Supervised Deep Learning

Marc'Aurelio Ranzato

Facebook A.I. Research

Supervised Learning

$\{(\boldsymbol{x}^i, y^i), i = 1 \ldots P\}$ training dataset

$\boldsymbol{x}^i$ i-th input training example

y^i i-th target label

P number of training examples

Goal: predict the target label of unseen inputs.

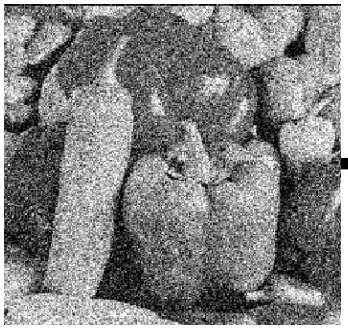
Supervised Learning: Examples

Classification

“dog”

classification

Denoising



regression

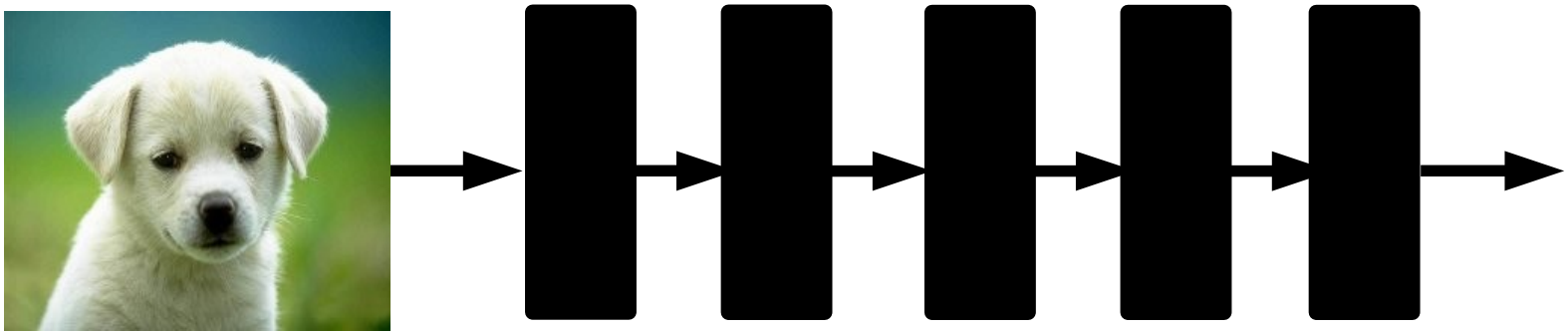
OCR

“2 3 4 5”

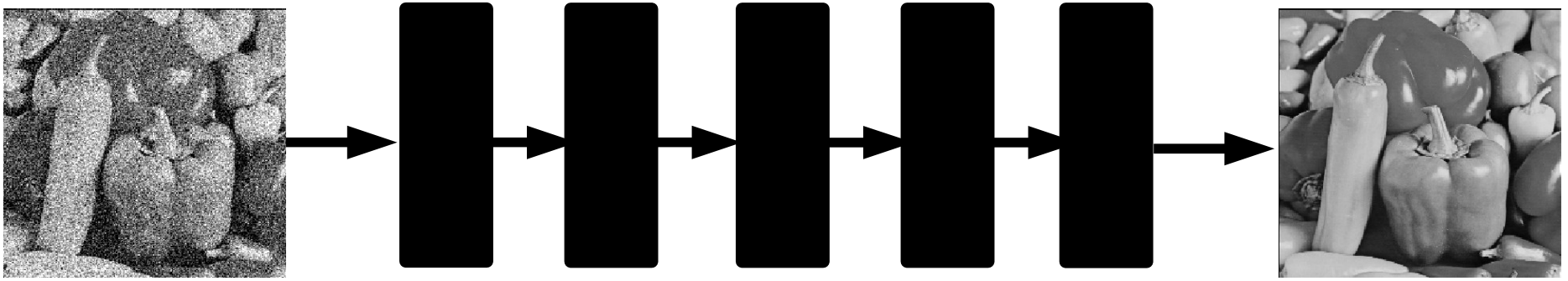
structured prediction

Supervised Deep Learning

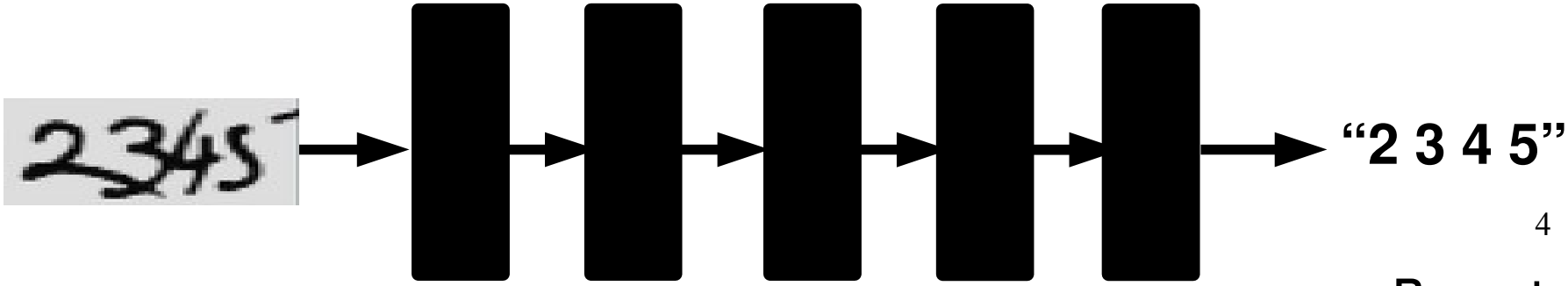
Classification



Denoising



OCR



Outline

- Supervised Neural Networks
- Convolutional Neural Networks
- Examples
- Tips

Neural Networks

Assumptions (for the next few slides):

- The input image is vectorized (disregard the spatial layout of pixels)
- The target label is discrete (classification)

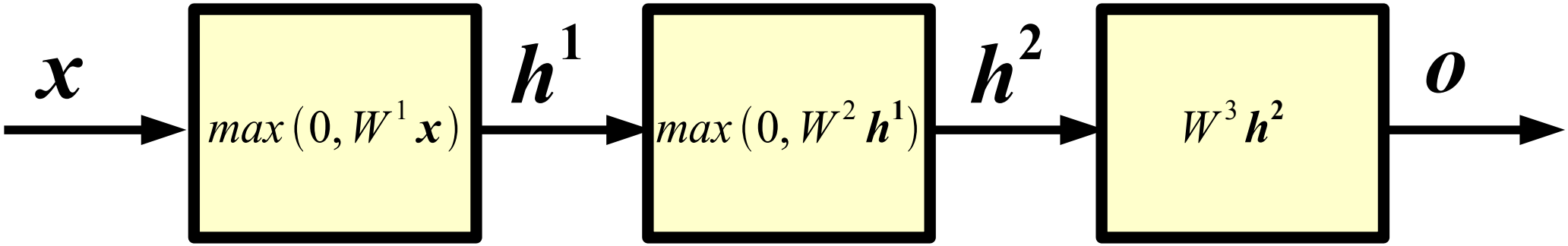
Question: what class of functions shall we consider to map the input into the output?

Answer: composition of simpler functions.

Follow-up questions: Why not a linear combination? What are the "simpler" functions? What is the interpretation?

Answer: later...

Neural Networks: example



$\boldsymbol{x}$ input

$\boldsymbol{h}^1$ 1-st layer hidden units

$\boldsymbol{h}^2$ 2-nd layer hidden units

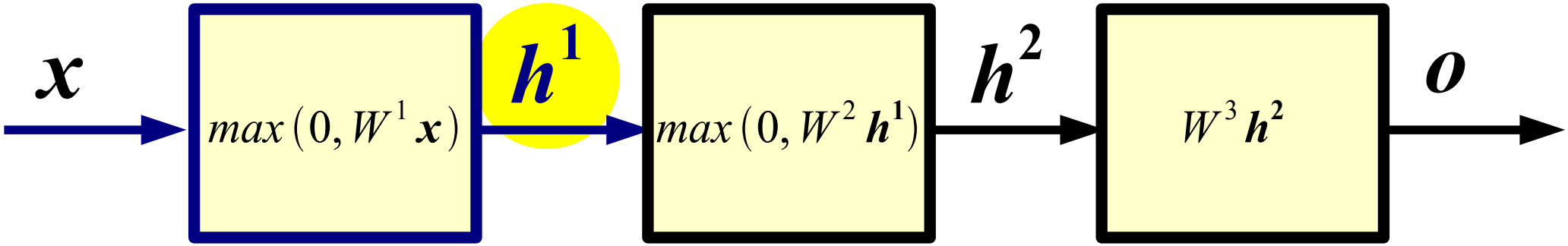
$\boldsymbol{o}$ output

Example of a 2 hidden layer neural network (or 4 layer network, counting also input and output).

Forward Propagation

Def.: Forward propagation is the process of computing the output of the network given its input.

Forward Propagation



$$\boldsymbol{x} \in R^{D} \qquad W^{1} \in R^{N_1 \times D} \qquad \boldsymbol{b}^{1} \in R^{N_1} \qquad \boldsymbol{h}^{1} \in R^{N_1}$$

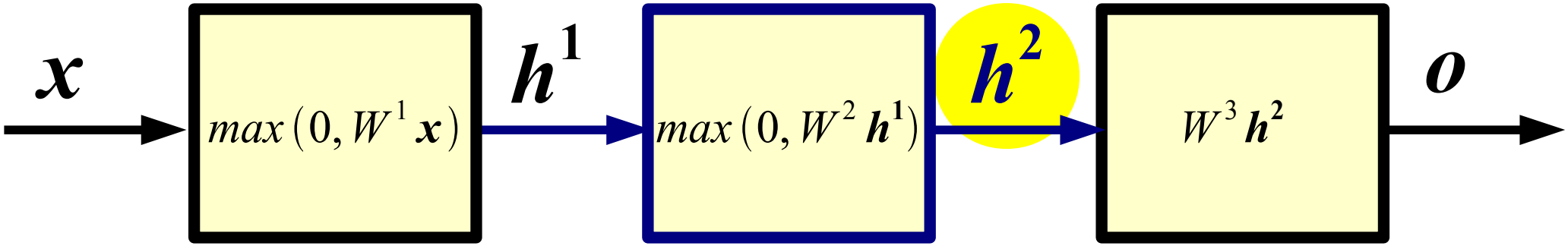
$$\boldsymbol{h}^{1} = max(0, W^{1}\boldsymbol{x} + \boldsymbol{b}^{1})$$

W^1 1-st layer weight matrix or weights

$\boldsymbol{b}^1$ 1-st layer biases

The non-linearity $u = max(0, v)$ is called **ReLU** in the DL literature.
Each output hidden unit takes as input all the units at the previous layer: each such layer is called "**fully connected**".

Forward Propagation



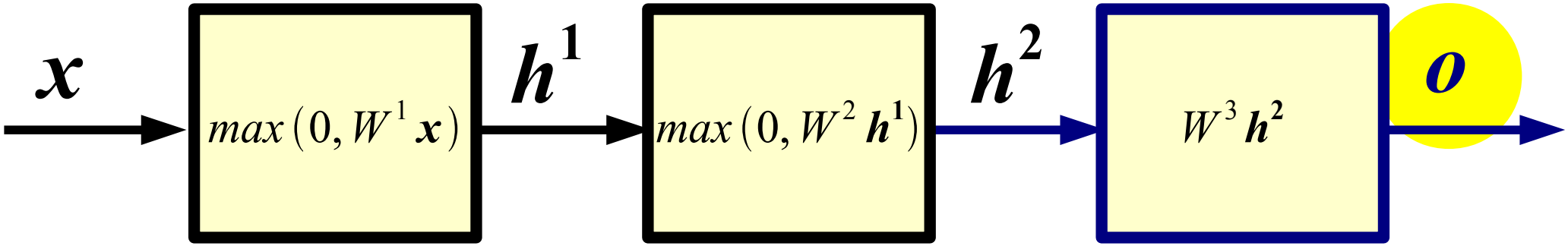
$$\boldsymbol{h^1} \in R^{N_1} \quad W^2 \in R^{N_2 \times N_1} \quad \boldsymbol{b^2} \in R^{N_2} \qquad \boldsymbol{h^2} \in R^{N_2}$$

$$\boldsymbol{h^2} = max\left(0, W^2 \boldsymbol{h^1} + \boldsymbol{b^2}\right)$$

W^2 2-nd layer weight matrix or weights

$\boldsymbol{b^2}$ 2-nd layer biases

Forward Propagation



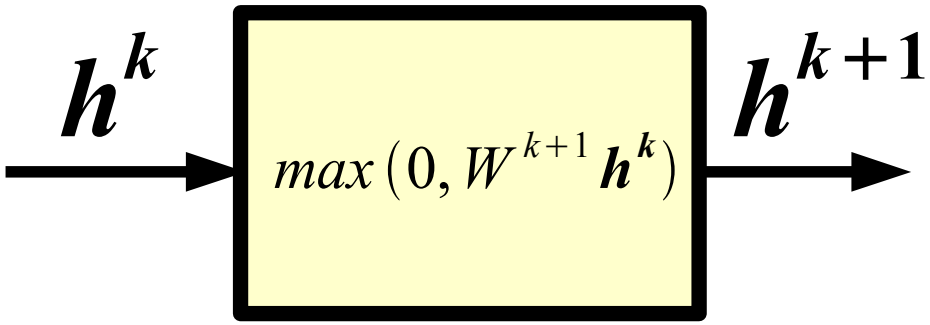
$$\boldsymbol{h^2} \in R^{N_2} \quad W^3 \in R^{N_3 \times N_2} \quad \boldsymbol{b^3} \in R^{N_3} \qquad \boldsymbol{o} \in R^{N_3}$$

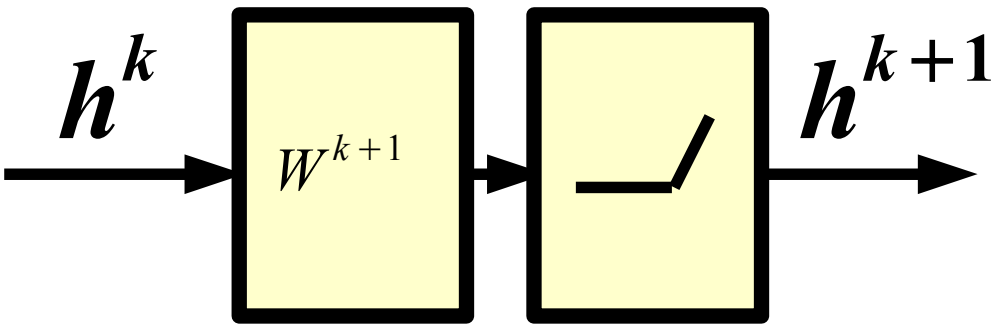
$$\boldsymbol{o} = max\left(0, W^3 \boldsymbol{h^2} + \boldsymbol{b^3}\right)$$

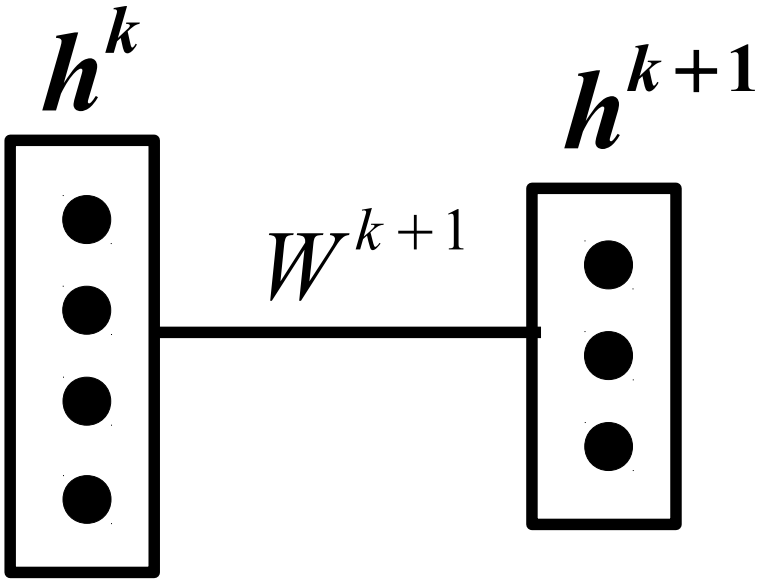
W^3 3-rd layer weight matrix or weights

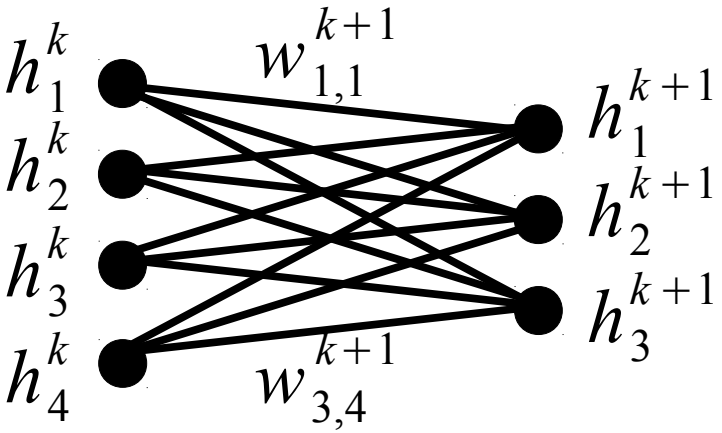
$\boldsymbol{b^3}$ 3-rd layer biases

Alternative Graphical Representation









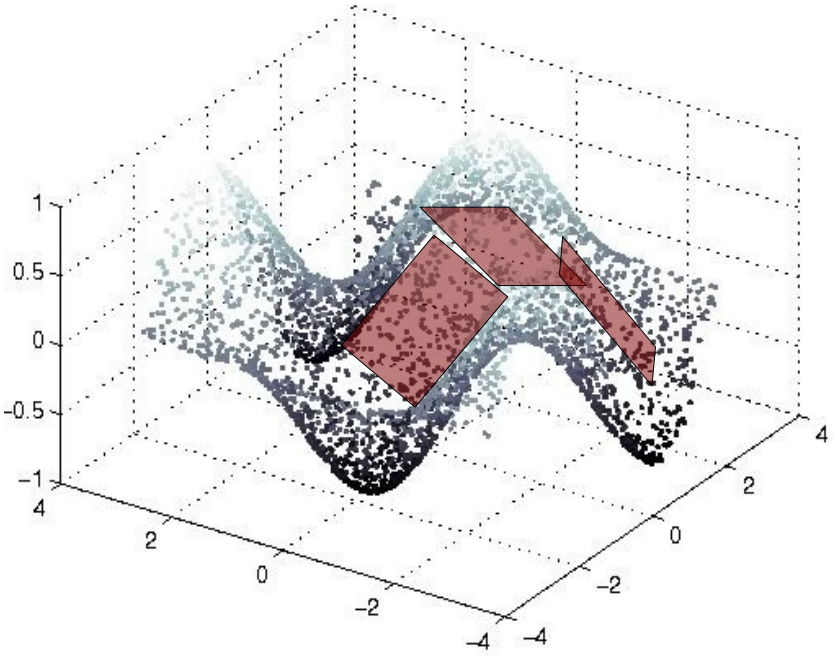
Interpretation

Question: Why can't the mapping between layers be linear?

Answer: Because composition of linear functions is a linear function. Neural network would reduce to (1 layer) logistic regression.

Question: What do ReLU layers accomplish?

Answer: Piece-wise linear tiling: mapping is locally linear.

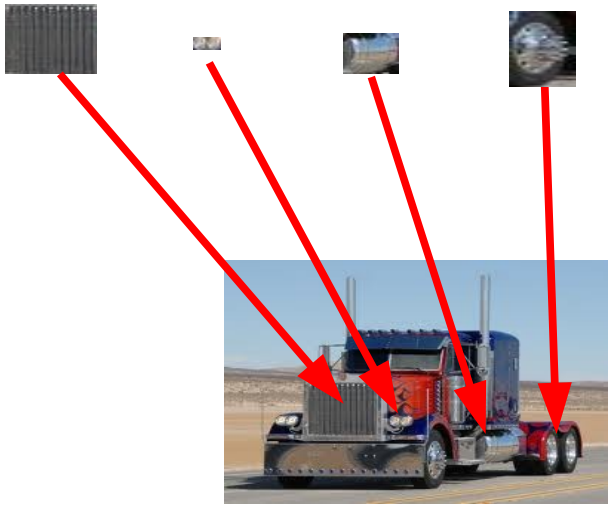


Interpretation

Question: Why do we need many layers?

Answer: When input has hierarchical structure, the use of a hierarchical architecture is potentially more efficient because intermediate computations can be re-used. DL architectures are efficient also because they use **distributed representations** which are shared across classes.

[0 0 **1** 0 0 0 0 **1** 0 0 **1** **1** 0 0 **1** 0 …] truck feature

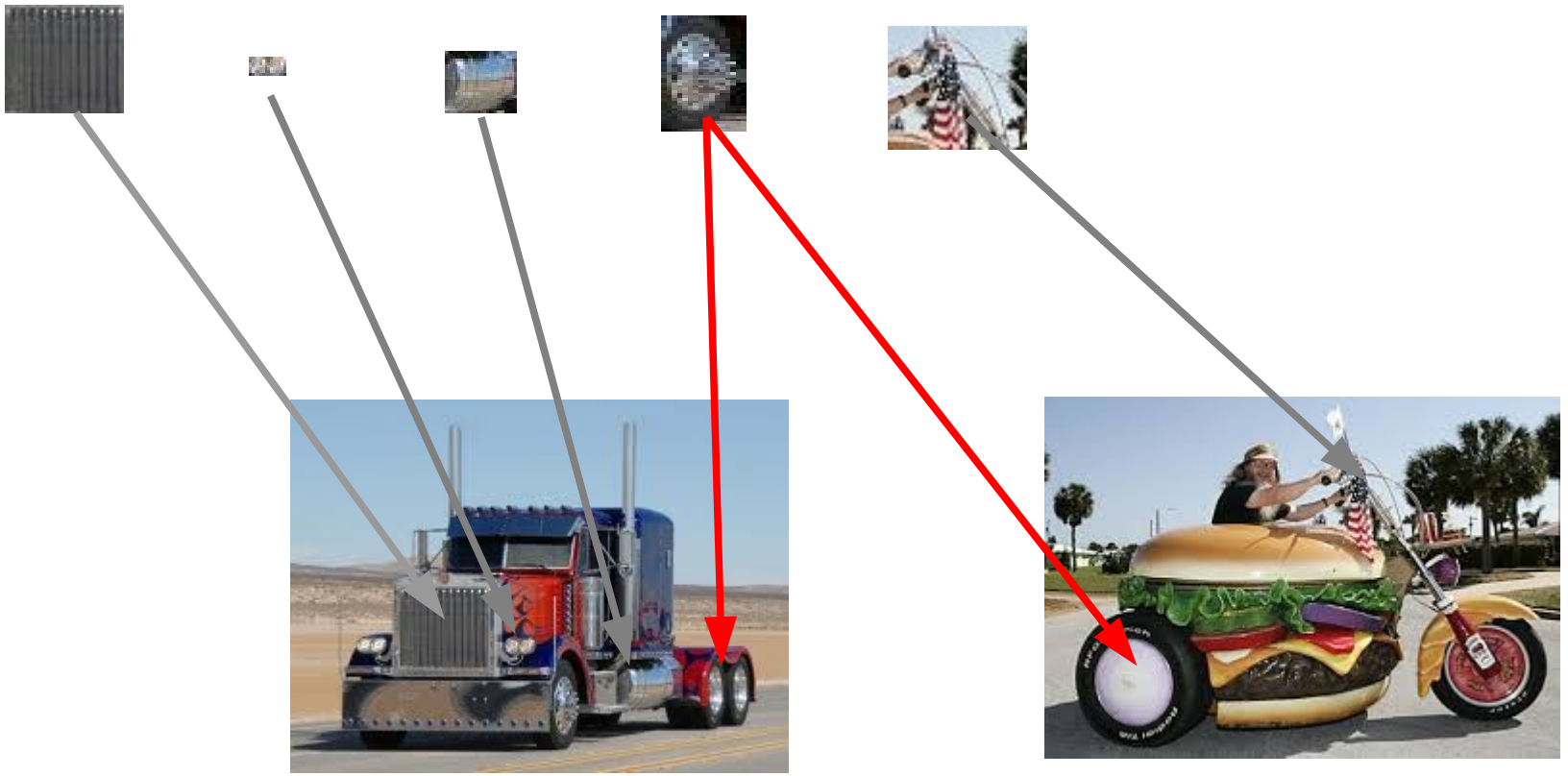


Exponentially more efficient than a 1-of-N representation (a la k-means)

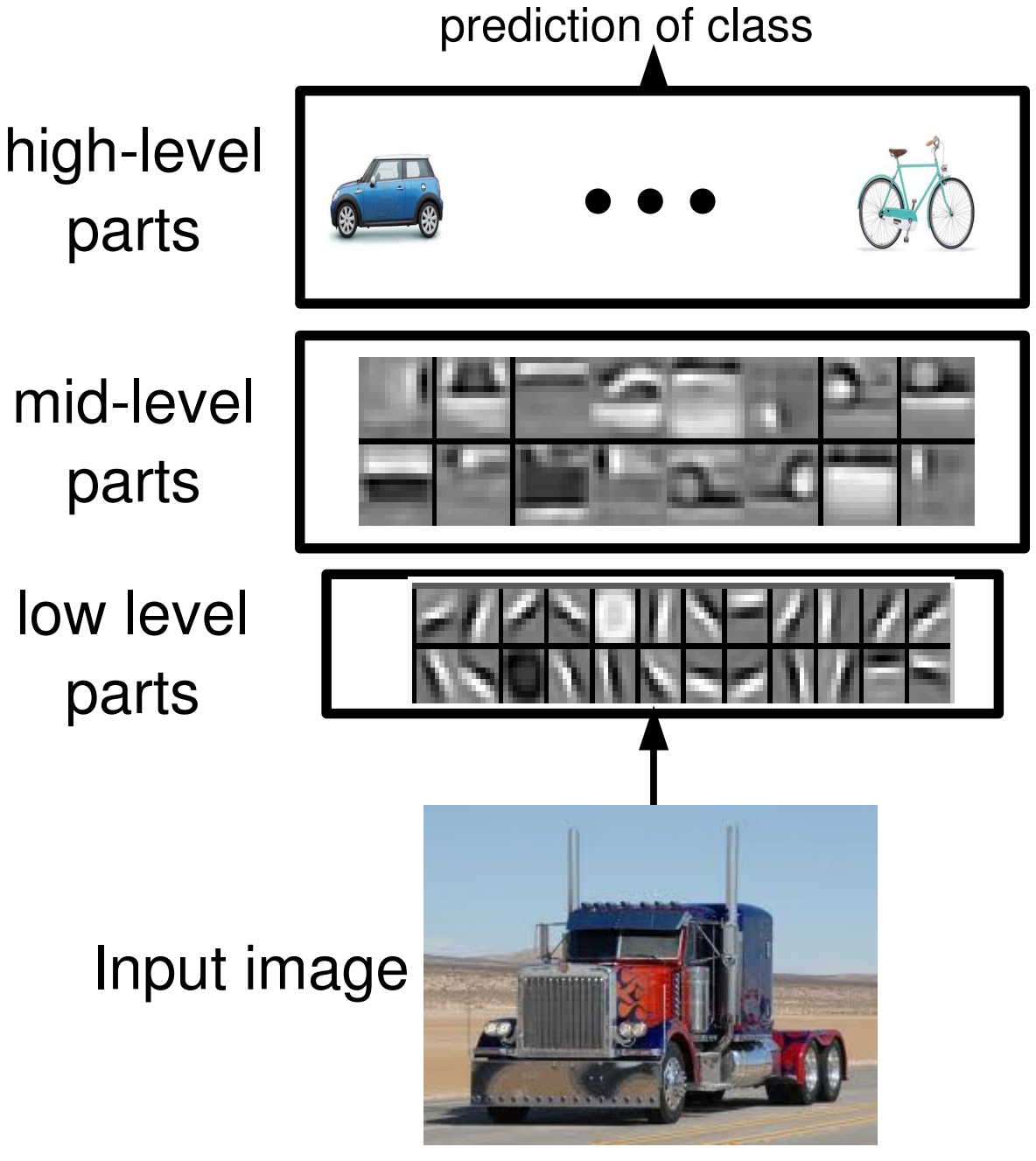
Interpretation

[1 1 0 0 0 1 0 **1** 0 0 0 0 1 1 0 1…] motorbike

[0 0 **1** 0 0 0 0 **1** 0 0 **1** 1 0 0 1 0 …] truck



Interpretation



- distributed representations
- feature sharing
- compositionality

Lee et al. “Convolutional DBN's ...” ICML 2009

Interpretation

Question: What does a hidden unit do?

Answer: It can be thought of as a classifier or feature detector.

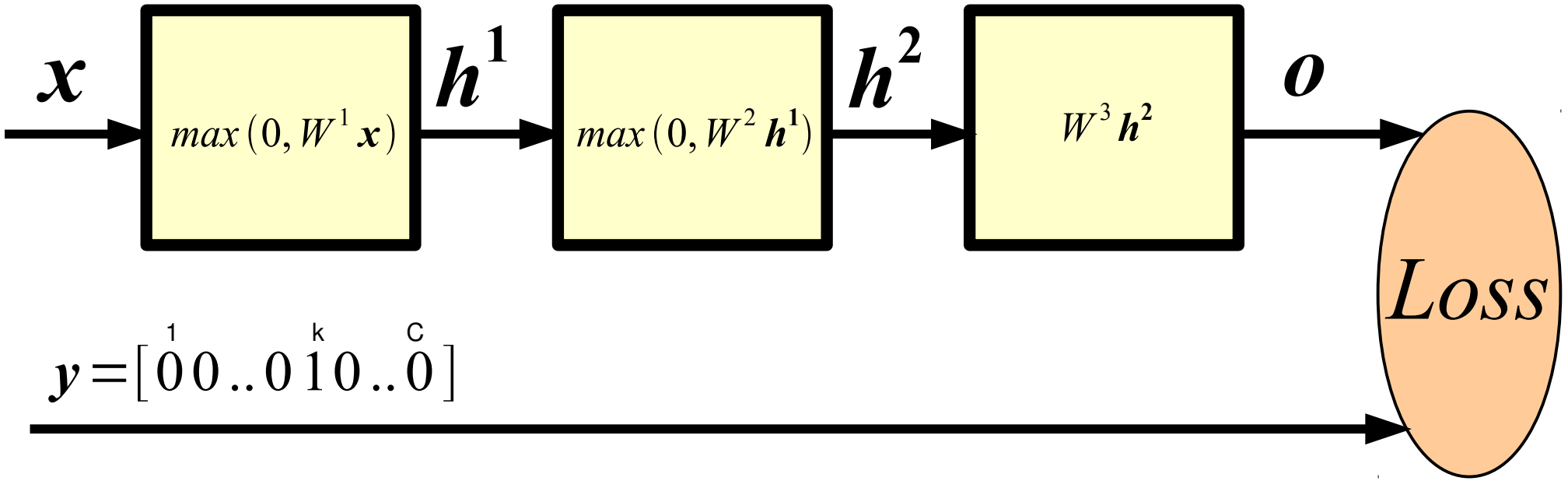
Question: How many layers? How many hidden units?

Answer: Cross-validation or hyper-parameter search methods are the answer. In general, the wider and the deeper the network the more complicated the mapping.

Question: How do I set the weight matrices?

Answer: Weight matrices and biases are learned.
First, we need to define a measure of quality of the current mapping.
Then, we need to define a procedure to adjust the parameters.

How Good is a Network?



Probability of class k given input (softmax):

$$p(c_k = 1|\boldsymbol{x}) = \frac{e^{o_k}}{\sum_{j=1}^{C} e^{o_j}}$$

(Per-sample) **Loss**; e.g., negative log-likelihood (good for classification of small number of classes):

$$L(\boldsymbol{x}, y; \boldsymbol{\theta}) = -\sum_j y_j \log p(c_j|\boldsymbol{x})$$

Training

Learning consists of minimizing the loss (plus some regularization term) w.r.t. parameters over the whole training set.

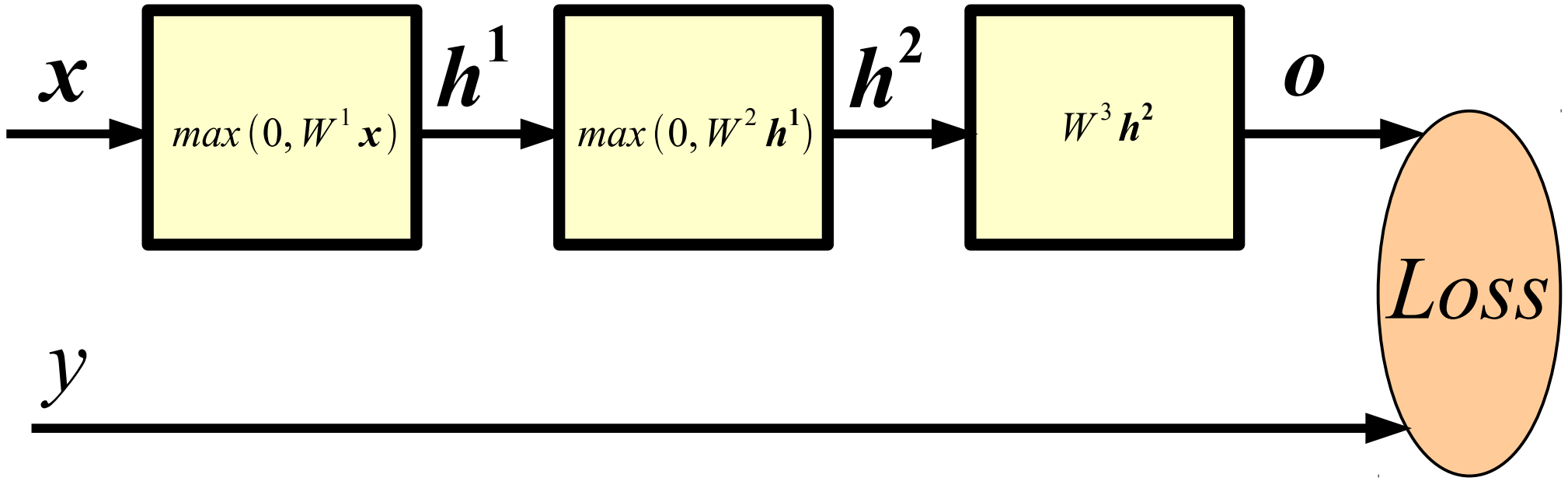
$$\boldsymbol{\theta}^* = arg\ min_{\boldsymbol{\theta}} \sum_{n=1}^{P} L\left(\boldsymbol{x}^{\boldsymbol{n}}, y^n; \boldsymbol{\theta}\right)$$

Question: How to minimize a complicated function of the parameters?

Answer: Chain rule, a.k.a. **Backpropagation**! That is the procedure to compute gradients of the loss w.r.t. parameters in a multi-layer neural network.

Rumelhart et al. “Learning internal representations by back-propagating..” Nature 1986

Key Idea: Wiggle To Decrease Loss



Let's say we want to decrease the loss by adjusting $W^1_{i,j}$.
We could consider a very small $\epsilon = 1\text{e-}6$ and compute:

$$L(\boldsymbol{x}, y; \boldsymbol{\theta})$$

$$L(\boldsymbol{x}, y; \boldsymbol{\theta} \setminus W^1_{i,j}, W^1_{i,j} + \epsilon)$$

Then, update:

$$W^1_{i,j} \leftarrow W^1_{i,j} + \epsilon \, sgn\left(L(\boldsymbol{x}, y; \boldsymbol{\theta}) - L(\boldsymbol{x}, y; \boldsymbol{\theta} \setminus W^1_{i,j}, W^1_{i,j} + \epsilon) \right)$$

Derivative w.r.t. Input of Softmax

$$p(c_k = 1|\boldsymbol{x}) = \frac{e^{o_k}}{\sum_j e^{o_j}}$$

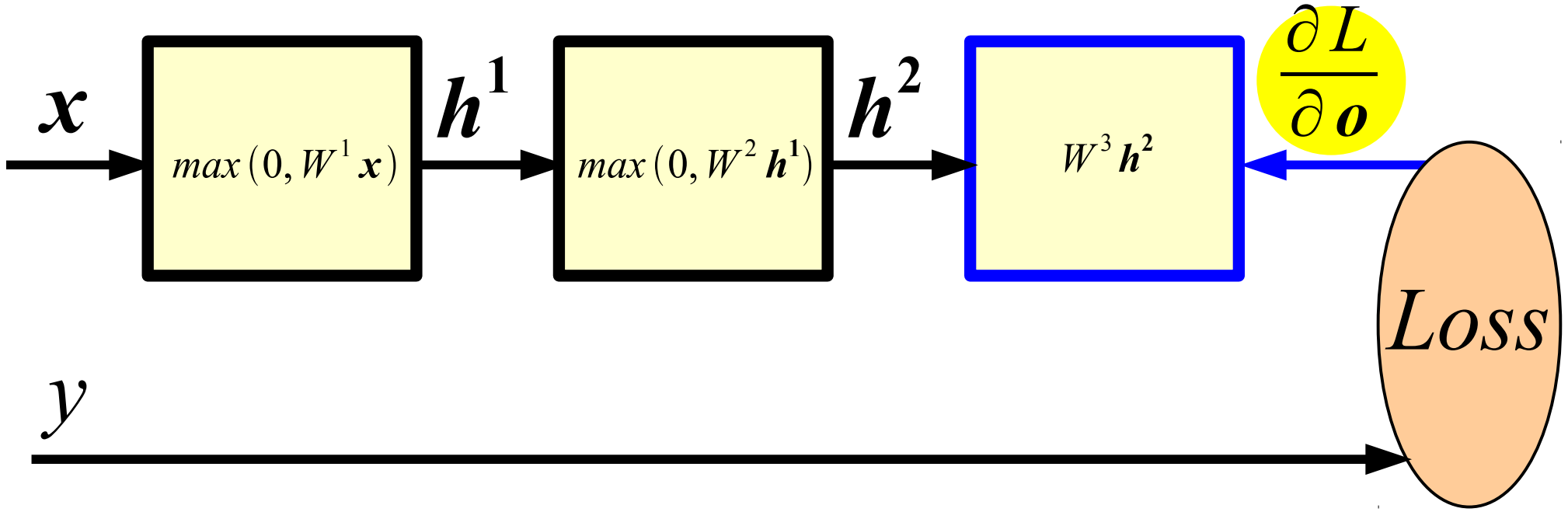
$$L(\boldsymbol{x}, y; \boldsymbol{\theta}) = -\sum_j y_j \log p(c_j|\boldsymbol{x}) \qquad \boldsymbol{y} = [\overset{1}{0}\,0\,..\,0\,\overset{k}{1}\,0\,..\,\overset{C}{0}]$$

By substituting the fist formula in the second, and taking the derivative w.r.t. $\boldsymbol{o}$ we get:

$$\frac{\partial L}{\partial o} = p(c|\boldsymbol{x}) - \boldsymbol{y}$$

HOMEWORK: prove it!

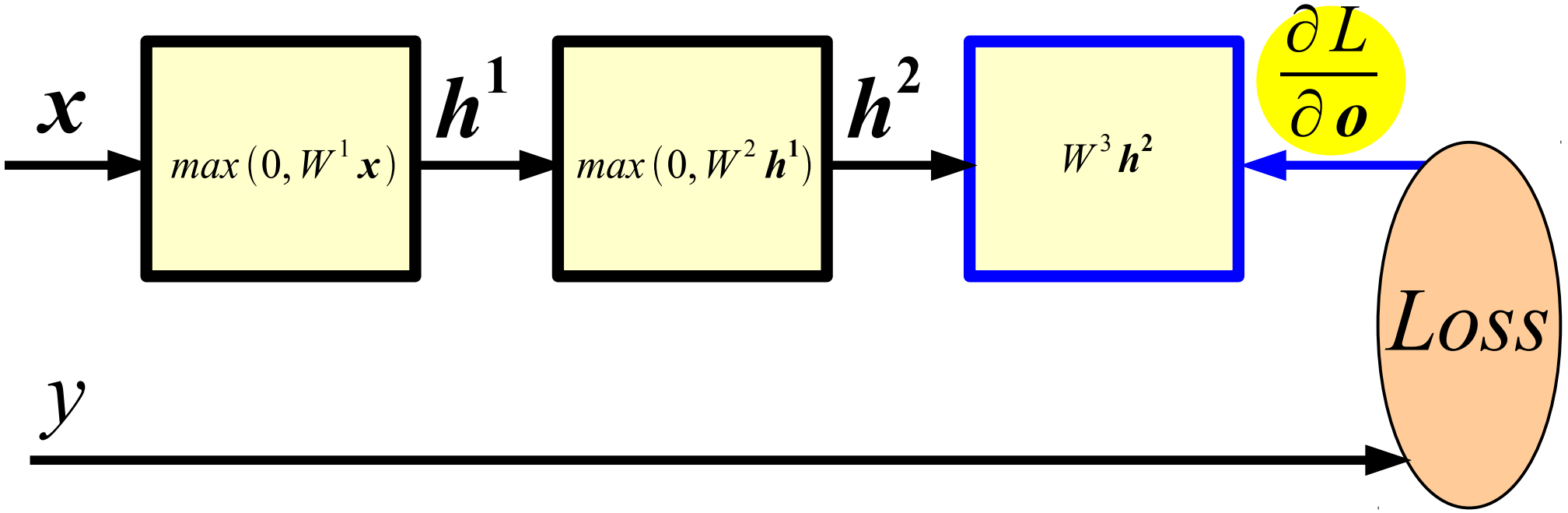
Backward Propagation



Given $\partial L / \partial \boldsymbol{o}$ and assuming we can easily compute the Jacobian of each module, we have:

$$\frac{\partial L}{\partial W^3} = \frac{\partial L}{\partial \boldsymbol{o}} \frac{\partial \boldsymbol{o}}{\partial W^3} \qquad \frac{\partial L}{\partial \boldsymbol{h}^2} = \frac{\partial L}{\partial \boldsymbol{o}} \frac{\partial \boldsymbol{o}}{\partial \boldsymbol{h}^2}$$

Backward Propagation

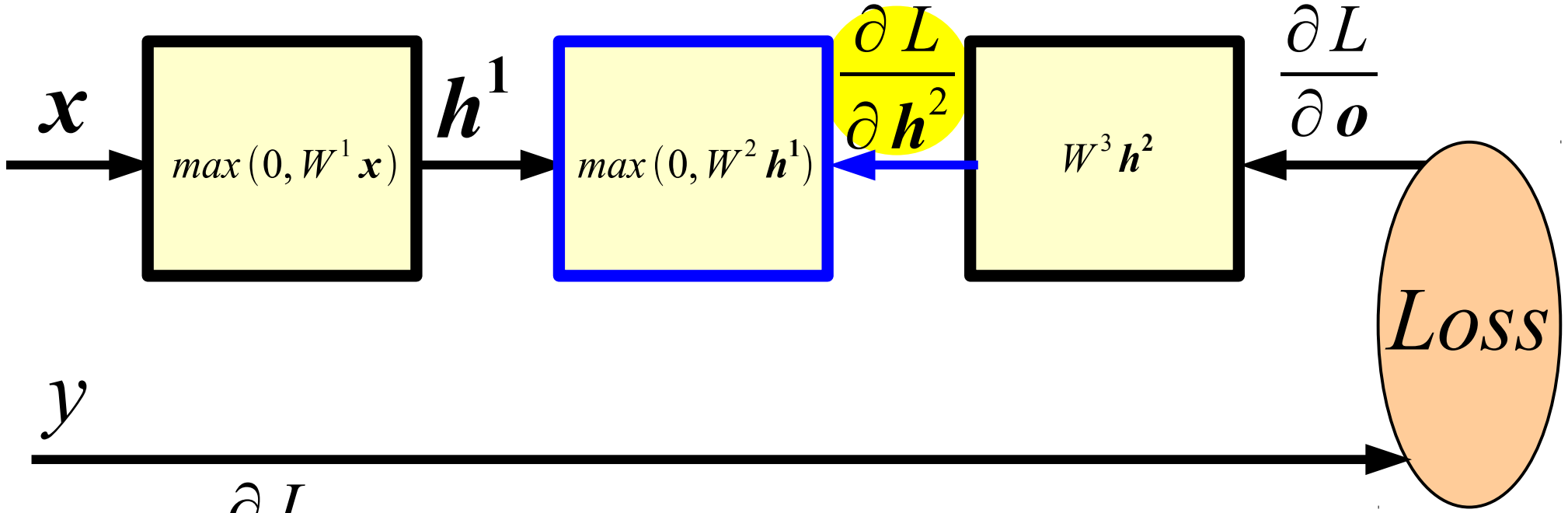


Given $\partial L / \partial \boldsymbol{o}$ and assuming we can easily compute the Jacobian of each module, we have:

$$\frac{\partial L}{\partial W^3} = \frac{\partial L}{\partial \boldsymbol{o}} \frac{\partial \boldsymbol{o}}{\partial W^3} \qquad \frac{\partial L}{\partial \boldsymbol{h}^2} = \frac{\partial L}{\partial \boldsymbol{o}} \frac{\partial \boldsymbol{o}}{\partial \boldsymbol{h}^2}$$

$$\frac{\partial L}{\partial W^3} = (p(c|\boldsymbol{x}) - \boldsymbol{y})\ \boldsymbol{h}^{2\,T} \qquad \frac{\partial L}{\partial \boldsymbol{h}^2} = W^{3\,T}(p(c|\boldsymbol{x}) - \boldsymbol{y})$$

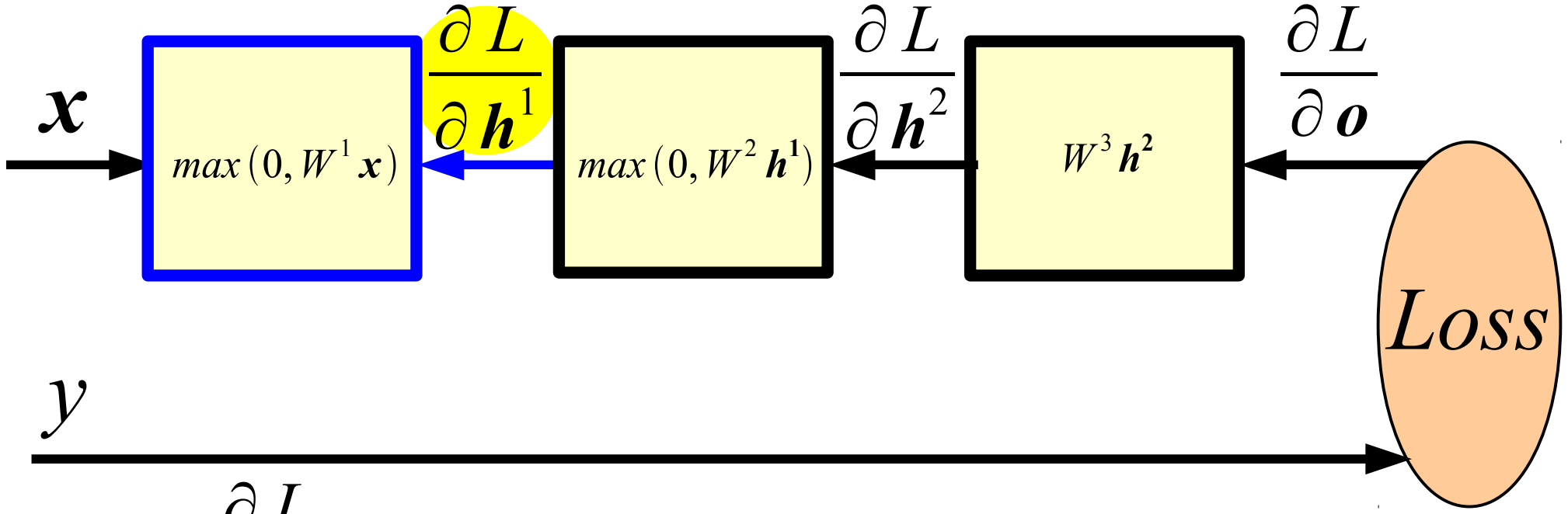
Backward Propagation



Given $\frac{\partial L}{\partial h^2}$ we can compute now:

$$\frac{\partial L}{\partial W^2} = \frac{\partial L}{\partial h^2} \frac{\partial h^2}{\partial W^2} \qquad \frac{\partial L}{\partial h^1} = \frac{\partial L}{\partial h^2} \frac{\partial h^2}{\partial h^1}$$

Backward Propagation



Given $\frac{\partial L}{\partial h^1}$ we can compute now:

$$\frac{\partial L}{\partial W^1} = \frac{\partial L}{\partial h^1} \frac{\partial h^1}{\partial W^1}$$

Backward Propagation

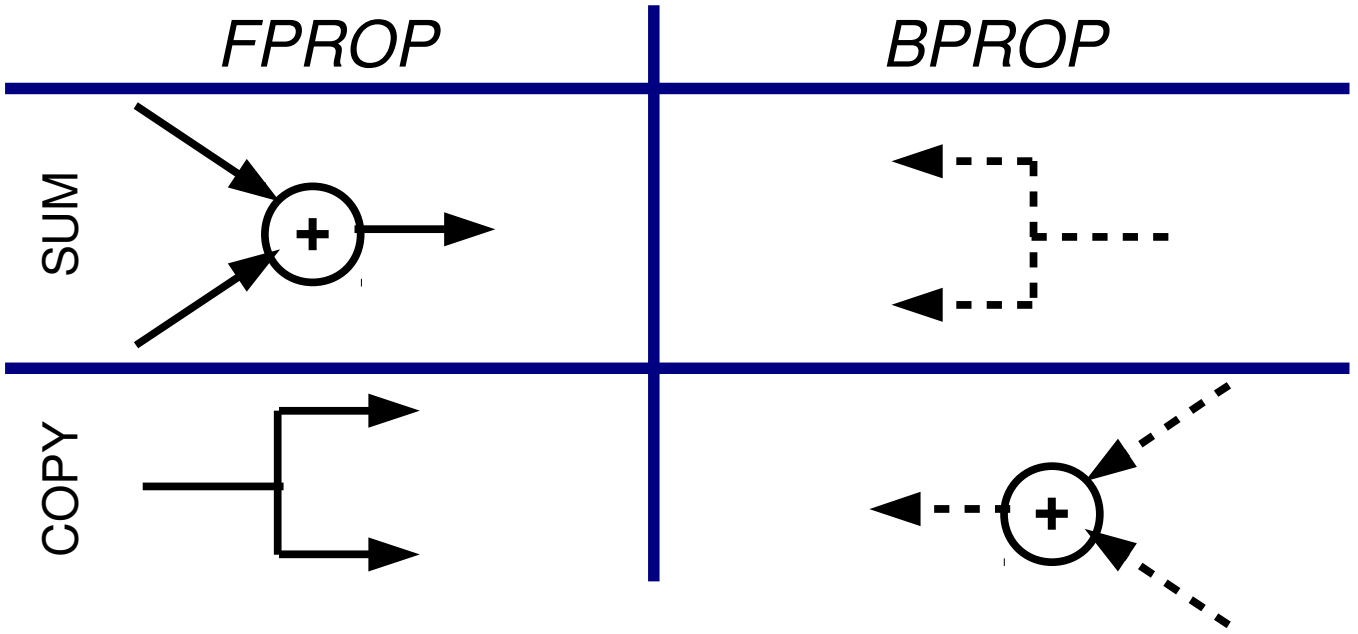
Question: Does BPROP work with ReLU layers only?

Answer: Nope, any a.e. differentiable transformation works.

Question: What's the computational cost of BPROP?

Answer: About twice FPROP (need to compute gradients w.r.t. input and parameters at every layer).

Note: FPROP and BPROP are dual of each other. E.g.,:



Optimization

Stochastic Gradient Descent (on mini-batches):

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{\partial L}{\partial \boldsymbol{\theta}}, \eta \in (0,1)$$

Stochastic Gradient Descent with Momentum:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \boldsymbol{\Delta}$$

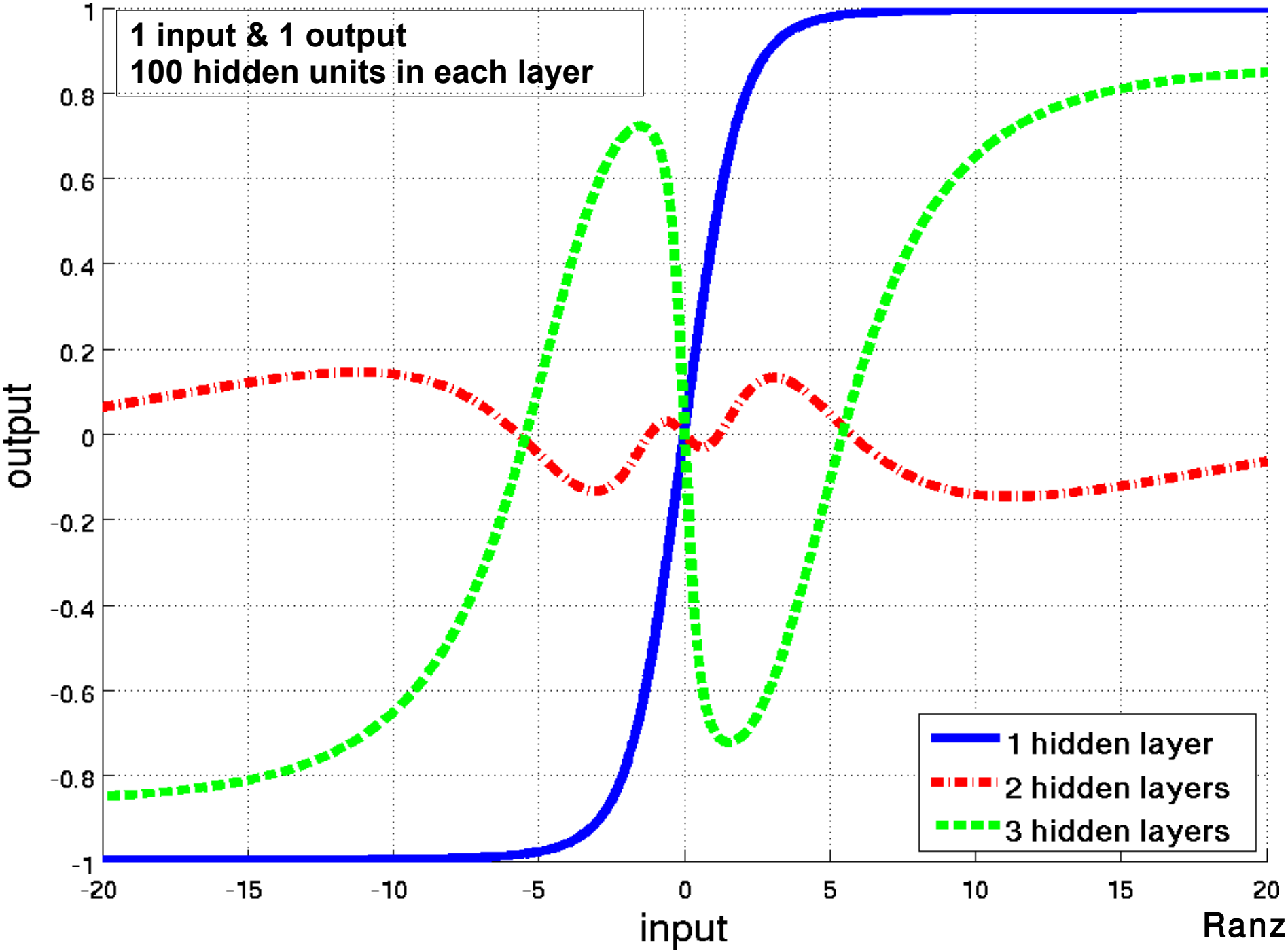
$$\boldsymbol{\Delta} \leftarrow 0.9 \boldsymbol{\Delta} + \frac{\partial L}{\partial \boldsymbol{\theta}}$$

Note: there are many other variants...

Toy Code (Matlab): Neural Net Trainer

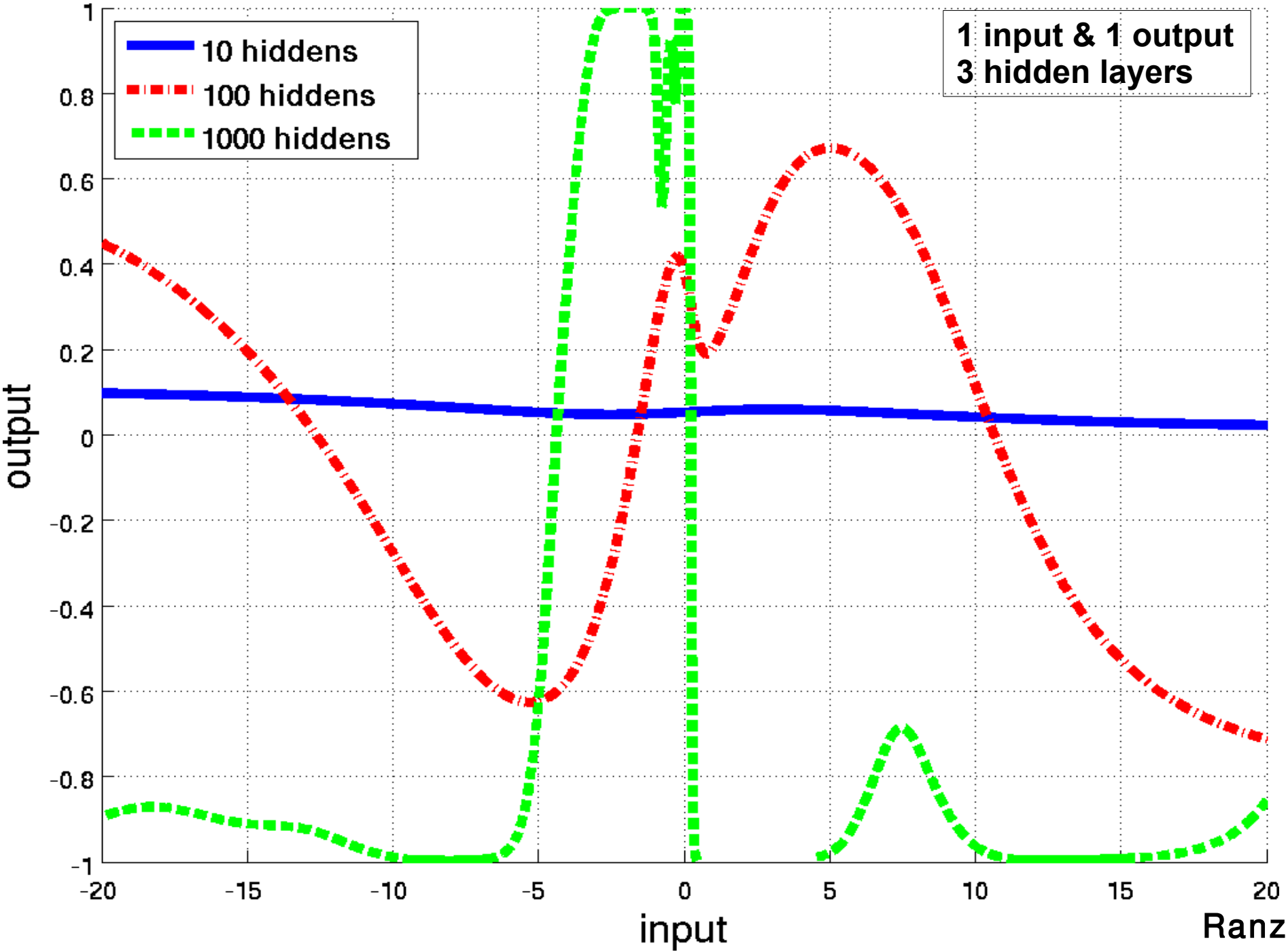
```
% F-PROP
for i = 1 : nr_layers - 1
  [h{i}  jac{i}]  =  nonlinearity(W{i} * h{i-1} +  b{i});
end
h{nr_layers-1}  =  W{nr_layers-1} * h{nr_layers-2}  +  b{nr_layers-1};
prediction  =  softmax(h{l-1});


% CROSS ENTROPY LOSS
loss  =  -  sum(sum(log(prediction)  .*  target)) / batch_size;


% B-PROP
dh{l-1}  =  prediction  -  target;
for i = nr_layers - 1 : -1 : 1
  Wgrad{i}  =  dh{i} * h{i-1}';
  bgrad{i}  =  sum(dh{i}, 2);
  dh{i-1}  =  (W{i}' * dh{i})  .*  jac{i-1};
end


% UPDATE
for i = 1 : nr_layers - 1
  W{i}  =  W{i}  -  (lr / batch_size)  *  Wgrad{i};
  b{i}  =  b{i}  -  (lr / batch_size)  *  bgrad{i};
end
```

Toy Example: Synthetic Data

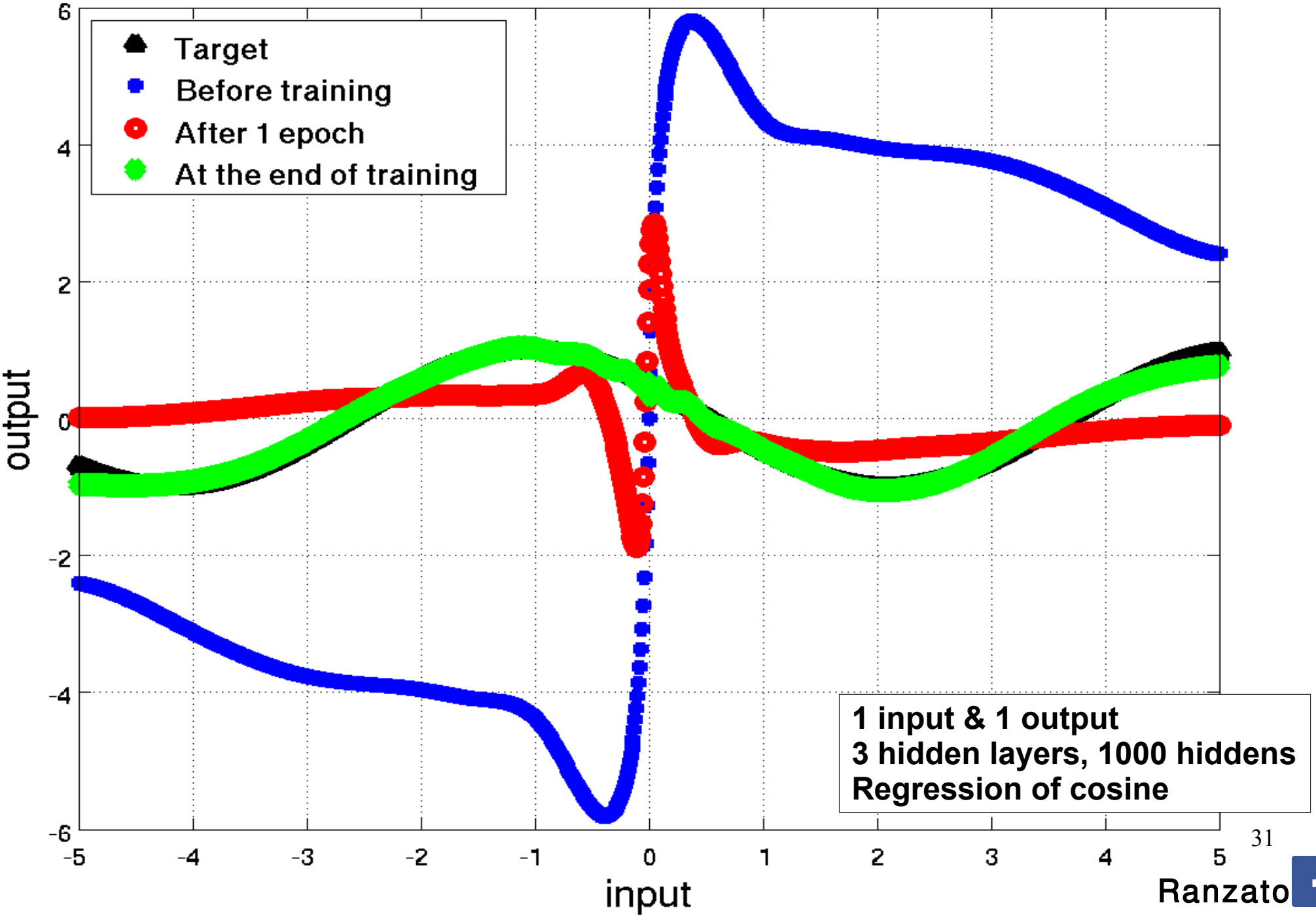


Ranzato

Toy Example: Synthetic Data



Toy Example: Synthetic Data

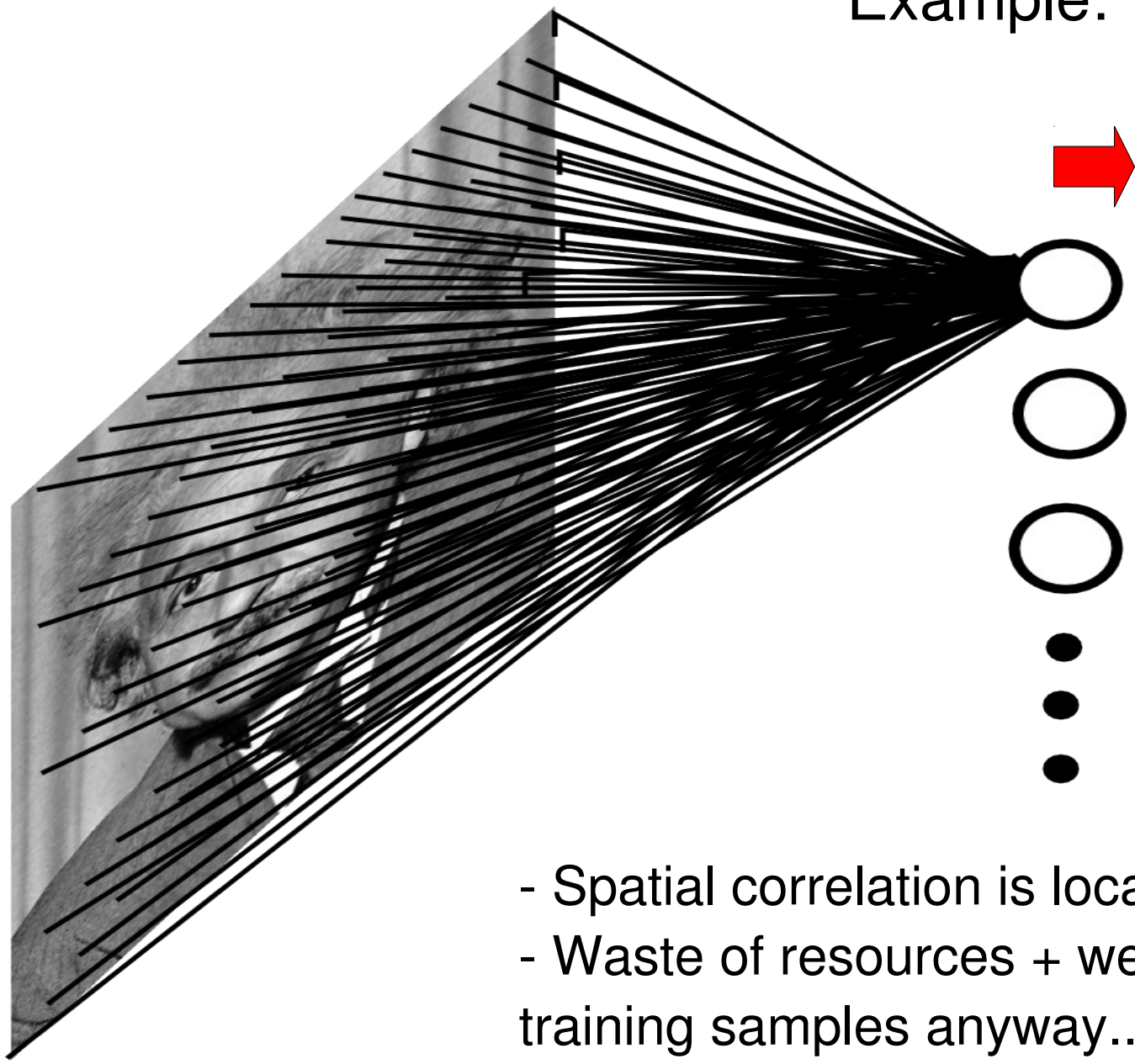


1 input & 1 output
3 hidden layers, 1000 hiddens
Regression of cosine

Outline

- Supervised Neural Networks
- Convolutional Neural Networks
- Examples
- Tips

Fully Connected Layer



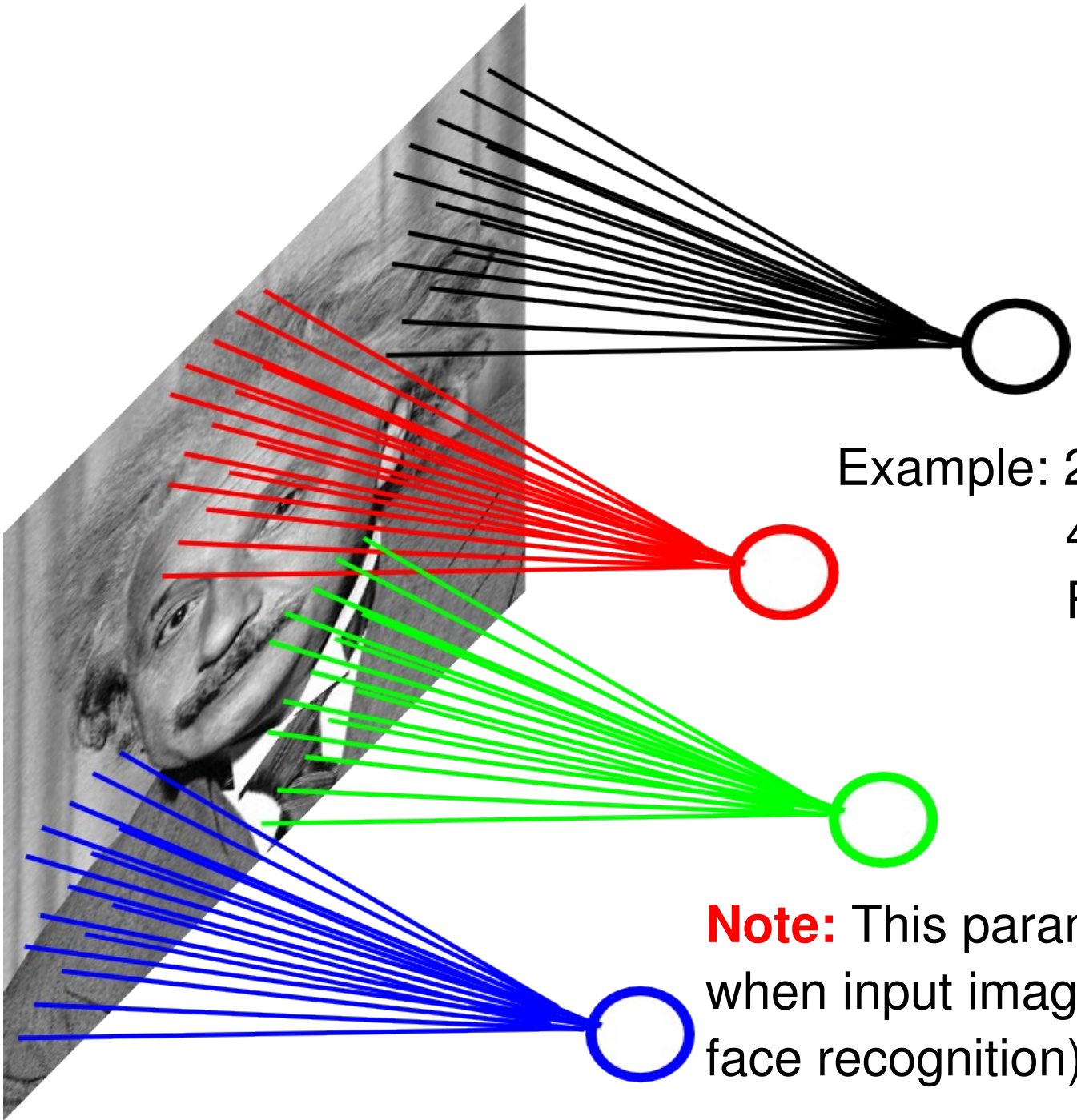
Example: 200x200 image

40K hidden units

~2B parameters!!!

- Spatial correlation is local
- Waste of resources + we have not enough training samples anyway..

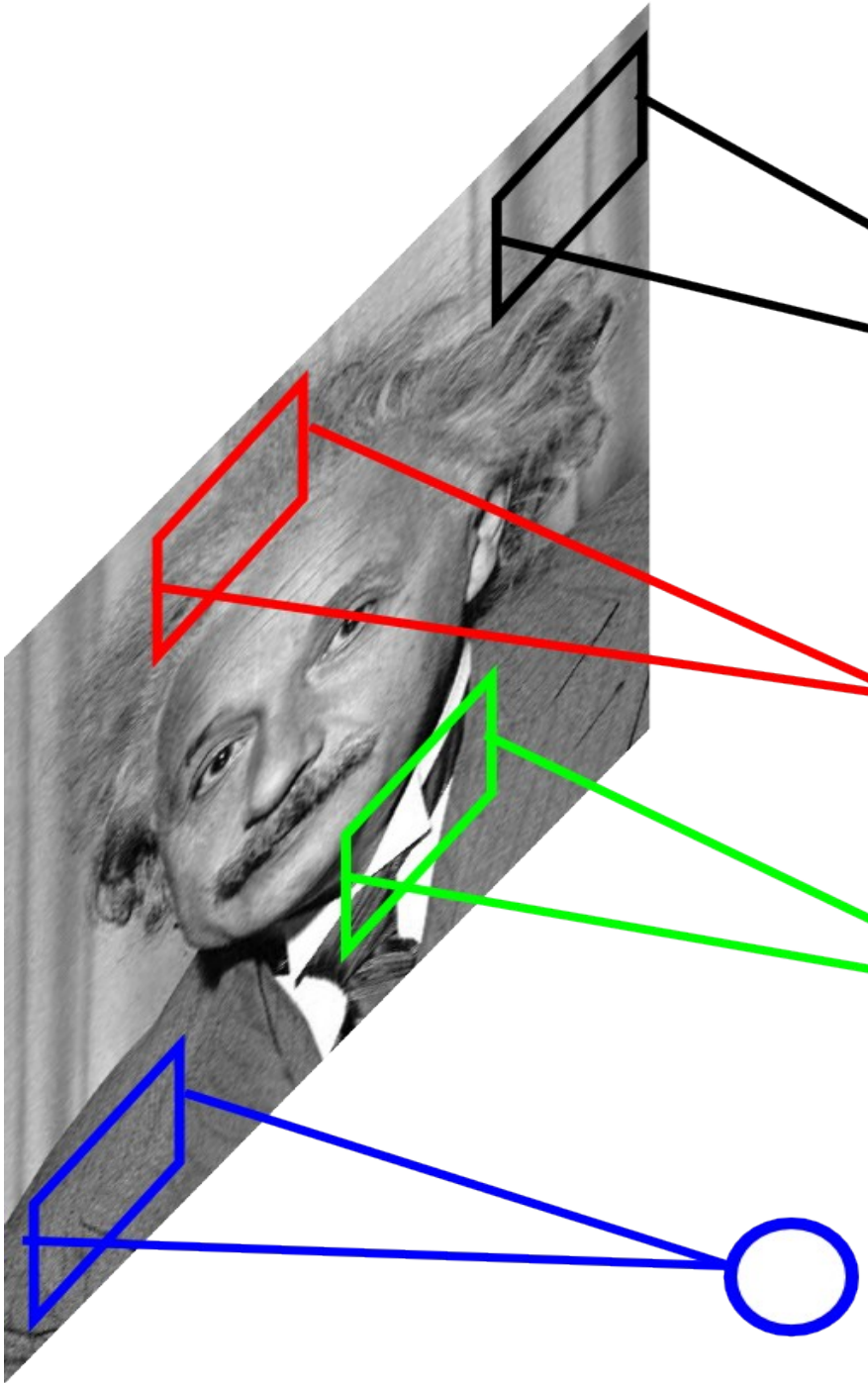
Locally Connected Layer



Example: 200x200 image
40K hidden units
Filter size: 10x10
4M parameters

Note: This parameterization is good when input image is registered (e.g., face recognition).

Locally Connected Layer

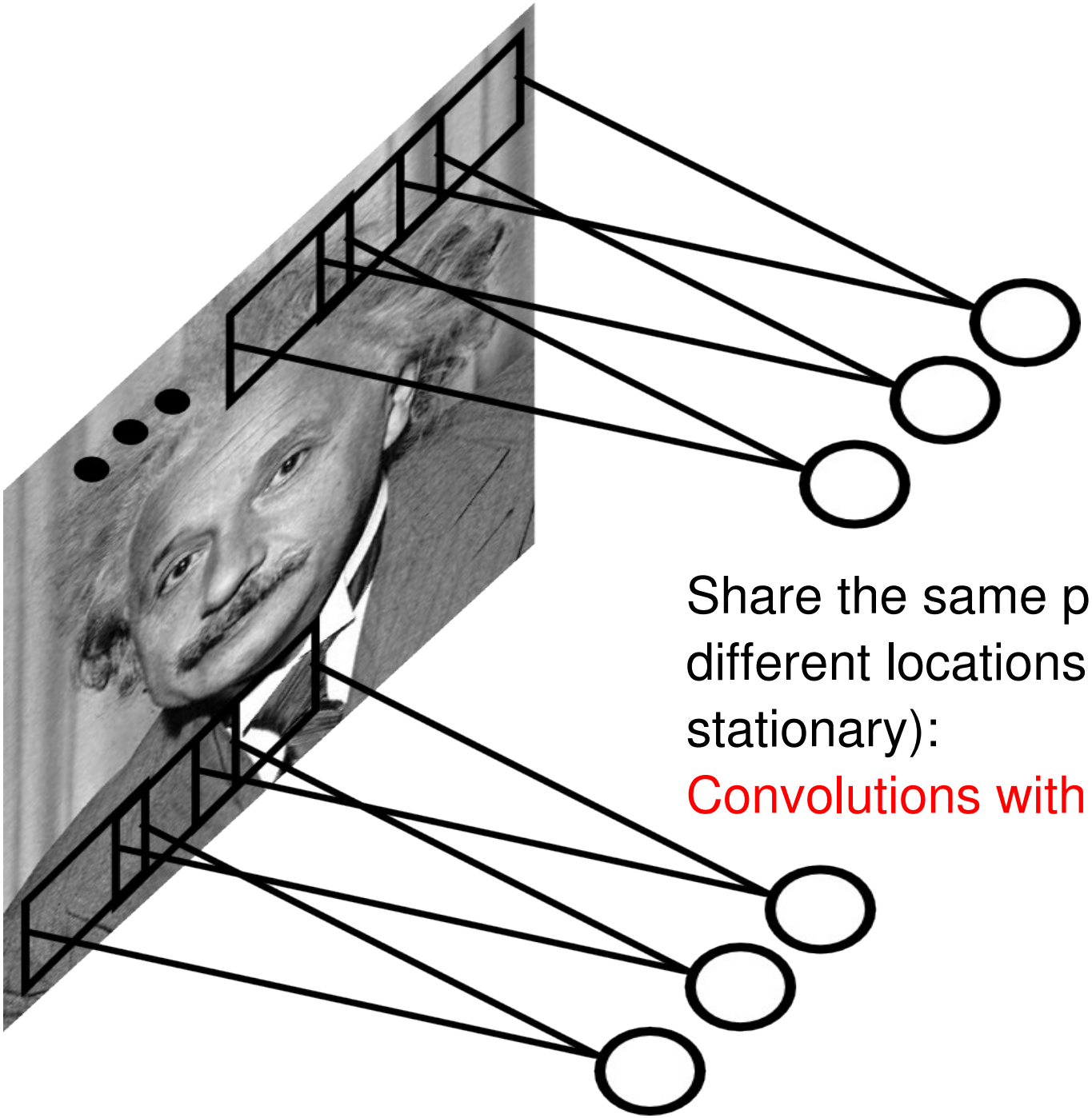


STATIONARITY? Statistics is similar at different locations

Example: 200x200 image
40K hidden units
Filter size: 10x10
4M parameters

Note: This parameterization is good when input image is registered (e.g., face recognition).

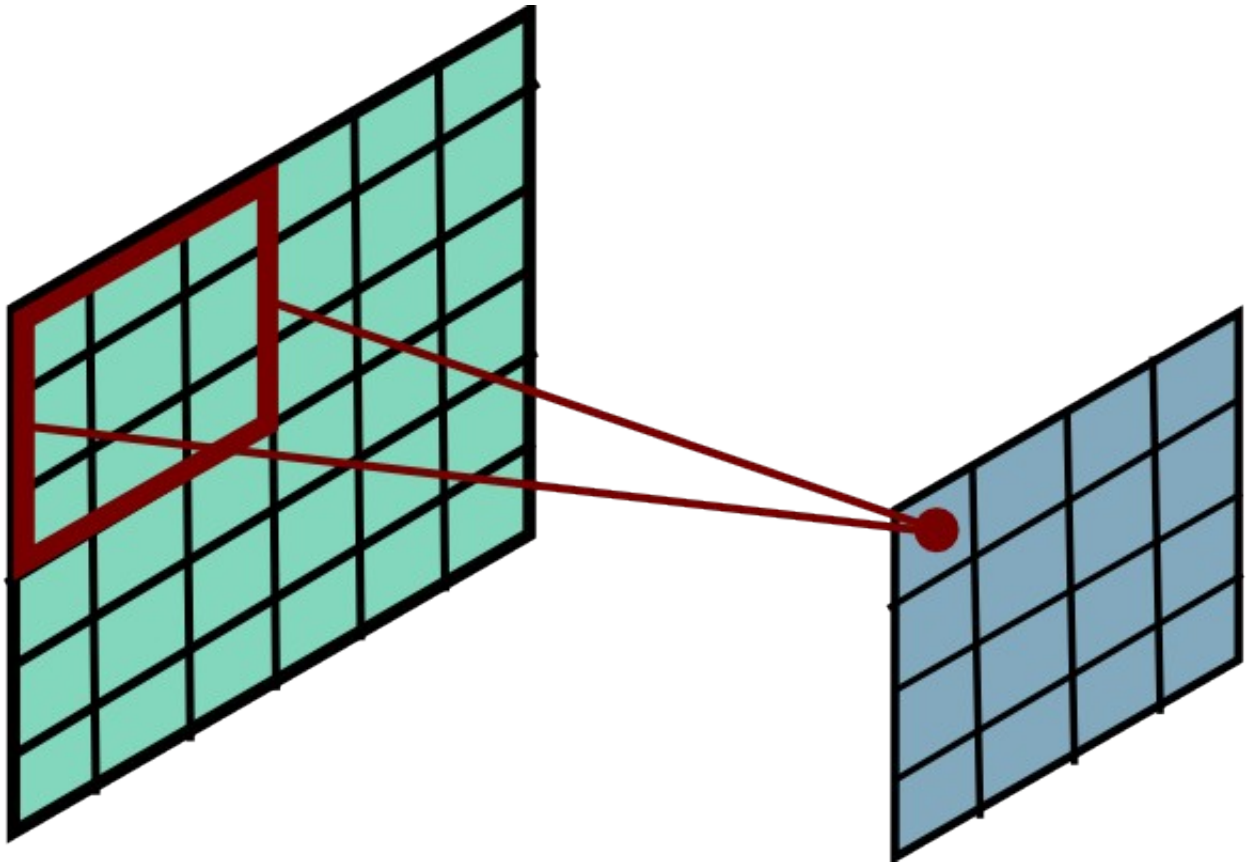
Convolutional Layer



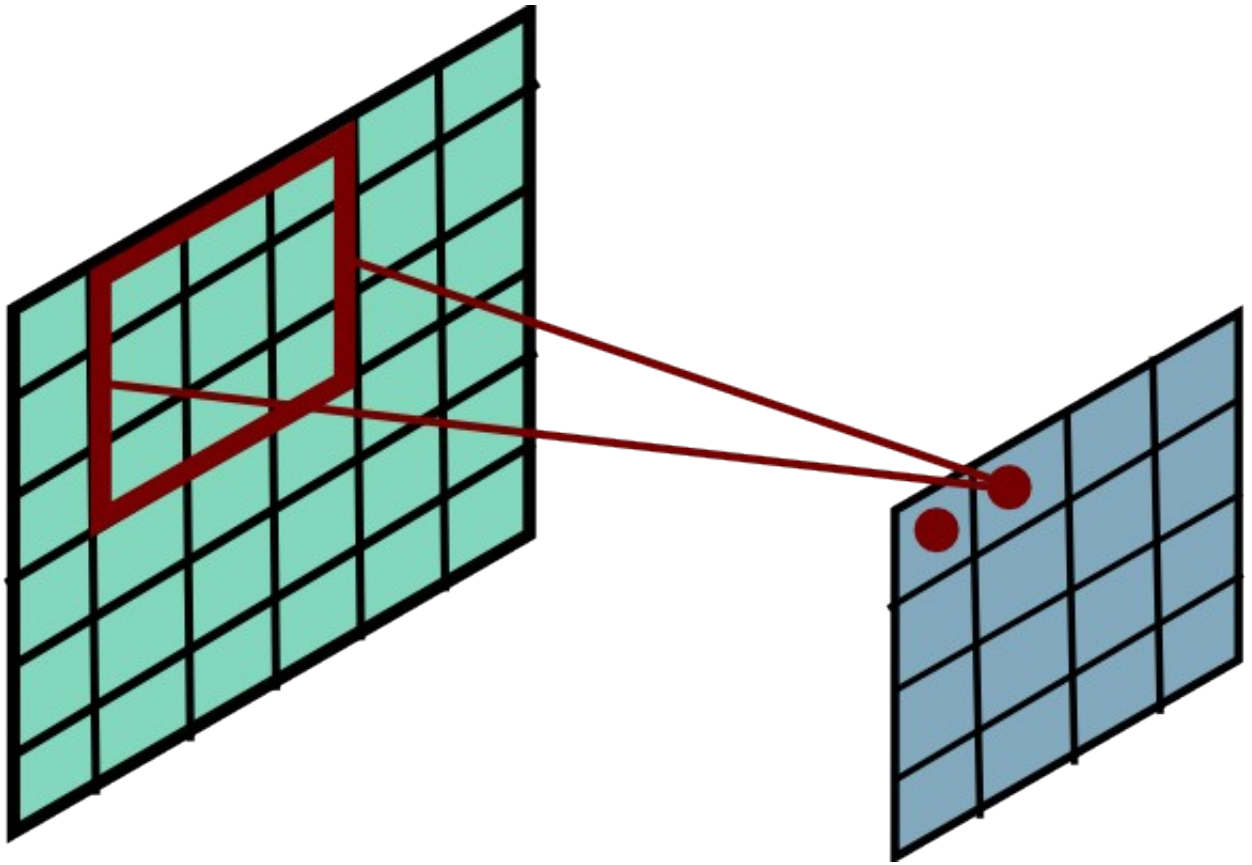
Share the same parameters across different locations (assuming input is stationary):

Convolutions with learned kernels

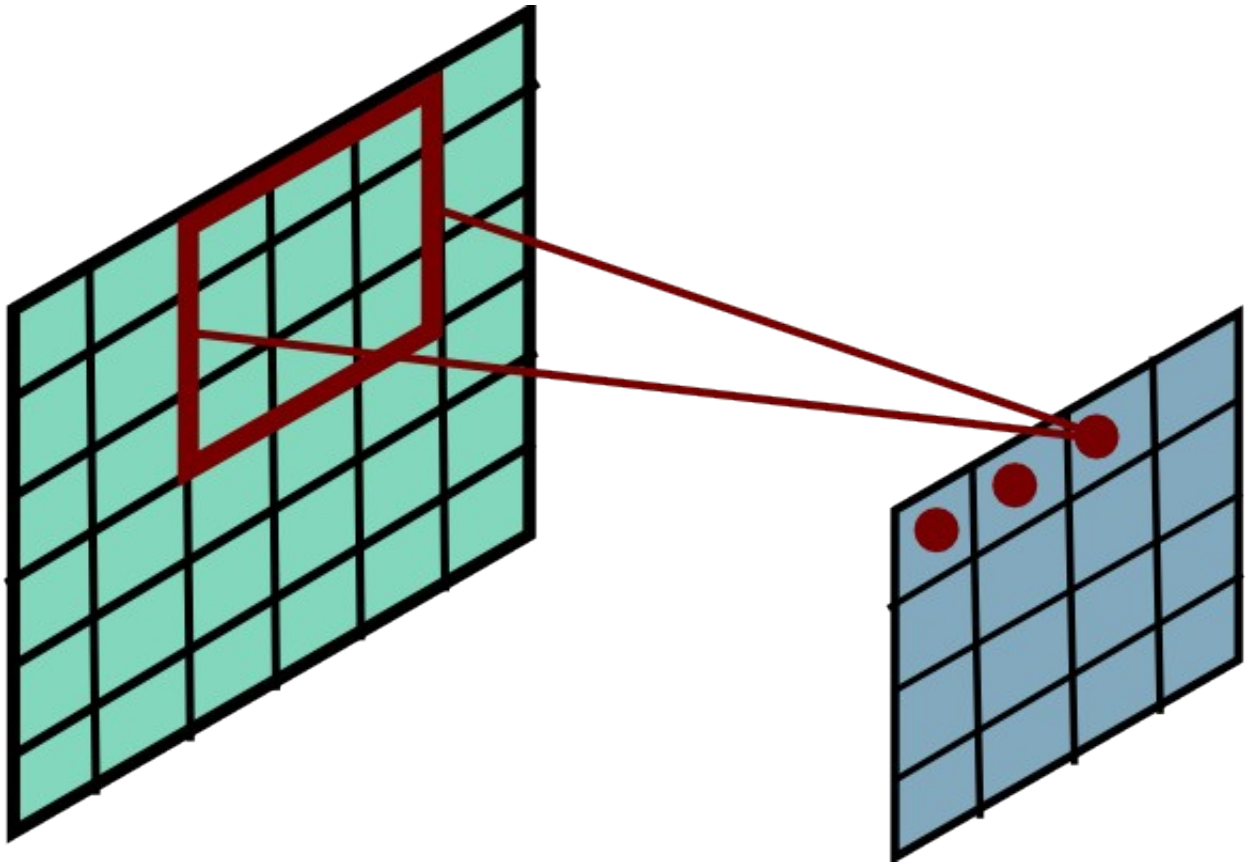
Convolutional Layer



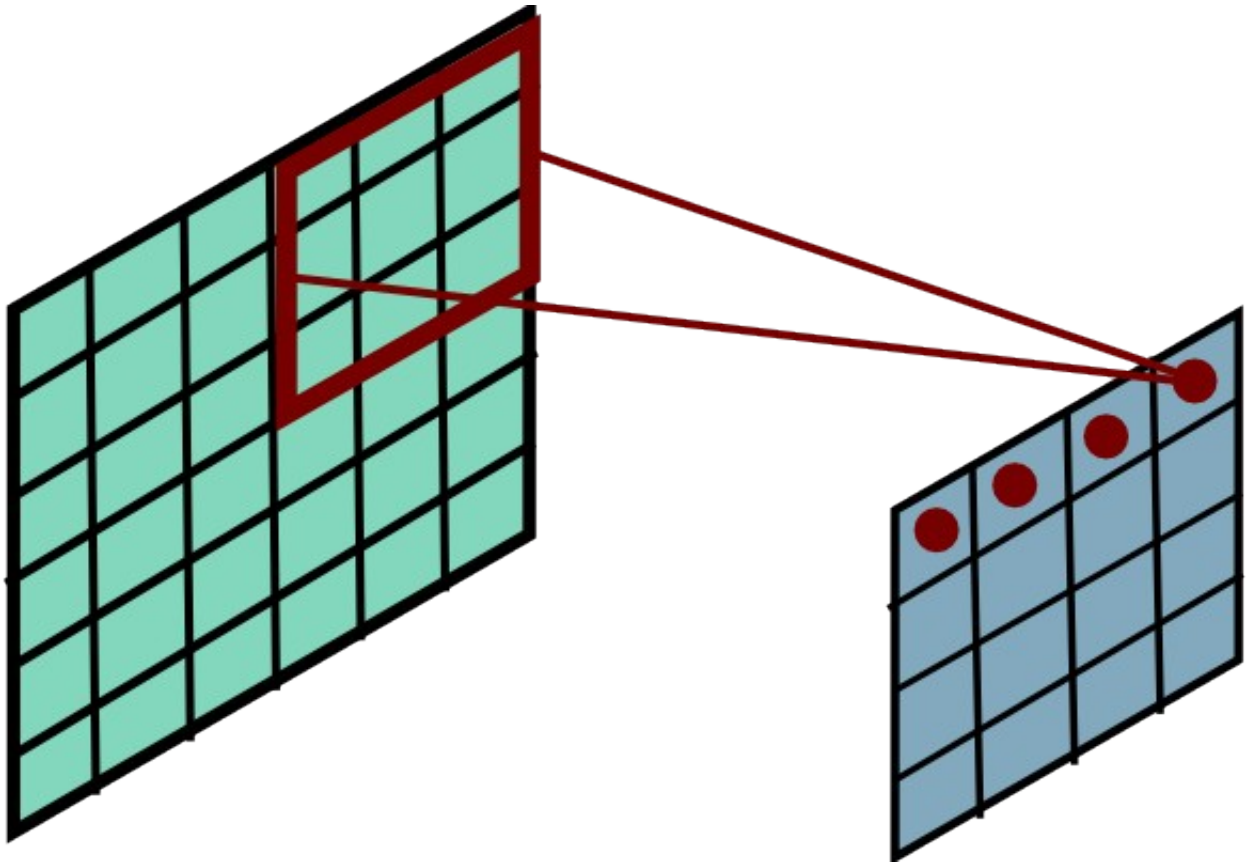
Convolutional Layer



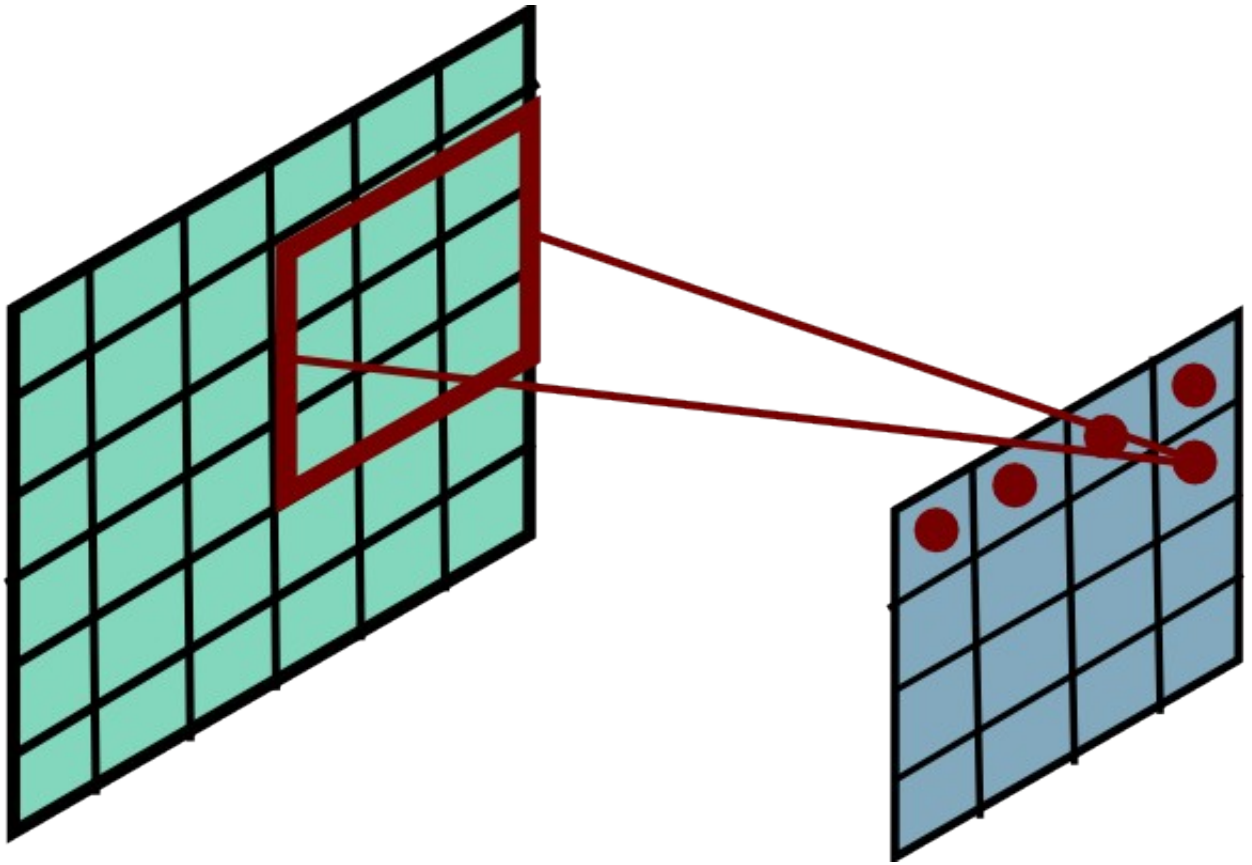
Convolutional Layer



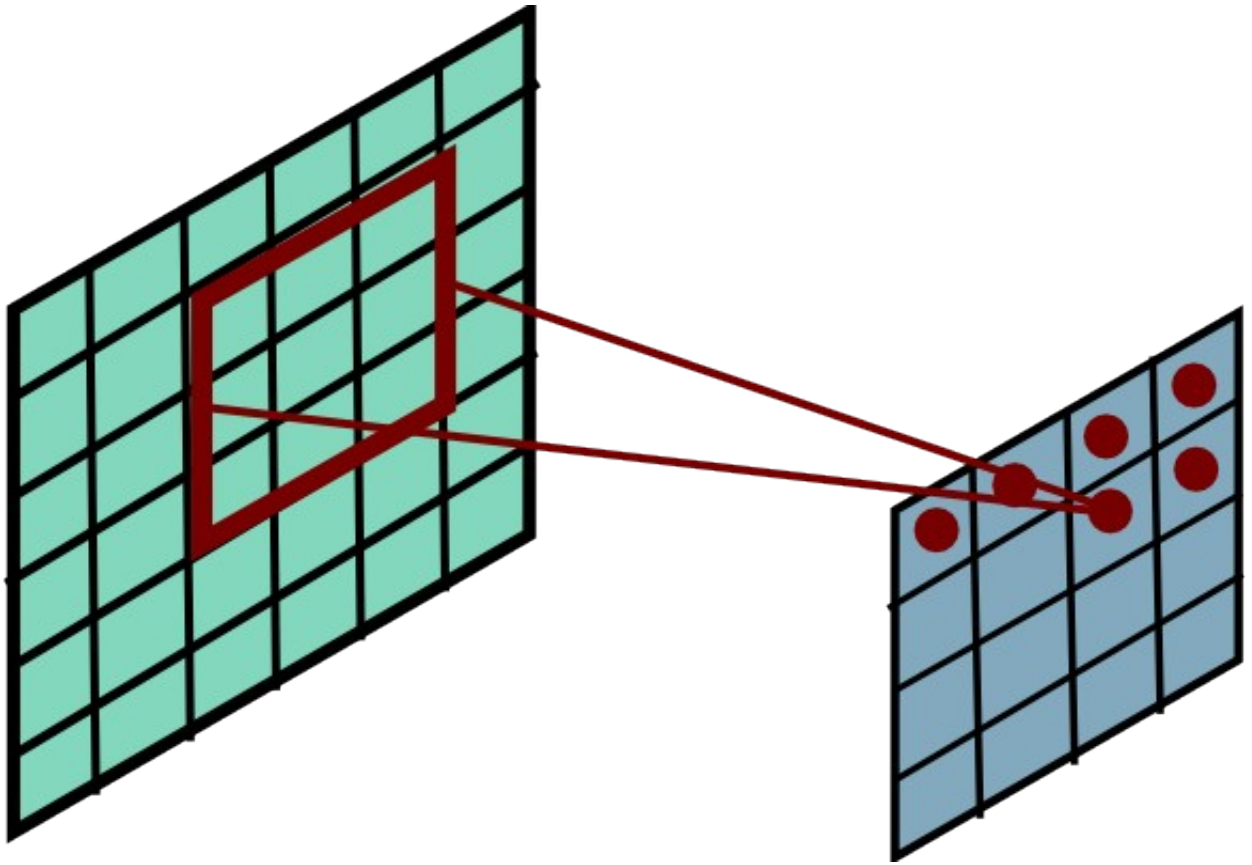
Convolutional Layer



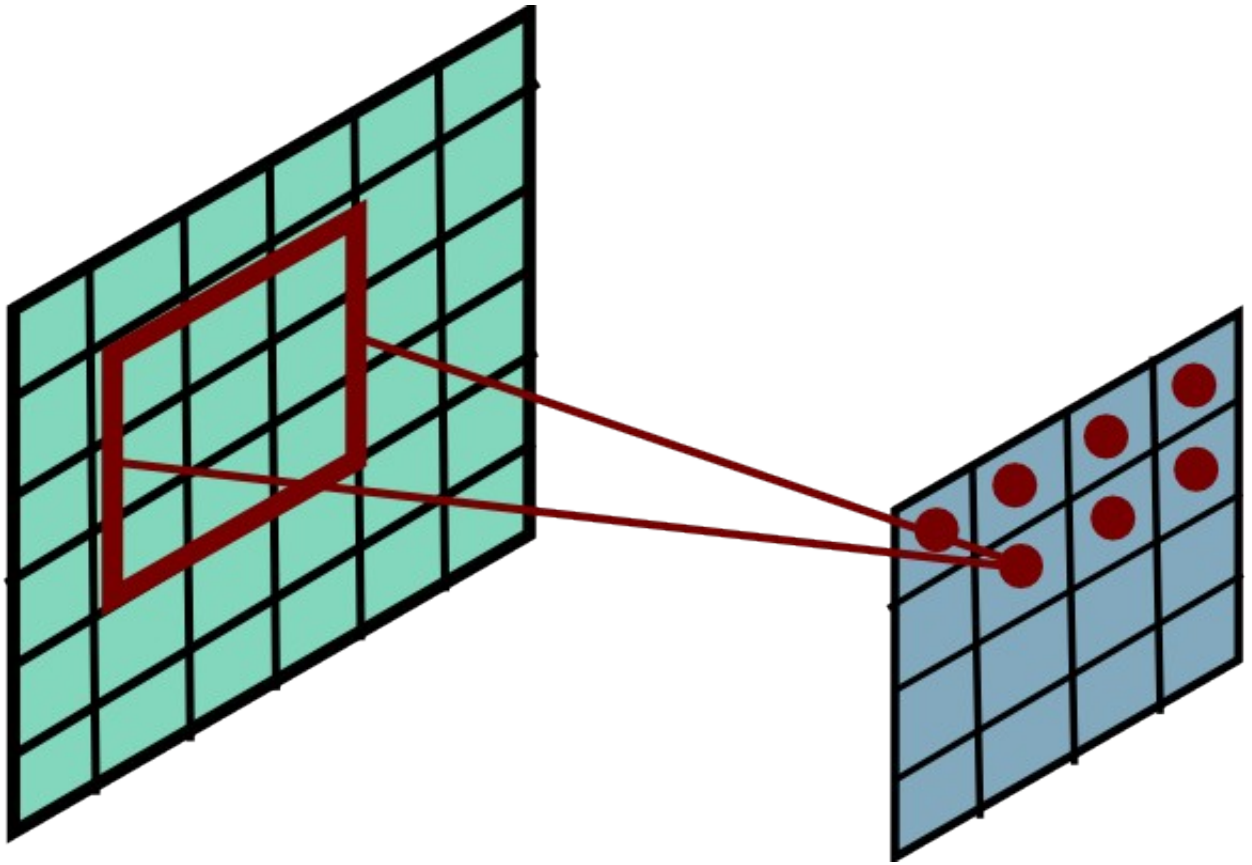
Convolutional Layer



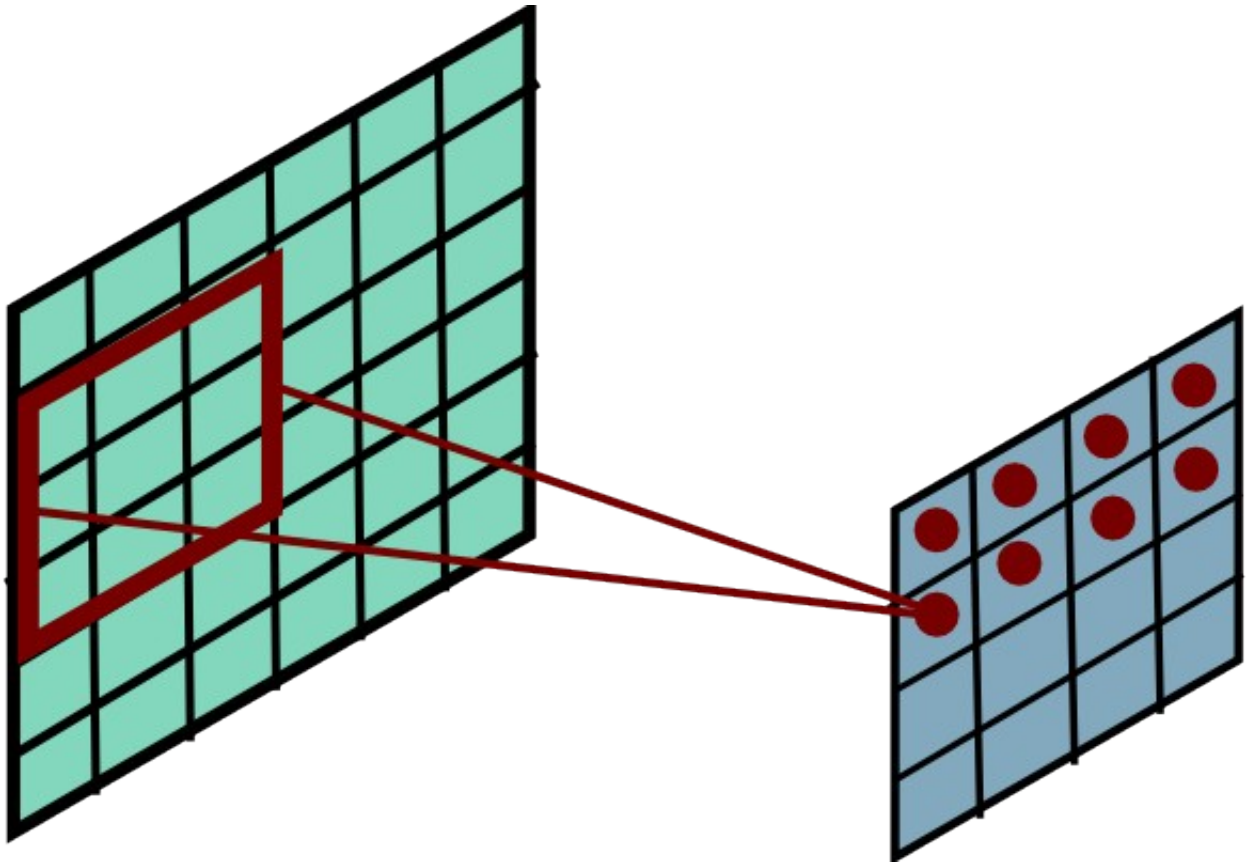
Convolutional Layer



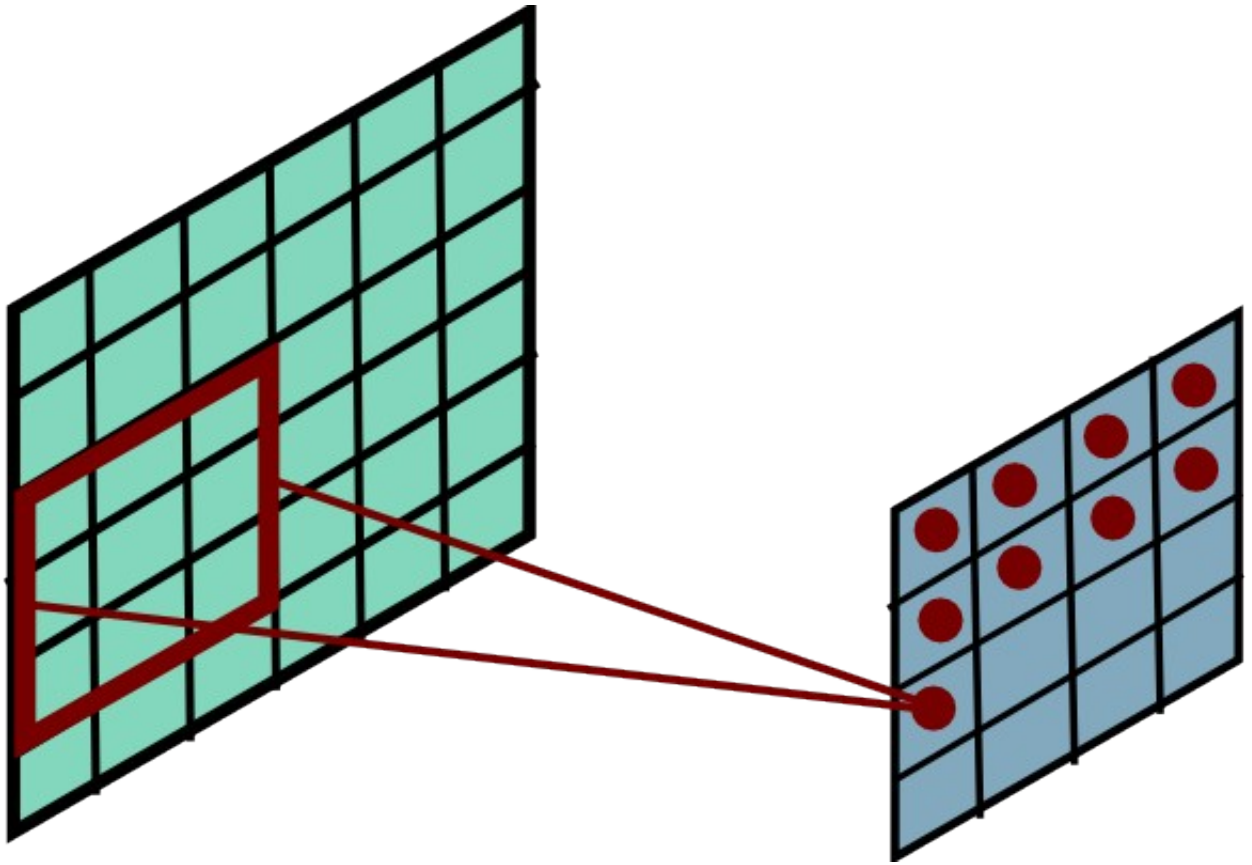
Convolutional Layer



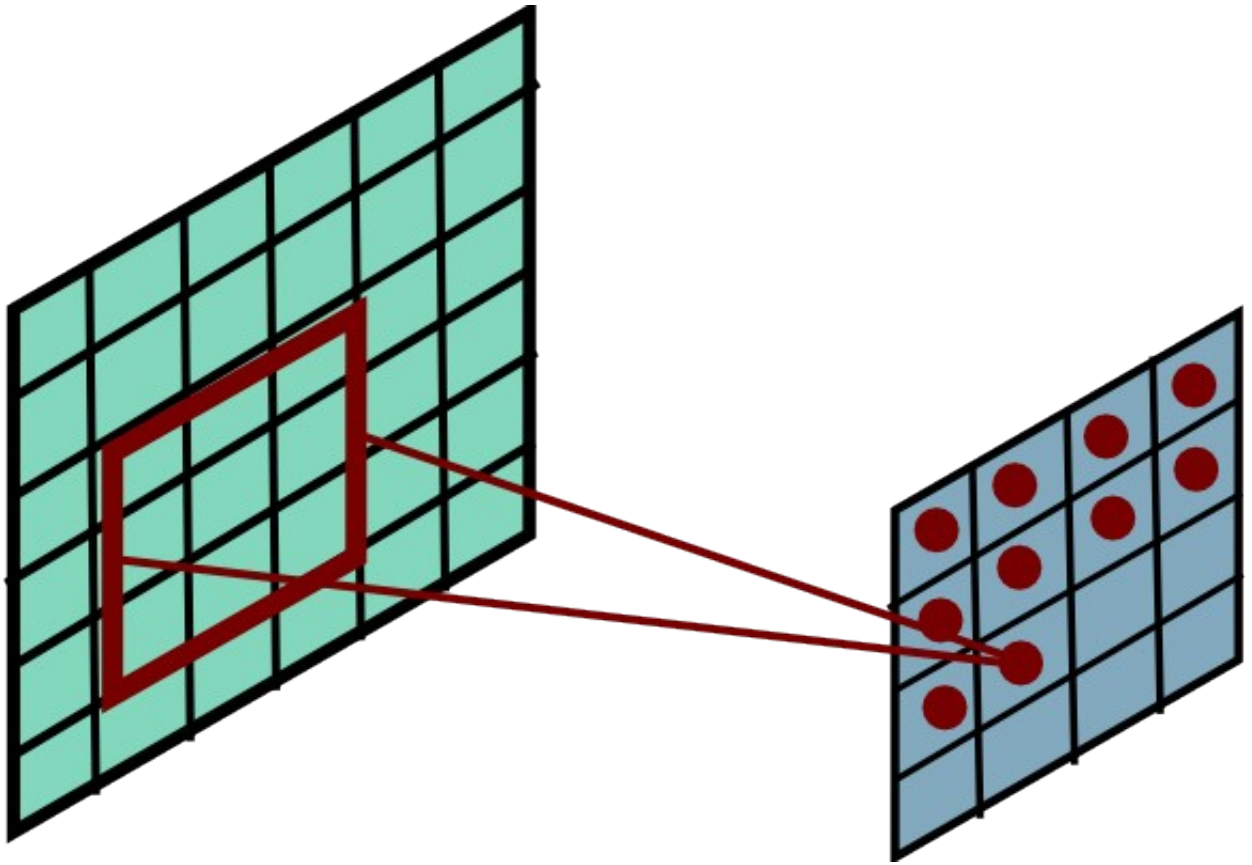
Convolutional Layer



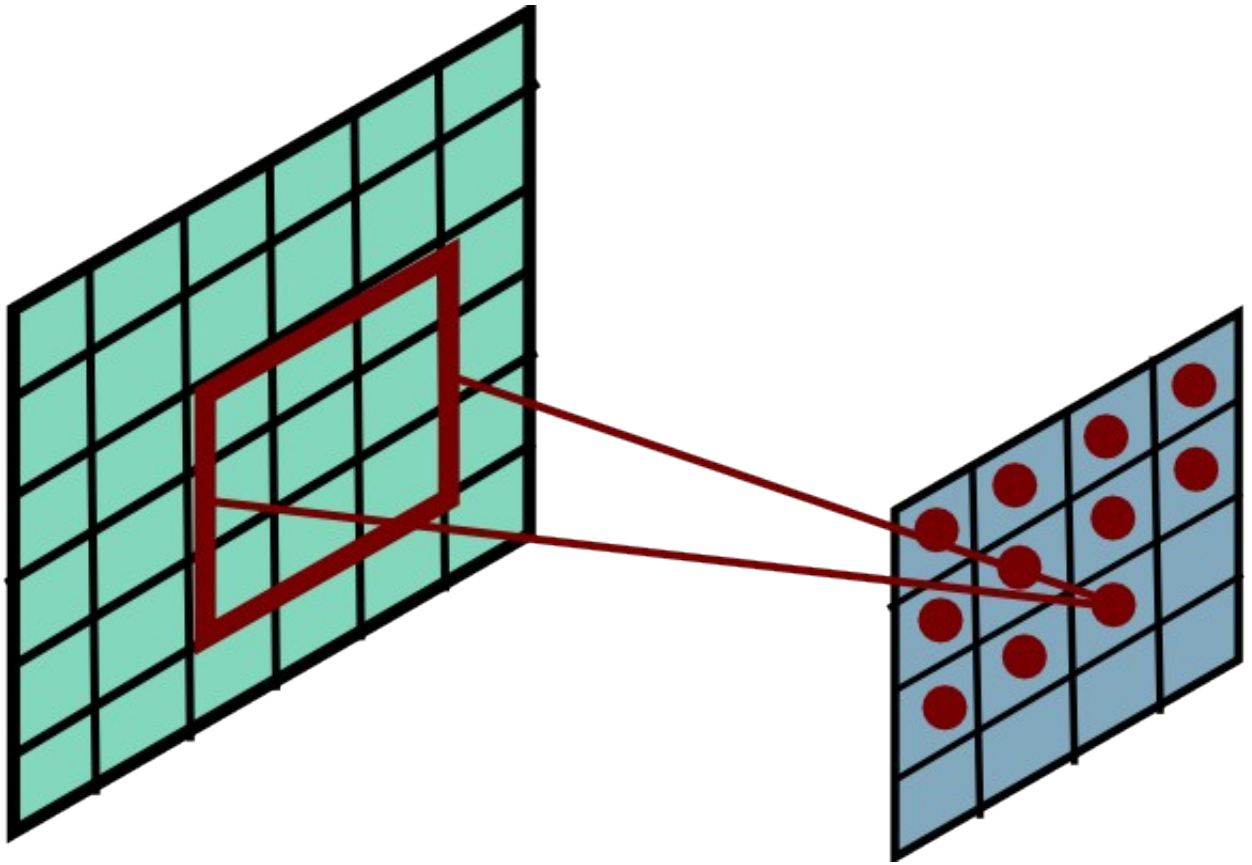
Convolutional Layer



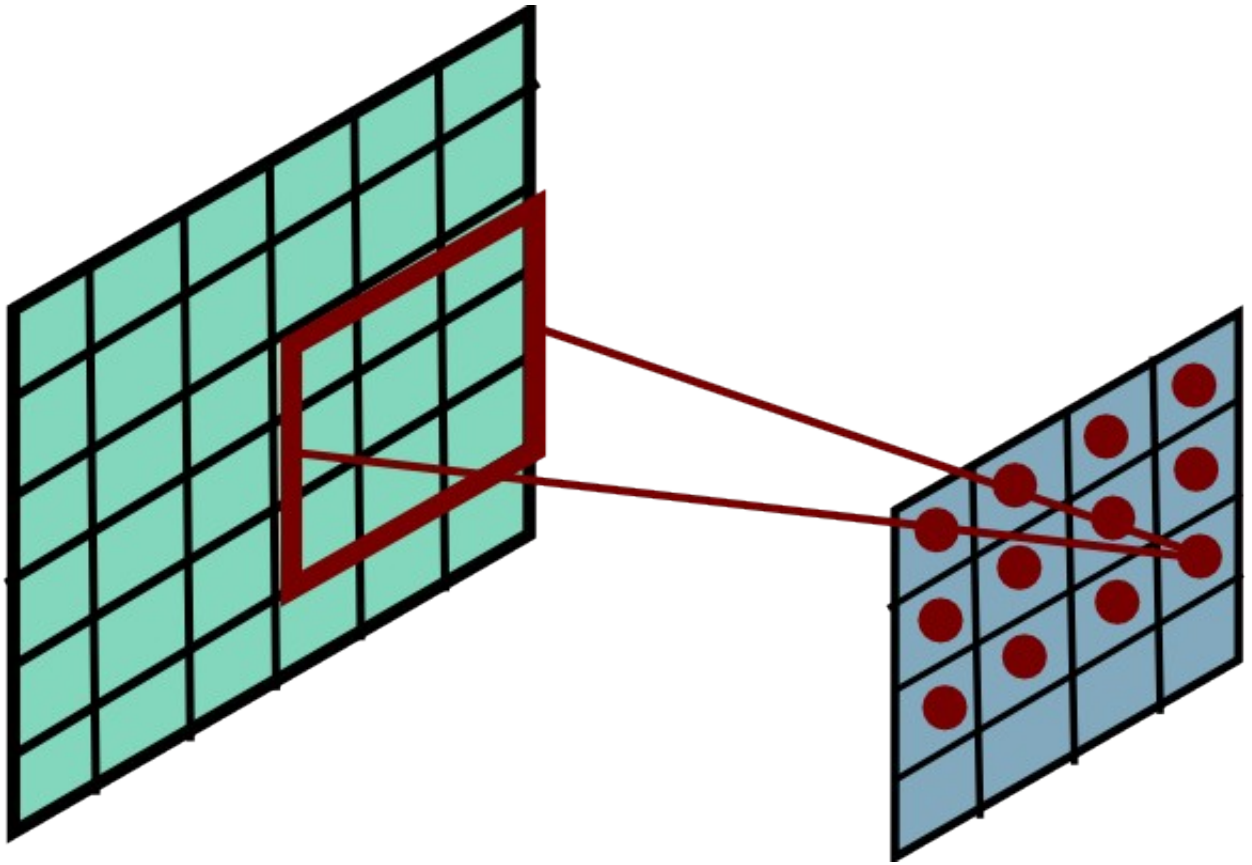
Convolutional Layer



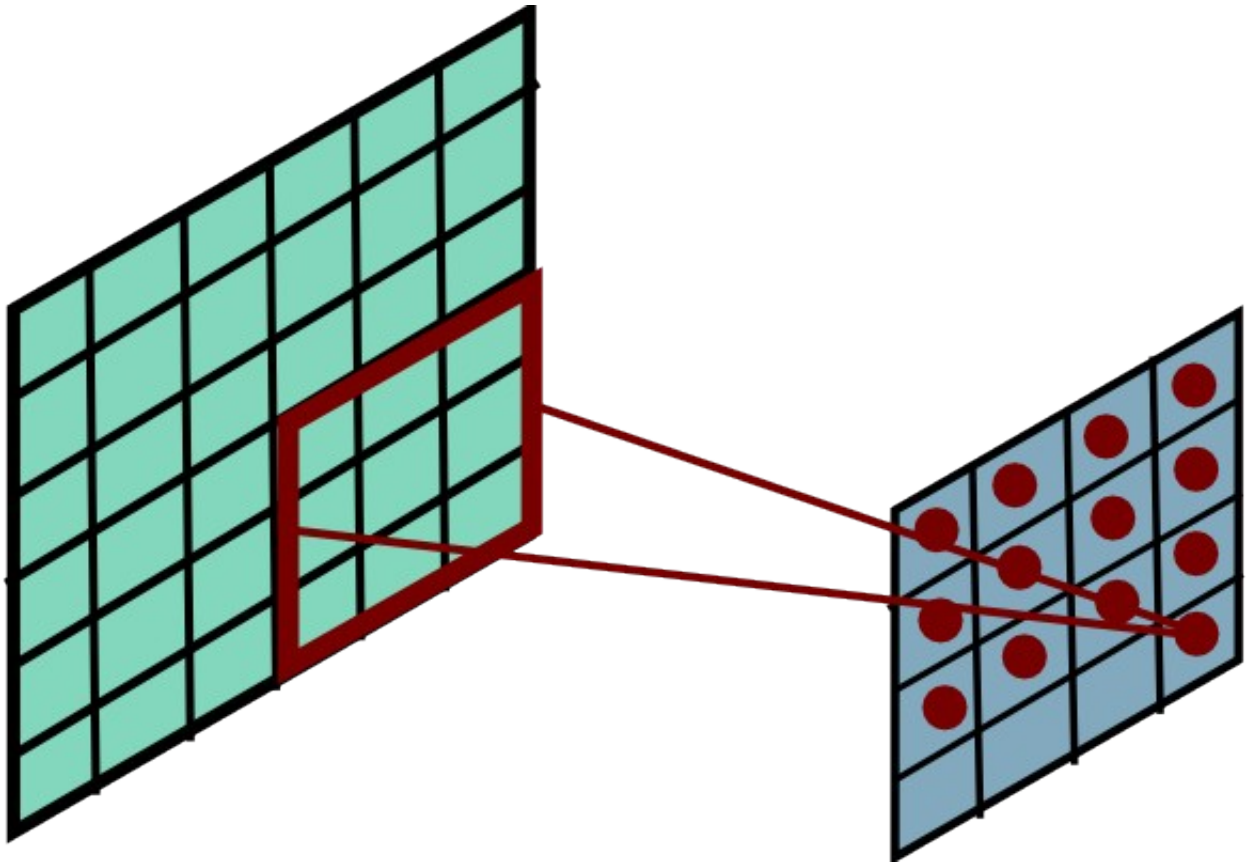
Convolutional Layer



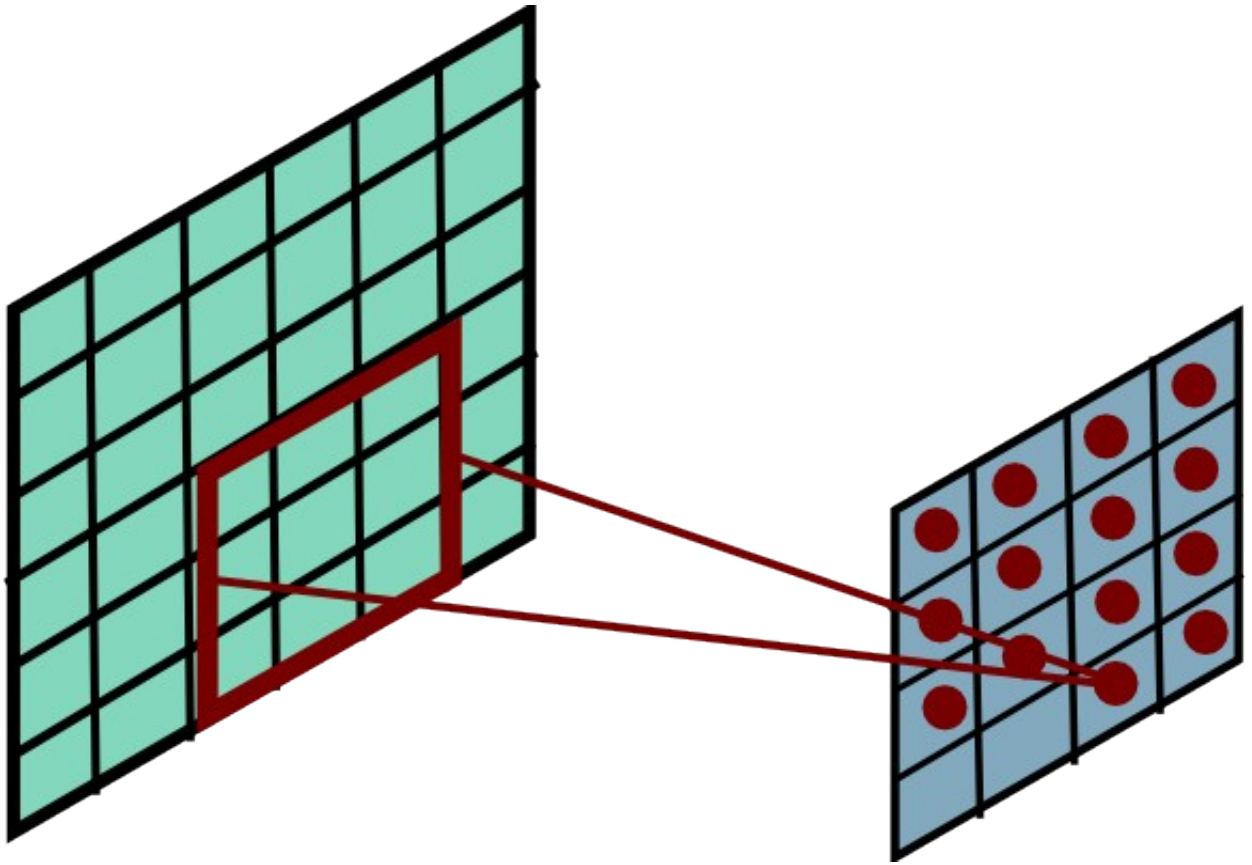
Convolutional Layer



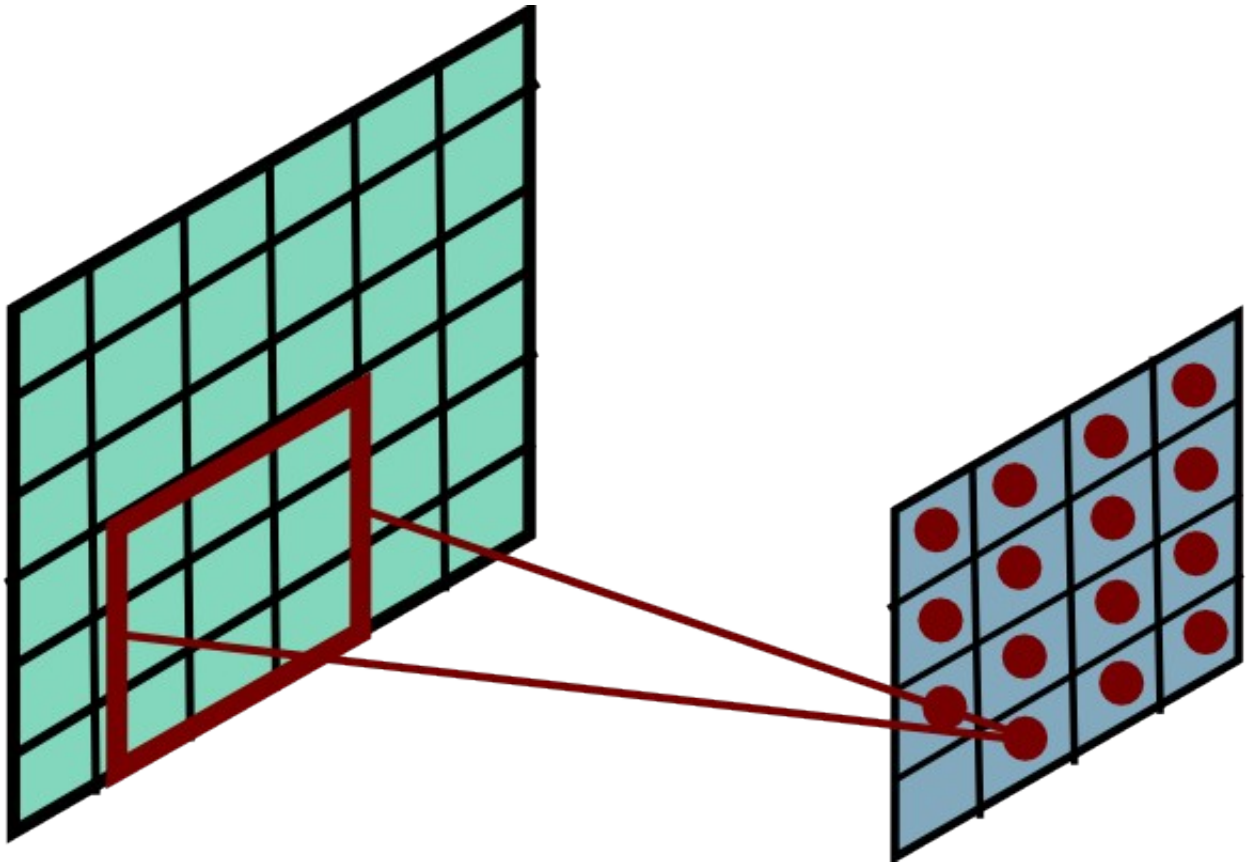
Convolutional Layer



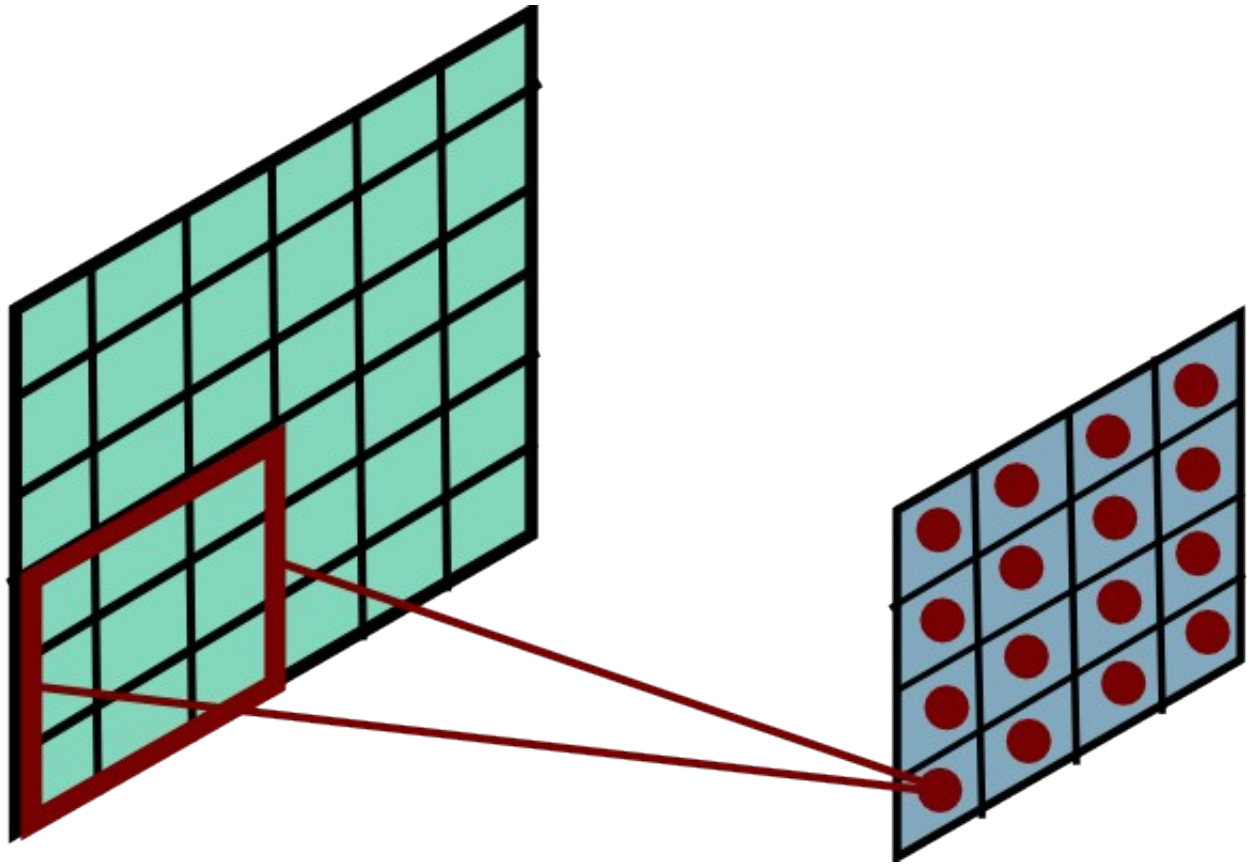
Convolutional Layer



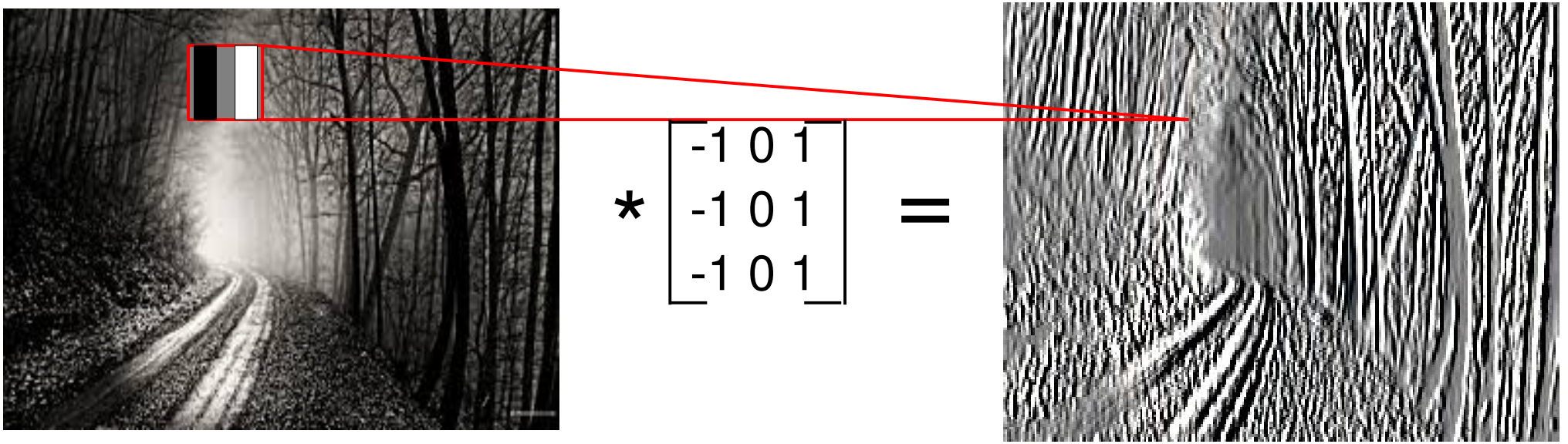
Convolutional Layer



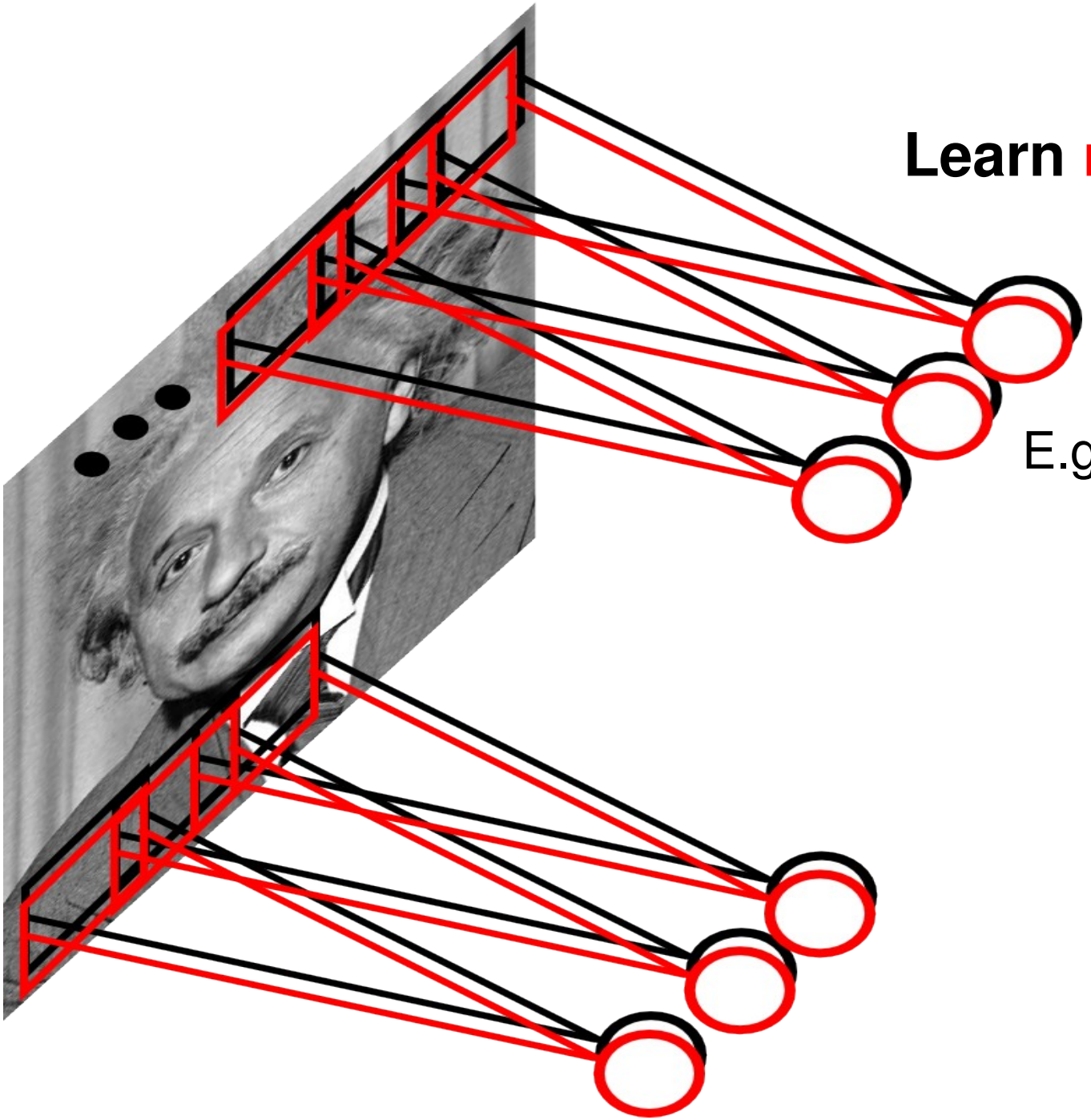
Convolutional Layer



Convolutional Layer



Convolutional Layer



Learn multiple filters.

E.g.: 200x200 image
100 Filters
Filter size: 10x10
10K parameters

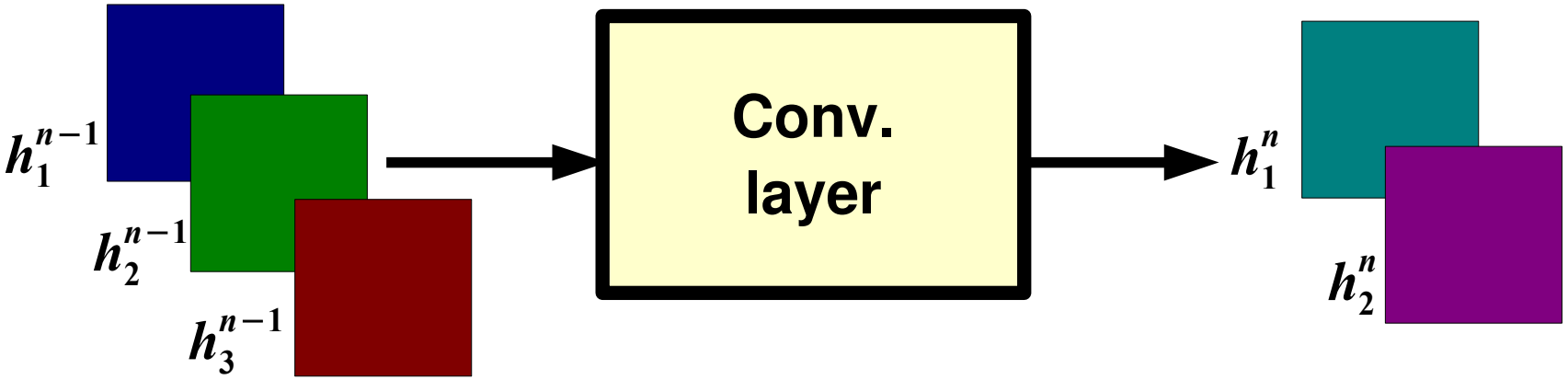
Convolutional Layer

$$h_j^n = max\left(0, \sum_{k=1}^{K} h_k^{n-1} * w_{kj}^n\right)$$

output feature map

input feature map

kernel



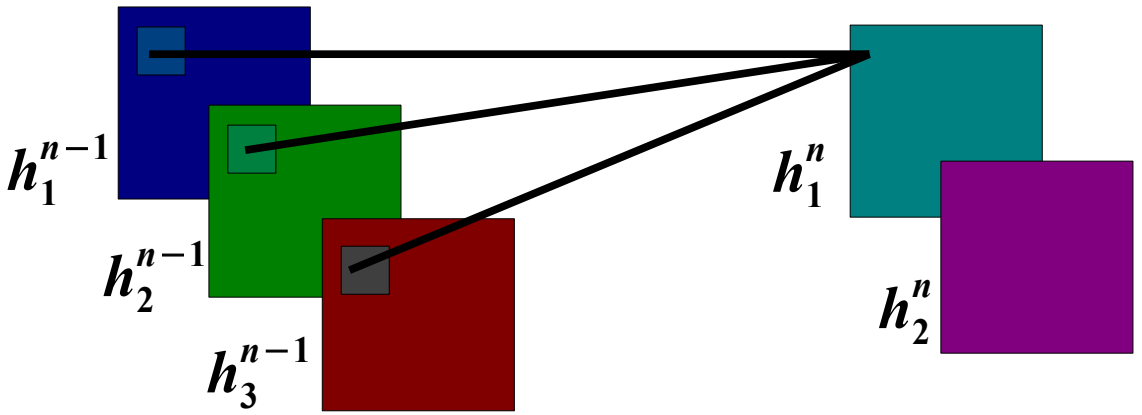
Convolutional Layer

$$h_j^n = max\left(0, \sum_{k=1}^{K} h_k^{n-1} * w_{kj}^n\right)$$

output feature map

input feature map

kernel



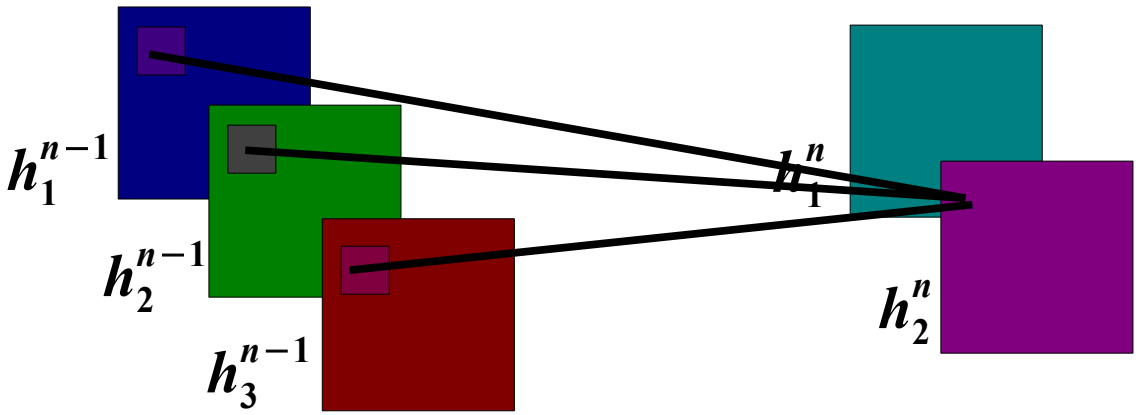
Convolutional Layer

$$h_j^n = max\left(0, \sum_{k=1}^{K} h_k^{n-1} * w_{kj}^n\right)$$

output feature map

input feature map

kernel



Convolutional Layer

Question: What is the size of the output? What's the computational cost?

Answer: It is proportional to the number of filters and depends on the stride. If kernels have size KxK, input has size DxD, stride is 1, and there are M input feature maps and N output feature maps then:
- the input has size M@DxD
- the output has size N@(D-K+1)x(D-K+1)
- the kernels have MxNxKxK coefficients (which have to be learned)
- cost: M*K*K*N*(D-K+1)*(D-K+1)

Question: How many feature maps? What's the size of the filters?

Answer: Usually, there are more output feature maps than input feature maps. Convolutional layers can increase the number of hidden units by big factors (and are expensive to compute).
The size of the filters has to match the size/scale of the patterns we want to detect (task dependent).

Key Ideas

A standard neural net applied to images:

- scales quadratically with the size of the input
- does not leverage stationarity

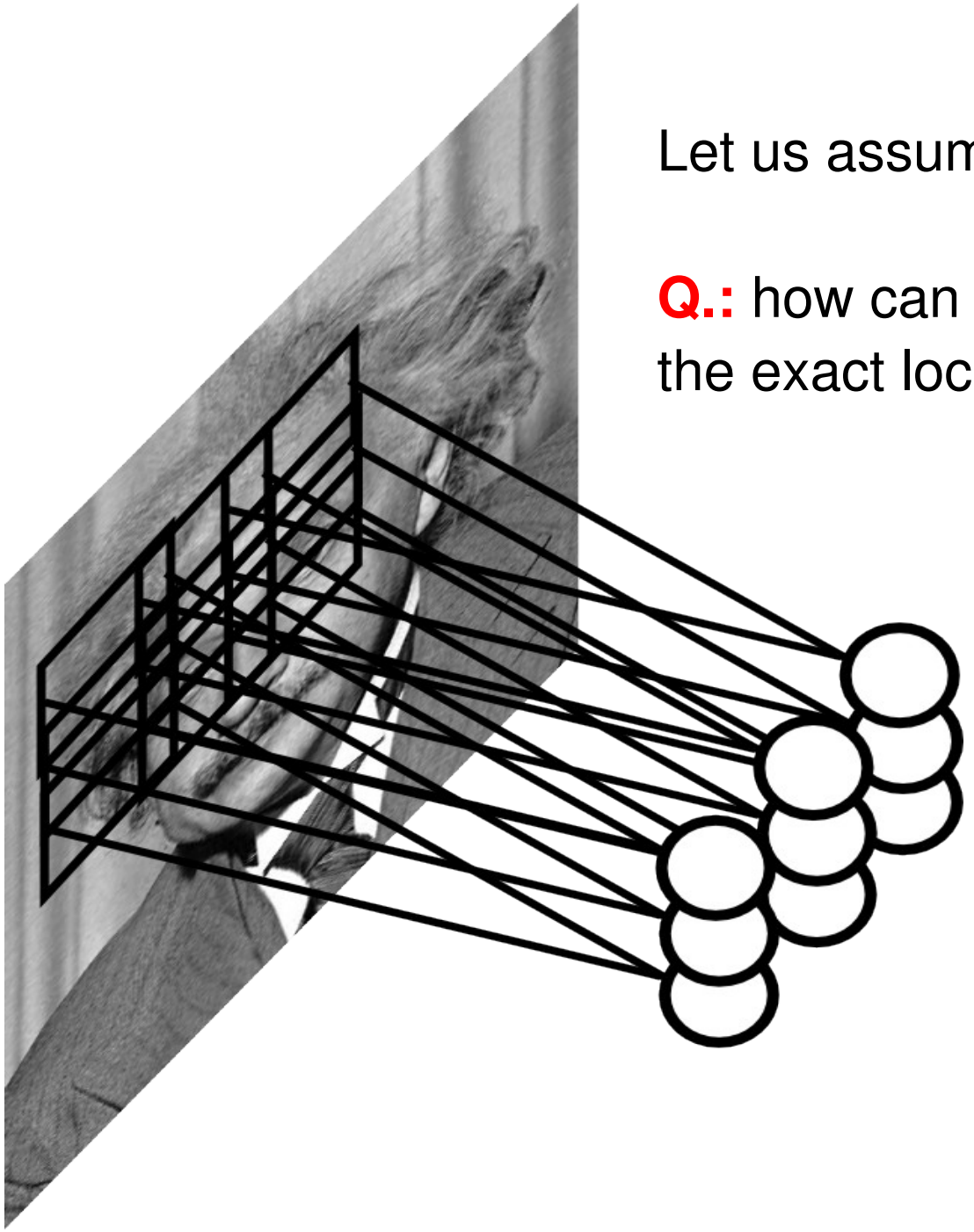
Solution:

- connect each hidden unit to a small patch of the input
- share the weight across space

This is called: **convolutional layer.**
A network with convolutional layers is called **convolutional network.**

LeCun et al. “Gradient-based learning applied to document recognition” IEEE 1998

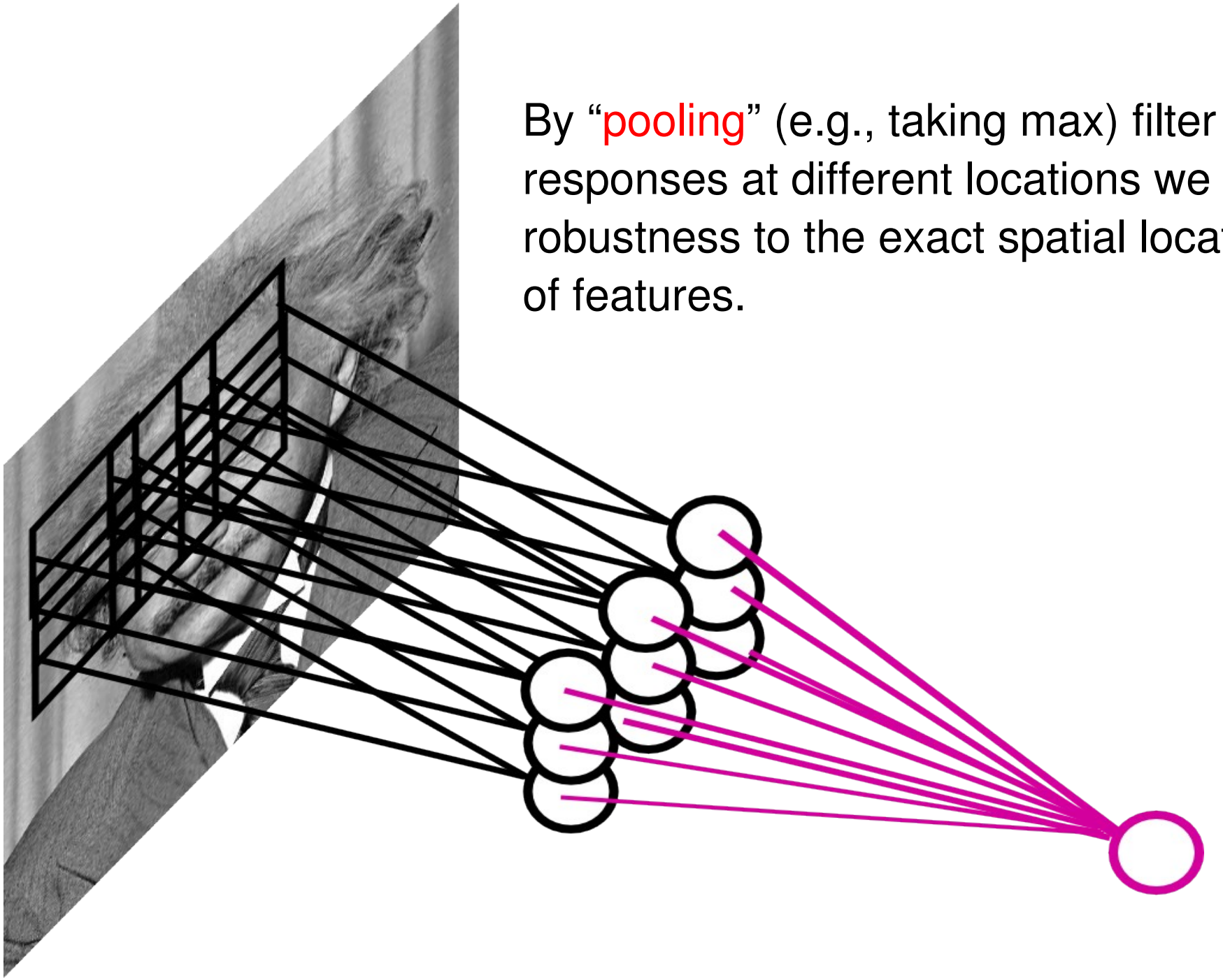
Pooling Layer



Let us assume filter is an “eye” detector.

Q.: how can we make the detection robust to the exact location of the eye?

Pooling Layer



By “pooling” (e.g., taking max) filter responses at different locations we gain robustness to the exact spatial location of features.

Pooling Layer: Examples

Max-pooling:

$$h_j^n(x,y) = max_{\bar{x}\in N(x),\, \bar{y}\in N(y)} h_j^{n-1}(\bar{x},\bar{y})$$

Average-pooling:

$$h_j^n(x,y) = 1/K \sum_{\bar{x}\in N(x),\, \bar{y}\in N(y)} h_j^{n-1}(\bar{x},\bar{y})$$

L2-pooling:

$$h_j^n(x,y) = \sqrt{\sum_{\bar{x}\in N(x),\, \bar{y}\in N(y)} h_j^{n-1}(\bar{x},\bar{y})^2}$$

L2-pooling over features:

$$h_j^n(x,y) = \sqrt{\sum_{k\in N(j)} h_k^{n-1}(x,y)^2}$$

Pooling Layer

Question: What is the size of the output? What's the computational cost?

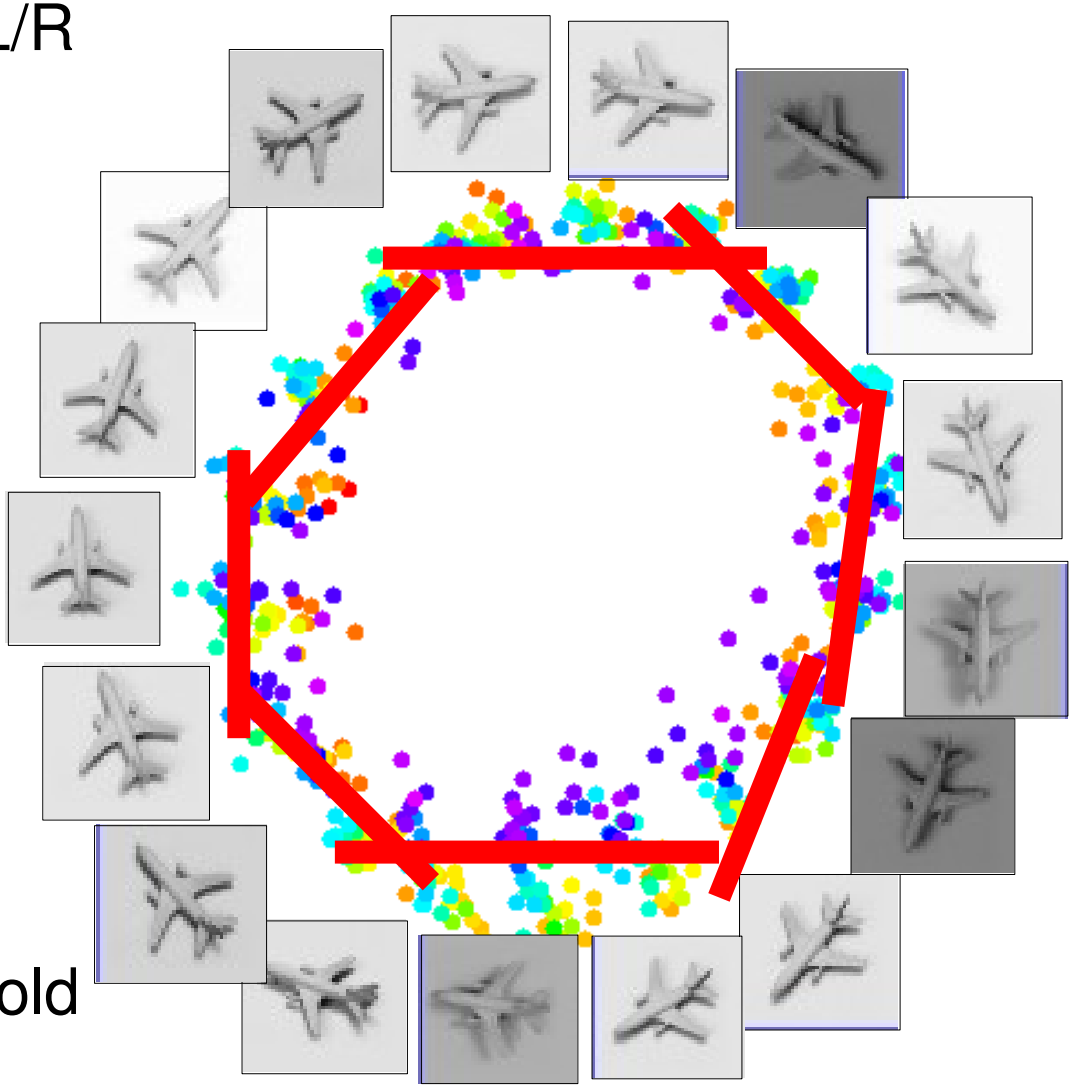
Answer: The size of the output depends on the stride between the pools. For instance, if pools do not overlap and have size KxK, and the input has size DxD with M input feature maps, then:
- output is M@(D/K)x(D/K)
- the computational cost is proportional to the size of the input (negligible compared to a convolutional layer)

Question: How should I set the size of the pools?

Answer: It depends on how much “invariant” or robust to distortions we want the representation to be. It is best to pool slowly (via a few stacks of conv-pooling layers).

Pooling Layer: Interpretation

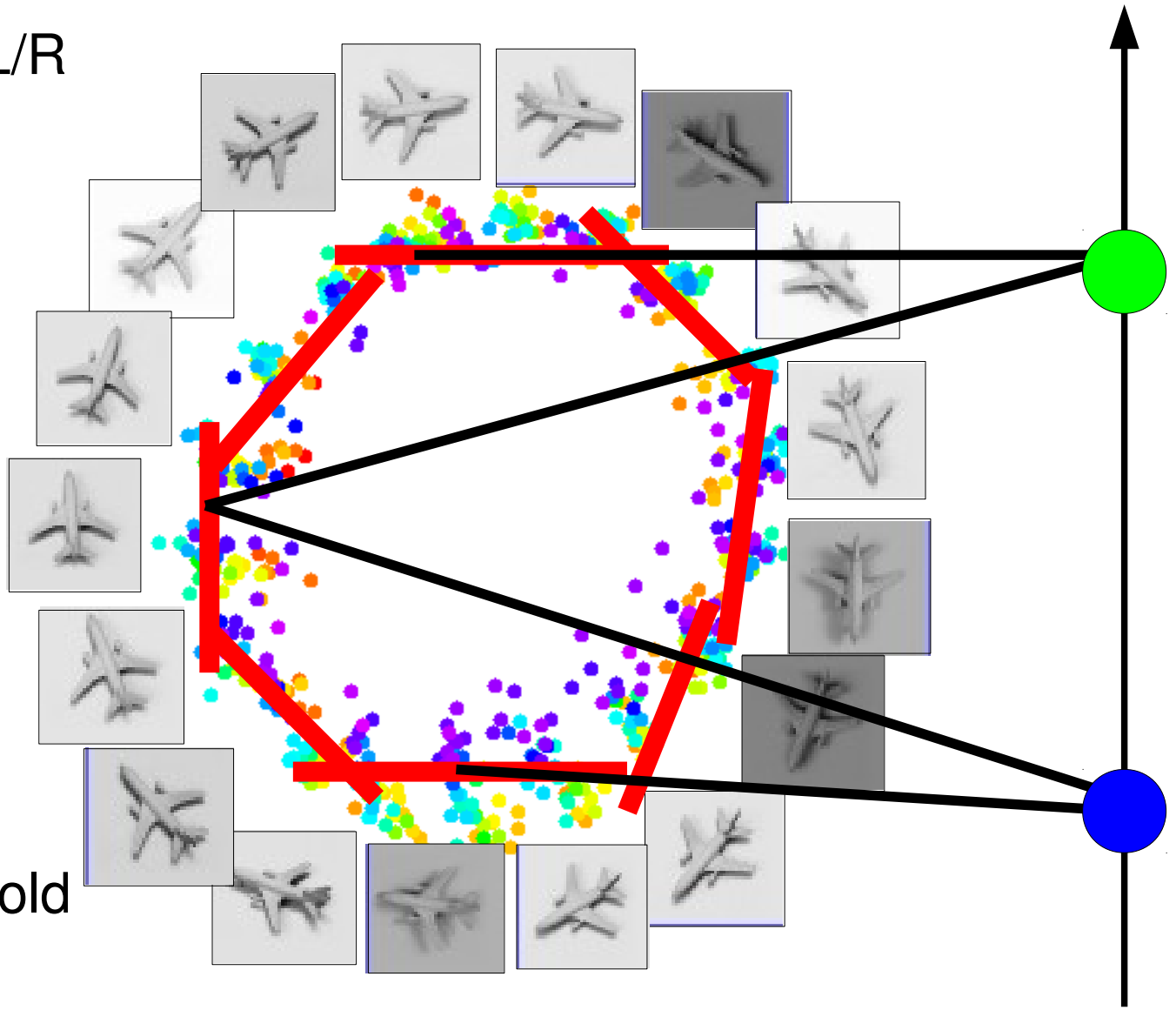
Task: detect orientation L/R



Conv layer:
linearizes manifold

Pooling Layer: Interpretation

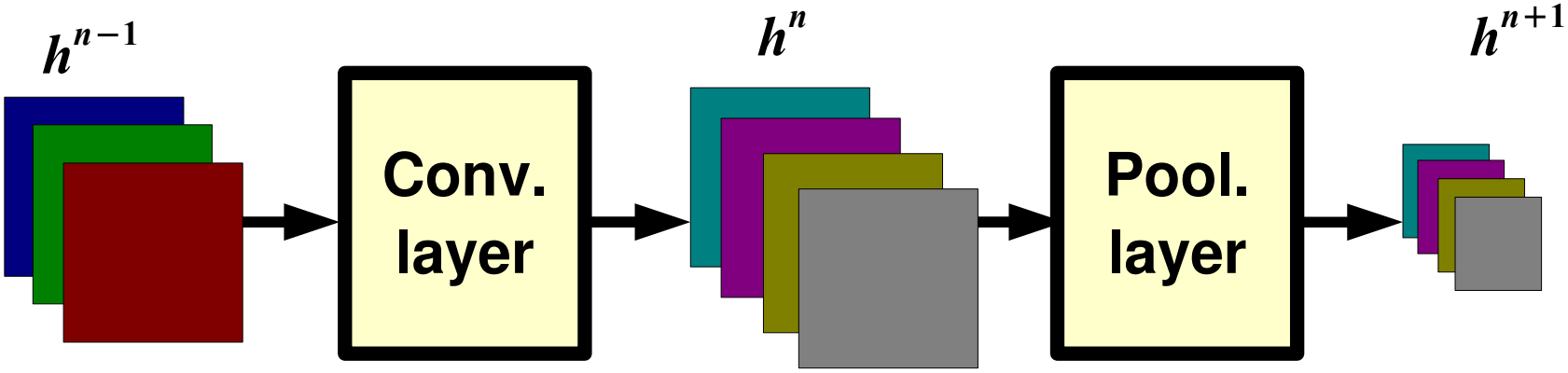
Task: detect orientation L/R



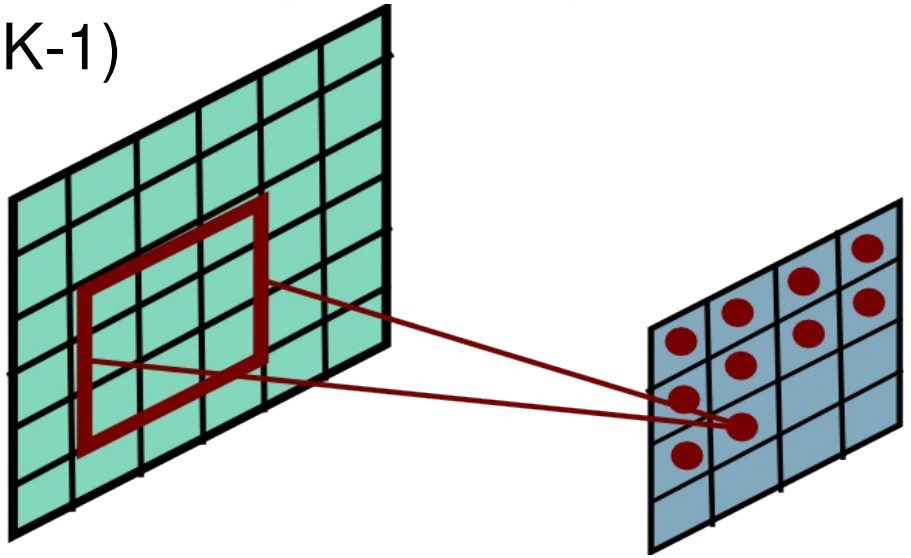
Conv layer:
linearizes manifold

Pooling layer:
collapses manifold

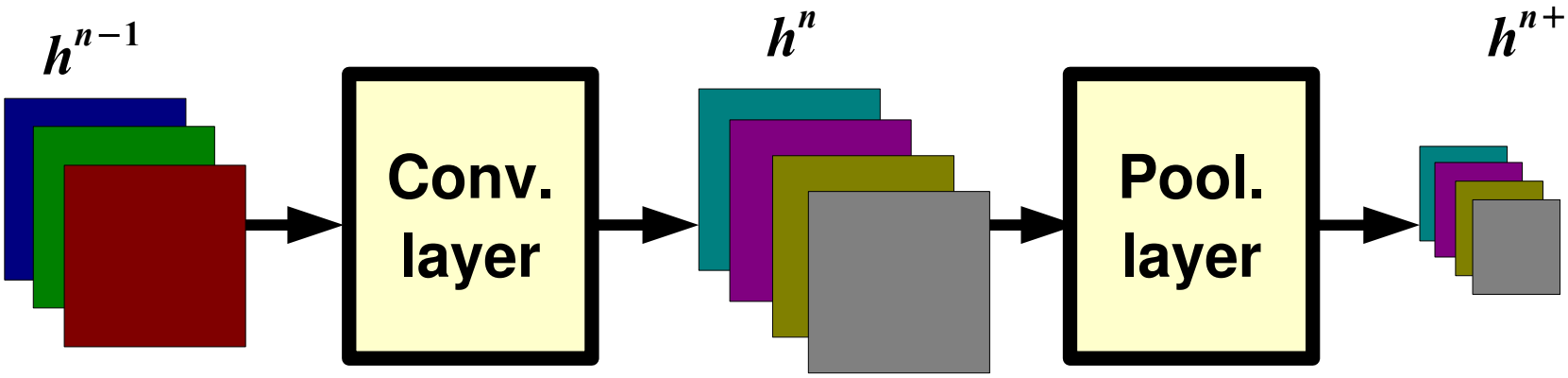
Pooling Layer: Receptive Field Size



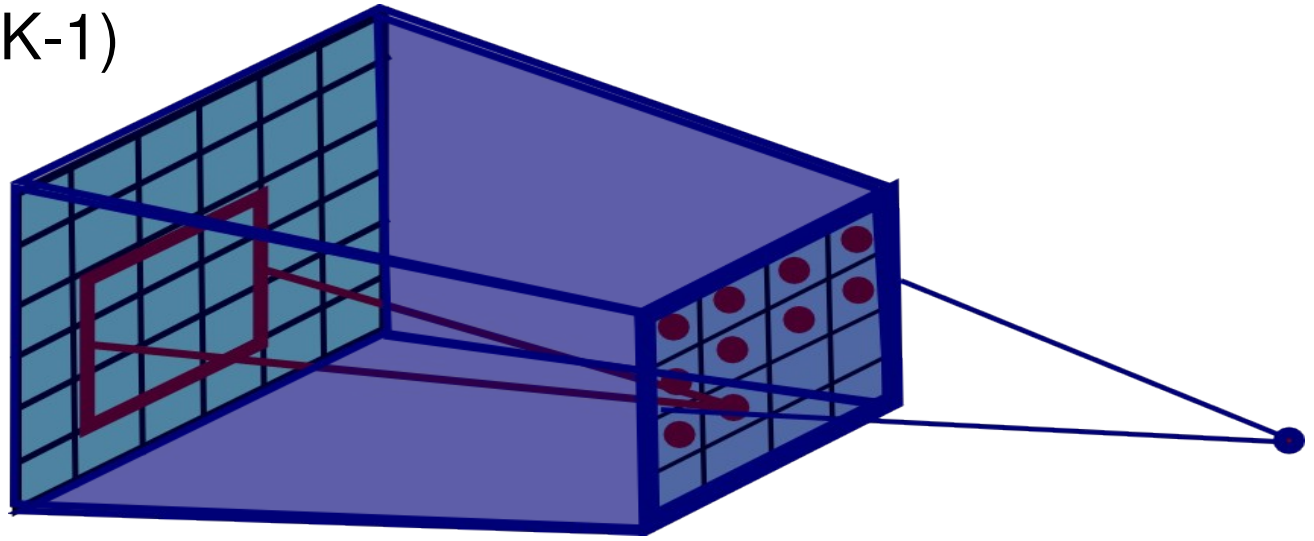
If convolutional filters have size KxK and stride 1, and pooling layer has pools of size PxP, then each unit in the pooling layer depends upon a patch (at the input of the preceding conv. layer) of size: (P+K-1)x(P+K-1)



Pooling Layer: Receptive Field Size

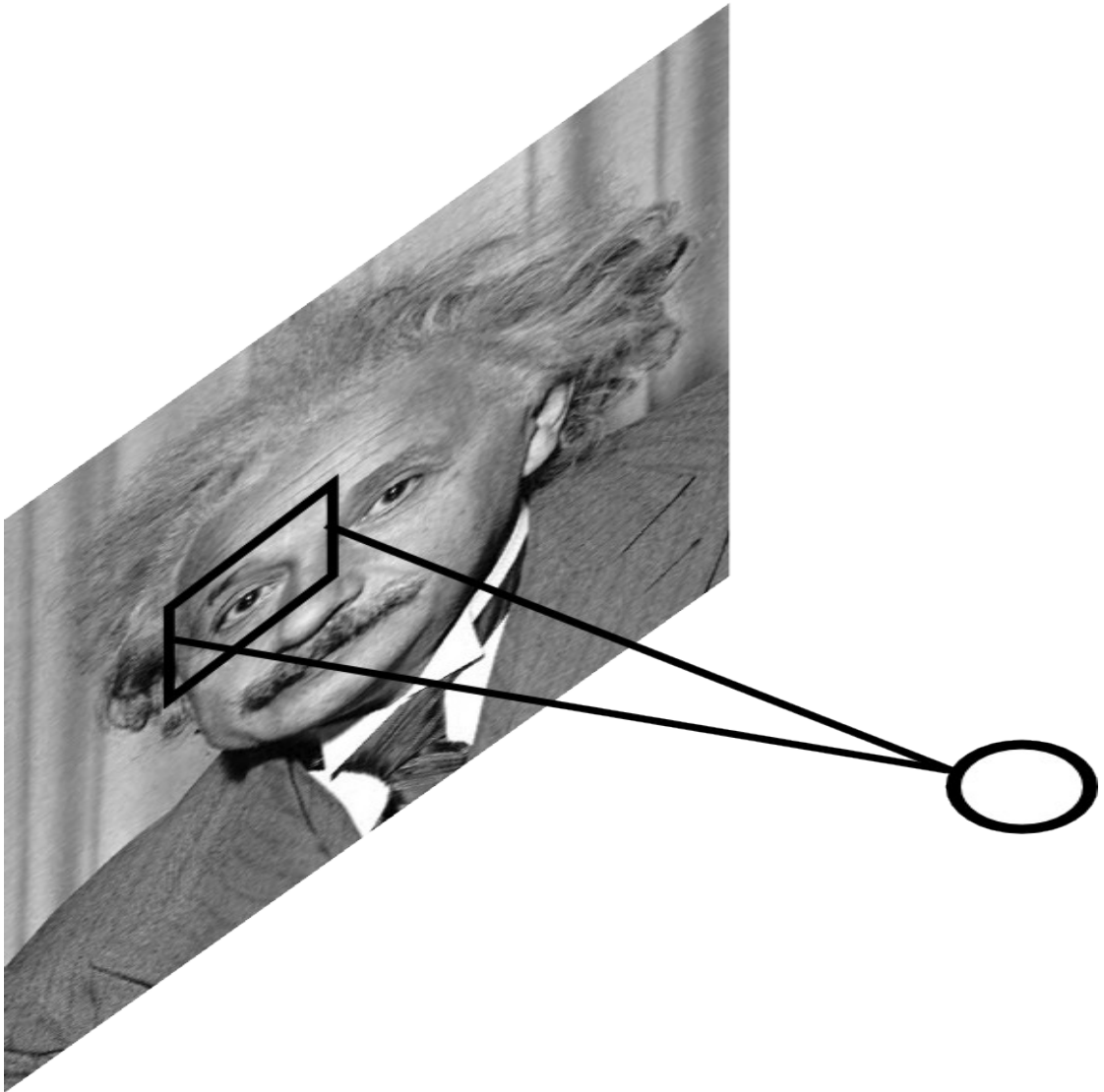


If convolutional filters have size KxK and stride 1, and pooling layer has pools of size PxP, then each unit in the pooling layer depends upon a patch (at the input of the preceding conv. layer) of size: (P+K-1)x(P+K-1)



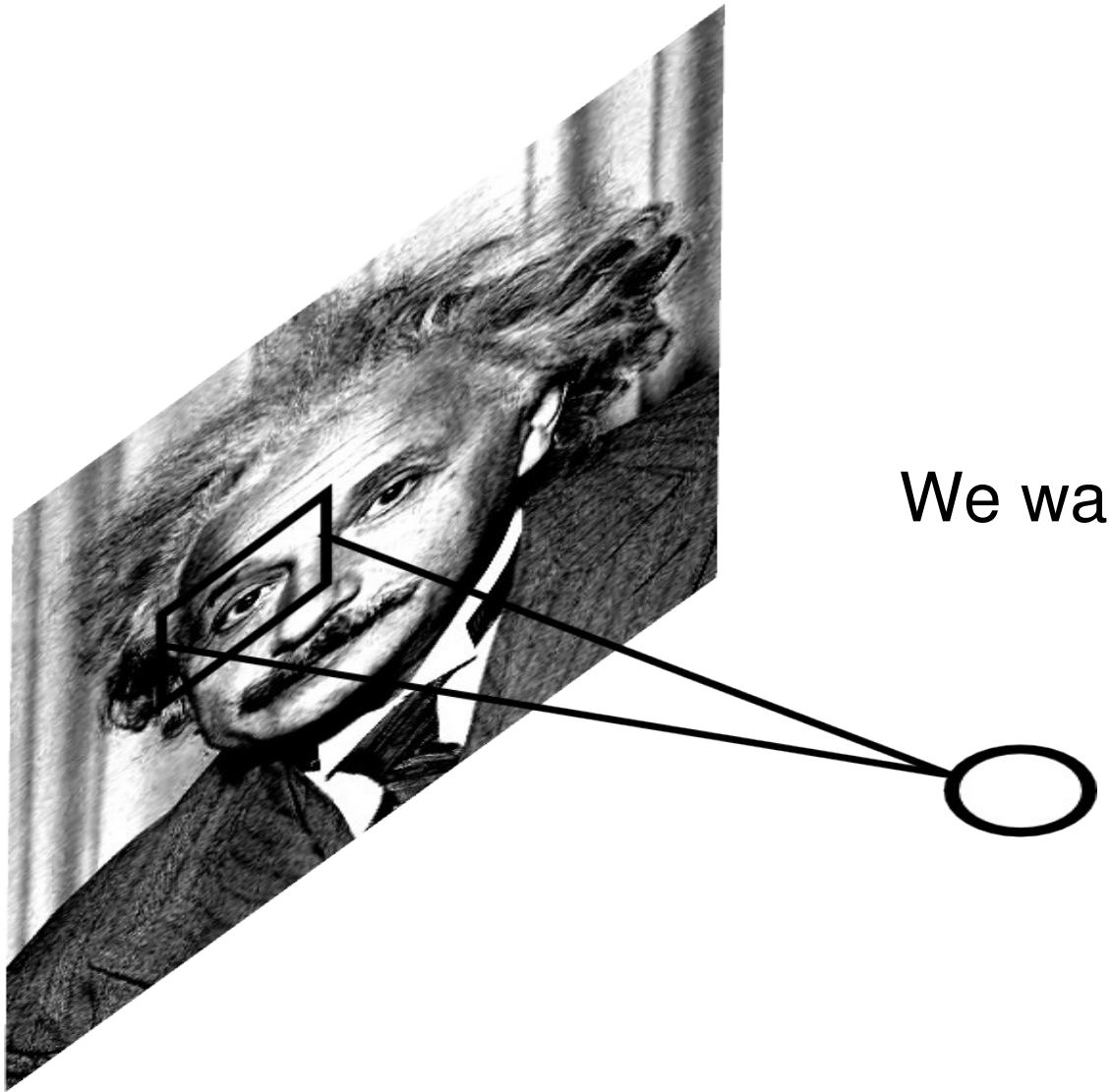
Local Contrast Normalization

$$h^{i+1}(x,y)=\frac{h^{i}(x,y)-m^{i}(N(x,y))}{\sigma^{i}(N(x,y))}$$



Local Contrast Normalization

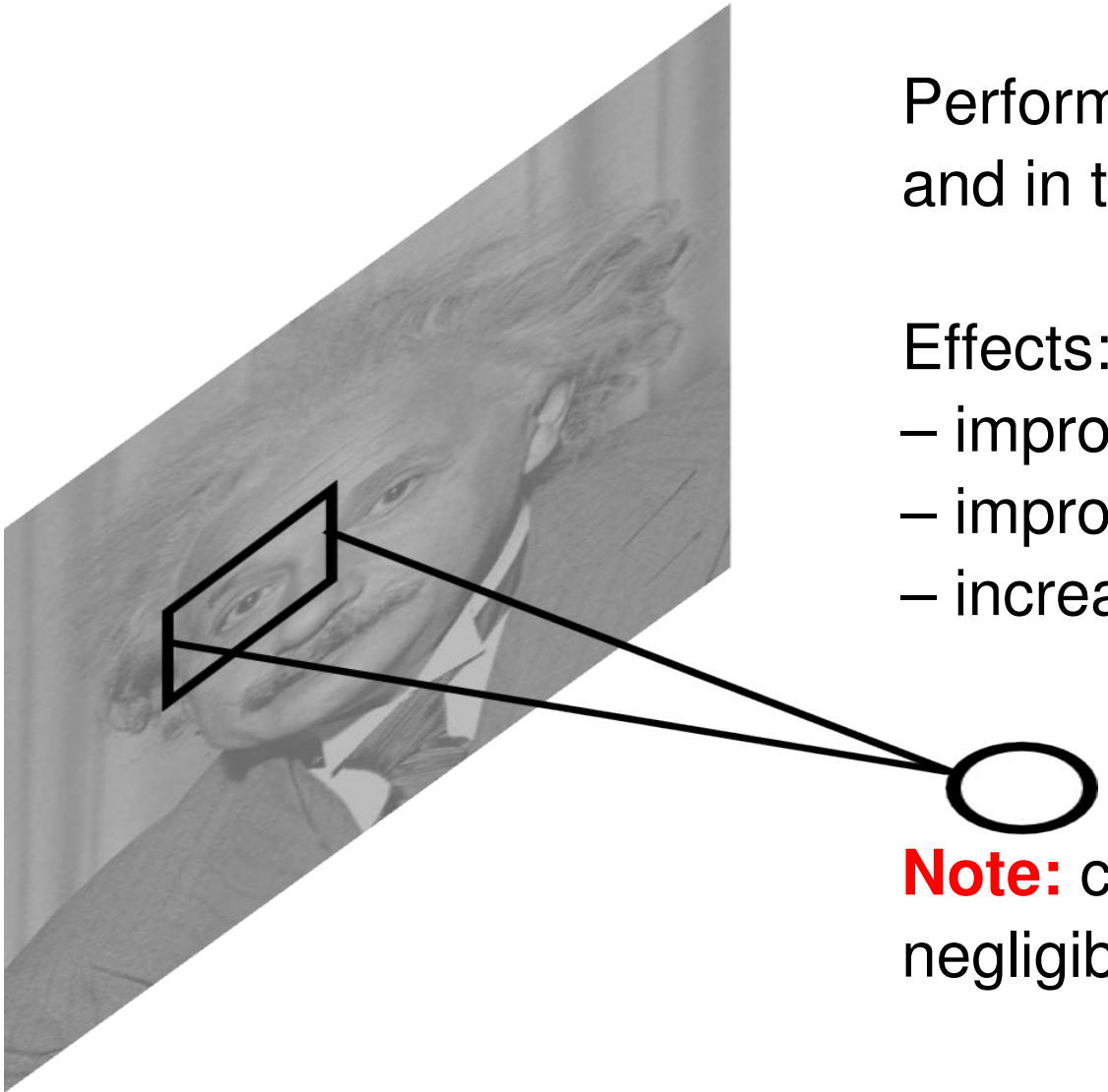
$$h^{i+1}(x,y)=\frac{h^{i}(x,y)-m^{i}(N(x,y))}{\sigma^{i}(N(x,y))}$$



We want the same response.

Local Contrast Normalization

$$h^{i+1}(x,y)=\frac{h^{i}(x,y)-m^{i}(N(x,y))}{\sigma^{i}(N(x,y))}$$



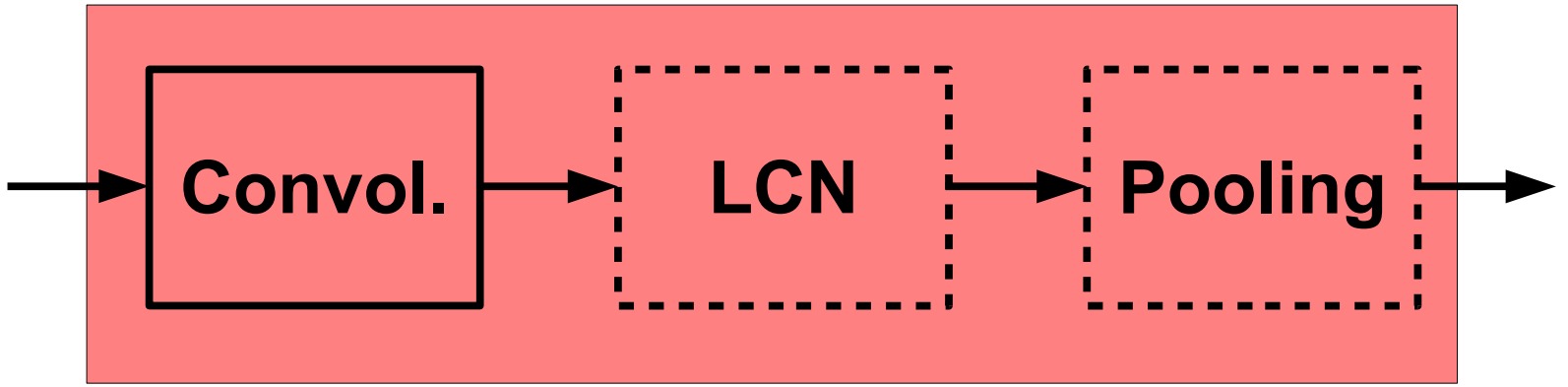
Performed also across features and in the higher layers..

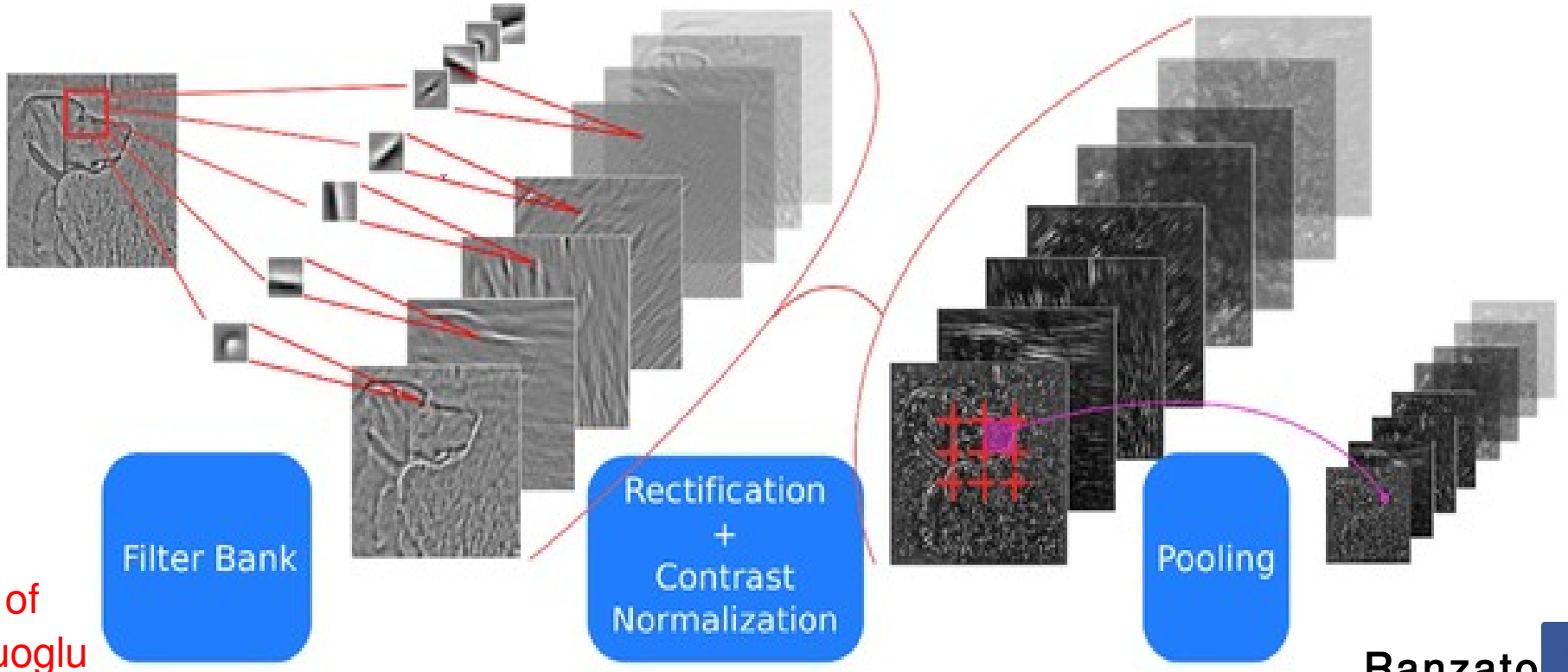
Effects:
- improves invariance
- improves optimization
- increases sparsity

Note: computational cost is negligible w.r.t. conv. layer.

ConvNets: Typical Stage

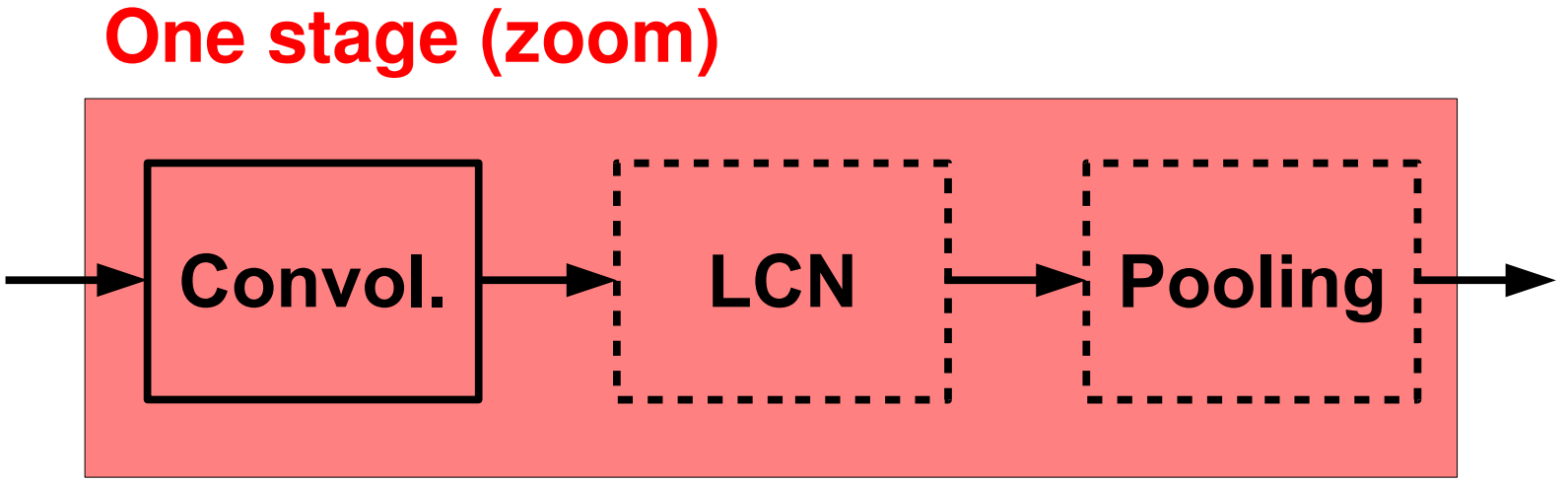
One stage (zoom)





courtesy of
K. Kavukcuoglu

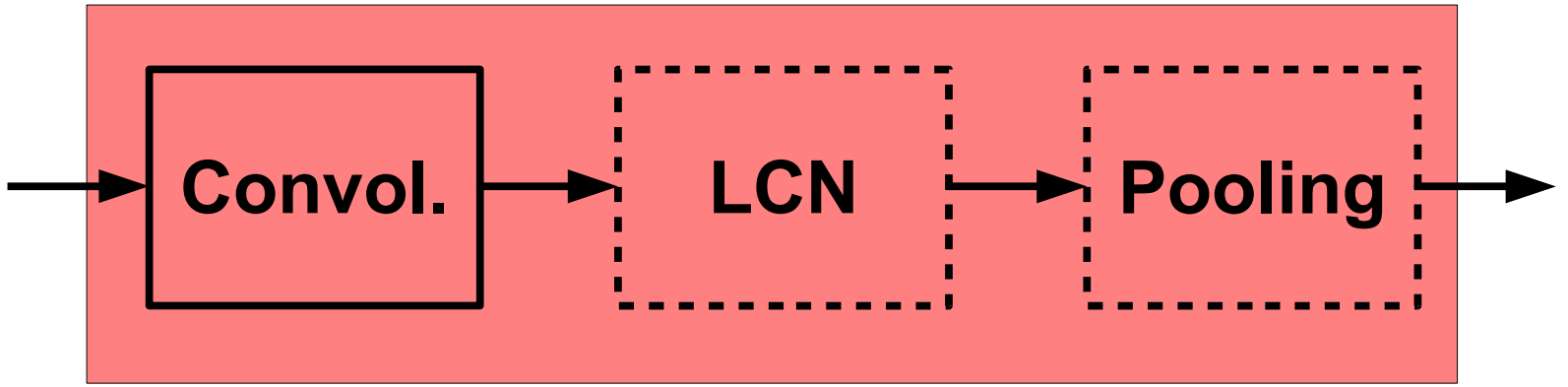
ConvNets: Typical Stage



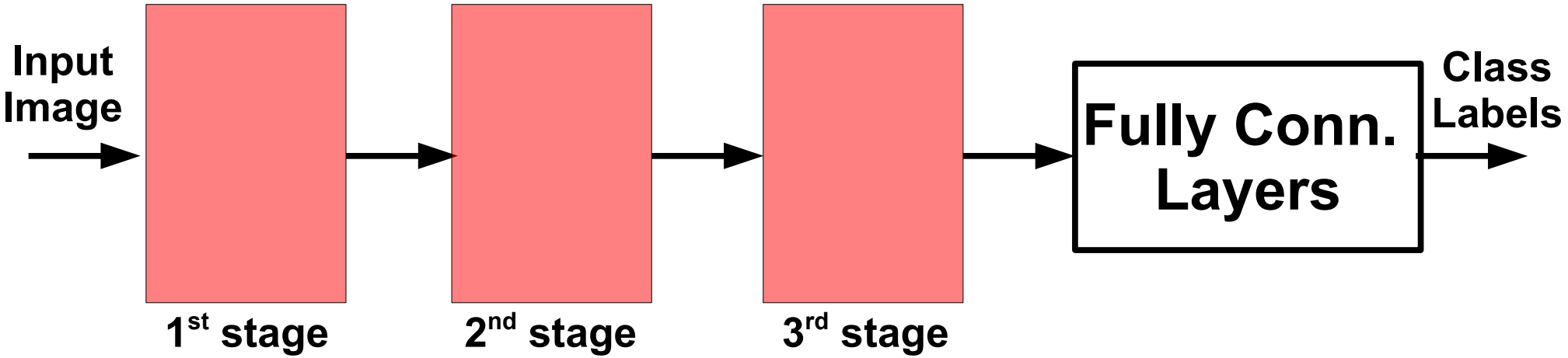
Conceptually similar to: SIFT, HoG, etc.

ConvNets: Typical Architecture

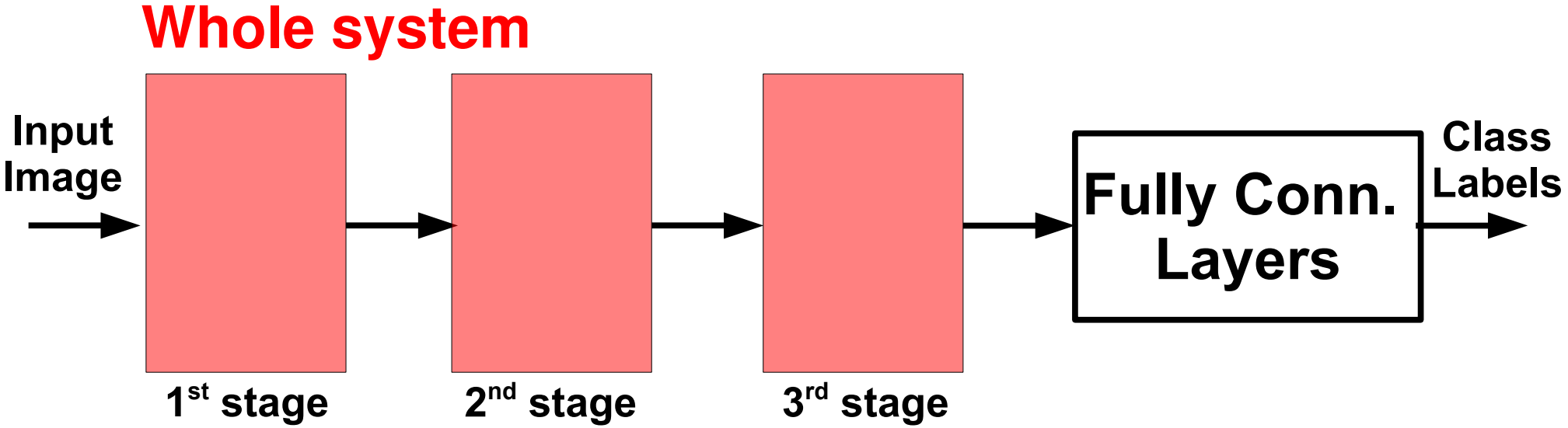
One stage (zoom)



Whole system



ConvNets: Typical Architecture



Conceptually similar to:

SIFT → K-Means → Pyramid Pooling → SVM

Lazebnik et al. "...Spatial Pyramid Matching..." CVPR 2006

SIFT → Fisher Vect. → Pooling → SVM

Sanchez et al. "Image classifcation with F.V.: Theory and practice" IJCV 2012

ConvNets: Training

All layers are differentiable (a.e.).
We can use standard back-propagation.

Algorithm:

Given a small mini-batch

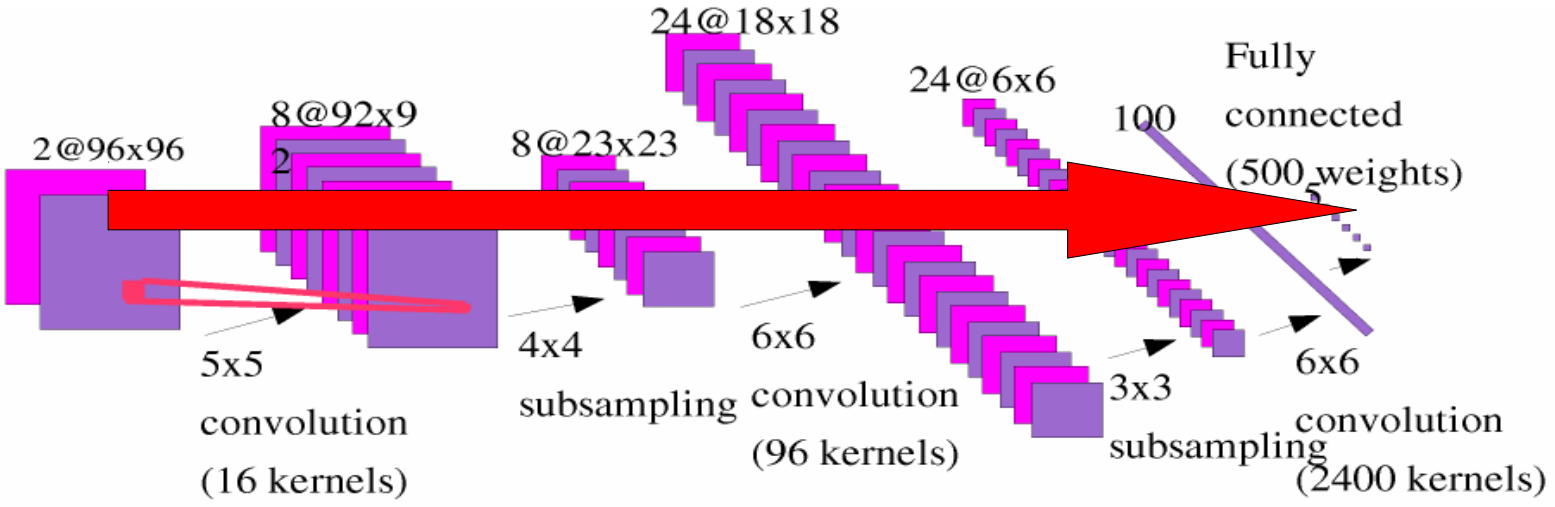
- F-PROP

- B-PROP

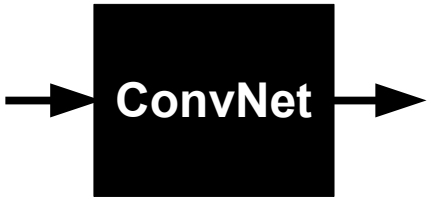
- PARAMETER UPDATE

ConvNets: Test

At test time, run only is forward mode (FPROP).



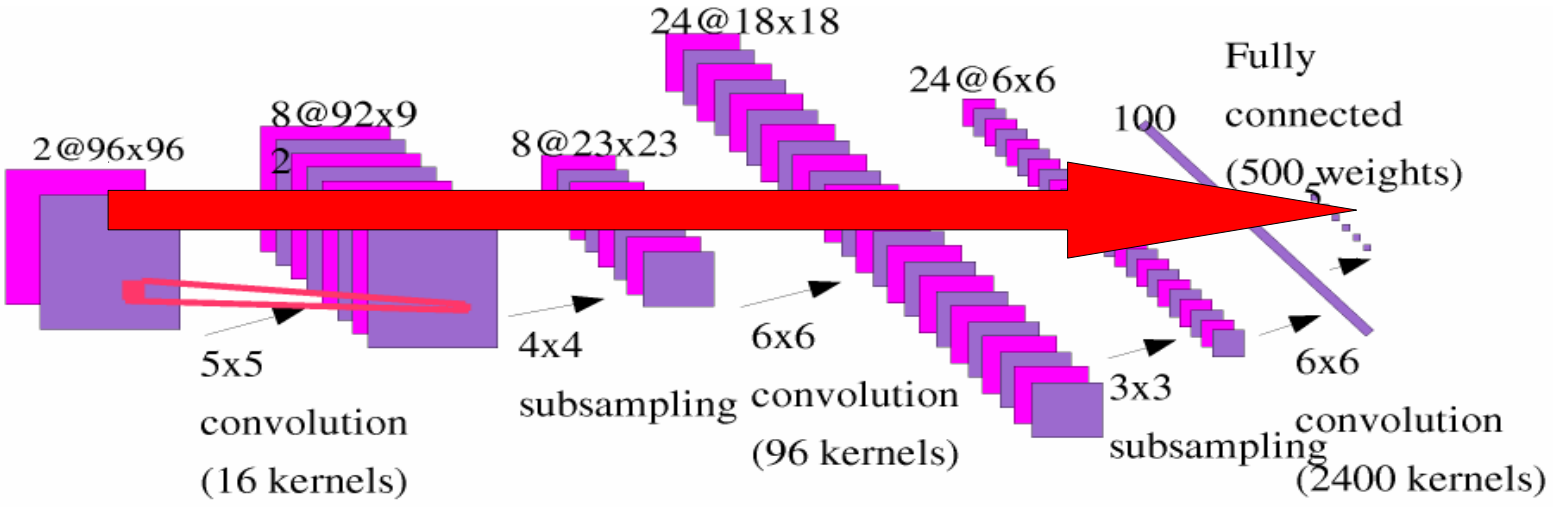
Naturally, convnet can process larger images at little cost.



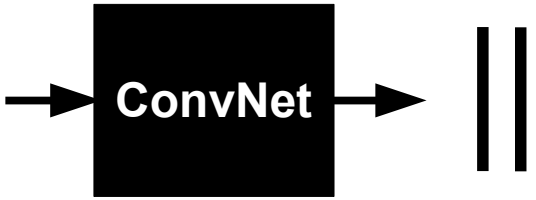
Traditional methods use inefficient sliding windows.

ConvNets: Test

At test time, run only is forward mode (FPROP).



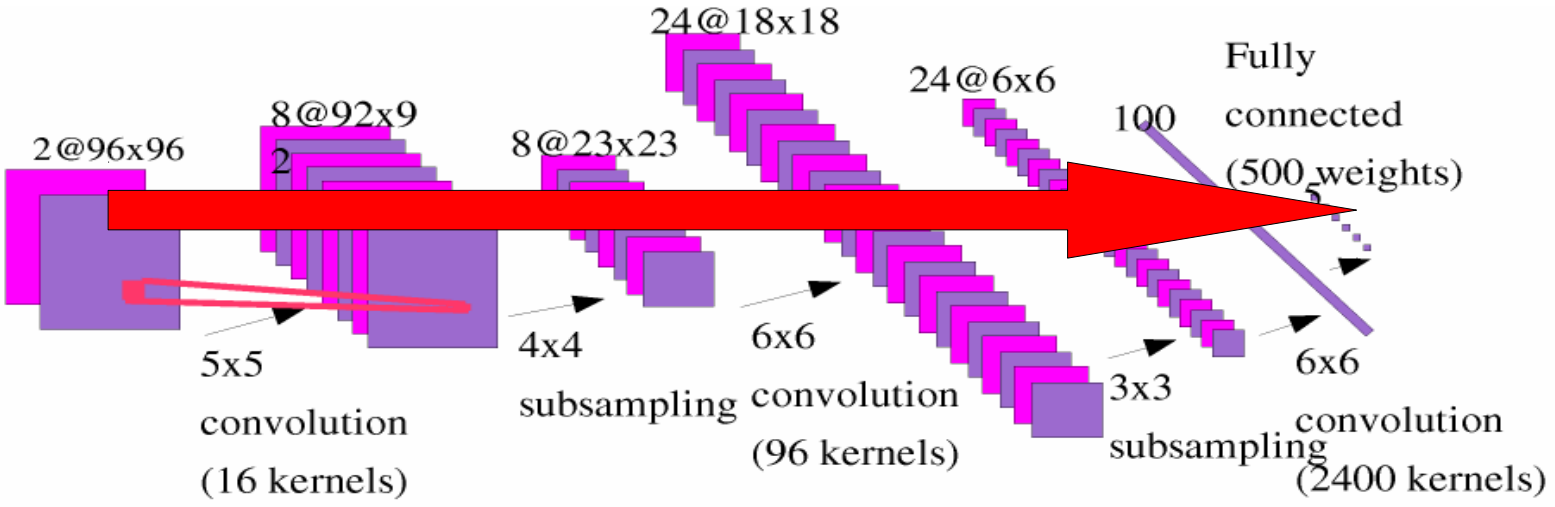
Naturally, convnet can process larger images at little cost.



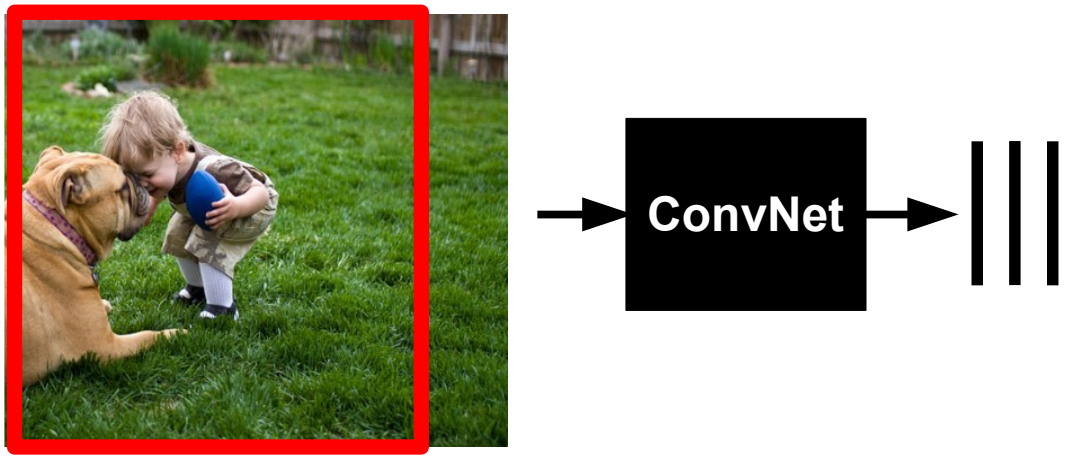
Traditional methods use inefficient sliding windows.

ConvNets: Test

At test time, run only is forward mode (FPROP).



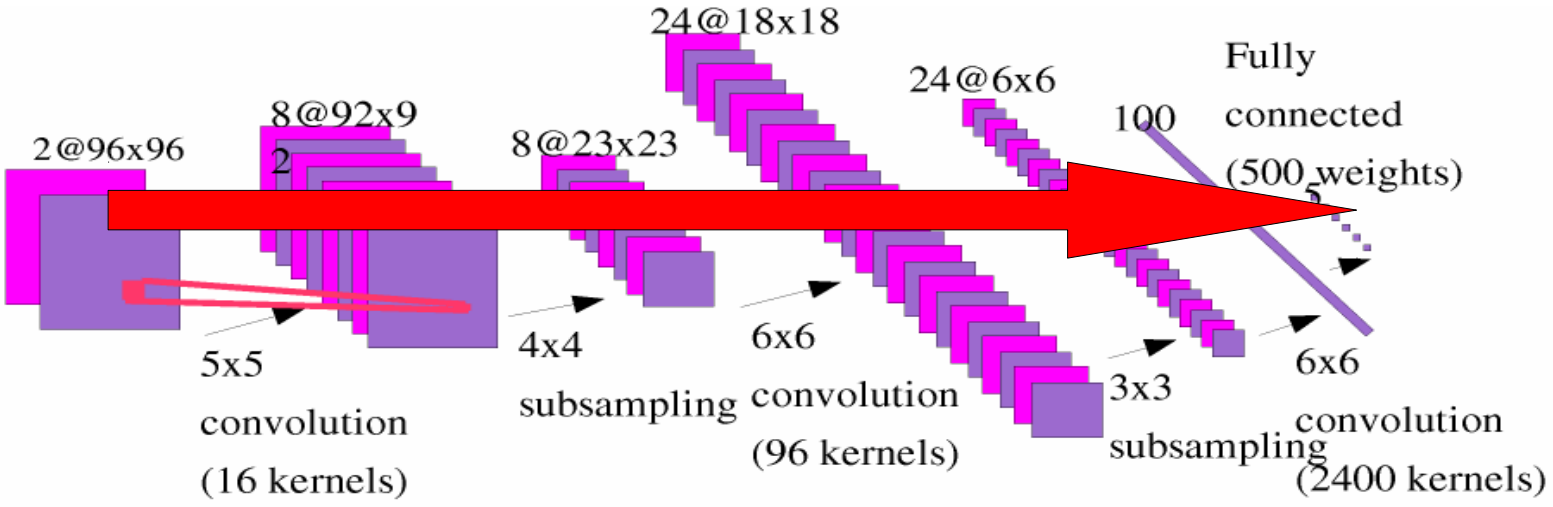
Naturally, convnet can process larger images at little cost.



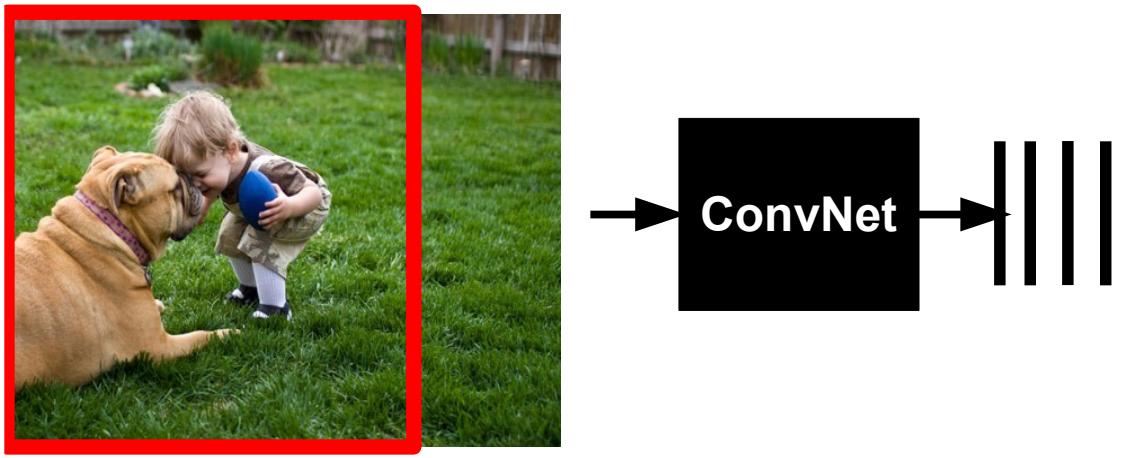
Traditional methods use inefficient sliding windows.

ConvNets: Test

At test time, run only is forward mode (FPROP).



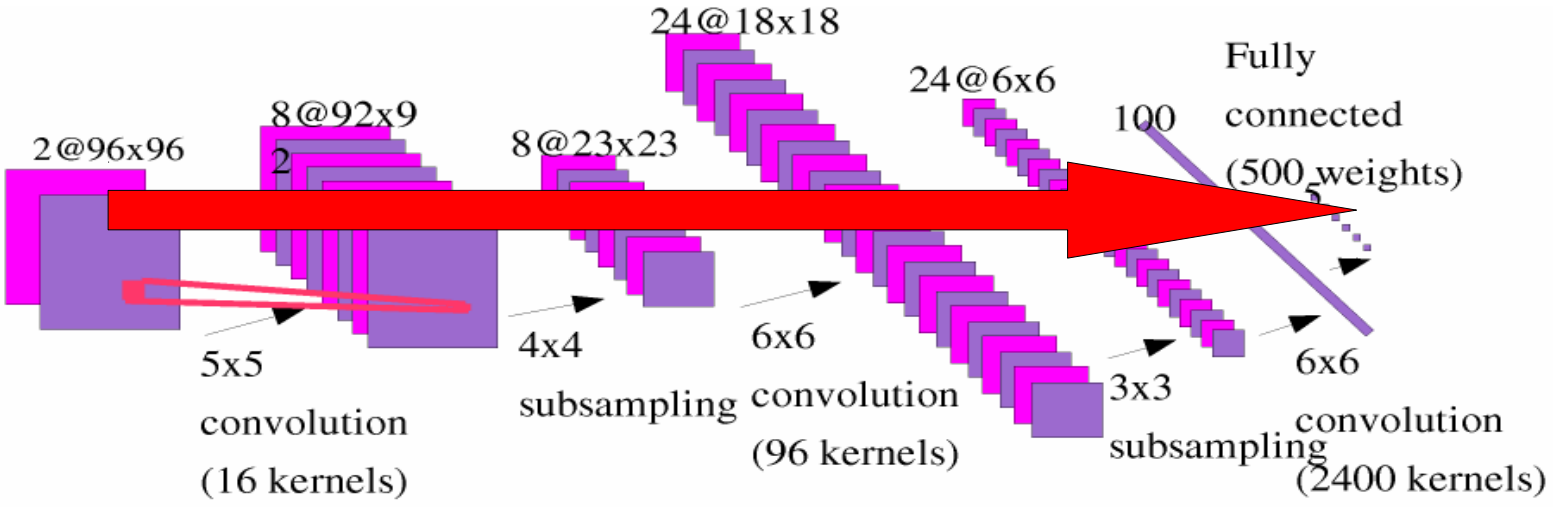
Naturally, convnet can process larger images at little cost.



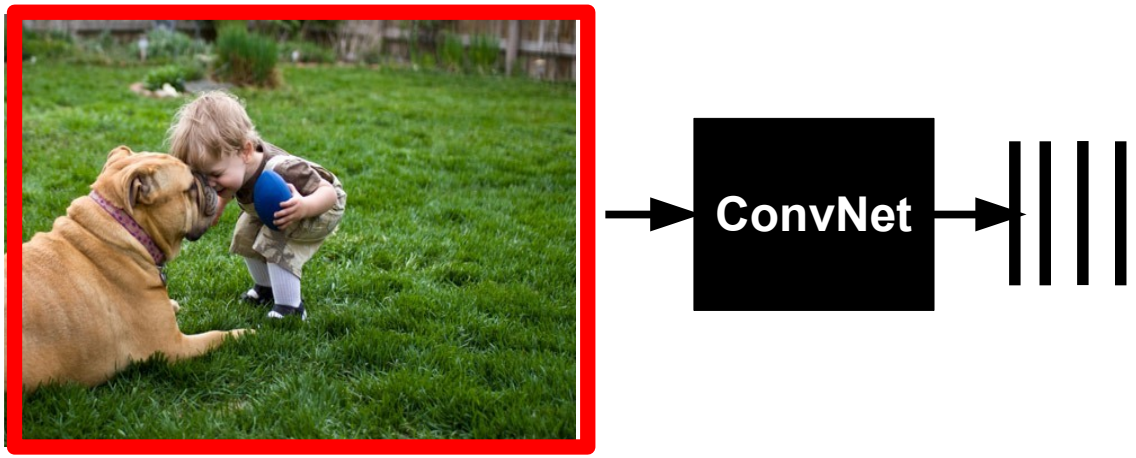
Traditional methods use inefficient sliding windows.

ConvNets: Test

At test time, run only is forward mode (FPROP).



Naturally, convnet can process larger images at little cost.



ConvNet: unrolls convolutions over bigger images and produces outputs at several locations.

Outline

- Supervised Neural Networks
- Convolutional Neural Networks
- Examples
- Tips

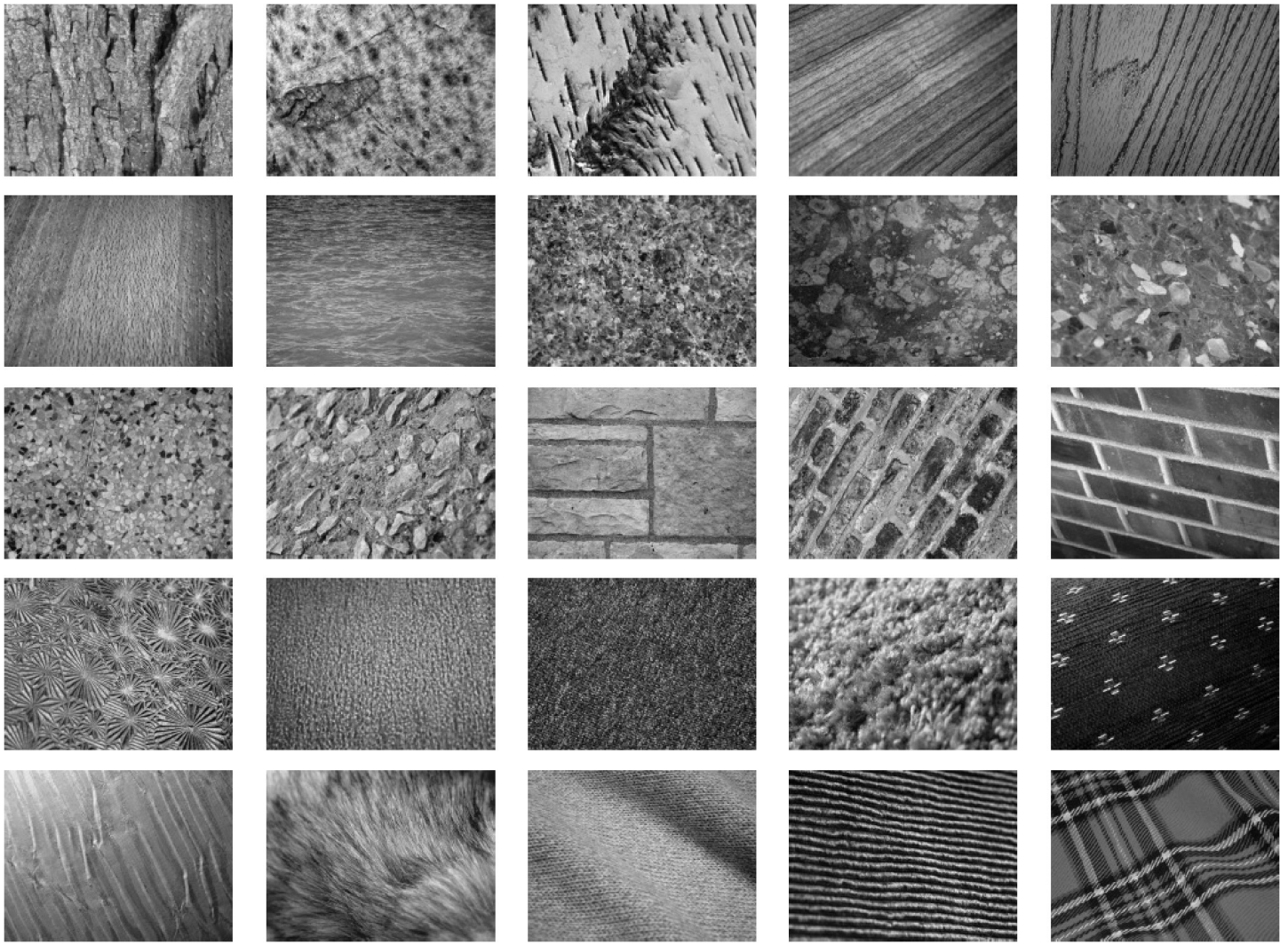
CONV NETS: EXAMPLES

- **OCR / House number & Traffic sign classification**

Ciresan et al. “MCDNN for image classification” CVPR 2012
Wan et al. “Regularization of neural networks using dropconnect” ICML 2013
Jaderberg et al. “Synthetic data and ANN for natural scene text recognition” arXiv 2014

CONV NETS: EXAMPLES

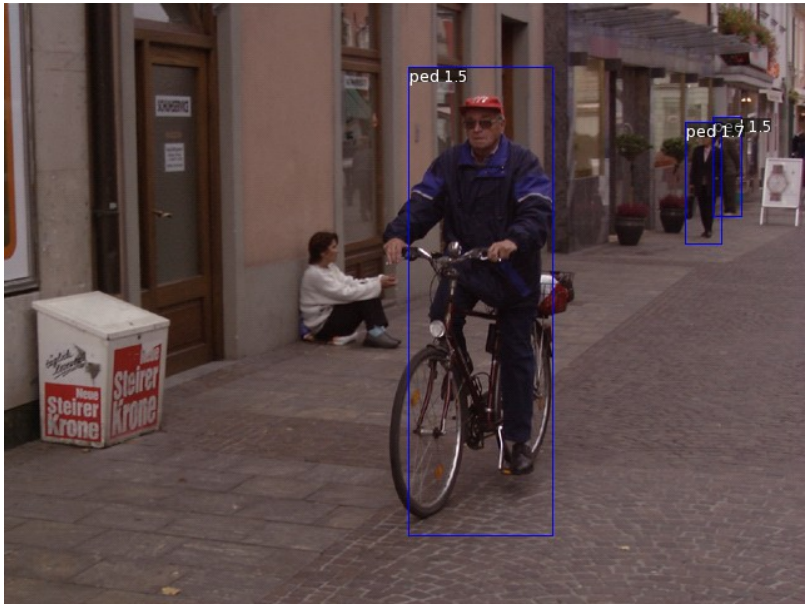
- **Texture classification**

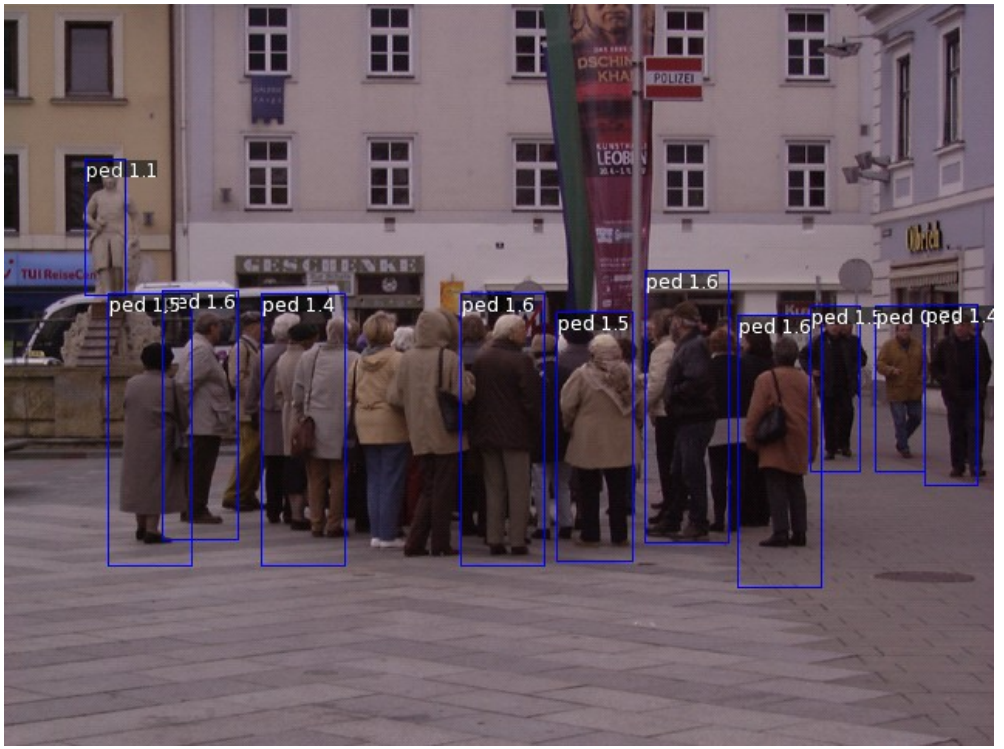


Sifre et al. “Rotation, scaling and deformation invariant scattering...” CVPR 2013

CONV NETS: EXAMPLES

- **Pedestrian detection**

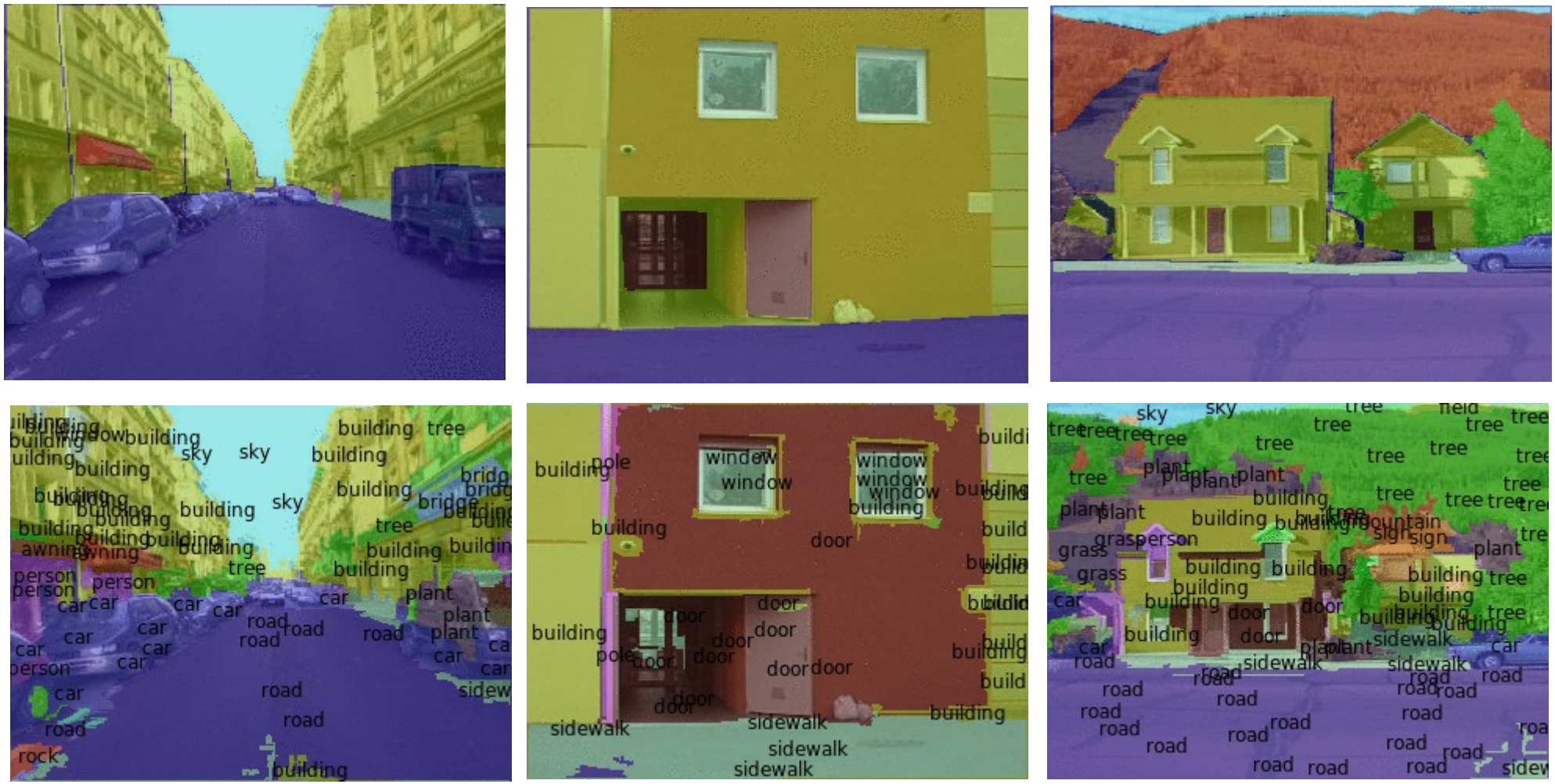




Sermanet et al. “Pedestrian detection with unsupervised multi-stage..” CVPR 2013

CONV NETS: EXAMPLES

- **Scene Parsing**

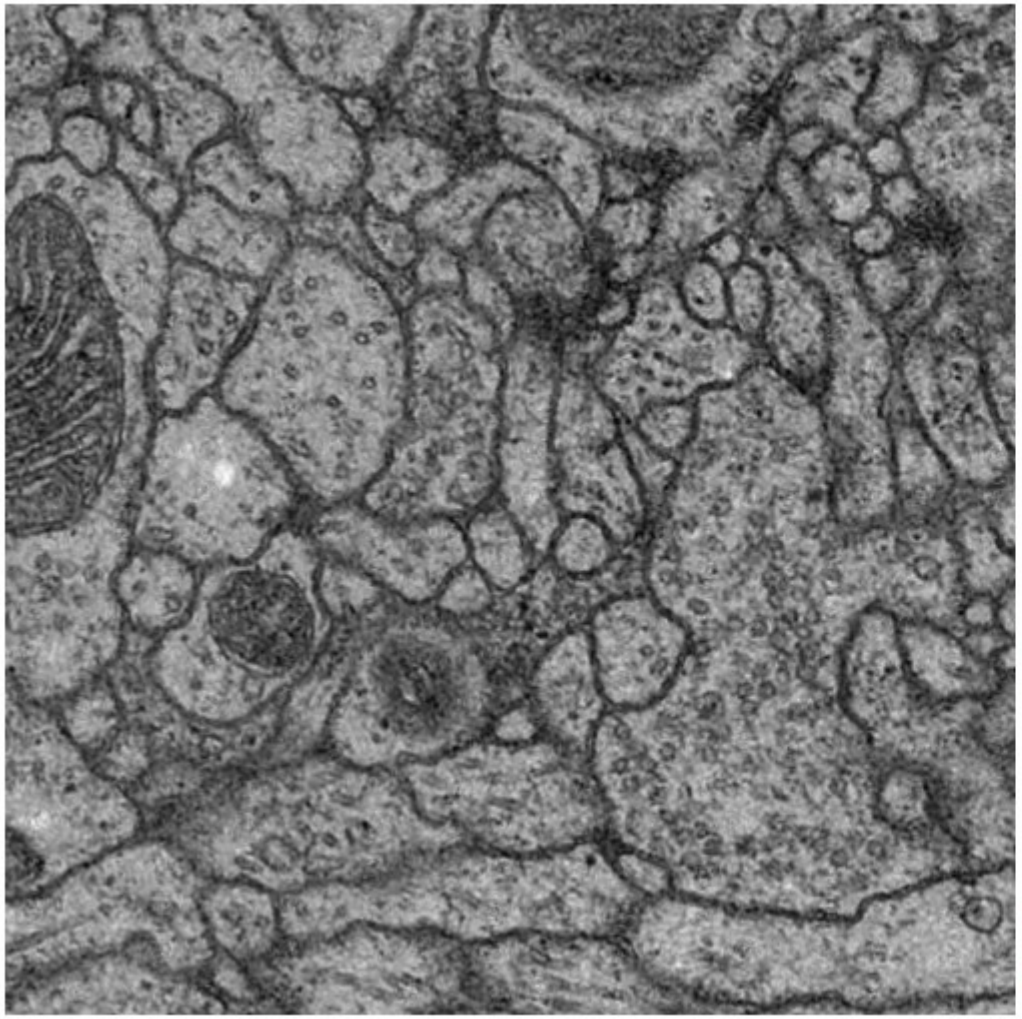


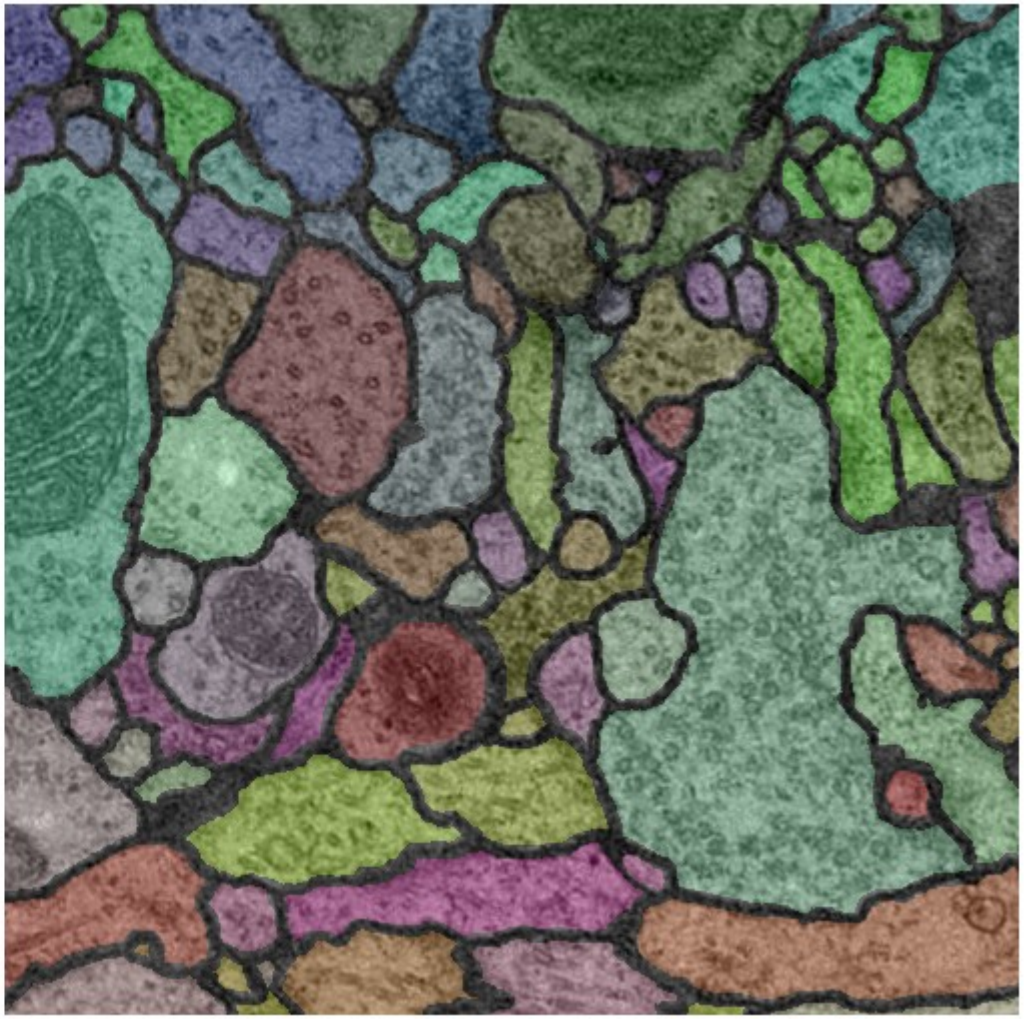
Farabet et al. “Learning hierarchical features for scene labeling” PAMI 2013

Pinheiro et al. “Recurrent CNN for scene parsing” arxiv 2013

CONV NETS: EXAMPLES

- **Segmentation 3D volumetric images**





Ciresan et al. “DNN segment neuronal membranes...” NIPS 2012
Turaga et al. “Maximin learning of image segmentation” NIPS 2009

CONV NETS: EXAMPLES

- **Action recognition from videos**

Taylor et al. “Convolutional learning of spatio-temporal features” ECCV 2010
Karpathy et al. “Large-scale video classification with CNNs” CVPR 2014
Simonyan et al. “Two-stream CNNs for action recognition in videos” arXiv 2014

CONV NETS: EXAMPLES

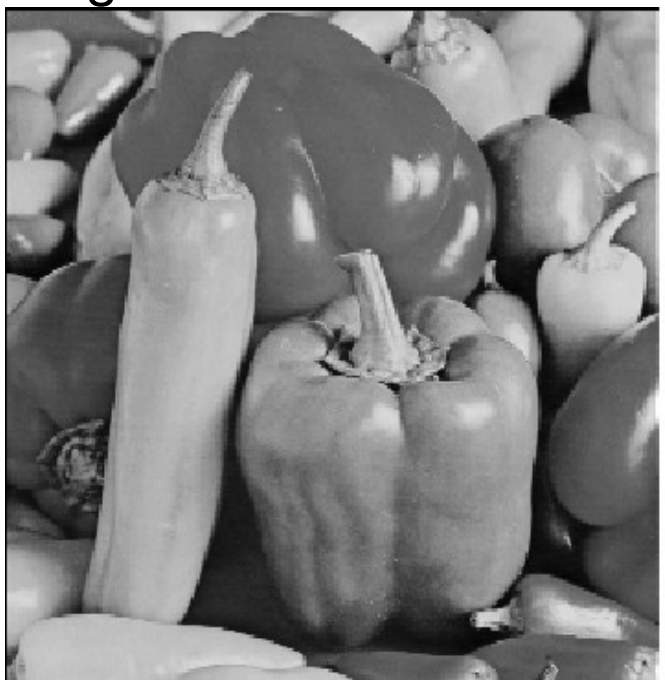
- **Robotics**

Sermanet et al. “Mapping and planning ...with long range perception” IROS 2008

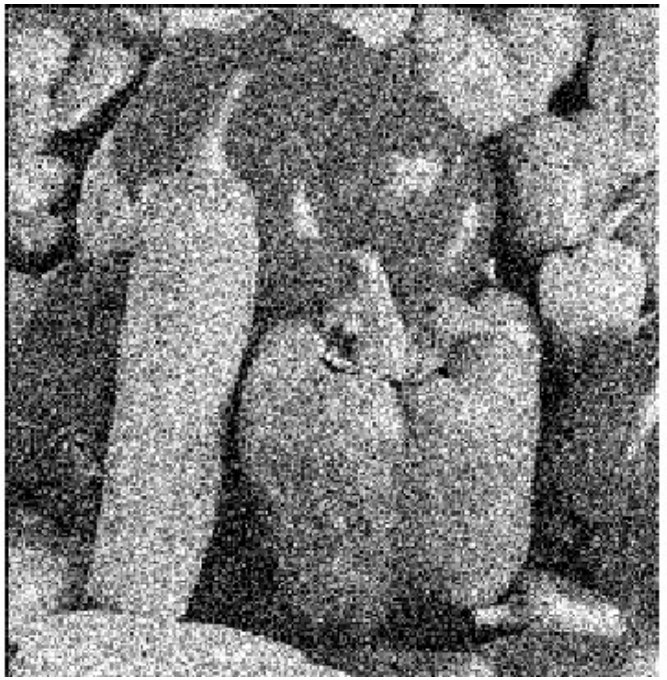
CONV NETS: EXAMPLES

- **Denoising**

original



noised

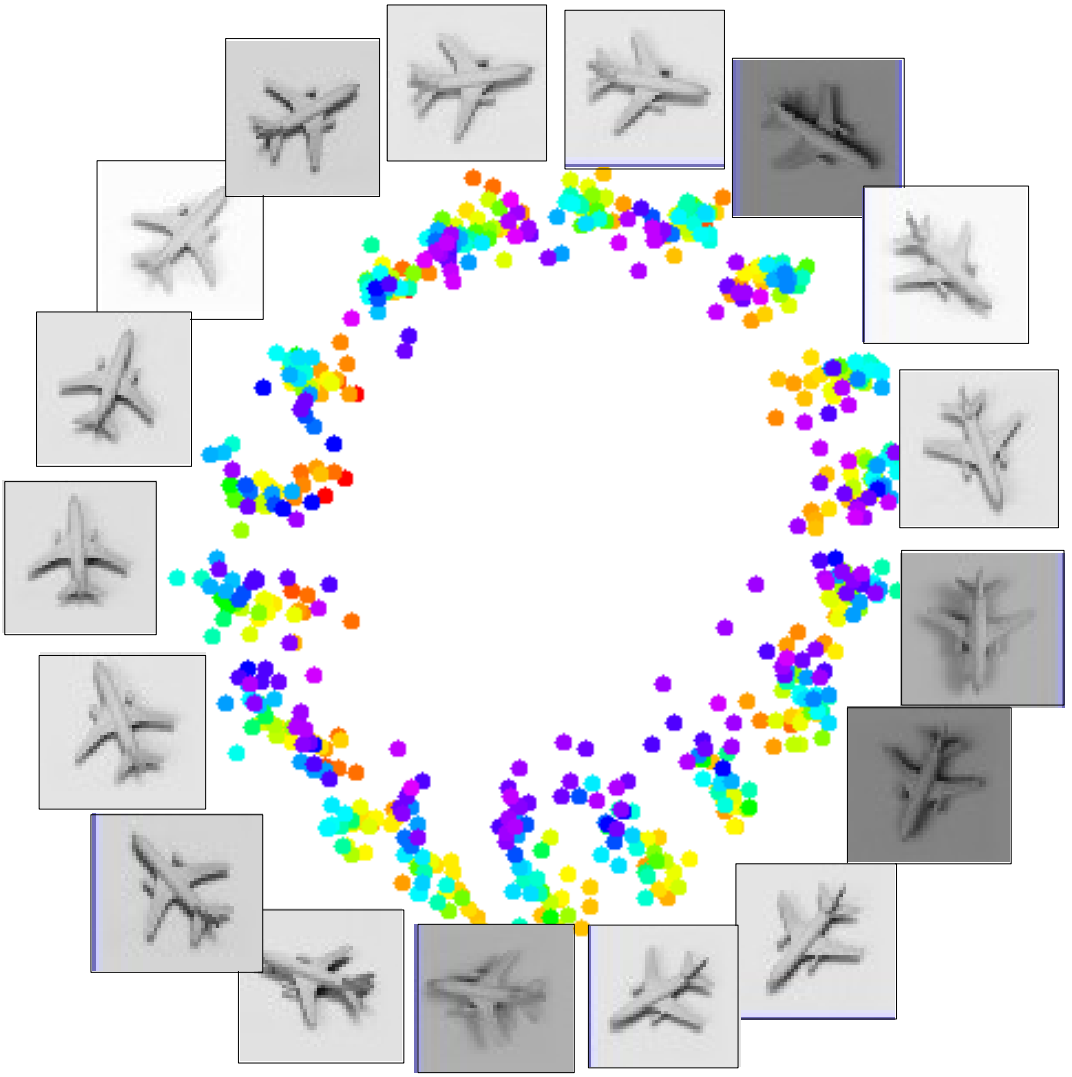


denoised

Burger et al. “Can plain NNs compete with BM3D?” CVPR 2012

CONV NETS: EXAMPLES

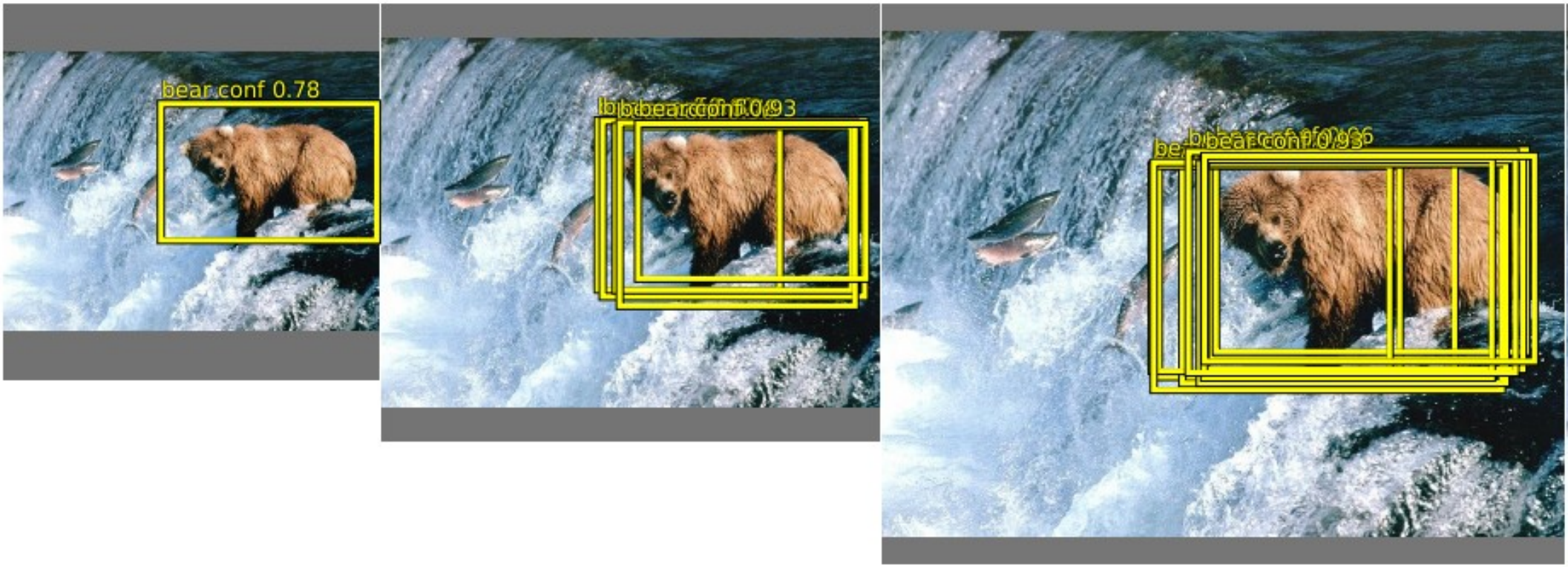
- **Dimensionality reduction / learning embeddings**



Hadsell et al. “Dimensionality reduction by learning an invariant mapping” CVPR 2006

CONV NETS: EXAMPLES

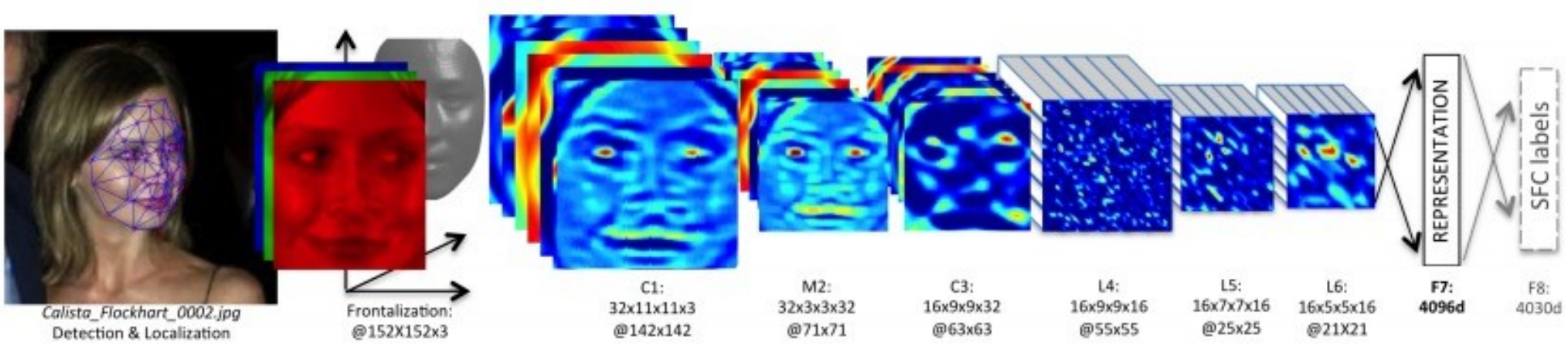
- **Object detection**



Sermanet et al. “OverFeat: Integrated recognition, localization, ...” arxiv 2013
Girshick et al. “Rich feature hierarchies for accurate object detection...” arxiv 2013
Szegedy et al. “DNN for object detection” NIPS 2013

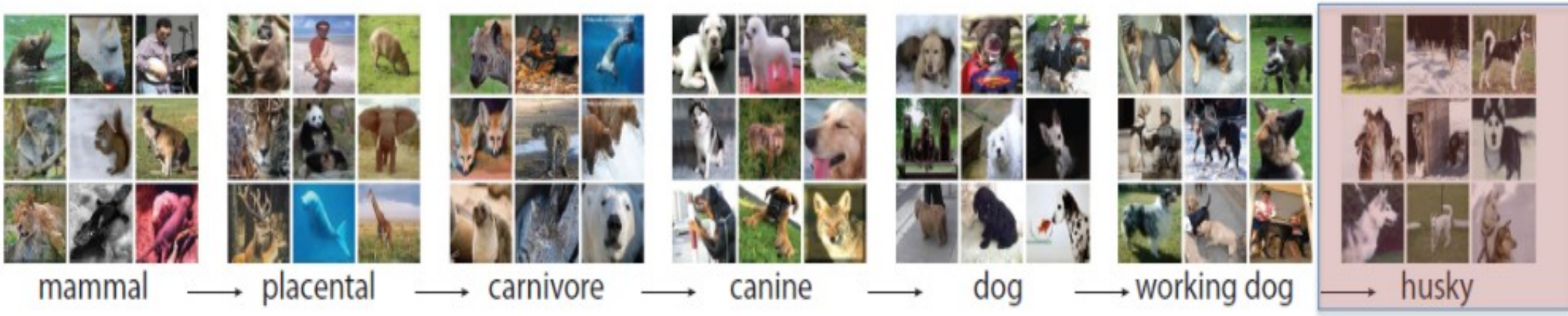
CONV NETS: EXAMPLES

- **Face Verification & Identification**



Taigman et al. "DeepFace..." CVPR 2014

Dataset: ImageNet 2012



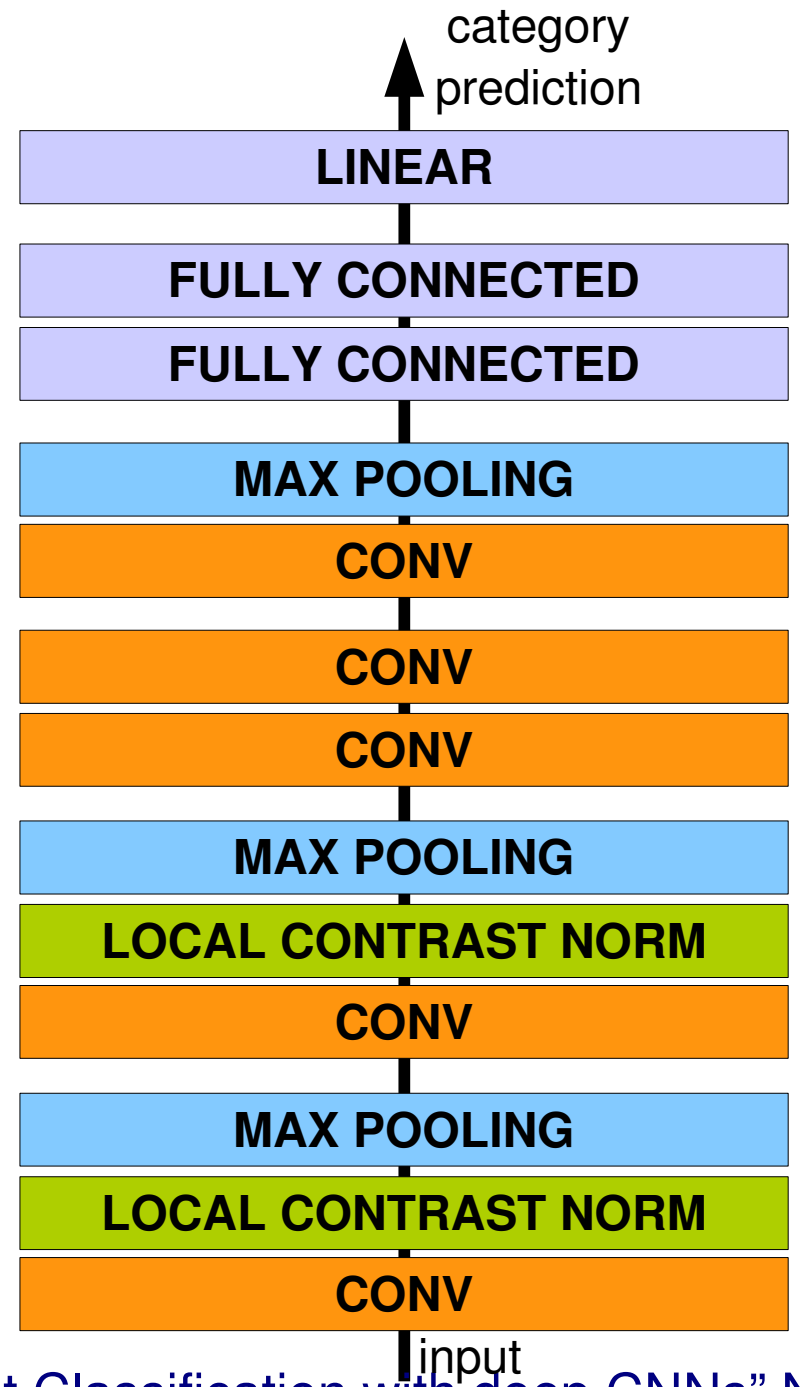
- S: (n) Eskimo dog, **husky** (breed of heavy-coated Arctic sled dog)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
      - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
        - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
          - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
            - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
              - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
                - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
                  - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                    - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
                      - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                        - S: (n) living thing, animate thing (a living (or once living) entity)
                          - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                            - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                              - S: (n) physical entity (an entity that has physical existence)
                                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Deng et al. “Imagenet: a large scale hierarchical image database” CVPR 2009

ImageNet

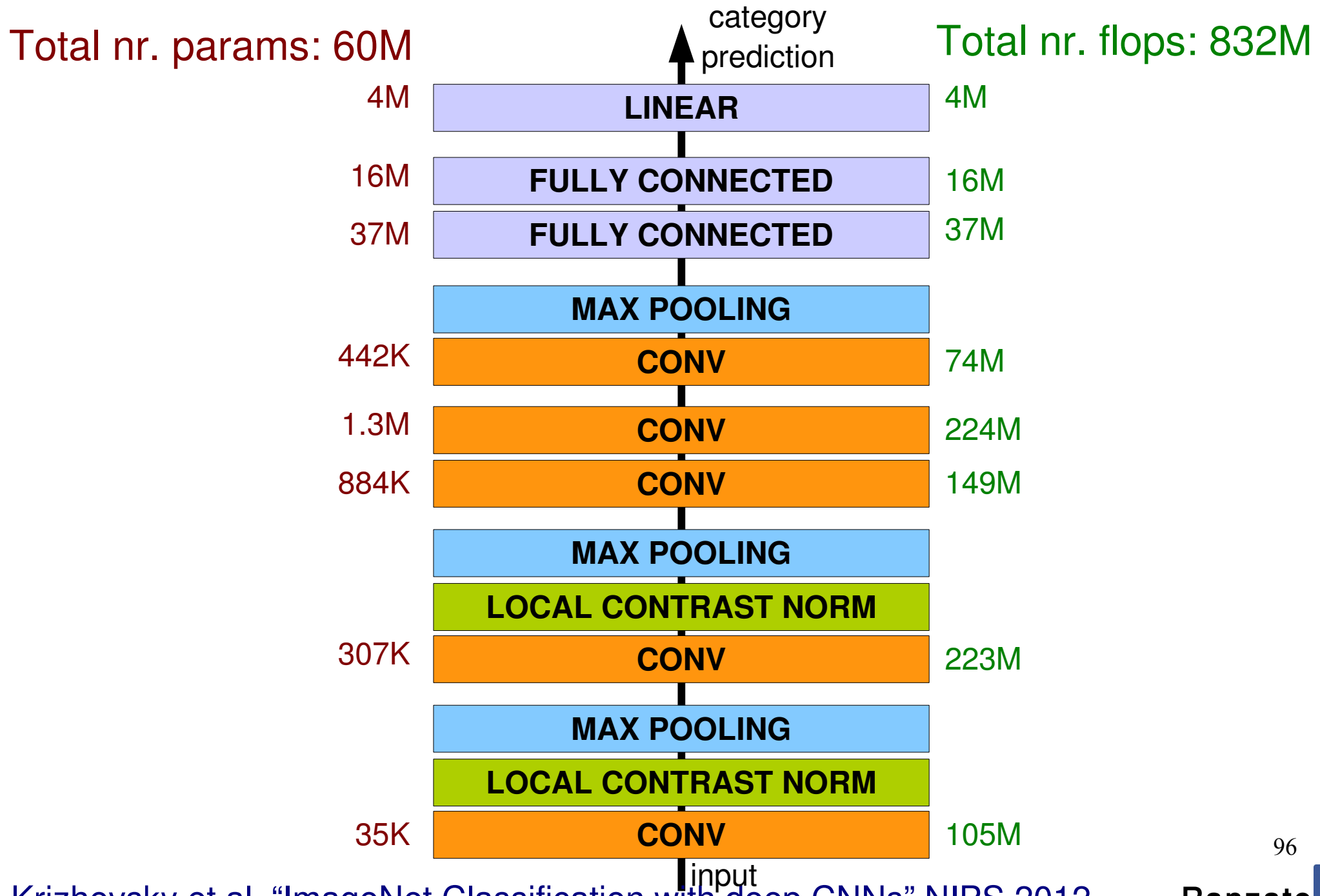
Examples of hammer:

Architecture for Classification



Krizhevsky et al. “ImageNet Classification with deep CNNs” NIPS 2012

Architecture for Classification



Krizhevsky et al. “ImageNet Classification with deep CNNs” NIPS 2012

Optimization

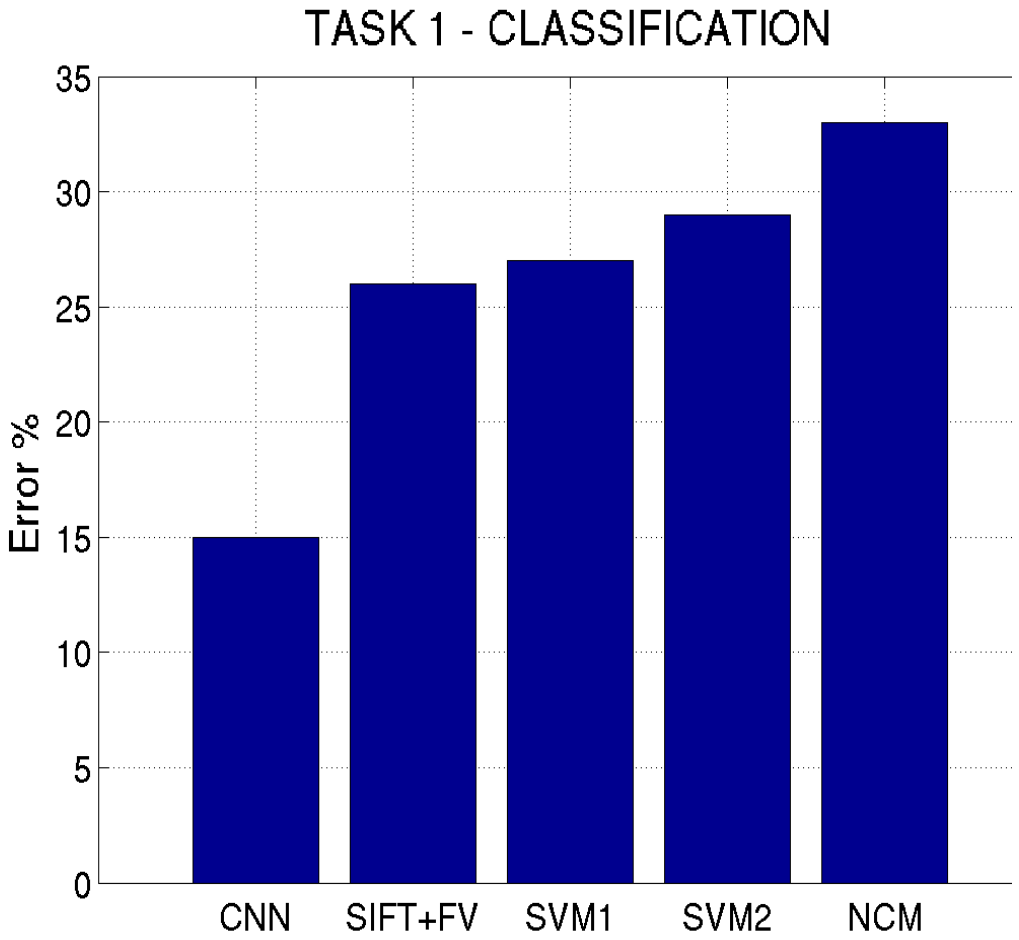
SGD with momentum:

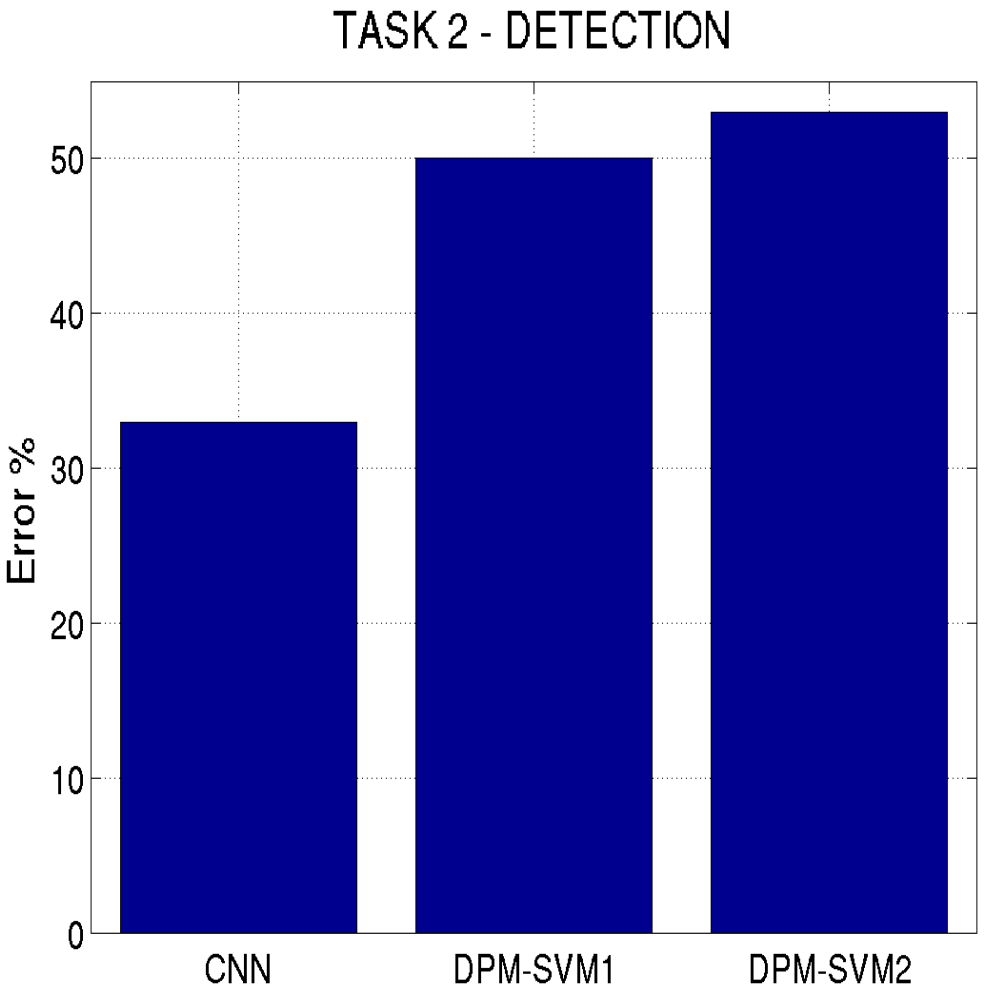
- Learning rate = 0.01
- Momentum = 0.9

Improving generalization by:

- Weight sharing (convolution)
- Input distortions
- Dropout = 0.5
- Weight decay = 0.0005

Results: ILSVRC 2012





Krizhevsky et al. “ImageNet Classification with deep CNNs” NIPS 2012

mite

mite
black widow
cockroach
tick
starfish

container ship

container ship
lifeboat
amphibian
fireboat
drilling platform

motor scooter

motor scooter
go-kart
moped
bumper car
golfcart

leopard

leopard
jaguar
cheetah
snow leopard
Egyptian cat

grille

convertible
grille
pickup
beach wagon
fire engine

mushroom

agaric
mushroom
jelly fungus
gill fungus
dead-man's-fingers

cherry

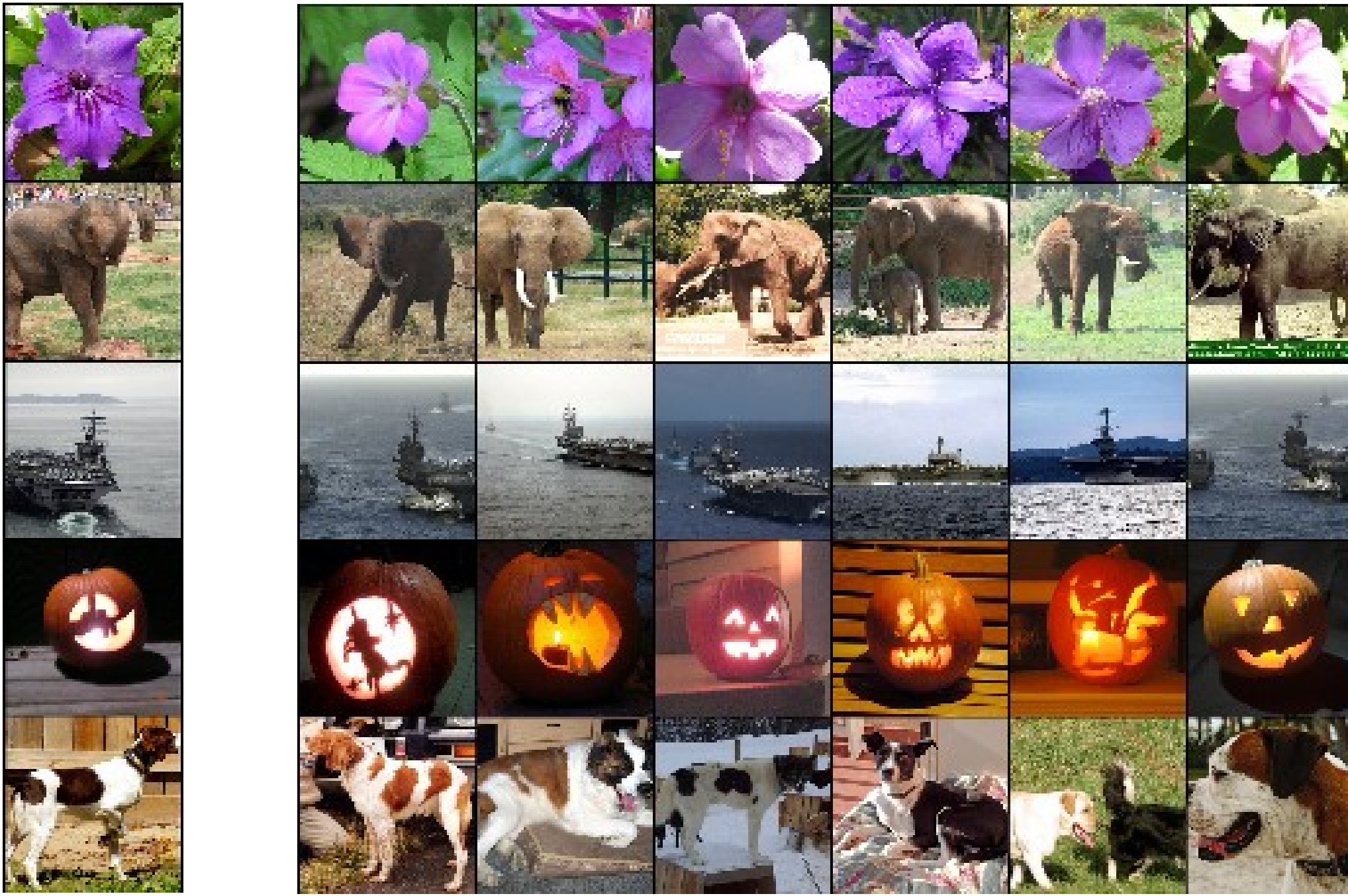
dalmatian
grape
elderberry
ffordshire bullterrier
currant

Madagascar cat

squirrel monkey
spider monkey
titi
indri
howler monkey

TEST IMAGE

RETRIEVED IMAGES



Demo of classifier by Matt Zeiler & Rob Fergus:

http://horatio.cs.nyu.edu/

File Edit View History Bookmarks Tools Help

Image Classifier Demo

horatio.cs.nyu.edu/index.html

matt zeiler horatio

Most Visited Getting Started Latest Headlines Click Here!

Image Classifier Demo | Demo | About | Terms

Image Classifier Demo

NYU

Upload your images to have them classified by a machine! Upload multiple images using the button below or dropping them on this page. The predicted objects will be refreshed automatically. Images are resized such that the smallest dimension becomes 256, then the center 256x256 crop is used. More about the demo can be found here .

+ Upload Images | Remove All | Show help tips

I agree to the Terms of Use

Predicted objects:

1. Lion, King Of Beasts, Panthera Leo (0.34)
2. Hartebeest (0.19)
3. Hyena, Hyaena (0.16)
4. Arabian Camel, Dromedary, Camelus Dromedarius (0.06)
5. Dingo, Warrigal, Warragal, Canis Dingo (0.04)

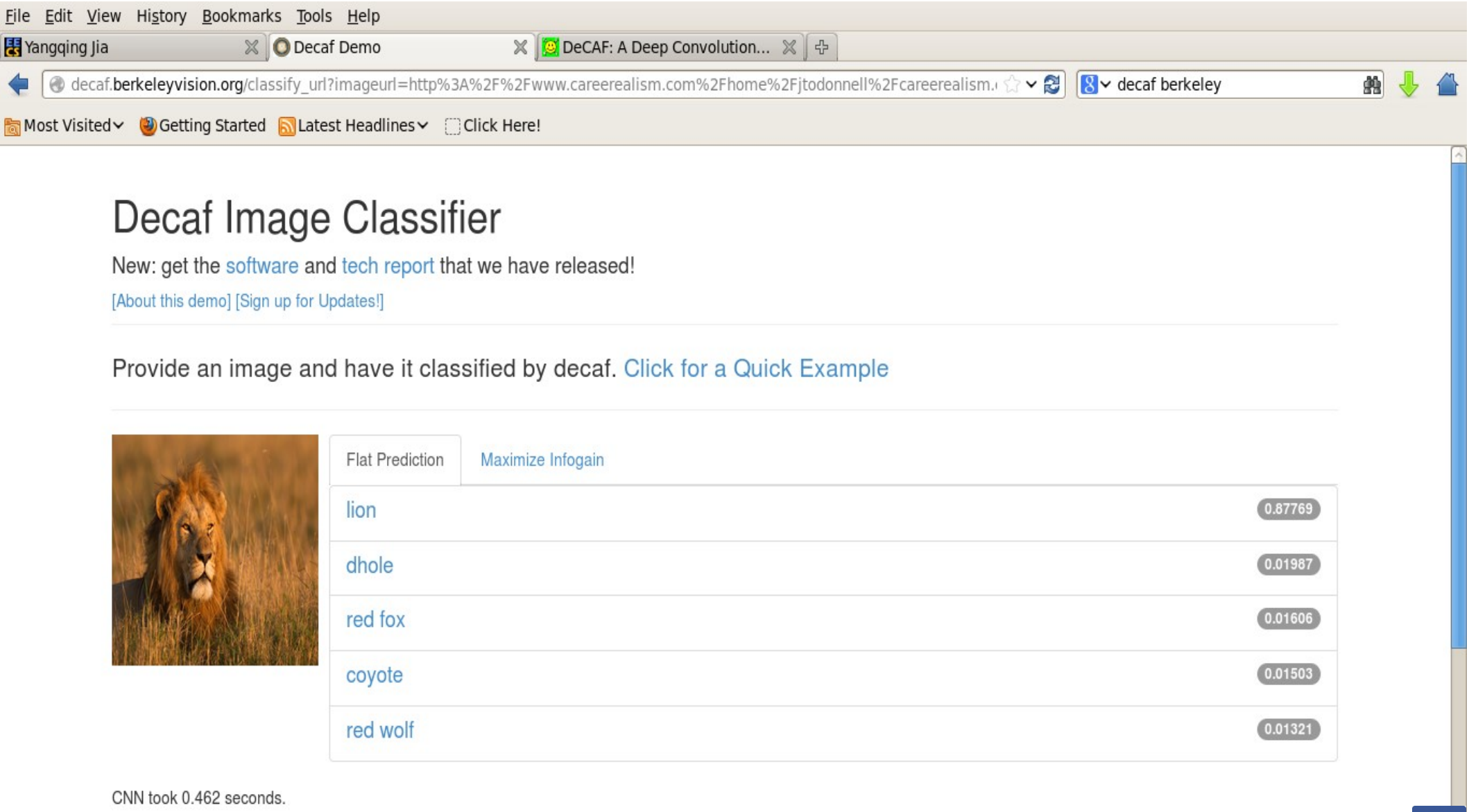
Remove

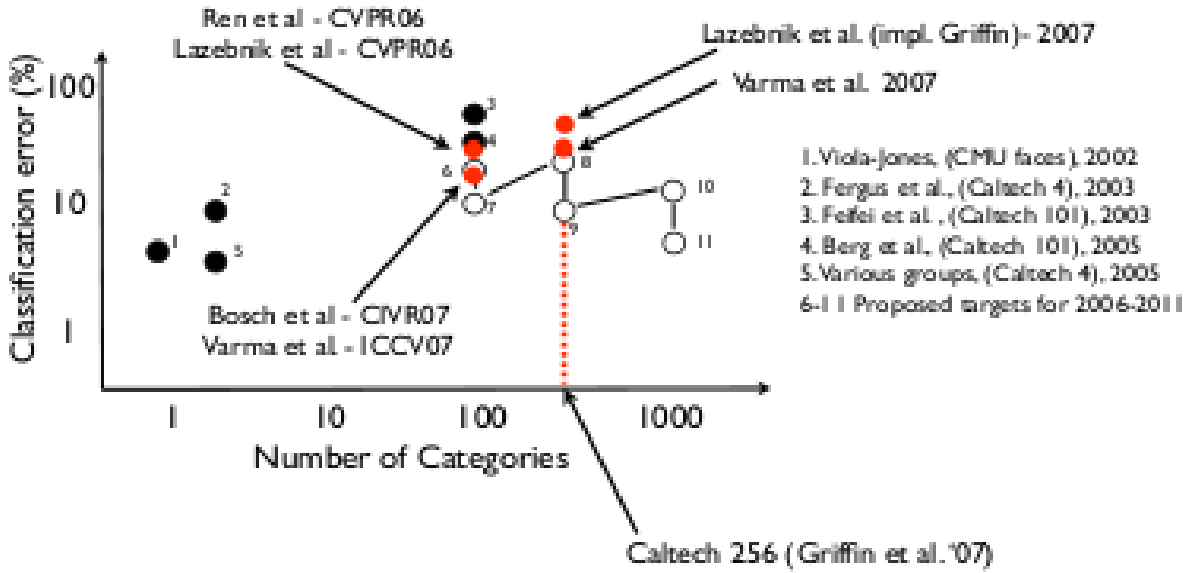
Other objects:

Ranzato

Demo of classifier by Yangqing Jia & Trevor Darrell:

http://decafberkeleyvision.org/





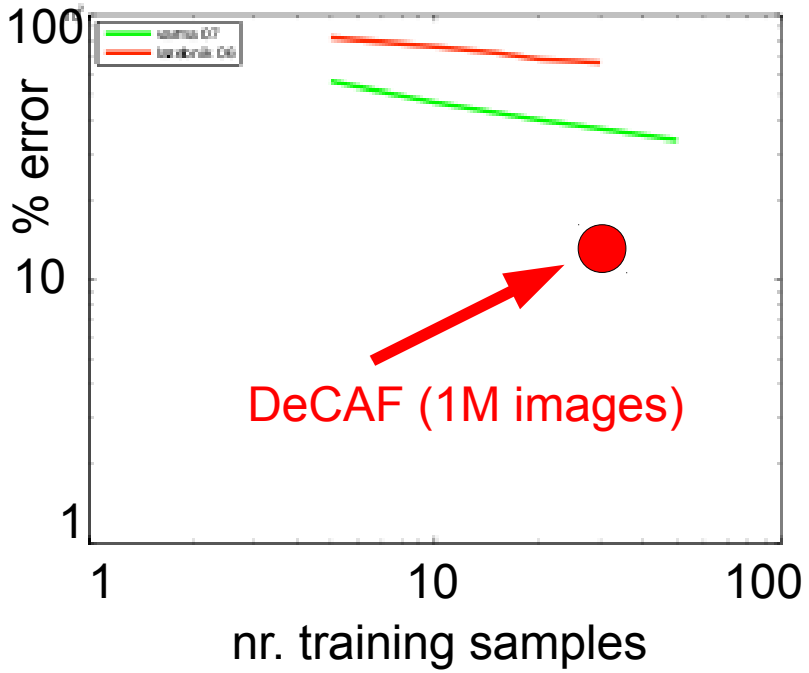


Figure 3: How well are we doing? (Left) Classification performance has seen steady improvement in the last few years, both in the number of categories on which algorithms are tested and in classification error rates. (Right) Performance of the best 2006 [Lazebnik et al., 2006] and the best 2007 algorithm [Varma, 2007] are compared here (classification error rates vs number of training examples). One may notice the significant year-on-year progress (see also left panel). Extrapolation enthusiasts may calculate that 10^8 training examples would be sufficient to achieve 1% error rates with current algorithms. Furthermore, if the pace of year-on-year progress is constant on this log scale chart, 1% error rates with 30 training examples will be achieved in 8-10 years.

Excerpt from Perona Visual Recognition 2007

Donahue, Jia et al. DeCAF arXiv 1310.1531 2013

Outline

- Supervised Neural Networks
- Convolutional Neural Networks
- Examples
- Tips

Choosing The Architecture

- Task dependent
- Cross-validation
- [Convolution $\rightarrow$ LCN $\rightarrow$ pooling]* + fully connected layer
- The more data: the more layers and the more kernels
  - Look at the number of parameters at each layer
  - Look at the number of flops at each layer
- Computational resources
- Be creative :)

How To Optimize

- SGD (with momentum) usually works very well

- Pick learning rate by running on a subset of the data

  Bottou "Stochastic Gradient Tricks" Neural Networks 2012

  - Start with large learning rate and divide by 2 until loss does not diverge
  - Decay learning rate by a factor of ~1000 or more by the end of training

- Use  non-linearity

- Initialize parameters so that each feature across layers has similar variance. Avoid units in saturation.

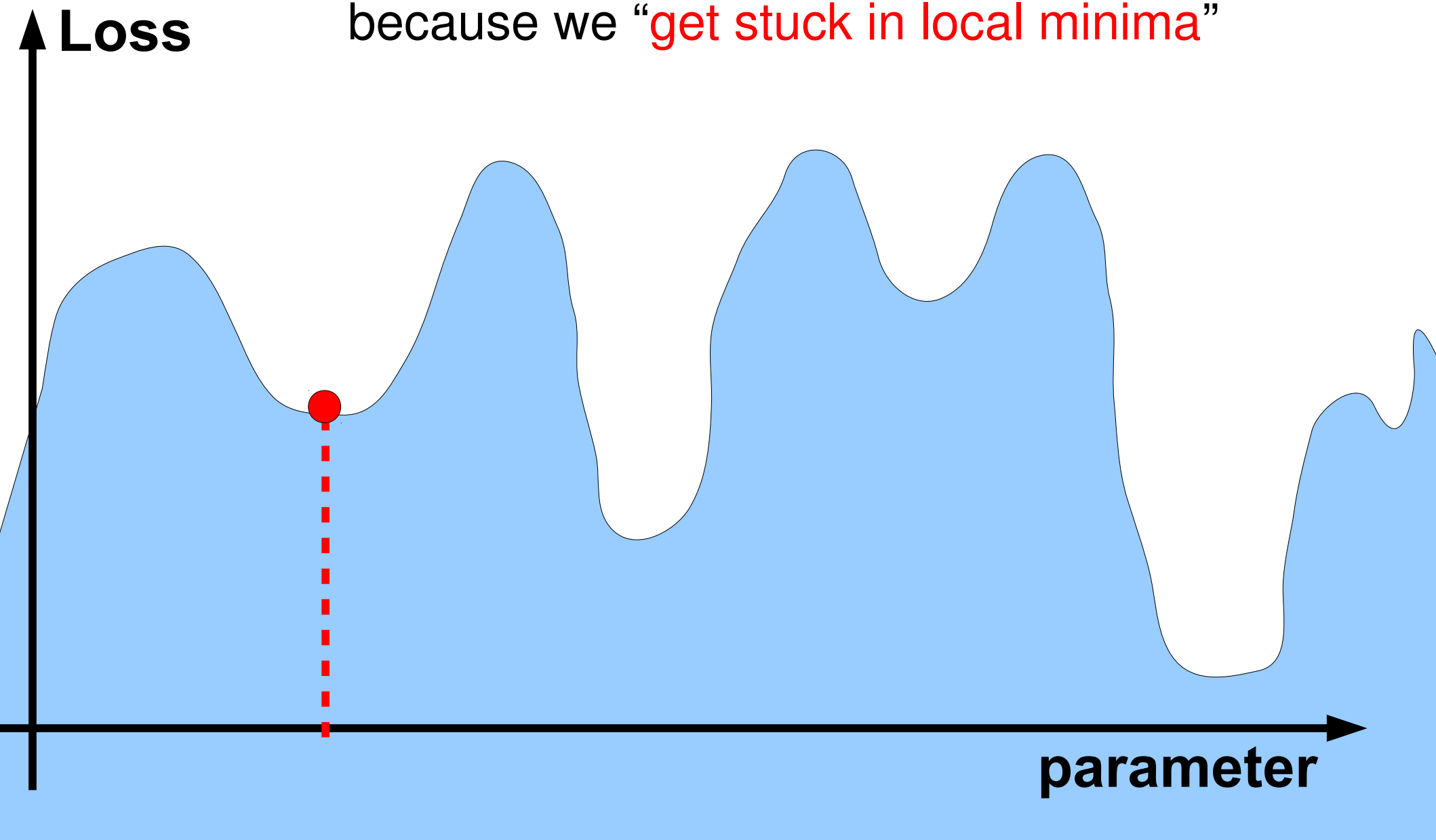
Improving Generalization

- Weight sharing (greatly reduce the number of parameters)
- Data augmentation (e.g., jittering, noise injection, etc.)
- Dropout

  Hinton et al. “Improving Nns by preventing co-adaptation of feature detectors” arxiv 2012
- Weight decay (L2, L1)
- Sparsity in the hidden units
- Multi-task (unsupervised learning)

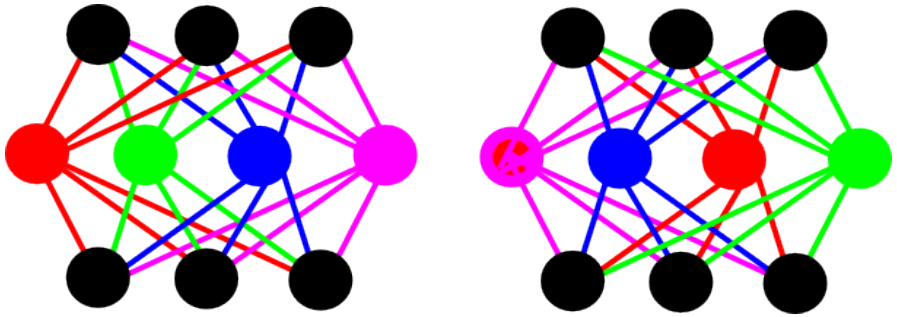
ConvNets: till 2012

Common wisdom: training does not work because we “get stuck in local minima”

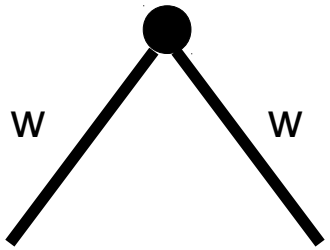


ConvNets: today

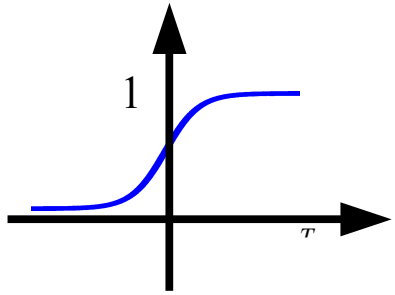
Local minima are all similar, there are long plateaus, it can take long time to break symmetries.



input/output invariant to permutations



breaking ties between parameters



Saturating units

Dauphin et al. “Identifying and attacking the saddle point problem..” arXiv 2014

Neural Net Optimization is...

ConvNets: today

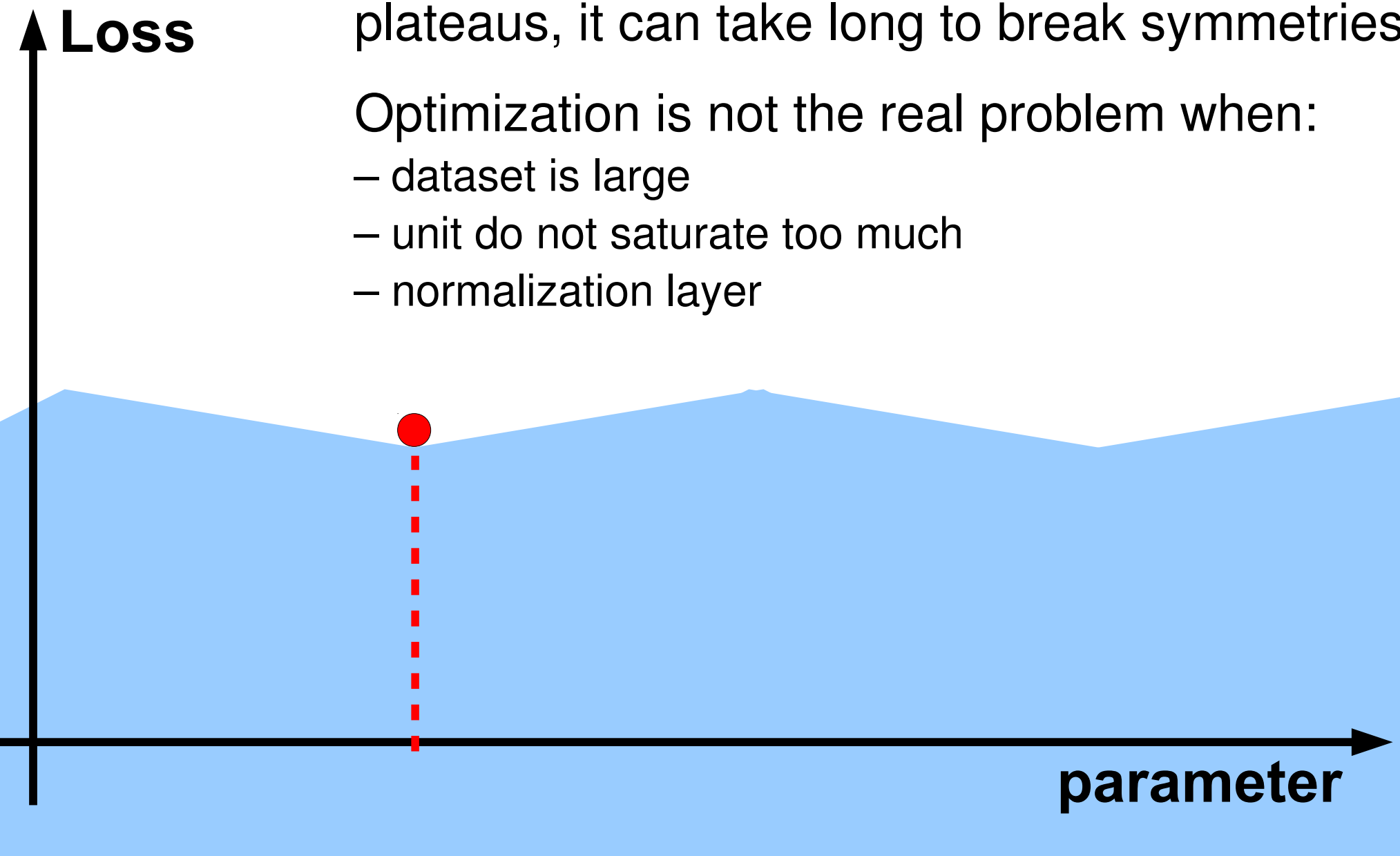
Local minima are all similar, there are long plateaus, it can take long to break symmetries.

Optimization is not the real problem when:

– dataset is large

– unit do not saturate too much

– normalization layer



ConvNets: today

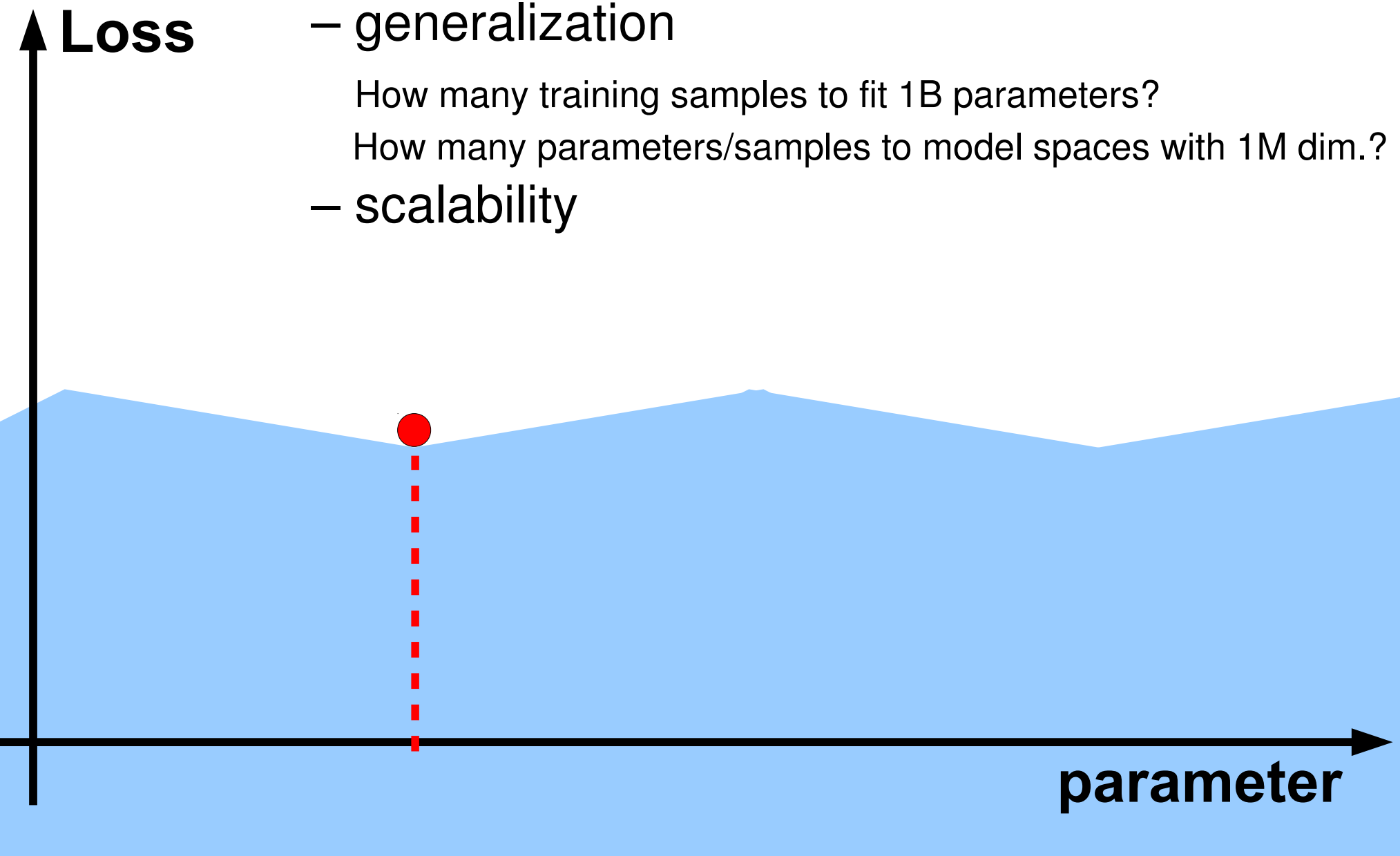
Today's belief is that the challenge is about:

– generalization

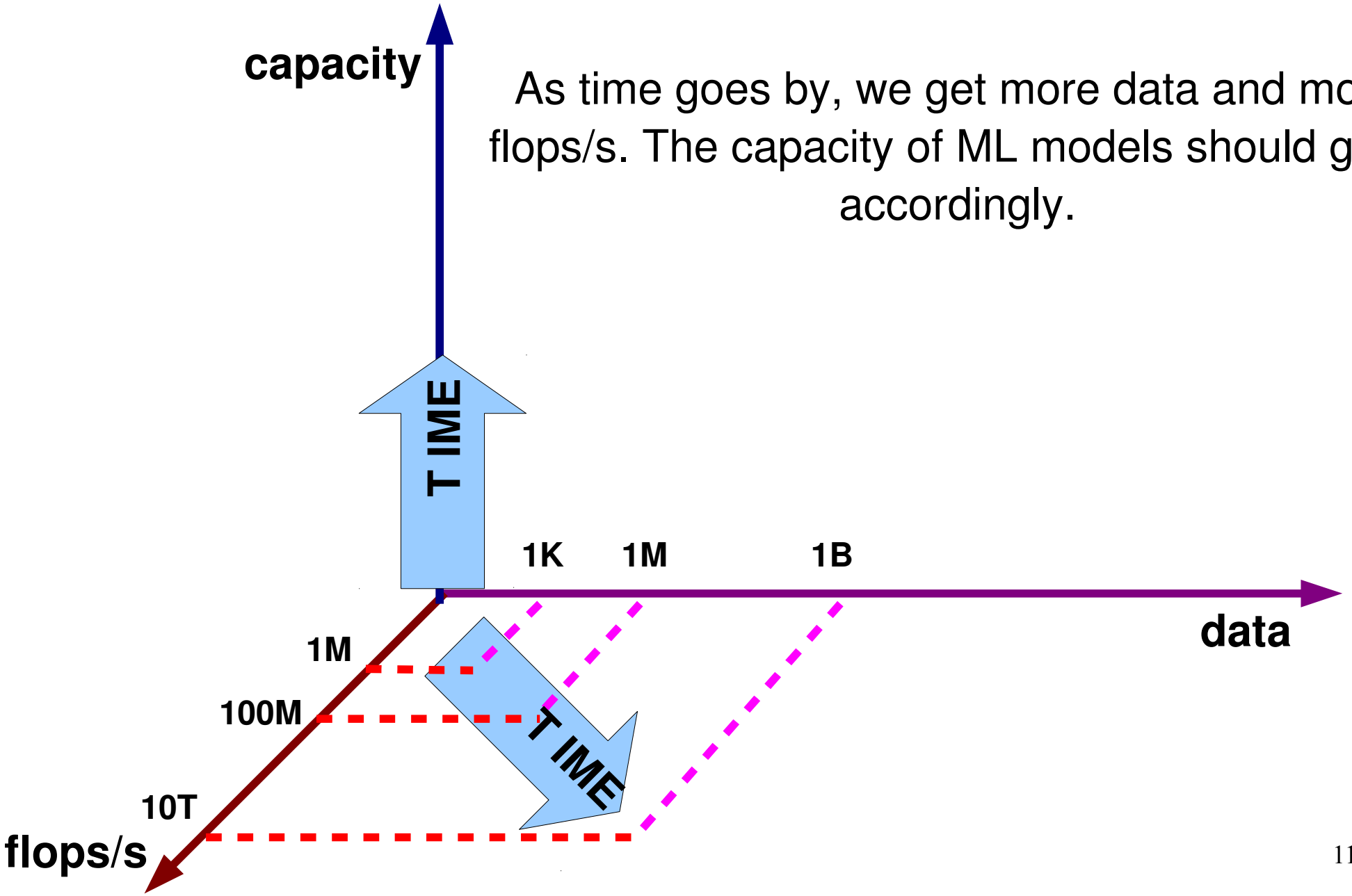
How many training samples to fit 1B parameters?

How many parameters/samples to model spaces with 1M dim.?

– scalability

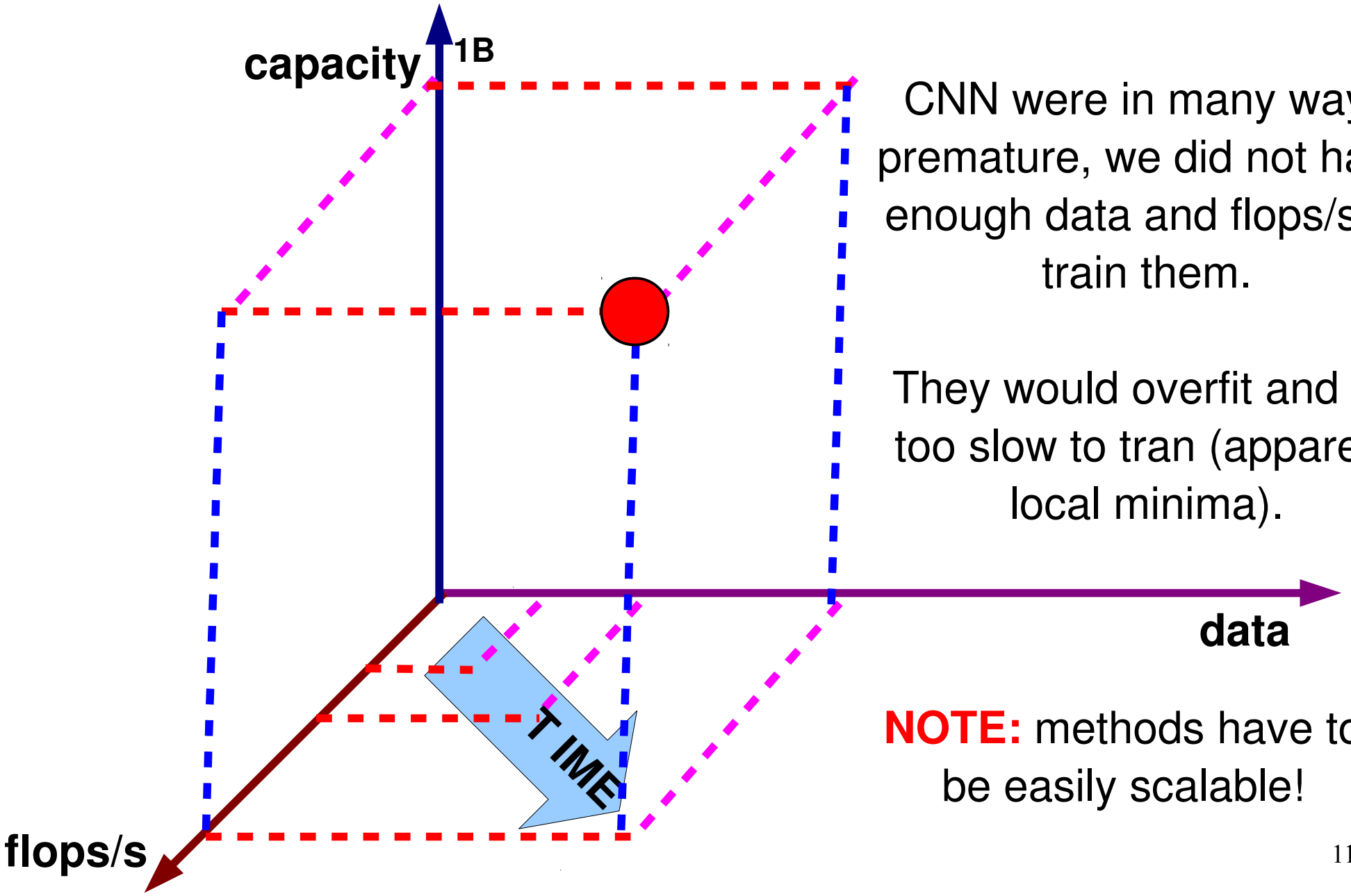


ConvNets: Why so successful today?



As time goes by, we get more data and more flops/s. The capacity of ML models should grow accordingly.

ConvNets: Why so successful today?



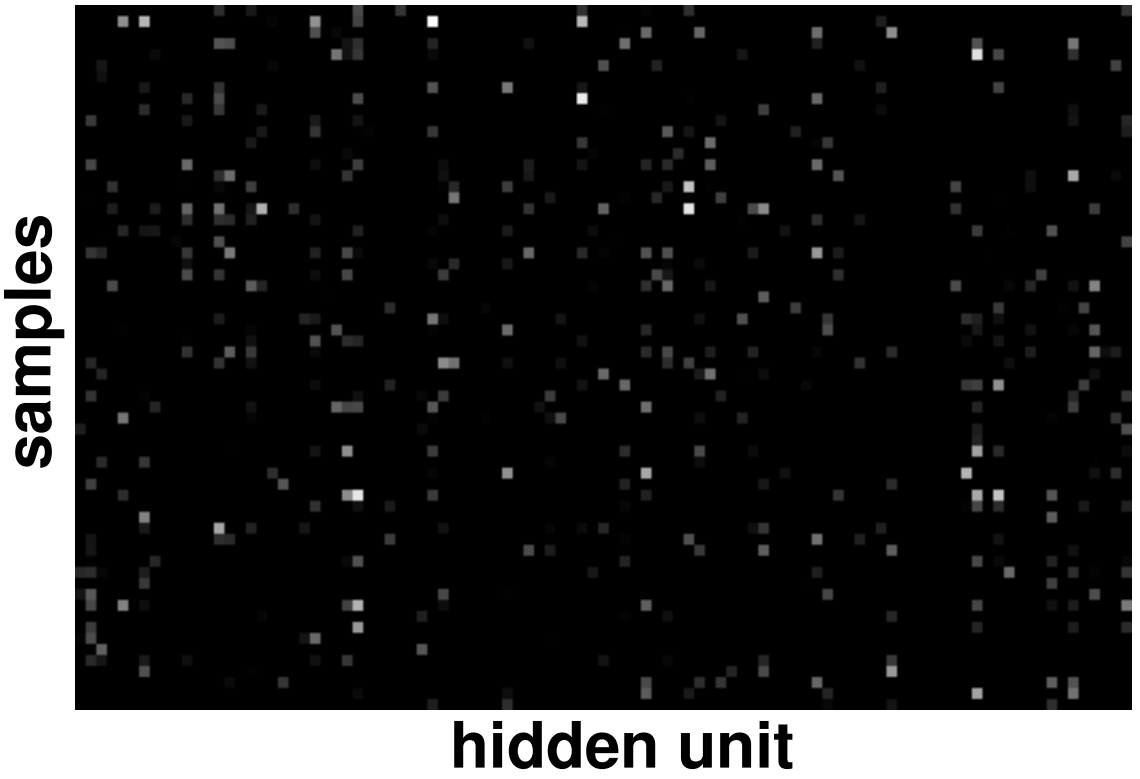
CNN were in many ways premature, we did not have enough data and flops/s to train them.

They would overfit and be too slow to tran (apparent local minima).

NOTE: methods have to be easily scalable!

Good To Know

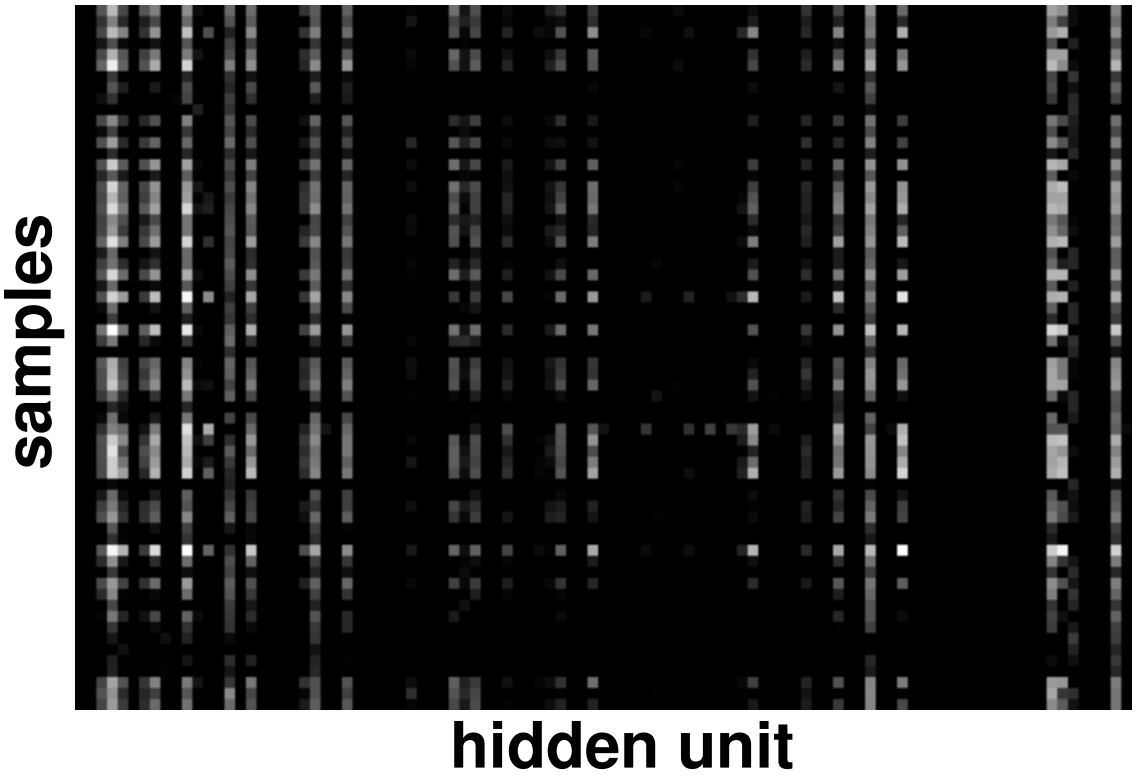
- Check gradients numerically by finite differences
- Visualize features (feature maps need to be uncorrelated) and have high variance.



Good training: hidden units are sparse across samples and across features.

Good To Know

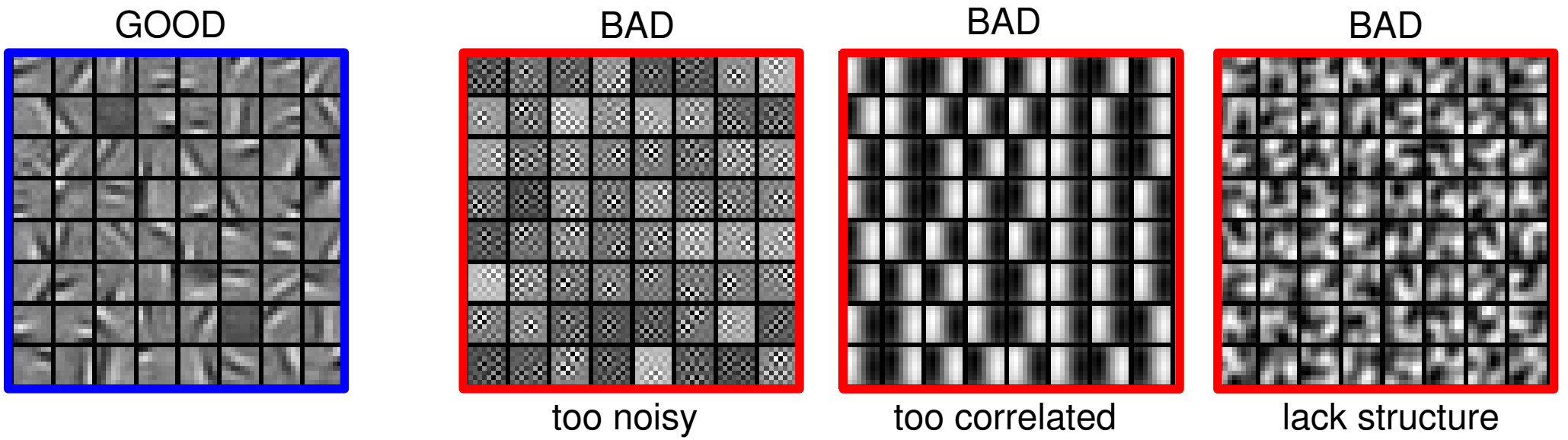
- Check gradients numerically by finite differences
- Visualize features (feature maps need to be uncorrelated) and have high variance.



Bad training: many hidden units ignore the input and/or exhibit strong correlations.

Good To Know

- Check gradients numerically by finite differences
- Visualize features (feature maps need to be uncorrelated) and have high variance.
- Visualize parameters



Good training: learned filters exhibit structure and are uncorrelated.

Zeiler, Fergus "Visualizing and understanding CNNs" arXiv 2013
Simonyan, Vedaldi, Zisserman "Deep inside CNNs: visualizing image classification models.." ICLR 2014

Good To Know

- Check gradients numerically by finite differences
- Visualize features (feature maps need to be uncorrelated) and have high variance.
- Visualize parameters
- Measure error on both training and validation set.
- Test on a small subset of the data and check the error $\rightarrow 0$.

What If It Does Not Work?

- Training diverges:
  - Learning rate may be too large → decrease learning rate
  - BPROP is buggy → numerical gradient checking
- Parameters collapse / loss is minimized but accuracy is low
  - Check loss function:
    - Is it appropriate for the task you want to solve?
    - Does it have degenerate solutions? Check "pull-up" term.
- Network is underperforming
  - Compute flops and nr. params. → if too small, make net larger
  - Visualize hidden units/params → fix optmization
- Network is too slow
  - Compute flops and nr. params. → GPU,distrib. framework, make net smaller

Summary

- Supervised learning: today it is the most successful set up.

ConvNets are used for a great variety of tasks.

- Optimization
  - Don't we get stuck in local minima? No, they are all the same!
  - In large scale applications, local minima are even less of an issue.
- Scaling
  - GPUs
  - Distributed framework (Google)
  - Better optimization techniques
- Generalization on small datasets (curse of dimensionality):
  - data augmentation
  - weight decay
  - dropout
  - unsupervised learning
  - multi-task learning

SOFTWARE

Torch7: learning library that supports neural net training

http://www.torch.ch
http://code.cogbits.com/wiki/doku.php (tutorial with demos by C. Farabet)
https://github.com/sermanet/OverFeat

Python-based learning library (U. Montreal)

- http://deeplearning.net/software/theano/ (does automatic differentiation)

Efficient CUDA kernels for ConvNets (Krizhevsky)

– code.google.com/p/cuda-convnet

Caffe (Yangqing Jia)

– http://caffe.berkeleyvision.org

REFERENCES

Convolutional Nets

– LeCun, Bottou, Bengio and Haffner: Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324, November 1998

- Krizhevsky, Sutskever, Hinton “ImageNet Classification with deep convolutional neural networks” NIPS 2012

– Jarrett, Kavukcuoglu, Ranzato, LeCun: What is the Best Multi-Stage Architecture for Object Recognition?, Proc. International Conference on Computer Vision (ICCV'09), IEEE, 2009

- Kavukcuoglu, Sermanet, Boureau, Gregor, Mathieu, LeCun: Learning Convolutional Feature Hierachies for Visual Recognition, Advances in Neural Information Processing Systems (NIPS 2010), 23, 2010

– see yann.lecun.com/exdb/publis for references on many different kinds of convnets.

– see http://www.cmap.polytechnique.fr/scattering/ for scattering networks (similar to convnets but with less learning and stronger mathematical foundations)

– see http://www.idsia.ch/~juergen/ for other references to ConvNets and LSTMs.

REFERENCES

Applications of Convolutional Nets

– Farabet, Couprie, Najman, LeCun. Scene Parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers", ICML 2012

– Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala and Yann LeCun: Pedestrian Detection with Unsupervised Multi-Stage Feature Learning, CVPR 2013

- D. Ciresan, A. Giusti, L. Gambardella, J. Schmidhuber. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. NIPS 2012

- Raia Hadsell, Pierre Sermanet, Marco Scoffier, Ayse Erkan, Koray Kavackuoglu, Urs Muller and Yann LeCun. Learning Long-Range Vision for Autonomous Off-Road Driving, Journal of Field Robotics, 26(2):120-144, 2009

– Burger, Schuler, Harmeling. Image Denoisng: Can Plain Neural Networks Compete with BM3D?, CVPR 2012

– Hadsell, Chopra, LeCun. Dimensionality reduction by learning an invariant mapping, CVPR 2006

– Bergstra et al. Making a science of model search: hyperparameter optimization in hundred of dimensions for vision architectures, ICML 2013

REFERENCES

Latest and Greatest Convolutional Nets

– Girshick, Donahue, Darrell, Malick. “Rich feature hierarchies for accurate object detection and semantic segmentation”, arXiv 2014

–Karpathy, Toderici, Shetty, Leung, Sukthankar, FeiFei “Large-scale video classification with convolutional neural networks, CVPR 2014

- Cadieu, Hong, Yamins, Pinto, Ardila, Solomon, Majaj, DiCarlo. “DNN rival in representation of primate IT cortex for core visual object recognition”. arXiv 2014

- Erhan, Szegedy, Toshev, Anguelov “Scalable object detection using DNN” CVPR 2014

- Dauphin, Pascanu, Gulcehre, Cho, Ganguli, Bengio “Identifying and attacking the saddle point problem in high-dimensional non-convex optimization” arXiv 2014

- Razavian, Azizpour, Sullivan, Carlsson “CNN features off-the-shelf: and astounding baseline for recognition” arXiv 2014

REFERENCES

Deep Learning in general

– deep learning tutorial slides at ICML 2013

– Yoshua Bengio, Learning Deep Architectures for AI, Foundations and Trends in Machine Learning, 2(1), pp.1-127, 2009.

– LeCun, Chopra, Hadsell, Ranzato, Huang: A Tutorial on Energy-Based Learning, in Bakir, G. and Hofman, T. and Schölkopf, B. and Smola, A. and Taskar, B. (Eds), Predicting Structured Data, MIT Press, 2006

THANK YOU