Revisiting the Evaluation of Uncertainty Estimation and Its Application to Explore Model Complexity-Uncertainty Trade-Off

Yukun Ding[1], Jinglan Liu[1], Jinjun Xiong[2], Yiyu Shi[1]
[1] University of Notre Dame
[2] IBM Thomas J. Watson Research Center
{yding5, jliu16, yshi4}@nd.edu, jinjun@us.ibm.com

Abstract

Accurately estimating uncertainties in neural network predictions is of great importance in building trusted DNNs-based models, and there is an increasing interest in providing accurate uncertainty estimation on many tasks, such as security cameras and autonomous driving vehicles. In this paper, we focus on the two main use cases of uncertainty estimation, i.e.*, selective prediction and confidence calibration. We first reveal potential issues of commonly used quality metrics for uncertainty estimation in both use cases, and propose our new metrics to mitigate them. We then apply these new metrics to explore the trade-off between model complexity and uncertainty estimation quality, a critically missing work in the literature. Our empirical experiment results validate the superiority of the proposed metrics, and some interesting trends about the complexity-uncertainty trade-off are observed.*

1. Introduction

Deep neural networks (DNNs) have been widely used in vision tasks and achieved remarkable performance improvement. A major challenge in adopting DNNs to real-world mission-critical applications such as the medical image segmentation, is the lack of self-awareness and the tendency to fail silently [16]. In contrast, human's awareness of prediction uncertainty enables, for example, human radiologists to conduct further investigations whenever they are in doubt for a diagnosis based on computed tomography (CT) images, and human drivers to slow down whenever they cannot clearly recognize an object. In order for DNNs to gain human's trust in making critical decisions, especially in mission-critical scenarios, we need to equip DNNs with self-awareness on a par with its task competency. Most recently, much effort has been devoted to providing an accurate quantified score representing the uncertainty of every prediction, where wrongly predicted instances are expected to be assigned with low confidence scores and correctly predicted ones are expected to be assigned with high confidence scores [1] [5, 13, 29, 37, 40].

The competency awareness of DNNs is commonly realized in two use cases of uncertainty estimation: *selective prediction* [9, 24, 30, 33] and *confidence calibration* [13, 15, 22, 32, 38, 39]. For the selective prediction, the obtained confidence scores are thresholded and the model can abstain from making predictions on samples with low confidence scores to achieve higher accuracy on the remaining part [14]. For instance, in automatic segmentation of medical images, it is desired that the machine segments the common and easy area of medical images and refers the area with unusual appearance to the radiologists to ensure an extremely high accuracy [38]. In this case, the confidence score is expected to be used for separating correct predictions and wrong predictions, and the popular quality metrics used to evaluate uncertainty estimation are *Area Under Receiver Operating Characteristic curve* (AUROC) and *Area Under Precision-Recall curve* (AUPR) [3, 14, 29]. For the confidence calibration, the aim is to provide a confidence score that approximates the empirical probability of a prediction being correct [13, 38]. For instance, in autonomous driving, human intervention is often not available in a timely manner and the high-level planning module will need such calibrated confidence score of pedestrian detection for instant decision making. In this case, common quality metrics are *Expected Calibration Error* (ECE) and *Maximum Calibration Error* (MCE) [13, 22, 38].

Despite the recent advancements, we show that the quality metrics of neural network uncertainty estimation used by most existing works could be problematic, potentially leading to unfair comparisons, confusing results, and/or undesired learning behaviors. Specifically, for the selective prediction, we show that even minimal changes in prediction models can make the commonly used evaluation based on

[1] Confidence is the additive inverse of uncertainty with respect to 1, so they are used interchangeably in the literature.

AUROC and AUPR meaningless or even misleading. To address this issue, we propose to use a different metric, called *Area Under Risk-Coverage* (AURC) curve as the primary metric for selective prediction. We show that AURC is the only reliable metric among AUROC, AUPR, and AURC, when the underlying prediction model changes, and is consistent with AUROC and AUPR when the underlying prediction model stays the same. As for confidence calibration, we show that, because of the basic binning strategy employed, the commonly used evaluation metrics ECE and MCE cannot expose some large calibration error even in the high confidence area. Moreover, they are vulnerable to internal compensation and inaccurate accuracy estimation in each confidence interval, which leads to poor robustness and inferior accuracy. Therefore, we propose a new binning strategy, called *adaptive binning*, for the evaluation by ECE and MCE, and empirically show its superiority.

While the complexity-accuracy trade-off of DNN-based models has been extensively studied in the literature, the effect of model complexity on uncertainty estimation quality is almost unknown. However, with the prevalence of DNNs in real-world applications, it is ever more important for model designers to seek the best trade-off between cost and different aspects of model performance under various resources constraints. Therefore, a better understanding of the uncertainty-related performance changes with model complexity is required. We first give some theoretical analysis of the relation between the selective prediction and the confidence calibration. Then we use our proposed new metrics to explore the effect of model complexity on the uncertainty-related model performance. Our study serves two purposes. First, it validates the effectiveness and robustness of our proposed evaluation metrics. Second, it provides the first empirical study on how uncertainty-related model performance changes with model complexity. From our study, we observe that, interestingly, estimation quality changes significantly with model complexity for selective prediction, but is insensitive to model complexity for confidence calibration.

In summary, the main contributions of this paper are as follows:

- We identify the potential issues of commonly used quality metrics for uncertainty estimation in both selective prediction and confidence calibration, and propose new metrics that provide more reliable and informative evaluations.

- As an application and validation of the proposed metrics, we provide the first exploration of complexity-uncertainty trade-off, and show some interesting observations.

2. Related Works

Uncertainty Estimation. Various methods exist in the literature to estimate the uncertainty of neural network predictions [6, 11, 15, 19, 27]. The most popular approaches include softmax probability [14, 13], Monte Carlo dropout [8, 36], and learned confidence estimation [5, 28]. The uncertainty estimation can either explicitly affect the model during the training process [7, 22, 26] or work as a post-processing step that does not affect the underlying prediction models [3, 13]. Note that in our definition and analysis, we did not make any assumption on how the confidence score is obtained or any correlation between the prediction and the confidence score. They can be obtained by any prediction model and any uncertainty estimation method.

Evaluation Methods. The commonly used evaluation metrics for selective prediction are AUROC and AUPR [3, 14, 29]. Recently, E-AURC has been used to evaluate the uncertainty estimation quality in a selective prediction scenario [11]. However, the E-AURC has exactly the same problem with AUROC and AUPR, because the accuracy difference is not considered, which is detailed in Section 4.1. It is meaningless when the underlying prediction model varies, which happens in many cases, *e.g.* comparing different models, evaluating algorithms that alter the training process [30], and when either MC-dropout or ensemble is used for uncertainty estimation [24, 33].

Currently, the confidence calibration quality of neural network-based models are evaluated by ECE and MCE [13, 22, 32, 38, 39]. In order to minimize the ECE, a differentiable proxy to ECE named MMCE is used for calibration-aware network training [22]. Negative Log Likelihood (NLL) and Brier Score are used as indirect and supplementary measurements in some works [13, 39]. We note that Brier Score and NLL are not suitable as primary metrics for confidence calibration, because they prefer better prediction rather than better calibration by design. ECE and MCE have been extended from the binary setting to the multi-class setting in [34, 42]. In either way, the computation of ECE and MCE heavily depends on the binning strategy which is the focus of this work. Equal-size binning where every bin has a same number of samples was proposed as a remedy for the known issues of the common fixed equal-size binning [34, 42]. However, we show that although equal-size binning helps, it is still not flexible enough to deal with highly non-uniform confidence distribution. [32] uses a Bayesian score to average a number of models with equal-size binning. Such modeling averaging is orthogonal to our binning method. In addition, it is used as a calibration method to improve the performance measured by ECE that uses the conventional equal-range binning.

3. Problem Setting

We put our discussion in a general classification setting. See [21] for recent advance in the regression setting. Following [22], we denote $\mathcal{Y} = \{1, 2, \dots, K\}$ as the set of class labels, $\mathcal{X}$ as the input space, $\mathcal{D}$ as the data distribution, and $N_\theta(y|x)$ as the probability distribution of model predictions with input x, and model parameters θ. For each input sample x_i and true label y_i, the model gets a predicted label $\hat{y}_i = \text{argmax}_{y \in \mathcal{Y}} N_\theta(y|x_i)$ and a confidence score r_i. If $\hat{y}_i = y_i$, which means the prediction is correct, we have the correctness score $c_i = 1$. Otherwise, $c_i = 0$. Then the distribution over r and c on $\mathcal{D}$ can be denoted as $P_{\theta,\mathcal{D}}(r, c)$.

Selective Prediction. In selective prediction, with a confidence score r_i for each input x_i and a threshold t, the input from dataset X and the prediction $\hat{Y}$ are split to $X_h = \{x_i | r_i >= t\}$, $X_l = \{x_i | r_i < t\}$ and $\hat{Y}_h = \{\hat{y}_i | r_i >= t\}$, $\hat{Y}_l = \{\hat{y}_i | r_i < t\}$ respectively. The model abstains from making prediction on X_l. Ideally, $\hat{Y}_l$ contains all wrong predictions and $\hat{Y}_h$ contains all correct predictions so that the error is avoided with the minimal cost. In this case, r_i is used for separating correct predictions and wrong predictions, which is a binary classification problem and therefore the common quality metrics are AUROC and AUPR [3, 14, 29, 30].

Confidence Calibration. Confidence calibration aims to give a confidence score $r \in [0, 1]$ that directly reflects the probability of the prediction being correct. The difference between the probability of correct prediction $E_{P_{\theta,\mathcal{D}}(c|r)}[c]$ and the confidence score r is defined as the calibration error. Consequently, the expected calibration error (ECE) and maximum calibration error (MCE) are defined as:

$$\text{ECE}(P_{\theta,\mathcal{D}}) = E_{P_{\theta,\mathcal{D}}(r)}[|E_{P_{\theta,\mathcal{D}}(c|r)}[c] - r|] \tag{1}$$

$$\text{MCE}(P_{\theta,\mathcal{D}}) = \max_{r \in [0,1]} |E_{P_{\theta,\mathcal{D}}(c|r)}[c] - r| \tag{2}$$

Practically, given a finite number of samples in $D \sim P_{\theta,\mathcal{D}}$, ECE and MCE are calculated by partitioning the $[0, 1]$ range to n bins according to a binning strategy. For every bin, an average accuracy and an average confidence are calculated using all samples inside. The difference between the average accuracy and the average confidence is the calibration error, which is denoted as calibration gap in [13]. The standard practice is to use n equal-range bins where n is chosen as 10 in the literature [13, 15, 22, 32, 38, 39]. Specifically, the partition is defined as $B_j = [\frac{j-1}{n}, \frac{j}{n}]$, $j = \{1, \dots, n\}$. It is possible to use different binning strategy such that these bins are not uniformly distributed and the definition of B_j will change accordingly. Given $D_j = \{x_i | r_i \in B_j\}$, ECE and MCE are computed as $\hat{\text{ECE}}$ and $\hat{\text{MCE}}$ by:

$$\hat{\text{ECE}}(P_{\theta,\mathcal{D}}) = \frac{1}{|D|} \sum_{j=1}^{n} |\sum_{x_i \in D_j} c_i - \sum_{x_i \in D_j} r_i| \tag{3}$$

$$\hat{\text{MCE}}(P_{\theta,\mathcal{D}}) = \max \frac{1}{|D_j|} |\sum_{x_i \in D_j} c_i - \sum_{x_i \in D_j} r_i| \tag{4}$$

It is worth mentioning that both ECE and MCE are proper scoring rules but not strictly proper scoring rules [12]. However, even strictly proper scoring rules do not guarantee a reliable evaluation and comparison [31]. Their potential issues are discussed in Section 4.

In addition to the quantified metric, Reliability Diagram [13, 22, 38, 39] is used as a standard qualitative analysis tool in the literature. It plots the empirical accuracy in each bin and the calibration error. Such a diagram not only visualizes the calibration error at different confidence intervals but also shows how the ECE and the MCE are calculated.

4. Evaluation Metrics: Issues & Solutions

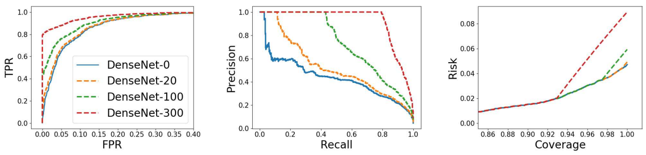
In this section, we discuss some issues of the existing quality metrics for uncertainty estimation that may lead to unfair comparison or neglected problems, and then present new metrics to mitigate them. In all figure captions and tables, ↑ means the higher the better, and ↓ means the lower the better.

4.1. Selective Prediction

We remark that comparing AUROC and AUPR is fair only when the underlying prediction models are the same. There was no proper treatment used in the literature when comparing uncertainty estimation methods with different prediction models as shown in recent works [11, 29, 30]. Below we use an example to show that even a small difference in the underlying prediction model could make the comparison of AUROC and AUPR meaningless if not misleading. Then we discuss the advantage of the proposed AURC over AUROC and AUPR.

In order to know the relative performance of models as a priori, we build an illustrative example based on a real-world network. We train a 100-layer DenseNet on Cifar10 with standard settings and denote it as DenseNet. For a given test dataset X, we further define X_c and X_p as $X_c = \{x_i | c_i = 1\}$ and $X_p = \{x_i | r_i > t_m, x_i \in X_c\}$ where t_m is a threshold such that $m = |X_p|$. Consider there is a network named DenseNet-m that makes the same predictions with DenseNet for all samples in X except for that in X_p and DenseNet-m has $c_i = 0$ for all $x_i \in X_p$. In other words, DenseNet-m is the same with DenseNet except that DenseNet-m makes m more wrong predictions in the most certain predictions out of the total 10^4 samples. By varying the value of m, we get different DenseNet-m. Note that DenseNet is equivalent to DenseNet-0. With this setup, DenseNet-m with smaller m has equal or higher accuracy than that with bigger m at any threshold t in selective prediction. Therefore, it is convincing to conclude that bigger m indicates worse selective prediction quality. A proper evaluation metric is expected to correctly reflect the relative performance of different variants of DenseNet-m.

We plot ROC curves and PR curves of DenseNet-m with different values of m in Figure 1. Remember that both ROC curves and PR curves are the higher the better. Both curves suggest a questionable result that DenseNet-m with a bigger m is better. The reason is that both AUROC and AUPR only measure a model's ability to distinguish correct and wrong predictions while assuming the numbers of correct and wrong predictions are the same. An accuracy change of 0.2% (comparing DenseNet-0 and DenseNet-20) could give a significant improvement on AUPR while the actual performance is getting worse. Therefore, when the underlying prediction models are different, AUROC and AUPR fail to correctly reflect the model's actual performance change.



(a) ROC curve↑ (b) PR curve↑ (c) RC curve ↓

Figure 1: Evaluation curves of DenseNet-m. We term the misclassified/classified samples as positive/negative samples following the literature.

The observation is also confirmed in quantitative results. In Table 1, the quantitative results of both AUROC and AUPR suggest that DenseNet-m with bigger m performs better which contradicts the prior knowledge.

Table 1: Quantitative comparison of AUROC, AUPR and AURC. Similar to the results in the literature, AUROC's change is relatively small because of the class imbalance.

Model	Acc.↑	AUROC↑	AUPR↑	AURC↓
DenseNet-0	95.30	93.71	43.36	0.438
DenseNet-20	95.10	94.17	52.36	0.439
DenseNet-100	94.30	95.80	72.80	0.456
DenseNet-300	92.30	98.08	91.64	0.605

In summary, evaluating models with AUPR and AUROC not only fails to provide a fair comparison, but also implicitly encourages the bad practice of reducing model accuracy in designing uncertainty estimation methods.

To mitigate this issue, we propose to do the evaluation with the Risk-Coverage (RC) curve instead. The coverage denotes the percentage of the input processed by the model without human intervention and the risk denotes the level of risk of these model prediction. Formally,

$$coverage = \frac{|X_h|}{|X|} \tag{5}$$

$$risk = \mathcal{L}(\hat{Y}_h) \tag{6}$$

where $\mathcal{L}$ is a loss function measuring the prediction quality. For classification, the 0/1 loss is commonly used [9] as it measures the classification accuracy.

The risk-coverage curve reflects the nature of the selective prediction very well by definition as the motivation of the selective prediction is to reduce the coverage of the model in order to achieve higher accuracy.

The RC curves of DenseNet-m are shown in Figure 1c. Quantitative comparison of AUROC, AUPR, and AURC are shown in Table 1. Note that although the RC curve has been used in the literature to demonstrate selective classification [9], using AURC as the evaluation metric of uncertainty estimation for selective prediction is first proposed in this work. In both qualitative and quantitative results, only AURC gives the correct performance ranking which is in line with the prior knowledge. The reason is that AURC by definition naturally evaluates the combined results of prediction and uncertainty estimation without the assumption that the prediction models are the same. Without AURC, even if the accuracy, AUROC, and AUPR are reported together, it is still unknown that which model in Table 1 performs the best in selective prediction. Therefore, AURC provides the only reliable evaluation when the underlying prediction models are different.

We further show that AURC is still a good metric when the accuracy of underlying prediction models are the same, because it correctly recognizes the better model just like AUROC and AUPR. The well-known connection and consistency between ROC curve and PR curve are established in [4] by proving that the curve of one model dominates the curve of another model in ROC space, if and only if it also dominates the other in the PR space. In this paper, we show that the RC curve shares the same inherent connection with ROC curve and PR curve by giving Theorem 1. The proof is given in the supplementary.

Theorem 1. *For any two models A and B of the same accuracy and their uncertainties measured by arbitrary methods (which can be different for A and B), the curve of A dominates that of B in the ROC space, if and only if the curve of A dominates that of B in the Risk-Coverage space.*

Another issue with the evaluation practice we identified is the poor generality out of classification tasks. For example, the image segmentation quality cannot be properly evaluated by pixel-wise accuracy. Consequently, even if the underlying prediction models are the same, AUROC and AUPR still fail to accurately reflect the performance of selective prediction. In contrast, the AURC can be easily extended to this case by using a suitable $\mathcal{L}$ for domain-specific performance measure. For example, when measuring the image segmentation quality, the risk can be defined as $1 - Dice$ where $Dice$ is a commonly used quality metric for image segmentation.

In selective prediction, each low confidence prediction leads to a special process such as processing by a bigger model, examining by human experts, or making conservative decisions by the controller. Such operation is usually "expensive" and thus the coverage directly determines the overall operation cost that is an important metric in a selective prediction application. However, this metric is not available from the conventional ROC curve or PR curve. In contrast, RC curve merges two accuracy axes to one and adds the cost axis to show the cost-performance trade-off which arguably makes it easier for human administrators to choose an operating point.

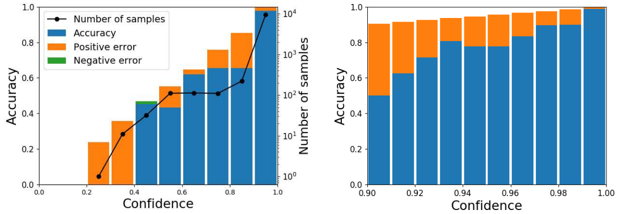
In summary, using AURC as a primary evaluation metric has the following advantage. (i) when the underlying prediction models are the same, AURC is an effective quality metric to indicate the performance of selective prediction; (ii) when the prediction models are different which happens a lot in the literature due to the emerging trend of uncertainty-aware training [30, 11, 29, 22], using AURC instead of AUPR and AUROC prevents unfair and potential misleading comparison. (iii) AURC can be generalized to distinct tasks with task-specific evaluation metrics while AUPR and AUROC cannot. (iv) AURC is an alternative optimization objective to directly maximize the performance of selective prediction and helps to avoid weighing multiple objective terms in related work [30, 29, 22]. (v) AURC directly shows the cost-performance trade-off in selective prediction which is not visible in the conventional ROC curve or PR curve.

4.2. Confidence Calibration

The accuracy and reliability of Reliability Diagrams, ECE, and MCE highly depend on the underlying binning strategy, whose limitations are explained as below.

Undetectable Error. We use the same original DenseNet used above as a real-world example to show the problem of the commonly used ECE, MCE, and Reliability Diagrams. The maximum softmax probability [14] is used as the confidence score. As shown in Figure 2a, the calibration error on $[0.9, 1]$ is as small as 0.0215 which means the average error between the confidence and accuracy is 2.15%. One would expect that for confidence in this interval, the accuracy is very close to the confidence. However, as shown in Figure 2b, for input samples with confidence in $[0.9, 0.91]$, the accuracy is only 50%. For input samples with confidence in $[0.9, 0.96]$, the accuracy is lower than 73%. This problem can mainly be attributed to the large bin range and the highly non-uniform distribution of confidence as shown in Figure 2a.

Most samples have a confidence score in $[0.98, 1.0]$ and the calibration error on that range is small. Then an average view makes the high calibration error on $[0.9, 0.96]$ undetectable by normal Reliability Diagrams, ECE, and MCE. Note that the big calibration error on $[0.9, 0.96]$ by no means should be tolerated, because it has a higher sample density than all nine bins on its left and ignoring it may jeopardize mission-critical systems.



(a) Normal setting with 10 bins

(b) Diagrams on $[0.9, 1]$

Figure 2: Undetectable error in Reliability Diagrams. In all Reliability Diagrams, positive error means confidence is larger than accuracy.

Internal Compensation. Even if the confidence distribution is relatively uniform, we remark that "internal compensation" can happen inside a bin and the ECE obtained is overly optimistic. As can be seen from Figure 2a, the error is not always positive or negative. In fact, the different sign of the error also exists inside a bin. This makes the computed ECE lower than a more accurate one computed based on a higher resolution. We conclude this effect in Proposition 1 and the complete proof is provided in the supplementary.

Proposition 1. *For any bin selection,* $\hat{ECE}(P_{\theta,\mathcal{D}}) = ECE(P_{\theta,\mathcal{D}})$ *if and only if for any bin* B_j*,* $E_{P_{\theta,\mathcal{D}}(c|r_k)}[c] \geq r_k$ *for all* $r_k \in B_j$ *or* $E_{P_{\theta,\mathcal{D}}(c|r_k)}[c] \leq r_k$ *for all* $r_k \in B_j$*. Otherwise,* $\hat{ECE}(P_{\theta,\mathcal{D}}) < ECE(P_{\theta,\mathcal{D}})$*.*

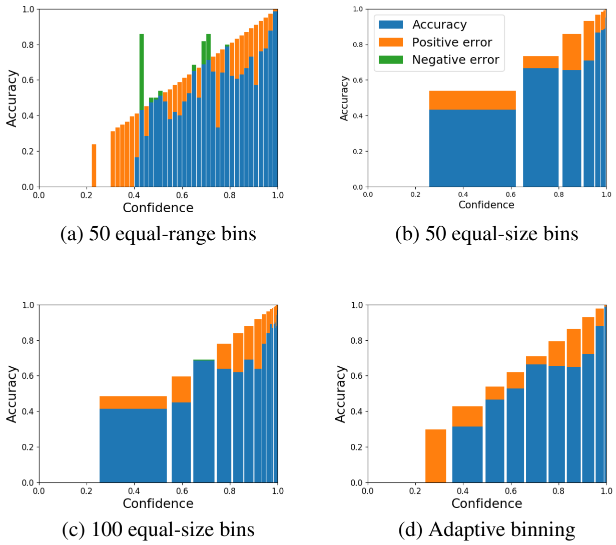
With the assumption that $E_{P_{\theta,\mathcal{D}}(c|r_k)}[c]$ is available, the range of bins should be as small as possible to recover the actual ECE. However, $E_{P_{\theta,\mathcal{D}}(c|r_k)}[c]$ is only available with enough samples which leads to the problem of inaccurate accuracy estimation.

Inaccurate Accuracy Estimation. A relatively straightforward solution to the aforementioned problems is to increase the number of bins to get higher resolution on the confidence scores. However, using more bins does not always get better results. Even though $\frac{1}{|D_j|}\sum_{x_i \in D_j} c_i$ is an unbiased and consistent estimator of $E_{P_{\theta,\mathcal{D}}(c|r)}[c]$ for

$x_i \in D_j$, with more bins $|D_j|$ becomes smaller and the inaccurate approximation of $E_{P_{\theta,\mathcal{D}}(c|r)}[c]$ leads to inaccurate ECE. In fact, another loophole of the Reliability Diagrams and MCE is that, $\frac{1}{|D_j|}\sum_{x_i \in D_j} c_i$ may not provide accurate estimation for $E_{P_{\theta,\mathcal{D}}(c|r)}[c]$ when $|D_j|$ is small. A real-world example is shown in Figure 3a. In order to make the low accuracy around 0.95 visible, the number of bins has to be increased from 10 to 50. The significant fluctuation is a result of the inaccurate accuracy when the number of samples in the bin is small.

We remark that although the inaccurate accuracy estimation can be solved if there are excessive samples, it cannot fix the undetectable error and internal compensation, because the underlying reason is the highly nonuniform confidence distribution instead of limited samples. A seemingly feasible remedy is to use equal-size binning [34, 42], where each bin has the same number of samples, instead of equal-range binning. However, when the bin lies in a confidence region where samples are sparse, the resulting confidence range may be too large and less informative. On the other hand, when it lies in a region where samples are dense, the range of bins becomes too small and accuracy estimation is suboptimal. This can be seen from Figure 3b and Figure 3c, when 50 and 100 bins are used respectively. In both Figure 3b and Figure 3c, it can be seen that a very wide bin exists in the low confidence region while an excessive number of bins reside in high confidence region.

To tackle this challenge, we resort to an adaptive binning strategy, where the number of samples in a bin is adaptive to the distribution of the samples in the confidence range. We achieve a dynamic balance between the resolution and the accuracy estimation by associating the number of samples in each bin with the range of the bin. In this way, more samples can be included for better accuracy estimation when the samples are dense and fewer samples will be used to avoid too large range when the samples are sparse. Specifically, we use $n = 0.25\left(\frac{Z_{\alpha/2}}{\epsilon}\right)^2$ to estimate the number of samples needed to estimate the accuracy for each bin where ϵ is the error margin, $Z_{\alpha/2}$ is the Z-score of a standard normal distribution and $1-\alpha$ is the confidence interval. Even though there are still two hyper-parameters, we find that the result is not sensitive to these parameters in a wide range due to its high robustness. We use an 80% confidence interval and let ϵ equal to the width of the confidence range of the bin in all experiments. We denote the resulting new adaptive metrics as AECE and AMCE in the rest of this paper. To make it easy for other researchers to use this adaptive binning, we provide the details and our implementation as an open-source tool at https://github.com/yding5/AdaptiveBinning. The computation overhead is minimal and the complexity is still $\mathcal{O}(|D|)$, so there is no scalability issue. The result on the same network is shown in Figure 3d, which achieves the best results in terms of capturing the calibration error and alleviating all the aforementioned problems. More examples of adaptive binning are shown in the supplementary. We hope researchers can consider to use this new quality metric in confidence calibration in the future.



(a) 50 equal-range bins

(b) 50 equal-size bins

(c) 100 equal-size bins

(d) Adaptive binning

Figure 3: Reliability Diagrams of various binning methods.

Theoretically, the issues discussed above can also happen in other probabilistic forecast problems, *e.g.* weather forecast [23] where similar evaluation metrics such as ECE and Brier Score are used [18, 44]. Furthermore, consistency bar is used to partially solve the inaccurate accuracy estimation issue by indicating the fluctuations of the observed frequencies (accuracy in our context) caused by the limited samples in each bin with the *consistency resampling* technique [1]. However, consistency bar can only indicate a reasonable fluctuation range of the frequencies that a reliable calibration would likely fall into. It does not change the binning and cannot help with the poor estimation quality when the number of samples is small or the range is not appropriate. The reason that these solutions cannot fully solve the problem has two parts. First, the neural network uncertainty can be more non-uniformly distributed compared with the weather forecast, especially when the weather forecast confidence scores are clustered to 11 uniformly distributed options in $[0, 1]$ [2]. Second, uncertainty for neural networks are more critical, especially in high confidence area *e.g.* the difference between the accuracy of 0.95 and 0.99 can be much more significant in pedestrian detection in an autonomous vehicle than that in the weather forecast.

5. Effect of Model Complexity on Uncertainty Estimation

In this section, we apply the proposed evaluation metrics in a series of experiments to validate the effectiveness and robustness of our proposed evaluation metrics and provide the first empirical study on how uncertainty-related model performance is affected by model complexity.

5.1. Relation Between the Two Use Cases

Before delving into the effect of model complexity on selective prediction and confidence calibration, it is interesting to analyze whether the effect would be similar for the two cases, which may help to justify the different trends observed from the experiments in Section 5.2.

When the prediction model is given, the performance of selective prediction is solely determined by the relative ranking of r_i and r_j for $c_i = 1$ and $c_j = 0$. The specific values of r do not matter due to the thresholding mechanism. Even though a good threshold may be unknown, existing statistical methods are available for finding a desirable threshold [9]. In contrast, for confidence calibration, the specific value of r does matter. The quality of the given confidence score is evaluated based on the difference between the confidence score and the expected accuracy of the samples with this score. As such, we expect that **performance of selective prediction and confidence calibration are not necessarily correlated.** One can construct illustrative examples to show that the confidence score r can be perfect for one case but bad for the other. When $r_i = 0.5$ for all $x_i \in \mathcal{D}$ and $\frac{\sum_{x_i \in \mathcal{D}} c_i}{|\mathcal{D}|} = \frac{1}{2}$, we have ECE$(P_{\theta,\mathcal{D}}) = 0$ but selective prediction is not feasible. When $r \in \{0.9, 1\}$, $c_i = 1$ for all $r_i = 1$ and $c_j = 0$ for all $r_j = 0.9$, selective prediction achieves the best possible result but ECE$(P_{\theta,\mathcal{D}})$ is as high as 0.9.

Proposition 2 further shows that a confidence estimation is perfect in both cases, if and only if it perfectly knows the correctness for each prediction which is almost impossible in practice. The proof is given in the supplementary.

Proposition 2. *The uncertainty estimation r is perfect for both selective prediction and confidence calibration, if and only if for all samples $r \in \{0, 1\}$, $E_{P_{\theta,\mathcal{D}}(c|r=0)}[c] = 0$, and $E_{P_{\theta,\mathcal{D}}(c|r=1)}[c] = 1$.*

The results imply that given limited learning capability, there exists a trade-off between two aspects of uncertainty estimation and the model should be optimized for the specific use case.

5.2. Experiment Results

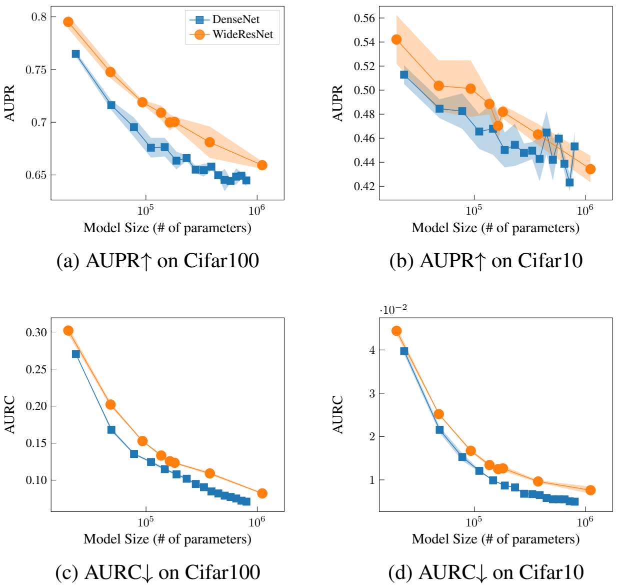
We evaluate the uncertainty estimation quality of a series of models for selective prediction and confidence calibration in image classification and medical image segmentation. There are different uncertainty estimation methods available. In this work, we use maximum softmax probability [14] and temperature scaling [13] for selective prediction and confidence calibration as they are popular and shown to be competitive with more complex approaches [41, 3, 10].

We first evaluate two popular networks DenseNet [17] and WideResNet [43] on Cifar10 and Cifar100 [20] to cover different levels of difficulty and accuracy. For DenseNet, we keep the growth rate at 12 and reduce its depth from 100 to 10. For WideResNet, we change the widen factor of a 16-layer network and a 28-layer network for a comparable number of parameters with DenseNet.

The performance of selective prediction increases with model size. For selective prediction, we first show how the conventional AUPR metric changes with the model size, and the results are shown in Figure 4a and Figure 4b. It is shown that AUPR decreases with the model size, indicating networks' decreasing capability to differentiate wrongly predicted samples and correctly prediction samples. The reason is that wrong predictions with high confidence scores, an issue known as over-confidence in high capacity neural networks [25], are usually caused by inherent learning limitation or data similarity instead of network capacity. As a result, although higher capacity models have fewer wrong predictions, they are increasingly concentrated in the high confidence area, which in turn makes accurate uncertainty estimation harder. However, this is against the intuition that larger networks learn probability distribution better and thus behave better in uncertainty estimation. The reason is that the impact of the original full-coverage accuracy in selective prediction is not taken into consideration. If we use the proposed AURC instead as shown in Figure 4c and Figure 4d, the estimation quality becomes consistent with common expectation, showing the increased performance of selective prediction with increasing model size.

The performance of confidence calibration is insensitive to the model size. For confidence calibration, the uncertainty estimation quality measured by AECE and AMCE is shown in Figure 5. We also plot the results measured by ECE and MCE in the same figures to validate the discussion. We find that the estimation quality remains almost flat and does not show a strong trend with the model complexity in terms of AECE and AMCE. This is a mixed result of a number of factors including model accuracy, the effectiveness of temperature scaling, and the confidence distribution.

In terms of the effect of adaptive binning, it is observed that AECE is generally bigger than ECE by a small margin. The reason is that different binning methods lead to different levels of internal compensation as discussed above. The adaptive binning used by AECE creates 12.6 bins and 21.2 bins on average for Cifar10 and Cifar100 respectively, which are more than the 10 equal-range bins



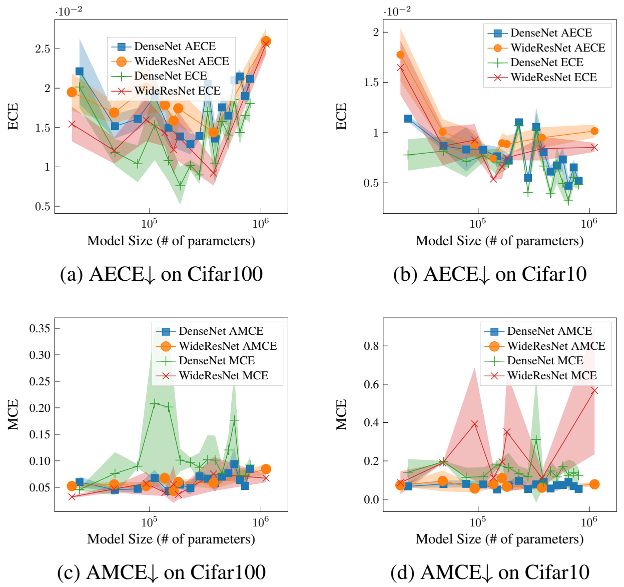
(a) AUPR↑ on Cifar100

(b) AUPR↑ on Cifar10

(c) AURC↓ on Cifar100

(d) AURC↓ on Cifar10

Figure 4: Effect of model complexity on selective prediction.



(a) AECE↓ on Cifar100

(b) AECE↓ on Cifar10

(c) AMCE↓ on Cifar100

(d) AMCE↓ on Cifar10

Figure 5: Effect of model complexity on confidence calibration.

used in ECE [13, 22, 38]. Note that Cifar100 gets more bins than Cifar10 because Cifar100 is more difficult and the confidence distribution is significantly flatter, which naturally enables more bins with accurate accuracy estimation. This further validates the superiority of the adaptive binning. Note that AECE is very close to ECE even when the underline bins are very different, because most of the samples are in the bin with the highest confidence (the case is more severe when the model has high accuracy) and these samples dominate the value of ECE and AECE. The advantage of using adaptive binning is much more significant in reliability diagrams and AMCE compared with the "expected" calibration error.

Meanwhile, as shown in Figure 5c, MCE is close to AMCE for WideResNet but significantly bigger than AMCE in some cases for DenseNet. The reason is that DenseNet tends to have confidence distributions that are more concentrated in the high confidence area. As a result, the inaccurate accuracy estimation in the bins with a small number of samples is exposed, and this leads to some undesired big calibration error. This is further validated in the results on Cifar10 where both WideResNet and DenseNet have more non-uniform confidence distributions because of the easier task. As shown in Figure 5d, the MCE for both WideResNet and DenseNet are unstable and significantly higher than AMCE indicating an even worse situation caused by the inaccurate accuracy estimation. In summary, the comparison between the baseline and adaptive binning validates our discussion and design intuition for adaptive binning.

We further validate the evaluation metric on medical image segmentation where uncertainty estimation is crucial. We trained a group of U-Nets [35] with standard practice and different widths on the Multi-Modality Whole Heart Segmentation dataset [45] . The observations are similar to the classification experiments except for a few differences. Firstly, because the network is optimized for better Dice instead of pixel-wise accuracy, the AUPR is not meaningful in this case. Secondly, the difference between ECE and AECE is almost zero. The small size of the dataset make the model trend to overfitting. There are also a lot of background voxels that are easy to predict. These two factors lead to the fact that most predictions have a confidence score close to 1. Then the calibration error at the high confidence area dominate the ECE, despite different binning strategy. However, MCE is still very unstable. This also validates that an excessive number of samples along cannot solve the issues of binning strategy. Detailed results are shown in the supplementary.

6. Conclusions

Understanding the quality of uncertainty estimation is critical when applying DNNs to real-world vision problems. We focus on two main use cases of uncertainty estimation, *i.e.*, selective prediction and confidence calibration. We identified the issues with the existing metrics for uncertainty estimation that may lead to unreliable or misleading results, and proposed new justified metrics to mitigate these issues. Finally, we validated the new metrics by exploring the effect of model complexity on uncertainty estimation while showing that selective prediction and confidence calibration have different complexity-uncertainty trade-offs.

References

[1] J Eric Bickel and Seong Dae Kim. Verification of the weather channel probability of precipitation forecasts. *Monthly Weather Review*, 136(12):4867–4881, 2008. 6

[2] Jochen Bröcker and Leonard A Smith. Increasing the reliability of reliability diagrams. *Weather and forecasting*, 22(3):651–661, 2007. 6

[3] Tongfei Chen, Jiří Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam. Confidence scoring using whitebox metamodels with linear classifier probes. *arXiv preprint arXiv:1805.05396*, 2018. 1, 2, 3, 7

[4] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006. 4

[5] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 1, 2

[6] Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018. 2

[7] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*, pages 9175–9186, 2018. 2

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 2

[9] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in neural information processing systems*, pages 4878–4887, 2017. 1, 4, 7

[10] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. *arXiv preprint arXiv:1901.09192*, 2019. 7

[11] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. *ICLR 2019*, 2018. 2, 3, 5

[12] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. 3

[13] 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-Volume 70*, pages 1321–1330. JMLR. org, 2017. 1, 2, 3, 7, 8

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 2, 3, 5, 7

[15] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. In *Advances in Neural Information Processing Systems*, pages 917–926, 2018. 1, 2, 3

[16] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017. 1

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 7

[18] Ian T Jolliffe and David B Stephenson. *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons, 2012. 6

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 2

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 7

[21] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2801–2809, 2018. 3

[22] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2810–2819, 2018. 1, 2, 3, 5, 8

[23] Tze Leung Lai, Shulamith T Gross, David Bo Shen, et al. Evaluating probability forecasts. *The Annals of Statistics*, 39(5):2356–2382, 2011. 6

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017. 1, 2

[25] Kimin Lee, Changho Hwang, Kyoung Soo Park, and Jinwoo Shin. Confident multiple choice learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2014–2023. JMLR. org, 2017. 7

[26] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 2

[27] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2

[28] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems*, pages 10622–10632, 2019. 2

[29] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018. 1, 2, 3, 5

[30] Amit Mandelbaum and Daphna Weinshall. Distance-based confidence score for neural network classifiers. *arXiv preprint arXiv:1709.09844*, 2017. 1, 2, 3, 5

[31] Edgar C Merkle and Mark Steyvers. Choosing a strictly proper scoring rule. *Decision Analysis*, 10(4):292–304, 2013. 3

[32] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 1, 2, 3

[33] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–663. Springer, 2018. 1, 2

[34] Jeremy Nixon, Mike Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *arXiv preprint arXiv:1904.01685*, 2019. 2, 6

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 8

[36] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019. 2

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic nighttime image segmentation with synthetic stylized data, gradual adaptation and uncertainty-aware evaluation. *arXiv preprint arXiv:1901.05946*, 2019. 1

[38] Jörg Sander, Bob D de Vos, Jelmer M Wolterink, and Ivana Išgum. Towards increased trustworthiness of deep learning segmentation methods on cardiac mri. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094919. International Society for Optics and Photonics, 2019. 1, 2, 3, 8

[39] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Confidence calibration in deep neural networks through stochastic inferences. *arXiv preprint arXiv:1809.10877*, 2018. 1, 2, 3

[40] Avanti Shrikumar and Anshul Kundaje. Calibration with bias-corrected temperature scaling improves domain adaptation under label shift in modern neural networks. *arXiv preprint arXiv:1901.06852*, 2019. 1

[41] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019. 7

[42] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B Schön. Evaluating model calibration in classification. *arXiv preprint arXiv:1902.06977*, 2019. 2, 6

[43] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 7

[44] Michaël Zamo and Philippe Naveau. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2):209–234, 2018. 6

[45] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Medical image analysis*, 58:101537, 2019. 8