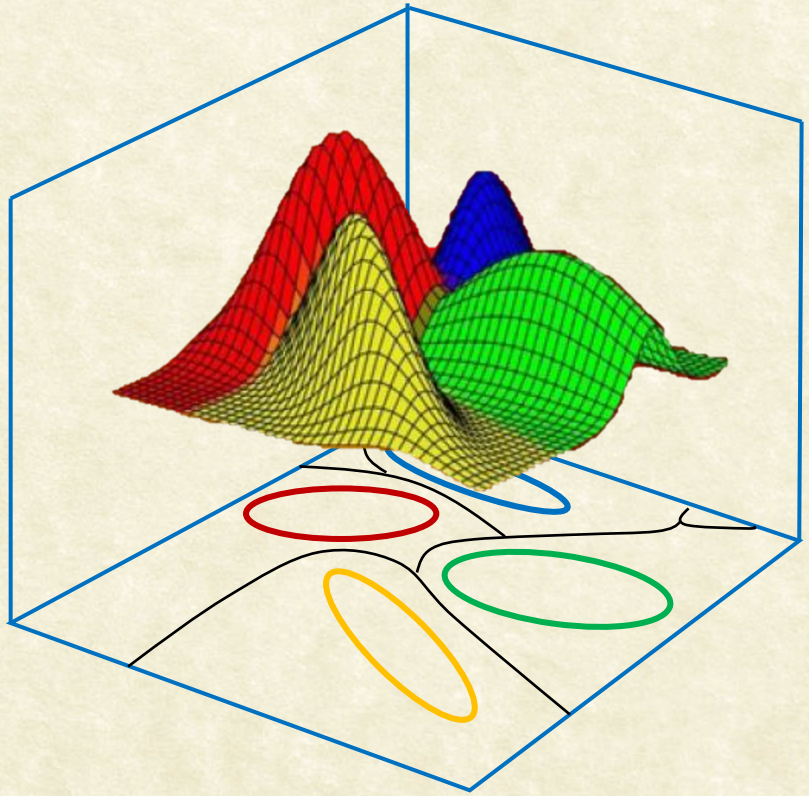
CS7.403: Statistical Methods in AI

Monsoon 2022:

Convolutional Neural Network



Anoop M. Namboodiri

Biometrics and Secure ID Lab, CVIT,

IIIT Hyderabad

Convolutional NN

The Structure

Terminologies

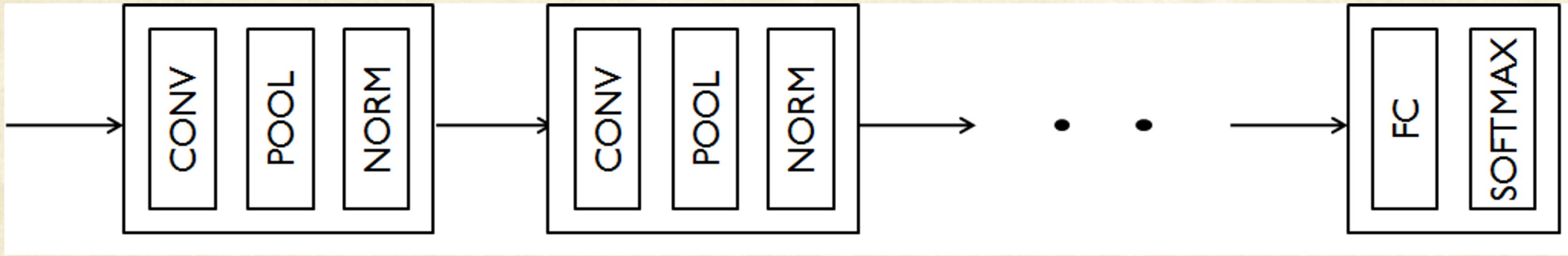
- # Input Channels
- # Output channels

- Feature Maps/Channels
- Filters/Weights
- Filter Size/Window Size

- Stride
- Pooling (Max/Average)
- Fully Connected Layer
- Soft-Max
- Normalization
- Flattening
- Convolution Layer

Typical Architecture

- A typical deep convolutional network

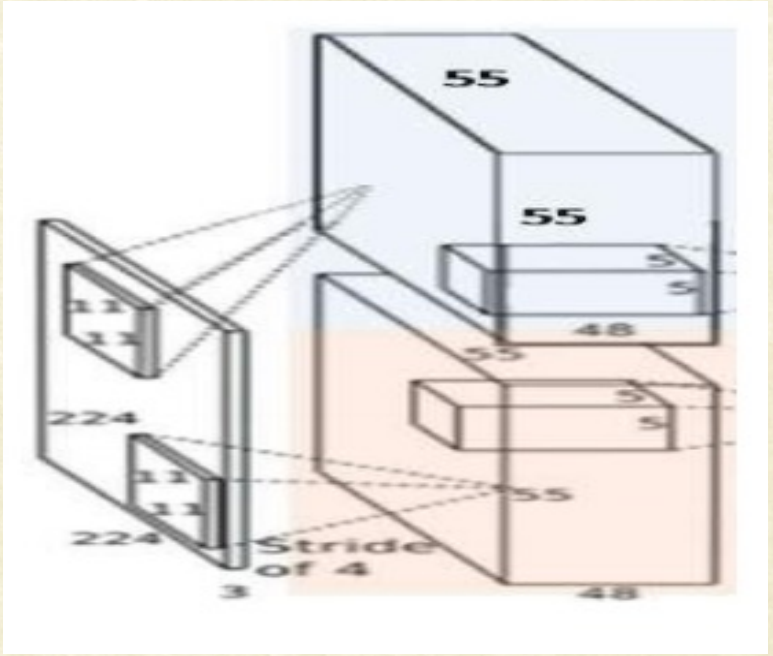


- Other layers
  - Pooling
  - Normalization
  - Fully connected
  - etc.

Parameter Calculation

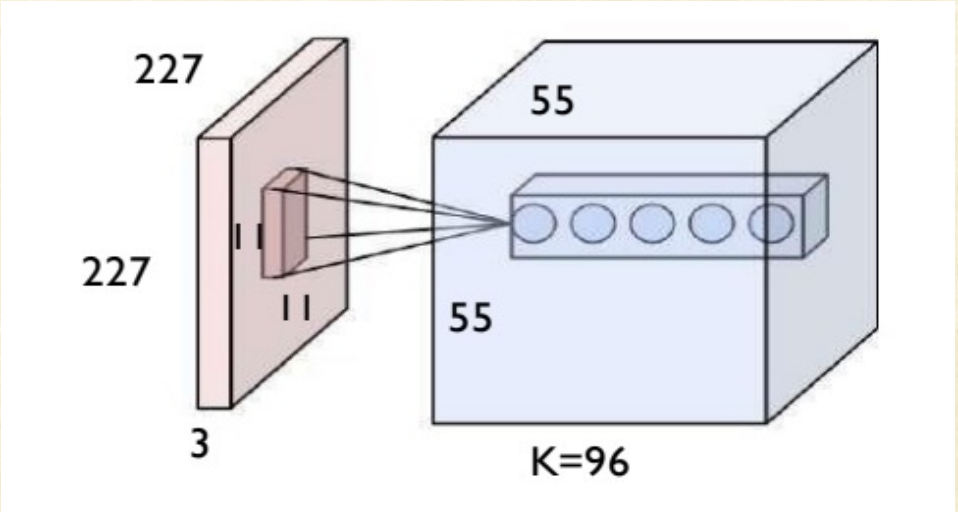
- Filter Size: F
- Input volume streams: D
- **#** filters: K
- **#** parameters in a layer is **(F .F .D) .K**

  Example:
- For layer 1, Input images are 227 x 227 x 3
- F = 11 and K = 96
- Each filter has 11 x 11 x 3 = 363 and 1 (bias) i.e., 364 weights
- # weights = 364 x 96 = **35 K** (approx.)

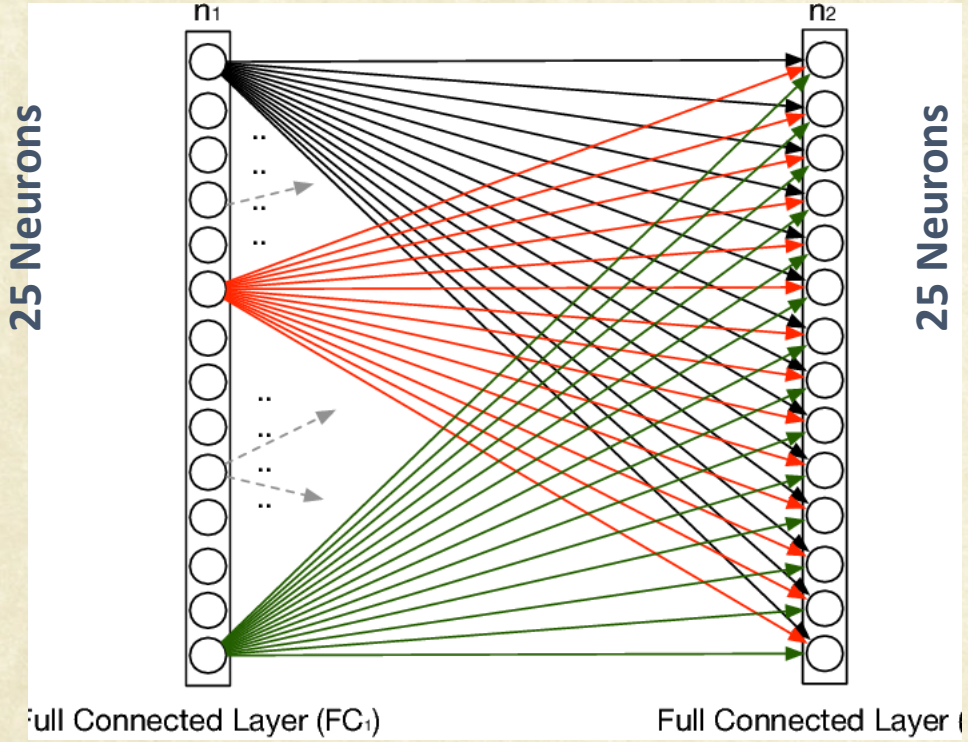


Parameter Calculation

- Stride S
- Zero padding P
- Input Size: W_1 x H_1 x D_1
- Output Size: **W_2 x H_2 x D_2**
- W_2 = [(W_1 – F + 2P) / S] + 1 and D_2 = K
- S = 4, W = 227, F =11, P = 0, D_2 = 96
- W_2 = (227 -11)/4 + 1 = 55
- Output Size: **55 x 55 x 96**

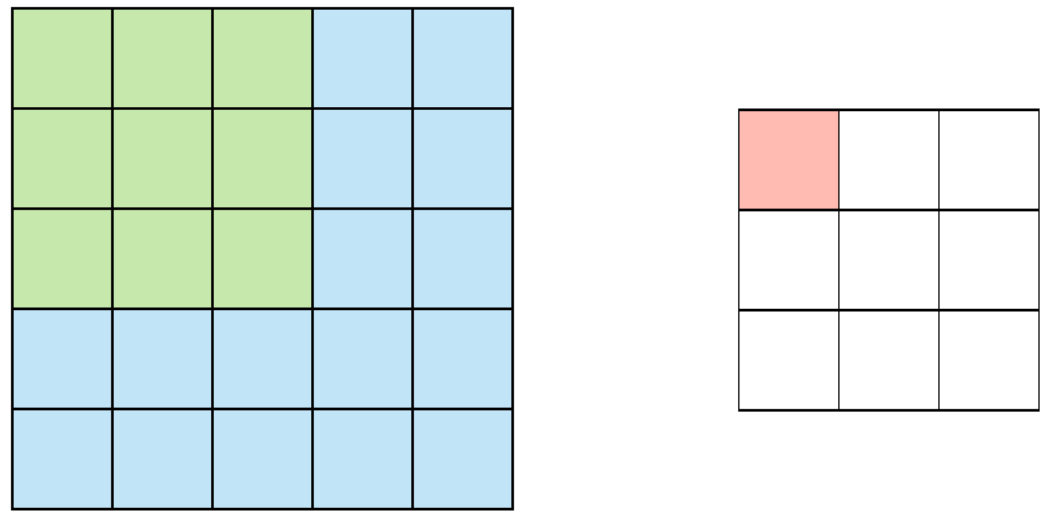


Fully Connected vs. Convolutional Layer



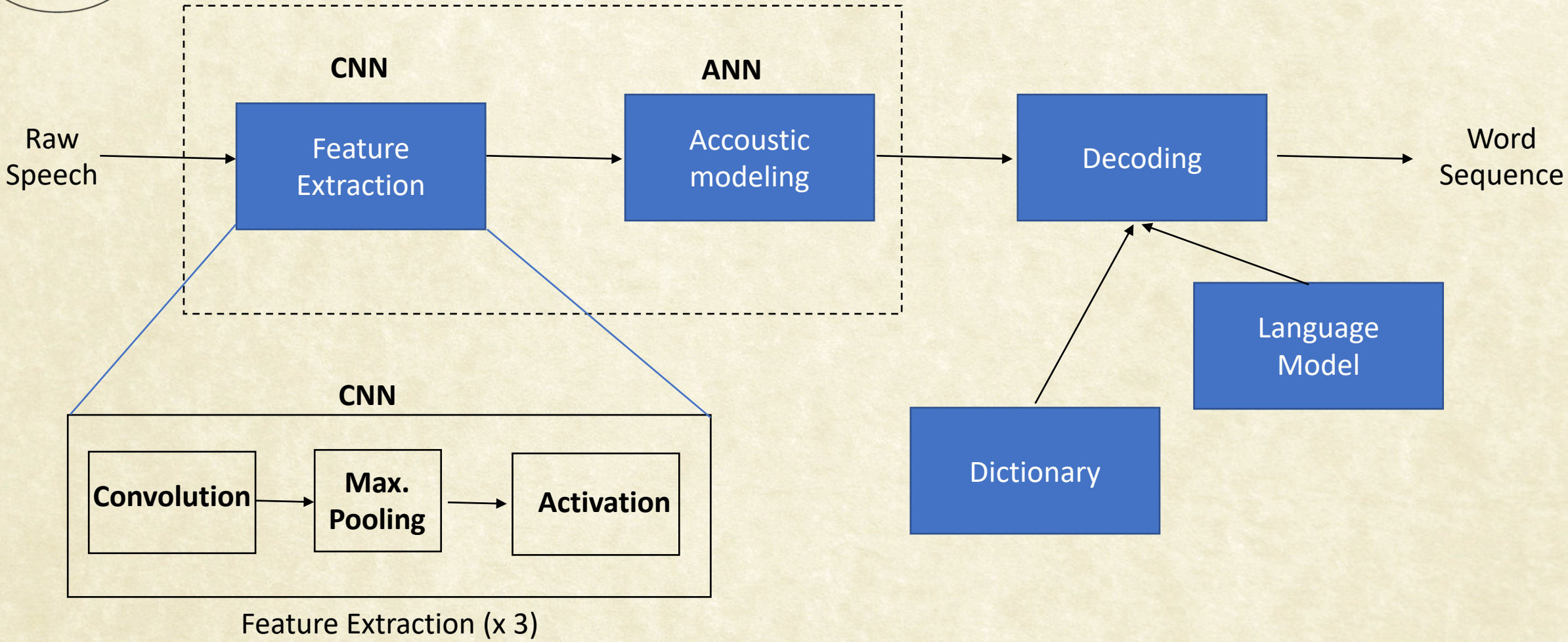
- Consider a 5*5 image (i.e., 25 pixels/values)
- **#Neurons** in L_1, L_2 = n_1, n_2 = 25
- **Weights** = $n_1 * n_2$ = 25 * 25 = 625
- **Operations**:
  - Multiplications(M) = n_2 per each neuron in L_1 = 25
  - Additions (A) = $n_2 - 1$ per neuron in L_1 = 24
  - Total operations = (M+A)* n_1 = (25+24)*25 = 1225

Fully Connected vs. Convolutional Layer



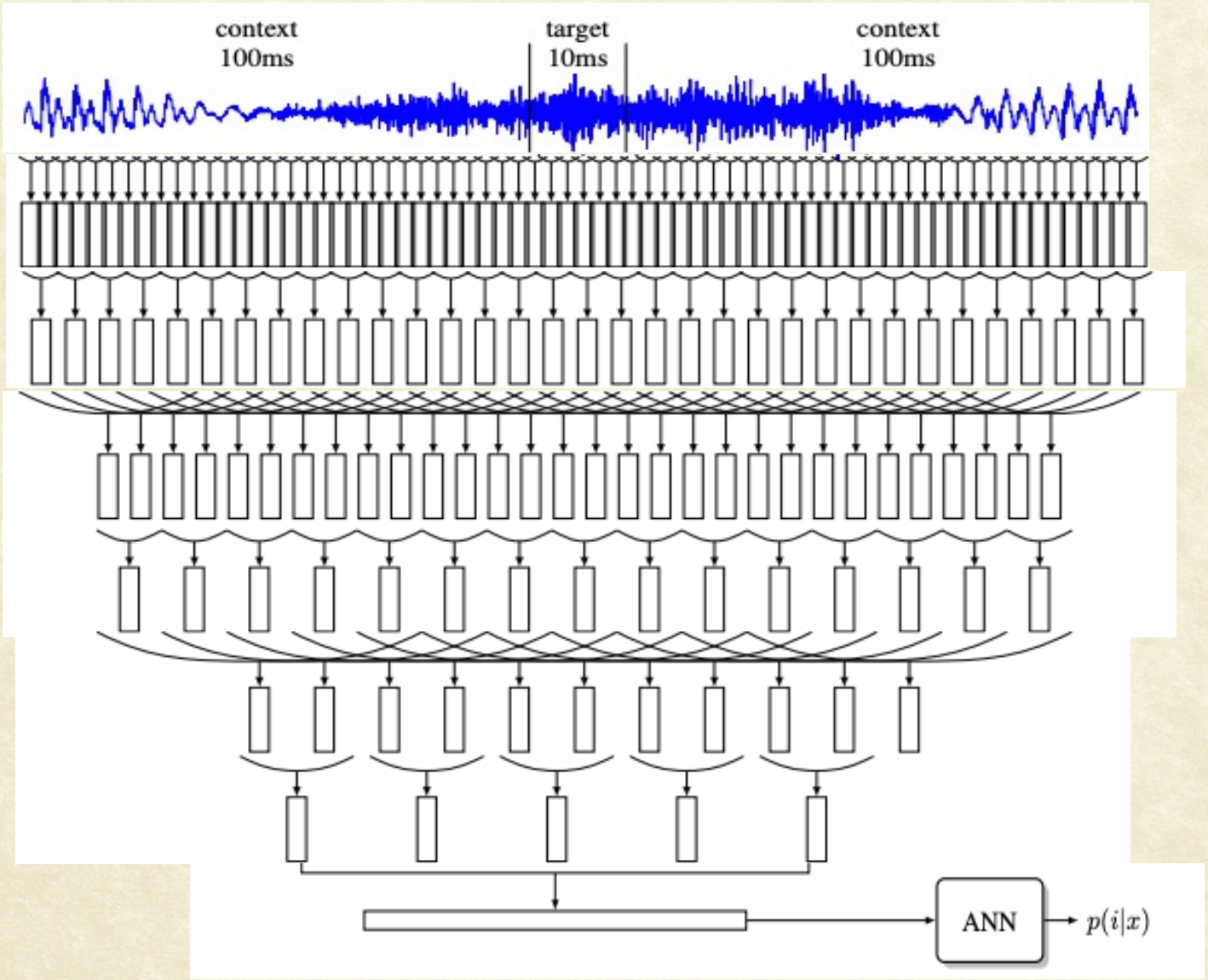
- Consider a 5 * 5 Image
- **#Filters** (n)= 25
- **Filter size** (K)= 3*3 (9 weights)
- **Stride**(S)= 1
- **Total steps** to cover the image (TS) = 9
- **Total Weights** = K * n = 9*25 = 225
- **Operations**
  - Multiplications (M)= 9 per filter at one step
  - Additions (A) = 8 per filter at one step
  - **Total Operations** = (M+A) * TS * n
    = (8+9)*9*25
    = **3825**

Speech Recognition with CNN



Dimitri Palaz, Mathew Magimai-Doss and Ronan Collobert, “Convolutional Neural Networks-based Continuous Speech Recognition using Raw Speech Signal”, ICASSP 2015, pp.4295-4299.

Detailed View



Convolution

Max pooling

Convolution

Max pooling

Convolution

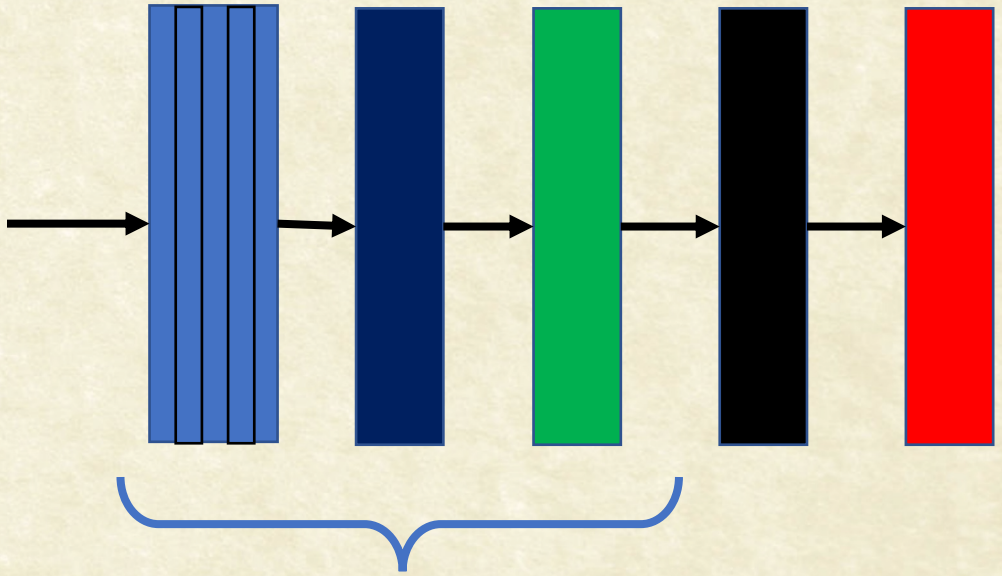
Max pooling

WSJ Word Error Rate:
64 % vs 5.6 % for
MFCC+ANN vs CNN

https://pdfs.semanticscholar.org/d224/6d41fe4ca1cb1f13eaa7bec0b86b95fdd8ac.pdf

Summary: Layers of a Neural Network

- Based on the connection pattern and operations, we can think of a layer in a Neural Network as:
  - Convolutional
    - A Layer can have multiple Channels
  - Non-Linear (often not drawn)
  - Max-Pooling
  - Fully Connected
  - Soft Max



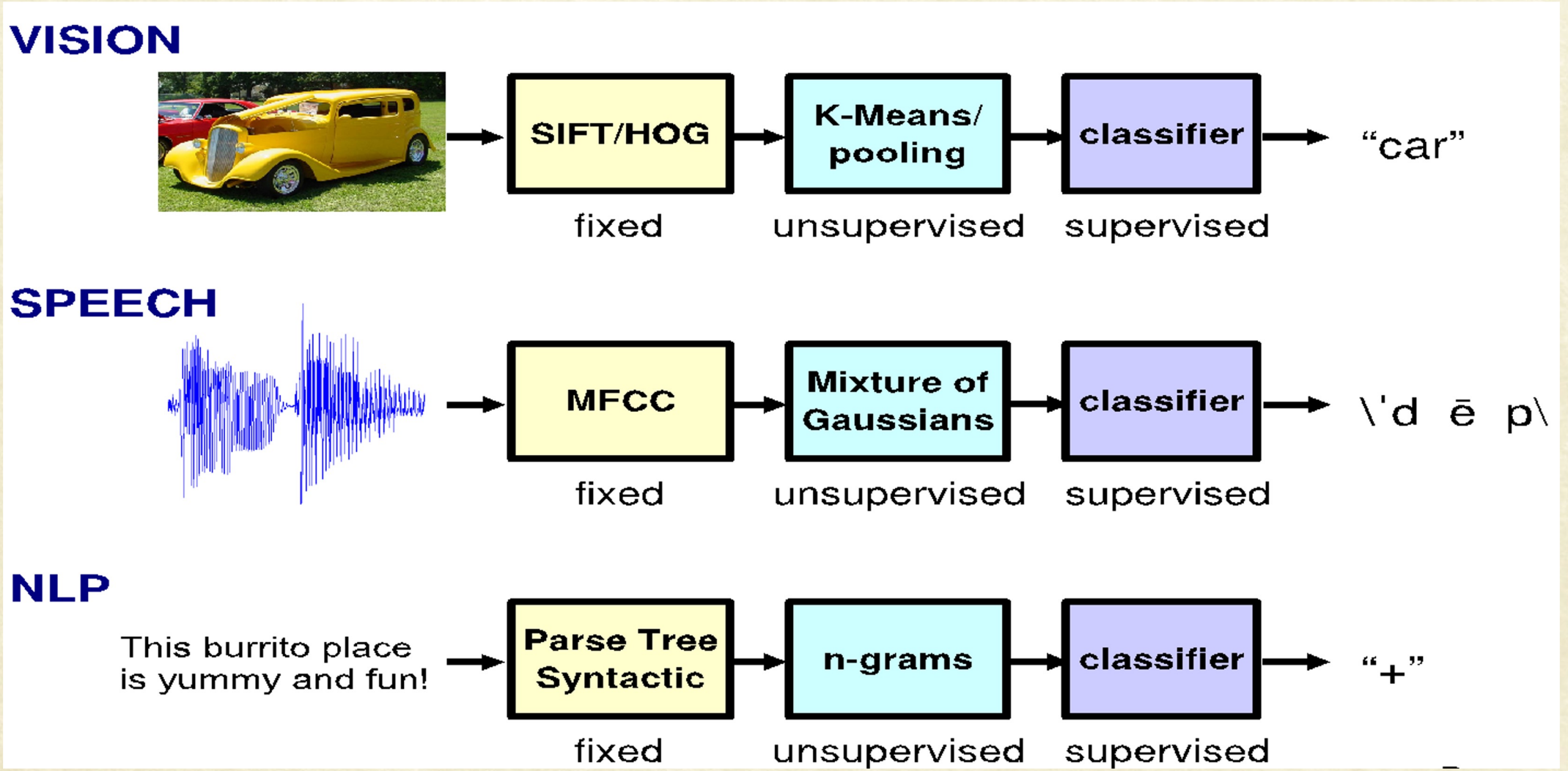
This is often repeated multiple times

Questions?

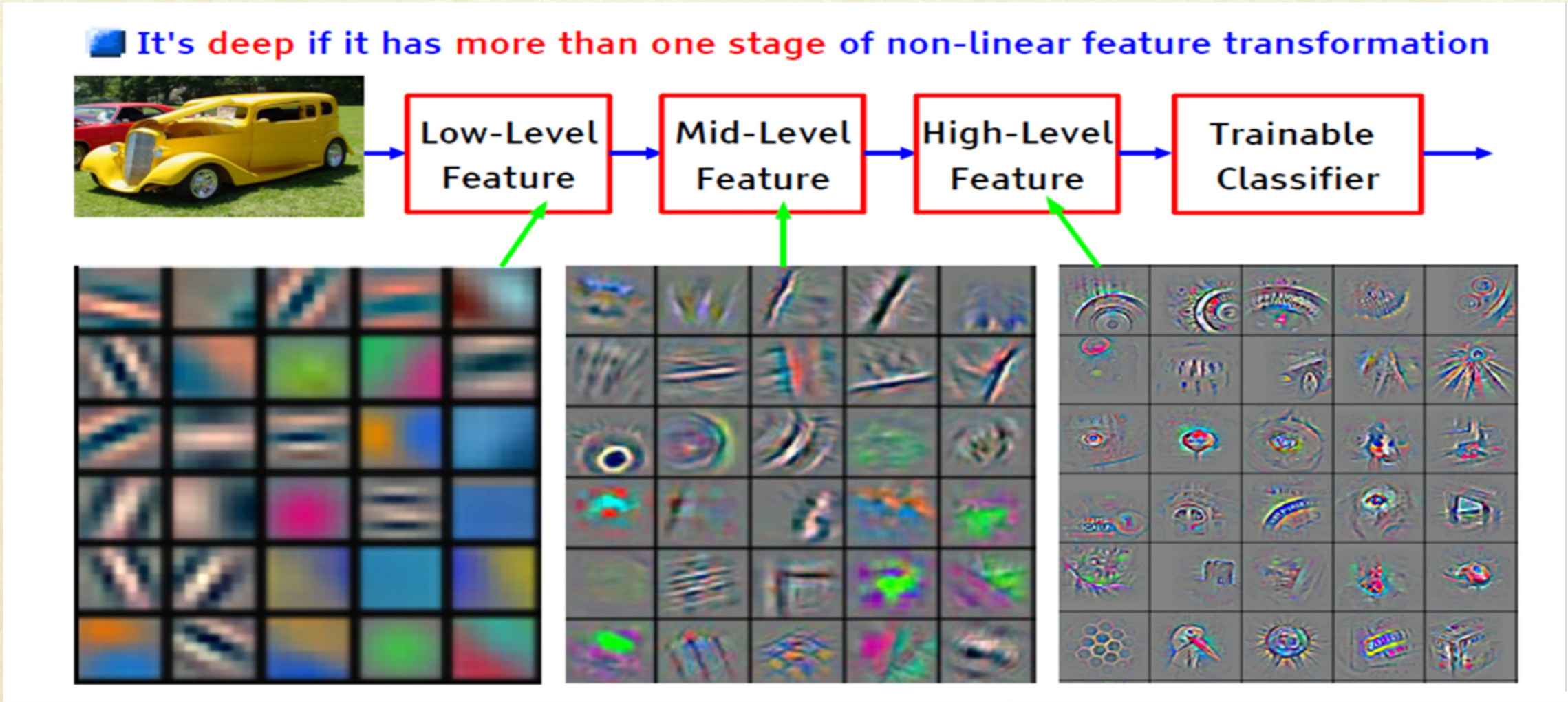
Visualizing CNNs

What did it learn?

Common Pipeline: Till Then



Deep Learnt Features



Learn the full pipeline

VISION

pixels → edge → texton → motif → part → object

SPEECH

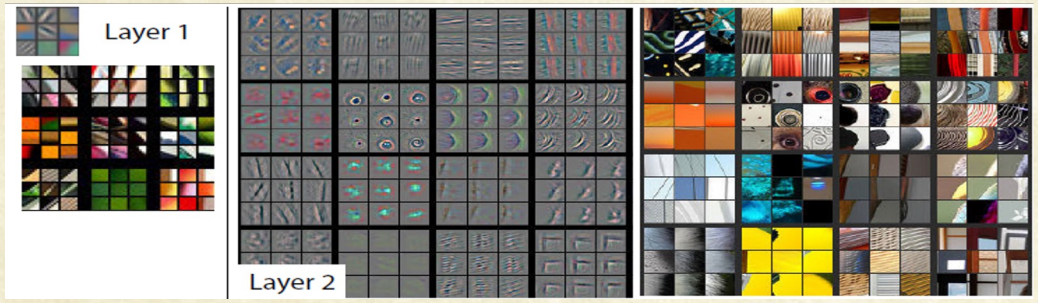
sample → spectral band → formant → motif → phone → word

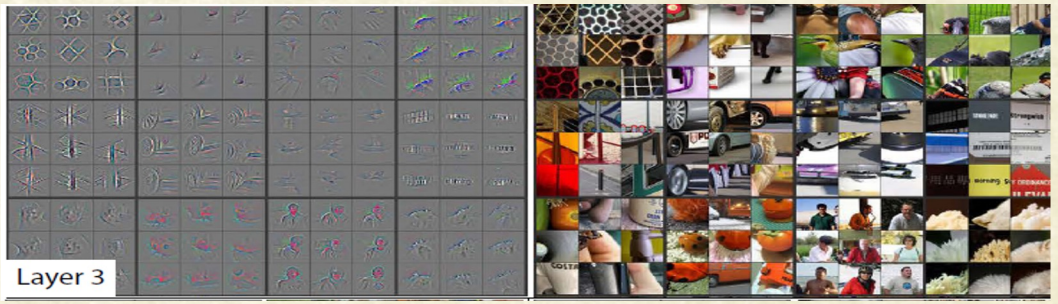
NLP

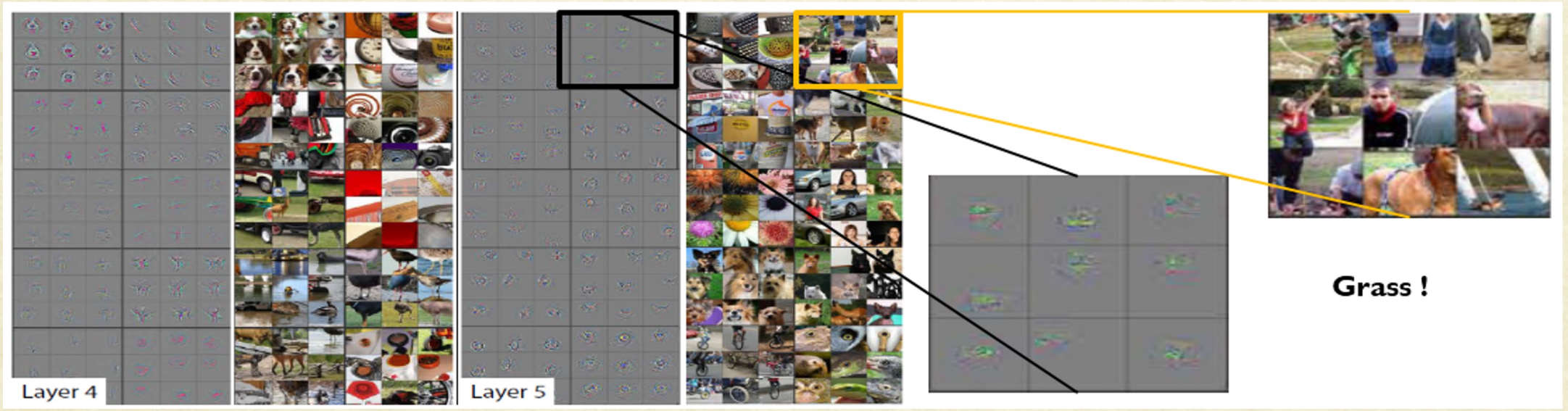
character → word → NP/VP/.. → clause → sentence → story

Visualizing CNNs

A. How do I interpret the learned filters?







Source: Zeiler e.t. al. ECCV'14

Zeiler and Fergus , Visualizing and Understanding Convolutional Networks,. ECCV 2014

Early Layers Converge Faster

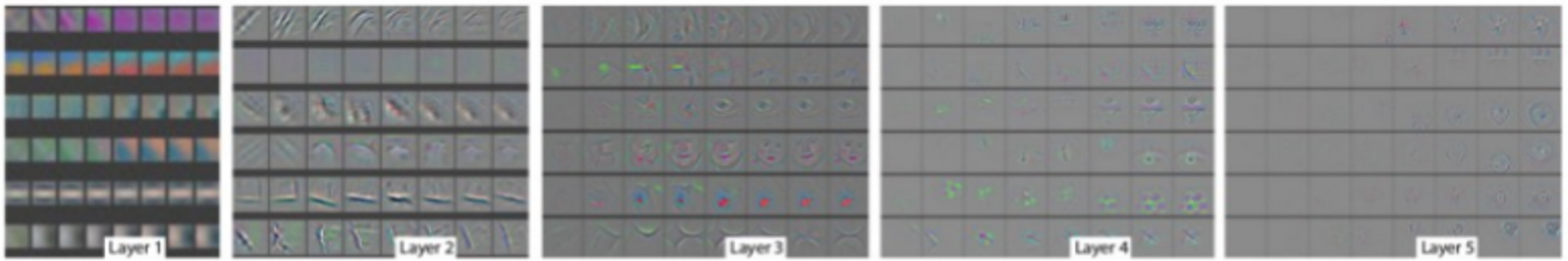
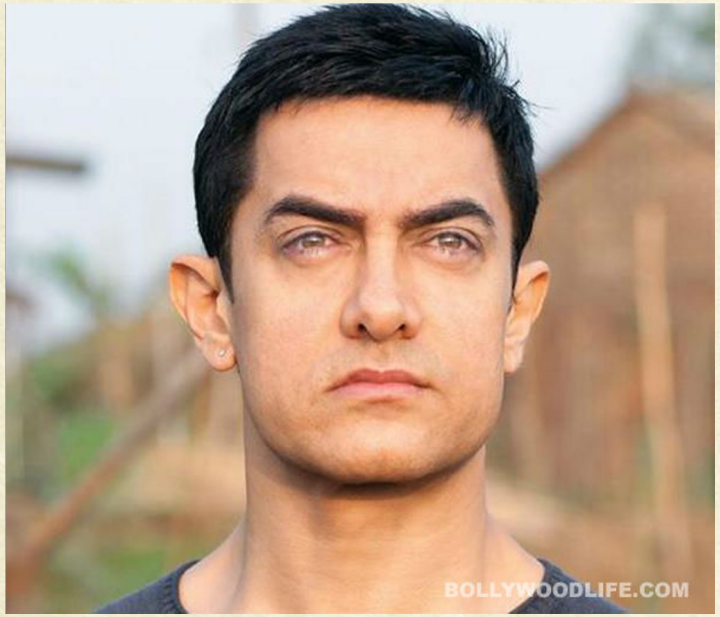
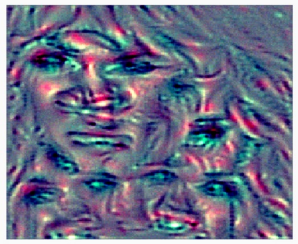


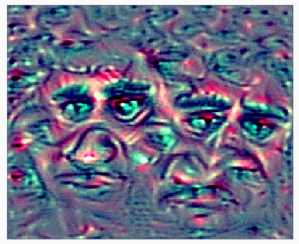
Figure: Evolution of randomly chosen subset of model features generated using deconvnet through training at epoch 1, 2, 5, 10, 20, 30, 40, 64.

Deep Dream




(a) Class 93


(b) Class 301

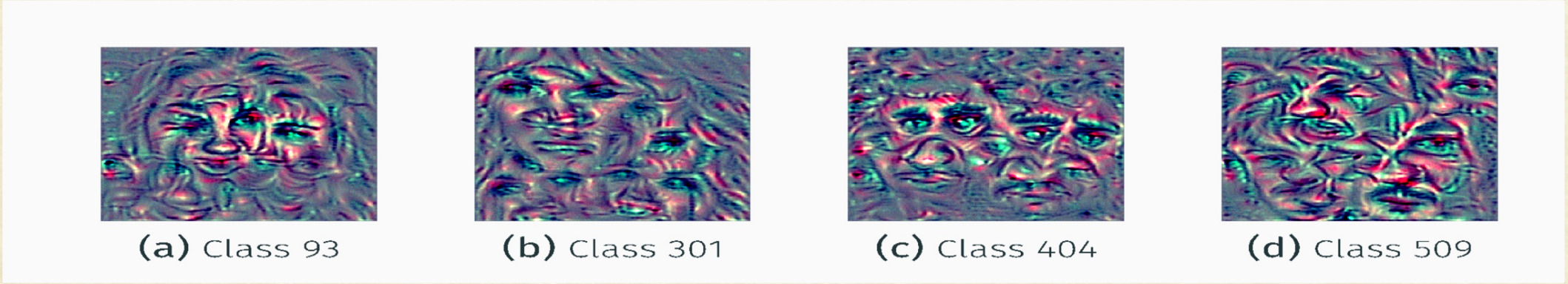

(c) Class 404


(d) Class 509

Simonyan et al. NIPS 2014,
Mahendran et al. CVPR 2015

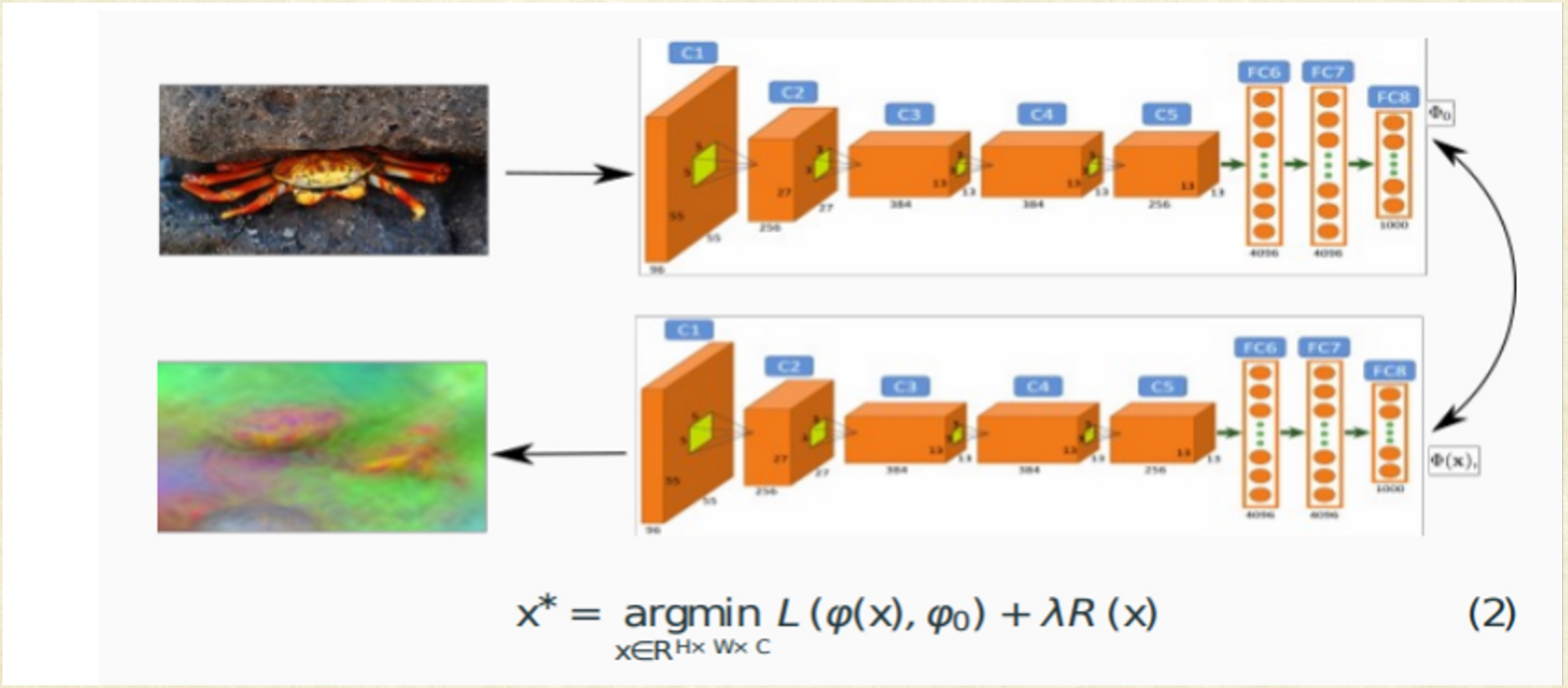
Visualizing CNNs

- Class Model Visualization
  - Find an L_2 normalized image which maximizes the C_l class score
  - Initialize with mean image.
  - Back-propagate



Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. CoRR 2014

Inverting Specific Representation



Aravindh Mahendran and Andrea Vedaldi, Understanding Deep Image Representations by Inverting Them, CVPR'15

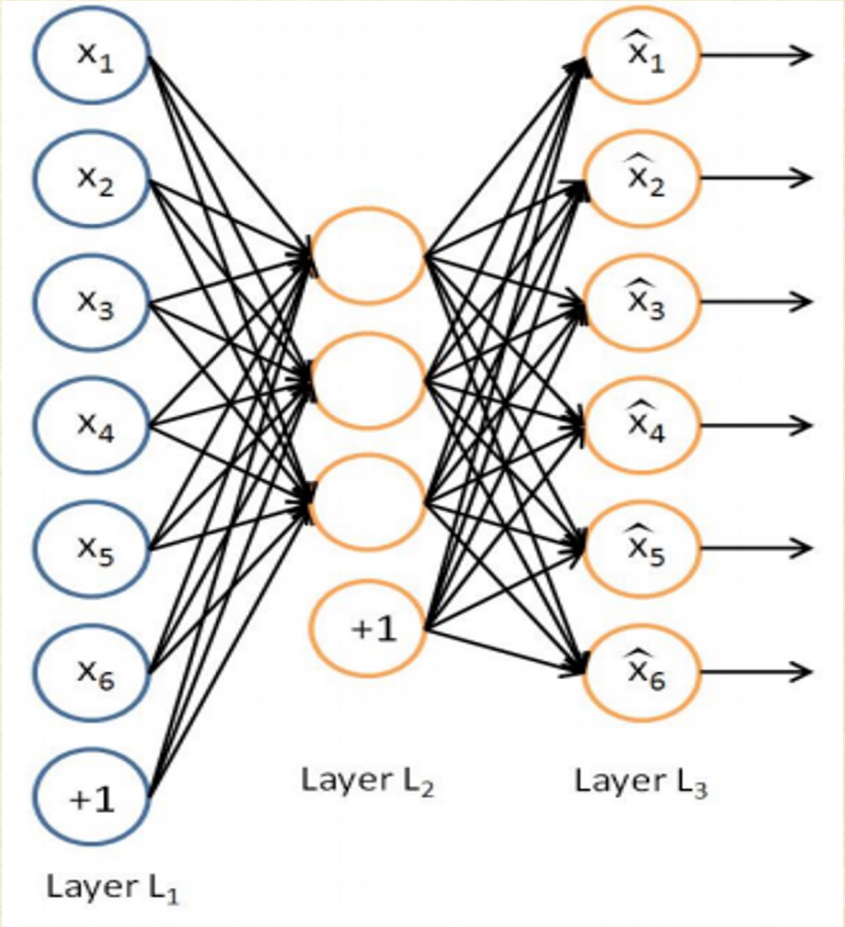
Questions?

Auto Encoders

Learning without Labels

Auto-encoder

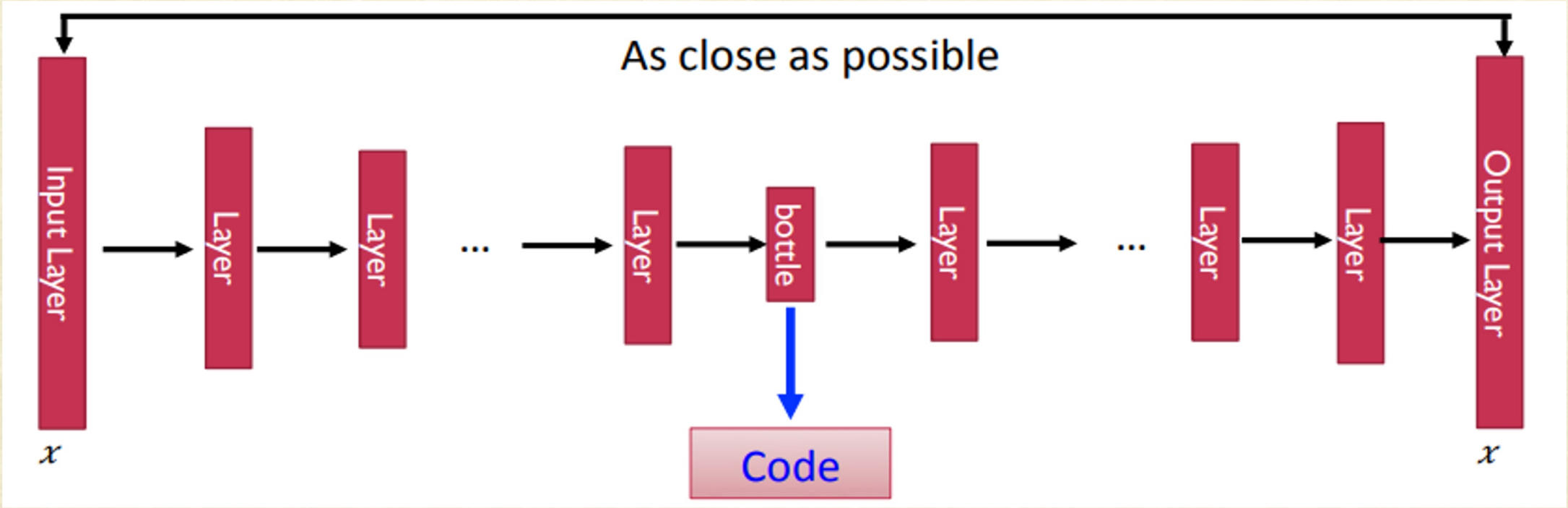
- Similar to MLP
- Input is same as output
- Network learns to reconstruct.
- “Bottleneck” layer learns a compact representation.



Deep Auto-Encoders

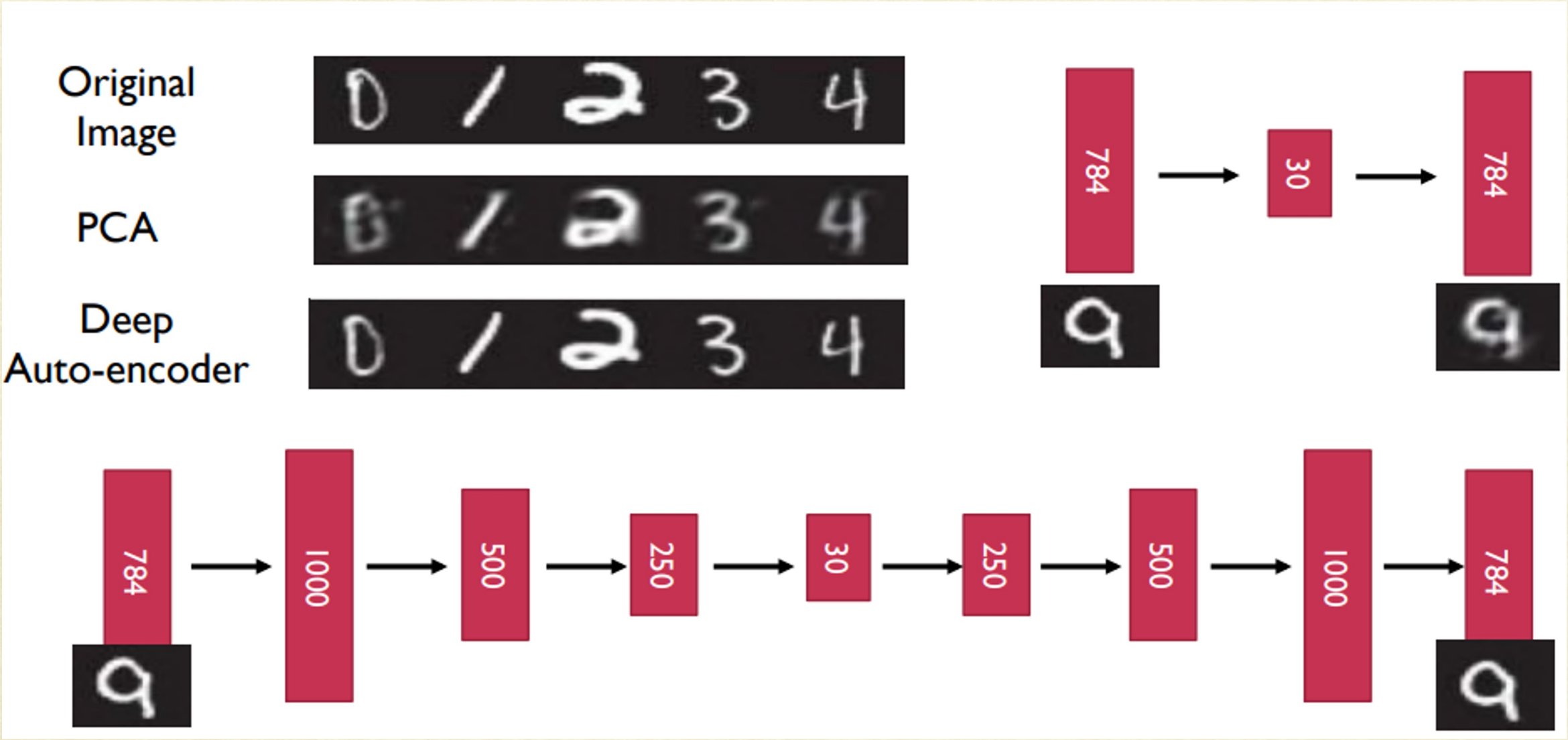
- Of course, the auto-encoder can be deep

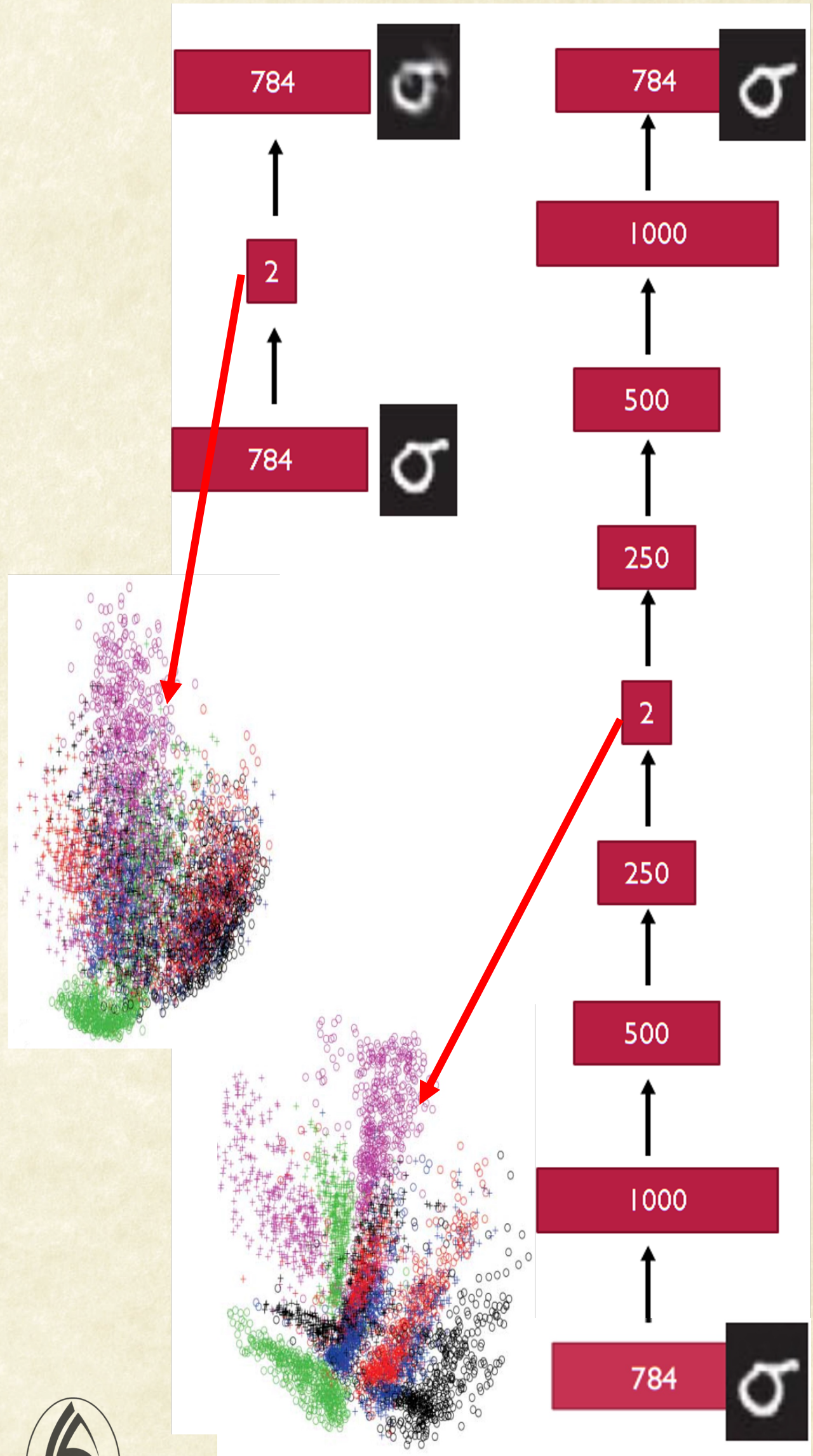
Symmetric is not necessary



Reference: Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507

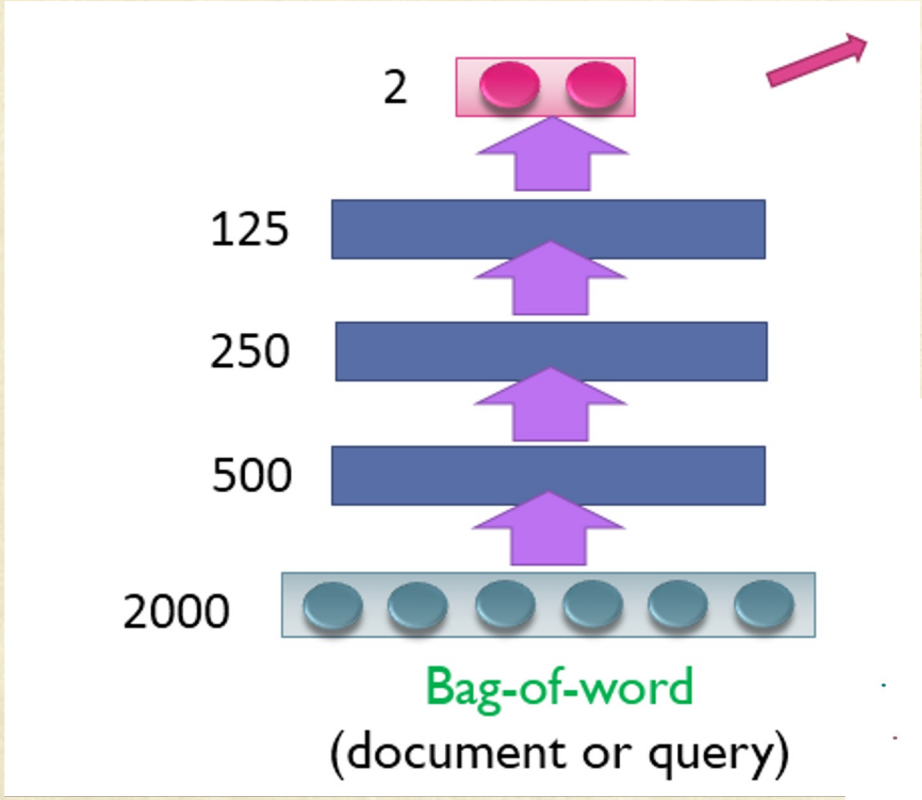
Deep Auto-Encoders

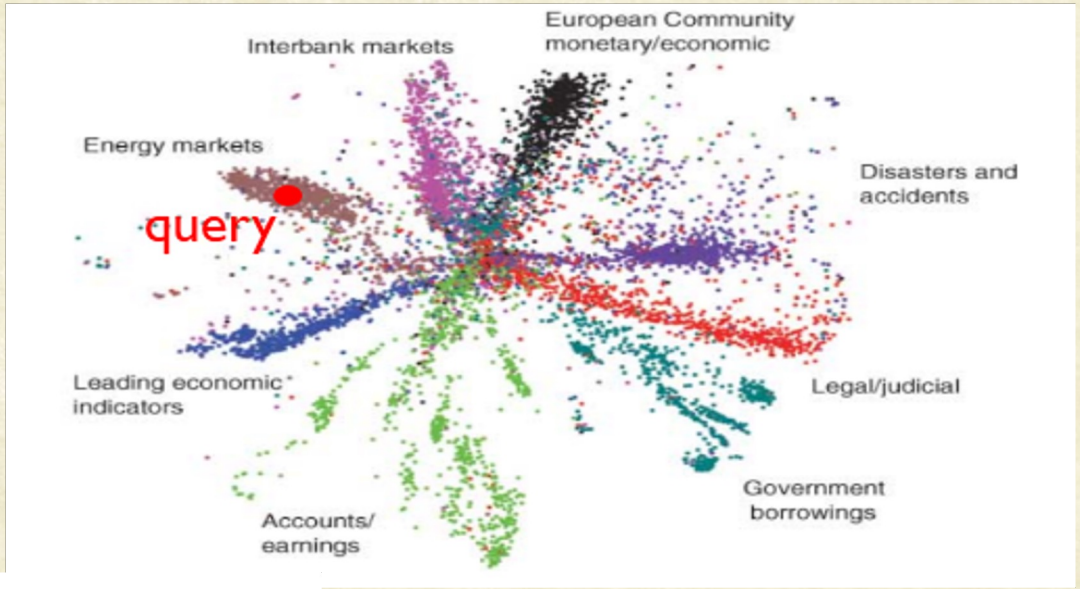


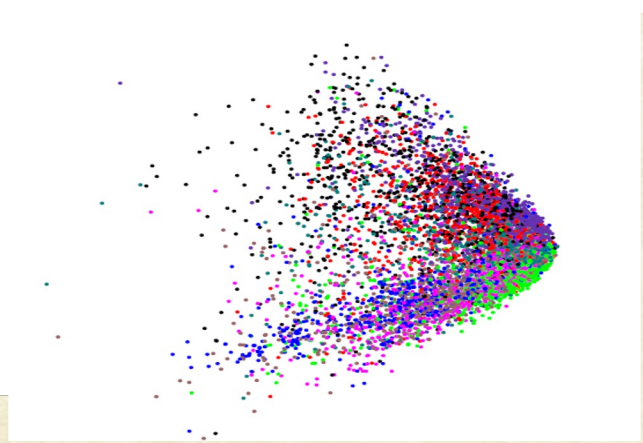

784
2
784
784
1000
500
250
2
250
500
1000
784

Auto-encoder – Text Retrieval

The documents talking about the same thing will have close code.



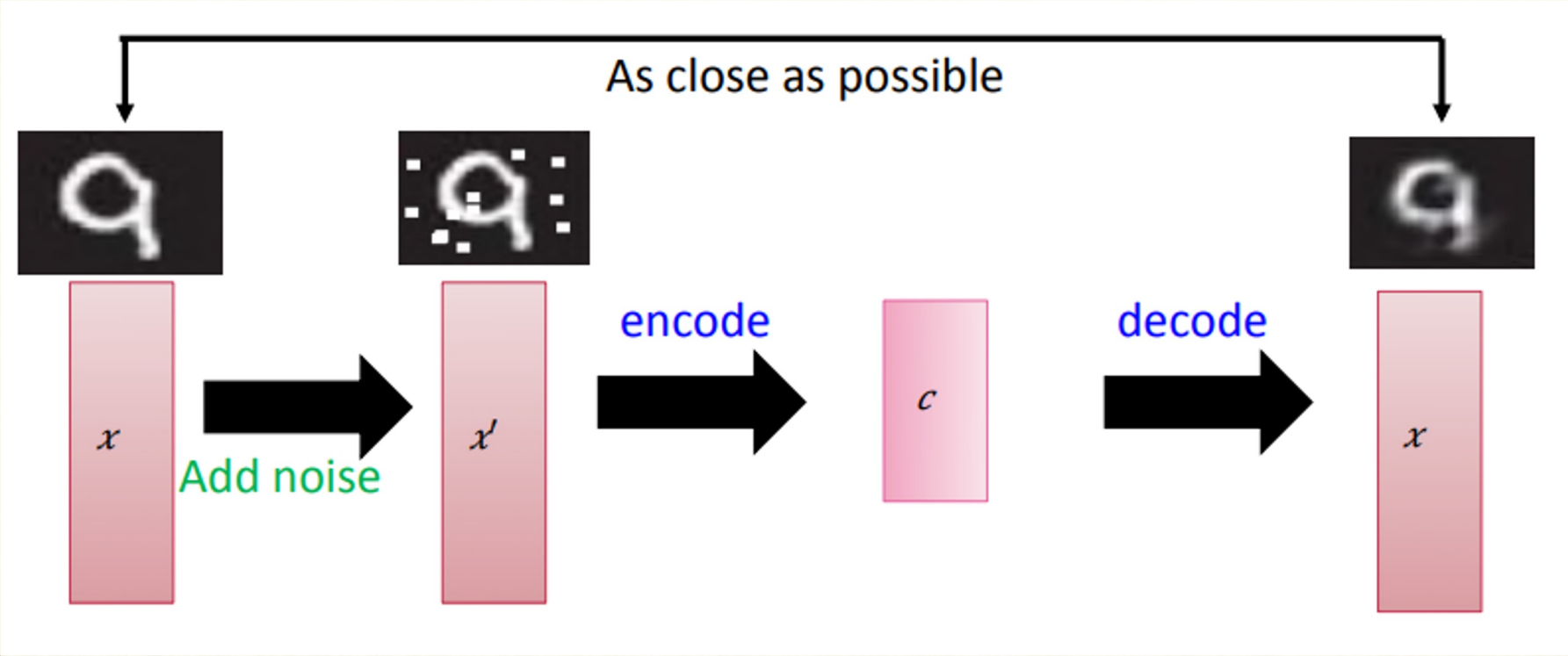




LSA: project documents to 2 latent topics

Auto-encoder

- De-noising auto-encoder



Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." *ICML,* 2008.

Questions?

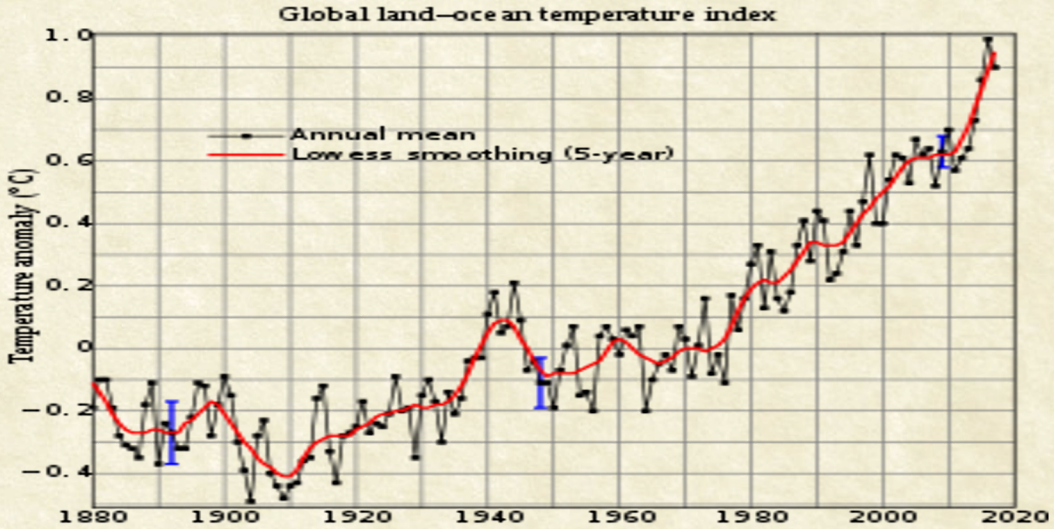
Time Series Models

Remember the Past

Examples

BSE SENSEX

Global Land Ocean temperature



Examples

Day	No. of Packets of Milk sold
Monday	90
Tuesday	88
Wednesday	85
Thursday	75
Friday	72
Saturday	90
Sunday	102

Year	Population(in Million)
1921	251
1931	279
1941	319
1951	361
1961	439
1971	548
1981	685

Time Series

- Time series is a sequence of observations often ordered in time.
- Popular Problem: Given a sequence, predict future samples.
- Applications:
  - Meteorology,
  - Finance,
  - Marketing etc.

Notation and Problem

- Notation: x[0], x[1], x[2], …, x[N].
- X[t], Where t is the time or index in the sequence.
- Assumption: Measurement at time t depends on three previous ones.
  - i.e., t-1, t-2 and t-3
- Why 3? We can have a different number.

Data

Raw Data	
Time	Sample
1	X_1
2	X_2
3	X_3
4	X_4
5	X_5
6	X_6
7	X_7

Rearranged Data			
Feature-1	Feature-2	Feature-3	Y_i
X_1	X_2	X_3	X_4
X_2	X_3	X_4	X_5
X_3	X_4	X_5	X_6
X_4	X_5	X_6	X_7

Feature Vector	
Feature	Y_i
V_1	X_4
V_2	X_5
V_3	X_6
V_4	X_7

A Simple Model

- X[t] = w1 X[t-1] + w2 X[t-2] + w3 X[t-3] + n
  - Where n is noise.
- Problem:
  - Given the sequence X[0], X[1], ... X[N]
  - Find coefficients w1, w2, w3
- Find the coefficients w1,w2,w3 such that prediction error is minimal.

Performance Metrics For Time Series Data

- We need a way to compare different time series techniques for a given data set.
- Four common techniques are:
  - mean absolute deviation,
  - mean absolute percent error,
  - the mean square error,
  - root mean square error.

$$X_i : ACTUAL$$

$$\hat{X}_i : PREDICTED$$

$$\mathrm{MAD} = \sum_{i=1}^{n} \frac{\left|X_i - \hat{X}_i\right|}{n}$$

$$\mathrm{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{\left|X_i - \hat{X}_i\right|}{\hat{X}_i}$$

$$\mathrm{MSE} = \sum_{i=1}^{n} \frac{\left(X_i - \hat{X}_i\right)^2}{n}$$

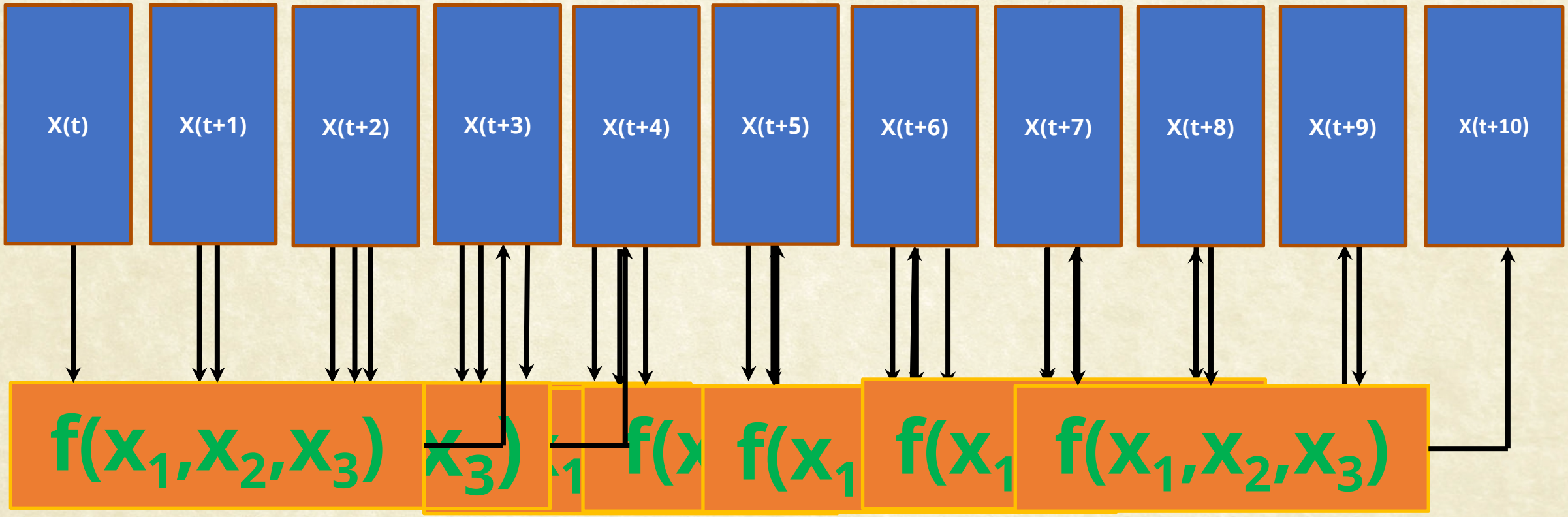
$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

More Powerful Model

- $X_t = f(W, X_{t-1}, X_{t-2}, X_{t-3}) + n$
- Problem:
  - Given the sequence $X_0, X_1, \ldots\ldots X_N$
  - Find coefficients W
- Data may be modeled as in the above linear case.
- f() may be seen as a MLP

$$\min W \sum_{t=3}^{N} (X_t - f(W, X_{t-1}, X_{t-2}, X_{t-3}))^2$$

Time series prediction

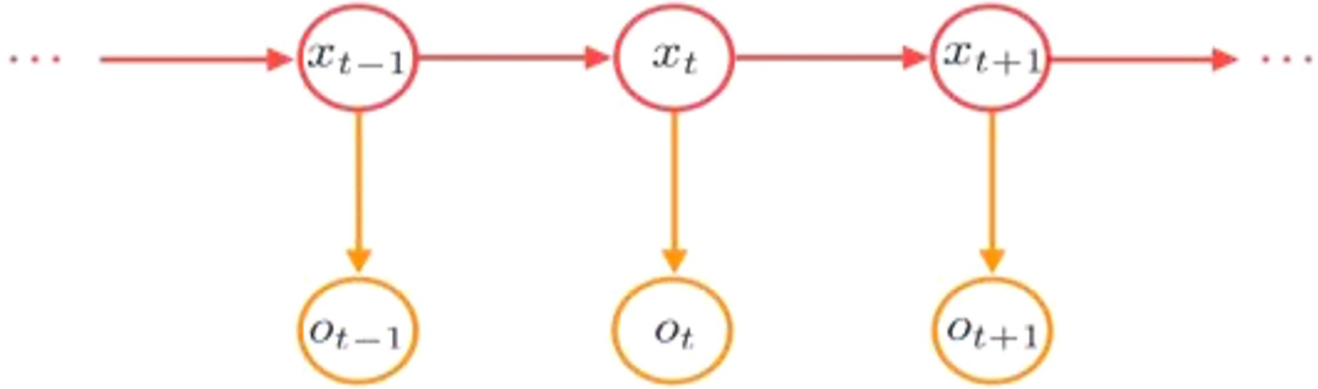


Summary

- Predicting future samples is a new problem
- However, solution is similar to what we know.
  - Cast as regression.
- Model can be linear
  - Linear regression
- Or nonlinear
  - MLP
- On how many past samples, the future sample will depend?
  - Order/model to be guessed?

Questions?

Dynamic Systems

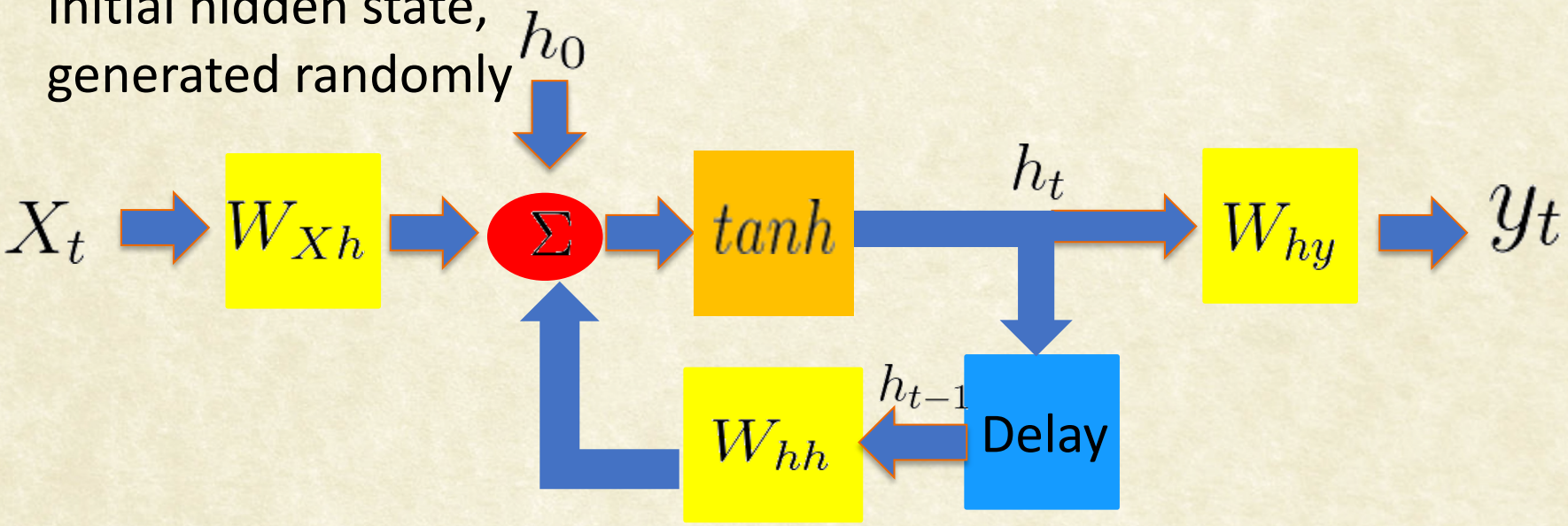


$$x_{t+1} = Ax_t + \epsilon \qquad \epsilon \sim N(0, Q)$$
$$o_t = Cx_t + \upsilon \qquad \upsilon \sim N(0, R)$$

$$\theta^{\star} = \{A, C, Q, R, x_1\}$$

Peep into RNNs

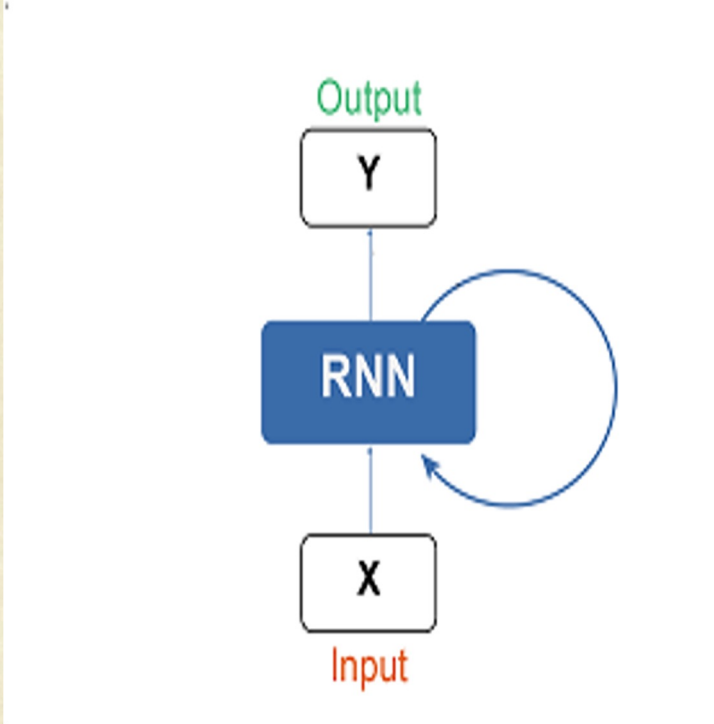
Initial hidden state, generated randomly

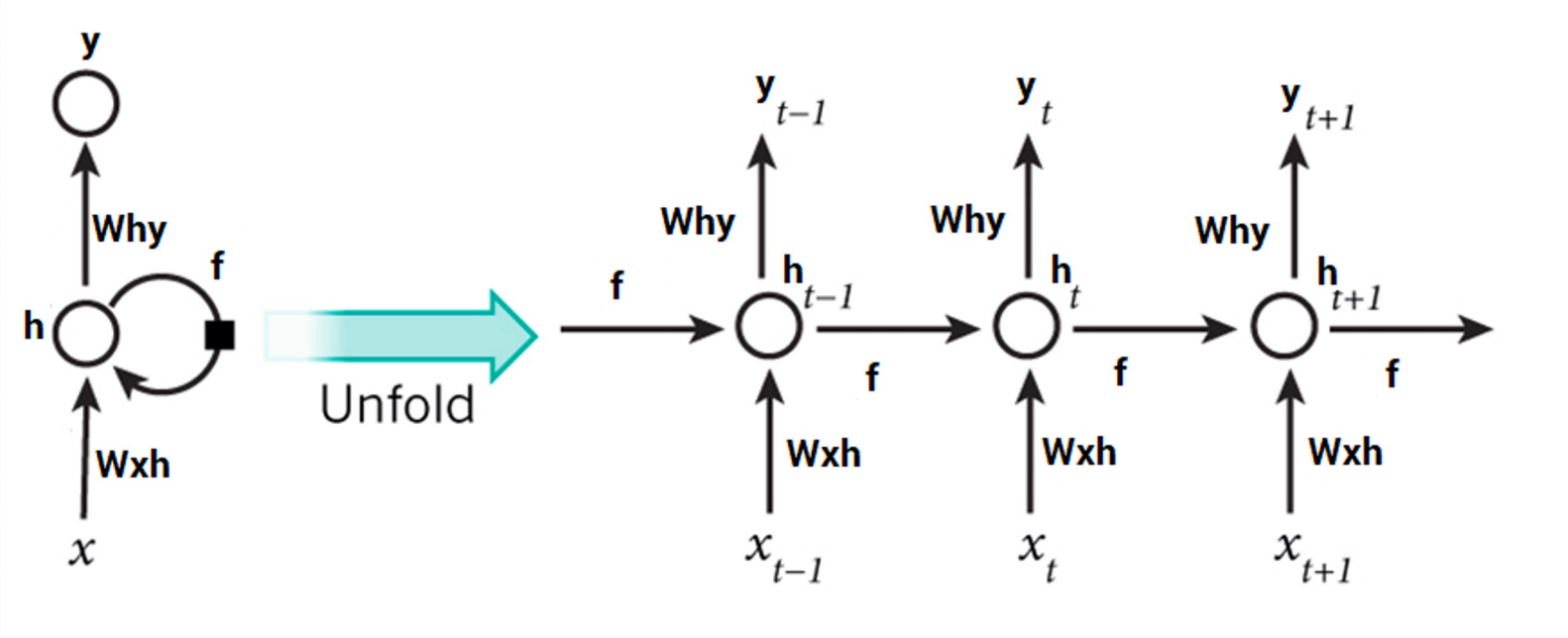


$$h_t = tanh(W_{hh}h_{t-1} + W_{Xh}X_t)$$

$$y_t = W_{hy}h_t$$

Peep into RNNs : (more later)





$$h_t = tanh(W_{hh}h_{t-1} + W_{Xh}X_t)$$

$$y_t = W_{hy}h_t$$

Questions?