Long-Tailed Partial Label Learning via Dynamic Rebalancing

Feng Hong[1] Jiangchao Yao[1,2,†] Zhihan Zhou[1] Ya Zhang[1,2] Yanfeng Wang[1,2,†]

[1]CMIC, Shanghai Jiao Tong University [2]Shanghai AI Lab

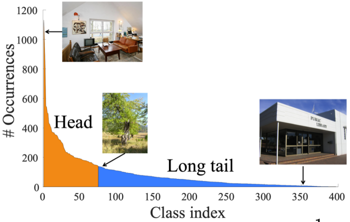
Annotation ambiguity

Real-world data naturally suffers from inherent label ambiguity.

Annotator 1: Korat
Annotator 2: Russian Blue

The History of LeBron James and Stephen Curry's Rivalrous Friendship

Partial Label Learning & A New Challenge



The long-tailed distribution[1]

- **Partial label learning (PLL)**
  - The annotation for each sample is an ambiguous set containing the groundtruth and other confusing labels.
- **The ideal assumption behind PLL**
  - The collected data is approximately uniformly distributed regarding classes.
  - However, Real-world natural sources usually follow the long-tailed law.
- **A new challenge: long-tailed partial label learning (LT-PLL)**
  - The algorithmic robustness to both category imbalance and label ambiguity.
  - Tail samples cannot be correctly recognized even in training.
  - No available class prior.

[1] Learning to Model the Tail. NeurIPS 2017.

Related Work - PLL

Partial label learning (PLL) and long-tailed learning (LT) independently study partial aspects of LT-PLL.

Partial label learning (PLL)

- Key challenge
  - Label disambiguation: detecting the groundtruth from the candidate label set
- Existing work
  - Average-based methods[1]
  - Graph-based methods[2]
  - Self-training methods[3]

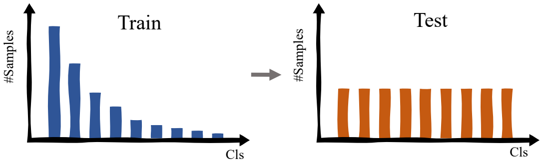
[1] Learning from partial labels. JMLR 2011.
[2] GM-PLL: graph matching based partial label learning. TKDE, 2021.
[3] Provably consistent partial-label learning. NeurIPS 2020.

Related Work - LT

Partial label learning (PLL) and long-tailed learning (LT) independently study partial aspects of LT-PLL.

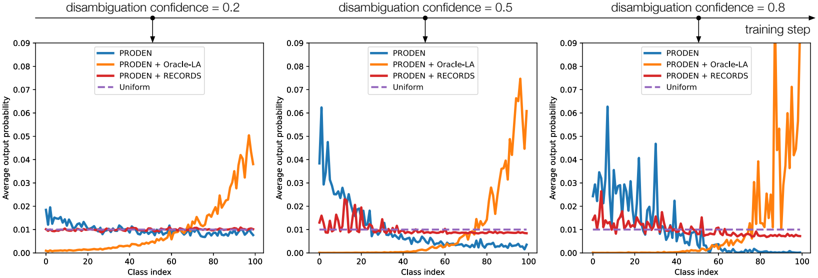
Long-tailed learning (LT)

- Key challenge
  - Rebalancing: learning a balanced model from imbalanced data
- Existing work
  - Re-sampling[1]
  - Re-weighting[2]
  - Transfer learning[3]
  - Logit adjustment[4]



[1]Decoupling representation and classifier for long-tailed recognition. ICLR 2020.
[2]Class-balanced loss based on effective number of samples. CVPR 2019.
[3]Feature transfer learning for face recognition with under-represented data. CVPR 2019.
[4]Long-tail learning via logit adjustment. ICLR 2021.

Motivation



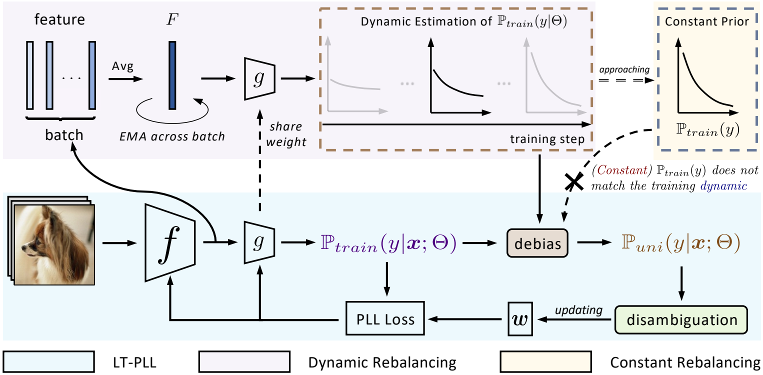
Observation: The prediction imbalance of PLL baselines (blue curve) is not significant at the early stage and gradually increases with the label disambiguation.

Constant rebalancing (orange curve) (LA: $z^y_{uni}(\boldsymbol{x}) = z^y(\boldsymbol{x}) - \log \mathbb{P}_{train}(y)$):

1. No available class prior $\mathbb{P}_{train}(y)$. 2. It does not consider the dynamics of label disambiguation and leads to failure.

Dynamic rebalancing (red curve): A dynamic rebalancing method that considers the label disambiguation process can be intuitively more effective.

RECORDS: Rebalancing for Dynamic Bias



Dynamic rebalancing adapted to model training is more friendly to weakly supervised paradigms like PLL.

RECORDS can be easily plugged into the current PLL methods in an end-to-end manner.

RECORDS

- Dynamic rebalancing paradigm

$$\begin{aligned}\mathbb{P}_{uni}(y|\boldsymbol{x};\Theta) &\propto \mathbb{P}(\boldsymbol{x}|y;\Theta)\cdot\mathbb{P}_{train}(y|\Theta)\ /\ \mathbb{P}_{train}(y|\Theta)\\ &\propto \mathbb{P}_{train}(y|\boldsymbol{x};\Theta)\ /\ \mathbb{P}_{train}(y|\Theta)\\ &\propto \mathrm{softmax}(z^y(\boldsymbol{x}) - \log\mathbb{P}_{train}(y|\Theta)),\end{aligned}$$

- NWGM approximation & momentum updates

$$\begin{aligned}\mathbb{P}_{train}(y|\Theta) = \mathbb{E}_{\boldsymbol{x}_i\in\mathcal{D}_{train}}\mathrm{softmax}(z^y(\boldsymbol{x}_i)) &\overset{NWGM}{\approx} \mathrm{softmax}(\mathbb{E}_{\boldsymbol{x}_i\in\mathcal{D}_{train}}z^y(\boldsymbol{x}_i))\\ &= \mathrm{softmax}(g^y(\mathbb{E}_{\boldsymbol{x}_i\in\mathcal{D}_{train}}f(\boldsymbol{x}_i;\theta);\boldsymbol{W})).\end{aligned}$$

$$F \leftarrow mF + (1-m)\mathbb{E}_{\boldsymbol{x}_i\in Batch}f(\boldsymbol{x}_i;\theta).$$

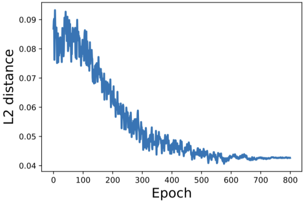
- Final implementation

$$z^y_{uni}(\boldsymbol{x}) = z^y(\boldsymbol{x}) - \log\mathrm{softmax}(g^y(F;\boldsymbol{W})).$$

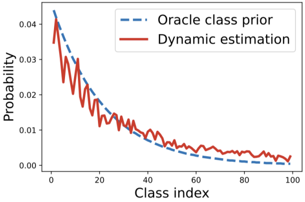
Relation between Dynamic & Constant Rebalancing

Proposition

Let $\tilde{h} = h_{\tilde{\Theta}}$ *be the optimal classifier on the basis of the label disambiguation. If the small ambiguity degree condition*[1] *satisfies, the* L_2 *distance between* $\mathbb{P}_{train}(y)$ *and* $\mathbb{P}_{train}(y|\tilde{\Theta})$ *given* $\tilde{h}$ *is bounded as* $L_2\left(\tilde{h}\right) < \frac{4}{(\ln 2 - \ln(1+\eta))N}(d_H(\ln 2N + 2\ln C) - \ln\delta + \ln 2)$ *with probability at least* $1-\delta$.



(a) L_2 distance during training



(b) Estimated class distribution

[1] Learning from partial labels. JMLR 2011.

Experiments

Main results

Table 1: Top-1 accuracy on three benchmark datasets. Bold indicates the superior results.

	CIFAR-10-LT						CIFAR-100-LT						PASCAL VOC
Imbalance ratio ρ	50			100			50			100			
Ambiguity q	0.3	0.5	0.7	0.3	0.5	0.7	0.03	0.05	0.07	0.03	0.05	0.07	
CORR	76.12	56.45	41.56	66.38	50.09	38.11	42.29	38.03	36.59	38.39	34.09	31.05	24.43
+ Oracle-LA post-hoc	80.70	58.49	43.44	72.96	54.64	41.66	46.94	40.76	39.07	41.49	36.79	33.32	34.12
+ Oracle-LA	36.27	17.61	12.77	29.97	15.80	11.75	22.56	5.59	3.12	11.37	3.32	1.98	52.51
+ RECORDS	**82.57**	**80.28**	**67.24**	**77.66**	**72.90**	**57.46**	**48.06**	**45.56**	**42.51**	**42.25**	**40.59**	**38.65**	**56.46**
vs. CORR	+6.45	+23.83	+25.68	+11.28	+22.81	+19.35	+5.77	+7.53	+5.92	+3.86	+6.40	+7.60	+32.03
PRODEN	73.12	54.45	41.37	63.55	47.37	38.06	39.23	35.45	33.90	34.52	32.04	29.40	22.39
+ Oracle-LA post-hoc	77.41	57.14	42.91	70.71	48.79	41.38	43.40	38.64	35.82	38.40	35.20	31.92	31.53
+ Oracle-LA	27.18	16.97	11.52	19.51	14.11	11.17	12.37	4.09	2.64	6.79	2.73	1.98	48.33
+ RECORDS	**79.48**	**76.73**	**65.31**	**72.15**	**65.22**	**52.26**	**44.56**	**41.31**	**39.26**	**39.13**	**37.23**	**35.26**	**52.65**
vs. PRODEN	+6.36	+22.28	+23.94	+8.60	+17.85	+14.2	+5.33	+5.86	+5.36	+4.61	+5.19	+5.86	+30.26
LW	70.11	37.67	22.73	64.78	39.57	23.54	35.54	29.50	27.86	31.58	28.09	24.65	19.41
+ Oracle-LA post-hoc	74.34	40.27	25.34	69.60	42.34	27.35	35.47	28.80	27.27	31.03	26.96	23.20	21.06
+ Oracle-LA	41.90	21.36	15.28	25.75	20.35	14.24	30.37	14.43	4.79	30.30	5.08	2.70	51.53
+ RECORDS	**76.02**	**57.39**	**40.28**	**71.18**	**57.23**	**41.24**	**36.56**	**31.67**	**29.39**	**33.00**	**28.85**	**25.64**	**53.09**
vs. LW	+5.91	+19.72	+17.55	+6.40	+17.66	+17.70	+1.02	+2.17	+1.53	+1.42	+0.76	+0.99	+33.68
CAVL	56.73	40.27	18.52	54.28	38.97	17.28	29.63	17.31	8.34	28.29	25.39	8.20	17.25
+ Oracle-LA post-hoc	55.23	39.76	18.34	51.37	37.28	14.58	29.65	14.86	5.76	28.34	26.27	5.80	22.27
+ Oracle-LA	22.16	14.97	11.50	18.29	14.23	10.67	17.31	4.36	2.83	7.24	2.55	2.03	50.78
+ RECORDS	**67.27**	**61.23**	**40.71**	**64.35**	**58.27**	**37.38**	**42.25**	**36.53**	**29.13**	**36.93**	**31.49**	**24.98**	**53.07**
vs. CAVL	+10.54	+20.96	+22.19	+10.07	+19.30	+20.1	+12.62	+19.22	+14.27	+8.64	+6.10	+16.78	+35.82

Experiments

Fine-grained analysis

Table 2: Fine-grained analysis on CIFAR-100-LT with $\rho = 100$ and $q \in \{0.03, 0.05, 0.07\}$. Many/Medium/Few corresponds to three partitions on the long-tailed data.

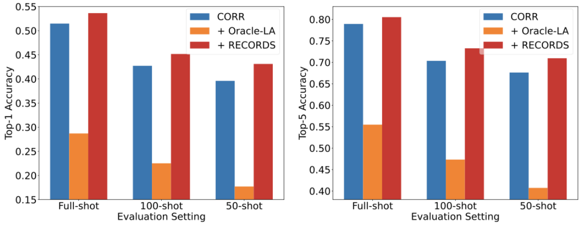
Method	$q = 0.03$				$q = 0.05$				$q = 0.07$			
	Many	Medium	Few	Overall	Many	Medium	Few	Overall	Many	Medium	Few	Overall
CORR	68.43	37.40	4.50	38.39	67.51	29.60	0.33	34.09	68.86	19.80	0.07	31.05
+ Oracle-LA post-hoc	70.37	41.89	7.33	41.49	70.46	33.40	1.47	36.79	69.77	24.86	0.67	33.32
+ Oracle-LA	11.03	12.34	10.63	11.37	0.34	4.46	5.47	3.32	0.00	0.71	5.77	1.98
+ RECORDS	66.37	42.54	13.77	**42.25**	68.49	40.20	8.50	**40.59**	69.97	36.71	4.37	**38.65**
vs. CORR	-2.06	+5.14	+9.27	+3.86	+0.98	+10.60	+8.17	+6.50	+1.11	+16.91	+4.30	+7.60

Experiments

Further analysis

Table 5: Comparison with other dynamic strategies on CIFAR-10-LT and CIFAR-100-LT.

	CIFAR-10-LT						CIFAR-100-LT					
Imbalance ratio ρ	50			100			50			100		
Ambiguity q	0.3	0.5	0.7	0.3	0.5	0.7	0.03	0.05	0.07	0.03	0.05	0.07
CORR	76.12	56.45	41.56	66.38	50.09	38.11	42.29	38.03	36.59	38.39	34.09	31.05
+ Temp Oracle-LA	81.37	43.62	18.10	76.09	25.88	16.11	47.44	43.46	29.75	41.78	39.19	33.69
+ Epoch RECORDS	75.43	70.27	59.50	69.38	63.12	47.85	46.54	43.07	38.28	41.58	37.14	34.38
+ RECORDS	**82.57**	**80.28**	**67.24**	**77.66**	**72.90**	**57.46**	**48.06**	**45.56**	**42.51**	**42.25**	**40.59**	**38.65**



(a) Linear Probing

Summary

- **Challenge:** We delve into the LT-PLL scenario, and identify its several challenges that cannot be addressed and even lead to failure by the straightforward combination of the current LT and PLL methods.
- **Methodology:** We propose a novel RECORDS for LT-PLL that conducts the dynamic adjustment to rebalance the training without requiring any prior about the class distribution.
- **Theoretical Understanding:** The theoretical and empirical analysis show that the dynamic parametric class distribution is asymmetrically approaching to the oracle class distribution but more friendly to label disambiguation.
- **Lightweight:** Our method is orthogonal to existing PLL methods and can be easily plugged into the current PLL methods in an end-to-end manner.

Codes: https://github.com/MediaBrain-SJTU/RECORDS-LTPLL