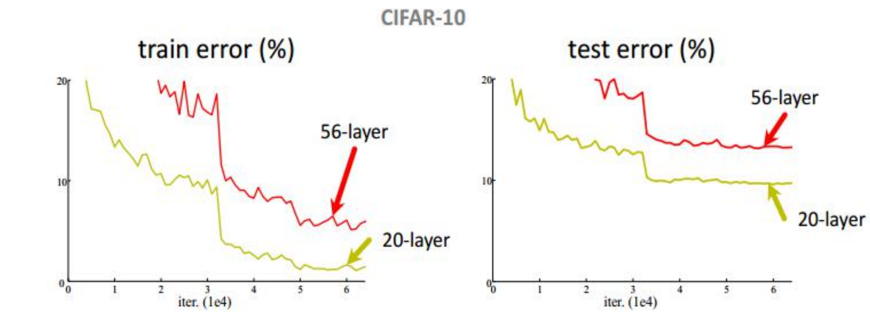
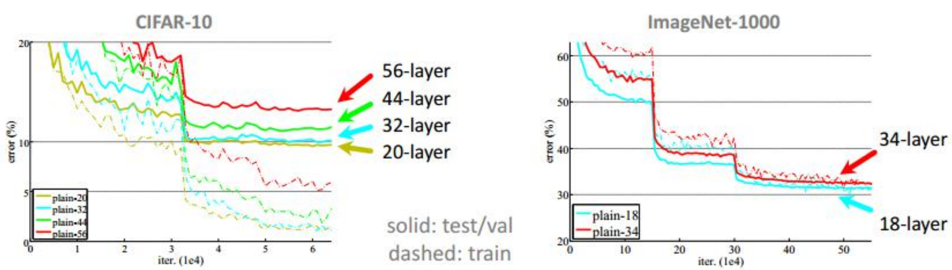
Deep Residual Networks

su
2018/1116

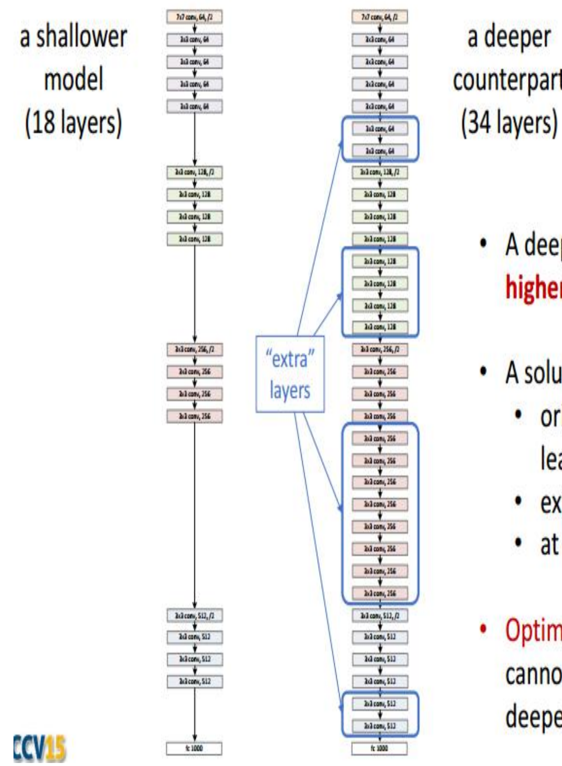
Simply stacking layers?



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



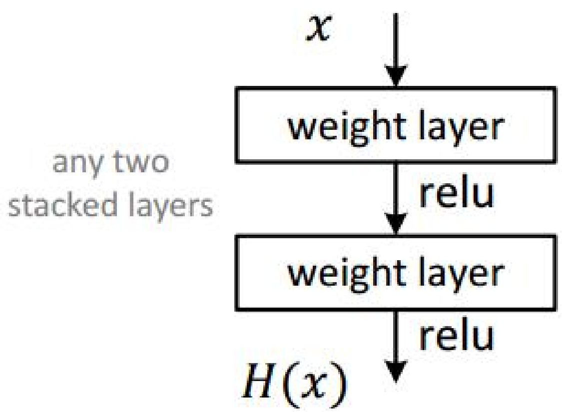
- “Overly deep” plain nets have **higher training error**
- A general phenomenon, observed in many datasets



- 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

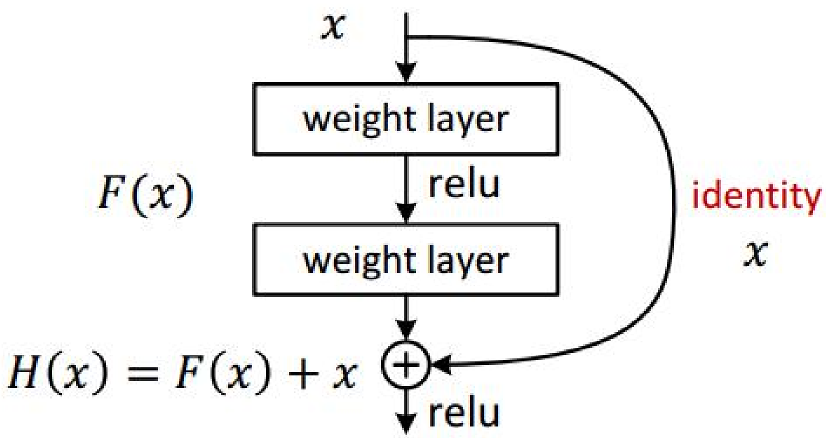
- Plaint net



$H(x)$ is any desired mapping,

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

- Residual net



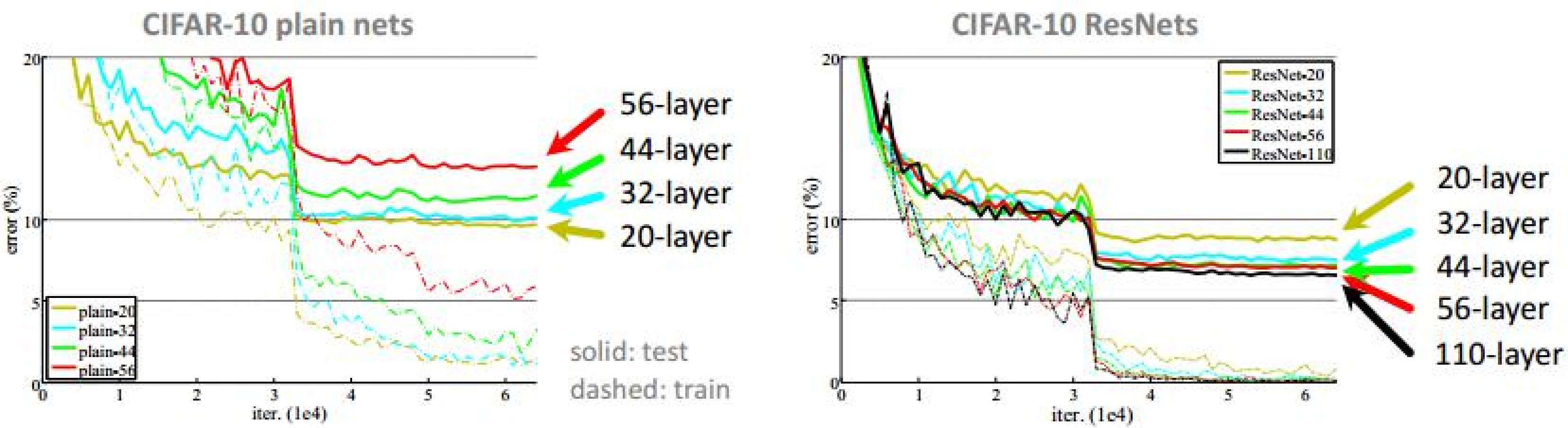
$H(x)$ is any desired mapping,

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

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

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

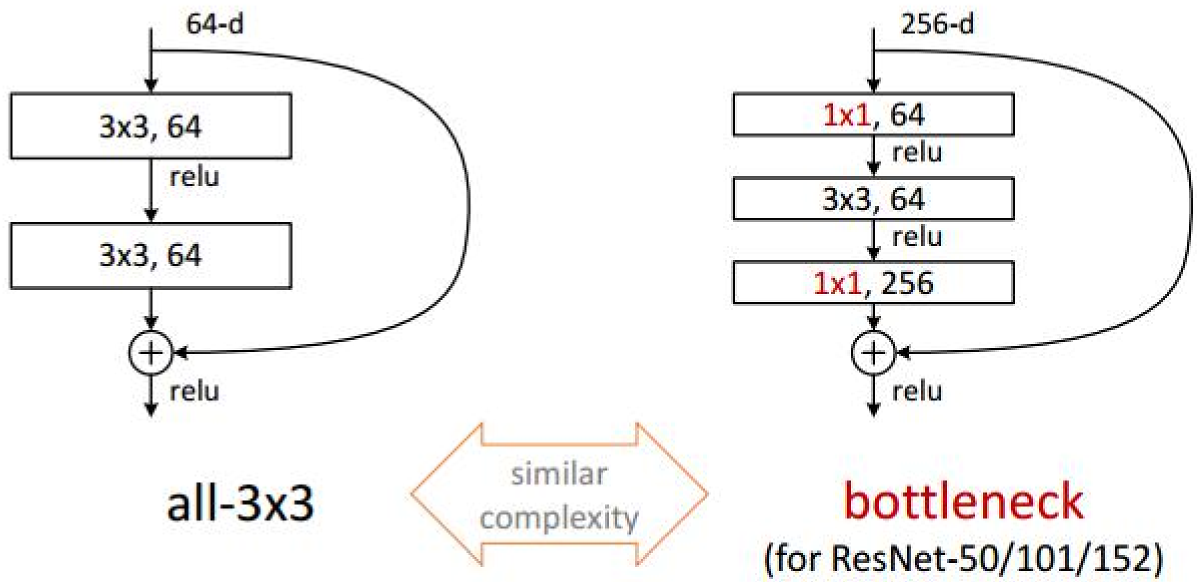
Result on CIFAR-10



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

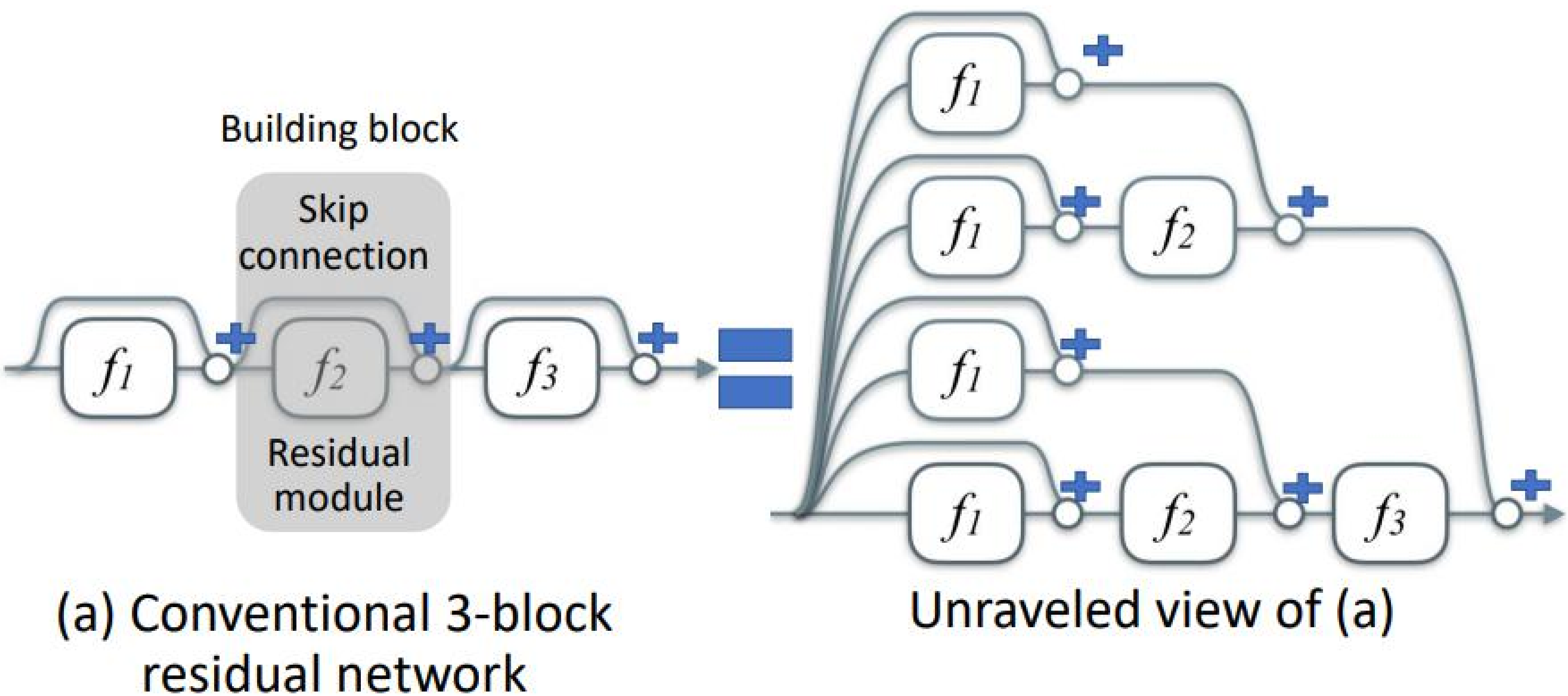
ImageNet experiments

- A practical design of going deeper



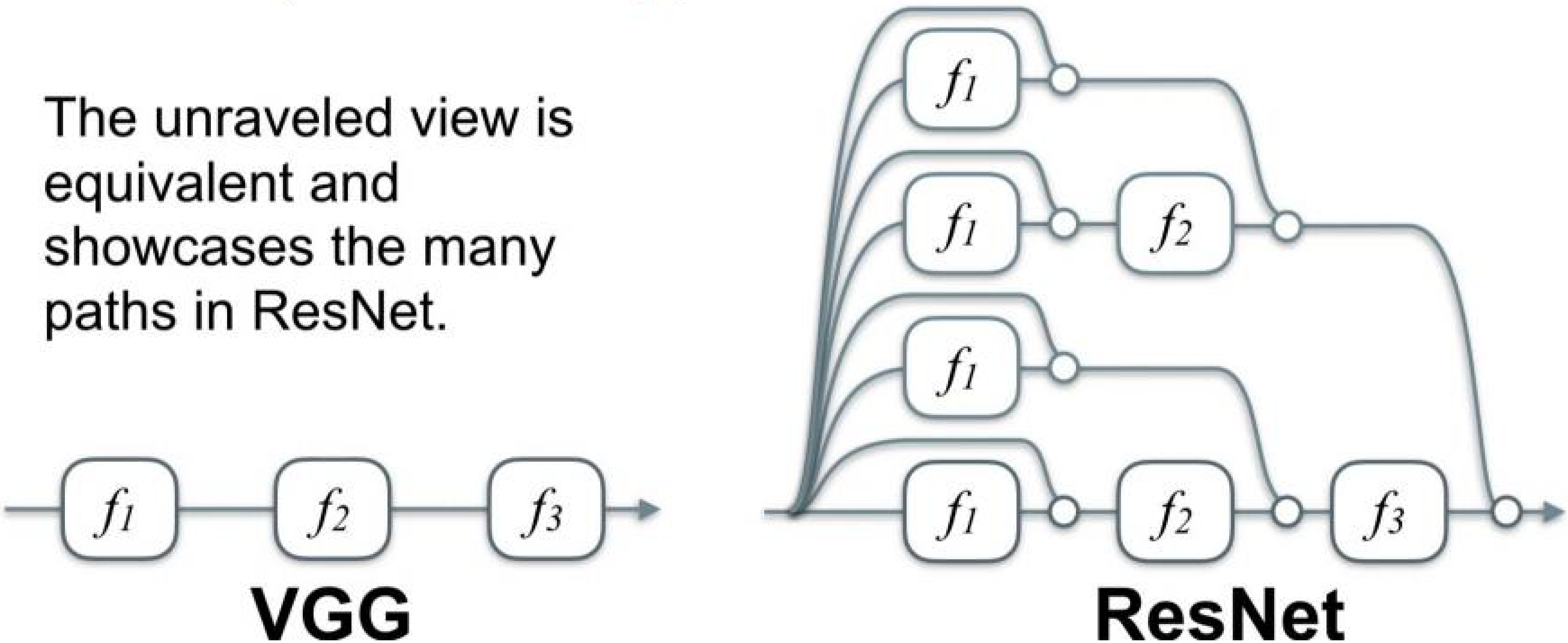
ICCV15

Why does this work?

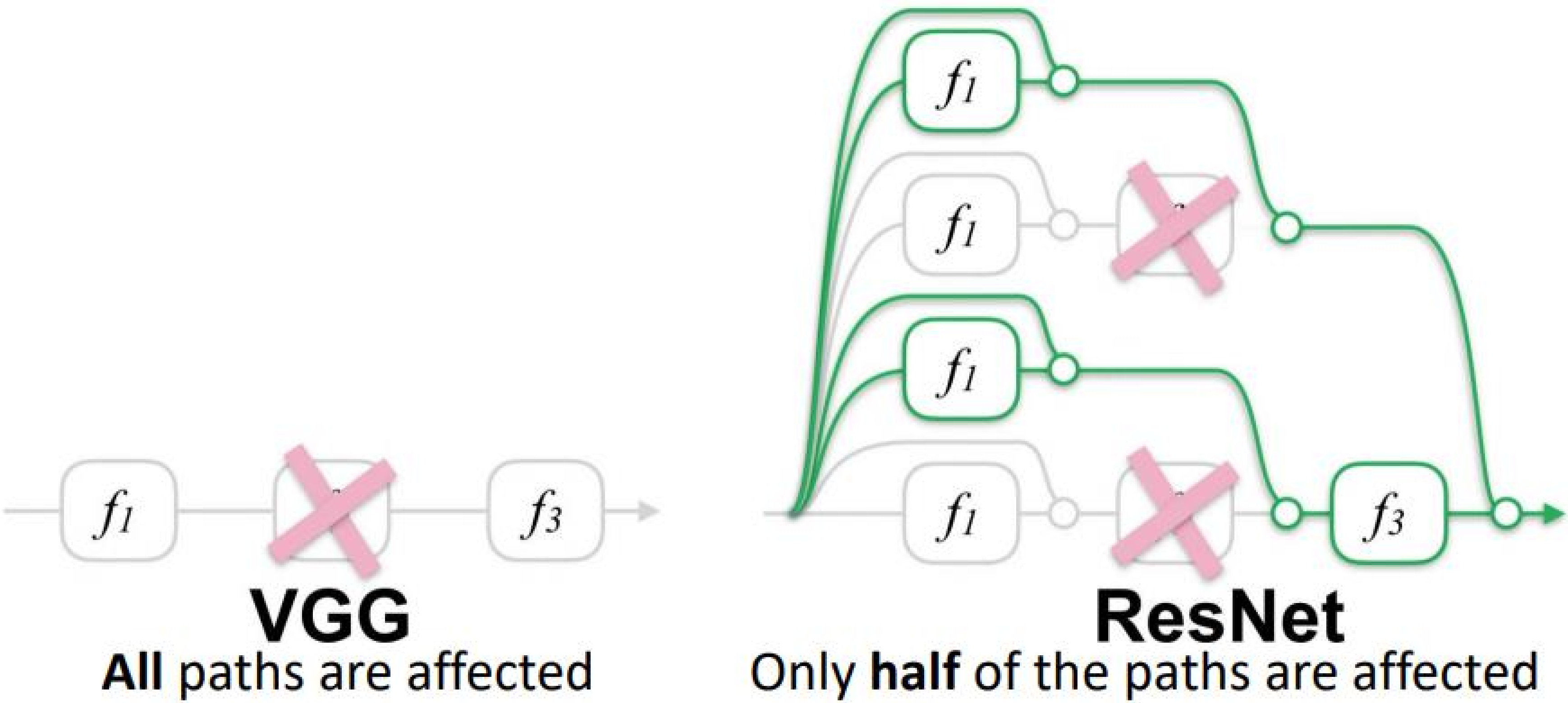


Why does this work?

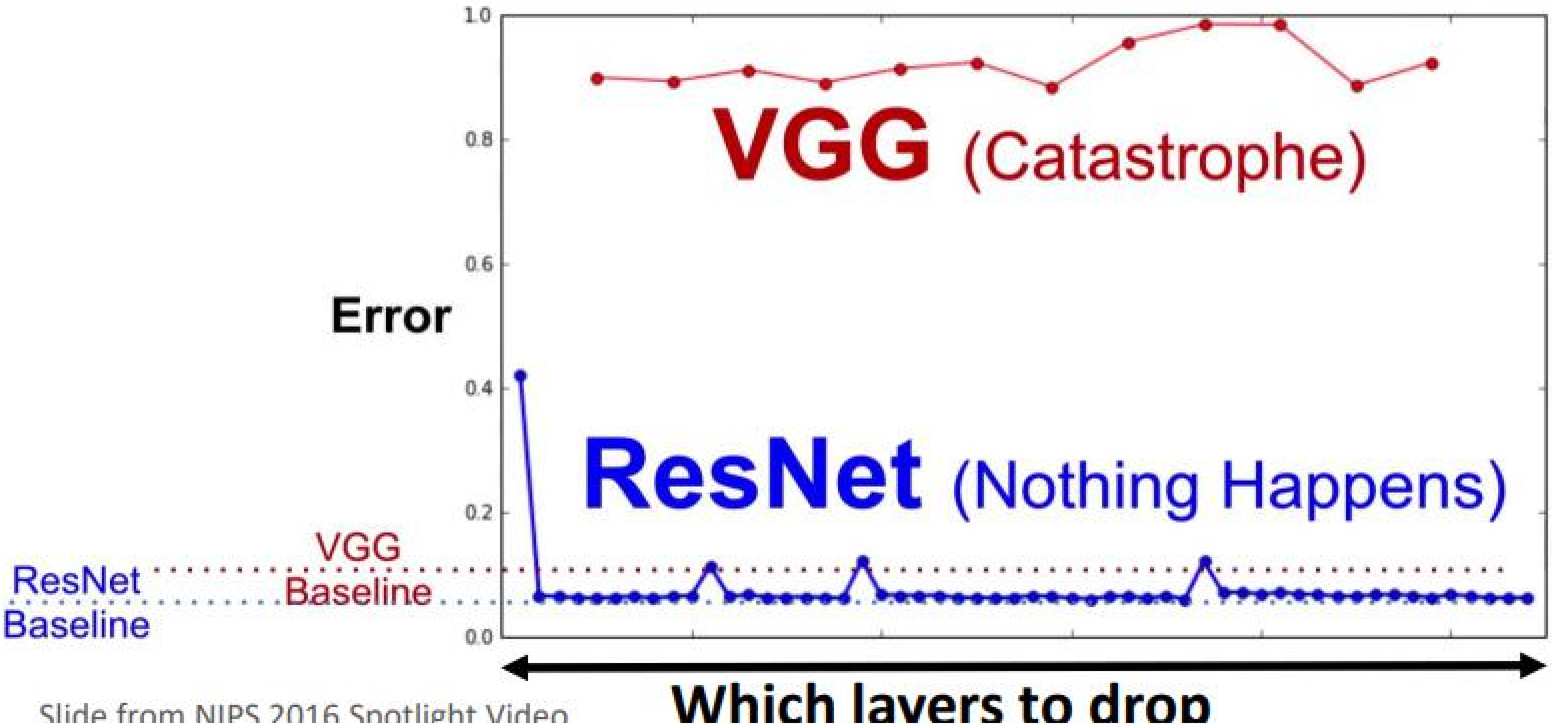
The unraveled view is equivalent and showcases the many paths in ResNet.



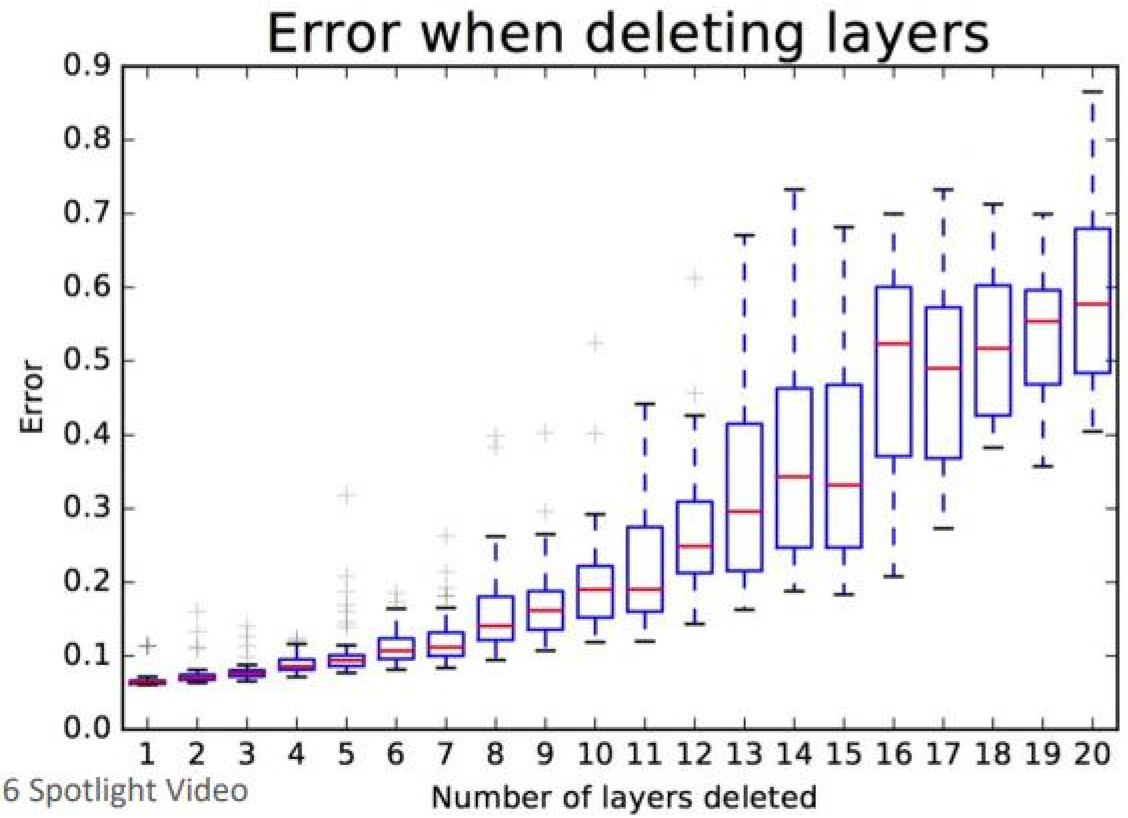
Deletion of one layer at test time



Deletion of one layer at test time



Deletion of several layers



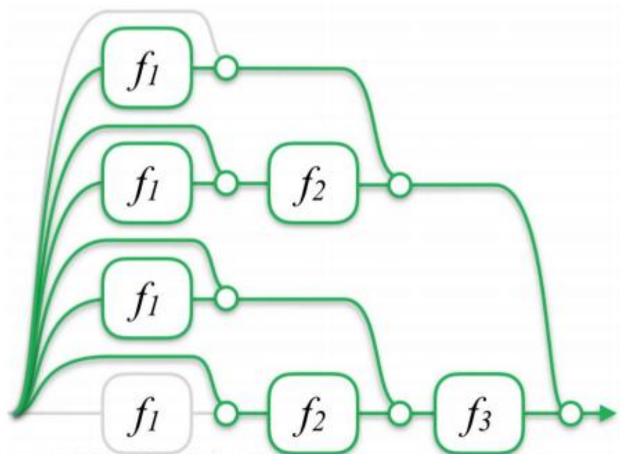
Slide from NIPS 2016 Spotlight Video

Conclusion 1

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

Path Length

Distribution of path length



rom NIPS 2016 Spotlight Video

There are very few **short paths…**

And very few **long paths…**

Most paths are **medium length!**

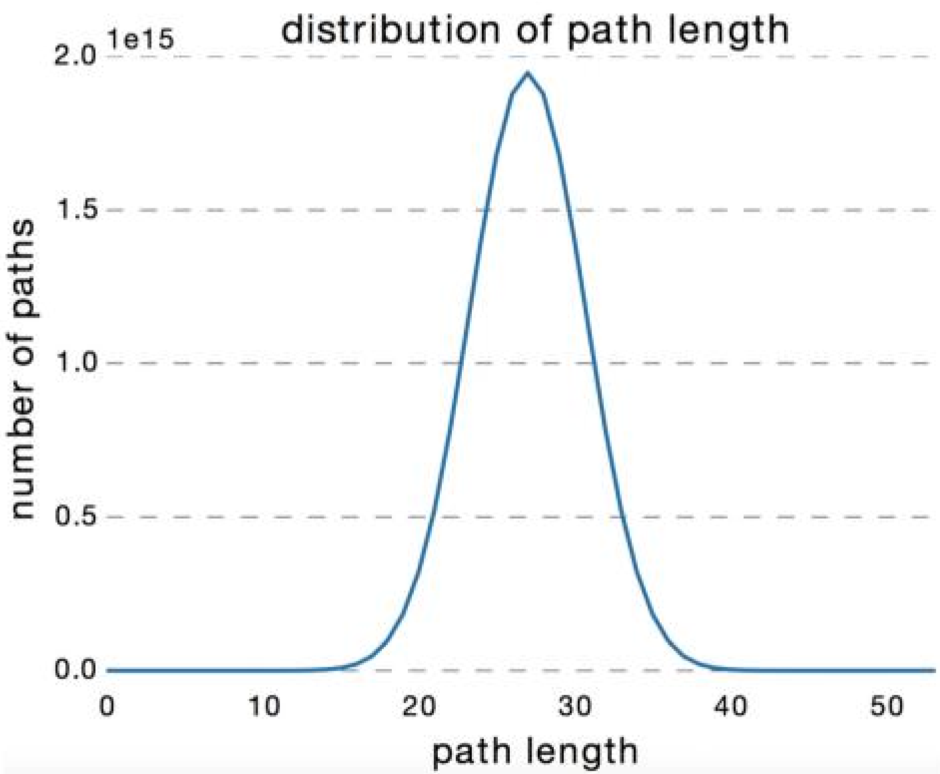
Residual networks contain many paths.

Previous networks have a single path.

Only short paths contribute gradient during training.

Vanishing gradient suppresses gradient from long paths.

Distribution of path lenth



There are very few **short paths…**

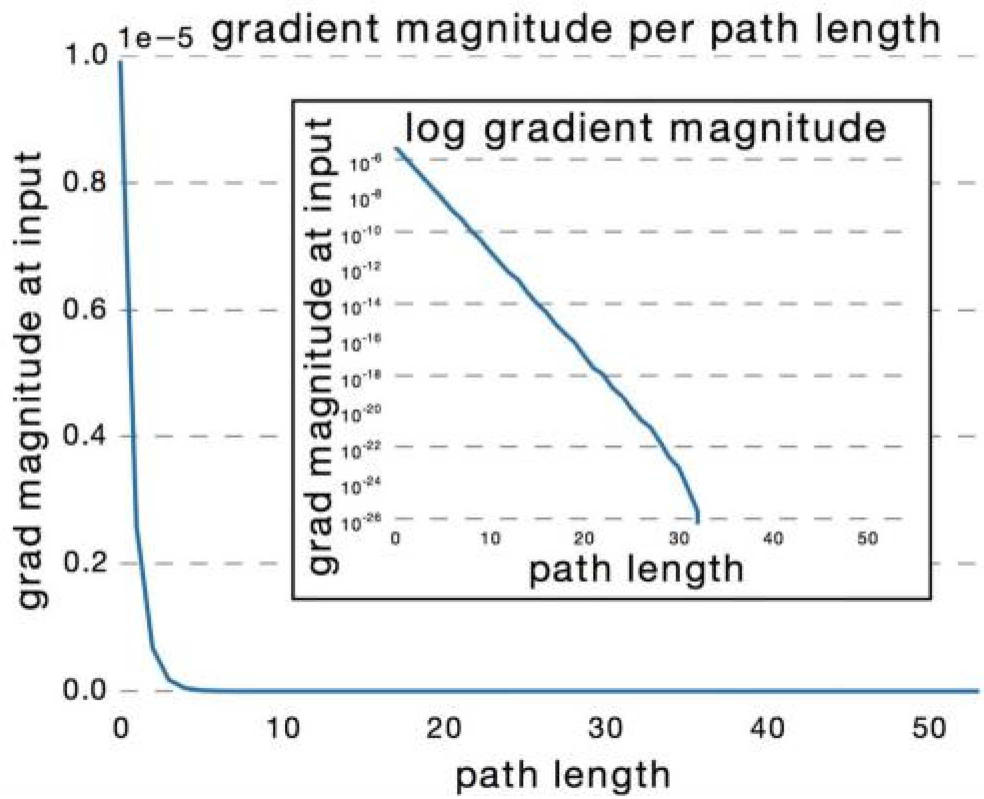
And very few **long paths…**

Most paths are **medium length!**

Paths length follows a **binomial distribution.**

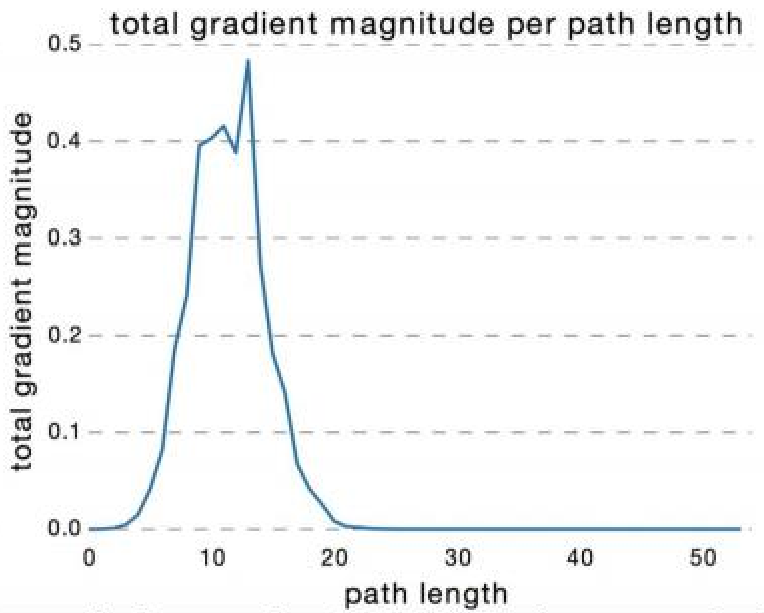
Vanishing gradient

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



rom NIPS 2016 Spotlight Video

Gradient during training with path length



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

Conclusion 2

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

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

Reference:

- Residual Networks Behave Like Ensembles of Relatively Shallow Networks

- Deep Residual Learning for Image Recognition