Spring 2024

CMPE 258-01 Deep Learning

Dr. Kaikai Liu, Ph.D. Associate Professor

Department of Computer Engineering

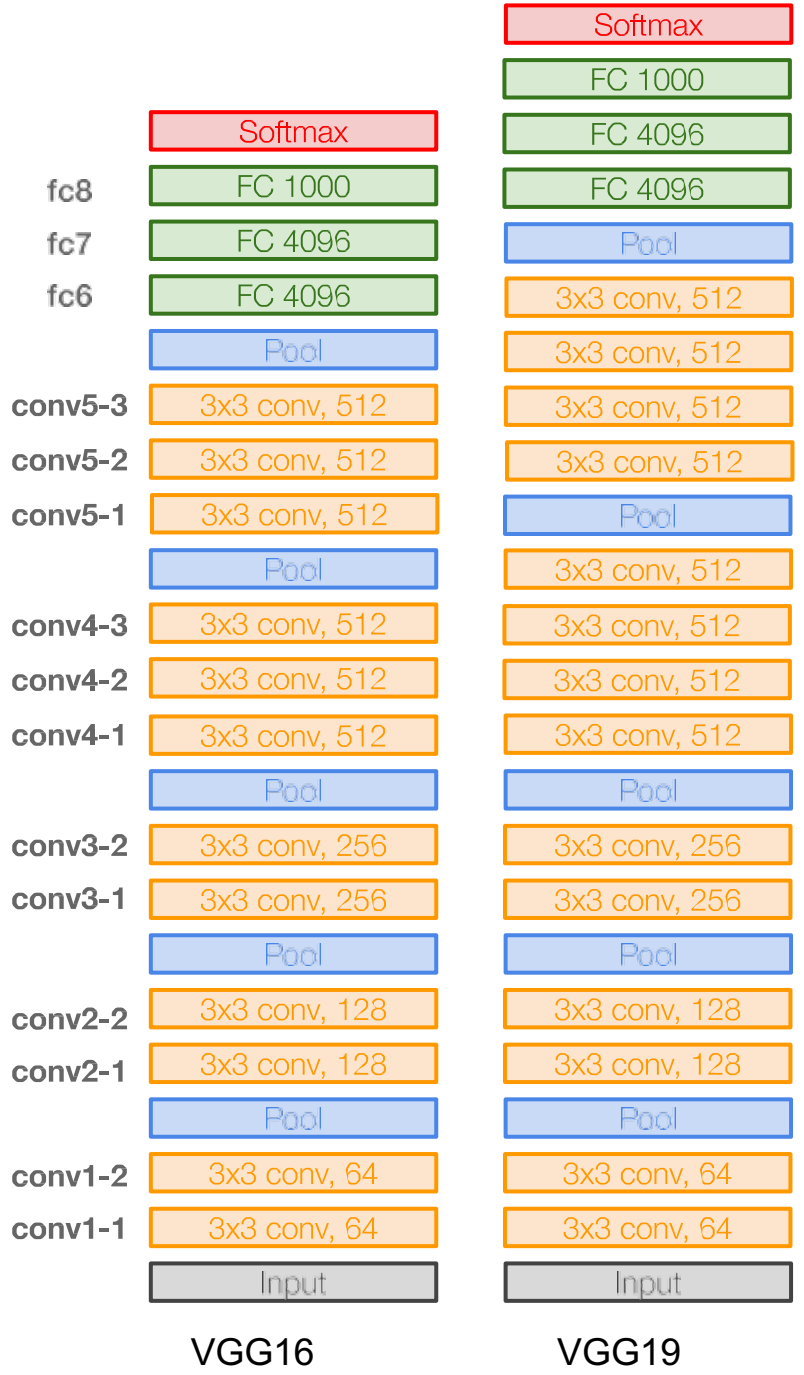
San Jose State University

Email: kaikai.liu@sjsu.edu

Website: https://www.sjsu.edu/cmpe/faculty/tenure-line/kaikai-liu.php

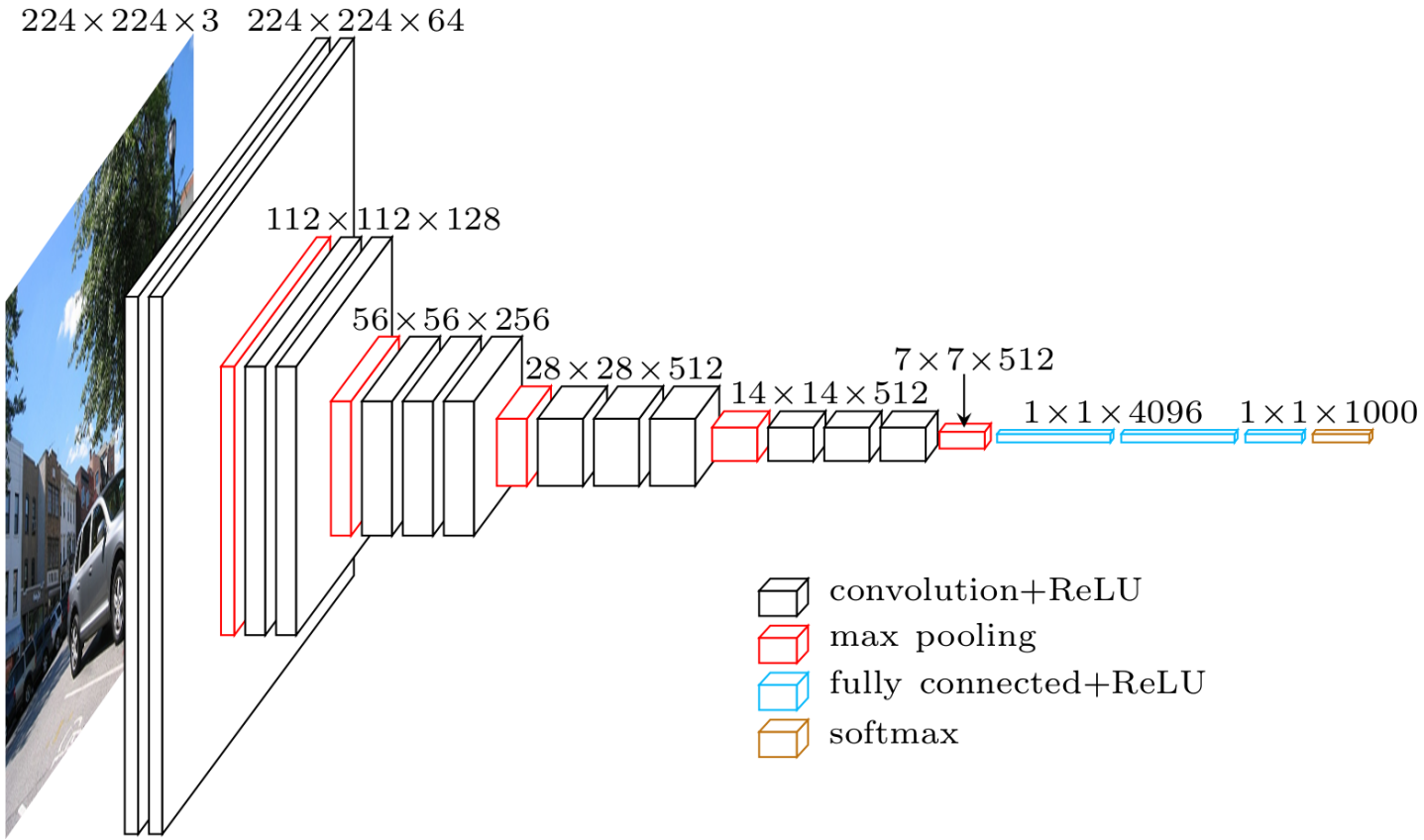
VGGNet (2014)

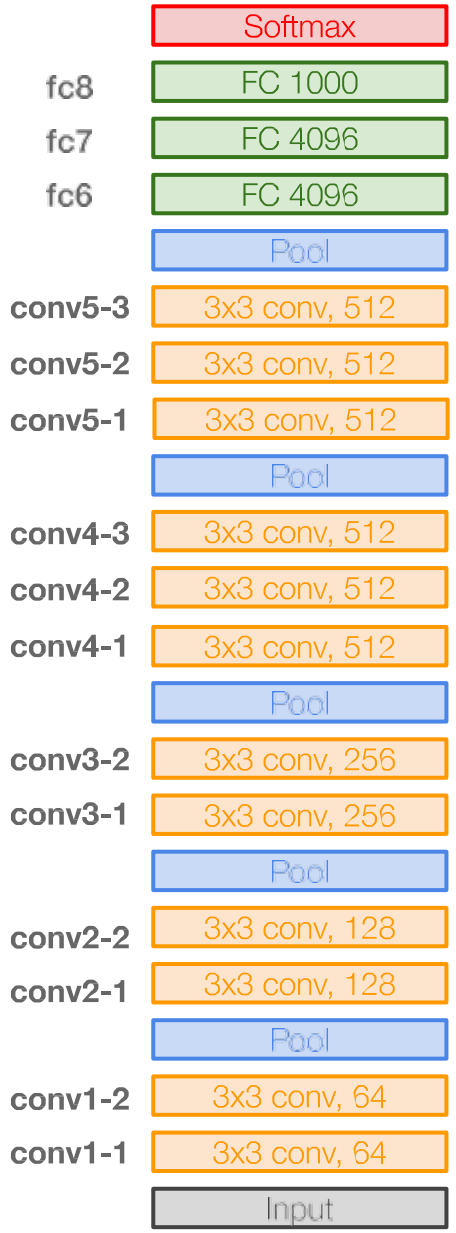
- Simonyan, K.; Zisserman, A., ***Very deep convolutional networks for large-scale image recognition***, International Conference on Learning Representations (2015)
  - Stack many convolutional layers before pooling
  - Use "same" convolutions (3x3) to avoid resolution reduction
  - Up to 19 layers
  - The main principle is that a stack of three 3×3 conv. layers are similar to a single 7×7 layer. And maybe even better! Because they use three non-linear activations in between (instead of one), which makes the function more discriminative.
    - $3*(3^2)C^2=27\times C^2$ weights, compared to a 7×7 conv. layer that would require $1*(7^2)\,C^2 = 49C^2$ parameters (81% more)



VGG16

- VGG16
  - the input is a 224x224x3 tensor (that means a 224x224 pixel RGB image)

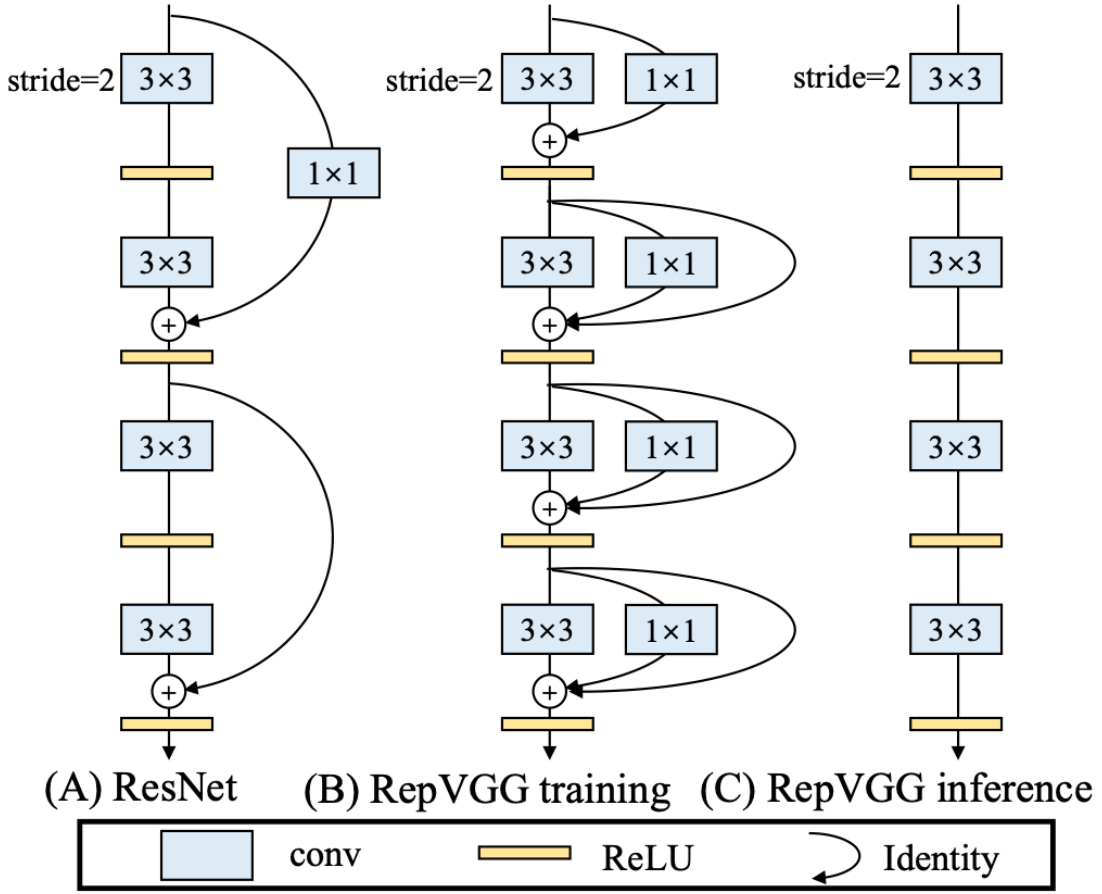




VGG16

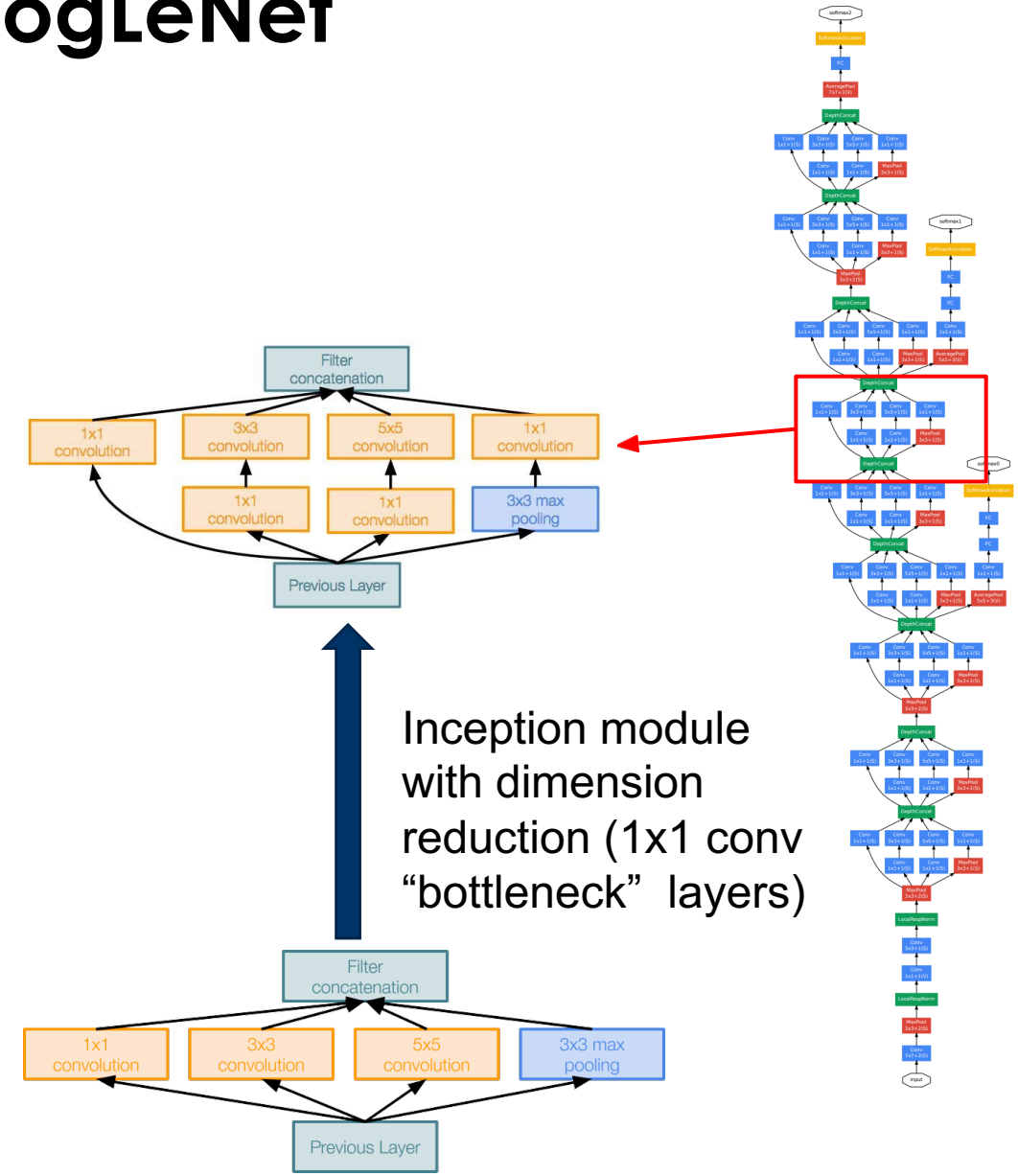
RepVGG

- RepVGG: Making VGG-style ConvNets Great Again
  - Paper (2021): https://arxiv.org/abs/2101.03697
  - It is challenging for a plain model to reach a comparable level of performance as the multi-branch architectures
  - Since the benefits of multi-branch architecture are all for training and the drawbacks are undesired for inference, we propose to decouple the training-time multi-branch and inference-time plain architecture via structural re-parameterization, which means converting the architecture from one to another via transforming its parameters.



Case Study: GoogLeNet

- GoogLeNet: 22 layers
  - ILSVRC'14 classification winner (6.7% top 5 error)
  - part of these layers are a total of 9 inception modules
  - The input layer of the GoogLeNet architecture takes in an image of the dimension 224 x 224
- Efficient "Inception" module: Multi-scale feature extraction
  - "Inception module": design a good local network topology (network within a network) and then stack these modules on top of each other
  - using kernels of various sizes in parallel and thus increasing the width of the model instead of depth
  - extract both bigger and smaller features simultaneously



Naive Inception module

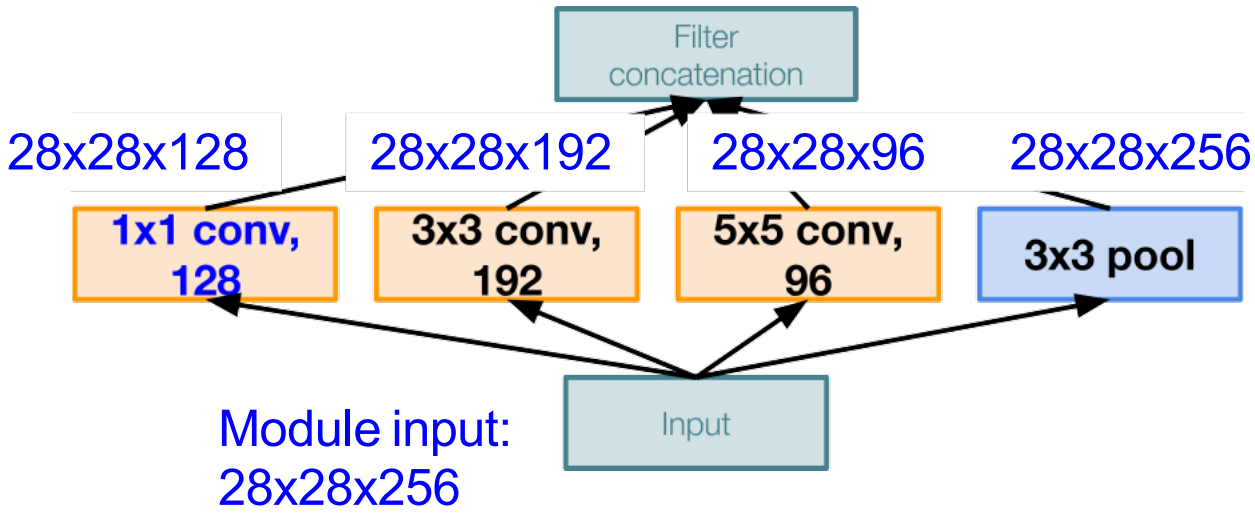
Case Study: GoogLeNet

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



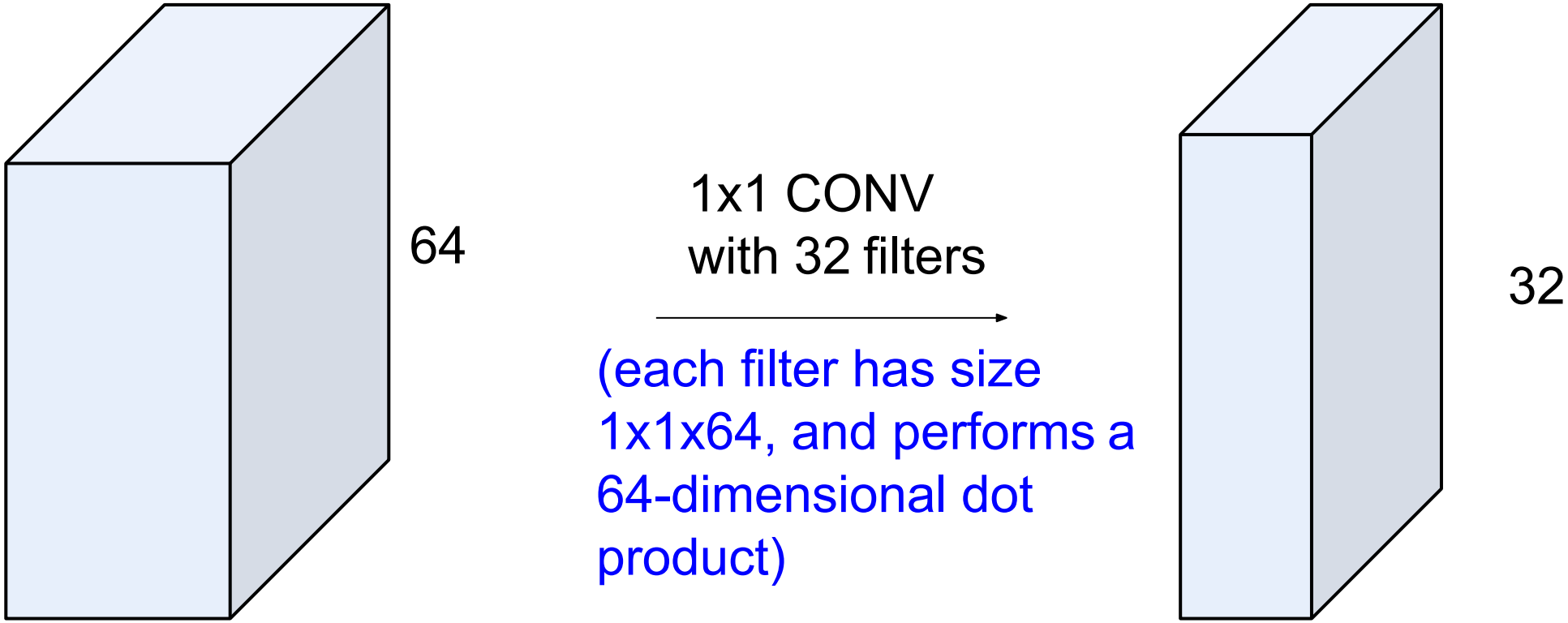
Naive Inception module

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!

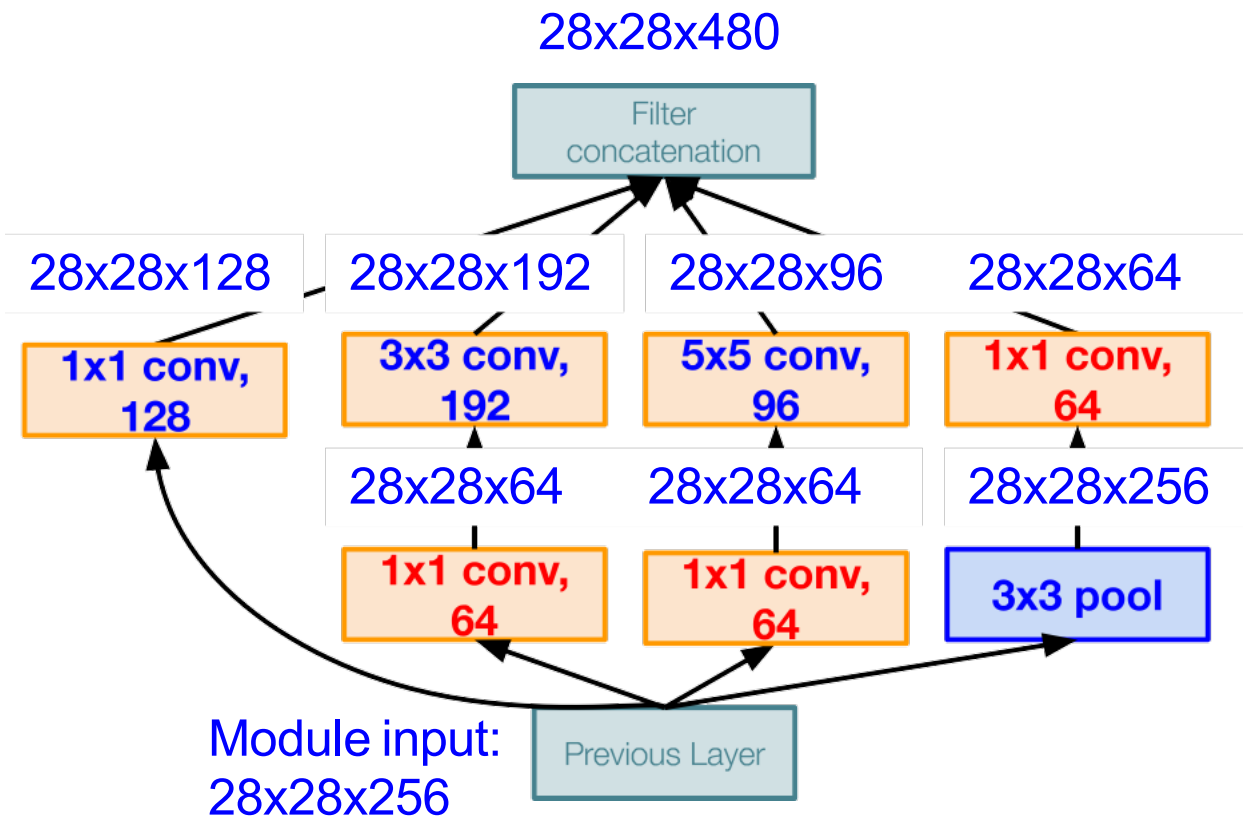
Reminder: 1x1 convolutions



Preserves spatial dimensions, reduces depth!

Projects depth to lower dimension (combination of feature maps)

Case Study: GoogLeNet



Inception module with dimension reduction

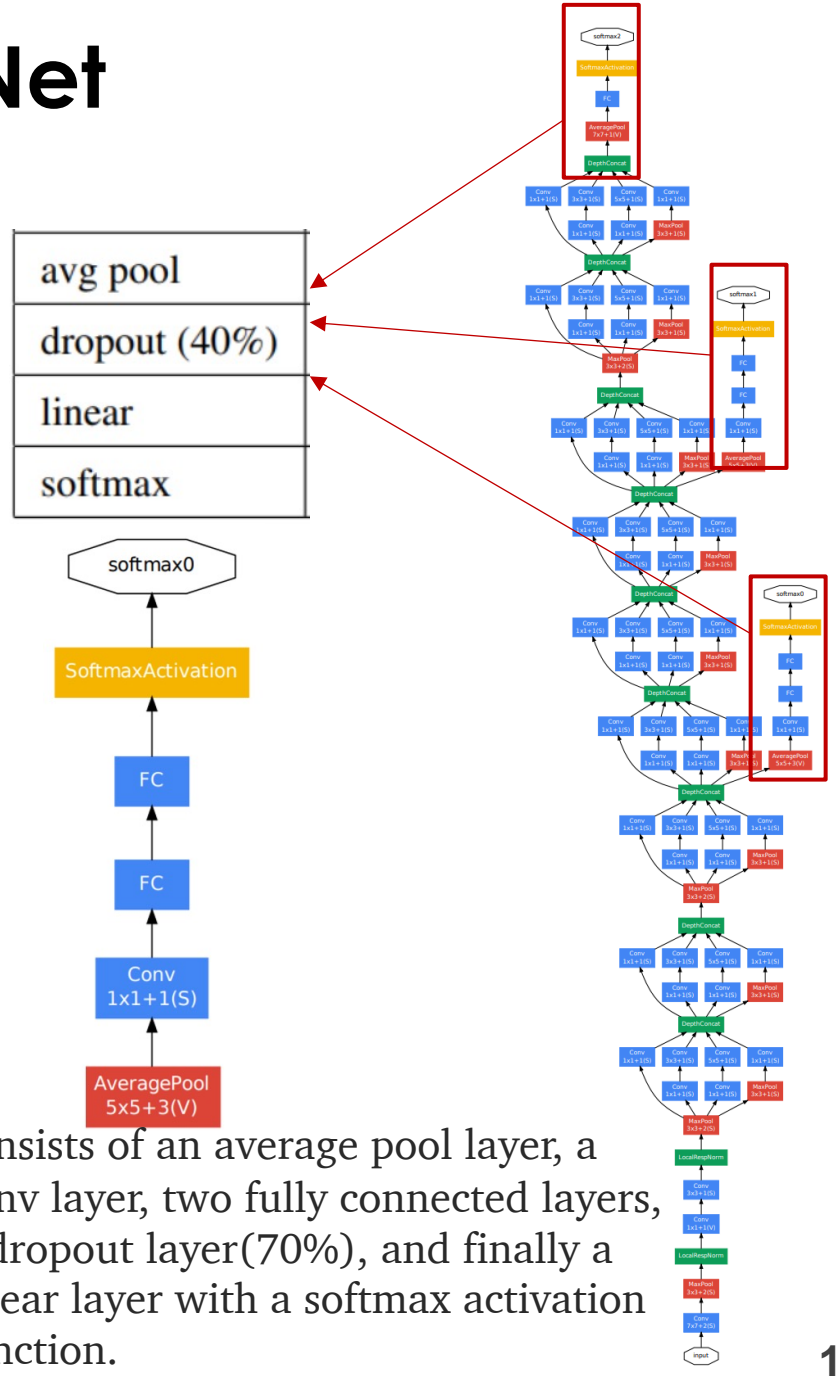
Using same parallel layers as naive example, and adding "1x1 conv, 64 filter" bottlenecks:

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

Case Study: GoogLeNet

- GoogLeNet: classifier
  - The average pooling layer takes a mean across all the feature maps produced by the last inception module and reduced the input height and width to 1x1
  - A dropout layer(40%) is utilised just before the linear layer. The dropout layer is a regularisation technique that is used during training to prevent overfitting of the network.
- Auxilary Classifiers
  - One main problem of an extensive network is that they suffer from vanishing gradient descent.
  - Vanishing gradient descent occurs when the update to the weights that arises from backpropagation is negligible within the bottom layers as a result of relatively small gradient value.
  - Auxilary Classifiers are added to the intermediate layers of the architecture, namely the third(Inception 4[a]) and sixth (Inception4[d])
  - Auxilary Classifiers are only utilised during training and removed during inference. The purpose of an auxiliary classifier is to perform a classification based on the inputs within the network's midsection and add the loss calculated during the training back to the total loss of the network.



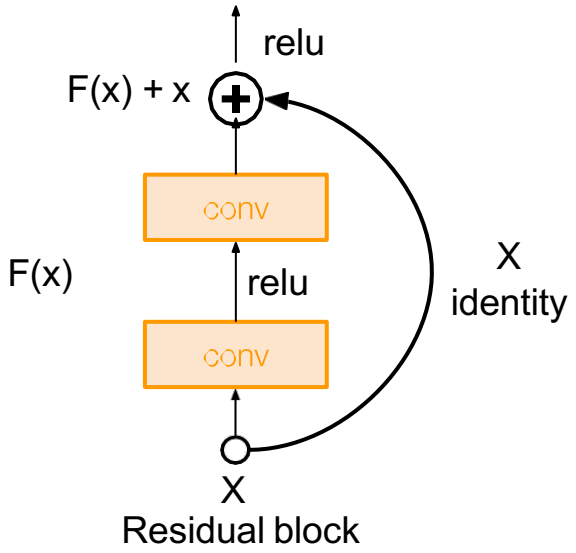
consists of an average pool layer, a conv layer, two fully connected layers, a dropout layer(70%), and finally a linear layer with a softmax activation function.

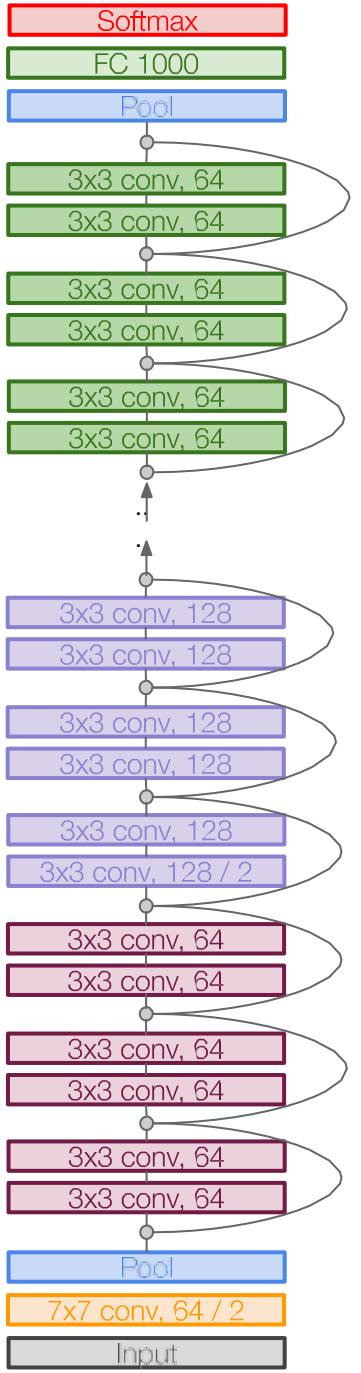
Inception Model

- Inception V2: Rethinking the Inception Architecture for Computer Vision
  - Factorize 5x5 and 7x7 (in InceptionV3) convolutions to two and three 3x3 sequential convolutions respectively. This improves computational speed. This is the same principle as VGG.
  - They used spatially separable convolutions. Simply, a 3x3 kernel is decomposed into two smaller ones: a 1x3 and a 3x1 kernel, which are applied sequentially.
  - They tried to distribute the computational budget in a balanced way between the depth and width of the network.
  - They added batch normalization.
- Inceprion V4: https://arxiv.org/abs/1602.07261

ResNet (2015)

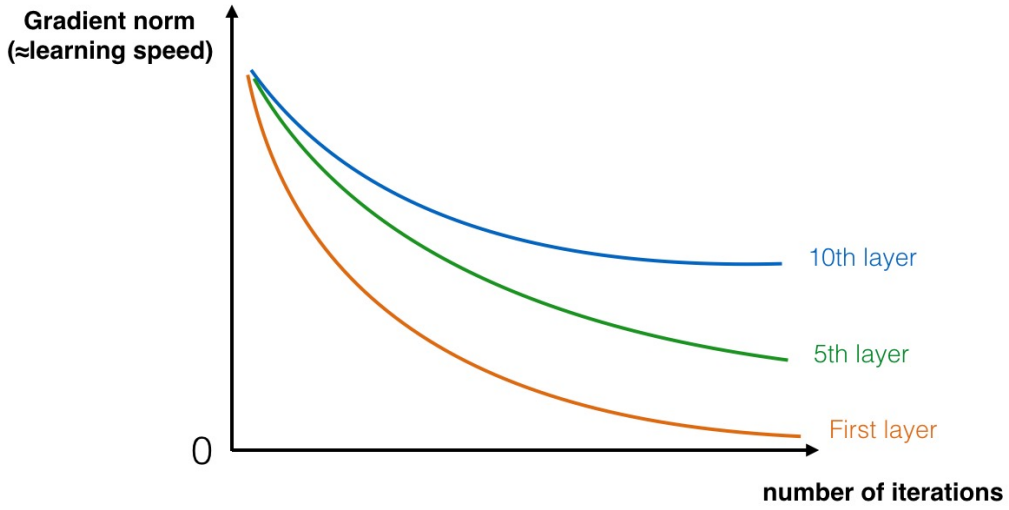
- He, K. et al., ***Deep residual learning for image recognition***, IEEE conference on computer vision and pattern recognition (2016)
  - Very deep networks using residual connections
  - 152-layer model for ImageNet
  - ILSVRC'15 classification winner (3.57% top 5 error)
  - Swept all classification and detection competitions in ILSVRC'15 and COCO'15!

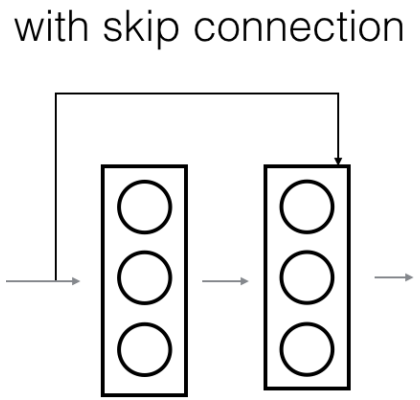




ResNet (2015)

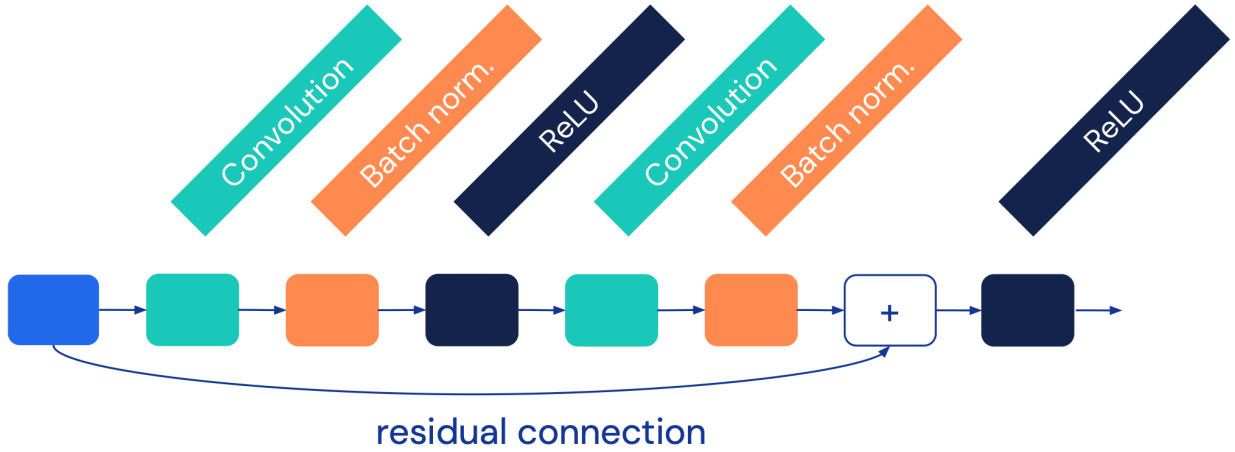
- ResNet
  - Problem: during gradient descent, as you backprop from the final layer back to the first layer, you are multiplying by the weight matrix on each step, and thus the gradient can decrease exponentially quickly to zero.
  - In ResNets, a "shortcut" or a "skip connection" allows the gradient to be directly backpropagated to earlier layers
  - Nowadays, ResNet architectures have mostly replaced VGG as a base network for extracting features

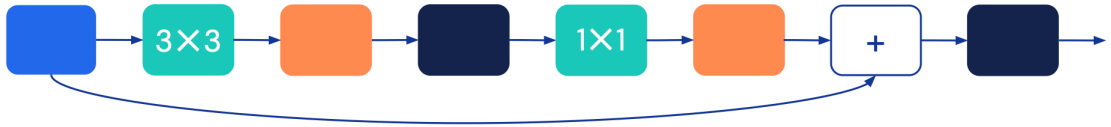


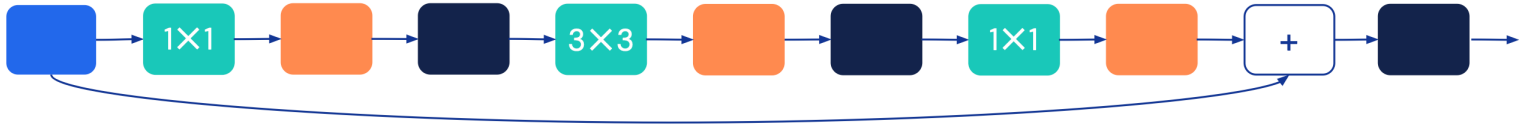


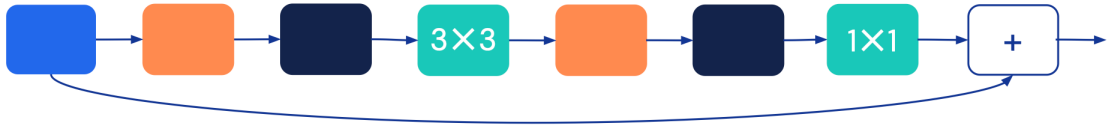
ResNet (2015)

- Residual connections facilitate training deeper networks
- Different flavours
  - ResNet V2 (bottom) avoids all nonlinearities in the residual pathway
  - Up to 152 layers









Thank You

Address:
ENG257, SJSU

Email Address:
Kaikai.liu@sjsu.edu