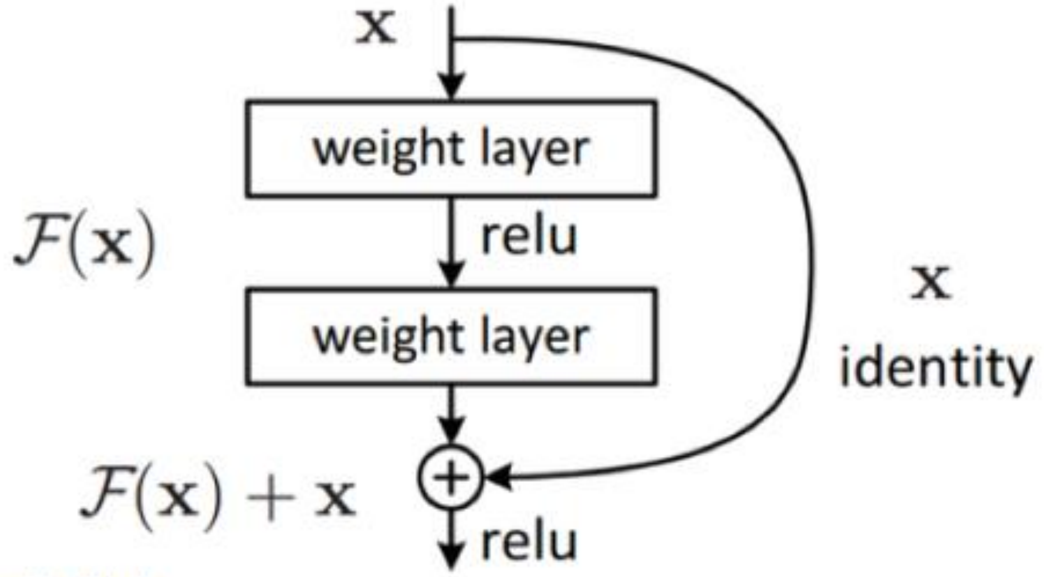
ResNets
DenseNets
HighwayNets



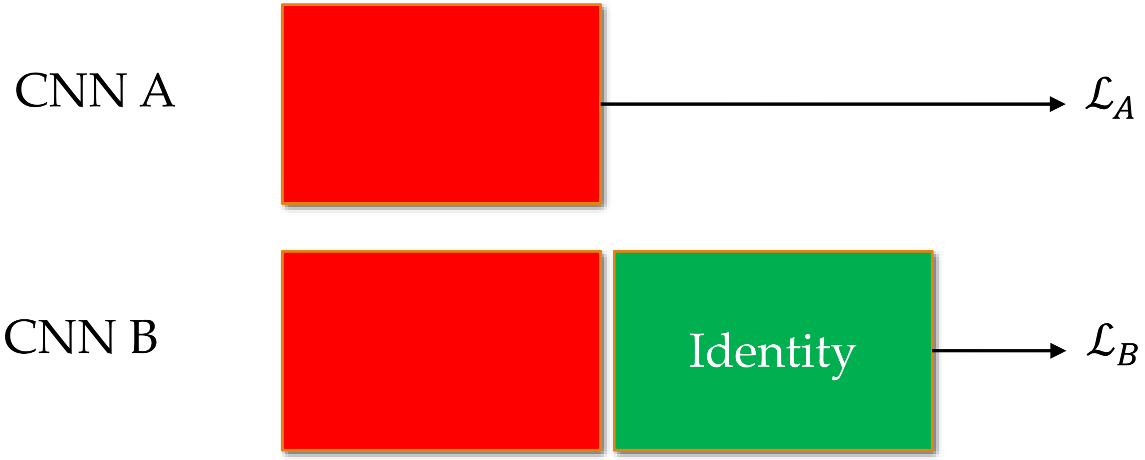
A residual block

Some facts

- The first truly Deep Network, going deeper than 1,000 layers
- The first deep architecture to gracefully go deeper than a few dozen layers
  - Not simply getting more GPUs, more training time, adding classifiers, etc
- Smashed Imagenet, with a ~3% error (with ensembles)
- Won all object classification, detection, segmentation, etc. challenges

Hypothesis

- **Hypothesis:** Can we have a very deep network at least as accurate as averagely deep networks?
- **Thought experiment:** Let's assume two almost identical convnets A, B
  - B is the same as A, just with extra "identity" layers
  - Identity layers pass information unchanged → their errors *should* be similar
  - Thus, there is at least one Convnet B as good as A w.r.t. training error



Testing the hypothesis

- Trainng a shallow and a deeper architecture
- The deeper model does worse **in training error**!
- **Performance degradation** not by overfitting → just harder optimization
- Assuming optimizers are doing their job fine
  - not all networks are the same as easy to optimize

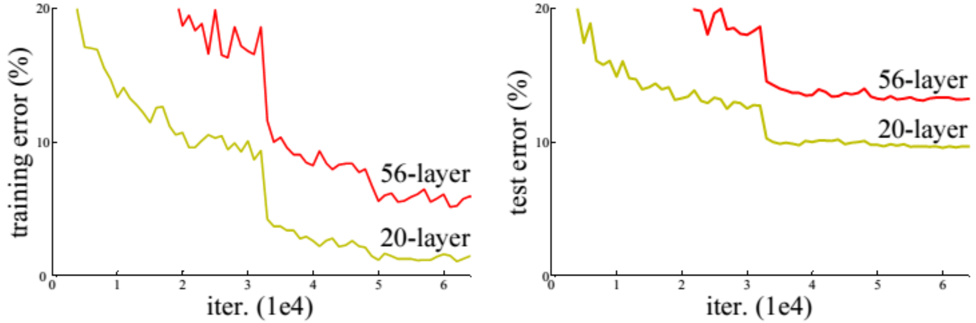


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

Residual connections to the rescue

- Add to your module output $F(x)$ the input x

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

- If dimensions don't match zero padding or a projection layer

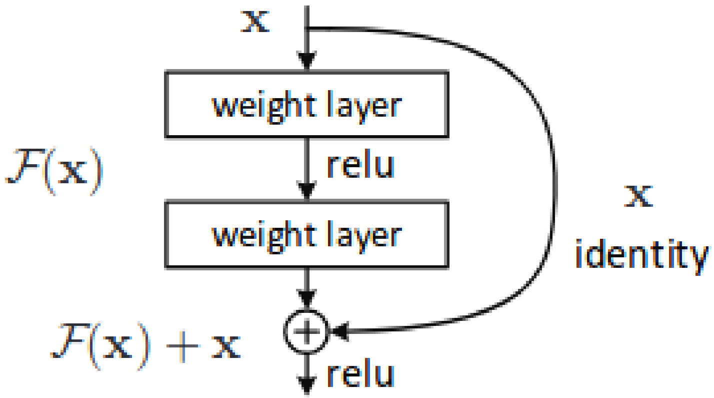
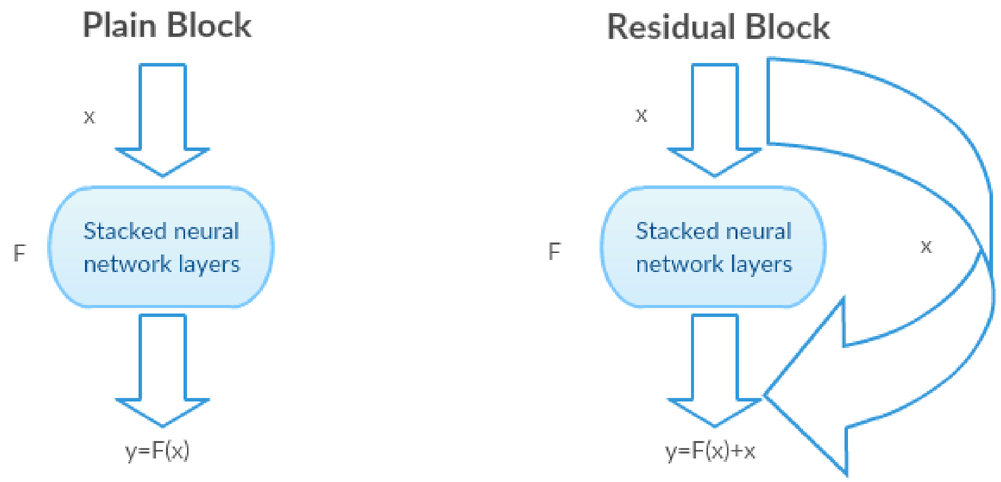

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual learning: a building block.



Hard to get F(x)=x and make y=x an identity mapping

Easy to get F(x)=0 and make y=x an identity mapping

No degradation anymore

- Without residual connections deeper networks attain worse scores

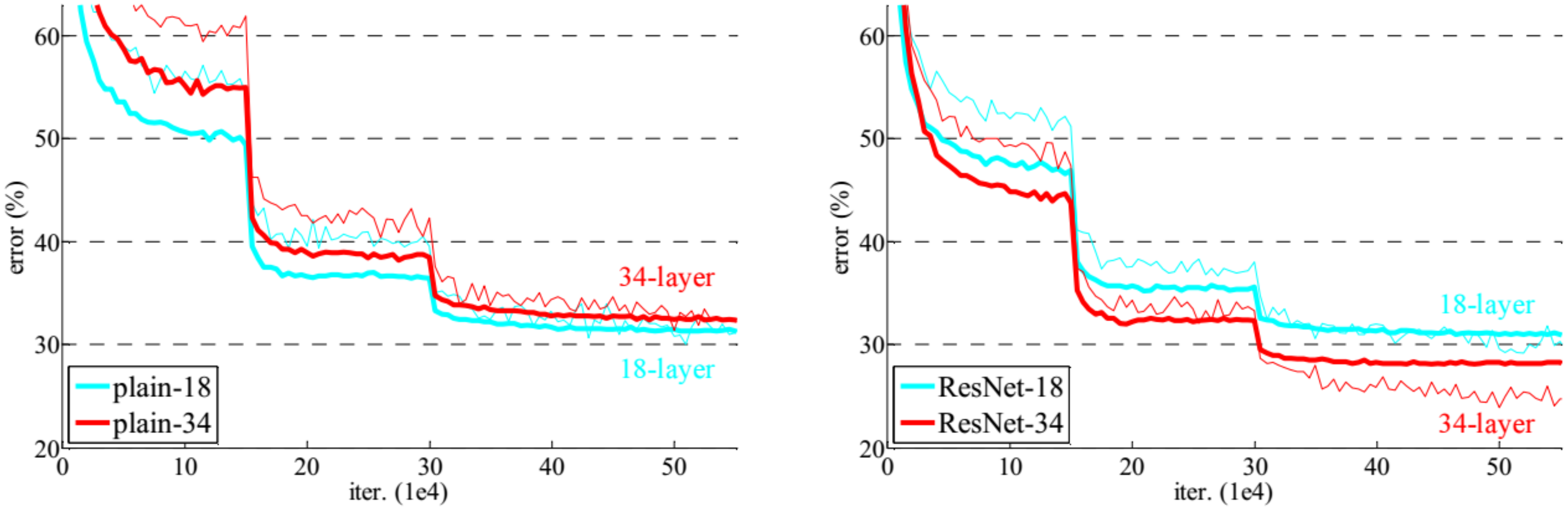
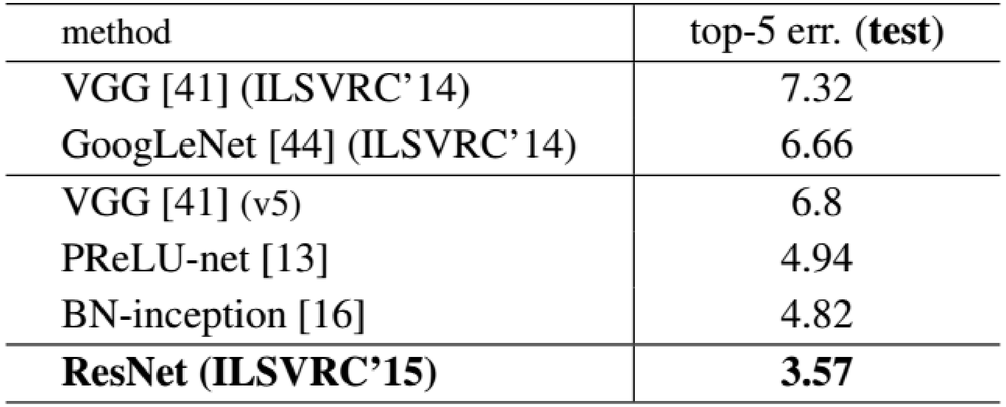


Figure 4. Training on **ImageNet**. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

ResNet breaks records

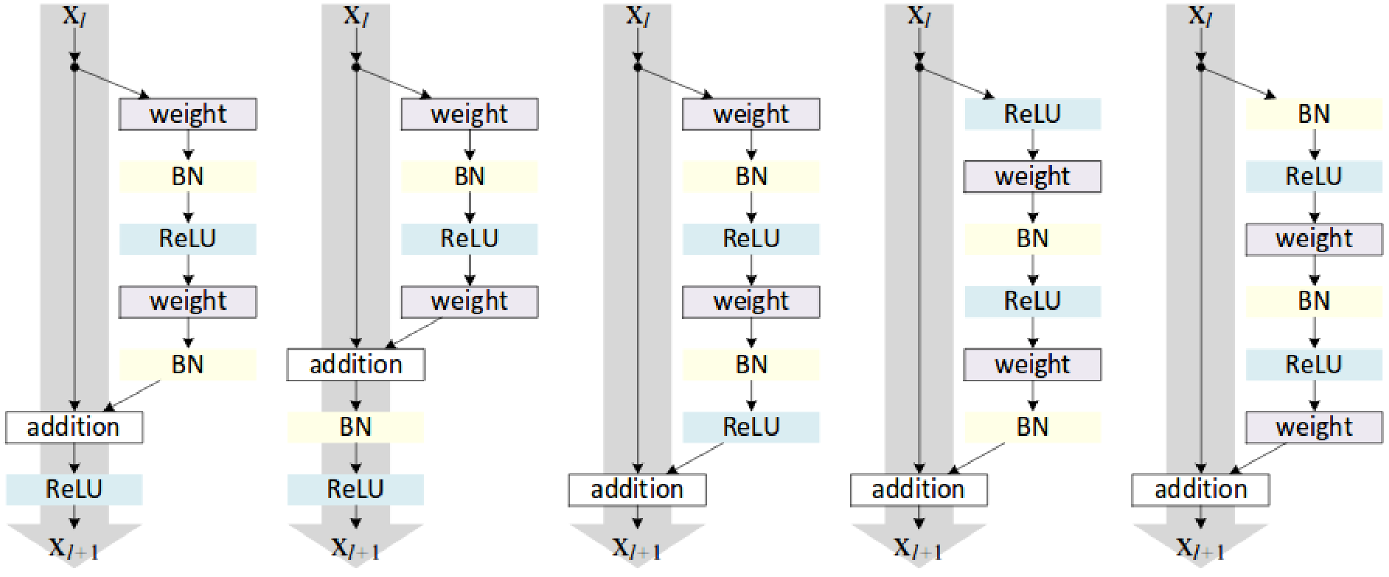
- Ridiculously low error in ImageNet
- Up to 1000 layers ResNets trained
  - Previous deepest network ~30-40 layers on simple datasets



method	top-5 err. (**test**)
VGG [41] (ILSVRC'14)	7.32
GoogLeNet [44] (ILSVRC'14)	6.66
VGG [41] (v5)	6.8
PReLU-net [13]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	**3.57**

Table 5. Error rates (%) of **ensembles**. The top-5 error is on the test set of ImageNet and reported by the test server.

ResNet architectures & ResNeXt



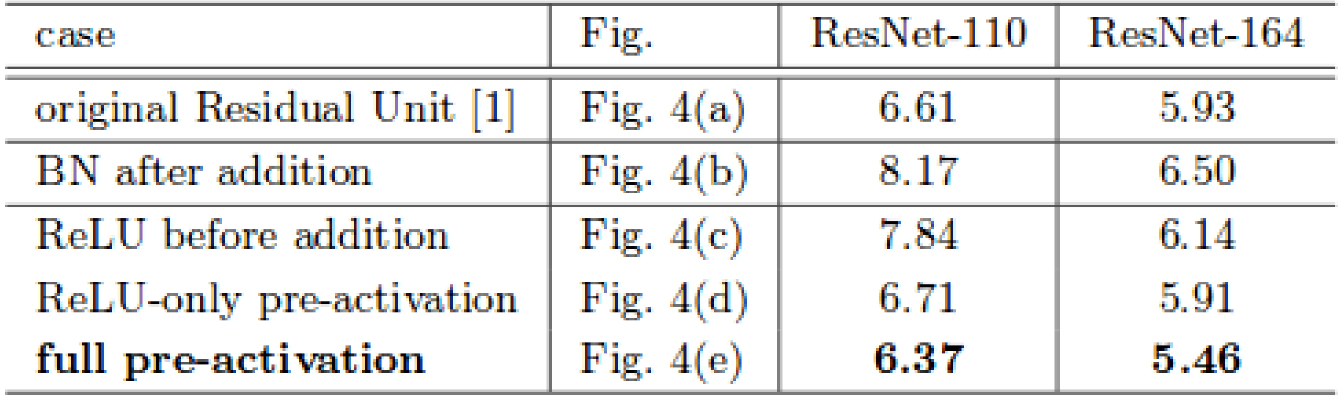
(a) original

(b) BN after addition

(c) ReLU before addition

(d) ReLU-only pre-activation

(e) **full pre-activation**



case	Fig.	ResNet-110	ResNet-164
original Residual Unit [1]	Fig. 4(a)	6.61	5.93
BN after addition	Fig. 4(b)	8.17	6.50
ReLU before addition	Fig. 4(c)	7.84	6.14
ReLU-only pre-activation	Fig. 4(d)	6.71	5.91
full pre-activation	Fig. 4(e)	**6.37**	**5.46**

ResNeXt

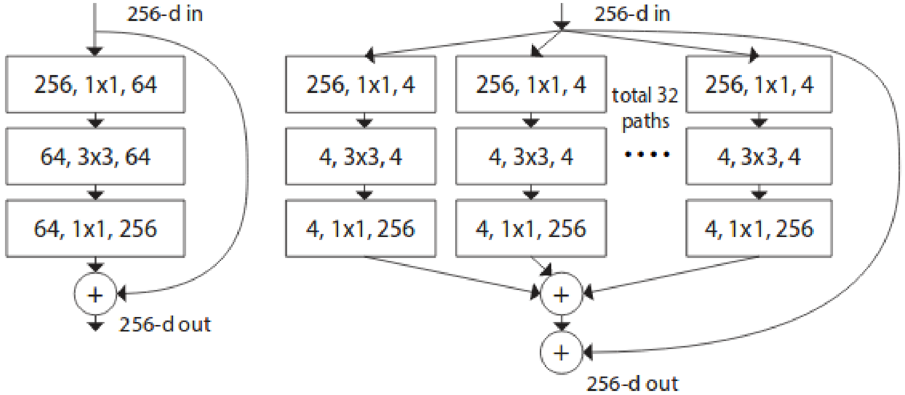
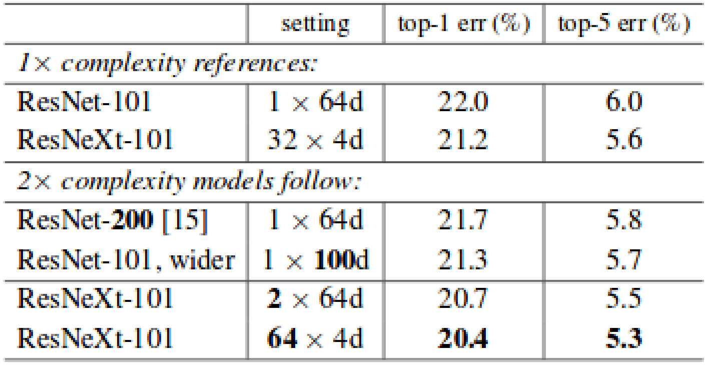


Figure 1. **Left**: A block of ResNet [14]. **Right**: A block of ResNeXt with cardinality = 32, with roughly the same complexity. A layer is shown as (# in channels, filter size, # out channels).



	setting	top-1 err (%)	top-5 err (%)
1× complexity references:			
ResNet-101	1 × 64d	22.0	6.0
ResNeXt-101	32 × 4d	21.2	5.6
2× complexity models follow:			
ResNet-**200** [15]	1 × 64d	21.7	5.8
ResNet-101, wider	1 × **100**d	21.3	5.7
ResNeXt-101	**2** × 64d	20.7	5.5
ResNeXt-101	**64** × 4d	**20.4**	**5.3**

Table 4. Comparisons on ImageNet-1K when the number of FLOPs is increased to 2× of ResNet-101's. The error rate is evaluated on the single crop of 224×224 pixels. The highlighted factors are the factors that increase complexity.

Aggregated Residual Transformations for Deep Neural Networks, Xie et al., 2016

CNNs and residual connections: insights

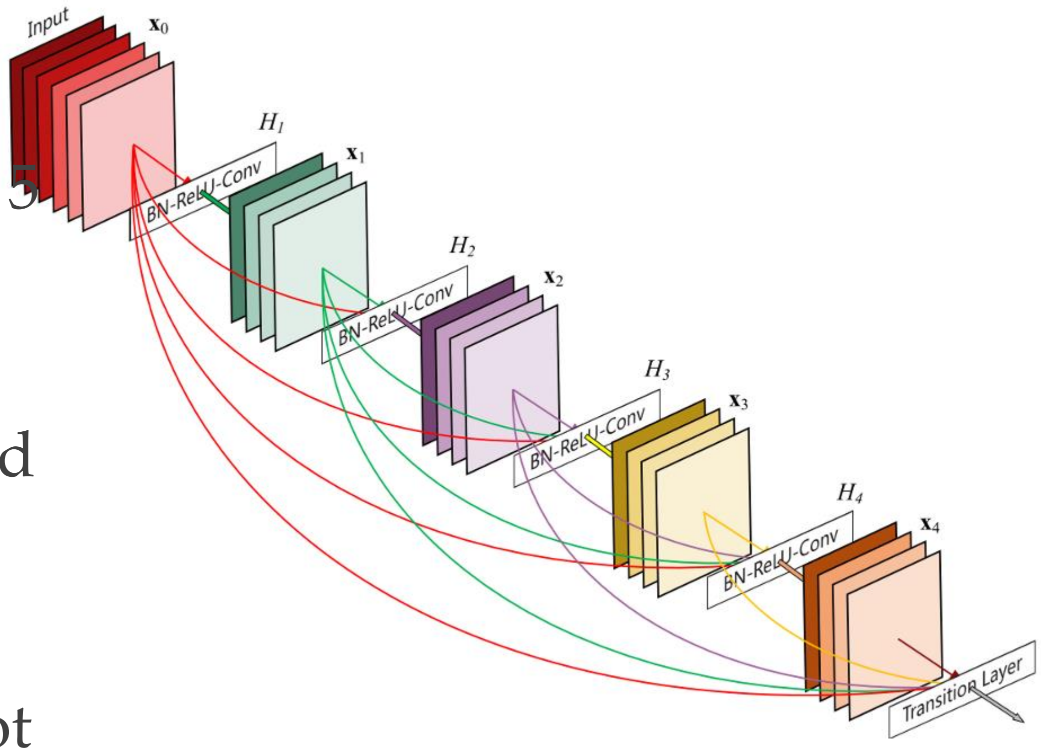
- BatchNorms absolutely necessary because of vanishing gradients
- Identity shortcuts cheaper and almost equal to project shortcuts
- Networks with skip connections converge faster
  - Compare to the same network without skip connections
- Generally, skip/residual connections are an asset for deeper architectures

DenseNet

- Add skip connections to multiple forward layers

$$y = h(x_l, x_{l-1}, \dots, x_{l-n})$$

- Assume layer 1 captures edges, while layer 5 captures faces (and other stuff)
- Why not have a layer that combines both faces and edges (e.g. to model a scarred face)
- Standard ConvNets do not allow for this
  - Layer 6 combines only layer 5 patterns, not lower



HighwayNet

- Similar to ResNets, but with a learnable gate per skip connection

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot (1 - T(x, W_T))$$

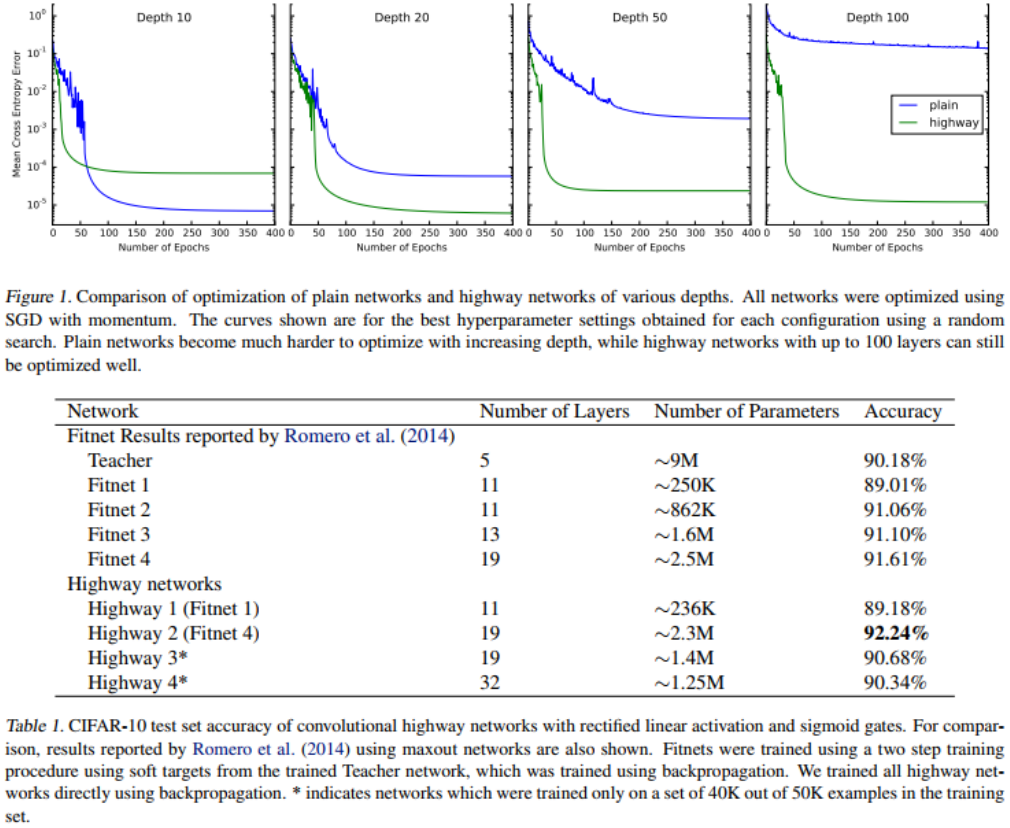


Figure 1. Comparison of optimization of plain networks and highway networks of various depths. All networks were optimized using SGD with momentum. The curves shown are for the best hyperparameter settings obtained for each configuration using a random search. Plain networks become much harder to optimize with increasing depth, while highway networks with up to 100 layers can still be optimized well.

Network	Number of Layers	Number of Parameters	Accuracy
Fitnet Results reported by Romero et al. (2014)			
Teacher	5	~9M	90.18%
Fitnet 1	11	~250K	89.01%
Fitnet 2	11	~862K	91.06%
Fitnet 3	13	~1.6M	91.10%
Fitnet 4	19	~2.5M	91.61%
Highway networks			
Highway 1 (Fitnet 1)	11	~236K	89.18%
Highway 2 (Fitnet 4)	19	~2.3M	**92.24%**
Highway 3*	19	~1.4M	90.68%
Highway 4*	32	~1.25M	90.34%

Table 1. CIFAR-10 test set accuracy of convolutional highway networks with rectified linear activation and sigmoid gates. For comparison, results reported by Romero et al. (2014) using maxout networks are also shown. Fitnets were trained using a two step training procedure using soft targets from the trained Teacher network, which was trained using backpropagation. We trained all highway networks directly using backpropagation. * indicates networks which were trained only on a set of 40K out of 50K examples in the training set.

Srivastava, Greff, Schmidhuber, Highway Networks, 2015