Using neural networks to select what data to label

Christian Fuglede, Freja Tusindfryd Dollas, Søren Stange

s204206, s204248, s204229 Technical University of Denmark

Introduction

Neural networks have revolutionized problem-solving but in order to attain good results vast amounts of labeled data is required. In many fields, like medicine, where labeling requires expert validation, the labeling process is both expensive and time-consuming. Therefore the possibility of training models with little to none labelled data is a field of interest. Active Learning offers a way of identifying informative data points to label and after labeling iteratively improving the model by adding the carefully chosen data points to the training data. We seek to improve the accuracy of deep learning models trained on image data, specifically the ResNet18 model with the MNIST- and CIFAR-10 data, by implementing Active Learning with uncertainty sampling, targeting data points near the decision boundary. Additionally we explore Bayesian Neural Networks for enhanced uncertainty estimates and clustering methods to pinpoint high-value data points, to attain a high accuracy at a low cost.

Key points

- We train the **ResNet-18** model using the two different image data-sets, **CIFAR-10** and **MNIST**, to compare performance in respect to the training data.
- We train the **ResNet-18** model on different sizes sets of labeled data to show how accuracy is related to the training set size.
- We train the ResNet18 model using **margin-based, uncertainty sampling, K-means clustering** and **Bayesian active deep learning** respectively as different methods to locate which data points to label and compare the performance of these active learning approaches.
- From the generated results we see a clear advantage for using Active learning for selecting what data points to label in comparison to a randomized approach.

ResNet18 model

The ResNet-18 model is an 18-layer convolutional network presented in 2015 [1]. It belongs to the **ResNet family**, which stands for "Residual Networks". The network consists of 1 initial convolutional layer followed by 16 layers organized as 4 groups each consisting of 2 blocks with every block having 2 convolutional layers, and finally a fully connected layer. Every block has a residual connection that adds the input of the block to the output of the block (after the second convolution). This in turn introduces the nice trait, that input of a layer can bypass one ore more subsequent layers. This helps mitigate the problem of vanishing gradients, enabling the training of very deep networks by maintaining the flow of gradients during backpropagation. In this project, we have chosen the Adam optimizer to reduce the loss which is calculated with the Cross entropy loss function.

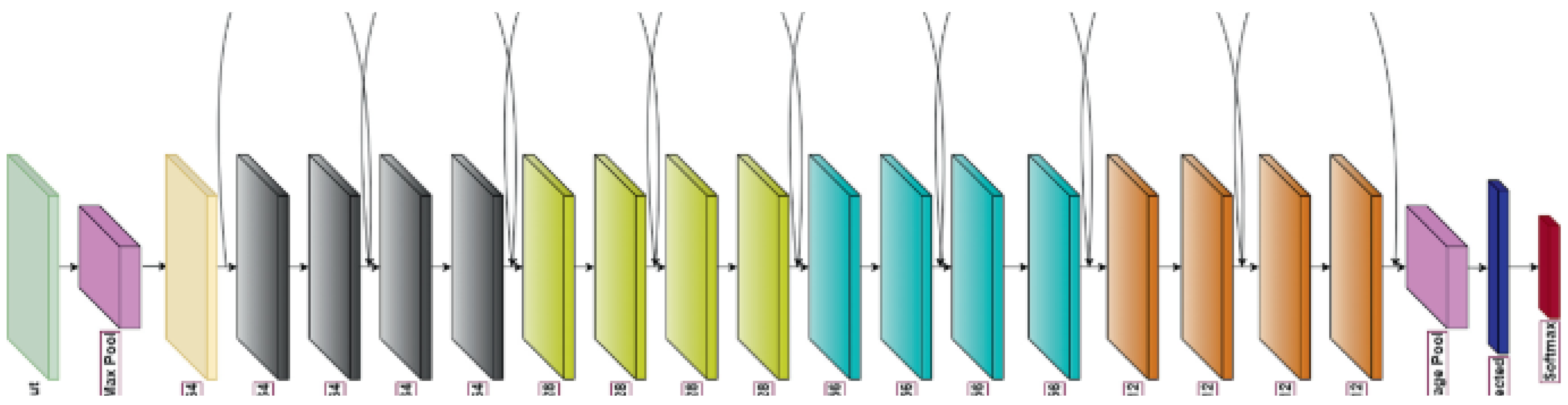

Figure 1: ResNet18 architecture representation. [3]

CIFAR-10 and MNIST datasets

The Modified National Institute of Standards and Technology database (MNIST database) is a database of 70.000 images of handwritten digits from 0 to 9 with labels, whereas 60.000 are training data and 10.000 are test data (28x28 grayscale)[2]. The CIFAR-10 (Canadian Institute for Advanced Research, 10 classes) is a database of 60.000 images (32x32 RGB) evenly distributed over 10 classes with labels, whereas 50.000 are training data and 10.000 are test data. In Figure 2 the accuracy of ResNet-18 models trained on different training set sizes on both datasets are visualized. For MNIST the accuracy becomes somewhat constant at around 1500 training set datapoints, while the accuracy for CIFAR-10 still increases at training set size 5500. This suggests that the models need to see a lot more examples of the CIFAR-10 to be able to differentiate the pictures and label them into the right classes, while models trained on MNIST data only need relatively few images to be able to classify the images correctly. However, for both datasets it is clear that larger training set size gives better accuracy. This shows the need for labeling data to attain larger training sets, but to minimize cost it is a field of interest to also minimize the amount of labeled data. To do this and not compromise of the quality of the predictions, active learning is an interesting tool since it can help to determine which datasets to label first and therefore maybe minimize the amount of labeling needed.

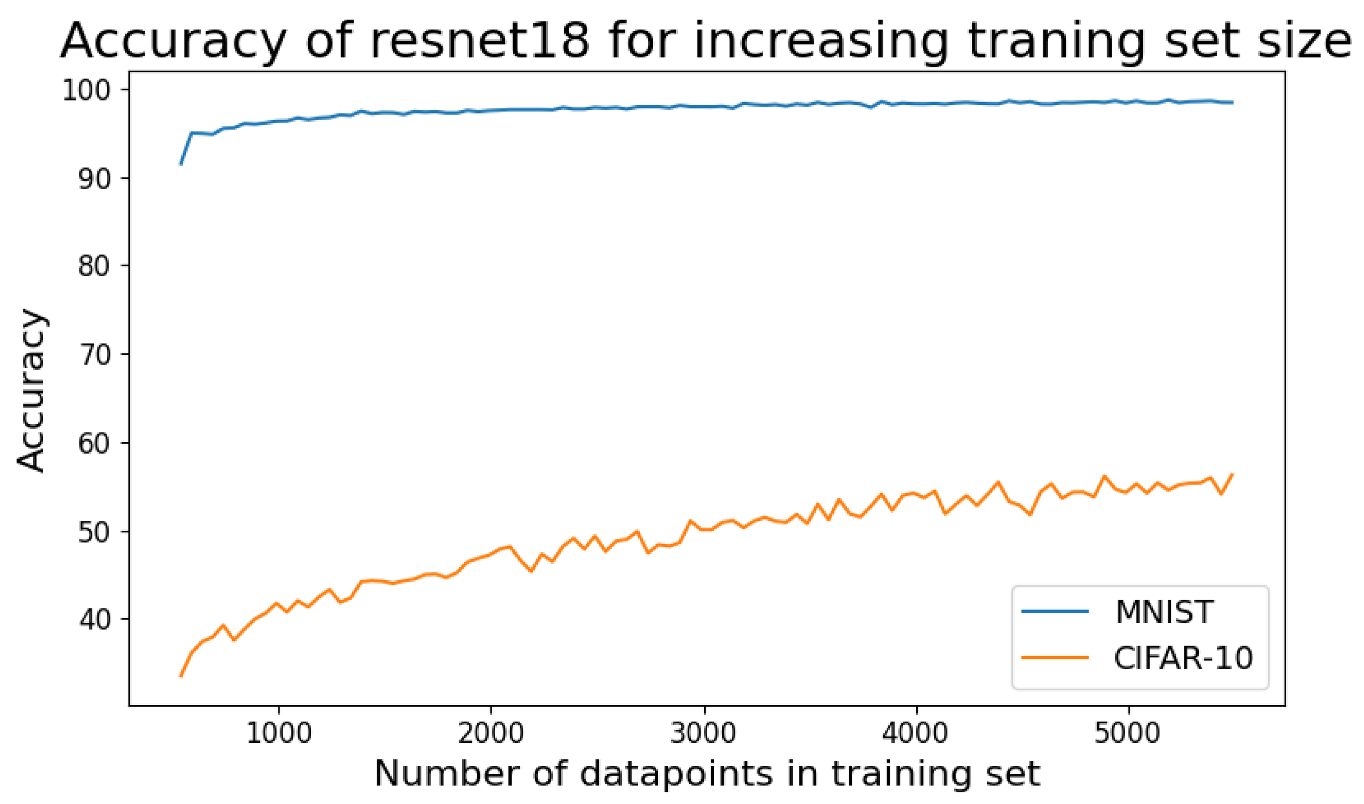

Figure 2: The accuracy of ResNet-18 visualized with increasing training set size for MNIST and CIFAR-10 respectively.

Active Learning methods

To test our active learning methods we use four different active learning approaches. Additionally we use an approach to compare if the active learning methods add to the performance. Therefore the baseline is designed to have as many additionally labeled datapoints to train on as the active learning approaches, however these datapoints are randomly chosen. In **margin-based sampling** the datapoints to label are chosen by looking at the prediction certainty of the highest and second highest predicted class. Points where these values are closest are then requested to be labeled. In **uncertainty sampling** the datapoints to label are chosen by looking at the prediction certainty of the highest predicted class. Points that have the lowest value for the highest predicted class are then requested to be labeled. In **K-means clustering** the datapoints are grouped in **12** clusters, and we seek to find boundary-points by determining the distance from every point to the assigned cluster centroid and the second closest centroid. The points with the smallest difference in these distances are then requested to be labeled. In **Bayesian** active deep learning[4] we seek to improve uncertainty estimates, using MC Dropout to simulate a stochastic learning pattern, datapoints with the highest entropy are selected for labeling, focusing on uncertain samples to enhance performance efficiently compared to the random baseline.

Model performance

In figure 3 and 4 we have tested the four different active learning approaches against a baseline model. We have chosen to run 150 label iterations, and for each label iteration we label 0.1% of the data points chosen by the active learning approach. We then compare these approaches to a baseline model, which is trained the number of data points we have after the last label iteration. However, these data points are chosen randomly and not through the iterative active learning cycle. We use a batch size of 64 when training, and each label iteration is trained on 100 epochs. In figure 3 and 4 the accuracy of the baseline and four active learning models are visualized for the CIFAR-10 and MNIST dataset respectively. It is clear that for the same amount of data points the baseline is outperformed by the active learning models. For both datasets the accuracy of the 4 active learning approaches are very similar. Additionally, for the CIFAR-10 dataset the accuracy of the 4 active learning approaches are not much better than that of the baseline model. For the same size training data the active learning approaches have atmost a 4% better accuracy than that of the baseline. For the MNIST dataset the accuracy of the 4 active learning approaches are better than the baseline for a much smaller dataset. This means that the amount of labeled data can be minimized when using this approach for this dataset and therefore a higher accuracy can be attained for a smaller cost.

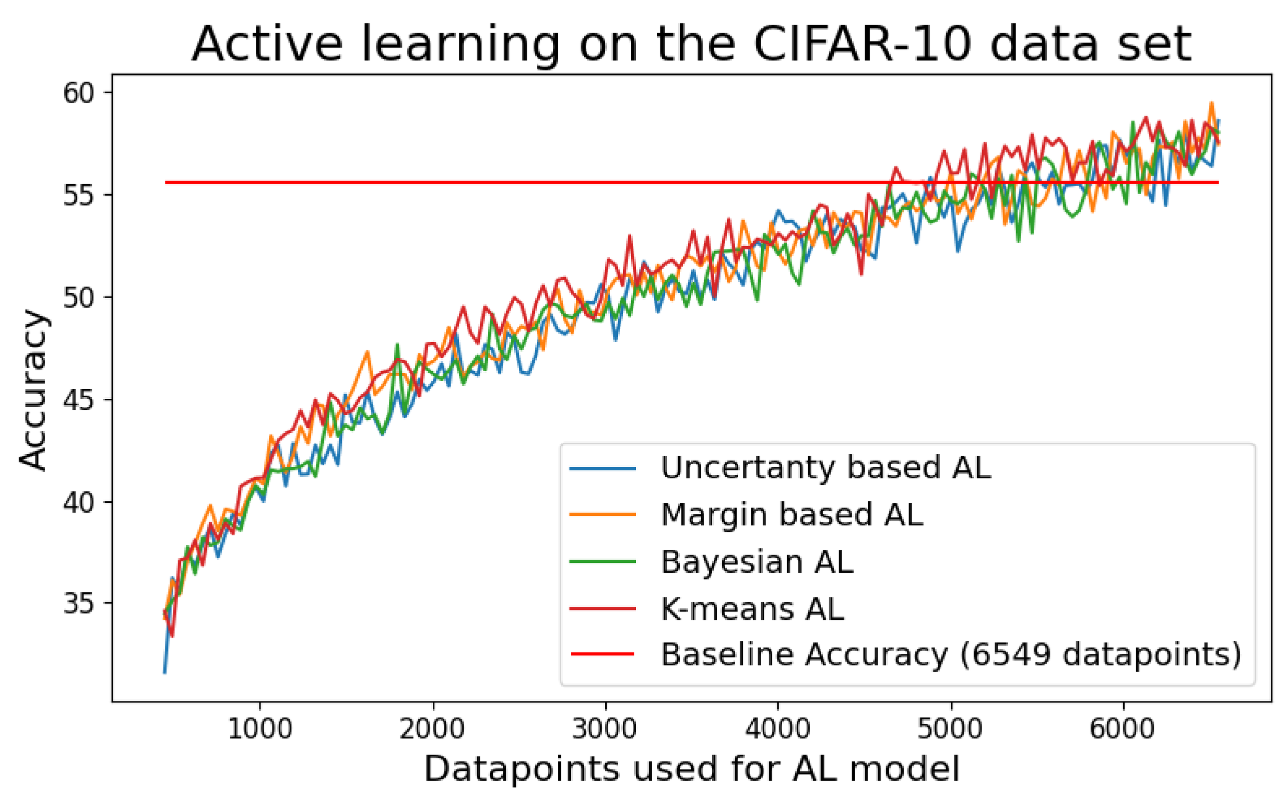

Figure 3: Accuracy of ResNet-18 trained on different sized CIFAR-10 datasets with data-points chosen using 5 different approaches

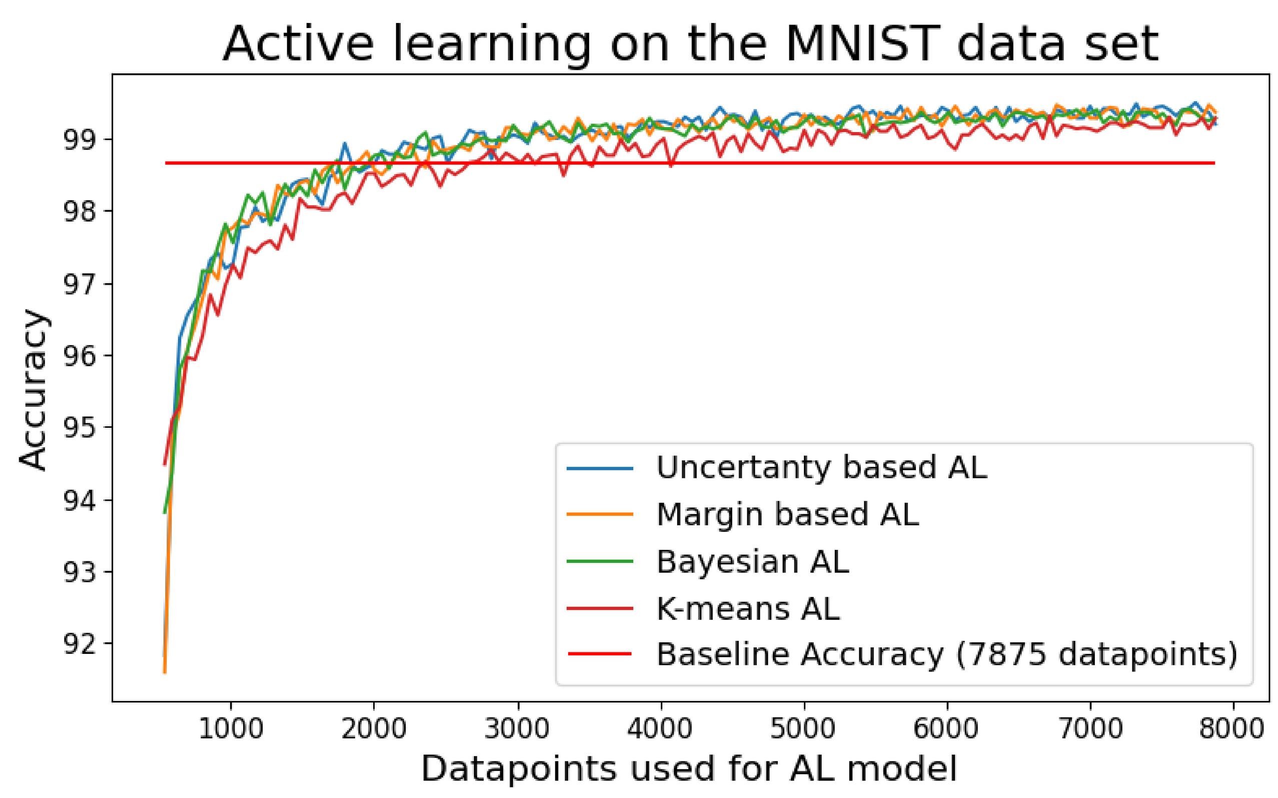

Figure 4: Accuracy of ResNet-18 trained on different sized MNIST datasets with data-points chosen using 5 different approaches

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[2] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, 1998. URL http://yann.lecun.com/exdb/mnist/. Accessed: 2024-12-08.

[3] N. Salam and T. Jemima Jebaseeli. Integrating resnet18 and yolov4 for pedestrian detection. In S. Roy, D. Sinwar, N. Dey, T. Perumal, and J. M. R. S. Tavares, editors, *Innovations in Computational Intelligence and Computer Vision*, pages 49–62, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-99-2602-2.

[4] M. Surajiwale. Doing more with less using bayesian active learning, 2020. URL https://product.hubspot.com/blog/bayesian-active-learning?fbclid=IwY2xjawGxRS1leHRuA2FlbQIxMAABHXdPpSTVdcquOaQH1JDkxOS_8w0zK3WNg7dbscw7MIiIHIs93ItNJ9QRzg_aem_3KIQHdTO5fnM_0ZdAp747A.