Multi-Classifier Adversarial Optimization for Active Learning

Lin Geng, Ningzhong Liu*, and Jie Qin*

School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
{lingeng, ningzhongliu, jie.qin}@nuaa.edu.cn

Abstract

Active learning (AL) aims to find a better trade-off between labeling costs and model performance by consciously selecting more informative samples to label. Recently, adversarial approaches have emerged as effective solutions. Most of them leverage generative adversarial networks to align feature distributions of labeled and unlabeled data, upon which discriminators are trained to better distinguish between them. However, these methods fail to consider the relationship between unlabeled samples and decision boundaries, and their training processes are often complex and unstable. To this end, this paper proposes a novel adversarial AL method, namely multi-classifier adversarial optimization for active learning (MAOAL). MAOAL employs task-specific decision boundaries for data alignment while selecting the most informative samples to label. To fulfill this, we introduce a novel classifier class confusion (C^3) metric, which represents the classifier discrepancy as the inter-class correlation of classifier outputs. Without any additional hyper-parameters, the C^3 metric further reduces the negative impacts of ambiguous samples in the process of distribution alignment and sample selection. More concretely, the network is trained adversarially by adding two auxiliary classifiers, reducing the distribution bias of labeled and unlabeled samples by minimizing the C^3 loss between classifiers, while learning tighter decision boundaries and highlighting hard samples by maximizing the C^3 loss. Finally, the unlabeled samples with the highest C^3 loss are selected to label. Extensive experiments demonstrate the superiority of our approach over state-of-the-art AL methods in terms of image classification and object detection.

Introduction

Over the past decade, the emergence of large-scale annotated datasets (Deng et al. 2009) and the development of deep learning techniques have brought prosperity to the field of computer vision (Badrinarayanan, Kendall, and Cipolla 2017; Li et al. 2021; He et al. 2016). However, high-quality annotations are both time- and resource-consuming, hindering the application of convolutional neural networks (CNNs) in realistic scenarios. This dilemma between performance and costs has given rise to active learning (AL) (Settles 2009), which aims to maximize the model performance with a limited annotation budget, by choosing the most suitable samples to annotate from a large unlabeled dataset.

Conventional AL approaches can be categorized into query synthesis or pool-based methods. In query synthesis methods, the most informative samples are selected using generative models (Mayer and Timofte 2020; Zhu and Bento 2017). Pool-based methods can be further divided into uncertainty-based methods (Ducoffe and Precioso 2018; Houlsby et al. 2011), representation-based methods (Sener and Savarese 2018; Caramalau, Bhattarai, and Kim 2021), and the combination of both (Kuo et al. 2018; Liu and Ferrari 2017). More recently, several pool-based methods (Sinha, Ebrahimi, and Darrell 2019; Kim et al. 2021) have also resorted to generative models (*e.g.*, variational auto-encoders (VAEs) (Kingma and Welling 2014) and generative adversarial networks (GANs) (Goodfellow et al. 2014)) to select uncertain samples for annotation. For instance, (Sinha, Ebrahimi, and Darrell 2019) trained the VAE and discriminator using an adversarial approach, where the VAE generates latent space representations of labeled and unlabeled data, and the discriminator is utilized as a binary classifier to determine the uncertainty of the input samples. These adversarial methods are effective in aligning the feature distributions of labeled and unlabeled data. Nevertheless, on the one hand, class discriminability cannot be guaranteed in the adversarial learning process; on the other hand, adversarial training is very sensitive to the selection of hyper-parameters between discriminator and generator (Berthelot, Schumm, and Metz 2017), leading to a complex and unstable training process. Therefore, it is highly desirable to develop an active learning method that could inherit the merits of the adversarial spirit but circumvent the tedious training procedure.

Inspired by the recent success of the methods exploiting classifier discrepancies (Fu et al. 2021; Cho et al. 2022; Saito et al. 2018), in this paper, we propose a novel, easy-to-train yet effective adversarial active learning approach, namely multi-classifier adversarial optimization for active learning (MAOAL). In particular, MAOAL plays the min-max game between the feature generator G, the main classifier C, and two auxiliary classifiers C_1 and C_2, with the aim of minimizing/maximizing the class-level discrepancy. As illustrated in Fig. 2, by fixing C_1 and C_2, we update G and C to learn more discriminative features, ensuring the consistency of the distribution between labeled and unlabeled samples. Ad-

*Corresponding authors.

versarially, C_1 and C_2 are optimized while fixing G and C to learn tighter decision boundaries to highlight informative unlabeled samples far from the labeled distribution.

It is worth noting that ADS (Fu et al. 2021) also mines the discrepancy between two adversarial classifiers; however, it defines the classifiers' discrepancy as the l_1 distance between predictions, which only evaluates the difference between the classifiers' predictions with respect to the same class while ignoring the correlation between different classes. As a result, confusing features on unlabeled data may be generated, thus decreasing the feature discriminability and the model accuracy. In this work, we address the above shortcoming by further considering the correlation between different classes through a delicately-designed classifier class confusion (C^3) metric. Specifically, the C^3 metric measures the confusion level between the predictions obtained by different classifiers. Without adding additional hyper-parameters, it is simply reflected in a correlation matrix and efficiently computed from well-calibrated bi-classifier predictions. In addition, we add a main classifier whose decision hyperplane is between the two auxiliary classifiers, enlarging the distance between the support vectors and the decision boundary. To summarize, the proposed MAOAL trains triple classifiers by optimizing the proposed C^3 loss in an adversarial fashion, aligning the feature distribution of labeled and unlabeled data as well as learning tighter decision boundaries for informative sample acquisition.

The main contributions of this paper are three-fold:

- We propose multi-classifier adversarial optimization for active learning (MAOAL), which trains multiple classifiers in an adversarial manner to consider tighter inter-class decision boundaries in the process of aligning labeled and unlabeled feature distributions.
- We propose a novel classifier discrepancy metric, namely classifier class confusion (C^3), to enhance classifier determinacy and prediction diversity by adversarially optimizing the C^3 loss while pushing ambiguous samples near the decision boundary to effectively drive the sample selection process.
- Extensive experiments on both image classification and object detection tasks demonstrate that our approach consistently outperforms state-of-the-art active learning methods.

Related Work

This paper focuses on pool-based active learning scenarios (Haussmann et al. 2020; Yuan et al. 2021). Pool-based methods typically employ the current model to predict each unannotated data point to obtain a ranking metric of the informativeness for each sample on the unlabeled dataset, and then choose the top-N samples based on this metric to annotate by the oracle(s). Existing methods can be roughly divided into three categories: uncertainty-based methods, representation-based methods, and hybrid methods that combine the two.

Uncertainty-based Methods. Intuitively, the predictive uncertainty of the model reflects the informativeness of data samples which can be estimated with different methods, such as (Ebrahimi et al. 2020; Gorriz et al. 2017) based on probabilistic models, (Joshi, Porikli, and Papanikolopoulos 2009; MacKay 1992) based on information entropy, and (Brinker 2003) by measuring the distance between samples and the decision boundary. In more recent works, (Gal, Islam, and Ghahramani 2017) proposed to use dropout layers to estimate the uncertainty of a neural network's prediction for sample query. (Yoo and Kweon 2019) presented a learning loss prediction module to estimate the loss of unlabeled data to track uncertainty.

Representation-based Methods. Representation-based methods (Yang et al. 2015; Caramalau, Bhattarai, and Kim 2021) try to select a set of diverse samples that can represent the entire dataset well. (Gissin and Shalev-Shwartz 2019) proposed the discriminative active learning (DAL) to train a binary classifier to discriminate between labeled and unlabeled samples so as to select the most representative sample. Core-set (Sener and Savarese 2018) was a typical representation-based approach, which selected the samples based on the core-set distance of intermediate features.

Hybrid Methods. Hybrid methods (Wang et al. 2016; Agarwal et al. 2020; Liu and Ferrari 2017) combine uncertainty with representation. For instance, BatchBALD (Kirsch 2019) increased the diversity of the selected data by a traceable approximate mutual information sampling method. Badge (Ash et al. 2020) inherited the Core-set (Sener and Savarese 2018) method and combined it with the Bald (Houlsby et al. 2011) and experimented on several models.

Adversarial Methods. In the recent literature, adversarial active learning (Sinha, Ebrahimi, and Darrell 2019; Zhang et al. 2020; Wang et al. 2020; Kim et al. 2021) has trained a generative adversarial network (GAN) (Goodfellow et al. 2014) structured auxiliary networks that introduced variational auto-encoders (VAEs) (Kingma and Welling 2014) to learn a low-dimensional latent space and discriminate the labeled and unlabeled samples to select the unlabeled data most different from the labeled ones. However, these AL methods brought an unstable training process and additional computational costs.

Our method absorbs some ingredients from ADS (Fu et al. 2021) and MCDAL (Cho et al. 2022). However, it is noteworthy that MAOAL significantly differs from these two works from the following perspectives. First, compared to ADS, we propose a novel discrepancy metric C^3 replacing the l_1 distance to potentially discover more ambiguous unlabeled data near the decision boundary. Then, besides using two classifiers, we add a main classifier to increase the distance between the uncertain unlabeled samples and the labeled ones, facilitating the sample acquisition process. Second, unlike MCDAL that only exploits the classifier discrepancy in the sample acquisition stage, we adversarially optimize the C^3 metric and take full advantage of unlabeled data in the whole training process while simultaneously considering the determinacy of the classifier and the distributional alignment between labeled and unlabeled data.

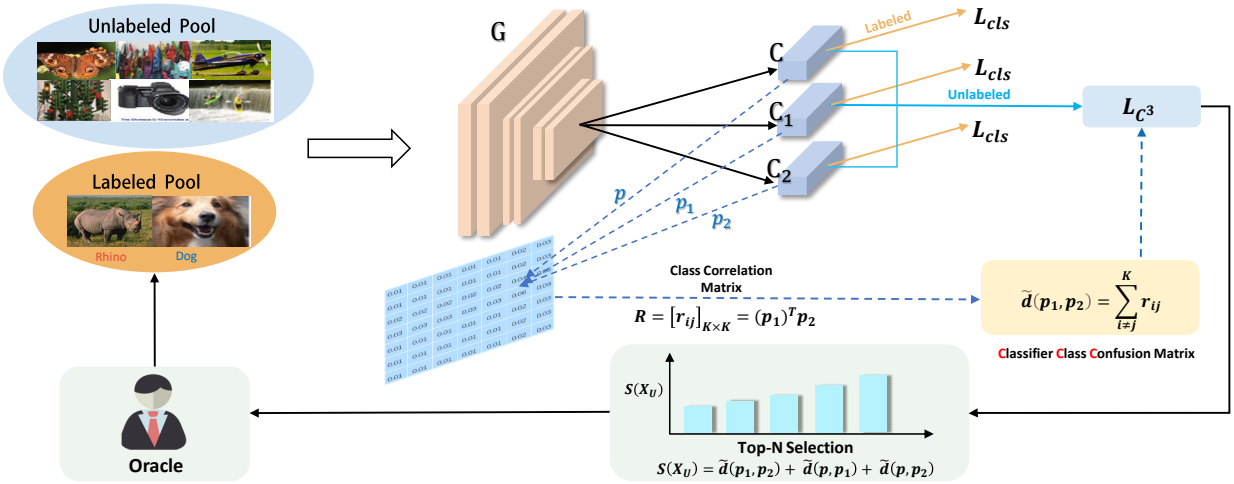


Figure 1: The overall framework of the proposed MAOAL. First, we construct a class correlation matrix from classifier outputs to obtain a novel metric w.r.t. the discrepancy between classifiers, namely the classifier class confusion (C^3) metric. Then, the model adversarially optimizes the C^3 loss to align labeled and unlabeled data while generating a larger sampling region. Finally, we design the sample acquisition function to further select the most informative samples.

Proposed Method

Overview

In this section, we consider a typical pool-based AL scenario. Let (X_L, Y_L) denote the pool of labeled data, and X_U denote a large pool of unlabeled data; the goal is to select the most informative samples from the unlabeled pool X_U by the sample acquisition function $S(x)$. These selected samples can be evaluated as the ones with the most significant performance gain on the main task model F when labeled. Specifically, in the i-th active learning iteration, we select N samples from X_U^i and move them to the labeled pool (X_L^i, Y_L^i) after annotation, and then update the unlabeled and labeled pool to train and evaluate F. Iterations are repeated until the model performance satisfies customer requirements, or the annotated budget is exhausted.

Next, we introduce the proposed MAOAL in detail. Fig. 1 illustrates the overall framework. The network architecture is composed of a feature generator G, which accepts input from X_U and X_L, a main classifier C, and two auxiliary classifiers C_1 and C_2, that take features from G. We introduce a novel classifier class confusion (C^3) metric to measure the discrepancy between classifiers, by training labeled and unlabeled data in an adversarial manner, and perform sample acquisition by using the proposed C^3 metric.

Classifier Class Confusion

The two auxiliary classifiers C_1 and C_2 output a K-dimensional vector of logits. Then the Softmax function is used to obtain class probabilities through the vector:

$$p_1 = \text{softmax}(C_1(G(x_U))), \tag{1}$$

$$p_2 = \text{softmax}(C_2(G(x_U))), \tag{2}$$

$$\sum_{j=1}^{K} p_i^j = 1, \quad \forall i \in 1, 2, \tag{3}$$

where $p_1, p_2 \in \mathbb{R}^{1\times K}$ are the K-dimensional probabilistic outputs of C_1, C_2, and K is the number of possible categories. Previous methods (Fu et al. 2021; Cho et al. 2022) take the absolute value of the difference between the two classifiers' probabilistic outputs (*i.e.*, l_1 distance) as the discrepancy between two classifiers. However, the l_1 distance only considers the classifier similarity on the same class and ignores the correlation between different classes. For example, when the outputs of classifiers C_1, C_2 are $p_1 = [0.33, 0.33, 0.34]$ and $p_2 = [0.33, 0.33, 0.34]$, respectively, though the l_1 distance between them is equal to zero, such predictions are prone to confusion between different classes. Motivated by the calculation process of the self-correlation matrix (Jin et al. 2020), we find that the class correlation of two classifiers' predictions for the same instance can be naturally represented by the product between one classifier's prediction and the other's transposition. Thereby, we define the classifier correlation matrix $R \in \mathbb{R}^{K\times K}$ between two classifiers as:

$$\begin{aligned} R = [r_{ij}]_{K\times K} &= (p_1)^T p_2 \\ &= \begin{bmatrix} p_1^1 \\ p_1^2 \\ \vdots \\ p_1^K \end{bmatrix} \left[p_2^1, p_2^2, \cdots, p_2^K\right], \end{aligned} \tag{4}$$

where $r_{ij} = p_1^i p_2^j$, $i, j = 1, 2, \cdots, K$, is the element in the i-th row and j-th column of R. For the matrix R, the main diagonal element indicates the intra-class correlation, which is the probability product of two classifiers assigning a sample to the same class; the off-diagonal element indicates the inter-class correlation or confusion, which is the probability product of the same instance being divided into different categories by two classifiers. For convenience, we define the overall intra-class correlation as I_a and the overall

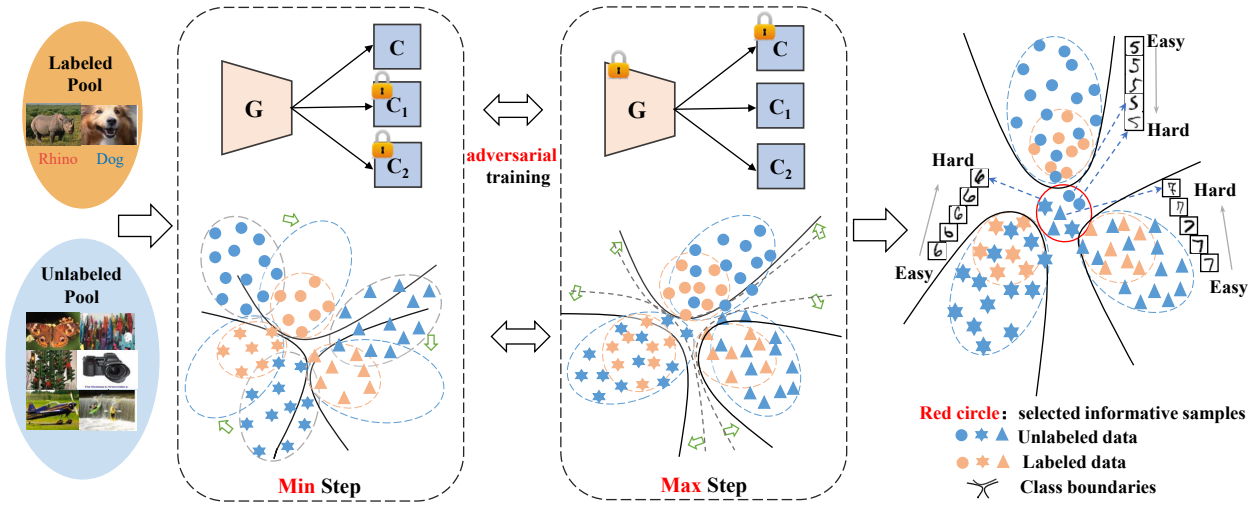


Figure 2: Adversarial training steps of our method. The generator G learns to minimize the C^3 loss in the Min step (fixing C_1 and C_2), and the classifiers learn to maximize the C^3 loss in the Max step (fixing G and C).

inter-class correlation as I_e, respectively:

$$I_a = \sum_{i=j=1}^{K} r_{ij};\ I_e = \sum_{i \neq j}^{K} r_{ij}. \tag{5}$$

According to Eq. (3), I_a and I_e satisfy $I_a + I_e = 1$. For the unlabeled dataset, the predictions usually yield a relatively small I_a and a large I_e due to the lack of supervised training. I_e can be seen as containing all probabilities that the two classifiers have inconsistent predictions, so it can be used to measure the prediction discrepancy between the two classifiers. Thus, we define the classifier class confusion (C^3) metric as follows:

$$\tilde{d}(p_1, p_2) = I_e. \tag{6}$$

Notably, the C^3 metric can be implemented with one line of code and has no additional hyper-parameters. Moreover, the computational complexity of C^3 is the same as the l_1 distance; hence, it can be efficiently implemented.

Multi-Classifier Adversarial Optimization

We devise our multi-classifier adversarial optimization algorithm based on the proposed C^3 metric.

For the network training, we first train the feature generator network G and all three classifiers C, C_1, and C_2 on the labeled dataset by optimizing the multi-class cross-entropy loss. Let θ_G and $\theta_C/\theta_{C_1}/\theta_{C_2}$ denote the parameters of the generator G and the classifiers $C/C_1/C_2$, respectively, the objective function is given as follows:

$$\min_{\theta_C, \theta_{C_1}, \theta_{C_2}, \theta_G} L_{cls}(X_L, Y_L), \tag{7}$$

$$L_{cls} = \mathbb{E}_{(x_L, y_L) \in (X_L, Y_L)} \left[-\sum_{k=1}^{K} \mathbb{1}[k = y_L] log p^k(y|x_L) \right], \tag{8}$$

where $p^k(y|x)$ denotes a probability element of prediction output p for class k, $\mathbb{1}$ is an indicator that equals 1 if a statement is true and 0 otherwise.

Due to the distribution divergence between labeled and unlabeled datasets, it is difficult for the model trained on the labeled set to classify unlabeled samples directly. To take full advantage of the unlabeled samples, in the following, we use an adversarial method to train two auxiliary classifiers C_1, C_2 on unlabeled data, extending their distance to separate them from the original class boundaries while aligning the feature distributions of labeled and unlabeled data. To achieve this, we use the designed C^3 loss to measure the classifier discrepancy as an adversarial loss, as follows:

$$L_{adv} = \mathbb{E}_{x_U \in X_U} \left[L_{C^3} \right], \tag{9}$$

$$L_{C^3} = \tilde{d}(p_1, p_2) + \tilde{d}(p, p_1) + \tilde{d}(p, p_2), \tag{10}$$

where $p/p_1/p_2$ are the probabilistic outputs of $C/C_1/C_2$, respectively. We play the min-max game based on the following objective function:

$$\min_{\theta_G, \theta_C} \max_{\theta_{C_1}, \theta_{C_2}} L_{adv}(X_U). \tag{11}$$

Min Step. By minimizing the C^3 loss of the classifiers, the obtained features of unlabeled data can have strong discriminability, which increases the consistency of the distribution between labeled and unlabeled samples. Specifically, we train the feature generator G and classifier C to minimize the adversarial loss of the fixed classifiers C_1 and C_2 as follows:

$$\min_{\theta_G, \theta_C} L_{adv}(X_U). \tag{12}$$

Max Step. We maximize the C^3 loss of the classifiers to learn tighter decision boundaries, while highlighting informative unlabeled samples that are uncertain and far from the labeled distribution. Specifically, we train the classifiers C_1,

C_2 for the fixed generator G and classifier C. The objective function is given as follows:

$$\max_{\theta_{C_1},\theta_{C_2}} L_{adv}(X_U). \tag{13}$$

By minimizing the C^3 loss to train the feature extractor G and classifier C, the negative impact of ambiguous samples on feature learning is avoided while aligning labeled and unlabeled data in the feature space as much as possible. By maximizing the C^3 loss, we adapt the two auxiliary classifiers to form a tighter decision boundary while highlighting the information-rich unlabeled samples, as illustrated in Fig. 2.

Sample Acquisition

After several iterations of training, the decision hyperplane of the main classifier C lies between those of C_1 and C_2, which makes the distance between the support vectors and the decision boundary larger. The two classifiers C_1 and C_2 have a tighter decision boundary that results in a larger region in the feature space, which we refer to as the sampling region. The C^3 loss between the outputs of classifiers is larger, *i.e.*, the samples with large prediction discrepancy are located in this sampling region, and these samples are both uncertain and far from the labeled distribution. Consequently, labeling these samples improves the model performance the most. Therefore, we define the sample acquisition function as follows:

$$S(X_U) = \tilde{d}(p_1, p_2) + \tilde{d}(p, p_1) + \tilde{d}(p, p_2). \tag{14}$$

We employ the sample acquisition function to quantify how informative each sample is, and the top-N samples are selected to be labeled by the oracle(s). We summarize the above training process of MAOAL in Algorithm 1.

Algorithm 1: The training process of multi-classifier adversarial optimization for active learning (MAOAL)

Input: Labeled pool (X_L, Y_L), Unlabeled pool X_U.
Parameter: Network parameters θ_G, classifiers' parameters θ_C, θ_{C_1} and θ_{C_2}.

for $iteration$ **do**
  for $epoch$ **do**
    if $epoch == 0$ **then**
      Train G, C, C_1, C_2 on (X_L, Y_L) using Eq. (7);
    end if
    Train G, C on X_U using Eq. (12);
    Train C_1, C_2 on X_U using Eq. (13);
    Train G, C, C_1, C_2 on (X_L, Y_L) using Eq. (7);
  end for
  Select samples using Eq. (14);
  Update (X_L, Y_L) and X_U.
end for

Experiments

We evaluate MAOAL against various state-of-the-art AL methods with respect to two computer vision tasks, *i.e.*, image classification and object detection, on four benchmark datasets. To verify the performance of each active learning method, we report the averaged results based on three runs.

Image Classification

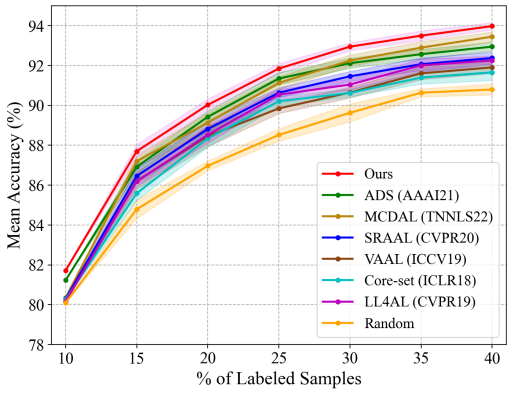
Datasets. For image classification, we evaluate our method on three classical datasets, including CIFAR-10, CIFAR-100 (Krizhevsky 2009), and Caltech-101 (Li Fei-Fei, Fergus, and Perona 2006). Both CIFAR-10 and CIFAR-100 contain 60,000 images of 32x32x3 pixels, with 50,000 images for training and 10,000 for testing. CIFAR-10 contains 10 classes with 6,000 images per class, while CIFAR-100 has 100 classes with 600 images per class. Caltech-101 consists of 9146 images divided into 101 categories, with about 40 to 800 images in each class.

Compared Methods. We compare MAOAL with state-of-the-art approaches including Core-set (Sener and Savarese 2018), LL4AL (Yoo and Kweon 2019), VAAL (Sinha, Ebrahimi, and Darrell 2019), SRAAL (Zhang et al. 2020), ADS (Fu et al. 2021)and MCDAL (Cho et al. 2022). Several previous works (Sinha, Ebrahimi, and Darrell 2019; Yoo and Kweon 2019) have shown that classical methods, such as (Gal and Ghahramani 2016; Gal, Islam, and Ghahramani 2017; Beluch et al. 2018), exhibit similar or worse performance than random sampling in experiments, so we only additionally choose random sampling as a baseline.

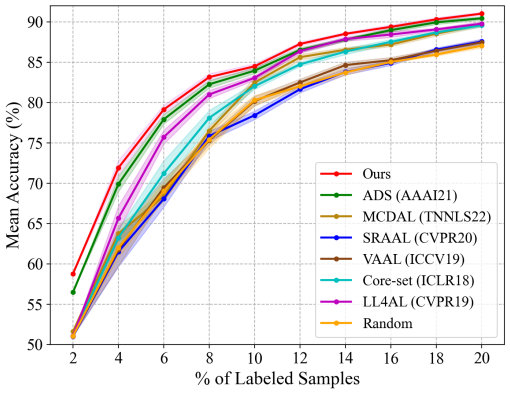
Experimental Settings. We use ResNet-18 (He et al. 2016) as the backbone network for all image classification tasks and only fine-tune the final feature layer and the fully connected layer. For all classification datasets, we randomly select 10% samples from the entire dataset to initialize the labeled pool, and the rest is considered the unlabeled pool. In each iteration of the current model training, we select 5% samples from the unlabeled pool until the portion of labeled samples reaches 40%. In addition, to validate the performance of our method at a relatively small budget of the labeled set, for CIFAR-10, we start training with 1000 labeled images and iterate for 10 cycles, adding 1000 images for each iteration, and finally reaching 20% of the labeled samples. For each learning iteration, we train the model for 200 epochs using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.1, a momentum of 0.9, a weight decay of 0.0005, and a batch size of 128. After 80% of the training epochs, the learning rate is decreased to 0.01.

Performance on CIFAR-10. Fig. 3(a) shows that MAOAL outperforms the current state-of-the-art methods in all the stages with notable margins, especially in the late stages. When labeled data rates are 30%, 35%, and 40%, the mean accuracies of MAOAL are 92.95%, 93.50%, and 93.98%, respectively, which are 0.7%, 0.6%, and 0.53% higher than the second-best method (MCDAL). The experimental results show the superiority of MAOAL on a dataset with a small number of categories.

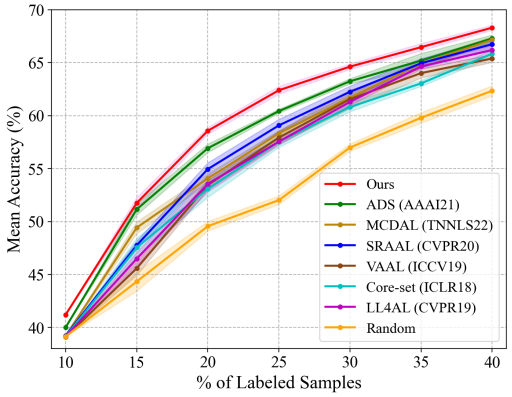
When the labeling budget is restricted, *e.g.*, one may only be able to annotate 20% of the data instead of 40%, MAOAL is also proven to be beneficial and shows performance improvement at every stage, particularly at the early iterations, as shown in Fig. 3(b). MAOAL achieves 91.03% accuracy



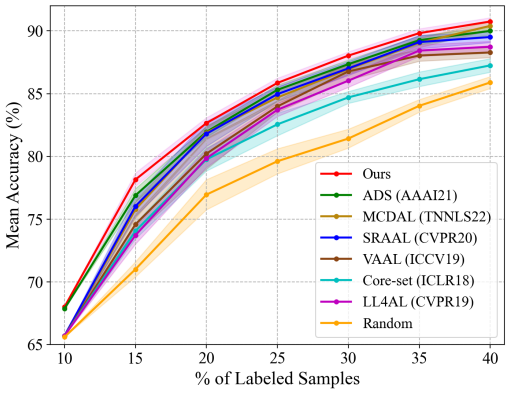
(a) CIFAR-10 when the labeling budget is 40%



(b) CIFAR-10 when the labeling budget is 20%

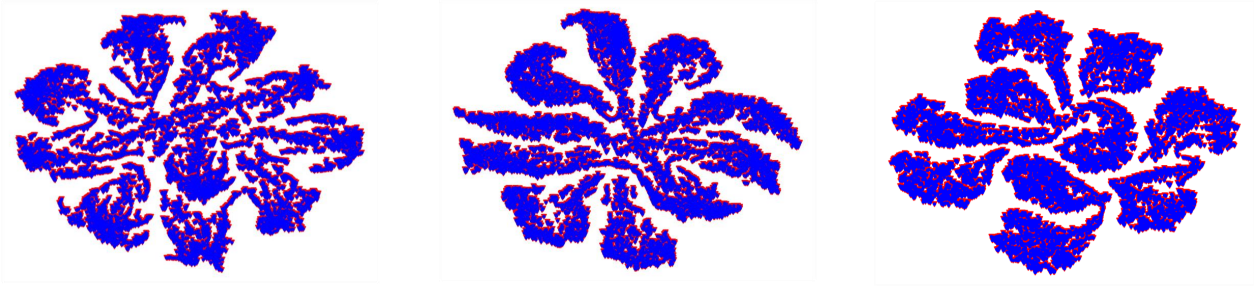


(c) CIFAR-100



(d) Caltech-101

Figure 3: Results in image classification on CIFAR-10, CIFAR-100, and Caltech-101.



(a) Iteration 1 (b) Iteration 4 (c) Iteration 7

Figure 4: t-SNE visualization of the features generated by MAOAL on CIFAR-10. Red points are the sampled points to label.

with 20% samples, which is very close to that on the full training set. The results show that when the labeling budget is limited, MAOAL can select informative samples using a small training set and achieve higher accuracy.

Performance on CIFAR-100. CIFAR-100 has 10 times more classes than CIFAR-10 and is a more challenging dataset. As shown in Fig. 3(c), when data rates are 15%, 25%, and 35%, the mean accuracies of MAOAL are 51.75%, 62.40%, and 66.47%, respectively, which are 0.60%, 1.97%, and 1.27% higher than the second-best method (ADS). Overall, MAOAL outperforms other state-of-the-art methods at all sampling stages, demonstrating a considerable performance gap, and shows fairly decent performance at 40% labeled data.

L_{C^3}	L_1	Ent.	$S()$	Bi-cls	Tri-cls	Accuracy (%) on Proportion (%) of Labeled Samples						
						10	15	20	25	30	35	40
						39.14	44.35	49.56	52.12	56.99	59.80	62.35
	✓		✓	✓		37.22	47.84	54.04	59.74	62.85	63.78	65.85
✓			✓	✓		40.45	50.15	57.60	61.96	64.93	66.10	68.24
✓		✓		✓		39.48	49.58	56.62	61.32	63.05	64.40	66.24
✓		✓			✓	39.77	50.16	57.10	61.50	62.36	64.45	66.73
✓			✓		✓	**41.20**	**51.75**	**58.57**	**62.40**	**64.63**	**66.47**	**68.31**

Table 1: Results of our proposed method with/without the proposed components/structures. L_{C^3} and L_1 respectively denote using the proposed C^3 loss and the l_1 distance as the classifier discrepancy metric. Ent. and $S()$ respectively denote using the mean entropy sampling and our designed sampling strategy $S()$ to select samples. Bi-cls denotes using two classifiers in the network, while Tri-cls denotes training the network with three classifiers, *i.e.*, one main classifier and two auxiliary ones.

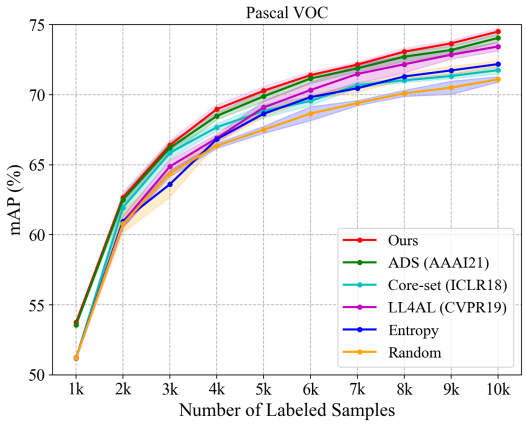


Figure 5: Results in object detection on PASCAL VOC.

Performance on Caltech-101. The Caltech-101 dataset is considerably smaller compared to CIFAR-100, including images of much higher resolution (*i.e.*, 300×200) and a different number of images per class. Therefore, it is a rather difficult dataset. As can be seen from Fig. 3(d), our MAOAL outperforms other state-of-the-art methods at all stages, with the best performance at 40% labels, achieving 90.75% accuracy. These results show the capability of our MAOAL in a real-world setting with unbalanced data.

Visualization. Fig. 4 depicts the t-SNE visualization of features learned by MAOAL on CIFAR-10. Blue and red points indicate unlabeled and labeled samples, respectively. With the increasing training iterations, we can observe that feature distributions gradually present good clustering results. The learned features align the labeled and unlabeled samples with 10 clusters with clear boundaries. Points with red color are sampled points for labeling.

Object Detection

Datasets and Settings. Pascal VOC (Everingham et al. 2010) contains 20 object categories, consisting of the VOC 2007 trainval set, the VOC 2012 trainval set, and the VOC 2007 test set. We use the trainval sets of VOC 2007 and VOC 2012 datasets for training, which contain 5011 and 11540 images. We follow LL4AL to adopt SSD (Liu et al. 2016) with VGG-16 (Simonyan and Zisserman 2015) as the base detector, where 1,000 images in the training set are selected as the initially labeled subset and 1000 images are selected at each acquisition cycle. We learn the model set for 300 epochs with the mini-batch size of 32. The learning rate for the first 240 epochs is 0.001 and decreased to 0.0001 for the last 60 epochs. We compare MAOAL with random sampling, entropy sampling, Core-set, LL4AL, and ADS.

Performance. Fig. 5 illustrates our results of object detection on VOC compared to previous approaches. From the figure, we can see that MAOAL outperforms all other methods throughout the process.

Ablation Study

To evaluate the effect of our proposed components/structures, we conduct an ablation study on CIFAR-100 using ResNet-18. As shown in Tab. 1, when using the proposed C^3 loss as the classifier discrepancy and the sampling acquisition function (Eqs. (6) and (14)), our method yields substantially higher performance at all AL stages compared to using the l_1 distance and the mean entropy sampling. When training the network with three classifiers, a main classifier and two auxiliary classifiers, MAOAL significantly boosts the performance in all training iterations. This confirms the effectiveness of the proposed method in generating a larger sampling area and highlighting hard samples.

Conclusion

In this paper, we propose a multi-classifier adversarial active learning algorithm, MAOAL, that learns a discriminative representation on the unlabeled data in an adversarial game. We calculate the prediction discrepancy between classifiers by introducing a novel metric, classifier class confusion (C^3). We adversarially optimize the C^3 loss by adding two auxiliary classifiers to align the distribution of labeled and unlabeled data while pushing the hard samples to the decision boundaries, which facilitates selecting more informative samples. The experimental results on four datasets demonstrate the effectiveness of the proposed method.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62276129) and the Natural Science Foundation of Jiangsu Province (No. BK20220890).

References

Agarwal, S.; Arora, H.; Anand, S.; and Arora, C. 2020. Contextual Diversity for Active Learning. In *ECCV*.

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *ICLR*.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. PAMI*, 39(12): 2481–2495.

Beluch, W. H.; Genewein, T.; Nurnberger, A.; and Kohler, J. M. 2018. The Power of Ensembles for Active Learning in Image Classification. In *IEEE CVPR*.

Berthelot, D.; Schumm, T.; and Metz, L. 2017. BEGAN: Boundary Equilibrium Generative Adversarial Networks. *arXiv preprint arXiv:1703.10717*.

Brinker, K. 2003. Incorporating Diversity in Active Learning with Support Vector Machines. In *ICML*.

Caramalau, R.; Bhattarai, B.; and Kim, T.-K. 2021. Sequential Graph Convolutional Network for Active Learning. In *IEEE CVPR*.

Cho, J. W.; Kim, D.-J.; Jung, Y.; and Kweon, I. S. 2022. MCDAL: Maximum Classifier Discrepancy for Active Learning. *IEEE Trans. NNLS*, Early Access: 1–11.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE CVPR*.

Ducoffe, M.; and Precioso, F. 2018. Adversarial Active Learning for Deep Networks: A Margin Based Approach. *arXiv preprint arXiv:1802.09841*.

Ebrahimi, S.; Elhoseiny, M.; Darrell, T.; and Rohrbach, M. 2020. Uncertainty-Guided Continual Learning with Bayesian Neural Networks. In *ICLR*.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2): 303–338.

Fu, M.; Yuan, T.; Wan, F.; Xu, S.; and Ye, Q. 2021. Agreement-Discrepancy-Selection: Active Learning with Progressive Distribution Alignment. In *AAAI*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *ICML*.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep Bayesian Active Learning with Image Data. In *ICML*.

Gissin, D.; and Shalev-Shwartz, S. 2019. Discriminative Active Learning. *arXiv preprint arXiv:1907.06347*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*.

Gorriz, M.; Carlier, A.; Faure, E.; and Giro-i-Nieto, X. 2017. Cost-Effective Active Learning for Melanoma Segmentation. In *NIPS Workshop*.

Haussmann, E.; Fenzi, M.; Chitta, K.; Ivanecky, J.; Xu, H.; Roy, D.; Mittel, A.; Koumchatzky, N.; Farabet, C.; and Alvarez, J. M. 2020. Scalable Active Learning for Object Detection. In *IEEE IV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE CVPR*.

Houlsby, N.; Huszár, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian Active Learning for Classification and Preference Learning. *arXiv preprint arXiv:1112.5745*.

Jin, Y.; Wang, X.; Long, M.; and Wang, J. 2020. Minimum Class Confusion for Versatile Domain Adaptation. In *ECCV*.

Joshi, A. J.; Porikli, F.; and Papanikolopoulos, N. 2009. Multi-Class Active Learning for Image Classification. In *IEEE CVPR*.

Kim, K.; Park, D.; Kim, K. I.; and Chun, S. Y. 2021. Task-Aware Variational Adversarial Active Learning. In *IEEE CVPR*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114v10*.

Kirsch, A. 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. In *NIPS*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.

Kuo, W.; Häne, C.; Yuh, E.; Mukherjee, P.; and Malik, J. 2018. Cost-Sensitive Active Learning for Intracranial Hemorrhage Detection. In *MICCAI*.

Li, J.; Tang, S.; Zhu, L.; Shi, H.; Huang, X.; Wu, F.; Yang, Y.; and Zhuang, Y. 2021. Adaptive Hierarchical Graph Reasoning with Semantic Coherence for Video-and-Language Inference. In *IEEE ICCV*.

Li Fei-Fei; Fergus, R.; and Perona, P. 2006. One-Shot Learning of Object Categories. *IEEE Trans. PAMI*, 28(4): 594–611.

Liu, B.; and Ferrari, V. 2017. Active Learning for Human Pose Estimation. In *IEEE ICCV*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *ECCV*.

MacKay, D. J. C. 1992. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4): 590–604.

Mayer, C.; and Timofte, R. 2020. Adversarial Sampling for Active Learning. In *IEEE WACV*.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In *IEEE CVPR*.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *ICLR*.

Settles, B. 2009. Active Learning Literature Survey. Technical report, University of Wisconsin-Madison.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational Adversarial Active Learning. In *IEEE ICCV*.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin, L. 2016. Cost-Effective Active Learning for Deep Image Classification. *IEEE Trans. CSVT*, 27(12): 2591–2600.

Wang, S.; Li, Y.; Ma, K.; Ma, R.; Guan, H.; and Zheng, Y. 2020. Dual Adversarial Network for Deep Active Learning. In *ECCV*.

Yang, Y.; Ma, Z.; Nie, F.; Chang, X.; and Hauptmann, A. G. 2015. Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization. *IJCV*, 113(2): 113–127.

Yoo, D.; and Kweon, I. S. 2019. Learning Loss for Active Learning. In *IEEE CVPR*.

Yuan, T.; Wan, F.; Fu, M.; Liu, J.; Xu, S.; Ji, X.; and Ye, Q. 2021. Multiple Instance Active Learning for Object Detection. In *IEEE CVPR*.

Zhang, B.; Li, L.; Yang, S.; Wang, S.; Zha, Z.-J.; and Huang, Q. 2020. State-Relabeling Adversarial Active Learning. In *IEEE CVPR*.

Zhu, J.-J.; and Bento, J. 2017. Generative Adversarial Active Learning. *arXiv preprint arXiv:1702.07956*.