COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

Very Deep Convolutional Networks For Large-Scale Image Recognition, ICLR2015

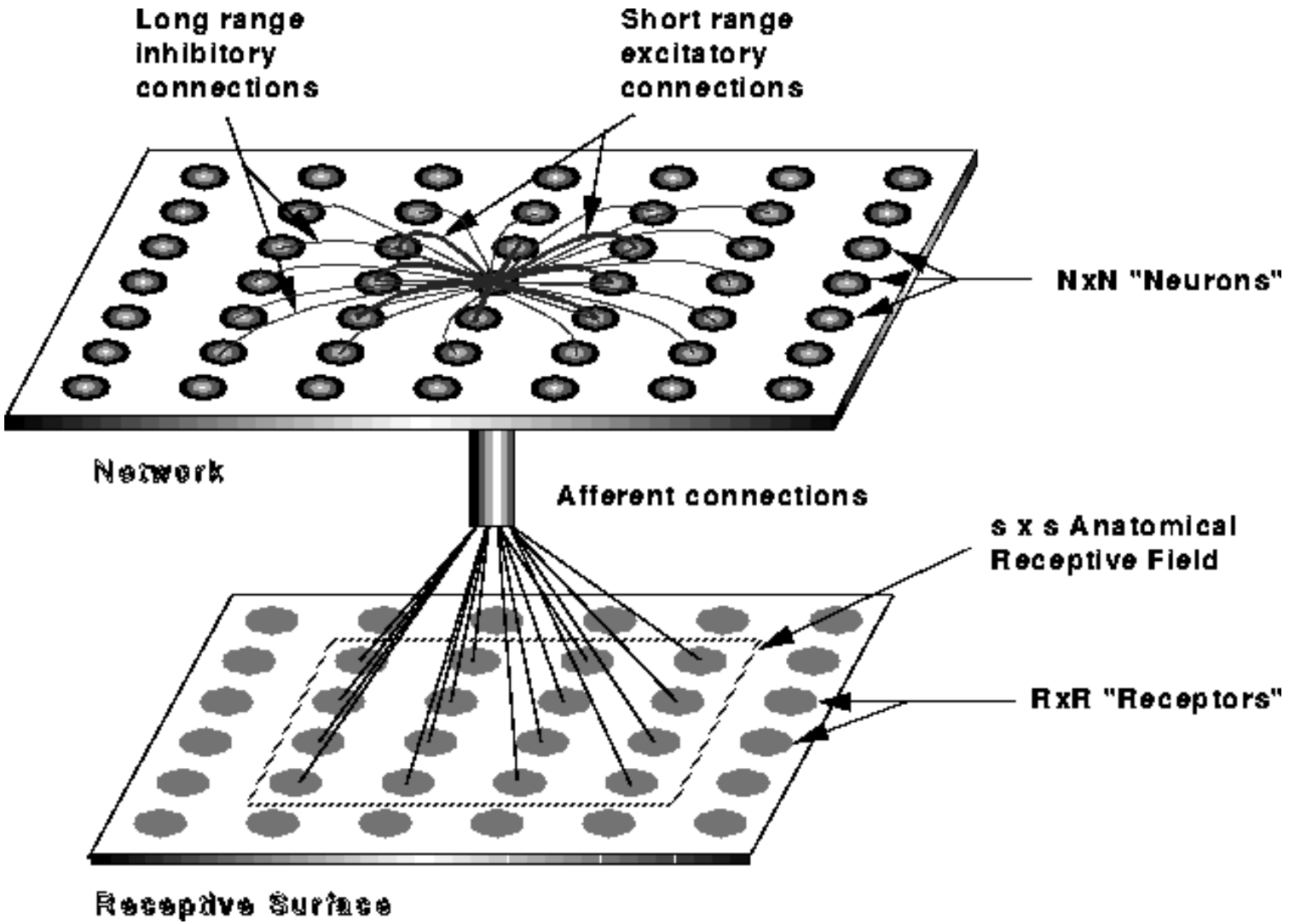
COMPUTER VISION LAB

22211226 JHNam

CONTENT

- Introduction
- ConvNet Configurations
- Classification Framework
- Classification Results

- **Receptive Window Size**



- **Architecture**
  - Input size : $224 \times 224 \times 3$
  - Preprocess : Substracting the mean RGB value from each pixel.
  - Filter size : 3×3 → Smallest size to capture the notion of left/right, up/down, center
  - Stride : 1 pixel
  - Spatial Padding : The spatial resolution is preserved after convolution : 1 pixel

$$n_{out} = \left[\frac{n_{in} + 2p - k}{s}\right] + 1$$

    - n_{in} : #input features
    - n_{out} : #output features
    - k : convolution kernel size
    - p : convolution padding size
    - s : convolution stride size

$$n_{in} = \left[\frac{n_{in} + 2p - 3}{1}\right] + 1 = n_{in} + 2p - 2 \rightarrow 2p = 2 \rightarrow p = 1$$

  - Spatial Pooling : Five max-pooling layer : 2×2 window with stride 2
  - 3 Fully Connect Layer : $4096 \rightarrow 4096 \rightarrow 1000$ = #class
  - Activation function : ReLU

- **Configurations**

Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channels⟩". The ReLU activation function is not shown for brevity.

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					

- **Discussion**
  - 1. Receptive field : Stack of two 3×3 conv layer > One 5×5 conv layer
    - Stack of two 3×3 conv layer = One 7×7 conv layer > One 5×5 conv layer
  - 2. #Parameters → More Deeper!!
    - Three 3×3 conv layer has C channels : $27C^2$
    - One 7×7 conv layer has C channels : $49C^2$

- **Training**
  - Experiment Setting
    - Batch Size : 256
    - Optimizer
      - Stochastic Gradient Descent with momentum 0.9
      - weight decay : 5×10^{-4}
    - DropOut : $p = 0.5$
    - Learning rate
      - 0.01 → Decreased by a factor of 10 when the validation set accuracy stopped improving
  - Augmentation : Horizontal Flip & Random RGB Color Shift

- **Training**
  - Training Image Size
    - S : Smallest side of an isotropically-rescaled training image.
    - Two fixed scale S
      - 1. Single-scale training
        $$S = 256 \rightarrow S = 384$$
      - 2. Multi-scale training

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

- **Dataset**
  - ILSVRC-2012 Dataset

COMPUTER VISION LAB.,
School of Artificial Intelligence,
Inha University

- **Single Scale Evaluation**
  - Test Image Size
    - $Q = S$
    - $Q = 0.5(S_{\min} + S_{\max}), S \in [S_{\min},\ S_{\max}]$

- **Single Scale Evaluation**
  - Analysis
    - 1. Deeper Model → Lower top1 & top5 error
    - 2. Same Depth : C(1×1 Conv) < D(3×3 Conv) | Receptive field : C < D → Spatial Information ↑
    - 3. Additional Nonlinearity 1×1 Conv : C > B
    - 4. B(3×3 Conv) > B(5×5 Conv)

Table 3: **ConvNet performance at a single test scale.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

- **Multi Scale Evaluation**
  - Analysis
    - 1. Single Scale < Multi Scale

Table 4: **ConvNet performance at multiple test scales.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256; 512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**