Multimodal Fusion using Dynamic Hybrid Models
Mohamed R. Amer, Behjat Siddiquie, Saad Khan, Ajay Divakaran and Harpreet Sawhney
SRI International, Princeton, NJ

**Problem and Motivation**

- **Goal**: Detect Multimodal events in time varying sequences.
- **Application**: Analysis of Human behaviors and emotions: Facial expressions, paralinguistics, eye gaze, hand gestures, head motion etc.
- **Temporal dynamics within and across modalities is key to modeling and capturing affect.**

**Approach**

- Staged hybrid model: exploits the strength of discriminative classifiers along with the representational power of generative models.

**Staged-Dynamic Hybrid Model**

*Why Staged Hybrid Dynamic Model?*

- (Staged) training each model separately, where the discriminative model trained on representations learned by the generative model.
- (Hybrid) exploiting the generative model’s expressiveness and the discriminative model’s classification power.
- (Dynamic) Modeling the temporal content of time varying data is important.

**Generative – Multimodal Conditional Restricted Boltzmann Machines:**

- Multimodal CRBMs consists of single CRBMs, and fusion CRBM.
- $p_c(v, h, v_{z,1}) = \exp[-E_c(v, h, v_{z,1})]/Z(\theta_0)$
- $Z(\theta_0) = \sum_{v, h} \exp[-E_c(v, h, v_{z,1})]$, \( \theta_0 = \{a, b, A, B, W\}$
- Each of the single CRBMs captures the representation of one modality.
- $E_c(v, h, v_{z,1} | \theta_0) = \sum_{i} E_c(v_i, h_i, v_{z,1} | \theta_0)$
- Fused features classified by the CRF

**Discriminative – Conditional Random Fields:**

- The CRF operates on the features learned using MMCRBM.
- $p_d(y | h^f, \theta_0) = \exp[E_d(y | h^f, \theta_0)]/Z(\theta_0)$
- $Z(\theta_0) = \sum_{h^f} \exp[E_d(y | h^f, \theta_0)]$
- $E_d(y | h^f, \theta_0) = \sum_{i} \omega_i f_i(y_i, 1, y_i, h_i^f) + \sum_{i,j} \alpha_i f_{ij}(y_i, h_i^f)$

**Hybrid Dynamic Model**

- $p_{hyb}(v | v_{z,1}, h_{z,1}) = p_c(v | v_{z,1}, h_{z,1}) \cdot p_d(v | v_{z,1}, h_{z,1})$

**Learning**

- Generative – Contrastive Divergence: \( \theta_0 = \{a, b, A, B, W\} \)
- Discriminative – Max. Likelihood Estimation: \( \theta_0 = \{\omega', \omega''\} \)

**Inference (Bottom-Up)**

- Activate each modality of the MMCRBM:
- $p_c(h^o_i = 1 | v^o_i, v^m_i) \sim \sigma(c_i^o + \sum_{o} c_{oo} v^o_o)$
- Activate the fusion layer:
- $p_c(h^f_i = 1 | h^o_{-i-M}, h^m_{-i-M}) \sim \sigma(c^f_i + \sum_{o} h^o_{-i-M} \omega^f_o)$
- Fused features classified by the CRF

**Results and Conclusions**

**Average Classification Accuracy**

- Compare our approach against the relevant baselines and the state-of-the-art on the AVEC, AVLetters and CUAVE datasets.

<table>
<thead>
<tr>
<th>Model/Dataset</th>
<th>AVEC-A</th>
<th>AVEC-C</th>
<th>AVLetters-A</th>
<th>AVLetters-C</th>
<th>CUAVE-A</th>
<th>CUAVE-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF-RBM</td>
<td>68.1</td>
<td>65.5</td>
<td>60.9</td>
<td>56.3</td>
<td>45.3</td>
<td>42.8</td>
</tr>
<tr>
<td>SVM-RBM (10%)</td>
<td>69.7</td>
<td>69.7</td>
<td>65.0</td>
<td>57.5</td>
<td>41.7</td>
<td>39.1</td>
</tr>
<tr>
<td>SVM-RBM (50%)</td>
<td>71.7</td>
<td>70.7</td>
<td>65.4</td>
<td>57.2</td>
<td>42.7</td>
<td>40.3</td>
</tr>
</tbody>
</table>

**Missing Data:**

- **within Modality**
- **across Modalities**

**Conclusions**

- Hybrid Dynamic Model: effective for classifying sequential data from multiple heterogeneous modalities.
- Discriminative Model (CRF): Models long range temporal dynamics.

**Acknowledgements**

- Defense Advanced Research Projects Agency
- Army Research Office