Hierarchical Bayesian Neural Networks for Personalized Classification

Ajjen Joshi\textsuperscript{1}, Soumya Ghosh\textsuperscript{2}, Margrit Betke\textsuperscript{1}, Hanspeter Pfister\textsuperscript{3}

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

**Problem Statement**

- Building robust classifiers trained on data susceptible to group-specific variations is a challenging problem.
- We develop flexible hierarchical Bayesian models that parameterize group-specific conditional distributions via multi-layered Bayesian neural networks and use it for personalized gesture recognition.

**Hierarchical Bayesian Neural Networks**

- Given a dataset $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$; each subject is endowed with its own conditional distribution

\[ p(y_n | x_n, \theta) = \prod_{g=1}^G \prod_{i=1}^N p(y_{g,i} | \theta_{g,i}, \tau_{g}) \]

- The joint distribution is given by:

\[ p(\theta, W, y, x) = \prod_{g=1}^G \prod_{i=1}^N p(y_{g,i} | \theta_{g,i}, \tau_{g}) \]

**Inference**

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

\[ q(\theta, W, \mathcal{T} | \phi) = q(\theta_0 | \phi_0) \prod_{g=1}^G q(W_g | \phi_g)q(\tau_{g}^{-1/2} | \phi_{g}) \]

- The ELBO is then maximized with respect to the variational parameters using doubly stochastic Variational Bayes.

\[ \mathcal{L}(\phi) = E_{q_{\phi}}[\ln p(\theta_0, W, \mathcal{T}, y | x, z, \tau_0, \theta)] - E_{q_{\phi}}[\ln q(\theta_0, W, \mathcal{T} | \phi)] \]

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

- Predictions on held-out data are made via Monte Carlo estimates of the posterior predictive distribution.

**Results**

We test our method on the MSRC-12 Gesture Dataset (~4900 gestures, 12 unique gestures, 30 subjects).

1. **Benefits of local reparameterization**

2. **Model flexibility**

3. **Personalization**

\[ \{W_g\}_{g=1}^{G+1} \]

- Given a model trained on $\mathcal{D}$, we only update $W_{G+1}$ while keeping everything else fixed.

- To best utilize limited labeling resources, we adopt the Bayesian Active Learning by Disagreement (BALD) algorithm to adaptively select training instances for the new group.

\[ x_\ell = \arg\max_{x \in X_{pool}} \mathcal{H}[y | x, \mathcal{D}] - \mathbb{E}_{W_{G} \sim p(W_g | \mathcal{D})} \mathcal{H}[y | x, W_g] \]