S^{3}VAADA: Submodular Subset Selection for Virtual Adversarial Active Domain Adaptation

Harsh Rangwani Arihant Jain* Sumukh K Aithal* R. Venkatesh Babu
Video Analytics Lab, Indian Institute of Science, Bengaluru, India
harshr@iisc.ac.in, arihantjain@iisc.ac.in, sumukhaithal6@gmail.com, venky@iisc.ac.in

Abstract

Unsupervised domain adaptation (DA) methods have focused on achieving maximal performance through aligning features from source and target domains without using labeled data in the target domain. Whereas, in the real-world scenario's it might be feasible to get labels for a small proportion of target data. In these scenarios, it is important to select maximally-informative samples to label and find an effective way to combine them with the existing knowledge from source data. Towards achieving this, we propose S^{3}VAADA which i) introduces a novel submodular criterion to select a maximally informative subset to label and ii) enhances a cluster-based DA procedure through novel improvements to effectively utilize all the available data for improving generalization on target. Our approach consistently outperforms the competing state-of-the-art approaches on datasets with varying degrees of domain shifts. The project page with additional details is available here: https://sites.google.com/iisc.ac.in/s3vaada-iccv2021/.

1. Introduction

Deep Neural Networks have shown significant advances in image classification tasks by utilizing large amounts of labeled data. Despite their impressive performance, these networks produce spurious predictions hence suffer from performance degradation when used on images that come from a different domain [26] (e.g., model trained on synthetic data (source domain) being used on real-world data (target domain)). Unsupervised Domain Adaptation (DA) [9, 18, 5, 28, 14, 15] approaches aim to utilize the labeled data from source domain along with the unlabeled data from the target domain to improve the model's generalization on the target domain. However, it has been observed that the performance of Unsupervised DA models often falls short in comparison to the supervised methods [37], which leads to their reduced usage for performance critical applications. In such cases, it might be possible to label some of the target data to improve the performance of the model.

*Equal Contribution

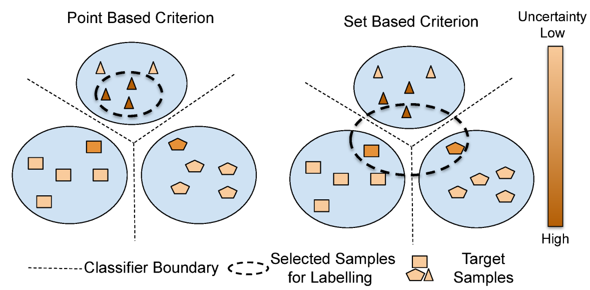


Figure 1. We pose sample selection for labeling in **Active Domain Adaptation** as an informative subset selection problem. We propose an information criterion to provide *score for each subset* of samples for labeling. Prior works (See t-SNE for AADA [35] in Sec. 1 of supplementary material) which use a point-based criterion *(i.e. score each sample independently)* to select samples suffer from redundancy. As our set-based criterion is aware of the other samples in the set, it avoids redundancy and selects diverse samples.

In such a case, the dilemma is, *"Which samples from the target dataset should be selected for labeling?"*. Active Learning (AL) [7, 32] approaches aim to provide techniques to select the maximally informative set for labeling, which is then used for training the model. However, these approaches do not effectively use the unlabeled data and labeled data present in various domains. This objective contrasts with Unsupervised DA objective that aims to use the unlabeled target data effectively. In practice, it has been found that just naively using AL and fine-tuning offers sub-optimal performance in presence of domain shift [35].

Another question that follows sample selection (or sampling) is, *"How to effectively use all the data available to improve model performance?"*. Unsupervised DA approaches based on the idea of learning invariant features for both the source and target domain have been known to be ineffective in increasing performance when additional labeled data is present in target domain [27]. Semi-Supervised DA

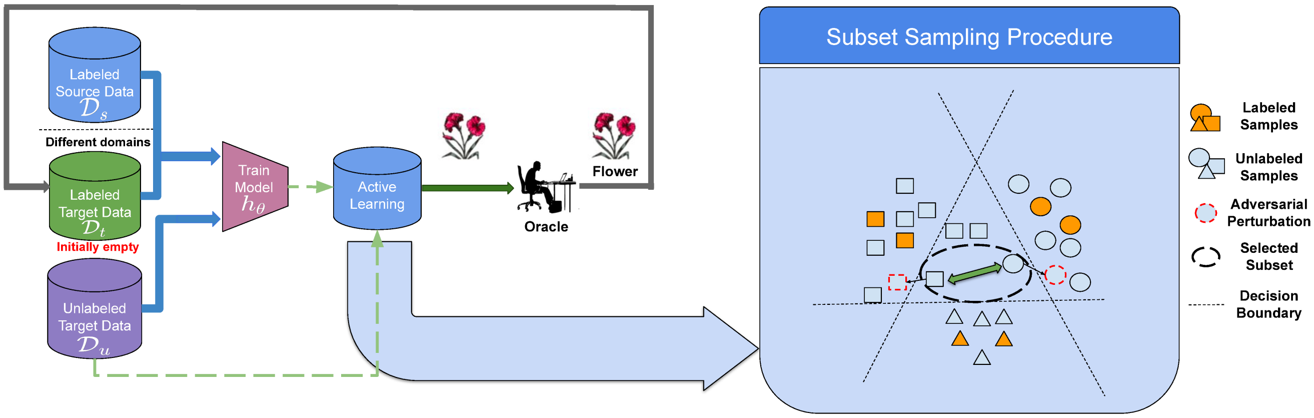


Figure 2. Overview of Submodular Subset Selection for Virtual Adversarial Active Adaptation (S^{3}VAADA). Step 1: We select a subset of samples which are uncertain (i.e., prediction can change with small perturbation), diverse and representative in feature space (see Fig. 3 for details). Step 2: The labeled samples and the unlabeled samples are used by proposed VAADA adaptation procedure to obtain the final model. The above two steps are iteratively followed selecting B samples in each cycle, till the annotation budget is exhausted.

(SSDA) [27, 17] methods have been developed to mitigate the above issue, but we find their performance plateau's as additional data is added (Sec. 5.4). This is likely due to the assumption in SSDA of only having a small amount of labeled data per-class (i.e., few shot) in target domain which is restrictive.

The Active Domain Adaptation (Active DA) paradigm introduced by Rai et al. [25] aims to first effectively select the informative-samples, which are then used by a complementary DA procedure for improving model performance on target domain. The state-of-the-art work of AADA [35] aim to select samples with high value of $p_{target}(x)/p_{source}(x)$ from domain discriminator, multiplied by the entropy of classifier which is used by DANN [9] for adaptation. As the AADA criterion is a point-estimate, it is unaware of other selected samples; hence the samples selected can be redundant, as shown in Fig. 1.

In this work, we introduce Submodular Subset Selection for Virtual Adversarial Active Domain Adaptation (S^{3}VAADA) which proposes a set-based informative criterion that provides scores for each of the subset of samples rather than a point-based estimate for each sample. As the information criteria is aware of other samples in the subset, it tends to avoid redundancy. Hence, it is able to select diverse and uncertain samples (shown in Fig. 1). Our subset criterion is based on the cluster assumption, which has shown to be widely effective in DA [8, 16]. The subset criterion is composed of a novel uncertainty score (Virtual Adversarial Pairwise (VAP)) which is based on the idea of the sensitivity of model to small adversarial perturbations. This is combined with a distance based metrics such that the criterion is submodular (defined in Sec. 3.1). The submodularity of the criterion allows usage of an efficient algorithm [20] to obtain the optimal subset of samples. After obtaining the labeled data, we use a cluster based domain adaption scheme based on VADA [33]. Although VADA, when naively used is not able to effectively make use of the additional target labeled data [29], we mitigate this via two modifications (Sec. 4.2) which form our Virtual Active Adversarial Domain Adaptation (VAADA) procedure.
In summary, our contributions are:

- We propose a novel set-based information criterion which is aware of other samples in the set and aims to select uncertain, diverse and representative samples.
- For effective utilization of the selected samples, we propose a complementary DA procedure of VAADA which enhances VADA's suitability for active DA.
- Our method demonstrates state-of-the-art active DA results on diverse domain adaptation benchmarks of Office-31, Office-Home and VisDA-18.

2. Related Work

Domain Adaptation: One of the central ideas in DA is minimizing the discrepancy in two domains by aligning them in feature space. DANN [9] achieves this by using domain classifier which is trained through an adversarial min-max objective to align the features of source and target domain. MCD [29] tries to minimize the discrepancy by making use of two classifiers trained in an adversarial fashion for aligning the features in two domains. The idea of semi-supervised domain adaptation by using a fraction of labeled data is also introduced in MME [27] approach, which induces the feature invariance by a MinMax Entropy objective. Another set of approaches uses the cluster assumption to cluster the samples of the target domain and source domain. In our work, we use ideas from VADA (Virtual Adversarial Domain Adaptation) [33] to enforce cluster assumption.

Active Learning (AL): The traditional AL methods are iterative schemes which obtain labels (from oracle or experts) for a set of informative data samples. The newly labeled samples are then added to the pool of existing labeled data and the model is trained again on the labeled data. The proposed techniques can be divided into two classes: 1) **Uncertainty Based Methods**: In this case, model uncertainty about a particular sample is measured by specific criterion like entropy [40] etc. 2) **Diversity or Coverage Based Methods**: These methods focus on selecting a diverse set of points to label in order to improve the overall performance of the model. One of the popular methods, in this case, is Core-Set [30] which selects samples to maximize the coverage in feature space. However, recent approaches like BADGE [1] which use a combination of uncertainty and diversity, achieve state-of-the-art performance. A few task-agnostic active learning methods [34, 42] have also been proposed.

Active Domain Adaptation: The first attempt for active domain adaptation was made by Rai et al. [25], who use linear classifier based criteria to select samples to label for sentiment analysis. Chattopadhyay et al. [4] proposed a method to perform domain adaptation and active learning by solving a single convex optimization problem. AADA (Active Adversarial Domain Adaptation) [35] for image based DA is a method which proposes a hybrid informativeness criterion based on the output of classifier and domain discriminator used in DANN. The criterion used in AADA for selecting a batch used is a point estimate, which might lead to redundant sample selection. We introduce a set-based informativeness criterion to select samples to be labeled. CLUE [24] is a recent concurrent work which selects samples through uncertainty-weighted clustering for Active DA.

3. Background

3.1. Definitions and Notations

Definitions: We first define a set function $f(S)$ for which input is a set S. A submodular function is a set function $f : 2^{\Omega} \to \mathbb{R}$, where 2^{Ω} is the power set of set Ω which contains all elements. The submodular functions are characterized by the property of diminishing returns i.e., addition of a new element to smaller set must produce a larger increase in f in comparison to addition to a larger set. This is mathematically stated as for every $S_1, S_2 \subseteq \Omega$ having $S_1 \subseteq S_2$ then for every $x \in \Omega \backslash S_2$ the following property holds:

$$f(S_1 \cup \{x\}) - f(S_1) \geq f(S_2 \cup \{x\}) - f(S_2) \quad (1)$$

This property is known as the *diminishing returns* property.
Notations Used: In the subsequent sections we use $h_\theta(x)$ as softmax output of the classifier, $h_\theta(x)$ is a composition of $f_\theta \circ g_\theta(x)$ where, $g_\theta(x)$ is the function that maps input to embedding and f_θ does final classification. The domain discriminator is a network $D_\phi(g_\theta(x))$ which classifies the sample into source and target domain which adversarially aligns the domains. We use $\mathcal{D}$ for combined data from both domains and use symbols of $\mathcal{D}_s$ and $\mathcal{D}_t$ for labeled data from source and target domain respectively. $\mathcal{D}_u$ denotes the unlabeled target data. In active DA, we define budget B as number of target samples selected from $\mathcal{D}_u$ and added to $\mathcal{D}_t$ in each cycle.
Active Domain Adaptation: In each cycle, we first perform DA using $\mathcal{D}_s$ and $\mathcal{D}_t$ as the source and $\mathcal{D}_u$ as the target. Active Learning techniques are then utilized to select B most informative samples from $\mathcal{D}_u$ which is then added to $\mathcal{D}_t$. This is performed for C cycles.

3.2. Cluster Assumption

Cluster assumption states that the decision boundaries should not lie in high density regions of data samples, which is a prime motivation for our approach. For enforcing cluster assumption we make use of two additional objectives from VADA [33] method. The first objective is the minimization of conditional entropy on the unlabeled target data $\mathcal{D}_u$. This is enforced by using the following loss function:

$$L_c(\theta; \mathcal{D}_u) = -\mathbb{E}_{x \sim \mathcal{D}_u}[h_\theta(x)^T \ln h_\theta(x)] \quad (2)$$

The above objective ensures the formation of clusters of target samples, as it enforces high-confidence for classification on target data. However due to large capacity of neural networks, the classification function learnt can be locally non Lipschitz which can allow for large change in function value with small change in input. This leads to unreliable estimates of the conditional entropy loss L_c. For enforcing the local Lipschitzness we use the second objective, which was originally proposed in Virtual Adversarial Training (VAT) [19]. It ensures smoothness in the ϵ norm ball enclosing the samples. The VAT objective is given below:

$$L_v(\theta; \mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[\max_{||r|| \leq \epsilon} D_{KL}(h_\theta(x) || h_\theta(x + r))] \quad (3)$$

4. Proposed Method

In Active Domain Adaptation, there are two distinct steps i.e., sample selection (i.e. sampling) followed by Domain Adaptation which we describe below:

4.1. Submodular Subset Selection

4.1.1 Virtual Adversarial Pairwise (VAP) Score

In our model architecture, we only use a linear classifier and a softmax over domain invariant features $f_\theta(x)$ for classification. Due to the linear nature, we draw inspiration from

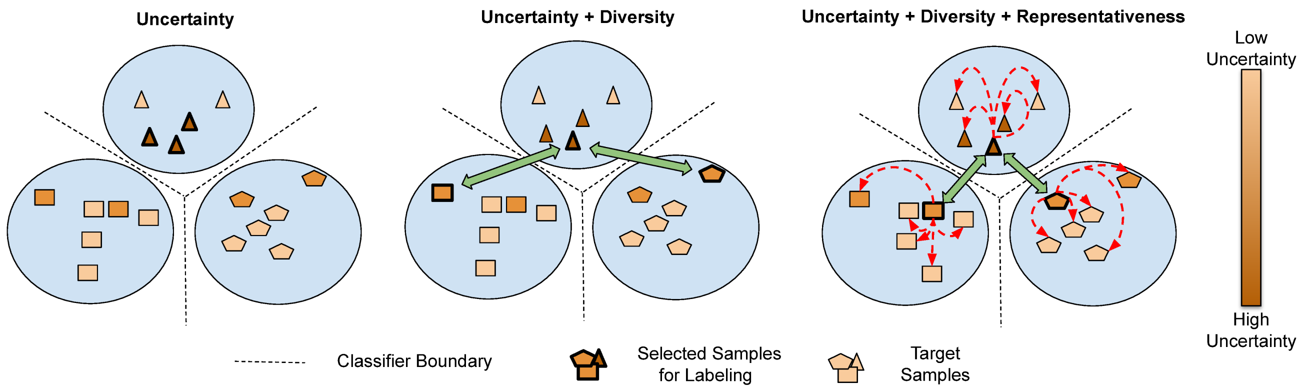


Figure 3. Our sampling technique incorporates uncertainty, diversity and representativeness. Just using uncertainty can lead to *redundant* sample selection as shown on left. Whereas, incorporating diversity to ensure a large distance between selected samples may lead to selection of *outliers*. Our sampling technique avoids outliers by selecting uncertain samples which are representative of the clusters.

SVM (Support Vector Machines) based AL methods which demonstrate that the samples near the boundary are likely to be support vectors, hence are more informative than other samples. There is also the theoretical justification behind choosing samples which are near the boundary in case of SVMs [36]. Hence we also aim to find the vectors which are near to the boundary by adversarially perturbing each sample x. We use the following objective to create perturbation:

$$\underset{||r_i||\leq\epsilon}{max}\ D_{KL}(h_\theta(x)||h_\theta(x+r_i)) \tag{4}$$

Power Method [10] is used for finding the adversarial perturbation r_i which involves initialization of random vector. As we aim to find vectors which are near the decision boundary, there can be cases where a particular sample may lie close to multiple decision boundaries as we operate in the setting of multi-class classification. Hence we create the perturbation r_i for N number of random initializations. This is done to select samples which can be easily perturbed to a diverse set of classes and also increase the reliability of the uncertainty estimate. We use the mean pairwise KL divergence score of probability distribution as a metric for measuring the uncertainty of sample. This is defined as Virtual Adversarial Pairwise (VAP) score given formally as:

$$\begin{aligned} VAP(x) = \frac{1}{N^2}\Bigg(&\sum_{i=1}^{N} D_{KL}(h_\theta(x)||h_\theta(x+r_i)) \\ &+ \sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N} D_{KL}(h_\theta(x+r_i)||h_\theta(x+r_j))\Bigg) \end{aligned} \tag{5}$$

The first term corresponds to divergence between perturbed input and the original sample x, the second term corresponds to diversity in the output of different perturbations. The approach is pictorially depicted on the right side in Fig. 2. For VAP score to be meaningful we assume that cluster assumption holds and the function is smooth, which makes VAADA a complementary DA approach to our method.

4.1.2 Diversity Score

Just using VAP score for sampling can suffer from the issue of multiple similar samples being selected from the same cluster. For selecting the diverse samples in our set S we propose to use the following diversity score for sample x_i which is not present in S.

$$d(S, x_i) = \min_{x\in S} D(x, x_i) \tag{6}$$

where D is a function of divergence. In our case we use the KL Divergence function:

$$D(x_j, x_i) = D_{KL}(h_\theta(x_j)||h_\theta(x_i)) \tag{7}$$

4.1.3 Representativeness Score

The combination of above two scores can ensure that the diverse and uncertain samples are selected. But this could still lead to selection of outliers as they can also be uncertain and diversely placed in feature space. For mitigating this we use a term based on facility location problem [41] which ensures that selected samples are placed such that they are representative of unlabeled set. The score is mathematically defined as:

$$R(S, x_i) = \sum_{x_k\in\mathcal{D}_u} \max(0, s_{ki} - \max_{x_j\in S} s_{kj}) \tag{8}$$

The s_{ij} corresponds to the similarity between sample x_i and x_j. We use the similarity function $-\ln(1 - BC(h_\theta(x_i), h_\theta(x_j))$ where $BC(p, q)$ is the Bhattacharya coefficient [2] defined as $\sum_k \sqrt{p_k q_k}$ for probability distributions p and q.

4.1.4 Combining the Three Score Functions

We define the set function $f(S)$ by defining the gain as a convex combination of $VAP(x_i)$, $d(S, x_i)$ and $R(S, x_i)$.

$$\begin{aligned} f(S \cup \{x_i\}) - f(S) = \alpha VAP(x_i) + \beta d(S, x_i) \\ + (1 - \alpha - \beta) R(S, x_i) \end{aligned} \tag{9}$$

Here $0 \leq \alpha, \beta, \alpha + \beta \leq 1$ are hyperparameters which control relative strength of uncertainty, diversity and representativeness terms. We normalize the three scores before combining them through Eq. 9.

Lemma 1: The set function $f(S)$ defined by Eq. 9 is submodular.

Lemma 2: The set function $f(S)$ defined by Eq. 9 is a non decreasing, monotone function.

We provide proof of the above lemmas in Sec. 2 of supplementary material. Overview of the overall sampling approach is present in Fig. 3.

4.1.5 Submodular Optimization

As we have shown in the previous section that the set function $f(S)$ is submodular, we aim to select the set S satisfying the following objective:

$$\underset{S:|S|=B}{max} f(S) \tag{10}$$

For obtaining the set of samples S to be selected, we use the greedy optimization procedure. We start with an empty set S and add each item iteratively. For selecting each of the sample (x_i) in the unlabeled set, we calculate the gain of each the sample $f(S \cup \{x_i\}) - f(S)$. The sample with the highest gain is then added to set S. The above iterations are done till we have exhausted our labeling budget B. The performance guarantee of the greedy algorithm is given by the following result:

Theorem 1: Let S^* be the optimal set that maximizes the objective in Eq. 10 then the solution S found by the greedy algorithm has the following guarantee (See Supp. Sec. 2):

$$f(S) \geq \left(1 - \frac{1}{e}\right) f(S^*) \tag{11}$$

Insight for Diversity Component: The optimization algorithm with $\alpha = 0$ and $\beta = 1$ degenerates to greedy version of diversity based Core-Set [31] (i.e., K-Center Greedy) sampling. Diversity functions based on similar ideas have also been explored for different applications in [12, 3]. Further details are provided in the Sec. 3 of supplementary material.

4.2. Virtual Adversarial Active Domain Adaptation

Discriminator-alignment based Unsupervised DA methods fail to effectively utilize the additional labeled data present in target domain [29]. This creates a need for modifications to existing methods which enable them to effectively use the additional labeled target data, and improve generalization on target data. In this work we introduce VAADA (Virtual Adversarial Active Domain Adaptation) which enhances VADA through modification which allow it to effectively use the labeled target data.

We have given our subset selection procedure to select samples to label (i.e., $\mathcal{D}_t$) in Algo. 1 and in Fig. 2. For aligning the features of labeled ($\mathcal{D}_s \cup \mathcal{D}_t$) with $\mathcal{D}_u$, we make use of domain adversarial training (DANN) loss functions given below:

$$L_y(\theta; \mathcal{D}_s, \mathcal{D}_t) = \mathbb{E}_{(x,y)\sim(\mathcal{D}_s \cup \mathcal{D}_t)}[y^T \ \ln h_\theta(x)] \tag{12}$$

$$\begin{aligned} L_d(\theta; \mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_u) = \underset{D_\phi}{sup}\, \mathbb{E}_{x\sim\mathcal{D}_s\cup\mathcal{D}_t}[\ln D_\phi(f_\theta(x))] \\ + \mathbb{E}_{x\sim\mathcal{D}_u}[\ln(1 - D_\phi(f_\theta(x)))] \end{aligned} \tag{13}$$

As our sampling technique is based on cluster assumption, for enforcing it we add the Conditional Entropy Loss defined in Eq. 2. Additionally, for enforcing Lipschitz continuity by Virtual Adversarial Training, we use the loss defined in Eq. 3. The final loss is obtained as:

$$\begin{aligned} L(\theta; \mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_u) = L_y(\theta; \mathcal{D}_s, \mathcal{D}_t) + \\ \lambda_d L_d(\theta; \mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_u) \ + \lambda_s L_v(\theta; \mathcal{D}_s \cup \mathcal{D}_t) + \\ \lambda_t(L_v(\theta; \mathcal{D}_u) + L_c(\theta; \mathcal{D}_u)) \end{aligned} \tag{14}$$

The λ-values used are the *same for all our experiments* and are mentioned in the Sec. 5 of supplementary material.

Differences between VADA and VAADA: We make certain important changes to VADA listed below, which enables VADA [33] to effectively utilize the additional supervision of labeled target data and for VAADA procedure:

1) High Learning Rate for All Layers: In VAADA, we use the same learning rate for all layers. In a plethora of DA methods [29, 18] a lower learning rate for initial layers is used to achieve the best performance. We find that although this practice helps for Unsupervised DA it hurts the Active DA performance (experimentally shown in Sec. 5 of supplementary material).

2) Using Gradient Clipping in place of Exponential Moving Average (EMA): We use gradient clipping for all network weights to stabilize training whereas VADA uses EMA for the same. We find that clipping makes training of VAADA stable in comparison to VADA and achieves a significant performance increase over VADA.

We find VAADA is able to work robustly across diverse datasets. It has been shown in [29] that VADA, when used out of the box, is unable to get gains in performance when used in setting where target labels are also available for training. This also agrees with our observation that VAADA significantly outperforms VADA in Active DA scenario's (demonstrated in Fig. 10, with additional analysis in Sec. 5 of supplementary material).

Algorithm 1: S[3]VAADA: Submodular Subset Selection for Virtual Adversarial Active Domain Adaptation

Require: Labeled source $\mathcal{D}_s$; Unlabeled target $\mathcal{D}_u$; Labeled target $\mathcal{D}_t$; Budget per cycle B; Cycles C; Model with parameters θ; Parameters α, β

Ensure: Updated model parameters with improved generalization ability on target domain

```
Train the model according to Eq. 14
for cycle ← 1 to C do
    S ← ∅
    for iter ← 1 to B do
        x* = argmax_{x ∈ D_u \ S} f(S ∪ {x}) − f(S)
        S ← S ∪ {x*}
    end for
    Get ground truth labels l_S for samples in S from oracle
    D_t ← D_t ∪ (S, l_S)
    D_u ← D_u \ S
    Train the model according to Eq. 14
end for
```

5. Experiments

5.1. Datasets

We perform experiments across multiple source and target domain pairs belonging to 3 diverse datasets, namely Office-31 [26], Office-Home [39], and VisDA-18 [23]. We have specifically not chosen any DA task using real world domain as in those cases the performance maybe higher due to ImageNet initialization not due to adaptation techniques. In **Office-31** dataset, we evaluate the performance of various sampling techniques on DSLR $\rightarrow$ Amazon and Webcam $\rightarrow$ Amazon, having 31 classes. The **Office-Home** consists of 65 classes and has 4 different domains belonging to Art, Clipart, Product and Real World. We perform the active domain adaptation on Art $\rightarrow$ Clipart, Art $\rightarrow$ Product and Product $\rightarrow$ Clipart. **VisDA-18** image classification dataset consists of two domains (synthetic and real) with 12 classes in each domain.

5.2. Experimental Setup

Following the common practice in AL literature, we first split the target dataset into train set and validation set with 80% data being used for training and 20% for validation. In all the experiments, we set budget size B as 2% of the number of images in the target training set, and we perform five cycles ($C = 5$) of sampling in total. At the end of all cycles, 10% of the labeled target data will be used for training the model. This setting is chosen considering the practicality of having a small budget of labeling in the target domain

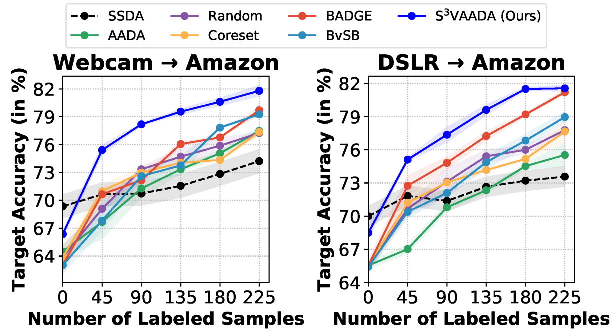


Figure 4. Active DA target accuracy on two adaptation tasks from Office-31 dataset. S[3]VAADA consistently outperforms BADGE [1], AADA [35] and SSDA (MME* [27]) techniques.

and having access to unlabeled data from the target domain. We use ResNet-50 [11] as the feature extractor g_θ, which is initialized with weights pretrained on ImageNet. We use SGD Optimizer with a learning rate of 0.01 and momentum (0.9) for VAADA training. For the DANN experiments, we follow the same architecture and training procedure as described in [18]. In all experiments, we set α as 0.5 and β as 0.3. We use PyTorch [21] with automatic mixed-precision training for our implementation. Further experimental details are described in the Sec. 6 of supplementary material. We report the mean and standard error of the accuracy of the 3 runs with different random seeds.

In AADA [35] implementation, the authors have used a different architecture and learning schedule for DANN which makes comparison of increase in performance due to Active DA, over unsupervised DA intricate. In contrast we use ResNet-50 architecture and learning rate schedule of DANN used in many works [18, 6]. We first train DANN to reach optimal performance on Unsupervised Domain Adaptation and then start the active DA process. This is done as the practical utility of Active DA is the performance increase over Unsupervised DA.

5.3. Baselines

It has been shown by Su et al. [35] that for active DA performing adversarial adaptation through DANN, with adding newly labeled target data into source pool works better than fine-tuning. Hence, we use DANN for all the AL baselines described below:

1. **AADA** (Importance weighted sampling) [35]: This state-of-the-art active DA method incorporates uncertainty by calculating entropy and incorporates diversity by using the output of the discriminator.
2. **BvSB** (Best vs Second Best a.k.a. margin) [13]: It uses the difference between the probabilities of the highest and second-highest class prediction as the the metric of uncertainty, on which low value indicates high uncertainty.

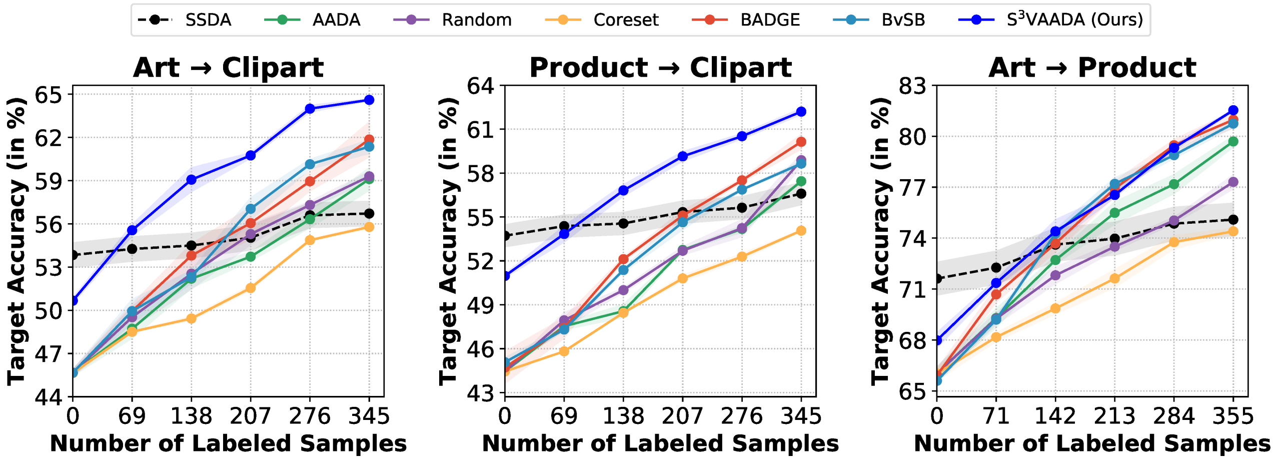


Figure 5. Active DA performance on three different Office-Home domain shifts. We see a significant improvement through S^{3}VAADA in two difficult adaptation tasks of Art → Clipart (left) and Product → Clipart (middle) .

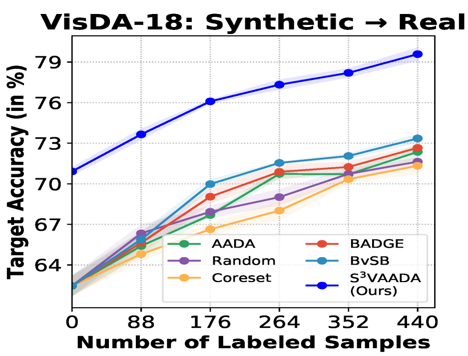


Figure 6. Active DA Results on VisDA-18 dataset.

3. **BADGE** [1]: BADGE incorporates uncertainty and diversity by using the gradient embedding on which k-MEANS++ [38] algorithm is used to select diverse samples. BADGE method is currently one of the state-of-the-art methods for AL.
4. K**-Center (Core-Set)** [30]: Core-Set selects samples such that the area of coverage is maximized. We use greedy version of Core-Set on the feature space (g_θ). It is a diversity-based sampling technique.
5. **Random**: Samples are selected randomly from the pool of unlabeled target data.

Semi-Supervised DA: We compare our method against recent method of MME* [27] with ResNet-50 backbone on Office datasets, using the author's implementation[1]. In each cycle target samples are randomly selected, labeled and provided to the MME* method for DA.

5.4. Results

Fig. 4 shows the results on Office-31 dataset, S^{3}VAADA outperforms all other techniques. On Webcam → Amazon shift, it shows significant improvement of 9% in the target accuracy with just 45 labeled samples. S^{3}VAADA gets 81.8% accuracy in the last cycle which is around 15% more than the unsupervised DA performance, by using just 10% of the labeled data. On DSLR → Amazon shift, VAADA follows a similar trend and performs better than all other sampling techniques. On Office-Home dataset on the harder domain shifts of Art → Clipart and Product → Clipart, S^{3}VAADA produces a significant increase of 3%-5% and 2%-6% respectively across cycles, in comparison to other methods (Fig. 5). On the easier Art → Product shift, our results are comparable to other methods.

Large Datasets: On the VisDA-18 dataset, where the AADA method is shown to be ineffective [35] due to a severe domain shift. Our method (Fig. 6) is able to achieve significant increase of around 7% averaged across all cycles, even in this challenging scenario. For demonstrating the scalability of our method to DomainNet [22], we also provide the results of one adaptation scenario in Sec. 10 of supplementary material.

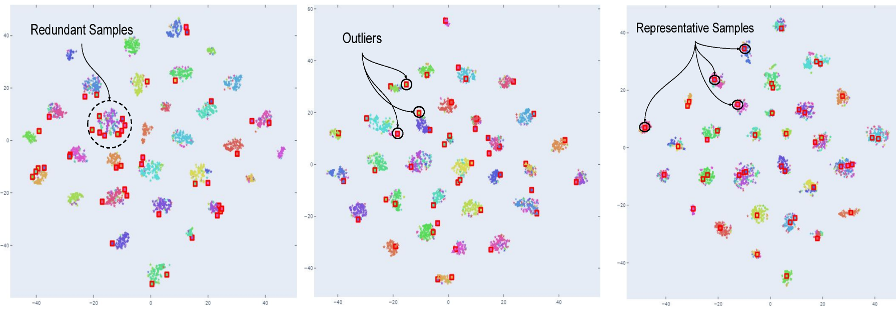
Semi-Supervised DA: From Figs. 4 and 5 it is observed that performance of MME* saturates as more labeled data is added, in contrast, S^{3}VAADA continues to improve as more target labeled data is added.

6. Analysis of S^{3}VAADA

Visualization: Fig. 7 shows the analysis of samples selected for uncertainty **U**, diversity **D** and representativeness **R** criterion, which depicts the *complementary* preferences of the three criterion.

Sensitivity to α and β : Fig. 8 shows experiments for probing the effectiveness of each component (i.e., uncertainty, diversity and representativeness) in the information criteria. We find that just using Uncertainty ($\alpha = 1$) and Diversity

[1] https://github.com/VisionLearningGroup/SSDA_MME



(a) Uncertainty (**U**) (b) Diversity (**D**) (c) Representativeness (**R**)

Figure 7. Feature space visualization using t-SNE with selected samples in Red. Using **Uncertainty** leads to *redundant* samples from same cluster, whereas using **Diversity** leads to only diverse boundary samples being selected which maybe *outliers*. Sampling using **Representativeness** prefers samples near the cluster center, hence we use a combination of these complementary criterion as our criterion.

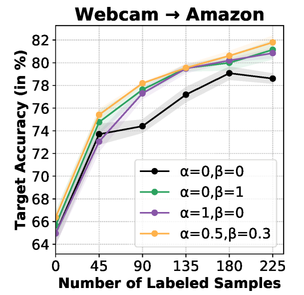


Figure 8. Trade off between Uncertainty, Diversity and Representativeness (i.e., Parameter sensitivity to α, β).

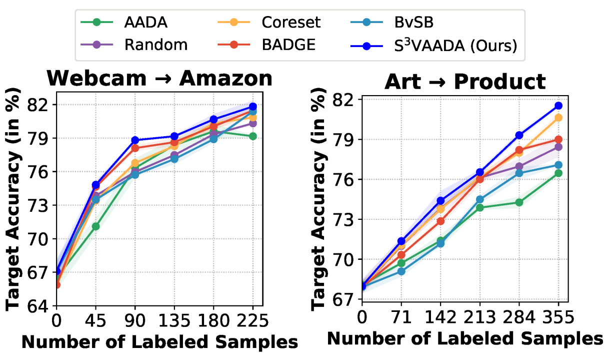


Figure 9. Ablation on sampling methods on different domain shifts. In both cases, we train the sampling techniques via VAADA. Our method (S^{3}VAADA) consistently outperforms all other sampling methods.

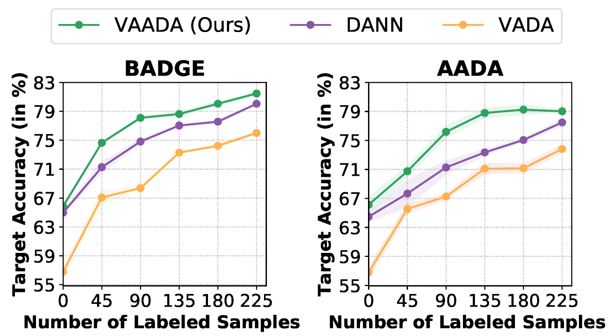


Figure 10. Comparison of different DA methods for Active DA on Webcam $\rightarrow$ Amazon.

($\beta = 1$) provide reasonable results when used individually. However, the individual performance remain sub-par with the hybrid combination (i.e., $\alpha = 0.5$, $\beta = 0.3$). We use value of $\alpha = 0.5$ and $\beta = 0.3$ across all our experiments, hence our sampling does not require parameter-tuning specific to each dataset.

Comparison of Sampling Methods: For comparing the different sampling procedures we fix the adaptation technique to VAADA and use different sampling techniques. Fig. 9 shows that our sampling method outperforms others in both cases. In general we find that hybrid approaches i.e., Ours and BADGE perform *robustly* across domain shifts.

Comparison of VAADA: Fig. 10 shows performance of VAADA, DANN and VADA when used as adaptation procedure for two sampling techniques. We find that a significant improvement occurs for all the sampling techniques in each cycle for VAADA comparison to DANN and VADA. The $\geq 5\%$ improvement in each cycle, shows the importance of proposed improvements in VAADA over VADA.

We provide additional analysis on *convergence, budget size and hyper parameters* in the Sec. 4 of supplementary material. Across our analysis we find that S^{3}VAADA works robustly in different DA scenario.

7. Conclusion

We formulate the sample selection in Active DA as optimal informative subset selection problem, for which we propose a novel submodular information criteria. The information criteria takes into account the uncertainty, diversity and representativeness of the subset. The most informative subset obtained through submodular optimization is then labeled and used by the proposed adaptation procedure VAADA. We find that the optimization changes introduced for VAADA significantly improve Active DA performance across all sampling schemes. The above combination of sampling and adaptation procedure constitutes S^{3}VAADA, which consistently provides improved results over existing methods of Semi-Supervised DA and Active DA.

Acknowledgements: This work was supported in part by PMRF Fellowship (Harsh), SERB (Project:STR/2020/000128) and UAY (Project:UAY, IISC_010) MHRD, Govt. of India.

References

[1] 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*, 2020. 3, 6, 7

[2] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946. 4

[3] Shayok Chakraborty, Vineeth Balasubramanian, and Sethuraman Panchanathan. Adaptive batch mode active learning. *IEEE transactions on neural networks and learning systems*, 26(8):1747–1760, 2014. 5

[4] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *International conference on machine learning*, pages 253–261. PMLR, 2013. 3

[5] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai. Adversarial-learned loss for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3521–3528, 2020. 1

[6] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai. Adversarial-learned loss for domain adaptation. *arXiv*, abs/2001.01046, 2020. 6

[7] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994. 1

[8] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9944–9953, 2019. 2

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1, 2

[10] H. Goluba and Henk A. van der Vorstb. Eigenvalue computation in the 20 th century gene. 2000. 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[12] KJ Joseph, Krishnakant Singh, Vineeth N Balasubramanian, et al. Submodular batch selection for training deep neural networks. *arXiv preprint arXiv:1906.08771*, 2019. 5

[13] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009. 6

[14] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, Rahul M V, and R. Venkatesh Babu. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[15] Jogendra Nath Kundu, Naveen Venkat, Rahul M V, and R. Venkatesh Babu. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[16] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 91–100, 2019. 2

[17] Da Li and Timothy Hospedales. Online meta-learning for multi-source and semi-supervised domain adaptation. In *European Conference on Computer Vision*, pages 382–403. Springer, 2020. 2

[18] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1645–1655, 2018. 1, 5, 6

[19] T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. 3

[20] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978. 2

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[22] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 7

[23] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2102–21025, 2018. 6

[24] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings, 2020. 3

[25] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010. 2, 3

[26] 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. 1, 6

[27] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International

Conference on Computer Vision, pages 8050–8058, 2019. 1, 2, 6, 7

[28] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations*, 2018. 1

[29] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2, 5

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

[31] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016. 5

[32] Burr Settles. Active learning literature survey. 2009. 1

[33] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018. 2, 3, 5

[34] 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. 3

[35] 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 (WACV)*, March 2020. 1, 2, 3, 6, 7

[36] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001. 4

[37] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 1

[38] Sergei Vassilvitskii and David Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006. 7

[39] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[40] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014. 3

[41] Wikipedia contributors. Facility location problem — Wikipedia, the free encyclopedia, 2021. [Online; accessed 17-March-2021]. 4

[42] 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. 3