Exploring the Limits of Epistemic Uncertainty Quantification in Low-Shot Settings

Matias Valdenegro-Toro. Email: matias.valdenegro@dfki.de

German Research Center for Artificial Intelligence, Bremen, Germany

German Research Center for Artificial Intelligence

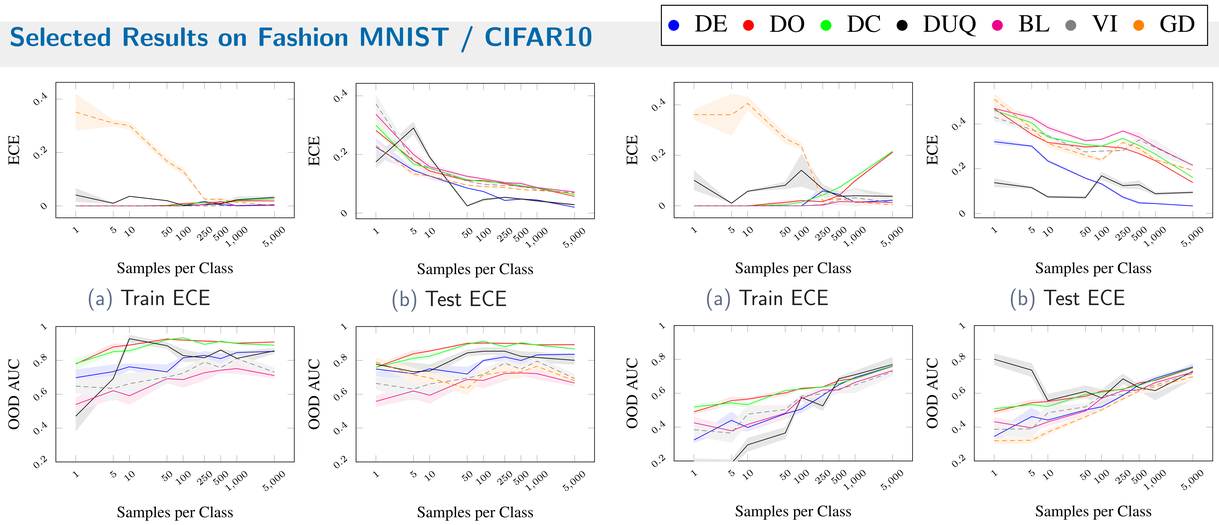
Summary

- **Motivation:** Bayesian Deep Learning promises good uncertainty estimates, but methods often rely in approximations, and real-world datasets have issues not present in academic benchmarks (like CIFAR10, Fashion MNIST, ImageNet, etc), such as low number of samples. Evaluating the quality of output uncertainty is difficult as there are no labels. In this paper we evaluate uncertainty quantification methods as the size of the training set is varied, to simulate real-world datasets.
- **Approach:** We take random subsamples of CIFAR10 and Fashion MNIST training sets and train several uncertainty methods (7 in total), evaluating on the corresponding test set. We measure accuracy, expected calibration error, entropy, maximum probability, and out of distribution detection AUC (with SVHN and MNIST), as the training set size is varied.
- **Contributions:** We compare uncertainty methods across different training set sizes, showing that confidences do not accurately portray model uncertainty. We show that ECE and OOD detection degrades with small training sets. We provide evidence for practitioners to select uncertainty methods and give future research directions.

Key Takeaways from Overall Results

- All methods except gradient, across all training set sizes, are well calibrated on the training set, but miscalibrated on the test set, and calibration improves with training set size.
- DUQ is less confident when SPC is low, which indicates that it correctly gauges its own uncertainty, other methods seem to be overconfident.
- Gradient-based methods seems to behave strangely, with poor Test/OOD AUC performance, and the worse calibration error both in train and test sets.
- Ensembles are competitive in terms of accuracy and calibration error, but do not perform as well in some OOD detection scenarios (Test/OOD).
- It is not clear if maximum probability or entropy is the best for out of distribution detection.
- There is no method that clearly outperforms all others across varied SPC values. Some methods work very well for calibration, but are outperformed in different out of distribution detection settings.
- Selection of uncertainty methods should be done carefully, considering the training set size for a given task.

Selected Results on Fashion MNIST / CIFAR10



(a) Train ECE (b) Test ECE (c) Entropy Test/OOD (d) Maxprob Test/OOD

Figure: Fashion MNIST with MNIST as OOD

(a) Train ECE (b) Test ECE (c) Entropy Test/OOD (d) Maxprob Test/OOD

Figure: CIFAR10 with SVHN as OOD

Toy Classification on Two Moons Dataset

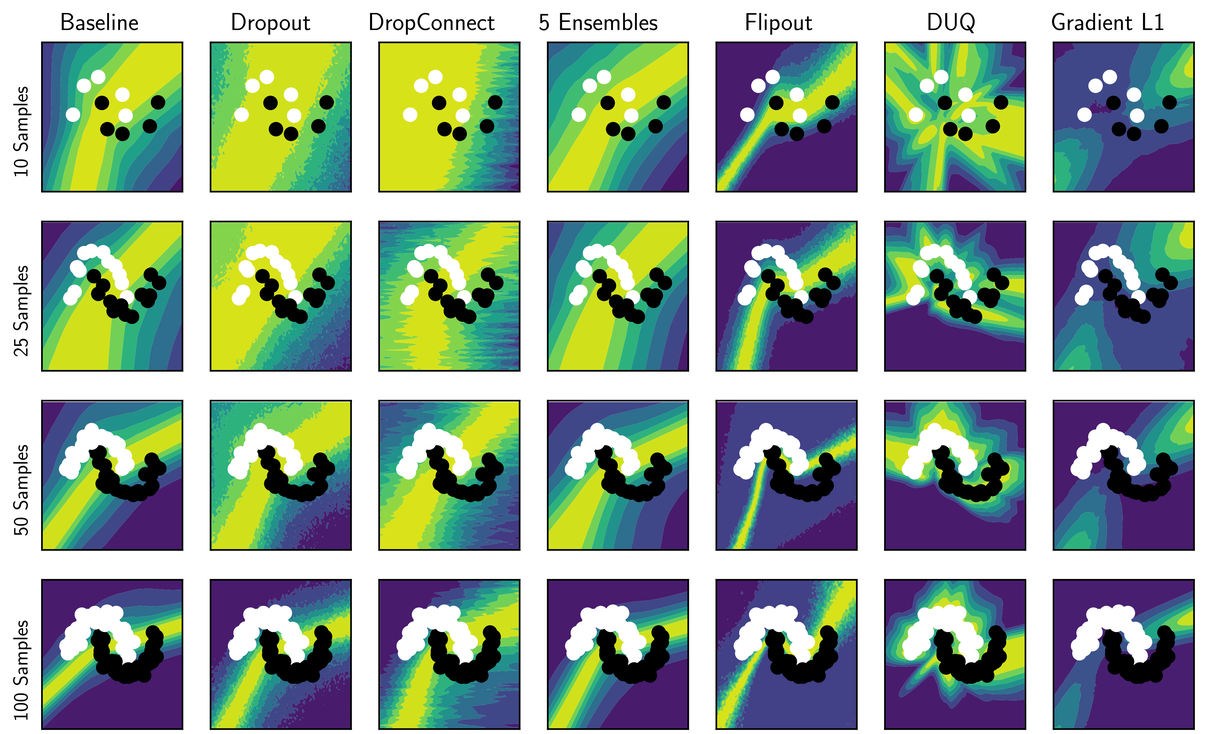


Figure: Qualitative comparison of two moons classification, yellow indicates high entropy, and blue indicates low entropy. For DUQ the plot indicates distance instead of uncertainty.

Toy Regression of $\sin(x) + \epsilon$ with $\epsilon \sim N(0, 0.15(1 + e^{-x})^{-1})$

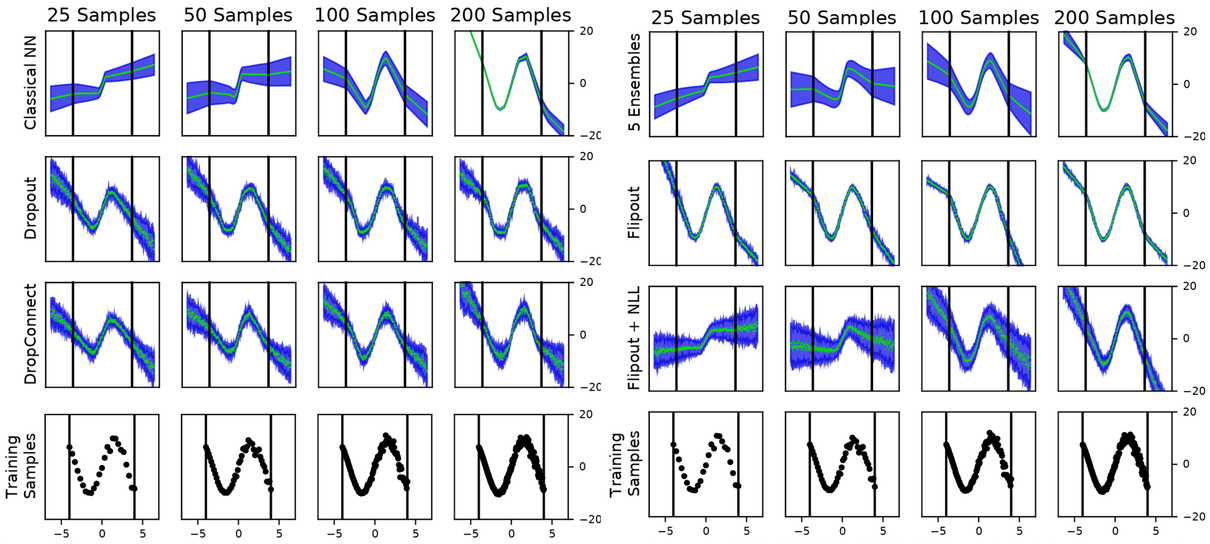


Figure: Qualitative comparison of toy regression as training set size is varied at $s \in [25, 50, 100, 200]$. The two black lines indicate the limits of the training set, while the out of distribution test set ranges at $[-7, -4] \cup [4, 7]$