Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation

Boqing Gong BOQINGGO@USC.EDU
Department of Computer Science, University of Southern California, Los Angeles, CA 90089

Kristen Grauman GRAUMAN@CS.UTEXAS.EDU
Department of Computer Science, University of Texas at Austin, Austin, TX 78701

Fei Sha FEISHA@USC.EDU
Department of Computer Science, University of Southern California, Los Angeles, CA 90089

Abstract

Learning domain-invariant features is of vital importance to unsupervised domain adaptation, where classifiers trained on the source domain need to be adapted to a different target domain for which no labeled examples are available. In this paper, we propose a novel approach for learning such features. The central idea is to exploit the existence of *landmarks*, which are a *subset* of labeled data instances in the source domain that are distributed most similarly to the target domain. Our approach automatically discovers the landmarks and use them to bridge the source to the target by constructing provably easier auxiliary domain adaptation tasks. The solutions of those auxiliary tasks form the basis to compose invariant features for the original task. We show how this composition can be optimized discriminatively *without* requiring labels from the target domain. We validate the method on standard benchmark datasets for visual object recognition and sentiment analysis of text. Empirical results show the proposed method outperforms the state-of-the-art significantly.

1. Introduction

Learning algorithms often rely heavily on the assumption that data used for training and testing are drawn from the same distribution. However, the validity of this assumption is frequently challenged in real-world applications. For example, in computer vision, recent studies have shown that object classifiers optimized on one benchmark dataset often exhibit significant degradation in recognition accuracy when evaluated on another one (Torralba & Efros, 2011; Perronnin et al., 2010). The culprit is clear: the visual appearance of even the same object varies significantly across different datasets as a result of many factors, including imaging devices, photographers' preferences, or illumination. These idiosyncrasies often cause a substantial mismatch between the training and the testing distributions. Similarly, in text analysis, we might want to train a document classifier on one corpus (e.g., product reviews on kitchen appliances) and apply to another one such as reviews on books (Blitzer et al., 2007). The two corpora thus have mismatched distributions of words and their usages, such that the trained classifier would not perform well.

How can we build classifiers that are robust to mismatched distributions? This is the domain adaptation problem, where the training data comes from a *source* domain and the testing data comes from a different *target* domain (Shimodaira, 2000; Daumé & Marcu, 2006; Pan & Yang, 2010; Gretton et al., 2009). When some labeled data from the target is accessible, the problem is similar to semi-supervised learning and is referred to as *semi-supervised domain adaptation* (Daumé et al., 2010; Bergamo & Torresani, 2010; Saenko et al., 2010). In contrast, when there is no labeled data from the target domain to help learning, the problem is called *unsupervised domain adaptation* (Blitzer et al., 2007; 2006; Gopalan et al., 2011; Gong et al., 2012; Chen et al., 2011).

Unsupervised domain adaptation is of great impor-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

tance to real-world applications. For instance, suppose we want to allow mobile phone users to take pictures and recognize objects in environments specific to their lifestyles. While both the camera phones and the users' environments introduce idiosyncrasies in the images, it would be highly desirable if users did not have to label any captured image data; the recognition system ought to adapt automatically from existing labeled vision datasets, such as LabelMe or ImageNet (Russell et al., 2008; Deng et al., 2009).

While appealing, unsupervised domain adaptation is especially challenging. For example, the common practice of discriminative training is generally not applicable. Without labels, it is not even clear how to define the right discriminative loss on the target domain! Similarly, it is also difficult to perform model selection (e.g., tuning regularization coefficients).

Thus, to enable domain adaptation, we need to determine how domains are related (Pan & Yang, 2010; Quionero-Candela et al., 2009). One extensively studied paradigm is to assume that there is a domain-invariant feature space (Blitzer et al., 2007; 2006; Gopalan et al., 2011; Blitzer et al., 2011; Chen et al., 2011; Ben-David et al., 2007; 2010; Pan et al., 2009). In this space, the source and target domains have the same (or similar) marginal distributions, and the posterior distributions of the labels are the same across domains too. Hence, a classifier trained on the labeled source would likely perform well on the target. Several ways of measuring distribution similarities have been explored and theoretical analysis shows that the performance of the classifier on the target is indeed positively correlated with those similarities (Ben-David et al., 2010; Mansour et al., 2009a;b).

Despite such progress, existing approaches so far have only been limited to macroscopically examining the distribution similarity by tuning to statistical properties of the samples as a whole — when comparing distributions, all the samples are used. This notion is stringent, as it requires all discrepancies to be accounted for and forces learning inefficiently (or even erroneously) from "hard" cases that might be just outliers to the target domains.

In contrast, we will leverage the key insight that ***not all instances are created equally in terms of adaptability***. Thus, we will examine distribution similarity microscopically at the instance level; our approach plucks out and exploits the most desirable instances to facilitate adaptation. Identifying those instances requires comparing all possible subsets from the source domain to the target. We show how this can be addressed with tractable optimization. In what follows, we summarize the main idea behind our approach. After describing it in detail in section 2, we contrast it to related work in section 4.

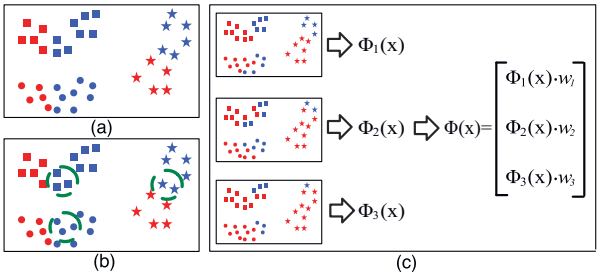

Figure 1. Sketch of the main idea of our approach (best viewed in color). (a) The original domain adaptation (DA) problem where instances in red are from the target and in blue from the source. (b) **Landmarks**, shown inside the green circles, are data instances from the source that can be regarded as samples from the target (section 2.1). (c) Multiple auxiliary tasks are created by augmenting the original target with landmarks, which switches their color (domain association) from blue to red (section 2.2). Each task gives rise to a new feature representation. These representations are combined discriminatively to form domain-invariant features for the original DA problem (section 2.3).

Main idea Our approach centers around the notion of **landmarks**. Landmarks are defined as a subset of labeled instances from the source domain. These instances are distributed similarly to the target domain. Thus, they are expected to function as a conduit *connecting* the source and target domains to facilitate adaptation. As an intuitive example, suppose we want to recognize objects placed in two types of environments: homes (as the source) and offices (as the target). Conceivably, only certain images from the source — such as those taken in home offices — could also be regarded as samples from the target domain. Such landmark images thus might have properties that are shared by both domains. These properties in turn can guide learning algorithms to search for invariant features. Fig. 2 displays several discovered landmark images for the vision datasets we use in this work.

Leveraging the existence of landmarks and their properties, we create a cohort of auxiliary tasks where landmarks explicitly bridge the source and target domains. Specifically, in those auxiliary tasks, the original target domain is augmented with landmarks, blurring the distinction across domains. Thus, those tasks are *easier* to solve than the original problem. We show this is indeed true both theoretically and empirically.

The auxiliary tasks offer multiple views of the original problem. In particular, each task differs by how its landmarks are selected, which in turn is determined by how the similarity among instances is measured. In

this work, we measure similarities at multiple scales (of distances). Thus, each view provides a different perspective on the adaptation problem by being robust to idiosyncrasies in the domains at different granularities.

The solutions of the auxiliary tasks give rise to multiple domain-invariant feature spaces that can be characterized by linear positive semidefinite kernel functions. We parameterize invariant features for the original adaptation problem with those auxiliary kernels. We show how the corresponding learning problem is equivalent to multiple kernel learning. We learn the kernel discriminatively to minimize classification errors on the landmark data instances, which serve as a proxy to discriminative loss on the target domain. Fig. 1 schematically illustrates the overall approach.

Contributions We contribute to domain adaptation by proposing a novel landmark-based approach. The key insight is to use landmarks to create auxiliary tasks that inform the original problem. We show i) how to automatically identify landmarks; ii) how to construct easier auxiliary domain adaptation tasks; iii) how to combine the solutions of auxiliary tasks discriminatively to solve the original domain adaptation problem (cf. section 2.3); iv) strong empirical results on standard benchmark datasets for object recognition and sentiment analysis, outperforming state-of-the-art algorithms by a significant margin (cf. section 3).

2. Proposed Approach

The key insight of our approach is that not all instances are equally amenable to adaptation. In particular, only certain instances bridge the source and target domains, owing to their statistical properties. We aim to identify and exploit them for adaptation.

To this end, we propose a landmark-based approach that consists of three steps that will be described in turn: i) identifying and selecting the landmark instances; ii) constructing multiple auxiliary tasks using landmarks and inferring the corresponding domain-invariant feature spaces, one for each auxiliary task; iii) discriminatively learning the final domain-invariant feature space that is optimized for the target domain.

2.1. Landmarks

Landmarks are data points from the source domain; however, given how they are distributed, they look like they could be samples from the target domain too (cf. Fig. 1 for a schematic illustration, and Fig. 2 in section 3 for exemplar images of visual objects identified as landmarks in vision datasets). The intuition behind our approach is to use these landmarks to bridge the source and the target domains.

How can we identify those landmarks? At the first glance, it seems that we need to compare all possible subsets of training instances in the source domain to the target. We will show in the following this seemingly intractable problem can be relaxed and solved with tractable convex optimization.

Let $\mathcal{D}_S = \{(\boldsymbol{x}_m, y_m)\}_{m=1}^{\mathsf{M}}$ denote M data points and their labels from the source domain. Likewise, we use $\mathcal{D}_T = \{\boldsymbol{x}_n\}_{n=1}^{\mathsf{N}}$ for the target domain.

Landmark selection To identify landmarks, we use M indicator variables $\boldsymbol{\alpha} = \{\alpha_m \in \{0,1\}\}$, one for each data point in the source domain. If $\alpha_m = 1$, then $\boldsymbol{x}_m$ is regarded as a landmark. Our goal is to choose among all possible configurations of $\boldsymbol{\alpha} = \{\alpha_m\}$ such that the distribution of the ***selected*** data instances are maximally similar to that of the target domain.

To determine whether the two distributions are similar, we use a non-parametric two-sample test (other approaches are also possible, including building density estimators when the dimensionality is not high). Specifically, we use a nonlinear feature mapping function $\phi(\cdot)$ to map $\boldsymbol{x}$ to a Reproducing Kernel Hilbert Space and compare the difference in sample means (Gretton et al., 2006). We choose $\boldsymbol{\alpha}$ such that the difference is minimized, namely,

$$\min_{\boldsymbol{\alpha}} \left\| \frac{1}{\sum_m \alpha_m} \sum_m \alpha_m \phi(\boldsymbol{x}_m) - \frac{1}{\mathsf{N}} \sum_n \phi(\boldsymbol{x}_n) \right\|_{\mathcal{H}}^2 . \quad (1)$$

Furthermore, we impose the constraint that labels be *balanced* in the selected landmarks. Concretely,

$$\frac{1}{\sum_m \alpha_m} \sum_m \alpha_m y_{mc} = \frac{1}{\mathsf{M}} \sum_m y_{mc}, \quad (2)$$

where y_{mc} is an indicator variable for $y_m = c$. The right-hand-side of the constraint is simply the prior probability of the class c, estimated from the source.

We stress that the above criterion is defined on landmarks, which are a *subset* of the source domain, as the sample mean is computed *only* on the selected instances (cf. the denominator $\sum_m \alpha_m$ in eq. (1)). This is very different from other approaches that have used similar nonparametric techniques for comparing distributions (Pan et al., 2009; Gretton et al., 2009). There they make stronger assumptions that all data points in the source domain need to be collectively distributed similarly to the target domain. Furthermore, they do not impose the balance constraint of eq. (2). Our results will show that these differences are crucial to the success of our approach.

Eq. (1) is intractable due to the binary constraints on α_m. We relax and solve it efficiently with convex optimization. We define new variables β_m as $\alpha_m \left(\sum_m \alpha_m\right)^{-1}$. We relax them to live on the simplex $\Delta = \{\boldsymbol{\beta} : 0 \leq \beta_m \leq 1, \sum_m \beta_m = 1\}$. Substituting $\{\beta_m\}$ into eq. (1) and its constraints, we arrive at the following quadratic programming problem:

$$\begin{array}{ll} \min_{\boldsymbol{\beta}\in\Delta} & \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\beta} - 2/\mathsf{N}\,\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{B}\mathbf{1} \\ \text{s.t.} & \sum_m \beta_m y_{mc} = 1/\mathsf{M}\sum_m y_{mc}, \quad \forall\ c, \end{array} \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{\mathsf{M}\times\mathsf{M}}$ denotes the kernel matrix computed over the source domain, and $\boldsymbol{B} \in \mathbb{R}^{\mathsf{M}\times\mathsf{N}}$ denotes the kernel matrix computed between the source domain points and target domain points. The optimization is convex, as the kernel matrix $\boldsymbol{A}$ is positive semidefinite.

We recover the binary solution for α_m by finding the support of β_m, ie, $\alpha_m = \textsc{threshold}(\beta_m)$. In practice, we often obtain *sparse* solutions, supporting our modeling intuition that only a subset of instances in the source domain is needed to match the target domain.

Multiscale analysis The selection of landmarks depends on the kernel mapping $\phi(\boldsymbol{x})$ and its parameter(s). For theoretical reasons, we use Gaussian RBF kernels, defined as follows:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\{-(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)/\sigma^2\}, \tag{4}$$

where the metric $\boldsymbol{M}$ is positive semidefinite. We experimented with several choices — details in section 3.

The bandwidth σ is a scaling factor for measuring distances and similarities between data points. Since we regard landmarks as likely samples from the target domain, σ determines how much the source and the target are similar to each other at different granularities. A small σ will attenuate distances rapidly and regard even close points as being dissimilar. Thus, it is likely to select a *large* number of points as landmarks in order to match distributions. A large σ will have the opposite effect. Fig. 2 illustrates the effect of σ.

Instead of choosing one σ in the hope that one scale fits all, we devise a multiscale approach. We use a set $\{\sigma_q \in [\sigma_{min},\ \sigma_{max}]\}_{q=1}^{\mathsf{Q}}$. For each σ_q, we compute the kernel according to eq. (4) and solve eq. (3) to obtain the corresponding landmarks $\mathcal{L}^q = \{(\boldsymbol{x}_m, y_m) : \alpha_m = 1\}$. Using multiple scales adds the flexibility of modeling data where similarities cannot be measured in one homogeneous scale. For example, the category of GRIZZLY BEAR is conceivably much closer to GREY BEAR than to POLAR BEAR, and so similarities among all three are better modeled at two scales.

Each set of landmarks (one set per scale) gives rise to a different perspective on the adaptation problem by suggesting which instances to explore to connect the source and the target. We achieve this connection by creating auxiliary tasks, as we describe next.

2.2. Auxiliary tasks

Constructing auxiliary tasks Imagine we create a new source domain $\mathcal{D}_S^q = \mathcal{D}_S \setminus \mathcal{L}^q$ and a new target domain $\mathcal{D}_T^q = \mathcal{D}_T \bigcup \mathcal{L}^q$, where the $\mathcal{L}^q$ is removed from and added to the source and target domains, respectively. We do not use $\mathcal{L}^q$'s labels at this stage yet.

Our auxiliary tasks are defined as Q domain adaptation problems, $\mathcal{D}_S^q \rightarrow \mathcal{D}_T^q$. The auxiliary tasks differ from the original problem $\mathcal{D}_S \rightarrow \mathcal{D}_T$ in an ***important*** aspect: the new tasks should be "easier", as the existence of landmark points ought to aid the adaptation. This is illustrated by the following theorem, stating that the discrepancy between the new domains is smaller than the original.

Theorem 1 *Let $P_S(X)$ and $P_T(X)$ denote the distributions of the original source and the target domains, respectively. For the auxiliary task, assume the new target distribution is modeled as a mixture distribution $Q_T(X) = (1-\mu)P_T(X) + \mu P_S(X)$ where $\mu \in [0,\ 1)$. In other words, the landmarks increase the component of $P_S(X)$ in the target domain. Thus,*

$$\begin{array}{rcl} KL(P_S(X)\|Q_T(X)) & \leq & (1-\mu)KL(P_S(X)\|P_T(X)) \\ & \leq & KL(P_S(X)\|P_T(X)), \end{array} \tag{5}$$

where $KL(\cdot\|\cdot)$ stands for the Kullback-Leibler divergence. In words, the new target distribution is closer to the source distribution.

The proof appeals to KL-divergence's convexity in its arguments. Details are in the Supplementary Material.

With the reduced discrepancy between $P_S(X)$ and $Q_T(X)$, we can apply the analysis in (Mansour et al., 2009b) (Lemma 1) to show that classifiers applied to $Q_T(X)$ attain a smaller generalization error bound than those applied to $P_T(X)$. Intuitively, the increased similarity between the new domains is also closely related to the increased difficulty of distinguishing which domain a data point is sampled from. More formally, if we were to build a binary classifier to classify a data point into one of the two categories SOURCE versus TARGET, we would expect the accuracy to drop when we compare the original to the auxiliary tasks. The accuracy — also named as *A-distance* — is closely related to how effective domain adaptions can be (Blitzer et al., 2007). A high accuracy is indicative of a highly contrasting pair of domains, and thus is possibly due to many domain-specific features capturing each domain's individual characteristics.

These insights motivate our design of auxiliary tasks: they conceivably have low accuracy for binary classification as the landmarks blend the two domains, discouraging the use of domain-specific features. We describe next how to extract domain-invariant ones using the solutions of those easy problems as a *basis*.

Learning basis from auxiliary tasks For every pair of auxiliary domains, we use the geodesic flow kernel (GFK), a state-of-the-art algorithm for unsupervised domain adaptation (Gong et al., 2012), to compute domain-invariant features. The GFK is particularly adept at measuring domain-invariant distances among data points, as exemplified by its superior performance in nearest-neighbor classifiers. Thus, it is especially suitable for the final stage of our approach when we use Gaussain RBF kernels to compose complex domain-invariant features (cf. 2.3).

We give a brief description of that method in the following. (We omit the index q for brevity in notation.)

The GFK technique models the domain shift by modeling each domain with a d-dimensional linear subspace and embedding them onto a Grassmann manifold. Specifically, let $\boldsymbol{P}_S, \boldsymbol{P}_T \in \mathbb{R}^{\mathsf{D}\times\mathsf{d}}$ denote the basis of the PCA subspaces for each of the two domains, respectively. The Grassmann manifold $\mathbb{G}(\mathsf{d},\mathsf{D})$ is the collection of all d-dimensional subspaces of the feature vector space $\mathbb{R}^{\mathsf{D}}$. We infer the optimal d with the automatic procedure in (Gong et al., 2012).

The geodesic flow $\{\boldsymbol{\Phi}(t) : t \in [0,\ 1]\}$ between $\boldsymbol{P}_S$ and $\boldsymbol{P}_T$ on the manifold parameterizes a path connecting the two subspaces. Every point on the flow is a basis of a d-dimensional subspace. In the beginning of the path, the subspace is similar to $\boldsymbol{P}_S = \boldsymbol{\Phi}(0)$ and in the end of the flow, the subspace is similar to $\boldsymbol{P}_T = \boldsymbol{\Phi}(1)$. We project the original feature $\boldsymbol{x}$ into these subspaces and view the flow as a collection of infinitely many features varying gradually from the source to the target domain: $\boldsymbol{z}^\infty = \{\boldsymbol{\Phi}(t)^{\mathrm{T}}\boldsymbol{x} : t \in [0,\ 1]\}$.

Using the new feature representation for learning will force the classifiers to be less sensitive to domain differences and to use domain-invariant features. Particularly, the inner products of the new features give rise to a positive semidefinite kernel:

$$\begin{aligned} G(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \langle \boldsymbol{z}_i^\infty, \boldsymbol{z}_j^\infty \rangle \\ &= \boldsymbol{x}_i^{\mathrm{T}} \int_0^1 \boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^{\mathrm{T}}\, dt\ \boldsymbol{x}_j = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x}_j. \end{aligned} \tag{6}$$

The matrix $\boldsymbol{G}$ can be computed efficiently using singular value decomposition (Gong et al., 2012). Note that computing $\boldsymbol{G}$ does not require any labeled data.

The domain-invariant feature space is extracted as the mapping $\Phi_q(\boldsymbol{x}) = \sqrt{\boldsymbol{G}_q}\boldsymbol{x}$. In the following, we describe how to integrate the spaces — one for each auxiliary task — *discriminatively* so that the final feature space is optimal for the target.

2.3. Discriminative learning

In this final step, we reveal the second use of landmarks beyond constructing auxiliary tasks. We will use their labels to learn *discriminative* domain-invariant features for the target domain. Concretely, we compose the features for the original adaptation problem with the auxiliary tasks' features as a basis.

We scale and concatenate those features $\{\sqrt{w_q}\Phi_q(\boldsymbol{x})\}_{q=1}^{\mathsf{Q}}$ into a super-feature vector $\boldsymbol{f}$. Learning $\{w_q\}$ is cast as learning a convex combination of all kernels $\boldsymbol{G}_q$ (Lanckriet et al., 2004),

$$\boldsymbol{F} = \sum_q w_q \boldsymbol{G}_q,\ \text{s.t.}\quad w_q \geq 0 \text{ and } \sum_q w_q = 1. \tag{7}$$

We use the kernel $\boldsymbol{F}$ in training a SVM classifier and the labels of the landmarks $\{\mathcal{L}^q\}$, i.e., $\mathcal{D}_{\text{TRAIN}} = \sum_q \mathcal{L}^q$ to optimize $\{w_q\}$ discriminatively. We use $\mathcal{D}_{\text{DEV}} = \mathcal{D}_S \setminus \mathcal{D}_{\text{TRAIN}}$ be a validation dataset for model selection. Since $\mathcal{D}_{\text{TRAIN}}$ consists of landmarks that are distributed similarly to the target, we expect the classification error on $\mathcal{D}_{\text{TRAIN}}$ to be a good proxy to that of the target.

2.4. Summary

To recap our approach: i) at each granularity σ_q, we automatically select *landmarks* — individual instances that are distributed most similarly to the target; ii) we then construct *auxiliary* tasks and use their solutions as a basis for composing domain-invariant features; iii) we learn features *discriminatively*, using classification loss on the landmarks as a proxy to the discriminative loss on the target.

3. Experimental Results

We evaluate the proposed method on benchmark datasets extensively used for domain adaptation in the contexts of object recognition (Gopalan et al., 2011; Gong et al., 2012; Saenko et al., 2010; Kulis et al., 2011) and sentiment analysis (Blitzer et al., 2007). We compare the proposed method to several competitive ones. Empirical results show that our method outperforms all prior techniques in almost all cases.

3.1. Object recognition

We use 4 datasets of object images: CALTECH (Griffin et al., 2007), AMAZON, WEBCAM, and

Table 1. Recognition accuracies on 9 pairs of source/target domains are reported. C: CALTECH, A: AMAZON, W: WEBCAM, D: DSLR. The proposed method (LANDMARK) performs the best on 8 out of 9 pairs, among all unsupervised methods.

%	A→C	A→D	A→W	C→A	C→D	C→W	W→A	W→C	W→D
NO ADAPTATION	41.7	41.4	34.2	51.8	54.1	46.8	31.1	31.5	70.7
TCA (Pan et al., 2009)	35.0	36.3	27.8	41.4	45.2	32.5	24.2	22.5	80.2
GFS (Gopalan et al., 2011)	39.2	36.3	33.6	43.6	40.8	36.3	33.5	30.9	75.7
GFK (Gong et al., 2012)	42.2	42.7	40.7	44.5	43.3	44.7	31.8	30.8	75.6
SCL (Blitzer et al., 2006)	42.3	36.9	34.9	49.3	42.0	39.3	34.7	32.5	83.4
KMM (Huang et al., 2007)	42.2	42.7	42.4	48.3	53.5	45.8	31.9	29.0	72.0
METRIC (Saenko et al., 2010)	42.4	42.9	**49.8**	46.6	47.6	42.8	38.6	33.0	**87.1**
LANDMARK (**ours**)	**45.5**	**47.1**	**46.1**	**56.7**	**57.3**	**49.5**	**40.2**	**35.4**	75.2

DSLR (Saenko et al., 2010). Each dataset is treated as a separate domain: images in AMAZON came from online catalogs, images in DSLR and WEBCAM were captured by a digital SLR camera and a webcam with high and low resolutions, respectively. 10 object classes are common to all 4 datasets. The number of images per class ranges from 15 (in DSLR) to 30 (WEBCAM), and to 100 (CALTECH and AMAZON). Due to its small size, DSLR is not used as a source domain. We experiment extensively on the remaining 9 possible domain pairs.

We follow the previously reported protocols for preparing features (Saenko et al., 2010). SURF features are quantized into an 800-bin histogram with codebooks computed via K-means on a subset of images from AMAZON. The histograms are standardized such that each dimension is zero-mean and unit-standard deviation within each domain, and are publicly available.[1]

We compare to several leading approaches and variants of our own approach. We follow the standard procedures for selecting models and tuning hyperparameters. Whenever other approaches do not state clearly the tuning process, we give them the benefit of the doubt by reporting their best results by revealing the target domain labels to those algorithms. (Our method does not use those labels to tune its models.)

The bandwidth parameters σ_q for the Gaussian RBF kernels used for selecting landmarks (cf. section 2.1) are chosen as $\sigma_q = 2^q \sigma_0$, where $q \in \{-6, -5, \cdots, 5, 6\}$. The σ_0 is the median distance computed over all pairwise data points, cf. eq (4). The metric $\boldsymbol{M}$ for computing the distances is chosen to be the kernel from the GFK method (Gong et al., 2012) using *all* instances.

Recognition accuracies Table 1 reports object recognition accuracies on the *target* under 9 pairs of source and target domains. We contrast the proposed approach (LANDMARK) to the methods of transfer component analysis (TCA) (Pan et al., 2009), geodesic flow sampling (GFS) (Gopalan et al., 2011), the GFK (GFK) (Gong et al., 2012), structural correspondence learning (SCL) (Blitzer et al., 2006), kernel mean matching (KMM) (Huang et al., 2007), and a metric learning method (METRIC) (Saenko et al., 2010) for *semi-supervised* domain adaptation, while label information (1 instance per category) from the target domains is used. We also report the baseline results of NO ADAPTATION, where we use source-only data and the original features to train classifiers.

Our approach LANDMARK clearly performs the best on almost all pairs, even the METRIC method which has access to labels from the target domains. The only significant exception is on the pair WEBCAM → DSLR. Error analysis reveals that the two domains are very similar, containing images of the same object instance with different imaging resolutions. As such, many data points in WEBCAM have been selected as landmarks, leaving very few instances for model selection during the discriminative training. Addressing this issue is left for future work.

Detailed analysis on landmarks The notion of landmarks is central to our approach. In what follows, we further examine its utility in domain adaptation. We first study whether *automatically selected* landmarks coincide with our modeling intuition, ie, that they look like samples from the target domain.

Fig. 2 confirms our intuition. It displays several landmarks selected from the source domain AMAZON when the target domain is WEBCAM. The top-left panels display representative images from the HEADPHONE and MUG categories from WEBCAM, and the remaining panels display images from AMAZON, including both landmarks and non-landmarks.

When the scale σ is large, the selected landmarks are very similar in visual appearance to the representative images. As the scale decreases, landmarks with greater variance start to show. This is particularly pronounced

[1] `http://www-scf.usc.edu/~boqinggo/da.html`

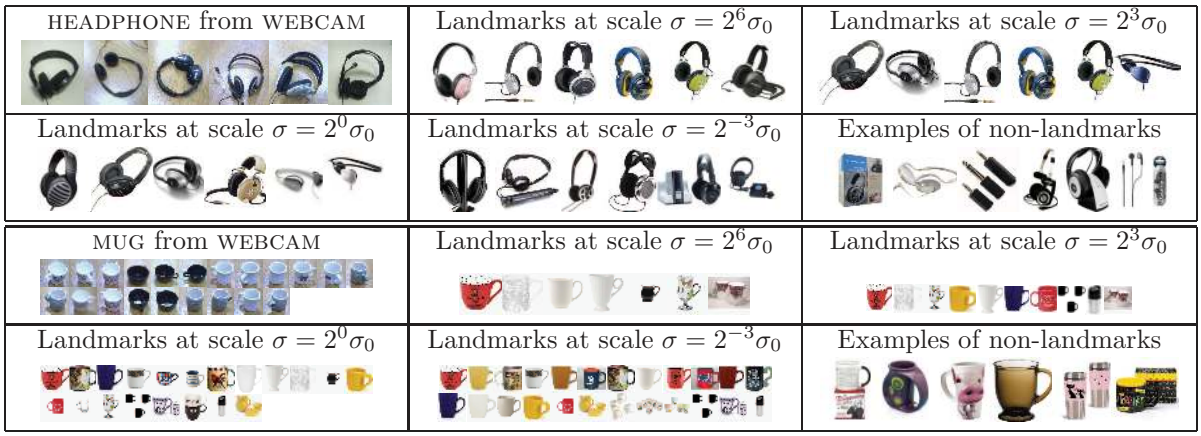

Figure 2. Landmarks selected from the source domain AMAZON for the target domain WEBCAM, as well as non-landmarks (best viewed in color). As the scale decreases, images with greater variance in appearance are selected, as expected.

Table 2. Contrasting LANDMARK to several variants, illustrating the importance of our landmark selection algorithm.

%	A→C	A→D	A→W	C→A	C→D	C→W	W→A	W→C	W→D
LANDMARK (**ours**)	**45.5**	47.1	46.1	**56.7**	**57.3**	49.5	**40.2**	**35.4**	**75.2**
RAND. SEL.	44.5	44.5	41.9	53.8	49.9	49.5	39.8	34.1	74.2
SWAP	41.3	**47.8**	37.6	46.2	42.0	46.1	38.2	32.2	70.1
UNBALANCED	37.0	36.9	38.3	55.3	49.0	**50.1**	39.4	34.9	73.9
EUC. SEL.	44.5	44.0	41.0	50.2	40.1	45.1	39.1	34.5	67.5

at $2^{-3}\sigma_0$. Nonetheless, they still look far more likely to be from the target WEBCAM domain than non-landmark images (see bottom-right panels). Note that the non-landmark images for the HEADPHONE category contain images such as earphones, or headphones in packaging boxes. Similarly, non-landmark images in the MUG category are more unusually shaped ones.

In Table 2, we contrast our method to some of its variants, illustrating quantitatively the novelty and significance of using landmarks to facilitate adaptation.

First, we study the adverse effect of selecting incorrect images as landmarks. The row of RAND. SEL. displays results of randomly selecting landmarks, as opposed to using the algorithm proposed in section 2.1. (The number of random landmarks is the average number of "true" landmarks chosen in LANDMARK.) The averaged accuracies over 10 rounds are reported (Standard errors are reported in the Suppl). LANDMARK outperforms the random strategy, often by a significant margin, validating the automatic selection algorithm.

The SWAP row in Table 2 gives yet another strong indication of how landmarks could be viewed as samples from the target. Recall that landmarks are used as *training* data in the final stage of our learning algorithm to infer the domain-invariant feature space (cf. section 2.3). Other instances, ie, non-landmarks in the source, are used for model selection. This setup follows the intuition that as landmarks are mostly similar to the target, they are a better proxy than non-landmarks for optimizing discriminative loss for the target.

When we swap the setup, the accuracies drop significantly, except on the pair $A \rightarrow D$ (compare the rows SWAP and LANDMARK). This once again establishes the unique and extremely valuable role of landmarks.

We also study the usefulness of the class balancing constraint in eq. (2), which enforces that the selected landmarks obey the class prior distribution. Without it, some classes dominate and would result in poor classification results on the target domain. This is clearly evidenced in the row of UNBALANCED where accuracies drop significantly after we remove the constraint.

Finally, we study the effect of using GFK to measure distribution similarity, as in eq. (4). The row of EUC. SEL. reports the results of using the conventional Euclidean distance, illustrating the striking benefit of using GFK (in the row of LANDMARK). While using nonparametric two-sample tests to measure distribution similarity has been previously used for domain adaptation (e.g., kernel mean matching, cf. the row of KMM in Table 1), selecting a proper kernel has received little attention, despite its vital importance. Our comparison to EUC. SEL. indicates that measuring distribution similarity *across* domains is greatly enhanced with a kernel revealing domain-invariant features.

Table 3. Sentiment classification accuracies on target domains. K: KITCHEN, D: DVD, B: BOOKS, E: ELECTRONICS

%	K→D	D→B	B→E	E→K
NO ADAPTATION	72.7	73.4	73.0	81.4
TCA	60.4	61.4	61.3	68.7
GFS	67.9	68.6	66.9	75.1
GFK	69.0	71.3	68.4	78.2
SCL	72.8	76.2	75.0	82.9
KMM	72.2	78.6	76.9	**83.5**
METRIC	70.6	72.0	72.2	77.1
LANDMARK (**ours**)	**75.1**	**79.0**	**78.5**	**83.4**

3.2. Sentiment analysis

Next, we report experimental results on the task of cross-domain sentiment analysis of text. We use the Amazon dataset described in (Blitzer et al., 2007). The dataset consists of product reviews on kitchen appliances, DVDs, books, and electronics. There are 1000 positive and 1000 negative reviews on each product type, each of which serves as a domain. We reduce the dimensionality to use the top 400 words which have the largest mutual information with the labels. We have found this preprocessing does not reduce performance significantly, while being computationally advantageous. We use bag-of-words as features.

In Table 3, we compare our LANDMARK method to leading methods for domain adaptation, including TCA (Pan et al., 2009), GFS (Gopalan et al., 2011), GFK (Gong et al., 2012), SCL (Blitzer et al., 2006), KMM (Huang et al., 2007), METRIC (Saenko et al., 2010), as well as the baseline NO ADAPTATION.

Note that while SCL and KMM improve over the baseline, the other three methods underperform. Nonetheless, our method outperforms almost all other methods. Most interestingly, our method improves GFK significantly. We attribute its advantages to two factors: using multiple scales to analyze distribution similarity while GFK uses the "default" scale, and using landmarks to *discriminatively* learn invariant features.

3.3. Summary of supplementary material

We provide more detailed results including standard errors and comparison to other methods. We also study the benefits of having multiple auxiliary tasks. We show that while individual auxiliary tasks can lead to improved performance in adaptation, combining multiple of them with the multiple kernel learning framework (cf section 2.3) improves further. Namely, constructing auxiliary tasks using multiple scales to reflect similarities at different granularities yields different views of the adaptation problem, and the learning framework successfully exploits them.

4. Related Work

Learning domain-invariant feature representations has been extensively studied in the literature (Ben-David et al., 2007; Blitzer et al., 2006; 2007; Daumé, 2007; Chen et al., 2011; Pan et al., 2009). However, identifying and using instances that are distributed similarly to the target to bridge the two domains has never been explored before.

Our approach is also very different from transductive-style domain adaptation methods (Bergamo & Torresani, 2010; Chen et al., 2011). We partition the source domain into two disjoint subsets, only once for each auxiliary task. In those methods, however, the target and the source domains are merged iteratively.

Kernel mean matching has previously been used to *weigh* samples from the source (Huang et al., 2007; Pan et al., 2009; Gretton et al., 2009) to correct the mismatch between the two domains. We *select* samples as our landmarks. Those prior works typically do not yield sparse solutions (of the weights), and thus do not perform *selection*. Additionally, the inclusion of the balancing constraint in our formulation of eq. (3) is crucial, as evidenced in our comparison to those methods in experimental studies (cf. Table 1 and Table 3). Without it, some classes could be underrepresented in selected landmarks, leading to poor performance.

5. Conclusion

Distribution similarity is central to learning invariant features across domains. While existing approaches focus on treating all samples as a whole block, we have proposed an instance-based approach. At the core is the idea of exploiting landmarks, which are data instances from the source that are distributed similarly to the target. The landmarks enable analyzing distribution similarity on multiple scales, hypothesizing a basis for invariant features, and discriminatively learning features. On benchmark tasks in both vision and text processing, our method consistently outperforms others, often by large margins. Thus, our approach has broad application potential to other tasks and domains. For future work, we plan to advance in this direction further, for example, proposing other mechanisms to identify and select landmarks.

Acknowledgements

This work is partially supported by DARPA D11AP00278 and NSF IIS-1065243 (B. G. and F. S.), and ONR ATL #N00014-11-1-0105 (K. G.).

References

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *NIPS*, 2007.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Bergamo, A. and Torresani, L. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, 2010.

Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

Blitzer, J., Foster, D., and Kakade, S. Domain adaptation with coupled subspaces. In *AISTATS*, 2011.

Chen, M., Weinberger, K.Q., and Blitzer, J.C. Co-training for domain adaptation. In *NIPS*, 2011.

Daumé, H. Frustratingly easy domain adaptation. In *ACL*, 2007.

Daumé, H. and Marcu, D. Domain adaptation for statistical classifiers. *JAIR*, 26(1):101–126, 2006.

Daumé, H., Kumar, A., and Saha, A. Co-regularization based semi-supervised domain adaptation. In *NIPS*, 2010.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

Gopalan, R., Li, R., and Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. In *NIPS*. 2006.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Scholkopf, B. Covariate shift by kernel mean matching. In Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N.D. (eds.), *Dataset Shift in Machine Learning*. MIT Press, 2009.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.

Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., and Scholkopf, B. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

Kulis, B., Saenko, K., and Darrell, T. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. El, and Jordan, M. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009a.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Multiple source adaptation and the rényi divergence. In *UAI*, 2009b.

Pan, S.J. and Yang, Q. A survey on transfer learning. *Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Pan, S.J., Tsang, I.W., Kwok, J.T., and Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. NN*, (99):1–12, 2009.

Perronnin, F., Senchez, J., and Liu, Y. Large-scale image categorization with explicit data embedding. In *CVPR*, 2010.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N.D. *Dataset shift in machine learning*. The MIT Press, 2009.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. LabelMe: a database and web-based tool for image annotation. *IJCV*, 77:157–173, 2008.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *ECCV*, 2010.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Torralba, A. and Efros, A.A. Unbiased look at dataset bias. In *CVPR*, 2011.