CSCI567 Machine Learning (Spring 2018)

Michael Shindler

Lecture 9: February 7 2018

Outline

1. Administration
2. Review of last lecture
3. Finishing last lecture
4. Convolution neural networks

Outline

1. Administration
2. Review of last lecture
3. Finishing last lecture
4. Convolution neural networks

Administrative Stuff

- Homework due soon
  - Problem Set and Programming are different
  - Grace Days are allowed *for programming only*
  - When submitting programming, **do not force push**.
- Homework 2 release soon
  - If you force push programming one, you risk losing homework 2
  - If you do, let us know, we will replace it (small penalty)
  - *Do not ask someone else for theirs*
    - Different people's start might be slightly different

Outline

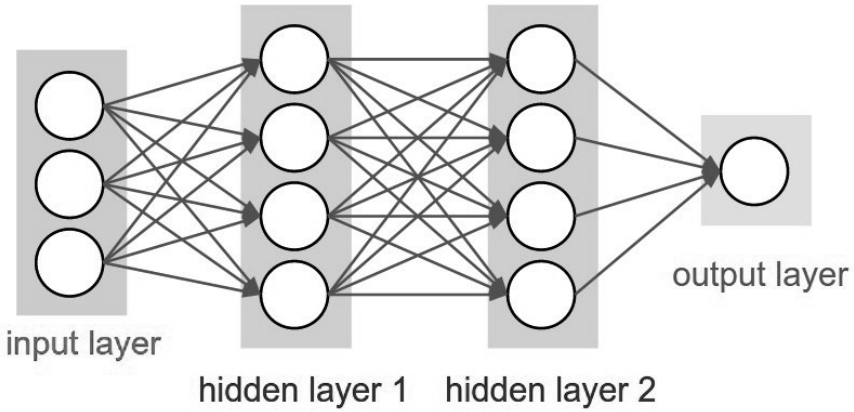
1 Administration

2 Review of last lecture

3 Finishing last lecture

4 Convolution neural networks

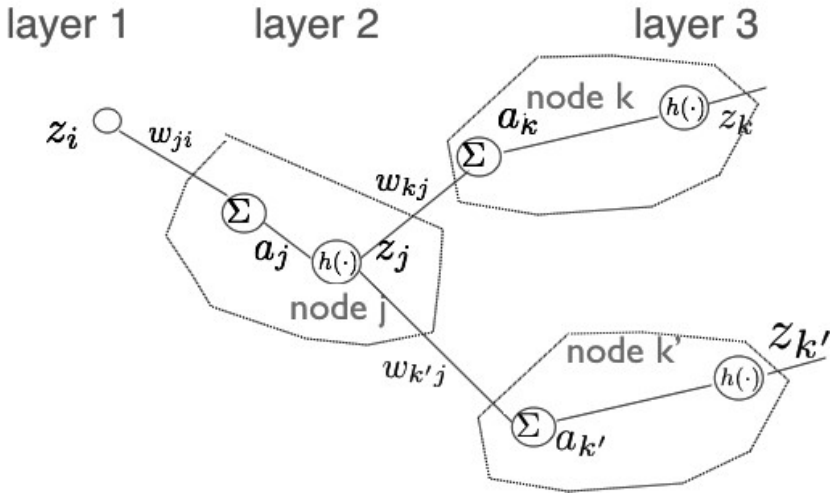
Neural nets



Key steps (essentially, chain rule in calculus)

To compute

$$\frac{\partial \ell}{\partial w_{ji}} = \frac{\partial \ell}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = z_i \frac{\partial \ell}{\partial a_j}$$



as w_{ji} affects only a_j

Passthru the nonlinear transfer function

$$\frac{\partial \ell}{\partial a_j} = \frac{\partial \ell}{\partial z_j}\frac{\partial z_j}{\partial a_j} = h'(a_j)\frac{\partial \ell}{\partial z_j}$$

Collect errors from next layer

$$\frac{\partial \ell}{\partial z_j} = \sum_k \frac{\partial \ell}{\partial a_k}\frac{\partial a_k}{\partial z_j} = \sum_k \frac{\partial \ell}{\partial a_k} w_{kj}$$

Recursion

$$\frac{\partial \ell}{\partial a_j} = h'(a_j)\sum_k \frac{\partial \ell}{\partial a_k} w_{kj}$$

Outline

1 Administration

2 Review of last lecture

3 Finishing last lecture
- Regularization for Neural Nets
- Optimization for Neural Networks
- Summary

4 Convolution neural networks

Overfitting is very possible

Methods for overcoming overfitting

- Regularization
- Early stopping
- Data augmentation
- Inject noise

Regularization: weight decay in neural nets

Original loss function, eg:

$$\ell(\boldsymbol{w}) = -\sum_n \sum_k y_{nk} \log p(y = k|x_n)$$

Adding L_2 norm to regularize

$$\ell'(\boldsymbol{w}) = \ell(\boldsymbol{w}) + \frac{1}{2}\lambda \|\boldsymbol{w}\|_2^2$$

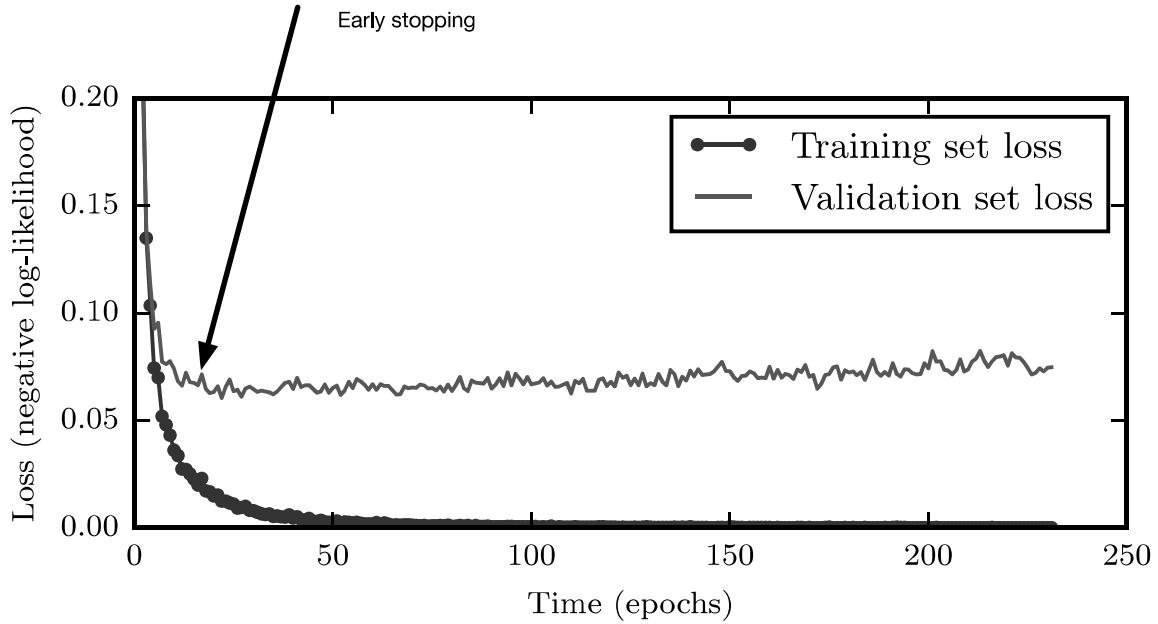
Gradient changed to

$$\frac{\partial \ell'(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{\partial \ell(\boldsymbol{w})}{\partial \boldsymbol{w}} + \lambda \boldsymbol{w}$$

Effect: (slowly) decaying $\boldsymbol{w}$ to zero.

Early stopping

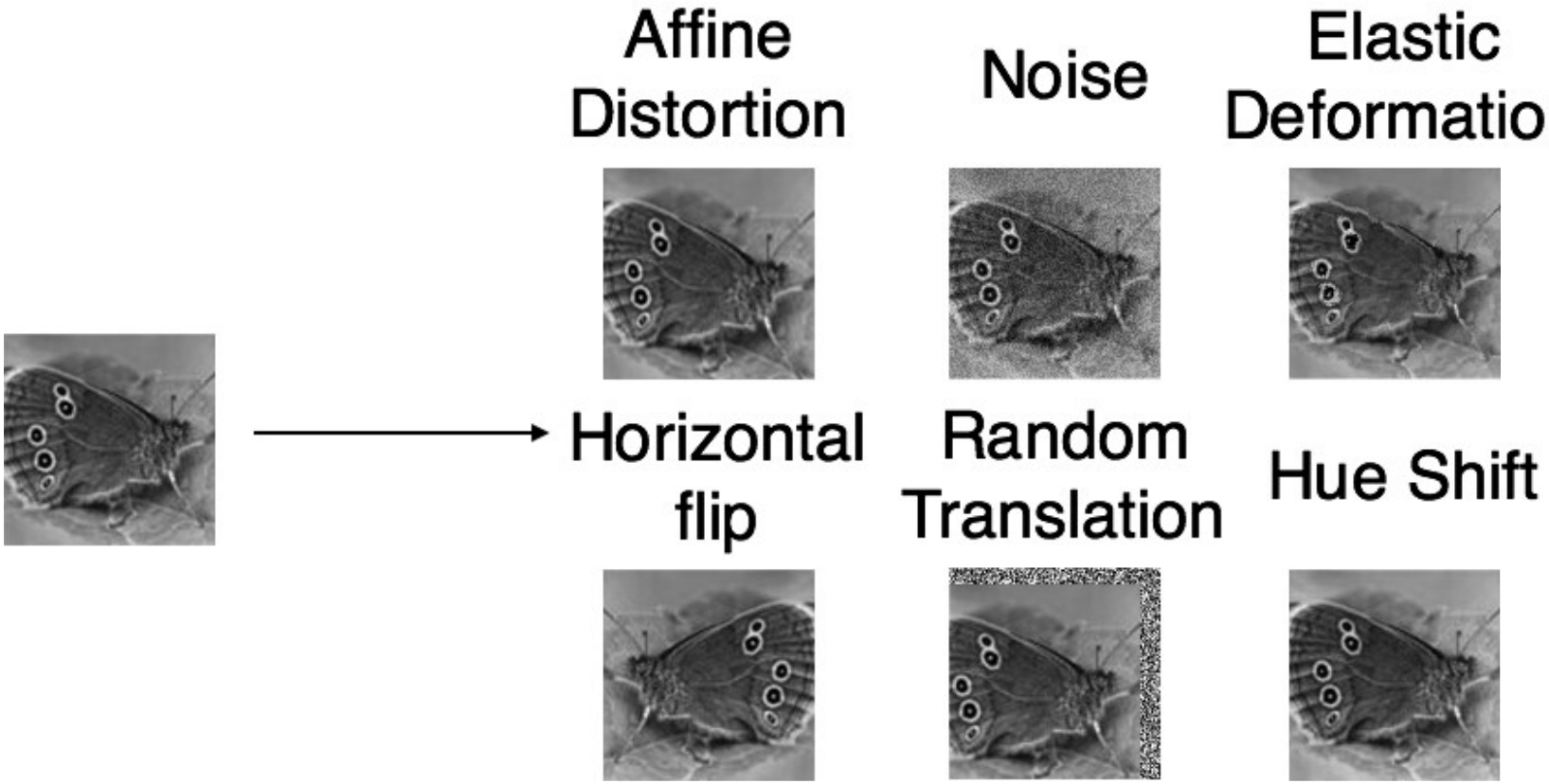
Stopping training before the training error is reduced too much



Thus, we should *monitor* the change of performance on the validation set every iteration.

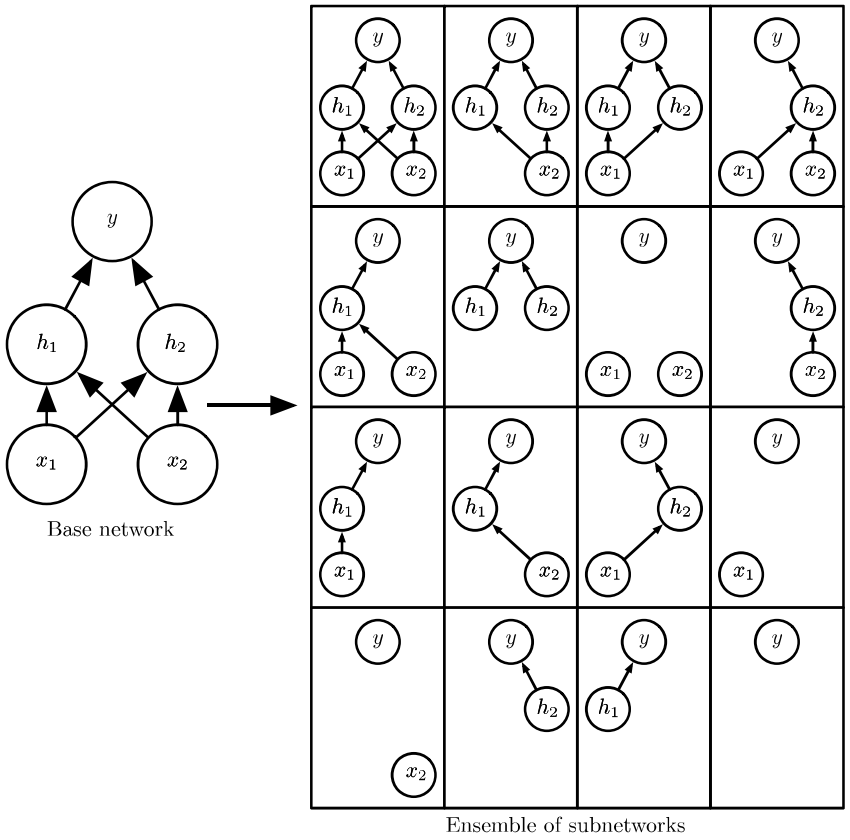
Data Augmentation

Exploiting prior knowledge to add more labeled data



Injecting noise

During training, randomly delete nodes (this is called *dropout*)



Optimization in neural networks is challenging

Common approaches

- Better initialization
- Stochastic gradient descent with minibatch
- Stochastic gradient descent with momentum
- Adaptive learning rate

Stochastic Gradient Descent with minibatch

Recipe

- Decide a minibatch size m
- Initialize $\boldsymbol{w}$ to $\boldsymbol{w}^{(0)}$; set $t = 0$; choose $\eta > 0$
- Loop *until convergence*
  1. random choose m training samples $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_m$
  2. Compute their contribution to the gradient

$$\boldsymbol{g} = \frac{1}{m} \sum_i \frac{\partial \ell(\boldsymbol{x}_m, y_m)}{\partial \boldsymbol{w}}$$

  3. Update the parameters $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \boldsymbol{g}$
  4. $t \leftarrow t + 1$

How to choose m?

Stochastic Gradient Descent with Momentum

Intuition

Stochastic gradient descent is "greedy" in using the current minibatch of samples.

Can we use the past gradient to help us? Namely how the parameters move in the past (the velocity) could be useful to us.

Recipe

- Decide a minibatch size m, a momentum strength of $\alpha \in (0, 1)$
- Initialize $\boldsymbol{w}$ to $\boldsymbol{w}^{(0)}$, initialize the velocity vector $\boldsymbol{v}$ (say, 0); set $t = 0$; choose $\eta > 0$
- Loop *until convergence*
  1. random choose m training samples $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_m$
  2. Compute their contribution to the gradient

$$\boldsymbol{g} = \frac{1}{m} \sum_i \frac{\partial \ell(\boldsymbol{x}_m, y_m)}{\partial \boldsymbol{w}}$$

  3. Update the velocity

$$\boldsymbol{v} \leftarrow \alpha \boldsymbol{v} - \eta \boldsymbol{g}$$

  4. Update the parameters $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \boldsymbol{v}$
  5. $t \leftarrow t + 1$

How history of update can affect present

Assume initial velocity is 0

time	velocity $\boldsymbol{v}$	how much parameters are to be updated
0	$\mathbf{0}$	$-\eta \boldsymbol{g}^1$
1	$-\eta \boldsymbol{g}^1$	$-\alpha\eta \boldsymbol{g}^1 - \eta \boldsymbol{g}^2$
2	$-\alpha\eta \boldsymbol{g}^1 - \eta \boldsymbol{g}^2$	$-\alpha^2\eta \boldsymbol{g}^1 - \alpha\eta \boldsymbol{g}^2 - \eta \boldsymbol{g}^3$

$\boldsymbol{g}^1$ stands for the gradient computed on the first minibatch and $\boldsymbol{g}^2$ stands for the gradient computed on the second batch, so on and so forth.

Challenge At time t, what is the parameter vector (if the initial parameter vector is $\boldsymbol{w}^0$) – expressed in all the $\boldsymbol{g}^t$s?

Summary

- Deep neural networks are hugely popular.
- They achieve stellar performance on many problems.
- They do need a lot of data points to work well.
- They do have optimization headaches to overcome: local optima, convergence, hyperparameters (minibatch size, step size, momentum, etc)
- They do tend to require heavy computation: use a GPU for massive parallel computing
- Selecting good models takes time: how many hidden layers, how many units/neurons each layer etc

Good reference: Goodfellow, Bengio and Courville's textbook *Deep Learning* `www.deeplearningbook.org`

Acknowledgements

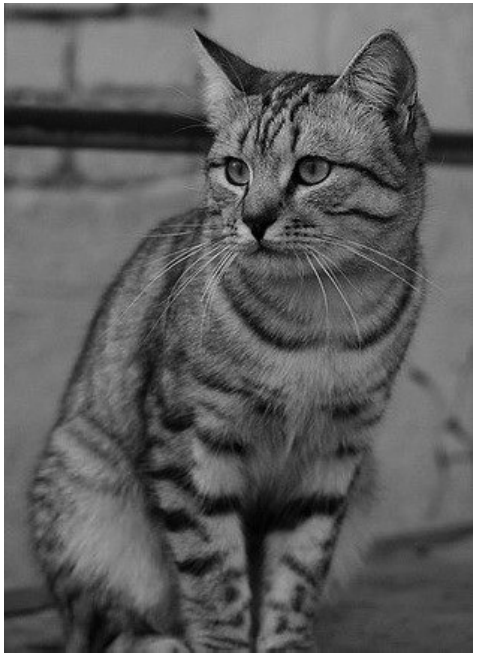
The lecture slides borrow *heavily* from the following sources:

- Stanford Course Cs231n: Convolutional Neural Networks for Visual Recognition Spring 2017 (taught by Prof. Fei-Fei Li and her students).
  The website `http://cs231n.stanford.edu/` provides also very valuable notes, demo codes, and pointers to useful information
- Dr. Ian Goodfellow's lectures on neural networks and deep learning.
  The website`http://deeplearningbook.org` provides very valuable lecture notes, pointers to video-tapped lectures and other useful information

Outline

1. Administration
2. Review of last lecture
3. Finishing last lecture
4. Convolution neural networks
   - Motivation
   - Architecture

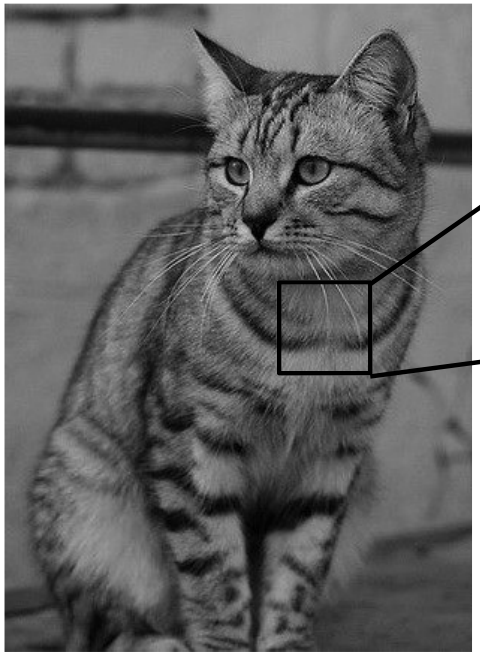
Image Classification: A core task in Computer Vision



(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

cat

The Problem: Semantic Gap



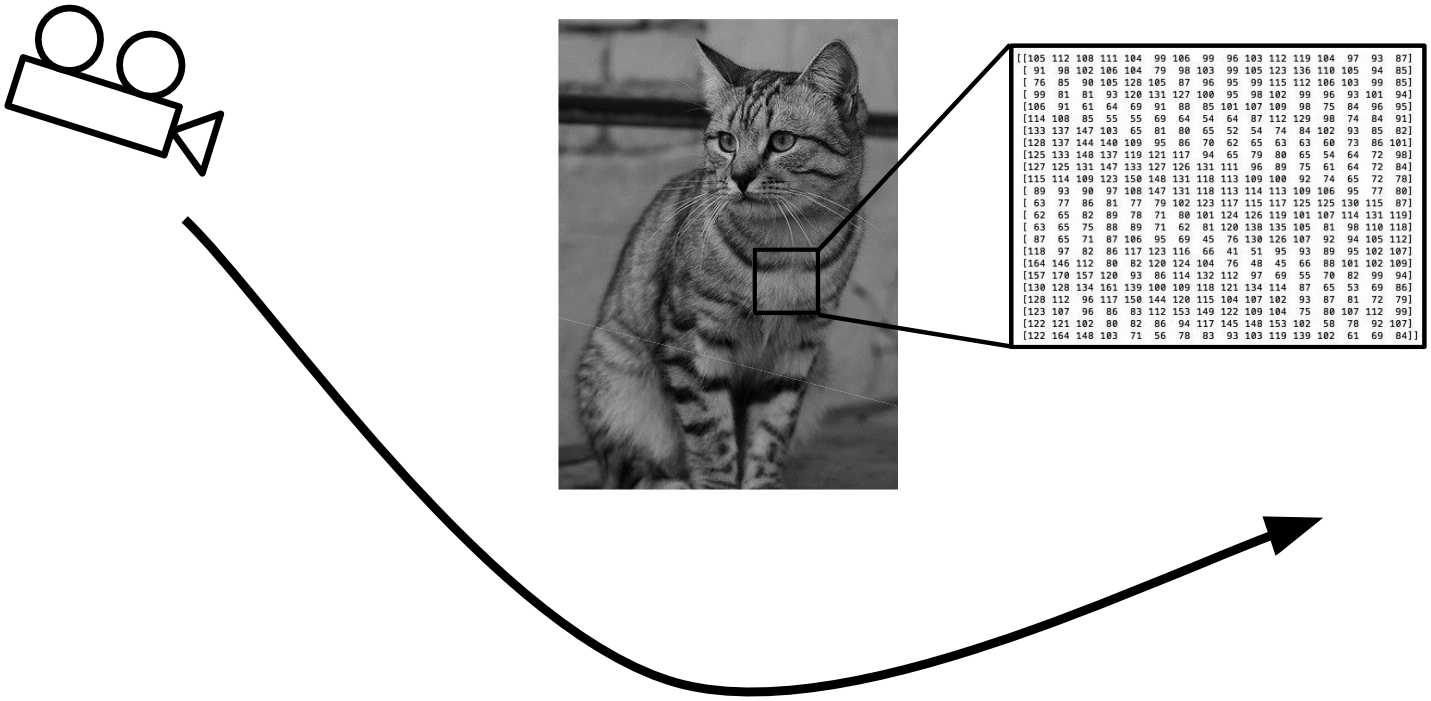

```
[[105 112 108 111 104  99 106  99  96 103 112 119 104  97  93  87]
 [ 91  98 102 106 104  79  98 103  99 105 123 136 110 105  94  85]
 [ 76  85  90 105 128 105  87  96  95  99 115 112 106 103  99  85]
 [ 99  81  81  93 120 131 127 100  95  98 102  99  96  93 101  94]
 [106  91  61  64  69  91  88  85 101 107 109  98  75  84  96  95]
 [114 108  85  55  55  69  64  54  64  87 112 129  98  74  84  91]
 [133 137 147 103  65  81  80  65  52  54  74  84 102  93  85  82]
 [128 137 144 140 109  95  86  70  62  65  63  63  60  73  86 101]
 [125 133 148 137 119 121 117  94  65  79  80  65  54  64  72  98]
 [127 125 131 147 133 127 126 131 111  96  89  75  61  64  72  84]
 [115 114 109 123 150 148 131 118 113 109 100  92  74  65  72  78]
 [ 89  93  90  97 108 147 131 118 113 114 113 109 106  95  77  80]
 [ 63  77  86  81  77  79 102 123 117 115 117 125 125 130 115  87]
 [ 62  65  82  89  78  71  80 101 124 126 119 101 107 114 131 119]
 [ 63  65  75  88  89  71  62  81 120 138 135 105  81  98 110 118]
 [ 87  65  71  87 106  95  69  45  76 130 126 107  92  94 105 112]
 [118  97  82  86 117 123 116  66  41  51  95  93  89  95 102 107]
 [164 146 112  80  82 120 124 104  76  48  45  66  88 101 102 109]
 [157 170 157 120  93  86 114 132 112  97  69  55  70  82  99  94]
 [130 128 134 161 139 100 109 118 121 134 114  87  65  53  69  86]
 [128 112  96 117 150 144 120 115 104 107 102  93  87  81  72  79]
 [123 107  96  86  83 112 153 149 122 109 104  75  80 107 112  99]
 [122 121 102  80  82  86  94 117 145 148 153 102  58  78  92 107]
 [122 164 148 103  71  56  78  83  93 103 119 139 102  61  69  84]]
```

What the computer sees

An image is just a big grid of numbers between [0, 255]:

e.g. 800 x 600 x 3
(3 channels RGB)

Challenges: Viewpoint variation



All pixels change when the camera moves!

Challenges: Illumination

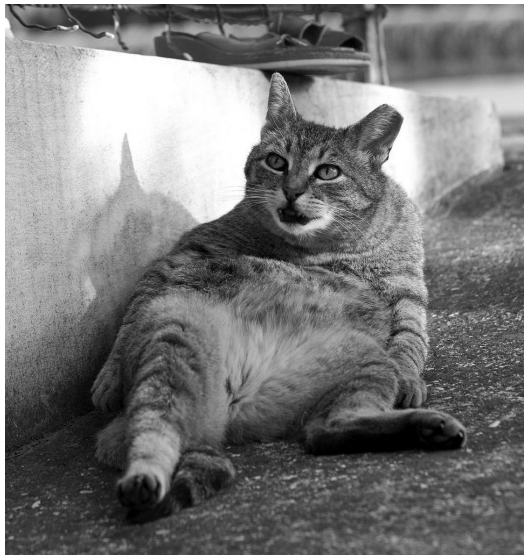
This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

Challenges: Deformation



This image by Umberto Salvagnin is licensed under CC-BY 2.0

This image by Umberto Salvagnin is licensed under CC-BY 2.0

This image by sare bear is licensed under CC-BY 2.0

This image by Tom Thai is licensed under CC-BY 2.0

Challenges: Occlusion

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

This image by jonsson is licensed under CC-BY 2.0

Challenges: Background Clutter

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

Challenges: Intraclass variation

This image is CC0 1.0 public domain

Fundamental problems in vision and learning

The key challenge

How to train a visual recognition model to be resilient to all those variations?

In terms of our nonlinear basis functions, those functions must be able to map the images to an invariant representation (with respect to pose, illumination, etc). But *how to get such representations?*

Main ideas

- We will need a lot of data that exhibits those variations – otherwise the models cannot automatically conjure up invariances.
- We will need powerful models to capture the invariances.

A bit of history:

Hubel & Wiesel,

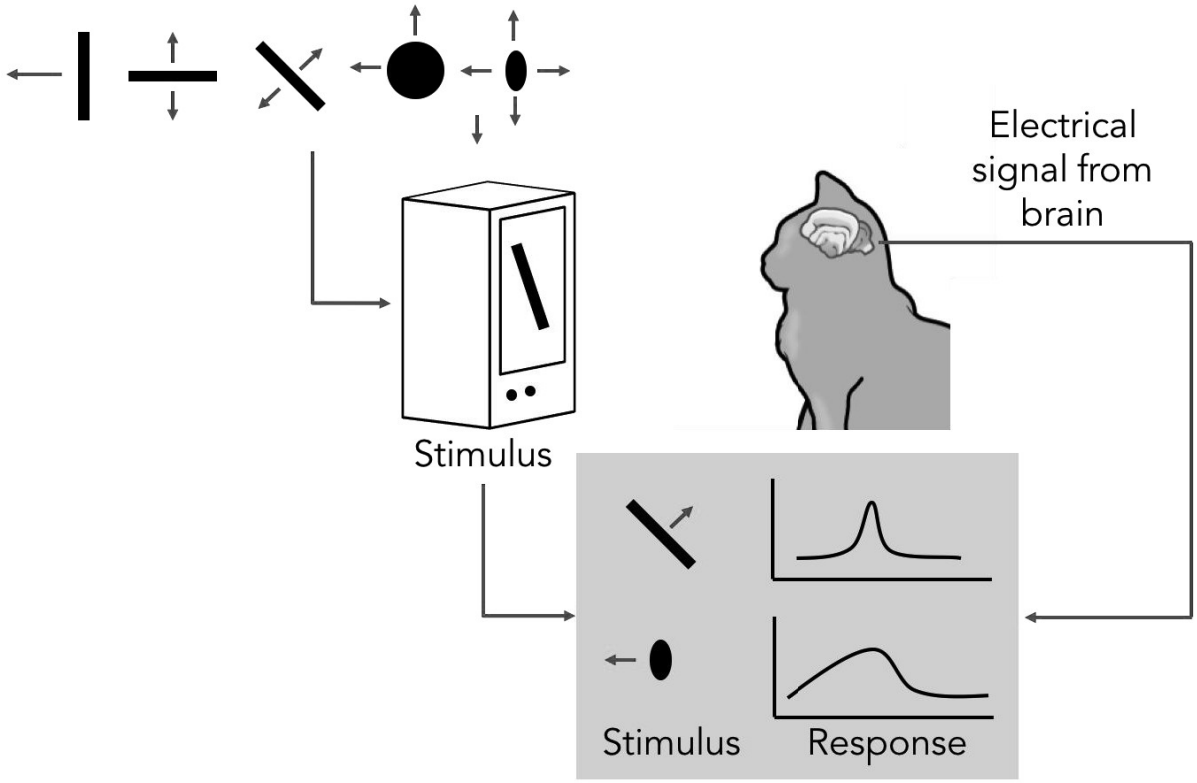
1959

RECEPTIVE FIELDS OF SINGLE NEURONES IN
THE CAT'S STRIATE CORTEX

1962

RECEPTIVE FIELDS, BINOCULAR INTERACTION
AND FUNCTIONAL ARCHITECTURE IN
THE CAT'S VISUAL CORTEX

1968...



Hierarchical organization

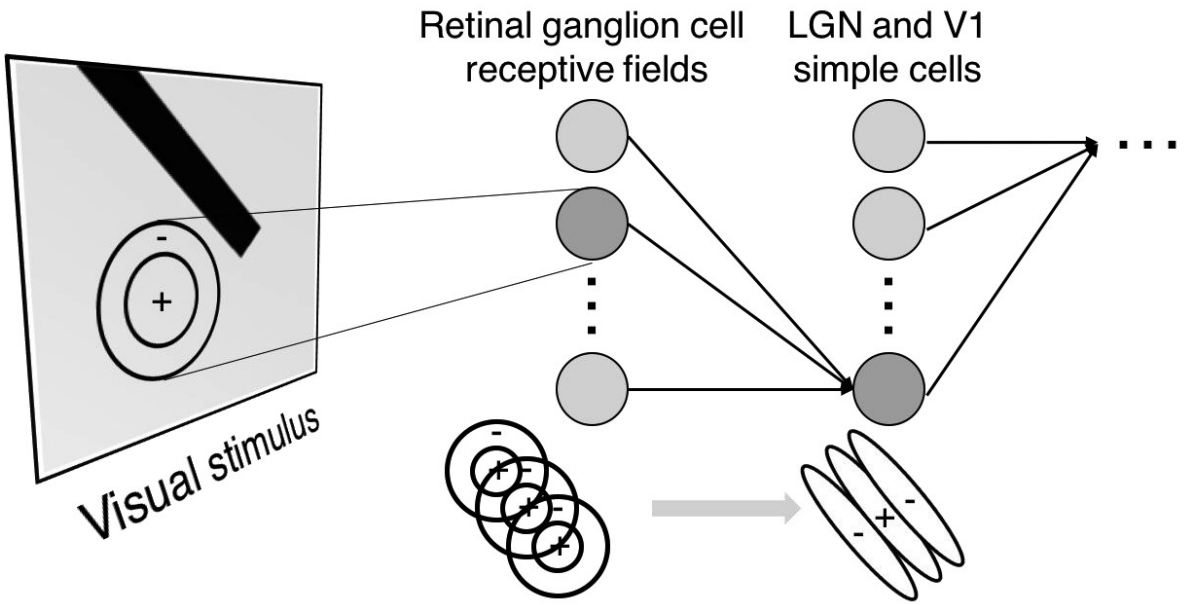
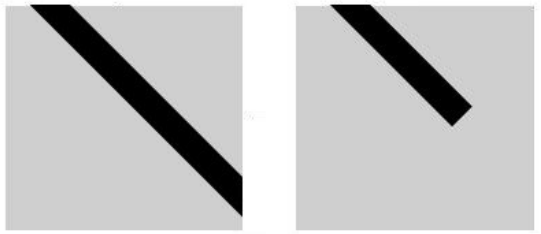


Illustration of hierarchical organization in early visual pathways by Lane McIntosh, copyright CS231n 2017

Simple cells:
Response to light orientation

Complex cells:
Response to light orientation and movement

Hypercomplex cells:
response to movement with an end point



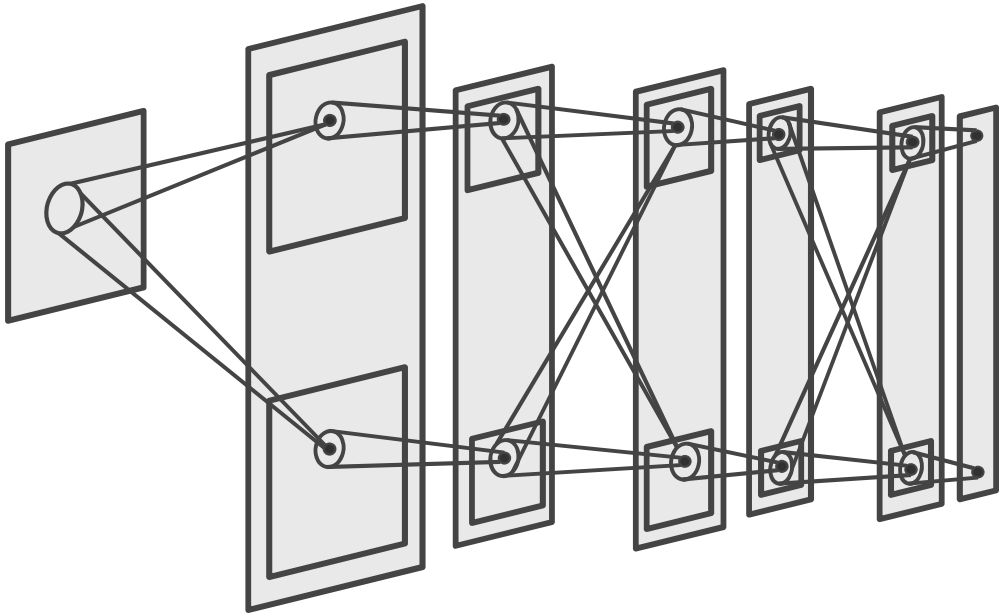
No response

Response (end point)

An attempt to mimic the biological pipeline

A bit of history:

Neocognitron
[Fukushima 1980]

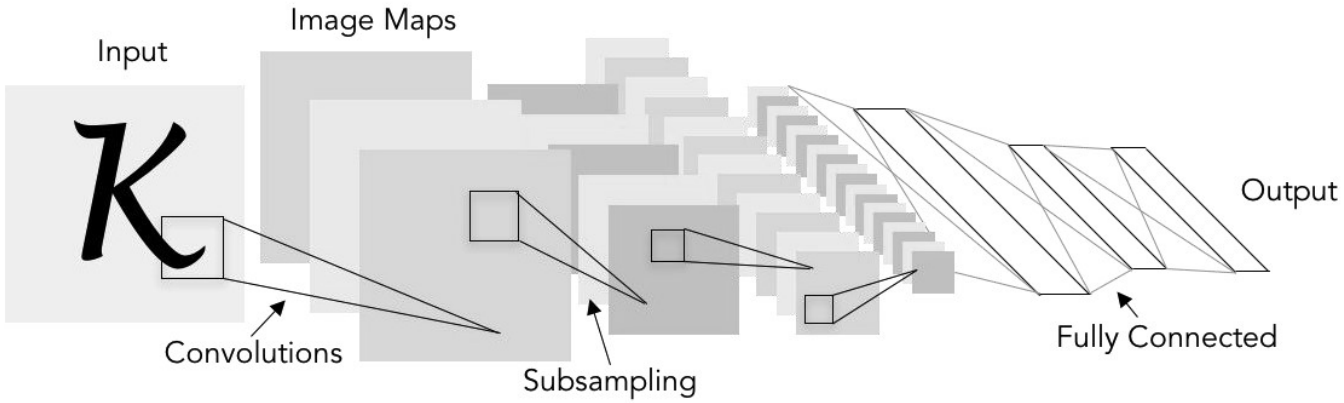


"sandwich" architecture (SCSCSC…)
simple cells: modifiable parameters
complex cells: perform pooling

A bit of history:

Gradient-based learning applied to document recognition

[LeCun, Bottou, Bengio, Haffner 1998]



LeNet-5

A bit of history:

ImageNet Classification with Deep Convolutional Neural Networks

[Krizhevsky, Sutskever, Hinton, 2012]

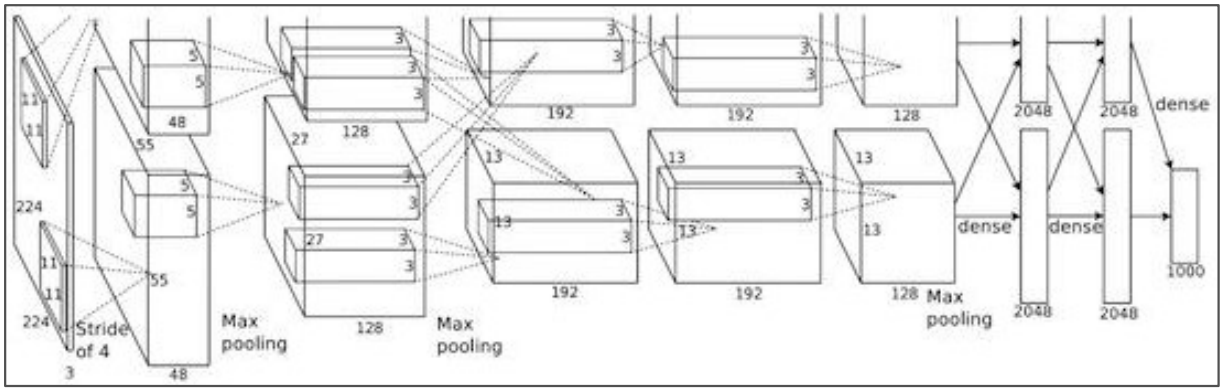


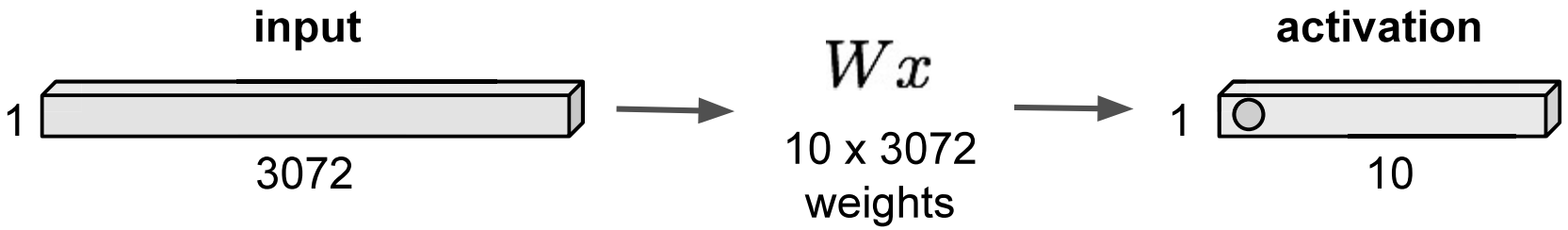
Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

"AlexNet"

What is wrong with standard NN for image as input?

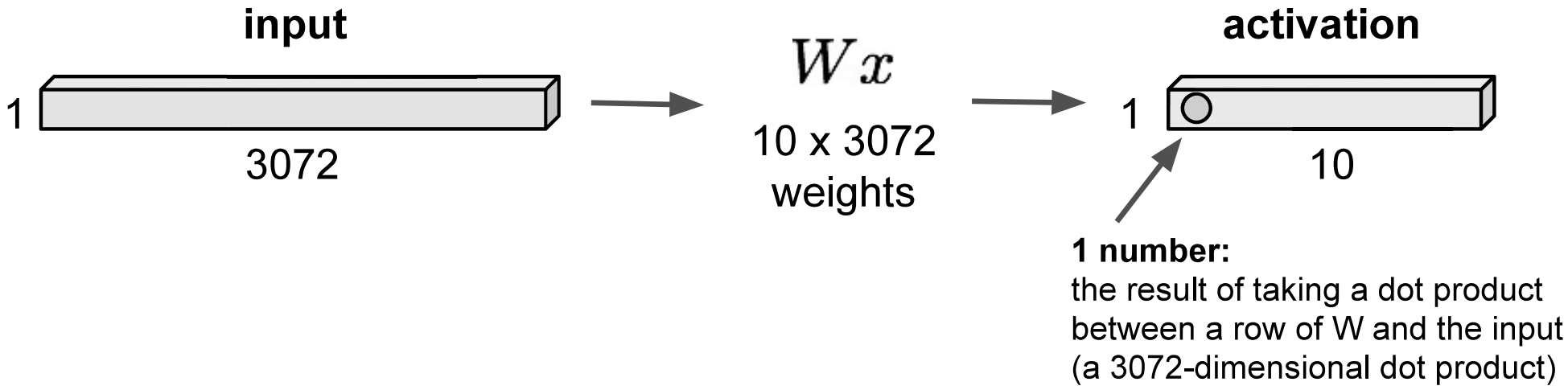
Fully Connected Layer

32x32x3 image -> stretch to 3072 x 1



Fully Connected Layer

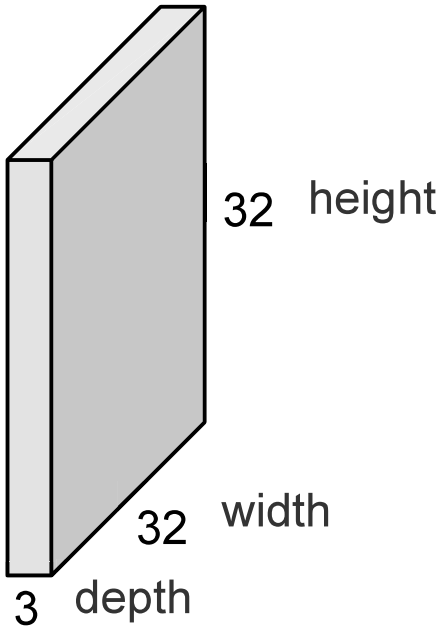
32x32x3 image -> stretch to 3072 x 1



Convolution with filters to preserve spatial structure

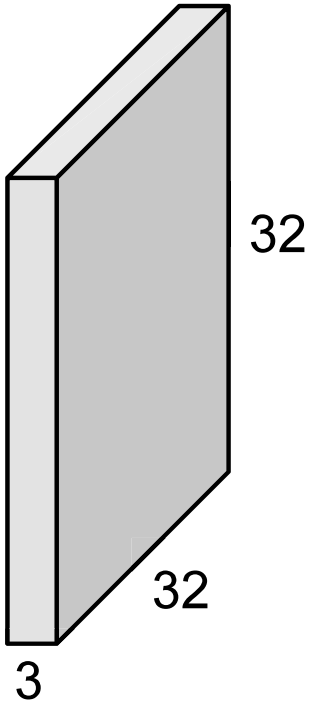
Convolution Layer

32x32x3 image -> preserve spatial structure

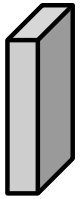


Convolution Layer

32x32x3 image



5x5x3 filter

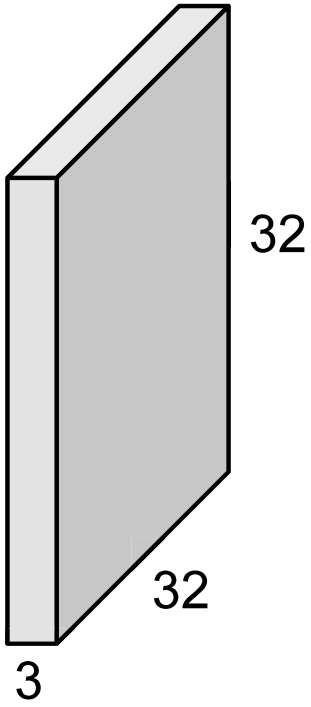


Convolve the filter with the image
i.e. “slide over the image spatially,
computing dot products”

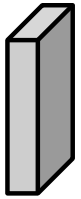
Convolution Layer

Filters always extend the full depth of the input volume

32x32x3 image

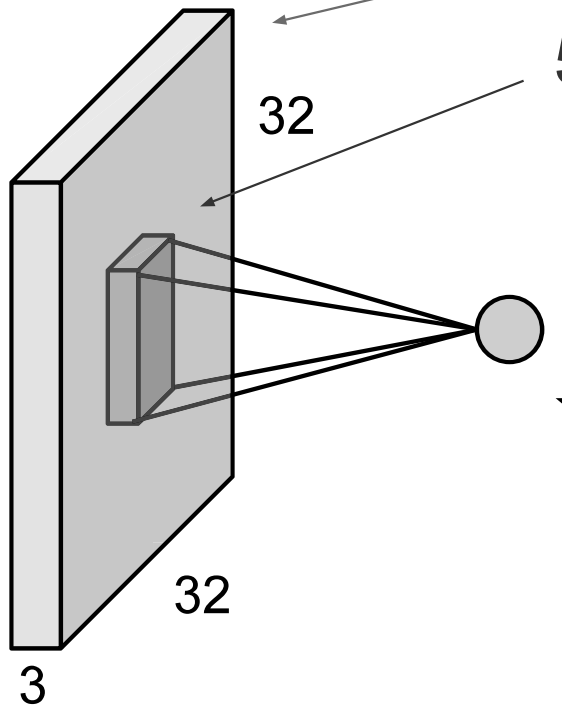


5x5x3 filter



Convolve the filter with the image i.e. “slide over the image spatially, computing dot products”

Convolution Layer



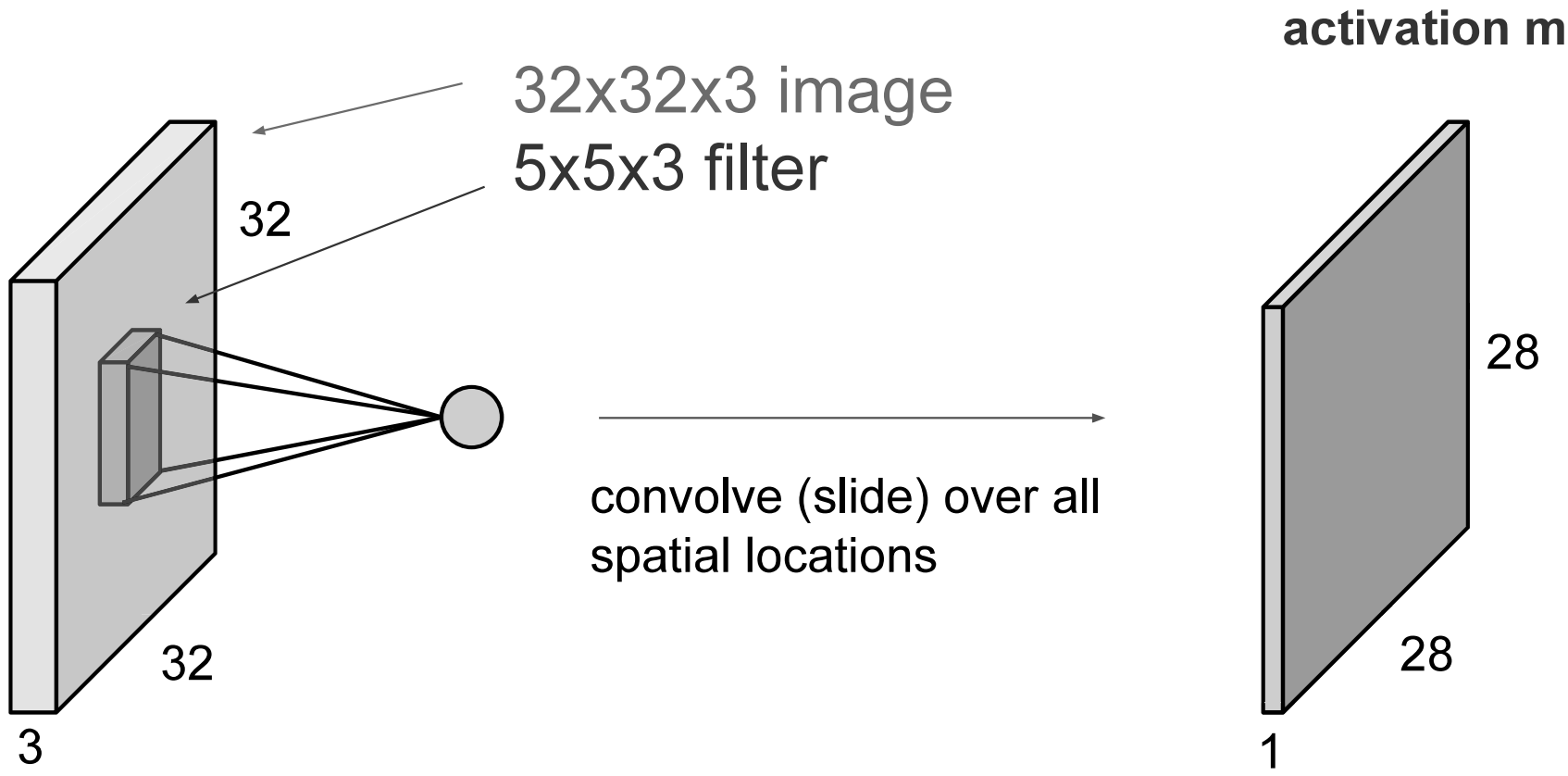
32x32x3 image

5x5x3 filter w

1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the image
(i.e. 5*5*3 = 75-dimensional dot product + bias)

$$w^T x + b$$

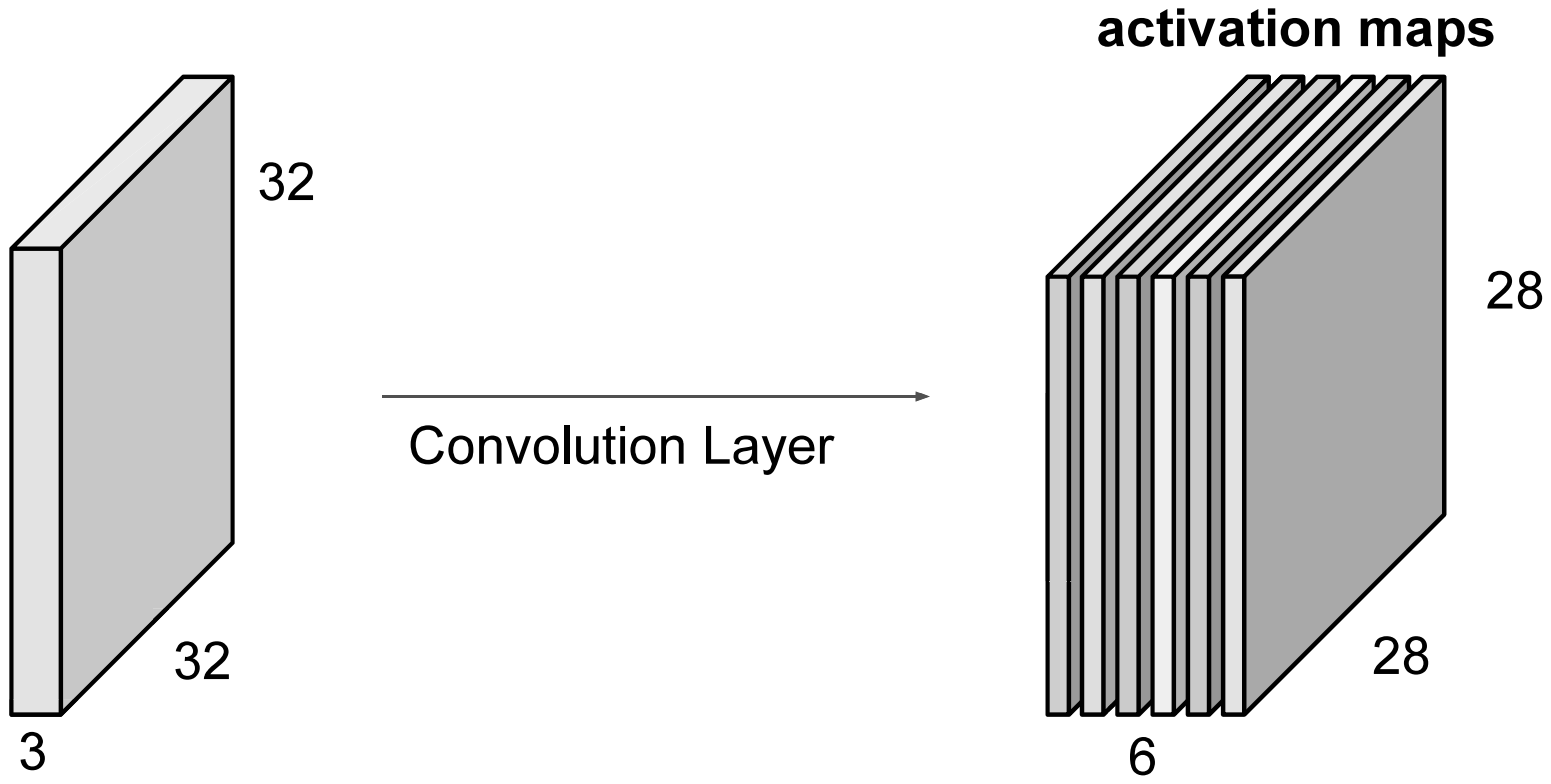
Convolution Layer



Convolution Layer

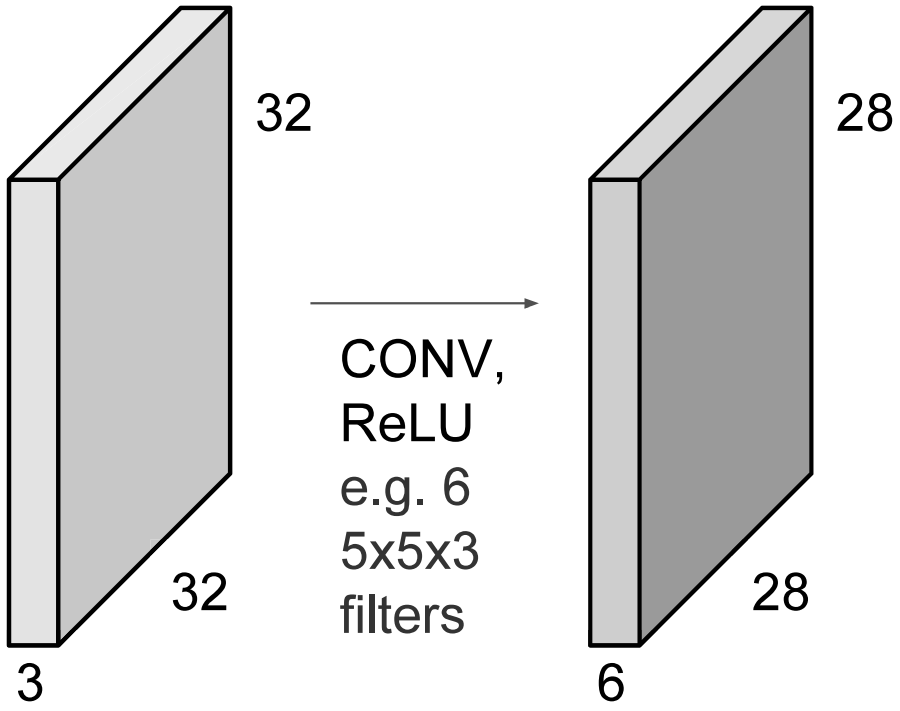
consider a second, green filter

For example, if we had 6 5x5 filters, we'll get 6 separate activation maps:

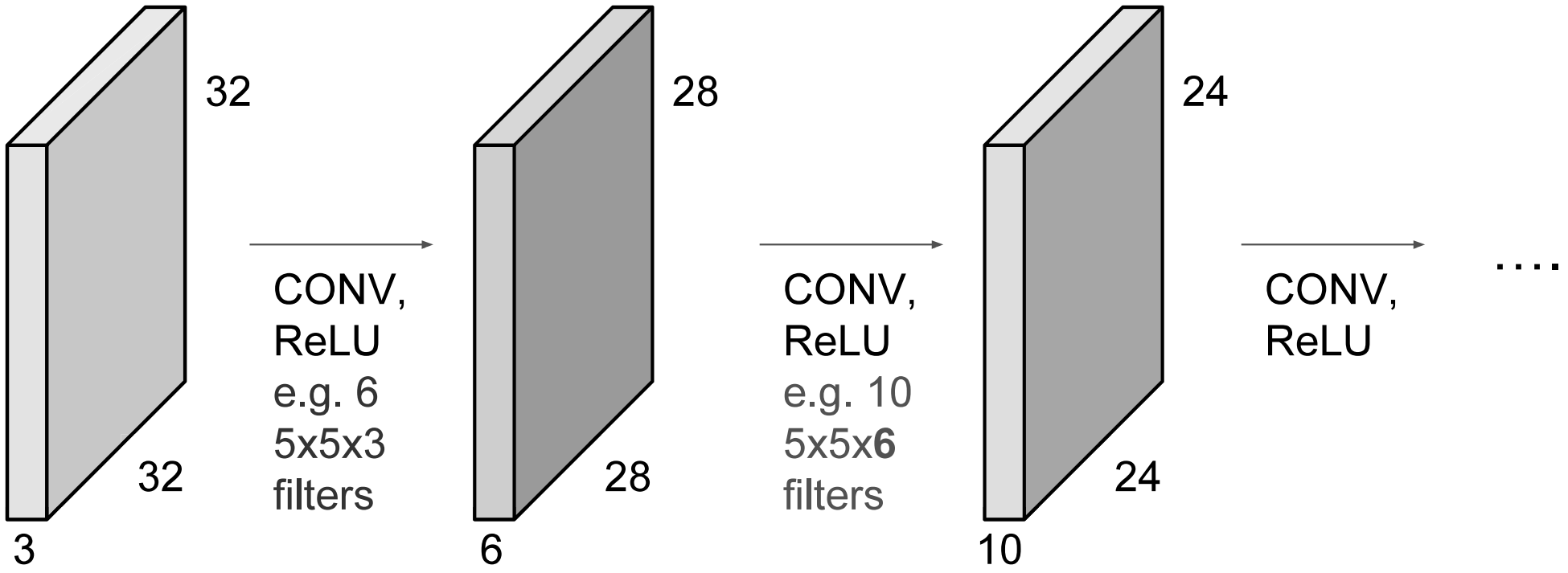


We stack these up to get a "new image" of size 28x28x6!

Preview: ConvNet is a sequence of Convolution Layers, interspersed with activation functions



Preview: ConvNet is a sequence of Convolutional Layers, interspersed with activation functions

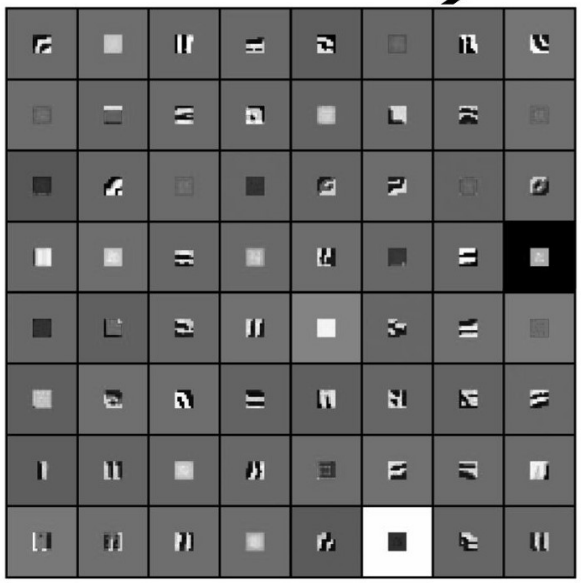


Preview

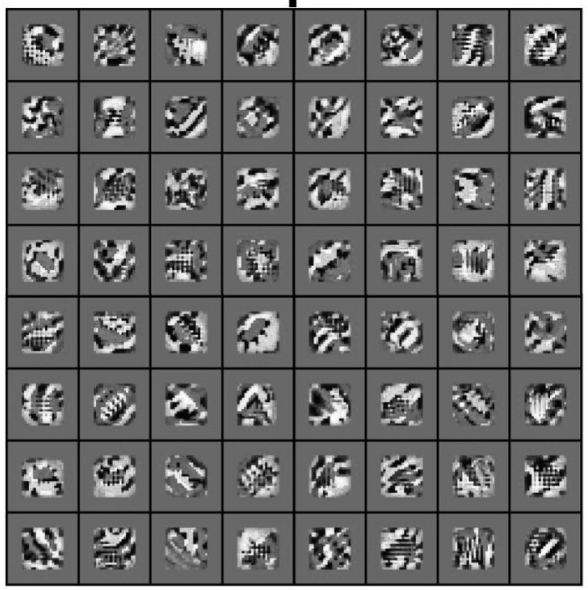
[Zeiler and Fergus 2013]

Visualization of VGG-16 by Lane McIntosh. VGG-16 architecture from [Simonyan and Zisserman 2014].

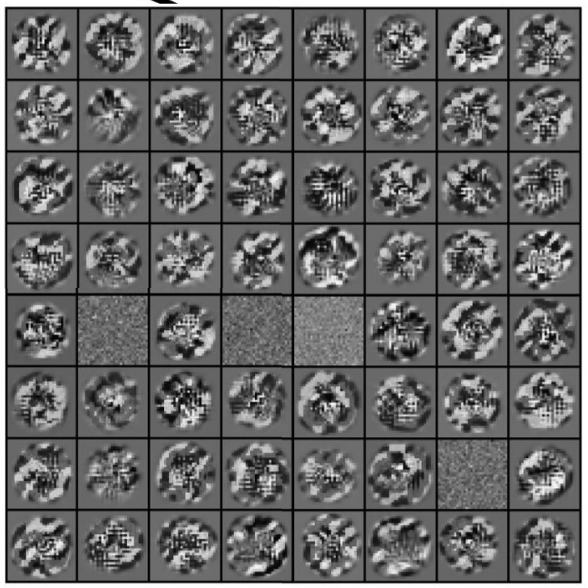
Low-level features → Mid-level features → High-level features → Linearly separable classifier



VGG-16 Conv1_1

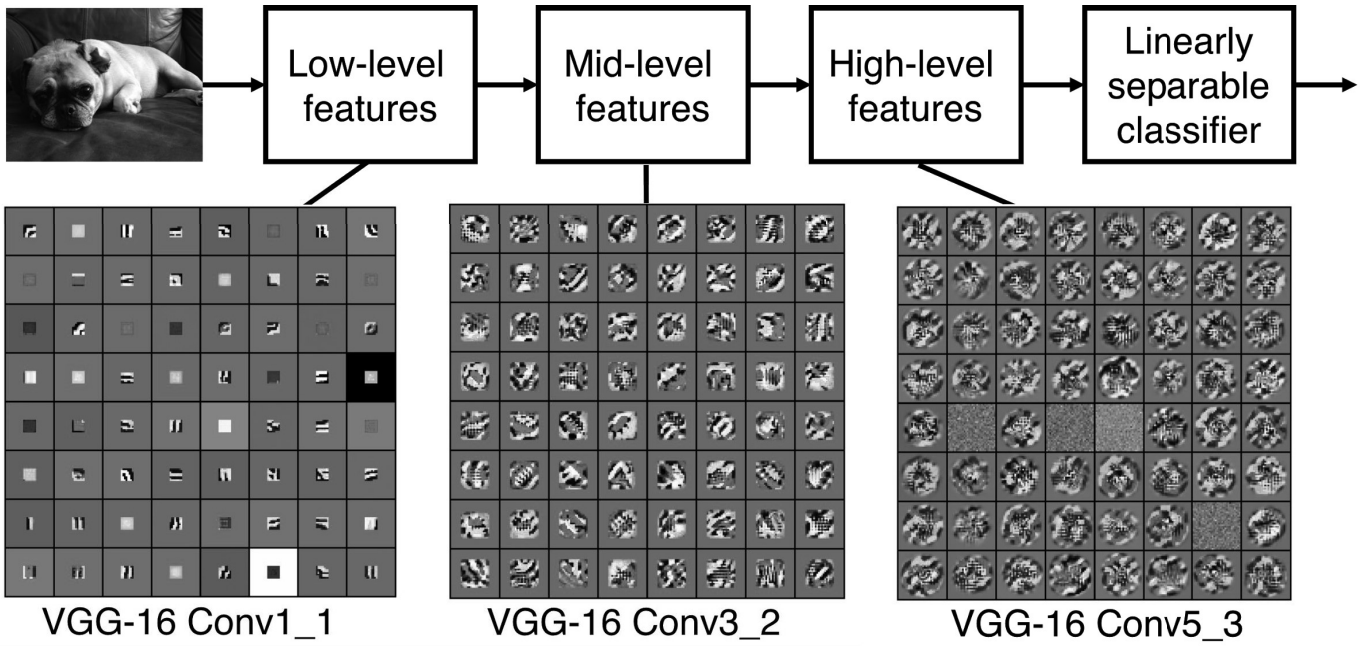


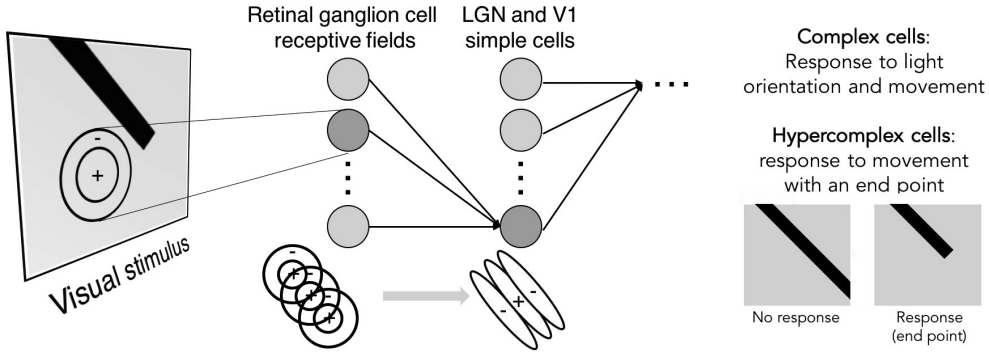
VGG-16 Conv3_2

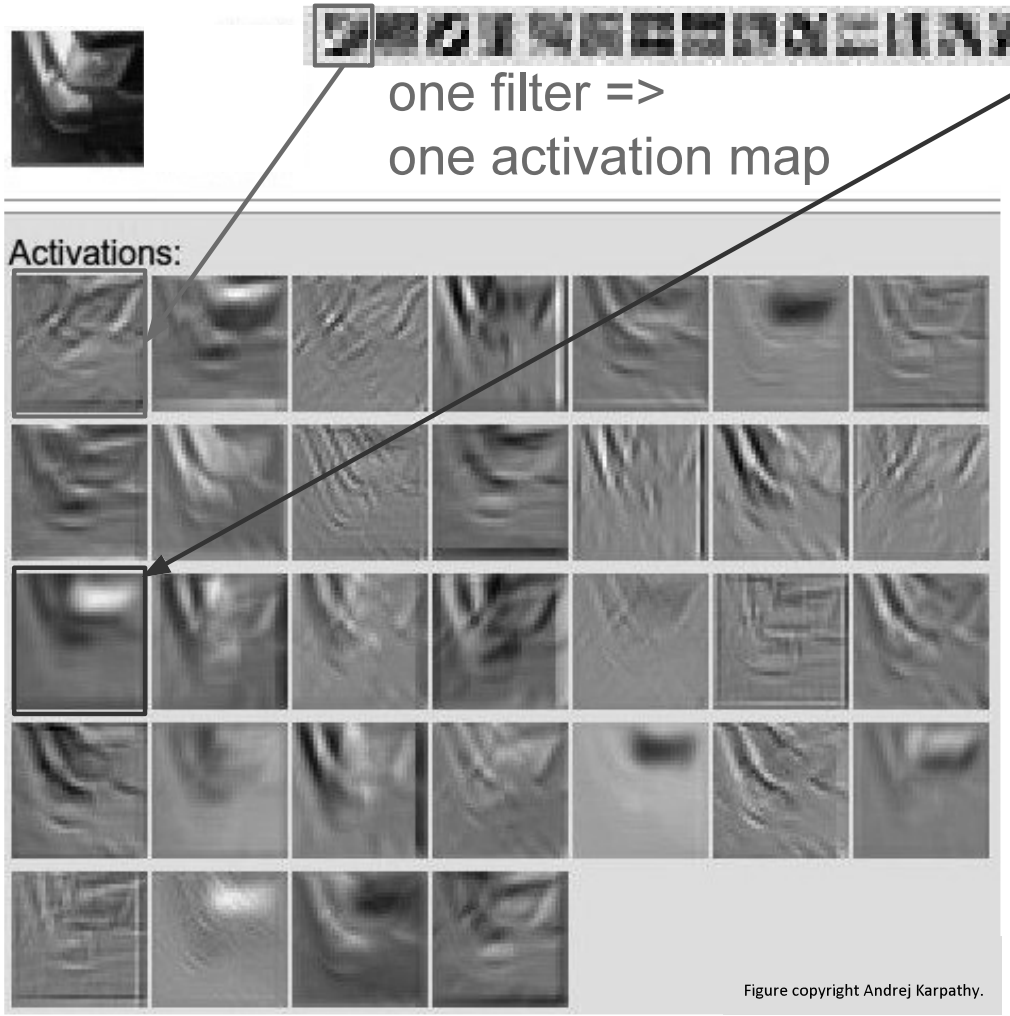


VGG-16 Conv5_3

Preview


Low-level features
Mid-level features
High-level features
Linearly separable classifier
VGG-16 Conv1_1
VGG-16 Conv3_2
VGG-16 Conv5_3


Retinal ganglion cell receptive fields
LGN and V1 simple cells
Visual stimulus
Complex cells:
Response to light orientation and movement
Hypercomplex cells:
response to movement with an end point
No response
Response (end point)



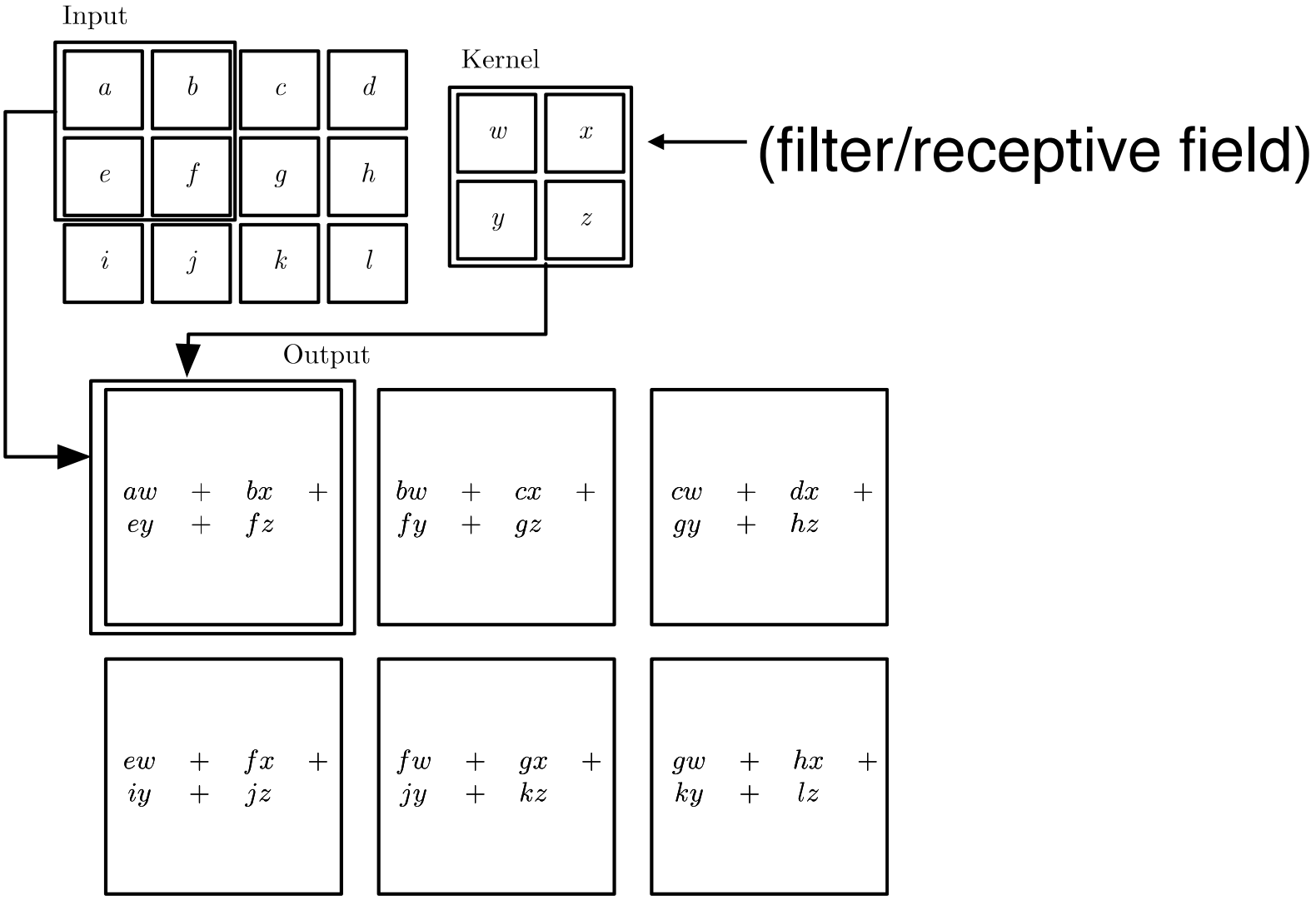
example 5x5 filters
(32 total)

We call the layer convolutional because it is related to convolution of two signals:

$$f[x,y] * g[x,y] = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f[n_1,n_2] \cdot g[x-n_1, y-n_2]$$

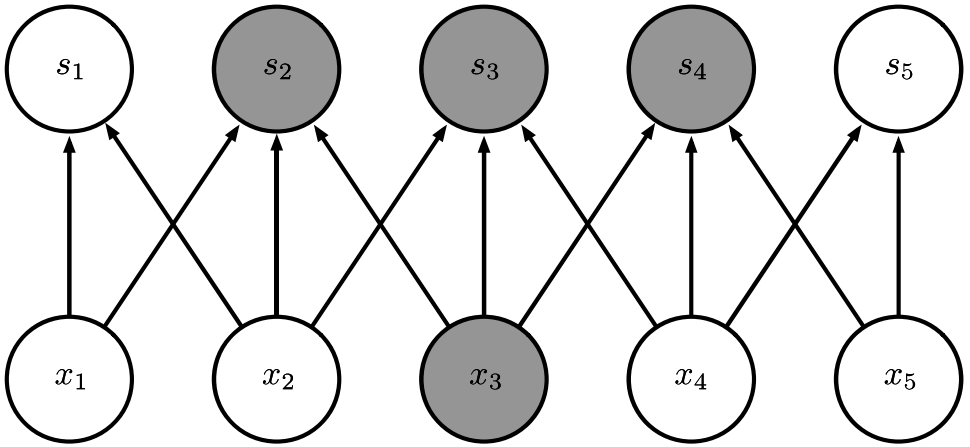
elementwise multiplication and sum of a filter and the signal (image)

2D Convolution

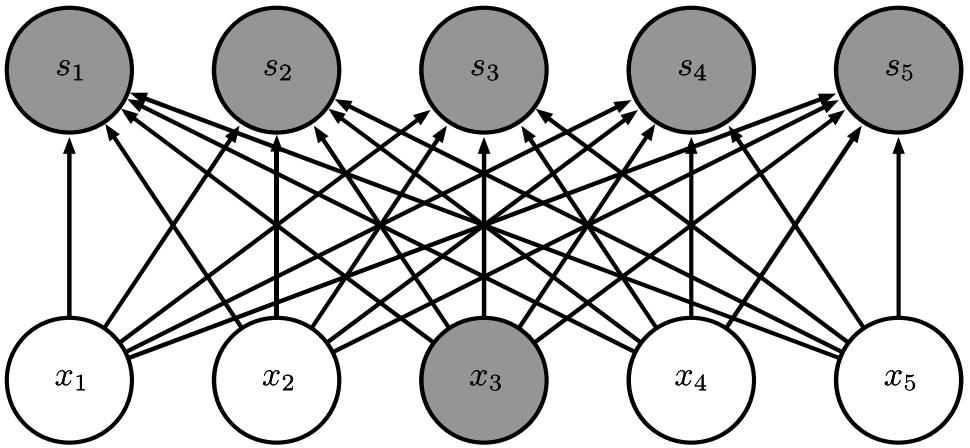


Local Receptive Field Leads to Sparse Connectivity (affects less)

Sparse connections due to small convolution kernel

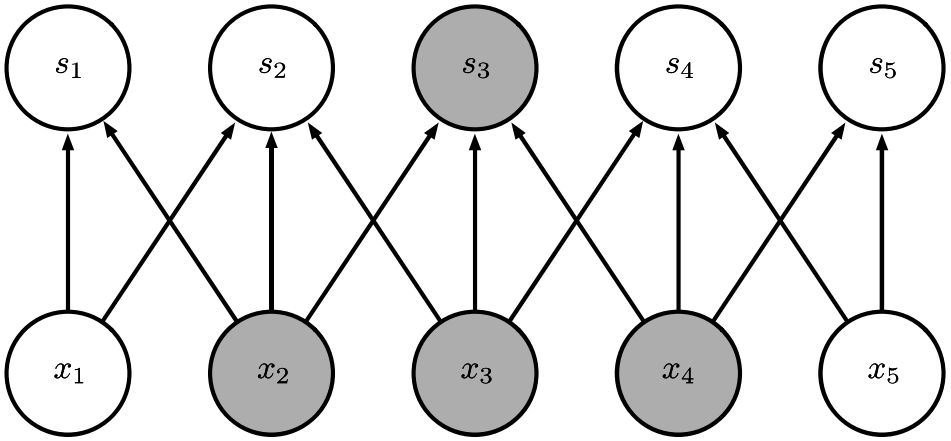


Dense connections



Sparse connectivity: being affected by less

Sparse connections due to small convolution kernel



Dense connections

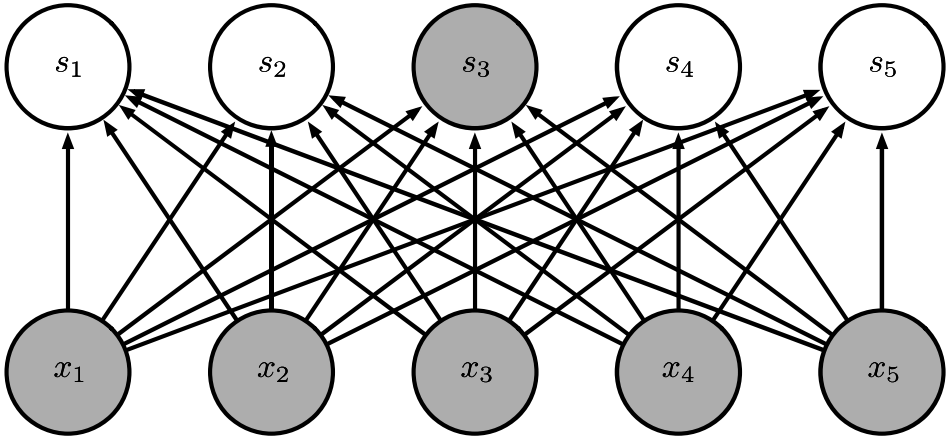


Figure 9.3

However, “depth” by stacking layers creates larger receptive fields

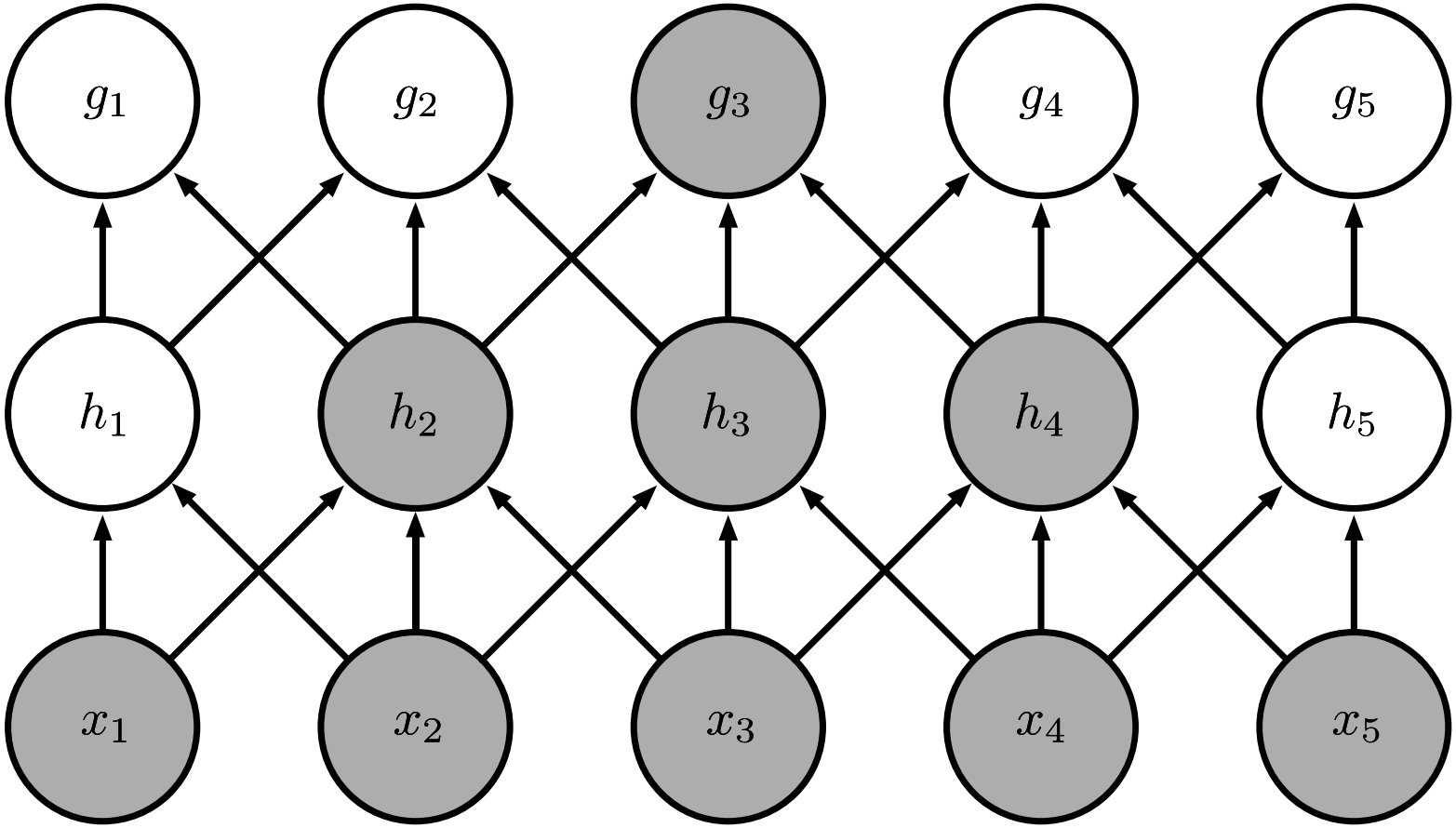
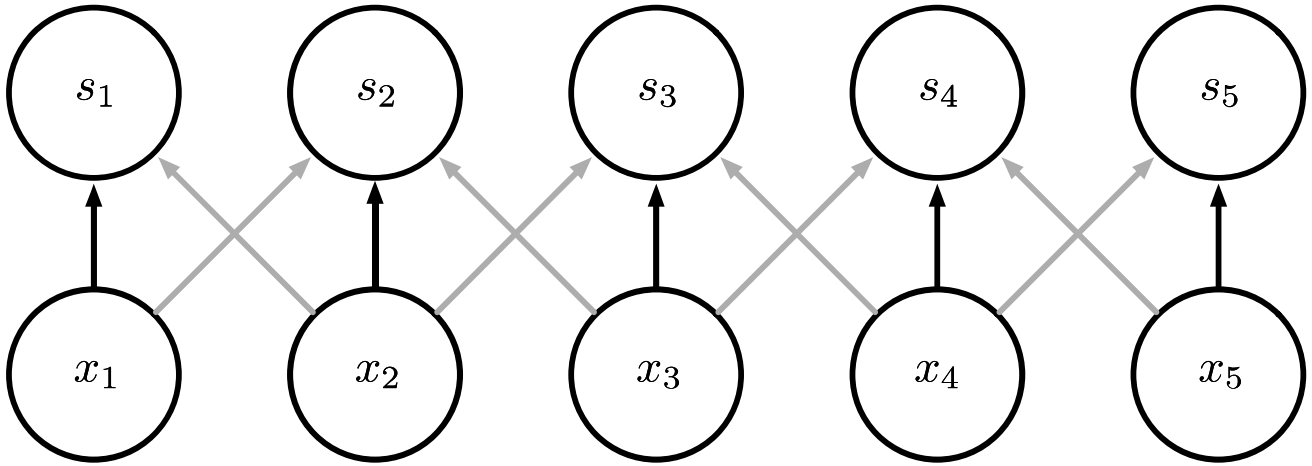


Figure 9.4

Parameter Sharing

Convolution shares the same parameters across all spatial locations



Traditional matrix multiplication does not share any parameters

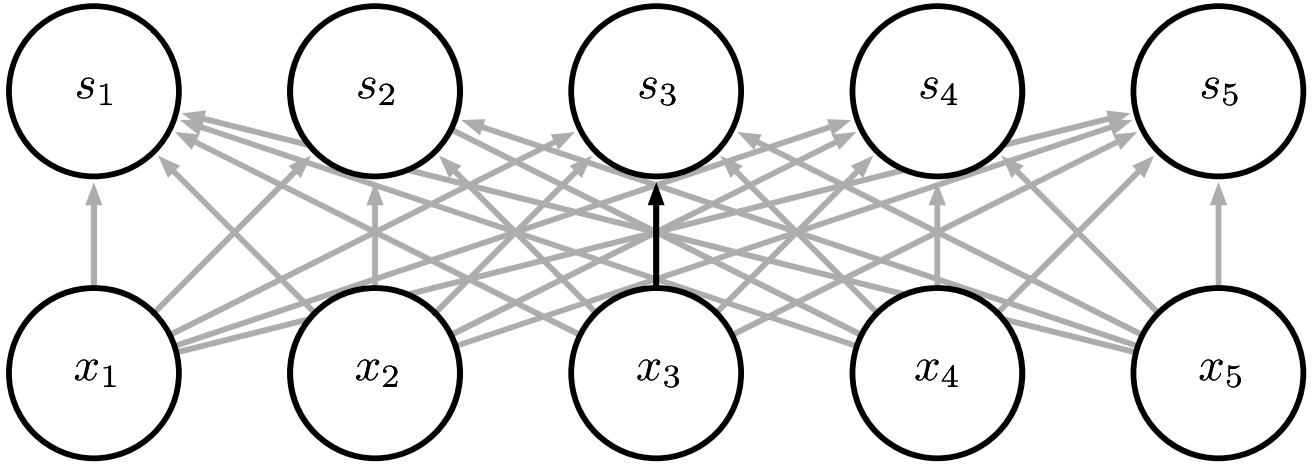
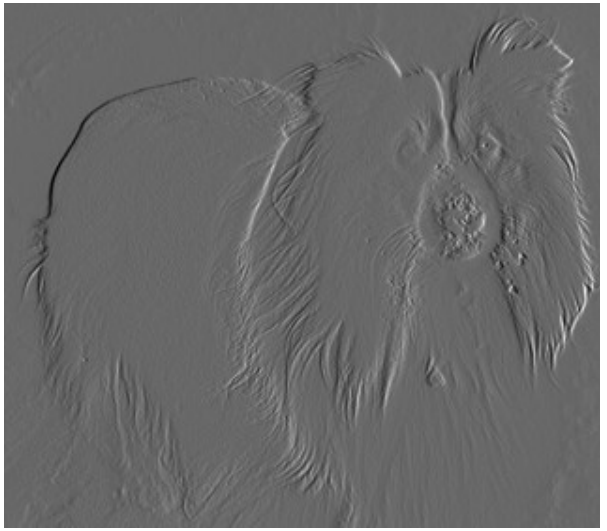


Figure 9.5

A simple example why filtering is useful

Input

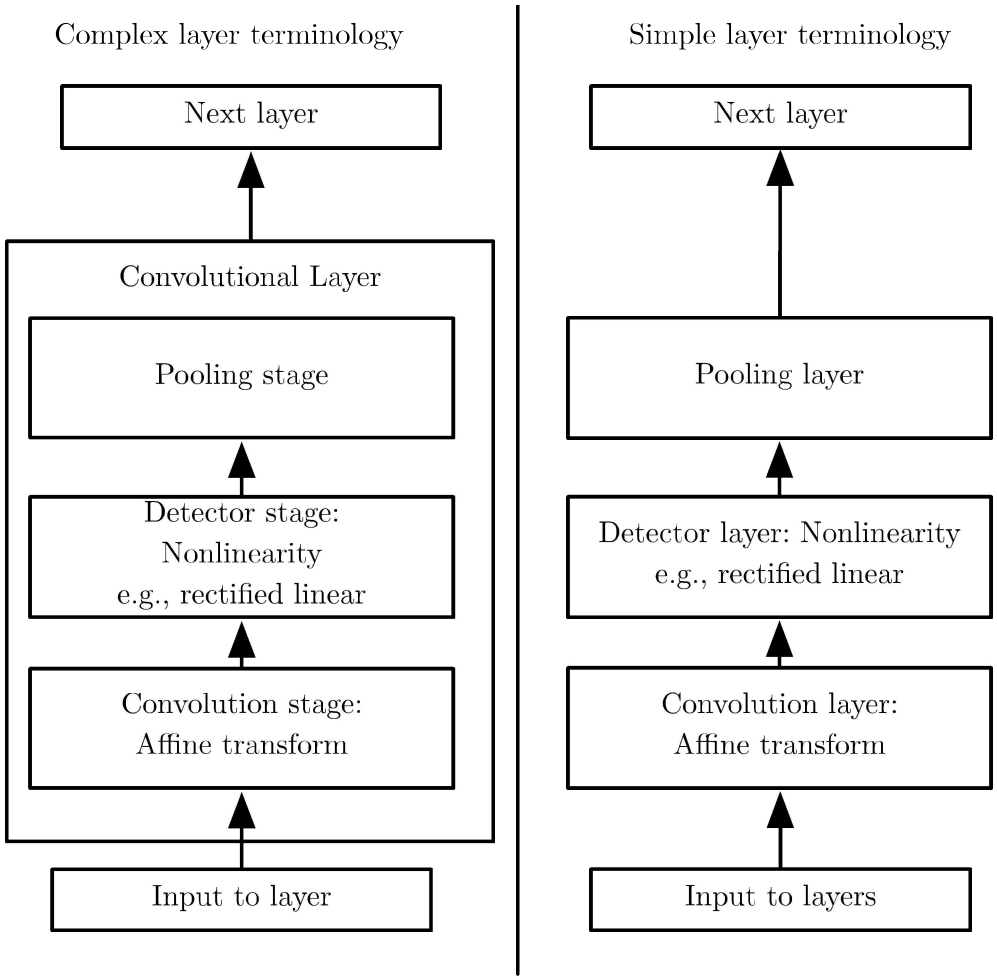


Output

1	-1

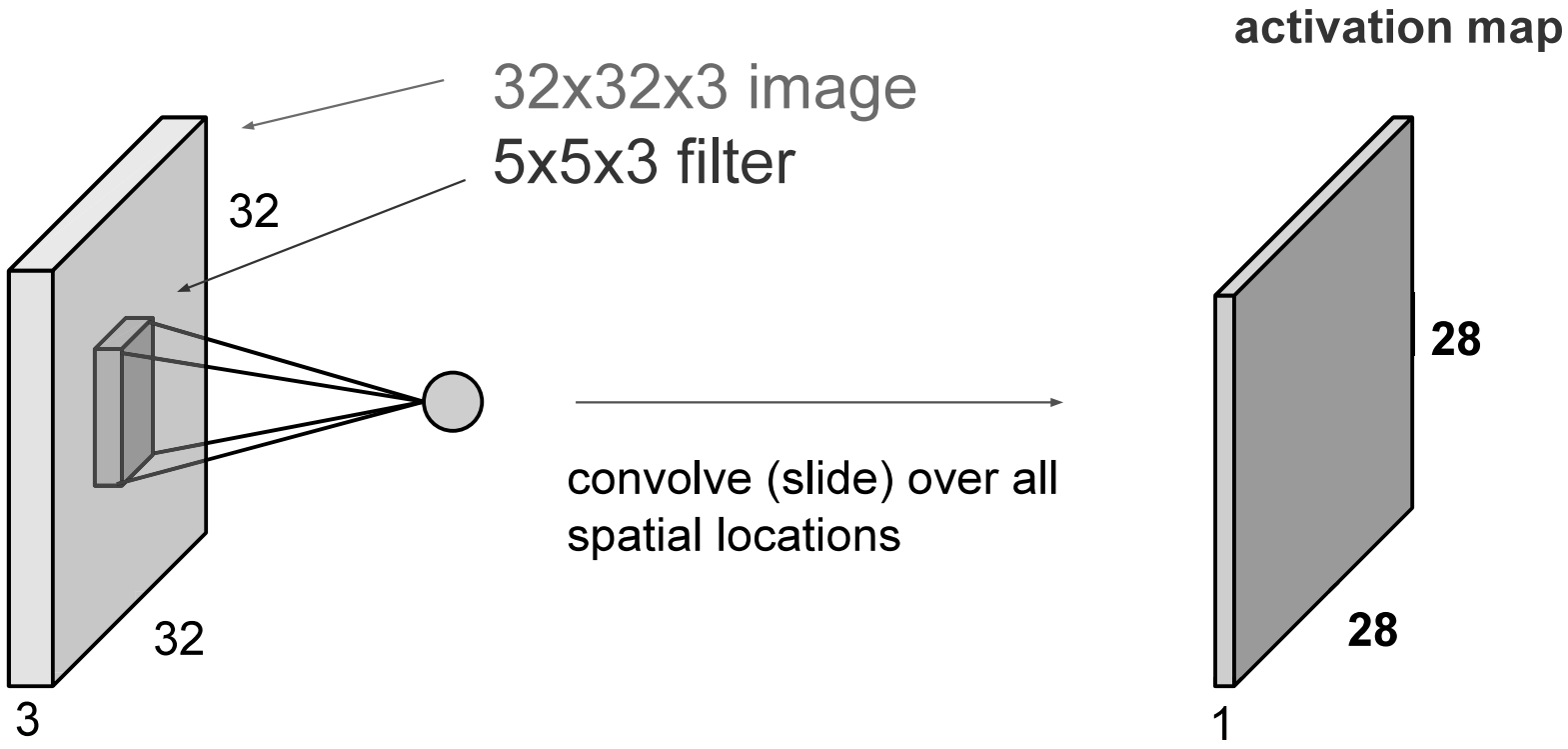
Kernel

Compounding layers as convolution neural nets (CNNs)

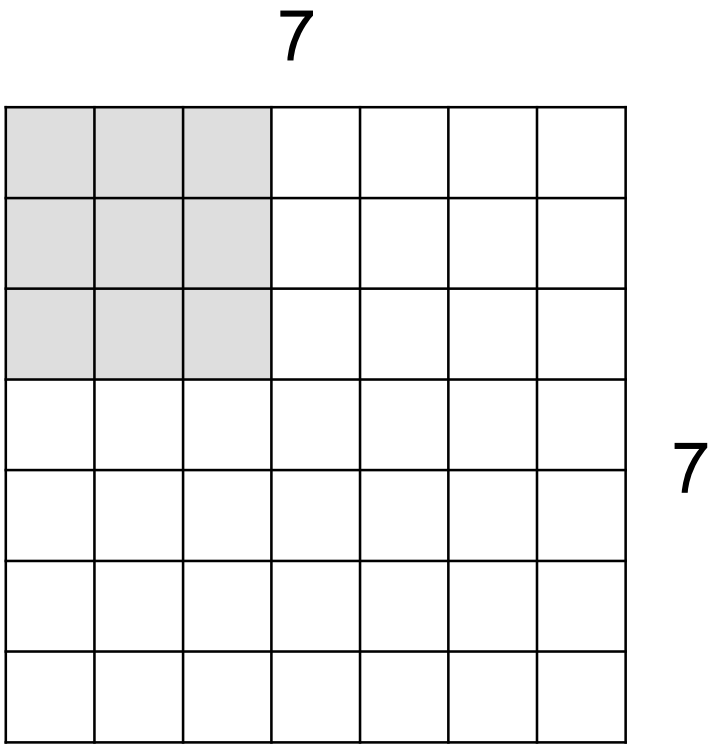


Stride and padding for convolution

A closer look at spatial dimensions:

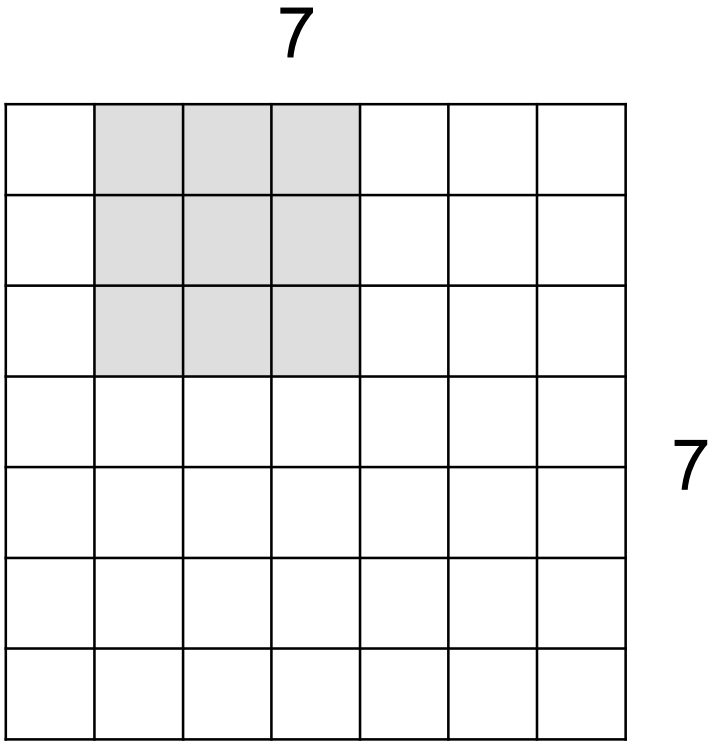


A closer look at spatial dimensions:



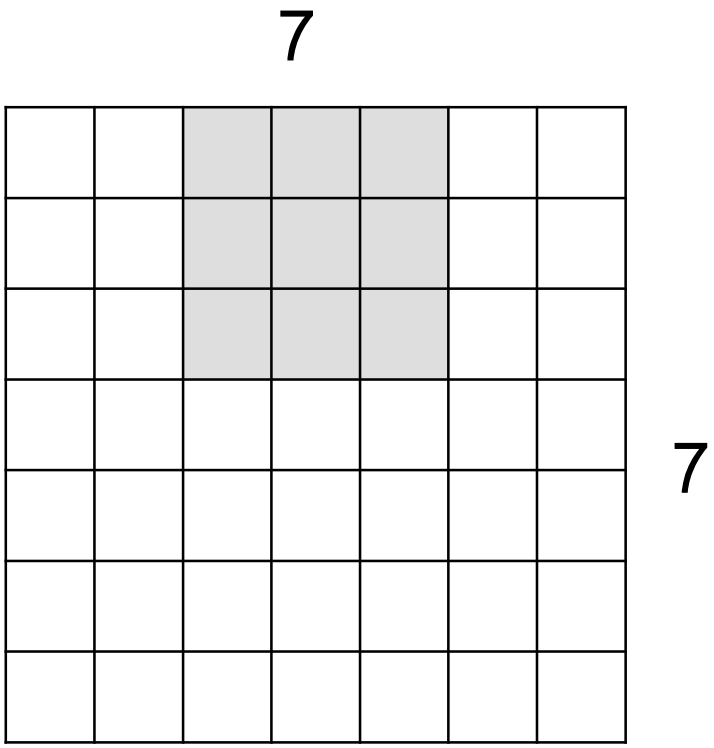
7x7 input (spatially)
assume 3x3 filter

A closer look at spatial dimensions:



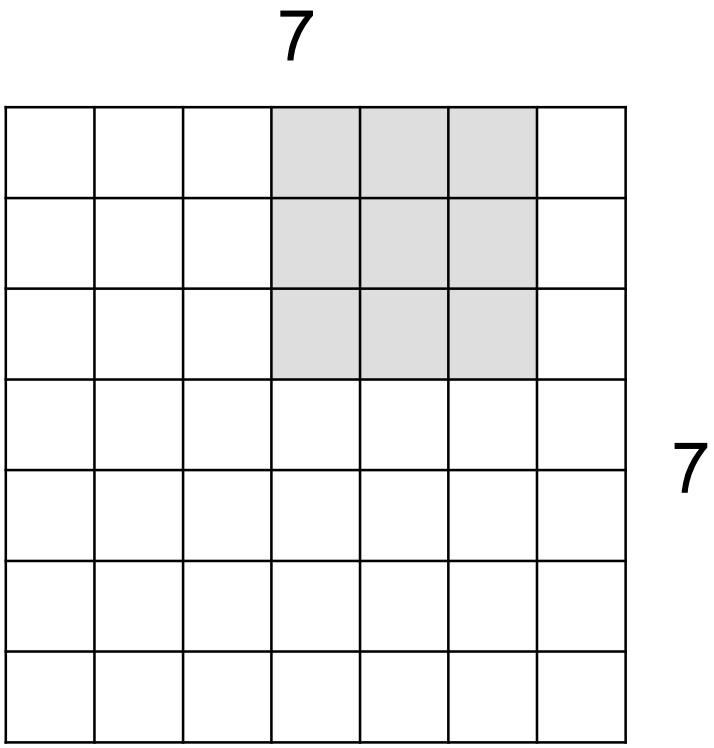
7x7 input (spatially)
assume 3x3 filter

A closer look at spatial dimensions:



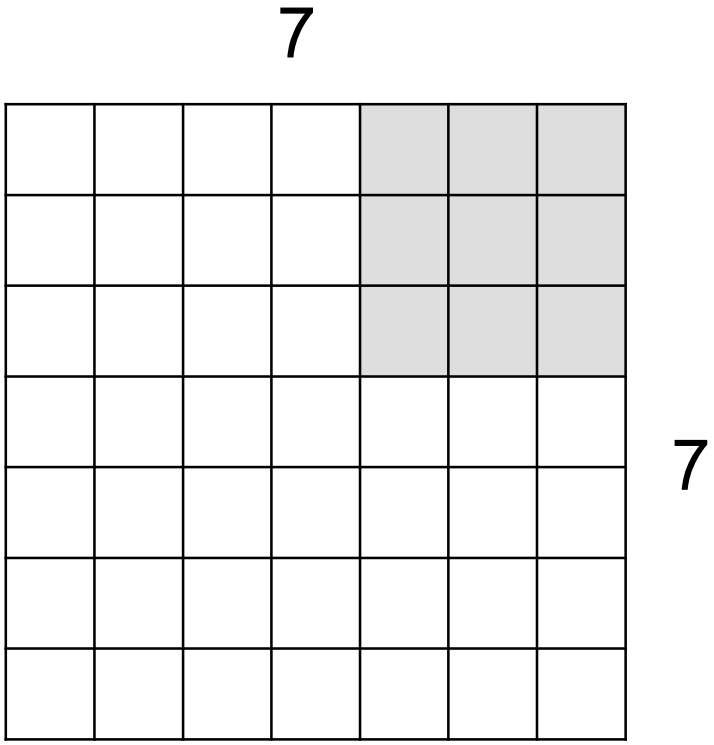
7x7 input (spatially)
assume 3x3 filter

A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter

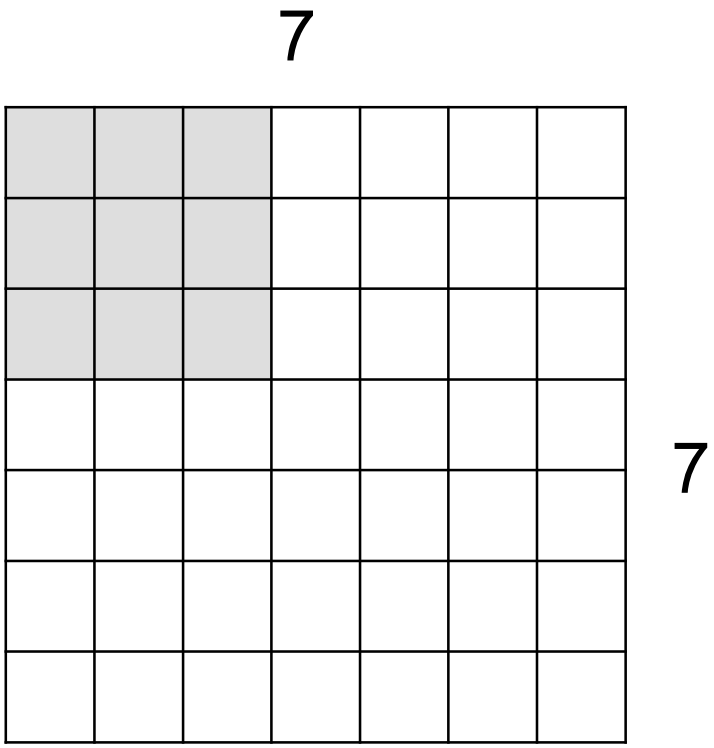
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter

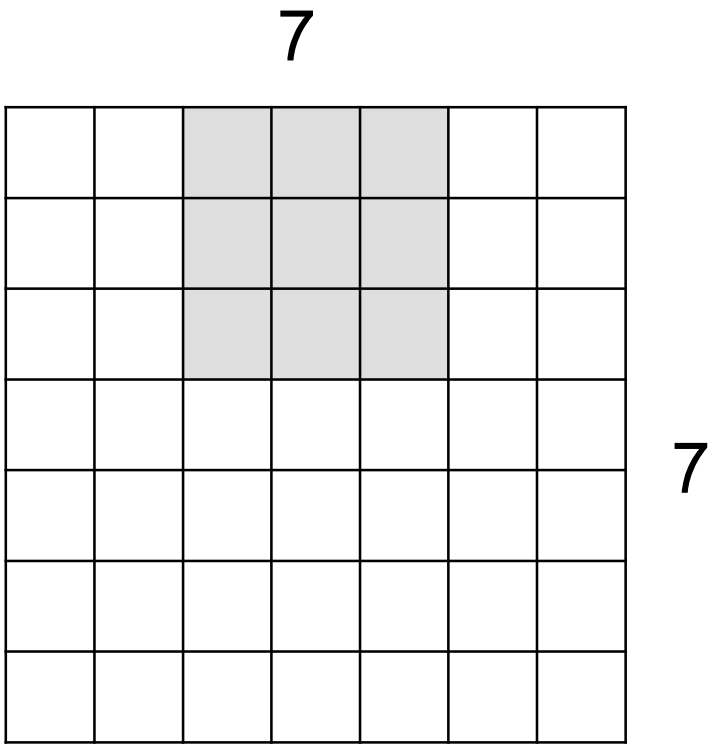
=> 5x5 output

A closer look at spatial dimensions:



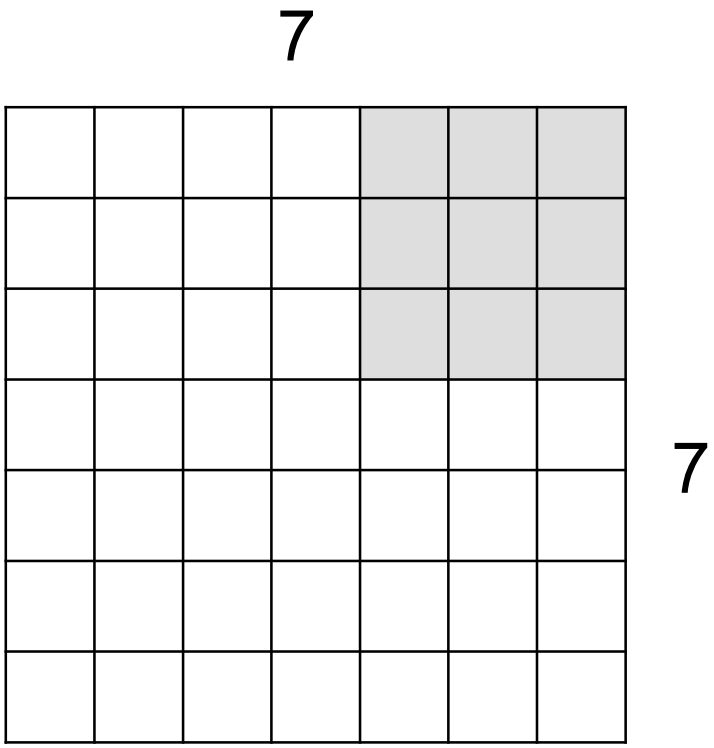
7x7 input (spatially)
assume 3x3 filter
applied **with stride 2**

A closer look at spatial dimensions:



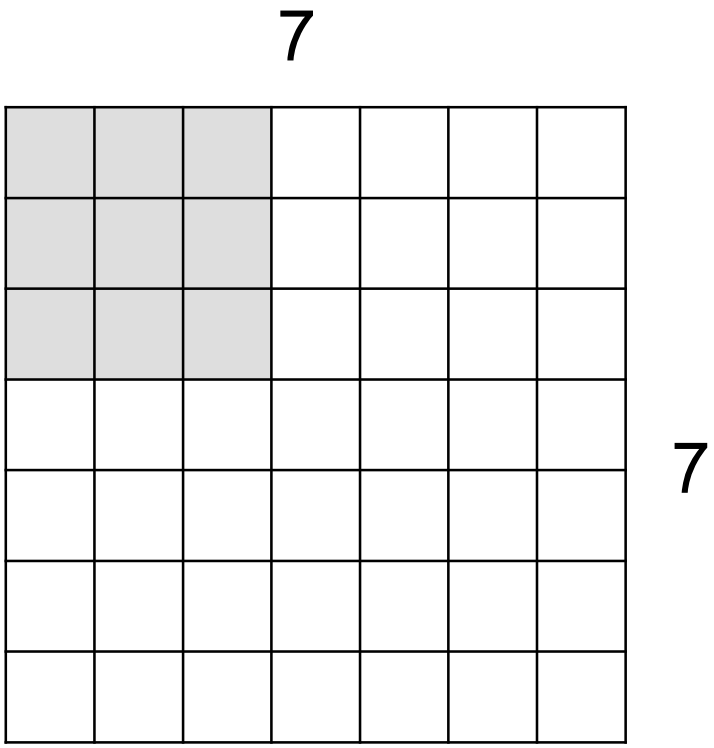
7x7 input (spatially)
assume 3x3 filter
applied **with stride 2**

A closer look at spatial dimensions:



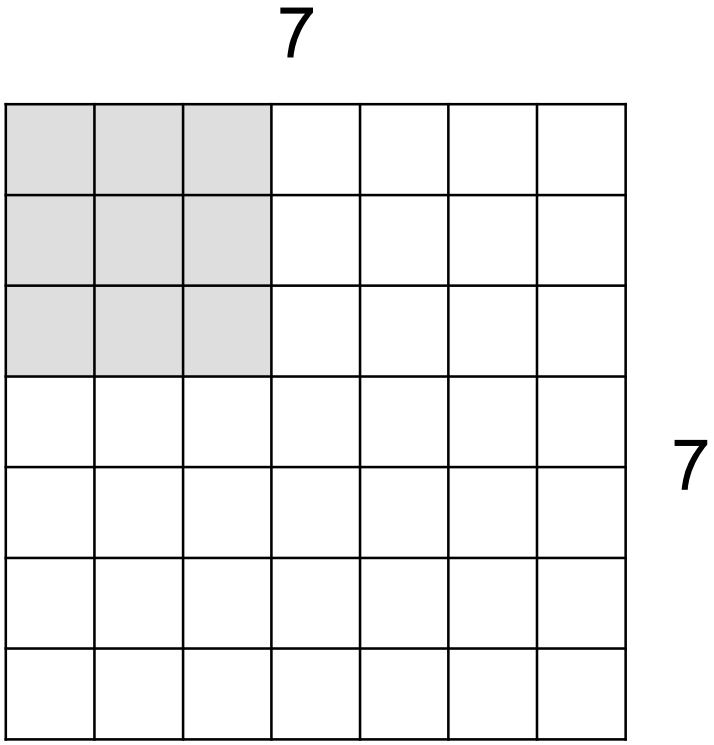
7x7 input (spatially)
assume 3x3 filter
applied **with stride 2**
=> 3x3 output!

A closer look at spatial dimensions:



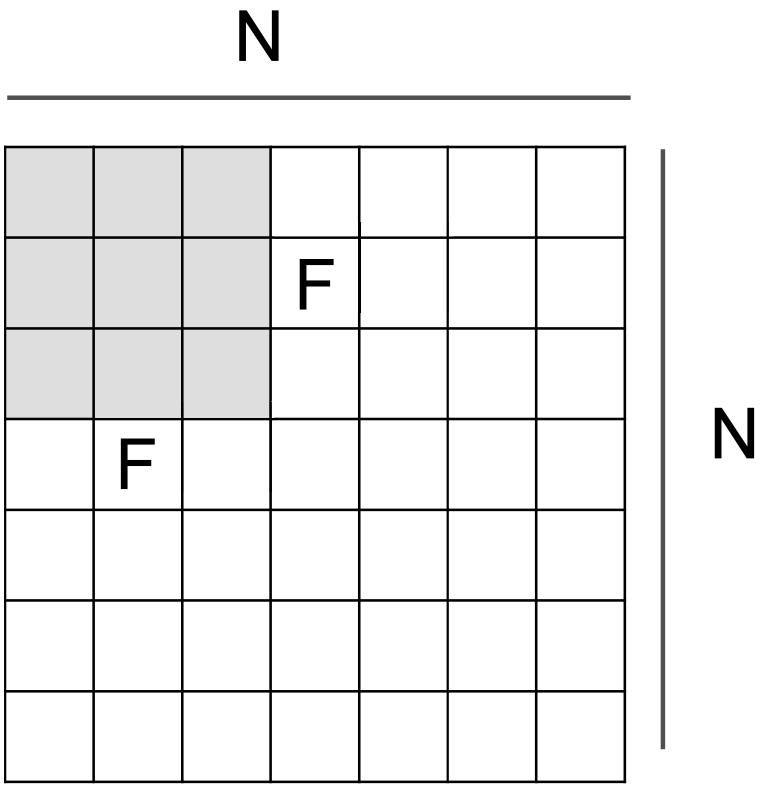
7x7 input (spatially)
assume 3x3 filter
applied **with stride 3?**

A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter
applied **with stride 3?**

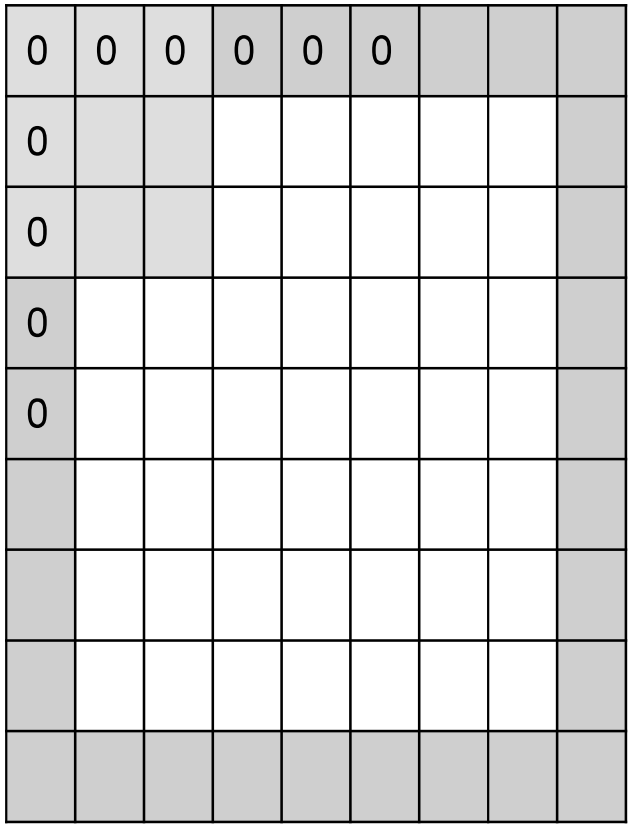
doesn't fit!
cannot apply 3x3 filter on
7x7 input with stride 3.



Output size:
(N - F) / stride + 1

e.g. N = 7, F = 3:
stride 1 => (7 - 3)/1 + 1 = 5
stride 2 => (7 - 3)/2 + 1 = 3
stride 3 => (7 - 3)/3 + 1 = 2.33 :\

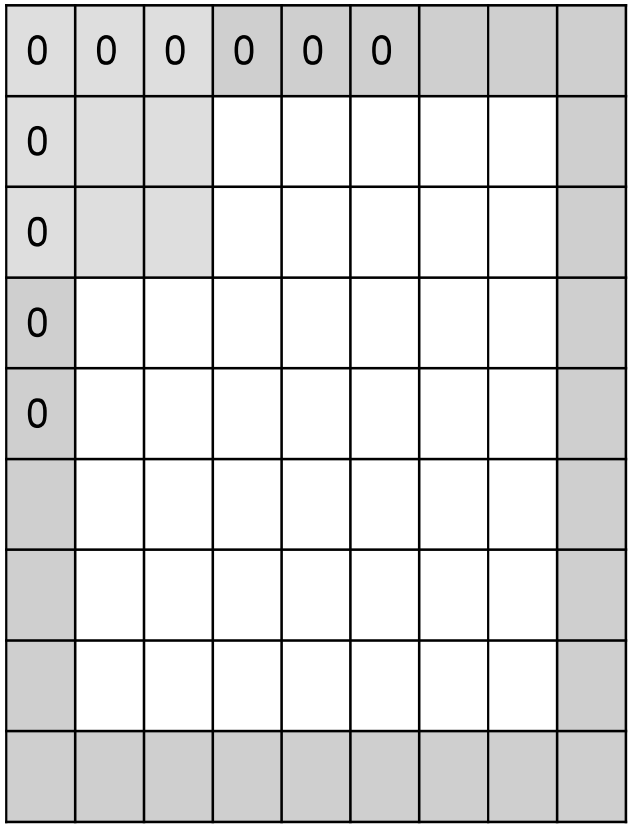
In practice: Common to zero pad the border



e.g. input 7x7
3x3 filter, applied with **stride 1**
pad with 1 pixel border => what is the output?

(recall:)
(N - F) / stride + 1

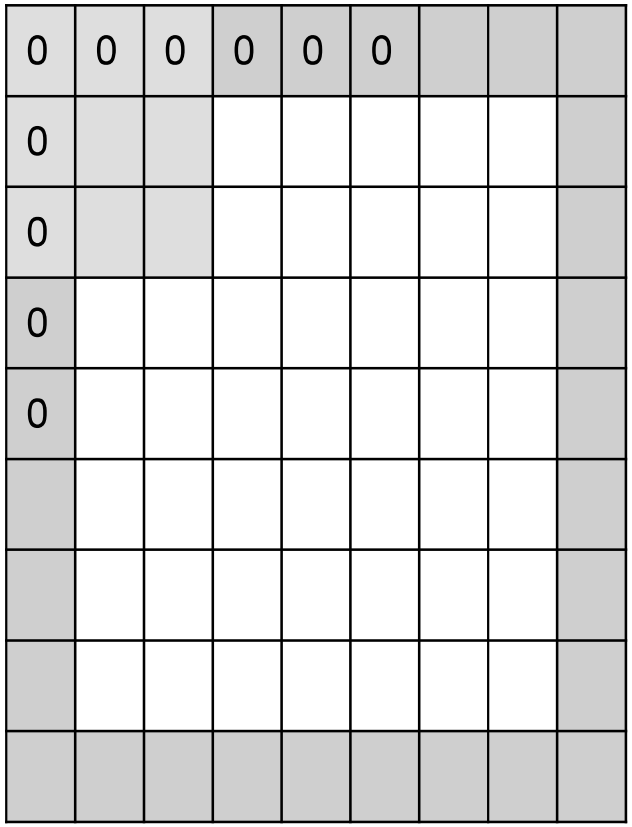
In practice: Common to zero pad the border



e.g. input 7x7
3x3 filter, applied with **stride 1**
pad with 1 pixel border => what is the output?

7x7 output!

In practice: Common to zero pad the border

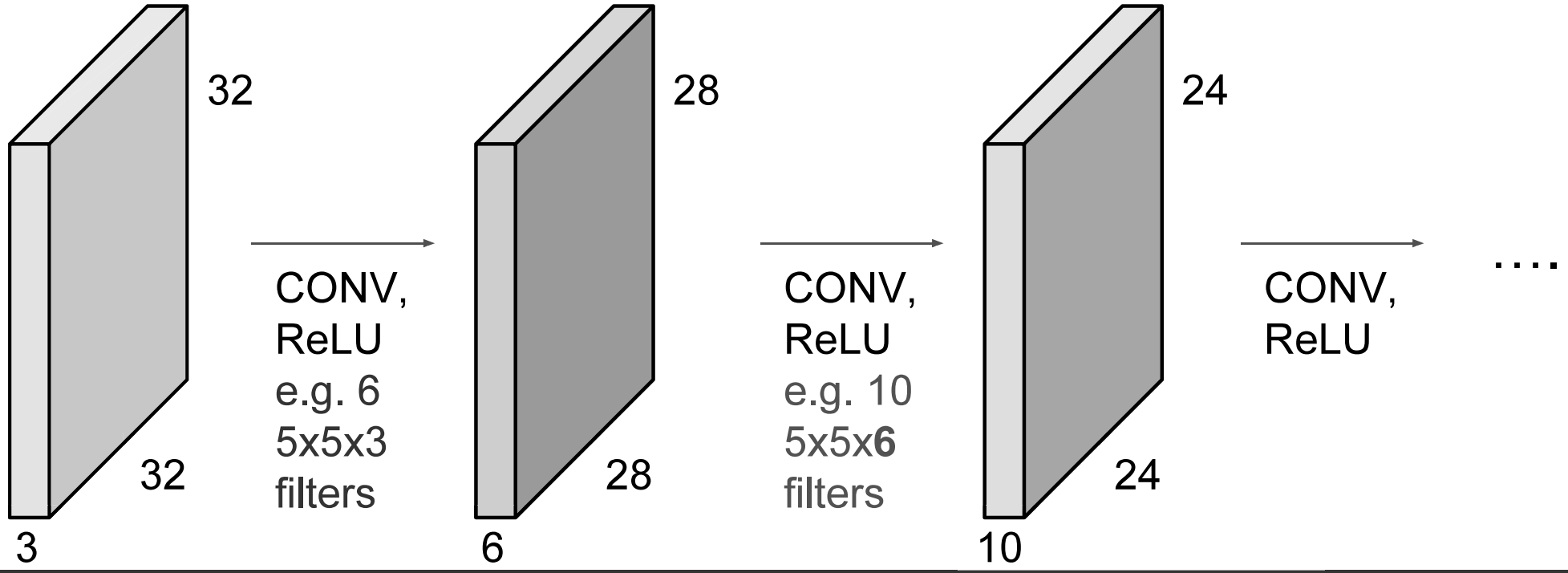


e.g. input 7x7
3x3 filter, applied with **stride 1**
pad with 1 pixel border => what is the output?

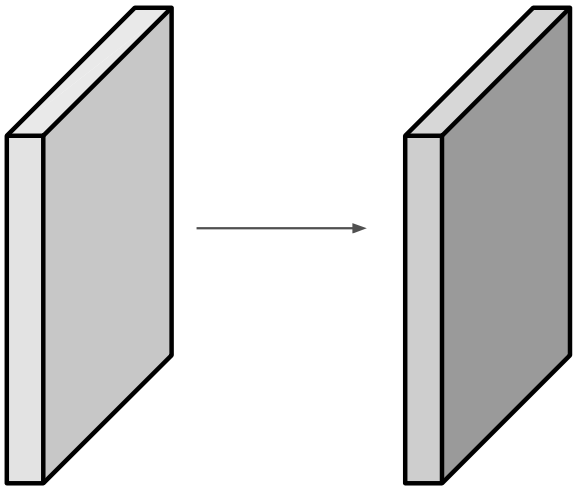
7x7 output!
in general, common to see CONV layers with stride 1, filters of size FxF, and zero-padding with (F-1)/2. (will preserve size spatially)
e.g. F = 3 => zero pad with 1
F = 5 => zero pad with 2
F = 7 => zero pad with 3

Remember back to…

E.g. 32x32 input convolved repeatedly with 5x5 filters shrinks volumes spatially! (32 -> 28 -> 24 ...). Shrinking too fast is not good, doesn't work well.



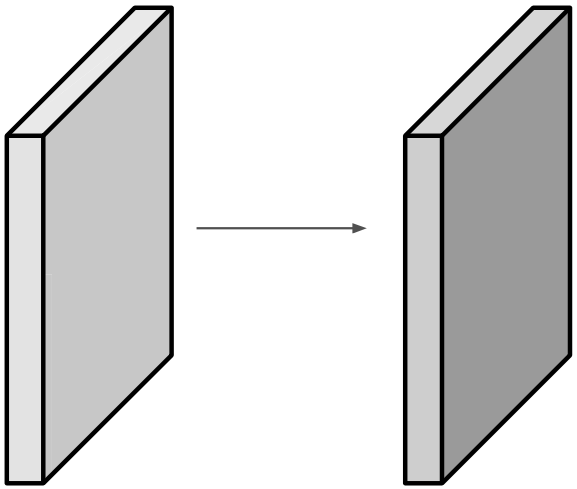
Examples time:



Input volume: **32x32x3**
10 5x5 filters with stride 1, pad 2

Output volume size: ?

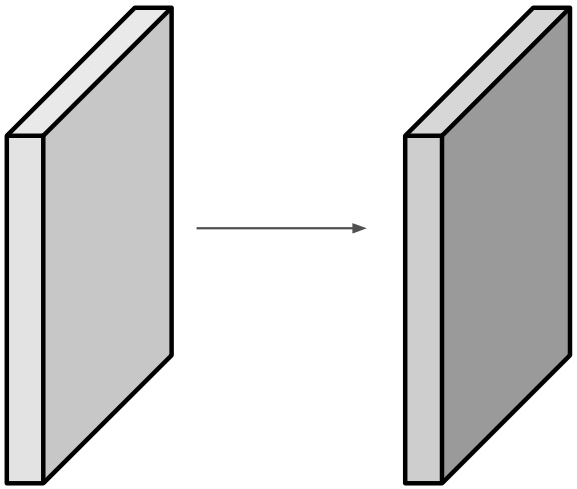
Examples time:



Input volume: **32x32x3**
10 5x5 filters with stride 1, pad 2

Output volume size:
(32+2*2-5)/1+1 = 32 spatially, so
32x32x10

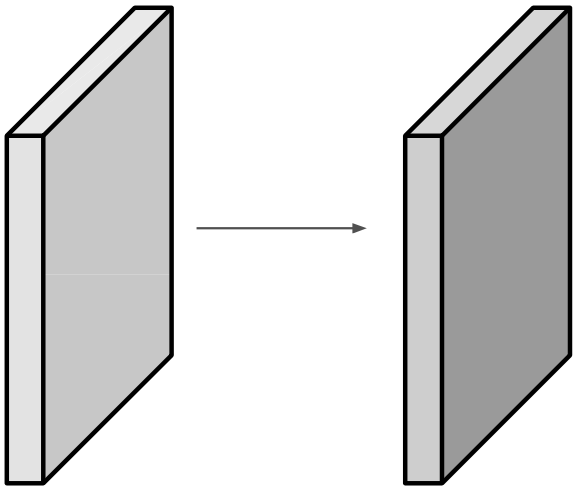
Examples time:



Input volume: **32x32x3**
10 5x5 filters with stride 1, pad 2

Number of parameters in this layer?

Examples time:

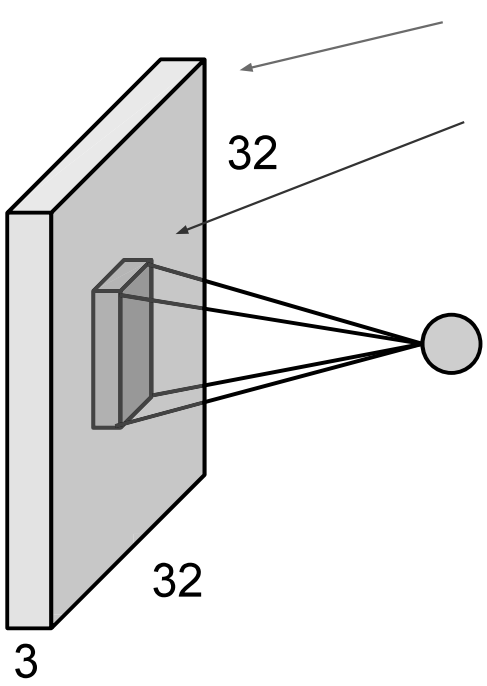


Input volume: **32x32x3**
10 5x5 filters with stride 1, pad 2

Number of parameters in this layer?
each filter has 5*5*3 + 1 = 76 params (+1 for bias)
=> 76*10 = **760**

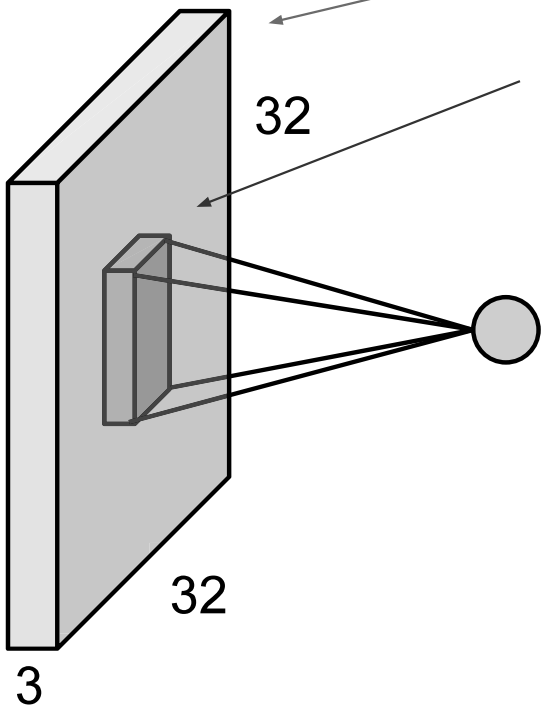
Adding more filters

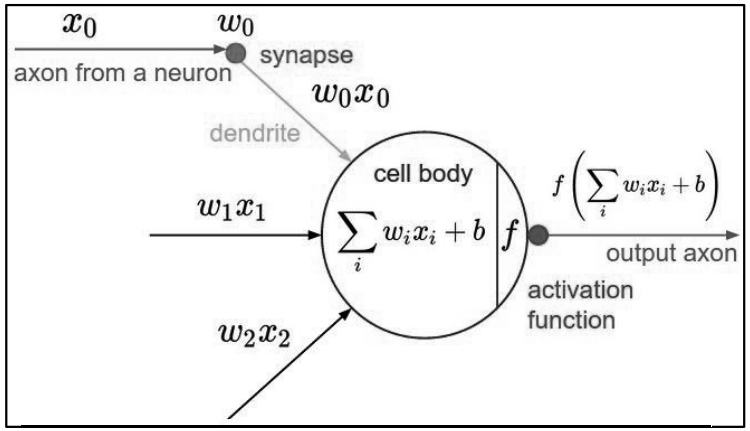
The brain/neuron view of CONV Layer



1 number:
the result of taking a dot product between the filter and this part of the image
(i.e. 5*5*3 = 75-dimensional dot product)

The brain/neuron view of CONV Layer

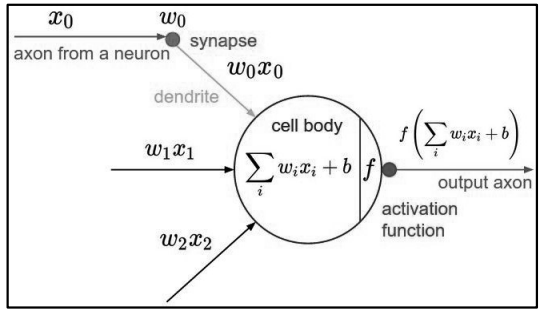


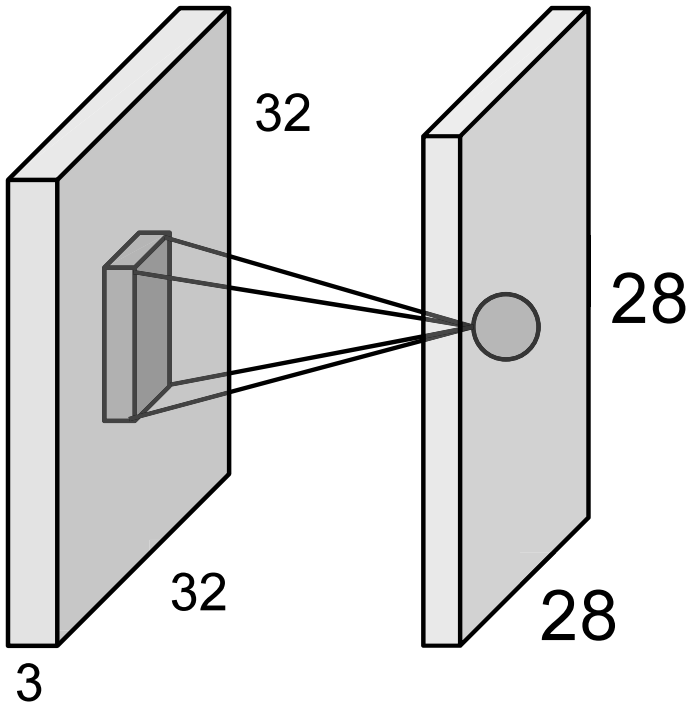


It's just a neuron with local connectivity...

1 number:
the result of taking a dot product between the filter and this part of the image
(i.e. 5*5*3 = 75-dimensional dot product)

The brain/neuron view of CONV Layer



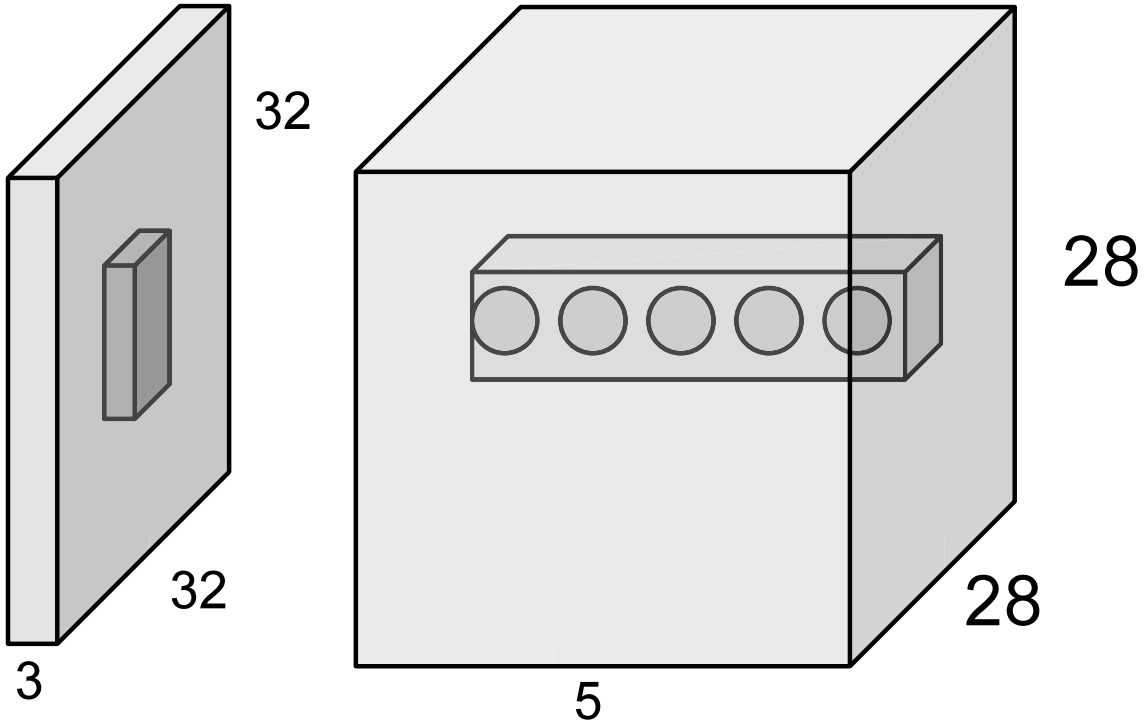


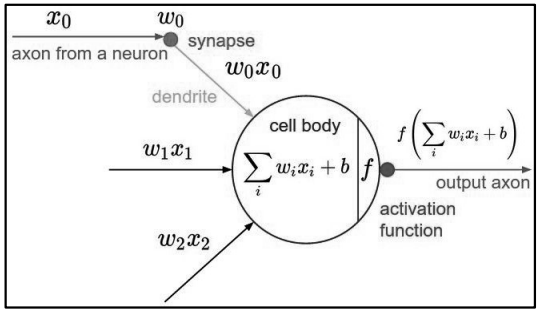
An activation map is a 28x28 sheet of neuron outputs:

1. Each is connected to a small region in the input
2. All of them share parameters

"5x5 filter" -> "5x5 receptive field for each neuron"

The brain/neuron view of CONV Layer



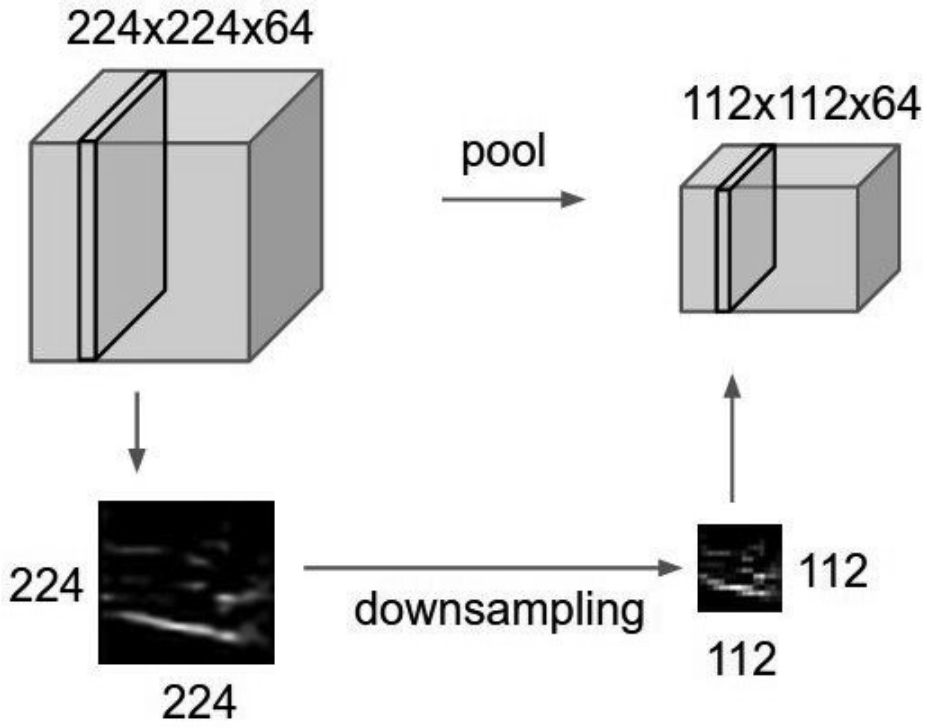


E.g. with 5 filters,
CONV layer consists of neurons arranged in a 3D grid (28x28x5)

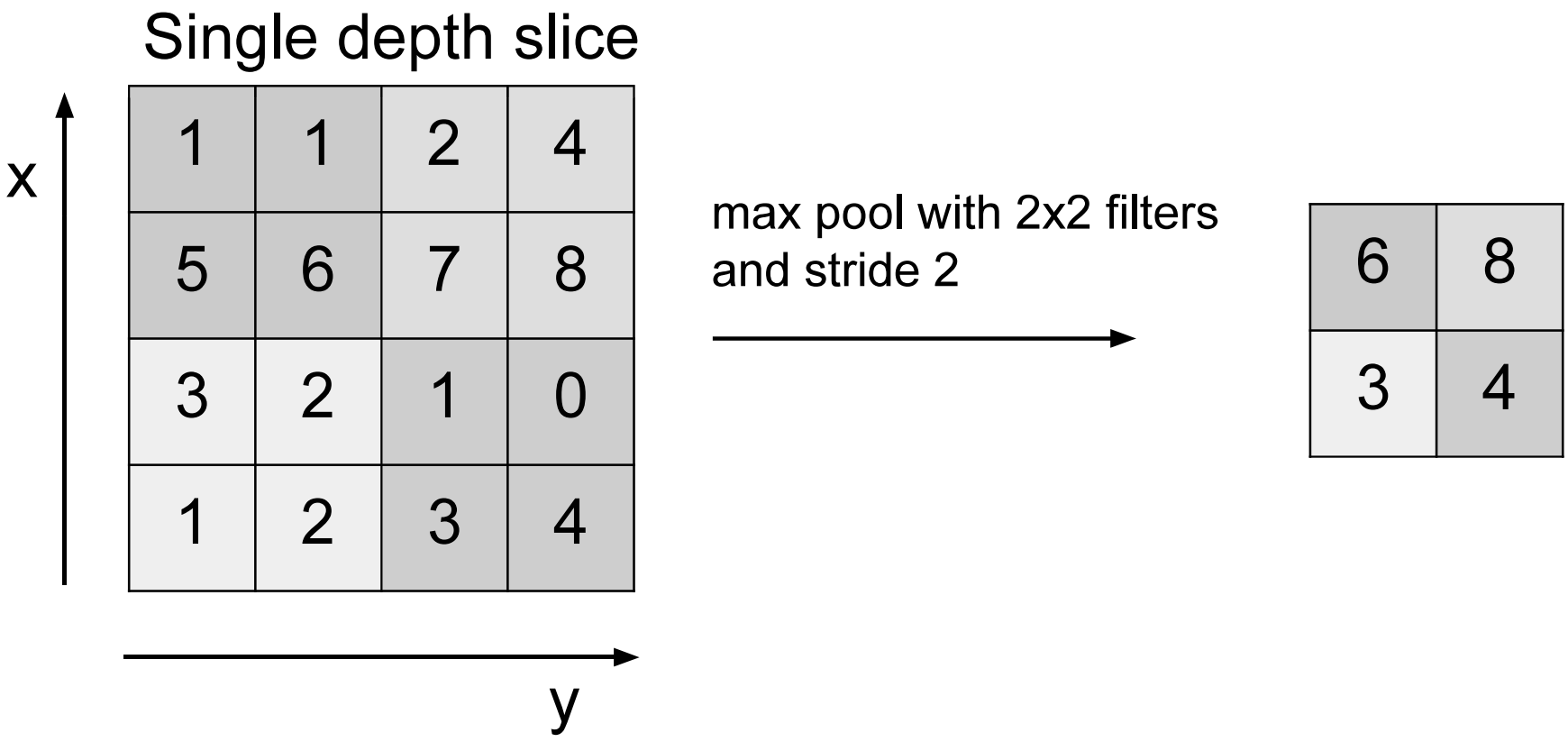
There will be 5 different neurons all looking at the same region in the input volume

Pooling layer

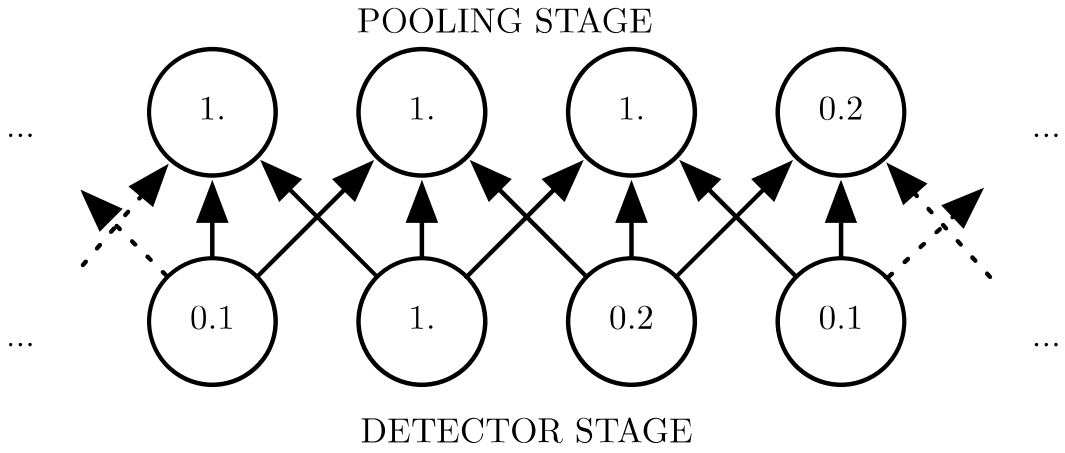
- makes the representations smaller and more manageable
- operates over each activation map independently:

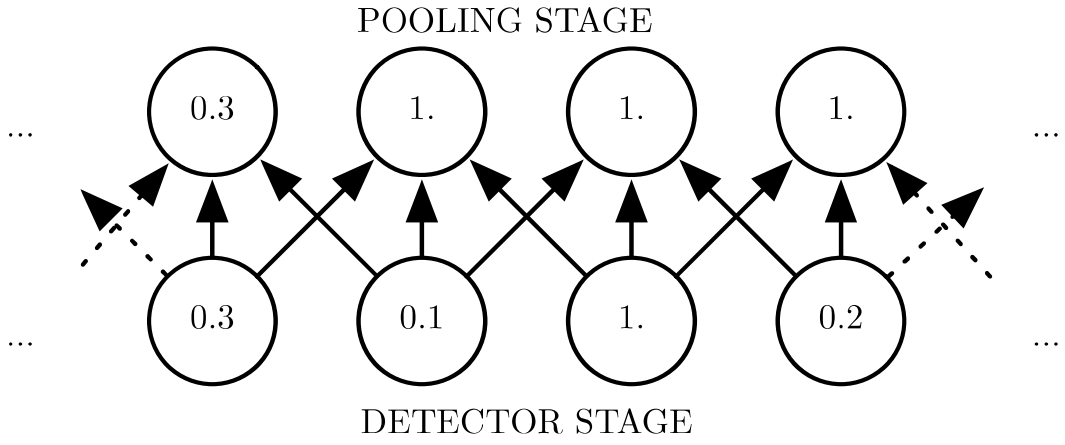


MAX POOLING


Single depth slice
x
y
1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
max pool with 2x2 filters
and stride 2
6 8
3 4

Max Pooling and Invariance to Translation





Cross-Channel Pooling and Invariance to Learned Transformations

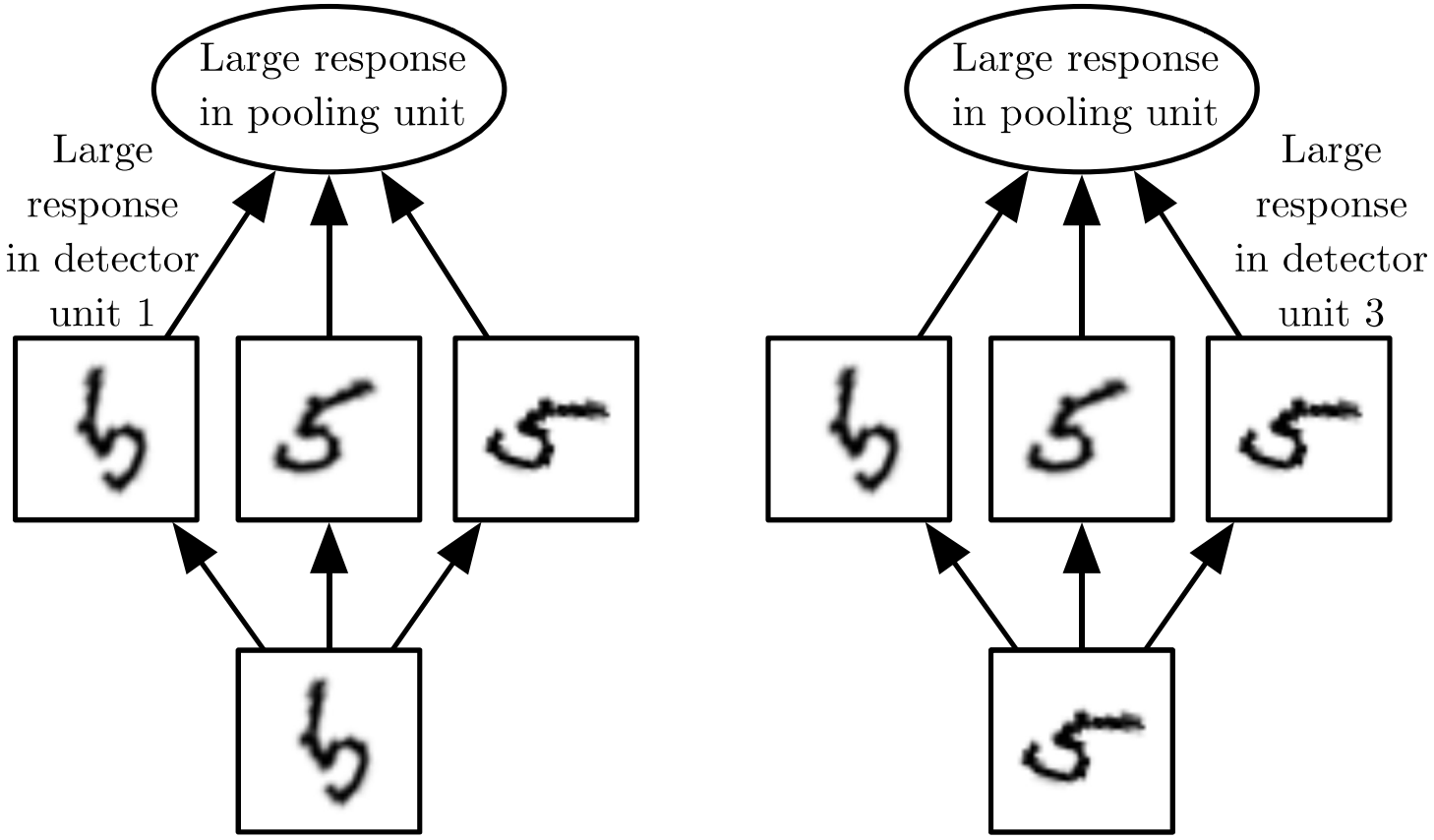


Figure 9.9

Pooling with Downsampling

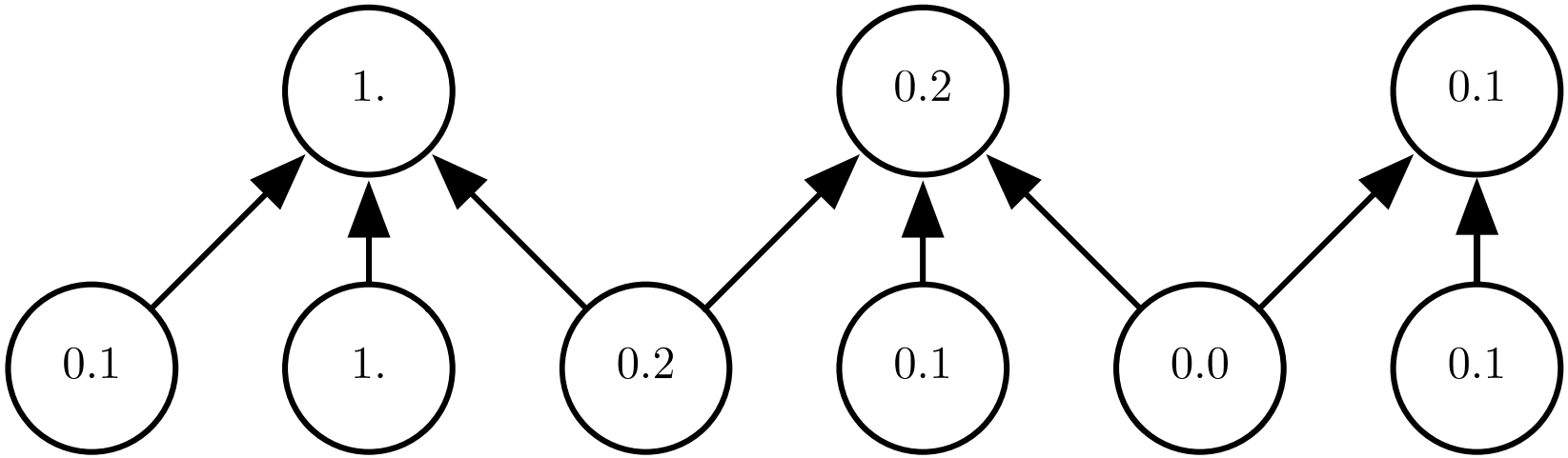


Figure 9.10

Putting things together to classify

Summary

- ConvNets stack CONV,POOL,FC layers
- Trend towards smaller filters and deeper architectures
- Trend towards getting rid of POOL/FC layers (just CONV)
- Typical architectures look like
  [(CONV-RELU)*N-POOL?]*M-(FC-RELU)*K,SOFTMAX
  where N is usually up to ~5, M is large, 0 <= K <= 2.
  - but recent advances such as ResNet/GoogLeNet challenge this paradigm