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

1 Hierarchical Bayesian Neural Networks

Building robust classifiers trained on data susceptible to group or subject-specific variations is a challenging yet common problem in pattern recognition. Hierarchical models allow sharing of statistical strength across groups while preserving group-specific idiosyncrasies, and are commonly used for modeling such grouped data [3]. We develop flexible hierarchical Bayesian models that parameterize group-specific conditional distributions $p(y_g \mid x_g, \mathcal{W}_g)$ via multi-layered Bayesian neural networks. Sharing of statistical strength between groups allows us to learn large networks even when only a handful of labeled examples are available. We leverage recently proposed doubly stochastic variational Bayes algorithms to infer a full posterior distribution over the weights while scaling to large architectures. We find the inferred posterior leads to both improved classification performance and to more effective active learning for iteratively labeling data. Finally, we demonstrate *state-of-the-art* performance on the MSRC-12 Kinect Gesture Dataset [2].

Model Given a dataset $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$ containing data from G related but distinct groups, we aim to learn the functional mapping between features ($x_n \in \mathbb{R}^D$) and responses ($y_n \in \mathcal{Y}$). We endow each group with its own conditional distribution, parameterized by multi-layered feedforward neural network. Assuming the distribution factorizes over data instances, we have, $p(\mathbf{y} \mid \mathcal{W}, \mathbf{z}, \mathbf{x}) = \prod_{n=1}^N \prod_{g=1}^G p(y_n \mid f(\mathcal{W}_g, x_n))^{\mathbf{1}[z_n=g]}$. Here, z_n, assumed to be observed during training, is a G dimensional categorical random variable capturing the group membership of data instance n and $\mathcal{W} = \{\mathcal{W}_g\}_{g=1}^G$ where $\mathcal{W}_g$ is the set of group-specific weights parameterizing a neural network f. We place factorized Gaussian priors on $\mathcal{W}_g$ with independent group-specific variances, $p(\mathcal{W}_g \mid \mathcal{W}_0, \tau_g) = \prod_{l,i,j}^{L,V_{l-1},V_l} \mathcal{N}(w_{ij,l}^g \mid w_{ij,l}^0, \tau_g^{-1})$ to model our assumption that each group's functional mapping is an independently corrupted version of a common latent mapping. We further place a Gaussian prior on the mean weights $p(\mathcal{W}_0 \mid \tau_0) = \prod_{l,i,j}^{L,V_{l-1},V_l} \mathcal{N}(w_{ij,l}^0 \mid 0, \tau_0^{-1})$, where i, j, l denote the weight indices of the neural network, and a half-normal prior on the group-specific standard deviations, $p(\tau_g^{-1/2} \mid v) = \text{Half-Normal}(\tau_g^{-1/2} \mid 0, v)$. We fix the hyper-parameter v to a large value to encode our lack of apriori knowledge about τ_g (Figure 1).

Inference We approximate the intractable posterior with a fully factorized approximation, $q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi) = q(\mathcal{W}_0|\phi_0) \prod_{g=1}^G q(\mathcal{W}_g|\phi_g) q(\tau_g^{-1/2}|\phi_{\tau_g})$ and maximize the expected lower bound (ELBO) $\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\ln p(\mathcal{W}_0, \mathcal{W}, \mathcal{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \tau_0, v)] - \mathbb{E}_{q_\phi}[\ln q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi)]$ with respect to the variational parameters ϕ. We rely on doubly stochastic variational Bayes [1, 8] to cope with the intractable expectations in the ELBO. In computing the Monte Carlo estimate of the gradients, we use the local reparameterization trick [6], which leads to significant improvements (Figure 1). We resort to Monte Carlo estimates of the posterior predictive density for predictions on held-out data. Unobserved group memberships of held out data are inferred via an inference network [7, 4].

Personalization We define personalization as the process of adapting the model to data from a new, previously unseen group. We focus on the case when a small (one-shot) number of data instances $\mathcal{D}_{G+1}$ from the new group $G+1$ are made available for adaptation. We learn a group-specific

Workshop on Bayesian Deep Learning, NIPS 2016, Barcelona, Spain.

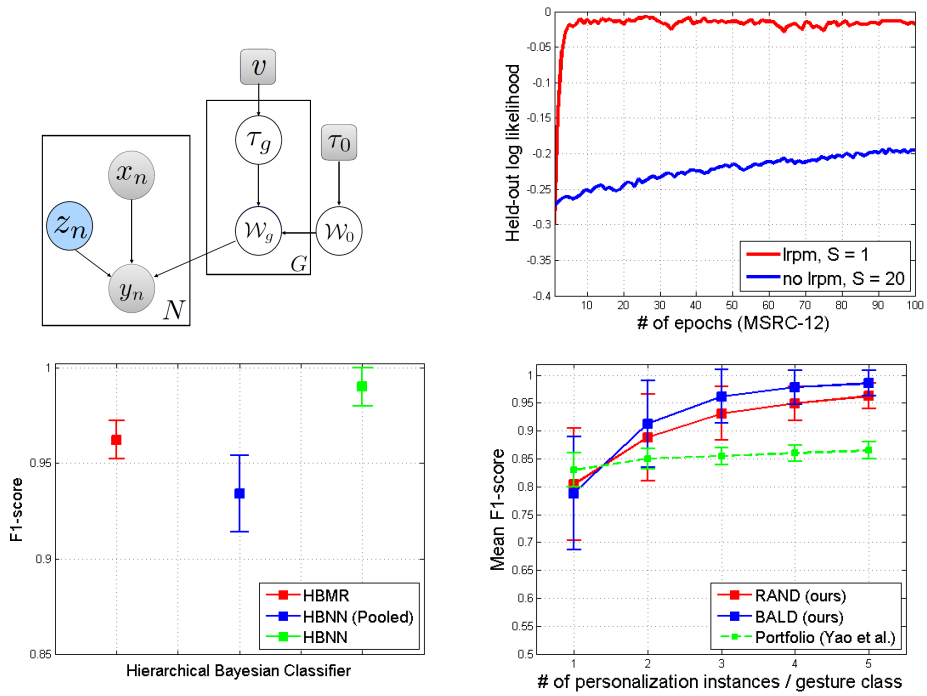

Figure 1: **TOP** *Left*: Hierarchical Bayesian Neural Network. The node shaded blue indicates that it is observed during training but may be latent at test. *Right*: Benefits of local re-parameterization: The held-out log likelihood is plotted against iterations with and without using local re-parameterization. **Bottom** *Left*: Mean F-scores plotted for Hierarchical Bayesian Multinomial Regression (HBMR), Hierarchical Bayesian Neural Network (HBNN) and a Bayesian Neural Network (HBNN-Pooled) when data across all subjects is pooled. *Right*: One-shot personalization results: Comparison of F-scores when the HBNN model is personalized using random samples (RAND) to when it is updated using BALD along with one of the results reported by Yao et al. [9].

model $\mathcal{W}_{G+1} \mid \mathcal{D}_{G+1}$ from the new training instances. The learning can be performed efficiently by observing that $\{\mathcal{W}_g\}_{g=1}^{G+1}$ are conditionally independent given $\mathcal{W}_0$. Thus, given a model trained on $\mathcal{D}$, we only update $\mathcal{W}_{G+1}$ while keeping the estimates $\{\mathcal{W}_g\}_{g=1}^{G} \mid \mathcal{D}$ and $\mathcal{W}_0 \mid \mathcal{D}$ fixed.

Acquiring new training instances $\mathcal{D}_{G+1}$ is often made expensive by the onerous process of labeling them. To best utilize limited labeling resources, we explore the use of active learning. Recent advances in active learning exploit posterior distributions over parameters to outperform traditional maximum entropy based search [5]. Here, we adopt a recently proposed algorithm, Bayesian active learning by disagreement (BALD) [5] to adaptively select training instances for group $G + 1$.

2 Results and Discussion

We test our method on the MSRC-12 Gesture Dataset [2], containing ∼4900 gesture instances with 12 gestures collected from 30 subjects. We normalized the temporally segmented gestures into 10 frames and concatenated the coordinates of 20 body joints over the normalized gesture segment to create 600-dimensional input features. We used 5 random 75/25 splits of the dataset to test different versions of our framework: Hierarchical Bayesian Multinomial Regression (HBMR), Hierarchical Bayesian Neural Network (HBNN) with 2 hidden layers, each with 400 activation nodes, and an HBNN with all training data pooled (HBNN-Pooled) into 1 group (Figure 1). For our personalization setup, we used leave-one-subject-out cross-validation, where we personalized models pre-trained on 29 subjects with a pool of 7 randomly selected gestures from each class from the test subject. Our BALD method outperforms the personalization schemes of Yao et al. [9], who train a portfolio of random forests, from which the most apt classifier is chosen given personalization instances. Thus, we show the benefit of using our hierarchical framework (HBNN) over a generic classifier (HBNN-Pooled) and demonstrate that our personalization approach (BALD) outperforms a competing method [9].

References

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1613–1622, 2015.

[2] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746. ACM, 2012.

[3] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[4] Samuel J Gershman and Noah D Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2014.

[5] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[6] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, 2015.

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

[8] Michalis Titsias and Miguel Lázaro-gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979, 2014.

[9] Angela Yao, Luc Van Gool, and Pushmeet Kohli. Gesture recognition portfolios for personalization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1923–1930. IEEE, 2014.