Improving model calibration with accuracy versus uncertainty optimization

Ranganath Krishnan, Omesh Tickoo (Intel Labs)

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- Obtaining reliable and accurate uncertainty quantification from deep neural networks is important to build robust AI systems
- A well-calibrated model should be accurate when it is certain about its prediction and indicate high uncertainty when it is likely to be inaccurate
- Uncertainty calibration is a challenging problem as there is no ground truth available for uncertainty estimates
- Our solution: Accuracy versus Uncertainty Calibration (AvUC)

Accuracy versus Uncertainty Calibration

- We propose an optimization method that leverages the relationship between accuracy and uncertainty as anchor for uncertainty calibration in deep neural network classifiers (Bayesian and non-Bayesian)
- We propose differentiable approximation to *accuracy vs uncertainty* (AvU) measure [Mukhoti and Gal 2018] and introduce trainable *accuracy vs uncertainty calibration* (AvUC) loss function. In this work, AvU utility function is optimized during training for obtaining well-calibrated uncertainties along with improved accuracy

		Uncertainty: certain	Uncertainty: uncertain
Accuracy	accurate	AC	AU
	inaccurate	IC	IU

$$\mathcal{L}_{\text{AvUC}} := \log\left(1+\frac{n_{AU}+n_{IC}}{n_{AC}+n_{IU}}\right)$$

where;

$$n_{AU} = \sum_{i \in \left\{\begin{smallmatrix}\tilde{y}_i = y_i \text{ and} \\ u_i > u_{\text{th}}\end{smallmatrix}\right\}} p_i \odot \tanh(u_i) \quad ; \quad n_{IC} = \sum_{i \in \left\{\begin{smallmatrix}\tilde{y}_i \neq y_i \text{ and} \\ u_i \leq u_{\text{th}}\end{smallmatrix}\right\}} (1-p_i) \odot (1-\tanh(u_i))$$

$$n_{AC} = \sum_{i \in \left\{\begin{smallmatrix}\tilde{y}_i = y_i \text{ and} \\ u_i \leq u_{\text{th}}\end{smallmatrix}\right\}} p_i \odot (1-\tanh(u_i)) \quad ; \quad n_{IU} = \sum_{i \in \left\{\begin{smallmatrix}\tilde{y}_i \neq y_i \text{ and} \\ u_i > u_{\text{th}}\end{smallmatrix}\right\}} (1-p_i) \odot \tanh(u_i)$$

- We use AvUC loss as an additional utility-dependent penalty term to accomplish the task of improving uncertainty calibration relying on the theoretically sound loss-calibrated approximate inference framework [Lacoste-Julien et al. 2011, Cobb et al. 2018]
- Loss-calibrated evidence lower bound (ELBO) in stochastic variational inference (SVI) given by equation below. We refer to this method as **SVI-AvUC**.

$$\mathcal{L} := \underbrace{\underbrace{-\mathbb{E}_{q_\theta(w)}[\log p(y|x,w)]}_{\text{expected negative log likelihood}} + \underbrace{\mathrm{KL}[q_\theta(w)||p(w)]}_{\text{Kullback-Leibler divergence}}}_{\mathcal{L}_{\text{ELBO}}\text{(negative ELBO)}} + \beta \underbrace{\log\left(1+\frac{n_{AU}+n_{IC}}{n_{AC}+n_{IU}}\right)}_{\mathcal{L}_{\text{AvUC}}\text{(AvUC loss)}}$$

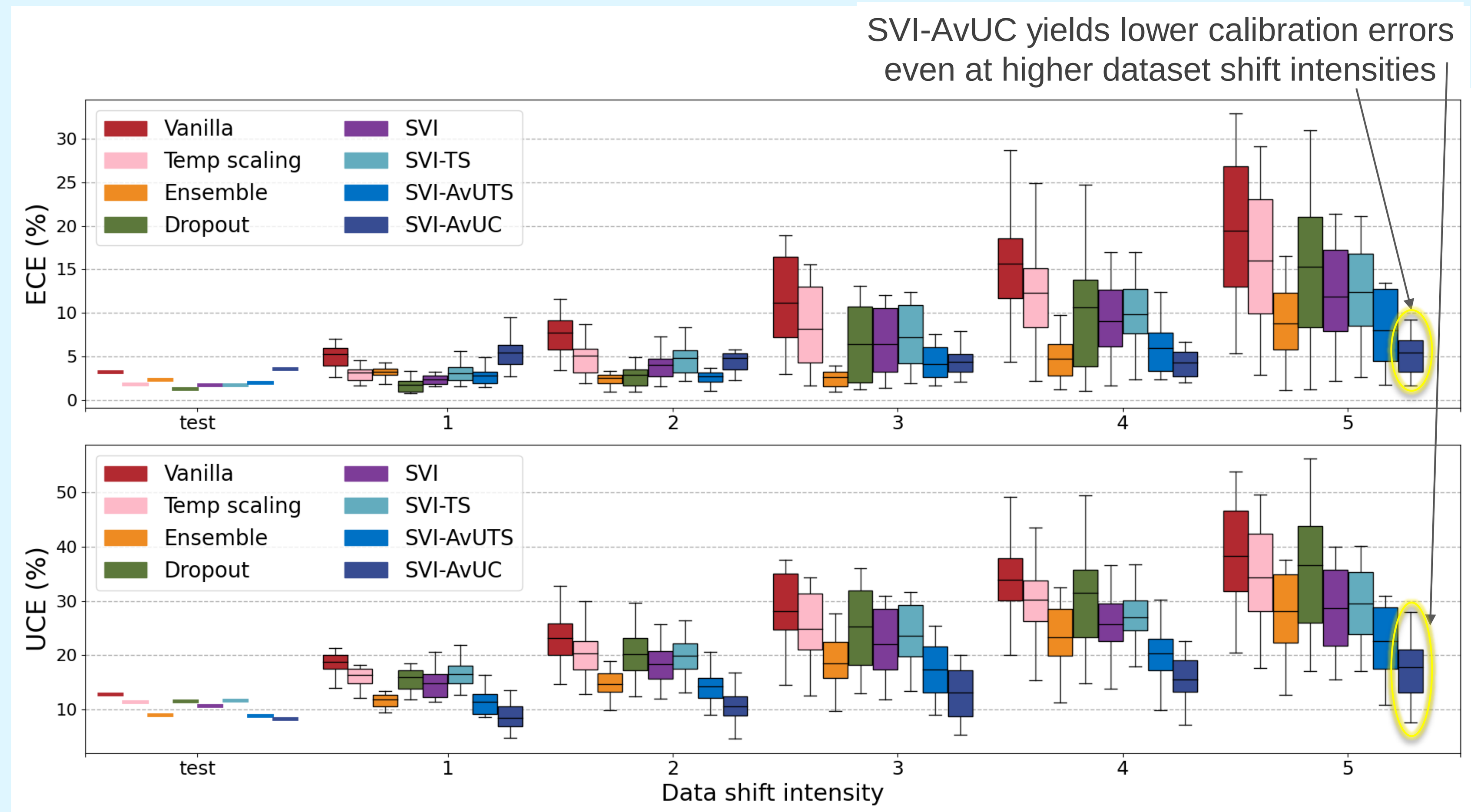
- We also propose a simple post-hoc uncertainty calibration on pretrained models with temperature scaling [Guo et al. 2017] by replacing negative log-likelihood loss with AvUC loss. We refer to this method applied to pretrained SVI model as **SVI-AvUTS**.

Experiments and Results

- We perform empirical evaluation of proposed methods SVI-AvUC and SVI-AvUTS on large-scale image classification task under distributional shift, comparing with various high performing Bayesian and non-Bayesian methods
- We evaluate the model calibration; model performance with respect to confidence and uncertainty estimates; and the distributional shift detection performance
- Dataset/Models: ImageNet/ResNet-50; CIFAR10/ResNet-20, SVHN (OOD)

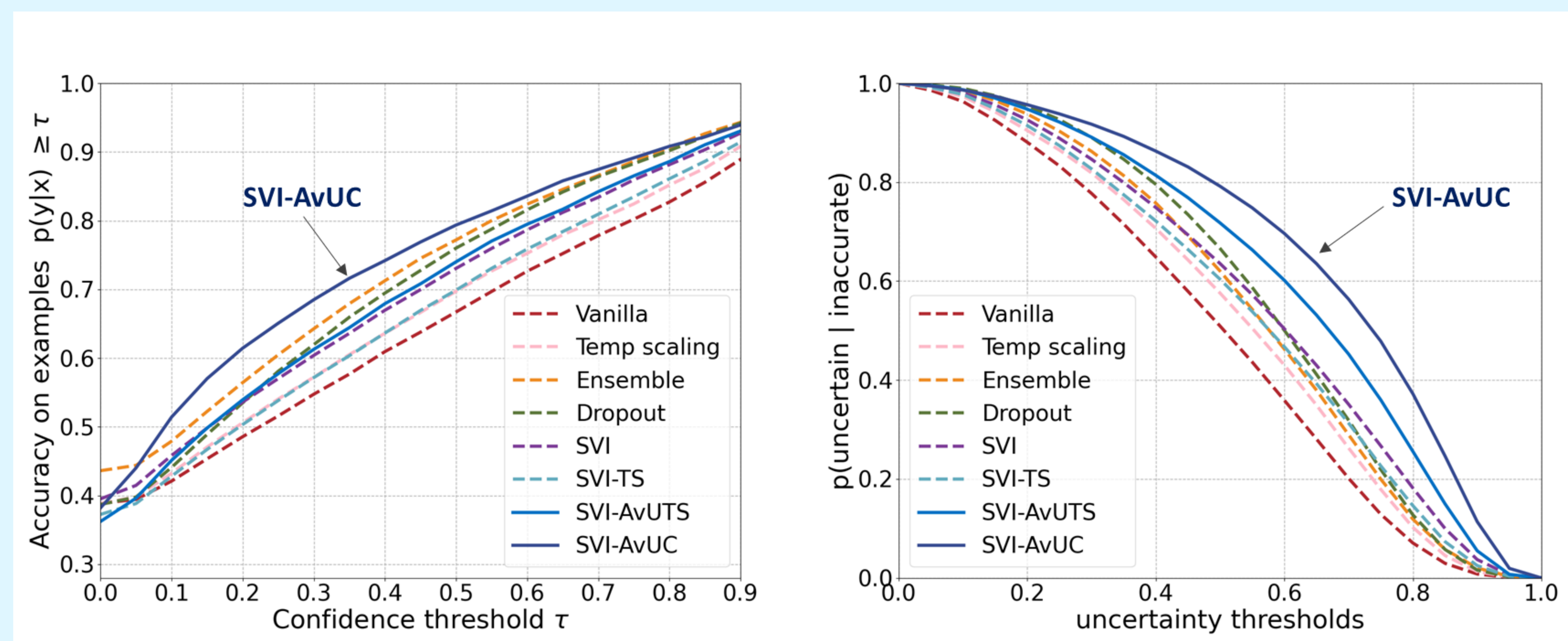
  ImageNet-C and CIFAR10-C [Hendrycks and Dietterich 2019] for evaluation of model calibration under dataset shift [Snoek et al. 2019].

Model calibration under dataset shift



ImageNet: Model calibration comparison using ECE and UCE on test data and under dataset shift. At each shift intensity level, the boxplot summarizes the results across 16 different dataset shift types. A well-calibrated model should provide lower calibration errors even at increased dataset shift. Similar observation on CIFAR10.

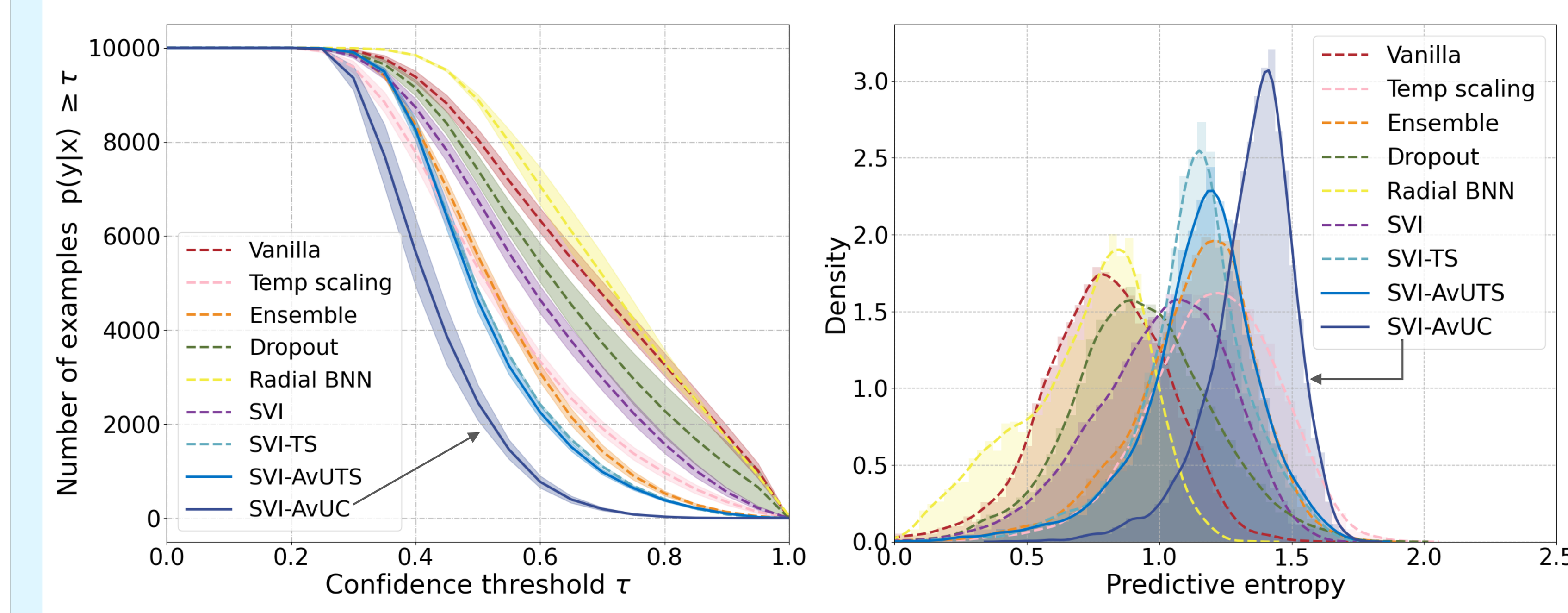
Model performance wrt confidence and uncertainty estimates



Confidence vs Accuracy (↑) Probability of being uncertain given inaccurate (↑)

ImageNet: SVI-AvUC is more accurate at higher confidence and more uncertain when making inaccurate predictions under distributional shift, compared to other methods.

Model reliability towards out-of-distribution (OOD) data



Count vs Confidence (↓) Predictive uncertainty density histogram

OOD evaluation with SVHN data on the model trained with CIFAR10. SVI-AvUC has lesser number of examples with higher confidence and provides higher predictive uncertainty estimates on out-of-distribution data.

Distributional shift detection performance

Method	ImageNet (Dataset shift detection) AUROC ↑	Detection accuracy↑	AUPR ↑	CIFAR10 (OOD detection) AUROC ↑	Detection accuracy↑	AUPR ↑
Vanilla	93.36	86.08	92.82	96.53	91.60	97.23
Temp scaling	93.71	86.47	93.21	96.65	92.14	97.39
Ensemble	95.49	88.82	95.31	95.78	91.47	96.95
Dropout	96.38	89.98	96.16	91.48	86.84	93.99
SVI	96.40	90.03	95.97	93.94	87.87	95.30
SVI-TS	96.61	90.45	96.24	90.81	87.59	93.84
SVI-AvUTS	96.89	90.93	96.58	93.79	89.39	95.49
SVI-AvUC	**97.60**	**92.07**	**97.39**	**99.35**	**97.16**	**99.50**

Distributional shift detection using predictive uncertainty estimates. For dataset shift detection on ImageNet, test data corrupted with Gaussian blur of intensity level 5 is used. SVHN is used as out-of-distribution (OOD) data for OOD detection on model trained with CIFAR10. SVI-AvUC outperforms across all the metrics.

Conclusion

- We proposed novel optimization methods AvUC and AvUTS for improving uncertainty calibration in deep neural networks
- We introduced a trainable uncertainty calibration loss that can be used as an additional utility-dependent penalty term and combined with existing losses
- Uncertainty calibration is important for reliable and informed decision making in safety critical applications, we envision AvUC as a step towards advancing probabilistic deep neural networks in providing well-calibrated uncertainties along with improved accuracy
- We empirically demonstrated the proposed method yield state-of-the-art model calibration under increased dataset shift and outperforms in distributional shift detection
- Code available at https://github.com/IntelLabs/AVUC