Dual Adversarial Network for Deep Active Learning

Shuo Wang[1,2], Yuexiang Li[2(✉)], Kai Ma[2], Ruhui Ma[1(✉)], Haibing Guan[1], and Yefeng Zheng[2]

[1] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
ruhuima@sjtu.edu.cn
[2] Tencent Jarvis Lab, Shenzhen, China
vicyxli@tencent.com

Abstract. Active learning, reducing the cost and workload of annotations, attracts increasing attentions from the community. Current active learning approaches commonly adopted uncertainty-based acquisition functions for the data selection due to their effectiveness. However, data selection based on uncertainty suffers from the overlapping problem, i.e., the top-K samples ranked by the uncertainty are similar. In this paper, we investigate the overlapping problem of recent uncertainty-based approaches and propose to alleviate the issue by taking representativeness into consideration. In particular, we propose a dual adversarial network, namely DAAL, for this purpose. Different from previous hybrid active learning methods requiring multi-stage data selections i.e., step-by-step evaluating the uncertainty and representativeness using different acquisition functions, our DAAL learns to select the most uncertain and representative data points in one-stage. Extensive experiments conducted on three publicly available datasets, i.e., CIFAR10/100 and Cityscapes, demonstrate the effectiveness of our method—a new state-of-the-art accuracy is achieved.

Keywords: Active learning · Generative adversarial network · Unsupervised video summarization · Deep learning

1 Introduction

Benefiting from large-scale annotated datasets, deep learning has shown its great success in various computer vision tasks such as image classification, object detection, and semantic segmentation. Yet, the annotation of large-scale datasets is extremely laborious and costly to obtain, especially for the dense pixel-level annotation and the one requiring experienced annotators to tackle (e.g., medical images). For this reason, semi-supervised learning methods [3,18,25,29,30] and

S. Wang—Intern at Tencent Jarvis Lab.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12369, pp. 680–696, 2020.
https://doi.org/10.1007/978-3-030-58586-0_40

unsupervised learning methods [1,6,28,44] attract increasing attention. However, given a fixed amount of data, their performance is still bound to that of fully-supervised learning.

Active learning (AL) that incrementally queries the most informative samples from the data pool to reduce the overall annotation effort has thus emerged as a promising research avenue for the use of deep learning [9]. Among recent AL-related works, pool-based AL methods [10,32,36,41], which iteratively select data points from a large unlabeled data pool for annotation according to the acquisition function, are the most successful. Accurate estimation of data informativeness is the core of pool-based AL. Many researches focused on exploring effective acquisition functions to achieve this goal, which can be classified to two categories—uncertainty-based and representation-based. The AL using uncertainty-based acquisition functions [36,41] prefer to select samples confusing the classifier (i.e., high uncertainty), while the representation-based ones [39,45] select samples best representing the unlabeled pool. The estimation of data informativeness is not comprehensive in either term of uncertainty or representativeness. Therefore, some studies [39,45] proposed the hybrid strategies. However, their inferior performance compared to the uncertainty-based approaches [36,41] illustrates that the benefit of representativeness is not actually exploited.

Recent years witnessed the success of adversarial networks and several studies [11,36] tried to apply the adversarial learning for more accurate estimation of data informativeness. For example, Sinha et al. [36] proposed the VAAL by using a variational autoencoder (VAE) [20] and a discriminator, where the VAE embedded the labeled and unlabeled images to a latent space and the discriminator was utilized as a binary classifier to measure the uncertainty of the input samples. However, the selected samples may not be the most informative ones for the task model performing the target task (e.g., image classification and semantic segmentation) since the VAAL is fully task-agnostic.

In contrast, Yoo et al. [41] employed the loss of task model as the criterion to estimate the contribution of data made to the target task—a loss prediction module was proposed to estimate the loss of task model for the unlabeled data. Since the loss prediction module directly utilizes features from the task model, it gives more accurate estimation of the informativeness of current input to the task model. However, this loss prediction module suffers from the overlapping problem [32,41]—the information contained in the top K selected samples is similar. The solution [32,41] to this problem is performed in a two-stage manner—a random subset with a certain size from the unlabeled pool is firstly created to ensure the diversity of data, and then apply uncertainty-based acquisition function to that subset for data point selection. Although this simple random subset selection (RSS) strategy can alleviate the overlapping problem, its effectiveness to other uncertainty-based AL methods [36] has not been explored. Current multi-stage AL frameworks, i.e., step-by-step evaluating the uncertainty and representativeness using different methods, are another potential solutions to the overlapping problem. However, they are time-consuming for data selection and difficult to provide proper estimation of data informativeness due to the trade-off between

two evaluation methods. Therefore, a one-stage AL method without suffering from the overlapping problem is worthwhile to develop.

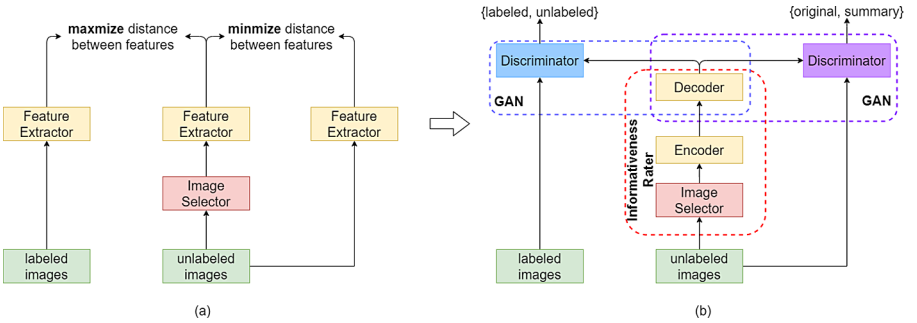

Fig. 1. Overview: (a) Our goal is to simultaneously select the most uncertain and representative images for the task model from the unlabeled pool. The image selector is required to select the samples maximizing the distance to labeled images (uncertainty), while minimizing the distance to unlabeled ones (representativeness). (b) The generative adversarial framework is proposed to assist image selector to properly measure the distances between deep features.

To this end, we propose a one-stage dual adversarial network for active learning, namely DAAL, to accurately select the most informative data points from the unlabeled pool for the task model by simultaneously considering the uncertainty and representativeness. Inspired by the unsupervised video summarization [15,26] in which a frame selector was proposed to learn a representative summary[1] of the original video, we propose to use an image selector as the acquisition function to find a sparse and representative subset from the unlabeled pool, which is simultaneously with high uncertainty for the task model.

The overview of our approach is illustrated in Fig. 1 (a). The image selector is required to find samples with larger feature distance to the labeled samples (i.e., high uncertainty to the classifier) and smaller distance to the unlabeled data (i.e., good representativeness to the unlabeled data). However, specifying a suitable distance of deep features is difficult [23]. Therefore, we use a generative adversarial (GAN) [13] framework to assist the image selector in the distance measurement between deep feature representations. As shown in Fig. 1 (b), samples selected by the image selector are sent to a VAE, which embeds the features of selected samples into the same latent space and then reconstructs them. The reconstructed features are fed to two discriminators, which encourage the image selector to simultaneously take uncertainty (labeled/unlabeled) and representativeness (original/summary) into consideration during data selection. The informativeness rater (consisting of image selector and VAE) and discriminators are

[1] Summary is a sparse subset of video frames which optimally represent the input video.

framed as a multi-player competition, similar to GAN [13]. The informativeness rater is trained to trick the discriminators. The performances of informativeness rater and discriminators are improved by iteratively optimization.

In summary, our contributions are manifold. First, we propose a one-stage dual adversarial network for active learning, namely DAAL, which can simultaneously learn to select the most uncertain and representative data points from the unlabeled pool. Second, our designed DAAL is more effective to alleviate the overlapping problem, compared to the approaches using random subset selection and other hybrid methods. Last but not least, extensive experiments conducted on three publicly available datasets (CIFAR10/100 and Cityscapes) show that our approach outperforms the benchmarking methods and achieves a new state-of-the-art.

2 Related Work

Active Learning. The methods in the area of active learning can be roughly classified to two categories—uncertainty-based and representation-based methods.

Uncertainty-Based Methods. The core idea is to find the samples, which are difficult for the classifier to correctly classify (i.e., with high uncertainty to the classifier). These methods can be further categorized to Bayesian and non-Bayesian frameworks. For Bayesian active learning methods, probabilistic models such as Bayesian neural networks [7] and Gaussian processes [19] are used to estimate the uncertainty of samples. Houlsby et al. [16] proposed a Bayesian active learning, which used the mutual information of the training examples as a proxy uncertainty measurement for sample selection. For non-Bayesian active learning methods, the sample uncertainty can be measured in various ways such as the distance between samples and the decision boundary [4], information entropy [17] and risk expectation [38]. In more recent works, Gal et al. [9] proposed to utilize dropout layers to estimate the uncertainty of the prediction yielded by a neural network for sample query. Yoo et al. [41] proposed to use an auxiliary loss prediction module to learn the target loss of inputs jointly with the training phase and samples with high predicted losses are selected. Sinha et al. [36] proposed a framework (namely VAAL) for active learning, consisting of a variational autoencoder (VAE) and generative adversarial network (GAN). The probability of discriminator is seen as the uncertainty estimation for sample selection.

Representation-Based Methods. This kind of approaches aims to constitute a set of diverse samples, which are the most representative of the entire dataset. Sener et al. [31] proposed a core-set selection method, which selected the samples minimizing the Euclidean distance between the selected data and the unlabeled data pool in the feature space. There are also some hybrid methods [8,40] taking both uncertainty and diversity into account.

Active Learning for Semantic Segmentation. Semantic segmentation is one of the most prevailing tasks for active learning due to its expensive annotation, which has been broadly investigated in recent studies [24,34,36,39]. Yang et al. [39] proposed a hybrid framework, namely suggestive annotation (SA), combining the measurements of uncertainty and representativeness. This framework estimated the uncertainty of data points using an ensemble of models and measured the representativeness using the core-set approach [31].

Variational Autoencoder. Autoencoders are commonly used to effectively learn a feature representation for various tasks [2]. Variational autoencoder is a variant of autoencoder, which defines a posterior distribution over the observed data, given an unobserved latent variable. A VAE is used to embed the labeled and unlabeled images into the same latent space by VAAL [36]. Given $e \sim p_e(e)$ as a priori over the unobserved latent variable, we can formulate the objective function of VAE with observed data x as:

$$\mathcal{L}_{VAE} = -\log \frac{p(x|e)p(e)}{q(e|x)} = \underbrace{-\log(p(x|e))}_{\mathcal{L}_{recon}} + \underbrace{D_{KL}(q(e|x)||p(e))}_{\mathcal{L}_{prior}} \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence; $q(e|x)$ is the probability of observing e given x; $p_e(e)$ is the standard normal distribution; and $p(x|e)$ is the conditional generative distribution for x.

Generative Adversarial Network. The typical generative adversarial network (GAN) [13] consists of a generator network and a discriminator. The generator generates data simulating an unknown distribution and the discriminator network aims to distinguish the generated/fake samples from the real ones. The generator and the discriminator are alternately trained to force the generator fitting the real data distribution while maximizing the probability of the discriminator making a mistake. Given x as the true data, $e \sim p_e(e)$ as the prior input noise, and $\hat{x} = G(e)$ as the generated sample, the objective function of a typical GAN can be formulated as:

$$\min_G \max_D \left[\mathbb{E}_x[\log D(x)] + \mathbb{E}_e[\log(1 - D(\hat{x}))]\right] \tag{2}$$

where the discriminator D is trained to maximize the probability of real/fake classification and the generator G is trained to minimize $\log(1 - D(\hat{x}))$.

3 Method

In this section, we introduce the proposed DAAL in details. Specific descriptions of network architecture are introduced in Sect. 3.1. In Sect. 3.2, the detailed training procedure of the DAAL is illustrated.

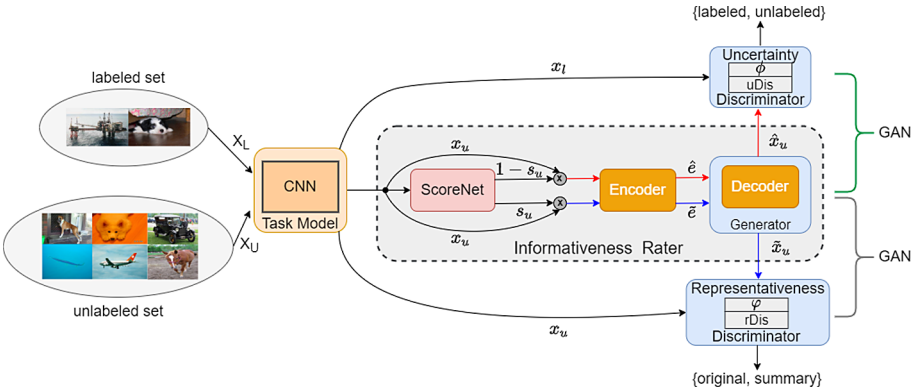

Fig. 2. Major components of our approach. The informativeness rater consists of a ScoreNet (image selector) and a VAE (encoder-decoder architecture). Given the features encoded by the task model for labeled (x_l) and unlabeled images (x_u), the ScoreNet assigns importance score s_u to x_u. The encoder encodes the sample with a pair of features ($\hat{e}$, $\tilde{e}$), which are fed to the decoder for reconstruction ($\hat{x}_u$, $\tilde{x}_u$). The reconstructed features are sent to the dual discriminators, respectively. The uDis is required to classify whether the $\hat{x}_u$ belongs to labeled pool or not. The rDis aims to identify the reconstructed feature $\tilde{x}_u$ from the original one x_u. The decoder/generator and dual discriminators are adversarially trained until the uDis cannot discriminate between labeled and unlabeled data points and the rDis is not able to distinguish between the summary and original datasets.

3.1 Dual Adversarial Network for Deep Active Learning

The detailed information of our DAAL including the information flow is illustrated in Fig. 2. The ScoreNet and VAE form an independent function unit, namely informativeness rater, which cooperates with two different discriminators to construct the dual adversarial network. To accurately measure the informativeness of input samples for the task model, the features of labeled and unlabeled images encoded by the task model are adopted as input of our DAAL.

Informativeness Rater. Given the deep features of images from the unlabeled pool ($X_U = \{x_u : u = 1, ..., N\}$) generated by the task model, the ScoreNet assigns a relative importance score ($s = \{s_u : s_u \in [0, 1], u = 1, ..., N\}$) to each of them. Original input features x_u are weighted using these scores. Note that we use $1 - s_u$ and s_u as the weights for uncertainty and representativeness branches, respectively, to ensure the optimization of these two terms in the same direction. These weighted deep features are sent to a VAE which consists of an encoder and a decoder/generator. The encoder maps the inputs to the features ($e = \{\hat{e}, \tilde{e}\}$) in the same latent space, while the decoder/generator reconstructs the embedded features ($x_{\text{recon}} = \{\hat{x}_u, \tilde{x}_u\}$). The ScoreNet adopts a simple architecture, consisting of a 5-layer multi-layer perceptron (MLP) with Xavier initialization [12], to map the input feature to a 1×1 score vector. Both the encoder and decoder are

neural networks with 7-layer MLP. A dropout layer [37] with the dropout ratio of 0.4 is added to the end of each MLP layer to avoid overfitting.

Dual Discriminators. The dual discriminators are utilized to measure the distance between the input features and their reconstructions given by the generator. The deep features, i.e., x_l, $\hat{x}_u$ and x_u, $\tilde{x}_u$, are fed to the dual discriminators, respectively, for different purposes. Specifically, *uncertainty discriminator* (uDis) takes x_l and $\hat{x}_u$ as input, and aims to distinguish which pool (i.e., 'labeled' or 'unlabeled') the features belong to. *Representativeness discriminator* (rDis) tasks x_u and $\tilde{x}_u$ as input, and aims to classify them into two distinct classes (i.e., original or summary). The dual discriminators in the GAN adopt the same architecture to ScoreNet without dropout layers. The class 'summary' represents the reconstruction of weighted deep features of the input batch. If the discriminator cannot distinguish the summary batch from the original one, the images with high scores are seen to have good representativeness to the small unlabeled pool.

Training Strategy: Multi-player Competition. Our DAAL involves two adversarial games between the VAE and dual discriminators, which iteratively optimize the ScoreNet for accurate data selection. Due to the score $s \in [0, 1]$, the features with scores closed to 1 are easier for decoder/generator to reconstruct. For the uDis, if a well-reconstructed $\hat{x}_u$ fools it, the lower score ($1 - s_u$ as the weights for $\hat{x}_u$) should be maintained. The VAE and uDis optimize the ScoreNet via adversarial training. Simultaneously, as aforementioned, the adversarial learning between VAE and rDis encourage the ScoreNet to assign larger scores to the samples representative to the unlabeled pool. Therefore, the importance scores yielded by our ScoreNet can simultaneously evaluate the uncertainty and representativeness of data points—a higher score intrinsically represents the image with both larger uncertainty and representativeness. During the sample selection of active learning, samples with top k largest scores ranked by the ScoreNet are selected from the unlabeled pool for annotation.

3.2 Training Procedure of DAAL

Denote the network weights of ScoreNet, encoder and decoder of VAE and the dual discriminator (uDis and rDis) as $\{w_s; w_e, w_d; w_u, w_r\}$. The training procedure of DAAL is summarized in Algorithm 1. The proposed DAAL is supervised by four loss functions, which are 1) prior loss $\mathcal{L}_{prior}$ (as defined in Eq. 1) for the encoder of VAE, 2) reconstruction loss $\mathcal{L}_{recon}$ for VAE, 3) GAN loss $\mathcal{L}_{GAN}$, and 4) sparsity loss $\mathcal{L}_{\text{sparsity}}$ for the ScoreNet.

Reconstruction Loss $\mathcal{L}_{\text{recon}}$. Instead of using the standard reconstruction loss for autoencoder networks, i.e., $||x - \hat{x}||_2$ where x and $\hat{x}$ are the input and corresponding reconstruction, respectively, we follow the practice in [23] to use the last output layer of the discriminators for the calculation of $\mathcal{L}_{\text{recon}}$. Denote the

Algorithm 1. Training dual adversarial network

Input: Features (X_L and X_U) encoded by the task model for labeled and unlabeled images, respectively.
Output: Learned parameters $\{w_s, w_e, w_d, w_u, w_r\}$.
Function:
$f(x; w)$: forward the input x through neural network (w).
update(.): backward to update the neural network weights.
$\mathcal{L}(.)$: loss function.
Procedure:
Initialize all parameters $\{w_s, w_e, w_d, w_u, w_r\}$
for batch (x_l, x_u) from X_L and X_U **do**
    $s_u \leftarrow f(x_u; w_s)$ *// select images*
    $(\hat{e}, \tilde{e}) \leftarrow f((x_u, s_u); w_e)$ *// encoding*
    $(\hat{x}_u, \tilde{x}_u) \leftarrow f((\hat{e}, \tilde{e}); w_r)$ *// reconstruction*
    // Updates using stochastic gradient:
    $\{w_s, w_e\} \leftarrow update(\mathcal{L}_{recon}(x_u, \hat{e}, \tilde{e}) + \mathcal{L}_{prior}(x_u, \hat{e}, \tilde{e}) + \mathcal{L}_{sparsity}(x_u, s_u))$
    $\{w_d\} \leftarrow update(\mathcal{L}_{\text{recon}}(x_u, \hat{e}, \tilde{e}) + \mathcal{L}_{\text{GAN}}(x_u, x_l, \hat{x}_u, \tilde{x}_u))$
    $\{w_u, w_r\} \leftarrow update(\mathcal{L}_{\text{GAN}}(x_u, x_l, \hat{x}_u, \tilde{x}_u))$ *// maximization update*
end for

output of the last hidden layer of uDis and rDis as $\phi(x_u)$ and $\varphi(x_u)$, for input x_u, respectively. Given embedded features $\hat{e}$ and $\tilde{e}$ of input x_u, $\mathcal{L}_{\text{recon}}$ can be formulated as:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}[-\log p(\phi(x_u)|\hat{e})] + \mathbb{E}[-\log p(\varphi(x_u)|\tilde{e})] \tag{3}$$

where expectation $\mathbb{E}$ is approximated as the empirical mean of the training samples.

GAN loss $\mathcal{L}_{\text{GAN}}$. The adversarial learning between generator and dual discriminators is supervised by the GAN loss, which can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = &\log(\text{uDis}(x_l)) + \log(1 - \text{uDis}(\hat{x}_u)) \\ &+ \log(\text{rDis}(x_u)) + \log(1 - \text{rDis}(\tilde{x}_u)).\end{aligned} \tag{4}$$

where uDis(.) and rDis(.) represent the model functions of uncertainty discriminator and representativeness discriminator, respectively.

Sparsity Loss. The sparsity loss is a regularization term for the ScoreNet to prevent it from assigning equal importance to all data points. The sparsity loss consists of a length regularizer loss $\mathcal{L}_{LR}$ and a determinantal point process (DPP) loss $\mathcal{L}_{DPP}$ [22]. The $\mathcal{L}_{LR}$ limits the number of elements selected by the ScoreNet, while the $\mathcal{L}_{DPP}$ ensures the diversity of selected data points. The overall sparsity loss [35] can be defined as:

$$\mathcal{L}_{\text{sparsity}} = \mathcal{L}_{\text{LR}} + \mathcal{L}_{\text{DPP}} \tag{5}$$

where

$$\mathcal{L}_{\mathrm{LR}} = \left\| \sigma - \frac{1}{n} \sum_{t=1}^{n} s_t \right\|_2, \quad \mathcal{L}_{\mathrm{DPP}} = -\log(P(s_{x'})) \tag{6}$$

where σ represents the percentage of images for subset selection; $s_{x'}$ is the importance scores for a subset $x' \subset X_U$. The probability function P in $\mathcal{L}_{\mathrm{DPP}}$ can be written as:

$$P(s_{x'}; D) = \frac{|D(s_{x'})|}{|D + I|} \tag{7}$$

where $D \in R^{n \times n}$ with $D_{i,j} = s_i s_j x_i^T x_j$ and $D(s_{x'}) \in R^{\sigma n \times \sigma n}$ (i.e., a submatrix of D given $s_{x'}$); $|.|$ denotes determinant and I is the identity matrix.

4 Experiments

In this section, we evaluate the effectiveness of our DAAL on three publicly available datasets. The evaluation results are presented in Sects. 4.3 and 4.4. Furthermore, we conduct an in-depth investigation on the drawback of uncertainty-only AL approaches. The related results can be found in Sect. 4.5. Finally, we analyze the importance of each component in the DAAL network in Sect. 4.6.

4.1 Datasets

CIFAR10 [21] **and CIFAR100** [21]**.** We evaluate the proposed DAAL on CIFAR10 and CIFAR100 datasets for image classification. Both datasets contain 60,000 images with a uniform size of 32 × 32 pixels. CIFAR10 and CIFAR100 have 10 and 100 categories, respectively. The training set and test set consist of 50,000 images and 10,000 images, respectively. The average classification accuracy is adopted as the metric for performance evaluation in this task.

Cityscapes [5]**.** Our DAAL is further evaluated on the Cityscapes dataset for semantic segmentation. Cityscapes is a large scale driving video dataset which contains 3,475 frames with instance segmentation annotations of 19 classes. The images have a uniform size of 2048 × 1024 pixels. The Cityscapes dataset is separated to the public training set and test set. For fair comparison, we adopt the public dataset partition in our experiments. The mean IoU is utilized to evaluate the performance of semantic segmentation.

4.2 Experimental Settings

Consistent to the existing approaches [36,41], we randomly select 10% samples from the training set for labeling and use them as the initial labeled pool at the beginning of the experiments. The rest of the training data set is treated as the unlabeled pool. In each iteration of active learning, we augment the labeled pool with 5% of the whole training set by selecting samples from the unlabeled pool for oracles to annotate using the acquisition function.

Baseline Methods. We involve various previous AL approaches as baselines for comparison, including common AL approaches for both image classification and semantic segmentation, e.g., Entropy [33], Learning Loss (LL) [41], Core-set [32], and VAAL [36], and task-specific ones for semantic segmentation only, e.g., MC Dropout [14] and Suggestive Annotation [32]. The results using *random sampling* is also reported for comparison. We notice that the idea of SA is closed to our DAAL, which both considers the uncertainty and representativeness for data point selection. However, SA is a two-step hybrid ensemble method, using the bootstrapping and Core-set for uncertainty and representativeness estimation, respectively. In contrast, our DAAL can select the samples in both terms of uncertainty and representativeness in one-step.

Implementation Details. For the image classification task, we use Wide-Resnet-Network [43] with depth = 28, width = 2 (WRN-28-2) as the backbone of the task model, while the dilated residual network [42] is used for the semantic segmentation task. The dual adversarial network is implemented using the PyTorch toolbox. For fair comparison, the baselines adopt the same training protocol. To alleviate the influence caused by the random nature of a neural network and the random initial labeled pool, all experimental results reported are the average of three repeated experiments. Note that we evaluate all methods without any data augmentation. One reason is that we expect to select the most representative subset of the raw dataset rather than the enlarged dataset by data augmentation. The other is that as reported in [27], the performance of AL methods with/without using data augmentation differs drastically, which is difficult for a fair comparison.

Image Classification. The task model and our DAAL are trained for 150 epochs. The stochastic gradient descent (SGD) optimizer is adopted to supervise the training of the task model. The initial learning rate is set to 0.1 and decreases to 0.01 after 80 epochs and 0.001 after 120 epochs, respectively. For the training of our dual adversarial network, the Adam optimizer is used with the learning rate of 5×10^{-4}. The batch size during adversarial learning is set to 128 and σ of Eq. 6 is set to 0.2.

Semantic Segmentation. The task model and our DAAL are trained for 50 epochs and 100 epochs, respectively, using the Adam optimizer. The learning rate is set to 5×10^{-4} and σ of Eq. 6 is set to 0.2.

4.3 Image Classification on CIFAR10/100

We compare the proposed DAAL with the baselines on CIFAR10/100. The evaluation results are presented in Fig. 3. On CIFAR10, the WRN-28-2 trained with the samples selected by our DAAL achieves the highest average classification accuracy, e.g., $84.17 \pm 0.24\%$ with 40% data from the training dataset, which is close to the accuracy yielded by training on the entire dataset (i.e.,

$88.34 \pm 0.43\%$). As shown in Fig. 3, the average classification accuracy between different AL approaches is similar with an extremely small labeled pool (i.e., 15%). The underlying reason is that the task model is not well trained and samples in the unlabeled pool may contain similar informativeness for the task model. As the size of labeled pool increases, the DAAL shows its advantage on data selection and surpasses the runner-up (i.e., Learning Loss [41]). The random sampling strategy yields the lowest classification accuracy under most settings of labeled data amount.

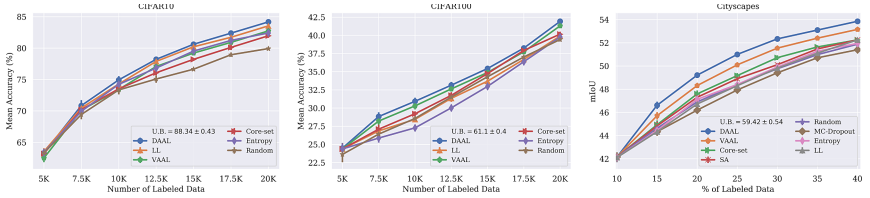

Fig. 3. Performance comparison of different AL approaches, including VAAL [36], Learning Loss (LL) [41], Core-set [32], Entropy [33], random sampling and our DAAL on CIFAR10/100. For semantic segmentation, the additional approaches, MC-Dropout [14] and Suggestive Annotation [32] are compared. The U.B. denotes the upper bound performance given by the task model trained on the entire data set.

Similar trends of improvement are observed on CIFAR100. Our DAAL achieves an average classification accuracy of $41.92 \pm 0.31\%$ with 40% training data, which is the highest record among the listed AL approaches. It is worthwhile to mention that only our DAAL provides consistent improvements on both CIFAR10 and CIFAR100. The runner-up approach (i.e., Learning Loss) on CIFAR10 achieves an average classification accuracy of $39.15 \pm 0.41\%$ on CIFAR100 with 40% training data, which ranks the fourth place—even lower than the random sampling strategy. The experimental results demonstrate that our DAAL can comprehensively estimate the informativeness of data points and select the samples with larger contributions to the optimization of model for the image classification task.

4.4 Semantic Segmentation on Cityscapes

We illustrate the performance of DAAL and the baseline methods on Cityscapes in Fig. 3. Our DAAL achieves an mIoU of $53.8 \pm 0.24\%$ by using only 40% labeled data, which is comparable to the performance of training on the entire dataset (i.e., $59.42 \pm 0.29\%$). As the proposed DAAL takes both the uncertainty and representativeness into consideration, it outperforms the uncertainty-only AL approach (e.g., VAAL and MC Dropout). Although the SA selects samples based on a combination term of uncertainty and representativeness, its performance is even lower than the uncertainty-only VAAL, which demonstrates the benefit of

the two terms (i.e., uncertainty and representativeness) is not fully exploited. The state-of-the-art AL approach, i.e., LL, which predicts the loss of neural network as its uncertainty estimation, yields a similar accuracy to random sampling on the Cityscapes test set. The reason for this phenomenon is that LL can provide excellent uncertainty estimation only for datasets containing fewer classes (e.g., CIFAR10). As the task difficulty increases, such as more classes (i.e., CIFAR100) or complicated targets (i.e., pixel-wise prediction on Cityscapes), the LL often fails to accurately predict the loss of neural network and consequently selects the less informative samples.

4.5 Performance Analysis

To further evaluate the effectiveness of our DAAL, we conduct experiments to compare the state-of-the-art AL approaches (Learning Loss [41] and VAAL [36]) with the proposed DAAL.

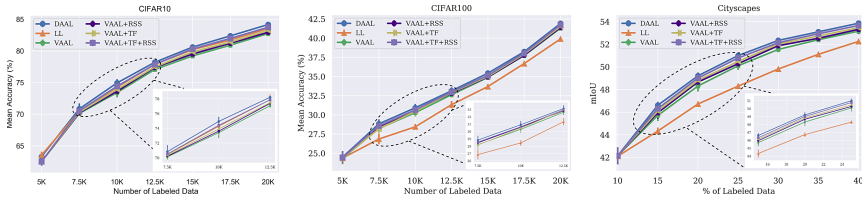

Fig. 4. Performance comparison of DAAL, original VAAL, VAAL using the deep features of the task model, and LL on CIFAR10/100. TF denotes the deep features of the task model and RSS represents the random subset selection.

Influence of Feature Representation in VAAL. We have analyzed the shortcomings of LL and VAAL on CIAFR10 and CIAFR100 and presented some explanations in previous section. As the VAAL does not involve the information of task model for data selection, the comparison between VAAL and our DAAL may be unfair. In this regard, we build a VAAL taking the features encoded by the task model as input (denoted as VAAL + TF), instead of the image data. The evaluation results on the three datasets are presented in Fig. 4. It can be observed that the accuracy of VAAL increases on CIFAR10 by using the features encoded by the task model, which achieves a comparable accuracy to LL. Furthermore, LL [41] constitutes a random subset with the size of 10,000 samples from unlabeled pool during each active learning iteration to ensure the diversity of selected samples, where the K-most uncertain samples are chosen. We evaluate this random subset strategy with VAAL. The variants are denoted as VAAL + RSS and VAAL + TF + RSS, respectively. As illustrated in Fig. 4, the performances of VAAL and VAAL + TF are both improved by using the random subset strategy. However, due to the lack of consideration of representativeness, the performances of VAAL variants are still inferior to our DAAL.

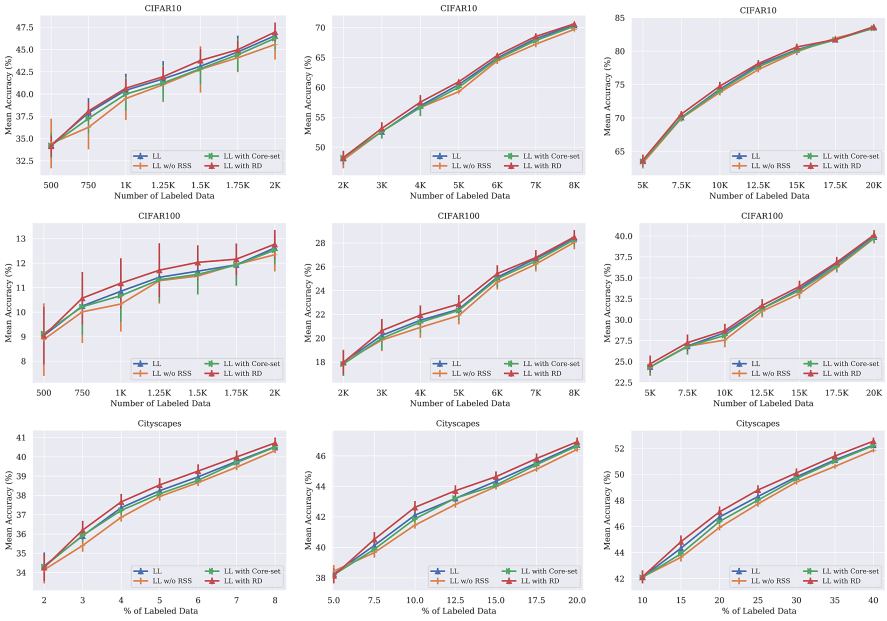

Fig. 5. The influence of random subset selection and using the representativeness discriminator to avoid the overlapping problem in different budget sizes (i.e., the number of samples selected for annotation in each iteration) on CIFAR10 (top), CIFAR100 (middle), and Cityscapes (bottom). The budget sizes are (250, 1000, 2500), (250, 1000, 2500) and (30, 75, 150) for the CIFAR10, CIFAR100, and Cityscapes dataset, respectively. RSS denotes random subset strategy and RD denotes the representativeness discriminator.

Overlapping Problem Occurring in Learning Loss. Several studies [32,41] stated that there was an overlapping problem occurring in the current uncertainty-only AL approaches. In particular, if the uncertainty-only AL approaches are asked to select K samples from the unlabeled pool, the information contained in the K selected samples may be similar, due to the single criterion (i.e., uncertainty to the classifier). When the budget for oracle annotation is small, this overlapping problem tends to be more severe. To verify this intuition, we further conduct experiments evaluating the performance of LL with different budget sizes, which are (250, 1000, 2500), (250, 1000, 2500), and (30, 75, 150) for CIFAR10, CIFAR100, and Cityscapes, respectively. The evaluation results are presented in Fig. 5. It can be observed that the influence caused by the overlapping problem decreases as the budget size increases—the LL and LL without random subset strategy (LL w/o RRS) achieve similar performance with the largest budget size.

To evaluate the benefit generated by integrating the additional criteria (e.g., representativeness) into the process of data selection, our representativeness dis-

criminator is added to LL without RSS, which forms an one-stage framework, denoted as LL + RD. To further evaluate the drawback of multi-stage hybrid approaches, the representative-based Core-set is integrated to LL, denoted as LL + Core-set, which first selects a representative subset by Core-set and then chooses the K-most uncertain samples from them by LL. It can be observed from Fig. 5 that without using RSS, the performance of LL significantly drops due to the overlapping problem. The use of Core-set can alleviate the overlapping problem, which achieves similar accuracies to LL + RSS. Oppositely, the representativeness discriminator remarkably boosts the accuracy of LL w/o RSS, which consistently surpasses the multi-stage approaches (LL + RSS and LL + Core-set) under different settings (e.g., budget size). The experimental results demonstrate the superiority of representativeness tackling the overlapping problem and the one-stage framework, which fully exploits the benefit of representativeness.

4.6 Ablation Study

The contribution of each component of DAAL on CIFAR10/100 and Cityscapes is illustrated in Fig. 6. The performance of DAAL degrades by removing either of the discriminators. The uDis-only DAAL yields similar accuracies to LL, while the rDis-only DAAL can only surpass the random sampling. Hence, the estimation of data informativeness is not comprehensive in either term of uncertainty and representativeness. Our one-stage DAAL is a more proper solution for the data selection during active learning.

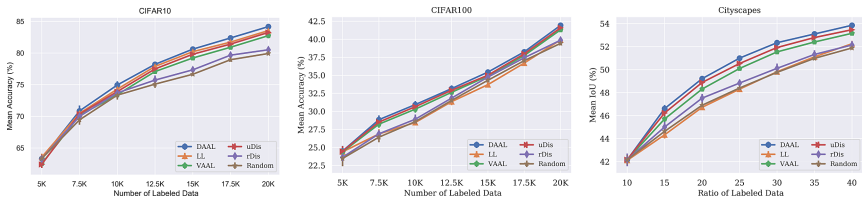

Fig. 6. Impact of the dual discriminators of the proposed DAAL. uDis and rDis represent using uncertainty discriminator or representativeness discriminator only to train the informativeness rater, respectively.

5 Conclusion

In this paper, we proposed a novel one-stage pool-based active learning approach, namely DAAL, which learns to select samples with high uncertainty to the classifier and good representativeness to the unlabeled samples. The proposed AL framework involves a informativeness rater and dual discriminators. Through the adversarial learning between the informativeness rater and discriminators, our

DAAL framework is able to comprehensively estimate the data informativeness to the optimization of the task model. Extensive experiments were conducted on three publicly available datasets (i.e., CIFAR10/100 and Cityscapes). The experimental results showed that our DAAL surpassed the state-of-the-art AL approaches.

References

1. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 37–45 (2015)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5050–5060 (2019)
4. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: Proceedings of the 20th International Conference on Machine Learning (ICML 2003), pp. 59–66 (2003)
5. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
6. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1422–1430 (2015)
7. Ebrahimi, S., Rohrbach, A., Darrell, T.: Gradient-free policy architecture search and adaptation. arXiv preprint arXiv:1710.05958 (2017)
8. Elhamifar, E., Sapiro, G., Yang, A., Shankar Sasrty, S.: A convex optimization framework for active learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 209–216 (2013)
9. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1183–1192. JMLR.org (2017)
10. Gao, M., Zhang, Z., Yu, G., Arik, S.O., Davis, L.S., Pfister, T.: Consistency-based semi-supervised active learning: towards minimizing labeling cost. arXiv preprint arXiv:1910.07153 (2019)
11. Gissin, D., Shalev-Shwartz, S.: Discriminative active learning. arXiv preprint arXiv:1907.06347 (2019)
12. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
13. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 2672–2680 (2014)
14. Gorriz, M., Carlier, A., Faure, E., Giro-i Nieto, X.: Cost-effective active learning for melanoma segmentation. arXiv preprint arXiv:1711.09168 (2017)
15. He, X., et al.: Unsupervised video summarization with attentive conditional generative adversarial networks. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2296–2304 (2019)
16. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745 (2011)

17. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2372–2379. IEEE (2009)
18. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 67–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_5
19. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. In: 2007 IEEE 11th International Conference on Computer Vision (ICCV), pp. 1–8. IEEE (2007)
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech report (2009)
22. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Found. Trends® Mach. Learn. **5**(2–3), 123–286 (2012)
23. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300 (2015)
24. Mackowiak, R., Lenz, P., Ghori, O., Diego, F., Lange, O., Rother, C.: Cereals-cost-effective region-based active learning for semantic segmentation. arXiv preprint arXiv:1810.09726 (2018)
25. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 181–196 (2018)
26. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial LSTM networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 202–211 (2017)
27. Mittal, S., Tatarchenko, M., Çiçek, Ö., Brox, T.: Parting with illusions about deep active learning. arXiv preprint arXiv:1912.05361 (2019)
28. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5898–5906 (2017)
29. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR), pp. 1742–1750 (2015)
30. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 3546–3554 (2015)
31. Sener, O., Savarese, S.: Active learning for convolutional neural networks: a core-set approach. arXiv preprint arXiv:1708.00489 (2017)
32. Sener, O., Savarese, S.: Active learning for convolutional neural networks: a core-set approach. In: International Conference on Learning Representations (2018)
33. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
34. Siddiqui, Y., Valentin, J., Nießner, M.: Viewal: active learning with viewpoint entropy for semantic segmentation. arXiv preprint arXiv:1911.11789 (2019)
35. Singh, A., Virmani, L., Subramanyam, A.: Image corpus representative summarization. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp. 21–29. IEEE (2019)

36. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5972–5981 (2019)
37. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
38. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. **2**, 45–66 (2001)
39. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 399–407. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_46
40. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. Int. J. Comput. Vision **113**(2), 113–127 (2015)
41. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 93–102 (2019)
42. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 472–480 (2017)
43. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
44. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: unsupervised learning by cross-channel prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1058–1067 (2017)
45. Zheng, H., et al.: Biomedical image segmentation via representative annotation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5901–5908 (2019)