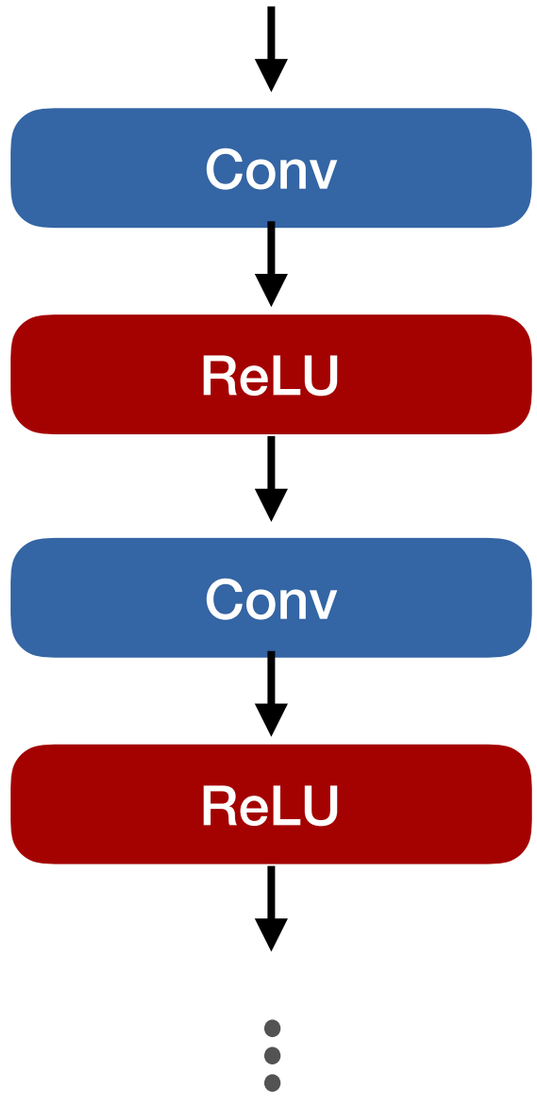
Residual connections

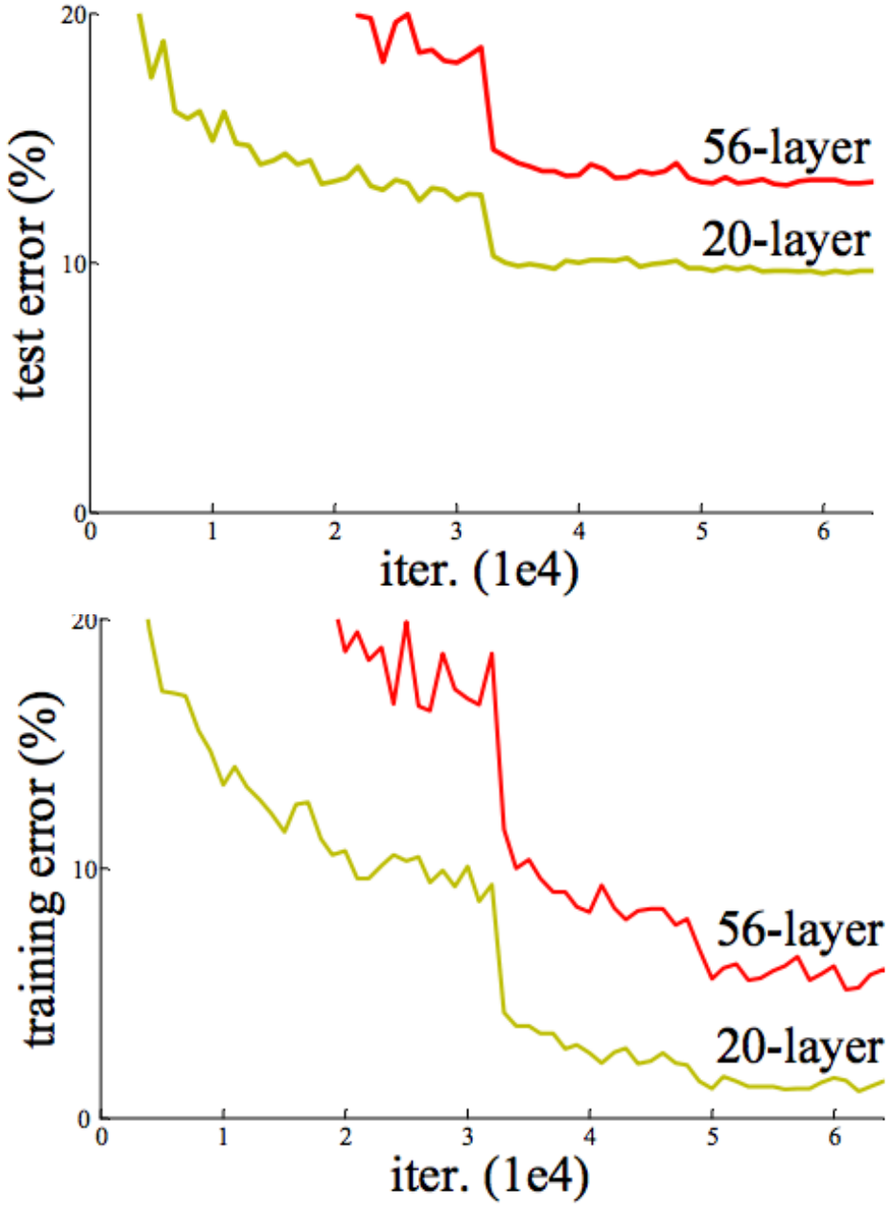
Deep networks

- Without normalization
  - Max depth 10-12
- With normalization
  - Max depth 20-30



What happens to deeper networks?

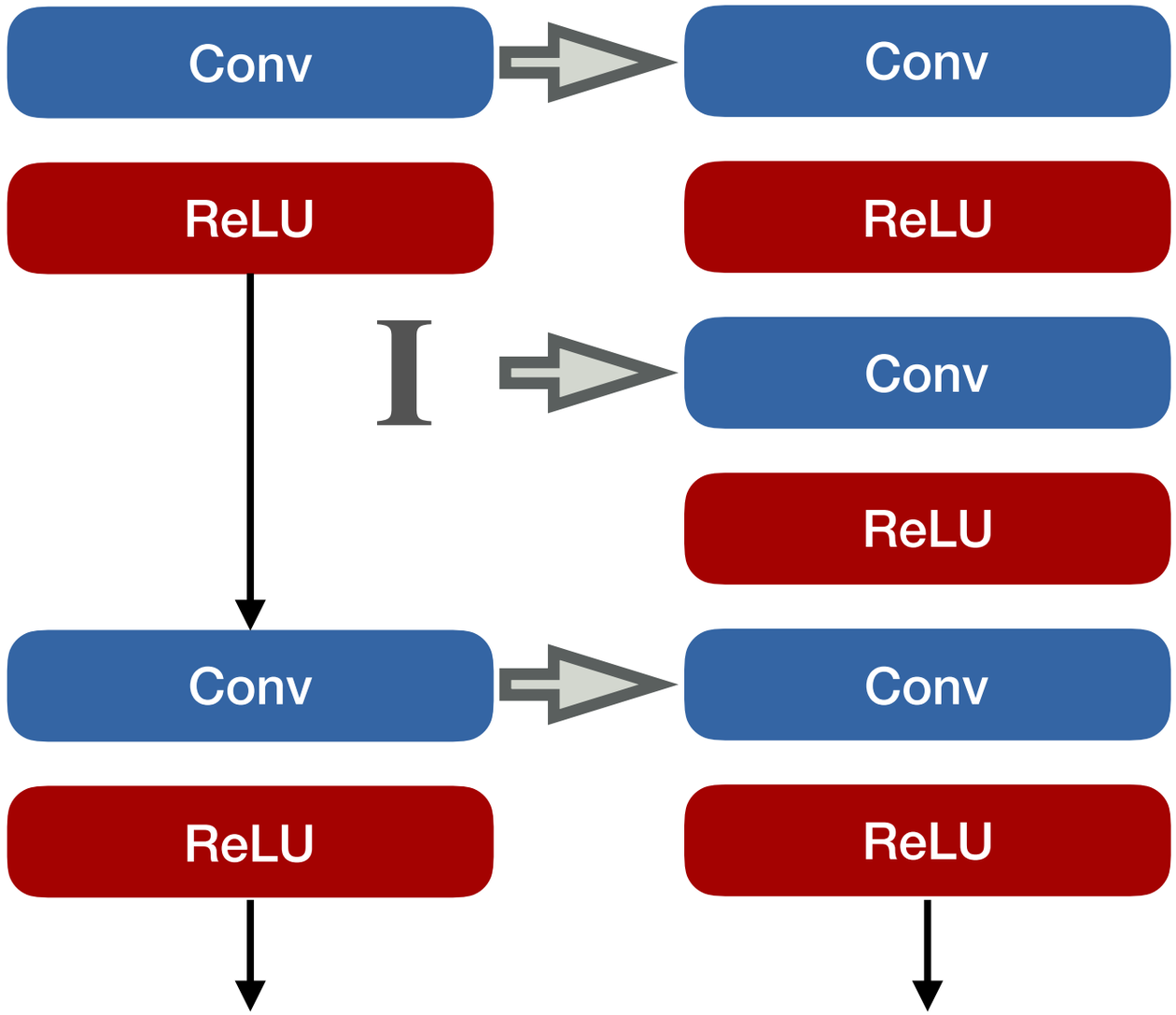
- It does not train well



[Figure source: Kaiming He et al., "Deep Residual Learning for Image Recognition", CVPR 2016]

What happens to deeper networks?

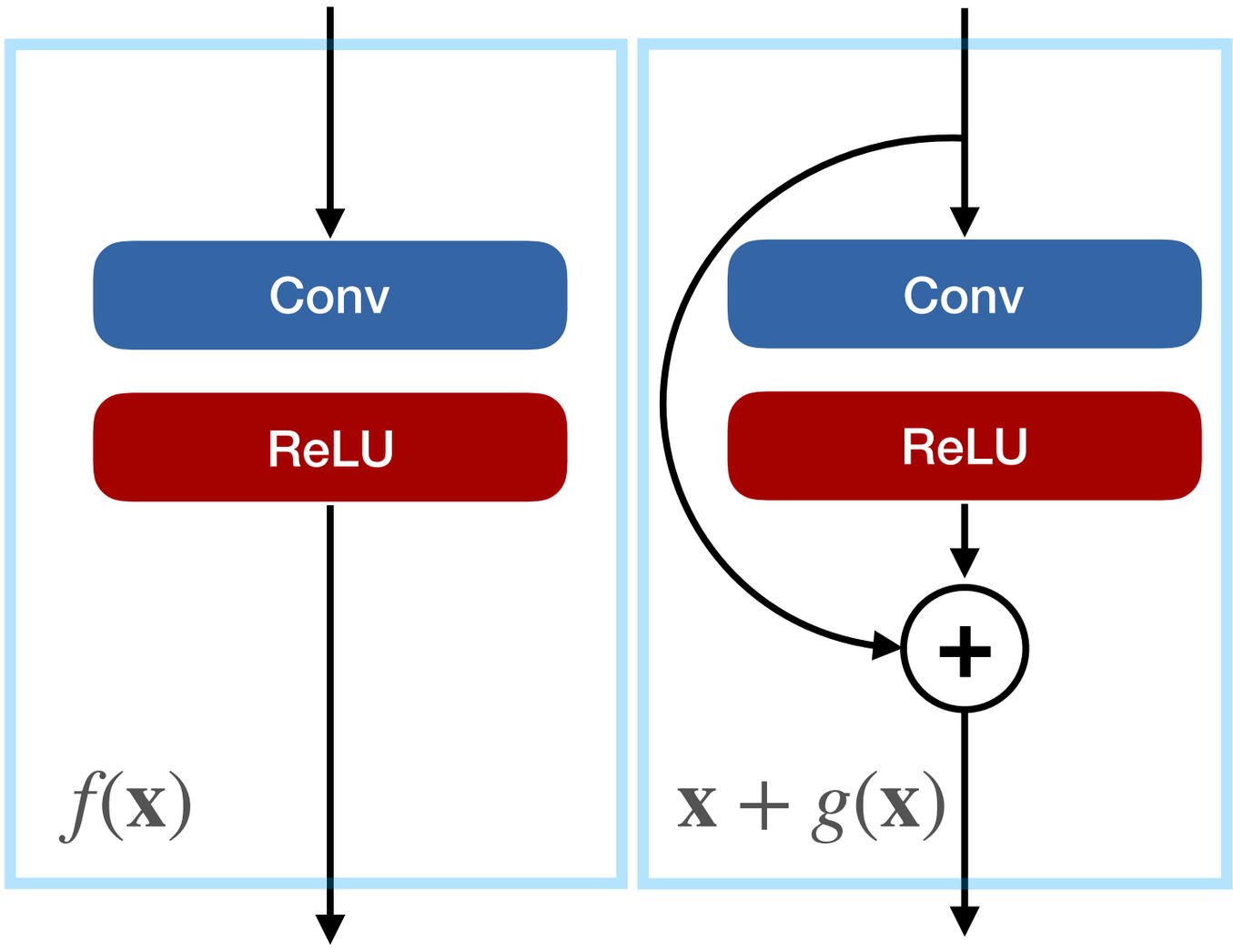
- Training a shallower network and adding identity layers works better



Solution: Residual connections

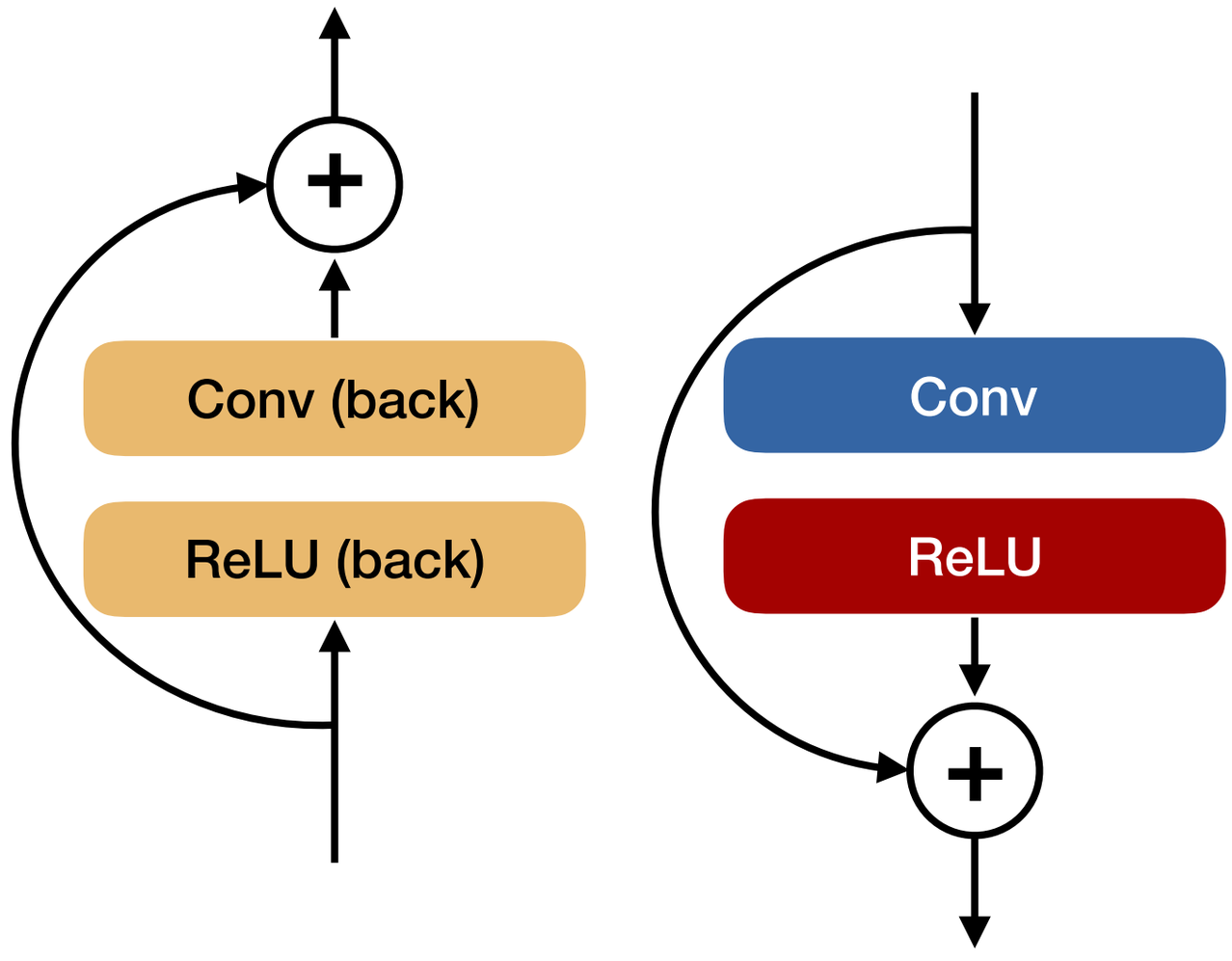
- Parametrize layers as

$$f(\mathbf{x}) = \mathbf{x} + g(\mathbf{x})$$

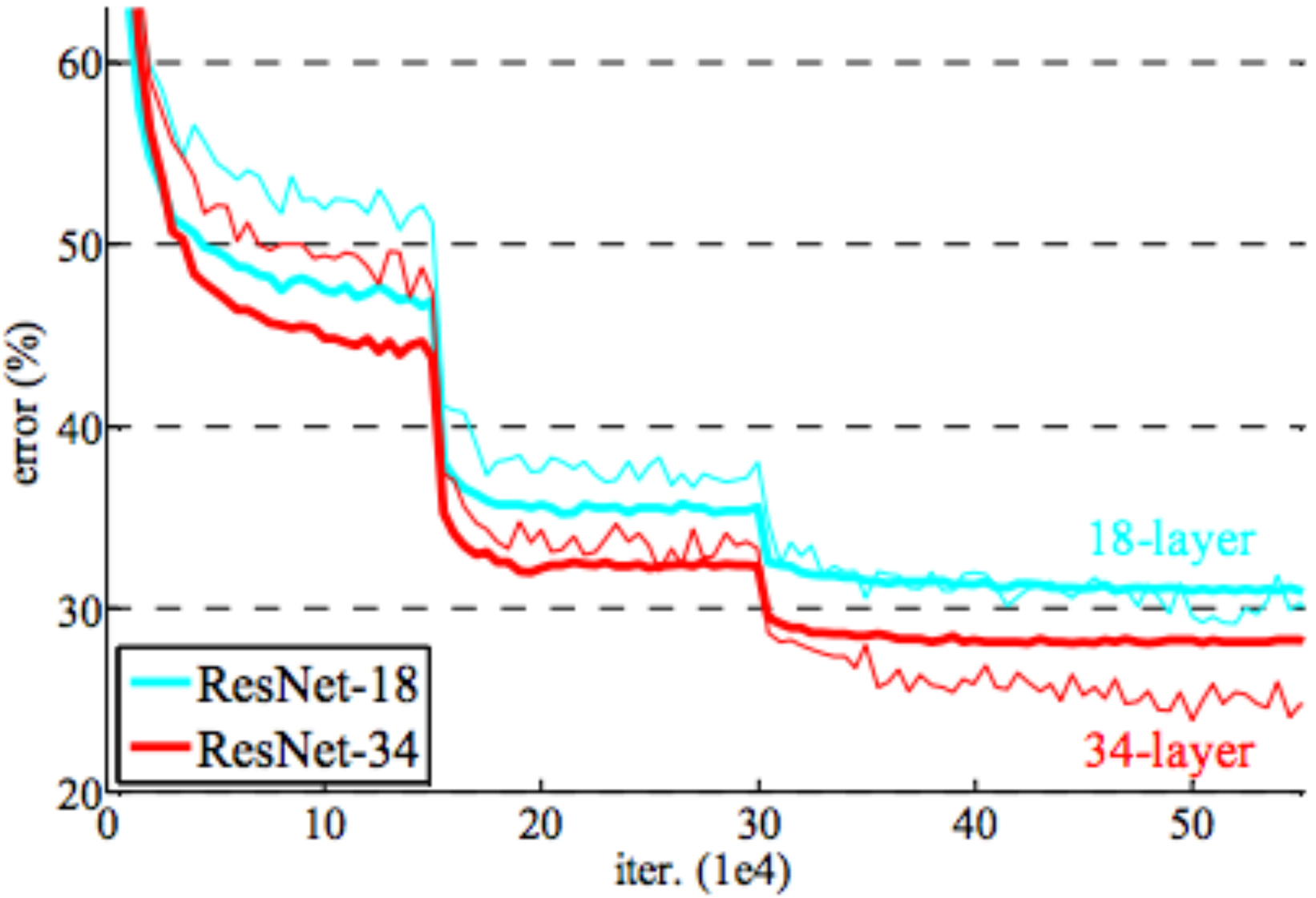


Fun fact

- Backward graph is symmetric



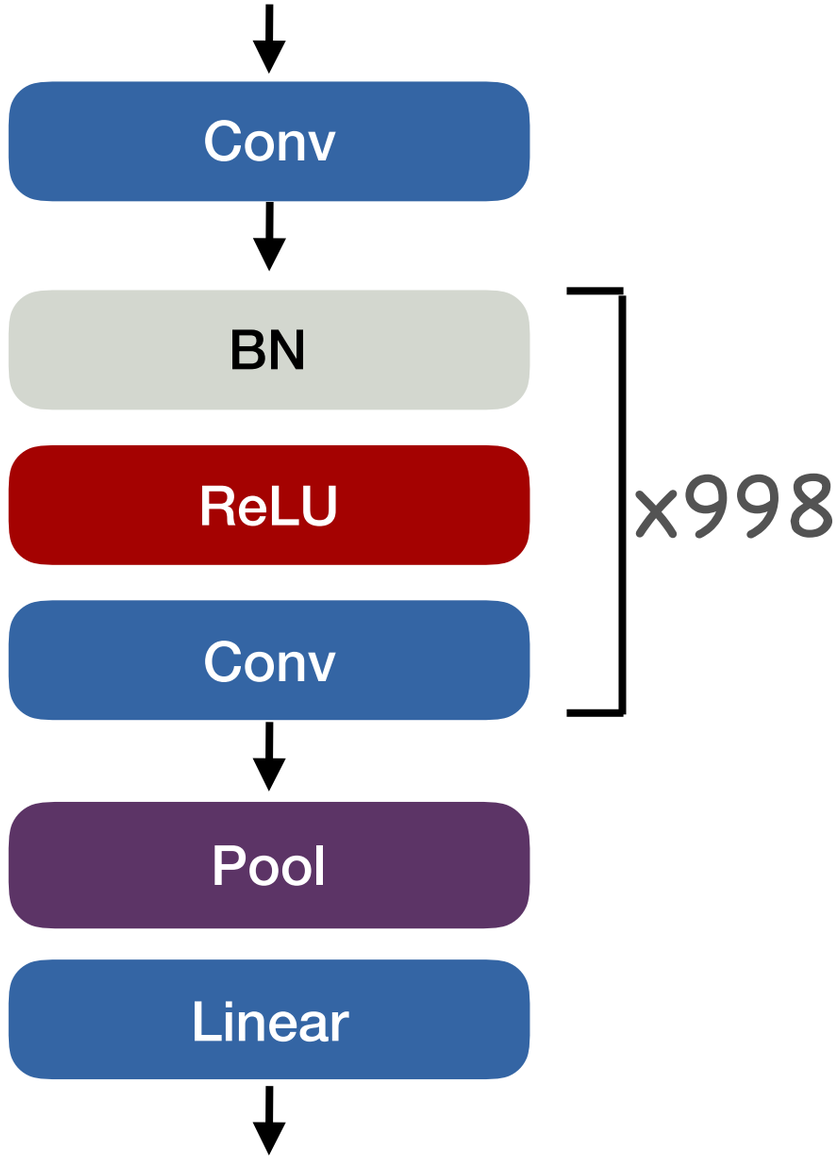
Residual Networks



[Figure source: Kaiming He et al., "Deep Residual Learning for Image Recognition", CVPR 2016]

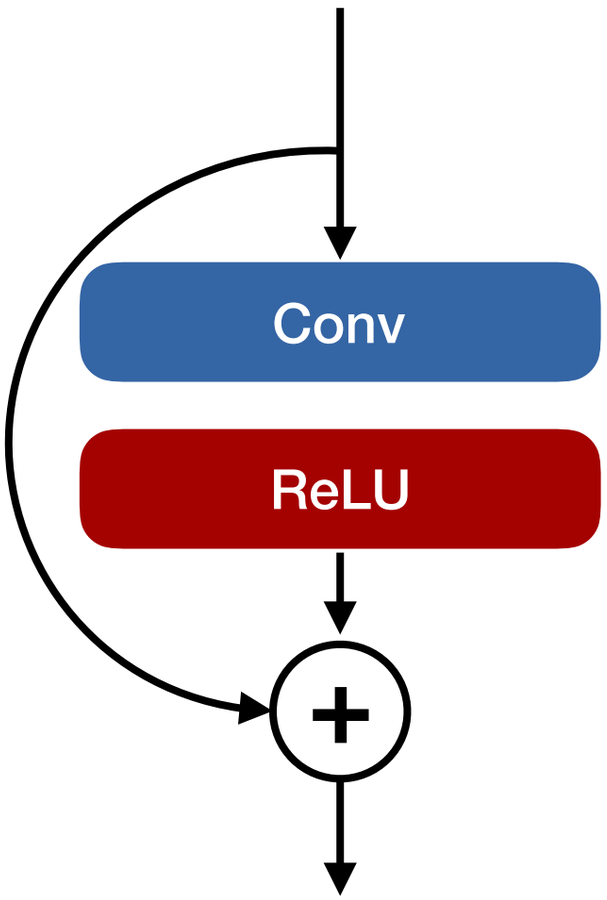
How well do residual connections work?

- Can train networks of up to 1000 layers



Why do residual connection work? - Practical answer

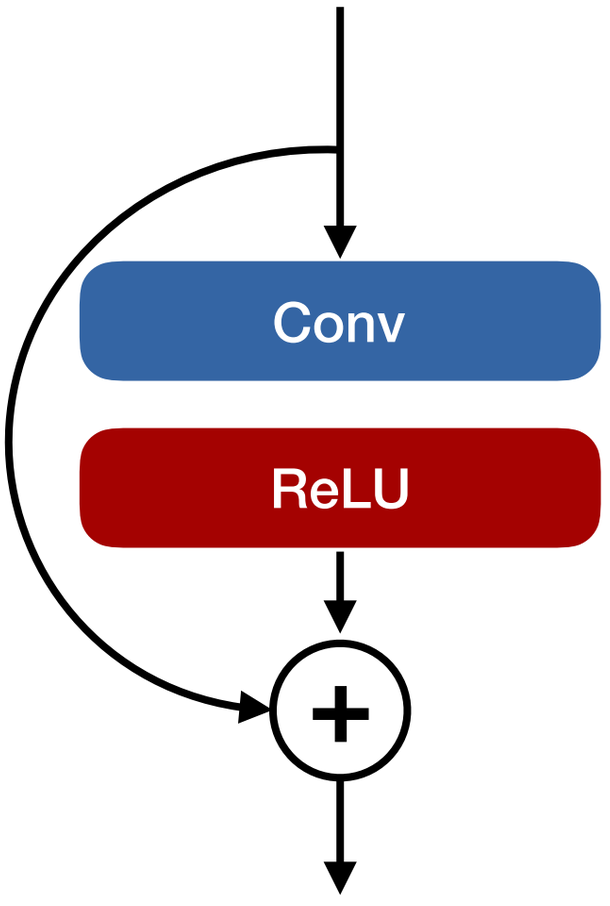
- Gradient travels further without major modifications (vanishing)
- Reuse of patters
  - Only update patterns
  - Dropping some layers does not even hurt performance
- Weights $\rightarrow 0$
  - Model $\rightarrow$ identity



[Gao Huang et al., "Deep Networks with Stochastic Depth", ECCV 2016]

Why do residual connection work? - Theoretical answer

- Without ReLU
  - Invertible functions
- Very wide
  - SGD find global optimum



[Moritz Hardt and Tengyu Ma, “Identity matters in deep learning”, ICLR 2017]

[Simon S. Du, et al., "Gradient Descent Finds Global Minima of Deep Neural Networks", ICML 2019]

Residual connections - Summary

- Used in most modern networks
- Allow for much deeper networks