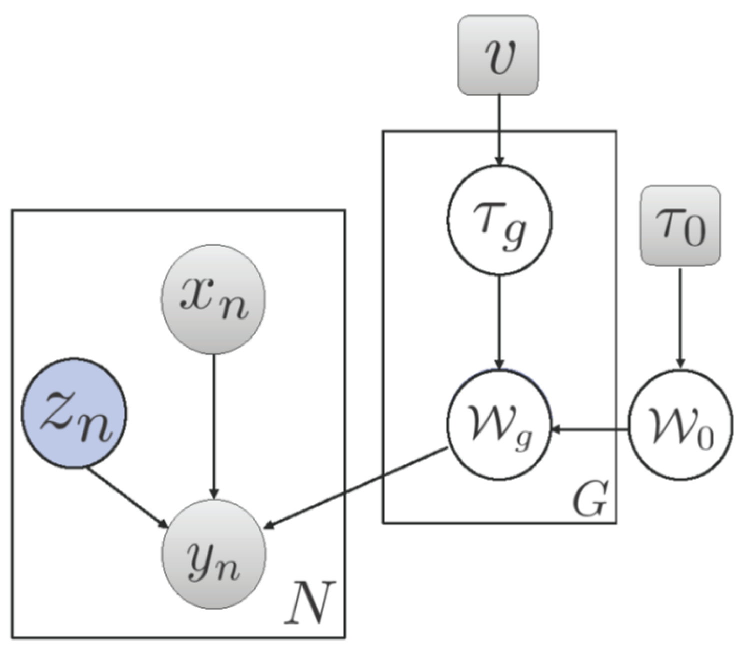
Hierarchical Bayesian Neural Networks for Personalized Classification

Ajjen Joshi[1], Soumya Ghosh[2], Margrit Betke[1], Hanspeter Pfister[3]
[1]Boston University, [2]IBM T.J. Watson Research Center, [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 $p(y_n \mid z_n = g, f(x_n, \mathcal{W}_g))$.

$$p(\mathcal{W}_g \mid \mathcal{W}_0, \tau_g) = \prod_{l=1}^{L}\prod_{i=1}^{V_{l-1}}\prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^{g} \mid w_{ij,l}^{0}, \tau_g^{-1})$$

$$p(\mathcal{W}_0 \mid \tau_0) = \prod_{l=1}^{L}\prod_{i=1}^{V_{l-1}}\prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^{0} \mid 0, \tau_0^{-1})$$

$$p(\tau_g^{-1/2} \mid v) = \text{Half-Normal}(\tau_g^{-1/2} \mid 0, v)$$

- The joint distribution is given by:

$$p(\mathcal{W}_0, \mathcal{W}, \mathcal{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \tau_0, v) = p(\mathcal{W}_0 \mid \tau_0^{-1}) \prod_{g=1}^{G} p(\gamma_g \mid v) p(\mathcal{W}_g \mid \mathcal{W}_0, \tau_g^{-1}) \prod_{n=1}^{N}\prod_{g=1}^{G} p(y_n \mid f(\mathcal{W}_g, x_n))^{\mathbf{1}[z_n = g]}$$

Inference

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

$$q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi) = q(\mathcal{W}_0 \mid \phi_0) \prod_{g=1}^{G} q(\mathcal{W}_g \mid \phi_g) q(\tau_g^{-1/2} \mid \phi_{\tau_g})$$

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

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\ln p(\mathcal{W}_0, \mathcal{W}, \mathcal{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \tau_0, v)] - \mathbb{E}_{q_\phi}[\ln q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi)]$$

- 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.

Personalization

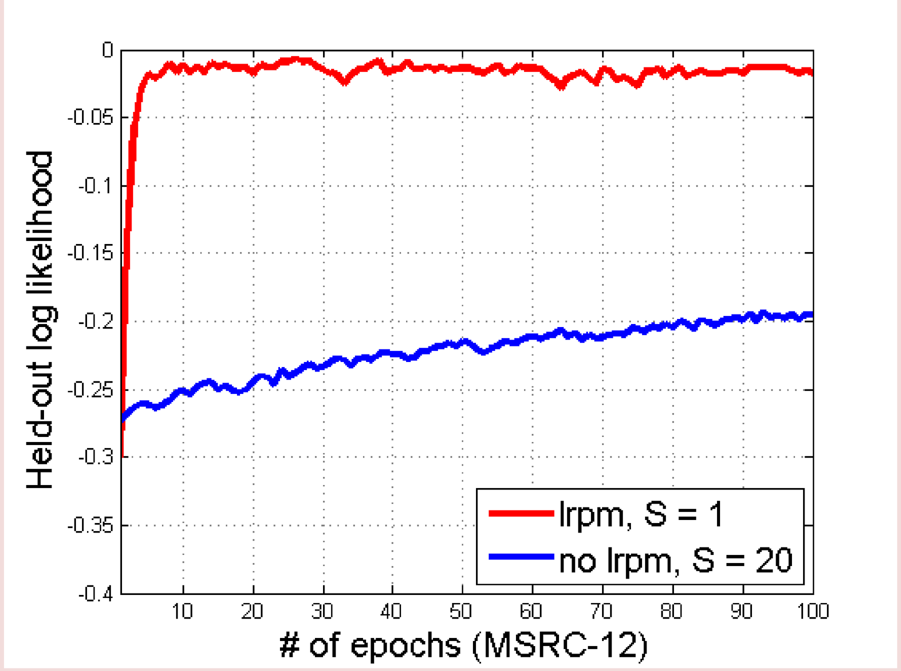
- $\{\mathcal{W}_g\}_{g=1}^{G+1}$ are conditionally independent given $\mathcal{W}_0$.
- Given a model trained on $\mathcal{D}$, we only update $\mathcal{W}_{G+1}$ while keeping everything else fixed.
- To best utilize limited labeling resources, we adopt the Bayesian Active Learning by Disagreement (BALD) algorithm to adaptively select training instances for the new group.

$$x_l = \operatorname*{argmax}_{x \in X_{pool}} \mathbb{H}[y \mid x, \mathcal{D}] - \mathbb{E}_{\mathcal{W}_g \sim p(\mathcal{W}_g \mid \mathcal{D})} \mathbb{H}[y \mid x, \mathcal{W}_g]$$

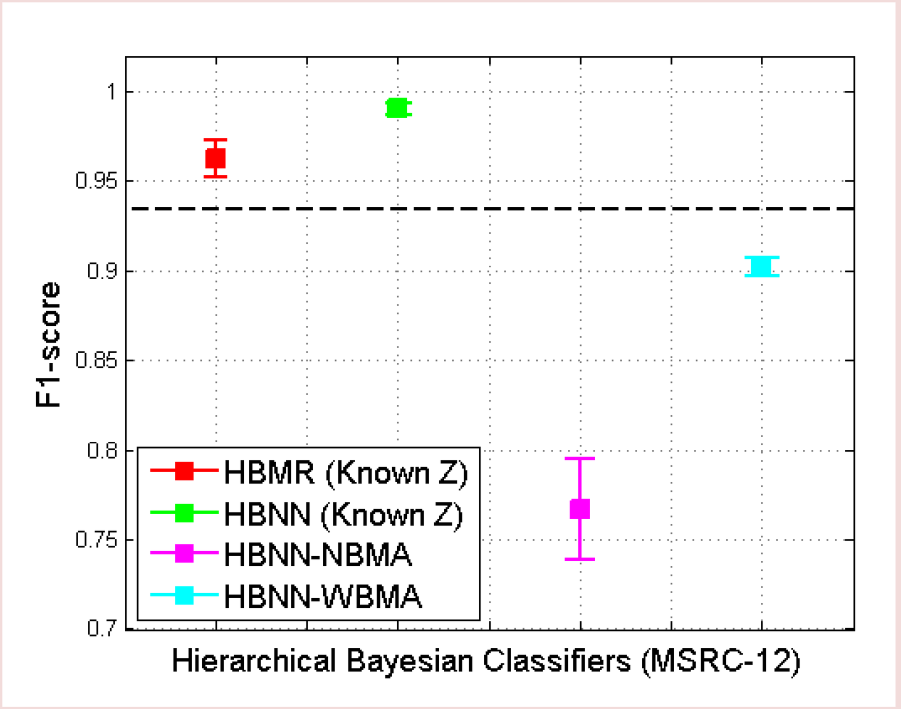
Results

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