On The Dark Side Of Calibration For Modern Neural Networks

Aditya Singh [1] **Alessandro Bay** [1] **Biswa Sengupta** [1] **Andrea Mirabile** [1]

Abstract

Modern neural networks are highly uncalibrated. It poses a significant challenge for safety-critical systems to utilise deep neural networks (DNNs), reliably. Many recently proposed approaches have demonstrated substantial progress in improving DNN calibration. However, they hardly touch upon refinement, which historically has been an essential aspect of calibration. We portray refinement as the separation between a DNN's correct and incorrect predictions. In this paper, we empirically highlight the downside of many modern calibration techniques. We find that many calibration approaches with the likes of label smoothing, mixup etc. lower the utility of a DNN by degrading its refinement. Even under natural data shift, this calibration-refinement trade-off holds for the majority of calibration methods. These findings call for an urgent retrospective into some popular pathways taken for modern DNN calibration.

1. Introduction

Guo et al. (2017) showed that modern deep neural networks are miscalibrated. This implies that the model's confidence in its estimate is not reflective of its accuracy. Typically, the output after a softmax layer of a neural network is interpreted as confidence (Hendrycks & Gimpel, 2017; Guo et al., 2017). Many studies have found that DNNs output 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. The output of a network is a probability distribution over K possible outcomes. The predicted category and predicted confidence are respectively denoted as

$$\hat{y}_i = argmax_{k \in \mathcal{Y}} P(Y = k | x_i, \theta) \quad (1)$$

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

where, $x \in \mathbb{R}^d$, $y_i \in \mathcal{Y} = \{1, 2, \dots K\}$ and L are the total number of samples in the dataset. Formally, calibration can be defined as

Definition 1.1 (Calibration)**.** A model P_θ is said to be calibrated if $\mathbb{P}(y_i = \hat{y}_i | c_i) = c_i \ \forall (x_i, y_i)$.

Refinement describes the separability of a binary classification problem (Murphy, 1973; Gneiting et al., 2007). To build trust, it can be interpreted as the separability of correct and incorrect predictions based on predicted confidence. This is very similar to how calibration is assessed using predicted estimate as the assigned confidence. Good refinement indicates an ordinal ranking of predictions which allows better segregation of correct predictions from incorrect ones (Moon et al., 2020). Such a ranking can then allow the user to find an appropriate operating point. Moreover, it also plays an important part in describing predictors' effectiveness. Formally,

Definition 1.2 (Refinement)**.** Let S^p and S^n denote correct and incorrect classification of a model on the test set D^t. Predictions are considered refined iff $c_i > c_j \ \forall x_i \in S^p \ , \ x_j \in S^n$.

This way refinement enforces a separation between the two sets of prediction. Degroot & Fienberg (1981) provide an alternative definition of refinement for calibrated classifiers.

Refinement and calibration have been an integral component for describing a trustworthy and reliable predictor(Gneiting et al., 2007; Murphy & Winkler, 1977; Van Calster et al., 2019), it raises an important question: 'How well do modern calibration approaches fare on refinement?'. The focus of our paper is to investigate this question.

2. Calibration & Refinement

We focus on the problem of calibration and refinement for a reduced binary setting. For a multi-class classification problem we form two groups, overall *correctly* classified samples (or positive category) and overall *incorrectly* classified samples (or negative category). We intend to measure

[1]CTO-SI, Zebra Technologies, London, United Kingdom. Correspondence to: Aditya Singh <aditya.singh@zebra.com>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

calibration and refinement within this reduced setting. A common metric often used to measure calibration in practice is the Expected calibration error (Naeini et al., 2015). It is measured as the difference between the accuracy and predicted confidences computed over several bins. Formally,

$$ECE \triangleq \sum_{m}^{M} \frac{|B_m|}{L} |A_m - C_m|, \tag{3}$$

where average confidence (C) and accuracy (A) is computed after splitting the predictions into predefined M bins sampled uniformly based on the predicted confidence and B_m is the number of total samples falling in bin m.
We consider area under the ROC curve (r) (Ling et al., 2003), as an appropriate choice of metric for measuring refinement of a model(Corbière et al., 2019). A common interpretation of r is that it denotes the expectation that a uniformly drawn random positive sample is ranked higher (higher confidence) than a uniformly drawn random negative sample. Hand & Till (2001) calculate r as:

$$r = \frac{R^p - |S^n| \times (|S^n| + 1)/2}{|S^p| \times |S^n|} \tag{4}$$

where, $R^p = \sum_{\forall x \in S^p} rank(x)$ and $rank(x)$ denotes the rank of prediction x in an increasingly sorted list of predictions based on associated confidence. It is straightforward to observe that r for a refined model will always be greater than an unrefined one (switching the rank of an incorrect prediction with the correct one decreases r).

2.1. Connecting ECE and r

Assumption: We assume that $A_m < C_m \forall m$. It implies that the network is over-confident in its prediction throughout. This is partly true in practice as for all deep neural networks the problem of calibration entails over-confident predictions(Thulasidasan et al., 2019). Also, we empirically observed that for networks trained on ImageNet (Deng et al., 2009) and CIFAR-100 (Krizhevsky, 2009), the number of bins for which $A_m <= C_m$ holds true are 80 and 95 respectively for $M = 100$.
Let, p_m and n_m represent positive and negative category samples in bin m respectively which implies $|S^p| = \sum_m p_m$ and $|S^n| = \sum_m n_m$. We can now describe the accuracy within a bin as $A_m = \frac{p_m}{p_m + n_m}$. Substituting all the above conversions to Equation (3), ECE is updated as

$$ECE = \sum_{m} \frac{(p_m + n_m)}{|S^p| + |S^n|} \left(C_m - \frac{p_m}{p_m + n_m} \right). \tag{5}$$

This can be further expanded to

$$ECE = \underbrace{\sum_{m} \frac{(p_m + n_m)}{|S^p| + |S^n|} C_m}_{I} - \underbrace{\sum_{m} \frac{p_m}{|S^p| + |S^n|}}_{II}. \tag{6}$$

I denotes the expected confidence of the predictions, $\mathbb{E}_{C \sim p_\theta(X)}[C]$, of the model, whereas II is the expected model accuracy, $\mathbb{E}[A]$. Equation (6) can thus be updated to

$$ECE = \mathbb{E}[C] - \mathbb{E}[A]. \tag{7}$$

For a binary classification task, it has been shown (Hernández-Orallo et al., 2012; Flach & Kull, 2015) that r and $\mathbb{E}[A]$ are linearly related averaged over all possible true-positive rates. They showed that:

$$\mathbb{E}[A] = \frac{P}{|S^p| + |S^n|}\left(1 - \frac{P}{|S^p| + |S^n|}\right)(2r - 1) + \frac{1}{2}, \tag{8}$$

where r is the area under the ROC curve. Substituting Equation (8) for $\mathbb{E}[A]$ in Equation (7) and re-arranging the terms gives us the final expression in the form of

$$ECE = \underbrace{\mathbb{E}[C]}_{\alpha} - r \underbrace{\frac{2PN}{(|S^p| + |S^n|)^2}}_{\beta} - \underbrace{\frac{P^2 + N^2}{2(|S^p| + |S^n|)^2}}_{\gamma}. \tag{9}$$

Traditionally, for strictly proper scoring rules such as the Brier score, the decomposition of the metric into calibration and refinement is well known. However, for ECE which is not a strict proper scoring rule, we have shown that the breakdown is into average predicted confidence and refinement under the applied assumption of bins-wide overconfidence. Moon et al. (2020) have shown that their refinement based approach improves calibration however, they do not provide the reasoning behind such an observation. Their observation can now be supported by the relationship described in Equation (9). Regularisation based calibration approaches only focus on lowering the confidence estimates (see appendix A.1).

3. Related Work

3.1. Calibration

The existing work can be categorised into the following 3 broad groups based on the commonalities between the approaches. **Regularisation** based approaches apply a calibrating penalty to the supervised learning objective. Approaches falling under this category are Entropy regularization(Pereyra et al., 2017), label smoothing(Müller et al., 2019), optimising a proxy for the calibration error metric (Kumar et al., 2018), and focal loss (Mukhoti et al., 2020). **Post-hoc** approaches rescale the confidence scores of an uncalibrated neural network to make it calibrated. Some of the recently proposed approaches are temperature scaling (Guo et al., 2017), scaling and binning calibration (Kumar et al., 2019), Dirichlet calibration (Kull et al., 2019), and beta calibration (Kull et al., 2017).In the last group, we list

	Method	VGG-16			ResNet-50			DenseNet-121		
		Brier (↓)	ECE (↓)	AUROC (↑)	Brier (↓)	ECE (↓)	AUROC (↑)	Brier (↓)	ECE (↓)	AUROC (↑)
CIFAR-10	Baseline	10.92±0.20	4.80±0.12	90.90±0.64	7.12±0.12	2.69±0.07	93.80±0.04	7.43±0.13	2.83±0.12	92.12±0.18
	TS	10.21±0.18	2.99±0.07	**90.66**±0.70	6.73±0.12	1.35±0.09	**93.71**±0.01	7.03±0.12	2.04±0.19	**91.60**±0.06
	ERL	10.58±0.08	4.29±0.01	**89.22**±0.40	7.01±0.18	2.37±0.02	**93.64**±0.49	7.19±0.01	2.18±0.04	**91.83**±0.18
	LS	10.60±0.10	3.93±0.26	**78.31**±1.36	**8.00**±0.19	**3.82**±0.17	**73.46**±1.41	**8.51**±0.04	2.80±0.08	**73.56**±1.71
	MX	9.79±0.18	3.75±1.01	**84.38**±1.99	6.35±0.15	2.67±0.27	**90.10**±1.36	6.82±0.01	**3.67**±0.36	**88.98**±0.02
	FL	10.88±0.16	3.48±0.06	**84.50**±0.50	**7.50**±0.04	1.73±0.06	**92.40**±0.60	**7.69**±0.12	1.71±0.06	**91.72**±0.15
CIFAR-100	Baseline	43.26±0.94	16.29±1.39	84.97±0.45	34.05±0.38	12.17±0.06	85.69±0.12	30.68±0.28	8.47±0.07	87.15±0.17
	TS	39.18±0.55	6.46±1.10	**84.31**±0.53	31.50±0.42	3.49±0.19	**85.07**±0.09	29.91±0.26	4.68±0.13	**86.39**±0.26
	ERL	42.09±0.72	13.96±1.43	**84.21**±0.40	32.56±0.59	9.80±0.12	**85.41**±0.19	30.50±0.33	6.10±0.11	**86.78**±0.12
	LS	38.69±0.11	8.77±0.06	**82.90**±0.12	32.07±0.20	6.21±0.13	**81.96**±0.64	**33.01**±0.33	8.22±0.02	**81.36**±0.50
	MX	38.18±0.64	7.68±1.17	**83.60**±0.51	30.27±0.53	6.02±1.57	85.78±0.24	28.93±0.03	3.55±0.09	**85.80**±0.32
	FL	39.68±0.49	9.05±0.48	**83.12**±0.26	30.61±0.33	3.98±0.48	85.91±0.24	29.90±0.03	3.06±0.16	**86.50**±0.01

Table 1. Joint evaluation for calibration and refinement. We highlight the values which are worse than the baseline in bold and red. The arrows indicate that higher (↑) and lower (↓) values are better respectively.

the remaining approaches. Mixup (Zhang et al., 2018; Thulasidasan et al., 2019) and AugMix (Hendrycks et al., 2020) utilise forms of data augmentation shown to improve the model's predictive uncertainty and calibration with added regularisation. Pre-training (Hendrycks et al., 2019a) and self-supervised learning (Hendrycks et al., 2019b) have also been highlighted to be beneficial in this regard.

For the scores utilised to assess calibration, the most commonly used are Brier score, negative log-likelihood (NLL), Expected Calibration Error (ECE) and Overconfidence Error (OE). Brier score (Brier, 1950) and NLL are strictly proper scoring rules (Gneiting & Raftery, 2007; Dawid & Musio, 2014). It has been shown that proper scoring rules decompose into calibration and refinement components (Murphy, 1973; Blattenberger & Lad, 1985). The presence of the refinement component describes the utility of the calibration approach. However, the implicit combination of the two can conceal the area of improvement. ECE and OE (Degroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005; Naeini et al., 2015) are not strictly proper scoring rules and are adapted from reliability diagrams for judging the calibration of the models.

3.2. Refinement

By refining prediction, methods seek to find a good ordinal ranking of predictions. This may or may not result in a calibrated model as it has not been studied for this problem extensively. Moon et al. (2020) incorporated 'Correctness Ranking Loss' to allow a DNN to learn appropriate ordinal rankings for classified samples. They also observed that their approach 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 a better ordinal ranking of predictions than the output of the network. ConfidNet (Corbière et al., 2019) incorporates the learning of this trust score as an additional branch in the network.

To evaluate refinement, the commonly used metrics are AUROC, AUPR, and FPR at 95%-TPR. AUROC can be interpreted as the probability of algorithm ranking a random positive sample higher than a random negative sample. Hence, if the AUROC is high, we observe fewer correctly classified samples ranked lower than incorrectly classified samples. AUPR and FPR at 95%-TPR provide similar information w.r.t refinement of a model.

4. Experiments

4.1. Implementation

To empirically verify our findings we employ the following calibration approaches in our study. Temperature Scaling (**TS**)Guo et al. (2017), Entropy Regularization (**ERL**): Pereyra et al. (2017), Label Smoothing (**LS**): Müller et al. (2019), Mixup (**MX**): Thulasidasan et al. (2019) and Focal Loss (**FL**): (Mukhoti et al., 2020). We compare these approaches to a cross-entropy trained model referred to as **baseline**. For the datasets we rely on CIFAR-10/100 (Krizhevsky, 2009) which has been used extensively in recent calibration studies. The neural network architectures chosen are Resnet-50 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2015) and DenseNet-121 (Huang et al., 2017). We report ECE and Brier score as calibration errors whereas, AUROC for refinement. All values provided are $\times 100$. We report mean and std. deviation over 3 trials where applicable. We report the accuracies in the supplementary document as we found them to be highly similar across different methods.

4.2. Calibration & Refinement

Unsurprisingly, the results in table 1 demonstrates that calibration approaches attain lower calibration errors. At times, it is also evident that the approaches only improve either the Brier score or ECE. Also, another important observation to note is the poor refinement performance than the baseline across the board. Out of all the approaches assessed, LS consistently acquires the lowest refinement performance. TS, scales the output confidence values based on the chosen temperature. Since, it does not re-order the predictions it

Method	CIFAR-10.1			CIFAR10.2		
	Brier (↓)	ECE (↓)	AUROC (↑)	Brier (↓)	ECE (↓)	AUROC (↑)
Baseline	24.50 ± 1.23	11.28 ± 0.71	86.23 ± 0.91	34.25 ± 0.73	16.32 ± 0.60	83.39 ± 1.72
TS	22.68 ± 1.18	8.16 ± 1.18	**86.10** ± 0.90	31.83 ± 0.73	13.07 ± 0.66	**83.06** ± 1.91
ERL	23.70 ± 0.58	10.16 ± 0.27	**84.32** ± 1.19	34.23 ± 1.26	15.60 ± 0.59	**81.85** ± 0.19
LS	23.60 ± 0.07	6.84 ± 0.09	**77.44** ± 0.91	32.37 ± 0.08	11.26 ± 0.19	**74.05** ± 2.85
MX	20.89 ± 0.71	6.87 ± 0.60	**84.76** ± 1.24	29.35 ± 0.75	9.77 ± 1.79	**80.25** ± 1.67
FL	**25.13** ± 0.69	10.51 ± 0.29	**80.55** ± 0.89	33.05 ± 0.36	14.61 ± 0.29	**77.74** ± 1.18

Table 2. Joint evaluation for calibration and refinement under natural data shift.

has the same final accuracy as that of the baseline. However, the scaling operation does not guarantee a better or worse ranking of predictions. Empirically it suggests that temperature scaling retains much of the baseline refinement as well. ERL provides the least improvement in terms of calibration and achieves slightly worse AUROC w.r.t the baseline. MX and FL provide moderate to low decay of refinement. Lastly, we also note the mismatch in improvements reported by ECE and Brier score. This variation is more frequent for LS. As the Brier score reports a single value encompassing calibration and refinement, depending on the degree of change for one attribute, it can drive the final narrative.

4.3. Refinement Under Natural Shift

In this experiment, we aim to assess the deterioration under natural distribution shift of the datasets. Natural shift implies a subtle change in scene composition, object types, lighting conditions, and many others (Taori et al., 2020). It is logical to assume that a DNN is bound to confront such images in the real world. Examples of naturally shifted datasets are CIFAR-10.1 (Recht et al., 2018) and CIFAR-10.2 (Lu et al., 2020). These datasets are collected following the identical process to that of the original reference dataset. Such datasets have been utilised to measure the lack of generalisation and robustness of many classification models (Taori et al., 2020; Ovadia et al., 2019). This is the first attempt at evaluating calibration-refinement under natural data-shift to the best of our knowledge.

4.3.1. Results

Table 2 shows the performance of models trained on original datasets and tested on shifted variants. We spot that the trend of worsening refinement continues for models under data shift as well. Similar to what we have already seen for LS, it also provides the lowest refinement performance under natural shift. A surprising observation to note is the poor performance of MX. MX as shown by Thulasidasan et al. (2019) performs well on out-of-distribution detection. However, when the data shift is not severe it appears that mixup provides no added benefit in terms of refinement. We also observe that calibration approaches provide better calibration than the baseline under the natural shift. This observation has not yet been highlighted in existing studies which focus on ood performance or some form of generalisation metric (relative accuracy) to investigate robustness of a model. For synthetic shifts, Ovadia et al. (2019) made a similar observation and noted that that calibration approaches to a certain extent improve calibration on corrupted images w.r.t the baseline.

5. Discussion & Conclusion

In this paper we have brought forth a downside of many calibration approaches. We believe refinement is an important aspect which communicates the usefulness of safety-critical DNNs. Discussed theoretically and empirically, we shed light on the current situation of calibration-refinement trade-off.

The derived relationship in equation 9 showed how improving refinement can help better calibrate the model. This provides justification for calibration observed for refinement approach of Moon et al. (2020). In the appendix (A.2), we show that calibration is induced by the refinement technique proposed by Corbière et al. (2019).

Another outcome of our analysis was that adding regularisation without realising its impact is not an ideal direction that we should be progressing. Many calibration approaches with the likes of label smoothing and mixup do this and as a result incorporate calibration at the cost of usefulness and practical utility of the models. In the future, we aim to focus on finding balanced calibration methods which preserve if not improve refinement of predictions.

We also noted the extension of calibration to naturally shifted data. Akin to the observations made by (Ovadia et al., 2019) on their evaluation on synthetically shifted datasets, we observed that existing solutions provide calibration on naturally shifted datasets as well. However, this calibration comes at a cost and as a result refinement aspect of the models is comparably poorer than their uncalibrated counterparts.

Apart from relying on ECE and Brier score, incorporating metrics like AUROC, AUPR etc. helps in further distinguishing useful calibration approaches. Additionally, many evaluation protocols have been proposed recently which extend the problem of calibration to a multi-class setting (Widmann et al., 2019). A natural extension will be to study refinement conjointly with calibration in a similar manner.

References

Blattenberger, G. and Lad, F. Separating the brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 1985.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 1950.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. In *NeurIPS*. 2019.

Dawid, A. P. and Musio, M. Theory and applications of proper scoring rules. 2014.

Degroot, M. and Fienberg, S. Assessing probability assessors: Calibration and refinement. 1981.

Degroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 1983.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Flach, P. A. and Kull, M. Precision-recall-gain curves: Pr analysis done right. In *NeurIPS*, 2015.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, 2017.

Hand, D. and Till, R. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 2001. doi: 10.1023/A:1010920819831.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *ICML*, 2019a.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019b.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. AugMix: A simple data processing method to improve robustness and uncertainty. *ICLR*, 2020.

Hernández-Orallo, J., Flach, P., and Ferri, C. A unified view of performance metrics: Translating threshold choice into expected classification loss. In *JMLR*, 2012.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017.

Jiang, H., Kim, B., Guan, M. Y., and Gupta, M. To trust or not to trust a classifier. In *NeurIPS*, 2018.

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

Kull, M., Silva Filho, T., and Flach, P. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 2017.

Kull, M., Nieto, M. P., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, 2019.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *ICML*, 2018.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019.

Ling, C., Huang, J., and Zhang, H. Auc: a statistically consistent and more discriminating measure than accuracy. *IJCAI*, 2003.

Lu, S., Nott, B., Olson, A., Todeschini, A., Vahabi, H., Carmon, Y., and Schmidt, L. Harder or different?a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

Moon, J., Kim, J., Shin, Y., and Hwang, S. Confidence-aware learning for deep neural networks. In *ICML*, 2020.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H., and Dokania, P. K. Calibrating deep neural networks using focal loss. 2020.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *NeurIPS*, 2019.

Murphy, A. H. A new vector partition of the probability score. *Journal of Applied Meteorology (1962-1982)*, 1973.

Murphy, A. H. and Winkler, R. L. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1977.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 2015.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *ICML*, 2005.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., and Hinton, G. E. Regularizing neural networks by penalizing confident output distributions. In *ICLR, Workshop Track Proceedings*, 2017.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? 2018. https://arxiv.org/abs/1806.00451.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2007.00644.

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., Macaskill, P., Moons, K. G. M., Van Calster, B., van Smeden, M., and Vickers, A. J. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 2019.

Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests in multi-class classification: A unifying framework. In *NeurIPS*, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

A. Appendix

A.1. Calibration with Label Smoothing

Label smoothing calibration can be written as

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{LS}, \tag{10}$$

where CE stands for cross-entropy and LS represents label smoothing contribution. Label smoothing contribution is the KL divergence between uniform distribution (U) and network's output distribution (P_θ). Formally,

$$\mathcal{L}_{LS} = -\mathrm{D}_{KL}(U||P_\theta). \tag{11}$$

LS is closely associated with ERL(Pereyra et al., 2017) and focal loss(Mukhoti et al., 2020). $\mathcal{L}_{LS}$ is defined as,

$$\mathcal{L}_{LS} = \sum_{i=0}^{i<N} -\underbrace{U(x_i)log(P_\theta(x_i))}_{I} + \underbrace{U(x_i)log(U(x_i))}_{II}, \tag{12}$$

where x_i is a sample input from a total of N sample points. The value for the uniform distribution is set before hand to a small constant ϵ thus making II a constant term. I is the term which is optimised and for a binary classification problem can be written as

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} \epsilon\ logc_i + \epsilon\ log(1-c_i) \\ \text{s.t.} \quad & 0 \leq c_i \leq 1, \end{aligned} \tag{13}$$

The above expression reaches a minimum value when $c_i = 0.5$. This shows that adding entropy regularization will encourage in confidence estimates to be close to 50% where the entropy is maximum. For multi-class classification, the minimum is achieved at $\frac{1}{K}$. Similar observation of lowering confidence holds if we expand equations for label smoothing. We discuss this in more detail in the supplemental (see Section A.1).

A.2. Calibration by Refinement

In this section we present the results of the refinement approach of Corbière et al. (2019). ConfidNet (CFN) learns as a post-processing step a point-estimate for new predictions. The pre-trained classification branch drives the classification of an input sample, and for estimating the confidence for the prediction, the estimate from the confidence branch is employed.

The authors highlight the refinement advantage over baseline and TrustScore (Jiang et al., 2018) by employing AUPR, AUROC, etc. We utilize the official source code and train VGG-16 (Simonyan & Zisserman, 2015) with batch normalization. We retain 10% of training data to validate CFN

training parameters and report the calibration and refinement results on the official test split for CIFARs (Krizhevsky, 2009). The results are reported over 3 independent runs of the experiment.

A.2.1. RESULT

	CIFAR-100		CIFAR-10	
	ECE(↓)	AUROC(↑)	ECE(↓)	AUROC(↑)
Baseline	19.12 ± 0.13	85.18 ± 0.21	5.38 ± 0.15	92.5 ± 0.01
CFN	$\mathbf{13.95 \pm 2.7}$	$\mathbf{86.0 \pm 0.18}$	$\mathbf{4.1 \pm 0.2}$	$\mathbf{92.55 \pm 0.1}$

Table 3. Calibration and refinement results aggregated over 3 runs. Values in bold font indicates the best value w.r.t the corresponding metric.

Results in Table 3 show the CFN performance in comparison to an uncalibrated and unrefined baseline. Not only does CFN provide better refinement, it is also able to reduce the calibration errors over the datasets. This provides further support to our understanding of calibrating a model by improving refinement.