Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Unsupervised Domain Adaptation with a Relaxed Covariate Shift Assumption

Tameem Adel
University of Manchester, UK
tameem.hesham@gmail.com

Han Zhao
Carnegie Mellon University, USA
han.zhao@cs.cmu.edu

Alexander Wong
University of Waterloo, Canada
a28wong@uwaterloo.ca

Abstract

Domain adaptation addresses learning tasks where training is performed on data from one domain whereas testing is performed on data belonging to a different but related domain. Assumptions about the relationship between the source and target domains should lead to tractable solutions on the one hand, and be realistic on the other hand. Here we propose a generative domain adaptation model that allows for modelling different assumptions about this relationship, among which is a newly introduced assumption that replaces covariate shift with a possibly more realistic assumption without losing tractability due to the efficient variational inference procedure developed. In addition to the ability to model less restrictive relationships between source and target, modelling can be performed without any target labeled data (unsupervised domain adaptation). We also provide a Rademacher complexity bound of the proposed algorithm. We evaluate the model on the Amazon reviews and the CVC pedestrian detection datasets.

Introduction

Domain adaptation focuses on modelling problems where the training and test distributions are different but related. The training and test domains are commonly referred to in the domain adaptation literature as the source and target domains, respectively. Domain diversity can emerge as a result of the scarcity of available labeled data from the target domain. It can as well be innate in the problem itself due to, for example, an ongoing change occurring to the source domain like in cases where the original source domain keeps changing over time. Domain adaptation aims at finding solutions for this kind of problem, where the training (source) data are generated from a distribution different from that of the test (target) data, by leveraging the available labeled data from the similar source domain (Ben-David et al. 2007).

More details about seminal works in the domain adaptation literature can be found in Section 1 in the supplementary.

The principal metric of a domain adaptation algorithm is the performance of the target learning hypothesis, which as a matter of fact depends on the labeled source data, (primarily unlabeled) target data and the relationship between the source and target domains. There can be multiple source domains to learn from (Mansour, Mohri, and Rostamizadeh 2009b), but the default setting is one source and one target domain. Regarding target data, several frameworks have access to a target sample containing labeled data instances, and a larger target unlabeled sample , e.g. Blitzer et al.; Oquab et al. (2008; 2014). In other frameworks, like Huang et al. (2006), referred to as unsupervised domain adaptation, the target data is fully unlabeled. The proposed model is an unsupervised domain adaptation framework, although it can incorporate target labeled data, if available, in a forthright manner.

The manner by which source labeled data are utilized in learning the target distribution and consequently the all-important target labels heavily depends on the relationship between the source and target domains. Such relationship is depicted by the learner's assumptions about the similarities/differences between source and target domains. In this regard, covariate shift represents one of the most widely used domain adaptation assumptions (Sugiyama and Mueller 2005), on which much of the domain adaptation research is based. Covariate shift states that the conditional labeling distributions of both the source and target domains are the same, while their respective marginal data distributions can be different (Storkey and Sugiyama 2006; Ben-David and Urner 2012; 2014). Covariate shift is a valid assumption in some problems, but it can as well be quite unrealistic for many other domain adaptation tasks where the conditional label distributions are not (or, more precisely, not guaranteed to be) identical. The simplification resulting from assuming identical labeling distributions facilitates the quest for a tractable learning algorithm, albeit possibly at the cost of reducing the expressiveness power of the representation, and consequently the accuracy of the resulting hypothesis.

We propose a probabilistic relaxation of the covariate shift assumption where the more uncertain a learner is about the source labeling of an instance, the more uncertain it is that its target conditional labeling probability has a value similar to the corresponding source conditional labeling probability (the more uncertain it is that the label remains the same across domains), and vice versa. We formalize this notion in a generative model that learns a representation from both the target unlabeled data and a probabilistic prospect of the change in their labels across domains, and we use such representation to learn the target labels. The proposed model incorporates a tractable inference procedure for a broad range of problems by exploiting the recent advances in vari-

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ational inference (Rezende, Mohamed, and Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013). We focus on promoting the proposed domain adaptation model for the introduced assumption, but the model is generic enough to substantiate other adaptation assumptions, and to be considered a generalization of some other adaptation frameworks. The proposed algorithm can be seen as a Bayesian, more generic and more scalable extension of Adel and Wong (2015).

Our main contributions are as follows:

- We propose a generalized generative modelling framework for unsupervised domain adaptation, which does not require any target labeled data.
- We develop a scalable learning algorithm that broadens the range of solution tractability with less restrictive and possibly more realistic domain adaptation assumptions than covariate shift.
- We propose a relaxed probabilistic version of covariate shift and exploit the proposed model in substantiating this assumption.
- Scalability and tractability of the model are achieved by developing a variational inference procedure tailored to domain adaptation, which is based on recent advances in variational Bayesian procedures, namely approximating the posterior via a recognition model (Rezende, Mohamed, and Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013; Kingma and Welling 2014; Kingma et al. 2014). We provide a lower bound on the marginal likelihood. To the best of our knowledge, this is the first variational inference procedure for domain adaptation.
- Our generalized domain adaptation model yields a rather flexible association between the adaptation assumptions and the learning solution so that conceptual modelling details and their implementations can each be scrutinized more easily and rather separately.
- A Rademacher complexity bound is derived for the proposed algorithm.
- The model is applied to benchmark domain adaptation datasets where its performance is demonstrated by comparison with the achieved state-of-the-art results. There are mainly two experiments: A sentiment analysis task on the Amazon reviews dataset and a pedestrian detection task on the CVC-02 and CVC-04 datasets.

The rest of the paper is organized as follows: The proposed relaxed covariate shift assumption is introduced in Section , followed by an illustration of the proposed unsupervised domain adaptation model. Section describes the developed variational inference procedure. A Rademacher generalization bound is introduced in Section , and finally the two experiments are presented in Section . The related work, the proof of the Rademacher complexity bound, among other details, can be found in the supplementary.

Generative Model for Domain Adaptation

Notation Let Y be a label set. We address classification problems; $\forall y \in Y, y \in \mathbb{Z}$, at the source and target domains. Denote the number of labels by k, $|Y| = k$. Let x be a data instance. Input to the domain adaptation learner is composed of a source sample, S, consisting of n labeled instances, $(x_i, y_i), i \in \{1, 2, \cdots, n\}$, and a target sample T of m unlabeled instances, $x_j, j \in \{1, 2, \cdots, m\}$. Define the loss function of two labeling hypotheses, h_1 and h_2, by $L(h_1(x), h_2(x)) : y \times y \to R$, and the expected loss over a distribution, P, as $\mathbb{L}_P(h_1, h_2) = E_{x \in P}(L(h_1(x), h_2(x)))$. Referring to the Bayes Optimal target labeling function as $h_T^{opt}(x)$, the main goal is to develop a learning hypothesis of the target domain, h_T, with a maximized classification accuracy, i.e. with a minimized expected target loss, $\mathbb{L}_{P_T}(h_T, h_T^{opt})$. We refer to the proposed model as `GenDA`.

We begin by describing our relaxed covariate shift-based model for unsupervised domain adaptation. We assume one source and one target domain. We refer to the marginal data distributions in source and target domains as $P_S(x)$ and $P_T(x)$, respectively. The joint data and label distributions of the two domains are denoted by $P_S(x, y)$ and $P_T(x, y)$. Denote by $l_S(x)$ the probability of the assigned source label to x, i.e. the probability $P_S(y_{\max}|x)$, with $y_{\max} = \text{argmax}_{y \in Y} P_S(y|x)$. After learning the source hypothesis, we learn the latent feature space, z, which is assumed to generate both the unlabeled target instances, $x \in T$, and their corresponding source labels along with the relationship between source and target domains (modelled in our algorithm by the relaxed covariate shift assumption). We then learn the target labels, $y = \text{argmax}_{y \in Y} P_T(y|x)$, of the target sample, $x \in T$, by inferring $P_T(y|z)$. Learning target labels through $P_T(y|z)$ is more informative than through $P_T(y|x)$ since z conveys information about both the unlabeled target data, x and the corresponding source labels (source labels of the target data) mapped onto the target via the relaxed covariate shift assumption, or more generally via the assumption(s) on the relationship between source and target labels.

Relaxed Covariate Shift

We do not assume covariate shift, i.e. $l_S(x)$ and $l_T(x)$ do not have to be equal for each x. Instead, we construct a model where learning the target hypothesis is based on a nonlinear interaction of two respects: The topology of the unlabeled target data and their source labels along with the introduced probabilistic domain adaptation assumption. In the proposed model, we introduce and employ an assumption that is a relaxation of covariate shift. Even though we focus on this introduced assumption, the proposed model is not exclusively tailored for a solitary particular assumption. On the contrary, other assumptions on the relationship between source and target domains can be modelled, as we briefly show in Section 2 in the supplementary.

Rather than rigidly assuming the standard covariate shift, the relaxed covariate shift assumption states that the uncertainty in the source label assigned to a data instance, x, manifested in the source conditional labeling probability, is proportionate to the value of the loss function L between its source and target labeling functions, $L(h_S(x), h_T(x))$. In other words, the more uncertain the source hypothesis about a labeling decision of an instance, the higher the probability its source and target labels are not identical. Assuming a 0-1 loss, $L(h_S(x), h_T(x))$, the relaxed covariate shift assumption is formulated as follows:

For $x \in T, P_S(x) \neq 0$, let $L(h_S(x), h_T(x)) \sim Bin(1, p(x))$, a binomial distribution with a parameter $p(x)$ referring to the probability of success, i.e. the probability that $L(h_S(x), h_T(x)) = 1$ (the probability that the source and target labels of x are not identical). The value of $p(x)$ follows a monotonically decreasing function of $|l_S(x) - 1/k|$, for which we choose the exponential decay function:

$$\begin{aligned} p(x) &= f(|l_S(x) - 1/k|) \qquad (1) \\ &= \mathrm{e}^{-\lambda(|l_S(x)-1/k|)}, \quad |l_S(x) - 1/k| \in [0, \frac{k-1}{k}] \end{aligned}$$

where $|l_S(x) - 1/k|$ signifies the uncertainty manifested in the source labeling decision, since $l_S(x) = 1/k$ is the minimum possible value for $l_S(x)$. Think of the binary classification case as an example, where the closer $l_S(x)$ to $1/k = 0.5$, the more uncertain the source labeling decision. The extremes in the spectrum of this uncertainty are: i), $l_S(x) = 1$ then $|l_S(x) - 1/k| = \frac{k-1}{k}$, which denotes total certainty (minimum uncertainty), and ii) $l_S(x) = 1/k$ leading to $|l_S(x) - 1/k| = 0$, denoting maximum uncertainty. Moving on through this spectrum from case ii) to case i), the value of $p(x)$ keeps diminishing, corresponding to a decline in the probability of having a 1 (rather than 0) loss between the source and target hypotheses for x, $L(h_S(x), h_T(x))$. The relaxed covariate shift is a generalization of the standard covariate shift assumption, since setting the probability of a 1 loss, $L(h_S(x), h_T(x)) = 1$, to be always zero, $p(x) = 0$, for all values of $|l_S(x) - 1/k|$ results in the standard covariate shift definition.

The measure of uncertainty developed in (1) is chosen because it is sound and also because it well captures the relationships between source and target domains in some applications (demonstrated by the experiments). However, other measures of uncertainty, such as the conditional entropy, -$\sum_x \sum_y P_S(x) P_S(y|x) \log P_S(y|x)$, can be used instead. The relationship between the loss, $L(h_S(x), h_T(x))$, and the uncertainty manifested by the source hypothesis, $|l_S(x) - 1/k|$ for $k = 2$, is illustrated in Figure 1 in the supplementary.

After learning parameters of the source hypothesis, the target sample, $x \in T$, is given as input to the source hypothesis resulting in $l_S(x)$ and the corresponding values of $p(x)$ from (1). Afterwards one learns the latent representation, z, that is assumed to generate both $x \in T$, and $p(x)$, where x conveys information about the topology of the target sample, and $p(x)$ conveys information about the corresponding source labels along with probabilistic labeling information according to the relaxed probabilistic covariate shift assumption. Based on the latent representation, a grouping followed by a label alignment technique is pursued to extract the target labels. Extending the proposed form of the covariate shift assumption to other relationships between source and target domains, as well as to other loss functions, is not arduous.

Learning

The graphical model depicting the proposed model is displayed in Figure 2 in the supplementary. For clarity of presentation, the variable x (whose marginals in both domains are $P_S(x)$ and $P_T(x)$), and y (whose conditional distributions in both domains are $P_S(y|x) = \frac{P_S(y,x)}{P_S(x)}$ and $P_T(y|x)$ are displayed twice in a slight abuse of notation.

To learn the parameter set, α, of the source discriminative classifier (hypothesis), the *source* training data, $(x, y) \in S, |S| = n$, is used to learn a discriminative classifier on the source domain (the upper plate of Figure 2 in the supplementary). A probabilistic output for each $y|x$ is obtained via the learnt source discriminative classifier so that $l_S(x)$ can be computed, followed by the computation of corresponding values of $p(x)$ by (1).

Now we have learnt the parameters, α, of the source hypothesis. The intuition behind the target hypothesis is that we have an unlabeled target sample, and we base the learning on two aspects: i) topology of the unlabeled target data instances, and ii) the corresponding source labels of the target sample, which belong to a related (source) domain. The relationship between these two aspects as well as their impact on the sought target label is modelled by the latent variable, z. The latent feature space, z, is a representation that allows for a non-linear transformation of: i) the observed target data, and of ii) the probabilistic labeling information through the corresponding source labels and the relaxed probabilistic covariate shift assumption. The latent feature space, z, represents such components in a more robust and lower-dimensional (than the original space of x and $p(x)$) space, which potentially makes the target data more easily separable into their labels. The latent variable z models the relationship between each $x \in T$ and their corresponding values of $p(x)$ expressing probabilistic information about the 0-1 loss, $L(h_S(x), h_T(x))$.

Instances of the *target* sample, $x \in T$, are given as input to the established source hypothesis (with parameters α) resulting in $l_S(x)$ for each $x \in T$. Based on $l_S(x)$, values of $p(x)$ are indicated by (1). For each $x \in T$, $p(x)$ determines the probability of the target label being different from the corresponding source label, i.e. the probability of $L(h_S(x), h_T(x)) = 1$. There is one z per each $x \in T$. The generative model utilized to learn the variable z corresponding to each data point $x, x \in T$, is as follows (note that θ refers to the generative parameters):

$$P_\theta(z) = \mathcal{N}(0, I)\text{: standard Gaussian } (\mu = 0, \sigma^2 = 1) \quad (2)$$

$$P_\theta(x, p(x)|z) = f(x, p(x); z, \theta) = \mathcal{N}(\mu_\theta(z), \sigma_\theta(z)) \quad (3)$$

The function $f(x, p(x); z, \theta)$ is a non-linear transformation of z, modelled by a neural network. The nonlinear relationship between z and $(x, p(x))$ renders the exact posterior computation intractable. Since the exact posterior, $P_\theta(z|x, p(x))$, is intractable, the variational posterior, $q_\phi(z|x, p(x))$ is computed via which approximate samples of z are used in the subsequent generative process including z and y^U. The latter is a variable that groups the z variables, where each z is a representation of $x, x \in T$, and the corresponding $p(x)$. Regarding the variational posterior, a recognition model (Rezende, Mohamed, and Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013) $q_\phi(z|x, p(x))$ is developed for modelling z. Note that ϕ refers to the variational parameters. More details about the variational inference procedure are given in Section . The Gaussian chosen in (2) is not a limitation of the domain adaptation model; assuming a Gaussian $P_\theta(z)$ facilitates the inference procedure as will be shown in (4).

The grouping operation (with parameters β in Figure 2 in the supplementary is applied to z, resulting in k groups, where each group is expressed by a different value of the grouping variable, y^U, using Expectation-Maximization (EM). By casting the resulting clusters into target labels, the following label alignment method is pursued: i) We construct a confusion matrix whose columns are the source labels of data points $x, x \in T$, (identified via $l_S(x)$), and rows are the clusters, y^U. ii) Out of values of each column, the row (cluster) with the largest number of members is considered the representative cluster of the respective column (class label), i.e. the cluster whose members are the target members of the source label at the column. The intuition behind this method counts on the validity of the relaxed covariate shift assumption, which considers the label change across domains to be unlikely in case $l_S(x)$ is large, i.e. $l_S(x) \to 1$. Its intuition also counts on the general consensus of domain adaptation that there should still be some similarity between source and target domains so that an overall change would not span all data points. Both source and target domains should still be similar enough to allow for the idea of adaptation across domains to be brought in, in the first place. Therefore, assuming that a majority of the data points in each class do not change labels across different domains is quite realistic (and still much more flexible than the assumption that no label change at all can take place across domains). Note that the term "grouping" is performed only w.r.t. z, which already has labeling information from $p(x)$, and in turn from $L(h_S(x), h_T(x))$, and for that it is not grouping (not fully unsupervised) w.r.t. the observed data, $x \in T$. The primary steps of the algorithm, referred to as `GenDA`, are shown in Algorithm 1.

Section 2 of the supplementary explains how some of the other domain adaptation models can be seen as special cases of the proposed model.

Algorithm 1 Generative Domain Adaptation Algorithm (`GenDA`)

```
Input: source $(x, y) \in S$, size $n$ + target $x \in T$, size $m$,
where $(x_i, y_i)$ is a data instance-label pair.
output: target label, $y$, of $x \in T$.
Source learning: Learn $P_S(y|x)$ from $(x, y) \in S$, by
establishing $h_S$; a source hypothesis.
Source labels of $x \in T$:
  - For $x \in T$, learn source labels using $h_S(x)$.
  - Using the resulting $l_S(x)$, compute:
    $p(x) = P(L(h_S(x), h_T(x)) = 1)$ by (1).
- Initialize $\theta$ and $\phi$ (generative & variational param.)
repeat
    Perform minibatch stochastic gradient ascent on
    $\mathcal{L}(x, p(x))$ to learn $\theta$ and $\phi$:
      - $z_i \sim q_\phi(z_i|x_i, p(x_i)), \forall x_i \in B$, $B$ is a random
    minibatch.
      - $\mathcal{L} = \sum_{i=1}^{|B|} \mathcal{L}(x_i, p(x))$.
      - Take derivatives of $\mathcal{L}$ w.r.t. $\theta$ and $\phi$.
      - Update $\theta$ and $\phi$ accordingly
until $\theta, \phi$ do not change
- Learn $P(y^U|z)$ for $x \in T$ by the EM algorithm.
- Learn target labels, $y$ of $x \in T$ from $y^U$ by the label
alignment procedure.
```

Variational Inference

Evidence Lower Bound (ELBO)

Due to the non-linear dependencies between the model components, exact computation of the posterior $P_\theta(z|x, p(x))$ is intractable. A scalable variational inference technique is developed and optimized here. It is based on a variational inference approach recently introduced in Rezende, Mohamed, and Wierstra; Hoffman et al. (2014; 2013), which takes into account producing a high-fidelity as well as computationally efficient variational approach to estimate posteriors. The pursued variational inference approach is referred to as a recognition model (Rezende, Mohamed, and Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013; Kingma and Welling 2014). We derive a lower bound on the marginal likelihood of the proposed generative domain adaptation model and utilize it in establishing the objective function used in computing a high-fidelity variational posterior, $q_\phi(z|x, p(x))$. For a data instance $x \in T$, and its corresponding $p(x)$, the variational bound, $\mathcal{L}(x, p(x))$ is:

$$\begin{aligned}&\log \int_z P_\theta(x, p(x), z)\, dz = \log E_{q_\phi(z|x,p(x))} \frac{P_\theta(z) P_\theta(x, p(x)|z)}{q_\phi(z|x, p(x))} \geq \\ &E_{q_\phi(z|x,p(x))}[\log P_\theta(z) + \log P_\theta(x, p(x)|z) - \log q_\phi(z|x, p(x))] = \\ &E_{q_\phi(z|x,p(x))}[\log P_\theta(x, p(x)|z)] - D_{KL}(q_\phi(z|x, p(x)) \| P_\theta(z)) = \\ &-\mathcal{L}(x, p(x)) \qquad (4)\end{aligned}$$

which yields the objective function used in optimizing the generative and variational parameters, θ and ϕ, respectively. For the recognition model of the latent variable, z, we assume a variational Gaussian $q_\phi(z|x, p(x))$:

$$q_\phi(z|x, p(x)) = \mathcal{N}(\mu_\phi(x, p(x)), \sigma_\phi(x, p(x))) \qquad (5)$$

where $\mu_\phi(x, p(x))$ and $\sigma_\phi(x, p(x))$ are modelled as neural networks. The neural networks used in the generative and variational inference models are multilayer perceptrons (MLPs) with the softplus activation function, $\log(1 + \mathrm{e}^x)$. Two hidden layers are used.

Back to (4), for the optimization of the objective function, we should differentiate $\mathcal{L}(x, p(x))$ w.r.t. generative and variational parameters, θ and ϕ, respectively. An analytical computation of the second term in the third line of (4), $D_{KL}(q_\phi(z|x, p(x)) \| P_\theta(z))$, is possible since both distributions of the KL-divergence, $P_\theta(z)$ and $q_\phi(z|x, p(x))$ (from (2) and (5)) are Gaussians. Regarding the first term of (4), $E_{q_\phi(z|x,p(x))}[\log P_\theta(x, p(x)|z)]$, the workaround for the tricky step of taking gradients w.r.t. ϕ is to reparameterize the latent variable $z = g_\phi(x, p(x), u)$ (Rezende, Mohamed, and Wierstra 2014), where $g_\phi(x, p(x), u)$ is a deterministic function and all the randomness comes through u. Since $z|x, p(x)$ is assumed to be Gaussian, we can apply a location-scale transformation, $z = g_\phi(x, p(x), u) =$

$\mu_\phi(x, p(x)) + \sigma_\phi(x, p(x)) \times u$, $u \in \mathcal{N}(0, I)$. The first term in (4) can then be reformulated as:

$$E_{q_\phi(z|x,p(x))}[\log P_\theta(x, p(x)|z)] = \tag{6}$$
$$E_{u\sim N(0,I)}[\log P_\theta(x, p(x)|\mu_\phi(x, p(x)) + \sigma_\phi(x, p(x)) \times u)]$$

Gradients of (6) are:

$$\nabla_{\theta,\phi}\, E_{q_\phi(z|x,p(x))}[\log P_\theta(x, p(x)|z)] = \tag{7}$$
$$E_{u\sim N(0,I)}[\nabla_{\theta,\phi}(\log P_\theta(x, p(x)|\mu_\phi(x, p(x)) + \sigma_\phi(x, p(x)) \times u))]$$

And they can be computed via Monte Carlo estimates of the expectation (Kingma and Welling 2014). Gradients are computed here by stochastic gradient descent (SGD) (Bottou 2010) and AdaGrad (Duchi, Hazan, and Singer 2010).

A Generalization Bound

Several generalization bounds for domain adaptation have been introduced into the literature. Some of them are based on specific notions of distance between domains like the $\mathcal{A}$-distance defined in Kifer, Ben-David, and Gehrke (2004) and used, most notably, in Ben-David et al. (2007), and like the discrepancy distance-based Rademacher complexity bound introduced in Mansour, Mohri, and Rostamizadeh (2009a). We derive a Rademacher complexity bound based on the introduced relaxed covariate shift assumption.

Define the discrepancy distance between the source and target domains as:

$$\text{dsc-dist}(S, T) = \max_{h,h' \in H} |\mathbb{L}_{P_S}(h, h') - \mathbb{L}_{P_T}(h, h')| \tag{8}$$

Theorem 1. *Let $h_S^\bullet \in H$ be the best hypothesis in H at the source domain, i.e. $h_S^\bullet \in \operatorname{argmin}_{h\in H} \mathbb{L}_{P_S}(h, h_S^{opt})$, where h_S^{opt} is the Bayes Optimal, and similarly define $h_T^\bullet$ over the target domain. For any target hypothesis, $h \in H$:*

$$\mathbb{L}_{P_T}(h, h_T^{opt}) - \mathbb{L}_{P_T}(h_T^\bullet, h_T^{opt}) \leq \tag{9}$$
$$\frac{1 - \mathrm{e}^{-\lambda(\frac{k-1}{k})}[(\lambda(k-1)/k) + 1]}{\lambda^2} + \textit{dsc-dist}(S, T) + \mathbb{L}_{P_S}(h_S^\bullet, h).$$

Proof. The proof of Theorem 1 can be found in Section 5 of the supplementary. □

Also, we can see from Theorem 1 that larger values of λ lead to a tighter generalization bound. As such, Theorem 1 supports the conclusions of (1) and Figure 1 in the supplementary that a larger λ leads to a smaller probability of $L(h_S^\bullet(x), h_T^\bullet(x)) = 1$, which signifies that the source and target domains are more similar to each other with larger λ values than cases with smaller λ:
Larger λ → More similar domains → Tighter bound.

In a way, that can be seen as if the true value of λ between a source and a target domain provides a rough estimate of a distance between the two domains.

Experiments

We evaluate the proposed GenDA algorithm on the Amazon reviews dataset, and on the CVC-02 and CVC-04 pedestrian detection datasets. We start with the note that descriptions of some issues related to the three discriminative learners used at the source domain and to setting the parameters of GenDA, the reverse cross-validation method used, and finally example images related to the CVC data, are all left, due to space limitation, to Section 6 in the supplementary.

Sentiment Analysis

The dataset used in this experiment is the Amazon reviews dataset (Blitzer, Dredze, and Pereira 2007), which represents one of the benchmark domain adaptation datasets. Its intact version contains more than 340,000 reviews describing 22 product types. Since its original form is immensely unbalanced, the version of the Amazon reviews data used in domain adaptation, which was first introduced by Blitzer, Dredze, and Pereira (2007), consists of 4 product types where each product types refers to a domain. The four domains are: i) books (BK), ii) DVDs (DV), iii) electronics (EL), and iv) kitchen appliances (KA). Each Amazon review originally had a rating from 0 to 5 stars. For comparability, the convention followed in domain adaptation is to change the review ratings into either positive (higher than 3 stars) or negative (less than or equal to 3 stars) (Chen et al. 2012; Glorot, Bordes, and Bengio 2011). Features are pre-processed with standard tf-idf (Salton and Buckley 1988). The selected feature set is composed of 5,000 features of unigrams and bigrams. All these settings are done in order to compare on common grounds with state-of-the-art.

There are 2,000 labeled reviews per each of the 4 domains. Number of unlabeled reviews ranges between 3,586 KA reviews and 5,945 DV reviews. The two classes of the data are balanced in each of the 4 domains, i.e. ratio of positive ratings is 50%. We perform 12 adaptation tasks, moving from one domain to another, e.g. EL → BK. Training on the source domain is performed on 2,000 labeled source data instances (reviews), i.e. the sample S. The target learner is given a set of 2,000 unlabeled target instances. The unlabeled target learning set is used as the sample, T, that is given as input to the source learner so that corresponding values of $p(x)$ can be computed according to (1), and then the non-linear transformation between the unlabeled sample, T along with $p(x)$ and the latent feature space, z, is learnt. Afterwards, Expectation-Maximization (EM) followed by the label alignment method are used to learn the target labels of the sample T, as illustrated in Section .

Results of the experiments performed on the 12 Amazon reviews adaptation tasks are displayed in Table 1. Comparisons with state-of-the-art results on the Amazon reviews data by Ganin et al. (2015) are based on classification accuracy. For each adaptation task, we compare to the best classification accuracy obtained by Ganin et al. (2015). A paired t-test with $p = 0.05$ is used to identify significance. The proposed algorithm GenDA outperforms the best result by Ganin et al. (2015) in 10 out of the 12 adaptation tasks (significantly so in 8 tasks). Concerning the source classifiers, the SVM classifier is clearly superior to the other two classifiers, and the logistic regression (LR) comes second in most of the tasks.

Pedestrian Detection

The task of pedestrian detection (Dollar et al. 2012; Enzweiler and Gavrila 2009) in images is challenging both due to diffi-

Table 1: Classification Accuracy of GenDA using 3 different source classifiers vs. state-of-the-art by taking the best available classification accuracy from Ganin et al. (2015), on the Amazon reviews dataset. Bold denotes the best result significantly outperforms all the competitors for such task.

Source-Target	**SVM GenDA**	**NN GenDA**	**LR GenDA**	**Ganin**
BK-DV	**81.2%**	72.4%	78.8%	79.9%
BK-EL	77.1%	75.1%	75.9%	**79.2%**
BK-KA	**84%**	74.7%	81.5%	81.6%
DV-BK	**77.1%**	70.2%	75.4%	75.5%
DV-EL	**78.9%**	73.9%	77.8%	78.6%
DV-KA	**85.1%**	71.4%	84.5%	82.2%
EL-BK	**74.5%**	65.5%	69.7%	72.7%
EL-DV	**77.3%**	76.1%	71.2%	76.5%
EL-KA	80.4%	78.1%	81%	**85.4%**
KA-BK	**81%**	71.4%	74.1%	72%
KA-DV	**74.4%**	70.1%	73.9%	74%
KA-EL	**85.1%**	79%	83.7%	84.3%

culties inherent in the task itself, like diversities in scene contents, poses, occlusions, etc, and due to the high risk arising from some related applications like driver assistance systems. For the core subtask of pedestrian classification (primarily used to decide whether or not a given image window contains a pedestrian), we apply our unsupervised domain adaptation algorithm, `GenDA`, to two pedestrian detection datasets, referred to as the CVC-02 and CVC-04 datasets. Input to a pedestrian detection task consists of images manually labeled where bounding boxes provide information about the location of pedestrians in pedestrian images, i.e. images containing pedestrians. The manual intervention required for such task is very exhaustive (Vazquez, Lopez, and Ponsa 2012). By having a virtual input, where pedestrian and pedestrian-free cropped images[1] come from virtual images, e.g. video games, unsupervised domain adaptation can be used by considering the labeled virtual images to be the source domain and by considering unlabeled real images to be the target domain. More details about the settings of this experiment is given in Section 7 of the supplementary.

As such, the pedestrian classifier is a binary classifier with labels: Pedestrian images, and pedestrian-free images. An image window is labeled as a pedestrian cropped image if the classification score is larger than a threshold, and is labeled as a pedestrian-free cropped image otherwise. Value of the threshold as well as parameters of the pyramidal sliding window (8×8 pixels), are set equivalent to their values in Vazquez, Lopez, and Ponsa (2012). The principal task of `GenDA` is to classify cropped target images into either a pedestrian or a pedestrian-free cropped image.

As source labeled data, we use the CVC-04 virtual-world pedestrian dataset (Vazquez et al. 2014), which consists of 1208 virtual pedestrians and 1220 (a subset of the available 6828) pedestrian-free cropped images. As target unlabeled data, where the labels are used only to assess accuracy at the end but never in learning, we use the CVC-02 real-world dataset (Geronimo et al. 2010). The dataset, CVC-02 was recorded using a camera based on 640×480 pixels resolution (Vazquez, Lopez, and Ponsa 2013). We use a set of 600 CVC-02 cropped images, evenly divided between the two classes, as the unlabeled target sample, T.

We compare our results to state-of-the-art unsupervised domain adaptation results on CVC datasets represented by Vazquez, Lopez, and Ponsa; Vazquez, Lopez, and Ponsa (2012; 2013), which are based on transductive SVM. We compare results based on the same metric used in Vazquez, Lopez, and Ponsa; Vazquez, Lopez, and Ponsa (2012; 2013), which is detection error tradeoff (DET) curve (Martin et al. 1997) showing the FPPI (false positive per image) rate vs. false negative (or missed detections) rate, for different thresholds. The smaller the area under the curve, the more accurate the algorithm. We implemented the algorithm in Vazquez, Lopez, and Ponsa (2012) with the help of coding excerpts taken from SVM$^{\text{light}}$ (Joachims 1999a; 1999b). In Figure 1, UDA refers to the unsupervised domain adaptation algorithm by Vazquez, Lopez, and Ponsa; Vazquez, Lopez, and Ponsa (2012; 2013). Results displayed in Figure 1 demonstrate an improvement induced by `GenDA` over UDA, which we believe is mainly due to two reasons: Flexibility of the introduced relaxed covariate shift assumption, and the power of the latent feature representation, Z. Note that we stick to the DET curve, shown in Figure 1, as a metric so that we can compare on common grounds with state-of-the-art. Two examples from the CVC-02 dataset are displayed in Section 7 of the supplementary.

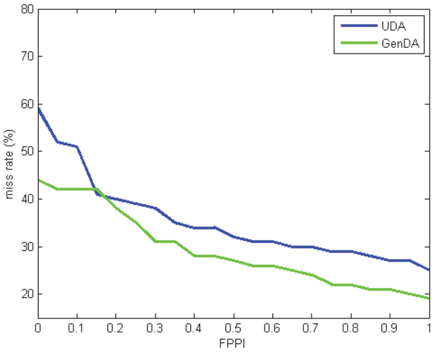

Figure 1: A DET curve showing FPPI vs. missed detections rate resulting from applying `GenDA` vs. its competitor UDA (Vazquez, Lopez, and Ponsa 2012; 2013) to 2 CVC pedestrian detection datasets. FPPI rate refers to false positive per image rate. The smaller the area under the curve, the more accurate the algorithm. Average area under the curve is 33.2 for UDA, and is 28.6 for `GenDA`.

Conclusion

We introduced a probabilistic relaxation of the covariate shift assumption, and utilized it as a part of a proposed unsupervised domain adaptation model. On the one hand, the assumption is flexible enough to depict several forms of relationships between the source and target domains. On the other hand, the proposed model is efficient due to the power of the induced latent representation and due to its inference procedure. The proposed domain adaptation model can also be used to model other domain adaptation assumptions, and that represents an imminent direction for future work.

[1] we use the term "cropped image" to refer to a part of an image on which pedestrian detection has already been applied. It is the same term used in the CVC data descriptions.

References

Adel, T., and Wong, A. 2015. A probabilistic covariate shift assumption for domain adaptation. *AAAI*.

Ben-David, S., and Urner, R. 2012. On the hardness of domain adaptation and the utility of unlabeled target samples. *ALT* 139–153.

Ben-David, S., and Urner, R. 2014. Domain adaptation – can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence* 70:185–202.

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

Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Wortman, J. 2008. Learning bounds for domain adaptation. *NIPS*.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Association for Computational Linguistics (ACL)*.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. *COMPSTAT* 177–186.

Chen, M.; Xu, Z.; Weinberger, K.; and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. *ICML*.

Dollar, P.; Wojek, C.; Schiele, B.; and Perona, P. 2012. Pedestrian detection: An evaluation of the state of the art. *TPAMI* 34:743–761.

Duchi, J.; Hazan, E.; and Singer, Y. 2010. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR* 12:2121–2159.

Enzweiler, M., and Gavrila, D. 2009. Monocular pedestrian detection: Survey and experiments. *TPAMI* 31:2179–2195.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2015. Domain-adversarial training of neural networks. *arXiv* (1505.07818).

Geronimo, D.; Sappa, A.; Ponsa, D.; and Lopez, A. 2010. 2d-3d based on-board pedestrian detection system. *Computer Vision and Image Understanding (Special Issue on Intelligent Vision Systems)*.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. *ICML*.

Hoffman, M.; Blei, D.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *JMLR* 14:1303–1347.

Huang, J.; Gretton, A.; Borgwardt, K.; Schoelkopf, B.; and Smola, A. 2006. Correcting sample selection bias by unlabeled data. *NIPS*.

Joachims, T. 1999a. *Making large-scale SVM learning practical*. MIT Press.

Joachims, T. 1999b. Transductive inference for text classification using support vector machines. *ICML*.

Kifer, D.; Ben-David, S.; and Gehrke, J. 2004. Detecting change in data streams. *VLDB* 180–191.

Kingma, D., and Welling, M. 2014. Auto-encoding variational Bayes. *ICLR*.

Kingma, D.; Rezende, D.; Mohamed, S.; and Welling, M. 2014. Semi-supervised learning with deep generative models. *NIPS* 28:3581–3589.

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

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009b. Learning from multiple sources. *NIPS* 23:1041–1048.

Martin, A.; Doddington, G.; Kamm, T.; Ordowski, M.; and Przybocki, M. 1997. The DET curve in assessment of detection task performance. *National Inst. of Standards and Technology Gathersburg*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks.

Rezende, D.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. *ICML*.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24(5):513–523.

Storkey, A., and Sugiyama, M. 2006. Mixture regression for covariate shift. *NIPS*.

Stuhlmuller, A.; Taylor, J.; and Goodman, N. 2013. Learning stochastic inverses. *NIPS* 27:3048–3056.

Sugiyama, M., and Mueller, K. 2005. Generalization error estimation under covariate shift. *Workshop on Information-Based Induction Sciences*.

Vazquez, D.; Lopez, A.; Marin, J.; Ponsa, D.; and Gomez, D. 2014. Virtual and real world adaptation for pedestrian detection. *TPAMI* 36(4):797–809.

Vazquez, D.; Lopez, A.; and Ponsa, D. 2012. Unsupervised domain adaptation of virtual and real worlds for pedestrian detection. *ICPR*.

Vazquez, D.; Lopez, A.; and Ponsa, D. 2013. *Domain adaptation of virtual and real worlds for pedestrian detection*. Ph.D. Dissertation, Universitat autonoma de Barceolona.