Adaptive Knowledge Transfer on Evolving Domains

Jun Wu, Hanghang Tong, Elizabeth Ainsworth, Jingrui He
University of Illinois Urbana-Champaign
{junwu3,htong,ainswort,jingrui}@illinois.edu

***Abstract*—In this paper, we study the dynamic transfer learning problem involving adaptive knowledge transfer from a static source domain to a time evolving target domain. One major challenge is the time evolving relatedness of the source domain and the current target domain as the target domain evolves over time. To address this challenge, we derive a generic error bound on the current target domain with flexible domain discrepancy measures. Moreover, we propose a label-informed $\mathcal{C}$-divergence to measure the shift of joint data distributions (over input features and output labels) across domains. The resulting tighter error bound with $\mathcal{C}$-divergence motivates us to develop a novel dynamic transfer learning algorithm `TransLATE`. Empirical results on various data sets confirm the effectiveness of our proposed algorithm in modeling the time evolving target domain.**

***Index Terms*—transfer learning, evolving domain, distribution discrepancy**

I. Introduction

Transfer learning [12], [16] has achieved significant success across multiple high-impact applications. It improves the prediction performance on a target domain with limited label information by leveraging the knowledge from a related source domain with abundant label information [3], [19]. However, in many real applications, the target domain is constantly evolving over time [14]. For example, the online movie reviews are changing over the years: some famous movies were not well received by the mainstream audience when they were first released, but became popular only years later (e.g., *Citizen Cane*, *Fight Club*, and *The Shawshank Redemption*); whereas the online book reviews typically do not bear this type of dynamics. Another example is regarding high-throughput plant phenotyping [18]: the relationship between the spectral reflectance of a leaf and some physiological/biochemical traits is smoothly changing with respect to the growing season; whereas its relationship with other traits might be relatively stable over time. It is challenging to transfer knowledge from the static source domain (e.g., the book reviews) to the time evolving target domain (e.g., the movie reviews).

Therefore, in this paper, we study the dynamic transfer learning from a static source domain to a continuously time evolving target domain [4], [6], [14], which has not attracted adequate attention from the research community and yet is commonly seen across many real applications. The unique challenge for dynamic transfer learning lies in the time evolving nature of the task relatedness between the static source domain and the time evolving target domain. Even though the change in the target data distribution in consecutive time stamps might be small, over time, the cumulative change in the target domain is likely to be significant and might even lead to negative transfer [15].

Existing theoretical analysis on transfer learning [1], [10] has revealed that the target error is typically bounded by the source error, the domain discrepancy of marginal data distributions, and the difference of labeling functions. However, it has been pointed out [20] that marginal feature distribution alignment might not guarantee the success of knowledge transfer in real world scenarios. This indicates that in the context of adaptive knowledge transfer on evolving domains, marginal feature distribution alignment would lead to the sub-optimal solution or even negative transfer [15], with undesirable prediction performance when directly transferring from the source domain $\mathcal{D}_S$ to the target domain $\mathcal{D}_{T_t}$ at the t^{th} time stamp. This paper aims to bridge the gap in terms of both the theoretical analysis and the empirical solutions for the target domain with a time evolving distribution. The main contributions of this paper are summarized as follows.

- A generic generalization error bound for adaptive knowledge transfer on evolving domains is derived with flexible domain divergence measures.
- We propose a label-informed domain discrepancy measure ($\mathcal{C}$-divergence) with its empirical estimate, which instantiates a tighter error bound for adaptive knowledge transfer.
- We design a novel adversarial Variational Auto-encoder algorithm (`TransLATE`) by empirically minimizing the $\mathcal{C}$-divergence based error upper bound.
- Experiments on various data sets verify the effectiveness of the proposed `TransLATE` algorithm.

The rest of the paper is organized as follows. The related work is summarized in Section II. Then we first introduce the problem definition in Section III, and we derive a generic error bound for dynamic transfer learning on evolving domains in Section IV. Formally, we propose a novel $\mathcal{C}$-divergence in Section V, followed by an instantiated error bound and a novel dynamic transfer learning algorithm in Section VI. Extensive experiments are provided in Section VII. Finally, we conclude the paper in Section VIII.

II. Related Work

Transfer learning [12], [21]–[23] improves the performance of a learning algorithm on the target domain by using the knowledge from the source domain. It is theoretically proven that the target error is well bounded [1], [19] when the source and target domains share the same label space (a.k.a. domain adaptation), followed by a wealth of practical algorithms [3], [9]. More recently, it is studied in the dynamic setting [4], [6], [8] where the data distribution of the target domain is evolving over time. The main challenge of transfer learning lies in the distribution shift across the source and target domains. It is notable that the distribution shift across domains can be measured under different assumptions, e.g., covariate shift assumption [12], label shift [7], etc. In this paper, we assume that the joint distribution over both input features and output labels would be shifted across domains and across time stamps. It is studied in previous work [11] by incorporating the label information for measuring the distribution shift across domains. Different from those works, our proposed $\mathcal{C}$-divergence is derived from the perspective of measurable set matching, thus shedding light on the empirical estimate of label-informed domain discrepancy from finite samples in practice. Moreover, we estimate the $\mathcal{C}$-divergence in a unified framework through the label-informed hidden representation. This is in sharp contrast to previous works which estimate the label-aware discrepancy over the label or conditional shift, or both (e.g., $p(x)$ and $p(y|x)$ in [7]).

III. Problem Setting

We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the input space and label space. Let $\mathcal{D}_S$ and $\mathcal{D}_T$ denote the source and target domains with data distribution $p_S(\mathbf{x}, y)$ and $p_T(\mathbf{x}, y)$ over $\mathcal{X} \times \mathcal{Y}$, respectively. When the target domain is evolving over time, we denote $\mathcal{D}_{T_j}$ to be the target domain at the j^{th} time stamp. Let $\mathcal{H}$ be a hypothesis class on $\mathcal{X}$, where a hypothesis is a function $h : \mathcal{X} \to \mathcal{Y}$. Following [4], [6], we formally define the problem of dynamic transfer learning as follows.

Definition 1. *Given a source domain $\mathcal{D}_S$ (available at time stamp $j = 1$) and a time evolving target domain $\{\mathcal{D}_{T_j}\}_{j=1}^t$ with time stamp j, dynamic transfer learning aims to improve the prediction function for target domain $\mathcal{D}_{T_{t+1}}$ using the knowledge from source domain $\mathcal{D}_S$ and the historical target domain $\mathcal{D}_{T_j} (j = 1, \cdots, t)$.*

Notice that the source domain $\mathcal{D}_S$ can be considered a special initial time stamp for the time-evolving target domain. Therefore, for notation simplicity, we will use $\mathcal{D}_{T_0}$ to represent the source domain in this paper. We assume that there are m_{T_0} labeled source examples drawn independently from a source domain $\mathcal{D}_{T_0}$ and m_{T_j} labeled target examples drawn independently from a target domain $\mathcal{D}_{T_j}$ at time stamp j.

IV. A Generic Error Bound

Given a static source domain and a time evolving target domain, dynamic transfer learning aims to improve the target prediction function over $\mathcal{D}_{T_{t+1}}$ using the source domain and historical target domain. We begin by considering the binary classification setting, i.e., $\mathcal{Y} = \{0, 1\}$. The source error of a hypothesis h can be defined as follows: $\epsilon_{T_0}(h) = \mathbb{E}_{(\mathbf{x},y)\sim p_{T_0}(\mathbf{x},y)}\left[\mathcal{L}(h(\mathbf{x}), y)\right]$ where $\mathcal{L}(\cdot,\cdot)$ is the loss function. Its empirical estimate using source labeled examples is denoted as $\hat{\epsilon}_{T_0}(h)$. Similarly, we define the target error $\epsilon_{T_j}(h)$ and the empirical estimate of the target error $\hat{\epsilon}_{T_j}(h)$ over the target distribution $p_{T_j}(\mathbf{x}, y)$ at time stamp j. A natural domain discrepancy measure over joint distributions on $\mathcal{X} \times \mathcal{Y}$ between features and class labels can be defined as follows:

$$d_1(\mathcal{D}_{T_0}, \mathcal{D}_T) = \sup_{Q \in \mathcal{Q}} \left|\Pr_{\mathcal{D}_{T_0}}[Q] - \Pr_{\mathcal{D}_T}[Q]\right| \qquad (1)$$

where $\mathcal{Q}$ is the set of measurable subsets under $p_{T_0}(\mathbf{x}, y)$ and $p_T(\mathbf{x}, y)$. Then, the error bound of dynamic transfer learning is given by the following theorem.

Theorem 2. *Assume the loss function $\mathcal{L}$ is bounded with $0 \le \mathcal{L} \le M$. Given a source domain $\mathcal{D}_{T_0}$ and historical target domain $\{\mathcal{D}_{T_i}\}_{i=1}^t$, for $h \in \mathcal{H}$, the target domain error $\epsilon_{T_{t+1}}$ on $\mathcal{D}_{T_{t+1}}$ is bounded as follows.*

$$\epsilon_{T_{t+1}}(h) \le \frac{1}{\bar{\mu}} \left(\sum_{j=0}^{t} \mu^{t-j} \epsilon_{T_j}(h) + M \sum_{j=0}^{t} \mu^{t-j} d_1(\mathcal{D}_{T_j}, \mathcal{D}_{T_{t+1}}) \right)$$

where $\mu \ge 0$ is a hyper-parameter[1] indicating the importance of source or historical target domain, and $\bar{\mu} = \sum_{j=0}^t \mu^{t-j}$.

In particular, we have the following arguments. (1) It is not tractable to accurately estimate d_1 from finite examples in real scenarios [1]; (2) This error bound could be much tighter when considering other advanced domain discrepancy measures, e.g., $\mathcal{A}$-distance [1], discrepancy distance [10], etc. (3) There are two special cases: when $\mu = 0$, the error bound of $\mathcal{D}_{T_{t+1}}$ would be simply determined by the newest historical target data $\mathcal{D}_{T_t}$, and on the other hand, if μ goes to infinity, $\mathcal{D}_{T_{t+1}}$ is largely determined by the source data $\mathcal{D}_{T_0}$ because intuitively the coefficient $\mu^{t-j}/\bar{\mu}$ of historical target domain data $\mathcal{D}_{T_j} (j = 1, \cdots, t)$ converges to zero.

Corollary 3. *With the assumption in Theorem 2 and assume the loss function $\mathcal{L}$ is symmetric (i.e., $\mathcal{L}(y_1, y_2) = \mathcal{L}(y_2, y_1)$ for $y_1, y_2 \in \mathcal{Y}$) and obeys the triangle inequality, then*

(1) if $\mathcal{A}$-distance [1] is adopted to measure the distribution shift, i.e., $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{T_0}, \mathcal{D}_T) = \sup_{h,h' \in \mathcal{H}} |\Pr_{\mathcal{D}_{T_0}}[h(\mathbf{x}) \ne h'(\mathbf{x})] - \Pr_{\mathcal{D}_T}[h(\mathbf{x}) \ne h'(\mathbf{x})]|$, we have:

$$\epsilon_{T_{t+1}}(h) \le \frac{1}{\bar{\mu}} \Big(\sum_{j=0}^{t} \mu^{t-j} \epsilon_{T_j}(h) + M \sum_{j=0}^{t} \mu^{t-j} \big(d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{T_j}, \mathcal{D}_{T_{t+1}}) + \frac{\lambda_j^*}{M} \big) \Big)$$

where $\lambda_j^* = \min_{h \in \mathcal{H}} \epsilon_{T_j}(h) + \epsilon_{T_{t+1}}(h)$.

(2) if discrepancy distance [10] is adopted, i.e., $d_{disc}(\mathcal{D}_{T_0}, \mathcal{D}_T) = \max_{h,h' \in \mathcal{H}} |\mathbb{E}_{\mathcal{D}_{T_0}}[\mathcal{L}(h(x), h'(x))] - \mathbb{E}_{\mathcal{D}_T}[\mathcal{L}(h(x), h'(x))]|$, we have:

$$\epsilon_{T_{t+1}}(h) \le \frac{1}{\bar{\mu}} \Big(\sum_{j=0}^{t} \mu^{t-j} \epsilon_{T_j}(h) + \sum_{j=0}^{t} \mu^{t-j} \big(d_{disc}(\mathcal{D}_{T_j}, \mathcal{D}_{T_{t+1}}) + \Omega_j \big) \Big)$$

where $\Omega_j = \mathbb{E}_{\mathcal{D}_{T_{t+1}}}[\mathcal{L}(h_j^*(\mathbf{x}), h_{t+1}^*(\mathbf{x}))] + \mathbb{E}_{\mathcal{D}_{T_j}}[\mathcal{L}(h_j^*(\mathbf{x}), y)] + \mathbb{E}_{\mathcal{D}_{T_{t+1}}}[\mathcal{L}(h_{t+1}^*(\mathbf{x}), y)]$, and $h_j^* = \arg\min_{h \in \mathcal{H}} \epsilon_{T_j}(h)$.

Remark: It is notable that (1) Corollary 3 would naturally degenerate to the standard error bounds in the static transfer learning setting [1], [10] with $t = 0$, and (2) in the special case where $\mu = 1/(t+1)$, the derived error bound with $\mathcal{A}$-distance in Corollary 3 coincides with previous work [8] in terms of source classification error and domain discrepancy between source domain and target domain at every time stamp.

[1] In this case, we assume $\mu^0 = 1$ for any $\mu \ge 0$.

The aforementioned domain discrepancy measures mainly focus on the marginal distribution over input features and have inspired a line of practical transfer learning algorithms [3], [19]. However, recent work [17], [20] pointed out that the minimization of marginal distributions cannot always guarantee the success of transfer learning in real scenarios. We propose to address this problem by incorporating the label information in the domain discrepancy measure in the next section.

V. Label-informed Domain Discrepancy

In this section, we introduce a novel label-informed domain discrepancy measure between the source domain $\mathcal{D}_{T_0}$ and target domain $\mathcal{D}_T$ and its empirical estimate.

A. $\mathcal{C}$-divergence

For a hypothesis $h \in \mathcal{H}$, we denote $I(h)$ to be the subset of $\mathcal{X}$ such that $\mathbf{x} \in I(h) \Leftrightarrow h(\mathbf{x}) = 1$. In order to estimate the label-informed domain discrepancy from finite samples in practice, instead of Eq. (1), we propose the following $\mathcal{C}$-divergence between $\mathcal{D}_{T_0}$ and $\mathcal{D}_T$, taking into consideration the joint distribution over features and class labels:

$$\begin{aligned} d_{\mathcal{C}}(\mathcal{D}_{T_0}, \mathcal{D}_T) = \sup_{h \in \mathcal{H}} \Big| &\mathrm{Pr}_{\mathcal{D}_{T_0}}[\{I(h), y = 1\} \cup \{\overline{I(h)}, y = 0\}] \\ &- \mathrm{Pr}_{\mathcal{D}_T}[\{I(h), y = 1\} \cup \{\overline{I(h)}, y = 0\}]\Big| \end{aligned} \tag{2}$$

where $\overline{I(h)}$ is the complement of $I(h)$.

We show that several existing domain discrepancy methods (e.g., [1]) can be seen as special cases of this definition by using the following relaxed covariate shift assumption.

Definition 4. *(Relaxed Covariate Shift Assumption) The source and target domains satisfy the relaxed covariate shift assumption if for any $h \in \mathcal{H}$,*

$$Pr_{\mathcal{D}_{T_0}}[y \mid I(h)] = Pr_{\mathcal{D}_T}[y \mid I(h)] = Pr[y \mid I(h)] \tag{3}$$

Notice that it would be equivalent to the covariance shift assumption [13] when $I(h)$ consists of only one example for all $h \in \mathcal{H}$.

Lemma 5. *With the relaxed covariate shift assumption, for any $h \in \mathcal{H}$, we have:*

$$\begin{aligned} d_{\mathcal{C}}(\mathcal{D}_{T_0}, \mathcal{D}_T) = \sup_{h \in \mathcal{H}} \Big| &\Big(Pr_{\mathcal{D}_{T_0}}[I(h)] - Pr_{\mathcal{D}_T}[I(h)]\Big) \cdot \mathcal{S}_h \\ &+ Pr_{\mathcal{D}_T}[y = 1] - Pr_{\mathcal{D}_{T_0}}[y = 1]\Big| \end{aligned}$$

where $\mathcal{S}_h = Pr[y = 1 | I(h)] - Pr[y = 0 | I(h)]$.

Remark: From Lemma 5, we can see that in the special case where $\mathcal{S}_h$ is a constant for all $h \in \mathcal{H}$ and $\mathrm{Pr}_{\mathcal{D}_T}[y = 1] = \mathrm{Pr}_{\mathcal{D}_{T_0}}[y = 1]$, the proposed $\mathcal{C}$-divergence degenerates to the prevalent $\mathcal{A}$-distance [1] defined on the marginal distribution of features. Generally speaking, $\mathcal{C}$-divergence can be considered as a weighted version of the $\mathcal{A}$-distance where the hypothesis whose characteristic function has a larger class-separability (i.e., $|\mathcal{S}_h|$) receives a higher weight. Intuitively, compared to $\mathcal{A}$-distance, $\mathcal{C}$-divergence pays less attention to class-inseparable regions in the input feature space, which provide irrelevant information for learning the prediction function in the target domain.

B. Empirical Estimate of $\mathcal{C}$-divergence

In practice, it is difficult to calculate the proposed $\mathcal{C}$-divergence based on Eq. (2) as it uses the true underlying distributions. Therefore, we propose an empirical estimate of the $\mathcal{C}$-divergence between $\mathcal{D}_{T_0}$ and $\mathcal{D}_T$ as follows. Assuming that the hypothesis class $\mathcal{H}$ is symmetric (i.e., $1 - h \in \mathcal{H}$ if $h \in \mathcal{H}$), the empirical $\mathcal{C}$-divergence is:

$$\begin{aligned} d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_0}, \hat{\mathcal{D}}_T) = 1 - \min_{h \in \mathcal{H}} \Big| &\frac{1}{m_{T_0}} \sum_{(\mathbf{x},y):h(\mathbf{x}) \neq y} \mathbb{I}[(\mathbf{x}, y) \in \hat{\mathcal{D}}_{T_0}] \\ &+ \frac{1}{m_T} \sum_{(\mathbf{x},y):h(\mathbf{x}) = y} \mathbb{I}[(\mathbf{x}, y) \in \hat{\mathcal{D}}_T]\Big| \end{aligned} \tag{4}$$

where $\hat{\mathcal{D}}_{T_0}$ and $\hat{\mathcal{D}}_T$ denote the source and target domains with finite samples, respectively. $\mathbb{I}[a]$ is the binary indicator function which is 1 if a is true, and 0 otherwise.

The following lemma provides the upper bound of the true $\mathcal{C}$-divergence using its empirical estimate.

Lemma 6. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over m_{T_0} labeled source examples $\mathcal{B}_{T_0}$ and m_T labeled target examples $\mathcal{B}_T$, we have:*

$$\begin{aligned} d_{\mathcal{C}}(\mathcal{D}_{T_0}, \mathcal{D}_T) \leq d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_0}, \hat{\mathcal{D}}_T) &+ \Big(\hat{\mathfrak{R}}_{\mathcal{B}_{T_0}}(L_H) + \hat{\mathfrak{R}}_{\mathcal{B}_T}(L_H)\Big) \\ &+ 3\left(\sqrt{\frac{\log \frac{4}{\delta}}{2m_{T_0}}} + \sqrt{\frac{\log \frac{4}{\delta}}{2m_T}}\right) \end{aligned}$$

where $\hat{\mathfrak{R}}_{\mathcal{B}}(L_H)(\mathcal{B} \in \{\mathcal{B}_{T_0}, \mathcal{B}_T\})$ denotes the Rademacher complexity [10] over $\mathcal{B}$ and $L_H = \{(\mathbf{x}, y) \to \mathbb{I}[h(\mathbf{x}) = y] : h \in \mathcal{H}\}$ is a class of functions mapping $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to $\{0, 1\}$.

VI. Proposed Algorithm

In this section, we derive the error bound of dynamic transfer learning based on our proposed $\mathcal{C}$-divergence, followed by a novel knowledge transfer algorithm (`TransLATE`).

A. Error Bound

The following theorem states that for a bounded loss function $\mathcal{L}$, the expected error of the newest target domain can be bounded in terms of the empirical classification error within the source and historical target domains, the empirical $\mathcal{C}$-divergence across domains as well as the empirical Rademacher complexity of the class of functions $L_H = \{(\mathbf{x}, y) \to \mathbb{I}[h(\mathbf{x}) = y] : h \in \mathcal{H}\}$.

Theorem 7. *Assume the loss function $\mathcal{L}$ is bounded with $0 \leq \mathcal{L} \leq M$. Given a source domain $\mathcal{D}_{T_0}$ and historical target domain $\{\mathcal{D}_{T_i}\}_{i=1}^t$, for $h \in \mathcal{H}$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, the target domain error $\epsilon_{T_{t+1}}$ on $\mathcal{D}_{T_{t+1}}$ is bounded as follows.*

$$\epsilon_{T_{t+1}}(h) \leq \frac{1}{\bar{\mu}} \left(\sum_{j=0}^{t} \mu^{t-j} \hat{\epsilon}_{T_j}(h) + M \sum_{j=0}^{t} \mu^{t-j} d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_j}, \hat{\mathcal{D}}_{T_{t+1}}) + M\Lambda\right)$$

where $\Lambda = \sum_{j=0}^{t} \mu^{t-j} \Big(\hat{\mathfrak{R}}_{\mathcal{B}_{T_j}}(L_H) + \hat{\mathfrak{R}}_{\mathcal{B}_{T_{t+1}}}(L_H) + 3\sqrt{\frac{\log \frac{8(t+1)}{\delta}}{2m_{T_j}}} + 3\sqrt{\frac{\log \frac{8(t+1)}{\delta}}{2m_{T_{t+1}}}} + \sqrt{\frac{\log \frac{4(t+1)}{\delta}}{2m_{T_j}}}\Big)$.

Remark: Compared to error bounds in Corollary 3 using existing domain divergence measures ([1], [10]), our bound consists of only data-dependent terms (e.g., empirical source error and $\mathcal{C}$-divergence), whereas previous error bounds are determined by the error terms involving the intractable labeling function or optimal target hypothesis (see Corollary 3).

B. TransLATE *Algorithm*

For dynamic transfer learning, we leverage both the source domain and historical target domain data to learn the prediction function for the current time stamp. To this end, we propose to minimize the error bound in Theorem 7 for learning the prediction function on $\mathcal{D}_{T_{t+1}}$. Furthermore, we learn a domain-invariant and time-invariant latent space such that the $\mathcal{C}$-divergence across domains and across time stamps could be minimized. To this end, we present an adversarial Variational Auto-encoder (VAE) algorithm with the objective function:

$$\mathcal{J}(T_0, T_1, T_2, \cdots, T_{t+1}) = \mathcal{L}_{clc}\left(T_{t+1}\right) + \sum_{j=0}^{t} \mu^{t-j} \big(\mathcal{L}_{clc}\left(T_j\right) + d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_j}, \hat{\mathcal{D}}_{T_{t+1}}) + \lambda \mathcal{L}_{ELBO}\left(T_j, T_{t+1}\right) \big) \tag{5}$$

where $\mathcal{L}_{clc}(T_j)$ represents the classification error over the labeled examples from $\mathcal{D}_{T_j}$, $d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_j}, \hat{\mathcal{D}}_{T_{t+1}})$ is the empirical estimate of $\mathcal{C}$-divergence across domains. Thus, the first term of Eq. (5) is the supervised classification error when there is a limited number of labeled examples in the target domain $\mathcal{D}_{T_{t+1}}$. The second and third terms are associated with $\hat{\epsilon}_{T_j}(h) + d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_j}, \hat{\mathcal{D}}_{T_{t+1}})$ in the error bound of Theorem 7. The third term $\mathcal{L}_{ELBO}(T_j, T_{t+1})$ is the variational bound in the VAE framework when learning the latent feature space and $\lambda > 0$ is a hyper-parameter. Note that in our framework, the variational term $\mathcal{L}_{ELBO}(\cdot, \cdot)$ aims to learn a label-informed feature representation for each example such that our $\mathcal{C}$-divergence could then be empirically estimated from the label-informed features across domains. In this case, we have $\mu \in [0, 1]$ because we assume that the data distribution of the target domain shifts smoothly over time.

Inspired by semi-supervised VAE [5], we propose to learn the feature space by maximizing the following likelihood.

$$\log p_\theta(\mathbf{x}, y) = \mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}, y) || p_\theta(\mathbf{z}|\mathbf{x}, y)\big) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},y)}[\log p_\theta(\mathbf{x}, y, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}, y)] \tag{6}$$

where ϕ and θ are the trainable parameters in the encoder and decoder respectively, and $\mathbf{z}$ is the latent feature representation of the input example $(\mathbf{x}, y)$. $\mathrm{KL}(\cdot||\cdot)$ is Kullback–Leibler divergence. The evidence lower bound (ELBO), a lower bound on this log-likelihood, can be written as follows.

$$\mathcal{E}_{\theta,\phi}(\mathbf{x}, y) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},y)}\left[\log p_\theta(\mathbf{x}, y|\mathbf{z})\right] + \mathrm{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}, y)||p(\mathbf{z})\right)$$

where $\mathcal{E}_{\theta,\phi}(\mathbf{x}, y) \leq \log p_\theta(\mathbf{x}, y)$. Similarly, we have the following ELBO to maximize the log-likelihood of $p_\theta(\mathbf{x})$ when the label is not available:

$$\mathcal{U}_{\theta,\phi}(\mathbf{x}) = \textstyle\sum_y \left(q_\phi(y|\mathbf{x}) \cdot \mathcal{E}_{\theta,\phi}(\mathbf{x}, y) - \mathbb{E}_{q_\phi(y|\mathbf{x})}\left[\log q_\phi(y|\mathbf{x})\right]\right) \tag{7}$$

where $p_\theta(\mathbf{x}, y, \mathbf{z}) = p_\theta(\mathbf{x}|y, \mathbf{z}) p_\theta(y|\mathbf{z}) p(\mathbf{z})$ with the prior Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, the ***variational bound*** $\mathcal{L}_{ELBO}(T_j, T_{t+1})$ is given below.

$$\mathcal{L}_{ELBO}(T_j, T_{t+1}) = - \sum_{i=1}^{m_{T_j} + m_{T_{t+1}}} \mathcal{E}_{\theta,\phi}(\mathbf{x}_i, y_i) - \sum_{i=1}^{u_{T_{t+1}}} \mathcal{U}_{\theta,\phi}(\mathbf{x}_i) \tag{8}$$

where $u_{T_{t+1}}$ is the number of unlabeled training examples from $\mathcal{D}_{T_{t+1}}$. Besides, the ***classification error*** $\mathcal{L}_{clc}(T_j, T_{t+1})$ can be expressed as follows.

$$\mathcal{L}_{clc}(T_j, T_{t+1}) = \textstyle\sum_{i=1}^{m_{T_j} + m_{T_{t+1}}} \mathcal{L}\left(y_i, q_\phi(\cdot|\mathbf{x}_i)\right) \tag{9}$$

where $q_\phi(\cdot)$ is the discriminative classifier formed by the distribution $q_\phi(y|\mathbf{x})$ in Eq. (7), and $\mathcal{L}(\cdot, \cdot)$ is the cross-entropy loss function in our experiments. To estimate the $\mathcal{C}$-divergence, we first define $\tilde{h}$ to be a two-dimensional characteristic function with $\tilde{h}(\mathbf{x}, y) = 1 \Leftrightarrow h(\mathbf{x}) = y \Leftrightarrow \{h(\mathbf{x}) = 1, y = 1\} \vee \{h(\mathbf{x}) = 0, y = 0\}$ for $h \in \mathcal{H}$. Then the empirical $\mathcal{C}$-divergence in Eq. (4) can be rewritten as follows.

$$d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_j}, \hat{\mathcal{D}}_{T_{t+1}}) = 1 - \min_{\tilde{h}} \Big| \frac{1}{m_{T_j}} \sum_{(\mathbf{x},y):\tilde{h}(\mathbf{x},y)=0} \mathbb{I}[(\mathbf{x}, y) \in \hat{\mathcal{D}}_{T_j}] + \frac{1}{m_{T_{t+1}}} \sum_{(\mathbf{x},y):\tilde{h}(\mathbf{x},y)=1} \mathbb{I}[(\mathbf{x}, y) \in \hat{\mathcal{D}}_{T_{t+1}}] \Big|$$

Note that the feature representation $\mathbf{z}$ learned by $q_\phi(\mathbf{z}|\mathbf{x}, y)$ captures both input feature and output label information of an example $(\mathbf{x}, y)$. Thus, the hypothesis $\tilde{h}$ can be considered as the composition of a feature extraction q_ϕ and a domain classifier $\mathcal{F}_j$, i.e, $\tilde{h}(\mathbf{x}, y) = \mathcal{F}_j(q_\phi(\mathbf{z}|\mathbf{x}, y))$. Formally, the ***empirical estimate of $\mathcal{C}$-divergence*** is given below.

$$d_{\mathcal{C}}(\hat{\mathcal{D}}_{T_j}, \hat{\mathcal{D}}_{T_{t+1}}) = 1 - \min_{\mathcal{F}_j} \Big| \frac{1}{m_{T_j}} \sum_{\mathbf{z}:\mathcal{F}_j(\mathbf{z})=0} \mathbb{I}[\mathbf{z} \in \hat{\mathcal{D}}_{T_j}] + \frac{1}{m_{T_{t+1}}} \sum_{\mathbf{z}:\mathcal{F}_j(\mathbf{z})=1} \mathbb{I}[\mathbf{z} \in \hat{\mathcal{D}}_{T_{t+1}}] \Big| \tag{10}$$

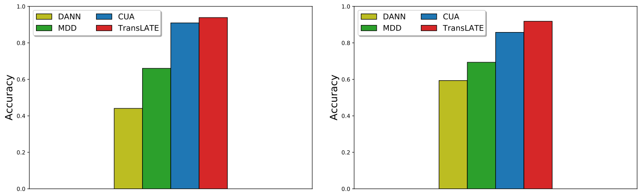
The benefits of TransLATE algorithm are twofold: first, it learns the latent feature space using both input $\mathbf{x}$ and output y; second, it minimizes a tighter error upper bound based on $\mathcal{C}$-divergence in Theorem 7. This algorithm can also be interpreted as a minimax game: VAE learns a domain-invariant and time-invariant feature space, whereas the domain classifier $\mathcal{F}_j$ distinguishes the examples from different domains and different time stamps. In this paper, we adopt the gradient reversal layer [3] when updating the parameters of domain classifier $\mathcal{F}_j$.

We further address two subtle issues. To be specific, we observe that (1) it is difficult to estimate the $\mathcal{C}$-divergence with only limited labeled target examples from $\mathcal{D}_{T_{t+1}}$; (2) when learning the latent features $\mathbf{z}$, combining the data $\mathbf{x}$ (e.g., one image) and class-label y directly might lead to over-emphasizing the data itself due to its high dimensionality compared to y. To address these problems, we propose the following *Pseudo-label Inference*, i.e., we infer the pseudo labels of unlabeled examples using the classifier $q_\phi(y|\mathbf{x})$ for each training epoch. Using labeled source and target examples as well as unlabeled target examples with inferred pseudo labels, $\mathcal{C}$-divergence is estimated in a balanced setting. Furthermore, to enforce the compatibility between $\mathbf{x}$ and y, we adopt a pre-encoder step to learn a dense representation for the input $\mathbf{x}$, and then learn the label-informed latent features $\mathbf{z}$.

VII. Experiment

A. Experiment Setup

1) Data Sets: We use the following data sets.



(a) Prediction performance
(b) Catastrophic forgetting

Fig. 1. Results on Caltran

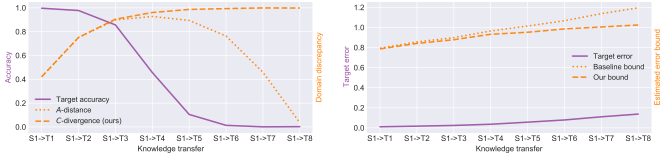
- **Synthetic Data:** The synthetic data set contains a set of source and target data points where the positive and negative samples are randomly sampled from two independent Gaussian distributions $\mathcal{N}([1.5\cos\theta, 1.5\sin\theta]^T, 0.5\cdot\mathbf{I}_{2\times 2})$ and $\mathcal{N}([1.5\cos{(-\theta)}, 1.5\sin{(-\theta)}]^T, 0.5\cdot\mathbf{I}_{2\times 2})$. We let $\theta = 0$ for the source domain (denoted as $S1$), and then the data points are rotated by setting θ as $\frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}, \pi$ to generate the target domain with time-evolving nature.
- **Object Recognition Data:** We use two object recognition data sets: Office-31[2] and Office-Home[3]. Office-31 has three domains: Amazon, DSLR and Webcam. Office-Home has four domains: Art, Product, Clipart and Real World. In this case, we simulate the time-evolving distribution on the target domain by constantly adding random salt&pepper noise and rotation into raw images over time.
- **Scene Classification Data:** Caltran[4] is a real-world image data set captured by a camera at an intersection for 12 days, and one day has over 100 images of 2 categories. In our experiments, we assume the first day as the source domain, and others as a time evolving target domain.

2) Baselines: We use the following baseline methods. (1) SourceOnly: training with only source data; (2) TargetERM: empirical risk minimization (ERM) on only target domain; (3) Static adaptation: DAN [9], DANN [3], and MDD [19]; (4) Dynamic adaptation: CUA [2], GST [6], and CIDA [14]; (5) TransLATE, and TransLATE$_\infty$ which is a static variant of TransLATE that directly transfers from source to the newest target domain. We set $\lambda = 0.1$ and $\mu = 1.0$ in the experiments.

B. Results

Tables I and Table II provide the results of TransLATE on Office-31 and Office-Home where the target classification accuracy in every time stamp is reported (e.g., 'T1' in Amazon $\rightarrow$ Webcam denotes the target domain Webcam in the first time stamp). It is observed that (1) the classification accuracy using SourceOnly algorithm significantly decreases on the evolving target domain due to the time evolving relatedness across domains; (2) the performance of static baselines is largely affected by the distribution shift in the evolving target domain; (3) TransLATE significantly outperforms TransLATE$_\infty$ and other competitors by a large margin (i.e., up to 30% improvement on the last time stamp of target domain).

Figure 1(a) shows the performance of TransLATE on the newest target domain of Caltran. It confirms the effectiveness of TransLATE on modeling the time evolving target domain. Besides, following [2], we investigate the catastrophic forgetting mitigation of TransLATE in Figure 1(b) where the average accuracy of TransLATE evaluated on historical data is provided. It demonstrates that TransLATE mitigates catastrophic forgetting.



(a) Comparison of domain discrepancy and target accuracy
(b) Comparison of different error bounds

Fig. 2. Model analysis

C. Analysis

1) Evaluation of $\mathcal{C}$-divergence: We compare the proposed $\mathcal{C}$-divergence with conventional domain discrepancy measure $\mathcal{A}$-distance [1] on the synthetic data set with an evolving target domain. We assume that the hypothesis space $\mathcal{H}$ consists of linear classifiers in the feature space. Figure 2(a) shows the domain discrepancy and target classification accuracy for each pair of source and target domains. We have the following observations. (1) The classification accuracy on the target domain significantly decreases from target domain T1 to T8. One explanation is that the joint distribution $p(x, y)$ on the time evolving target domain has gradually shifted. (2) The $\mathcal{A}$-distance increases from S1$\rightarrow$T1 to S1$\rightarrow$T4, and then decreases from S1$\rightarrow$T4 to S1$\rightarrow$T8. That is because it only estimates the difference of the marginal feature distribution $p(x)$ between the source and target domains. (3) The $\mathcal{C}$-divergence keeps increasing from S1$\rightarrow$T1 to S1$\rightarrow$T8, which indicates the decreasing task relatedness between the source and the target domains. Therefore, compared with $\mathcal{A}$-distance, $\mathcal{C}$-divergence better characterizes the transferability from the source to the target domains.

2) Evaluation of Error Bound: When there is only one time stamp involved in the target domain, Theorem 7 is reduced to the standard error bound in the conventional static transfer learning setting. We empirically compare this reduced error bound with the existing Rademacher complexity based error bound in [10]. We use the 0-1 loss function as $\mathcal{L}$ and assume that the hypothesis space $\mathcal{H}$ consists of linear classifiers in the feature space. Figure 2(b) shows the estimated error bounds and target error with the time evolving target domain (i.e., S1$\rightarrow$T1, $\cdots$, S1$\rightarrow$T8 in a new synthetic data set with a slower time evolving target domain to ensure that the baseline bound is meaningful most of the time) where we choose $h = h^*_{T_0}$. It demonstrates that our $\mathcal{C}$-divergence based error bound is much tighter than the baseline. Notice that when transferring source domain S1 to target domain T8, our error bound is largely determined by the $\mathcal{C}$-divergence, whereas the baseline is determined by the difference between the optimal source and target hypotheses. Furthermore, given any hypothesis $h \in \mathcal{H}$, we may not be able to estimate the baseline bound when the optimal hypothesis is not available.

[2] https://people.eecs.berkeley.edu/~jhoffman/domainadapt/
[3] http://hemanthdv.org/OfficeHome-Dataset/
[4] http://cma.berkeleyvision.org

TABLE I
CLASSIFICATION ACCURACY ON OFFICE-31 (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD)

	Amazon → Webcam					DSLR → Webcam				
	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
SourceOnly	0.7490	0.2255	0.2282	0.1275	0.1503	0.9651	0.4309	0.3329	0.1611	0.2027
TargetERM	0.5584	0.3933	0.4215	0.3396	0.3732	0.4966	0.4201	0.4188	0.3248	0.4067
DAN [9]	0.8537	0.5007	0.4993	0.3638	0.4470	0.9772	0.7302	0.6161	0.4765	0.5302
DANN [3]	0.8389	0.4993	0.4121	0.3973	0.3382	0.9651	0.7356	0.6416	0.4510	0.5490
MDD [19]	0.8940	0.6738	0.5490	0.5141	0.4295	0.9724	0.8738	0.7315	0.5047	0.5289
TransLATE$_\infty$	**0.9154**	0.6376	0.5758	0.4591	0.4846	0.9785	0.8591	0.7289	0.4926	0.5557
CUA [2]	0.8349	0.6805	0.6389	0.6456	0.6805	**0.9852**	0.8805	0.8792	0.8362	0.8617
GST [6]	0.8456	0.5987	0.6013	0.5584	0.5960	0.9739	0.8376	0.8134	0.7570	0.7865
CIDA [14]	0.8805	0.7638	0.7624	0.7195	0.7476	0.9812	0.8577	0.8376	0.7973	0.7960
TransLATE	**0.9154**	**0.8134**	**0.8081**	**0.7611**	**0.7826**	0.9785	**0.9235**	**0.9208**	**0.8886**	**0.9154**

TABLE II
CLASSIFICATION ACCURACY ON OFFICE-HOME (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD)

	Art → Real World					Clipart → Product				
	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
SourceOnly	0.7220	0.3947	0.3135	0.2650	0.3512	0.5944	0.1866	0.1342	0.1074	0.1550
TargetERM	0.5643	0.3297	0.3010	0.2582	0.3299	0.6033	0.3805	0.4232	0.3061	0.3791
DAN [9]	0.7341	0.4901	0.4193	0.3686	0.4597	0.7186	0.4201	0.3921	0.3352	0.4113
DANN [3]	0.7359	0.5092	0.4155	0.3850	0.4686	0.7063	0.4440	0.3694	0.3343	0.4303
MDD [19]	0.7435	0.5056	0.4331	0.3874	0.4686	0.7264	0.4765	0.3886	0.3514	0.4294
TransLATE$_\infty$	**0.7560**	0.5273	0.4575	0.4080	0.4850	**0.7411**	0.5017	0.4436	0.3634	0.4595
CUA [2]	0.7370	0.5732	0.5181	0.4932	0.5372	0.7143	0.4922	0.4431	0.4310	0.4879
GST [6]	0.7367	0.5283	0.4795	0.4681	0.4826	0.7285	0.5232	0.4782	0.4531	0.4943
CIDA [14]	0.7420	0.5643	0.4983	0.4896	0.5130	0.7226	0.5076	0.4334	0.4030	0.4362
TransLATE	**0.7560**	**0.6046**	**0.5447**	**0.5097**	**0.5459**	**0.7411**	**0.5747**	**0.5318**	**0.5009**	**0.5422**

VIII. CONCLUSION

In this paper, we study the dynamic transfer learning problem by deriving a generic error bound of dynamic transfer learning with flexible domain discrepancy measures. Then we propose a novel label-informed $\mathcal{C}$-divergence, which leads to an improved error bound. By minimizing this error bound, we further propose a novel adversarial Variational Auto-encoder algorithm TransLATE. The experimental results confirm the effectiveness of our TransLATE algorithm.

ACKNOWLEDGMENT

This work is supported by National Science Foundation under Award No. IIS-1947203, IIS-2117902, IIS-2137468, 1947135, 2134079, 1939725, and Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

REFERENCES

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 2010.

[2] A. Bobu, E. Tzeng, J. Hoffman, and T. Darrell. Adapting to continuously shifting domains. In *ICLR Workshop*, 2018.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016.

[4] J. Hoffman, T. Darrell, and K. Saenko. Continuous manifold based adaptation for evolving visual domains. In *CVPR*, 2014.

[5] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014.

[6] A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In *ICML*, 2020.

[7] T. Le, T. Nguyen, N. Ho, H. Bui, and D. Phung. Lamda: Label matching deep domain adaptation. In *ICML*, 2021.

[8] H. Liu, M. Long, J. Wang, and Y. Wang. Learning to adapt to evolving domains. *NeurIPS*, 2020.

[9] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

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

[11] M. Mohri and A. M. Medina. New analysis and algorithm for learning with drifting distributions. In *ALT*, 2012.

[12] S. J. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 2009.

[13] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000.

[14] H. Wang, H. He, and D. Katabi. Continuously indexed domain adaptation. In *ICML*, 2020.

[15] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell. Characterizing and avoiding negative transfer. In *CVPR*, 2019.

[16] J. Wu, E. A. Ainsworth, S. Wang, K. Guan, and J. He. Adaptive transfer learning for plant phenotyping. In *MLCAS*, 2021.

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

[18] C. R. Yendrek, T. Tomaz, C. M. Montes, Y. Cao, A. M. Morse, P. J. Brown, L. M. McIntyre, A. D. Leakey, and E. A. Ainsworth. High-throughput phenotyping of maize leaf physiological and biochemical traits using hyperspectral reflectance. *Plant physiology*, 2017.

[19] Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, 2019.

[20] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In *ICML*, 2019.

[21] Y. Zhou, F. Ma, J. Gao, and J. He. Optimizing the wisdom of the crowd: Inference, learning, and teaching. In *KDD*, 2019.

[22] Y. Zhou, L. Ying, and J. He. MultiC^2: an optimization framework for learning from task and worker dual heterogeneity. In N. V. Chawla and W. Wang, editors, *SDM*, 2017.

[23] Y. Zhou, L. Ying, and J. He. Multi-task crowdsourcing via an optimization framework. *TKDD*, 2019.