Machine Learning
for Media Data Analysis

Alexandros Iosifidis
Assistant Professor
Machine Learning and Computer Vision
Aarhus University, Department of Engineering
alexandros.iosifidis@eng.au.dk
Overview

Media Data Analysis applications
- Image analysis/recognition/segmentation
- Video analysis/segmentation
- Face recognition/verification and affective computing
- Human action recognition/localization/segmentation
- Domain-specific image/video analysis

Recent contributions
- Graph-based analysis/recognition/clustering
- Max-margin Classification
- Discriminant Learning
- Kernel-based learning
- Multi-view/modal Data Analysis
- Neural Network (Deep Learning) acceleration
- Data-driven (Deep) Architecture Learning

Extension to applications of other domains
Media Data Analysis applications
Human Face Analysis

Face recognition

Human Face Analysis

Face verification  
Face identification

Human Face Analysis

Affective Computing (Facial expression recognition)

Human Face Analysis

Visual Voice Activity Detection → Assign the correct face to the observed voice
We were the first to perform VVAD in the wild!

Human Face Analysis

Face – Sketch recognition

G. Cao, A. Iosifidis and M. Gabbouj, "Multi-modal Subspace Learning with Dropout regularization for Cross-modal Recognition and Retrieval ", IPTA 2016 (Best Student Paper Award)
Image Analysis

Salient object segmentation $\rightarrow$ Unsupervised (generic) case

C. Aytekin, A. Iosifidis and M. Gabbouj, "Probabilistic Saliency Estimation", Pattern Recognition, 2018
Image Analysis

Salient object segmentation $\rightarrow$ Supervised (User-directed case)

C. Aytekin, A. Iosifidis, S. Kiranyaz and M. Gabbouj, “Learning Graph Affinities for Spectral Graph-based Salient Object Detection”, Pattern Recognition, 2017
Human Action Recognition

How to combine action observations from various views/cameras
Restricted application scenario → Movie production

A. Iosifidis, A. Tefas and I. Pitas, "View-invariant action recognition based on Artificial Neural Networks", IEEE Transactions on Neural Networks and Learning Systems, 2012
Human Action Recognition

How to combine action observations from various views/cameras

Restricted application scenario → Movie production

Localization of people and classification of their actions

A. Iosifidis, A. Tefas and I. Pitas, "View-invariant action recognition based on Artificial Neural Networks", IEEE Transactions on Neural Networks and Learning Systems, 2012

Human Action Recognition

Human action recognition for assisted living of the elderly

Human Action Recognition

What can we do for complex actions in complex/cluttered scenes?
Human Action Recognition

What can we do for complex actions in complex/cluttered scenes?

We focus on Space-Time Interest Points (STIPs) and follow the similar approaches:

A. Iosifidis, A. Tefas and I. Pitas, "Distance-based Human Action Recognition using optimized class representations", Neurocomputing, 2015
Human Action Recognition

When enriched visual information is available → Stereo Cameras

Human Action Recognition

When enriched visual information is available → Stereo Cameras

Original DT-based description

Disparity-enhanced DT-based description

Video analysis

Object tracking

Video analysis

Object detection/recognition
Scene analysis

Multi-view/modal Data analysis

Image and Text (I2T and T2I) Retrieval

<table>
<thead>
<tr>
<th>Image Query</th>
<th>Text Query</th>
</tr>
</thead>
</table>
| ![Image](image1.jpg) | 1. A very big building with many windows and a clock on it.  
2. A very old tall building with a large clock tower sticking out of it.  
3. The clock tower stands high above the city.  
4. A clock that is on the side of a large building.  
5. The bridge is in front of a huge building with a clock tower in the middle of it. |
| Precision: 53.33% | Precision: 86.67% | Precision: 100% |
| (a) Query by original image feature | (b) Query by projected image feature | (c) Query by text |

<table>
<thead>
<tr>
<th>Image Query</th>
<th>Text Query</th>
</tr>
</thead>
</table>
| ![Image](image2.jpg) | 1. An open laptop sits on a desk in front of a window.  
2. An Apple laptop sitting on a wooden desk.  
3. An Apple laptop sitting on a wooden desk in an office.  
4. An Apple laptop on a desk in an office.  
5. A desk with a laptop sitting on top of it. |
| Precision: 60.00% | Precision: 86.67% | Precision: 66.67% |
| (a) Query by original image feature | (b) Query by projected image feature | (c) Query by text |

Domain-specific Media Data Analysis

Classification of Aquatic Macroinvertebrates (bugs in lakes)


Recent Contributions
**Discriminant Learning**

**Class-specific kernel Discriminant Analysis**
- Data represented as vectors
- Vector to vector transformation/mapping
- Find a data mapping that maximizes discrimination of the class of interest from the rest of the world

**Intra-class scatter** $S_i$

**Out-of-class scatter** $S_p$
Discriminant Learning

➢ Traditional CSKDA:
1. Data mapping to the feature space:

\[ x_i \in \mathbb{R}^D \rightarrow \phi(x_i) \in \mathcal{F} \]

2. Application of the linear projection in

\[
S_i = \Phi L_i \Phi^T \quad \quad \quad \quad S_p = \Phi L_p \Phi^T
\]

\[
W^T S_i W = A^T \Phi^T \Phi L_i \Phi^T \Phi A = A^T K L_i K A
\]

\[
W^T S_p W = A^T \Phi^T \Phi L_p \Phi^T \Phi A = A^T K L_p K A
\]

➢ \( A \) is calculated by applying eigenanalysis to the matrix

\((K L_p K)^{-1}(K L_i K)\) \(\leftarrow\) stability issues!
Discriminant Learning

We showed that the standard CSKDA solution is equivalent to:

\[
\hat{J} = \|W^T\Phi - T\|^2_F, \quad s.t.: \text{rank}(W) \leq d.
\]

\[
\hat{J} = \|B^T(Q^T\Phi) - T\|^2_F = \|B^T(A^TK) - T\|^2_F
\]

using data-independent targets $T$!

Benefits

› Much stable and fast solutions can be calculated
› Approximation schemes are readily available and extremely efficient/effective
› Incremental/Decremental solutions are available
› Hierarchical (deep) models for class-specific learning are now possible

Discriminant Learning

Deep Class-specific Discriminant Analysis model:

G. Cao, A. Iosifidis and M. Gabbouj, “Neural Class-Specific Projections for Face Verification”, IET Biometrics, 2017
Classification

Max-margin based classification
› Find the decision hyperplane discriminating the classes with maximum margin
› Theoretical guarantees for generalization error
› One (global) solution
Classification

➢ Binary classification:

\[
\min_{\mathbf{w},b} \frac{1}{2} \mathbf{w}^T \mathbf{w} + c \sum_{i=1}^{N} \xi_i,
\]

s.t.:

\[
y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, N
\]

\(y_i \in \{-1,1\}\) are the binary labels.

➢ Multi-class classification:

\[
\min_{\mathbf{w}_k, b_k} \sum_{k=1}^{K} \frac{1}{2} \mathbf{w}_k^T \mathbf{w}_k + c \sum_{i=1}^{N} \sum_{k \neq l_i} \xi_i^k
\]

s.t.:

\[
\mathbf{w}_{l_i}^T \mathbf{x}_i + b_{l_i} \geq \mathbf{w}_k^T \mathbf{x}_i + b_k + 2 - \xi_i^k, \quad \xi_i^k \geq 0, \quad i = 1, \ldots, N, \quad k \neq l_i.
\]
Classification

In order to increase class discrimination

› Discriminant data mapping
› Max-margin classification
Classification

**Discriminant max-margin based classification**

› We proved that these two processing steps can be applied at once!
› Max-margin to a discriminant (kernel) space

➢ Multi-class classification:

\[
\text{min} \sum_{k=1}^{K} \frac{1}{2} w_k^T w_k + c \sum_{i=1}^{N} \sum_{k \neq l_i} \xi_i^k + \sum_{k=1}^{K} \frac{\lambda}{2} w_k^T Sw_k
\]

s.t.:

\[
w_{l_i}^T x_i + b_{l_i} \geq w_k^T x_i + b_k + 2 - \xi_i^k, \quad \xi_i^k \geq 0, \quad i = 1, \ldots, N, \quad k \neq l_i
\]

➢ Joint optimization of \( K \) decision functions \( \{w_k, b_k\} \), \( k=1, \ldots, K \).

Probability-based Visual Saliency

We modeled visual saliency (for salient object segmentation) using probabilistic

\[
\arg\min_{P(x)} \left( \sum_i (P(x = x_i))^2 v_i + \frac{1}{2} \sum_{i,j} \left( P(x = x_i) - P(x = x_j) \right)^2 w_{i,j} \right)
\]

\[
\text{s.t. } \sum_i P(x = x_i) = 1.
\]

Properties

› Global optimum solution
› Generic framework for visual saliency:
   › Diffusion methods are special cases of PSE
   › PSE optimally refines the solution of QCut (SoA saliency estimation method)
› Now Saliency Estimation can be modelled as an One-Class Classification Problem

C. Aytekin, A. Iosifidis and M. Gabbouj, “Probabilistic Saliency Estimation”, Pattern Recognition, 2018
Supervised Visual Saliency

End-to-end learning for visual saliency that:
› Exploits successive learnable feature transformations
› Optimizes the end-to-end model using global information

C. Aytekin, A. Iosifidis, S. Kiranyaz and M. Gabbouj, “Learning Graph Affinities for Spectral Graph-based Salient Object Detection”, Pattern Recognition, 2018
Multi-view/modal Data Analysis

Generalized Multi-view Embedding

We showed that most multi-view embedding methods can be modeled using the Rayleigh Quotient

\[ J = \arg \max_W \frac{\text{Tr}(W^T PW)}{\text{Tr}(W^T QW)} \]

Multi-view/modal Data Analysis

Generalized Multi-view Embedding

We showed that most multi-view embedding methods can be modeled using the Rayleigh Quotient.

Based on this observation, we proposed a new MvLDA criterion incorporating inter-view and intra-view variance criteria.

\[
S_W = \sum_{i=1}^{V} \sum_{c=1}^{C} W_i^T X_i \left( I - \sum_{c=1}^{C} \frac{1}{N_c} e_c e_c^T \right) X_i^T W_i
\]

\[
S_W = \sum_{i=1}^{V} \sum_{c=1}^{C} W_i^T Q_{ii} W_i
\]

\[
S_B = \sum_{i=1}^{V} \sum_{j=1}^{V} \sum_{p=1}^{C} \sum_{q=1}^{C} (m_p^i - m_q^i)(m_p^j - m_q^j)^T
\]

\[
S_B = \sum_{i=1}^{V} \sum_{j=1}^{V} \sum_{p=1}^{C} \sum_{q=1}^{C} W_i^T X_i L_B X_j^T W_j
\]

Multi-view/modal Data Analysis

Generalized Multi-view Embedding

Neural Networks (Deep Learning)

Basic idea of neural network-based learning

- Data-driven fine-tuning the parameters of a network to regress inputs to targets using:
  - User-defined number of layers
  - User-defined activation functions (always the same for the entire network, except the last one)
  - User-defined neuron pooling operator (always the same for all neurons)
  - User-defined neuron nodal operator (always the same for all neurons)
Neural Networks (Deep Learning)

We proposed a Progressive Operational Feedforward Neural network learning approach

- Data-driven network’s architecture
- Data-driven network’s parameters tuning

Convolutional Neural Networks (Deep Learning)

**CNN architecture:**
- Convolutional layers
- Multilayer Perceptron (vector) layers
Convolutional Neural Networks (Deep Learning)

CNN architecture:
› Convolutional layers
› Multilayer Perceptron (vector) layers
Convolutional Neural Networks (Deep Learning)

**CNN architecture:**
- Convolutional layers
- Multilayer Perceptron (vector) layers

Real CNN architecture: CLs are tensors!
Convolutional Neural Networks (Deep Learning)

We proposed a tensor-based CNN filters modeling that can lead to:
› Lower number of network parameters (reduced memory footprint)
› Faster classification (reduced computational cost)

Convolutional Neural Networks (Deep Learning)

We proposed a tensor-based CNN filters modeling that can lead to:

› Lower number of network parameters (reduced memory footprint)
› Faster classification (reduced computational cost)

Extension to applications of other domains
Other applications

Stock prediction in financial markets


Other applications

University Ranking, Safety/hazard Assessment, Green Spaces, etc
Thank you for your attention!