KU LEUVEN

Meta-Cal: Well-controlled Post-hoc Calibration by Ranking

Xingchen Ma, Matthew Blaschko

Introduction

Problem

We want a classifier:

1. make accurate predictions
2. output calibrated posterior probabilities

Measure calibration:

1. top-label calibration (ECE, confidence calibrated)
2. classwise calibration (Marginal)

Current Solutions and Problems

Two categories:

1. preserve accuracy: temperature based (ensemble, local)
2. accuracy not preserved: Dirichlet calibration, GP calibration

Problems:

1. calibration map family not flexible
2. accuracy not controlled

Our Solution

MetaCal

Components:

1. a base calibrator
2. a ranking model

Advantages:

1. prove to be more flexible
2. controlled mis-coverage
3. controlled coverage-accuracy

Quality assurance (QA) system:

1. products accepted by QA satisfy requirements with high confidence: coverage-accuracy
2. no unnecessary rejection: mis-coverage

Algorithms

mis-coverage

Algorithm 1 Meta-Cal (miscoverage control)

1: **Input:** Training data set $\{(x_i, y_i)\}_{i=1}^n$, miscoverage rate tolerance α, base calibration model g_m, ranking model h.
2: **Output:** Binary classifier $\hat{\phi}$, Meta-Cal calibration model g.
3: Partition the training data set randomly into two parts. The first part has only negative ($\hat{Y} = Y$) samples. The second part contains both negative and positive samples ($\hat{Y} \neq Y$).
4: Compute ranking scores on the first part using the ranking model h. Compute threshold r^* based on α.
5: Construct a binary classifier $\hat{\phi}$ based on r^*.
6: Train a base calibration model g_m using samples whose scores are smaller than r^* among the second part.
7: Construct the final calibration map g using updated rules.

coverage-accuracy

Algorithm 2 Meta-Cal (coverage accuracy control)

1: **Input:** Training data set $\{(x_i, y_i)\}_{i=1}^n$, desired coverage accuracy β, base calibration model g_m, ranking model h.
2: **Output:** Binary classifier $\hat{\phi}$, Meta-Cal calibration model g.
3: Randomly split the training data set into two parts.
4: Estimate the coverage accuracy transformation $\hat{l}$ on the first part.
5: Compute a threshold $r^* = \hat{l}^{-1}(\beta)$ based on the estimated $\hat{l}$ and β.
6: Construct a binary classifier $\hat{\phi}$ based on r^*.
7: Train a base calibration model g_m using samples among the second part whose scores are smaller than r^*.
8: Construct the final calibration map g using updated rules.

Bounds

Mis-coverage:

$$\mathbb{P}\left(\left|\hat{F}_0(g) - F_0(g)\right| \geq \delta\right) \approx \mathbb{P}\left(|R_0 - F_0(g)| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2\sigma^2}\right)$$

Coverage-accuracy:

$$\mathbb{P}\left(\left|\hat{F}_1(g) - \beta\right| \geq \delta\right) \approx \mathbb{P}\left(|R_1 - \beta| \geq \delta\right) \leq 2\exp\left(-\frac{m_1\delta^2}{2\beta(1-\beta)}\right)$$

Results

Calibration Comparisons

Table 1. ECE comparison. *Uncal*, *TS*, *ETS*, *GPC*, *MetaMis*, *MetaAcc* denote no-calibration, temperature scaling, ensemble temperature scaling, Gaussian Process calibration, Meta-Cal under miscoverage rate constraint and Meta-Cal under coverage accuracy constraint respectively. Reported values are the average of 40 independent runs. All standard errors are less than $5e-4$.

Dataset	Network	Acc	Uncal	TS	ETS	GPC	MetaMis	MetaAcc
CIFAR10	DenseNet40	0.9242	0.05105	0.00510	0.00567	0.00634	0.00434	0.00355
	ResNet110	0.9356	0.04475	0.00781	0.00809	0.00684	0.00391	0.00441
	ResNet110SD	0.9404	0.04022	0.00439	0.00509	0.00364	0.00350	0.00315
	WideResNet32	0.9393	0.04396	0.00706	0.00712	0.00684	0.00485	0.00532
CIFAR100	DenseNet40	0.7000	0.21107	0.01067	0.01104	0.01298	0.01093	0.00793
	ResNet110	0.7148	0.18182	0.02037	0.02130	0.01348	0.01815	0.01441
	ResNet110SD	0.7283	0.15496	0.01043	0.01057	0.01265	0.01109	0.00733
	WideResNet32	0.7382	0.18425	0.01332	0.01351	0.00993	0.01332	0.01189
ImageNet	DenseNet161	0.7705	0.05531	0.02053	0.02064	NA	0.01388	0.01248
	ResNet152	0.7620	0.06290	0.02023	0.02004	NA	0.01360	0.01138

Bounds Verification

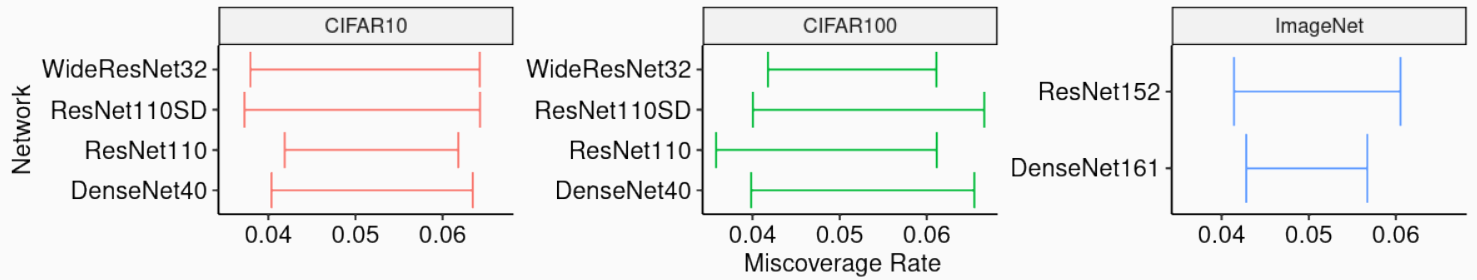

Figure 1. Empirical miscoverage rate. The error bars show ± 2 standard deviation of 40 independent runs. The desired miscoverage rates for CIFAR-10, CIFAR-100 and ImageNet are all set to be 0.05.

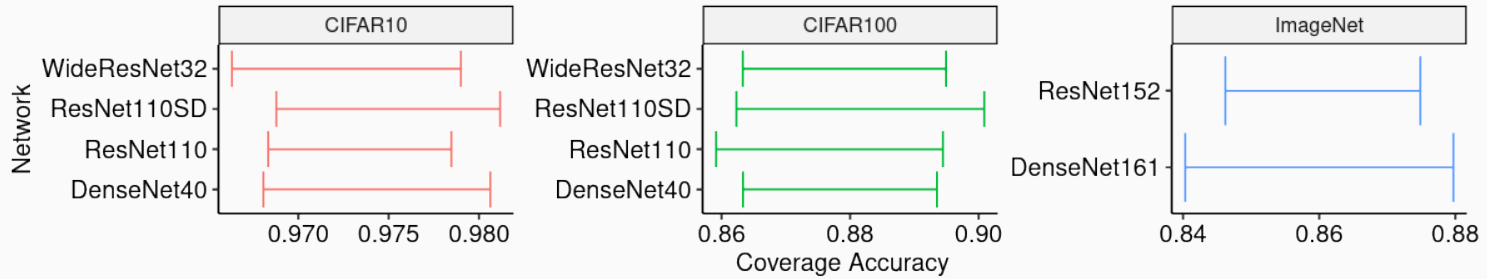

Figure 2. Empirical coverage accuracy. The error bars show ± 2 standard deviation of 40 independent runs. The desired coverage accuracy for CIFAR-10, CIFAR-100 and ImageNet are set to be 0.97, 0.87 and 0.85, respectively.

links

paper link

https://arxiv.org/abs/2105.04290

code link

https://github.com/maxc01/metacal