Context-sensitive Facial Expressivity Prediction by Multimodal Hierarchical Bayesian Neural Networks

Ajjen Joshi\textsuperscript{1}, Soumya Ghosh\textsuperscript{2}, Sarah Gunnery\textsuperscript{3}, Linda Tickle-Degnen\textsuperscript{2}, Stan Sclaroff\textsuperscript{1}, Margrit Betke\textsuperscript{1}

\textsuperscript{1}Boston University, \textsuperscript{2}IBM T.J. Watson Research Center, \textsuperscript{3}Tufts University

Overview

- We investigate whether contextual-information can be leveraged for the task of predicting facial expressivity in patients with Parkinson’s Disease.
- We experiment with two notions of context: (1) gender and (2) sentiment.
- We train hierarchical Bayesian neural networks with multimodal feature representations.

Contributions

- We demonstrate the benefits of using a framework that adapts to contextual information.

Hierarchical Bayesian Neural Networks

- Given a dataset \( D = \{ x_n, y_n \}_{n=1}^N \), each group is endowed with its own conditional distribution \( p(y_n | \pi_n, f(r_n, W_n)) \).

\[ p(W, y | \pi, f) = \prod_{n=1}^N p(y_n | \pi_n, f(r_n, W_n)) \]

\[ p(\pi | \eta) = \prod_{n=1}^N \prod_{r=1}^R \mathcal{N}(\pi_n^{(r)} | 0, \eta^{(r)}) \]

\[ p(\eta | \tau) = \mathcal{N}(\eta | 0, \tau_\eta^{1/2}) \]

- The joint distribution is given by:

\[ p(W, \pi, T, y | x_n, \tau, \eta, v) = p(W, \pi | \tau_\eta^{1/2}) \prod_{n=1}^N \prod_{r=1}^R p(\eta_n^{(r)} | \pi_n^{(r)}) p(\pi_n^{(r)} | \tau_\eta^{1/2}) \]

Inference

- We approximate the intractable posterior with a fully factorized variational approximation,

\[ q(W, \pi, T | \phi) = q(W_0 | \phi_0) \prod_{n=1}^N q(\pi_n | \phi) q(\tau_\eta^{1/2} | \phi_\tau) \]

- The Evidence Lower Bound (ELBO) is then maximized with respect to the variational parameters using variational Bayes.

\[ \mathcal{L}(\phi) = \mathbb{E}_q[\ln p(W, \pi, T, y | x_n, \tau, \eta, v)] - \mathbb{E}_q[\ln q(W, \pi, T | \phi)] \]

- In computing the Monte Carlo estimate of the gradients, we use the local reparameterization trick.

Model Pipeline

Results

- We test our method on a dataset of 772 short audio-video clips of patients with Parkinson’s Disease using 9-fold cross validation.
- We divide the dataset into context-sensitive groups.
- For each video clip we extract:
  a) Action Unit stats (AU-stats) to capture visual features
  b) MFCC stats (MFCC-stats) to capture audio features

1. Action Unit Analysis

2. Multi-modality

3. Context-sensitive Models