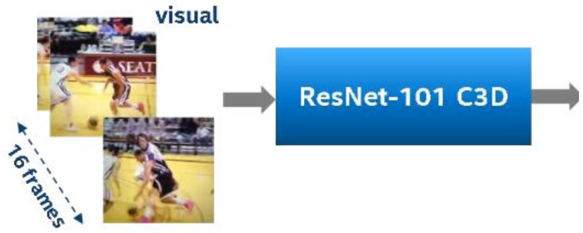
Uncertainty-aware Audiovisual Activity Recognition using Deep Bayesian Variational Inference

Mahesh Subedar*, Ranganath Krishnan*, Paulo Lopez Meyer, Omesh Tickoo, Jonathan Huang
{mahesh.subedar, ranganath.krishnan, paulo.lopez.meyer, omesh.tickoo, jonathan.huang}@intel.com
Intel Labs

Activity Recognition



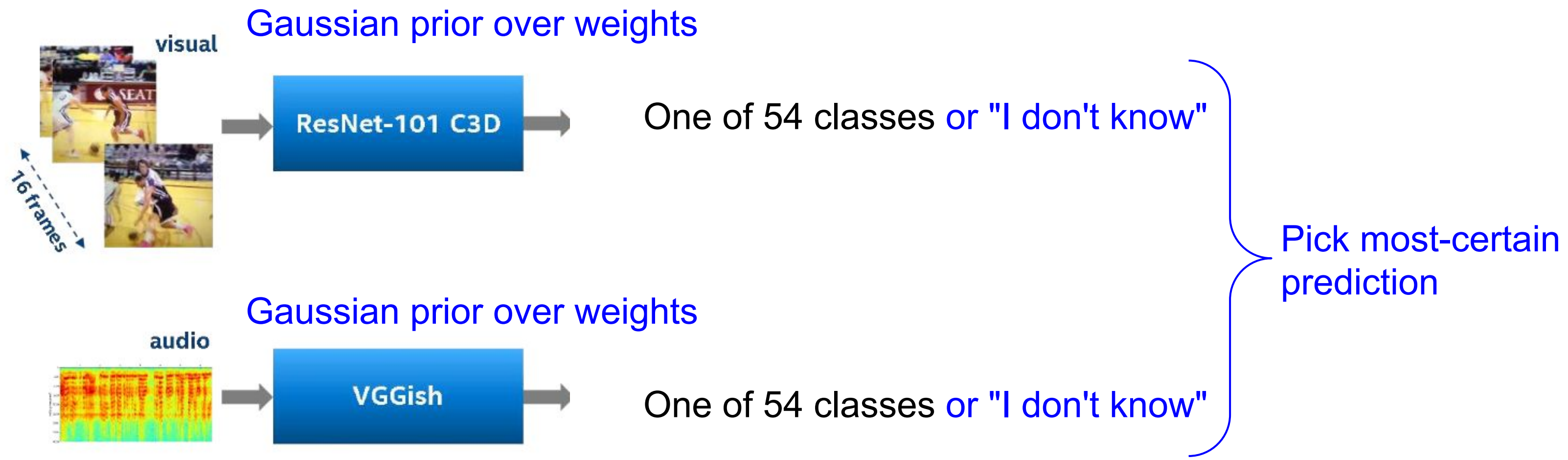
One of 54 classes

Activity Recognition using Deep Bayesian Variational Inference

Gaussian prior over weights

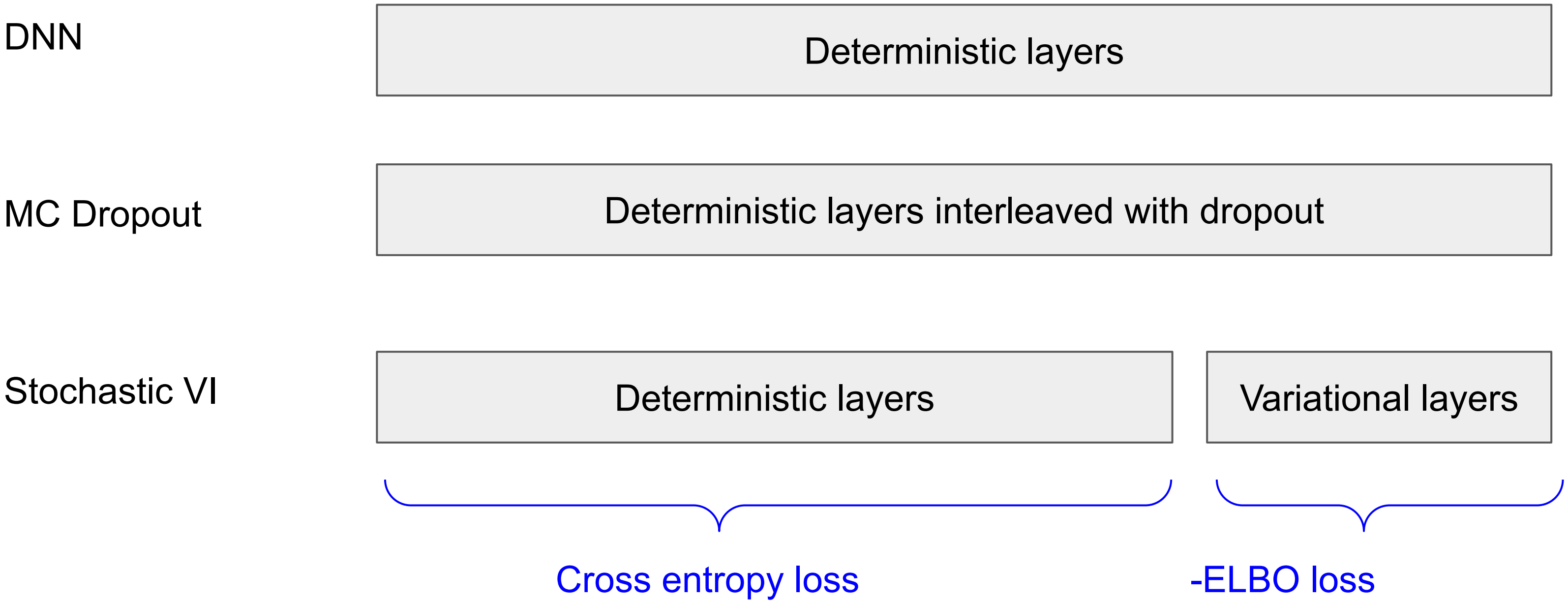
One of 54 classes or "I don't know"

Uncertainty-aware Audiovisual Activity Recognition using Deep Bayesian Variational Inference



Original dataset: 339 classes

This paper: 54 + 54 classes


DNN
Deterministic layers
MC Dropout
Deterministic layers interleaved with dropout
Stochastic VI
Deterministic layers
Variational layers
Cross entropy loss
-ELBO loss

Likelihood : Multinomial(NeuralNet(x, w))

???

Prior : Gaussian(0, α)

$$p(w|D) = \frac{p(y|x,w)p(w)}{p(y|x)} \tag{1}$$

Likelihood : Multinomial(NeuralNet(x, w))

???

Prior : Gaussian(0, α)

$$p(w|D) = \frac{p(y|x, w)p(w)}{p(y|x)} \tag{1}$$

SVI: Let $p(w|D) = \prod_i q(w_i|D) = \prod_i \text{Gaussian}(\mu_i, \sigma_i)$

Flipout: different noise-masks within batch

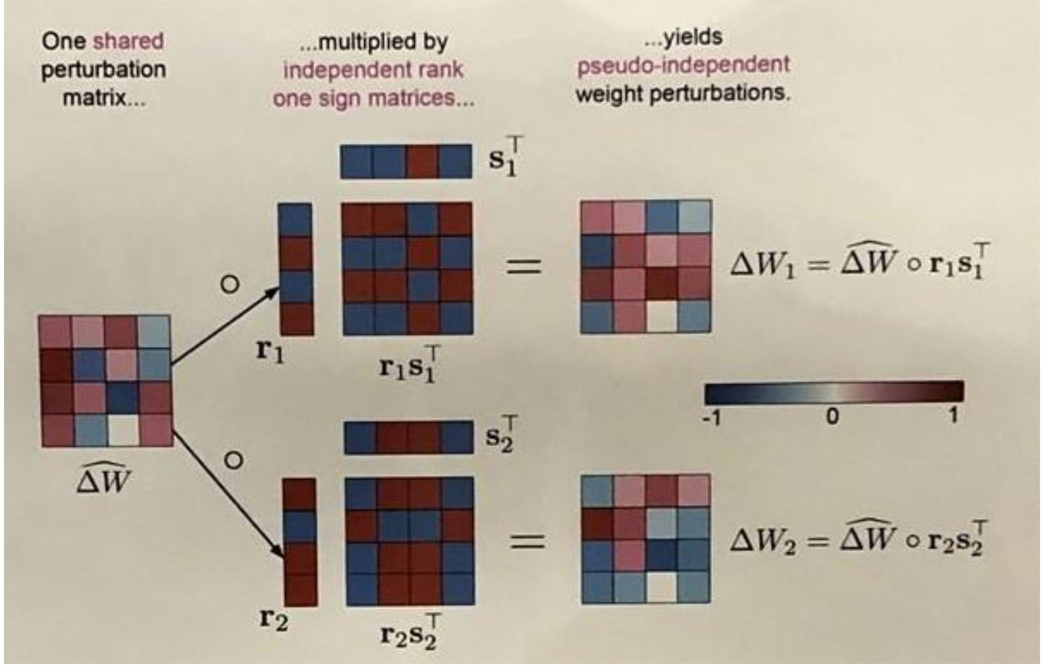
Problem:

SVI => Multiply with Gaussian Noise

Mask shared within batch => high variance

Solution:

Flipout yields lower variance estimates



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 3 shows the predictive distribution of the output y^* given new input x^*:

$$
\begin{aligned}
p(y^*|x^*, D) &= \int p(y^*|x^*, w)\, q_\theta(w) dw \\
p(y^*|x^*, D) &\approx \frac{1}{T} \sum_{i=1}^{T} p(y^*|x^*, w_i)\,, \quad w_i \sim q_\theta(w)
\end{aligned} \tag{3}
$$

1. Sample a $\mathbf{w_i}$ from distribution
2. $\mathbf{y_i}$ = forward pass with $\mathbf{w_i}$
3. Repeat **T** times and average

In [12, 26], modeling aleatoric and epistemic uncertainty is described. We evaluate the epistemic uncertainty using Bayesian active learning by disagreement (BALD) [21] for the activity recognition task. BALD quantifies mutual information between parameter posterior distribution and predictive distribution, as shown in Equation 4.

$$BALD := \underbrace{H(y^*|x^*, D)}_{\text{Entropy of prediction}} - \mathbb{E}_{p(w|D)}[\underbrace{H(y^*|x^*, w)}_{\text{Entropy of sample } \mathbf{w}}] \quad (4)$$

$$\mathrm{AvU} = \frac{n_{ac} + n_{iu}}{n_{ac} + n_{au} + n_{ic} + n_{iu}} \tag{8}$$

Correct prediction made

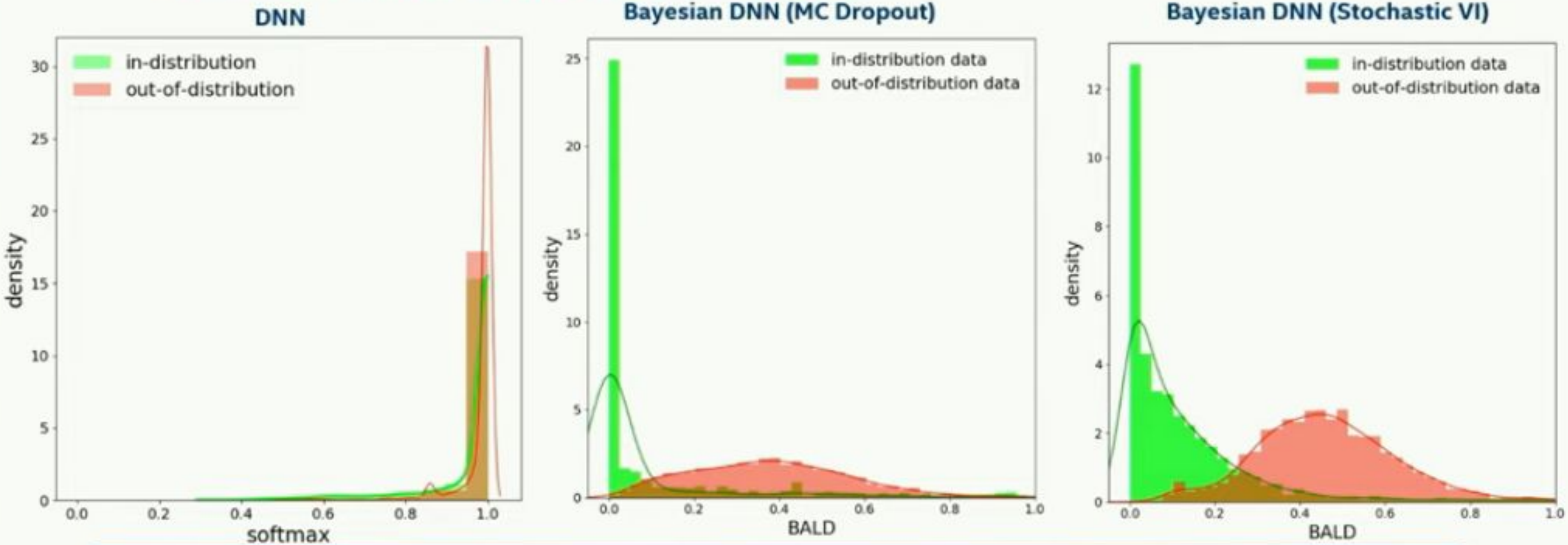
	certain	uncertain
accurate	n_{ac}	n_{au}
inaccurate	n_{ic}	n_{iu}

Wrong prediction avoided

Figure 4: Accuracy vs Uncertainty confusion matrix

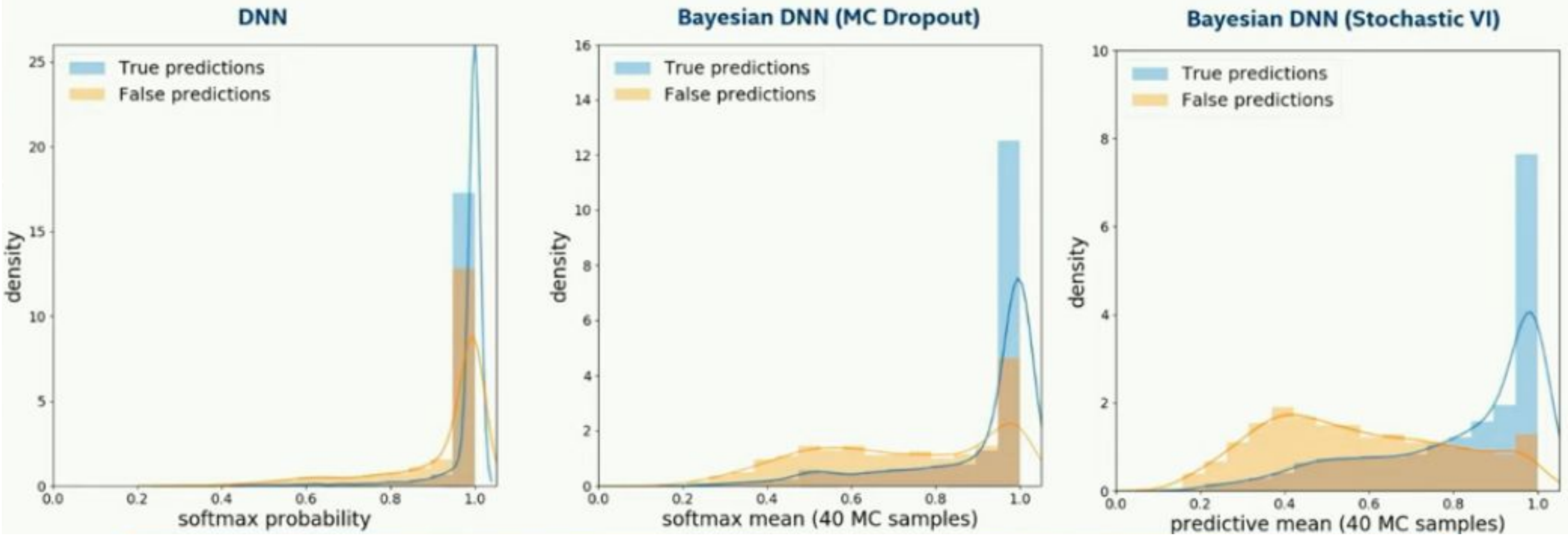
Results: Out-of-distribution detection

Dataset: Moments-in-Time (54 classes) in-distribution data
Moments-in-Time (54 classes) out-of-distribution data



Results: Confidence measures

Dataset: Moments-in-Time (54 classes) in-distribution data



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).