Label-Noise Robust Domain Adaptation

Xiyu Yu [1] **Tongliang Liu** [2] **Mingming Gong** [3] **Kun Zhang** [4] **Kayhan Batmanghelich** [5] **Dacheng Tao** [2]

Abstract

Domain adaptation aims to correct the classifiers when faced with distribution shift between source (training) and target (test) domains. State-of-the-art domain adaptation methods make use of deep networks to extract domain-invariant representations. However, existing methods assume that all the instances in the source domain are correctly labeled; while in reality, it is unsurprising that we may obtain a source domain with noisy labels. In this paper, we are the first to comprehensively investigate how label noise could adversely affect existing domain adaptation methods in various scenarios. Further, we theoretically prove that there exists a method that can essentially reduce the side-effect of noisy source labels in domain adaptation. Specifically, focusing on the generalized target shift scenario, where both label distribution P_Y and the class-conditional distribution $P_{X|Y}$ can change, we discover that the denoising Conditional Invariant Component (DCIC) framework can provably ensures (1) extracting invariant representations given examples with noisy labels in the source domain and unlabeled examples in the target domain and (2) estimating the label distribution in the target domain with no bias. Experimental results on both synthetic and real-world data verify the effectiveness of the proposed method.

1. Introduction

In the classical domain adaptation setting, given raw features $\{x_1^T, \cdots, x_n^T\}$ from a target domain, we aim to learn a function to predict the labels $\{y_1^T, \cdots, y_n^T\}$ using labeled data $\{(x_1^S, y_1^S), \cdots, (x_m^S, y_m^S)\}$ from a different but related source domain (Wu et al., 2019). Let X and Y be the variables of features and labels, respectively. In contrast to the standard supervised learning, the joint distributions P_{XY}^S and P_{XY}^T are different. For example, in medical data analysis, health record data collected from patients of different age groups or hospital locations often vary (Purushotham et al., 2017). Inferring invariant knowledge from a domain (e.g., an age group or a location) with a large set of labeled examples to another with unlabeled data is desirable (Raghu et al., 2019) since it is often laborious to obtain high-quality labels for clinical data (Dubois et al., 2017).

According to the assumptions about how the joint distribution P_{XY} shifts across domains, several domain adaptation scenarios have been studied. (1) Covariate shift assumes that the marginal distribution P_X changes but the conditional distribution $P_{Y|X}$ stays the same. In this situation, methods have been proposed to correct the shift in P_X, for instance, by importance reweighting (Huang et al., 2007) and invariant feature learning (Long et al., 2015; Kumagai et al., 2019; Meyerson & Miikkulainen, 2019; Chen et al., 2019a). (2) Model shift (Wang et al., 2014) assumes that P_X and $P_{Y|X}$ change independently. In this case, it also requires Y to be continuous, the change in $P_{Y|X}$ to be smooth, and some labeled data to be available in the target domain. (3) Target shift (Zhang et al., 2013a; Azizzadenesheli et al., 2019) assumes that P_Y shifts while $P_{X|Y}$ stays the same. In this scenario, P_X and $P_{Y|X}$ change dependently because their changes are caused by the change in P_Y. (4) Generalized target shift (Zhang et al., 2013a) assumes that $P_{X|Y}$ and P_Y change independently across domains, causing P_X and $P_{Y|X}$ to change dependently. An interpretation of the difference between these scenarios from a causal standpoint was also provided (Schölkopf et al., 2012a).

Additionally, the aforementioned domain adaptation methods extract invariant features across different domains based on a strong assumption; that is, the source domain labels are accurate. However, since accurately labeling training set tends to be expensive, time-consuming, and sometimes impossible, this assumption is often violated in practice. For example, in medical data analysis, due to the subjectivity of domain experts, insufficient discriminative information, and digitalization errors (Sáez et al., 2016), noisy labels are of-

[1]Department of Computer Vision Technology (VIS), Baidu Incorporation [2]UBTECH Sydney AI Centre, The University of Sydney [3]School of Mathematics and Statistics, University of Melbourne [4]Department of Biomedical Informatics, University of Pittsburgh [5]Department of Philosophy, Carnegie Mellon University. Correspondence to: Xiyu Yu <yuxiyu@baidu.com>, Tongliang Liu <tongliang.liu@sydney.edu.au>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

ten inevitable. In computer vision, to reduce the expensive human supervision, we often prefer directly transferring knowledge from easily obtainable but imperfectly labeled source data such as webly-labeled data or machine-labeled data to target data (Xu et al., 2016; Lee et al., 2018).

Therefore, in this paper, we consider the setting of domain adaptation that the observed labels in source domain are noisy. As such, we have no access to the true source distribution. One may think that this issue can be easy to solve by combining existing label-noise learning methods and domain adaptation methods. For example, simply applying the label-noise robust classifiers after extracting invariant features across domains by employing existing domain adaptation methods. However, for many setting, label noise will degenerate invariant feature extracting and the unlabeled data in the target domain is also helpful for denoising. Simple combination is therefore inefficient.

As expected, except the covariate shift scenario in which correcting the shift in P_X does not require label information, we can show that label noise can adversely affect most existing domain adaptation methods in different scenarios. Taking target shift as an example, by assuming $P_{X|Y}$ is invariant across domains, the shift in P_Y can be corrected by estimating the class ratio between P_Y^T and P_Y^S from a mixture proportion estimation problem (Zhang et al., 2013a; Iyer et al., 2014). However, when labels in source domain are corrupted, the information of $P_{X|Y}$ is unknown. Then, it is unclear whether the class ratio P_Y^T/P_Y^S can be estimated. Another example is generalized target shift. In this scenario, the estimated P_Y^T/P_Y^S can be possibly incorrect. Further, invariant features are often learned by matching distributions across domains which heavily rely on the estimate of P_Y^T/P_Y^S. As a result, label noise can lead to biased learning of features with incorrect estimate of P_Y^T/P_Y^S. Label noise also affects the learning in model shift, but we will not consider this case because we are concerned with discrete labels and the setting where no label exists in target domain.

To address this issue, we propose a label-noise robust domain adaptation method in the generalized target shift scenario. To deal with label noise, we propose a novel method to denoise conditional invariant components. Our method can provably identify the changes in distribution P_Y and extract the conditional invariant representations by reducing the side effect of label noise using both source and target data. Specifically, we construct a new distribution $P_{X'}^{\text{new}}$ which is marginalized from the weighted noisy source distribution $P_{\rho X',Y}^S$. Here, we denote P_ρ as the distributions associated with label noise. By matching $P_{X'}^{\text{new}}$ and $P_{X'}^T$, the conditional invariant components and P_Y^T are identifiable from the noisy source data and unlabeled target data. Moreover, in our denoising conditional invariant component framework, we can also theoretically ensure the convergence of the estimate of label distribution in target domain.

To verify the effectiveness of our method, we conduct comprehensive experiments on both synthetic and real-world data. The performance are evaluated on classification problems. For fair comparison, after extracting invariant features using domain adaptation methods, we train the robust classifier by employing the forward method in (Patrini et al., 2017). Compared with state-of-the-art domain adaptation methods, our method achieves superior performance.

2. Related Work

Classification with Label Noise. Learning with noisy labels in classification has been widely studied (Long & Servedio, 2008; Van Rooyen et al., 2015). These methods can be coarsely categorized into three categories, i.e., unbiased losses or risk minimizers (Natarajan et al., 2013; Xu et al., 2019a; Sukhbaatar et al., 2014; Patrini et al., 2017; Han et al., 2018a), bootstrapping losses (Arazo et al., 2019), label noise reweighting and cleansing (Jiang et al., 2018; Han et al., 2018b; Chen et al., 2019b; Thulasidasan et al., 2019; Nguyen et al., 2020). Learning with complementary labels (Xu et al., 2019b; Yu et al., 2018b; Ishida et al., 2017; L. Feng & Sugiyama, 2020; Y.-T. Chou & Sugiyama, 2020) can also be viewed as a special case of learning with label noise. They often exploit similar ideas when designing robust models. Depending on whether we explicitly model label noise using transition matrix, label noise-robust methods can also be classified into transition matrix based methods (B. Han & Sugiyama, 2020; Xia et al., 2019; 2020) and transition matrix-free methods (Yang et al., 2019; Han et al., 2018c; Cheng et al., 2020; Liu & Guo, 2020; Wu et al., 2020). Here,, our method belongs to the first category. However, the problem considered here is more challenging because the clean source domain distribution is not assumed to be identical to the target domain distribution. In contrast to classification with label noise, our method can learn invariant features across different domains, where both P_Y and $P_{X|Y}$ may change and the labels of the source data is corrupted. Reports on the general results obtained in this setting are scarce.

Traditional Generalized Target Shift Methods. Existing generalized target shift methods assume that there exists a transformation τ, e.g., location-scale transformation (Zhang et al., 2013a; Gong et al., 2016), such that the conditional distribution $P_{\tau(X)|Y}$ is invariant across domains. In this paper, we also assume that the conditional invariant components (CICs) exist. We aim to find a transformation τ such that $P^T(\tau(X)|Y) = P^S(\tau(X)|Y)$ as in (Gong et al., 2016) and to estimate $P^T(Y)$. However, we are given only samples drawn from the distribution P_X^T and the noisy distribution $P_{\rho XY}^S$, which makes the problem challenging.

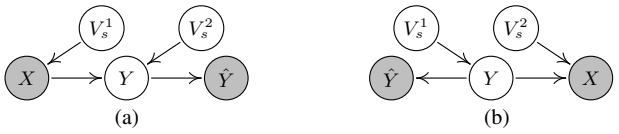


Figure 1. Possible situations of domain adaptation with label noise. V_s^1 and V_s^2 are independent domain-specific selection variables, leading to changing P_{XY} across domains. (a) Model shift: V_s^1 and V_s^2 change P_X and $P_{Y|X}$, respectively. (b) Generalized target shift: V_s^1 and V_s^2 change P_Y and $P_{X|Y}$, respectively. In the first scenario, X is a cause for Y, whilst in the second scenario, Y is a cause of X. If V_s^2 is not present, (a) reduces to covariate shift and (b) reduces to target shift. In our setting, the true labels Y in the source domain is unobservable. We only observe noisy labels $\hat{Y}$.

Note that our work is not a simple combination of traditional generalized target shift methods and robust classifiers. As aforementioned, simple combination of domain adaptation and label-noise robust classifier overlooks that the learning of invariant features can be affected by label noise, which thus produces biased results. In the setting where only noisy source data and unlabeled target data are available, learning τ becomes pretty challenging. This is because without clean label Y in both domains, no direct information is available to ensure the identity of conditional distributions $P(\tau(X)|Y)$. As such, τ is hard to learn. Moreover, it is challenging to estimate $P^T(Y)$ as briefly discussed in the introduction. Therefore, we proposed a novel denoising conditional invariant component framework. It is able to identify $P^T(Y)$ and conditional invariant components $\tau(X)$ from the noisy source data and unlabeled target data.

In this paper, the simple combinations of domain adaptation methods with robust classifiers are included as baselines in our experiments. Our method strongly outperforms the baselines, verifying that the superiority of the proposed method to extract invariant features across different domains.

3. The Effects of Label Noise

In this section, we examine the effects of label noise in four different domain adaptation scenarios, namely 1) covariate shift, 2) model shift, 3) target shift, and 4) generalized target shift. From a causal perspective, 1) and 2) assume that X causes Y, indicating that P_X and $P_{Y|X}$ contain no information about each other (Schölkopf et al., 2012b). In domain adaptation, the causal relation implies that changes in P_X are independent of changes in $P_{Y|X}$. If the change in $P_{Y|X}$ is large, then it is difficult to correct the shift in $P_{Y|X}$ because we often have no or scarce labels in the target domain. On the contrary, 3) and 4) assume that Y is the cause for X, implying that changes in P_Y and $P_{X|Y}$ are independent, while changes in P_X and $P_{Y|X}$ depend on each other. Figure 1 represents the causal relations between variables in domain adaptation using selection diagram defined in (Pearl & Bareinboim, 2011). Here, although the noisy label $\hat{Y}$ is usually generated after X is observed, we exploit the causal model $Y \to \hat{Y}$ according to the assumption that flip rates are independent of features, which is widely employed in the label noise setting (Natarajan et al., 2013; Patrini et al., 2017; Scott, 2015). The effects of label noise in different scenarios are also summarized as follows:

Covariate shift. In covariate shift (Huang et al., 2007; Zhang et al., 2013b), label noise has no effects on the correction of shift in P_X. However, after correcting the shift in P_X, one needs to take the effects of label noise into account when training a classifier on the source domain (Natarajan et al., 2013; Liu & Tao, 2016). This problem can be efficiently solved by a simple combination of label-noise learning and domain adaptation.

Model shift. In model shift (Wang et al., 2014), since P_X and $P_{Y|X}$ change independently, we can correct them separately. Similar to covariate shift, correcting P_X is not affected by label noise. However, correcting shift in $P_{Y|X}$ requires matching $P^S_{Y|X}$ and $P^T_{Y|X}$, which can be seriously harmed by label noise. In this scenario, since a small number of clean labels are assumed to be available in the target domain, $P_{Y|X}$ is often assumed to change smoothly across domains to reduce the estimation error. The smoothness constraint can reduce the effects of label noise to some extent if one directly matches $P^S_{\rho Y|X}$ and $P^T_{Y|X}$.

Target shift. In target shift (Iyer et al., 2014; Zhang et al., 2013a; Jiaxian Guo & Tao, 2020), it is required that $P^S_{X|Y} = P^T_{X|Y}$. The changes in P_Y are often corrected by matching the marginal distribution of the reweighted source domain $P^{\text{new}}_X = \sum_{i=1}^{c} P^S_{X|Y=i} P^S_{Y=i} \beta(Y=i)$ and the target domain P^T_X, where $\beta(Y=i) = P^T_{Y=i}/P^S_{Y=i}$ and c is the class number. In the presence of label noise, however, we only have access to $P^S_{\rho X|Y}$ and $P^S_{\rho Y}$ in the source domain. In this situation, the estimate of P^T_Y can be incorrect. Take binary problem as an example, let $P^T_X = \omega_{\rho 1} P^S_{\rho X|Y=1} + \omega_{\rho 2} P^S_{\rho X|Y=2}$, $P^T_X = \omega_1 P^S_{X|Y=1} + \omega_2 P^S_{X|Y=2}$, and $\pi_{ij} = P^S(Y=j|\hat{Y}=i), \forall i,j \in \{1,2\}$. In fact, ω_i and $\omega_{\rho i}$ respent $P^T_{Y=i}$ and $P^S_{\rho Y=i}$ $(i=1,2)$, respectively. Then,

Proposition 1. *We have* $\omega_{\rho i} = \omega_i, i = 1,2$ *only when* $\pi_{12}\omega_1 = \pi_{21}\omega_2$.

Here, $P^S_{\rho X|Y=1} = \pi_{11} P^S_{X|Y=1} + \pi_{12} P^S_{X|Y=2}$ and $P^S_{\rho X|Y=2} = \pi_{21} P^S_{X|Y=1} + \pi_{22} P^S_{X|Y=2}$ are known as mutually contaminated distributions (Menon et al., 2015). We can see, $\pi_i j$ and transition probability $P(\hat{Y}=j|Y=i)$, $i,j \in \{1,\cdots,c\}$ can be related via Bayes' rule. According to Proposition 1, in the special case where the label distribution of target domain is balanced and the label noise is symmetric, label noise does not affect the estimation of

P_Y^T. But in most cases, $\omega_i \neq \omega_{\rho i}$. This indicates that we cannot directly estimate P_Y^T from the noisy source data and unlabeled target data. Detailed proof of Proposition 1 can be found in the Supplementary Material.

Generalized target shift. In general target shift (Zhang et al., 2013a; Gong et al., 2016), $P_{X|Y}$ also changes across domains, but it changes independently of P_Y. A widely-employed approach is learning conditional invariant components that satisfy $P^S_{X'|Y} = P^T_{X'|Y}$. Under the assumption of conditional invariant components, many works jointly learn X' and $P^T(Y)$ by matching $P^{\text{new}}_{X'} = \sum_{i=1}^{c} P_{X'|Y=i} P^S_{Y=i} \beta(Y=i)$ and $P^T_{X'}$, which naturally requires the information of P^S_{XY} and P^T_X.

However, in the setting of label noise, similar to target shift, the estimates of invariant components and P_Y^T are very likely to be inaccurate if we directly use the noisy source distribution $P^S_{\rho XY}$ to correct distribution shift. Specifically, if we assume that X' is successfully learned, the estimate of P_Y^T may be incorrect as that in target shift. A wrong estimate of P_Y^T can in turn result in the biased learning of invariant representations as in (Gong et al., 2016).

In conclusion, we can observe that label noise is harmful for extracting invariant features and correcting distribution shift in most domain adaptation scenarios. We target to reduce these adverse effects of label noise in the following sections.

4. Label-Noise Robust Domain Adaptation

Here, we study a new domain adaptation setting in which (1) both $P_{X|Y}$ and P_Y change across different domains; (2) and we have access to only "noisy" observations $\{(x_1^S, \hat{y}_1^S), \cdots, (x_m^S, \hat{y}_m^S)\}$ in the source domain and unlabeled data $\{x_1^T, \cdots, x_n^T\}$ in the target domain. Here, $\hat{y}$ refers to a noisy label; and we consider the class-conditional label noise (Natarajan et al., 2013). The label noise is stochastically modeled via a transition probability $P(\hat{Y} = j|Y = i)$, i.e., the flip rate from clean label i to noisy label j. All these transition probabilities are summarized into a transition matrix Q, where $Q_{ij} = P(\hat{Y} = j|Y = i)$. The class-conditional label noise is the vast majority noise setting adopted in the label noise community. It has been widely used and been proved to be effective for evaluating label noise methods such as (Natarajan et al., 2013; Chen et al., 2019b).

In this section, we first study how to provably identify invariant feature across different domains and correct the distribution shift in the general target shift scenario with label noise. Then, an importance reweighting framework is introduced for correcting classifiers. Both our end-to-end deep domain adaptation model is finally presented.

4.1. Denoising Conditional Invariant Components

In the label noise setting, learning invariant features and P_Y^T is challenging due to that we can only observe the noisy labels but have no clean label Y in the source domain. To address this issue, we first introduce the conditional invariant components to ensure this problem being tractable. That is, we assume that for every d-dimensional data X, there exists a transformation $\tau : \mathbb{R}^d \to \mathbb{R}^{d'}$ satisfying

$$P^T_{\tau(X)|Y} = P^S_{\tau(X)|Y}, \tag{1}$$

where $X' = \tau(X) \in \mathbb{R}^{d'}$ are known as conditional invariant components (CICs) (Gong et al., 2016) across domains.

Since label noise makes existing domain adaptation methods ineffective, we propose a novel method to denoise the conditional invariant components. We find that if the information of label noise model is available, a unique relationship between $P^S_{\rho X'Y}$ and $P^T_{X'}$ can be built, which, in turn, is a clue for us to identify X'.

We observe that label noise does not affect the distribution of X'. Then, intuitively, if we marginalize out the variable $\hat{Y}$ of the noisy labels, we may achieve Eq. (1) by matching the marginal distribution $P_{X'}$. But we need some nontrivial strategies to make it possible. Specifically, we first construct a new distribution $P^{\text{new}}_{X'}$, which is marginalized from the reweighted distribution $P^S_{\rho X'Y}$ as follows,

$$\begin{aligned} P^{\text{new}}_{X'} &= \sum_{y'} \beta_\rho(\hat{Y} = y') P^S_\rho(X', \hat{Y} = y') \\ &= \sum_{y} \sum_{y'} \beta_\rho(\hat{Y} = y') P^S_\rho(X', Y = y, \hat{Y} = y'), \end{aligned} \tag{2}$$

where β_ρ are the weights for noisy labels. Note that, in the rest of this paper, when no ambiguity occurs, we use Y as the variable for both "clean" and "noisy" labels; otherwise, both Y and $\hat{Y}$ are used as variables for "clean" and "noisy" label, respectively.

Then, under mild conditions, by matching the distribution $P^T_{X'}$ with the new distribution $P^{\text{new}}_{X'}$, we can provably identify the invariant components $\tau(X)$:

Theorem 1. *Suppose the transformation τ satisfies that $P(\tau(X)|Y = i), i \in \{1, \cdots, c\}$ are linearly independent, and that the elements in the set $\{v_i P^S(\tau(X)|Y = i) + \lambda_i P^T(\tau(X)|Y = i); i \in \{1, \cdots, c\}; \forall v_i, \lambda_i \, (v_i^2 + \lambda_i^2 \neq 0)\}$ are linearly independent. Then, if $P^{new}_{X'} = P^T_{X'}$, we have $P^T_{X'|Y} = P^S_{X'|Y}$; and $\beta(Y = y) = \sum_{y'} P^S(\hat{Y} = y'|Y = y)\beta_\rho(\hat{Y} = y'), \forall y, y' \in \{1, \cdots, c\}$, where $\beta(Y = y) = P^T(Y = y)/P^S(Y = y)$.*

Please see the proof of Theorem 1 in the Supplementary Material. Note that the linearly independent property is a weak assumption which has been widely used as the basic condition for class ratio estimation (Gong et al., 2016).

Let $\mathbf{u} = [\beta(Y=1), \cdots, \beta(Y=c)]^\top$ and $\mathbf{u}_\rho = [\beta_\rho(Y=1), \cdots, \beta_\rho(Y=c)]^\top$. According to Theorem 1, we have $\mathbf{u} = Q\mathbf{u}_\rho$. In label noise, we assume that Q is usually diagonally dominant and invertible. Then, the relationship between β_ρ and β is uniquely determined, as well as the relationship between $P^S_{\rho\, X'Y}$ and $P^T_{X'}$. In this case, if Q is known and these two marginal distributions are successfully matched, we can (1) identify the conditional invariant components; (2) and learn β_ρ which indicates that the changes in the distribution P_Y is also identifiable. In practice, the transition matrix Q is not available, but we can estimate it by methods in (Liu & Tao, 2016; Patrini et al., 2017).

In Thoerem 1, we focus on the linear independence assumption on $P_{X'|Y}$. In the following section, we exploit β and Q to correct β_ρ such that we can correct the distribution shift directly on the unbiased estimators of clean distributions. But it is also interesting to note that this theorem indicates that the learning of conditional invariant components are not affected by label noise. Let π be the matrix in which $\pi_{ij} = P(Y=j|\hat{Y}=i)$. Again, π and Q are related by Bayes' rule. If Q is invertible, then it is easy to obtain that π is also invertible. In this condition, if we assume $P_{X'|Y=i}, \forall i \in \{1, \cdots, c\}$ are linear independent, then $P_{\rho X'|\hat{Y}=i}, \forall i \in \{1, \cdots, c\}$ are also linear independent. According to Theorem 1 in (Gong et al., 2016), we can see that the conditional invariant components can be identified by correcting the changes in $\beta_\rho(\hat{Y}=y)P_\rho(X', \hat{Y}=y)$. That is to say, we provably find that CIC method in (Gong et al., 2016) is robust to label noise when identifying conditional invariant components.

But this conclusion may be not empirically correct. In our experiments, we find that by correcting β_ρ to obtain an unbiased estimator of clean distributions, the proposed denoising maximum mean discrepancy (MMD) loss can perform better. The modified MMD loss is present as follows.

Denoising MMD Loss. To enforce the matching between $P^{\text{new}}_{X'}$ and $P^T_{X'}$, we employ the kernel mean matching of these two distributions and minimize the squared maximum mean discrepancy (MMD) loss:

$$\begin{aligned} &\|\mu_{P^{\text{new}}_{X'}}[\psi(X')] - \mu_{P^T_{X'}}[\psi(X')]\|^2 \\ &= \|\mathbb{E}_{X'\sim P^{\text{new}}_{X'}}[\psi(X')] - \mathbb{E}_{X'\sim P^T_{X'}}[\psi(X')]\|^2, \end{aligned} \tag{3}$$

where ψ is a kernel mapping. According to Eq. (2), we have

$$\mathbb{E}_{X'\sim P^{\text{new}}_{X'}}[\psi(X')] = \mathbb{E}_{(X',Y)\sim P^S_{\rho X'Y}}[\beta_\rho(Y)\psi(X')].$$

Therefore, minimizing Eq. (3) is equivalent to minimizing

$$\|\mathbb{E}_{(X',Y)\sim P^S_{\rho X'Y}}[\beta_\rho(Y)\psi(X')] - \mathbb{E}_{X'\sim P^T_{X'}}[\psi(X')]\|^2.$$

In practice, we can only observe the corruptly labeled source data $\{(x_1, \hat{y}^S_1), \cdots, (x_m, \hat{y}^S_m)\}$ and the unlabeled target data $\{x^T_1, \cdots, x^T_n\}$. Therefore, we approximate the expected kernel mean values by the empirical ones:

$$\|\frac{1}{m}\psi(\mathbf{x'}^S)\beta_\rho(\hat{\mathbf{y}}^S) - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\|^2, \tag{4}$$

where $\beta_\rho(\hat{\mathbf{y}}^S) = [\beta_\rho(\hat{y}_1), \cdots, \beta_\rho(\hat{y}_m)]^\top$; $\mathbf{x}'$ denotes the matrix of the invariant representations.

However, Eq. (4) is not explicitly formulated w.r.t. P^T_Y. If we directly optimizing Eq. (4) w.r.t. $\beta_\rho(\hat{\mathbf{y}}^S)$, it will result in incorrect β_ρ that violates the fact that $\beta_\rho(\hat{y})$ should be the same for the same $\hat{y}$. It is thus impossible to identify P^T_Y.

Therefore, we need to reparameterize the formulation by applying the relationship between β_ρ and P^T_Y in Theorem 1, i.e., $\beta_\rho(\hat{Y}=i) = \sum_{j=1}^c Q^{-1}_{ij}\frac{P^T(Y=j)}{P^S(Y=j)}$. It is also easy to derive that $[P^S(Y=1), \cdots, P^S(Y=c)]Q = [P^S_\rho(Y=1), \cdots, P^S_\rho(Y=c)]$. Given estimated $\hat{Q}$ and $[\hat{P}^S_\rho(Y=1), \cdots, \hat{P}^S_\rho(Y=c)]^\top$, we can construct the vectors $\mathbf{g}_i = [\frac{\hat{Q}^{-1}_{i1}}{\hat{P}^S(Y=1)}, \cdots, \frac{\hat{Q}^{-1}_{ic}}{\hat{P}^S(Y=c)}], i \in \{1, \cdots, c\}$. If $\hat{y}_k = i, \forall k \in \{1, \cdots, m\}$, define the matrix $G \in \mathbb{R}^{m\times c}$, where the k-th row of G is $\mathbf{g}_i$. Let $\beta_\rho(\hat{\mathbf{y}}^S) = G\alpha$. Then, α is an estimate of $[P^T(Y=1), \cdots, P^T(Y=c)]^\top$.

The denoising MMD loss now can be reparametrized as

$$\begin{aligned} &\|\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\|^2 \\ &= \frac{\alpha^\top G^\top \mathbf{K}^S G\alpha}{m^2} - \frac{2\mathbf{1}^\top \mathbf{K}^{T,S} G\alpha}{mn} + \frac{\mathbf{1}^\top \mathbf{K}^T \mathbf{1}}{n^2}, \end{aligned} \tag{5}$$

where $\mathbf{K}^S$ and $\mathbf{K}^T$ are the kernel matrix of $\mathbf{x'}^S$ and $\mathbf{x'}^T$, respectively; $\mathbf{K}^{T,S}$ is the cross kernel matrix. In this paper, the Gaussian kernel, i.e., $k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ is applied, where σ is the bandwidth.

Therefore, according to Theorem 1, optimizing the denoising MMD loss in Eq. (5) ensures us to identify the conditional invariant components and $P^T(Y)$.

A New Perspective on Denoising MMD Loss. Here, we discuss why using β and Q to correct β_ρ can be more helpful. By correcting β_ρ by β and Q, we actually provide an unbiased estimator of $\sum_y \beta(Y=y)P^S(X', Y=y)$. This proof is straightforward. We can easily prove that πQ is identity matrix when Q is invertible; and $\mathbf{u}_\rho = \pi\mathbf{u}$. Replace $\beta_\rho(\hat{Y})$ with $\beta(Y)$ using the above relationship. Then, we can easily obtain $P^{\text{new}}_{X'} = \sum_y \beta(Y=y)P^S(X', Y=y)$.

That is to say, by correcting β_ρ, we can build the direct relationship between $P^{T\prime}_X$ and $P^S_{X'|Y}$. This is very important because this enables us to directly correct the changes in $P(Y=y)P(X'|Y=y)$ and extract P^T_Y. Even though when Q is invertible, β_ρ is provably identifiable according to Theorem 1, the learning process is more difficult since the mixed noisy data are closer to each other especially when

only finite examples are given. This is why our denoising MMD loss can work better.

4.2. Importance Reweighting

After adapting invariant features, we can now correct the classifiers. Here, we aim to learn a hypothesis function $f^* : \mathbb{R}^{d'} \to \mathbb{R}^c$ from the noisy source data that can generalize well on the target data. Ideally, f^* minimizes the expected loss $\mathbb{E}_{(X',Y)\sim P^T_{X'Y}}[\ell(f(X'),Y)]$, where ℓ is the loss function; X' are the conditional invariant components.

In practice, we assume that f^* can predicts $P^T(Y|X')$ (Reid & Williamson, 2010; Patrini et al., 2017) and $\arg\max_{i\in\{1,\cdots,c\}} f^*_i$ predicts the label. Here, f^*_i is the i-th entry of f^*. To facilitate the learning of f^*, we first imagine that the target domain has the same label noise model as the source domain. Note that, this does not necessarily imply that label noise really exists in target domain because, in our setting, we even have no label information of target data. We can see, the minimizer $f^*_\rho = \arg\min_f \int \ell(f(X'),Y)P^T_\rho(X',Y)dX'dY$ is also assumed to be able to predict $P^T_\rho(Y|X')$. If the classifier f^*_ρ is found and Q is invertible, we can obtain f^* according to the following relationship:

$$\begin{aligned}&[P^T(Y=1|X'),\cdots,P^T(Y=c|X')]Q\\&=[P^T_\rho(Y=1|X'),\cdots,P^T_\rho(Y=c|X')].\end{aligned}\tag{6}$$

Thus, the problem remains to learn f^*_ρ, which can be obtained by exploiting the importance reweighting strategy:

$$\begin{aligned}f^*_\rho &= \arg\min_f \int \ell(f(X'),Y)P^T_\rho(X',Y)dX'dY\\&=\arg\min_f \int \frac{P^T_\rho(X',Y)}{P^S_\rho(X',Y)}\ell(f(X'),Y)P^S_\rho(X',Y)dX'dY.\end{aligned}$$

Since $P^T_\rho(X',Y)$ is constructed from $P^T(X,Y)$ by using the same transition matrix Q and $P^T(X'|Y)=P^S(X'|Y)$, we can easily have $P^T_\rho(X'|Y)=P^S_\rho(X'|Y)$ and thus

$$\begin{aligned}f^*_\rho &= \arg\min_f \int \frac{P^T_\rho(Y)}{P^S_\rho(Y)}\ell(f(X'),Y)P^S_\rho(X',Y)dX'dY\\&=\arg\min_f \int \gamma(Y)\ell(f(X'),Y)P^S_\rho(X',Y)dX'dY,\end{aligned}$$

where $\gamma(Y)=\frac{P^T_\rho(Y)}{P^S_\rho(Y)}$. In practice, only the training sample is observable, we thus minimize the empirical loss,

$$\hat{R}=\frac{1}{m}\sum_{i=1}^m \gamma(\hat{y}^S_i)\ell(f(x'^S_i),\hat{y}^S_i),\tag{7}$$

to find the approximated classifier f_ρ.

Instead of separately finding f^*_ρ by minimizing Eq. (7) and transiting f^*_ρ to f^* according to Eq. (6), in this paper, we employ the forward strategy proposed in (Patrini et al., 2017); that is, we directly minimize the following risk,

$$\hat{R}=\frac{1}{m}\sum_{i=1}^m \gamma(\hat{y}^S_i)\ell(Q^\top f(x'^S_i),\hat{y}^S_i),\tag{8}$$

As we know, by minimizing the risk $\hat{R}$, $Q^\top f(x'^S_i)$ can approximately predict $P^T_\rho(Y|X')$. Then, according to Eq. (6), $f(x'^S_i)$ can finally approximately predict $P^T(Y|X')$.

Note that, in practice, the ratio $\gamma(Y)$ is also unknown. But $P^S_\rho(Y)$ can be empirically estimated from the noisy source data, and $P^T(Y)$ is estimated by our denoising MMD loss, $P^T_\rho(Y)$ can also be computed according to the relationship similar to Eq. (6). In this way, $\gamma(Y)$ can be obtained.

4.3. The Overall Models

In order to extract conditional invariant components, the transformation τ varies from linear ones to non-linear ones depending on the complexity of input data space. Since linear model is similar except a two-stage procedure, we mainly present our end-to-end deep learning model. We modify the conventional deep neural network for classification, e.g., AlexNet (Krizhevsky et al., 2012), in two aspects: (1) Due to that the domain discrepancy becomes larger for the features in higher-level layers (Long et al., 2015; 2017), we impose the denoising MMD loss on a higher-level layer for extracting the invariant representations; (2) to learn a classifier robust to label noise, we add the forward procedure (Patrini et al., 2017) before the cross-entropy (CE) loss as in Eq. (8). Descriptions about linear model and the structure of deep model can be found in Supplementary Material.

Let h^l be the responses of the l-th hidden layer, $W_{1:l}$ be the parameters in the 1-th to l-th layers, and L be the total number of layers in our deep model. Suppose that we impose the denoising MMD loss on the features in the l-th layer; that is, $\tau(x_i)=h^l_i$. Then, the denoising MMD loss is

$$\hat{D}(W_{1:l},\alpha)=\|\frac{1}{m}\psi(\mathbf{h}^{l^S})G\alpha-\frac{1}{n}\psi(\mathbf{h}^{l^T})\mathbf{1}\|^2,\tag{9}$$

where $\mathbf{h}^l$ is the matrix of the responses of the l-th layer.

Denote $f(x_k)$ as the softmax output w.r.t. the input x_k. According to Eq. (8), the loss for classification is

$$\hat{R}(W_{1:L})=\frac{1}{m}\sum_{k=1}^m \gamma(\hat{y}^S_k)CE(Q^\top f(x^S_k),\hat{y}^S_k),\tag{10}$$

where $\gamma(\hat{y}^S_k)=\frac{\alpha^\top Q_{:i}}{P^S_\rho(Y=i)}$ if $\hat{y}^S_k=i$; $Q_{:i}$ denotes the i-th column of Q. Together with the regularization $\Omega(W_{1:L})$

(e.g., l_2 norm) of the parameters, our final model becomes

$$\min_{W_{1:L},\alpha} \hat{R}(W_{1:L}) + \lambda_1 \hat{D}(W_{1:l_1}, \alpha) + \lambda_2 \Omega(W_{1:L}),$$
$$\text{s.t.} \sum_{i=1}^{c} \alpha_i = 1; \alpha_i \geq 0, \forall i \in \{1, \cdots, c\}, \tag{11}$$

where λ_1 and λ_2 are the tradeoff parameters of denoising MMD loss and regularization, respectively. Again, by minimizing Eq. (11), if $Q^\top f(X)$ approximates $P_\rho^T(Y|X)$, then $f(X)$ approximates $P^T(Y|X)$. We can then successfully learn the classifier for the target data.

4.4. Convergence Analysis

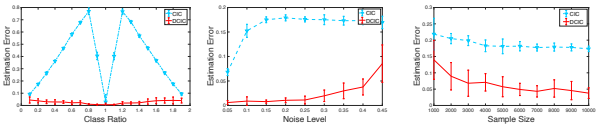
In this subsection, we study the convergence rates of the estimates to the true label noise rates and optimal class priors. Estimation of noise rate can be viewed as a mixture proportion estimation problem (Yu et al., 2018a; Ramaswamy et al., 2016; Yao et al., 2020). The convergence rate for the label noise rates has been well studied under the "anchor set" condition that for any y there exist x in the domain of X such that $P(Y = y|X) = 1$ and $P(Y = y'|X) = 0, \forall y' \neq y$, which is likely to be held in practice. For example, estimators with convergence guarantees has been proposed in (Liu & Tao, 2016). Recently, (Ramaswamy et al., 2016) exploited the "anchor set" condition in Hilbert space and designed estimators that can converge to the true label noise rates with an order of $O(m^{-\frac{1}{2}})$. Some work based on a weaker assumption, i.e, linearly independent assumption, is also proposed to estimate label noise, and a fast convergence is also guaranteed (Yu et al., 2018a). Therefore, we mainly focus on the convergence analysis of estimating class ratios.

To analyze the convergence rate of the estimated class prior $\hat{\alpha}$ to the optimal α^* in the presence of label noise, we first abuse the training samples $\{(x_1^S, \hat{y}_1^S), \cdots, (x_m^S, \hat{y}_m^S)\}$ and $\{x_1^T, \cdots, x_n^T\}$ as i.i.d. variables, respectively. Abuse W as the parameters related to the transformation τ and let $\mathcal{D}(W, \alpha) = \|\mathbb{E}\frac{1}{m}\psi(\mathbf{x}'^S)G\alpha - \mathbb{E}\frac{1}{n}\psi(\mathbf{x}'^T)\mathbf{1}\|^2$. We analyze the convergence rate by deriving an upper bound for $\mathcal{D}(W, \hat{\alpha}) - \mathcal{D}(W, \alpha^*)$ with fixed Q and W.

Theorem 2. *Given learned $\hat{Q}$ and $\hat{W}$, let the induced RKHS be universal and upper bounded that $\|\psi(\tau(x))\| \leq \wedge_{\hat{W}}$ for all x in the source and target domains, and let the entries of G be bounded that $|G_{ij}| \leq \wedge_{\hat{Q}}$ for all $i \in \{1, \cdots, m\}, j \in \{1, \cdots, c\}$. $\forall \delta > 0$, with probability at least $1 - \delta$, we have*

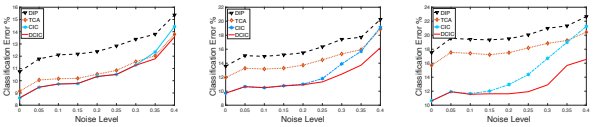
$$\mathcal{D}(\hat{W}, \hat{\alpha}) - \mathcal{D}(\hat{W}, \alpha^*)$$
$$\leq 8(\wedge_{\hat{Q}} + 1)^2 \wedge_{\hat{W}}^2 \sqrt{\frac{\sqrt{c}}{\sqrt{m}} + \frac{\sqrt{c}}{\sqrt{n}} + \sqrt{2(\frac{1}{m} + \frac{1}{n}) \log \frac{1}{\delta}}}.$$

See the proof of Theorem 2 in the Supplementary Material. Although the bound in Theorem 2 involves two fixed parameters, the result is informative if Q^* and W^* are given or $\hat{Q}$ and $\hat{W}$ quickly converges to Q^* and W^*, respectively. From previous analyses, we know that fast convergence rates for estimating label noise rate are guaranteed. However, the convergence of $\hat{W}$ to W^* is not guaranteed because the objective function is non-convex w.r.t. W. How to identify the transferable components $\tau(X)$ should be further studied.



(a) (b) (c)

Figure 2. The estimation error of β. (a), (b), and (c) present the estimate errors with the increasing class ratio $\beta(Y = 1)$, the increasing flip rate ρ, and the increasing sample size n, respectively.



(a) (b) (c)

Figure 3. The effectiveness of invariant components extraction. (a), (b), and (c) present the classification error with increasing flip rate ρ when $\beta_1 = 1.4$, 1.6, and 1.8, respectively.

5. Experiments

To show the robustness of our method to label noise, we conduct comprehensive evaluations on both simulated and real data. We first compare our method, denoising conditional invariant components (abbr. as DCIC hereafter), with CIC (Gong et al., 2016) on identifying the changes in P_Y given noisy observations. The effectiveness of our method is then verified on both synthetic and real data. We compare DCIC with the domain invariant projection (DIP) (Baktashmotlagh et al., 2013), transfer component analysis (TCA) (Pan et al., 2011), Deep Adaptation Networks (DAN) (Long et al., 2015) and CIC (Gong et al., 2016). In our experiments, the bandwidth σ of the Gaussian kernel is set to be the median value of the pairwise distances between all invariant (resp. raw) features for deep (resp. linear) model.

5.1. Synthetic Data

We use the linear model to verify the effectiveness of DCIC in two situations: (a) the estimation of class ratio β in the target shift (TarS) scenario given the true flip rates (i.e., transition probabilities); and (b) the evaluation of the extracted invariant components in the generalized target shift (GeTarS) scenario, with various class ratios and different label flip rates. In all experiments, the flip rates are estimated using the method proposed in (Liu & Tao, 2016). We repeat the experiments for 20 times and report the average scores.

We generate the binary classification training and test data from a 2-dimensional mixture of Gaussians (Gong et al., 2016), i.e., $x \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(\theta_i, \mathbf{\Sigma}_i)$ where the mean parameters $\theta_{ij}, j = 1, 2$ are sampled from the uniform distribution $\mathcal{U}(-0.25, 0.25)$ and the covariance matrices $\mathbf{\Sigma}_i$ are sampled from the Wishart distribution $\mathcal{W}(2 \times \mathbf{I}_2, 7)$. The class labels are the cluster indices. Under TarS, $P_{X|Y}$ remains the same. We only change the class priors across domains. Under GeTarS, we apply location and scale transformations on the features to generate target domain data. To get the noisy observations, we randomly flip the clean labels in the source domain with the same transition probability ρ.

First, we verify that with corrupted labels, the proposed DCIC can almost recover the correct class ratio under TarS. We set the source class prior $P^S(Y = 1)$ to 0.5. The target domain class prior $P^T(Y = 1)$ varies from 0.1 to 0.9 with step 0.1. The corresponding class ratio $\beta(Y = 1) = P^T(Y = 1)/P^S(Y = 1)$ varies from 0.2 to 1.8 with step 0.2. Then, we compare the proposed method with CIC (Gong et al., 2016) on finding the true class ratio β^* with noisy labels in source domain. We evaluate the performance by using the class ratio estimation error $\|\beta_{est} - \beta^*\|/\|\beta^*\|$, where β_{est} is the estimated class ratio vector. Figure 2(a) shows that DCIC can find the solutions close to the true β^* for various class ratios. In this experiment, given large label noise ($\rho = 0.4$), β estimated by CIC is close to the true one only when $\beta^*(Y = 1)$ is close to 0, 1, and 2. The estimation of CIC is accurate at $\beta^*(Y = 1) = 1$ because we set the class prior $P^S_{Y=1}$ to 0.5 in the clean source domain, which happens to make $P^S_{\rho Y} = P^S_Y$. If $P^S_{Y=1} \neq 0.5$, then $P^S_{\rho Y} \neq P^S_Y$, the estimated β will be wrong (see Section 3). CIC gives accurate results when $\beta^*(Y = 1)$ is close to 0, 2 because target domain collapses to a single class, rendering the estimated results trivially right. Figure 2(b) shows the superiority of the proposed method over CIC at different levels of label noise. When $\rho > 0.1$, CIC finds the incorrect solutions. However, our method can find a good solution even when ρ is close to 0.5. Figure 2(c) shows that the estimate of β improves as the sample size gets larger.

Second, under GeTarS, we evaluate whether our method can discover the invariant representations given the noisy source data and unlabeled target data. In these experiments, we fix the sample size to 500, and the class prior $P^S(Y = 1)$ to 0.5. We use classification accuracies to measure the performance. The results in Figure 3 show that our method is more robust to the label noise than DIP, TCA, and CIC.

5.2. Real Data

MNIST-USPS. USPS dataset is a handwritten digit dataset including ten classes 0-9 and contains 7,291 training images and 2,007 test images of size 16×16, which is rescaled to 28×28. MNIST shares the same 10 classes of digits which consist of 60,000 training images and 10,000 test images of size 28×28. In our experiments, these two datasets are resampled to construct the domain adaptation datasets in which the class priors P_Y across different domains vary. For MNIST, we assume that the class priors are unbalanced. For the first 5 classes, the class prior is set to 0.04. For the rest 5 classes, the class prior is equal to 0.16. For USPS, the class priors are balanced; that is, the class prior is set to 0.1 for each class. According to these class priors, we sample 5,000 images from both MNIST and USPS datasets to construct the new dataset mnist2usps. We switch the source/target pair to get another dataset usps2mnist. Same with (Patrini et al., 2017), in the source data, noise flips between the similar digits: $2 \to 7$, $3 \to 8$, $5 \leftrightarrow 6$, $7 \to 1$ with the transition probability $\rho = 0.2$ or 0.4. After the noisy data are obtained, we leave 10 percent of source data as validation set. The LeNet (LeCun et al., 1998) structure in Caffe's (Jia et al., 2014) MNIST tutorial is employed to train the model from scratch. Our denoising MMD loss is imposed on the first fully connected layer. In all experiments, l_2 regularization is applied and we set $\pi_1 = 1$ and $\pi_2 = 1e - 4$. The batch sizes for both source and target data are set to 100. The initial learning rate $r_0 = 0.01$ and is decayed exponentially according to $r_0(1 + 0.0001t)^{-0.75}$, where t is the index of current iteration. Each experiment is repeated 5 times.

Here, DCIC is compared with the baseline that training with source data only (SO), DAN, and CIC. These methods are integrated with the forward procedure in (Patrini et al., 2017) to reduce the effects of label noise. They are denoted as methods with "Forward Q (resp. $\hat{Q}$)" given the true (resp. estimated) transition matrix. Note that, DAN has verified that adapting more layers and using MK-MMD are more helpful. Here, we use single-layer adaptation and the modified vanilla MMD to compare with baselines, which futher verified the effectiveness of our method. We are also aware that DAN is for covariate shift problem, so we extended it to CIC to adress generalized target shift. CIC is also added in the first fully connected layer and the vanilla MMD loss is used. Further, the exploited CIC here is not the original one in (Gong et al., 2016) but the extension of DAN with idea from (Gong et al., 2016). The results are shown in Table 1. When label noise is present, CIC based methods cannot correctly estimate the class ratios, which adversely affects the identification of the invariant components. It thus performs worse than the DAN based methods in some cases. The latter, however, ignores the change of P_Y in different domains. In contrast, our method often gives better estimation of the class ratios and can effectively identify the invariant components, which leads to the higher performances.

VLCS. VLCS dataset (Torralba & Efros, 2011) consists of the images from five common classes: "bird", "car", "chair", "dog", and "person" in the datasets Pascal VOC 2007 (V),

Table 1. Classification accuracies and their standard deviations for USPS and MNIST datasets.

	mnist → usps ($\rho = 0.4$)	usps → mnist ($\rho = 0.4$)	mnist → usps ($\rho = 0.2$)	usps → mnist ($\rho = 0.2$)
SO+Forward Q	58.12 ± 0.32	61.02 ± 0.90	59.27 ± 1.51	65.90 ± 0.65
SO+Forward $\hat{Q}$	54.93 ± 2.23	60.80 ± 0.49	56.97 ± 1.36	65.51 ± 3.07
DAN+Forward Q	59.34 ± 5.43	64.68 ± 1.07	62.82 ± 1.15	67.05 ± 0.77
DAN+Forward $\hat{Q}$	54.76 ± 1.62	63.87 ± 0.84	61.28 ± 1.44	65.70 ± 1.24
CIC	65.23 ± 2.63	58.09 ± 2.17	66.70 ± 1.31	61.02 ± 3.96
CIC+Forward Q	65.37 ± 2.49	63.35 ± 4.43	66.84 ± 3.62	68.45 ± 0.91
CIC+Forward $\hat{Q}$	64.18 ± 1.49	62.78 ± 2.92	63.42 ± 0.99	67.99 ± 1.30
DCIC+Forward Q	$\mathbf{69.94 \pm 2.25}$	$\mathbf{68.77 \pm 2.34}$	$\mathbf{72.33 \pm 2.15}$	$\mathbf{70.80 \pm 1.59}$
DCIC+Forward $\hat{Q}$	$\mathbf{68.50 \pm 0.37}$	$\mathbf{66.78 \pm 1.53}$	$\mathbf{69.29 \pm 4.07}$	$\mathbf{70.47 \pm 2.29}$

Table 2. Classification accuracies and their standard deviations for VLCS dataset.

	VLS2C	LCS2V	VLC2S	VCS2L
SO+Forward Q	85.88 ± 2.17	62.07 ± 0.86	59.40 ± 1.37	49.34 ± 1.39
SO+Forward $\hat{Q}$	78.62 ± 4.36	59.49 ± 0.50	57.09 ± 1.81	49.14 ± 1.39
DAN+Forward Q	87.66 ± 2.37	64.37 ± 2.07	59.54 ± 0.83	51.07 ± 1.26
DAN+Forward $\hat{Q}$	84.69 ± 0.24	58.64 ± 1.91	57.51 ± 1.25	50.41 ± 1.20
CIC	75.15 ± 6.23	54.69 ± 0.96	53.61 ± 2.35	49.30 ± 0.48
CIC+Forward Q	86.83 ± 2.53	64.22 ± 0.27	60.36 ± 0.36	51.76 ± 0.82
CIC+Forward $\hat{Q}$	85.69 ± 1.76	59.80 ± 0.47	57.65 ± 0.60	50.33 ± 0.31
DCIC+Forward Q	$\mathbf{91.60 \pm 0.51}$	$\mathbf{65.67 \pm 0.37}$	$\mathbf{61.79 \pm 0.77}$	$\mathbf{52.47 \pm 0.50}$
DCIC+Forward $\hat{Q}$	87.28 ± 1.18	63.35 ± 0.37	58.88 ± 0.74	51.60 ± 1.48

LabelMe (L), Caltech (C), and SUN09 (S), respectively. For these four datasets, we first randomly select at most 300 images for each class to construct the new datasets, respectively. Then, we construct the domain adaptation datasets by using the leave-one-domain-out evaluation strategy. For example, in "VLS2C", the source data is the combination of the new Pascal VOC 2007, LabelMe, and SUN09 datasets. The target dataset is the new Caltech. In each source data, the labels flip from "person" to "car", "chair" to "person", and "dog" to "person" with the probability $\rho = 0.4$. We leave 30% of the source data as the validation set. Each experiment is repeated 5 times.

In this experiments, the source data is finetuned on the pretrained AlexNet (Krizhevsky et al., 2012) model with the parameters in conv1-conv3 layers being freezed. We impose our denoising MMD loss on the fc7 layer. As discussed in DAN, we also focus on high-level features because the transferability gap grows from low-level to high-level features, and that the gap becomes large in high-level ones. Further, high-level semantic features are more prone to be affected by wrong labels. However, if label noise possibly affects the low-level features, correcting the low-level features directly could be more powerful.

The batch sizes for both source and target data are 32. The initial learning rate is 0.001 and decayed exponentially according to $0.001(1 + 0.002t)^{-0.75}$. The results are shown in Table 2. Our proposed method also improves the performances of the compared baselines, which indicates the effectiveness of the proposed model to correct the shift in different domains even though the label noise is present.

6. Conclusion

We have studied domain adaptation with label noise. We found that label noise is detrimental to the performance of existing domain adaptation methods. In particular, when the label is the cause for the features, the estimate of target domain class distribution and conditional invariant representations can be unreliable. To alleviate the effects of label noise on domain adaptation, we have proposed the novel denoising MMD loss to improve the estimation of both target domain label distribution and conditional invariant components from the noisy source data and the unlabeled target data. We have provided both theoretical and empirical studies to demonstrate the effectiveness of our method.

Acknowledgements

This research was supported in part by Australian Research Council Projects, i.e., DE-190101473, FL-170100117, DP-180103424, IH-180100002, IC-190100031, and LE-200100049. This work was partially supported by NIH Award Number 1R01HL141813-01, NSF 1839332 Tripod+X, and SAP SE. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. We were also grateful for the computational resources provided by Pittsburgh Supercomputing grant number TG-ASC170024. Finally, thanks anonymous reviewers for their constructive comments.

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. Unsupervised label noise modeling and loss correction. In *ICML*. 2019.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *ICLR*, 2019.

B. Han, G. Niu, X. Y.-Q. Y. M. X. I. W. T. and Sugiyama, M. Sigua: Forgetting may make learning with noisy labels more robust. In *ICML*, 2020.

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M. Unsupervised domain adaptation by domain invariant projection. In *CVPR*, pp. 769–776, 2013.

Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., and Huang, J. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, June 2019a.

Chen, P., Liao, B., Chen, G., and Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*. 2019b.

Cheng, J., Liu, T., Ramamohanarao, K., and Tao, D. Learning with bounded instance-and label-dependent label noise. 2020.

Dubois, S., Romano, N., Jung, K., Shah, N., and Kale, D. C. The effectiveness of transfer learning in electronic health records data. In *ICLR Workship Track*, 2017.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *ICML*, pp. 2839–2848, 2016.

Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., and Sugiyama, M. Masking: A new perspective of noisy supervision. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *NIPS*, pp. 5836–5846. 2018a.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *NIPS*, pp. 8527–8537. 2018b.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NIPS*, pp. 8527–8537, 2018c.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. Correcting sample selection bias by unlabeled data. In *NIPS*, pp. 601–608, 2007.

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from complementary labels. In *NIPS*, pp. 5639–5649, 2017.

Iyer, A., Nath, J. S., and Sarawagi, S. Maximum Mean Discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *ICML*, pp. 530–538, 2014.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In *The 22nd ACM international conference on Multimedia (ACMMM)*, pp. 675–678. ACM, 2014.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*. 2018.

Jiaxian Guo, Mingming Gong, T. L.-K. Z. and Tao, D. Ltf: A label transformation framework for correcting label shift. In *ICML*, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105, 2012.

Kumagai, A., Iwata, T., and Fujiwara, Y. Transfer anomaly detection by inferring latent domain representations. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (eds.), *NIPS*, pp. 2467–2477. 2019.

L. Feng, T. Kaneko, B. H.-G. N. B. A. and Sugiyama, M. Learning from multiple complementary labels. In *ICML*, 2020.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, K.-H., He, X., Zhang, L., and Yang, L. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, 2018.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

Liu, Y. and Guo, H. Peer loss functions: Learning from noisy labels without knowing noise rates. 2020.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. In *ICML*, pp. 97–105, 2015.

Long, M., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. In *ICML*, pp. 608–615, 2008.

Menon, A., Van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *ICML*, pp. 125–134, 2015.

Meyerson, E. and Miikkulainen, R. Modular universal reparameterization: Deep multi-task learning across diverse domains. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (eds.), *NIPS*, pp. 7901–7912. 2019.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *NIPS*, pp. 1196–1204, 2013.

Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., and Brox, T. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2020.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

Pearl, J. and Bareinboim, E. Transportability of causal and statistical relations: A formal approach. In *AISTATS*, pp. 247–254, 2011.

Purushotham, S., Carvalho, W., Nilanon, T., and Liu, Y. Variational recurrent adversarial deep domain adaptation. In *ICLR*, 2017.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (eds.), *NIPS*, pp. 3342–3352. 2019.

Ramaswamy, H., Scott, C., and Tewari, A. Mixture Proportion Estimation via kernel embeddings of distributions. In *ICML*, pp. 2052–2060, 2016.

Reid, M. D. and Williamson, R. C. Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422, 2010.

Sáez, J. A., Krawczyk, B., and Woźniak, M. On the influence of class noise in medical data classification: Treatment using noise filtering methods. *Applied Artificial Intelligence*, 30(6):590–609, 2016.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *ICML*, 2012a.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012b.

Scott, C. A rate of convergence for Mixture Proportion Estimation, with application to learning from noisy labels. In *AISTATS*, pp. 838–846, 2015.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. Combating label noise in deep learning using abstention. In *ICML*. 2019.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR*, pp. 1521–1528. IEEE, 2011.

Van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In *NIPS*, pp. 10–18, 2015.

Wang, X., Huang, T.-K., and Schneider, J. G. Active transfer learning under model shift. In *ICML*, pp. 1305–1313, 2014.

Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., Liu, H., and Niu, G. Class2simi: A new perspective on learning with label noise. *arXiv preprint arXiv:2006.07831*, 2020.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. Domain adaptation with asymmetrically-relaxed distribution alignment. In *ICML*, 2019.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *NIPS*, pp. 6838–6849, 2019.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Parts-dependent label noise: Towards instance-dependent label noise. *arXiv preprint arXiv:2006.07836*, 2020.

Xu, Y., Cao, P., Kong, Y., and Wang, Y. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (eds.), *NIPS*, pp. 6222–6233. 2019a.

Xu, Y., Gong, M., Chen, J., Liu, T., Zhang, K., and Batmanghelich, K. Generative-discriminative complementary learning. *arXiv preprint arXiv:1904.01612*, 2019b.

Xu, Z., Huang, S., Zhang, Y., and Tao, D. Webly-supervised fine-grained visual categorization via deep domain adaptation. *TPAMI*, 40(5):1100–1113, 2016.

Y.-T. Chou, G. Niu, H.-T. L. and Sugiyama, M. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *ICML*, 2020.

Yang, H., Yao, Q., Han, B., and Niu, G. Searching to exploit memorization effect in learning from corrupted labels. *arXiv preprint arXiv:1911.02377*, 2019.

Yao, Y., Liu, T., Han, B., Gong, M., Niu, G., Sugiyama, M., and Tao, D. Towards mixture proportion estimation without irreducibility. *arXiv preprint arXiv:2002.03673*, 2020.

Yu, X., Liu, T., Gong, M., Batmanghelich, K., and Tao, D. An efficient and provable approach for mixture proportion estimation using linear independence assumption. In *CVPR*, 2018a.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *ECCV*, pp. 68–83, 2018b.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *ICML*, pp. 819–827, 2013a.

Zhang, K., Zheng, V., Wang, Q., Kwok, J., Yang, Q., and Marsic, I. Covariate shift in hilbert space: A solution via sorrogate kernels. In *ICML*, pp. 388–395, 2013b.