Uncertainty and Diversity in Deep Active Image Classification

Hariank Muthakana

CMU-CS-19-132

December 2019

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Aarti Singh, Chair
Barnabás Póczos

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Keywords: Active Learning, Computer Vision

Abstract

Deep neural networks have revolutionized computer vision, with state-of-the art performance across multiple tasks. An important part of training such networks is the availability of large, high-quality labeled datasets. This makes building new datasets a significant hurdle to approaching novel tasks or domains. In many cases, acquiring labels can be difficult, expensive, or time-consuming. Active learning seeks to improve label efficiency and lower overall labeling cost by allowing the learning system to intelligently pick samples to label. Active learning is well studied for classical machine learning models, but many of these approaches have been shown to be ineffective for deep models and modern image datasets. This raises the question of how to develop and use active strategies in these settings. In this work, we seek to build intuitions for deep active learning by conducting a comprehensive empirical analysis of existing approaches for image classification tasks. Critical to this analysis is the distinction between uncertainty and diversity-based strategies and how they perform in various settings. Our experiments show surprising results regarding the efficacy of existing approaches in commonly tested settings. We find that active learning is more useful in settings such as low data availability, class imbalance, and transfer learning. Finally, our results provide heuristics for the active learning practitioner to decide on a strategy to use, and more crucially whether to use active learning at all.

Acknowledgments

I would first like to thank my wonderful advisor Prof. Aarti Singh. Having worked with her since Fall 2017 on various projects, I can attest to her amazing mentorship and guidance, as well as her patience for working with undergraduate students. Her strong theoretical knowledge was consistently helpful in times when I felt lost. It was an absolute privilege to work with Prof. Singh and I hope to collaborate with her again in the future.

I would also like to thank Prof. Michael Tarr and the members of TarrLab. I worked on an additional research project involving neuroscience and machine learning in parallel to this project, and I found it immensely rewarding. TarrLab introduced me to fascinating problems in biologically-inspired modeling that I will forever find intriguing, and taught me how to be a better researcher.

Thank you to Barnabás Póczos for serving on my thesis committee and giving me great feedback. Thank you to my advisor Tracy Farbacher for answering my endless questions about the Masters program and this thesis. Finally, thank you to my friends and family for their constant love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning models have shown incredible ability for inference and prediction across a variety of domains. However, this ability is extremely dependent on the amount and quality of training data. As model complexity increases, so does the amount of training samples required. And while samples may be easy to collect, such as through crawling the web [50], in supervised learning we require a large number of labels as well.

Building such large datasets has become even more important with recent advances in deep neural networks. Although they have achieved successful results in high complexity sample domains like natural language and images, deep networks often require tens of thousands of examples or more. ImageNet, a popular large image classification dataset, currently contains 14 million images with accompanying labels collected through crowdsourcing [12]. Collecting such a labeled dataset for a novel task can thus be difficult, particularly if labeling is time-intensive or expensive. For example, medical images may require an expert opinion in order to obtain a label [1]. Active learning seeks to solve this problem by allowing learning systems to query labels for unlabeled samples during the learning process, and has shown successful results in many tasks including speech recognition, classification, and filtering [41].

There have been several results for active learning with classical machine learning models. One paradigm is uncertainty sampling, where the algorithm queries labels for samples that the model is uncertain about. Other paradigms include query-by-committee, in which the algorithm trains several models and queries the samples for which they disagree most, and density-weighted sampling, which queries the most representative samples in the input space. Of these, uncertainty sampling has been the most successful and commonly used [41].

However, many of these intuitions, paradigms, and results have not been successful for deep models. One issue is that deep models are unlikely to learn from single data points and generally require mini-batch learning to train efficiently and avoid local minima. This requires designing active algorithms which can query several samples at once. In addition, many uncertainty sampling methods have been specifically shown to perform poorly in the deep setting. Methods that have shown the most promise [40, 47] instead query diversely throughout the input space, which is unlike most of the existing paradigms. Even so, it is still not clear which active learning methods a practitioner would use with a deep model. In addition, evaluating active learning algorithms for deep networks is difficult due to long training times and noisy results. And critically, nearly all recent work trains in restricted settings which are often not the most realistic or

practical.

In this work, we empirically analyze deep active image classification algorithms across several different settings in order to address these concerns. We particularly focus on the differences between uncertainty and diversity-based methods across these settings. Additionally, we argue that the most common settings in which these algorithms are tested are not optimal for active learning, and explore what we believe are more interesting and practical settings.

Our main contributions are: (a) identifying settings where uncertainty-based active methods outperform diversity-based methods, (b) analyzing the effect of learned model representations on diversity-based methods, (c) understanding how deep active methods perform in common classification tasks of varying difficulty, and (d) helping the active learning practitioner identify the usefulness of active learning strategies in their particular setting.

Chapter 2

Background

In this section we provide an overview of existing active learning paradigms and approaches, both for classical models as well as deep models. In an active learning system, the learner queries for new labels in a systematic way to achieve the best performance. In many cases, we can achieve better performance with smart label queries than random selections, and indeed we evaluate these systems against a random sampling baseline ("passive" learning).

There are several high-level scenarios for active learning. In pool-based active learning, the most common scenario, samples come from a fixed unlabeled pool. When the algorithm queries a sample and receives a label, the sample moves to the labeled pool. Other scenarios exist, such as query-synthesis, in which the system can query *any sample* from the input space (even synthetically generated ones), and stream-based, in which the system receives candidate samples in a stream and chooses whether or not to label them [41].

We focus on pool-based active learning in this work. At each active acquisition step, we query the new sample(s) to label from the unlabeled pool. These samples are moved to the labeled pool, and the process is repeated with the next step

2.1 Classical Active Learning

There have been many successful active learning results for classical models. A common category of methods is uncertainty sampling, in which we query samples the learner is unsure about. These strategies implicitly select points near decision boundaries in order to improve performance. For models that estimate class probabilities, margin [39], and entropy [45] sampling are common ways of computing uncertainty. For maximum-margin classifiers, we can directly estimate distance to the margin [9, 49]. Uncertainty-based strategies also exist for nonparametric models such as decision trees [32] and nearest-neighbor classifiers [18, 33].

A related category of strategies is query-by-committee (QBC), where several learners are trained and the samples that the learners most disagree on are selected [44]. QBC strategies have been explored for naive Bayes [35] and hidden Markov models [3]. Several model-agnostic methods also exist [2, 36, 37]. Often these methods employ methods similar to uncertainty sampling in order to compute disagreement, such as [3] which extends entropy sampling to a committee of models. An issue with QBC methods is that we must train multiple models, but

[35, 42, 44] suggest that using as few as two models is sufficient.

Another important category of strategies is density-weighting, where the learner queries the most "representative" samples. This involves modeling the input space, which can be challenging, but avoids the problem of querying unnecessary outliers (which uncertainty sampling may do). It also allows for possibly leveraging the unlabeled samples, turning the active strategy into a form of semi-supervised learning. [42] employs a density-weighted strategy that estimates density using average similarity to nearby samples, and combines it with uncertainty sampling. The formulation allows for arbitrary uncertainty methods to be used, and for trading off between representativeness and uncertainty. [52] performs a similar tradeoff by using an integer programming formulation to querying informative samples while querying close to the input data distribution.

Other strategies aim to find samples that most improve expected error or induce the greatest model change. [29] introduces a Bayesian method for Gaussian processes to find queries that maximize expected model improvement. [43] uses expected gradient length as a proxy for model change. [38] proposes an error reduction strategy for naive Bayes. Again, efficiency is often a concern with some of these strategies, as we may have to retrain the model for every candidate query.

A significant characteristic of classical methods is strong theoretical results. One specific idea we are interested in in this work is understanding the conditions under which active learning algorithms will provide a potential advantage. [10] studies upper and lower bounds for active classification improvement for nonparametric active methods, as functions of data dimensionality, decision boundary complexity, and noise around the boundaries. Critically, they show that potential improvement degrades exponentially as dimensionality d or noise κ increases. [6] extends this result to parametric methods and shows a better lower bound on active classification improvement, but with a similar dependence on the dimension d. Together both works suggest that increased task difficulty, represented by higher data dimensionality and more complex boundaries, lowers the effectiveness of active learning. The various settings we study in this work are motivated by this idea.

A full overview of classical active learning techniques is out of the scope of this work, and we refer the reader to [41] for a comprehensive survey.

2.2 Deep Active Learning

As we move to settings with deep models and complex datasets, many of these active strategies have been empirically shown to break down. Several studies [7, 40] find that uncertainty sampling methods, are ineffective (i.e. unable to outperform random sampling) for nontrivial datasets. It has been shown that class scores from softmax outputs are often poor estimates of probability (the "calibration problem") [23]. This cripples uncertainty sampling strategies that leverage these scores like entropy and margin sampling. [22] finds that an expected gradient length strategy significantly underperforms for image datasets.

[40] argues that mini-batch training in deep learning is a large part of the reason why classical active learning algorithms underperform. Most classical methods query samples in a serial manner, but deep active strategies must query in batches since networks have difficulty learning

from single samples. Furthermore, within these batches, we would like the samples to be uncorrelated for the networks to learn well. Another issue is efficiency. Some methods are suited to small datasets and have undesirable time or space complexity. For example, [15, 24, 56] require optimization over $O(n^2)$ variables where n is the dataset size.

2.2.1 Diversity-based approaches

A recent trend in deep active methods is querying diversely throughout the sample space. This direct optimization of sample diversity is claimed to improve learning, and runs contrary to the uncertainty sampling intuitions that have been successful in classical methods. Coreset [40] formulates this approach geometrically through the k-Center problem [53]. The goal is to find a subset of points, which they call a "core-set", such that for all points, the maximum distance to the closest selected point is minimized. If X_u and X_ℓ are the set of unlabeled and labeled samples respectively, we aim to find a set $S \subseteq X_u$ of size k such that:

$$\min_{S \subseteq X_u} \max_i \min_{x_j \in X_u \cup X_\ell} \Delta(x_i, x_j)$$

The method includes a greedy solution as well as a robust solution with mixed integer programming - we use the greedy solution following the recommendation of [7], which showed that the performance difference was negligible. Coreset incorporates the current model (and by extension, the current labeled pool) by defining $\Delta(x_i, x_j)$ as the Euclidean distance between last-layer embeddings of x_i and x_j.

Other diversity-based methods seek to model the unlabeled sample space more directly, leveraging the large set of unlabeled samples. These methods are similar to density-sampling as they aim to query in areas of the sample space where labels are sparse. [22] train a binary classifier to distinguish between unlabeled and labeled samples. The samples that are predicted to be most likely from the unlabeled set are selected. [47] extends this by using the classifier as a discriminator. The discriminator is trained adversarially with a variational autoencoder (VAE) [30] - the VAE aims to trick the discriminator into classifying both unlabeled and labeled samples as labeled, while the discriminator aims to distinguish the two.

2.2.2 Uncertainty-based approaches

A few uncertainty-based methods have also been proposed for deep models. [14] aims to query samples near decision boundaries by taking labeled samples, finding their adversarial attacks, and querying unlabeled samples near the attacks. Combining Bayesian methods with deep models is a promising direction, as such methods have been sparsely used with classical models successfully with the exception of Gaussian processes. One recent approach [19, 20, 21] obtains posterior uncertainties using dropout masks. This method, called MC-dropout, involves getting uncertainty estimates by setting several random dropout masks and averaging the resulting network outputs. This approximates a query-by-committee strategy as the collection of masks implicitly defines an ensemble of models.

[7], which we call Ensembles, extends MC-dropout by explicitly using an ensemble of models. They experiment with various uncertainty estimation methods to compute disagreement, and

find that the best performing one is the *variation ratio*, defined as the proportion of predictions from the ensemble that are not equal to the modal prediction. If p_m is the most common class prediction for a sample across the ensemble and N is the number of networks, the uncertainty estimate for the sample would be

$$1 - \frac{p_m}{N}$$

The algorithm simply ranks unlabeled samples according to this metric and picks the top k. Ensembles has been shown to outperform MC-dropout and a variety of classical uncertainty sampling methods.

2.2.3 Hybrid approaches

Finally, some works have explored hybrid methods that incorporate both diversity and uncertainty. [27] combines uncertainty estimation with softmax entropy with an "informativeness" metric for pretrained networks. [4] computes embeddings for each sample based on induced gradients, and then uses k-means sampling to geometrically pick diversely from the space, similar to Coreset. They argue that since the gradient space gives information on both the magnitude and direction, picking diversely from this space yields useful, yet diverse sample queries. Such hybrid approaches are promising but have failed to outperform non-hybrid methods in standard deep settings. In this work we mainly focus on the distinction between purely diversity and uncertainty-based approaches.

2.3 Class Imbalance

An interesting setting for active learning is class imbalance, where one more more classes is rarer than the others. Nearly all practical classification problems are inherently imbalanced to some degree. In some cases, we are not even aware that classes are imbalanced, such as when there are hidden patterns in the data or when our selection of classes is poorly defined.

As imbalance increases, a carelessly trained model can achieve high accuracy on majority classes while ignoring minority classes. Although we can track metrics that tease out performance between classes like area under curve (AUC) and average precision (AP), it is still often nontrivial to improve minority class performance. A common approach is to force a balanced class distribution by oversampling minority classes and/or undersampling majority classes. However, both of these can cause issues - overfitting on minority class samples [11], or losing valuable majority class information [13, 28].

Active learning methods have high potential in this setting. A common approach for improving performance is to collect more data for the minority class. This "mining" problem involves finding probable minority class samples without labels, a clear application for active learning. Active learning could also be used to implicitly undersample the majority class by identifying the most useful samples. In this way, classical active methods are able to be used out-of-the-box in imbalanced scenarios. [5] and [17] show that SVM-based active learning strategies are able to query minority samples without any modifications. However, there are some issues with directly

using active methods. As imbalance becomes more harsh, minority samples may get missed entirely, worsening the problem for future label acquisitions. Several works propose to explicitly encourage minority selections. [8] modify the SVM loss to weight classes differently. [57] and [16] adaptively oversample minority class samples through query synthesis.

Once again, however, we see issues when moving to the deep setting. The problems that arise with imbalanced settings are exacerbated in data-hungry deep networks that need to learn highly nonlinear decision boundaries. Classical active learning methods, even those specifically designed for imbalanced datasets, have shown difficulty when applied to deep learning [5]. Certain task-specific deep methods have succeeded - for example, object detection is a common highly imbalanced task between positive (is an object) and negative (not an object) classes. Hard example mining [46] has been extensively studied for detection and related tasks, and can be seen as a form of active learning. Active learning has also been applied to the problem of fairness by emphasizing less-represented classes using diversity-based sampling [54, 55]. However, it is unclear how recent deep active methods for classification perform in this setting, as they usually test on balanced, standard datasets. We seek intuitions on how uncertainty and diversity-based methods perform in imbalanced settings in Chapter 4.4.

Chapter 3

Simple Classification Settings

3.1 Introduction

As mentioned in Chapter 2, a common intuition in classical active learning is to select points near decision boundaries. However, this becomes difficult when working with high-complexity data as decision boundaries are highly complex, and we have to simultaneously learn a representation for the data. We can alleviate this problem by studying settings where the classification task is easy enough for a network to learn a discriminative representation with few samples and where decision boundaries are simple. In this chapter, we focus on a simple synthetic dataset as well as two image classification datasets, and investigate the ability of uncertainty-based and diversity-based active strategies to identify samples near decision boundaries.

3.2 Experimental Setup

We briefly describe here the specific pool-based framework we use for experiments in this chapter and the following chapter. An initial labeled pool, which we denote as the *warmup set* is selected randomly from all samples. These are the only labels available before any active queries. At each step, we select the same number of samples from the unlabeled pool, denoted as the *query size*, and receive labels for them. Unless otherwise specified, the warmup set is initialized with the same number of samples as the query size. The model is initially trained on the warmup set, and the active algorithm has access to this trained model when selecting samples to query. We record the number of samples in the current labeled pool, and evaluate the trained model on a held-out test set. After a new set of labels is queried, this model is discarded, and a fresh model is trained on the new labeled set. The process continues until we are satisfied with the model's performance. We run each active strategy with 3 random seeds to reduce variability.

3.3 Synthetic Data

In order to motivate our work in this setting, we experiment with deep active algorithms on a synthetic dataset consisting of 5000 points sampled from two equal-variance 2D Gaussians, as

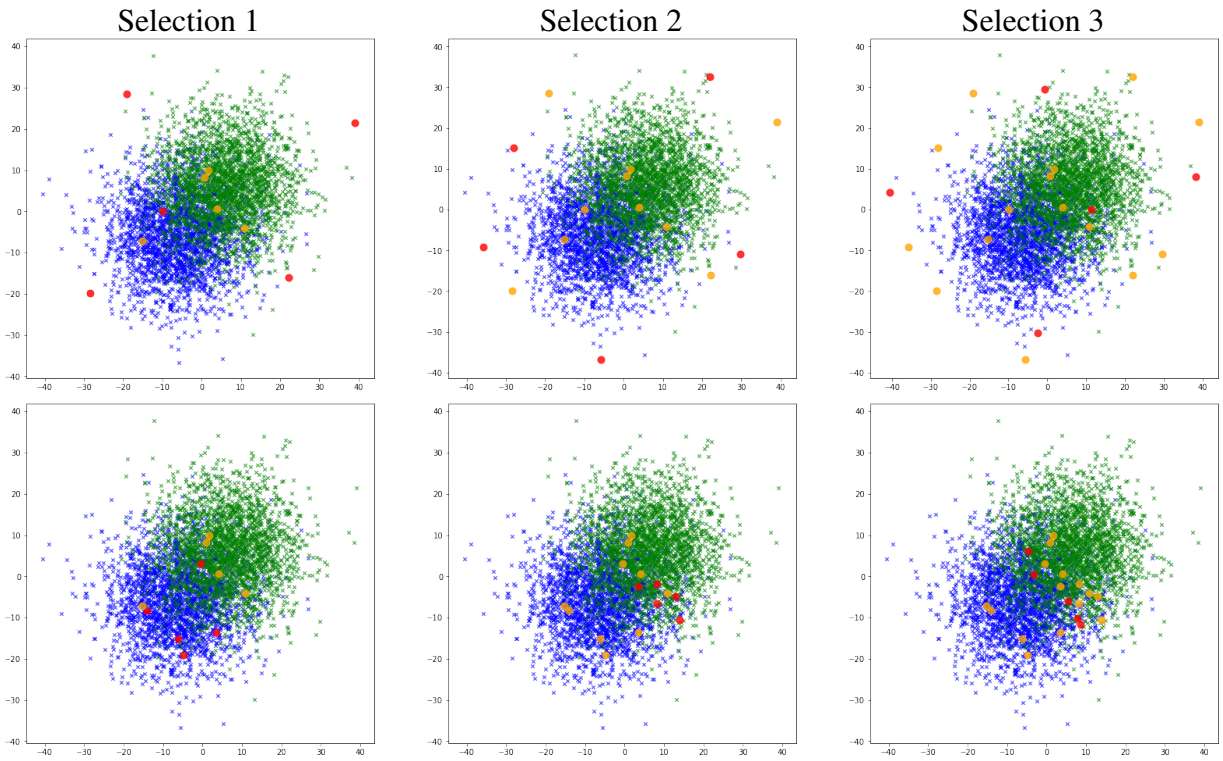

Figure 3.1: Active queries on 2-blob Gaussian dataset. First row: Coreset selections. Second row: Ensembles selections. Labeled points denoted in orange, Current step label queries in red.

shown in Figure 3.1. Each Gaussian represents one class, so this sets up a binary classification problem with a simple decision boundary. The low data complexity allows us to easily visualize selections in the original space. We train a simple multilayer perceptron (MLP) network with 2 layers and ReLU activation to solve this task, and simulate 3 active acquisition steps with random, diversity, and uncertainty sampling. We choose the Coreset [40] and Ensembles [7] strategies as described in Chapter 2 as our diversity and uncertainty candidates respectively. The warmup set consists of 5 randomly chosen labels, and each strategy queries 5 additional points at each step. After each selection, we evaluate accuracy on a held-out test set of 500 samples from the same synthetic dataset.

The results are visualized in Figure 3.1. We found that Coreset is unable to identify the points near the training boundary as it picks diversely throughout the space, causing more outlier selections. This issue persisted over all selections. In contrast, Ensembles picked nearly all of its points near the decision boundary. Table 3.1 shows test accuracy at each selection, including

Strategy	Selection 1	Selection 2	Selection 3
Coreset	**80.2**	83.4	80.3
Ensembles	79.4	**84.0**	**85.6**
Random	77.8	81.8	82.4

Table 3.1: Test accuracy for active strategies on 2D Gaussian dataset at each selection

accuracy of a random acquisition strategy. Coreset is hurt by outlier selections, causing accuracy to actually regress at the last selection. In contrast, Ensembles is able to leverage good boundary selections into a 3.2% improvement over random selection.

3.4 Natural Images

In moving to real datasets, we aim to find similar settings where a useful representation is quickly learned, causing decision-boundary points to be important when querying labels. We start with the MNIST handwritten digit dataset, which contains 60000 images. We run both Coreset and Ensembles with a query size of 600 (1% of the dataset), following the setup in [22]. We also replace the MLP with LeNet [31] to account for increased task difficulty. In order to slightly increase the difficulty of querying labels, we also test with harshly restricted data availability, using a low query size of 10 (0.2% of the dataset). Our results for both query sizes are shown in

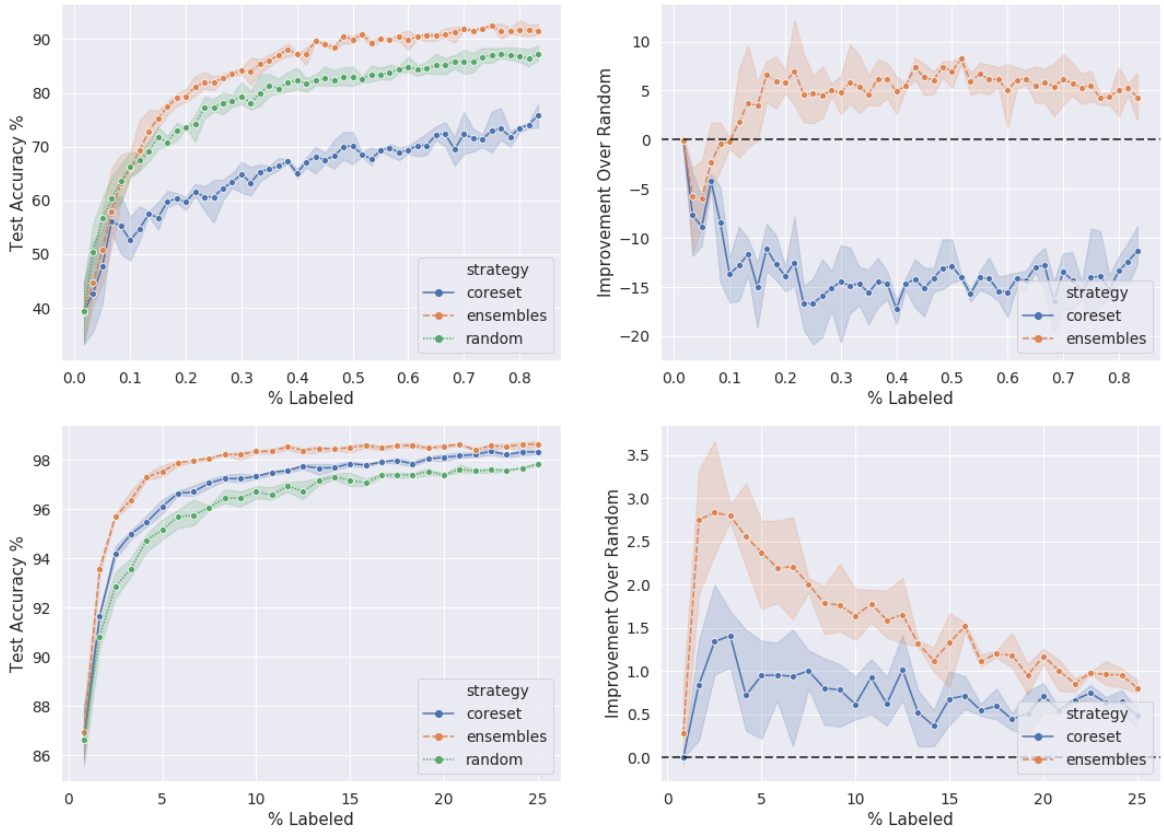

Figure 3.2: Left: MNIST test accuracy at each label query. Right: accuracy improvement over random sampling. Top: low query size (0.2%). Bottom: high query size (1%).

Figure 3.2. We found that both active approaches consistently outperformed random sampling at all queries, with Ensembles doing the best. In addition, all strategies achieved high accuracy very early, which is similar to the synthetic case. We also saw that diversity sampling was punished heavily when query size was low. This makes intuitive sense since the fewer points we

query diversely, the farther away they will be from each other, making it difficult to evenly pick throughout the space. In addition, Coreset relies heavily on the learnt representation - querying diversely from a poor representation space may not result in diverse points at all. Although MNIST is simple enough for a representation to be learnt with few samples, the extremely low query size may have caused a snowballing effect starting from a flawed initial representation.

For both query size settings, we see that we are able to quickly achieve extremely high performance on MNIST, even with less than 1% of the total labels. MNIST has been shown to be a redundant dataset where a fraction of the dataset is sufficient to learn the task [51]. [51] show that this is not the case for CIFAR-10, a 50000 image object recognition dataset. We test on CIFAR-10 to investigate if our findings extend to less redundant, higher-dimensional data. However, to keep the task simple, we reduce the number of classes to 2, using only the "horse" and "automobile" classes of CIFAR-10 as in [34]. This reduces the overall dataset size to 10000. We test Coreset and Ensembles with both a low query size of 10 (0.2%) and a high query size of 500 (5%).

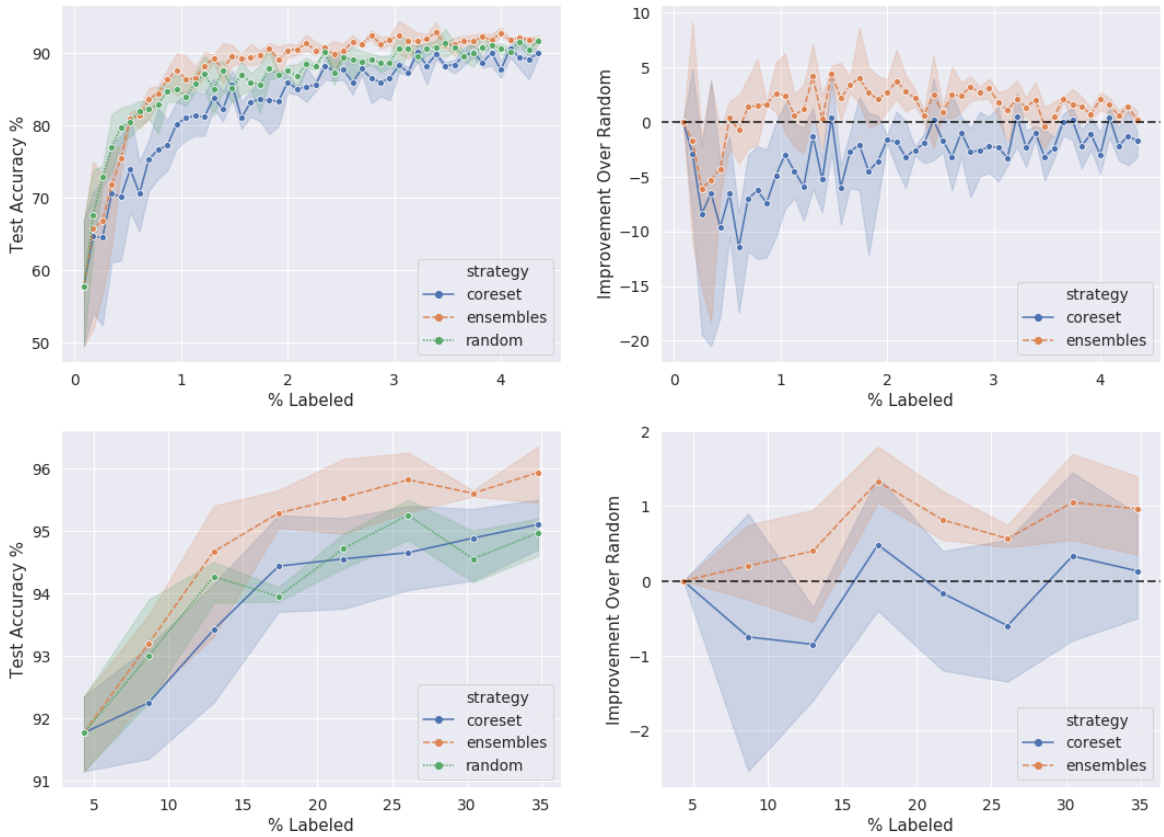

Figure 3.3: Left: CIFAR-10 2-class test accuracy at each label query. Right: accuracy improvement over random sampling. Top: low query size (0.2%). Bottom: high query size (5%).

Our results are shown in Figure 3.3. We again see the same trends in relation to Coreset and low query size, as well as Ensembles continuing to outperform the other two strategies. However, the margin of improvement was lower across all strategies. This introduces an idea that we will

see again in future settings: active learning yields less of a benefit when the task becomes more difficult. Table 3.2 summarizes our results across all datasets, showing the average accuracy improvement over all selections for each strategy.

Dataset/Strategy	Low query size % improvement	High query size % improvement
MNIST		
Coreset	-13.388	0.725
Ensembles	**4.584**	**1.547**
CIFAR-10 2-class		
Coreset	-3.158	-0.178
Ensembles	**1.338**	**0.667**

Table 3.2: Average Coreset and Ensembles accuracy improvement over all selections for simple classification settings

3.5 Discussion

Across all simple classification settings, uncertainty sampling was able to outperform diversity sampling. We noted that regardless of strategy, in all of these cases deep networks were able to achieve nearly 90% accuracy with extremely low data availability. Indeed, these settings may be unrealistic for an active learning scenario unless incremental performance beyond 90% is desired, or if label costs are extremely high. However, they give us empirical evidence that settings where uncertainty-based strategies outperform both diversity-based strategies and random selection do exist,

In the next chapter, we move to more practical datasets and settings while continuing to investigate ideas from this chapter like low data availability, representation strength, and relative task difficulty.

Chapter 4

Practical Classification Settings

4.1 Introduction

In this chapter, we study a variety of practical datasets and settings for deep active image classification. As described in Chapter 2, most existing work explores the setting of high-dimensional, difficult classification datasets with relatively large query sizes. However, we argue that this "vanilla" setting is often not only impractical, but also crippling to active approaches. We nevertheless experiment with uncertainty and diversity-based approaches in the vanilla setting, but also explore other settings where AL may have a larger benefit. We also look at more challenging settings involving low data availability.

4.2 Vanilla AL

For the vanilla setting, we tested Coreset and Ensembles on the CIFAR-100 dataset with a query size of 2500 (5%). CIFAR-100 is a object-recognition dataset similar to CIFAR-10, but with 100 classes instead of 10. In many deep AL scenarios, the label cost is high, leading to low data availability [48]. So, we also tested with a low query size of 50 (0.1%). To accommodate for the higher task difficulty, we use ResNets [25] for all experiments in this chapter.

We found that in the high query size setting, Coreset outperformed random selection. Diversity outperforming uncertainty-based approaches in this setting was a key result of Coreset [40], but the improvement we saw in practice was slight and much less than originally reported in the paper. However, in the low query size setting, both strategies failed to outperform random selection. Results are shown in Figure 4.1. Coreset's decline in performance was either due to the inability to learn a good representation or the low query size itself. Our results suggest that although this is the prevailing setting in previous work, it is an extremely challenging setting with small, if any, improvements in performance over random selection.

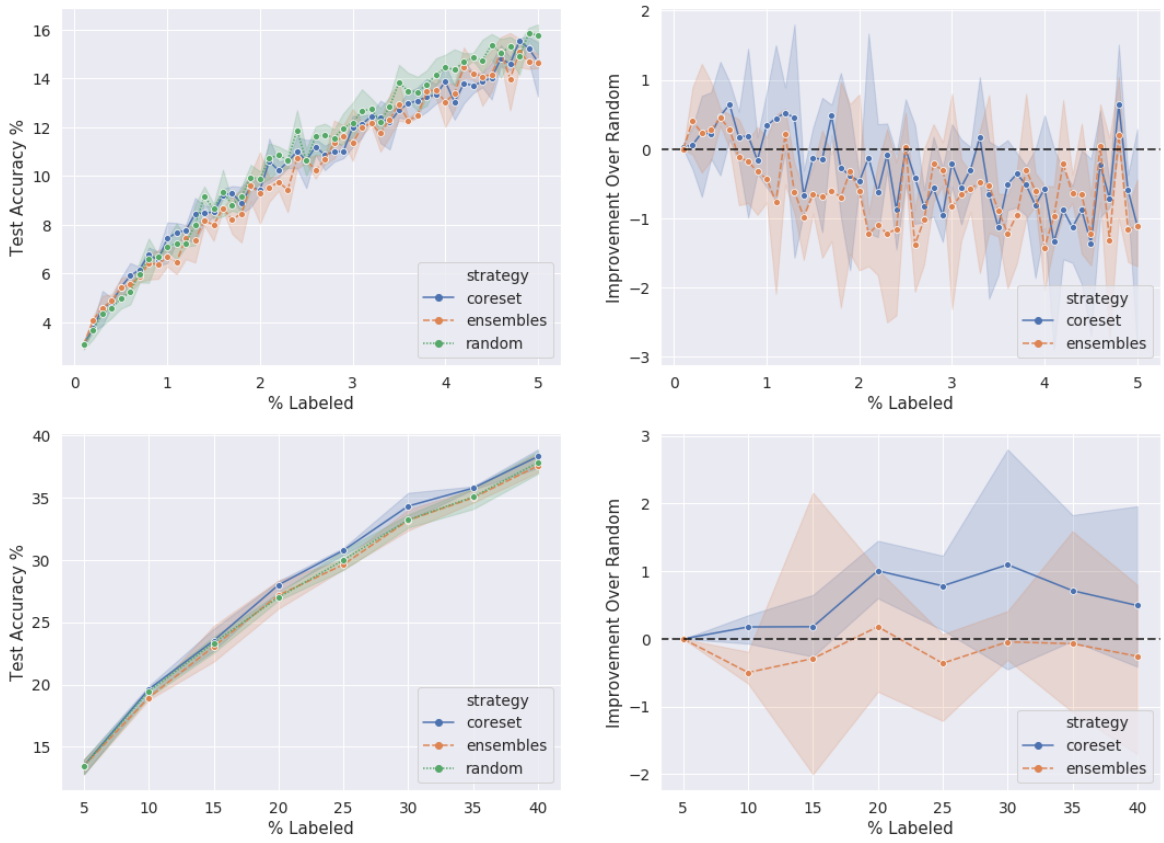

Figure 4.1: Left: CIFAR-100 test accuracy at each label query. Right: accuracy improvement over random sampling. Top: low query size (0.1%). Bottom: high query size (5%).

4.3 Transfer Learning

4.3.1 ImageNet Pretraining

Pretraining on ImageNet is a staple technique in deep image classification [26]. By using a network that has learned a more general object recognition representation space, we can ideally learn more sample-efficiently in a related target domain. We investigate the effects of replacing an untrained network in the vanilla setting with an ImageNet-pretrained network, with surprising results. Ideally this would lead to an improved representation which would improve Coreset's performance.

First, we simply run the vanilla setting but with a pretrained network, with results shown in Figure 4.2. For the high query size, we saw results similar to the vanilla setting. But for the low query size, while Ensembles was competitive with random sampling, Coreset heavily underperformed. We claim that this is because although the ImageNet representation is useful for classification, it is heavily biased towards ImageNet samples. And although the natural images in CIFAR-100 are similar, Coreset is unable to collect enough samples to unbias the representation. Since it is such a representation-reliant method, this leads to extremely poor selections.

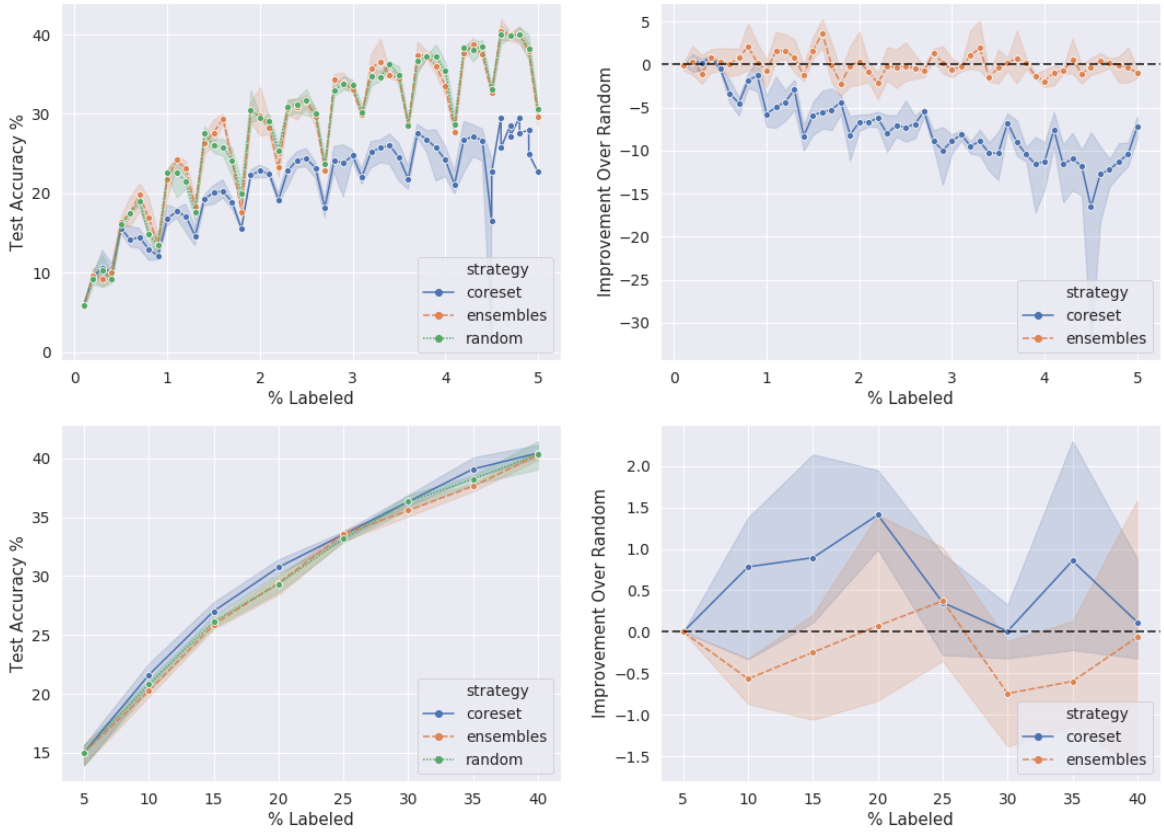

Figure 4.2: Left: CIFAR-100 (ImageNet pretrained) test accuracy at each label query. Right: accuracy improvement over random sampling. Top: low query size (0.1%). Bottom: high query size (5%).

4.3.2 Same-Dataset Pretraining

In order to confirm this claim, we eliminate the domain shift by replacing ImageNet pretraining with CIFAR-100 pretraining - that is, simply increasing the warmup size to a large percentage of the data. We repeat the analysis with a vanilla network and a warmup size of 15k (30%), with results in Figure 4.3. We saw that when a strong existing representation from the same dataset is present initially, Coreset was able to outperform random even for the low query size. Table 4.1 summarizes our results across the vanilla and pretrained settings, showing the average accuracy improvement over all selections for each strategy.

4.4 Class Imbalance

As discussed in Chapter 2.3, active learning is lucrative in imbalanced classification tasks for its ability to identify minority class samples. In this section we experiment with active learning strategies on imbalanced datasets ,with both natural and biological images, in order to see if these methods are able to work out-of-the-box in less-standard settings.

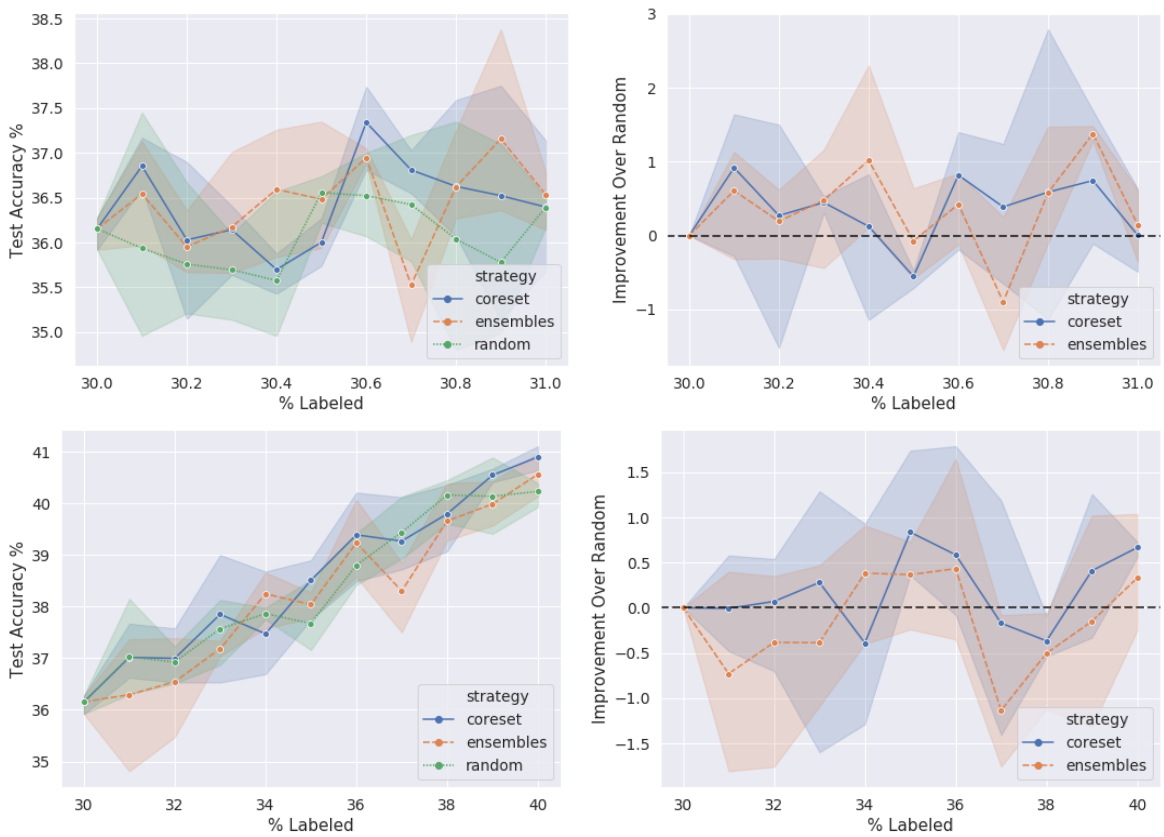

Figure 4.3: Left: CIFAR-100 (high warmup) test accuracy at each label query. Right: accuracy improvement over random sampling. Top: low query size (0.1%). Bottom: high query size (1%).

4.4.1 Natural Images

We manually construct an imbalanced binary classification task by marking the ”automobile” and ”truck” classes in CIFAR-10 with a positive label, and all other classes with a negative label. This sets up a ”car” vs. ”non-car” binary task, where the ”car” class is the minority. We additionally downsample the minority class to be 5% of the total labels, creating a 1:19 imbalance and further increasing the task difficulty.

In these settings, we are usually interested in improving performance of the minority class and not total accuracy. We instead evaluate average precision (AP) of the minority class. We experiment with Coreset and Ensembles using a low query size of 50 (0.1%), both with and without ImageNet pretraining.

Our results are shown in Figure 4.4. In the pretrained case, we saw similar results to [7], with Ensembles handily outperforming random sampling in test AP. We also saw this without pretraining. Figure 4.5 shows that Ensembles consistently selected the highest proportion of minority samples at each acquisition in both settings. Even with an extremely low query size, Ensembles was able to adequately identify and query a sufficient number of minority samples. Although the average AP improvement was only 0.03, the improvement increased with later selections. At the 6% data threshold, Ensembles provided a significant 0.15 increase in minority

Setting/Strategy	Low query size % improvement	High query size % improvement
Vanilla		
Coreset	**-0.309**	**0.556**
Ensembles	-0.564	-0.167
ImageNet pretrained		
Coreset	-7.126	**0.553**
Ensembles	**-0.002**	-0.221
High warmup		
Coreset	0.340	**0.174**
Ensembles	**0.350**	-0.162

Table 4.1: Average Coreset and Ensembles accuracy improvement over all selections for CIFAR-100 settings

class AP. On the other hand, Coreset was unable to outperform random sampling, with or without a prior representation.

4.4.2 Biological Images

ImageNet pretraining has also historically been used for target domains with non-natural images, such as biological images. [1] is such a dataset, containing high-resolution retina images in varying stages of diabetic retinopathy (DR). DR affects over 90 million people worldwide and is a leading cause of blindness. [7] designed an imbalanced classification problem for this dataset, and showed that Ensembles outperforms random sampling by identifying more minority samples at every acquisition.

We follow the same setup, collecting the images into 2 classes with a 1:19 imbalance, and pretraining our network on Imagenet. The minority class consists of retinal images with moderate-severe DR, while the majority class consists of all healthier retinal images. Figure 4.6 shows examples of both classes.

We run Ensembles and Coreset with a low query size of 20 images (0.1%). Our results are shown in Figure 4.7. We confirmed that Ensembles is able to outperform random sampling, but also show that it outperforms Coreset. Figure 4.8 shows that Ensembles again consistently selected the highest proportion of minority class samples at each acquisition.

Table 4.2 summarizes our results in the imbalanced setting, showing the average AP improvement over all selections for each strategy. We see that Ensembles performs the best across all settings, but with a much lower margin in the biological transfer learning settings vs. the natural transfer learning setting. We expect that this is due to higher task difficulty as we shift between image domains.

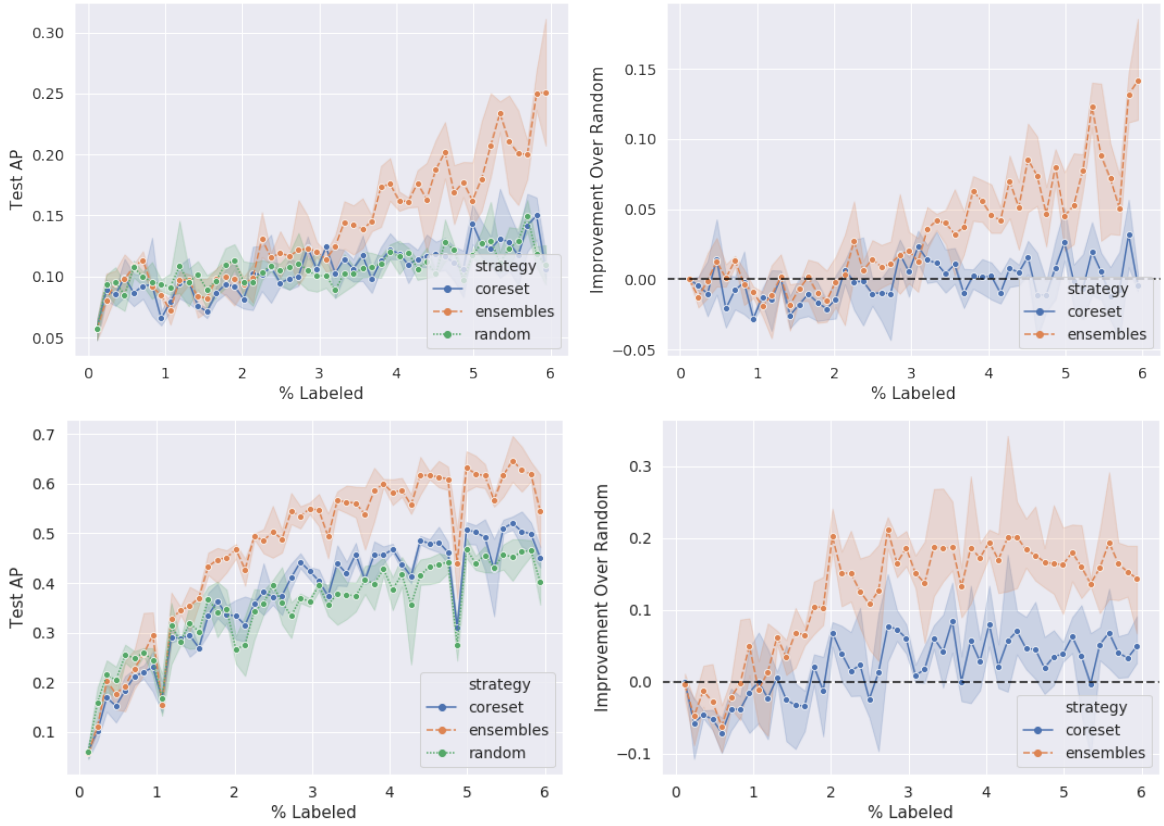

Figure 4.4: Left: CIFAR-10 2-class (1:19 imbalance) test AP at each label query. Right: AP improvement over random sampling. Top: vanilla. Bottom: ImageNet pretrained.

4.5 Discussion

In this chapter, we tested Ensembles and Coreset across a variety of practical settings. We found that low query sizes were a difficult setting for both strategies. In the vanilla and transfer learning cases both active strategy types struggled, and in the transfer learning setting with low data availability, Coreset highly underperformed. One setting where Coreset succeeded with low data availability was when a large initial labeled pool was present. This result is useful in rare cases such as when labeling cost is continually increasing. We can use a higher query size at the beginning of collection and lower it over time, using Coreset to query the labels. We also saw that, in general Coreset benefits from higher query sizes and is able to produce an accuracy improvement. This is true for both the vanilla and pretrained settings. However, the the average accuracy improvements we saw over random sampling in all of these settings was relatively low, at most 0.5%.

We note that once data complexity was high enough, it became difficult to consistently and evaluate active strategies in balanced classification settings. Networks must be retrained and re-evaluated at each query, so a true measure of performance at each data threshold would require a full hyperparameter search - an impractical approach when testing across various settings. Instead most existing work, and the approach we followed, uses a fixed set of hyperparameters for

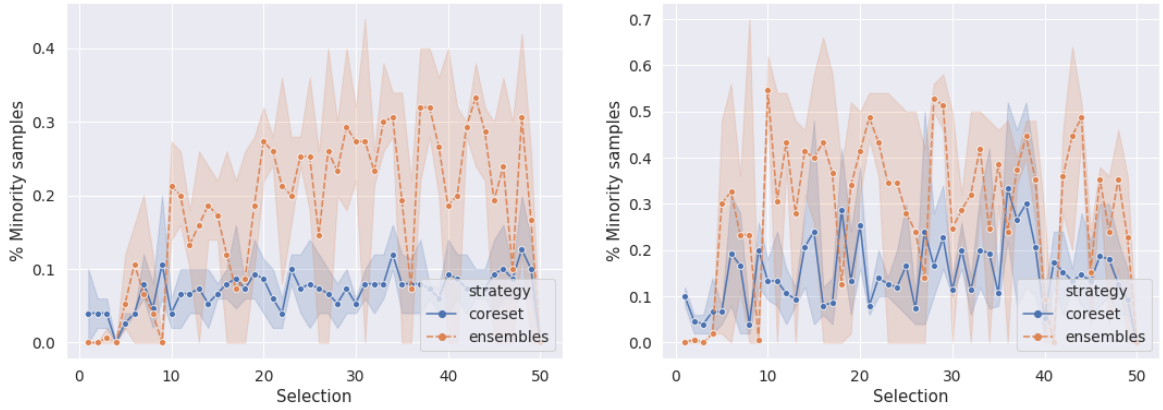

Figure 4.5: CIFAR-10 2-class (1:19 imbalance): proportion of minority class queried labels. Left: vanilla. Right: Imagenet pretrained.

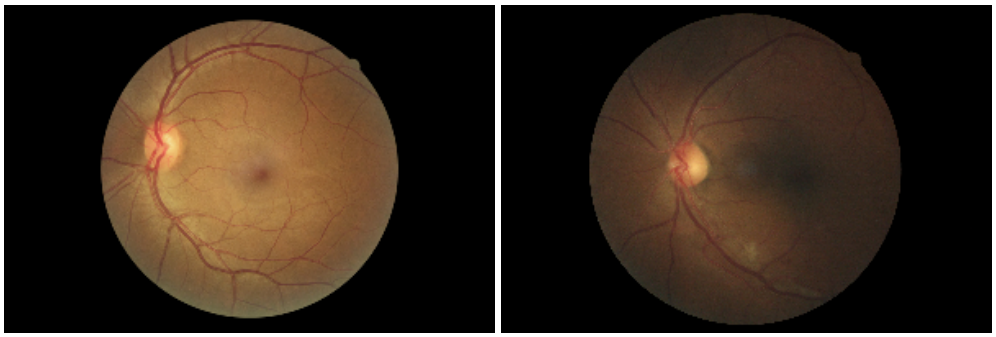

Figure 4.6: Diabetic Retinopathy image samples. Left: healthy (majority class). Right: unhealthy (minority class)

all steps. Ultimately this was reflected in inconsistent results, particularly in the vanilla setting where active methods seem to provide small, if any, accuracy improvements. These inconsistencies were echoed in the literature as well. [4] and [7] found that Coreset has inconsistent performance across different architectures and datasets. And our results were contradictory to [7], which claimed that Ensembles is able to outperform Coreset in the vanilla setting. Rigorously evaluating deep active methods may first require advances in deep learning theory.

Furthermore, our results across the 2D Gaussian, MNIST, CIFAR-10 2-class, and CIFAR-100 datasets suggest that in vanilla, balanced classification tasks, increased task complexity correlates with a decreased benefit of active learning. This mirrors results from classical active learning, as outlined in Chapter 2.1, and suggests that active learning may not be practically useful for difficult balanced classification settings unless incremental improvements in accuracy are highly valuable. Given these findings, our consistently results out-of-the-box in imbalanced settings with an uncertainty-based method are very promising. Ensembles performed the best in all tested scenarios, from vanilla binary classification to a challenging biological dataset, and in some cases with relatively large minority class AP improvement.

Figure 4.7: Left: Diabetic Retinopathy 2-class (1:19 imbalance) test AP at each label query. Right: AP improvement over random sampling.

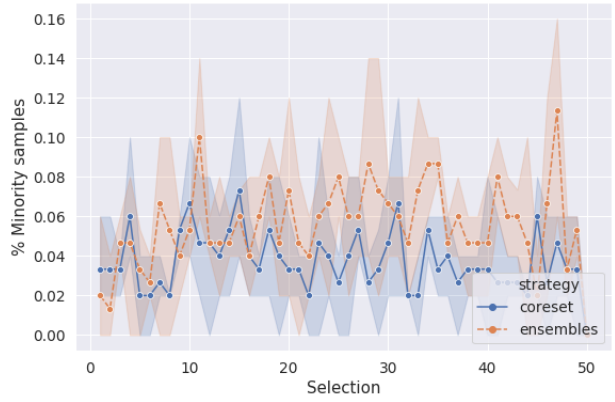

Figure 4.8: Diabetic Retinopathy 2-class (1:19 imbalance): proportion of minority class queried labels

Setting/Strategy	AP improvement
Vanilla CIFAR-10	
Coreset	-0.0015
Ensembles	**0.0325**
ImageNet pretrained CIFAR-10	
Coreset	0.0198
Ensembles	**0.1200**
Imagenet pretrained Diabetic Retinopathy	
Coreset	-0.0001
Ensembles	**0.0019**

Table 4.2: Average Coreset and Ensembles AP improvement over all selections for imbalanced settings

Chapter 5

Conclusions

In this work, we empirically analyzed deep active image classification algorithms across several different settings. We were motivated by the inability of classical active learning methods to succeed with deep models, the unintuitive nature of the methods that do succeed, and the narrowness of the settings that state-of-the-art methods are tested in. We focused on the performance of two state-of-the-art approaches, one a diversity-based strategy and the other a uncertainty-based strategy, while varying characteristics in the setting and data.

Our study found that while active learning is lucrative in simpler classification tasks, as data complexity increases, the commonly tested vanilla settings are poor candidates for active learning approaches. The small performance improvements we saw suggest that the use of an active strategy in these cases may only be practical if incremental accuracy improvements are highly valuable. Furthermore, when label cost is high, leading to low data availability, active approaches are unable to produce any improvement at all. We also noted that the diversity-based method is highly representation-dependent, and our study found at least one setting where this characteristic caused it to dangerously underperform. Finally, we saw that across a wide variety of settings, higher task complexity correlated with lower benefits of active learning, even for state-of-the-art methods. These findings suggest that the active learning practitioner should be wary of using active learning strategies out-of-the-box in these settings.

On the other hand, we saw that in imbalanced classification settings the uncertainty-based strategy was able to produce a significant performance improvement, outperforming both the diversity-based strategy and random sampling. This has two major implications. First, we argue that imbalanced classification should be the primary setting for developing and testing deep active learning algorithms. Imbalance is ubiquitous in real-world problems, and given the additional difficulty of robustly evaluating active approaches for deep networks, large performance improvements make it a promising setting for future active learning work. Second, the dominance of an uncertainty-based method over a diversity-based one in this setting allows us to revisit the existing literature on classical uncertainty sampling-based algorithms. We believe that even better active methods could be developed for imbalanced classification tasks using intuition from classical uncertainty-sampling methods.

In the future, we would like to also test hybrid strategies that incorporate both diversity and uncertainty. While some methods have been explored for the deep setting, they have generally not performed as well as non-hybrid approaches. It could be possible to combine Coreset and

Ensembles in a formulation that suppresses the flaws of both strategies. We would also like to test other practical changes to the setting, such as when acquired labels are noisy, or when the initial warmup labels are highly biased.

Bibliography

[1] Kaggle diabetic retinopathy detection training dataset. URL `https://www.kaggle.com/c/diabetic-retinopathy-detection`. 1, 4.4.2

[2] Naoki Abe and Hiroshi Mamitsuka. Query Learning Strategies Using Boosting and Bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 1–9, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-556-5. 2.1

[3] S. Argamon-Engelson and I. Dagan. Committee-Based Sample Selection for Probabilistic Classifiers. *Journal of Artificial Intelligence Research*, 11:335–360, November 1999. ISSN 1076-9757. doi: 10.1613/jair.612. arXiv: 1106.0220. 2.1

[4] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *arXiv:1906.03671 [cs, stat]*, June 2019. arXiv: 1906.03671. 2.2.3, 4.5

[5] Josh Attenberg, Etsy, and Brooklyn. Class Imbalance and Active Learning. 2011. 2.3

[6] Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 288–316, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. 2.1

[7] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The Power of Ensembles for Active Learning in Image Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00976. 2.2, 2.2.1, 2.2.2, 3.3, 4.4.1, 4.4.2, 4.5

[8] Michael Bloodgood and K. Vijay-Shanker. Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 137–140, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. 2.3

[9] Klaus Brinker. Incorporating Diversity in Active Learning with Support Vector Machines. page 8. 2.1

[10] R. M. Castro and R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, May 2008. ISSN 1557-9654. doi: 10.1109/TIT.

2008.920189. 2.1

[11] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 01 2002. doi: 10.1613/jair.953. 2.3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. page 8. 1

[13] Chris Drummond and Robert C. Holte. C 4 . 5 , class imbalance , and cost sensitivity : Why under-sampling beats oversampling. 2003. 2.3

[14] Melanie Ducoffe and Frederic Precioso. Adversarial Active Learning for Deep Networks: a Margin Based Approach. *arXiv:1802.09841 [cs, stat]*, February 2018. arXiv: 1802.09841. 2.2.2

[15] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S. Shankar Sasrty. A Convex Optimization Framework for Active Learning. In *2013 IEEE International Conference on Computer Vision*, pages 209–216, Sydney, Australia, December 2013. IEEE. ISBN 978-1-4799-2840-8. doi: 10.1109/ICCV.2013.33. 2.2

[16] Seyda Ertekin. Learning in extreme conditions: Online and active learning with massive, imbalanced and noisy data. In *PhD thesis, The Pennsylvania State University*, 2009. 2.3

[17] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: Active learning in imbalanced data classification. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 127–136, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321461. 2.3

[18] Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. Selective Sampling for Example-based Word Sense Disambiguation. *Computational Linguistics*, 24(4):573–597, 1998. 2.1

[19] Yarin Gal and Zoubin Ghahramani. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. *arXiv:1506.02158 [cs, stat]*, January 2016. arXiv: 1506.02158. 2.2.2

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv:1506.02142 [cs, stat]*, October 2016. arXiv: 1506.02142. 2.2.2

[21] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. *arXiv:1703.02910 [cs, stat]*, March 2017. arXiv: 1703.02910. 2.2.2

[22] Daniel Gissin and Shai Shalev-Shwartz. Discriminative Active Learning. September 2018. 2.2, 2.2.1, 3.4

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *arXiv:1706.04599 [cs]*, August 2017. arXiv: 1706.04599. 2.2

[24] Yuhong Guo. Active Instance Sampling via Matrix Partition. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 802–810. Curran Associates, Inc., 2010. 2.2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. arXiv: 1512.03385. 4.2

[26] Kaiming He, Ross Girshick, and Piotr DollÃ¡r. Rethinking ImageNet Pre-training. *arXiv:1811.08883 [cs]*, November 2018. arXiv: 1811.08883. 4.3.1

[27] Sheng-Jun Huang and Songcan Chen. Transfer Learning with Active Queries from Source Domain. page 7. 2.2.3

[28] Nathalie Japkowicz. Learning from imbalanced data sets: A comparison of various strategies. pages 10–15. AAAI Press, 2000. 2.3

[29] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning. In *IJCAI*, 2007. 2.1

[30] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014. arXiv: 1312.6114. 2.2.1

[31] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Ha. Gradient-Based Learning Applied to Document Recognition. page 46, 1998. 3.4

[32] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1994. 2.1

[33] Michael Lindenbaum, Shaul Markovitch, and Dmitry Rusakov. Selective Sampling for Nearest Neighbor Classifiers. *Machine Learning*, 54(2):125–152, February 2004. ISSN 0885-6125. doi: 10.1023/B:MACH.0000011805.60520.fe. 2.1

[34] Christoph Mayer and Radu Timofte. Adversarial Sampling for Active Learning. *arXiv:1808.06671 [cs, stat]*, August 2018. arXiv: 1808.06671. 3.4

[35] Andrew McCallum and Kamal Nigam. Employing EM and Pool-Based Active Learning for Text Classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 350–358, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-556-5. 2.1

[36] Prem Melville and Raymond J. Mooney. Diverse Ensembles for Active Learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 74–, New York, NY, USA, 2004. ACM. ISBN 978-1-58113-838-2. doi: 10.1145/1015330.1015385. event-place: Banff, Alberta, Canada. 2.1

[37] Ion Muslea, Steven Minton, and Craig A. Knoblock. Selective Sampling with Redundant Views. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 621–626. AAAI Press, 2000. ISBN 978-0-262-51112-4. 2.1

[38] Nicholas Roy and Andrew McCallum. Toward Optimal Active Learning Through Sampling Estimation of Error Reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 441–448, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-778-1. 2.1

[39] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active Hidden Markov Models for Information Extraction. In Frank Hoffmann, David J. Hand, Niall Adams, Douglas

Fisher, and Gabriela Guimaraes, editors, *Advances in Intelligent Data Analysis*, Lecture Notes in Computer Science, pages 309–318, Berlin, Heidelberg, 2001. Springer. ISBN 978-3-540-44816-7. doi: 10.1007/3-540-44816-0_31. 2.1

[40] Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. *arXiv:1708.00489 [cs, stat]*, August 2017. arXiv: 1708.00489. 1, 2.2, 2.2.1, 3.3, 4.2

[41] Burr Settles. Active Learning Literature Survey. page 67. 1, 2, 2.1

[42] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 1070, Honolulu, Hawaii, 2008. Association for Computational Linguistics. doi: 10.3115/1613715.1613855. 2.1

[43] Burr Settles, Mark Craven, and Soumya Ray. Multiple-Instance Active Learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1289–1296. Curran Associates, Inc., 2008. 2.1

[44] H. S. Seung, M. Opper, and H. Sompolinsky. Query by Committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 287–294, New York, NY, USA, 1992. ACM. ISBN 978-0-89791-497-0. doi: 10.1145/130385.130417. event-place: Pittsburgh, Pennsylvania, USA. 2.1

[45] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1948.tb01338.x. 2.1

[46] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 761–769, June 2016. doi: 10.1109/CVPR.2016.89. 2.3

[47] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational Adversarial Active Learning. *arXiv:1904.00370 [cs, stat]*, March 2019. arXiv: 1904.00370. 1, 2.2.1

[48] Asim Smailagic, Hae Young Noh, Pedro Costa, Devesh Walawalkar, Kartik Khandelwal, Mostafa Mirshekari, Jonathon Fagert, AdriÃ¡n GaldrÃ¡n, and Susu Xu. MedAL: Deep Active Learning Sampling Method for Medical Image Analysis. *arXiv:1809.09287 [cs]*, September 2018. arXiv: 1809.09287. 4.2

[49] Simon Tong and Daphne Koller. Support Vector Machine Active Learning with Applications to Text Classiïcation. page 22. 2.1

[50] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR 2011*, pages 1449–1456, June 2011. doi: 10.1109/CVPR.2011.5995430. 1

[51] Kailas Vodrahalli, Ke Li, and Jitendra Malik. Are All Training Examples Created Equal? An Empirical Study. *arXiv:1811.12569 [cs, stat]*, November 2018. arXiv: 1811.12569. 3.4

[52] Zheng Wang and Jieping Ye. Querying Discriminative and Representative Samples for Batch Mode Active Learning. page 9. 2.1

[53] Gert W. Wolf. Facility location: concepts, models, algorithms and case studies. Series:

Contributions to Management Science. *International Journal of Geographical Information Science*, 25(2):331–333, March 2011. ISSN 1365-8816. doi: 10.1080/13658816.2010.528422. 2.2.1

[54] P. Xie, R. Salakhutdinov, L. Mou, and E. P. Xing. Deep determinantal point process for large-scale multi-label classification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 473–482, Oct 2017. doi: 10.1109/ICCV.2017.59. 2.3

[55] Pengtao Xie, Aarti Singh, and Eric P. Xing. Uncorrelation and evenness: a new diversity-promoting regularizer. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3811–3820, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 2.3

[56] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G. Hauptmann. Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization. *Int. J. Comput. Vision*, 113(2):113–127, June 2015. ISSN 0920-5691. doi: 10.1007/s11263-014-0781-x. 2.2

[57] Jingbo Zhu and Eduard Hovy. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 783–790, Prague, Czech Republic, June 2007. Association for Computational Linguistics. 2.3