Uncertainty aware audiovisual activity recognition using deep Bayesian variational inference

Mahesh Subedar* Ranganath Krishnan* Paulo Lopez Meyer

Omesh Tickoo Jonathan Huang

Intel Labs

Abstract

Deep neural networks (DNNs) provide state-of-the-art results for a multitude of applications, but the approaches using DNNs for multimodal audiovisual applications do not consider predictive uncertainty associated with individual modalities. Bayesian deep learning methods provide principled confidence and quantify predictive uncertainty. Our contribution in this work is to propose an uncertainty aware multimodal Bayesian fusion framework for activity recognition. We demonstrate a novel approach that combines deterministic and variational layers to scale Bayesian DNNs to deeper architectures. Our experiments using in- and out-of-distribution samples selected from a subset of Moments-in-Time (MiT) dataset show a more reliable confidence measure as compared to the non-Bayesian baseline and the Monte Carlo dropout (MC dropout) approximate Bayesian inference. We also demonstrate the uncertainty estimates obtained from the proposed framework can identify out-of-distribution data on the UCF101 and MiT datasets. In the multimodal setting, the proposed framework improved precision-recall AUC by 10.2% on the subset of MiT dataset as compared to non-Bayesian baseline.

1. Introduction

Vision and audio are complementary inputs and fusing these modalities can greatly benefit an activity recognition application. Multimodal audiovisual activity recognition using deep neural network (DNN) architectures are not successful in modeling the inherent ambiguity in the correlation between two modalities. One of the modalities (e.g., sneezing in audio, writing in vision) can be more certain about the activity class than the other modality. It is important to model reliable uncertainty estimates for the individual modalities to benefit from multimodal fusion. Probabilistic Bayesian models provide principled ways to gain insight about data and capture reliable uncertainty estimates in predictions. Bayesian deep learning [1, 2] has allowed bridging DNNs and probabilistic Bayesian theory to leverage the strengths of both methodologies.

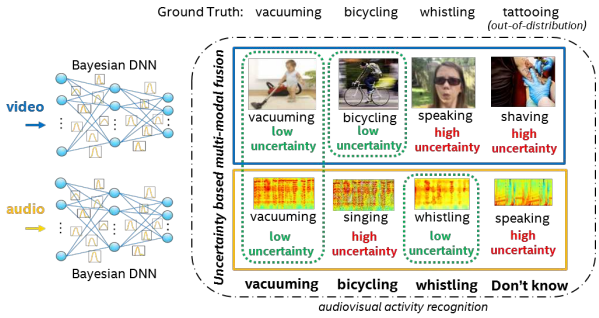


Figure 1: Uncertainty aware audiovisual activity recognition

Multimodal models have been proposed for audiovisual analysis tasks such as emotion recognition [3], audiovisual speech recognition [4], speech localization [5, 6], cross-modal retrieval [7]. These audiovisual methods apply joint modeling of the audio and vision inputs during the training phase for better generalizability of the models, but then use single modality during the inference phase. None of the methods listed here provide a quantifiable means to determine the relative importance of each modality.

Our main contributions in this work include: a) A multimodal fusion framework based on predictive uncertainty estimates applied to activity recognition: To the best of our knowledge, this is the first work on multimodal fusion based on uncertainty estimates using Bayesian deep learning with variational inference. (b) A scalable Bayesian variational inference by combining deterministic and variational layers in DNNs. (c) Identifying out-of-distribution data for activity recognition using uncertainty estimates: We demonstrate the uncertainty estimates obtained from the proposed architecture can identify out-of-distribution data in Moments-in-Time (MiT) and UCF-101 action recognition datasets.

*These two authors contributed equally.

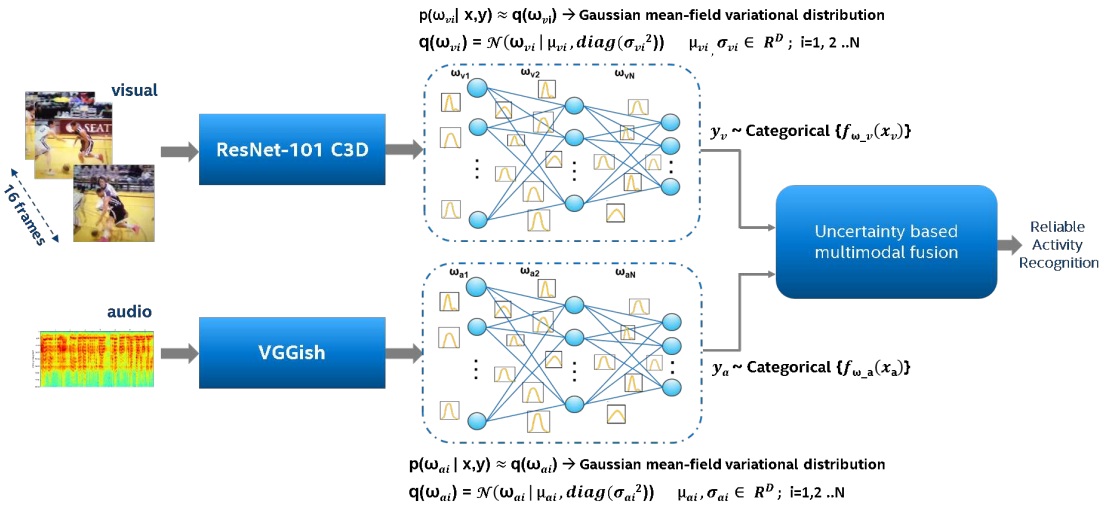


Figure 2: Bayesian audiovisual activity recognition: ResNet-101 C3D and VGGish DNN architectures are used to represent vision and audio information, respectively. The final layer of the DNN is replaced with three fully connected variational layers followed by categorical distribution. The Bayesian inference is applied to the variational layers through Monte Carlo sampling on the posterior of model parameters, which provides the predictive distribution.

2. Bayesian Multimodal DNN Architecture

We present a Bayesian multimodal fusion framework based on uncertainty estimates for audiovisual activity recognition. The block diagram of the proposed audiovisual activity recognition using Bayesian variational inference is shown in Figure 2. We use the ResNet-101 C3D [8] and VGGish [9] architectures for visual and audio modalities, respectively. We replace the final fully connected layer for both vision and audio DNN models with three fully connected variational layers followed by the categorical distribution.

The weight and bias parameters in the fully connected variational layers are modeled through mean-field Gaussian distribution, and the network is trained using Bayesian variational inference based on KL divergence [10, 11]. In order to learn the posterior distribution of model parameters w, we train Bayesian DNN with variational inference method. The objective is to optimize log evidence lower bound (ELBO) [12] as the cost function. The model parameters of the fully connected variational layers are parametrized by mean μ and variance σ^2, i.e. $q_\theta(w) = \mathcal{N}(w|\mu, \sigma^2)$. These parameters in the variational layers are optimized by minimizing the negative ELBO loss (L^v) [12]. We use Flipout [13], which is an efficient method that correlates the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each input. The parameters in deterministic layers are optimized using cross-entropy loss (L^d) [14]. The model parameters for variational and deterministic DNN layers are obtained by applying stochastic gradient descent optimizer [15] to the loss functions (details are in Appendix A). During prediction stage we perform multiple Monte Carlo forward passes on the final variational layers by sampling the parameters from learned posteriors to measure uncertainty estimates [16].

In [17], an accuracy vs uncertainty (AvU) metric is proposed obtained from the confusion matrix values: number of accurate-certain (n_{ac}), inaccurate-uncertain (n_{iu}), accurate-uncertain (n_{au}) and inaccurate-certain(n_{ic}) predictions. A reliable model will provide higher AvU score. An uncertainty threshold value that maximizes AvU metric from individual modalities is the optimal threshold, which is used for multimodal fusion. We perform average pooling of the audio-vision predictive distributions if the uncertainty measures in both modalities are below the optimal threshold values, else we rely on the single modality that has lower uncertainty measure. For comparison with the non-Bayesian baseline, we maintain the same model depth as the Bayesian DNN model and use three deterministic fully connected final layers for the non-Bayesian DNN model. Dropout layer is used after every fully connected layer to avoid over-fitting of the model. In the rest of the document, we refer the non-Bayesian DNN model as simply the DNN model.

3. Results

We analyze the model performance on the Moments-in-Time (MiT) [18] dataset. The MiT dataset consists of 339 classes, and each video clip is 3 secs (~90 frames) in length. In this work, we considered a subset of 54 classes as in-distribution and another 54 classes as out of distribution

Model	Top1 (%)	Top5 (%)
Vision		
DNN	52.65	79.79
Bayesian DNN (MC Dropout)	52.88	80.10
Bayesian DNN (Stochastic VI)	53.3	81.20
Audio		
DNN	34.13	61.68
Bayesian DNN (MC Dropout)	32.46	60.97
Bayesian DNN (Stochastic VI)	35.80	63.40
Audiovisual		
DNN	56.61	79.39
Bayesian DNN (MC-Dropout)	55.04	80.34
Bayesian DNN (Stochastic VI)	**58.2**	**83.8**

Table 1: Comparison of accuracies for DNN, Bayesian DNN MC Dropout and Stochastic Variational Inference (Stochastic VI) models applied to subset of MiT dataset (in-distribution classes).

samples. The selected dataset for both the categories include audio information.

We trained the ResNet101-C3D vision and VGGish audio architectures using the in-distribution MiT dataset, which includes ˜150K training and ˜5.3K validation samples. We select individual vision and audio paths from the model shown in Figure 2 to obtain single modality results. In the case of Bayesian DNN stochastic VI model, we perform multiple stochastic forward passes on the final three fully connected variational layers with Monte Carlo sampling on the weight posterior distributions. In our experiments, 40 forward passes provide reliable estimates above which the final results are not affected. Bayesian DNN model predictive mean is obtained by averaging the confidence estimates from the Monte Carlo sampling predictive distributions.

3.1. Uncertainty and confidence measures

We compare BALD uncertainty measure (details are in Appendix A) using in- and out-of-distribution classes from the subset of MiT dataset. The density histogram is a histogram with area normalized to one. The confidence measure density histogram plots for DNN model (Figure 5 (a)) indicate higher confidence for both in- and out-of-distribution classes. A peak is observed near higher confidence values for out-of-distribution samples indicating incorrect confidence predictions. The uncertainty estimates obtained from the Bayesian DNN models (Figure 5 (b) and (c)) indicate higher uncertainty for the out-of-distribution samples and lower uncertainty values for the in-distribution samples. A peak is observed near higher uncertainty values for out-of-distribution samples indicating reliable predictions.

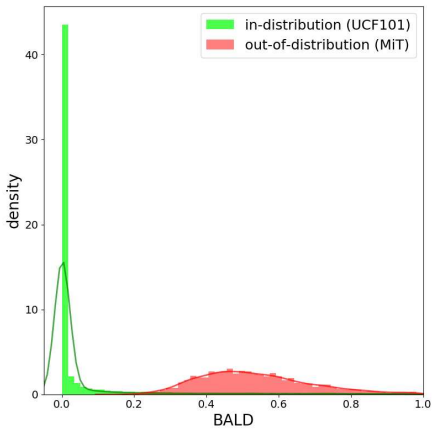


Figure 3: Density histogram of BALD uncertainty measure obtained from Bayesian DNN stochastic VI model.

3.2. Model performance comparison

The classification accuracy for MiT in-distribution samples is presented in Table 1. Bayesian DNN stochastic VI model consistently provides higher accuracies for individual and combined audio-vision modalities. Bayesian DNN stochastic VI audiovisual model provides an improvement of 9.2% top1 and 3.2% top5 accuracies over the Bayesian DNN visual model. Bayesian DNN stochastic VI model (audiovisual) provides an improvement of 2.8% top1 and 5.6% top5 accuracies over the baseline DNN model (audiovisual). The accuracies for Bayesian DNN MC dropout model are lower than the proposed Bayesian stochastic VI model.

Figure 4 shows the comparison of precision-recall (top) and ROC (bottom) plots using the confidence measures for DNN and Bayesian DNN models. It is observed from the plots that Bayesian DNN model consistently outperforms the DNN model for the individual modalities and also for the combined audiovisual modalities. The Precision-Recall AUC plot for the audiovisual Bayesian-DNN model shows an improvement of 10.2% over the audiovisual DNN model and an improvement of 9.5% over the vision only Bayesian DNN model.

We also compared the uncertainty estimates obtained from the proposed Bayesian DNN stochastic VI model using two separate datasets (UCF-101 as in-distribution and MiT as out-of-distribution). The comparison of uncertainty measures for in-distribution and out-of-distribution samples obtained from Bayesian DNN are shown in Figure 3. BALD density histogram indicates a clear separation of uncertainty estimates for in- and out-of-distribution samples.

These results confirm that the proposed Bayesian DNN stochastic VI model provides reliable confidence measure than the conventional DNN for the audiovisual activity recognition and can identify out-of-distribution samples.

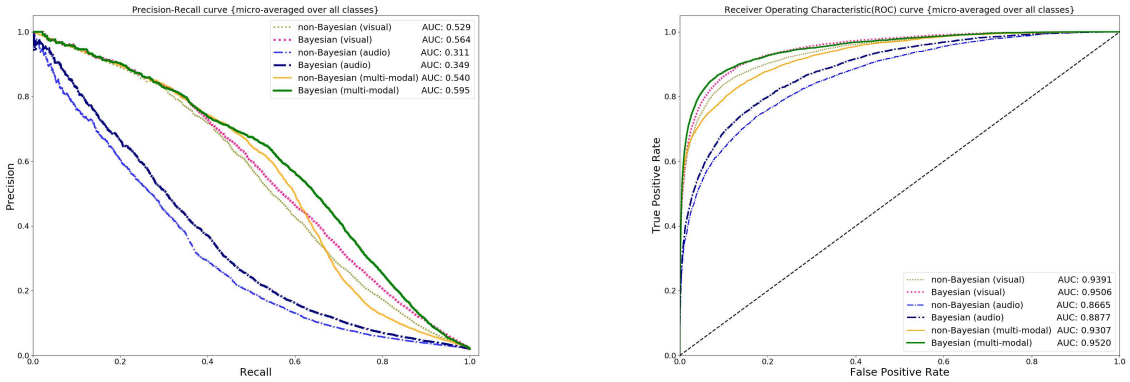
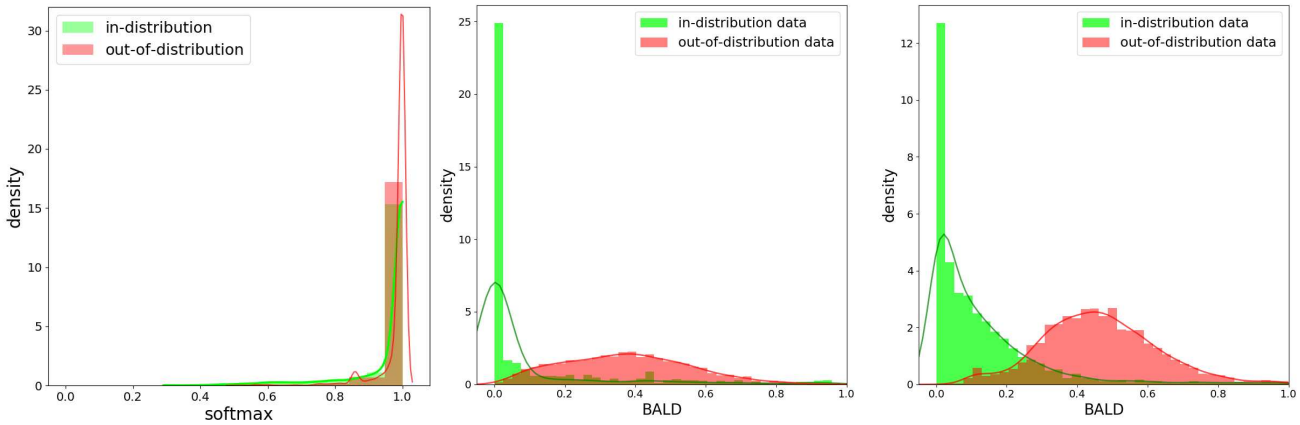


Figure 4: Precision-Recall (left) and ROC (right) plots micro-averaged over all the MiT in-distribution classes. The audiovisual Bayesian DNN model shows an improvement of 10.2% Precision-Recall AUC and 2.7% ROC AUC over the audiovisual DNN model.



(a) DNN confidence measure

(b) Bayesian DNN (MC Dropout) uncertainty measure

(c) Bayesian DNN (Stochastic VI) uncertainty measure

Figure 5: Density histograms obtained from in- and out-of-distribution samples for the subset of MiT dataset. (a) DNN confidence measure, (b) Bayesian DNN MC Dropout uncertainty measure and (c) Bayesian DNN (Stochastic VI) uncertainty measure.

References

[1] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. 1, 5

[2] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 2016. 1

[3] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79, 2012. 1

[4] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 1

[5] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. *arXiv preprint arXiv:1804.01665*, 2018. 1

[6] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *arXiv preprint arXiv:1804.03641*, 2018. 1

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 1

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 2

[9] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore,

Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017. 2

[10] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. *arXiv preprint arXiv:1401.0118*, 2013. 2, 5, 6

[11] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015. 2, 5, 6

[12] Christopher M Bishop. Pattern recognition and machine learning (information science and statistics) springer-verlag new york. *Inc. Secaucus, NJ, USA*, 2006. 2, 5, 6

[13] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018. 2, 6

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 2, 6

[15] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 2, 6

[16] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. 2, 6

[17] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 2

[18] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*, 2018. 2

[19] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011. 5

[20] Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011. 5

[21] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 5

[22] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999. 5

[23] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017. 5

Appendix A. Background

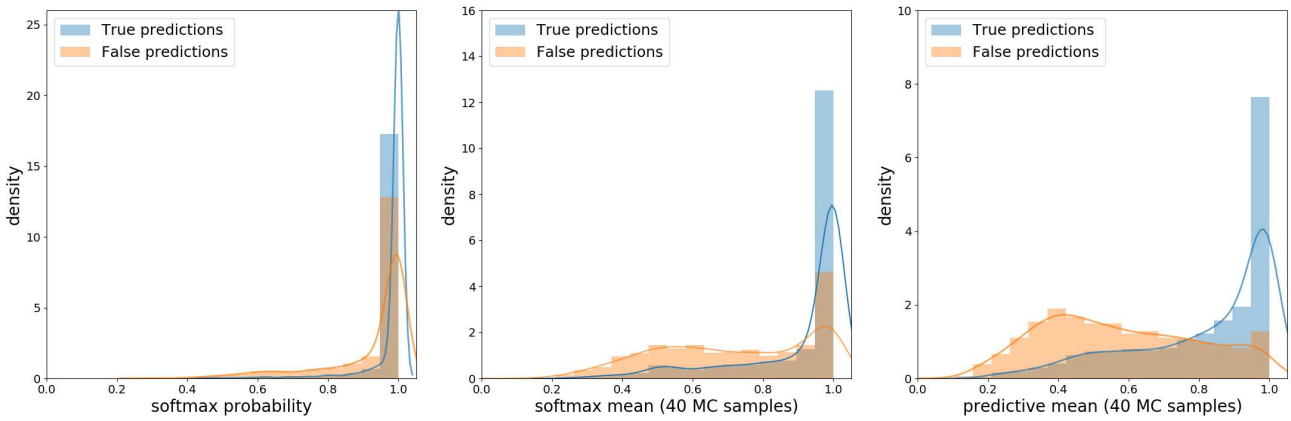
Given training dataset $D = \{x, y\}$ with inputs $x = x_1, ..., x_N$ and their corresponding outputs $y = y_1, ..., y_N$, in parametric Bayesian setting we would like to infer a distribution over parameters w as a function $y = f_w(x)$ that represents the DNN model. With the posterior for model parameters inferred during Bayesian neural network training, we can predict the output for a new data point by propagating over the model likelihood $p(y|x, w)$ while drawing samples from the learned parameter posterior $p(w|D)$. Computing the posterior distribution $p(w|D)$ is often intractable, some of the previously proposed techniques to achieve an analytically tractable inference include: (i) Markov Chain Monte Carlo (MCMC) sampling based probabilistic inference [1, 19] (ii) variational inference techniques to infer the tractable approximate posterior distribution around model parameters [10, 11, 20] and (iii) Monte Carlo dropout approximate inference [21].

Variational inference [22, 23] approximates a complex probability distribution $p(w|D)$ with a simpler distribution $q_\theta(w)$, parameterized by variational parameters θ while minimizing the Kullback-Leibler (KL) divergence [12]. Minimizing the KL divergence is equivalent to maximizing the log evidence lower bound [12, 21].

$$\mathcal{L} := \int q_\theta(w)\, log\, p(y|x, w)\, dw - KL[q_\theta(w)||p(w)] \tag{1}$$

Predictive distribution is obtained through multiple stochastic forward passes through the network during the prediction phase while sampling from the posterior distribution of network parameters through Monte Carlo estimators. Equation 2 shows the predictive distribution of the output y^* given new input x^*:

$$p(y^*|x^*, D) = \int p(y^*|x^*, w)\, q_\theta(w) dw$$
$$p(y^*|x^*, D) \approx \frac{1}{T} \sum_{i=1}^{T} p(y^*|x^*, w_i)\ , \quad w_i \sim q_\theta(w) \tag{2}$$



(a) DNN model (b) Bayesian DNN (MC Dropout) model (c) Bayesian DNN (Stochastic VI) model

Figure 6: Density histogram of confidence measures for subset of MiT dataset in-distribution true (correct) and false (incorrect) predictions:

where, T is number of Monte Carlo samples.

We evaluate the model uncertainty using Bayesian active learning by disagreement (BALD) [16] for the activity recognition task. BALD quantifies mutual information between parameter posterior distribution and predictive distribution, which captures model uncertainty, as shown in Equation 3.

$$BALD := H(y^*|x^*, D) - \mathbb{E}_{p(w|D)}[H(y^*|x^*, w)] \quad (3)$$

where, $H(y^*|x^*, D)$ is the predictive entropy given by:

$$H(y^*|x^*, D) = -\sum_{i=0}^{K-1} p_{i\mu}\, log\, p_{i\mu} \quad (4)$$

$p_{i\mu}$ is predictive mean probability of i^{th} class from T Monte Carlo samples and K is the total number of output classes.

Appendix B. Model training

The weights and bias parameters in the fully connected variational layers (shown in Figure 2) are modeled through mean-field Gaussian distribution, and the network is trained using Bayesian variational inference based on KL divergence [10, 11]. In order to learn the posterior distribution of model parameters w, we train Bayesian DNN with variational inference method. The objective is to optimize log evidence lower bound (ELBO) (Equation 1) as the cost function. The model parameters of the fully connected variational layers are parametrized by mean μ and variance σ^2, i.e. $q_\theta(w) = \mathcal{N}(w|\mu, \sigma^2)$. These parameters in the variational layers are optimized by minimizing the negative ELBO loss (L^v) [12]:

$$L^v = -\mathbb{E}_{q_\theta(w)}[log\, p(y|x, w)] + KL[q_\theta(w)||p(w)] \quad (5)$$

$$\mu_{i+1} \leftarrow \mu_i - \alpha\, \Delta_\mu L_i^v \qquad \sigma_{i+1} \leftarrow \sigma_i - \alpha\, \Delta_\sigma L_i^v$$

where, i is the training step, α is the learning rate, $\Delta_\mu L^v$ and $\Delta_\sigma L^v$ are gradients of the loss function computed w.r.t μ and σ, respectively. We use Flipout [13], which is an efficient method that correlates the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each input.

The parameters in deterministic layers are optimized using cross-entropy loss (L^d) [14] given by:

$$L^d = -\sum_c y_c \log \hat{y_c} \quad (6)$$

where, y_c and $\hat{y_c}$ are true and predicted label distributions, respectively. The model parameters for variational and deterministic DNN layers are obtained by applying stochastic gradient descent optimizer [15] to the loss functions given in Equation 5 and 6, respectively. During prediction stage we perform multiple Monte Carlo forward passes on the final variational layers by sampling the parameters from learned posteriors to measure uncertainty estimates using Equation 3.

Appendix C. Additional Results

The density histograms for the DNN confidence measure and Bayesian DNN uncertainty measure are plotted in Figure 6. In the case of false (incorrect) predictions, the DNN model still shows confidence measure density histograms peaked near 1.0. On the contrary, Bayesian DNN models show confidence measure density histograms skewed towards lower values indicating more reliable predictions. The proposed stochastic VI model shows a more pronounced peak towards lower values for false predictions indicating better predictive confidence measure than the MC dropout model.