IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125292

Active Learning Methods for Annotating Training Sets

Gorla Charan Sai Chowdhary, Suraj Rajshekhar Mukkannavar, Kushagra Gupta, Rajot Saha, Prof. Anala M R

Information Science and Engineering RV College of Engineering, Bangalore, India

Abstract— Active learning is a machine learning technique that identifies data that should be labeled by human annotators. This can be used to reduce the cost and time of labeling datasets, while still achieving high accuracy. Active learning works by iteratively training a machine learning model on a small set of labeled data, and then using the model to predict the labels of unlabeled data. The data points that the model is most uncertain about are then selected for labeling. This process is performed iteratively until the desired level of accuracy is achieved. Active learning has been shown to be effective for a variety of machine learning tasks, including text classification, image classification, and natural language processing. It is particularly well-suitedfor tasks where labeling data is expensive or time-consuming. In this study, we investigate the use of active learning with the CIFAR10, EuroSAT and FashionMNIST datasets. We compare a variety of active learningmethods, including Least Confidence, Margin Sampling and Entropy Sampling. We show that they can all improve the performance of the model over random sampling.

Keywords— Active learning , Human Labeling, Least Confidence, Margin Sampling, Entropy Sampling, CIFAR-10 , EuroSAT, CNN, Fashion MNIST.

I. INTRODUCTION

In machine learning, data labeling is the process of assigning labels to data points. This is a critical step in the machine learning process, as the quality of the labels will directly impact the performance of the machine learning model. However, data labeling can be a time-consuming and expensive process, especially for large datasets. Active learning is a technique that can be used to reduce the amountof data labeling required. In active learning, a machine learning model is trained on a small subset of labeled data[1].

The model is then used to predict the labels of unlabeled data points. The data points for which the model is least confident are then labeled by a human expert. This process isrepeated until the model is sufficiently trained. This can improve the performance of the machine learning model. It is because the model is trained on a more representative sample of the data. Active learning methods that are used in this work are Least Confidence, Margin Sampling and Entropy Sampling[5].

Least confidence is a method that selects the data points for which the model has the least confidence in its prediction.This strategy is based on the idea that data points for which the model is uncertain are more likely to be informative and thus more likely to improve the model's performance. In Margin sampling it selects the data points that are most uncertain for the current model. This is done by calculating the margin of each data point, which is the difference between the highest probability and second highest probability where the probability is the likeliness of the datapoint belonging to a particular class.The data points with the smallest margins are the most uncertain and are therefore the most valuable to label. Entropy sampling is a method of sampling data that is based on the entropy of the data. Entropy is a measure of the uncertainty of a data set, and it is calculated by counting the number of different values that a data point can take on. Entropy sampling selectsdata points with high entropy, which means that they are more uncertain and therefore more likely to be informative. These three methods are applied on CIFAR-10, EuroSAT andFashionMNIST Datasets.

II. RELATED WORK

A new active labeling method for Deep learning [5]: The proposed active learning method uses a Bayesian framework to select the most informative data points to query. The method was shown to outperform other active learning methods on the MNIST and CIFAR-10 datasets. The methodis likely to be effective in a variety of other applications where deep learning models are used.

Region-based active learning for efficient labeling in semantic segmentation [6]: Region-based active learning is a promising approach for reducing the amount of labeled data required to train semantic segmentation models.The method is based on a Bayesian framework and uses a combination of entropy and uncertainty measures to select informative regions. The method has been shown to be effective in a variety of settings and is likely to be even more effective as deep learning models become more complex.

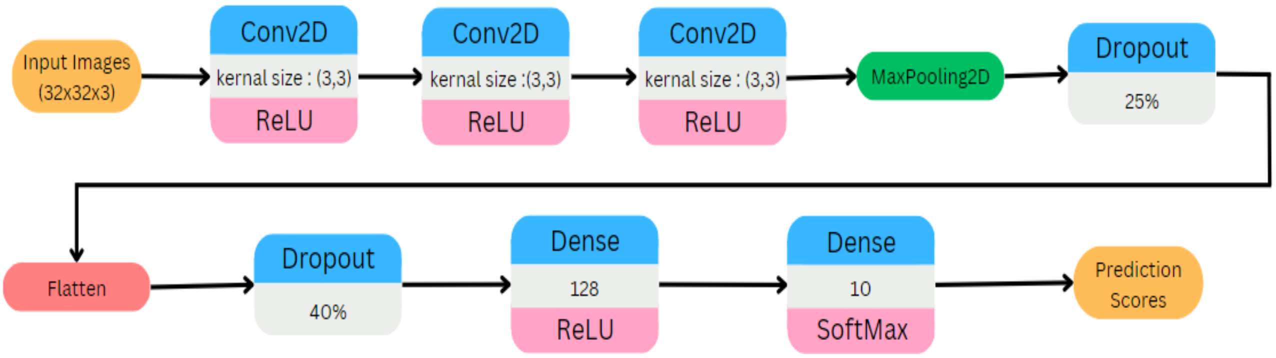

Fig. 1: Proposed Architecture

Reducing Label Effort: Self-Supervised meets Active Learning[7]: The proposed AL-SSL framework is a promising approach for reducing the amount of labeled data required to train machine learning models. The framework issimple to implement and can be used with a variety of machine learning models. The framework is also scalable and can be used with large datasets.

DIAL: Deep Interactive and Active Learning for Semantic Segmentation in Remote Sensing [8]: The proposed DIAL framework is a promising approach for semantic segmentation in remote sensing. The framework is simple to use and can be used with a variety of deep neural networks. The framework is also scalable and can be used with large datasets. Active learning with point supervision for cost-effective panicle detection in cereal crops [9]: The proposed active learning approach is a promising approach for cost-effective panicle detection in cereal crops. The approach is simple to implement and can be used with a variety of deep learning models. The approach is also scalable and can be used with large datasets.

Gradient and Log-based Active Learning for SemanticSegmentation of Crop and Weed for Agricultural Robots[10]: The proposed active learning method is a promising approach for reducing the amount of labeled data required to train semantic segmentation models for weed detection inagricultural robots. The method is simple to implement and can be used with a variety of deep learning models. Themethod is also scalable and can be used with large datasets. Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images [11]: The proposed active learning method is a promising approach for reducing the amount of labeled data required to train semantic segmentation models for land cover classification in satellite images. The method is simple to implement and can be usedwith a variety of deep learning models. The method is alsoscalable and can be used with large datasets.

Active learning for object detection in high resolution satellite images. [12] Their main objective was to apply two active learning techniques to segmentation of satellite images: Bayesian dropout and Core-Set. To evaluate the contribution of the uncertainty and core-set approaches, we compare them to two different baselines. The Core-Set approach seems to be the most effective method for the weak model. For the strong model, even if the Core-Set approachis also performing well the Bayesian dropout approach outperforms it by a large margin.

Dropout as Bayesian approximation: Deep learning representation of model uncertainty.[16] This influential paper demonstrates how dropout in deep neural networks can be interpreted as approximate Bayesian inference, enabling active learning methods to leverage model uncertainty estimates for effective instance selection.

Active learning based on diversity with applications to the detection of unusual classes in data streams. [23] This work investigates rare-class recognition in streaming data using diversity-based active learning techniques. In order to

DOI: 10.17148/IJARCCE.2023.125292

enhance the detection of unusual occurrences in changing data streams, it provides a novel active learning technique that maximizes the diversity of labeled cases.

A comparison study using active learning for graph classification. [20] A comparison of active learning methods for graph classification tasks is presented in this research report. It assesses and contrasts several active learning techniques on diverse graph datasets, offering insights into active learning's efficacy for graph-based tasks.

The power of ensembles in classifying images through active learning.[18] In this study, deep ensemble models and activelearning are combined. It demonstrates how ensemble-based active learning techniques may be used to do image classification tasks with less annotation work and higher classification accuracy.

Bayesian active learning for electronic health records-basedclinical decision assistance[19]. This study investigates the use of active learning strategies with electronic health records-based clinical decision support systems. It shows how Bayesian active learning can be used to enhance clinicalpredictions while reducing the need for expert annotation.

A deep reinforcement learning technique to active learning.
[17] Active learning is shown in this work using a deep reinforcement learning approach. In order to learn an agent's policy for choosing which instances to label,it formulates active learning as a sequential decision-making issue and applies a deep Q-network to the problem.

Active learning with inadequate information[22]. The purpose of this study is to examine active learning in situations where the annotator can only supply incomplete orrestricted information. It proposes brand-new active learning

methods that deal with missing or ambiguous label queries, making active learning more resilient and flexible to actual annotating scenarios.

III. DATASET DESCRIPTION

The CIFAR-10 dataset is a collection of 60,000 32x32 color images in 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The images are labeled withone of these 10 classes and are split into 50,000 training images and 10,000 test images. The CIFAR-10 dataset is a popular benchmark for machine learning algorithms that are used to classify images.The images in the CIFAR-10 dataset were collected from the web and resized to 32x32 pixels.

EuroSAT RGB dataset is a publicly available dataset of 27,000 labeled and geo-referenced satellite images covering 10 land use and land cover classes. The dataset is based on Sentinel-2 satellite images. The categories are Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, Sea and Lake, with eachcategory consisting of 2,700 images. The images have a spatial resolution of 64x64 pixels and are stored in RGB format.

The Fashion-MNIST dataset is a collection of 60,000 28x28 grayscale images of 10 fashion categories, along with a test set of 10,000 images. Each image is associated with a label from 10 classes: T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot. The dataset can be used as a drop-in replacement for the MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits

IV. PROPOSED ARCHITECTURE

In the architecture as shown Fig.1 the two dimensional Convolutional Neural Network used with RELU as activation function and 3x3 kernel, first the input is provided to the model. This is in the form of a 32x32x3 image. After three two dimensional Convolution layers a. Max pooling layer is provided with pool size of 2x2, as it selects the most dominant pixels in a block, and maps thatblock to that pixel value and overall matrix size reduces, without losing out on integrity and information. After this a Dropout layer of 25% is provided to control the overfitting of the model by randomly setting outgoing edges of hidden units to 0 at each updating phase.. After this we provide a flatten layer that converts the output of the convolutional layers into a single long feature vector. Dense layer is provided after the Dropout layer as each neuron in a dense layer is connected to every neuron in the previous layer and helps to learn complex relationships between the input and output data. Again a Dropout layer of 40% is provided to avoid further overfitting. Finally a Dense layer is provided with softmax as activation function.

DOI: 10.17148/IJARCCE.2023.125292

V. CONCLUSION AND FUTURE WORK

Active learning is a powerful technique that can be used to reduce the amount of data labeling required. Active learning can be a valuable tool for machine learning practitioners who are looking to improve the performance of their models while reducing the cost of data labeling. However, there are some challenges to applying active learning to these datasets. One challenge is that the labels of data points in these datasets are expensive to obtain. Anotherchallenge is that these datasets are not very diverse, which can make it difficult to train a model that generalizes well to new data. Despite these challenges, active learning is a promising approach for improving the performance of machine learning models on CIFAR10,EuroSAT and Fashion MNIST datasets. We believe that future research in active learning will focus on addressing the challenges of these datasets and developing even more effective active learning methods.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to our guide**Prof. Anala M R,** Department of Information Science, RV College of Engineering for her unwavering and consistent support at all times.

REFERENCES

[1] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach", International Conference on Learning Representations, 2018.

[2] S. Ravi and H. Larochelle, "Meta-learning for batch mode active learning", International Conference on Learning Representations workshop, 2018.

[3] L. Copa, D. Tuia, M. Volpi and M. Kanevski, "Unbiased query-by-bagging active learning for VHR image classification", Image and Signal Processing for RemoteSensing XVI, pp. 78300K, 2018.

[4] M. Li, R. Wang and K. Tang, "Combining SemiSupervised and active learning for hyperspectral image classification", Computational Intelligence andData Mining (CIDM) 2013 IEEE Symposium on, pp. 89-94, 2019

[5] D. Wang and Y. Shang, "A new active labeling method for deep learning," 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 2014, pp. 112-119, doi: 10.1109/IJCNN.2014.6889457.

[6] T. Kasarla, G. Nagendar, G. M. Hegde, V. Balasubramanian and C. V. Jawahar, "Region-based active learning for efficient labeling in semantic segmentation," 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2019, pp. 1109-1117, doi: 10.1109/WACV.2019.00123.

[7] J. Z. Bengar, J. van de Weijer, B. Twardowski and B. Raducanu, "Reducing Label Effort: Self-Supervised meets Active Learning," 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 1631-1639, doi: 10.1109/ICCVW54120.2021.00188.

[8] G. Lenczner, A. Chan-Hon-Tong, B. Le Saux, N. Luminari and G. Le Besnerais, "DIAL: Deep Interactive and Active Learning for Semantic Segmentation in Remote Sensing," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol.15, pp. 3376-3389, 2022, doi: 10.1109/JSTARS.2022.3166551.

[9] Chandra, A.L., Desai, S.V., Balasubramanian, V.N. et al. Active learning with point supervision for cost-effective panicle detection in cereal crops. Plant Methods 16, 34 (2020). https://doi.org/10.1186/s13007-020-00575-8.

[10] R. Sheikh, A. Milioto, P. Lottes, C. Stachniss, M. Bennewitz and T. Schultz, "Gradient and Log-based Active Learning for Semantic Segmentation of Crop and Weed for Agricultural Robots," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 1350-1356, doi: 10.1109/ICRA40945.2020.9196722.

[11] S. Desai and D. Ghose, "Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV),Waikoloa, HI, USA, 2022, pp. 1485-1495, doi: 10.1109/WACV51458.2022.00155.

[12] R. Sheikh, A. Milioto, P. Lottes, C. Stachniss, M. Bennewitz and T. Schultz, "Gradient and Log-based Active Learning for Semantic Segmentation of Crop and Weed for Agricultural Robots," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 1350-1356, doi: 10.1109/ICRA40945.2020.9196722.

[13] S. Desai and D. Ghose, "Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 1485-1495, doi: 10.1109/WACV51458.2022.00155.

[14] Settles, B. (2017). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

[15] Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach.

DOI: 10.17148/IJARCCE.2023.125292

International Conference on Learning Representations (ICLR)

[16] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conferenceon Machine Learning (ICML).

[17] Fang, Y., et al. (2017). Learning how to do active learning: A deep reinforcement learning approach. Advances in Neural Information Processing Systems (NeurIPS).

[18] Beluch, W. H., et al. (2018). The power of ensembles for active learning in image classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[19] Wei, L., et al. (2020). Bayesian active learning for clinical decision support using electronic health records. Journal of the American Medical Informatics Association (JAMIA).

[20] Siddiqui, S., et al. (2021). Active learning for graph classification: A comparative study. IEEE Transactions on Knowledge and Data Engineering (TKDE).

[21] Yang, T., et al. (2022). Efficient active learning with graph neural networks. International Conference on Machine Learning (ICML).

[22] Dasgupta, A., et al. (2020). Active learning with incomplete information. Advances in Neural Information Processing Systems (NeurIPS).

[23] Ash, S. D., et al. (2019). Diversity-based active learningwith applications to rare-class detection in data streams. Knowledge and Information Systems