Zero-Shot Learning via Joint Latent Similarity Embedding

Ziming Zhang [1,2,3] & Venkatesh Saligrama [1]

[1] Data Science & Machine Learning Lab, College of Engineering, Boston University, Boston, USA
[2] School of Automation, Huazhong University of Science and Technology, Wuhan, China
[3] MOE Key Laboratory of Image Information Processing and Intelligent Control, Huazhong University of Science and Technology, Wuhan, China

Zero-shot Learning Problem:

- **Training stage**: *source* domain attributes and *target* domain data corresponding to only a subset of classes (**seen** classes) are given.
- **Test stage**: *source* domain attributes for **unseen** (i.e. no training data provided) classes are then revealed.
- **Task**: for **unseen** data, match *target* domain unseen instances with *source* domain descriptors (classes).

Our Approach:

Key Insight---Binary Hypothesis Testing

- For a source and target domain instance pair:
  $H_0: y^{(s)} = y^{(t)}$ or $y^{(st)} \triangleq \mathbf{1}\{y_i = y_j\} = 1$
  $H_1: y^{(s)} \neq y^{(t)}$ or $y^{(st)} = 0$
- ***Class-independent*** test.

Optimal Test Statistic:

$$f(\mathbf{x}^{(s)}, \mathbf{x}^{(t)}) \triangleq \log p(y^{(st)} \mid \mathbf{x}^{(s)}, \mathbf{x}^{(t)}) \underset{\text{Diff}}{\overset{\text{Ident}}{\gtrless}} \theta$$

Shared Latent Models

- Decompose the posterior distribution with latent source and target domain variables $\mathbf{z}^{(s)}$ and $\mathbf{z}^{(t)}$:

$$p(y^{(st)}|\mathbf{x}^{(s)}, \mathbf{x}^{(t)}) = \sum_{\mathbf{z}^{(s)}} \sum_{\mathbf{z}^{(t)}} p(y^{(st)}, \mathbf{z}^{(s)}, \mathbf{z}^{(t)}|\mathbf{x}^{(s)}, \mathbf{x}^{(t)})$$

Markov Chain:

$$X^{(s)} \leftrightarrow Z^{(s)} \leftrightarrow Y \leftrightarrow Z^{(t)} \leftrightarrow X^{(t)}$$

Factorization

- Conditional independence given the underlying class Y

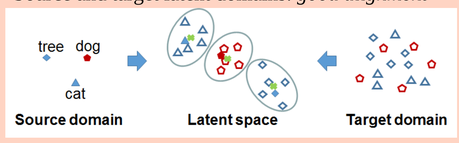
$$p(y^{(st)}|\mathbf{x}^{(s)}, \mathbf{x}^{(t)}) = \sum_{\mathbf{z}^{(s)}} \sum_{\mathbf{z}^{(t)}} p(\mathbf{z}^{(s)}|\mathbf{x}^{(s)}) p(\mathbf{z}^{(t)}|\mathbf{x}^{(t)}) p(y^{(st)}|\mathbf{z}^{(s)}, \mathbf{z}^{(t)})$$

  - ✓ Class-independent embedding
  - ✓ Class-independent similarity kernel
- Lower bound for computational efficiency

$$p(y^{(st)}|\mathbf{x}^{(s)}, \mathbf{x}^{(t)}) \geq \max_{\mathbf{z}^{(s)}, \mathbf{z}^{(t)}} p(\mathbf{z}^{(s)}|\mathbf{x}^{(s)}) p(\mathbf{z}^{(t)}|\mathbf{x}^{(t)}) p(y^{(st)}|\mathbf{z}^{(s)}, \mathbf{z}^{(t)})$$

Learning Principle

- Target latent domain: *good separation*.
- Source and target latent domains: *good alignment*



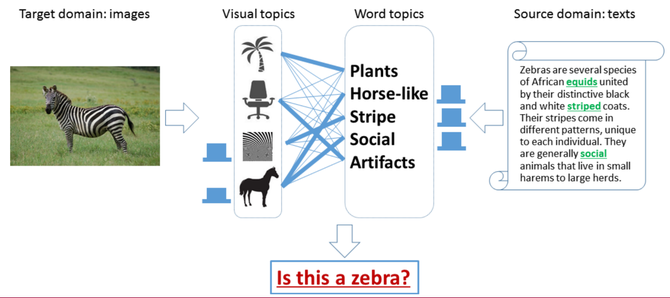
Procedure:

- Training:
  - ✓ Learn similarity function $f(x^{(s)}, x^{(t)})$.
- Testing:
  - ✓ Generate source and target domain latent vectors
- Decision function:
  - ✓ Multi-class classification:

$$y^* = \underset{c}{\operatorname{argmax}} f(x_c^{(s)}, x^{(t)})$$

Motivation:

- There exists a statistical relationship between latent co-occurrence patterns of corresponding source and target domain data pairs.
- This is a class-independent relationship => learn a "universal" similarity function



Parameterization I: Semantic Similarity Embedding (SSE) [ICCV 2015]

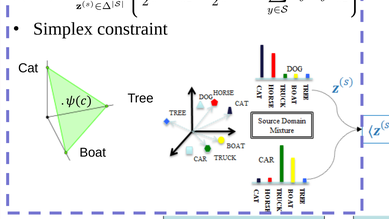
- **Intuition**: express source/target data as a mixture of seen class proportions.
- **Decision rule**: if the mixture portion from target domain is similar to that from source domain, then they must arise from the same class.

Source domain embedding ψ:

- Inspired by sparse coding

$$\psi(\mathbf{x}^{(s)}) \overset{\text{def}}{=} \underset{\mathbf{z}^{(s)} \in \Delta^{|\mathcal{S}|}}{\arg\min} \left\{ \frac{\gamma}{2} \|\mathbf{z}^{(s)}\|_2^2 + \frac{1}{2} \|\mathbf{x}^{(s)} - \sum_{y \in \mathcal{S}} \mathbf{x}_y^{(s)} z_y^{(s)}\|_2^2 \right\}$$

- Simplex constraint

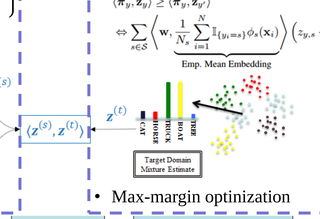


Target Domain Embedding π:

- $\pi_y(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi_y(\boldsymbol{x}) \rangle$
- Source-target domain alignment:

$$\langle \bar{\pi}_y, \mathbf{z}_y \rangle \geq \langle \bar{\pi}_{y'}, \mathbf{z}_{y'} \rangle$$

$$\Leftrightarrow \sum_{s \in \mathcal{S}} \left\langle \mathbf{w}, \underbrace{\frac{1}{N_s} \sum_{i=1}^{N} \mathbb{I}_{\{y_i = s\}} \phi_s(\mathbf{x}_i)}_{\text{Emp. Mean Embedding}} \right\rangle \left(z_{y,s} - z_{y',s} \right) \geq 0,$$



- Max-margin optimization

$\log p(\mathbf{z}^{(s)}|\mathbf{x}^{(s)})$ $\log p(y^{(st)}|\mathbf{z}^{(s)}, \mathbf{z}^{(t)})$ $\log p(\mathbf{z}^{(t)}|\mathbf{x}^{(t)})$

Parametrization II: Supervised Dictionary Learning (SDL) [CVPR 2016]

- **Intuition**: allow for *different* source and target domain latent spaces other than the semantic similarity space with *joint* learning of source/target domain embeddings.
- **Training**: estimate source domain dictionary **B**, target domain dictionary **D**, and cross-domain similarity matrix **W**.

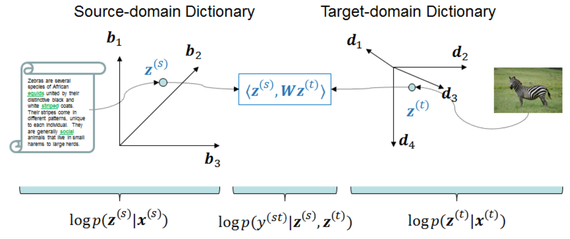
$$-\log p_B \overset{\text{def}}{=} \frac{\lambda_1^{(s)}}{2} \|\mathbf{z}_i^{(s)}\|_2^2 + \frac{\lambda_2^{(s)}}{2} \|\mathbf{x}_i^{(s)} - \mathbf{B}\mathbf{z}_i^{(s)}\|_2^2, \quad \text{s.t.} \quad \mathbf{z}^{(s)} > \mathbf{0}, \ \langle \mathbf{e}, \mathbf{z}^{(s)} \rangle = 1.$$

$$-\log p_D \overset{\text{def}}{=} \frac{\lambda_1^{(t)}}{2} \|\mathbf{z}_j^{(t)}\|_2^2 + \frac{\lambda_2^{(t)}}{2} \|\mathbf{x}_j^{(t)} - \mathbf{D}\mathbf{z}_j^{(t)}\|_2^2, \quad \text{s.t.} \quad \|\mathbf{D}_k\|_2^2 \leq 1, \ \forall k,$$

$$-\log p_W \overset{\text{def}}{=} \frac{\lambda_W}{2} \|\mathbf{W}\|_F^2 + \max\left\{0, 1 - \mathbf{1}_{y_{ij}^{(st)}} \langle \mathbf{z}_i^{(s)}, \mathbf{W}\mathbf{z}_j^{(t)} \rangle\right\}$$

  - ✓ Data-independent B, D, W
  - ✓ Alternate minimization
- **Testing:**

  Decoding in both domains to generate the source and target domain latent vectors. Similarity score in the latent spaces.

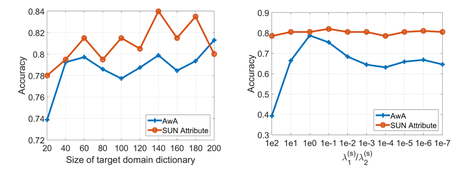


Experiments:

- **Dataset statistics**

Data	# img	# attr.	C/C'
aPY	15,339	64 (real)	20/12
AwA	30,475	85 (real)	40/10
CUB	11,788	312 (bin.)	150/50
SUN	14,340	102 (bin.)	707/10

Parameter Sensitivity

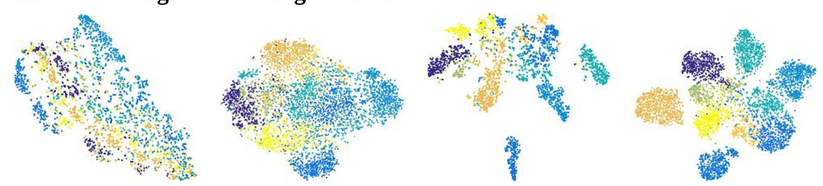


(a) (b)

- **Exp I: Zero-Shot Recognition**

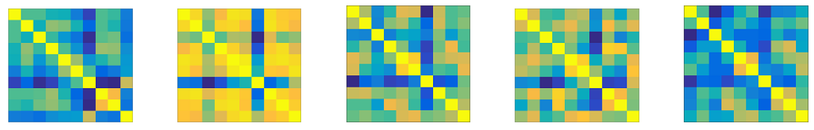
Method	aP&Y	AwA	CUB-200-2011	SUN Attribute	Ave.
Akata *et al.* [2]	-	61.9	***40.3***	-	-
Lampert *et al.* [19]	38.16	57.23	-	72.00	-
Romera-Paredes and Torr [32]	24.22±2.89	75.32±2.28	-	82.10±0.32	-
SSE-INT [44]	44.15±0.34	71.52±0.79	30.19±0.59	82.17±0.76	57.01
SSE-ReLU [44]	***46.23±0.53***	***76.33±0.83***	30.41±0.20	***82.50±1.32***	***58.87***
(i) init. $\forall \mathbf{z}_i^{(s)}, \forall \mathbf{z}_j^{(t)}$ + init. $\forall \mathbf{z}_{i'}^{(s)}, \forall \mathbf{z}_{j'}^{(t)}$ + Eq. 6	38.10±2.64	76.96±1.40	39.03±0.87	81.17±2.02	58.81
(ii) init. $\forall \mathbf{z}_i^{(s)}, \forall \mathbf{z}_j^{(t)}$ + init. $\forall \mathbf{z}_{i'}^{(s)}, \forall \mathbf{z}_{j'}^{(t)}$ + Eq. 7	38.20±2.75	80.11±1.13	41.07±0.81	81.33±1.76	60.20
(iii) init. $\forall \mathbf{z}_i^{(s)}, \forall \mathbf{z}_j^{(t)}$ + Alg. 2 + Eq. 6	47.29±1.45	74.92±2.51	38.94±0.81	80.67±2.57	60.46
(iv) init. $\forall \mathbf{z}_i^{(s)}, \forall \mathbf{z}_j^{(t)}$ + Alg. 2 + Eq. 7	47.79±1.83	77.37±0.39	40.91±0.86	80.83±2.25	61.73
(v) Alg. 1 + init. $\forall \mathbf{z}_{i'}^{(s)}, \forall \mathbf{z}_{j'}^{(t)}$ + Eq. 6	39.13±2.35	77.58±0.81	39.92±0.20	83.00±1.80	59.91
(vi) Alg. 1 + init. $\forall \mathbf{z}_{i'}^{(s)}, \forall \mathbf{z}_{j'}^{(t)}$ + Eq. 7	38.94±2.27	***80.46±0.53***	***42.11±0.55***	82.83±1.61	61.09
(vii) Alg. 1 + Alg. 2 + Eq. 6	50.21±2.90	76.43±0.75	39.72±0.19	83.67±0.29	62.51
(viii) Alg. 1 + Alg. 2 + Eq. 7	***50.35±2.97***	79.12±0.53	41.78±0.52	***83.83±0.29***	***63.77***

- **Visualization of target embedding on AwA**



(a) SSE: decaf (b) Ours: decaf (c) SSE: verydeep-19 (d) Ours: verydeep-19

- **Visualization of source domain embedding on AwA**



(a) Attributes (b) SSE: decaf (c) Ours: decaf (d) SSE: verydeep-19 (e) Ours: verydeep-19

- **Exp II: Zero-Shot Retrieval**

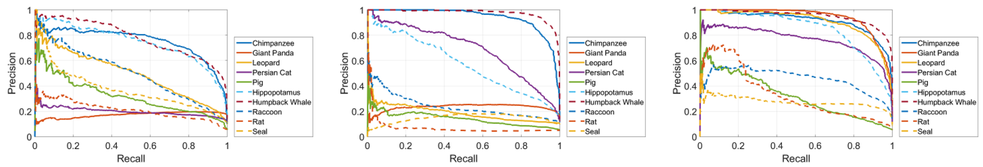
Table 3. Retrieval performance comparison (%) using mAP.

Method	aP&Y	AwA	CUB	SUN	Ave.
SSE-INT [44]	15.43	46.25	4.69	58.94	31.33
SSE-ReLU [44]	14.09	42.60	3.70	44.55	26.24
Ours	***38.30***	***67.66***	***29.15***	***80.01***	***53.78***

Table 4. Retrieval performance comparison (%) using AP on AwA.

Chim.	Panda	Leop.	Cat	Pig	Hipp.	Whale	Racc.	Rat	Seal	mAP
76.05	19.67	50.12	20.33	32.83	74.88	78.31	***50.52***	21.85	***37.96***	46.25
94.20	24.81	19.24	69.08	14.73	57.51	***97.56***	24.11	7.59	17.20	42.60
91.75	***94.06***	***91.09***	***76.95***	***33.00***	***84.85***	95.13	47.05	***34.58***	28.18	***67.66***

- **Precision-recall curves on AwA**



(a) SSE-INT (b) SSE-ReLU (c) Ours

Reference:

Zhang and Saligrama. "Zero-shot learning via semantic similarity embedding", ICCV. 2015.
Akata et al. Evaluation of output embeddings for fine-grained image classification. CVPR, 2015
Lampert et al. Attribute-based classification for zero-shot visual object categorization. PAMI, 2014.
Romera-Paredes and Torr. An embarrassingly simple approach to zero-shot learning. ICML, 2015.

Postdoc positions & code available: https://zimingzhang.wordpress.com/