Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift

Dmitry Protasov

MIPT, 2023

November 21, 2023

Motivation

Deep Neural Networks (DNNs) are increasingly used in high-stakes applications like medical diagnosis and autonomous driving. In such scenarios, accurately quantifying predictive uncertainty is as important as making point predictions

Contribution

This study focuses on the challenge of evaluating predictive uncertainty in machine learning models, especially under conditions of dataset shift. A large-scale empirical comparison of Bayesian and non-Bayesian methods for uncertainty estimation is presented.

Background

Notation and Problem Setup

Let $x \in \mathbb{R}^d$ represent a set of d-dimensional features and $y \in \{1, \ldots, k\}$ denote corresponding labels for k-class classification. The training dataset D consists of N i.i.d. samples $D = \{(x_n, y_n)\}_{n=1}^{N}$.

Data Generating Process

The true distribution $p^*(x, y)$, unknown and observed only through the samples D, assumed to be a discrete distribution over k classes.

OOD Inputs

Inputs sampled from a distribution $q(x, y)$ that is different from the training distribution $p^*(x, y)$, challenging the model's predictive certainty.

Background

We consider two kinds of shifts

Covariate Shift

Shifts in the distribution of test inputs, potentially leading to a degradation in the accuracy of the model's predictions.

Out-Of-Distribution (OOD) Dataset

A test set where the true label is not one of the known k classes, requiring the model to express higher predictive uncertainty.

Overview of Uncertainty Estimation Methods

Categories of Methods

There are three main categories of machine learning methods for uncertainty estimation and out-of-distribution (OOD) detection:

- Methods which deal with conditional probability $p(y|x)$ only
- Methods modeling the joint distribution $p(y, x)$
- Methods with an OOD-detection component in addition to $p(y|x)$

Note

Comparing across these classes of methods is challenging due to differences in modeling assumptions. However, the focus remains on methods that quantify predictive uncertainty.

Methods

Selected Methods

A subset of methods from the probabilistic deep learning literature was selected based on prevalence, scalability, and practical applicability:

- Maximum softmax probability (Vanilla)
- Post-hoc calibration by temperature scaling (Temp Scaling)
- Monte-Carlo Dropout
- Ensembles of networks
- Stochastic Variational Bayesian Inference (SVI)
- Bayesian inference for the parameters of the last layer only (LL)
  - Mean field stochastic variational inference (LL SVI)
  - Dropout on activations before the last layer (LL Dropout)

Model Uncertainty Metrics

Metrics to evaluate the quality of model uncertainty under dataset shift

- Negative Log-Likelihood (NLL) – can over-emphasize tail probabilities
- The Brier Score (BS) – insensitive to predicted probabilities associated with in/frequent events:

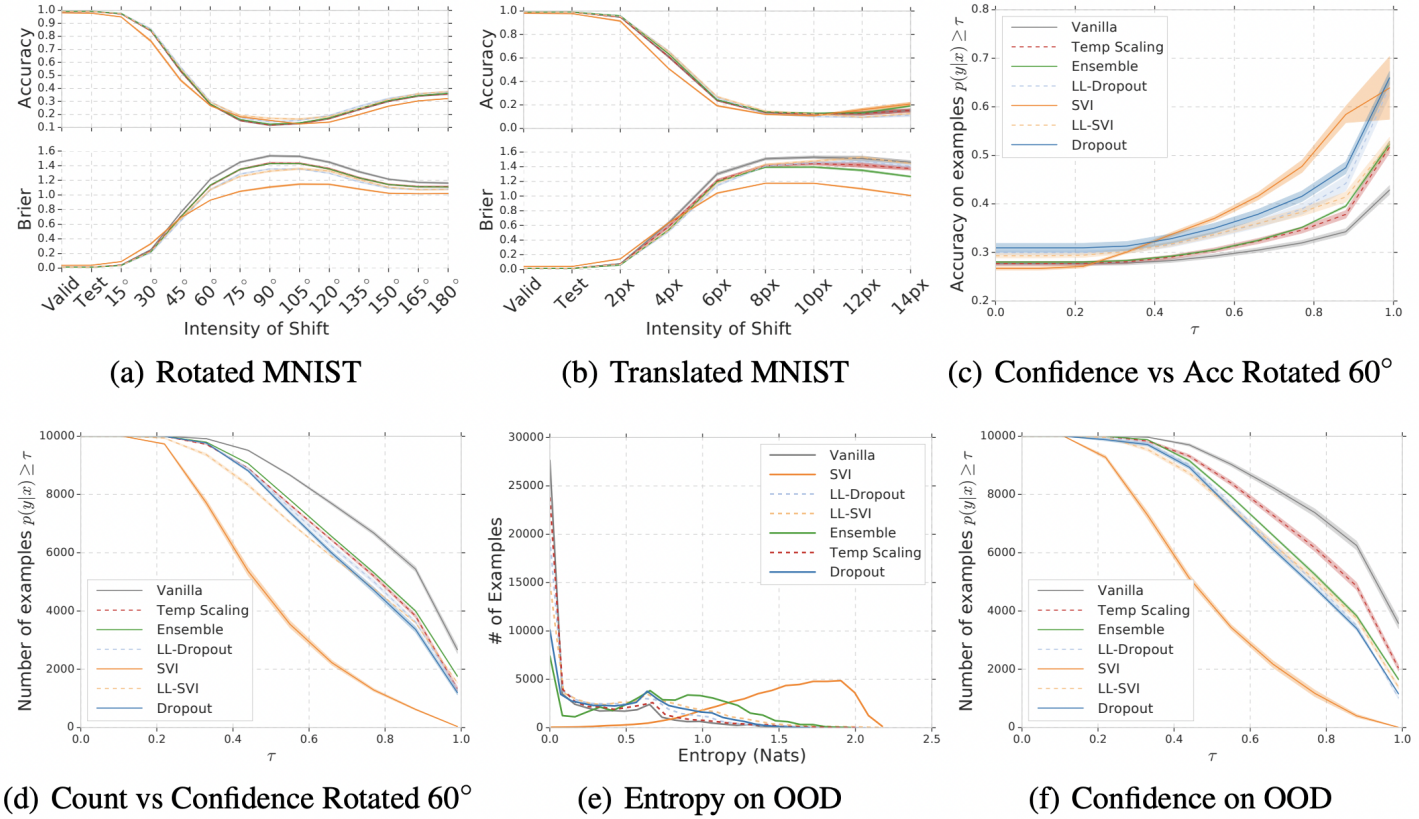
$$BS = \frac{1}{N}\sum_{n=1}^{N}\Big(p(y_n|x_n,\theta) - \delta(y_n - y)\Big)^2 \tag{1}$$

- The Expected Calibration Error (ECE) is:

$$ECE = \sum_{s=1}^{S}\frac{|B_s|}{N}|acc(B_s) - conf(B_s)| \tag{2}$$

where B_s are bins of predicted probabilities, and $acc(B_s)$, $conf(B_s)$ are accuracy and confidence of bin B_s, respectively. ECE does not monotonically increase as predictions approach ground truth.

Experiments: An illustrative example - MNIST



(a) Rotated MNIST (b) Translated MNIST (c) Confidence vs Acc Rotated 60°

(d) Count vs Confidence Rotated 60° (e) Entropy on OOD (f) Confidence on OOD

Figure: (a) and (b): accuracy and Brier score as the data is increasingly shifted. Shaded regions represent standard error over 10 runs. The predictive distributions of each method by looking at the confidence of the predictions in (c) and (d). The entropy and confidence of each method on entirely OOD data in (e) and (f). SVI has lower accuracy on the validation and test splits, but it is significantly more robust to dataset shift

Image Models: CIFAR-10 and ImageNet

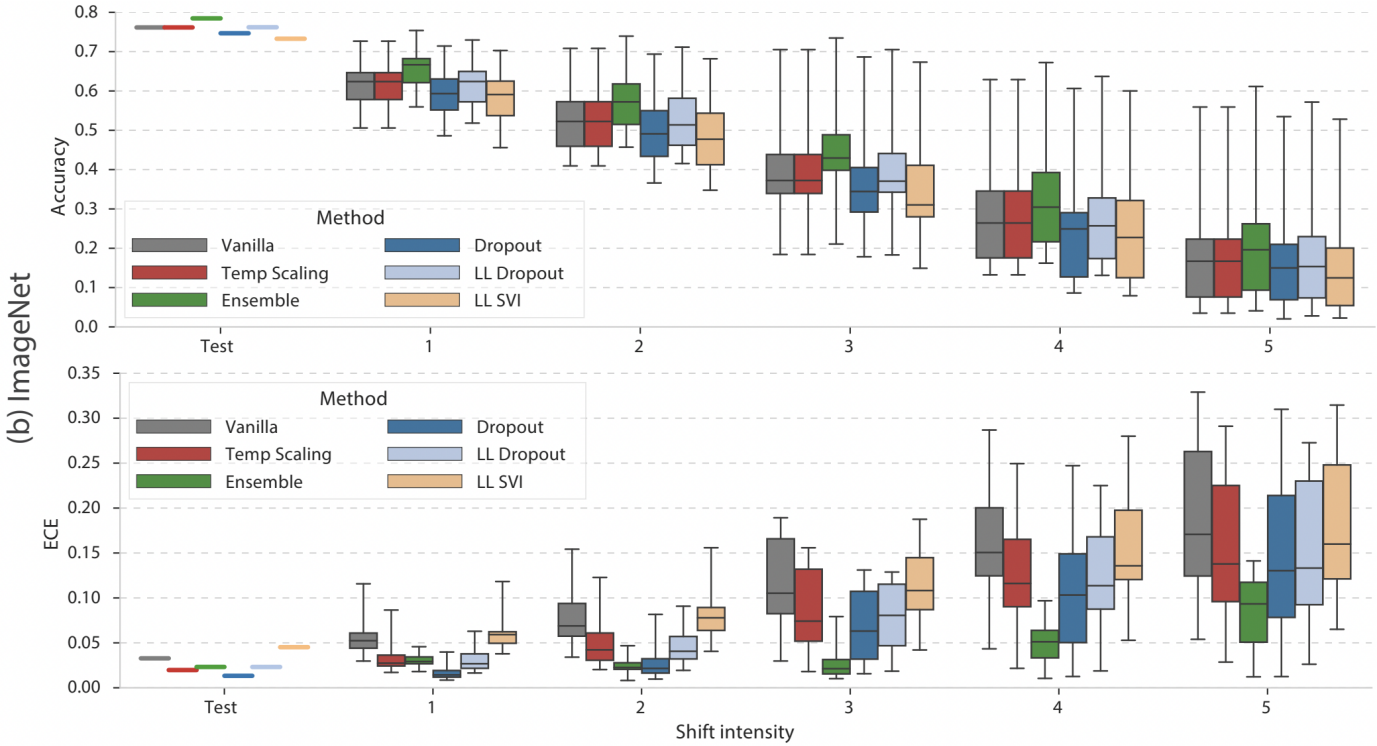
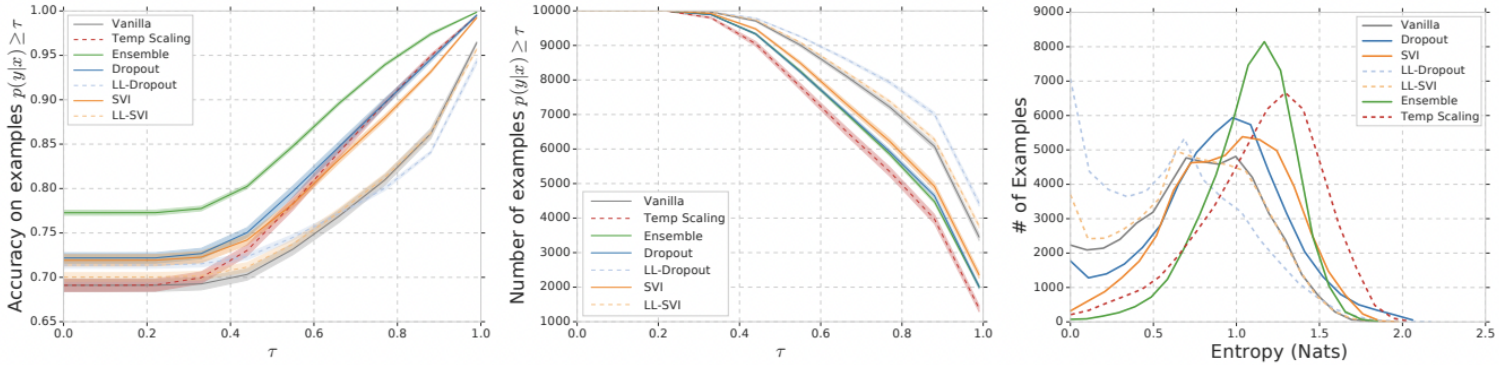
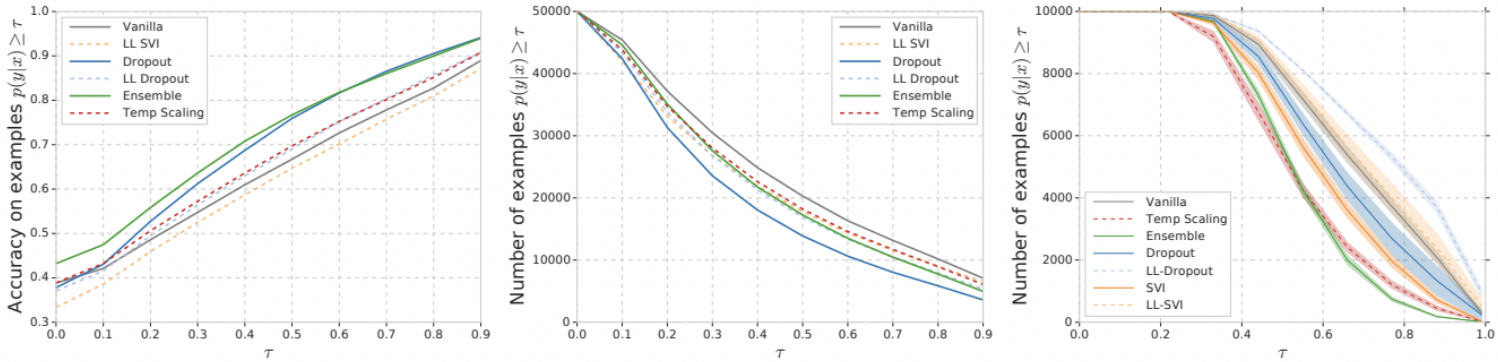


Figure: Calibration under distributional shift: comparison of accuracy and ECE under all types of corruptions on ImageNet. For each method we show the mean on the test set and summarize the results on each intensity of shift with a box plot. Each box shows the quartiles summarizing the results across all (16) types of shift while the error bars indicate the min and max across different shift types.

Image Models: CIFAR-10 and ImageNet



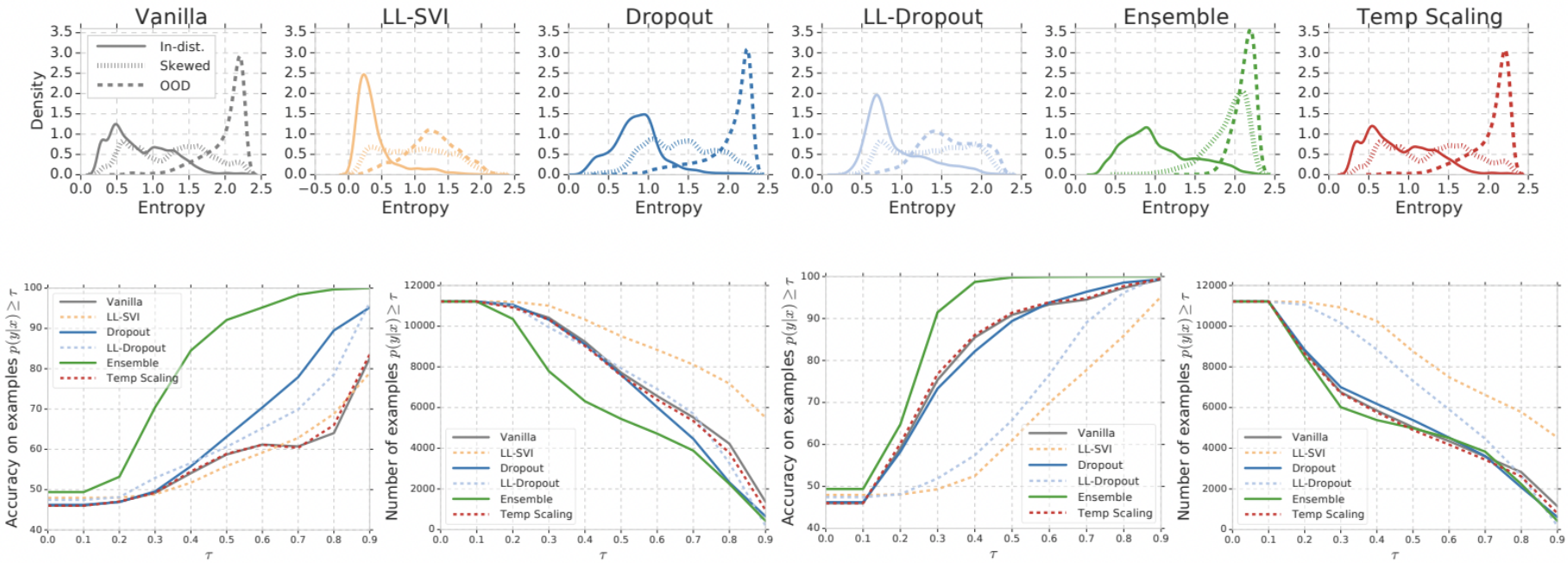
(a) CIFAR: Confidence vs Accuracy (b) CIFAR: Count vs Confidence (c) CIFAR: Entropy on OOD



(d) ImageNet: Confidence vs Acc (e) ImageNet: Count vs Confidence (f) CIFAR: Confidence on OOD

Figure: left: accuracy as a function of confidence
middle: the number of examples greater than given confidence values for Gaussian blur of intensity 3
right: histogram of entropy and confidences from CIFAR-trained models on a completely different dataset (SVHN) (right column)

Text Models



(a) Confidence vs Acc. (b) Confidence vs Count (c) Confidence vs Accuracy (d) Confidence vs Count

Figure: Top row: Histograms of the entropy of the predictive distributions for in-distribution (solid lines), shifted (dotted lines), and completely different OOD (dashed lines) text examples.
Bottom row: Confidence score vs accuracy and count respectively when evaluated for in-distribution and in-distribution shift text examples (a,b), and in-distribution and OOD text examples (c,d)

Ad-Click Model with Categorical Features

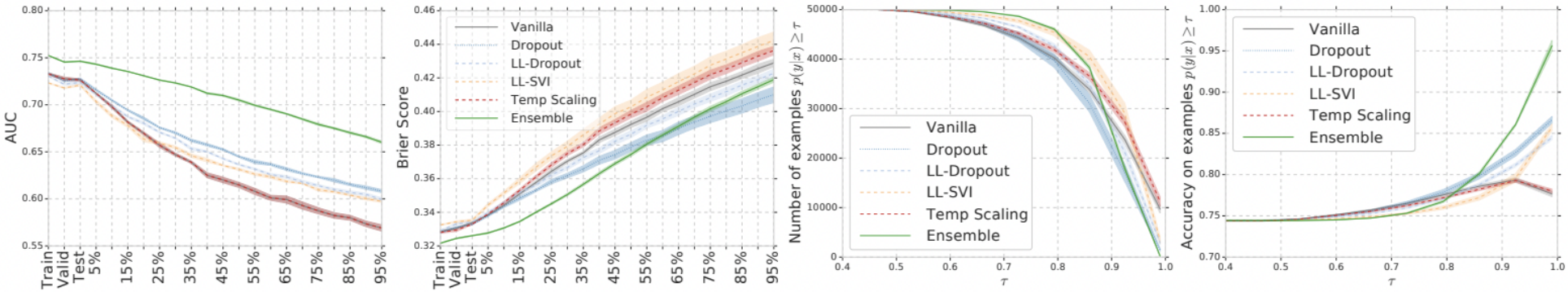


Figure: Results on Criteo: The first two plots show degrading AUCs and Brier scores with increasing shift while the latter two depict the distribution of prediction confidences and their corresponding accuracies at 75 % randomization of categorical features. SVI is excluded as it performed too poorly

Takeaways and Recommendations

Summary of Findings

- Uncertainty quality degrades with increasing dataset shift.
- Calibration on i.i.d. test data does not ensure calibration under dataset shift.
- Temperature scaling on i.i.d. validation sets leads to better-calibrated uncertainty than baseline methods, especially with greater shifts.
- Last layer Dropout shows less uncertainty than Dropout.
- SVI is effective for small datasets but struggles with larger ones
- The relative ordering of methods is mostly consistent across exp-s
- Deep ensembles generally outperform other methods

Recommendations for Future Work

It is recommended to explore more scalable and efficient methods for uncertainty quantification, especially under dataset shift. The benchmark set forth in this paper should inspire further research in this area.

Literature

1. **Main article** Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.