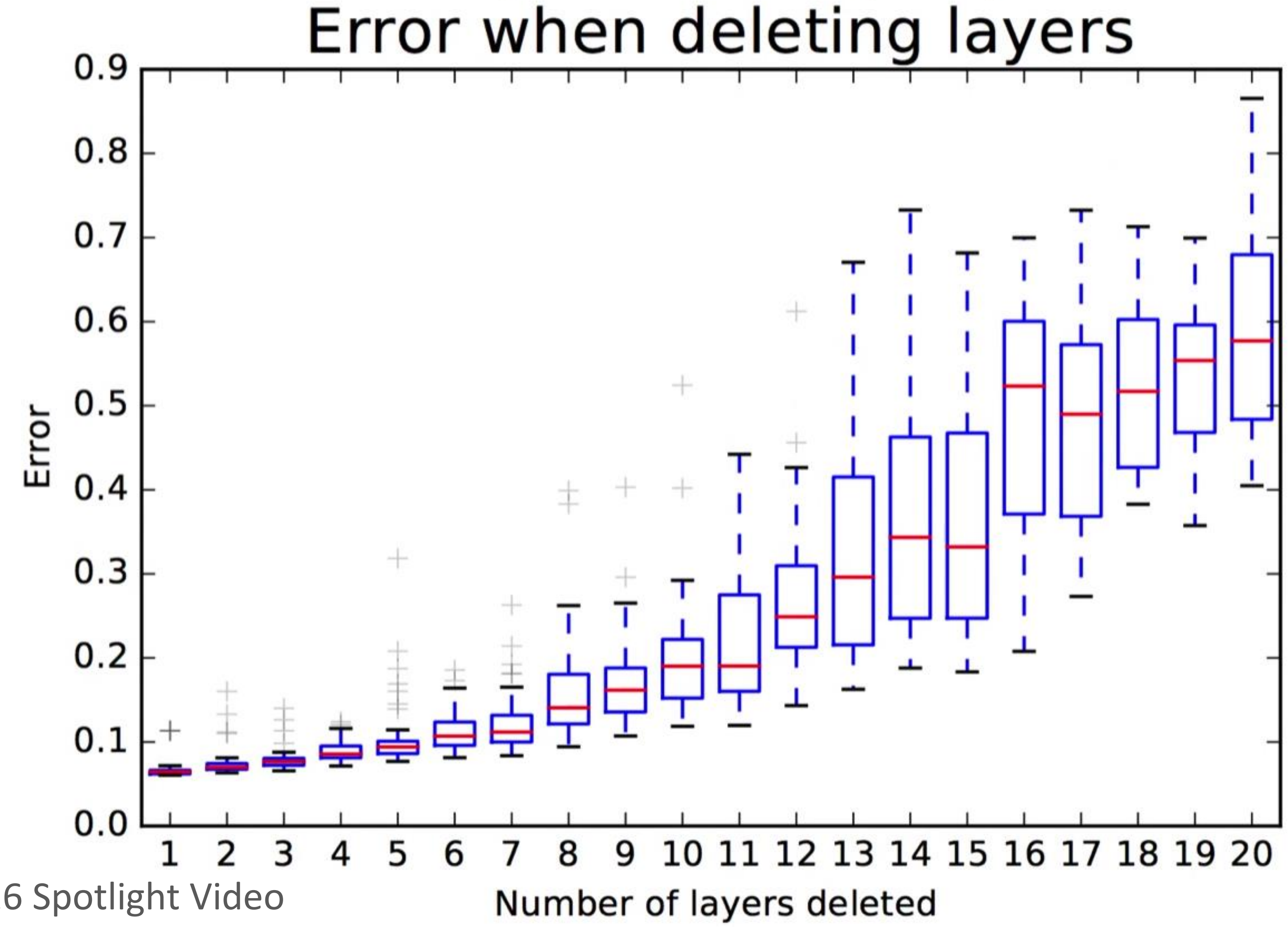


ResNet

Only **half** of the paths are affected

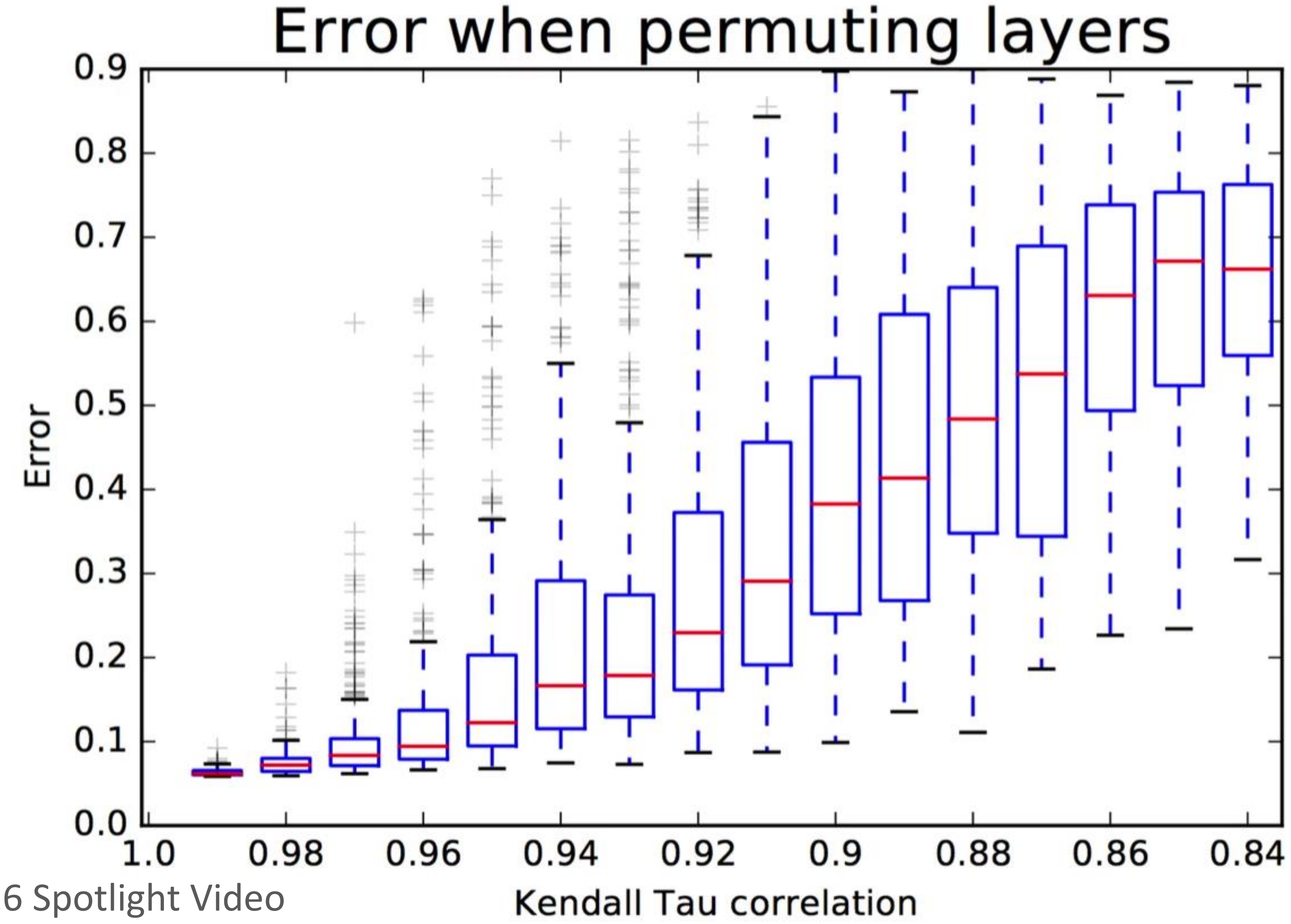
Slide from NIPS 2016 Spotlight Video

Performance varies smoothly when deleting **several** layers.



Slide from NIPS 2016 Spotlight Video

Performance varies smoothly when **re-ordering** layers.



Slide from NIPS 2016 Spotlight Video

Conclusion 1:

- Residual Networks consist of many paths.
- Although trained jointly, they do not strongly depend on each other: Ensemble-like behavior

Slide adapted from NIPS 2016 Spotlight Video

Key takeaways

Residual networks contain many paths.

Previous networks have a single path.

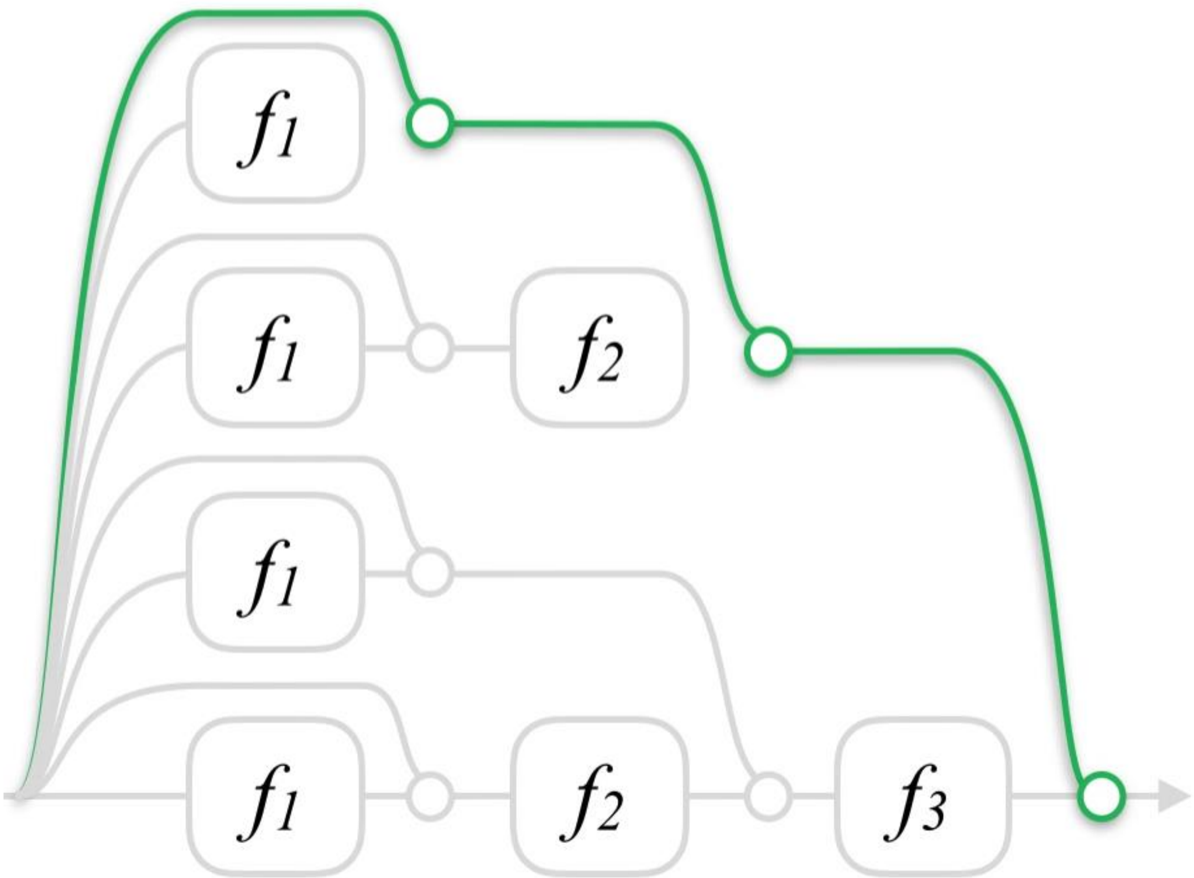
Only short paths contribute gradient during training.

Vanishing gradient suppresses gradient from long paths.

Slide from NIPS 2016 Spotlight Video

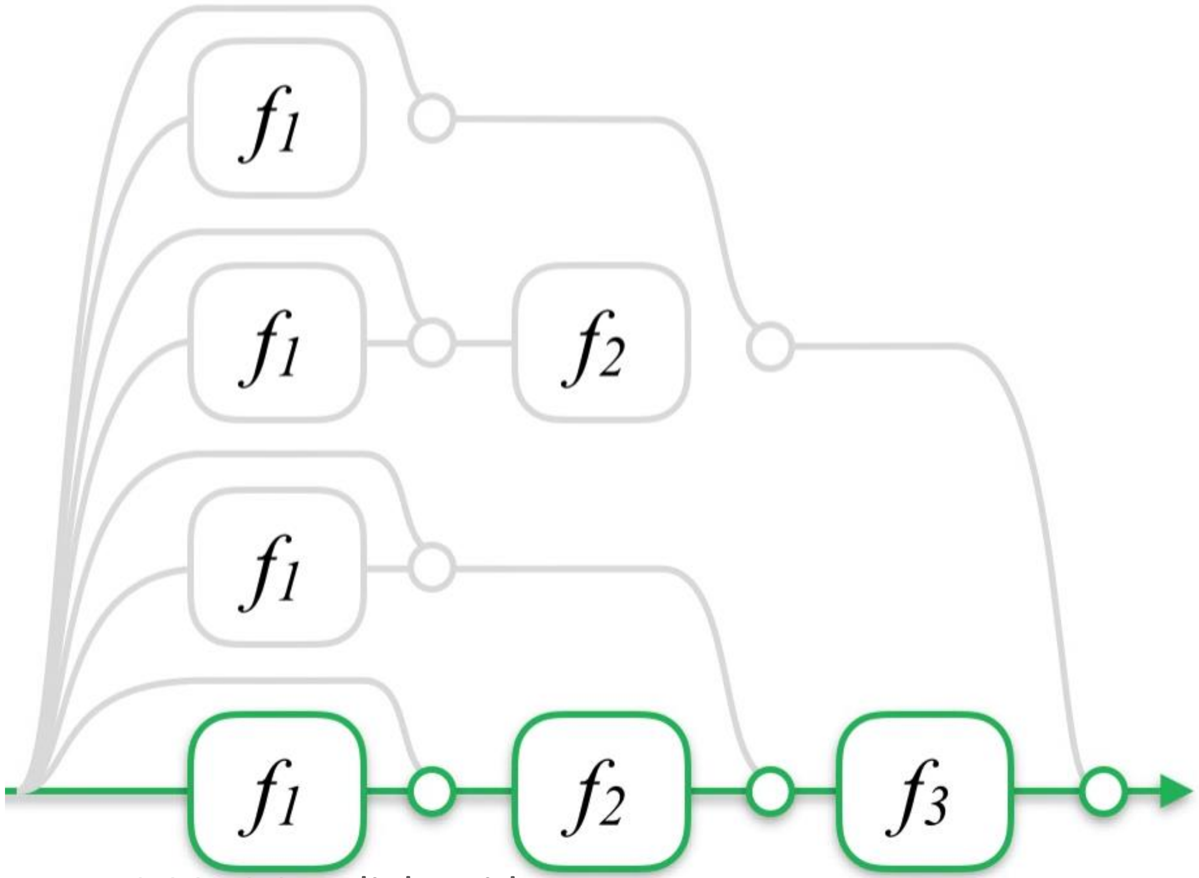
Distribution of path length

There are very few
short paths…



Slide from NIPS 2016 Spotlight Video

Distribution of path length

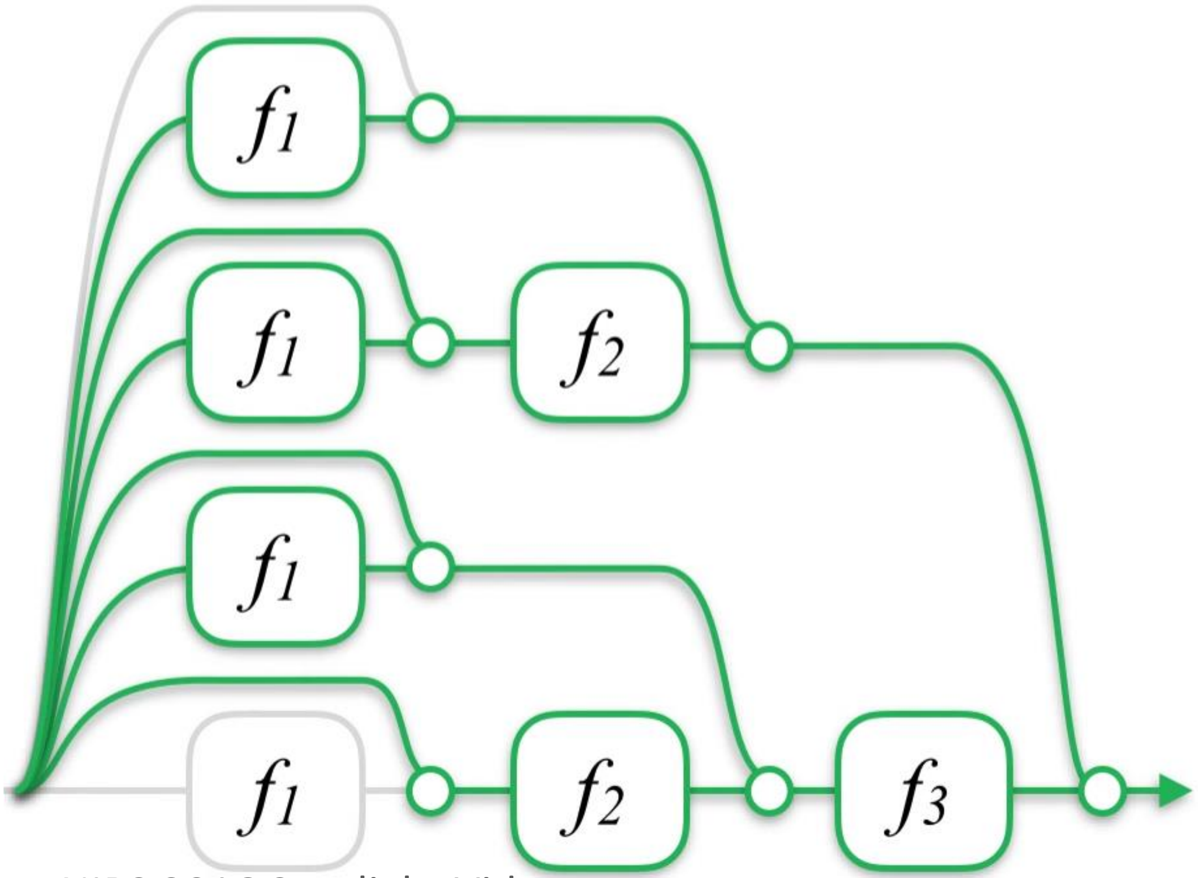


There are very few **short paths…**

And very few **long paths…**

Slide from NIPS 2016 Spotlight Video

Distribution of path length



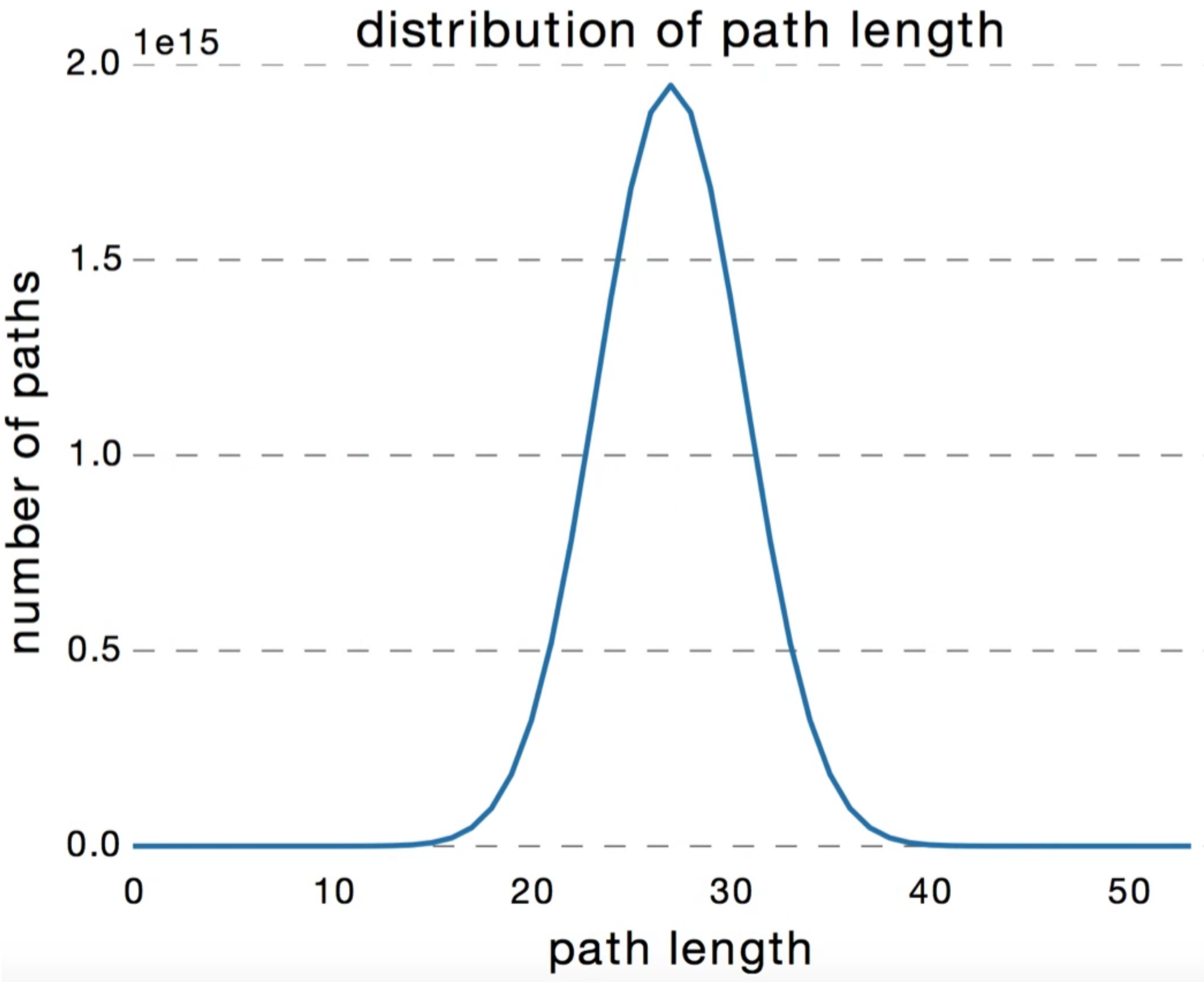
Slide from NIPS 2016 Spotlight Video

There are very few **short paths…**

And very few **long paths…**

Most paths are **medium length!**

Distribution of path length



There are very few **short paths…**

And very few **long paths…**

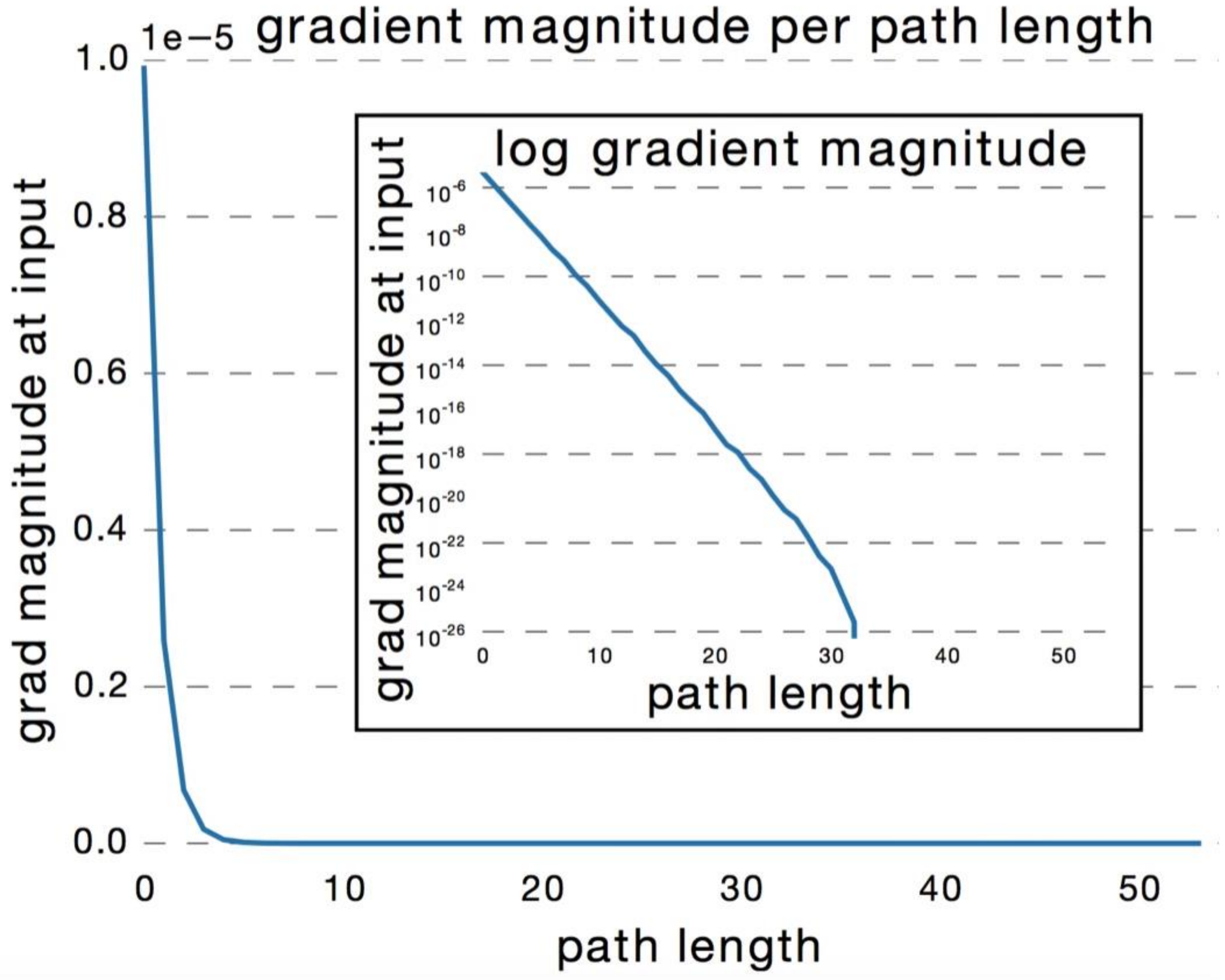
Most paths are **medium length!**

Paths length follows a **binomial distribution.**

Slide from NIPS 2016 Spotlight Video

Vanishing gradient

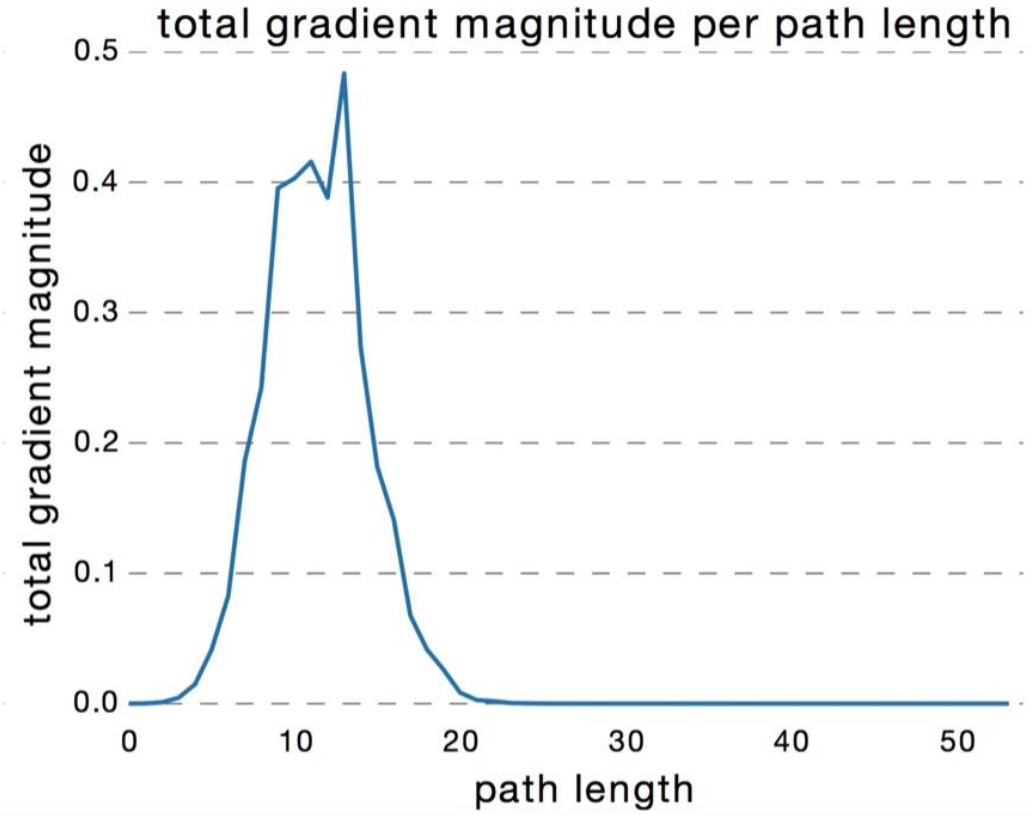
The gradient magnitude **decreases exponentially** with increasing path length.



Slide from NIPS 2016 Spotlight Video

Andreas Veit, Michael Wilber & Serge Belonie. NIPS 2016.

Gradient during training with respect to path lengths



Combining the path length distribution and the vanishing gradients, one can observe that most of the gradient comes from relatively short paths.

Slide from NIPS 2016 Spotlight Video

Conclusion 2:

- Most paths through a ResNet are relatively short.
- During training, gradients only flow through short paths.

Slide from NIPS 2016 Spotlight Video

Q & A

References

- Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult. " *IEEE Transactions on Neural Networks* 5(2)*,* 1994, pp 157–166.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): n. pag. Web.
- He, Kaiming, et al. "Identity Mappings in Deep Residual Networks." *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, 2016, pp. 630–645., doi:10.1007/978-3-319-46493-0_38.
- He, Kaiming. "ICML 2016 Tutorial on Deep Residual Networks." *ICML 2016 Tutorial on Deep Residual Networks*. N.p., n.d. Web. 14 Mar. 2017.
- Hochreiter, Sepp. "Untersuchungen zu dynamischen neuronalen netzen." *Master's thesis*, 1991, Institut für Informatik, Technische Universitat, München.
- Veit, Andreas, Michael J. Wilber, and Serge Belongie. "Residual Networks Behave Like Ensembles of Relatively Shallow Networks." *Advances in Neural Information Processing Systems* 29 (2016): n. pag. Web.
- Wilber, Michael. "Residual Networks Behave Like Ensembles of Relatively Shallow Networks." *YouTube*. YouTube, 21 Nov. 2016. Web. 14 Mar. 2017.
- Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks? " *Advances in Neural Information Processing Systems (*2014).