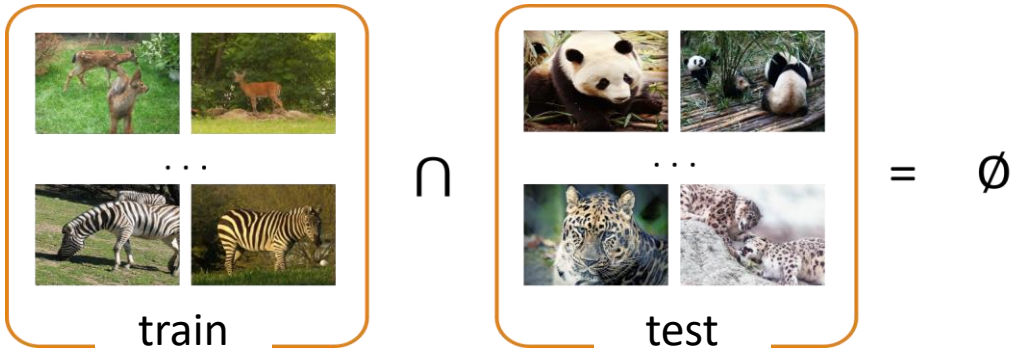
Multi-Label Zero-Shot Learning with Transfer-Aware Label Embedding Projection

Meng Ye[1] and Yuhong Guo[2]

[1] Computer and Information Sciences
Temple University, USA

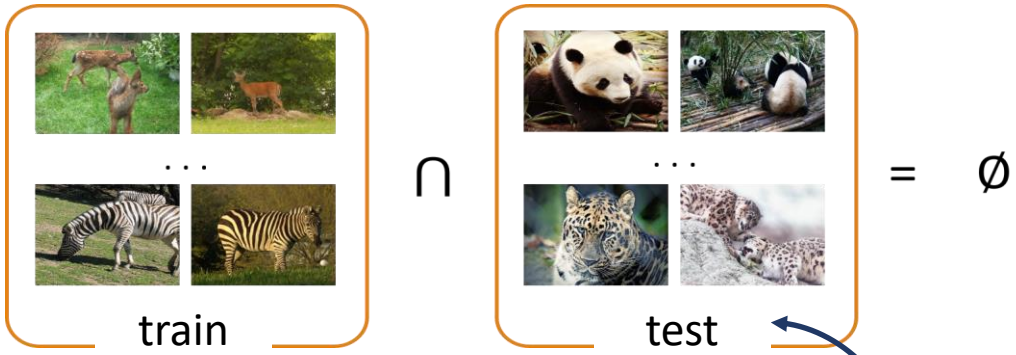
[2] School of Computer Science
Carleton University, Canada

Zero-Shot Learning

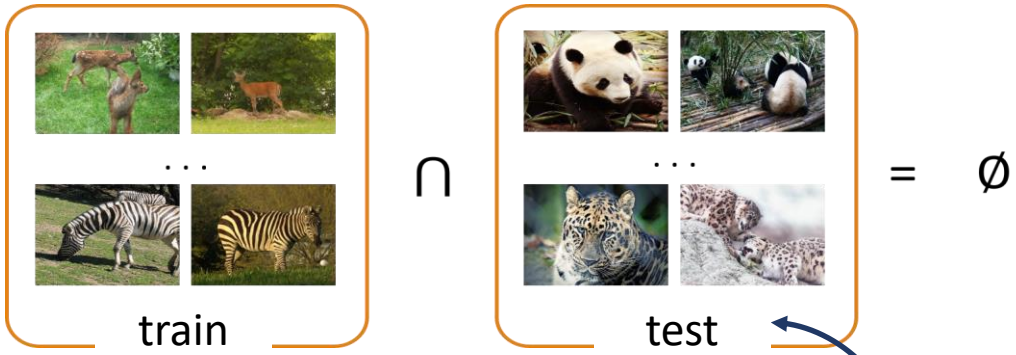


$\cap$ = $\emptyset$

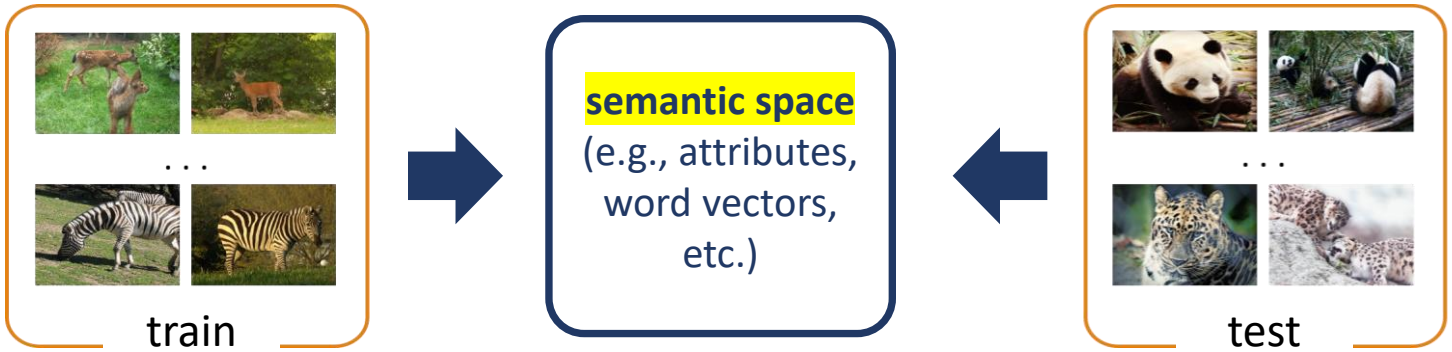
Zero-Shot Learning



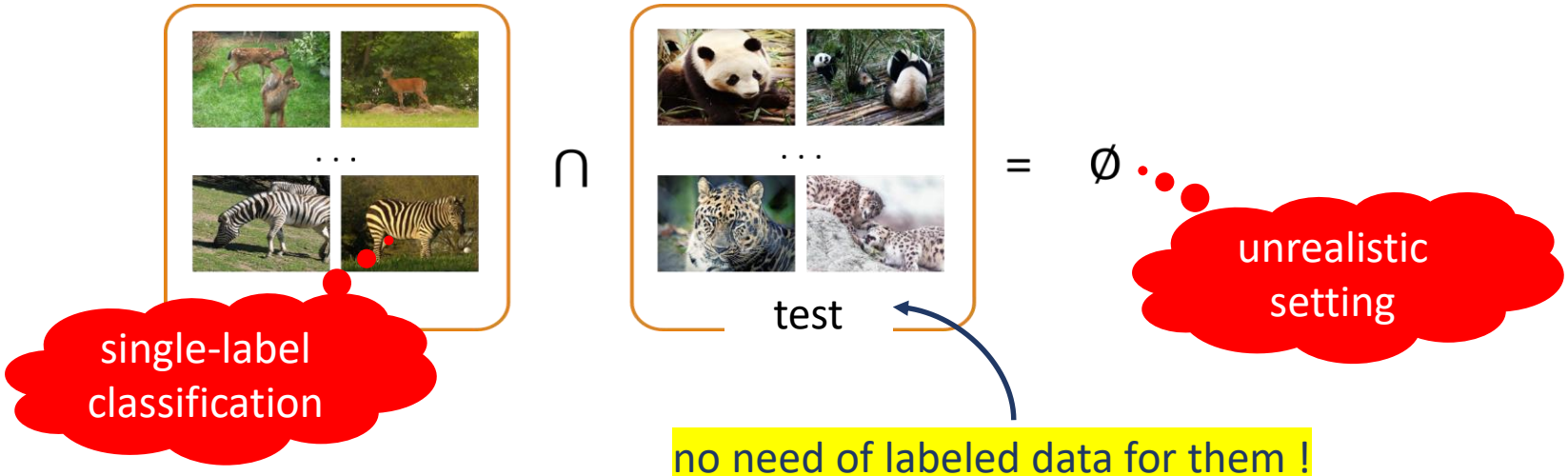
Zero-Shot Learning



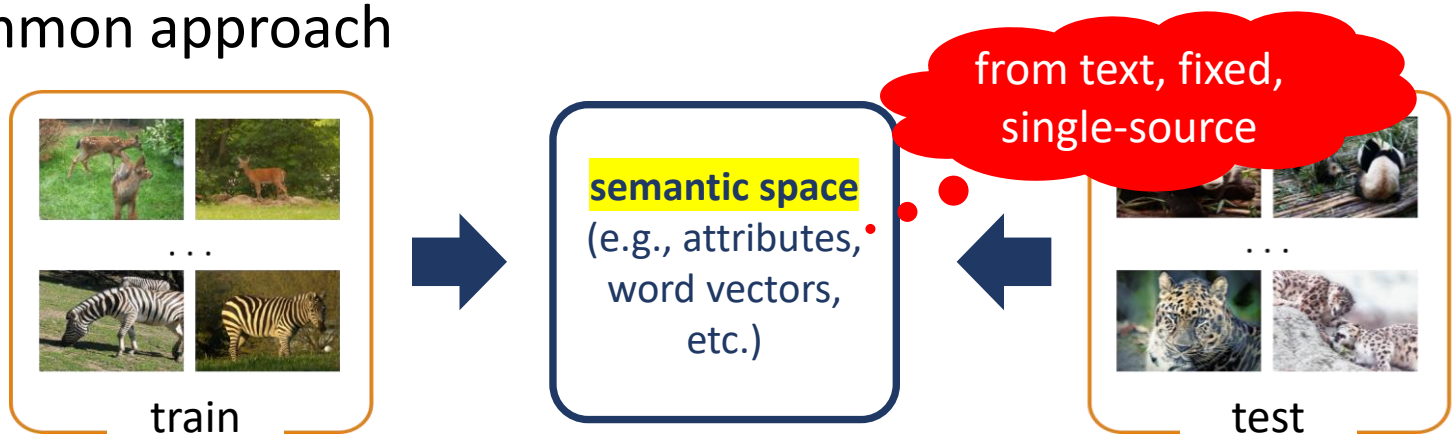
Common approach



Zero-Shot Learning

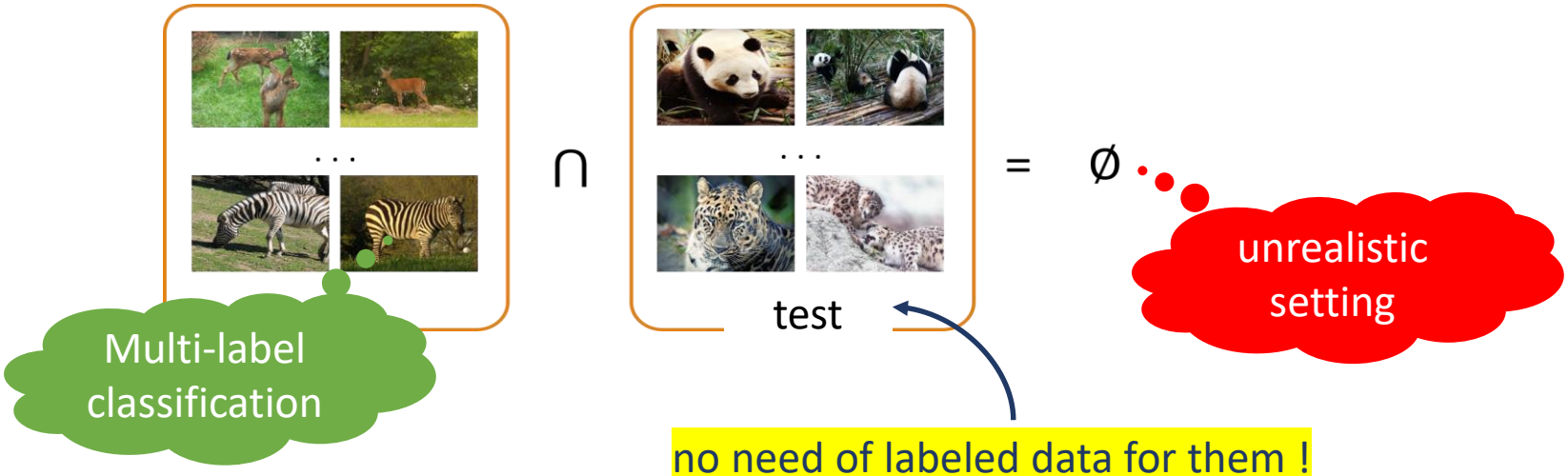


Common approach

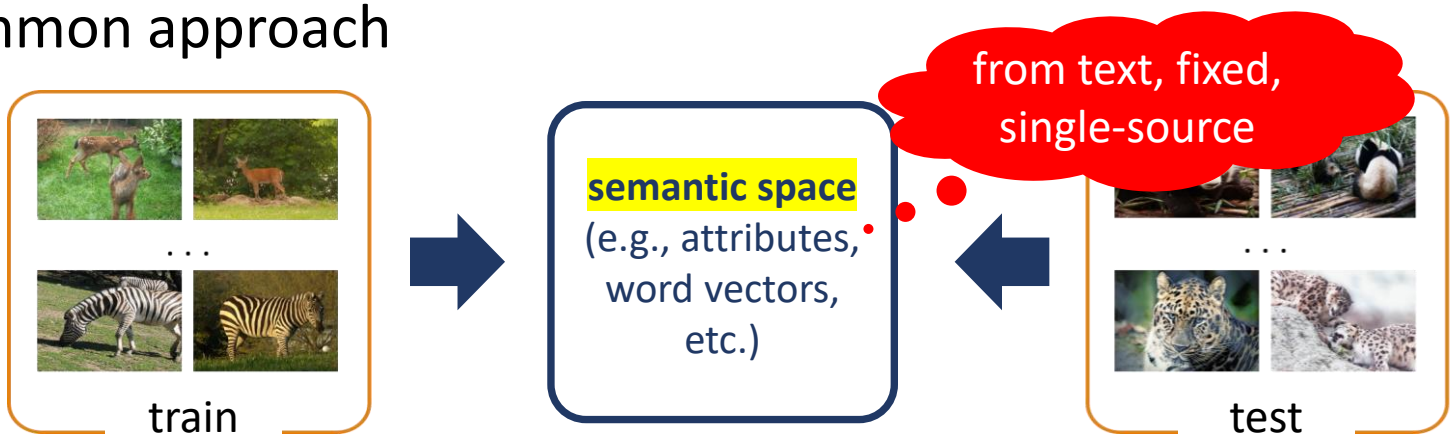


TEMPLE UNIVERSITY Carleton UNIVERSITY

Zero-Shot Learning

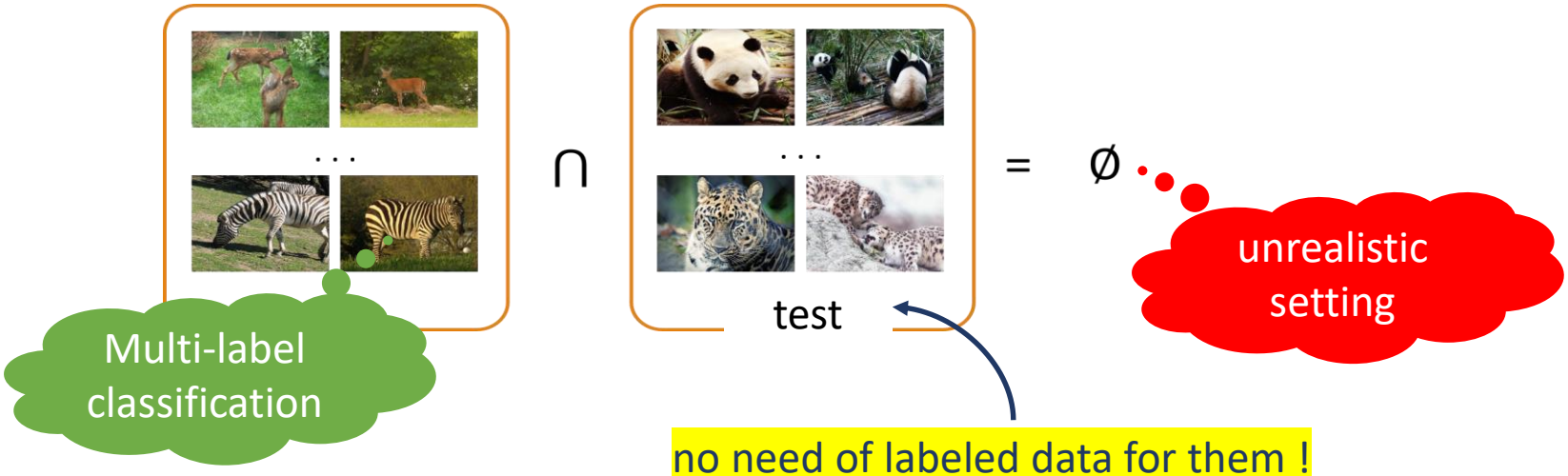


Common approach

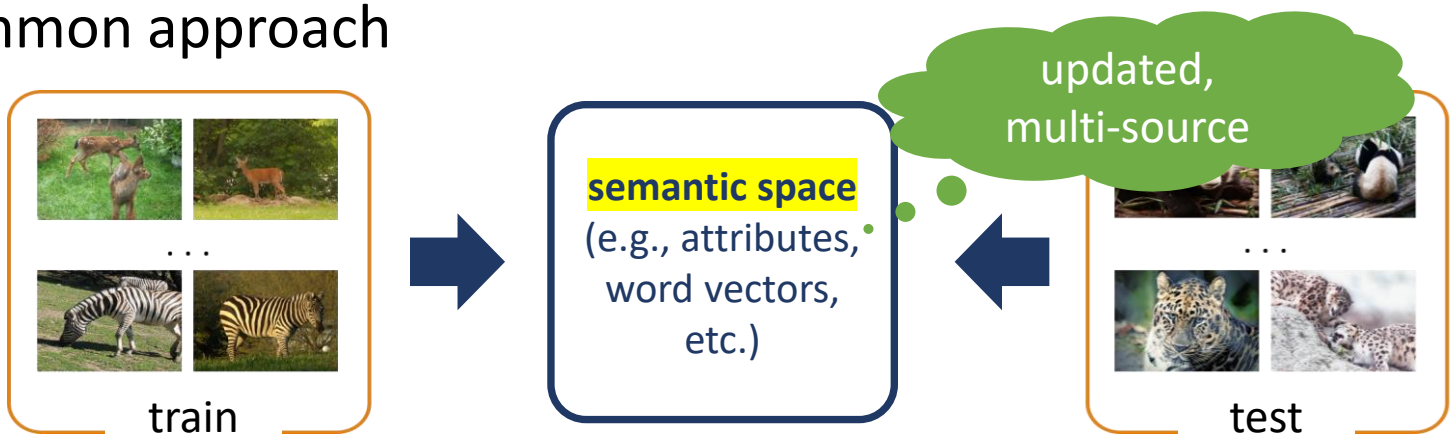


TEMPLE UNIVERSITY Carleton UNIVERSITY

Zero-Shot Learning

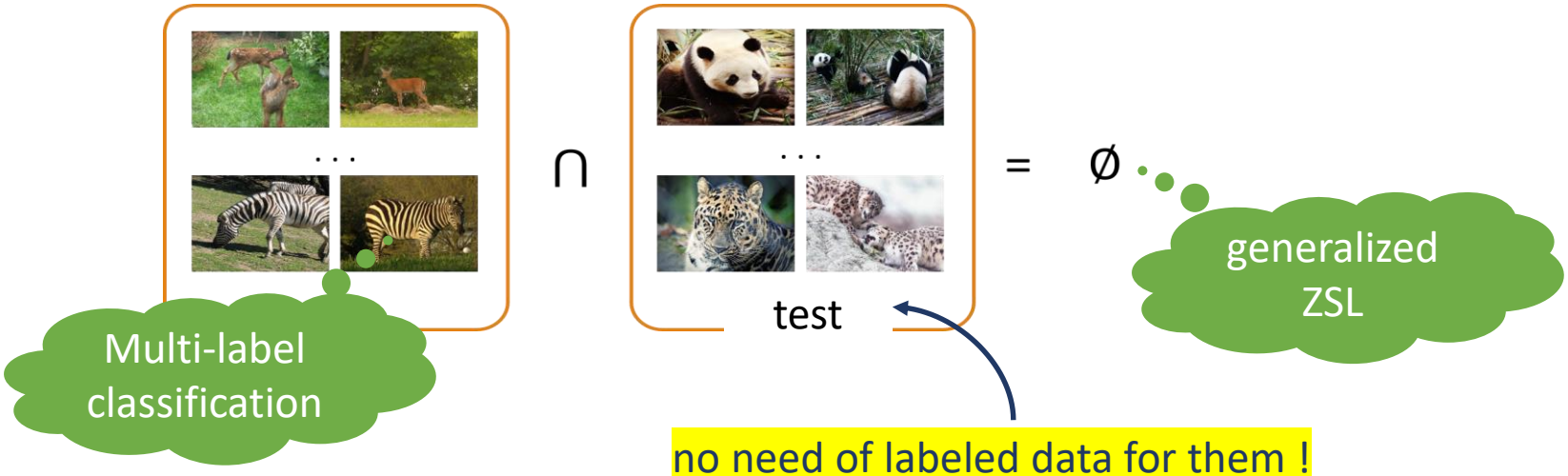


Common approach

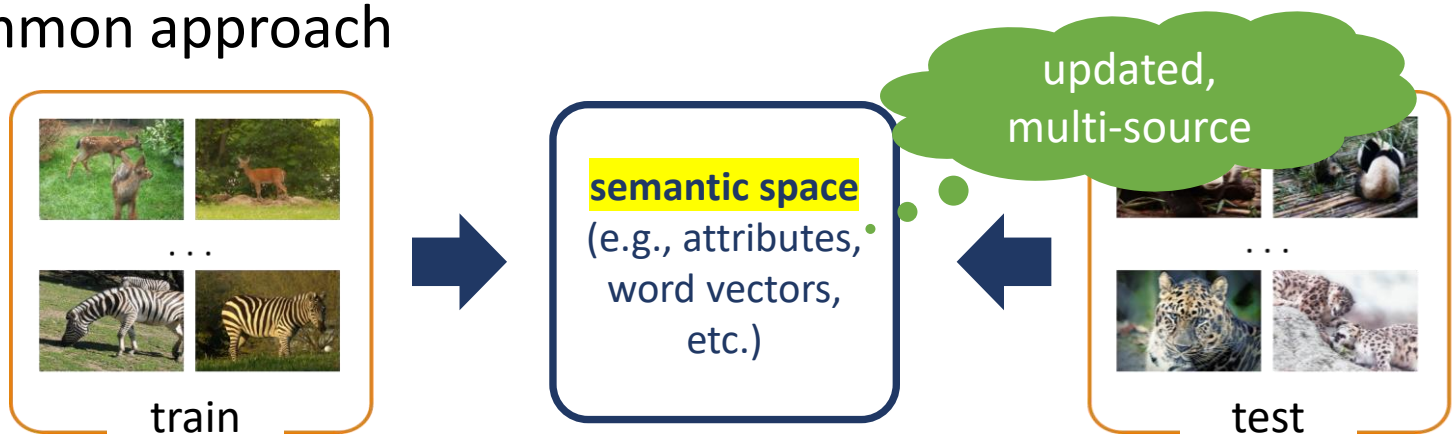


TEMPLE UNIVERSITY Carleton UNIVERSITY

Zero-Shot Learning



Common approach



TEMPLE UNIVERSITY Carleton UNIVERSITY

Problem Definition

Given

Training data: $D = (X, Y)$

Feature matrix: $X \in \mathbb{R}^{n \times d}$

Label matrix: $Y \in \{0, 1\}^{n \times L^s}$

Seen classes: $\mathcal{S} = \{1, 2, ..., L^s\}$

Unseen classes: $\mathcal{U} = \{L^s + 1, L^s + 2, ..., L\}$

Word embedding vectors: $M = [M^s; M^u] \in \mathbb{R}^{L \times m}$

TEMPLE UNIVERSITY Carleton UNIVERSITY

Problem Definition

Given

Training data: $D = (X, Y)$

Feature matrix: $X \in \mathbb{R}^{n \times d}$

Label matrix: $Y \in \{0, 1\}^{n \times L^s}$

Seen classes: $\mathcal{S} = \{1, 2, ..., L^s\}$

Unseen classes: $\mathcal{U} = \{L^s + 1, L^s + 2, ..., L\}$

Word embedding vectors: $M = [M^s; M^u] \in \mathbb{R}^{L \times m}$

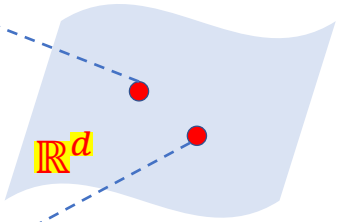
Goal

Learn a multi-label model from the training data that can predict unseen labels on test data

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Embedding projection

Tags:
Dining_table,chair

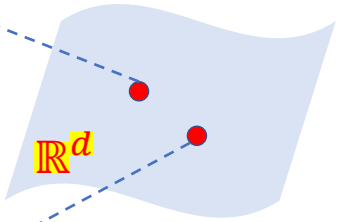


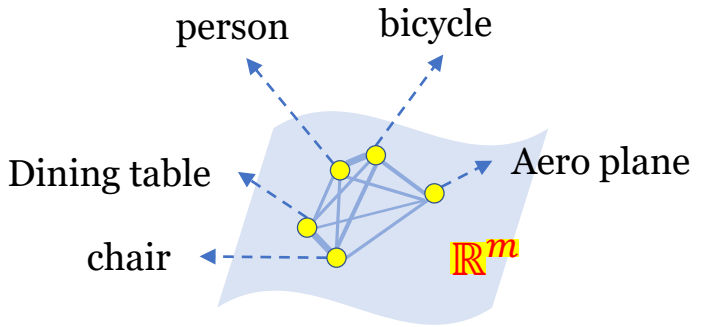
Tags:
person,bicycle

Approach – Embedding projection

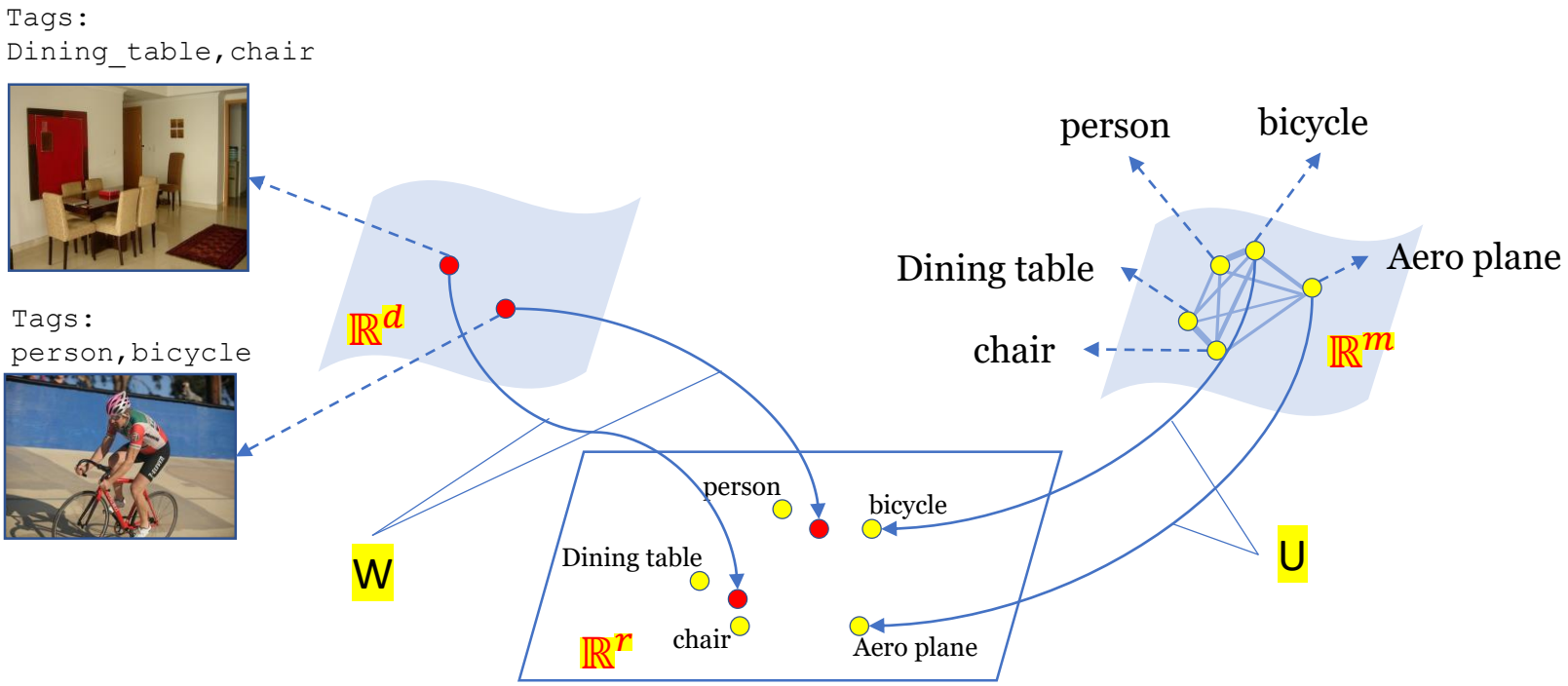
Tags:
Dining_table,chair

Tags:
person,bicycle





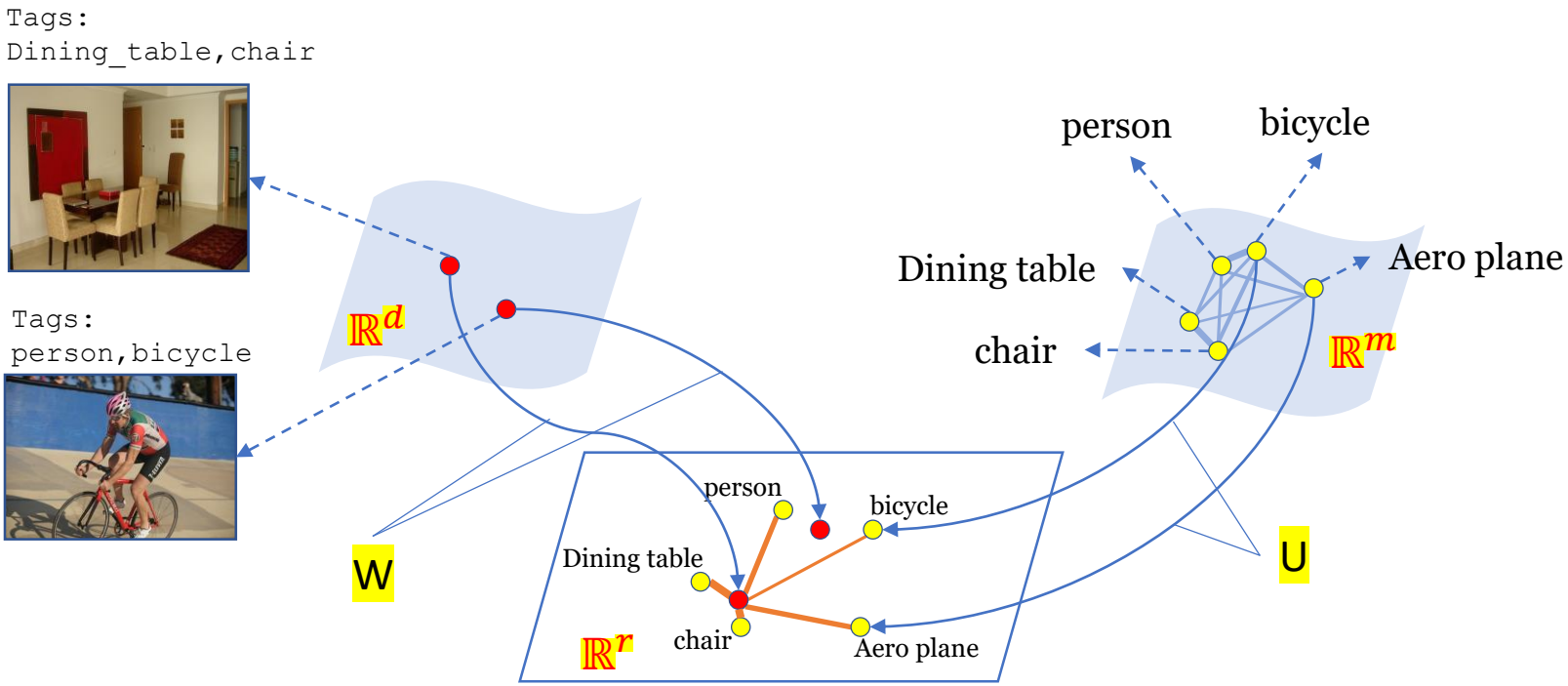
Approach – Embedding projection



$$\theta(X_i) = X_i W,\ W \in \mathbb{R}^{d \times r}$$

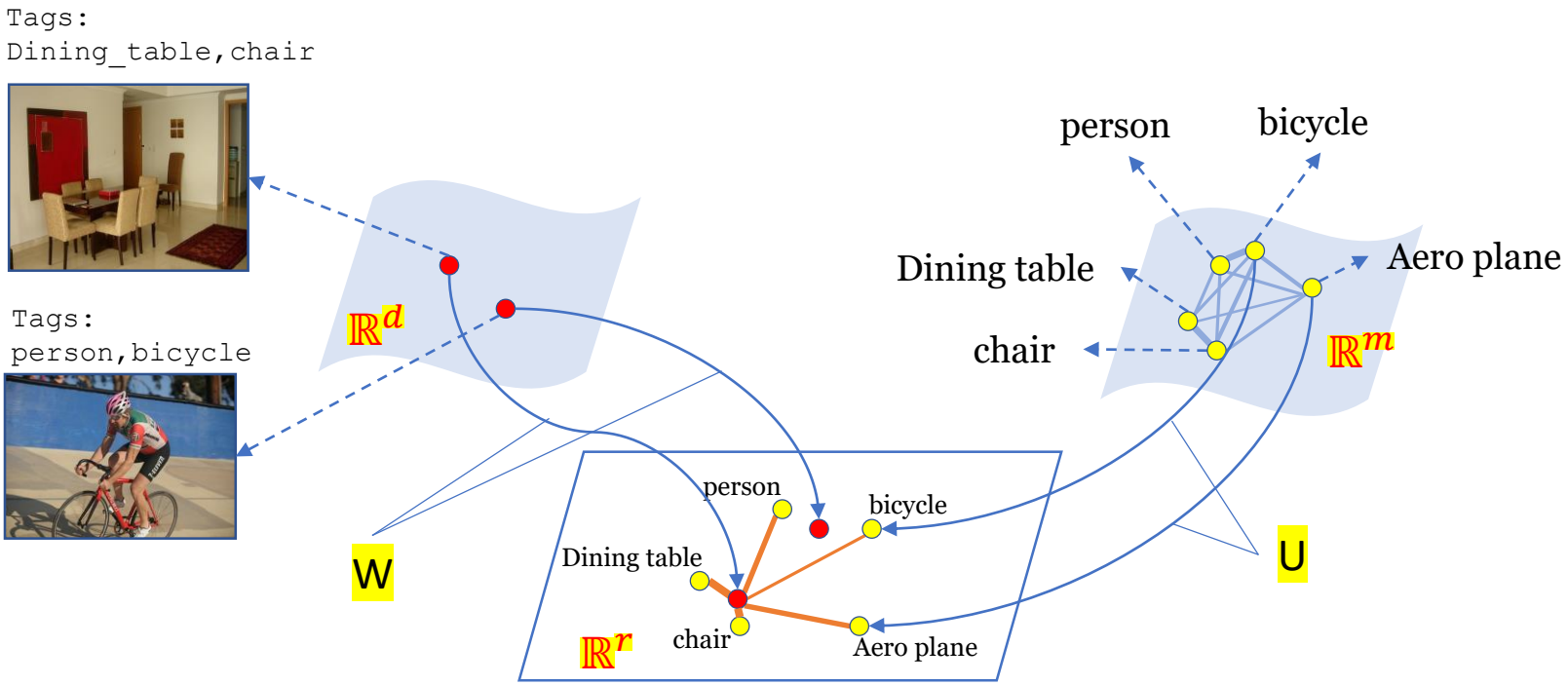
$$\phi(M_c) = M_c U,\ U \in \mathbb{R}^{m \times r}$$

Approach – Max-margin multi-label learning



$$F(i,c) = \theta(X_i)\phi(M_c)^\top$$

Approach – Max-margin multi-label learning

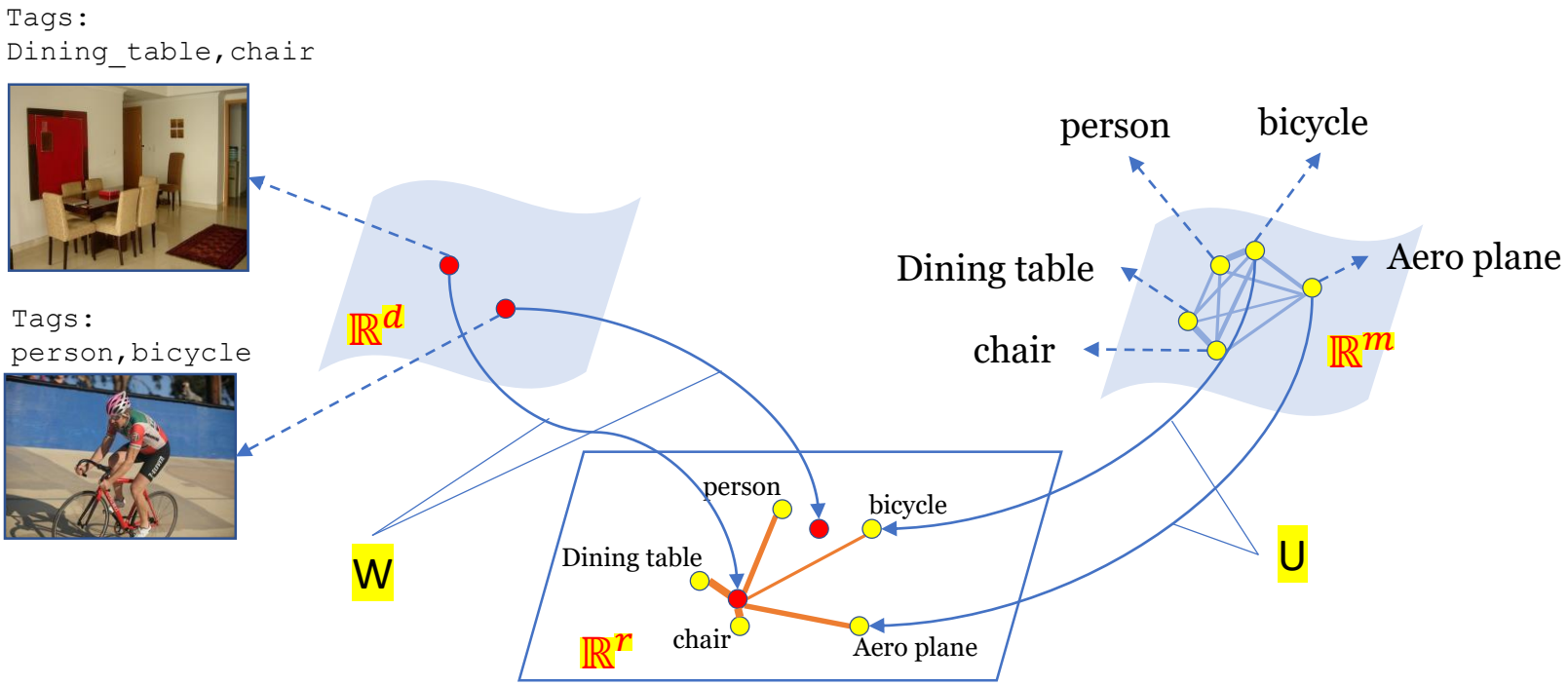


$$F(i,c) = \theta(X_i)\phi(M_c)^\top$$

$$\mathcal{L}(W,U;X_i,Y_i) = \max_{c \in Y_i} \left[1 + F(i,0) - F(i,c)\right]_+$$

$$+ \max_{\bar{c} \in \bar{Y}_i} \left[1 + F(i,\bar{c}) - F(i,0)\right]_+$$

Approach – Max-margin multi-label learning



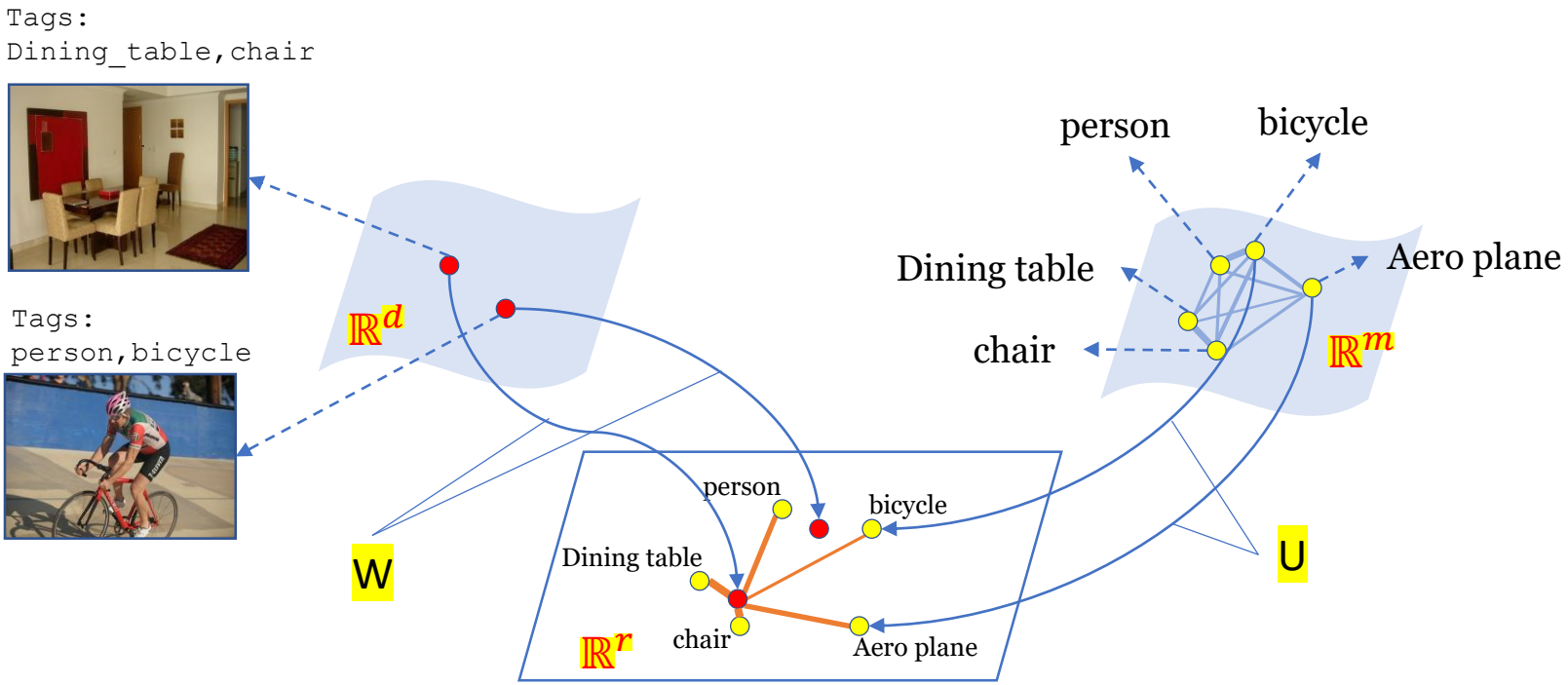
$$F(i,c) = \theta(X_i)\phi(M_c)^\top$$

$$\mathcal{L}(W,U;X_i,Y_i) = \max_{c\in Y_i}\left[1 + F(i,0) - F(i,c)\right]_+ + \max_{\bar{c}\in \bar{Y}_i}\left[1 + F(i,\bar{c}) - F(i,0)\right]_+$$

$$F(i,0) = X_i W_0$$

dummy score

Approach – Max-margin multi-label learning



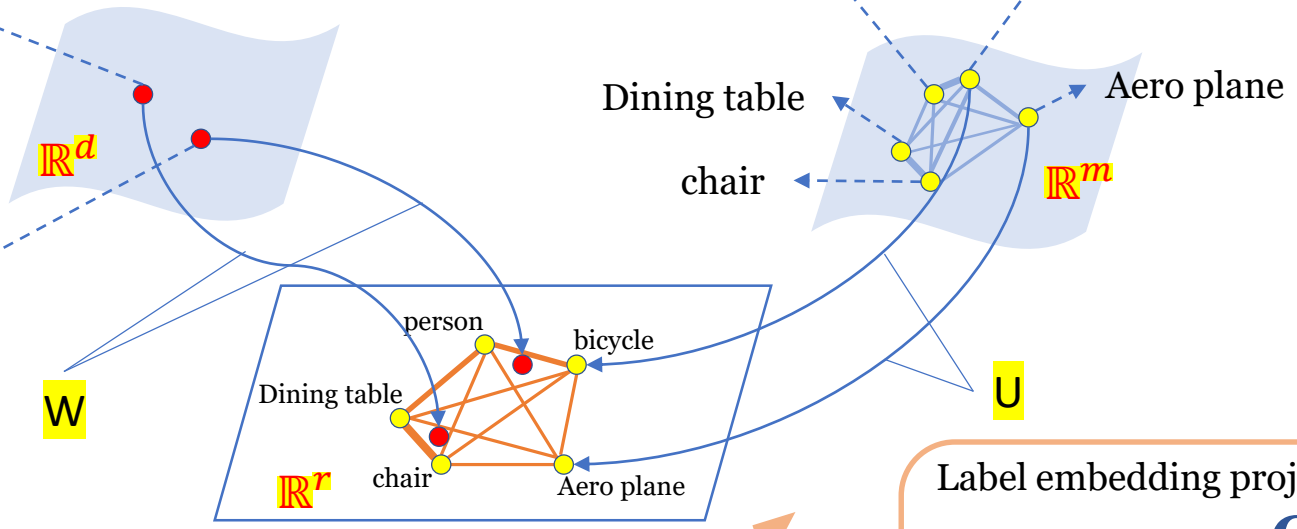
$$F(i,c) = \theta(X_i)\phi(M_c)^\top$$

$$\mathcal{L}(W,U;X_i,Y_i) = \max_{c\in Y_i}\left[1 + F(i,0) - F(i,c)\right]_+ + \max_{\bar{c}\in \bar{Y}_i}\left[1 + F(i,\bar{c}) - F(i,0)\right]_+$$

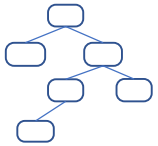
$$F(i,0) = X_i W_0$$

dummy score

Approach – Transfer-Aware Embedding Projection (TAEP)



Label embedding projection regularization **Q**

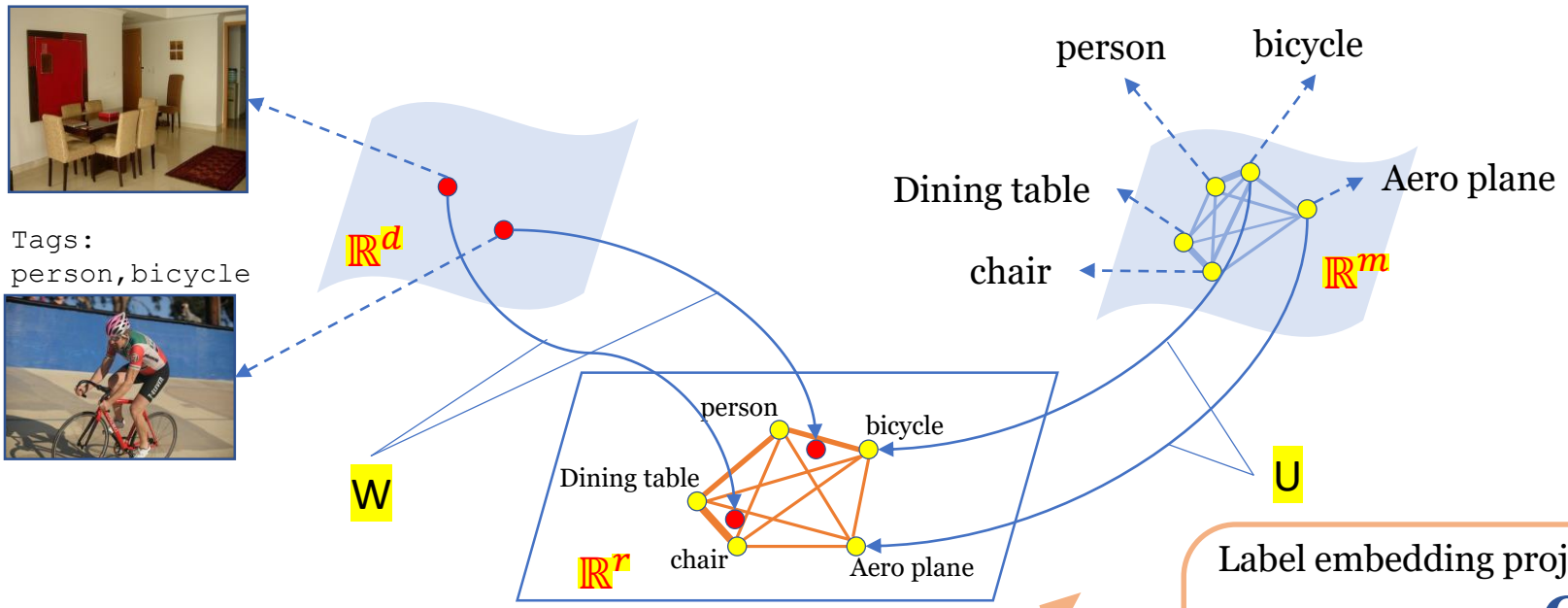


flickr

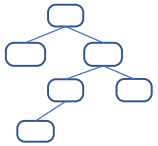
$$\mathcal{H}(U) = \frac{\gamma}{2L^u(L^u-1)} \sum_{i,j\in\mathcal{U},i\neq j} M_i U U^\top M_j^\top - \frac{\gamma}{2L^s L^u} \sum_{i\in\mathcal{S},j\in\mathcal{U}} M_i U U^\top M_j^\top$$

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Transfer-Aware Embedding Projection (TAEP)



Label embedding projection regularization **Q**



flickr

$$\mathcal{H}(U) = \frac{\gamma}{2L^u(L^u-1)} \sum_{i,j\in\mathcal{U},i\neq j} M_i U U^\top M_j^\top - \frac{\gamma}{2L^s L^u} \sum_{i\in\mathcal{S},j\in\mathcal{U}} M_i U U^\top M_j^\top$$

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Transfer-Aware Embedding Projection (TAEP)

$$\mathcal{H}(U) = \frac{\gamma}{2L^u(L^u-1)} \sum_{i,j\in\mathcal{U},i\neq j} M_i U U^\top M_j^\top - \frac{\gamma}{2L^s L^u} \sum_{i\in\mathcal{S},j\in\mathcal{U}} M_i U U^\top M_j^\top$$
$$= \frac{\gamma}{2}\mathrm{tr}\left(U^\top M^\top Q M U\right)$$

$$Q = \begin{bmatrix} \mathbf{0}_{L^s,L^s} & \frac{-1}{2L^sL^u}\mathbf{1}_{L^s,L^u} \\ \frac{-1}{2L^sL^u}\mathbf{1}_{L^u,L^s} & \frac{1}{L^u(L^u-1)}(\mathbf{1}_{L^u,L^u} - I_{L^u}) \end{bmatrix}$$

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Transfer-Aware Embedding Projection (TAEP)

Auxiliary information:

A matrix R, each entry R_{ij} denotes the similarity between a pair of labels (i, j)

Approach – Transfer-Aware Embedding Projection (TAEP)

Auxiliary information:

A matrix R, each entry R_{ij} denotes the similarity between a pair of labels (i, j)

Manifold regularization:

degree

adjacency

$$D = diag(R\mathbf{1}) \qquad Q^{\mathcal{A}} = I - D^{-1/2} R D^{-1/2}$$

normalized laplacian

$$\mathcal{A}(U) = \frac{\lambda}{2} \mathrm{tr}\left(U^\top M^\top Q^{\mathcal{A}} M U\right)$$

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Transfer-Aware Embedding Projection (TAEP)

Auxiliary information:

A matrix R, each entry R_{ij} denotes the similarity between a pair of labels (i, j)

Manifold regularization:

degree

adjacency

$$D = diag(R\mathbf{1}) \qquad Q^{\mathcal{A}} = I - D^{-1/2} R D^{-1/2}$$

normalized laplacian

$$\mathcal{A}(U) = \frac{\lambda}{2} \mathrm{tr}\left(U^\top M^\top Q^{\mathcal{A}} M U\right)$$

If a pair of label (i, j) was similar in auxiliary source of information, then they should still be similar after projection

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Transfer-Aware Embedding Projection (TAEP)

Advantages:

- $H(U)$ and $A(U)$ can be integrated together:

$$\mathcal{H}(U) = \frac{\gamma}{2}\mathrm{tr}\left(U^\top M^\top \left(Q + \frac{\lambda}{\gamma}Q^{\mathcal{A}}\right)MU\right)$$

1. G. A Miller. “WordNet: a lexical database for English”. In: Communications of the ACM 38.11 (1995). P. 39-41
2. M. Rohrbach et al. “What helps where – and why? Semantic relatedness for knowledge transfer”. In: CVPR. 2010.
3. T. Mensink, E. Gavves and C. GM Snoek. “Costa: Co-occurrence statistics for zero-shot classification”. In: CVPR. 2014.

Approach – Transfer-Aware Embedding Projection (TAEP)

Advantages:

- $H(U)$ and $A(U)$ can be integrated together:

$$\mathcal{H}(U) = \frac{\gamma}{2}\mathrm{tr}\left(U^\top M^\top \left(Q + \frac{\lambda}{\gamma}Q^{\mathcal{A}}\right)MU\right)$$

- Different sources can be incorporated easily:

WordNet[1] hierarchy: $R_{ij} = \frac{1}{path\_len(i,j)+1}$

shortest path

Flickr Image Hit-Count[2,3]: $R_{ij} = \frac{HC(i,j)}{HC(i)+HC(j)}$

Co-occurrence statistics

1. G. A Miller. "WordNet: a lexical database for English". In: **Communications of the ACM 38.11** (1995). P. 39-41
2. M. Rohrbach et al. "What helps where – and why? Semantic relatedness for knowledge transfer". In: **CVPR**. 2010.
3. T. Mensink, E. Gavves and C. GM Snoek. "Costa: Co-occurrence statistics for zero-shot classification". In: **CVPR**. 2014.

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Full objective

$$\min_{\substack{W, W_0, \xi, \eta, \\ U:\ U^\top U = I}} \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}(\|W\|_{\mathrm{F}}^2 + \|W_0\|^2) + \mathcal{H}(U)$$

$$\text{s.t.} \quad F(i,c) - F(i,0) \geq 1 - \boldsymbol{\xi}_i, \forall c \in Y_i, \forall i;\ \boldsymbol{\xi} \geq 0;$$

$$F(i,0) - F(i,\bar{c}) \geq 1 - \boldsymbol{\eta}_i, \forall \bar{c} \in \bar{Y}_i, \forall i;\ \boldsymbol{\eta} \geq 0$$

Approach – Full objective

$$\min_{\substack{W, W_0, \xi, \eta, \\ U:\ U^\top U = I}} \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}(\|W\|_{\mathrm{F}}^2 + \|W_0\|^2) + \mathcal{H}(U)$$

$$\text{s.t.} \quad F(i,c) - F(i,0) \geq 1 - \boldsymbol{\xi}_i, \forall c \in Y_i, \forall i;\ \boldsymbol{\xi} \geq 0;$$

$$F(i,0) - F(i,\bar{c}) \geq 1 - \boldsymbol{\eta}_i, \forall \bar{c} \in \bar{Y}_i, \forall i;\ \boldsymbol{\eta} \geq 0$$

- Maximize margins to encourage positive labels ranking higher than negative labels

TEMPLE UNIVERSITY Carleton UNIVERSITY

Approach – Full objective

$$\min_{\substack{W, W_0, \xi, \eta, \\ U:\ U^\top U = I}} \mathbf{1}^\top \xi + \mathbf{1}^\top \eta + \frac{\beta}{2}(\|W\|_{\mathrm{F}}^2 + \|W_0\|^2) + \mathcal{H}(U)$$

$$\text{s.t.} \quad F(i,c) - F(i,0) \geq 1 - \xi_i, \forall c \in Y_i, \forall i;\ \xi \geq 0;$$

$$F(i,0) - F(i,\bar{c}) \geq 1 - \eta_i, \forall \bar{c} \in \bar{Y}_i, \forall i;\ \eta \geq 0$$

- Maximize margins to encourage positive labels ranking higher than negative labels
- Minimize H(U) to refine label embeddings in the projected space

Approach – Full objective

$$\min_{\substack{W, W_0, \xi, \eta, \\ U:\, U^\top U = I}} \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}(\|W\|_{\mathrm{F}}^2 + \|W_0\|^2) + \mathcal{H}(U)$$

$$\text{s.t.} \quad F(i,c) - F(i,0) \geq 1 - \boldsymbol{\xi}_i, \forall c \in Y_i, \forall i;\ \boldsymbol{\xi} \geq 0;$$

$$F(i,0) - F(i,\bar{c}) \geq 1 - \boldsymbol{\eta}_i, \forall \bar{c} \in \bar{Y}_i, \forall i;\ \boldsymbol{\eta} \geq 0$$

- Maximize margins to encourage positive labels ranking higher than negative labels
- Minimize H(U) to refine label embeddings in the projected space

Approach – Test

$$F(i,c) = x_i^{test} W U^\top M_c^\top$$

then, rank all classes c w.r.t. $F(i, c)$

TEMPLE UNIVERSITY Carleton UNIVERSITY

Experiments – Settings

Dataset	PASCAL VOC2007	PASCAL VOC2012
# classes	20	20
# training images	5011	5717
# test images	4952	5823

Experiments – Settings

Dataset	PASCAL VOC2007	PASCAL VOC2012
# classes	20	20
# training images	5011	5717
# test images	4952	5823

- **Visual feature**: 4096-dim vector from VGG[1] pretrained on ImageNet
- **Label embeddings**: 300-dim word vectors trained by GloVe[2]
- **Evaluation metrics**:

  MiAP: mean AP scores averaged over each image[3]

  mi/ma-F1: mean micro/macro-F1 scores averaged over each label

1. K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: **arXiv preprint arXiv: 1409.1556** (2014)
2. J. Pennington, R. Socher and C. D Manning. "Glove: Global Vectors for Word Representation". In: **EMNLP**. 2014.
3. X. Li et al. "Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval". In: **ACM Computing Surveys (CSUR) 49.1** (2016), p. 14.

Experiments – Comparison methods

- **ConSE**[1]

  multinomial logistic regression + convex combination

- **LatEm-M**

  original LatEm[2] + multi-label ranking loss

- **DMP**[3] **/ Fast0Tag**[4]

  multi-label ZSL methods

1. M. Norouzi et al. “Zero-shot learning by convex combination of semantic embeddings”. In: **arXiv preprint arXiv: 1312.5650** (2013).
2. Y. Xian et al. “Latent embeddings for zero-shot classification”. In: **CVPR**. 2016.
3. Y. Fu et al. “Transductive Multi-label Zero-Shot Learning”. In: **BMVC**. 2014.
4. Y. Zhang, B. Gong and M. Shah. “Fast zero-shot image tagging”. In: **CVPR** 2016.

TEMPLE UNIVERSITY Carleton UNIVERSITY

Experiments – Comparison methods

- **ConSE**[1]

  multinomial logistic regression + convex combination

- **LatEm-M**

  original LatEm[2] + multi-label ranking loss

- **DMP**[3] **/ Fast0Tag**[4]

  multi-label ZSL methods

- **proposed**

  TAEP - only use matrix Q to regularize *U*
  TAEP-H - use Q + Q^A (WordNet **H**ierarchy) to regularize *U*
  TAEP-C - use Q + Q^A (Flickr Image Hit-**C**ount) to regularize *U*

1. M. Norouzi et al. "Zero-shot learning by convex combination of semantic embeddings". In: **arXiv preprint arXiv: 1312.5650** (2013).
2. Y. Xian et al. "Latent embeddings for zero-shot classification". In: **CVPR**. 2016.
3. Y. Fu et al. "Transductive Multi-label Zero-Shot Learning". In: **BMVC**. 2014.
4. Y. Zhang, B. Gong and M. Shah. "Fast zero-shot image tagging". In: **CVPR** 2016.

Experiments – Standard ZSL

Methods	VOC2007			VOC2012		
	MiAP	mi-F1	ma-F1	MiAP	mi-F1	ma-F1
ConSE	49.98	30.80	27.57	49.95	33.48	28.83
LatEm-M	52.45	35.32	36.69	51.44	35.74	36.33
DMP	53.52	36.70	40.44	52.92	35.73	41.04
Fast0Tag	52.39	35.01	36.76	52.29	34.23	35.38
TAEP	57.42	38.48	42.33	54.39	37.63	41.58
TAEP-C	**59.22**	**39.84**	**43.77**	**57.13**	**39.30**	**42.97**
TAEP-H	57.62	38.95	43.29	56.10	38.89	42.23

**average performance of 5 runs*

TEMPLE UNIVERSITY Carleton UNIVERSITY

Experiments – Standard ZSL

Methods	VOC2007			VOC2012		
	MiAP	mi-F1	ma-F1	MiAP	mi-F1	ma-F1
ConSE	49.98	30.80	27.57	49.95	33.48	28.83
LatEm-M	52.45	35.32	36.69	51.44	35.74	36.33
DMP	53.52	36.70	40.44	52.92	35.73	41.04
Fast0Tag	52.39	35.01	36.76	52.29	34.23	35.38
TAEP	57.42	38.48	42.33	54.39	37.63	41.58
TAEP-C	**59.22**	**39.84**	**43.77**	**57.13**	**39.30**	**42.97**
TAEP-H	57.62	38.95	43.29	56.10	38.89	42.23

**average performance of 5 runs*

- Specialized multi-label methods are generally better than adapted single-label methods

Experiments – Standard ZSL

Methods	VOC2007			VOC2012		
	MiAP	mi-F1	ma-F1	MiAP	mi-F1	ma-F1
ConSE	49.98	30.80	27.57	49.95	33.48	28.83
LatEm-M	52.45	35.32	36.69	51.44	35.74	36.33
DMP	53.52	36.70	40.44	52.92	35.73	41.04
Fast0Tag	52.39	35.01	36.76	52.29	34.23	35.38
TAEP	57.42	38.48	42.33	54.39	37.63	41.58
TAEP-C	**59.22**	**39.84**	**43.77**	**57.13**	**39.30**	**42.97**
TAEP-H	57.62	38.95	43.29	56.10	38.89	42.23

**average performance of 5 runs*

- Specialized multi-label methods are generally better than adapted single-label methods
- TAEP outperforms comparison methods

Experiments – Standard ZSL

Methods	VOC2007			VOC2012		
	MiAP	mi-F1	ma-F1	MiAP	mi-F1	ma-F1
ConSE	49.98	30.80	27.57	49.95	33.48	28.83
LatEm-M	52.45	35.32	36.69	51.44	35.74	36.33
DMP	53.52	36.70	40.44	52.92	35.73	41.04
Fast0Tag	52.39	35.01	36.76	52.29	34.23	35.38
TAEP	57.42	38.48	42.33	54.39	37.63	41.58
TAEP-C	**59.22**	**39.84**	**43.77**	**57.13**	**39.30**	**42.97**
TAEP-H	57.62	38.95	43.29	56.10	38.89	42.23

**average performance of 5 runs*

- Specialized multi-label methods are generally better than adapted single-label methods
- TAEP outperforms comparison methods
- TAEP-C and TAEP-H further increase the performance

TEMPLE UNIVERSITY Carleton UNIVERSITY

Experiments – Generalized ZSL

Methods	VOC2007			VOC2012		
	MiAP	mi-F1	ma-F1	MiAP	mi-F1	ma-F1
ConSE	64.10	42.11	32.29	62.85	41.17	35.72
LatEm-M	66.46	43.11	32.37	63.06	39.95	32.35
DMP	67.79	43.97	34.13	64.24	41.29	32.39
Fast0Tag	67.34	43.54	33.31	64.63	41.28	32.46
TAEP	68.16	43.61	35.29	64.67	40.60	34.07
TAEP-C	**69.87**	**44.75**	**35.62**	**65.33**	**42.10**	**36.74**
TAEP-H	69.74	44.55	35.56	65.10	41.39	35.95

**average performance of 5 runs*

- TAEP-C and TAEP-H outperform other methods
- Improvements are less significant than standard setting

TEMPLE UNIVERSITY Carleton UNIVERSITY

Experiments – Efficacy of TAEP

Experiments – Efficacy of TAEP

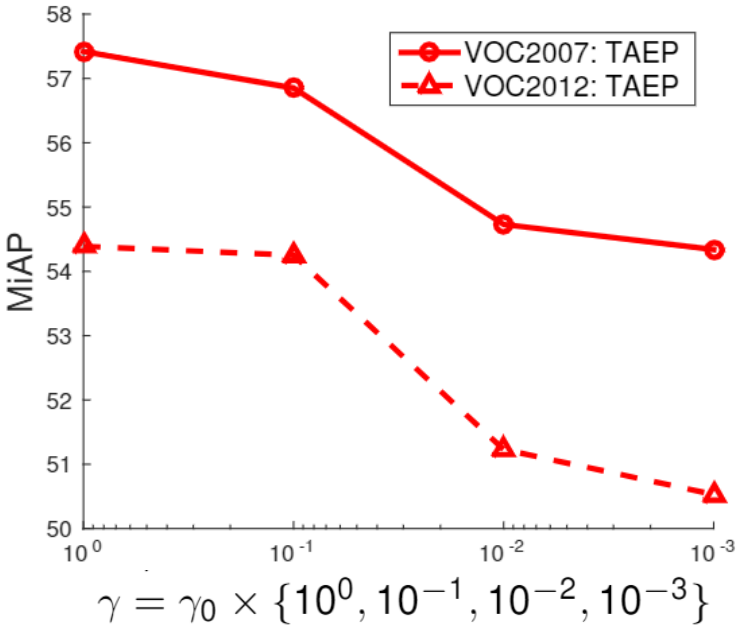
$$\mathcal{H}(U) = \frac{\gamma}{2}\mathrm{tr}\left(U^\top M^\top (Q + \frac{\lambda}{\gamma} Q^{\mathcal{A}}) M U\right)$$

γ and λ control the impact of $H(U)$ term:

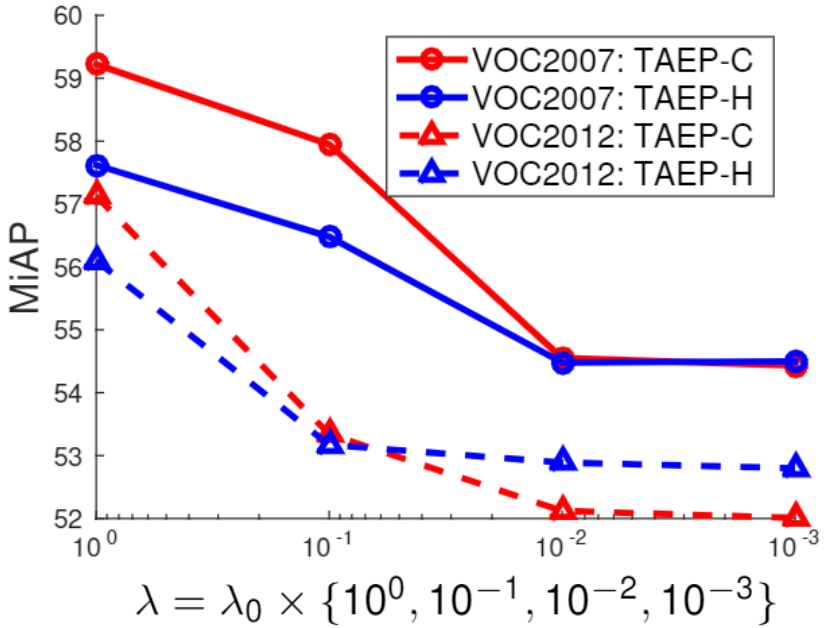
Experiments – Efficacy of TAEP

$$\mathcal{H}(U) = \frac{\gamma}{2}\mathrm{tr}\left(U^\top M^\top (Q + \frac{\lambda}{\gamma} Q^{\mathcal{A}}) MU\right)$$

γ and λ control the impact of $H(U)$ term:



$\gamma = \gamma_0 \times \{10^0, 10^{-1}, 10^{-2}, 10^{-3}\}$



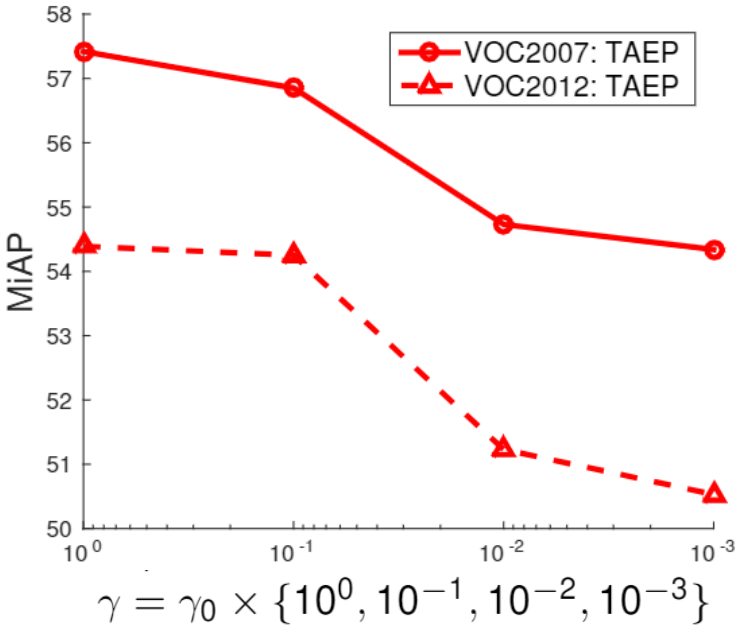
$\lambda = \lambda_0 \times \{10^0, 10^{-1}, 10^{-2}, 10^{-3}\}$

TEMPLE UNIVERSITY Carleton UNIVERSITY

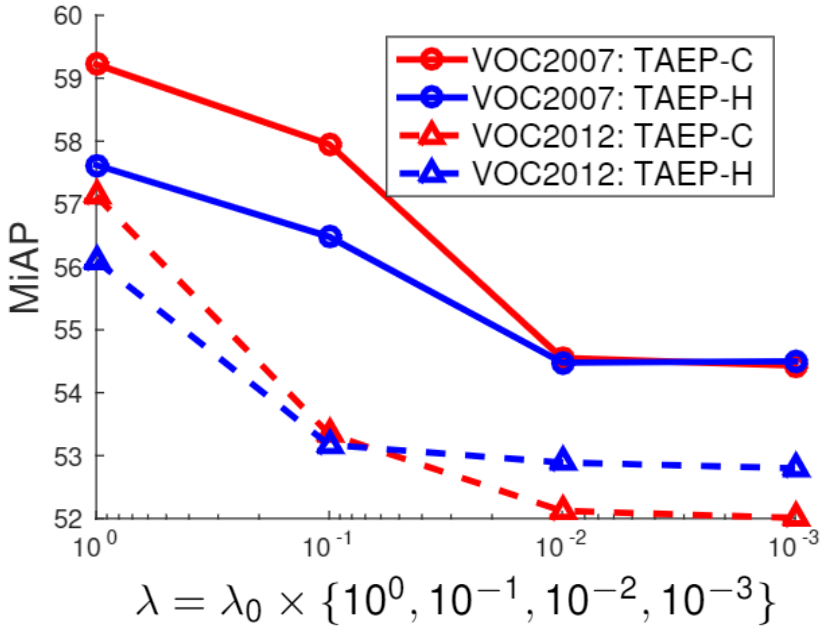
Experiments – Efficacy of TAEP

$$\mathcal{H}(U) = \frac{\gamma}{2}\mathrm{tr}\left(U^\top M^\top (Q + \frac{\lambda}{\gamma} Q^{\mathcal{A}}) MU\right)$$

γ and λ control the impact of *H*(*U*) term:



$\gamma = \gamma_0 \times \{10^0, 10^{-1}, 10^{-2}, 10^{-3}\}$



$\lambda = \lambda_0 \times \{10^0, 10^{-1}, 10^{-2}, 10^{-3}\}$

- Reducing the contribution of H(U) → performance drop.
- The TAEP regularization term is a useful component to facilitate the cross-class information in ZSL.

TEMPLE UNIVERSITY Carleton UNIVERSITY

Conclusions

- A max-margin approach to deal with multi-label ZSL problem
- Using auxiliary information to regularize label embedding projection
- Different auxiliary sources can be incorporated easily
- Experiments under both standard and generalized ZSL setting

Conclusions

- A max-margin approach to deal with multi-label ZSL problem
- Using auxiliary information to regularize label embedding projection
- Different auxiliary sources can be incorporated easily
- Experiments under both standard and generalized ZSL setting

Future Works

- Move on to large-scale datasets
- Exploit label relations
- Use powerful end-to-end deep model

TEMPLE UNIVERSITY Carleton UNIVERSITY

Thank You !


MULTI-LABEL ZERO-SHOT LEARNING WITH TRANSFER-AWARE LABEL EMBEDDING PROJECTION

Meng Ye* Yuhong Guo*

*Temple University, USA *Carleton University, Canada

ABSTRACT

1. INTRODUCTION

2. PROPOSED APPROACH

2.1. Problem Definition and Notations

2.2. Max-margin Multi-label Learning with Semantic Embedding Co-Projection

2.3. Transfer-Aware Label Embedding Projection

2.4. Integration of Auxiliary Information

3. EXPERIMENTS

3.1. Experimental Setting

3.2. Multi-label Zero-shot Learning Results

3.3. Impact of Label Embedding Regularization

4. CONCLUSION

5. REFERENCES

check our paper for more details !

Meng Ye
meng.ye@temple.edu