USING NEURAL NETWORKS TO SELECT WHAT DATA TO LABEL

Freja Tusindfryd Dollas (s204248), Christian Fuglede (s204206) & Søren Stange (s204229)

Technical University of Denmark

ABSTRACT

In the recent years there have been an ever-increasing interest in deep learning models and their capabilities in solving exceedingly difficult problems. From the first time the world heard about Chat-GPT to the Google DeepMind project winning the nobel prize in Chemistry in 2024[1], the interest has never been greater. However, these deep learning models are also known for the fact that they demand a lot of labeled data in order to get good performances. This data can be very expensive to attain, especially in fields such as medicine[2]. To combat this Active Learning (AL) is introduced. This serves as a method where the underlying neural network to be trained iteratively selects data points from an unlabeled pool of data. The chosen data points are then labeled in the hope that labeling exactly those chosen data points will result in an improved model performance. As few data points as possible are labeled, while still getting better performance with more data points. In this project four different AL methods are investigated: Margin-based sampling, uncertainty sampling, K-means clustering and Bayesian active learning. The results show that there definitely is a benefit in using AL models to label data, rather than using a randomized approach.

1. 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 ResNet-18 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.

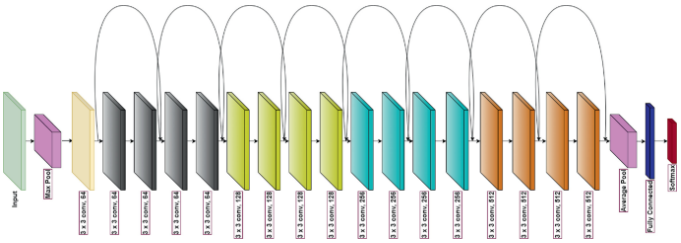

Fig. 1. ResNet-18 architecture representation. [4]

2. THE RESNET-18 MODEL

The ResNet-18 model is an 18-layer convolutional network presented in 2015 by Kaiming He et al. [3]. It belongs to the ResNet family, which stands for ”Residual Networks”. The ResNet-18 model contains 18 layers:

- An initial convolutional layer,
- 16 convolutional layers ordered in **4 groups** (shown as four squares with different colors in figure 1) with **2 blocks** each and every block having **2 convolutional layers**. Every block has a residual connection that adds the input of the block to the output of the block (after the second convolution),
- A fully connected layer.

The residual connections introduce the nice trait that input of a layer can bypass one or 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.

3. DATA SETS

In order to test how different active learning methods perform for improving model accuracy, we have chosen to do all of our experiments on two different image classification data sets: The MNIST data set and the CIFAR-10 data set.

Fig. 2. Example of the images in the MNIST data set.

Fig. 3. Example of the images in the CIFAR-10 data set.

3.1. MNIST

The Modified National Institute of Standards and Technology database (MNIST database) is a classic data set in deep learning litterature, and it often serves as a data set for those who are starting their journey in deep learing. It consists of 70.000 images of handwritten digits from 0 to 9, each image with a corresponding label stating what number it is. All of the images have been resized and centered such that they all have the same dimension which is 28x28 greyscale pixels. The data set is divided into a training set consisting of 60.000 images and a test set consisting of the remaining 10.000 images.[5]. We have depicted some examples of the MNIST dataset in Figure 2.

3.2. CIFAR-10

The Canadian Institute For Advanced Research (CIFAR) is a labeled subset of the 80 million tiny images dataset. It exists in two forms a CIFAR-10 and a CIFAR-100. The difference is that in the former there are 10 classes present and in the latter there are 100 classes present. In this project we have chosen to work with CIFAR-10 as the 10 classes serves for a good comparison to the MNIST dataset which in turn also has 10 classes. The data set consists of 60.000 images of dimension 32x32 RGB pixels. The classes are evenly distributed through the data set and consists of the classes: Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship and Truck. The data set is divided into 50.000 images in the training set and 10.000 images in the test set[6]. Examples of images from the CIFAR-10 data set are shown in Figure 3.

4. MOTIVATION FOR ACTIVE LEARNING

In Figure 4 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 data points, 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 data points to label first.

5. ACTIVE LEARNING METHODS

To compare 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 with a baseline designed to have as many additionally labeled data points to train on as the active learning approaches, however these data points are randomly chosen.

5.1. Margin-based sampling

We use the active learning approach margin-based sampling, which is a method where the data points to label are chosen by looking at the prediction certainty of the highest and second

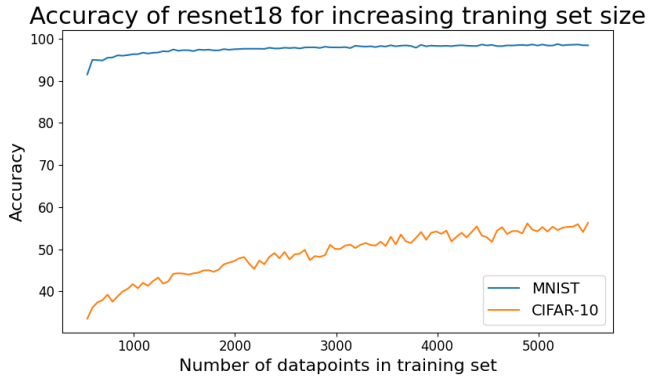

Fig. 4. Experiment with increasing trainings set size for both MNIST and CIFAR-10 in regards to their Accuracy.

highest predicted class. Points where these values are closest are then requested to be labeled[2].

5.2. Uncertainty sampling

We use the active learning approach uncertainty sampling, which uses the data points with the lowest prediction certainty of the highest predicted class, and chooses these for the next label iteration. Thus the data points which the model is most certain belong to a specific are assumed to be correctly labeled, and the data points which the model is most uncertain where belong are deemed to be the data points the model can learn the most from[2].

5.3. K-means clustering

We use K-means clustering to group the data points 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[7]. The points with the smallest difference in these distances are then requested to be labeled. We tested the performance of the model trained on MNIST-data when having 10 cluster, 25 clusters and many options in between. It makes no sense to have less than 10 clusters, because there are 10 classes in the data sets, however there might be things like subtypes and variations that can be identified by additional clusters.

5.4. Bayesian active deep learning

In Bayesian Active Deep Learning (BADL) we use Monte Carlo (MC) dropout to make a stochastic forward pass to get slightly different predictions in the neural network[8].

$$\text{Entropy} = -\sum_{i=1}^{C} P_i \log(P_i) \tag{1}$$

$$\text{Average Entropy} = \frac{1}{N}\sum_{k=1}^{N} \text{Entropy}^{(k)} \tag{2}$$

$$P_{\text{avg},i} = \frac{1}{N}\sum_{k=1}^{N} P_i^{(k)} \tag{3}$$

$$\text{Average Predictions} = -\sum_{i=1}^{C} P_{\text{avg},i} \log(P_{\text{avg},i}) \tag{4}$$

$$\text{Final Entropy} = \text{Average Predictions} - \text{Average Entropy} \tag{5}$$

In the equations above[9] we see how the final entropy score for a datapoint is calculated. We start with getting a prediction and an entropy of the prediction of the data points for each class. Then entropies are calculated from the predictions as seen in equation 1, here we have that i is index for the class in C classes. Then an average of the entropies are calculated in equation 2 with N being amount of epochs in the MC dropout. In equation 4 we calculate the average entropy across the classes from the average predictions in the MC dropout which makes us able to get the final entropy of a datapoint in equation 5. When this has run over all the data points, this gives us a sorted list of entropies for each data point

6. MODEL PERFORMANCE

In Figure 6 and Figure 5 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 6 and Figure 5 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 data sets 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 at most 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.

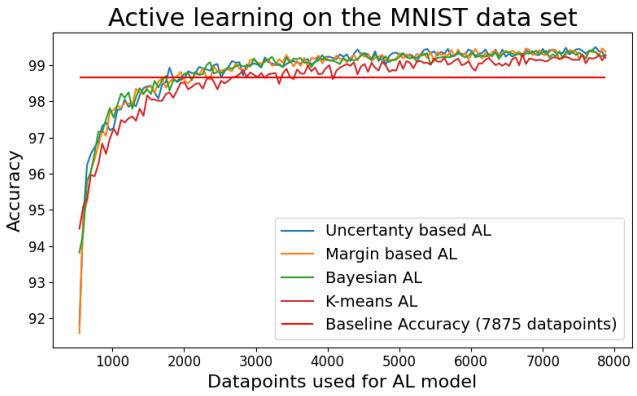

Fig. 5. Comparison of the accuracy of the differente AL methods on the MNIST data set.

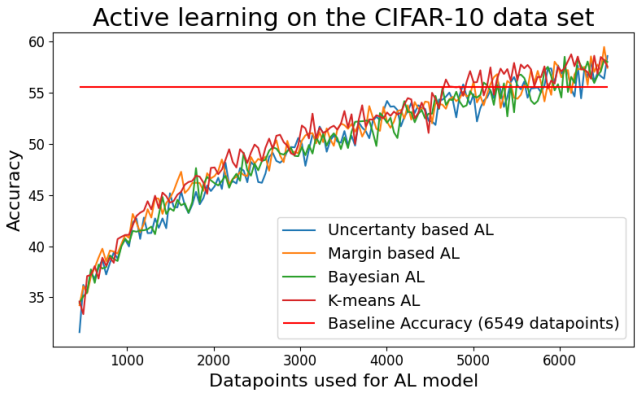

Fig. 6. Comparison of the accuracy of the differente AL methods on the CIFAR-10 data set.

7. DISCUSSION

From the results shown in 6 and Figure 5 shows that the amount of labeled data can be minimized when using an active learning approach for both data sets and still attain the same accuracy. This is certainly the case for the MNIST dataset, however also the CIFAR-10 model has better accuracy for the active learning approaches than the baselines. This means that overall higher accuracy can be attained for a smaller cost. However, these results can possibly be greatly improved in different ways.

In the ResNet-18 model there is architectural regularization in the way that batch normalization is applied after each convolutional layer and that the residual connections minimize overfitting. However the chosen optimizer, Adam with cross entropy, learning rate 0.0005, has no added regularization. This possibly would be an interesting feature to add and other parameters could as well be investigated. Additionally the batch size of 64 and the number of 100 epochs could also be revised. In K-means clustering the number of clusters, 12, is chosen based on performance of the model trained on MNIST data. However, it is used on the CIFAR-10 data as well, and here another choice of number of clusters could be relevant.

Also different active learning approaches could be implemented to reach an even better model performance and other data sets could be used to train and test the performance of models.

To attain better generalized results the choice of performance need to be considered. Here accuracy is used and this is limited in the way that it only measures how many correct predictions are made of all predictions without considering if any one class is underrepresented in correct predictions or other details that could be valuable in future work. Therefore it should be considered to add additional performance measures like visualizing the confusion matrix, the AUC-ROC for multiclass problems etc.

8. CONCLUSION

The field of active learning is very interesting in minimizing the costs of labeling data while retaining high performance. The performance of the four models in this work clearly has a better performance than a randomized approach of labeling data. This in turn means that companies developing deep learning models, can get more value for money by using AL methods when attaining data, rather that using a randomized approach.

9. GITHUB LINK

A link to the github repository is provided here:
https://github.com/fuglede30/deep_learning_project.[10]

10. REFERENCES

[1] The Royal Swedish Academy of Sciences, "The nobel prize in chemistry 2024," NobelPrize.org, October 2024, Press release.

[2] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang, "A survey of deep active learning," 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.

[4] Nader Salam and T. Jemima Jebaseeli, "Integrating resnet18 and yolov4 for pedestrian detection," in *Innovations in Computational Intelligence and Computer Vision*, Satyabrata Roy, Deepak Sinwar, Nilanjan Dey, Thinagaran Perumal, and João Manuel R. S. Tavares, Eds., Singapore, 2023, pp. 49–62, Springer Nature Singapore.

[5] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges, "The mnist database of handwritten digits," 1998, Accessed: 2024-12-08.

[6] Alex Krizhevsky, "Cifar-10 and cifar-100 datasets," 2009, Accessed: 2024-12-10.

[7] Fedor Zhdanov, "Diverse mini-batch active learning," 2019.

[8] Mukul Surajiwale, "Doing more with less using bayesian active learning," 2020.

[9] Zoubin Ghahramani 1 Yarin Gal, Riashat Islam, "Deep bayesian active learning with image data," 2017.

[10] Freja Dollas Søren Stange, Chistian Fuglede, "https://github.com/fuglede30/deep_learning_project," .