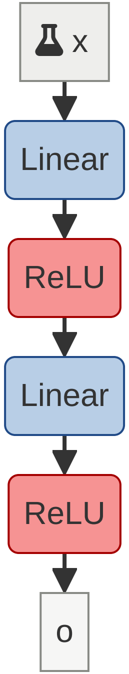
Residual Connections

Deep Networks

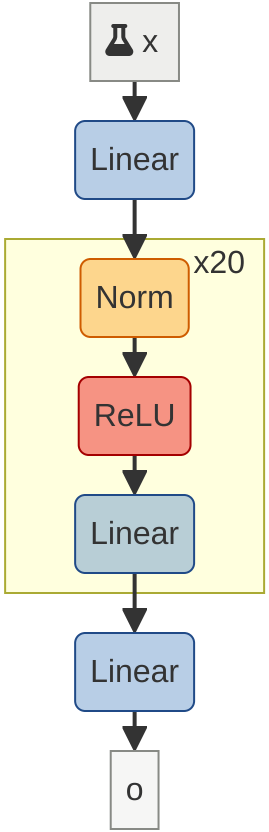
Without normalization

- Max depth 10-12



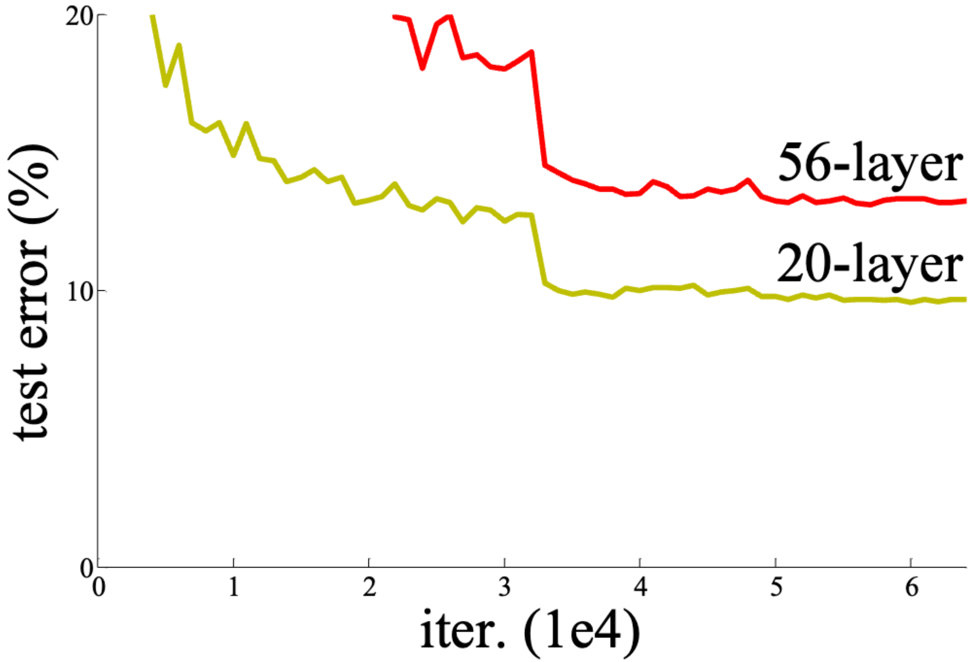
With normalization

- Max depth 20-30



What Happens to Deeper Networks?

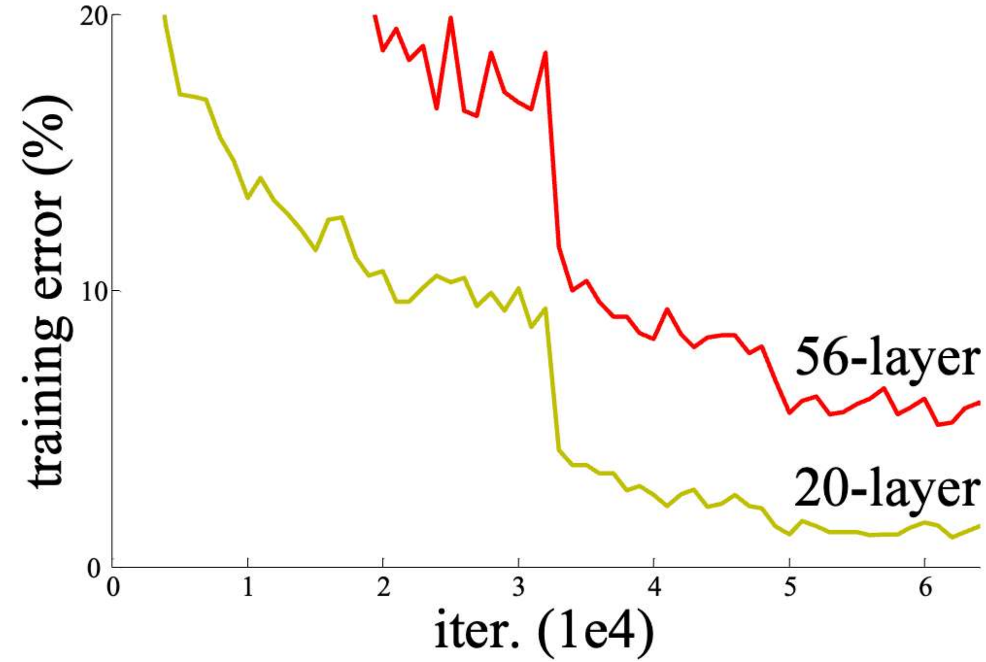
They don’t perform well!



Kaiming He et al., "Deep Residual Learning for Image Recognition", CVPR 2016 ↩

What Happens to Deeper Networks?

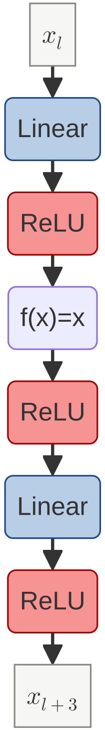
They don't even train well!



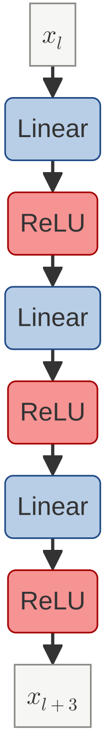
Kaiming He et al., "Deep Residual Learning for Image Recognition", CVPR 2016 ↩️

20 vs 50 Layers Networks

20 layers with identity-blocks



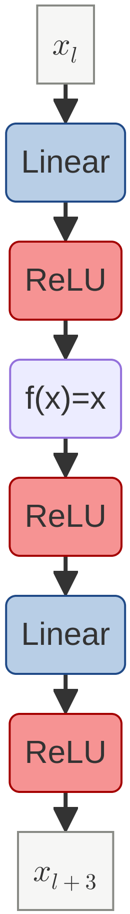
50 layers

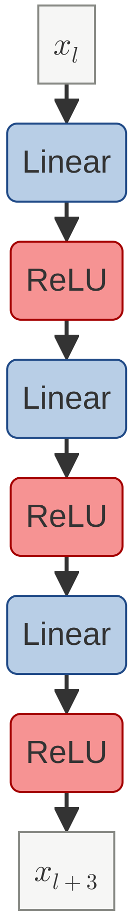


Why don't they train well?

Initial updates are hard

- Initial weights: Random Gaussian
- After a few layers:
  - Inputs look Gaussian (random noise)
  - Gradients look Gaussian (random noise)

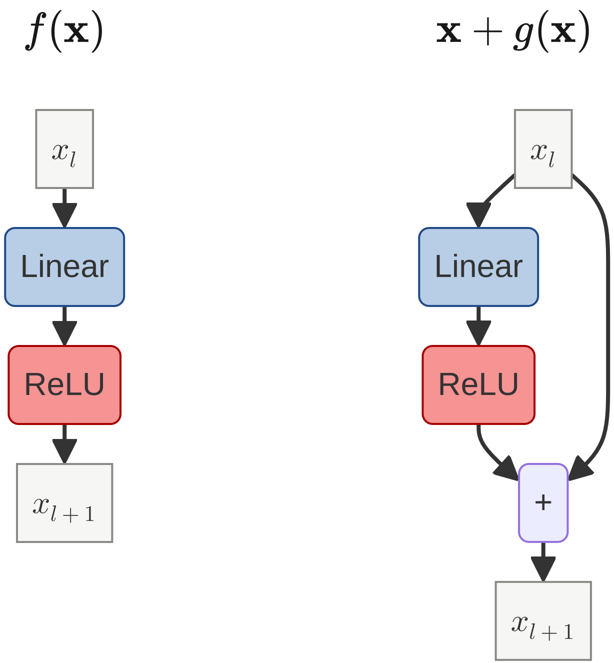




Solution: Residual Connections

Parameterize layers as

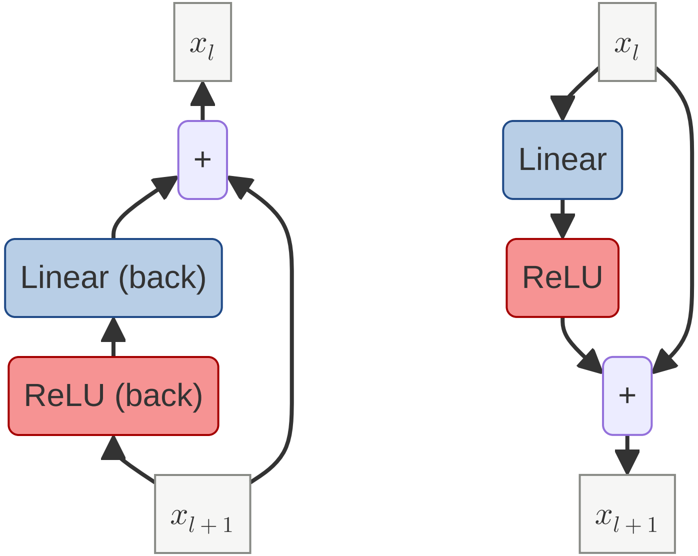
$$f(\mathbf{x}) = \mathbf{x} + g(\mathbf{x})$$



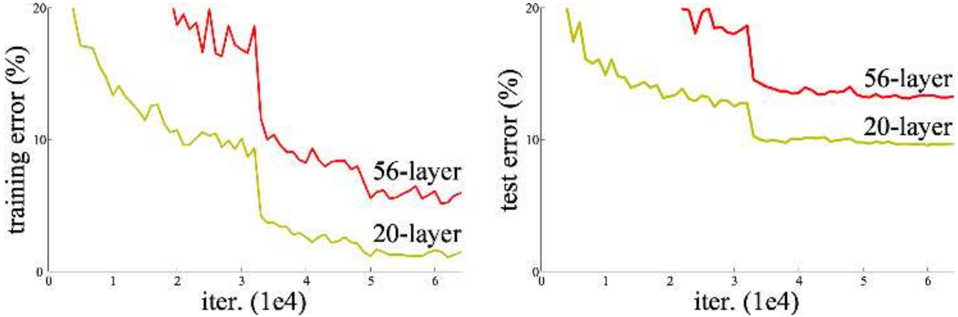
Kaiming He et al., "Deep Residual Learning for Image Recognition", CVPR 2016 ↩️

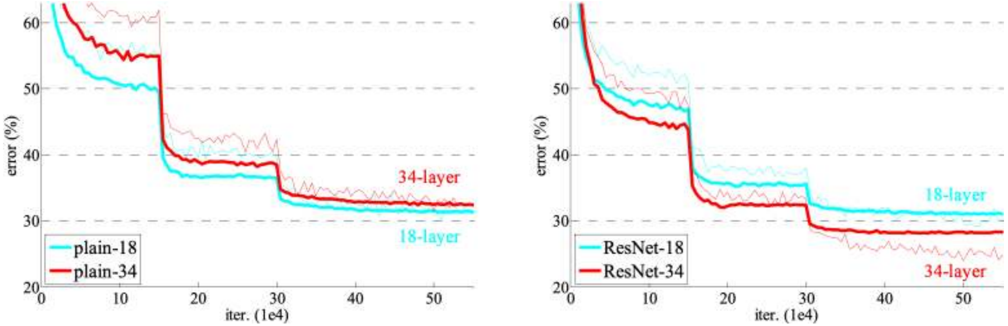
Fun Fact

Backward graph is symmetric



Residual Networks

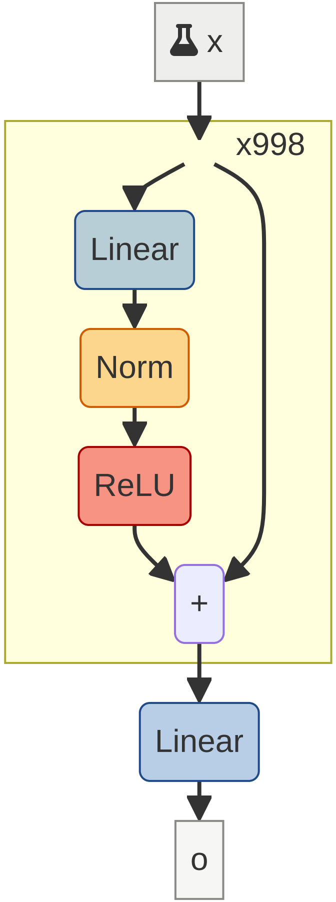




Kaiming He et al., "Deep Residual Learning for Image Recognition", CVPR 2016 ↩️

How Well Do Residual Connections Work?

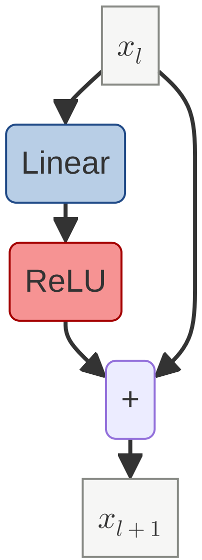
Can train networks of **up to 1000 layers**

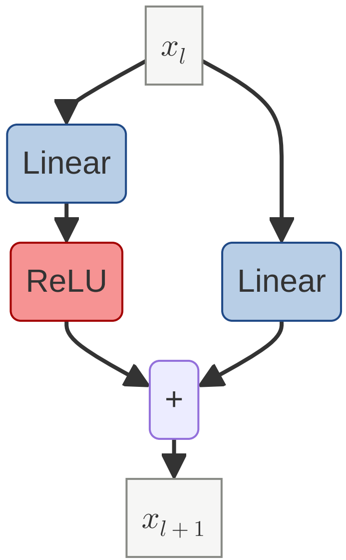


What if Input and Output Are Not the Same Size?

Add a linear layer to reshape

- Design network with same `input.shape = output.shape` for most blocks





Why Do Residual Connection Work? - Practical Answer

Gradient Travels Further

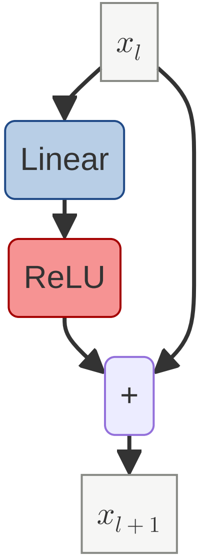
- Another way to prevent vanishing gradients

Reuse of Patterns

- Only update patterns

Can even drop some layers[1]:

- Dropping some layers still does well
- As weights $\implies 0$, model $\implies$ identity

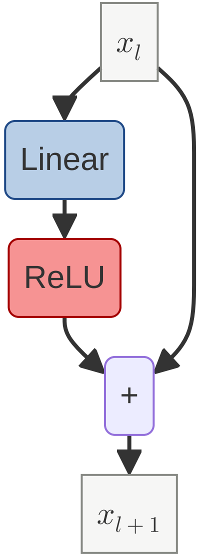


1. Gao Huang et al., "Deep Networks with Stochastic Depth", ECCV 2016 ↩︎

Why Do Residual Connection Work? - Theoretical Answer

Optimization [1,2]

- Invertibility
- Model capacity: very wide
- Simplified "loss landscape" for SGD



1. Simon S. Du, et al., "Gradient Descent Finds Global Minima of Deep Neural Networks", ICML 2019 ↩
2. Moritz Hardt and Tengyu Ma, “Identity matters in deep learning”, ICLR 2017 ↩

Residual TL;DR

Go deeper with residual connections

Residuals + Normalization fixes vanishing activations and gradients