Unified Evaluation Of Neural Network Calibration & Refinement

Aditya Singh, Alessandro Bay, Andrea Mirabile
Zebra Technologies, London, UK
{firstname.lastname}@zebra.com

Abstract

Network calibration aims at creating deep networks which have predictive confidence representative of their predictive accuracy. Refinement accounts for the degree of separation between a network's correct and incorrect predictions. Both of these properties are highly desired from a deep learning model being deployed in critical settings such as medical analysis, automated driving, etc. However, recent approaches proposed for one have been studied in isolation from the other. In this paper, we aim to evaluate these independently studied solutions together. Firstly, we derive a simple linear relation between the two problems, thereby, linking calibration and refinement. This implies, improving calibration can help achieve a refined model and on the flip side, approaches focused on finding better ordinal ranking of predictions can help in improving calibration of networks. Motivated by this finding, we jointly benchmark various recently proposed approaches for the tasks of calibration and refinement. We find that the existing refinement approaches also provide significant improvement on calibration of the model while maintaining high degree of refinement.

1 Introduction

Deep neural networks are known to be highly uncalibrated (Guo et al. 2017). This implies that the model's confidence in its estimate is not reflective of its accuracy. Specifically, many studies have found that the networks produce high confidences for incorrectly classified samples (Guo et al. 2017; Pereyra et al. 2017). For scenarios such as automated driving, medical image analysis etc. where one wishes to avoid failures at all cost, such highly confident incorrect predictions can prove fatal. As a result, calibration is a desired property of the deployed neural networks which is being actively studied in deep learning research.

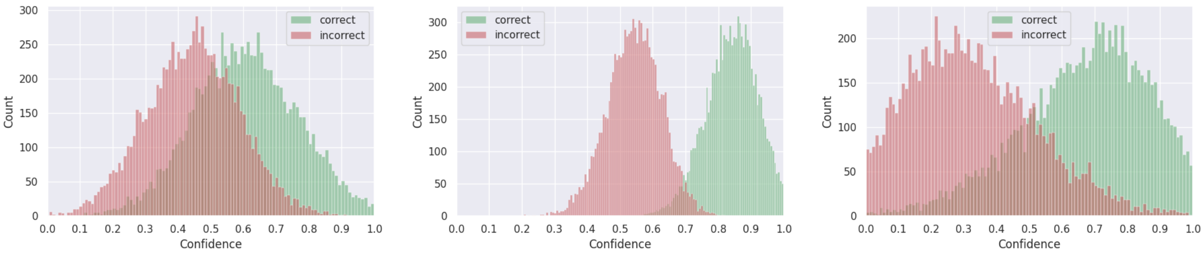
Refinement of a network's prediction is another such desired property. It has also been referred to as trustworthiness of a network (Jiang et al. 2018). Typically, the output after a softmax layer of a neural network is interpreted as confidence (Hendrycks and Gimpel 2017; Guo et al. 2017). The main focus is to find a scoring which provides trustworthy ordinal ranking of predictions or simply, a better segregation of incorrectly and correctly classified samples. Such a ranking can then allow the user to find an appropriate operating point based on refined scores to safely override a prediction.

Considering that both calibration and refinement allow the end-user to trust the predictions, the existing solutions rarely discuss these in tandem. Though commonly studied together in the domains of statistics (Murphy and Winkler 1977), meteorological forecast (Bröcker 2009), medical analysis (Gerds, Cai, and Schumacher 2008); for recent approaches proposed in the deep learning domain the joint importance has been sidelined for individual improvements. Before integrating both of these components directly into a study and propose a solution which preserves both of these properties, it is important to understand the underlying relationship between solving these two tasks. Subjectively, from Figure 1 we can assess that it is possible for a network to exhibit varying degree of association between the two properties. In (a) we have a classifier which is well calibrated but poorly refined. In this case, we can observe that the predictions are reliable towards the higher end of the confidence values. However, due to poor refinement there is a significant overlap between chunk of the correct and incorrect predictions. This causes majority of the correct predictions to be unreliable. For (b), we see that the predictions are well separated but not well calibrated. We can select an operating threshold for the network to make sure that we don't encounter many false-positives in practice, however, the remaining predictions being uncalibrated subsequently become unreliable. Case (c) shows an ideal scenario where the predictions are relatively well separated and calibrated.

The above visual inspection indicates a complex dynamic between the two desired properties. Also, it is possible for these to be satisfied concurrently which leads to a highly reliable network. However, currently the connection is not well established. As a result, in this paper, we aim to build a framework to understand calibration and refinement together. And, motivated by our findings, we assess the approaches proposed individually for each task under a unified setting.

Our contributions are as follows:

- We highlight the connection between Expected Calibration Error, a calibration metric, and area under the ROC curve computed for a classification task. This result forms the motivation for our cross-domain evaluation.



(a) Good calibration, poor refinement

(b) Poor calibration, good refinement

(c) Good Calibration, good refinement

Figure 1: Hypothetical classification results leading to different calibration and refinement scenarios. Figure a) has low calibration error ($ECE = 0.06$) but also low refinement score ($AUROC = 77.4\%$). In b) we have well refined outputs ($AUROC = 99.46\%$) but a poor calibration performance ($ECE = 0.18$). Lastly in figure c), the calibration error ($ECE = 0.08$) and the refinement score ($AUROC = 89\%$) is relatively better. The details of the metrics are provided in section 2.

.

- Based on the derived relationship, we discuss the cases where methods focusing on one task can positively or negatively impact the other.
- We evaluate respective state of the art methods which are studied in isolation under a unified setting.

The structure of the paper is as follows: In Section 2, we review the existing approaches proposed for calibration and improving refinement. In Section 3, we show that under reasonable assumptions the goal of minimising the calibration error falls in line with the goal of improving separability between correctly & incorrectly classified samples. We benchmark the state of the art methods across standard datasets to observe cross-domain impact in Section 4. We conclude our work in Section 5.

2 Related Work

Due to the high relevance of the problems addressed in our work, the amount of existing literature is abundant. We focus the discussion to non-bayesian approaches in this study.

2.1 Calibration

For shallower versions of neural networks Niculescu-Mizil and Caruana (2005) showed that the outputs are well-calibrated for a binary classification tasks. However, for deep neural networks sadly this is not the case. It has been shown that modern day networks are miscalibrated (Guo et al. 2017). Since then there have been many approaches to regularize the overconfident results of a model. The existing approaches introduce temperature scaling (Guo et al. 2017), negative entropy term to penalise the confident predictions (Pereyra et al. 2017), Mixup data-augmentation (Zhang et al. 2018; Thulasidasan et al. 2019) to reduce the predicted confidence estimate. Dirichlet based calibration (Kull et al. 2019) extends this idea to a classwise-calibration setting. The above mentioned approaches are either applied after the training(*post-hoc*) or applied while the learning is performed.

To measure the calibration of a classifier there are many metrics in practice. However, the ones predominantly used for modern neural networks are Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) (Naeini, Cooper, and Hauskrecht 2015). Kumar, Liang, and Ma (2020) proposed a post-hoc calibration approach mixing Platt scaling and histogram binning. They also extended the idea to a multi-class calibration. An alternative score often used is the Brier score (Brier 1950). It captures both calibrative and discriminative aspects of the predictions (Murphy 1973). However, the reported score is a summary of overall performance and can hide the underlying failures.

Majority of the solutions proposed for calibration only focus on lowering the predicted confidence estimate. They do not discuss the role of refining the predictions for improving calibration which is precisely our focus in this work.

2.2 Refinement

Obtaining meaningful confidence values from a network is a challenge which refinement seeks to solve. Many approaches have been proposed in this direction. Gal and Ghahramani (2016) used dropout (Srivastava et al. 2014) at test time to estimate predictive uncertainty by sampling predictions over multiple predictions. Lakshminarayanan, Pritzel, and Blundell (2017) use ensembles of neural networks to obtain useful confidence estimates. Moon et al. (2020) incorporated *Correctness Ranking Loss* (CRL) to allow network to learn ordinal rankings between classified samples. They also observed that CRL also helped in calibrating the network however, do not discuss the reasoning behind this observation. As a replacement for confidence estimate, Jiang et al. (2018) introduced TrustScore, which provides better ordinal ranking of predictions than the output of the network. They utilized the ratio between the distance from the sample to the nearest class different from the predicted class and the distance to the predicted class as the trust score. ConfidNet (Corbière et al. 2019) incorporates the learning of this 'trust' score as an additional branch in the network. In the post-hoc stage, ConfidNet branch of the classifier is trained to predict a confidence score which mimics the reliability of the network on its prediction.

Important metrics utilized to measure ordinal rankings are (i) area under the ROC curve (AUROC) (Corbière et al. 2019) (ii) area under the precision-recall curve (AUPR) (Moon et al. 2020) (iii) excess area under the risk–coverage curve (E-AURC) (Geifman, Uziel, and El-Yaniv 2019)

3 Calibration & Refinement

A dataset is composed of tuples of inputs and targets represented as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x \in \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{1, 2, \dots K\}$. We represent the learnable weights of a network as θ. The output of a network is a multinoulli distribution for K possible outcomes. The predicted category and predicted confidence are respectively as:

$$\hat{y}_i = \operatorname*{argmax}_{k \in \mathcal{Y}} P(Y = k | x_i, \theta) \tag{1}$$

$$c_i = \max_{k \in \mathcal{Y}} P(Y = k | x_i, \theta). \tag{2}$$

c_i is referred to as either the winning probability or maximum class probability.

We focus on the problem of a single-class calibration and single-class refinement. This suggests both the tasks deal with only 2 categories, which are overall *correctly* classified samples (or positive category) and overall *incorrectly* classified samples (or negative category). We first explain the metrics employed and then subsequently explain the assumptions we make.

Expected Calibration Error is measured as the difference between the expected accuracy and predicted confidence. Formally,

$$ECE = \sum_m \frac{|B_m|}{n} \left[|\mathbb{E}\left[A\right]_m - C_m| \right], \tag{3}$$

where average confidence (C) and accuracy (A) is computed after splitting the predictions in to predefined m bins based on the predicted confidence. The choice for m varies across literature, most common value is 15. Large number of bins increases the variability of the results (Nixon et al. 2019), whereas, small number of bins only provide a coarser indication of the miscalibration (Kumar, Liang, and Ma 2020). ECE is 0 for a well calibrated model.

AUROC captures the concept of ordinal ranking nicely. It denotes the expectation that a uniformly drawn random positive is ranked higher than a uniformly drawn random negative sample. AUROC (r), is formally defined as

$$r = \int_0^1 tpr \ \mathrm{d}fpr, \tag{4}$$

where tpr is the true positive rate and fpr is the false positive rate. The maximum value that r can attain is 1 representing an ideal ranking scenario.

Assumptions: We assume that the number of bins, $m = 1$, for computing ECE. By lowering the bin value we are relaxing the lower-bound of the true calibration error (Kumar, Liang, and Ma 2020). Another assumption we make is that $\mathbb{E}\left[A\right] < C$. This is true in practice as for all deep neural networks the problem of calibration entails over-confident predictions.

Utilizing the assumptions, we can rewrite equation (3) as:

$$ECE = C - \mathbb{E}\left[A\right]. \tag{5}$$

For a binary classification task, it has been shown (Hernández-Orallo, Flach, and Ferri 2012; Flach and Kull 2015) that r and $\mathbb{E}\left[A\right]$ are linearly related. They show that

$$\mathbb{E}\left[A\right] = \pi(1 - \pi)(2r - 1) + \frac{1}{2}, \tag{6}$$

where π is the proportion of positive samples in the data. We can subtract the average of the predicted confidence, C, on both sides of equation (6) to obtain

$$\mathbb{E}\left[A\right] - C = \pi(1 - \pi)(2r - 1) + \frac{1}{2} - C. \tag{7}$$

By replacing the resulting left hand side with ECE from equation (5) we get

$$ECE = -\left[\pi(1 - \pi)(2r - 1) + \frac{1}{2} - C\right]. \tag{8}$$

Rearranging the right hand side provides us

$$ECE = \alpha C - \beta r - \gamma, \tag{9}$$

where $\alpha \geq \beta > 0$ and $\gamma \geq 0$. The equalities hold when $\pi = \frac{1}{2}$, indicating an equal proportion of correct and incorrect predictions in the data.

Equation (9) indicates a linear relationship between ECE and AUROC under the applied assumptions from which we can draw a number of observations:

1. Reducing the average prediction confidence of an over-confident model helps in lowering the calibration error. Many of the existing work for calibration predominantly work along this direction.
2. Increasing r (or AUROC) can also help in reducing ECE. To the best of our knowledge, proposed approaches for calibration have not taken this direction or mentioned it explicitly.
3. Approaches which emphasize on refinement focus on improving the separability of the misclassified samples. Based on the linear relationship, we can expect them to have a positive impact on ECE.

4 Method

4.1 Selected Approaches

We first list the methods which we consider for joint evaluation and then proceed to enlist the chosen benchmark datasets and implementation details

Calibration

- **ERL** (Pereyra et al. 2017): This method penalises the output distribution of the network by adding a regularising term based on the entropy of the predicted estimates.
- **LS** (Müller, Kornblith, and Hinton 2019): In this approach, the target one-hot encoded vector is replaced by an ϵ-smoothed vector. This has been shown to increase the calibration of the network by reducing overconfident predictions.
- **Mixup** (Thulasidasan et al. 2019): The authors highlighted the undocumented effect of Mixup (Zhang et al. 2018) on the calibration of a network.

	CIFAR-100			STL-10			CIFAR-10		
	Accuracy(↑)	ECE(↓)	AUROC(↑)	Accuracy(↑)	ECE(↓)	AUROC(↑)	Accuracy(↑)	ECE(↓)	AUROC(↑)
Baseline	72.07 ± 0.2	19.12 ± 0.13	85.18 ± 0.21	81.61 ± 0.2	11.55 ± 0.19	85.53 ± 0.7	92.96 ± 0.2	5.38 ± 0.15	92.5 ± 0.01
ERL	72.40 ± 0.19	16.8 ± 0.1	85.19 ± 0.3	82.38 ± 0.29	9.6 ± 0.41	86.57 ± 0.16	93.23 ± 0.01	4.41 ± 0.07	92.11 ± 0.4
LS	72.92 ± 0.43	**5.76 ± 0.56**	81.49 ± 0.27	81.99 ± 0.45	5.64 ± 0.02	85.13 ± 0.3	93.07 ± 0.2	7.4 ± 0.18	82.36 ± 1.23
Mixup	**73.12 ± 0.18**	6.87 ± 1.81	82.96 ± 0.27	**82.94 ± 0.08**	**3.46 ± 0.33**	85.9 ± 0.14	**93.46 ± 0.18**	4.16 ± 1.2	86.72 ± 0.8
CFN	72.07 ± 0.2	13.95 ± 2.7	86.0 ± 0.18	81.61 ± 0.2	9.23 ± 1.02	**86.64 ± 0.4**	92.96 ± 0.2	4.1 ± 0.2	92.55 ± 0.1
CRL	71.5± 0.2	12.5 ± 1.1	**88.11 ± 0.16**	79.5 ± 0.4	6.34 ± 1.19	85.29 ± 0.68	93.05 ± 0.37	**1.87 ± 0.21**	**92.59 ± 0.42**

Table 1: Calibration and refinement results aggregated over 3 runs. ↑ and ↓ indicate that for a particular metric higher and lower values are better respectively. All values presented are in percentage. Values in bold font indicate the best values w.r.t the corresponding metrics.

Refinement

- **CFN** (Corbière et al. 2019): This approach relies on a post-hoc training of an alternative scoring criteria which reflects the network's 'trust' in its prediction
- **CRL** (Moon et al. 2020): The authors introduced a loss motivated to improve ordinal ranking of misclassified samples.

We represent a model trained traditionally without any calibration or refinement-based enhancements as **Baseline**.

Datasets We have selected 3 popular image classification datasets in our study. These are

- CIFAR-100 (Krizhevsky 2009)
- CIFAR-10 (Krizhevsky 2009)
- STL-10 (Coates, Ng, and Lee 2011)

Implementation Details The deep neural network architecture we use is the VGG-16 (with batch normalisation). Many of the selected approaches use it in their respective experiments hence it provides a relatively even ground for conducting the study. We first split the training data into 'training' and 'validation' in the ratio $9 : 1$. We utilize the validation set to save the best model while training. We perform each experiment 3 times and report the average and standard deviation of their performance. For computing ECE, we use $m = 15$. We use official implementations where applicable otherwise rely on our own Pytorch (Paszke et al. 2019) based implementation. All the methods are trained for 300 epochs with a starting learning rate of 0.1 reduced by a factor of 10 at epochs 150 and 250. CFN is trained using the baseline model as obtained in the previous step and fine-tuned using the schema of Corbière et al. (2019).

4.2 Results & Discussion

Table 1 contains the results from our effort for unified evaluation.

Accuracy is an important factor of classifier as we wish to seek a classifier that is accurate in its prediction. Overall, calibration approaches appear to provide better results in terms of accuracy. **Mixup** being the best among these. Refinement based approaches achieve comparable or marginally lower accuracies than their corresponding baselines. This highlights a potential pitfall of existing refinement approaches. Though, not many refinement approaches exist for deep networks at the moment, a potential aspect to consider would be to achieve accuracy on par with the baseline model.

For calibration, approaches specifically designed for the task provide the lowest calibration error in 2 out 3 datasets. Refinement approaches also provide comparable reduction in calibration error. This supports our theory of achieving reduction in calibration error when AUROC in terms of predictions is improved. For CIFAR-10, **CRL** achieves the lowest calibration error.

Refinement based approaches provide the best ordinal ranking of its prediction. **CRL** provides better separability for 2 out of 3 datasets. Calibration approaches perform either comparable or worse than the baseline model with **LS** and **Mixup** as the worst performers. This inconsistency in separability is alarming. Our hypothesis is that many of the calibration methods naively reduce the confidence for all predictions. This shifts the high density regions of the output distribution towards the low confidence values which improves calibration but has neutral or negative impact on refinement. What we seek is a classifier which retains majority of correct predictions with high confidence but outputs relatively lower confidence values for possible incorrect predictions. We would like to explore experimentally verify our hypothesis in future work and learn the root cause of this refinement degradation.

5 Conclusion

Calibration and refinement are two important attributes of a safety critical system. Our aim was to highlight the existing shortcomings of the individual approaches and provide a link between the two tasks. Calibration methods provide better accuracy and reduced calibration error, however in terms of refinement, they perform poorly. On the other hand, refinement approaches do provide better separability from misclassified predictions and comparable reduction in calibration error but, they often fail to match the baseline accuracy. We hope by the juxtaposition of respective state of art methods, we can encourage the community to focus on these problems simultaneously. We also showed that increasing refinement can assure reduced calibration error. This result can serve as an alternative route for future joint calibration and refinement approaches.

References

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78(1): 1–3. ISSN 0027-0644. doi:10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Bröcker, J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society* 135(643): 1512–1519. ISSN 1477-870X. doi:10.1002/qj.456. URL http://dx.doi.org/10.1002/qj.456.

Coates, A.; Ng, A.; and Lee, H. 2011. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *AISTATS*. https://cs.stanford.edu/~acoates/papers/coatesleeng_aistats_2011.pdf.

Corbière, C.; Thome, N.; Bar-Hen, A.; Cord, M.; and Pérez, P. 2019. Addressing Failure Prediction by Learning Model Confidence. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 2902–2913. Curran Associates, Inc. URL http://papers.nips.cc/paper/8556-addressing-failure-prediction-by-learning-model-confidence.pdf.

Flach, P. A.; and Kull, M. 2015. Precision-Recall-Gain Curves: PR Analysis Done Right. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 838–846. Cambridge, MA, USA: MIT Press.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. New York, New York, USA: PMLR. URL http://proceedings.mlr.press/v48/gal16.html.

Geifman, Y.; Uziel, G.; and El-Yaniv, R. 2019. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. URL https://openreview.net/forum?id=SJfb5jCqKm.

Gerds, T. A.; Cai, T.; and Schumacher, M. 2008. The Performance of Risk Prediction Models. *Biometrical Journal* 50(4): 457–479. doi:10.1002/bimj.200810443. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200810443.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1321–1330. JMLR.org.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *Proceedings of International Conference on Learning Representations* .

Hernández-Orallo, J.; Flach, P.; and Ferri, C. 2012. A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. volume 13, 2813–2869. JMLR.org. ISSN 1532-4435.

Jiang, H.; Kim, B.; Guan, M. Y.; and Gupta, M. 2018. To Trust or Not to Trust a Classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 5546–5557. Red Hook, NY, USA: Curran Associates Inc.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32 (NIPS 2019)*. Neural Information Processing Systems Foundation. doi:http://papers.nips.cc/paper/9397-beyond-temperature-scaling-obtaining-well-calibrated-multi-class-probabilities-with-dirichlet-calibration. URL https://nips.cc/Conferences/2019.

Kumar, A.; Liang, P.; and Ma, T. 2020. Verified Uncertainty Calibration.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30, 6402–6413. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Moon, J.; Kim, J.; Shin, Y.; and Hwang, S. 2020. Confidence-Aware Learning for Deep Neural Networks.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32, 4694–4703. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf.

Murphy, A. H. 1973. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* 12(4): 595–600. ISSN 0021-8952. doi:10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2.

Murphy, A. H.; and Winkler, R. L. 1977. Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 26(1): 41–47. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346866.

Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 2901–2907. AAAI Press. ISBN 0262511290.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting Good Probabilities with Supervised Learning. In *Proceedings* *of the 22nd International Conference on Machine Learning*, ICML '05, 625–632. New York, NY, USA: Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pereyra, G.; Tucker, G.; Chorowski, J.; Łukasz Kaiser; and Hinton, G. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15(1): 1929–1958. ISSN 1532-4435.

Thulasidasan, S.; Chennupati, G.; Bilmes, J.; Bhattacharya, T.; and Michalak, S. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. In *NeurIPS*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=r1Ddp1-Rb.