Learning Distinctive Margin toward Active Domain Adaptation

Ming Xie[1]* Yuxi Li[2]* Yabiao Wang[2] Zekun Luo[2] Zhenye Gan[2]
Zhongyi Sun[2] Mingmin Chi[1]† Chengjie Wang[2]† Pei Wang[3]
[1]Fudan University [2]Tencent Youtu Lab [3]NAOC CAS
{mxie20,mmchi}@fudan.edu.cn
{yukiyxli,caseywang,zekunluo,wingzygan,zhongyisun,jasoncjwang}@tencent.com
wangpei@nao.cas.cn

Abstract

Despite plenty of efforts focusing on improving the domain adaptation ability (DA) under unsupervised or few-shot semi-supervised settings, recently the solution of active learning started to attract more attention due to its suitability in transferring model in a more practical way with limited annotation resource on target data. Nevertheless, most active learning methods are not inherently designed to handle domain gap between data distribution, on the other hand, some active domain adaptation methods (ADA) usually requires complicated query functions, which is vulnerable to overfitting. In this work, we propose a concise but effective ADA method called Select-by-Distinctive-Margin (SDM), which consists of a maximum margin loss and a margin sampling algorithm for data selection. We provide theoretical analysis to show that SDM works like a Support Vector Machine, storing hard examples around decision boundaries and exploiting them to find informative and transferable data. In addition, we propose two variants of our method, one is designed to adaptively adjust the gradient from margin loss, the other boosts the selectivity of margin sampling by taking the gradient direction into account. We benchmark SDM with standard active learning setting, demonstrating our algorithm achieves competitive results with good data scalability. Code is available at https://github.com/TencentYoutuResearch/ActiveLearning-SDM

1. Introduction

The domain adaptation problem has been widely studied in transfer learning society, where adaptation algorithms are designed to generalize a model trained on source domain to a target domain with different data distribution [4]. In most of studies, the semantic labels from target domain are assumed to be unavailable [4, 8, 12, 13, 24] (UDA) or only few-shot of target samples are labeled [21, 23, 29](SSDA). However, in a more practical sense, although it is difficult to annotate all data in target domain, a moderate amount of labeled data should be acceptable given certain budget on annotations cost.

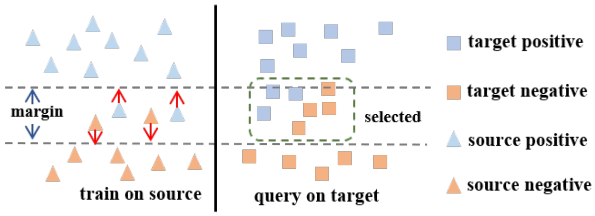


Figure 1. A simple conceptual illustration of our Select-by-Distinctive-Margin pipeline. Before each sampling step, a model is trained with a maximum margin objective, and unlabeled data lying in the margin with similar distance to different categorical centers are sampled to augment training data.

With this consideration, domain adaptation turns into an active learning problem (AL), which focusing on additionally labeling limited data to bring maximum improvement of machine learning algorithms [2,6,17,30,32,37,38]. However, currently most active learning algorithms are derived from a pure semi-supervised scenario, where the unlabeled data are assumed to conform to the same distribution as labeled data. These methods usually focus on designing a distinctive query function to depict how informative or representative an unlabeled data sample is, which highly relies on the uncertainty [6, 32] or structural distribution of data features [2, 30]. On the contrary, in a domain adaptation problem, the task model is initially trained with only source data and the query function is usually correlated to the prediction of task models, in this case, most of target data will be discriminated as uncertain regardless its location in fea-

*Both author contributed equally to this work. Work is completed during Ming Xie's internship at Tencent Youtu Lab.
†Corresponding author

ture space. Consequently, the sampling methods are prone to sample some target samples that are easily classified and make less effect on the biased decision boundaries.

Recently, there exist some researches aimed at appropriate data selection under the scenario of domain adaptation. However, these methods either design complicated and hand-crafted query function with deliberately designed architecture [11, 27], or select data in a tedious manner of high complexity [26]. These complicated design makes the query function and selection strategies easy to overfit to a certain transferring scenario and hard to be extended to more general cases. In addition, most of these methods simply exploit all source data equally during training [11, 26, 34], which is vulnerable to bias toward source domain and results unreliable query. Besides, few of studies above discuss the intrinsic relationship between their training objective and query function, ignoring the potential correlation between data of two domains during selection.

With the consideration above, in this paper, we propose to tackle domain adaptation problem with a simple but effective active learning strategy called Selective-by-Distinctive-Margin (SDM) by evaluating the distance from a data sample to different categorical clusters (as shown in Figure 1). Different from most of previous efforts focusing on selecting data through the uncertainty or diversity of pure unlabeled target data [11, 26, 27, 30, 32], SDM makes attempt to select unlabeled data via their relation to some "hard examples" from the source domain. However, instead of explicitly model such data relation, we implicitly depict the similarity between unlabeled samples and potential hard source samples via a simple maximum margin loss function. *Intuitively*, the margin loss will guide the network to maximize the distance between close examples from different categorical clusters in source domain, meanwhile ignore the affect from well classified source samples. This reversely helps detecting informative target samples still lying near the trained decision boundaries through a simple margin sampling query function. By collecting these data into training set, the manifold of decision boundaries can be further refined and generalized to target distribution. *Theoretically*, by analyzing with a simplified linear model, we confirm that model trained with margin loss can act like a Support Vector Machine [7], which collects only "hard examples" in source domain, and take these examples to detect unlabeled target data via the similarity in feature space.

In addition, derived from the simple SDM baseline, we further extend the strategy into two variants. For the training phase, for the sake of dynamically adjust gradient of margin loss to adapt to samples of different difficulties, we propose to extend the original margin loss to a dynamic form with adaptive modulation factor and max-logit reglularizer. On the other hand, during sample selection, to boost the selectivity, we take the first-order gradient of margin sampling function as additional guidance in query function, leading to select target samples which decreases the sampling function in the fastest direction with its estimated gradient. Further, both variants can further be combined together to construct more effective active learning pipeline.

Our SDM algorithm is evaluated on different domain adaption benchmarks like Office-Home [28] and Office-31 [35] under a classical active learning setting, besides, we also extend our method to a general active learning task on CIFAR-10 [18], demonstrating our approach can achieve state-of-the-art results with less query complexity and good data-scalablility. In a nutshell, our contributions can be summarized into three folds:

- We propose Select-by-Distinctive-Margin (SDM), a concise but effective active learning method for active domain adaptation, which consists of a maximum margin loss and a margin sampling function as a complete active learning cycle. Theoretical analysis is provided to show this SDM framework work like a SVM to take hard examples to mine informative targets.

- Derived from the SDM baseline, two variants are developed. One is designed for training phase to dynamically adjust margin loss gradient, the other is designed to enhance the selectivity with the help of first-order gradient of margin sampling function.

- Experiments conducted on several domain adaptation benchmarks show that our approach can achieve state-of-the-art results with limited annotation budget.

2. Related Work

Domain Adaptation. The goal of domain adaptation is to generalize a model trained on source domain to target data distribution [4]. The core issue of domain adaptation lies in the misalignment between feature and label space of source and target domain. To deal with this problem, previous domain adaptation focus on guiding a deep neural network to learn some domain invariant representation and classifiers. To be specific, the adversarial training [12, 24] is utilized to align feature distribution with a domain discriminator, regularizers like entropy constraint [13, 29] or maximum prediction rank [8] are applied to implicitly constrain the cross-domain feature space. Recently, there are also some works regard the domain alignment as minimizing the one-to-one optimal matching cost across two sets [10].

One common characteristic of methods above is that all of them assume the annotation in target domain is not accessible or only accessible for a few data, resulting unsupervised or semi-supervised domain adaptation setting. However, in a more practice scene, a moderate number of labeled data from target domain is usually allowed, and there are already some pseudo-label based methods demonstrating some properly labeled target data are powerfully enough to adapt a model from source to target domain [5,22,23,33]. As a result, new demand emerges to maximize the model transfer ability given a proper budget of annotated target data samples, which is highly overlapped with the study interests of Active Learning community.

Active Learning. The research of active learning aims at selecting proper samples to label and taking them to augment original training set and maixmum the improvement on model performance [31]. To measure the value of labeling a sample, a query function is usually designed to assign a query score to each sample for rank and selection. Classically, the query function is decided by the uncertainty metrics like entropy, score margin [3] or least confidence [20]. Recently, some advanced active learning pipelines are proposed, which are usually accompanied with deliberately designed training process, among which the Variational Auto Encoder is widely used to model the probability of erroneous prediction [6] or directly learn a binary classifier [32,38] or sample loss ranker [17,37] to select samples. Besides, there are other studies starting from the coverage of appended samples, and select data toward the objective of maximum diversity [2, 30]. All the methods above achieve promising performance on active learning task of consistent data distribution, however, none of them is designed with specific consideration of potential domain gap between labeled and unlabeled data. Consequently, these query function or sampling strategy are easy to select data with less training difficulty.

Active Domain Adaptation. AADA [34] is one of the earliest research to apply active learning technique specifically for domain adaptation, which applies a discriminator with cross-domain adversarial learning to construct sample query function. The work of [11, 27] consider the domain misalignment and design series of training objective and rules to measure the uncertainty and domainness of a target sample, [11] further proposes a randomize selection strategy to enhance the sample diversity. The method of CLUE [26] design a entropy weighted clustering algorithm to take both diversity and uncertainty of target data into an unified clustering framework.

Nevertheless, most of these approaches rely on scenario-specific prior and complicated query functions with series of hyper-parameters, making the methods easy to overfit to specific transfer scenarios and not general. Besides, there are some complicated operations like adversarial example [11,27] or clustering [26] with high complexity. In contrast, our SDM algorithm is simple in both training and data selection with insightful theoretical interpretation, by exploiting only some hard examples from source domain, our strategy can achieve promising results on different benchmarks.

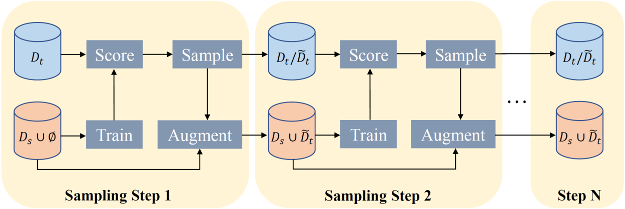


Figure 2. Illustration of active learning loop for domain adaptation

3. Approach

3.1. Problem Formulation

In the problem of Active Domain Adaptation, a labeled source domain is denoted as $\mathcal{D}_s = \{(x_s, y_s)\}$, with data x_s and its semantic label $y_s \in \{1, 2, \cdots, K\}$, where K is the number of class types, an unlabeled target domain is denoted as $\mathcal{D}_t = \{x_t\}$. Meanwhile, we denote a labeled target set as $\widetilde{\mathcal{D}}_t$, which is an empty set ϕ initially. With these initial data and a given annotation budget B, an active domain adaptation loop can be build as Figure 2. The unlabeled data is sampled several times, for each selected data $\hat{x}_t \in \mathcal{D}_t/\widetilde{\mathcal{D}}_t$, annotators will assign its label $\hat{y}_t$ to it, and $\widetilde{\mathcal{D}}_t$ is augmented with new labeled target data $\{(\hat{x}_t, \hat{y}_t)\}$ after each sampling step, then the model can be trained with $\mathcal{D}_s \cup \widetilde{\mathcal{D}}_t$, afterward the updated model is exploited to select new target data for annotation from set $\mathcal{D}_t/\widetilde{\mathcal{D}}_t$. The process repeats until appended number of target samples achieves the budget $|\widetilde{\mathcal{D}}_t| = B$. For the ease of representation, we denote our model as a composition of a feature extractor $g(\cdot)$ to extract data feature $\mathbf{f} = g(x)$ and a linear classifier $c(\cdot)$ to categorize a feature into class logit vector of size K.

3.2. Select by Distinctive Margin

3.2.1 Pipeline

In a classical paradigm, all labeled data from $\mathcal{D}_s \cup \widetilde{\mathcal{D}}_t$ can be utilized to train a new deep network, which is widely followed by previous ADA approaches [11, 26, 27]. However, such strategy makes trained model dominated and biased toward some salient area in source domain of high data density at early stages, and reversely prevent the query function from detecting informative target data.

To mitigate such source-oriented bias, we propose to exploit only "hard examples" from source domain to construct our training objective, since these examples are important to shape the decision boundaries with less domain-biased information. Therefore we design the categorical-wise margin loss to supervise the network output due to its inherent selective property

$$\mathcal{L}_m(x, y) = \sum_{i \neq y} [m - c(g(x))_y + c(g(x))_i]_+ \qquad (1)$$

where $[x]_+$ denotes zero-clip operation $max(0, x)$, the subscript y and i indicates the y-th and i-th entry of vectors, and m is a hyperparameter to control the expected margin width. From Eq (1), we see only samples with similar classification score between ground-truth class and other classes can contribute the gradient for deep network, thus the model will not be dominated by redundant source samples and be easier to transfer to target domain.

On the other hand, since the loss in Eq (1) explicitly enlarges the gap between different category clusters, it is natural to pay more attention to those samples with smaller gap between category-scores in target domain due to the impact they will have on current learned decision boundaries. As a result, a margin sampling query function is proposed to evaluate the importance of an unlabeled target sample

$$\mathbf{p} = \mathbf{softmax}(c(g(x_t))) \tag{2}$$

$$Q(x) = 1 - (\mathbf{p}_{1^*} - \mathbf{p}_{2^*}) \quad \forall x_t \in \mathcal{D}_t / \widetilde{\mathcal{D}}_t \tag{3}$$

where the subscript 1^* and 2^* indicates the index of maximum and second maximum value of a vector, before computing the query function, softmax operation is applied to map the logit vector to normalized probability to ensure the scale of $Q(x) \in (0, 1)$. The smaller category gap a sample has, the larger $Q(x)$ value is assigned. Therefore unlabeled target data can be re-ranked by the metric in Eq (3) and top ranked samples are labeled to augment training set.

3.2.2 Theoretical Insights

To further discuss how a marginal loss helps our model to select informative sample under the margin sampling query function, we simplified our model into a parameterized binary linear classification problem $c(g(x)) = [w_+, w_-]^T x$, where a data feature $x \in \mathcal{R}^D$ can only be categorized as positive or negative, under this setting, we prove that **the query function $Q(x)$ is correlated to the similarity between x and "hard examples" during training**. To be specific, we denote a training batch as $\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^-$, where $\mathcal{S}^+$ denotes the set of positive sample and $\mathcal{S}^-$ contains all negative samples in a batch, these training samples are exploited to train a binary linear classifier with positive weight w_+ and negative weight w_- via a margin loss as Eq (1), after training, a sample x can be discriminated via its predicted probability of belonging to a certain category

$$p^+(x) = \frac{e^{w_+^T x}}{e^{w_+^T x} + e^{w_-^T x}} \quad p^-(x) = \frac{e^{w_-^T x}}{e^{w_+^T x} + e^{w_-^T x}} \tag{4}$$

Data Selection. With the formulation in Eq (4), we define a Signed Local Similarity Indicator $\mathcal{I}(x; \mathcal{S})$ as Definition 1.

Definition 1 (Signed Local Similarity Indicator)**.** *Given a sample feature x, its Signed Local Similarity Indicator $\mathcal{I}(x; \mathcal{S})$ is defined as*

$$\mathcal{I}(x; \mathcal{S}) = \sum_{x_p \in \mathcal{S}^+} \delta(m > w_+^T x_p - w_-^T x_p) x_p^T x - \sum_{x_n \in \mathcal{S}^-} \delta(m > w_-^T x_n - w_+^T x_n) x_n^T x \tag{5}$$

where $\delta(\cdot)$ equals 1 if the condition inside holds otherwise equals 0.

From the definition, we see the indicator $\mathcal{I}(x; \mathcal{S})$ only focuses on the similarity between x and those labeled samples x_p, x_n close to classification boundary, i.e. samples that are vague for current classifiers to discriminate, when x manifests stronger similarity with vague positive samples x_p in batch $\mathcal{S}$, $\mathcal{I}(x; \mathcal{S})$ increases, in contrast, when x is closer to vague negative samples, $\mathcal{I}(x; \mathcal{S})$ gets smaller value. With the Definition 1, we claim that the Proposition 1 holds

Proposition 1. [‡] *If an unlabeled sample $x_t \in \mathcal{D}_t / \widetilde{\mathcal{D}}_t$ is measured by query function $Q(x_t)$ as Eq (3), after a gradient descending step on batch $\mathcal{S}$, then the following monotonicity holds*

- *if $p^+(x_t) > p^-(x_t)$, $Q(x_t)$ is decreasing monotonically with respect to $\mathcal{I}(x_t; \mathcal{S})$*
- *if $p^+(x_t) < p^-(x_t)$, $Q(x_t)$ is increasing monotonically with respect to $\mathcal{I}(x_t; \mathcal{S})$*

With the Proposition 1, we see our margin loss performs under a mechanism analogous to Support Vector Machine [7], where only a few hard examples (like the support vectors) are collected as component to decide the query function score of a target sample x_t, e.g. if the trained classifier predicts that sample x_t is more likely to be positive, i.e. $p^+(x_t) > p^-(x_t)$, then the closer x_t is to existing hard positive samples, the less query function value $Q(x_t)$ will be obtained, in contrast, when x_t is closer to some hard negative samples, margin loss will impose a larger $Q(x_t)$ score.

Transferability. It should be noticeable that SDM is not only suitable for data selection, but also helps domain transfer theoretically. Following the analysis of [29], we define a margin-based domain classifier space as

$$\mathcal{H} = \{h(x)\} = \left\{\delta(|w_+^T x - w_-^T x| \geq m) | w_+, w_- \in \mathcal{R}^D\right\} \tag{6}$$

then we can obtain the Proposition 2 to verify that SDM helps to shrink the domain gap [4] under certain assumption

Proposition 2. [‡] *For source and target data $x_s \sim \mathcal{P}_s, x_t \sim \mathcal{P}_t$, given the margin domain classifier family $\mathcal{H}$ of Eq (6), if $\mathcal{P}(h(x_t) = 1) \leq \mathcal{P}(h(x_s) = 1)$, then optimizing the binary margin loss is equivalent to minimize the upperbound of domain $\mathcal{H}$-divergency $d_{\mathcal{H}}(\mathcal{P}_s, \mathcal{P}_t)$ defined by [4].*

[‡] The proof can be found at supplementary material.

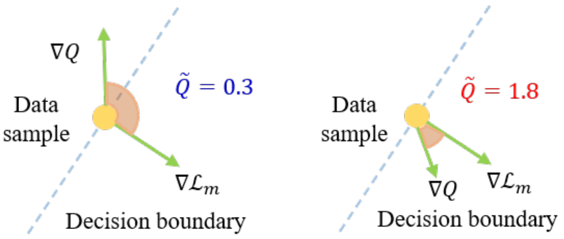


Figure 3. Examples of query with first-order differential margin. The left figure shows situation where the gradient direction from loss and query function diverge a lot. The right figure illustrates an example where the feature gradient from both loss and query function share similar update direction and yield high query score.

3.3. Variants

Dynamically Adjusted Margin Loss. Although the loss in Eq (1) can implicitly select hard source data, it still shows some flaws. First, all hard samples contributes equally in terms of the backward gradient. Besides, the margin constraint only considers the relative distance from samples to different class decision boundaries, ignoring constraint on the absolute score of ground-truth class label. With this consideration, we propose a dynamic version of margin loss to adaptively adjust the backward gradient in proportion to the margin size, and append a max-logit regularizer to ensure the gradient from ground-truth class will not vanish even if the margin is large enough than pre-defined m

$$\widetilde{\mathcal{L}}_m(x,y) = \sum_{i\neq y} \alpha_i \left[m - c(g(x))_y + c(g(x))_i\right]_+ - c(g(x))_y \tag{7}$$

$$\alpha_i = 1 - \frac{c(g(x))_y - c(g(x))_i}{m}$$

In Eq (7) we modify the rectified margin as a modulation factor α_i and take it to modulate the score of other classes except ground-truth. With this modulation, the loss term with smaller margin will be emphasized and generate larger gradient to push sample away from corresponding categorical clusters, helping our network adpatively focusing on hard source examples of different difficulties. Besides, a max-logit term is appended in Eq (7) to constrain our network to always assign large score to prediction on ground-truth class. This kind of variant is termed as “SDM-A”.

Query with Gradient Direction Consistency. To boost the selectivity of our query function, derived from the basic marge sampling in Eq (3), we further take its variation into account. Inspired by [2] which applies the weight variation to depict the data importance, we expect that the gradient from a newly appended sample will push its feature representation $\mathbf{f}$ toward direction that minimizes the margin sampling function as examplified in Figure 3, this is equivalent to ensure the gradient from both loss term and margin sampling manifests similar orientation in feature space

$$\widetilde{Q}(x) = Q(x) + \lambda \left\langle \nabla_{\mathbf{f}}\mathcal{L}_m(x,y), \nabla_{\mathbf{f}} Q_m(x)\right\rangle \tag{8}$$

where $\langle\cdot,\cdot\rangle$ is cosine-similarity metric, λ is a balance factor. However, it is not possible to acquire the annotation y of an unlabeled sample before selection, instead, we take the probabilistic gradient estimation $\nabla_{\mathbf{f}}\hat{\mathcal{L}}_m(x)$ which is consistent with margin sampling

$$\nabla_{\mathbf{f}}\hat{\mathcal{L}}_m(x) = \mathbf{p}_{1^*}\nabla_{\mathbf{f}}\mathcal{L}_m(x,1^*) + \mathbf{p}_{2^*}\nabla_{\mathbf{f}}\mathcal{L}_m(x,2^*) \tag{9}$$

where the notation $\mathbf{p}, 1^*, 2^*$ follow the same definition from Eq (3). Through the modified query function in Eq (8), the sampled data is not only close to decision boundaries of trained model, but also ensured to fast converged to a non-fuzzy state. This variant is termed as “SDM-G”. Besides, the two variants are not mutually exclusive to each other and can be exploited simultaneously to obtain a combined active learning pipeline as “SDM-AG”.

4. Experiments

4.1. Setup

Dataset and Metric. In our experiments, We first evaluate the performance of our framework on two mainstream domain adaptation benchmarks, Office-Home [28] and Office-31 [35]. Then we further extend our method to single-domain dataset CIFAR-10 [18] to validate the generality of SDM. The Office-31 dataset includes 3 different domain with imbalance image distribution, there are total 4110 images of 31 object categories. The Office-Home dataset is a more challenging benchmark consisting of 4 different domains and 65 different types of objects. CIFAR-10 is a widely used dataset for different machine learning tasks, there are total 50000 images for 10 common classes. Following the work of [11], for experiments on Office-31 and Office-Home, we report results on all transfer scenarios and average the accuracy on different scenarios for final comparison. Our active learning loop starts with data from only source domain, at each sampling step, 1% of target data is sampled, and totally 5 times of sampling steps are conducted. For CIFAR-10, our training process starts from 10% of full training data, at each sampling step, 5% of data is sampled and the budget is set as 30% of training data.

Implementation Detail. Our experiments are implemented with Pytorch framework. Following the setting of [11], we take the commonly used ResNet50 [14] architecture which is pre-trained on ImageNet [19] as our feature extractor and classifier. Different from some of previous methods for ADA [11, 26] combining unsupervised domain adaptation methods and trained with data from target domain, in our implementation, we avoid training on unlabeled data

Method	Office-Home												
	A → C	A → P	A → R	C → A	C → P	C → R	P → A	P → C	P → R	R → A	R → C	R → P	Avg
ResNet [14]	42.1	66.3	73.3	50.7	59.0	62.6	51.9	37.9	71.2	65.2	42.6	76.6	58.3
RAN	56.8	78.0	77.7	58.9	70.7	70.5	60.9	53.2	76.8	71.5	57.5	81.8	67.9
ENT	56.8	80.0	82.0	59.4	75.8	73.8	62.3	54.6	80.3	73.6	58.8	85.7	70.2
CONF	57.7	81.3	82.2	60.8	76.5	74.2	61.9	54.5	80.4	73.4	59.4	85.9	70.7
MAR	58.6	81.3	81.7	60.3	76.2	73.6	63.4	55.2	80.5	73.8	60.5	86.3	70.9
QBC [9]	56.9	78.0	78.4	58.5	73.3	69.6	60.2	53.3	76.1	70.3	57.1	83.1	67.9
Cluster [25]	56.0	76.8	78.1	58.4	72.6	69.2	58.4	51.2	75.4	70.1	56.4	82.4	67.1
AADA [34]	56.6	78.1	79.0	58.5	73.7	71.0	60.1	53.1	77.0	70.6	57.0	84.5	68.3
ADMA [15]	57.2	79.0	79.4	58.2	74.0	71.1	60.2	52.2	77.6	71.0	57.5	85.4	68.6
BADGE [2]	59.2	81.0	81.6	60.8	74.9	73.3	63.7	54.2	79.2	73.6	59.7	85.7	70.6
TQS [11]	58.6	81.1	81.5	61.1	76.1	73.3	61.2	54.7	79.7	73.4	58.9	86.1	70.5
SDM-AG	**61.2**	**82.2**	**82.7**	**66.1**	**77.9**	**76.1**	**66.1**	**58.4**	**81.0**	**76.0**	**62.5**	**87.0**	**73.1**

Table 1. Classification accuracy (%) on the Office-Home dataset with the budget of 5% data. Among the abbreviation, “RAN” is random sampling, “ENT” is entropy-based sampling, “CONF” is least confidence sampling and “MAR” is pure margin sampling.

Method	Office-31						
	A→W	A→D	W→A	W→D	D→A	D→W	Avg
ResNet [14]	81.5	75.0	63.1	95.2	65.7	99.4	80.0
RAN	87.1	84.1	75.5	98.1	75.8	99.6	86.7
UCN [16]	89.8	87.9	78.2	99.0	78.6	100.0	88.9
QBC [9]	89.7	87.3	77.1	98.6	78.1	99.6	88.4
Cluster [25]	88.1	86.0	76.2	98.3	77.4	99.6	87.6
AADA [34]	89.2	87.3	78.2	99.5	78.7	100.0	88.8
ADMA [15]	90.0	88.3	79.2	100.0	79.1	100.0	89.4
CLUE [26]	88.1	91.4	76.1	100.0	76.1	98.6	88.4
TQS [11]	92.2	92.8	80.4	100.0	80.6	100.0	91.0
SDM-AG	**93.5**	**94.8**	**81.9**	**100.0**	**81.9**	**100.0**	**92.0**

Table 2. Classification accuracy (%) on the Office-31 dataset with the budget of 5% data. “RAN” represents random sampling.

with any unsupervised learning technique for fairer comparison, this also makes our SDM suitable for both pooled and sequential setting of active learning. During the training process, we first train our network with initial data for 10 epochs with margin loss and an auxiliary cross entropy loss, after which we start our sampling steps. The sampling process is performed every two epochs until the labeled target data reaches the total budget. The learning rate is set to be 0.01 and batch size is set as 72. We set the hyper parameter margin m in Eq (1) to 1 and λ in Eq (8) to 0.01 in terms of detailed ablation studies.

4.2. Main Results

We compare our “SDM-AG” pipeline with other active learning approaches on different benchmarks. We take a ResNet50 trained with pure initial source data as our baseline method for comparison, methods with classical active learning strategies [2, 9, 16, 25] are taken into account, further, we also compare our methods with recent state-of-the-art ADA approaches [11, 15, 26, 34]. Besides, we also compare with some commonly used simple query functions like random sampling (RAN), entropy-based sampling (ENT), least confidence (CONF) and margin sampling (MAR).

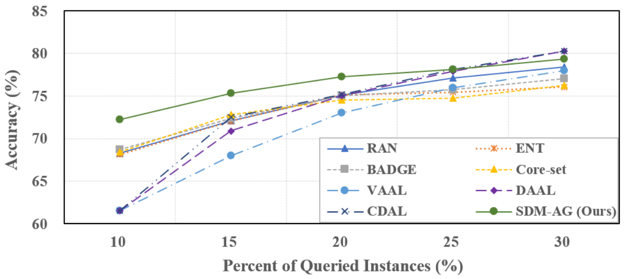


Figure 4. Experiment results on CIFAR-10 dataset from 10% training data to 30% training data. “RAN” is random sampling and “ENT” is entropy-based sampling.

The comparison results on Office-Home are presented in Table 1. From this table, we see our SDM-AG pipeline outperforms either classical active learning approaches or recent ADA methods designed with complicated selection strategies. To be specific, our SDM-AG method can bring about +2.6% performance gain in average accuracy over state-of-the-art active learning methods like TQS [11] or BADGE [2]. Further, it can be observed that in some harder scenarios with larger discrepancy between source and target (e.g. *C to A* and *P to A*), the improvement from our SDM-AG method is more salient. In total, our method can achieve +14.8% improvement on the average performance over the baseline with pure source data. Similar results can be found on the dataset of Office-31, which is listed in Table 2. Although some transferring scenarios in this benchmark is kind of saturated, it can still be observed that SDM-AG achieves substantial performance gain over other state-of-the-art method [11, 26, 34] on some challenging scenarios, and our simple pipeline can outperforms all compared

Method	Adjust	Gradient	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
Baseline			58.6	81.3	81.7	60.3	76.2	73.6	63.4	55.2	80.5	73.8	60.5	86.3	70.9
SDM			60.5	79.6	81.4	65.3	76.5	74.9	65.8	56.5	80.6	75.2	61.1	85.7	71.9
SDM-A	✓		60.7	81.5	82.1	65.7	76.8	**76.3**	**66.3**	58.1	80.2	75.2	**62.7**	86.6	72.7
SDM-G		✓	**61.2**	81.9	**82.7**	65.6	77.6	76.1	66.0	58.0	80.8	75.8	61.8	86.9	72.9
SDM-AG	✓	✓	**61.2**	**82.2**	**82.7**	**66.1**	**77.9**	76.1	66.1	**58.4**	**81.0**	**76.0**	62.5	**87.0**	**73.1**

Table 3. Ablation study with different configuration with 5% of target labeled data on Office-Home dataset. "Baseline" is a model trained with cross-entropy loss and selecting data with margin sampling. "Adjust" indicates dynamically adjusted margin loss. "Gradient" denotes first-order gradient consistency.

Sample Strategy	Training Loss	Acc	Δ
Entropy	Cross Entropy	70.24	+0.28
	Margin Loss	**70.52**	
Least Confidence	Cross Entropy	70.68	+0.54
	Margin Loss	**71.22**	
Margin Sample	Cross Entropy	70.94	**+0.98**
	Margin Loss	**71.92**	

Table 4. Comparison with different combination between different types of training loss and sampling strategies on Office-Home dataset. The "Acc" represents the averaged accuracy over all 12 transferring scenarios. Δ denotes improvement with margin loss.

methods in terms of averaged domain adaptation accuracy.

In addition, we also extend our experiment and comparison to a general active learning setting without domain gap on the benchmark of CIFAR-10. The results are evaluated after training of each sampling step and ploted in Figure 4. It can be observed that our SDM pipeline can still outperforms most of other state-of-the-art methods [2, 30, 32] regardless of numbers of labeled data and comparable to some newest AL algorithms [1,36]. It is also noticeable that when the number of queried number is small (e.g. 10% $\sim$ 20% of training data), SDM outperforms all competitors including recently proposed DAAL [36] or CDAL [1] by a large margin, demonstrating our SDM algorithm is more friendly to scenarios of active learning with low budget.

4.3. Detailed Analysis

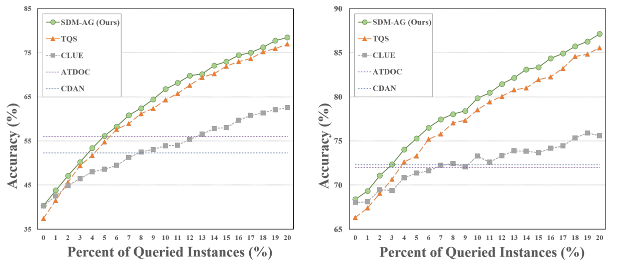
In this section, we analysis the components of our algorithm in detail. If not specified, the analysis is conducted on Office-Home with our default setting.

Improvement over Cross-Entropy Baseline. First we conduct experiments to investigate the superiority of SDM over a simple active learning baseline. To thie end, we design a baseline method where the network is trained with pure cross-entropy loss, but selecting samples with the same criterion of margin sampling as Eq (3). The comparison on different transferring scenarios is shown in Table 3. It is observed that our SDM paradigm, i.e. model trained with margin loss and selecting data by margin sampling achieves consistent improvement over cross-entropy baseline on most scenarios and overall performance. This observation indicates that the improvement is the results of the whole solution of SDM instead of a simple inclusion of sampling strategy.

Effectiveness of Different Variants. Next we investigate the improvement of different variants based on our SDM baseline. The results are listed in Table 3. From the table, we observe that both SDM-A and SDM-G can bring significant performance gain compared with simple SDM pipeline, demonstrating the improved dynamic margin loss and query function with gradient guidance can benifit the active learning process respectively. Besides, we see the combination of two variants, i.e. SDM-AG can further boost the averaged performance to at most 73.1% and achieve the best results on most of scenarios of Office-Home dataset.

Compatibility between Margin Loss and Sampling. In the discussion of Proposition 1, we have show training with margin loss is inherently helpful for margin sampling especially to mine informative data from target domain. In this section we further investigate this property with empirical results. To this end, we test different combination between training objective and query functions. For training loss, we investigate the margin loss and commonly used cross-entropy loss, as for sampling strategy, in addition to margin sampling, the commonly used least confidence and entropy sampling strategies are exploited. The test results are shown in Table 4, from the table we can conclude that: (1) Regardless of the sampling strategy we use, margin loss can bring improvement over pipeline trained with cross-entropy. (2) In terms of the performance gain (Δ in Table 4), the margin sampling strategy obtains the most gain in average accuracy, indicating that the margin loss is inherently suitable for a relative margin-based data selection strategy to mine informative data for domain transfer, which is consistent to Proposition 1.

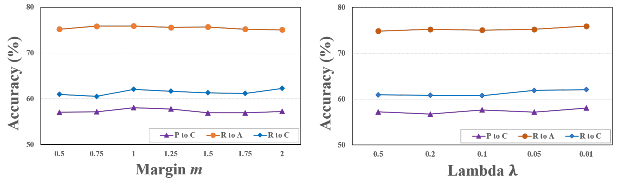
Variation with Different Budget Size. The annotation budget B is an important parameter for active learning since it decides the available target data to be labeled, therefore we test how the domain adaptation performance varies with



(a) P to C

(b) R to A

Figure 5. Performance variation with different budget size on different scnarios of Office-Home dataset.



(a) Analysis of m

(b) Analysis of λ

Figure 6. Sensitivity analysis on hyper-parameters of SDM on different scenarios of Office-Home dataset.

increasing budget. To make horizontal comparison, we also compare SDM with other two ADA methods TQS [11] and CLUE [26]. In addition, we also compare with recent unsupervised DA methods like CDAN [24] and ATDOC [23] to see the required sample number to achieve competitive results to these approaches. The budget size B is controlled within the range from 0% to 20% of target data. The results are plotted as curves in Figure 5. When compared with TQS and CLUE, our SDM-AG methods can achieve consistent improvement regardless of the budget size, this superiority is more obvious on the scenario of *R to A*, demonstrating that our method can steadily benefit from the growth of budget B and is not easy to saturate. Besides, when compared with ATDOC and CDAN, our methods can achieve the comparable results with a burden of only 5% target data labeled, demonstrating the efficiency of our algorithm.

Sensitivity of Hyper-parameter. We further test how the hyper-parameters in our SDM pipeline affect the overall performance of domain adaptation to see if our algorithm is sensitive to some parameters. To be specific, we tune the training margin m of Eq (7) and balance factor λ in Eq (8) within a tolerable range and test the accuracy on three scenarios of different difficulties (*P to C*, *R to C*, *R to A*). The results are ploted as curves in Figure 6. On all scenarios, we see the accuracy varies marginally with the tuned hyper-parameters. This observation demonstrate our approach is stable and not sensitive to specific hyper-parameters.

Complexity Analysis. Finally, we analyze the complexity and running time to confirm the claim that our SDM algorithm is a simple pipeline compared with other complicated ADA methods. To be specific, we compare the theoretical complexity and actual running time of one round of data query and sampling. We compare SDM with state-of-the-art clustering method [2, 26] and ranking method [11]. For all methods, we ignore the running time and complexity of network forward pass since this is the common step and consumes the same time, and for all rank-based methods, we assume a stable comparison sort algorithm is applied with the lower bound of complexity $\mathcal{O}(N \log N)$ to sort all data. The comparison results are listed in Table 5, readers can refer to the appendix material for more details about the complexity derivation of SDM. In Table 5, we see the rank-based methods do not rely on budget size B, resulting more efficient complexity. In terms of running time, SDM achieves 24.6× query speed compared with the nearest competitor [26] and is much faster than TQS [11], since TQS requires parsing results from multiple classifiers and running an additional discrimination network for domainess.

Method	Query Complexity	Time (s)
BADGE [2]	$\mathcal{O}(BNKD)$	11.47
CLUE [26]	$\mathcal{O}(tNBD)$	1.65
TQS [11]	$\mathcal{O}(NMK + N \log N)$	2.19
SDM-AG (ours)	$\mathcal{O}(NKD + N \log N)$	**0.067**

Table 5. Comparison between complexity and running time of different methods. B is the budget size, K is number of classes, D is feature dimension, N is number of target samples, t is the clustering iteration in [26] and M is the committee size in [11].

5. Conclusion

In this paper, aimed at the active domain adaptation problem, we propose a simple but effective solution termed as Select-by-Distinctive-Margin (SDM). We provide theoretical analysis to show how a model trained with margin loss select informative data, and further propose two variants to enhance the model training and data sampling. Comprehensive experiment results demonstrate that our algorithm is a concise, stable and superior solution toward the active domain adaptation problem.

Acknowledgement

This work is supported by following foundations: National Natural Science Foundation of China No. 11988101, Natural Science Foundation of China under contract 62171139, National Key R&D Program under contract (2017YFA0402600), National Natural Science Foundation of China under grant U2031117, the Youth Innovation Promotion Association CAS (id. 2021055), CAS Project for Young Scientists in Basic Reasearch (grant YSBR-006) and the Cultivation Project for FAST Scientific Payoff and Research Achievement of CAMS-CAS.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020. 7

[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2019. 1, 3, 5, 6, 7, 8

[3] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007. 3

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010. 1, 2, 4

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[6] Jongwon Choi, Kwang Moo Yi, Jihoon Kim, Jinho Choo, Byoungjip Kim, Jinyeop Chang, Youngjune Gwon, and Hyung Jin Chang. Vab-al: Incorporating class imbalance and difficulty with variational bayes for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6758, 2021. 1, 3

[7] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 2, 4

[8] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020. 1, 2

[9] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995. 6

[10] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 447–463, 2018. 2

[11] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7272–7281, 2021. 2, 3, 5, 6, 8

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 1, 2

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances Neural Information Processing Systems*, pages 529–536, 2004. 1, 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 770–778, 2016. 5, 6

[15] Sheng-Jun Huang, Jia-Wei Zhao, and Zhao-Yang Liu. Cost-effective training of deep cnns with active model adaptation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1580–1588, 2018. 6

[16] Ajay J Joshi, Fatih Porikli, and Nikolaos P Papanikolopoulos. Scalable active learning for multiclass image classification. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2259–2273, 2012. 6

[17] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8166–8175, 2021. 1, 3

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012. 5

[20] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994. 3

[21] Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1104–1113, 2021. 1

[22] Kai Li, Chang Liu, Handong Zhao, Yulun Zhang, and Yun Fu. Ecacl: A holistic framework for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8578–8587, October 2021. 3

[23] Jian Liang, Dapeng Hu, and Jiashi Feng. Domain adaptation with auxiliary target domain-oriented classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2021. 1, 3, 8

[24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1647–1657, 2018. 1, 2, 8

[25] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004. 6

[26] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8505–8514, October 2021. 2, 3, 5, 6, 8

[27] Harsh Rangwani, Arihant Jain, Sumukh K Aithal, and R. Venkatesh Babu. S3vaada: Submodular subset selection for virtual adversarial active domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7516–7525, October 2021. 2, 3

[28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 2, 5

[29] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 1, 2, 4

[30] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 7

[31] Burr Settles. Active learning literature survey. *Science*, 10(3):237–304, 1995. 3

[32] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019. 1, 2, 3, 7

[33] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[34] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020. 2, 3, 6

[35] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 2, 5

[36] Shuo Wang, Yuexiang Li, Kai Ma, Ruhui Ma, Haibing Guan, and Yefeng Zheng. Dual adversarial network for deep active learning. In *European Conference on Computer Vision*, pages 680–696. Springer, 2020. 7

[37] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019. 1, 3

[38] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3