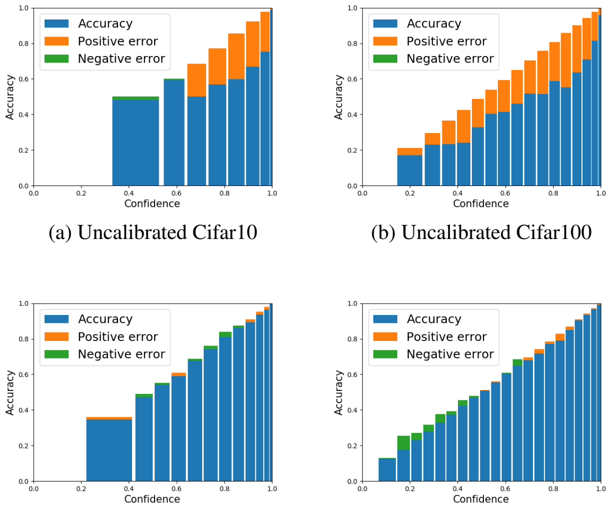
Supplementary Material for Revisiting the Evaluation of Uncertainty Estimation and Its Application to Explore Model Complexity-Uncertainty Trade-Off

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

A. Examples of Adaptive Binning

Figure 1 shows the Reliability Diagram of DenseNet on Cifar10 and Cifar100 datasets before and after model calibration using adaptive binning. It can be seen that the calibration with temperature scaling significantly reduces the calibration error. For a more difficult dataset and a calibrated model, more bins are used automatically.



(a) Uncalibrated Cifar10

(b) Uncalibrated Cifar100

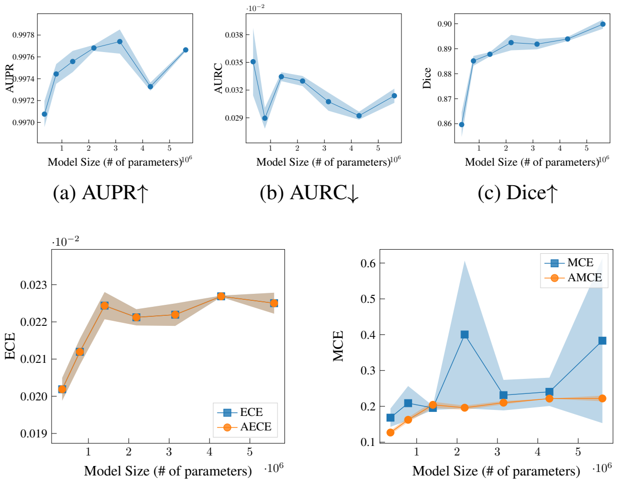
(c) Calibrated Cifar10

(d) Calibrated Cifar100

Figure 1: Reliability Diagrams of various models.

B Experiment Results of Medical Image Segmentation

Figure 2 shows the experiment results on the Multi-Modality Whole Heart Segmentation dataset.



(a) AUPR↑

(b) AURC↓

(c) Dice↑

(d) AECE↓

(e) AMCE↓

Figure 2: Effect of model complexity on uncertainty estimation in medical image segmentation.

C. Proofs

C.1 Proof of Theorem 1

Theorem 1. *For any two networks 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.*

Proof. Proof by contradiction. Since the wrong predictions are positive samples and correct predictions are negative samples. Having the same accuracy means that the two networks have the same number of positive and negative samples. Denote TN, TP, FN, FP, TPR, and FPR as true negative, true positive, false negative, false positive, true positive rate, and false positive rate respectively. Then we have

$$coverage = \frac{TN + FN}{TN + FN + TP + FP} \tag{1}$$

$$risk = \frac{FN}{TN + FN} \tag{2}$$

Suppose the curve of B dominates that of A in the ROC space but not in the Risk-Coverage space. Then there exists a point a on the curve of network A and a point b on the curve of B such that $coverage_a = coverage_b$ and $risk_a < risk_b$.

From $coverage_a = coverage_b$, we have $TN_a + FN_a = TN_b + FN_b$. Since $\frac{FN_a}{(TN_a+FN_a)} < \frac{FN_b}{(TN_b+FN_b)}$, we have $FN_a < FN_b$ and $TN_a > TN_b$.

Remember that the numbers of positive and negative samples are equal. Therefore we have $FP_a + TN_a = FP_b + TN_b$ and $TP_a + FN_a = TP_b + FN_b$. Then we obtain $FP_a < FP_b$ and $TP_a > TP_b$. Then we have $TPR_a > TPR_b$ and $FPR_a < FPR_b$.

This contradicts the fact that the curve of B is higher than that of A in the ROC space. The other direction can be proved in the same way. □

C.2 Proof of Proposition 1

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, $\hat{ECE}(P_{\theta,\mathcal{D}}) < ECE(P_{\theta,\mathcal{D}})$.*

Proof. For clarity, we reuse n, B_j and D_j as the number of bins, the range of bin, and the sample set where $j = \{1, \ldots, n\}$. Note that here the bin selection no longer needs to be a uniform partition. In order to make ECE($P_{\theta,\mathcal{D}}$) meaningful, we assume there are enough samples and $E_{P_{\theta,\mathcal{D}}(c|r_k)}[c]$ is a solvable value. Denote the number of different values of r as m and these different values of r as r_k where $k = \{1, \ldots, m\}$. Then we partition D to m bins $D_k = \{x_i | r_i = r_k\}$ so that each bin only has one unique r value. The ground truth ECE can be written as:

$$\text{ECE}(P_{\theta,\mathcal{D}}) = \frac{1}{|D|} \sum_{k=1}^{m} |D_k| |E_{P_{\theta,\mathcal{D}}(c|r_k)}[c] - r_k| \tag{3}$$

Then we have

$$\begin{aligned} \hat{\text{ECE}}(P_{\theta,\mathcal{D}}) &= \frac{1}{|D|} \sum_{j=0}^{n} | \sum_{x_i \in D_j} c_i - \sum_{x_i \in D_j} r_i | \quad (4) \\ &= \frac{1}{|D|} \sum_{j=0}^{n} | \sum_{r_k \in B_j} |D_k| (E_{P_{\theta,\mathcal{D}}(c|r_k)}[c] - r_k) | \end{aligned}$$

Note that

$$\sum_{r_k \in B_j} |D_k| (E_{P_{\theta,\mathcal{D}}(c|r_k)}[c] - r_k) \leq \sum_{r_k \in B_j} |D_k| \left| (E_{P_{\theta,\mathcal{D}}(c|r_k)}[c] - r_k) \right|$$

and they are equal 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 $r_k \in B_j$. Together with Equation 3, we conclude the proof. □

C.3 Proof of Proposition 2

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 $E_{P_{\theta,\mathcal{D}}(c|r=1)}[c] = 1$.*

Proof. Given $r \in \{0, 1\}$, $E_{P_{\theta,\mathcal{D}}(c|r=0)}[c] = 0$, and $E_{P_{\theta,\mathcal{D}}(c|r=1)}[c] = 1$, it follows trivially that $E_{P_{\theta,\mathcal{D}}(r)}[|E_{P_{\theta,\mathcal{D}}(c|r)}[c] - r|] = 0$ and $r_a > r_b$ for any $c_a = 1, c_b = 0$.

On the other side, if $E_{P_{\theta,\mathcal{D}}(r)}[|E_{P_{\theta,\mathcal{D}}(c|r)}[c] - r|] = 0$, we have $E_{P_{\theta,\mathcal{D}}(c|r)}[c] = r$. If there exists a x_i that $r_i \in (0, 1)$, then we $E_{P_{\theta,\mathcal{D}}(c|r_i)}[c] \in (0, 1)$.

Consequently, $|\{x | c = 0, r = r_i\}| > 0$ and $|\{x | c = 1, r = r_i\}| > 0$. Then for any two samples $x_a \in \{x | c = 1, r = r_i\}$ and $x_b \in \{x | c = 0, r = r_i\}$, we have $c_a = 1$, $c_b = 0$ and $r_a = r_b$ that contradict with the fact that $r_a > r_b$ for $c_a = 1$, $c_b = 0$. Therefore, $x_i \in \{0, 1\}$. Using $E_{P_{\theta,\mathcal{D}}(c|r)}[c] = r$, it follows immediately that $E_{P_{\theta,\mathcal{D}}(c|r=0)}[c] = 0$, and $E_{P_{\theta,\mathcal{D}}(c|r=1)}[c] = 1$. □