Multimodal context analysis and prediction

Valeria Tomaselli (valeria.tomaselli@st.com)
Sebastiano Battiato
Giovanni Maria Farinella
Tiziana Rotondo (PhD student)
• Context analysis vs prediction
• Multimodality
• Objective and Challenges
• Some notions on Neural Networks
• State of the Art
• Our work (preliminary analysis)
• Conclusions and Future work
UNDERSTANDING

It is the first step to build intelligent machines that provide assistive support to people’s everyday life

You’ve an accident. I’m calling for help!

PREDICTION

It will enable many real-world applications:

• foreseeing abnormal situations before they happen;

• improving the way machines interact with humans

You’re having an accident. I’m slowing down your car!
Multi-modality

- Humans perceive the context around them through a variety of stimuli.

- Machines can also exploit data coming from different sensors, such as camera, microphone, accelerometers, gyroscope, etc.;
• Each input modality is characterized by distinct statistical properties;

• A joint representation can benefit from the combination of multiple modalities;

To merge different modalities to improve context understanding and prediction
Challenges

• Handle multi-modal data
  • Dataset availability
  • Manage different sampling rates;
  • Select the best fusion strategy
    • Early fusion
    • Late fusion

• Prediction:
  • Learn a representation space suitable for prediction;
  • Achieve performances comparable to context understanding
Some notions on Neural Networks
Consider a supervised learning problem, where we have access to labeled training examples \((x^{(i)}, y^{(i)})\). Neural networks give a way of defining a complex, non-linear form of hypotheses \(h_{W,b}(x)\), with parameters \(W, b\) we can fit to our data.

The neuron is a computational unit that outputs

\[
h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{3} W_i x_i + b)
\]

Where \(f\) is called activation function (sigmoid, ReLU, etc.)

A neural network puts together several neurons

Training in 2 phases:

- Feed forward: \(a^{(l+1)} = f(W^{(l)} a^{(l)} + b^{(l)})\)
- Backpropagation:
  - Suppose we have a fixed training set \(\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}\)
  - We define the overall cost function to be:
    \[
    J(W, b) = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{2} \| h_{W,b}(x^{(i)}) - y^{(i)} \|^2 \right) + \frac{\lambda}{2} \sum_{l=1}^{n_l-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (W_{ji}^{(l)})^2
    \]

Our goal is to minimize \(J(W, b)\) as a function of \(W\) and \(b\).
Neural networks are the state-of-the-art algorithm to understand complex sensory data such as images, videos, speech, audio, voice, music, etc.
State of the Art
Paper 1: Multimodal Deep Learning (1/2)

• Authors: J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng

• International Conference on Machine Learning, 2011

• Input: sequences of video frames and sequences of audio signals.

• They use different models for feature learning:
  • They train Restricted Boltzmann Machines for (a) audio and (b) video separately as a baseline;
  • The shallow model (c) is limited and they find that this model is unable to capture correlations across the modalities;
  • The bimodal deep belief network (DBN) model (d) is trained in a greedy layer-wise fashion by first training models (a) & (b).
• Then, they use deep autoencoder models. A video-only model is shown in (a) where the model learns to reconstruct both modalities given only video as the input. A similar model can be drawn for the audio-only setting.

• They also train the (b) bimodal deep autoencoder in a denoising fashion, using an augmented dataset with examples that require the network to reconstruct both modalities given only one.
Paper 2: Anticipating Daily Intention using On-Wrist Motion Triggered Sensing (1/2)

• Authors: Tz-Ying Wu, Ting-An Chien, Cheng-Sheng Chan, Chan-Wei Hu, Min Sun, ICCV 2017

• Input: sequences of video frames and action sequences.

• The authors use this model for predicting daily intentions;

• Their model consists of an RNN with LSTM cell encoder and a Policy Network. At each frame, RNN will generate an anticipated intention according to a new embedded representation $g$ and the previous hidden state $h$ of the RNN;

• The policy will generate the motion-trigger decision $a$ for next frame, based on motion representation and the hidden state $h$ of the RNN.
The joint loss is: \( L = L^P + \lambda L^A \)

They use exponential loss to train their RNN-based model. The anticipation loss \( L^A \) is defined as

\[
\sum_{t=1}^{T} L^A_t = \sum_{t=1}^{T} - \log p_t(y^{gt}) \cdot e^{\log(0.1) \frac{T-t}{T}}
\]

where \( y^{gt} \) is the ground truth intention and \( T \) is the time when intention reached.

They define a policy loss function

\[
L^P = -\frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \log(\pi(a^k_t \mid (h^k_t, f^k_{m,t}); W_P)) \cdot R^k_t
\]

where \( \{a^k_t\}_t \) is the \( k^{th} \) sequence of trigged patterns sampled from \( \pi(\cdot) \), \( K \) is the number of sequences, and \( T \) is the time when intention reached. \( R^k_t \) is the reward of the \( k^{th} \) sampled sequence at time \( t \) computed from

\[
R = \begin{cases} 
    p_t(y^{gt}) \cdot R^+ \cdot \left(1 - \frac{n}{T}\right), & \text{if } y = y^{gt} \\
    p_t(y^{gt}) \cdot R^- \cdot \frac{n}{T}, & \text{if } y \neq y^{gt}
\end{cases}
\]
Paper 3: Jointly Learning Energy Expenditures and Activities using Egocentric Multimodal Signals (1/2)

- Authors: Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, Li Fei-Fei, CVPR 2017
- Input: $\mathcal{V} = \{v_1, v_2, \ldots, v_T\}$ sequence of video frames; $\mathcal{A} = \{a_1, a_2, \ldots, a_T\}$ sequence of triaxial acceleration signals;
- The authors train a model for jointly perform activity detection and energy expenditure regression;
- The heart rate signal is used as a self-supervision to derive energy expenditure
A multi-task loss is used to jointly optimize activity detection $y^t_a$ and energy expenditure regression $y^t_e$ at each time step.

Training data $(x_t, y^a_t, y^e_t)$, where $x_t \in \mathbb{R}^d$ is the input feature vector, $y^a_t \in \mathbb{R}^{24}$ is the ground truth activity label, and $y^e_t \in \mathbb{R}$ is the derived energy expenditure ($y^e_t = \alpha HR_t + \beta \text{weight} + \gamma HR_t \text{weight}$);

Loss Function: $L = L_{act} + \lambda L_{EE}$, where the first term $L_{act}$ is a cross-entropy loss for activity detection and the second term $L_{EE}$ is a Huber loss for energy expenditure regression

$$L_{EE} = \begin{cases} \frac{1}{2} r^2 & \text{if } |r| \leq \delta \\ \delta \left( |r| - \frac{1}{2} \delta \right) & \text{otherwise} \end{cases}$$

where $r = y^e_t - \widehat{y}^e_t$. 

Paper 3: Jointly Learning Energy Expenditures and Activities using Egocentric Multimodal Signals (2/2)
Our work (preliminary analysis)
Chosen multi-modal dataset:
Stanford-ECM Dataset (from Paper 3)

• Stanford-ECM Dataset comprises 31 hours of egocentric video (113 videos) augmented with heart rate and acceleration data;

• The lengths of the individual videos covered a diverse range from 3 minutes to about 51 minutes in length;

• The mobile phone collected egocentric video at 720x1280 resolution and 30fps, as well as triaxial acceleration at 30Hz;

• The wrist-worn heart rate sensor was used to capture the heart rate every 5 seconds (0.2 Hz);

• Annotated with 25 different activities

• It has been conceived for multimodal classification (not prediction)
ECM Dataset Adaptation for Prediction

- Identified only 2 types of transitions, suitable for performing training/test:
  - Unknown/Activity
  - Activity/Unknown

- The dataset has been cut around transitions
  - 64 frames before and after transitions

- Selected 9 activities, besides Unknown
  - Bicycling
  - Playing With Children
  - Walking
  - Strolling
  - Food Preparation
  - Talking Standing
  - Talking Sitting
  - Sitting Tasks
  - Shopping
Features extraction

• Features have been separately computed before and after each transition (Unknown/Activity and Activity/Unknown);

• Visual features:
  • Features extracted from pool5 layer of Inception CNN, pretrained on ImageNet;
  • \( x_t^v = CNN_{\theta_c}(v_t) \), 1024-dimensional feature vector

• Acceleration features (sliding window size=32):
  • Time-domain: mean, standard deviation, skewness, kurtosis, percentiles;
  • Frequency domain: spectral entropy;

• Heart rate features:
  • mean and standard deviation
We represent features in a temporal pyramid with three levels (level 0, level 1 and level 2).

The top level $j = 0$ is a histogram (mean) over the full temporal extent of a data, the next level ($j=1$) is the concatenation of two histograms obtained by temporally segmenting the video into two halves, and so on.

Feature vector size is 7434
- Video: 1024 x 7;
- Acceleration: 36 x 7;
- Heart rate: 2 x 7
• SVM (Support Vector Machine) trained on feature vectors
  • **Classification**: classify the activity from feature vectors extracted from “Activity” chunk;
  • **Prediction**: predict the activity from feature vectors extracted from the “Unknown” chunk before the activity
Baseline Results

• Training and test with all the input combinations
• Prediction: 10% accuracy drop w.r.t. classification

<table>
<thead>
<tr>
<th>Inputs (feature size)</th>
<th>Classification</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear SVM</td>
<td>Rbf SVM</td>
</tr>
<tr>
<td>Acceleration (252)</td>
<td>30%</td>
<td>46.67%</td>
</tr>
<tr>
<td>Heart rate (14)</td>
<td>33.33%</td>
<td>30%</td>
</tr>
<tr>
<td>Video (7168)</td>
<td>76.67%</td>
<td>76.67%</td>
</tr>
<tr>
<td>Video + Acceleration (7420)</td>
<td>76.67%</td>
<td><strong>83.33%</strong></td>
</tr>
<tr>
<td>Video + Heart rate (7182)</td>
<td>76.67%</td>
<td>76.67%</td>
</tr>
<tr>
<td>Acceleration + Heart rate (266)</td>
<td>33.33%</td>
<td>46.67%</td>
</tr>
<tr>
<td>Video + Acceleration + Heart rate (7434)</td>
<td>76.67%</td>
<td>76.67%</td>
</tr>
</tbody>
</table>
Conclusions & Future Work

• Conclusions:
  • Prediction seems a feasible task;
  • Multi-modality improves both classification and prediction;

• Future work:
  • Fill the gap between classification and prediction accuracies
  • Populate an ST Multimodal dataset, exploiting Bluecoin sensors and video from a smartphone
Thank you