Improving Classifier Confidence using Lossy Label-Invariant Transformations: Supplementary Material

A Brier score

Brier score (Brier, 1950) is defined as Equation 5.

$$Brier = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(p_{ik} - y_{ik})^2 \tag{5}$$

where, N is the dataset size, K is the number of classes, p_{ik} is the confidence for the label k of the i^{th} data, and y_{ik} is 1 if the true label for i^{th} data is k otherwise 0. Here, we normalized Brier score by dividing it by the number of classes as used in Kull et al. (2019). The formal definition of the normalized Brier score is shown in Equation 6.

$$Normalized\ Brier = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}(p_{ik} - y_{ik})^2 \tag{6}$$

B Flowchart for Runtime Confidence Calculation using ReCal

After ReCal learns a calibration map, it can process runtime calibration as illustrated in Figure 4. It consists of two steps: initialization step, iterative group-wise calibration. In initialization step, it computes the base logits for original inputs and transformed inputs by the sampled transformation during the learning process. After the initialization step, it repeats 1) Find a group number for the given input, 2) Calibrate logits using given temperature parameter. The detail explanation about this process is in Section 5.

C Datasets and Models

This section describes the datasets and models for our experiments. In detail, it describes the details about the datasets, *e.g.,* the dataset size, and the number of classes. It also explains the list of models for each dataset and how we obtain the models.

Datasets. We perform experiments on three datasets: CIFAR10/100 (Krizhevsky et al., 2009), ImageNet (Deng et al., 2009). CIFAR10/100 contain 10/100 classes images whose dimension is a $32 \times 32 \times 3$. Its original dataset size is 50,000/10,0000 for training/test, and 5,000 images are sampled from the training set as a validation set. ImageNet has 1,000 classes images of $224 \times 224 \times 3$. The original dataset size is 1.3M/50,000 for training/validation, and 25,000 images are sampled from the validation set as a test set.

Models. We investigate various models for each dataset. For CIFAR10/100, we use DenseNet40 (Huang et al., 2017), LeNet5 (LeCun et al., 1998), ResNet110 (He et al., 2016), ResNet110 SD (Huang et al., 2016), and WRN-28-10 (Zagoruyko and Komodakis, 2016). We acquire codes for DenseNet40 from a github repository (Veit et al., 2017), ResNet110 and WRN-28-10 from a github repository (Yang et al., 2019), and implement other models. For ImageNet, we use DenseNet161 (Huang et al., 2017) and ResNet152 (He et al., 2016), obtained from PyTorch.

D Additional Results

In this section, we display more results for Section 4 and Section 6. For Section 4, we present the detail ECE and number of images of each group for each transformation type and parameter, and for Section 6, we show the test error rate and learning time of a calibration map.

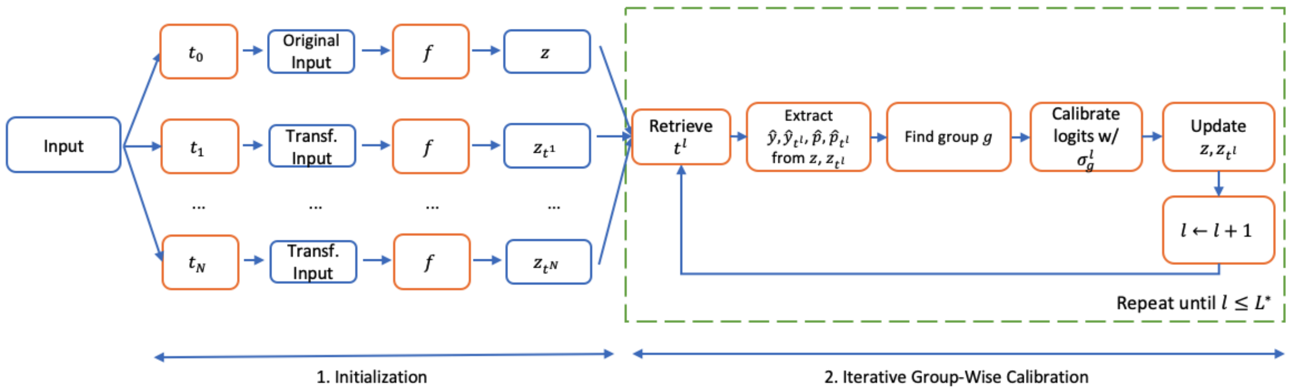

Figure 4: Runtime Confidence Calculation using ReCal

D.1 Detail Result for Lossy Label-Invariant Grouping

Table 4 and 5 show ECE and number of images of each group for different parameters using zoom-out transformation and brightness transformation. The second row represents ECE and number of images of test data, and following rows show the ECE and the number of images of each group for each transformation parameter. The bold and italic numbers mean the best and worst result among four groups, respectively.

As shown in Table 4, with zoom-out transformation, group 4 has the best ECE, *i.e.,* this group requires less adjustment and group 1 has the worst ECE *i.e.,* this group requires more adjustment, and group 2 and 3 have the medium range of ECE, for the most of transformation parameters except 0.1x and 0.2x. The similar pattern with more variability is observed with brightness transformation as displayed in Table 5. With brightness transformation, group 4 has the best ECE for the transformation less than 0.8x, and group 1 has the worst ECE for the transformation less than 0.5x. Unlike zoom-out transformation, group 2 also have the worst ECE for about the half of transformation parameters. However, each group still show different ECE compared to other groups, which supports our idea of group-wise calibration.

Table 4 and 5 show the number of images in each group. For different parameters, the image distribution over groups are different, and zoom-out transformation shows more variability than brightness transformation. Based on this observation, we design an ReCal which can incorporate multiple parameters as described in Section 5. Lastly, for zoom-out transformation with the scale factor of 0.1x, group 3 has only three images. With this small amount of images, calibration can overfit the data, and we address this issue as described in Section 5.

D.2 Additional Comparison Results

Besides ECE and Brier Score, we also compare test error rate and learning time of a calibration map. Table 6 and 7 show the test error rate, and the learning time, respectively. For tables, bold numbers mean the best results and underlined numbers represent the second-best results.

Error rate. We calculate test error rate to compare the accuracy preserving properties. As shown in Table 6, temperature scaling (TS), ReCal do not change the original accuracy, *i.e.,* their error rate is equal to an uncalibrated classifier's one. However, vector scaling (VS), MS-ODIR, and Dir-ODIR change the original accuracy. Without a consistent pattern, all of those calibration algorithms increase or decrease the error rate depending on the dataset and model. In detail, vector scaling decreases the original classifier's accuracy except the ImageNet experiments. MS-ODIR hurts the original classifier's accuracy except for DenseNet40 and ResNet110 SD on CIFAR10 and DenseNet161, ResNet152 on ImageNet. Dir-ODIR worsen the original classifier's accuracy except for LeNet5 on CIFAR10, LeNet5/Resnet110/ResNet110 SD on CIFAR100.

Learning Time. We display the learning time of a calibration map for each algorithm in Table 7. Temperature scaling (TS) is always the fastest calibration algorithm followed by vector scaling (VS). The next fastest one is ReCal, and we think ReCal can be applied to ImageNet in terms of the learning time. Specifically, it takes 50,730 seconds or 14.1 hours for DenseNet161 on ImageNet and 71,254 seconds or 19.8 hours for ResNet152 on ImageNet. On the other hand, MS-ODIR and Dir-ODIR are slower than other calibration algorithms because it basically calibrate many models to find appropriate its hyper-parameters.

Table 4: Grouping Image Using Zoom-Out Transformation

		ECE		Count	
	Test Data	0.020069		25000	
		Incr.	Not Incr.	Incr.	Not Incr.
0.9x	Change	*0.047142*	0.040512	1578	2328
	No Change	0.025389	**0.020825**	8250	12844
0.8x	Change	*0.048417*	0.033766	1830	3084
	No Change	0.025770	**0.020072**	7038	13048
0.7x	Change	*0.036925*	0.034266	1938	4149
	No Change	0.029248	**0.019347**	6109	12804
0.67x	Change	*0.040322*	0.032636	1990	4690
	No Change	0.028604	**0.018290**	5507	12813
0.5x	Change	*0.044078*	0.026399	2272	8444
	No Change	0.028710	**0.016006**	3744	10540
0.4x	Change	*0.044421*	0.022210	2041	12630
	No Change	0.034228	**0.016316**	2096	8233
0.33x	Change	*0.054198*	0.019901	1837	16635
	No Change	0.038278	**0.017077**	1067	5461
0.2x	Change	0.073773	**0.019371**	791	23446
	No Change	*0.205023*	0.030856	58	705
0.1x	Change	0.069130	**0.019659**	344	24596
	No Change	*0.161533*	0.098585	3	57

Table 5: Grouping Image Using Brightness

		ECE		Count	
	Test Data	0.020069		25000	
		Incr.	Not Incr.	Incr.	Not Incr.
0.9x	Change	0.066956	*0.077188*	315	356
	No Change	**0.021815**	0.022545	11404	12925
0.8x	Change	0.059497	*0.073143*	556	625
	No Change	**0.023280**	0.023908	11020	12799
0.7x	Change	0.041225	*0.052961*	750	951
	No Change	0.025298	**0.022240**	10623	12676
0.67x	Change	0.039053	*0.061532*	817	1060
	No Change	0.024897	**0.022962**	10437	12686
0.5x	Change	0.046058	*0.053280*	1172	1583
	No Change	0.026930	**0.022807**	9355	12890
0.4x	Change	*0.045266*	0.040179	1332	2109
	No Change	0.027172	**0.022091**	8462	13097
0.33x	Change	*0.048402*	0.042858	1446	2572
	No Change	0.025383	**0.023382**	7676	13306
0.2x	Change	*0.052800*	0.035715	1578	4337
	No Change	0.027635	**0.017540**	5394	13691
0.1x	Change	*0.040800*	0.029543	1198	9341
	No Change	0.034183	**0.016133**	2360	12101

Table 6: Test Error Rate (%)

Dataset	Model	Uncal.	TS	VS	MS-ODIR	Dir-ODIR	ReCal (z, .1-.9, 20)	ReCal (z, .5-.9, 10)	ReCal (b, .1-.9, 20)
CIFAR10	DenseNet40	8.25	8.25	8.31	**8.22**	8.31	8.25	8.25	8.25
CIFAR10	LeNet5	27.23	27.23	27.33	27.24	**27.20**	27.23	27.23	27.23
CIFAR10	ResNet110	**6.90**	**6.90**	7.06	7.06	7.03	**6.90**	**6.90**	**6.90**
CIFAR10	ResNet110 SD	9.62	9.62	9.76	**9.59**	9.64	9.62	9.62	9.62
CIFAR10	WRN 28-10	**4.06**	**4.06**	4.10	4.13	4.10	**4.06**	**4.06**	**4.06**
CIFAR100	DenseNet40	**31.84**	**31.84**	32.27	32.00	31.89	**31.84**	**31.84**	**31.84**
CIFAR100	LeNet5	62.34	62.34	62.66	62.58	**62.22**	62.34	62.34	62.34
CIFAR100	ResNet110	30.48	30.48	30.94	30.80	**30.46**	30.48	30.48	30.48
CIFAR100	ResNet110 SD	29.90	29.90	29.98	29.91	**29.89**	29.90	29.90	29.90
CIFAR100	WRN 28-10	**20.10**	**20.10**	20.29	20.47	20.51	**20.10**	**20.10**	**20.10**
ImageNet	DenseNet161	22.55	22.55	22.49	**22.10**	23.07	22.55	22.55	22.55
ImageNet	ResNet152	21.31	21.31	21.22	**20.96**	21.63	21.31	21.31	21.31

Table 7: Learning Time (sec)

Dataset	Model	TS	VS	MS-ODIR	Dir-ODIR	ReCal (z, .1-.9, 20)
CIFAR10	DenseNet40	**2.94**	31.10	77353.63	43001.99	84.04
CIFAR10	LeNet5	**1.86**	12.06	42830.58	37001.63	110.79
CIFAR10	ResNet110	**2.21**	26.65	70702.87	45836.87	38.85
CIFAR10	ResNet110 SD	**4.35**	26.52	85859.16	54783.42	58.74
CIFAR10	WRN 28-10	**7.68**	28.22	67955.20	36386.26	49.62
CIFAR100	DenseNet40	**14.03**	26.31	320284.77	134317.54	136.23
CIFAR100	LeNet5	**9.63**	26.10	109645.75	83324.48	97.77
CIFAR100	ResNet110	**8.63**	26.61	300360.19	134317.54	97.29
CIFAR100	ResNet110 SD	**13.24**	26.73	276767.31	126100.97	604.12
CIFAR100	WRN 28-10	**14.23**	25.60	161327.35	85532.50	125.84
ImageNet	DenseNet161	**865.40**	285.73	379487.45	276553.98	50730.17
ImageNet	ResNet152	**754.51**	342.50	215746.16	229493.41	71254.34