Convolutional Neural Networks

Convolution

- Convolution = Spatial filtering

$$(a \star b)[i,j] = \sum_{i',j'} a[i',j']b[i-i',j-j']$$

- Different filters (weights) reveal a different characteristics of the input.

$\ast \frac{1}{8}$

0	1	0
1	4	1
0	1	0

Convolution

- Convolution = Spatial filtering

$$(a \star b)[i,j] = \sum_{i',j'} a[i',j']b[i-i',j-j']$$

- Different filters (weights) reveal a different characteristics of the input.

$*$

0	-1	0
-1	4	-1
0	-1	0

Convolution

- Convolution = Spatial filtering

$$(a \star b)[i,j] = \sum_{i',j'} a[i',j']b[i-i',j-j']$$

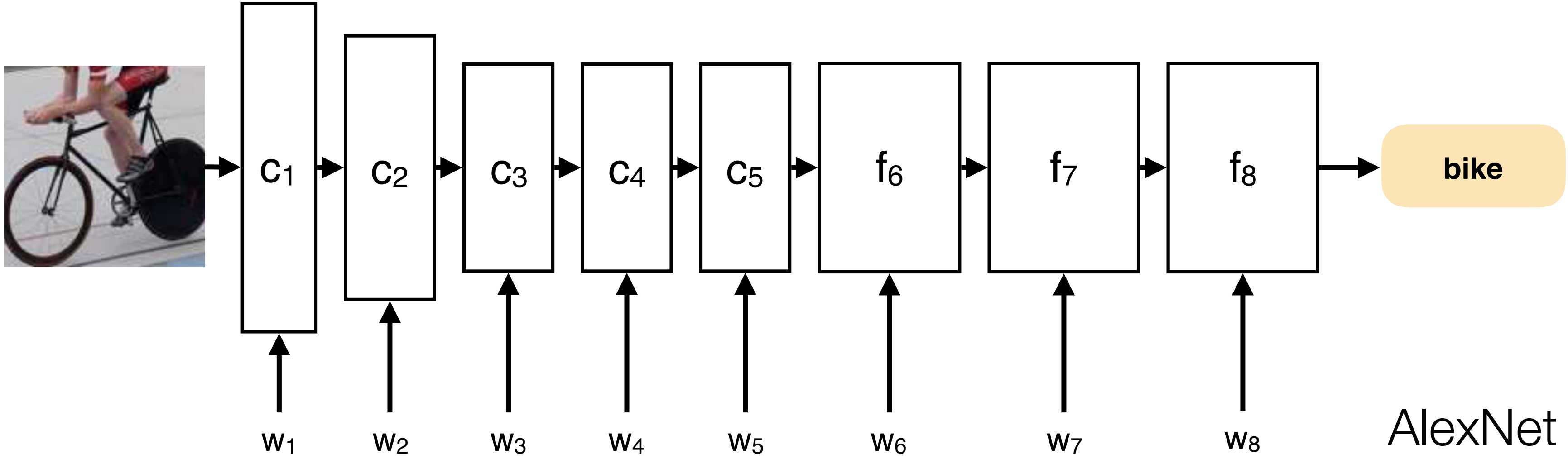
- Different filters (weights) reveal a different characteristics of the input.

$*$

1	0	-1
2	0	-2
1	0	-1

CNNs - A review

- A neural network model that consists of a sequence of local & translation invariant layers
  - Many identical copies of the same neuron: Weight/parameter sharing
  - Hierarchical feature learning

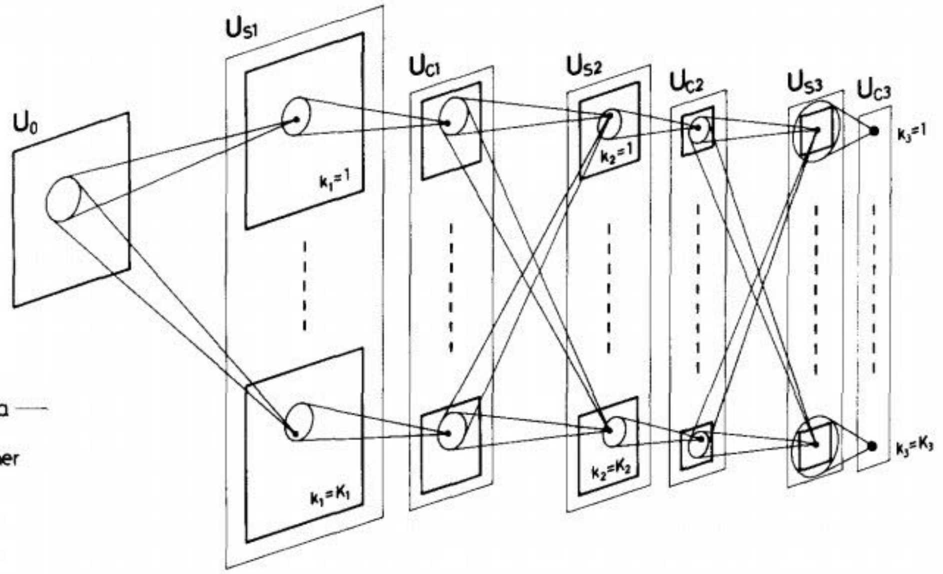


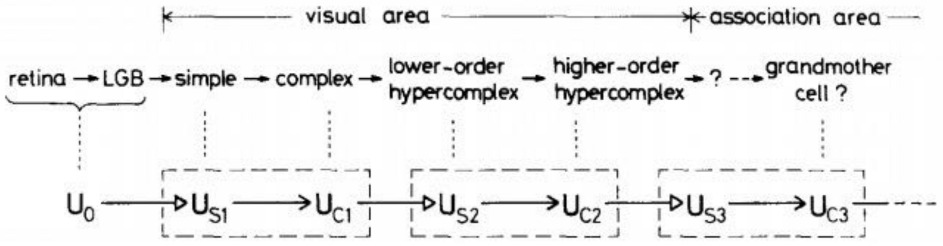
AlexNet

Image credit: Andrea Vedaldi

CNNs - A bit of history

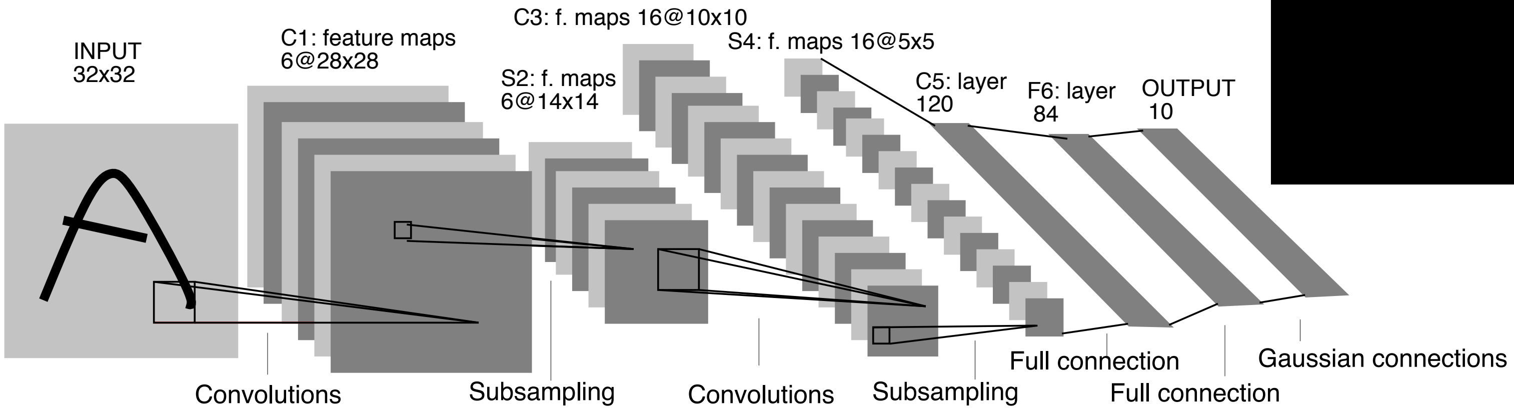
- Neurocognitron model by Fukushima (1980)
- The first convolutional neural network (CNN) model
- so-called “sandwich” architecture
  - simple cells act like filters
  - complex cells perform pooling
- Difficult to train
  - No backpropagation yet





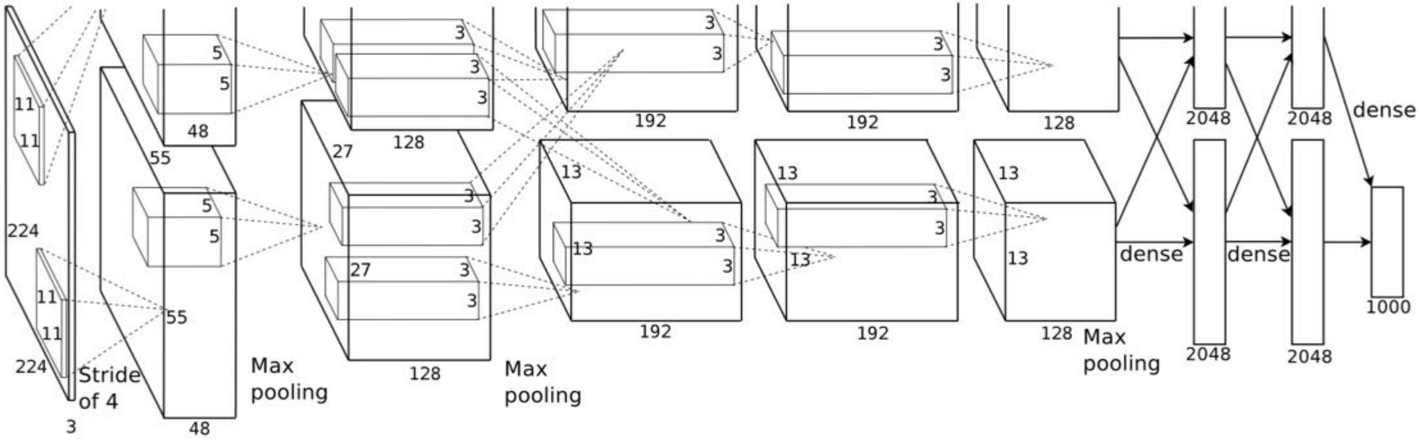
CNNs - A bit of history

- Gradient-based learning applied to document recognition [LeCun, Bottou, Bengio, Haffner 1998]
- LeNet-5 model



CNNs - A bit of history

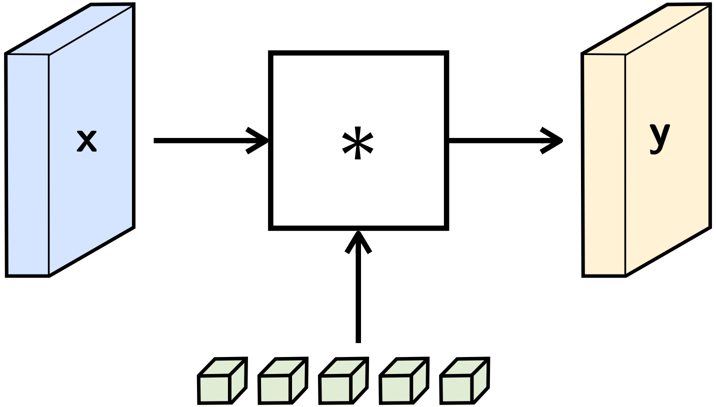
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. *Imagenet classification with deep convolutional neural networks*. In Proc. NIPS, 2012.
- AlexNet model



Convolutional layer

- Learn a filter bank (a set of filters) once
- Use them over the input data to extract features

$$\boxed{\mathbf{y} = F * \mathbf{x} + b}$$

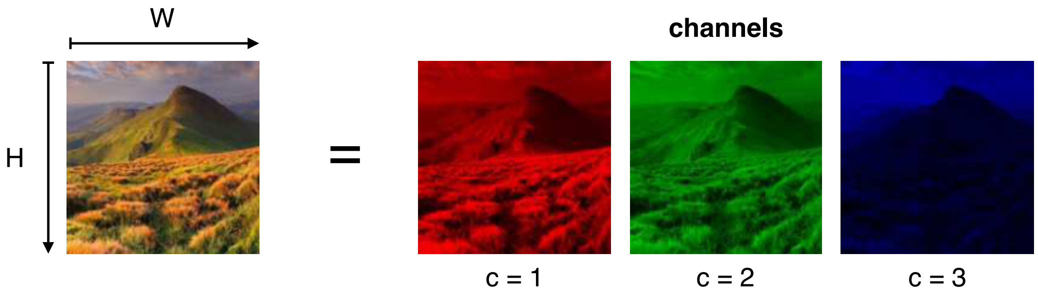


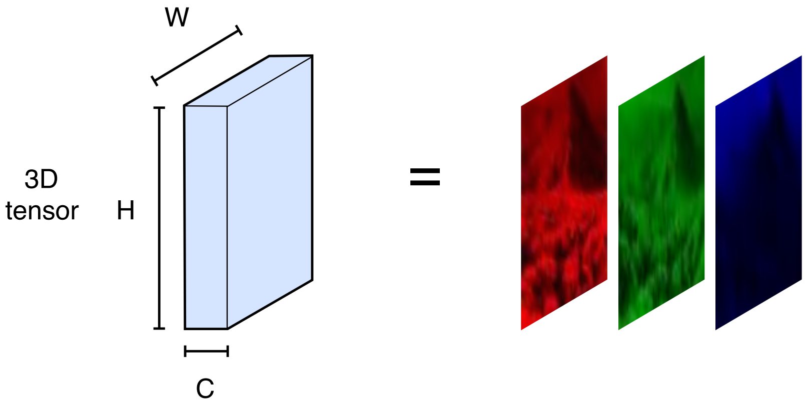
input data **x** filter bank *F* output data **y**

Image credit: Andrea Vedaldi

Data = 3D Tensor

- There is a vector of feature channels (e.g. RGB) at each spatial location (pixel).



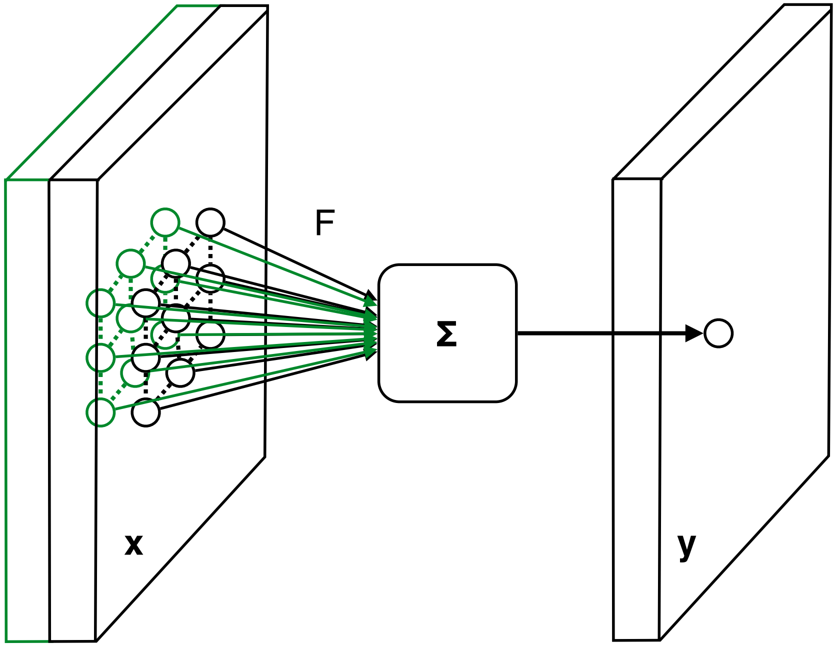


Convolution with 3D filters

- Each filter acts on multiple input channels

Local
Filters look locally

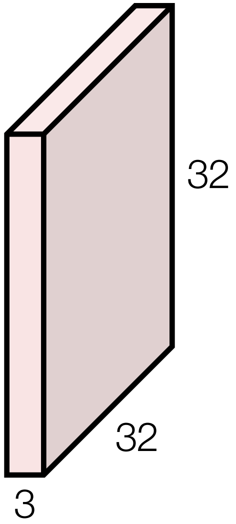
Translation invariant
Filters act the same everywhere



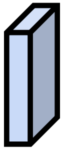
Slide credit: Andrea Vedaldi

Convolutional Layer

32x32x3 input

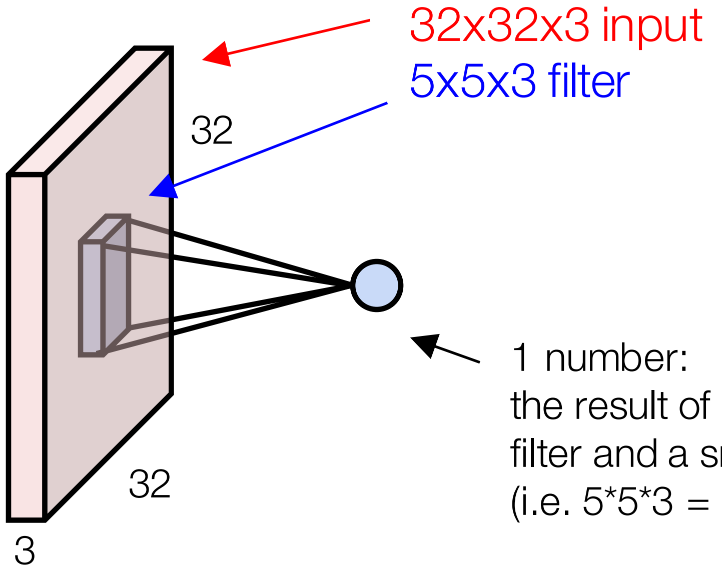


5x5x3 filter



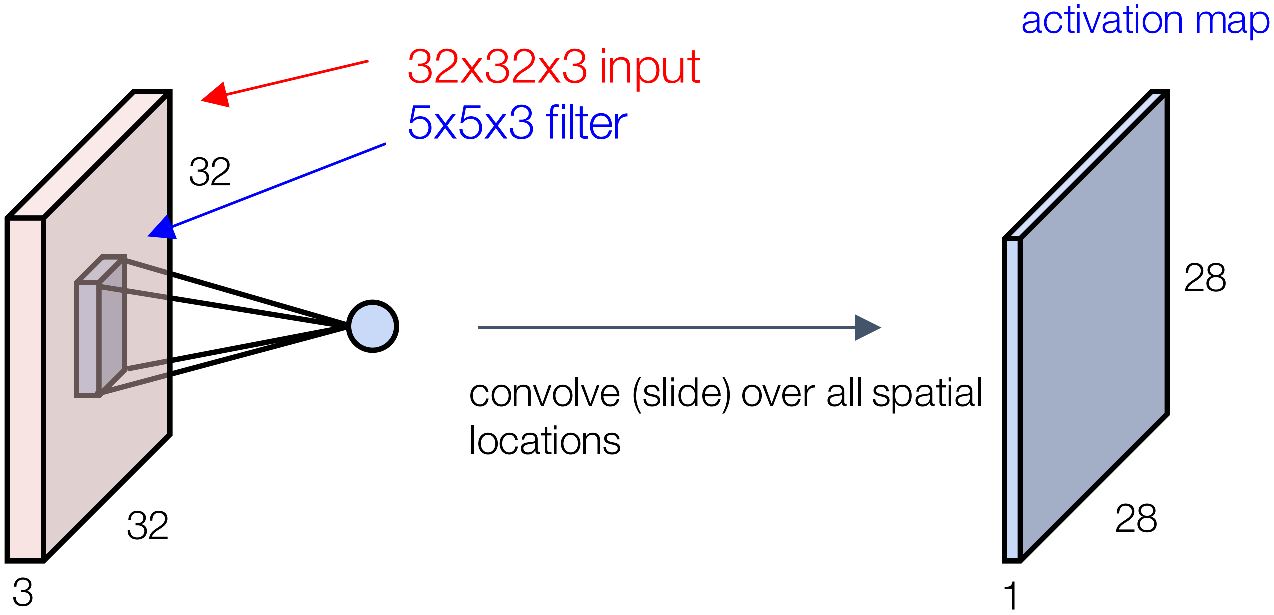
Convolve the filter with the input
i.e. “slide over the image spatially,
computing dot products”

Convolutional Layer



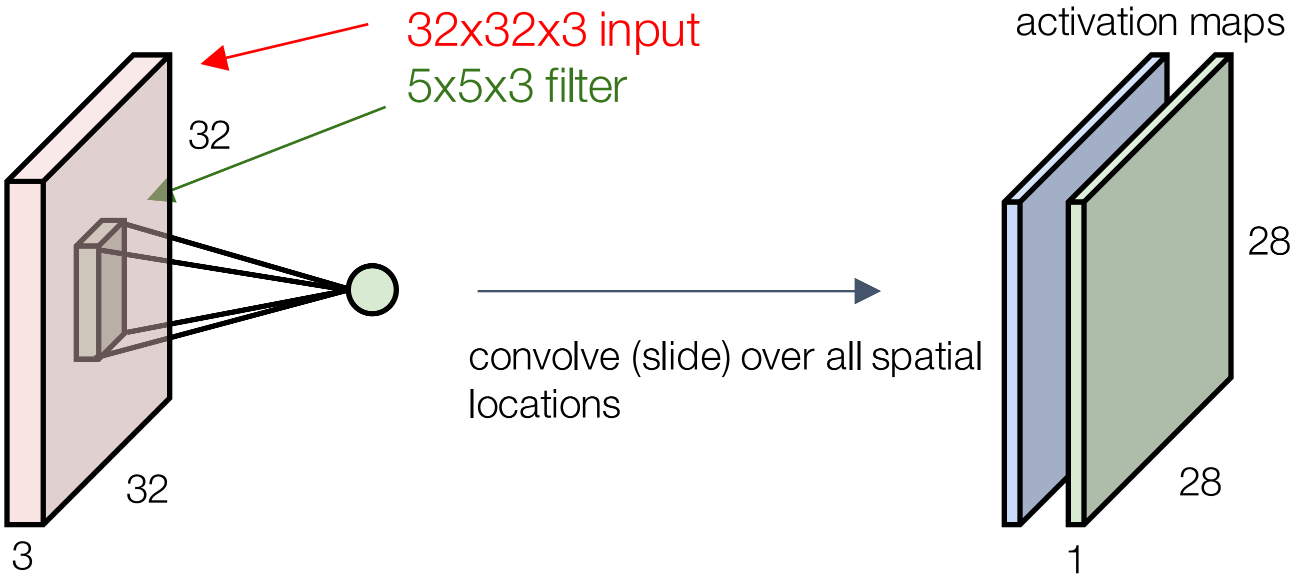
1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the input
(i.e. 5*5*3 = 75-dimensional dot product + bias)

Convolutional Layer



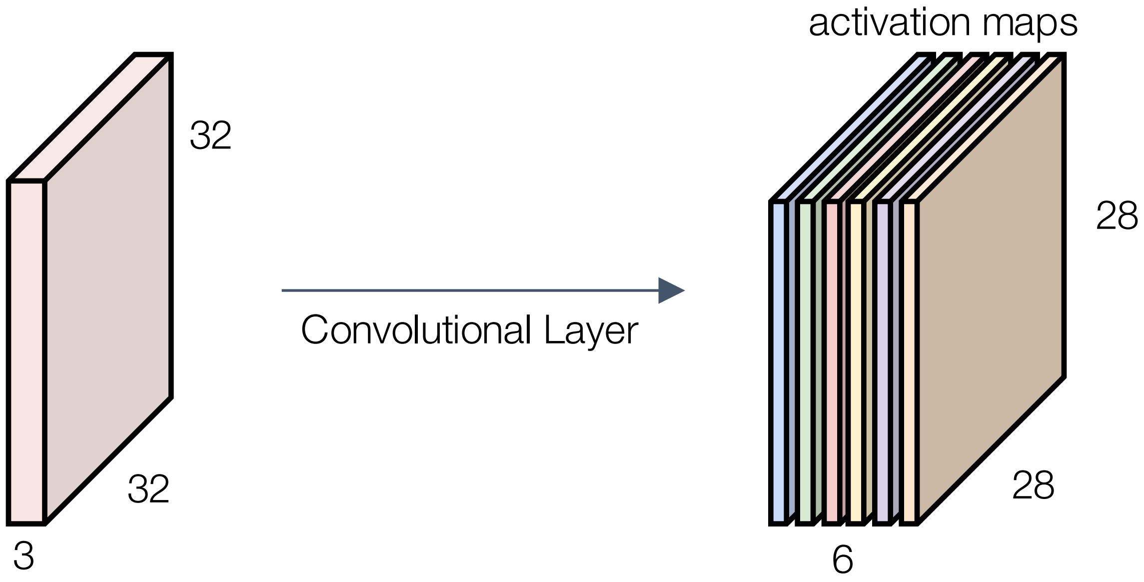
Convolutional Layer

consider a second, green filter



Convolutional Layer

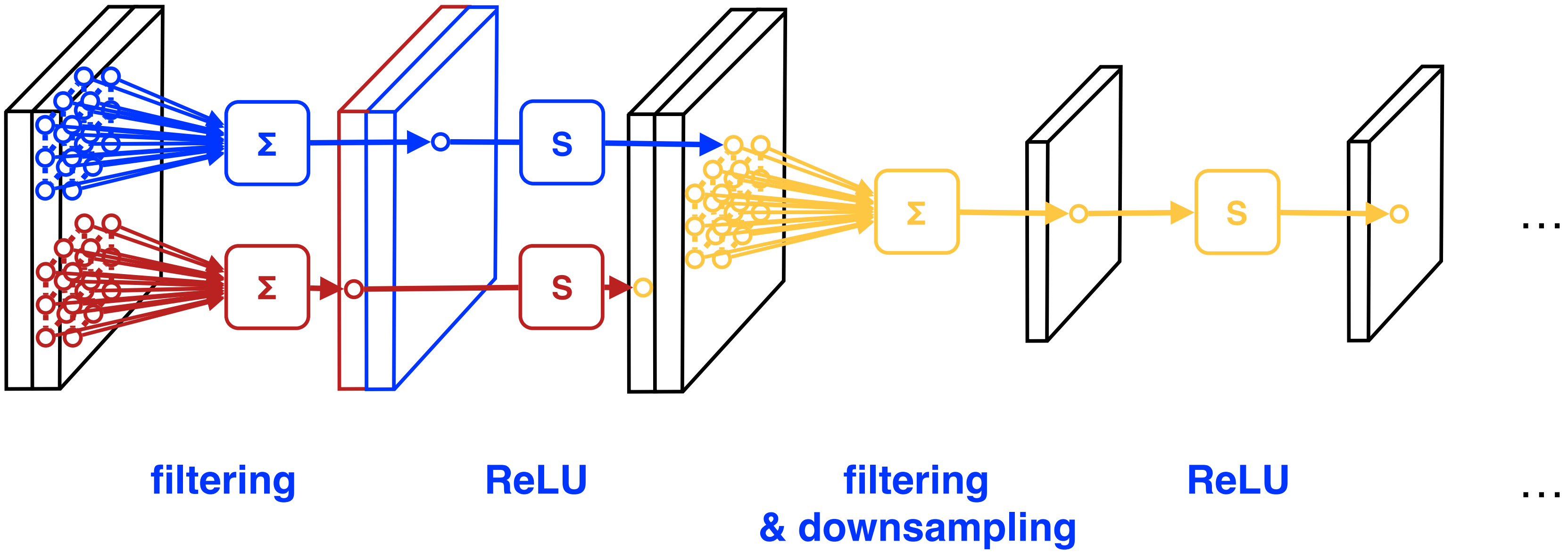
- Multiple filters produce multiple output channels
- For example, if we had 6 5x5 filters, we'll get 6 separate activation maps:



We stack these up to get an output of size 28x28x6.

Linear / non-linear chains

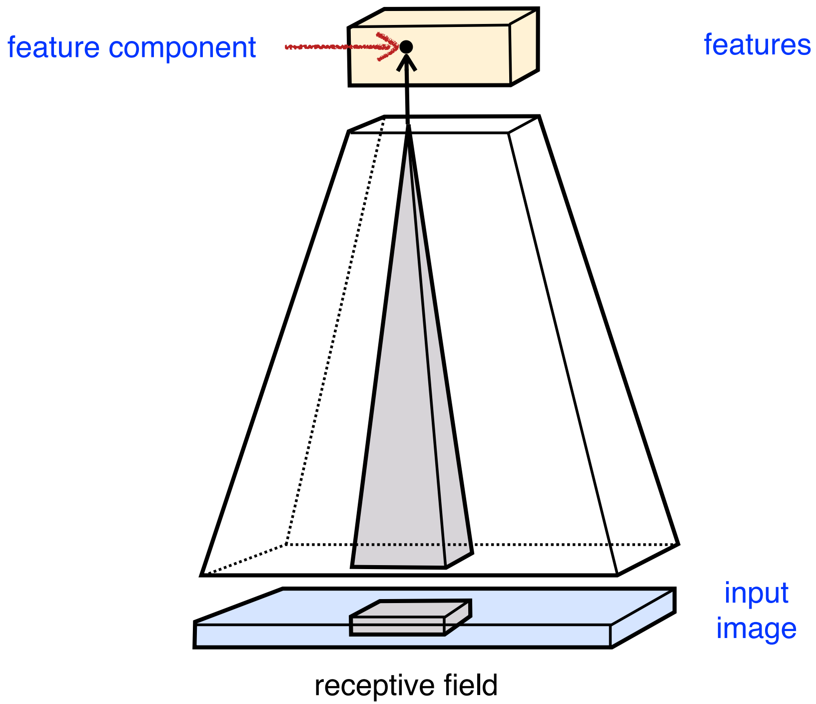
- The basic blueprint: The sandwich architecture
- Stack multiple layers of convolutions



Slide credit: Andrea Vedaldi

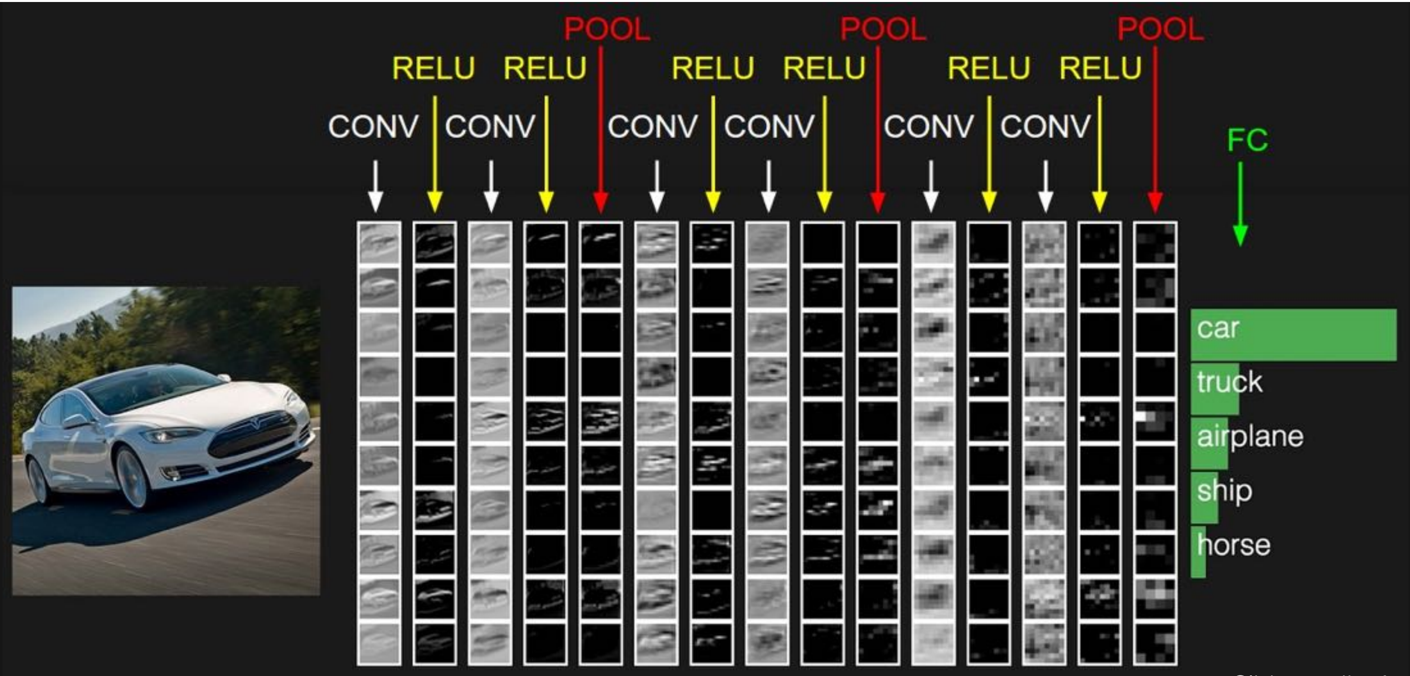
Convolutional layers

- Local receptive field
- Each column of hidden units looks at a different input patch



Feature Learning

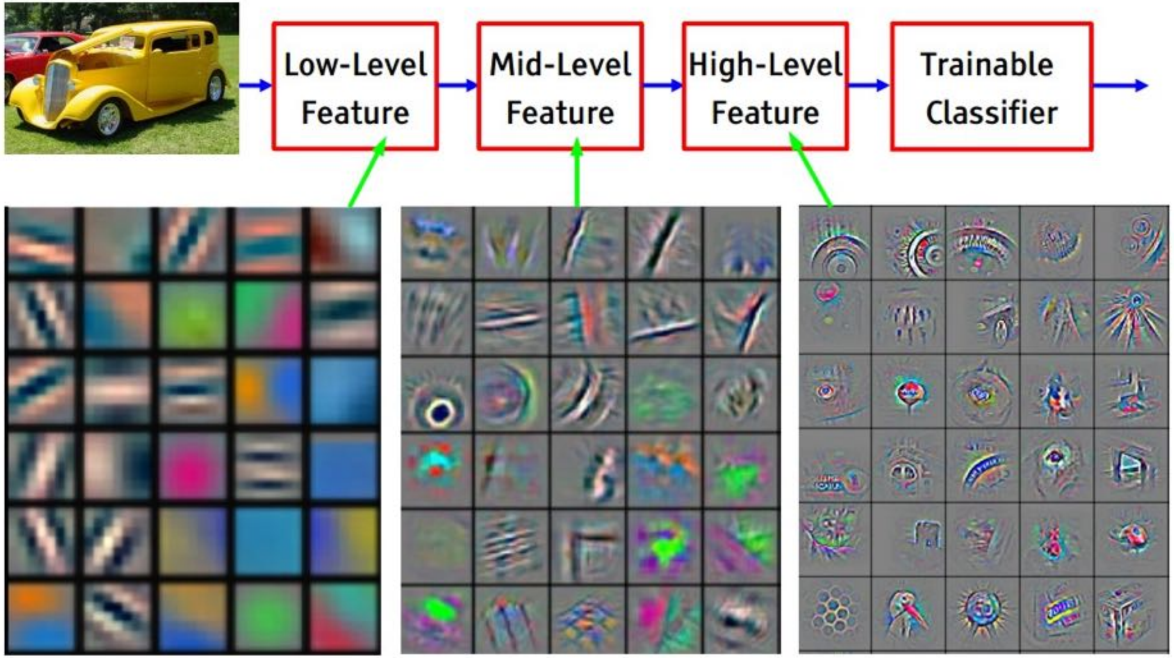
- Hierarchical layer structure allows to learn hierarchical filters (features).



Slide credit: Andrej Karpathy

Feature Learning

- Hierarchical layer structure allows to learn hierarchical filters (features).

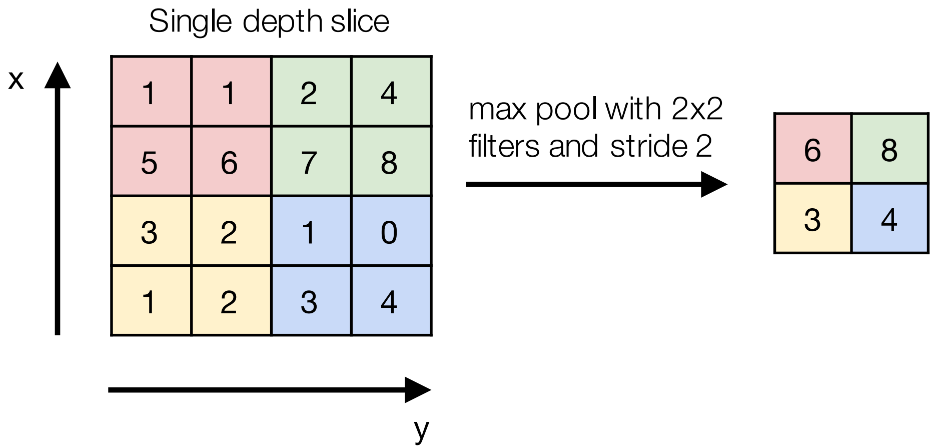


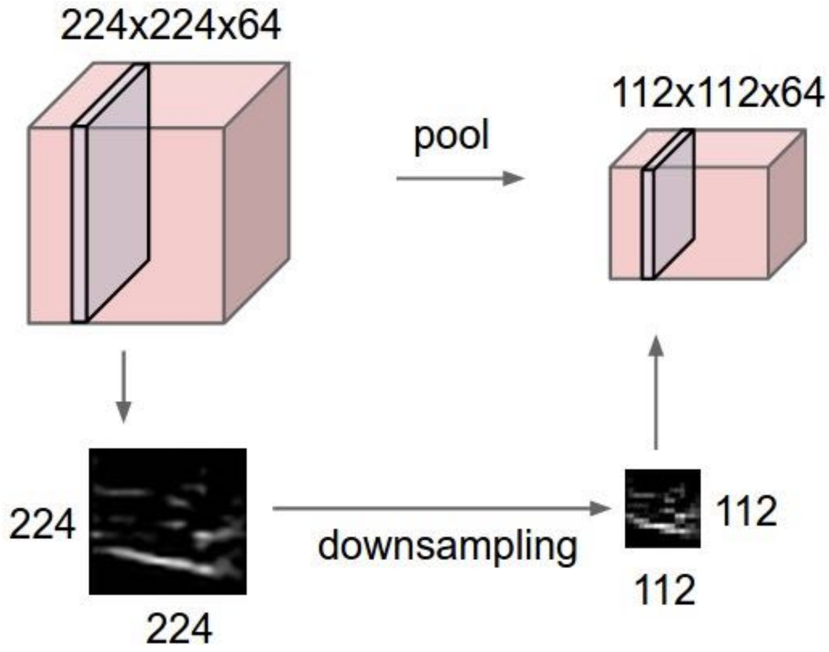
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Slide credit: Yann LeCun

Pooling layer

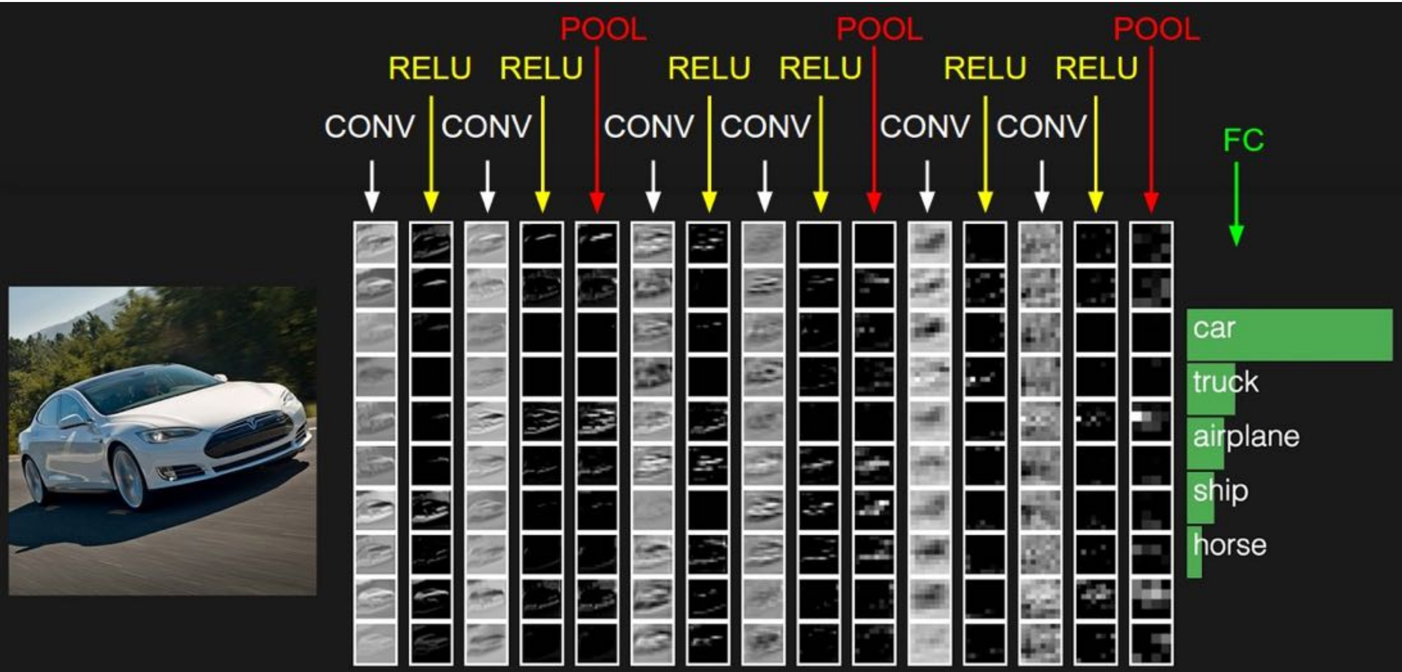
- makes the representations smaller and more manageable
- operates over each activation map independently:
- Max pooling, average pooling, etc.





Fully connected layer

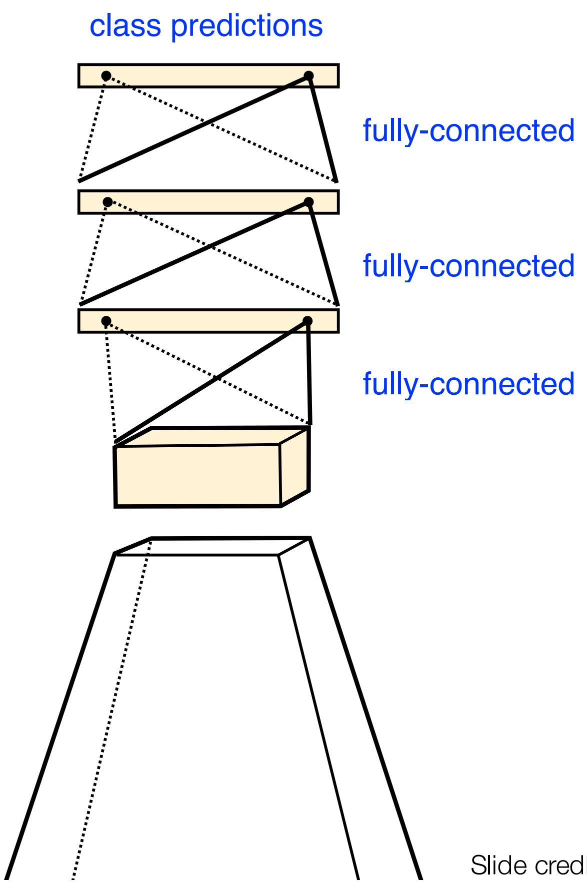
- contains neurons that connect to the entire input volume, as in ordinary Neural Networks



Slide credit: Andrej Karpathy

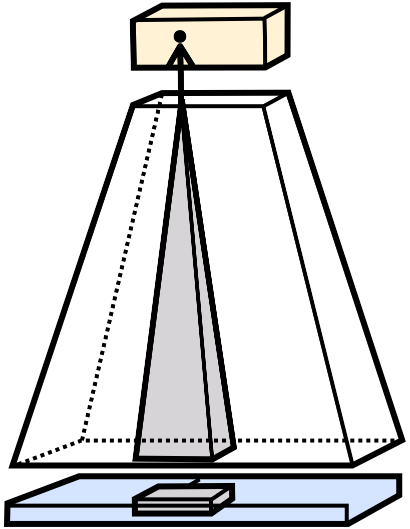
Fully connected layers

- Global receptive field
- Each hidden unit looks at the entire image

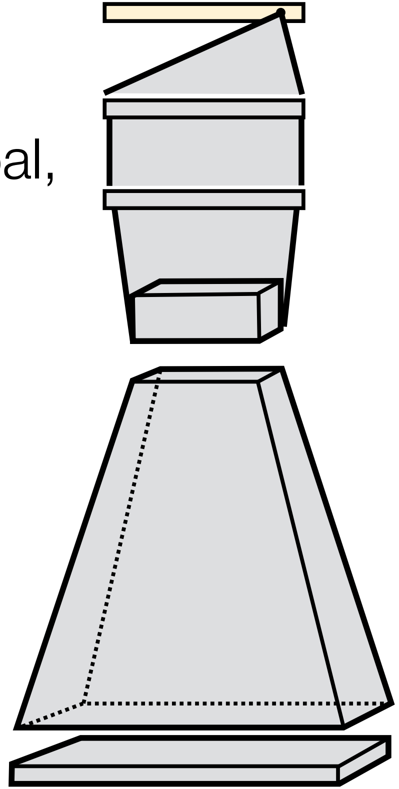


Convolutional vs Fully connected

- **Convolutional layers**:
  Responses are spatially selective, can be used to localize things.

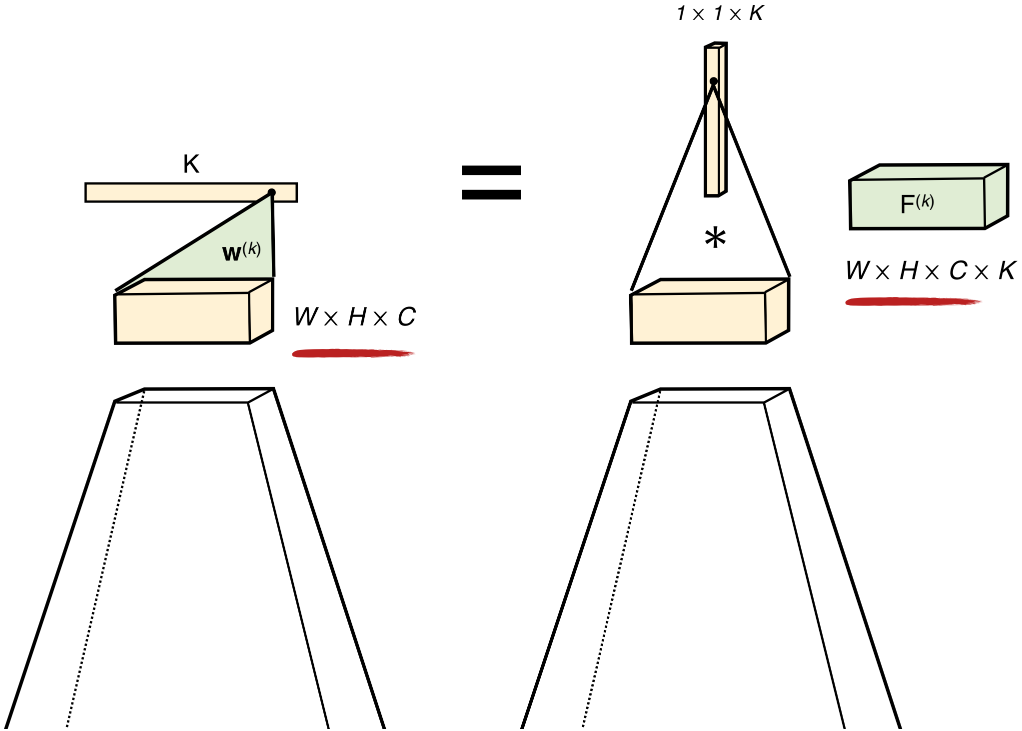


- **Fully connected layers:**
  Responses are global, do not characterize well position



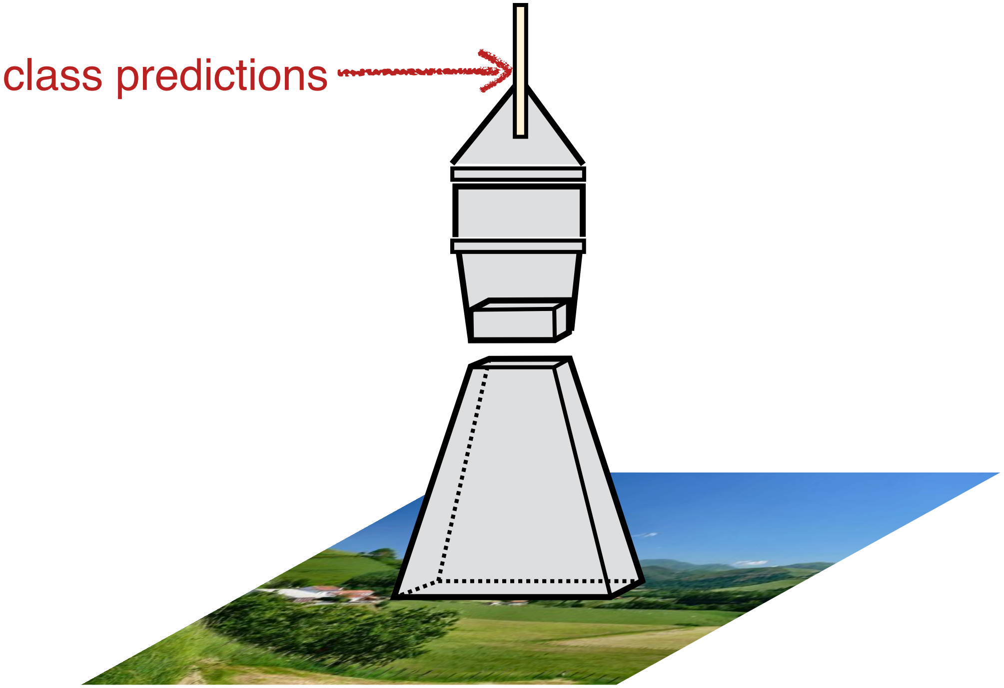
Fully connected layer = large filter

- Fully connected layer can be interpreted as a very large filter who spans the whole input data



Fully-convolutional neural networks

- Proposed for pixel-level labeling (e.g. semantic segmentation)



CNN Demo

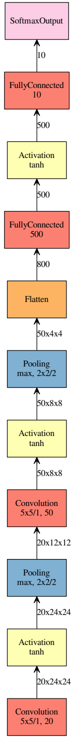
- ConvNetJS demo: training on CIFAR-10
- http://cs.stanford.edu/people/karpathy/convnetjs/demo/cifar10.html

CNNs - Years of progress

- From LeNet (1998) to ResNet (2015)

How deep is enough?

LeNet (1998)



2 convolutional layers

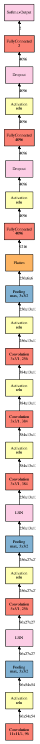
2 fully connected layers

How deep is enough?

LeNet (1998)

2 convolutional layers
2 fully connected layers

AlexNet (2012)



5 convolutional layers
3 fully connected layers

How deep is enough?

LeNet (1998) AlexNet (2012) VGGNet-M (2013)

How deep is enough?

LeNet (1998)

AlexNet (2012)

VGGNet-M (2013)

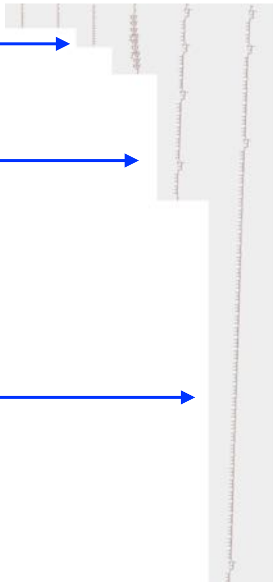
GoogLeNet (2014)

How deep is enough?

GoogLeNet (2014)
VGG-VD-16 (2014)
VGG-M (2013)
AlexNet (2012)

ResNet 50 (2015)
ResNet 152 (2015)

16 convolutional layers
50 convolutional layers
152 convolutional layers



Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet classification with deep convolutional neural networks*. In Proc. NIPS, 2012.

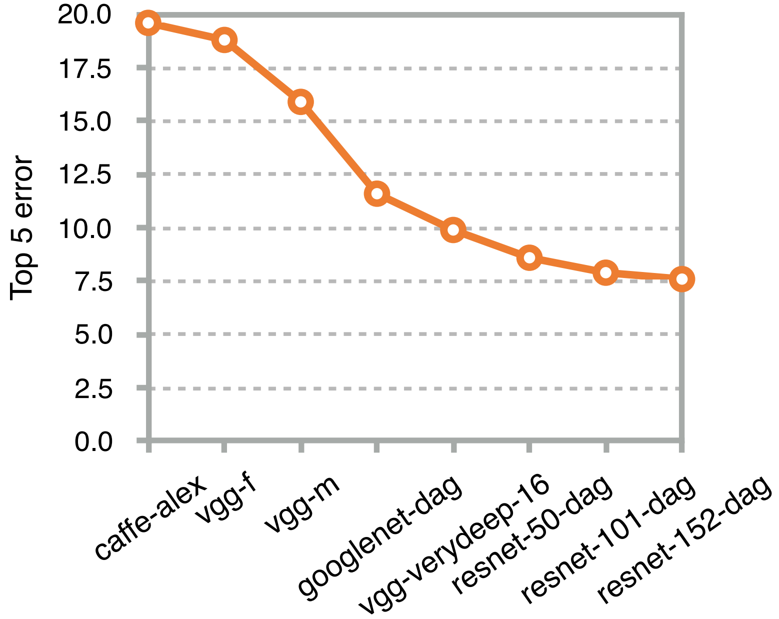
C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. *Going deeper with convolutions*. In Proc. CVPR, 2015.

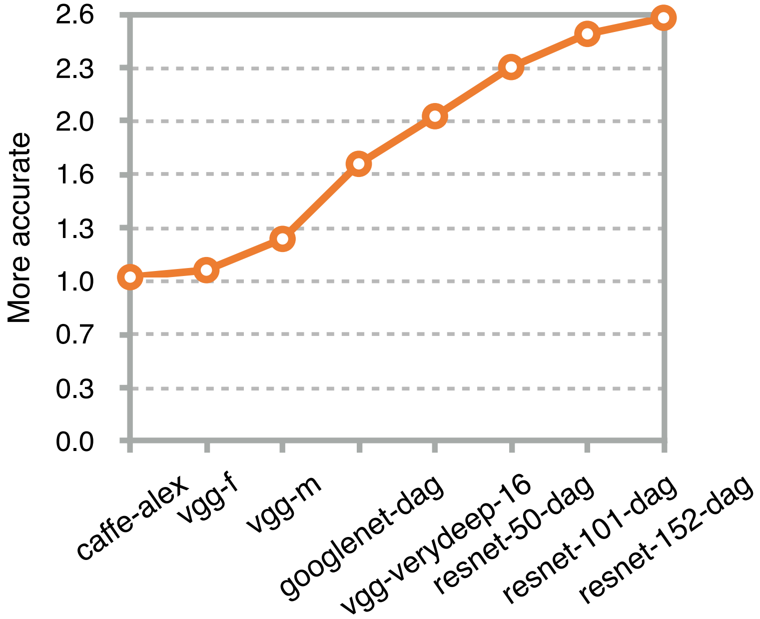
K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. In Proc. ICLR, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*. In Proc. CVPR, 2016.

Accuracy

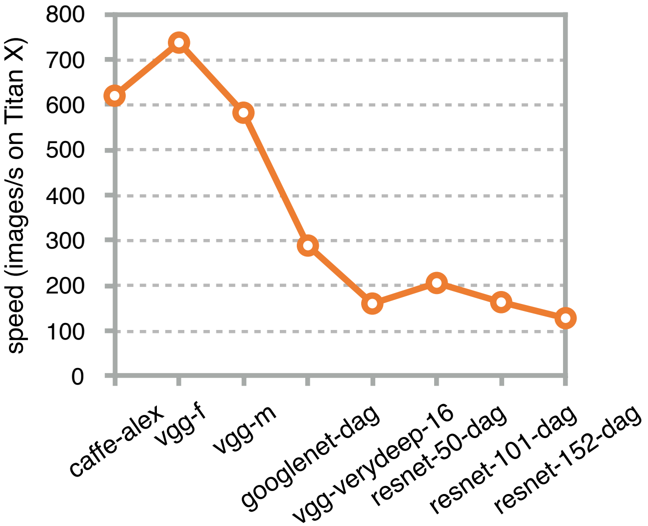
- 3 × more accurate in 3 years

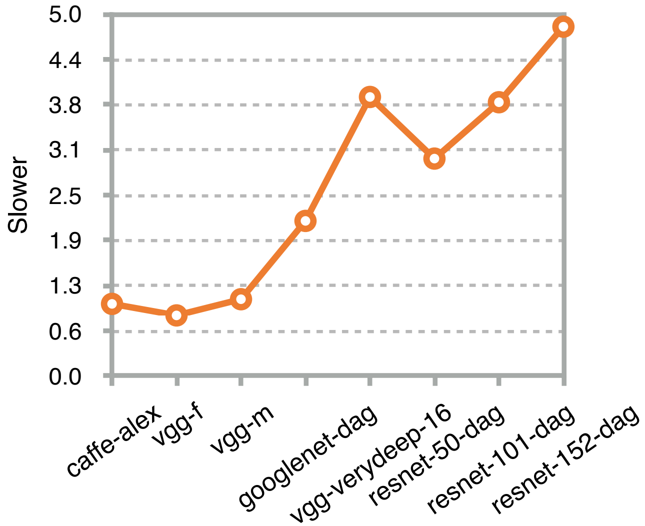




Speed

- 5 × slower

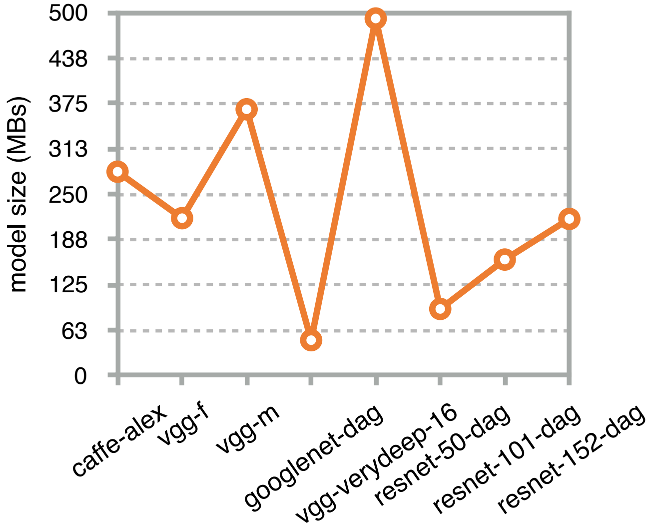


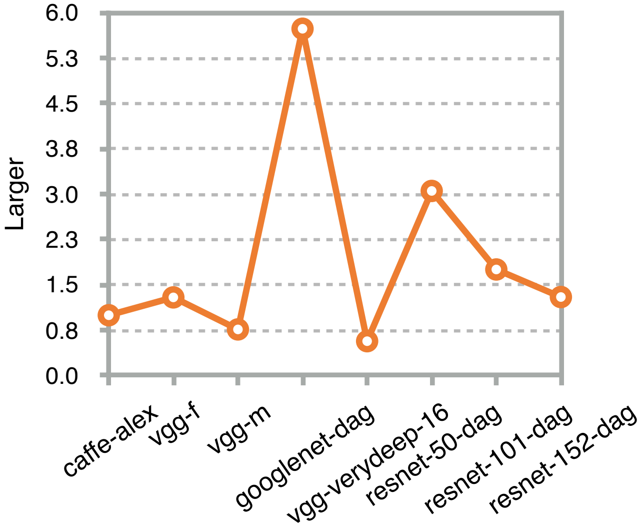


Remark: 101 ResNet layers same size/speed as 16 VGG-VD layers
Reason: far fewer feature channels (quadratic speed/space gain)
Moral: optimize your architecture

Model size

- Num. of parameters is about the same





Remark: 101 ResNet layers same size/speed as 16 VGG-VD layers
Reason: far fewer feature channels (quadratic speed/space gain)
Moral: optimize your architecture

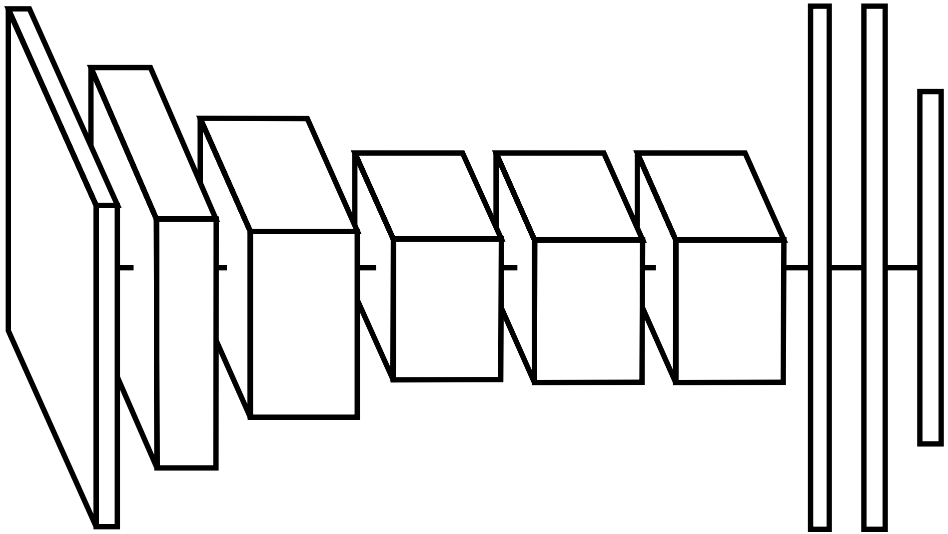
Beyond CNNs

- Do features extracted from the CNN generalize other tasks and datasets?
  - Donahue et al. (2013), Chatfield et al. (2014), Razavian et al. (2014), Yosinski et al. (2014), etc.

- CNN activations as deep features
- Finetuning CNNs

CNN activations as deep features

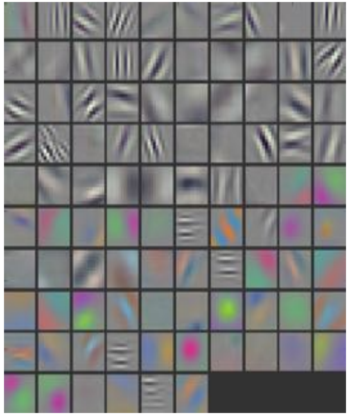
- CNNs discover effective representations. Why not to use them?



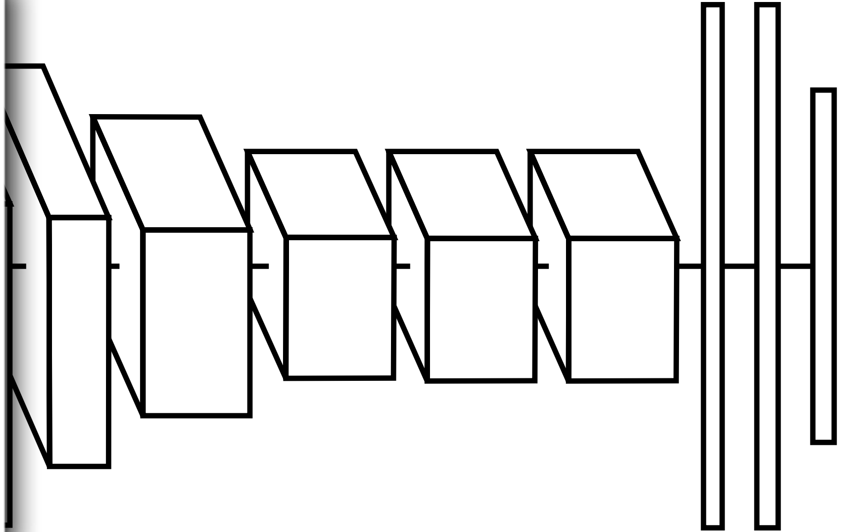
Slide credit: Jason Yosinski

CNN activations as deep features

- CNNs discover effective representations. Why not to use them?

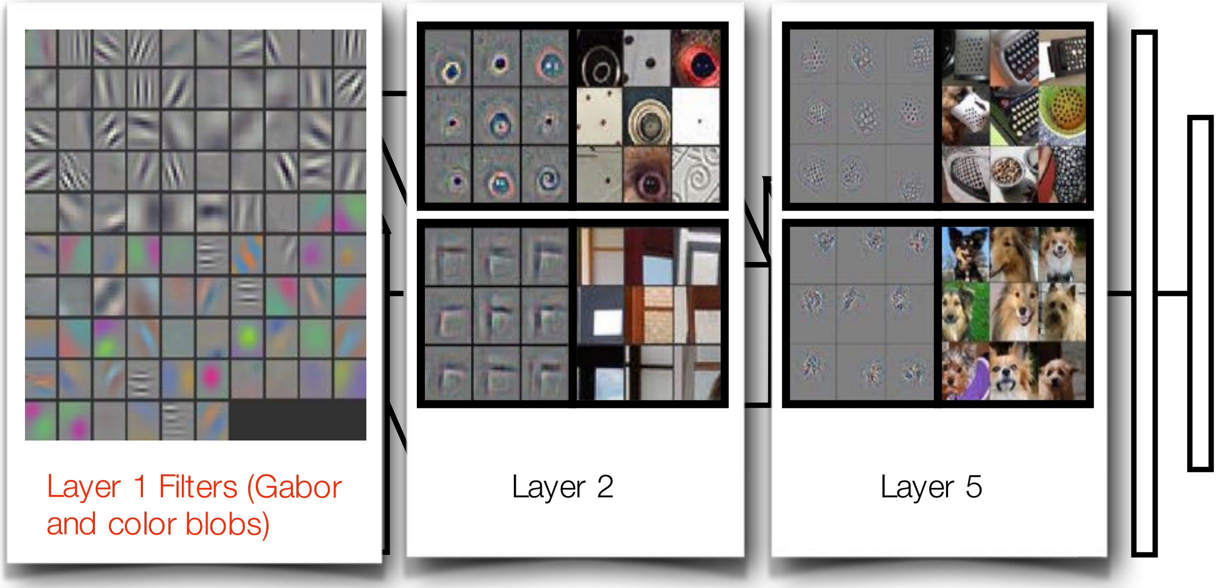


Layer 1 Filters (Gabor and color blobs)



CNN activations as deep features

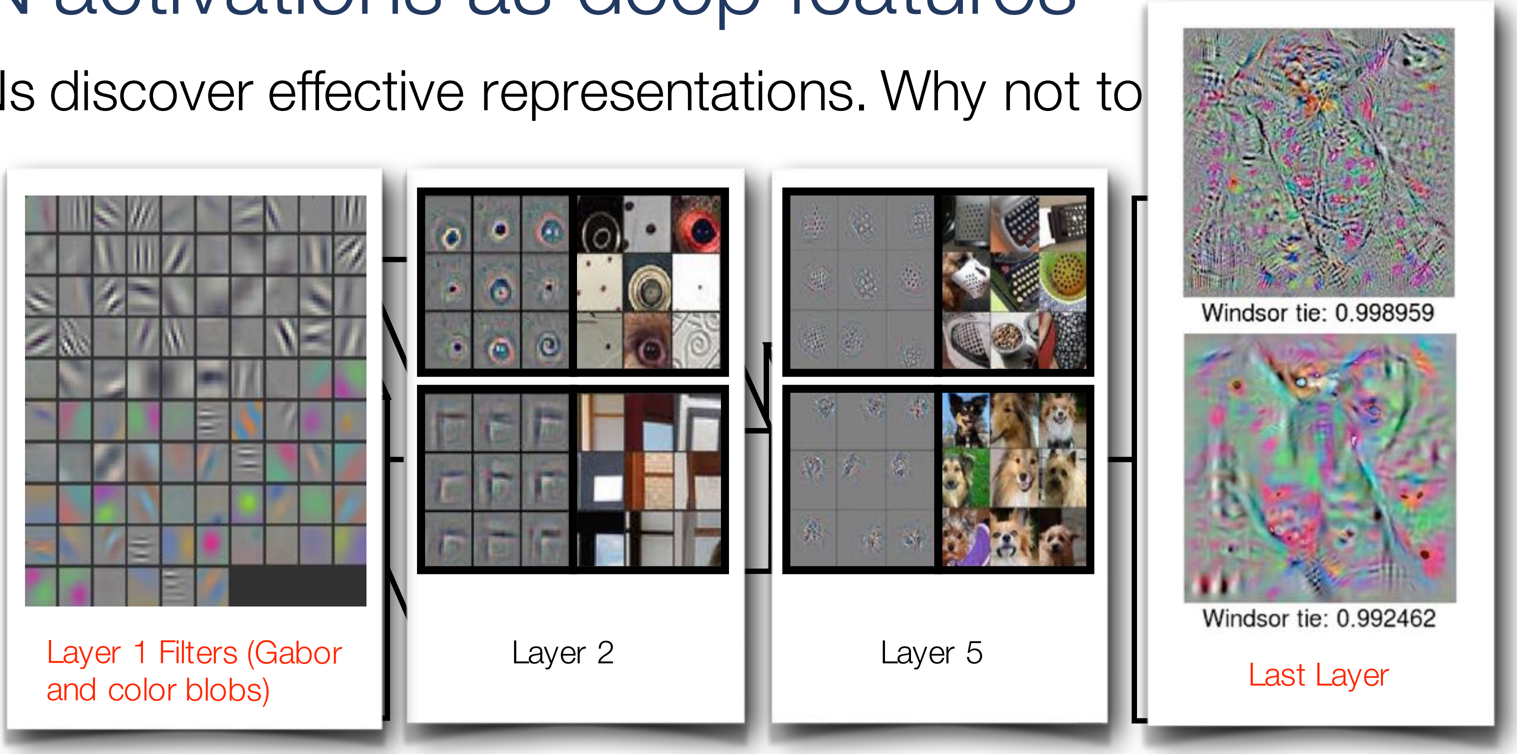
- CNNs discover effective representations. Why not to use them?



Zeiler et al., 2014

CNN activations as deep features

- CNNs discover effective representations. Why not to



Layer 1 Filters (Gabor and color blobs)

Layer 2

Layer 5

Last Layer

Zeiler et al., 2014

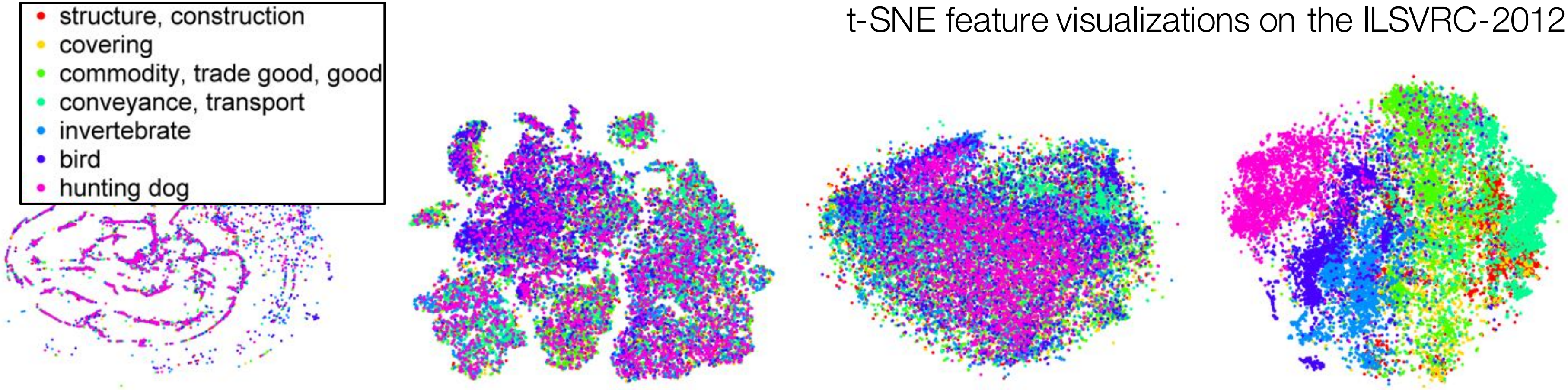
Nguyen et al., 2014

Slide credit: Jason Yosinski

CNNs as deep features

- CNNs discover effective representations. Why not to use them?

t-SNE feature visualizations on the ILSVRC-2012



LLC GIST Conv-1 activations Conv-6 activations

DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, [Donahue et al.,'14]

Transfer Learning with CNNs

- A CNN trained on a (large enough) dataset generalizes to other visual tasks

Learning visual features from Large Weakly supervised Data, [Joulin et al.,'15]

Slide credit: Joan Bruna

Transfer Learning with CNNs

- Keep layers 1-7 of our ImageNet-trained model fixed
- Train a new softmax classifier on top using the training images of the new dataset.

1. Train on Imagenet

2. Small dataset: feature extractor

Freeze these

Train this

3. Medium dataset: finetuning

more data = retrain more of the network (or all of it)

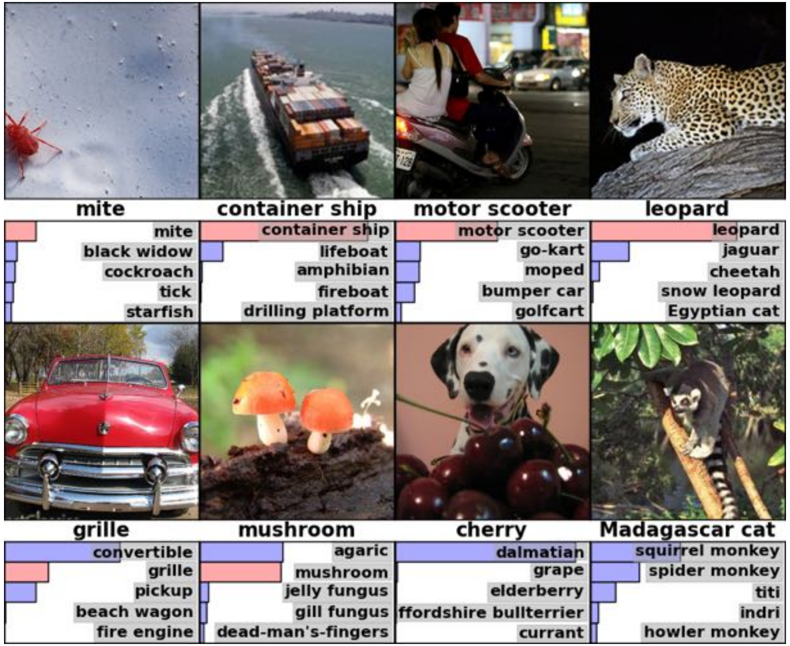
Freeze these

tip: use only ~1/10th of the original learning rate in finetuning top layer, and ~1/100th on intermediate layers

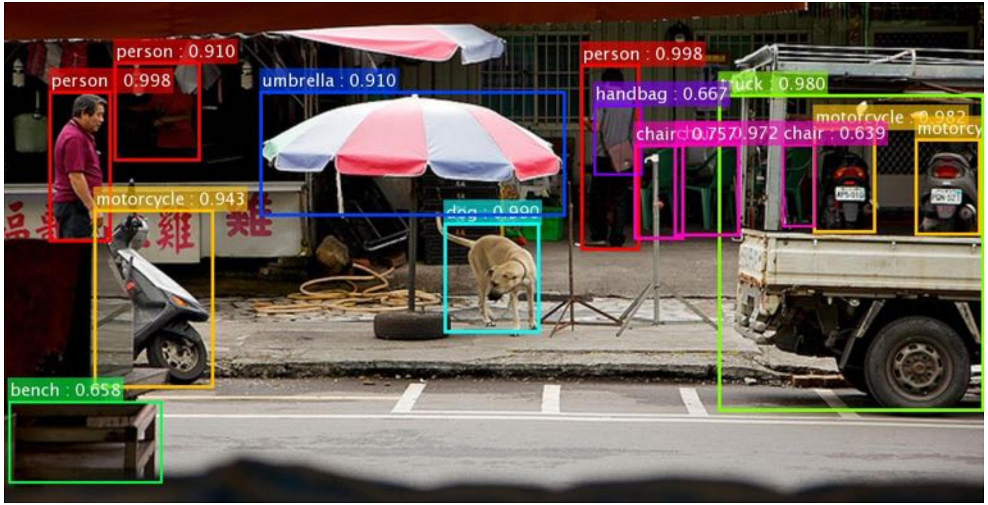
Train this

Slide credit: Andrej Karpathy

CNNs in Computer Vision

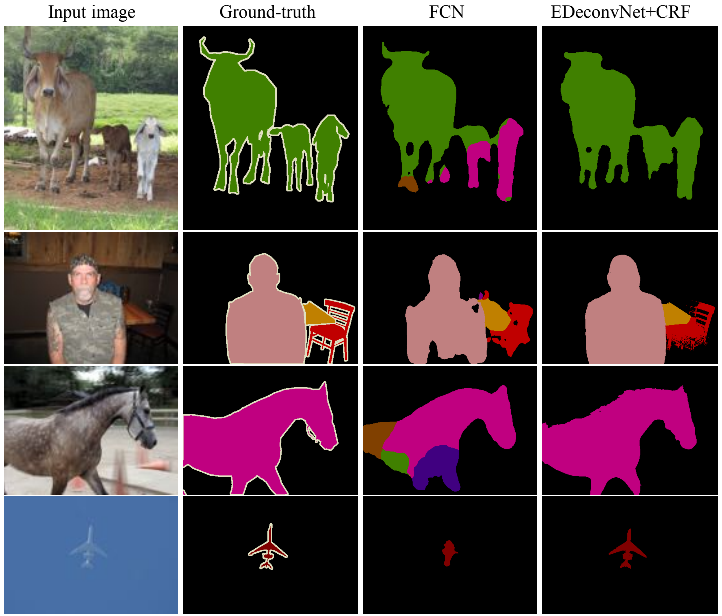


Classification (Krizhevsky et al., 2012)



Object detection (Ren et al., 2015)

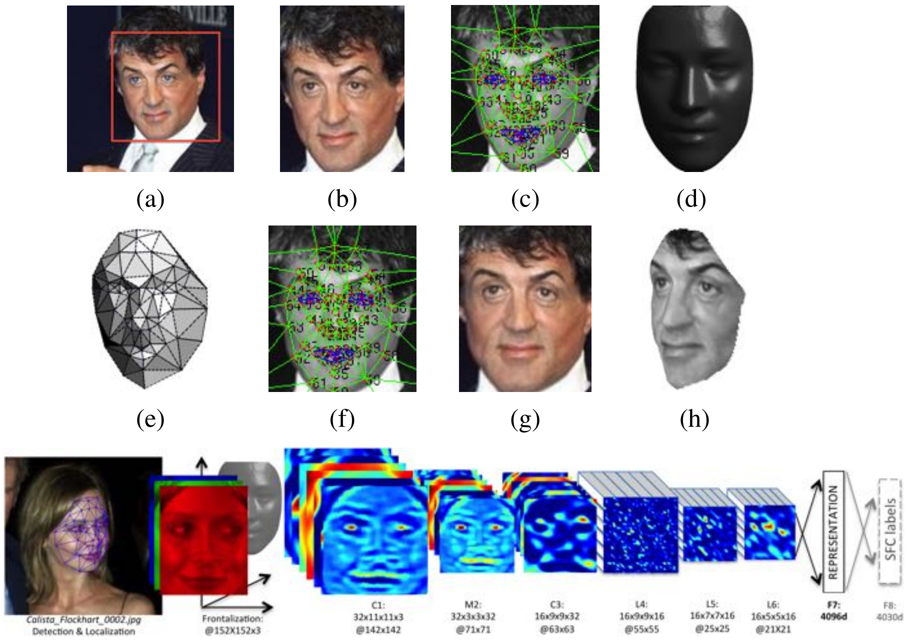
CNNs in Computer Vision



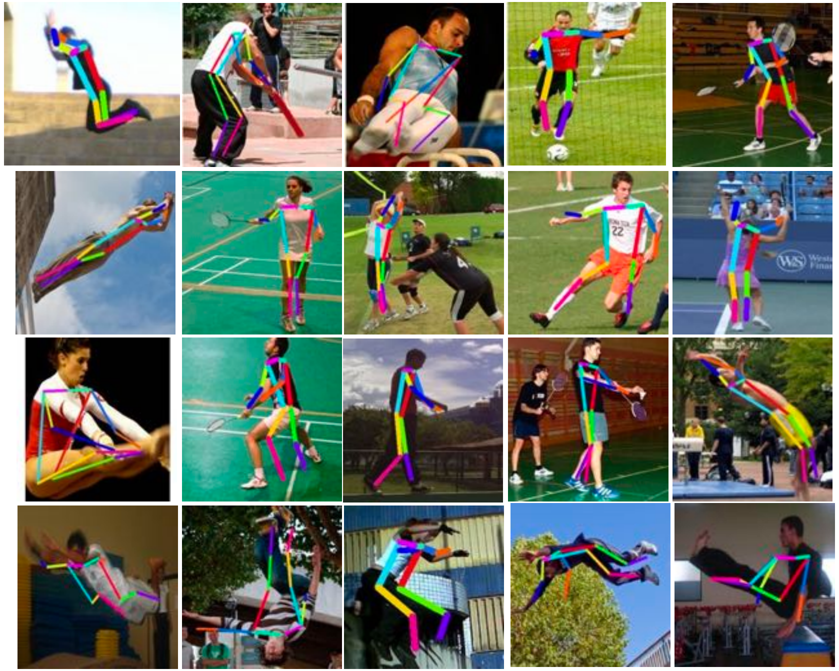
Semantic Segmentation (Noh et al., 2015)

Multi-Instance Segmentation (He and Gould, 2014)

CNNs in Computer Vision



Face recognition (Taigman et al., 2014)



Pose estimation (Toshev and Szegedy, 2014)

CNNs in Computer Vision

Text detection and retrieval (Jaderberg et al., 2016)

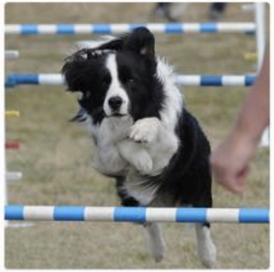
CNNs in Computer Vision

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."



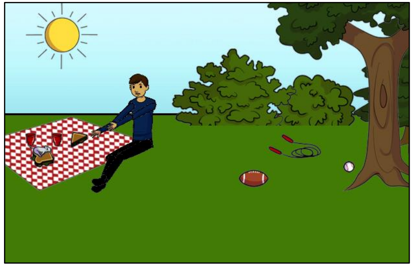
"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

Image Captioning (Karpathy and Fei-Fei, 2015)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Visual Question Answering (Antol et al., 2015)