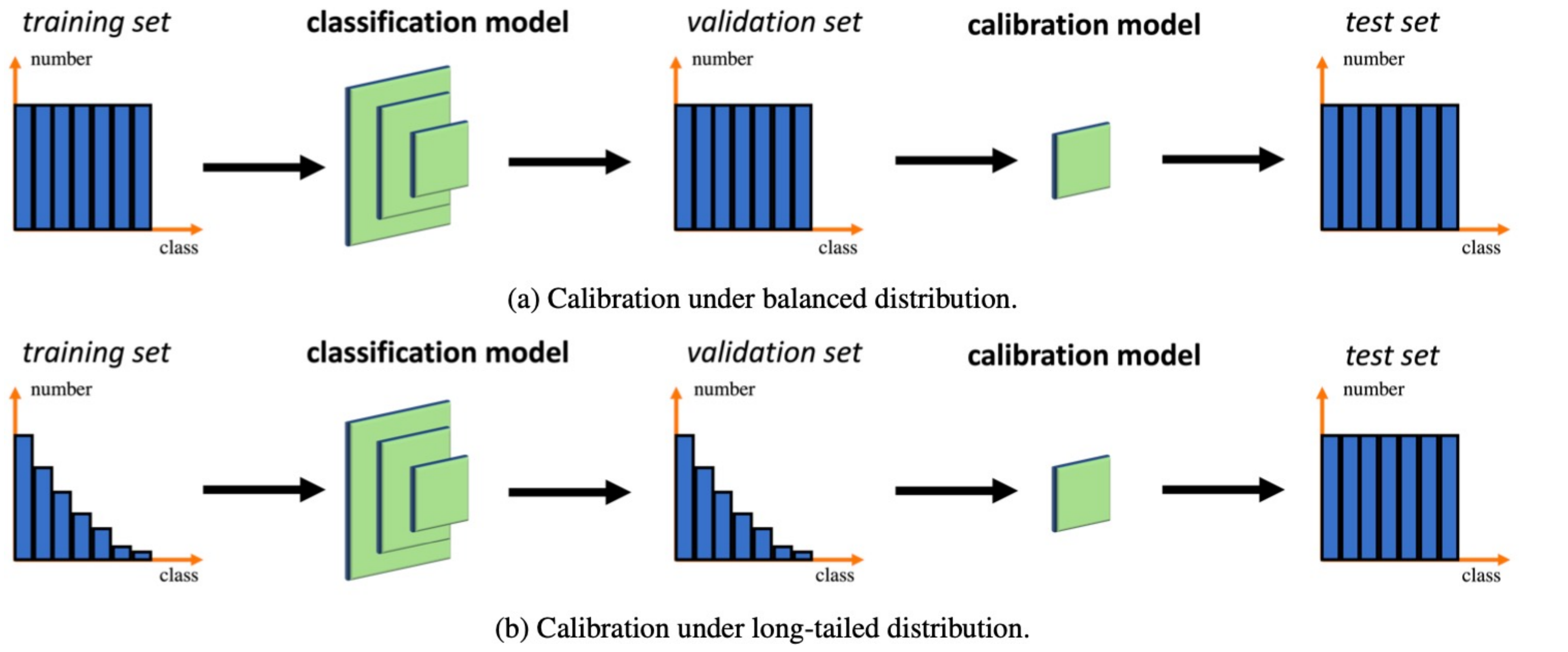
Transfer Knowledge from Head to Tail: Uncertainty Calibration under Long-tailed Distribution

Jiahao Chen, Bing Su†
Gaoling School of Artificial Intelligence, Renmin University of China

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation:

Problem: calibration under long-tailed distribution.
Training set: Long-tailed distribution.
Validation set: Long-tailed distribution.
Test set: Balanced distribution.



Approach:

The Optimization target of temperature scaling:

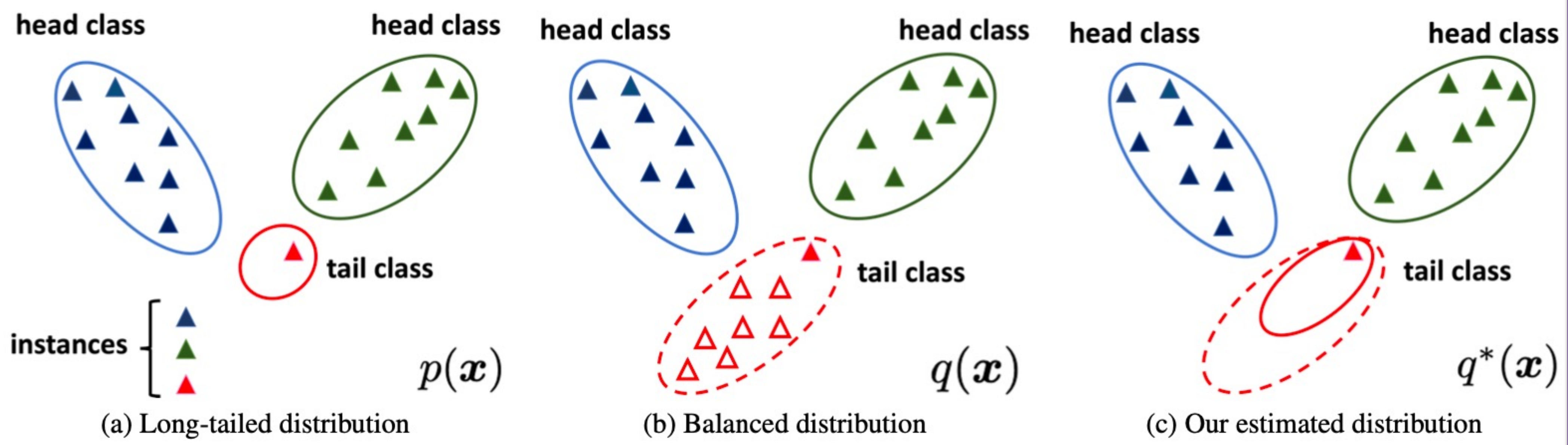
$$T^* = \arg\min_T \mathbb{E}_p[\mathcal{L}(s(\boldsymbol{z}_i/T), y_i)]$$

Ensure: validation set and Test set in the same distribution.

Not satisfied under the long-tailed distribution.

Considering importance-weight based method:

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



Question: how to realize calibration?

Key idea: transfer knowledge from head to tail

Step 1: Estimate the feature distribution of each class.

Step 2: Calculate the attention between head and tail classes.

$$\boldsymbol{d}_c^k = Wasserstein(p_c(\boldsymbol{x}), p_k(\boldsymbol{x})) \quad \boldsymbol{s}_c = softmax(-\frac{\boldsymbol{d}_c}{\sqrt{dim(\boldsymbol{f})}})$$

Step 3: Estimate the calibrated probability function.

$$\left.\begin{aligned}\boldsymbol{\mu}_{c^*} &= \alpha\boldsymbol{\mu}_c + (1-\alpha)\sum_{k\in\mathcal{A}_{head}} \boldsymbol{s}_c^k\boldsymbol{\mu}_k \\ \sqrt{\boldsymbol{\Sigma}}_{c^*} &= \alpha\sqrt{\boldsymbol{\Sigma}}_c + (1-\alpha)\sum_{k\in\mathcal{A}_{head}} \boldsymbol{s}_c^k\sqrt{\boldsymbol{\Sigma}}_k\end{aligned}\right\} \mathcal{N}(\boldsymbol{\mu}_{c^*}, \boldsymbol{\Sigma}_{c^*})$$

Step 4: Estimate the importance weight.

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

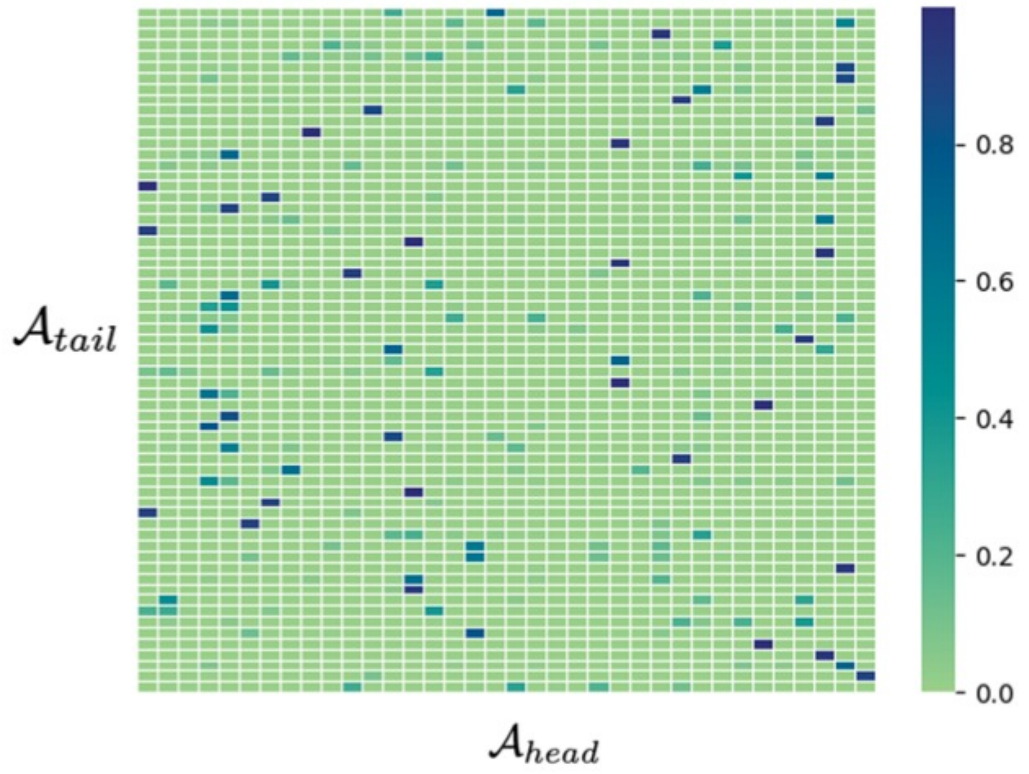
Step 5: Learn the temperature with the importance weights.

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

Experiment:

CIFAR-10-LT, with imbalanced factor 100, 50, 10.

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**



Visualization of attention.

Each row denotes the vector $\boldsymbol{s}$. It is shown that each tail class has more than one similar head classes and their knowledge will be transferred to corresponding tail class .

Conclusion:

We propose a novel importance weight-based strategy to achieve post-processing calibration under long-tailed distribution. The tackled problem differs from traditional calibration tasks as the validation set follows a long-tailed distribution, while the test data distribution is balanced. The importance weight strategy is used to re-weight instances of tail classes. We enhance the estimation of tail class distributions by transferring knowledge from head classes.