Data Importance-Based Active Learning for Limited Labels

Pouya Pezeshkpour
University of California
Irvine, CA
pezeshkp@uci.edu

Zhengli Zhao
University of California
Irvine, CA
zhengliz@uci.edu

Sameer Singh
University of California
Irvine, CA
sameer@uci.edu

Abstract

Active learning aims to reduce the required labeled data by selecting which unlabeled instances should be labeled next for a given model, instead of picking instances randomly. To select the instances to label, most existing approaches primarily rely on some form of uncertainty in the current predictions, as obtaining labels for these instances is expected to be the most useful for the current model. Despite the success of these approaches, uncertainty sampling does not directly evaluate what will change if an instance is labeled. In this work, we introduce a formulation of active learning that directly estimates this effect of adding an instance to the training data on the model, and uses this so called data importance *to select the next instance to label. In particular, we use a recently introduced representer point representation of the model to efficiently estimate the importance of samples on the model predictions, and use it for active learning by selecting points that will have the largest impact. We evaluate our model on both synthetic and real-world datasets, demonstrating the utility of data importance in active learning for the limited label regime.*

1. Introduction

Given that recent machine learning models often rely on large labeled datasets, it is difficult to adapt them to novel tasks and labels without considerable effort (and time) for gathering labels. To quickly train machine learning models for the task of interest, active learning [8, 2, 3] iteratively selects the instance to be labeled next that will be the *most informative* to the current model. These active learning methods have achieved tremendous success on a variety of classification tasks improving their performance using less number of labeled data.

The primary goal of active learning is to select the instance that, when labeled, will be the most informative for the current model. Estimating the effect of each instance is difficult not only because the label of the instance is unknown, but also because it is challenging to efficiently estimate the actual effect of including the instance into the training data. Many active learning algorithms thus primarily rely on using the model uncertainty as a proxy of informativeness to the model [7, 1, 11], since getting *any* one label for such instances should be informative. Although these methods improve model performance on a limited amount of labeled data, they are often not consistently better than just randomly selecting which sample to label [1, 9]. The inability of uncertainty based approaches to improve classifiers' performance dramatically compared to random selection on some samples might be a result of the inherent difficulty of those samples and not their role in better classifying the data points which is not distinguishable in terms of uncertainty. As a result, there is a need for active learning algorithms that go beyond just uncertainty and efficiently estimate the informativeness of the samples.

In this paper, we introduce a data importance-based batch active learning algorithm, i.e., identifying the most informative samples to label at each step using their influence on the predictions of the model. In particular, data importance of a training sample is its importance (or contribution) for a specific target prediction. For active learning, we are interested in choosing the most informative samples based on their accumulative influence on the current training data. More specifically, we choose data points that disagree with current training samples the most using importance as our metric. In the past few years, there have been many attempts to provide an efficient approximation for data importance such as influence function [5] and Shapley values [6], however they are quite slow. Instead, we build upon the representer point selection [10] that estimates the influence of each training sample on the prediction by decomposing the output for the target as a weighted sum for each training point, capturing their importance for the target prediction.

We evaluate our proposed methods through the following experiments. First, on a synthetic dataset generated to demonstrate shortcomings of random and uncertainty based active learning algorithms, our model significantly outperforms these baselines. Second, we evaluate our model on a low-label setting for CIFAR-100, demonstrating that our approach in identifying informative samples is considerably effective when very few labeled instances are available.

2. Importance Based Active Learning (IBAL)

Before describing data importance's application to active learning by identifying the most informative samples, we first introduce the representer point selection model. Then, we define importance score for each point, and adapt the representer point selection to an active learning scenario. Finally, to eliminate inherently difficult instances, we combine this importance metric with a distance-based score.

Representer Point Selection As shown in Yeh et al. [10], the output for a target sample of any classifier that has a linear output layer and L2 regularization for that layer's weights can be represented by a formulation that assigns an "importance" to each instance in the training data, i.e.:

$$\phi(x_t, \theta^*) = \sum_{i}^{n} k(x_t, x_i, \alpha_i) \tag{1}$$

where $\phi(.)$ represent the output score of model for the target sample x_t, $\alpha_i = \frac{1}{-2\lambda_n} \frac{\partial \mathcal{L}(x_i, y_i, \theta)}{\partial \phi(x_i, \theta)}$ and $k(x_t, x_i, \alpha_i) = \alpha_i f_i^T f_t$. Further, $\mathcal{L}(x_i, y_i, \theta)$ denotes the loss term for any training sample, and f_i/f_t denote the model's output *before* the last layer for training sample x_i and target x_t, respectively. This $k(x_t, x_i, \alpha_i)$ thus represents the importance of training sample x_i on the target prediction x_t.

Importance Based Active Learning Intuitively, the most informative samples for updating the model are the unlabeled samples whose labeling would disagree the most with our current model predictions. We want to estimate, for each unlabeled data, how *important* it would be for the current model, if it was in the training data. If a data point seems to be important for the current model over existing training points, it suggests that the data point is more similar to them, and thus adding it to the training data will have minimal effect. To capture this intuition, we define the aggregate importance of each unlabeled instance on the training data by using representer point selection and treating the current training data as our target predictions and the pool of unlabeled samples as our "training" data, and identify the most informative unlabeled sample as one that is least consistent with the training data, as:

$$\begin{aligned} &\text{argmin}_{x_i} \text{IBAL}(x_i) = \\ &\text{argmin}_{x_i} \sum_{x_t \in \text{L-set}} \sum_{y \in \text{C}} p(y|x_i) k(x_t, x_i, \alpha_i|y) \end{aligned} \tag{2}$$

where L-set is the set of all previously labeled samples, and C is the set of all possible classes. Further $k(x_t, x_i, \alpha_i|y)$ denote the representer value for class y. Thus, minimizing $\text{IBAL}(x_i)$ identifies samples that disagree the most with current labeled samples. To form the batch, we choose the top-B samples minimizing the above equation (B is our labeling budget for each iteration).

Distance Score To avoid choosing samples that are both of low influence and close to previously labeled ones, i.e. samples that are challenging but not helpful for the model, we introduce additional measurement of distance as:

$$\text{D}(x_i) = \text{argmin}_{x_t \in \text{L-set}} d(x_i, x_t) \tag{3}$$

where d measures the euclidean distance between representations of data points. In our implementation, we utilize the encoder part of an autoencoder trained on all the samples, to project data onto latent space. We use two CNN layers for both the encoder and the decoder of the autoencoder.

Active Learning Objective To combine the importance and distance scores, we first normalize the distance score (divided by the maximum value), and then calculate the final score for each sample as:

$$\text{IBAL-D}(x_i) = \text{IBAL}(x_i) - \gamma \text{D}(x_i) \tag{4}$$

where γ is a hyperparameter that we tune on a validation dataset. As a result, we select the most informative data points in each step through the following optimization:

$$\text{argmin}_{x_i} \text{IBAL-D}(x_i) \tag{5}$$

capturing both importance and distance.

3. Experiments

We evaluate our model on both a synthetic dataset for intuitive illustration, and the commonly-used CIFAR-100 dataset for comparing with baseline methods. For our synthetic dataset, we use a simple model of two linear layers with the relu activation, while for CIFAR-100, we use resnet-18 [4] as our classifier.

Synthetic data To demonstrate the shortcomings of uncertainty based AL, and of randomly picking points, we generate a synthetic dataset as visualized in Figure 1. We want to capture two fundamental issues of baseline methods: (1) imbalanced data affects the efficiency of randomly choosing samples, and (2) uncertainty based methods only focus on samples with low certainty which might be inherently challenging (as opposed to poorly represented) points. We consider 2000, 500, and 50 points for our three classes respectively. As the baselines we consider 1) sample data points uniformly at random and 2) entropy-based

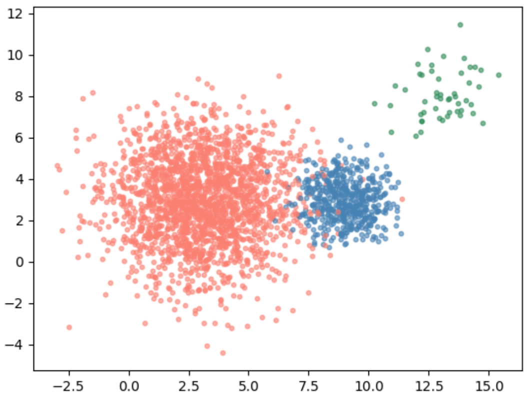


Figure 1: Scatter plot of synthetic dataset (2 features, 3 classes) that captures the shortcomings of random sampling (imbalanced classes) and uncertainty based active learning (class overlap).

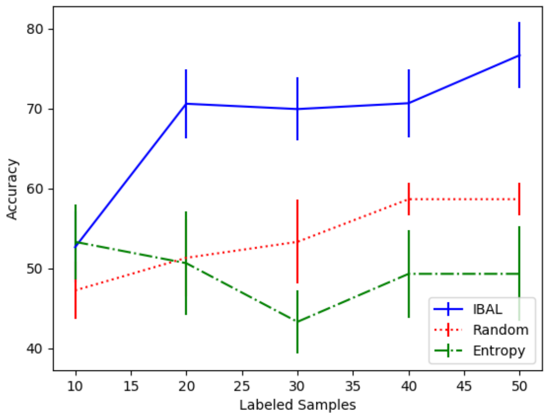


Figure 2: Accuracy on synthetic dataset.

active learning representing uncertainty based approaches. The average accuracy (over 5 different random seeds) plot of our model is reported in Figure 2. As it shows, our importance-based method outperforms baselines by a large margin, demonstrating its capability in identifying informative samples more efficiently.

CIFAR-100 We also carry out additional experiments on CIFAR-100 dataset to evaluate the performance of our method on a real dataset. In addition to random and entropy baselines, we also consider combining our distance score with entropy, solving the optimization problem: $\text{argmax}_{x_i} \text{Entropy}(x_i) + \gamma D(x_i)$. The average accuracy results (over 5 different random seeds) of the active learning algorithm on CIFAR-100 are shown in Figure 3. Although CIFAR-100 is a well-balanced dataset making it hard to improve upon the random baselines, as the figure shows, using data importance to identify informative samples can be beneficial to designing a better active learning algorithm. Further, our final model with 250 labeled samples outperforms all the baselines. To show that the importance-based measure (IBAL from Eq (2)) is qualitatively different from uncertainty measure for active learning, we plot the normalized IBAL score vs two uncertainty estimates (entropy and classifier's confidence) for CIAFR-100 data in Figures 4a and 4b. As it shows, clearly there is a significant difference between these two approaches, suggesting IBAL is capturing a different active learning phenomenon.

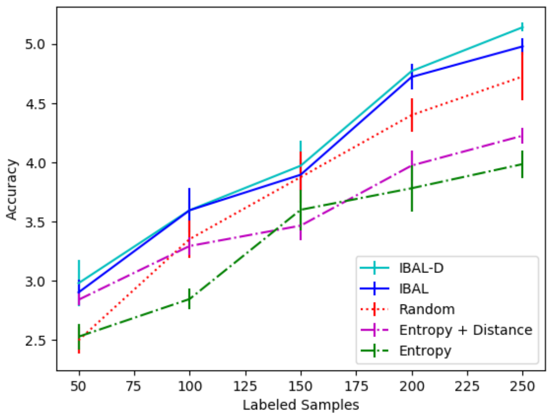
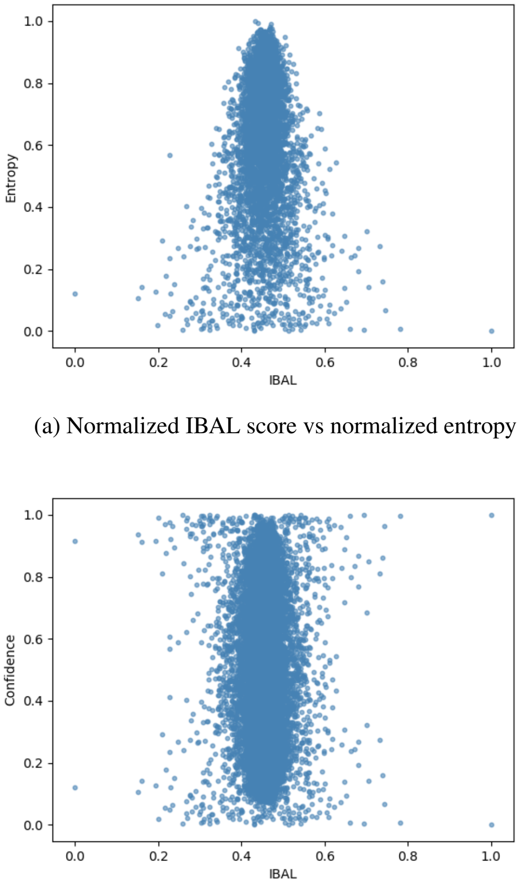


Figure 3: Accuracy on CIFAR-100 dataset.

4. Conclusion and Future Work

Motivated by the need for more efficient active learning algorithms, we present a novel approach for active learning incorporating the sample importance. We adopt the previously introduced representer point selection to introduce a data importance-based score for active learning and combine it with diversity to develop an efficient active learning algorithm. Evaluating the accuracy performance of our model on synthetic and real-world data with a quite limited number of labeled samples, we demonstrate that our method can outperform several baselines, shedding light upon the effectiveness of combining importance and diversity.

For future work, we seek to incorporate the representer point selection into the importance score in a more principled manner. As for diverse sampling, we will conduct comparisons to other methods of diversity on more datasets. Moreover, we will analyze further on the assumptions of our problem setup, and carry out thorough ablation studies on the components of our proposed active learning pipeline.



(a) Normalized IBAL score vs normalized entropy

(b) Normalized IBAL's score vs classifier's confidence

Figure 4: Scatter plots of IBAL scores versus uncertainty estimates, on CIFAR100 dataset

Acknowledgments

We would like to thank the anonymous reviewers for their detailed feedback and suggestions. This work is sponsored in part by the Air Force Research Laboratory (AFRL), and DARPA. The views expressed are those of the authors and do not reflect the official policy or position of the funding agencies.

References

[1] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. 1

[2] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415, 2014. 1

[3] Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017. 1

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017. 1

[6] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017. 1

[7] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 1

[8] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009. 1

[9] S. Sinha, S. Ebrahimi, and T. Darrell. Variational adversarial active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5972–5981, 2019. 1

[10] C.-K. Yeh, J. Kim, I. E.-H. Yen, and P. K. Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9291–9301, 2018. 1, 2

[11] F. Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019. 1