LETTER

# Learning from Multiple Sources via Multiple Domain Relationship

**Zhen LIU**[†a)], **Junan YANG**[†], **Hui LIU**[†], ***Nonmembers***, ***and*** **Jian LIU**[††], ***Student Member***

**SUMMARY** Transfer learning extracts useful information from the related source domain and leverages it to promote the target learning. The effectiveness of the transfer was affected by the relationship among domains. In this paper, a novel multi-source transfer learning based on multi-similarity was proposed. The method could increase the chance of finding the sources closely related to the target to reduce the "negative transfer" and also import more knowledge from multiple sources for the target learning. The method explored the relationship between the sources and the target by multi-similarity metric. Then, the knowledge of the sources was transferred to the target based on the smoothness assumption, which enforced that the target classifier shares similar decision values with the relevant source classifiers on the unlabeled target samples. Experimental results demonstrate that the proposed method can more effectively enhance the learning performance.

***key words:*** *transfer learning, multiple source transfer, domain similarity, manifold assumption*

## 1. Introduction

Transfer learning [1], [2] can effectively exploit and transfer the knowledge from different but similar source domains for target domain learning, which has been applied in many real-world applications. For the single-source domain setting, much work has been developed [1]. In general, the effectiveness of the knowledge from a source to the target depends on how they are related. The stronger the relationship, the more usable will be the source knowledge. On the other hand, brute force transferring in case of weak relationships may lead to performance deterioration of the target learning, i.e., "negative transfer". Often in practice, one may be offered more than one source domain for learning. It is wasteful if we only use one source for learning.

We propose a novel multi-similarity based multi-source transfer learning method ($(MS)^2TL$). The method explores the relationships between the sources and the target by multi-similarity metric. Then, the knowledge of the sources is transferred to the target based on the smoothness assumption, which enforces the requirement that the target classifier shares similar decision values with the relevant source classifiers on the unlabeled target samples. We also employ a sparsity-regularizer based on the $\varepsilon$-insensitive loss to enforce the sparsity of the target classifier with the support vectors only from the target domain such that the label prediction on any test sample is very fast. Furthermore, $(MS)^2TL$ only needs the pre-learned source classifiers when training the target classifier, which is suitable for the large dataset. $(MS)^2TL$ can not only improve the ability to avoid "negative transfer" but also explore more knowledge from the sources for the target learning. Experimental results demonstrate that the proposed method can more effectively enhance the learning performance.

Manuscript received January 11, 2016.
Manuscript revised March 8, 2016.
Manuscript publicized April 11, 2016.

[†]The authors are with Electronic Engineering Institution, Huangshang Road #460, Hefei 230037, P.R. China.
[††]The author is with the College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Xinmofan Road #66, Nanjing 210003, P.R. China.
a) E-mail: ahulz@163.com

DOI: 10.1587/transinf.2016EDL8008

## 2. Proposed Algorithm

Let us represent $D^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ as the $s$th source domain, where $\mathbf{x}_i^s$ and $y_i^s$ are the feature vector and label, $N_s$ is the number of samples in $D^s$. $s = 1, \ldots, M$ and $M$ is the number of the sources. The target domain $D^T$ includes a labeled set $D_l^T = \{(\mathbf{x}_i^T, y_i^T)\}_{i=1}^{N_l}$ and an unlabeled set $D_u^T = \{\mathbf{x}_i^T\}_{i=N_l+1}^{N_T}$, where $N_l$ is the number of labeled samples and $N_T$ is the total number. $f^s$ is the pre-learned source classifier in $D^s$ and $f^T$ is the target classifier. Any types of classifier can be readily used as $f^s$'s. For the target sample $\mathbf{x}_i^T$, we denote the decision values as $f_i^T = f^T(\mathbf{x}_i^T)$ and $f_i^s = f^s(\mathbf{x}_i^T)$.

### 2.1 Multi-Source Transfer Manifold Regularizer Based on Multi-Similarity

Here we define the similarities among domains at two levels, i.e., "domain-domain" and "sample-domain", as shown in Fig. 1.

To measure the similarities of "domain-domain", we define the similarity weight $\gamma_s$ of the $s$th source $D^s$ as

$$\gamma_s = \exp(-MMD^2(D^s, D^T)/\beta_1) \tag{1}$$

where $MMD(D^s, D^T)$ is the maximum mean discrepancy (MMD) [3] for measuring the data distributions between the

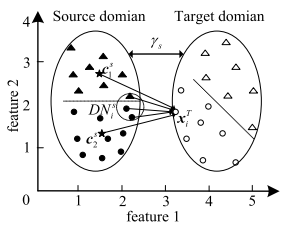



**Fig. 1** Multi-similarity

$s$th source and the target. $\beta_1 > 0$ is fixed as the mean of $MMD$ among domains.

To describe the relevance further in detail, we concern the similarities at the level of "sample-domain". First, two kinds of distance are defined, i.e., $DN_i^s$ and $DC_i^s$.

$DN_i^s$ is the average distance of the target sample $\mathbf{x}_i^T$ to its neighbors in the $s$th source domain $D^s$.

$$DN_i^s = (1/N_k)\sum_{k=1}^{N_k} d(\mathbf{x}_i^T, \mathbf{x}_k^s) \tag{2}$$

where $\mathbf{x}_k^s$ is the $k$th neighbor of $\mathbf{x}_i^T$ in $D^s$, $N_k$ is the number of neighbors, $d(\cdot)$ is a general distance metric.

$DC_i^s$ is the minimum distance of the target sample $\mathbf{x}_i^T$ to the class centers in the $s$th source domain $D^s$.

$$DC_i^s = \min_j d(\mathbf{x}_i^T, \mathbf{c}_j^s) \tag{3}$$

where $\mathbf{c}_j^s$ is the mean of the $j$th class samples in $D^s$.

If $DN_i^s$ is small, $\mathbf{x}_i^T$ is more likely to occur in $D^s$, which can be regarded as the similarity between $\mathbf{x}_i^T$ and $D^s$ from the perspective of the marginal distribution. If $DC_i^s$ is small and $\mathbf{x}_i^T$ is most close to $\mathbf{c}_j^s$, $\mathbf{x}_i^T$ probably belongs to the $j$th class in $D^s$, which can be treated as the similarity from the perspective of the conditional distribution. To take into account them comprehensively, we compute the $d_i^s = 0.5(DN_i^s + DC_i^s)$ as the final distance metric from $\mathbf{x}_i^T$ to $D^s$. Then, we have the similarity weight $A_{is}$ of $\mathbf{x}_i^T$ in the $s$th source domain at the level of "sample-domain".

$$A_{is} = \exp(-(d_i^s)^2/\beta_2) \tag{4}$$

where $\beta_2$ is fixed as the mean of $d_i^s$ in the whole target set.

Motivated from the manifold assumption [4], we similarly assume that the target classifier $f^T$ should have similar decision values on the unlabeled target samples with the prelearned classifiers $f^s$'s from the relevant source domains. Thus, the multi-source transfer manifold regularizer $\Omega_D(f^T)$ is given as follows.

$$\Omega_D(f^T) = \frac{1}{2}\sum_{s=1}^{M}\gamma_s\sum_{i=N_l+1}^{N_T} A_{is}(f_i^T - f_i^s)^2 \tag{5}$$

where $\gamma_s$ and $A_{is}$ are defined in (1) and (4) respectively. Unlike the traditional manifold regularizer which concentrates on the low-dimensional manifold embedded in the high-dimensional space [4], $\Omega_D(f^T)$ is used to connect the sources and the target through $\gamma_s$ and $A_{is}$. As in Fig. 2, if $\gamma_s$ and $A_{is}$ are large, the decision values of $f^T$ and $f^s$ on $\mathbf{x}_i^T$ will be similar. Thus, we can transfer the knowledge from the sources to the target under the assumption of "domain relevance- decision constraint".

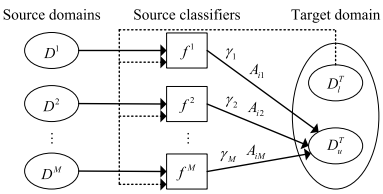



**Fig. 2** Transfer learning based on the multi-similarities.

## 2.2 Multi-Similarity Based Multi-Source Transfer Learning and Its Solution

Assume $f^T$ admits a form of $f^T(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x}) + b$. To minimize the loss of the labeled target data as well as the multi-source transfer manifold regularizer defined on the unlabeled target data simultaneously, the proposed framework $(MS)^2TL$ is then formulated as follows

$$\min_{f^T} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\lambda_L}{2}\sum_{i=1}^{N_l}(f_i^T - y_i^T)^2 + \lambda_D\Omega_D(f^T) \tag{6}$$

where the first term controls the complexity of $f^T$, the second is a loss function of $f^T$ on the labeled target samples, the third is the multi-source transfer manifold regularizer on the unlabeled target samples, and $\lambda_L$, $\lambda_D > 0$ are the regularization parameters.

To solve (6) efficiently, the $\varepsilon$-insensitive loss function in SVR [5] is introduced into (6). Then, we have

$$\begin{aligned}\min_{f_i^T,\mathbf{w},b} &\frac{1}{2}\|\mathbf{w}\|^2 + \frac{\lambda_L}{2}\sum_{i=1}^{N_l}(f_i^T - y_i^T)^2 \\ &+\lambda_D\Omega_D(f^T) + C\sum_{i=1}^{N_T}\ell_\varepsilon(\mathbf{w}'\phi(\mathbf{x}_i) + b - f_i^T)\end{aligned} \tag{7}$$

where the $\varepsilon$-insensitive loss function $\ell_\varepsilon(t) = |t| - \varepsilon$ if $|t| > \varepsilon$, otherwise 0. $C$ is the regularization parameter. Since $\ell_\varepsilon(\cdot)$ is non-smooth, (7) is usually transformed as a constrained optimization problem

$$\begin{aligned}\min_{f_i^T,\mathbf{w},,b,\xi_i,\xi_i^*} &\frac{1}{2}\|\mathbf{w}\|^2 + \frac{\lambda_L}{2}\sum_{i=1}^{N_l}(f_i^T - y_i^T)^2 \\ &+\lambda_D\Omega_D(f^T) + C\sum_{i=1}^{N_T}(\xi_i + \xi_i^*)\end{aligned} \tag{8}$$

$$s.t.\ \mathbf{w}'\phi(\mathbf{x}_i^T) + b - f_i^T \le \varepsilon + \xi_i, \xi_i \ge 0 \tag{9}$$

$$f_i^T - \mathbf{w}'\phi(\mathbf{x}_i^T) - b \le \varepsilon + \xi_i^*,\ \xi_i^* \ge 0 \tag{10}$$

$$i = 1, \ldots, N_T$$

By introducing the Lagrange multipliers $\alpha_i$'s and $\eta_i$'s (resp., $\alpha_i^*$'s and $\eta_i^*$'s) for the constraints in (9) (resp., (10)), we arrive at the following dual formulation

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} &\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)'\tilde{\mathbf{K}}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + \tilde{\mathbf{y}}'(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + \varepsilon\mathbf{1}'_{N_T}(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) \\ s.t.\quad &\mathbf{1}'_{N_T}\boldsymbol{\alpha} = \mathbf{1}'_{N_T}\boldsymbol{\alpha}^*, \mathbf{0}_{N_T} \le \boldsymbol{\alpha}, \boldsymbol{\alpha}^* \le C\mathbf{1}_{N_T}\end{aligned} \tag{11}$$

where $\mathbf{0}_{N_T}$, $\mathbf{1}_{N_T}$ is the column vectors of all zeros and all ones, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{N_T}]'$, $\boldsymbol{\alpha}^* = [\alpha_1^*, \ldots, \alpha_{N_T}^*]'$, $\tilde{\mathbf{K}} = \mathbf{K} + diag(\mathbf{q})$, $\mathbf{K} = \boldsymbol{\Phi}'\boldsymbol{\Phi}$, $\boldsymbol{\Phi} = [\phi(\mathbf{x}_1^T), \ldots, \phi(\mathbf{x}_{N_T}^T)]$, $\mathbf{q} = [q_1, \ldots, q_{N_T}]'$. If $i = 1, \ldots, N_l$, $\tilde{y}_i = y_i^T$ and $q_i = 1/\lambda_L$, otherwise $\tilde{y}_i = 1/(\sum_{s=1}^M \gamma_s A_{is})\sum_{s=1}^M \gamma_s A_{is} f_i^s$ and $q_i = 1/(\lambda_D\sum_{s=1}^M \gamma_s A_{is})$. Since the dual form of (11) is similar to that of $\varepsilon$-SVR [6], the objective function in (7) can be solved efficiently by using state-of-the-art SVM solvers such as LIBSVM [7]. For any test sample $\mathbf{x}$, the decision value of the target classifier $f^T$ is

$$f^T(\mathbf{x}) = \mathbf{w}'\phi(\mathbf{x}) + b = \sum_{i:a_i - a_i^* \ne 0}(a_i^* - a_i)k(\mathbf{x}_i^T, \mathbf{x}) + b \tag{12}$$

which is only a linear combination of $k(\mathbf{x}_i^T, \mathbf{x})$'s without involving any source classifiers. According to the Karush Kuhn Tucker conditions, if a target sample $\mathbf{x}_i^T$ has the value $|\mathbf{w}'\phi(\mathbf{x}_i^T)+b-f_i^T| < \varepsilon$, then its coefficient $(\alpha_i^*-\alpha_i)$ in (12) becomes zero. Therefore, with the $\varepsilon$-insensitive loss function, the computation for the prediction using the sparse representation in (12) can be greatly reduced.

## 3. Experimental Results

The experiments use three datasets as the sources [8]: Amazon (images from online merchants), Webcam (low-resolution images by a web camera), and DSLR (high-resolution images by a digital SLR camera). Caltech-256 [8] is used as the target domain. There are totally 10 classes of target images which are common to all four datasets with 8 to 151 samples per category per domain, and 2533 images in total. Figure 3 highlights the differences among these domains with example images from the category of Computer-monitor.

We extract the 4096 dimensional DeCAF$_6$ features [9] from the raw images. Then, these features from different domains are used to learn a classification model for the target. In the default setting, we set $\lambda_L = \lambda_D = 1$, $C = 1$, and $N_k = 8$. Gaussian kernel (i.e., $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-d^2(\mathbf{x}_i, \mathbf{x}_j)/(2\sigma^2))$ is used as the default kernel where the kernel parameter $\sigma$ is set as the mean distance between samples in the target domain. In the target domain, $n$ samples per class are randomly selected as the labeled target set which is set as 0, 2, 4, 6, 10, 15, and 20. The experiments are repeated for 20 times with different samples. The average classification accuracy is used as the evaluation measure.

### 3.1 Performance of Our Proposed Method

If we only concern the similarity at the level of "domain-domain", namely, set all $A_{is}$'s equal to 1, (MS)$^2$TL would be similar to the DAM algorithm [6]. Thus, we compare our method (MS)$^2$TL with DAM in the experiments. LS-SVM classifier is pre-learned in every source domains and used as the source classifiers $f^s$'s. The classification accuracies of (MS)$^2$TL compared with the Base and DAM are recorded in Table 1. The Base means that the source classifiers are used to predict the unlabeled target samples directly and the average accuracy is the final result. The highest accuracy among different methods is highlighted in bold.

To show that (MS)$^2$TL could use any type types of classifier as $f^s$'s, we also conduct the experiments by using Naïve Bayes as the source classifiers, as in Table 2.

Both in Table 1 and Table 2, (MS)$^2$TL can effectively improve the accuracy compared with other methods. The results demonstrate that (MS)$^2$TL could better explore the relevant relationship between the sources and the target, and transfer more knowledge from the sources to promote the target learning. As the Base method uses the source classifiers directly without considering the difference between domains, its results are always not good. The accuracies of (MS)$^2$TL and DAM generally increase along with the increasing of $n$ (the number of the labeled target samples per class). The performance of (MS)$^2$TL is better when Naïve Bayes is used as $f^s$'s rather than LS-SVM. However, the difference in performance between the two cases becomes smaller as $n$ increases. It can be concluded that (MS)$^2$TL depends more on the labeled target samples if $f^s$'s are LS-SVM.

(a) Amazon b) Webcam (c) DSLR (d) Caltech-256

**Fig. 3** Images of Computer-monitor in different domains.

### 3.2 Parameter Analysis

In this section, we evaluate the performance variations with respect to the regularization parameters $C$, $\lambda_L$, $\lambda_D$, and the number of neighbors $N_k$. When evaluating the performance variations with respect to one parameter, we fix the other parameters as their default values. We choose LS-SVM as the source classifiers since it also has the parameter $C$.

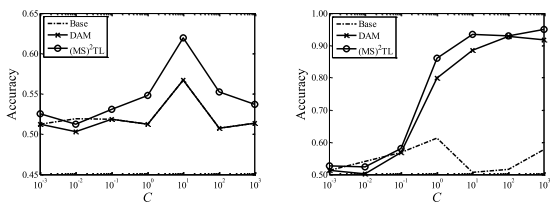
First, we consider the performance variations w.r.t. $C$, as in Fig. 4. In the experiments, $C$ is set as $10^{-3}$, $10^{-2}$, $10^{-1}$, 1, 10, $10^2$, and $10^3$. We observe that (MS)$^2$TL is better than other methods by using different $C$'s in most cases. If there is no labeled samples in the target domain (i.e., $n = 0$), DAM has no improvements compared with Base while (MS)$^2$TL

**Table 1** Classification accuracies when $f^s$'s are LS-SVM.

| $n$ | Base | DAM | (MS)$^2$TL |
|---|---|---|---|
| *0* | 0.5125±0.0000 | 0.5125±0.0000 | **0.5482±0.0000** |
| *2* | 0.4902±0.0173 | 0.4902±0.0093 | **0.5355±0.0061** |
| *4* | 0.5465±0.0214 | 0.5519±0.0114 | **0.5677±0.0069** |
| *6* | 0.5940±0.0197 | 0.6974±0.0126 | **0.7085±0.0083** |
| *10* | 0.6124±0.0205 | 0.7992±0.0147 | **0.8602±0.0107** |
| *15* | 0.5865±0.0227 | 0.9121±0.0156 | **0.9504±0.0111** |
| *20* | 0.5753±0.0263 | 0.9477±0.0189 | **0.9657±0.0128** |

**Table 2** Classification accuracies when $f^s$'s are Naïve Bayes.

| $n$ | Base | DAM | (MS)$^2$TL |
|---|---|---|---|
| *0* | 0.6836±0.0000 | 0.7254±0.0000 | **0.8000±0.0000** |
| *2* | 0.7023±0.0201 | 0.7622±0.0119 | **0.7708±0.0089** |
| *4* | 0.6764±0.0204 | 0.7239±0.0123 | **0.7764±0.0103** |
| *6* | 0.6534±0.0217 | 0.7505±0.0147 | **0.8528±0.0104** |
| *10* | 0.6275±0.0223 | 0.8577±0.0152 | **0.9483±0.0112** |
| *15* | 0.6488±0.0254 | 0.9515±0.0149 | **0.9555±0.0121** |
| *20* | 0.7015±0.0312 | 0.9708±0.0192 | **0.9798±0.0109** |




(a) $n$=0 (b) $n$=10

**Fig. 4** Classification accuracies of all methods with different $C$.

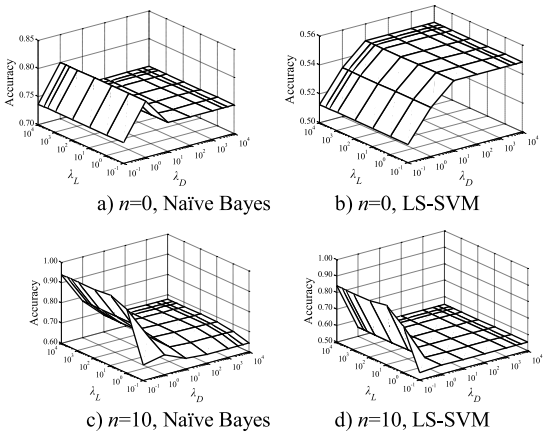



**Fig. 5** Classification accuracy with different $\lambda_L$ and $\lambda_D$.

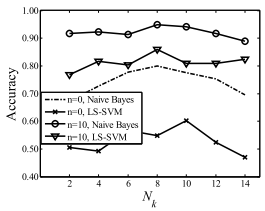



**Fig. 6** Classification accuracy with different $N_k$.

still achieves the highest accuracy in most cases. In the case of labeled target samples exist (i.e., $n = 10$), the performances of DAM and (MS)$^2$TL tend to saturate when $C$ becomes large while the classification results of Base are always not good.

The performance variations w.r.t. different $\lambda_L$ and $\lambda_D$ are shown in Fig. 5. Specifically, we set $\lambda_L$ and $\lambda_D$ as 0.1, 1, 10, $10^2$, $10^3$, $3*10^3$, $5*10^3$, and $10^4$ respectively. We observe that the performance of (MS)$^2$TL changes more dramatically along with the variation of $\lambda_D$ compared with $\lambda_L$. It demonstrates that the regularizer $\Omega_D(f^T)$ has a big influence on the performance of (MS)$^2$TL. Compared with the two settings (i.e., $n = 0$ or 10), we also observe that (MS)$^2$TL achieves the highest accuracy at a larger value of $\lambda_D$ when there is no labeled target samples. It can be explained that (MS)$^2$TL depends more on $\Omega_D(f^T)$ when no labeled target samples exist.

We show the performances of (MS)$^2$TL by using different $N_k$ in Fig. 6, where $N_k$ is set as 2, 4, 6, 8, 10, 12, and 14. In both two settings (i.e., $n = 0$ and 10), the performance of (MS)$^2$TL depends on the setting of $N_k$. Especially, this dependence is evident when no labeled target samples exist. This may because that (MS)$^2$TL will depend more on the sources if there is no labeled target sample, then $N_k$ will have a bigger influence since $N_k$ is a key parameter for the knowledge transfer. In most cases, the learning performance will be badly hurt if $N_k$ is too large or too small. The reason can be concluded as: if $N_k$ is set too small, the local scope can not cover all the affinitive examples; on the contrary, if $N_k$ is fixed beyond normal scope, the similarity measure may suffer interfere from the false distribution of the irrelevant data. Thus, fixing the value of $N_k$ at [6, 10] is recommended.

## 4. Conclusions

In this paper, a novel multi-source transfer learning method called (MS)$^2$TL is proposed. (MS)$^2$TL can import more knowledge from multiple sources for the target learning and also increase the chance of finding the sources closely related to the target to reduce the "negative transfer". The method explores the relevance between domains by a multi-similarity metric. Then, the knowledge of the sources is transferred to the target based on the smoothness assumption. We also employ the $\varepsilon$-insensitive loss regularizer such that the label prediction on any test sample is very fast. What's more, (MS)$^2$TL only needs the pre-learned source classifiers when training the target classifier, which is suitable for a large dataset. Comprehensive experiments clearly demonstrate the effectiveness of our method.

## Acknowledgments

Both authors would like to acknowledge the support of National High-tech R&D Program (863 Program), and Anhui Provincial Natural Science Foundation (NO.1308085QF99, NO.1408085MKL46).

### References

[1] S.J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowledge and Data Engineering, vol.22, no.10, pp.1345–1359, Oct. 2010.
[2] S. Sun, H. Shi, and Y. Wu, "A survey of multi-source domain adaptation," Information Fusion, vol.24, pp.84–92, July. 2015.
[3] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.-P. Kriegel, B. Scholkopf, and A.J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," Bioinformatics, vol.22, no.14, pp.49–57, July 2006.
[4] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," J. Mach. Learn. Res., vol.7, pp.2399–2434, Dec. 2006.
[5] I.W. Tsang and J.T. Kwok, "Large-scale sparsified manifold regularization," in Advances in Neural Information Processing Systems 19, Cambridge, pp.1401–1408, 2007.
[6] L. Duan, D. Xu, and I.W.-H. Tsang, "Domain adaptation from multiple sources: a domain-dependent regularization approach," IEEE Trans. Neural Networks Learn. Syst., vol.23, no.3, pp.504–518, March 2012.
[7] C.C. Chang and C.J. Lin, LIBSVM: A Library for Support Vector Machines (2001). [Online]. Available: http://www.csie.ntu.edu.tw/˜cjlin/libsvm
[8] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow klernel for unsupervised domain adaptation," CVPR'12 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Washington, pp.2066–2073, 2012.
[9] J. Donahue, Y. Jia, O. Vinyals, et al., "DeCAF: A deep convolutional activation feature for generic visual recognition," Proc. International Conf. Machine Learning, pp.647–655, Beijing, 2014.