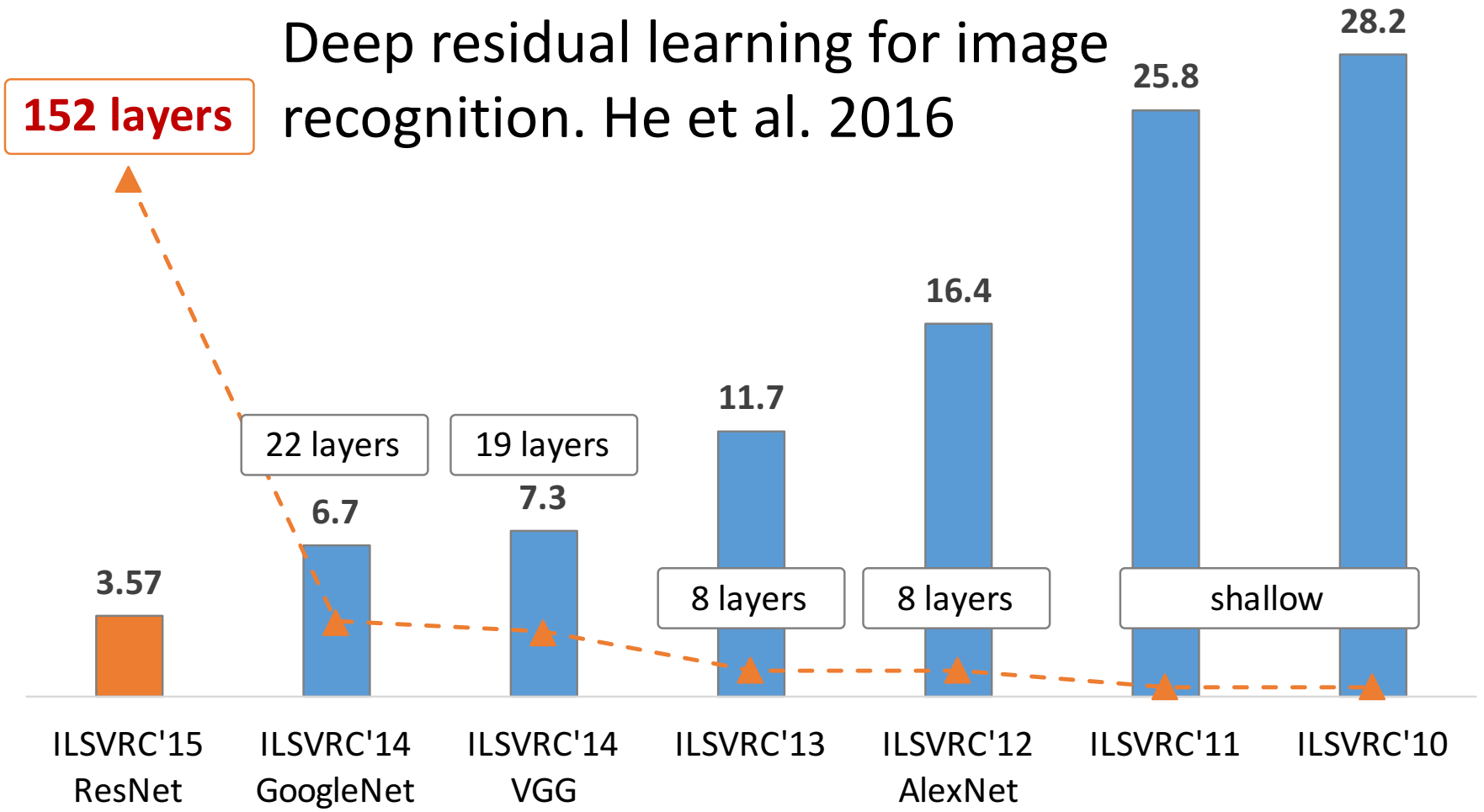
Tricks To Train Very Deep Nets

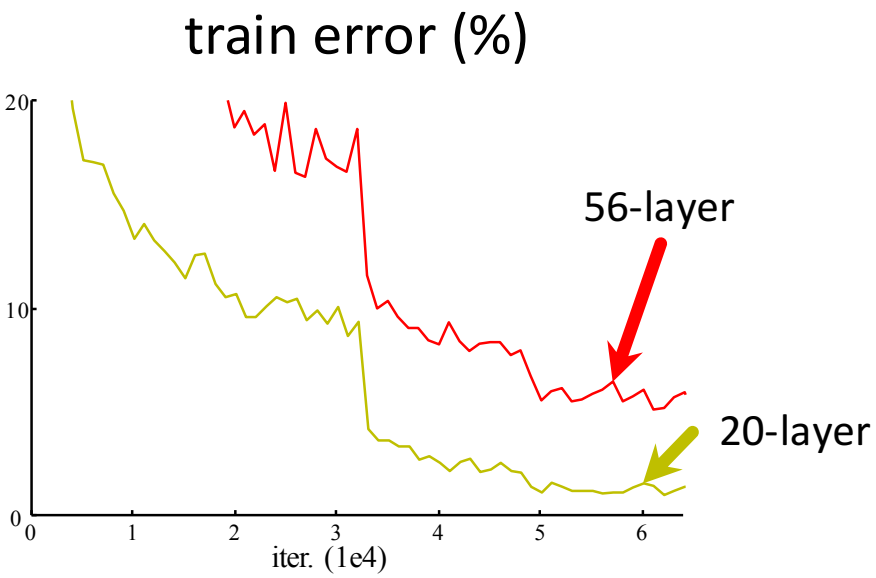
Thang Vu

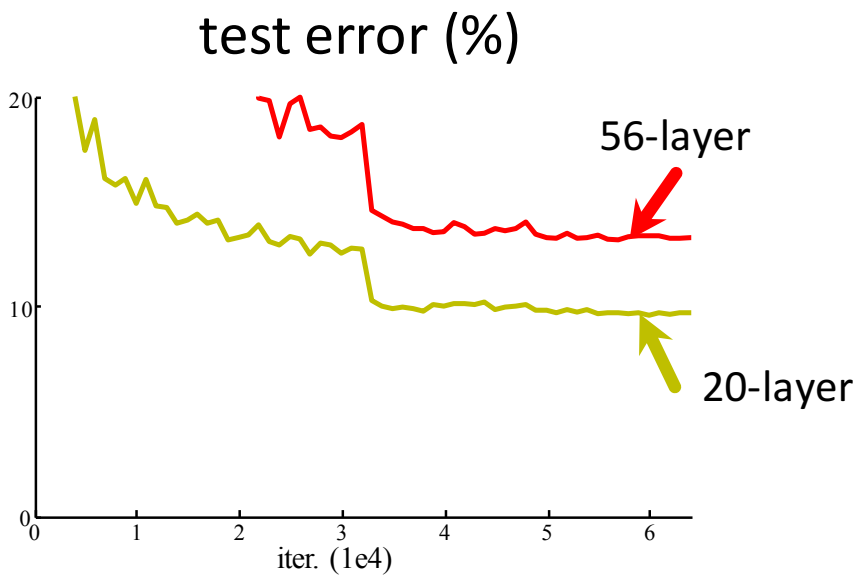
Very Deep Network



Simply Stacking Layer

CIFAR-10





He et al. 2016

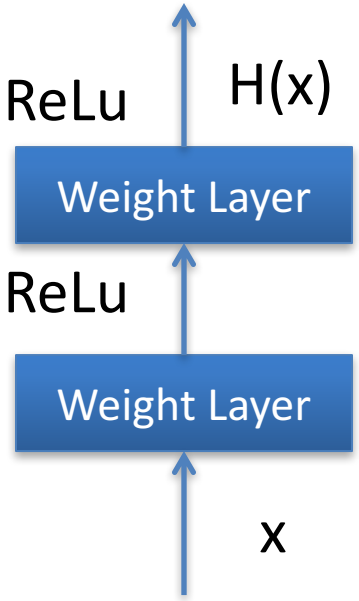
Deep Residual Network

Shortcut connections
- No add. parameters
- No extra computational complexity

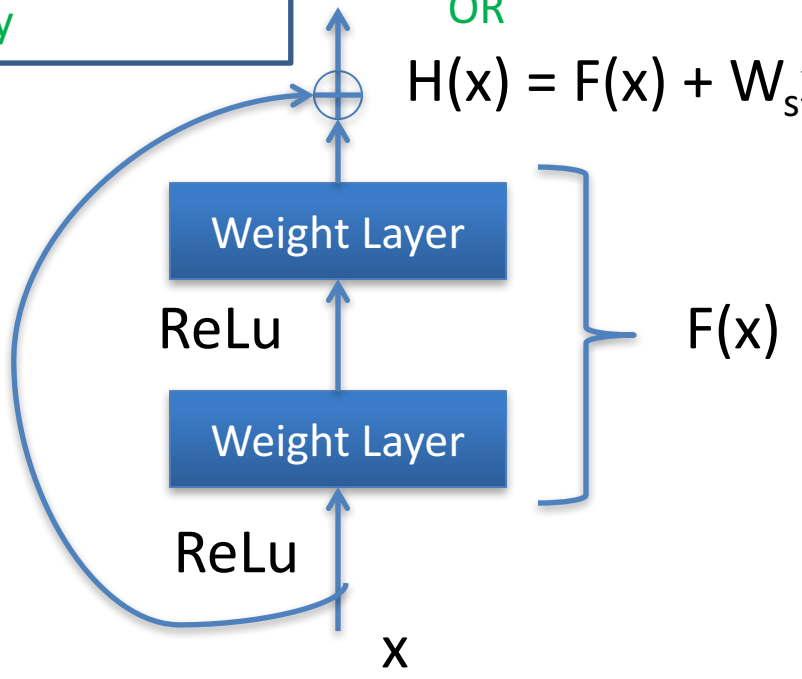
$H(x) = F(x) + x$

OR

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



Plain net



Residual net

Deep Residual Network

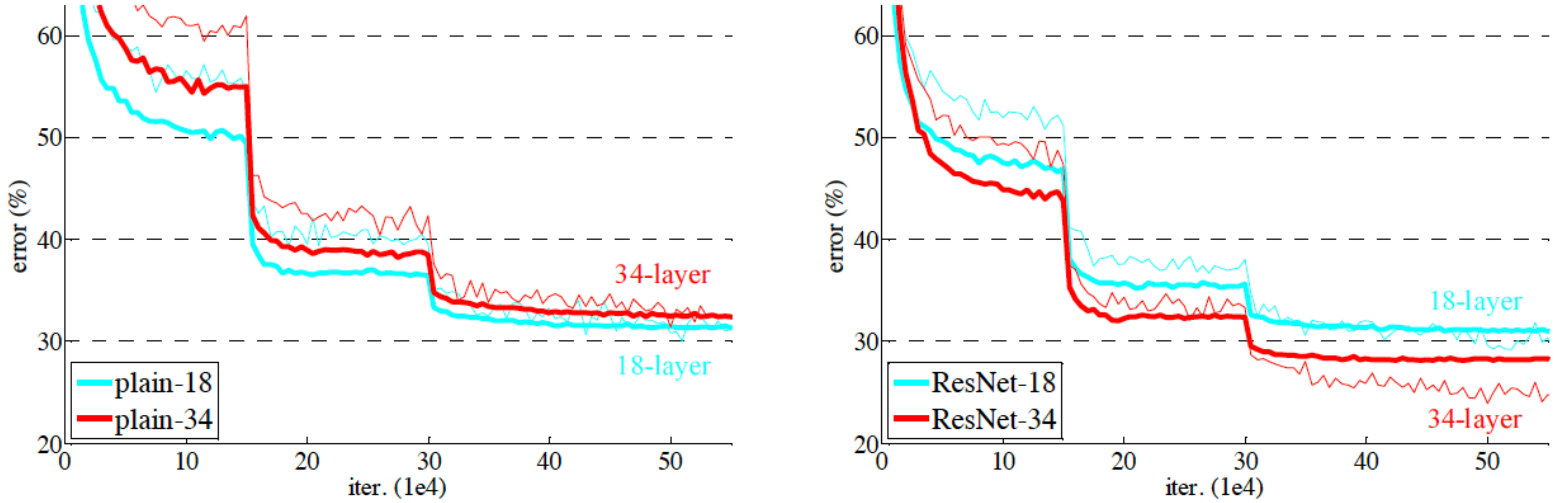


Figure 4. Training on **ImageNet**. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

Deep residual learning for image recognition. He et al. 2016

Deep Residual Network

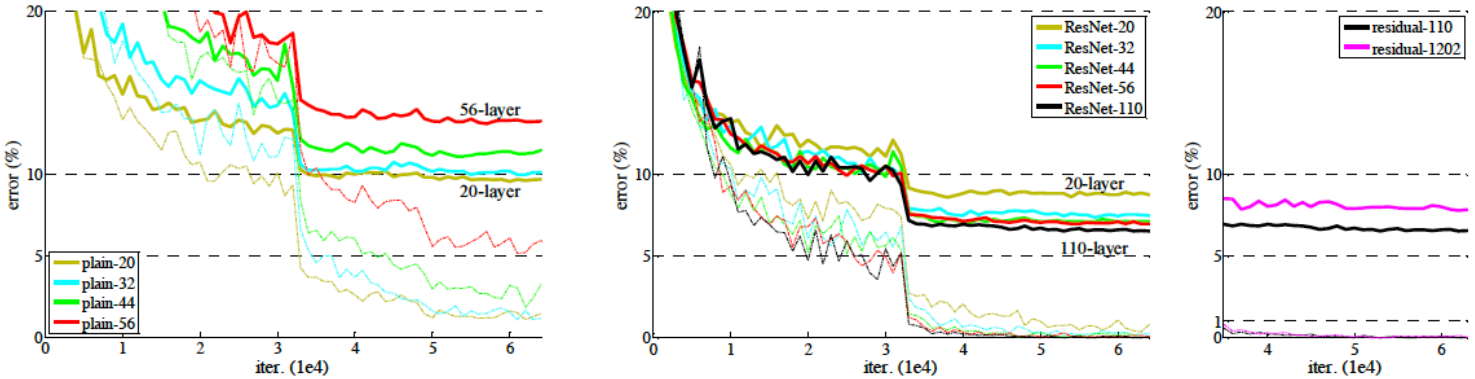


Figure 6. Training on **CIFAR-10**. Dashed lines denote training error, and bold lines denote testing error. **Left**: plain networks. The error of plain-110 is higher than 60% and not displayed. **Middle**: ResNets. **Right**: ResNets with 110 and 1202 layers.

Deep residual learning for image recognition. He et al. 2016

Very Deep CNN for Speech Recognition

- Microsoft ASR system 2017
- This system achieve a 5.1% WER on the 2000 Switchboard evaluation set

Models	Devset WER	Test WER
BLSTM	11.5	8.3
ResNet	10.0	8.2
LACE	11.2	8.1
CNN-BLSTM	11.3	8.4
BLSTM+ResNet+LACE	9.8	7.2
BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2

Very Deep CNN for Text Classification

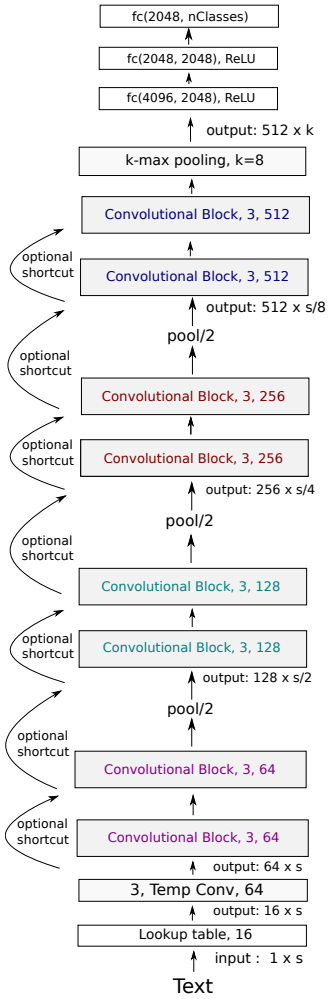


Figure 1: VDCNN architecture.

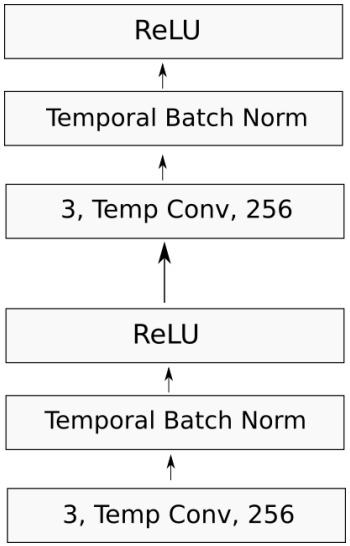


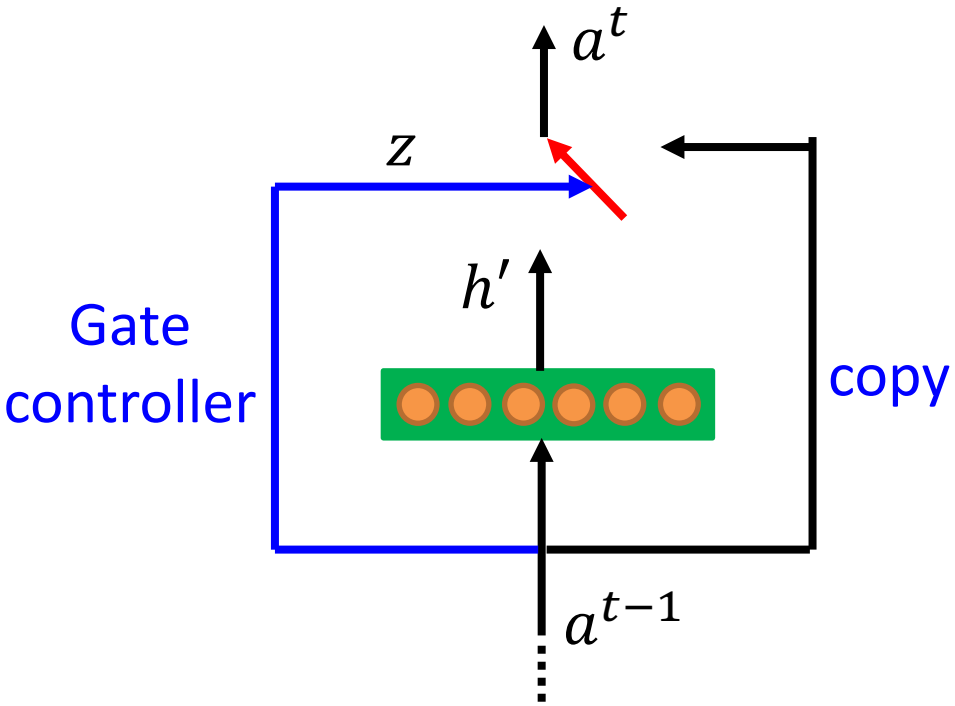
Figure 2: Convolutional block.

Very deep convolutional networks for text classification – Facebook AI, 2017

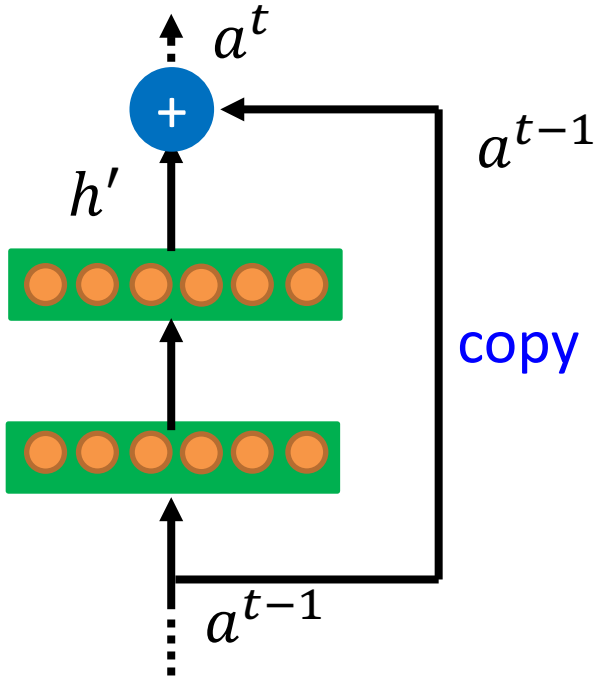
Highway Network

$h' = f(Wa^{t-1})$ $z = \sigma(W'a^{t-1})$ $a^t = z \odot a^{t-1} + (1 - z) \odot h$

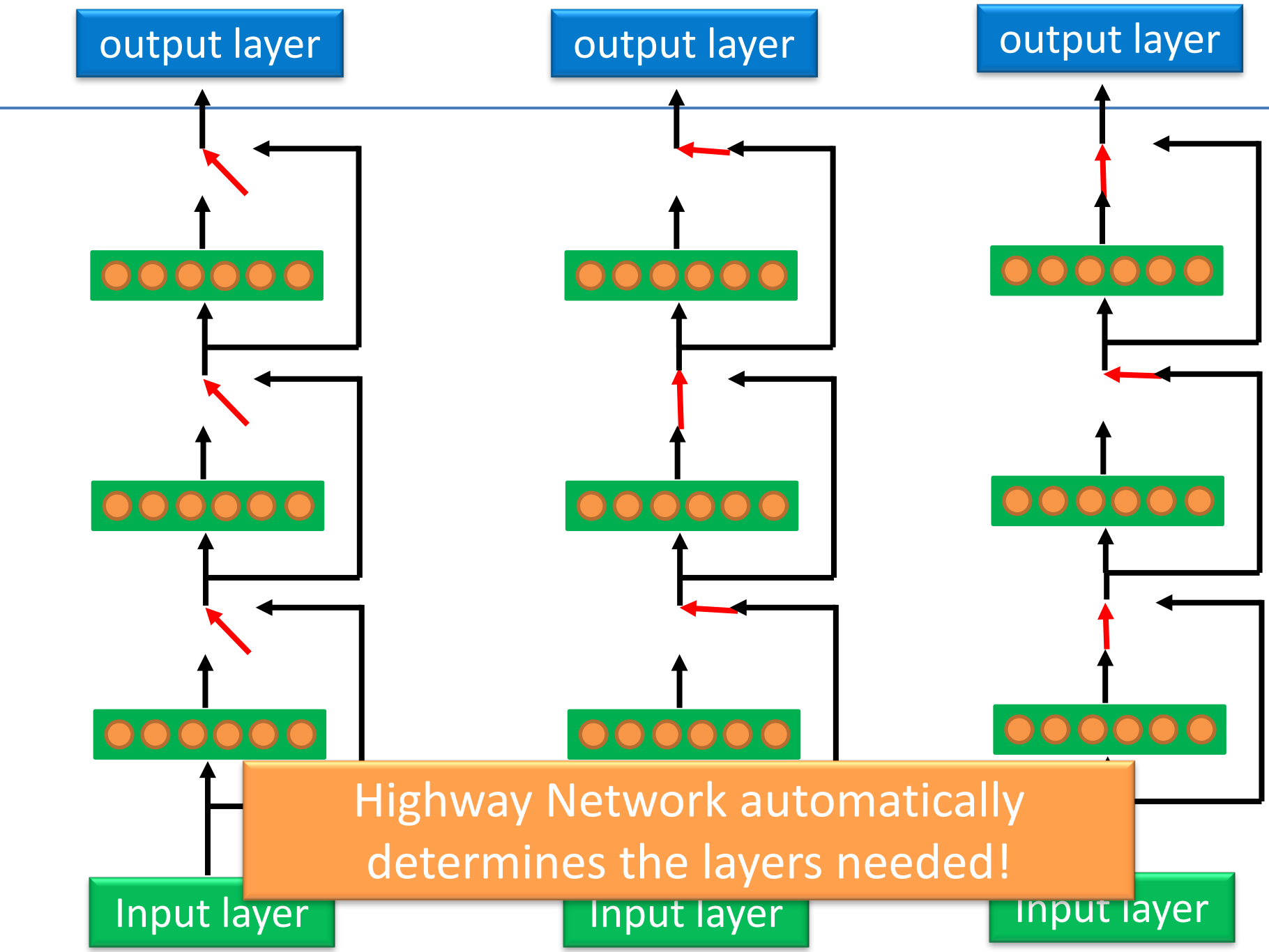
- **Highway Network**



- **Residual Network**



Training Very Deep Networks, 2015


output layer
output layer
output layer
Highway Network automatically determines the layers needed!
Input layer
Input layer
Input layer