Assessing the Quality of Uncertainty Estimates in Deep Learning

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

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

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

Security Administration under contract DE-NA0003525.

Motivation

- Standard deep learning (DL) methods give predictions without associated measures of uncertainty.
- From a statistical perspective, predictions without a measure of prediction variance are incomplete. From a practical perspective, the usefulness of uncertainty estimates is clear.
- Developing methods for including uncertainty estimates along with the powerful prediction capabilities of DL is currently an active and important area of research.

Motivation

Our initial motivation for this work stems from analysis of a simulated hyperspectral image data set.

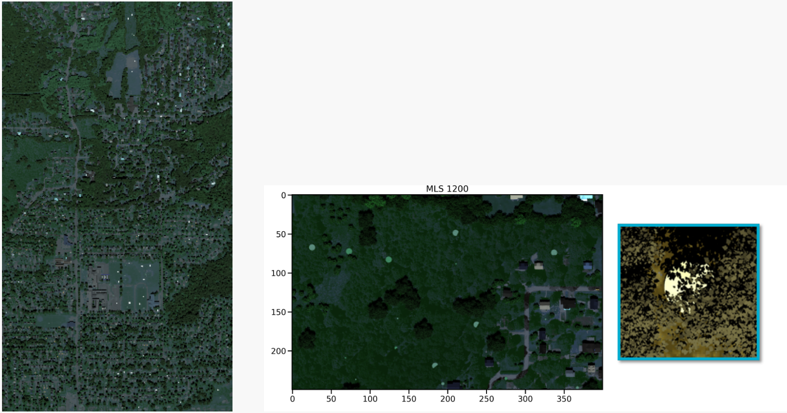


Figure: Example pseudo color rendering of image with green discs (left), zoomed-in region (center), and a single disc partially obstructed from view (right).

Motivation

A critical question arose as part of this analysis: how to assess the quality of the estimated uncertainty?

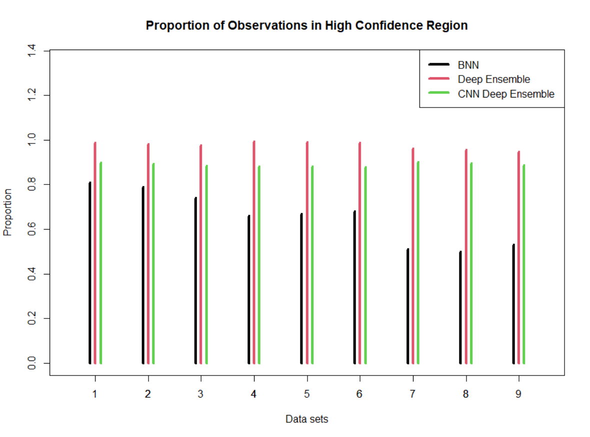


Figure: Proportion of predictions in which a model is 'highly confident' for three different models trained and evaluated on nine different hyperspectral images. This demonstrates the problems of (1) different models producing very different uncertainties and (2) overconfidence in model predictions.

Motivation

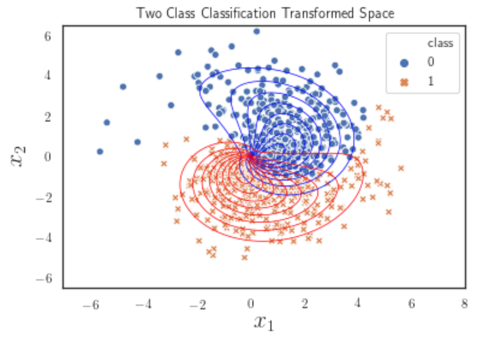
- From a practical standpoint, these models are producing drastically different levels of confidence in their predictions. Assessing the quality of the uncertainty estimates becomes critical in high-consequence applications, especially as the number of predictions to be made increases.
- Our goal is to develop a framework for the principled assessment of uncertainty estimates from deep learning models.

Interval Coverage and Width

- One approach is to evaluate the coverage and width of the prediction intervals.
- Our intervals are attempting to capture the probability of a given pixel (voxel) containing a target. The class probabilities are not available to us, so we are unable to assess coverage using the data at hand.
- This is a problem with classification problems in general. Further, simulating data sets with similar properties where ground truth class probabilities are known is a difficult challenge in itself.

Interval Coverage and Width

To get some idea of how various methods performed with regard to interval coverage and width, we simulated a two-dimensional two-class classification (TCC) data set where class probabilities are known for every (x_1, x_2) pair.



Interval Coverage and Width Results

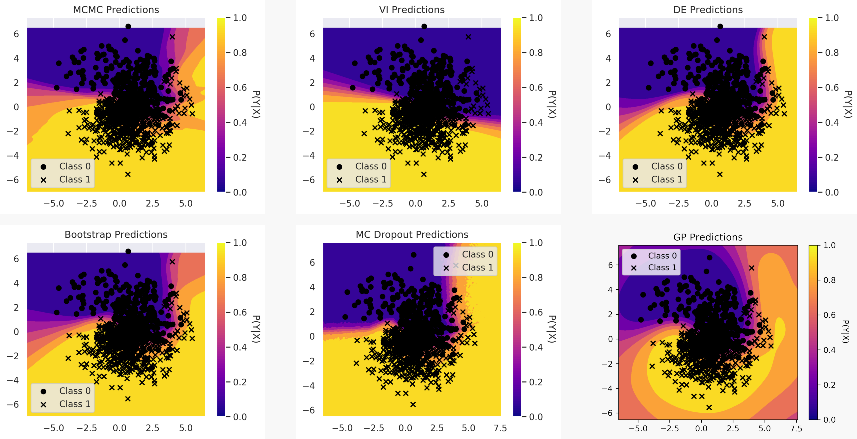
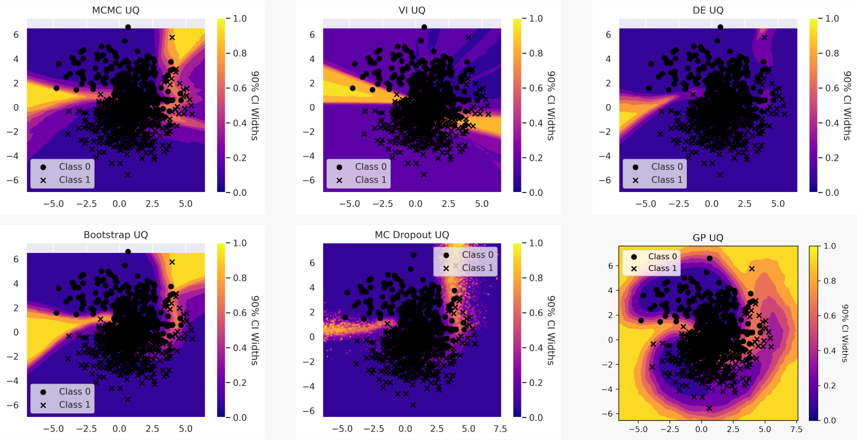


Figure: Prediction surfaces for each model on one TCC simulation.

Interval Coverage and Width Results



MCMC UQ

VI UQ

DE UQ

Bootstrap UQ

MC Dropout UQ

GP UQ

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

Interval Coverage and Width Results

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

Table: TCC interval coverage and width results for 90% prediction intervals. Coverage and width values are averaged (1) over all test observations for a given simulated data set and (2) over all simulated data sets. The standard deviation, averaged over all simulated data sets, is given in parentheses.

Interval Coverage and Width Results

- As we might expect based on results from the hyperspectral image analysis, we see a large discrepancy in terms of prediction interval quality.
- While informative, we do not claim that these conclusions apply generally. We believe the clearest lesson from this analysis is that more work on assessing the quality of uncertainty estimates is needed.

Approximating Coverage and Width

- As mentioned, simulating realistic data sets with known ground truth class probabilities is a difficult obstacle for complex data sets (such as the hyperspectral image data set we considered). This in turn can prevent us from assessing uncertainty quality through coverage and width.
- We are currently working on another approach that will allow us to approximate coverage and width. The idea is to use a generative model to simulate the data for us. With more powerful generative models (e.g. GANs), there is potential to accurately represent even complicated data sets.
- We illustrate this idea with the same two-class classification data set used previously.

Approximating Coverage and Width

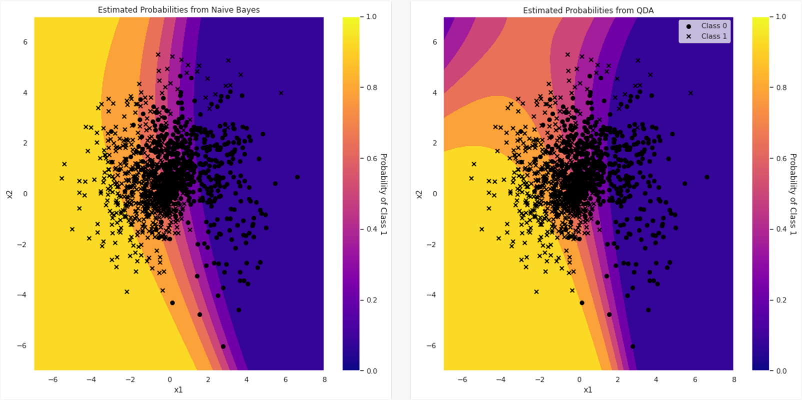


Figure: Estimated probability distributions generated by Naive Bayes (Left) and QDA (Right) with TCC data set overlaid.

Approximating Coverage and Width

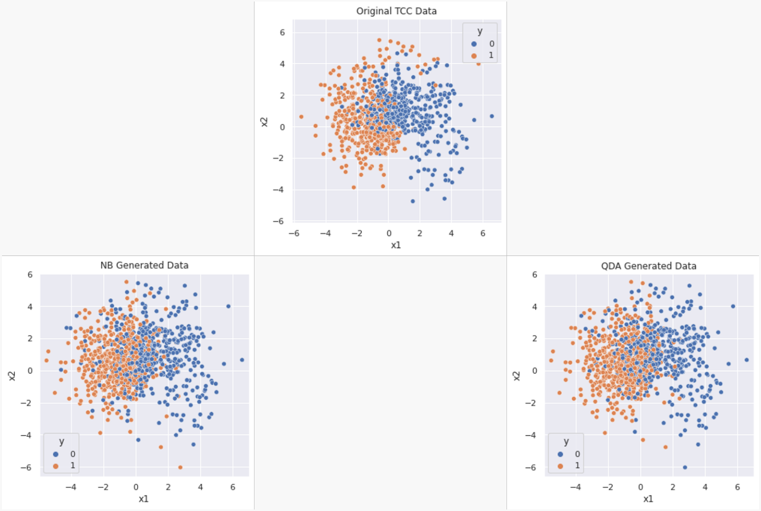


Figure: Original data (top) with Naive Bayes generated data (bottom left) and QDA generated data (bottom right). In a real application, the ground truth class probabilities for the top data set are unknown but known for the two bottom data sets.

Approximating Coverage and Width

- With the generated data sets, we know the underlying class probabilities that generated the data, so we can obtain interval coverage and width estimates from the *generated data*.
- While the two methods shown here are simple and may not work well on complex data sets, more sophisticated generative models are capable of replicating much more complex data.

Additional Approaches

- A method for calibrating epistemic uncertainty has been developed and has shown promising results in improving the accuracy of a model's predictions (this has been developed by Chris Qian, student of Dr. Feng Liang at UIUC)
- We are currently working with conformal prediction as coverage on a real data set *can* be estimated for conformal prediction sets. We are generating conformal prediction sets from the same models we used to assess coverage and width.
- We are also exploring other existing metrics (e.g., expected calibration error) to better understand what exactly each metric describes and how each should be used in understanding the quality of uncertainty estimates produced by a given model.

Thank you!

- Questions?
- Please feel free to contact me: jradams@sandia.gov