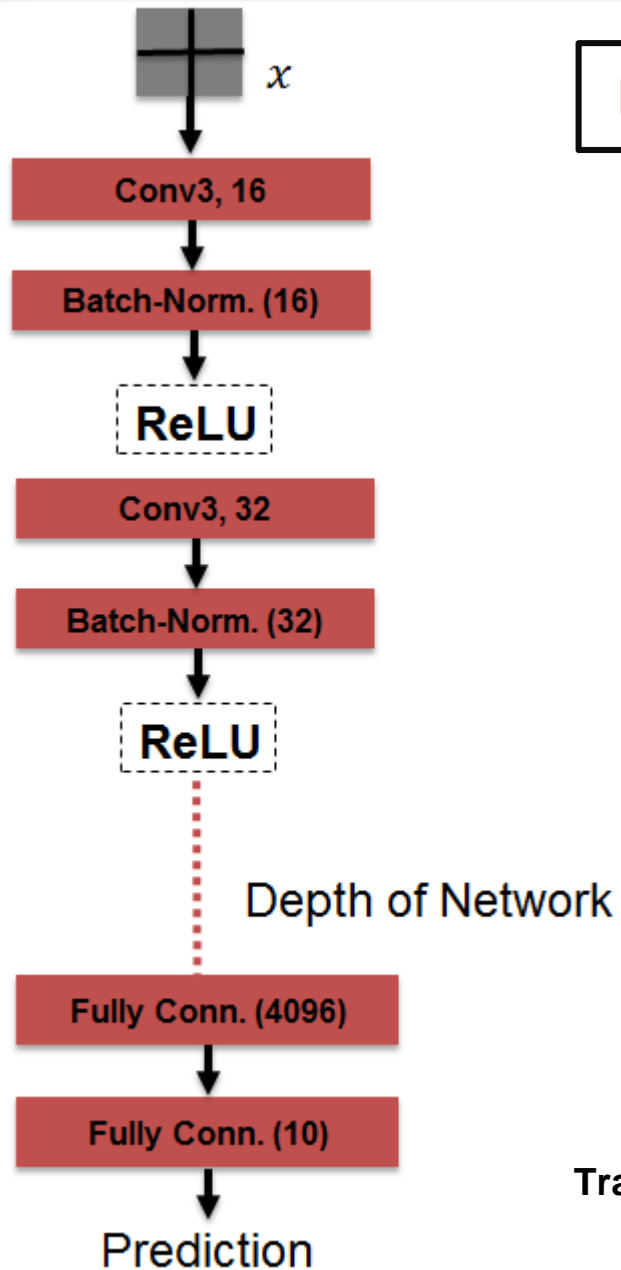# IE 598: Final Project
Deep Residual Learning for Image Recognition

Team Members: Anish Shenoy
Dhruv Gelda
Joshua Schiller
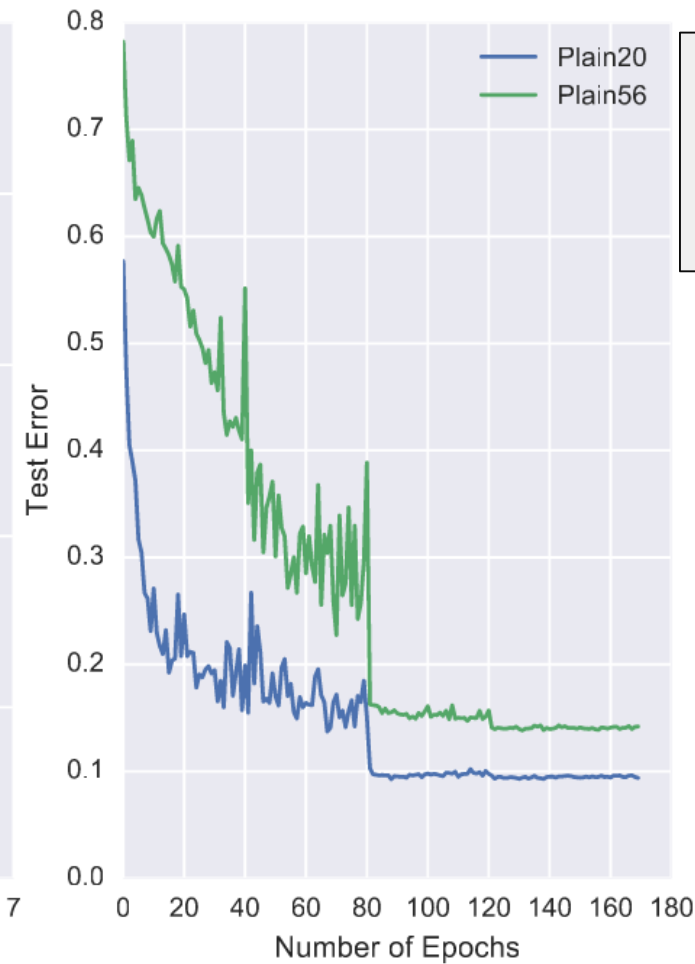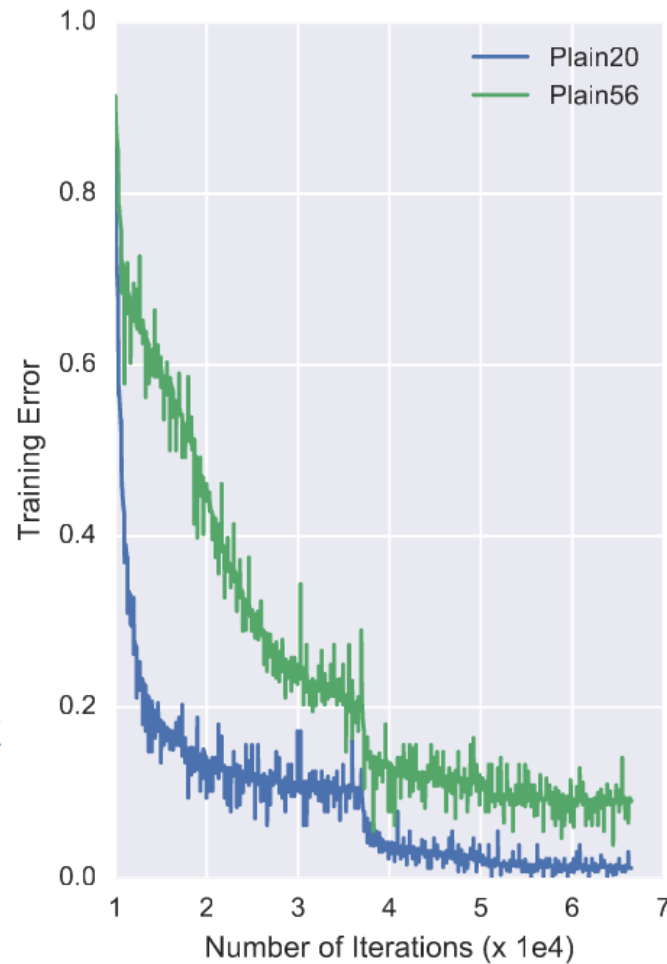**Pramod Srinivasan**

University of Illinois at Urbana-Champaign

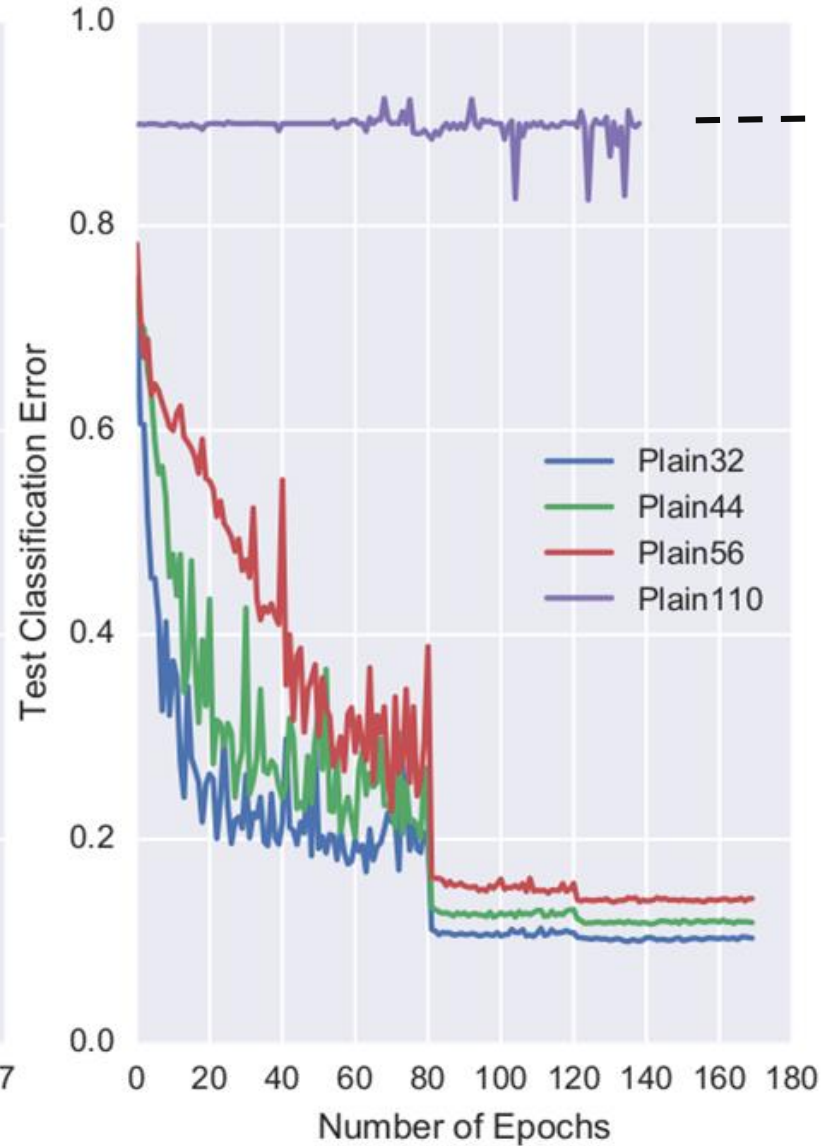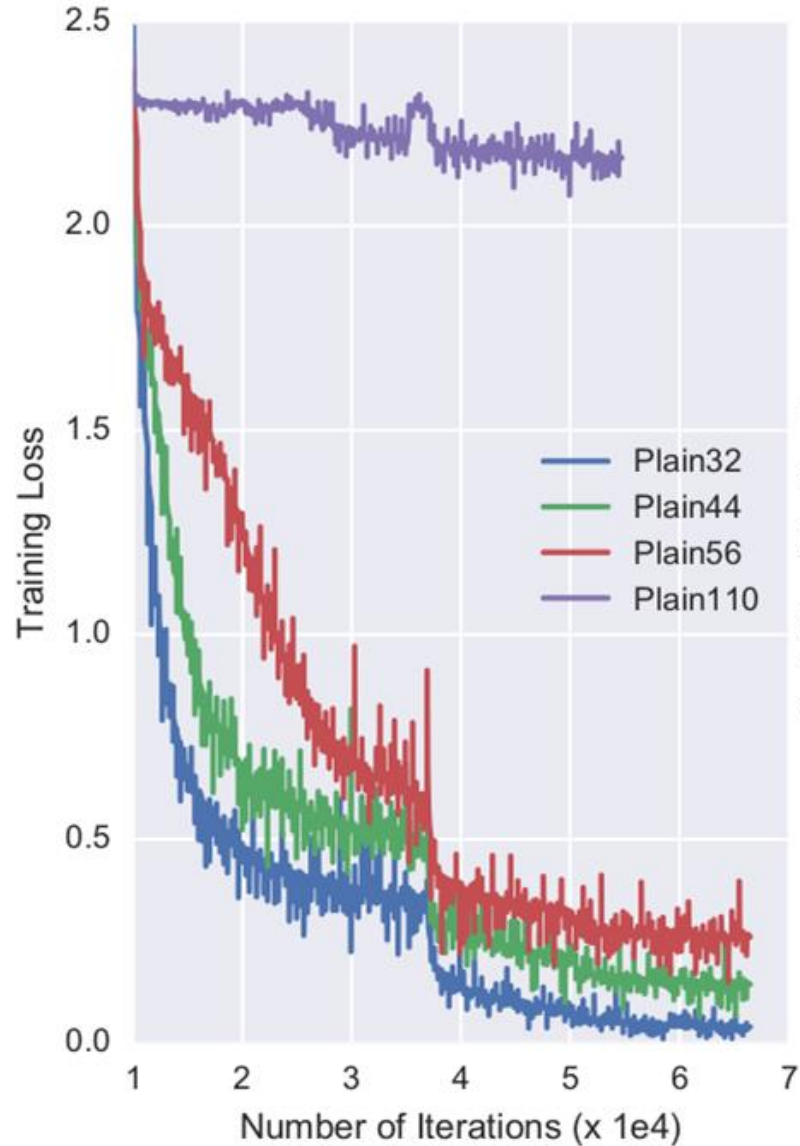**Is learning better models equivalent to stacking more layers ?**

Vanishing / exploding gradients ✗

Overfitting ✗

Training error (left) and **Test error** (right) on CIFAR - 10 with 20 and 56 layer networks.

Degradation of 110 layer network

No. of Layers	Test accuracy
20 Layer	90.61
32 Layer	88.22
44 Layer	87.49
56 Layer	85.81
110 Layer	9.98

Plain Network

Residual Network

x

x

Weight Layer

Weight Layer

ReLU

Identity Mapping x

ReLU

Weight Layer

Weight Layer

Shortcut connections perform an identity mapping

$+$

$H(x)$

$H(x) = x + F(x)$
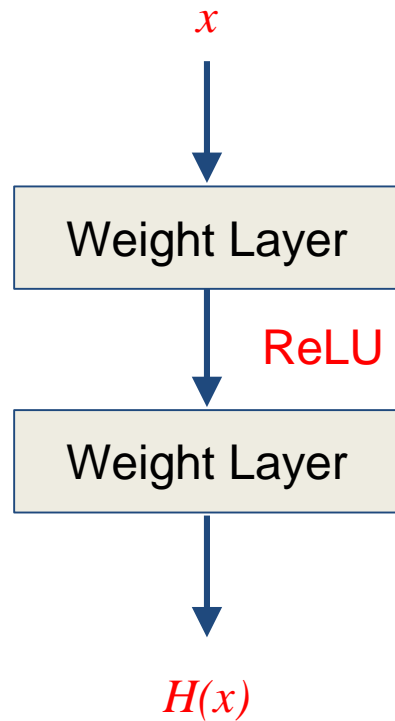
H(x) is the desired mapping that is to be learnt
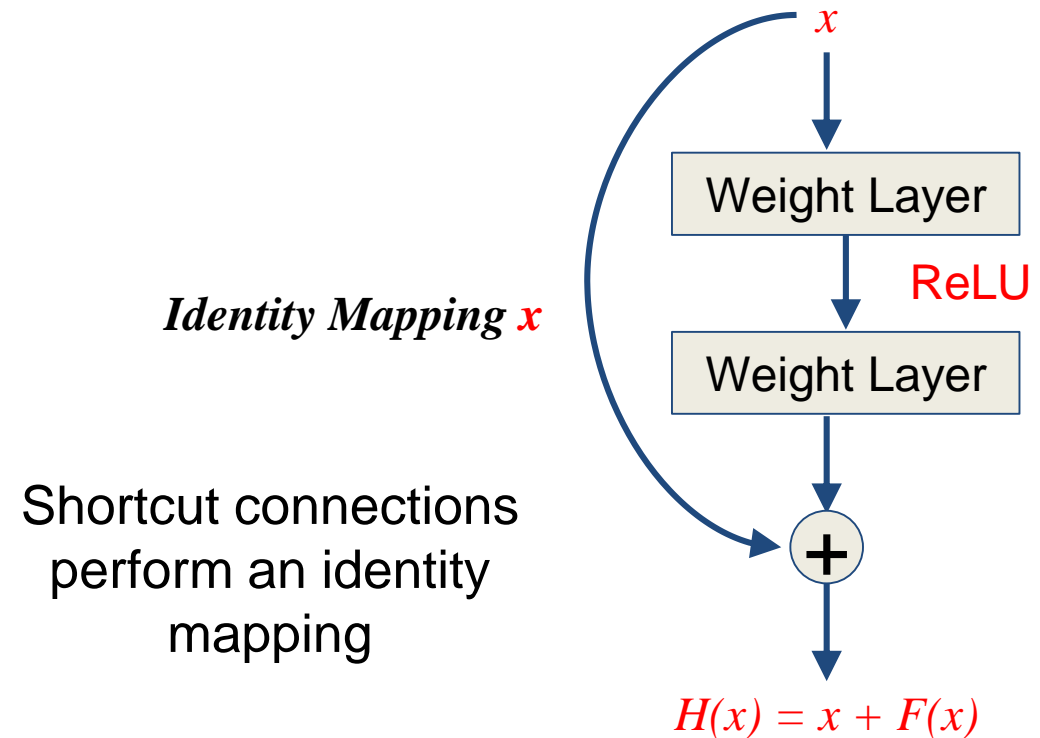
New network learns the residual F(x)
- Network has same depth but with no extra parameter
  - Pushing the residual to zero is easier

Module
(nin channels, nout channels)

x

Conv 3, nout

BN, ReLU

Conv 3, nout

BN

+

ReLU

n	3	5	7	9	18
Number of layers	20	32	44	56	110

Block
(nin channels, nout channels)
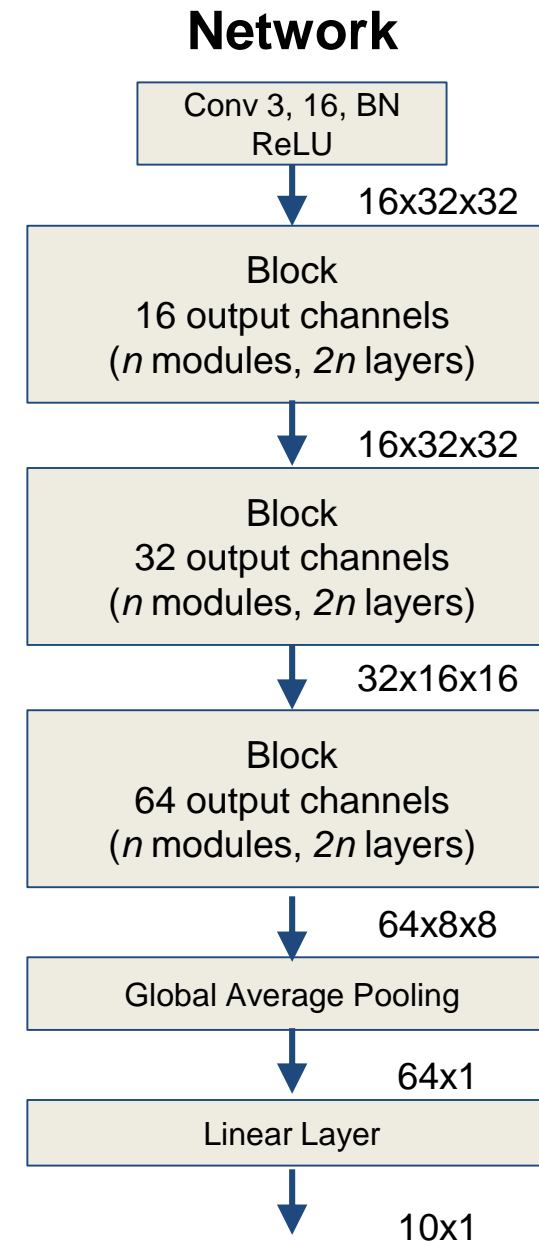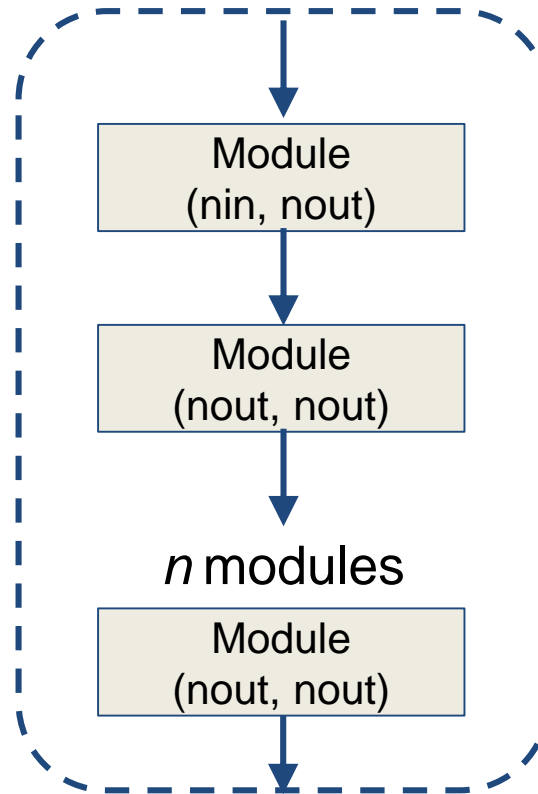
Module
(nin, nout)

Module
(nout, nout)

n modules

Module
(nout, nout)

Network

Conv 3, 16, BN
ReLU

16x32x32

Block
16 output channels
(n modules, $2n$ layers)

16x32x32

Block
32 output channels
(n modules, $2n$ layers)

32x16x16

Block
64 output channels
(n modules, $2n$ layers)

64x8x8

Global Average Pooling

64x1

Linear Layer

10x1

6n+2
Layers
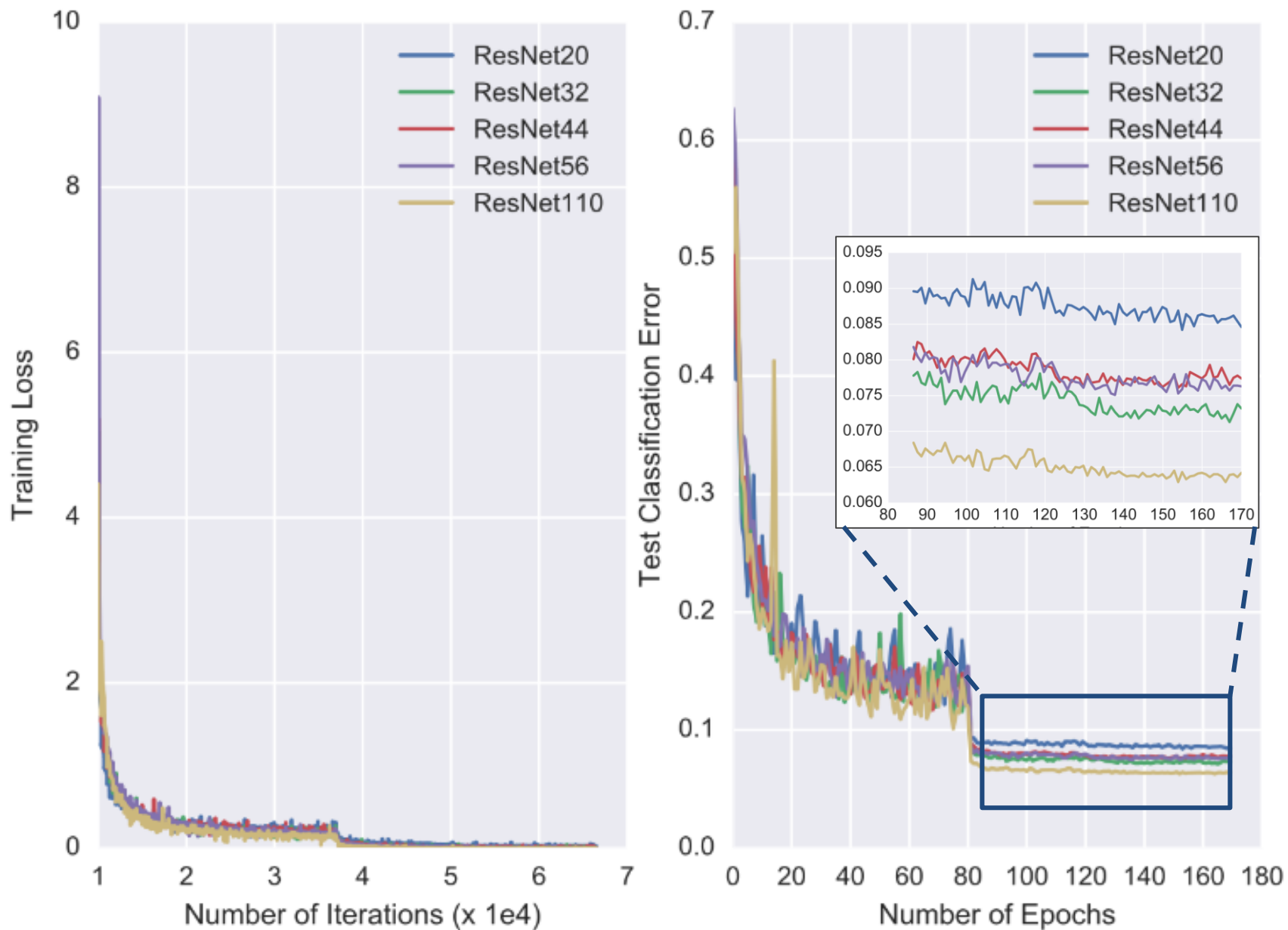
Architecture	Test accuracies (paper)	Test accuracies (our)
20 Layer	91.25	91.53
32 Layer	92.49	92.66
44 Layer	92.83	92.2
56 Layer	93.03	92.30
110 Layer	93.57	93.59

Exploring Beyond
Simple Residual Networks

Motivation
Is there any benefit to fitting residual layers to a residual?

Procedure
Shortcut paths added to bypass blocks of modules

Results
- Shallower and deeper networks perform best
- Residuals of residuals appear to cause overfitting of the model

Potential Reasons
- Shallower models have less adjustable parameters
- Longer shortcuts might have less effect on the model

x

3 x 32 x 32

Conv3, 16, BN, ReLU

16 x 32 x 32

Block
16 output channels
(*n* modules, 2*n* layers)

+ ReLU

16 x 32 x 32

Block
32 output channels
(*n* modules, 2*n* layers)

+ ReLU

32 x 16 x 16

Block
64 output channels
(*n* modules, 2*n* layers)

+ ReLU

64 x 8 x 8

Global Average Pooling

64 x 1

Linear Layer

10 x 1

# Layers	Acc.
20 Layer	90.05
32 Layer	88.32
44 Layer	88.82
56 Layer	86.03
110 Layer	90.89

Motivation
Is there any benefit to fitting residual layers to a residual?
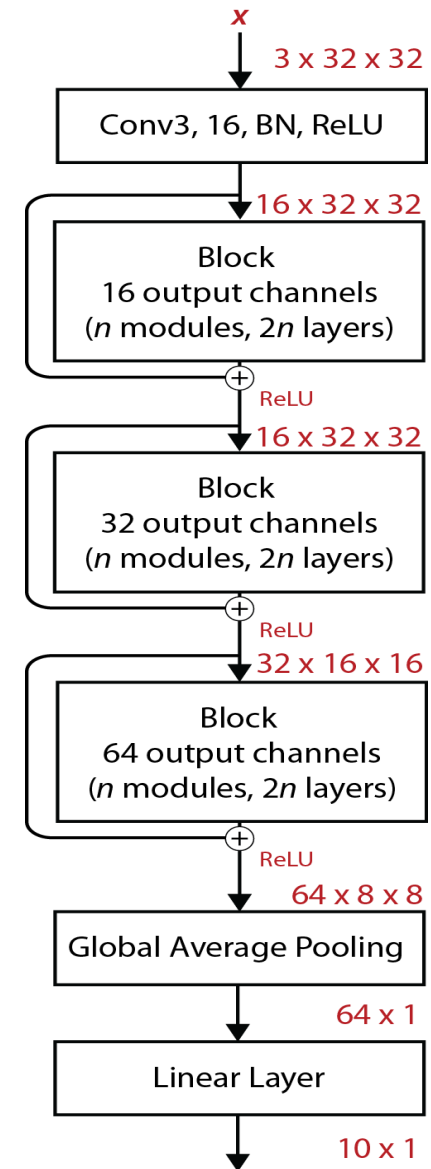
Procedure
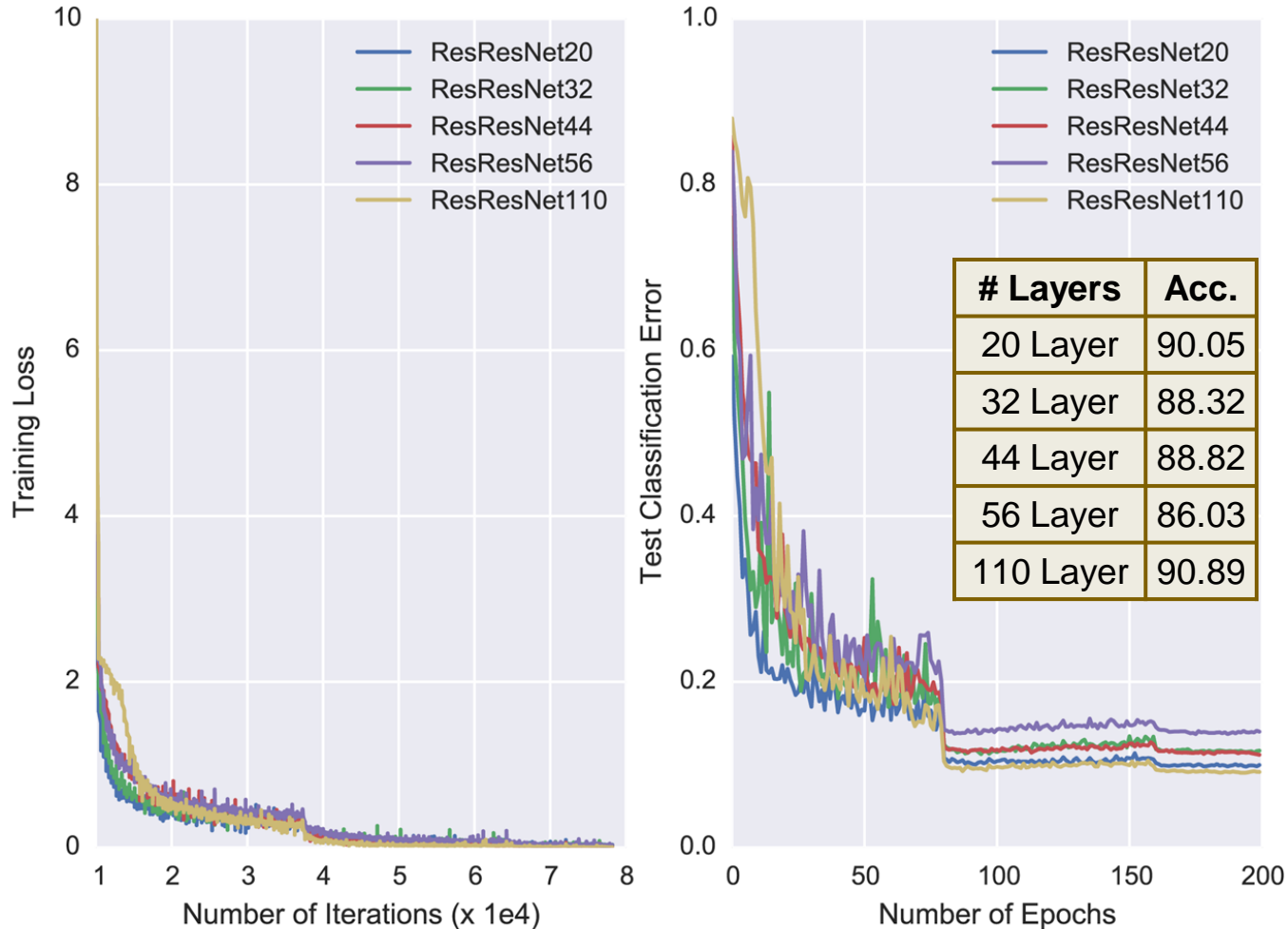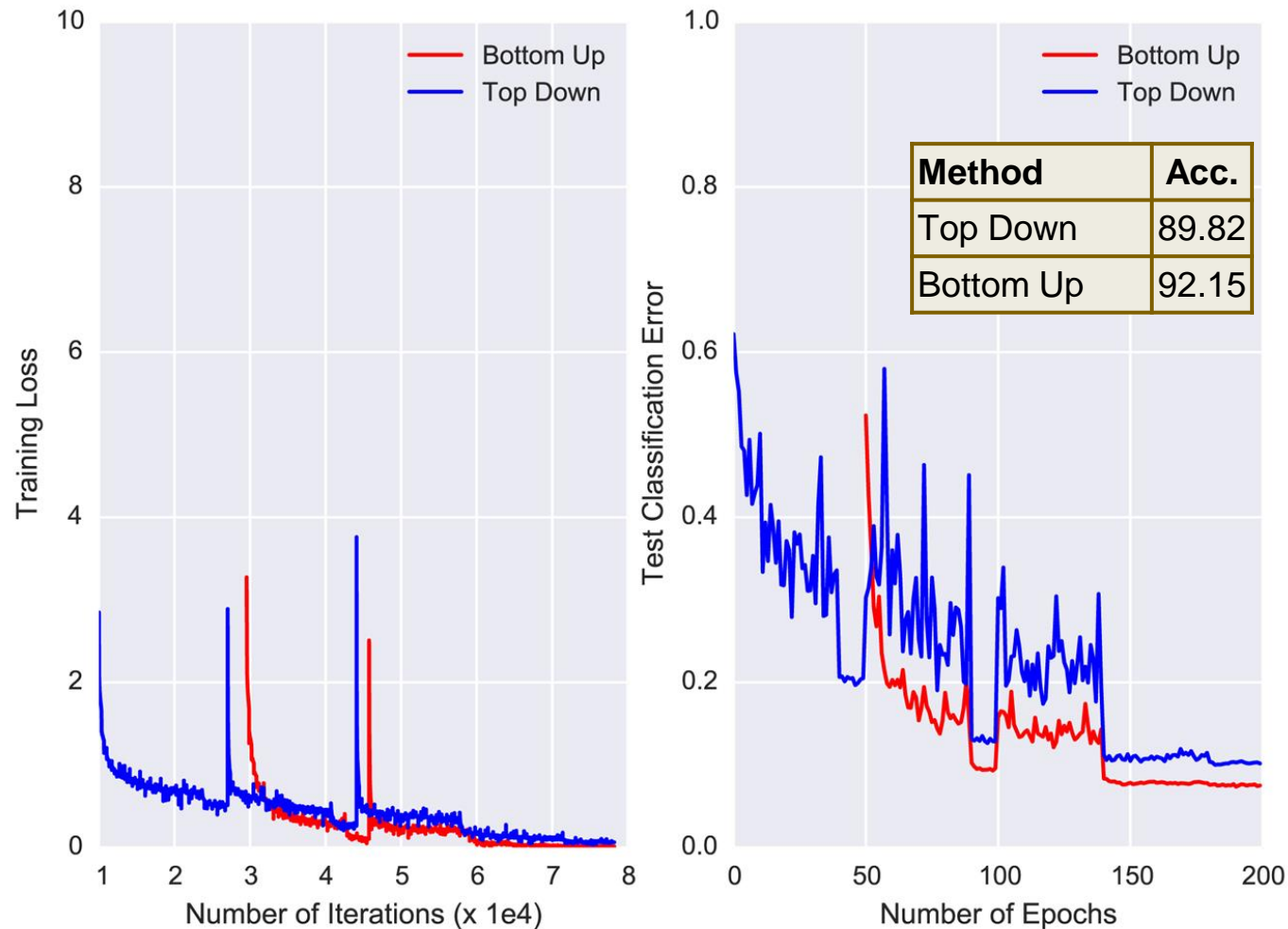Shortcut paths added to bypass blocks of modules

Results
- Shallower and deeper networks perform best
- Residuals of residuals appear to cause overfitting of the model

Potential Reasons
- Shallower models have less adjustable parameters
- Longer shortcuts might have less effect on the model

Method	Acc.
Top Down	89.82
Bottom Up	92.15

Motivation
Is it possible to get a better initialization by pretraining a subset of the model layers beforehand?

Methodology
Sequentially added blocks of 5 modules

Results
- Top-down performed worse
- Bottom-up had no effect

Potential Reasons
- Shallower layers might be more plastic than deeper ones

Thank you for listening!

Questions?

[1] He, Kaiming, et al. "Deep residual learning for image recognition." *arXiv preprint arXiv:1512.03385* (2015).

[2] Zhang, Ke, et al. "Residual Networks of Residual Networks: Multilevel Residual Networks." *arXiv preprint arXiv:1608.02908* (2016).

[3] He, Kaiming, et al. "Identity mappings in deep residual networks." *arXiv preprint arXiv:1603.05027* (2016).

[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*(2014).