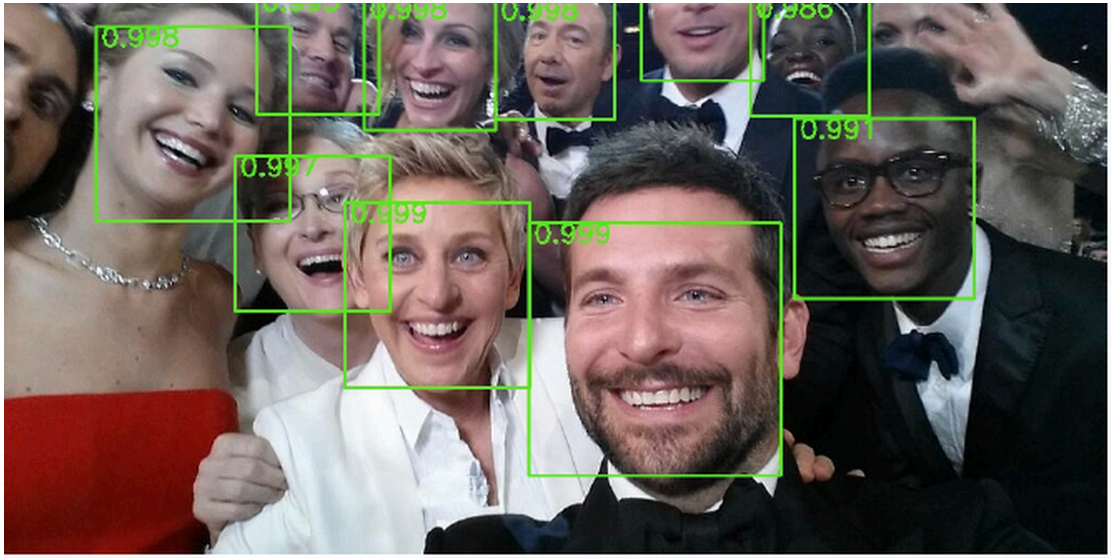
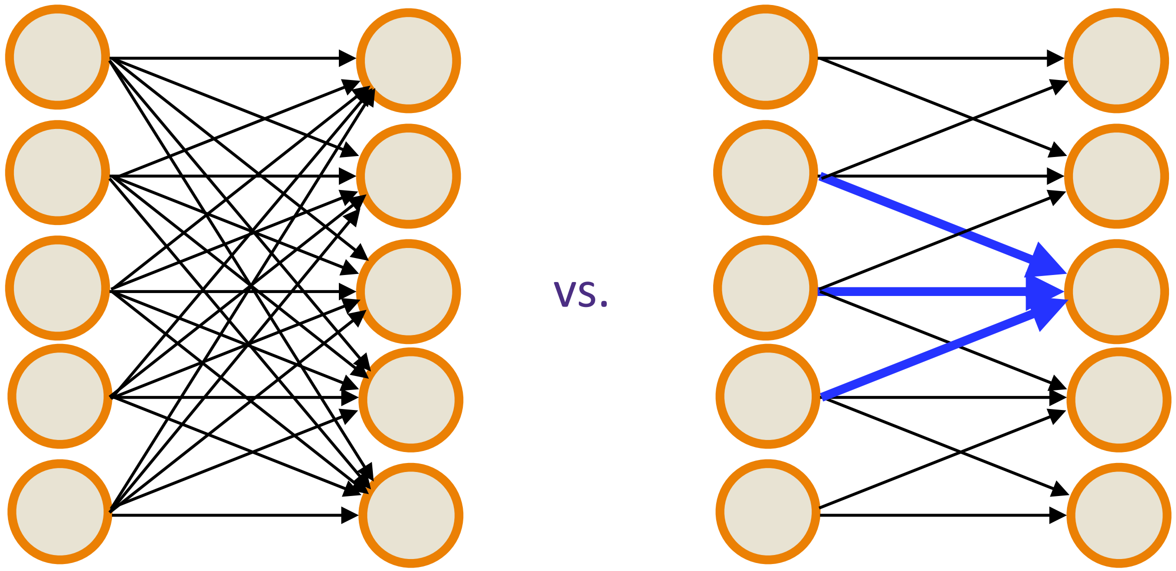
Convolutional Neural Networks

Neural Network Architecture

Objects are often **localized in space** so to find the faces in an image, not every pixel is important for classification—makes sense to drag a window across an image.

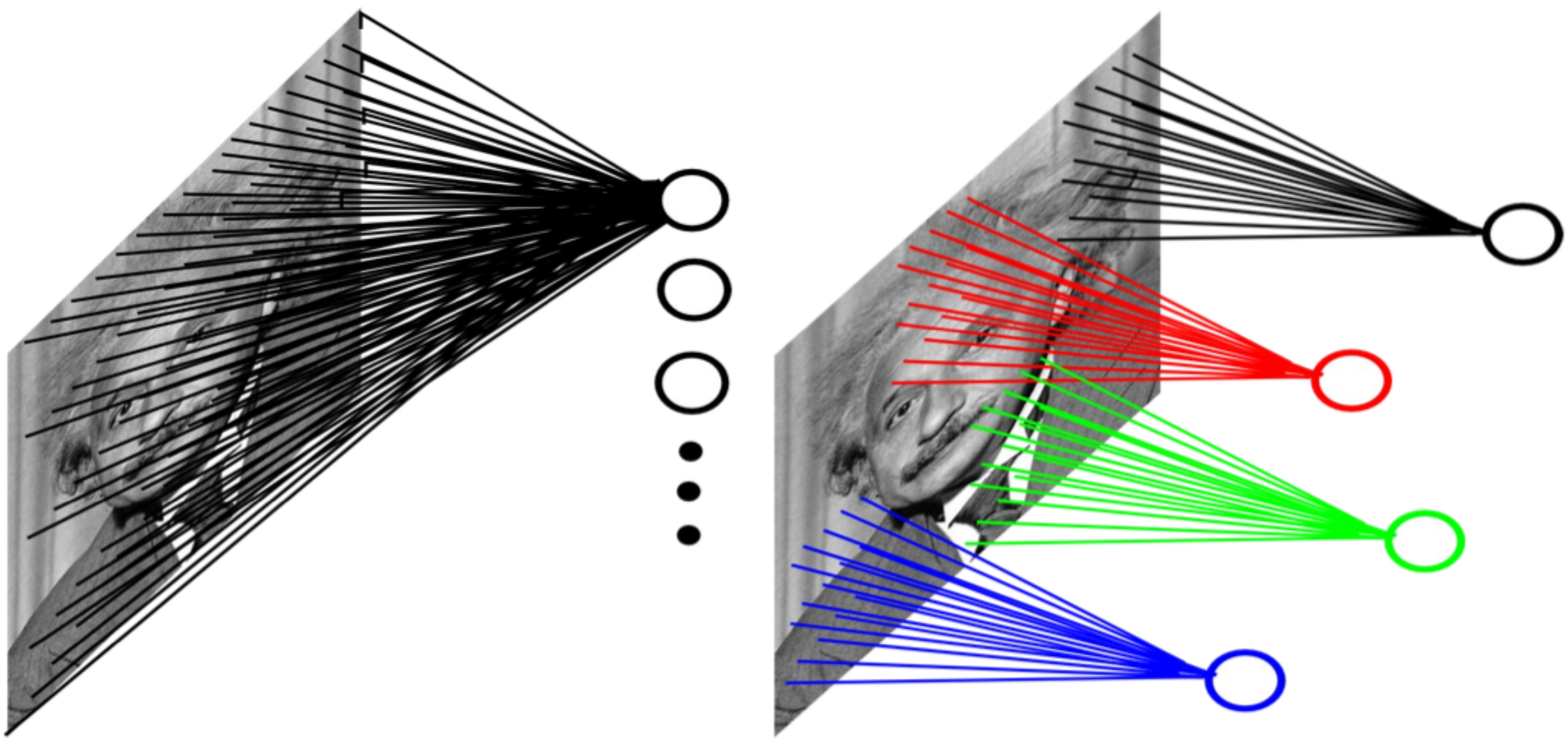


Similarly, to identify edges or other local structure, it makes sense to only look at **local information**



2d Convolution Layer

- **Example: 200x200 image**
  - Fully-connected, 400,000 hidden units = 16 billion parameters
  - Locally-connected, 400,000 hidden units 10x10 fields = 40 million params
  - Local connections capture local dependencies



Convolution of images (2d convolution)

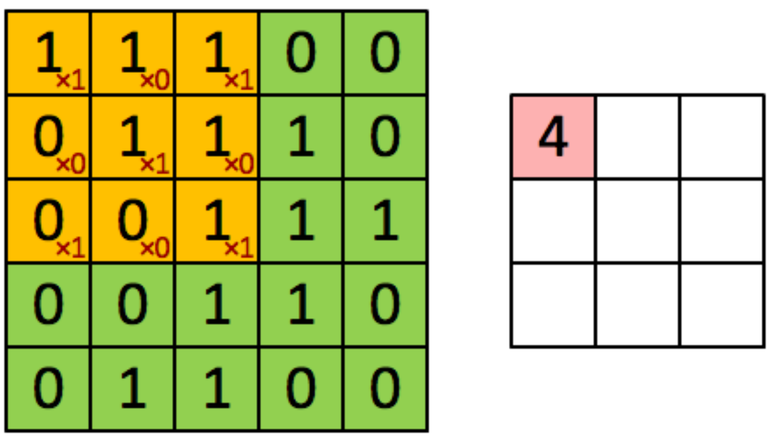
$$(I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n)$$

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Image I

1	0	1
0	1	0
1	0	1

Filter K

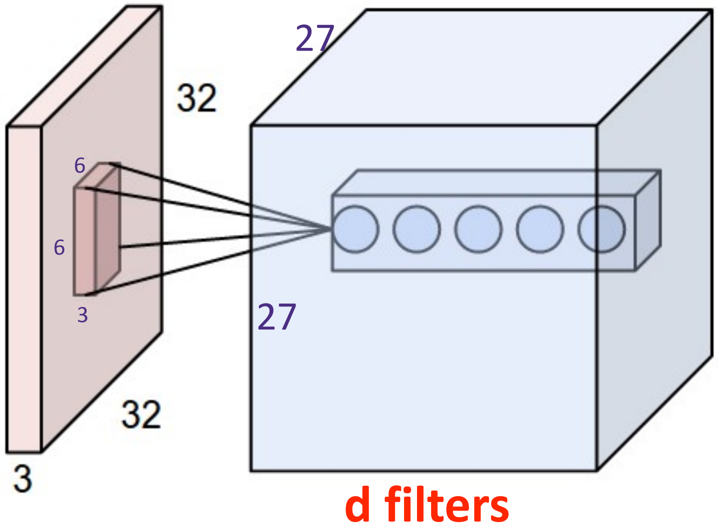


Image

Convolved Feature

$I * K$

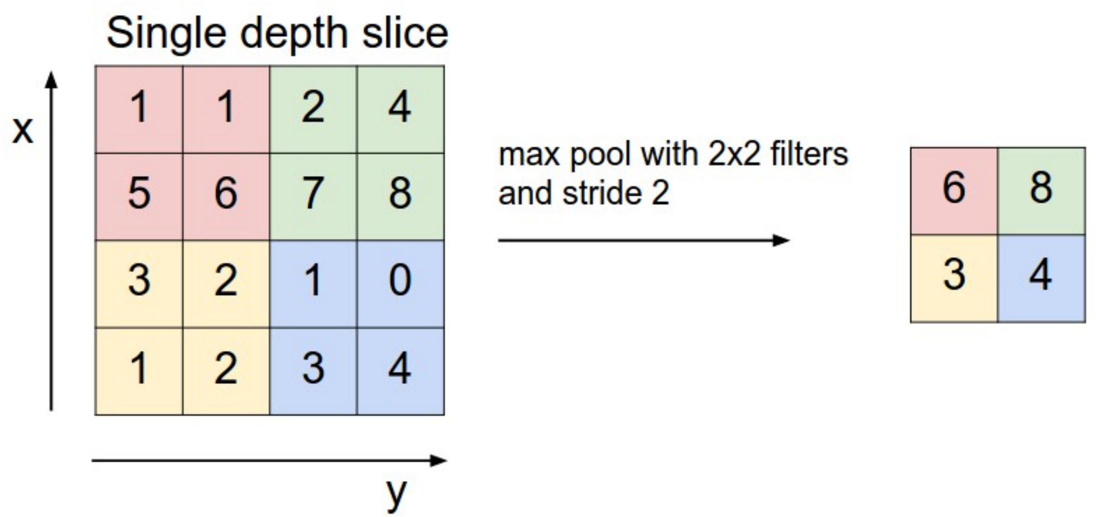
Stacking convolved images

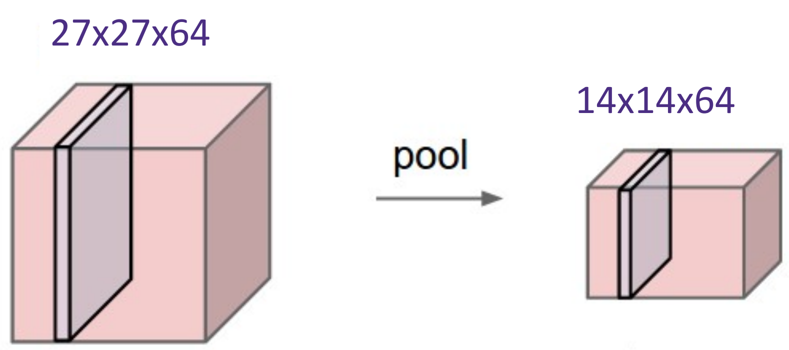


Repeat with d filters!

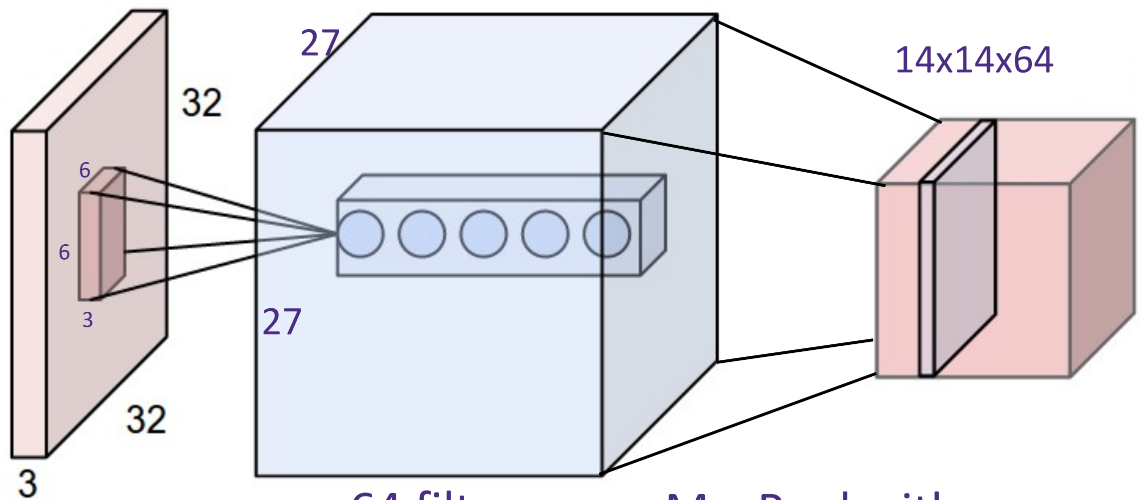
Pooling

Pooling reduces the dimension and can be interpreted as “This filter had a high response in this general region”





Pooling Convolution layer

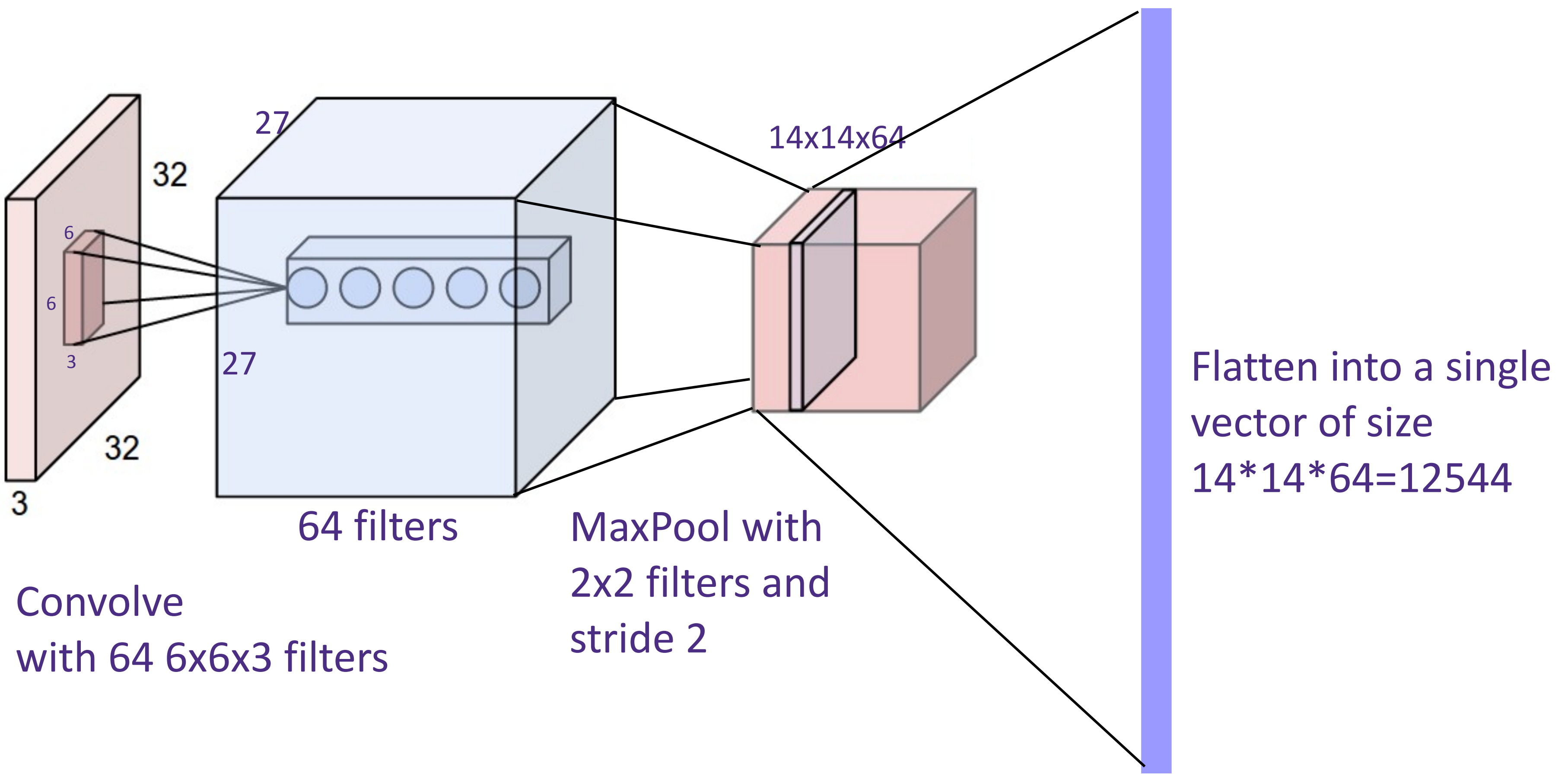


64 filters

MaxPool with
2x2 filters and
stride 2

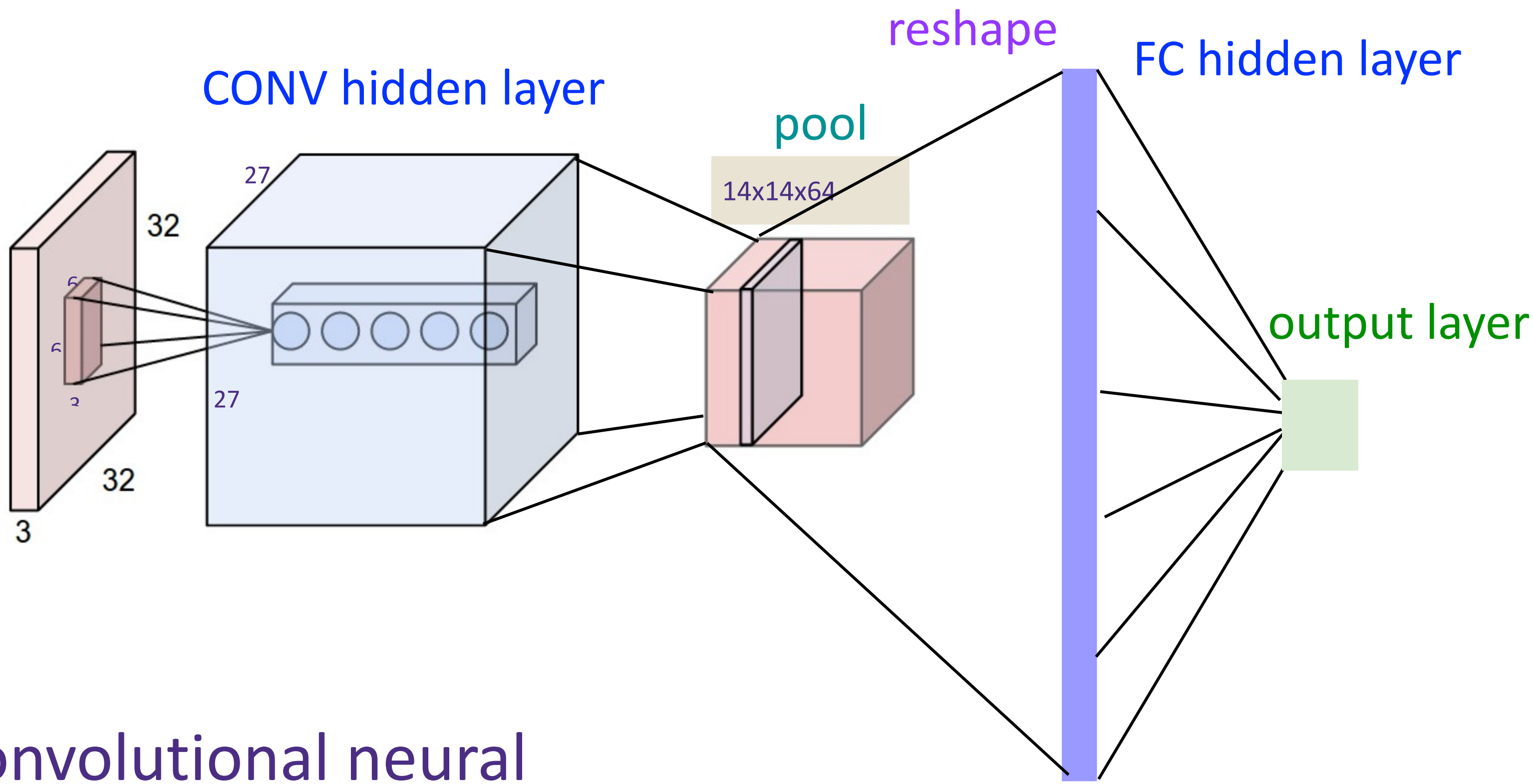
Convolve
with 64 6x6x3 filters

Flattening



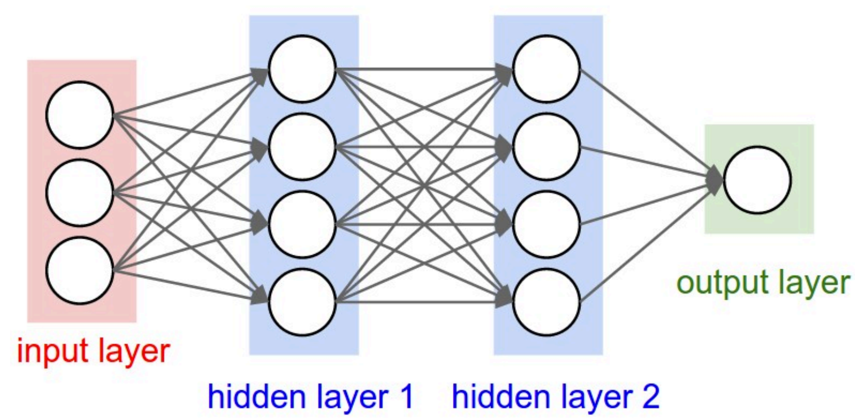
27
32
6
6
3
27
32
3
14x14x64
64 filters
MaxPool with
2x2 filters and
stride 2
Convolve
with 64 6x6x3 filters
Flatten into a single
vector of size
14*14*64=12544

Training Convolutional Networks

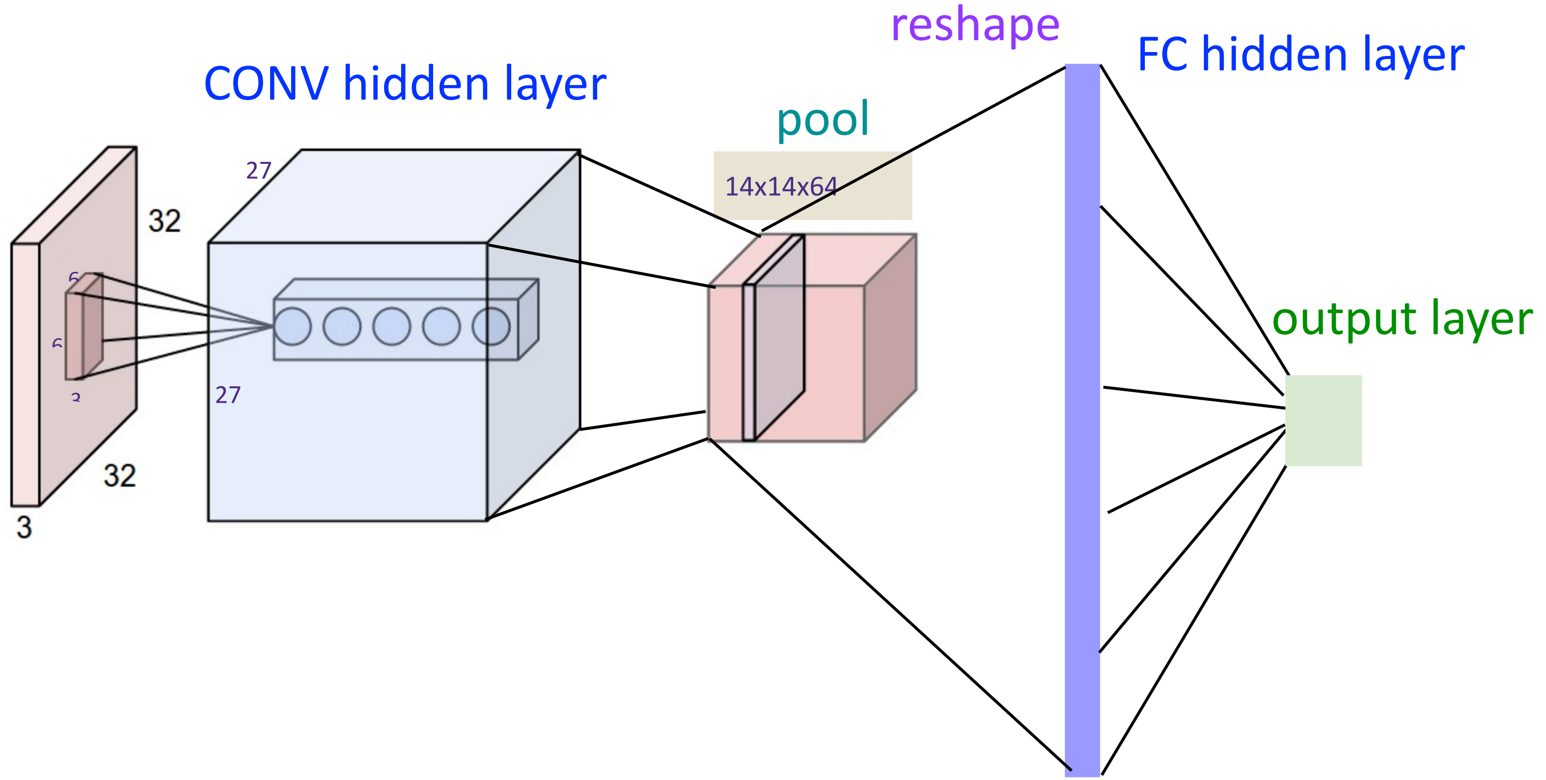


Recall: Convolutional neural networks (CNN) are just regular fully connected (FC) neural networks with some connections removed.

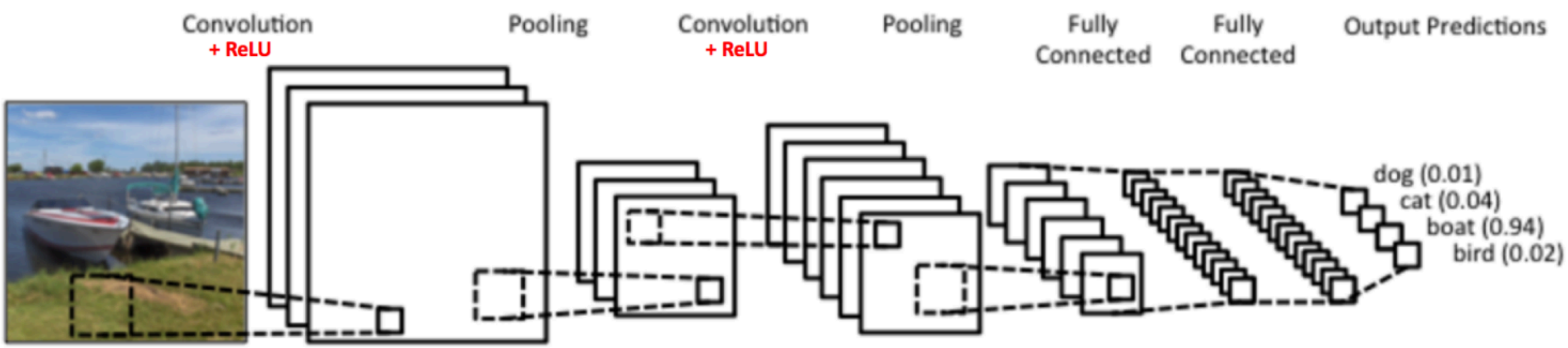
Train with SGD!

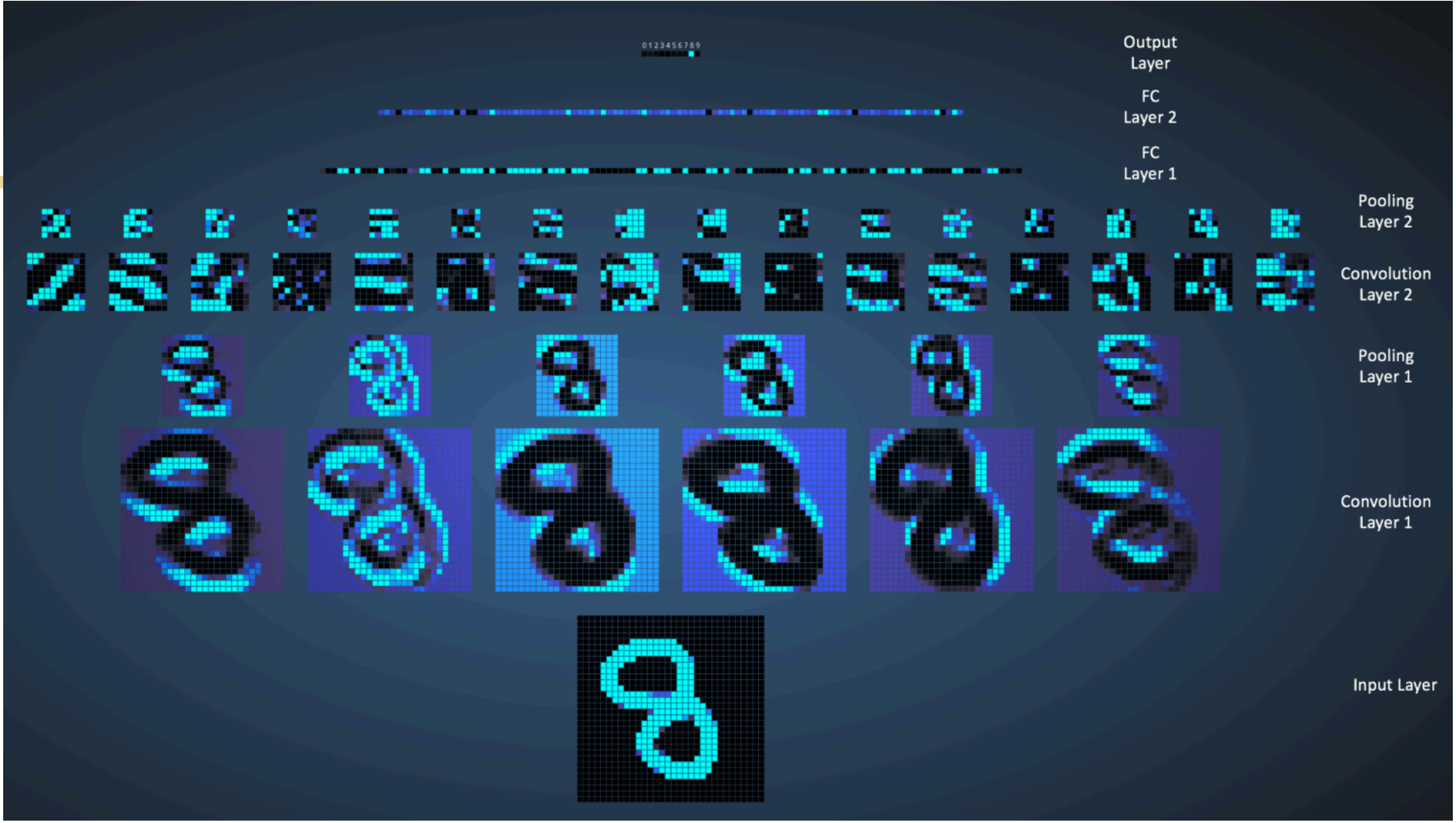


Training Convolutional Networks

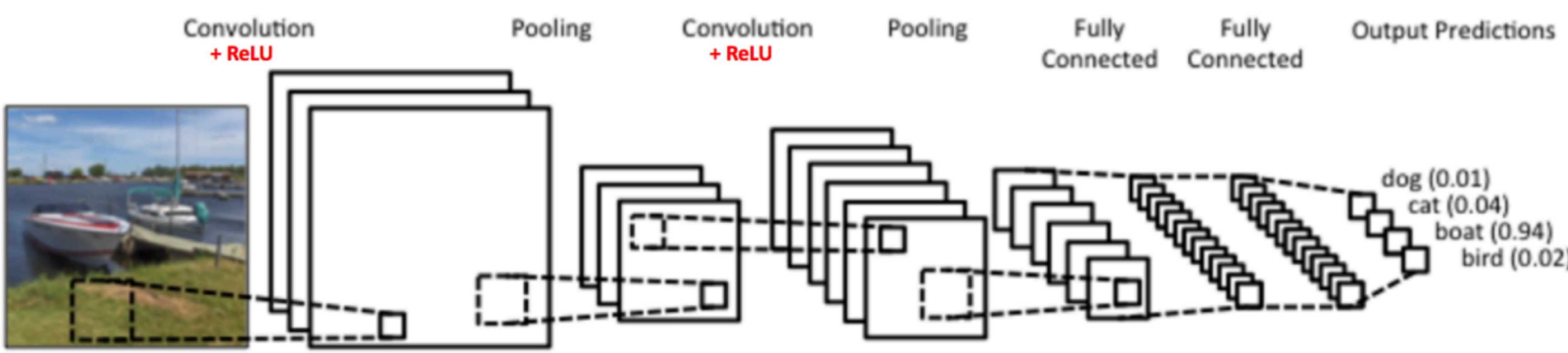


Real example network: LeNet





Real example network: LeNet



Famous CNNs

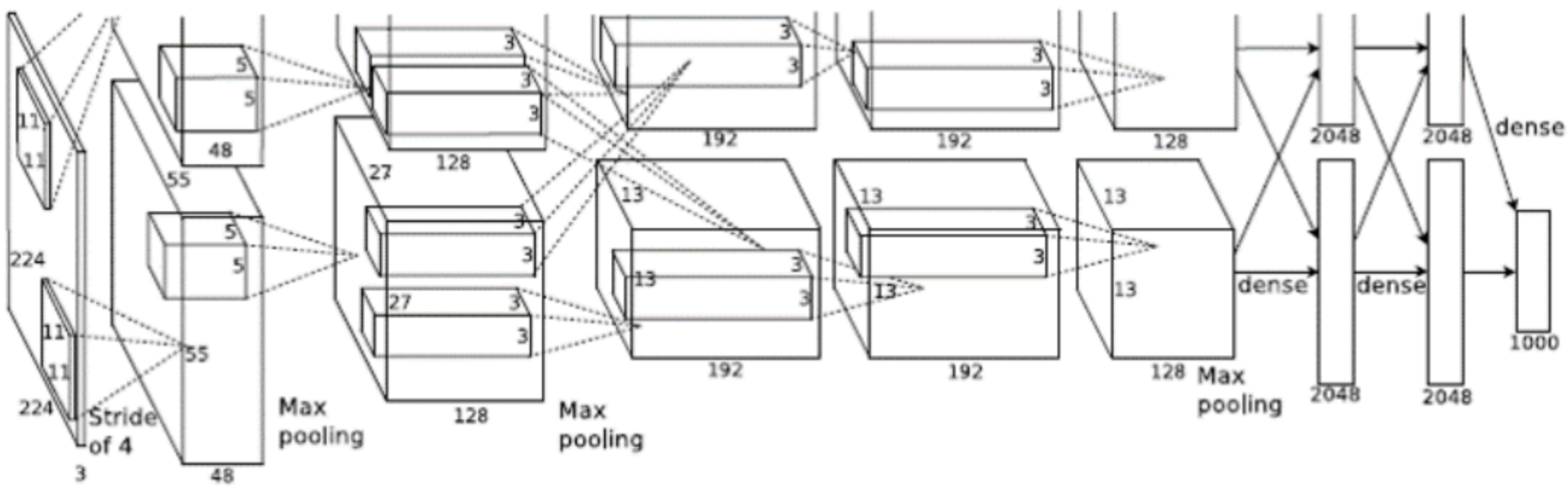
ImageNet Dataset

~14 million images, 20k classes

Deng et al. “Imagenet: a large scale hierarchical image database” ‘09

AlexNet

Breakthrough on ImageNet: ~the beginning of deep learning era



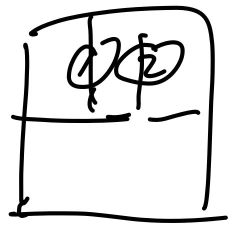
Krizhevsky, Sutskever, Hinton "ImageNet Claasification with Deep Convolutional Neural Networks", NIPS 2012. TPH of Award

AlexNet

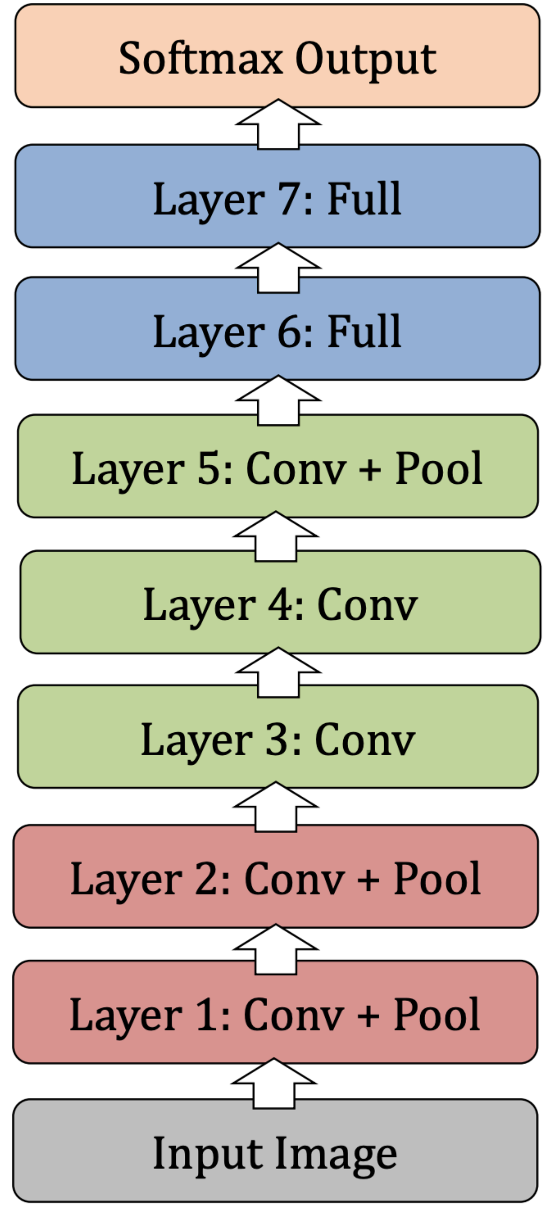
over-params

8 layers, ~60M parameters

Top5 error: 18.2%



Techniques used:
ReLU activation, overlapping pooling, dropout, ensemble (create 10 patches by cropping and average the predictions), data-augmentation (intensity of RGB channels)



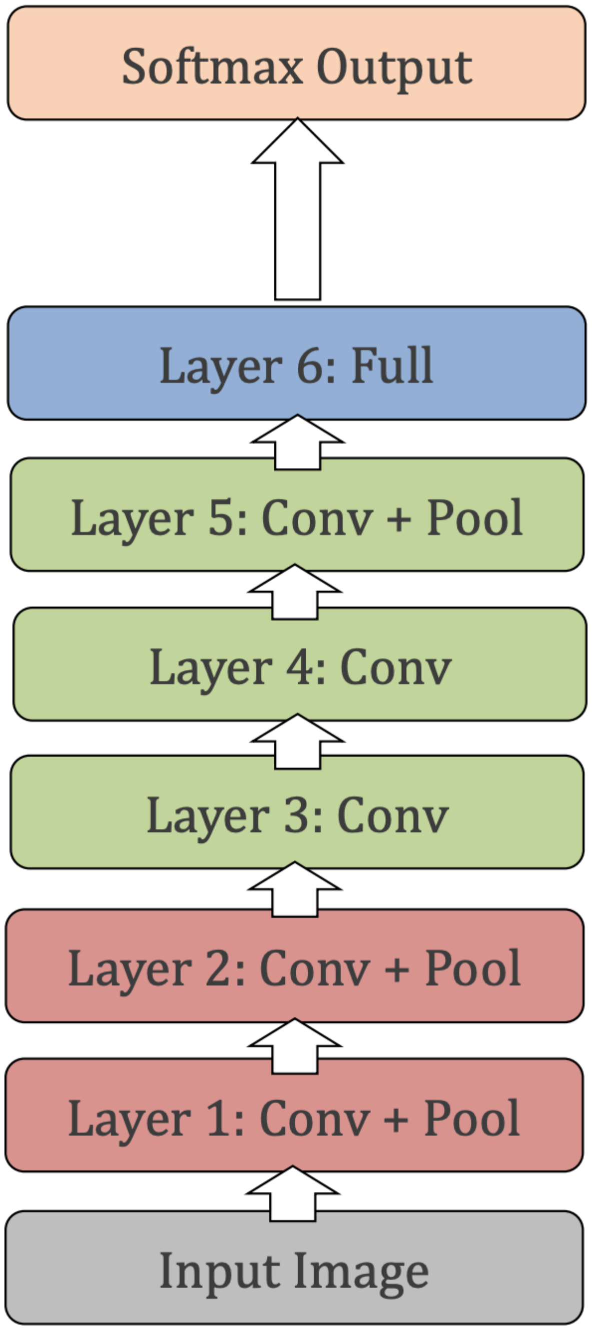
[From Rob Fergus' CIFAR 2016 tutorial]

AlexNet

Remove top fully-connected layer 7

Drop ~16 million parameters

1.1% drop in performance



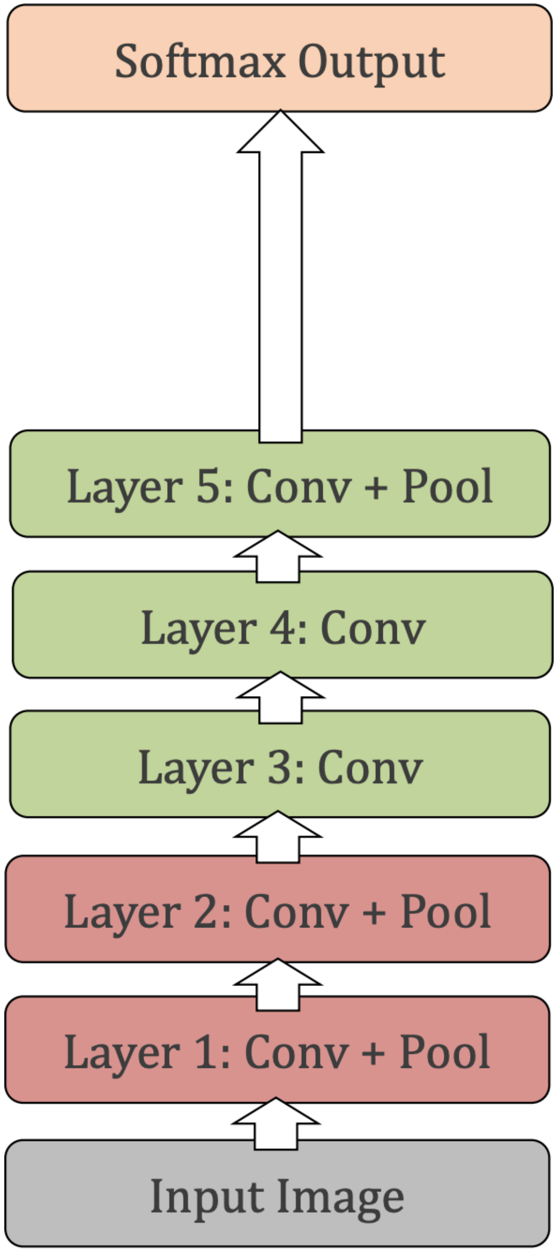
[From Rob Fergus' CIFAR 2016 tutorial]

AlexNet

Remove both fully connected layers 6 and 7

Drop ~50 million parameters

5.7% drop in performance



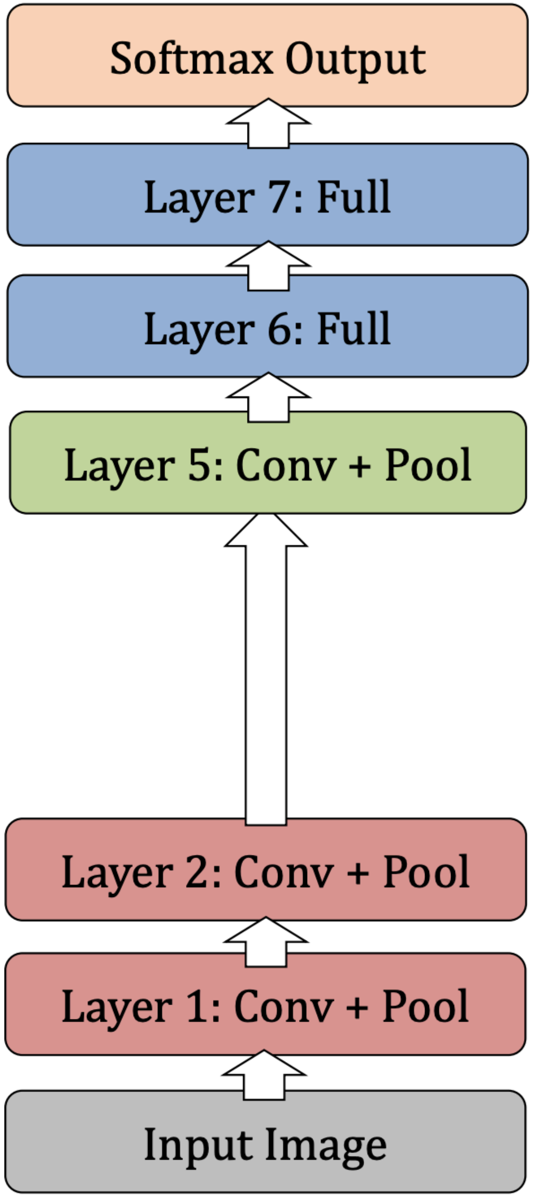
[From Rob Fergus' CIFAR 2016 tutorial]

AlexNet

Remove upper convolutio / feature extractor layers (layer 3 and 4)

Drop ~1 million parameters

3% drop in performance



[From Rob Fergus' CIFAR 2016 tutorial]

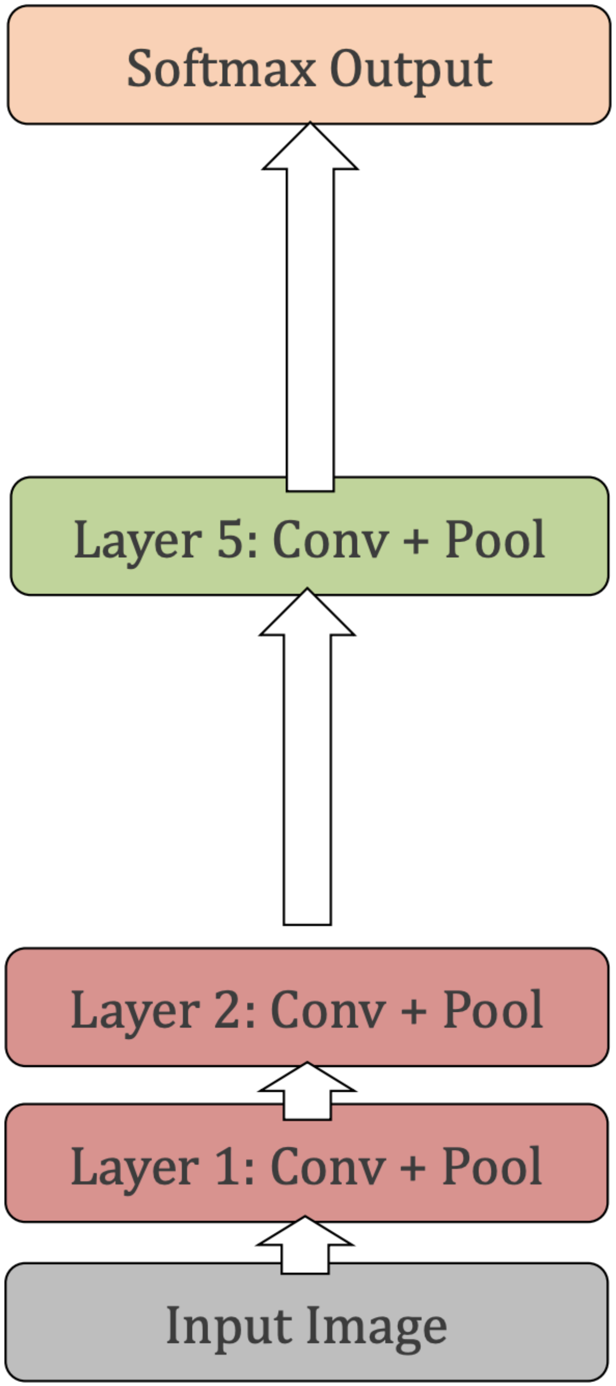
AlexNet

Remove top fully connected layer 6,7 and upper convolution layers 3,4.

33.5% drop in performance.

Depth of the network is the key.

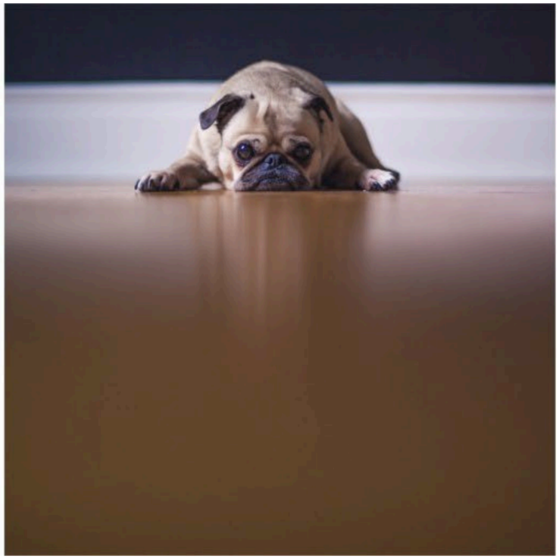
[From Rob Fergus' CIFAR 2016 tutorial]



GoogLeNet

Motivation: multiscale nature of images





7x7, 5x5

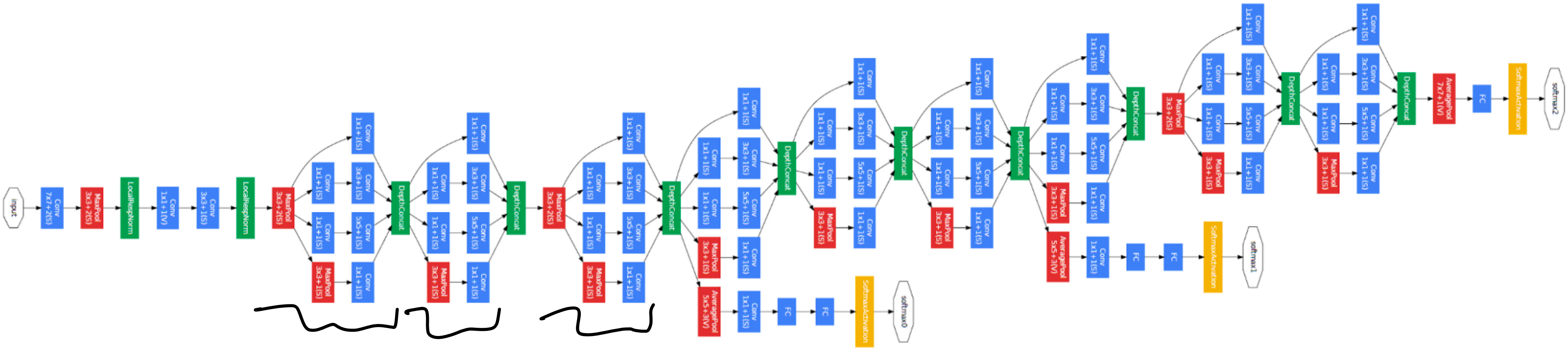
1x1, 3x3

Large kernel for global features, and **smaller kernel** for local features.

Idea: have multiple different-size kernels at any layer.

[Going Deep with Convolutions, Szegedy et al. '14]

GoogLeNet

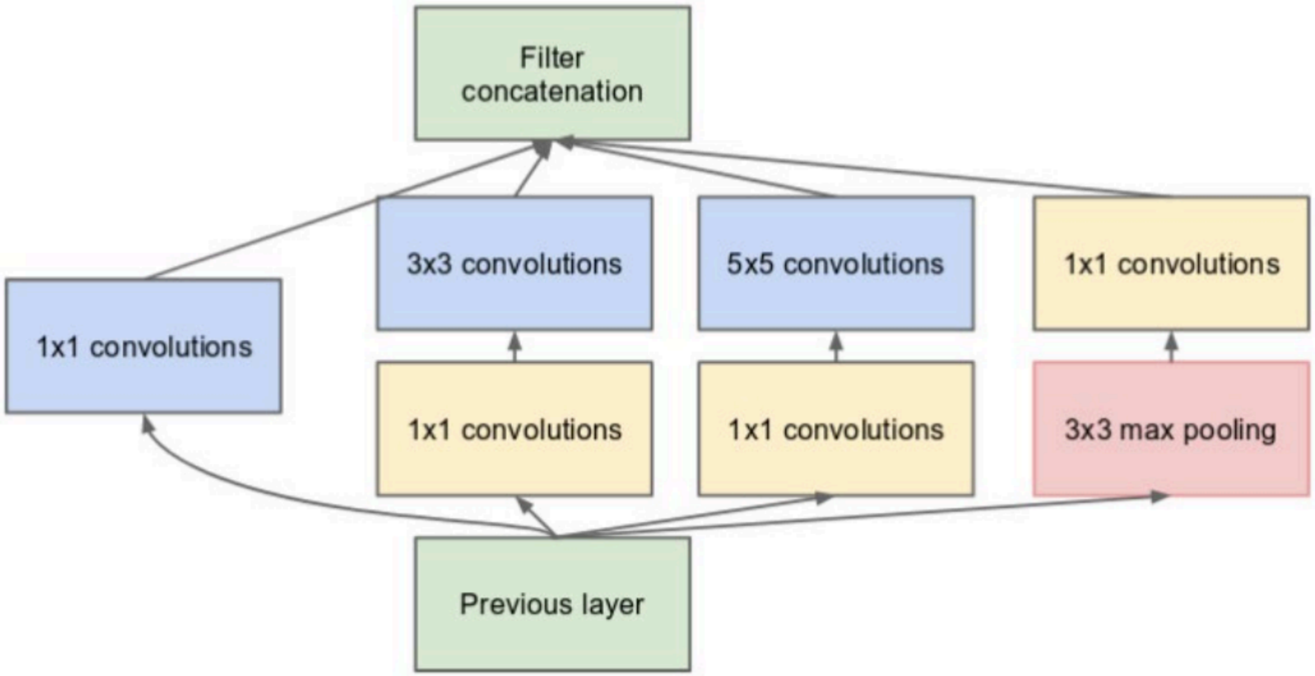


Large kernel for global features, and **smaller kernel** for local features.

Idea: have multiple different-size kernels at any layer.

[Going Deep with Convolutions, Szegedy et al. ’14]

Inception Module



Multiple filter scales at each layer

Dimensionality reduction to keep computational requirements down

[Going Deep with Convolutions, Szegedy et al. ’14]

Residual Networks

Motivation: extremely deep nets are hard to train (gradient explosion/ vanishing)

learning decay

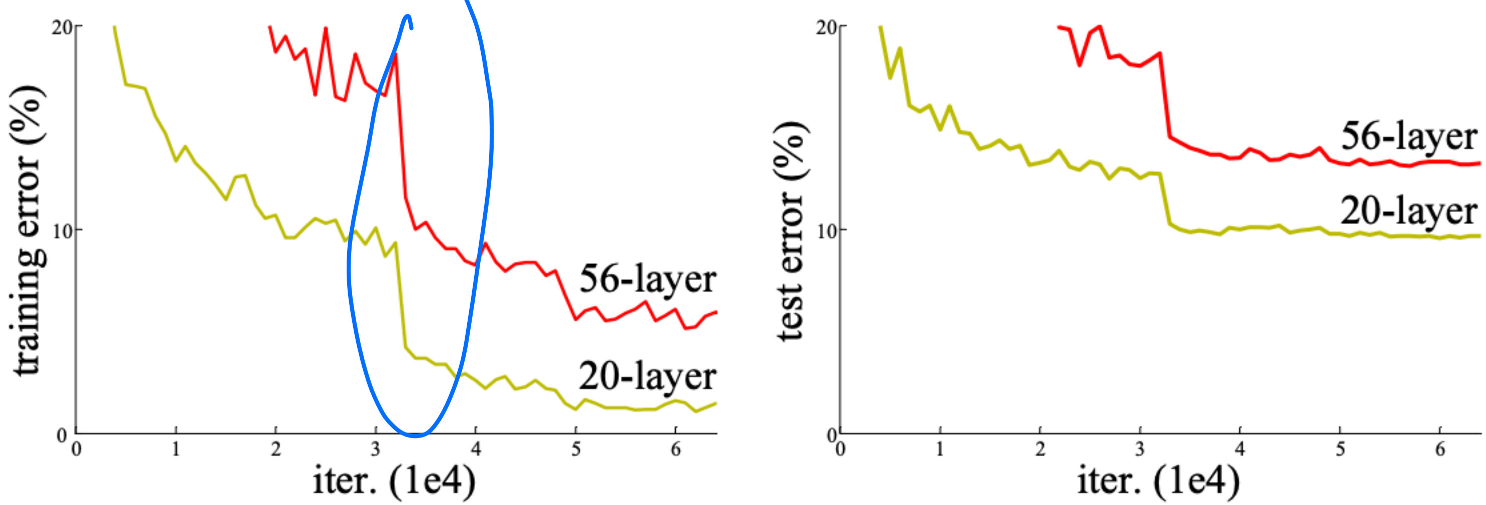


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

[He, Zhang, Ren, Sun, '16]

Residual Networks

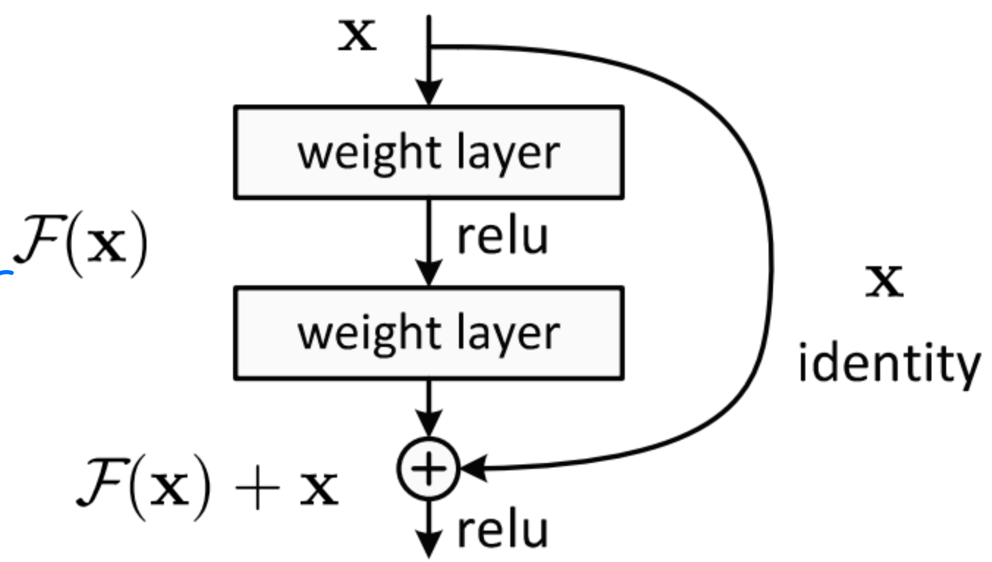
ReLU(ReLU(x)) ...

ReLU(x)

Idea: identity shortcut, skip one or more layers.

Justification: network can easily simulate shallow network ($F \approx 0$), so performance should not degrade by going deeper.

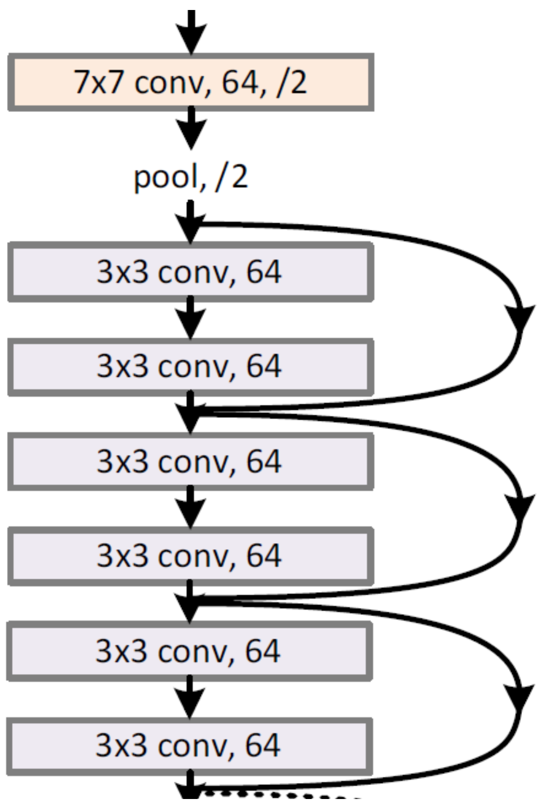
block

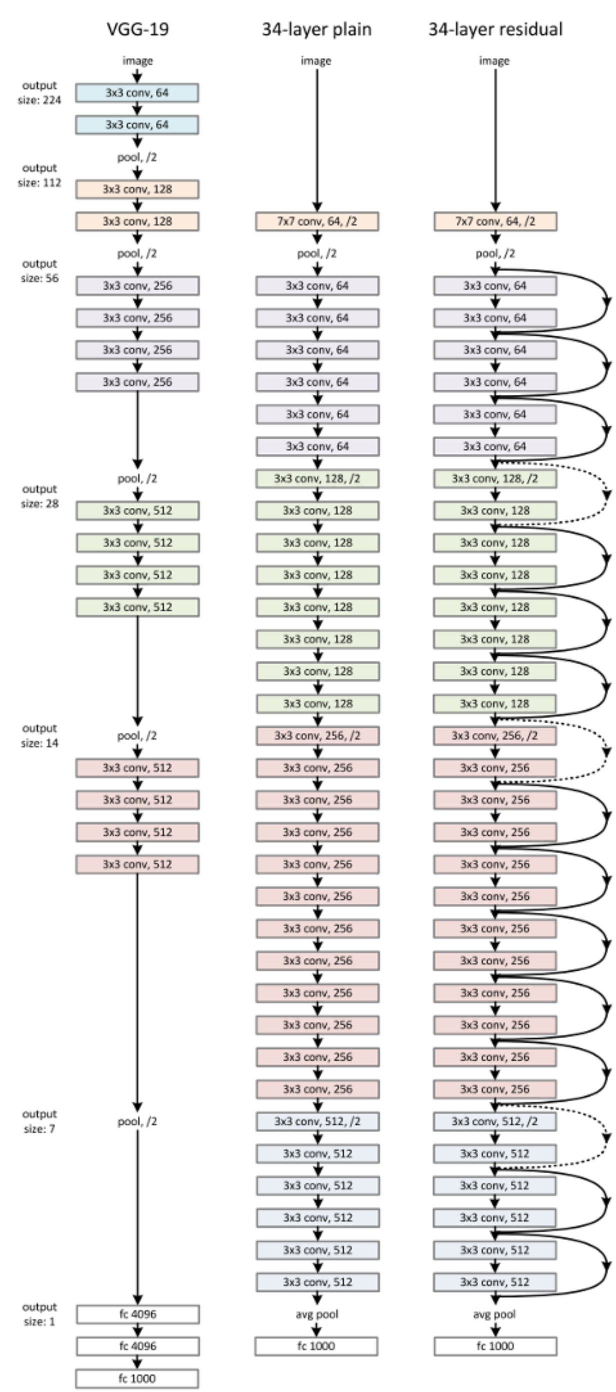


[He, Zhang, Ren, Sun, '16]

Residual Networks

- 3.57% top-5 error on ImageNet
- First deep network with > 100 layers.
- Widely used in many domains (AlphaGo)

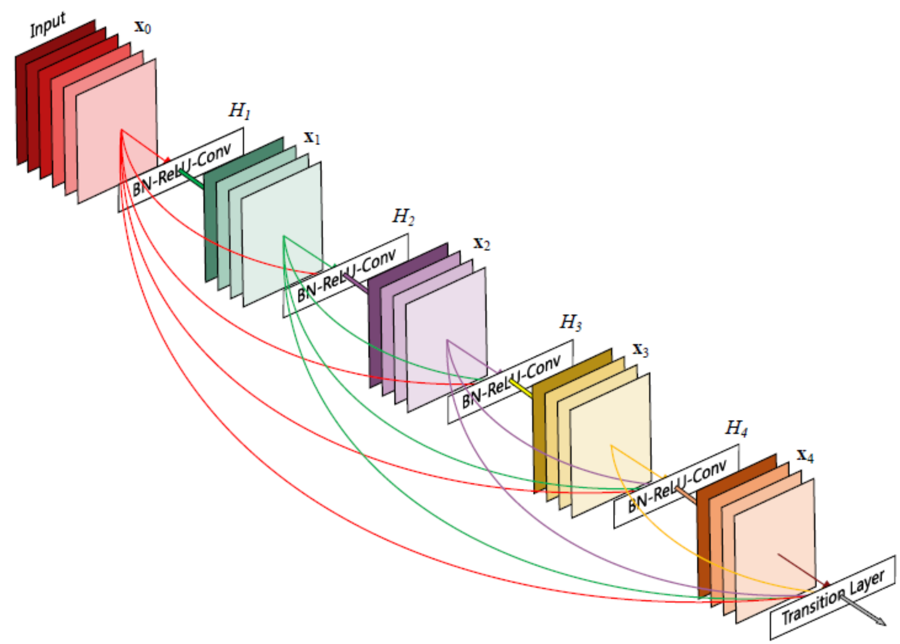




[He, Zhang, Ren, Sun, ’16]

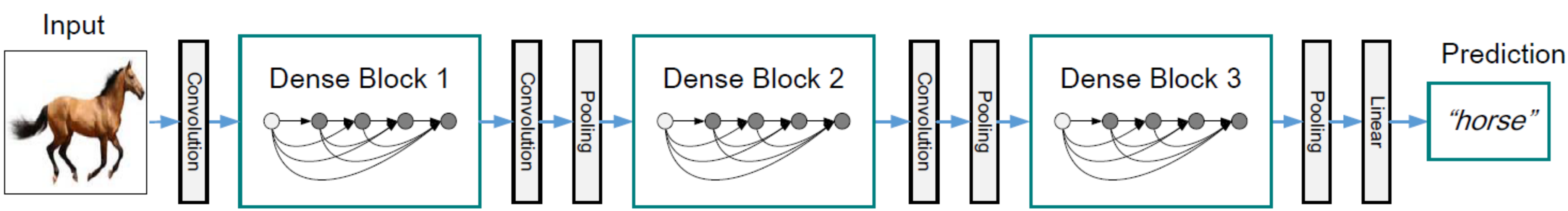
Densely Connected Network

Idea: explicit forward output of layer to all future layers (by concatenation)



Intuition: helps vanishing gradients, encourage reuse features (reduce parameter count)

Issues: network maybe too wide, need to be careful about memory consumption



~~[He, Zhang, Ren, Sun, '16]~~

Neural Architecture / Hyper-Parameter Search

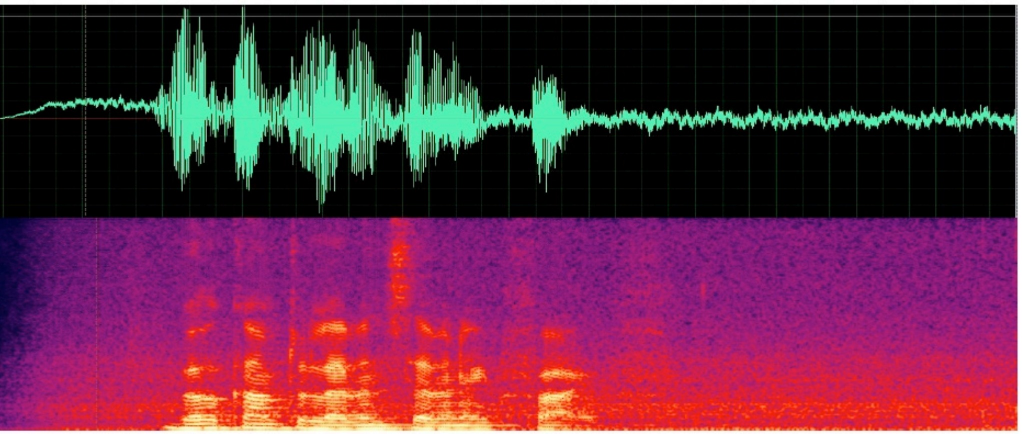
Many design choices:
- Number of layers, width, kernel size, pooling, connections, etc.
- Normalization, learning rate, batch size, etc.

Strategies:
- Grid search
- Random search [Bergestra & Bengio ’12]
- Bandit-based [Li et al. ’16]
- Gradient-based (DARTS) [Liu et al. ’19]
- Neural tangent kernel [Xu et al. ’21]
- …

Recurrent Neural Networks

Sequence Data

audio



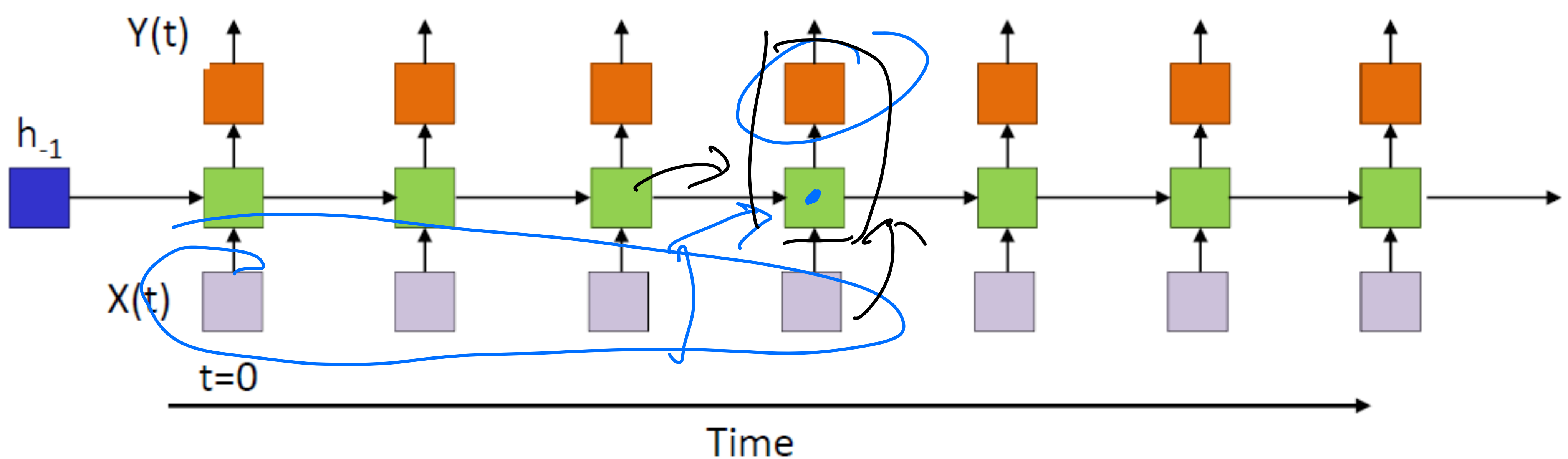
time series

HMM, POMDP

State-Space Model

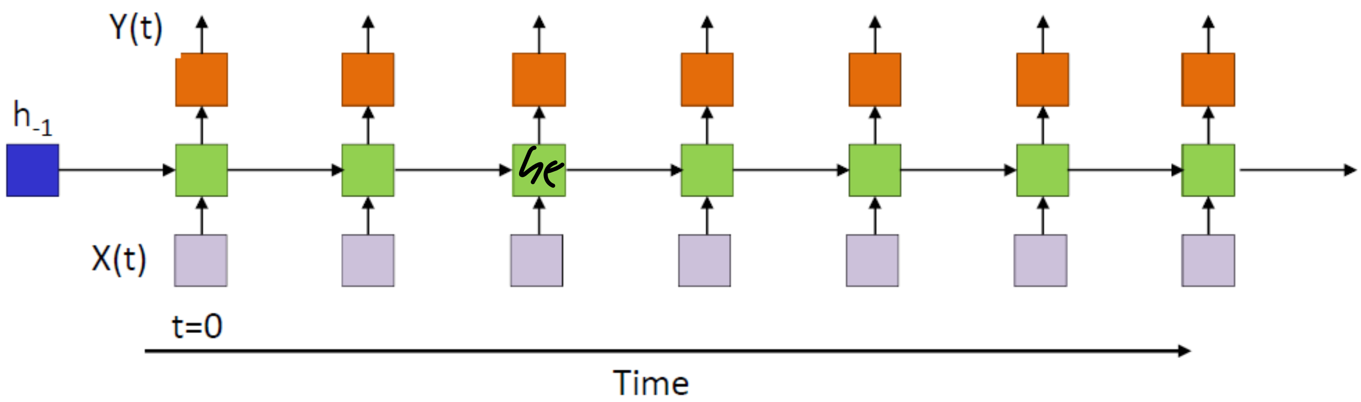
- h_t: hidden state
- X_t: input
- Y_t: output
- $Y_t, h_t = f(h_{t-1}, X_t; \theta)$
- h_{-1}: initial state

Y(t)



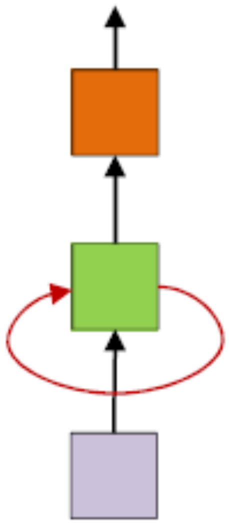
Recurrent Neural Network

- h_t: hidden state
- X_t: input
- Y_t: output
- $Y_t, h_t = f(h_{t-1}, X_t; \theta)$
- h_{-1}: initial state



Fully-connect NN vs. RNN

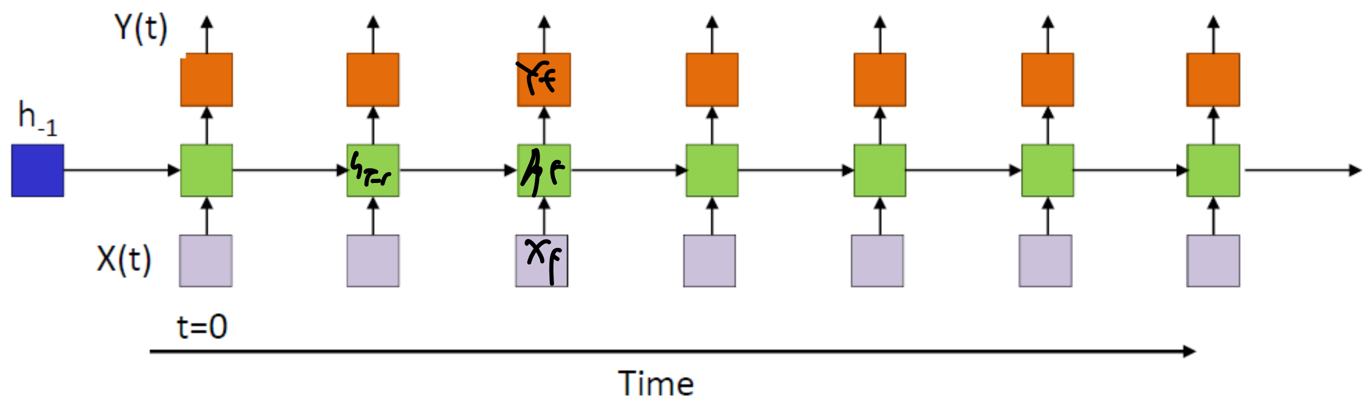
- h_t: a vector summarizes all past inputs (a.k.a. "memory")
- h_{-1} affects the entire dynamics (typically set to zero)
- X_t affects all the outputs and states after t
- Y_t depends on $X_0, \ldots, X_t$



FC: $(X_0, \ldots X_T) \to (Y_0, \ldots, Y_T)$

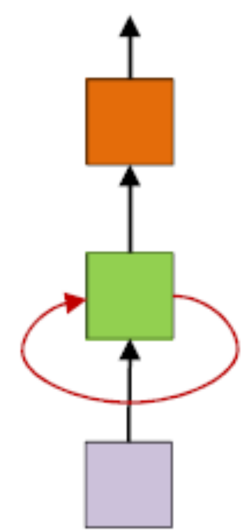
Recurrent Neural Network

- h_t: hidden state
- X_t: input
- Y_t: output
- $Y_t, h_t = f(h_{t-1}, X_t; \theta)$
- h_{-1}: initial state

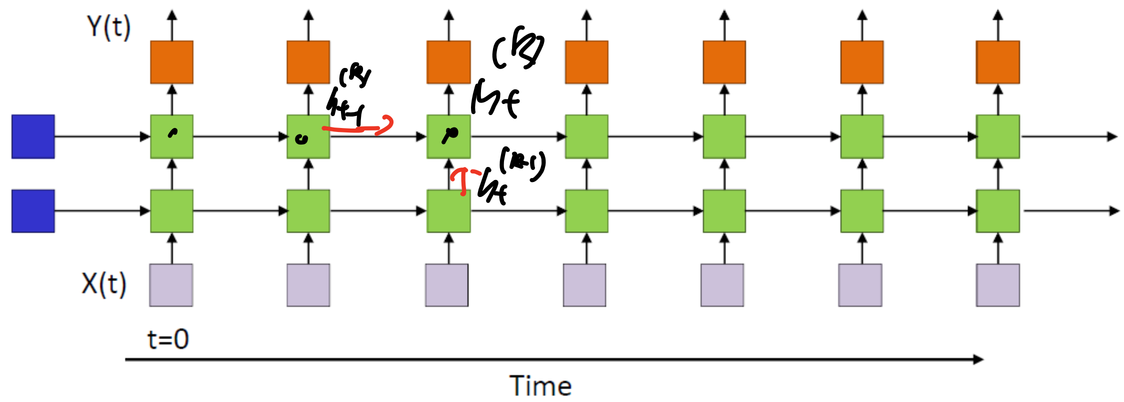


Fully-connect NN vs. RNN

- RNN can be viewed as repeated applying fully-connected NNs
- $h_t = \sigma_1(W^{(1)}X_t + W^{(11)}h_{t-1} + b^{(1)})$
- $Y_t = \sigma_2(W^{(2)}h_t + b^{(2)})$
- σ_1, σ_2 are activation functions (sigmoid, ReLU, tanh, etc)

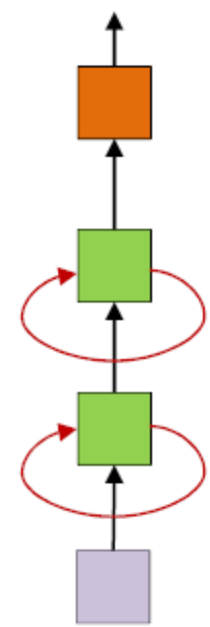


Recurrent Neural Network



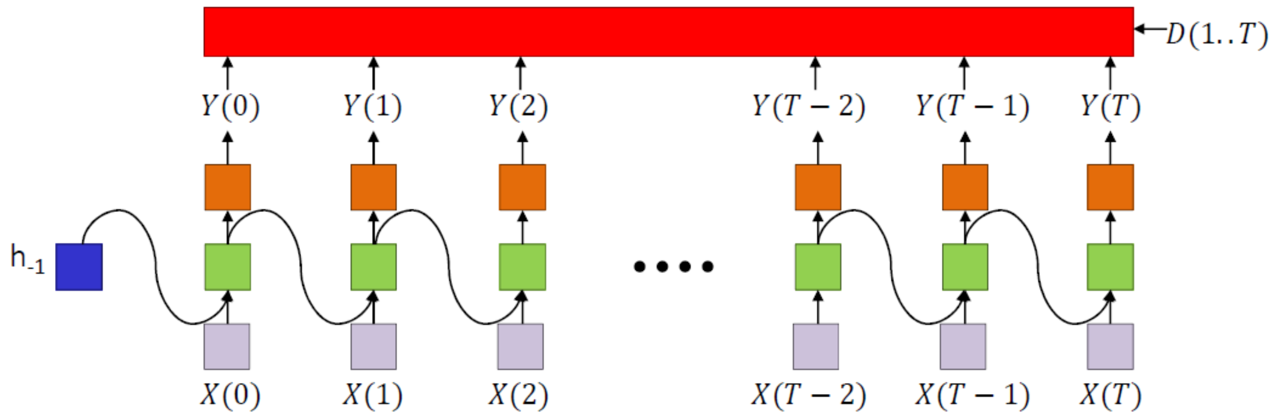
Stack K layers of fully-connected NN

- $h_t^{(k)}$: hidden state
- X_t: input
- Y_t: output
- $h_t^{(1)} = f_1^{(1)}(h_{t-1}^{(1)}, X_t; \theta)$
- $h_t^{(k)} = f_1^{(k)}(h_{t-1}^{(k)}, h_t^{(k-1)}; \theta)$
- $Y_t = f_2(h_t^{(K)}; \theta)$
- $h_{-1}^{(k)}$: initial states



Training Recurrent Neural Network

- h_t: hidden state
- X_t: input
- Y_t: output
- $Y_t, h_t = f(h_{t-1}, X_t; \theta)$
- h_{-1}: initial state



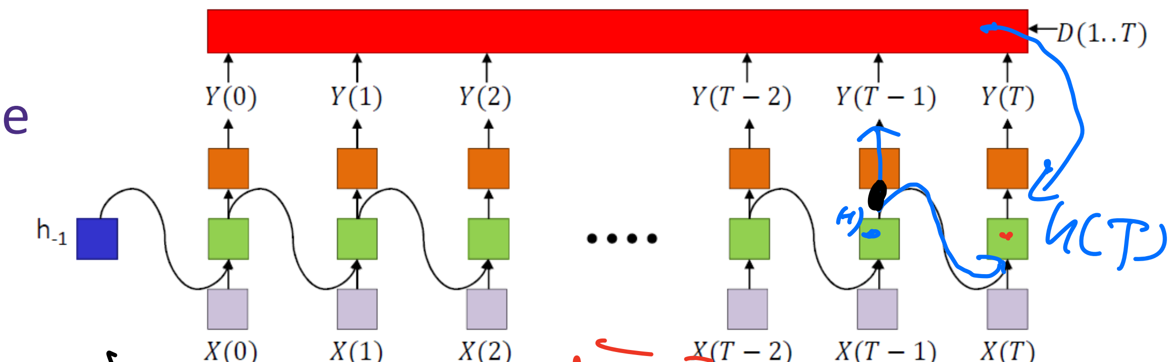
$V_t = Y_t$

- Data: $\{(X_t, D_t)\}_{t=1}^{T}$ (RNN can handle more general data format)
- Loss $L(\theta) = \sum_{t=1}^{T} \ell(Y_t, D_t)$
- Goal: learn θ by gradient-based method
  - Back propagation

Back Propagation Through Time

$Z_j^{(1)}$

- $h_t = \sigma_1(W^{(1)}X_t + W^{(11)}h_{t-1} + b^{(1)})$
- $Y_t = \sigma_2(W^{(2)}h_t + b^{(2)})$
- $Z_t^{(1)}$: pre-activation of hidden state ($h_t = \sigma_1(Z_t^{(1)})$)
- $Z_t^{(2)}$: pre-activation of output ($Y_t = \sigma_2(Z_t^{(2)})$)



$t = T$

(1) Start from $\frac{dL}{dY(T)}$

(2) $\frac{dL}{dZ_j^{(2)}(T)} = \frac{dL}{dY_j(T)} \cdot \frac{\partial Y_j(T)}{\partial Z_j^{(2)}(T)}$

(3) $\frac{dL}{dW_{ij}^{(2)}} \mathrel{+}= \frac{dL}{dZ_j^{(2)}(T)} \cdot h_i(T)$

d init $= 0$

$$\frac{dL}{dW_{ij}^{(2)}} = \sum_{t=0}^{T} \frac{dL}{dZ_j^{(2)}(t)} \cdot h_i(t)$$

$$\frac{dL}{dW_{ij}^{(1)}} = \sum_{t=0}^{T} \frac{dL}{dZ_j^{(1)}(t)} \cdot X_i(t)$$

$$\frac{dL}{dW_{ij}^{(11)}} = \sum_{t=0}^{T} \frac{dL}{dZ_j^{(1)}(t)} \cdot h_i(t-1)$$

Back Propagation Through Time

(4) $\frac{dL}{dh_i(T)} = \sum_j \frac{dL}{dz_j^{(2)}(T)} \cdot \frac{\partial z_j^{(2)}(T)}{\partial h_i(T)} = \sum_j W_{ij}^{(2)} \cdot \frac{dL}{dz_j^{(2)}(T)}$

(5) $\frac{dL}{dz_j^{(1)}(T)} = \frac{dL}{dh_j(T)} \cdot \frac{\partial h_j(T)}{\partial z_j^{(1)}(T)} = \frac{dL}{dh_j(T)} \cdot \sigma_1'(z_j^{(1)}(T))$

(6) $\frac{dL}{dw_{ij}^{(1)}} \mathrel{+}= \frac{dL}{dz_j^{(1)}(T)} \cdot X_i(T)$, $\frac{dL}{dw_{ij}^{(1)}} \mathrel{+}= \frac{dL}{dz_j^{(1)}(T)} \cdot h_i(T)$

Step T-1

(1) $\frac{dL}{dY(T-1)}$ / $\frac{dL}{dz_j^{(2)}(T-1)} = \frac{dL}{dY_j(T-1)} \cdot \frac{\partial Y_j(T-1)}{\partial z_j^{(2)}(T-1)}$

(2) $\frac{dL}{dW_{ij}^{(2)}} \mathrel{+}= \frac{dL}{dY_j(T-1)} \cdot h_j(T-1)$

Back Propagation Through Time

(3) $$\frac{dL}{dh_i(T-1)} = \sum_j W_{ij}^{(2)} \frac{dL}{dz_j^{(2)}(T-1)} + \underbrace{\sum_j W_{ij}^{(11)} \cdot \frac{dL}{dz_j^{(1)}(T)}}$$

(4) $$\frac{dL}{dz_j^{(1)}(T-1)} = \frac{dL}{dh_j(T-1)} \cdot \frac{\partial h_j(T-1)}{\partial z_j^{(1)}(T-1)}$$

(5) $$\frac{dL}{dw_{ij}^{(1)}} \mathrel{+}= \frac{dL}{dz_j^{(1)}(T-1)} \cdot x_i(T-1)$$

$$\frac{dL}{dw_{ij}^{(11)}} \mathrel{+}= \frac{dL}{dz_j^{(1)}(T-1)} \cdot h_i(T-2)$$

Repeat for $t = T-2, T-3, \ldots$

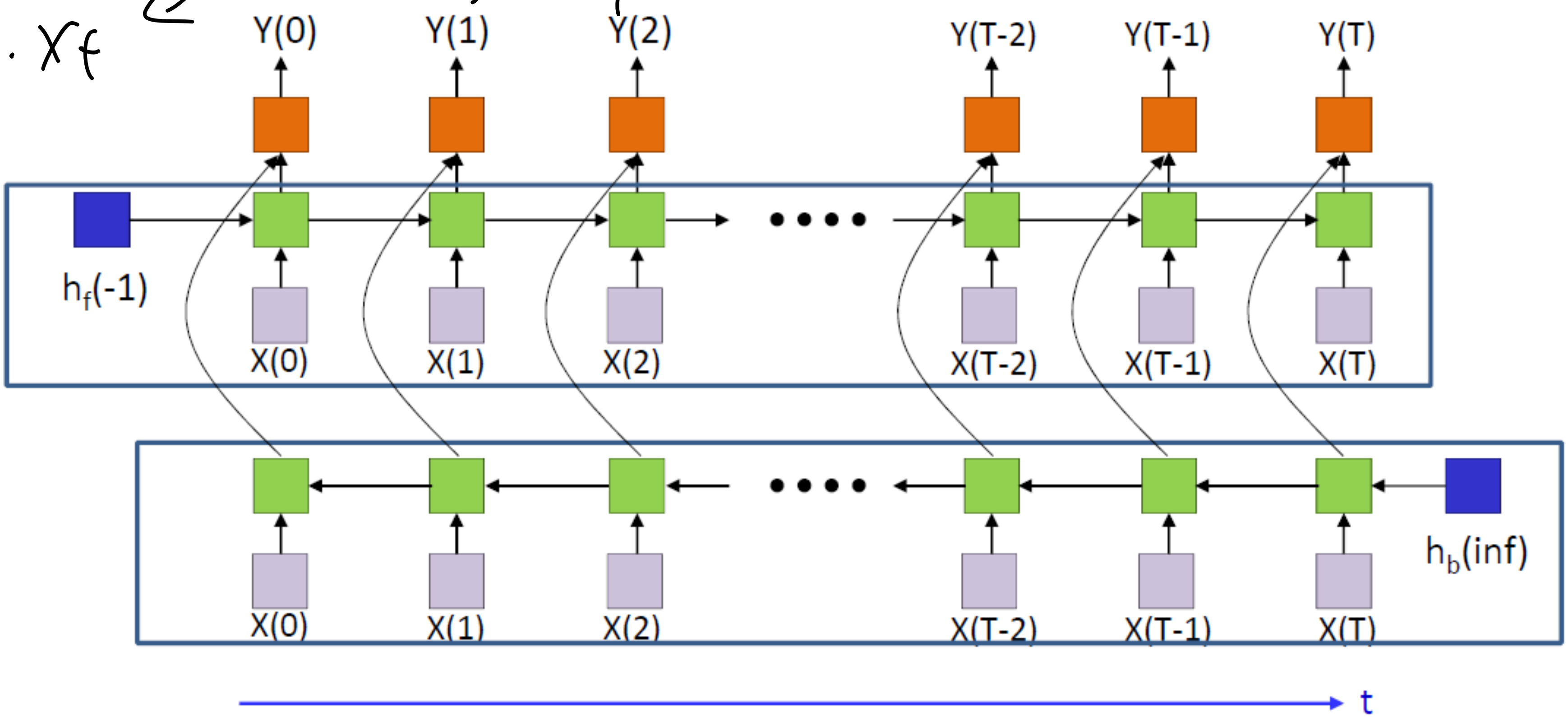
Extensions

What if Y_t depends on the entire inputs?

- Biredictional RNN:
    - AN RNN for forward dependencies: t= 0,…,T
    - An RNN for backward dependencies: t= T,…0
    - $Y_t = f_2(h_t^f, h_t^b; \theta)$

$X_t, \dots X_T$

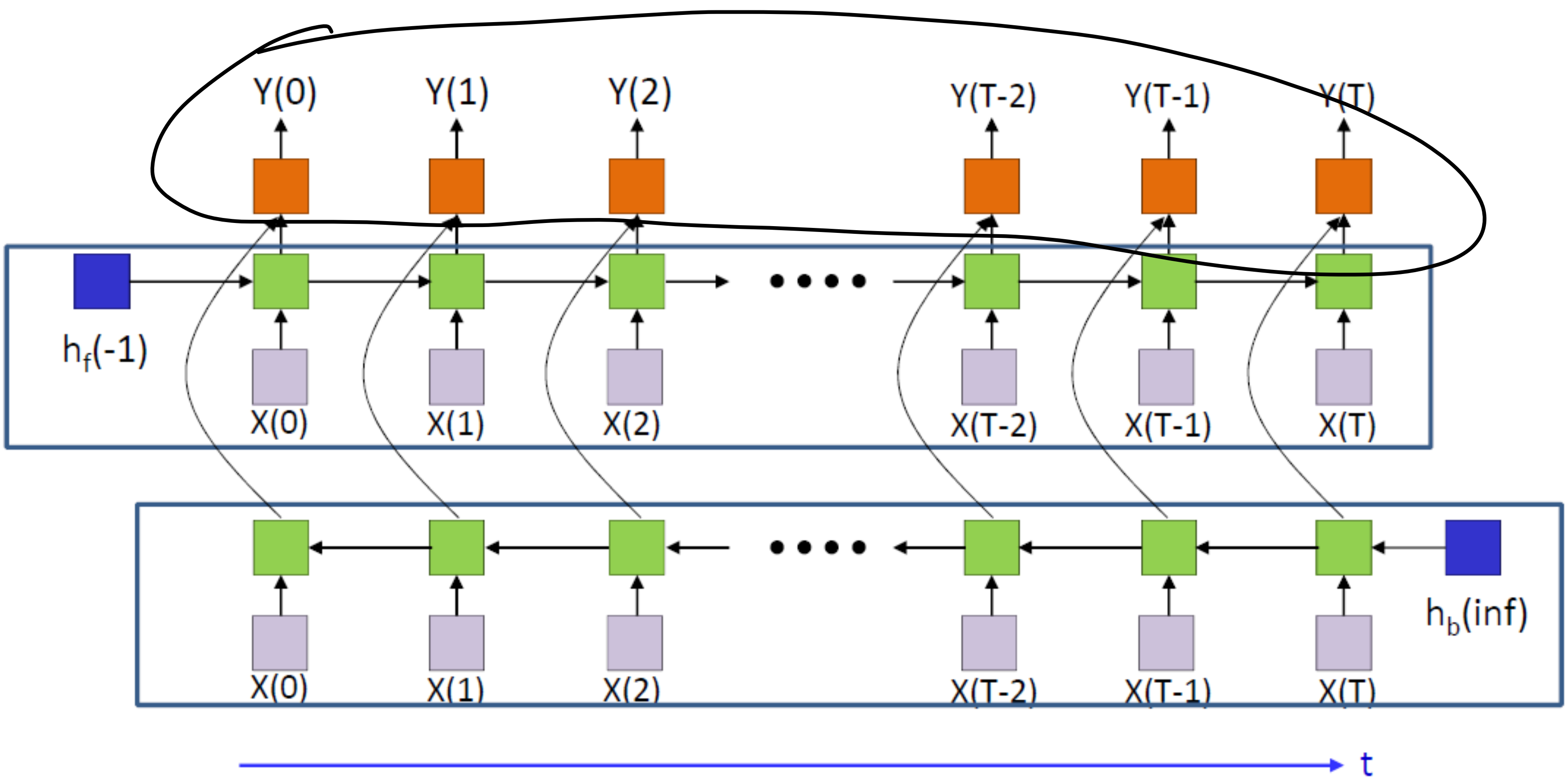
$X_0, \dots X_t$



Extensions

RNN for sequence classification (sentiment analysis)

- $Y = \max_t Y_t$
- Cross-entropy loss



Practical issues of RNN

Linear RNN derivation

$$\sigma(z) = z$$

- $h_t = W^{(11)} h_{t-1} + W^{(1)} \cdot x_t$
- $h_k = W^{(1)} \cdot x_k + W^{(11)} \cdot h_{k-1}$

$$= W^{(1)} \cdot x_k + W^{(11)} \left(W^{(1)} \cdot x_{k-1} + W^{(11)} \cdot h_{k-2} \right)$$

$$\cdots$$

$$= \left(W^{(11)} \right)^{k+1} \cdot h_{-1} + \sum_{j=0}^{k} \left(W^{(11)} \right)^{k-j} W^{(1)} \cdot x_j$$

exp()

if $\lambda_{max}(W^{(11)}) > 1 \rightarrow$ exp large

$< 1 \rightarrow$ exp small

cannot $\Rightarrow$ capture long-term info

Practical issues of RNN: training

Gradient explosion and gradient vanishing

- $z_t = W^{(11)} h_{t-1} + W^{(1)} x_t$
- $h_t = \sigma(z_t)$
- $L_k(\theta) = L(y_k, o_k)$
- $\frac{\partial L_k}{\partial h} \propto \underbrace{(W^{(11)})^k}_{\text{exp large or small}} \underbrace{\prod_{t=0}^{k} \sigma'(z_t)}_{<1 \text{ forgetting}}$

exp large or small

<1 forgetting

>1 → exp

Techniques for avoiding gradient explosion

- Gradient clipping

- Identity initialization

- Truncated backprop through time
  - Only backprop for a few steps