VALSE 2019

April 11 – 14, 2019

Learning from Web Data using Adversarial Discriminative Neural Networks for Fine-Grained Classification (AAAI'2019)

Xiaoxiao Sun, Liyi Chen, Jufeng Yang *Nankai University*

Introduction

While most existing works learn from web data by reducing the noisy level of web data, we address this problem by overcoming the dataset gap between the web and well-labelled datasets. Specifically, an adversarial discriminative loss is used to advocate representation coherence between the two kinds of data.

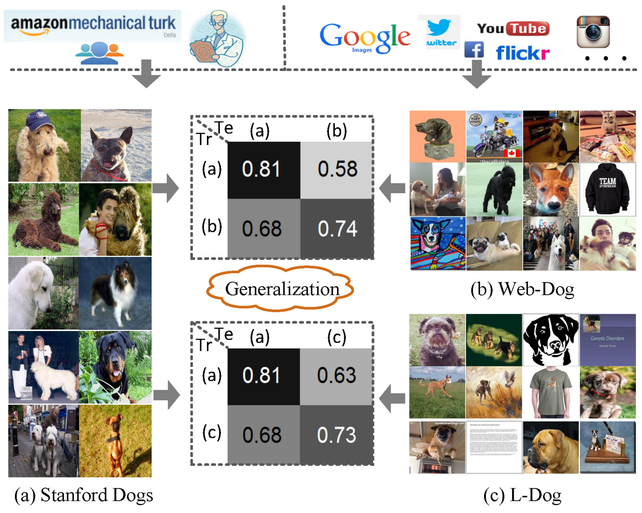

Figure1: Examples from the Stanford Dogs dataset, the web images (Web-Dog) and the L-Dog dataset. Note the Stanford Dogs dataset is well-labeled by users, while the other two datasets are labeled with keywords on the web. The cross dataset training-testing accuracies are shown in the center, where "Tr" and "Te" indicate training set and test set respectively. The gap between the results of Tr and Te on the same dataset and those of "Tr" and "Te" on different datasets shows that these datasets are not generalized well to each other.

Our contribution:

- We propose a jointly optimized deep architecture towards reducing the influence of dataset gap between easily acquired web images and the well-labeled data from standard datasets.
- We conduct extensive experiments on the Food-101, Stanford Dogs and MIT Indoor 67 datasets. The results show that the proposed method is simple yet powerful and achieves state-of-the-art classification results on the three datasets.

Method

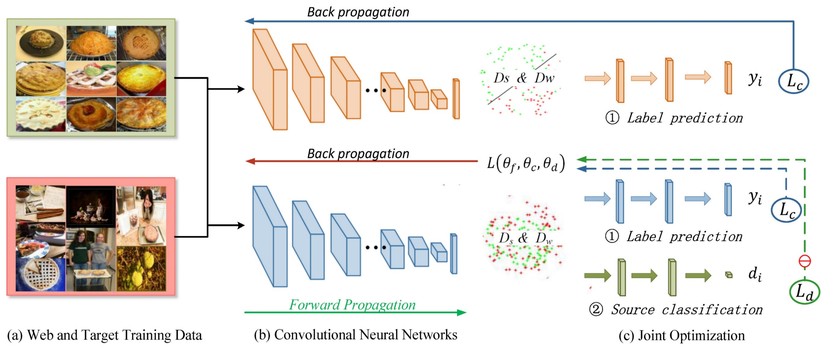

Figure 2: Overview of the proposed method. (a) web and target training data is the input to (b) convolutional neural networks. (c) joint optimization for label prediction and source classification. $\boldsymbol{L_c}$ is the loss for label prediction and $\boldsymbol{L_d}$ is the loss for source classification. $\boldsymbol{L(\theta_f, \theta_c, \theta_d)}$ is the joint loss, which influences the parameters by back propagation. $\boldsymbol{L_d}$ is preceded by a minus sign, so $\boldsymbol{\theta_f}$ aims to maximize $\boldsymbol{L_d}$, which means that the feature from the shared convolutional layers becomes more and more consistent for web and target data

➢Standard Classification

- using the standard softmax loss for object classification during both training and testing stages:

$$L_{\mathbf{c}}(x, y; \boldsymbol{\theta}_{\mathbf{f}}, \boldsymbol{\theta}_{\mathbf{c}}) = -\left[\sum_{j=1}^{C} \mathbf{1}(y=j) \log \frac{e^{\{\boldsymbol{\theta}_{\mathbf{f}}, \boldsymbol{\theta}_{\mathbf{c}}\}_j^\top x}}{\sum_{k=1}^{C} e^{\{\boldsymbol{\theta}_{\mathbf{f}}, \boldsymbol{\theta}_{\mathbf{c}}\}_k^\top x}}\right]$$

➢Source Classification

- distinguishing the data from the standard and web datasets during the training stage
- maximizing the log-likelihood loss for domain classification:

$$L_{\mathbf{d}}(d; g(x, \boldsymbol{\theta}_{\mathbf{d}}, \boldsymbol{\theta}_{\mathbf{f}})) = \sum_{i=1}^{M^b} -d_i \log(g(x, \boldsymbol{\theta}_{\mathbf{d}}, \boldsymbol{\theta}_{\mathbf{f}})) - (1-d_i) \log(1 - g(x, \boldsymbol{\theta}_{\mathbf{d}}, \boldsymbol{\theta}_{\mathbf{f}}))$$

➢Multi-task Learning

- minimizing the joint loss function:
- parameter λ controls the trade-off between the two losses

$$L(\boldsymbol{\theta}_{\mathbf{f}}, \boldsymbol{\theta}_{\mathbf{c}}, \boldsymbol{\theta}_{\mathbf{d}}) = L_{\mathbf{c}} - \lambda L_{\mathbf{d}}$$

Experiment

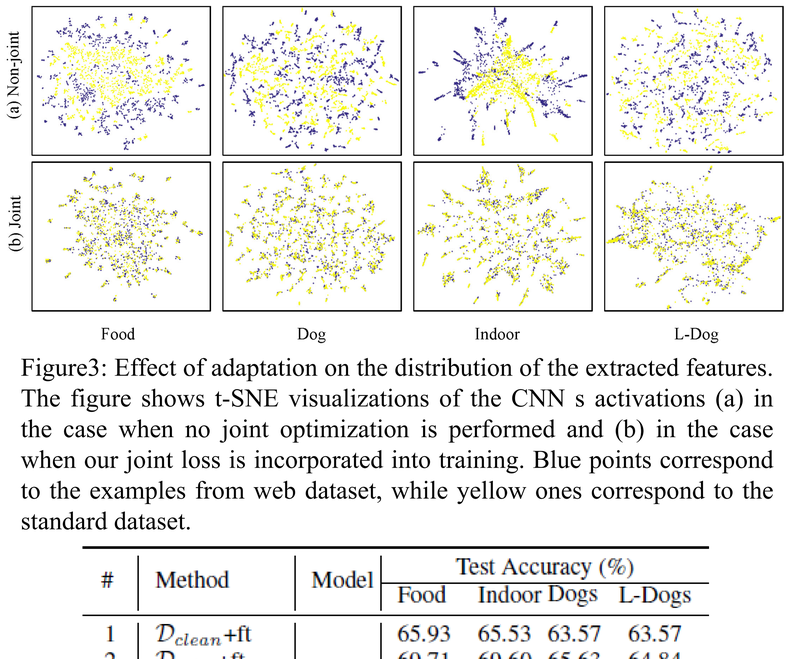

Figure3: Effect of adaptation on the distribution of the extracted features. The figure shows t-SNE visualizations of the CNN s activations (a) in the case when no joint optimization is performed and (b) in the case when our joint loss is incorporated into training. Blue points correspond to the examples from web dataset, while yellow ones correspond to the examples from the standard dataset.

#	Method	Model	Test Accuracy (%) Food	Indoor	Dogs	L-Dogs
1	$\mathcal{D}_{clean}$+ft	AlexNet	65.93	65.53	63.57	63.57
2	$\mathcal{D}_{mix}$ +ft		69.71	69.60	65.63	64.84
3	$\mathcal{D}_{filter}$+ft		69.89	66.25	67.95	66.16
4	Bottom-up		70.29	66.79	72.17	71.59
5	Pseudo-label		69.36	67.35	70.32	71.01
6	Weakly		71.10	67.82	73.88	73.64
7	Ours		**73.78**	**71.21**	**75.26**	**74.58**
8	$\mathcal{D}_{clean}$+ft	CaffeNet	66.61	65.24	63.19	63.19
9	$\mathcal{D}_{mix}$ +ft		69.25	68.00	66.08	65.34
10	$\mathcal{D}_{filter}$+ft		68.48	63.53	69.56	65.90
11	Boosting		72.53	65.56	73.49	73.28
12	PGM		73.14	65.29	72.63	71.83
13	WSL		73.21	65.58	73.52	73.79
14	Ours		**74.78**	**68.40**	**74.93**	**75.25**
15	$\mathcal{D}_{clean}$+ft	VggNet	74.32	71.81	78.29	78.68
16	$\mathcal{D}_{mix}$ +ft		76.98	72.00	81.03	77.54
17	$\mathcal{D}_{filter}$+ft		78.24	72.04	79.70	79.57
18	Harnessing		79.02	72.48	78.45	79.92
19	Ours		**82.94**	**76.12**	**84.92**	**82.55**
20	$\mathcal{D}_{clean}$+ft	ResNet50	84.31	79.63	80.51	80.51
21	$\mathcal{D}_{mix}$ +ft		85.21	82.35	81.43	82.07
22	$\mathcal{D}_{filter}$+ft		86.10	81.32	82.62	83.61
23	Goldfince		86.75	83.47	85.90	85.48
24	Ours		**89.35**	**84.59**	**87.38**	**86.64**

Table1: Classification accuracies on four datasets with different methods. D_clean and D_mix represent clean data and mixed data. D_filter denotes that the web data is filtered from D_mix. The "ft" means the fine-tuning process. L-Dogs refers to the web dataset from Goldfince is used for boosting the performance of Stanford Dogs.