Webly Supervised Learning Meets Zero-shot Learning: A Hybrid Approach for Fine-grained Classification

Li Niu, Ashok Veeraraghavan, and Ashu Sabharwal

CPVPR 2018

Park, MinKyu

2019.11.8

Dongguk University

Artificial Intelligence Laboratory

Why choose Webly supervised learning?

FUTURE DIRECTIONS

Combination with other learning paradigms. As a learning paradigm, zero-shot learning can be combined with other learning paradigms to solve a wider range of problems. For example, in [109], zero-shot learning is combined with *Webly supervised learning* to classify classes that just have noisy labeled images obtained from the Web. There are also some works on applying zero-shot learning methods to the paradigm of *few-shot learning* [2, 51, 66, 100, 138].

Besides the above classification problems, zero-shot learning can also be combined with machine learning approaches for other purposes. In [53], zero-shot learning is used as the prior of *active learning* to enhance the learning process. In [63], it is combined with *lifelong learning* to learn new tasks with only descriptions. In [55, 133, 167], zero-shot learning is combined with *hashing* and forms *zero-shot hashing* problems which aim to hash images of unseen classes. In [61, 71, 111], it is combined with *reinforcement learning* to better handle new tasks, new domains and new scenarios. More combinations of zero-shot learning with other learning paradigms can be explored in future research.

A Survey of Zero-Shot Learning: Settings, Methods, and Applications

Fine-grained image recognition?

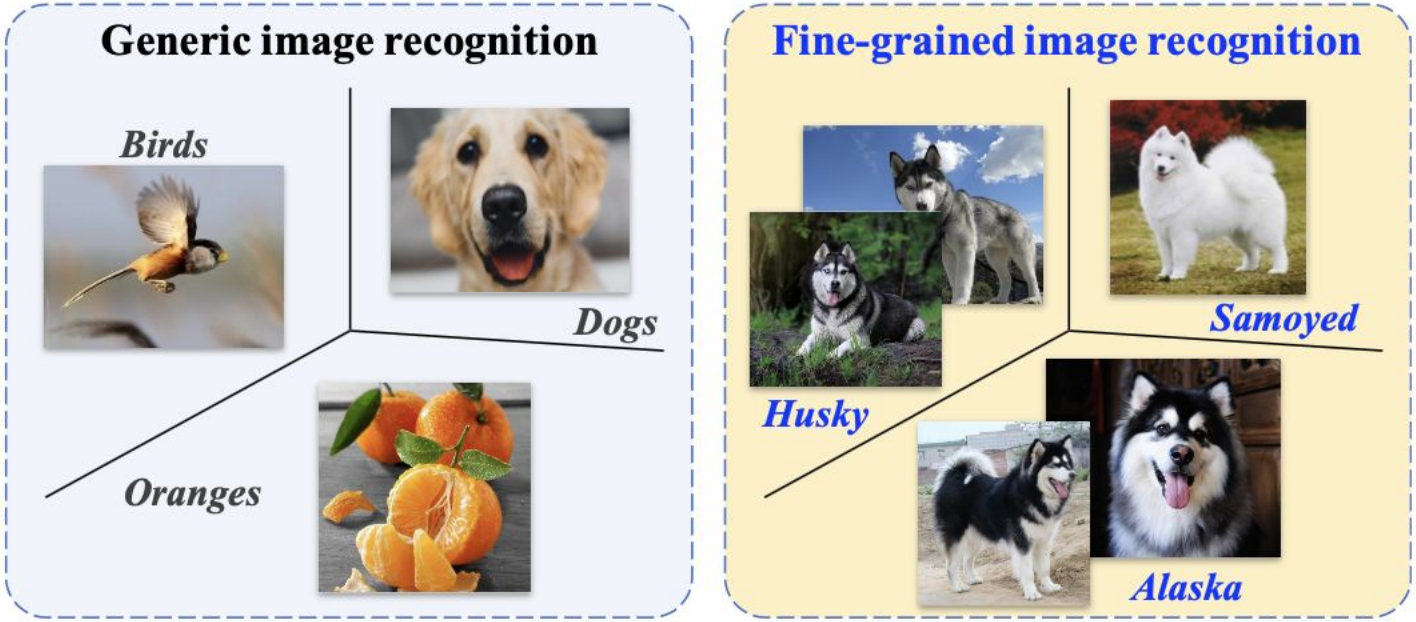
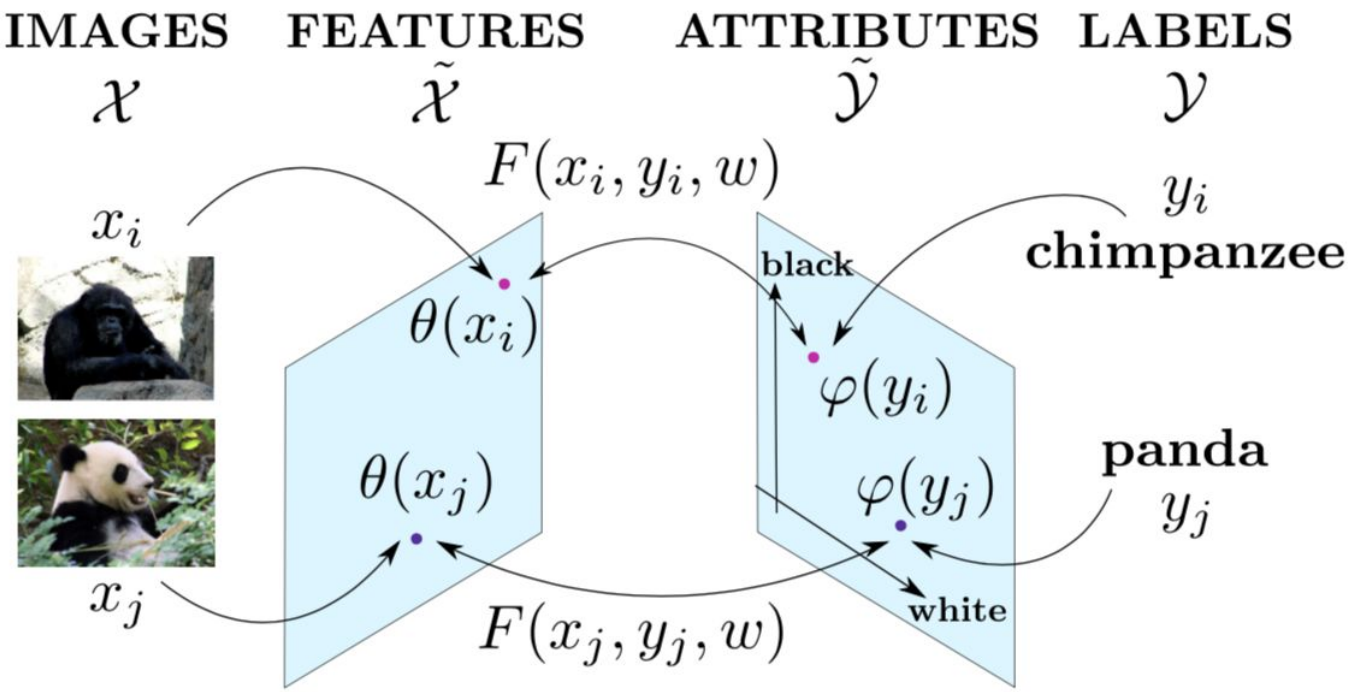


Figure 1: Fine-grained image analysis *vs.* generic image analysis (taking the recognitiont task for an example).

Deep learning for fine-grained image analysis: A survey

Label-Embedding for Image Classification



Label-Embedding for Image Classification

Motivation

- *targets at distinguishing subtle distinctions among various subordinate categories*

  ⇒ ***the scarcity*** *of well-labeled training images*

- reasons

  high demand of professional knowledge

  the number of subcategories belonging to one category is generally very huge

⇒ lack of well-labeled training images becomes **a critical issue** for fine-grained classification

Introduction

1. utilize freely available web images without human annotation(WSL)

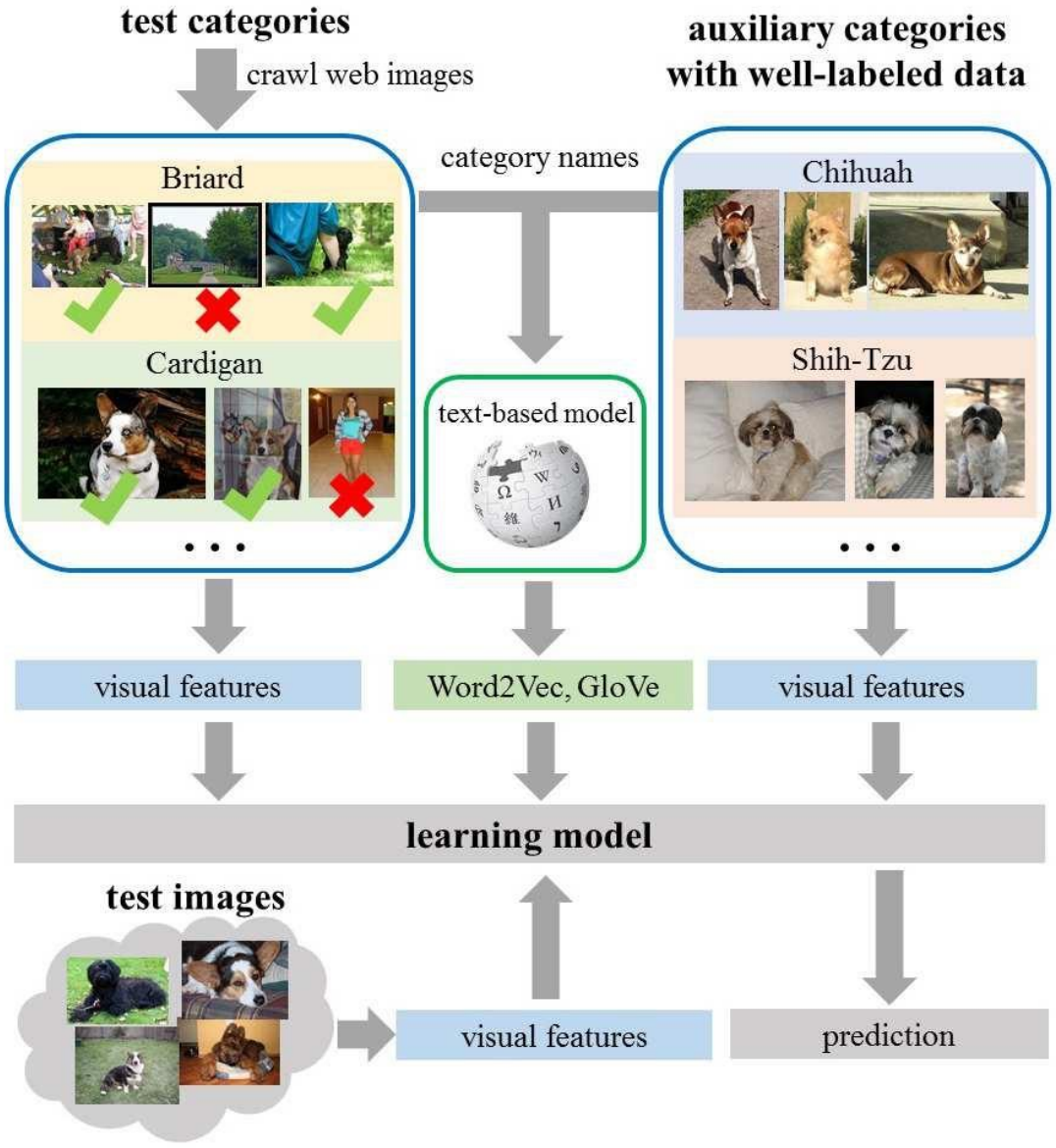
   *the labels of web images are **very noisy** and the data **distribution** between web images and test images are considerably different*

2. only annotate some fine-grained categories and transfer the knowledge to other fine-grained categories, which falls into the scope of zero-shot learning (ZSL)

   *the performance **gap** between ZSL and traditional supervised learning is still very large*

WSL + ZSL

Our framework



Our Formulation

A^{-1} : the inverse matrix of A , I : the identity matrix, O : zero matrix
$\langle A, B \rangle$: inner product , $(A \circ B)$: element-wise product

Train images C^a

Test images C^t

Web images C^w

Fully-supervised

Weakly-supervised

Visual feature $\begin{cases} X^a \in R^{d \times n^a} \\ X^t \in R^{d \times n^t} \\ X^w \in R^{d \times n^w} \end{cases}$

Well-labeled training data $A^a \in R^{m \times n^a}$

Test data $A^t \in R^{m \times n^t}$

Web data $A^w \in R^{m \times n^w}$

Semantic representation $\begin{cases} \bar{A}^a \in R^{m \times C^a} \\ \bar{A}^t \in R^{m \times C^t} \end{cases}$

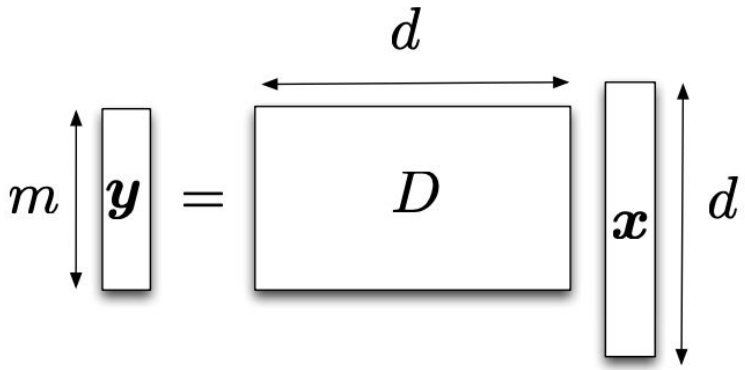
d is the dimension of visual feature,
n^a is the number of training images,
m-dim sematic representation

Comparing A^t with $\bar{A}^t$

Sparse Coding

- Approximate $\boldsymbol{y}$ using D and $\boldsymbol{x}$:

$$\boldsymbol{y} = D\boldsymbol{x}.$$



Basic Formulation(Matrix and Lagrangian function)

- Given observation signal Y, find latent signal/coefficients X and basis D minimizing the objective:

$$\min_{D,X} \|Y - DX\|_2^2 + \lambda \|X\|_{0,\infty}^{col}, \quad \lambda > 0.$$

Norm

임의의 $1 \leq p \leq \infty$에 대하여, 유클리드 공간 $\mathbb{R}^n$ 위에 다음과 같은 노름 $\|\cdot\|_p$을 정의할 수 있으며, 이를 ℓ^p **노름**이라고 한다.

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

여기서 $p = 2$인 경우는 표준적인 유클리드 노름

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

이다. 만약 $p = \infty$일 경우는 **상한 노름**(영어: supremum norm)

$$\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$$

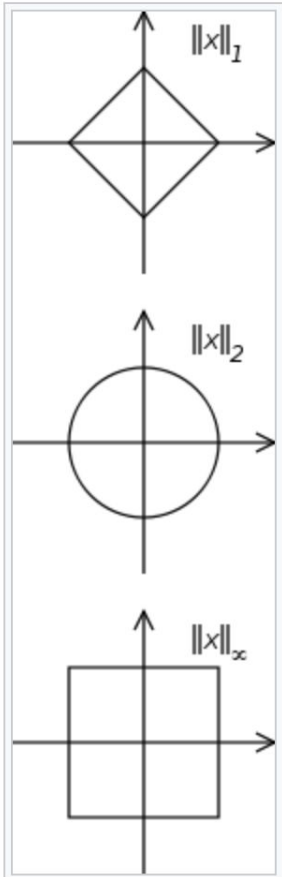
이 된다. $p = 1$인 경우는 **맨해튼 노름**

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

이 된다.

ℓ^p 노름 말고도 유클리드 공간 위에 수많은 노름들을 정의할 수 있다. 예를 들어, $\mathbb{R}^4$ 위에는 다음과 같은 노름이 존재한다.

$$\|x\| = 2|x_1| + \sqrt{3|x_2|^2 + \max(|x_3|, 2|x_4|)^2}$$



서로 다른 노름 공간에서 정의된 단위원.

https://eehoeskrap.tistory.com/227

Knowledge Transfer

*from **fully-supervised categories** to **weakly-supervised categories***

- First Stage

Learn the dictionary of fully-supervised categories

$$\min_{\mathbf{D}^a} \quad \frac{1}{2}\|\mathbf{X}^a - \mathbf{D}^a\mathbf{A}^a\|_F^2 + \frac{1}{2}\|\mathbf{D}^a\|_F^2, \qquad (1)$$

two visual-semantic dictionaries D^a and $D^t \in R^{d\times m}$

Unsupervised domain adaptation for zero-shot learning, in ICCV 2015

Knowledge Transfer

*from **fully-supervised categories** to **weakly-supervised categories***

- First Stage

$$\min_{\mathbf{D}^t,\mathbf{A}^t} \frac{1}{2}\|\mathbf{X}^t-\mathbf{D}^t\mathbf{A}^t\|_F^2+\frac{\lambda_1}{2}\|\mathbf{D}^t-\mathbf{D}^a\|_F^2+\lambda_2\|\mathbf{A}^t\|_*, \quad (2)$$

- Minimize the mapping error on the test images
- Enforce D^t to be close to D^a
- Expect A^t to be low-rank (convex approximation of rank function)

Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*

Utilizing Noisy Web Images

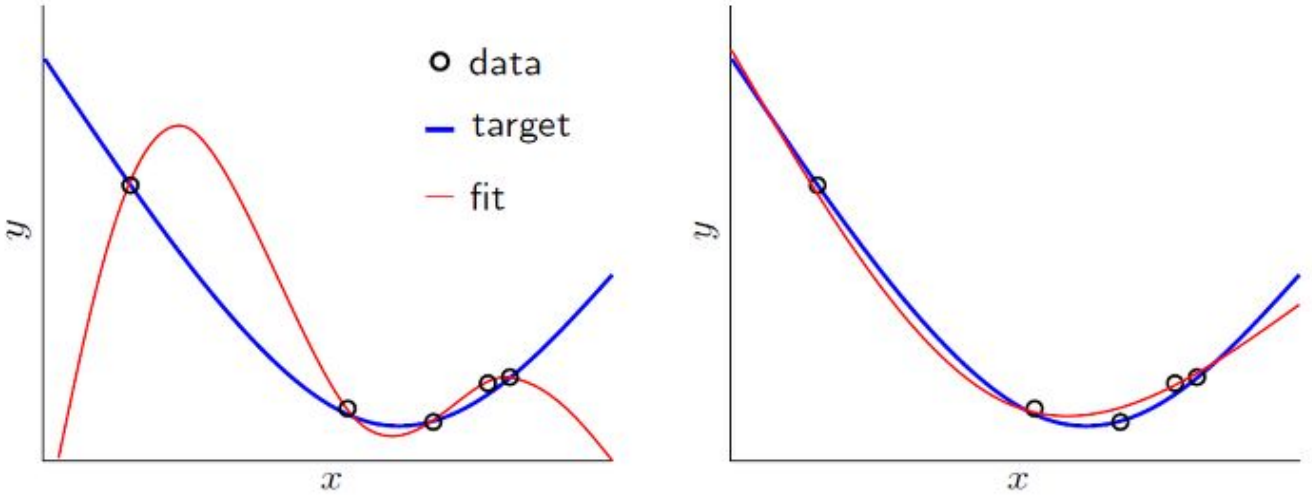
- The domain shift between **web images** and **test images**

Maximum Mean Discrepancy (MMD) [15] based regularizer

$$\left\| \frac{1}{n^w} \mathbf{X}^w \boldsymbol{\theta} - \frac{1}{n^t} \mathbf{X}^t \mathbf{1} \right\|^2$$

To reduce the distance between the center of weighted web images and the center of test images

Regualrization



(a) without regularization (b) with regularization

Regularization : 모델 복잡도에 대한 패널티로 Regularization은 Overfitting 을 예방하고 Generalization(일반화) 성능을 높이는데 도움을 줌. 종류로는 L1 Regularization, L2 Regularization, Dropout, Early stopping 등이 있음

https://enginius.tistory.com/476

Utilizing Noisy Web Images

- The label noise of web images

The group-lasso regularizer

$$\|(\mathbf{X}^w - \mathbf{D}^t \mathbf{A}^w)\mathbf{\Theta}\|_{2,1}$$

we actually leverage auxiliary categories to help tackle the label noise of web images

Optimization

$$\min_{\mathbf{D}^t,\mathbf{A}^t,\boldsymbol{\theta}} \quad \frac{1}{2}\|\mathbf{X}^t-\mathbf{D}^t\mathbf{A}^t\|_F^2+\frac{\lambda_1}{2}\|\mathbf{D}^t-\mathbf{D}^a\|_F^2+\lambda_2\|\mathbf{Z}^t\|_*$$

$$+\frac{\lambda_3}{2}\|\frac{1}{n^w}\mathbf{X}^w\boldsymbol{\theta}-\frac{1}{n^t}\mathbf{X}^t\mathbf{1}\|^2+\lambda_4\|\mathbf{E}^w\|_{2,1}, \qquad (5)$$

$$\text{s.t.} \quad \mathbf{1}'\boldsymbol{\theta}=n^w, \quad \mathbf{0}\le\boldsymbol{\theta}\le b\mathbf{1},$$

$$\mathbf{E}^w=(\mathbf{X}^w-\mathbf{D}^t\mathbf{A}^w)\boldsymbol{\Theta}, \quad \mathbf{Z}^t=\mathbf{A}^t.$$

A novel solution based on inexact Augmented Lagrange Multiplier (ALM)

intermediate variable $\boldsymbol{E^w}$ to replaces $(X^w - D^t A^W)\,\Theta$

$\boldsymbol{Z^t = A^t}$

Optimization

- Minimize the following augmented Lagrangian function

$$
\begin{aligned}
\mathcal{L}_{\substack{\mathbf{D}^t,\mathbf{A}^t,\mathbf{Z}^t \\ \mathbf{E}^w,\boldsymbol{\theta}\in\boldsymbol{\mathcal{S}}}} =& \frac{1}{2}\|\mathbf{X}^t-\mathbf{D}^t\mathbf{A}^t\|_F^2+\frac{\lambda_1}{2}\|\mathbf{D}^t-\mathbf{D}^a\|_F^2+\lambda_2\|\mathbf{Z}^t\|_* \\
&+\frac{\lambda_3}{2}\|\frac{1}{n^w}\mathbf{X}^w\boldsymbol{\theta}-\frac{1}{n^t}\mathbf{X}^t\mathbf{1}\|^2+\lambda_4\|\mathbf{E}^w\|_{2,1} \\
&+\frac{\mu}{2}\|\mathbf{E}^w-(\mathbf{X}^w-\mathbf{D}^t\mathbf{A}^w)\boldsymbol{\Theta}\|_F^2+\left\langle\mathbf{R},\mathbf{E}^w-(\mathbf{X}^w-\mathbf{D}^t\mathbf{A}^w)\boldsymbol{\Theta}\right\rangle \\
&+\frac{\mu}{2}\|\mathbf{A}^t-\mathbf{Z}^t\|_F^2+\left\langle\mathbf{T},\mathbf{A}^t-\mathbf{Z}^t\right\rangle, \qquad (6)
\end{aligned}
$$

$$\boldsymbol{\mathcal{S}} = \{\theta|1'\theta = n^w, 0 \le \theta \le b1\}$$

$\boldsymbol{\mu}$: a penalty parameter

$\{\boldsymbol{R},\boldsymbol{T}\}$: Lagrangian multipliers

Optimization

Algorithm 1 Solving (5) with inexact ALM

Input: $\mathbf{X}^a, \mathbf{A}^a, \mathbf{X}^w, \mathbf{A}^w, \mathbf{X}^t, \mathbf{D}^a$.
Initialize $\mathbf{R} = \mathbf{O}, \mathbf{T} = \mathbf{O}, \boldsymbol{\theta} = \mathbf{1}, \mathbf{D}^t = \mathbf{D}^a, \rho = 0.1, \mu = 0.1, \mu_{max} = 10^6, \nu = 10^{-5}, N_{iter} = 10^6$.
for $t = 1 : N_{iter}$ **do**
    Update $\mathbf{E}^w$ by using (7).
    Update $\mathbf{Z}^t$ by using (9).
    Update $\mathbf{D}^t$ by using (11).
    Update $\mathbf{A}^t$ by using (12).
    Update $\boldsymbol{\theta}$ by solving (15).
    Update $\mathbf{R}$ by $\mathbf{R} = \mathbf{R} + \mu(\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t\mathbf{A}^w)\boldsymbol{\Theta})$.
    Update $\mathbf{T}$ by $\mathbf{T} = \mathbf{T} + \mu(\mathbf{A}^t - \mathbf{Z}^t)$.
    Update the parameter μ by $\mu = \min(\mu_{max}, (1+\rho)\mu)$.
    Break if $\|\mathbf{E}^w - (\mathbf{X}^w - \mathbf{D}^t\mathbf{A}^w)\boldsymbol{\Theta}\|_\infty < \nu$ and $\|\mathbf{A}^t - \mathbf{Z}^t\|_\infty < \nu$.
end for
Output: $\mathbf{A}^t$.

Experiments - Fine-grained Image Classification

- Dataset

  Caltech-UCSD Bird (CUB)

  Scene UNderstanding (SUN) attribute dataset

  Stanford Dogs dataset

  Flickr image dataset : queries to collect the top ranked 100 images from Flickr website for each category

- Features:

  Visual features

  - 4,096-dim output of the 6-th layer of the pretrained VGG model

  Semantic representations:

  - Two types of word vectors Word2Vec and GloVe
  - Train language models based on the latest Wikipedia corpus, with the word vector dimension being 400
  - Concatenate the word vectors, leading to an 800-dim vector for each category

Table 1: Accuracies (%) of different methods on three datasets. The best results are highlighted in boldface.

Dataset	CUB	SUN	Dogs	Avg
LR	68.39	62.50	77.67	69.52
KMM	70.54	64.00	79.16	71.23
GFK	70.37	62.50	79.51	70.79
SA	68.67	63.00	80.18	70.62
TCA	68.56	63.00	80.22	70.59
CORAL	69.04	63.50	80.37	70.97
NEIL	69.08	63.00	80.16	70.74
Bergamo and Torresani	70.13	64.00	78.64	70.93
WSDG	70.61	66.00	80.20	72.27
Sukhbaatar *et al.*	70.47	64.50	81.15	72.04
Xiao *et al.*	70.92	65.50	81.67	72.69
ESZSL	38.08	65.00	37.21	46.77
LatEm	35.15	66.50	35.99	45.88
SJE	42.65	71.50	34.85	49.67
DAP/IAP	28.91	57.50	33.15	39.85
Changpinyo *et al.*	41.83	72.00	39.91	51.25
Li *et al.*	32.36	72.50	43.15	49.34
Kodirov *et al.*	47.53	71.00	47.32	55.28
Zhang and Saligrama	44.08	76.50	48.09	56.23
Xu *et al.*	45.72	71.50	39.85	52.36
Shojaee and Baghshah	46.68	71.00	48.82	55.50
WSL+ZSL	72.21	78.50	81.90	77.53
Ours_WSL	69.42	65.50	80.43	71.78
Ours_ZSL	47.94	71.50	47.70	55.71
Ours_sim1	72.72	83.50	85.04	80.42
Ours_sim2	76.00	79.50	83.75	79.75
Ours	**76.47**	**84.50**	**85.16**	**82.04**

simply learns a linear regressor based on web training images

Domain adaptation(DA) baselines

WSL baselines

ZSL baselines

$$\min_{\mathbf{D}^t,\mathbf{A}^t,\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{X}^t-\mathbf{D}^t\mathbf{A}^t\|_F^2+\frac{\lambda_1}{2}\|\mathbf{D}^t-\mathbf{D}^a\|_F^2+\lambda_2\|\mathbf{Z}^t\|_* + \frac{\lambda_3}{2}\|\frac{1}{n^w}\mathbf{X}^w\boldsymbol{\theta}-\frac{1}{n^t}\mathbf{X}^t\mathbf{1}\|^2+\lambda_4\|\mathbf{E}^w\|_{2,1}, \quad (5)$$
$$\text{s.t.} \quad \mathbf{1}'\boldsymbol{\theta}=n^w, \quad \mathbf{0}\le\boldsymbol{\theta}\le b\mathbf{1},$$
$$\mathbf{E}^w=(\mathbf{X}^w-\mathbf{D}^t\mathbf{A}^w)\boldsymbol{\Theta}, \quad \mathbf{Z}^t=\mathbf{A}^t.$$

$\lambda_1 = 0$
$\lambda_3 = \lambda_4 = 0$
$\lambda_2 = 0$
$\lambda_4 = 0$

Experiments - Utilizing More Web Images

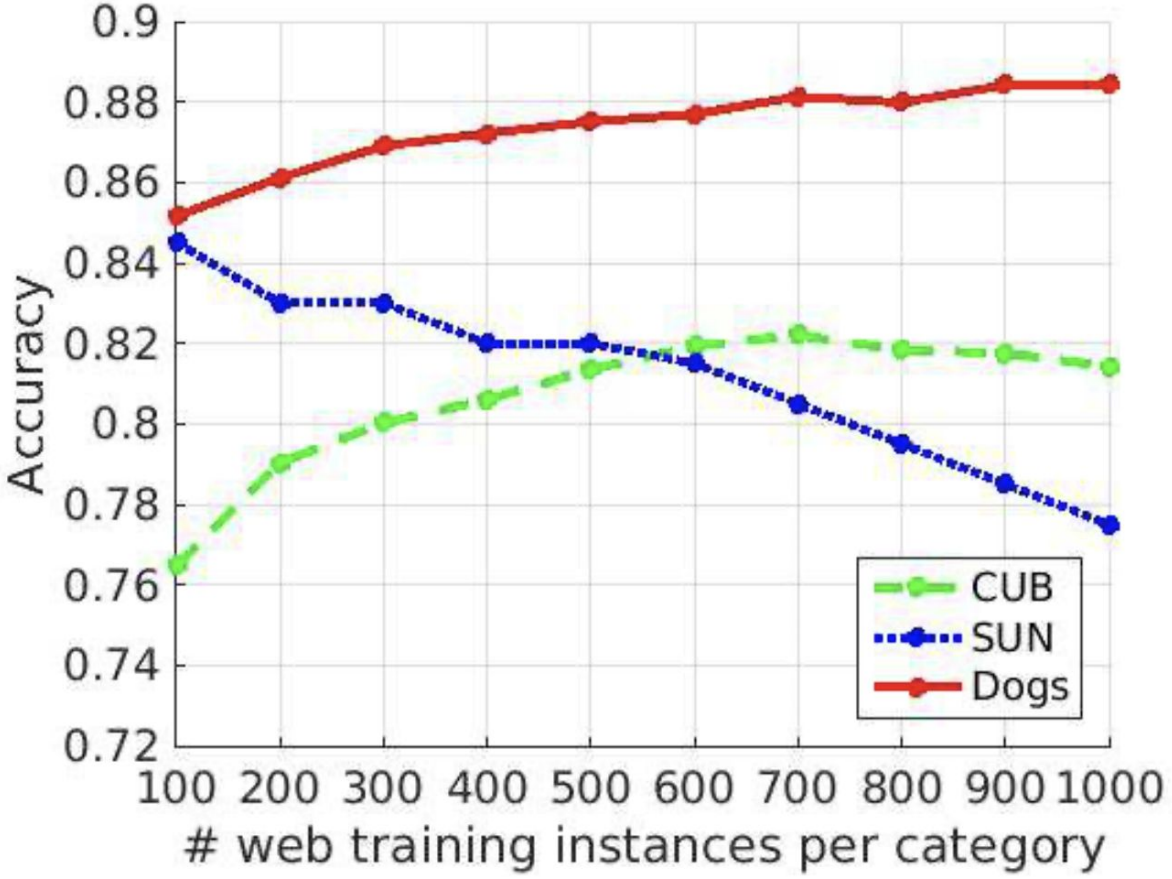


Figure 2: The performance variation of our method w.r.t. different numbers of web training images per category.

Experiments - Qualitative Analysis of Learnt Weights θ

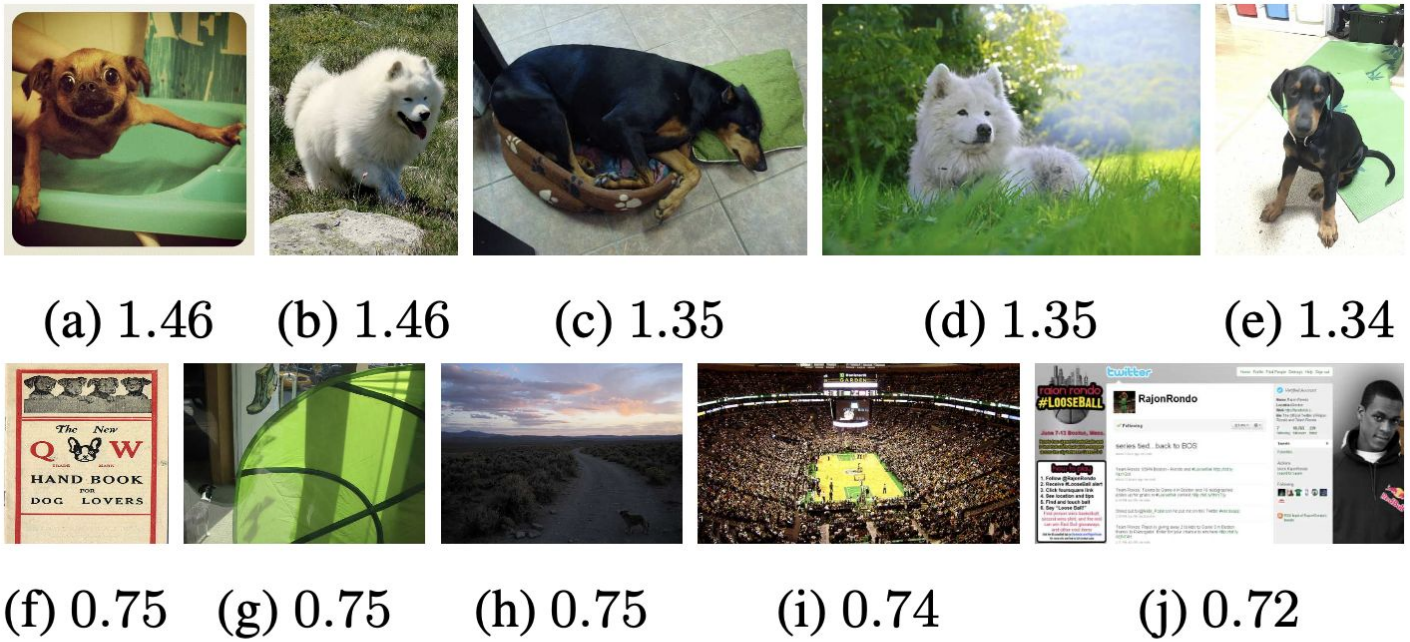


Figure 3: The web images in the top (*resp.*, bottom) row are associated with 5 highest (*resp.*, lowest) weights based on the learnt weight vector $\boldsymbol{\theta}$.

Extension to Generalized Setting

Table 2: Accuracies (%) of different methods on three datasets under the generalized setting. The best results are highlighted in boldface.

Dataset	CUB	SUN	Dogs	Avg
LR_mix	55.27	32.03	53.74	47.01
WSL+LR	57.60	35.11	55.13	49.28
Chao *et al.*	25.75	20.77	31.53	26.02
Ours	**59.60**	**36.00**	**65.89**	**53.83**

Conclusion

- A new learning scenario for fine-grained image classification by jointly utilizing web data and auxiliary labeled categories.

- Develop a novel learning model, which unifies WSL and ZSL in one formulation with an efficient and effective solution.