

Creation Date: 04/01/2020 21:18

Last Updated: 09/01/2020 13:33

Lec 5: Convolutional Neural Networks

Lecture 5:

Convolutional Neural Networks

Stanford

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 1

University April 18, 2017

- In CNN we **maintain spatial structure**.

A bit of history...

The **Mark I Perceptron** machine was the first implementation of the perceptron algorithm.

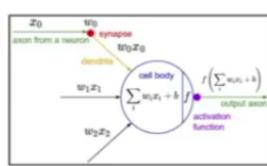
The machine was connected to a camera that used 20×20 cadmium sulfide photocells to produce a 400-pixel image.

recognized letters of the alphabet

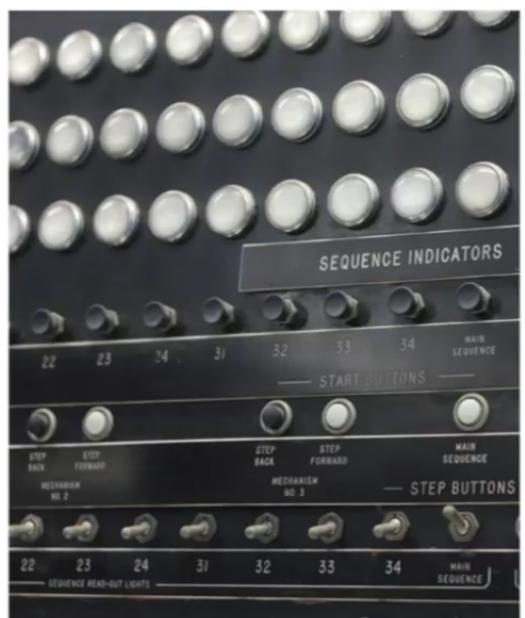
$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

update rule:

$$w_i(t+1) = w_i(t) + \alpha(d_j - y_j(t))x_{j,i},$$



Frank Rosenblatt, ~1957: Perceptron



This image by Rocky Acolat is licensed under CC BY NC

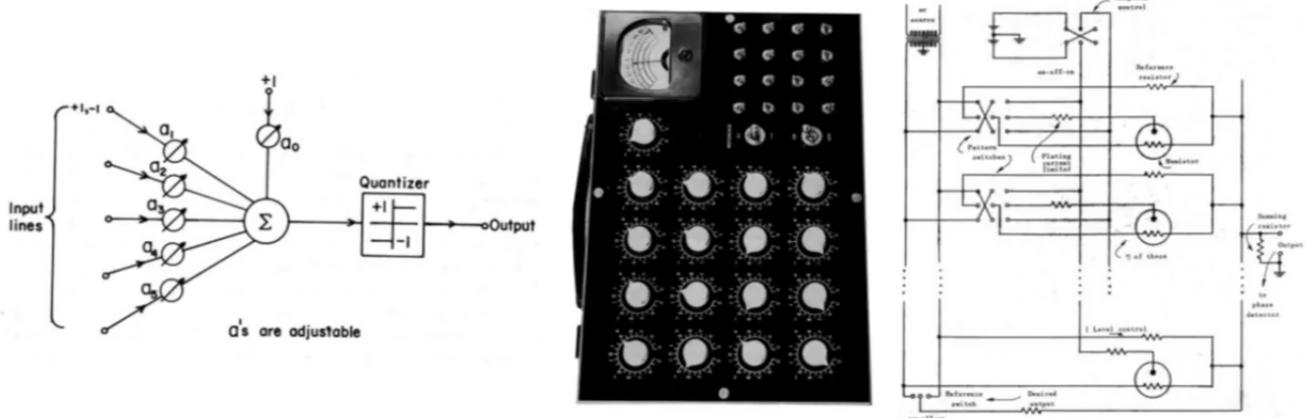
Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 5

University April 18, 2017

- During this time backprop technique was not yet created.
- We took the weights and adjusted them to work for the machine.

A bit of history...



Widrow and Hoff, ~1960: Adaline/Madaline

These figures are reproduced from [Widrow 1960, Stanford Electronics Laboratories Technical Report](#) with permission from [Stanford University Special Collections](#).

Stanford

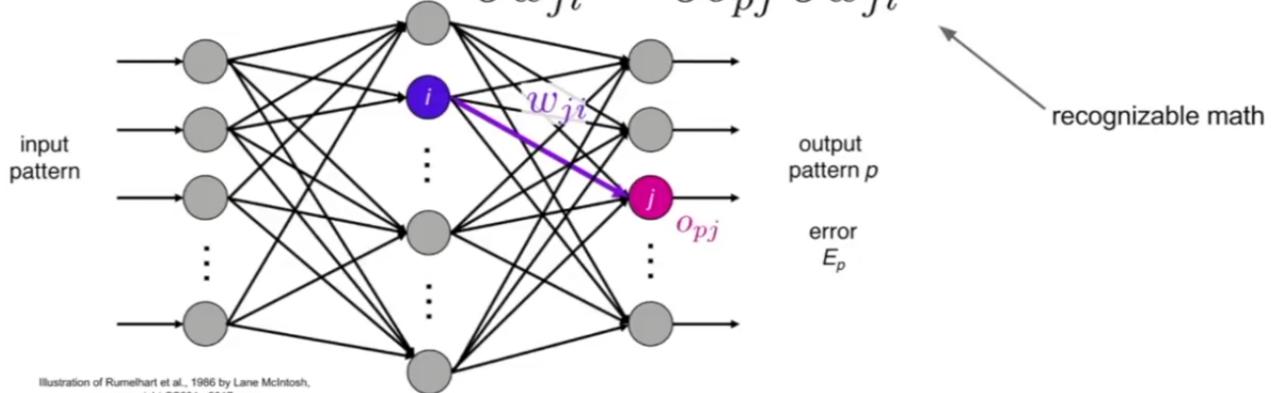
Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 6 April 18, 2017

- In Adaline/Madaline they formed multi-layer perceptrons.
- Backpropagation was introduced for the first time in 1986 by Rumelhart

A bit of history...

$$\frac{\partial E_p}{\partial w_{ji}} = \frac{\partial E_p}{\partial o_{pj}} \frac{\partial o_{pj}}{\partial w_{ji}}$$



Rumelhart et al., 1986: First time back-propagation became popular

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 7 April 18, 2017

- AI winter sets in.

- But during those time the NN was not able to scale up to deeper and larger networks.

A bit of history...

[Hinton and Salakhutdinov 2006]

Reinvigorated research in Deep Learning

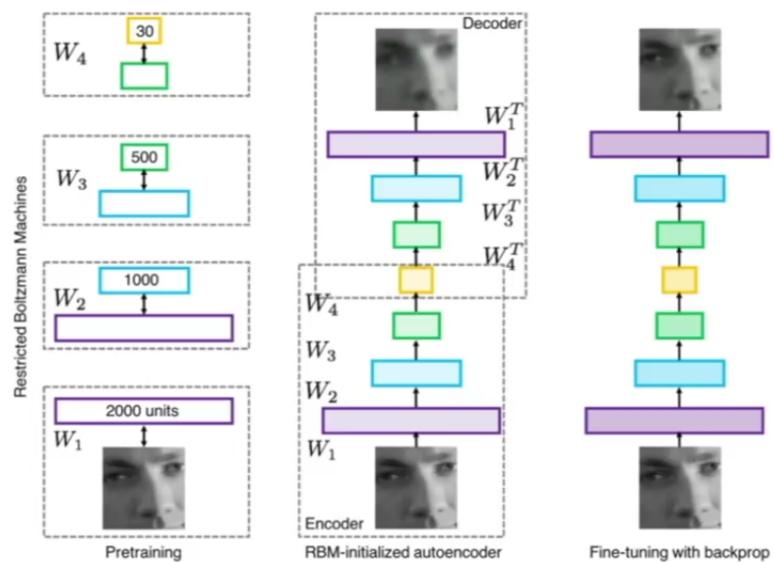


Illustration of Hinton and Salakhutdinov 2006 by Lane McIntosh, copyright CS231n 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 8 University April 18, 2017

- Here, each of the hidden layers are restricted boltzmann machines.
- After getting the result from the hidden layers, they were fed into the NN

First strong results

Acoustic Modeling using Deep Belief Networks

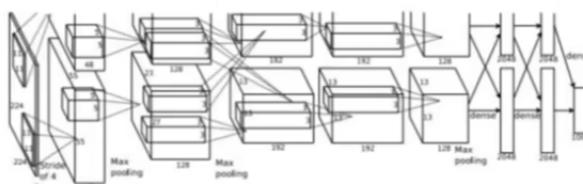
Abdel-rahman Mohamed, George Dahl, Geoffrey Hinton, 2010

Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition

George Dahl, Dong Yu, Li Deng, Alex Acero, 2012

Imagenet classification with deep convolutional neural networks

Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, 2012



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

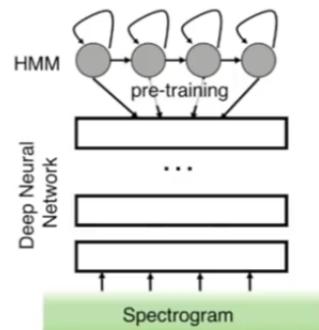
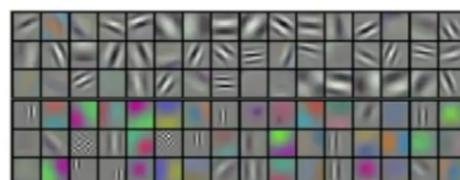


Illustration of Dahl et al. 2012 by Lane McIntosh, copyright CS231n 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 9 University April 18, 2017



Stanford

- It took almost 6 years to have a breakthrough result in NN (AlexNet)

A bit of history:

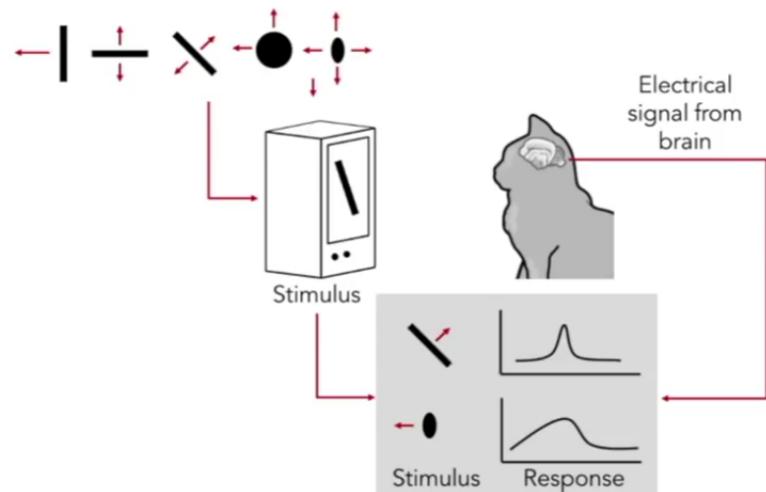
**Hubel & Wiesel,
1959**

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...



Cat image by CNX OpenStax is licensed under CC BY 4.0; changes made

Stanford

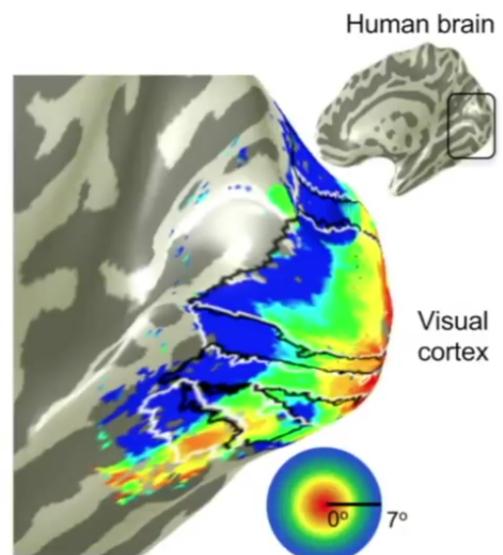
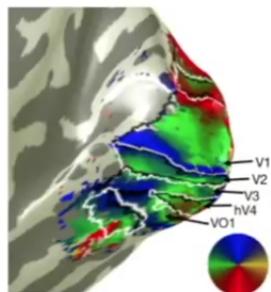
Lecture 5 - 10 University April 18, 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

- Electrodes were put inside the cat's brain and various visual stimulus were given to check the response.

A bit of history

Topographical mapping in the cortex:
nearby cells in cortex represent
nearby regions in the visual field



Retinotopy images courtesy of Jesse Gomez in the Stanford Vision & Perception Neuroscience Lab.

Stanford

Lecture 5 - 11 University April 18, 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

Hierarchical organization

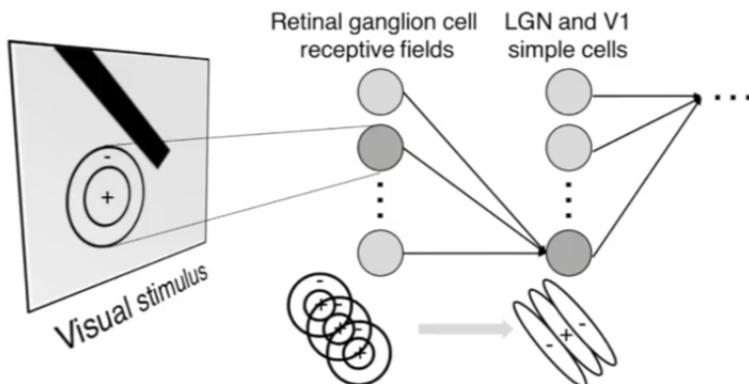


Illustration of hierarchical organization in early visual pathways by Lane McIntosh, copyright CS231n 2017

Simple cells:
Response to light orientation

Complex cells:
Response to light orientation and movement

Hypercomplex cells:
response to movement with an end point



Fei-Fei Li & Justin Johnson & Serena Yeung

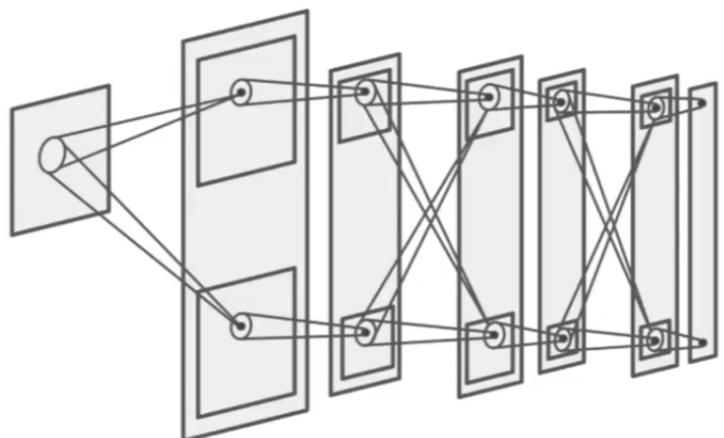
Lecture 5 - 12 University April 18, 2017

Stanford

University April 18, 2017

A bit of history:

Neurocognitron [Fukushima 1980]



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

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 13 University April 18, 2017

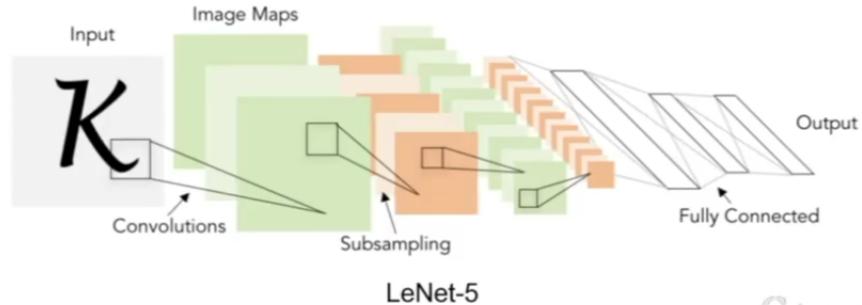
Stanford

University April 18, 2017

A bit of history:

Gradient-based learning applied to document recognition

[LeCun, Bottou, Bengio, Haffner 1998]

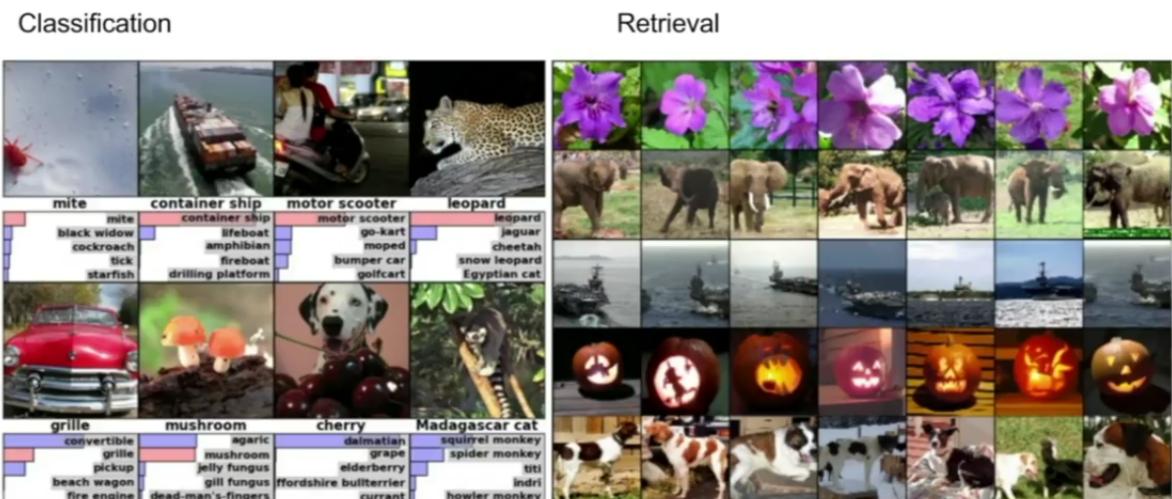


Fei-Fei Li & Justin Johnson & Serena Yeung

Stanford University April 18, 2017

- This was used in postal services to read the pincode.

Fast-forward to today: ConvNets are everywhere

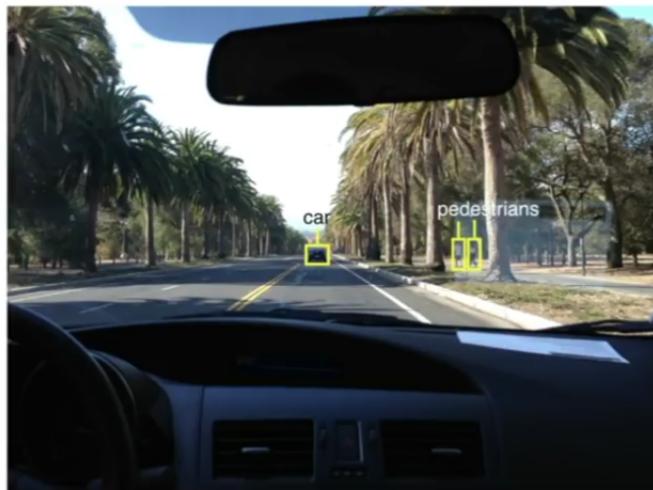


Figures copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

Fei-Fei Li & Justin Johnson & Serena Yeung

Stanford University April 18, 2017

Fast-forward to today: ConvNets are everywhere



self-driving cars

Photo by Lane McIntosh. Copyright CS231n 2017.



This image by GBPublic_PR is licensed under CC-BY 2.0

NVIDIA Tesla line

(these are the GPUs on rye01.stanford.edu)

Note that for embedded systems a typical setup would involve NVIDIA Tegras, with integrated GPU and ARM-based CPU cores.

Stanford

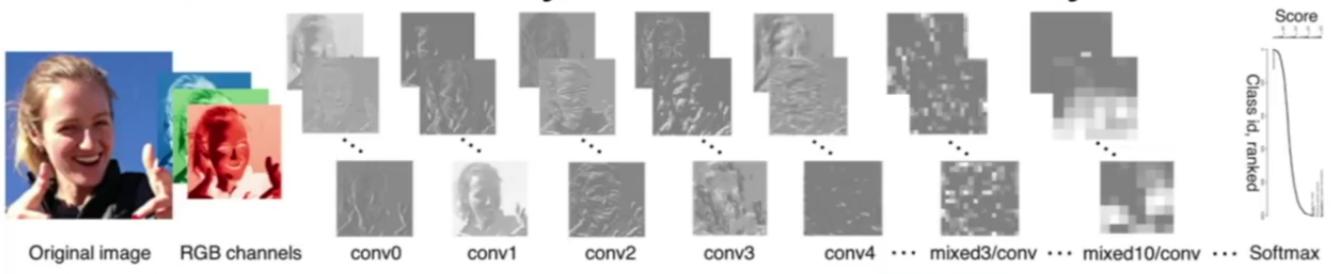
University April 18, 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

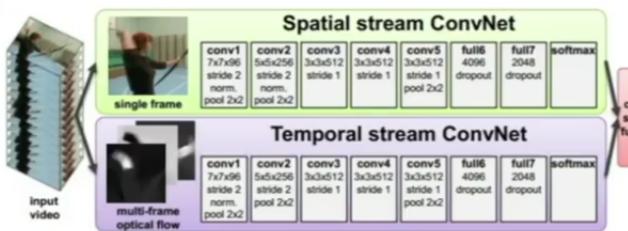
Lecture 5 - 18

April 18, 2017

Fast-forward to today: ConvNets are everywhere



Activations of [inception-v3 architecture](#) [Szegedy et al. 2015] to image of Emma McIntosh, used with permission. Figure and architecture not from Taigman et al. 2014.



Figures copyright Simonyan et al., 2014.
Reproduced with permission.

Illustration by Lane McIntosh,
photos of Katie Cumnock
used with permission.

Stanford University April 18, 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 19

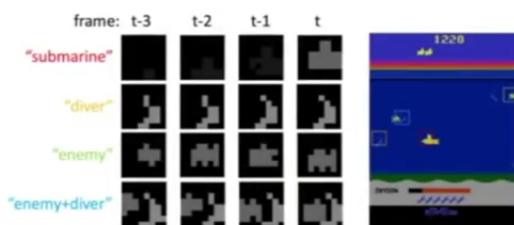
April 18, 2017

Fast-forward to today: ConvNets are everywhere

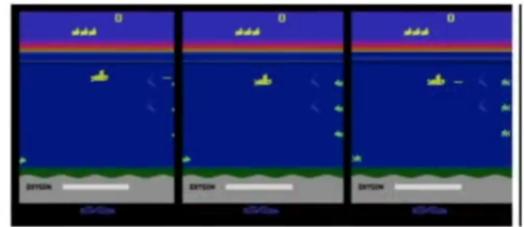


[Toshev, Szegedy 2014]

Images are examples of pose estimation, not actually from Toshev & Szegedy 2014. Copyright Lane McIntosh.



[Guo et al. 2014]

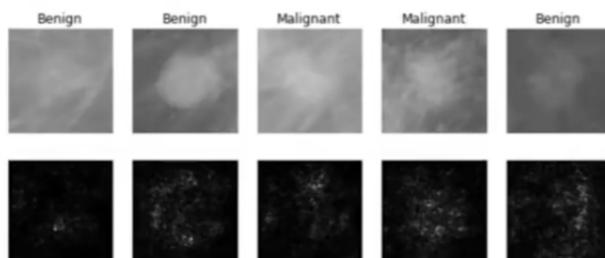


Figures copyright Xiaoxiao Guo, Salinifer Singh, Hon-Jak Lee, Richard Lewis, and Xiaoshi Wang, 2014. Reprinted with permission.

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 20 University April 18, 2017

Fast-forward to today: ConvNets are everywhere



[Levy et al. 2016]

Figure copyright Levy et al. 2016.
Reproduced with permission.



[Dieleman et al. 2014]

From left to right: public domain by NASA, usage permitted by
ESA/Hubble, public domain by NASA, and public domain.



[Sermanet et al. 2011]
[Ciresan et al.]

Stanford University

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 21 University April 18, 2017

This image by Christin Khan is in the public domain and originally came from the U.S. NOAA.



Whale recognition, Kaggle Challenge

Photo and figure by Lane McIntosh; not actual example from Mnih and Hinton, 2010 paper.



Mnih and Hinton, 2010

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 22 April 18, 2017

Stanford

University

Image Captioning

[Vinyals et al., 2015]
[Karpathy and Fei-Fei, 2015]



A white teddy bear sitting in the grass



A man in a baseball uniform throwing a ball



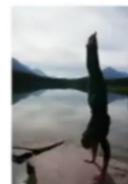
A woman is holding a cat in her hand



A man riding a wave on top of a surfboard



A cat sitting on a suitcase on the floor



A woman standing on a beach holding a surfboard

All images are CC0 Public domain:
<https://pixabay.com/en/luggage-antique-cat-1643010/>
<https://pixabay.com/en/teddy-plush-bear-cute-teddy-bear-1623436/>
<https://pixabay.com/en/surf-wave-summer-sport-librait-1668716/>
<https://pixabay.com/en/woman-female-model-portrait-adult-983967/>
<https://pixabay.com/en/handstand-lake-meditation-196008/>
<https://pixabay.com/en/baseball-player-shortstop-infield-1045263/>

Captions generated by Justin Johnson, Laing Neuraltalk

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 23 April 18, 2017

Stanford

University



Figures copyright Justin Johnson, 2015. Reproduced with permission. Generated using the Inceptionism approach from a [blog post](#) by Google Research.

Original image is CC0 public domain
Starry Night and Tree Roots by Van Gogh are in the public domain
Bokeh Image is in the public domain
Stylized images copyright Justin Johnson, 2017; reproduced with permission

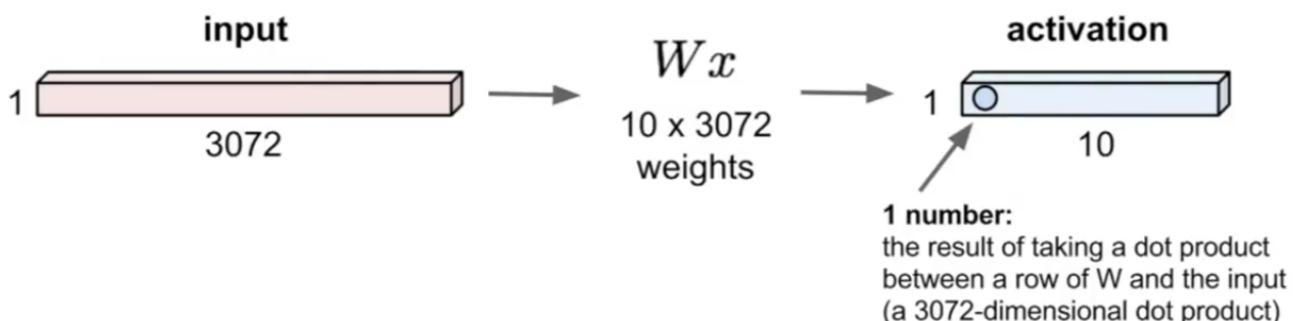
Gatys et al, "Image Style Transfer using Convolutional Neural Networks", CVPR 2016
Gatys et al, "Controlling Perceptual Flow in Image Style Transfer", CVPR 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 24 University April 18, 2017

Fully Connected Layer

32x32x3 image \rightarrow stretch to 3072 x 1



Fei-Fei Li & Justin Johnson & Serena Yeung

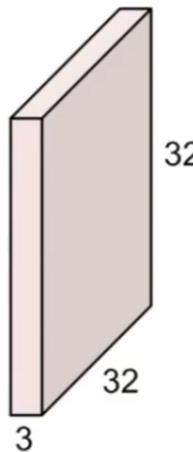
Stanford University April 18, 2017

Lecture 5 - 27

- Fully Connected layer doesn't preserve spatial structure since we just flatten the image into a vector.

Convolution Layer

32x32x3 image



5x5x3 filter



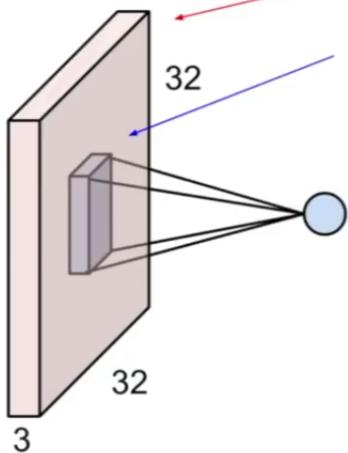
Filters always extend the full depth of the input volume

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

Convolution Layer

32x32x3 image

5x5x3 filter w

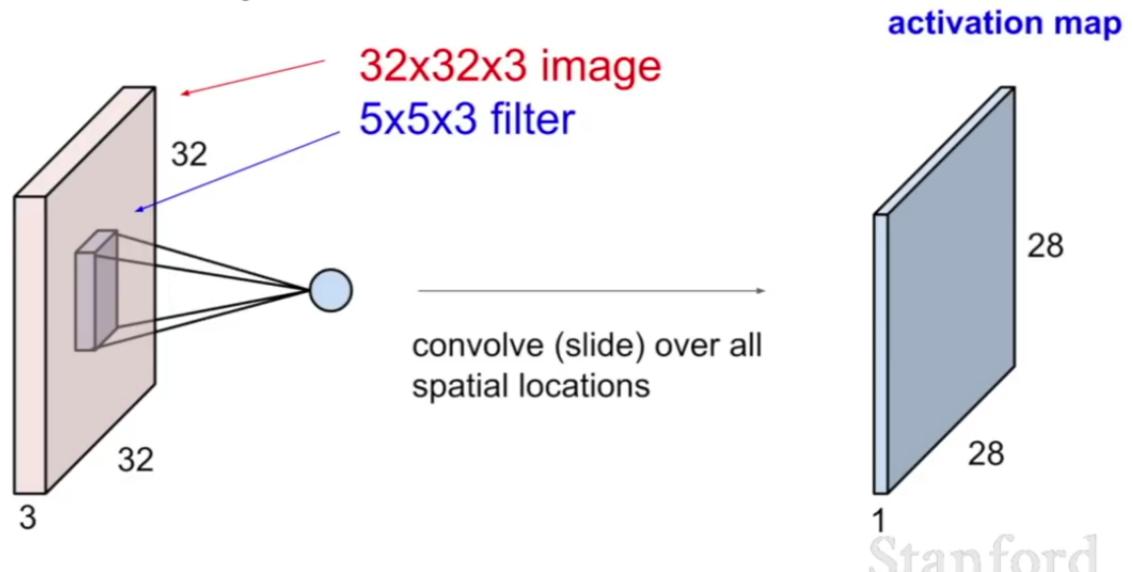


1 number:

the result of taking a dot product between the filter and a small 5x5x3 chunk of the image
(i.e. $5 \times 5 \times 3 = 75$ -dimensional dot product + bias)

$$w^T x + b$$

Convolution Layer

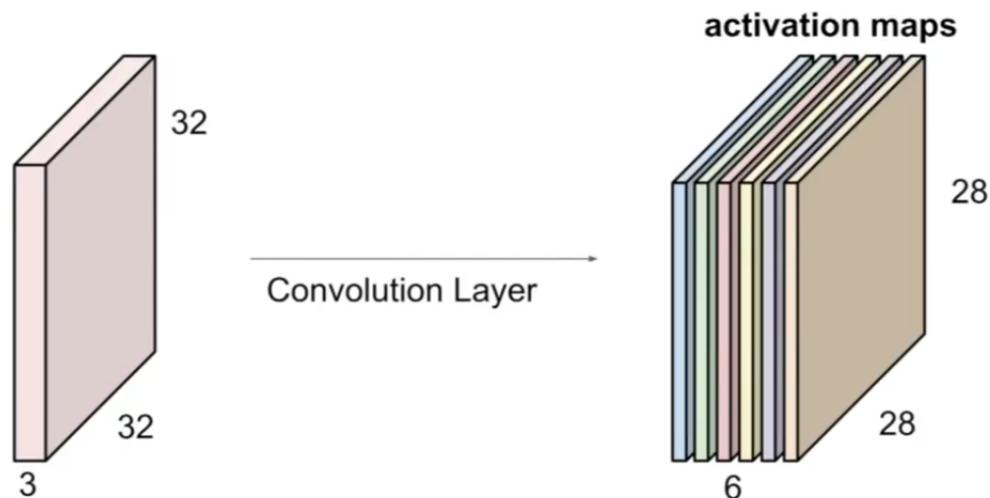


Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 32

Stanford University April 18, 2017

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



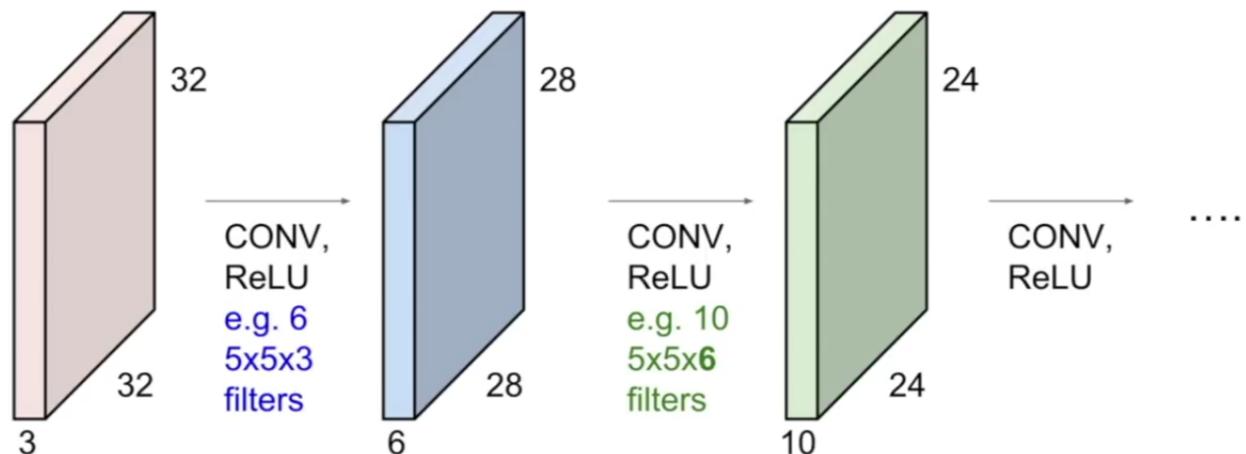
We stack these up to get a “new image” of size 28x28x6!

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 34

Stanford University April 18, 2017

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



Fei-Fei Li & Justin Johnson & Serena Yeung

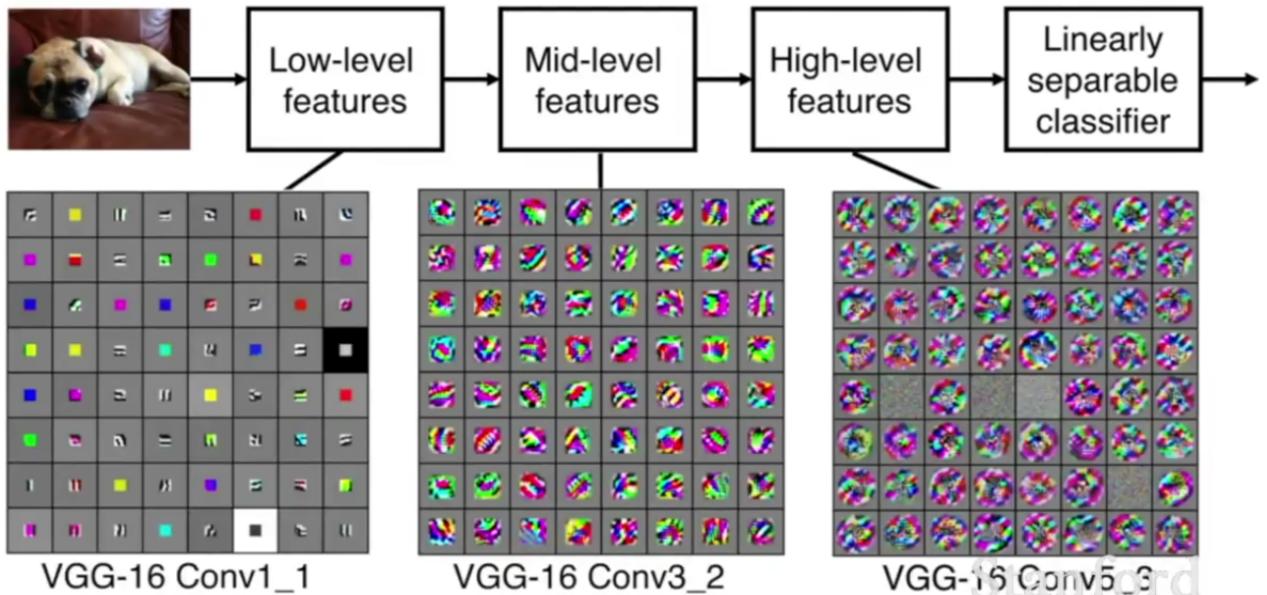
Lecture 5 - 36

Stanford University April 18, 2017

Preview

[Zeiler and Fergus 2013]

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



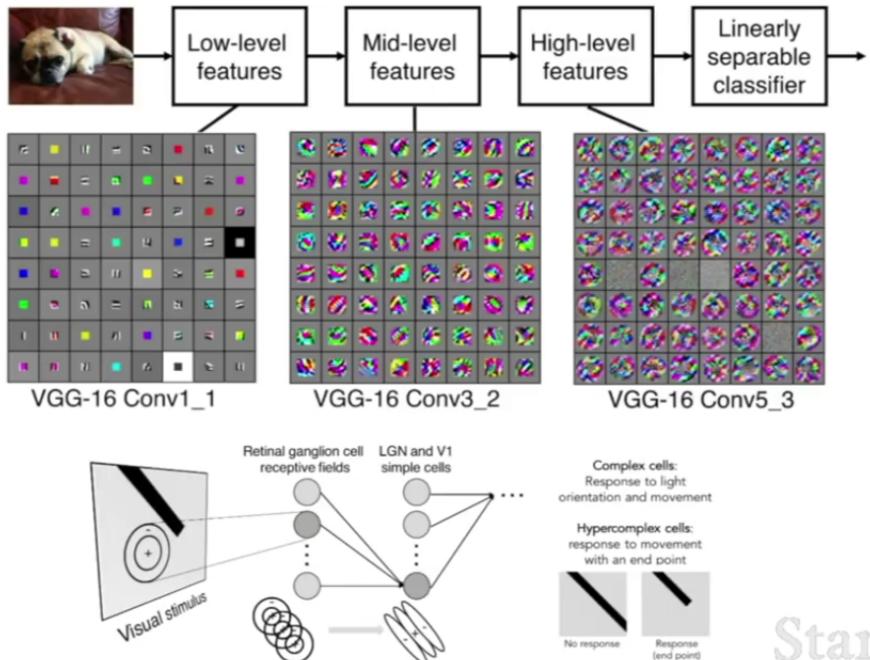
Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 37

Stanford University April 18, 2017

- So the activation maps when stacked form layers, and the layer at the beginning learn simple, low-level features whereas deeper layers learn more complex features.

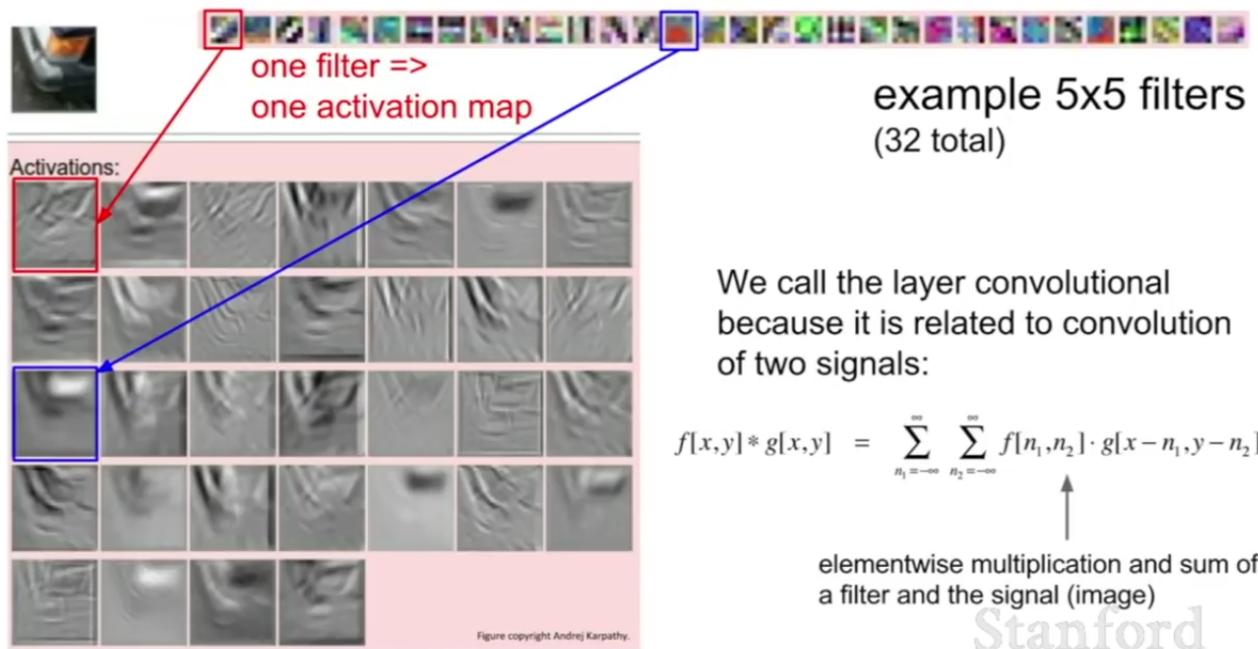
Preview



Fei-Fei Li & Justin Johnson & Serena Yeung

Stanford University April 18, 2017
Lecture 5 - 38

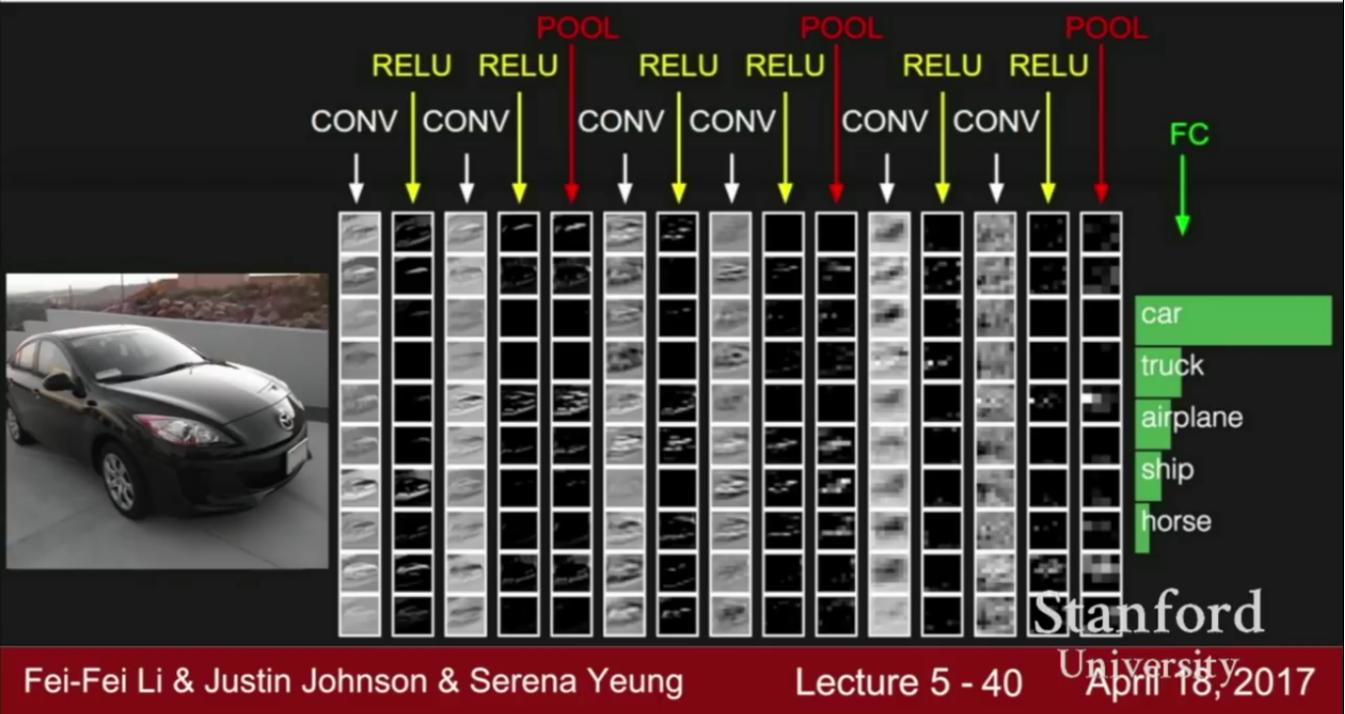
- We can notice that the NN mimics the earlier experiment.



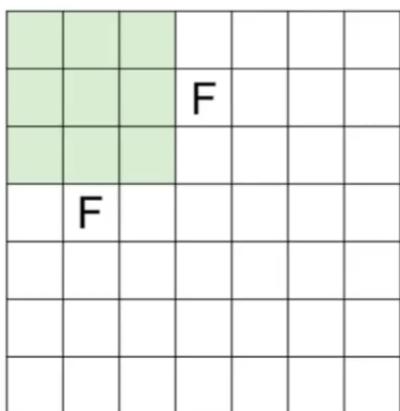
Fei-Fei Li & Justin Johnson & Serena Yeung

Stanford University April 18, 2017
Lecture 5 - 39

preview:



N



Output size:
 $(N - F) / \text{stride} + 1$

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

Fei-Fei Li & Justin Johnson & Serena Yeung

Stanford University
Lecture 5 - 52 April 18, 2017

In practice: Common to zero pad the border

0	0	0	0	0	0		
0							
0							
0							
0							

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. $F = 3 \Rightarrow$ zero pad with 1

$F = 5 \Rightarrow$ zero pad with 2

$F = 7 \Rightarrow$ zero pad with 3

Stanford

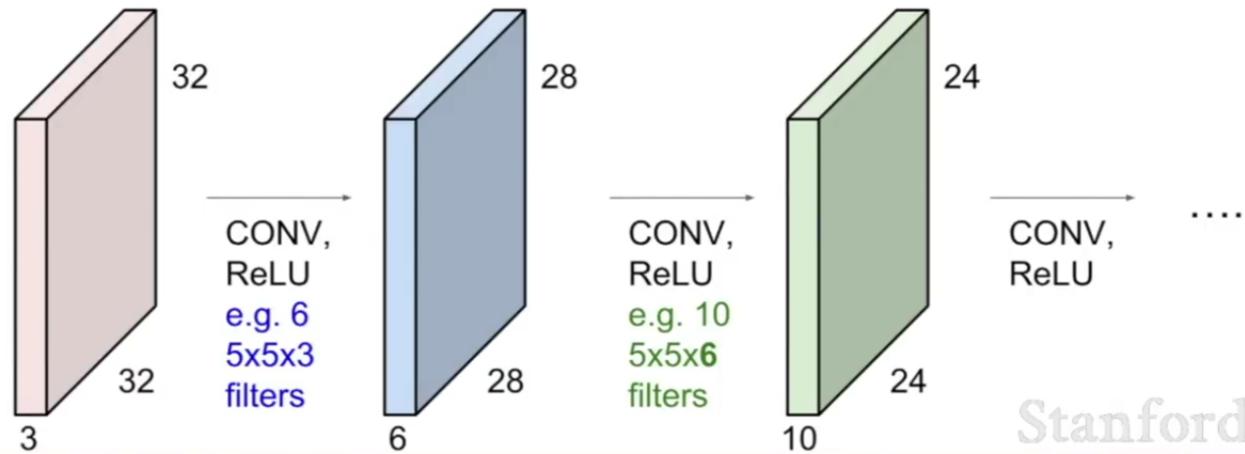
University April 18, 2017

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 55

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.



Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 56

Stanford

University April 18, 2017

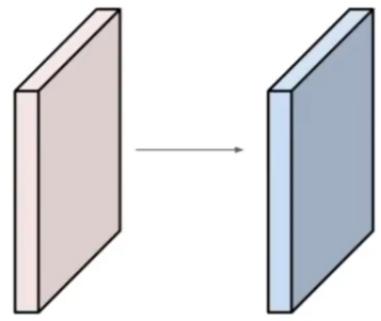
- Shrinking too fast is not good, doesn't work well.
- Hence padding is used, and max-pool is done after some depths.

- IMPORTANT

Examples time:

Input volume: **32x32x3**

10 5x5 filters with stride 1, pad 2



Number of parameters in this layer?

each filter has $5 \times 5 \times 3 + 1 = 76$ params (+1 for bias)

$$\Rightarrow 76 \times 10 = 760$$

Stanford

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 60

University April 18, 2017

Summary. To summarize, the Conv Layer:

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires four hyperparameters:
 - Number of filters K ,
 - their spatial extent F ,
 - the stride S ,
 - the amount of zero padding P .
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
 - $W_2 = (W_1 - F + 2P)/S + 1$
 - $H_2 = (H_1 - F + 2P)/S + 1$ (i.e. width and height are computed equally by symmetry)
 - $D_2 = K$
- With parameter sharing, it introduces $F \cdot F \cdot D_1$ weights per filter, for a total of $(F \cdot F \cdot D_1) \cdot K$ weights and K biases.
- In the output volume, the d -th depth slice (of size $W_2 \times H_2$) is the result of performing a valid convolution of the d -th filter over the input volume with a stride of S , and then offset by d -th bias.

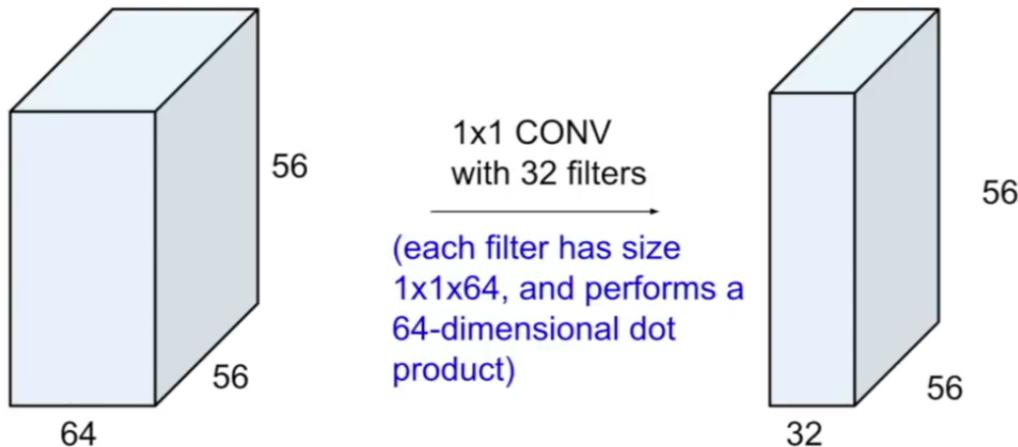
Stanford

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 61

University April 18, 2017

(btw, 1x1 convolution layers make perfect sense)



Fei-Fei Li & Justin Johnson & Serena Yeung

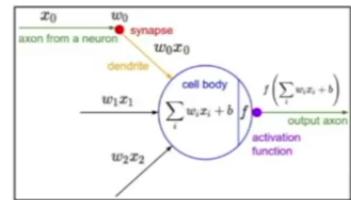
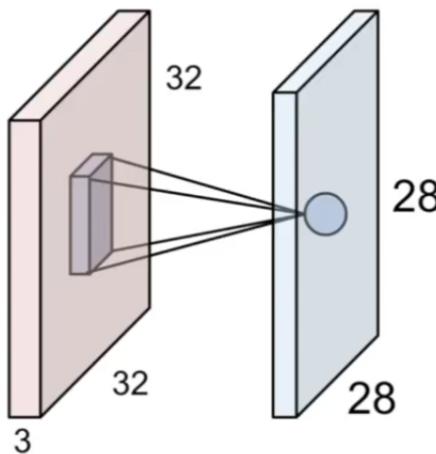
Lecture 5 - 63

Stanford

University April 18, 2017

- Receptive Field:

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”

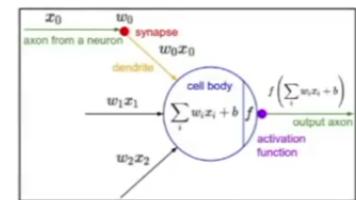
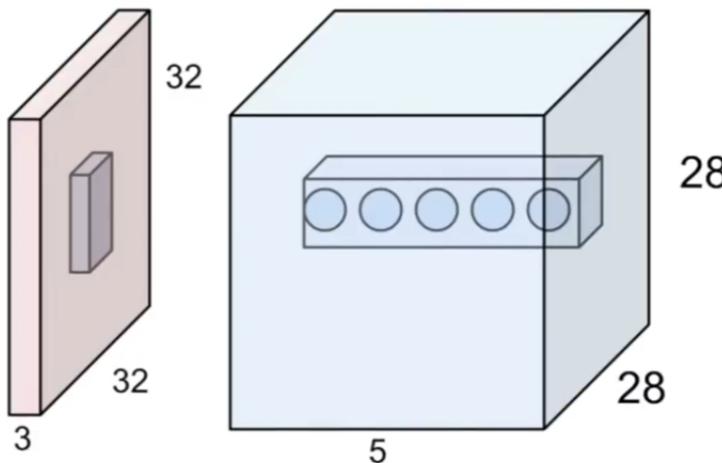
Stanford

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 68

University April 18, 2017

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

Stanford

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 69

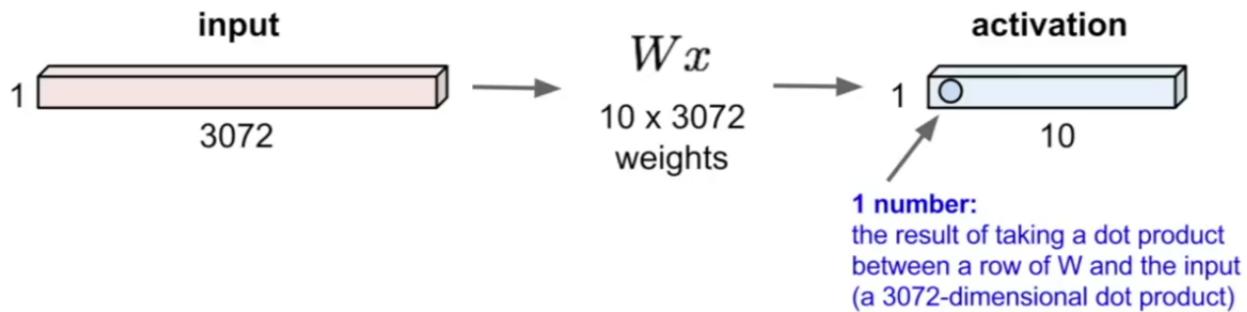
University April 18, 2017

- Here we can notice that in the activation map, different filters were applied to the same spatial location.

Reminder: Fully Connected Layer

32x32x3 image -> stretch to 3072 x 1

Each neuron looks at the full input volume



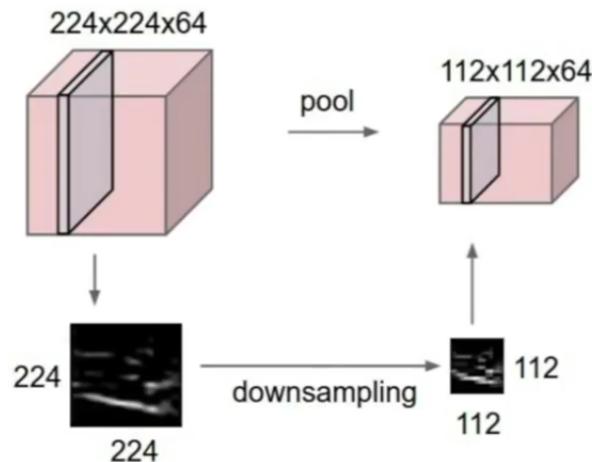
Fei-Fei Li & Justin Johnson & Serena Yeung

Stanford University April 18, 2017

Lecture 5 - 70

Pooling layer

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



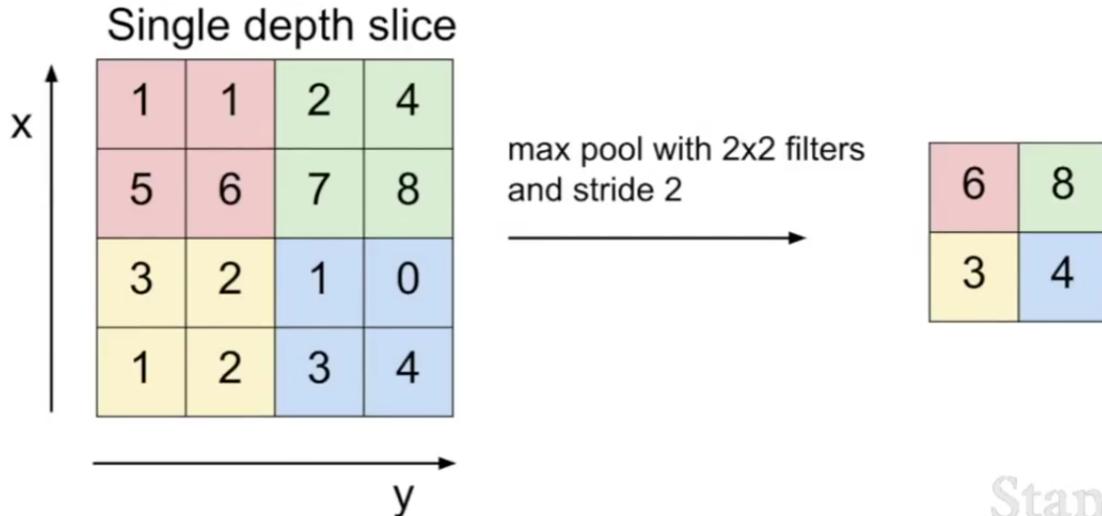
Fei-Fei Li & Justin Johnson & Serena Yeung

Stanford University April 18, 2017

Lecture 5 - 72

- Pooling doesn't change the depth, it only changes the spatial dimensions.

MAX POOLING



Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 73 Stanford University April 18, 2017

Stanford

University
April 18, 2017

- While pooling, we usually avoid overlapping with enough large stride.
- **Max Pooling is done because it mimics the notion whether any neuron fired in the considered region.**

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires three hyperparameters:
 - their spatial extent F ,
 - the stride S ,
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
 - $W_2 = (W_1 - F)/S + 1$
 - $H_2 = (H_1 - F)/S + 1$
 - $D_2 = D_1$
- Introduces zero parameters since it computes a fixed function of the input
- Note that it is not common to use zero-padding for Pooling layers

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 74 Stanford University April 18, 2017

Stanford

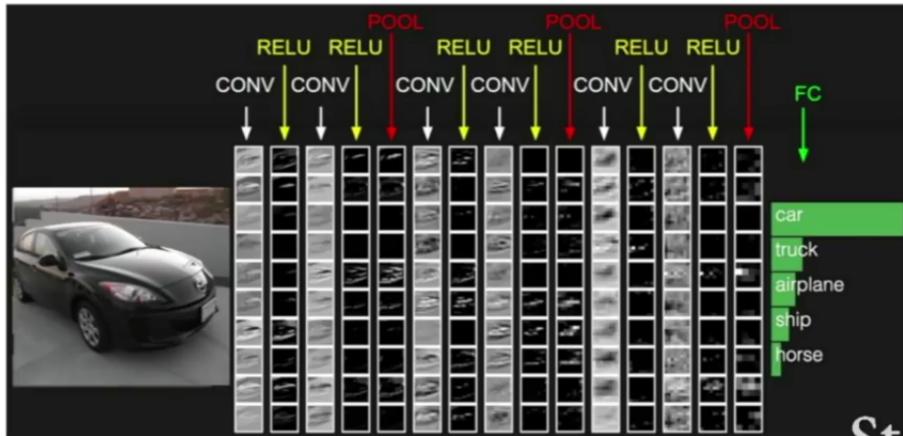
University
April 18, 2017

- Pooling layers have 0 parameters.

- The depth remains same after pooling since the pooling is done on each layer separately.

Fully Connected Layer (FC layer)

- Contains neurons that connect to the entire input volume, as in ordinary Neural Networks



Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 76 University April 18, 2017

- For FULLY CONNECTED LAYERS, we flatten the convolutional output of size (w,h,d) into a vector of size $(wxhxd) \times 1$ vector and then connect each of the neurons to the FC layer.
- Even though the final layers will have less number of neurons the information present in them will be the accumulated information passed through the entire network.*

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, $0 \leq K \leq 2$.
 - but recent advances such as ResNet/GoogLeNet challenge this paradigm

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 5 - 78 University April 18, 2017