Semantics-Preserving Locality Embedding for Zero-Shot Learning

Shih-Yen Tao[1], Yao-Hung Hubert Tsai[2], Yi-Ren Yeh[3], Yu-Chiang Frank Wang[4]

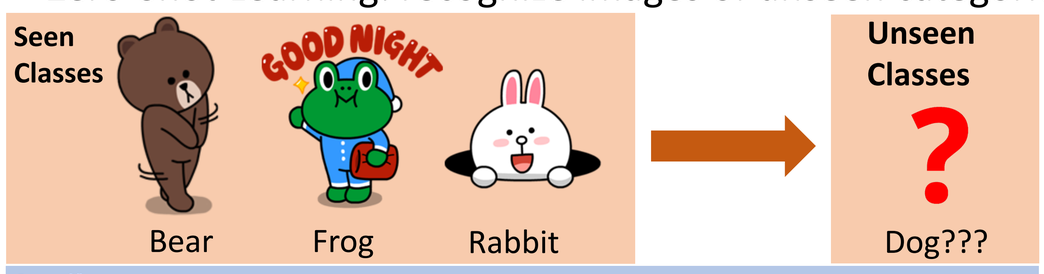
{[1]Language Technologies Institute, [2]Machine Learning Department}, Carnegie Mellon University, USA

[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



Attributes	Bear	Frog	Rabbit	Dog???
Black Nose	O	X	X	O
Long Leg	O	O	X	X
Furry	O	X	O	O

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

Highlights of Our Method

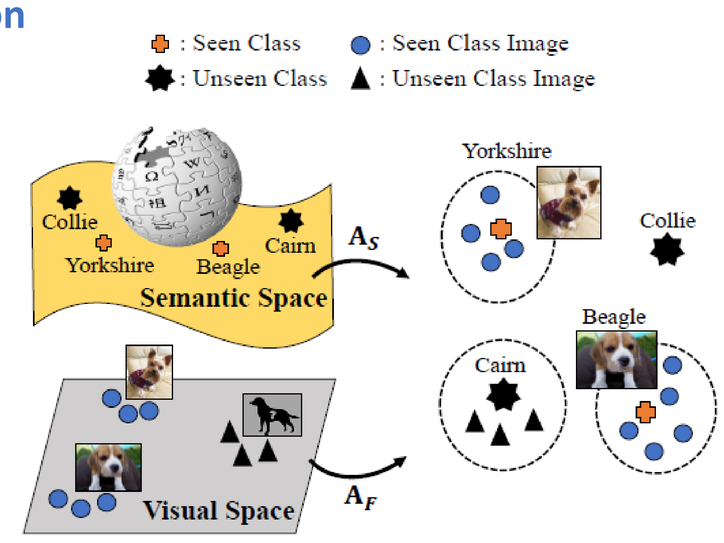
- Matching cross-domain concepts via subspace learning
- Semantics-preserving locality embedding exploits the locality of within-class image data with semantics info jointly embedded.
- Applicable in both inductive & transductive settings

Related Works

- **Inductive ZSL**
  ESZSL [ICML'15], LatEm [CVPR'16], SSE [ICCV'15], Sync [CVPR'16], JLSE [CVPR'16], SOC [NIPS'09], Devise [NIPS'13]
- **Transductive ZSL** (i.e., unseen test data presented & semantic vectors known)
  TMV[PAMI'15], SMS[AAAI'16]

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: disjoint label sets $L = \{1,2,..,C\}$ and $L^U = \{1^U, 2^U, \dots, C^U\}$
- ✓ Semantic vectors for seen/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 $A_S \epsilon R^{d_S \times d_k}$, respectively.
- ✓ Result in improved separation btw projected data of different labels.

Semantics-Preserving Locality Embedding

Objective function:

$$\min E_C(A_S, A_F) + \rho_1 E_S(A_F) + \rho_2 \sigma(A_S, A_F), s.t. ZHZ^T = I,$$

where $Z = [A_S^T S, A_F^T X]$, $\sigma(A_S, A_F)$ as L_2 regularizer, and H centering matrix.

Concept matching $E_C(A_S, A_F)$:

- Visual concept: Class mean
- Semantic concept: Semantic vector (1 for each class)

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

Within-class locality $E_S(A_F)$:

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

From Inductive to Transductive ZSL

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:

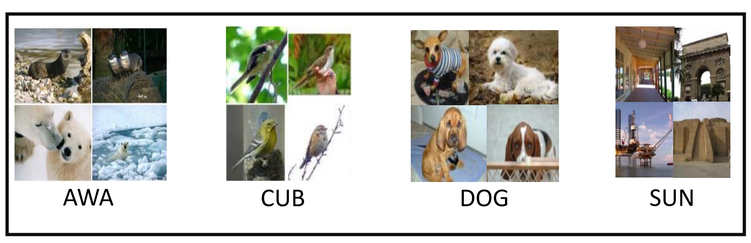
- ✓ Predict pseudo labels and update transformations iteratively

Experiments

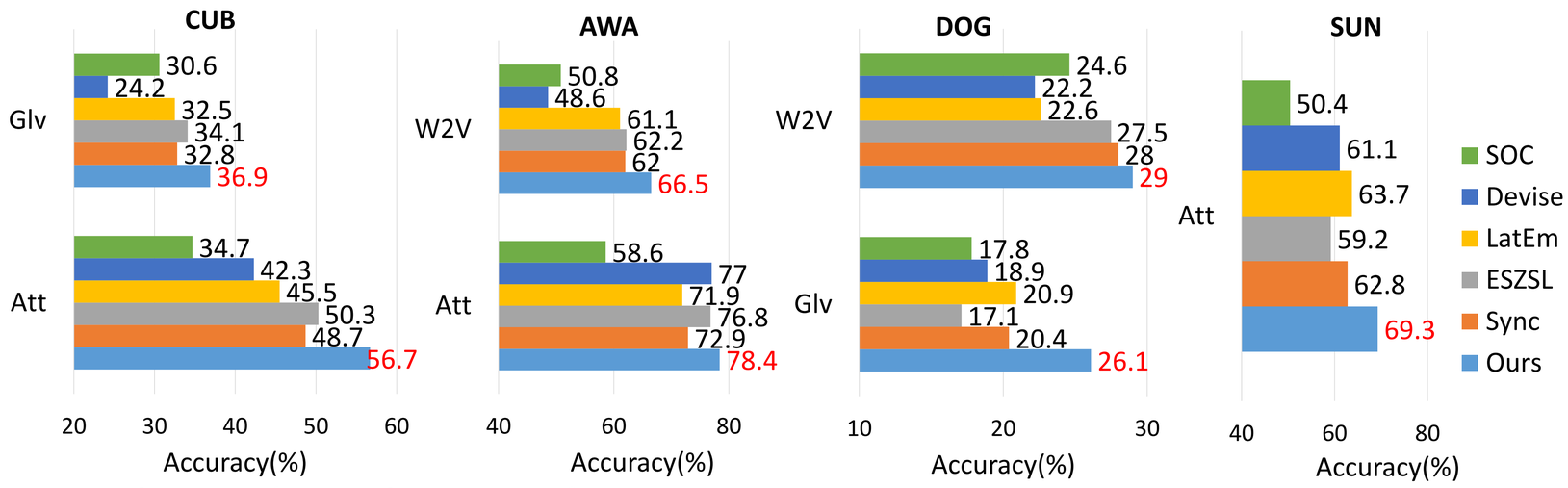
Datasets

	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	-

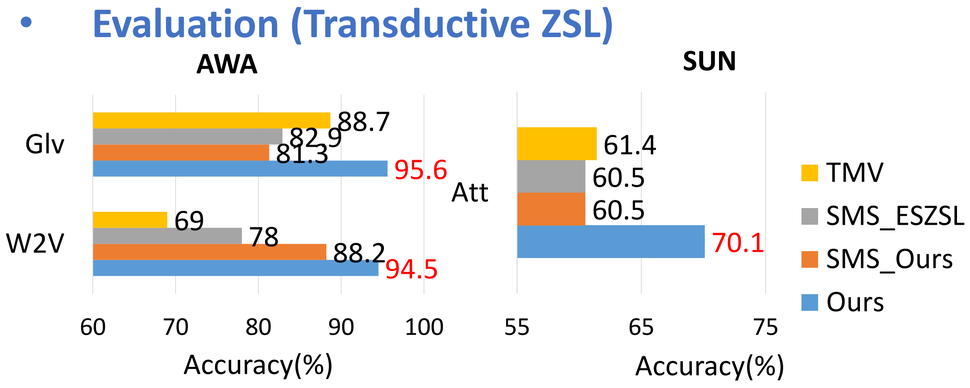
- ✓ Visual features: 1024-dim GoogLeNet feature
- ✓ Evaluation: Classification acc. of unseen classes



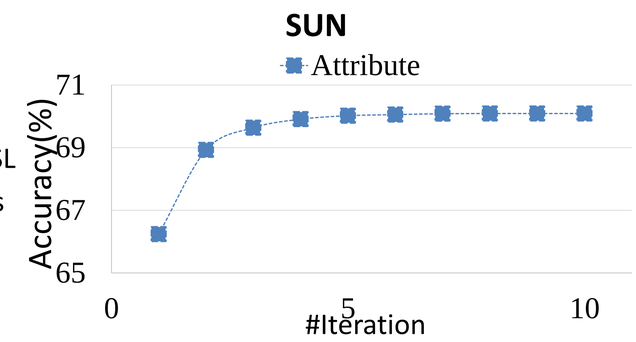
Evaluation (Inductive ZSL)



Evaluation (Transductive ZSL)

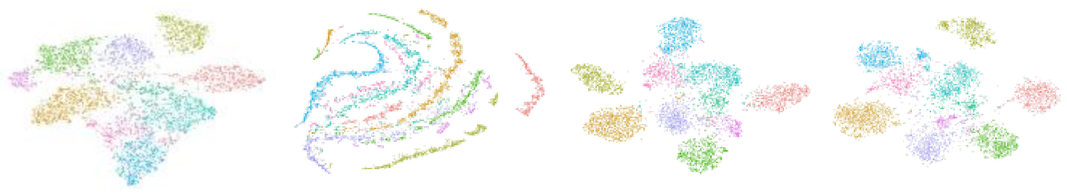


Convergence Analysis



Visualization

- ✓ Different colors denote different classes



Conclusions

- Semantics-preserving locality embedding:
  Concept matching + within-class locality
- Our method improves the separation between data of distinct semantic info, and thus is particularly preferable for ZSL.
- Our method can be easily generalized to the transductive setting.
- Promising inductive/transductive ZSL results on benchmark datasets.