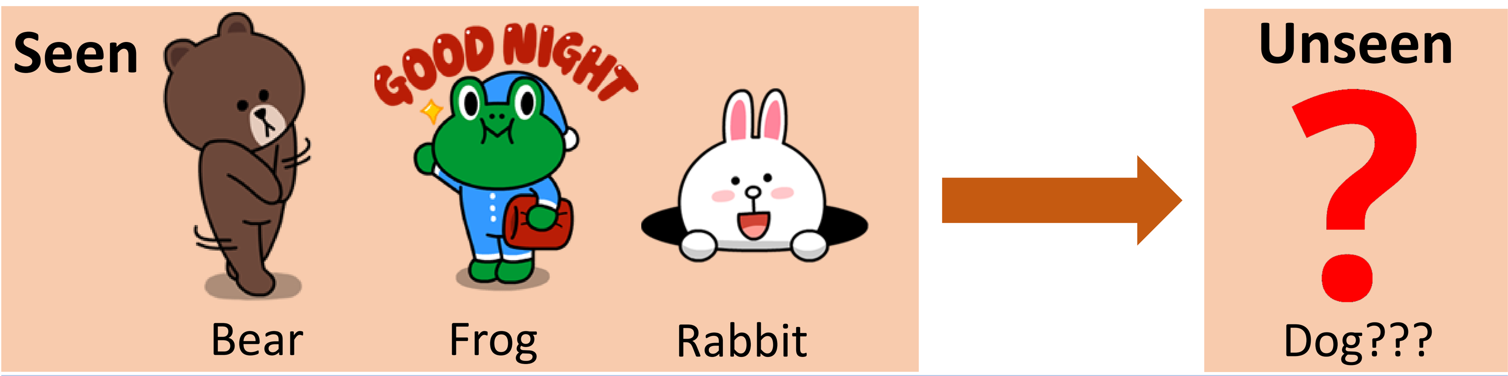
Semantics-Preserving Locality Embedding for Zero-Shot Learning

Carnegie Mellon University Language Technologies Institute
MACHINE LEARNING DEPARTMENT

Shih-Yen Tao[1], Yao-Hung Hubert Tsai[2], Yi-Ren Yeh[3], Yu-Chiang Frank Wang[4]
{[1]Language Technology Institute, [2]Machine Learning Department}, Carnegie Mellon University, USA
[3]Department of Mathematics, National Kaohsiung Normal University, Taiwan
[4]Department of Electrical Engineering, National Taiwan University, Taiwan

Introduction

- Zero-Shot Learning: Recognize images of unseen categories



Attribute	Bear	Frog	Rabbit	Dog???
Black Nose	o	x	x	o
Long Leg	o	o	x	x
Furry	o	x	o	o

- Each class is represented by a semantic vector
- ✓ Supervised: Attributes
- ✓ Unsupervised: Word2Vec, Glove, Wordnet Vector

Highlights

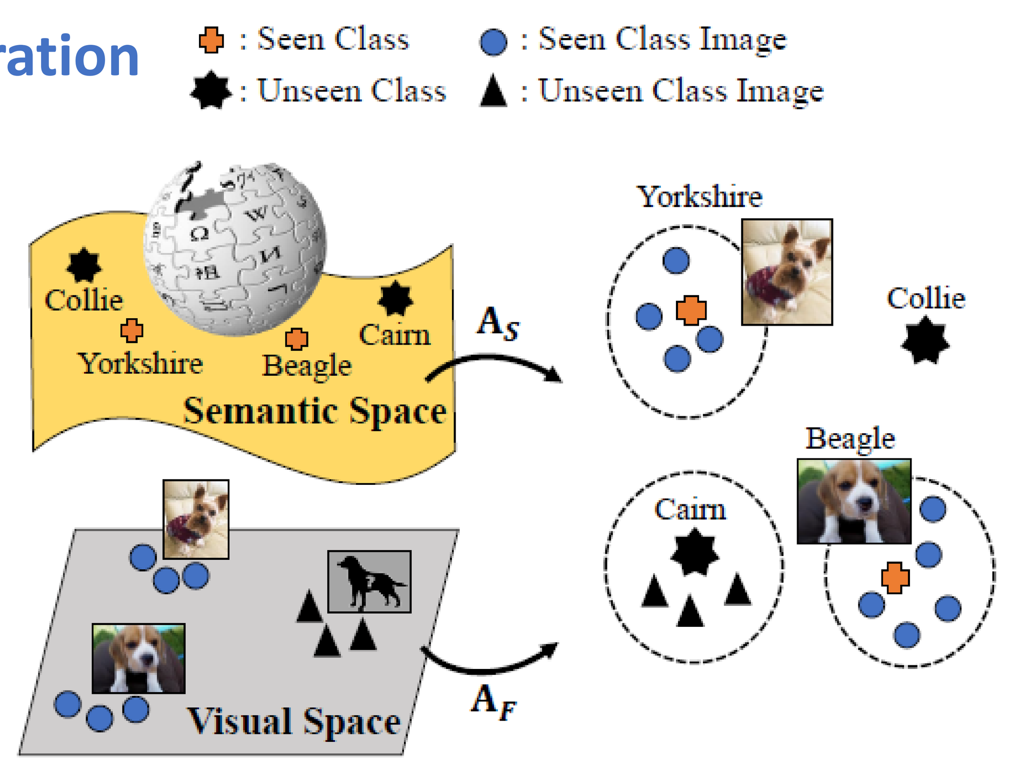
- Subspace learning via matching cross-domain concepts
- Semantics-Preserving Locality Embedding exploits the locality of within-class image data with the semantics jointly embedded.
- Work in both inductive & transductive settings
- Closed-form solution via eigen-decomposition

Related Works

- Inductive setting
- ✓ ESZSL[1], LatEm[2], SSE[3], Sync[4], JLSE[5], SOC[6], Devise[7]
- Transductive setting
- ✓ TMV[8], SMS[9]

Approach

Illustration



Notations

- ✓ Seen image data $D = \{X, Y\} = \{x_i, y_i\}_{i=1}^{N}, x_i \epsilon R^{d_f}$
- ✓ Unseen image data $D^U = \{X^U, Y^U\} = \{x_i^U, y_i^U\}_{i=1}^{N^U}, x_i^U \epsilon R^{d_f}$
- ✓ Y and Y^U come from disjoint label sets $L = \{1,2,\ldots,C\}$ and $L^U = \{1^U, 2^U, \ldots, C^U\}$
- ✓ Semantic vectors for seen and unseen classes $S = \{s_i \epsilon R^{d_s}\}_{i=1}^{C}, S^U = \{s_i^U \epsilon R^{d_s}\}_{i=1^U}^{C^U}$

Goal

- ✓ Find transformations $A_F \epsilon R^{d_F \times d_k}$ and $A_S \epsilon R^{d_S \times d_k}$ for visual and semantic space
- ✓ Zero-shot classification can be done in the resulting subspace

Semantics-Preserving Locality Embedding

- ✓ Objective function:

$$\min E_C(A_S, A_F) + \rho_1 E_S(A_F) + \rho_2 \sigma(A_S, A_F) \longleftarrow \text{l}_2 \text{ regularizier}$$
$$s.t.\, ZHZ^T = I, \longleftarrow \text{Maximize the variance of projected data}$$
$$where\ Z = [A_S^T S, A_F^T X], H\ is\ the\ centering\ matrix$$

- ✓ Concept matching: Extract cross-domain common concept
  - ◆ Visual concept: Class mean
  - ◆ Semantic concept: Semantic vector

$$E_C(A_S, A_F) = \sum_{j=1}^{C} \left\| A_S^T s_j - \frac{1}{N_j} \sum_{i=1}^{N_j} A_F^T x_i^j \right\|^2$$

- ✓ Within-class locality: More compact of the local structure in same label

$$E_S(A_F) = \frac{1}{2} \sum_{j=1}^{C} \{ \frac{1}{N_j^2} \sum_{i=1}^{N_j} \sum_{k=1}^{N_j} \| A_F^T x_i^j - A_F^T x_k^j \|^2 \}$$

- ✓ Remark: This results in improved separation between projected images of different labels

Zero-Shot Classification

$$y(x^U) = argmax \frac{\langle A_F^T x^U, A_S^T s_j^U \rangle}{\| A_F^T x^U \| \| A_S^T s_j^U \|}$$

From Inductive to Transductive Zero-Shot Learning

- ✓ Semantic vectors and images of unseen classes are represented in training stage
- ✓ Objective function:

$$\min E_C(A_S, A_F) + E_C^U(A_S, A_F) + \rho_1 \{ E_S(A_F) + E_S^U(A_F) \} + \rho_2 \sigma(A_S, A_F)$$

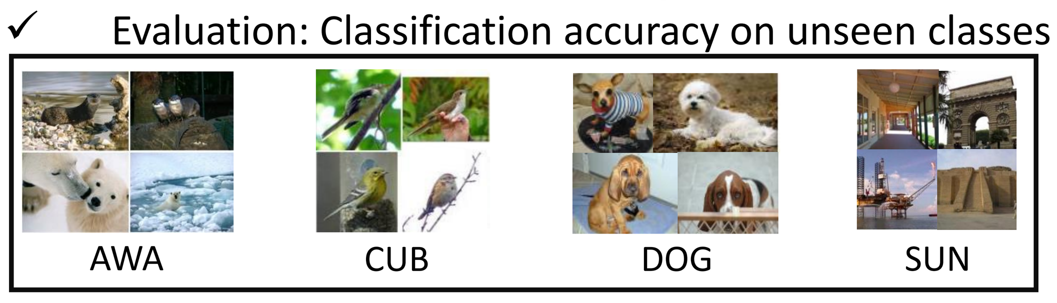
- ✓ Self-learning strategy: Update predicted label and transformations iteratively

Experiments

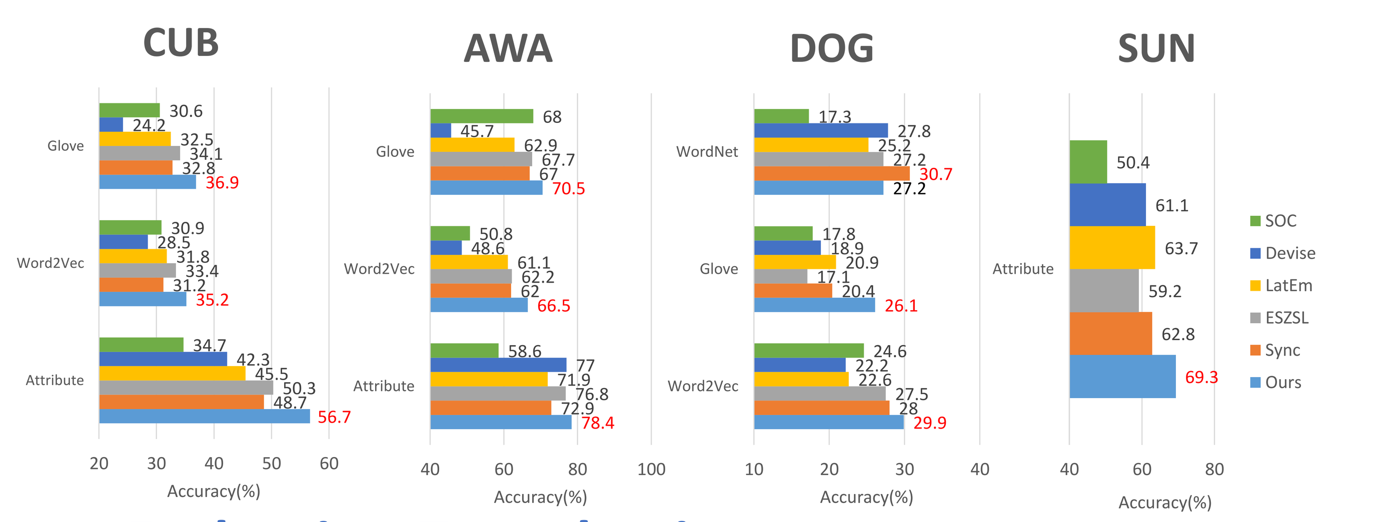
Datasets

	AWA	CUB	DOG	SUN
# of seen classes	40	150	85	645/646
# of unseen classes	10	50	28	72/71
# of images	30473	11786	19499	14340
Dim of Attributes	-	312	85	102
Dim of Word2Vec	400	400	400	-
Dim of Glove	400	400	200	-
Dim of Wordnet	-	-	163	-

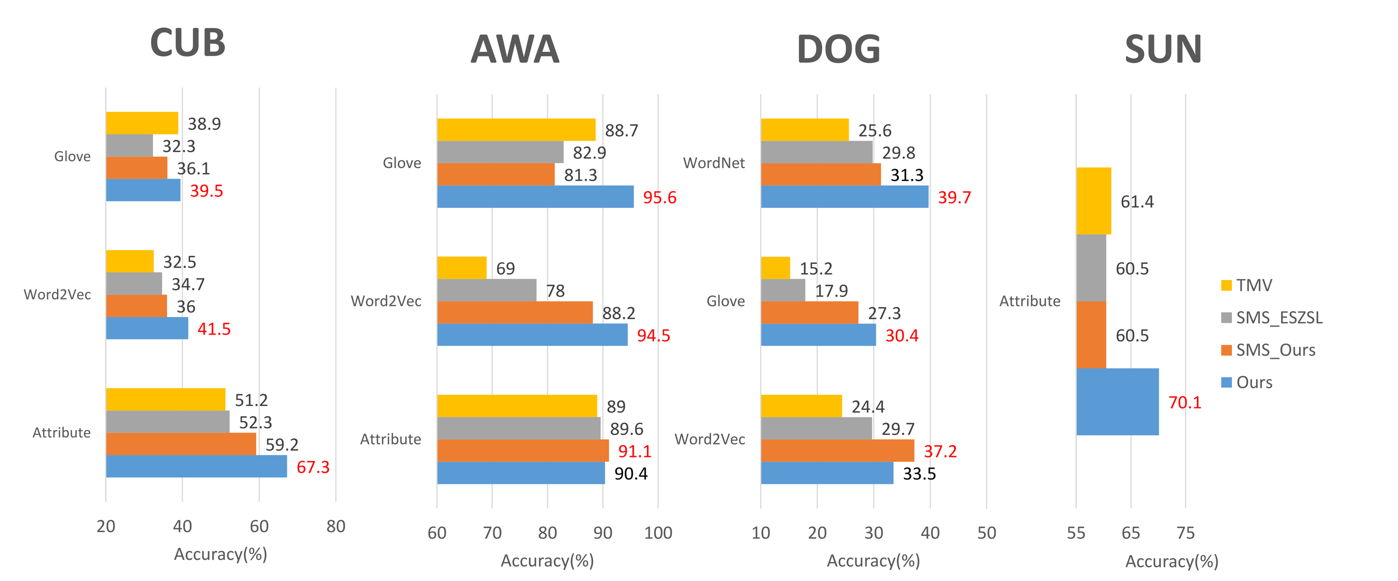
- ✓ Visual features: 1024-dim GoogLeNet feature
- ✓ Evaluation: Classification accuracy on unseen classes



Evaluation-Inductive

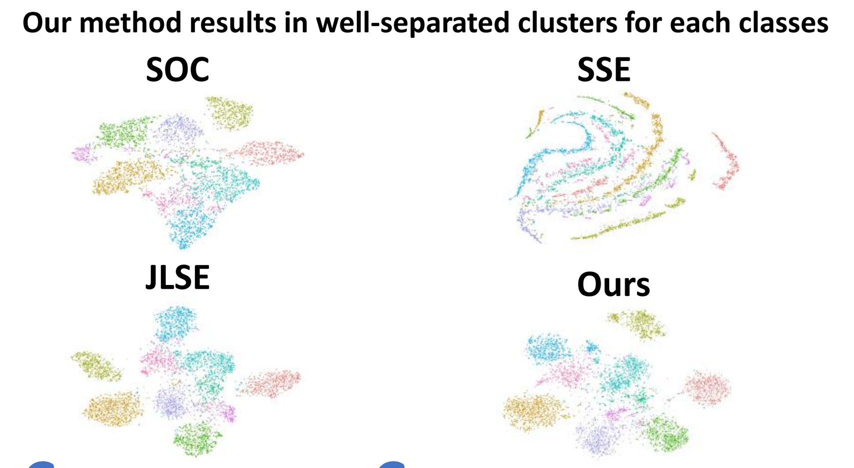


Evaluation-Transductive



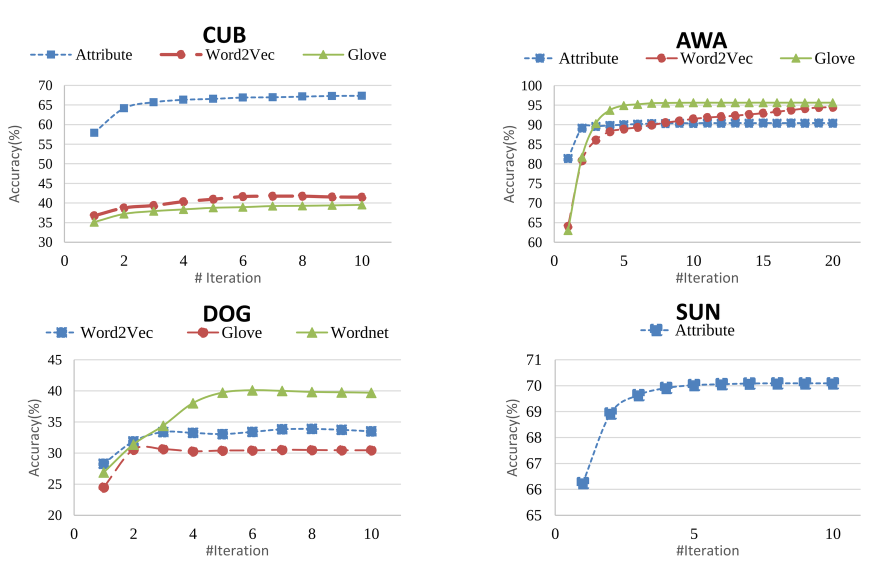
Visualization on AWA

- ✓ **t-SNE Visualization of different subspace learning methods**
- ✓ **Different colors represent different classes**
- ✓ **Our method results in well-separated clusters for each classes**



Convergence Curve

- ✓ Convergence analysis of the transductive setting



Conclusions

- Semantics-Preserving Locality Embedding for zero-shot classification task
- With-in class locality term improves separation between semantic data
- Our method can be easily generalized to transductive setting
- Promising results for both settings on four benchmark datasets

References

[1] Bernardino Romera-Paredes et al. An embarrassingly simple approach to zero-shot learning. In ICML, 2015.
[2] Yongqin Xian et al. Latent embeddings for zero-shot classification. In CVPR, 2016.
[3] Ziming Zhang et al. Zero-shot learning via semantic similarity embedding. In ICCV, 2015.
[4] Soravit Changpinyo et al. Synthesized classifiers for zero-shot learning. In CVPR, 2016.
[5] Ziming Zhang et al. Zero-shot learning via joint latent similarity embedding. In CVPR, 2016.
[6] Mark Palatucci et al. Zero-shot learning with semantic output codes. In NIPS, 2009.
[7] Andrea Frome et al. Devise: A deep visual-semantic embedding model. In NIPS, 2013.
[8] Yanwei Fu et al. Transductive multi-view zero-shot learning. TPAMI, 2015.
[9] Yuchen Guo et al. Transductive zeroshot recognition via shared model space learning. In AAAI, 2016.