Uncertainty-aware Audiovisual Activity Recognition using Deep Bayesian Variational Inference

Mahesh Subedar*, Ranganath Krishnan*, Paulo L. Meyer, Omesh Tickoo, Jonathan Huang

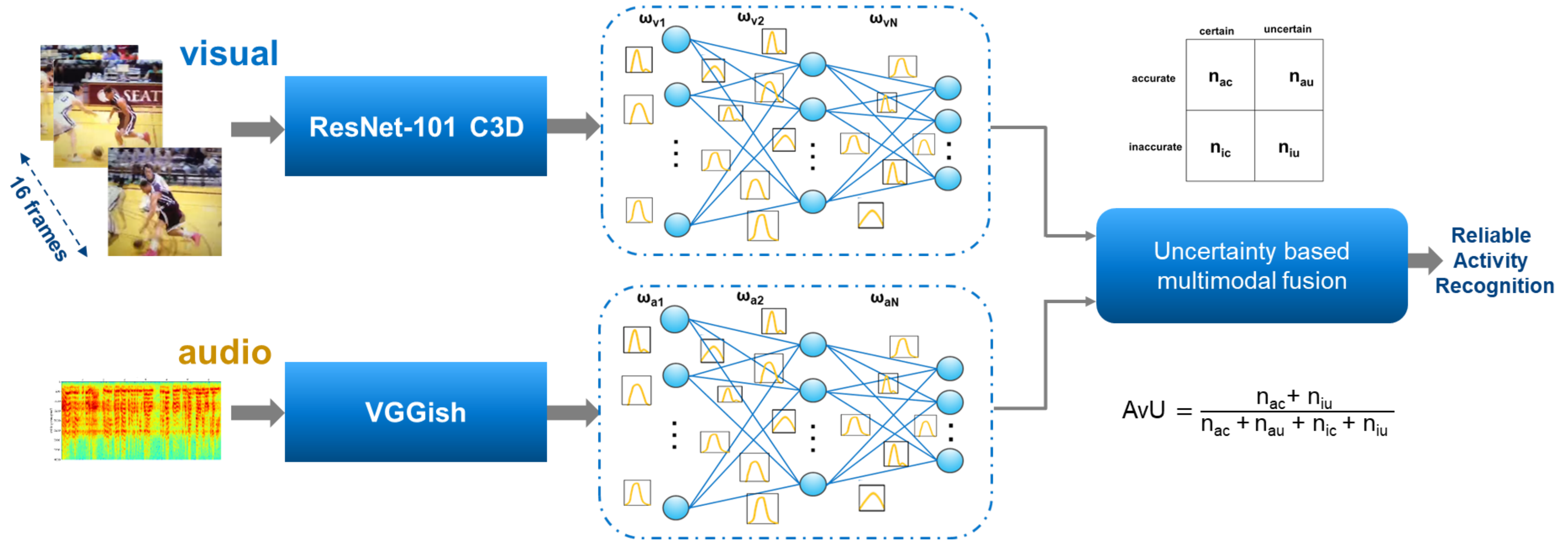
Intel Labs, USA

*equal contribution

ICCV 2019
Seoul, Korea

Bayesian Multi-modal fusion

Efficient multimodal fusion should intelligently understand the relative significance of each modality, or fallback to reliable modes of sensing. To design robust and reliable multimodal AI systems, it is essential to quantify uncertainty estimates from individual modalities in deep neural network (DNN) for effective multimodal fusion. We illustrate our proposed method applied to activity recognition with vision and audio modalities, and show it performs better than DNN baseline and Monte Carlo (MC) dropout method.



	certain	uncertain
accurate	n_{ac}	n_{au}
inaccurate	n_{ic}	n_{iu}

$$AvU = \frac{n_{ac} + n_{iu}}{n_{ac} + n_{au} + n_{ic} + n_{iu}}$$

$p(\omega_i \mid x,y) \approx q(\omega_i)$ → Gaussian mean-field variational distribution

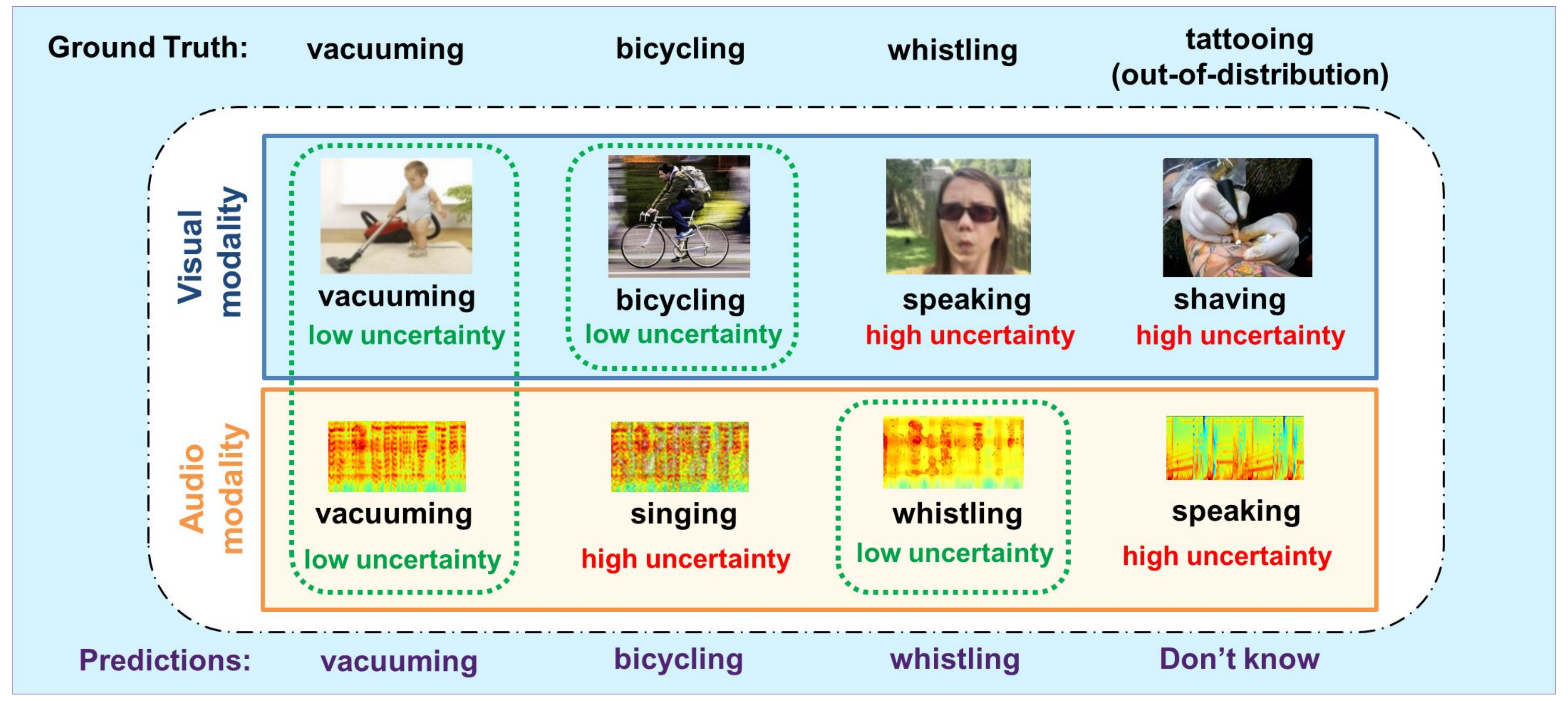
$q(\omega_i) = \mathcal{N}(\omega_i \mid \mu_i, diag(\sigma_i^2))$ $\mu_i, \sigma_i \in R^D$; i=1,2 ..N

Negative ELBO loss:

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

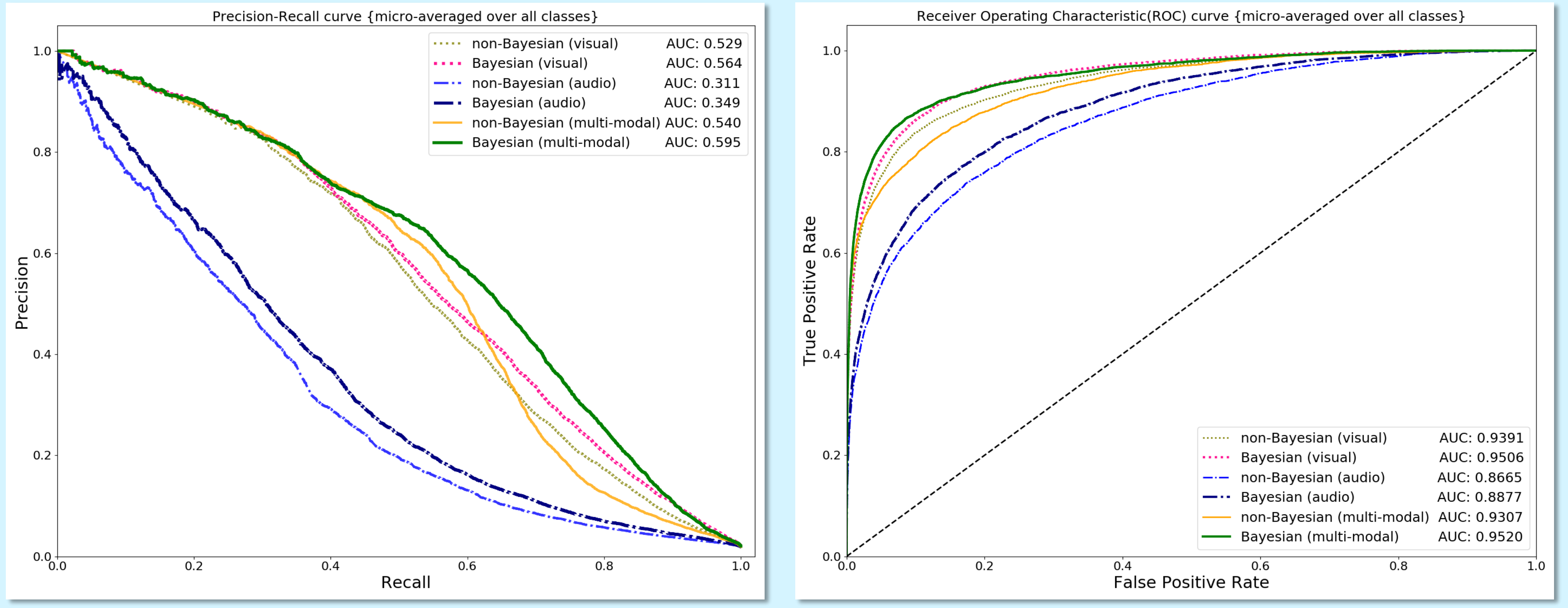
Contributions:

- ❑ Bayesian Multimodal fusion framework based on uncertainty estimates applied to activity recognition
- ❑ Demonstrate scalable stochastic variational inference (VI) in large-scale Bayesian DNN models on real-world tasks by combining deterministic and variational layers
- ❑ Identifying out-of-distribution data for activity recognition using uncertainty estimates



Uncertainty based multi-modal fusion (audiovisual activity recognition)

Results

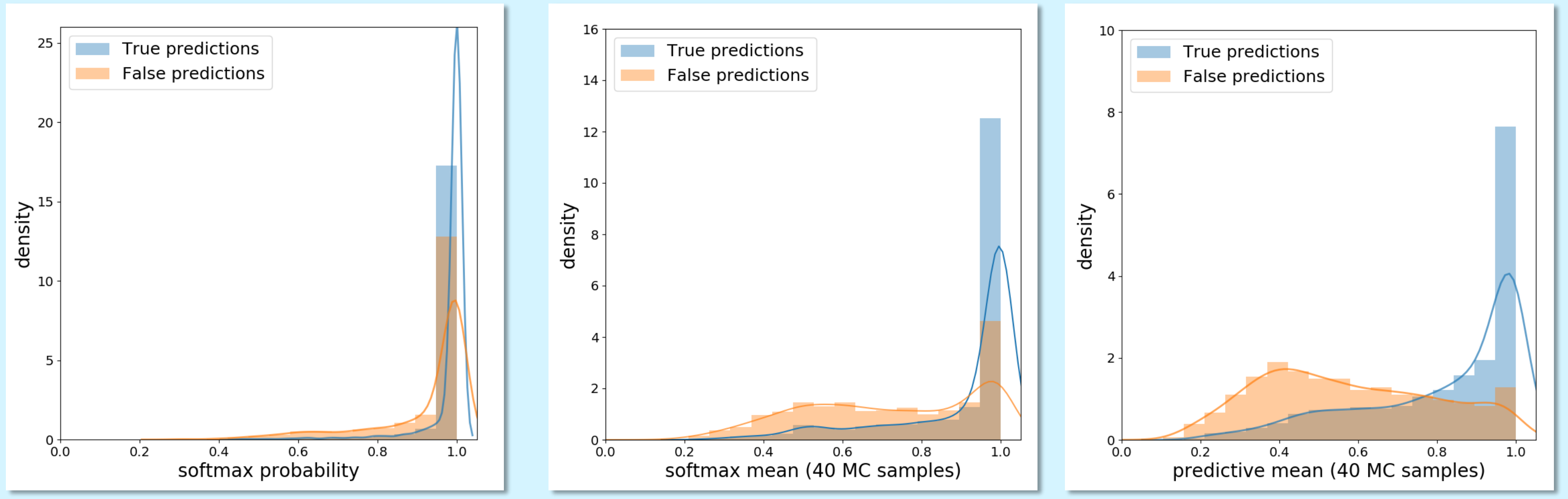


Comparison of Precision-Recall (left) and ROC (right) plots for DNN and Bayesian DNN models

DNN

Bayesian DNN (MC dropout)

Bayesian DNN (Stochastic VI)



Density histograms of confidence measures shows conventional DNN tend to be overconfident for false (incorrect) predictions, while Bayesian DNNs indicate lower confidence for false predictions signifying that Bayesian DNNs are transparent when they do not know.

- The proposed Bayesian multimodal fusion method achieves precision-recall AUC improvement of **10.2%** over the non-Bayesian baseline.
- Optimal uncertainty value that maximizes Accuracy vs. Uncertainty (AvU) metric from individual modalities is used for multimodal fusion.
- Our method indicates higher model uncertainty towards unseen activity classes, showcasing the capability to reliably identify out-of-distribution data through uncertainty quantification.

Comparison of validation accuracies on subset of Moments-in-Time (MiT) dataset (in-distribution data)

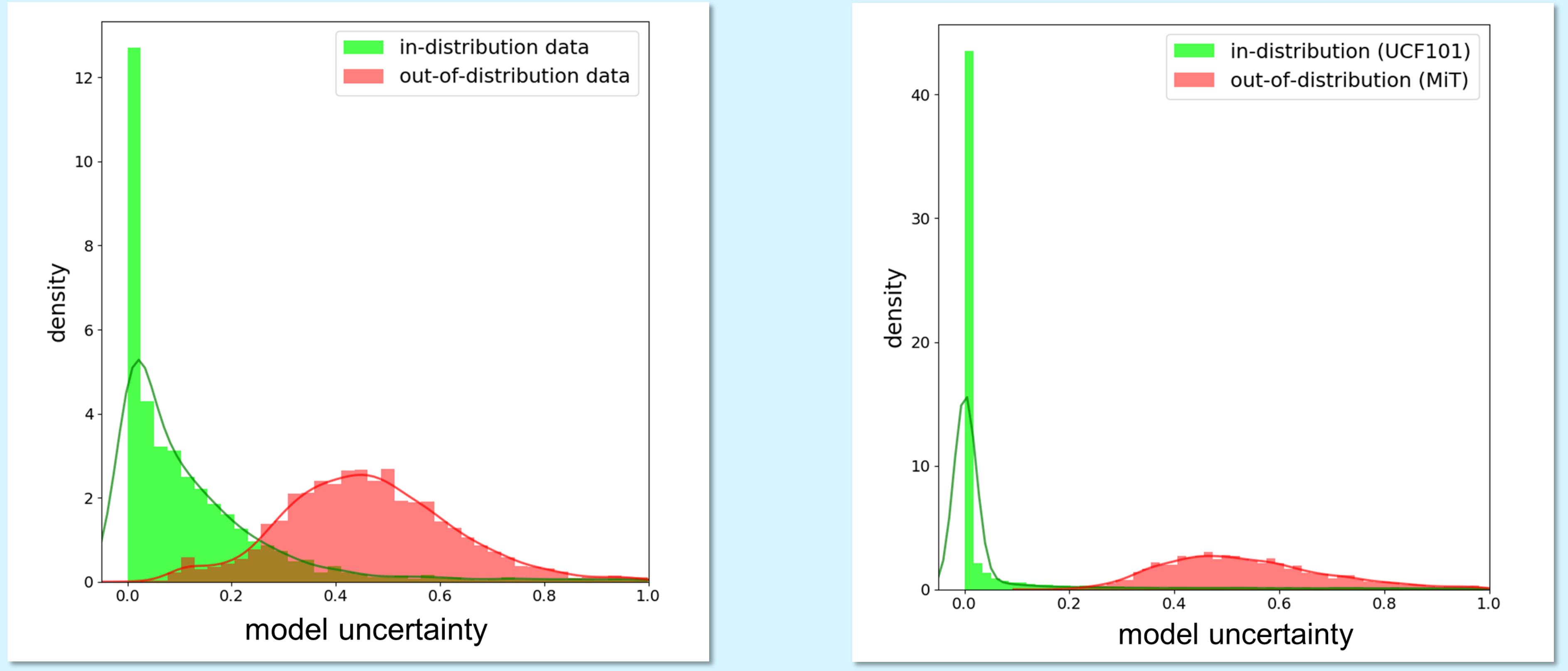
Model	Vision		Audio		Audiovisual	
	Top 1 (%)	Top 5 (%)	Top 1 (%)	Top 5 (%)	Top 1 (%)	Top 5 (%)
DNN	52.65	79.79	34.13	61.68	56.61	79.39
Bayesian DNN (MC Dropout)	52.28	80.10	32.46	60.97	55.04	80.34
Bayesian DNN (Stochastic VI)	53.3	81.20	35.80	63.40	**58.2**	**83.8**

---- proposed method: Multimodal fusion based on uncertainty estimates from Bayesian DNN (Stochastic VI)

Bayesian DNN (Stochastic VI)

in- and out-of-distribution subsets from MiT

in- and out-of-distribution (UCF101 and MiT)



Density histograms of model uncertainty estimates [BALD (Bayesian active learning by disagreement)]. Shows higher model uncertainty for out-of-distribution data, signifying that proposed method have the capability of expressing "don't know" when not confident about predictions on unseen activity classes.

Conclusions

- Reliable uncertainty quantification in Bayesian DNNs benefits multimodal fusion. The precision-recall AUC improvement for audiovisual activity recognition with our proposed method is attributed to efficient multimodal fusion.
- The uncertainty estimates obtained by combining deterministic and variational layers in Bayesian DNNs can reliably identify out-of-distribution data.
- The proposed Bayesian multimodal fusion framework can be extended to other real-world multimodal applications.