VISUAL RECOGNITION

ANUSH SANKARAN
IBM RESEARCH

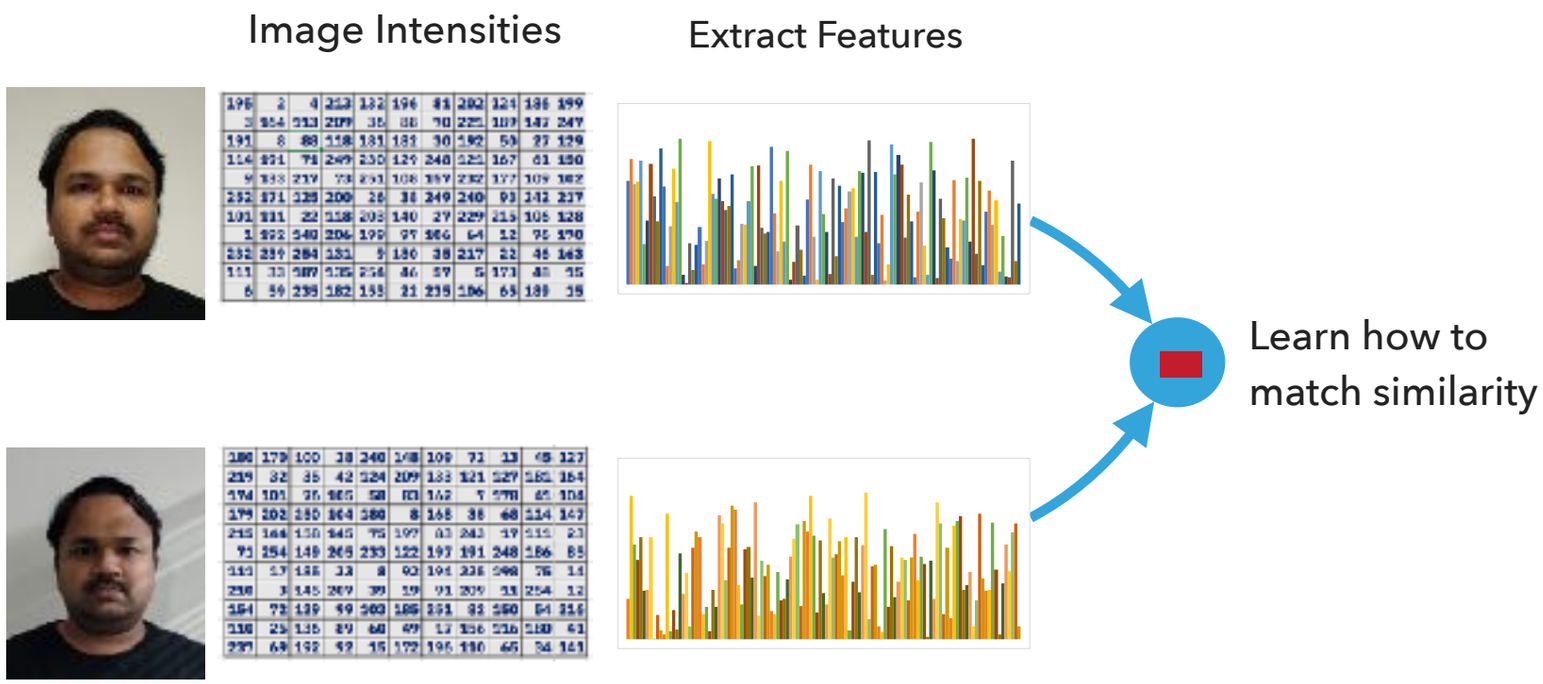
Most of the slide content are borrowed from CS231n. Thanks to the original authors

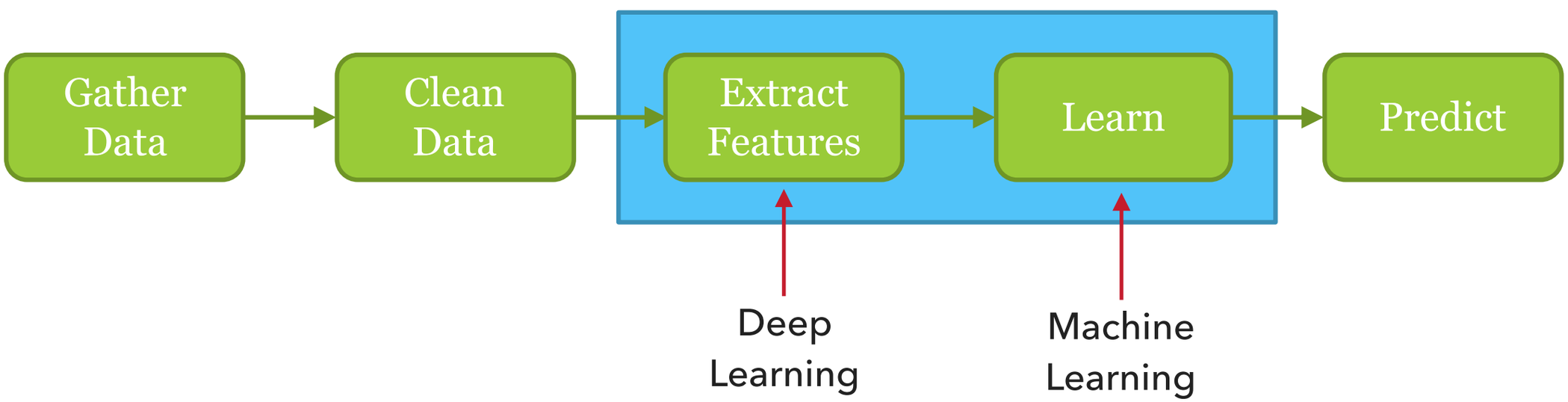
RECAP: WHAT IS THE PROBLEM?

- You are a genius …
- Only if a machine or an algorithm could do it …

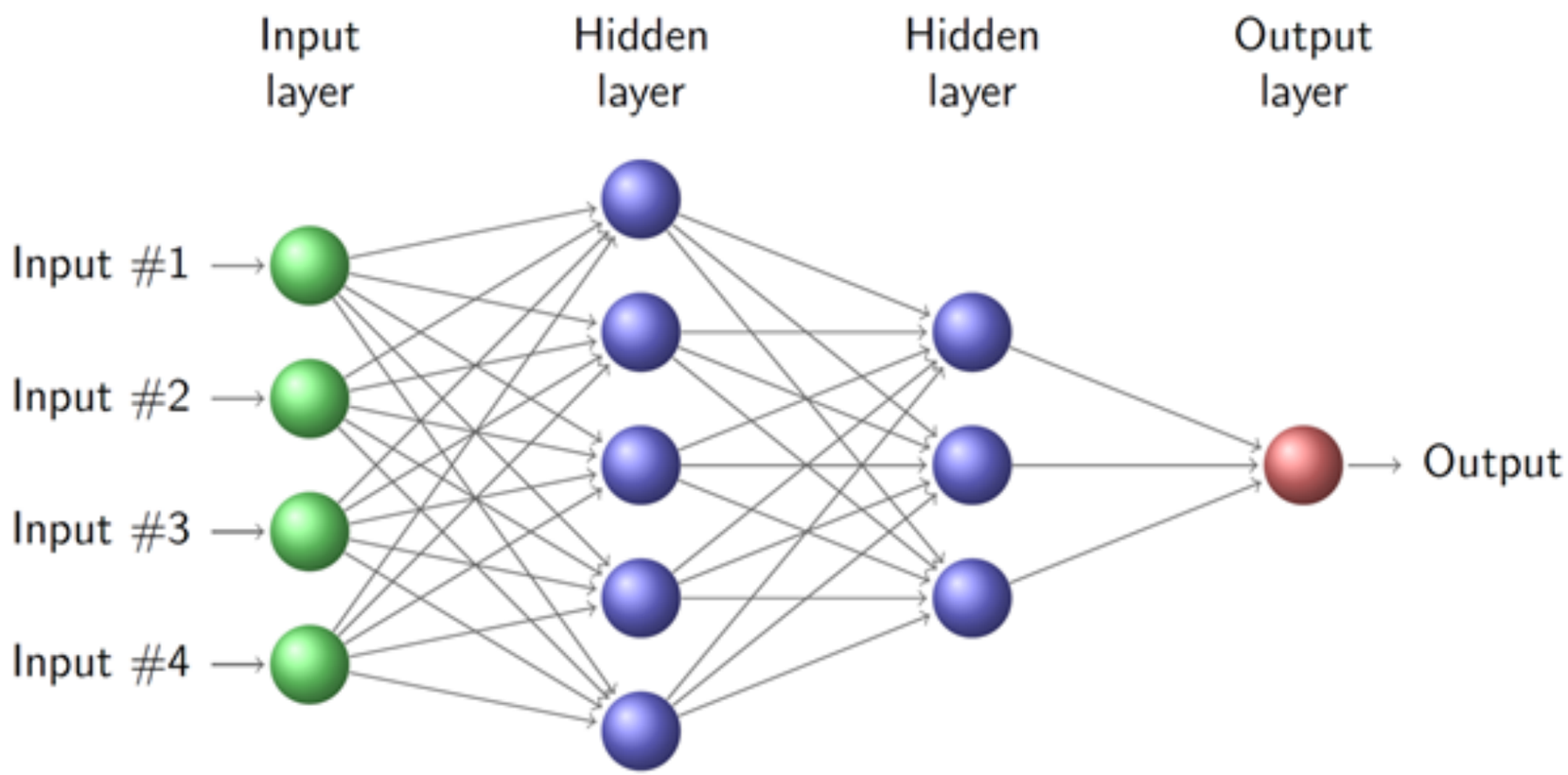
- Make a computer understand the image …
- Understand the background & environment …
- Classify various objects in the image …

RECAP: WHAT IS THE SOLUTION PIPELINE

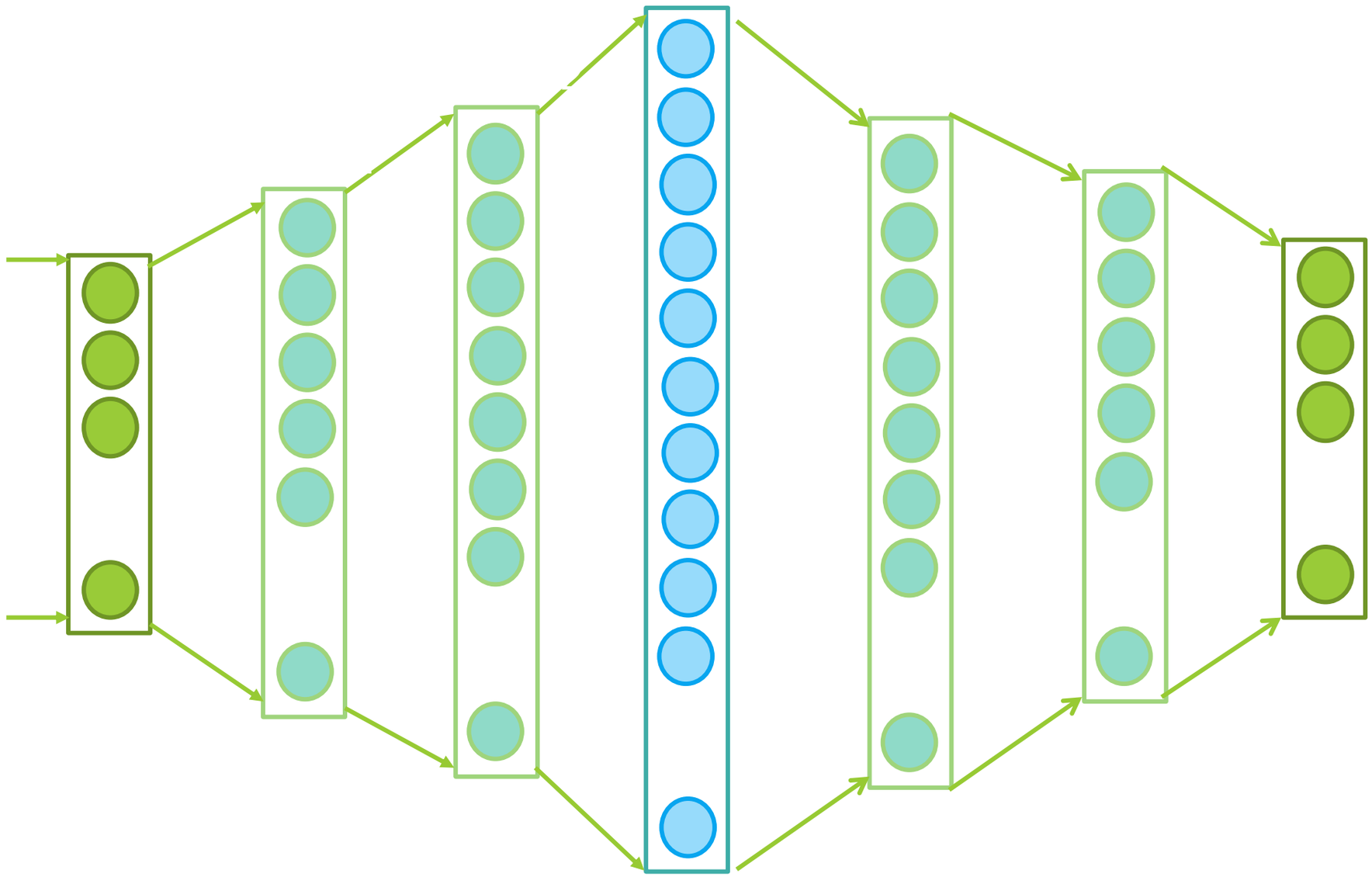




RECAP: MULTI LAYER PERECEPTRON



RECAP: STACKED AUTOENCODER



RECAP: DIFFERENT TYPES OF DEEP LEARNING

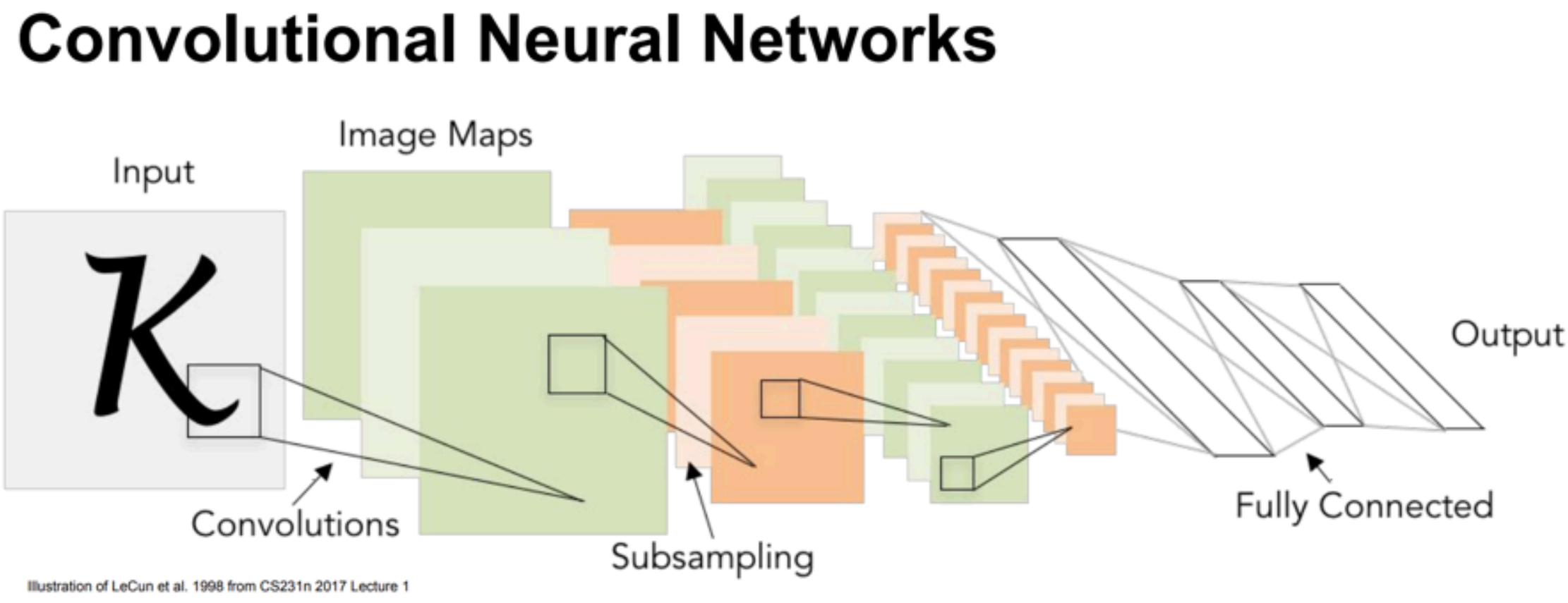
- Input data is a vector Stacked
  - Autoencoder
  - Restricted Boltzmann Machines
  - Deep Belief Networks

- Input data is multi-dimensional
  - Convolutional Neural Network
  - Example: Image analysis, Object recognition

- Input data is time-series
  - Recurrent Neural Network
  - Example: Language Modelling, Speech recognition

RECAP: CONVOLUTIONAL NEURAL NETWORK

Where we are now...



RECAP: WHAT HYPER-PARAMETERS TO PLAY WITH?

- Network architecture
  - How many layers?
  - What sequence of layers?
  - What is the depth of the model?
- Learning rate
  - Decay rate
  - Update strategy
- Regularization
  - L2 / Dropout strength
- Which loss function?
  - hinge/ L2/ Cross-entropy
- Which optimization?
  - SGD/ adam/ RMSPROP

RECAP: SUMMARY

- Activation Functions (use ReLU)
- Data Preprocessing (images: subtract mean)
- Weight Initialization (use Xavier/He init)
- Batch Normalization (use)
- Babysitting the Learning process
- Hyperparameter Optimization
  (random sample hyperparams, in log space when appropriate)

ASSIGNMENT #1 (DEADLINE: 14TH APRIL)

- Part 1: Implement Convolution2D, Pooling2D, Fully Connected Layer
  - Implement Conv2D, Pool2D, Fully connected layers from scratch without using Keras
  - Take a trained Conv. Autoencoder (using Keras, the one discussed in the class) & Save the weights as Numpy
  - Make a forward pass with your implemented layers and reproduce the actual output
  - Compare the output of your implementation with Keras implementation
- Part 2: Implement only decoder
  - Write a decoder to convert random noise into image (MNIST images)
  - Use Keras for this. For any random vector, I provide, the decoder should output a MNIST like image

ASSIGNMENT #2 (DEADLINE: 21 APRIL) TEAM SIZE (2–3)

- Consider InceptionV3 model, CIFAR-10 dataset
- Hyperparameters to explore:
  - Layer activation function
  - Data preprocessing, normalisation
  - Dropout (where to apply, what is the strength)
  - BatchNorm (where to apply)
  - Loss function
  - Learning rate and learning strategy
- Report the final best set of hyper parameters set that gave you the best accuracy.
- You will be evaluated on what values you explored for each of the parameter and in what order? Please do explain how you baby-sit the learning process.
- Please note that there is no single correct answer to this question.

EASY LOGISTICS: GITHUB

- Github Repo: https://github.com/goodboyanush/iiit-bangalore-march-april-2019
- Lectures slides, Hands-on notebooks, Assignment solutions
- Have any doubt in my lectures or assignments?
  - Go ahead and create an issue in the repo!
  - I will try to answer them asap!
  - Everyone will be benefitted by the questions asked by one student

MOST IMPORTANT CHALLENGE

CNN Model has lot of weight parameters to learn…

"You need lots of data to train/use a CNN model"

MOST IMPORTANT CHALLENGE

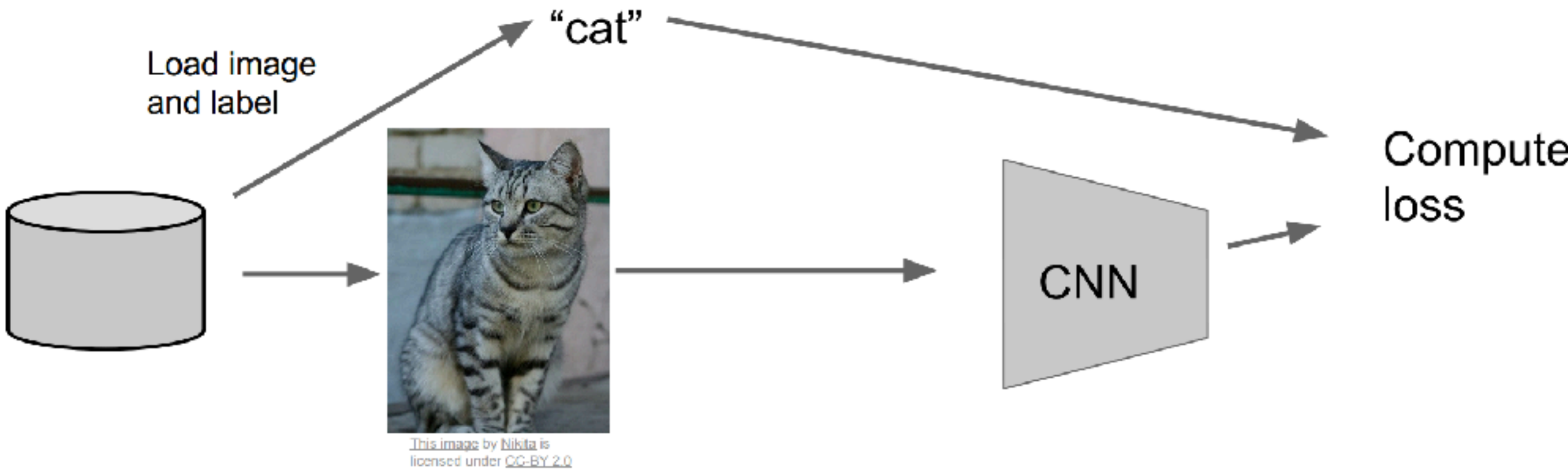
CNN Model has lot of weight parameters to learn…

"You need lots of data to train/use a CNN model"

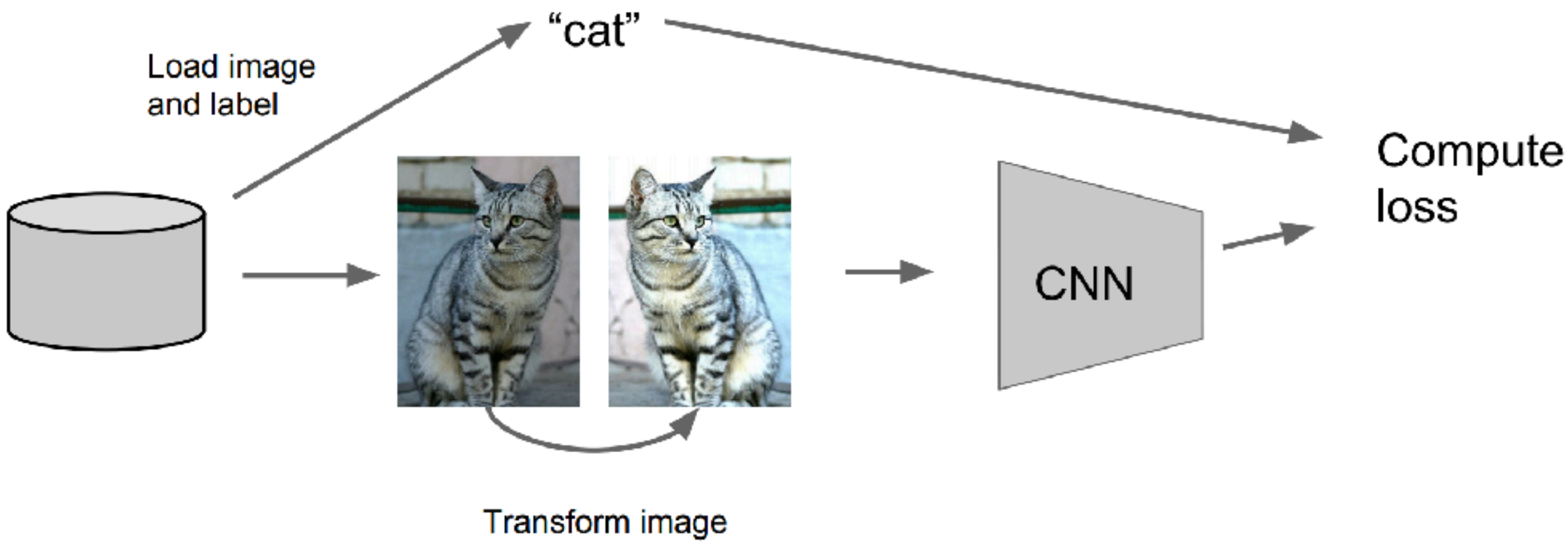
BUSTED

1. Data Augmentation
2. Transfer Learning

DATA AUGMENTATION

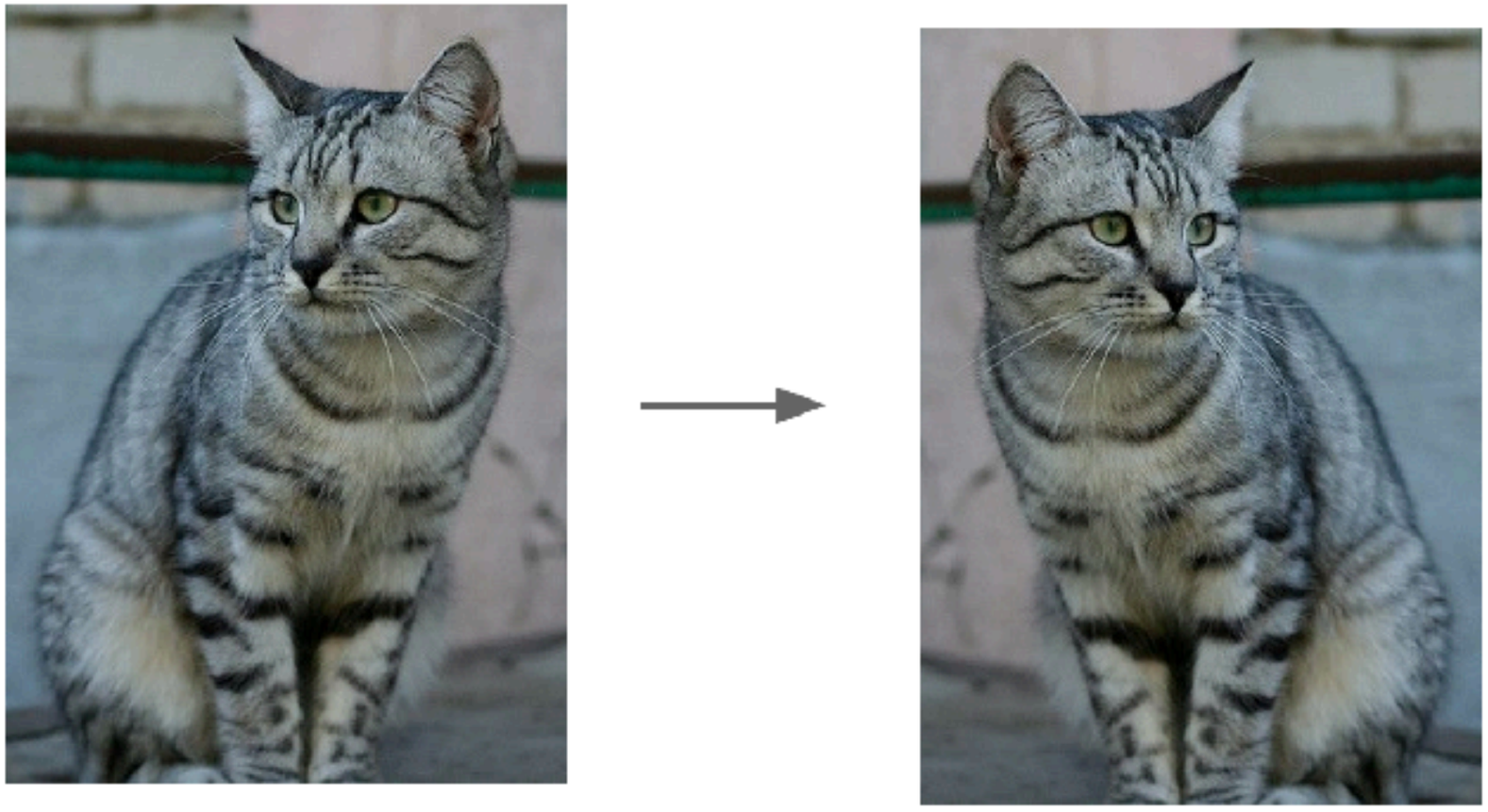


DATA AUGMENTATION



DATA AUGMENTATION

Horizontal Flip



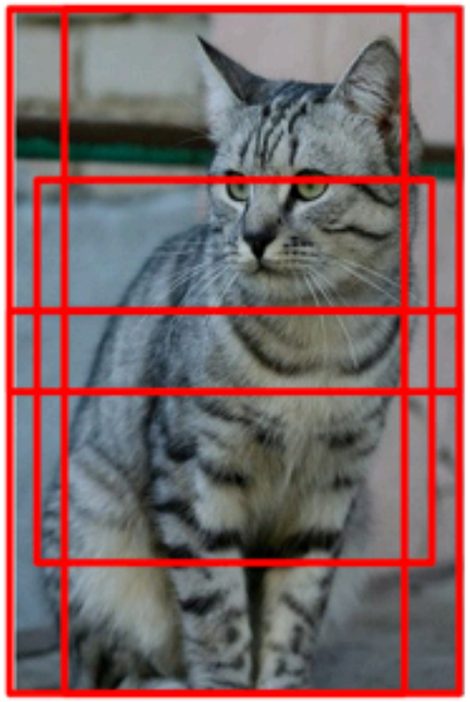
DATA AUGMENTATION

Random crops and scales

Training: sample random crops / scales

ResNet:

1. Pick random L in range [256, 480]
2. Resize training image, short side = L
3. Sample random 224 x 224 patch



Testing: average a fixed set of crops

ResNet:

1. Resize image at 5 scales: {224, 256, 384, 480, 640}
2. For each size, use 10 224 x 224 crops: 4 corners + center, + flips

DATA AUGMENTATION

Data Augmentation

Color Jitter

Simple: Randomize contrast and brightness

More Complex:

1. Apply PCA to all [R, G, B] pixels in training set
2. Sample a "color offset" along principal component directions
3. Add offset to all pixels of a training image

(As seen in *[Krizhevsky et al. 2012]*, ResNet, etc)

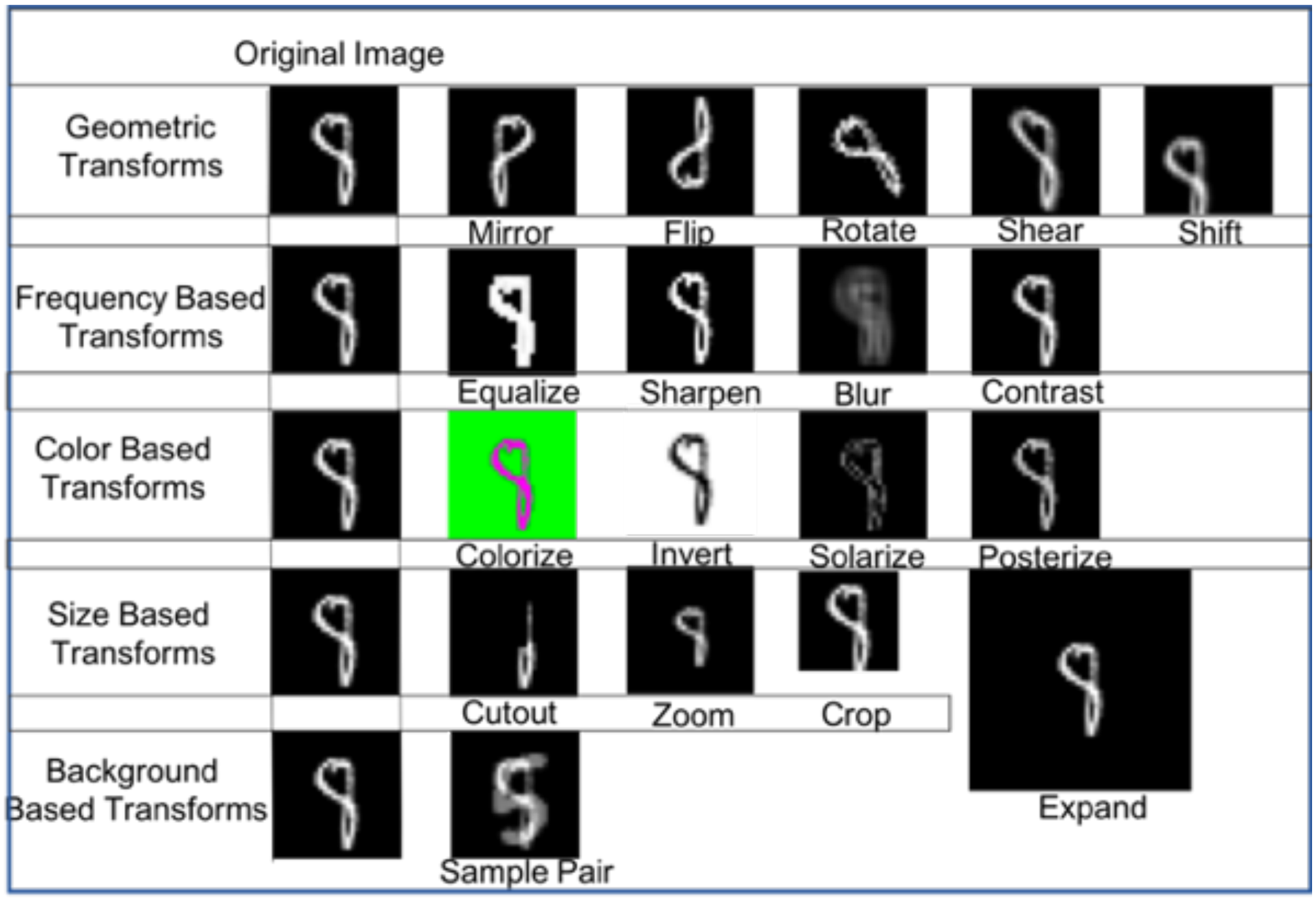
DATA AUGMENTATION

Get creative for your problem!

Random mix/combinations of :

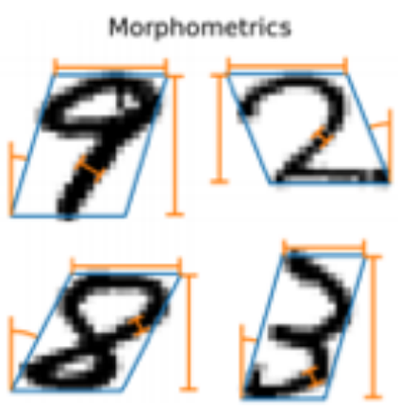
- translation
- rotation
- stretching
- shearing,
- lens distortions, … (go crazy)

DATA AUGMENTATION

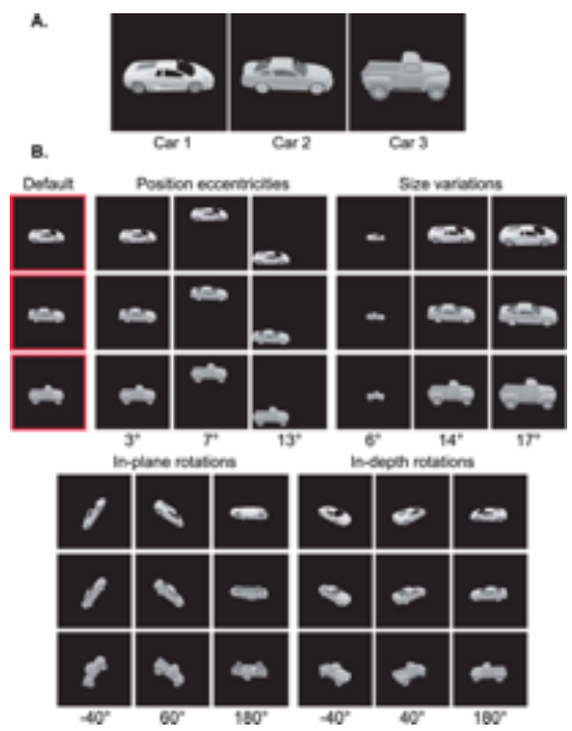


DATA AUGMENTATION

Morphological Variations

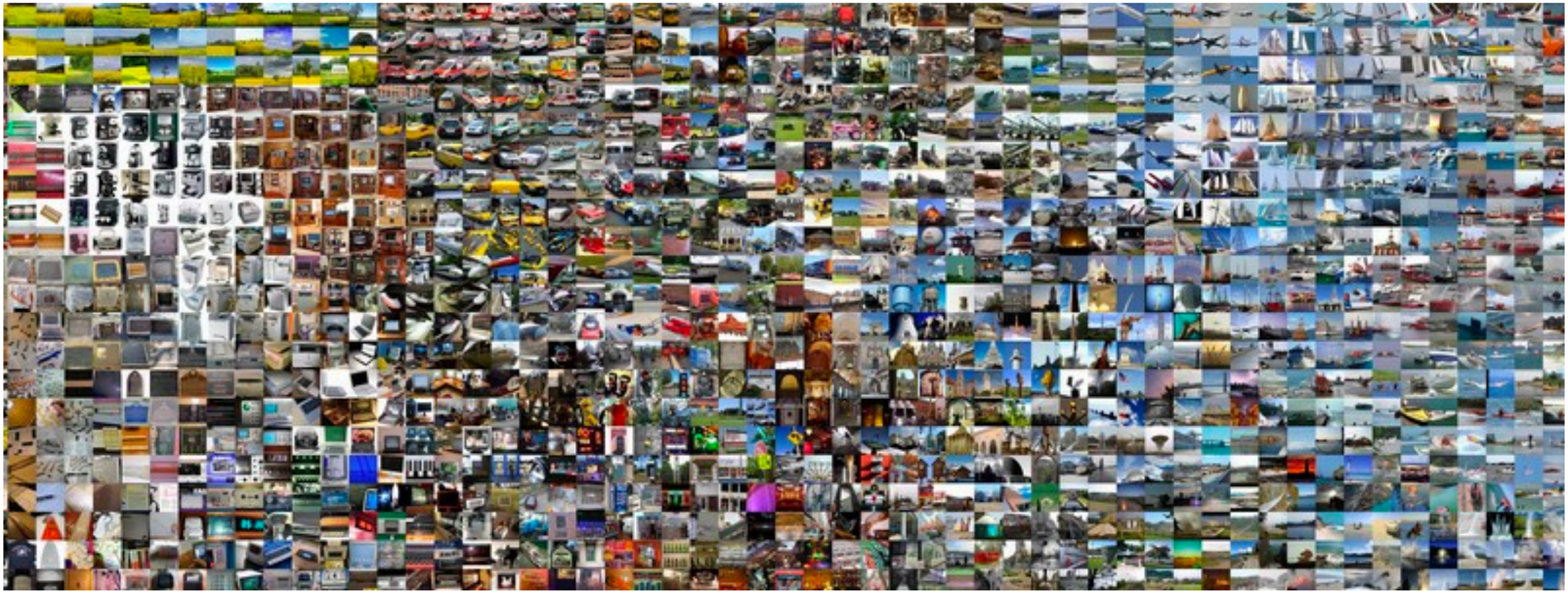


Geometric Variations



Color Variations

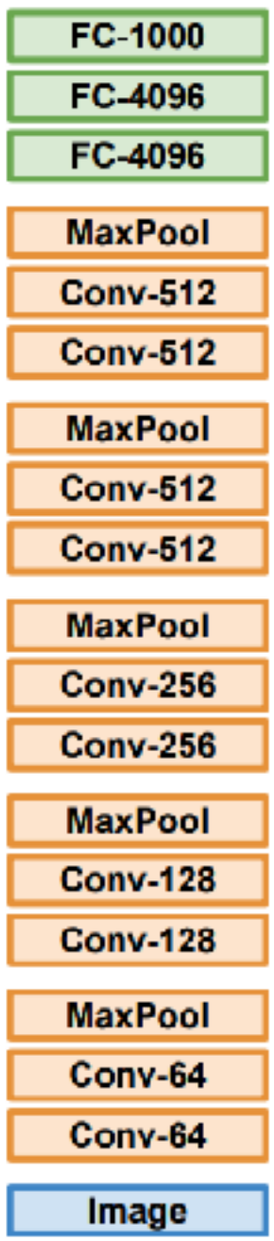
DATASETS: IMAGENET



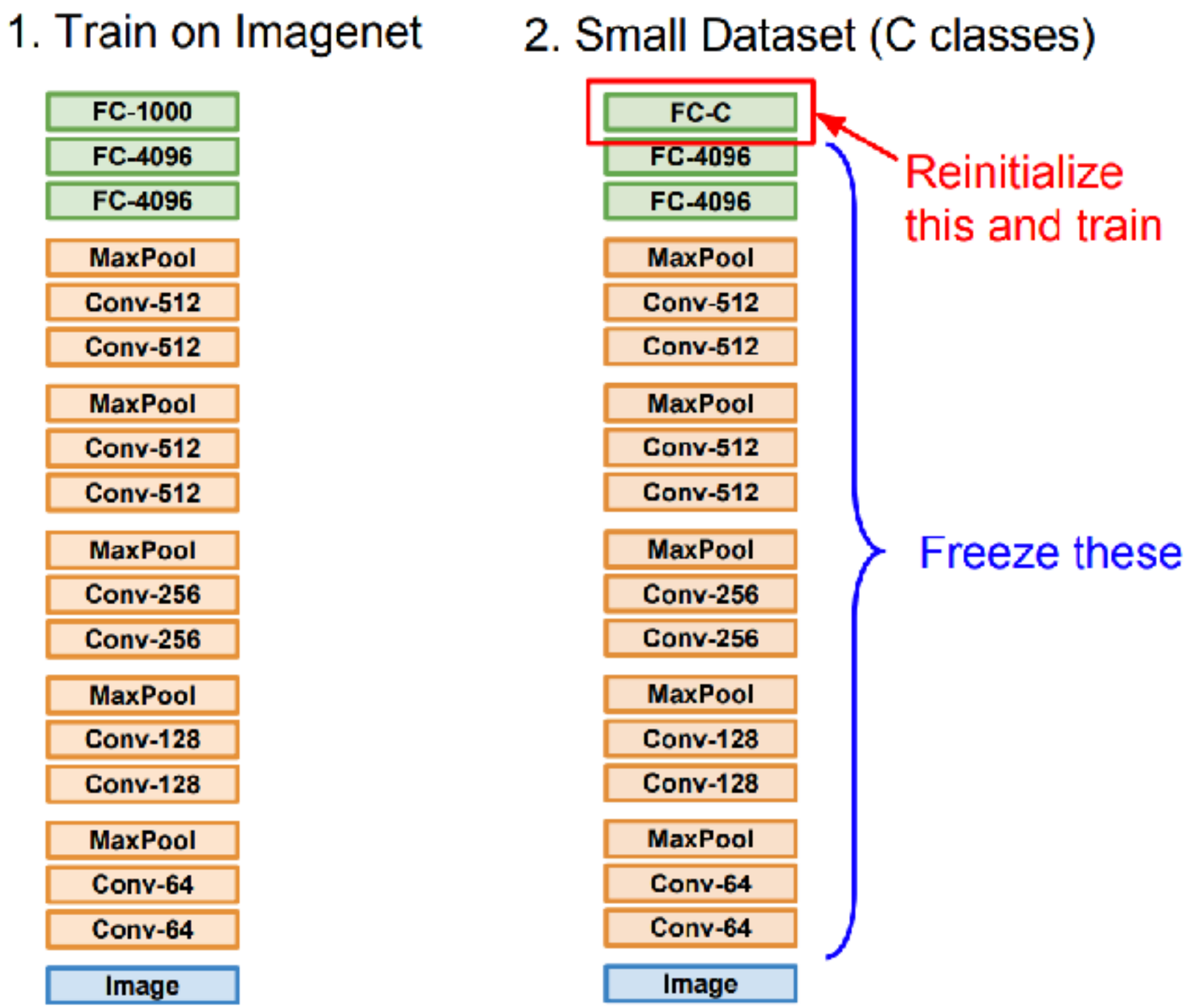
- 256*256*3 color images
- More than 1000 classes
- Manually annotated in a crowdsource manner

TRANSFER LEARNING

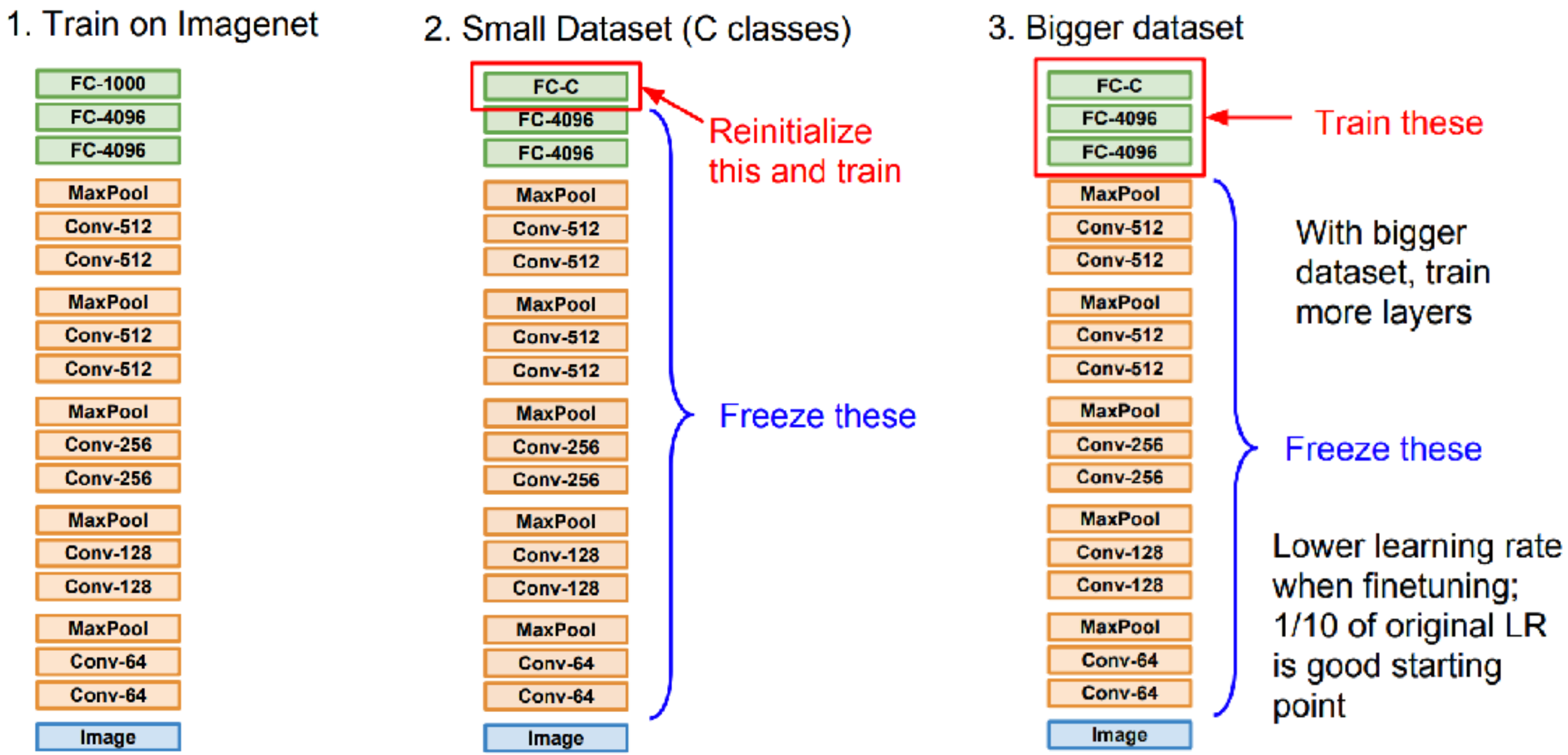
1. Train on Imagenet



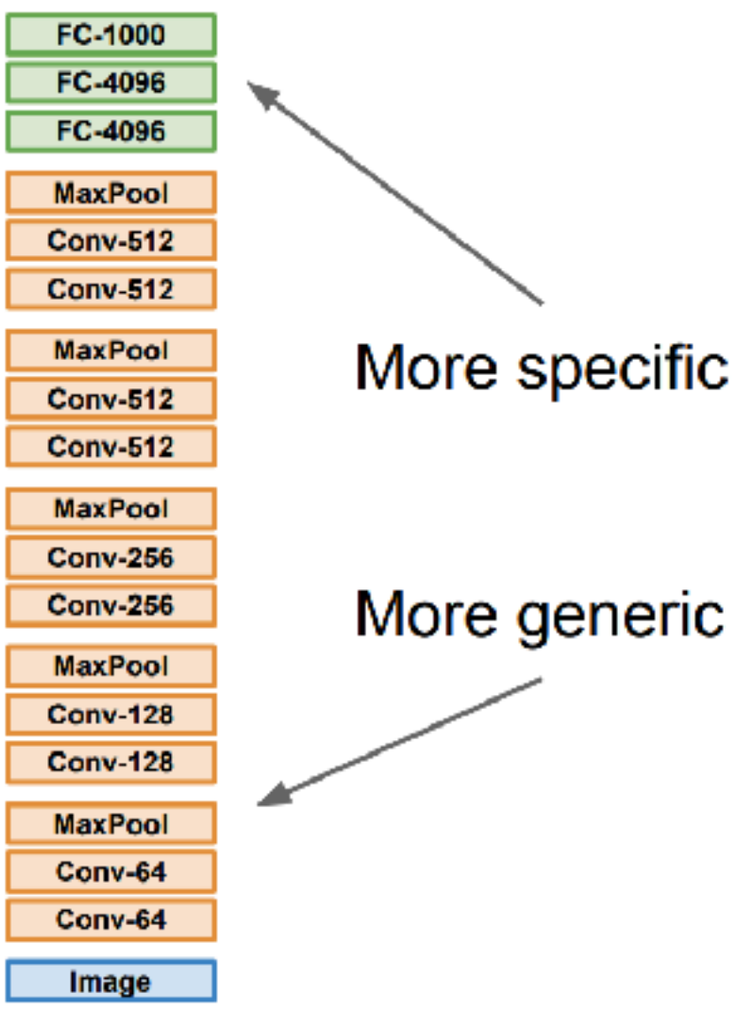
TRANSFER LEARNING



TRANSFER LEARNING



TRANSFER LEARNING



	very similar dataset	very different dataset
very little data	Use Linear Classifier on top layer	You're in trouble… Try linear classifier from different stages
quite a lot of data	Finetune a few layers	Finetune a larger number of layers

DIFFERENT CNN ARCHITECTURES

Case Studies

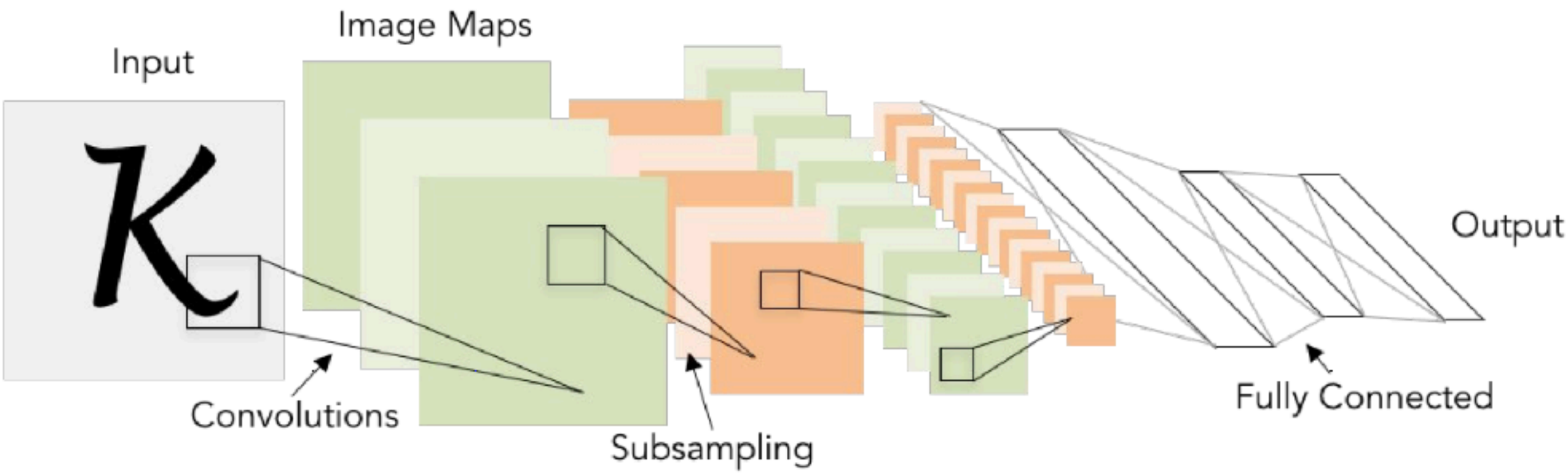
- AlexNet
- VGG
- GoogLeNet
- ResNet

Also....

- NiN (Network in Network)
- Wide ResNet
- ResNeXT
- Stochastic Depth
- Squeeze-and-Excitation Network
- DenseNet
- FractalNet
- SqueezeNet
- NASNet

LENET-5: 1998

[LeCun et al., 1998]

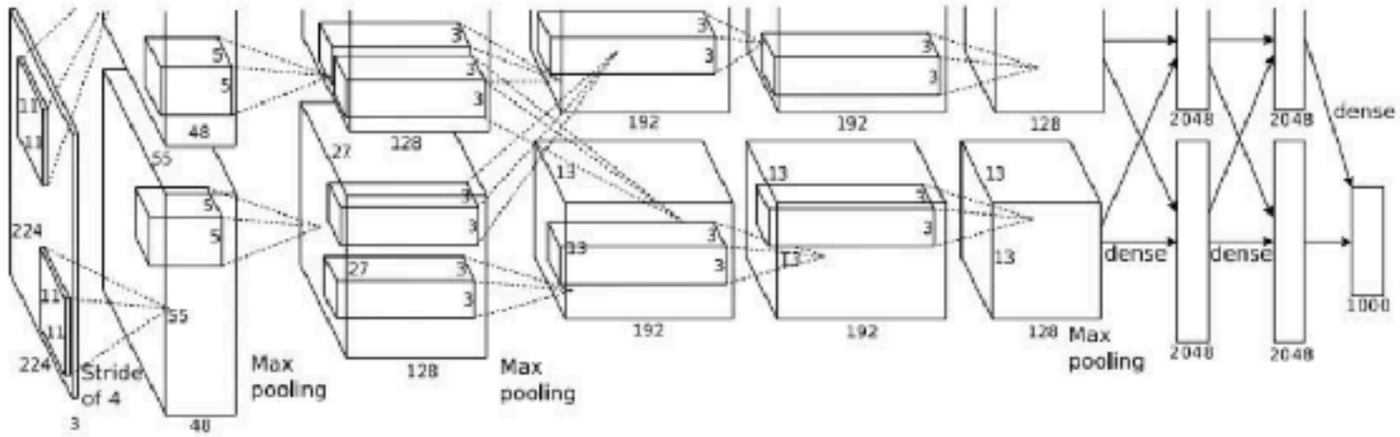


Conv filters were 5x5, applied at stride 1
Subsampling (Pooling) layers were 2x2 applied at stride 2
i.e. architecture is [CONV-POOL-CONV-POOL-FC-FC]

ALEXNET: 2012

Case Study: AlexNet

[Krizhevsky et al. 2012]

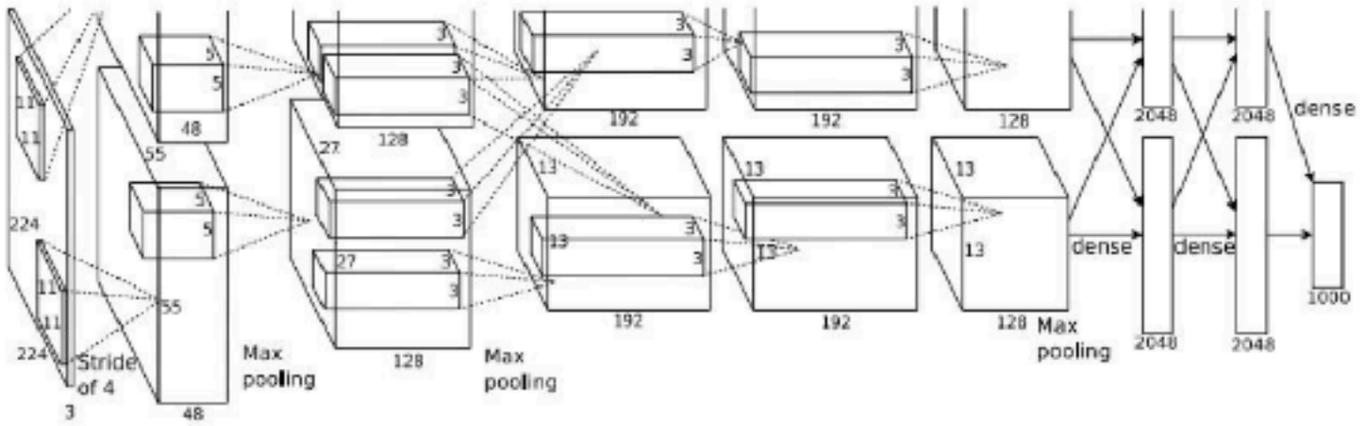


Architecture:
CONV1
MAX POOL1
NORM1
CONV2
MAX POOL2
NORM2
CONV3
CONV4
CONV5
Max POOL3
FC6
FC7
FC8

ALEXNET: 2012

Case Study: AlexNet

[Krizhevsky et al. 2012]

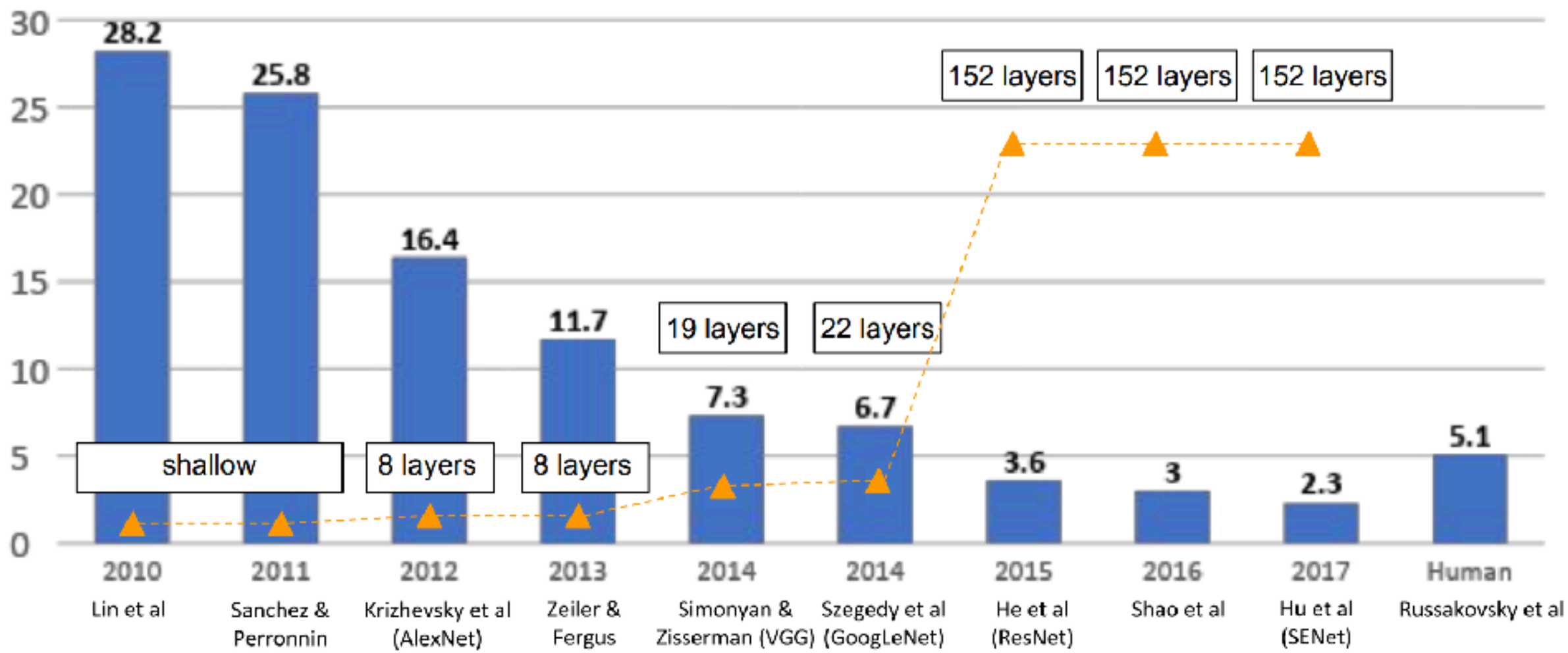


Full (simplified) AlexNet architecture:
[227x227x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores)

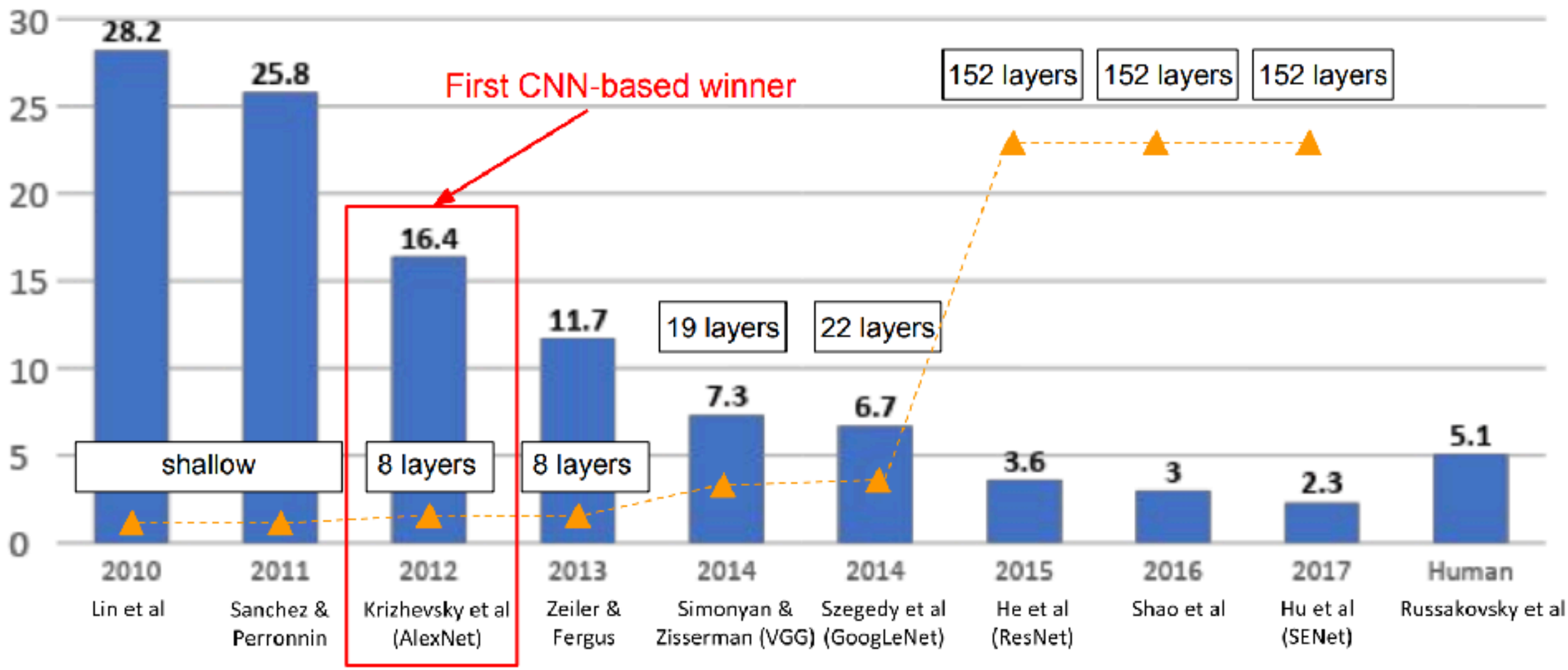
Details/Retrospectives:
- first use of ReLU
- used Norm layers (not common anymore)
- heavy data augmentation
- dropout 0.5
- batch size 128
- SGD Momentum 0.9
- Learning rate 1e-2, reduced by 10 manually when val accuracy plateaus
- L2 weight decay 5e-4
- 7 CNN ensemble: 18.2% -> 15.4%

Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

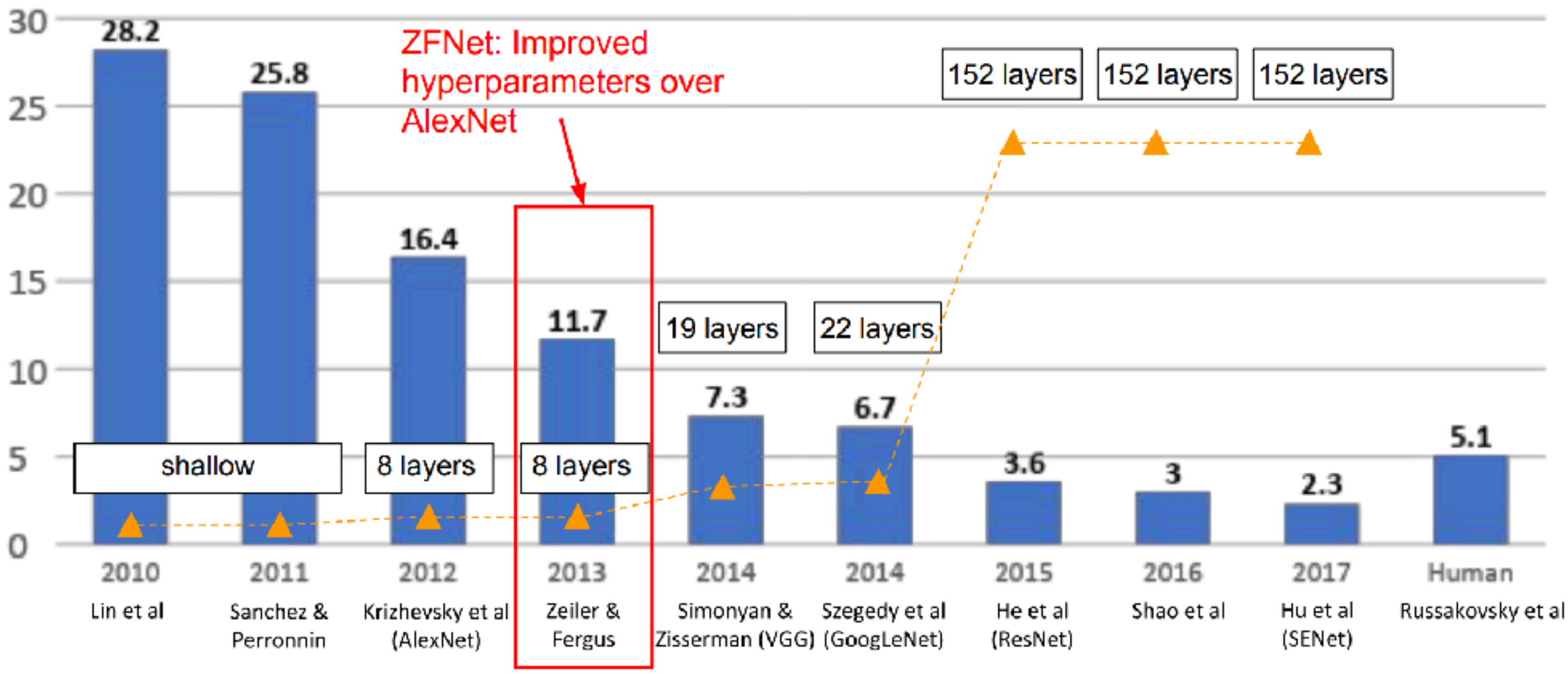
IMAGENET CHALLENGE WINNERS



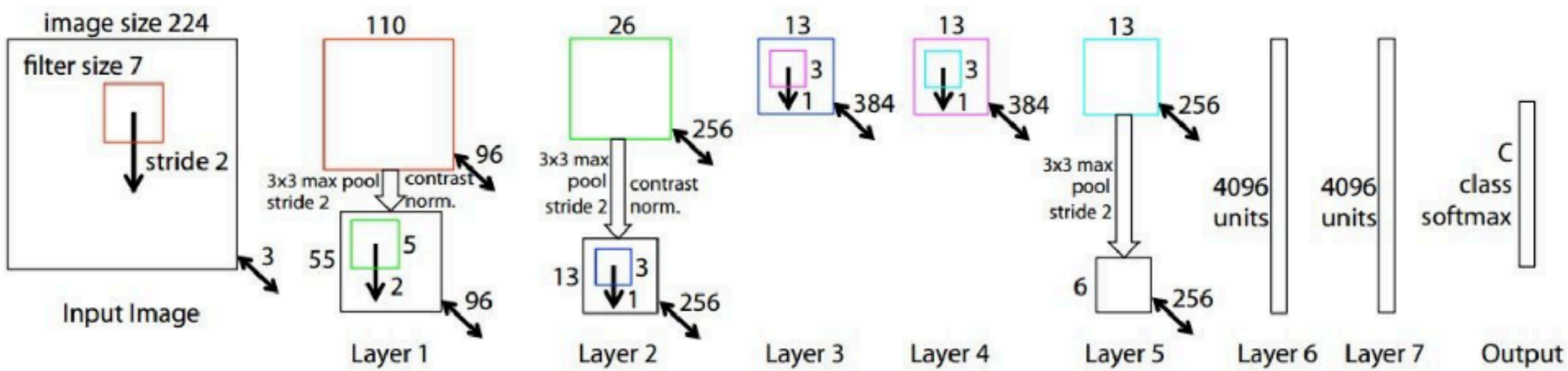
IMAGENET CHALLENGE WINNERS



IMAGENET CHALLENGE WINNERS



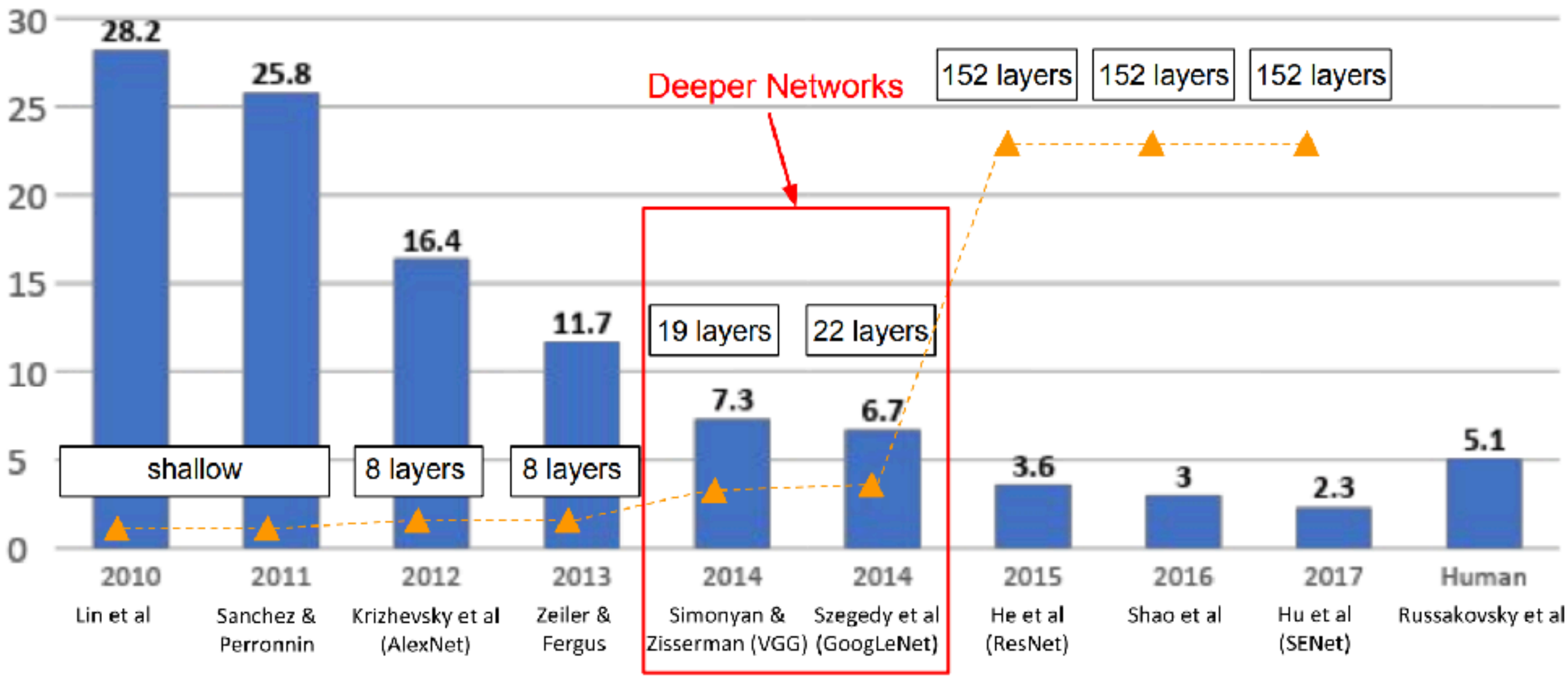
ZF-NET: 2013



AlexNet but:
CONV1: change from (11x11 stride 4) to (7x7 stride 2)
CONV3,4,5: instead of 384, 384, 256 filters use 512, 1024, 512

ImageNet top 5 error: 16.4% -> 11.7%

IMAGENET CHALLENGE WINNERS



VGGNET: 2014

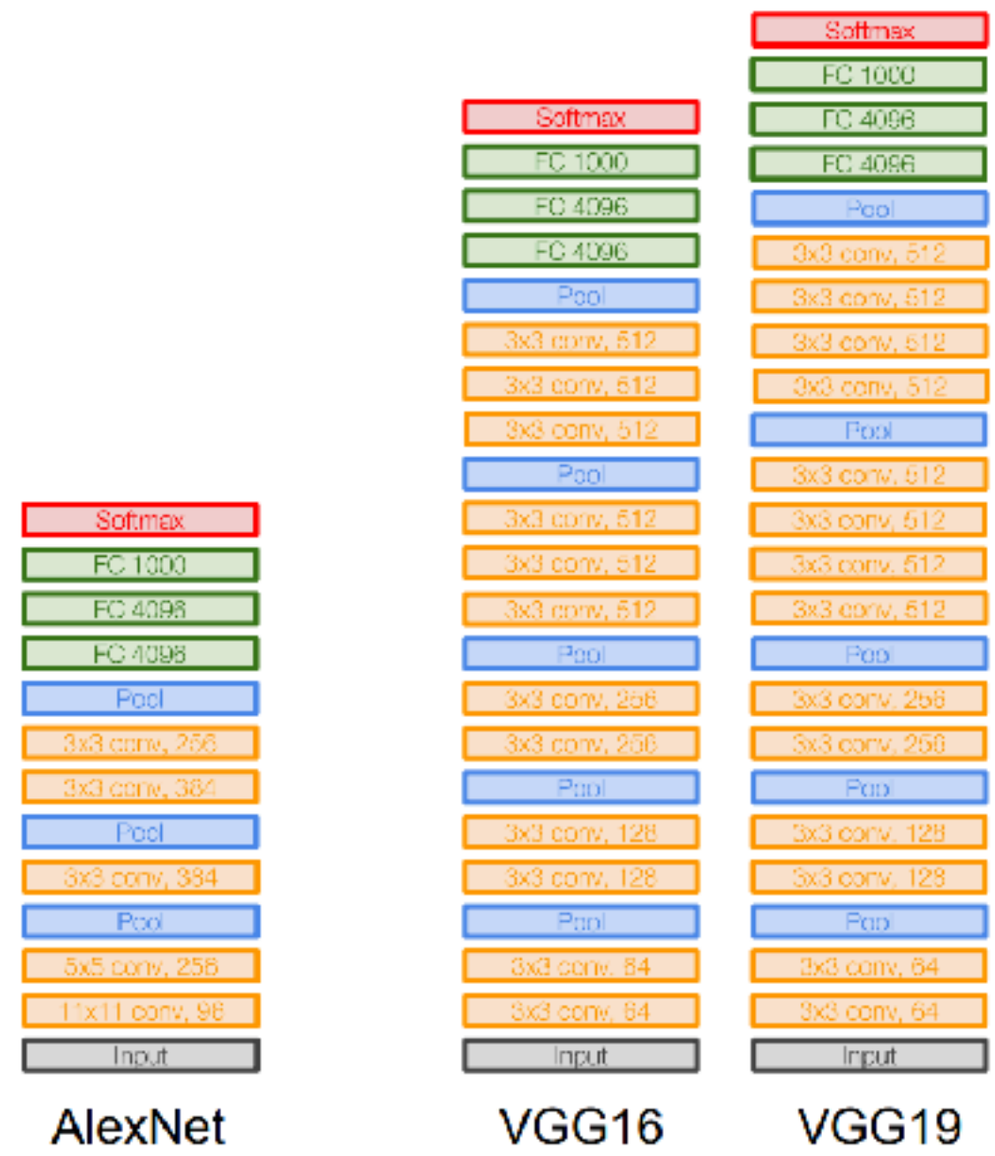
[Simonyan and Zisserman, 2014]

Small filters, Deeper networks

8 layers (AlexNet)
-> 16 - 19 layers (VGG16Net)

Only 3x3 CONV stride 1, pad 1
and 2x2 MAX POOL stride 2

11.7% top 5 error in ILSVRC'13
(ZFNet)
-> 7.3% top 5 error in ILSVRC'14



VGGNET: 2014

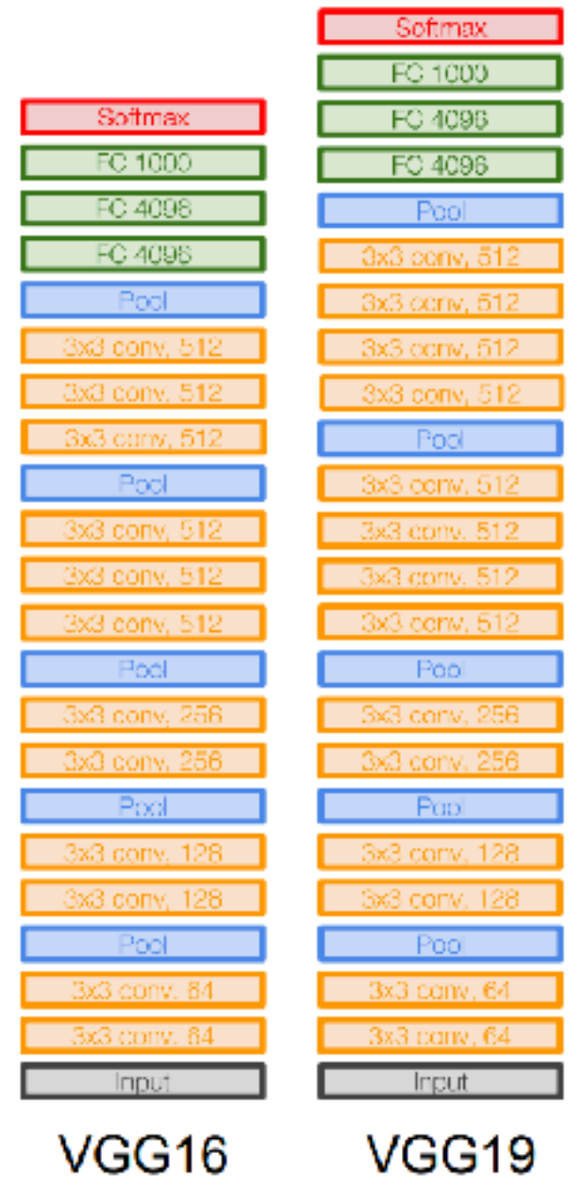
[Simonyan and Zisserman, 2014]

Q: Why use smaller filters? (3x3 conv)

Stack of three 3x3 conv (stride 1) layers has same **effective receptive field** as one 7x7 conv layer

But deeper, more non-linearities

And fewer parameters: $3 * (3^2C^2)$ vs. 7^2C^2 for C channels per layer



VGGNET: 2014

(not counting biases)

INPUT: [224x224x3] memory: 224*224*3=150K params: 0
CONV3-64: [224x224x64] memory: **224*224*64=3.2M** params: (3*3*3)*64 = 1,728
CONV3-64: [224x224x64] memory: **224*224*64=3.2M** params: (3*3*64)*64 = 36,864
POOL2: [112x112x64] memory: 112*112*64=800K params: 0
CONV3-128: [112x112x128] memory: 112*112*128=1.6M params: (3*3*64)*128 = 73,728
CONV3-128: [112x112x128] memory: 112*112*128=1.6M params: (3*3*128)*128 = 147,456
POOL2: [56x56x128] memory: 56*56*128=400K params: 0
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*128)*256 = 294,912
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*256)*256 = 589,824
POOL2: [28x28x256] memory: 28*28*256=200K params: 0
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*256)*512 = 1,179,648
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*512)*512 = 2,359,296
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*512)*512 = 2,359,296
POOL2: [14x14x512] memory: 14*14*512=100K params: 0
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
POOL2: [7x7x512] memory: 7*7*512=25K params: 0
FC: [1x1x4096] memory: 4096 params: 7*7*512*4096 = **102,760,448**
FC: [1x1x4096] memory: 4096 params: 4096*4096 = 16,777,216
FC: [1x1x1000] memory: 1000 params: 4096*1000 = 4,096,000

Note:

Most memory is in early CONV

Most params are in late FC

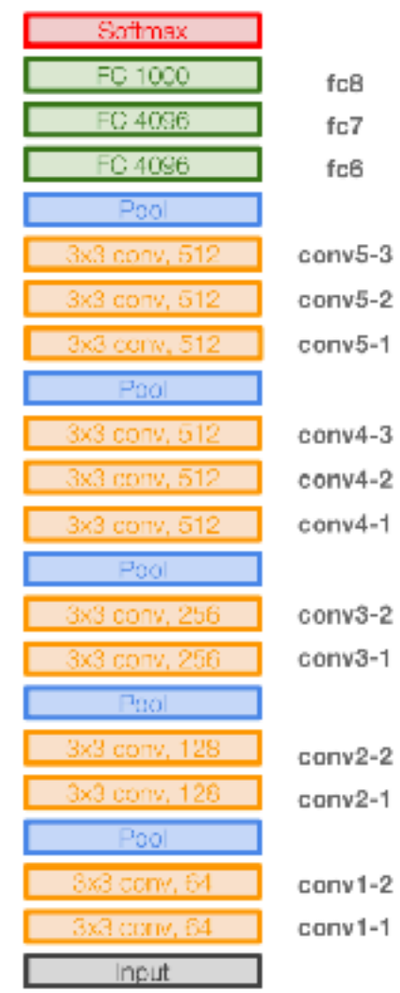
TOTAL memory: 24M * 4 bytes ~= 96MB / image (only forward! ~*2 for bwd)
TOTAL params: 138M parameters

VGGNET: 2014

(not counting biases)

INPUT: [224x224x3] memory: 224*224*3=150K params: 0
CONV3-64: [224x224x64] memory: 224*224*64=3.2M params: (3*3*3)*64 = 1,728
CONV3-64: [224x224x64] memory: 224*224*64=3.2M params: (3*3*64)*64 = 36,864
POOL2: [112x112x64] memory: 112*112*64=800K params: 0
CONV3-128: [112x112x128] memory: 112*112*128=1.6M params: (3*3*64)*128 = 73,728
CONV3-128: [112x112x128] memory: 112*112*128=1.6M params: (3*3*128)*128 = 147,456
POOL2: [56x56x128] memory: 56*56*128=400K params: 0
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*128)*256 = 294,912
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*256)*256 = 589,824
CONV3-256: [56x56x256] memory: 56*56*256=800K params: (3*3*256)*256 = 589,824
POOL2: [28x28x256] memory: 28*28*256=200K params: 0
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*256)*512 = 1,179,648
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*512)*512 = 2,359,296
CONV3-512: [28x28x512] memory: 28*28*512=400K params: (3*3*512)*512 = 2,359,296
POOL2: [14x14x512] memory: 14*14*512=100K params: 0
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512] memory: 14*14*512=100K params: (3*3*512)*512 = 2,359,296
POOL2: [7x7x512] memory: 7*7*512=25K params: 0
FC: [1x1x4096] memory: 4096 params: 7*7*512*4096 = 102,760,448
FC: [1x1x4096] memory: 4096 params: 4096*4096 = 16,777,216
FC: [1x1x1000] memory: 1000 params: 4096*1000 = 4,096,000

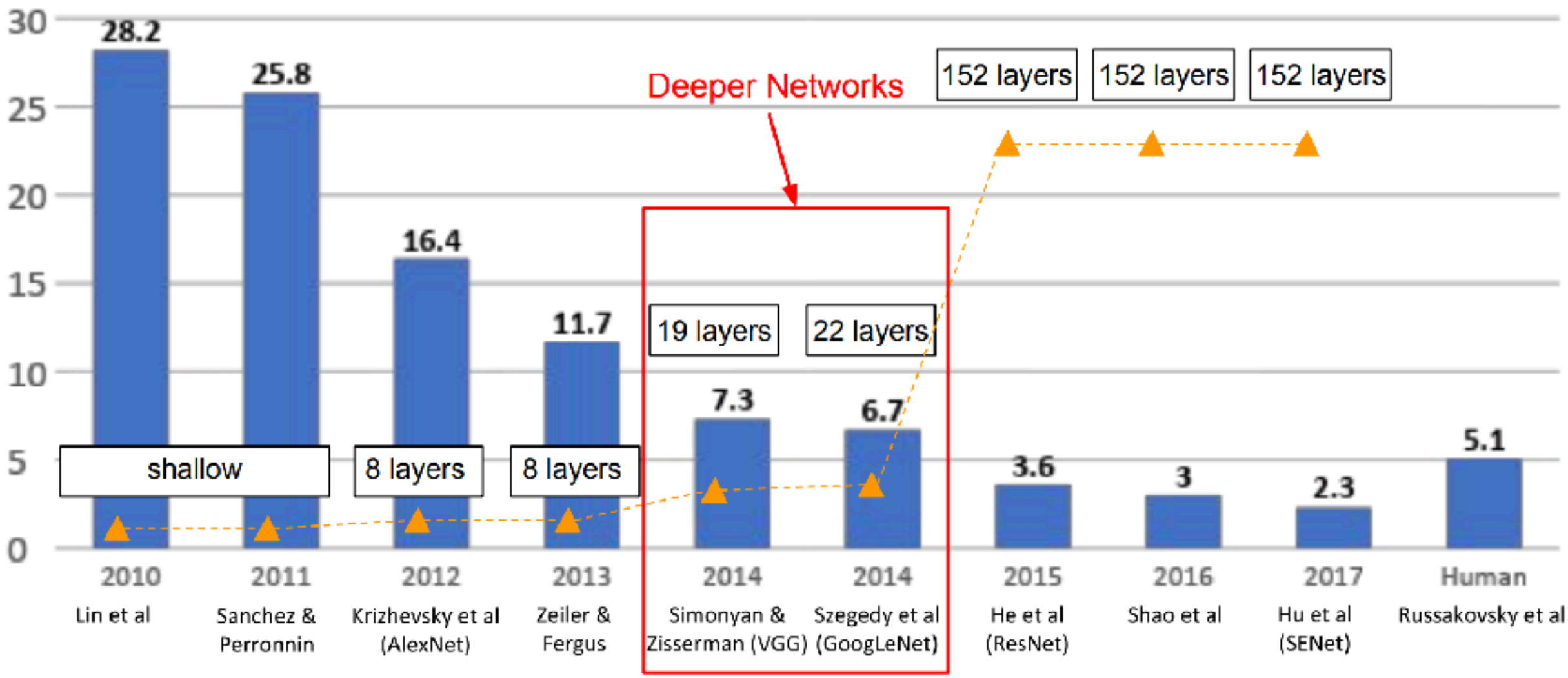
TOTAL memory: 24M * 4 bytes ~= 96MB / image (only forward! ~*2 for bwd)
TOTAL params: 138M parameters



VGG16

Common names

IMAGENET CHALLENGE WINNERS

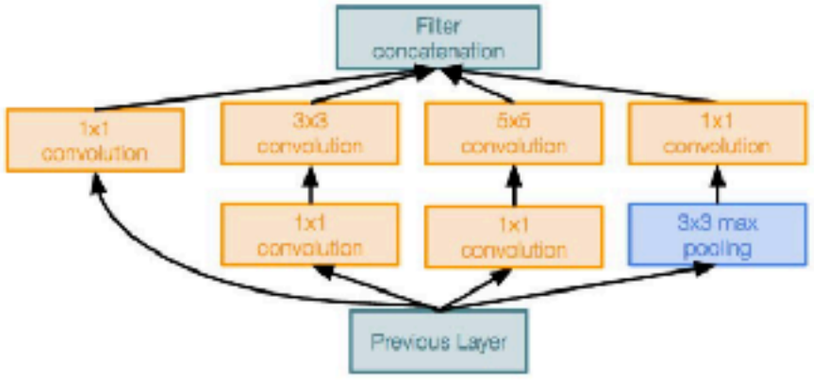


GOOGLENET: 2014

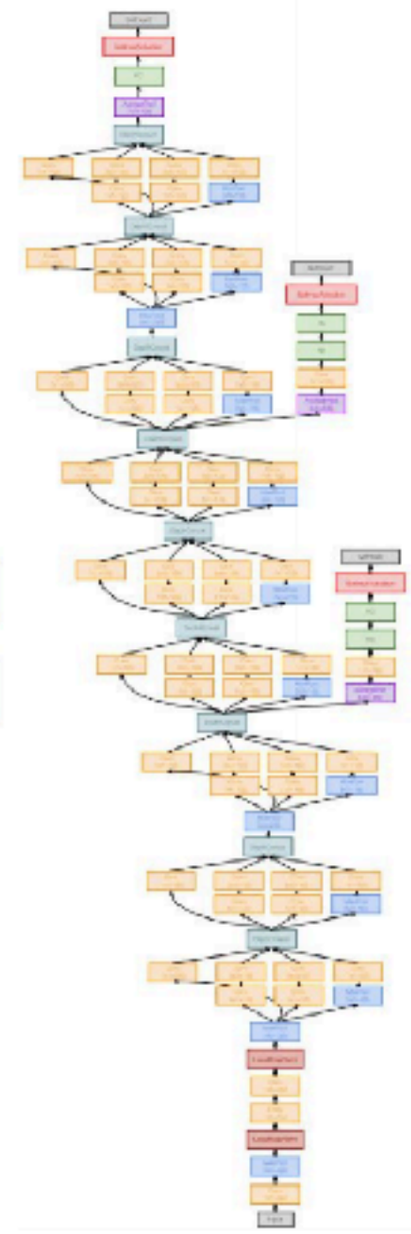
[Szegedy et al., 2014]

Deeper networks, with computational efficiency

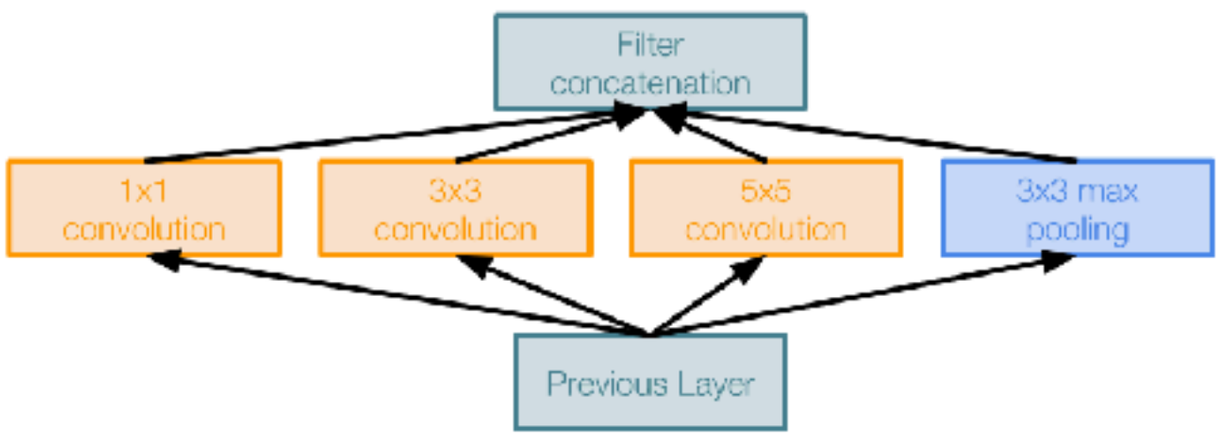
- 22 layers
- Efficient "Inception" module
- No FC layers
- Only 5 million parameters!
  12x less than AlexNet
- ILSVRC'14 classification winner
  (6.7% top 5 error)



Inception module



GOOGLENET: INCEPTION MODULE



Naive Inception module

Apply parallel filter operations on the input from previous layer:

- Multiple receptive field sizes for convolution (1x1, 3x3, 5x5)
- Pooling operation (3x3)

Concatenate all filter outputs together depth-wise

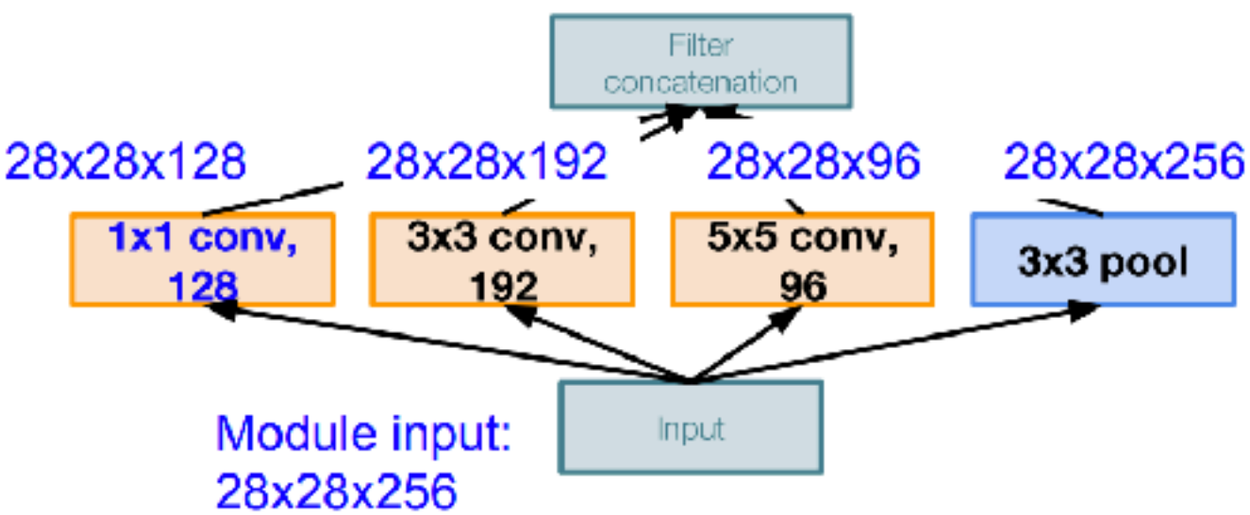
GOOGLENET: INCEPTION MODULE

Q: What is the problem with this?
[Hint: Computational complexity]

Example:

Q3:What is output size after filter concatenation?

28x28x(128+192+96+256) = 28x28x672



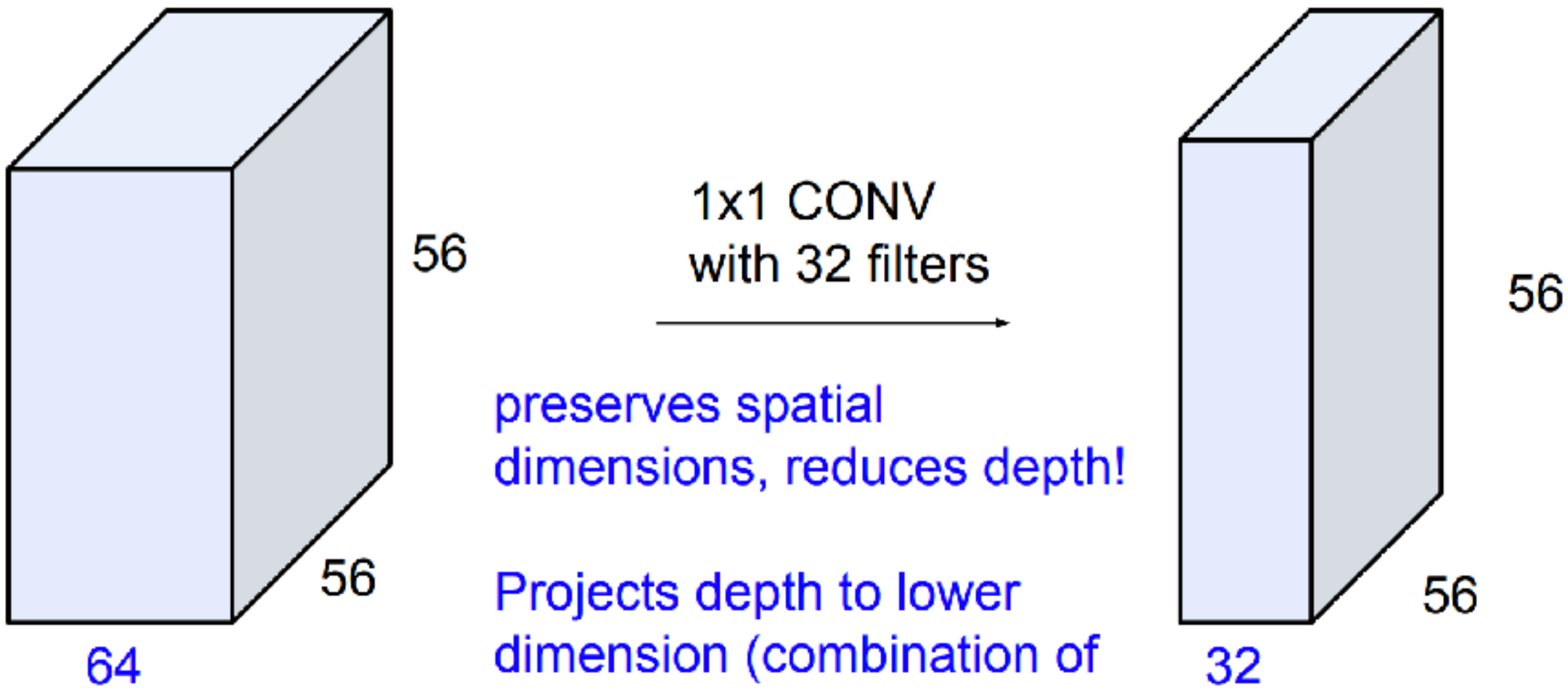
Naive Inception module

Conv Ops:
[1x1 conv, 128] 28x28x128x1x1x256
[3x3 conv, 192] 28x28x192x3x3x256
[5x5 conv, 96] 28x28x96x5x5x256
Total: 854M ops

Very expensive compute

Pooling layer also preserves feature depth, which means total depth after concatenation can only grow at every layer!

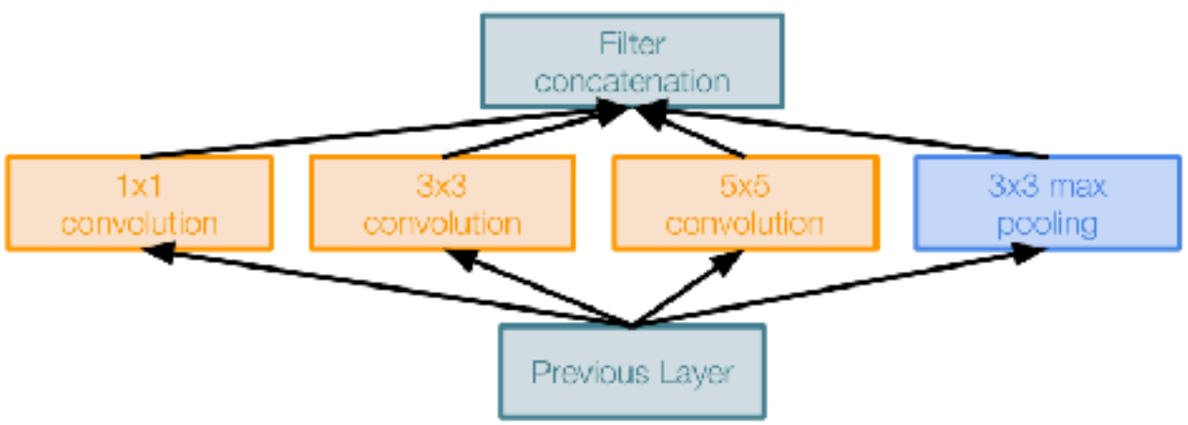
(1 X 1) CONVOLUTIONAL LAYER



preserves spatial dimensions, reduces depth!

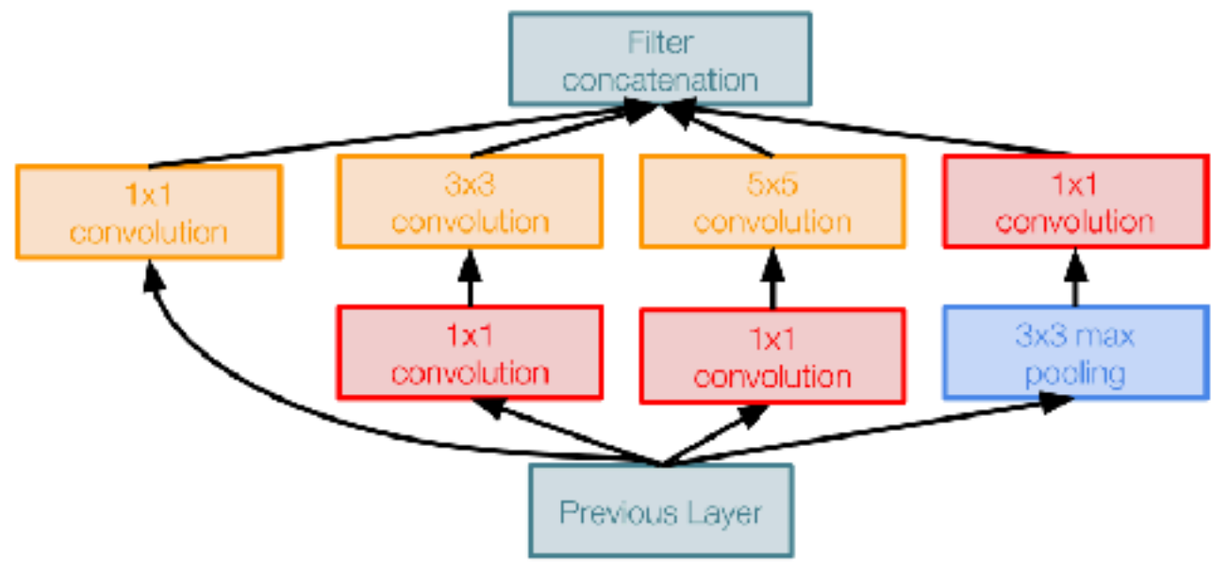
Projects depth to lower dimension (combination of feature maps)

GOOGLENET: INCEPTION MODULE



Naive Inception module

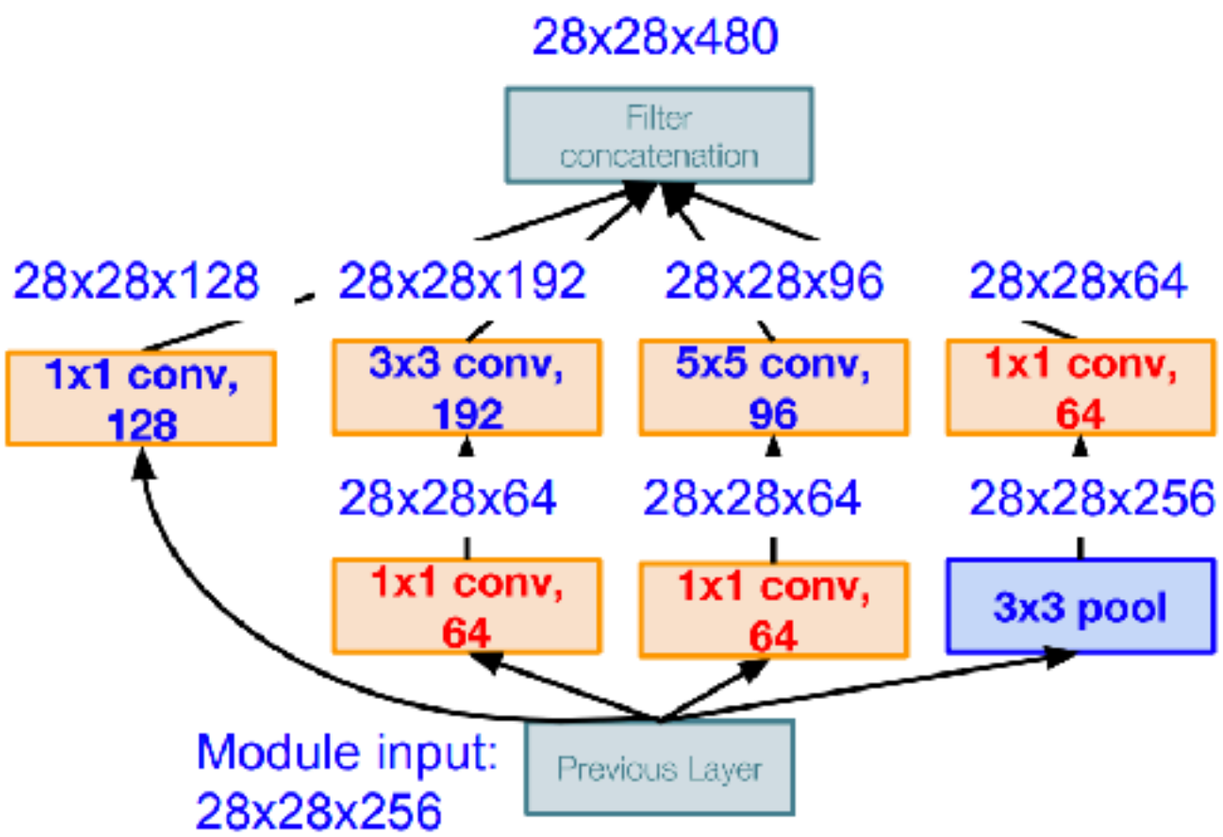
1x1 conv "bottleneck" layers



Inception module with dimension reduction

GOOGLENET: INCEPTION MODULE

[Szegedy et al., 2014]



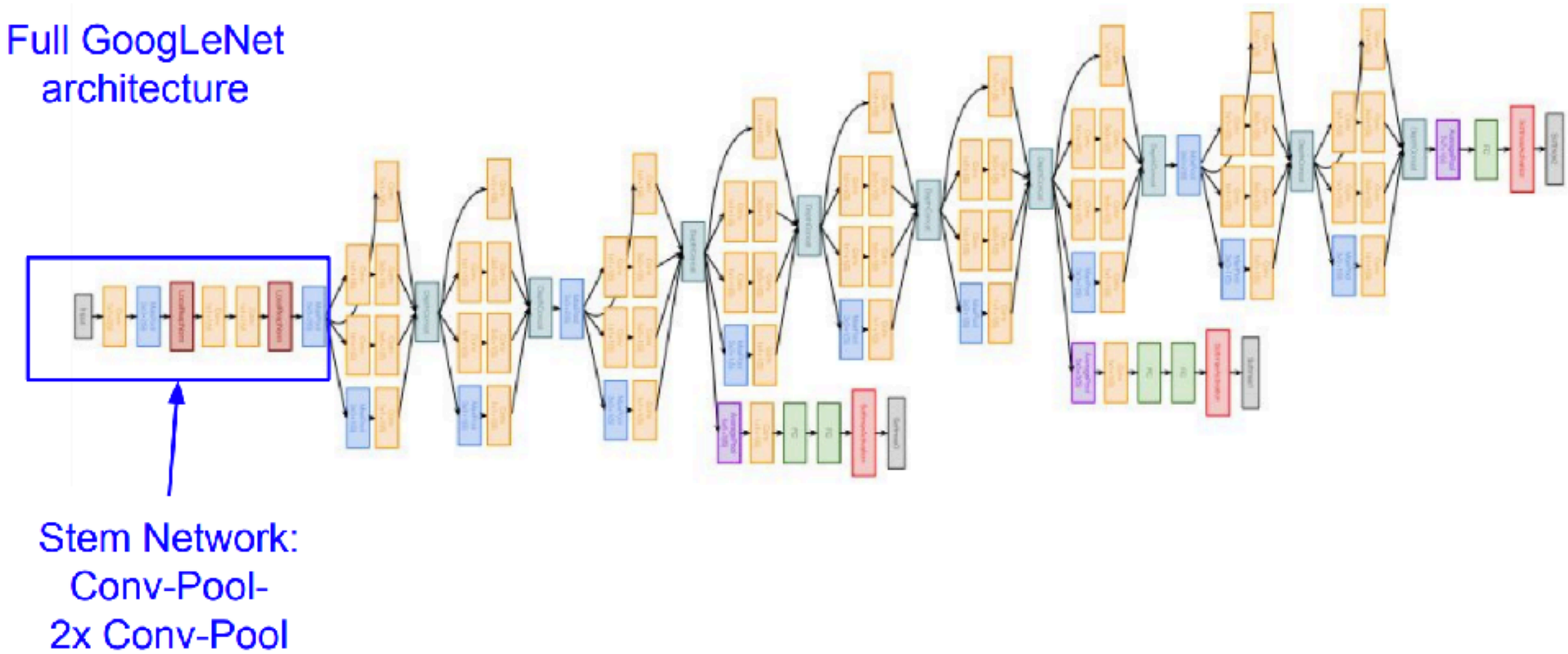
Inception module with dimension reduction

Using same parallel layers as naive example, and adding "1x1 conv, 64 filter" bottlenecks:

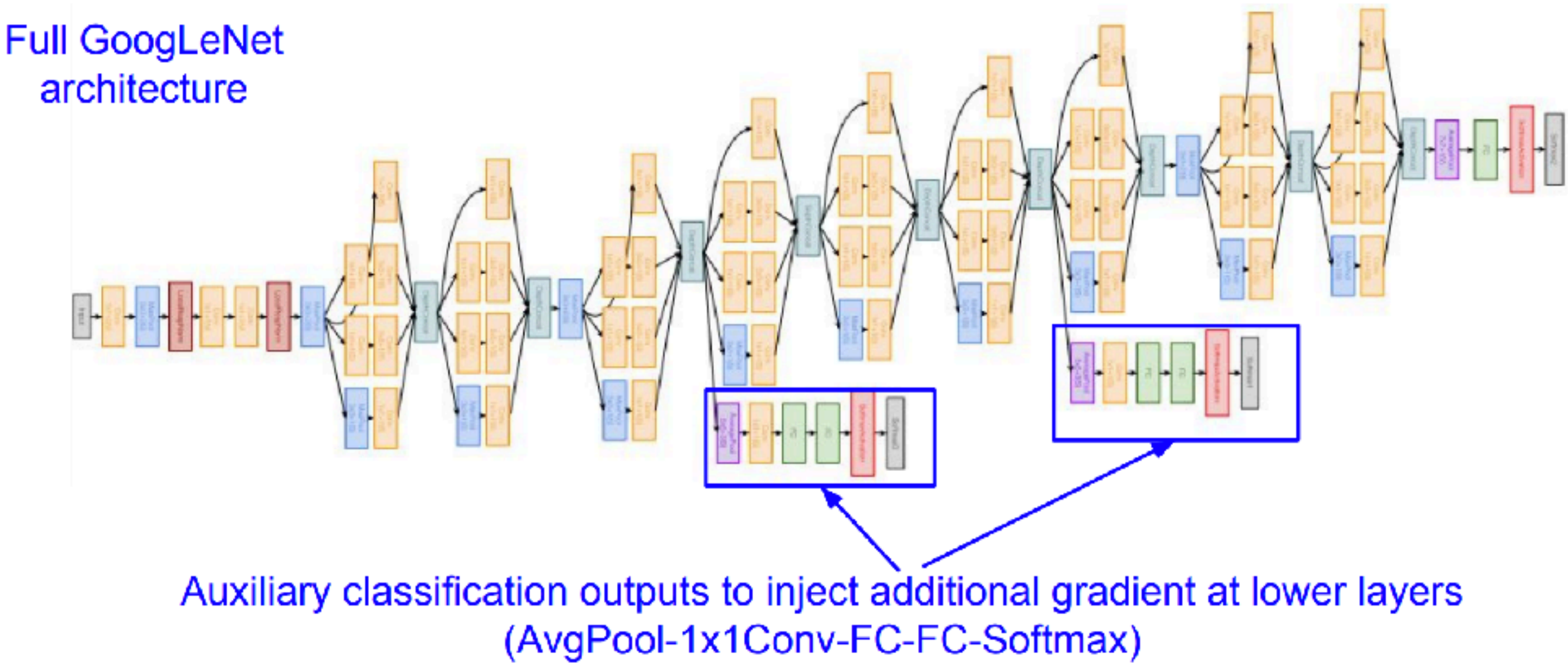
Conv Ops:
[1x1 conv, 64] 28x28x64x1x1x256
[1x1 conv, 64] 28x28x64x1x1x256
[1x1 conv, 128] 28x28x128x1x1x256
[3x3 conv, 192] 28x28x192x3x3x64
[5x5 conv, 96] 28x28x96x5x5x64
[1x1 conv, 64] 28x28x64x1x1x256
Total: 358M ops

Compared to 854M ops for naive version
Bottleneck can also reduce depth after pooling layer

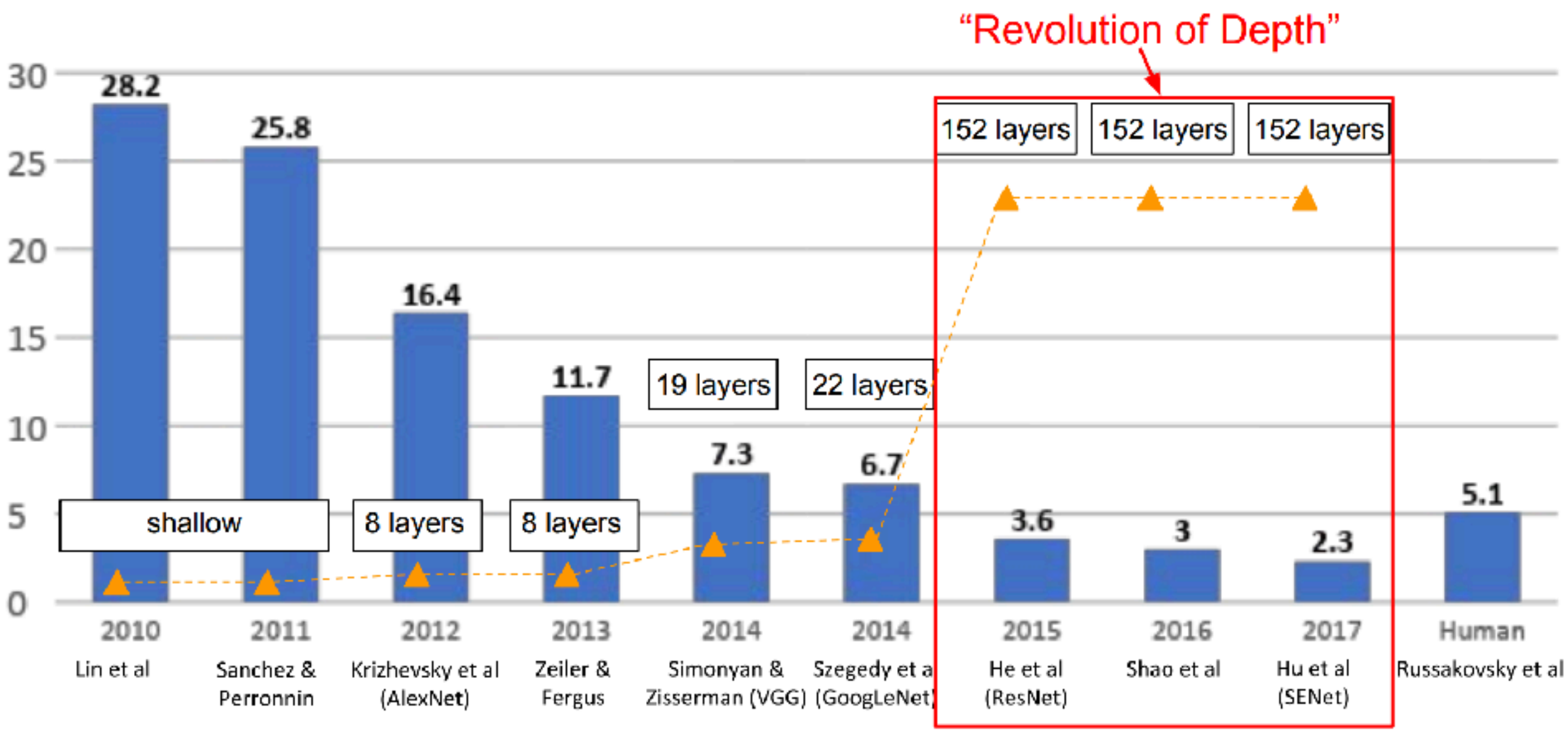
GOOGLENET: FULL ARCHITECTURE



GOOGLENET: FULL ARCHITECTURE



IMAGENET CHALLENGE WINNERS

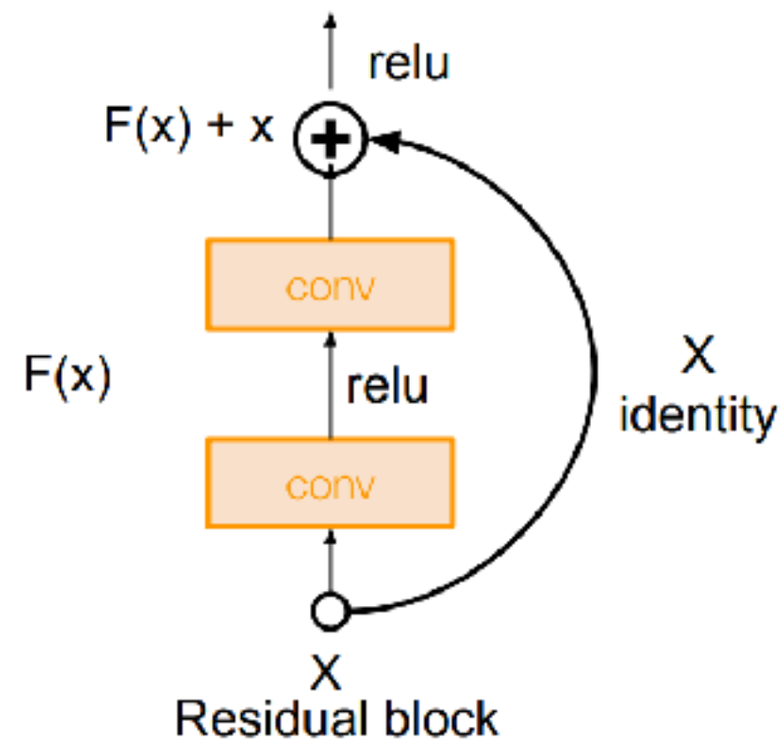


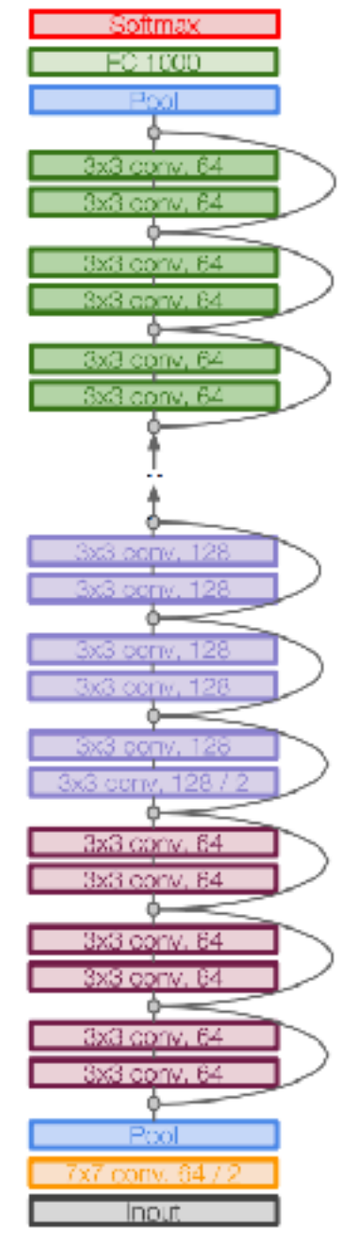
RESNET: 2015

[He et al., 2015]

Very deep networks using residual connections

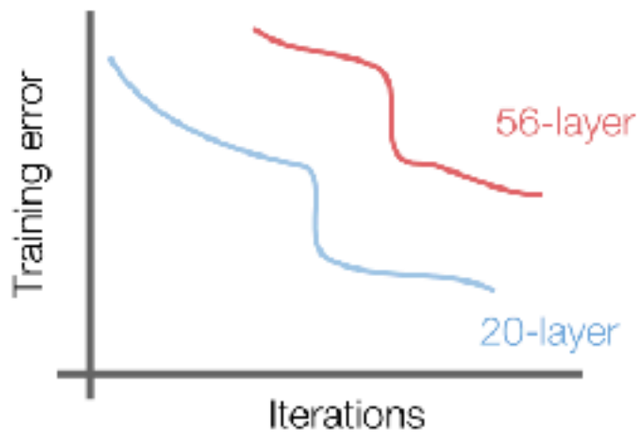
- 152-layer model for ImageNet
- ILSVRC'15 classification winner (3.57% top 5 error)
- Swept all classification and detection competitions in ILSVRC'15 and COCO'15!

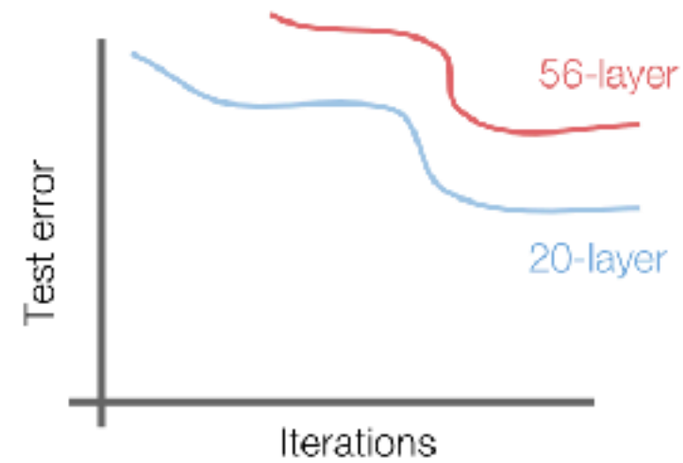




WHAT IS THE PROBLEM WITH DEEPER MODELS?

What happens when we continue stacking deeper layers on a “plain” convolutional neural network?





56-layer model performs worse on both training and test error
-> The deeper model performs worse, but it’s not caused by overfitting!

WHAT IS THE SOLUTION?

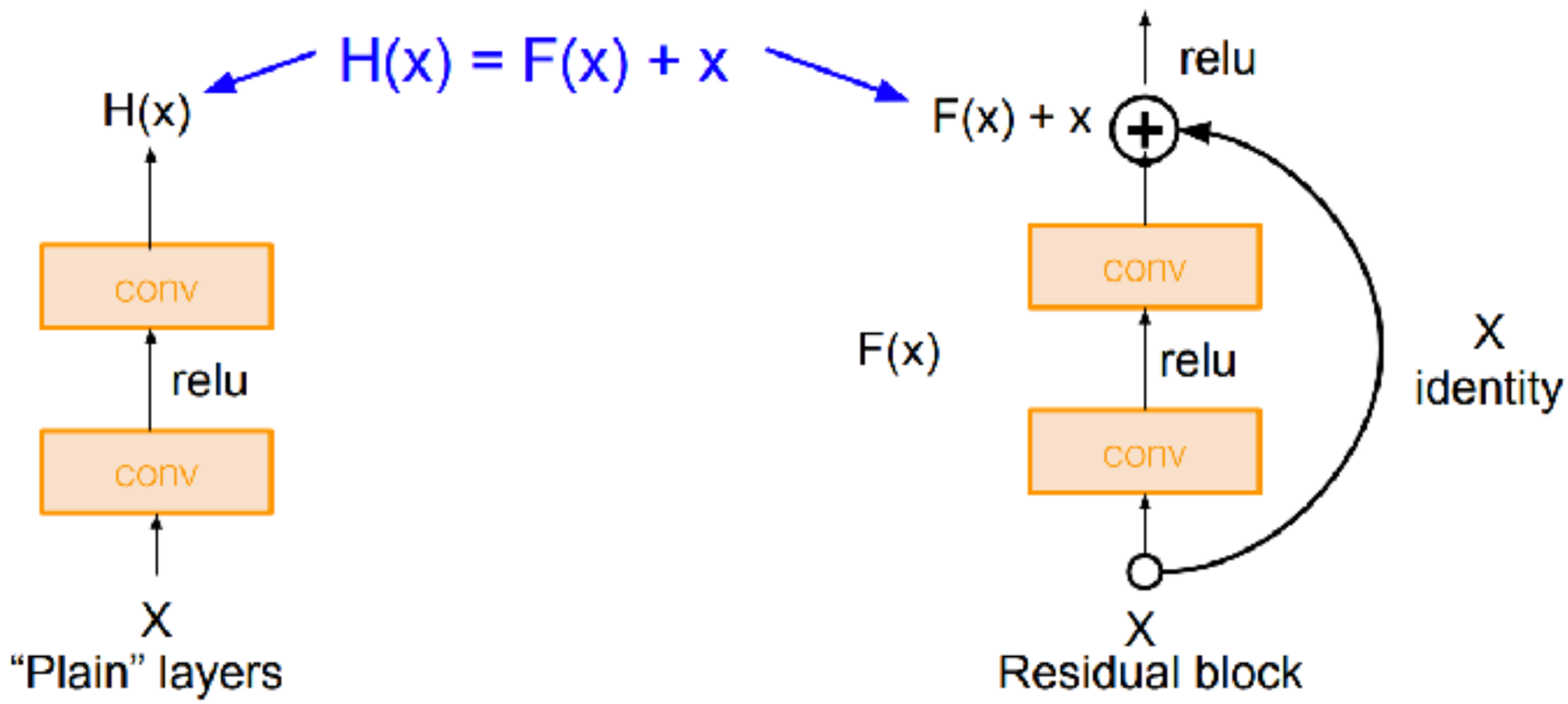
Hypothesis: the problem is an *optimization* problem, deeper models are harder to optimize

The deeper model should be able to perform at least as well as the shallower model.

A solution by construction is copying the learned layers from the shallower model and setting additional layers to identity mapping.

WHAT IS THE SOLUTION?

Solution: Use network layers to fit a residual mapping instead of directly trying to fit a desired underlying mapping



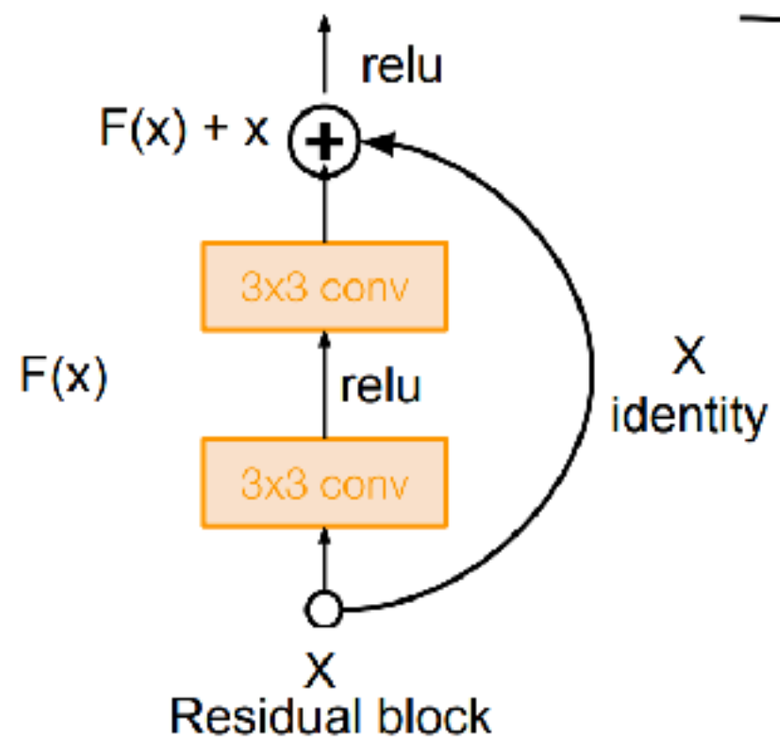
Use layers to fit residual F(x) = H(x) - x instead of H(x) directly

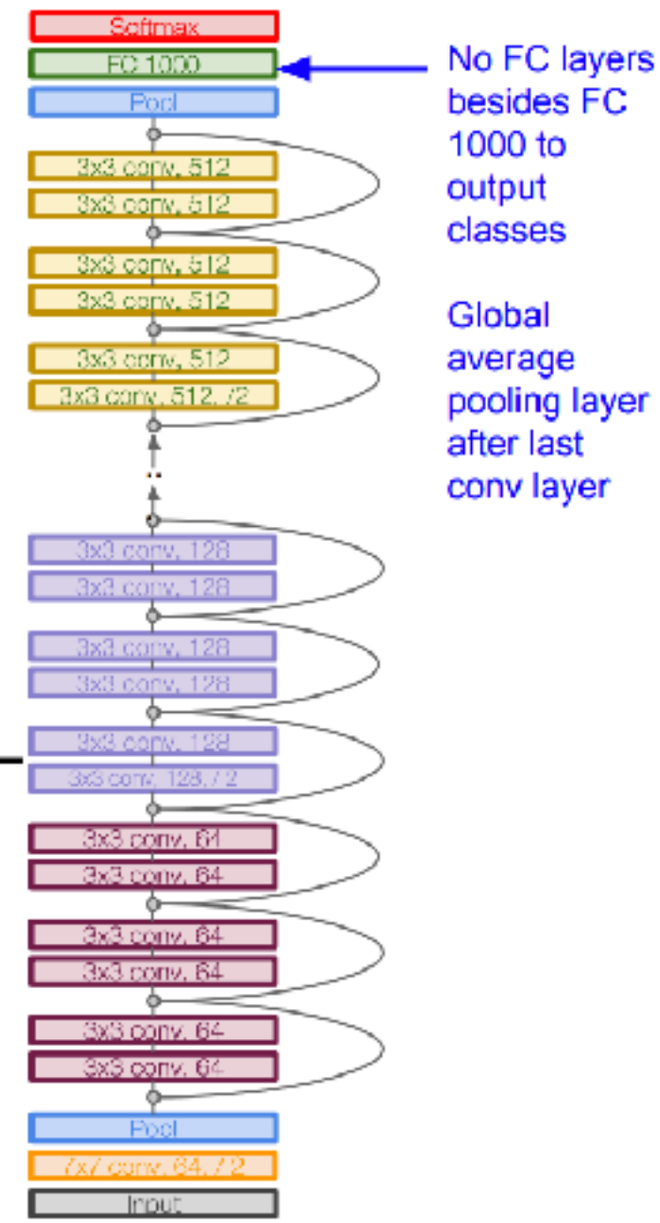
RESNET: 2015

[He et al., 2015]

Full ResNet architecture:

- Stack residual blocks
- Every residual block has two 3x3 conv layers
- Periodically, double # of filters and downsample spatially using stride 2 (/2 in each dimension)
- Additional conv layer at the beginning
- No FC layers at the end (only FC 1000 to output classes)

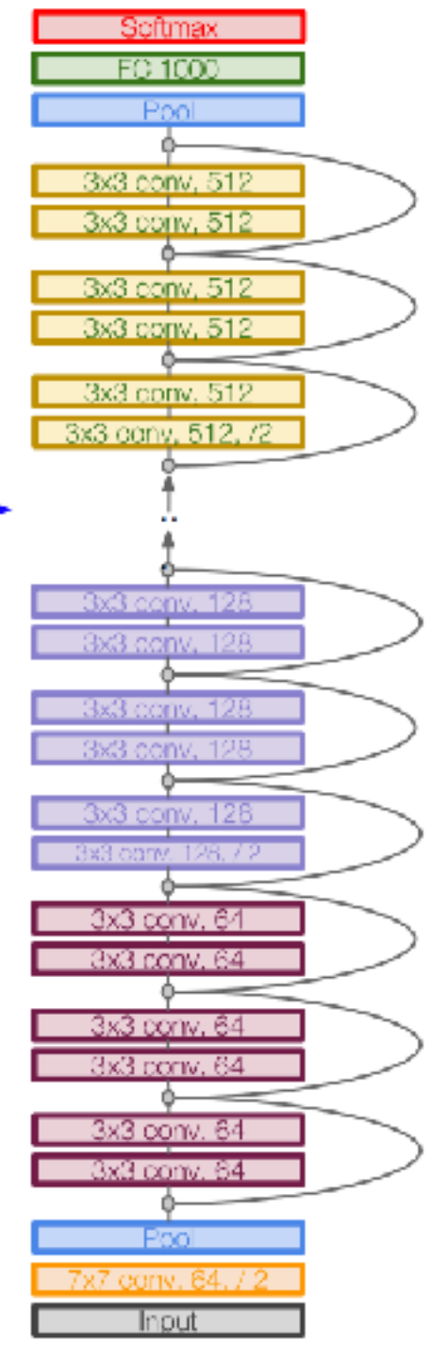




RESNET: 2015

[He et al., 2015]

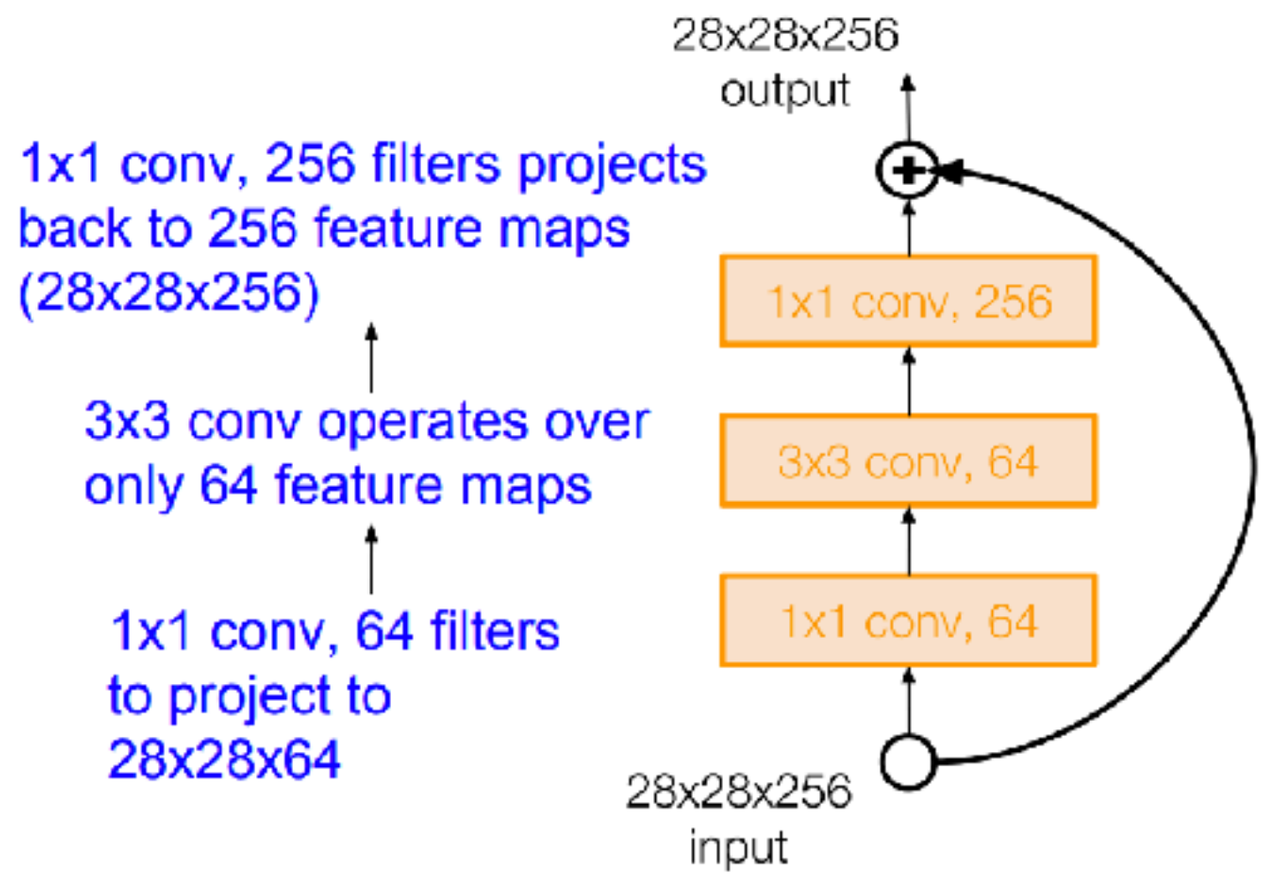
Total depths of 34, 50, 101, or 152 layers for ImageNet



RESNET: 2015

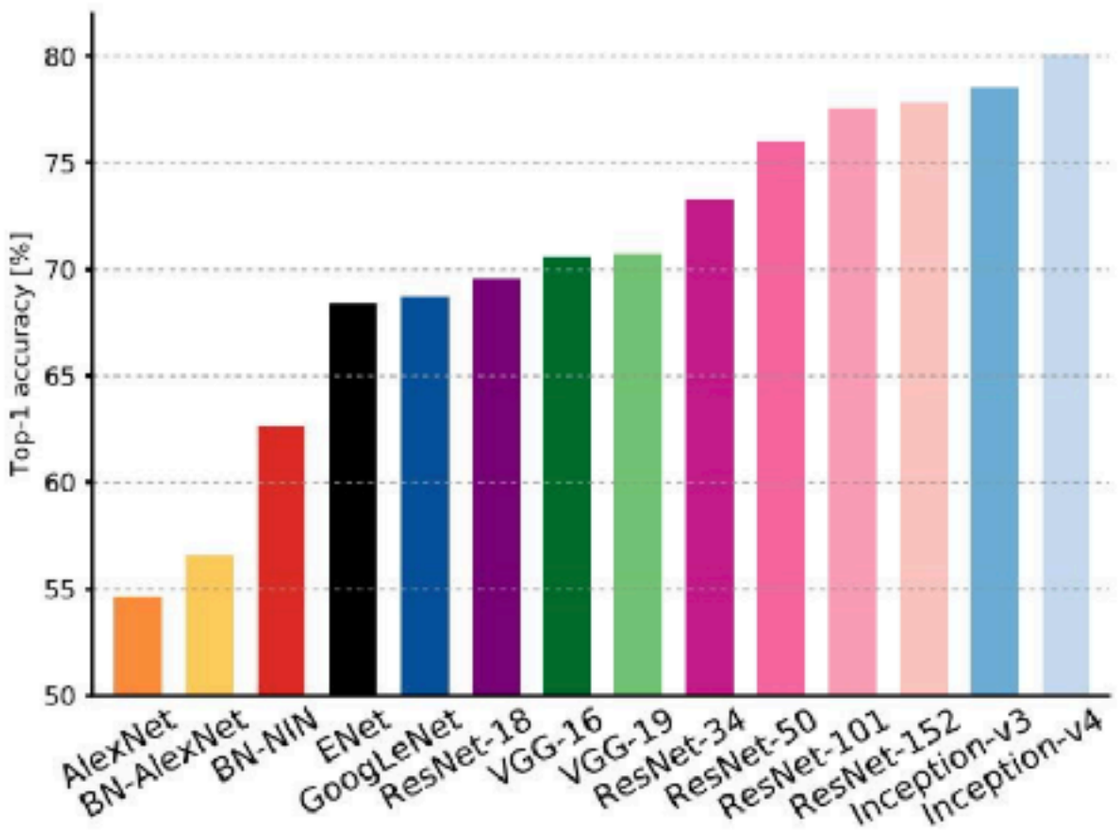
[He et al., 2015]

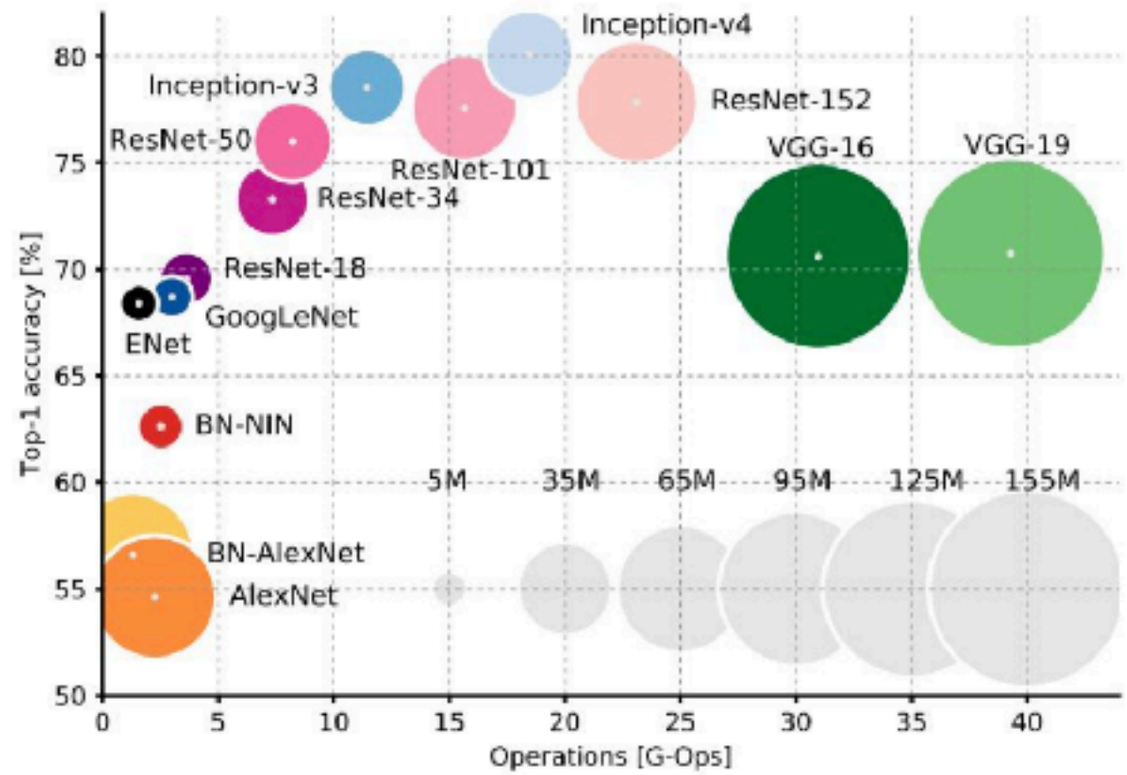
For deeper networks (ResNet-50+), use "bottleneck" layer to improve efficiency (similar to GoogLeNet)



COMPARISON OF MODELS

Comparing complexity...





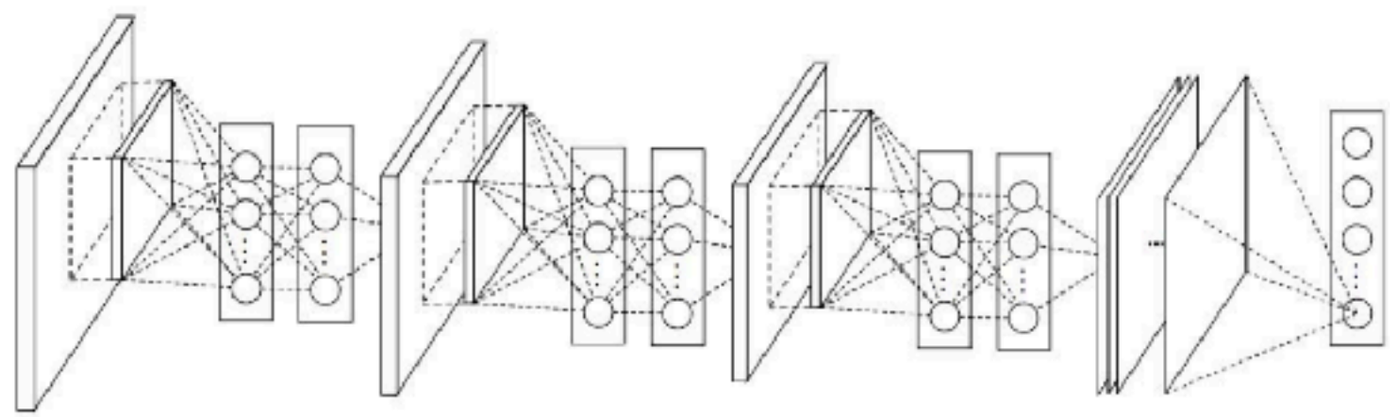
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

OTHER POPULAR MODELS

NETWORK IN NETWORK (NIN)

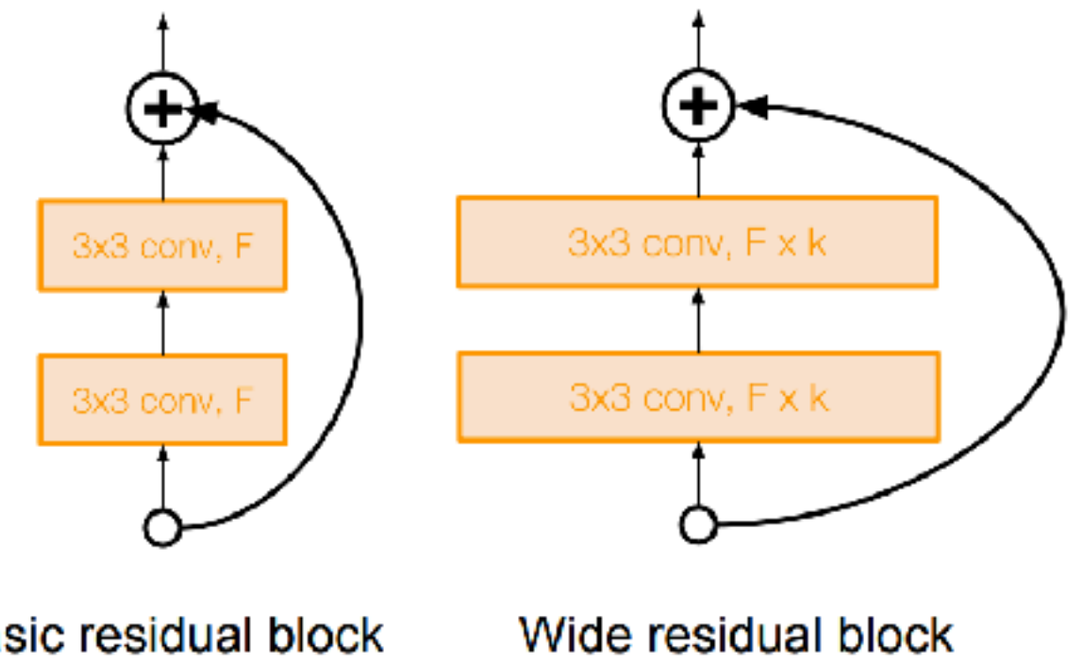
- Mlpconv layer with "micronetwork" within each conv layer to compute more abstract features for local patches
- Micronetwork uses multilayer perceptron (FC, i.e. 1x1 conv layers)
- Precursor to GoogLeNet and ResNet "bottleneck" layers
- Philosophical inspiration for GoogLeNet

(a) Linear convolution layer (b) Mlpconv layer



WIDE RESIDUAL NETWORK

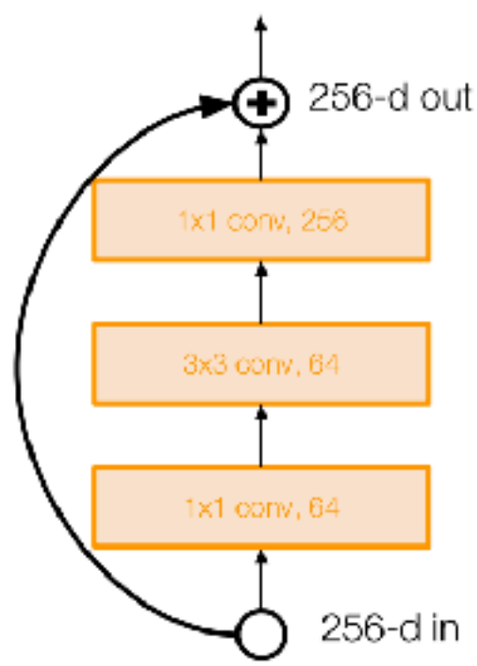
- Argues that residuals are the important factor, not depth
- User wider residual blocks (F x k filters instead of F filters in each layer)
- 50-layer wide ResNet outperforms 152-layer original ResNet
- Increasing width instead of depth more computationally efficient (parallelizable)

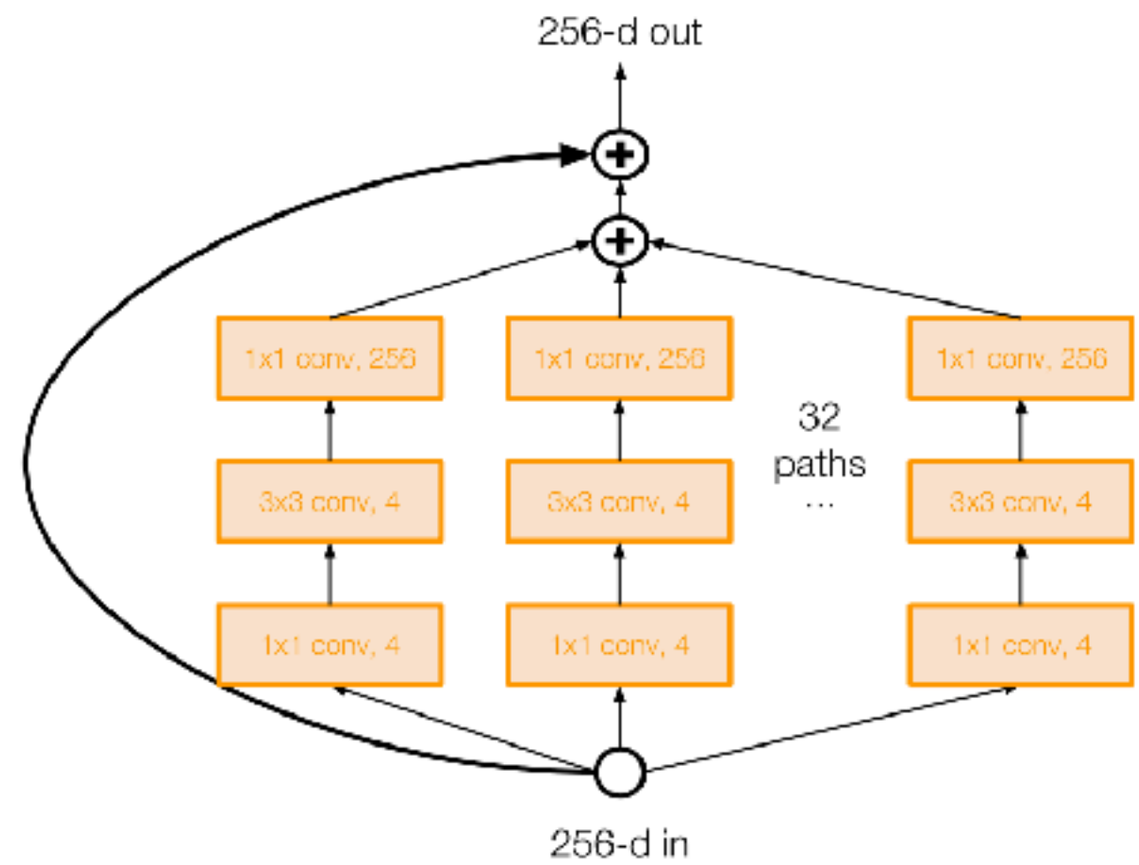


RESNEXT: 2016

[Xie et al. 2016]

- Also from creators of ResNet
- Increases width of residual block through multiple parallel pathways ("cardinality")
- Parallel pathways similar in spirit to Inception module





MODEL ENSEMBLE ALWAYS WORKS!

"Good Practices for Deep Feature Fusion"

[Shao et al. 2016]

- Multi-scale ensembling of Inception, Inception-Resnet, Resnet, Wide Resnet models
- ILSVRC'16 classification winner

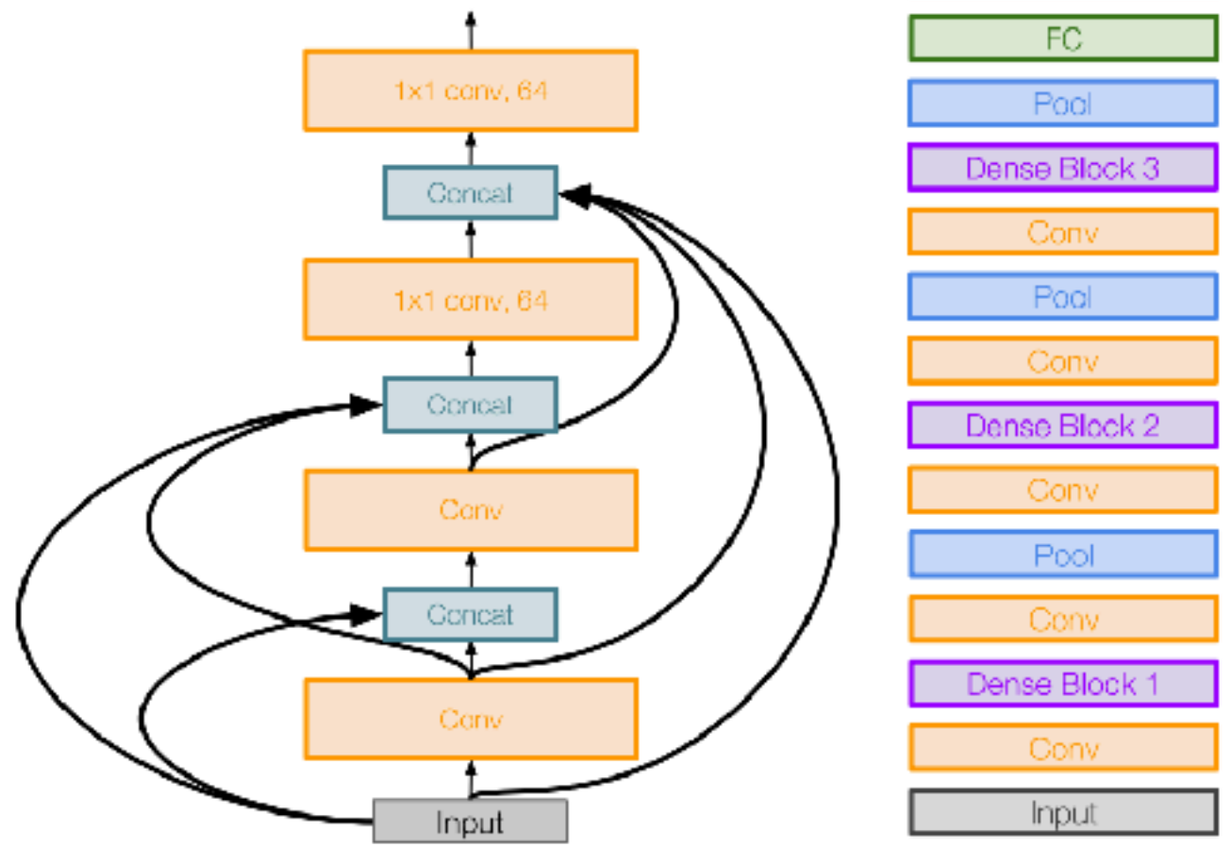
	Inception-v3	Inception-v4	Inception-Resnet-v2	Resnet-200	Wrn-68-3	Fusion（Val.）	Fusion（Test）
Err. (%)	4.20	4.01	3.52	4.26	4.65	2.92 (-0.6)	2.99

DENSENET: 2017

Densely Connected Convolutional Networks

[Huang et al. 2017]

- Dense blocks where each layer is connected to every other layer in feedforward fashion
- Alleviates vanishing gradient, strengthens feature propagation, encourages feature reuse



Dense Block

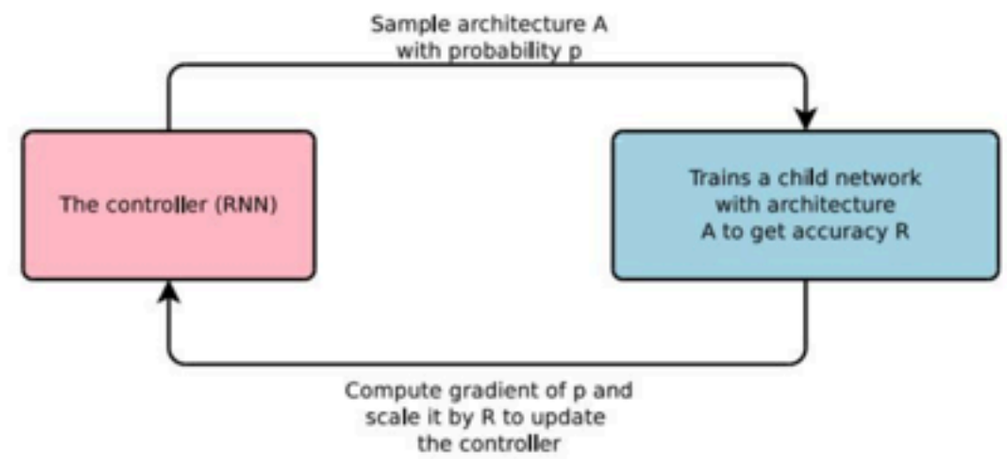
NASNET

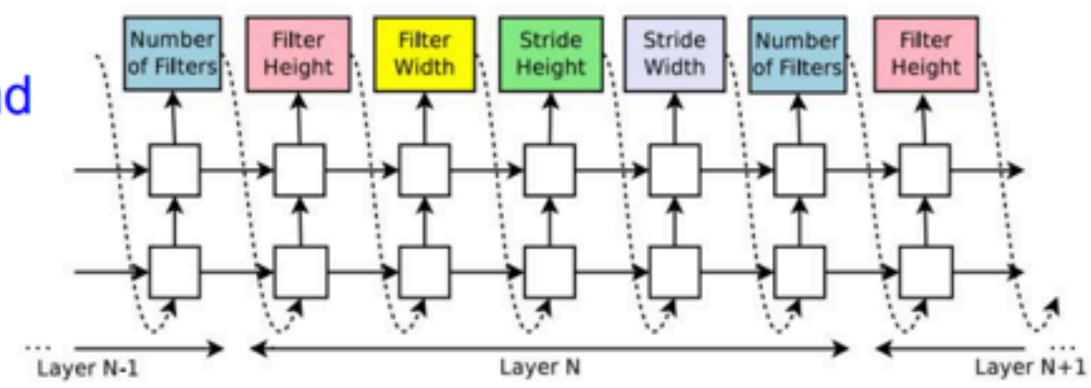
Meta-learning: Learning to learn network architectures...

Neural Architecture Search with Reinforcement Learning (NAS)

[Zoph et al. 2016]

- "Controller" network that learns to design a good network architecture (output a string corresponding to network design)
- Iterate:
  1) Sample an architecture from search space
  2) Train the architecture to get a "reward" R corresponding to accuracy
  3) Compute gradient of sample probability, and scale by R to perform controller parameter update (i.e. increase likelihood of good architecture being sampled, decrease likelihood of bad architecture)





HANDS ON

- Can we implement CNN models in cloud?