Matrix Tri-Factorization with Manifold Regularizations for Zero-shot Learning

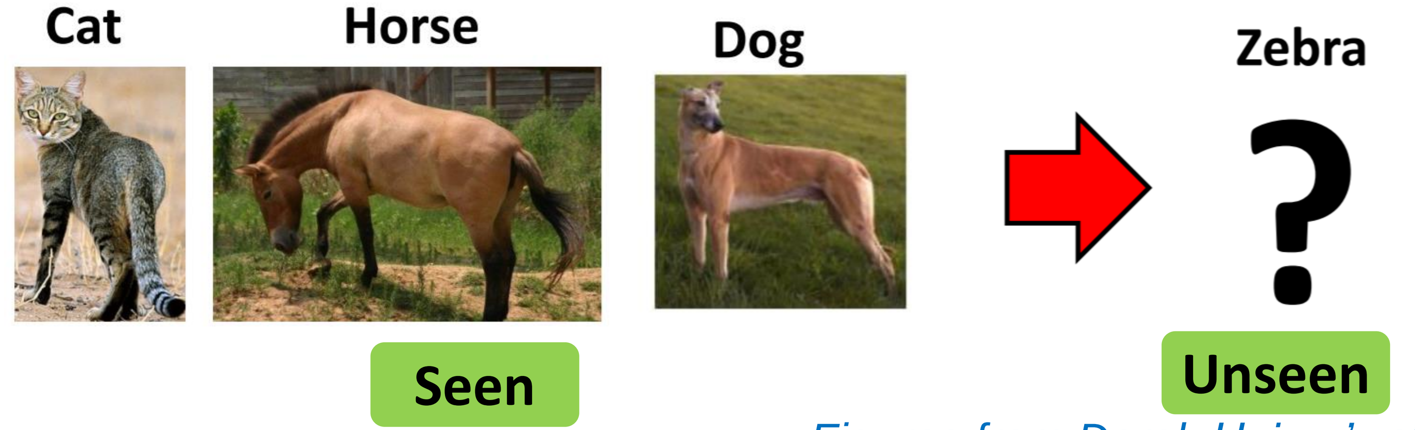
Xing Xu, Fumin Shen, Yang Yang, Dongxiang Zhang, Heng Tao Shen and Jingkuan Song
Center For Future Media & School of Computer Science and Engineering
University of Electronical Science and Technology of China, China

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

未来媒体研究中心
CENTER FOR FUTURE MEDIA

Background

- **Zero-shot learning (ZSL)**



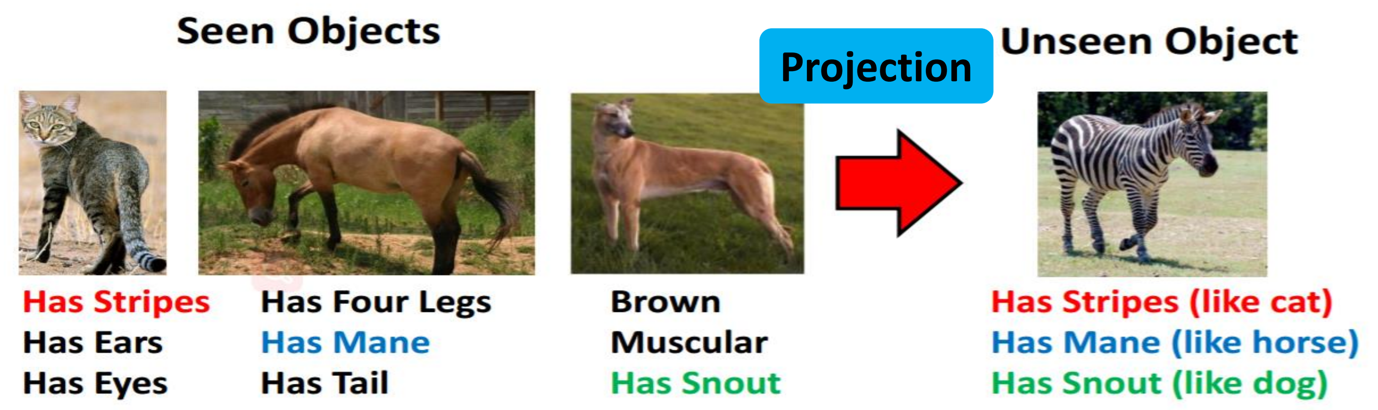
Figures from Derek Hoiem's slides

Two types of classes
- **Seen**: with labeled instances(for training only)
- **Unseen**: without instances (for testing only)

Goal
- Recognizing unseen object classes based on the knowledge learned from seen object classes during training.

Semantic embeddings
- **Q:** How to relate seen and unseen classes?
  A: Utilizing semantic embeddings (attributes, word vectors) to describe each object class (including seen and unseen ones).

Knowledge transfer to unseen classes
- Learning a projection from the visual feature space to the semantic embedding space based on seen classes, and apply it to unseen classes.

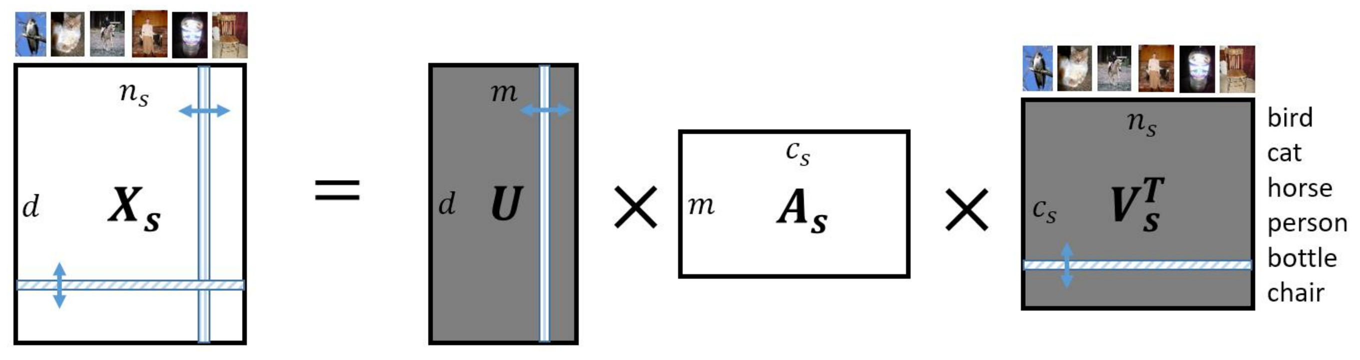


Challenges in learning the projection
- The intrinsic manifold structure in the semantic embeddings of classes is not well explored.
- Projection shift problem exists due to the different distribution of seen and unseen classes.

Proposed Method

- **Matrix Tri-Factorization with Manifold Regularization (MFMR)**

Learning the projection from visual features of seen classes



$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X}_s - \mathbf{U}\mathbf{A}_s\mathbf{V}_s^\top\|^2 + R(\mathbf{U}) + R(\mathbf{V})$$

- $\boldsymbol{U} \in \boldsymbol{R}^{d\times m}$ is the projection, each $\boldsymbol{u}_i$ represents a visual feature cluster for each semantic embedding (e.g., attributes).
- $\boldsymbol{A}_s \in \boldsymbol{R}^{m\times c_s}$ is the semantic embeddings (e.g., attributes).
- $\boldsymbol{V}_s \in \boldsymbol{R}^{n_s\times c_s}$, each $\boldsymbol{v}_i$ represents an instance cluster for each seen class.

Modeling the manifold structure of seen classes instances

- Two graph regularizers are introduced based on the row- and column-wise decomposition of the visual feature matrices, to preserve geometric strucutures.

- Column-wise (instance graph regularizer)

$$R^I_\pi = \frac{1}{2}\sum_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|^2 (\mathbf{W}^I_\pi)_{ij} = tr(\mathbf{V}_\pi^\top(\mathbf{Q}^I_\pi - \mathbf{W}^I_\pi)\mathbf{V}_\pi)$$

$$(\mathbf{W}^I_\pi)_{ij} = \begin{cases} cos(\mathbf{x}^\pi_{*i}, \mathbf{x}^\pi_{*j}), & if\ \mathbf{x}^\pi_{*i} \in N_p(\mathbf{x}^\pi_{*j})\ or\ \mathbf{x}^\pi_{*j} \in N_p(\mathbf{x}^\pi_{*i}) \\ 0, & otherwise \end{cases}$$

- Row-wise (feature graph regularizer)

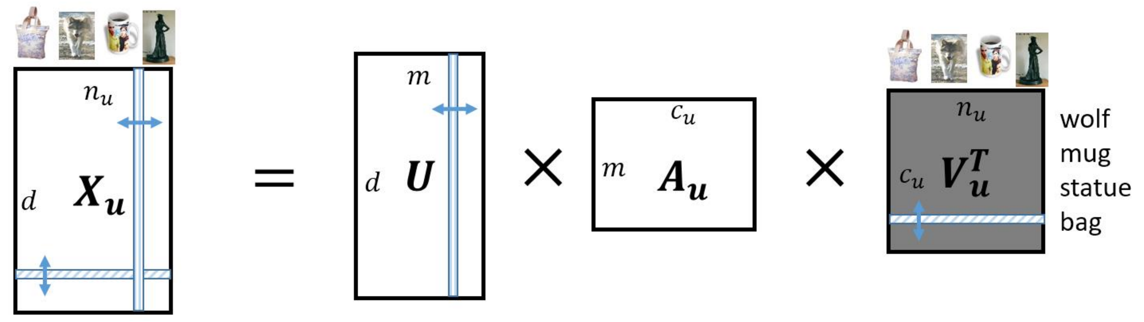
$$R^F_\pi = \frac{1}{2}\sum_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|^2 (\mathbf{W}^F_\pi)_{ij} = tr(\mathbf{U}_\pi^\top(\mathbf{Q}^F_\pi - \mathbf{W}^F_\pi)\mathbf{U}_\pi)$$

$$(\mathbf{W}^F_\pi)_{ij} = \begin{cases} cos(\mathbf{x}^\pi_{i*}, \mathbf{x}^\pi_{j*}), & if\ \mathbf{x}^\pi_{i*} \in N_p(\mathbf{x}^\pi_{j*})\ or\ \mathbf{x}^\pi_{j*} \in N_p(\mathbf{x}^\pi_{i*}) \\ 0, & otherwise, \end{cases}$$

Predicting the categories of unseen classes instances

- Simple prediction scheme (MFMR): $\mathbf{y}^u_j = \arg\min_l dist(\mathbf{U}^{-1}\mathbf{x}^u_j, \mathbf{a}^u_l), \quad l \in [1, c_u]$

- Joint prediction scheme (MFMR-joint):
  --- Exploiting the manifold structure in unseen classes instances



$$\min_{\mathbf{V}} \|\mathbf{X}_u - \mathbf{U}\mathbf{A}_u\mathbf{V}_u^\top\|^2 + R(\mathbf{V}_u) \Longrightarrow \mathbf{Y}_u = \arg\max_l(\mathbf{V}^l_u), \quad l \in [1, c_u]$$

Optimization Details can be referred to our paper (id: 1440)

Experiments

- **Experimental settings**

Datasets

	AwA	CUB	aPY	SUN
Num. Samples (Seen / Unseen)	24,295 / 6,180	8,855 / 2,933	12,695 / 2,644	14,140 / 200
Num. Classes (Seen / Unseen)	40 / 10	150 / 50	20 / 12	707 / 10
Num. Attributes	85	312	64	102
Image Feature	4096D VGG (CNN feature)			

Evaluation tasks and metrics
- Configuration: Conventional & Generalized ZSL
- Tasks: zero-shot classification & retrieval
- Metrics: Mean Average Precision (MAP)

Compared methods
- DAP[CVPR'09], ALE [CVPR'13], ESZSL [ICML'15], TMV-HLP [TPAMI'15], SSE [ICCV'15], JSLE [CVPR'2016], SynC [CVPR'16]

- **Results on "Conventional" Setting**

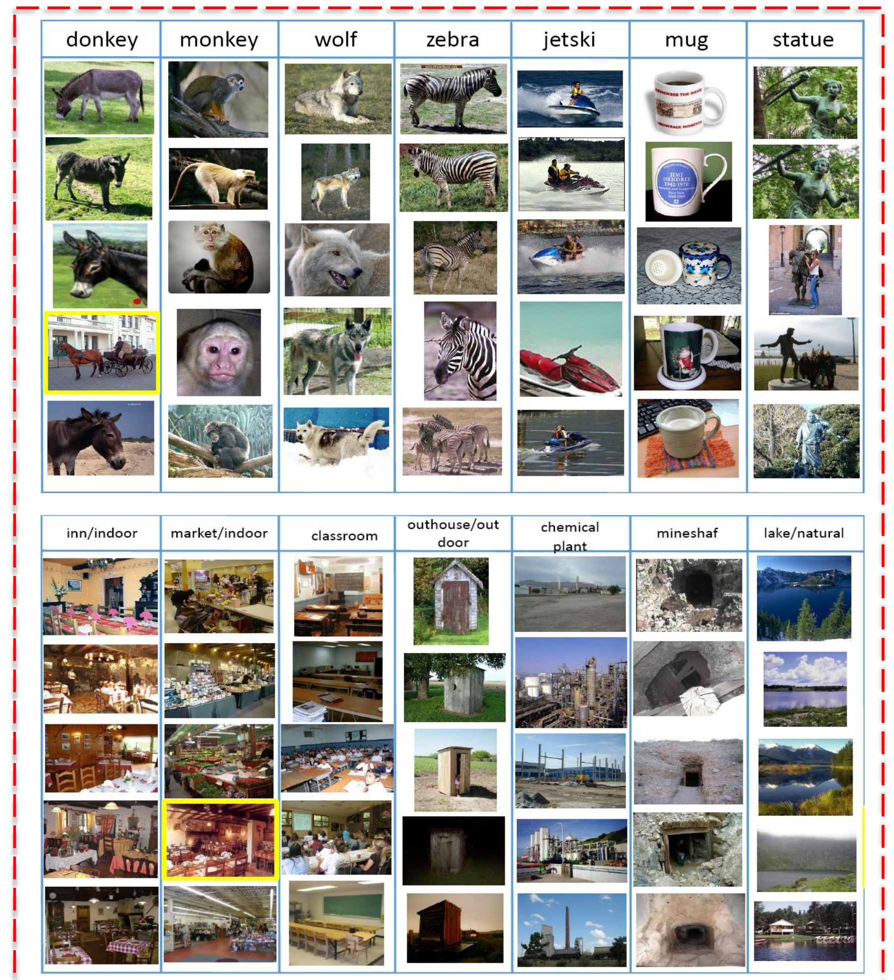
MAP scores on Zero-shot classification task

	AwA	CUB	aPY	SUN	Average
DAP	57.2	-	38.2	72.0	-
ALE	61.9	40.3	-	-	-
ConSE	61.6	-	37.6	-	-
ESZSL	75.3	-	24.2	82.1	56.5
TMV-HLP	80.5	47.9	-	-	-
SSE-INT	71.5	30.2	44.2	82.2	57.0
SSE-ReLU	76.3	30.4	46.2	82.5	58.9
JSLE	79.1	41.8	50.3	83.8	63.8
SynC-ova	77.3	48.8	47.2	79.5	63.2
SynC-struct	78.8	50.3	48.9	81.5	64.8
MFMR	79.8	47.7	48.2	84.0	64.9
MFMR-joint	**83.5**	**53.6**	**56.8**	**84.5**	**69.6**

MAP scores on Zero-shot retrieval task

	AwA	CUB	aPY	SUN	Average
SSE-INT	46.2	4.7	15.4	58.9	31.3
SSE-ReLU	44.6	3.7	14.1	44.6	26.2
JSLE	66.5	23.9	32.7	76.5	49.9
Sync-ova	64.3	30.4	29.6	72.1	49.1
Sync-struct	65.4	34.3	30.4	74.3	51.1
MFMR	70.8	30.6	45.6	77.4	56.2
MFMR-joint	**82.8**	**47.5**	**55.9**	**83.2**	**67.4**

- The propose MFMR and MFMR-joint are superior to the state of the art methods on "conventional" zero-shot classification and retrieval task.



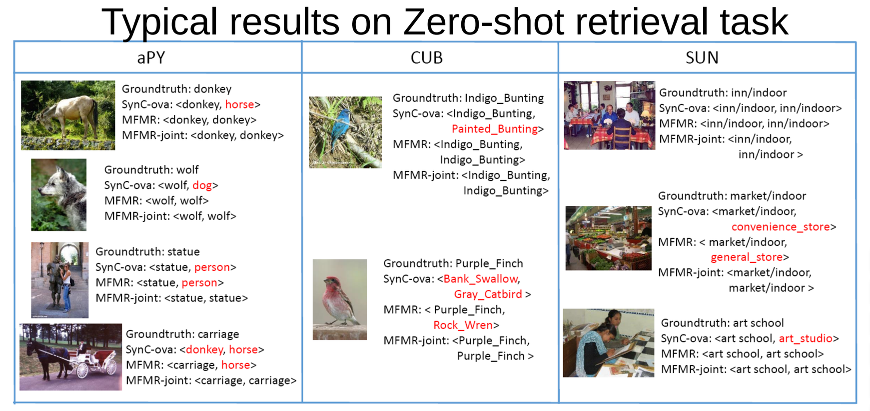
Typical retrieval result of our MFMR-joint on aPY (above) and SUN (below)

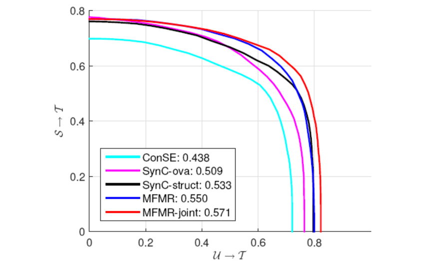
- **Results on "Generalized" Setting**

MAP scores on Zero-shot retrieval task

	U->U	S->S	U->T	S->T	AUSUC
ConSE	72.1	72.1	9.8	69.8	0.438
SynC-ova	76.4	77.6	1.1	75.7	0.509
SynC-struct	79.6	76.8	1.8	**76.1**	0.533
MFMR	79.9	76.1	13.4	75.6	0.550
MFMR-joint	**81.2**	**76.9**	**18.4**	75.6	**0.571**

Typical results on Zero-shot retrieval task



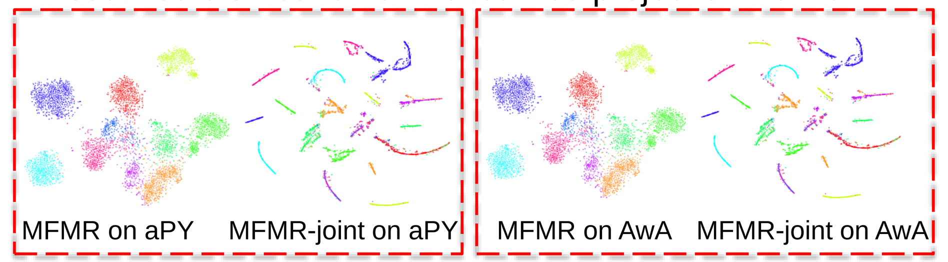


The Seen-Unseen accuracy Curve (SUC)

- The propose MFMR and MFMR-joint consistently perform the best on "Generalized" setting.
- The propose MFMR and MFMR-joint predict more accurate results on "generalized" zero-shot classification task.

- **Detailed analysis**

Visualized distribution of the learned projection via t-SNE



- The proposed MFMR-joint well explores the manifold structures of the unseen target data X_t and obtain discriminative embeddings V_t.