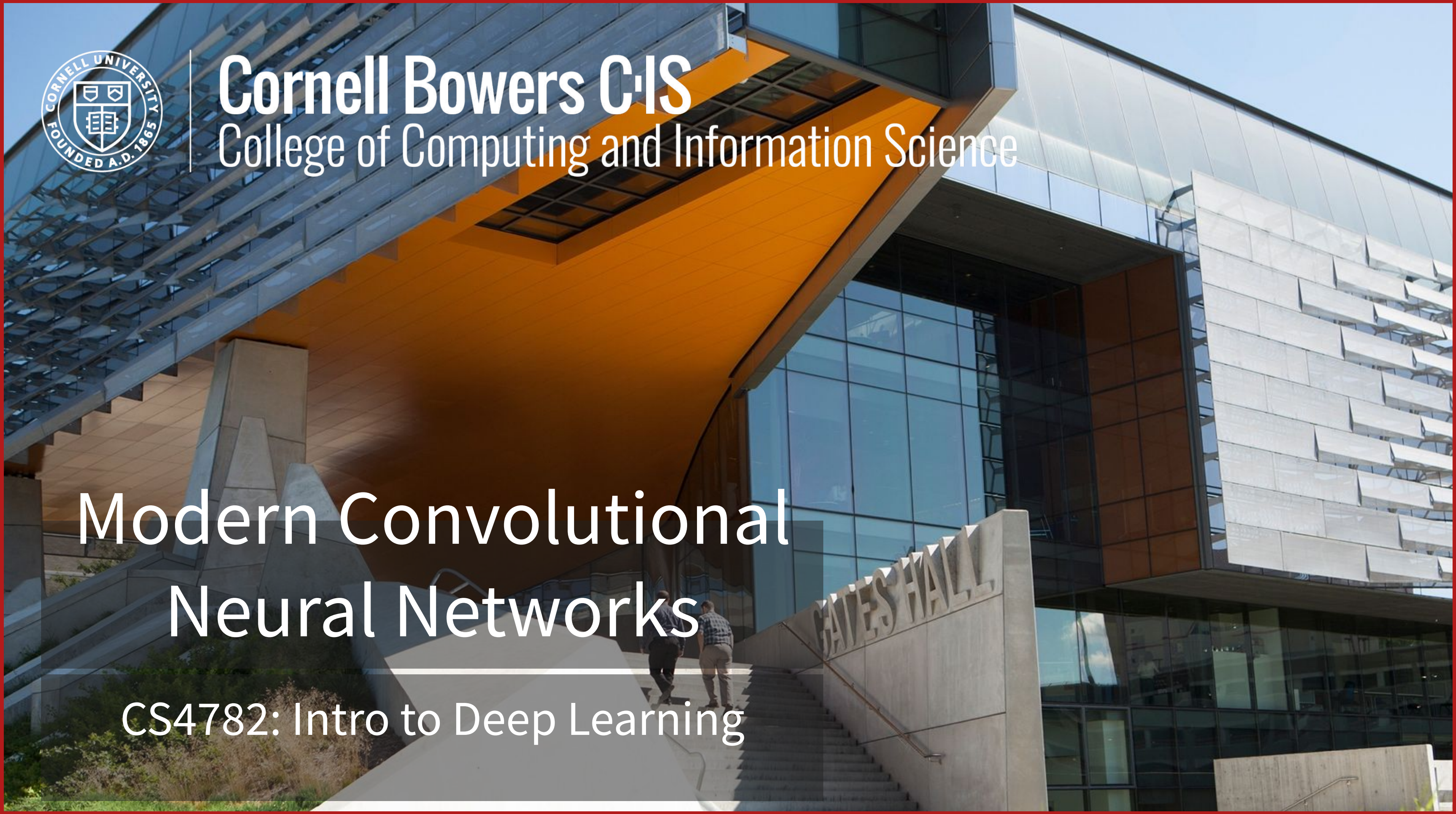


Cornell Bowers CIS
College of Computing and Information Science

Modern Convolutional Neural Networks

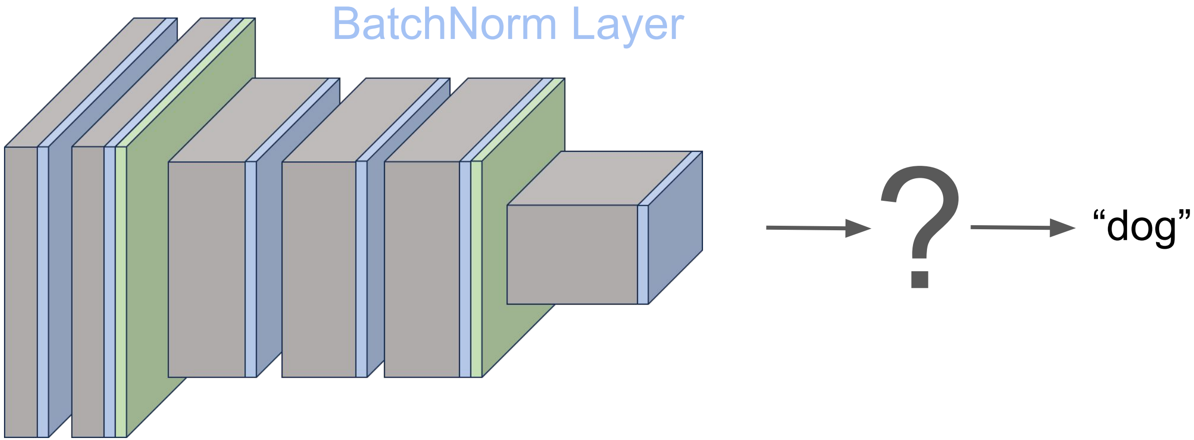
CS4782: Intro to Deep Learning

Cornell Bowers CIS

Review: Convolutional Neural Networks (CNNs)

✅ Convolutions	Maintain spatial relation between pixels Reduce number of parameters through weight sharing
✅ Pooling	Captures key information from across different areas of the feature maps Together with convolutions allows for translational invariance
✅ **BatchNorm**	Increases speed and stability of training

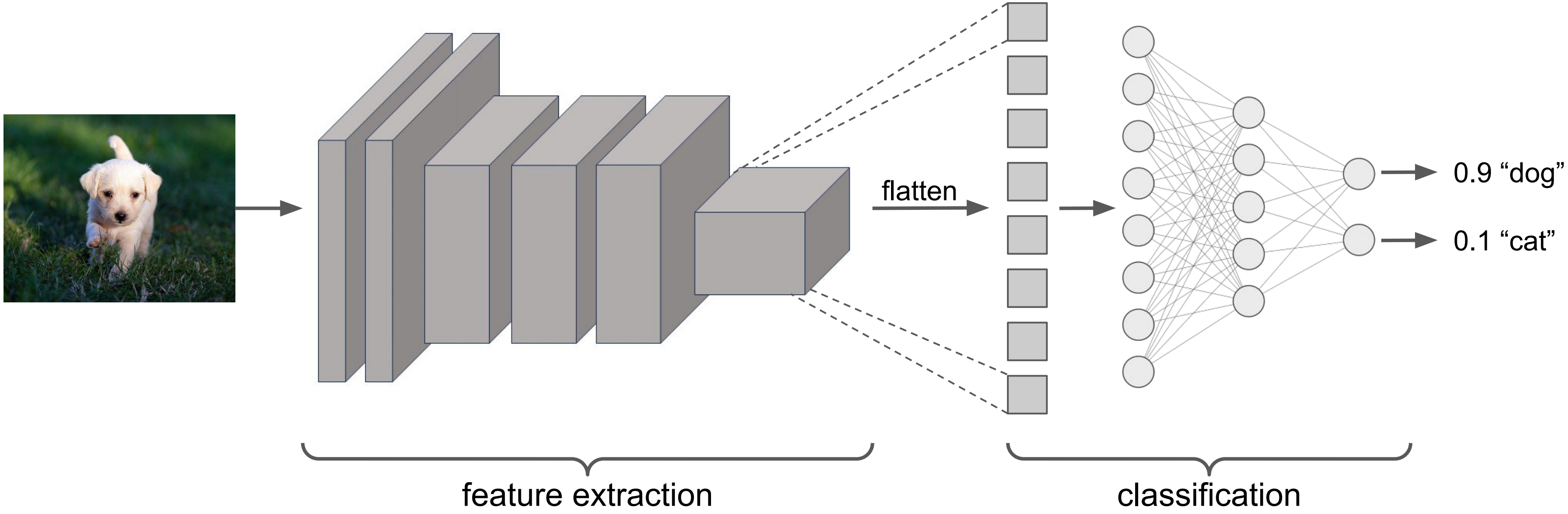

input image



Cornell Bowers C·IS

Image Classification

- Important: Everything is differentiable!
- Can calculate gradient of the loss with backpropagation
    - Train with SGD/Adam/etc.
    - Learn convolutional filters and classification head end-to-end!

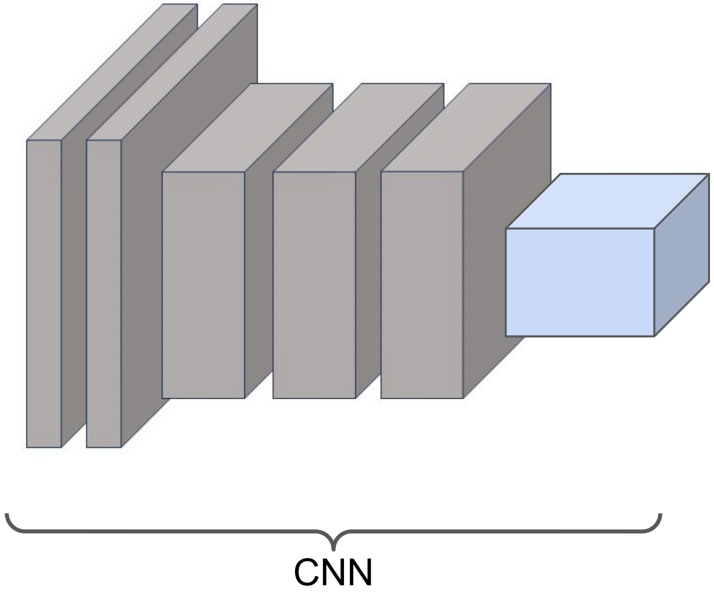


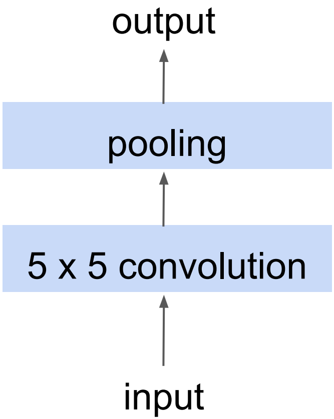
Discuss: Padding

- Given a 5x5 feature map and a 3x3 convolution:
  - How much padding do I need to maintain the spatial size of the feature map (i.e., 5x5)?
- What about when using a 5x5 convolution?

Cornell Bowers C·IS

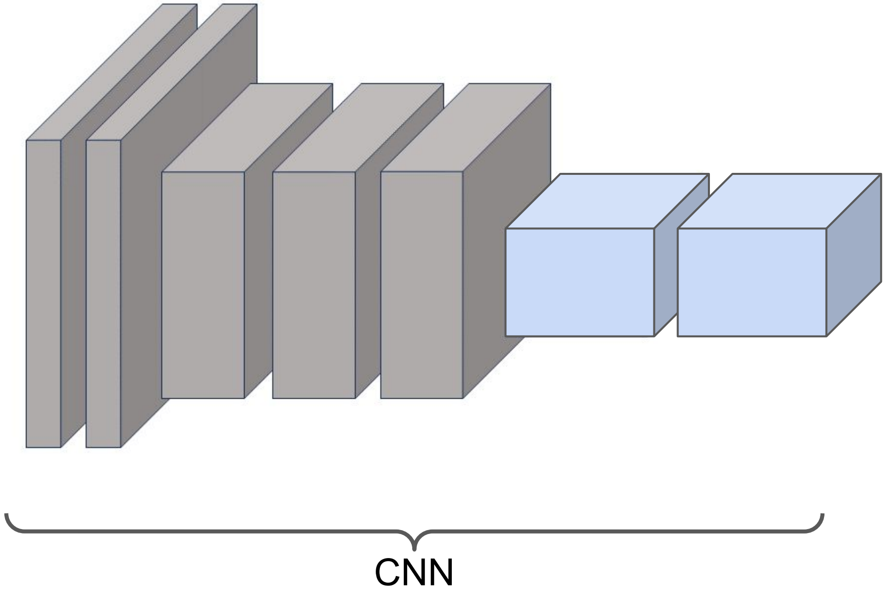
Deeper CNN Architectures

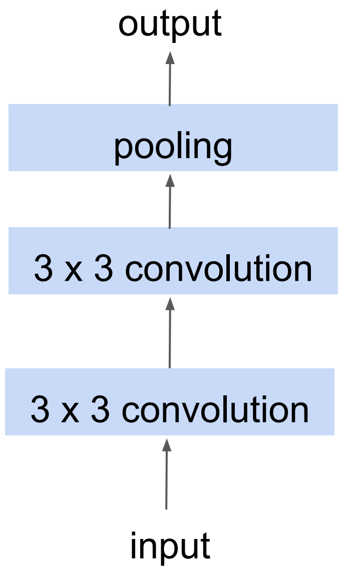




Cornell Bowers C·IS

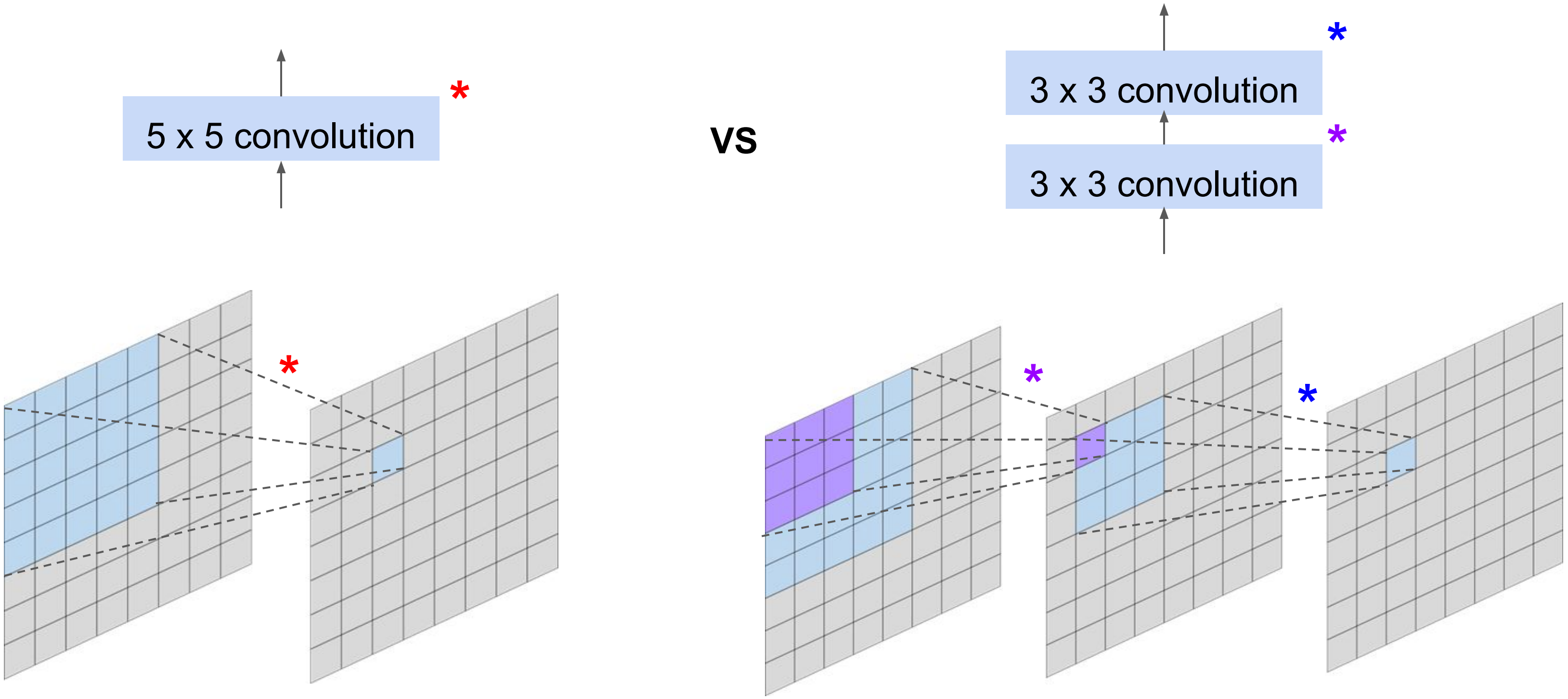
Deeper CNN Architectures



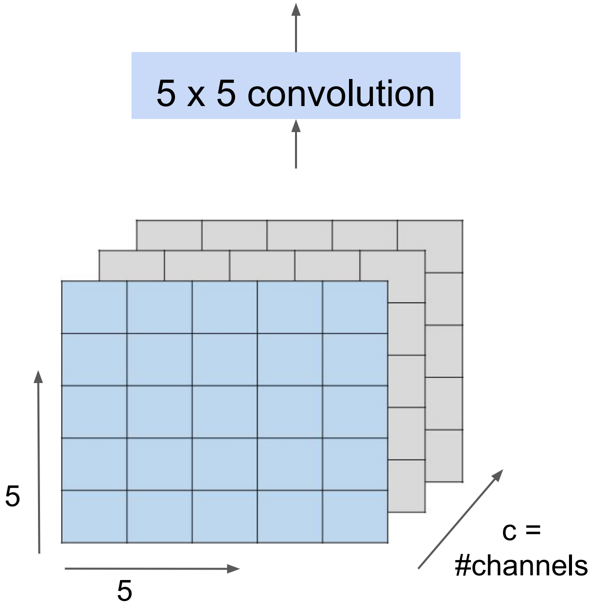


Cornell Bowers C·IS

Deeper CNN Architectures

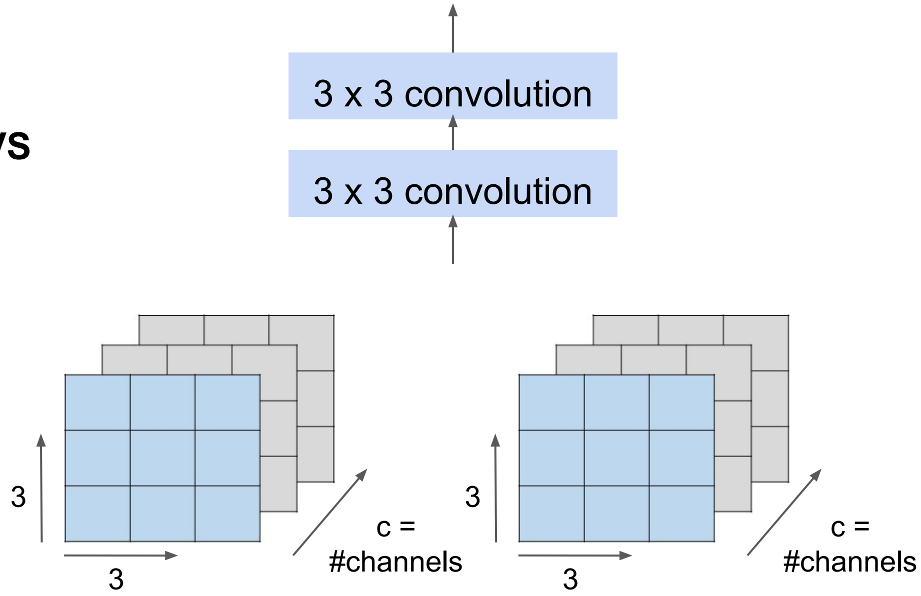


Deeper CNN Architectures



$$5 * 5 * c^2$$
$$= 25c^2 \text{ parameters}$$

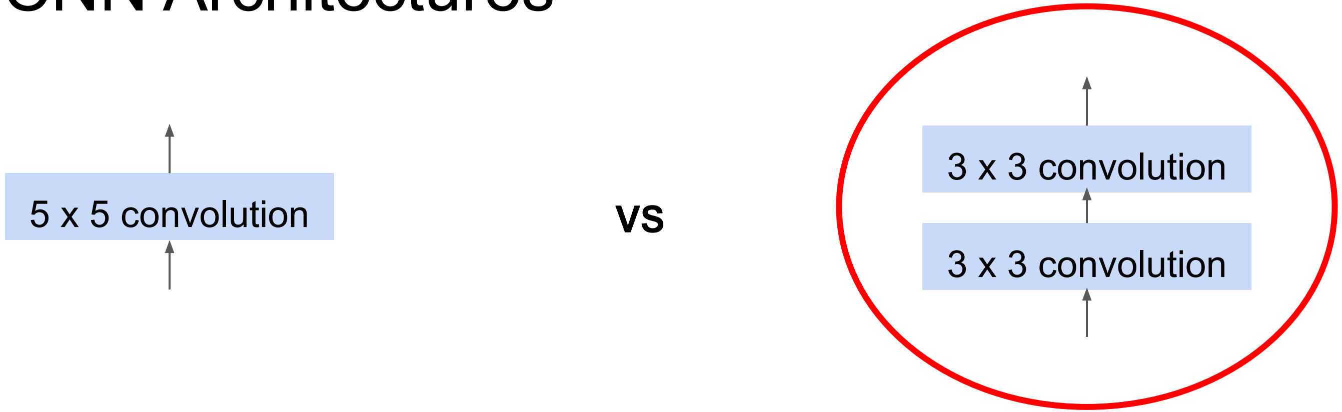
vs



$$2 * 3 * 3 * c^2$$
$$= 18c^2 \text{ parameters}$$

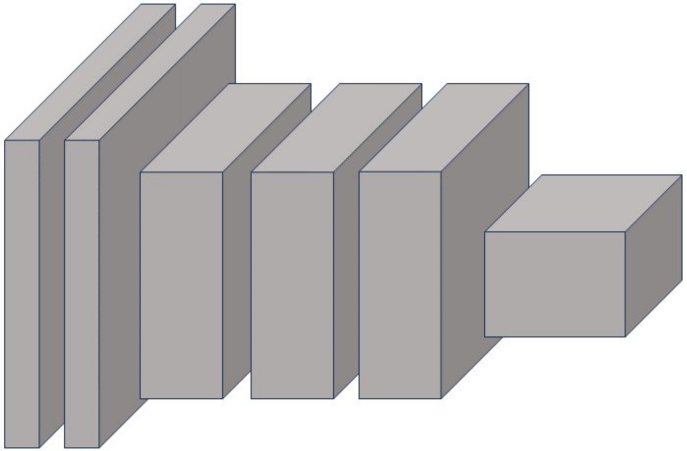
Cornell Bowers C·IS

Deeper CNN Architectures



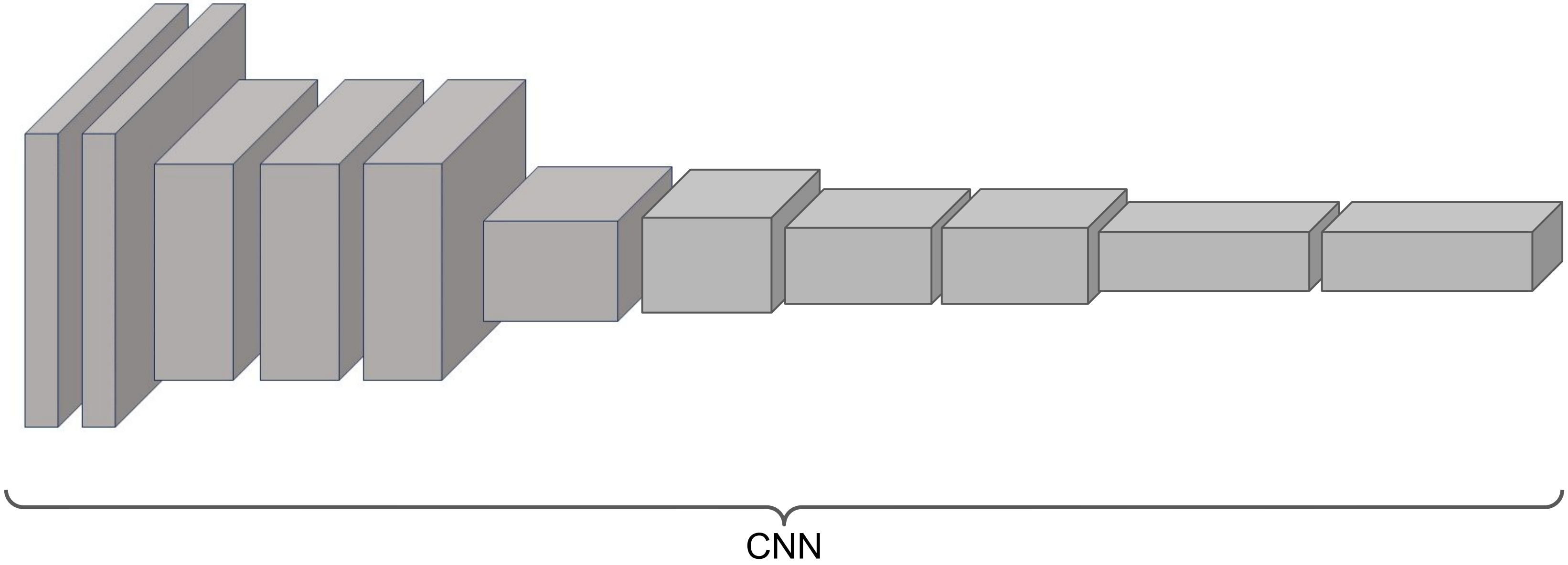
Performed better!

Deeper == better

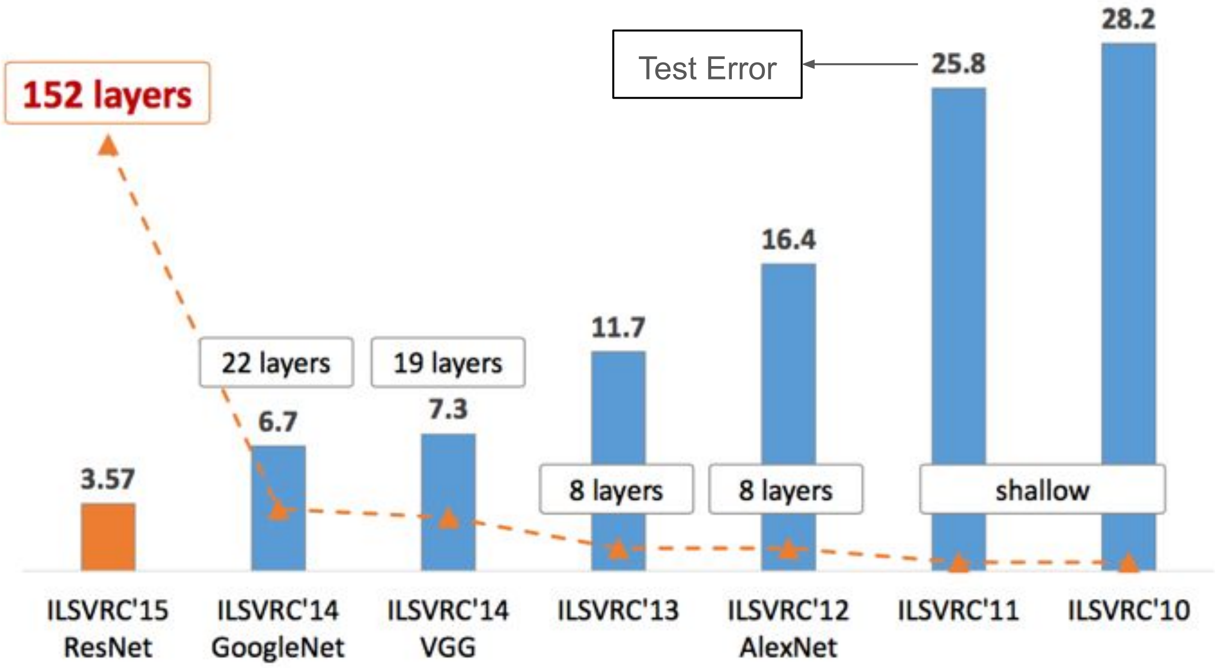


CNN

Deeper == better



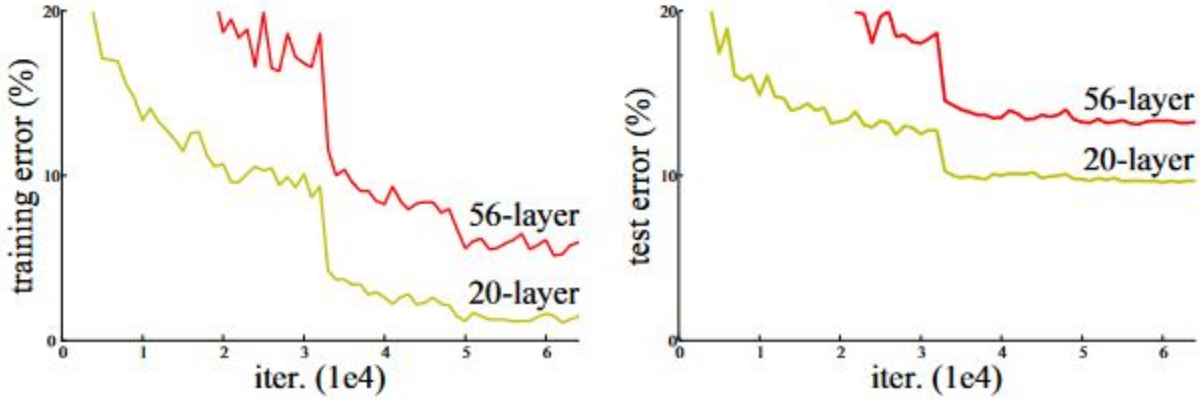
ImageNet Classification Challenge: Deeper == better



[Nguyen, Kien & Fookes, Clinton & Ross, Arun & Sridharan, Sridha. (2017). Iris Recognition with Off-the-Shelf CNN Features: A Deep Learning Perspective. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2784352.]

Cornell Bowers C·IS

Deeper == better?



56 layer CNN has higher training and test error than 20 layer CNN on CIFAR-10 dataset for image classification

[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

Deeper != better

- Long training times

- Vanishing gradient problem
  - Recall backpropagation to update weights

$$\frac{\partial z}{\partial z_i} = \frac{\partial z}{\partial z_{n-1}} \frac{\partial z_{n-1}}{\partial z_{n-2}} \cdots \frac{\partial z_{i+1}}{\partial z_i}$$

  - If each term <<< 1, gradient "vanishes" as the entire multiplication goes towards 0
  - => Weights not updated properly

Cornell Bowers C·IS

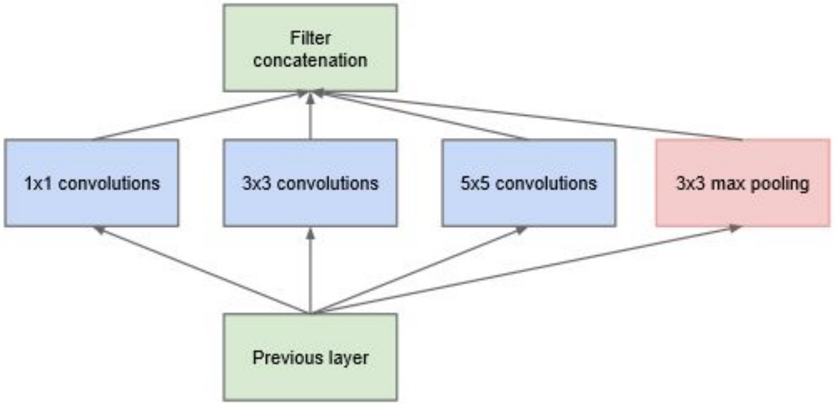
GoogLeNet/Inception Net

Goal: given a fixed computational budget, optimize the depth and width of the network

=> Deeper networks with computational efficiency

[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

Inception Module



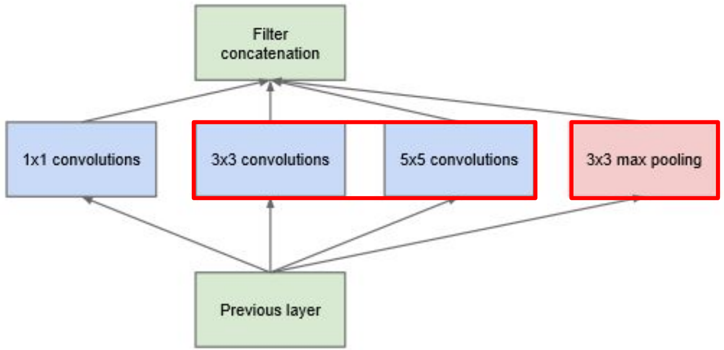
Inception module = main building blocks

[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

Cornell Bowers CIS

Inception Module

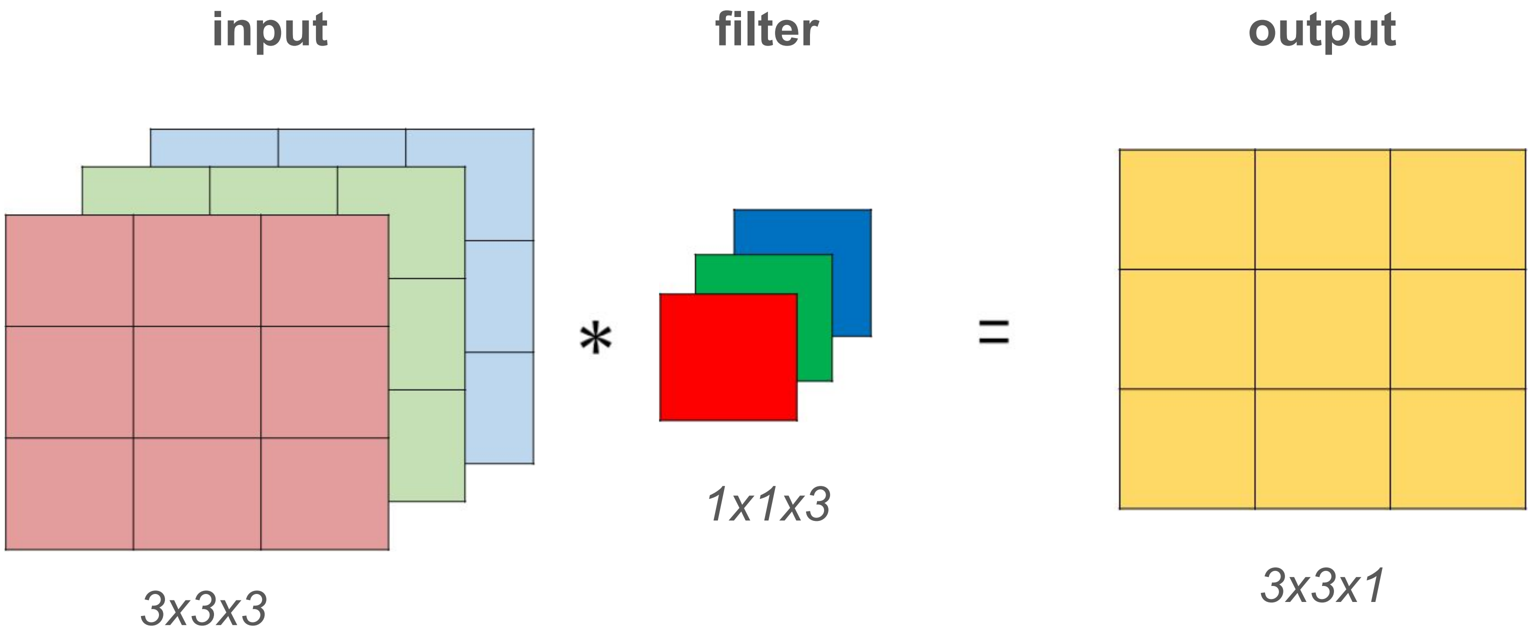
Still expensive!



- 3x3 and 5x5 convolutions have large number of operations
- Output of pooling layer increases the output channel dimension when concatenated

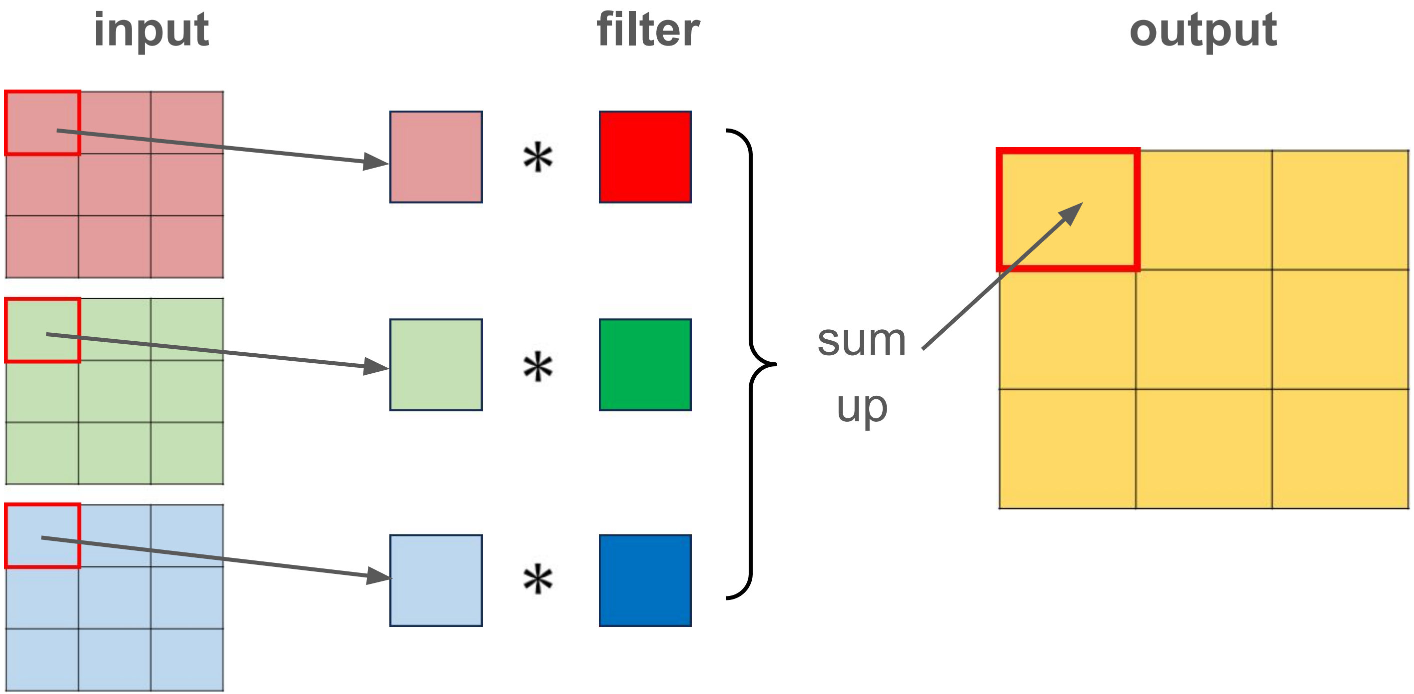
[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

Slight Detour: 1x1 convolutions



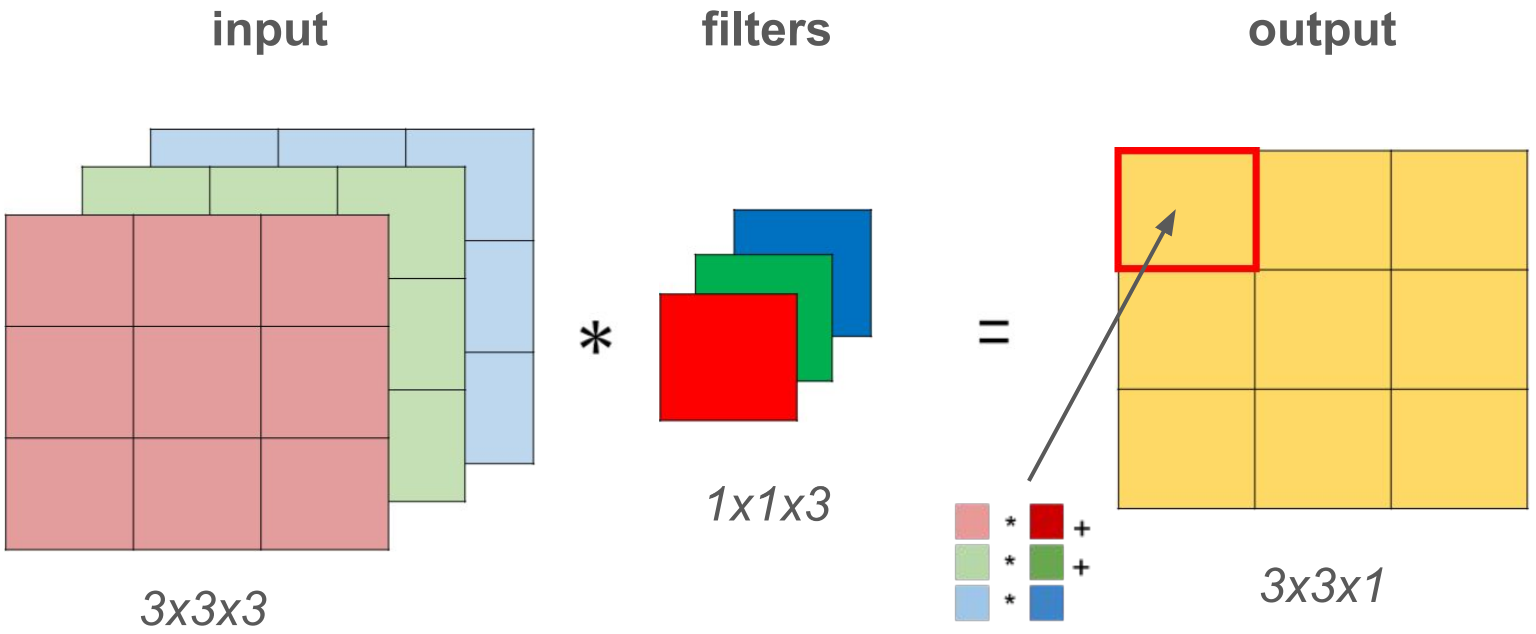
Cornell Bowers C·IS

Slight Detour: 1x1 convolutions



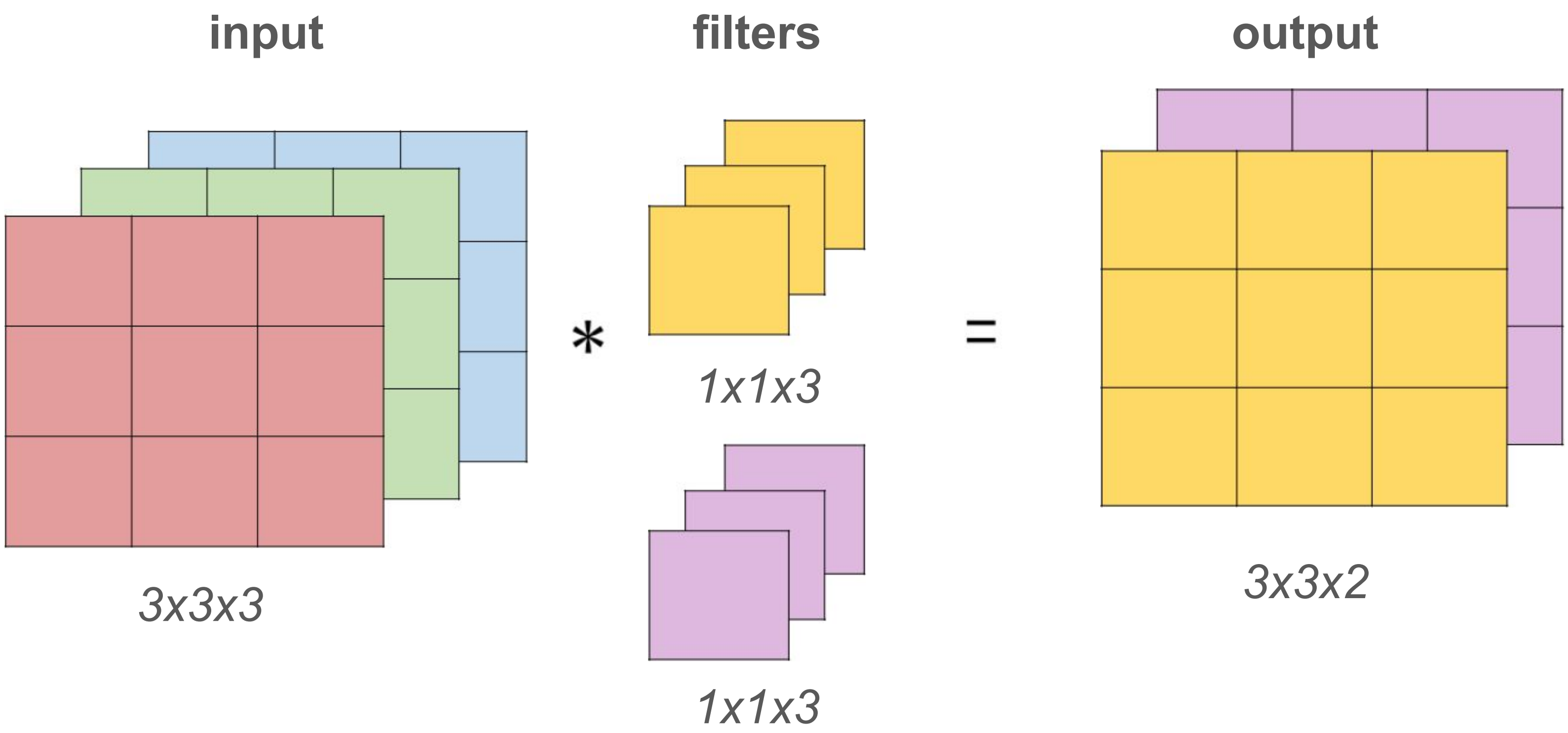
Cornell Bowers C·IS

Slight Detour: 1x1 convolutions

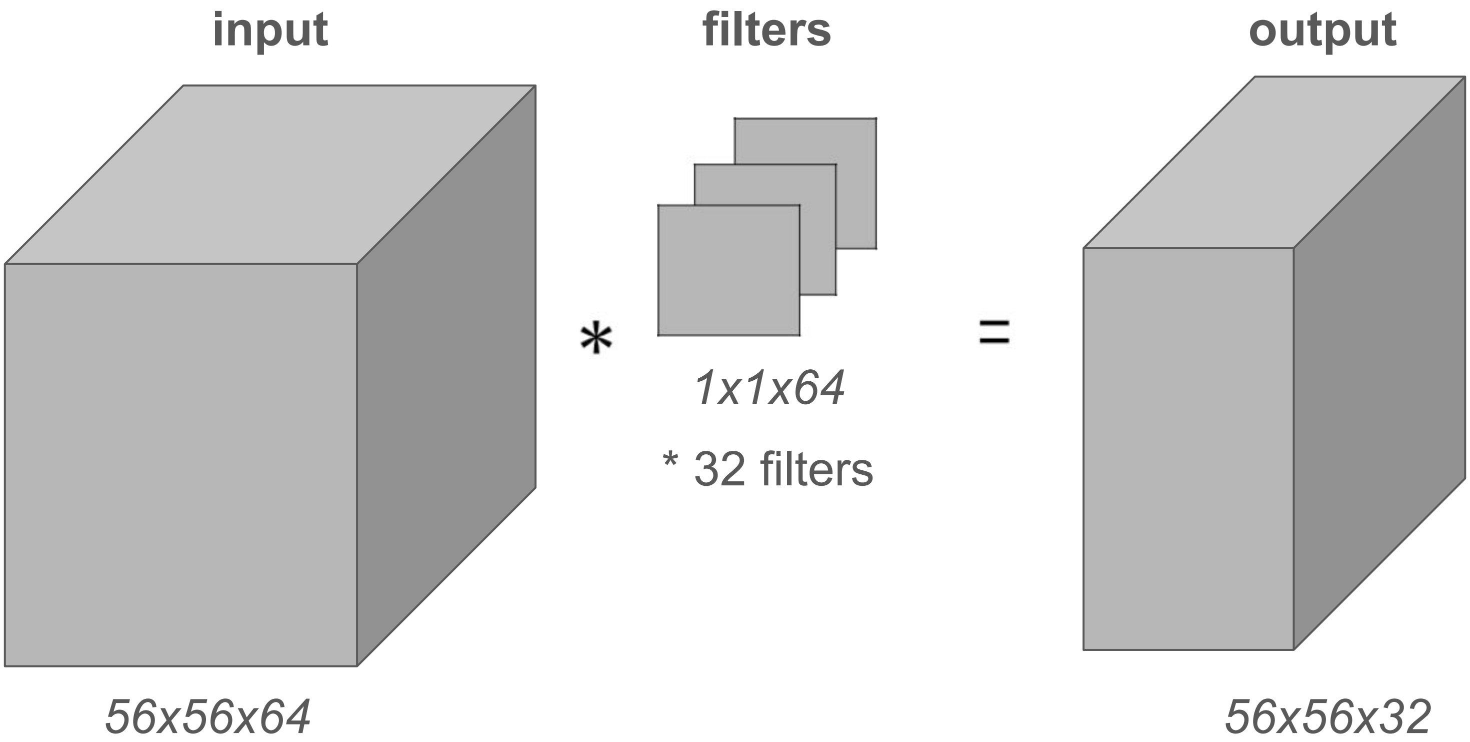


Cornell Bowers CIS

Slight Detour: 1x1 convolutions

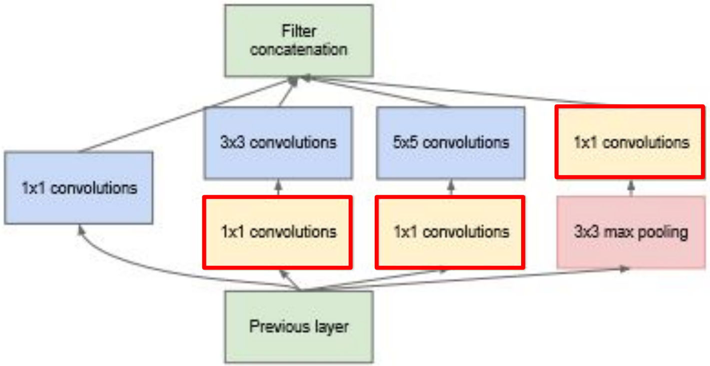


Slight Detour: 1x1 convolutions



Inception Module

Solution: Inception module with dimension reduction



- “Bottleneck” with 1x1 convolutions to reduce dimensions

[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

Discuss: Impact of Dimension Reduction

Assume you have an input feature map with 256 dimensions.

Compare the parameter counts from:

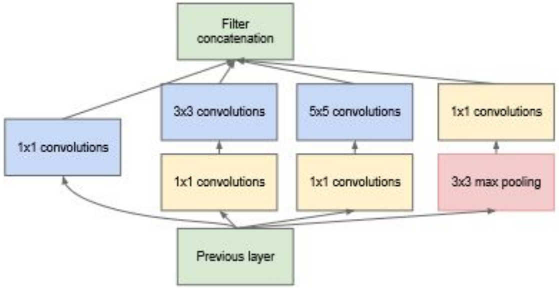
1. 3x3 conv with 256 filters

2. 1x1 conv with 64 filters → 3x3 conv with 64 filters → 1x1 conv with 256 filters

Cornell Bowers C·IS

GoogLeNet Architecture

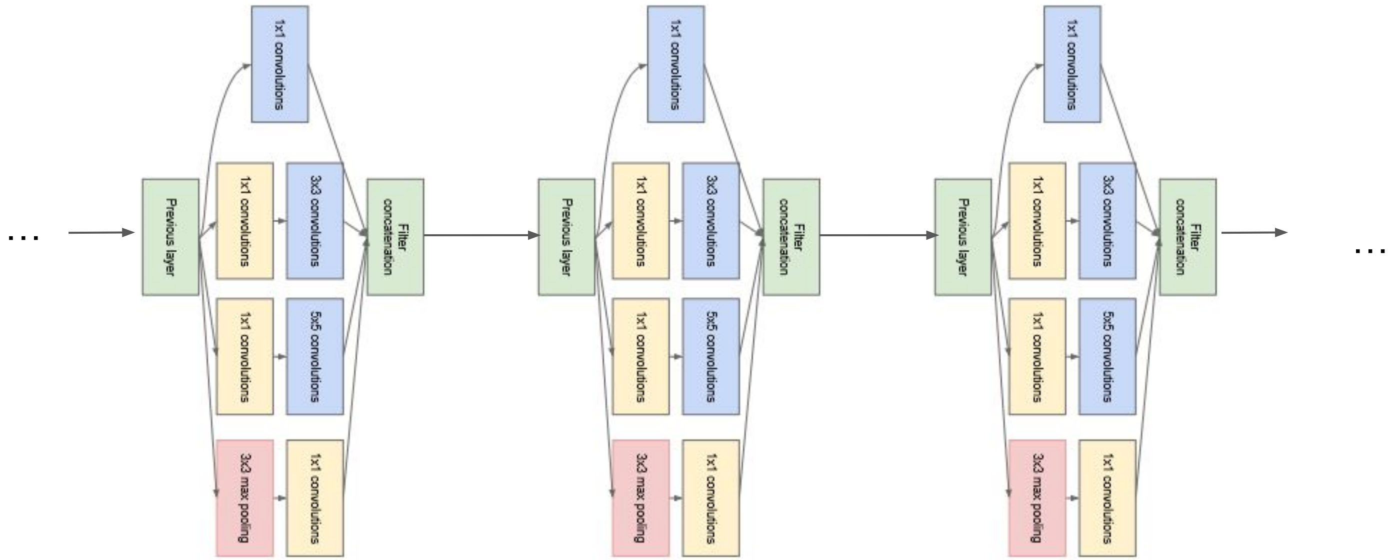
Key idea: stack inception modules together



[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

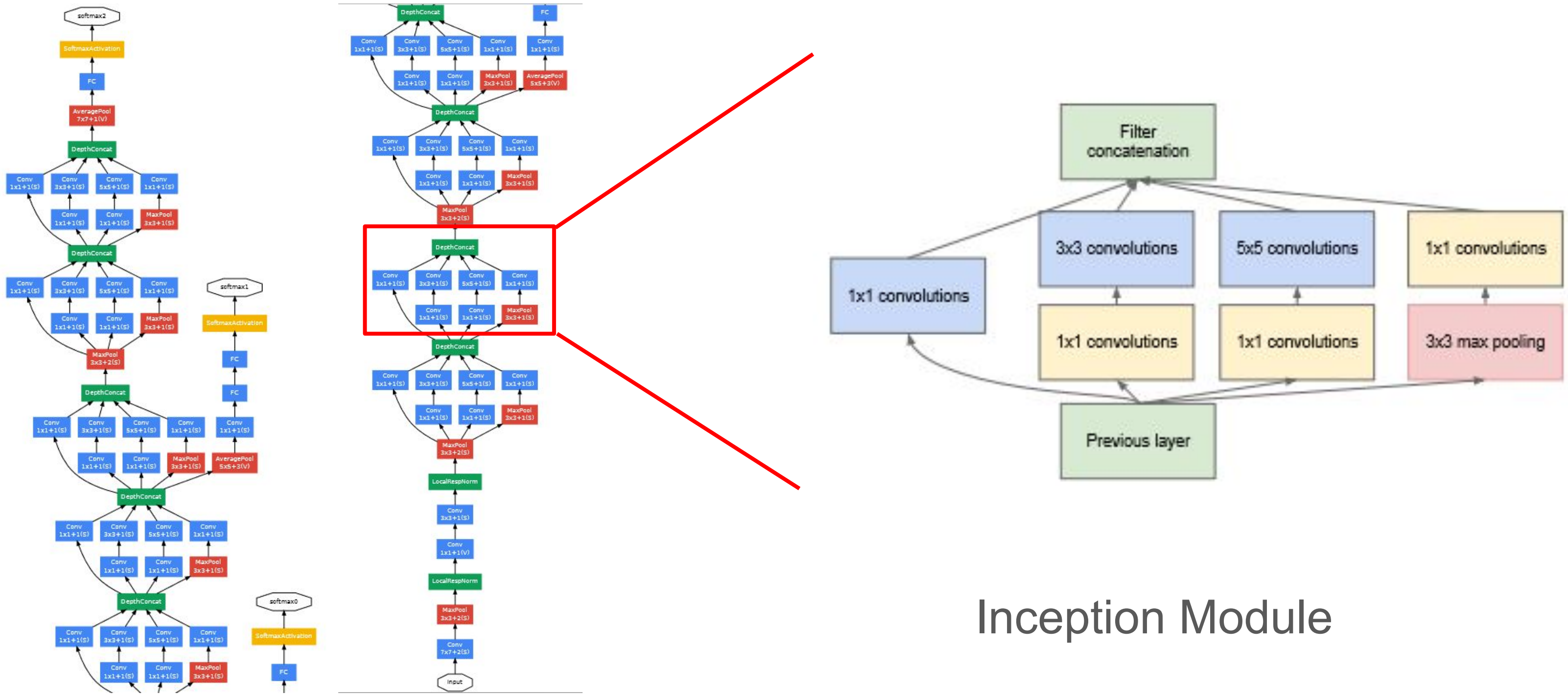
GoogLeNet Architecture

Key idea: stack inception modules together



[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

The Entire GoogLeNet Architecture

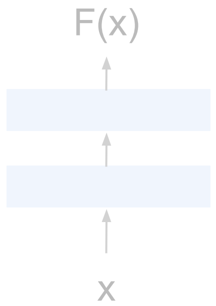


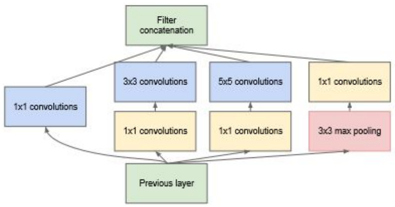
Inception Module

[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

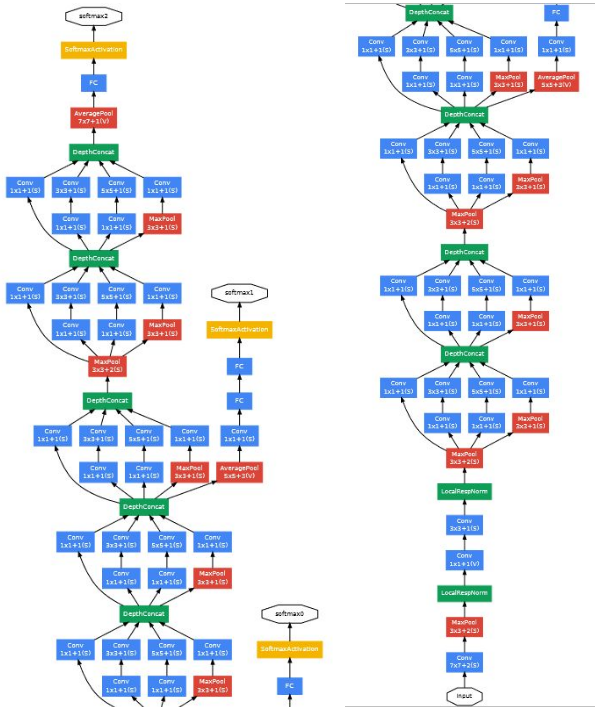
CNN Architectures

"Plain" CNN	**GoogLeNet**
Simple connection from previous to next layer	1x1, 3x3, 5x5 convolutions and pooling between each layer

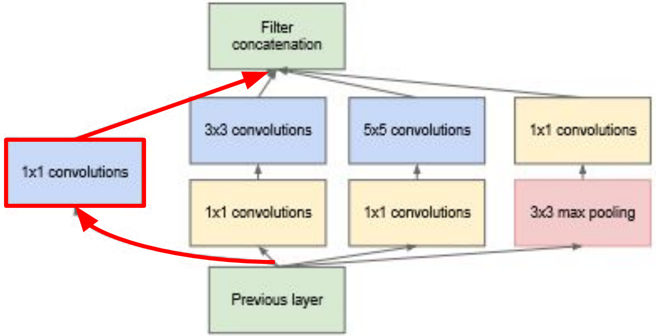




The Entire GoogleNet Architecture



Very complicated - how exactly did this architecture solve the problem?

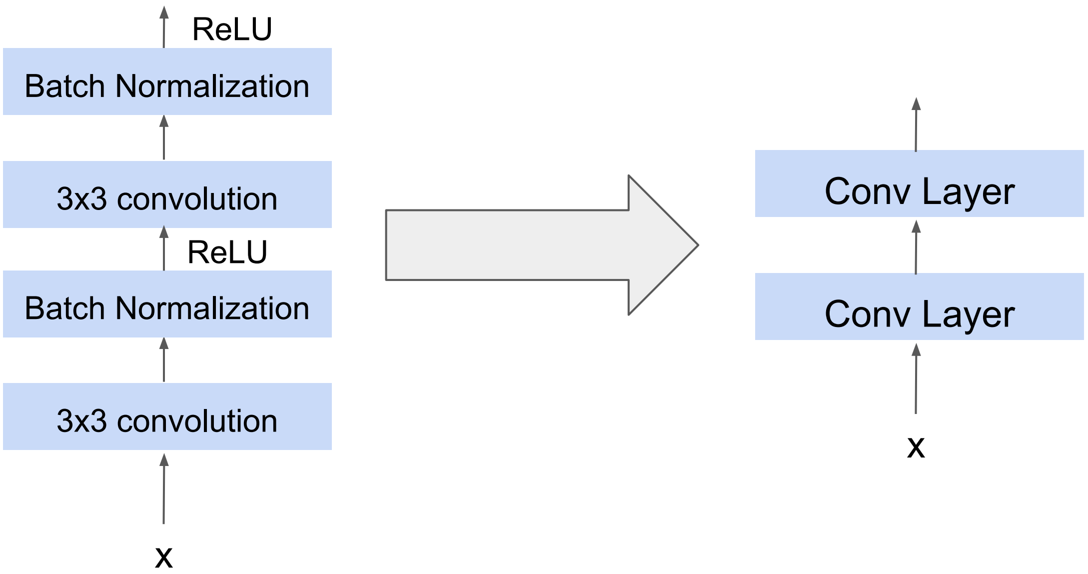


Residual connections

[Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.]

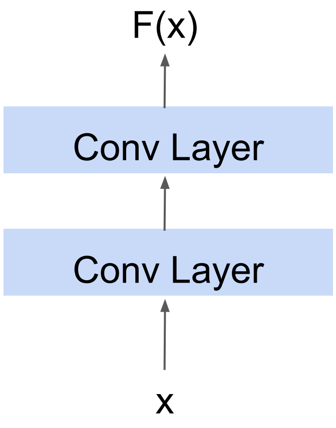
Cornell Bowers CIS

Aside: Conv Layer Abstraction

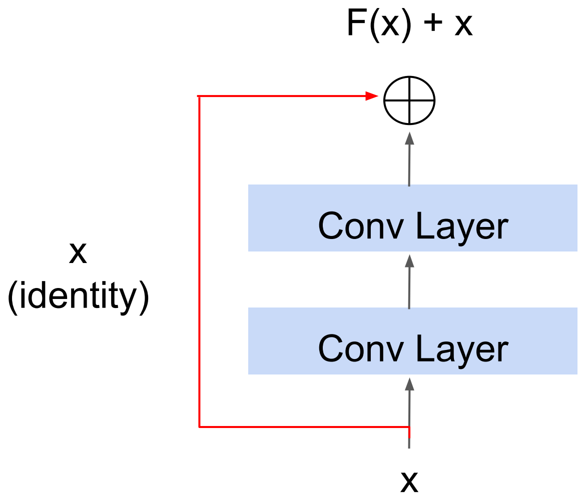


Residual Connections

aka skip connections - add an identity mapping to the output function



“Plain” layers

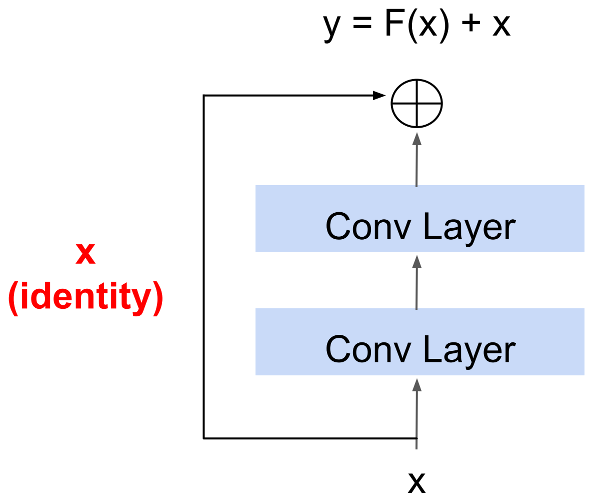


Residual Blocks

[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

Cornell Bowers C·IS

Residual Connections



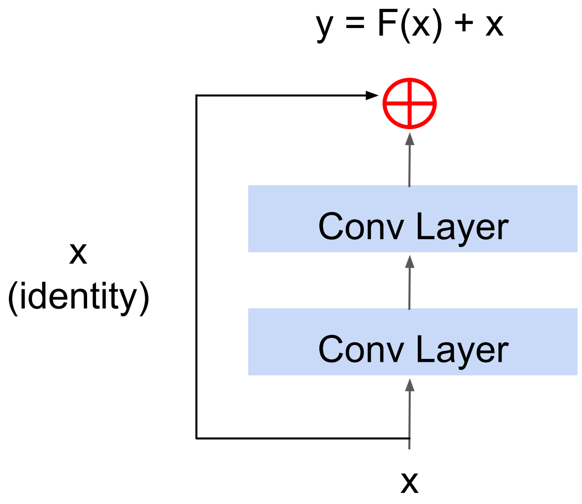
Residual Blocks

Identity mapping

- can propagate features forward
- only learn *difference* of feature maps

[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

Residual Connections



Residual Blocks

Additive component of identity

- alleviates vanishing gradients

$$\frac{\delta L}{\delta x} = \frac{\delta L}{\delta y} * \frac{\delta y}{\delta x} = \frac{\delta L}{\delta y}\boxed{(F'(x))}$$

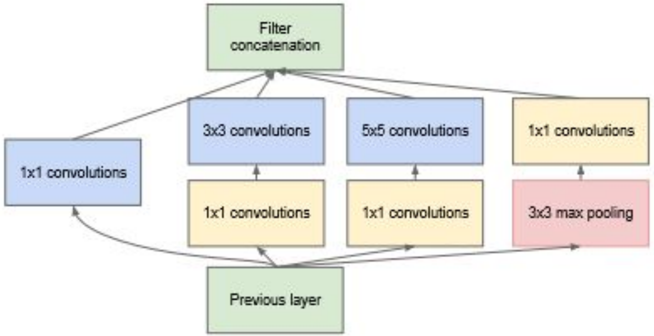
Plain

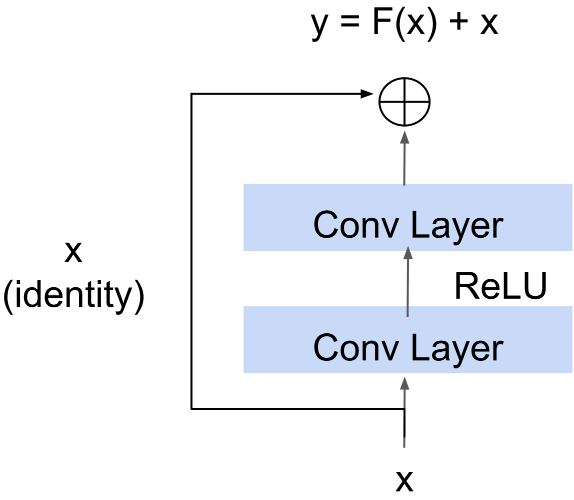
$$\frac{\delta L}{\delta x} = \frac{\delta L}{\delta y} * \frac{\delta y}{\delta x} = \frac{\delta L}{\delta y}\boxed{(1 + F'(x))}$$

ResNet

[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

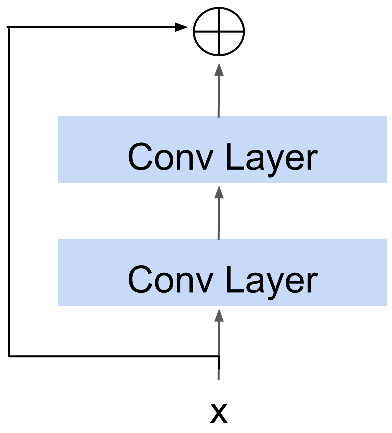
Discuss: Spot (and explain) the difference





ResNet

Stack residual blocks together!

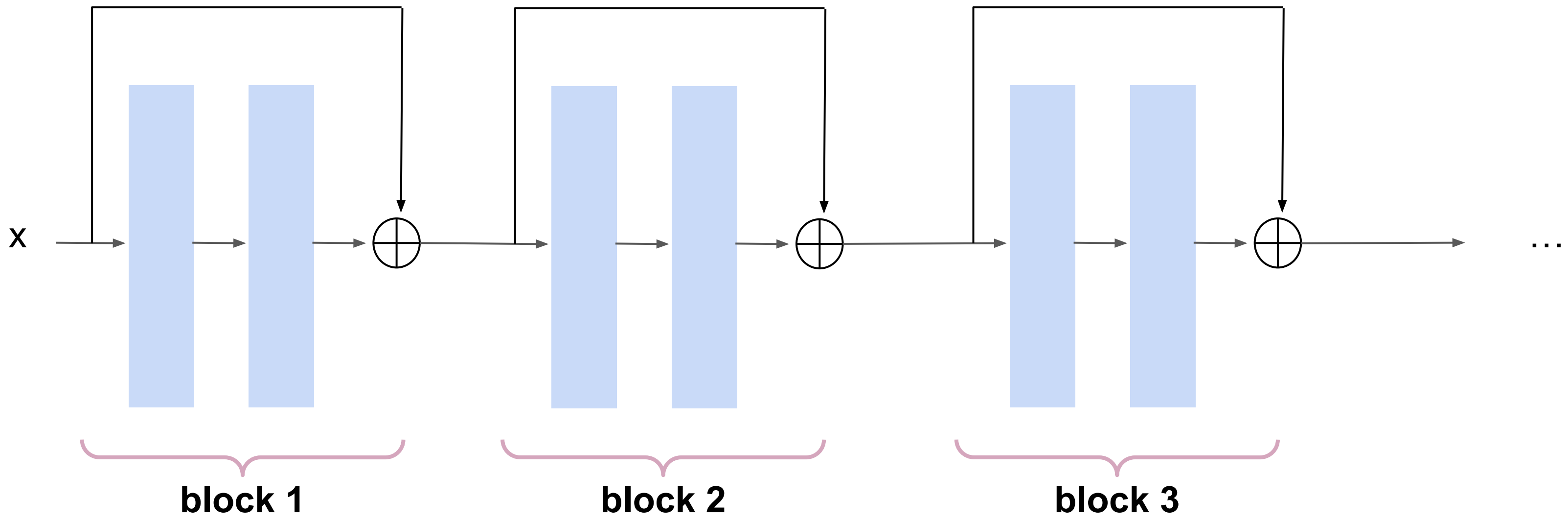


[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

Cornell Bowers CIS

ResNet

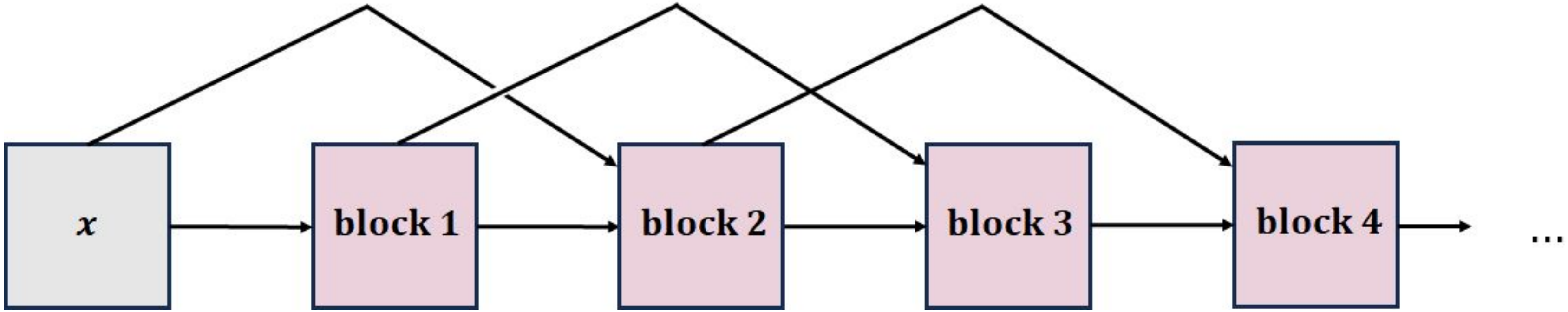
Stack residual blocks together!



[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

ResNet

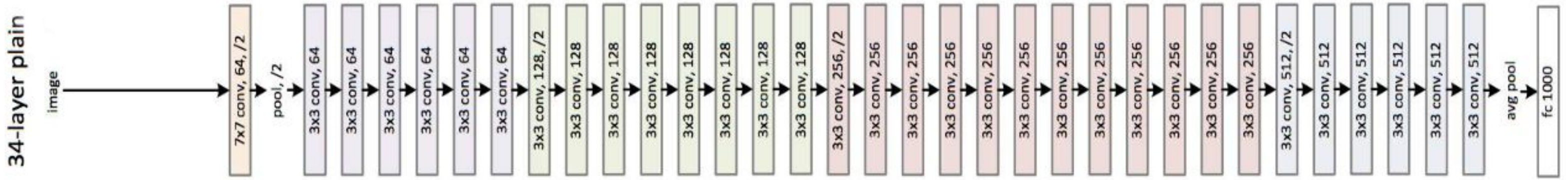
Stack residual blocks together!



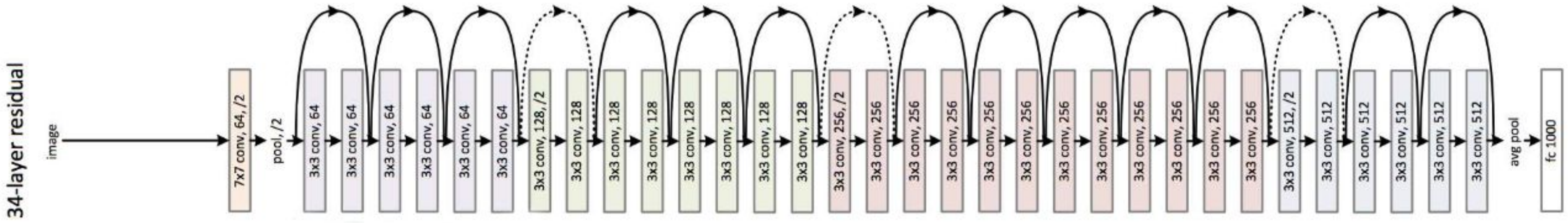
[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

Full ResNet Architecture

“Plain” Network



ResNet



[He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.]

Deeper == better

Can train deeper models!

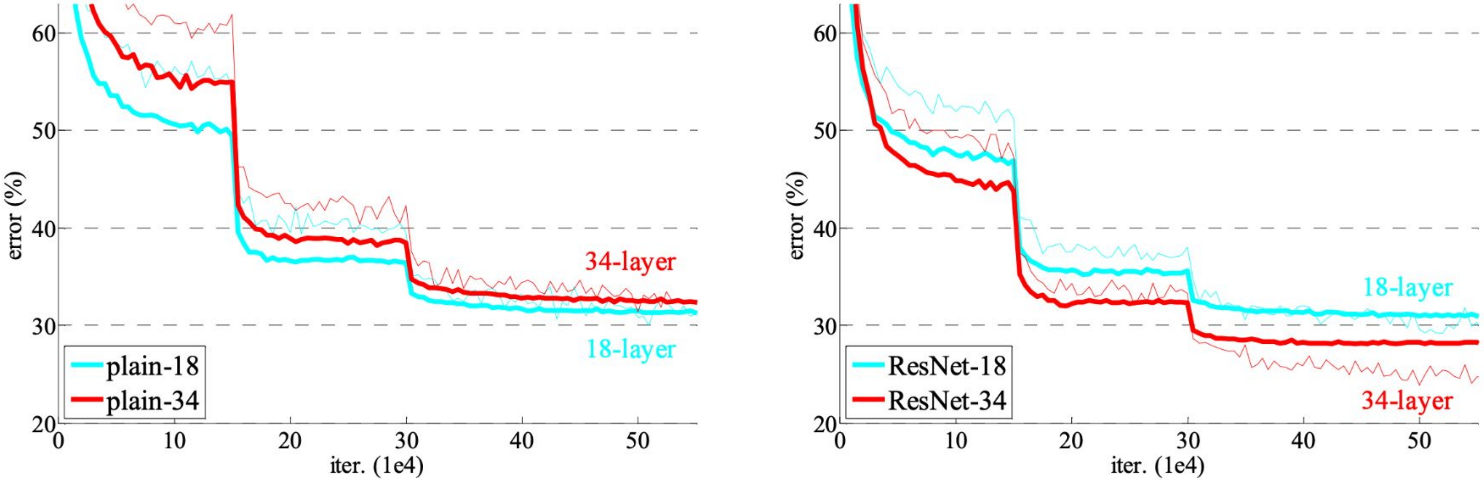
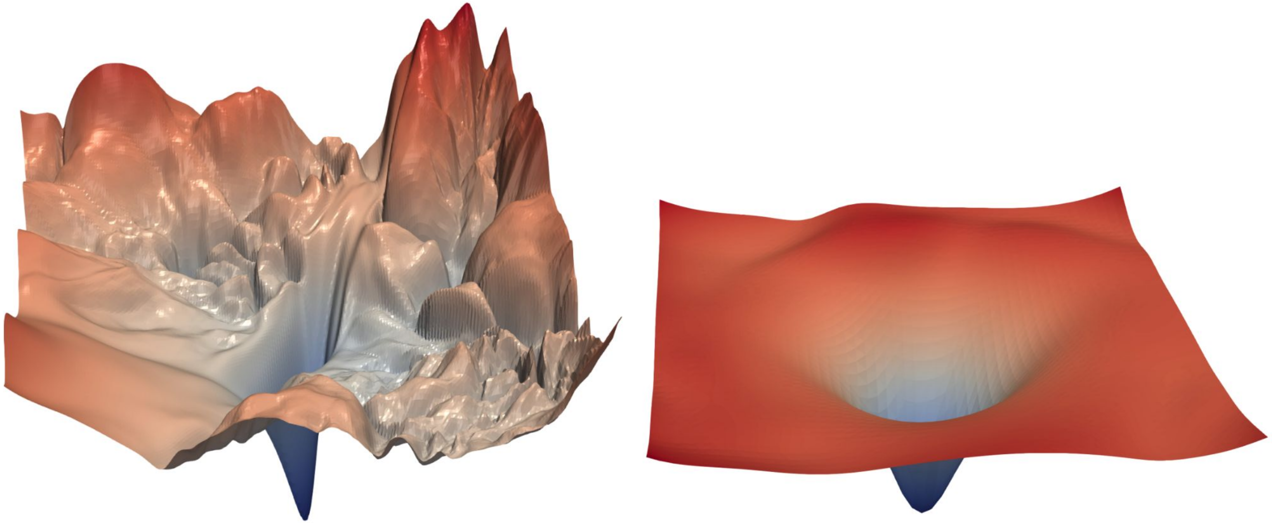


Figure 4. Training on **ImageNet**. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

Visualizing the Effect of Skip Connections

Makes optimization easier!



(a) without skip connections (b) with skip connections

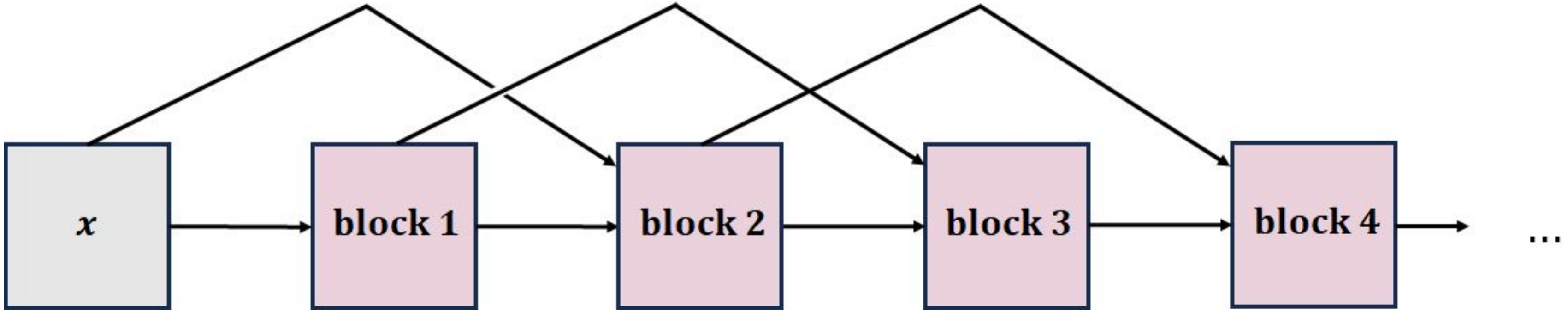
Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

[Li, Hao, et al. "Visualizing the loss landscape of neural nets." Advances in neural information processing systems 31 (2018).]

Stochastic Depth

During training, randomly drop Residual Blocks using skip connections

Like dropout but with residual blocks instead of individual neurons

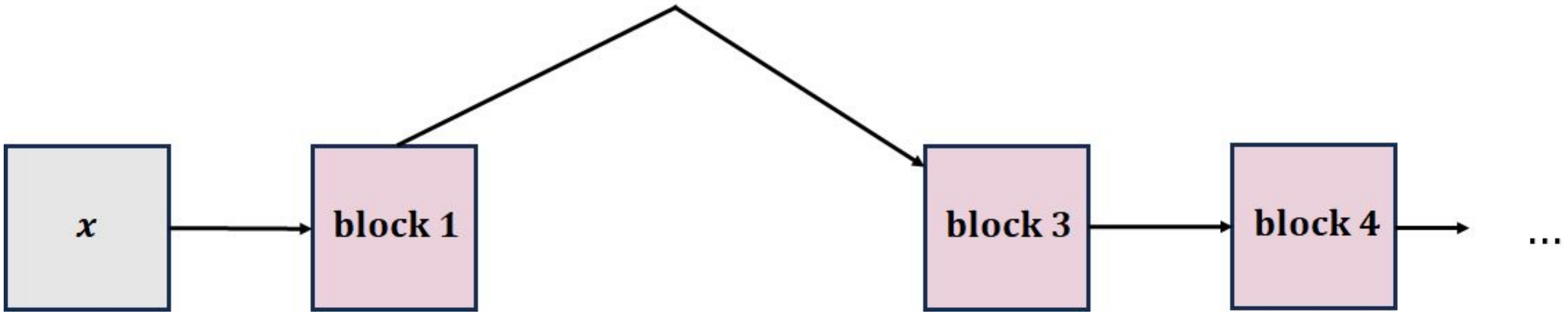


[Huang, Gao, et al. "Deep networks with stochastic depth." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14.* Springer International Publishing, 2016.]

Stochastic Depth

During training, randomly drop Residual Blocks using skip connections

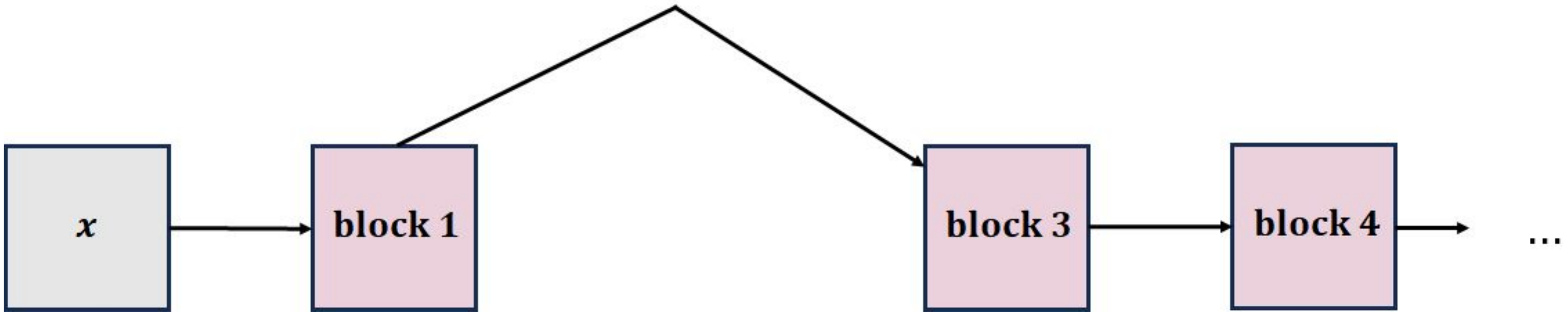
Like dropout but with residual blocks instead of individual neurons



[Huang, Gao, et al. "Deep networks with stochastic depth." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14.* Springer International Publishing, 2016.]

Stochastic Depth

Another benefit: robustness/mitigating overfitting



[Huang, Gao, et al. "Deep networks with stochastic depth." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14.* Springer International Publishing, 2016.]

Cornell Bowers C·IS

Stochastic Depth

Increases training loss, but… decreases test error!

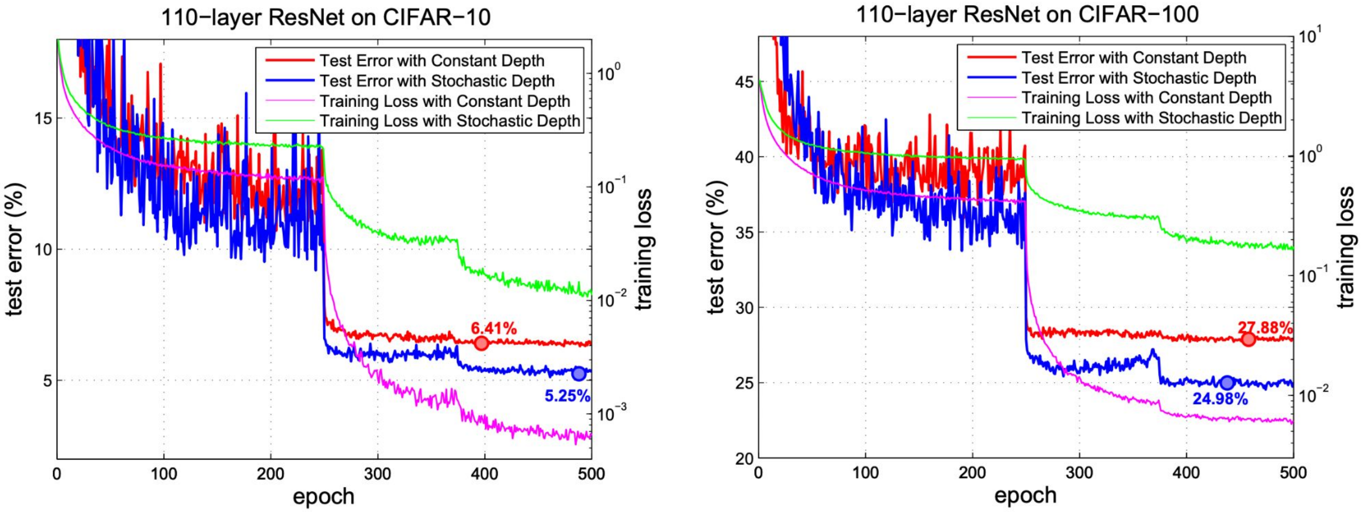
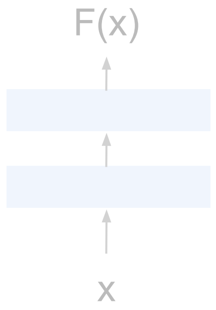


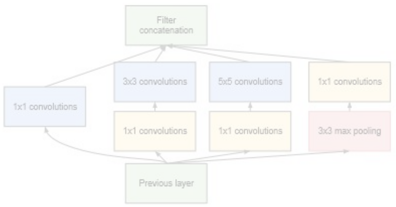
Fig. 3. Test error on CIFAR-10 (*left*) and CIFAR-100 (*right*) during training, with data augmentation, corresponding to results in the first two columns of Table 1.

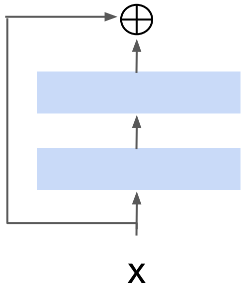
[Huang, Gao, et al. "Deep networks with stochastic depth." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14.* Springer International Publishing, 2016.]

CNN Architectures

"Plain" CNN	GoogLeNet	ResNet
Simple connection from previous to next layer	1x1, 3x3, 5x5 convolutions and pooling between each layer	Skip connections Add output of previous layer to next layer

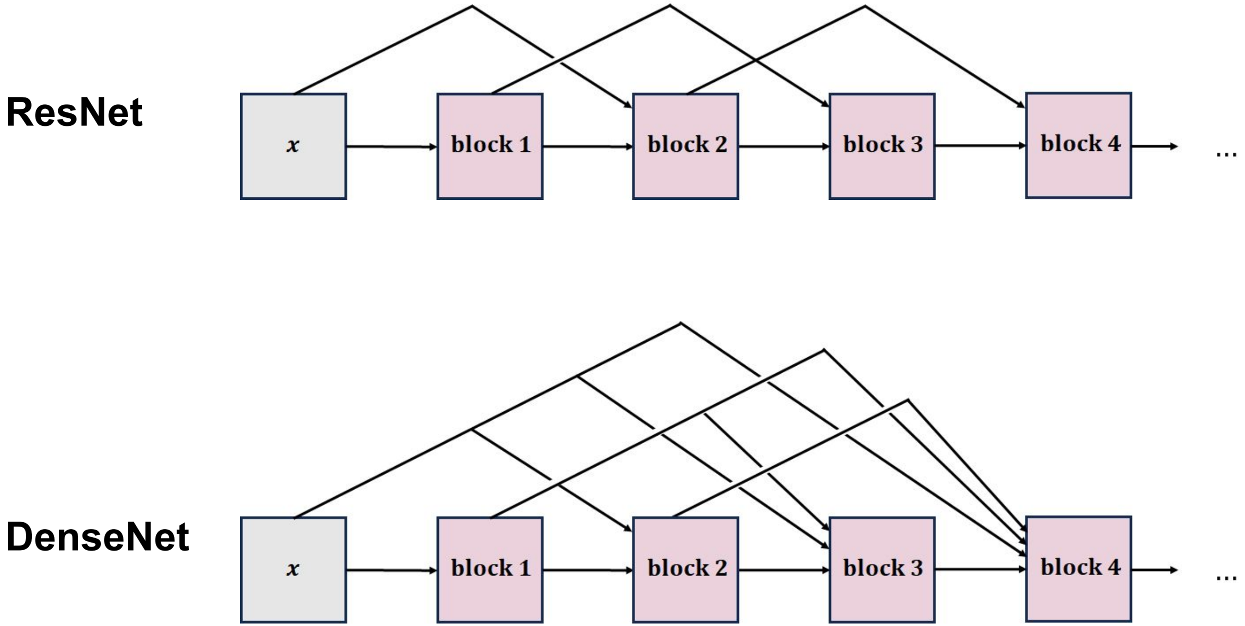






Cornell Bowers CIS

From ResNets to DenseNets



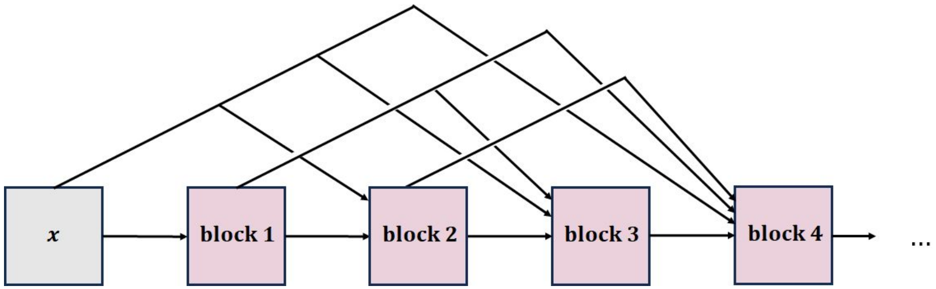
[Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.]

Cornell Bowers C·IS

Dense Connections

Each layer has access to every other layer before it, which:

- maximizes information flow
- allows for feature-map reuse
- less parameters to learn
- alleviates vanishing gradient

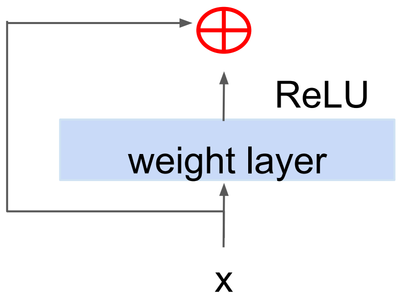


[Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.]

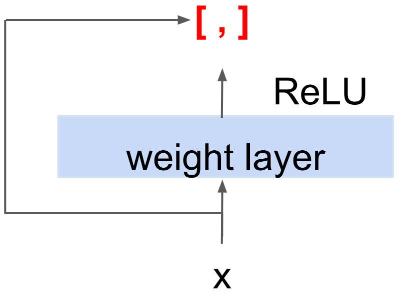
Cornell Bowers C·IS

Dense Blocks

To create dense connections, dense blocks use the same structure as residual blocks, but <u>concatenate</u> (denoted by [,]) inputs instead of simply adding them



Residual Blocks



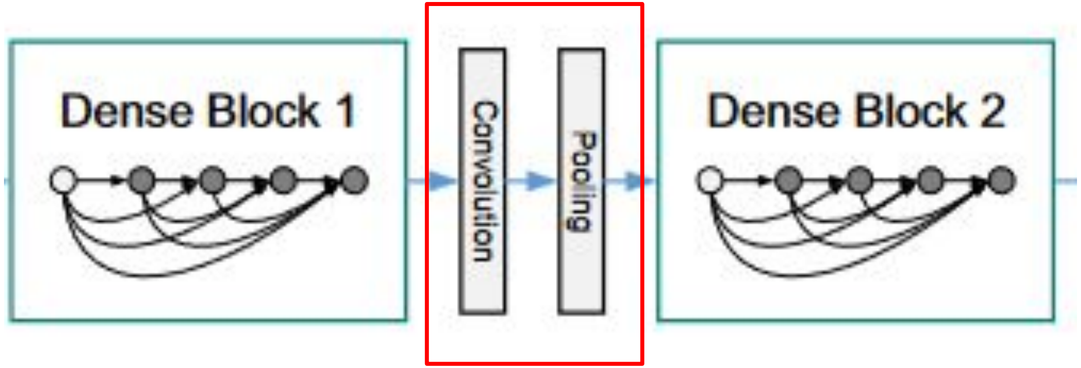
Dense Blocks

[Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.]

Transition Layers

Each dense block increases the number of dimensions

Maintain size of dimension with 1x1 convolutions and pooling (transition layer)

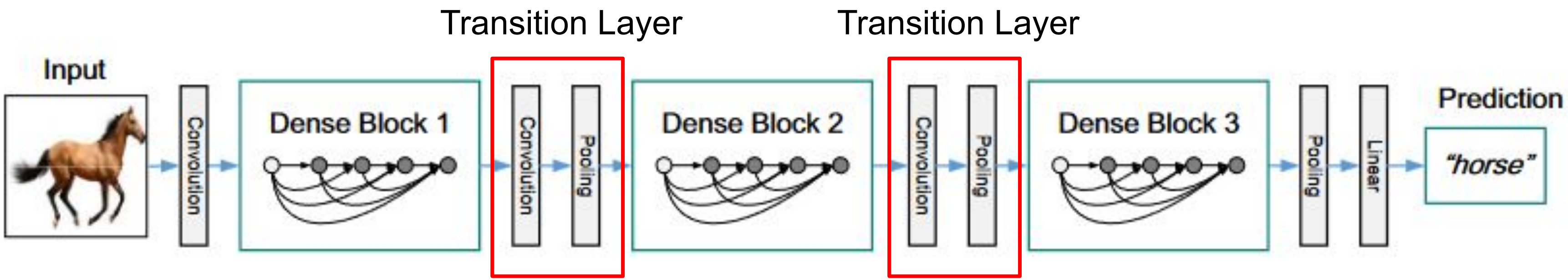


Transition Layer

[Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.]

DenseNet Architecture

Stack Dense Blocks together with transition layers in between each Dense Block

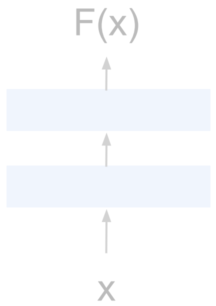


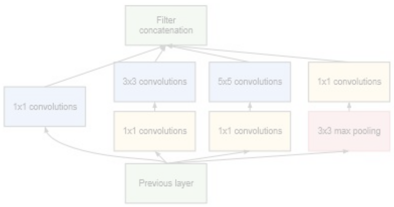
[Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.]

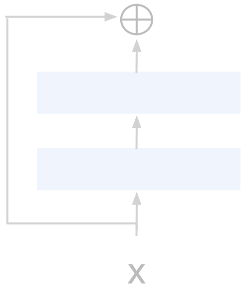
Cornell Bowers C·IS

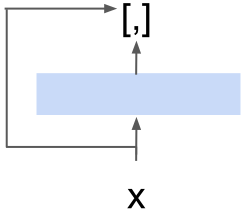
CNN Architectures

"Plain" CNN	GoogLeNet	ResNet	**DenseNet**
Simple connection from previous to next layer	1x1, 3x3, 5x5 convolutions and pooling between each layer	Skip connections Add output of previous layer to next layer	Dense connections Concatenate output of previous layer to next layer









Cornell Bowers CIS

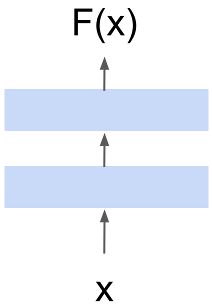
Model Comparison - error rates

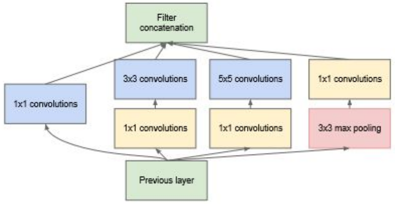
Method	Depth	Params	C10	C10+	C100	C100+	SVHN
Network in Network [22]	-	-	10.41	8.81	35.68	-	2.35
All-CNN [31]	-	-	9.08	7.25	-	33.71	-
Deeply Supervised Net [20]	-	-	9.69	7.97	-	34.57	1.92
Highway Network [33]	-	-	-	7.72	-	32.39	-
FractalNet [17]	21	38.6M	10.18	5.22	35.34	23.30	2.01
with Dropout/Drop-path	21	38.6M	7.33	4.60	28.20	23.73	1.87
ResNet [11]	110	1.7M	-	6.61	-	-	-
ResNet (reported by [13])	110	1.7M	13.63	6.41	44.74	27.22	2.01
ResNet with Stochastic Depth [13]	110	1.7M	11.66	5.23	37.80	24.58	1.75
	1202	10.2M	-	4.91	-	-	-
Wide ResNet [41]	16	11.0M	-	4.81	-	22.07	-
	28	36.5M	-	4.17	-	20.50	-
with Dropout	16	2.7M	-	-	-	-	1.64
ResNet (pre-activation) [12]	164	1.7M	11.26*	5.46	35.58*	24.33	-
	1001	10.2M	10.56*	4.62	33.47*	22.71	-
DenseNet ($k = 12$)	40	1.0M	**7.00**	5.24	**27.55**	24.42	1.79
DenseNet ($k = 12$)	100	7.0M	**5.77**	**4.10**	**23.79**	**20.20**	1.67
DenseNet ($k = 24$)	100	27.2M	**5.83**	**3.74**	**23.42**	**19.25**	**1.59**
DenseNet-BC ($k = 12$)	100	0.8M	**5.92**	4.51	**24.15**	22.27	1.76
DenseNet-BC ($k = 24$)	250	15.3M	**5.19**	**3.62**	**19.64**	**17.60**	1.74
DenseNet-BC ($k = 40$)	190	25.6M	-	**3.46**	-	**17.18**	-

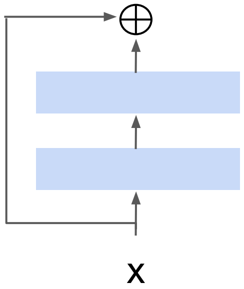
[Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.]

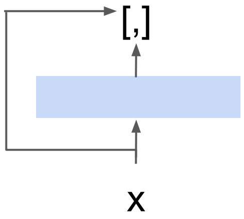
Summary of Models

"Plain" CNN	Google Net	ResNet	DenseNet
Simple connection from previous to next layer	1x1, 3x3, 5x5 convolutions and pooling between each layer	Skip connections Add output of previous layer to next layer	Dense connections Concatenate output of previous layer to next layer









Summary

- Deep CNNs outperform shallow CNNs
- But…
    - Harder optimization problem!
- Residual (and dense) connections make training easier!
    - Can train networks with 100s of layers!
- Stochastic depth let’s you train deeper networks faster
    - 1000+ layers!
- In general…
    - Build large networks as stacks of (many!) simple building blocks