Weak Adaptation Learning: Addressing Cross-domain Data Insufficiency with Weak Annotator

Shichao Xu, Lixu Wang, Yixuan Wang, Qi Zhu
Northwestern University, Evanston, USA
{shichaoxu2023, lixuwang2025, yixuanwang2024}@u.northwestern.edu, qzhu@northwestern.edu

Abstract

Data quantity and quality are crucial factors for data-driven learning methods. In some target problem domains, there are not many data samples available, which could significantly hinder the learning process. While data from similar domains may be leveraged to help through domain adaptation, obtaining high-quality labeled data for those source domains themselves could be difficult or costly. To address such challenges on data insufficiency for classification problem in a target domain, we propose a weak adaptation learning (WAL) approach that leverages unlabeled data from a similar source domain, a low-cost weak annotator that produces labels based on task-specific heuristics, labeling rules, or other methods (albeit with inaccuracy), and a small amount of labeled data in the target domain. Our approach first conducts a theoretical analysis on the error bound of the trained classifier with respect to the data quantity and the performance of the weak annotator, and then introduces a multi-stage weak adaptation learning method to learn an accurate classifier by lowering the error bound. Our experiments demonstrate the effectiveness of our approach in learning an accurate classifier with limited labeled data in the target domain and unlabeled data in the source domain.

1. Introduction

Machine Learning (ML) techniques, especially those based on deep neural networks, have shown great promises in many applications, to a large extent due to their abilities in studying and memorizing the knowledge embedded in high-quality training data [12]. Having a large number of data samples with accurate labels could enable effective supervised learning methods for improving ML model performance. However, it may be difficult to collect many data samples in some problem domains or scenarios, such as for the training of autonomous vehicles during extreme weather (e.g., fog, snow, hail) and natural disasters (e.g., mudflow), or for search and rescue robots during forest fire and earthquake. One possible solution to such problem of data unavailability is using data from other similar domains to train the target domain model and then fine-tune it with limited target domain data, i.e., through domain adaptation. Taking the aforementioned cases as examples, while there may not be much data in hailing weather, we could collect data in days with heavy rain; while it may be difficult to find images during earthquakes for large parts of America, we could collect images in Japan, where earthquakes occur more often in a different environment. However, obtaining a large amount of high-quality labeled data in these source domains could still be challenging and costly.

To address the above data insufficiency challenges across domains, we consider leveraging low-cost weak annotators that can automatically generate large quantity of labeled data based on certain labeling rules/functions, task-specific heuristics, or other methods (which may be inaccurate to some degree). More specifically, our approach considers the following setting for classification problems: There is a small amount of data samples with accurate labels collected for the target domain, which is called target domain data or target data in this paper for simplicity. There is also a large amount of unlabeled data that can be acquired from a similar but different source domain (i.e., there exists domain discrepancy), which is called source (domain) data in this paper. Finally, there is a weak annotator that can produce weak (possibly inaccurate) labels on data samples. Our objective is to learn an accurate classifier for the target domain based on the labeled target data, the initially-unlabeled source data, and the weak annotator.

The problem we are considering here is related but different from Semi-Supervised Learning (SSL) [39, 9, 23] and Unsupervised Domain Adaptation (UDA) [24, 8, 47, 7]. In the setting of SSL, the available training data consists of two parts – one has accurate labels while the other is unlabeled, and the two parts are drawn from the same distribution in terms of training features. This is different from our problem, where there exists domain discrepancy across
the source and target domains. The objective of UDA is to adapt a model to perform well in the target domain based on labeled data in the source domain and unlabeled data in the target domain. This is again very different from our problem, where the source domain data is initially unlabeled and assigned with inaccurate labels by a weak annotator, while the target domain data has labels but its quantity is small. Another related field is Positive-unlabeled Learning (PuL) [20, 5], an approach for sample selection. The training data of PuL also consists of two parts – positive and negative data, and the task is to learn a binary classifier to filter out samples that are similar to the positive data from a large amount of negative data. However, the current PuL approaches usually conduct experiments in a single data set rather than multiple domains with feature discrepancy.

To solve our target problem, we first develop a theoretical analysis on the error bound of a trained classifier with respect to the data quantity and the weak annotator performance. We then propose a Weak Adaptation Learning (WAL) method to learn an accurate classifier by lowering the error bound. The main idea of WAL is to obtain a cross-domain representation for both source domain and target domain data, and then use the labeled data to estimate the classification error/distance between the weak annotator and the ideally optimal classifier in the target domain. Next, all the data is re-labeled based on such estimation of weak annotator classification error. Finally, the newly-relabeled data is used to learn a better classifier in the target domain.

Our work makes the following contributions:

- We address the challenge of data insufficiency in domain adaptation with a novel weak adaptation learning approach that leverages unlabeled source domain data, limited number of labeled target domain data, and a weak annotator.
- Our approach includes a theoretical analysis on the error bound of the trained classifier and a multi-stage WAL method that improves the classifier accuracy by lowering such error bound.
- We compare our approach with various baselines in experiments with domain discrepancy setting on several digit datasets and the VisDA-C dataset, and study the cases without domain discrepancy on the CIFAR-10 dataset. We also conduct ablation studies on the impact from the weak annotator accuracy and the quantity of labeled data samples to further validate our ideas.

2. Related Work

We introduce related works in the topics about weakly- and semi-supervised learning, and the importance of sample quantity here. You can also find more related works about domain adaptation in the supplementary materials.

2.1. Weakly- and Semi-Supervised Learning

Weakly Supervised Learning is a large concept that may have multiple problem settings [51]. The problem we consider in this paper is related to the incomplete supervision setting that is often addressed by Semi-Supervised Learning (SSL) approaches. Standard SSL solves the problem of training a model with a few labeled data and a large amount of unlabeled data. Some of the widely-applied methods [39, 9, 36, 2] assign pseudo labels to unlabeled samples and then perform supervised learning. And there are works that address the noises in the labels of those samples [31, 11, 25]. Our target problem is related to SSL with inaccurate supervision, but is different since we consider the feature discrepancy between the (unlabeled) source data and the (labeled) target data – a case that occurs often in practice but has not been sufficiently addressed.

Positive-unlabeled Learning (PuL) is usually regarded as a sub-problem of SSL. Its goal is to learn a binary classifier to distinguish positive and negative samples from a large amount of unlabeled data and a few positive samples. Several works [20, 5] can achieve great performance on selecting samples that are similar to the positive data, and there are also works using samples selected by PuL to perform other tasks [49, 26].

2.2. Importance of Sample Quantity

The training of machine learning models, especially deep neural networks, often requires a large amount of data samples. However, in many practical scenarios, there is not sufficient training data to feed the learning process, degrading the model performance sharply [41, 16, 48]. Many approaches have been proposed to make up for the lack of training samples, e.g., data re-sampling [44], data augmentation [37], metric learning and meta learning [3, 4, 42, 45]. And there are works [33, 1, 3, 46] conducting theoretical analysis on the relation between training data quantity and model performance. These analyses are usually in the form of bounding the prediction error of the models and provide valuable information on how the sample quantity of training data affects the model performance. In our work, we also perform a theoretical analysis on the error bound of the trained model, with respect to not only the data quantity but also the performance of the weak annotator.

3. Theoretical Analysis

3.1. Problem Definition and Formulation

We consider the task of classification, where the goal is to predict labels for samples in the target domain. Two types of supporting data can be accessed for training the model – source domain data and target domain data. The source domain data samples are initially unlabeled and come from a joint probability distribution \( Q^s \). They can be labeled by a
weak annotator $h^w$ (which may be inaccurate) and denoted as $D_s = \{(x_i, y_i)\}_{i=1}^{N_s}$, where $N_s$ is the number of source data samples. The target domain data $D_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ consists of $N_t$ samples collected from the target distribution $Q^t$. Note that $Q^t$ may be different from $Q^S$. And we use $Q_X^s$, $Q_Y^s$, $Q_X^t$, $Q_Y^t$ to represent the marginal distributions of the source and target domains, respectively. Moreover, as stated before, we consider the case where there is only a small amount of target domain data, i.e., $N_t \ll N_s$.

Our goal is to learn an accurate classifier for the target domain. The classifier is initialized from a parameter distribution $\mathcal{H}$, which denotes the hypothesis parameter space of all possible classifiers.

In the following analysis, we will define the classification risk of a classifier and then derive its bound. According to the PAC-Bayesian framework [30, 10], the expected classification risk of a classifier drawn from a distribution $\mathcal{Q}$ that depends on the training data can be strictly bounded. Let $h_{\Theta}$ denote a learned classifier from the training data, and its parameter $\Theta$ is drawn from $\mathcal{Q}$. We consider that the prior parameter distribution $\mathcal{H}$ over the hypothesis is independent of the training data. And given a $\delta$ with the probability $\geq 1 - \delta$ over the training data set of size $m$, the expected error of $h_{\Theta}$ can be bounded as follows [29]:

$$L(h_{\Theta}) \leq \hat{L}(h_{\Theta}) + \sqrt{\hat{L}(h_{\Theta}) \cdot \Omega + \Omega}$$

$$\Omega = \frac{2 (KL(Q \| \mathcal{H}) + \ln \frac{m}{\delta})}{m - 1}$$

Here $L(h_{\Theta})$ is the expected error of $h$ over parameter $\Theta$, and $\hat{L}(h_{\Theta})$ is the empirical error computed from the training set $\{(x_i, y_i)\}_{i=1}^{m} \sim \mathcal{L}(x_i, y_i)$, where $\mathcal{L}$ denotes the loss of a single training sample. In Eq. (1), $KL(Q \| \mathcal{H})$ represents the Kullback-Leibler (KL) divergence between parameter distribution $Q$ and $\mathcal{H}$. For any two distributions $p, q$, the specific form of their KL divergence is

$$KL(p \| q) = -\mathbb{E}[p \cdot \ln \frac{q}{p}]$$

In most cases of mini-batch training, the training loss $\hat{L}(h_{\Theta})$ is much smaller than $\Omega$, and thus we can get a further bound as follows [33]:

$$L(h_{\Theta}) \leq \hat{L}(h_{\Theta}) + 4 \sqrt{\frac{KL(Q \| \mathcal{H}) + \ln \frac{2m}{\delta}}{m}}$$

Then if we denote the model parameters of $h$ before the training that are drawn from $\mathcal{H}$ as $\Theta^p$, the KL divergence can be written as $KL(Q \| \mathcal{H}) = -\mathbb{E}[\Theta \cdot (\ln \Theta^p - \ln \Theta)]$. As aforementioned, $h_{\Theta}$ is trained with the training data set from $h_{\Theta^p}$, and we consider that the training is optimized by gradient-based method. Thus, we can formulate that $\Theta = \Theta^p + \nabla \hat{L}(h_{\Theta^p})$. Here we omit the learning rate to simplify the formula.

The PAC-Bayesian error bound is valid for any parameter distribution $\mathcal{H}$ that is independent of the training data, and any method of optimizing $\Theta^p$ dependent on the training set [33]. Therefore, in order to simplify the problem, we instatiate the bound as setting $\mathcal{H}$ to confrom to a Gaussian distribution with zero mean ($\mu_\mathcal{H} = 0$) and $\sigma_\mathcal{H}^2$ variance. This simplification is the same as previous PAC-Bayesian works [33, 34]. We further assume that the parameter change of the overall model during training can also be regarded as conforming to an empirical Gaussian distribution. This Gaussian distribution is independent of model parameters if we regard the parameter updates induced by gradient back-propagation as accumulated random perturbations, i.e., each training sample corresponds to a small perturbation [34]. And we denote the mean and the variance of a single training sample as follows:

$$\mu \triangleq \mathbb{E}[\nabla_{\Theta^p} \mathcal{L}(x, y)]$$

$$\sigma^2 \triangleq \mathbb{E}[(\nabla_{\Theta^p} \mathcal{L}(x, y) - \mu)(\nabla_{\Theta^p} \mathcal{L}(x, y) - \mu)^T]$$

Then, the specific formula of KL divergence to any two Gaussian distributions $p \sim N(\mu_1, \sigma_1^2)$ and $q \sim N(\mu_2, \sigma_2^2)$ is written as follows:

$$KL(p \| q) = \ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

**Theorem 1.** For a classifier parameter distribution $\mathcal{H} \sim N(0, \sigma_\mathcal{H}^2)$ that is independent of the training data with size $m$, and a posterior parameter distribution $\mathcal{Q}$ learned from the training data set, if we assume $\mathcal{Q} \sim N(\mu_\mathcal{Q}, \sigma_\mathcal{Q}^2)$ and consider $\Theta^p$, $\Theta$ as drawn from $\mathcal{H}$, $\mathcal{Q}$ respectively ($\Theta = \Theta^p + \nabla \hat{L}(h_{\Theta^p})$), the KL divergence of $\mathcal{Q}$ and $\mathcal{H}$ is bounded with symbols defined in Eq. (3) as follows:

$$KL(Q \| \mathcal{H}) \leq \frac{\sigma^2 + \mu^2}{2\sigma_\mathcal{H}^2}$$

The detailed proof of Theorem 1 is presented in our Supplementary Materials. With the above risk definition, the risk of $h$ with respect to the target data distribution $Q^t$ is

$$R^t(h) = \mathbb{E}_{(x, y) \sim Q^t}[\mathcal{L}(h(x), y)] = L(h_{\Theta}) \cdot Q^t$$

Besides, we define the Classification Distance of two classifiers $h_1$ and $h_2$ under the same domain distribution $\mathbb{P}$ as

$$CD_{\mathbb{P}}(h_1, h_2) = \mathbb{E}_{x \sim \mathbb{P}}[\mathcal{L}(h_1(x), h_2(x))]$$

Moreover, the Discrepancy Distance of two domains is defined as in [28]: $\forall h_1, h_2$, the discrepancy distance between the distributions of two domains $\mathbb{P}, \mathbb{Q}$ is

$$DD(\mathbb{P}, \mathbb{Q}) = \sup_{h_1, h_2 \in \mathcal{H}} |CD_{\mathbb{P}}(h_1, h_2) - CD_{\mathbb{Q}}(h_1, h_2)|$$

For further analysis, we also define two operators in a parameter distribution $\mathcal{H}$:

- $\oplus$: $\forall h_1, h_2 \in \mathcal{H}$, and $\forall x \in \mathbb{P}$, a new classifier $h_3 = h_1 \oplus h_2$ can be acquired by conducting operator $\oplus$ on $h_1$ and $h_2$, and $h_3(x) = h_1(x) + h_2(x)$.
3.2. Error Bound Analysis

Let $h^*$ and $h^{\alpha}$ denote the ideal classifiers that perform optimally on the source data and target data, respectively:

$$h^* = \arg\min_{h \in Q} R^s(h), \quad h^{\alpha} = \arg\min_{h \in Q} R^t(h)$$

(9)

In our approach, we design a classifier that learns the discrepancy between the weak annotator and the ground truth (details will be introduced in Section 4), and we denote it as $d$ drawn from $Q$. Thus, we can get a model that is the product of conducting the aforementioned $\oplus$ operator on $h$ and $d$, i.e., $h \oplus d$. Here $h$ is designed for approximating the weak labels. And for the risk of $h \oplus d$, we can obtain the following relation:

**Theorem 2.** For all L1 (Mean Absolute Error [17]), L2 (Mean Squared Error [14]) and their non-negative combination loss functions (Huber Loss [50], Quantile Loss [21], etc.), the classification risk of aforementioned $h \oplus d$ can be formulated as follows:

$$R^t(h \oplus d) = E_{Q_X^t}\mathcal{L}(h \oplus d, h^w \oplus h^{\alpha} \oplus h^w)$$

$$\leq E_{Q_X^t}\mathcal{L}(h, h^w) + E_{Q_X^t}\mathcal{L}(d, h^{\alpha} \oplus h^w)$$

(10)

Please refer to our Supplementary Materials for detailed proof of Theorem 2. Then if we consider that the training loss $\tilde{L}(h)$ (which equals the average loss of all training samples) is hardly influenced by the sample quantity, and it is the same for the discrepancy between two domains [27], we can split the error bound of $h \oplus d$ into two parts, where one part, denoted as $\Delta$, is not influenced by the sample quantity and the other is related to the sample quantity. According to Eq. (2), these two parts can be written as follows (the detailed derivation of inequalities starting from Eq. (10) can be found in the Supplementary Materials):

$$R^t(h \oplus d) = E_{Q_X^t}\mathcal{L}(h \oplus d, h^w \oplus h^{\alpha} \oplus h^w)$$

$$\leq \Delta + 4 \frac{KL_d}{N_t} + 4 \frac{KL_h}{N_s} + 12 \frac{\ln(2N_s)}{N_t} + 8 \frac{\ln(2N_s)}{N_s}$$

(11)

where $\Delta = 2\tilde{L}_t(h^w) + \tilde{L}_t(d) + \tilde{L}_s(h)$

$$+ \tilde{L}_s(h^w) + DD(Q_X^t, Q_X^s)$$

Here $KL_d$ and $KL_h$ denote KL divergences between trained $d$, $h$ and $\mathcal{H}$ respectively. According to Theorem 1, this KL divergence term is influenced by the training, especially impacted by the sample quantity. We will discuss the insights obtained from this error bound in the next section, and then introduce our weak adaptation learning process that is inspired by those insights.

4. Weak Adaptation Learning Method

4.1. Observation from Error Analysis

Based on the error bound derived in Eq. (11), we can put efforts into the following ideas in our approach to improve the classifier performance in the target domain:

- **Performance of annotator** ($2\tilde{L}_t(h^w) + \tilde{L}_s(h^w)$): The supervision provided by the weak annotator can guide the model to better target the given task. Ideally, we want $h^w$ to produce more accurate labels for both source and target data, reducing $2\tilde{L}_t(h^w)$ and $\tilde{L}_s(h^w)$ simultaneously. Practically though, we may just be able to make the annotator perform better on the source domain and cannot do much with the target domain.

- **Discrepancy between domains** ($DD(Q_X^t, Q_X^s)$): Designing loss to quantify the discrepancy between the source and target domains is well studied in Domain Adaptation. In our approach, we propose a novel inter-domain loss (called Classified-MMD) to minimize $DD(Q_X^t, Q_X^s)$, similar to the introduction later.

- **Quantity of source and target samples** ($N_s, N_t$): First, the learning of $d$ needs the supervision of the ground truth, and thus we can only use the labeled target data to train $d$. Then, in our method, $h$ is designed to approximate the weak annotator, and therefore it may see enough that we just use the source data to train $h$. However, to further reduce $KL_h$, according to Theorem 1, we also use target samples to train $h$, which increases the sample size of training data. Moreover, since the sample quantity of source data is much larger than that of target data (i.e., $N_t \ll N_s$), $\sqrt{KL_h/N_t}$ in Eq. (11) dominates over $\frac{KL_d}{N_t}$ and in the case of $\delta \leq 2/\epsilon$, $12 \frac{\ln(2N_s)}{N_t}$ also strictly dominates over $8 \frac{\ln(2N_s)}{N_s}$. As the result, the terms influenced by a few target samples dominates the overall error risk. Therefore, directly applying $h \oplus d$ to the target domain will still be impacted by the insufficient samples. However, note that $h \oplus d$ can produce more accurate labels for the source data than the weak annotator. Therefore, we add a final step in our learning process that utilizes re-labeled source data and conducts supervised learning with such augmentation.

4.2. Learning Process

In this section, we present the detailed process of our weak adaptation learning (WAL) method, which is designed based on the observations from the above error bound analysis. The overview of our WAL process is shown in Figure 1. The designed network consists of three parts – (\Phi_0, \Phi_1, \Phi_2). \Phi_0 can be seen as a shared feature network for both source
and target data, using typical classification networks such as VGG, ResNet, etc. \( \Phi_1 \) consists of three fully-connected layers that follow the output of \( \Phi_0 \). And we denote the combination of \( \Phi_0 \) and \( \Phi_1 \) as \( F_1 \). \( \Phi_2 \) consists of two fully-connected layers that follow the output of \( \Phi_0 \). The combination of \( \Phi_0 \) and \( \Phi_2 \) is denoted as \( F_2 \). The detailed network architecture is shown in the Supplementary Materials. The workflow of our method is shown in Algorithm 1.

Algorithm 1 The workflow of Weak Adaptation Learning.

1: Initialize parameters of network components \( \Phi_0, \Phi_1, \Phi_2 \).
2: Obtain dataset \( D \) from the source and target data with the help of weak annotator \( \mathbf{h}^w \).
3: Train \( F_1 = \Phi_1 \circ \Phi_0 \) using \( D \), with loss function following equation \( \mathcal{L} = \mathcal{L}_{KL} + \alpha \mathcal{L}_{cmmd} \).
4: Fix the parameters of \( \Phi_1 \) and use \( F_2 = \Phi_2 \circ \Phi_0 \) to fit the distance of the optimal classifier for target data \( \mathbf{h}^w \) and the weak annotator \( \mathbf{h}^w \) with the target data.
5: Generate a new dataset using both source and target data. The new labels are calculated by \( y^{\text{new}} = \mathbf{h}^w(\mathbf{x}) + \Phi_2(\mathbf{h}^w(\mathbf{x}), \Phi_0(\mathbf{x})) \).
6: Initialize parameters of \( \Phi_0, \Phi_1, \Phi_2 \).
7: Fix \( \Phi_2 \) and train \( F_1 \) using the new dataset. The loss function follows \( \mathcal{L} = \mathcal{L}_{KL} + \alpha \mathcal{L}_{cmmd} \).
8: Output classifier \( F_1 \).

Stage 1: The first goal we step on is to obtain a common representation for both the source and target data, which helps us encode the inputs while mitigating the domain discrepancy in the feature representation. We gather all the unlabeled source data and the target data without their labels and use weak annotator \( \mathbf{h}^w \) to assign a label for each data sample \( \mathbf{x}_i \) and \( y^w_i = \mathbf{h}^w(\mathbf{x}_i) \). We denote the dataset obtained in this way as \( D = \{(\mathbf{x}_i, y^w_i)\}_{i=1}^{N_s+N_t} \). Then we fix \( \Phi_2 \) and only consider the left part of the network, which is \( F_1 = \Phi_1 \circ \Phi_0 \). It is normally trained by supervised learning using the dataset \( D \) for \( c_p \) training epochs, and uses the following loss function:

\[
\mathcal{L} = \mathcal{L}_{KL} + \alpha \mathcal{L}_{cmmd} \tag{12}
\]

In this loss function, there are two loss terms and the hyper-parameter \( \alpha \) is a scaling factor to balance the scale of two loss functions (we set it as 0.0001 in our experiments). The first term \( \mathcal{L}_{KL} \) is the Kullback-Leibler (KL) divergence loss, stated as follows:

\[
\mathcal{L}_{KL} = KL(y^1_{\text{pre}} || y^w) = KL(\Phi_1 \circ \Phi_0(\mathbf{x}) || \mathbf{h}^w(\mathbf{x})) \tag{13}
\]

where \( y^1_{\text{pre}} \) is the output prediction value of \( F_1 \) and \( y^w \) is the corresponding weak label produced by the weak annotator \( \mathbf{h}^w \). The second term \( \mathcal{L}_{cmmd} \) aims to mitigate the domain
discrepancy of the source and target domain at the feature representation level in the neural networks. Based on the basic MMD loss introduced by [43], we further change it into the version with data labels. We call this loss function as Classified-MMD loss (corresponding to the subscript \(cmmd\)), which is defined as:

\[
\mathcal{L}_{cmmd} = \frac{1}{M} \sum_{i=1}^{M} \left( \frac{1}{|D_X^{(S)}|} \sum_{x_s \in D_X^{(S,i)}} F_1(x_s) - \frac{1}{|D_X^{(T)}|} \sum_{x_t \in D_X^{(T,i)}} F_1(x_t) \right)
\]

(14)

where \(M\) is the number of classes, \(D_X\) is the data from the produced dataset \(D\) without labels, and \(D_X^{(S,i)}\) is the source data selected from \(D_X\) with \(\arg\max(y^w) = i\). Then, we utilize target data with its accurate labels to continue to train the network component \(F_1\) under the loss function \(\mathcal{L}_{KL}\) for \(ep_t\) training epochs, which helps further fine-tune the feature we learned through accurate labels of the target data.

**Stage 2:** After finishing training in Stage 1, the next step is to estimate the distance of the optimal classifier for target data \(\Phi^w\) and the weak annotator \(\Phi^w\). We estimate this distance through available target data with accurate labels. We adopt the parameters trained from Stage 1 and train network component \(F_2 = \Phi_2 \circ \Phi_0\) using the target data \(D_t\). For an input data sample \(x\), it is brought into both \(\Phi_0\) and the weak annotator as their input. Then \(\Phi_2\) takes the output feature of \(\Phi_0(x)\) and \(\Phi^w(x)\) as input feature (these two features are concatenated as the input feature of \(\Phi_2\)). For data sample \((x_t, y_t) \in target\ dataset\ D_t\), the learning of \(F_2\) uses the following classifier discrepancy loss function:

\[
\mathcal{L}_{MSE} = \| \Phi_2(\Phi^w(x_t), \Phi_0(x_t)) - (y_t - \Phi^w(x_t)) \|^2
\]

(15)

The network is trained for \(ep_3\) training epochs.

**Stage 3:** The third step is to generate a new dataset \(D_{new}\) through the obtained network \(F_2\) above. Specifically, we collect data \(x\) from both source data and target data, and we re-label these data based on the weak annotator and \(F_2\) obtained from the previous steps:

\[
D_{new} = \{(x, y_{new}) | x \in D_S, y_{new} = \Phi^w(x) + \Phi_2(\Phi^w(x), \Phi_0(x))\}
\]

(16)

**Stage 4:** In the last step, we focus on \(F_1 = \Phi_1 \circ \Phi_0\) again. We fix the parameters of network component \(\Phi_2\) and train \(F_1\) using the new dataset \(D_{new}\) obtained in Stage 3. To avoid introducing feature bias from the previous steps, we clean all previous network weights and re-initialize the whole network before training. The training lasts for \(ep_4\) epochs, and the loss function for this step is \(\mathcal{L} = \mathcal{L}_{KL} + \alpha \mathcal{L}_{cmmd}\), which is the same as the function in Stage 1. Finally, we get the final model \(F_1\) as the desired classifier.

To sum up, in Stage 1, we learn the model \(h\) with the help of the weak annotator to decrease the empirical loss \(\bar{L}_s(h)\), and the CMMD loss will reduce the term \(DD(Q_X^t, Q_X^s)\). Stage 2 uses a new classifier \(d\) to learn the classification distance corresponding to the term \(\bar{L}_t(d)\). The Stage 3 uses the annotator and the learned \(d\) to give more accurate labels than those given solely by the annotator. Then in Stage 4, the model is trained by the relabeled data, making both \(\bar{L}_s(h)\) and \(DD(Q_X^t, Q_X^s)\) be further decreased. The setting of the hyper-parameters used in this section can found in the Supplementary Materials.

5. Experimental Results

The supplementary materials can be found from https://arxiv.org/abs/2102.07358

5.1. Dataset

The experiments are conducted on three application scenarios, the digits recognition with domain discrepancy (SVHN[32], MNIST[6] and USPS[15] digit datasets), object detection with domain discrepancy (VisDA-C[35]), and object detection without domain discrepancy (CIFAR-10[22]). For space, we introduce details of these datasets in the Supplementary Materials.

5.2. Training Setting

All experiments are conducted on a server with Ubuntu 18.04 LTS with NVIDIA TITAN RTX GPU cards. The implementation is based on the Pytorch framework. The hyper-parameter \(\alpha\) mentioned above is set to \(1e-4\). We use the standard Adam optimizer [19] for optimizing the learning. The network architectures, the learning rate for each part of the network components, the training epoch setting, as well as other hyper-parameters are specified in the Supplementary Materials. And we get weak annotators in different performance by applying early stop for the training. The implementation details of weak annotators can also be found in the Supplementary Materials.

5.3. Baseline Experiments Setting

We conduct comparison experiments with the following baselines. Baseline \(B_{wa}\) is the performance of the weak annotator chosen in the experiments in the target domain. Baseline \(B_1\) is training \(F_1\) only with target data. Baseline \(B_f\) is a fine-tuning result. It takes the same model as \(F_1\) and first uses source domain data and weak labels generated by the weak annotator to train it. Then it uses target domain data to fine-tune the last three layers. Baseline \(B_{f3}\) is also a fine-tuning result. The difference is that instead of fine-tuning the last three layers, it trains all network parameters.

As introduced before, our problem is related to the Semi-Supervised Learning (SSL) and the Semi-Supervised Domain Adaptation (SSDA). For SSL, although we can replace the unlabeled data with samples drawn from another domain instead of the target domain, we cannot find a good
way to incorporate the weak annotator into SSL methods for fair comparison with our approach. For SSDA, we were able to extend it to our setting for comparison. Specifically, we add 1,000 unlabeled target samples (plus 1,000 labeled target samples, and this setting will be changed accordingly in digits recognition to keep consistent settings) to meet the semi-supervised requirement, and we apply weak annotator to produce weak labels instead of accurate ones for source data. We compare our approach with the following SSDA baselines: FAN [18], MME [40], ENT [13], S+T [4, 38]. Note that to the best of our knowledge, there is no previous work with exactly the same problem setting as ours. The above changes aim at making the comparison as fair as possible. Another thing that is worth to mention is that most SSDA methods conduct adaptation on the ImageNet pre-trained models, which introduces a lot of irrelevant data information from the ImageNet dataset. Thus, we disable the pre-training and only allow training with the available data.

### 5.4. Results of Digits Recognition

We evaluate our methods on the digit recognition datasets: SVHN (S), MNIST (M), and USPS (U). According to the results shown in Table 1, when the weak annotator performs much worse than the model learned only from the provided target data $B_t$ ($B_{wa} = 73.28\%$ on M $\rightarrow$ U, 73.28\% on S $\rightarrow$ U and 76.41\% on S $\rightarrow$ M), its corresponding baseline $B_{f_1}$ is also lower than $B_t$, and only the second fine-tuning method $B_{f_2}$ is better than or competitive with $B_t$. This indicates that the feature learned from the source domain data and with weak labels introduce data bias, and this bias can be mitigated when the parameters from the front layers are fine-tuned by the target data.

Overall, we can clearly see that with 15,000 source domain data, limited number of labeled target domain data (second line), and a weak annotator, our method can outperform all the baselines in Table 1 with 80.00\% on M $\rightarrow$ S, 95.99\% on M $\rightarrow$ U, 96.36\% on S $\rightarrow$ U and 97.24\% on S $\rightarrow$ M.

### 5.5. Results of Object Recognition

The results of various methods on the VisDA-C dataset are presented in Figure 2. In this task, we utilize the synthetic images as the source domain dataset, and the real-world images as the target domain dataset. And we can see from the table that the performance of the network trained only with the target data is merely 32.86\%. Then, when the weak annotator is provided, it can help two fine-tune baselines $B_{f_1}$ and $B_{f_2}$ reach 27.67\% and 35.03\% respectively. As for the SSDA baselines, all of them perform very badly, and they are provided with more target samples with no labels. The best SSDA methods FAN can only achieve 32.99\%. Our method can provide a result of 40.83\%, which again exceeds all baselines above. Besides, we also provide additional experiment results using a different weak annotator in the supplementary materials.

Moreover, we also test on the scenario without domain discrepancy using the CIFAR-10 dataset. We randomly select 10,000 data samples from the dataset as the source data and another 1,000 samples as the target data. The result is included in Table 2. As we can see, when the weak annotator is given at 48.96\% accuracy, the model trained only with the target data can reach 30.46\%, while our method nearly doubles the performance and hits 61.71\%, which exceeds all other baselines.

### 5.6. Ablation Study

We also study how the quantity of target domain samples and the performance of the weak annotator affect the overall performance of our method. To reduce the impact of domain discrepancy when we study these two factors, we conduct the ablation study on CIFAR-10.

#### 5.6.1 Sample quantity of target samples

As presented in Figure 3, the horizontal axis indicates the number of target domain data, and the vertical axis shows
the performance of our model using the corresponding number of target domain samples. When keeping the weak annotator the same as Section 5.5 and fixing the sample quantity of the source data as 10,000, the precision of the model grows as the number of target domain data increases. And it will gradually get saturated when there is enough target domain data. This saturation phenomenon can be explained as the second derivative of $\sqrt{KL/N}$ and $\sqrt{\ln \frac{2^{N}}{N}}$ for $N$ is positive while the first derivative is negative. And according to the curve, we can observe that the performance improvement when the target data is less than the source data is relatively higher than the case when there is more target data. The reason for this can be found in our theoretical analysis, i.e., when the sample quantities of source and target data become closer, terms impacted by the quantity of target data will not dominate over the error bound.

![Figure 3. The performance of our learned model under different quantities of target domain samples.](image)

### 5.6.2 Performance of weak annotator

Figure 4 shows the curve of how the performance of our model changes with respect to the precision of the weak annotator. As shown in the figure, when the weak annotator performs the worst with accuracy of 23.79%, our model can reach 42.29%, which is a relatively significant improvement. And as the precision of the weak annotator increases, our model performs better accordingly. Interestingly, the improvement curve in Figure 4 is approximately linear, which demonstrates that it is reasonable to linearly add the terms of the weak annotator in the error bound.

![Figure 4. The performance of our learned model under different accuracy of the weak annotator.](image)

### 6. Conclusion

In this work, we present a novel approach leveraging weak annotator to address the data insufficiency challenge in domain adaptation, where only a small amount of data samples is available in the target domain and the data samples in the source domain are unlabeled. Our weak adaptation approach includes a theoretical analysis that derives the error bound of a trained classifier with respect to the data quantity and the performance of the weak annotator, and a multi-stage learning process that improves classifier performance by lowering the error bound. Our approach shows significant improvement over baselines on cases with or without domain discrepancy in various data sets.

### 7. Acknowledgement

We gratefully acknowledge the support from NSF grants 1834701, 1839511, 1724341, 2038853, and ONR grant N00014-19-1-2496.

### References


32] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6