CNN Architectures

Machine Learning

(Largely based on slides from Fei-Fei Li & Justin Johnson & Serena Yeung)

Prof. Sandra Avila

Institute of Computing (IC/Unicamp)

MC886, October 21, 2019

Today's Agenda

- CNN Architectures
  - LeNet (1998)
  - AlexNet (2012)
  - ZFNet (2013)
  - VGGNet (2014)
  - GoogLeNet (2014)
  - ResNet (2015)

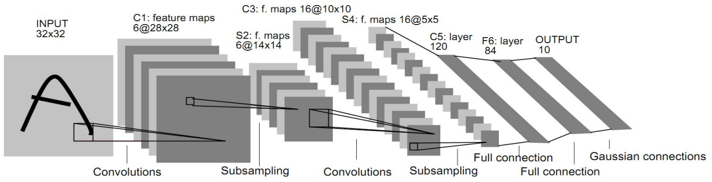
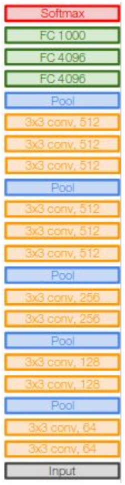
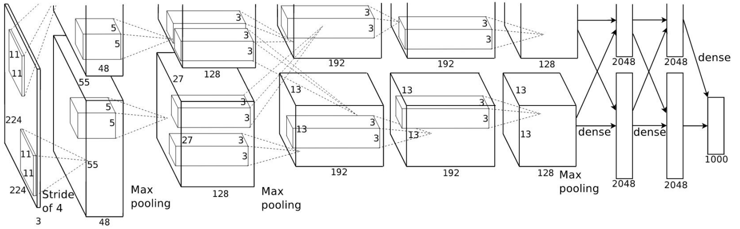
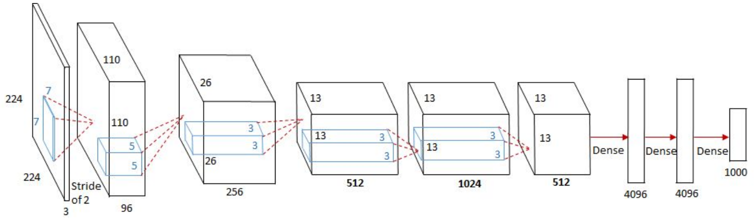
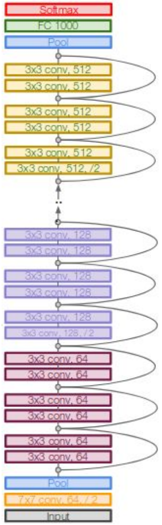


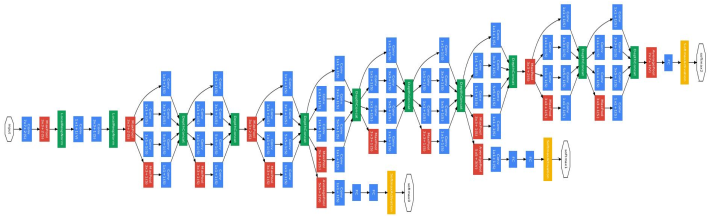
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.











Today’s Agenda

- CNN Architectures
    - LeNet (1998)
    - AlexNet (2012)
    - ZFNet (2013)
    - VGGNet (2014)
    - GoogLeNet (2014)
    - ResNet (2015)

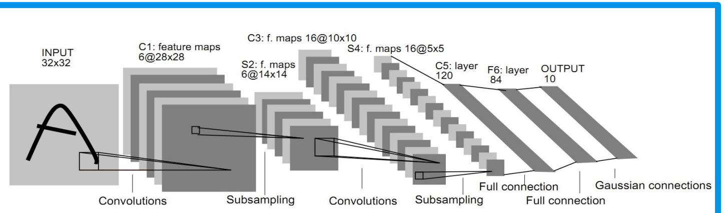
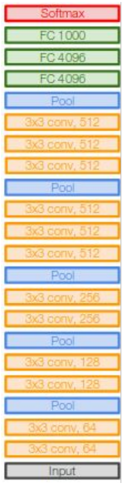
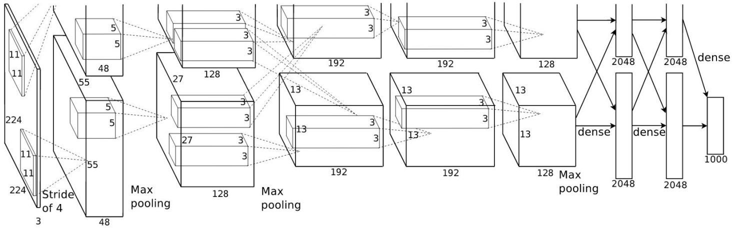
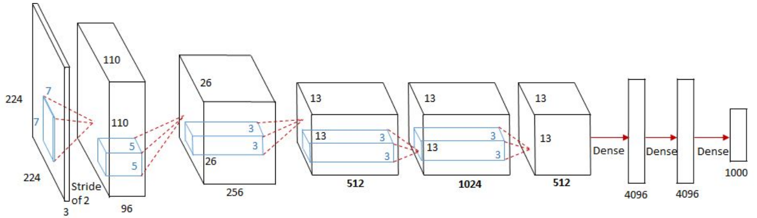
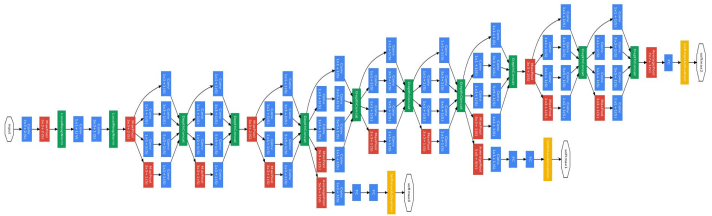


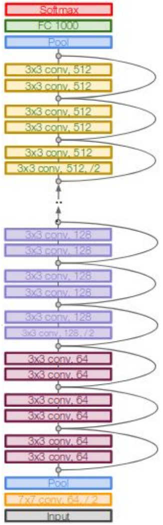
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.



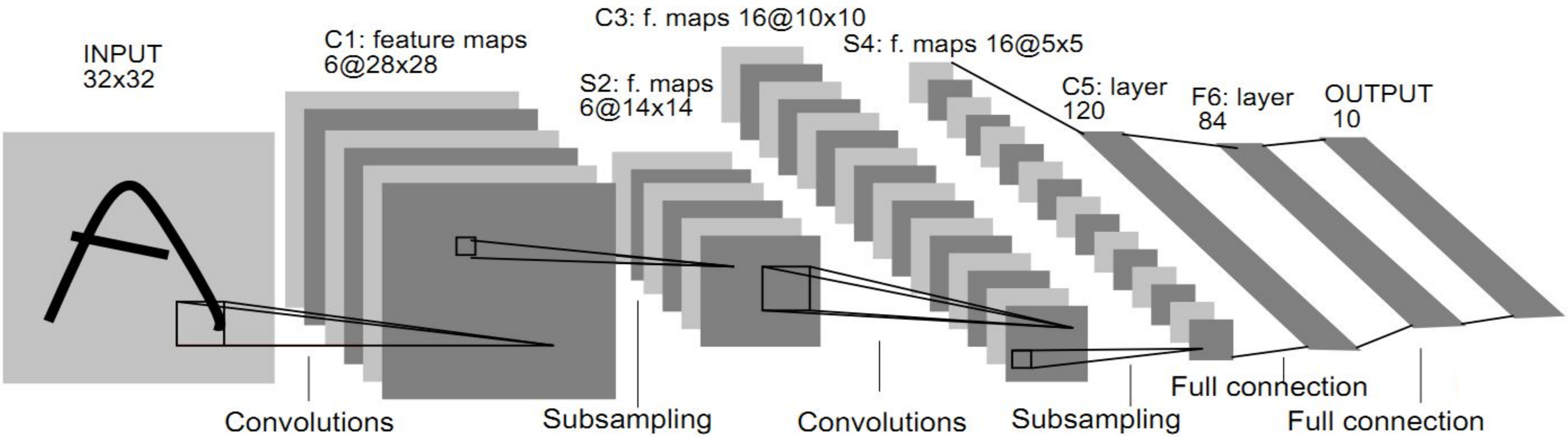








LeNet-5 [LeCun et al., 1998]



Convolution filters: 5x5 with stride 1

Subsampling (Pooling) layers: 2x2 with stride 2

[CONV-POOL-CONV-POOL-FC-FC]

Today's Agenda

- CNN Architectures
  - LeNet (1998)
  - **AlexNet (2012)**
  - ZFNet (2013)
  - VGGNet (2014)
  - GoogLeNet (2014)
  - ResNet (2015)

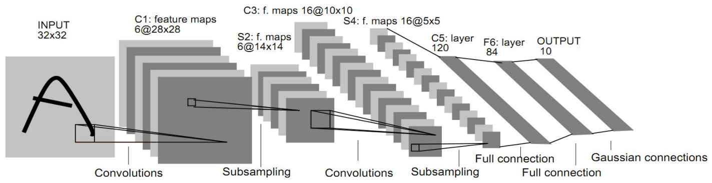
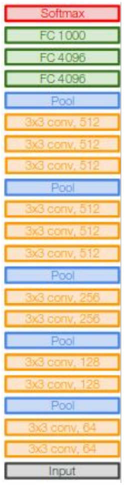
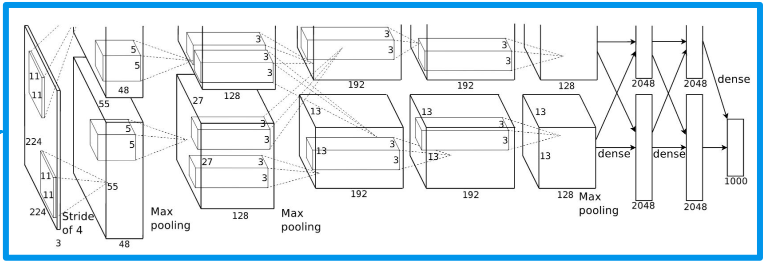
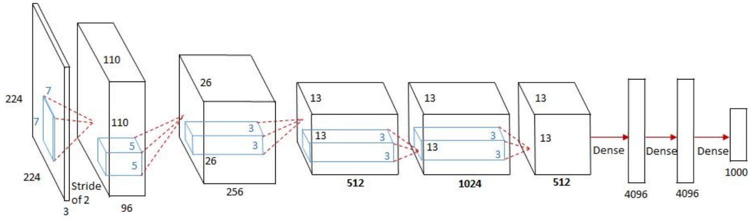
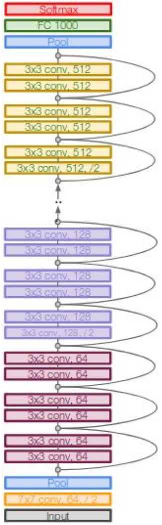


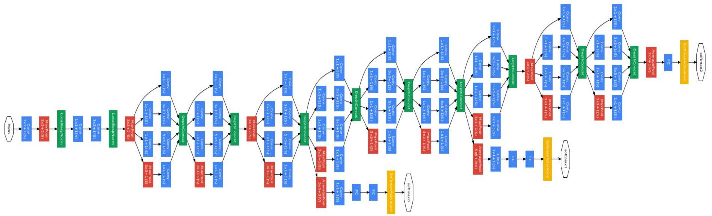
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

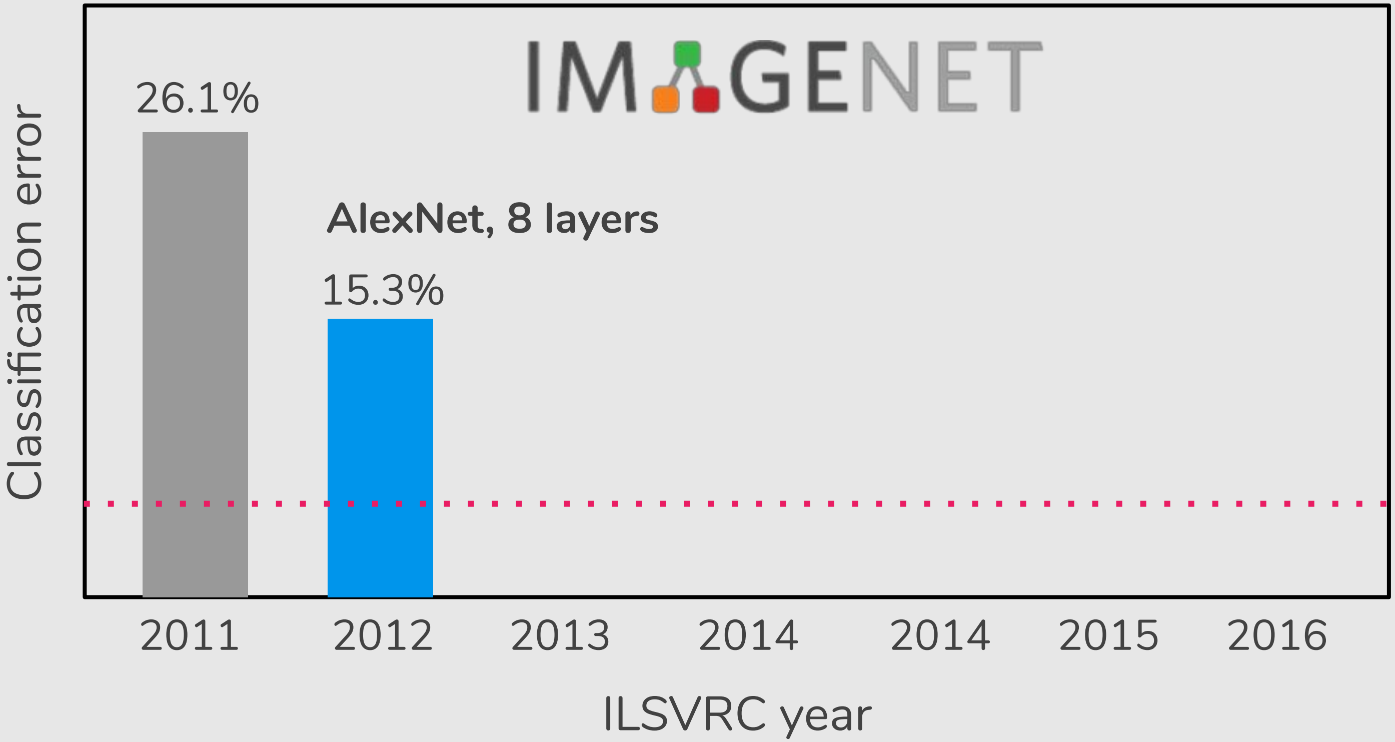






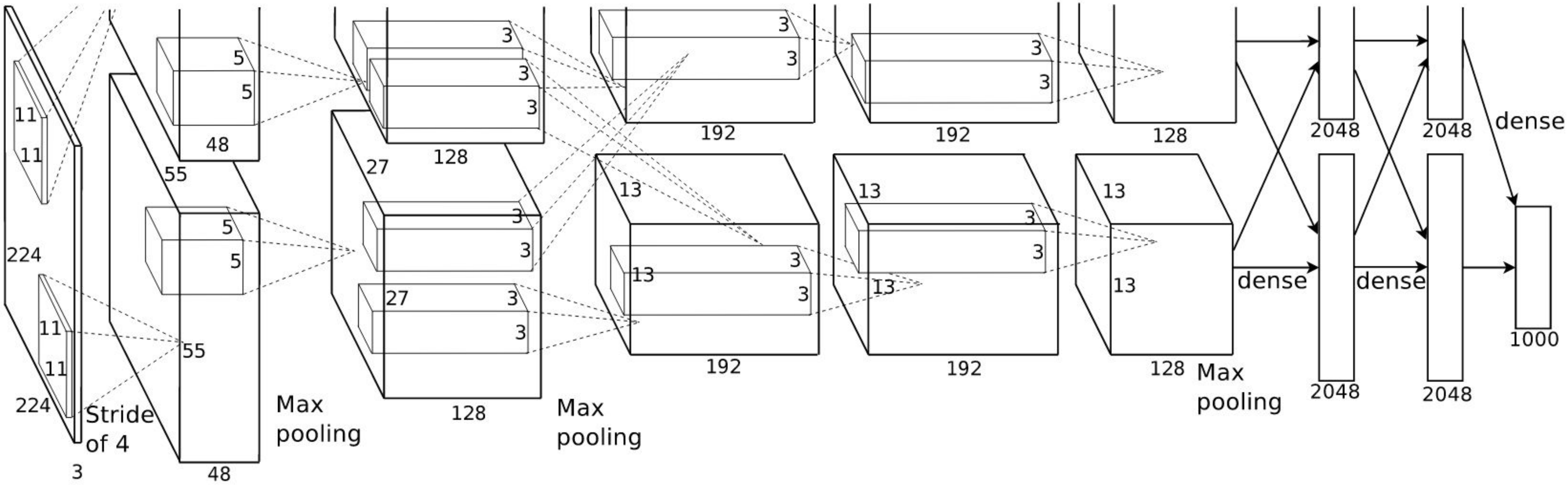






“ImageNet classification with deep convolutional neural networks”. NIPS, 2012.

AlexNet [Krizhevsky et al., 2012]



“ImageNet Classification with Deep Convolutional Neural Networks”, NIPS 2012.

AlexNet [Krizhevsky et al., 2012]

Details:

- 60 million learned parameters
- first use of ReLU
- used Norm layers (not common anymore)
- heavy data augmentation
- dropout 0.5
- batch size 128
- 7 CNN ensemble: 18.2% -> 15.3%
- 5-6 days to train on 2 GTX 580 3GB GPUs

Today's Agenda

- CNN Architectures
  - LeNet (1998)
  - AlexNet (2012)
  - **ZFNet (2013)**
  - VGGNet (2014)
  - GoogLeNet (2014)
  - ResNet (2015)

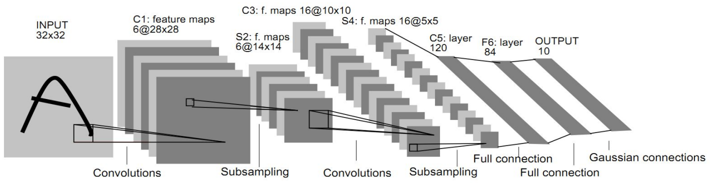
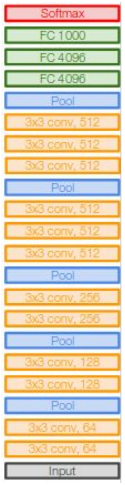
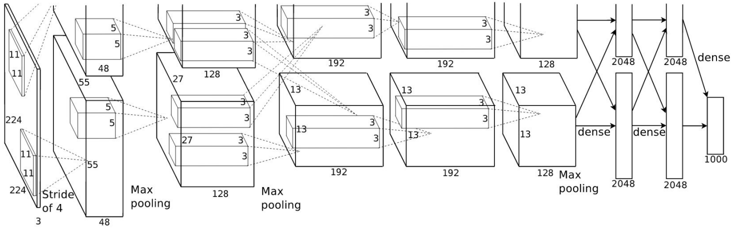
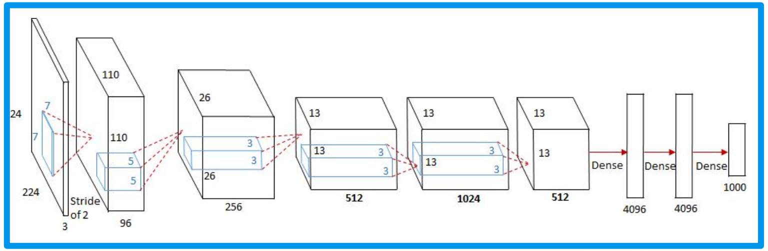
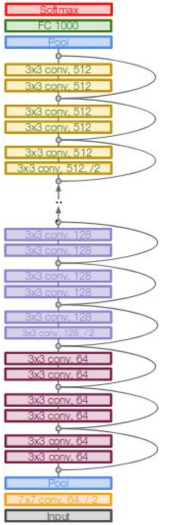


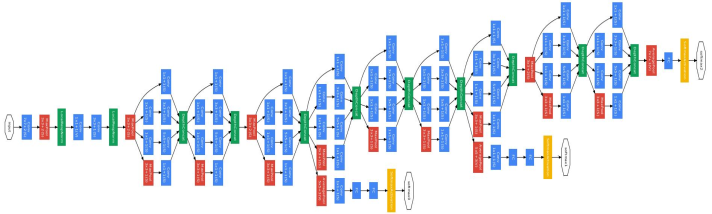
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

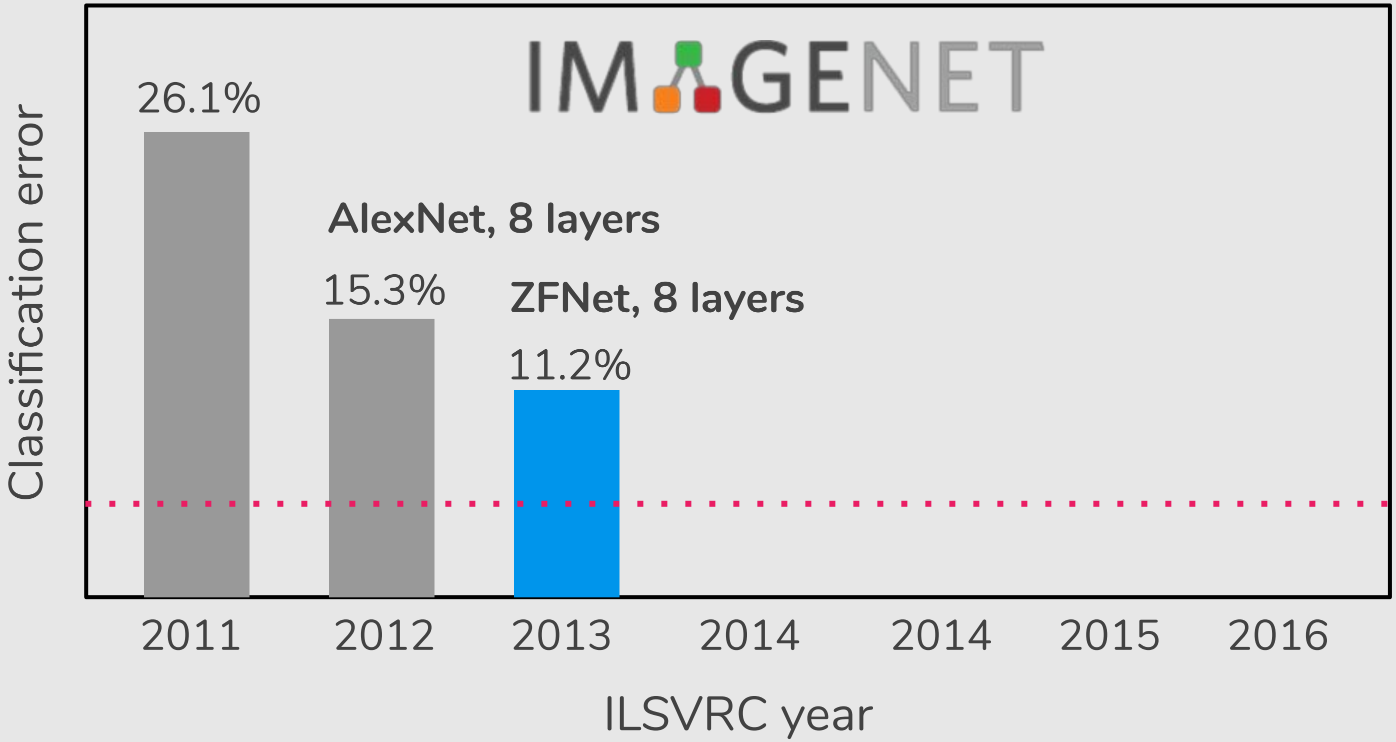






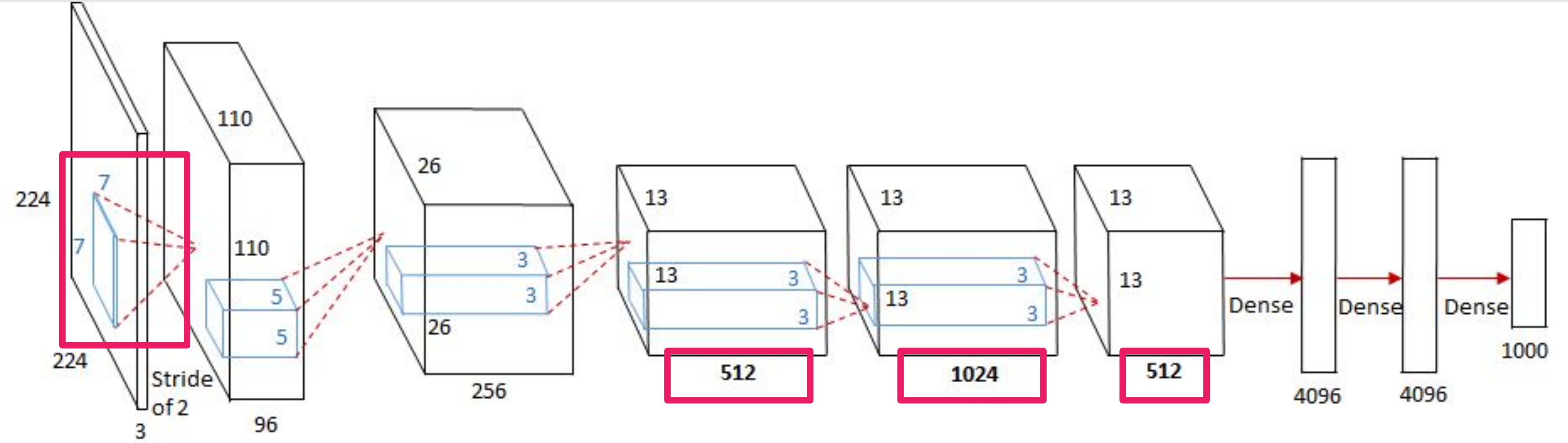






“Visualizing and Understanding Convolutional Networks”, ECCV 2014, https://cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf

ZFNet [Zeiler & Fergus, 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

Today's Agenda

- CNN Architectures
  - LeNet (1998)
  - AlexNet (2012)
  - ZFNet (2013)
  - **VGGNet (2014)**
  - GoogLeNet (2014)
  - ResNet (2015)

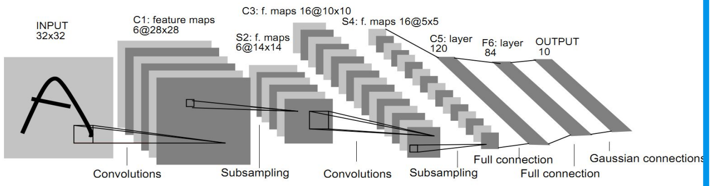
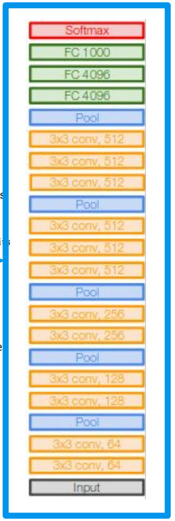
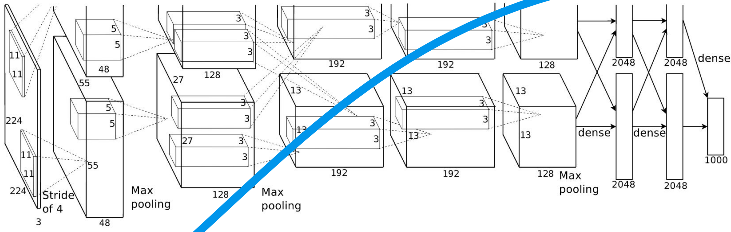
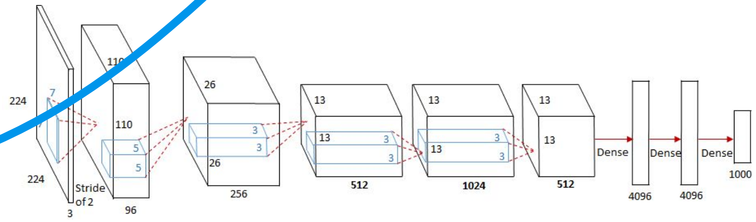
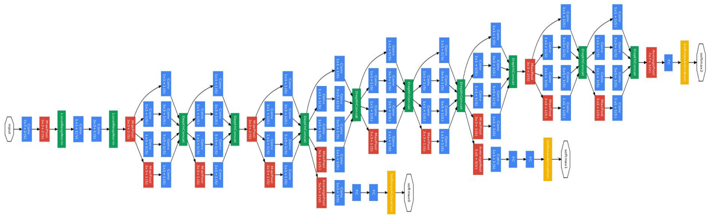


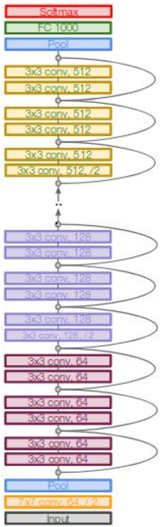
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

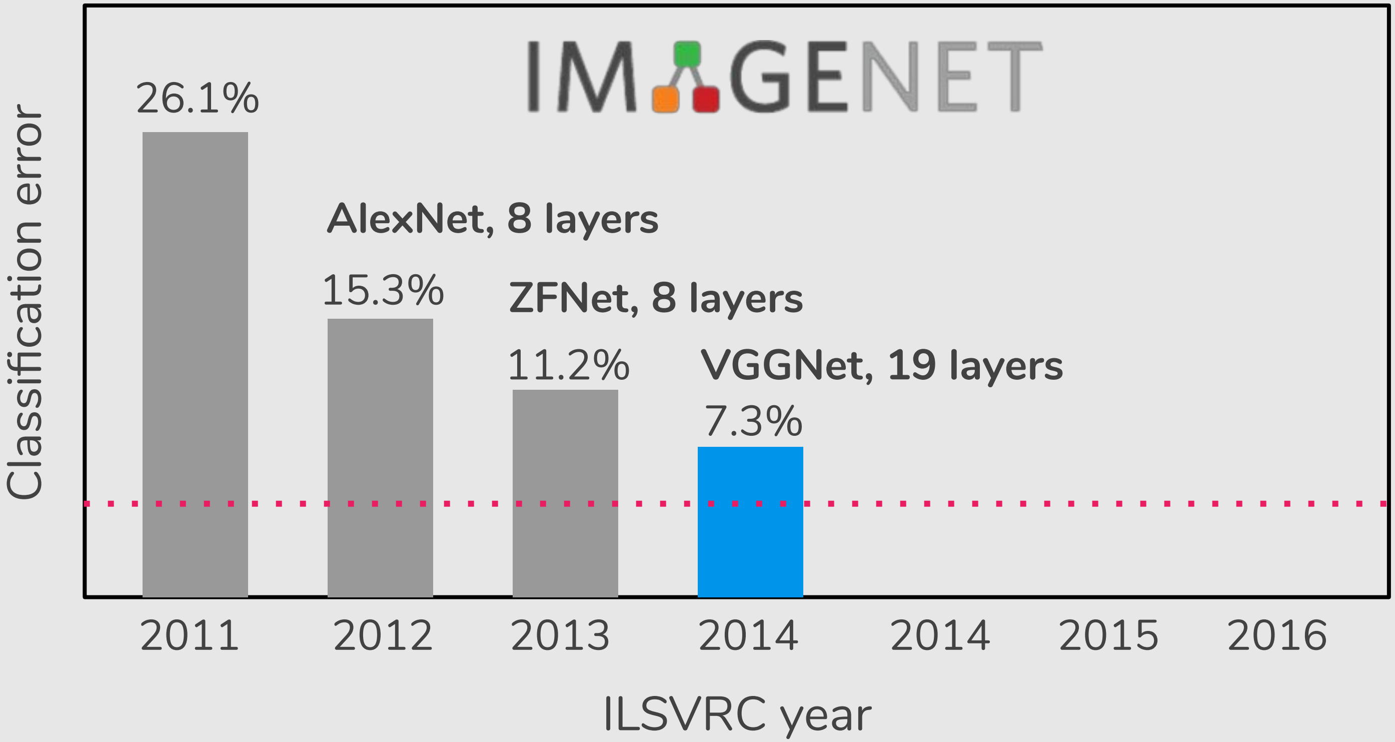












“Very Deep Convolutional Networks for Large-Scale Image Recognition”, https://arxiv.org/pdf/1409.1556

VGGNet [Simonyan & Zisserman, 2014]

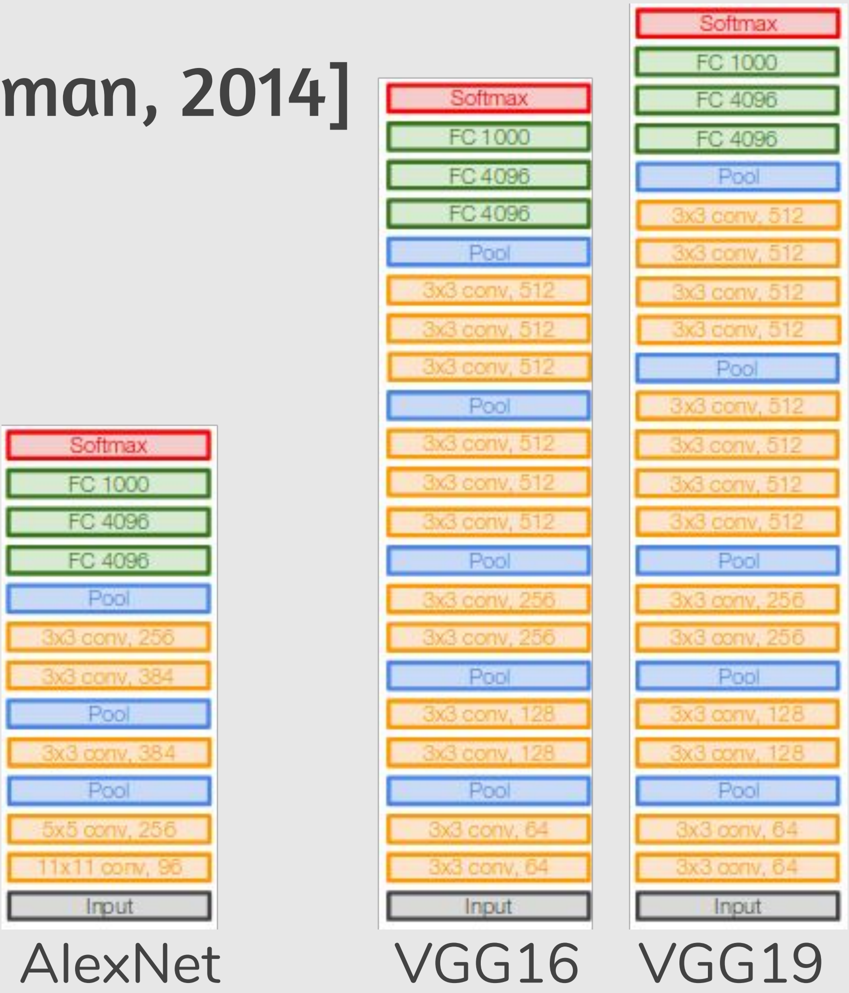
Small filters, Deeper networks

8 layers (AlexNet)

16-19 layers (VGG16Net)

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

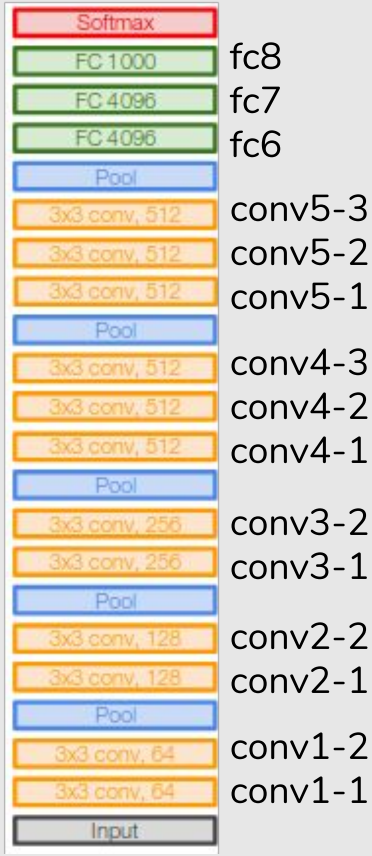
11.2% in ILSVRC'13 (ZFNet)
7.3% in ILSVRC'14



VGGNet [Simonyan & Zisserman, 2014]

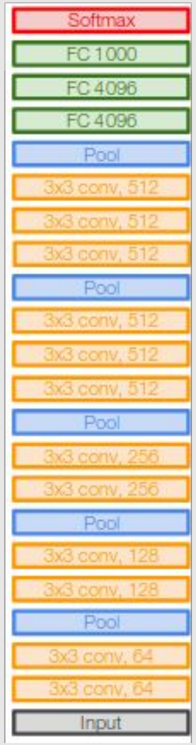
Details:

- 138M parameters
- 2nd in classification, 1st in localization
- Use VGG16 or VGG19 (VGG19 only slightly better, more memory)
- Use ensembles for best results
- FC7 features generalize well to other tasks



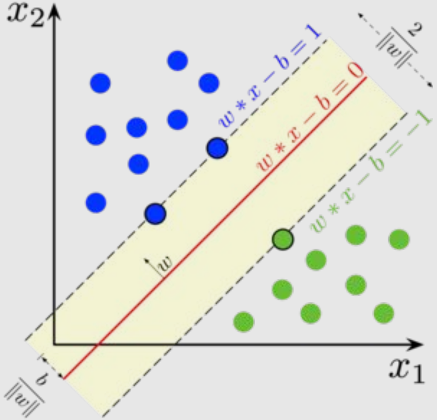
VGG16

VGGNet [Simonyan & Zisserman, 2014]



[0.01 0.8 1 0.5 ... 0.3 0.07 0 0.4 0.6 0 0]
4096-d

Train a classifier (e.g., SVM)



VGG as Feature Extractor

Today's Agenda

- CNN Architectures
  - LeNet (1998)
  - AlexNet (2012)
  - ZFNet (2013)
  - VGGNet (2014)
  - **GoogLeNet (2014)**
  - ResNet (2015)

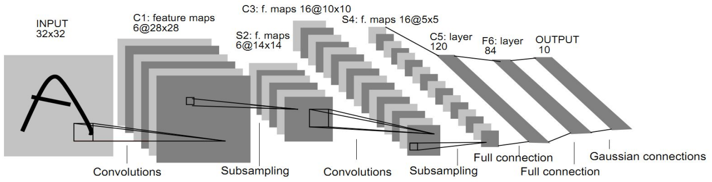
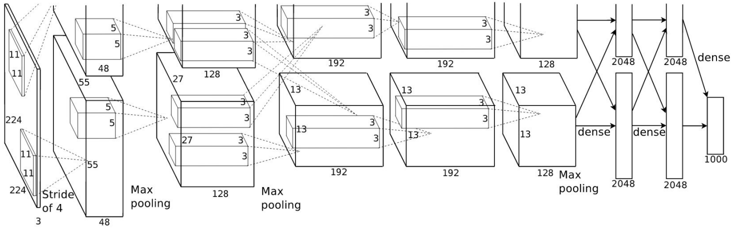
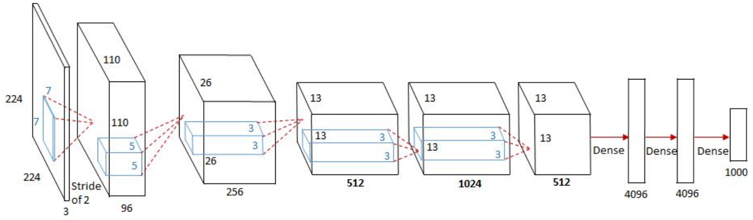
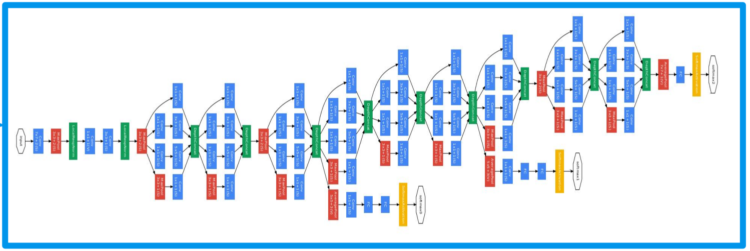
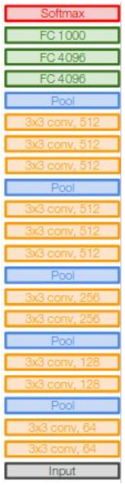


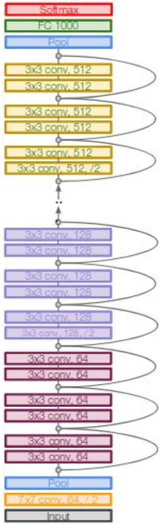
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

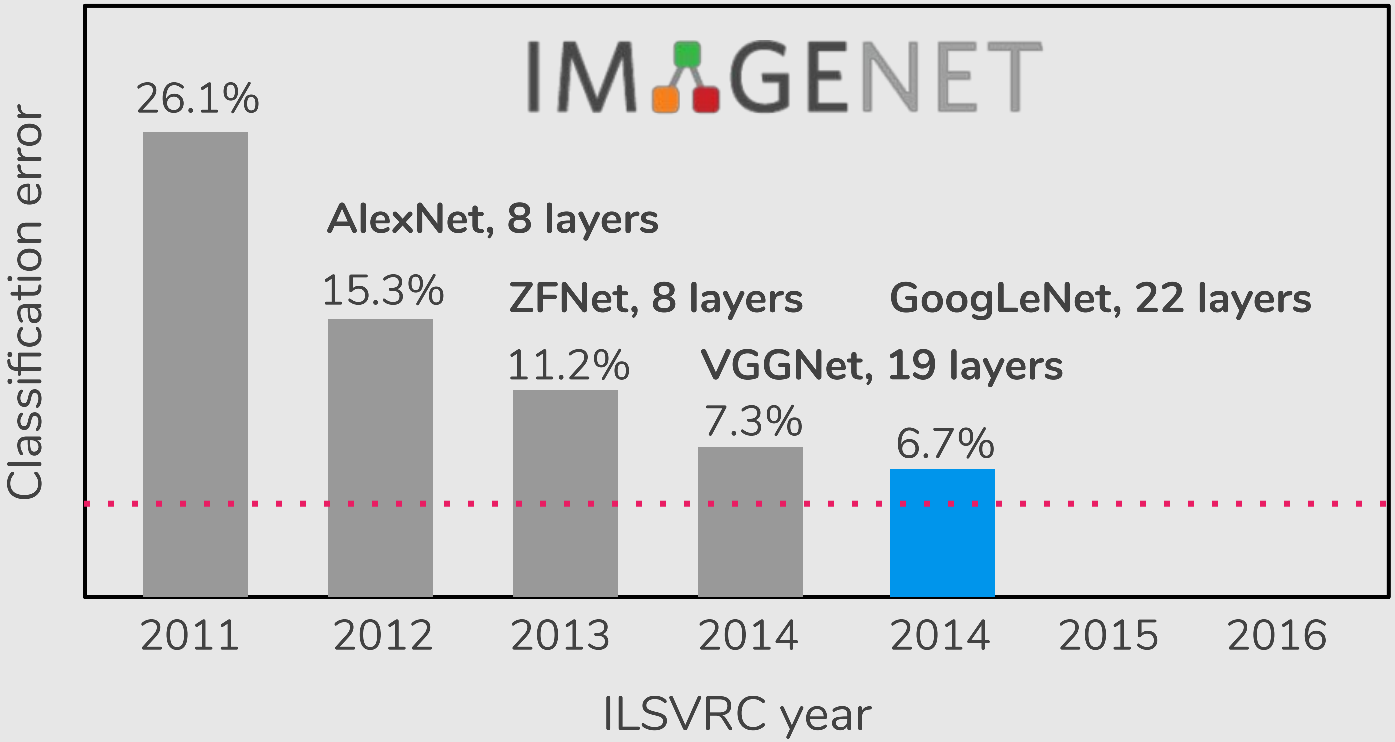










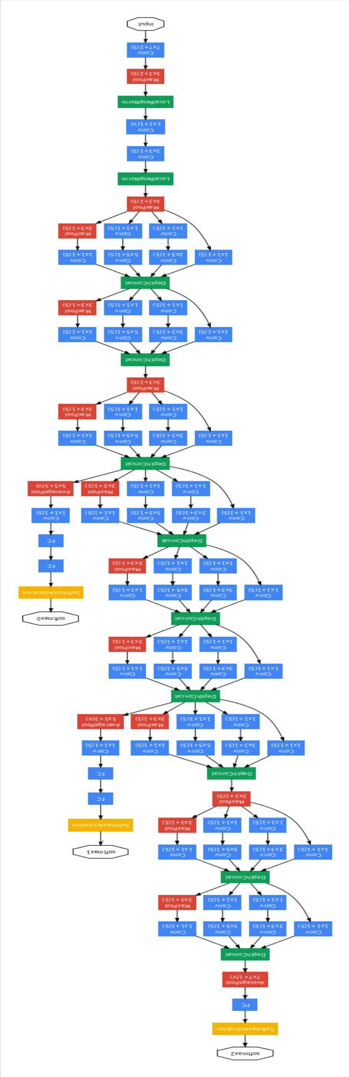


“Very Deep Convolutional Networks for Large-Scale Image Recognition”, https://arxiv.org/pdf/1409.1556

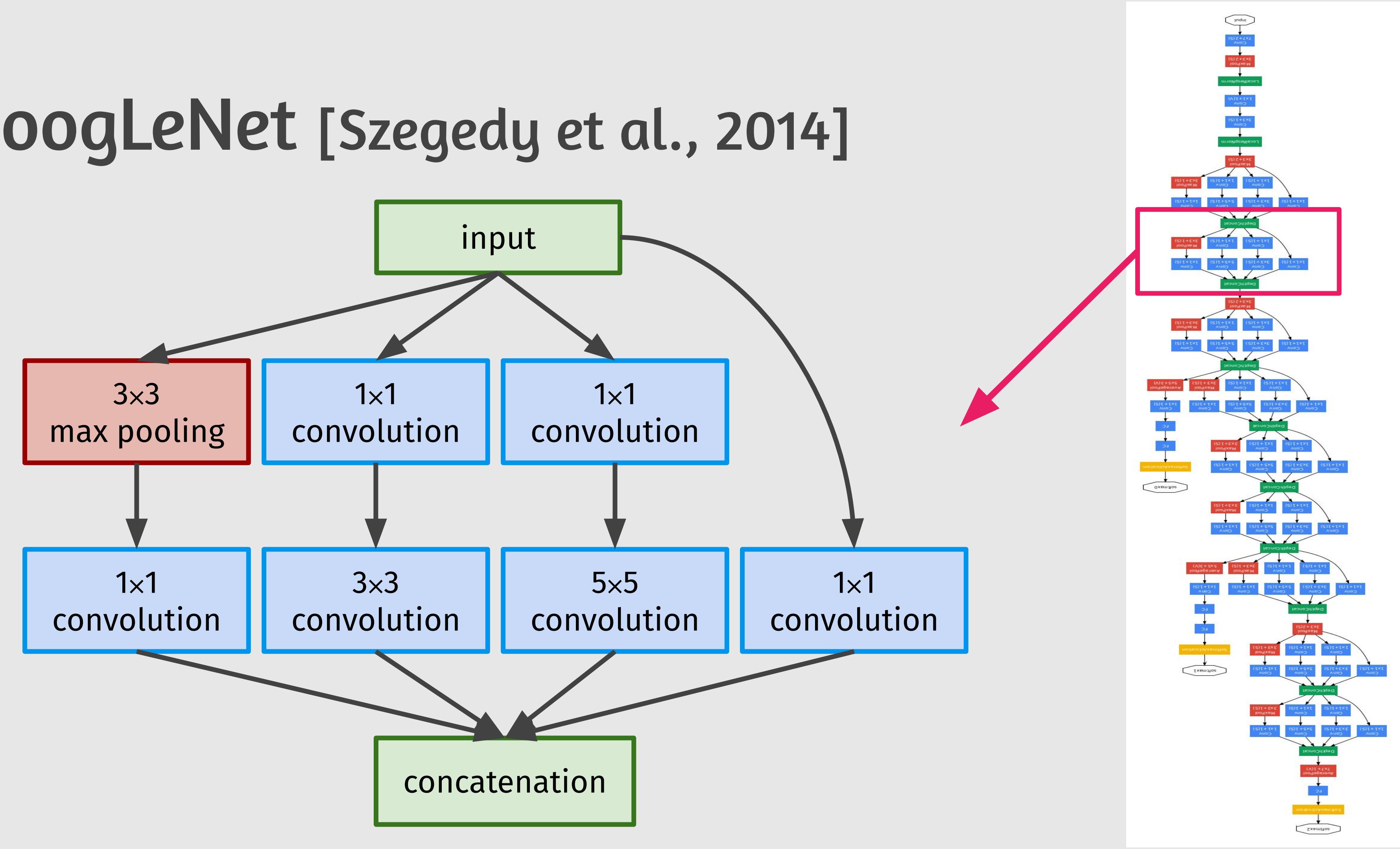
GoogLeNet [Szegedy et al., 2014]

Deeper networks, with computational efficiency

- 22 layers
- Inception module
- Only 5 million parameters!
  12x less than AlexNet



GoogLeNet [Szegedy et al., 2014]



input

3×3
max pooling

1×1
convolution

1×1
convolution

1×1
convolution

3×3
convolution

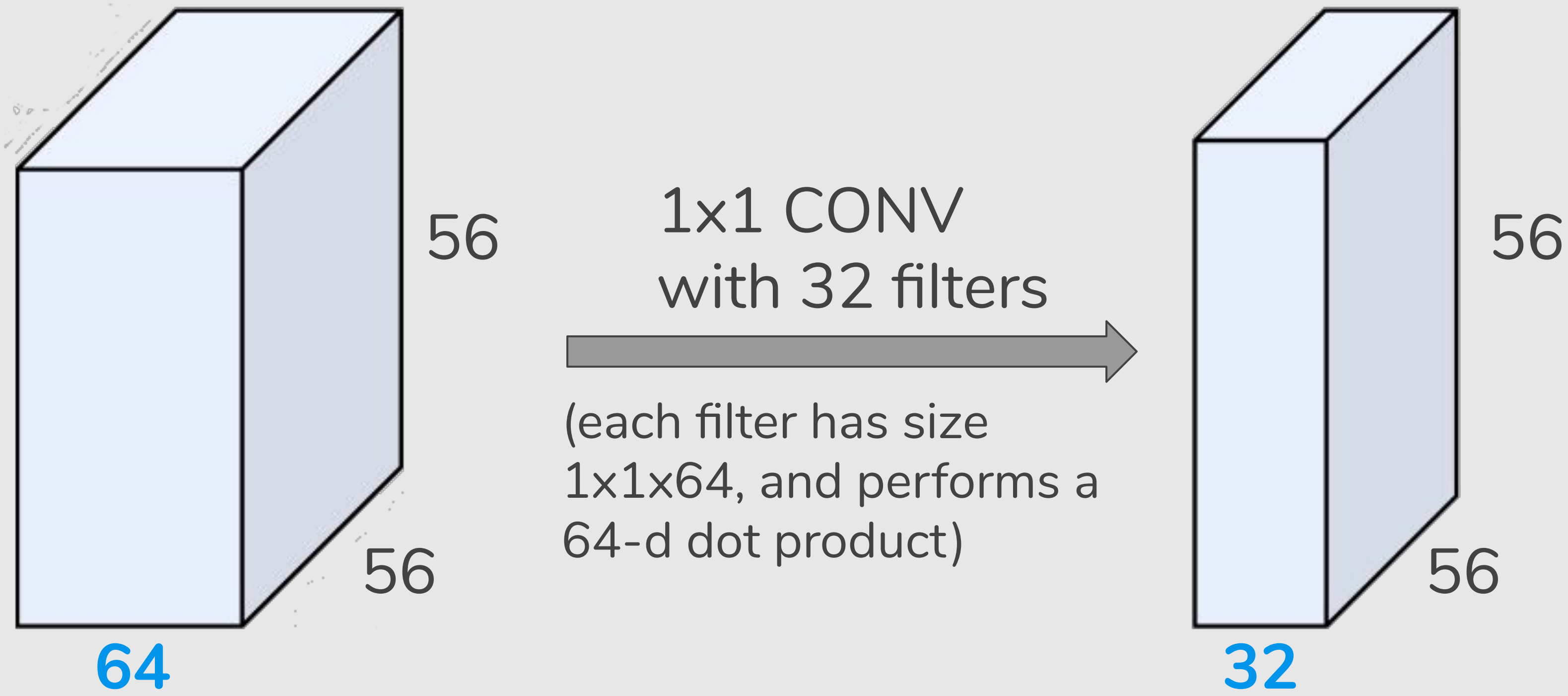
5×5
convolution

1×1
convolution

concatenation

GoogLeNet [Szegedy et al., 2014]

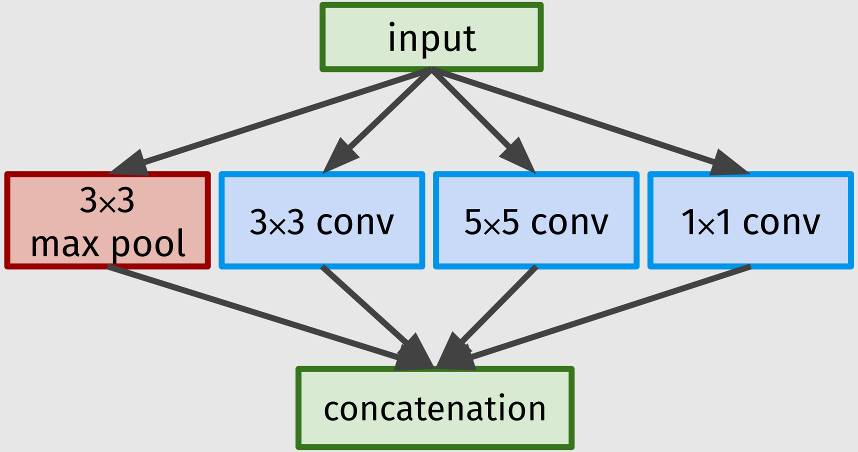
Reminder: 1x1 convolutions



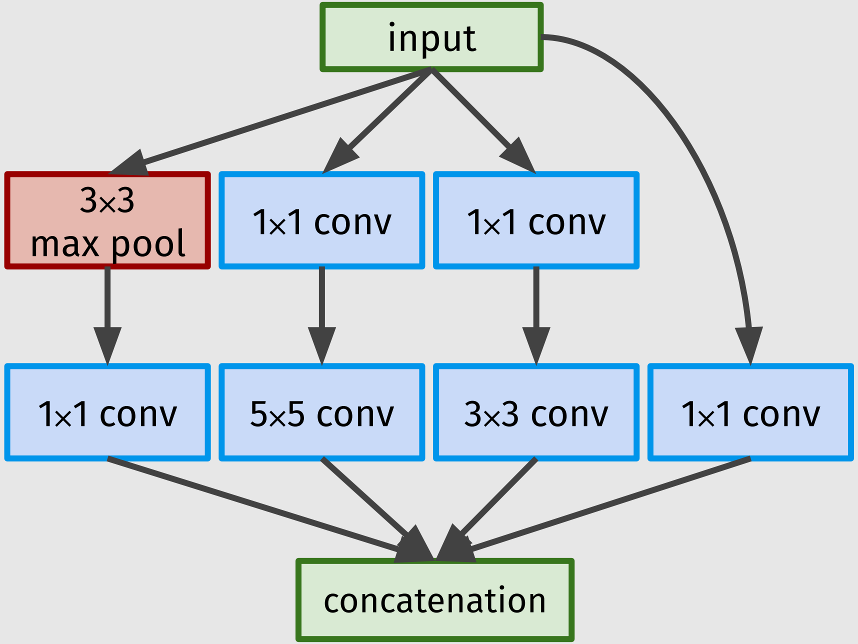
1x1 CONV
with 32 filters

(each filter has size 1x1x64, and performs a 64-d dot product)

GoogLeNet [Szegedy et al., 2014]



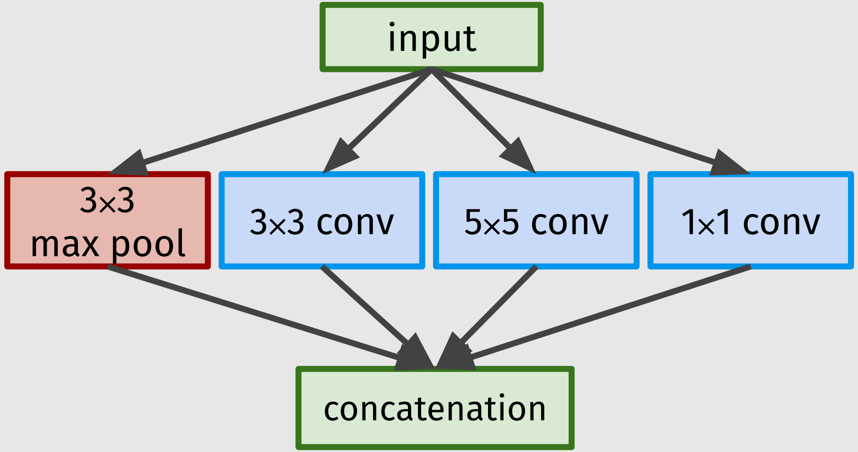
Naive Inception Module



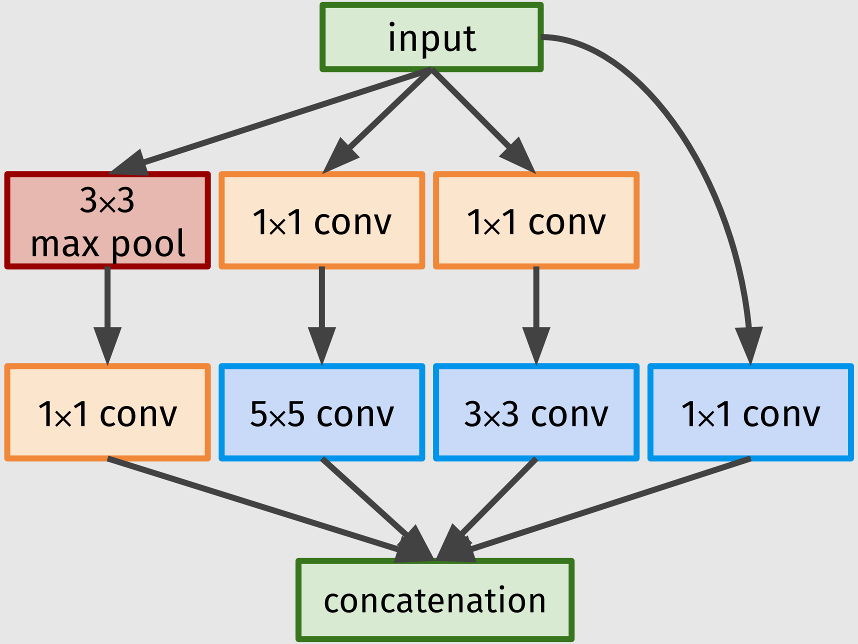
Inception Module

GoogLeNet [Szegedy et al., 2014]

1x1 conv
“bottleneck” layers

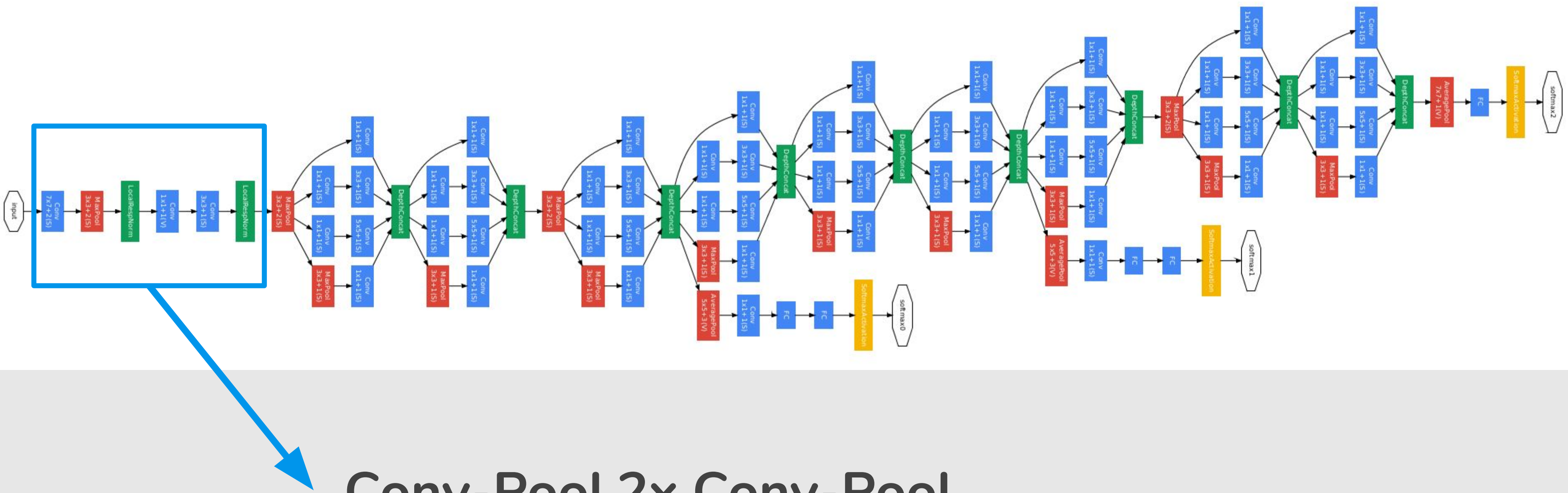


Naive Inception Module



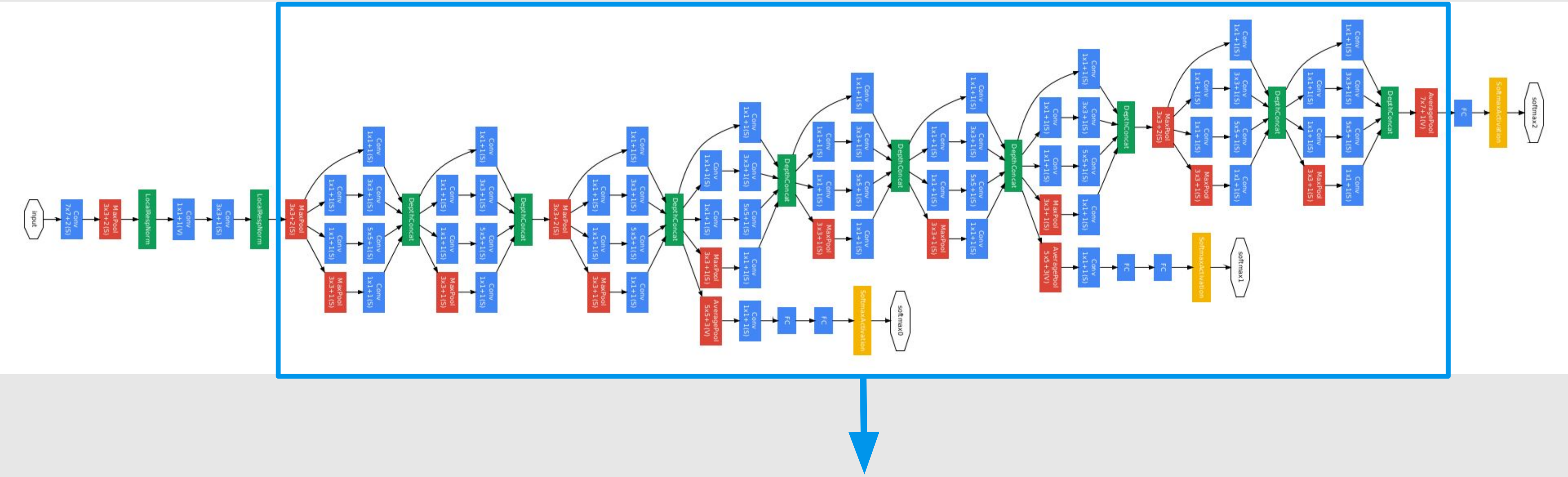
Inception Module

GoogLeNet [Szegedy et al., 2014]



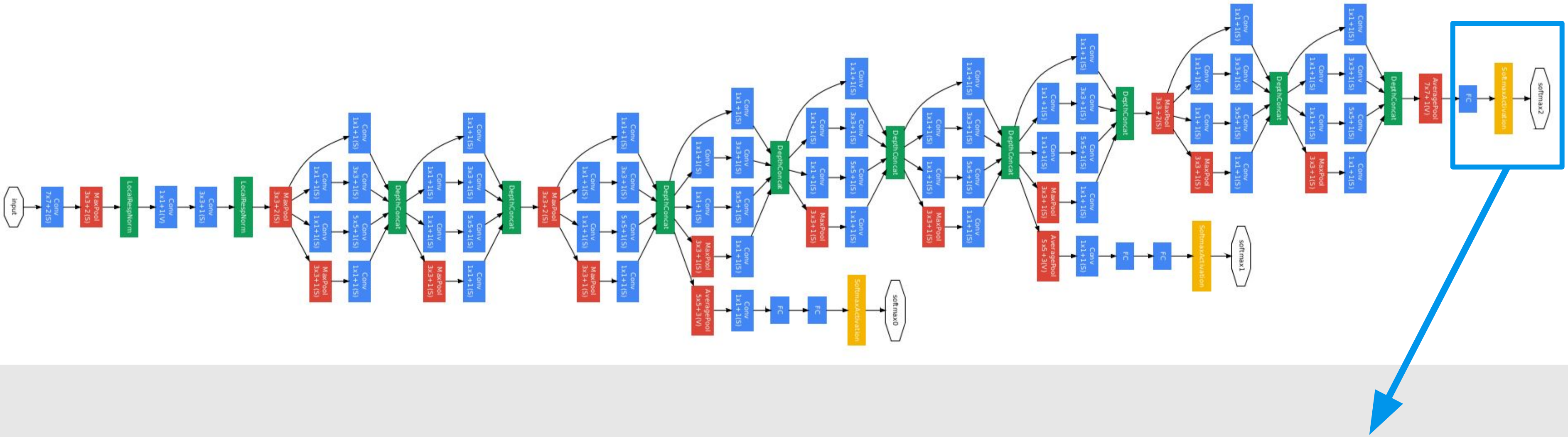
Conv-Pool 2x Conv-Pool

GoogLeNet [Szegedy et al., 2014]



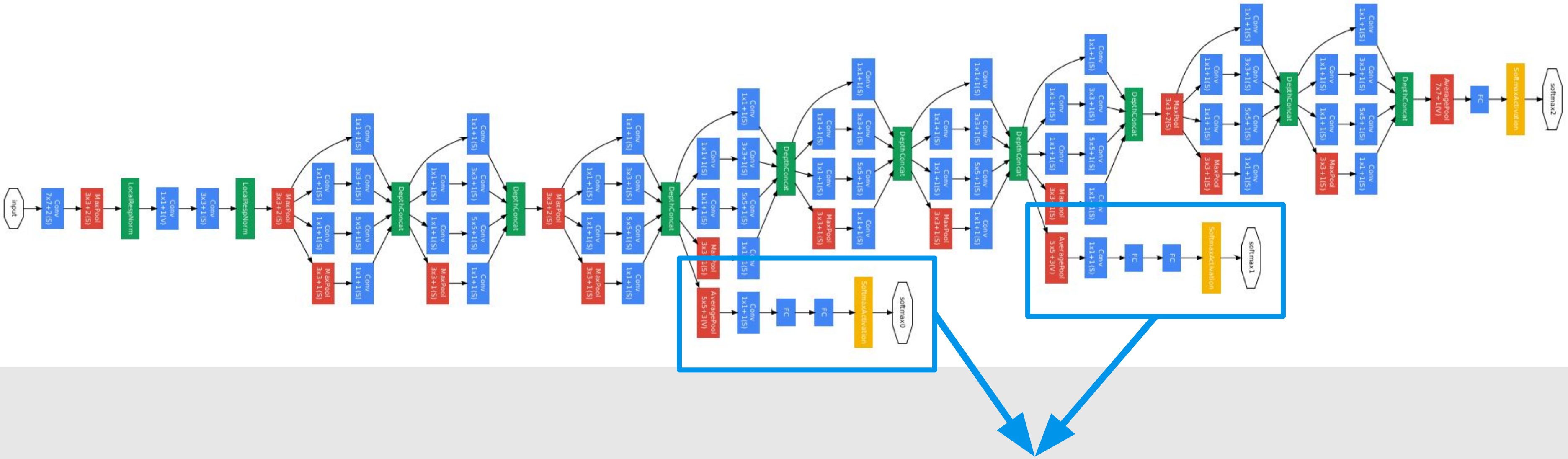
Stacked Inception Modules

GoogLeNet [Szegedy et al., 2014]



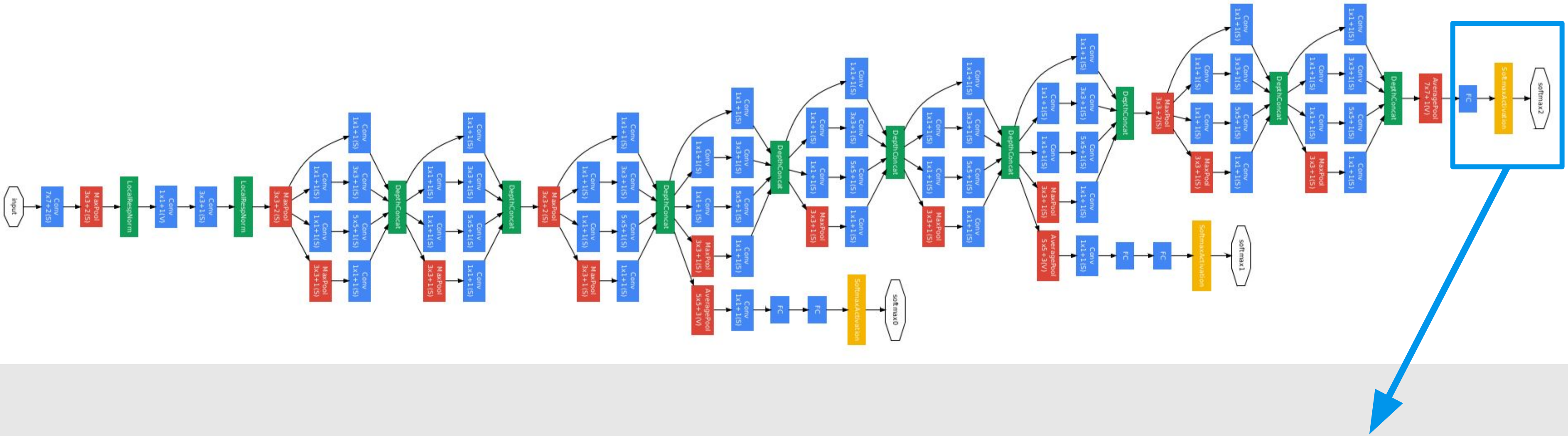
Classifier Output

GoogLeNet [Szegedy et al., 2014]



Auxiliary Classifiers

GoogLeNet [Szegedy et al., 2014]



The total loss function is a weighted sum of the auxiliary loss and the real loss.

```
total_loss = real_loss + 0.3*aux_loss_1 + 0.3*aux_loss_2
```

GoogLeNet [Szegedy et al., 2014]

- GoogLeNet has **9 inception modules** stacked linearly.
- It is **22 layers deep** (27, including the pooling layers).
- It uses **global average pooling** at the end of the last inception module.
- GoogLeNet = Inception v1
- Inception v2, v3, v4, Inception-ResNet v1, v2: https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202

Today’s Agenda

- CNN Architectures
  - LeNet (1998)
  - AlexNet (2012)
  - ZFNet (2013)
  - VGGNet (2014)
  - GoogLeNet (2014)
  - **ResNet (2015)**

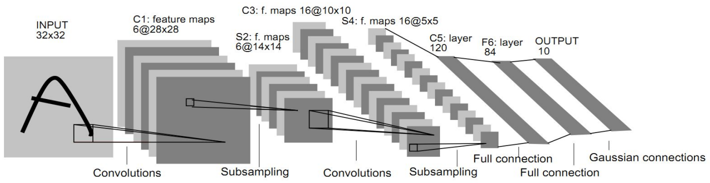
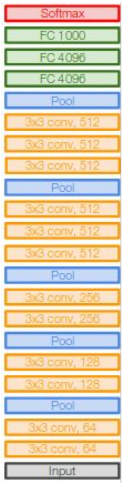
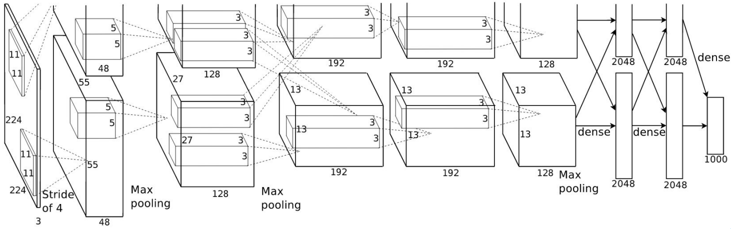
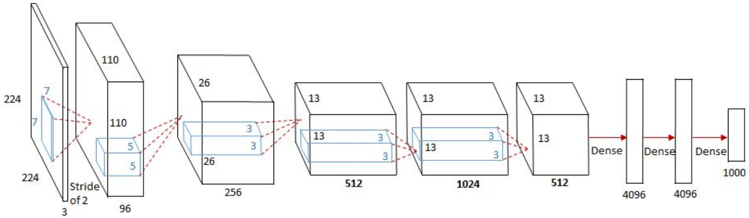
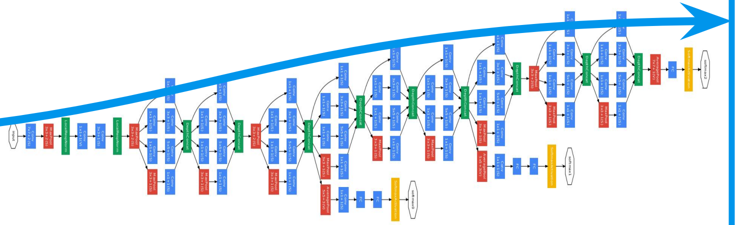


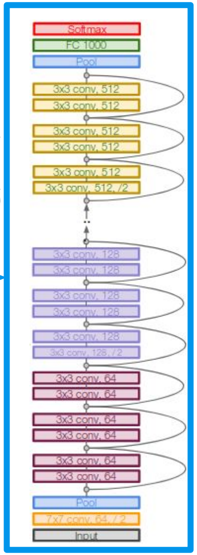
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

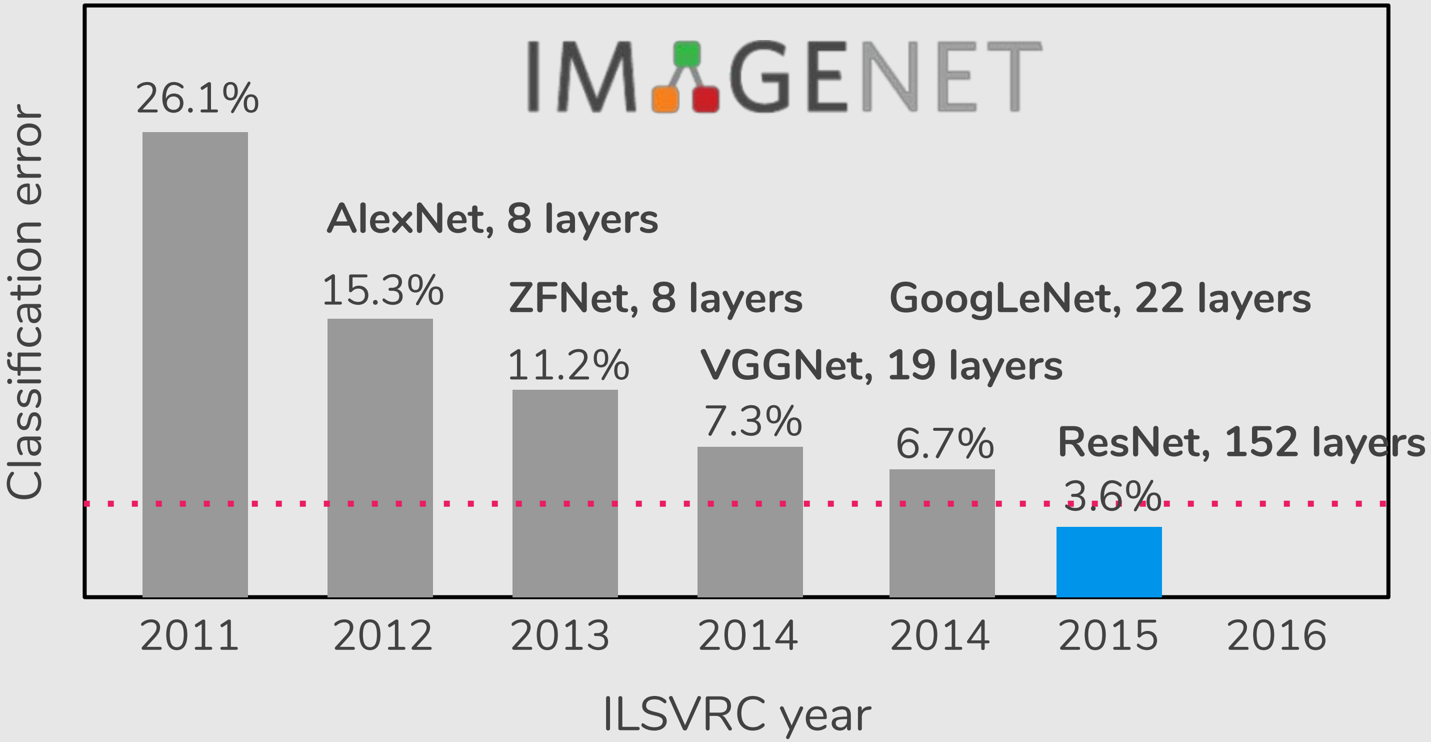












“Deep Residual Learning for Image Recognition”, CVPR 2016, https://arxiv.org/pdf/1512.03385

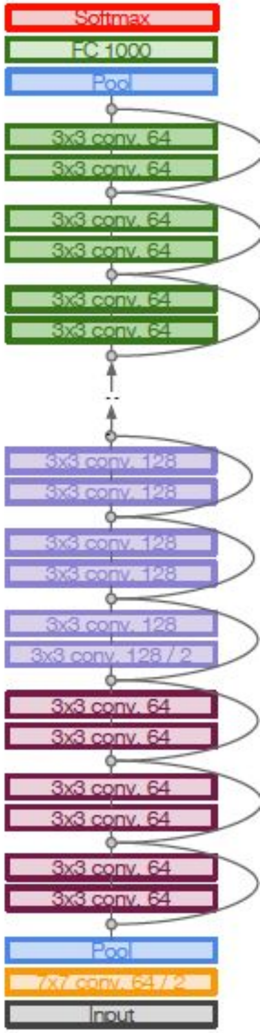
ResNet [He et al., 2015]

ResNet @ ILSVRC & COCO 2015 Competitions

1st place in ALL five main tracks

- ImageNet Classification: “Ultra-deep” 152-layer nets
- ImageNet Detection: 16% better than 2nd
- ImageNet Localization: 27% better than 2nd
- COCO Detection: 11% better than 2nd
- COCO Segmentation: 12% better than 2nd

ResNet [He et al., 2015]



Very deep networks using residual connections

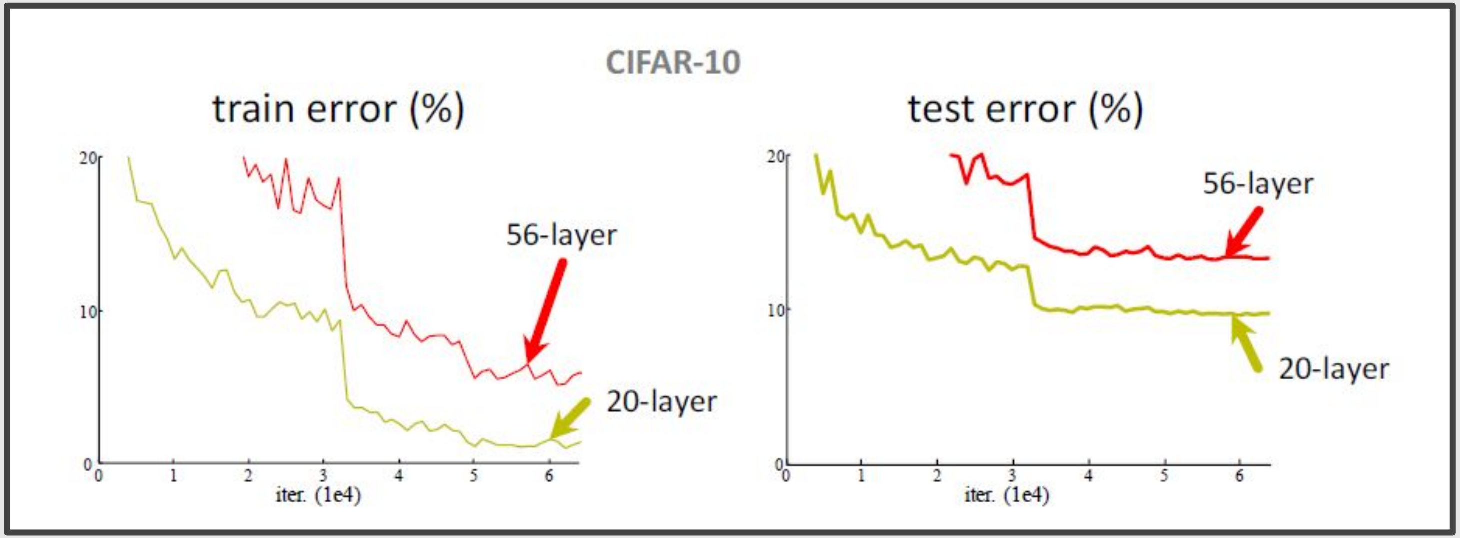
- 152-layer model for ImageNet
- ILSVRC’15 classification winner (3.57% top 5 error)

ResNet [He et al., 2015]

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

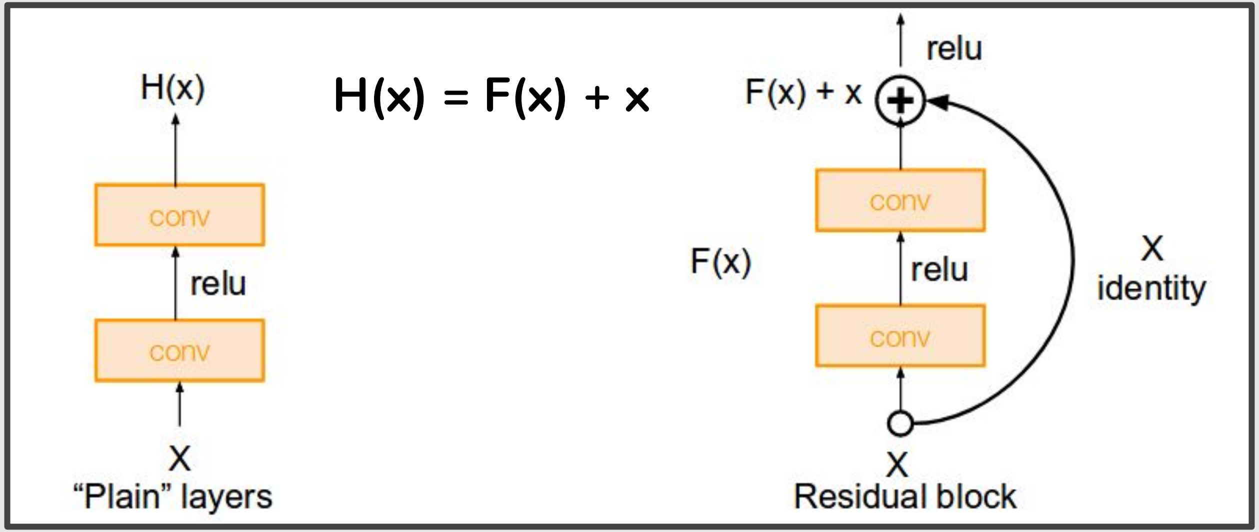
ResNet [He et al., 2015]

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



ResNet [He et al., 2015]

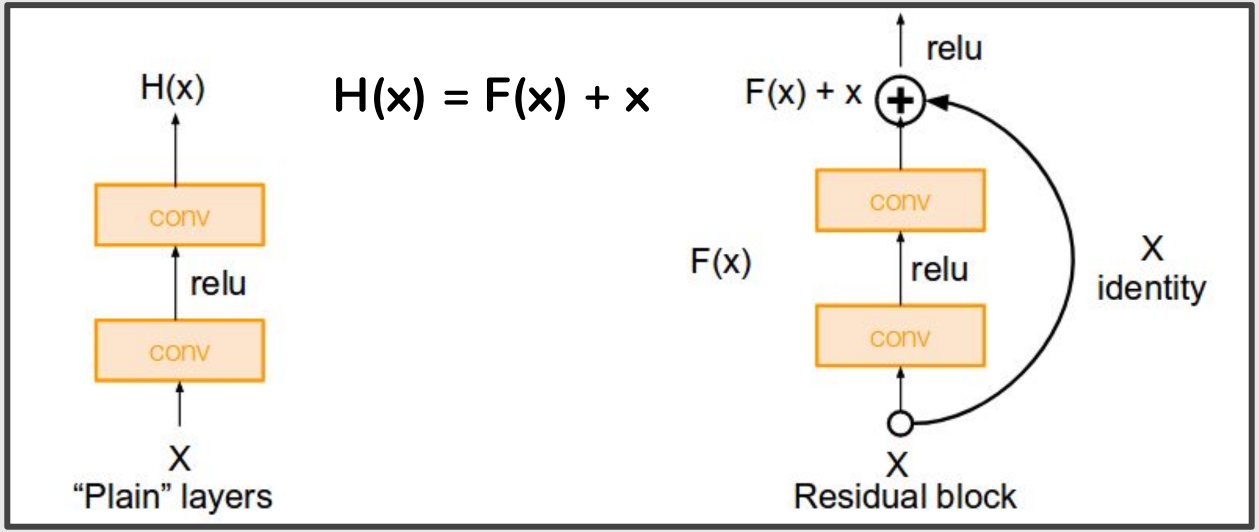
Solution: Use network layers to fit a residual mapping instead of directly trying to fit a desired underlying mapping



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

ResNet [He et al., 2015]

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

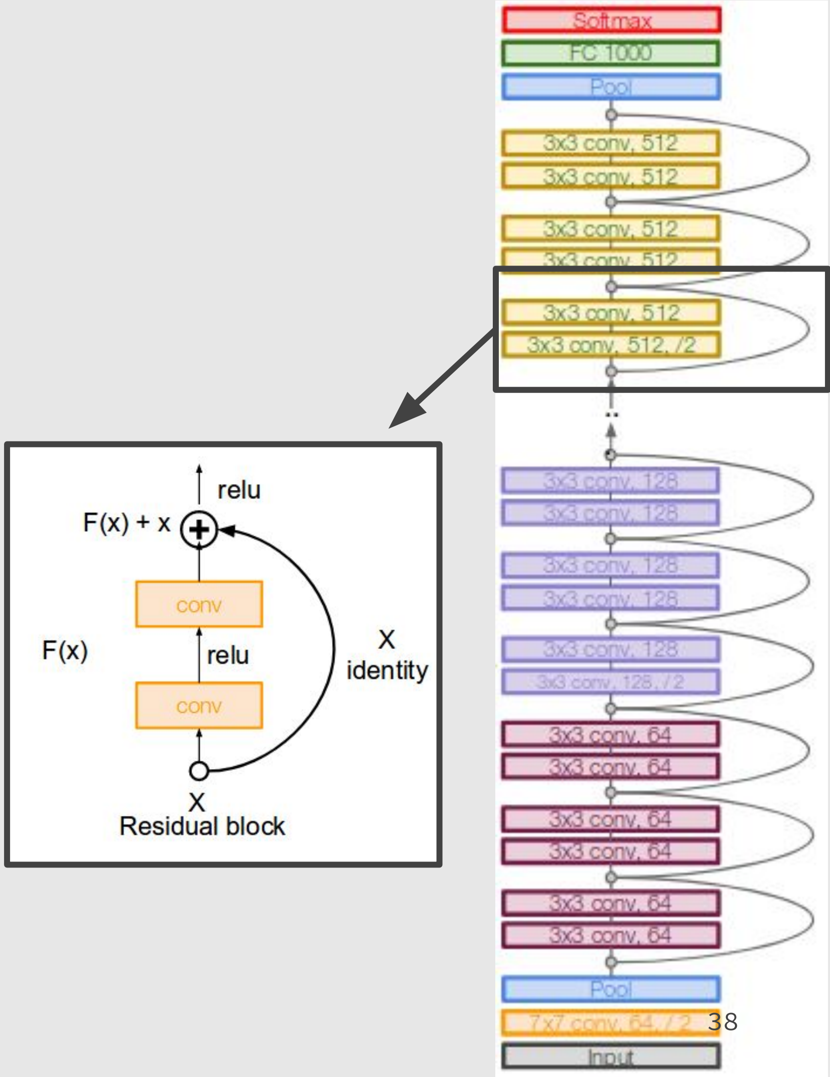


Use layers to fit residual F(x) = H(x) - x instead of H(x) directly

ResNet [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)



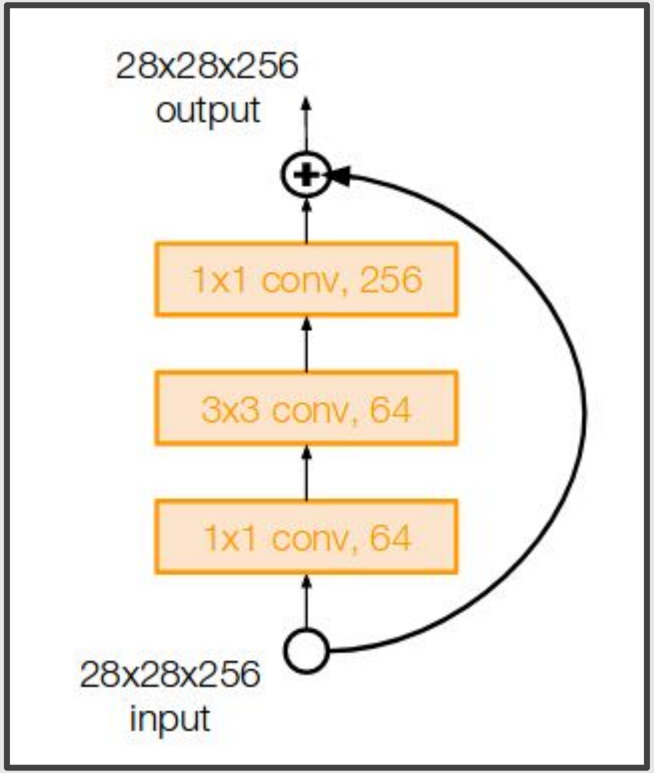
ResNet [He et al., 2015]

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

1x1 conv, 256 filters projects back to 256 feature maps (28x28x256)

3x3 conv operates over only 64 feature maps

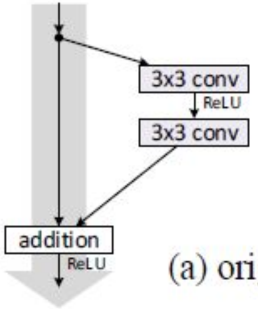
1x1 conv, 64 filters to project to 28x28x64



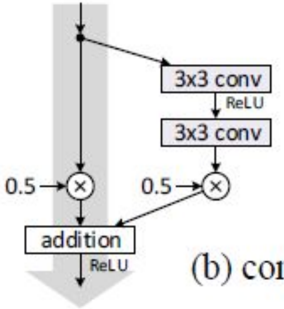
* ResNet-110 on CIFAR-10

$h(x) = x$

error: 6.6%



(a) original



(b) constant scaling

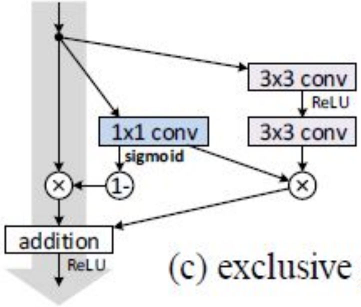
$h(x) = 0.5x$

error: 12.4%

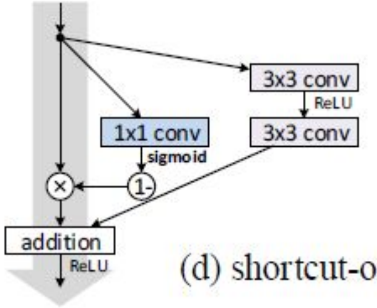
$h(x) = \text{gate} \cdot x$

error: 8.7%

*similar to "Highway Network"



(c) exclusive gating



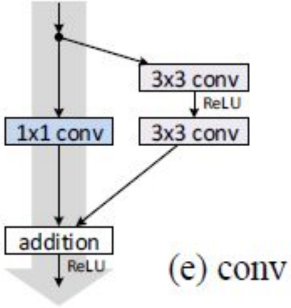
(d) shortcut-only gating

$h(x) = \text{gate} \cdot x$

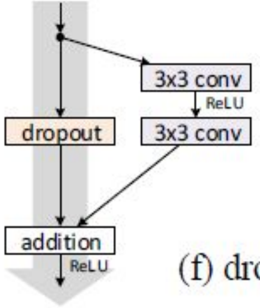
error: 12.9%

$h(x) = \text{conv}(x)$

error: 12.2%



(e) conv shortcut

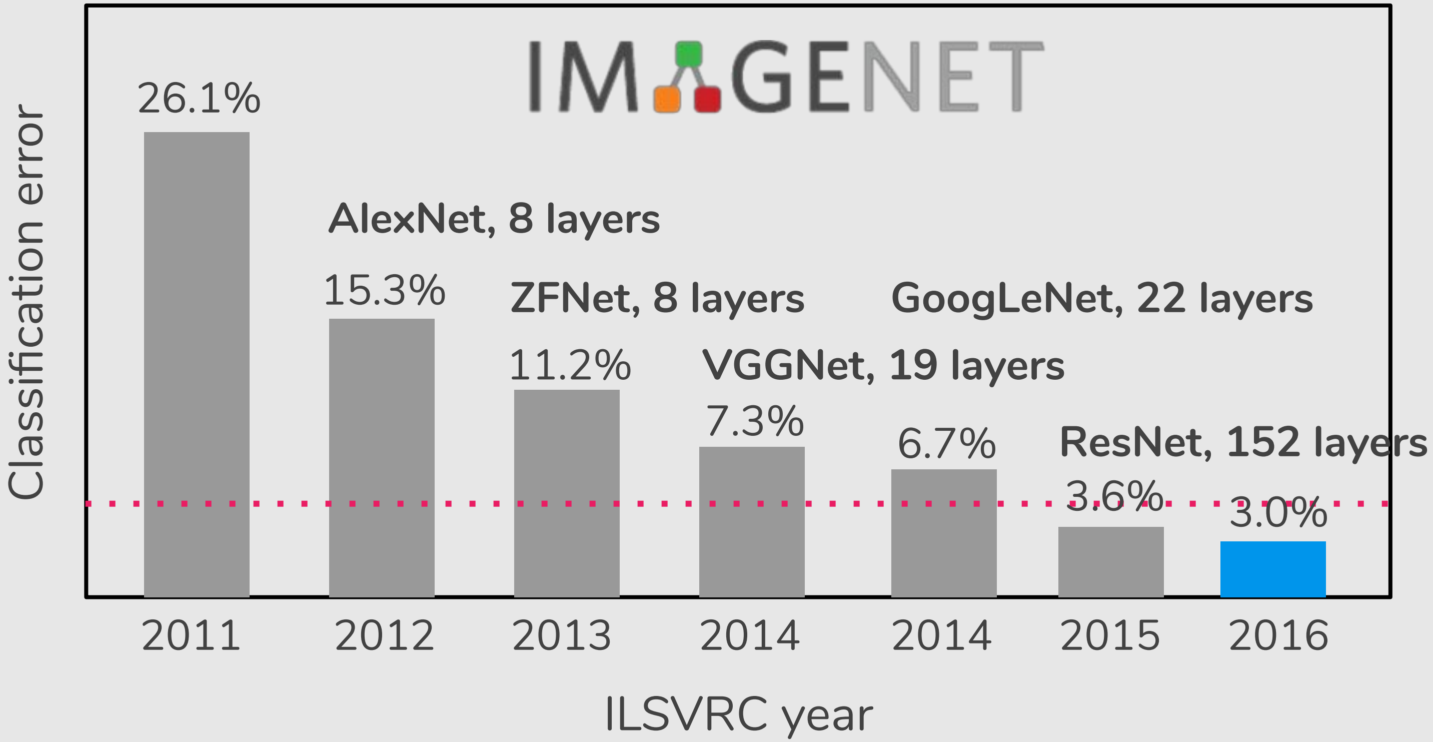


(f) dropout shortcut

$h(x) = \text{dropout}(x)$

error: > 20%

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Identity Mappings in Deep Residual Networks". arXiv 2016.



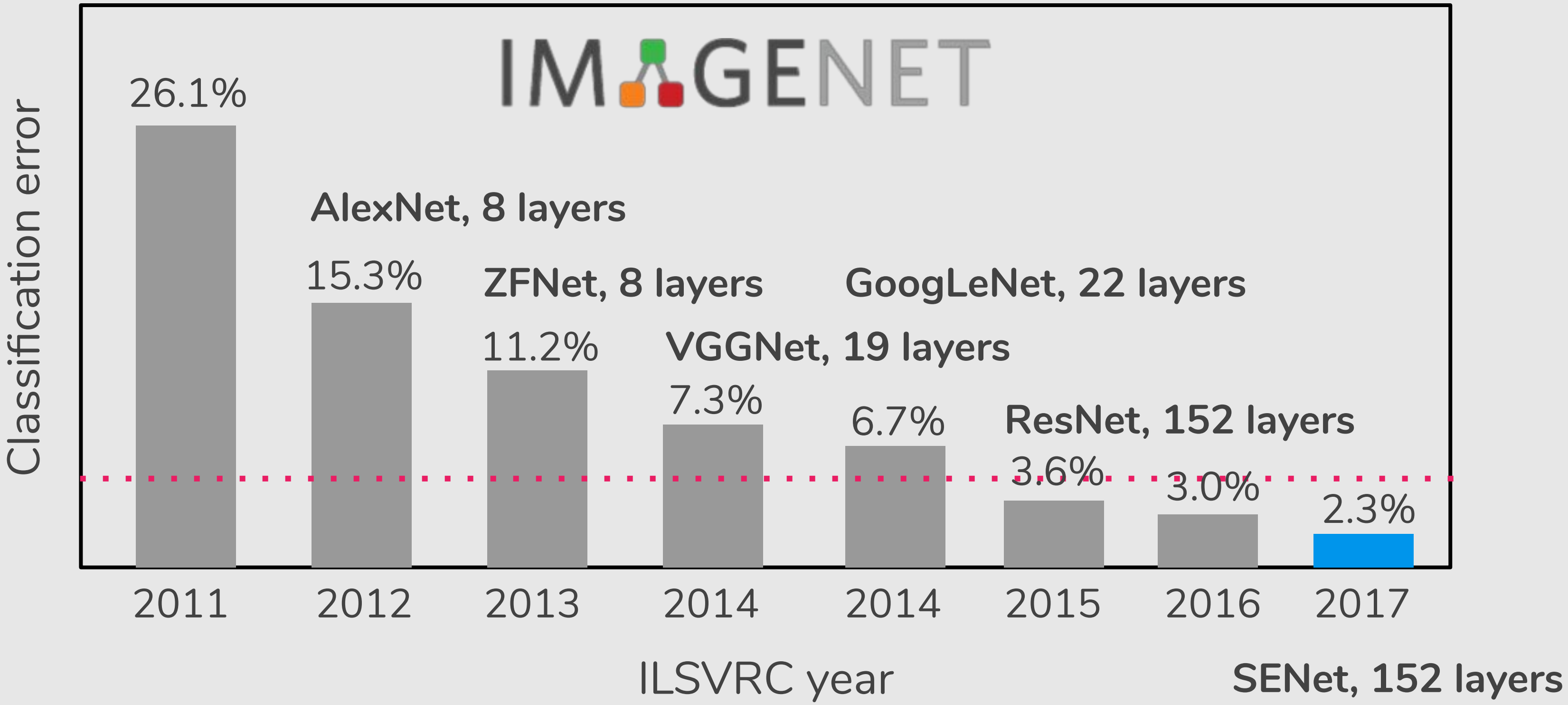
"Good Practices for Deep Feature Fusion", ECCV 2016,
http://image-net.org/challenges/talks/2016/Trimps-Soushen@ILSVRC2016.pdf (Slides only)

Good Practices for Deep Feature Fusion

[Shao et al., 2016]

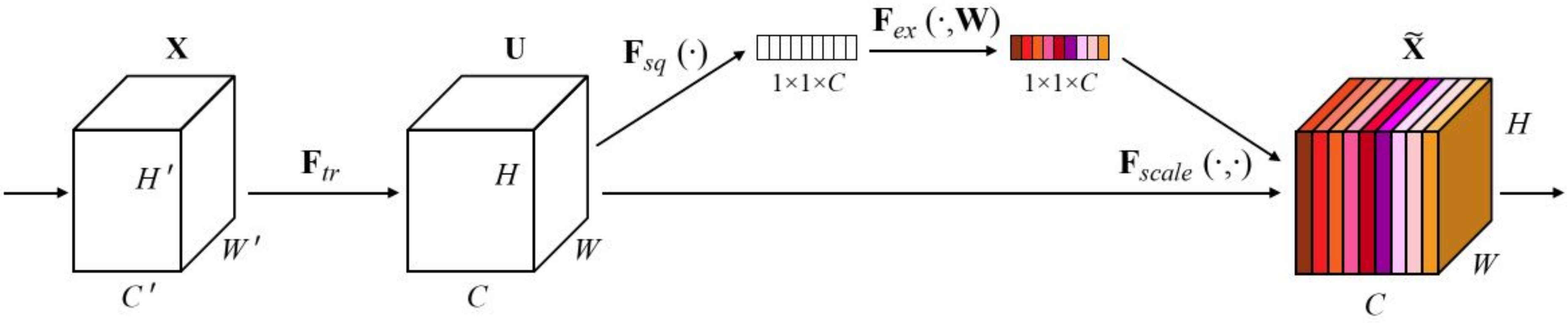
- Training
  - Multi-scale augmentation & large mini-batch size
- Testing
  - Multi-scale & flip & dense fusion

	Error (%)
Inception-v3	4.20
Inception-v4	4.01
Inception-ResNet-v2	3.52
ResNet-200	4.26
Wrm-68-3	4.65
Fusion (Test)	2.99

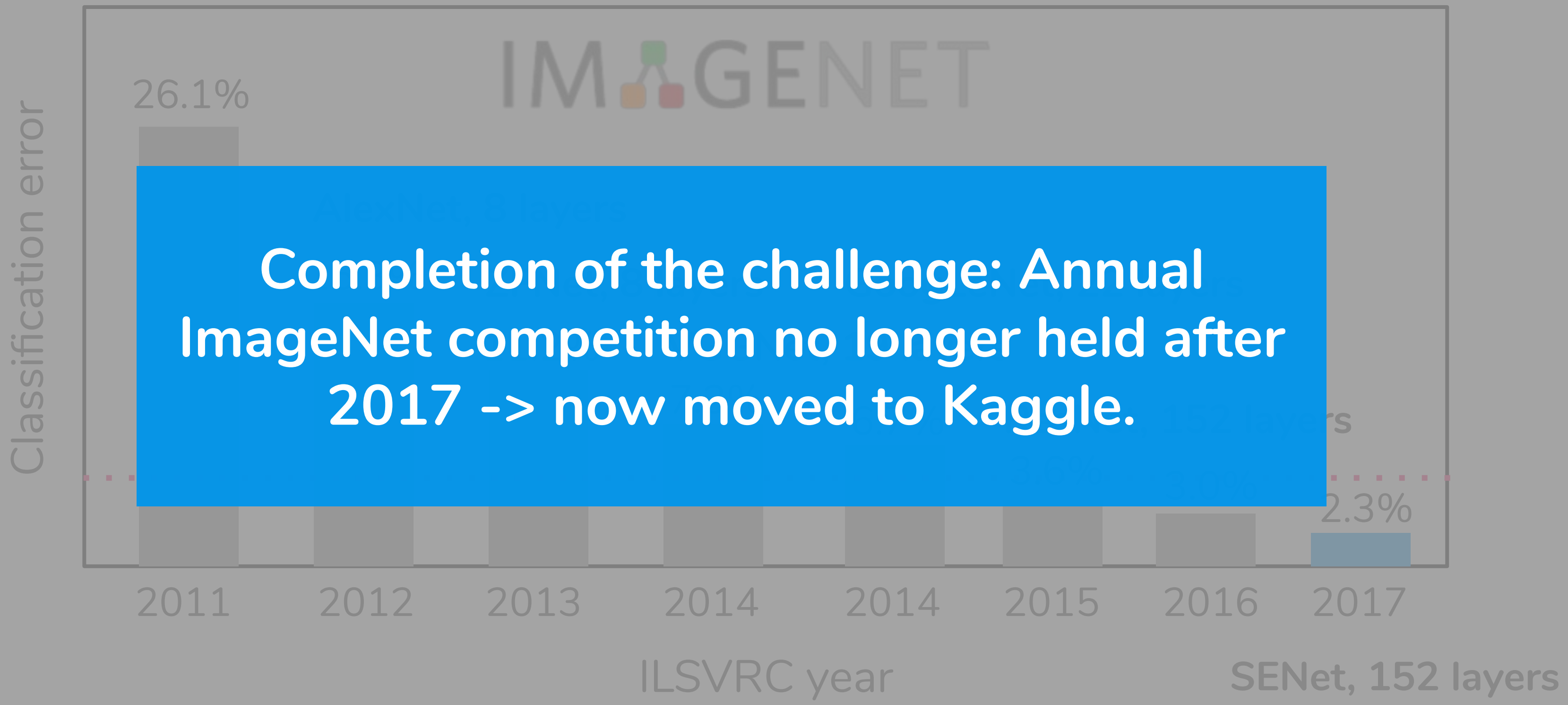


“Squeeze-and-Excitation Networks”, CVPR 2018, https://arxiv.org/pdf/1709.01507

SENets [Hu *et al.* 2017]



Add a "feature recalibration" module that **learns** to **adaptively reweight feature maps**.



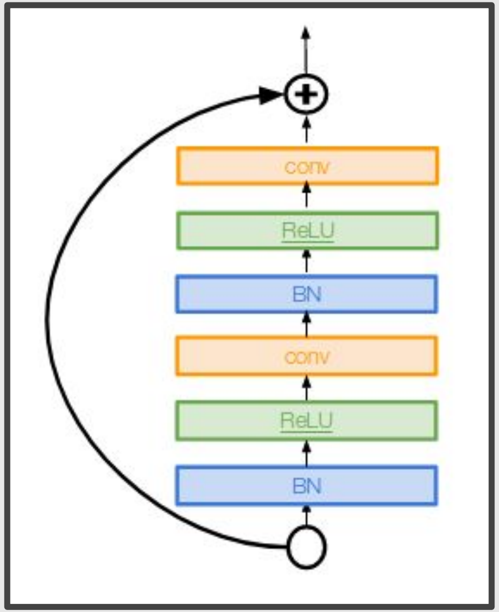
Completion of the challenge: Annual ImageNet competition no longer held after 2017 -> now moved to Kaggle.

“Squeeze-and-Excitation Networks”, CVPR 2018, https://arxiv.org/pdf/1709.01507

Improving ResNet ...

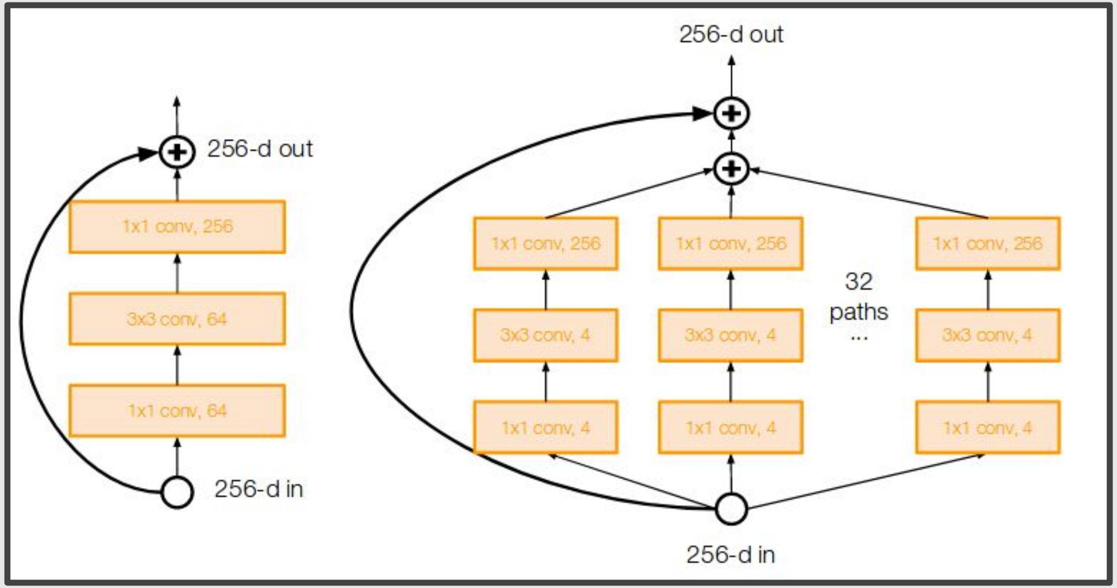
Identity Mappings in Deep Residual Networks [He et al., 2016]

- Creates a more direct path for propagating information throughout network (moves activation to residual mapping pathway)
- Gives better performance



Improving ResNet ...

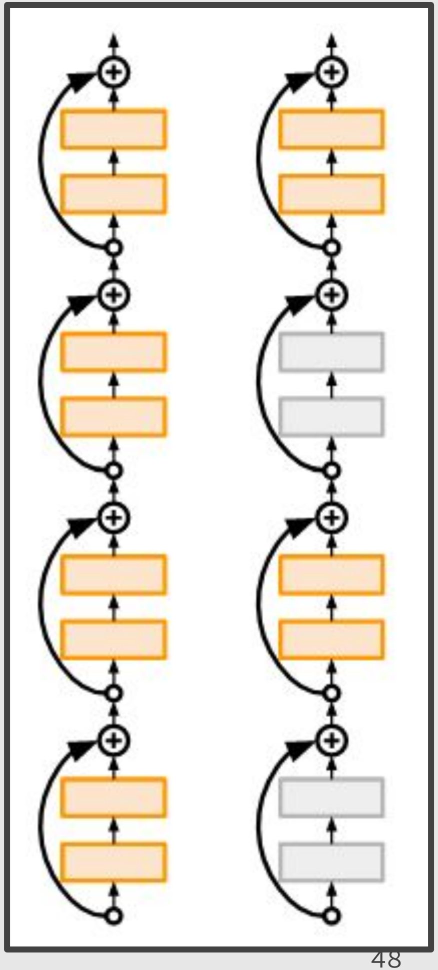
Aggregated Residual Transformations for Deep Neural Networks (**ResNeXt**) [Xie et al., 2016]



Improving ResNet ...

Deep Networks with Stochastic Depth
[Huang et al., 2016]

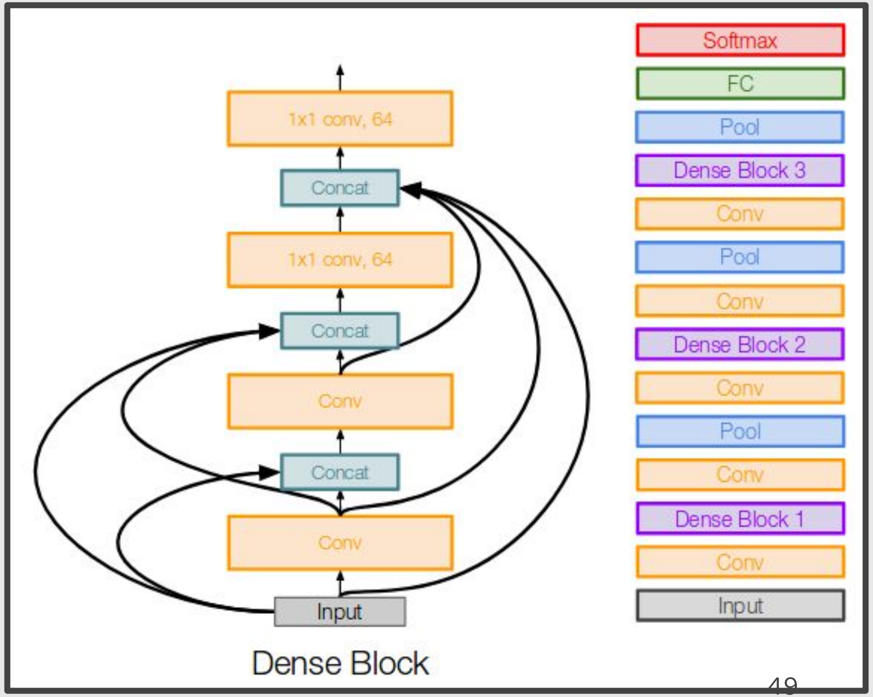
- Motivation: reduce vanishing gradients
- **Randomly drop** a subset of layers during each training pass
- Bypass with identity function

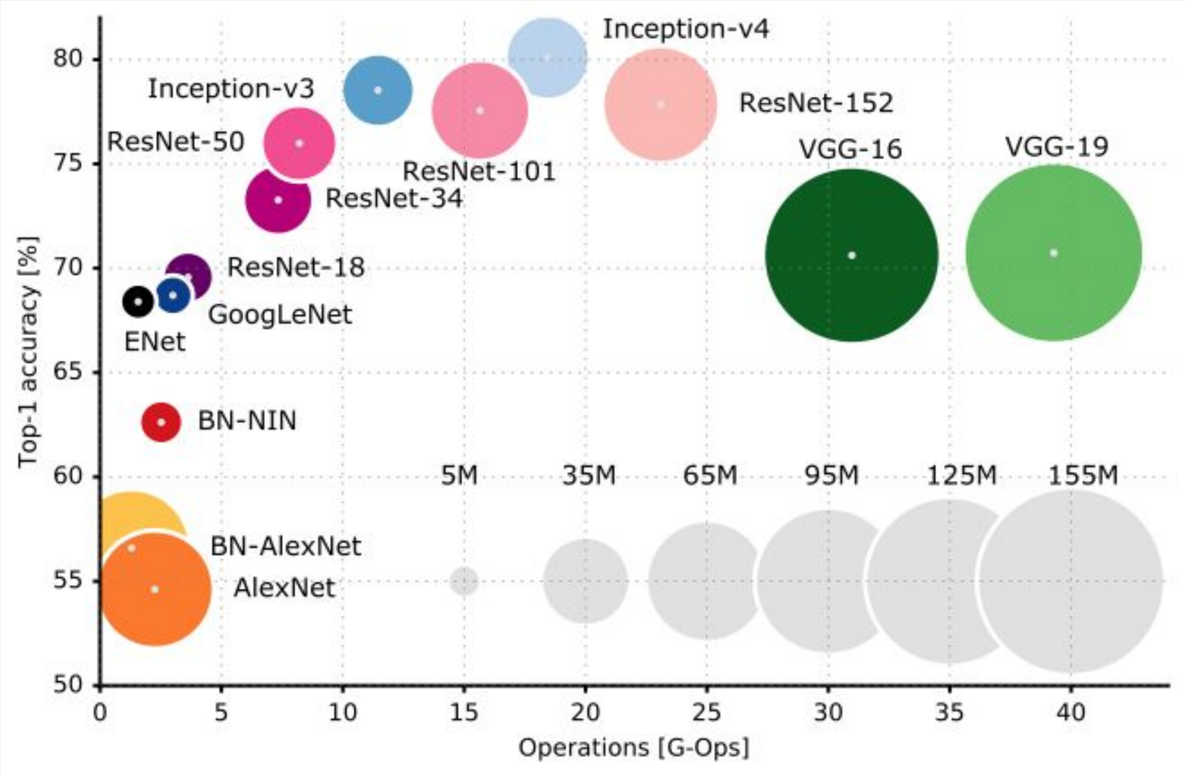


Beyond ResNet ...

Densely Connected Convolutional Networks (**DenseNet**) [Huang et al., 2017]

- Each layer is connected to every other layer in feedforward fashion

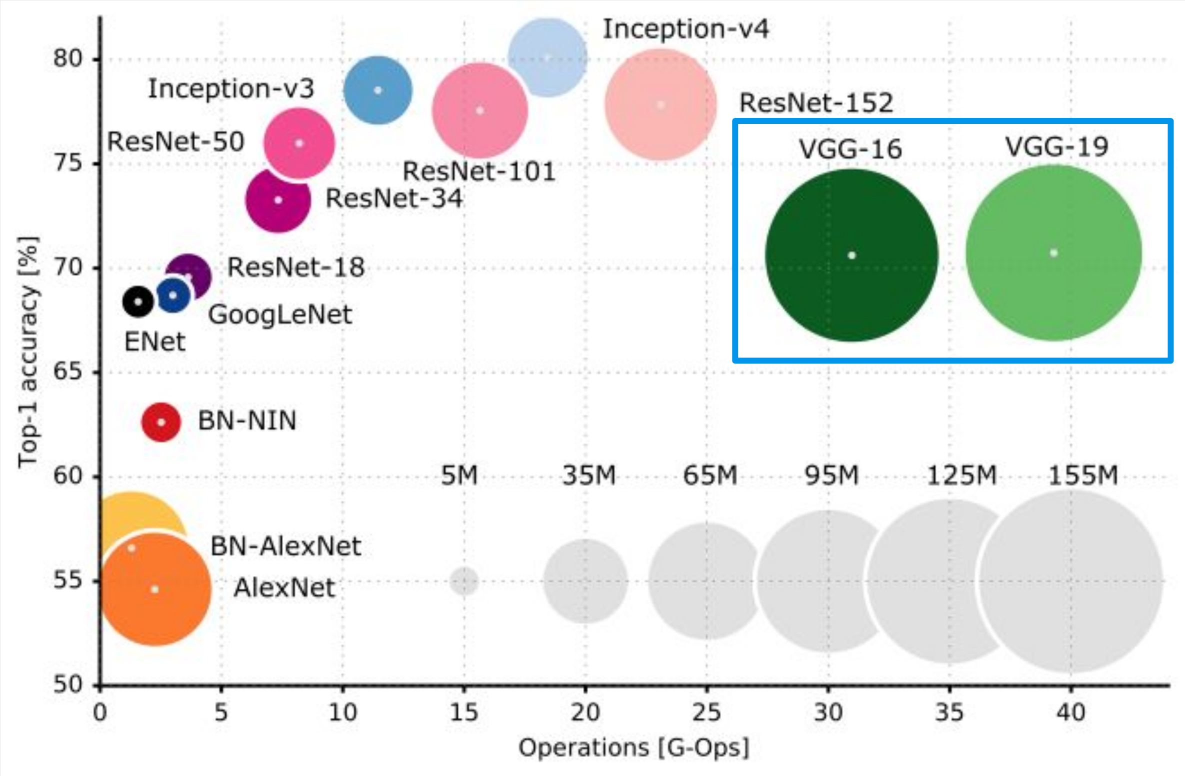




The size of the blobs is proportional to the number of network parameters.

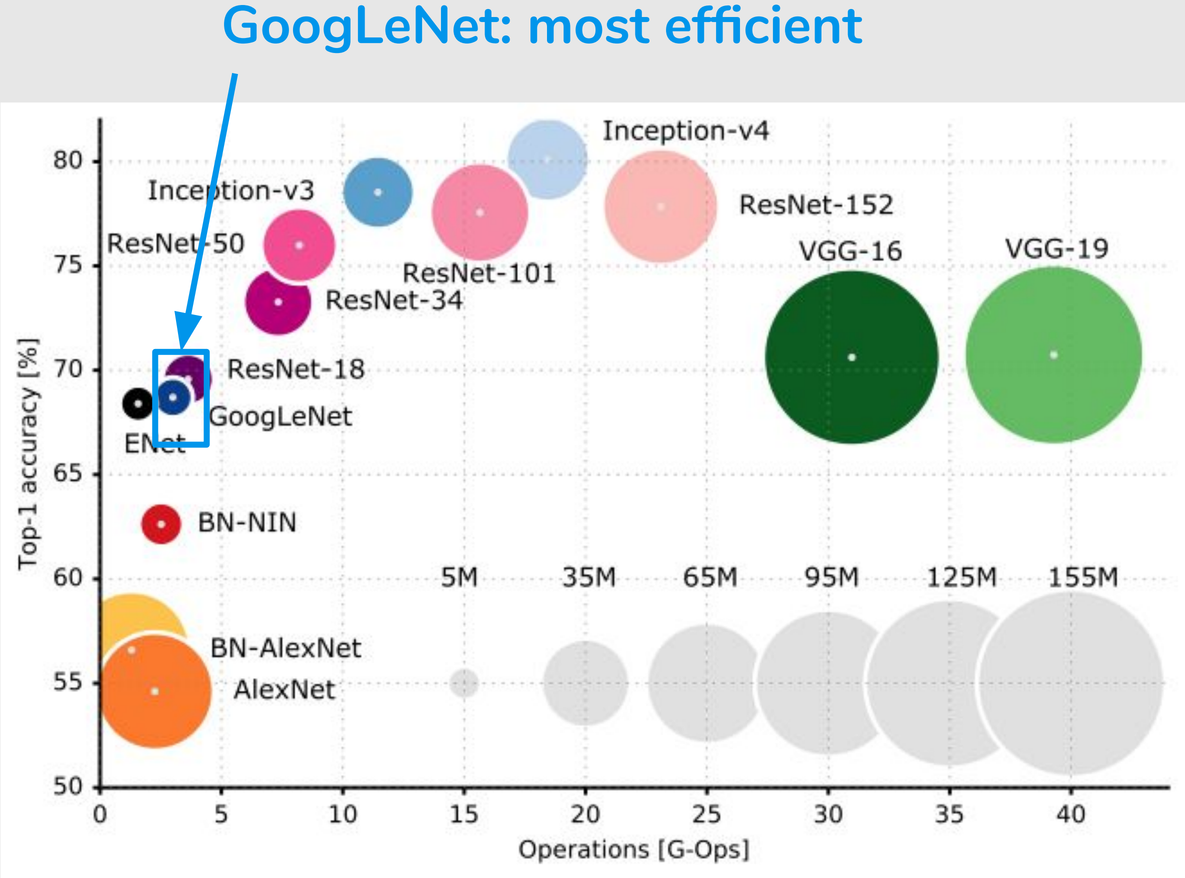
https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba

VGG: Highest memory, most operations



The size of the blobs is proportional to the number of network parameters.

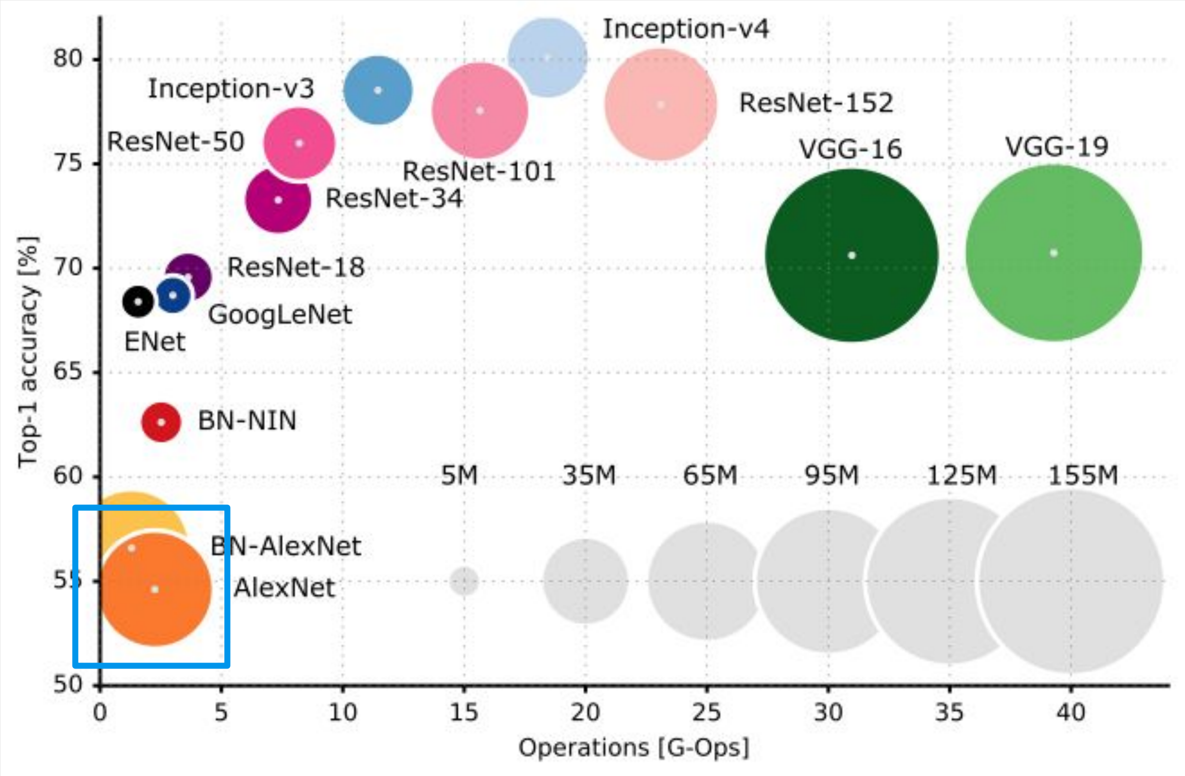
https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba



The size of the blobs is proportional to the number of network parameters.

https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba

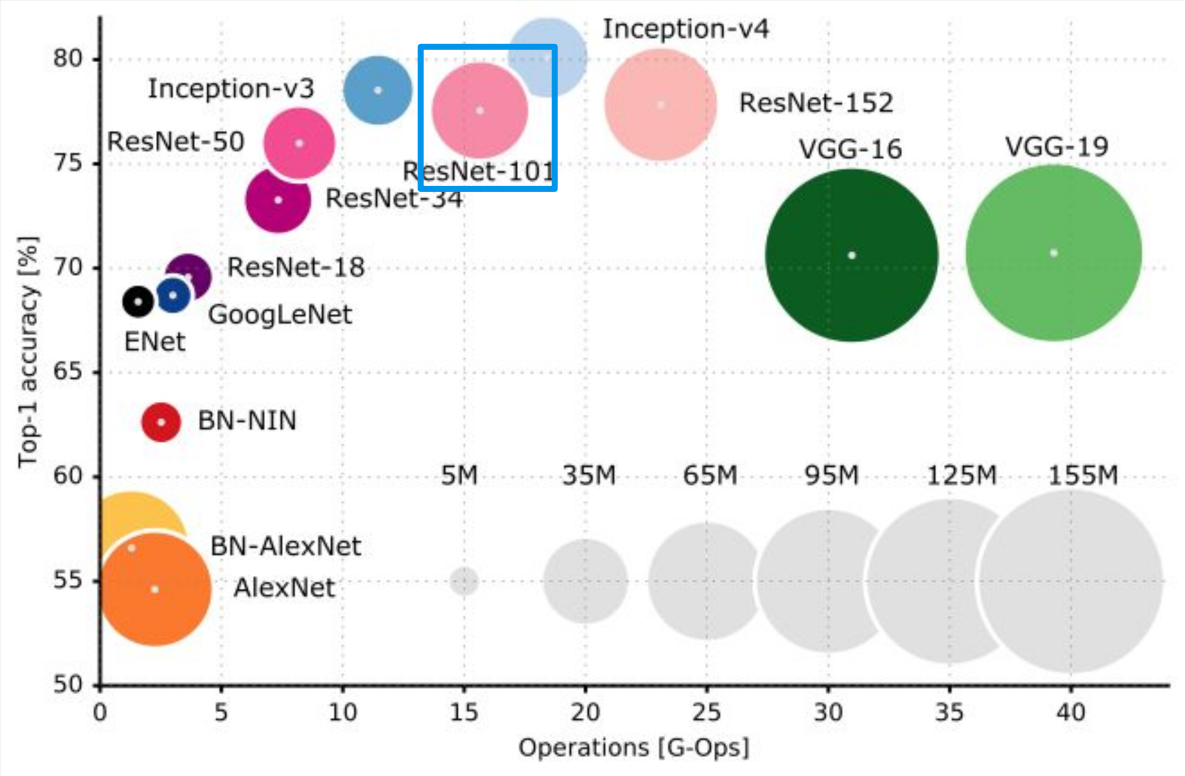
AlexNet: Smaller compute, still memory heavy, lower accuracy



The size of the blobs is proportional to the number of network parameters.

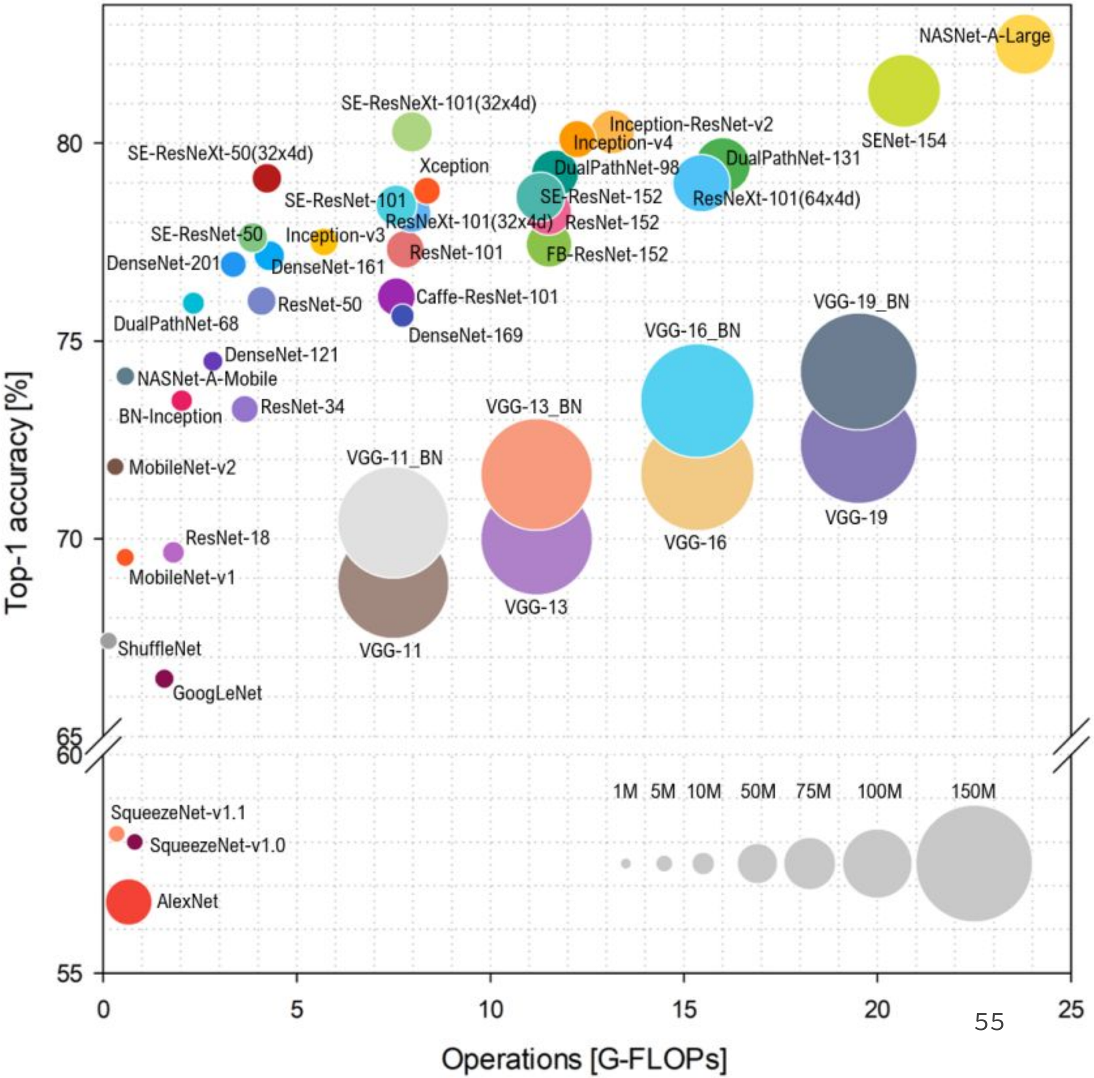
https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba

ResNet: Moderate efficiency depending on model, highest accuracy



The size of the blobs is proportional to the number of network parameters.

https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba

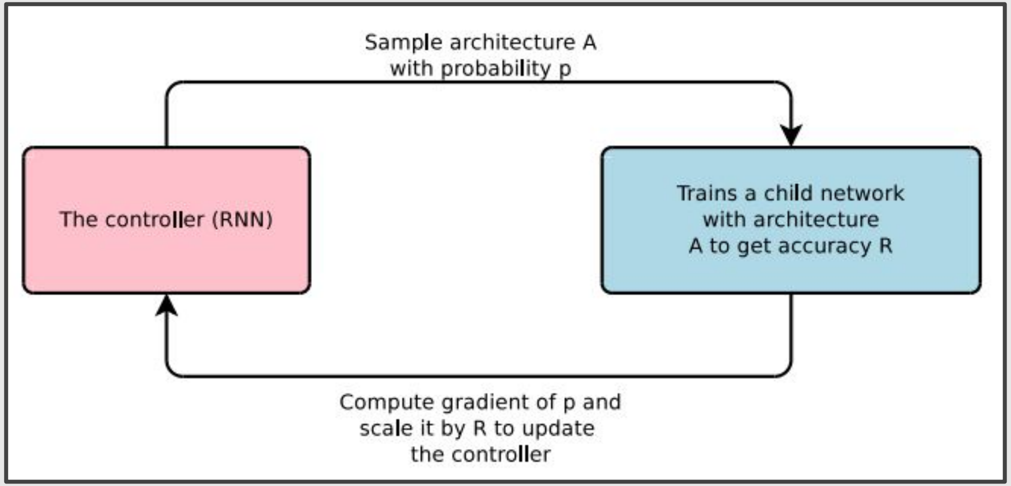


“Benchmark Analysis of Representative Deep Neural Network Architectures”, Nov. 2018. https://doi.org/10.1109/ACCESS.2018.2877890

Learning to learn network architectures...

Neural Architecture Search with Reinforcement Learning (NASNet) [Zoph and V. Le, 2016]

- “Controller” network that learns to design a good network architecture (output a string corresponding to network design)



Summary: CNN Architectures

- Many popular architectures available in **model zoos**
- **ResNet and SENet** currently good defaults to use
- Networks have gotten increasingly deep over time
- Many other aspects of network architectures are also **continuously being investigated and improved**
- Even more recent trend towards **meta-learning**