# Contextual Diversity for Active Learning

Sharat Agarwal*[1], Himanshu Arora*†[2], Saket Anand[1], and Chetan Arora[3]

[1] IIIT-Delhi, India, {sharata,anands}@iiitd.ac.in
[2] Flixstock Inc., himanshu@flixstock.com
[3] Indian Institute of Technology Delhi, India, chetan@cse.iitd.ac.in

**Abstract.** Requirement of large annotated datasets restrict the use of deep convolutional neural networks (CNNs) for many practical applications. The problem can be mitigated by using active learning (AL) techniques which, under a given annotation budget, allow to select a subset of data that yields maximum accuracy upon fine tuning. State of the art AL approaches typically rely on measures of visual diversity or prediction uncertainty, which are unable to effectively capture the variations in spatial context. On the other hand, modern CNN architectures make heavy use of spatial context for achieving highly accurate predictions. Since the context is difficult to evaluate in the absence of ground-truth labels, we introduce the notion of contextual diversity that captures the confusion associated with spatially co-occurring classes. Contextual Diversity (CD) hinges on a crucial observation that the probability vector predicted by a CNN for a region of interest typically contains information from a larger receptive field. Exploiting this observation, we use the proposed CD measure within two AL frameworks: (1) a core-set based strategy and (2) a reinforcement learning based policy, for active frame selection. Our extensive empirical evaluation establish state of the art results for active learning on benchmark datasets of Semantic Segmentation, Object Detection and Image classification. Our ablation studies show clear advantages of using contextual diversity for active learning. The source code and additional results are available at https://github.com/sharat29ag/CDAL.

## 1 Introduction

Deep convolutional neural networks (CNNs) have acheived state of the art (SOTA) performance on various computer vision tasks. One of the key driving factors for this success has been the effort gone in preparing large amounts of labeled training data. As CNNs become more popular, they are applied to diverse tasks from disparate domains, each of which may incur annotation costs that are task as well as domain specific. For instance, the annotation effort in image classification is substantially lower than that of object detection or semantic segmentation in images or videos. Similarly, annotations of RGB images may be cheaper than

* Equal contribution.
† Work done while the author was at IIIT-Delhi.

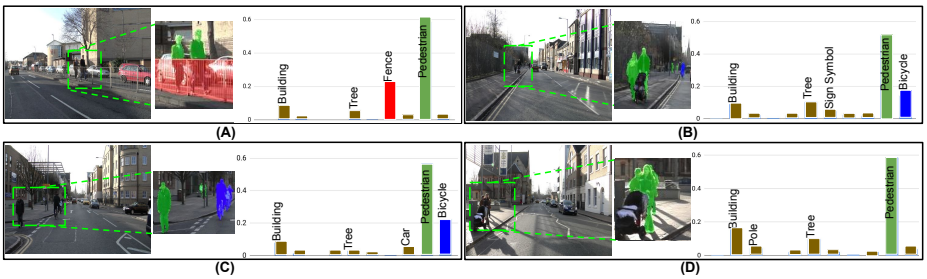



**Fig. 1.** Illustration showing 4 frames from Camvid. Each subfigure shows the full RGB image, region of interest with ground truth overlaid, and the average probability for the 'pedestrian' class with bars color coded by class. We observe that the confusion reflected by the average probability vector corresponding to a class in a frame is also influenced by the object's background. Notice the confusion of pedestrian class with fence in (A) and with bicycle in (C), each of which appear in the neighborhood of a pedestrian instance. We propose a novel *contextual diversity* based measure that exploits the above structure in probability vectors to help select images containing objects in diverse backgrounds. Including this set of images for training helps improving accuracy of CNN-based classifiers, which rely on the local spatial neighborhoods for prediction. For the above example our contextual diversity based selection picks $\{(A), (C), (D)\}$ as opposed to the set $\{(B), (C), (D)\}$ picked by a maximum entropy based strategy (best viewed in color).

MRI/CT images or Thermal IR images, which may require annotators with specialized training.

The core idea of Active Learning (AL) is to leverage the current knowledge of a machine learning model to select most *informative* samples for labeling, which would be more beneficial to model improvement compared to a randomly chosen data point [33]. With the effectiveness of deep learning (DL) based models in recent years, AL strategies have been investigated for these models as well. Here, it has been shown that DL models trained with a fraction of available training samples selected by active learning can achieve nearly the same performance as when trained with all available data [32,42,36]. Since DL models are expensive to train, AL strategies for DL typically operate in a batch selection setting, where a set of images are selected and annotated followed by retraining or fine-tuning of the model using the selected set.

Traditional AL techniques [20,21,34,25,14] have mostly been based on *uncertainty* and have exploited the ambiguity in the predicted output of a model. As most measures of uncertainty employed are based on predictions of individual samples, such approaches often result in highly correlated selections in the batch AL setting. Consequently, more recent AL techniques attempt to reduce this correlation by following a strategy based on the *diversity* and *representativeness* of the selected samples [37,40,18]. Existing approaches that leverage these cues are still insufficient in adequately capturing the spatial and semantic context within an image and across the selected set. Uncertainty, typically measured through entropy, is also unable to capture the class(es) responsible for the resulting un-

certainty. On the other hand, visual diversity and representativeness are able to capture the semantic context across image samples, but are typically measured using global cues in a feature space that do not preserve information about the spatial location or relative placement of the image's constituent objects.

Spatial context is an important aspect of modern CNNs, which are able to learn discriminative semantic representations due to their large receptive fields. There is sufficient evidence that points to the brittleness of CNNs as object locations, or the spatial context, in an image are perturbed [31]. In other words, a CNN based classifier's misclassification is not simply attributed to the objects from the true class, but also to other classes that may appear in the object's spatial neighborhood. This crucial observation also points to an important gap in the AL literature, where existing measures are unable to capture uncertainty arising from the diversity in spatial and semantic context in an image. Such a measure would help select a training set that is diverse enough to cover a *variety of object classes* and their *spatial co-occurrence* and thus improve generalization of CNNs. The objective of this paper is to achieve this goal by designing a novel measure for active learning which helps select frames having objects in diverse contexts and background. Figure 1 describes an illustrative comparison of some of the samples selected by our approach with the entropy based one.

In this paper, we introduce the notion of contextual diversity, which permits us to unify the model prediction uncertainty with the diversity among samples based upon spatial and semantic context in the data. We summarize our contributions below:

- We introduce a novel information-theoretic distance measure, Contextual Diversity (CD), to capture the diversity in spatial and semantic context of various object categories in a dataset.
- We demonstrate that using CD with core-set based active learning [32] almost always beats the state of the art across three visual recognition tasks: semantic segmentation, object detection and image classification. We show an improvement of 1.1, 1.1, and 1.2 units on the three tasks, over the state of the art performance achieving 57.2, 73.3, and 80.9 respectively.
- Using CD as a reward function in an RL framework further improves the AL performance and achieves an improvement of 2.7, 2.1, and 2.3 units on the respective visual recognition tasks over state of the art (57.2, 73.3, and 80.9 respectively).
- Through a series of ablation experiments, we show that CD complements existing cues like visual diversity.

## 2 Related Work

Active learning techniques can be broadly categorized into the following categories. Query by committee methods operate on consensus by several models [2,10]. However, these approaches in general are too computationally expensive to be used with deep neural networks and big datasets. Diversity-based approaches identify a subset of a dataset that is sufficiently representative of the

entire dataset. Most approaches in this category leverage techniques like clustering [30], matrix partitioning [13], or linkage based similarity[3]. Uncertainty based approaches exploit the ambiguity in the predicted output of a model. Some representative techniques in this class include [20,21,34,25,14]. Some approaches attempt to combine both uncertainty and diversity cues for active sample selection. Some notable works in this category include [22,37,40,18]. Recently, generative models have also been used to synthesize informative samples for Active Learning [46,29,28]. In the following, we give a detailed review of three recent state of the art AL approaches applied to vision related tasks. We compare with these methods later in the experiment sections over the three visual recognition tasks.

**Core-Set.** Sener and Savarese [32] have modeled active learning as a *core-set* selection problem in the feature space learned by convolutional neural networks (CNNs). Here, the core-set is defined as a selected subset of points such that the union of $\mathbb{R}^n$-balls of radius $\delta$ around these points contain *all* the remaining unlabeled points. The main advantage of the method is in its theoretical guarantees, which claim that the difference between the loss averaged over all the samples and that averaged over the selected subset does not depend on the *number of samples* in the selected subset, but only on the radius $\delta$. Following this result, Sener and Savarese used approximation algorithms to solve a facility location problem using a Euclidean distance measure in the feature space. However, as was noted by [36], reliance on Euclidean distance in a high-dimensional feature space is ineffective. Our proposed contextual diversity measure relies on KL divergence, which is known to be an effective surrogate for distance in the probability space [6]. Due to distance like properties of our measure, the proposed approach, named *contextual diversity based active learning using core-sets* (CDAL-CS), respects the theoretical guarantees of core-set, yet does not suffer from curse of dimensionality.

**Learning Loss.** Yoo and Kweon [42] have proposed a novel measure of uncertainty by learning to predict the loss value of a data sample. They sampled data based on the ranking obtained on the basis of predicted loss value. However, it is not clear if the sample yielding the largest loss, is also the one that leads to most performance gain. The samples with the largest loss, could potentially be outliers or label noise, and including them in the training set may be misleading to the network. The other disadvantage of the technique is that, there is no obvious way to choose the diverse samples based upon the predicted loss values.

**Variational Adversarial Active Learning (VAAL).** Sinha et al. [36] have proposed to use a VAE to map both the labeled and unlabeled data into a latent space, followed by a discriminator to distinguish between the two based upon their latent space representation. The sample selection is simply based on the output probability of the discriminator. Similar to [42], there seem to be no obvious way to choose diverse samples in their technique based on the discriminator score only. Further, there is no guarantee that the representation learnt by their VAE is closer to the one used by the actual model for the task. Therefore, the

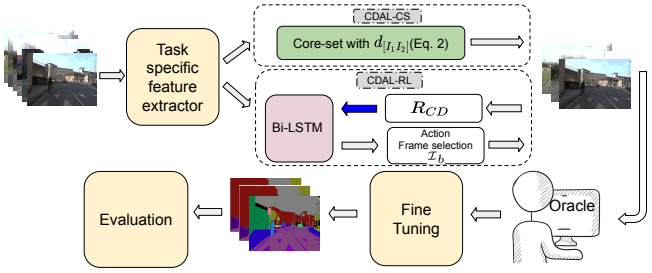



**Fig. 2.** The architecture for the proposed frame selection technique. Two of the strengths of our technique are its unsupervised nature and its generalizability to variety of tasks. The frame selection can be performed in either way by CDAL-CS or CDAL-RL modules. Based on the visual task, a pre-trained model can be readily integrated. The top scoring frames are selected to be annotated and are used to fine tune the model to be evaluated over the main task.

most informative frame for the discriminator need not be the same for the target model as well. Nonetheless, in the empirical analysis, VAAL demonstrates state of the art performance among the other active learning techniques for image classification and semantic segmentation tasks.

**Reinforcement Learning for Active Learning.** Recently, there has been an increasing interest in application of RL based methods to active learning. RALF [7] takes a meta-learning approach where the policy determines a weighted combination of pre-defined uncertainty and diversity measure, which is then used for sample selection. Both [39] and [24] train the RL agents using ground truth based rewards for one-shot learning and person re-identification separately. This requires their method to have a large, annotated training set to learn the policy, and therefore is hard to generalize to more annotation heavy tasks like object detection and semantic segmentation. In [16], an RL framework minimizes time taken for object-level annotation by choosing between bounding box verification and drawing tasks. Fang et al. [9] design a special state space representation to capture uncertainty and diversity for active learning for text data. This design makes it harder to generalize their model to other tasks. Contrary to most of these approaches, our RL based formulation, CDAL-RL, takes a task specific state representation and uses the contextual diversity based reward that combines uncertainty and diversity in an unsupervised manner.

## 3 Active Frame Selection

One of the popular approaches in semi-supervised and self-supervised learning is to use *pseudo-labels*, which are labels as predicted by the current model [4,19,37]. However, directly using pseudo-labels in training, without appropriately accounting for the uncertainty in the predictions could lead to overfitting and confirmation bias [1]. Nonetheless, the class probability vector predicted by

a model contains useful information about the model's discriminative ability. In this section, we present the proposed *contextual diversity* (CD), an information-theoretic measure that forms the basis for Contextual Diversity based Active Learning (CDAL). At the heart of CD is our quantification of the model's predictive uncertainty defined as a mixture of softmax posterior probabilities of pseudo-labeled samples. This mixture distribution effectively captures the spatial and semantic context over a set of images. We then derive the CD measure, which allows us to select diverse and uncertain samples from the unlabeled pool for annotation and finally suggest two strategies for active frame selection. First (CDAL-CS), inspired by the core-set [32] approach and the second (CDAL-RL) using a reinforcement learning framework. An overview of our approach to Active Learning is illustrated in Fig. 2.

### 3.1 Contextual Diversity

Deep CNNs have large receptive fields to capture sufficient spatial context for learning discriminative semantic features, however, it also leads to feature interference making the output predictions more ambiguous [31]. This spatial pooling of features adds to confusion between classes, especially when a model is not fully trained and has noisy feature representations. We quantify this ambiguity by defining the *class-specific confusion*.

Let $C = \{1, \ldots, n_C\}$ be the set of classes to be predicted by a Deep CNN based model. Given a region $r$ within an input image $\mathbf{I}$, let $\boldsymbol{P}_r(\widehat{y} \mid \mathbf{I}; \boldsymbol{\theta})$ be the *softmax probability vector* as predicted by the model $\boldsymbol{\theta}$. For convenience of notation, we will use $\boldsymbol{P}_r$ instead of $\boldsymbol{P}_r(\widehat{y}|\mathbf{I}; \boldsymbol{\theta})$ as the subscript $r$ implies the conditioning on its constituent image $\mathbf{I}$ and the model $\boldsymbol{\theta}$ is fixed in one step of sample selection. These regions could be pixels, bounding boxes or the entire image itself depending on the task at hand. The pseudo-label for the region $r \subseteq \mathbf{I}$ is defined as $\widehat{y}_r = \arg\max_{j \in C} \boldsymbol{P}_r[j]$, where the notation $\boldsymbol{P}_r[j]$ denotes the $j^{th}$ element of the vector. We emphasize that this abstraction of regions is important as it permits us to define overlapping regions within an image and yet have different predictions, thereby catering to tasks like object detection. Let $\mathcal{I} = \cup_{c \in C} \mathcal{I}^c$ be the total pool of *unlabeled* images, where $\mathcal{I}^c$ is the set of images, each of which have at least one region classified by the model into class $c$. Further, let $\mathcal{R}^c_{\mathbf{I}}$ be the set of regions within image $\mathbf{I} \in \mathcal{I}^c$ that are assigned a pseudo-label $c$. The collection of *all* the regions that the model believes belong to class $c$ is contained within the set $\mathcal{R}^c_{\mathcal{I}} = \cup_{\mathbf{I} \in \mathcal{I}^c} \mathcal{R}^c_{\mathbf{I}}$. We assume that for a sufficiently large unlabeled pool $\mathcal{I}$, there will be a non-empty set $\mathcal{R}^c_{\mathcal{I}}$. For a given model $\boldsymbol{\theta}$ over the unlabeled pool $\mathcal{I}$, we now define the class-specific confusion for class $c$ by the following mixture distribution $\boldsymbol{P}^c_{\mathcal{I}}$

$$\boldsymbol{P}^c_{\mathcal{I}} = \frac{1}{|\mathcal{I}^c|} \sum_{\mathbf{I} \in \mathcal{I}^c} \left[ \frac{\sum_{r \in \mathcal{R}^c_{\mathbf{I}}} w_r \boldsymbol{P}_r(\widehat{y} \mid \mathbf{I}; \boldsymbol{\theta})}{\sum_{r \in \mathcal{R}^c_{\mathbf{I}}} w_r} \right] \tag{1}$$

with $w_r \geq 0$ as the mixing weights. While the weights could take any non-negative values, we are interested in capturing the predictive uncertainty of the

model. Therefore, we choose the weights to be the Shannon entropy of $w_r = \mathbb{H}(\boldsymbol{P}_r) = -\sum_{j \in C} \boldsymbol{P}_r[j] \log_2 \boldsymbol{P}_r[j] + \epsilon$, where $\epsilon > 0$ is a small constant and avoid any numerical instabilities. If the model were perfect, $\boldsymbol{P}^c_{\mathcal{I}}$ would be a one-hot encoded vector[1], but for an insufficiently trained model $\boldsymbol{P}^c_{\mathcal{I}}$ will have a higher entropy indicating the confusion between class $c$ and all other classes ($c' \in C, c' \neq c$). We use $\boldsymbol{P}^c_{\mathbf{I}}$ to denote the mixture computed from a single image $\mathbf{I} \in \mathcal{I}^c$.

As discussed in Sec. 1, in CNN based classifiers, this uncertainty stems from spatial and semantic context in the images. For instance, in the semantic segmentation task shown in Fig. 1, the model may predict many pixels as of class 'pedestrian' ($c$ = pedestrian) with the highest probability, yet it would have a sufficiently high probability of another class like 'fence' or 'bicycle'. In such a case, $\boldsymbol{P}^c_{\mathcal{I}}[j]$ will have high values at $j = \{\text{fence, bicycle}\}$, reflecting the chance of confusion between these classes across the unlabeled pool $\mathcal{I}$. As the predictive ability of the model increases, we expect the probability mass to get more concentrated at $j = c$ and therefore reduce the overall entropy of $\boldsymbol{P}^c_{\mathcal{I}}$. It is easy to see that the total Shannon's entropy $h_{\mathcal{I}} = \sum_{c \in C} \mathbb{H}(\boldsymbol{P}^c_{\mathcal{I}})$ reduces with the cross-entropy loss.

Annotating an image and using it to train a model would help resolve the confusion constituent in that image. Based on this intuition, we argue that the annotation effort for a new image is justified only if its inclusion increases the informativeness of the selected subset, i.e., when an image captures a *different kind* of confusion than the rest of the subset. Therefore, for a given pair of images $\mathbf{I}_1$ and $\mathbf{I}_2$, we quantify the disparity between their constituent class-specific confusion by the *pairwise contextual diversity* defined using a symmetric KL-divergence as

$$d_{[\mathbf{I}_1, \mathbf{I}_2]} = \sum_{c \in C} \mathbb{1}^c(\mathbf{I}_1, \mathbf{I}_2) \left(0.5 * \mathrm{KL}(\boldsymbol{P}^c_{\mathbf{I}_1} \parallel \boldsymbol{P}^c_{\mathbf{I}_2}) + 0.5 * \mathrm{KL}(\boldsymbol{P}^c_{\mathbf{I}_2} \parallel \boldsymbol{P}^c_{\mathbf{I}_1})\right). \tag{2}$$

In Eq. (2), $\mathrm{KL}(\cdot \parallel \cdot)$ denotes the KL-divergence between the two mixture distributions. We use the indicator variable denoted by $\mathbb{1}^c(\cdot)$ that takes a value of one only if both $\mathbf{I}_1, \mathbf{I}_2 \in \mathcal{I}^c$, otherwise zero. This variable ensures that the disparity in class-specific confusion is considered only when both images have at least one region pseudo-labeled as class $c$, i.e., when both images have a somewhat reliable measure of confusion w.r.t. class $c$. This confusion disparity accumulated over all classes is the pairwise contextual diversity measure between two images. Given that the KL-divergence captures a distance between two distributions, $d_{[\mathbf{I}_1, \mathbf{I}_2]}$ can be used as a distance measure between two images in the probability space. Thus, using pairwise distances, we can take a core-set [32] style approach for sample selection. Additionally, we can readily aggregate $d_{[\mathbf{I}_m, \mathbf{I}_n]}$ over the selected batch of images, $\mathcal{I}_b \subseteq \mathcal{I}$ to compute the aggregate contextual diversity

$$d_{\mathcal{I}_b} = \sum_{\mathbf{I}_m, \mathbf{I}_n \in \mathcal{I}_b} d_{[\mathbf{I}_m, \mathbf{I}_n]}. \tag{3}$$

[1] We ignore the unlikely event where the predictions are perfectly consistent over the large unlabeled pool $\mathcal{I}$, yet different from the *true* label.

We use this term as the primary reward component in our RL framework. In addition to the intuitive motivation of using the contextual diversity, we show extensive comparisons in Sec. 4 and ablative analysis in Sec. 5.

### 3.2 Frame Selection Strategy

**CDAL-CS.** Our first frame selection strategy is contextual diversity based active learning using core-set (CDAL-CS), which is inspired by the theoretically grounded core-set approach [32]. To use core-set with Contextual Diversity, we simply replace the Euclidean distance with the pairwise contextual diversity (Eq. 2) and use it in the K-Center-Greedy algorithm [32, Algo. 1], which is reproduced in the supplementary material for completeness.

**CDAL-RL.** Reinforcement Learning has been used for frame selection [39,45] for tasks like active one-shot learning and video summarization. We use contextual diversity as part of the reward function to learn a Bi-LSTM-based policy for frame selection. Our reward function comprises of the following three components.

**Contextual Diversity ($R_{cd}$).** This is simply the aggregated contextual diversity, as given in Eq. (3), over the selected subset of images $\mathcal{I}_b$.

**Visual Representation ($R_{vr}$).** We use this reward to incorporate the visual representativeness over the whole unlabeled set using the image's feature representation. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be the feature representations of an image $\mathbf{I}_i \in \mathcal{I}$ and of $\mathbf{I}_j \in \mathcal{I}_b$ respectively, then

$$R_{vr} = \exp\left(-\frac{1}{|\mathcal{I}|}\sum_{i=1}^{|\mathcal{I}|}\min_{j\in\mathcal{I}_b}\left(\|\mathbf{x}_i - \mathbf{x}_j\|_2\right)\right) \tag{4}$$

This reward prefers to pick images that are spread out across the feature space, akin to $k$-medoid approaches.

**Semantic Representation ($R_{sr}$).** We introduce this component to ensure that the selected subset of images are reasonably balanced across all the classes and define it as

$$R_{sr} = \sum_{c\in C}\log\left(|\mathcal{R}^c_{\mathcal{I}_b}|/\lambda\right) \tag{5}$$

Here, $\lambda$ is a hyper-parameter that is set to a value such that a selection that has substantially small representation of a class ($|\mathcal{R}^c_{\mathcal{I}_b}| \ll \lambda$) gets penalized. We use this reward component only for the semantic segmentation application where certain classes (e.g., 'pole') may occupy a relatively small number of regions (pixels).

We define the total reward as $R = \alpha R_{cd} + (1-\alpha)(R_{vr}+R_{sr})$ and use it to train our LSTM based policy network. To emphasize the CD component in the reward function we set $\alpha$ to 0.75 across all tasks and experiments. The precise value of $\alpha$ does not influence results significantly as shown by the ablation experiments reported in the supplementary.

### 3.3 Network Architecture and Training

The contextual diversity measure is agnostic to the underlying *task network* and is computed using the predicted softmax probability. Therefore in Sec. 4, our task network choice is driven by reporting a fair comparison with the state-of-the-art approaches on the respective applications. In the core-set approach [32], images are represented using the feature embeddings and pairwise distances are Euclidean. Contrarily, our representation is the mixture distribution computed in Eqn. (1) over a single image and the corresponding distances are computed using pairwise contextual diversity in Eqn. (2).

For CDAL-RL, we follow a policy gradient based approach using the REINFORCE algorithm [38] and learn a Bi-LSTM *policy network*, where the reward used is as described in the previous section. The input to the policy network at a given time step is a representation of each image extracted using the task network. This representation is the vectorized form of an $n_C \times n_C$ matrix, where the columns of the matrix are set to $\boldsymbol{P}_{\mathbf{I}}^c$ for all $c \in C$ such that $\mathbf{I} \in \mathcal{I}^c$, and zero vectors otherwise. The binary action (select or not) for each frame is modeled as a Bernoulli random variable using a sigmoid activation at the output of the Bi-LSTM policy network. The LSTM cell size is fixed to 256 across all experiments with the exception of image classification, where we also show results with a cell size of 1024 to accommodate for a larger set of 100 classes. For REINFORCE, we use learning rate $= 10^{-5}$, weight decay $= 10^{-5}$, max epoch $= 60$ and #episodes $= 5$. We achieve the best performance when we train the policy network from scratch in each iteration of AL, however, in Sec. 5 we also analyze and compare other alternatives. It is worth noting that in the AL setting, the redundancy within a large unlabeled pool may lead to multiple subsets that are equally good selections. CDAL-RL is no exception and multiple subsets may achieve the same reward, thus rendering the specific input image sequence to our Bi-LSTM policy network, irrelevant.

## 4 Results and Comparison

We now present empirical evaluation of our approach on three visual recognition tasks of semantic segmentation, object detection and image classification[2].

**Datasets.** For semantic segmentation, we use Cityscapes [5] and BDD100K [44]. Both these datasets have 19 object categories, with pixel-level annotation for 10K and 3475 frames for BDD100K and Cityscapes respectively. We report our comparisons using the mIoU metric. For direct comparisons with [42] over the object detection task, we combine the training and validation sets from PASCAL VOC 2007 and 2012 [8] to obtain 16,551 unlabeled pool and evaluate the model performance on the test set of PASCAL VOC 2007 using the mAP metric. We evaluate the image classification task using classification accuracy as the metric over the CIFAR-10 and CIFAR-100 [17] datasets, each of which have 60K images evenly divided over 10 and 100 categories respectively.

[2] Additional results and ablative analysis is presented in the supplementary.

**Compared Approaches.** The two recent works [42,36] showed state of the art AL performance on various visual recognition tasks and presented a comprehensive empirical comparison with prior work. We follow their experimental protocol for a fair comparison and present our results over all the three tasks. For the semantic segmentation task, we use the reported results for VAAL and its other competitors from [36], which are core-set [32], Query-by-Committee (QBC) [18], MC-Dropout [12] and Suggestive Annotation (SA) [41]. We refer to our contextual diversity based approaches as CDAL-CS for its core-set variant and CDAL-RL for the RL variant, which uses the combined reward $R$ as defined in Sec.3.2. The object detection experiments are compared with learn loss [42] and its competitors – core-set, entropy based and random sampling – using results reported in [42]. For the image classification task, we again compare with VAAL, core-set, DBAL [11] and MC-Dropout. All the CDAL-RL results are reported after averaging over three independent runs. In Sec. 5 we demonstrate the strengths of $CD$ through various ablative analysis on the Cityscapes dataset. Finally, in the supplementary material, we show further comparisons with region based approaches [15,27], following their experimental protocol on the Cityscapes dataset.

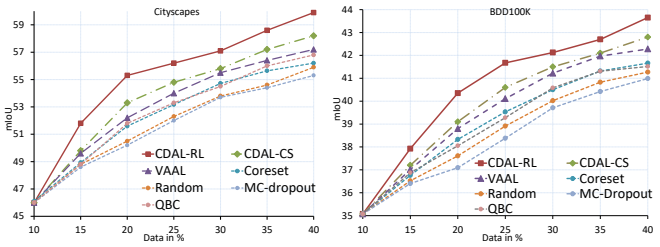



**Fig. 3.** Quantitative comparison for the Semantic Segmentation problem over Cityscapes **(left)** and BDD100K **(right)**. Note: DRN results 62.95% and 44.95% mIoU on 100% data for Cityscapes and BDD respectively (best viewed in color).

### 4.1 Semantic Segmentation

Despite the tediousness associated with semantic segmentation, there are limited number of works for frame-level selection using AL. A recent approach applied to this task is VAAL [36], which achieves state-of-the-art performance while presenting a comprehensive comparison with previously proposed approaches. We follow the experimental protocol of VAAL [36], and also use the same backbone model of dilated residual network (DRN) [43] for a fair comparison. As in their case, the annotation budget is set to 150 and 400 for Cityscapes and BDD100K respectively. The evaluation metric is mIoU. For each dataset, we evaluate the performance of an AL technique at each step, as the number of samples to be

selected are increased from 10% to 40% of the unlabeled pool size, in steps of 5%. Fig. 3 shows the comparison over the two datasets.

We observe that for both challenging benchmarks, the two variants of CDAL comprehensively outperform all other approaches by a significant margin. Our CDAL-RL approach can acheive current SOTA 57.2 and 42.8 mIoU by reducing the labeling effort by 300 and 800(10%) frames on cityscapes and BDD100k respectively. A network's performance on this task is the most affected by the spatial context, due to the fine-grained spatial labeling necessary for improving the mIoU metric. We conclude that the CD measure effectively captures the spatial and semantic context and simultaneously helps select the most informative samples. There exist region-level AL approaches to semantic segmentation, where only certain regions are annotated in each frame [15,27]. Our empirical analysis in the supplementary material shows that our CDAL based frame selection strategy is complementary to the region-based approaches.

### 4.2 Object Detection

For the object detection task, we compare with the learning loss approach [42] and the competing methods therein. For a fair comparison, we use the same base detector network as SSD [23] with a VGG-16 [35] backbone and use the same hyperparameter settings as described in [42].

Fig. 4 shows the comparisons, where we see in most cases, both variants of CDAL perform better than the other approaches. During the first few cycles of active learning, i.e., until about 5K training samples are selected for annotation, CDAL performs nearly as well as coreset, which outperforms all the other approaches. In the later half of the active learning cycles with 5K to 10K selected samples, CDAL variants outperform all the other approaches including coreset [32]. CDAL-RL achieved 73.3 mAP using 8k data where learning loss [42] achieved it by 10k hence reducing 2k labeled samples.

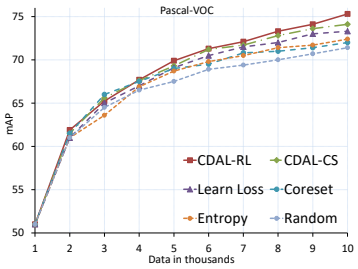



**Fig. 4.** Quantitative comparison for Object Detection over PASCAL-VOC dataset. We follow the experimental protocol of the learning loss method [42]. Note: SSD results 77.43% mAP on 100% data of PASCAL-VOC(07+12).

### 4.3 Image Classification

One of the criticisms often made about the active learning techniques is their relative difficulty in scaling with the number of classes in a given task. For example it has been reported in [36], that core-set [32] does not scale well for large number of classes. To demonstrate the strength of contextual diversity cues when the number of classes is large, we present the evaluations on the image classification task using CIFAR-10 and CIFAR-100. Fig. 5 shows the comparison.

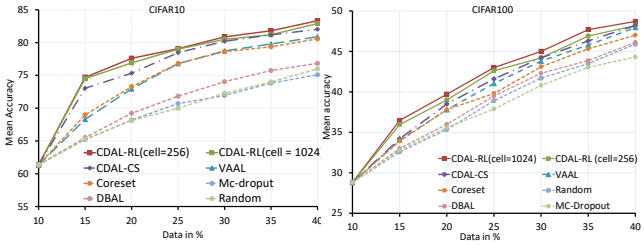



**Fig. 5.** Quantitative comparison for Image Classification over CIFAR-10 **(left)** and CIFAR-100 **(right)**. CDAL-RL(cell=$n$) indicates that the LSTM policy network has a cell size $n$. Note: VGG16 results 90.2% and 63.14% accuracy on 100% data for CIFAR10 and CIFAR100 respectively (best viewed in color).

It is clear that CDAL convincingly outperforms the state of the art technique, VAAL [36] on both the datasets. We can see that CDAL-RL can achieve ∼81% accuracy on CIFAR10 by using 5000 (10%) less samples than VAAL [36] and similarly 2500 less samples are required in CIFAR100 to beat SOTA of 47.95% accuracy. These results indicate that CDAL can gracefully scale with the number of classes, which is not surprising as CD is a measure computed by accumulating KL-divergence, which scales well with high-dimensions unlike the Euclidean distance. It is worth noting that an increase in the LSTM cell size to 1024, helps improve the performance on CIFAR-100, without any significant effect on the CIFAR-10 performance. A higher dimension of the LSTM cell has higher capacity which better accommodates a larger number of classes. For completeness, we include more ablations of CDAL for image classification in the supplementary.

We also point out that in image classification, the entire image qualifies as a *region* (as defined in Sec. 3.1), and the resulting mixture $\boldsymbol{P}_{\mathbf{I}}^{c}$ comprising of a single component still captures confusion arising from the spatial context. Therefore, when a batch $\mathcal{I}_b$ is selected using the contextual diversity measure, the selection is diverse in terms of classes and their confusion.

## 5 Analysis and Ablation Experiments

In the previous section, we showed that contextual diversity consistently outperforms state of the art active learning approaches over three different visual recognition tasks. We will now show a series of ablation experiments to demonstrate the value of contextual diversity in an active learning problem. Since active learning is expected to be the most useful for the semantic segmentation task with highest amount of annotation time per image, we have chosen the task for our ablation experiments. We have designed all our ablation experiments on the Cityscapes dataset using the DRN model in the same settings as in Sec. 4.1.

**Reward Component Ablation**. We first investigate the performance of various components of the reward used in our approach. Fig. 6(left) shows the performance of CDAL in three different reward settings: only contextual diversity ($R = R_{cd}$), contextual diversity and visual representation (CD+VR, i.e.,

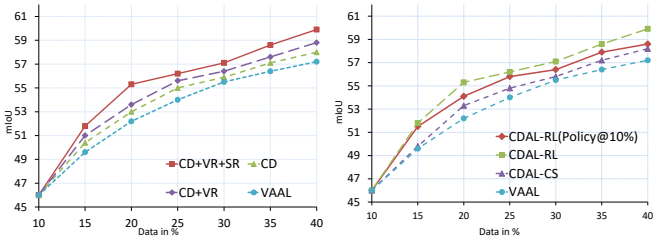


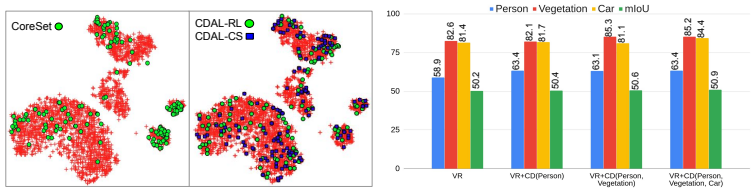

**Fig. 6. (left)** Ablation with individual reward components on Cityscapes. **(right)** Cityscapes results when the CDAL-RL policy was learned only once in the first iteration with 10% randomly selected frames (best viewed in color).

$R = \alpha R_{cd} + (1-\alpha)R_{vr}$) and all the three components including semantic representations (CD+VR+SR, i.e., $R = \alpha R_{cd} + (1-\alpha)(R_{vr} + R_{sr})$). It is clear that contextual diversity alone outperforms the state of the art, VAAL [36], and improves further when the other two components are added to the reward function. As mentioned in Sec. 3.2, the value of $\alpha = 0.75$ was not picked carefully, but only to emphasize the CD component, and remains fixed in all experiments.

**Policy Training Analysis.** Our next experiment analyzes the effect of learning the Bi-LSTM-based policy only once, in the first AL iteration. We train the policy network using the *randomly selected* 10% and use it in each of the AL iterations for frame selection without further fine-tuning. The results are shown in Fig. 6(right), where we can see that this policy denoted by CDAL-RL(policy@10%), still outperforms VAAL and CDAL-CS in *all* iterations of AL. Here CDAL-RL is the policy learned under the setting in Sec. 4.1, where the policy network is trained from scratch in each AL iteration. An interesting observation is the suitability of the contextual diversity measure as a reward, and that it led to learning a meaningful policy even with randomly selected data.

**Visualization of CDAL-based Selection.** In Fig. 7(a), we show t-SNE plots [26] to visually compare the distribution of the points selected by the CDAL variants and that of core-set. We use the Cityscapes training samples projected into the feature space of the DRN model. The red points in the plots show the unlabeled samples. The left plot shows green points as samples selected by core-set and the right plot shows green and blue points are selected by CDAL-RL and CDAL-CS respectively. It is clear that both variants of the contextual diversity based selection have better spread across the feature space when compared with the core-set approach, which is possibly due to the distance concentration effect as pointed by [36]. This confirms the limitation of the Euclidean distance in a high-dimensional feature space corresponding to the DRN model. On the other hand, CDAL selects points that are more uniformly distributed across the entire feature space, reflecting better representativeness capability.

**Class-wise Contextual Diversity Reward.** The CD is computed by accumulating the symmetric KL-divergence (cf. Eq. (2)) over all classes. Therefore, it is possible to use the $R_{cd}$ reward only for a few, and not all, the classes. Fig. 7(b) shows the segmentation performance as we incorporate the contextual diversity




(a) (b)

**Fig. 7.** (a) t-sne plots comparison with [32] on Cityscapes: CoreSet (left), and CDAL (right) (b) Performance analysis when CD reward is computed for an increasing number of classes.(best viewed in color)

(CD) from zero to three classes. The initial model is trained using only the visual representation reward ($R_{vr}$) and is shown as the leftmost group of color-coded bars. As we include the $R_{cd}$ term in the reward with the CD only being computed for the *Person* class, we see a substantial rise in the IoU score corresponding to *Person*, as well as a marginal overall improvement. As expected, when we include both, the *Person* and *Vegetation* classes in the CD reward, we see substantial improvements in both the classes. The analysis indicates that the contextual diversity reward indeed helps mitigating class-specific confusion.

**Limitations of CDAL.** While we show competitive performance of the policy without retraining (Fig.6(right)), for best performance retraining at each AL iteration is preferred. For large datasets, this requires larger unrolling of the LSTM network incurring more computational and memory costs. Another limitation of CDAL in general is in the case of image classification, where the entire image is treated as a single region and thus is unable to fully leverage spatial context.

## 6 Conclusion

We have introduced a novel contextual diversity based measure for active frame selection. Through experiments for three visual recognition tasks, we showed that contextual diversity when used as a distance measure with core-set, or as a reward function for RL, is a befitting choice. It is designed using an information-theoretic distance measure, computed over the mixture of softmax distributions of pseudo-labeled data points, which allows it to capture the model's predictive uncertainty as well as class-specific confusion. We have only elaborated the promising empirical results in this paper, and plan to investigate deeper theoretical interpretations of contextual diversity that may exist.

### Acknowledgement

The authors acknowledge the partial support received from the Infosys Center for Artificial Intelligence at IIIT-Delhi. This work has also been partly supported by the funding received from DST through the IMPRINT program (IMP/2019/000250).

## References


1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning (2019) 5
2. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9368–9377 (2018) 3
3. Bilgic, M., Getoor, L.: Link-based active learning. In: NIPS Workshop on Analyzing Networks and Learning with Graphs (2009) 4
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision (2018) 5
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 9
6. Dabak, A.G.: A geometry for detection theory. In: PhD Thesis, Rice Unviersity (1992) 4
7. Ebert, S., Fritz, M., Schiele, B.: Ralf: A reinforced active learning formulation for object class recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3626–3633 (2012) 5
8. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 9
9. Fang, M., Li, Y., Cohn, T.: Learning how to active learn: A deep reinforcement learning approach. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 595–605 (2017) 5
10. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1183–1192. JMLR. org (2017) 3
11. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1183–1192. JMLR. org (2017) 10
12. Gorriz, M., Carlier, A., Faure, E., Giro-i Nieto, X.: Cost-effective active learning for melanoma segmentation. arXiv preprint arXiv:1711.09168 (2017) 10
13. Guo, Y.: Active instance sampling via matrix partition. In: Advances in Neural Information Processing Systems. pp. 802–810 (2010) 4
14. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2372–2379. IEEE (2009) 2, 4
15. Kasarla, T., Nagendar, G., Hegde, G., Balasubramanian, V., Jawahar, C.: Region-based active learning for efficient labeling in semantic segmentation. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1109–1118 (Jan 2019). 10, 11
16. Konyushkova, K., Uijlings, J., Lampert, C.H., Ferrari, V.: Learning intelligent dialogs for bounding box annotation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 5
17. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009) 9
18. Kuo, W., Häne, C., Yuh, E.L., Mukherjee, P., Malik, J.: Cost-sensitive active learning for intracranial hemorrhage detection. In: Medical Image Computing and

Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III. pp. 715–723 (2018) 2, 4, 10
19. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning (WREPL) (2013) 5
20. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine learning proceedings 1994, pp. 148–156. Elsevier (1994) 2, 4
21. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR94. pp. 3–12. Springer (1994) 2, 4
22. Li, X., Guo, Y.: Adaptive active learning for image classification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 859–866 (2013) 4
23. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision (ECCV) 11
24. Liu, Z., Wang, J., Gong, S., Lu, H., Tao, D.: Deep reinforcement active learning for human-in-the-loop person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 5
25. Luo, W., Schwing, A., Urtasun, R.: Latent structured active learning. In: Advances in Neural Information Processing Systems. pp. 728–736 (2013) 2, 4
26. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008) 13
27. Mackowiak, R., Lenz, P., Ghori, O., Diego, F., Lange, O., Rother, C.: CEREALS - cost-effective region-based active learning for semantic segmentation. In: British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018 (2018) 10, 11
28. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Reyes, M.: Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 580–588. Springer (2018) 4
29. Mayer, C., Timofte, R.: Adversarial sampling for active learning. arXiv preprint arXiv:1808.06671 (2018) 4
30. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proceedings of the twenty-first international conference on Machine learning. p. 79. ACM (2004) 4
31. Rosenfeld, A., Zemel, R.S., Tsotsos, J.K.: The elephant in the room. CoRR **abs/1808.03305** (2018) 3, 6
32. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018) 2, 3, 4, 6, 7, 8, 9, 10, 11, 14
33. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1), 1–114 (2012) 2
34. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1070–1079. Association for Computational Linguistics (2008) 2, 4
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015) 11

36. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 2, 4, 10, 11, 12, 13
37. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Trans. Cir. and Sys. for Video Technol. **27**(12), 2591–2600 (2017) 2, 4, 5
38. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**(34), 229256 (1992) 9
39. Woodward, M., Finn, C.: Active one-shot learning. In: NIPS Deep RL Workshop (2017) 5, 8
40. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 399–407. Springer (2017) 2, 4
41. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 399–407. Springer (2017) 10
42. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 2, 4, 9, 10, 11
43. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 472–480 (2017) 10
44. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 (2018) 9
45. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward (2018) 8
46. Zhu, J.J., Bento, J.: Generative adversarial active learning. arXiv preprint arXiv:1702.07956 (2017) 4