Deep Networks - VGGNet, GoogLeNet

Pattern Recognition & Machine Learning Laboratory
Ji-Sang Hwang, Aug 3, 2021

VGGNet & GoogLeNet

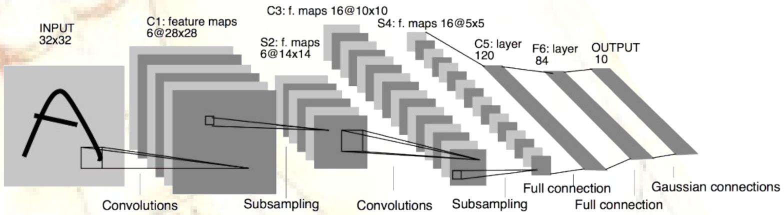
- **Introduction**
  - Improvement in object classification and detection capabilities with deep learning and convolutional networks (ConvNets)
  - Progress is a consequence of new ideas, algorithms and improved network architectures
- **Discussion**
  - Receptive smaller window size and smaller stride of the first convolutional layer [*Zeiler & Fergus*, 2013; *Sermanet et al.*, 2014]
  - Training and testing networks densely over the whole image and over multiple scales [*Sermanet et al.*, 2014; *Howard*, 2014]
  - These papers,
    - Address another important aspect of ConvNet architecture design
    - Deep / Depth
      - Increased network depth
      - A new level of organization in the form of the "Inception module"

Meme of the Inception

Zeiler, M. D. and Fergus, R. "Visualizing and understanding convolutional networks", *ECCV*, 2014.
Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks", *ICLR*, 2014.

Related Work

- **Contributions**
  - **LeNet-5 [*LeCun et al.,* 1989]**
    - **Standard structure of Convolution Neural Networks (CNN)**
      - Stacked convolution layers (optionally followed by contrast normalization and max-pooling)
      - One or more fully-connected layers
      - For large datasets,
        - » Increase the number of layers
        - » Increase layer size
        - » Using dropout to address overfitting



Architectures of LeNet-5

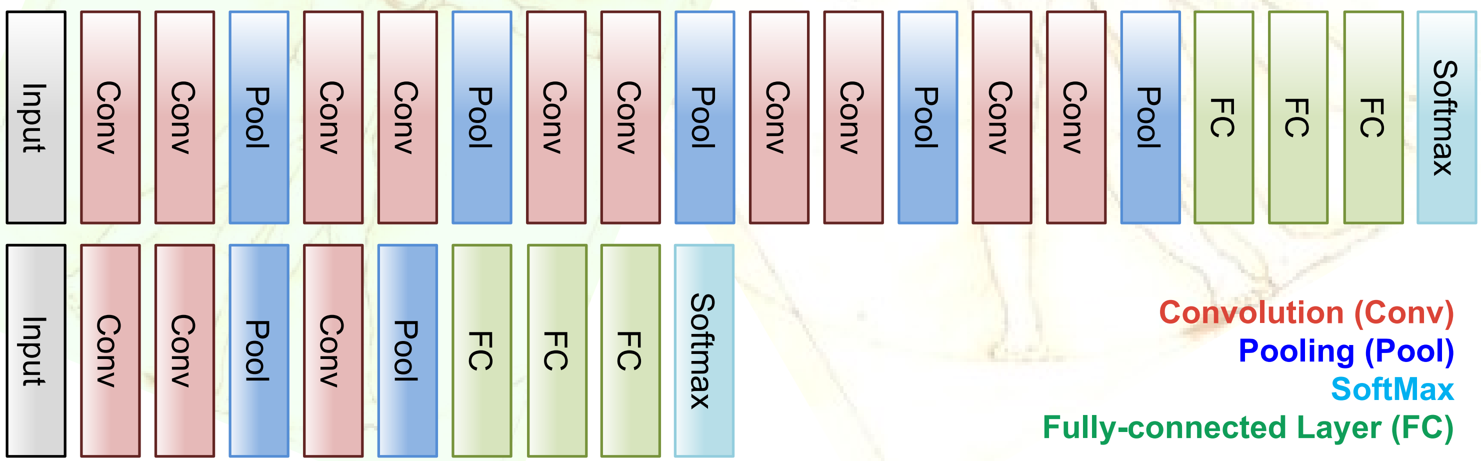
  - **Network-in-Network [*Lin et al.,* 2013]**
    - **Increase the representational power of neural networks**
    - **Add Additional convolutional layers to the network for increasing its depth and adding non-linearity**
  - **Regions with Convolution Neural Networks (R-CNN) [*Girshick et al.,* 2014]**
    - **Utilizing low-level cues in order to generate object location proposals in a cate-gory-agnostic fashion**
    - **Using CNN classifiers to identify object categories at those location**

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. "Backpropagation applied to handwritten zip code recognition", *Neural Computation*, 1989.
M. Lin, Q. Chen, and S. Yan. "Network in network", *CoRR*, 2013.
R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation", *Computer Vision and Pattern Recognition,* 2014. *CVPR* 2014. *IEEE Conference on*, 2014.

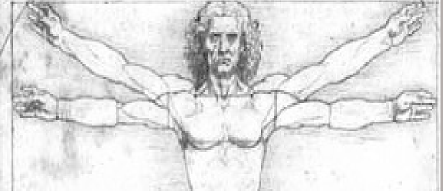
Very Deep Convolutional Networks for Large-Scale Image Recognition [*K. Simonyan et al.*] (1/2)

Architecture of VGGNet

- **Use convolution filter (smallest size to capture the notion of left/right, up/down, center)**
  - **Reason of using convolution filter**
    - **3 non-linear rectification layers make the decision function more discriminative**
    - **Decrease the number of parameters**
      - 3-layer convolution stack : , **: Channel**
      - 1-layer convolution stack :
- **A stack of convolutional layers is followed by 3 Fully-Connected layers**
- **Hidden layers are equipped with rectification (Rectified Linear Unit (ReLU))**



Architectures of VGGNet-13 (Top) and AlexNet (Bottom)



Very Deep Convolutional Networks for Large-Scale Image Recognition [*K. Simonyan et al.*] (2/2)

- **Differ only in the depth**
  - **From 11 weight layers in the network A to 19 weight layers in the network E**
  - **Using Local response normalization (LRN) does not improve on the model a without any normalization layer**
  - **convolution filter is a way to increase the non-linearity of decision function without affecting the receptive fields of the convolutional layers**
  - **Using pre-initialized layers to prohibit stalling learning due to instability of gradient in deep nets**
    - **Initialized first 4 convolutional layers and the last 3 fully-connected layers of network A**
    - **Did not decrease the learning rate for pre-initialized layers when training another networks**

Table of ConvNet configurations

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table of number of parmaters (in millions)

Network	A, A-LRN	B	C	D	E
Number of parameters	133M	133M	134M	138M	144M

Going Deeper with Convolutions [*C. Szegedy et al.*] (1/2)

- **Method**
  - Architecture of The Inception
    - Consider how an optimal local sparse structure of a convolutional network can be approximated and covered by readily available dense components
    - Problems of Naïve Version
      - A modest number of convolutions can be prohibitively expensive
      - Leading to a computational blow up within a few stage
    - Solving problems with convolutional layer
      - Using 'bottleneck' layers to compute reductions before the expensive and convolutions
      - Including the use of rectified linear activation for adding non-linearity

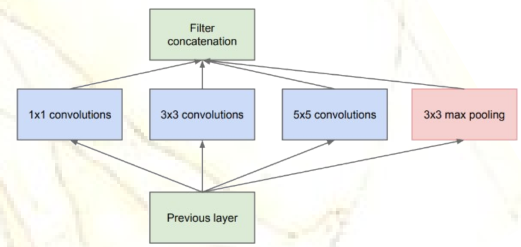
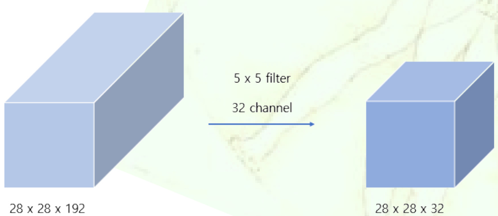
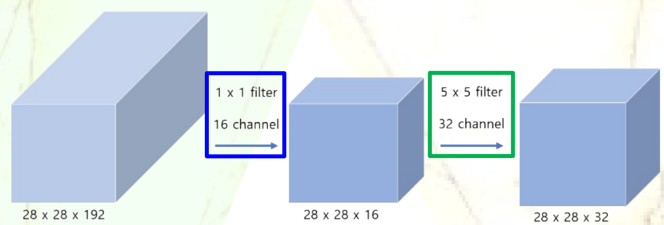


Image of inception module(Naïve Version)



5x5 convolution



5x5 convolution with 1x1 convolution

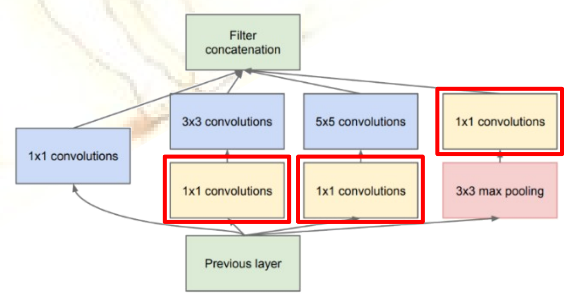
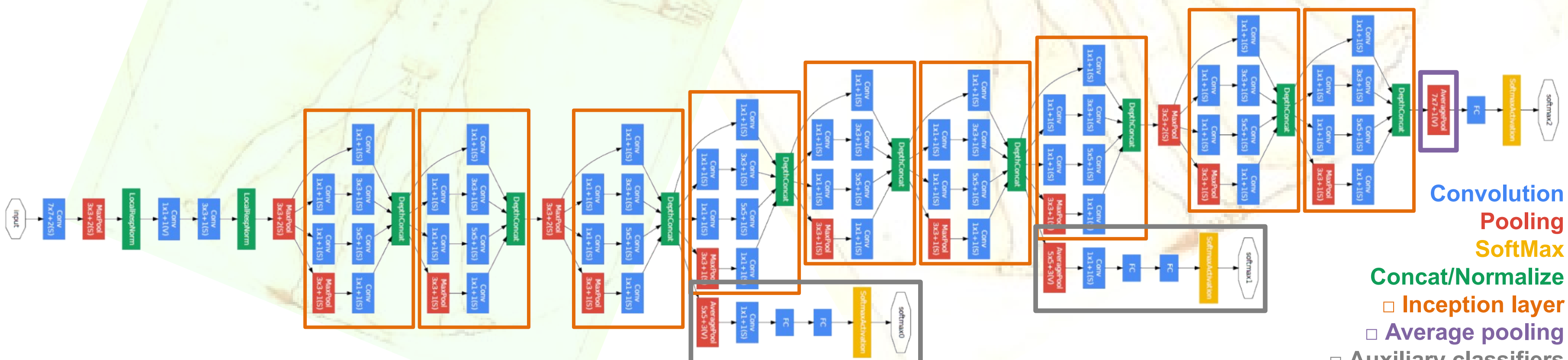


Image of inception module(Dimension reduction)

Going Deeper with Convolutions [*C. Szegedy et al.*] (2/2)

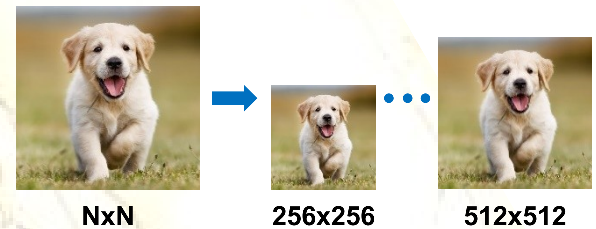
- **Architecture of GoogLeNet**
  - 22 layers deep when counting only layers with parameters
  - The use of 'average pooling' before the classifier enables to easily adapt net-works to other label sets
  - Adding 'auxiliary classifiers' to combat the vanishing gradient problems while providing regularization
    - An average pooling layer with filter size and stride 3
    - A convolution with 128 filters for dimension reduction and rectified linear activa-tion
    - A fully connected layer with 1024 units and rectified linear activation
    - A linear layer with softmax loss as the classifier



GoogLeNet architecture

Training Models

- **Method**
  - **Training VGGNet**
    - **Using stochastic gradient descent (SGD) with momentum**
      - Batch size : 256 / momentum : 0.9
    - **Regularized by weight decay and dropout**
      - L2 penalty multiplier :
      - Dropout ratio : 0.5 (First 2 fully-connected layers)
    - **Randomly cropped from rescaled training images**
      - 1 crop per image per SGD iteration
      - Single-scale training
        - » Fix Scale () : and
        - » Pretrained with and trained with initial learning rate of
      - Multi-scale training (Called scale jittering)
        - » Rescaled by randomly sampling from a certain range



Example of Multi-scale training

  - **Training GoogLeNet**
    - **Using asynchronous stochastic gradient descent with momentum**
      - Momentum : 0.9
    - **Regularized by fixed learning rate schedule**
      - Decreasing the learning rate by 4% every 8 epochs

Conclusion

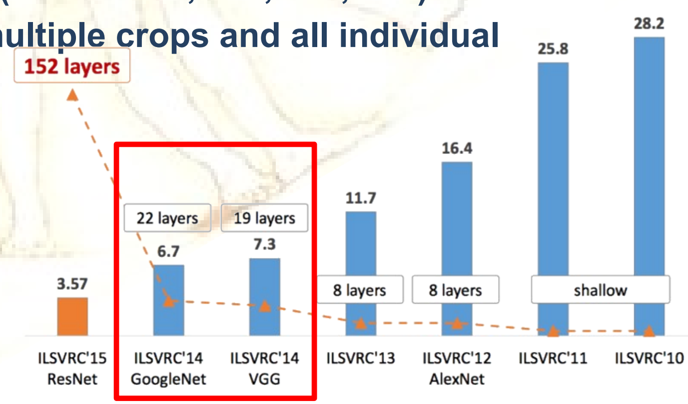
- **Result**
  - ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 Classification Challenge Result

Table of classification performance in ILSVRC 2014

Team	Year	Place	Error (top-5)	Uses external data	Layers	Parms
GoogLeNet	2014	1st	6.67%	No	22 layers	5 M
VGG	2014	2nd	7.32%	No	19 layers	144 M

- **Setup of GoogLeNet**
  - Trained independently 7 versions of same GoogLeNet model and performed ensemble prediction with them
  - Aggressive cropping approach during testing (Resize 256, 288, 320, 352)
  - The softmax probabilities are averaged over multiple crops and all individual classifiers to obtain the final prediction

- **Conclusion**
  - **GoogLeNet**
    - Significant quality gain at a modest increase of computational requirements to shallower and narrower architectures
  - **VGGNet**
    - Importance of dept in visual representations



Result of classification performance in ILSVRC