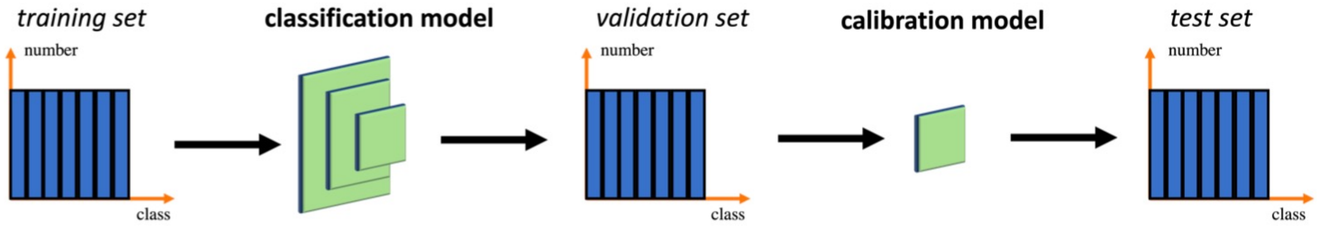
Transfer Knowledge from Head to Tail: Uncertainty Calibration under Long-tailed Distribution

Jiahao Chen, Bing Su

Gaoling School of Artificial Intelligence, Renmin University of China
Beijing Key Laboratory of Big Data Management and Analysis Methods

Tag:THU-AM-332

Overview



(a) Calibration under balanced distribution.

(b) Calibration under long-tailed distribution.

Training set: balanced
Validation set: balanced
Test set: balanced

Training set: long-tailed
Validation set: long-tailed
Test set: balanced

We investigate the problem of *calibration under long-tailed distribution*.

Background

Calibration:
If the Eq.1 is satisfied, the model is called perfect calibrated

$$\mathbb{P}(\hat{y}_i = y_i | \hat{p}_i = p) = p \qquad \forall p \in [0,1] \qquad (1)$$

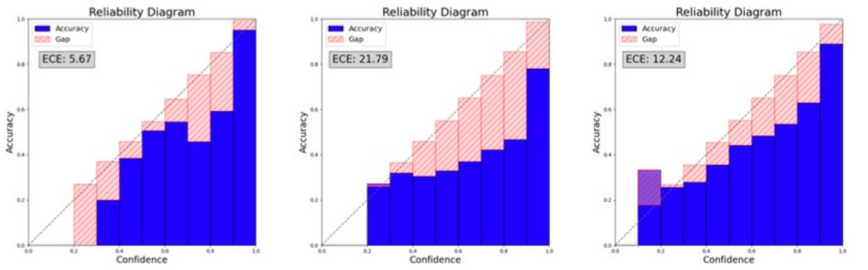
$$\hat{p}_i = \max softmax(\boldsymbol{z}_i) \quad \hat{y}_i = \underset{\{1,2,\cdots,C\}}{\arg\max}\; softmax(\boldsymbol{z}_i)$$

For example, 20% of all predictions with a confidence score of 80% should be false.

Temperature scaling[1]:

$$T^* = \underset{T}{\arg\min}\, \mathbb{E}_p[\mathcal{L}(s(\boldsymbol{z}_i/T), y_i)]$$

Validation set and test set should be in the same distribution.
Not satisfied long-tailed calibration.



(a) Validation set (b) Test set (c) Temperature scaling

[1] *ChuanGuo, GeoffPleiss, YuSun, and Kilian Q Weinberger. On calibration of modern neural networks. In International conference on machine learning 2017*

Importance weight

Importance weight-based method

Source distribution $p(x)$ and target distribution $q(x)$

$$\begin{aligned}\mathbb{E}_q[\mathcal{L}(s(\boldsymbol{z}_i/T), y_i)] &= \int_q q(\boldsymbol{x}_i)\mathcal{L}(s(\boldsymbol{z}_i/T), y_i)dx \\ &= \int_p \frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} p(\boldsymbol{x}_i)\mathcal{L}(s(\boldsymbol{z}_i/T), y_i)dx \\ &= \mathbb{E}_p[w(\boldsymbol{x}_i)\mathcal{L}(s(\boldsymbol{z}_i/T), y_i)]\end{aligned}$$

If we know the probability of each sample in the source distribution (*long-tailed distribution*), and in the target distribution (*balanced distribution*), we can achieve long-tailed calibration.

Q： how to acquire the probability under the target distribution?

Our method: Utilize the knowledge from head class to estimate importance weight $w(x)$

Long-tailed calibration

Five steps to realize calibration

step1: Estimate the feature distribution of each class.

step2: Calculate attentions between head and tail classes.

$$\boldsymbol{d}_c^k = Wasserstein(p_c(\boldsymbol{x}), p_k(\boldsymbol{x}))$$

$$\boldsymbol{s}_c = softmax(-\frac{\boldsymbol{d}_c}{\sqrt{dim(\boldsymbol{f})}})$$

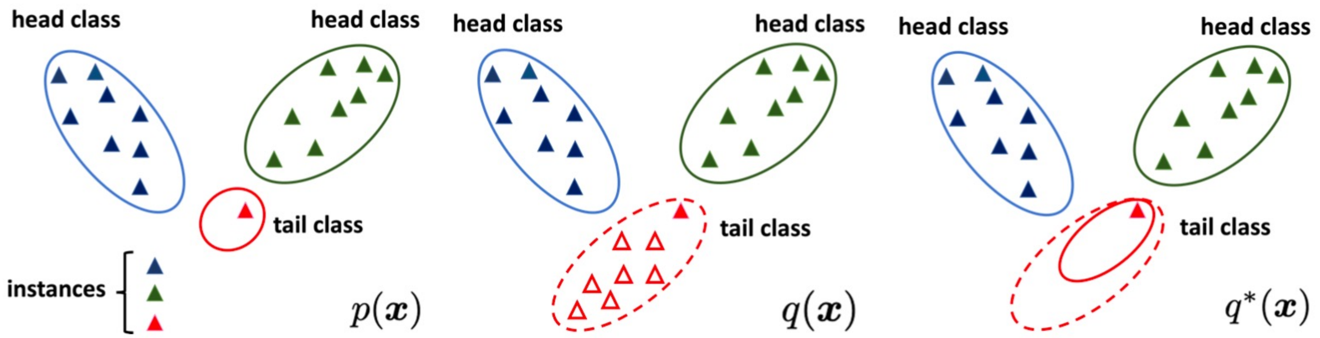
step3: Estimate the calibrated probability function.

$$\boldsymbol{\mu}_{c^*} = \alpha\boldsymbol{\mu}_c + (1-\alpha)\sum_{k\in\mathcal{A}_{head}} \boldsymbol{s}_c^k\boldsymbol{\mu}_k$$

$$\sqrt{\boldsymbol{\Sigma}}_{c^*} = \alpha\sqrt{\boldsymbol{\Sigma}}_c + (1-\alpha)\sum_{k\in\mathcal{A}_{head}} \boldsymbol{s}_c^k\sqrt{\boldsymbol{\Sigma}}_k$$

step4: Estimate the importance weight

$$w^*(\boldsymbol{x}_i) = \begin{cases} 1 & y_i \in \mathcal{A}_{head} \\ min(max(\frac{q^*_{y_i}(\boldsymbol{x}_i)}{p_{y_i}(\boldsymbol{x}_i)}, \eta_1), \eta_2) & y_i \in \mathcal{A}_{tail} \end{cases}$$



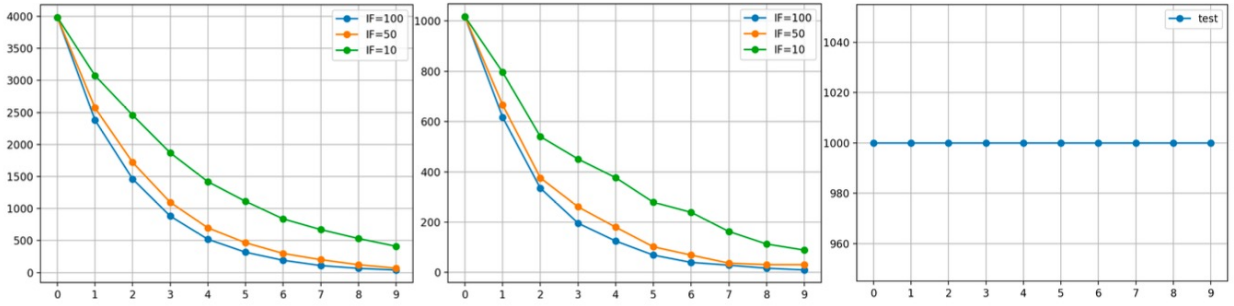
(a) Long-tailed distribution (b) Balanced distribution (c) Our estimated distribution

step5: Learn the temperature with the importance weights.

$$T^* = \arg\min_T \mathbb{E}_p[w^*(\boldsymbol{x}_i)\mathcal{L}(s(\boldsymbol{z}_i/T), y_i)]$$

Experiments

- CIFAR-10-LT
  - Original CIFAR-10
  - CIFAR-10.1
  - CIFAR-10.1-C
  - CIFAR-F
- MNIST-LT
  - Original MNIST
  - SVHN
  - USPS
  - Digital-S
- CIFAR-100-LT
  - Original CIFAR-100
- ImageNet-LT
  - Balanced Test set



(a) Training set (b) Validation set (c) Test set

We constitute the long-tailed training set, validation set and balnced Test set to verify our method.

Experiments

- CIFAR-10-LT results

IF	Dataset	Method								
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours
IF=100	CIFAR-10	21.79	12.24	12.16	11.64	12.36	13.36	12.13	11.65	**9.84**
	CIFAR-10.1	28.97	16.75	16.70	16.65	17.13	17.93	16.78	15.71	**13.86**
	CIFAR-10.1-C	58.22	43.01	43.00	43.05	43.34	43.83	42.53	41.98	**39.58**
	CIFAR-F	29.22	15.27	15.24	15.52	15.75	16.23	15.45	14.18	**12.15**
IF=50	CIFAR-10	17.36	7.65	8.04	8.22	9.75	9.45	7.55	7.78	**3.99**
	CIFAR-10.1	22.79	10.36	10.99	11.72	13.35	12.70	10.32	10.82	**5.74**
	CIFAR-10.1-C	55.52	38.66	39.9	40.16	41.58	40.76	38.94	39.39	**33.09**
	CIFAR-F	25.37	11.30	12.21	12.67	14.39	13.37	11.4	11.76	**6.64**
IF=10	CIFAR-10	8.39	2.23	1.64	2.03	2.29	2.42	2.49	2.01	**1.00**
	CIFAR-10.1	13.80	4.87	4.25	4.54	5.38	5.23	5.63	4.66	**3.95**
	CIFAR-10.1-C	48.31	32.77	31.07	32.11	32.29	31.94	33.16	31.37	**29.98**
	CIFAR-F	19.73	8.15	6.80	8.42	8.97	8.13	8.54	7.10	**5.97**

- CIFAR-100-LT results

Model	Dataset	Method								
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours
ResNet-32	CIFAR-100	20.38	2.50	2.10	6.07	9.35	5.92	6.74	3.27	**1.50**
DenseNet-40	CIFAR-100	16.00	3.43	2.51	5.57	8.42	5.76	5.96	2.73	**2.37**
VGG-19	CIFAR-100	27.86	3.81	2.36	6.35	10.35	6.66	8.03	3.82	**1.99**

- MNIST-LT results

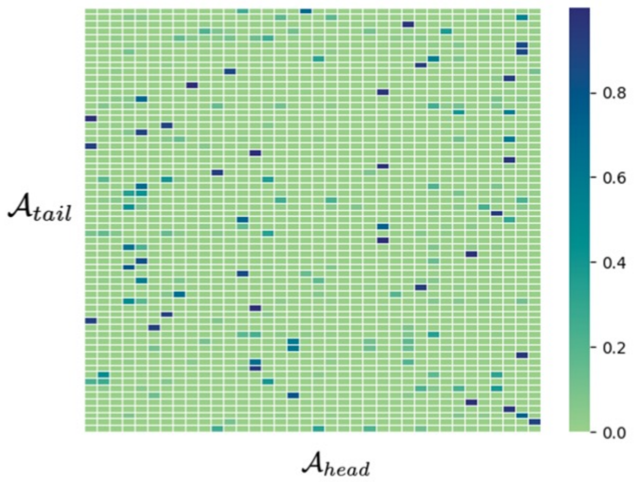
IF	Dataset	Method								
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours
IF=100	MNIST	2.52	1.27	1.84	2.82	2.84	1.84	1.92	1.76	**1.08**
	SVHN	16.06	7.20	11.62	21.25	22.18	14.93	9.59	13.67	**6.09**
	USPS	15.00	9.52	12.25	13.25	13.62	10.58	10.10	11.44	**8.40**
	Digital-S	32.10	22.13	27.35	30.13	31.01	27.48	23.34	27.60	**20.28**
IF=50	MNIST	1.12	0.85	1.14	1.53	1.54	1.02	1.01	1.12	**0.79**
	SVHN	**2.32**	3.95	3.33	11.42	12.15	2.63	9.43	**2.32**	4.53
	USPS	11.21	8.14	12.81	11.89	11.91	10.54	8.57	11.21	**8.02**
	Digital-S	15.22	10.81	17.81	20.96	21.81	13.64	16.74	15.18	**10.34**
IF=10	MNIST	0.56	0.23	**0.21**	0.50	0.52	0.23	0.25	0.41	0.36
	SVHN	5.75	6.76	6.94	8.10	**4.51**	5.31	7.00	5.31	7.43
	USPS	8.29	4.81	4.60	6.59	6.98	4.76	5.12	5.88	**4.55**
	Digital-S	13.55	8.21	8.09	15.37	13.34	8.31	7.67	8.24	**7.37**

- ImageNet-LT results

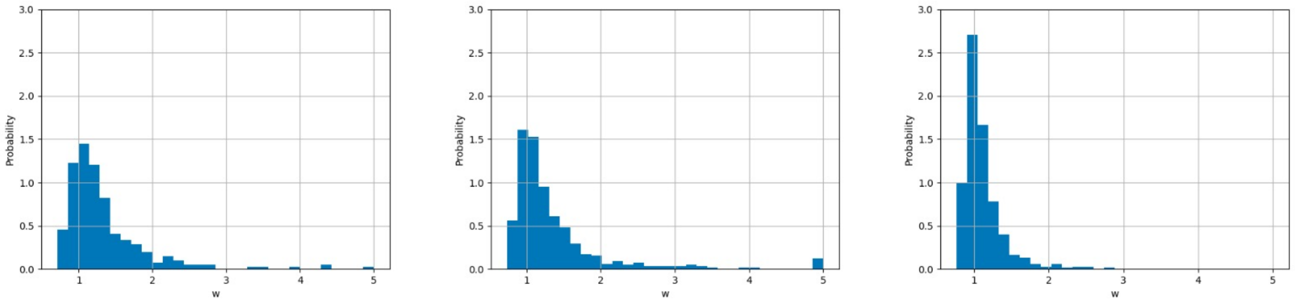
Model	Dataset	Method								
		Base	TS	ETS	TS-IR	IR	IROvA	SBC	GPC	Ours
ResNet-50	ImageNet	10.18	6.72	6.06	10.23	11.15	7.63	9.12	5.46	**3.45**

Experiments

- Visualization of the attention map

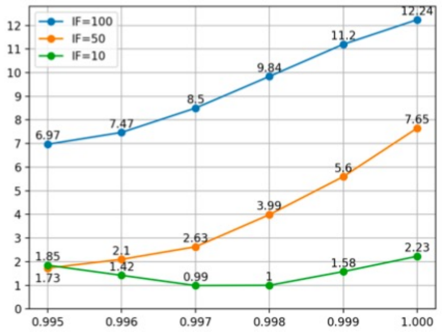


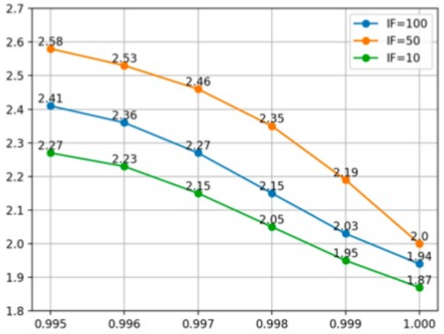
- Visualization of the distribution of $w(x)$



(a) IF=100 (b) IF=50 (c) IF=10

- Ablation of the hyper-parameter λ





- Ablation study of the different transferring strategies

Dataset	Uniform	OneHot	Ours
CIFAR-10	8.10	7.42	6.97
CIFAR-10.1	11.72	10.91	10.4
CIFAR-10.1-C	37.09	36.17	35.59
CIFAR-F	9.91	9.02	8.46

Conclusion

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Our contribution:
  - We explore the problem of calibration under long-tailed distribution, which has important practical implications but is rarely studied. We apply the importance weight strategy to enhance the estimation of tail classes for more accurate calibration.
  - We propose an importance weight estimation method by viewing distributions of head classes as prior for distributions of tail classes. For each tail class, our method estimates its probability density function from the distribution calibrated by head classes and calculates the importance weight to realize balanced calibration.
  - We conduct extensive experiments on the CIFAR-10-LT, CIFAR-100-LT , MNIST- LT, ImageNet-LT datasets and the results demonstrate the effectiveness of our method.

THANKS