Can you trust your model’s uncertainty?

Evaluating Predictive Uncertainty Under Dataset Shift

Yaniv Ovadia, Emily Fertig, Jie Ren, Zack Nado,
D Sculley, Sebastian Nowozin, Joshua Dillon,
Balaji Lakshminarayanan, Jasper Snoek

Google AI
DeepMind

1. Motivation

- Modern ML classifiers assume data was drawn i.i.d. from the target data distribution.
- In practice, deployed models are evaluated on non-stationary data distributions.
  - **Distributions shift** (over time, seasonality, online trends , sensor degradation, etc.).
  - They are exposed to completely OOD data.
- We study the behavior of the predictive distributions of a variety of modern deep classifiers under (realistic) dataset shift.
  - Degradation of accuracy is expected, but do models remain calibrated?
  - Do models become increasingly uncertain under shift? Is uncertainty robust to shift?
  - Does calibration on the validation set help?
- We present a benchmark for uncertainty.

2. Modeling Methods

We tested a handful of scalable and well-known methods that attempt to account for uncertainty due to incomplete data (i.e. epistemic uncertainty).

- **Vanilla:** Baseline neural net model
- **Temp-Scaling:** Post-hoc calibration by temperature scaling using an in-distribution validation set.
- **Dropout:** Monte-Carlo Dropout.
- **Ensembles:** Ensembles of M networks trained independently from random initializations
- **SVI:** Stochastic Variational Bayesian Inference.
- **LL:** Approx. Bayesian inference for parameters of the last layer only (i.e. **LL-SVI**, **LL-Dropout**).

3. Metrics

In addition to reporting model accuracies, we also use the following metrics to evaluate predictive distributions

Expected Calibration Error (ECE)
- Computed as the average gap between within-bucket accuracy and within-bucket predicted probability for S buckets.
- Does not reflect accuracy (predicting class frequencies gives perfect calibration).

Negative Log-Likelihood (NLL)
- Proper scoring rule.
- Can overemphasize tail probabilities
- Commonly used to evaluate the quality of model uncertainty.

Brier Score
- Also a proper scoring rule.
- Quadratic penalty is more tolerant of low-probability errors than log

$$\mathrm{BS} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\left[p(y|\mathbf{x}_n,\theta)-\delta(y-y_n)\right]^2$$

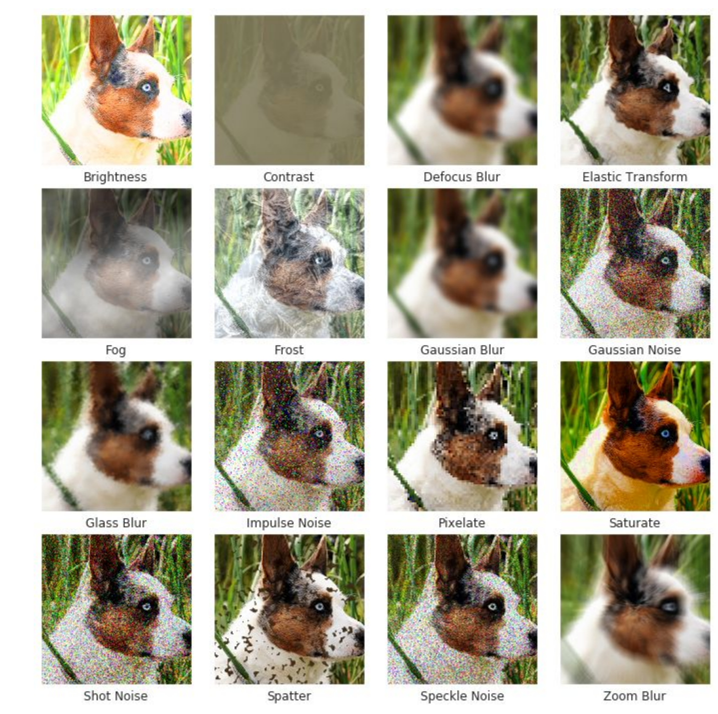
We also plot **accuracy-vs-confidence** to visualize the accuracy tradeoff when using prediction confidence as an OOD score.

Some experiments evaluated predictions on fully OOD examples; for this, we compare **distributions of predictive entropy**.

4. Datasets

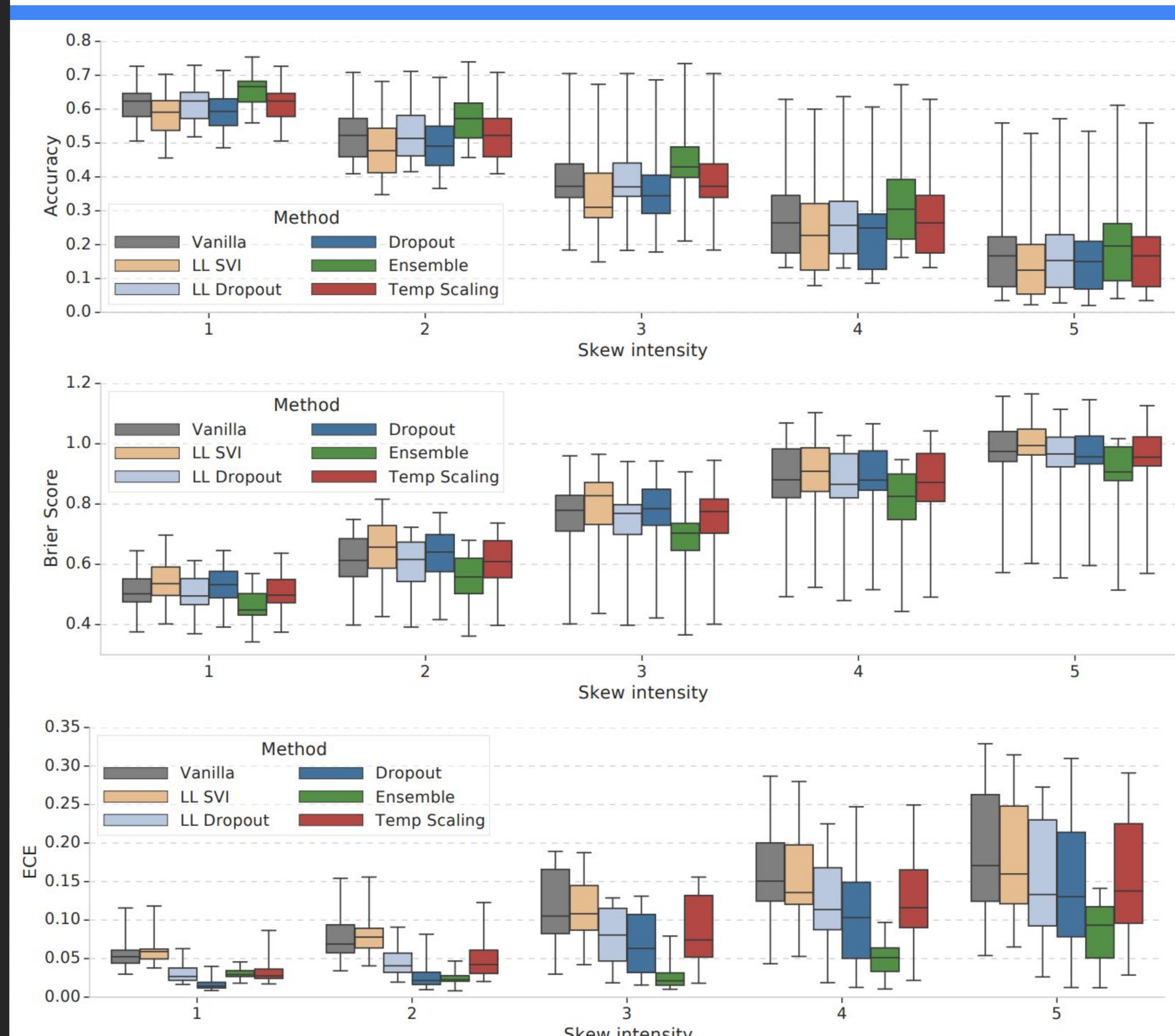
We tested datasets of different modalities and types of shift.:

- ImageNet
  - 16 different skew types of 5 intensities (from [Hendrycks and Dietterich, 2019])
  - Fully out-of-distribution (OOD) images Celeb-A
- CIFAR-10
  - 16 different skew types of 5 intensities (from [Hendrycks and Dietterich, 2019])
  - Fully OOD data from Streetview Housing Numbers
- Text
  - 20 Newsgroups (even classes as in-distribution, odd classes as shifted data)
  - Fully OOD text from LM1B
- Criteo Kaggle Display Ads Challenge
  - Skewed by randomizing categorical features with probability p (simulates token churn in non-stationary categorical features).



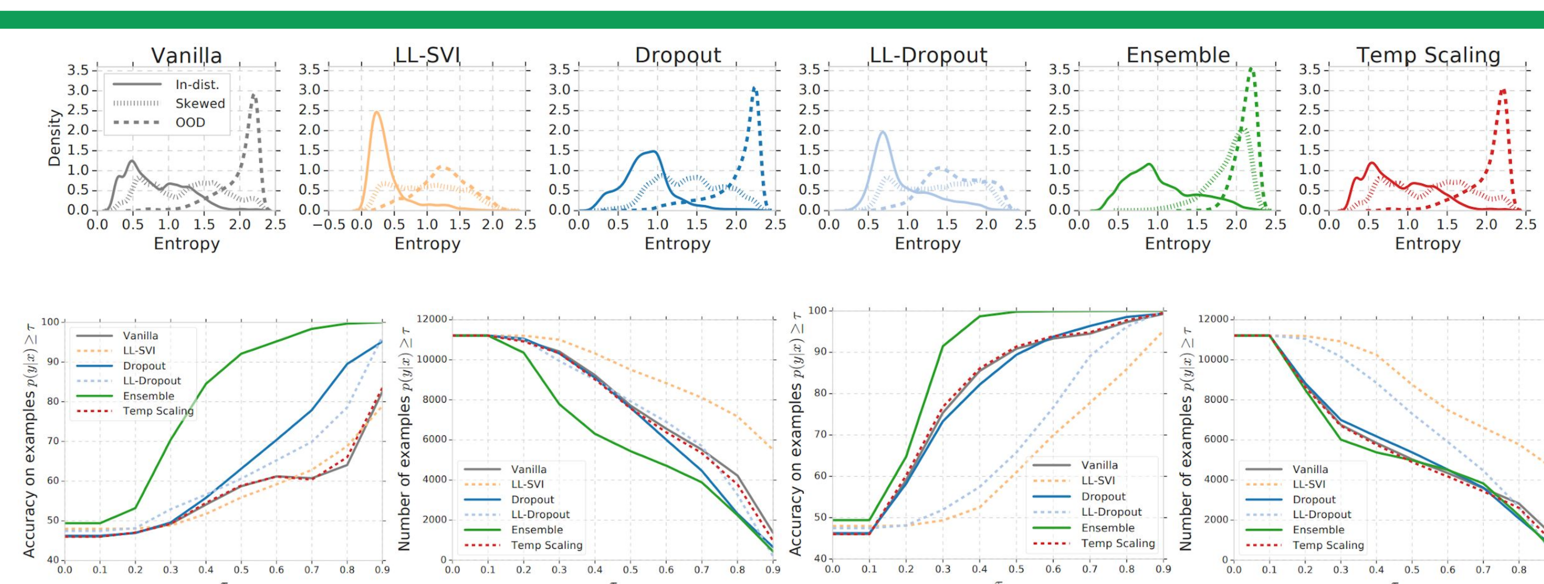
Hendrycks and Dietterich, 2019

5. Results: ImageNet



- Quality of uncertainty consistently degrades with increasing dataset shift regardless of the method.
- Better calibration and accuracy on i.i.d. test dataset does not usually translate to better calibration under dataset shift.
- Post-hoc calibration (on i.i.d validation) with temperature scaling leads to well-calibrated uncertainty on i.i.d. test and small values of skew, but is outperformed by methods that take epistemic uncertainty into account as the skew increases.
- Deep ensembles seem to perform the best across most metrics and be more robust to dataset shift.

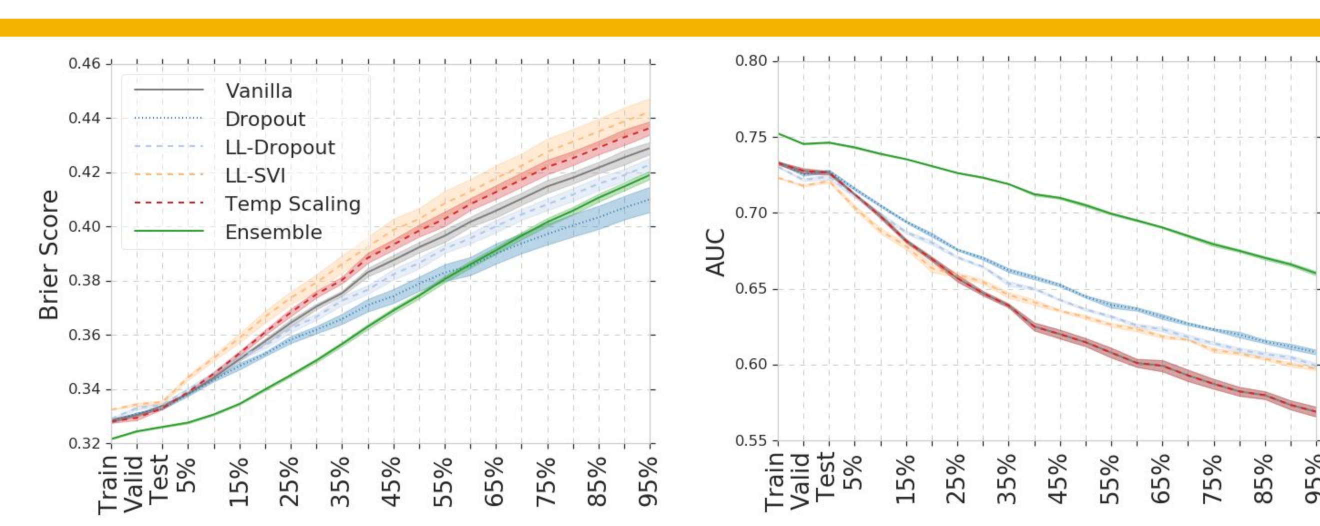
6. Results: Text-Classification



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

- All methods show increased entropy on skewed / OOD text.
- (a, b) correspond to a 50/50 mix of in-distribution and skewed text.
- (c, d) correspond to a 50/50 mix of in-distribution and fully-OOD text.

7. Results: Criteo Ad-Click Prediction



- Ensembles perform the best, but Brier score degrades rapidly.
- Both Dropout variants improve over Vanilla, and their Brier scores see less deterioration as skew increases.
- Temp Scaling led to worse Brier scores under skew.

8. Additional Observations

- SVI is promising on MNIST/CIFAR but difficult to use on larger datasets (e.g. ImageNet) and complex architectures (e.g. LSTMs).
- Relative ordering of methods is mostly consistent (except for MNIST) across our experiments.
- Deep ensembles seem to perform the best across most metrics and be more robust to dataset shift; relatively small ensemble size (e.g. 5) may be sufficient.

ArXiv Version: https://arxiv.org/abs/1906.02530