SIGIR 2016

Pisa, Italy

Self-Paced Cross-Modal Subspace Learning

Jian Liang, Zhihang Li, Dong Cao, Ran He

Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, CASIA, Beijing, China

Jingdong Wang

Microsoft Research
Asia, Beijing, China

CRIPAC 智能感知与计算研究中心 Center for Research on Intelligent Perception and Computing

CASIA

THE CHINESE ACADEMY OF SCIENCES

Microsoft Research 微软亚洲研究院

Outline

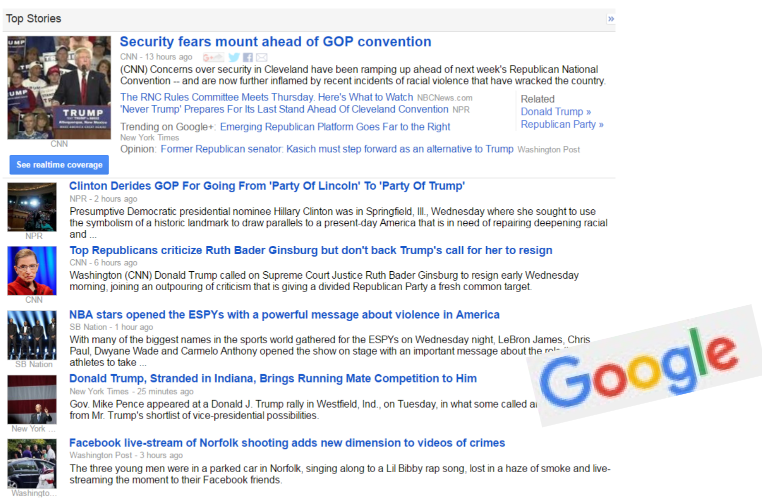
- Background
- Self-Paced Learning
- Proposed Method
- Experiments
- Conclusions

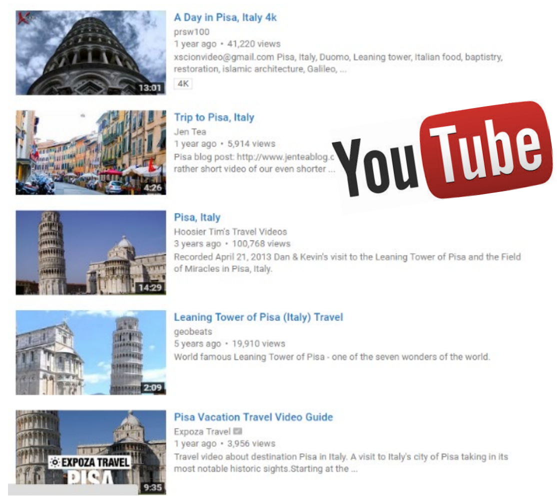
Outline

- Background
- Self-Paced Learning
- Proposed Method
- Experiments
- Conclusions

Motivation

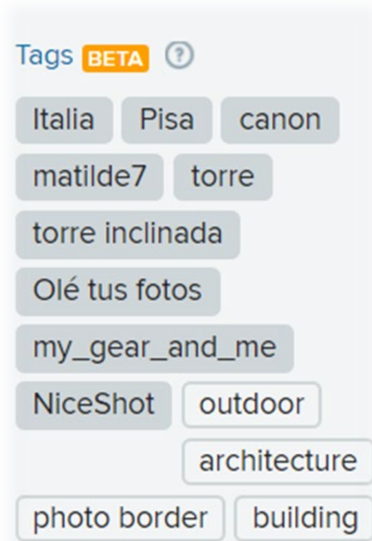
Massive explosion of rich “content” emerges on the web.





Here we main focus on **visual** and **textual** descriptions for cross-modal learning.

Problem Definition (cont’d)

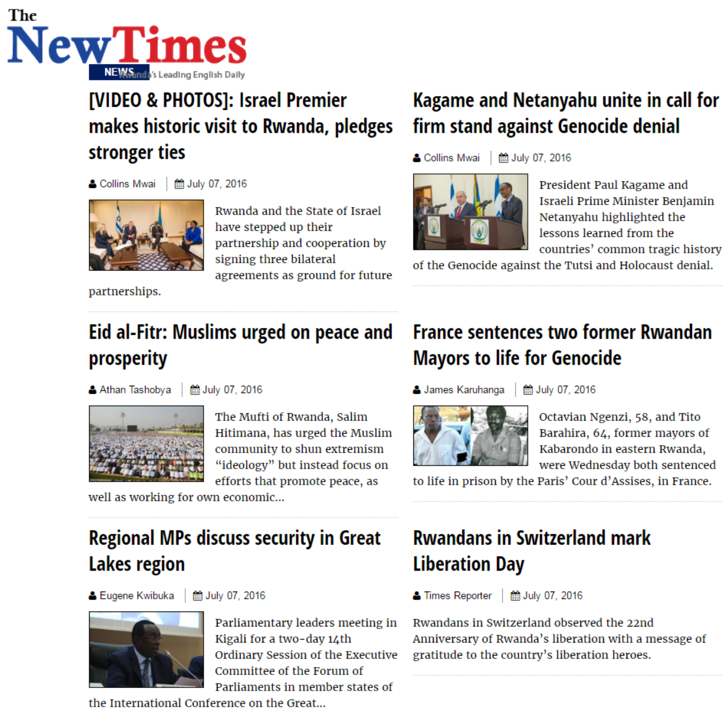


“Pisa”– text cues – Pisa

If no such exact tags exists for these images, how we can find them?

“Pisa”– (implicitly) – Pisa

Problem Definition


The NewTimes
NEWS Rwanda's Leading English Daily

[VIDEO & PHOTOS]: Israel Premier makes historic visit to Rwanda, pledges stronger ties

Collins Mwai | July 07, 2016

Rwanda and the State of Israel have stepped up their partnership and cooperation by signing three bilateral agreements as ground for future partnerships.

Kagame and Netanyahu unite in call for firm stand against Genocide denial

Collins Mwai | July 07, 2016

President Paul Kagame and Israeli Prime Minister Benjamin Netanyahu highlighted the lessons learned from the countries' common tragic history of the Genocide against the Tutsi and Holocaust denial.

Eid al-Fitr: Muslims urged on peace and prosperity

Athan Tashobya | July 07, 2016

The Mufti of Rwanda, Salim Hitimana, has urged the Muslim community to shun extremism "ideology" but instead focus on efforts that promote peace, as well as working for own economic...

France sentences two former Rwandan Mayors to life for Genocide

James Karuhanga | July 07, 2016

Octavian Ngenzi, 58, and Tito Barahira, 64, former mayors of Kabarondo in eastern Rwanda, were Wednesday both sentenced to life in prison by the Paris' Cour d'Assises, in France.

Regional MPs discuss security in Great Lakes region

Eugene Kwibuka | July 07, 2016

Parliamentary leaders meeting in Kigali for a two-day 14th Ordinary Session of the Executive Committee of the Forum of Parliaments in member states of the International Conference on the Great...

Rwandans in Switzerland mark Liberation Day

Times Reporter | July 07, 2016

Rwandans in Switzerland observed the 22nd Anniversary of Rwanda's liberation with a message of gratitude to the country's liberation heroes.

Query

...ur, was accompanied by his wife Sara Netanyahu, and a delegation o... overnment officials and business leaders from Israel.

...he three agreements are in the areas of joint declaration of intent on ...nnovation, visa exemption for holders of diplomatic passports, and ...int declaration in the field of agriculture.

...ddressing reporters after the signing, President Kagame said the ...ooperation between the two states was partly informed by the...

× × × √

Which pictures do we mostly want?

Given one sentence for example, if these news/videos/images above have no associated textual tags, how can we discover the most **related** heterogeneous content?

Outline

- Background
- Self-Paced Learning
- Proposed Method
- Experiments
- Conclusions

Cognitive Science

How human/animal learns: First input easy samples and gradually involve more into training from easy to complex

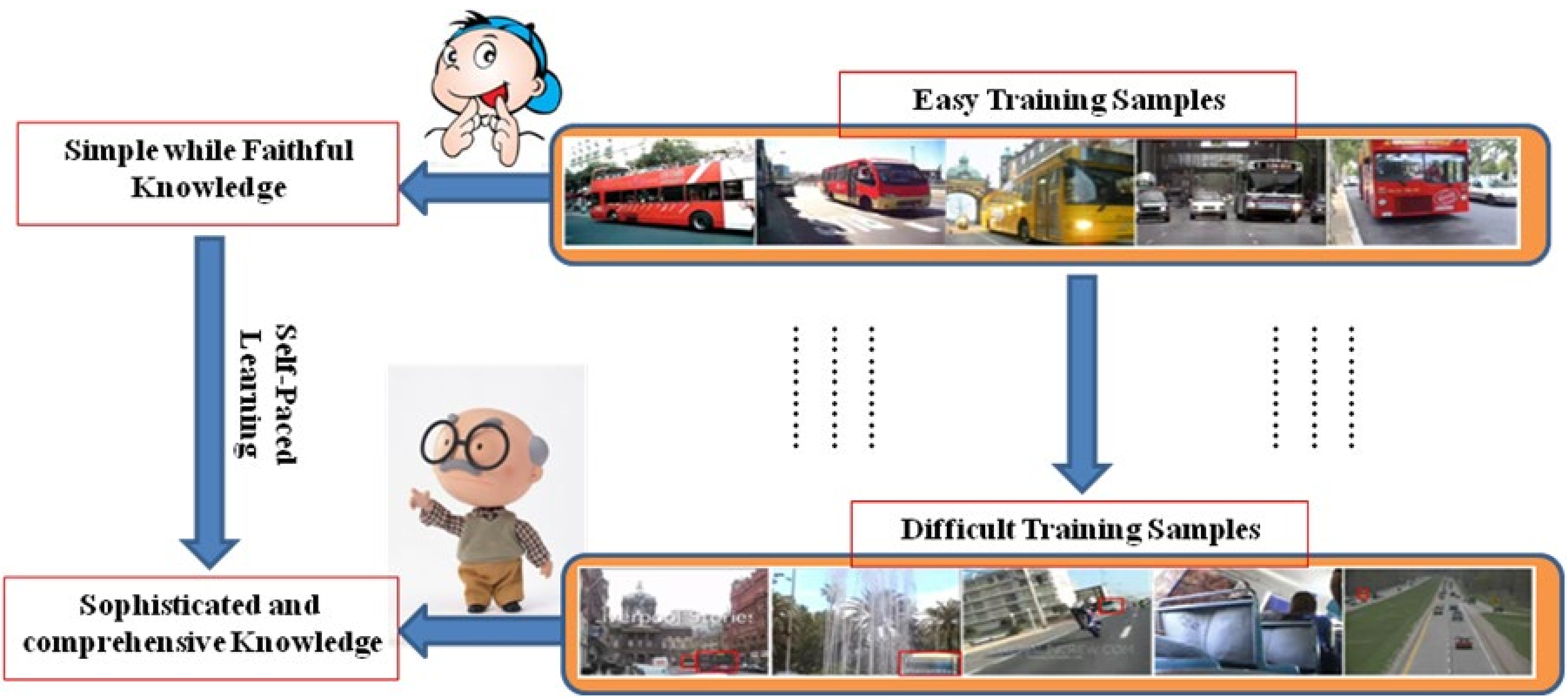


Image taken from http://gr.xjtu.edu.cn/web/dymeng/6

History

Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al 2010) is a recently proposed learning paradigm that is inspired by the learning process of humans and animals.

The samples are not learned randomly but organized in a meaningful order which illustrates from **easy** to gradually more **complex** ones.

Easy sample ➡ **smaller loss** to the already learned model.

Complex sample ➡ **bigger loss** to the already learned model.

Basic Self-Paced Model

$$\min_{w,v\in[0,1]^n} E(w,v;\lambda) = \boxed{\sum_{i=1}^{n} v_i L\big(y_i, g(x_i,w)\big)} + p(w) + f(\mathbf{v};\lambda)$$

where

Loss function term

$$f(\mathbf{v};\lambda) = -\boldsymbol{\lambda}\sum_i v_i$$

One of the most simplified self-paced regularizers proposed in (Kumar et al. 2010)

Following works (Jiang et al. 2015, Zhao et al. 2015) proposed more extension of explicit self-paced regularizers, and implicit regularizers are further investigated in (Fan et al. 2016) .

Basic Self-Paced Model (cont'd)

$$\min_{w,v\in[0,1]^n} E(w,v;\lambda) = \boxed{\sum_{i=1}^{n} v_i L\big(y_i, g(x_i,w)\big)} + p(w) + f(\mathbf{v};\lambda)$$

Loss function term

where

$$f(\mathbf{v};\lambda) = -\boldsymbol{\lambda}\sum_i v_i$$

One of the most simplified self-paced regularizers proposed in (Kumar et al. 2010)

Optimization Algorithm (Alternating Search)

- Fixing w,

$$v_i = \begin{cases} 1 & \text{if } \ell_{\mathrm{i}} \le \lambda, \\ 0 & \text{if } \ell_{\mathrm{i}} > \lambda. \end{cases}$$

- Fixing v, it turns out to be a standard classification sub-problem.

Basic Self-Paced Model (cont'd)

$$\min_{w,v\in[0,1]^n} E(w,v;\lambda) = \boxed{\sum_{i=1}^{n} v_i L\big(y_i, g(x_i,w)\big)} + p(w) + f(\mathbf{v};\lambda)$$

Loss function term

where

$$f(\mathbf{v};\lambda) = -\boldsymbol{\lambda}\sum_{i} v_i$$

One of the most simplified self-paced regularizers proposed in (Kumar et al. 2010)

Expected Advantages:

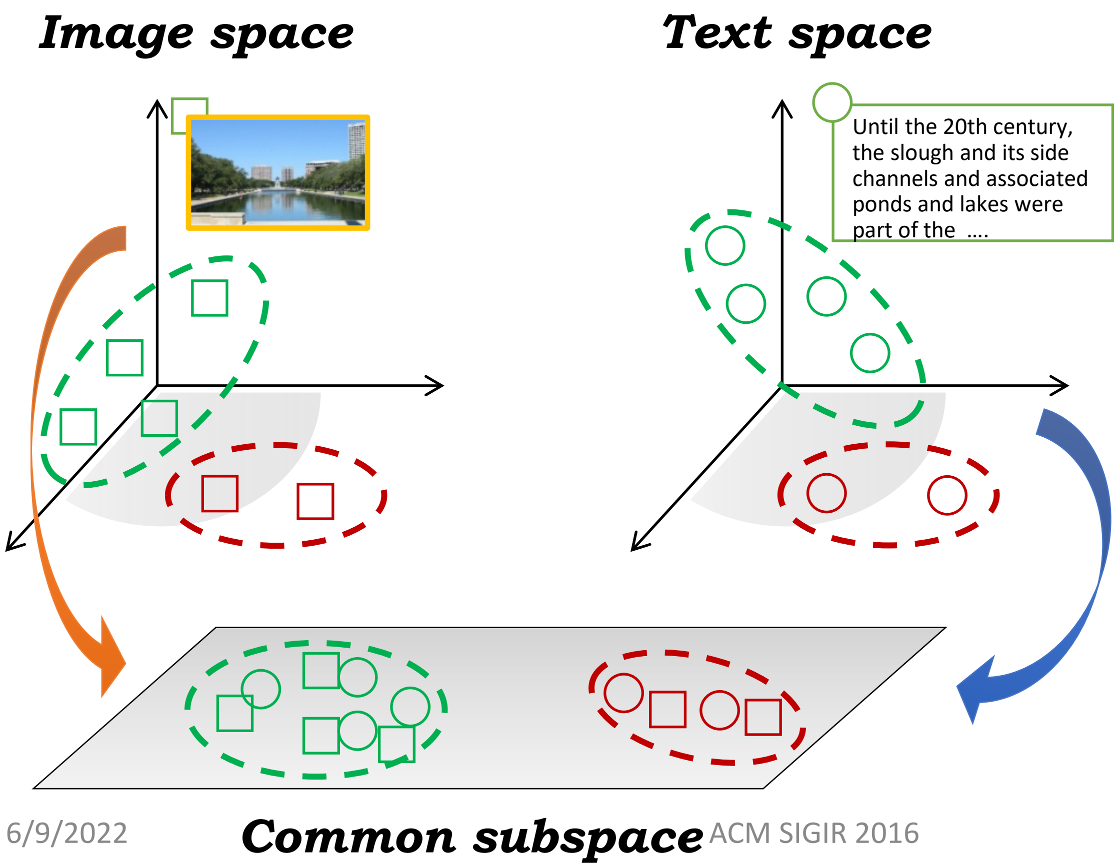
- ✓ Help find a better local minima (as a regularizer)
- ✓ Speed the convergence of training towards the global minimum (for convex problem)

Outline

- Background
- Self-Paced Learning
- Proposed Method
- Experiments
- Conclusions

Prior Work

Probability based learning algorithms, subspace based learning algorithms and deep learning based algorithms have been developed to alleviate this gap (Wang et al. 2015).



Supervised method

GMLDA (Sharma et al. 2012)

- Intra-class and Inter-class constraints

Unsupervised method

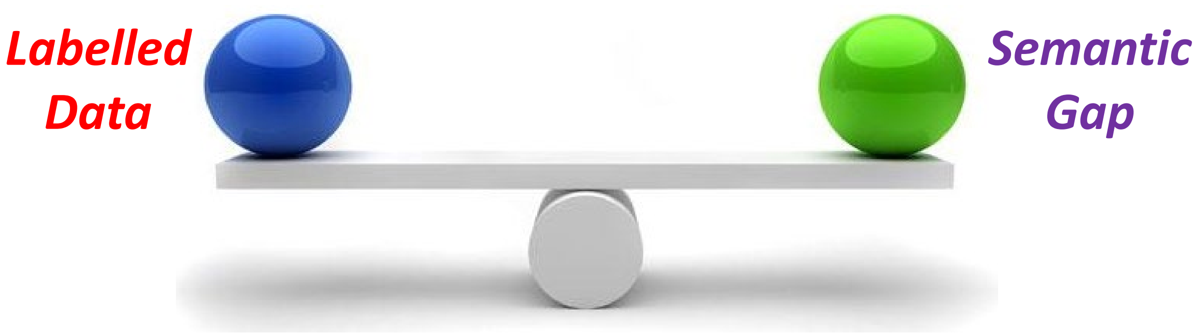
CCA (Hardoon et al. 2004)

- Pairwise constraints

Prior Work (cont'd)

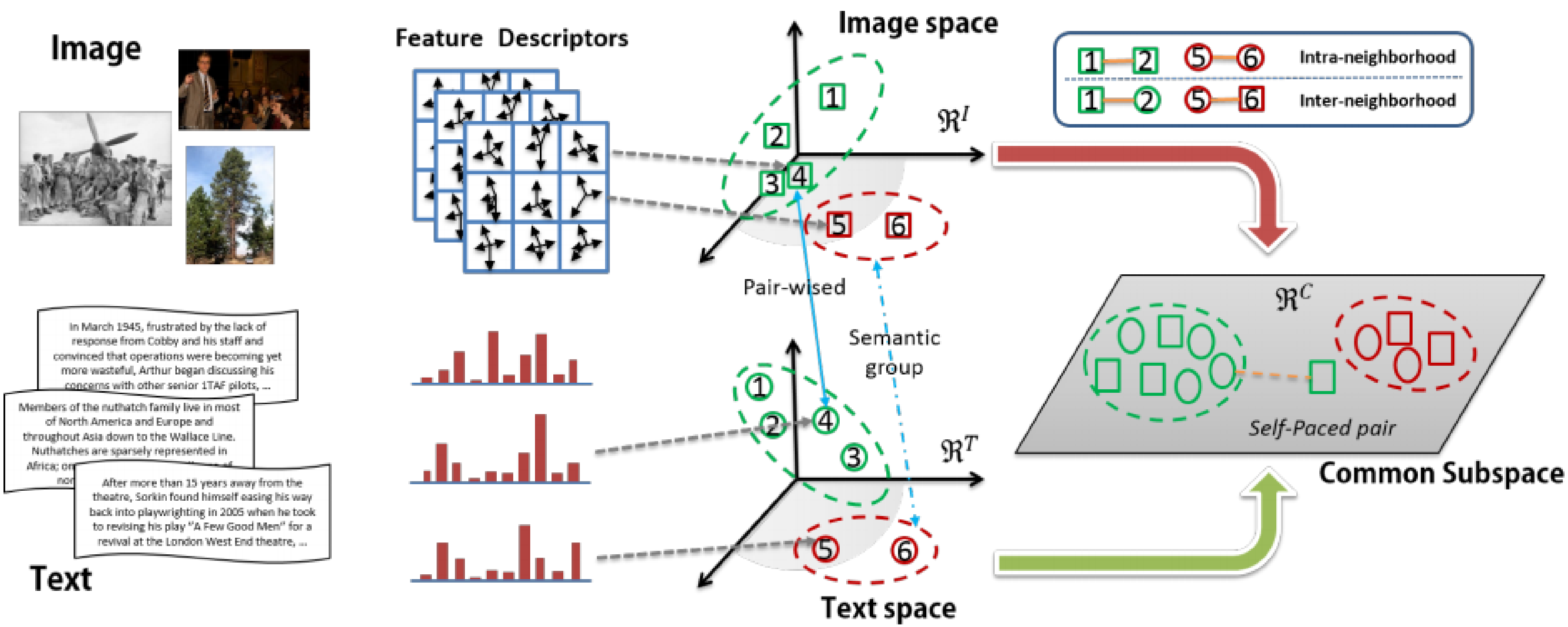
- Summary

Category	Methods	Merits	Demerits
Unsupervised methods	CCA, PLS, BLM etc.	No command for label information	Not effectively alleviate the semantic gap
Supervised methods	CDFE, GMLDA, GMMFA, LCFS etc.	Using label information to bridge the semantic gap	Need lots of labelled data



Framework

- Semantic labels are expensive;
- In the semantic level, hard heterogeneous pairs and mismatching pairs exist.



Formulation

$$\min_{\mathbf{U}_a,\mathbf{U}_b} \boxed{\|\mathbf{U}_a^T\mathbf{X}_a - \mathbf{U}_b^T\mathbf{X}_b\|_F^2} + \Phi(\mathbf{U}_a, \mathbf{U}_b)$$

$$\Rightarrow \quad \min_{\mathbf{U}_a,\mathbf{U}_b,\mathbf{Y}} \boxed{\sum_{p\in\{a,b\}} ||\mathbf{U}_p^T\mathbf{X}_p - \mathbf{Y}\|_F^2} + \Phi(\mathbf{U}_a, \mathbf{U}_b)$$

Pseudo Labels

$$s.t.\ \mathbf{Y} \in \{0,1\}^{c\times n}, \sum_i Y_{i,j} = 1, \forall j \in [1, n].$$

$$\Rightarrow \quad \min_{\mathbf{U}_a,\mathbf{U}_b,\mathbf{Y},v} \sum_{p\in\{a,b\}} \sum_{i=1}^{n} v_i\, \ell_{p,i} + \beta \sum_{p\in\{a,b\}} ||\mathbf{U}_p||_F^2 + f(\mathbf{v}; k)$$

Self-Paced Manner

$$s.t.\ \mathbf{Y} \in \{0,1\}^{c\times n}, \sum_i^{c} Y_{i,j} = 1, \forall j \in [1, n].$$

Formulation (cont'd)

Multimodal Locality Preserving Term

$$\min_{\mathbf{U}_a,\mathbf{U}_b} \|\mathbf{U}_a^T\mathbf{X}_a - \mathbf{U}_b^T\mathbf{X}_b\|_F^2 + \boldsymbol{\Phi}(\mathbf{U}_a,\mathbf{U}_b)$$

Similarity Matrices

$$W = \begin{bmatrix} \gamma W^a & W^{ab} \\ W^{ba} & \gamma W^b \end{bmatrix}$$

$$\begin{aligned}\boldsymbol{\Phi}(\mathbf{U}_a,\mathbf{U}_b) &= \mathcal{L}_{inter} + \gamma\mathcal{L}_{intra} \\ &= \sum_{p\in\{a,b\}} tr(\mathbf{U}_p^T\mathbf{X}_p\mathbf{L}_{pq}\mathbf{X}_q^T\mathbf{U}_q)\end{aligned}$$

Intra-modal Similarity

Gaussian kernel function

$$W_{ij}^p = \begin{cases} d(\mathrm{x}_p^i, \mathrm{x}_p^j), & \mathrm{x}_p^j \in N_r\left(\mathrm{x}_p^i\right) or\ \mathrm{x}_p^i \in N_r\left(\mathrm{x}_p^j\right) \\ 0, & otherwise. \end{cases}$$

Inter-modal Similarity

$$W^{ab} = W^{ba} = \mathbf{Y}^T\mathbf{Y}$$

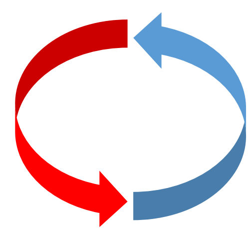
Optimization Algorithm

$$\min_{\mathbf{U}\{a,b\},\mathbf{v},\mathbf{Y}} \sum_{p\in\{a,b\}} \sum_{i=1}^{n} v_i\, \ell_{p,i} + \alpha\mathcal{L}(\mathbf{U}_a, \mathbf{U}_b) + \beta \sum_{p\in\{a,b\}} ||\mathbf{U}_p||_F^2 + f(\mathbf{v}; k)$$

$$s.t.\ \mathbf{Y} \in \{0,1\}^{c\times n}, \sum_{i}^{c} Y_{i,j} = 1, \forall j \in [1, n].$$

Alternating Minimization Methods

- ☐ *Initialing v and Y;*
- ☐ *Fixing Y and v, solve U_a, U_b;*
- ☐ *Fixing U_a, U_b and v, solve Y;*
- ☐ *Fixing U_a, U_b and Y, solve v;*
- ☐ *Update similarity matrices W and parameter $k \leftarrow u_* k$*



Optimization Algorithm (cont'd)

a) Solve U_a and U_b , when **Y**, **v** *are fixed*.

$$\min_{\mathbf{U}_p} ||(\mathbf{U}_p^T \mathbf{X}_p - \mathbf{Y})\mathbf{V}||_F^2$$
$$+ \alpha \sum_p \sum_q Tr(\mathbf{U}_p^T \mathbf{X}_p \mathbf{L}_{pq} \mathbf{X}_q^T \mathbf{U}_q) + \beta ||\mathbf{U}_p||_F^2$$

Convex subproblem

Differentiating the objective function with respect to $\mathbf{U}_p$ and setting it to zero.

Optimization Algorithm (cont'd)

b) Solve Y, when $\mathbf{U}_a, \mathbf{U}_b$, v are fixed.

$$\min_{\mathbf{Y}} \sum_{p \in \{a,b\}} ||(\mathbf{U}_p^T \mathbf{X}_p - \mathbf{Y})\mathbf{V}||_F^2 + 2\alpha Tr(\mathbf{U}_a^T \mathbf{X}_a \mathbf{Y}^T \mathbf{Y} \mathbf{X}_b^T \mathbf{U}_b)$$

$$s.t. \quad \mathbf{Y} \in \{0,1\}^{c \times n}, \sum_{i}^{c} Y_{i,j} = 1, \forall j \in [1, n].$$

Y has the discrete constraints, resulting in a NP hard problem. Inspired by (Shen et al. 2015), we can optimize Y column by column, i.e., optimize one column of Y with all the other columns fixed.

c) Solve v, when $\mathbf{U}_a, \mathbf{U}_b$, Y are fixed.

Compute the loss of each sample and determine v.

Outline

- Background
- Self-Paced Learning
- Proposed Method
- Experiments
- Conclusions

Datasets

Dataset	# Class	# Training/test instances	# Image/text features
VOC	20	2,808/2,841	512/399
Wiki	10	1,300/1,566	128/10
Wiki++	10	1,300/1,566	4,096/5,000
LabelMe	8	1,600/1,086	512/470

Wiki

After more than 15 years away from the theatre, Sorkin found himself easing his way back into playwrighting in 2005 when he took to revising his play "A Few Good Men" for a revival at the London West End theatre, the Haymarket. It had been a while since he had originally written the play and so he gave it a polish.

Chelsea ‘s highest appearance-maker is ex-captain Ron Harris, who played in 795 first-class games for the club between 1961 and 1980.For the appearance and goalscoring records of all Chelsea players, see This record is unlikely to be broken in the near future; Chelsea’ s current highest appearance-maker is Frank Lampard …

VOC

'person'
'person'
'frame'
'frame'
'tvmonitor'
'table'
'plate'
'tvstand'

'horse'
'person'
'mountain'
'grass'

Evaluation Criteria

- *Mean Average Precision (MAP)*
- *Precision-Scope curve*
- *Precision-Recall curve*

- *The cosine distance is utilized to measure the similarity of projected feature pairs.*
- *The higher MAP indicates the better performance.*

Baseline Methods

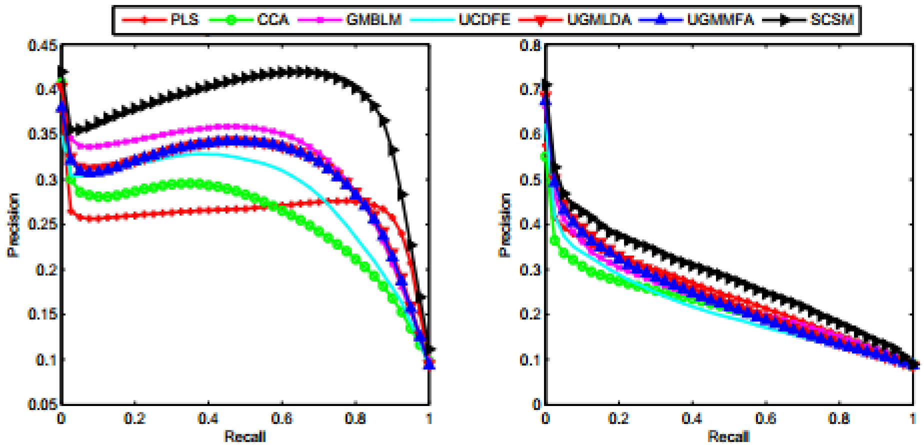
- *Unsupervised methods*
  - *CCA (Hardoon et al. 2004), PLS (Sharma and Jacobs 2011), GMBLM (Sharma et al. 2012)*
  - *UCDFE, UGMLDA, UGMMFA (the unsupervised version of CDFE, GMLDA and GMMFA, which assume the clustering result as pseudo semantic labels)*
- *Supervised methods*
  - *CDFE (Lin and Tang 2006), GMLDA (Sharma et al. 2012), GMMFA (Sharma et al. 2012), LCFS (Wang et al. 2013) and JFSSL (Wang et al. 2015)*

Performances

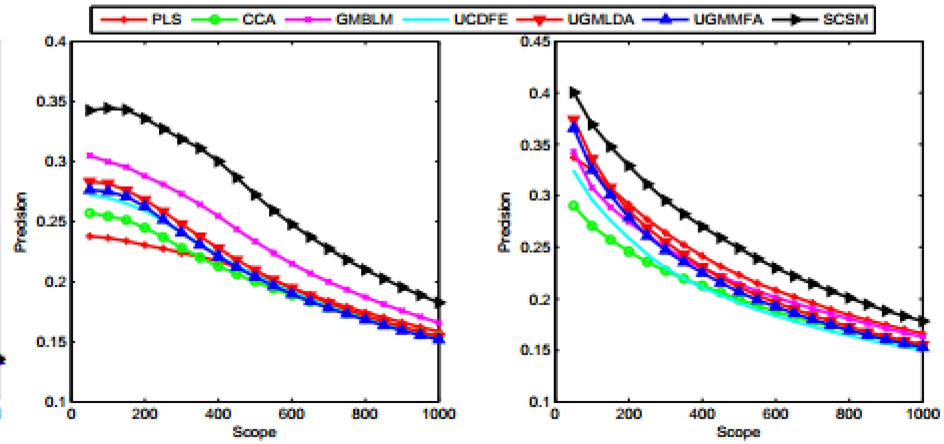
- *MAP scores of **unsupervised** SCSM and other methods on all four datasets.*

Dataset & Methods	Pascal VOC			Wiki			Wiki++			LabelMe		
	Image	Text	Avg	Image	Text	Avg	Image	Text	Avg	Image	Text	Avg
CCA	0.250	0.212	0.231	0.251	0.199	0.225	0.347	0.310	0.329	0.268	0.236	0.252
PLS	0.256	0.241	0.249	0.262	0.174	0.218	0.304	0.329	0.317	0.522	0.435	0.478
GMBLM	0.312	0.232	0.272	0.255	0.204	0.229	0.347	0.318	0.333	0.515	0.466	0.490
UCDFE	0.279	0.209	0.244	0.224	0.184	0.204	0.333	0.301	0.317	0.570	0.594	0.582
UGMMFA	0.298	0.232	0.265	0.269	0.211	0.240	0.340	0.310	0.325	0.512	0.499	0.505
UGMLDA	0.301	0.239	0.270	0.272	0.215	0.244	0.340	0.318	0.325	0.542	0.536	0.539
CDFE	0.306	0.227	0.267	0.260	0.209	0.234	0.397	0.344	0.370	0.685	0.725	0.705
GMMFA	0.327	0.259	0.293	0.273	0.219	0.246	0.409	0.362	0.386	**0.719**	**0.724**	**0.722**
GMLDA	0.324	0.260	0.292	0.273	0.218	0.246	0.409	0.362	0.386	0.716	0.720	0.718
LCFS [35]	0.344	0.267	0.306	0.280	0.214	0.247	0.413	0.384	0.404	-	-	-
JFSSL [35]	**0.361**	**0.280**	**0.320**	**0.306**	**0.228**	**0.267**	**0.428**	**0.396**	**0.412**	-	-	-
SCSM	**0.375**	**0.282**	**0.329**	**0.274**	**0.217**	**0.245**	**0.423**	**0.381**	**0.402**	**0.641**	**0.672**	**0.656**

Performances (cont'd)

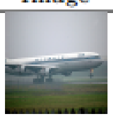
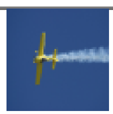
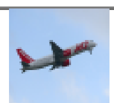
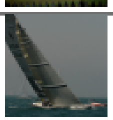
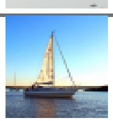
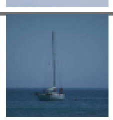
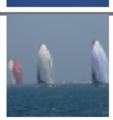
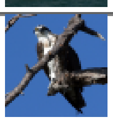
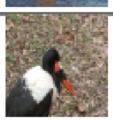
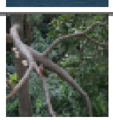
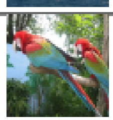
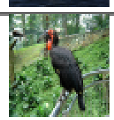
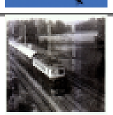


(a) precision-recall curve

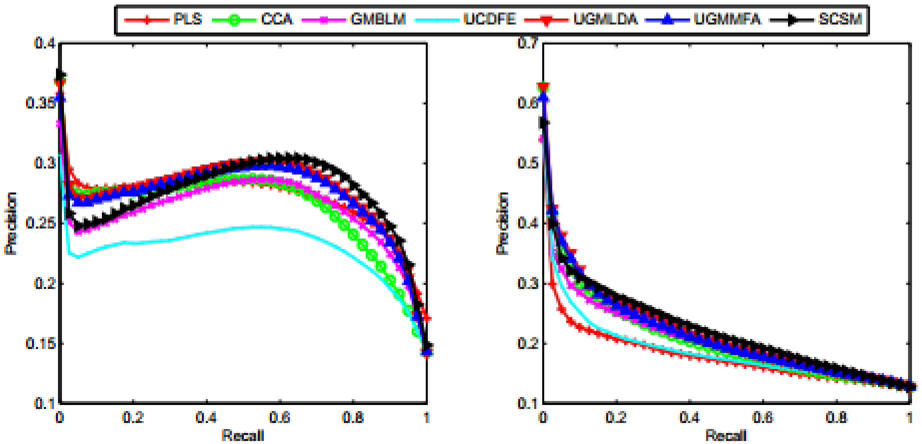


(b) precision-scope curve

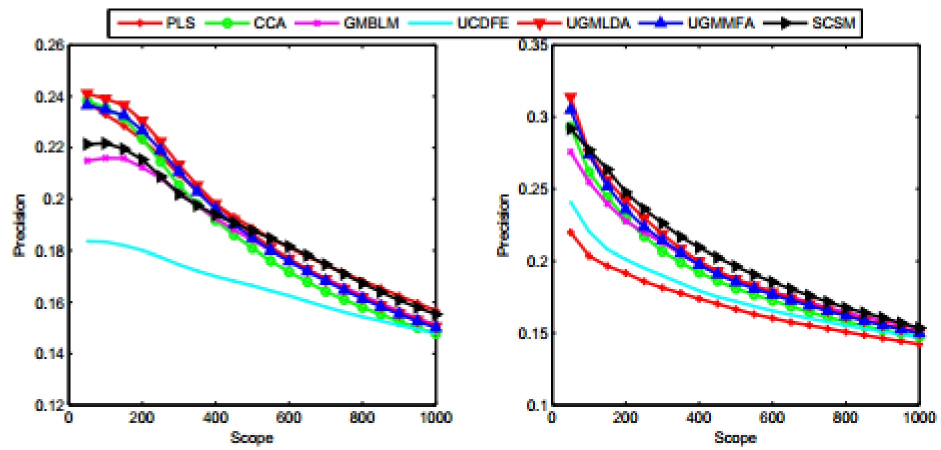
- ➢ *Retrieval examples by user tags queries on the* *PASCAL-VOC dataset* *by the proposed SCSM.*

Class	Tags queries	True Image	Matching Images			
aeroplane	aeroplane grass door window wheel					
boat	boat rigging water sky wave					
bird	bird tree sky					
train	train rope railroad pole tree					

Performances (cont'd)

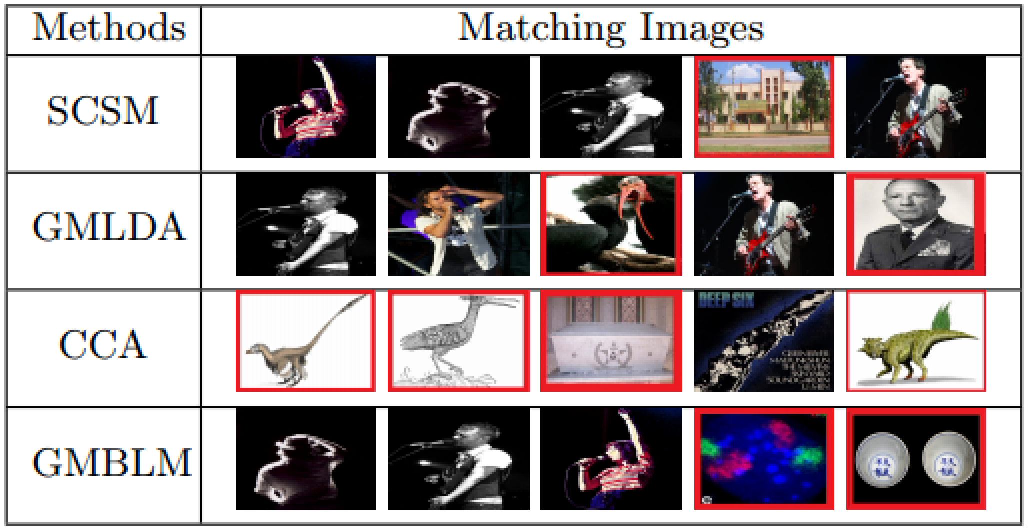


(a) precision-recall curve

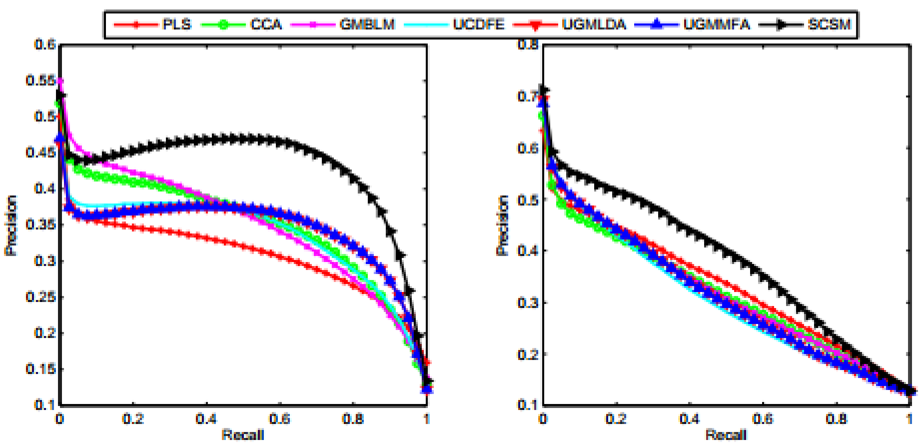


(b) precision-scope curve

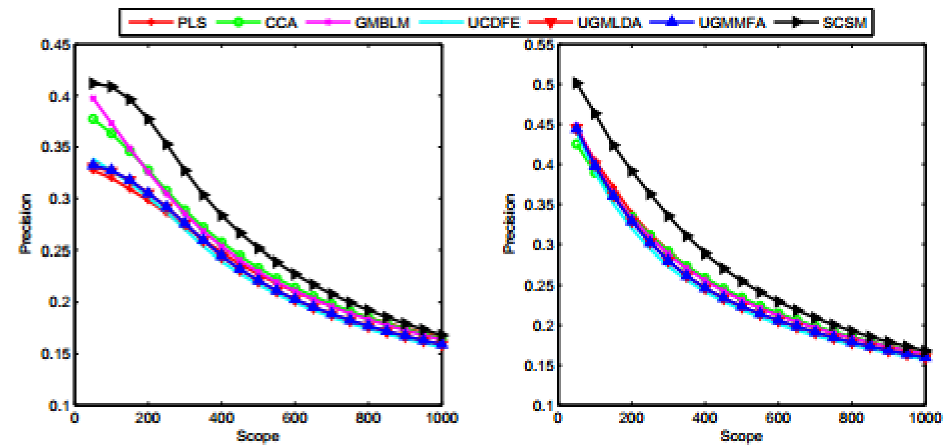
- *Retrieval examples by text query "In 1994, Angus and ……" on the* ***Wiki dataset*** *by the proposed SCSM.*



Performances (cont'd)

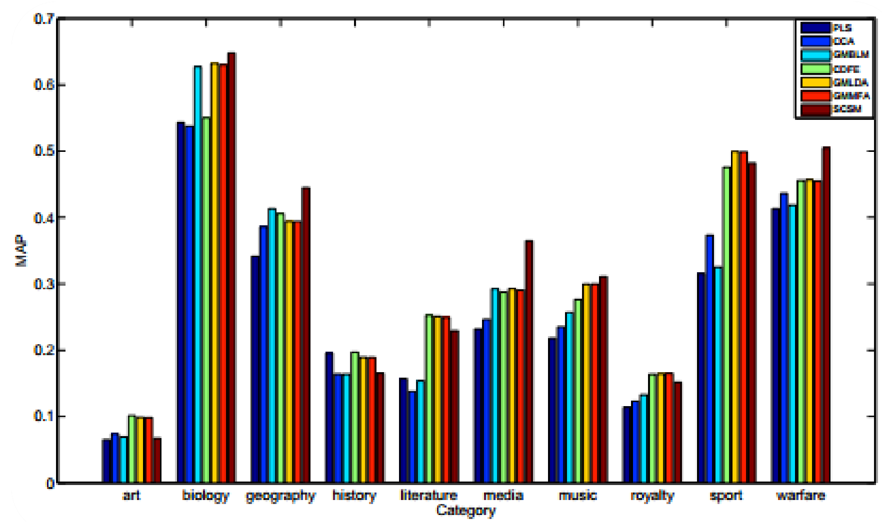


(a) precision-recall curve

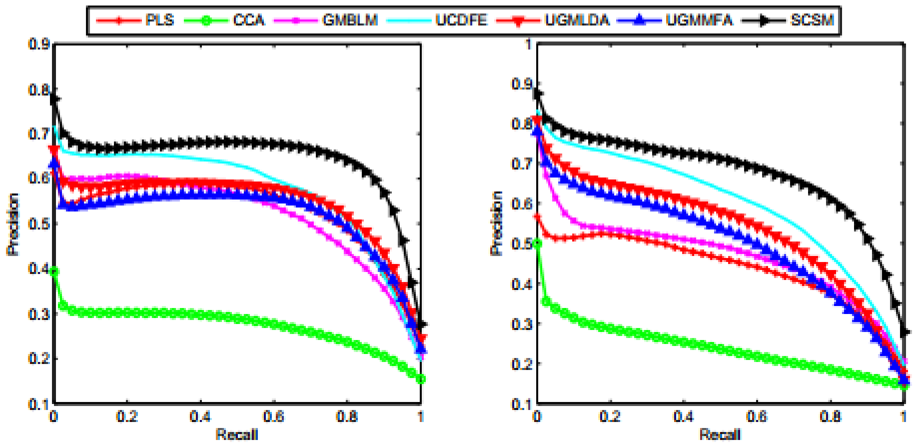


(b) Precision-scope curve

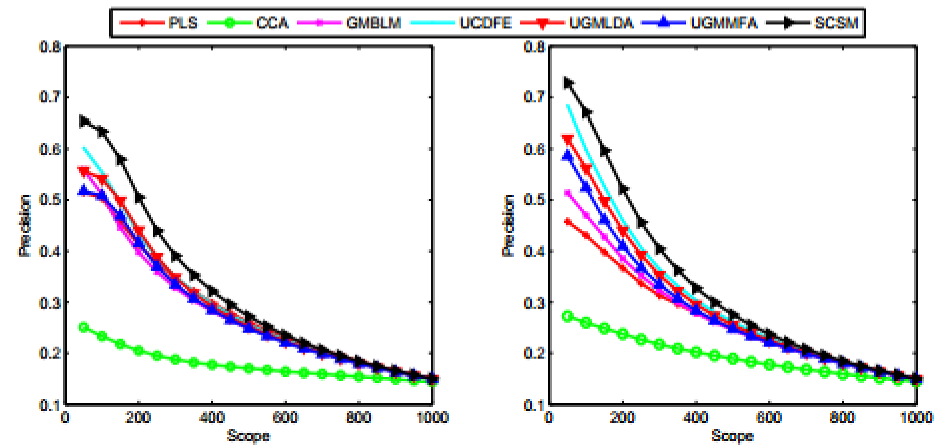
- ➢ *MAP of different methods on the* *Wiki++ dataset* *w.r.t. each category.*



Performances (cont'd)

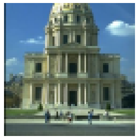


(a) precision-recall curve



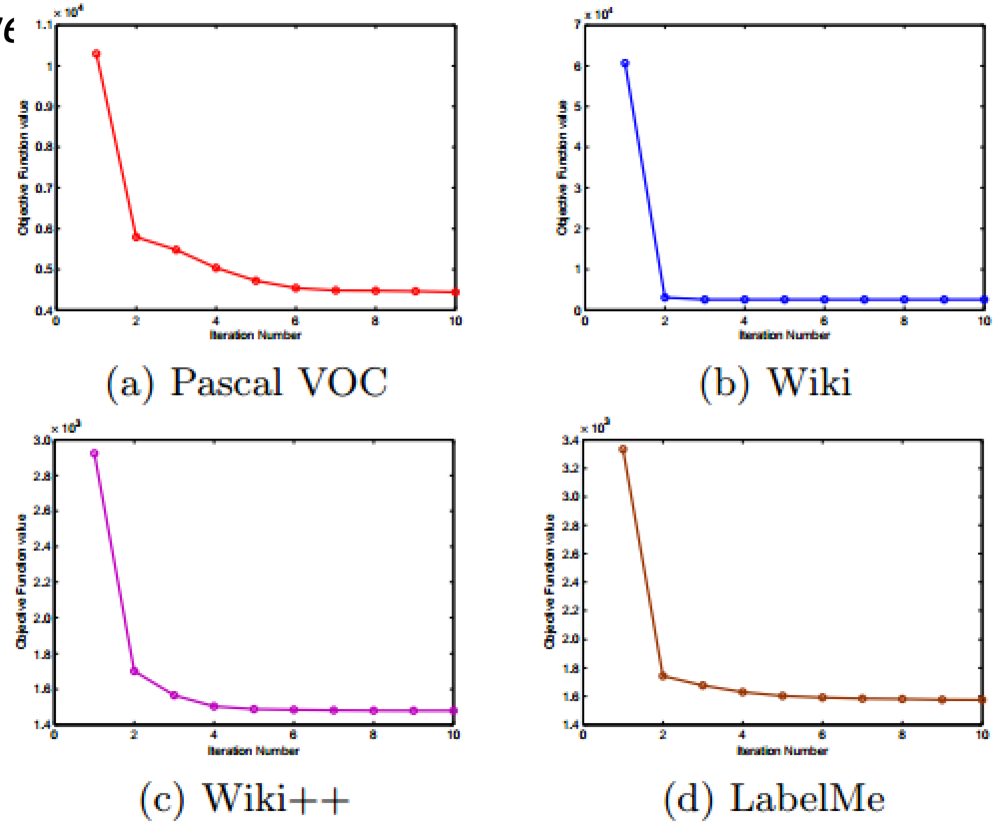
(b) precision-scope curve

- *Retrieval examples by image queries on the LabelMe dataset by the proposed SCSM and supervised GMLDA.*

Image queries	SCSM	GMLDA
	"water sea building sky sand beach mountain rainbow", "sand beach water sea sky", "sky mountain ocean water water sea sand beach trees", "water sea sky waves"	"sky mountain field river water trees", "sea water sky sand beach building tree", "sky buildings sea water", "sky buildings sea water"
	"sky field road tree shrub", "sky desert ground trees hill", "sky snowy mountain hill buildings", "sky mountain field grass cow"	"sky path shrub trees stone", "sky mountain field river water trees", "sky mountain field desert shrub plant", "sky desert ground trees hill"
	"sky skyscraper building", "building trees sidewalk road bus", "sky buildings skyscraper", "sky skyscraper building"	"sky building wall step person walking", "sky building car box plants", "sky skyscraper buildings car side sidewalk", "city river water building skyscraper dock"

Convergence Analysis

➢ *It is obvious to prove that the employed alternative minimization strategy can converge to a local optimum. However, under the self-paced framework, our learning algorithm is hard to guarantee the global conve*

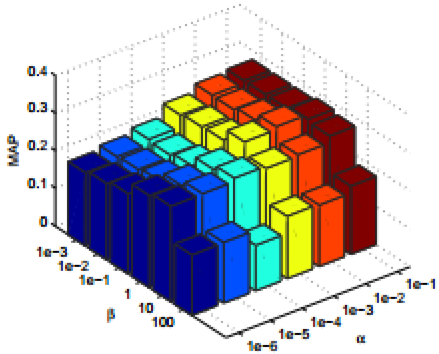


(a) Pascal VOC (b) Wiki

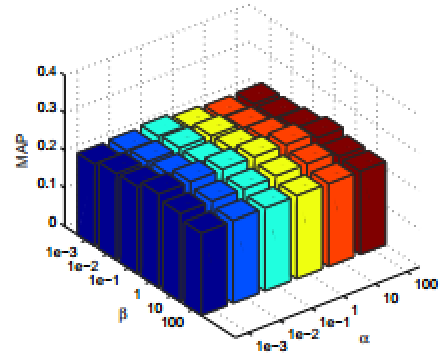
(c) Wiki++ (d) LabelMe

Parameter Analysis

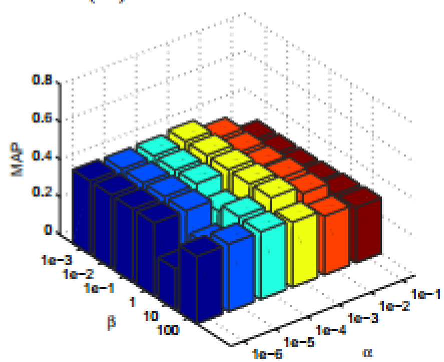
➢ *The analysis on parameter sensitivity shows that SCSM is very robust to model parameters which can achieve stable and superior performance under a wide range of parameter values.*



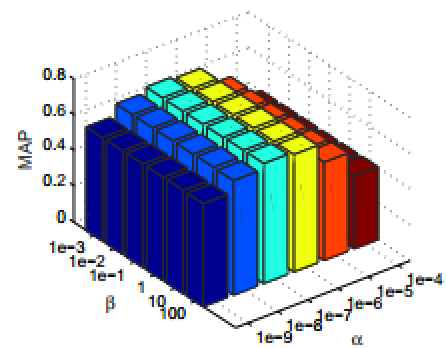
(a) Pascal VOC



(b) Wiki



(c) Wiki++



(d) LabelMe

Outline

- Background
- Self-Paced Learning
- Proposed Method
- Experiments
- Conclusions

Conclusions

Experimental results show that, our SCSM not only surpasses the unsupervised methods (e.g., CCA, GMBLM), but also outperforms some supervised methods (e.g., CDFE, GMLDA and LCFS). As an unsupervised method, the reasons for the better performance of our SCSM may lie in threefold.

- ✓ *One image often contain several semantic concepts. Hence the grouping results may be a complementary when feature representations are enough powerful.*
- ✓ *Pseudo group label is inaccurate. Since clustering is a non-convex problem, self-paced learning can help avoid the local minima.*
- ✓ *Besides, multimodal graph helps preserve the inter- and intra-similarities in the subspace.*

References

[Hardoon et al. 2004] Hardoon, David R., Sandor Szedmak, and John Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods." Neural computation 16, no. 12 (2004): 2639-2664

[Lin and Tang 2006] Lin, Dahua, and Xiaoou Tang. "Inter-modality face recognition." In European Conference on Computer Vision, pp. 13-26. Springer Berlin Heidelberg, 2006.

[Bengio et al. 2009] Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston. "Curriculum learning." In Proceedings of the 26th annual international conference on machine learning, pp. 41-48. ACM, 2009.

[Kumar et al. 2010] Kumar, M. Pawan, Benjamin Packer, and Daphne Koller. "Self-paced learning for latent variable models." In Advances in Neural Information Processing Systems, pp. 1189-1197. 2010.

[Sharma and Jacobs 2011] Sharma, Abhishek, and David W. Jacobs. "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch." In Computer Vision and Pattern Recognition , pp. 593-600. IEEE, 2011.

[Sharma et al. 2012] Sharma, Abhishek, Abhishek Kumar, Hal Daume, and David W. Jacobs. "Generalized multiview analysis: A discriminative latent space." In Computer Vision and Pattern Recognition, pp. 2160-2167. IEEE, 2012.

[Wang et al. 2013] Wang, Kaiye, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. "Learning coupled feature spaces for cross-modal matching." In Proceedings of the IEEE International Conference on Computer Vision, pp. 2088-2095. 2013.

[Jiang et al. 2015] Jiang, Lu, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. "Self-Paced Curriculum Learning." In Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[Wang et al. 2015] Wang, Kaiye, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. "Joint Feature Selection and Subspace Learning for Cross-modal Retrieval." (2015).

[Shen et al. 2015] Shen, Fumin, Chunhua Shen, Wei Liu, and Heng Tao Shen. "Supervised discrete hashing." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 37-45. 2015.

[Zhao et al. 2015] Zhao, Qian, Deyu Meng, Lu Jiang, Qi Xie, Zongben Xu, and Alexander G. Hauptmann. "Self-Paced Learning for Matrix Factorization." In Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[Fan et al. 2016] Fan, Yanbo, Ran He, Jian Liang and Bao-gang Hu. “Self-Paced Learning: an Implicit Regularization Perspective.” arXiv preprint arXiv:1606.00128 (2016).

谢谢！
Thank you!

Questions?