EPiC Series in Computing

Volume 65, 2019, Pages 152–162

GCAI 2019. Proceedings of the 5th Global
Conference on Artificial Intelligence

EPiC
Computing

On the Robustness of Active Learning

Lukas Hahn[1,2], Lutz Roese-Koerner[2], Peet Cremer[2], Urs Zimmermann[2],
Ori Maoz[2], and Anton Kummert[1]

[1] University of Wuppertal,
Department of Electrical Engineering
Wuppertal, Germany
lukas.hahn@uni-wuppertal.de
[2] Aptiv,
Wuppertal, Germany
firstname.lastname@aptiv.com

Abstract

Active Learning is concerned with the question of how to identify the most useful samples for a Machine Learning algorithm to be trained with. When applied correctly, it can be a very powerful tool to counteract the immense data requirements of Artificial Neural Networks. However, we find that it is often applied with not enough care and domain knowledge. As a consequence, unrealistic hopes are raised and transfer of the experimental results from one dataset to another becomes unnecessarily hard.

In this work we analyse the robustness of different Active Learning methods with respect to classifier capacity, exchangeability and type, as well as hyperparameters and falsely labelled data. Experiments reveal possible biases towards the architecture used for sample selection, resulting in suboptimal performance for other classifiers. We further propose the new "Sum of Squared Logits" method based on the Simpson diversity index and investigate the effect of using the confusion matrix for balancing in sample selection.

1 Introduction

The term Active Learning describes the field of selecting samples from a given pool of data in order to subsequently train a Machine Learning algorithm with. This can be done for two major reasons: Firstly, deciding which subset of collected data will be annotated in order to create training and validation set for a supervised Machine Learning task. While it can be comparatively easy and inexpensive to record and gather sensor data and ever decreasing cost makes it affordable to possibly neglect storage expenses, reliable ground truth annotation still requires manual labour and is therefore the crucial factor. In industrial application of Machine Learning to various tasks, budget and time constraints play a significant role and performance can depend on choosing the best n samples to train on.

Secondly, one could think of using Active Learning methods as a form of regularization. While increasing the number of available training samples is in general regarded as helpful, certain factors can lead to an impaired performance when doing so. The more objects in a

D. Calvanese and L. Iocchi (eds.), GCAI 2019 (EPiC Series in Computing, vol. 65), pp. 152–162

recognition task are standardized, the more redundant information is potentially added to the dataset with each new sample, which can result in worsened generalization. Active Learning methods can also be applied to sanitize a dataset from falsely labelled samples, as a suitable strategy will not pick samples with a conspicuous difference between label and prediction.

However, despite the great potential of Active Learning it also bears significant risks. If applied in an incorrect way it could lead to a sub-optimal sample selection, and, in the worst case, rendering the complete Machine Learning task unsuccessful. In order to point out how to avoid these pitfalls, we examine a set of known Active Learning query strategies, as well as some extensions of our own, and their performance on various different image classification datasets. We then view their performance under different aspects including changing hyperparameters, influence of falsely labelled data and the replaceability of varying CNN architectures. Eventually we regard the performance of the same strategies when applied to a problem of hierarchical classifiers. Our main contributions are:
1.) A robustness investigation of state-of-the-art Active Learning strategies with respect to the impact of falsely labelled data, hyperparameters and the impact of changing the classifier model during the selection phase. 2.) An extension of the Active Learning method based on Entropy computation using the Simpson Diversity. 3.) Theoretical insights and experimental results for Active Learning on Hierarchical Neural Networks.

2 Related Work

An overview of methods from the pre Deep Learning area can be found in the very comprehensive review of [18]. Many approaches originating from that time (e.g. Uncertainty sampling, Margin based sampling, Entropy Sampling, ...) have been later adapted to neural networks. Additional examples for this include the approach of [16], who applied a Monte Carlo method to compute an estimated error reduction that can be used for sample selection as well as clustering approaches like those described in [14] and [3].

[22] and [15] propose a semi-supervised approach. They use Active Learning to query samples which the network has not yet understood and use label propagation to also utilize well understood samples with ”pseudo-labels”.

In the field of supervised learning, [10] used a Bayes approach to distil a Monte Carlo approximation of the posterior predictive density for sample selection. In the theoretical work of [9], Active Learning was rephrased as a convex optimisation problem and the balancing of the selection of samples with high diversity and those that are very representative for a subset are discussed. Unlike many other methods, the core-set approach of [17] does not use the output layer of a network for Active Learning. Instead they solve a relaxed k-centres problem to minimize the maximal distance to the closest cluster centre for each sample in a space that is spanned by the neurons of a hidden layer of a network. As discussed later, this approach has a very high independence of the actual classes of a network, which can be helpful when dealing with hierarchical networks [23] for example.

[6] introduced the concept of live-dropout to Active Learning. The idea is to approximate the behaviour of an ensemble of Bayesian estimators by activating dropout during inference and multiple forward passes. They furthermore developed an Active Learning framework which is able to use this and other deep Bayesian methods. In the same line of thought, [4] investigated live dropout and Query-by-committee methods. However, [2] used ensembles of CNNs with identical architectures but different weight initiations to show that ensembles work better than ”ensemble approximation methods” like the above mentioned MC dropout of [6] or approaches based on geometric distributions like [17].

Some recent approaches also utilize ”meta” knowledge for Active Learning. [5] introduced ”Policy based Active Learning”. There, reinforcement learning is used for stream based Active Learning in a language processing setting. This is very similar to the approach of [1] who proposed ”Learning Algorithms for Active Learning”. They also used Reinforcement Learning to jointly learn a data representation, an item selection heuristic and a method for constructing prediction functions from labelled training sets. [8] reuse knowledge from previously annotated datasets to improve the Active Learning performance.

3 Methods

In the following we review existing methods from the field of pool-based Active Learning and propose a suggestion of our own. Given a classification model θ and a dataset $\mathcal{D}$, consisting of a feature and label pair $\langle x \in X, y \in Y \rangle$, such an algorithm has the following structure:

```
Input : $\mathcal{L} \subset \mathcal{D} = \{\langle x, y \rangle\}$ : Labelled set,
        $\mathcal{U} = \mathcal{D} \setminus \mathcal{L} = \{\langle x, ? \rangle\}$ : Unlabelled set,
        $\theta$ : Classification model,
        $\phi$: Active Learning query function
while $|\mathcal{U}| > 0 \wedge$ no stopping criterion do
    $\theta = train(\mathcal{L})$;
    for all $\langle x, ? \rangle \in \mathcal{U}$ do
        Compute Active Learning metric $\phi(x)$ under $\theta$;
    end
    Choose $x^\star$ with highest magnitude of $\phi(x^\star)$ ;
    Annotate $y^\star$;
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{\langle x^\star, y^\star \rangle\}$;
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\langle x^\star, y^\star \rangle\}$;
end
```

Algorithm 1: Pool-Based Active Learning.

Considering a large dataset, one can query numerous samples at once. This set of the chosen samples is denoted by $\mathcal{B} \subset U$ [9].

We take a closer look at uncertainty sampling, a strategy that selects samples the classifier is uncertain about. In this context uncertainty means a low confidence for the predicted class that is given by $\hat{y} = \text{argmax}_y \; P_\theta(y|x)$. We consider three commonly used uncertainty measures:

(a) *Least Confident:* $x^\star_{LC} = \text{argmax}_x \, (1 - P_\theta(\hat{y}|x))$

(b) *Margin:* $x^\star_M = \text{argmin}_x \, (P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x))$

(c) *Entropy:* $x^\star_H = \text{argmax}_x \, (-\sum_i \; P_\theta(y_i|x) \log P_\theta(y_i|x))$

(a): Considering only one class label, the sample $x^\star_{LC}$ with the least confident label prediction is selected. (b): Margin sampling includes information about the second most certain prediction. The algorithm queries the sample $x^\star_M$ with the smallest difference between the two most probable class labels. (c): For multi class tasks, it is relevant to consider all label confidences. For each sample every class probability is weighted with its information content and summed up. The algorithm queries the sample with the highest entropy $x^\star_H$ [18].

For the following experiments we implement eight query strategies.
Based on Least Confident (a):

Naive Certainty (NC) Low: Select n samples with the minimal maximal activation in classifier logits. Since basing the decision only on the one highest activated neuron is a very straightforward approach, we call this family of strategies the "Naive" methods.

NC Range: Select n samples within a certain range of the classifier logits' activation (e.g. $[0.1, 0.9]$).

NC Diversity: Select n samples with the minimal maximal activation in classifier logits and additionally prevent that similar samples are chosen by calculating the diversity of the samples below the threshold compared to those already included in the training set.

NC Balanced: Select n samples with the minimal maximal activation in classifier logits and balance the class distribution using the reciprocal value of the classification confusion matrix obtained with the previous training set. Terminates if one class contains no more samples to be drawn.

Based on Margin (b):

Margin: Select n samples with the smallest difference of the two highest firing logits.

Based on Entropy (c):

Entropy High: Select n samples with the highest entropy.

Sum of Squared Logits (SOSL): Select n samples with the highest Simpson diversity index $D = 1 - \sum_i (l_i)^2$ [20] (cf. 3.1).

Core Set Greedy: A similarity measure in the embedding space. Creates a core set by approximating the problem of distributing k-centres in n points, such that the minimal distance of all points to the nearest centre is maximized. Select n samples for which the minimum distance to all samples which are already part of the training set is maximized (cf. [17]).

3.1 Sum of Squared Logits (SOSL) Method

In Active Learning, we require a measure of how sure the classifier is that its class decision during inference is accurate. One possibility for such an accuracy-of-inference measure is to analyze the distribution of logits. Within the trained model of the classifier, the logits can be interpreted as probabilities that the inferred sample belongs to the class associated of the respective logit. If the logits are strongly biased in favour of a certain class, it is very likely that the given sample belongs to the class corresponding to the strongest logit. On the contrary, if the logits do not show a clear preference for a certain class, there is a high risk that taking the class of the strongest logit results in a false prediction. In other words, to which degree the distribution of logits tends towards peaks rather than an equipartition indicates how accurate the inference is going to be.

In previous literature, the Shannon entropy [19] has been frequently used as a measure of how peaked or equipartitioned a distribution is. A valid strategy for Active Learning could then be to sort out those samples, for which the Shannon entropy $H = -\sum_i l_i \log(l_i)$, with l_i being the values of the logits, is particularly high. However, a shortcoming of this approach is that it does not adequately account for the situation when the the distribution of logits is admittedly strongly peaked, but with peaks on more than one class logit. Such a situation can easily arise in samples, when they belong to classes showing similarities and the classifier's model does not yet feature a clear decision boundary between them. In such a case, the distribution of logits is still far away from an equipartition, resulting in a relatively low value for the Shannon entropy H. Thus, although labelling these samples would be particularly valuable for fleshing out the decision boundary and allowing the classifier to better separate between classes, they would not be added to Active Learning training set.

To overcome these shortcomings of the Shannon entropy H as a measure for characterizing the distribution of logits l_i, we propose to use the Simpson diversity index $D = 1 - \sum_i (l_i)^2$ [20] instead. The closer the distribution l_i is to an equipartition, the larger D becomes. If the l_i shows a strong peak at a certain i, D is close to zero. Finally, if the l_i are strongly peaked among several classes, D will have a small-to-moderate value between zero and one. The latter property of D in particular allows to select those samples for labelling, for which the classifier can narrow the class decision down to a few classes, among which it is still unsure. The Active Learning strategy is then to select in each iteration the n samples with highest D.

4 Experiments and Results

We conduct a series of experiments with the query strategies presented in section 3 on six different datasets for image classification (cf. Table 1). These consist of the well-known digit classification set MNIST [12] and the thereof inspired dataset of the Latin alphabet CoMNIST [21] and clothing classification Fashion-MNIST [24], as well as general object classification CIFAR-10 [11] and the house number collection SVHN [13]. We furthermore evaluate strategies on a private dataset of 33 different classes of traffic signs (TSR) represented through small grey scale images.

4.1 General Performance

Before we analyse the robustness of the presented query strategies, we compare their general performance on the datasets presented above. For each dataset we use a distinct plain feed-forward CNN. Only for CIFAR-10 we use an implementation of ResNet50 [7]. As we are not aiming to find the best architecture for a certain problem but to identify the most promising samples, we choose the number of layers and channels according to the approximate complexity of the task and select learning rates and batch sizes in commonly used ranges.

For all of these experiments, we start with a training set of 100 samples per class of the particular dataset. We train the CNN for up to 1000 epochs with an early stopping of 200. For this purpose we split 10% of the training set into an additional "development set". It is not used for training but to validate classification over the course of the training. This is done to obviate an overfitting-like bias with the use of early stopping. Of course the validation accuracy is then determined on the original test set of the respective dataset, using the best network weights acquired during training according to the development set accuracy.

	CIFAR-10	CoMNIST	Fashion-MNIST	MNIST	SVHN	TSR
Classes	10	26	10	10	10	33
Image Size	32×32	32×32	28×28	28×28	32×32	34×34
Channels	3	1	1	1	3	1
Training Samples	50 000	9 918	60 000	60 000	73 257	265 774
Validation Samples	10 000	1 300	10 000	10 000	26 032	66 443

Table 1: Characteristics of the datasets used for the experiments.

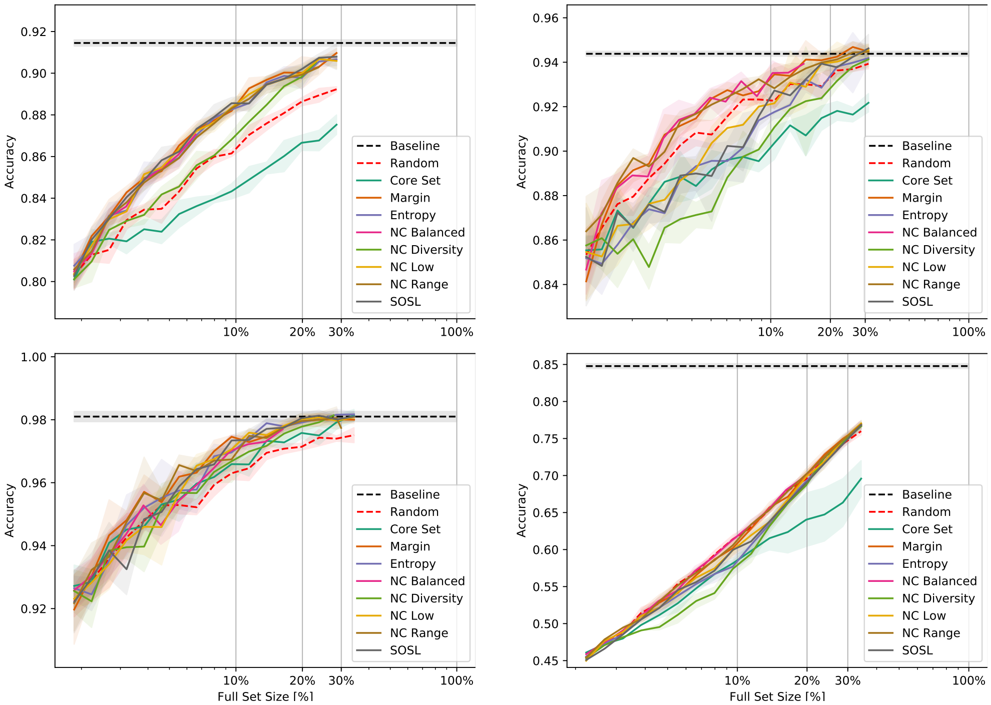

Figure 1: Classification accuracy over training set size for all strategies on Fashion-MNIST (top left), TSR (top right), MNIST (bottom left) and CIFAR10 (bottom right). The plotted value is the median of five runs and the shaded area denotes one standard deviation.

This network is also the one used to then select new samples to be added to the training set utilizing the query strategies. With each iteration we increase the number of samples in the training set by 20%. In all cases we conduct five repetitions per strategy and dataset for statistical significance. To reduce the computational burden, we iteratively draw new samples until we have reached approximately a third of the full size of the respective training set.

Figure 1 illustrates the results of the evaluation of all query strategies. Nearly all findings show a benefit of Active Learning methods and at least some of the query strategies are either hitting the baseline, or are close to it, around the 30% mark. For CIFAR-10 however, this is not true. None of the methods show any profit for this dataset and are in line with the random sample selection, resulting in a nearly perfectly linear increase in accuracy. This does not come as a surprise, as CIFAR-10 has very diverse representations of its classes and seems to contain no redundant information.

4.2 Changing Hyperparameters and Falsely Labelled Data

As hyperparameter optimisation is very important in fine-tuning the performance of Machine Learning algorithms, we analyse how much changes in these parameters influence the usability of the Active Learning methods shown.
Figure 2 shows the effect of altering the learning rate over two magnitudes and the batch size up to a factor of 16, for experiments on MNIST. All methods behave very robustly and do not show to be influenced by these alterations.

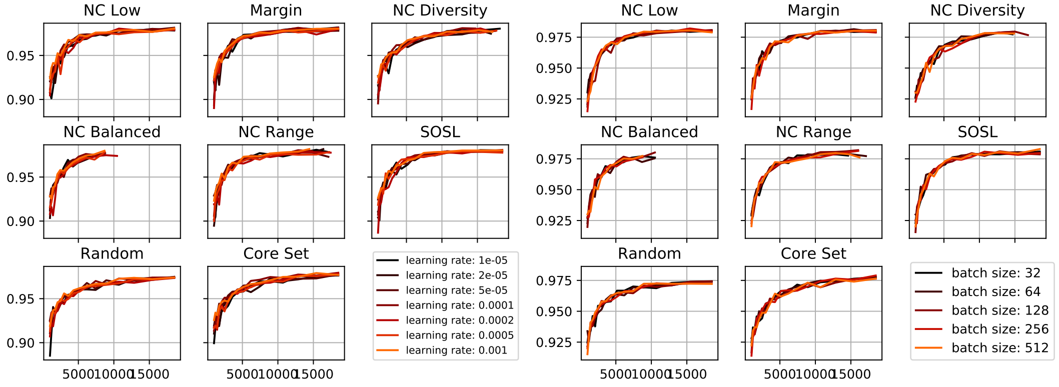

Figure 2: Changes in learning rate (left) and batch size (right) for various Active Learning strategies on MNIST. The plotted value is the median of five runs.

Since it can be expected that human annotation, especially in large scale labelling of sensor data, is never perfectly accurate, it is interesting to investigate how this might interfere with the applicability of Active Learning. In Figure 3 (left) we show results for an experiment where we purposely introduced false labels into the Fashion-MNIST training set. It can clearly be seen, that methods relying on a diversity criterion (NC Diversity, Core Set) suffer the most, since their selection process prevents similar samples from being chosen and therefore it can be harder to correct the negative impact that the selection of a wrongly labelled sample would have. Please note that these strategies also show the highest sensitivity to changes in dropout (cf. Figure 3 right).

4.3 Replaceability of Classifiers

In the application of Machine Learning, especially in product context, successive refinement of the algorithm is very common. A CNN architecture might be adjusted several times over the course of development or a production process, to optimise the performance or to adapt to changes in the dataset or external restrictions like computational resources. We investigate how the usability of Active Learning might be influenced, if data selection is done by a different network than the one eventually targeted for classification performance. For this purpose we implemented three CNNs, referred to as Min, Med and Max in the following, of different capacity to iteratively select samples from Fashion-MNIST with the query strategies as described above. We then perform a cross-training, where every network is trained with the selections of the others and its own. To ensure comparability, we use the same initial dataset of 100 samples per class for all classifiers and repeat calculations five times.

Figure 4 shows the results for selected strategies. Apart from information about the replaceability of classifiers, these results can show how the classifier capacity itself influences the applicability of Active Learning strategies. For the example of NC Balanced we can note a bias for the own selection performing best with the Max and Min classifier, while the medium-sized one shows indifference. The "weaker" the network gets, the better the performance of the random selection becomes. For the SOSL, this becomes even more clear. While the selection of the Max classifier is still definitely the best for itself, the smaller networks show the best performance with the randomized set.

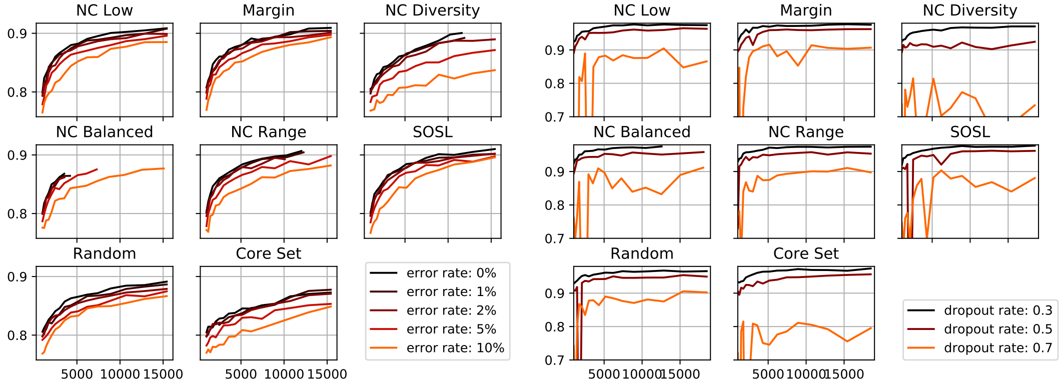

Figure 3: Active Learning strategy results for various synthetic labelling error rates on Fashion-MNIST (left) and different dropout regularization rates on MNIST (right). The plotted value is the median of five runs.

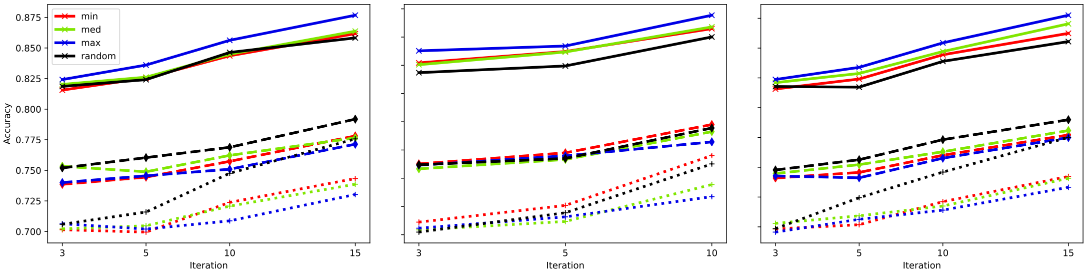

Figure 4: Results (mean of five runs) of cross training of different classifiers, Max (solid line), Med (dashed), Min (dotted), with the sample selection of the other networks and its own (blue, green, red), compared against a random selection (black), for the strategies Entropy High (left), NC Balanced (mid) and SOSL (right). Evaluations are made after 3, 5, 10 and 15 Iterations of querying new samples, except for NC Balanced which already terminated before the 15th iteration.

The results with Entropy High are very similar, but the gaps become even more obvious. *Max* now shows a very clear preference for the own selection compared to any other and the performance of the Active Learning strategy selection on the *Min* network is now more than three percentage points behind random.

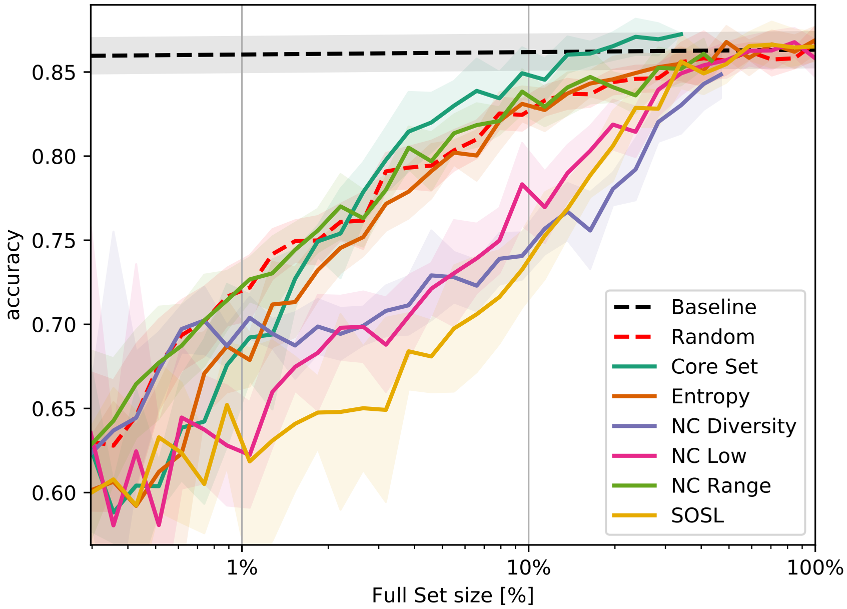

Figure 5: Classification accuracy over training set size for different Active Learning methods on hierarchical classifier for hand gesture classification. The plotted value is the median of ten runs and the shaded area denotes one standard deviation.

4.4 Hierarchical Classifiers

To complete our Active Learning robustness study, we examine a neural network structure different from the straightforward CNNs in the preceding sections.

Hierarchical or cascaded classifiers do not use a single label per sample but a whole label tree (cf. [23]). Consequently, label vectors consist of one of the three following options per class: ”1, 0 or not applicable” and each sample belongs to exactly one class per hierarchy level. Furthermore, during the learning phase each class is treated independent of all others. If we have an n-class classification problem, n ”1-vs-all classifiers” are trained.

This renders all Active Learning strategies which rely on quantifying the uncertainty of the logits useless. All of them (e.g. Naive Certainty, Margin) implicitly rely on the assumption that labels with two possible states are used. As the neurons that belong to classes marked as ”not applicable” are not considered during backpropagation (cf. [23]) they can take arbitrarily high values and thus confuse the mentioned Active Learning methods. As can be seen in Figure 5 this can even result in worse performance than random sampling. However, we can show, that methods, which work in the embeddings space (like the Core Set method), are not effected and thus are also employable for hierarchical neural networks.

4.4.1 Used Dataset

In all experiments with the hierarchical classifier we use a private dataset that consists of 12 classes which depict different poses of a human hand (e.g. ”One finger”, ”Two fingers”, ”Fist Thumb Left”, etc.). We use a training set of 670 000, a development set of 75 000 and a test set containing 8 000 grey scale images of size 22 × 46.

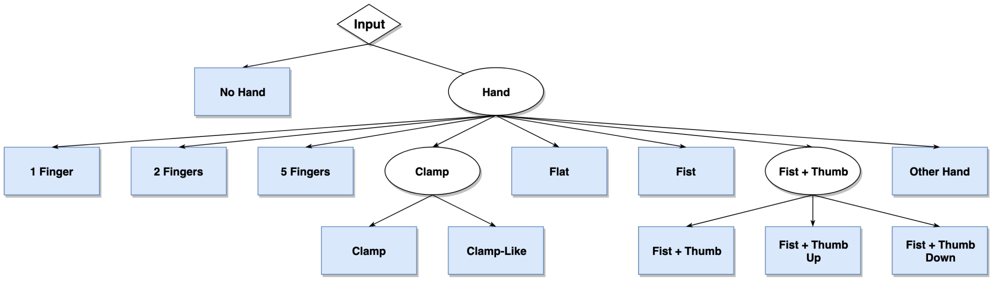

Figure 6: Hierarchical labels for hand gesture recognition. Blue boxes denote the 12 classes.

As depicted in Figure 6, we use three levels of hierarchy: 1.) ”Hand”/”No hand”, 2.) Class, 3.) Subclass. A sample of ”Fist Thumb Left” e.g. would have the labels ”Hand + ”Fist Thumb” + ”Fist Thumb Left”. Especially the neurons of the subclasses often have the label ”not applicable” as each subclass belongs to only one class.

5 Conclusion

We have presented a study on the robustness of Active Learning. While we show that even plain methods can bring a notable profit in different image classification applications, we emphasise, that prior knowledge about the data and the Machine Learning algorithm in use is essential for successful application. As seen in 4.1, methods that work well on a number of datasets might suddenly fail on a different one and certain data collections might be inherently unsuitable for this kind of active data selection. Although many changes in hyperparameters and erroneous labels might not influence the performance of particular strategies on one hand (cf. 4.2), classifier changes on the other can by all means (cf. 4.3). Critical alterations in the way a Machine Learning tasks is tackled, like switching from a straightforward to a hierarchical classifier (cf. 4.4), can turn the all previous findings upside down.

These findings underline, that Active Learning can be a helpful tool in data science, but has to be used with knowledge about the targeted utilisation. We aim to continue our endeavours in this field and expand the considerations to segmentation problems and ways to automatically provide assessment on promising combinations of data, Machine Learning algorithms and Active Learning strategies, to avoid possible pitfalls like the ones presented in this work.

References

[1] Philip Bachman, Alessandro Sordoni, and Adam Trischler. Learning algorithms for active learning. *CoRR*, abs/1708.00088, 2017.

[2] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 208–215, New York, NY, USA, 2008. ACM.

[4] Melanie Ducoffe and Frédéric Precioso. Active learning strategy for CNN combining batchwise dropout and query-by-committee. In *25th European Symposium on Artificial Neural Networks, ESANN 2017, Bruges, Belgium, April 26-28, 2017*, 2017.

[5] Meng Fang, Yuan Li, and Trevor Cohn. Learning how to active learn: A deep reinforcement learning approach. *CoRR*, abs/1708.02383, 2017.

[6] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *ICML*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[8] Fabian Caba Heilbron, Joon-Young Lee, Hailin Jin, and Bernard Ghanem. What do i annotate next? an empirical study of active learning for action localization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[9] Maya Kabkab, Azadeh Alavi, and Rama Chellappa. DCNNs on a diet: Sampling strategies for reducing the training set size. *CoRR*, abs/1606.04232, 2016.

[10] Anoop Korattikara, Vivek Rathod, Kevin Murphy, and Max Welling. Bayesian dark knowledge. In *NIPS*, 2015.

[11] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report 1648, University of Toronto, 2009.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE, 86(11):2278-2324*, 1998.

[13] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[14] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 79–, New York, NY, USA, 2004. ACM.

[15] Matthias Rottmann, Karsten Kahl, and Hanno Gottschalk. Deep bayesian active semi-supervised learning. *CoRR*, abs/1803.01216, 2018.

[16] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

[17] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[18] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010.

[19] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[20] Edward H. Simpson. Measurement of diversity. *Nature*, 163, 1949.

[21] Gregory Vial. Comnist: Cyrillic-oriented mnist. In *Github*, 2017.

[22] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Trans. Cir. and Sys. for Video Technol.*, 27(12):2591–2600, December 2017.

[23] Patrick Weyers, Alexander Barth, and Anton Kummert. Driver state monitoring with hierarchical classification. In *21st Int. Conference on Intelligent Transportation Systems (ITSC)*, pages 3239–3244, 11 2018.

[24] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. Technical Report cs.LG/1708.07747, arXiv, 2017.