Towards Unbiased Calibration using Meta-Regularization

Cheng Wang
Amazon

Jacek Golebiowski
distil labs
(work done at Amazon)

What is calibration?

Confidence calibration is the problem of predicting probability estimates representative of the true correctness likelihood

Predicted probability (confidence): the probability of a data point x having label y as predicted by the classifier
Observed probability (accuracy): the fraction of data points with the correct label assignment

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$$

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

$$\mathrm{ECE} = \sum_{n=1}^{N} \frac{|b_n|}{m} |\mathrm{acc}(b_n) - \mathrm{conf}(b_n)|$$

$$\mathrm{MCE} = \max_{n \in 1,\dots,N} |\mathrm{acc}(b_n) - \mathrm{conf}(b_N)|$$

How to calibrate: Motivation

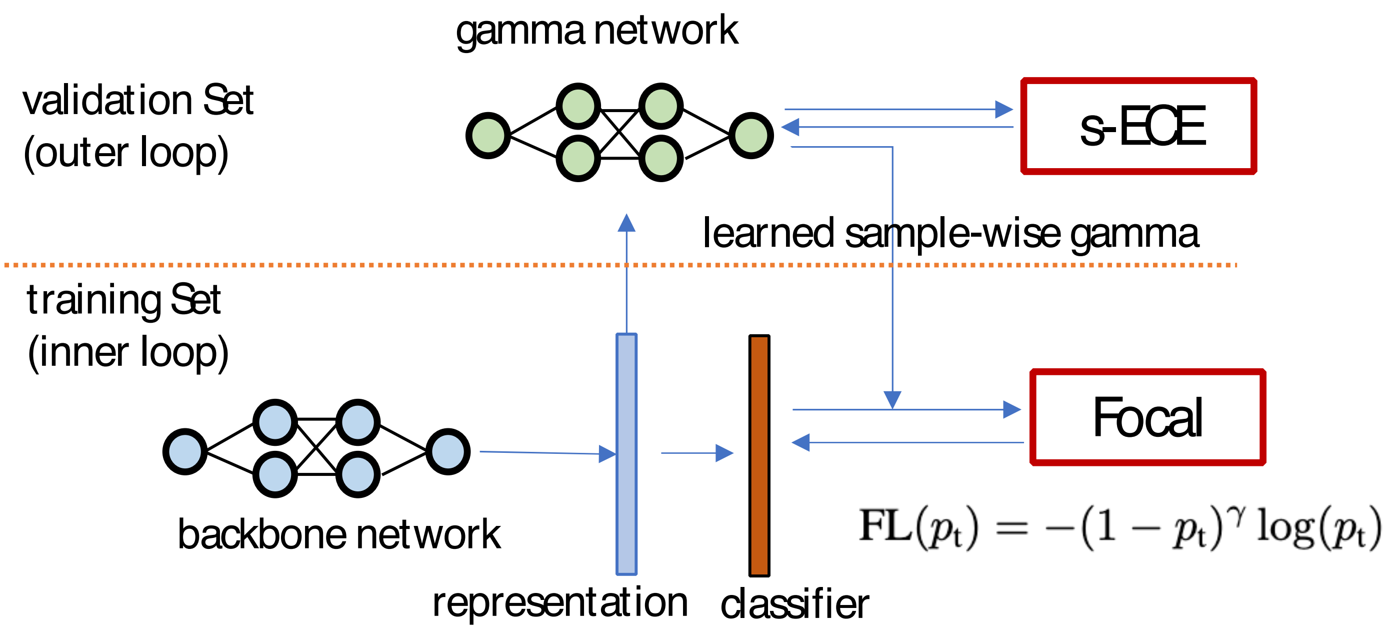
$$\mathrm{FL}(p_t) = -\underbrace{\boxed{(1 - p_t)^{\gamma}}}_{\text{Focal modification}}\underbrace{\boxed{\log(p_t)}}_{\text{Cross-entropy}}.$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise,} \end{cases}$$

Tsung-Yi Lin et al. Focal Loss for Dense Object Detection

How to calibrate: Motivation

- gamma-Net: predict the right hyper parameter of the focal loss (trained with meta learning)
- SECE: Improve model calibration via a differentiable calibration proxy.



How to calibrate: Focal loss with γ-Net

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t).$$

$\mathbf{x} \in \mathbb{R}^{b \times d}$ (b: batch size, d: hidden dimension)

$\mathbf{A} \in \mathbb{R}^{d \times k}$

$$\boxed{\begin{aligned} \mathbf{a} &= \mathbf{x} \cdot \mathbf{A}, \quad \in \mathbb{R}^{b \times k} \\ \mathbf{p} &= \textsc{Softmax}(a), \quad \in \mathbb{R}^{b \times k} \\ \tilde{\mathbf{x}} &= \mathbf{p} \cdot \mathbf{A}^{\top}, \quad \in \mathbb{R}^{b \times d} \end{aligned}}$$

$$\boxed{\gamma = |\tilde{\mathbf{x}} \cdot \mathbf{W}| / \tau, \quad \in \mathbb{R}^{b \times 1}}$$

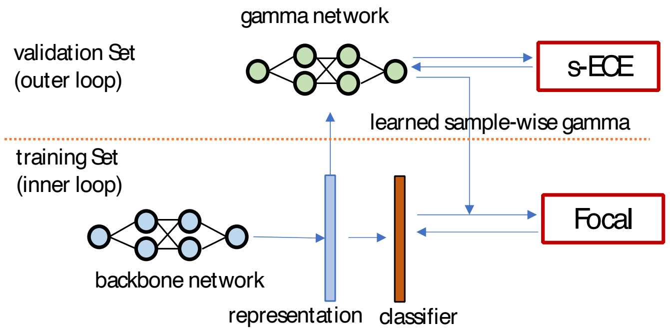
How to calibrate: sECE

$$\mathrm{ECE} = \sum_{n=1}^{N} \frac{|b_n|}{m} \left| \boxed{\mathrm{acc}(b_n)} - \boxed{\mathrm{conf}(b_n)} \right|$$

$$sACC(b_i) = \sum_{j \in \{1,\ldots,M\}} \mathrm{acc}(x_j) k(p_i, p_j)$$

$$sECE = \sum_{n=1}^{M} \frac{1}{M} \left| sACC(n) - \mathrm{conf}(n) \right|$$

How to calibrate: meta learning

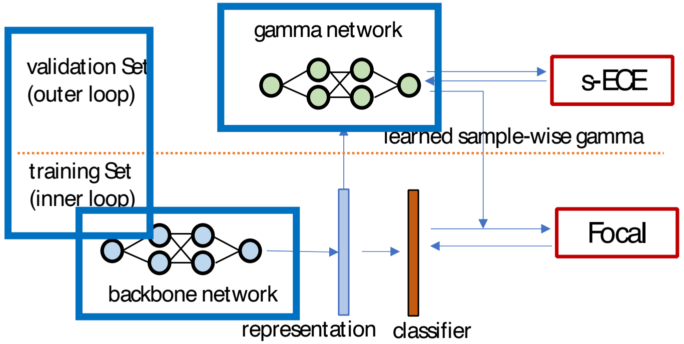


Algorithm 1: Meta optimization with γ-Net and s-ECE

Input: f^c and f^γ with initialized θ and ϕ respectively
Output: Optimized θ and ϕ
Data: Training data D_{train} and validation data D_{val}

while θ *not converged* **do**
- $(\mathbf{x}_i^t, \mathbf{y}_i^t) \sim D_{train}$: Sample a mini-batch of training data
- $(\mathbf{x}_i^v, \mathbf{y}_i^v) \sim D_{val}$: Sample a mini-batch of validation data
- Find $\gamma = f^\gamma(\mathbf{x}_i^t)$
- Compute $\mathcal{L}_\gamma^f(f^c(\mathbf{x}_i^t), \mathbf{y}_i^t)$ based on γ
- Use $\nabla_\theta \mathcal{L}_\gamma^f$ to update θ, parameters of f^c
- Compute auxiliary loss s-ECE$(f^c(\mathbf{x}_i^v), \mathbf{y}_i^v)$
- Use $\nabla_\phi s$-ECE to update ϕ, parameters of f^γ

How to calibrate: meta learning

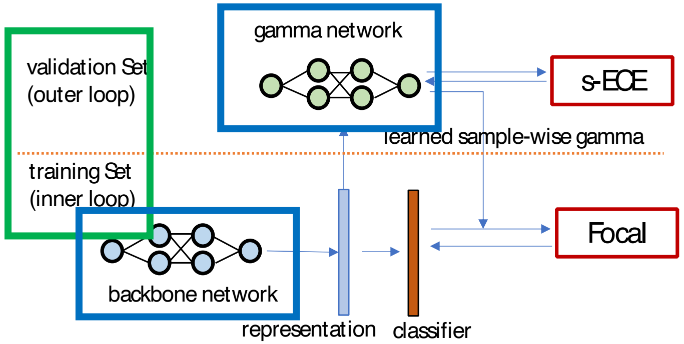


Algorithm 1: Meta optimization with γ-Net and s-ECE

Input: f^c and f^γ with initialized θ and ϕ respectively
Output: Optimized θ and ϕ
Data: Training data D_{train} and validation data D_{val}

```
while θ not converged do
    (x_i^t, y_i^t) ~ D_train: Sample a mini-batch of training data
    (x_i^v, y_i^v) ~ D_val: Sample a mini-batch of validation data
    Find γ = f^γ(x_i^t)
    Compute L_γ^f(f^c(x_i^t), y_i^t) based on γ
    Use ∇_θ L_γ^f to update θ, parameters of f^c
    Compute auxiliary loss s-ECE(f^c(x_i^v), y_i^v)
    Use ∇_φ s-ECE to update φ, parameters of f^γ
```

How to calibrate: meta learning

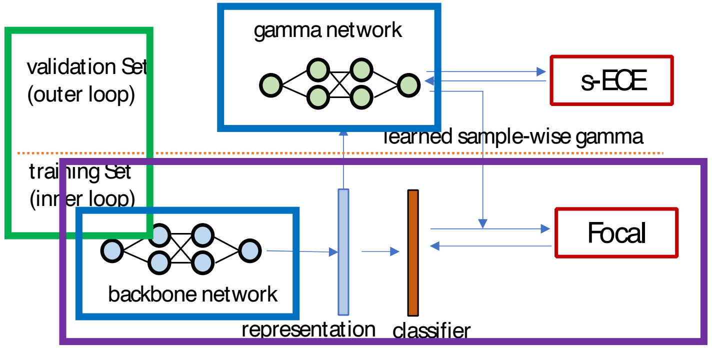


Algorithm 1: Meta optimization with γ-Net and s-ECE

Input: f^c and f^γ with initialized θ and ϕ respectively
Output: Optimized θ and ϕ
Data: Training data D_{train} and validation data D_{val}

while *θ not converged* **do**
- $(\mathbf{x}_i^t, \mathbf{y}_i^t) \sim D_{train}$: Sample a mini-batch of training data
- $(\mathbf{x}_i^v, \mathbf{y}_i^v) \sim D_{val}$: Sample a mini-batch of validation data
- Find $\gamma = f^\gamma(\mathbf{x}_i^t)$
- Compute $\mathcal{L}_\gamma^f(f^c(\mathbf{x}_i^t), \mathbf{y}_i^t)$ based on γ
- Use $\nabla_\theta \mathcal{L}_\gamma^f$ to update θ, parameters of f^c
- Compute auxiliary loss s-ECE$(f^c(\mathbf{x}_i^v), \mathbf{y}_i^v)$
- Use $\nabla_\phi s$-ECE to update ϕ, parameters of f^γ

How to calibrate: meta learning

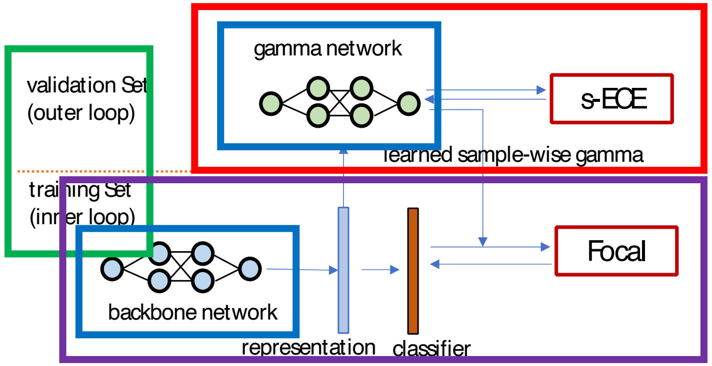


Algorithm 1: Meta optimization with γ-Net and s-ECE

Input: f^c and f^γ with initialized θ and ϕ respectively
Output: Optimized θ and ϕ
Data: Training data D_{train} and validation data D_{val}

while *θ not converged* **do**
- $(\mathbf{x}_i^t, \mathbf{y}_i^t) \sim D_{train}$: Sample a mini-batch of training data
- $(\mathbf{x}_i^v, \mathbf{y}_i^v) \sim D_{val}$: Sample a mini-batch of validation data
- Find $\gamma = f^\gamma(\mathbf{x}_i^t)$
- Compute $\mathcal{L}_\gamma^f(f^c(\mathbf{x}_i^t), \mathbf{y}_i^t)$ based on γ
- Use $\nabla_\theta \mathcal{L}_\gamma^f$ to update θ, parameters of f^c
- Compute auxiliary loss s-ECE$(f^c(\mathbf{x}_i^v), \mathbf{y}_i^v)$
- Use $\nabla_\phi s$-ECE to update ϕ, parameters of f^γ

How to calibrate: meta learning



Algorithm 1: Meta optimization with γ-Net and s-ECE

Input: f^c and f^γ with initialized θ and ϕ respectively
Output: Optimized θ and ϕ
Data: Training data D_{train} and validation data D_{val}

while θ *not converged* **do**
- $(\mathbf{x}_i^t, \mathbf{y}_i^t) \sim D_{train}$: Sample a mini-batch of training data
- $(\mathbf{x}_i^v, \mathbf{y}_i^v) \sim D_{val}$: Sample a mini-batch of validation data
- Find $\gamma = f^\gamma(\mathbf{x}_i^t)$
- Compute $\mathcal{L}_\gamma^f(f^c(\mathbf{x}_i^t), \mathbf{y}_i^t)$ based on γ
- Use $\nabla_\theta \mathcal{L}_\gamma^f$ to update θ, parameters of f^c
- Compute auxiliary loss s-ECE$(f^c(\mathbf{x}_i^v), \mathbf{y}_i^v)$
- Use $\nabla_\phi s$-ECE to update ϕ, parameters of f^γ

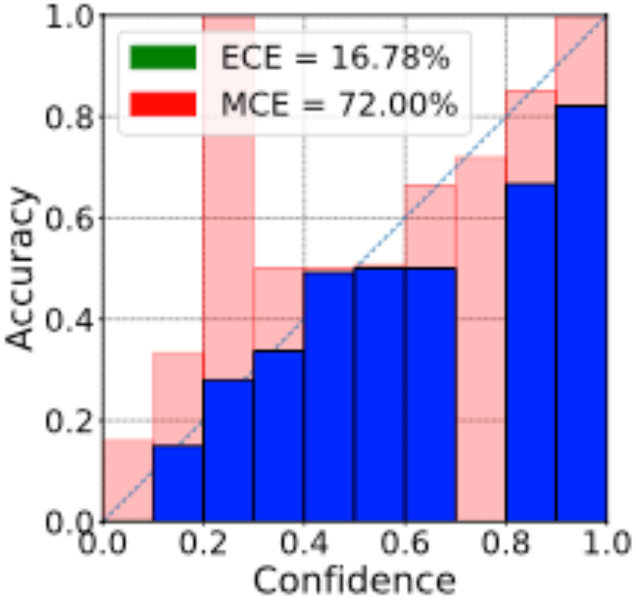
Results

Methods	Error	NLL	ECE	MCE	ACE	Classwise ECE
			CIFAR 10			
CE	4.812 ± 0.122	0.335 ± 0.01	4.056 ± 0.092	33.932 ± 5.433	4.022 ± 0.136	0.848 ± 0.023
CE (TS)	4.812 ± 0.122	0.211 ± 0.005	3.083 ± 0.140	26.695 ± 2.959	3.046 ± 0.157	0.656 ± 0.022
Focal	4.874 ± 0.100	0.207 ± 0.005	3.193 ± 0.104	28.034 ± 5.702	3.174 ± 0.098	0.690 ± 0.018
FLSD	4.916 ± 0.074	0.211 ± 0.005	6.904 ± 0.462	**19.246 ± 11.071**	6.805 ± 0.446	1.465 ± 0.088
LS (0.05)	4.744 ± 0.126	0.232 ± 0.003	2.900 ± 0.085	24.860 ± 8.599	3.985 ± 0.154	0.727 ± 0.009
LS(0.1)	4.918 ± 0.085	0.266 ± 0.004	7.566 ± 0.41	16.033 ± 3.783	7.611 ± 0.161	1.637 ± 0.056
Mixup(α=1.0)	**4.126 ± 0.068**	0.273 ± 0.033	12.863 ± 3.2	20.739 ± 4.205	12.833 ± 3.161	2.678 ± 0.615
MMCE	4.808 ± 0.082	0.333 ± 0.012	4.027 ± 0.082	41.647 ± 10.275	4.013 ± 0.091	0.845 ± 0.014
CE-DECE	5.194 ± 0.161	0.301 ± 0.038	4.106 ± 0.402	41.346 ± 13.325	4.088 ± 0.395	0.868 ± 0.074
CE-SECE	5.222 ± 0.168	0.289 ± 0.027	4.062 ± 0.241	50.81 ± 21.705	4.049 ± 0.251	0.852 ± 0.040
FL$_\gamma$-DECE	5.434 ± 0.095	**0.193 ± 0.009**	2.257 ± 0.787	56.633 ± 23.856	2.396 ± 0.669	**0.557 ± 0.165**
FL$_\gamma$-SECE	5.428 ± 0.144	**0.193 ± 0.010**	**2.138 ± 0.819**	22.725 ± 5.756	**2.357 ± 0.541**	**0.556 ± 0.165**
			CIFAR 100			
CE	22.570 ± 0.438	0.997 ± 0.014	8.380 ± 0.336	23.250 ± 2.436	8.347 ± 0.344	0.233 ± 0.006
CE (TS)	22.570 ± 0.438	0.959 ± 0.008	5.388 ± 0.393	13.454 ± 2.377	5.360 ± 0.315	0.208 ± 0.003
Focal	22.498 ± 0.214	0.900 ± 0.007	5.044 ± 0.203	12.454 ± 0.893	5.015 ± 0.207	0.203 ± 0.004
FLSD	22.656 ± 0.113	0.876 ± 0.005	5.956 ± 0.804	14.716 ± 1.387	5.958 ± 0.802	0.241 ± 0.008
LS (0.05)	21.810 ± 0.172	1.070 ± 0.011	8.108 ± 0.346	20.268 ± 1.536	8.106 ± 0.346	0.272 ± 0.006
LS(0.1)	22.244 ± 0.155	1.052 ± 0.011	4.754 ± 0.709	17.228 ± 0.923	4.777 ± 0.647	0.239 ± 0.004
Mixup(α=1.0)	**21.210 ± 0.227**	0.917 ± 0.017	9.716 ± 0.754	16.01 ± 1.335	9.722 ± 0.740	0.315 ± 0.011
MMCE	22.490 ± 0.143	1.021 ± 0.007	8.713 ± 0.245	23.565 ± 1.141	8.670 ± 0.305	0.238 ± 0.004
CE-DECE	23.406 ± 0.323	1.148 ± 0.006	7.309 ± 0.245	22.565 ± 1.446	7.253 ± 0.315	0.241 ± 0.002
CE-SECE	23.448 ± 0.302	1.153 ± 0.015	7.668 ± 0.330	24.261 ± 1.614	7.609 ± 0.295	0.244 ± 0.002
FL$_\gamma$-DECE	23.712 ± 0.204	0.888 ± 0.009	**1.879 ± 0.440**	8.271 ± 2.651	**1.838 ± 0.371**	0.195 ± 0.005
FL$_\gamma$-SECE	23.686 ± 0.377	**0.877 ± 0.004**	1.940 ± 0.365	**7.480 ± 1.867**	**1.939 ± 0.379**	**0.192 ± 0.006**
			Tiny ImageNet			
CE	40.110 ± 0.110	1.838 ± 0.171	8.059 ± 1.296	15.73 ± 1.905	8.006 ± 1.282	0.154 ± 0.001
Focal	39.415 ± 0.625	1.896 ± 0.009	7.600 ± 0.309	13.771 ± 0.897	7.469 ± 0.301	0.152 ± 0.002
FLSD	39.705 ± 0.075	1.904 ± 0.025	14.501 ± 1.078	21.528 ± 2.116	14.501 ± 1.078	0.202 ± 0.006
LS (0.1)	**39.395** ± 0.305	2.185 ± 0.001	16.777 ± 0.476	29.088 ± 1.835	16.901 ± 0.460	0.199 ± 0.001
Mixup(α=1.0)	39.890 ± 0.271	1.932 ± 0.054	12.133 ± 2.069	31.440 ± 0.968	12.028 ± 2.079	0.193 ± 0.009
MMCE	40.310 ± 0.100	1.826 ± 0.177	8.206 ± 1.219	16.802 ± 2.339	8.165 ± 1.269	**0.149** ± 0.001
CE-DECE	41.350 ± 0.000	2.228 ± 0.033	10.694 ± 0.503	20.888 ± 0.430	10.553 ± 0.553	0.160 ± 0.000
CE-SECE	41.005 ± 0.145	2.213 ± 0.058	10.928 ± 1.125	21.362 ± 2.526	10.912 ± 1.069	0.157 ± 0.003
FL$_\gamma$-DECE	40.635 ± 0.095	**1.826 ± 0.007**	5.944 ± 1.090	11.543 ± 1.990	6.077 ± 1.095	0.155 ± 0.007
FL$_\gamma$-SECE	40.850 ± 0.140	1.829 ± 0.005	**5.794** ± 0.756	**11.477** ± 1.563	**5.848** ± 0.751	0.156 ± 0.005

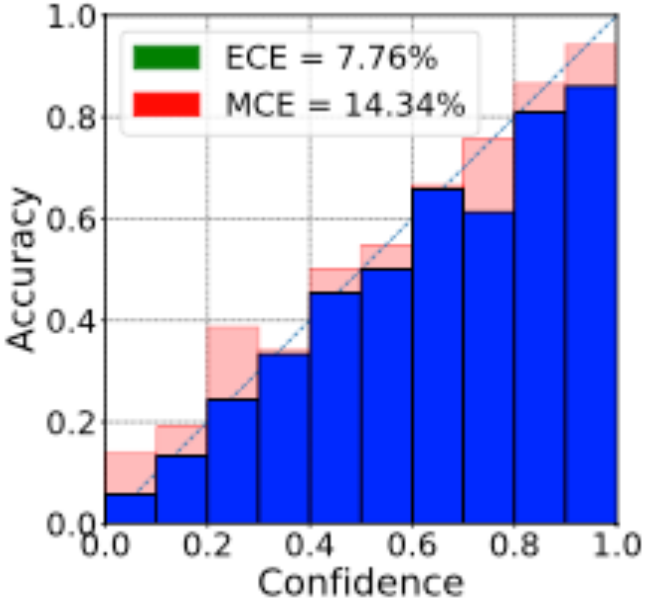
- Our approach (FL r-SECE) achieves lower errors across multiple calibration metrics.
- Our approach (FL r-SECE) achieves comparable predictive performance.

For the two production model candidates, which one do you pick?

Uncalibrated
(offline accuracy = 83.8%)



Calibrated
(offline accuracy = 81.2%)



Thank you !