Supplementary Material: Personalizing Gesture Recognition Using Hierarchical Bayesian Neural Networks

Ajjen Joshi[1] Soumya Ghosh[2] Margrit Betke[1] Stan Sclaroff[1] Hanspeter Pfister[3]
[1]Boston University [2]IBM T.J. Watson Research Center [3]Harvard University
[1]{ajjendj, betke, sclaroff}@bu.edu [2]ghoshso@us.ibm.com [3]pfister@seas.harvard.edu

Benefits of Local Reparameterization

For all three datasets, on fifteen random 75-25 split of the data, we trained a Hierarchical Bayesian Neural Network for 100 epochs, with and without using local reparameterization. When not using local reparameterization, we approximated the ELBO using 20 Monte Carlo samples whereas when using local reparameterization, we only used 1 sample. We plot the mean logarithm of the ELBO versus the number of training epochs (Figure 1) and observe that the ELBO curves for the model that employs local reparameterization is much higher than the model that doesn't, suggesting the model can learn a better approximation of its parameters much faster.

Effects of Modifying Depth of Model

We used a leave-one-subject-out cross validation scheme, where we personalized models pre-trained on $G-1$ subjects with a pool of 7 randomly selected gestures per class from each test subject using HBNNs with 1 and 3 hidden layers. We plot the mean F1-scores for different personalization schemes against the number of personalization instances per gesture for the different HBNN architectures (Figure 2a). We observe that models personalized using BALD outperforms models personalized using RAND for all architectures. The HBNN model with 1 hidden layer performs comparably to the best-performing HBNN with 2 hidden layers. However, the HBNN model with 3 hidden layers performs worse due to overfitting.

Comparing Classification Against Baselines

Wang et al. [2] introduced a feature representation based on using a kernel matrix to model nonlinear relationships among the features and demonstrated *state-of-the-art* results on the MSRC-12 dataset. Joshi et al. [1] demonstrated *state-of-the-art* results on the NATOPS dataset by combining appearance-based and skeleton-based features and training a random forest classifier. On the NATOPS dataset, we compared the performance of our HBNN classifier against these benchmarks. We find that our results outperform Joshi et al.'s random forest approach and is comparable with Wang et al.'s method (Figure 2b). We note here that our method along with the random forest approach used identical raw features to train a classifier, whereas Wang et al.'s method focused on building a feature representation before training an SVM model. While our method should automatically discover useful feature representations from the data, there are benefits of using engineered feature representations in the absence of abundant training data.

Comparing Personalization Against Baselines

We also compared our personalization results with the aforementioned benchmarks using a leave-one-subject-out cross validation scheme and plotting their mean scores (Figure 3). For all three datasets, our personalization models had 2 hidden layers, each with 400 activation nodes. We can observe the benefits of personalization with very few data instances per gesture class. For the MSRC dataset, our personalization model outperforms the best-performing baseline (Covariance features [2]) on average when provided with only 2 gesture examples per class. For the NATOPS dataset, our personalization model outperforms the best-performing baseline (Random Forest [1]) on average when provided with 6 gesture instances per class. For the ChaLearn dataset, our personalization model outperforms the random forest baseline when provided with 5 gesture instances per class.

References

[1] A. Joshi, C. Monnier, M. Betke, and S. Sclaroff. A random forest approach to segmenting and classifying gestures. In *2015 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2015)*. IEEE, 2015. 1

[2] L. Wang, J. Zhang, L. Zhou, C. Tang, and W. Li. Beyond covariance: Feature representation with nonlinear kernel matrices. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4570–4578, 2015. 1

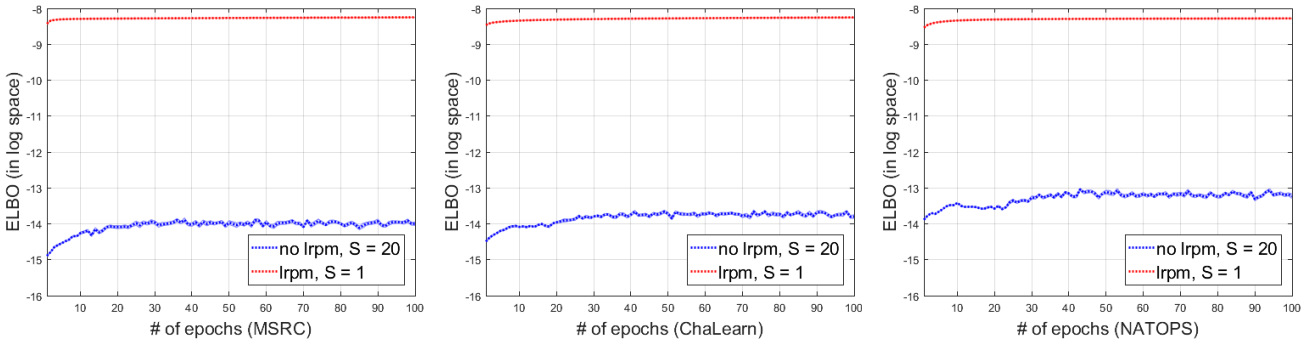

Figure 1. The mean logarithm of the expected lower bound (ELBO) versus the number of training epochs, for 15 random 75-25 splits of the data, when the model uses local reparameterization (lprm) and when it doesn't (no lrpm). For all three datasets, MSRC-12 (left), ChaLearn 2013 (middle) and NATOPS (right), the model reaches a faster convergence when using local reparameterization.

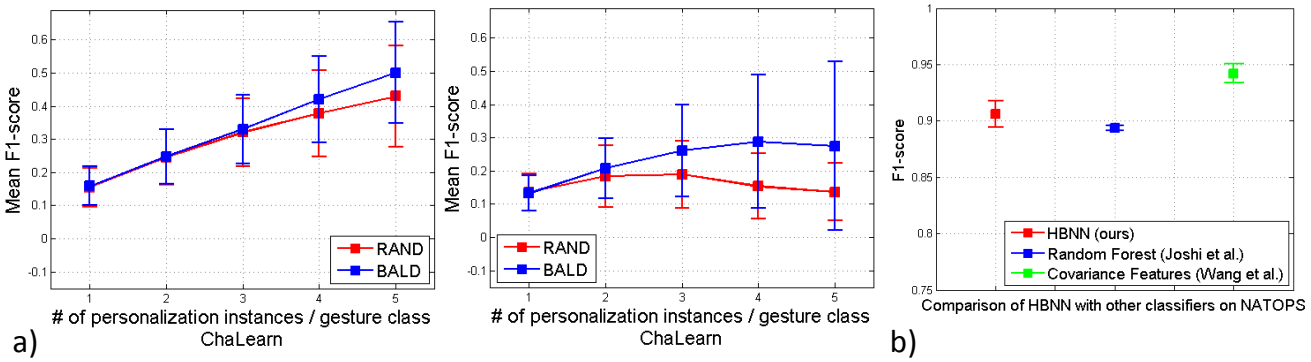

Figure 2. a) The mean F1-scores for different personalization schemes plotted against number of personalization instances per gesture for different HBNN architectures for the ChaLearn dataset for different HBNN architectures: HBNN with one hidden layer (left), and HBNN with three hidden layers (right). b) The mean F1-scores for our classifier, compared against 2 baselines for the NATOPS dataset, trained using 5 random 75-25 splits of the dataset.

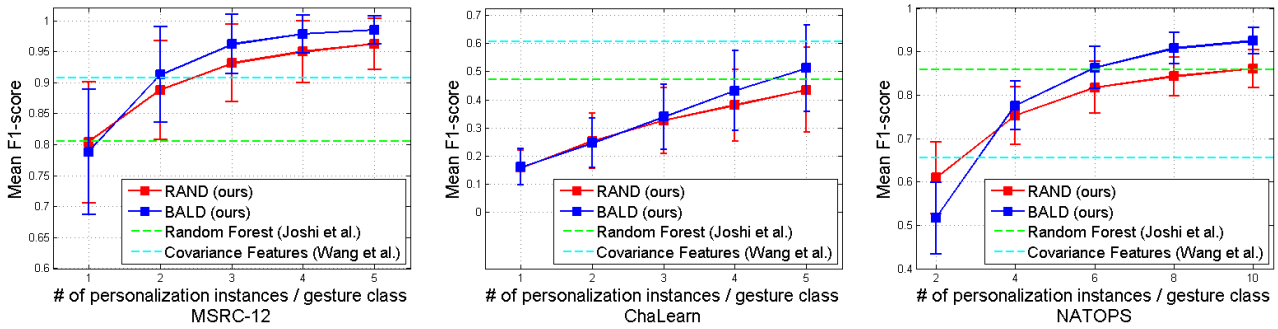

Figure 3. The mean F1-scores for different personalization schemes plotted against number of personalization instances per gesture, compared against 2 baselines for all three datasets: MSRC-12 (left), ChaLearn (middle) and NATOPS (right).