ONLY GOD

1402-2023

mansoori@shirazu.ac.ir

Neural Network & Deep Learning

Deep Learning

CSE & IT Department

School of ECE

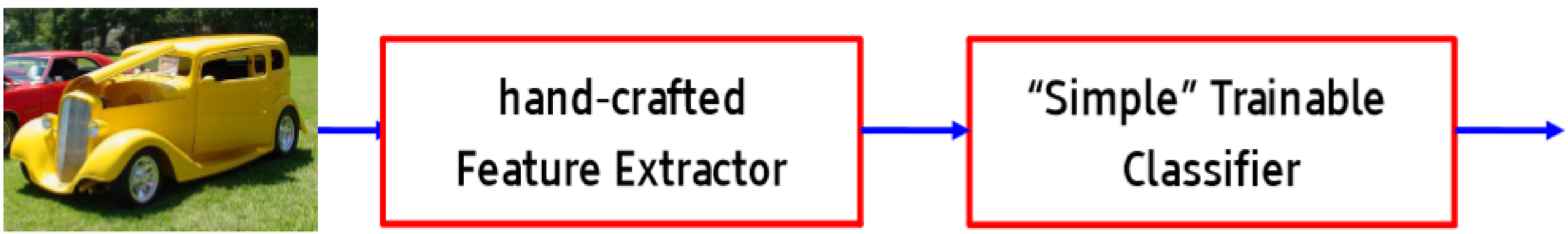
Shiraz University

Traditional Learning

mansoori@shirazu.ac.ir

Traditional model of pattern recognition (PR) (since late 50's)

- Fixed/engineered features + trainable classifier



Traditional vs Deep Learning

mansoori@shirazu.ac.ir

- Traditional PR: Fixed/Handcrafted feature extractor

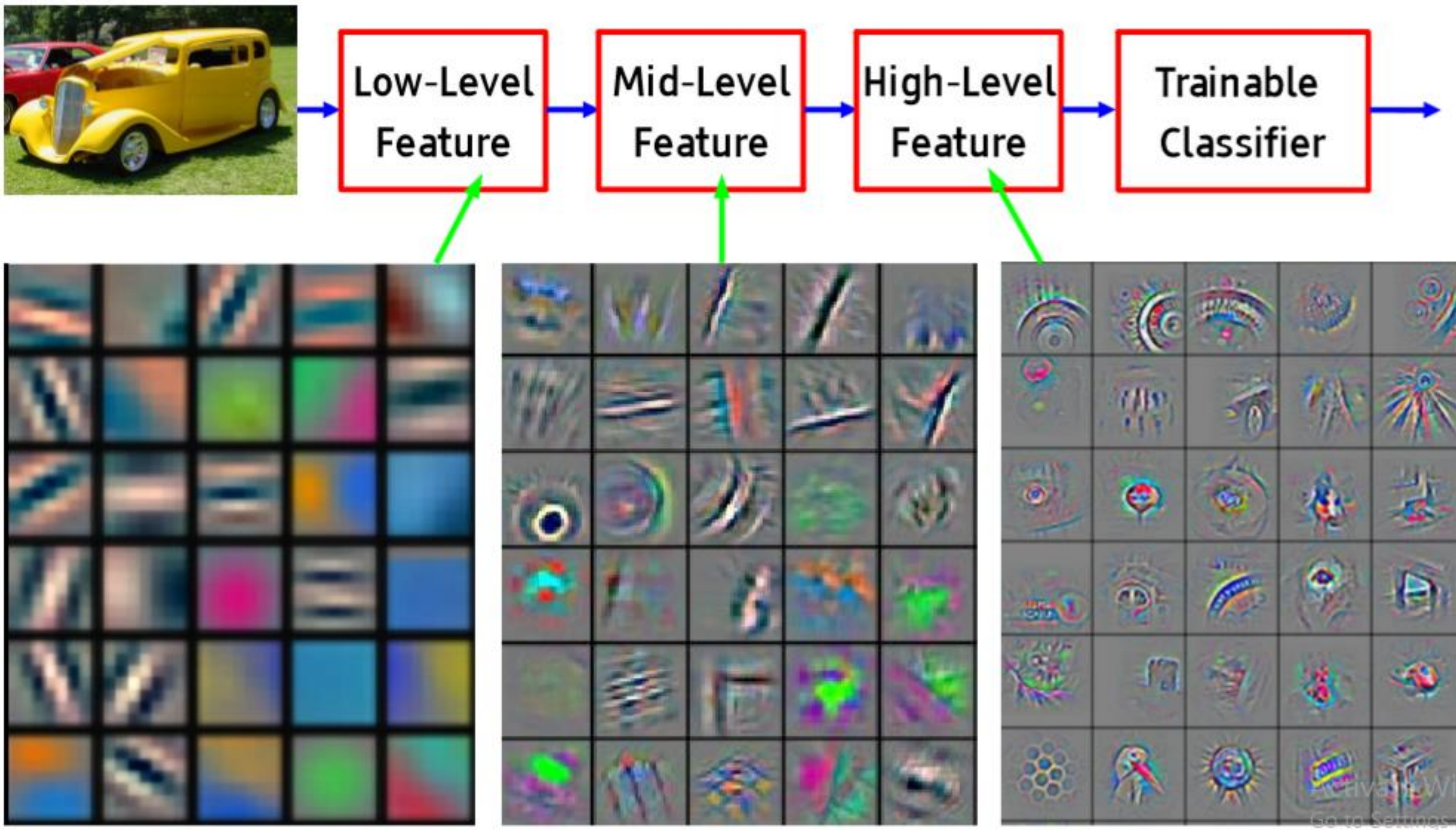
- Modern PR: Unsupervised mid-level features

- Deep learning PR: Representations are hierarchical and

Deep Learning

mansoori@shirazu.ac.ir

- Learns hierarchical representations



Trainable Feature Hierarchy

mansoori@shirazu.ac.ir

Hierarchy of representations with increasing abstraction level

- Each stage is a kind of trainable feature transform
- Image recognition

  pixel --> edge --> texton --> motif --> part --> object

- Text

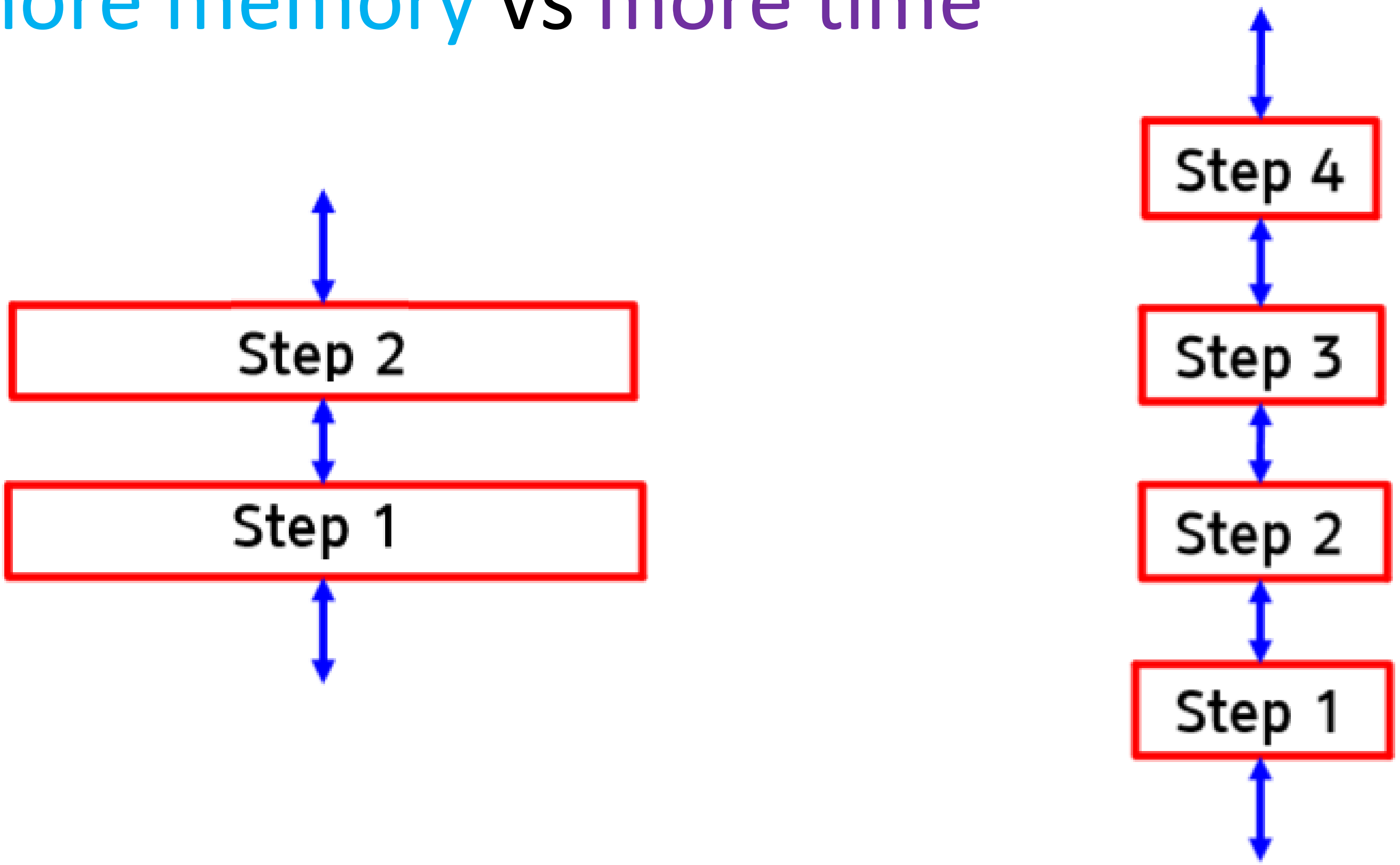
  character --> word --> word group --> clause --> sentence --> story

- Speech

  sample --> spectral band --> sound --> ... --> phone --> phoneme --> word

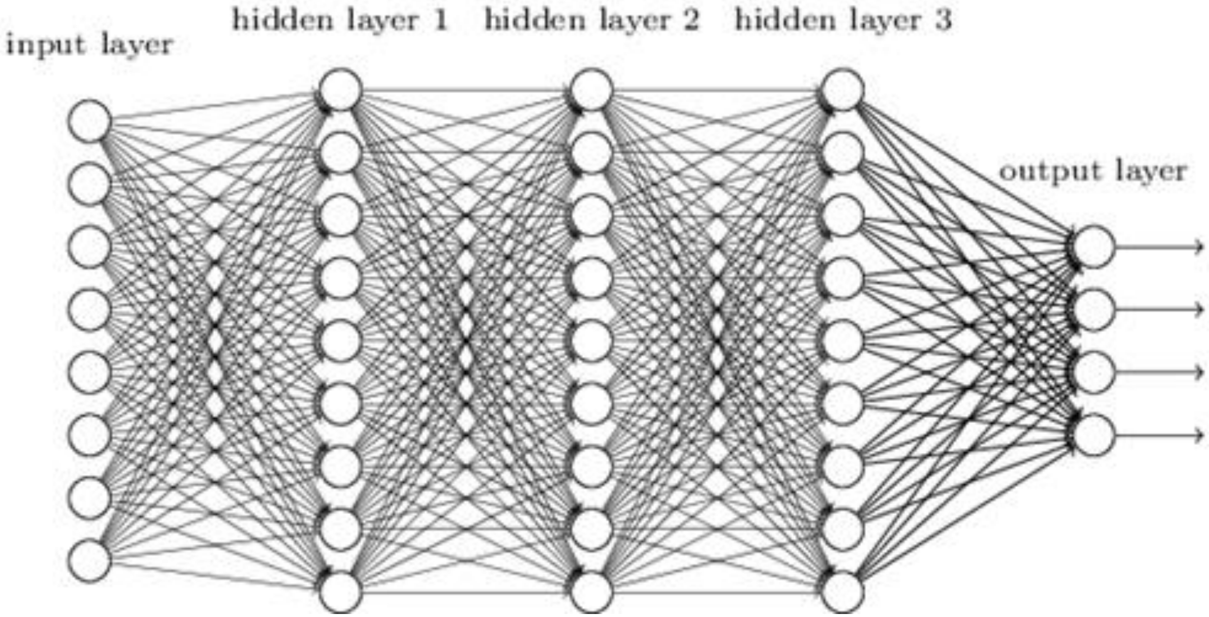
Shallow vs Deep

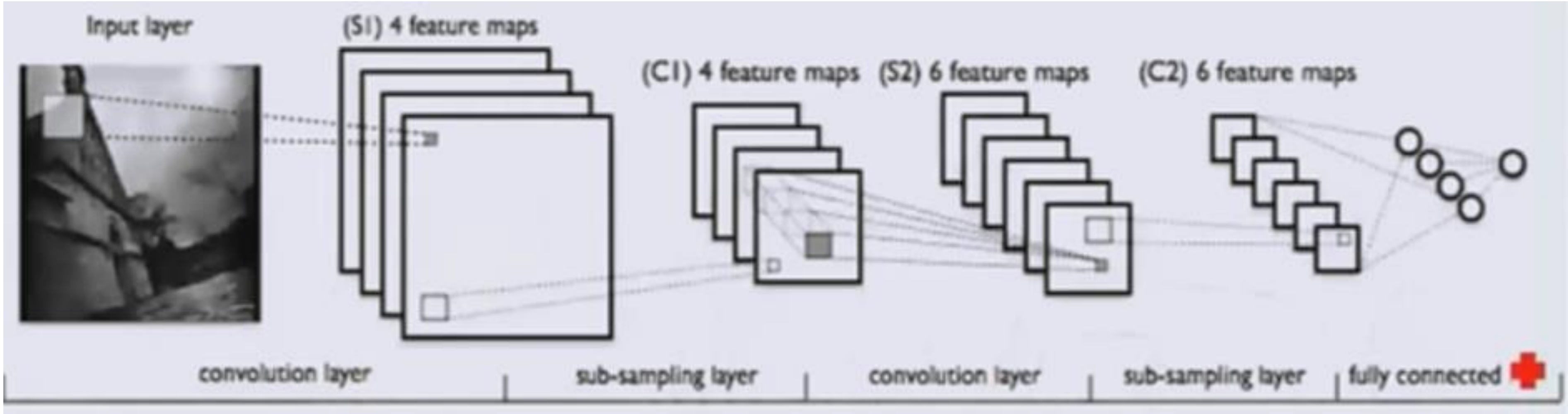
- shallow and wide vs deep and narrow
- more memory vs more time



Deep Networks

m
a
n
s
o
o
r
i @shirazu.ac.ir





DL difficulties (Unavailability of Data)

mansoori@shirazu.ac.ir

- Labeled data is often scarce
- Training on insufficient data would result in over-fitting
- Unlabeled data is cheap and plentiful
- Unlabeled data is used to learn good initials for weights in all layers (except final layer)
- Labeled data is used to fine-tune weights in all layers
- Often results in much better classifiers being learned

DL difficulties (Local Optima)

mansoori@shirazu.ac.ir

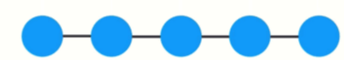
- Training a shallow network using supervised learning, results in weights converging to reasonable outputs
- Training a network using supervised learning involves solving a highly non-convex optimization problem
- In deep networks, this problem turns out to be rife with bad local optima
- So, training with gradient descent (conjugate gradient , …) no longer works well
- New optimization methods like adadelta, adagrad, … are popularly used for training deep networks

DL difficulties (Vanishing Gradient Problem)

- Using back-propagation to compute derivatives, gradients propagated backwards rapidly diminish in magnitude as depth of network increases

- Derivative of overall cost w.r.t weights in earlier layers is very small

- Weights of earlier layers change slowly and these layers fail to learn much



- This vanishing gradient problem occurs in gradient-based learning methods with certain activation functions
- New activation functions are proposed for solving this problem