模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Missing multi-label learning with global high-rank &/Or local Low-rank assumptions

陈松灿

(s.chen@nuaa.edu.cn || http://parnec.nuaa.edu.cn)

Works with Zhongchen Ma（马忠臣） & Xiang Li（李想）

2023-11-1

Outline

01

Introduction

Multi-label learning Formulation

Multi-Label Learning （MLL）

Given $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \{y_1, y_2, \cdots y_k\}$. Goal $f: \mathcal{X} \mapsto 2^{\mathcal{Y}}$

Categorization of MLL

Categorization by label (relative) completeness :

Complete MLL

完整是**相对的**，

缺失是**绝对的!**

Missing MLL

Ignored: Open scenarios ➔ (known & seen, unknown & seen)

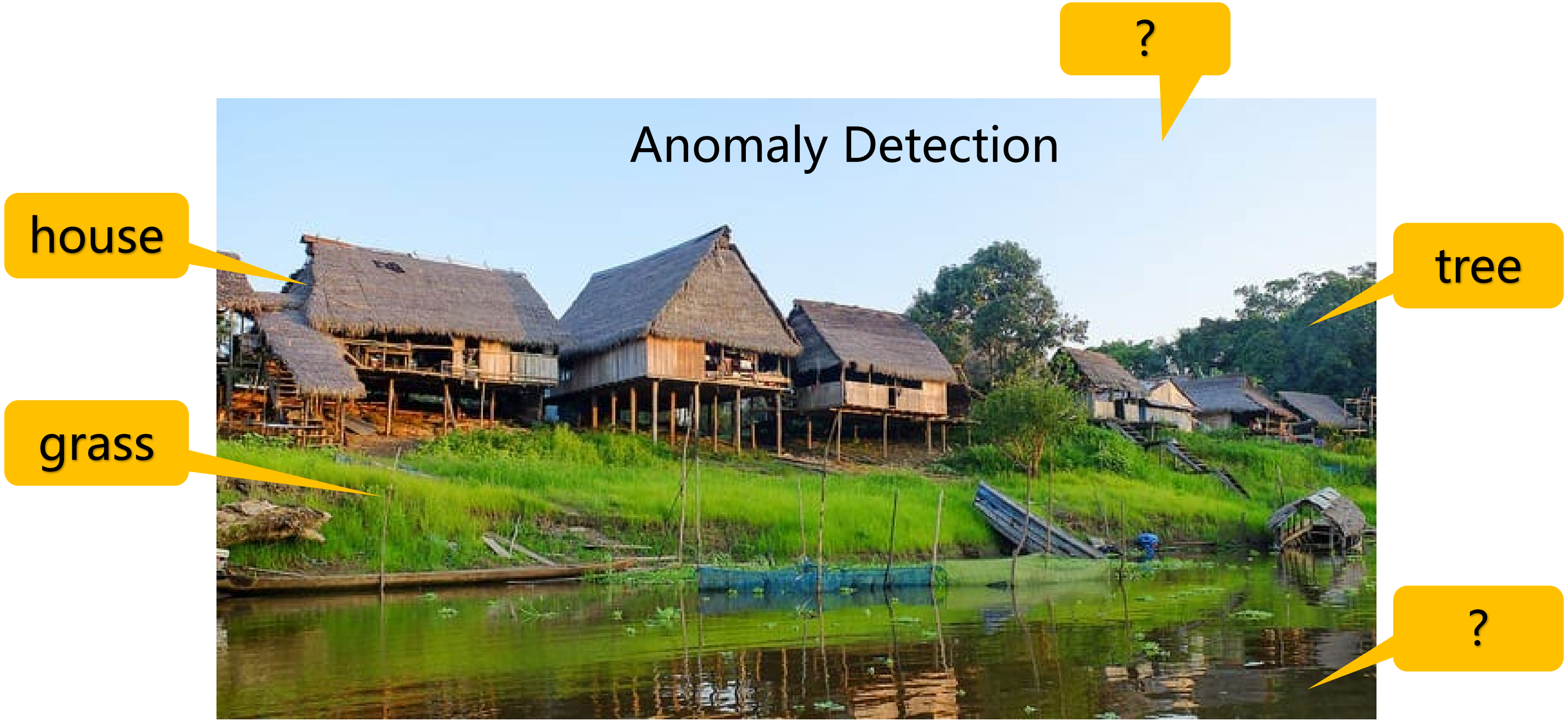
02

Missing MLL:

(Global) High-Rankness on Single View

Multi-label learning with Missing labels:

A more practical and challenging scenario: **M**issing **M**ulti-label **L**earning (**MML**)



(Relative) Causes of label missing:

- Expensive cost of fully labeling

- Limited knowledge (w.r.t lablers)

- Interests in *partial* labels (w.r.t lablers)

Challenges of <u>Missing</u> ML:

- Performance degradation of fully supervised methods
- Complex relationships among multiple labels
- $\mathcal{O}(2^k)$ (k is # of classes) vs $\mathcal{O}(k)$ (multi-class)

Methods of addressing MML:

1. Discarding the samples with missing labels

2. Label Completion
- Pre-processing methods (e.g. ML-MG ICCV 2015)
- Transductive methods (e.g. lrMMC AAAI 2015)
- Synchronized methods (e.g. GLOCAL TKDE 2018)

Ma ZC, Chen SC. Expand globally, shrink locally: Discriminant multi-label learning with missing labels[J]. Pattern Recognition, 2021

Keys of addressing & completing MML:

1. Correlations among multiple labels (unknown)
   - **Local** (label correlations shared by a **subset** of samples)
   - **Global** (label correlations shared by **all** the samples)

2. Structural information of multi-label
   - Sparse (a few positive labels)
   - **Low-rank (Popular assumption)**
   - **(Global) High-rank**

3. Effective representation learning

Low-rank based label completion methods: Matrix Completion

Probably the most popular MML methods.

- Theoretical foundations (only under the assumption)
- Label co-occurrence
- Optimization convenience

Popular = Reasonable for usage ?

Not exactly!

Candès E J, Recht B. Exact matrix completion via convex optimization. Foundations of Comput. mathematics, 2009, 9(6): 717-772. 6725

Low-rank (unknown) based MML methods in single view:

Low rank **E**mpirical risk minimization for **M**ulti-label **L**earning (**LEML**)

$$\hat{Z} = \arg\min_{Z} J_{\Omega}(Z) = \sum_{(i,j)\in\Omega} \ell(Y_{ij}, f^{j}(\boldsymbol{x}_i; Z)) + \lambda \cdot r(Z),$$

$$s.t.\ \operatorname{rank}(Z) \le k,$$

Explicitly upper-bounded rank for Z !

Let $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}^{T}$ and $r(\boldsymbol{Z}) = ||\boldsymbol{Z}||_{tr}$, where $\boldsymbol{Z} \in \mathbb{R}^{d\times L}$ is the coefficient matrix and $\boldsymbol{W} \in \mathbb{R}^{d\times k}$, $\boldsymbol{H} \in \mathbb{R}^{L\times k}$ (L is the number of labels), then we have,

$$J_{\Omega}(W,H) = \sum_{(i,j)\in\Omega} \ell(Y_{ij}, \boldsymbol{x}_i^{T} W \boldsymbol{h}_j) + \frac{\lambda}{2}\left(\|W\|_F^2 + \|H\|_F^2\right)$$

Yu H F, Jain P, Kar P, et al. Large-scale multi-label learning with missing labels. ICML. 2014: 593-601. 534

Low-rank based MML methods in single view:

Low-Rank label Correlation for Multi-Label classification (**ML-LRC**)

$$\min_{W,S,E} \|XW - YS\|_F^2 + \lambda_1\|W\|_F^2 + \lambda_2\|S\|_* + \lambda_3\|E\|_{2,1}$$
$$\text{s.t. } Y = YS + E$$

Implicit rank for S !

where $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{W} \in \mathbb{R}^{d\times c}, \boldsymbol{Y} \in \mathbb{R}^{n\times c}, \boldsymbol{S} \in \mathbb{R}^{c\times c}$ is the label correlation matrix and $\boldsymbol{E} \in \mathbb{R}^{n\times c}$ is the error matrix.

Questable: in aligning XW and YS, why to use different norms for W and S respectively?

Xu L, Wang Z, Shen Z, et al. Learning low-rank label correlations for multi-label classification with missing labels. ICDM,2014.

Common Problem with Low-rank based MML methods :

However, mostly violating the reality in applications!

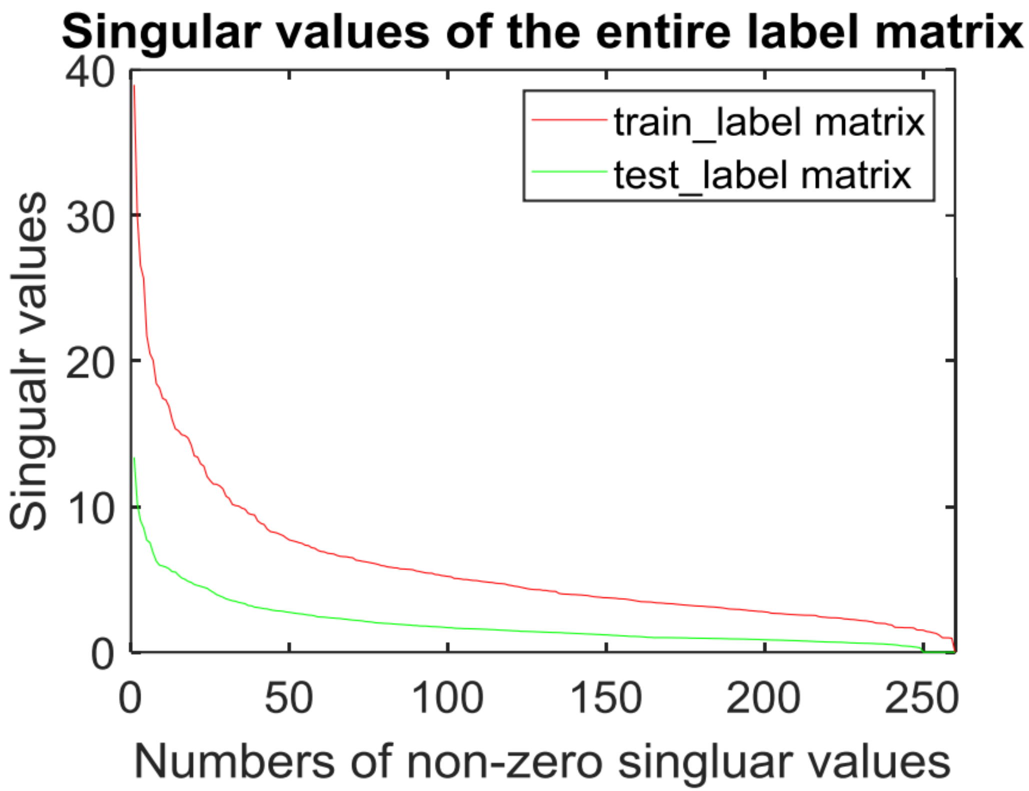
Our findings:

(Global) High-rankness of the whole/global multi-label matrix.

- **Intuitively**, samples in real datasets with multiple labels usually both are **diverse** and contain **dissimilar** labels.
- **Mathematically**, as entries of the label matrix take **binary values**, it is unlikely for this matrix to be low-rank.

02 Missing MLL: High-Rankness on Single View

The **evidence** of high-rankness:



Corel5k dataset

Statistics of the Datasets

datasets	n	c	#avg	train_rank	test_rank
Corel5k	4999	260	3.396	259	249
Espgame	20770	268	4.686	268	268
IAPRTC12	19627	291	5.719	291	291
Mirflickr	25000	38	4.716	38	38
Pascal07	9963	20	1.465	20	20

Rank Drift **between training and test sets**

These benchmarks are also used in iMVWL (IJCAI 2018), LSA-MML (AAAI 2018), SIMM (IJCAI 2019), LCBM (TPAMI 2021).

Our starting points:

Explicit structural information
- Local low-rank for label **subset**
- Global high-rank for **whole** label set

Implicit label correlations
- Similar labels are **strongly correlated**
- Dissimilar labels are **weakly correlated**, or even uncorrelated

02 Missing MLL: High-Rankness on Single View

An intuitive description:

label matrix

sub-label matrices

Y

	sky	cloud	tree	sea	fish
x_1	1	1	-1	-1	-1
x_2	-1	-1	1	1	-1
x_3	-1	-1	-1	1	1
x_4	1	1	-1	1	-1
x_5	1	-1	1	-1	-1

rank(Y) = 5

global structure of multiple labels

sky

	sky	cloud	tree	sea	fish
x_1	**1**	1	-1	-1	-1
x_4	**1**	1	-1	1	-1
x_5	**1**	-1	1	-1	-1

rank(sky) = 3

cloud

	sky	cloud	tree	sea	fish
x_1	1	**1**	-1	-1	-1
x_4	1	**1**	-1	1	-1

rank(cloud) = 2

• • •

fish

	sky	cloud	tree	sea	fish
x_3	-1	-1	-1	1	**1**

rank(fish) = 1

local structure of multiple labels

Global High Rank

Local low Rank

02 Missing MLL:
High-Rankness on Single View

Our formulation:

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathcal{R}_{\boldsymbol{\Omega}}(\mathbf{XW}) - \tilde{\mathbf{Y}}\|_F^2 + \lambda_d \left(\underbrace{\boxed{\sum_{k=1}^{c} \|\mathbf{X}_k\mathbf{W}\|_*}}_{\text{Local low-rank}} - \underbrace{\boxed{\|\mathbf{XW}\|_*}}_{\text{Global high-rank}} \right)$$

Empirical Risk

where $\mathbf{X}_k$ denotes the training samples associated with label k.

(Intersections of $\mathbf{X}_k$ and $\mathbf{X}_j$ are not empty, $k \neq j$)

Containing **only one** hyper-parameter! However, a concern: can the regularization term be non-negative?

Ma Z, Chen S. Expand globally, shrink locally: Discriminant multi-label learning with missing labels. Pattern Recognition, 2021.

A theorem:

Theorem 1. Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ be matrices of the same row dimensions, and $[\mathbf{A}, \mathbf{C}]$ be concatenation of $\mathbf{A}$ and $\mathbf{C}$, Likewise for $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$ and $[\mathbf{B},\mathbf{C}]$. Then, we have

$$\|[\mathbf{A}, \mathbf{B}, \mathbf{C}]\|_* \leq \|[\mathbf{A}, \mathbf{C}]\|_* + \|[\mathbf{B}, \mathbf{C}]\|_*$$

$\sum_{k=1}^{c} ||\mathbf{X}_k||_*$ is an upper bound of $||\mathbf{XW}||_*$

Ma Z, Chen S. Expand globally, shrink locally: Discriminant multi-label learning with missing labels. Pattern Recognition, 2021

Compared methods:

LEML (ICML 2014)

$$\hat{Z} = \arg\min_{Z} J_{\Omega}(Z) = \sum_{(i,j)\in\Omega} \ell(Y_{ij}, f^j(\boldsymbol{x}_i; Z)) + \boxed{\lambda} \cdot r(Z),$$
$$s.t.\ \ \mathrm{rank}(Z) \leq \boxed{k}$$

One explicit hyper-parameter
One implicit hyper-parameter

ML-LRC (ICDM 2014)

$$\min_{W,S,E} \|XW - YS\|_F^2 + \boxed{\lambda_1}\|W\|_F^2 + \boxed{\lambda_2}\|S\|_* + \boxed{\lambda_3}\|E\|_{2,1}$$
$$\text{s.t. } Y = YS + E$$

Three explicit hyper-parameters

GLOCAL (TKDE 2018)

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{Z}}} \|\boldsymbol{J}\circ(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})\|_F^2 + \boxed{\lambda}\|\boldsymbol{V} - \boldsymbol{W}^\top\boldsymbol{X}\|_F^2$$
$$+ \sum_{m=1}^{g}\left(\frac{\boxed{\lambda_3} n_m}{n}\mathrm{tr}\left(\boldsymbol{F}_0^\top\boldsymbol{Z}_m\boldsymbol{Z}_m^\top\boldsymbol{F}_0\right) + \boxed{\lambda_4}\mathrm{tr}(\boldsymbol{F}_m^\top\boldsymbol{Z}_m\boldsymbol{Z}_m^\top\boldsymbol{F}_m)\right)$$
$$+ \boxed{\lambda_2}\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})$$
$$\text{s.t.}\quad \mathrm{diag}(\boldsymbol{Z}_m\boldsymbol{Z}_m^\top) = \boldsymbol{1}, m = 1, 2, \ldots, g.$$

Four explicit hyper-parameters

LSML (Inf. Sci. 2019)

$$\min_{\mathbf{W},\mathbf{C}} \frac{1}{2}\|\mathbf{XW} - \mathbf{YC}\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{YC} - \mathbf{Y}\|_F^2 + \lambda_2\|\mathbf{C}\|_1 + \lambda_3\|\mathbf{W}\|_1 + \lambda_4\mathrm{tr}(\mathbf{WLW}^T)$$
$$s.t.\ \ \mathbf{C} \succeq 0$$

Four explicit hyper-parameters

Experimental results:

Results for learning with **full** labels.

		LEML	ML-LRC	GLOCAL	LSML	DM2L-l	DM2L-nl
Art	Rkl	0.167	0.161	0.152	0.127	0.139	**0.125**
	Auc	0.835	0.841	0.85	**0.876**	0.864	0.875
	Cvg	6.305	5.636	5.852	4.978	5.409	**4.965**
	Ap	0.596	0.484	0.61	**0.625**	0.623	0.61
Bus	Rkl	0.056	0.04	0.047	0.043	0.046	**0.039**
	Auc	0.945	**0.962**	0.954	0.958	0.956	0.961
	Cvg	3.153	**2.333**	2.656	2.521	2.674	2.419
	Ap	0.864	0.885	0.877	0.881	0.882	**0.887**
Rec	Rkl	0.1788	0.149	0.159	0.152	0.147	**0.136**
	Auc	0.8252	0.855	0.845	0.853	0.857	**0.864**
	Cvg	5.0164	4.082	4.542	4.365	4.2	**3.983**
	Ap	0.6027	0.575	0.618	0.631	**0.635**	0.633
Enr	Rkl	0.172	0.121	0.117	0.136	0.131	**0.109**
	Auc	0.83	0.882	0.885	0.866	0.871	**0.892**
	Cvg	20.37	15.477	**15.430**	18.346	17.714	15.718
	Ap	0.589	0.603	0.632	0.634	0.588	**0.648**
Ima	Rkl	0.203	0.182	0.18	0.181	0.193	**0.138**
	Auc	0.797	0.819	0.82	0.819	0.807	**0.862**
	Cvg	1.069	0.996	0.992	0.993	1.044	**0.827**
	Ap	0.758	0.781	0.783	0.783	0.76	**0.833**
Soc	Rkl	0.106	0.08	0.078	0.062	0.065	**0.053**
	Auc	0.894	0.92	0.922	0.938	0.935	**0.947**
	Cvg	5.62	4.09	4.012	3.544	3.711	**3.077**
	Ap	0.723	0.596	0.666	0.777	0.78	**0.783**
Cor	Rkl	0.164	0.152	0.175	**0.142**	0.146	0.143
	Auc	0.836	0.848	0.825	**0.860**	0.854	0.857
	Cvg	47.896	45.377	49.909	**42.811**	43.673	43.323
	Ap	0.328	0.293	0.306	0.335	0.328	**0.339**
Tmc	Rkl	0.046	0.044	0.046	0.047	0.046	**0.030**
	Auc	0.954	0.956	0.954	0.954	0.955	**0.970**
	Cvg	2.874	2.701	2.855	2.88	2.838	**2.273**
	Ap	0.833	0.821	0.833	0.833	0.833	**0.886**

Low-rank based methods

LEML (ICML 2014)

ML-LRC (ICDM 2014)

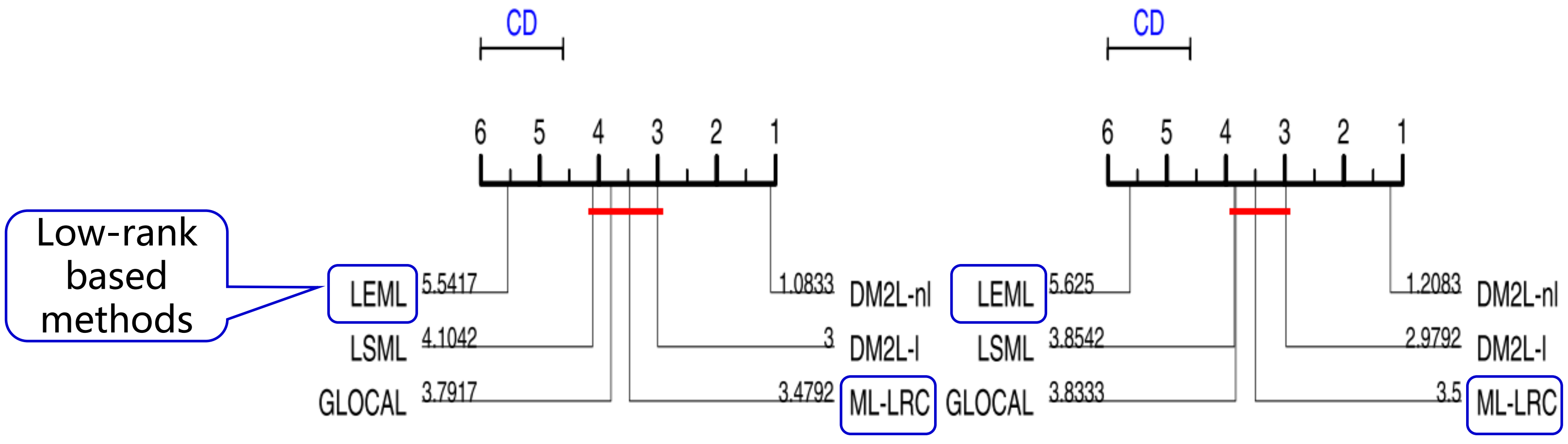
GLOCAL (TKDE 2018)

LSML (Inf. Sci. 2019)

Results for learning with missing labels.

		LEML	ML-LRC	GLOCAL	LSML	DM2L-l	DM2L-nl
Art	Rkl	0.199	0.164	0.16	0.145	0.138	**0.119**
	Auc	0.804	0.839	0.843	0.857	0.864	**0.881**
	Cvg	7.212	5.711	6.079	5.617	5.382	**4.810**
	Ap	0.556	0.479	0.6	0.608	0.621	**0.623**
Bus	Rkl	0.072	0.048	0.045	0.054	0.043	**0.037**
	Auc	0.93	0.953	0.957	0.947	0.958	**0.963**
	Cvg	3.905	2.8	2.532	3.097	2.571	**2.317**
	Ap	0.835	0.876	0.877	0.867	0.882	**0.889**
Rec	Rkl	0.211	0.15	0.161	0.177	0.149	**0.131**
	Auc	0.793	0.855	0.843	0.828	0.856	**0.870**
	Cvg	5.746	4.122	4.527	4.967	4.216	**3.834**
	Ap	0.556	0.577	0.601	0.6	0.628	**0.635**
Enr	Rkl	0.2	0.156	0.127	0.182	0.133	**0.123**
	Auc	0.802	0.846	0.875	0.82	0.869	**0.877**
	Cvg	23.336	19.006	**16.545**	21.974	17.82	17.2
	Ap	0.56	0.557	0.622	0.55	0.578	**0.631**
Ima	Rkl	0.213	0.185	0.184	0.186	0.222	**0.141**
	Auc	0.787	0.815	0.816	0.814	0.778	**0.859**
	Cvg	1.107	1.006	1.005	1.013	1.154	**0.831**
	Ap	0.748	0.778	0.779	0.778	0.72	**0.825**
Soc	Rkl	0.13	0.084	0.079	0.081	0.064	**0.052**
	Auc	0.87	0.916	0.921	0.919	0.936	**0.948**
	Cvg	6.661	4.393	4.088	4.500	3.634	**2.975**
	Ap	0.697	0.595	0.671	0.757	0.776	**0.780**
Cor	Rkl	0.166	0.155	0.181	0.146	0.15	**0.143**
	Auc	0.834	0.845	0.819	**0.858**	0.850	0.857
	Cvg	48.697	46.478	51.574	43.637	44.996	**43.137**
	Ap	0.326	0.284	0.300	0.334	0.323	**0.346**
Tmc	Rkl	0.047	0.045	0.047	0.047	0.046	**0.031**
	Auc	0.953	0.955	0.953	0.953	0.955	**0.969**
	Cvg	2.922	2.811	2.906	2.928	2.824	**2.337**
	Ap	0.832	0.831	0.83	0.832	0.83	**0.881**

Experimental results: Significance (Nemenyi) Test



(a) Ranking Loss (b) AUC

(c) Coverage (d) Average Precison

Code is available at https://github.com/John986/Multi-label-Learning-with-Missing-Labels.

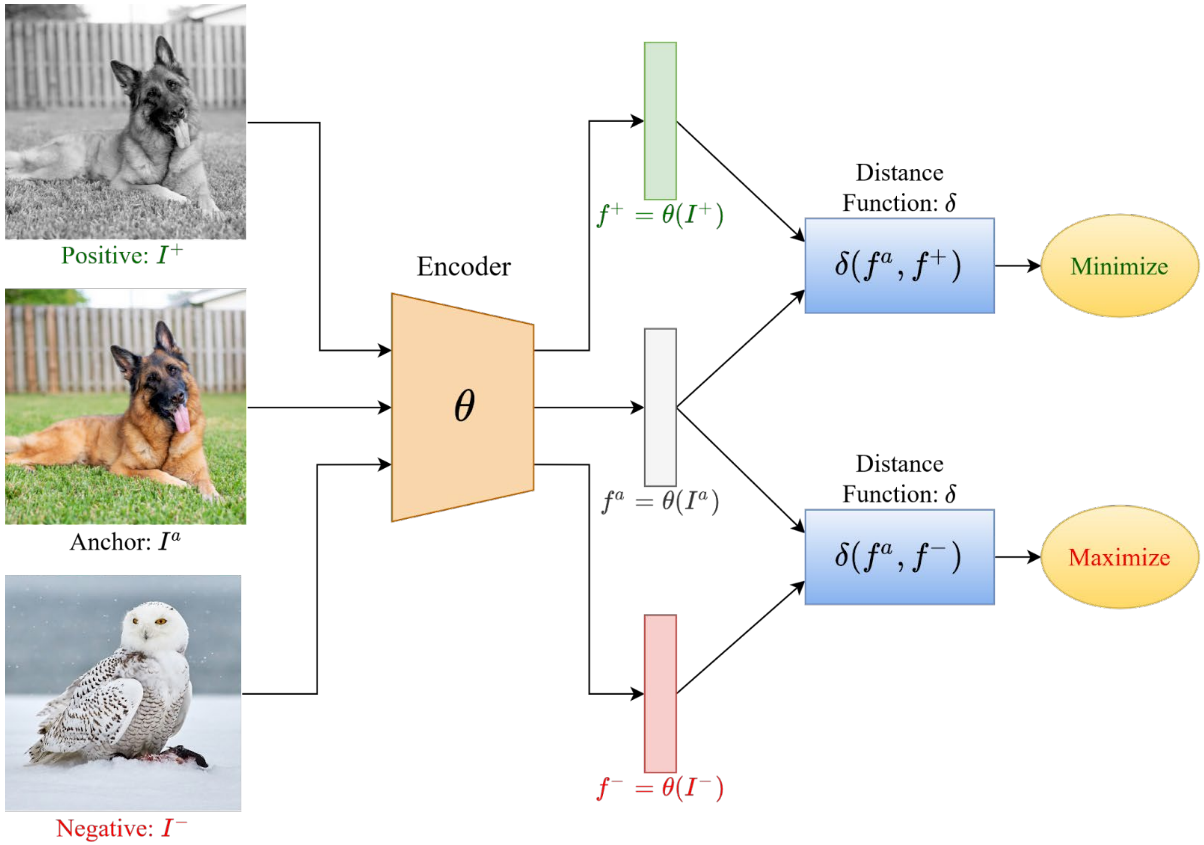
03

Missing MLL:

(Global) High-rankness for Contrastive Learning

What is Contrastive Learning?

- Contrastive learning is a machine learning technique used to **learn general features** of a dataset without labels by **teaching** a model which data points are **similar or different**.



Advantage: Contrastive learning **excels** on a wide range of tasks, such as Image Classification, Semantic Segmentation.

03 Missing MLL:
High-rankness for Contrastive Learning

ParNeC

Methods

Contrastive Learning
(示例间的2阶对比)

- Unsupervised Learning
  - **MoCo** (CVPR 2020)
  - SimCLR (ICML 2020)
  - SwAV (NeurIPS 2020)
  - BYOL (NeurIPS 2020)
- Supervised Learning
  - **SupCon** (NeurIPS 2020) (**Single label)**
  - MulCon (ArXiv.2021) **(Full Multi-label)**
  - GZSL (CVPR 2021)

K. He, H. Fan, et al. Momentum contrast for unsupervised visual representation learning (MoCo), CVPR, 2020. 9092

Khosla, Prannay, et al. Supervised contrastive learning (SupCon). NeurIPS, 2020: 18661-18673. 3004

Son D.Dao, et al, Contrast Learning Visual Attention for Multi Label Classification, arXiv:2107.11626, 15

Junwen Bai, etal. Gaussian Mixture Variational Autoencoder with Contrastive Learning for ML Classification, icml2022. 6

New challenges: Contrastive learning meets Missing labels

- **False** contrastive instances **unfavorable** , due to difficulty in **defining** the positive and negative instances to **contrast** a given anchor image in **multi-label scenario**.

E.g., an anchor instance x with label y

$$x, y = \{\lambda_1, \lambda_2\}$$

Is another instance $x', y' = \{\lambda_1\}$ the positive or negative contrastive instance?

➔确实是**其中一个标记**的反例，但**未必是其整体**的反例！

还有可能的是，同一示例会产生正例和负例的矛盾配对！

One Solution:

- Learn **individual label-specific** embedding for each image

$$g_i = MultiAttBlock(U, r_i, r_i)$$

where $r_i = \mathrm{Enc}(x_i) \in \mathbb{R}^{C\times H\times W}$, each row of $U \in \mathbb{R}^{L\times C}$ is a class-specific the embedding, $g_i \in \mathbb{R}^{L\times D}$ represents the **label-level embeddings**

- **(Full) Mul**ti-**l**abel Classification with **Con**trastive Loss (MulCon)

$$L_{con}^{ij} = \frac{-1}{|P(i,j)|} \sum_{z_p \in P(i,j)} \log \frac{\exp(z_{ij} \cdot z_p/\tau)}{\sum_{z_a \in A(i,j)} \exp(z_{ij} \cdot z_a/\tau))}$$

where $z_{ij} = \mathrm{Proj}(g_{ij}) \in \mathbb{R}^{d_z}$, $I = \{z_{ij} \in Z | y_{ij} = 1\}$, $A(i,j) = I \setminus z_{ij}$

Dinh S D D E Z, Cai P J. Multi-Label Image Classification with Contrastive Learning, arXiv:2107.11626, 15.

03 Missing MLL: High-rankness for Contrastive Learning

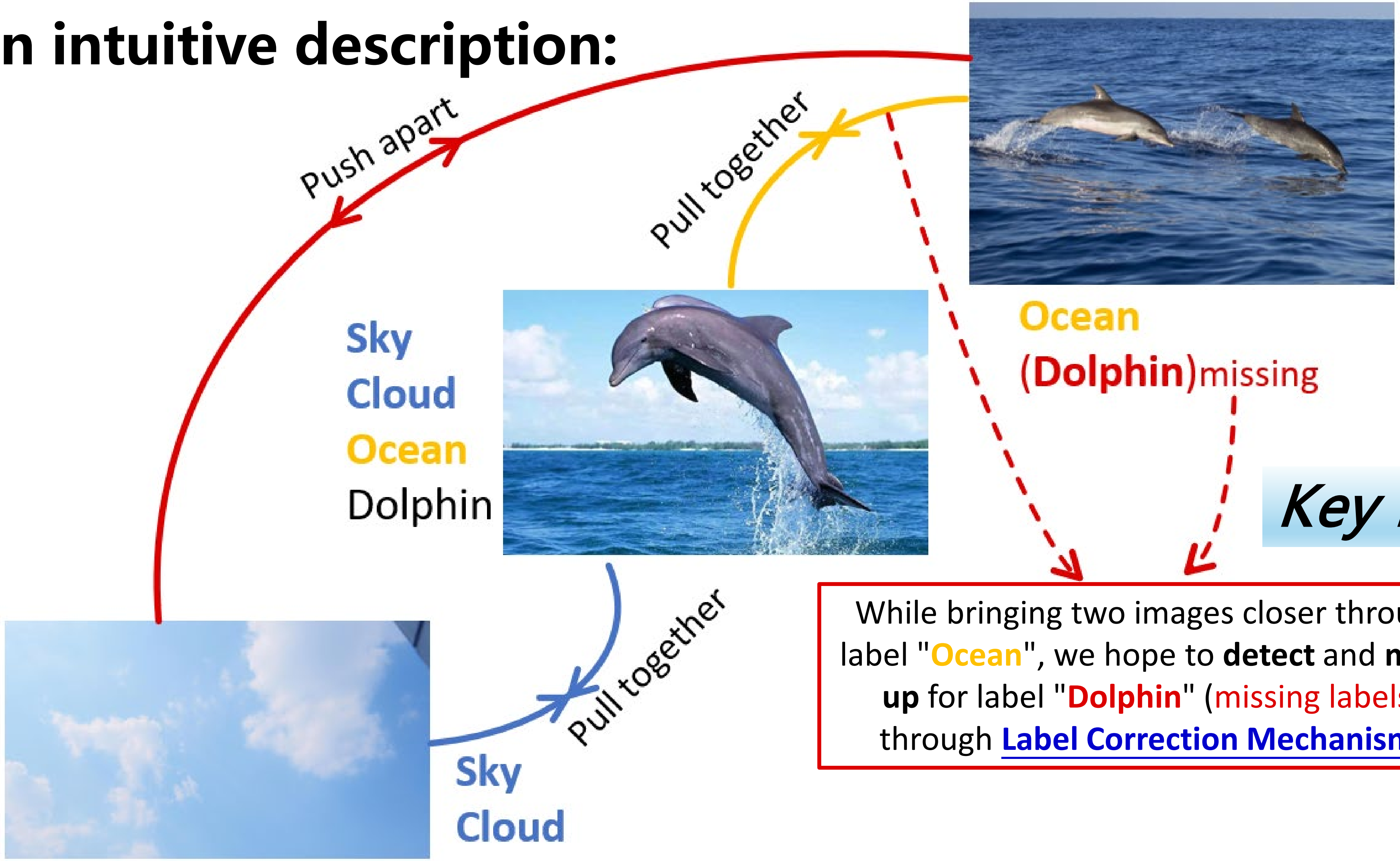
However, it meets new challenges for Missing scenario:

- How to **incorporate label dependency**, which has been shown to be helpful in dealing with missing labels, into the **contrastive learning formulation**.

- How to **incorporate** the **Missing Label Correction Mechanism** into the **learning process** to better solve the Missing MLL task.

03 Missing MLL: High-rankness for Contrastive Learning

An intuitive description:



Key Point!

While bringing two images closer through label "Ocean", we hope to **detect** and **make up** for label "**Dolphin**" (missing labels) through **Label Correction Mechanism**.

03 Missing MLL: High-rankness for Contrastive Learning

Our starting points:

Define contrast pairs in a **multi-label scenario**

- Accurately bring images close to their true positive images and **false (假) negative images**
- Far from the true (真) negative images of images.

Preserve label structure

- Local low-rank
- Global high-rank

Join our label correction mechanism using DNN nature (性质)

Our formulation (as an <u>independent</u> Regularization term)

$$\mathcal{L}_{CLML} = \boxed{\sum_{k=1}^{C} \| \{\mathbf{Z}_i | \widetilde{y}_{ik} = +1, \mathbf{Z}_i \in \mathbf{Z}\} \|_*} - \boxed{\|\mathbf{Z}\|_*}$$

$$\widetilde{y}_{ik} = \begin{cases} y_{ij}^o \ , & N_e < N_E \\ LaCo(y_{ij}^o) \ , & N_e \geq N_E \end{cases}$$

$$LaCo(y_{ij}^o) = \begin{cases} +1 \ , & f(\mathbf{Z}_i)_j \geq \delta \\ y_{ij}^o \ , & f(\mathbf{Z}_i)_j < \delta \end{cases}$$

Pull together <u>similar samples</u> while enforcing **low rank constraint**

Push apart <u>dissimilar samples</u> while enforcing **high rank constraint**

<u>LaCo</u>= <u>La</u>bel <u>Co</u>rrection Using DNN nature

where $\boldsymbol{Z}_k = \mathbb{E}(\boldsymbol{X}_k; \theta)$ is <u>the sub-matrix of deep embedding belonging to label k</u>.

<u>Notice</u>: Strictly speaking, CLML is ***essentially*** a contrast among <u>the subset of instances</u> **<u>rather than generally</u>** <u>individual instances</u>!

Ma, Zhongchen et al. Label Structure Preserving Contrastive Embedding for Multi-Label Learning with Missing Labels. arXiv: 2209.01314, TIP major revision

03 Missing MLL: High-rankness for Contrastive Learning

Algorithm:

Algorithm 1 Label Structure Preserving Contrastive Embedding

Input: Training data matrix $\mathbf{X}$, label matrix $\mathbf{Y}^o$, deep embedding network $\mathbf{Z} = \mathbb{E}(\mathbf{X}, \theta)$ and deep multi-label classifier $f(\mathbf{Z}, \phi)$, trade-off parameter λ.

Output: The well trained deep model $\mathbb{E}(\cdot, \theta)$ and $f(\cdot, \phi)$

for $N_e = \{1, \cdots, N_{epoch}\}$ **do**
- **for** each minibatch $\mathbf{X}_b$ and $\mathbf{Y}_b^o$ **do**
  - $\mathbf{Z}_b = \mathbb{E}(\mathbf{X}_b, \theta)$
  - Correct the false negative labels in $\mathbf{Y}_b^o$ according to $LaCo(y_{ij}^o) = \begin{cases} +1\ , & f(\mathbf{Z}_i)_j \geq \delta \\ y_{ij}^o\ , & f(\mathbf{Z}_i)_j < \delta \end{cases}$
  - Calculate contrastive loss $\mathcal{L}_{CL}(\mathbf{Z}_b, \mathbf{Y}_b^o)$ according to $\mathcal{L}_{CLML} = \sum_{k=1}^{C} \| \{\mathbf{Z}_i | \widetilde{y}_{ik} = +1, \mathbf{Z}_i \in \mathbf{Z}\} \|_* - \|\mathbf{Z}\|_*$
  - Calculate total loss according to $\min_{\theta,\phi} \mathcal{L}_{classification}(\mathbf{X}, \mathbf{Y}^o, \theta, \phi) + \lambda \cdot \mathcal{L}_{CLML}(\mathbf{X}, \mathbf{Y}^o, \theta)$
  - Backpropagation
- **end for**

end for

03 Missing MLL: High-rankness for Contrastive Learning

Advantages:

- CLML is the **first** contrastive regularization term proposed for **Missing MLL**. It can relatively accurately bring images close to their **true positive images** and **false negative images**, far away from **their true negative** images.

- The **global and local label dependencies** are naturally **preserved** in CLML, allowing the label correlation to be used more effectively to solve the Missing MLL task.

03 Missing MLL: High-rankness for Contrastive Learning

Compared methods:

BCE (Classic)

$$L_n = -w_n[y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)]$$

Classical classification loss function

Focal (ICCV 2017)

$$L_{Focal} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t)$$

Solve the problem of class-imbalance

Hill (arXiv 2022)

$$L_{Hill} = -w(p) \times MSE$$
$$= -(\lambda - p)p^2$$

re-weight negatives in the shape of a Hill to alleviate the effect of false negatives

SPLC (arXiv 2022)

$$L_{SPLC}^{+} = loss^{+}(p)$$
$$L_{SPLC}^{-} = \mathbb{I}(p \le \tau)loss^{-}(p) + (1 - \mathbb{I}(p \le \tau))loss^{+}(p)$$

use the loss derived from the maximum likelihood criterion under an approximate distribution of missing labels

03 Missing MLL: High-rankness for Contrastive Learning

ParNeC

Experimental results:

TABLE II: Compared results on COCO dataset with varied missing label ratios

Method		BCE (full labels)	BCE	BCE+CLML	Focal [44]	Focal+CLML	Hill [17]	Hill+CLML	SPLC [17]	SPLC+CLML
75% labels left	mAP ↑	80.3	76.8	**78.0**	77.0	**78.3**	78.8	**79.6**	78.4	**80.4**
	CP↑	80.8	85.1	**86.2**	83.8	**86.0**	**73.6**	72.8	72.6	**75.6**
	CR↑	70.3	58.1	**58.7**	59.4	**61.0**	74.4	**76.3**	**75.1**	74.6
	CF1↑	74.9	67.7	**68.5**	68.4	**69.7**	73.6	**74.1**	73.2	**74.8**
	OP↑	84.3	90.1	**90.9**	88.6	**89.1**	**76.4**	74.6	74.0	**79.1**
	OR↑	74.2	58.7	**59.3**	59.8	**61.2**	78.3	**80.3**	**79.3**	78.0
	OF1↑	78.9	71.1	**71.8**	71.4	**72.6**	77.3	77.3	76.6	**78.5**
single label	mAP↑	-	68.6	**69.5**	70.2	**71.8**	73.2	**74.0**	73.2	**74.0**
	CP↑	-	88.6	**89.1**	88.2	**88.9**	79.7	**83.0**	**83.8**	80.9
	CR↑	-	33.0	**33.5**	36.0	**37.4**	**58.0**	55.7	53.1	**58.7**
	CF1↑	-	43.8	**44.2**	47.0	**48.6**	**65.5**	64.2	61.6	**65.5**
	OP↑	-	93.9	**94.8**	93.4	**93.9**	85.3	**88.7**	**90.1**	86.4
	OR↑	-	23.6	**24.5**	26.6	**28.3**	**58.7**	55.0	53.8	**60.5**
	OF1↑	-	37.7	**38.9**	41.4	**43.5**	**69.5**	67.8	67.4	**71.2**

even exceeds performance of BCE trained under full labels !

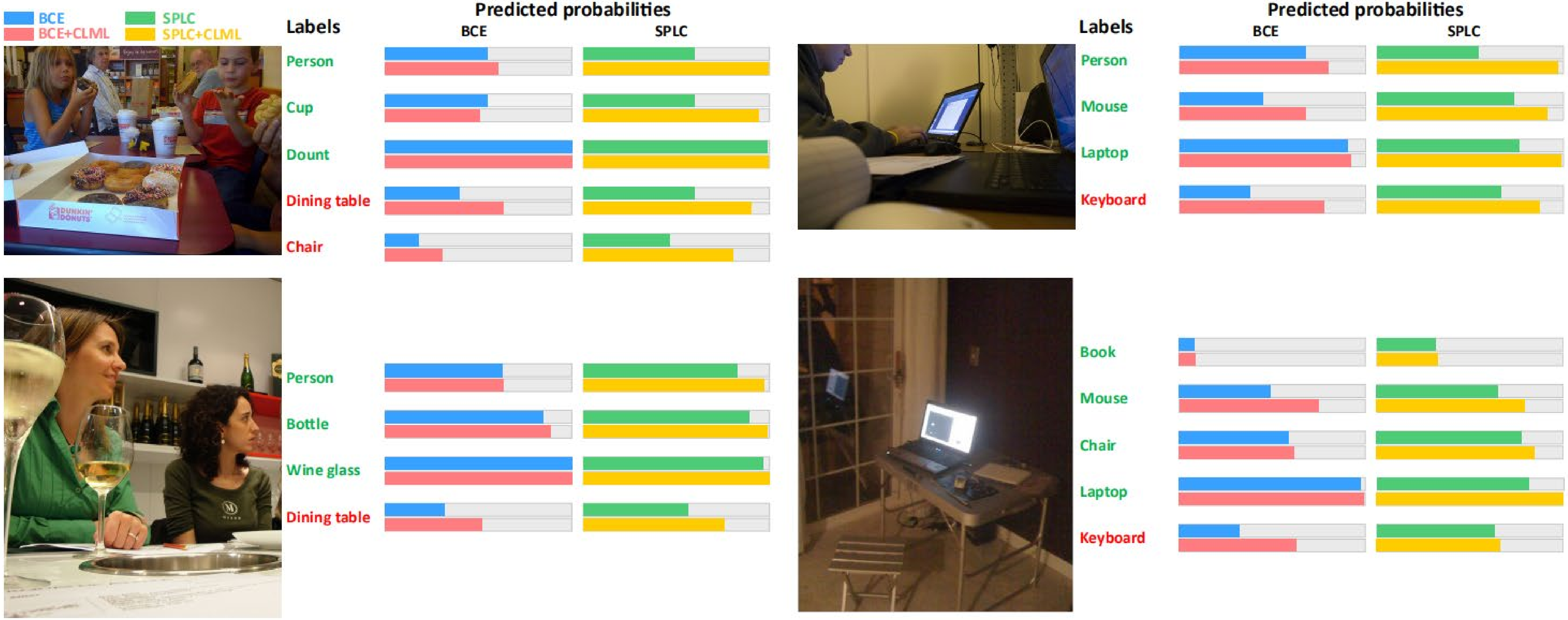
Dataset is annotated by only one label!

Our method outperforms SOTAs.

03 Missing MLL:
High-rankness for Contrastive Learning

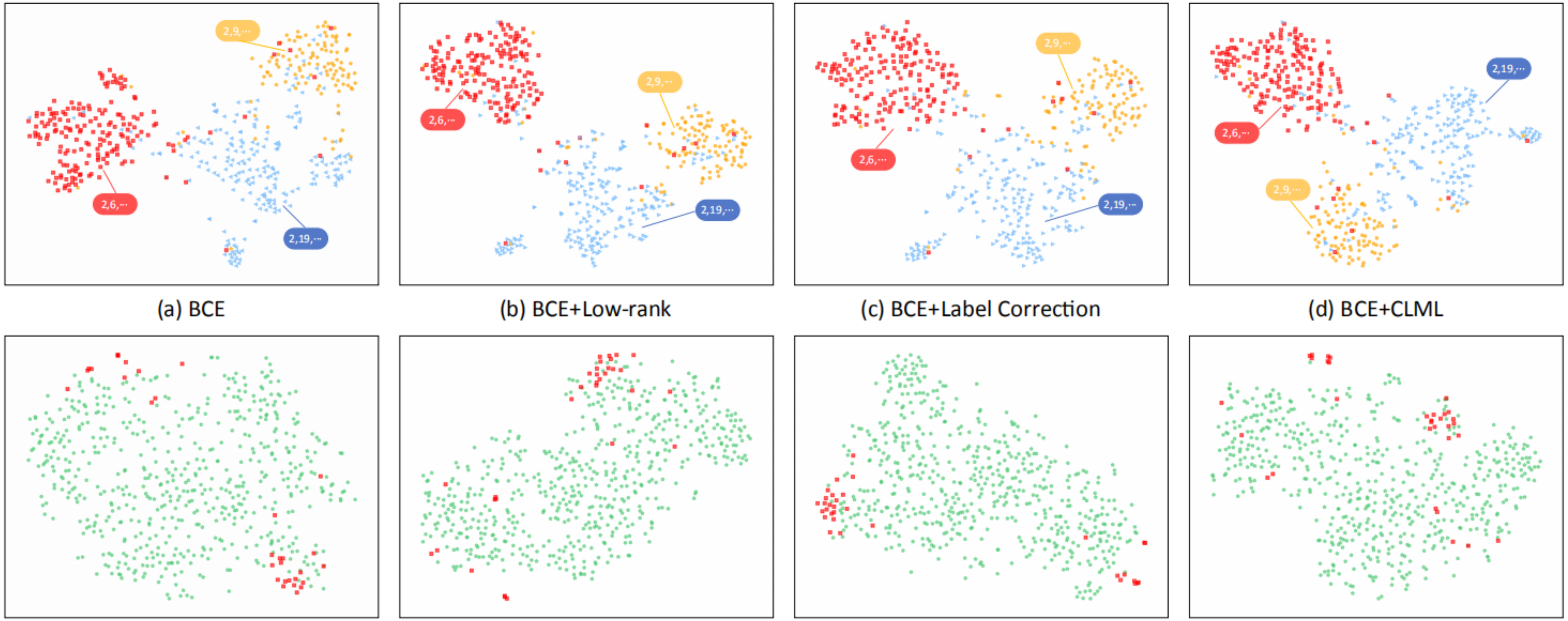
Visualization results:

Our CLML can effectively improve the prediction probability of missing labels:



03 Missing MLL: High-rankness for Contrastive Learning

Ablation Study



(a) BCE (b) BCE+Low-rank (c) BCE+Label Correction (d) BCE+CLML

(e) BCE (f) BCE+Low-rank (g) BCE+Label Correction (h) BCE+CLML

Our CLML has great advantages in **maximizing** **inter-class variance**, **minimizing** **intra-class variance** and **mining missing labels**.

04

Missing MLL:
(Global) High-Rankness on Multi-View

04 Missing MLL: High-Rankness on Multi-View

Why multi-view?

- Objects in **real-world** are often represented by multiple views
- Multi-view multi-label learning is still relatively **under-studied**
- Utilizing multi-view information can **improve performance**

Zhao J, Xie X, Xu X, et al. Multi-view learning overview: Recent progress and new challenges. Information Fusion, 2017.
Huang Y, Du C, Xue Z, et al. What Makes Multimodal Learning Better than Single (Provably). arXiv preprint 2021.

New scenario:

Totally non-aligned multi-view with **incomplete** multi-view.

Definition 1. Given a multi-view multi-label data set Ω, suppose that $\Omega = \{\mathbf{X}^{(i)}\}_{i=1}^{V}$ contains V different views, where $\mathbf{X}^{(i)} = \left[\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \cdots, \mathbf{x}_n^{(i)}\right] \in \mathbb{R}^{n \times d_i}$ is the feature matrix of the i-th view, n and d_i are the numbers of samples and the dimensions of features of the i-th view, respectively. If samples across all views are **totally unpaired**, i.e., the m-th sample of the i-th view $\mathbf{x}_m^{(i)}$ and the m-th sample of the j-th view $\mathbf{x}_m^{(j)}$ are distinct samples, for all $m \in \{1,2,\cdots,n\}$, $i,j \in \{1,2,\cdots,V\}$ and $i \neq j$. Then these views are called **non-aligned views**.

Li X, Chen S. A Concise yet Effective Model for Non-Aligned Incomplete Multi-view and Missing Multi-label Learning. TPAMI, 2022.

04 Missing MLL: High-Rankness on Multi-View

New challenges:

- Explicitly **complementary information** among multi-views can hardly be exploited. (incomplete + non-aligned multi-view)

- **Completion** of the incomplete views is hard to be tractable even if possible. (incomplete + non-aligned multi-view)

- **Label information** for the correspondence among views is quite limited in the MML. (non-aligned multi-view + missing labels)

Difference between full labels and missing labels:

- **Full labels:** The issue of non-aligned multi-view is no more challenging as we can align the views by those shared/common labels.

- **Missing labels:** The shared/common labels are limited in this case.

➡ Correspondence among views is **difficult**.

➡ Problem with non-aligned and incomplete views is **more severe**.

➡ Information about multi-label needs to be further mined.

(Global) Low-rank based MML methods with multi-view:

incomplete **M**ulti-**V**iew **W**eak-label **L**earning (**iMVWL**)

$$\min_{\{\mathbf{U}_v,\mathbf{V},\mathbf{W},\mathbf{S}\}} \sum_{v=1}^{n_v} \left\| \mathbf{O}^v \odot \left(\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T \right) \right\|_F^2, \\ + \alpha \| \mathbf{M} \odot (\mathbf{VWS} - \mathbf{Y}) \|_F^2 + \beta \| \mathbf{S} \|_*$$

where $\mathbf{X}_v \in \mathbb{R}^{n \times d_v}$, $\mathbf{U}_v \in \mathbb{R}^{d_v \times k}$, $\mathbf{V} \in \mathbb{R}^{n \times k}$. $\mathbf{O}^v \in \mathbb{R}^{n \times d_v}$ and $\mathbf{M} \in \mathbb{R}^{n \times c}$ are the indicator matrices for the missing views and labels, $\boldsymbol{S} \in \mathbb{R}^{c \times c}$ is the label correlation matrix.

Tan Q, Yu G, Domeniconi C, et al. Incomplete multi-view weak-label learning, IJCAI. 2018: 2703-2709.

Our starting points:

Samples among views can be bridged *implicitly* through the **common or shared labels**.

Structures (**local low-rank, global high-rank**) of missing multi-labels still hold in the new non-aligned incomplete multi-view setting.

04 Missing MLL: High-Rankness on Multi-View

An intuitive description:

label matrix

			sky	cloud	tree	sea	fish
view 1	A	$x_1^{(1)}$	1	1	-1	-1	-1
		$x_2^{(1)}$	-1	-1	1	1	-1
		$x_3^{(1)}$	-1	-1	-1	1	1
		$x_4^{(1)}$	1	1	-1	1	-1
		$x_5^{(1)}$	1	-1	1	-1	-1
view 2	B	$x_1^{(2)}$	1	1	-1	1	-1
		$x_2^{(2)}$	1	-1	1	-1	-1
		$x_3^{(2)}$	-1	-1	1	1	-1
		$x_4^{(2)}$	1	1	-1	-1	-1
		$x_5^{(2)}$	-1	-1	-1	1	1

rank(C) = rank([A;B])=5

global structure of multiple labels

sub-label matrices

		sky	cloud	tree	sea	fish
sky1	$x_1^{(1)}$	1	1	-1	-1	-1
	$x_4^{(1)}$	1	1	-1	1	-1
	$x_5^{(1)}$	1	-1	1	-1	-1
sky2	$x_1^{(2)}$	1	1	-1	1	-1
	$x_2^{(2)}$	1	-1	1	-1	-1
	$x_4^{(2)}$	1	1	-1	-1	-1

rank(sky) = rank([sky1;sky2])=3

		sky	cloud	tree	sea	fish
cloud1	$x_1^{(1)}$	1	1	-1	-1	-1
	$x_4^{(1)}$	1	1	-1	1	-1
cloud2	$x_1^{(2)}$	1	1	-1	1	-1
	$x_4^{(2)}$	1	1	-1	-1	-1

rank(cloud) = rank([cloud1;cloud2])=2

…

		sky	cloud	tree	sea	fish
fish1	$x_3^{(1)}$	-1	-1	-1	1	1
fish2	$x_5^{(2)}$	-1	-1	-1	1	1

rank(fish) = rank([fish1;fish2])=1

local structure of multiple labels

The global and local structures of the multiple labels

It can be directly generalized to the case of **more than two views.**

Our formulation:

$$\min_{\mathbf{W}^{(i)}} \frac{1}{2}\sum_{i=1}^{V}\left\|\mathbf{P}^{(i)} \odot \left(\mathbf{X}^{(i)}\mathbf{W}^{(i)} - \mathbf{Y}^{(i)}\right)\right\|_F^2$$
$$+ \lambda\left(\underbrace{\sum_{k=1}^{c}\left\|[\mathbf{X}_k^{(1)}\mathbf{W}^{(1)};\mathbf{X}_k^{(2)}\mathbf{W}^{(2)};\cdots;\mathbf{X}_k^{(V)}\mathbf{W}^{(V)}]\right\|_*}_{\text{Local low-rank}}\right.$$
$$\left.\underbrace{-\left\|[\mathbf{X}^{(1)}\mathbf{W}^{(1)};\mathbf{X}^{(2)}\mathbf{W}^{(2)};\cdots;\mathbf{X}^{(V)}\mathbf{W}^{(V)}]\right\|_*}_{\text{Global high-rank}}\right).$$

where $\mathbf{X}^{(i)} \in \mathbb{R}^{n\times d_i}$ is the feature matrix of the i-th view, $\mathbf{X}_k^{(i)}$ is the **sub-matrix** of $\mathbf{X}^{(i)}$ which consists of samples corresponding to the k-th label observed in the i-th view.

The intersection of $\mathbf{X}_k^{(i)}$ w.r.t. k is non-empty.

A theorem:

Theorem 2. Let $\mathbf{X}_k^{(1)}\mathbf{W}^{(1)},\ \mathbf{X}_k^{(2)}\mathbf{W}^{(2)},\ \cdots,\ \mathbf{X}_k^{(V)}\mathbf{W}^{(V)}(k=1,\ 2,\cdots,c)$be matrices with the same column dimension, where $\mathbf{X}_k^{(i)}$ is a sub-matrix of $\mathbf{X}^{(i)}\ (i=1,\ 2,\cdots,V)$. If **(a)**$\forall i \in \{i=1,\ 2,\cdots,V\}$, the vertical concatenation of $\mathbf{X}_1^{(i)}\mathbf{W}^{(i)}$ to $\mathbf{X}_c^{(i)}\mathbf{W}^{(i)}$ contains all rows of $\mathbf{X}^{(i)}\mathbf{W}^{(i)}$ and **(b)** $\forall k,h \in \{i=1,\ 2,\cdots,c\}$, $k \neq h$, at least one of the intersection between $\mathbf{X}_k^{(i)}\mathbf{W}^{(i)}$ and $\mathbf{X}_h^{(i)}\mathbf{W}^{(i)}$ is non-empty, then we have

$$\sum_{k=1}^{c}\left\|\mathbf{X}_k^{(1)}\mathbf{W}^{(1)};\mathbf{X}_k^{(2)}\mathbf{W}^{(2)};\cdots;\mathbf{X}_k^{(V)}\mathbf{W}^{(V)}\right\|$$
$$\geq \left\|\mathbf{X}^{(1)}\mathbf{W}^{(1)};\mathbf{X}^{(2)}\mathbf{W}^{(2)},\cdots;\mathbf{X}^{(V)}\mathbf{W}^{(V)}\right\|_* .$$

➡ Trivial solutions can be **avoided!**

Advantages:

Addressing the three issues: **missing labels, incomplete views, and non-aligned views** simultaneously with just **one** hyper-parameter.

Designing an efficient ADMM algorithm (linear computational complexity w.r.t. the number of samples) which can **handle large scale data**.

Our method (**without view-alignment**) outperforms SOTAs (with view-alignment) on five real datasets.

Compared methods:

iMSF KDD 2012

$$\min_{\beta} \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i} L(x_j^i, y_j^i, \beta_i) + \boxed{\lambda}\sum_{s=1}^{S}\sum_{k=1}^{p_s}\left\|\beta_{I(s,k)}\right\|_2$$

One explicit hyper-parameter

LabelMe IJCAI 2013

$$\min_{Q,H} f = -Tr\left(Y^\top QH\right) + \boxed{\theta_1}Tr\left(HL_wH^\top\right) + \boxed{\theta_2}Tr\left(H^\top Q^\top L_\rho QH\right)$$

Two explicit hyper-parameters

MVL-IV TIP 2015

$$\min_{U,W,Z} \quad \frac{1}{2}\sum_{i=1}^{m}\|U_i\boxed{W} - Z_i\|_F^2$$
$$\text{s.t.} \quad \mathcal{P}_{O_i}(Z_i) = \mathcal{P}_{O_i}(X_i), \quad \forall i \in [1, m].$$

One explicit hyper-parameter (NMF)

lrMMC AAAI 2015

$$\min_{B,P,\theta} \boxed{\mu}\|B\|_* + \|\boxed{P}B - A\|_F^2 + \frac{\boxed{\gamma}}{2}\|\theta\|_2^2$$
$$\text{s.t. } \theta_i \geq 0, \sum \theta_i = 1, i = 1, \ldots, m.$$

Two explicit hyper-parameters

iMVWL IJCAI 2018

$$\min_{\{\mathbf{U}_v,\mathbf{V},\mathbf{W},\mathbf{S}\}} \sum_{v=1}^{n_v}\left\|\mathbf{O}^v \odot \left(\mathbf{X}_v - \boxed{\mathbf{V}}\mathbf{U}_v^T\right)\right\|_F^2, + \boxed{\alpha}\|\mathbf{M} \odot (\mathbf{VWS} - \mathbf{Y})\|_F^2 + \boxed{\beta}\|\mathbf{S}\|_*$$

Two explicit hyper-parameters

Experimental results:

Low-rank based methods

Results of incomplete multi-view (50%) and missing multi-label (50%)

dataset	metrics	lrMMC	MVL-IV	LabelMe	iMSF	iMVWL	NAIM^{3}L
Corel5k	1-HL(%)	95.40(0.00)	95.40(0.00)	94.60(0.00)	94.30(0.00)	97.84(0.02)	**98.70**(0.01)
	1-RL(%)	76.20(0.20)	75.60(0.10)	63.80(0.30)	70.90(0.50)	86.50(0.33)	**87.84**(0.21)
	AP(%)	24.00(0.20)	24.00(0.10)	20.40(0.20)	18.90(0.20)	28.31(0.72)	**30.88**(0.35)
	AUC(%)	76.30(0.20)	76.20(0.10)	71.50(0.10)	66.30(0.50)	86.82(0.32)	**88.13**(0.20)
Pascal07	1-HL(%)	88.20(0.00)	88.30(0.00)	83.70(0.00)	83.60(0.00)	88.23(0.38)	**92.84**(0.05)
	1-RL(%)	69.80(0.30)	70.20(0.10)	64.30(0.40)	56.80(0.00)	73.66(0.93)	**78.30**(0.12)
	AP(%)	42.50(0.30)	43.30(0.20)	35.80(0.30)	32.50(0.00)	44.08(1.74)	**48.78**(0.32)
	AUC(%)	72.80(0.20)	73.00(0.10)	68.60(0.50)	62.00(0.10)	76.72(1.20)	**81.09**(0.12)
ESPGame	1-HL(%)	97.00(0.00)	97.00(0.00)	96.70(0.00)	96.40(0.00)	97.19(0.01)	**98.26**(0.01)
	1-RL(%)	77.70(0.10)	77.80(0.00)	68.30(0.20)	72.20(0.20)	80.72(0.14)	**81.81**(0.16)
	AP(%)	18.80(0.00)	18.90(0.00)	13.20(0.00)	10.80(0.00)	24.19(0.34)	**24.57**(0.17)
	AUC(%)	78.30(0.10)	78.40(0.00)	73.40(0.10)	67.40(0.30)	81.29(0.15)	**82.36**(0.16)
IAPRTC12	1-HL(%)	96.70(0.00)	96.70(0.00)	96.30(0.00)	96.00(0.00)	96.85(0.02)	**98.05**(0.01)
	1-RL(%)	80.10(0.00)	79.90(0.10)	72.50(0.00)	63.10(0.00)	83.30(0.27)	**84.78**(0.11)
	AP(%)	19.70(0.00)	19.80(0.00)	14.10(0.00)	10.10(0.00)	23.54(0.39)	**26.10**(0.13)
	AUC(%)	80.50(0.00)	80.40(0.10)	74.60(0.00)	66.50(0.10)	83.55(0.22)	**84.96**(0.11)
Mirflickr	1-HL(%)	83.90(0.00)	83.90(0.00)	77.80(0.00)	77.50(0.00)	83.98(0.28)	**88.15**(0.07)
	1-RL(%)	80.20(0.10)	80.80(0.00)	77.10(0.10)	64.10(0.00)	80.60(1.11)	**84.40**(0.09)
	AP(%)	44.10(0.10)	44.90(0.00)	37.50(0.00)	32.30(0.00)	49.48(1.24)	**55.08**(0.18)
	AUC(%)	80.60(0.10)	80.70(0.00)	76.10(0.00)	71.50(0.10)	79.44(1.46)	**83.71**(0.06)

iMSF KDD 2012

LabelMe IJCAI 2013

MVL-IV TIP 2015

lrMMC AAAI 2015

iMVWL IJCAI 2018

Our method (**without view-alignment**) outperforms SOTAs (**with view-alignment**).

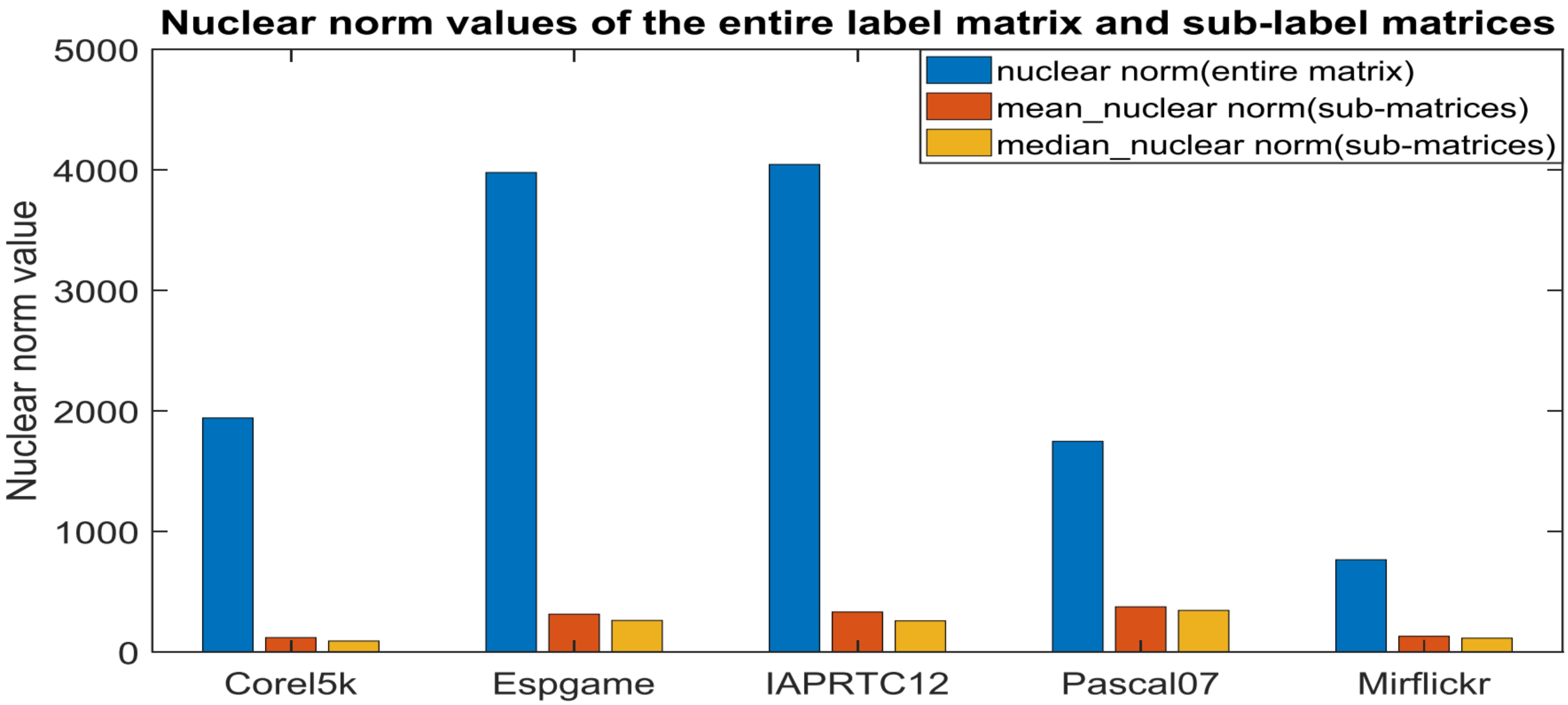
Ablation study:

Without regularization → NAIM³L-I

Only with low-rank term → NAIM³L-II

datasets	metrics	NAIM^{3}L-I	NAIM^{3}L-II	NAIM^{3}L
Corel5k	1-HL(%)	**98.70**(0.00)	**98.70**(0.00)	**98.70**(0.01)
	1-RL(%)	82.73(0.20)	83.54(0.21)	**87.84**(0.21)
	AP(%)	30.20(0.40)	30.47(0.36)	**30.88**(0.35)
	AUC(%)	82.99(0.20)	83.80(0.21)	**88.13**(0.20)
Pascal07	1-HL(%)	92.83(0.00)	92.83(0.00)	**92.84**(0.05)
	1-RL(%)	77.29(0.18)	77.35(0.17)	**78.30**(0.12)
	AP(%)	48.64(0.35)	48.66(0.35)	**48.78**(0.32)
	AUC(%)	79.99(0.17)	80.55(0.17)	**81.09**(0.12)
ESPGame	1-HL(%)	**98.26**(0.00)	**98.26**(0.00)	**98.26**(0.01)
	1-RL(%)	79.63(0.20)	79.80(0.11)	**81.81**(0.16)
	AP(%)	24.28(0.20)	24.34(0.16)	**24.57**(0.17)
	AUC(%)	80.04(0.20)	80.24(0.13)	**82.36**(0.16)
IAPRTC12	1-HL(%)	**98.05**(0.00)	**98.05**(0.00)	**98.05**(0.01)
	1-RL(%)	82.52(0.00)	82.70(0.00)	**84.78**(0.11)
	AP(%)	25.71(0.10)	25.76(0.10)	**26.10**(0.13)
	AUC(%)	82.56(0.10)	82.76(0.10)	**84.96**(0.11)
Mirflickr	1-HL(%)	**88.15**(0.00)	**88.15**(0.00)	**88.15**(0.07)
	1-RL(%)	84.05(0.00)	84.10(0.00)	**84.40**(0.09)
	AP(%)	54.95(0.20)	54.98(0.16)	**55.08**(0.18)
	AUC(%)	83.33(0.00)	83.39(0.00)	**83.71**(0.06)

Validation of high/low ranks:



05

Summary

New findings:

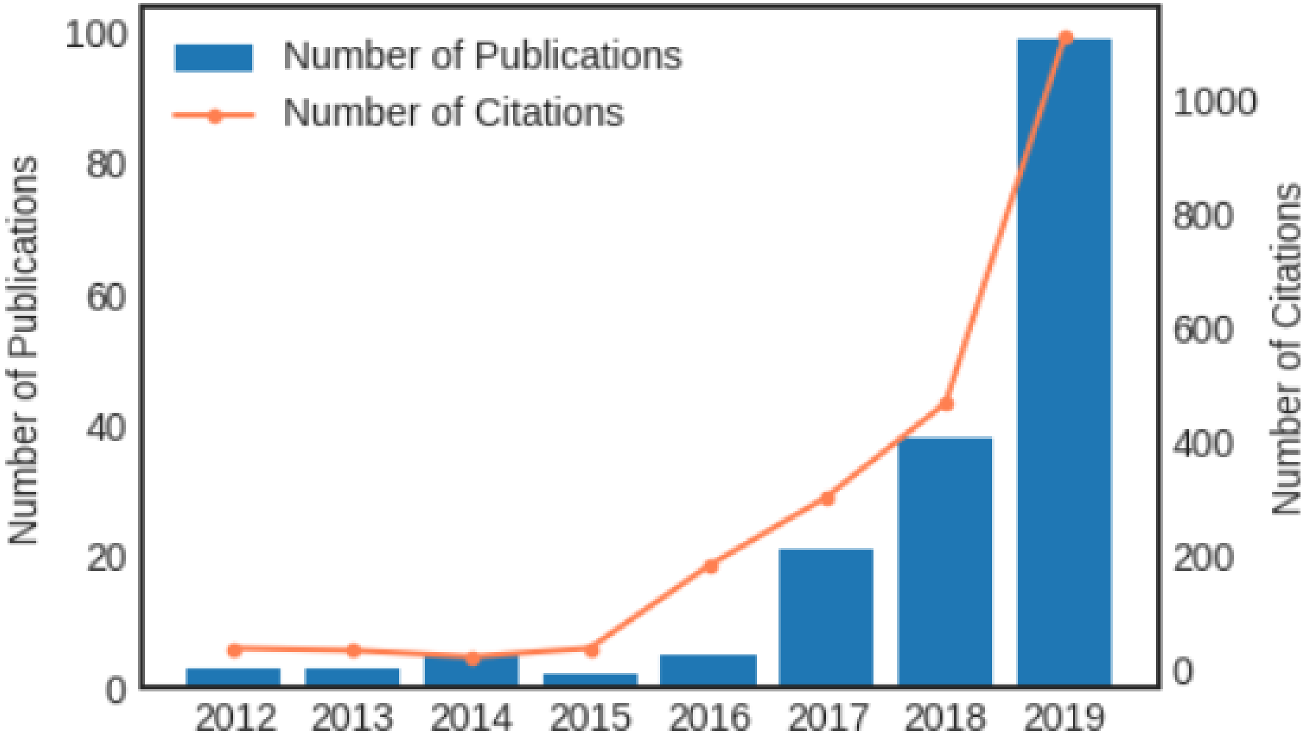
① The **global high-rankness** of multi-label matrix.

② The new **non-aligned** multi-view scenario.

③ Effective **Label Structure Preserving Contrastive** Embedding for MML.

Future direction:

Self-supervised learning: publications + citations.



Huge attention!

Yann Lecun. Self-Supervised Learning. AAAI-2020.

Future direction:

<u>High-order Multi-Label Contrastive Learning</u>.

Supervised contrastive learning. (Multi-class) √

Multi-label contrastive learning. (Multi-label) √

High-order Multi-label contrastive learning. ?

P. Khosla, P. Teterwak, et al. , Supervised contrastive learning, NerIPS 2020. 3004
J. Song and S. Ermon. Multi-label contrastive predictive coding, NerIPS 2020. 47
ZC Ma et al, Label Structure Preserving Contrastive Embedding **for Multi-Label Learning with Missing Labels**, TIP Major Revision
Zhang, Shu, et al. Use All The Labels: A Hierarchical Multi-Label Contrastive Learning Framework. CVPR. 2022. 27

Challenges:

- Directly utilizing contrastive learning can **hardly improve** performance in multi-label case.
- The **high-order label correlation** of multi-labels makes it **more difficult** to define contrastive pairs.

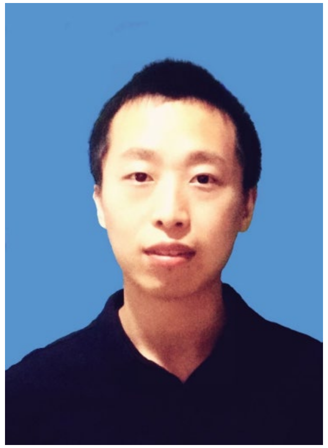
Possible solutions:

- Contrastive learning + label completion (**high-order high-rank**)
- **Semi/weakly-supervised** contrastive learning

Dao S D, Zhao E, et al. Contrast Learning Visual Attention for Multi Label Classification. arXiv 2107.11626, 3
Mingkai Zheng, etal, Weakly supervised contrastive learning, ICCV2021. 74

05 Summary

Songcan Chen
陈松灿



Zhongchen Ma
马忠臣

Xiang Li
李想

Our website: **http://parnec.nuaa.edu.cn/**

Ma Z, Chen S. Expand globally, shrink locally: Discriminant multi-label learning with missing labels[J]. Pattern Recognition, 2021

Code is available at https://github.com/John986/Multi-label-Learning-with-Missing-Labels

Li X, Chen S. A Concise yet Effective Model for Non-Aligned Incomplete Multi-view and Missing Multi-label Learning[J]. TPAMI 2021

Code is available at https://github.com/EverFAITH/NAIM3L

ZC Ma et al, Label Structure Preserving Contrastive Embedding for Multi-Label Learning with Missing Labels, TIP, Major revision.
Code is available at https://github.com/chuangua/ContrastiveLossMLML

THANK YOU!

Q & A