A Unified Meta-Learning Framework for Dynamic Transfer Learning

Jun Wu , Jingrui He
University of Illinois Urbana-Champaign
{junwu3, jingrui}@illinois.edu

Abstract

Transfer learning refers to the transfer of knowledge or information from a relevant source task to a target task. However, most existing works assume both tasks are sampled from a stationary task distribution, thereby leading to the sub-optimal performance for dynamic tasks drawn from a non-stationary task distribution in real scenarios. To bridge this gap, in this paper, we study a more realistic and challenging transfer learning setting with dynamic tasks, i.e., source and target tasks are continuously evolving over time. We theoretically show that the expected error on the dynamic target task can be tightly bounded in terms of source knowledge and consecutive distribution discrepancy across tasks. This result motivates us to propose a generic meta-learning framework `L2E` for modeling the knowledge transferability on dynamic tasks. It is centered around a task-guided meta-learning problem with a group of meta-pairs of tasks, based on which we are able to learn the prior model initialization for fast adaptation on the newest target task. `L2E` enjoys the following properties: (1) effective knowledge transferability across dynamic tasks; (2) fast adaptation to the new target task; (3) mitigation of catastrophic forgetting on historical target tasks; and (4) flexibility on incorporating any existing static transfer learning algorithms. Extensive experiments on various image data sets demonstrate the effectiveness of the proposed `L2E` framework.

1 Introduction

Transfer learning [Pan and Yang, 2009] aims to leverage the knowledge of a source task to improve the generalization performance of a learning algorithm on a target task. The knowledge transferability across tasks can be theoretically guaranteed under mild assumptions, even when no labeled training examples are available in any target task [Ben-David *et al.*, 2010; Ghifary *et al.*, 2016; Acuna *et al.*, 2021]. One key assumption is that source and target tasks are sampled from a stationary task distribution. The resulting relatedness between tasks allows transferring knowledge from a source task with adequate labeled data to a target task with little or no labeled data. However, the learning task might be evolving over time [Mohri and Medina, 2012] in real scenarios. For example, the data distribution of clothes images in Amazon is changing over the years due to the varying fashion trend [He and McAuley, 2016], thus resulting in a time-evolving image recognition task. Another example is the IMDb's film rating system where the rating scores of a film change in different time periods due to the dynamic user preference [Rafailidis and Nanopoulos, 2015], thus leading to a time-evolving film recommendation task. Such application scenarios would challenge the conventional static transfer learning algorithms [Pan and Yang, 2009] due to the dynamic task relatedness.

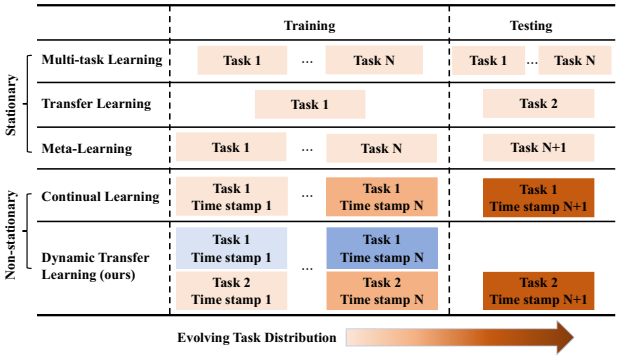


Figure 1: Illustration of transfer learning on dynamic tasks (best viewed in color)

Recent works [Hoffman *et al.*, 2014; Liu *et al.*, 2020; Wang *et al.*, 2020; Kumar *et al.*, 2020] have studied continuous transfer learning with a static source task and a time evolving target task. They revealed that the prediction function on the newest target task can be learned by aggregating the knowledge from the labeled source data and historical unlabeled target data. Nevertheless, in real scenarios, both source and target tasks could be changing over time. In this case, those works will lead to the sub-optimal solution due to the under-explored source knowledge. To the best of our knowledge, very little effort has been devoted to modeling the knowledge transferability from a labeled dynamic source task to an unlabeled dynamic target task.

To bridge this gap, in this paper, we study the dynamic transfer learning problem with dynamic source and target tasks sampled from a non-stationary task distribution. As

shown in Figure 1, conventional knowledge transfer problems focus on either static tasks (e.g., multi-task learning [Sener and Koltun, 2018], transfer learning [Pan and Yang, 2009] and meta-learning [Finn *et al.*, 2017]) or one dynamic task (e.g., continual learning [Li and Hoiem, 2017]), whereas we aim to transfer the knowledge from a dynamic source task to a dynamic target task. More specifically, we focus on the learning scenario where both source and target tasks always share the same class-label space (a.k.a. domain adaptation [Pan and Yang, 2009]) at any time stamp. We are able to show that the generalization error bounds of dynamic transfer learning can be derived under the following assumptions. First, the class labels of the source task are available at any time stamp. Second, the source and target tasks are related at the initial time stamp. Third, the data distributions of both source and target tasks are continuously changing over time. The theoretical results indicate that the target error is bounded in terms of flexible domain discrepancy measures (e.g., $\mathcal{H}$-divergence [Ben-David *et al.*, 2010], discrepancy distance [Mansour *et al.*, 2009], f-divergence [Acuna *et al.*, 2021], etc.) across tasks and across time stamps. This motivates us to propose a generic meta-learning framework `L2E` for dynamic transfer learning. It reformulates the dynamic source and target tasks into a set of meta-pairs of consecutive tasks, and then learns the prior model initialization for fast adaptation on the newest target task. The effectiveness of `L2E` is empirically verified on various image data sets. The main contributions of this paper are summarized as follows:

- We derive the error bounds for dynamic transfer learning with time-evolving source and target tasks.
- We propose a generic meta-learning framework (`L2E`) for transfer learning on dynamic tasks by minimizing the error upper bounds with flexible divergence measures.
- Extensive experiments on public data sets confirm the effectiveness of our proposed `L2E` framework.

The rest of the paper is organized as follows. We review the related work in Section 2, followed by our problem setting in Section 3. In Section 4, we derive the error bounds of continuous transfer learning and then present the `L2E` framework. The extensive experiments and discussion are provided in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

Transfer Learning: Transfer learning [Pan and Yang, 2009] improves the generalization performance of a learning algorithm under distribution shift. Most existing algorithms [Zhang *et al.*, 2019; Acuna *et al.*, 2021; Wu and He, 2021] assume the relatedness of static source and target tasks in order to guarantee the success of knowledge transfer. The most related problem to our work is the meta-transfer learning [Sun *et al.*, 2019], which learns to adapt to a set of few shot learning tasks sampled from a stationary task distribution. However, our work focuses on the non-stationary task distribution where a sequence of meta-pairs of consecutive tasks could be formulated for dynamic transfer learning.

Continual Learning: Continual learning aims to learn a model on a new task using knowledge from experienced tasks. Conventional continual learning algorithms [Li and Hoiem, 2017] focused on mitigating the catastrophic forgetting when learning the prediction function on one time-evolving task. In contrast, continuous transfer learning [Hoffman *et al.*, 2014; Bobu *et al.*, 2018; Wu and He, 2020; Liu *et al.*, 2020; Wang *et al.*, 2020; Kumar *et al.*, 2020] transferred the knowledge from a labeled static source task to an unlabeled time-evolving target task. Our work would further extend it to the transfer learning setting with dynamic source and target tasks.

Meta-learning: Meta-learning [Finn *et al.*, 2017; Fallah *et al.*, 2020], or learning to learn, leverages the knowledge from a set of prior tasks for fast adaptation to unseen tasks. It assumes that all the tasks follow a stationary task distribution. More recently, it has been extended into the online learning setting [Finn *et al.*, 2019] where a sequence of tasks are sampled from non-stationary task distributions. However, those works focused on improving the prediction performance with the accumulated data, whereas we aim to explore the knowledge transferability between dynamic source and target tasks.

3 Problem Setting

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input feature space and output label space respectively. We consider the dynamic transfer learning problem[1] with dynamic source task $\{\mathcal{D}_j^s\}_{j=1}^N$ and target task $\{\mathcal{D}_j^t\}_{j=1}^N$ with time stamp j. In this case, we assume that there are m_j^s labeled training examples $D_j^s = \{(\mathbf{x}_{ij}^s, y_{ij}^s)\}_{i=1}^{m_j^s}$ in the j^{th} source task and no labeled training examples in the target task. Let m_j^t be the number of unlabeled training examples $D_j^t = \{\mathbf{x}_{ij}^t\}_{i=1}^{m_j^t}$ in the j^{th} target task. Furthermore, each task $\mathcal{D}_j^s$ ($\mathcal{D}_j^t$) is associated with a task-specific labeling function f_j^s (f_j^t). Let $\mathcal{H}$ be the hypothesis class on $\mathcal{X}$ where a hypothesis is a function $h : \mathcal{X} \to \mathcal{Y}$. $\mathcal{L}(\cdot, \cdot)$ is the loss function such that $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The expected classification error on the task $\mathcal{D}_j$ (either source or target) is defined as $\epsilon_j(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_j}[\mathcal{L}(h(\mathbf{x}), y)]$ for any $h \in \mathcal{H}$, and its empirical estimate is given by $\hat{\epsilon}_j(h) = \frac{1}{m_j}\sum_{i=1}^{m_j}\mathcal{L}(h(\mathbf{x}_{ij}), y_{ij})$.

Formally, our dynamic transfer learning problem can be defined as follows.

Definition 1. ***(Dynamic Transfer Learning)*** *Given labeled dynamic source tasks $\{\mathcal{D}_j^s\}_{j=1}^N$ and unlabeled dynamic target tasks $\{\mathcal{D}_j^t\}_{j=1}^N$, dynamic transfer learning aims to learn the prediction function for the newest target task $\mathcal{D}_{N+1}^t$ by leveraging the knowledge from historical source and target tasks.*

In dynamic transfer learning, we have the following mild assumptions. (1) The class labels of the source task are available at any time stamp. Specially, when the source task is static, it is naturally degenerated into a typical continuous transfer learning problem [Hoffman *et al.*, 2014; Bobu *et al.*, 2018; Wu and He, 2020; Liu *et al.*, 2020]. Moreover, when the target task is also static, it would be a stan-

[1] In this paper, we assume that all the tasks share the same output label space $\mathcal{Y}$ for simplicity. Besides, we use $\mathcal{D}_j^s$ to represent both the time-specific task (i.e., source task at the j^{th} time stamp) and its data distribution (i.e., probability distribution of the source task at the j^{th} time stamp over $\mathcal{X} \times \mathcal{Y}$) for notation simplicity.

dard transfer learning problem [Pan and Yang, 2009]. (2) The source and target tasks are related at the initial time stamp $j = 1$. That is, those tasks might not be related in the following time stamps $j > 1$, as their data distributions can be evolving towards different patterns. (3) The data distributions of both source and target tasks are continuously changing over time.

4 The Proposed Framework

In this section, we first derive the error bounds for dynamic transfer learning, and then propose a generic meta-learning framework (L2E) for modeling the knowledge transferability across dynamic tasks.

4.1 Error Bounds

Following [Ben-David *et al.*, 2010], we consider a binary classification problem with $\mathcal{Y} \in \{0, 1\}$ for simplicity. Before deriving the generalization error bound for a dynamic target task, we first introduce some basic concepts below.

Definition 2. *(L^1-**divergence** [Ben-David* et al.*, 2010]) The L^1-divergence between two distributions $\mathcal{D}$ and $\mathcal{D}'$ over $\mathcal{X}$ is defined as follows.*

$$d_1(\mathcal{D}, \mathcal{D}') := 2 \sup_{Q \in \mathcal{Q}} |\Pr_{\mathcal{D}}[Q] - \Pr_{\mathcal{D}'}[Q]| \tag{1}$$

where $\mathcal{Q}$ is the set of measurable subsets under $\mathcal{D}$ and $\mathcal{D}'$.

Definition 3. *(μ-**admissibility**) A loss function $\mathcal{L}(\cdot, \cdot)$ is μ-admissible if there exists $\mu > 0$ such that for all $\mathbf{x} \in \mathcal{X}$, $y, y' \in \mathcal{Y}$ and $h, h' \in \mathcal{H}$, the following inequalities hold.*

$$|\mathcal{L}(h'(\mathbf{x}), y) - \mathcal{L}(h(\mathbf{x}), y)| \leq \mu\, |h'(\mathbf{x}) - h(\mathbf{x})|$$

$$|\mathcal{L}(h(\mathbf{x}), y') - \mathcal{L}(h(\mathbf{x}), y)| \leq \mu\, |y' - y|$$

In the context of dynamic transfer learning, the following theorem states that the expected target error of the newest target task can be bounded in terms of historical source and target knowledge.

Theorem 1. *Assume that the loss function $\mathcal{L}(\cdot, \cdot)$ is μ-admissible and obeys the triangle inequality. Given a class of functions $\mathcal{H}_{\mathcal{L}} = \{(\mathbf{x}, y) \mapsto \mathcal{L}(h(\mathbf{x}), y) : h \in \mathcal{H}\}$, for any $\delta > 0$ and $h \in \mathcal{H}$, with probability at least $1 - \delta$, the expected error ϵ^t_{N+1} for the newest target task $\mathcal{D}^t_{N+1}$ is bounded by*

$$\epsilon^t_{N+1}(h) \leq \frac{1}{2N} \sum_{j=1}^{N} \left(\hat{\epsilon}^s_j(h) + \hat{\epsilon}^t_j(h)\right) + \frac{N+2}{2} \left(\tilde{d} + \tilde{\lambda}\right) + \tilde{\Re}(\mathcal{H}_{\mathcal{L}}) + \frac{\mu}{N} \sqrt{\frac{\log \frac{1}{\delta}}{2\tilde{m}}}$$

where $\tilde{d} = \mu \cdot \max\big\{ \max_{1 \leq j \leq N-1} d_1(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}), d_1(\mathcal{D}^s_1, \mathcal{D}^t_1), \max_{1 \leq j \leq N} d_1(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}) \big\}$, $\tilde{\lambda} = \mu \cdot \max\big\{ \max_{1 \leq j \leq N-1} \lambda_*(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}), \lambda_*(\mathcal{D}^s_1, \mathcal{D}^t_1), \max_{1 \leq j \leq N} \lambda_*(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}) \big\}$ and λ_* measures the difference of the labeling functions across task and across time stamps, i.e., $\lambda_*(\mathcal{D}^s_1, \mathcal{D}^t_1) = \min\{\mathbb{E}_{\mathcal{D}^s_1}[|f^s_1(\mathbf{x}) - f^t_1(\mathbf{x})|], \mathbb{E}_{\mathcal{D}^t_1}[|f^s_1(\mathbf{x}) - f^t_1(\mathbf{x})|]\}$. $\tilde{\Re}(\mathcal{H}_{\mathcal{L}})$ is a Rademacher complexity term (see Appendix for more details) and $\tilde{m} = \sum_{j=1}^{N}(m^s_j + m^t_j)$ is the total number of training examples from source and historical target tasks.

This theorem reveals that the expected error on the newest target task can be bounded in terms of (i) the empirical errors of historical source and target tasks; (ii) the maximum of the distribution discrepancy (e.g., L^1-divergence) across tasks and across time stamps; (iii) the maximum of the labeling difference across tasks and across time stamps; and (iv) the average Rademacher complexity [Mansour *et al.*, 2009] of the class of functions $\mathcal{H}_{\mathcal{L}} = \{(\mathbf{x}, y) \mapsto \mathcal{L}(h(\mathbf{x}), y) : h \in \mathcal{H}\}$ over all the tasks.

However, it has been observed that (1) L^1-divergence cannot be accurately estimated from finite samples of arbitrary distributions [Ben-David *et al.*, 2010]; and (2) the generalization error bound with L^1-divergence is not very tight because L^1-divergence involves all the measurable subsets over $\mathcal{X}$. Therefore, we would like to derive much tighter error bounds with existing domain divergence measures over either marginal feature space (e.g., $\mathcal{H}$-divergence [Ben-David *et al.*, 2010], discrepancy distance [Mansour *et al.*, 2009], f-divergence [Acuna *et al.*, 2021] and Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2012]) or joint feature and label space (e.g., $\mathcal{C}$-divergence [Wu and He, 2020]). It is notable [Acuna *et al.*, 2021] that the generic f-divergence subsumes many popular divergences, including Margin Disparity Discrepancy [Zhang *et al.*, 2019], Jensen-Shannon (JS) divergence, Pearson χ^2 divergence, etc.

Corollary 1. *Assume that the loss function $\mathcal{L}(\cdot, \cdot)$ is μ-admissible and 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*

*(a) when using f-**divergence** [Acuna* et al.*, 2021][2], denoted by d_f, the error bound of Theorem 1 holds with*

$$\tilde{d} = \max\Big\{ \max_{1 \leq j \leq N-1} d_f(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}), d_f(\mathcal{D}^s_1, \mathcal{D}^t_1), \max_{1 \leq j \leq N} d_f(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}) \Big\}$$

$$\tilde{\lambda} = \max\Big\{ \max_{1 \leq j \leq N-1} \lambda_*(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}), \lambda_*(\mathcal{D}^s_1, \mathcal{D}^t_1), \max_{1 \leq j \leq N} \lambda_*(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}) \Big\}$$

where $\lambda_*(\mathcal{D}^s_1, \mathcal{D}^t_1) = \min_{h \in \mathcal{H}} \epsilon^s_1(h) + \epsilon^t_1(h)$.

*(b) when using $\mathcal{C}$-**divergence** [Wu and He, 2020] (measuring the distribution discrepancy over joint data distribution on $\mathcal{X} \times \mathcal{Y}$), denoted by $d_{\mathcal{C}}$, the error bound of Theorem 1 holds with*

$$\tilde{d} = \mu \cdot \max\Big\{ \max_{1 \leq j \leq N-1} d_{\mathcal{C}}(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}), d_{\mathcal{C}}(\mathcal{D}^s_1, \mathcal{D}^t_1), \max_{1 \leq j \leq N} d_{\mathcal{C}}(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}) \Big\}$$

$$\tilde{\lambda} = 0$$

The theoretical results above motivate us to develop a dynamic transfer learning framework by empirically minimizing the generalization error bounds with flexible domain discrepancy measures (see the next subsection).

4.2 L2E Framework

Following the theoretical results [Ben-David *et al.*, 2010], a typical transfer learning paradigm on static source and target tasks aims to minimize the static generalization error bound

[2]Note that we have similar error bounds when using other marginal domain discrepancy measures (e.g., $\mathcal{H}$-divergence [Ben-David *et al.*, 2010], discrepancy distance [Mansour *et al.*, 2009], or MMD [Gretton *et al.*, 2012]), so we omit the details here (see Appendix for more illustration)

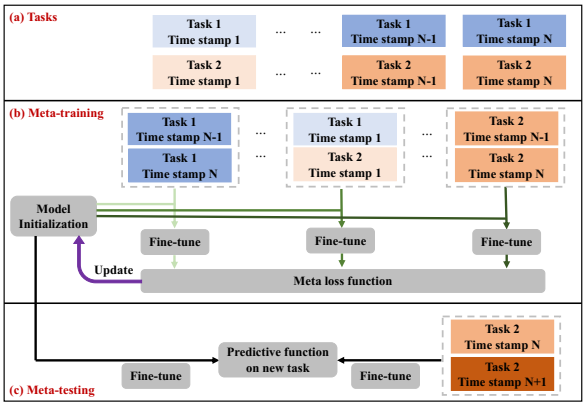


Figure 2: Illustration of our proposed L2E framework (best viewed in color). It consists of three phases: (i) reformulating the meta-pairs of consecutive tasks; (ii) learning the prior model initialization; (iii) fine-tuning on the new task.

involving empirical source classification error and domain discrepancy across tasks as follows.

$$\min_{\theta} J(\theta) = \hat{\epsilon}_s(\theta) + \gamma \cdot \hat{d}(\mathcal{D}^s, \mathcal{D}^t; \theta) \tag{2}$$

where θ is the trainable parameters and γ is a trade-off parameter between the empirical source error and the domain discrepancy. The second term $\hat{d}(\mathcal{D}^s, \mathcal{D}^t; \theta)$ aims to match the data distribution of source and target tasks by learning the domain-invariant latent representation for every input example. Then the predictive function learned by the first term $\hat{\epsilon}_s(\theta)$ on the source task could be applied to the target task directly. In this case, existing works [Ganin *et al.*, 2016; Acuna *et al.*, 2021] have attempted to instantiate different domain discrepancy measures. Nevertheless, when the source and target tasks are evolving over time, it might be suboptimal when directly transferring the labeled source task to the newest unlabeled target task. That is because the success of knowledge transfer cannot be guaranteed if the task relatedness becomes weaker over time [Rosenstein *et al.*, 2005].

In this paper, we propose a generic meta-learning framework L2E for transferring the knowledge from a dynamic source task to a dynamic target task. It leverages the knowledge from both historical source and target tasks to improve the predictive performance on the newest target task. Following our generalization error bounds in Corollary 1, we have the following objective function for learning the predictive function of $\mathcal{D}^t_{t+1}$ on the $(N+1)^{\text{th}}$ time stamp.

$$\begin{aligned}\min_{\theta} J(\theta) = &\sum_{j=1}^{N} \left(\hat{\epsilon}^s_j(\theta) + \hat{\epsilon}^t_j(\theta)\right) + \gamma \cdot \Big(\hat{d}(\mathcal{D}^s_1, \mathcal{D}^t_1; \theta) \\ &+ \sum_{j=1}^{N-1} \hat{d}(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}; \theta) + \sum_{j=1}^{N} \hat{d}(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}; \theta)\Big)\end{aligned} \tag{3}$$

where $\hat{d}(\cdot, \cdot; \theta)$ is the empirical distribution discrepancy estimated from finite samples. It can be instantiated with any existing domain discrepancy measures discussed in Corollary 1. We would like to point out that the error bound of Corollary 1 is derived in terms of the maximum distribution discrepancy across tasks and across time stamps. But there is no prior knowledge regarding the time stamp with the maximum distribution discrepancy. Therefore, in our framework, we propose to minimize all the distribution discrepancies across tasks and across time stamps.

However, the learned model on the newest target task $\mathcal{D}^t_{N+1}$ might have the issue of catastrophic forgetting such that it performs badly on historical target tasks when updating the new target task. To solve this problem, we would like to learn optimal prior model initialization shared across all the target tasks such that this model can be efficiently fine-tuned on both new and historical target tasks with just a few updates. Figure 2 illustrates our proposed L2E framework with three crucial components: meta-pairs of tasks, meta-training, and meta-testing.

Meta-pairs of tasks: With the assumption that the source and target tasks are continuously evolving over time, the framework of Eq. (3) is equivalent to sequentially optimizing with respect to the adjacent tasks using standard transfer learning of Eq. (2) as follows.

$$\begin{aligned}\min_{\theta} J(\theta) = &\sum_{j=1}^{N-1} \left(\hat{\epsilon}^s_{j+1}(\theta) + \gamma \cdot \hat{d}(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}; \theta)\right) \\ &+ \left(\hat{\epsilon}^s_1(\theta) + \gamma \cdot \hat{d}(\mathcal{D}^s_1, \mathcal{D}^t_1; \theta)\right) \\ &+ \sum_{j=1}^{N} \left(\hat{\epsilon}^t_j(\theta) + \gamma \cdot \hat{d}\left(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}; \theta\right)\right)\end{aligned} \tag{4}$$

Thus we would like to simply split all the tasks into a set of meta-pairs consisting of two consecutive tasks as shown in Figure 2, and learn the prior model initialization with those meta-pairs of tasks. Different from previous works [Hoffman *et al.*, 2014; Liu *et al.*, 2020] with only static source task, we argue that the evolution pattern of the source task can also help improve the performance of L2E on the newest target task (see more empirical analysis in our experiments). In Subsection 4.3, we provide more discussion about the construction of meta-pairs of tasks.

Meta-training: Let $\zeta_j(\theta) = \hat{\epsilon}^t_j(\theta) + \gamma \cdot \hat{d}(\mathcal{D}^t_j, \mathcal{D}^t_{j+1}; \theta)$, $\zeta_0(\theta) = \hat{\epsilon}^s_1(\theta) + \gamma \cdot \hat{d}(\mathcal{D}^s_1, \mathcal{D}^t_1; \theta)$ and $\zeta_{-j}(\theta) = \hat{\epsilon}^s_{j+1}(\theta) + \gamma \cdot \hat{d}(\mathcal{D}^s_j, \mathcal{D}^s_{j+1}; \theta)$. Then Eq. (4) has a simplified expression $\min_\theta J(\theta) = \sum_{k=1-N}^{N} \zeta_k(\theta)$, where ζ_k denotes the objective function of standard transfer learning across tasks and across time stamps. Moreover, it can be formulated as a meta-learning problem [Finn *et al.*, 2017]. That is, the optimal model initialization is learned from historical source and target tasks such that it can be adapted to the newest target task with a few updates. To be more specific, we randomly split the training data from every historical source or target task into one training set $\mathcal{D}^{tr}_k$ and one validation set $\mathcal{D}^{val}_k$. Let ζ^{tr}_k (ζ^{val}_k) be the loss function of ζ_k on the training (validation) set. The model initialization $\tilde{\theta}^*_N$ can be learned as follows.

$$\begin{aligned}\tilde{\theta}^*_N &\leftarrow \arg\min_{\theta} \sum_{k=1-N}^{N-1} \zeta^{val}_k(M_k(\theta)) \\ M_k(\theta) &\leftarrow \theta - \alpha \cdot \nabla_\theta \zeta^{tr}_k(\theta)\end{aligned} \tag{5}$$

where $M_k : \theta \rightarrow \theta_k$ is a mapping function which learns the optimal task-specific parameter θ_k from model initialization θ. Following the model-agnostic meta-learning

(MAML) [Finn *et al.*, 2017], $M_k(\theta)$ can be instantiated by one or a few gradient descent updates. Here $\alpha \geq 0$ is the learning rate of the inner loop when training on a specific task. In addition, the training examples of historical target tasks are not labeled. Thus, we propose to sequentially learn the predictive function for every historical target task and generate the pseudo-labels of unlabeled examples as follows.

$$\begin{aligned} \tilde{\theta}^*_{j-1} &\leftarrow \arg\min_{\theta} \sum_{k=1-N}^{j-2} \zeta_k^{val}(M_k(\theta)) \\ \hat{y}^t_{ij} &\leftarrow p\left(y|\mathbf{x}^t_{ij}; M_{j-1}(\tilde{\theta}^*_{j-1})\right) \end{aligned} \tag{6}$$

where $\hat{y}^t_{ij}$ is the predicted pseudo-label of input example $\mathbf{x}^t_{ij}$ from the historical target task $\mathcal{D}^t_j$ ($j = 1, \cdots, N$). Notice that the training examples with incorrect pseudo-labels might lead to the accumulation of misclassification errors on the new target tasks over time. To mitigate this issue, we propose to select those examples with high prediction confidence. Specifically, we estimate the entropy of the predicted class probability of target examples, and then choose the top $p\%$ with the lowest entropy values. We empirically evaluate this sampling strategy in the experiments.

Meta-testing: The optimal parameters θ_{N+1} on the newest target task $\mathcal{D}^t_{N+1}$ could be obtained as follows.

$$\theta_{N+1} = M_N(\tilde{\theta}^*_N) \leftarrow \tilde{\theta}^*_N - \alpha \cdot \nabla_{\tilde{\theta}^*_N} \zeta_N \left(\tilde{\theta}^*_N\right) \tag{7}$$

where $\tilde{\theta}^*_N$ is the optimized model initialization learned in the meta-training phase.

The intuition of this meta-learning framework L2E can be illustrated as follows. The evolution of dynamic source and target tasks can be represented as a sequential knowledge transfer process across time stamps. But from the perspective of transfer learning [Pan and Yang, 2009], it would be an asymmetric knowledge transfer process for every time stamp, with the goal of maximizing the prediction performance on the new task. That explains why continuous knowledge transfer [Bobu *et al.*, 2018; Liu *et al.*, 2020; Kumar *et al.*, 2020] might have the issue of catastrophic forgetting. However, the continuous evolution of source and target tasks indicates that there might exist some common knowledge transferred across all time stamps. For instance, no matter how the fashion of clothes images changes over time, it follows the basic designs (e.g., shape) for different types of clothes. This common knowledge can be captured by the prior model initialization in our meta-learning framework. It then enables the fast adaptation on the newest target task with only a few model updates.

4.3 Discussion

In this work, we assume that the source and target tasks are continuously evolving over time. In other words, the data distribution of the source or target task is similar at the adjacent time stamps. This naturally motivates us to design the meta-pairs of tasks using adjacent tasks. However, it is possible that there exist other related tasks at the non-adjacent time stamps. One extreme case is that when the evolution of task distribution is negligible, it will be close to the transfer learning scenarios with multiple sources [Zhao *et al.*, 2018; Wen *et al.*, 2020]. In this case, any two historical tasks can be considered as the meta-pair of tasks in our L2E framework. By combining with the existing task transferability measures [Tran *et al.*, 2019], L2E can better identify meta-pairs of related tasks even when we have no prior knowledge regarding the evolution of task distribution. We would like to leave this to our future work.

Note that compared with continuous transfer learning algorithms [Bobu *et al.*, 2018; Wang *et al.*, 2020; Liu *et al.*, 2020], our proposed L2E framework has the following benefits. (1) It leverages the evolution knowledge from both historical source and target tasks. (2) It could be efficiently adapted to the newest target task with just a few updates. (3) It mitigates the catastrophic forgetting on historical target tasks by learning the prior model initialization shared across meta-pairs of tasks (that is, it could preserve the predictive performance on historical tasks by fine-tuning on those tasks with a few updates). (4) It is flexible to incorporate any existing static transfer learning algorithms [Pan and Yang, 2009] by instantiating $\zeta(\cdot)$ of Eq. (5) accordingly.

5 Experiments

5.1 Experiment Setup

Data Sets: We used three publicly available image data sets: Office-31 (with 3 tasks: Amazon, Webcam and DSLR), Image-CLEF (with 4 tasks: B, C, I and P) and Caltran. For Office-31 and Image-CLEF, there are 5 time stamps in the source task and 6 time stamp in the target task (see Appendix for more details). Caltran contains the real-time images captured by a camera at an intersection for two weeks.

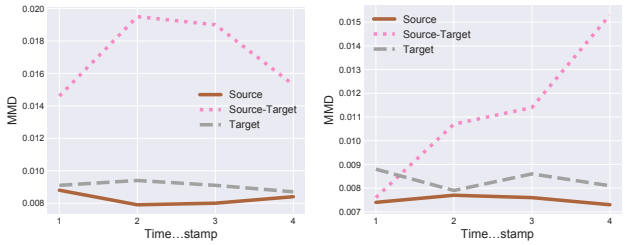
Baseline Methods: The comparison baselines are given below: (1) static adaptation methods: SourceOnly that trains only on the source task, DANN [Ganin *et al.*, 2016], and MDD [Zhang *et al.*, 2019]; (2) multi-source adaptation methods: MDAN [Zhao *et al.*, 2018], M3SDA [Peng *et al.*, 2019], and DARN [Wen *et al.*, 2020]; (3) continuous adaptation methods: CUA [Bobu *et al.*, 2018], TransLATE [Wu and He, 2020], GST [Kumar *et al.*, 2020], and our L2E with JS-divergence. In this case, we merge all the labeled source data into a large one, and then transfer its knowledge to the newest target task for static adaptation methods. For fair comparison, all the multi-source and continuous adaptation methods used both historical source and target data for knowledge transfer, and the target selection strategy for choosing high-quality target examples with pseudo-labels.

Configuration: We adopted the ResNet-18 [He *et al.*, 2016] pretrained on ImageNet as the base network for feature extraction, and set $\gamma = 0.1$ and $p = 80$ for all the experiments[3].

5.2 Results

Figure 3 shows the results of domain discrepancy via MMD, including (i) the evolution of source task, i.e., $\hat{d}(\mathcal{D}^s_j, \mathcal{D}^s_{j+1})$, (ii) the evolution of target task, i.e., $\hat{d}(\mathcal{D}^t_j, \mathcal{D}^t_{j+1})$, and (iii) the evolution of task relatedness, i.e., $\hat{d}(\mathcal{D}^s_j, \mathcal{D}^t_j)$. It indicates that in both Office-31 and Image-CLEF, the source and target tasks are changing smoothly, whereas the relatedness between source and target tasks is decreasing over time. Table 1, Table 2 and Figure 4 provide the transfer learning results on the dynamic tasks where the classification accuracies

[3] https://github.com/jwu4sml/L2E



(a) Webcam → DSLR (b) B → P

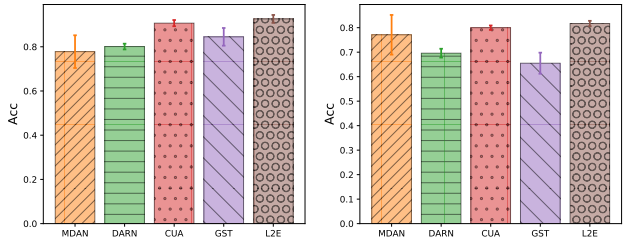
Figure 3: Task evolution on Office-31 (Webcam → DSLR) and Image-CLEF (B → P) where "Source": $\hat{d}(\mathcal{D}_j^s, \mathcal{D}_{j+1}^s)$, "Source-Target": $\hat{d}(\mathcal{D}_j^s, \mathcal{D}_j^t)$, and "Target": $\hat{d}(\mathcal{D}_j^t, \mathcal{D}_{j+1}^t)$

Method	Amazon → Webcam		Webcam → DSLR	
	Acc	H-Acc	Acc	H-Acc
SourceOnly	$0.168_{\pm 0.023}$	$0.355_{\pm 0.021}$	$0.776_{\pm 0.020}$	$0.880_{\pm 0.026}$
DANN	$0.317_{\pm 0.028}$	$0.450_{\pm 0.038}$	$0.837_{\pm 0.009}$	$0.905_{\pm 0.011}$
MDD	$0.319_{\pm 0.010}$	$0.447_{\pm 0.007}$	$0.852_{\pm 0.024}$	$0.909_{\pm 0.012}$
MDAN	$0.495_{\pm 0.027}$	$0.574_{\pm 0.011}$	$0.878_{\pm 0.003}$	$0.916_{\pm 0.010}$
M3SDA	$0.487_{\pm 0.031}$	$0.568_{\pm 0.032}$	$0.813_{\pm 0.045}$	$0.858_{\pm 0.030}$
DARN	$0.450_{\pm 0.018}$	$0.494_{\pm 0.033}$	$0.692_{\pm 0.035}$	$0.742_{\pm 0.027}$
CUA	$0.481_{\pm 0.005}$	$0.580_{\pm 0.009}$	$0.847_{\pm 0.030}$	$0.887_{\pm 0.036}$
TransLATE	$0.500_{\pm 0.013}$	$\mathbf{0.584}_{\pm \mathbf{0.020}}$	$0.862_{\pm 0.021}$	$0.900_{\pm 0.018}$
GST	$0.435_{\pm 0.014}$	$0.448_{\pm 0.002}$	$0.839_{\pm 0.018}$	$0.796_{\pm 0.008}$
L2E (ours)	$\mathbf{0.519}_{\pm \mathbf{0.016}}$	$0.581_{\pm 0.014}$	$\mathbf{0.893}_{\pm \mathbf{0.007}}$	$\mathbf{0.951}_{\pm \mathbf{0.001}}$

Table 1: Transfer learning accuracy on Office-31

Method	I → C		B → P	
	Acc	H-Acc	Acc	H-Acc
SourceOnly	$0.256_{\pm 0.013}$	$0.506_{\pm 0.013}$	$0.241_{\pm 0.004}$	$0.432_{\pm 0.010}$
DANN	$0.361_{\pm 0.002}$	$0.577_{\pm 0.008}$	$0.270_{\pm 0.009}$	$0.426_{\pm 0.003}$
MDD	$0.410_{\pm 0.011}$	$0.615_{\pm 0.011}$	$0.282_{\pm 0.031}$	$0.424_{\pm 0.015}$
MDAN	$0.620_{\pm 0.034}$	$0.772_{\pm 0.001}$	$0.368_{\pm 0.051}$	$0.508_{\pm 0.023}$
M3SDA	$0.559_{\pm 0.028}$	$0.743_{\pm 0.008}$	$0.386_{\pm 0.023}$	$0.524_{\pm 0.024}$
DARN	$0.553_{\pm 0.025}$	$0.763_{\pm 0.017}$	$0.390_{\pm 0.020}$	$0.517_{\pm 0.014}$
CUA	$0.575_{\pm 0.010}$	$0.737_{\pm 0.014}$	$0.358_{\pm 0.033}$	$0.508_{\pm 0.003}$
TransLATE	$0.637_{\pm 0.012}$	$0.764_{\pm 0.004}$	$0.402_{\pm 0.030}$	$0.548_{\pm 0.009}$
GST	$0.392_{\pm 0.013}$	$0.539_{\pm 0.030}$	$0.320_{\pm 0.013}$	$0.306_{\pm 0.018}$
L2E (ours)	$\mathbf{0.659}_{\pm \mathbf{0.020}}$	$\mathbf{0.804}_{\pm \mathbf{0.008}}$	$\mathbf{0.435}_{\pm \mathbf{0.036}}$	$\mathbf{0.573}_{\pm \mathbf{0.021}}$

Table 2: Transfer learning accuracy on Image-CLEF



(a) Acc (b) H-Acc

Figure 4: Transfer learning accuracy on Caltran

on the newest target task (Acc) and all the historical target tasks (H-Acc) are reported. We run all the experiments five times and report the mean and standard deviation of classification accuracies (the best results are indicated in bold). It can be observed that: (1) compared to static methods, the multi-source and continuous adaptation methods can achieve much better classification performance by leveraging the historical knowledge; (2) our proposed framework L2E outperforms state-of-the-art baselines in both the newest target task and all the historical target tasks. This confirms that L2E mitigates the issue of catastrophic forgetting on historical tasks when learning the new task.

Method	Acc	H-Acc
L2E w/o source evolution	$0.375_{\pm 0.044}$	$0.565_{\pm 0.018}$
L2E w merged source	$0.348_{\pm 0.070}$	$0.512_{\pm 0.024}$
L2E w/o historical target	$0.205_{\pm 0.014}$	$0.459_{\pm 0.016}$
L2E w all pairs	$0.387_{\pm 0.031}$	$0.569_{\pm 0.010}$
L2E	$\mathbf{0.435}_{\pm \mathbf{0.036}}$	$\mathbf{0.573}_{\pm \mathbf{0.021}}$

Table 3: Ablation study on Image-CLEF (B → P)

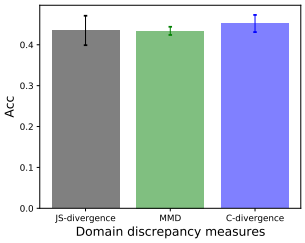


Figure 5: Impact of domain discrepancy measures

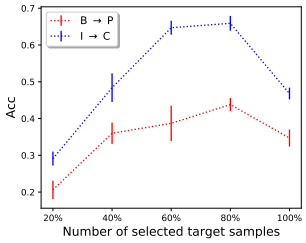


Figure 6: Impact of sampling selection strategy

5.3 Case Studies

Table 3 reports the results of several variants on Image-CLEF (B → P): (i) L2E w/o source evolution: using only source data at the initial time stamp; (ii) L2E w merged source: merging all source data into a large one; (iii) L2E w/o historical target: transferring the dynamic source tasks to the newest target task directly without historical target knowledge. It is observed that the evolution knowledge from historical source and target tasks could indeed improve the performance of L2E in dynamic transfer learning. Besides, we also consider to generate the meta-pairs from any two historical tasks (indicated in "L2E w all pairs" in Table 3). It could not outperform L2E with only the meta-pairs from consecutive tasks. One explanation is that it might generate the unrelated meta-pairs of tasks. Figure 5 shows the results of L2E instantiated with JS-divergence [Ganin *et al.*, 2016; Acuna *et al.*, 2021], MMD [Long *et al.*, 2015] and $\mathcal{C}$-divergence [Wu and He, 2020] on Image-CLEF (B → P). It indicates that our L2E framework is flexible to incorporate any domain discrepancy measures. Figure 6 shows the impact of sampling selection ratio $p\%$ on L2E where the classification accuracies on the newest target task (Acc) are reported on Image-CLEF. It confirms that selecting the historical target examples with high confidence positively affects the performance of L2E. Thus we choose $p = 80$ for our experiments.

6 Conclusion

In this paper, we study the transfer learning problem with dynamic source and target tasks. We show the error bounds of dynamic transfer learning on the newest target task in terms of historical source and target task knowledge. Then we propose a generic meta-learning framework L2E by minimizing the error upper bounds. Empirical results demonstrate the effectiveness of the proposed L2E framework.

Acknowledgements

This work is supported by National Science Foundation under Award No. IIS-1947203, IIS-2117902, IIS-2137468, 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

[Acuna *et al.*, 2021] David Acuna, Guojun Zhang, Marc T. Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *ICML*, 2021.

[Ben-David *et al.*, 2010] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 2010.

[Bobu *et al.*, 2018] Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. In *ICLR Workshop*, 2018.

[Fallah *et al.*, 2020] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *AISTATS*, 2020.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[Finn *et al.*, 2019] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *ICML*, 2019.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[Ghifary *et al.*, 2016] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *TPAMI*, 2016.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

[He and McAuley, 2016] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, pages 507–517, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hoffman *et al.*, 2014] Judy Hoffman, Trevor Darrell, and Kate Saenko. Continuous manifold based adaptation for evolving visual domains. In *CVPR*, 2014.

[Kumar *et al.*, 2020] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *ICML*, 2020.

[Li and Hoiem, 2017] Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 2017.

[Liu *et al.*, 2020] Hong Liu, Mingsheng Long, Jianmin Wang, and Yu Wang. Learning to adapt to evolving domains. *NeurIPS*, 2020.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[Mansour *et al.*, 2009] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

[Mohri and Medina, 2012] Mehryar Mohri and Andres Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *ALT*, 2012.

[Pan and Yang, 2009] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 2009.

[Peng *et al.*, 2019] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

[Rafailidis and Nanopoulos, 2015] Dimitrios Rafailidis and Alexandros Nanopoulos. Modeling users preference dynamics and side information in recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(6):782–792, 2015.

[Rosenstein *et al.*, 2005] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To transfer or not to transfer. In *NIPS workshop*, 2005.

[Sener and Koltun, 2018] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, pages 525–536, 2018.

[Sun *et al.*, 2019] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019.

[Tran *et al.*, 2019] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, 2019.

[Wang *et al.*, 2020] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *ICML*, 2020.

[Wen *et al.*, 2020] Junfeng Wen, Russell Greiner, and Dale Schuurmans. Domain aggregation networks for multi-source domain adaptation. In *ICML*, 2020.

[Wu and He, 2020] Jun Wu and Jingrui He. Continuous transfer learning with label-informed distribution alignment. *arXiv preprint arXiv:2006.03230*, 2020.

[Wu and He, 2021] Jun Wu and Jingrui He. Indirect invisible poisoning attacks on domain adaptation. In *KDD*, 2021.

[Zhang *et al.*, 2019] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, 2019.

[Zhao *et al.*, 2018] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *NeurIPS*, 2018.