ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200257

Improving Unsupervised Domain Adaptation with Variational Information Bottleneck

Yuxuan Song[1], Lantao Yu[2], Zhangjie Cao[2], Zhiming Zhou[1]
Jian Shen[1], Shuo Shao[1], Weinan Zhang[1], Yong Yu[1]

Abstract. Domain adaptation aims to leverage the supervision signal of source domain to obtain an accurate model for target domain, where the labels are not available. To leverage and adapt the label information from source domain, most existing methods employ a feature extracting function and match the marginal distributions of source and target domains in a shared feature space. In this paper, from the perspective of information theory, we show that representation matching is actually an *insufficient* constraint on the feature space for obtaining a model with good generalization performance in target domain. We then propose variational bottleneck domain adaptation (VBDA), a new domain adaptation method which improves feature transferability by explicitly enforcing the feature extractor to ignore the task-irrelevant factors and focus on the information that is essential to the task of interest for both source and target domains. Extensive experimental results demonstrate that VBDA significantly outperforms state-of-the-art methods across three domain adaptation benchmark datasets.

1 Introduction

Deep learning has shown impressive abilities on solving numerous machine learning tasks. Most of recent advances heavily rely on the access to huge amount of labeled data and the assumption that both training and test data are sampled from the same underlying distribution. However, there are many application scenarios where the labeled data for the task of interest (target domain) is hard to obtain, while another correlated domain (source domain) with non-negligible dissimilarity consists of sufficient annotated data. Hence, there is strong motivation to leverage the supervision signal from source domain to help build an effective model in target domain. Learning an accurate predictive model for target domain with the presence of covariate shift [34] (*i.e.*, the input data distributions of source and target domains are different) is known as domain adaptation. In this paper, we focus on a general and challenging setting where no label information is available in target domain, which is termed as unsupervised domain adaption.

Recent advances in deep learning stimulate a fruitful line of domain adaptation works, which leverage deep neural networks to infer the latent variables and match the marginal distributions of source and target domains in the latent space [7, 16, 18]. Inspired by Generative Adversarial Networks [9], an adversarial domain adaptation mechanism is utilized in [7, 37, 20, 39]. This mechanism involves a two-player game between a discriminator and a feature extractor: the domain discriminator is trained to tell whether the samples come from source or target domain, while the feature extractor is trained to maximize the discriminator's classification error. Essentially, adversarial domain adaptation methods seek to minimize the Jensen-Shannon divergence between source and target distribution of latent features.

However, it has been shown that matching the marginal distribution in latent feature space is not strong enough for ensuring the essential information to be transferred [40]. It is possible that the learned mapping is misled by the domain invariant yet task-irrelevant factors and fails to capture the semantic information. Consider the following example, the adaptation task is to recognize animals in the pictures, and in source domain, most of the sheep are appearing with the grassland as the background and most of the horses are appearing with the animal house as the background; while in target domain the background configuration of the two species are random and marginal distributions of the background are the same. In cases like this, directly matching the marginal feature distributions could result in that the domain-invariant yet task-irrelevant information, (*i.e.* the background in the above example), outweigh the task-relevant information, which then lead to worse performance on target domain (also known as negative transfer [23]).

To tackle the lack of semantic alignment, many recent works proposed to enhance the label information of target domain based on some strong assumptions [20, 32, 39, 28]. One of the most widely used hypotheses is the cluster assumption [10] (also known as low density separation assumption), which states that the data instances are distributed into several separate clusters and samples in the same cluster share the same label. However, the cluster assumption is actually too strong and inappropriate for many practical scenarios, and directly using the cluster assumption could bring non-negligible undesired effects [32].

In this paper, inspired by the information bottleneck principle, we propose a simple yet effective regularization technique for domain adaptation methods by combining conditional entropy minimization and variational information bottleneck, which enforces the feature extractor to ignore the irrelevant factors and focus on the essential information for the task of interest (*i.e.*, the sufficient statistics for determining the parameters of the predictive models). Our method tends to learn a balanced and clean representation space (*i.e.*, no information preference on source or target domain and less irrelevant factors), which improves the generalization ability of the predictive model and renders strong yet widely used assumptions such as the cluster assumption more realistic. We further provide a theoretical analysis on the generalization error bound in Section 4.3. Extensive experimental results demonstrate that our model outperforms state-of-the-art methods across three domain adaptation benchmark

[1] Shanghai Jiao Tong University, email: songyuxuan@apex.sjtu.edu.cn
[2] Stanford University, email: lantaoyu@cs.stanford.edu

datasets [38, 27, 22].

2 Related Works

In this section, we discuss several most relevant works in the field of domain adaptation. [7] and [16] proposed to project the source and target domain into a common representation space, and encouraged the corresponding marginal feature distribution to be matched under the guidance of some distance or divergence. Adversarial techniques based on the framework of GAN [9] are widely explored in the literatures of domain adaptation [29, 12, 37], which corresponds to minimizing the symmetric Jensen-Shannon divergence. However, [19] pointed out that adversarial domain adaptation methods which only match the marginal distribution are problematic and insufficient for successful adaptation. To address this limitation, various methods have been proposed. For example, [19, 17] proposed to match the joint distribution instead of purely matching the marginal; [8] introduced a decoder architecture for capturing the semantic information; and [12] utilized cycle consistency constraints to preserve semantic information. However, the main limitation of these methods is that, although the semantic information is enhanced, the learned representation is still likely to preserve domain-invariant factors that are *irrelevant* to the predictive task, which may mislead the semantic alignment especially when training samples are not sufficient enough. In the animal recognition example mentioned in Introduction section, the background is the domain-invariant yet *irrelevant* factors. The learned representation tended to preserve the background information due to the fact that the background has statistically dependency with the class label in source domain and the marginal distribution of background is invariant between the source and target domain. And irrelevant information will disturb the predictive task on target domain. Hence there is strong motivation to enforce the feature extractor to only focus on the essential information for the task of interest and ignore as much irrelevant factors as possible, no matter they are domain-invariant or not. Inspired by this intuition, we propose to regularize domain adaptation models with information bottleneck principle [35], which seeks to find the optimal tradeoff between representation accuracy and compression. Since information bottleneck method has been successfully applied to supervised learning [1], generative modeling [13, 25] and reinforcement learning [25], in the context of domain adaptation, we propose to exploit it to preserve sufficient statistics and remove irrelevant factors in the learned representations. While [21] also augment domain adaptation with information bottleneck, they focus on a specific scenario, where an auxiliary data view (*e.g.*, skeleton data for gestures and bounding box for objects) is available and the information bottleneck is incoporated to leverage these additional data view. In contrast, our method seeks to provide a new regularization technique for general unsupervised domain adaptation with deep neural networks.

On the other hand, to counter the lack of attention on target semantic information, conditional entropy minimization [10] is widely used in unsupervised domain adaptation [18, 20]. These methods are based on the cluster assumption that, the decision boundary should not cross high density regions, but instead lie in low density regions [3]. In other words, it assumes that the data instances are distributed into several separate clusters, and samples in the same cluster share the same class label. However, it should be noted that the cluster assumption can be too strong to be satisfied in many practical scenarios, which will bring undesired effects to the stability of training and performance of the models. Essentially, the cluster assumption in the representation space is satisfied only when the learned representations merely preserve semantic information that is relevant to the predictive task, while our variational bottleneck domain adaptation framework intrinsically seeks to find such a clean representation space which renders the cluster assumption more realistic and achieves better feature transferability.

3 Background & Notations

3.1 Domain Adaptation

To describe a domain, we introduce a joint data distribution $p(x, y)$ with which we define both the marginals and conditionals. Let $p_s(x_s, y_s)$ denote the underlying joint data distribution of the data instance x_s and the corresponding label y_s for source domain, and let $p_s(x_s)$ denote the marginal distribution of x_s. $p_t(x_t, y_t)$ and $p_t(x_t)$ are defined analogously for target domain. In feature-based unsupervised domain adaptation, our objective is to train a classifier $f_{\phi,\theta} = h_\phi \circ g_\theta$ which can perform well on target domain. Specifically, $g_\theta : \mathcal{X} \to \mathcal{Z}$ is the feature extractor, which is a projection function from data space $\mathcal{X}$ to latent feature space $\mathcal{Z}$, and $h_\phi : \mathcal{Z} \to \mathcal{P}(\mathcal{Y})$ is a classification function on the representation space, where $\mathcal{P}(\mathcal{Y})$ denotes the set of probability distributions over the label set $\mathcal{Y}$. To address the covariate shift problem, many domain adaption methods are proposed to minimize the following objective motivated by the theory in [2, 7]:

$$\mathbb{E}_{(x_s,y_s)\sim p_s} \mathcal{L}_c(f_{\phi,\theta}(x_s), y_s) + \lambda d(q_s(z_s;\theta), q_t(z_t;\theta)) \quad (1)$$

Here z_s and z_t are latent representations for source and target domain; q_s and q_t are the marginal distributions of z_s and z_t, which is implicitly defined by the marginals $p_s(x_s), p_t(x_t)$ and the deterministic mapping g_θ; $\mathcal{L}_c$ is the cross entropy loss for training a classifier; $d(\cdot,\cdot)$ is some divergence or distance measure between two distributions and λ is the weighting factor. For instance, in [16], the divergence is realized as maximum mean discrepancy (MMD) and in many adversarial domain adaptation methods [12, 7, 20], the Jensen-Shannon divergence between q_s and q_t is minimized within an adversarial learning framework [9]:

$$\min_\theta \max_\omega \mathbb{E}_{x_s\sim p_s} \log D_\omega(g_\theta(x_s)) + \mathbb{E}_{x_t\sim p_t} \log(1 - D_\omega(g_\theta(x_t)))$$

where D_ω is a domain classifier on the representation space. Intuitively, the domain classifier is trained to distinguish the latent representations of source domain from that of target domain, while the feature extractor is jointly trained to confuse the discriminator by maximizing its classification error. At optimality, the marginal distributions of latent representations will be matched and the learned representations will be domain-invariant.

3.2 Information Bottleneck Principle

Let random variable X denote the original signal and random variable Y denote an output variable (*e.g.* desired label), whose information we want to preserve. Given their joint distribution $p(X, Y)$, assuming the statistical dependence between Y and X, the mutual information $I(X;Y)$ measures the mutual dependence between these two random variables. In this case, Y implicitly determines both the relevant and irrelevant features in X. The information bottleneck(IB) method seeks to find an optimal representation of X which captures the relevant part and filters out the irrelevant part.

Formally, in the context of information bottleneck, we are interested in finding the relevant part of X with respect to Y, denoted by Z, the *minimal sufficient statistics* of X with respect to Y. Thus we assume the following *Markov chain*: $Y \to X \to Z$ and we can

obtain the optimal representation by minimizing $I(X;Z)$ under a constraint on $I(Y;Z)$ (to ensure the predictive ability of Z).

The objective of finding the optimal representation can be further formulated as the maximization of the following Lagrangian [36]:

$$L(p(z|x)) = I(Y;Z) - \beta I(X;Z) \tag{2}$$

subject to the Markov chain constraint. Here the positive Lagrangian multiplier β represents a tradeoff between the complexity of the representation ($I(X;Z)$) and the amount of preserved relevant information ($I(Y;Z)$). In essence, information bottleneck principle explicitly enforces the learned representation Z to only preserve the information in X that is useful to the prediction of Y, *i.e.*, the *minimal sufficient statistics* of X with respect to Y.

In this paper, under the framework of information bottleneck principle, we propose a novel domain adaptation method which enforces the feature extractor to focus on the relevant factors implicitly defined by the task, and provide a thorough analysis of the benefits brought by our method both emipirically and theoretically.

4 Method

4.1 Motivations

[7] claimed that a successful adaptation can be achieved when the source domain classification error and the domain confusion loss are both small, which can be realized through optimizing the objective in Equation (1).

From the perspective of information preference, we can reformulate the objective in Equation (1) and understand the weakness of the constraint in a more straightforward way. To begin with, we split the loss function in Equation (1) into two terms, $\mathbb{E}_{(x_s,y_s)\sim p_s}\mathcal{L}_c(f_{\phi,\theta}(x_s), y_s)$ and $\lambda d(q_s(z_s;\theta), q_t(z_t;\theta))$. In the following, we will show that minimizing the first term is equivalent to maximizing a variational lower bound of the mutual information between learned representations and the labels in source domain (*i.e.*, $I(Y_s;Z_s)$), and minimizing the second term corresponds to finding the domain-invariant features. To see these, let us first rewrite the negative of cross entropy loss as:

$$\begin{aligned}
&-\mathbb{E}_{(x_s,y_s)\sim p_s}\mathcal{L}_c(f_{\phi,\theta}(x_s), y_s) \\
=&\mathbb{E}_{(x_s,y_s)\sim p_s}\left[\int p_\theta(z_s|x_s)\log h_\phi(y_s|z_s)dz_s\right] \\
=&\int p_s(x_s,y_s)p_\theta(z_s|x_s)\log h_\phi(y_s|z_s)dx_s\,dy_s\,dz_s
\end{aligned} \tag{3}$$

where $p_\theta(z_s|x_s)$ denotes the conditional distribution implied by the projection function g_θ (when g_θ is a deterministic projection, $p_\theta(z_s|x_s)$ corresponds to a delta distribution with non-zero density at $z = g_\theta(x_s)$). With the Markov chain assumption introduced in the Information Bottleneck Principle section , Equation (3) can be rewritten as:

$$\begin{aligned}
&\int p_s(x_s,y_s)p_\theta(z_s|x_s,y_s)\log h_\phi(y_s|z_s)dx_s\,dy_s\,dz_s \\
=&\int p_\theta(x_s,y_s,z_s)\log h_\phi(y_s|z_s)dx_s\,dy_s\,dz_s \\
=&\int p_\theta(y_s,z_s)\log h_\phi(y_s|z_s)dy_s\,dz_s \\
\leq&\int p_\theta(y_s,z_s)\log p_\theta(y_s|z_s)dy_s\,dz_s = I(Y_s;Z_s) - H(Y_s)
\end{aligned}$$

Here, the inequality holds for the fact that $D_{\mathrm{KL}}(p_\theta(y|z)\|h_\phi(y|z)) \geq 0$. Since $H(Y)$ is a constant in our optimization procedure of θ and ϕ, we know that minimizing the first term in Equation (1) corresponds to maximizing a lower bound of $I(Y_s;Z_s)$.

The second term $\lambda d(q_s(z_s;\theta), q_t(z_t;\theta))$ accounts for matching the marginal distribution of latent variables under the guidance of some distance or divergence. One notable example is the optimization of Jensen-Shannon divergence with adversarial training. Essentially, this constraint seeks to find the domain-invariant features of X. However, it should be noted that matching the marginals of latent features is *agnostic* to the task of interest, which implies that the preserved domain-invariant features is likely to contain factors that are irrelevant to the prediction of desired labels. From learning theory [33] , we know that when the sample size is finite, the irrelevant factors (for the predictive task) in the noisy inputs can decrease the generalization ability of the models. We provide a formal discussion about the generalization error bound in Theoretical Analysis section. From this perspective, we know that one direction to improve domain adaptation models is to add more constraints on the representation space so that the preserved features will not only be domain-invariant, but also relevant to the task of interest.

On the other hand, due to the supervised learning objective in source domain, the learned representation with Equation (1) will intrinsically tend to capture the relationship between data instances and labels from source domain, while taking less attention on target domain. To take the label information for target domain where the exact label is not available into account during the feature learning, the cluster assumption can be adapted [18, 3], where the input distribution is assumed to contain separated data clusters and that data samples in the same cluster share the same class label. Cluster assumption introduces an inductive bias where we are seeking decision boundaries that do not go through high-density regions, which can be implemented through the following conditional entropy minimization:

$$\mathcal{L}_{ce} = \mathbb{E}_{x_t\sim p_t(x_t), z_t\sim p_\theta(z_t|x_t)}\left[\int h_\phi(y_t|z_t)\log h_\phi(y_t|z_t)dy_t\right] \tag{4}$$

Note that the cluster assumption is satisfied when the learned representations only preserve semantic information that is relevant to the predictive task, it is strongly motivated to find a clean representation space with information bottleneck to justify the use of strong assumptions in domain adaptation methods.

4.2 Variational Bottleneck Domain Adaption

Inspired by conditional entropy minimization in semi-supervised learning [3, 10] and deep variational information bottleneck [35, 1], to achieve better generalization ability, we propose a new regularization mechanism for domain adaptation, which explicitly enforces the feature extractor to only preserve the minimal sufficient statistics of the input data with respect to the labels for both source and target domain.

As discussed in the Motivations section, from the perspective of information bottleneck principle, we know that the objective in Equation (1) lacks a constraint for minimizing the mutual information between X and Z:

$$I_\theta(X;Z) = \int p_\theta(x,z)\log\frac{p_\theta(z|x)}{p_\theta(z)}dx\,dz \tag{5}$$

However, it should be noted that in general, directly computing and optimizing $I(X;Z)$ is computationally intractable[1], as it

requires solving an integral over latent feature space. To achieve tractability, we follow the methods proposed in [1] and instead optimize a tractable variational upper bound:

$$\begin{aligned} I_\theta(X;Z) &= \int p(x)p_\theta(z|x)\log\frac{p_\theta(z|x)}{p_\theta(z)}dx\,dz \\ &= \mathbb{E}_{x\sim p(x)}D_{\mathrm{KL}}(p_\theta(z|x)\|r(z)) - D_{\mathrm{KL}}(p_\theta(z)\|r(z)) \\ &\le \mathbb{E}_{x\sim p(x)}D_{\mathrm{KL}}(p_\theta(z|x)\|r(z)) \\ &\triangleq I_U(X;Z). \end{aligned} \tag{6}$$

Here, $r(z)$ is the prior distribution of latent features and $p_\theta(z)$ denotes the marginal distribution implied by $p(x)$ and conditional distribution $p(z|x)$, and the inequality holds for the fact that $D_{\mathrm{KL}}(p_\theta(z)\|r(z)) \ge 0$.

To incorporate the above variational information bottleneck, with abuse of notation, we introduce a stochastic feature extracting function $g_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$, which maps a sample x to a stochastic representation $z \sim g_\theta(z|x)$. Now we can add the following terms to the objective in order to enforce the feature extractor to only preserve task-relevant factors:

$$\mathbb{E}_{x_s\sim p_s}D_{\mathrm{KL}}(g_\theta(z|x_s)\|r(z)) + \mathbb{E}_{x\sim p_t}D_{\mathrm{KL}}(g_\theta(z|x_t)\|r(z))$$

In our experiments, the stochastic feature extracting function is realized as a Gaussian distribution $g_\theta(z|x) = \mathcal{N}(z|g_\theta^\mu(x), g_\theta^\Sigma(x))$, where $g_\theta(x)$ outputs the mean μ and diagonal covariance matrix Σ of z. When $r(z)$ allows for the computation of Kullback-Leibler divergence analytically, the upper bound in Equation (6) can be easily optimized. Thus we choose $r(z)$ to be a standard normal distribution, $r(z) = \mathcal{N}(0, I)$. Note that although the objective here shares similar mathematical form with the KL regularization term in Variational Autoencoder (VAE) [14], the motivation and interpretation of the objectives are related but different. As a generative model, VAE consists of a pre-determined prior $p(z)$ for the latent variables and a stochastic decoder $p(x|z)$ for reconstruction. The amortized encoder $q(z|x)$ is introduced as a variational approximation to the true posterior $p(z|x) = p(x|z)p(z)/p(x)$ and the resulting evidence lower bound (ELBO) works as a tractable lower bound for the log-likelihood objective. While in the variational information bottleneck, the $r(z)$ is introduced to derive a tractable upper bound for minimizing the mutual information term. Note that the equality in Equation (6) holds only when $p_\theta(z) = r(z)$. Therefore, by choosing a simple realization of $r(z)$ such as standard normal distribution, we are also introducing an inductive bias of regularizing the marginal distribution of the learned representations (*i.e.*, $p_\theta(z)$) to be as simple as possible.

Putting things together, the final objective function in our framework can be written as:

$$\begin{aligned} \mathcal{L}(\theta,\phi) = \; & \mathbb{E}_{(x_s,y_s)\sim p_s, z_s\sim g_\theta(z|x_s)}\mathcal{L}_c(h_\phi(z_s), y_s) + \\ & \lambda_d \cdot d(q_s(z_s;\theta), q_t(z_t;\theta)) + \lambda_{ce}\cdot\mathcal{L}_{ce} + \\ & \lambda_s \cdot \mathbb{E}_{x_s\sim p_s}D_{\mathrm{KL}}(g_\theta(z|x_s)\|r(z)) + \\ & \lambda_t \cdot \mathbb{E}_{x_t\sim p_t}D_{\mathrm{KL}}(g_\theta(z|x_t)\|r(z)) \end{aligned} \tag{7}$$

Here, $\mathcal{L}_c$ is the classification loss; $q_s(z_s)$ and $q_t(z_t)$ are implicit marginal distributions induced by the marginal distributions $p_s(x_s), p_t(x_t)$ and the conditional distribution $g_\theta(z|x)$; $\mathcal{L}_{ce}$ is the conditional entropy term defined in Equation (4); $\lambda_d, \lambda_{ce}, \lambda_s$ and λ_t are hyperparameters controlling the optimization tradeoff among each term. Note that there is a stochastic structure in the model, we utilize the reparameterization trick introduced in [14] to back-propagate unbiased estimated gradients through single example.

4.3 Theoretical Analysis

In this section, we analyze the theoretical properties of our proposed method.

Theorem 1 ([2]) *Let $\mathcal{H}$ be the hypothesis space, Given (X_s, ϵ_s) and (X_t, ϵ_t) as the two domains and their corresponding test error functions. Then for any $h \in \mathcal{H}$, we have:*

$$\epsilon_t(h) \le \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(X_s, X_t) + \epsilon_s(h) + \min_{h'\in\mathcal{H}}\epsilon_t(h') + \epsilon_s(h')$$

Here $d_{\mathcal{H}\Delta\mathcal{H}}$ represents a discrepancy measure between source and target domain with respect to a hypothesis space $\mathcal{H}$, which is defined as:

$$\begin{aligned} & d_{\mathcal{H}\Delta\mathcal{H}}(X_s, X_t) = \\ & 2\sup_{h,h'\in\mathcal{H}}\|\mathbb{E}_{x\sim X_s}\left[h(x)\neq h'(x)\right] - \mathbb{E}_{x\sim X_t}\left[h(x)\neq h'(x)\right]\|. \end{aligned} \tag{8}$$

For a fixed hypothesis space $\mathcal{H}$, $d_{\mathcal{H}\Delta\mathcal{H}}(X_s, X_t)$ is the intrinsic difference between source and target domain, which is fixed and determined by the characteristics of the data distributions. Now we will show that how the $I(X;Z)$ term from information bottleneck principle can help minimize the test error term, *i.e.* $\epsilon_s(h)$ and $\min_{h'\in\mathcal{H}}\epsilon_t(h') + \epsilon_s(h')$ in Theorem 1.

Theorem 2 ([31]) *For any probability distribution $p(x, y)$, with a probability of at least $1-\delta$ over the draw of the sample of size m from $p(x, y)$, $\hat{I}(X;Z)$ and $\hat{I}(Y;Z)$ are the empirical estimate of the mutual information $I(X;Z)$ and $I(Y;Z)$. Then for any Z,*

$$\begin{aligned} & |I(Y;Z) - \hat{I}(Y;Z)| \le \\ & \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)}{m}}(C_1\log(m)\sqrt{|\mathcal{Z}I(X;Z)|} \\ & + C_2|\mathcal{Z}|^{3/4}(I(X;Z))^{1/4} + C_3\hat{I}(X;Z)) \end{aligned} \tag{9}$$

where C, C_1, C_2 and C_3 are constants. $|\mathcal{Z}|$ and $|\mathcal{Y}|$ correspond to the cardinality of variables Z and Y.

Theorem. 2 shows that the $|I(Y;Z) - \hat{I}(Y;Z)|$ which is a measure of difference between training and test error is bounded by a monotonic function of $I(X;Z)$. Essentially, it is true that minimizing $I(X;Z)$ will minimize the generalization error, but this is not enough. A degenerate case is $I(X;Z) = 0$, in which case the prediction is random, although the difference between training and test error is zero. So we also need to make sure both the training error and the generalization error is small. We can decrease $\epsilon_s(h)$ with information bottleneck (IB) principle, since we are explicitly minimizing the training error in source domain and the generalization error in both domains. For $\min_{h'\in\mathcal{H}}\epsilon_t(h') + \epsilon_s(h')$, ideally IB will not harm predictive ability by just removing irrelevant factors, so the combined training error $\min_{h'\in\mathcal{H}}\hat{\epsilon_t}(h') + \hat{\epsilon_s}(h')$ should be the same with or without IB. While the combined test error is the sum of combined training error and combined generalization error, we are also able to reduce the combined test error.

5 Experiments

We conduct experiments on various visual domain adaptation benchmarks including **Office-31**, **Office-home** and **Digits**, to compare our approach against state-of-the-art deep domain adaptation methods.

Table 1. Classification accuracy (%) on Office-31 (ResNet50)

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50 [11]	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
RTN [18]	84.5±0.2	96.8±0.1	99.4±0.1	77.5±0.3	66.2±0.2	64.8±0.3	81.6
DANN [7]	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
ADDA [37]	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
MADA[24]	90.0± 0.1	97.4 ± 0.1	99.6 ± 0.1	87.8 ± 0.2	70.3 ± 0.3	66.4 ± 0.3	85.2
SimNet[26]	88.6±0.5	98.2 ± 0.2	99.7±0.2	85.3 ± 0.3	**73.4** ± 0.8	**71.6** ± 0.6	86.2
GTA [30]	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4 ±0.4	86.5
VBDA	**92.1**±0.1	**98.6**±0.1	**100.0**±.0	**93.2**±0.2	69.4±0.1	69.1±0.3	**87.0**

Table 2. Accuracy (%) on Office-Home for unsupervised domain adaptation (ResNet50)

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 [11]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [16]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [7]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [19]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN [17]	**49.0**	69.3	74.5	54.4	66.0	68.4	55.6	**48.3**	75.9	68.4	**55.4**	80.5	63.8
VBDA	45.6	**70.7**	**75.0**	**58.1**	**70.0**	**68.8**	**56.1**	45.8	**76.2**	**69.1**	53.8	**81.3**	**64.21**

Table 3. Classification accuracies (%) on digits datasets.

Source Domain Target Domain	M U	U M	S M
UNIT[15]	**96.0**	93.6	90.5
CyCADA [12]	95.6	96.5	90.4
RAAN [4]	89.0	92.1	89.2
CDAN [17]	95.6	**98.0**	89.2
VBDA(ours)	**96.0**	**98.0**	**93.8**

5.1 Setup

Office-31 [27] is a widely-used dataset for visual domain adaptation, with 4,652 images and 31 categories from three distinct domains: Amazon (**A**), which contains images downloaded from amazon.com, Webcam (**W**) and DSLR (**D**), which contain images taken by web camera and digital SLR camera respectively. We denote the three domains as **A**, **W** and **D**. By permuting the 3 domains, we get 6 domain adaptation tasks.

Office-home [38] is a better organized and more difficult dataset than Office-31, which consists of 15,500 images in 65 object classes in office and home settings. It consists of four extremely dissimilar domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**), and Real-World images (**Rw**). There are 12 domain adaptation tasks by permuting the 4 domains.

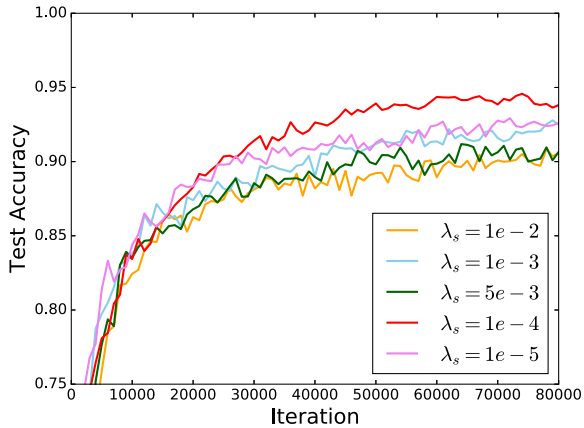
Digits We also explore three digits datasets of varying difficulty, **MNIST**, **SVHN** and **USPS**. Following the evaluation protocol of CyCADA [12], we investigate the following three tasks: USPS to MNIST (**U** → **M**), MNIST to USPS (**M** → **U**) and SVHN to MNIST (**S** → **M**).

We follow the standard protocols for evaluating unsupervised domain adaptation [16, 7]. In the experiments, we observed that the hyperparameters (λ_d, λ_s, λ_t, λ_{ce}) are easy to choose and work well across multiple tasks. Specifically, we keep a fixed weight λ_d for domain adversarial loss and we choose the value of λ_s, λ_t, λ_{ce} from a small candidate set, *i.e.*, $\{0.1, 0.01\}$. The hyperparameters λ_s,λ_t for variational information bottleneck are selected according to the entropy of domain. For example, the higher-entropy domain tends to hold more irrelevant information and needs a larger mutual information regularization weight. The experiments on **Office-31** and **Office-home** is implemented based on ResNet-50 [11] pretrained on the ImageNet dataset [5]. As for the digits dataset, we train our models with a small CNN [6].

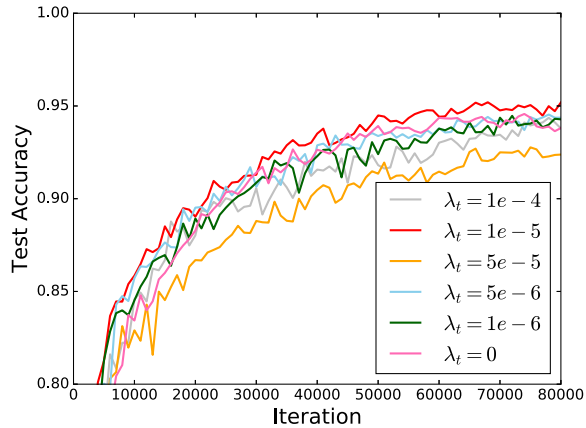
5.2 Results

The results on the Office-31 dataset are reported in Table 1. For fair comparison, the baselines are directly reported from their original papers if the protocol is the same. Our VBDA model remarkably outperforms all comparison methods on most of the tasks. Notably, the model performance are remarkably improved on the hard task, *e.g.*, $A \rightarrow W$, $A \rightarrow D$, where the two domain are significantly different. The interpretation follows that the variations of the source and target domain in these tasks are substantially different, and the task-irrelevant information are the main obstacles for adapting model. Thus this demonstrate that VBDA is good at eliminating these factors and focusing on the essential information for the task of interest. The performance is also further promoted on the relatively easy tasks, such as $D \rightarrow W$ and $W \rightarrow D$. However, the model performance on the tasks, $W \rightarrow A$ and $D \rightarrow A$ are slightly lower than some approaches. This is due to the fact that, the average number of images for 31 classes in Webcam and DSLR are only 26 and 16, which are much lower than the number of bins for representing the image distribution and make the empirically estimated mutual information bounds not reliable enough for applying effective information bottleneck.

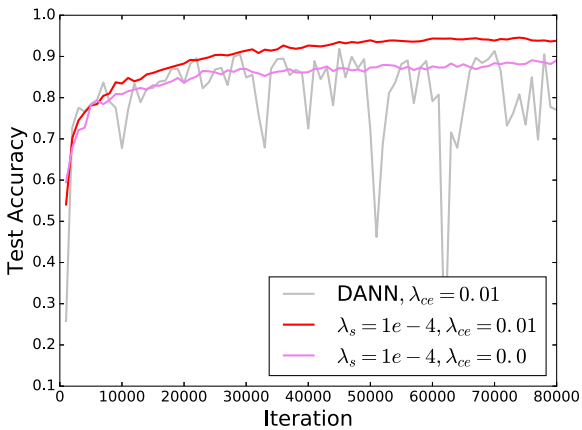
The results on the Office-home can be found in Table 2. The VBDA method significantly promotes the accuracy on most domain adaptation tasks and outperforms CDAN, a state-of-the-art method on this dataset by 0.41% on average. The Office-home is a more challenging dataset, which has four domains with larger domain gap



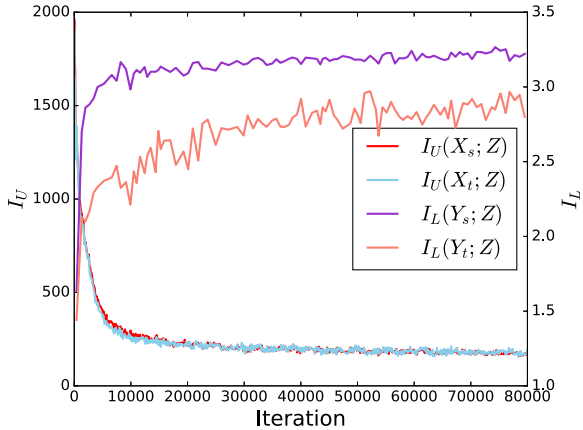
(a) Test accuracy for different λ_s.($\lambda_t = 0$)



(b) Test accuracy for different λ_t.($\lambda_s = 1e-4$)



(c) DANN+CE vs VBDA



(d) $I_U(X;Z)$ and $I_L(Y;Z)$

Figure 1. The sensitivity of the accuracy w.r.t the value of t_h (left) and t_l (right).

and more categories. The information difference between the four domains are more obvious, *i.e.* **Rw** and **Ar** contains much more redundant information than **Cl** and **Pr** for classification task, and the information bottleneck can help control the information flow flexibly to learn *clean* representation for adaptation and classification. The desirable performance on such challenging domain adaption tasks highlights the effectiveness of matching essential information by utilizing information bottleneck principle.

The results on digits datasets are shown in Table (3). In task MNIST→USPS and USPS→MNIST, VBDA performs better or at least comparably with previous methods. And on the more challenging setting, SVHN→MNIST, our model promotes the existing methods by 3.3%. In particular, VBDA outperforms CyCADA, a state-of-the-art pixel-level adaptation method, which further proves the efficacy of VBDA.

5.3 Analysis and Ablation Study

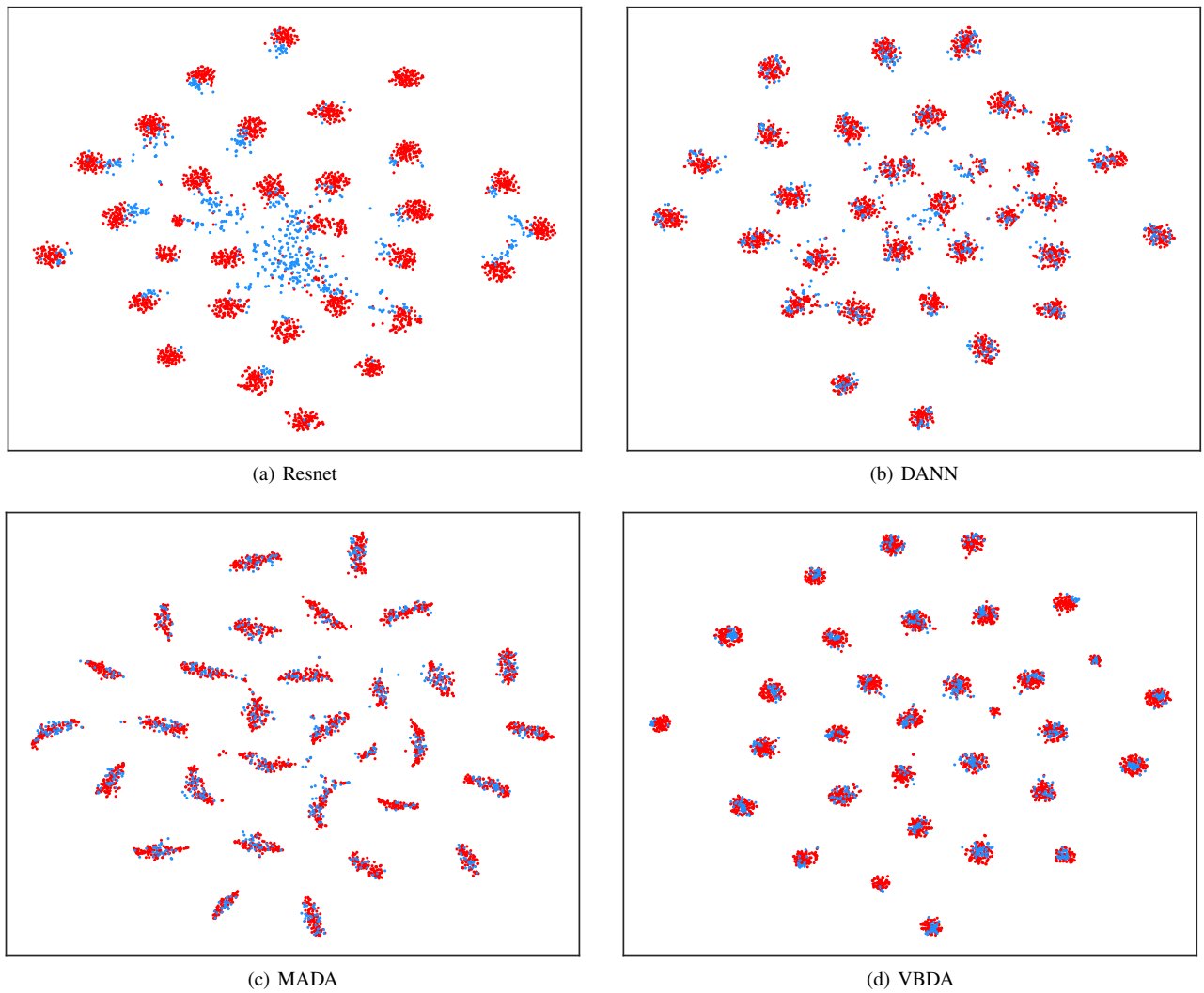
To make a distinction between the utility of two main components of VBDA: the conditional entropy term and the information bottleneck term, we conduct a case study on task SVHN→MNIST. We can observe that in Fig (1(c)), with conditional entropy term only(DANN+CE), the training is quite unstable; with bottleneck term only, the training is stable while the performance declines; with both terms, the model converges stably to a best test accuracy on target domain.

We also conduct ablation studies on hyper-parameter learning for λ_s and λ_t in task SVHN→MNIST. λ_t is preferred to be smaller than λ_s, since SVHN has more irrelevant information to be penalized than MNIST. From Fig (1(a)), we can observe, the accuracy suffers with a too large λ_s. As λ_s becomes larger, we forget more about the input and the learned representation start to become more and more indistinguishable. And the best performance is achieved with an intermediate value of λ_s, in this case, the best setting is $\lambda_s = 1e-4$. Similar phenomenon can be observed on the λ_t in Fig (1(b)).

And the mutual information changes during the optimization is showed in the Fig (1(d)). As we can observe, the mutual information between the representation and label, *i.e.*, $I(Y_s; Z)$ and $I(Y_t; Z)$, are both improved during the training and the mutual information upper bound, *i.e.*, $I_U(X_s; Z)$ and $I_U(X_t; Z)$, between input and representation gradually declined, which indicates that more semantic information has been embedded and more nuisances have been removed in the representation space.

5.4 Feature Visualization

The t-SNE visualization of representation in task A→W (31 classes) is illustrated in Fig (2). Note that the source and target representation is not aligned well by Resnet. DANN can match the marginal feature distribution, but there are still target points near or across the class



(a) Resnet (b) DANN

(c) MADA (d) VBDA

Figure 2. The t-SNE visualization of Resnet, DANN, MADA and VBDA on task **A** $\rightarrow$ **W** (red:**A**,blue:**W**)

boundary. MADA aligns the source and target domain and discriminates categories better, but each class are more scattered than that in VBDA and some target points deviate from the corresponding cluster center. VBDA has clearer cluster boundary and more compact and centered clusters, demonstrating that information irrelevant to classification is filtered by the proposed variational information bottleneck and only information relevant to classification is preserved.

6 Conclusions

In this paper, we proposed Variation Bottleneck Domain Adaptation (VBDA), a simple yet effective regularization mechanism for unsupervised domain adaptation. VBDA enhances semantic information and removes irrelevant factors in the learned representation space, which improves generalization ability and renders strong hypothesis such as cluster assumption more realistic. Comprehensive experiments demonstrate that the proposed approach achieves state-of-the-art performance on various domain adaptation benchmarks.

ACKNOWLEDGEMENTS

This work is sponsored by APEX-YITU Joint Research Program. The corresponding author Yong Yu and the team thank the support of National Natural Science Foundation of China (61702327, 61772333, 61632017, 81771937).

REFERENCES

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, 'Deep variational information bottleneck', *arXiv preprint arXiv:1612.00410*, (2016).
[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, 'A theory of learning from different domains', *Machine learning*, (2010).
[3] Olivier Chapelle and Alexander Zien, 'Semi-supervised classification by low density separation.', in *AISTATS*. Citeseer, (2005).
[4] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty, 'Re-weighted adversarial adaptation network for unsupervised domain adaptation', in *CVPR*, (2018).
[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', (2009).

[6] Geoffrey French, Michal Mackiewicz, and Mark Fisher, 'Self-ensembling for domain adaptation', *arXiv preprint arXiv:1706.05208*, (2017).
[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, 'Domain-adversarial training of neural networks', *JMLR*, (2016).
[8] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li, 'Deep reconstruction-classification networks for unsupervised domain adaptation', in *ECCV*. Springer, (2016).
[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial nets'.
[10] Yves Grandvalet and Yoshua Bengio, 'Semi-supervised learning by entropy minimization', in *NIPS*, pp. 529–536, (2005).
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *CVPR*, (2016).
[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell, 'Cycada: Cycle-consistent adversarial domain adaptation', *arXiv preprint arXiv:1711.03213*, (2017).
[13] Insu Jeon, Wonkwang Lee, and Gunhee Kim, 'Ib-gan: Disentangled representation learning with information bottleneck gan', (2018).
[14] Diederik P Kingma and Max Welling, 'Auto-encoding variational bayes', *arXiv preprint*, (2013).
[15] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, 'Unsupervised image-to-image translation networks', in *NIPS*, (2017).
[16] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan, 'Learning transferable features with deep adaptation networks', *arXiv preprint arXiv:1502.02791*, (2015).
[17] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan, 'Conditional adversarial domain adaptation', in *NeurIPS*, (2018).
[18] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan, 'Unsupervised domain adaptation with residual transfer networks', in *NIPS*, (2016).
[19] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan, 'Deep transfer learning with joint adaptation networks', in *ICML*. JMLR. org, (2017).
[20] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei, 'Label efficient learning of transferable representations acrosss domains and tasks', in *NIPS*, (2017).
[21] Saeid Motiian and Gianfranco Doretto, 'Information bottleneck domain adaptation with privileged information for visual recognition', in *European Conference on Computer Vision*, pp. 630–647. Springer, (2016).
[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, 'Reading digits in natural images with unsupervised feature learning', (2011).
[23] Sinno Jialin Pan and Qiang Yang, 'A survey on transfer learning', *TKDE*, (2010).
[24] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang, 'Multi-adversarial domain adaptation', in *AAAI*, (2018).
[25] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine, 'Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow', *arXiv preprint*, (2018).
[26] Pedro O Pinheiro, 'Unsupervised domain adaptation with similarity learning', in *CVPR*, (2018).
[27] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, 'Adapting visual category models to new domains', in *ECCV*. Springer, (2010).
[28] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada, 'Asymmetric tri-training for unsupervised domain adaptation', *arXiv preprint arXiv:1702.08400*, (2017).
[29] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada, 'Asymmetric tri-training for unsupervised domain adaptation', in *ICML*. JMLR. org, (2017).
[30] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa, 'Generate to adapt: Aligning domains using generative adversarial networks', in *CVPR*, (2018).
[31] Ohad Shamir, Sivan Sabato, and Naftali Tishby, 'Learning and generalization with the information bottleneck', *Theoretical Computer Science*, (2010).
[32] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon, 'A dirt-t approach to unsupervised domain adaptation', *arXiv preprint arXiv:1802.08735*, (2018).
[33] Eduardo D Sontag, 'Vc dimension of neural networks', *NATO ASI Series F Computer and Systems Sciences*, **168**, 69–96, (1998).
[34] Masashi Sugiyama and Motoaki Kawanabe, *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*, MIT press, 2012.
[35] Naftali Tishby, Fernando C Pereira, and William Bialek, 'The information bottleneck method', *arXiv preprint physics/0004057*, (2000).
[36] Naftali Tishby and Noga Zaslavsky, 'Deep learning and the information bottleneck principle', in *ITW*. IEEE, (2015).
[37] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, 'Adversarial discriminative domain adaptation', in *CVPR*, (2017).
[38] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan, 'Deep hashing network for unsupervised domain adaptation', in *CVPR*, (2017).
[39] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen, 'Learning semantic representations for unsupervised domain adaptation', in *ICML*, (2018).
[40] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon, 'On learning invariant representation for domain adaptation', *arXiv preprint arXiv:1901.09453*, (2019).