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

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

Google AI DeepMind

1. Motivation

- We typically assume that the test data is i.i.d. sampled from the same distribution as training data (e.g. cross-validation).
- In practice, deployed models are evaluated on non-stationary data distributions.
  - **Distributions shift** (over time, seasonality, online trends , sensor degradation, etc.).
  - They may be asked to predict on **out-of-distribution (OOD)** inputs.
- We study the behavior of the predictive distributions of a variety of modern deep classifiers under (realistic) dataset shift.
  - Degradation of accuracy is expected under dataset shift, but do models remain calibrated?
  - Do models become increasingly uncertain under shift?
- We present an open-source benchmark for uncertainty in deep learning.

2. Modeling Methods

We tested popular methods for uncertainty quantification.

- **Vanilla:** Baseline neural net model [Hendrycks & Gimpel, 2016]
- **Temperature-Scaling:** Post-hoc calibration by temperature scaling using an in-distribution validation set *[Guo et al., 2017]*.
- **Dropout:** Monte-Carlo Dropout *[Gal & Ghahramani, 2016]*.
- **Deep Ensembles:** Ensembles of *M* networks trained independently from random initializations *[Lakshminarayanan et al., 2017]*
- **SVI:** Stochastic Variational Bayesian Inference.
- **Last Layer variants:** Approximate Bayesian inference for parameters of the last layer only (i.e. **LL-SVI**, **LL-Dropout**).

3. Evaluation Metrics

In addition to accuracy, we also use the following metrics.

Calibration measures how well predicted confidence (probability of correctness) aligns with the observed accuracy.

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

Negative Log-Likelihood (NLL)
- Proper scoring rule.
- Can overemphasize tail probabilities

Brier Score

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

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

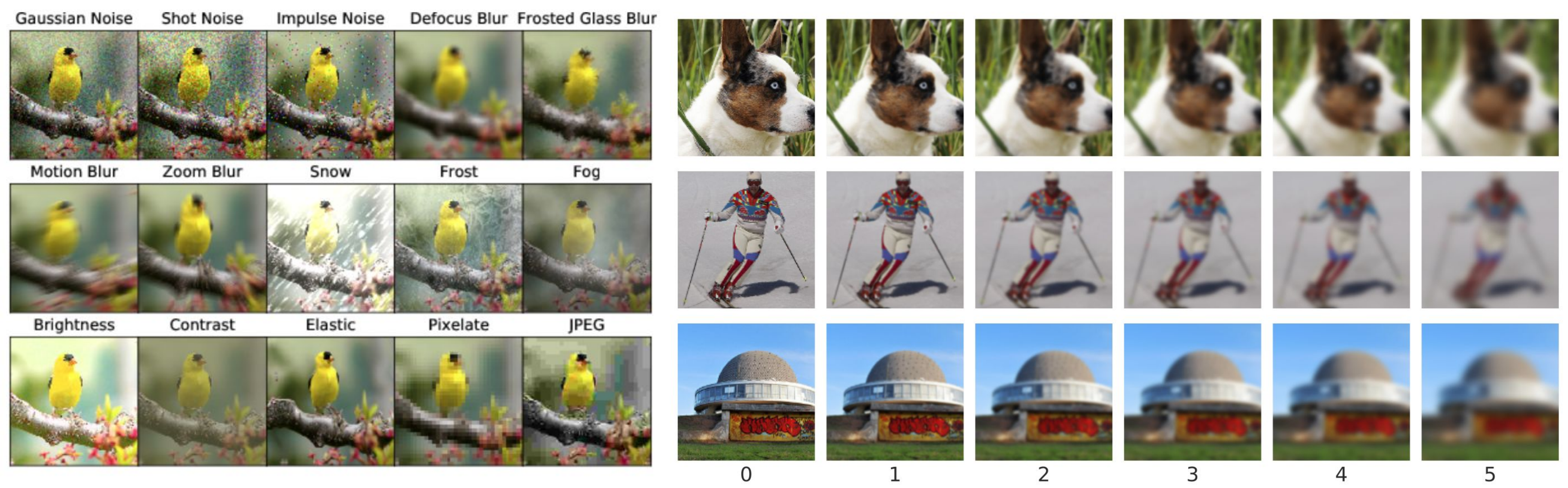
Accuracy-vs-confidence to visualize the accuracy tradeoff when using prediction confidence as an OOD score.

Distributions of predictive entropy on OOD datasets.

4. Datasets

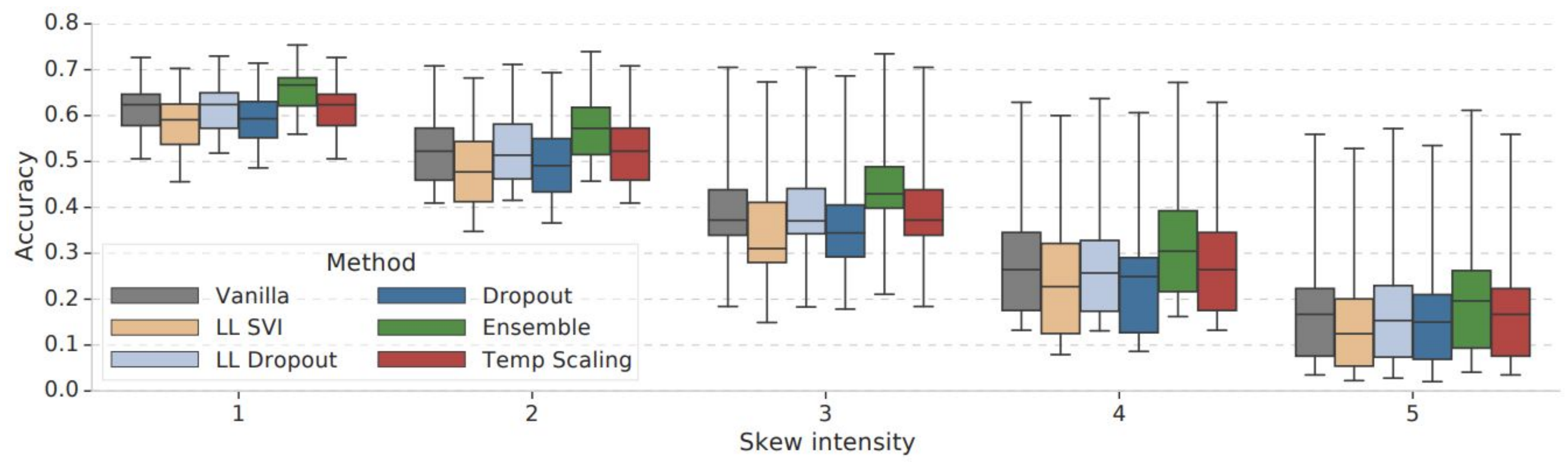
We tested datasets of different modalities and types of shift:

- Image classification on CIFAR-10 and ImageNet *(CNNs)*
  - 16 different skew types of 5 intensities *[Hendrycks & Dietterich, 2019]*
  - Train on ImageNet and Test on OOD images from Celeb-A
  - Train on CIFAR-10 and Test on OOD images from SVHN
- Text classification *(LSTMs)*
  - 20 Newsgroups (even classes as in-distribution, odd classes as shifted data)
  - Fully OOD text from LM1B
- Criteo Kaggle Display Ads Challenge *(MLPs)*
  - Skewed by randomizing categorical features with probability *p* (simulates token churn in non-stationary categorical features).

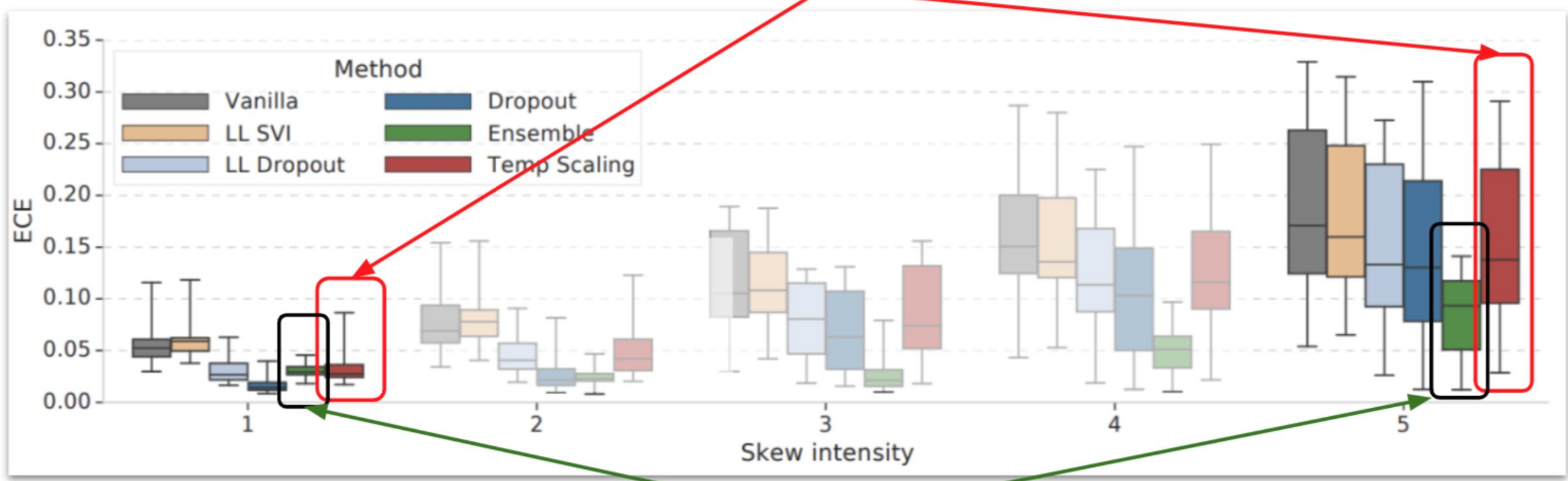


ImageNet-C *[Hendrycks & Dietterich, 2019]*. Left: types of corruptions and Right: Varying intensity.

5. Results: ImageNet



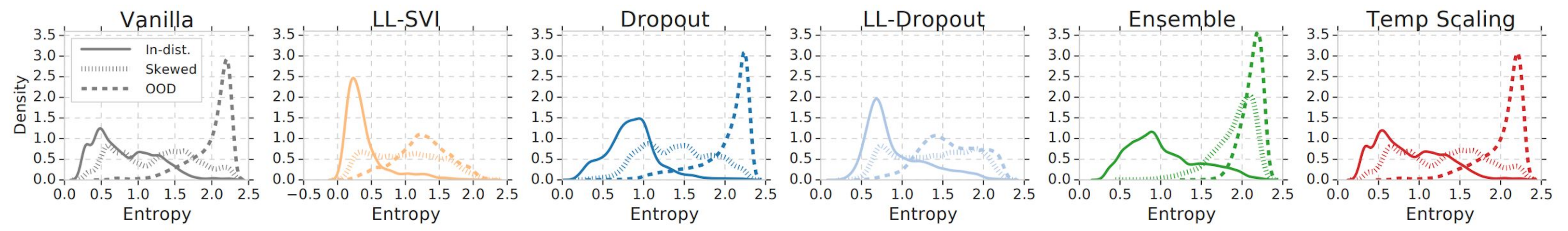
Temperature scaling is well-calibrated on i.i.d. test, but not calibrated under dataset shift

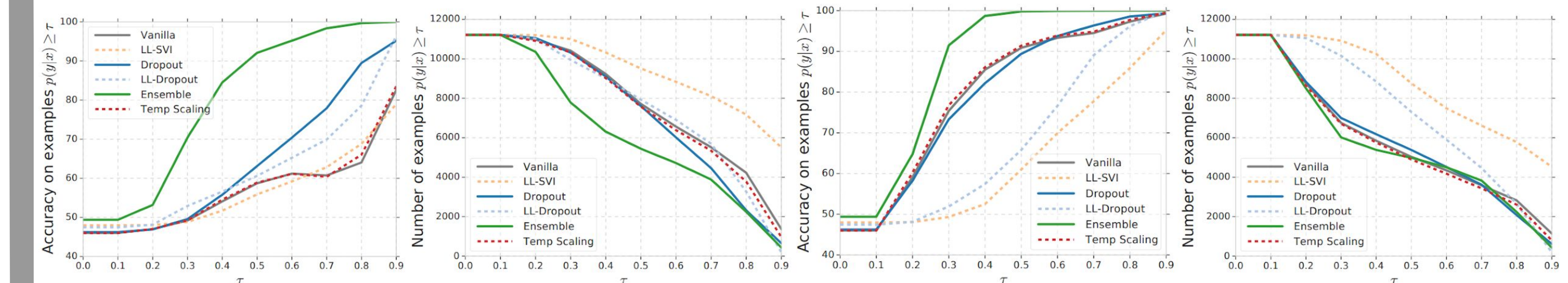


Ensembles are consistently among the best performing methods, especially under dataset shift

- Accuracy degrades with increasing dataset shift regardless of the method (as expected), but lower accuracy is not reflected in model's uncertainty.
- Similar trends on CIFAR-10.
- Ordering consistent when evaluating predictive entropy on OOD inputs.

6. Results: Text-Classification

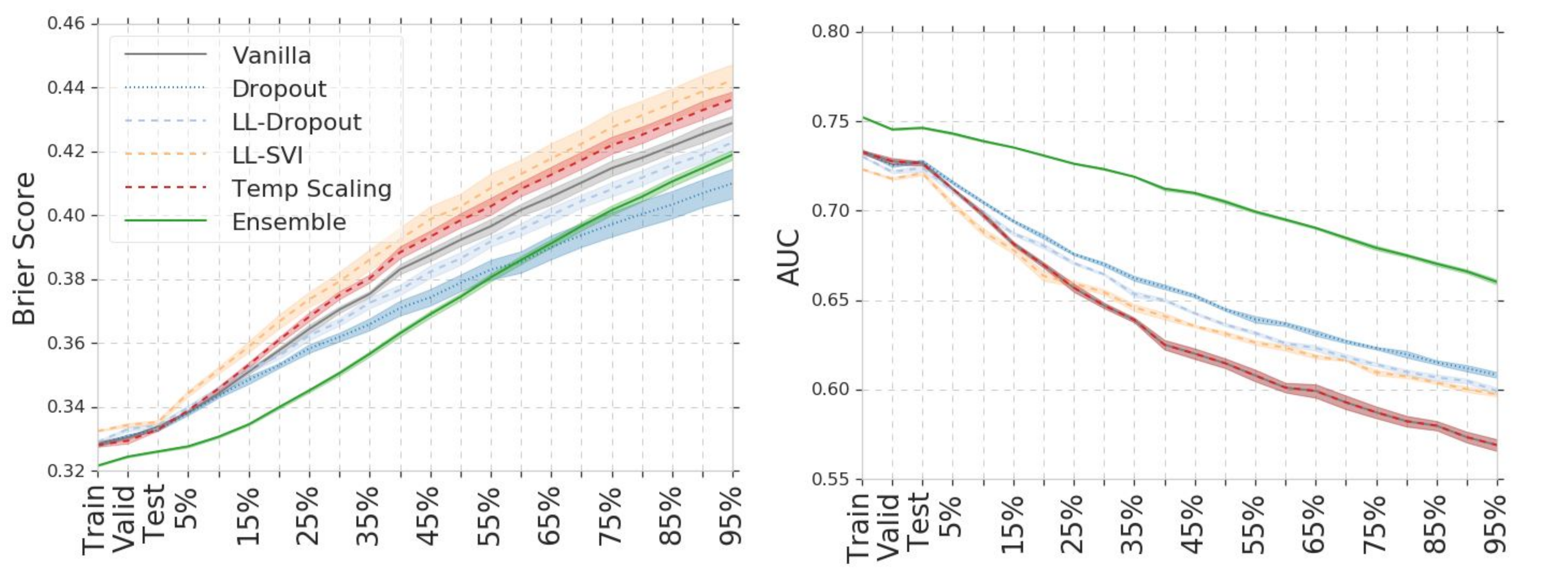




(a) Confidence vs Acc. (b) Confidence vs Count (c) Confidence vs Accuracy (d) Confidence vs Count

- (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.
- All methods generally exhibit higher entropy on skewed / OOD text.
- Confidence vs Accuracy curves show difference between the methods.

7. Results: Criteo Ad-Click Prediction



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

8. Take Home Messages

- Uncertainty under dataset shift is an important research challenge.
- Better calibration and accuracy on i.i.d. test dataset does not usually translate to better calibration under dataset shift.
- Bayesian neural nets (SVI) are 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)
- Deep ensembles are more robust to dataset shift & consistently perform the best across most metrics; relatively small ensemble size (e.g. 5) is sufficient.

Predictions and Code available online:

https://console.cloud.google.com/storage/browser/uq-benchmark-2019

https://github.com/google-research/google-research/tree/master/uq_benchmark_2019