Clarinet: A One-step Approach Towards Budget-friendly Unsupervised Domain Adaptation

Yiyang Zhang\textsuperscript{1,2,*}, Feng Liu\textsuperscript{2,*}, Zhen Fang\textsuperscript{2}, Bo Yuan\textsuperscript{1}, Guangquan Zhang\textsuperscript{2} and Jie Lu\textsuperscript{2,†}

\textsuperscript{1}Shenzhen International Graduate School, Tsinghua University
\textsuperscript{2}Centre for Artificial Intelligence, University of Technology Sydney

zhangyiy18@mails.tsinghua.edu.cn, feng.liu@uts.edu.au, zhen.fang@student.uts.edu.au, yuanb@sz.tsinghua.edu.cn, \{guangquan.zhang, jie.lu\}@uts.edu.au

Abstract

In unsupervised domain adaptation (UDA), classifiers for the target domain are trained with massive true-label data from the source domain and unlabeled data from the target domain. However, it may be difficult to collect fully-true-label data in a source domain given limited budget. To mitigate this problem, we consider a novel problem setting where the classifier for the target domain has to be trained with complementary-label data from the source domain and unlabeled data from the target domain named budget-friendly UDA (BFUDA). The key benefit is that it is much less costly to collect complementary-label source data (required by BFUDA) than collecting the true-label source data (required by ordinary UDA). To this end, complementary label adversarial network (CLARINET) is proposed to solve the BFUDA problem. CLARINET maintains two deep networks simultaneously, where one focuses on classifying complementary-label source data and the other takes care of the source-to-target distributional adaptation. Experiments show that CLARINET significantly outperforms a series of competent baselines.

1 Introduction

Domain Adaptation (DA) aims to train a target-domain classifier with data in source and target domains [Yan et al., 2017; Zhou et al., 2019b]. Based on the availability of data in the target domain (e.g., fully-labeled, partially-labeled and unlabeled), DA is divided into three categories: supervised DA [Sukhija et al., 2016], semi-supervised DA [Ao et al., 2017; Zhou et al., 2019a] and unsupervised DA (UDA) [Gong et al., 2018; Liu et al., 2020; Saito et al., 2017; Fang et al., 2019]. In practice, UDA methods have been applied to many real-world problems, such as object recognition [Agresti et al., 2019; Deng et al., 2019; Zhao et al., 2019].

UDA methods train a target-domain classifier with massive true-label data from the source domain (true-label source data) and unlabeled data from the target domain (unlabeled target data). Existing works in the literature can be roughly categorised into the following three groups: integral-probability-metrics based UDA [Long et al., 2015]; adversarial-training based UDA [Ganin et al., 2016; Long et al., 2018]; and causality-based UDA [Gong et al., 2016; Gong et al., 2018; Zhang et al., 2015]. Since adversarial-training based UDA methods extract better domain-invariant representations via deep networks, they usually have good target-domain accuracy [Sankaranarayanan et al., 2018].

However, the success of UDA still highly relies on the scale of true-label source data. Namely, the target-domain accuracy of a UDA method (e.g., CDAN) decays when the scale of true-label source data decreases. Hence, massive true-label source data are inevitably required by UDA methods, which is very expensive and prohibitive especially when the budget...
is limited [Liu et al., 2019]. This circumstance may hinder the promotion of DA to more areas.

While determining the correct label from many candidates is laborious, choosing one of the incorrect labels (i.e., complementary labels) would be much easier and thus less costly, especially when we have many candidates [Ishida et al., 2017]. For example, suppose that we need to annotate labels of a bunch of animal images from 1,000 candidates. One strategy is to ask crowd-workers to choose the true labels from 1,000 candidates, while the other is to judge the correctness of a label randomly given by the system from the candidates. Apparently, the cost of the second strategy is much lower than that of the first one [Ishida et al., 2019].

This brings us a novel problem setting, budget-friendly UDA (abbreviated as BFUDA), which aims to transfer knowledge from complementary-label source data to unlabeled target data (Figure 1). We describe this problem setting using the word budget-friendly since, compared to ordinary UDA, we can greatly save the labeling cost by annotating complementary labels in the source domain rather than annotating true labels [Ishida et al., 2017; Yu et al., 2018]. Please note that, existing UDA methods cannot handle BFUDA, as they require fully-true-label source data [Liu et al., 2020; Saito et al., 2017] or at least 20% true-label source data (demonstrated in [Liu et al., 2019; Shu et al., 2019]).

A straightforward but weak solution to BFUDA is a two-step approach, which sequentially combines complementary-label learning (CLL) methods and existing UDA methods (green dash line in Figure 1).1 CLL methods are used to assign pseudo labels for complementary-label source data. Then, we can train a target-domain classifier with pseudo-label source data and unlabeled target data using existing UDA methods. Nevertheless, pseudo-label source data contain noise, which may cause poor domain-adaptation performance of such two-step approach [Liu et al., 2019].

Therefore, we propose a powerful one-step solution to BFUDA, complementary label adversarial network (CLARINET). It maintains two deep networks trained by adversarial way simultaneously, where one can accurately classify complementary-label source data, and the other can discriminate source and target domains. Since Long et. al. [2018] and Song et. al. [2009] have shown that multimodal structures of distributions can only be captured sufficiently by the cross-covariance dependency between the features and classes (i.e., true labels), we set the input of domain discriminator $D$ as the outer product of feature representation (e.g., $g_s$ in Figure 2) and mapped classifier prediction (e.g., $T(f_s)$ in Figure 2).

Due to the nature of complementary-label classification, predicted probability of each class (i.e., each element of $f_s$, Figure 2) is relatively close. According to [Song et al., 2009], this kind of predicted probabilities could not provide sufficient information to capture the multimodal structure of distributions. To fix it, we propose a mapping function $T$ to make the predicted probabilities more scattered (i.e., $T(f_s)$, Figure 2) than previous ones (i.e., $f_s$, Figure 2). By doing so, the mapped classifier predictions can better indicate their choice. In this way, we can take full advantage of classifier predictions and effectively align distributions of two domains. Our ablation study (see Table 2) verifies that $T$ indeed helps improve the target-domain accuracy.

We conduct experiments on 6 BFUDA tasks and compare CLARINET with a series of competent baselines. Empirical results demonstrated that CLARINET can effectively transfer knowledge from complementary-label source data to unlabeled target data and is superior to all baselines.

### 2 Budget-friendly Unsupervised Domain Adaptation

In this section, we propose a novel problem setting, called budget-friendly unsupervised domain adaptation (BFUDA), and prove a learning bound of this new problem. Then, we show how BFUDA brings benefits to domain adaptation field.

#### 2.1 Problem Setting

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a feature (input) space and $\mathcal{Y} := \{1, \ldots, K\}$ be a label (output) space. A domain is defined as follows.

**Definition 1** (Domains for BFUDA). Given random variables $X_s, X_t \in \mathcal{X}, Y_s, Y_t \in \mathcal{Y}$, the source and target domains are joint distributions $P(X_s, Y_s)$ and $P(X_t, Y_t)$, where the joint distributions $P(X_s, Y_s) \neq P(X_t, Y_t)$ and $P(Y_s = c | Y_s = c) = 0$ for all $c \in \mathcal{Y}$.

Then, we propose BFUDA problem as follows.

**Problem 1** (BFUDA). Given independent and identically distributed (i.i.d.) labeled samples $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ drawn from the source domain $P(X_s, Y_s)$ and i.i.d. unlabeled samples $D_t = \{x_i^t\}_{i=1}^{n_t}$ drawn from the target marginal distribution $P(X_t)$, the aim of BFUDA is to train a classifier $F : \mathcal{X} \rightarrow \mathcal{Y}$ with $D_s$ and $D_t$ such that $F$ can accurately classify target data drawn from $P(X_t)$.

---

1. We implement this two-step approach and take it as a baseline.
It is clear that it is impossible to design a suitable learning procedure without any assumptions on \( P(X_s, Y_s) \). In this paper, we use the assumption for unbiased complementary learning proposed by [Ishida et al., 2019; Ishida et al., 2017]:

\[
P(Y_s = k|X_s) = \frac{1}{K-1} \sum_{c=1, c \neq k}^{K} P(Y_s = c|X_s),
\]

for all \( k, c \in \mathcal{Y} = \{1, \ldots, K\} \) and \( c \neq k \).

### 2.2 Learning Bound of BFUDA

This subsection presents a learning bound of BFUDA. Practitioner may safely skip it. The \( \ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+ \) is a loss function. The \( \text{decision function} \) is a vector-valued function \( h : \mathcal{X} \to \mathbb{R}^K \) and \( h_k \) is the \( k \)-th element of \( h \). The complementary risk for \( h \) with respect to \( \ell \) over \( P(X_s, Y_s) \) is

\[
L_s(h) = \mathbb{E}(\ell(h(X_s), Y_s)),
\]

where the risks for decision function \( h \) with respect to loss \( \ell \) over implicit distribution \( P(X_s, Y_s), P(X_t, Y_t) \) are:

\[
L_s(h) = \mathbb{E}(\ell(h(X_s), Y_s)), L_t(h) = \mathbb{E}(\ell(h(X_t), Y_t)).
\]

Then, we introduce our main theorem as follows.

**Theorem 1.** Given a loss function \( \ell \) and a hypothesis \( \mathcal{H} \subset \{h : \mathcal{X} \to \mathbb{R}^K\} \), then under unbiased assumption, for any \( h \in \mathcal{H} \), we have

\[
L_t(h) \leq L_s(h) + \frac{1}{2} d_\mathcal{H}^2(P_{X_s}, P_{X_t}) + \lambda,
\]

where \( L_s(h) := \sum_{k=1}^{K} \int_{\mathcal{X}} \ell(h(x), k) dP_{X_s} - (K-1)L_{\mathcal{Y}}(h) \)

\( P_{X_s}, P_{X_t} \) are source and target marginal distributions, \( \lambda = \min_{h \in \mathcal{H}} R_s(h) + R_t(h) \), and \( d_\mathcal{H}^2(P_{X_s}, P_{X_t}) \) is the distribution discrepancy defined in [Ben-David et al., 2010].

**Proof.** We firstly investigate the connection between \( L_s(h) \) and \( L_t(h) \) under unbiased assumption in Eq. (1). Given \( K \times K \) matrix \( Q \) whose diagonal elements are 0 and other elements are 1/\( K \), we represent the unbiased assumption by \( \eta = Q\eta \), where \( \eta = [P(Y_s = 1|X_s), \ldots, P(Y_s = K|X_s)]^T \) and \( \eta = [P(Y_s = 1|X_t), \ldots, P(Y_s = K|X_t)]^T \). Note that \( Q \) has inverse matrix \( Q^{-1} \) whose diagonal elements are \( -(K-2) \) and other elements are 1. Thus, we have that

\[
Q^{-1}\eta = \eta. \tag{2}
\]

According to Eq. (2), we have \( P(Y_s = k|X_s) = 1 - (K-1)P(Y_s = |X_s) \), which implies that

\[
L_s(h) = \sum_{k=1}^{K} \int_{\mathcal{X}} \ell(h(x), k) dP_{X_s} - (K-1)L_{\mathcal{Y}}(h). \tag{3}
\]

Hence, \( L_s(h) = \mathbb{E}(h) \). Combining Eq. (3) with the domain adaptation bound presented in Ben-David et al., 2010]

\[
L_t(h) \leq L_s(h) + \frac{1}{2} d_\mathcal{H}^2(P_{X_s}, P_{X_t}) + \lambda,
\]

we prove this theorem.

In this bound, the first term, i.e., Eq. (3), is the source classification error based on complementary-label source data. The second term is the distribution discrepancy distance between two domains and \( \lambda \) is the difference in labeling functions across the two domains. The empirical form of Eq. (3) is known as complementary-label loss (see Eq. (5)).

### 2.3 How does BFUDA Bring Benefits to DA Field?

Collecting true-label data is always expensive in the real world. Thus, learning from less expensive data [Kumar et al., 2017; Sakai et al., 2017] has been extensively studied in machine learning field, including label-noise learning [Han et al., 2018b; Han et al., 2018a], complementary-label learning [Ishida et al., 2019; Yu et al., 2018; Ishida et al., 2017] and so on. Among all these research directions, obtaining complementary labels is a cost-effective option. As described in the previous works mentioned above, compared with choosing the true class out of many candidate classes precisely, collecting complementary labels is obviously much easier and less costly. In addition, a classifier trained with complementary-label data is equivalent to a classifier trained with true-label data as shown in [Ishida et al., 2019].

At present, the success of DA still highly relies on the scale of true-label source data, which is a critical bottleneck. With limited budget, it is unrealistic to obtain enough true-label source data and thus cannot achieve a good distribution adaptation result. For the same budget, we can get multiple times more complementary-label data than the true-label data. In addition, the adaptation scenario is limited to some commonly used dataset, as they have sufficient true labels to support distributional adaptation. Thus if we can reduce the labeling cost in the source domain, for example, by using complementary-label data to replace true-label data (i.e. BFUDA), we can promote DA to more fields. Due to existing UDA methods require at least 20% true-label source data [Shu et al., 2019], they cannot handle BFUDA problem. To address BFUDA problem directly, we propose a powerful solution, CLARINET, as follows.

### 3 CLARINET: One-step BFUDA Approach

The proposed CLARINET (Figure 2) mainly consists of feature extractor \( G \), label predictor \( F \) and domain discriminator \( D \). Furthermore, we add a mapping function \( T \) between the label predictor \( F \) and domain discriminator \( D \) to take full advantage of classifier predictions. In this section, we first introduce two losses used to train CLARINET, and then the whole training procedures of CLARINET is presented.

#### 3.1 Loss Function in CLARINET

In this subsection, we show how to compute the two losses mentioned above in CLARINET after obtaining mini-batch \( \overline{d}_s \) from \( \overline{d}_s \) and \( d_t \) from \( d_t \).

**Complementary-label Loss.** We first divided \( \overline{d}_s \) into \( K \) disjoint subsets according to the complementary labels in \( \overline{d}_s \),

\[
\overline{d}_s = \bigcup_{k=1}^{K} \overline{d}_s,k, \quad \overline{d}_s,k = \{ (x^i_s, k) \}_{i=1}^{\pi_{s,k}}, \tag{4}
\]

where \( \overline{d}_s,k \cap \overline{d}_s,k' = \emptyset \) if \( k \neq k' \) and \( \pi_{s,k} = |\overline{d}_s,k| \). Then, following Eq. (3), the complementary-label loss on \( \overline{d}_s,k \) is

\[
\overline{L}_s(G, F, \overline{d}_s,k) = -(K-1) \frac{\pi_{F}}{\pi_{s,k}} \sum_{i=1}^{\pi_{s,k}} \ell(F \circ G(x^i_s), k)
\]

\[
+ \sum_{j=1}^{K} \frac{\pi_{F}}{\pi_{s,j}} \sum_{l=1}^{\pi_{s,j}} \ell(F \circ G(x^l_s), k), \tag{5}
\]

\[
\overline{L}_t(G, F, \overline{d}_s,k) = -(K-1) \frac{\pi_{F}}{\pi_{s,k}} \sum_{i=1}^{\pi_{s,k}} \ell(F \circ G(x^i_s), k)
\]

\[
+ \sum_{j=1}^{K} \frac{\pi_{F}}{\pi_{s,j}} \sum_{l=1}^{\pi_{s,j}} \ell(F \circ G(x^l_s), k),
\]
where ℓ can be any loss and we use the cross-entropy loss, \( \pi_k \) is the proportion of the samples complementary-labeled \( k \). The total complementary-label loss on \( \overline{d}_s \) is as follows.

\[
\overline{L}_s(G, F, \overline{d}_s) = \sum_{k=1}^{K} L_s(G, F, d_{s,k}).
\]  

As shown in Section 2.2, the complementary-label loss (i.e., Eq. (6)) is an unbiased estimator of the true-label data risk. Namely, the minimizer of complementary-label loss agrees with the minimizer of the true-label data risk with no constraints on the loss ℓ and model \( F \circ G \) [Ishida et al., 2019].

**Remark 1.** Due to the negative part in \( \overline{L}_s(G, F, \overline{d}_s) \), minimizing it will cause over-fitting [Kiryo et al., 2017]. To overcome this problem, we use a correctional way [Ishida et al., 2019] to minimize \( \overline{L}_s(G, F, \overline{d}_s) \) (lines 7-13 in Algorithm 1).

### 3.2 Training Procedure of CLARINET

Based on the two losses proposed in Section 3.1, in CLARINET, we try to solve the following optimization problem,

\[
\min_{G,F} \overline{L}_s(G, F, \overline{D}_s) - \lambda L_{adv}(G, F, \overline{D}_s, D_t),
\]

\[
\min_{D} L_{adv}(G, F, D, \overline{D}_s, D_t),
\]

where \( D \) tries to distinguish the samples from different domains by minimizing \( L_{adv} \), while \( F \circ G \) wants to maximize the \( L_{adv} \) to make domains indistinguishable. To solve the minimax optimization problem in Eq. (10), we add a gradient reversal layer [Ganin et al., 2016] between the domain discriminator and the classifier, which multiplies the gradient by a negative constant (-\( \lambda \)) during the back-propagation. \( \lambda \) is a hyper-parameter between the two losses to tradeoff source risk and domain discrepancy.

The training procedures of CLARINET are shown in Algorithm 1. First, we initialize the whole network (line 1) and shuffle the training set (line 3). During each epoch, after minibatch \( \overline{d}_s \) and \( d_t \) are fetched (line 5), we divide the source mini-batch \( \overline{d}_s \) into \( \{\overline{d}_{s,k}\}_{k=1}^{K} \) using Eq. (4) (line 6). Then, \( \{\overline{d}_{s,k}\}_{k=1}^{K} \) are used to calculate the complementary-label loss for each class (i.e., \( \sum_{k=1}^{K} \sum_{x \in \overline{d}_{s,k}} \omega_s(x) \log(D(\mathbf{g}(x))) \) (line 7).

**Algorithm 1 CLARINET: One-step BFUDA Approach**

**Input:** \( \overline{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}, \overline{D}_s = \{(x_i^t, y_i^t)\}_{i=1}^{n_t} \).

**Parameters:** learning rate \( \gamma_1 \) and \( \gamma_2 \), epoch \( T_{max} \), start epoch \( T_s \), iteration \( N_{max} \), class number \( K \), tradeoff \( \lambda \), network parameter \( \theta_{G,F,D} \) and \( \theta_D \).

**Output:** the neural network \( F \circ G \), namely the target domain classifier for \( D_t \).

1: Initialize \( \theta_{G,F} \) and \( \theta_D \);
2: for \( t = 1, 2, \ldots, T_{max} \) do
3: Shuffle the training set \( \overline{d}_s, D_t \);
4: for \( N = 1, 2, \ldots, N_{max} \) do
5: Fetch mini-batch \( \overline{d}_s, d_t \) from \( \overline{D}_s, D_t \);
6: Divide \( \overline{d}_s \) into \( \{\overline{d}_{s,k}\}_{k=1}^{K} \);
7: Calculate \( \{ L_s(G, F, \overline{d}_{s,k})\}_{k=1}^{K} \) using Eq. (5), and \( L_s(G, F, d_t) \) using Eq. (6);
8: if \( \min_k \{ L_s(G, F, \overline{d}_{s,k})\}_{k=1}^{K} \geq 0 \) then
9: Update \( \theta_{G,F} = \theta_{G,F} - \gamma_1 \nabla L_s(G, F, \overline{d}_s) \);
10: end if
11: else
12: Update \( \theta_{G,F} = \theta_{G,F} + \gamma_1 \nabla L_{neg} \);
13: end if
14: if \( t > T_s \) then
15: Calculate \( L_{adv}(G, F, D, \overline{d}_s, d_t) \) using Eq. (9);
16: Update \( \theta_D = \theta_D - \gamma_2 \nabla L_{adv}(G, F, D, \overline{d}_s, d_t) \);
17: Update \( \theta_{G,F} = \theta_{G,F} + \gamma_2 \lambda \nabla L_{adv}(G, F, D, \overline{d}_s, d_t) \);
18: end if
19: end for
20: end for

where \( \omega_s(x) \) and \( \omega_t(x) \) are \( 1 + e^{-H(F,G,x)} \), \( \mathbf{g}(x) \) is \( G(x) \odot T(F \circ G(x)) \) and \( \overline{d}_s[X] \) is the feature part of \( \overline{d}_s \).
4 Experiments

Based on five commonly used datasets: MNIST (M), USPS (U), SVHN (S), MNIST-M (m) and SYN-DIGITS (Y), we verify efficacy of CLARINET on 6 BFUDA tasks: $M \rightarrow U$, $U \rightarrow M$, $S \rightarrow M$, $M \rightarrow m$, $Y \rightarrow M$ and $Y \rightarrow S$. Note that, we generate complementary-label data according to [Ishida et al., 2019].

4.1 Baselines

We compare CLARINET with the following baselines: gradient ascent complementary label learning (GAC) [Ishida et al., 2019], namely non-transfer method, and several two-step methods, which sequentially combine GAC with UDA methods (including DAN [Long et al., 2015], DANN [Ganin et al., 2016] and CDAN [Long et al., 2018]). In two-step approach, GAC method is first used to assign pseudo labels for complementary-label source data. Then, we train the classifier with pseudo-label source data and unlabeled target data using UDA methods. Thus, we have four possible baselines: GAC, GAC+DAN, GAC+DANN and GAC+CDAN. For two-step methods, they share the same pseudo-label source data on each task. Note that, in this paper, we use the entropy conditioning variant of CDAN.

4.2 Experimental Setup

We design our feature extractor $G$, label predictor $F$ and domain discriminator $D$ according to the architecture from previous works. More precisely, we pick the structures of feature extractor $G$ from [Ganin et al., 2016; Long et al., 2018]. The label predictor $F$ and domain discriminator $D$ all share the same structure in all tasks, following CDAN [Long et al., 2018]. We follow the standard protocols for unsupervised domain adaptation and compare the average classification accuracy based on 5 random experiments. For each experiment, we take the result of the last epoch. The batch size is set to 128 and the number of epochs is set to 500. SGD optimizer (momentum = 0.9, weight decay = $5 \times 10^{-5}$) is with an initial learning rate of 0.005 in adversarial network and $5 \times 10^{-5}$ in classifier. In mapping function $T$, $l$ is set to 0.5. We update $\lambda$ according to [Long et al., 2018]. For parameters of each
baseline, we all follow the original settings. We implement all methods with default parameters by PyTorch. The code of CLARINET is available at github.com/Yiyang98/BFUDA.

4.3 Results on BFUDA Tasks

Table 1 reports the target-domain accuracy of 5 methods on 6 BFUDA tasks. As shown, our CLARINET performs best on each task and the average accuracy of CLARINET is significantly higher than those of baselines. Compared with GAC method, CLARINET successfully transfers knowledge from complementary-label source data to unlabeled target data. Since CDAN has shown much better adaptation performance than DANN and DAN [Long et al., 2018], GAC+CDAN should outperform other two-step methods on each task. However, on the task $U \rightarrow M$, the accuracy of GAC+CDAN is much lower than that of GAC+DANN. This abnormal phenomenon shows that the noise contained in pseudo-label source data significantly reduces transferability of existing UDA methods. Namely, we cannot obtain the reliable adaptation performance by using two-step BFUDA approach.

Figure 3 shows the target-domain accuracy of two-step methods and CLARINET on 6 BFUDA tasks when increasing epochs. It is clear to see that the accuracy of CLARINET gradually and steadily increases and eventually converges, achieving the best accuracy on each task. The accuracy of GAC+CDAN always reaches plateau quickly. For GAC+DANN, its accuracy is unstable and significantly drops after certain epochs on $M \rightarrow U$, $M \rightarrow m$ tasks. While the accuracy of GAC+DAN is relatively stable but not satisfactory.

4.4 Ablation Study

Finally, we conduct experiments to show the contributions of components in CLARINET. We consider following baselines:

- CLARINET by Algorithm 1, while replacing $L_s(G, F, D_s)$ by cross-entropy loss.
- GAC+DAN: train CLARINET by Algorithm 1, without adversarial network.
- GAC+DANN: train CLARINET by Algorithm 1, without adversarial network.
- GAC+CDAN: train CLARINET by Algorithm 1, without adversarial network.
- CLARINET: train CLARINET by Algorithm 1, without the mapping function $T$.

C w/ $L_{CE}$ uses the cross-entropy loss to replace the complementary-label loss. The target-domain accuracy of C w/ $L_{CE}$ will show if UDA methods can address the BFUDA problem. Comparing CLARINET with C w/o $T$ reveals whether the mapping function $T$ takes effect. As shown in Table 2, the target-domain accuracy of C w/ $L_{CE}$ is much lower than that of other methods. Namely UDA methods cannot handle BFUDA directly. Although C w/o $T$ achieves better accuracy than two-step methods, its accuracy is still worse than CLARINET’s. Thus, the mapping function $T$ can help improve the adaptation performance under BFUDA.

5 Conclusion

This paper presents a new problem setting for the domain adaptation field, called budget-friendly unsupervised domain adaptation (BFUDA), which exploits economical complementary-label source data instead of expensive true-label source data. Since existing UDA methods cannot address BFUDA problem, we propose a novel one-step BFUDA approach, called complementary label adversarial network (CLARINET). Experiments conducted on 6 BFUDA tasks confirm that CLARINET effectively achieves distributional adaptation from complementary-label source data to unlabeled target data and outperforms competitive baselines.

Acknowledgements

The work presented in this paper was supported by the Australian Research Council (ARC) under FL190100149 and DP170101632. We would like to thank the anonymous reviewers for their thoughtful comments. The first author particularly thanks the support by UTS-CAI during her visit.
References

[Agresti et al., 2019] G. Agresti, H. Schäfer, P. Sartor, and P. Zanuttigh. Unsupervised domain adaptation for tof data denoising with adversarial learning. In CVPR, pages 5584–5593, 2019.

[Ao et al., 2017] S. Ao, X. Li, and C.X. Ling. Fast generalized distillation for semi-supervised domain adaptation. In AAAI, pages 1719–1725, 2017.

[Ben-David et al., 2010] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. MLJ, 79(1-2):151–175, 2010.

[Deng et al., 2019] Z. Deng, Y. Luo, and J. Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In ICCV, pages 9944–9953, 2019.

[Fang et al., 2019] Zhen Fang, Jie Lu, Feng Liu, and Guangquan Zhang. Unsupervised domain adaptation with sphere retracting transformation. In IJCNN, pages 1–8. IEEE, 2019.

[Ganin et al., 2016] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. JMLR, 17(59):1–35, 2016.

[Gong et al., 2016] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In ICML, pages 2839–2848, 2016.

[Gong et al., 2018] M. Gong, K. Zhang, B. Huang, C. Glymour, D. Tao, and K. Batmanghelich. Causal generative domain adaptation networks. CoRR, abs/1804.04333, 2018.

[Han et al., 2018a] B. Han, J. Yao, G. Niu, M. Zhou, I. Tsang, Y. Zhang, and M. Sugiyama. Masking: A new perspective of noisy supervision. In NeurIPS, pages 5836–5846, 2018.

[Han et al., 2018b] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In NeurIPS, pages 8527–8537, 2018.

[Ishida et al., 2017] T. Ishida, G. Niu, W. Hu, and M. Sugiyama. Learning from complementary labels. In NeurIPS, pages 5639–5649, 2017.

[Ishida et al., 2019] T. Ishida, G. Niu, A. Menon, and M. Sugiyama. Complementary-label learning for arbitrary losses and models. In ICML, pages 2971–2980, 2019.

[Kiryo et al., 2017] R. Kiryo, G. Niu, M.C. du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In NeurIPS, pages 1675–1685, 2017.

[Kumar et al., 2017] A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In NeurIPS, pages 5534–5544, 2017.

[Liu et al., 2019] F. Liu, J. Lu, B. Han, G. Niu, G. Zhang, and M. Sugiyama. Butterfly: A panacea for all difficulties in wildly unsupervised domain adaptation. In NeurIPS LTS Workshop, 2019.

[Liu et al., 2020] F. Liu, G. Zhang, and J. Lu. Heterogeneous domain adaptation: An unsupervised approach. TNINLS, Early Access, 2020.

[Long et al., 2015] M. Long, Y. Cao, J. Wang, and M.I. Jordan. Learning transferable features with deep adaptation networks. In ICML, page 97–105, 2015.

[Long et al., 2018] M. Long, Z. Cao, J. Wang, and M.I. Jordan. Conditional adversarial domain adaptation. In NeurIPS, pages 1640–1650, 2018.

[Saito et al., 2017] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. In ICML, pages 2988–2997, 2017.

[Sakai et al., 2017] T. Sakai, M.C. du Plessis, G. Niu, and M. Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In ICML, pages 2998–3006, 2017.

[Sankaranarayanan et al., 2018] S. Sankaranarayanan, Y. Balaji, C.D. Castillo, and R. Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In CVPR, pages 8503–8512, 2018.

[Shu et al., 2019] Y. Shu, Z. Cao, M. Long, and J. Wang. Transferable curriculum for weakly-supervised domain adaptation. In AAAI, pages 4951–4958, 2019.

[Song et al., 2009] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In ICML, pages 961–968, 2009.

[Sukhija et al., 2016] S. Sukhija, N.C. Krishnan, and G. Singh. Supervised heterogeneous domain adaptation via random forests. In IJCAI, pages 2039–2045, 2016.

[Yan et al., 2017] Y. Yan, W. Li, M. K. P. Ng, M. Tan, H. Wu, H. Min, and Q. Wu. Learning discriminative correlation subspace for heterogeneous domain adaptation. In IJCAI, pages 3252–3258, 2017.

[Yu et al., 2018] X. Yu, T. Liu, M. Gong, and D. Tao. Learning with biased complementary labels. In ECCV, pages 69–85, 2018.

[Zhang et al., 2015] K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In AAAI, pages 3150–3157, 2015.

[Zhao et al., 2019] S. Zhao, H. Fu, M. Gong, and D. Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In CVPR, pages 9788–9798, 2019.

[Zhou et al., 2019a] J. T. Zhou, S. J. Pan, and I. W. Tsang. A deep learning framework for hybrid heterogeneous transfer learning. AJJ, 275:310 – 328, 2019.

[Zhou et al., 2019b] J. T. Zhou, H. Zhang, D. Jin, and X. Peng. Dual adversarial transfer for sequence labeling. TPAMI, Early Access, 2019.