Uncertainty Quantification and Deep Ensembles

Rahul Rahaman Alexandre H. Thiery

Department of Statistics and Data Science
National University of Singapore

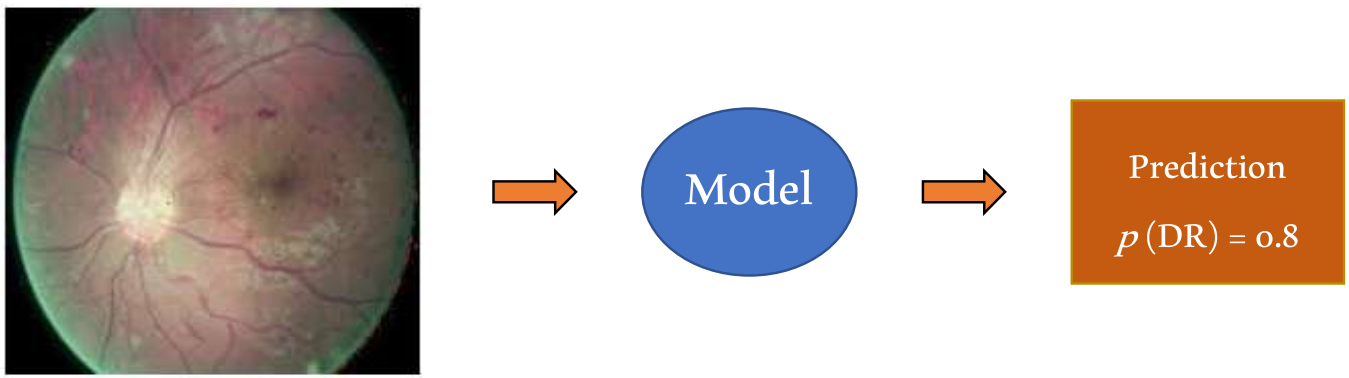
October 16, 2021

Contributions

- Contrary to popular belief, model averaging does not necessarily lead to better calibration. In fact in general it pushes the model towards under-confidence. In cases when the individual models are calibrated or under-confident, model averaging worsens the calibration performance.
- This observation is not limited to low-data setting, augmentations, or Deep ensembles. We show similar phenomenon for higher amount of data, calibrated models without the use of augmentation, and also other model averaging techniques.
- Simple post-processing calibration techniques can be tweaked to combat the under-confidence of ensembles. The ordering of pooling and calibration is extremely important, and can significantly impact model calibration.

Introduction: Uncertainty

Let us think of a model that predicts whether a patient have Diabetic Retinopathy (DR) by looking at fundus retina photographs.



After seeing an image, let us say that the model outputs a probability $p = 0.8$. What can a user (e.g. a Doctor) do with this number? Should one be 80% sure that the patient has DR, equivalently is it true that in 8 out of 10 such cases the patient will indeed have DR?

Introduction: Measures

- For a partition $0 = c_0 < \ldots < c_M = 1$ of the unit interval and a labelled set $\{x_i, y_i\}_{i=1}^N$, set $B_m = \{i : c_{m-1} < \widehat{p}(x_i) \leq c_m\}$ and $\text{acc}_m = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\widehat{y}(x_i) = y_i)$ and $\text{conf}_m = \frac{1}{|B_m|} \sum_{i \in B_m} \widehat{p}(x_i)$. The quantity ECE is defined as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \big| \, \text{conf}_m - \text{acc}_m \, \big|.$$

- Another widely used metric for calculating calibration is Brier Score [Bri50], which calculates the L_2 distance between the predictions $\text{p}(x)$ and its corresponding one-hot encoded target $\bar{y}$,

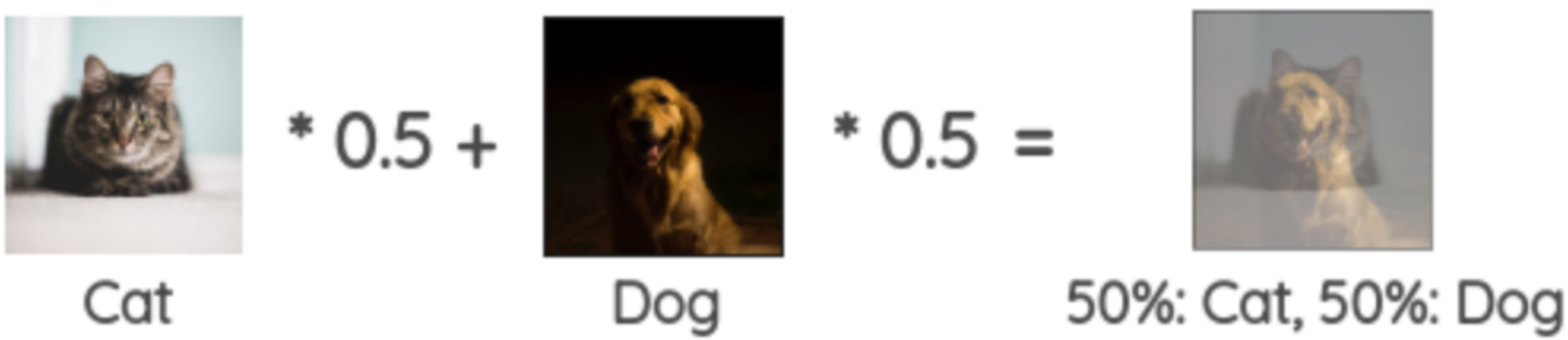
$$\text{Brier} := \frac{1}{N} \sum_{i=1}^{N} \|\text{p}(x_i) - \bar{y}_i\|_2^2$$

Methods: Augmentation

- Standard augmentation strategies include rotations, translations, brightness and contrast manipulations. Even simple augmentations help combat over-confidence.
- *Mixup* augmentation strategy [ZCDL17] augments a pair $(x, \overline{y}) \in \mathcal{X} \times \Delta_C$ to a different version $(x_\star, \overline{y}_\star)$ which is defined as

$$x_\star = \gamma\, x + (1-\gamma)\, x_J \qquad \text{and} \qquad \overline{y}_\star = \gamma\, \overline{y} + (1-\gamma)\, \overline{y}_J \qquad (1)$$

for a random coefficient $\gamma \in (0, 1)$.



Methods: Model Averaging

- Ensembling methods leverage a set of models by combining them into a aggregated model.
- Many practical model averaging techniques exist for Deep Learning such as Deep Ensembles [LPB17], SWAG [IPG+18, MGI+19], MC-DropOut [GG16] e.t.c.
- The general belief about Model Averaging is that, it naturally leads to Calibration.
- The common setups do not study already calibrated or under-confident models. Also, the interaction between augmentation and ensembling is missing.

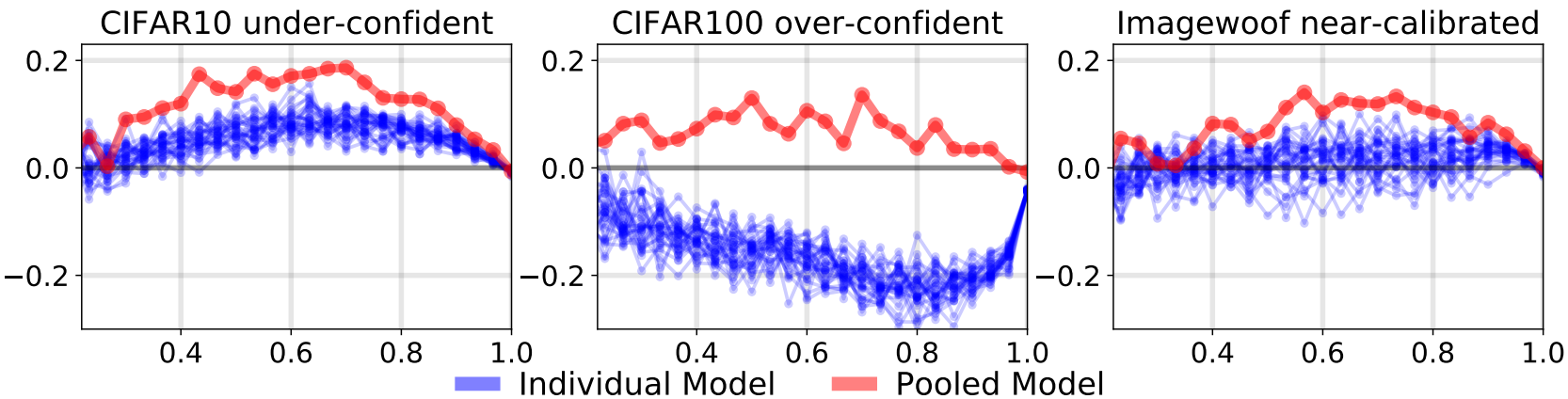
Methods: Post-processing Calibration

- *Temperature Scaling* is the simplest post-processing methods, that transforms the probabilistic outputs $\mathsf{p}(x) \in \Delta_C$ into a tempered version $\mathrm{Scale}[\mathsf{p}(x), \tau] \in \Delta_C$ defined through the scaling function

$$\mathrm{Scale}(\mathsf{p}, \tau) \equiv \sigma_{\mathrm{SM}}(\log \mathsf{p}/\tau),$$

  for a temperature parameter $\tau > 0$.
- The optimal parameter $\tau_\star > 0$ is usually found by minimizing a proper-scoring rules [GR07], often chosen as the negative log-likelihood, on a validation dataset.

Model Averaging and Calibration



- Studies show that Deep Ensembles, often leads to more accurate and better-calibrated predictions [LPB17, BC17, LPC$^+$15].
- Our findings suggest that instead Deep Ensembles (even others such as SWAG, MC-DropOut) always pushes the predictions towards under-confidence.
- We plot the calibration curve of 30 individual models (blue) and their final ensemble (red).

Underlying reasons

- Looking at the entropy functional, $\mathcal{H}(\mathrm{p}) = -\sum_{k=1}^{C} p_k \log p_k$, it can be seen that it is concave on the probability simplex Δ_C. Tempering with a temperature $\tau > 1$ will increase entropy, as can be proved by examining the derivative of the function $\tau \mapsto \mathcal{H}[\mathrm{p}^{1/\tau}]$.
- In a binary classification, with p_X being the model prediction for observation X, the measure

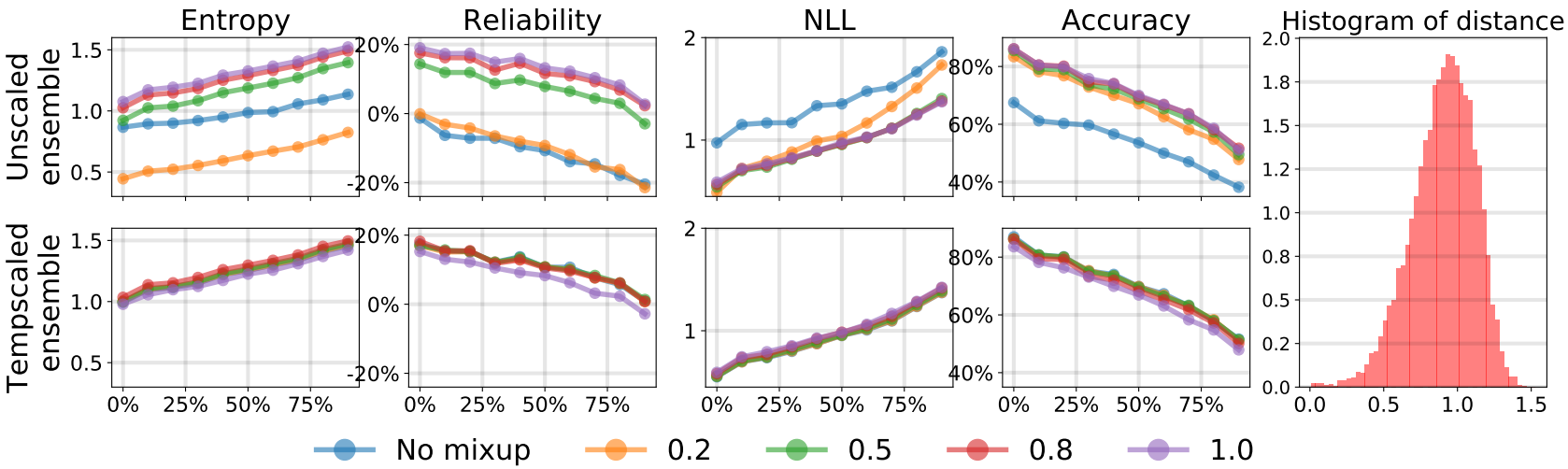
$$\mathrm{DC} := \mathbb{E}\Big[\big(1_{\{Y=1\}} - p_X\big)^2 - p_X(1-p_X)\Big].$$

  This metric captures deviation from calibration and can be shown to always decrease for an ensemble.

Distance To Training Data

- We first consider a mapping $\Phi : \mathbb{R}^{32,32} \to \mathsf{S}^{128}$, where $\mathsf{S}^{128} \subset \mathbb{R}^{128}$ denotes the unit sphere in $\mathbb{R}^{128}$, that maps an image to a low dimensional representation.
- We use the distance $d(x, y) = \|\Phi(x) - \Phi(y)\|_2$ between the 128-dimensional representations of the CIFAR10 images x and y. The distance of a test image x to the training dataset is defined as $\min\{d(x, y_i) \ : \ y_i \in \mathcal{D}_{\text{train}}\}$.

Distance To Training Data



- Deep Ensembles trained on $N = 1000$ CIFAR10 samples with different mixup strength. The 'x-axis' denotes distance percentile.
- All the metrics degrade with distance from training set.
- samples far away from training set are more over-confident than samples that are near.
- increasing the strength of mixup augmentation in general leads to better metrics.
- post-processing temperature scaling for the individual models almost washes-out all the differences due to mixup.

Possible Options

(A) Do nothing and hope that the averaging process intrinsically leads to better calibration

(B) Calibrate each individual network before aggregating all the results

(C) Simultaneously aggregate and calibrate the probabilistic forecasts of each individual model.

(D) Aggregate first the estimates of each individual model before eventually calibrating the pooled estimate.

Method [C], [D]

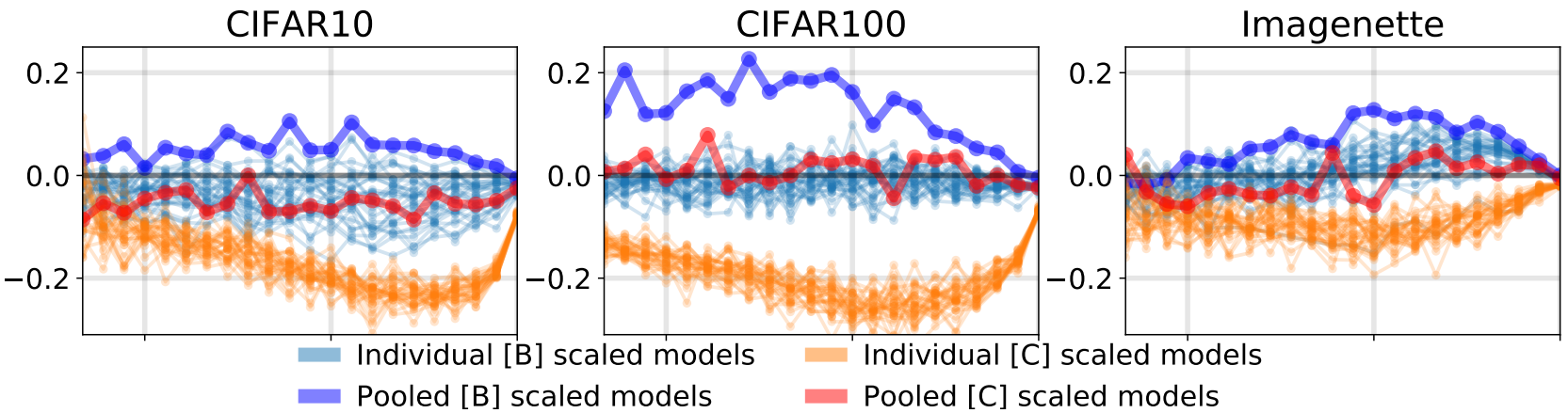
- **[C] Joint-pool-calibrate:** Learn the optimal temperature $\tau_\star$ jointly with pooling. $\tau_\star$ is found by minimizing a proper scoring rule $\text{Score}(\cdot)$ on a validation set $\mathcal{D}_{\text{valid}}$

$$\tau_\star = \arg\min \left\{ \tau \mapsto \frac{1}{\mathcal{D}_{\text{valid}}} \sum_{i \in \mathcal{D}_{\text{valid}}} \text{Score}(\text{p}_i^\tau, y_i) \right\},$$

  where $\text{p}_i^\tau = \textbf{Agg}\big[\text{Scale}(\text{p}^{1:K}(x_i), \tau)\big]$.
- **[D] Pool-then-calibrate:** Pool the predictions first and then fit a temperature $\tau_\star$ by a minimizing $\text{Score}(\cdot)$ on a $\mathcal{D}_{\text{valid}}$. Given a set $\text{p}^{1:K}$ of $K \geq 2$ predictions, the final prediction is defined as

$$\text{p}_\star \equiv \text{Scale}\left[\textbf{Agg}(\text{p}^{1:K}), \tau_\star\right].$$

Importance of the Pooling and Calibration order



- Calibration curve (x-axis confidence, y-axis Accuracy - confidence)
- **(light blue)** each model calibrated with one temperature per model (i.e. individually temperature scaled),
- **(dark blue)** ensemble of individually temperature scaled models (method **[B]**),
- **(orange)** each model scaled with a global temperature $\tau_\star$ obtained with method **[C]**,
- **(red)** final prediction of method **[C] Joint-pool-calibrate**.

Performance comparison

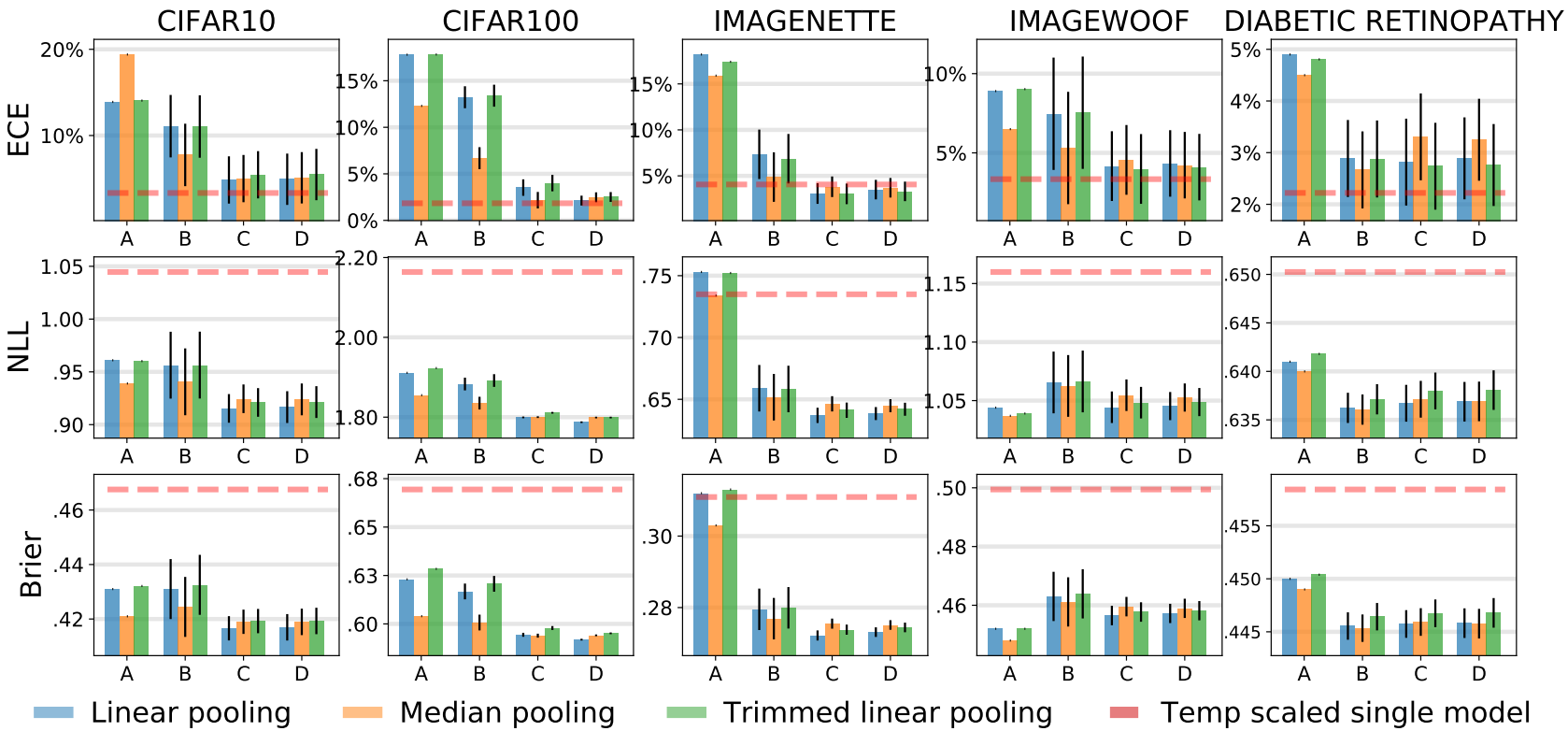


Figure: *Performance of different pooling strategies (A-D) with $K = 30$ models trained with mixup-augmentation ($\alpha = 1$) across multiple datasets. Experiments were executed 50 times on the same training data but different validation sets. The dashed red line represents a baseline performance when a single model was training with mixup augmentation ($\alpha = 1$) and post-processed with temperature scaling.*

Impact on low data setting

Metric	(Ours) 30 models temp scaled Augment + mixup	30 models mixup Augment	single model mixup Augment	single model no mixup Augment	single model no mixup no Augment
test acc	69.92 ± .04	**70.67**	66.45 ± .61	63.73 ± .51	49.85 ± .66
test ECE	**3.3** ± 1.9	13.9	7.03 ± .7	20.7 ± .4	23.4 ± 1.0
test NLL	**0.910** ± .012	0.961	1.03 ± .13	1.509 ± .017	1.770 ± .045
test BRIER	**0.414** ± .002	0.431	0.463 ± .005	0.556 ± .006	0.718 ± .009

Results on CIFAR10 1000 samples. The table breaks down individual component and justifies why it is necessary to employ both ensemble and mixup to achieve significant boost in performance especially in low-data regime. It also shows the model's journey from extreme over-confidence to calibration, then to extreme under-confidence and finally to calibrated and powerful ensemble.

Conclusions

- Contrary to popular belief, model averaging does not necessarily lead to better calibration. In fact in general it pushes the model towards under-confidence. In cases when the individual models are calibrated or under-confident, model averaging worsens the calibration performance.
- This observation is not limited to low-data setting, augmentations, or Deep ensembles. We show similar phenomenon for higher amount of data, calibrated models without the use of augmentation, and also other model averaging techniques.
- Simple post-processing calibration techniques can be tweaked to combat the under-confidence of ensembles. The ordering of pooling and calibration is extremely important, and can significantly impact model calibration.


Hamed Bonab and Fazli Can.
Less is more: a comprehensive framework for the number of components of ensemble classifiers.
arXiv preprint arXiv:1709.02925, 2017.


Glenn W Brier.
Verification of forecasts expressed in terms of probability.
Monthly weather review, 78(1):1–3, 1950.


Y. Gal and Z. Ghahramani.
Dropout as a bayesian approximation.
International Conference on Machine Learning, 2016.


Tilmann Gneiting and Adrian E Raftery.
Strictly proper scoring rules, prediction, and estimation.
Journal of the American statistical Association, 102(477):359–378, 2007.


P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson.
Averaging weights leads to wider optima and better generalization.

Uncertainty in Artificial Intelligence, 2018.


B. Lakshminarayanan, A. Pritzel, and C. Blundell.
Simple and scalable predictive uncertainty estimation using deep ensembles.
31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.


Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra.
Why m heads are better than one: Training a diverse ensemble of deep networks.
arXiv preprint arXiv:1511.06314, 2015.


W. Maddox, T. Garipov, P. Izmailov, D. Vetrov, and A. G. Wilson.
A simple baseline for bayesian uncertainty in deep learning.
arXiv preprint arXiv:1902.02476, 2019.


Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz.
mixup: Beyond empirical risk minimization.
CoRR, abs/1710.09412, 2017.