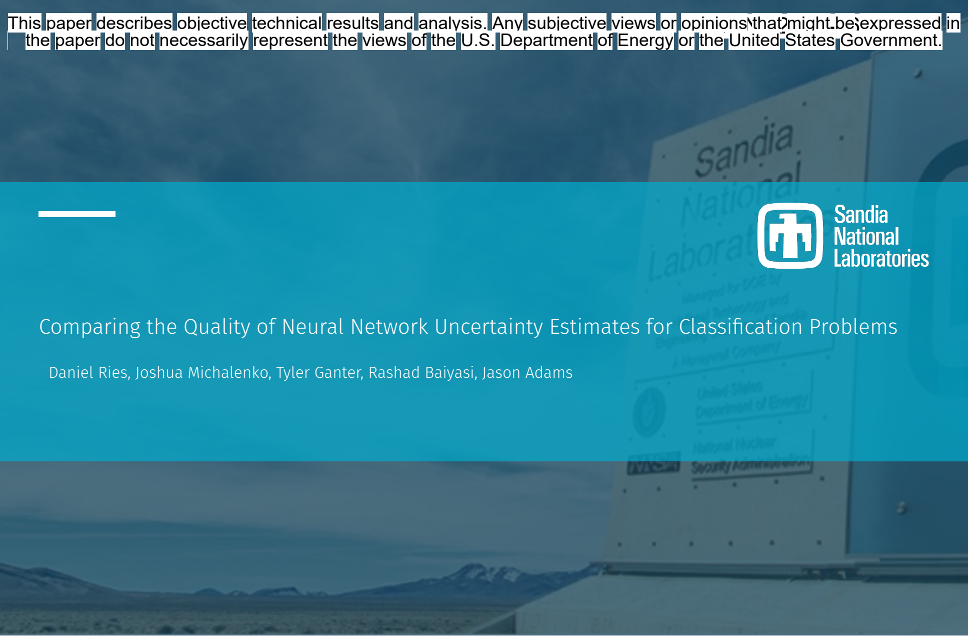


Comparing the Quality of Neural Network Uncertainty Estimates for Classification Problems

Daniel Ries, Joshua Michalenko, Tyler Ganter, Rashad Baiyasi, Jason Adams

This work was supported by the Laboratory Directed Research and Development programat Sandia National Labora-

Security Administration under contract DE-NA0003525.

boratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LL
Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE

Introduction

- Standard deep learning (DL) methods give predictions without associated measures of uncertainty.
- Developing methods for including uncertainty estimates along with the powerful prediction capabilities of DL is currently an active and important area of research.
- High-quality uncertainty quantification (UQ) is a critical aspect of applying DL methods to high-consequence applications. Our paper investigates the quality of the UQ produced by several UQ-enabled DL methods.

Motivating Application

- Our motivating application for this work is the analysis of a high-fidelity simulated hyperspectral image data set. We are interested in detecting small targets (green discs of varying size) that have been placed throughout the images.
- Our models output the estimated probability that a pixel contains target for each pixel in the image. The UQ-enabled DL methods used also allow us to construct credible intervals (CIs) for the class probabilities at each pixel.

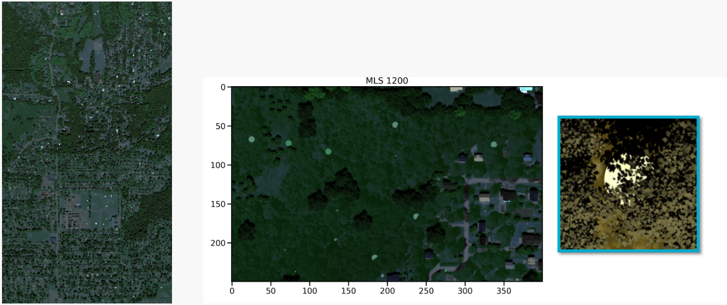


Figure: Example pseudo color rendering of image with green discs (left), zoomed-in region (center), and a single disc partially obstructed from view (right).

Motivating Application

UQ-enabled DL methods used:

- Bayesian neural network (BNN) trained using Markov Chain Monte Carlo (MCMC). Denoted as BNN-MCMC hereafter.
- BNN trained using Variational Inference (VI). Denoted as BNN-VI.
- Deep ensemble with 100 ensembles, denoted as DE.
- Bootstrap neural network - computationally the same as DE except that the training data for each ensemble is a bootstrap resample of the original training data - with 100 ensembles. Denoted as Bootstrap.
- Monte Carlo dropout with 100 ensembles, denoted as MC dropout.

The network architecture for each of these methods was a fully connected, two-layer network with 10 nodes per layer.

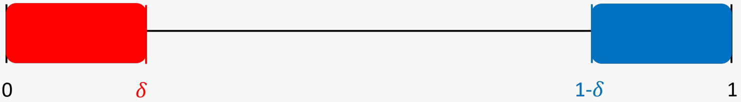
High Confidence Sets

- We want to know for which pixels the model is highly confident it has made the correct prediction. To do this, we used the high confidence sets (HCS's) introduced in Ries et al. [2022].
- The HCS is defined as

$$\Omega = \{i : (\mathcal{B}_{\pi_i}(\alpha)_{LB} > 1 - \delta \cup \mathcal{B}_{\pi_i}(\alpha)_{UB} < \delta)\}$$

where $\mathcal{B}_{\pi_i}(\alpha)_{LB}$ and $\mathcal{B}_{\pi_i}(\alpha)_{UB}$ are the lower and upper bounds of a $(1-\alpha)\%$ CI for π_i, and δ is a probability threshold. We use $\alpha = \delta = 0.2$.

A given pixel is in the HCS if the associated $(1-\alpha)\%$ CI lower bound is in the blue region or if the upper bound is in the red region.



- Daniel Ries, Jason Adams, and Joshua Zollweg. Target detection on hyperspectral images using MCMC and VI trained Bayesian neural networks. *Proceedings of the IEEE Aerospace Conference*, 2022.

Motivating Application Results

Method	Proportion of Pixels in HC Set
BNN-MCMC	0.81
BNN-VI	0.27
DE	0.71
Bootstrap	0.78
MC Dropout	0.74

Table: Proportion of test set pixels in HCS for Megascene for each model.

- Our test image contained over 1.5 million pixels. Thus the 10% difference between BNN-MCMC and DE corresponds to a difference in HCS size of roughly 150,000 pixels.
- While HCS's can be useful in reducing analyst burden, they rely on high-quality UQ. The natural question arising from these results is which of these models is producing the highest quality UQ?

UQ Quality

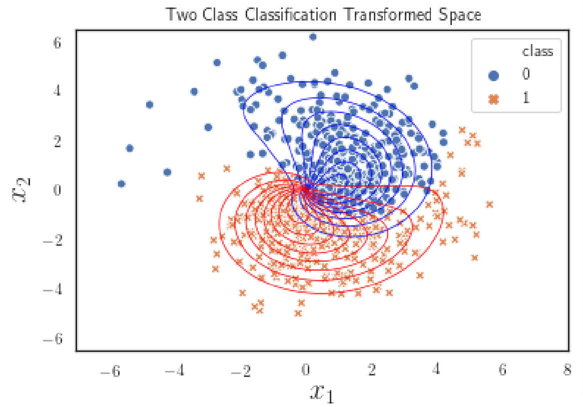
Existing metrics for measuring UQ quality:

- Interval coverage - this assesses the central claim of a CI. Given a large number of samples, a $(1-\alpha)\%$ CI should contain the true value (class probability in our case) in $(1-\alpha)\%$ of the samples.
- Interval width - if two methods both produce intervals with adequate coverage, we say the method producing the narrower intervals is better.
- Expected calibration error (ECE) - measures the difference between model accuracy and estimated probability values over bins that discretize the interval $[0, 1]$.

While ECE can be computed for any given classification model, interval coverage requires knowledge of the ground truth class probabilities to compute. Interval width without interval coverage is not meaningful. For this reason, we investigate our methods on a *simulated* data set where the ground truth class probabilities are known.

Simulated Two-class Classification Data Set

Our two-dimensional two-class classification (TCC) data set is simulated from a transformed Gaussian mixture. The key aspect of this data set is that ground-truth class probabilities are known at each (x_1, x_2) pair. Thus all of our UQ quality metrics can be computed.



UQ Quality Metrics Experiment

- All of the same methods, including the same network architecture, used on the motivating example were trained on the TCC data. Additionally, we trained a Gaussian process classification model as a non-DL comparison.
- 100 instances of TCC were simulated and each of the methods were trained on each instance. Estimated class probabilities and 90% CI's were obtained in each instance on a grid of points over the input space.
- From the estimated class probabilities, interval coverage, interval width, and ECE metrics were computed. Metrics are reported as average values over all grid points in all 100 TCC instances. Monte Carlo standard error is also given.

Simulation Results - Estimated Probabilities

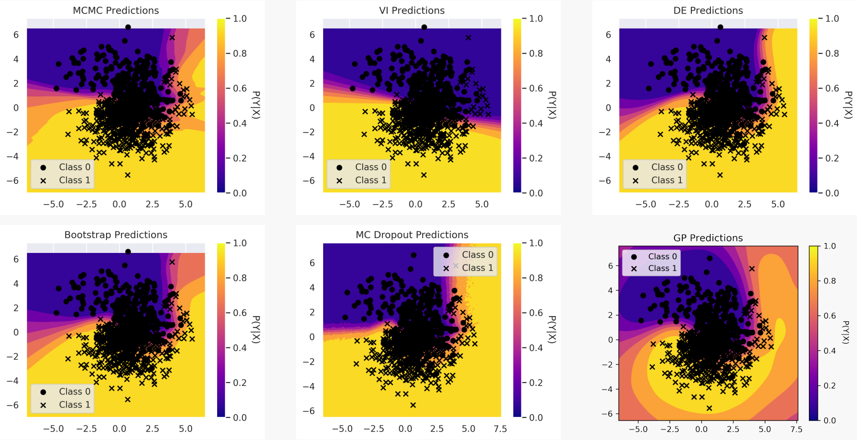


Figure: Prediction surfaces for each model on one TCC simulation.

Simulation Results - UQ

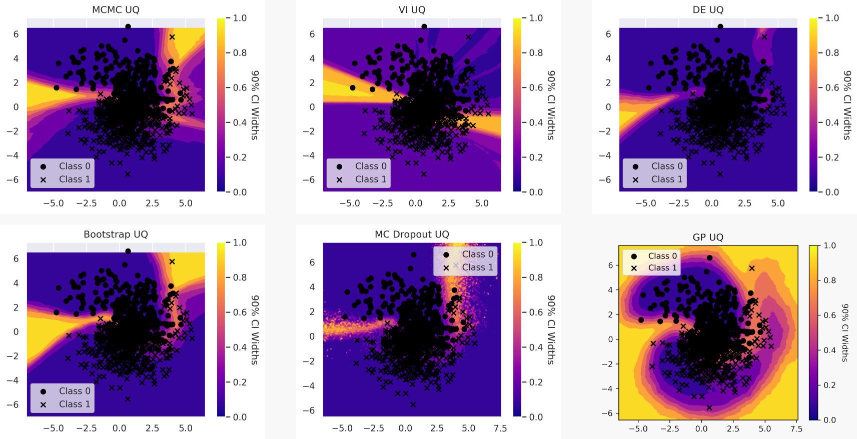


Figure: Uncertainties for each model via 90% prediction interval widths on one TCC simulation.

Simulation Results

Method	Coverage	Width	ECE
BNN-MCMC	**0.91 (0.04)**	$\mathbf{0.22(0.01)}^*$	**0.04 (0.01)**
BNN-VI	0.59 (0.17)	0.38 (0.07)	0.08 (0.02)
DE	0.48 (0.09)	0.09 (0.01)	**0.04 (0.01)**
Bootstrap	0.84 (0.06)	0.25 (0.02)	**0.04 (0.01)**
MC Dropout	0.67 (0.08)	0.15 (0.02)	**0.04 (0.01)**
GP	0.98 (0.02)	0.36 (0.02)	0.05 (0.01)

Table: TCC Simulation results. Bolded values indicate best metric in each column. The asterisk indicates the best interval width, given the nominal coverage was met (nominal rate = 0.9).

Conclusions

In making these conclusions, we first note that the simulated data and models implemented are fairly simplistic. However, some concerning issues clearly arise in this experiment, and we do not anticipate these issues to resolve themselves in the presence of more complex data or models.

- The Bootstrap NN appears to be providing much higher quality UQ than the DE. Because the Bootstrap does not require any additional computational burden beyond a DE, it seems to be a reasonable choice if higher quality UQ is desired.
- While ECE has been demonstrated to be a useful metric, it is clearly not telling the whole story. Essentially, it tells us that class probabilities are being accurately predicted, but it cannot tell us if the variability of those estimated probabilities is being adequately estimated.

Conclusions

- We advocate for caution when using VI. The speedup over MCMC methods can't be ignored, but in our experience, expert tuning is needed for VI implementations. Improved software and VI methodology could be fruitful areas for future research.
- Interval coverage and width clearly convey important information about UQ quality that is not being captured by ECE. But because they require ground truth, we also advocate for further research into the development of UQ quality metrics that can be used on real data.