Adaptive Calibrator Ensemble: Navigating Test Set Difficulty in Out-of-Distribution Scenarios

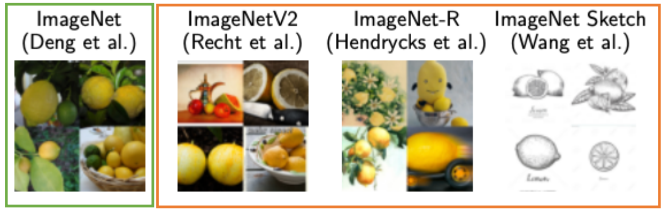
Yuli Zou[1] Weijian Deng[2] Liang Zheng[2]

[1]Hong Kong Polytechnic University [2]Australian National University

ICCV23 PARIS

Out-of-distribution (OOD) Calibration

Distribution Shift: test samples are from a different distribution than the calibration set

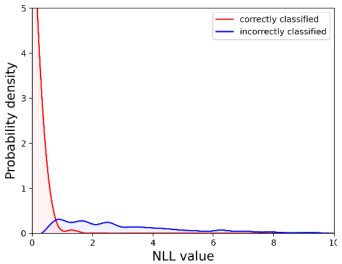


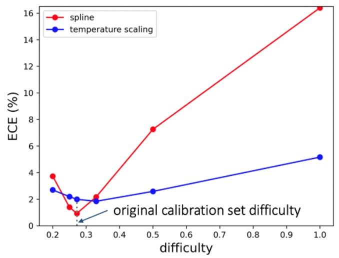
Calibration Set OOD Test Sets

Post-hoc calibration methods fall short under distribution shifts

Tentative Explanation: Calibration Set Difficulty

Difficulty: the ratio of the number of incorrectly classified samples to that of correctly classified samples





Observations: an individual sample matters in classification loss; calibration objective depends on dataset difficulty

Why OOD calibration fails?
The difficulty levels are different between calibration and OOD test sets, leading to distinct optimal calibration functions

Adaptive Calibrator Ensemble (ACE)

Step1: Seeking two calibration sets (D_o, D_h) with extreme difficulty levels: an ID difficulty level (d_o) and a high difficulty level (d_h);

Step2: Training two calibrators $\mathbf{f}_o$ and $\mathbf{f}_h$ on D_o and D_h, respectively. Then the logits z_o and z_h are obtained;

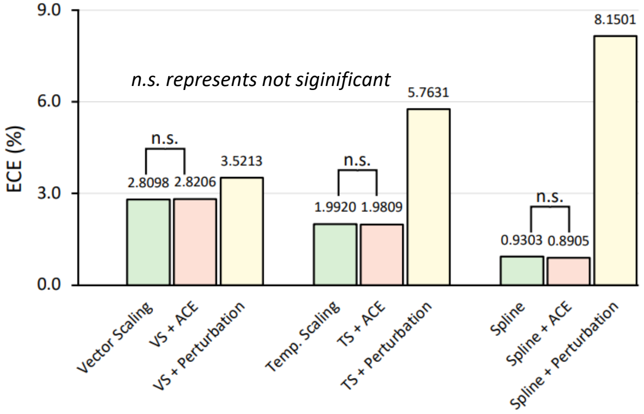
Step3: An adaptive weighting average scheme to fuse the output of calibrators trained on the two extreme calibration sets:

$$z_{\text{cal}} = \alpha z_o + (1-\alpha) z_h$$

We use average confidence score to indicate the OOD degree of test set. Thus, we compute the weight α as:

$$\alpha = \frac{\text{avgConf}(D_{\text{test}})}{\text{avgConf}(D_o)}$$

Results on In-Distribution Test Set



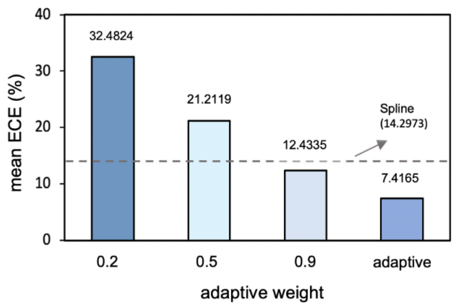
ACE **does not compromise** in-distribution calibration performance

Results on Out-of-Distribution Test Sets

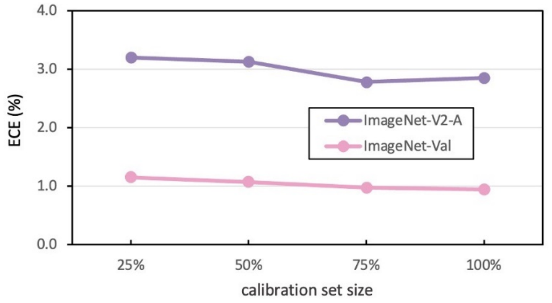
Methods	ImgNet-V2-A	ImgNet-V2-B	ImgNet-V2-C	ImgNet-S	ImgNet-R	ImgNet-Adv
uncalibrated	9.5016	6.2311	4.3117	24.6332	17.8621	50.8544
Vector Scaling	6.8068	4.2184	2.9258	20.3726	14.5037	44.7593
+ ACE	5.6291 ±0.0397 ▲	3.7742 ±0.0237 ▲	3.1141 ±0.0150 ▼	15.8747 ±0.0252 ▲	10.6343 ±0.0356 ▲	40.5773 ±0.0491 ▲
Temp. Scaling	4.4413	2.7309	1.6831	15.7879	10.4797	42.6302
+ ACE	3.5615 ±0.0028 ▲	2.5692 ±0.0013 ▲	1.7021 ±0.0001 ▼	10.3915 ±0.0092 ▲	6.7458 ±0.0083 ▲	38.0651 ±0.0114 ▲
Spline	4.5321	**1.8034**	1.3357	19.6392	13.1116	45.3623
+ ACE	**2.8201** ±0.0283 ▲	2.0235 ±0.0154 ▼	**1.0550** ±0.0092 ▲	**6.9264** ±0.0864 ▲	**6.8533** ±0.0011 ▲	**31.0926** ±0.0422 ▲

ACE **improves** calibration methods on out-of- distribution datasets

Component Analysis



The adaptive weight α achieves **lower** meanECE over various OOD test sets and ID test set than fixed value



ACE method is **stable** when simultaneously reduce the size of D_o and D_h by a certain percentage

Code is avaliable at https://github.com/insysgroup/Adaptive-Calibrators-Ensemble.git