Edge Domain Adaptation through Stepwise Cross-Domain Distillation

Taisei Yamana[1,a)] Yuko Hara-Azumi[1,b)]

Abstract: Machine learning is now required to be built on embedded systems to realize edge-AI devices, where not only weight reduction but also accuracy degradation that stems from domain shift need to be addressed. This paper proposes Stepwise Cross-Domain Distillation (SCDD) that employs unsupervised domain adaptation for lightweight models. By distilling knowledge from a pre-domain-adapted large model stepwisely through a teaching assistant model, the final lightweight student model can effectively achieve good accuracy in a target domain. We also provide insights obtained through quantitative evaluations to improve stepwise knowledge distillation in various domain shifts.Code is available at https://github.com/TaiseiYamana/SCDD.git

Keywords: Domain adaptation, knowledge distillation, deep learning, edge AI

1. Introduction

Machine learning traditionally has a trend of utilizing wide and deep network structures trained by large scale dataset and requires many computational resources. Along with the advance in Internet of Things and artificial intelligence (AI) technologies, another trend that increases the demands of edge AI (i.e., deploying machine learning models on edge devices like mobile devices) is arising. One of difficulties in edge AI stems from *domain shift* [14], which is a gap between training data distribution and environments where edge devices are deployed, leading to accuracy drop in edge AI. Additionally, as edge devices generally have severe constraints of computational resources, it is required to compress the machine learning models into computationally efficient and lightweight models, which may further reduce accuracy. An effective solution to mitigate accuracy degradation in edge AI is to create dataset for individual target environment by labeling thousands of data manually and utilize it in training. However, this approach is obviously inefficient and impractical.

The aforementioned problem was first defined as edge domain adaptation (EDA) [17], where model weight reduction and adaptation should be dealt with simultaneously. Then, a framework called MobileDA was proposed to resolve EDA. With cross-domain distillation which combines unsupervised domain adaptation and knowledge distillation, MobileDA enabled to obtain a lightweight model for the target environment. However, there is still room for improvement in MobileDA, mainly in terms of three issues as follows; (1) a teacher model is not trained for the target domain and good knowledge is not distilled to a student model, (2) the UDA method in MobileDA does not sufficiently improve accuracy of student model, and (3) the size gap between the teacher and student models, which affects knowledge distillation, is not taken into account.

In this paper, we propose a novel stepwise cross-domain distillation (SCDD) method that tackles with the above three issues in MobileDA. Specifically, we pre-train the teacher model in the target domain to improve distilled knowledge (solution to the issue (1)). Also, student model's accuracy in the target domain is improved by introducing Minimum Class Confusion (MCC) [7] in cross-domain distillation (solution to the issue (2)). Finally, an intermediate-sized Teacher Assistant (TA) [13] model is inserted in the knowledge distillation path between teacher and student models so that stepwise distillation is conducted to reduce the size gap (solution to the issue (3)).

The contributions of this paper are summarized as follows:

- We propose a novel knowledge distillation method to distill knowledge from domain-adapted teacher models to lightweight student models.
- We improve cross-domain distillation by introducing MCC to increase the accuracy of the student model in the target domain compared with the original MobileDA.
- We propose stepwise cross-domain distillation (SCDD) that makes training for lightweight student models under the target domain effective by accounting for the size gap between teacher and student models.

The remainder of this paper is organized as follows: Section 2 briefly reviews knowledge distillation, unsupervised domain adaptation, and MobileDA. Section 3 presents our proposed stepwise cross-domain distillation. Section 4 describes our experimental setup and results. Section 5 concludes this paper.

2. Preliminaries

This section briefly reviews baseline technologies that will be applied to our proposed method and understand the problems in MobileDA to be tackled.

[1] Tokyo Institute of Technology, Meguro, Tokyo 152–8552, Japan
[a)] yamana.t.ab@m.titech.ac.jp
[b)] hara@cad.ict.e.titech.ac.jp

2.1 Knowledge Distillation (KD)

Conventionally, machine learning has achieved high inference accuracy for a target classification task by using large models. However, it is difficult to deploy the functions of these machine learning models on edge devices because of the large models. Various research has been conducted to reduce the weight by model compression [3], [6], [5].

Knowledge Distillation (KD) is one of the most popular model compression methods [5]. KD compresses knowledge from large models into a single and lightweight model. In general, the large model is called teacher, and the lightweight model is called student. The student model can be trained to mimic the output of the high-performance teacher model in order to achieve better accuracy than the normal training process. KD using the outputs of the models scaled by temperature, called soft label. The soft label makes the distribution soft, so student can efficiently learn the teacher's knowledge. As the accuracy of the teacher model increases, the accuracy of the student model also improves.

2.2 Unsupervised Domain Adaptation (UDA)

In practical cases, using different datasets for training and testing happens frequently. Hereafter, we refer to the datasets used for the former and latter as source domain and target domain, respectively. In such cases, these two domains may have significant property differences called "domain shift." Domain shift is a common problem in image classification tasks that can result in poor inference accuracy. Fig. 1 shows some examples of the Office-31 dataset [14] that represents domain shift. While each domain shares the same class objects, we can visually see different input properties such as angles and luminance.

A conventional approach to resolve domain shift is to utilize transfer learning with data labeled in the target environment. However, this imposes a laborious effort for developers to label a huge amount of data in each target domain. Therefore, unsupervised domain adaptation (UDA) has been studied as an alternative approach that uses a dataset from the source domain and unlabeled images from the target domain to adapt to the target. Thanks for the convenience in dataset preparation, UDA has been focused and studied in various ways (e.g., [1], [7], [10], [11], [12], [16]). They commonly adapt to the target domain using a loss that equalizes the output of the model in the source domain and target domain. Distilling the Knowledge in a Neural Network (DANN) [1] uses a domain discriminator to make the two domains indistinguishable. Deep-CORAL [16] minimizes the distance between the co-variant matrices of the model outputs for the two domains. Minimum Class Confusion (MCC) [7] minimizes class confusion of the target domain to improve accuracy. Since new methods are often proposed that are better than the adversary methods, there is a need to improve the approach to applying a particular DA method accordingly.

2.3 MobileDA

MobileDA [17] which is cross- domain distillation combining KD and UDA to solve the EDA problem was proposed by Yang et al. Cross- domain distillation performs UDA on lightweight models while performing KD using inferred labels in the target domain of a pre-trained large model. In MobileDA, Deep-CORAL was adopted into UDA method to avoid computation overloads. MoibleDA has been shown that computationally efficient and lightweight models achieve better results than advanced UDA methods [1], [7], [10], [11], [12], [16].

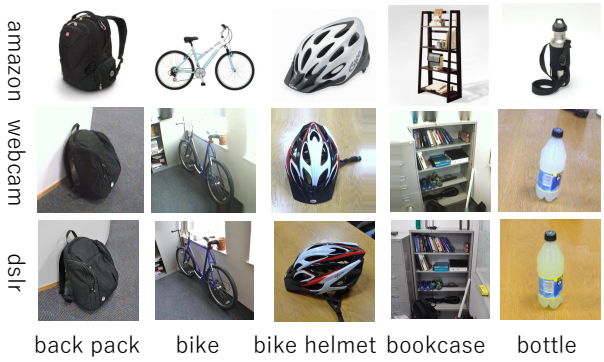

Fig. 1: Office-31 dataset samples for five classes from three domains

MobileDA still have room for improvement in three-fold:(1) The teacher model is pre-trained only in the source domain and is not adapted to the target domain. If the teacher model does not perform well in the target domain, the distilled knowledge may not be suitable for the target domain, which would degrade the student model's performance in the target domain. (2) Necessity of reconsidering the UDA method. There is a possibility that the accuracy of the student model can be improved by introducing a UDA method that is more improving accuracy than Deep-CORAL and less computationally expensive. (3) The size gap between the teacher and student models is not taken into account. In general, a large model is adopted for the teacher for higher accuracy, while a lightweight model is adopted for the intended student. Therefore, the size gap between models is a problem. If this gap is large, KD may not be performed effectively.

2.4 Teacher Assistant Knowledge Distillation (TAKD)

One drawback in KD is that distillation gets ineffective when the gap between the student and teacher scales is larg. The reason is mainly two-fold: (1) Even if the teacher is a rich model,the student does not have the sufficient capacity to well mimic the teacher's behavior. And, (2) the softness of the soft target is reduced when the certainty of the inference in the teacher model to the input is increased. In other words, it weakens knowledge transfer as it undermines the information that the soft target has about similar classes.

To overcome the aforementioned gap issue, a distillation framework called Teacher Assistant knowledge Distillation (TAKD) [13] was proposed. TAKD fills the gap in scale by introducing a Teacher Assistant (TA) as an intermediate model in the distillation process between the teacher and student models. The TA model should be smaller than the teacher and larger than the student. The distillation is done in a stepwise manner such that the teacher distills the knowledge to the TA, and then the TA distills the knowledge to the students. This strategy mitigates the gap issue and increases the effectiveness of knowledge transfer, thereby improving the accuracy of the student model over the

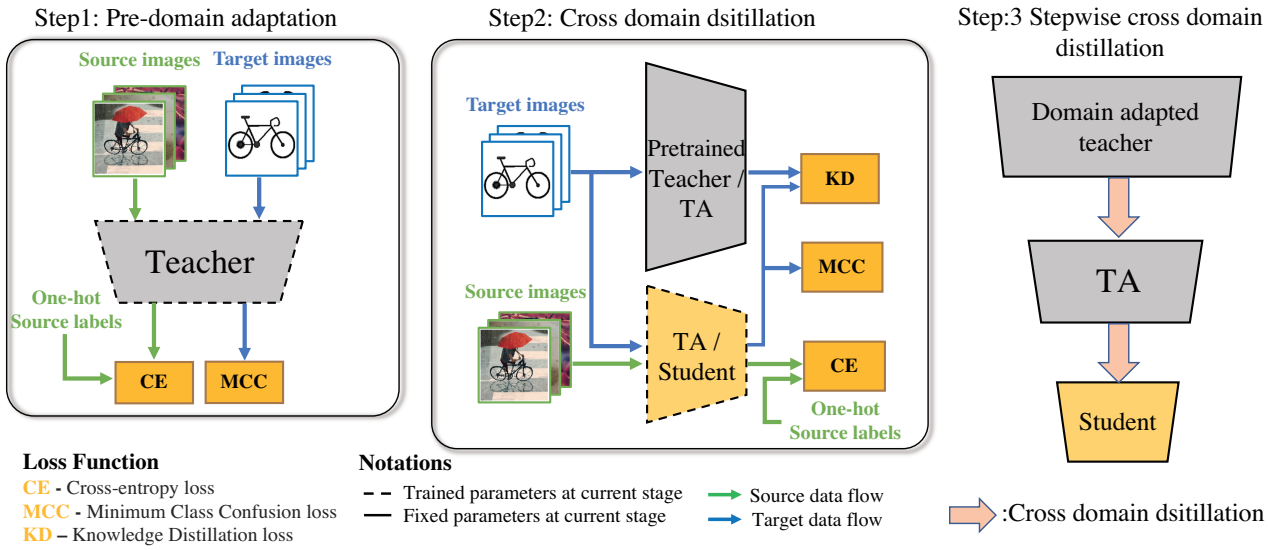

Fig. 2: Stepwise Cross-Domain Distillation

traditional KD. In addition, the use of multiple TAs between the teacher and student models is effective for the segmentation of the scale gap.

3. Stepwise cross-domain distillation (SCDD)

In this work, in order to resolve the aforementioned issues that reside in MobileDA, we propose a novel cross-domain distillation (SCDD) method that can stepwisely distill knowledge from the teacher to student models. First, to improve the accuracy of the teacher model under the target domain, domain adaptation is introduced to pre-train the teacher model (described in Section 3.2). Next, to improve the effect of UDA in cross-domain distillation, MCC is applied to cross-domain distillation (described in Section 3.3). Finally, to reduce the size gap between teachers and students, stepwise knowledge distillation using TA is introduced to cross-domain distillation (described in Section 3.4).

Here we describe our target problem and assumptions and explain how each technique contributes to achieve our goal.

3.1 Problem Definition and Notation

We consider data from source and target domains in unsupervised domain adaptation. The source data indicate accessible dataset with labels, denoted as $\mathcal{D}_s = \{x_s^i, y_s^i\}_{i=1}^{N_s}$. The target data can get from new environment of edge device and are only captured without labels, denoted as $\mathcal{D}_t = \{x_t^i\}_{i=1}^{N_t}$. N_s, N_t are the number of data in each domain. Consider that the source and target domains share the same task. However if there is domain shift between these two domains, mismatch between the input's marginal distributions of the two domains occurs. Under such a condition, even if the model trained with $\mathcal{D}_s$ predicts the target label $\{\hat{y}_t^i\}_{i=1}^{N_t}$, accuracy may be significantly degraded.

We assume that a lightweight model implemented on an edge device tries to predict the label $\{\hat{y}_t^i\}_{i=1}^{N_t}$ of the target domain in an environment with domain shift. Although UDA can be used to improve the accuracy of the lightweight models, it is difficult to obtain sufficient improvement due to the low amount of parameters in the model. Therefore, in cross-domain distillation, a large-scaled, high-performance teacher model Φ^T is used to train the target lightweight student model Φ^S. The teacher model Φ^T is trained in advance.

In our training method, we use the probability vector of a model Φ^f for a sample x, rescaled by temperature τ, where f takes either student or teacher. Let the model's logit be $\Phi^f(x) = z^f = \{z_1^f, \ldots, z_{|C|}^f\}$, where $|C|$ indicates number of classes. The k-th rescaled probability vector p^f is formulated as follow:

$$p_k^f(\tau, x) = \frac{\exp(z_k^f/\tau)}{\sum_{j=1}^{|C|} \exp(z_j^f/\tau)}, \tag{1}$$

when $\tau = 1$, we can use a normal probability vector. In KD, rescaled vector by temperature is called soft label.

3.2 Pre-domain adaptation

In MobileDA, a pre-trained teacher model is used to train a student model under the target domain. Because MobileDA's teacher model is pre-trained with $\mathcal{D}_s$ only, the teacher's accuracy in the target domain may not be good. This can be one factor that reduces the student's accuracy in the target domain as the teacher's poor knowledge for the target domain would be distilled to the student. From this perspective, our proposed method employs UDA in pre-training of the teacher model Φ^T for adapting it to the target domain. The optimization problem of training the teacher model Φ^T is formulated as follow:

$$\begin{aligned} \min_{\Phi^T} L(x_s, y_s, x_t; \Phi^T) &= L_{CE}(p^T(1, x_s), y_s; \Phi^T) \\ &\quad + \mu L_{MCC}(x_t; \Phi^T), \end{aligned} \tag{2}$$

$$L_{CE}(p^f(1, x_s), y_s; \Phi^f) = \sum_{j=1}^{|C|} -y_j \log p_j^f(1, x_s) \tag{3}$$

where μ is a trade-off for DA, and the classification loss L_{CE} uses a general cross-entropy loss.

Minimum class confusion (MCC) loss: From the perspective of knowledge distillation, we aim to improve the student's accuracy by increasing the teacher's accuracy under the target domain

in advance. For this purpose, it is essential to adopt a DA method with a high adaptive effect. In order to reduce the learning computation cost, a method with low computational load is preferred. Consequently, we adopt an MCC loss based on minimization of class confusion. Class confusion happens when the probability value of a model's output becomes ambiguous between similar classes. Because class confusion become large when domain shift causes accuracy degradation, domain adaptation is performed by quantifying the class confusion and minimize it. MCC has not only an adaptive capability to achieve the top accuracy among various DA methods including Deep CORAL, which was utilized in the original MobileDA, but also high convergence speed. In MCC loss, class confusion is quantified from the model's logit for a sample of the target domain. MCC loss is defined as follows:

$$L_{MCC}(x_t;\Phi^f) = \frac{1}{|C|}\sum_{j=1}^{|C|}\sum_{j'\neq j}^{|C|}|\tilde{C}_{jj'}|, \tag{4}$$

$$\tilde{C}_{jj'} = \frac{C_{jj'}}{\sum_{j''=1}^{|C|} C_{jj''}} \tag{5}$$

where C is quantified alignment of class confusion and normalize to $\tilde{C}$. Here we briefly explain the calculation of $\tilde{C}$. As mentioned above, we use the model's logits for the sample of the target domain to calculate C. According to [2], as deep neural network tends to make overconfident predictions, we rescale the probability vector by temperature τ_{mcc}. Next, for more accurate quantification, class confusion is highlighted by uncertainty weighting. Uncertainty weights W can be formulated as follows:

$$H(p^f(\tau_{mcc},x_t)) = -\sum_{j=1}^{|C|} p_j^f(\tau_{mcc},x_t)\log p_j^f(\tau_{mcc},x_t), \tag{6}$$

$$W = \frac{B(1+\exp(-H(p^f(\tau_{mcc},x_t))))}{\sum^{B}(1+\exp(-H(p^f(\tau_{mcc},x_t))))}, \tag{7}$$

where H is entropy function, B is batch size. After entropy is adjusted by laplace smoothing [15], it is normalize in the batch. Finally, using the uncertainty weights and the scaled probability vector, C is calculated by the following equation:

$$C = p^f(\tau_{mcc},x_t)^\intercal W p^f(\tau_{mcc},x_t). \tag{8}$$

3.3 Cross-domain distillation with adapted teacher

In our method, we propose a new cross-domain distillation based on MobileDA. The our method uses MCC, which was also used in the pre-domain adaptation of the teacher model, to adapt the student model to the target domain. This is the same reason of pre-domain adaptation, because MCC has higher improvement accuracy than Deep-CORAL in MobileDA and low computational load. Soft label is calculated from logits of pre-domain adapted teacher model when samples of the target domain is input to teacher model and use as ground true labels for distillation Loss. The optimization problem of training of student model Φ^S is formulated as follows:

$$\begin{aligned}\min_{\Phi^S} L(x_s,y_s,x_t;\Phi^S,\Phi^T) &= L_{CE}(p^S(1,x_s),y_s;\Phi^S)\\ &+ \mu L_{MCC}(x_t;\Phi^S)\\ &+ \lambda L_{KD}(x_t;\Phi^S,\Phi^T)\end{aligned} \tag{9}$$

Algorithm 1 Cross-domain distillation with adapted teacher

Pretrain: MCC train the Teacher model Φ^T in Eq.2
1: **for** each epoch **do**
2: Obtain the high-confidence samples x_{hc} from x_t
3: Calculate the CE Loss L_{CE} in Eq.3
4: Calculate the MCC Loss L_{MCC} in Eq.4
5: Calculate the KD loss L_{KD} in Eq.10
6: Optimize the total loss in Eq.9
7: **end for**
Output: The student model Φ^S in the IoT device

$$L_{KD}(x_t;\Phi^S,\Phi^T) = \tau_{kd}^2\sum_{j=1}^{|C|} p_j^T(\tau_{kd},x_t)\log\frac{p_j^T(\tau_{kd},x_t)}{p_j^S(\tau_{kd},x_t)} \tag{10}$$

where μ and λ are coefficients of the trade-off between the distillation and domain adaptation losses, respectively. Kullback-Leibler (KL) divergence loss [8] is used as L_{KD}. Motivated by self-training, we pre-select samples from the target domain where the predicted category confidence score is higher than a threshold γ similarly to MobileDA.

Alg. 1 describes the cross-domain distillation with adapted teacher. The teacher model is adapted to target domain with MCC before training student model (line 1). The student is trained by using the trained teacher model (lines 2 to 7). First, the data x_t from the target domain is input to the trained teacher model, and the input data whose probability outputs above the threshold is selected as x_{hc}. Next, Loss (L_{CE}, L_{MCC}, L_{KD}) are calculated from x_{hc} and source domain dataset D_s. Next, the student model is tuned in Eq.9. Repeat this process for any epoch. Finally, the trained student model is deployed to the target IoT device (line 8).

3.4 Stepwise knowledge distillation

TAKD [13] focuses on the impact of the size gap between the teacher and student models and effectively implements knowledge distillation into the student model via an intermediate model. To address negative factors introduced by the scale gap as described in Section 2.4, we present a stepwise distillation using Teacher Assistant (TA) into Alg. 1.

As shown in the right side of Fig. 1, the cross-domain distillation of our method is conducted from the teacher to the TA, and then to the student. Here the TA adopts a model that is medium in size between the teacher and student models. By this means, good knowledge on the target domain in the teacher model is transferred to the student stepwisely. Finally, the trained, lightweight student model can be deployed on the target IoT device.

4. Experimental Results

In this section, we first describe our evaluation dataset and setup, and then discuss the effectiveness of our proposed method over existing works.

4.1 Dataset

We validate our proposed method on two domain datasets, Office-31 and Office-Home.

Office-31 is the most commonly used dataset for domain adaptation of image classification: it contains 31 categories and con-

Table 1: Distillation paths (⇒ indicates cross-domain distillation).

Method	Distillation
TA	ResNet50⇒ResNet34
Student w/o TA	ResNet50⇒ResNet18
Student w/ TA (Ours)	ResNet50⇒ResNet34⇒ResNet18

Table 2: Number of parameters and MACs in each model.

Model	Params (M)	MACs (M)
ResNet50	23.57	4109.53
ResNet34	21.30	3670.77
ResNet18	11.19	1818.57
AlexNet	57.13	710.72

sists of 2817 images from Amazon (A), 795 images from webcam (W), and 498 images from DSRL (D). The W and D domains are very similar, but A contains most images in three domains. The dataset size of Office-31 and and the number of images in the three domains are disproportionate. We consider all cross-domains (i.e., A→W, D→W, W→D, A→D, D→A, and W→A).

Office-Home is also a domain dataset for image classification. It contains 65 categories and consists of 2,427 artistic images (Ar), 4,365 clip art (Cl), 4,439 product images (pr), and 4,357 real-world images (Rw) in four different domains. Here we also consider all cross-domain (i.e., Ar→Cl, Ar→Pr, Ar→Rw, Cl→Ar, Cl→Pr, Cl→Rw, Pr→Ar, Pr→Cl, Pr→Rw, Rw→Ar, Rw→Cl, and Rw→Pr). The adaptation on Office-Home is more difficult than Office-31 because Office-Home has 65 categories and enormous domain shifts.

4.2 Setup

In our evaluation, we conducted two experiments to evaluate the effectiveness of pre-domain adaptation and SCDD. In both experiments, we train the target student model on a domain dataset and evaluate accuracy for the target domain data. The unlabeld target domain data is used for test but only 80 % of them is used for training to make a fair experimental condition with MobileDA.

First, in the evaluation of pre-domain adaptation, we compared three methods – MobileDA, MobileDA with a pre-domain adapted teacher model, and our proposed method. Here Office-31 is set as the dataset. To compare with previous work, we selected ResNet34 [4] as the teacher model and AlexNet [9] as the student model. Next, in the SCDD evaluation, we adopted the ResNet series [4] for all models due to their similar network structures. We selected ResNet50 as teacher, ResNet34 as TA and ResNet18 as student to conduct our proposed cross-domain distillation. We evaluated the accuracy effect on the student model in the target domain by TA. Several distillation paths were evaluated as shown in Table 1.

We implemented SCDD based on pytorch and employed the ImageNet pre-trained parameters as the model initialization. The whole model was trained with the learning rate of η for the classifier and $\frac{\eta}{10}$ for the encoder. The initial learning rate η_0 was set to 0.001 and updated with $\eta_{n+1} = \eta_n/(1.001 \times n)^{0.9}$, where n is iteration. We adopted mini-batch stochastic gradient descent (SGD) with momentum of 0.9 and weight decay of 0.001. We also set the losses as $\lambda = 1$ and $\mu = 1$, the temperature $\tau_{mcc} = 2.5$, and the threshold $\gamma = 0.7$. In pre-domain adaptation experiments, we set $\tau_{kd} = 4$, and in SCDD experiments, we set $\tau_{kd} = 8$. These parameters are adopted empirically. The parameter sizes and MACs in each model are shown in Table 2. It is clear that the computational efficiency increases from the teacher model to the student model.

Table 3: Accuracy comparison of teacher models (ResNet34) on Office-31.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
Source Only	71.82	97.86	99.60	76.10	60.45	60.70	77.76
Deep-CORAL	77.74	96.60	99.80	76.10	61.70	61.16	78.85
MCC	**87.15**	**98.49**	**99.80**	**87.22**	**70.61**	**72.56**	**85.97**

4.3 Results

As shown in Table 3, the accuracy of the teacher model in the target domain is improved by using the domain adaptation method compared with the source only. The MCC we employed improves the accuracy over Deep-CORAL used in MobileDA. As shown in Table 4, where the pre-domain adapted teacher model is used for cross-domain distillation, the average accuracy of the student model is improved by 0.13% in the same MobileDA. As the accuracy of the teacher model increases, the correctness of the soft labels on the teacher's side used in the distillation loss increases, thus training of the students model is improved. Moreover, our proposed method with MCC-adapted teacher improves the accuracy by 7.84 % from MobileDA with Source only.

In terms of teacher pre-training, MCC is better than Deep-CORAL because it can be trained with lower batch size and higher accuracy than Deep-CORAL. To confirm the effectiveness of the cross-domain distillation employing MCC, we compared it to MobileDA under the condition that trained model with Source only is set as teacher. We can see that cross-domain distillation employing MCC is 0.33 % higher accuracy of student model than MobileDA employing Deep-CORAL one.

Next, as shown in Table 5, in the experiment of SCDD on the OfficeHome dataset, adapting TA to the proposed cross-domain distillation improves the accuracy of the student model under particular domain shifts. For example, in Rw→Cl, the TA improves the accuracy of the student model by 0.83%. However, there are another particular domain shifts where TA produces poor training results. In Rw→Cl, there is a 0.87% decrease in accuracy even though TA is used. Comparing the accuracy of the teacher to TA, accuracy of TA is better than teacher even though the TA size is smaller than Teacher. However, accuracy of Student distilled form TA may became deteriorate even when TA have better accuracy than teachers. We need to find out the cause of this phenomenon in order to improve SCDD.

4.4 Analysis and ideas for improvement

In this sub section, we will discuss the trends and causes of each domain shifts while analyzing the experimental results of SCDD. From these considerations, we will mention suggestions for improvement of the SCDD.

First, we investigated the structural compatibility of the ResNet34 and ResNet18 models, as shown in Figure.5. When ResNet34 is set as the TA, the change in accuracy of ResNet18 from ResNet34 is decreased under most domain shifts. However,

Table 4: Accuracy comparisons of student models (AlexNet) between MobileDA and CDD by MCC on Office-31.

Method	Teacher's method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
Source Only		49.69	93.46	95.78	47.79	32.69	33.01	58.74
MobileDA	Source only	76.86	97.48	99.80	76.31	61.34	60.56	78.73
	Deep-CORAL	**78.11**	96.73	99.60	**78.71**	60.21	60.03	**78.90**
CDD by MCC	Source only	76.98	97.48	99.60	77.51	61.77	60.99	79.06
	MCC	**91.45**	**98.74**	**99.80**	**87.15**	**69.58**	**72.70**	**86.57**

when ResNet34 is set as the teacher, the change in accuracy of ResNet18 from ResNet34 approximated to change in accuracy when ResNet18 is trained directly using ResNet50 as a teacher. Therefore, the poor results of cross-domain distillation from the TA to the students indicate that there is no structural causality between the ResNet34 and ResNet18. Therefore, the reason for the poor results of SCDD is not the structural relationship between the models set up, but the difference in the output data of TA and teacher models.

Next, to get a trend of the SCDD results, the accuracy improvement by TA with descending order is shown in Table 6. The number of soft labels of TA, selected by the threshold is less than teacher's soft label under most domain shifts. This indicates that there is a difference in the outputs of the teacher and TA. There are three possible factor where SCDD can improve accuracy; (1) where there is a sufficient number of target samples available, (2) where the number of soft labels of the TA selected by the threshold is not much less than the number of soft labels of the teacher, and (3) where the TA's accuracy is sufficiently high. The reason for (1) is the more data of the target domain is available, the more generic the model becomes to unknown samples in the target domain. Therefore, The reason for (2) is the effect of TA distillation is inferior to that of teacher because it impairs (1). Finally, the reason for (3) is that when the accuracy of TA is high, the number of correct soft labels increases. This shows an improvement in the quality of training of student models.

For example, in Rw→Cl, the accuracy of TA is low, but the accuracy is improved because the number of data in the target domain is large enough, and the number of soft labels of TA is larger than that of the teacher. In the case of Rw→Pr, even though the number of soft labels of TA is much smaller than that of the teacher, the number of data in the target domain is large and the accuracy of TA is high. Those factors indicates that accuracy of student is improved. When Ar is specified as the target domain, the number of datasets in Ar is too smalle. So the cross-domain distillation from any source domain to Ar will result in a decrease in accuracy.

Considerations for the number of soft labels selected: The reason why the number of soft labels selected for TA decreases from the teacher in most domain shifts is that the TA's prediction became to a uniform distribution than the teacher and the peak against predict category is weak. Therefore, even if the accuracy of TA is improved, the number of labels that exceed the threshold are reduced. Observe the MCC loss of the TA and the teacher as shown in Table 6. The MCC loss quantifies the class confusion in the model. So the larger MCC loss, the more ambiguous the predictions are among similar classes. This indicates that the predictions are uniformly distribute and having weak peaks. In the only case where the number of soft labels increased, Rw→Cl, the MCC Loss of TA is less than that of the teacher. Therefore, the accuracy of the students is improved due to the increased soft labels of the TA In the domain shifts where the number of soft labels decreased, the MCC loss increased. It is clear that the prediction of TA is closer to uniform distribution than that of the teacher, and the decrease in the number of labels occurs even though the accuracy is improved. Therefore, the improvement in accuracy by TA may not be achievable due to the loss of the factor (2) for improvement.

Discussion for improvement: Expecting that increasing the number of soft labels on TAs will improve SCDD, we discuss two possible approaches for future improvement. The first one is to increase the trade-off ratio of the MCC loss. By increasing the ratio, we can induce the MCC loss to be smaller in training than usual. However, it is necessary to check the impact of lowering the rate of other loss trade-offs, and investigation for empirical knowledge should be conducted. The second point is the adjustment of the threshold.When student model is trained with TAs, threshold is set lower than when training TAs with teachers, in order to gain more soft labels. However, there is a possibility that the number of wrong soft labels will increase. So effect of wrong soft labels needs to be investigated as well.

5. Conclusions

In this paper, we proposed SCDD to solve the problems of MobileDA. Our method stepwisely distills the knowledge of a teacher model and excellent student models in targeted environments can be trained. The approach using a pre-domain adapted teacher is better than the approach without pre-domain adaptation. The stepwise distillation approach further improved the accuracy of the student models in some domains. An investigation of the SCDD results for each domain showed that improving the number of soft labels could improve the accuracy of student in the future. Future work will improving the iterative distillation approach to extend the current methodology.

Acknowledgment

This work is partially supported by JSPS KAKENHI JP20H04154.

References

[1] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V.: Domain-adversarial training of neural networks, *The journal of machine learning research*, Vol. 17, No. 1, pp. 2096–2030 (2016).

[2] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q.: On calibration of modern neural networks, *International Conference on Machine Learn-*

Table 5: Accuracy (mean) of SCDD on Office-Home datasets.

Method (Model)	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
Teacher (ResNet50)	52.21	77.95	80.79	66.01	75.04	76.68	62.22	51.80	80.45	72.19	56.27	82.90	69.54
TA (ResNet34)	53.54	78.26	80.51	66.50	75.24	76.75	63.00	52.55	80.22	72.27	56.72	83.51	69.92
Student (ResNet18) w/o TA	52.92	77.70	80.08	65.35	75.06	76.08	61.10	51.98	79.71	71.73	56.40	83.28	69.28
Student (Resnet18) w/ TA (Ours)	52.78	**78.26**	**80.12**	64.48	**75.11**	75.74	60.86	51.68	**79.96**	71.12	**57.23**	**83.51**	69.24
Improvement by TA	-0.14	**0.44**	**0.04**	-0.87	**0.05**	-0.34	-0.24	-0.30	**0.25**	-0.61	**0.81**	**0.23**	

Table 6: Accuracy comparisons in SCDD experiment.

Domain shift	Teacher ResNet50	Student ResNet18	Differential accuracy	Teacher ResNet34	Student ResNet18	Differential accuracy	TA ResNet34	Student ResNet18	Differential accuracy
Ar→Cl	52.21	52.92	**0.71**	52.42	52.94	**0.52**	53.54	52.78	-0.76
Ar→Pr	77.95	77.70	-0.25	75.44	75.42	-0.02	78.26	78.26	0.00
Cl→Ar	66.01	65.35	-0.06	62.67	61.60	-1.07	66.50	64.48	-2.02
Cl→Pr	75.04	75.06	**0.02**	74.68	75.49	**0.81**	75.24	75.11	-0.13
Pr→Ar	51.80	51.98	**0.18**	51.78	51.62	-0.16	52.55	51.68	-0.87
Pr→Cl	64.21	64.02	-0.19	62.69	64.85	-0.15	64.85	63.86	-0.99
Ave	64.21	64.02	-0.19	62.84	62.69	-0.15	64.85	63.86	-0.99

Table 7: Comparisons on the number of soft labels selected by teachers and TAs

Domain Shift	Model	Accuracy	MCC loss	Selected labels	Differential labels	Improvement by TA
Rw→Cl	Teacher	56.27	0.52	3112	29	**0.80**
	TA	56.72	0.44	3139		
Ar→Pr	Teacher	77.95	0.45	3239	-43	**0.44**
	TA	78.26	0.47	3250		
Pr→Rw	Teacher	80.45	0.42	3291	-20	**0.25**
	TA	80.22	0.43	3271		
Rw→Pr	Teacher	82.90	0.42	3412	-87	**0.23**
	TA	83.51	0.52	3325		
Cl→Pr	Teacher	75.04	0.48	3262	-41	**0.05**
	TA	75.24	0.48	3221		
Ar→Rw	Teacher	80.79	0.46	3203	-52	**0.04**
	TA	80.51	0.48	3161		
Ar→Cl	Teacher	52.21	0.66	2673	-48	-0.14
	TA	53.54	0.61	2625		
Pr→Ar	Teacher	62.22	0.49	1759	-15	-0.24
	TA	63.00	0.48	1744		
Pr→Cl	Teacher	51.80	0.48	3112	-47	-0.30
	TA	52.55	0.51	3065		
Cl→Rw	Teacher	76.68	0.44	3241	-87	-0.34
	TA	76.75	0.46	3154		
Rw→Ar	Teacher	72.19	0.46	1801	-4	-0.61
	TA	72.27	0.47	1797		
Cl→Ar	Teacher	66.01	0.48	1776	-76	-0.87
	TA	66.50	0.52	1700		

ing, PMLR, pp. 1321–1330 (2017).

[3] Han, S., Mao, H. and Dally, W. J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, *arXiv preprint arXiv:1510.00149* (2015).

[4] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).

[5] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS 2014 Deep Learning Workshop* (2015).

[6] Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R. and Bengio, Y.: Quantized neural networks: Training neural networks with low precision weights and activations, *The Journal of Machine Learning Research*, Vol. 18, No. 1, pp. 6869–6898 (2017).

[7] Jin, Y., Wang, X., Long, M. and Wang, J.: Minimum class confusion for versatile domain adaptation, *European Conference on Computer Vision*, Springer, pp. 464–480 (2020).

[8] Kim, T., Oh, J., Kim, N., Cho, S. and Yun, S.-Y.: Comparing Kullback-Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation, *arXiv preprint arXiv:2105.08919* (2021).

[9] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, Vol. 25, pp. 1097–1105 (2012).

[10] Long, M., Cao, Y., Wang, J. and Jordan, M.: Learning transferable features with deep adaptation networks, *International conference on machine learning*, PMLR, pp. 97–105 (2015).

[11] Long, M., Cao, Z., Wang, J. and Jordan, M. I.: Conditional adversarial domain adaptation, *arXiv preprint arXiv:1705.10667* (2017).

[12] Long, M., Zhu, H., Wang, J. and Jordan, M. I.: Deep transfer learning with joint adaptation networks, *International conference on machine learning*, PMLR, pp. 2208–2217 (2017).

[13] Mirzadeh, S.-I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A. and Ghasemzadeh, H.: Improved Knowledge Distillation via Teacher Assistant (2019).

[14] Saenko, K., Kulis, B., Fritz, M. and Darrell, T.: Adapting visual category models to new domains, *European conference on computer vision*, Springer, pp. 213–226 (2010).

[15] Sankaranarayanan, S., Balaji, Y., Castillo, C. D. and Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512 (2018).

[16] Sun, B. and Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation, *European Conference on Computer Vision*, p. 443–450 (2016).

[17] Yang, J., Zou, H., Cao, S., Chen, Z. and Xie, L.: MobileDA: Toward edge-domain adaptation, *IEEE Internet of Things Journal*, Vol. 7, No. 8, pp. 6909–6918 (2020).