Subtype-aware Unsupervised Domain Adaptation for Medical Diagnosis

Xiaofeng Liu,1,2 Xiongchang Liu,2,3 Bo Hu,2,4 Wenxuan Ji,5 Fangxu Xing,1 Jun Lu*, 2 Jane You,6 C.-C. Jay Kuo,7 Georges El Fakhri,1 Jonghye Woo1

1 Dept. of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
2 Dept. of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
3 China University of Mining and Technology, China
4 Beijing University of Posts and Telecommunications, China
5 School of Artificial Intelligence, Nankai University, China
6 Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong
7 Dept. of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA

Abstract

Recent advances in unsupervised domain adaptation (UDA) show that transferable prototypical learning presents a powerful means for class conditional alignment, which encourages the closeness of cross-domain class centroids. However, the cross-domain inner-class compactness and the underlying fine-grained subtype structure remained largely unexplored. In this work, we propose to adaptively carry out the fine-grained subtype-aware alignment by explicitly enforcing the class-wise separation and subtype-wise compactness with intermediate pseudo labels. Our key insight is that the unlabeled subtypes of a class can be divergent to one another with different conditional and label shifts, while inheriting the local proximity within a subtype. The cases with or without the prior information on subtype numbers are investigated to discover the underlying subtype structure in an online fashion. The proposed subtype-aware dynamic UDA achieves promising results on a medical diagnosis task.

Introduction

The goal of unsupervised domain adaptation (UDA) is to transfer knowledge learned from a label-rich domain to new unlabeled target domains (Saito et al. 2017; Liu et al. 2020c; Liu 2020; Zou et al. 2019). The conventional adversarial training and maximum mean discrepancy (MMD) based methods propose to align the marginal distribution of sample \( x \) in the feature space, i.e., \( p(f(x)) \), where \( f(\cdot) \) is a feature extractor. Given the Bayes’ theorem \( p(f(x)|y) = \frac{p(y|f(x))p(f(x))}{p(y)} \), suppose that there are no concept and label shifts (i.e., \( p(y|f(x)) \) and \( p(y) \) are the same for two domains), then the conditional distribution \( p(f(x)|y) \) can be aligned by aligning \( p(f(x)) \) (Liu et al. 2021a). Nonetheless, the label shift \( p(y) \) is quite common in real-world applications, which indicates the label proportion is different (Zhao et al. 2019).

Recently, transferable prototypical networks (TPN) (Pan et al. 2019) is proposed to promote the source domain class separation with a cross-entropy (CE) loss, and match the class centroids of source and target samples to perform class-wise alignment. However, the CE loss in the source domain cannot minimize the inner-class variation (Liu et al. 2016). The class-wise separation in the to-be tested target domain cannot be well supported by the centroid closeness objective. As shown in Fig. 1 left, although the class centroids are well aligned, the sparsely distributed target samples can be easily misclassified.

One way to tackle this is by simply enforcing the cross-domain inner-class feature distribution compactness. However, in many cases, the unlabeled subtypes in a class can be diverse, and form an underlying local distribution. For instance, different cancer subtypes may have significantly diverse patterns (Yeoh et al. 2002). In such circumstances, the shared pattern among two different subtypes may not be exclusive for class-level discrimination. In these applications, it would be more reasonable and effective to exploit the subtype-wise patterns. Moreover, unsupervised deep clustering (Caron et al. 2018) empirically assigns 10× more clusters of the class to contain diverse subtypes. Recent works also show the fine-grained label can be helpful for the coarse classification (Chen et al. 2019b).

Moreover, domain shifts can be different w.r.t. subtypes, which leads to subtype conditional shift. Besides, the incidence of disease subtypes is usually varied across differ-
ent regions, which leads to subtype label shift. The proportion difference at the subtype-level can usually be more significant than the class-level (Wu et al. 2019). This motivates us to extend the concept of class conditional and label shifts (Kouw 2018) to the fine-grained subtype-level (i.e., \( p(f(x)|k) \) and \( p(k) \) vary across domains for subtype \( k \)). Therefore, a more realistic presumption of UDA can be both the class and subtype conditional and label shifts.

In this work, we resort to the feature space metric learning with intermediate pseudo labels to adaptively achieve both class-wise separation and cross-domain subtype-wise compactness. We first propose an online clustering scheme to explore the underlying subtype structure in an unsupervised fashion. With the prior knowledge of subtype numbers, concise k-means clustering can be simply applied, by assigning \( k \) to the subtype numbers. However, the subtype can be challenging to define, due to different taxonomy. We thereby further expand on our framework to unknown subtype numbers by capturing the underlying subtype structure with an adaptive sub-graph scheme using a reliability-path. With a few meta hyperparameters shared between clusters, the sub-graph scheme can be scalable to several classes and subtypes. To explicitly enforce the subtype-wise distribution compactness, the involved samples of a subtype are expected to be close to their subtype centroid in the feature space.

Our main contribution can be summarized as follows:

- We propose to adaptively explore the subtype-wise conditional and label shifts in UDA without the subtype labels, and explicitly enforce the subtype-aware compactness.
- We systematically investigate the cases with or without the prior information on subtype numbers. Our reliability-path based sub-graph scheme can effectively explore the underlying subtype local distribution with a few meta hyperparameters in an online fashion.
- We empirically validate its effectiveness on a multi-view congenital heart disease (CHD) diagnosis task with an efficient multi-view processing network and achieve promising performance.

**Related Work**

In recent years, big data drives the fast development of deep learning, which has transformed many fields, such as computer vision and medical image analysis (Han et al. 2020; Liu et al. 2019a, 2020f). Deep learning has drastically transformed the way in which features are extracted and then fed into a prediction model into simultaneously learning both features and a prediction model in an end-to-end fashion (Che et al. 2019; Liu et al. 2021c; Xing et al. 2021). The effectiveness of deep learning has been demonstrated in many computer vision tasks, such as classification, detection, and segmentation (Liu et al. 2019f, 2020d, 2018a, 2019c). In addition, to date, conventional machine learning research in medical image analysis has relied on hand-crafted features (Maraci et al. 2017; Liu et al. 2018c, 2017a) and expert decision rules (De Fauw et al. 2018). End-to-end deep learning approaches have also shown promising performance in many disease diagnosis tasks (Liu et al. 2019d, 2018e). For example, (Litjens et al. 2019) attempts to analyze ultrasonic data with deep learning and provides a diagnosis suggestion.

In recent years, big data drives the fast development of deep learning, which has transformed many fields, such as computer vision and medical image analysis (Han et al. 2020; Liu et al. 2019a, 2020f). Deep learning has drastically transformed the way in which features are extracted and then fed into a prediction model into simultaneously learning both features and a prediction model in an end-to-end fashion (Che et al. 2019; Liu et al. 2021c; Xing et al. 2021). The effectiveness of deep learning has been demonstrated in many computer vision tasks, such as classification, detection, and segmentation (Liu et al. 2019f, 2020d, 2018a, 2019c). In addition, to date, conventional machine learning research in medical image analysis has relied on hand-crafted features (Maraci et al. 2017; Liu et al. 2018c, 2017a) and expert decision rules (De Fauw et al. 2018). End-to-end deep learning approaches have also shown promising performance in many disease diagnosis tasks (Liu et al. 2019d, 2018e). For example, (Litjens et al. 2019) attempts to analyze ultrasonic data with deep learning and provides a diagnosis suggestion.
the dynamic memory framework is robust to the pseudo label noise and subtype undersampling.

Methodology

In this work, we consider the UDA task, where we have a source domain \( p^s(x, y) \) and a target domain \( p^t(x, y) \), and our learning framework has access to a labeled source set \( \{(x_i^s, y_i^s)\} \) drawn from \( p^s(x, y) \) and an unlabeled target set \( \{(x_i^t)\} \) drawn from \( p^t(x, y) \). The class label space of \( y_i \in \{1, 2, \ldots, N\} \) is shared for both source and target domains. We use \( n \) to index \( N \) classes. For class \( n \), we assume that there are \( K_n \) underlying subtypes indexed with \( k \in \{1, 2, \ldots, K_n\} \). UDA aims to build a good classifier in the target domain \( p^t(x, y) \) with the following theorem:

**Theorem 1** For a hypothesis \( h \) drawn from \( \mathcal{H} \), \( e^t(h) \leq e^s(h) + \frac{1}{2}d_{\mathcal{H}}(\{s\}) + \min_{h \in \mathcal{H}} e^s(h, l_s) + e^t(h, l_t) \).

Here, \( e^s(h) \) and \( e^t(h) \) denote the expected loss with hypothesis \( h \) in the source and target domains, respectively. Considering that the disagreement between labeling function \( l_s \) and \( l_t \) is small by optimizing \( h \) with the source data (Ben-David et al. 2007), the UDA focuses on minimizing the cross-domain divergence \( d_{\mathcal{H}}(\{s\}) \) in the feature space of \( f(x_i^s) \) and \( f(x_i^t) \).

Instead of aligning \( p^s(f(x)) \) and \( p^t(f(x)) \) (Kouw 2018), the prototypical networks propose to match the class centroids (Pan et al. 2019). However, the decision boundary of the low-density distributed target sample can be difficult to define, and the inherent subtype structure is underexplored.

On the embedding space, we expect the close proximity of the subtype centroid pairs and link the smallest rank first. The source & target center \( c_n^s \) and \( c_n^t \) with \( \mathcal{L}_{\text{class}} = \frac{1}{N} \sum_{n=1}^{N} ||c_n^s - c_n^t||_2^2 \), which is not sensitive to label shift, since it only chooses the representative centroids of the source and target domain placement. We expect the close proximity of \( c_n^s \) and \( c_n^t \) with \( \mathcal{L}_{\text{class}} = \frac{1}{N} \sum_{n=1}^{N} ||c_n^s - c_n^t||_2^2 \), which is not sensitive to label shift, since it only chooses the representative centroids of the source and target domain placement.

The source & target center \( c_n^s \) and \( c_n^t \) with \( \mathcal{L}_{\text{class}} = \frac{1}{N} \sum_{n=1}^{N} ||c_n^s - c_n^t||_2^2 \), which is not sensitive to label shift, since it only chooses the representative centroids of the source and target domain placement. We expect the close proximity of \( c_n^s \) and \( c_n^t \) with \( \mathcal{L}_{\text{class}} = \frac{1}{N} \sum_{n=1}^{N} ||c_n^s - c_n^t||_2^2 \), which is not sensitive to label shift, since it only chooses the representative centroids of the source and target domain placement.

Subtype-aware Alignment with \( K_n \) Prior

If the subtype numbers \( K_n \) of class \( n \) is known (e.g., 4 subtypes in the CHD disease dataset), we can achieve feature space class-independent clustering with the concise \( K \)-means, by defining \( K \) to be \( K_n \).

We denote \( K_n \) clustered subtypes with \( k \in \{1, 2, \ldots, K_n\} \), and calculate the source and target subtype centroids \( \mu_k^s \) and \( \mu_k^t \), respectively. However, \( K \)-means does not assign the specific class label to each cluster. To correlate the source and target clusters, we rank the distance of \( K_n^s \) subtype centroid pairs and link the smallest rank first.

Because of the imbalance distribution of subtypes and possible label shift, we assign the subtype centroids of both the source and target samples with \( \mu_k^s = \frac{\mu_k^s + \mu_k^t}{2} \) instead of averaging all of the source and target samples in subtype \( k \). Therefore, each subtype in both source and target domains contributes equally to \( \mu_k^s \). Then, we enforce all of the samples in subtype \( k \) to be close to the subtype centroid \( \mu_k^s \). For the sake of simplicity, we omit the class notation. The subtype compactness objective \( \mathcal{L}_{\text{sub}}^k \) can be

\[
\frac{1}{M_k^s} \sum_{i=1}^{M_k^s} ||f(x_i^s) - \mu_k^s||_2^2 + \frac{1}{M_k^t} \sum_{i=1}^{M_k^t} ||f(x_i^t) - \mu_k^t||_2^2,
\]

where \( M_k^s \) and \( M_k^t \) are the numbers of source and target samples in subtype \( k \), respectively, to balance the subtype

\[1 \quad \sum_{i=1}^{M_k^s} f(x_i^s) \]

\[2 \quad \sum_{i=1}^{M_k^t} f(x_i^t) \]

\[3 \quad \sum_{i=1}^{M_k^s} f(x_i^s) \]

The source & target center \( c_n^s \) and \( c_n^t \) are not robust to label shift. Note that \( c_n^s \) will change if we simply double the involved source/target samples.
The learned representations are expected to form compact clusters for class $n$, while each cluster does not need to be far away from one another.

**Reliability-path based Sub-graph Construction**

Defining or estimating $K_n$ can be difficult in many applications. Setting $K_n$ of all classes as $N$ hyper-parameters requires costly trials considering the diverse value range of different classes. Note that with a deterministic encoder $f$, the sub-type information can be transferred for different datasets. The class-wise matching and sub-type-wise compactness can be aggregated as a hierarchical alignment loss $\frac{1}{K_n} \sum_{n=1}^{K_n} \sum_{i=1}^{N} (\alpha L^{class}_k + \beta L^{sub}_k)$, where $\alpha$ and $\beta$ are the balancing parameters.

To further eliminate the effect of noise and undersampled subtypes on a batch, we only select the sub-graphs with more than $m$ nodes as the valid subtype clusters.

After exploring the $K_n$ subtypes in the source domain and calculating their centroids $\mu_k$, we assign each target sample with the pseudo label of class $n$ to the subtype with the most similar centroid (i.e., $\min_k \| f(x^t) - \mu_k \|^2$).

Considering the relatively low confidence or reliability of pseudo target labels (Zou et al. 2019; Gu, Sun, and Xu 2020), we adopt a simple online semi-hard mining scheme to select the target sample in a subtype. The cross-domain margin $\tau$ is used to define a circle at the center of $\mu_k$. For the target sample with the initial pseudo subtype label $k$, we choose these samples to distribute within the circle. We note that some target samples may be densely distributed around the circle boundary, and it is not reasonable to cut them apart simply. Therefore, we also resort to the reliability-path to involve the closely distributed neighboring target samples. The sub-graph construction can be robust to missing subtypes in the source or target domains caused by undersampling, since $m$ filters the unreliable source cluster out, and the self-labeling with semi-hard mining of $f(x^t)$ rejects the additional subtypes in the sampled target domain. The operation is illustrated in Fig. 2.

With the reliability-path connected $M^t_k$ source samples and the refined $M^t_k$ target samples in subtype $k$, we calculate $\mu^t_k = \frac{\mu^t_k + \mu^s_k}{2}$, and enforce the subtype-wise compactness with $L^{sub}_k$ as in Eq. (2).

The online sub-graph construction and alignment have three hyperparameters, including $\epsilon$, $\tau$, and $m$ that are shared for all classes and their subtypes, which can be regarded as the meta-knowledge across clusters. Moreover, we can simplify $\epsilon$ to be the constant 1, and change it to any other positive value results only in the matrices being multiplied by corresponding factors (Liu et al. 2017b). The range of $m$ can also be narrow and similar among different subtypes/classes.

**Optimization and Implementation**

The modern neural networks usually extract a high-dimensional vector, e.g., 4,096 or 2,048-dimensional features, as their representation, thereby demanding high memory and time complexity in subsequent clustering. To remedy this, deep clustering (Caron et al. 2018) proposes to perform dimension reduction via Principal Component Analysis (PCA) for the extracted features from all of the samples in a dataset. PCA, however, is not applicable anymore in our online SubUDA. The feature representations are extracted in different training iterations with different time stamps, which can have incompatible statistics. It is also computationally demanding to carry out PCA for all iterations. Accordingly, a non-linear head-layer with the structure of $fc -> bn -> relu -> dropout -> fc -> relu$ is adopted in order to reduce high dimensional features into 256 dimensions. As well, it is simultaneously optimized in each online SubUDA iteration. The non-linear head-layer is eliminated for the subsequent operations, e.g., calculating the L2 distance between features.

In order to prevent the subtype clustering from collapsing to a few subtype groups, (Caron et al. 2018) makes uniform sampling in all of the epochs, which is difficult in our online UDA setting, due to the missing subtype and target class labels. We thus propose a concise approach for SubUDA via re-weighting the loss with $\omega_k \propto \frac{1}{\sqrt{M^t_k + M^s_k}}$, according to the number of samples in the $k$-th subtype. Therefore, samples in smaller clusters are accounted more for the loss, which thereby pushes the classification boundary away to incorporate as more samples as possible. The optimization objective
can be summarized as
\[
\mathcal{L} = \mathcal{L}_{CE}^{class} + \frac{1}{N} \sum_{i=1}^{N} (\alpha \mathcal{L}_{CE}^{class} + \beta \frac{1}{K_n} \sum_{k=1}^{K_n} \mathcal{L}_{sub}^k).
\] (3)

For the classification in testing, we utilize the centroids of training features as prototypical (Pan et al. 2019).

**Experiments**

We carried out experiments using CHD data to evaluate the effectiveness of our approach. We implemented our method and comparison methods using PyTorch and set \(\alpha = 1\), \(\lambda = 0.5\), and \(\beta = 0.5\) consistently.

CHD is one of the most common types of birth defect, which usually results in the death of neonates. Therefore, early medical care and treatment can be helpful, which requires an efficient and accurate diagnosis. In clinical practice, clinicians rely on echocardiograms from five heart views. Collected raw data include echocardiogram videos, and a key-frame of each view is usually extracted for the assessment. Specifically, the collected views are from the parasternal long-axis (PSLAX), parasternal short-axis (PSSAX), apical four chambers (A4C), subxiphoid long-axis (SXLAX), and suprasternal long-axis (SSLAX).

To quantify the effect of subtype structure, four subtypes of CHD, including atrial septal defect (ASD), ventricular septal defect (VSD), patent ductus artery (PDA), and tetralogy of Fallot (TOF), are labeled by two clinicians or intra-operative records. We note that the fine-grained subtype label is not used in training, since large-scale labeling can be costly in clinical practice, while the normal/patient label can be relatively easy to acquire by primary clinicians. Of note, this work focuses on exploring the subtype-aware alignment for the conventional class-wise discriminative model.

Specifically, we evaluated our method on five-view echocardiogram datasets collected from two medical centers. We used 1,608 labeled source subjects (normal/patient) from Beijing Children’s Hospital (BCH) and 800 unlabeled target subjects from Beth Israel Deaconess Medical Center (BIDMC). Each dataset consists of echocardiograms from five views that are sufficient for the diagnosis.

Abnormal regions are likely to be different from one subtype to another; as such, different subtypes can be easily detected from different views. Considering the large inner-class variation of the patient class, it is reasonable to enforce the subtype-wise compactness. In Fig. 4 left, we can see that the source CHD samples of a subtype also tend to be distributed closely, demonstrating the underlying inner-subtype similarity. With our subtype-wise compactness objective as shown in Fig. 4 right, both the source and target samples are grouped into the high-density region w.r.t. subtypes.

**Data Collection of Cardiac Ultrasound Images**

- **BCH source domain.** A total of 1,608 echocardiogram datasets, including 823 healthy controls, 209 VSD, 276 ASD, 124 TOF, and 176 PDA, were collected using PHILIPS iE 33. The chest of each patient was exposed to the echocardiogram with the supine position. We set our transducer frequency between 3 to 8 MHz. We collected the five standard 2D views, i.e., PSLAX, PSSAX, A4C, SXLAX, and SSLAX.

- **BIDMC target domain.** A total of 800 echocardiogram datasets, including 300 healthy controls, 150 VSD, 150 ASD, 100 TOF, and 100 PDA, were collected. Similarly, the chest of each patient was exposed to the echocardiogram with the supine position. We used PHILIPS EPIQ 7C for echocardiogram imaging and set its transducer frequency from 3 to 8 MHz. We also collected the five standard 2D views, i.e., PSLAX, PSSAX, A4C, SXLAX, and SSLAX.

Notably, discrepancies between these two medical centers include imaging devices (PHILIPS iE 33 vs. EPIQ 7C), patient populations, and clinicians’ echocardiogram imaging experience, which introduced domain shifts.

We also extracted the key frame from the video of each view. The heart is a dynamic organ and has different shape within a heart cycle. Thus, we chose a time frame that the clear defects were visible. Specifically, the key frame was corresponding to the isovolumic relaxation phase.

We selected 80% and 20% subjects from the target domain for training and testing, respectively. Note that we only selected the subjects with all the key frames from the five views as our testing data. In addition, the key frames shown in training were not used for testing in a subject-independent manner.
The selected color key frames were first translated to the grey images, since the echocardiogram region has only single-channel 2 gray value information. Then, we cropped the region of interest (ROI) with a redefined sector mask. Since a typical input to convolutional neural networks (CNNs) has the size of $128 \times 128$, we resized the masked ROI to the $128 \times 128$ image. The five views were stacked following a specific sequence (i.e., SLAX, PSSAX, A4C, SXLAX, and SSLAX) to form our five channels training sample.

### Backbone Network Structure for the CHD Diagnosis Task

Training a network with massive parameters using a limited number of training data samples usually results in over-fitting, which causes a problem in medical image analysis. To alleviate the data constraints in our CHD diagnosis task, we propose to adopt the Depthwise Separable Convolution (DSC) (Howard et al. 2017) as our backbone to reduce the to-be-learned parameters. This strategy was introduced in the MobileNet (Howard et al. 2017) with a lightweight implementation. Notably, as a comparison, the DSC based CNNs with 1.32M weights performed similarly to AlexNet (Simonyan and Zisserman 2014) with 60M weights trained using the ImageNet dataset.

More specifically, the convolution operation in conventional CNNs is separated to the depth-wise and point-wise stages. The depth-wise stage processes each channel independently, while conventional CNNs process all channels together. Then, the $1 \times 1$ convolution is used in the subsequent point-wise stage. We adapt the two-stage convolution for our grayscale image analysis.

To fuse the information from the five views, we adopted the multi-channel DSC network. Following the MobileNet (Howard et al. 2017), we configured the first layer as the conventional convolution operation and set the stride as two. We also adopted two fully connected layers with the dimension of 1024 and 128, respectively. The structure of our backbone is shown in Fig. 6, and detailed in Table 1.

In addition to the multi-channel scheme, another feasible choice for multi-view information aggregation is the multi-branch network (Lee, Lee, and Kim 2016). However, our multi-channel framework has a few strengths over the multi-branch network. First, our multi-channel scheme is able to learn multi-view fusing in all of the layers adaptively, rather than simply concatenating each view in the late layer (Lee, Lee, and Kim 2016). Second, our multi-channel framework only uses a single forward model, which has much fewer to-be-learned weights. Considering that the echocardiograms from five views can share some similarity, the convolutional filters trained in each view may potentially be useful for one another. Thus, our framework with less to-be-learned weights can efficiently deal with the problem of small dataset size and the requirement of large memory in the implementation.

The class-level classification in our task aims to differentiate healthy controls from patients, which can be formulated as a binary classification problem. We thereby apply the sigmoid unit as our output layer, and adopt the binary CE loss as the supervision signal. We report the accuracy with the threshold of 0.5.

### Evaluations

For comparison, we re-implemented the current state-of-the-art methods with the same backbone and experiment setting, where we chose the batch size to 64. The results are shown in Table 2. Considering the imbalance of normal and patient proportion in the testing set, we also provide the area under
Table 2: Experimental results for CHD. ↑ larger is better.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%) ↑</th>
<th>AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>76.4±0.12</td>
<td>0.721±0.005</td>
</tr>
<tr>
<td>MCD (Saito et al. 2018)</td>
<td>88.6±0.15</td>
<td>0.856±0.003</td>
</tr>
<tr>
<td>GTA</td>
<td>90.9±0.17</td>
<td>0.873±0.005</td>
</tr>
<tr>
<td>CRST (Zou et al. 2019)</td>
<td>93.2±0.09</td>
<td>0.882±0.006</td>
</tr>
<tr>
<td>TPN (Pan et al. 2019)</td>
<td>93.4±0.14</td>
<td>0.885±0.004</td>
</tr>
<tr>
<td>SubUDA (K_n = 4)</td>
<td>96.2±0.13</td>
<td>0.910±0.003</td>
</tr>
<tr>
<td>SubUDA (K_n = 1)</td>
<td>94.7±0.11</td>
<td>0.902±0.004</td>
</tr>
<tr>
<td>SubUDA-(\omega_k) (K_n = 4)</td>
<td>95.4±0.10</td>
<td>0.903±0.005</td>
</tr>
<tr>
<td>SubUDA-(\omega_k) (K_n = 4)</td>
<td>96.0±0.13</td>
<td>0.908±0.004</td>
</tr>
<tr>
<td>SubUDA-DR (K_n = 4)</td>
<td>96.2±0.11</td>
<td>0.911±0.002</td>
</tr>
<tr>
<td>SubUDA-SG (m = 8)</td>
<td>96.0±0.12</td>
<td>0.907±0.004</td>
</tr>
<tr>
<td>SubUDA-SG-(\tau) (m = 8)</td>
<td>95.5±0.14</td>
<td>0.902±0.003</td>
</tr>
</tbody>
</table>

The receiver operating characteristic curve (AUC) metric in addition to the accuracy metric.

MCD (Saito et al. 2018) and GTA (Sankaranarayanan et al. 2018) are the typical adversarial training frameworks to align the marginal distribution \(p(x)\) at feature level or image level, respectively. The self-training is used to alternatively update the pseudo label of target samples and the network parameters (Zou et al. 2019). We note that the compared methods (Saito et al. 2018; Sankaranarayanan et al. 2018; Zou et al. 2019; Wu, Inkpen, and El-Roby 2020) use the fully-connected classifier after the encoder. The TPN (Pan et al. 2019) uses class centroids as a classifier. It proposes to align class centroids of the source and target sample to achieve the conditional alignment w.r.t. \(p(x|y)\). Our SubUDA outperformed the state-of-the-art methods w.r.t. both the accuracy and AUC by a large margin, by introducing the subtype-aware constraint. The results indicate that the online subtype compactness can effectively help the classification in the target domain.

The domain adaptation theory suggests proxy \(A\)-distance (Ben-David et al. 2007) as a measure of cross-domain discrepancy (Saito, Ushiku, and Harada 2017). In Fig. 7, we compare our SubUDA with the other state-of-the-art methods, and the smaller discrepancy has been observed by using the explicit compactness objective in our SubUDA.

For the ablation study, with \(K_n = 1\), the subtype-wise alignment was reduced to the class-wise compactness. Besides, we used the suffix -DR, -\(\omega_k\), and -\(\tau\) to denote the SubUDA without dimension reduction head, subtype balance weight, and semi-hard target mining, respectively. Furthermore, the suffix -\(\omega_k\) denotes using \(\mu_{k}^{n} = \sum \omega_{k}^{n} f(x_{i}^{n}) / \sum \omega_{k}^{n} \) as the subtype centroid, which is not robust to the subtype label shift. SubUDA-DR took \(4 \times \) clustering time, but the improvement was marginal. Therefore, we recommend using the dimension reduction head.

SubUDA-SG in Table 2 used the reliability-path based online sub-graph to replace \(k\)-means. With appropriate \(m = 8\), the adaptively learned clustering achieved comparable performance to the \(k\)-means with \(K_n = 4\). In Fig. 7 right, we can see that the semi-hard mining scheme in SubUDA-SG is not sensitive to \(\tau\) for a large range, since the network can flexibly learn to adjust the ratio of \(\epsilon\) and \(\tau\) in mapping space (Liu et al. 2017b). We note that too strict semi-hard mining (i.e., too small \(\tau\)) can degenerate our SubUDA to conventional class centroids matching, since no target samples are selected to form a subtype cluster.

In Fig. 5, we provide a sensitive analysis of hyperparameter \(K_n\) and \(m\) for two kinds of online clustering schemes. Consensus clustering can assess the clustering stability (Monti et al. 2003), and the optimal clustering is usually achieved in the elbow position of the area under CDF changes, where the CDF is for the consensus matrices\(^2\). We can see that the peak of accuracy usually coincides with the best clustering consensus metric, which indicates the good subtype clustering can boost the SubUDA performance. The choice of \(K_n = 4\) also matches our prior knowledge of the CHD patient subtypes.

Since we have four clear subtypes in this task, using the concise \(k\)-means can be a straightforward solution. However, Fig. 5 right shows that the accuracy curve can be robust for a relatively large range of \(m\), which is promising for the hyperparameter tuning of the case without the prior information of subtype numbers.

### Conclusion

In this work, we presented a new UDA approach with a more realistic assumption that the subtype-wise conditional and label shifts widely exist, and can be adaptively aligned without the subtype label. We systematically investigate a flexible yet principled solution for the case with and without the prior knowledge of subtype numbers. Rather than the concise \(k\)-means, we further extend our framework with an online sub-graph scheme using the reliability-path, which can be scalable to many classes and subtypes with a few meta hyperparameters. The effectiveness of explicitly enforcing the subtype-aware compactness has been successfully demonstrated in the CHD transfer task.

### Acknowledgments

We thank the anonymous reviewers for their constructive comments.

\(^2\)https://github.com/ZigaSajovic/Consensus_Clustering
References


Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In ICML.


Liu, X.; Che, T.; Lu, Y.; and Yang, C. 2020a. AUTO3D: Novel view synthesis through unsupervisedly learned variational viewpoint and global 3D representation. ECCV.


Saito, K.; Ushiku, Y.; and Harada, T. 2017. Asymmetric tri-training for unsupervised domain adaptation. In ICML.


