Two Sides of Mis-Calibration: Identifying Over and Under-Confidence Prediction for Network Calibration (Supplementary Material)

Shuang Ao[1] **Stefan Rueger**[2] **Advaith Siddharthan**[3]

[1,2,3]Knowledge Media Institute., The Open University, Milton Keynes, UK

A IDENTIFYING UNDER-CONFIDENCE

Table A1: Results of average under-confidence mis-calibration score (UC MCS) and average over-confidence miscalibration score (OC MCS) for baseline, TS and our proposed method cwMCS TS. All results are shown in percentage for clarity. Best results for each row are shown in bold. The value in the bracket shows the percentage of class being under or over-confident.

		Baseline		TS		cwMCS TS	
Dataset	Model	UC MCS (%)	OC MCS (%)	UC MCS (%)	OC MCS (%)	UC MCS (%)	OC MCS (%)
IN	ViT	-4.2 (48.3)	7.4 (51.7)	-5.2 (61.3)	7.2 (38.7)	**-3.7 (58.3)**	**0.5 (41.7)**
	SwinT	-11.6 (85.8)	9.2 (14.2)	-5.3 (64)	9.2 (36)	**-0.5 (65.7)**	**8.7 (34.3)**
	DeiT	-10.4 (79.6)	9.0 (20.4)	-5.2 (54.7)	9.6 (45.3)	**-3.2 (55.3)**	**5.6 (44.7)**
	CaiT	-6.7 (67.6)	9.8 (32.4)	-5.2 (57.6)	9.6 (42.4)	**-3.2 (58.2)**	**7.6 (41.8)**
	BeiT	-9.0 (81.9)	8.9 (18.1)	-5.4 (65.3)	8.5 (34.7)	**-4.1 (62.2)**	**8.2 (37.8)**
	CoaT	-10.9 (83.1)	8.7 (16.9)	-6.8 (55.6)	9.0 (44.4)	**-5.4 (57.8)**	**8.0 (42.2)**
	CrossViT	-9.6 (76.8)	9.5 (23.2)	-5.5 (56.1)	9.6 (43.9)	**-3.5 (56.1)**	**8.6 (43.9)**
	ConvMix	-19.0 (90.4)	**8.1 (9.6)**	-8.9 (61.5)	8.6 (38.5)	**-5.7 (59.7)**	8.4 (40.3)
	ConvNext	-5.9 (59.8)	9.3 (40.2)	-5.3 (53.8)	9.2 (46.2)	**-4.3 (51.6)**	**9.0 (48.4)**
	ResNet34	**-4.8 (40.1)**	9.9 (59.9)	-6.4 (53.8)	8.6 (46.2)	-6.3 (52)	**8.5 (48)**
	DenseNet121	**-5.2 (43.7)**	9.3 (56.3)	-6.3 (55.4)	**8.4 (44.6)**	-6.4 (55.2)	**8.4 (44.8)**
	VGG16	**-5.3 (40.5)**	8.8 (59.5)	-6.3 (53.9)	8.1 (46.1)	-6.2 (54)	**8.0 (46)**
	EfficientNet	-17.0 (90.3)	8.4 (9.7)	-16.0 (87.4)	8.2 (12.6)	**-8.8 (53.6)**	**1.2 (46.4)**
Tiny-IN	ResNet34	**-3.1 (11.5)**	10.6 (88.5)	-5.8 (58)	6.3 (42)	-5.4 (52)	**4.3 (48)**
	DenseNet121	-1.0 (2.5)	13.4 (97.5)	-5.5 (58.5)	7.6 (41.5)	**-5.2 (56.5)**	**6.6 (43.5)**
	VGG16	**-0.8 (1.5)**	15.9 (98.5)	-5.2 (57.5)	5.1 (42.5)	-4.9 (52.5)	**3.1 (47.5)**
C100	Res34	**-0.1 (5)**	13.4 (95)	-5.2 (61)	6.2 (39)	-5.0 (57)	**4.2 (43)**
	DenseNet121	**-1.3 (1)**	15 (99)	-5.7 (55)	6.3 (46)	-4.3 (58)	**5.6 (42)**
	VGG16	**-1.3 (3)**	10.3 (97)	-4.4 (56)	8.7 (44)	-4.4 (56)	**3.7 (44)**
C10	ResNet34	0.0 (0)	4.1 (100)	**-1.4 (30)**	2.1 (70)	-1.5 (30)	**1.5 (70)**
	DenseNet121	0.0 (0)	11.3 (100)	-1.6 (40)	5.4 (60)	**-0.2 (30)**	**5.2 (70)**
	VGG16	0.0 (0)	7.1 (100)	-0.9 (40)	2.8 (60)	**-0.7 (50)**	**0.2 (50)**

Table A1 illustrates the results of mean under-confidence and mean over-confidence scores, as well as the percentage of classes with different confidence statuses correspondingly. For ImageNet dataset with transformers variants, most of the classes are under-confident with baselines, where the absolute value of mean under-confidence score is higher than the mean over-confidence score. The model with the highest percentage of under-confident classes is ConvMix, where only 10 percent of classes are over-confident. When it comes to CNNs, over and under-confident classes are more balanced. Surprisingly, EfficientNet has a similar behavior as Convmix, where the percentage of under-confident classes are much higher than over-confident ones. After applying TS, the percentage of over and under-confident classes are more balanced, and our proposed method cwMCS TS keeps this trend. Compared to baseline, our cwMCS TS method almost halves over and under-confidence scores, whereas TS only makes a slight change of them. For Tiny-ImageNet, CIFAR100 and CIFAR10

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

datasets with CNNs, more than 90 percent of classes are over-confident in baselines, with none of the classes under-confident for CIFAR10 dataset. However, more than half of the classes become under-confident after applying TS, indicating that TS can overly calibrate models. Our proposed method cwMCS TS significantly improves the mean over-confidence score and contributes to better calibration for under-confident classes.