UNCERTAINTY FOR DEEP IMAGE CLASSIFIERS ON OUT-OF-DISTRIBUTION DATA.

Anonymous authors
Paper under double-blind review

ABSTRACT

In addition to achieving high accuracy, in many applications, it is important to estimate the probability that a model prediction is correct. Predictive uncertainty is particularly important on out-of-distribution (OOD) data where accuracy degrades. However, models are typically overconfident, and model calibration on OOD data remains a challenge. In this paper we propose a simple post hoc calibration method that significantly improves on benchmark results (Ovadia et al., 2019) on a wide range of corrupted data. Our method uses outlier exposure to properly calibrate the model probabilities.

1 PREDICTIVE UNCERTAINTY

When a machine learning model makes a prediction, we want to know how confident (or uncertain) we should be about the result. Uncertainty estimates are useful for both in distribution and out-of-distribution (OOD) data. Predictive uncertainty addresses this challenge by endowing model predictions with estimates of class membership probabilities. The baseline method for predictive uncertainty is to simply use the softmax probabilities of the model as a surrogate for class membership probabilities (Hendrycks & Gimpel, 2017). Other approaches include temperature scaling (Guo et al., 2017), dropout (Gal & Ghahramani, 2016; Srivastava et al., 2014), and model ensembles (Lakshminarayanan et al., 2017), as well as Stochastic Variational Bayesian Inference (SVBI) for deep learning (Blundell et al., 2015; Graves, 2011; Louizos & Welling, 2016; 2017; Wen et al., 2018), among others.

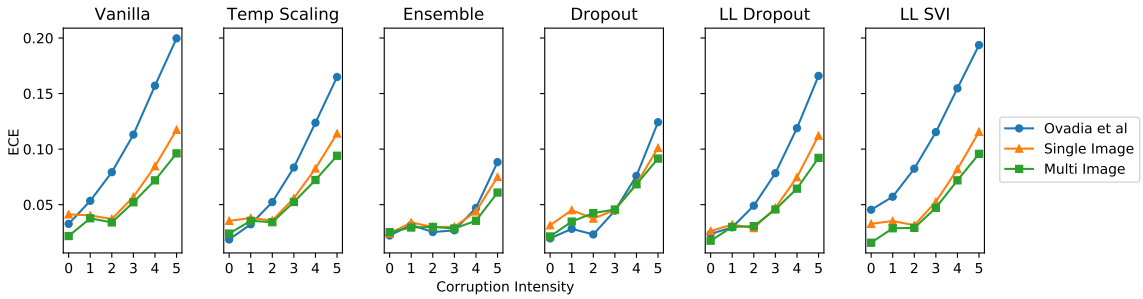
Our starting point is the work of Ovadia et al. (2019) which provides a large-scale benchmark of existing state-of-the-art methods for evaluating uncertainty on classification problems under dataset shift. We focus here on image classification problems, in particular the CIFAR-10 and Imagenet 2012 datasets with corruptions by Hendrycks & Dietterich (2019). One of main take-aways from the work of Ovadia et al. (2019) is that, unsurprisingly, the quality of the uncertainty predictions deteriorates significantly along with the dataset shift.

All methods suffer from some degree of calibration error, which is the difference between predicted error rates and actual error rates, as measured by collecting data into bins based on $p_{\max}$ bins. The standard measurement of calibration error is the expected calibration error (ECE) Guo et al. (2017), although other measures have been used (see Nguyen et al. (2015), Hendrycks & Gimpel (2017)), including the Brier score DeGroot & Fienberg (1983), which is also used in Ovadia et al. (2019).

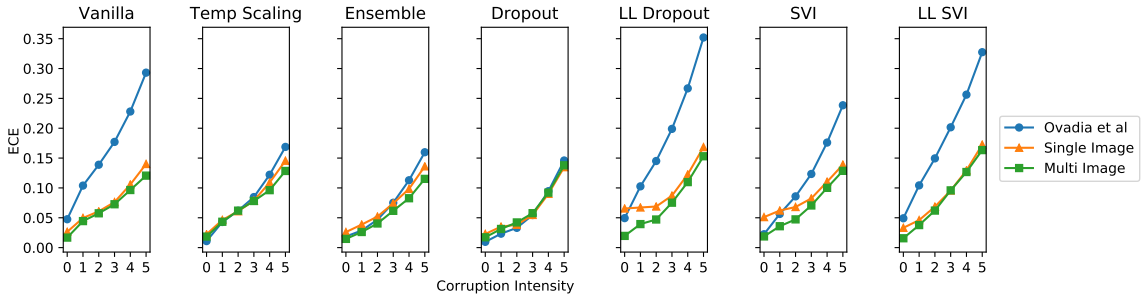
1.1 OUR RESULTS

We improve model calibration by adding a simple statistical calibration step to each of the models, converting the model outputs into calibrated probability estimates. We focus on the probability of correct classification, using the $p_{\max}$ values. We use several surrogate calibration datasets, each chosen to represent typical corruption intensities. The calibration is done by estimating the level of corruption of a given image (or sample of images). Then we convert the model $p_{\max}$ values to error probabilities, based on the estimated corruption intensity.

In Figure 1 we compare each benchmark uncertainty method with our two methods. The single image method determines the appropriate calibration set using only a single image. The multiple image method uses a sample of images (all drawn from the same corruption level and type) to



(a) ImageNet



(b) CIFAR-10

Figure 1: Comparison of the benchmark implementation (Ovadia et al., 2019), versus our single and multiple image methods. Mean Expected Calibration Error (ECE) across different corruptions types, for fixed corruption intensity going from 0 to 5. Each box represents a different uncertainty method. The ECE decreases across almost all methods and levels of intensity with the greatest improvement at the higher intensities. See Tables 1 and 2 in the Appendix for numerical comparisons.

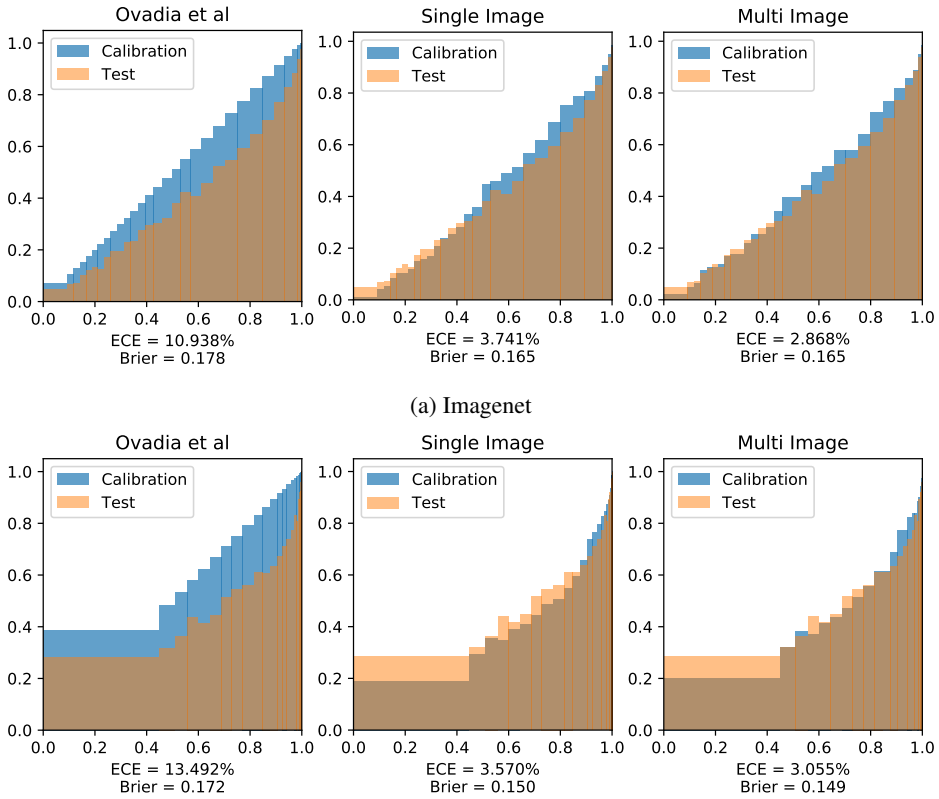
choose the calibration set. More images allow for a better choice of calibration set, further reducing the calibration error.

As shown in the figure, the ECE decreases across almost all methods and levels of intensity with the greatest improvement at the higher intensities. The Brier scores give similar results (see Figure 4 below). We also reproduce a figure in Ovadia et al. (2019) which gives a whisker plot of the distribution of the values.

Our results are based on the fact that data distribution shift typically leads to overconfident models Nguyen et al. (2015). We make the observation that overconfident models have $p_{\max}$ values which are *above* the true probability, which correspond to a shifted $p_{\max}$ distribution shift. Thus we use $p_{\max}$ distribution shift as a surrogate for data distribution shift. This allows us to significantly reduce calibration error using a purely *statistical approach*. In practice, we perform the model recalibration for different corruption intensities based simply on the $p_{\max}$ distribution shift, which we detect using surrogate corrupted datasets to calibrate the models on other corruptions. The calibrated probabilities are visualized using histograms in Figure 2.

1.2 Other Related Work

Models trained on a given dataset are unlikely to perform as well on a shifted dataset (Hendrycks & Dietterich, 2019). Moreover, there are inevitable tradeoffs between accuracy and robustness (Chun et al., 2020). Training models against corruptions can fail to make models robust to new corruptions (Vasiljevic et al., 2016; Geirhos et al., 2018). Hendrycks et al. (2019) deals with anomaly detection, the task of distinguishing between anomalous and in-distribution data. They propose an approach called Outlier Exposure (OE) that consists in training anomaly detectors on an auxiliary dataset of outliers. Similarly in Hendrycks et al. (2020), the authors propose AUGMIX, a method that improves both robustness and uncertainty measures by exposing the model to perturbed images during training.



(a) Imagenet

(b) CIFAR-10

Figure 2: Visualization of calibration errors for the Vanilla method on corrupted images on ImageNet and CIFAR-10 using the elastic transform corruption with intensity 4. The x-axis corresponds to the $p_{\max}$ values and the y-axis to the confidence estimates of p^{correct}. The blue histogram is the calibration probabilities, the orange is the test probabilities, and the brown is where both overlap. The ECE is large (lower is better) for the Vanilla method at higher corruption levels, due to the probability shift. The Brier score (lower is better) is also improved. The gap between the orange and blue curves represents the calibration error. Notice that we used 30 equally sized bins, so in the CIFAR-10 plot, there are very few values below .4, which is why the first bin is wide.

2 Background

2.1 Classification and Label Probabilities

Predictive uncertainty seeks to estimate the probability that an input x belongs to each class,

$$p_k^{class}(x) = \mathbb{P}\left[y = k \mid x\right], \qquad \text{for each } k \in \mathcal{Y}. \tag{1}$$

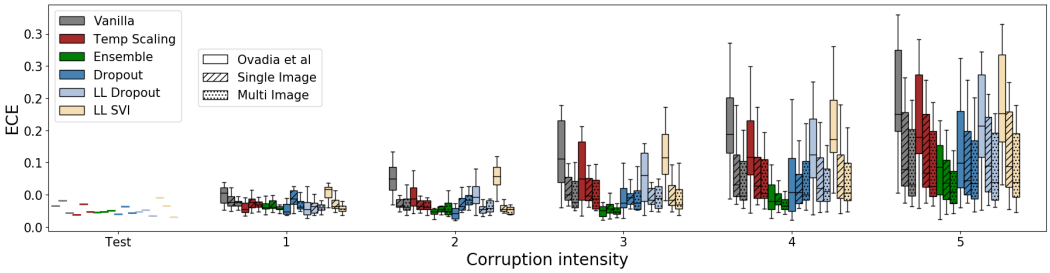
Here we write $x \in \mathcal{X}$ for data, $y \in \mathcal{Y} = \{1, \ldots, K\}$ for labels.

In the benchmark methods, the softmax of the model outputs, $p^{\text{softmax}}(x) = \text{softmax}(f(x))$ are used as a surrogate for the class probabilities. In the case of the vanilla and temperature scaling methods, $f(x)$ is simply the model outputs. Similarly, for the Ensemble or Dropout methods, $p^{\text{softmax}}(x)$ represents the average of the probability vectors over the multiple models or queries of the model, respectively.

Generally speaking, these softmax values are not an accurate prediction of the class probabilities $p_k^{class}(x)$ (Domingos & Pazzani, 1996). Here we focus on the *correct classification*

$$p^{\text{correct}}(x) = \mathbb{P}\left[y = \hat{y}(x)\right] \tag{2}$$

(a) ImageNet



(b) CIFAR-10

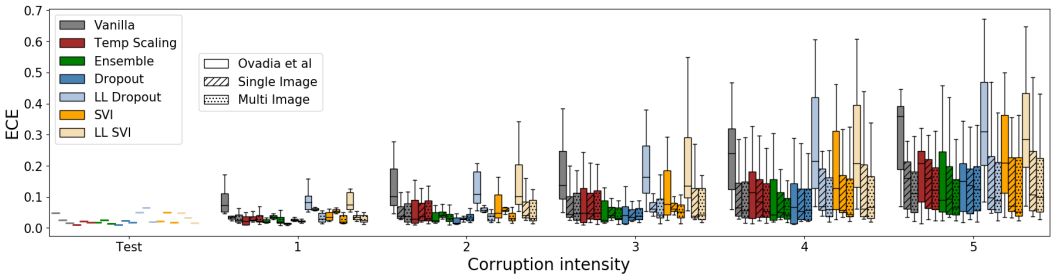


Figure 3: Comparison of the benchmark implementation (Ovadia et al., 2019) versus our single and multiple image methods. Expected Calibration Error (ECE) distribution across different corruptions types, for fixed corruption intensity going from 0 to 5. Each box represents a different uncertainty method.

where the classification of the model is given by $\hat{y}(x) = \arg\max f_i(x)$. Guo et al. (2017) showed that $p_{\max} = \max_i p_i^f$ usually overestimates p^{correct}. We can extend our method to top 5 correctness, as well as to other quantities of interest, such as, using different binning methods, based on structured predictors (Kuleshov & Liang, 2015).

3 METHOD

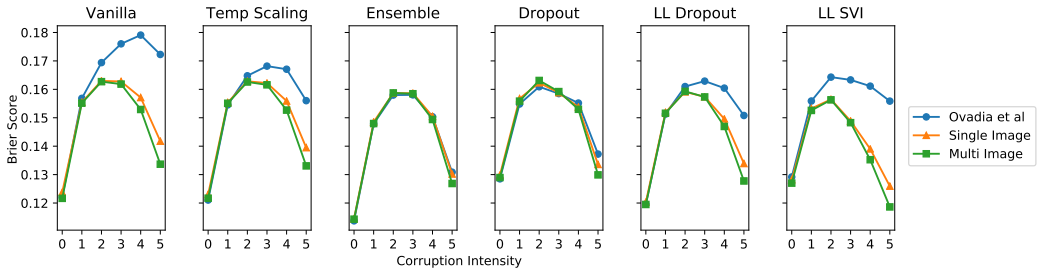
Problem definition We are given a model (or ensemble of models) trained on a given dataset ρ_{train}. We want to have calibrated error estimates on an unknown (different) data set ρ_{deploy}, which could have different levels of corruption. Since one model cannot be calibrated on all of the possible (different) datasets, we want to allow for multiple calibrations, and apply them accordingly.

We study two cases: (i) we have a single image drawn from an unknown distribution, or (ii) we have multiple images, each drawn from the same unknown distribution. In the latter case, we use the full test set (we obtain similar results using 100 images).

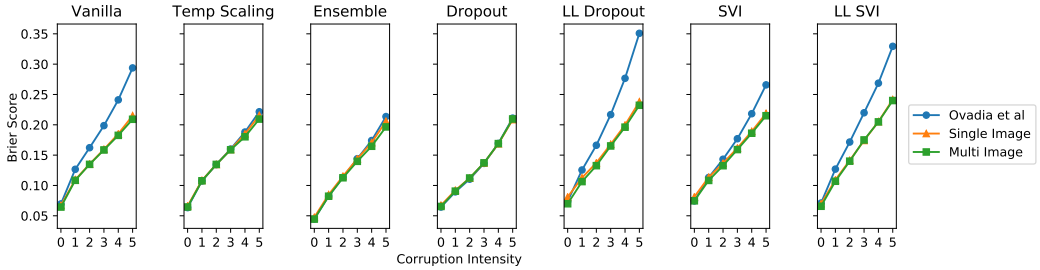
3.1 CALIBRATING FOR DATASET SHIFT

We choose C distinct calibration sets generated from shifted distributions $\rho^{\text{CAL},j}$. These sets are chosen to have different representative degrees of corruption intensity. Each calibration set leads to a different uncertainty estimate for a given $p_{\max}$ value. We adaptively choose the calibration set given an image (Single Image Method) or a test set of images (Multiple Image Method).

For each calibration dataset $S^{\text{CAL},j}$, we convert the $p_{\max}$ values on that calibration set into the p_{correct} values, as follows.



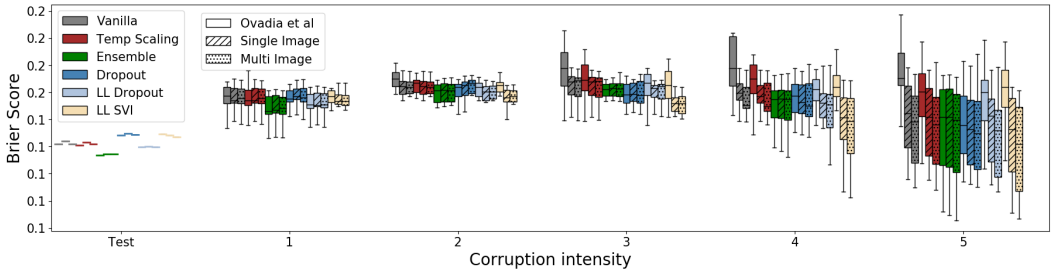
(a) ImageNet



(b) CIFAR-10

Figure 4: Comparison of the benchmark implementation (Ovadia et al., 2019) versus our single and multiple image methods. Mean Brier score across different corruptions types, for fixed corruption intensity going from 0 to 5. Each box represents a different uncertainty method. See Tables 3 and 4 for numerical comparisons.

(a) ImageNet



(b) CIFAR-10

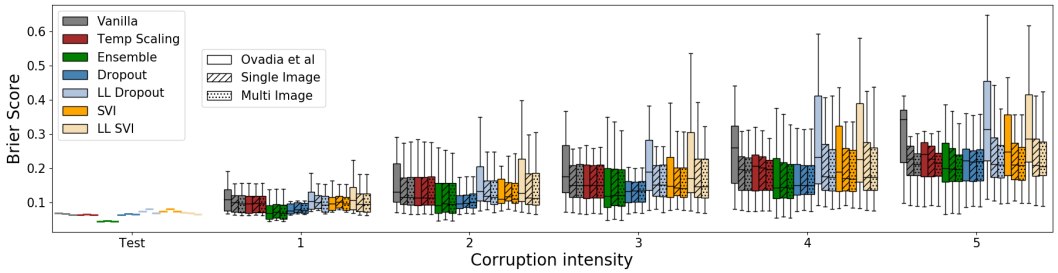


Figure 5: Comparison of the benchmark implementation (Ovadia et al., 2019) versus our single and multiple image methods. Brier score distribution across different corruptions types, for fixed corruption intensity going from 0 to 5. Each box represents a different uncertainty method. See Tables 3 and 4 in the Appendix for numerical comparisons.

(i) Evaluate the model $p_{\max}$ values on the calibration set, $P_m^{\mathrm{CAL},j} = \{p_{\max}(x) \mid x \in S_m^{\mathrm{CAL},j}\}$. Record the probability density $h^{\mathrm{CAL},j}(x)$, of the $p_{\max}$ values as a histogram, by binning the $p_{\max}$ using equally sized bins.

(ii) Define the calibrated model probabilities by

$$p_{\mathrm{correct}}^{\mathrm{CAL},j}(x) = \mathbb{P}\left[y = \hat{y}(x) \mid p_{\max}(x)\right], \tag{3}$$

using ground truth labels on the calibration set $S^{\mathrm{CAL},j}$. These probabilities are represented using a histogram. See (Oberman et al., 2020) for more details.

Single Image Method: Given a single image x drawn from an unknown distribution ρ_{deploy}.

(i) Compute the likelihood, $L_j(p_{\max}(x))$ of $p_{\max}$ under each of the calibration models, using the probability density $h^{\mathrm{CAL},j}(x)$. The probability that the $p_{\max}$ value came from a given calibration set, making the standard assumption that the *a priori* likelihoods of the calibration sets are all equal, is

$$q_i(p_{\max}) = \frac{L_j(p_{\max})}{\sum_{j=1}^{C} L_j(p_{\max})}$$

(ii) Then the calibrated probability, conditional on each of the calibration sets, is given by

$$p_{\mathrm{correct}}^{\mathrm{deploy}}(x) = \sum_{j=1}^{C} q_j(p_{\max})\, p_{\mathrm{correct}}^{\mathrm{CAL},j}(x)$$

Multiple Image Method: Given a sample of images $S_m^{\mathrm{deploy}} = \{x_1, \ldots, x_m\}$ drawn from an unknown distribution, ρ_{deploy}, where $m > 1$.

(i) Record the corresponding model $p_{\max}$ values, $P_m^{\mathrm{deploy}} = \{p_{\max}(x) \mid x \in S_m^{\mathrm{deploy}}\}$, and compute the mean μ.

(ii) Compare to the means μ_j of P_m^{CAL} and find the closest mean for the calibration sets, $i = \arg\min_j |\mu - \mu_j|$. Set

$$p_{\mathrm{correct}}^{\mathrm{deploy}}(x) = p_{\mathrm{correct}}^{\mathrm{CAL},i}(x).$$

3.2 Practical implementation of our method

In Ovadia et al. (2019) the additional methods used include (LL) Approximate Bayesian inference for the parameters of the last layer only (Riquelme et al., 2018), (LL SVI) Mean field stochastic variational inference on the last layer, (LL Dropout) Dropout only on the activations before the last layer. We refer to Ovadia et al. (2019) for more details on how each method was implemented. All these methods ultimately use the softmax probabilities as a surrogate for the class probabilities. The difference between the methods is how these probabilities are obtained.

The distributional shift on ImageNet used 16 corruption types with corruption intensity on a scale from 0 to 5: various forms of Noise, and Blur as well as Pixelate, Saturate, Brightness, Contrast, Fog and Frost, etc. See Figure S3 in (Ovadia et al., 2019).

We used the published softmax value of each of the methods from the benchmark dataset in Ovadia et al. (2019). We selected Contrast as the corruption to use for calibration (removing it from the test set). We calibrated our models using equally sized bins, on each of the calibration sets determined by combinations of intensity given by: {0}, {5,0}, {5,4,0}, {5,4,3,0}, {5,4,3,2,0}, and {5,4,3,2,1,0}. Heuristically we always want clean images in our calibration set while having different shifted means. For both CIFAR-10 and Imagenet, we used 5000 images, across each of the intensity levels, for calibration. A more sophisticated combination of calibration sets could lead to improvements.

We reproduced the figures in Ovadia et al. (2019) with a small adjustment: we measure the ECE and Brier score just for p_{correct}, rather than for p_k^{class}. However this made a negligible difference to the values. In addition, ECE can depend on the binning procedures (equally spaced or equally

Table 1: Comparison on Imagenet of the benchmark implementation (Ovadia et al., 2019) versus our single and multiple image methods. Numerical values of the means of ECE scores across different corruptions types, for fixed corruption intensity going from 0 to 5.

Method		Test	Corruption Intensity 1	2	3	4	5
Vanilla	(Ovadia et al.)	0.0327	0.0534	0.0792	0.1130	0.1571	0.1997
	(Single Image)	0.0413	0.0403	0.0371	0.0570	0.0846	0.1173
	(Multi Image)	**0.0218**	**0.0377**	**0.0340**	**0.0521**	**0.0719**	**0.0962**
Temp Scaling	(Ovadia et al.)	**0.0187**	**0.0323**	0.0523	0.0835	0.1237	0.1649
	(Single Image)	0.0354	0.0380	0.0358	0.0558	0.0825	0.1140
	(Multi Image)	0.0240	0.0355	**0.0342**	**0.0524**	**0.0722**	**0.0940**
Ensemble	(Ovadia et al.)	**0.0222**	0.0318	**0.0253**	**0.0269**	0.0470	0.0883
	(Single Image)	0.0240	0.0340	0.0299	0.0301	0.0439	0.0749
	(Multi Image)	0.0253	**0.0294**	0.0299	0.0286	**0.0355**	**0.0609**
Dropout	(Ovadia et al.)	**0.0196**	**0.0282**	**0.0232**	**0.0446**	0.0759	0.1242
	(Single Image)	0.0314	0.0451	0.0375	0.0452	0.0687	0.1011
	(Multi Image)	0.0214	0.0348	0.0424	0.0456	**0.0683**	**0.0915**
LL Dropout	(Ovadia et al.)	0.0236	**0.0296**	0.0491	0.0783	0.1189	0.1659
	(Single Image)	0.0263	0.0321	**0.0289**	0.0471	0.0747	0.1121
	(Multi Image)	**0.0176**	0.0300	0.0306	**0.0457**	**0.0644**	**0.0921**
LL SVI	(Ovadia et al.)	0.0454	0.0571	0.0824	0.1153	0.1547	0.1937
	(Single Image)	0.0327	0.0353	0.0315	0.0525	0.0820	0.1156
	(Multi Image)	**0.0158**	**0.0287**	**0.0291**	**0.0472**	**0.0719**	**0.0957**

sized). Equally sized bins are more effective for calibration since they reduce statistical error. They do however lead to different bin edges on different calibration sets, which required combining the bins. This can be done by refinement (which leads to ugly plots) or simple by one dimensional density estimation (Wasserman, 2006, Chapter 6), which we used here.

3.3 Discussion of the results

Both our single and multi image methods improve on the benchmark widely across methods and corruption intensity levels, as shown in Figure 1 and Tables 1, and 2, which report the mean ECE scores for each model and across different corruption types, for fixed corruption intensity going from 0 to 5. The multi image method performs better than the single image with a few exceptions. This is a natural consequence of using more images to better estimate the corruption level. While the ensemble method remains overall the best method, the gap to other methods is significantly reduced.

We point out that our calibration for the ensemble method was limited by a technical limitation. The published data set Ovadia et al. (2019) released only the ensemble averages (presumably for practical reasons since the database was large). However the individual model $p_{\max}$ values were lost (similarly for Dropout). When we recalibrate based on the average we have lost information, so we cannot calibrate as effective as if the individual model $p_{\max}$ were available. We note that instead of requiring class predictions for all classes, our method only requires $p_{\max}$ values, which is a considerably smaller data set. Moreover, we compile these values into a histogram with 30 bins, makes the added cost of our method negligible.

4 Conclusions

Increasingly we are asking models trained on a given dataset to perform on out of distribution data. Our work focused on uncertainty estimates, in particular, an estimate of the probability that our model classification is correct. In contrast to most deep uncertainty work, we use a purely statistical approach to reduce the calibration error of deep image classifiers under dataset shift. The approach is model agnostic, so it can be applied to future models.

Table 2: Comparison on CIFAR-10 of the benchmark implementation Ovadia et al. (2019) versus our single and multiple image methods. Numerical values of the means ECE scores across different corruptions types, for fixed corruption intensity going from 0 to 5.

Method		Test	Corruption Intensity				
			1	2	3	4	5
Vanilla	(Ovadia et al.)	0.0475	0.1039	0.1388	0.1771	0.2280	0.2931
	(Single Image)	0.0264	0.0496	0.0608	0.0762	0.1057	0.1401
	(Multi Image)	**0.0168**	**0.0444**	**0.0575**	**0.0726**	**0.0965**	**0.1206**
Temp Scaling	(Ovadia et al.)	**0.0110**	**0.0429**	0.0618	0.0847	0.1221	0.1687
	(Single Image)	0.0224	0.0463	**0.0609**	0.0788	0.1090	0.1455
	(Multi Image)	0.0185	0.0433	0.0619	**0.0781**	**0.0962**	**0.1284**
Ensemble	(Ovadia et al.)	0.0187	0.0287	0.0468	0.0752	0.1128	0.1598
	(Single Image)	0.0260	0.0386	0.0520	0.0742	0.0983	0.1361
	(Multi Image)	**0.0146**	**0.0261**	**0.0404**	**0.0616**	**0.0827**	**0.1152**
Dropout	(Ovadia et al.)	**0.0097**	**0.0233**	**0.0331**	**0.0546**	0.0945	0.1460
	(Single Image)	0.0229	0.0351	0.0383	0.0548	**0.0900**	**0.1347**
	(Multi Image)	0.0173	0.0313	0.0422	0.0577	0.0921	0.1379
LL Dropout	(Ovadia et al.)	0.0494	0.1026	0.1451	0.1989	0.2667	0.3520
	(Single Image)	0.0655	0.0672	0.0689	0.0871	0.1226	0.1684
	(Multi Image)	**0.0196**	**0.0394**	**0.0471**	**0.0753**	**0.1098**	**0.1530**
SVI	(Ovadia et al.)	0.0224	0.0564	0.0859	0.1234	0.1760	0.2385
	(Single Image)	0.0511	0.0621	0.0680	0.0824	0.1103	0.1389
	(Multi Image)	**0.0187**	**0.0358**	**0.0473**	**0.0706**	**0.1001**	**0.1286**
LL SVI	(Ovadia et al.)	0.0491	0.1042	0.1495	0.2016	0.2563	0.3274
	(Single Image)	0.0332	0.0461	0.0685	**0.0957**	0.1294	0.1725
	(Multi Image)	**0.0157**	**0.0378**	**0.0620**	0.0959	**0.1267**	**0.1634**

Previous work has shown that uncertainty estimates degrade on corrupted data, as measured by the expected calibration error. The greater the mismatch between training data and deployment data, the greater the degradation of uncertainty estimates.

We overcome this limitation introducing a method which allows given model to be better calibrated to multiple corruption intensities. Our method works by no longer requiring that model outputs approximate class probabilities. We add a simple extra calibration step, and detect the level of corruption of data, which allows use calibrations tuned to the corruption level of the data.

REFERENCES

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622. PMLR, 2015.

Sanghyuk Chun, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. *arXiv preprint arXiv:2003.03879*, 2020.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Pedro Domingos and Michael Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classi er. In *Proc. 13th Intl. Conf. Machine Learning*, pp. 105–112, 1996.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/gal16.html`.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in neural information processing systems*, pp. 7538–7550, 2018.

Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2348–2356. Curran Associates, Inc., 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1321–1330, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Volodymyr Kuleshov and Percy Liang. Calibrated structured prediction. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3474–3482, 2015. URL `http://papers.nips.cc/paper/5658-calibrated-structured-prediction`.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 6405–6416, 2017.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. volume 48 of *Proceedings of Machine Learning Research*, pp. 1708–1716, New York, New York, USA, 20–22 Jun 2016. PMLR.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2218–2227, 2017.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640. URL https://doi.org/10.1109/CVPR.2015.7298640.

Adam Oberman, Chris Finlay, Alexander Iannantuono, and Tiago Salvador. Calibrated top-1 uncertainty estimates for classification by score based models. In *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 13991–14002. Curran Associates, Inc., 2019.

Carlos Riquelme, George Tucker, and Jasper Roland Snoek. Deep bayesian bandits showdown. In *Proceedings of the International Conference on Learning Representations*, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014.

Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In *International Conference on Learning Representations*, 2018.

A Tables of Brier Metrics

Table 3 and Table 4 report the mean Brier scores for each model and dataset across different corruption types, for fixed corruption intensity going from 0 to 5. The Brier scores can be computed directly from the data, without binning. The ranking provided by the Brier scores is quite similar that provided by the ECE. On ImageNet the only difference is Ensemble at corruption level 1. On CIFAR-10 there were two ranking differences.

Table 3: Comparison on Imagenet of the benchmark implementation (Ovadia et al., 2019) versus our single and multiple image methods. Numerical values of the means of the Brier scores across different corruptions types, for fixed corruption intensity going from 0 to 5.

Method		Test	Corruption Intensity 1	2	3	4	5
Vanilla	(Ovadia et al.)	0.1220	0.1568	0.1694	0.1760	0.1791	0.1723
	(Single Image)	0.1238	0.1554	0.1630	0.1628	0.1571	0.1418
	(Multi Image)	**0.1216**	**0.1551**	**0.1627**	**0.1618**	**0.1529**	**0.1337**
Temp Scaling	(Ovadia et al.)	**0.1210**	**0.1545**	0.1648	0.1682	0.1671	0.1560
	(Single Image)	0.1231	0.1554	0.1628	0.1622	0.1559	0.1395
	(Multi Image)	0.1217	0.1551	**0.1626**	**0.1616**	**0.1527**	**0.1331**
Ensemble	(Ovadia et al.)	**0.1137**	**0.1479**	**0.1580**	**0.1580**	0.1504	0.1309
	(Single Image)	0.1145	0.1485	0.1587	0.1586	0.1505	0.1302
	(Multi Image)	0.1143	0.1480	0.1587	0.1584	**0.1494**	**0.1269**
Dropout	(Ovadia et al.)	**0.1285**	**0.1548**	**0.1609**	**0.1585**	0.1552	0.1372
	(Single Image)	0.1298	0.1567	0.1623	0.1588	0.1538	0.1336
	(Multi Image)	0.1290	0.1558	0.1631	0.1592	**0.1530**	**0.1299**
LL Dropout	(Ovadia et al.)	0.1196	**0.1513**	0.1610	0.1629	0.1604	0.1508
	(Single Image)	0.1203	0.1519	**0.1591**	0.1574	0.1496	0.1339
	(Multi Image)	**0.1195**	0.1516	0.1593	**0.1573**	**0.1469**	**0.1278**
LL SVI	(Ovadia et al.)	0.1292	0.1559	0.1643	0.1633	0.1611	0.1559
	(Single Image)	0.1282	0.1532	0.1565	0.1489	0.1390	0.1259
	(Multi Image)	**0.1270**	**0.1526**	**0.1563**	**0.1483**	**0.1352**	**0.1186**

Table 4: Comparison of CIFAR-10 of the benchmark implementation Ovadia et al. (2019) versus our single and multiple image methods. Numerical values of means Brier scores across different corruptions types, for fixed corruption intensity going from 0 to 5.

Method		Test	Corruption Intensity 1	2	3	4	5
Vanilla	(Ovadia et al.)	0.0696	0.1266	0.1622	0.1986	0.2411	0.2938
	(Single Image)	0.0671	0.1095	0.1352	0.1596	0.1850	0.2151
	(Multi Image)	**0.0644**	**0.1085**	**0.1348**	**0.1585**	**0.1826**	**0.2091**
Temp Scaling	(Ovadia et al.)	**0.0635**	**0.1075**	**0.1344**	0.1600	0.1881	0.2215
	(Single Image)	0.0661	0.1084	0.1346	0.1592	0.1850	0.2155
	(Multi Image)	0.0645	0.1077	0.1347	**0.1587**	**0.1802**	**0.2095**
Ensemble	(Ovadia et al.)	0.0447	0.0824	0.1135	0.1441	0.1741	0.2136
	(Single Image)	0.0474	0.0848	0.1152	0.1439	0.1706	0.2059
	(Multi Image)	**0.0446**	**0.0822**	**0.1127**	**0.1398**	**0.1646**	**0.1965**
Dropout	(Ovadia et al.)	**0.0644**	**0.0894**	**0.1103**	**0.1364**	0.1692	0.2110
	(Single Image)	0.0672	0.0920	0.1122	0.1371	**0.1687**	**0.2085**
	(Multi Image)	0.0654	0.0906	0.1123	0.1372	0.1690	0.2096
LL Dropout	(Ovadia et al.)	0.0739	0.1255	0.1663	0.2167	0.2766	0.3508
	(Single Image)	0.0810	0.1127	0.1372	0.1679	0.1997	0.2380
	(Multi Image)	**0.0698**	**0.1064**	**0.1326**	**0.1651**	**0.1961**	**0.2322**
SVI	(Ovadia et al.)	**0.0744**	0.1131	0.1432	0.1770	0.2183	0.2658
	(Single Image)	0.0812	0.1135	0.1366	0.1617	0.1891	0.2188
	(Multi Image)	0.0747	**0.1083**	**0.1325**	**0.1592**	**0.1861**	**0.2150**
LL SVI	(Ovadia et al.)	0.0710	0.1270	0.1716	0.2198	0.2686	0.3294
	(Single Image)	0.0698	0.1092	0.1414	**0.1741**	0.2052	0.2415
	(Multi Image)	**0.0657**	**0.1069**	**0.1400**	0.1750	**0.2048**	**0.2400**