Lab Seminar: 2022. 07. 19.

Going Deeper With Convolutions

(Szegedy et al., CVPR 2015)

IDEALAB

Improving lives through learning

JongHyeon Kim

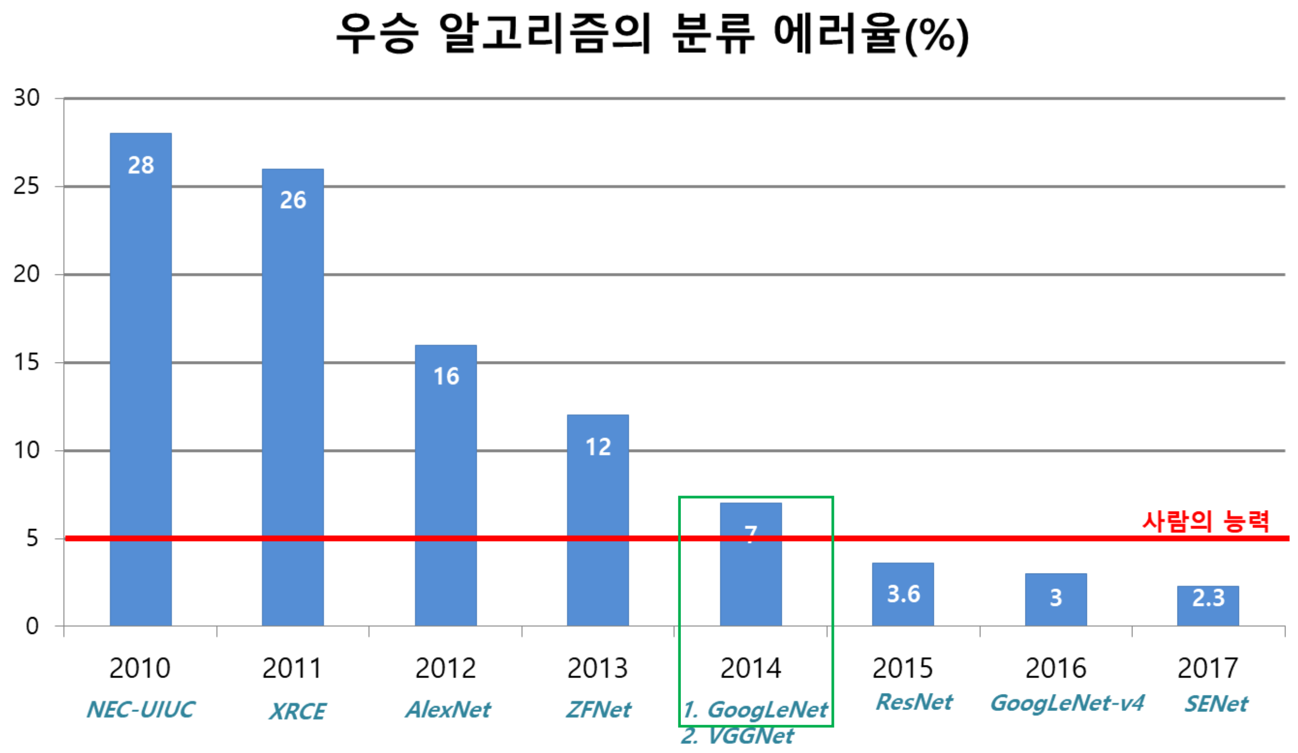
School of Computer Science/Department of AI Convergence Engineering

Gyeongsang National University (GNU)

경상국립대학교 Gyeongsang National University

Contents

Introduction



Introduction

- GoogLeNet: 12x fewer parameters than AlexNet
  - Synergy from deep architecture & classical computer vision (like R-CNN)
- Flexible architecture: for mobile & embedded computing
  - Under 1.5b multiply-adds at inference time
- Code name: Inception
  - From inception movie script "*we need to go deeper*"

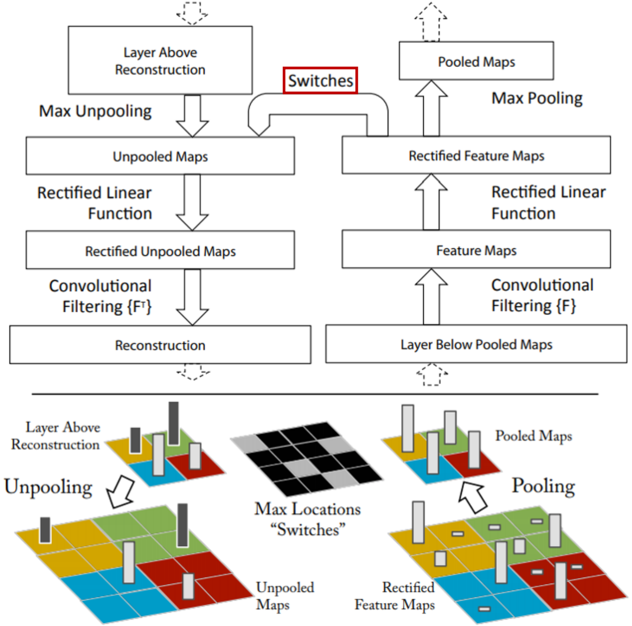
    - Deep means: new architecture (Inception Module), network depth

Related Work

- Visualizing and Understanding Convolutional Networks (Zelier et al., ECCV 2014; ZFNet)
  - Based on AlexNet (filter size: 11x11 -> 7x7, stride: 4 -> 2)
  - Deconvolutional Network (pixel -> feature mapping, pool-relu-conv)

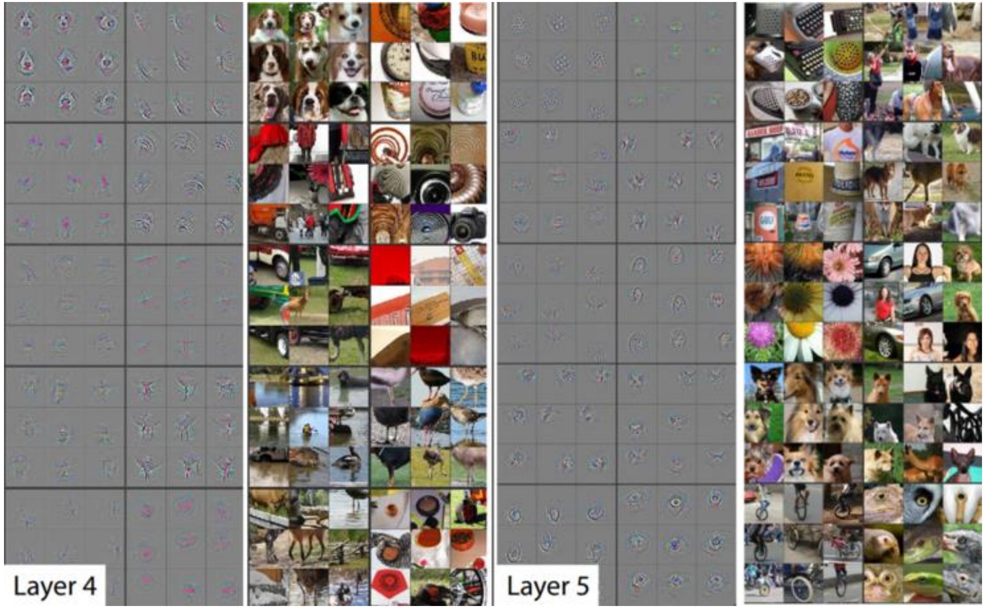
Related Work

- Visualizing and Understanding Convolutional Networks (Zelier et al., ECCV 2014; ZFNet)
  - Switch: memorize most high stimulation locations when max pooling



Related Work

- Visualizing and Understanding Convolutional Networks (Zelier et al., ECCV 2014; ZFNet)
  - Visualize convolutional layers with Switch



Motivation & High Level Considerations

- Simple Solution to Improving Performance (Standard)
  - Increase network size (depth, number of levels, width, ..)
    - Easy and safe way of training higher quality models
    - Works well large amount of labeled data

- Drawbacks
  - size = larger number of parameters -> easy to overfit with small labeled data (need to creation of train data; bottleneck)
  - Dramatically increased use of computational resources
    - Inefficient when most weights end up to be close to zeros

- Drawbacks

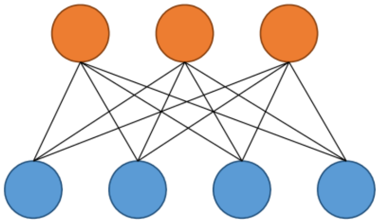
(a) Siberian husky

(b) Eskimo dog

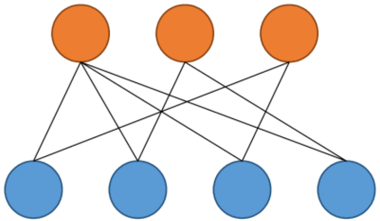
Motivation & High Level Considerations

- Dense vs. Sparse

Densely connected



Sparsely connected



Dense Matrix

1	2	31	2	9	7	34	22	11	5
11	92	4	3	2	2	3	3	2	1
3	9	13	8	21	17	4	2	1	4
8	32	1	2	34	18	7	78	10	7
9	22	3	9	8	71	12	22	17	3
13	21	21	9	2	47	1	81	21	9
21	12	53	12	91	24	81	8	91	2
61	8	33	82	19	87	16	3	1	55
54	4	78	24	18	11	4	2	99	5
13	22	32	42	9	15	9	22	1	21

Sparse Matrix

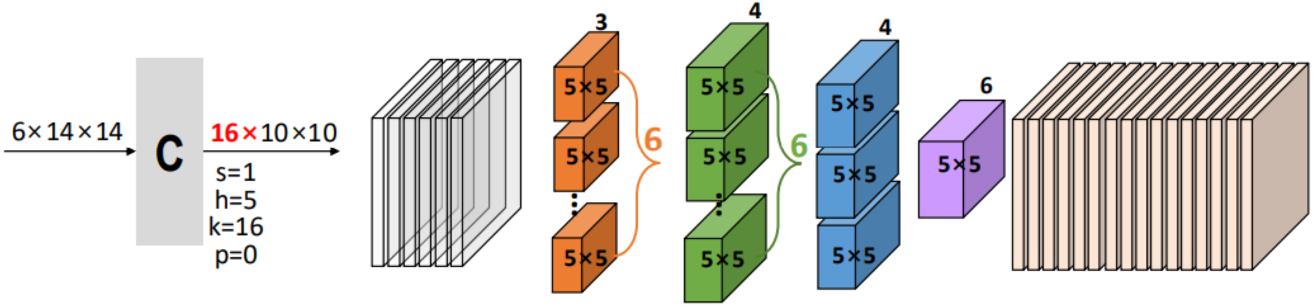
1	.	3	.	9	.	3	.	.	.
11	.	4	.	.	.	.	.	2	1
.	.	1	.	.	.	4	.	1	.
8	.	.	.	3	1	.	.	.	.
.	.	.	9	.	.	1	.	17	.
13	21	.	9	2	47	1	81	21	9
.	.	.	.	.	.	.	.	.	.
.	.	.	.	19	8	16	.	.	55
54	4	.	.	.	11	.	.	.	.
.	.	2	.	.	.	.	22	.	21

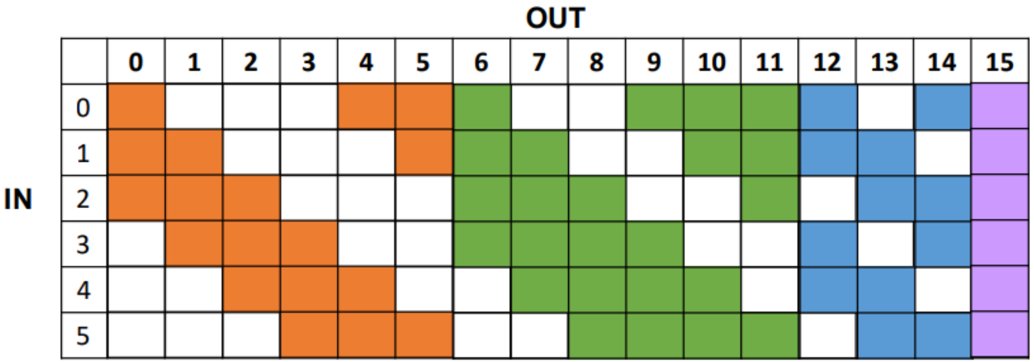
Motivation & High Level Considerations

- Non-uniform sparse data structure: inefficient
  - 100x reduced arithmetic operations: ineffective (cache miss, memory overhead)
- Dense data structure: fast, steadily improving
  - CPU/GPU improvement
  - Highly tuned
  - Numerical libraries for extremely fast dense matrix multiplication

Motivation & High Level Considerations

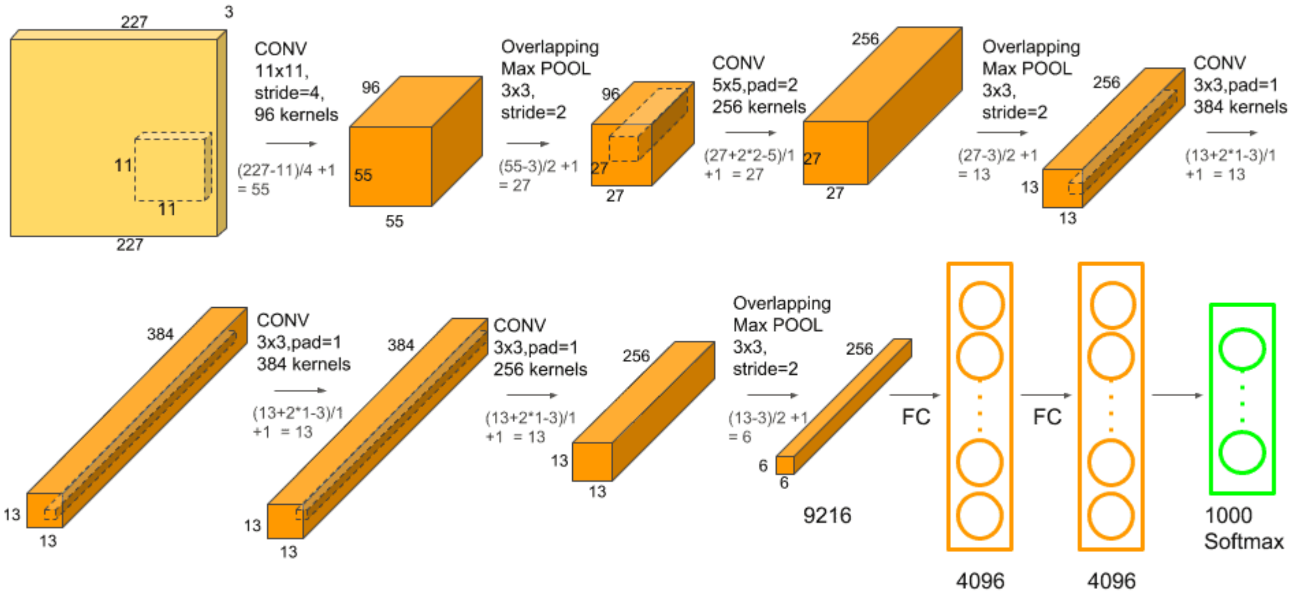
- Recap) LeNet-5 vs. AlexNet





Motivation & High Level Considerations

- Recap) LeNet-5 vs. AlexNet

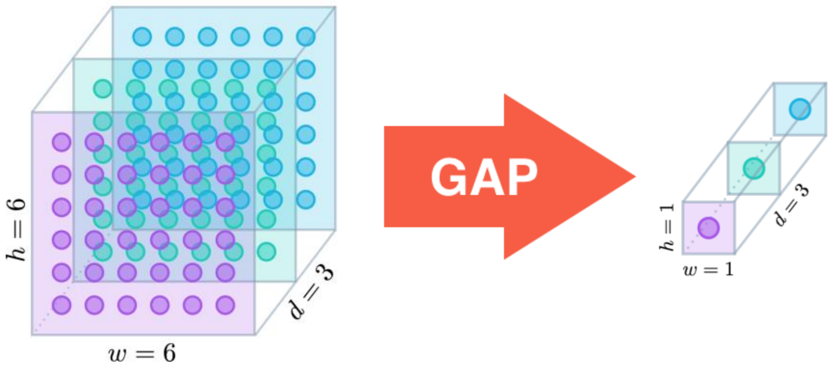


Motivation & High Level Considerations

- Provable Bounds for Learning Some Deep Representations (Arora et al., ICML 2013)
  - Sparse Matrix -> Dense Sub-Matrix
  - Represent with sparse deep NN -> Large probability distribution of dataset => optimal network (from mimicking biological systems)
    - Analyzing correlation statistics of the activations of the last layer
    - Clustering neurons with highly correlated outputs
      - Hebbian Principle: *neurons that fire together, wire together*

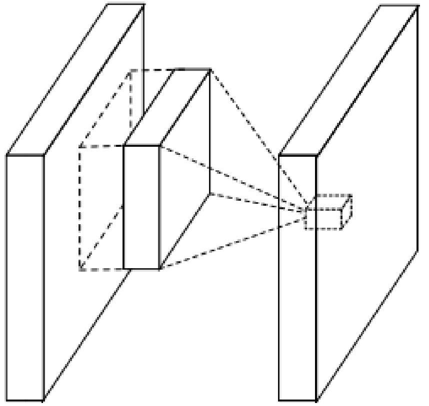
Motivation & High Level Considerations

- Network in Network (Lin et al., ICLR 2014)
  - Traditional Conv: generalized (for data path) linear model (GLM)
  - MLPConv: non-linearity
  - Global Average Pooling (GAP): due to enough features from MLPConv, prevents overfitting
  - 1x1 conv

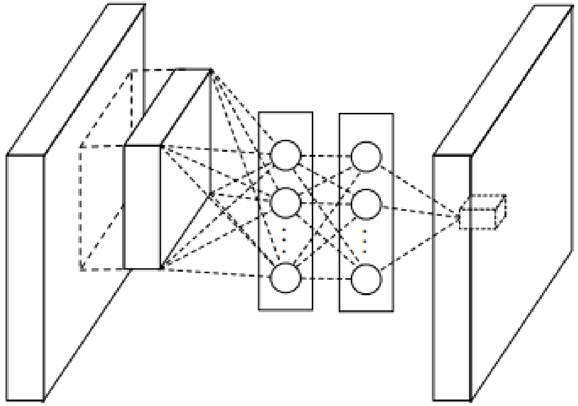


Motivation & High Level Considerations

- Network in Network (Lin et al., ICLR 2014)
  - MLPConv



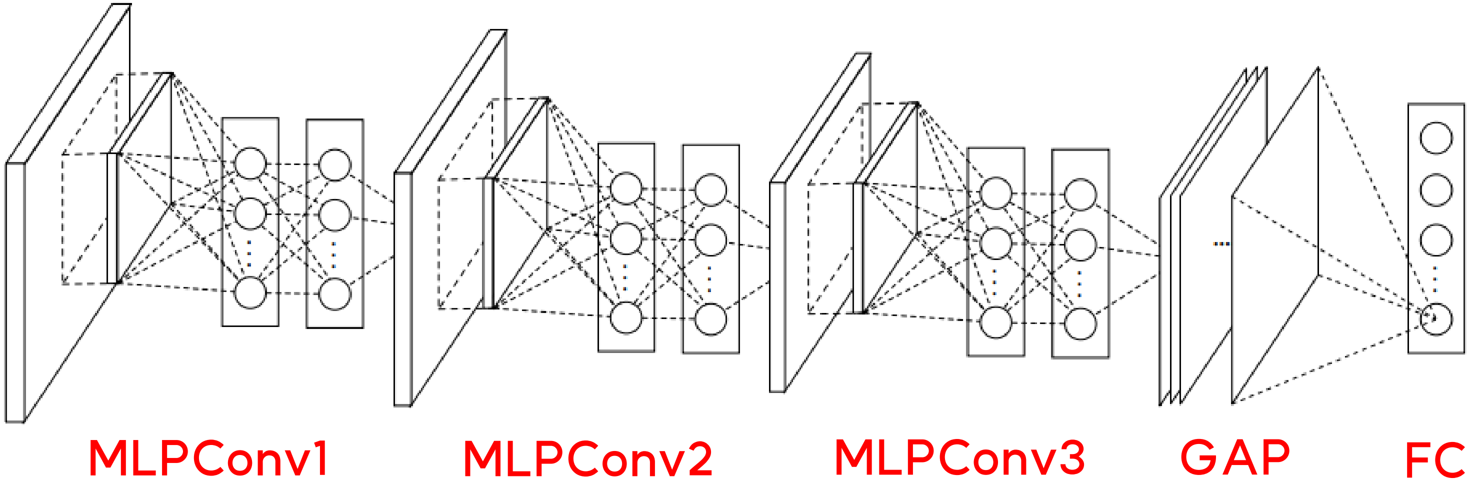
(a) Linear convolution layer



(b) Mlpconv layer

Motivation & High Level Considerations

- Network in Network (Lin et al., ICLR 2014)
  - NiN architecture

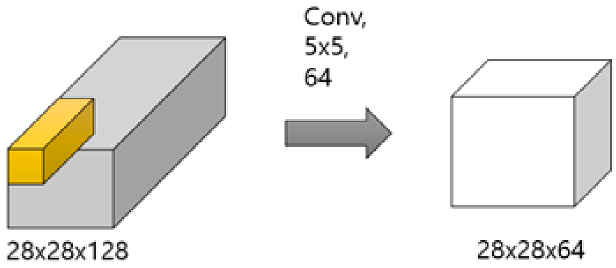


Motivation & High Level Considerations

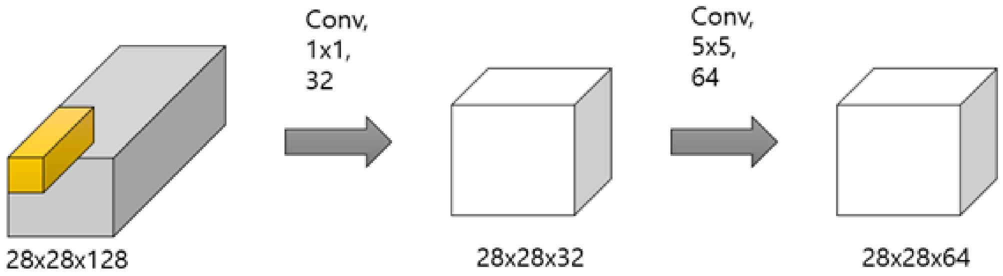
- 1x1 Convolution
  - Convolution Layer with 1x1 Filter Size
  - Number of channel (hyperparameter) adjust
  - Reduce parameters (efficient)
  - Non-linearity -> more expressive
  - More deeper network

Motivation & High Level Considerations

- **1x1 Convolution**



#params = 28x28x64x5x5x128 = 160M



#params = 28x28x32x128x1x1 = 4.8M

#params = 28x28x64x5x5x32 = 40M

#total = 44.8M

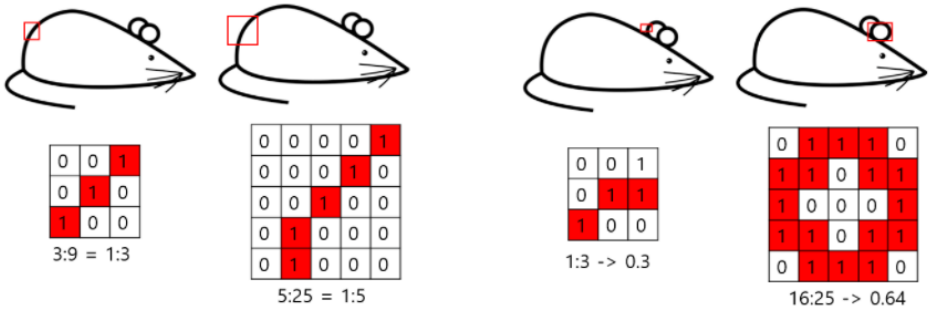
Architectural Details

- Main Idea
  - Approximate to optimal local sparse structure -> dense component
    - Clustering sparse matrix -> dense sub-matrix
  - Each unit of previous layer -> region of input image (assume)
    - Lower layer (near to input layer) -> correlated units are concentrated in specific region

Architectural Details

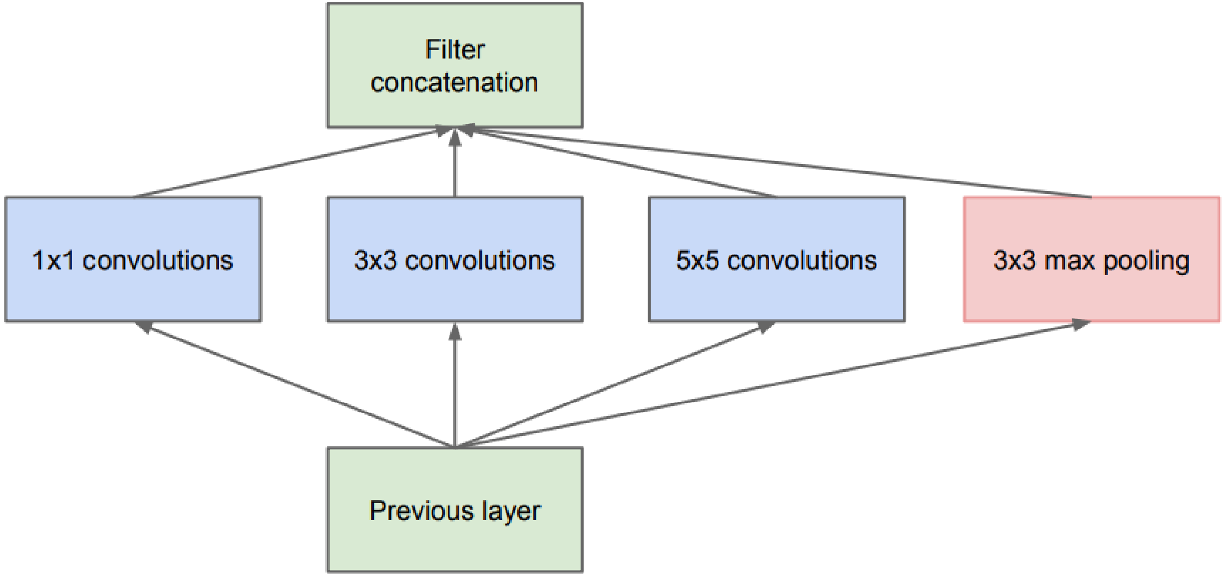
- **Main Idea**
  - **Correlation in images**
    - Color, texture, … -> Local Features
      - can be covered by 1x1 convolution

  - **Filter size: related to correlated unit rate**
    - Various feature maps
      - 1x1, 3x3, 5x5 convolution



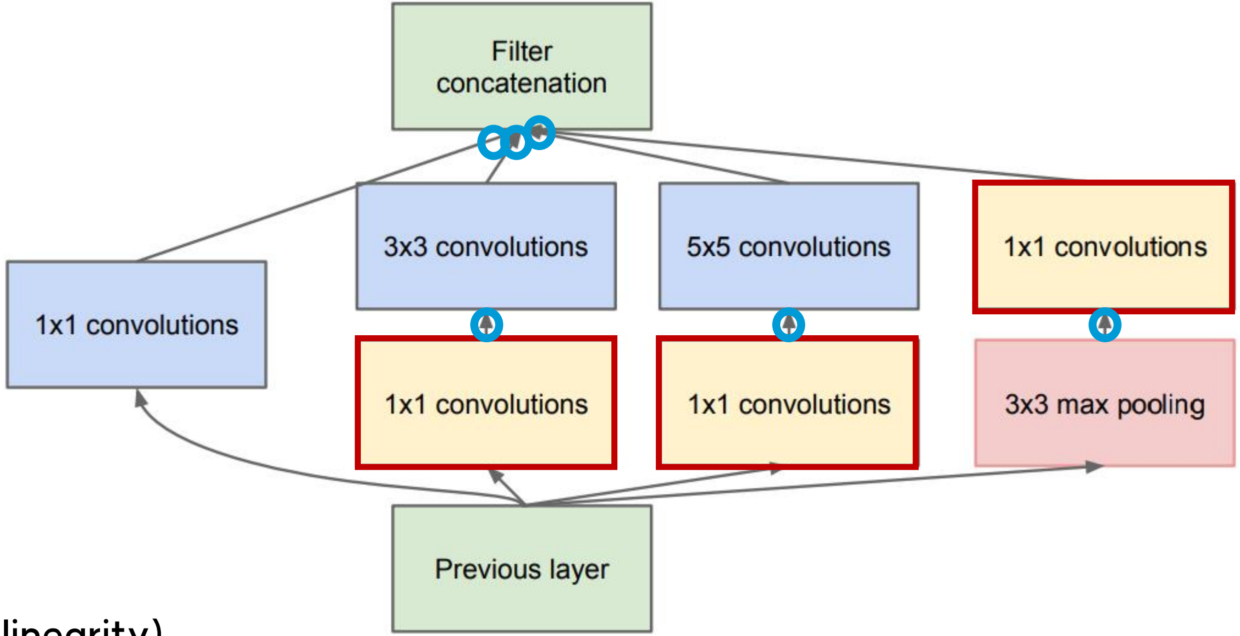
Architectural Details

- Inception module, naive version: expensive



Architectural Details

- Inception module, with dimension reductions (1x1 conv)



ReLU (non-linearity)

1x1 conv (dimension reduction)

Architectural Details

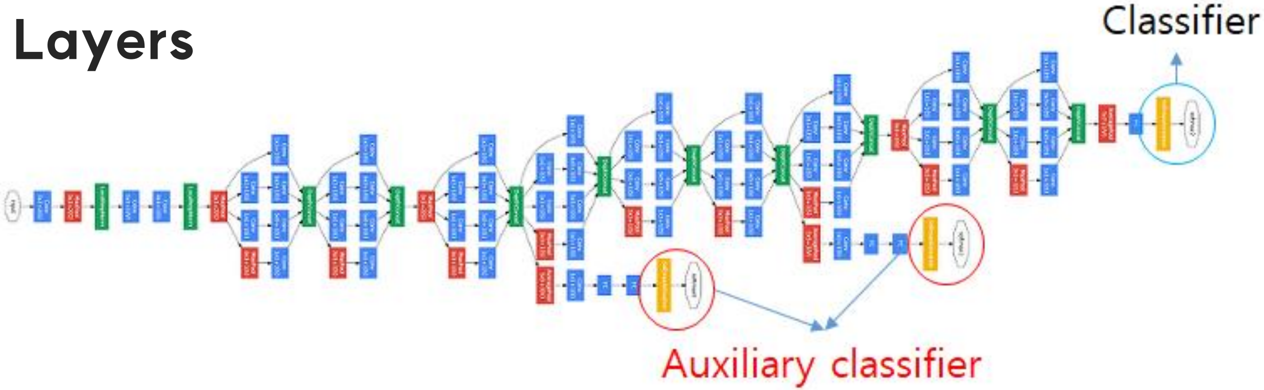
- Dimension Reduction
  - Capacity -> Reduce
    - Each filter have highly correlation

      – doesn't matter

- Advantages in Inception Module
  - Can increase units for each stage without computational complexity
  - Various feature: 1x1, 3x3, 5x5 convolution

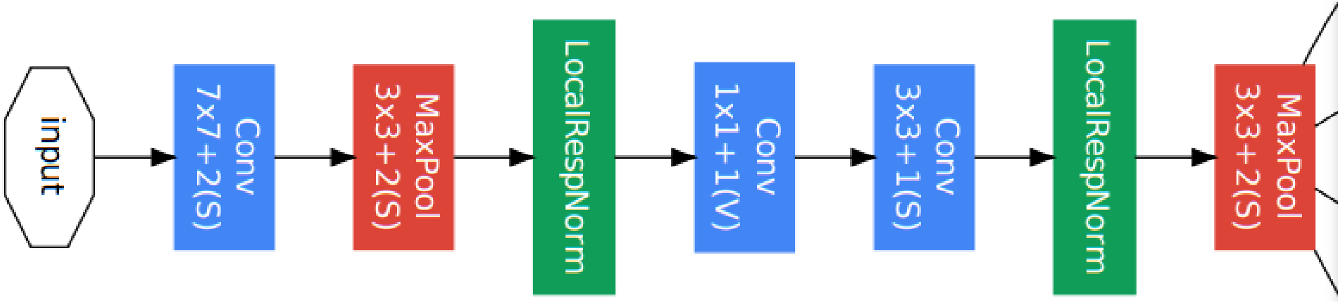
GoogLeNet

- Mean subtraction (AlexNet)
- ReLU: after all the convolutions
- Receptive field: 224x224 RGB color channels
- 22 Layers



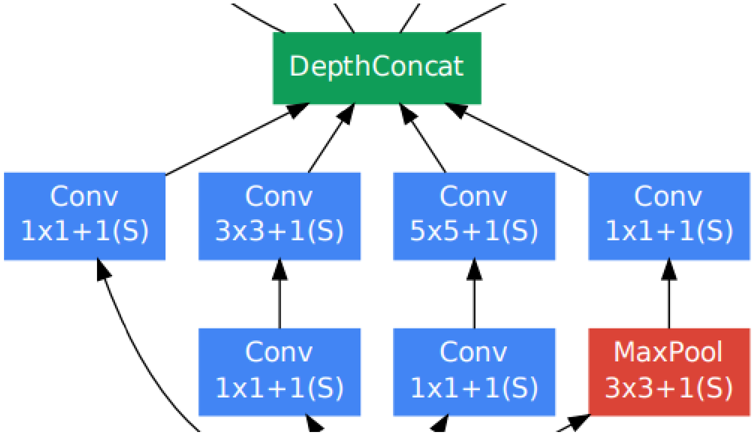
GoogLeNet

- 1. Lower Layers (near to input)
  - Simple Conv
    - For memory efficiency



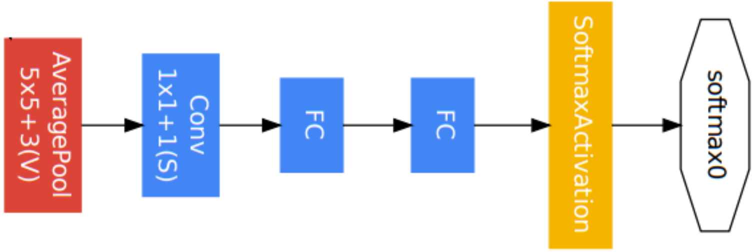
GoogLeNet

- 2. Inception Module
  - Local unit with parallel branches
  - 1x1 “Bottleneck” layers (to reduce channel dimension)



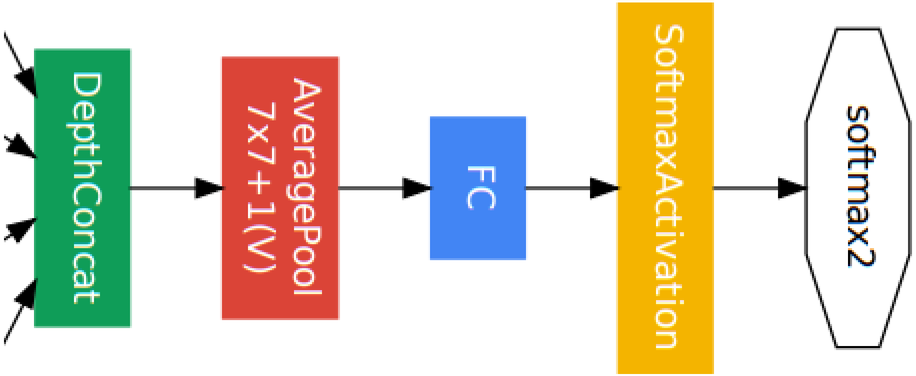
GoogLeNet

- 3. Auxiliary Classifier: Train Only
  - Deeper network: gradient vanishing problem
    - Encourage discrimination in the lower stages in the classifier
      - increase backpropagate gradient signal
      - provide additional regularization
  - Add to total loss with weighted by 0.3 (0.3 auxiliary + 0.7 inception block output) -> prevent affecting to weights



GoogLeNet

- 4. Global Average Pooling
  - No additional parameters (just pooling)
    - Adapted before last classifier (fully-connected layer)



GoogLeNet

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Results

- ILSVRC 2014 Classification Challenge
  - Trained 7 versions of the same GoogLeNet -> ensemble
    - Same initialization, learning rate

      – difference: sampling strategy
  - Aggressive cropping approach -> 256, 228, 320, 352 resize
    - Take the left, center, and right square

      – 4 corners and the center 224x224 crop + 224x224 resize (with horizontal flip)

      – 4 * 3 * 6 * 2 = 144 per image
  - Average over multiple crops (softmax probablities)

Results

- ILSVRC 2014 Classification Challenge

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

Results

- ILSVRC 2014 Classification Challenge

Number of models	Number of Crops	Cost	Top-5 error	compared to base
1	1	1	10.07%	base
1	10	10	9.15%	-0.92%
1	144	144	7.89%	-2.18%
7	1	7	8.09%	-1.98%
7	10	70	7.62%	-2.45%
7 **Final Submission**	144	1008	6.67%	-3.45%

Conclusions

- Evidence for improving neural networks for computer vision
  - Approximate optimal sparse structure -> dense structure
- A little increasement of computational cost, significant quality gain
  - Compared to shallower, less wide networks
- Expect result of similar quality with similar depth and width networks
  - GoogLeNet -> more efficient

Discussions

- 1x1 convolution
  - Channel reduction effect -> feature compression
    - Data loss
- Auxiliary classifier
    - More convolution is better ? E.g) add 7x7 conv

GNU

경상국립대학교

Gyeongsang National University

Improving lives through learning

IDEALAB