Introduction to Machine Learning summary

Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 3.5

Artificial Intelligence &
Information Analysis Lab

Introduction to Machine Learning

- **Supervised learning**
  - **Classification/recognition/identification, Identity verification**
  - **Regression, Object detection**
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Introduction to Machine Learning

General notations:

- $\mathbf{x} \in \mathbb{R}^n$: ML model input feature vector.
- $\mathbf{y} \in \mathbb{R}^m$: target label vector.
- $\hat{\mathbf{y}} \in \mathbb{R}^m$: predicted (estimated) ML model output vector.
- N: number of examples in the dataset $\mathcal{D}$.
- n: input vector dimensionality
- m: output dimensionality (e.g. number of classes).

- ***ML model***: a learnable function typically of the form $\hat{\mathbf{y}} = \boldsymbol{f}(\mathbf{x}; \boldsymbol{\theta})$.
  - Its structure may be predefined.
  - Its parameter vector $\boldsymbol{\theta}$ is typically learned through training, by optimizing an error function $J(\mathbf{x}, \boldsymbol{\theta})$.

Classification/Recognition/ Identification

- Given a set of classes $\mathcal{C} = \{\mathcal{C}_i, i = 1, \dots, m\}$ and a sample $\mathbf{x} \in \mathbb{R}^n$, the ML model $\hat{\mathbf{y}} = \boldsymbol{f}(\mathbf{x}; \boldsymbol{\theta})$ predicts a class label vector $\hat{\mathbf{y}} \in [0, 1]^m$ for input sample $\mathbf{x}$, where $\boldsymbol{\theta}$ are the learnable model parameters.

- Essentially, a probabilistic distribution $P(\hat{\mathbf{y}}|\mathbf{x})$ is computed.
- Interpretation: likelihood of the given sample $\mathbf{x}$ belonging to each class $\mathcal{C}_i$.

- Single-target classification:
  - classes $\mathcal{C}_i, i = 1, \dots, m$ are mutually exclusive: $||\hat{\mathbf{y}}||_1 = 1$.
- Multi-target classification:
  - classes $\mathcal{C}_i, i = 1, \dots, m$ are not mutually exclusive : $||\hat{\mathbf{y}}||_1 \geq 1$.

Supervised Learning

- A sufficient large training sample set $\mathcal{D}$ is required for Supervised Learning (regression, classification):

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N\}.$$

- $\mathbf{x}_i \in \mathbb{R}^n$: n –dimensional input (feature) vector of the i-th training sample.
- $\mathbf{y}_i$: its target label (output).
- Target vector $\mathbf{y}$ can be:
    - real-valued vector: $\mathbf{y} \in [0, 1]^m, \mathbf{y} \in \mathbb{R}^m$;
    - binary-valued vector $\mathbf{y} \in \{0,1\}^m$ or even categorical.

Classification/Recognition/ Identification

- ***Training***: Given N pairs of training samples $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in [0,1]^m$, estimate $\boldsymbol{\theta}$ by minimizing a loss function: $\min_{\boldsymbol{\theta}} J(\mathbf{y}, \hat{\mathbf{y}})$.

- ***Inference/testing***: Given N_t pairs of testing examples $\mathcal{D}_t = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N_t\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in [0,1]^m$, compute (*predict*) $\hat{\mathbf{y}}_i$ and calculate a performance metric, e.g., classification accuracy.

Classification/Recognition/ Identification

Optimal step between training and testing:

- ***Validation***: Given N_v pairs of testing examples (different from either training or testing examples) $\mathcal{D}_v = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N_v\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in [0,1]^m$, compute (predict) $\hat{\mathbf{y}}_i$ and validate using a performance metric.
- ***k-fold cross-validation*** (optional):
- Use only a percentage $(100 - \frac{100}{k})\%$, of the data for training and the rest for validation ($\frac{100}{k}\%$, e.g., 20%). Repeat it k times, until all data used for training and testing).
- Example: for 5-fold validation, 5 rounds each using:
  - 80% of the data for training and 20% for testing.

Classification

***Two-class classification*:**

- Two class ($m = 2$) and multiple class ($m > 2$) classification.
- Example: *Face detection* (two classes).

- Two class (binary) classification
  - One (binary) hypothesis to be tested:

$$\mathcal{H}_1: \quad \mathbf{x} \in \mathcal{C}_1, \qquad \mathcal{H}_2: \quad \mathbf{x} \in \mathcal{C}_2.$$

Classification

Multiclass Classification $(m > 2)$**:**

- Multiple $(m > 2)$ hypotheses testing: choose a winner class out of m classes.
- Binary hypothesis testing:
  - One class against all: m binary hypotheses.
    - one must be proven true.
  - Pair-wise class comparisons: $m(m-1)/2$ binary hypotheses.

Artificial Intelligence & Information Analysis Lab

Face Recognition/identification

Problem statement:

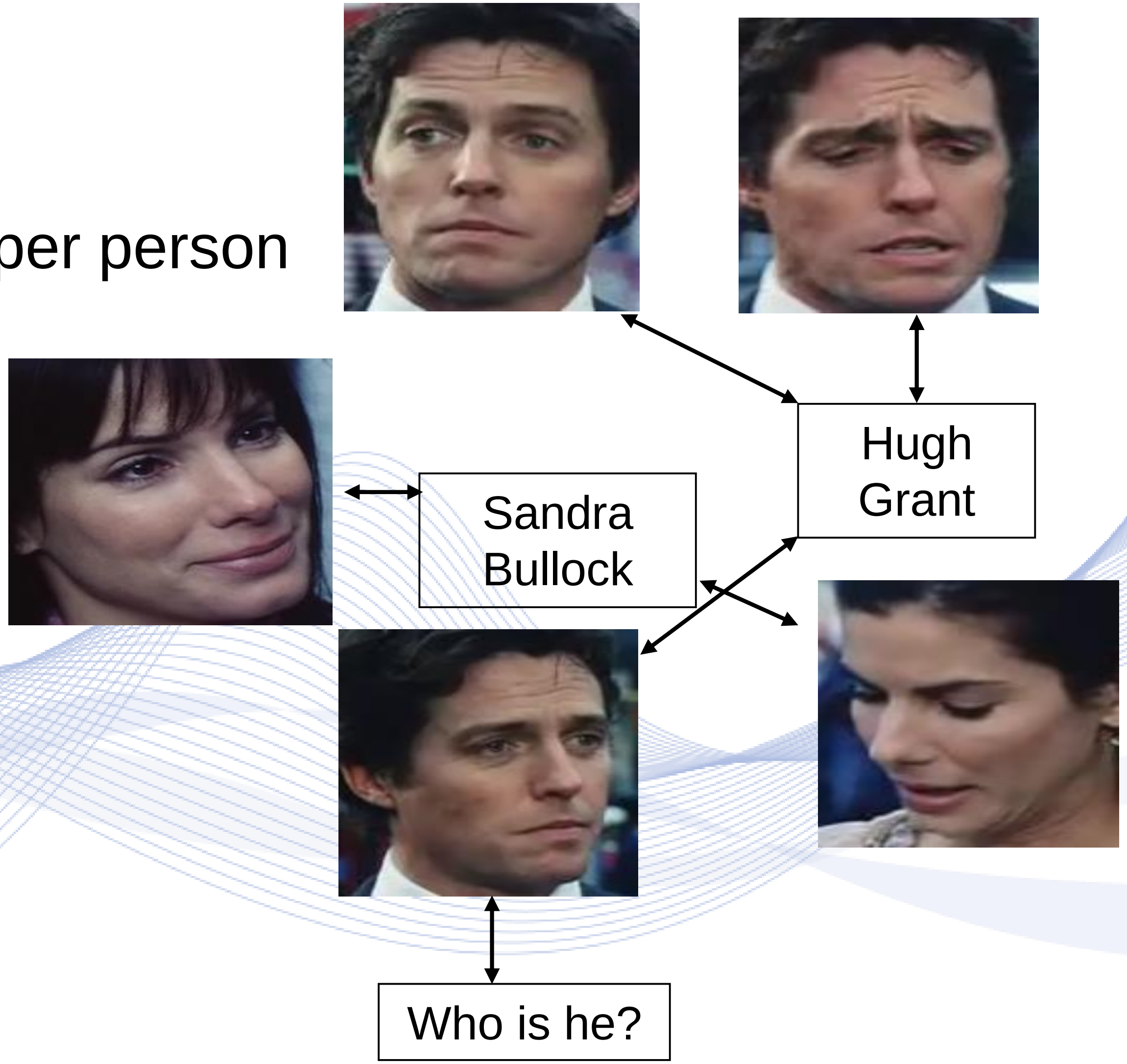
- To identify a face identity
- Input for training: several facial ROIs per person
- Input for inference: a facial ROI
- Inference output: the face id

- Supervised learning
- Applications:

  Biometrics

  Surveillance applications

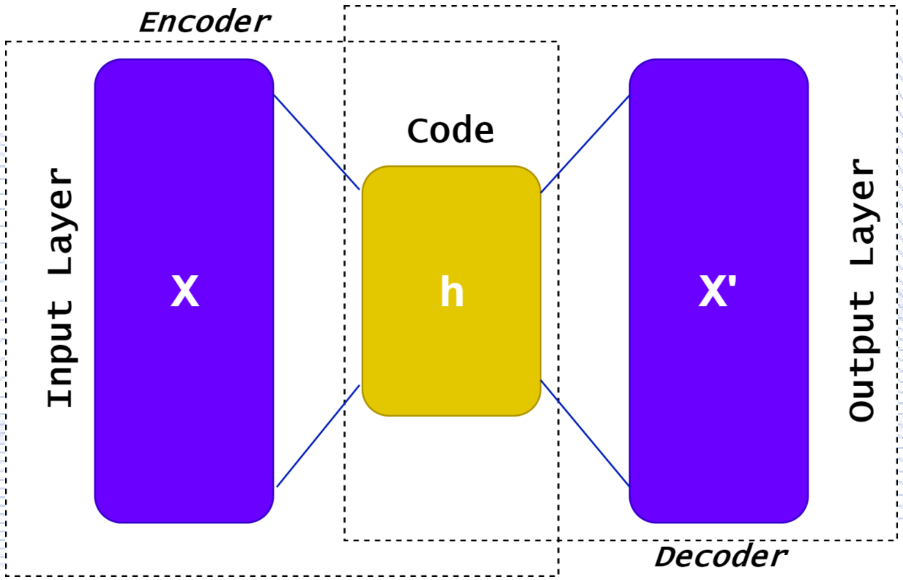
  Video analytics



Autoencoders

Given a sample $\mathbf{x} \in \mathbb{R}^n$ and a function $\mathbf{y} = \boldsymbol{f}(\mathbf{x}; \boldsymbol{\theta})$, the model output $\mathbf{y}$ should be equal to the model input $\mathbf{x}$:

- **Training**: Given N pairs of training examples $\mathcal{D} = \{\mathbf{x}_i, i = 1, \dots, N\}$, where $\mathbf{x}_i = \mathbf{y}_i \in \mathbb{R}^n$, estimate $\boldsymbol{\theta}$ by minimizing a loss function: $\min_{\boldsymbol{\theta}} J(\mathbf{x}, \hat{\mathbf{y}})$.

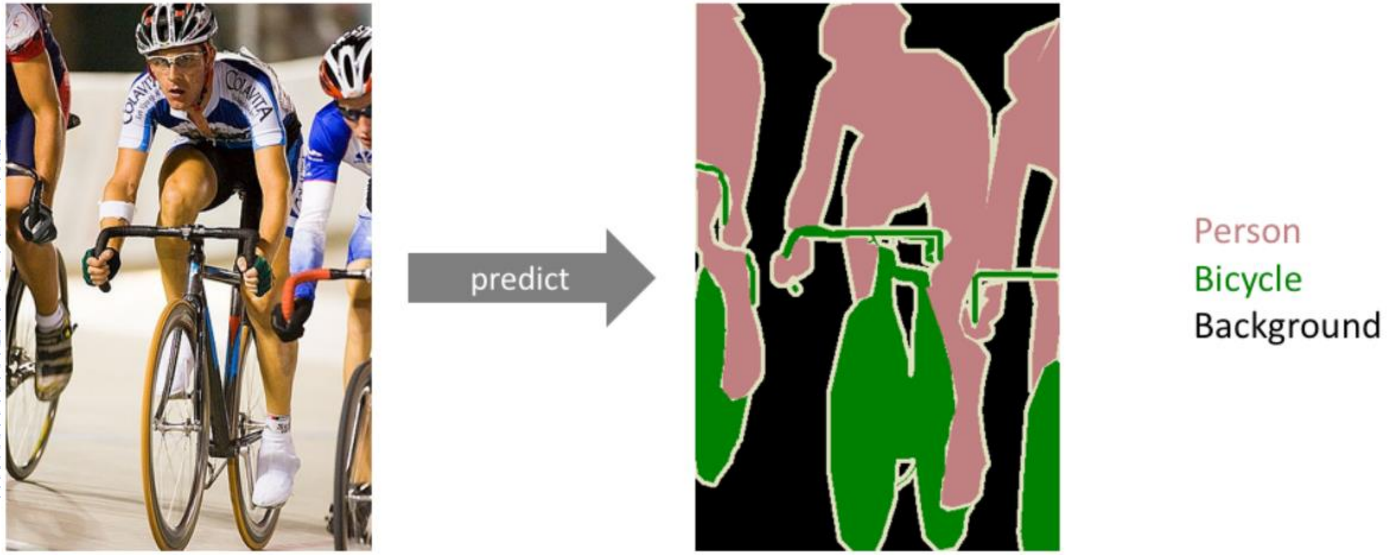


Autoencoder structure.

Image segmentation

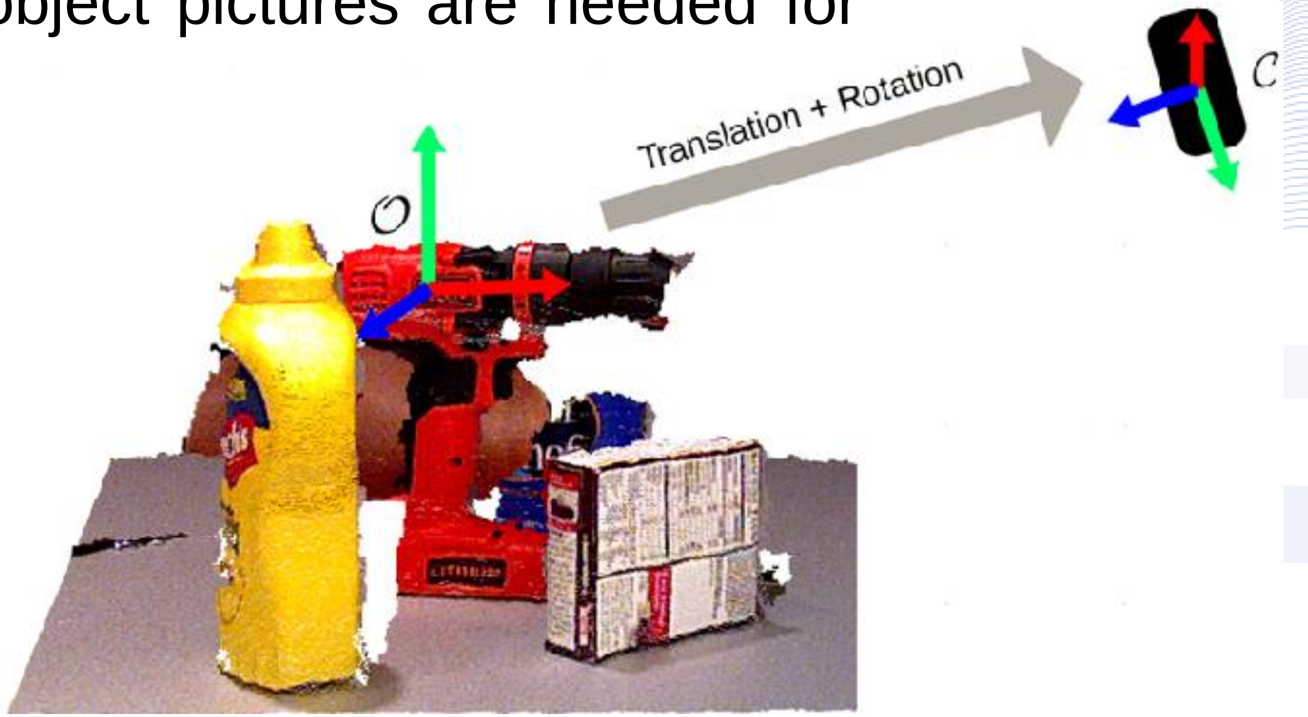
Given a region class label set $\mathcal{C} = \{\mathcal{C}_i, i = 1, \dots, m\}$, an image $\mathbf{x} \in \mathbb{R}^n$ must be segmented in m regions resulting in a segmentation map $\mathbf{y} \in \mathbb{R}^{n \times m}$.

- the ML model $\hat{\mathbf{y}} = \boldsymbol{f}(\mathbf{x}; \boldsymbol{\theta})$ predicts a segmentation map $\hat{\mathbf{y}} \in \mathbb{R}^{n \times m}$, where a class label vector $\hat{\mathbf{y}}_j \in \mathbb{R}^m$ is assigned to each image pixel $j = 1, \dots, n$ of the input image sample $\mathbf{x}$ minimizing the error $\min_{\boldsymbol{\theta}} J(\mathbf{y}, \hat{\mathbf{y}})$.
- Pixel-level classification.



6D object pose regression

- ***Object pose***: 3 3D object translation and 3 3D rotation parameters vs camera coordinate system.
- A ML model receives the object image and directly regresses its pose.
- Only a set of pose-annotated object pictures are needed for ML model training.



Multi-task Machine Learning

- The same ML model $\mathbf{y} = \boldsymbol{f}(\mathbf{x}; \boldsymbol{\theta})$ is optimized to learn performing multiple tasks, e.g.:
  - Object recognition
  - Region-of-Interest (bounding box) regression
  - Region segmentation
  - Depth regression.
- Output: $\mathbf{y} = [\mathbf{y}_1^T | \ldots | \mathbf{y}_M^T]^T$ for M different tasks.

- Optimization of a joint cost function:

$$\min_{\boldsymbol{\theta}} J(\mathbf{y}, \hat{\mathbf{y}}) = \alpha_1 J_1(\mathbf{y}, \hat{\mathbf{y}}) + \cdots + \alpha_M J_M(\mathbf{y}, \hat{\mathbf{y}}).$$

Object Detection

- Object detection = classification + localization:
- Find **what** is in a picture as well as **where** it is.

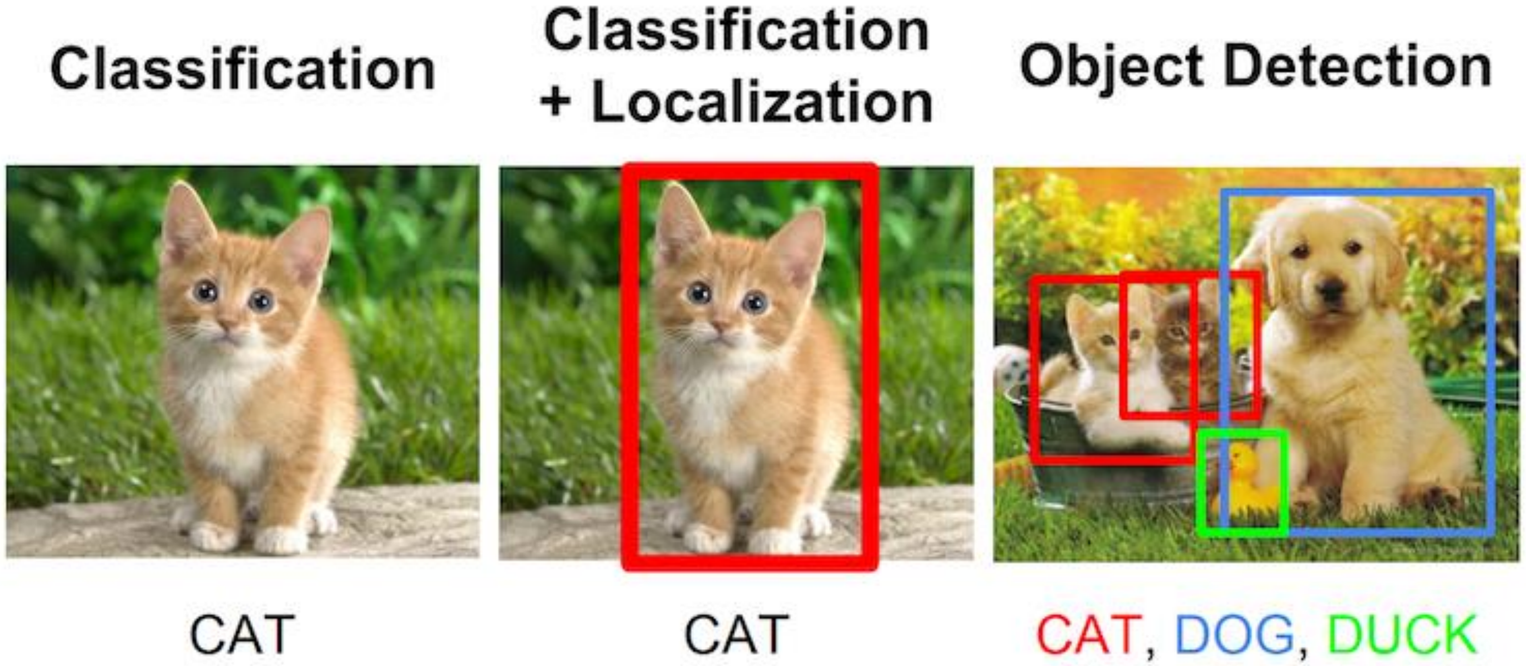


Figure: http://cs231n.stanford.edu/slides/2016/winter1516_lecture8.pdf

Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- **Unsupervised learning**
  - **Clustering**
  - **Dimensionality reduction, data retrieval**
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Unsupervised Learning

- In ***unsupervised learning,*** the ML model is provided with samples containing exclusively input feature vectors, without neither labels nor any information about the specific desired output:

$$\mathcal{D} = \{\mathbf{x}_i, i = 1,2, \dots, N\}$$

- $\mathbf{x} \in \mathbb{R}^n$.
- Unsupervised learning-based models are used for discovering the underlying structure of the data.

Clustering

- **Input:** A predefined number of clusters $\mathcal{C} = \{\mathcal{C}_i, i = 1, 2, \dots, m\}$ and a set of unlabeled samples $\mathcal{D} = \{\mathbf{x}_i, i = 1, 2, \dots, N\}$ $\mathbf{x}_i \in \mathbb{R}^n$.
  - Number of clusters m may be unknown.
- **Output:** Sample set $\mathcal{D} = \{\mathbf{x}_i, i = 1, 2, \dots, N\}$ partition to m clusters $\mathcal{C}_i, i = 1, \dots, m$
  - Cluster samples are similar and dissimilar to the samples of other clusters based on similarity/distance metric $\| \cdot \|$.
- Basically, clustering involves unlabeled data according to feature similarities.

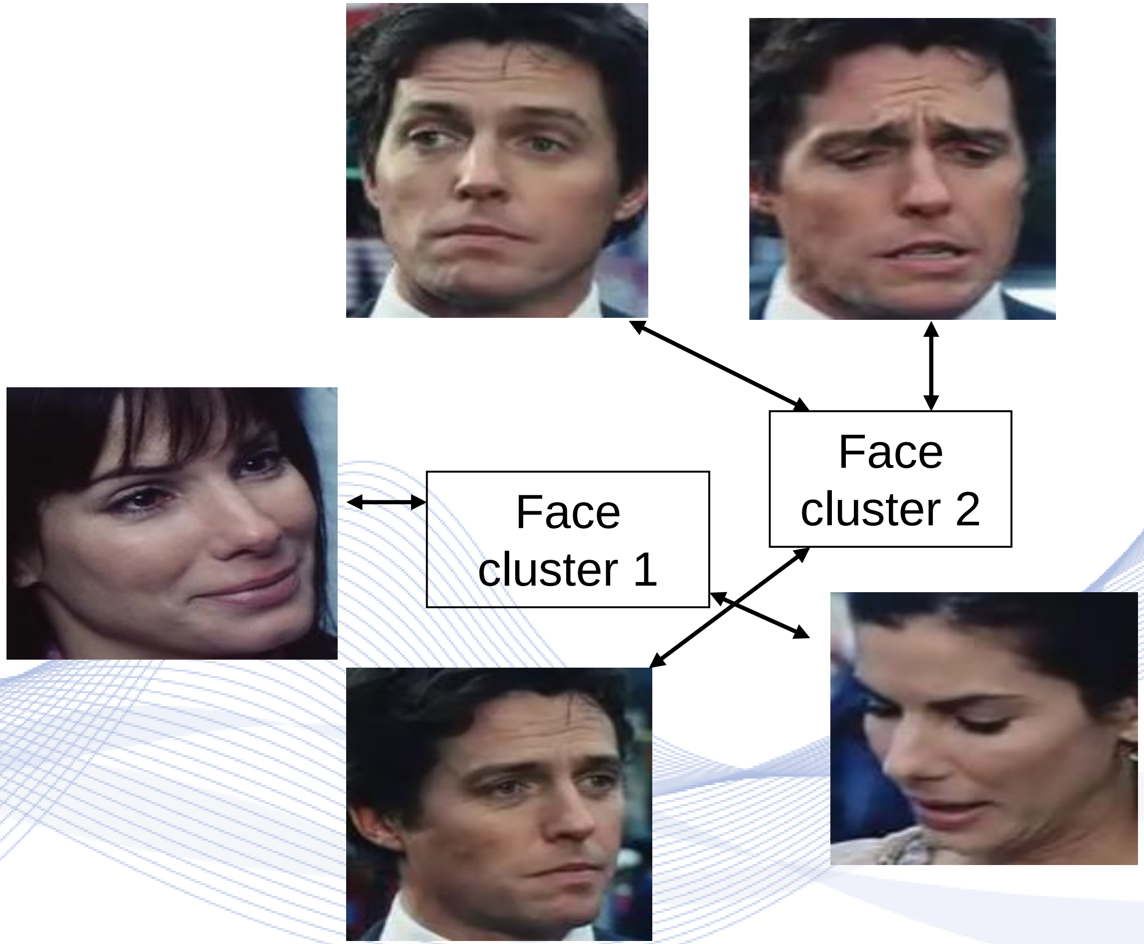
Face clustering

Problem statement:

- To cluster facial images
- Input: many facial ROIs
- Output: facial image clusters

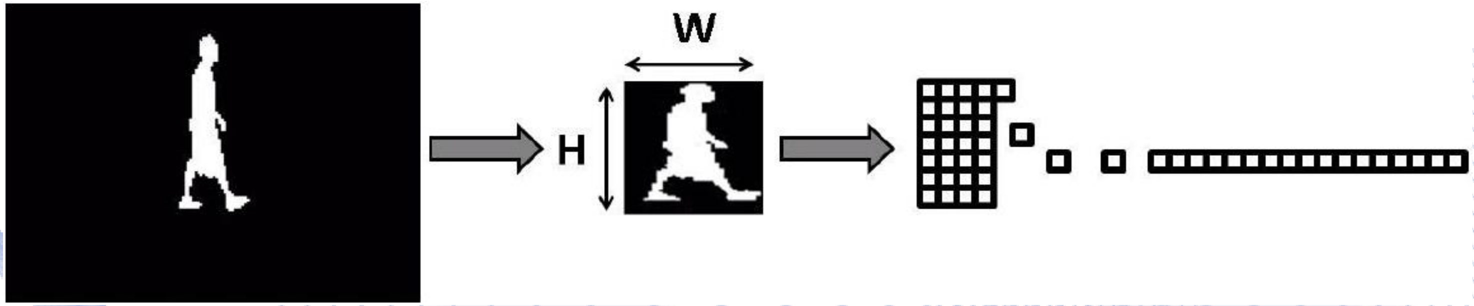
- Unsupervised learning
- Applications:

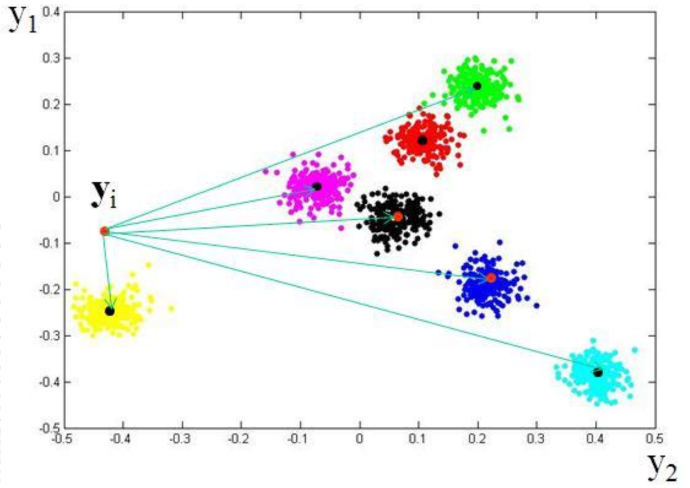
  Biometrics
  Surveillance applications
  Video analytics



Dimensionality Reduction

- Example: Human posture visualization.
- Dimensionality reduction from $\mathbf{p} \in \mathbb{R}^{HW}$ to $\mathbf{y} \in \mathbb{R}^2$





Binary human body image

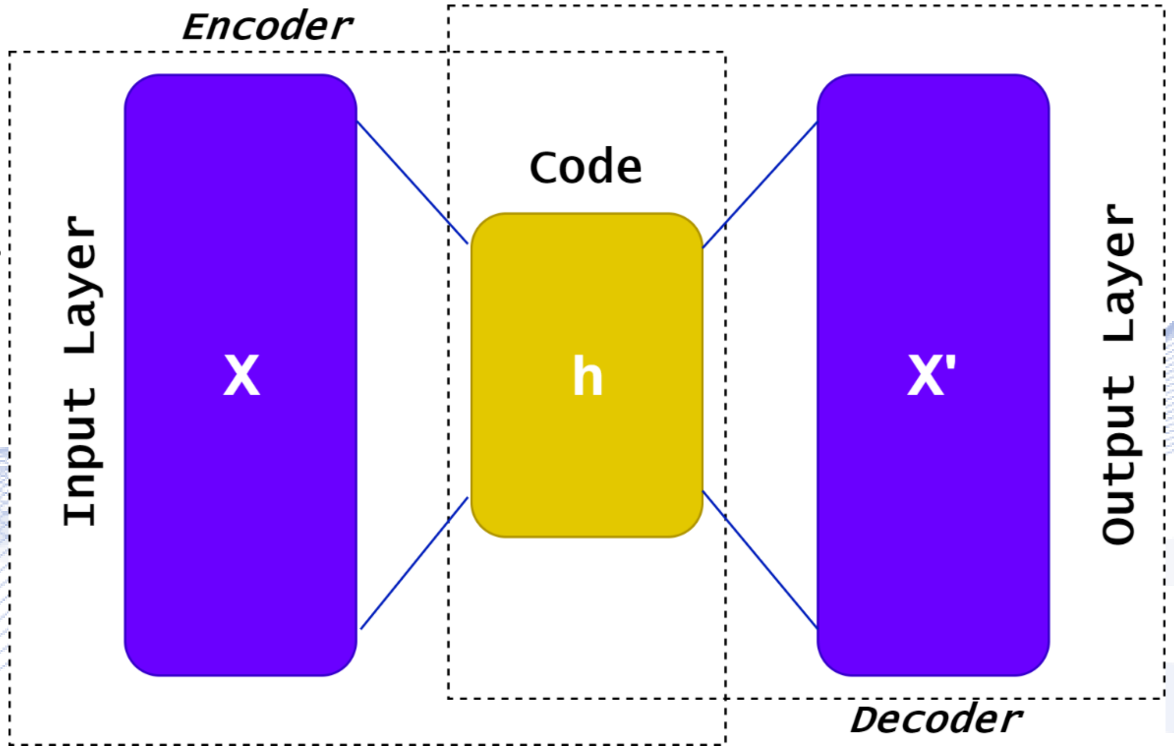
Posture image of fixed size

Posture vector $\mathbf{p} \in \mathbb{R}^{HW}$

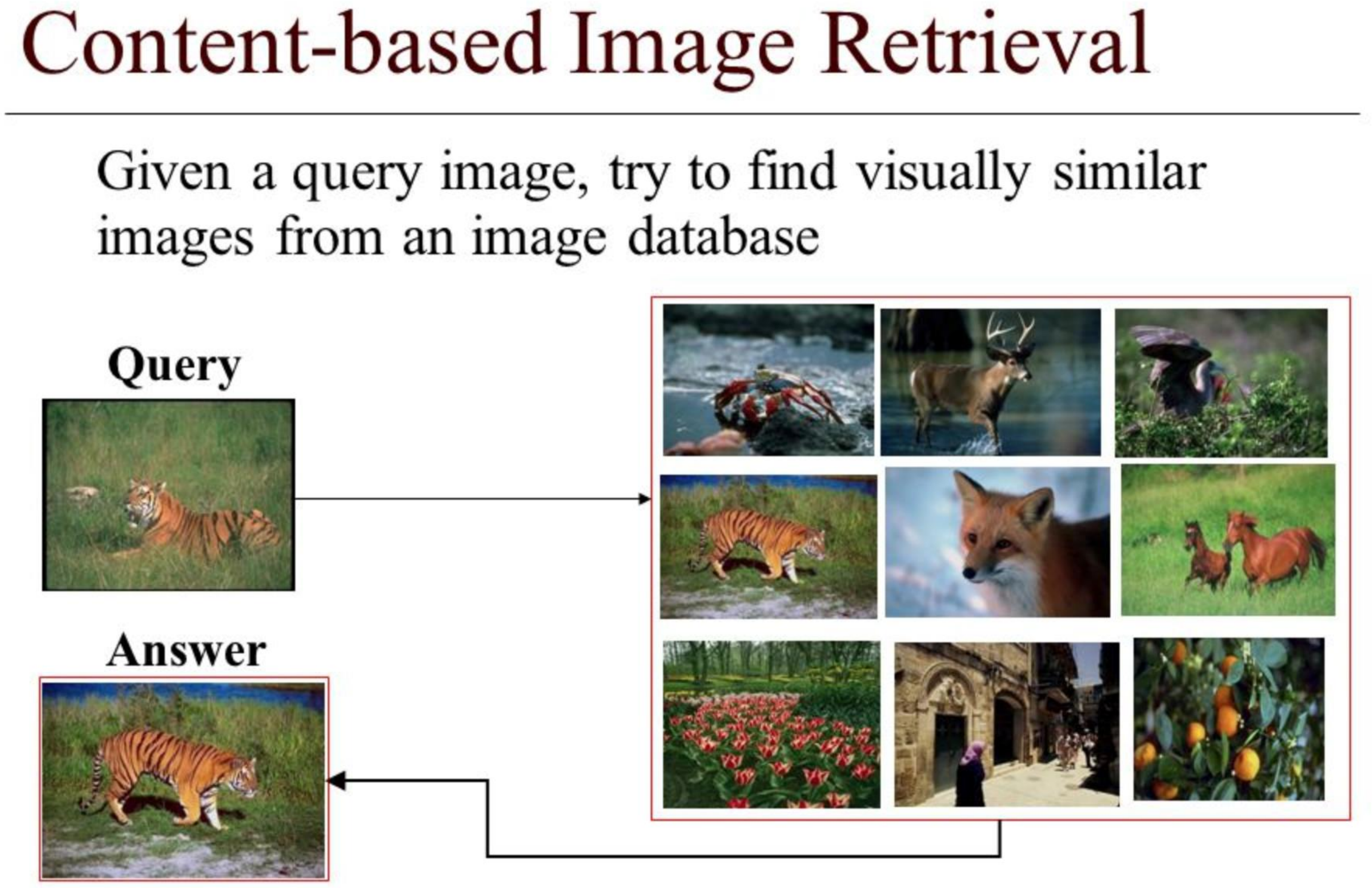
Posture visualization $\mathbf{y} \in \mathbb{R}^2$

Dimensionality Reduction

- Multidimensional scaling.
- Principal Component Analysis.
- Linear Discriminant analysis.
- Independent Component Analysis.
- Autoencoders.



Data Retrieval



http://slideplayer.com/slide/3415344/12/images/6/Content-based+Image+Retrieval.jpg

Person re-identification

Definition

- Refers to the problem of associating/matching images of the same person taken:
  - from different cameras or
  - from the same camera in different occasions (e.g., night day)
- It can be solved as a data retrieval problem.

Example

Artificial Intelligence & Information Analysis Lab

Introduction to Machine Learning

- Supervised learning
    - Classification/recognition/identification, Identity verification
    - Regression, Object detection
- Unsupervised learning
    - Clustering
    - Dimensionality reduction, data retrieval
- **Self-supervised learning**
- Semi-supervised learning
    - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Self-Supervised Learning

- ***Self-supervised learning*** resembles supervised learning.
- It relies on pairs of input-outputs, $(\mathbf{x}_i, \mathbf{y}_i)$ for ML model training.
- However, it does not require an explicit form of target labels $\mathbf{y}_i$.

- Instead, the necessary supervisory information is extracted from the input feature structure and correlations.

Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- **Semi-supervised learning**
  - **Label propagation**
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Semi-Supervised Learning

Semi-supervised learning:

- Combination of supervised and unsupervised learning.
- It relies on the existence of a large amount of training data, whose minority contains output information (data labels).
- Training dataset $\mathcal{D}$ consists of:
  - a set of N_1 labeled training examples, $\mathcal{D}_1 = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, N_1\}$.
  - a set of N_2 unlabeled examples, $\mathcal{D}_2 = \{\mathbf{x}_i, i = 1, \dots, N_2\}$., where $N_1 \ll N_2$:

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2.$$

- It is particularly useful for exploiting data structure (geometry) information.

Facial label propagation

Problem statement:

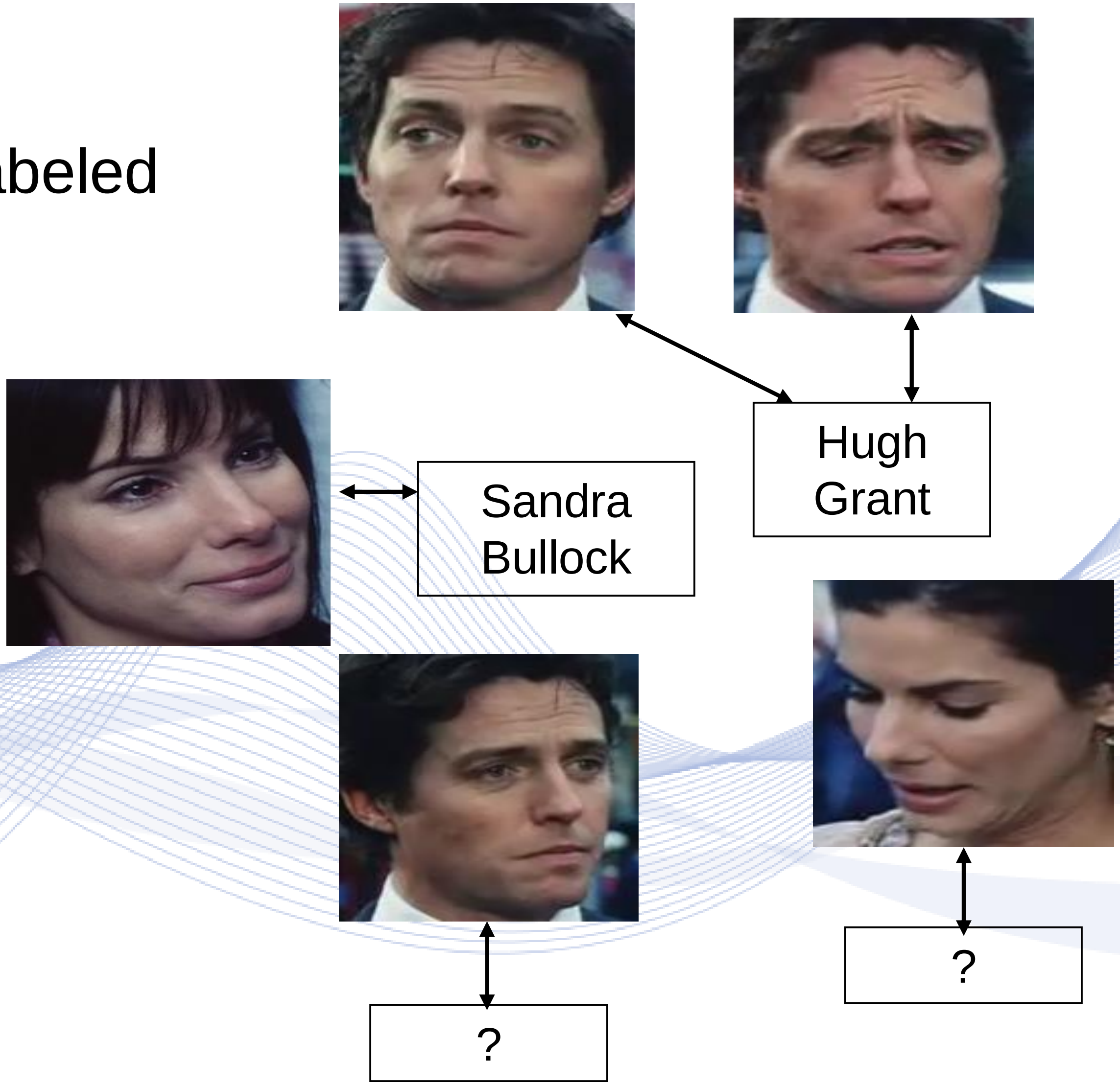
- To transfer labels from labeled to unlabeled facial images
- Input: a) labeled facial ROIs,
  b) unlabeled facial ROIs
- Output: facial image labels

- Semi-supervised learning
- Applications:

  Biometrics

  Surveillance applications

  Video analytics

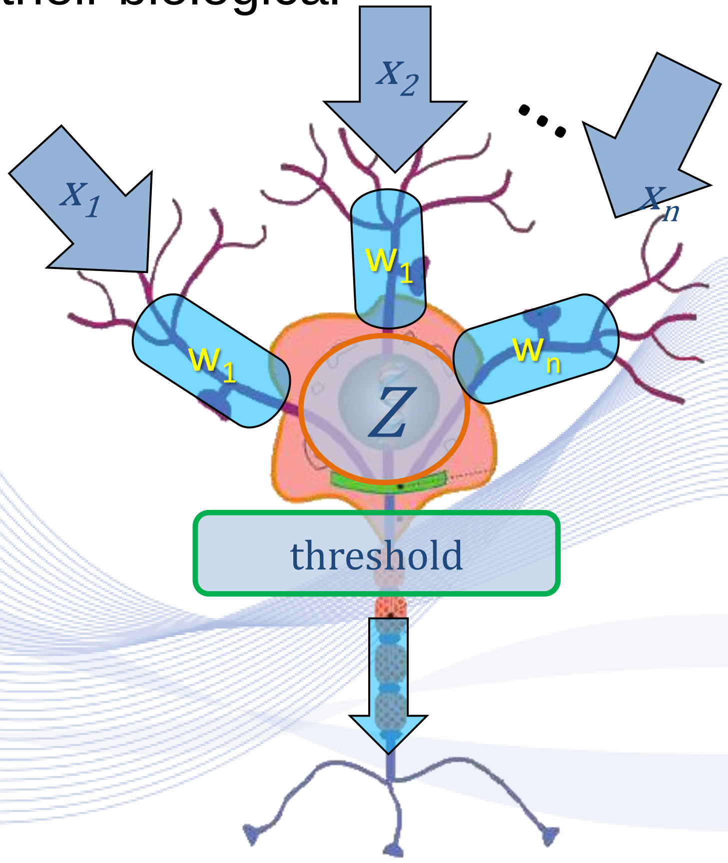


Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- **Artificial Neural Networks**
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Artificial Neural Networks

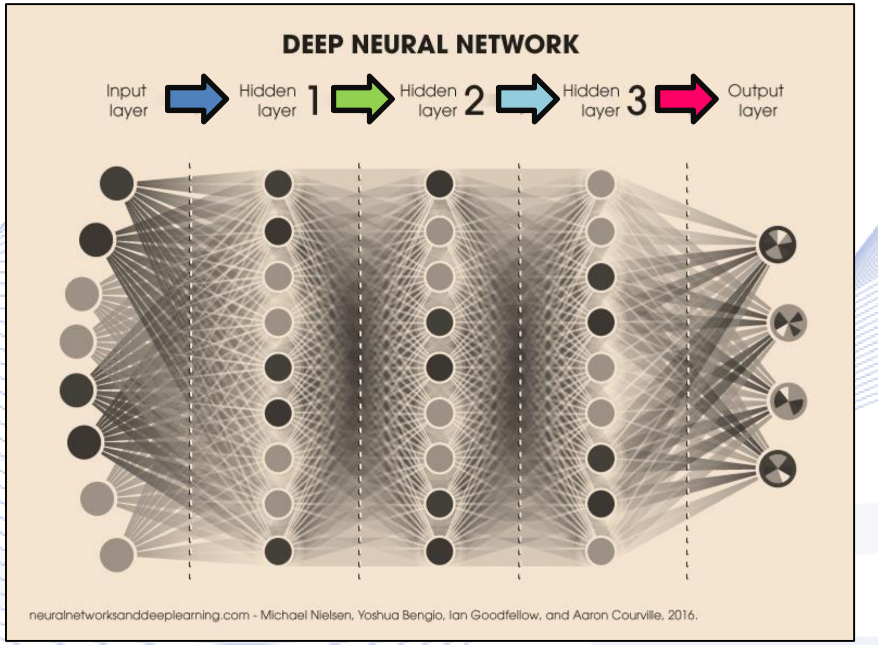
- Artificial neurons are mathematical models loosely inspired by their biological counterparts.
- Incoming signals: $\mathbf{x} = [x_1, x_2, \dots, x_n]^T,\ x_i \in \mathbb{R}$.
- Synaptic weights: $\mathbf{w} = [w_1, w_2, \dots, w_n]^T ,\quad w_i \in \mathbb{R}$.
- Synaptic integration: $Z = \sum_{i=1}^{N} w_i x_i = \mathbf{w}^T \mathbf{x}$.
- Output nonlinearity.
- ANNs have a layered structure:
  - Each layer consists of artificial neurons.
  - They learn a function $\hat{\mathbf{y}} = \boldsymbol{f}(\mathbf{x}; \boldsymbol{\theta})$ during training.



Deep Neural Networks

Definition

- *Deep Neural Networks (DNNs)* have a count of layers (depth) $L \geq 3$.
  - There are multiple hidden layers with regard to the MLP reference model.
  - Typically, first layers are convolutional, latter ones are fully connected (CNNs).



Deep Neural Network with $L = 4$

Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- Artificial Neural Networks
- **Adversarial Machine Learning**
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Adversarial Machine Learning

Adversarial machine learning:

- Given a class label set $\mathcal{C} = \{\mathcal{C}_i, i = 1, \dots, m\}$ and a trained ML model $\hat{\mathbf{y}} = \boldsymbol{f}(\mathbf{x}; \boldsymbol{\theta}), \hat{\mathbf{y}} \in [0,1]^m$
- find a perturbation $\mathbf{p}$, so that a perturbed test sample instance $\mathbf{x}_p = \mathbf{x} + \mathbf{p}$ (*adversarial sample*) is wrongly classified by the ML model as: $\hat{\mathbf{y}}_p = \boldsymbol{f}(\mathbf{x}_p; \boldsymbol{\theta})$, where $\hat{\mathbf{y}}_p \neq \hat{\mathbf{y}}$.

- *ML training set augmentation*: during the training process apart from using real samples $\mathbf{x}_i, i = 1, \dots, N$ in the training set, we also include their perturbed instances $\mathbf{x}_{p_i}$, so that both $\mathbf{x}_i$ and $\mathbf{x}_{p_i}$ are correctly classified.
- Adversarial training works as a regularization technique, in order to derive a more robust ML model.

Introduction to Machine Learning

- Supervised learning
    - Classification/recognition/identification, Identity verification
    - Regression, Object detection
- Unsupervised learning
    - Clustering
    - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
    - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- **Generative Machine Learning**
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- **Temporal Machine learning (RNN)**
- Continual Learning
- Reinforcement Learning
- Adaptive learning

Recurrent Neural Networks

- An RNN typically processes temporal information:
  - signals/ time sequences.
- It consists of recurrent neurons.
- A recurrent neuron takes into consideration the stored information from the past inputs(hidden state).

$\mathbf{x}_t$: input instance.

$\mathbf{h}_{t-1}$: hidden state.

φ : activation function.

$\hat{\mathbf{y}}_t$: the output.

t: is representing the time.

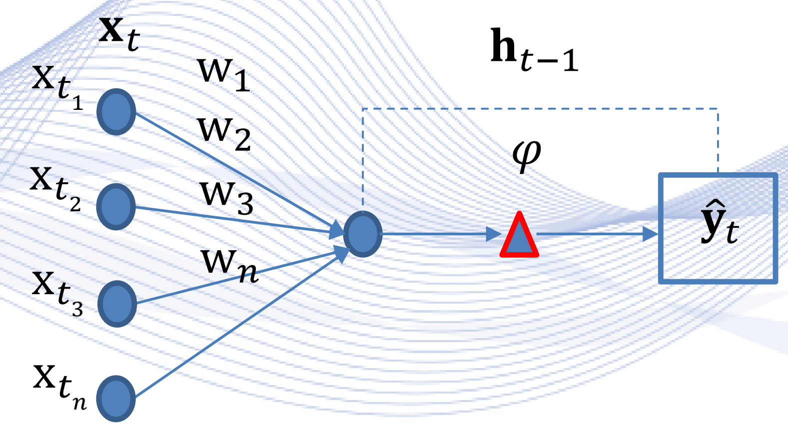


Fig.5 Recurrent artificial neuron

Artificial Intelligence & Information Analysis Lab

Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- **Continual Learning**
- Reinforcement Learning
- Adaptive learning

Continual Learning

- ***Continual learning (Incremental Learning, Life-long Learning)***:
  - The training example set $\mathcal{D}_t = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1,2, \dots, N\}$ changes over time t
    - with the addition of new samples
    - deletion of some old samples.
  - The ML model is incrementally trained (NOT from scratch);
  - The learning takes place, whenever new examples emerge;
  - It adjusts what has been learned according to the new examples;
  - It does not assume the availability of a sufficient training set, before the learning process starts.

- Catastrophic forgetting.

Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- **Reinforcement Learning**
- Adaptive learning

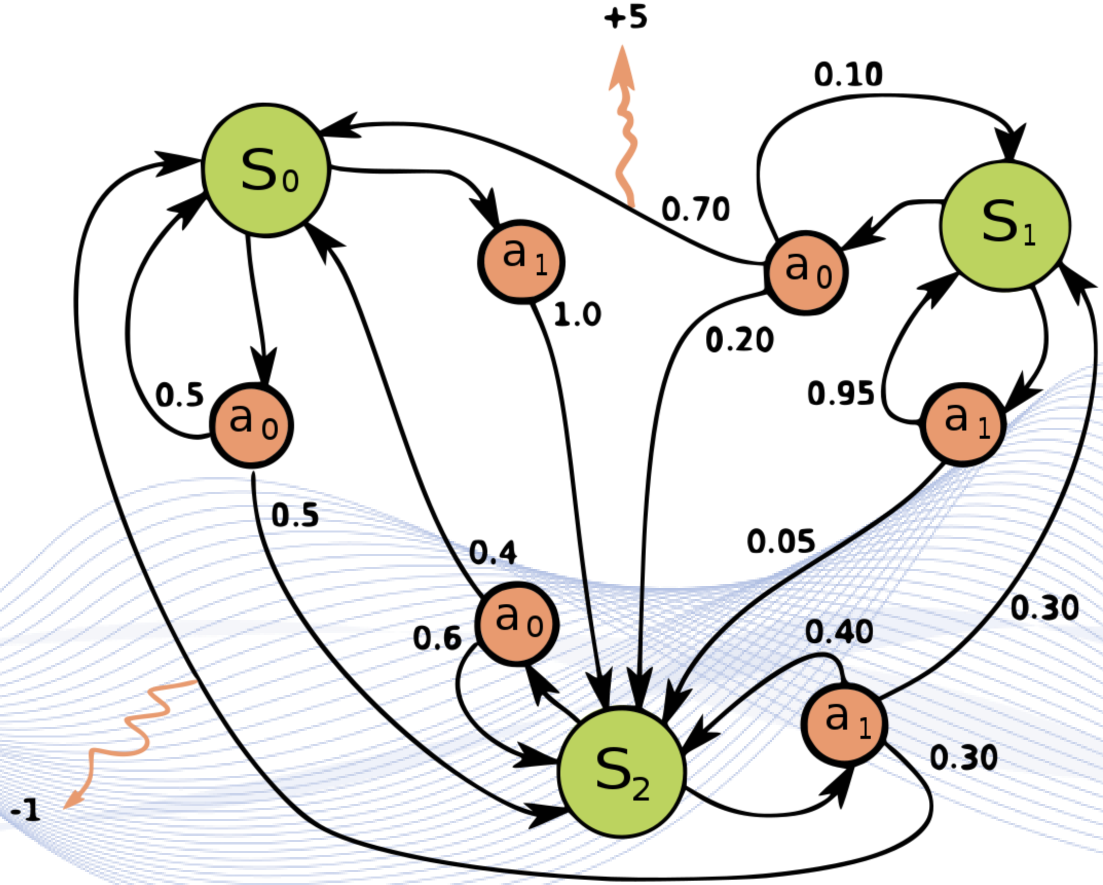
Reinforcement Learning

- ***Reinforcement Learning***: interaction scheme between an ML agent and his environment, in order to maximize some notion of cumulative rewards.

- Given a finite set of states $\mathcal{S} = \{s_i, i = 1,2, \dots, N_s\}$, a finite set of actions $\mathcal{A} = \{a_i, i = 1,2, \dots, N_a\}$, a reward function $R_a(s_i, s_j)$ and a probability function $P_a(s_j, r | s_i, a)$, where r is a reward, the goal of an RL model is to find a policy that maximizes a cumulative reward signal.

- ***Experience replay:*** Online reinforcement learning, based on remembering and reusing past experiences.

Reinforcement Learning



Introduction to Machine Learning

- Supervised learning
  - Classification/recognition/identification, Identity verification
  - Regression, Object detection
- Unsupervised learning
  - Clustering
  - Dimensionality reduction, data retrieval
- Self-supervised learning
- Semi-supervised learning
  - Label propagation
- Artificial Neural Networks
- Adversarial Machine Learning
- Generative Machine Learning
- Temporal Machine learning (RNN)
- Continual Learning
- Reinforcement Learning
- **Adaptive learning**

Adaptive learning

- Knowledge Distillation:
  - The input/output pairs of a trained teacher ML model (typically large and heavyweight) are employed for training a student ML model (typically smaller and initially untrained).
- Domain adaptation
  - Adaptation of an ML model trained on one task-specific source domain (dataset) to a different target domain (dataset).
  - The data of the two domains typically follow different pdfs.
  - The model/data are adapted, so that task-specific knowledge is maintained in the different domains.
- Transfer learning
  - An already pre-trained ML model is re-trained using new data to improve performance in the new (and old) domain/task of interest.

Adaptive learning

- Bio-inspired learning:
  - Bio-inspiration for fundamental learning mechanisms, e.g., based on memory or synaptic plasticity.
- Curiosity-driven learning:
  - Identification of important information to incorporate new knowledge and reduce uncertainty.
- Activation Pattern Analysis
  - Determining ML model behavior/response on novel test data.

Adaptive learning

- Federated learning/Collaborative learning
  - Decentralized ML model training across multiple nodes with local data samples only, without data exchange across nodes.

- Ensemble Learning
  - The analysis results from multiple different DNN models are weighed and combined to reach a more accurate aggregate prediction.

References

[BIS2006] C.M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[GOD2016] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016

[THE2003] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Elsevier, 2003.

Q & A

Thank you very much for your attention!

Contact: Prof. I. Pitas
pitas@csd.auth.gr