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

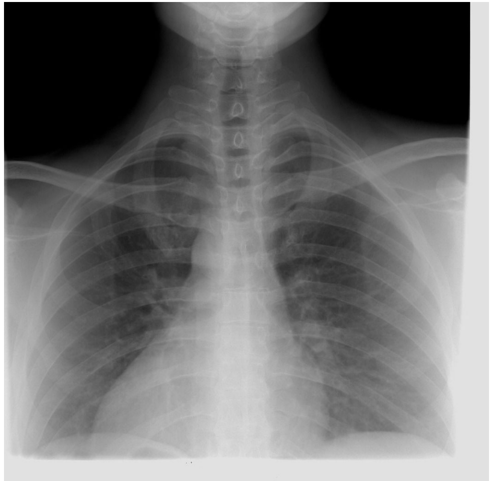
Uncertainty?

A motivating scenario

Deep learning is starting to show promise in radiology

- If output “probabilities” are passed on to doctors, can they be used to make medical decisions?
  - Does 0.3 chance of positive mean what they think it does?

- What happens when the model sees something it hasn’t seen before?
  - What if the camera lens starts to degrade?
  - One-in-a-million patient?
  - Does the model know what it doesn’t know?



Benchmarking Uncertainty

- This work: benchmarking uncertainty in modern deep learning models
    - Particularly as the input data changes from the training distribution - “covariate shift”
- We focus on classification probabilities
    - Are the numbers coming out of our deep learning classifiers (softmax) meaningful?
    - Can we treat them as probabilities?
        - If so we have a notion of uncertainty - e.g. entropy of the output distribution.
        - The model can express that is unsure (e.g. 0.5 chance of rain).
    - Probabilities allow us to make informed decisions downstream.

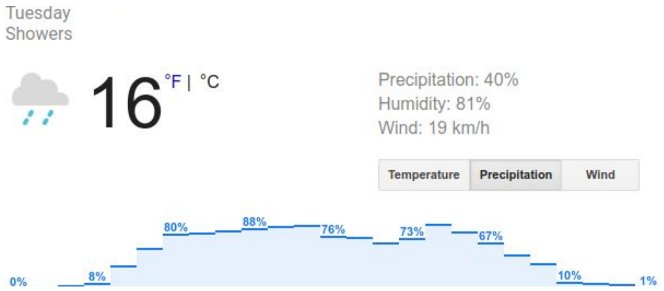
How do we measure the quality of uncertainty?

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).

Proper scoring rules

- See: *Strictly Proper Scoring Rules, Prediction and Estimation, Gneiting & Raftery, JASA 2007*
- Negative Log-Likelihood (NLL)
  - Can overemphasize tail probabilities



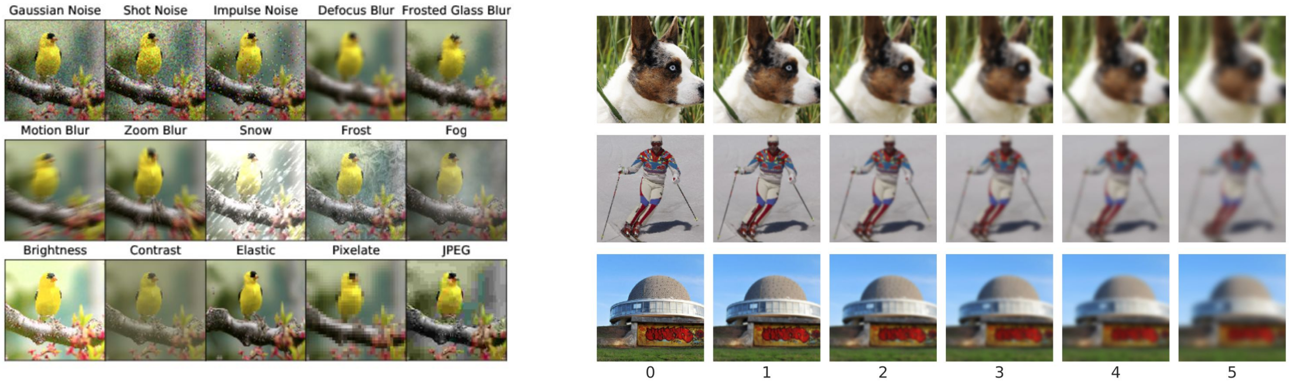
- Brier Score
  - Also a proper scoring rule.
  - Quadratic penalty is more tolerant of low-probability errors than log.

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

Dataset Shift

- Typically we assume training and test data are i.i.d. from the same distribution
  - Proper scoring rules suggest good calibration on test data

- In practice, often violated for test data
  - Distributions shift
  - What does this mean for uncertainty? Does the model know?



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

Datasets

We tested datasets of different modalities and types of shift:

- Image classification on CIFAR-10 and ImageNet *(CNNs)*
  - 16 different shift 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)*
  - Shifted by randomizing categorical features with probability p (simulates token churn in non-stationary categorical features).

Methods for Uncertainty (Non-Bayesian)

- Vanilla Deep Networks (baseline)
  - e.g. ResNet-20, LSTM, MLP, etc.

- Post-hoc Calibration
  - Re-calibrate on the validation set
  - Temperature Scaling (Guo et al., *On Calibration of Modern Neural Networks*, ICML 2017)

$$p(y_i|x) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

- Ensembles
  - Lakshminarayanan et al, *Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles*, NeurIPS, 2017.

(Approximately) Bayesian Methods

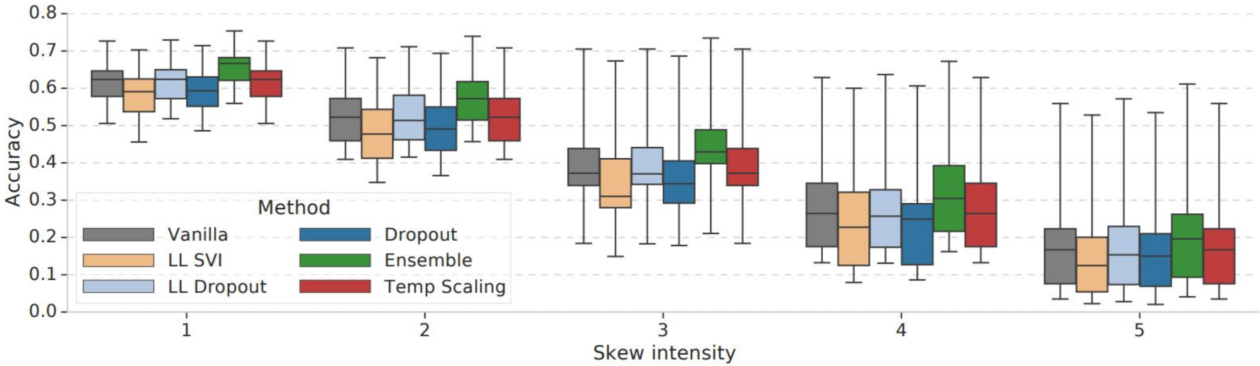
- Monte-Carlo Dropout
  - Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, Gal & Ghahramani, 2016

- Stochastic Variational Inference (mean field SVI)
  - e.g. Weight Uncertainty in Neural Networks, Blundell et al, ICML 2015

- What if we're just Bayesian in the last layer?
  - e.g. Snoek et al., Scalable Bayesian Optimization, ICML 2015
  - Last-layer Dropout
  - Last-layer SVI

Results - Imagenet

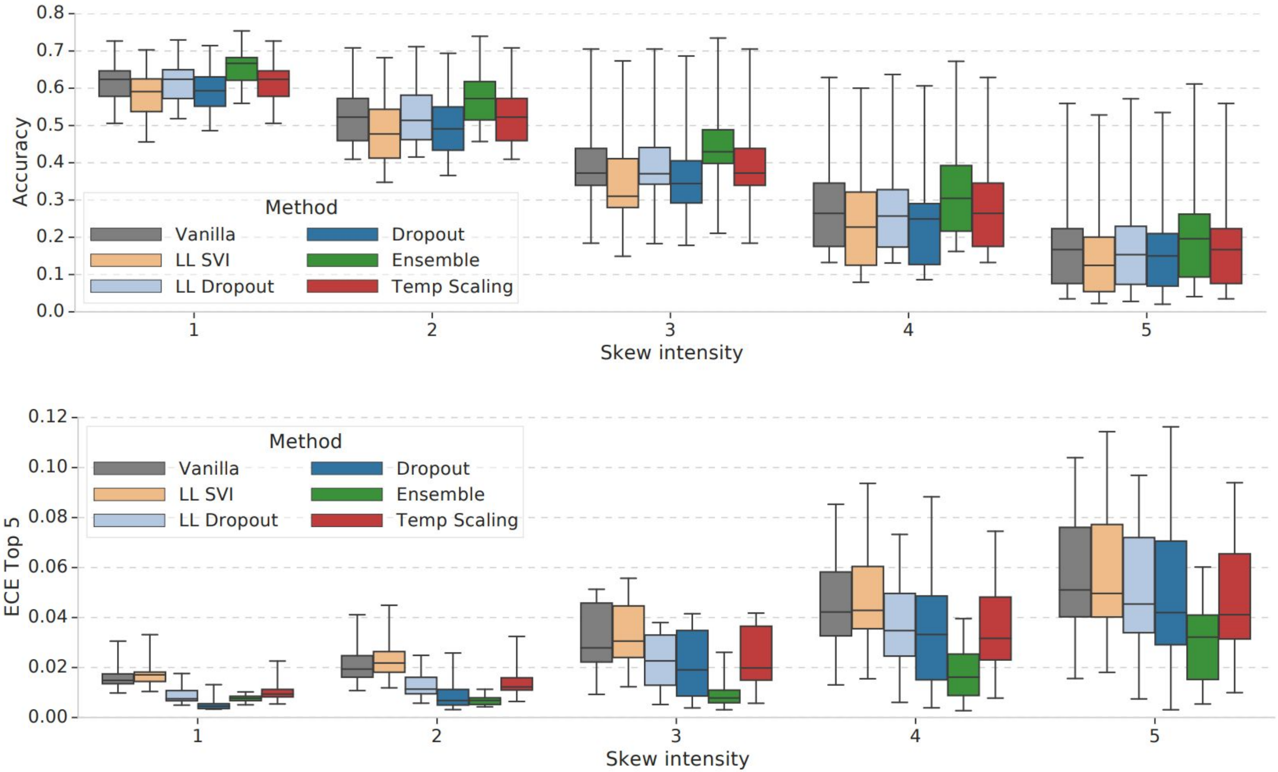
Accuracy degrades under shift



But does our model know it’s doing worse?

Results - Imagenet

Accuracy degrades under shift



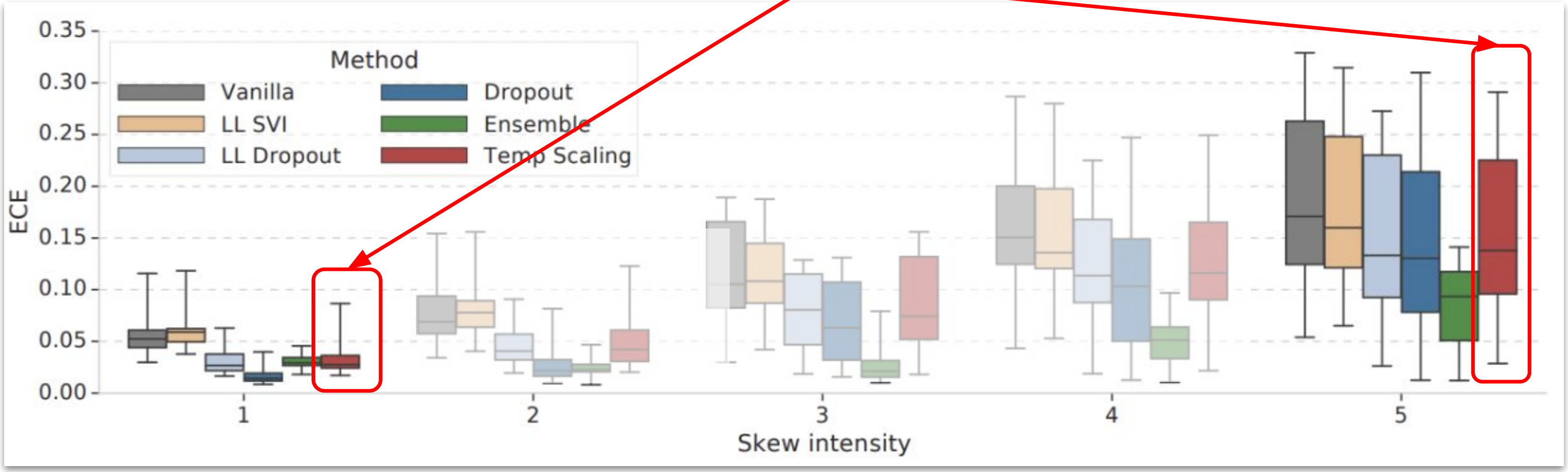
But does our model know it's doing worse?

- Not really...

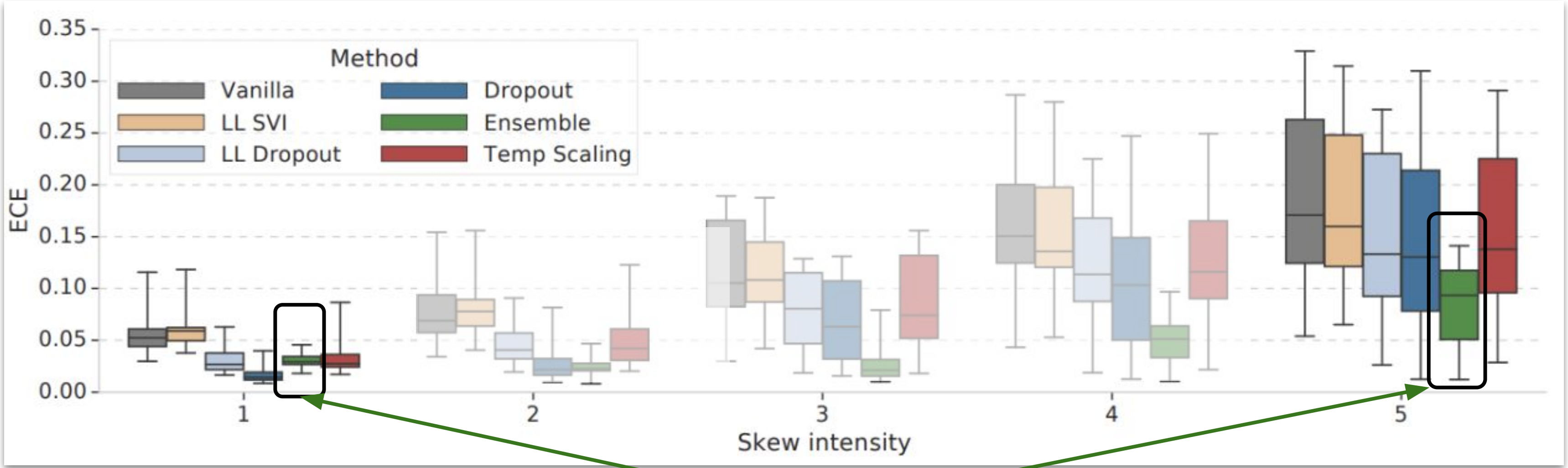
Google AI DeepMind

Traditional calibration methods are misleading

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

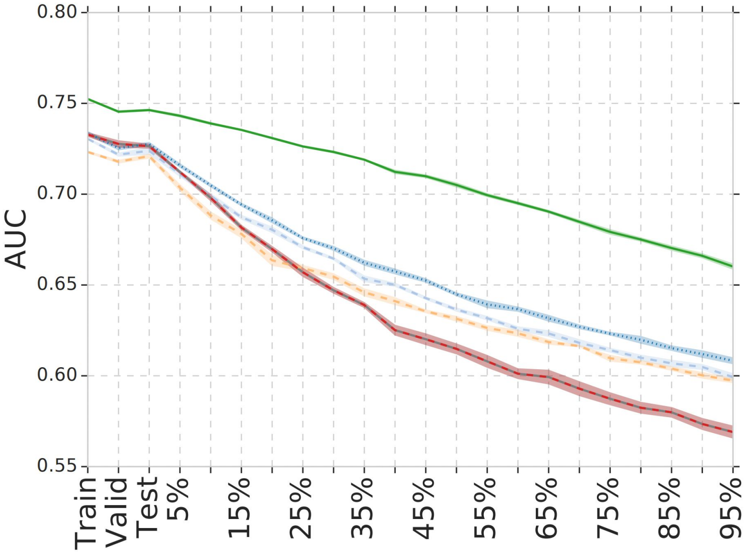


Ensembles work surprisingly well



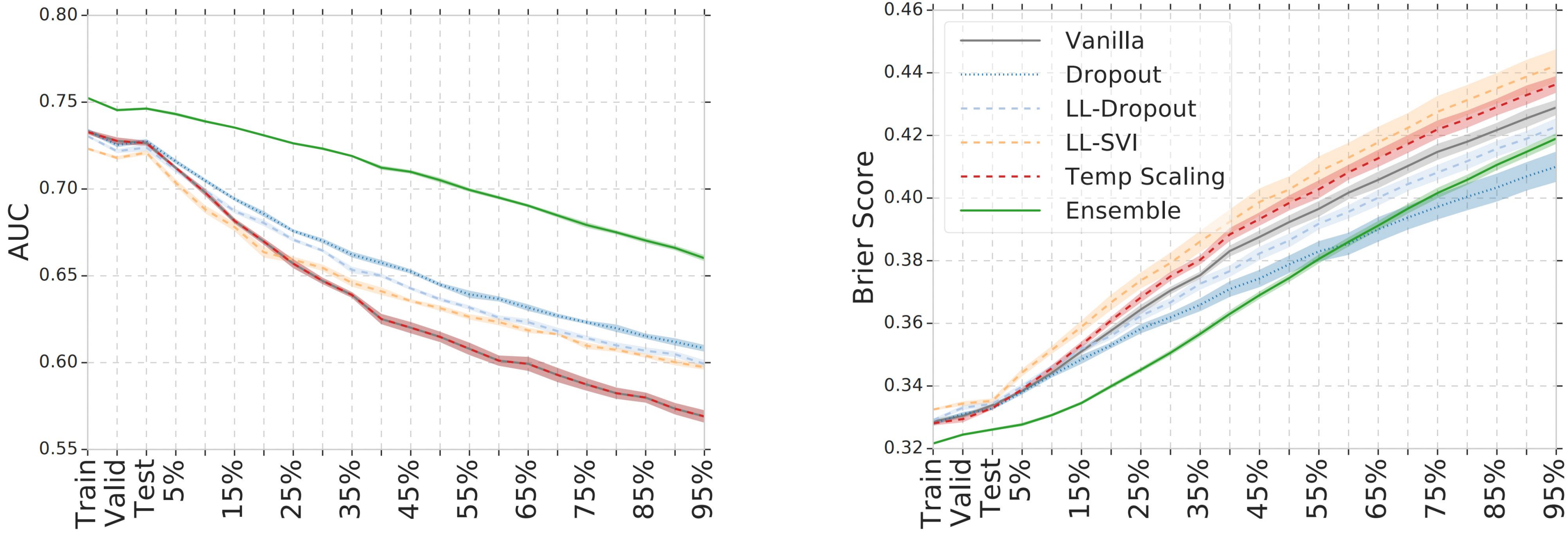
Ensembles are consistently among the best performing methods, especially under dataset shift

Criteo Ad-Click Prediction - Kaggle



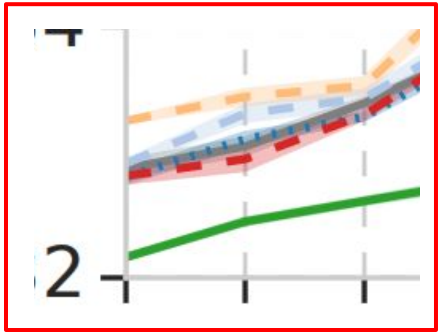
- Accuracy degrades with shift
- What about uncertainty?

Criteo Ad-Click Prediction - Kaggle

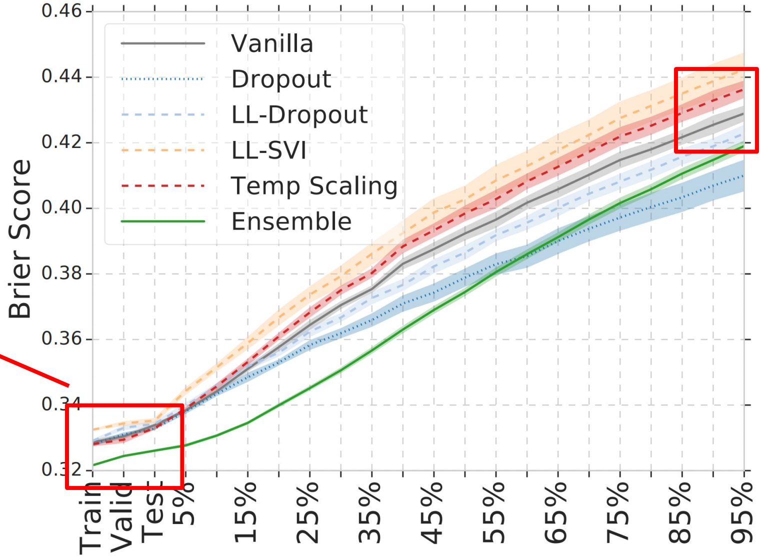


- Ensembles perform the best again, but Brier score degrades rapidly with shift.

Criteo Ad-Click Prediction - Kaggle



Temp scaling is better
than vanilla on the test set



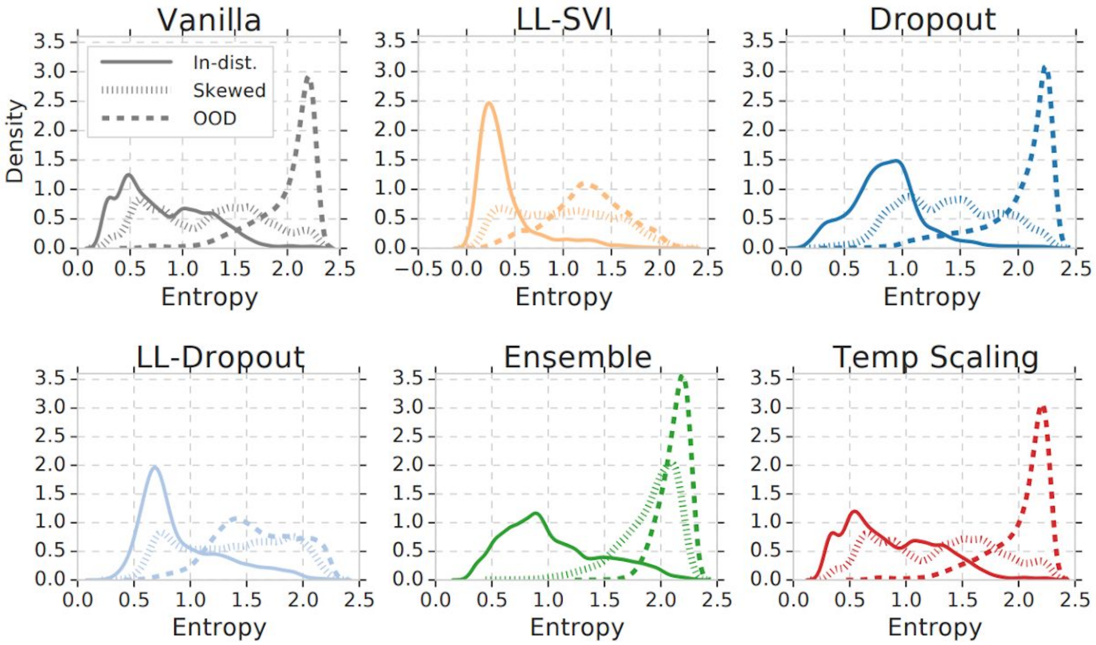
But worse
under shift!

- Post-hoc calibration (temp. scaling) actually makes things worse under dataset shift.

Results Text-Classification

What if we look at predictive entropy on the test set, shifted data and completely out-of-distribution data?

It's hard to disambiguate shifted from in-dist using a threshold on entropy...



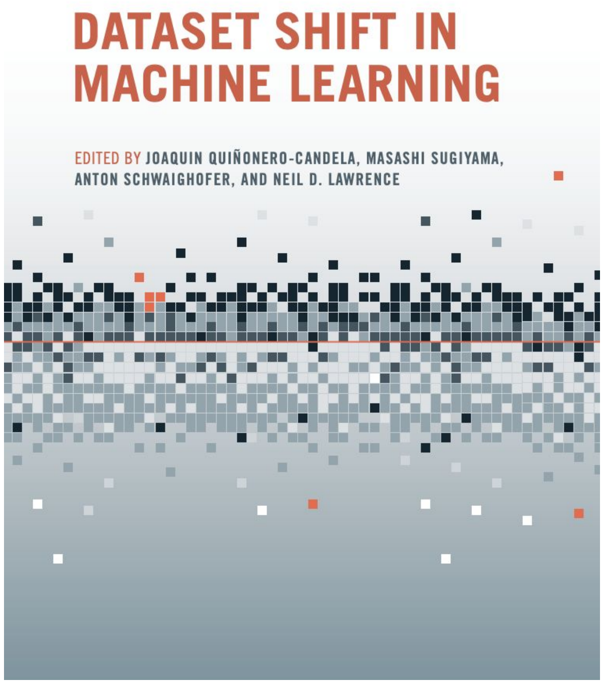
Take home messages

1. Uncertainty under dataset shift is worth worrying about.
2. Better calibration and accuracy on i.i.d. test dataset does not usually translate to better calibration under dataset shift.
3. Bayesian neural nets (e.g. SVI) are promising on MNIST/CIFAR but difficult to use on larger datasets (e.g. ImageNet) and complex architectures (e.g. LSTMs).
4. Relative ordering of methods is mostly consistent (except for MNIST)
5. Deep ensembles are more robust to dataset shift & consistently perform the best across most metrics; relatively small ensemble size (e.g. 5) is sufficient.

Take home messages

- Dataset shift is not new in ML!
  - *Dataset Shift in Machine Learning*, Sugiyama et al., 2009
  - But largely ignored in deep learning…

- We can learn a lot from revisiting pre-deep learning era work



Thanks!

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

https://arxiv.org/abs/1906.02530

Code + Predictions available online

https://github.com/google-research/google-research/tree/master/uq_benchmark_2019
Short URL: https://git.io/Je0Dk