JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

THU-PM-322

Label Information Bottleneck for Label Enhancement

Qinghai Zheng, Jihua Zhu, Haoyu Tang

- Github codes: https://github.com/qinghai-zheng/LIBLE

Qinghai Zheng, Fuzhou University, Email： zhengqinghai@fzu.edu.cn, qhxjtu@163.com homepage： https://qinghai-zheng.github.io/

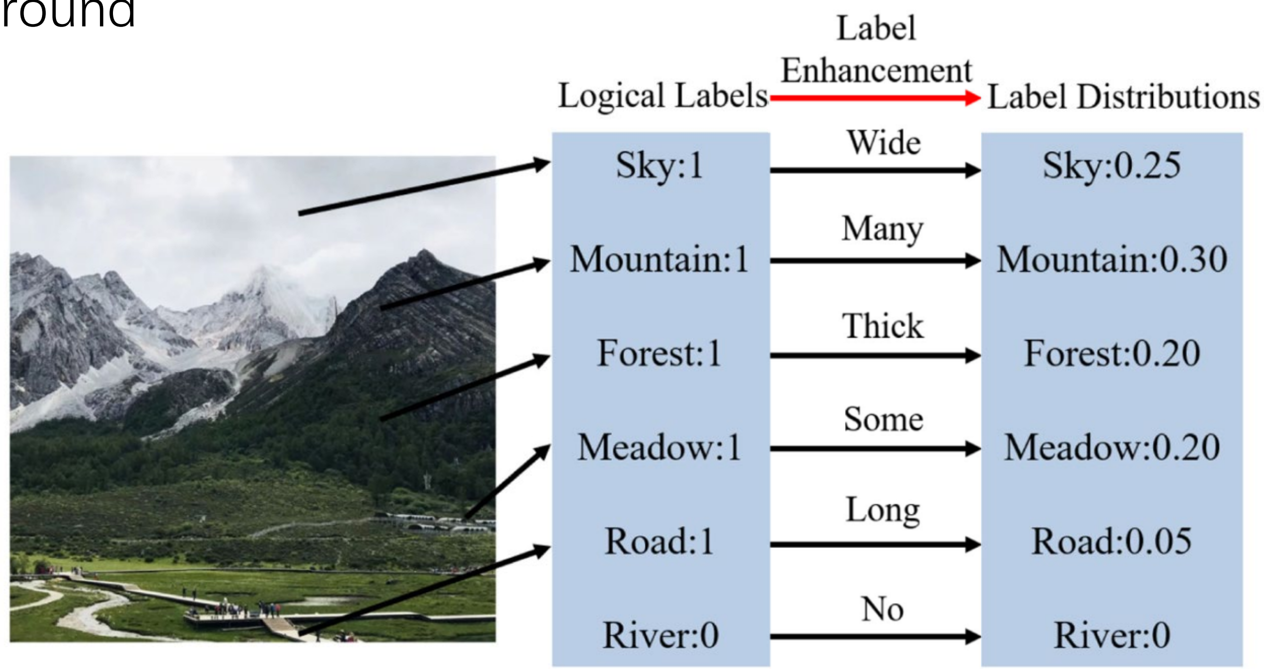
CONTENT

✓ Label Information Bottleneck for Label Enhancement

- Motivation
- Contributions
- The Proposed Framework
- Comparison with Existing LE Methods
- Experimental Results

Qinghai Zheng, Fuzhou University, Email: zhengqinghai@fzu.edu.cn, qhxjtu@163.com homepage: https://qinghai-zheng.github.io/

✓ Background



- It is unpractical to annotate data with label distributions manually.
- Most existing datasets in the field of computer vision and machine learning are annotated by single-label or multi-labels
- It is promising to recover the desired label distributions exactly from existing logical labels

✓ Existing LE Methods

- The objectives of most existing LE methods (GLLE, LESC, gLESC, and LEVI) can be concisely summarized as follows:

$$\min_{\theta} \ \|f_\theta(\boldsymbol{X}) - \boldsymbol{L}\|_F^2 + \gamma reg(f_\theta(\boldsymbol{X}))$$

1. GLLE: $reg(f_\theta(\boldsymbol{X})) = \sum q_{i,j} \|f_\theta(\boldsymbol{x}_i) - f_\theta(\boldsymbol{x}_j)\|_2^2$

2. LESC and gLESC: $reg(f_\theta(\boldsymbol{X})) = \|f_\theta(\boldsymbol{X}) - f_\theta(\boldsymbol{X})\boldsymbol{G}\|_F^2$

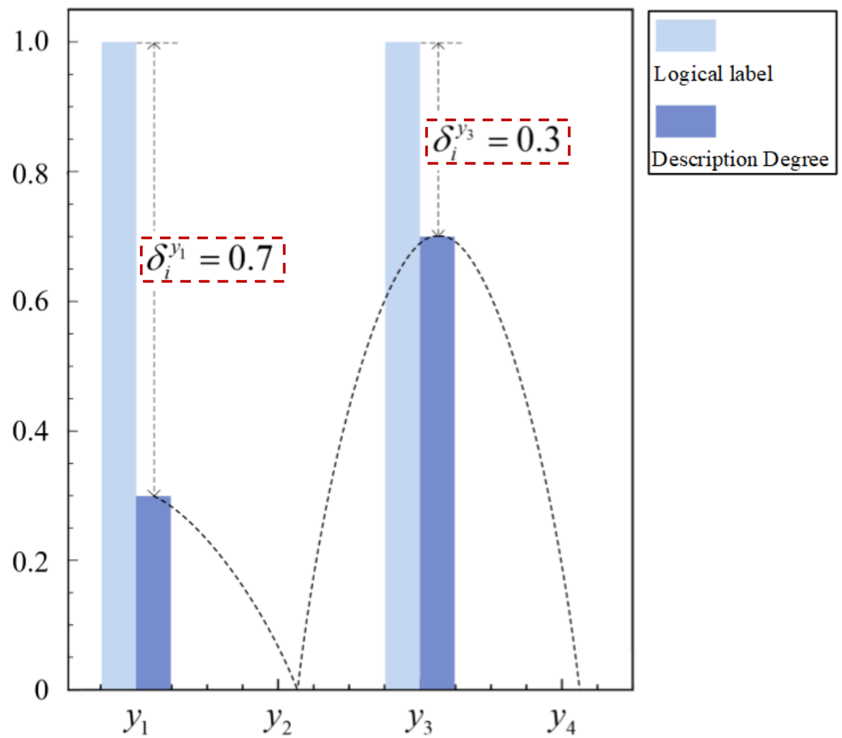
- Limitations:

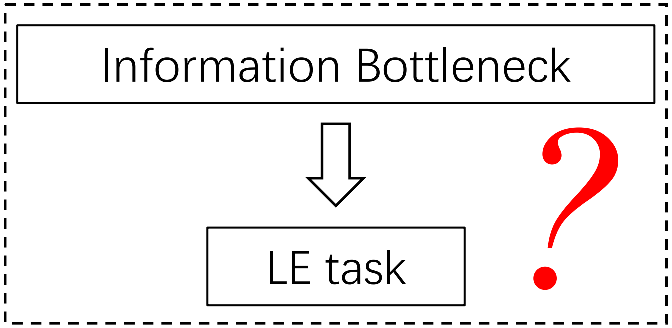
1. neglect the label irrelevant information contained in $\boldsymbol{X}$
2. Require the extra constraint: $\|\boldsymbol{d} - \boldsymbol{l}\|_2^2$

Qinghai Zheng, Fuzhou University, Email: zhengqinghai@fzu.edu.cn, qhxjtu@163.com homepage: https://qinghai-zheng.github.io/

✓ New idea for the task of LE

- We deal with the LE from the perspective of information theory
- We decompose the label relevant information into:
  1. the information about the assignments of labels to instance
  2. the information about the label gaps between logical labels and distribution labels



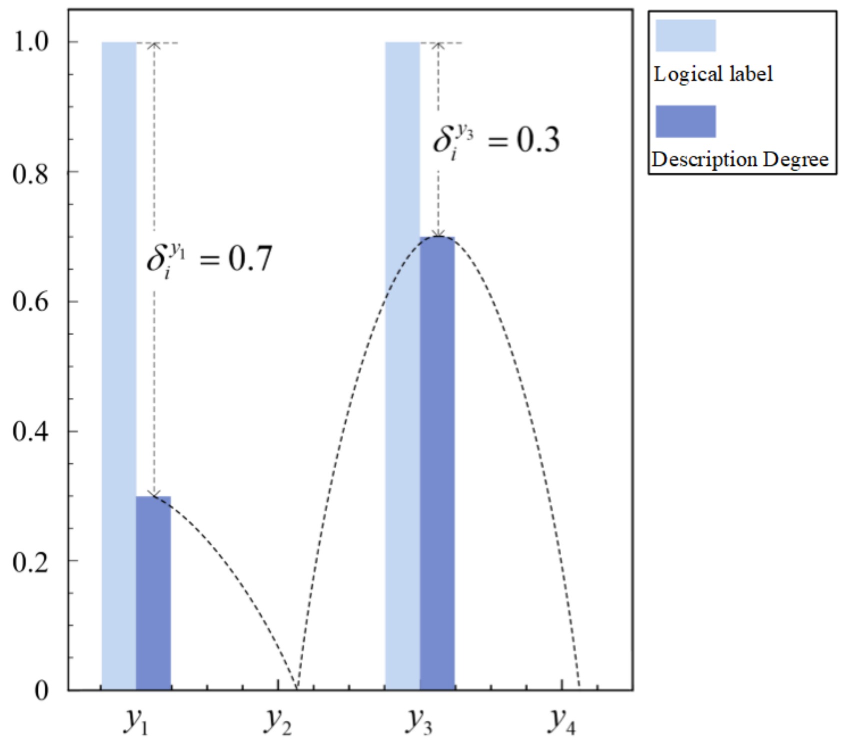


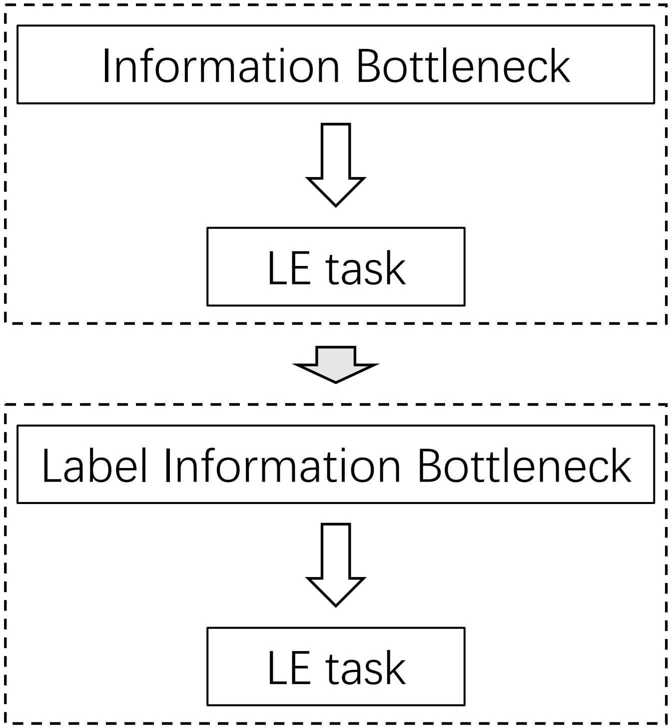
Qinghai Zheng, Fuzhou University, Email: zhengqinghai@fzu.edu.cn, qhxjtu@163.com homepage: https://qinghai-zheng.github.io/

CONTRIBUTIONS

✓ New idea for the task of LE

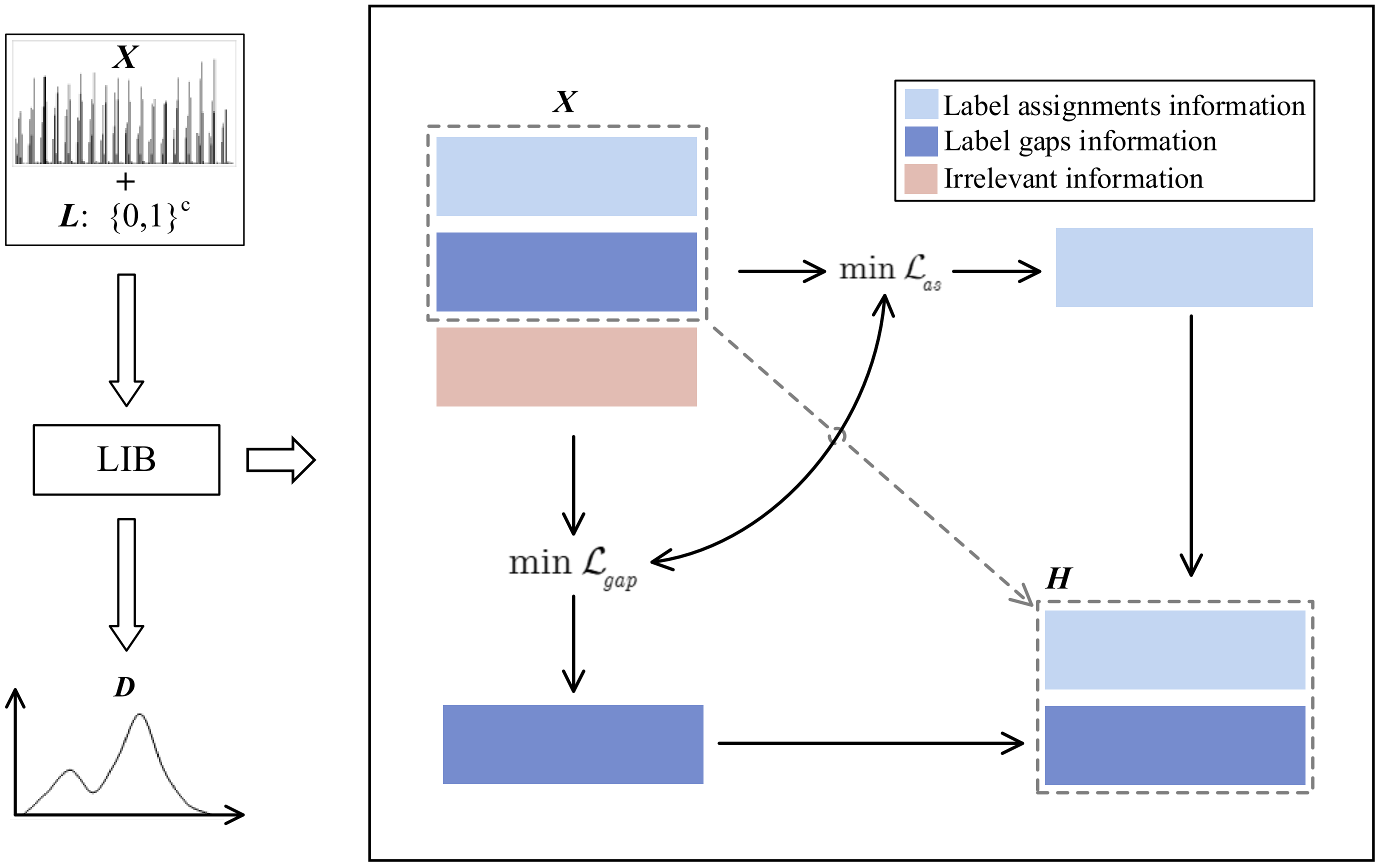
- We deal with the LE from the perspective of information theory
- We decompose the label relevant information into:
  1. the information about the assignments of labels to instance
  2. the information about the label gaps between logical labels and distribution labels





✓ Framework of LIB

- Fully consider the label relevant information during the LE process



✓ Objective of LIB

- LIB formulates the LE problem as the following two joint processes:
  1. Learn the representation with the label relevant information
  2. Recover label distributions based on the learned representation

$$\min_{\boldsymbol{H}} \mathcal{L}_{as} + \alpha\mathcal{L}_{gap}, \text{ s.t.}, I(\boldsymbol{X}, \boldsymbol{H}) \leqslant I_c$$

- $\mathcal{L}_{as}$: label assignmens information modeling.
  † It excavates the information about the assignments of labels to the instance

- $\mathcal{L}_{gap}$: label gaps information modeling.
  † It investigates the information about the label gaps between the logical labels and distribution labels

- $I(\boldsymbol{X}, \boldsymbol{H}) \leqslant I_c$: label irrelevant information modeling.

✓ Objective of LIB

- $\mathcal{L}_{as}$: $\mathcal{L}_{as} = -I(\boldsymbol{H}, \boldsymbol{L}) \Rightarrow \mathcal{L}_{as} = -\sum_{\boldsymbol{h}} \sum_{\boldsymbol{l}} p(\boldsymbol{h}, \boldsymbol{l}) \log \frac{p(\boldsymbol{l}|\boldsymbol{h})}{p(\boldsymbol{l})}$

$$\mathrm{KL}(p(\boldsymbol{l}|\boldsymbol{h})||q(\boldsymbol{l}|\boldsymbol{h})) = \sum_{\boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{h}) \log \frac{p(\boldsymbol{l}|\boldsymbol{h})}{q(\boldsymbol{l}|\boldsymbol{h})} \geqslant 0$$

$$\Rightarrow \sum_{\boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{h}) \log p(\boldsymbol{l}|\boldsymbol{h}) \geqslant \sum_{\boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{h}) \log q(\boldsymbol{l}|\boldsymbol{h}),$$

$$\mathbb{E}_{p(\boldsymbol{l})}[-\log p(\boldsymbol{l})] = -\sum_{\boldsymbol{l}} p(\boldsymbol{l}) \log p(\boldsymbol{l}) \geqslant 0,$$

1. KL divergence and the entropy are positive
2. Markov chain $L \leftarrow X \rightarrow H$

$$\mathcal{L}_{as} \leqslant -\sum_{\boldsymbol{x}} \sum_{\boldsymbol{l}} \sum_{\boldsymbol{h}} p(\boldsymbol{x}, \boldsymbol{l}) p(\boldsymbol{h}|\boldsymbol{x}) \log q(\boldsymbol{l}|\boldsymbol{h})$$

- $\mathcal{L}_{gap}$: $\mathcal{L}_{gap} = I(\boldsymbol{\Delta}|\boldsymbol{H}) = -\log p(\boldsymbol{\Delta}|\boldsymbol{H})$

$$= -\sum_{\boldsymbol{\delta}} \sum_{\boldsymbol{h}} \log p(\boldsymbol{\delta}|\boldsymbol{h})$$

$$= -\sum_{\boldsymbol{l}} \sum_{\boldsymbol{h}} \log p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h})$$

$$\max_{\boldsymbol{\Delta}} \log p(\boldsymbol{H}, \boldsymbol{\Delta})$$
$$\Rightarrow \max_{\boldsymbol{\Delta}} \log p(\boldsymbol{\Delta}|\boldsymbol{H}) + \log p(\boldsymbol{H})$$
$$\Rightarrow \max_{\boldsymbol{\Delta}} \log p(\boldsymbol{\Delta}|\boldsymbol{H}).$$

† We consider the conditional self-information here

† It can be also interpreted and derived from the view of the probability distribution

✓ Objective of LIB

- $I(\boldsymbol{X},\boldsymbol{H}) \leqslant I_c$: label irrelevant information modeling.

$$I(\boldsymbol{X},\boldsymbol{H}) = \sum_{\boldsymbol{x}}\sum_{\boldsymbol{h}} p(\boldsymbol{x},\boldsymbol{h}) \log \frac{p(\boldsymbol{h}|\boldsymbol{x})}{p(\boldsymbol{h})}$$

$$\Downarrow$$

$$\begin{aligned} I(\boldsymbol{X},\boldsymbol{H}) &\leqslant \sum_{\boldsymbol{x}}\sum_{\boldsymbol{h}} p(\boldsymbol{x},\boldsymbol{h}) \log \frac{p(\boldsymbol{h}|\boldsymbol{x})}{q(\boldsymbol{h})} \\ &= \sum_{\boldsymbol{x}}\sum_{\boldsymbol{l}} p(\boldsymbol{x},\boldsymbol{l}) \mathrm{KL}(p(\boldsymbol{h}|\boldsymbol{x})||q(\boldsymbol{h})) \end{aligned}$$

- The objective of LIB can be formulated as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{as} + \alpha\mathcal{L}_{gap} + \beta I(\boldsymbol{X},\boldsymbol{H}) \\ &\leqslant -\sum_{\boldsymbol{x}}\sum_{\boldsymbol{l}}\sum_{\boldsymbol{h}} p(\boldsymbol{x},\boldsymbol{l}) p(\boldsymbol{h}|\boldsymbol{x},\boldsymbol{l}) \log q(\boldsymbol{l}|\boldsymbol{h}) \\ &\quad - \alpha \sum_{\boldsymbol{l}}\sum_{\boldsymbol{h}} \log p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h}) \\ &\quad + \beta \sum_{\boldsymbol{x}}\sum_{\boldsymbol{l}} p(\boldsymbol{x},\boldsymbol{l}) \mathrm{KL}(p(\boldsymbol{h}|\boldsymbol{x})||q(\boldsymbol{h})). \end{aligned}$$

✓ Objective of LIB

- The Objective of LIB can be formulated as follows：

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{as} + \alpha\mathcal{L}_{gap} + \beta I(\boldsymbol{X}, \boldsymbol{H}) \\ &\leqslant -\sum_{\boldsymbol{x}}\sum_{\boldsymbol{l}}\sum_{\boldsymbol{h}} p(\boldsymbol{x}, \boldsymbol{l}) p(\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{l}) \log q(\boldsymbol{l}|\boldsymbol{h}) \\ &\quad - \alpha\sum_{\boldsymbol{l}}\sum_{\boldsymbol{h}} \log p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h}) \\ &\quad + \beta\sum_{\boldsymbol{x}}\sum_{\boldsymbol{l}} p(\boldsymbol{x}, \boldsymbol{l}) \mathrm{KL}(p(\boldsymbol{h}|\boldsymbol{x})||q(\boldsymbol{h})).\end{aligned}$$

† Use the empirical Monte Carlo approximation of sampling

$$\mathcal{L}_{LIB} = \frac{1}{n}\sum_{i=1}^{n}[-\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{x}_i) \log q(\boldsymbol{l}_i|\boldsymbol{h}) + \beta \mathrm{KL}(p(\boldsymbol{h}|\boldsymbol{x}_i)||q(\boldsymbol{h}))] - \alpha\sum_{\boldsymbol{l}}\sum_{\boldsymbol{h}} \log p(\boldsymbol{l} - \hat{\boldsymbol{d}}|\boldsymbol{h})$$

† Reparameterization trick

† Variational Inference

$$\begin{aligned}&\min_{\theta_{en},\theta_{de},\theta_{gd},\theta_{ld}} \mathcal{L}_{LIB} \\ \Rightarrow &\min_{\theta_{en},\theta_{de},\theta_{gd},\theta_{ld}} \frac{1}{n}\sum_{\boldsymbol{l}}[\frac{1}{2}||\boldsymbol{\mu}_{\boldsymbol{l}|\boldsymbol{h}} - \boldsymbol{l}||_2^2 \\ &+\alpha(\frac{1}{2}(\boldsymbol{l} - \hat{\boldsymbol{d}})^T(\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{h}}^{-2}\boldsymbol{I})(\boldsymbol{l} - \hat{\boldsymbol{d}}) + \log\det(\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{h}}^2\boldsymbol{I}))] \\ &+\frac{\beta}{2}\sum_{\boldsymbol{x}}[\boldsymbol{\mu}_{\boldsymbol{h}|\boldsymbol{x}}^T\boldsymbol{\mu}_{\boldsymbol{h}|\boldsymbol{x}} + \mathrm{tr}(\boldsymbol{\sigma}_{\boldsymbol{h}|\boldsymbol{x}}^2\boldsymbol{I}) - \log\det(\boldsymbol{\sigma}_{\boldsymbol{h}|\boldsymbol{x}}^2\boldsymbol{I})]\end{aligned}$$

✓ Main difference between LIB and existing methods

- LIB deals with the problem of LE from the perspective of information bottleneck
- $\|\boldsymbol{d}-\boldsymbol{l}\|_2^2$ ← Extra constraint： information in the label distributions is inherited from the initial logical labels.

⇕

$$\frac{1}{2}(\boldsymbol{l}-\hat{\boldsymbol{d}})^T(\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{h}}^{-2}\boldsymbol{I})(\boldsymbol{l}-\hat{\boldsymbol{d}})+\log\det(\boldsymbol{\sigma}_{\boldsymbol{\delta}|\boldsymbol{h}}^{2}\boldsymbol{I})$$

It can be deduced by excavating the label relevant information about the label gaps between logical labels and label distributions reasonably.

Experimental Results

✓ Metrics and datasets

$$D_{\text{Chebyshev}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \max_i \left| d^{y_i} - \hat{d}^{y_i} \right|,$$

$$D_{\text{Canberra}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sum_{i=1}^{c} \frac{\left| d^{y_i} - \hat{d}^{y_i} \right|}{d^{y_i} + \hat{d}^{y_i}},$$

$$D_{\text{Clark}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sqrt{\sum_{i=1}^{c} \frac{\left(d^{y_i} - \hat{d}^{y_i} \right)^2}{\left(d^{y_i} + \hat{d}^{y_i} \right)^2}},$$

$$D_{\text{Kullback-Leibler}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sum_{i=1}^{c} d^{y_i} \ln \frac{d^{y_i}}{\hat{d}^{y_i}},$$

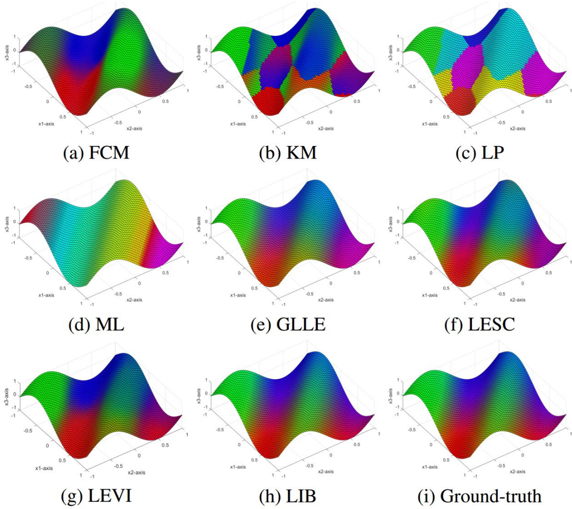
$$S_{\text{Cosine}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \frac{\sum_{i=1}^{c} d^{y_i} \hat{d}^{y_i}}{\sqrt{\sum_{i=1}^{c} \left(d^{y_i} \right)^2} \sqrt{\sum_{i=1}^{c} \left(\hat{d}^{y_i} \right)^2}},$$

$$S_{\text{Intersection}}(\boldsymbol{d}, \hat{\boldsymbol{d}}) = \sum_{i=1}^{c} \min \left(d^{y_i}, \hat{d}^{y_i} \right).$$

Dataset	# dimension q	# instance n	# labels c
Artificial_toy	3	2601	3
Movie	1869	7755	5
SBU-3DFE	243	2500	6
SJAFFE	243	213	6
Yeast-alpha	24	2465	18
Yeast-cdc	24	2465	15
Yeast-cold	24	2465	4
Yeast-diau	24	2465	7
Yeast-dtt	24	2465	4
Yeast-elu	24	2465	14
Yeast-heat	24	2465	6
Yeast-spo	24	2465	6
Yeast-spo5	24	2465	3
Yeast-spoem	24	2465	2

Qinghai Zheng, Fuzhou University, Email： zhengqinghai@fzu.edu.cn, qhxjtu@163.com homepage： https://qinghai-zheng.github.io/

✓ Some results on the toy dataset and real-world datasets



(a) FCM (b) KM (c) LP
(d) ML (e) GLLE (f) LESC
(g) LEVI (h) LIB (i) Ground-truth

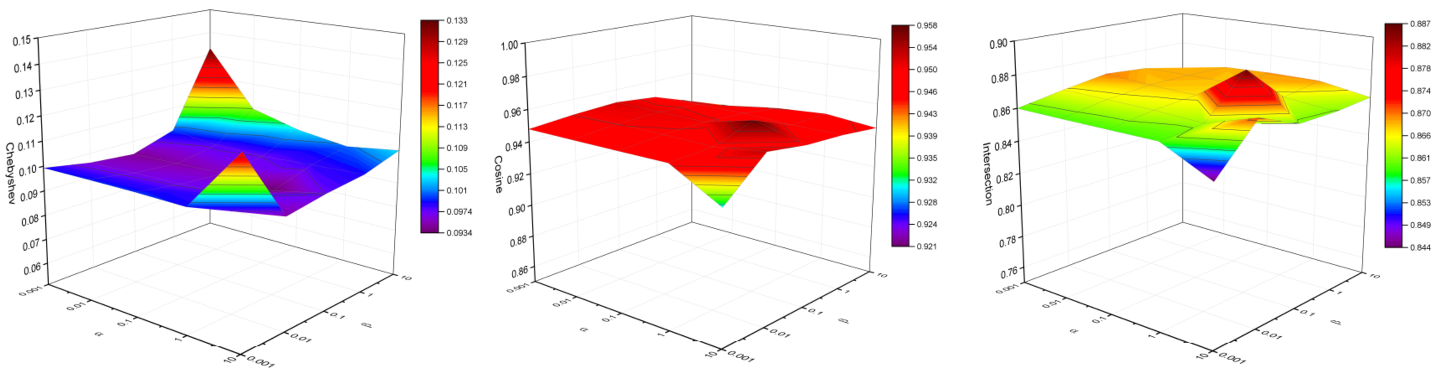
Metric	Chebyshev ↓							
Method	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB
Movie	0.230	0.234	0.161	0.164	0.122	0.121	0.110	**0.107**
SUB-3DFE	0.135	0.238	0.123	0.233	0.126	0.122	0.095	**0.094**
SJAFFE	0.132	0.214	0.107	0.186	0.087	**0.069**	0.075	0.071
Yeast-alpha	0.044	0.063	0.040	0.057	0.020	0.015	**0.012**	0.017
Yeast-cdc	0.051	0.076	0.042	0.071	0.022	0.019	**0.016**	0.017
Yeast-cold	0.141	0.252	0.137	0.242	0.066	0.056	0.082	**0.054**
Yeast-diau	0.124	0.152	0.099	0.148	0.053	**0.042**	0.044	0.049
Yeast-dtt	0.097	0.257	0.128	0.244	0.052	0.043	0.084	**0.034**
Yeast-elu	0.052	0.078	0.044	0.072	0.023	0.019	**0.017**	0.018
Yeast-heat	0.169	0.175	0.086	0.165	0.049	0.046	0.052	**0.039**
Yeast-spo	0.130	0.175	0.090	0.171	0.062	0.060	0.055	**0.053**
Yeast-spo5	0.162	0.277	0.114	0.273	0.099	0.092	0.091	**0.076**
Yeast-sopem	0.233	0.408	0.163	0.403	0.088	0.087	0.115	**0.069**
Avg.Rank	6.077	8.000	5.000	6.846	3.769	2.308	2.463	**1.538**

Metric	Chebyshev ↓		Clark ↓		Canberra ↓		Kullback-Leibler ↓		Cosine ↑		Intersection ↑	
Method	LIB_{gap}	LIB	LIB_{gap}	LIB	LIB_{gap}	LIB	LIB_{gap}	LIB	LIB_{gap}	LIB	LIB_{gap}	LIB
Movie	0.120	**0.107**	0.563	**0.517**	1.029	**0.920**	0.099	**0.077**	0.938	**0.955**	0.834	**0.859**
SUB-3DFE	0.130	**0.094**	0.395	**0.297**	0.849	**0.611**	0.079	**0.041**	0.923	**0.958**	0.846	**0.887**
SJAFFE	0.113	**0.071**	0.391	**0.262**	0.816	**0.531**	0.066	**0.027**	0.938	**0.973**	0.860	**0.909**
Yeast-alpha	0.018	**0.017**	0.281	**0.275**	0.920	**0.893**	0.010	**0.009**	0.991	**0.992**	0.950	**0.951**
Yeast-cdc	0.019	**0.017**	0.254	**0.242**	0.782	**0.747**	0.009	**0.008**	0.991	**0.992**	0.948	**0.951**
Yeast-cold	0.061	**0.017**	0.162	**0.146**	0.280	**0.250**	0.016	**0.012**	0.985	**0.988**	0.930	**0.938**
Yeast-diau	0.050	**0.049**	0.288	**0.273**	0.659	**0.621**	0.025	**0.022**	0.977	**0.979**	0.908	**0.913**
Yeast-dtt	0.045	**0.034**	0.124	**0.092**	0.217	**0.158**	0.010	**0.005**	0.991	**0.995**	0.946	**0.961**
Yeast-elu	0.019	**0.018**	0.237	**0.224**	0.714	**0.670**	0.009	**0.008**	0.992	**0.992**	0.949	**0.952**
Yeast-heat	0.045	**0.039**	0.193	**0.165**	0.388	**0.327**	0.014	**0.011**	0.986	**0.990**	0.936	**0.946**
Yeast-spo	0.059	**0.053**	0.253	**0.224**	0.523	**0.454**	0.025	**0.019**	0.976	**0.982**	0.914	**0.925**
Yeast-spo5	0.097	**0.076**	0.193	**0.158**	0.300	**0.241**	0.032	**0.021**	0.971	**0.983**	0.903	**0.924**
Yeast-sopem	0.088	**0.069**	0.130	**0.104**	0.181	**0.144**	0.027	**0.018**	0.977	**0.985**	0.912	**0.931**

Experimental Results

✓ Some results on the toy dataset and real-world datasets

Metric	Chebyshev ↓								Clark ↓							
Method	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB	FCM	KM	LP	ML	GLLE	LESC	LEVI	LIB
Movie	0.230	0.234	0.161	0.164	0.122	0.121	0.110	**0.107**	0.859	1.766	0.913	1.140	0.569	0.564	0.551	**0.517**
SUB-3DFE	0.135	0.238	0.123	0.233	0.126	0.122	0.095	**0.094**	0.482	1.907	0.580	1.848	0.391	0.378	0.303	**0.297**
SJAFFE	0.132	0.214	0.107	0.186	0.087	**0.069**	0.075	0.071	0.522	1.874	0.502	1.519	0.377	0.276	0.290	**0.262**
Yeast-alpha	0.044	0.063	0.040	0.057	0.020	0.015	**0.012**	0.017	0.821	3.153	1.185	3.088	0.337	**0.253**	0.319	0.275
Yeast-cdc	0.051	0.076	0.042	0.071	0.022	0.019	**0.016**	0.017	0.739	2.885	1.014	2.825	0.306	0.251	0.323	**0.242**
Yeast-cold	0.141	0.252	0.137	0.242	0.066	0.056	0.082	**0.054**	0.433	1.472	0.503	1.440	0.176	0.152	0.269	**0.146**
Yeast-diau	0.124	0.152	0.099	0.148	0.053	**0.042**	0.044	0.049	0.838	1.886	0.788	1.844	0.296	**0.224**	0.295	0.273
Yeast-dtt	0.097	0.257	0.128	0.244	0.052	0.043	0.084	**0.034**	0.329	1.477	0.499	1.446	0.143	0.119	0.294	**0.092**
Yeast-elu	0.052	0.078	0.044	0.072	0.023	0.019	**0.017**	0.018	0.579	2.768	0.973	2.711	0.295	0.241	0.317	**0.224**
Yeast-heat	0.169	0.175	0.086	0.165	0.049	0.046	0.052	**0.039**	0.580	1.802	0.568	1.764	0.213	0.199	0.288	**0.165**
Yeast-spo	0.130	0.175	0.090	0.171	0.062	0.060	0.055	**0.053**	0.520	1.811	0.558	1.768	0.266	0.258	0.277	**0.224**
Yeast-spo5	0.162	0.277	0.114	0.273	0.099	0.092	0.091	**0.076**	0.395	1.059	0.274	1.036	0.197	0.185	0.209	**0.158**
Yeast-sopem	0.233	0.408	0.163	0.403	0.088	0.087	0.115	**0.069**	0.401	1.028	0.272	1.004	0.132	0.129	0.182	**0.104**
Avg.Rank	6.077	8.000	5.000	6.846	3.769	2.308	2.463	**1.538**	5.385	8.000	5.615	7.000	3.385	1.923	3.462	**1.231**



(a) Chebyshev ↓ (b) Cosine ↑ (c) Intersection ↑

Qinghai Zheng, Fuzhou University, Email： zhengqinghai@fzu.edu.cn, qhxjtu@163.com homepage： https://qinghai-zheng.github.io/

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Thank you

Qinghai Zheng, Fuzhou University, Email： zhengqinghai@fzu.edu.cn, qhxjtu@163.com homepage： https://qinghai-zheng.github.io/