Transductive Zero-Shot Learning with Visual Structure Constraint

Ziyu Wan Dongdong Chen Yan Li Xingguang Yan Junge Zhang Yizhou Yu Jing Liao

ziyuwan2-c@my.cityu.edu.hk github.com/raywzy/VSC

Introduction

- **Zero-Shot Learning:** Given N_s source labeled data samples $D_s \equiv \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and $y_i^s \in Y_s = \{1, \dots, S\}$ which is the corresponding label within total S source classes. We also have N_u unlabeled target samples $D_u \equiv \{x_i^u\}_{i=1}^{N_u}$ which are from $Y_u = \{S+1, \dots, S+U\}$. Each class is represented with a pre-defined auxiliary attribute vector $a_z \in A$. The goal of ZSL is to predict the label $y_i^u \in Y_u$ given x_i^u with no labeled training data.
- **Transductive Setting:** Assume the semantic information and visual features of all target classes are known in advance. Alleviate the domain shift problem according to this prior

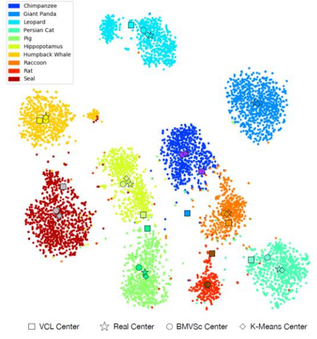
Motivation

- Discriminativity of pre-trained CNN:

  Visual features could be separated automatically.
- Domain Shift:

  Synthetic centers deviate from real centers.
- Learn a better mapping :

  Since no point-wise constraint could be directly utilized, how to capture the structure information to alleviate the domain shift problem?



Method

1. VCL (Visual Center Learning)

Semantic Space to Visual Space: MSE Loss of Source Domain:

$$c_i^{syn,s} = \sigma_2(w_2^T \sigma_1(w_1^T a_i^s)) \qquad \mathcal{L}_{MSE} = \frac{1}{S}\sum_{i=1}^{S} \|c_i^{syn,s} - c_i^s\|_2^2 + \lambda\Psi(w_1, w_2)$$

2. CDVSc (Chamfer-Distance-based VSC)

Inspired by 3D point clouds task, we propose use the symmetric Chamfer-distance to solve the structure matching problem .

$$\mathcal{L}_{CD} = \sum_{x \in C^{syn,u}} min_{y \in C^{clu,u}} \|x - y\|_2^2 + \sum_{y \in C^{clu,u}} min_{x \in C^{syn,u}} \|x - y\|_2^2$$

3. BMVSc (Bipartite-Matching-based VSC)

Consider one important prior for ZSL: The matching relation should conform strict one-to-one principle. To address the existed many-to-one matching in CDVSc, we propose to optimize the min-weight perfect matching problem using KM algorithm.

$$\mathcal{L}_{BM} = \min_X \sum_{i,j} dis_{ij} x_{ij}, \qquad s.t. \quad \sum_j x_{ij} = 1, \sum_i x_{ij} = 1, x_{ij} \in \{0,1\}$$

4. WDVSc (Wasserstein-Distance-based VSC)

Global optimal matching is not always valid especially when the approximated visual centers are not accurate enough. From the perspective of discrete distributions, we use the Wasserstein distance to measure the distance.

$$\mathcal{L}_{WD} = \min_X \sum_{i,j} dis_{ij} x_{ij} - \epsilon H(X) \text{ where } H(X) = -\sum_{ij} x_{ij} log x_{ij}$$

The Sinkhorn iterations could be written as

$$u^{(k+1)} = \frac{a}{Kv^{(k+1)}}, \qquad v^{(k+1)} = \frac{b}{K^T u^{(k+1)}}$$

Final objective: $\mathcal{L}_{WDVSc} = \mathcal{L}_{MSE} + \beta \times \mathcal{L}_{WD}$

Experiment

Experimental results on AwA1, AwA2, CUB, SUN72 and SUN10

Table 1: Quantitative comparisons of MCA (%) under standard splits (SS) in conventional ZSL setting. **I**: Inductive, **T**: Transductive, **O**: Our method, Bold: Best, Blue: Second best, V: VGG, R: ResNet, G: GoogLeNet

	Method	Features	AwA1	AwA2	CUB	SUN72	SUN10
I	CONSE [24]	R	63.6	67.9	36.7	44.2	–
	SSE [36]	V	76.3	–	30.4	–	82.5
	JLSE [37]	V	80.5	–	42.1	–	83.8
	SynC [4]	R	72.2	71.2	54.1	59.1	–
	SAE [16]	R	80.6	80.7	33.4	42.4	–
	SCoRe [23]	V	82.8	–	59.5	–	–
	f-CLSWGAN [33]	R	69.9	–	61.5	62.1	–
T	SP-ZSR [38]	V	92.0	–	53.2	–	86.0
	DSRL [34]	V	87.2	–	57.1	–	85.4
	DMaP [19]	V+G+R	90.5	–	67.7	–	–
	VZSL [31]	V	94.8	–	66.5	–	87.8
	QFSL [30]	V	–	84.1	61.2	–	–
O	**VCL**	V	81.7	82.6	58.2	58.8	87.2
	CDVSc	V	89.6	93.3	69.9	59.7	90.6
	BMVSc	V	92.7	94.0	70.8	61.3	89.7
	WDVSc	V	92.9	94.2	71.0	62.3	91.2
	VCL	R	82.0	82.5	60.1	63.8	89.6
	CDVSc	R	94.3	93.9	**74.2**	64.5	90.5
	BMVSc	R	95.9	**96.8**	73.6	66.2	91.7
	WDVSc	R	**96.2**	96.7	**74.2**	**67.8**	**92.2**

Table 2: Quantitative comparisons under the proposed splits (PS).

Method	AwA2	CUB	SUN72	Ave.
CONSE [24]	44.5	34.3	38.8	39.2
DeViSE [7]	59.7	52.0	56.5	56.0
SJE[2]	61.9	53.9	53.7	56.5
SynC [4]	46.6	55.6	56.3	52.8
SAE [16]	54.1	33.3	40.3	42.5
SCoRe [23]	69.5	61.0	51.7	60.7
LDF[20]	–	69.2	–	–
PSR-ZSL[3]	63.8	56.0	61.4	60.4
DCN [21]	–	56.2	61.8	–
VCL	61.5	59.6	59.4	60.1
CDVSc	78.2	71.7	61.2	70.3
BMVSc	81.7	71.0	62.2	71.6
WDVSc	**87.3**	**73.4**	**63.4**	**74.7**

Table 3: Quantitative comparisons under generalized ZSL setting.

	AwA2			CUB		
Method	acc_{y_u}	acc_{y_s}	H	acc_{y_u}	acc_{y_s}	H
CONSE [24]	0.5	**90.6**	1.0	1.6	72.2	3.1
SSE [36]	8.1	82.5	14.8	8.5	46.9	14.4
DeViSE [7]	17.1	74.7	27.8	23.8	53.0	32.8
SJE[2]	8.0	73.9	14.4	23.5	59.2	33.6
ESZSL[28]	5.9	77.8	11.0	12.6	63.8	21.0
SynC [4]	10.0	90.5	18.0	11.5	70.9	19.8
ALE[1]	14.0	81.8	23.9	23.7	62.8	34.4
PSR-ZSL[3]	20.7	73.8	32.3	24.6	54.3	33.9
VCL	21.4	89.6	34.6	15.6	**86.3**	26.5
CDVSc	66.9	88.1	76.0	37.0	84.6	51.4
BMVSc	71.9	88.2	79.2	33.1	86.1	47.9
WDVSc	**76.4**	88.1	**81.8**	**43.3**	85.4	**57.5**

Reference

1. Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In ECCV, 2016.
2. L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In CVPR, 2017.
3. M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In NeurIPS, 2013.