

Joint cross-domain classification and subspace learning

Basura Fernando¹, Tatiana Tommasi¹, and Tinne Tuytelaars¹

¹ESAT-PSI/VISICS - iMinds, KU Leuven, Kasteelpark Arenberg 10 , B-3001 Leuven, Belgium

Abstract

Domain adaptation aims at adapting a prediction function trained on a source domain, for a new different but related target domain. Recently several subspace learning methods have proposed adaptive solutions in the unsupervised case, where no labeled data are available for the target. Most of the attention has been dedicated to searching a new low-dimensional domain-invariant representation, leaving the definition of the prediction function to a second stage. Here we propose to learn both *jointly*. Specifically we learn the source subspace that best matches the target subspace while at the same time minimizing a regularized misclassification loss. We provide an alternating optimization technique based on stochastic sub-gradient descent to solve the learning problem and we demonstrate its performance on several domain adaptation tasks.

1 Introduction

In real world applications having a probability distribution mismatch between the training and the test data is more often the rule than an exception. Think about part of speech tagging across different text corpora [5], localization over time with wifi signal distributions that get easily outdated [33], or biological models to be used across different subjects [31]. Computer vision methods are also particularly challenged in this respect: real world conditions may alter the image statistics in many complex ways (lighting,

pose, background, motion blur etc.), to not even mention the difference in quality of the acquisition device (e.g. resolution), or the high number of possible artificial modifications obtained by post-processing (e.g. filtering). Due to this large variability, any learning algorithm trained on a *source* set regardless of the final *target* data will most likely produce poor, unsatisfactory results.

Domain Adaptation (DA) techniques propose to overcome these issues and make use of information coming from both source and target domains during the learning process to adapt the classifier automatically. In the *unsupervised* case, where no labeled samples are provided for the target, one of the most extensively studied paradigms consists in assuming the existence of a domain-invariant low dimensional feature space and searching for it. In general all the techniques based on this idea focus on transforming the representation of the source and target samples to maximize some notion of similarity between them [12, 13, 10]. However in this way the classification task is left aside and the prediction model is learned only in a second stage. As thoroughly discussed in [2, 24], the choice of the feature representation able to reduce the domain divergence is indeed a crucial factor for the adaptation. Nevertheless it is not the only one. What would happen if we project the source and target data into a low dimensional space that makes the two domains very much alike but loses the discriminative information? As an extreme case, both the domains can be projected to the exact same uninformative point. A common latent subspace where the difference between the marginal distributions of the two domains is minimized can be identified disregarding the conditional distribution of

¹Paper is under consideration at Pattern Recognition Letters.

the data, which however, is essential for training a discriminative classifier that generalizes from source to target. If it happens that in several subspaces the domains have similar marginals, would a classifier perform equally well in all of them? Here we answer these questions by presenting an algorithm that learns jointly both a low dimensional representation and a reliable classifier by optimizing a trade-off between the source-target similarity and the source training error.

2 Related Work

DA has received a lot of attention in the last years and there are two main scenarios studied in the literature. In the *semi-supervised* setting a few labeled samples are provided for the target domain besides a large amount of annotated source data. In the *unsupervised* setting, the available target samples are unlabeled.

In the first case, by knowing the class to which some target data belong, it is possible to impose similarity and dissimilarity constraints across samples through metric learning [20, 28], or optimize a classifier over both domains [9, 18, 7] even with extensions to cases where more than two domains are available [8, 17]. The unlabeled part of the target has also been used for co-regularization [21] with the aim of better integrate the source and the target classifiers.

In the more challenging unlabeled setup most of the approaches resort to the estimating the data distributions and minimizing a distance measure between them while re-weighting/selecting the samples [31, 11] or while looking for a new representation [1]. The Maximum Mean Discrepancy (MMD) [14] which maps two sets of data to a reproducing Kernel Hilbert Space has been largely used as distance measure between two distributions. Although endowed with nice properties, the choice of the kernel and kernel parameters is critical and if non-optimal can lead to very poor estimate of the distribution distance [15]. Dictionary learning methods have also been used with the goal of defining new representations that overcome the domain shift [25, 29]. Another promising direction is that of *subspace modeling*. As for dictio-

nary learning, the approaches presented in this framework are mostly linear, but can be easily extended to non-linear spaces through explicit feature mappings [32].

In [4] Canonical Correlation Analysis (CCA) has been applied to find a coupled domain-invariant subspace. Principal Component Analysis (PCA) is also widely used for subspace generation and several methods have proposed to exploit intermediate subspaces to link the source and the target data. This idea was introduced in [13] where the path across the domains is defined as a geodesic curve over a Grassmann manifold. This approach has been further extended in [12] where all the intermediate subspaces are integrated to define a cross-domain similarity measure. Despite the intuitive characterization of the problem, it is not clear why all the subspaces along this path should yield meaningful representations. Recently the Subspace Alignment (SA) method [10] demonstrated that it is possible to map directly from the source to the target subspace without necessarily passing through intermediate steps. Overall, the main focus of all the unsupervised subspace methods is on the domain invariance of the final data representation and less attention has been dedicated to its discriminative power. A first attempt in this direction has been done in [12] by substituting the use of PCA over the source subspace with Partial Least Squares (PLS).

Our work fits in this context and go over coding discriminativeness in the subspace representation. We propose here to integrate subspace and max-margin learning, jointly looking for the optimal low-dimensional projection and for the optimal classifier in an unsupervised DA setting. We underline here that, although solutions similar in spirit have been exploited in the semi-supervised setup [18, 7], we consider the specific case of subspace modeling and the unsupervised case where no access to the target labels is available, not even for hyperparameter cross validation.

In the following sections we define the notation that will be used in the rest of the paper (section 3) and we briefly revise the theory of learning from different domains together with the SA method on which we build. We then introduce our approach

(section 4) followed by an extensive experimental analysis that shows its effectiveness on several domain adaptation tasks (section 5). We conclude with a final discussion and sketching possible directions for future research (section 6).

3 Problem Setup and Background

Let us consider a classification problem where the data instances are in the form $(\mathbf{x}_i, y_i)$. Here $\mathbf{x}_i \in \mathbb{R}^D$ is the feature vector for the i -th sample and $y_i \in \{1, \dots, C\}$ is the corresponding label. We assume that n_s labeled training samples are drawn from a source distribution, while a set of n_t unlabeled test samples come from a different target distribution such that for the joint probabilities $\mathcal{D}_s = P(\mathbf{x}^s, y^s)$, $\mathcal{D}_t = P(\mathbf{x}^t, y^t)$ it holds $\mathcal{D}_s \neq \mathcal{D}_t$.

A bound on target domain error Theoretical studies on DA allowed to establish the conditions in which a classifier trained on the source data can be expected to perform well on the target data. Specifically, the following generalization bound on the target error ϵ_t has been demonstrated in [2]:

$$\epsilon_t(h) \leq \epsilon_s(h) + d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \lambda. \quad (1)$$

Here h indicates the predictor function, while $\mathcal{H}$ is the hypothesis class from which the predictor has been chosen. In words, the bound states that a low target error can be guaranteed if the source error $\epsilon_s(h)$, a measure of the domain distribution divergence $d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t)$, and the error λ of the ideal joint hypothesis on the two domains are small. Specifically with $h^* = \operatorname{argmin}_{h \in \mathcal{H}} (\epsilon_t(h) + \epsilon_s(h))$ the joint error is $\lambda = \epsilon_t(h^*) + \epsilon_s(h^*)$ and its value is supposed to be low if adaptation is possible. All the three terms on the right hand side of (1) contribute to the adaptability of the classifier.

Subspace Alignment (SA) Given the source and the target domains, we can specify their low-dimensional intrinsic structure by the corresponding orthonormal basis set. Specifically we use $S \in \mathbb{R}^{D \times d}$

for the source subspace and $T \in \mathbb{R}^{D \times d}$ for the target subspace where d is the subspace dimensionality. When using PCA, the basis correspond to the top d eigenvectors of the data's covariance matrix. Alternatively, discriminative subspaces can be determined by using the source labels with PLS or Linear Discriminant Analysis (LDA). In this last case the resulting subspace dimensionality depends on the number of classes by $d = C - 1$.

The domain shift can be reduced by aligning the associated basis sets and minimizing the following Bregman matrix divergence [10]

$$F(M) = \|SM - T\|_F^2. \quad (2)$$

Since the Frobenius norm $\|\cdot\|_F$ is invariant to orthonormal operations, we can apply $S^\top$ to both the terms of the difference and demonstrate that the optimal transformation matrix has a closed form solution $M = S^\top T \in \mathbb{R}^{d \times d}$. Thus, it is possible to define a new *target aligned* coordinate system $U = SM = SS^\top T$ where $U \in \mathbb{R}^{D \times d}$. This approach has shown promising results for visual cross-domain classification tasks outperforming other subspace adaptive methods [10]. However, on par with its competitors [13, 12], it keeps the domain adaptive process (learning M) and the classification process (e.g. learning the SVM model) separated, focusing only on the distribution divergence term of the bound in (1).

4 Proposed Approach

Under the covariate shift assumption we can consider the error λ in (1) as negligible. Thus, to optimize the target performance, we can concentrate on the two remaining terms on the right hand side of the bound and search jointly for a subspace representation that minimizes the domain divergence and for the best source classification model in the same space. We name our method *Joint cross-domain Classification and Subspace Learning* (JCSL) and we presents its details below.

Given a fixed target subspace basis $T \in \mathbb{R}^{D \times d}$ we

minimize the following regularized risk functional

$$F(U, \mathbf{w}) = \|\mathbf{w}\|_2^2 + \alpha \|U - T\|_F^2 + \beta \sum_i^{n_s} \mathcal{L}(\mathbf{x}_i^s, y_i^s, \mathbf{w}, U) \quad (3)$$

Here $\mathbf{w} \in \mathbb{R}^d$ is a linear source classifier, while $U \in \mathbb{R}^{D \times d}$ is a matrix that projects the source data to the target aligned subspace. In the defined objective function, the regularization terms aim at separately optimizing over the two variables $\mathbf{w}$ and U , while the loss function $\mathcal{L}$ leverages over their combination. In particular we use a domain adaptive regularization for U following the idea of the SA method. However, in our case it is not necessary to specify a priori the source subspace S which is now optimized together with the alignment transformation matrix M in a single step. Note that if the source and target data can be considered as belonging to the same domain (no domain shift) our method will automatically provide $U = T$ boiling down to standard learning in the shared subspace. We followed previous literature in using PCA to define the target subspace T [10, 12]. Besides having demonstrated empirical good results, this choice has a theoretical motivation. We can write the mutual information between the target and the source as $MI(\text{source}; \text{target}) = H(\text{target}) - KL(\text{source} || \text{target})$, where $KL(\cdot || \cdot)$ is the Kullback-Leibler divergence and $H(\cdot)$ is the entropy. Projecting the target data to the subspace T maximizes the entropy $H(\text{target})$, while our objective function (3) minimizes the domain divergence. Hence, we expect to get a high mutual information between source and target.

For our analysis we focus on classification tasks and we choose the hinge loss: $\mathcal{L}(\mathbf{x}_i^s, y_i^s, \mathbf{w}, U) = \max\{0, 1 - \mathbf{x}_i^{s\top} U \mathbf{w} y_i^s\}$, but other loss functions can be used for different cross-domain applications. Minimizing (3) jointly over $(U, \mathbf{w})$ is a non-convex problem and finding a global optimum is generally intractable. However we can apply alternated minimization for U and $\mathbf{w}$ resulting in two interconnected convex problems that can be efficiently solved by stochastic subgradient descent. The iterative procedure terminates when the algorithm converges, showing a negligible change of either U or $\mathbf{w}$ between two consecutive iterations. The formulation

holds for a binary classifier but can easily be used in its one-vs-all multiclass extension that highly benefits from the choice of the stochastic variant of the optimization process. If we indicate the classification score of class y for the target sample $\mathbf{x}_i^t$ with $s(\mathbf{x}_i^t, \mathbf{w}_y) = \mathbf{x}_i^{t\top} T \mathbf{w}_y$, the final prediction can be written as $y_i^* = \arg\max_y (s(\mathbf{x}_i^t, \mathbf{w}_y))$. The learning strategy is summarized in Algorithm 1.

Algorithm 1 JCSL

Input: step size η and batch size γ for stochastic sub-gradient descent
Output: $U^*, \mathbf{w}^*$

- 1: Initialize $U \leftarrow S, \mathbf{w} \leftarrow 0, k \leftarrow 0$
- 2: **while** not converged **do**
- 3: $k \leftarrow k + 1$
- 4: calculate the partial derivatives:
- 5: $\frac{\partial F(U, \mathbf{w})}{\partial U} = 2\alpha(U - T) - \beta \sum_{i=1}^{\gamma} \Gamma_i$
- 6: $\Gamma_i = \begin{cases} \mathbf{x}_i^{s\top} \mathbf{w} y_i^s & \text{if } (\mathbf{x}_i^{s\top} U \mathbf{w} y_i) < 1 \\ 0 & \text{otherwise} \end{cases}$
- 7: $\frac{\partial F(U, \mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{w} - \beta \sum_{i=1}^{\gamma} \Theta_i$
- 8: $\Theta_i = \begin{cases} \mathbf{x}_i^{s\top} U y_i^s & \text{if } (\mathbf{x}_i^{s\top} U \mathbf{w} y_i) < 1 \\ 0 & \text{otherwise} \end{cases}$
- 9: Fix U , identify the optimal $\mathbf{w}$:
- 10: $\mathbf{w}_k \leftarrow \mathbf{w}_{k-1} - \eta \left(\frac{\partial F(U, \mathbf{w})}{\partial \mathbf{w}} \right)_{\mathbf{w}_{k-1}}$
- 11: Fix $\mathbf{w}$, identify the optimal U :
- 12: $U_k \leftarrow U_{k-1} - \eta \left(\frac{\partial F(U, \mathbf{w})}{\partial U} \right)_{U_{k-1}}$
- 13: **end while**

5 Experiments

We validate our approach over several domain adaptation tasks. In the following we first describe our experimental setting and then we report on the obtained results with a detailed analysis on how JCSL effectively optimizes over the source classification error and the cross-domain similarity.

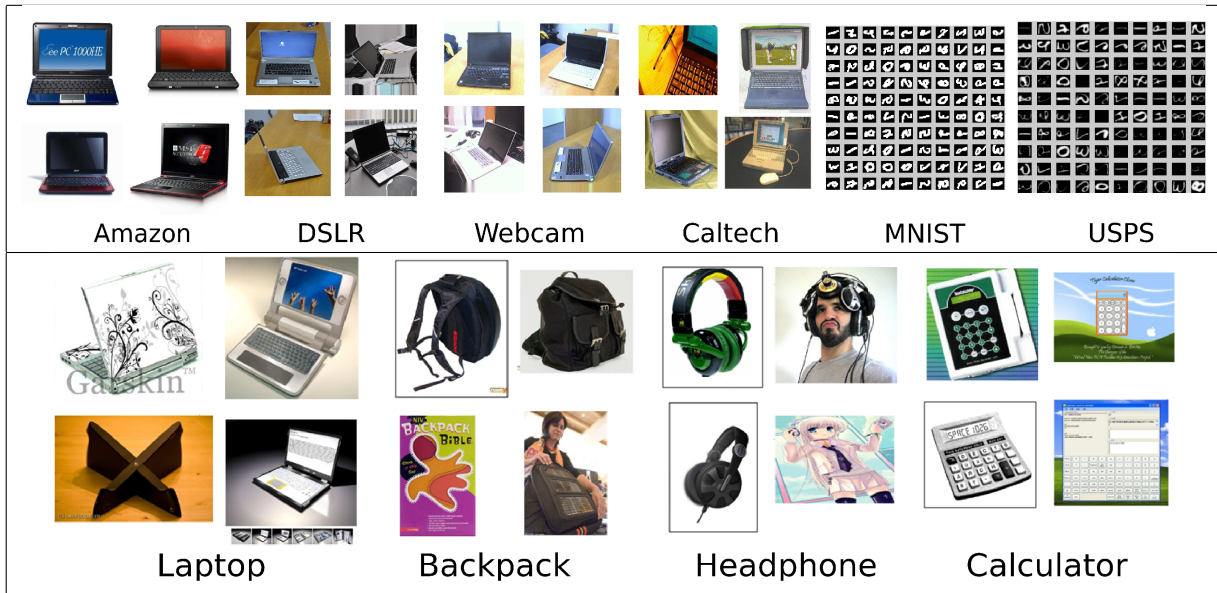


Figure 1: Top line: examples from Office+Caltech dataset and the MNIST+USPS dataset. Bottom line: weakly labeled images from Bing dataset.

5.1 Datasets, baselines and implementation details

We choose three image datasets (see Figure 1) and a wifi signal dataset.

Office + Caltech [12]. This dataset was created by combining the Office dataset [28] with Caltech256 [16] and it contains images of 10 object classes over four domains: **A**maz^on, **D**slr, **W**ebcam and **C**altech. Amazon consists of images from online merchants’ catalogues, while Dslr and Webcam domains are composed by respectively high and low resolution images. Finally, Caltech corresponds to a subset of the original Caltech256. We use the features provided by Gong et al. [12] already used in several previous publications: SURF descriptors quantized into histograms of 800 bag-of-visual words and standardized by z-score normalization. All the 12 possible source-target domain pairs are considered. We use the data splits provided by Hoffman et al. [18].

MNIST [22] + **USPS** [19]. This dataset combines two existing image collections of digits present-

ing different gray scale data distributions. Specifically they share 10 classes of digits. We randomly selected 1800 images from USPS and 2000 images from MNIST. By following [23] we uniformly re-scale all images to size 16×16 and we use the gray-scale pixel values as feature vectors. Both domains are alternatively used as source and target.

Bing+Caltech [3]. In this dataset, weakly annotated images from the Bing search engine define the source domain while images of Caltech256 are used as target. We run experiments varying the number of categories (5, 10, 15, 20, 25 and 30) and the number of source examples per category (5 and 10) using the same train/test split adopted in [3]. As typically done for this dataset, Classemes features are used as image representation [3].

WiFi [33]. This dataset was used in the 2007 IEEE ICDM contest for domain adaptation. The goal is to estimate the location of mobile devices based on the received signal strength (RSS) values from different access points. The domains correspond to two different time periods during which the collected RSS

Table 1: Recognition rate (%) results over the Office+Caltech and MNIST+USPS datasets.

DA Problem	NA	PCA_T	SA	GFK	TCA	TSL	JCSL
A $\rightarrow$ C	38.1 $\pm$ 2.6	41.1 $\pm$ 1.7	43.4 $\pm$ 3.2	43.2 $\pm$ 3.7	43.5 $\pm$ 3.2	40.4 $\pm$ 0.9	42.6 $\pm$ 0.9
A $\rightarrow$ D	32.9 $\pm$ 2.8	37.5 $\pm$ 1.4	44.7 $\pm$ 2.6	43.7 $\pm$ 2.8	38.8 $\pm$ 2.1	40.8 $\pm$ 1.7	42.5 $\pm$ 3.2
A $\rightarrow$ W	36.8 $\pm$ 2.9	39.1 $\pm$ 2.8	40.3 $\pm$ 2.9	41.3 $\pm$ 2.0	41.0 $\pm$ 1.0	41.1 $\pm$ 2.3	47.6 $\pm$ 2.1
C $\rightarrow$ A	39.5 $\pm$ 1.0	40.6 $\pm$ 2.8	39.3 $\pm$ 1.8	39.9 $\pm$ 1.5	42.5 $\pm$ 1.2	43.0 $\pm$ 4.2	44.3 $\pm$ 1.2
C $\rightarrow$ D	38.8 $\pm$ 2.4	40.9 $\pm$ 1.3	44.0 $\pm$ 2.0	42.0 $\pm$ 2.5	42.1 $\pm$ 2.5	40.4 $\pm$ 3.4	46.5 $\pm$ 1.5
C $\rightarrow$ W	37.8 $\pm$ 1.4	35.9 $\pm$ 3.2	37.3 $\pm$ 3.3	41.8 $\pm$ 3.8	39.4 $\pm$ 2.5	40.9 $\pm$ 2.4	46.5 $\pm$ 2.0
D $\rightarrow$ A	24.4 $\pm$ 1.7	30.6 $\pm$ 2.7	35.7 $\pm$ 2.3	31.0 $\pm$ 2.7	30.4 $\pm$ 2.1	39.6 $\pm$ 1.2	41.3 $\pm$ 0.9
D $\rightarrow$ C	30.5 $\pm$ 2.1	37.9 $\pm$ 1.3	41.5 $\pm$ 1.4	40.9 $\pm$ 2.8	36.7 $\pm$ 2.6	33.4 $\pm$ 2.1	35.1 $\pm$ 0.9
D $\rightarrow$ W	60.9 $\pm$ 3.1	67.9 $\pm$ 2.8	58.6 $\pm$ 2.1	60.5 $\pm$ 3.8	64.5 $\pm$ 3.2	73.7 $\pm$ 1.5	74.2 $\pm$ 3.6
W $\rightarrow$ A	29.7 $\pm$ 1.7	34.1 $\pm$ 1.7	34.7 $\pm$ 0.9	33.1 $\pm$ 1.2	34.6 $\pm$ 1.4	38.0 $\pm$ 1.6	43.1 $\pm$ 1.0
W $\rightarrow$ C	34.1 $\pm$ 1.6	38.1 $\pm$ 1.3	36.7 $\pm$ 1.2	37.7 $\pm$ 1.8	39.6 $\pm$ 2.2	30.4 $\pm$ 1.3	36.1 $\pm$ 2.0
W $\rightarrow$ D	71.1 $\pm$ 2.6	74.6 $\pm$ 0.9	69.5 $\pm$ 2.4	75.3 $\pm$ 2.6	77.3 $\pm$ 2.7	66.9 $\pm$ 1.6	66.2 $\pm$ 2.9
AVG.	39.6 $\pm$ 2.2	43.2 $\pm$ 2.0	43.8 $\pm$ 2.2	44.2 $\pm$ 2.6	44.2 $\pm$ 2.2	44.0 $\pm$ 2.0	47.2 $\pm$ 1.9
MNIST $\rightarrow$ USPS	45.4	45.1	48.6	34.6	40.8	43.5	46.7
USPS $\rightarrow$ MNIST	33.3	33.4	22.2	22.6	27.4	34.1	35.5
AVG.	39.4	39.2	35.4	28.6	34.1	38.8	41.1

values may have different distributions. The dataset contains 621 labeled examples collected during time period A (source) and 3128 unlabeled examples collected during time period B (target). There are 248 locations in total, each considered as a class. All source samples are used during the training and the classification accuracy is measured using all test samples.

We benchmark JCSL against four state of the art subspace-based DA methods, namely Geodesic Flow Kernel (**GFK**) [12], Transfer Subspace Learning (**TSL**) [30], Transfer Component Analysis (**TCA**) [27] and the Subspace Alignment method (**SA**) [10]. As a preliminary evaluation we compared the results obtained by GFK and SA when the basis of the source and target subspaces were obtained with PLS and LDA. Although performing similarly on average, PLS shows less stability than LDA with large changing in the outcome for small variations of the subspace dimensionality d . This can be explained considering the difficulty of finding the best d that jointly maximizes the source data/label coherence and minimizes the source/target shift. Thus, for our experiments we rely on the more stable LDA for the source. On the

other hand, the target subspace is always obtained by applying PCA. For all the methods the final classifier is a linear SVM with the C parameter cross-validated on the source over the range $\{0.001, 0.01, 0.1, 1.0, 10\}$. As a further baseline we also consider the source classifier learned with no adaptation in the original feature space (**NA**) and in the target subspace (**PCA_T**).

Our JCSL has three main parameters (α, β, d) that are chosen through cross validation on the source. We searched for α, β in the same range indicated before for C . The parameter d was tuned in $\{10, 20, \dots, 100\}$ both for JCSL and for the baselines **PCA_T** and **TSL**. We implemented the stochastic sub-gradient descent using a step size of $\eta = 0.1$ and a batch size of $\gamma = 10$. The alternating optimization converges for less than 100 iterations and we can obtain the results for any of the source-target pairs of the Office+Caltech in 2 minutes using a modern desktop computer (2.8GHz cpu, 4Gb of ram, 1 core).

5.2 Results - Office+Caltech and MNIST+USPS

The obtained results over the Office+Caltech and MNIST+USPS datasets are presented in Table 1.

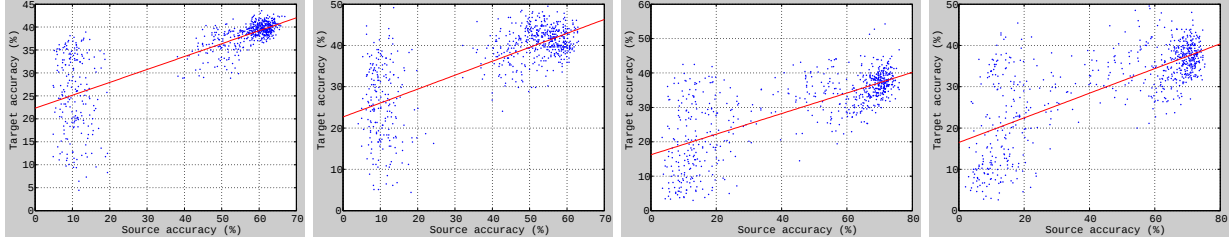


Figure 2: Target accuracy vs source accuracy over domain adaptation problem (Amazon $\rightarrow$ Caltech), (Amazon $\rightarrow$ Webcam), (MNIST $\rightarrow$ USPS) and (USPS $\rightarrow$ MNIST) obtained using our method by changing the parameters α , β and d .

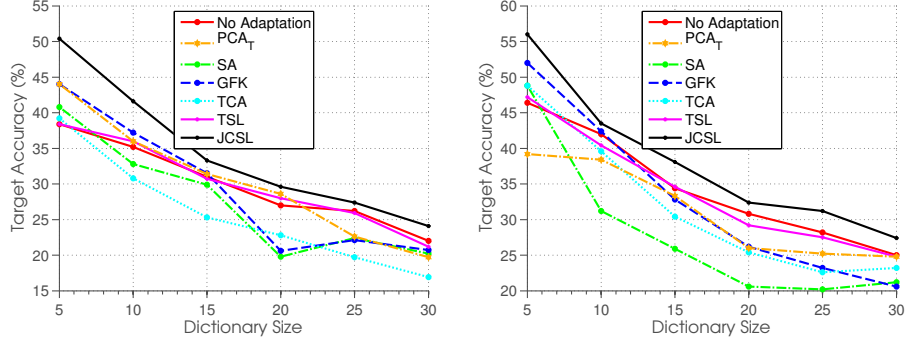


Figure 3: Experimental results on Bing+Caltech obtained when using 5 (left) and 10 (right) samples per class in the source.

Overall JCSL outperforms the considered baselines in 8 source-target pairs out of 14 and shows the best average results over the two datasets. Thus, we can state that by minimizing a trade-off between source-target similarity and the source classification error pays off compared to only reducing the cross-domain representation divergence. Still SA shows an advantage with respect to JCSL in a few of the considered cases most probably because it can exploit the discriminative LDA subspace. With respect to JCSL, TCA seems to work particularly well when the domain shift is small (i.e. Amazon $\rightarrow$ Caltech, Dslr $\rightarrow$ Webcam). Interestingly JCSL is the only method that consistently outperforms NA over MNIST+USPS.

Parameter analysis To better understand the performance of JCSL we analyze how the target ac-

curacy varies with respect to the source accuracy while changing the learning parameters α , β and d . The plots in Figure 2 consider four domain adaptation problems, namely (Amazon $\rightarrow$ Caltech), (Amazon $\rightarrow$ Webcam), (MNIST $\rightarrow$ USPS) and (USPS $\rightarrow$ MNIST)¹. All of them present two main clusters. On the left, when the source accuracy is low, the target accuracy is uniformly distributed. This behavior mostly appears when β is very small and α has a high value: this indicates that minimizing only $\|U - T\|_F^2$ does not guarantee stable results on the target task. On the other hand, in the second cluster the source accuracy is highly correlated with the target accuracy. On average for the points in this region both the domain divergence term and the misclassification

¹Analogous results are obtained for all the remaining source-target pairs.

Table 2: $\mathcal{H}\Delta\mathcal{H}$ analysis. Lower values indicate lower cross-domain distribution discrepancy.

Space	A→C	A→W	MNIST→USPS	USPS→MNIST
Original features	74.82	90.18	100.00	100.00
SA	65.96	56.56	55.78	55.74
JCSL	65.76	54.97	57.03	53.28

loss obtain low values. The final JCSL result with the optimal $(U^*, \mathbf{w}^*)$ always appear in this area. Overall the dimensionality of the subspace d seems to have only a moderate influence on the final results indicating that the choice of this parameter is less relevant for JCSL than what is for the other subspace adaptive methods. The red line reported on the plots is obtained by least-square fitting over the source and target accuracies and presents an analogous trend for all the considered source-target pairs. This is an indication that when domains are adaptable (negligible λ in (1)) our method is able to find a good source classifier as well as a source representation that generalizes to the target domain.

Measuring the domain shift For the same domain pairs considered above we also evaluate empirically the $\mathcal{H}\Delta\mathcal{H}$ divergence measure defined in [2]. This is obtained by learning a linear SVM that discriminates between the source and target instances, respectively pseudo-labeled with $+1$ and -1 . We separated each domain into two halves and use them for training and test when learning a linear SVM model. A high final accuracy indicates high domain divergence. Due to the considered one-vs-all multiple-class setting, our JCSL learns class-specific source subspaces U_y , so we calculate the average $\mathcal{H}\Delta\mathcal{H}$ divergence and compare it with the divergence obtained when using the original features and the SA target aligned subspace representation. The results in Table 2 indicate that SA and JCSL produce comparable results in terms of domain-shift reduction suggesting that the main advantage of JCSL comes from the learned classifier.

Table 3: Classification accuracy obtained over WiFi localization dataset [33].

NA	SA	GFK	TCA	JCSL
16.6	17.3	17.3	19.0	20.2

5.3 Results - Bing+Caltech

Due to the way in which it was defined, Bing+Caltech can be considered as a much more challenging testbed for unsupervised domain adaptation compared to the other used datasets (see also Figure 1). At the same time it also corresponds to one of the most realistic scenarios where domain adaptation is needed: we have access to only a limited number of noisy labeled source images obtained from the web and we want to use them to classify over a curated collection of object images. For this problem exploiting at the best all the available information is crucial. Specifically, since the source is not fully reliable, coding its discriminative information in the representation (e.g. through LDA or PLS) may be misleading. On the other hand, using the subspace of the non-noisy target data to guide the learning process can be much more beneficial.

As shown in Figure 3, JCSL is the only method that consistently improves over the non-adaptive approach independently from the number of considered classes. TSL is always equivalent to NA, while the other subspace methods, although initially helpful for problems with few classes, lose their advantage over NA when the number of classes increases. This behavior is almost equivalent when using both 5 and 10 source samples per class.

5.4 Results - WiFi Localization

To demonstrate the generality of the proposed algorithm, we evaluate JCSL also on non-visual data. Specifically, the results on the WiFi-localization task are reported in table 3. The obtained classification accuracy confirm the value of our method over the other subspace-based techniques.

6 Conclusions

Motivated by the theoretical results of Ben-David et al. [2], in this paper we proposed to integrate the learning process of the source prediction function with the optimization of the invariant subspace for unsupervised domain adaptation. Specifically, JCSL learns a representation that minimizes the Bregman matrix divergence between the source subspace and the target subspace while optimizing the classification model. Extensive experimental results have shown that, by taking advantage of the described principled combination and without the need of passing through the evaluation of the data distributions, JCSL outperform several other subspace domain adaptation methods that focus only on the representation part. Recently several works have demonstrated that Convolutional Neural Network performance are robust to domain shift [6, 26]. Reasoning at high level we can identify the cause of such a robustness on the same idea at the basis of JCSL : deep architectures learn jointly a discriminative representation *and* the prediction function. The highly non-linear transformation of the original data coded into the CNN activation values can also be used as input data descriptors for JCSL with the aim of obtaining a combined effect. As future work we plan to evaluate principled ways to find automatically the best subspace dimensionality d using low-rank optimization methods.

Acknowledgments

The authors acknowledge the support of the FP7 EC project AXES.

References

- [1] Mahsa Baktashmotlagh, Mehrtaash Harandi, Brian Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV 2013*, 2013.
- [2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*. 2007.
- [3] Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, December 2010.
- [4] John Blitzer, Dean Foster, and Sham Kakade. Domain adaptation with coupled subspaces. In *AISTATS*, 2011.
- [5] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.
- [6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531:1–8, 2013.
- [7] Lixin Duan, Ivor W. Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, March 2012.
- [8] Lixin Duan, Ivor W. Tsang, Dong Xu, and Tat-Seng Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, pages 289–296.
- [9] Lixin Duan, Ivor W. Tsang, Dong Xu, and Stephen J. Maybank. Domain transfer svm for video concept detection. In *CVPR*, pages 1375–1381, 2009.
- [10] Basura Fernando, Amaury Habrard, Marc Sebban, Tinne Tuytelaars, et al. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.
- [11] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013.

- [12] Boqing Gong, Yuan Shi, Fei Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.
- [13] R. Gopalan, Ruonan Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.
- [14] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel method for the two sample problem. In *NIPS*, pages 513–520, 2007.
- [15] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, 2012.
- [16] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.
- [17] Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *ECCV*, 2012.
- [18] Judy Hoffman, Erik Rodner, Jeff Donahue, Kate Saenko, and Trevor Darrell. Efficient learning of domain-invariant image representations. In *ICLR*, 2013.
- [19] J.J. Hull. A database for handwritten text recognition research. *PAMI*, 16(5):550–554, May 1994.
- [20] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.
- [21] Abhishek Kumar, Avishek Saha, and Hal Daume. Co-regularization based semi-supervised domain adaptation. In *NIPS*, pages 478–486, 2010.
- [22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.
- [23] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jianguang Sun, and P.S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013.
- [24] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.
- [25] Jie Ni, Qiang Qiu, and Rama Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *CVPR*, June 2013.
- [26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.
- [27] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. In *IJCAI*, 2009.
- [28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.
- [29] Sumit Shekhar, Vishal M. Patel, Hien V. Nguyen, and Rama Chellappa. Generalized domain-adaptive dictionaries. In *CVPR*, June 2013.
- [30] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *KDE*, 22:929–942, 2010.
- [31] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. In *NIPS*, 2011.
- [32] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence*, 34(3), 2011.
- [33] Qiang Yang, Sinno Jialin Pan, and Vincent Wenchen Zheng. Estimating location using wi-fi. *IEEE Intelligent Systems*, 23(1):8–13, 2008.