The power of ensembles for active learning in image classification

William H. Beluch
BCAI[*]

Tim Genewein
BCAI

Andreas Nürnberger
University Magdeburg

Jan M. Köhler
BCAI

Abstract

Deep learning methods have become the de-facto standard for challenging image processing tasks such as image classification. One major hurdle of deep learning approaches is that large sets of labeled data are necessary, which can be prohibitively costly to obtain, particularly in medical image diagnosis applications. Active learning techniques can alleviate this labeling effort. In this paper we investigate some recently proposed methods for active learning with high-dimensional data and convolutional neural network classifiers. We compare ensemble-based methods against Monte-Carlo Dropout and geometric approaches. We find that ensembles perform better and lead to more calibrated predictive uncertainties, which are the basis for many active learning algorithms. To investigate why Monte-Carlo Dropout uncertainties perform worse, we explore potential differences in isolation in a series of experiments. We show results for MNIST and CIFAR-10, on which we achieve a test set accuracy of 90% *with roughly 12,200 labeled images, and initial results on ImageNet. Additionally, we show results on a large, highly class-imbalanced diabetic retinopathy dataset. We observe that the ensemble-based active learning effectively counteracts this imbalance during acquisition.*

1. Introduction

Convolutional neural networks (CNNs) have become the state-of-the-art method for image classification tasks, achieving superior performance on well-known benchmark datasets like MNIST ([41]), CIFAR ([38]) or ImageNet ([51]) where CNNs were shown in [58], [17], and [55] to outperform human classification accuracy.

One shortcoming of CNNs is that they require large datasets for training which often come with a high labeling effort, resulting in a major hurdle in domains where labels can only be acquired by time-consuming and expensive manual labeling. Particularly for medical images labeling requires well-trained specialists, such that labeling resources and not computational power or model capacity become the main bottleneck for the rapid applicability of CNN models. Under a limited financial budget or if only a few experts are available, techniques to reduce labeling effort become important. The goal of such techniques is to find the minimally required set of labeled images in order to reach a certain classification accuracy. This problem can be formally addressed with the framework of active learning ([6]). Starting with an initial (small) data set to train a model, new data-points to be labeled (e.g. by a human expert) are selected with a so-called acquisition function. This function ranks unlabeled data by "how desirable" label information is expected to be for each data-point. Commonly used acquisition functions are based on criteria such as uncertainty of prediction ([16], [32], [62]), coverage of the data space ([52]), unanimity of a committee ([30]), expected error, or variance reduction ([25], [33]). Typically, a small number of highest-ranking unlabeled data-points are selected for labeling and subsequently added to the training set to train the model afresh. This procedure is repeated resulting in a step-wise increase of labeled data.

While active learning (AL) has a long history in machine learning, there is currently little literature on active learning for CNNs. Most traditional acquisition functions cannot be straightforwardly used since they do not scale to high-dimensional image data ([57]), or they rely on good uncertainty estimates for unlabeled data which are hard to obtain with standard CNNs. Methods which are applicable to CNNs have been introduced in ([16], [52], [59], [61]).

In this paper we use an ensemble of CNNs as a scalable approach for deriving well-behaved uncertainty estimates for unlabeled data (see also [49] and [40] for a comparable approach). These uncertainties are used to evaluate different acquisition functions proposed in the literature. We compare ensemble-based active learning with the two state-of-the art methods: a Bayesian deep learning approach (Monte Carlo Dropout [16]), and a density-based approach ([52]). Our experiments on MNIST, CIFAR-10 and a real-world large-scale medical image dataset for diabetic retinopathy (DR) detection ([11]) show that the en-

[*] Bosch Center for Artificial Intelligence. First name.last name@de.bosch.com

semble based approach consistently outperforms the other two approaches. On CIFAR-10 after acquiring $14,500$ images we report an accuracy of 91.5% compared to 88.4% and 88.2% for the methods by [16] and [52]. For the DR use case we achieve an AUC of 0.983 (a 1.8 units increase over the random baseline) for training on $21,000$ images. A recent paper on DR classification ([20]) achieves an AUC of 0.991 training on over $100,000$ images. The data-set is highly imbalanced (19.34% unhealthy examples) and the ensemble-based method selects the more informative, unhealthy examples for labeling with an increased probability.

In additional experiments we shed some light on why the ensemble-based uncertainties lead to better performance compared to MC Dropout uncertainties ([16]). We find that the ensemble uncertainties are better calibrated (see section 4), and subsequently investigate potential causes (increased model capacity, wider variety of randomness, increased diversity of ensemble members). Recent literature propose to split the uncertainty over network predictions into a data-dependent aleatoric and a parameter-dependent epistemic uncertainty ([36], [9]). Selecting data-points for labeling based only on epistemic model-uncertainty seems promising, however, we find that using epistemic uncertainty alone in our acquisition functions does not yield better results. One potential shortcoming of ensemble approaches is that training multiple networks can become computationally very costly. We investigate how a recently proposed method to overcome this problem, snapshot ensembles ([28]), performs in the active learning setting.

2. Related Work

Active learning The survey [53] gives an overview of the many AL strategies developed so far, though, it does not include any work on AL for deep neural networks. Active learning concepts relevant for this paper include uncertainty sampling ([42], [32], [35]), query-by-committee ([30], [5]) and density-based approaches ([52], [61], [46]).

Active learning for high-dimensional data Most of the methods proposed in the literature do not use deep learning models. [32] use an SVM classifier where uncertainties are calculated based on probabilistic outputs over the class label, with entropy and best-vs-second-best approaches as acquisition functions. [42] combine uncertainty and information density, the latter being extracted via a Gaussian process using precomputed SIFT features. [63] use a Gaussian random field model to combine active and semi-supervised learning. [26] use expected information gain as acquisition function and evaluate on various datasets using a SVM.

Ensembles for (deep) neural networks (to estimate uncertainty) Combining predictions of an ensemble of learners for improving task performance in neural networks goes back many years [21]. Today, ensembles are widely used in machine learning (see [10] for a review) and deep learning ([39], [23], [18]). Besides increasing task accuracy, [40], [49], and [48] use ensembles to estimate prediction uncertainty of deep neural networks in the context of outlier detection and reinforcement learning.

Uncertainty estimation in neural networks By taking the variance over softmax outputs of an ensemble of networks, and optionally using bagging to increase the network diversity, [48] show how to obtain uncertainty measures with neural networks. Another variant is shown in [40], who add an extra head to the network which is trained to account for a proportion of the variance ([47]). This allows one to use the combined variance over predictions as an uncertainty estimate, using a single network.

Bayesian neural networks model parameter-uncertainty explicitly by learning (Gaussian) posterior probabilities over weights, inducing uncertainty over network activations and thus predictions ([44], [45]). Recent approaches based on stochastic variational inference have been applied successfully to deep networks ([19]). [1] extend the approach of [19] by inducing a Gaussian mixture prior over the weights while minimizing an objective that also regularizes the network. A different approach uses additive Gaussian noise on the gradients during back-propagation for obtaining uncertainties over the posterior weight distribution ([37]). [24] introduce expectation back-propagation for approximate variational inference in deep neural networks.

Uncertainty used for AL [15] show how dropout can be used for obtaining posterior uncertainties over network predictions, and specifically for CNNs in [14]. The authors also use these uncertainties for AL in [16] on MNIST and a small medical data set. Uncertainty estimates are obtained by sampling from the average softmax output of multiple forward passes with random dropout masks—known as Monte Carlo (MC) Dropout. An interpretation is that each dropout mask produces one member of an ensemble of networks and averaging over multiple such forward passes is similar to having a full ensemble. In the experimental section, we compare MC Dropout uncertainties for active learning against uncertainties produced by a full ensemble.

[52] argue that uncertainty-based methods for AL are ineffective for CNNs and instead propose an acquisition function that uses geometric arguments in the data-space (essentially trying to cover a diverse set of points by considering distances between points and clusters of points). In our experiments we also compare against this method.

AL for CNNs Besides [16] there are three other AL approaches for CNNs to the best of our knowledge: [52] use a density approach to cover the entire space of unlabeled data points using a geometric based similarity function between images. [59] use the entropy of the softmax output for selecting images to label and additionally automatically label high confidence samples (pseudo-labeling). This approach is outperformed by [52]. Finally, [34] extend a method based on the expected model output change principle to deep neural networks. This method is computationally expensive and its performance is similar to [59].

3. Methodology

3.1. Pool-Based Active Learning

In pool-based AL, there exists a large unlabeled pool of data U, and an initial, small labeled set of data L. In each step of the process, a model M is trained on L, and an acquisition function $a(U, M)$ chooses s points to be labeled by an external oracle and added to L. This process is repeated, training M from scratch with the newly incorporated labeled data, until a certain budget of labeled data is exhausted or until a certain model performance is reached.

We explore different acquisition functions used for regression and image classification. For the former, $a(U, M)$ in this study is always based on the predictive variance of the model output. For the latter, either the softmax output vector of a neural network is used as an input to an uncertainty based acquisition function, or the outputs of the last fully-connected layer in the network are used as feature vectors to calculate image similarities for a density-based approach.

3.2. Uncertainty estimation and approaches

This study uses two recent methods to obtain well-behaved uncertainty estimates from deep neural networks: Monte-Carlo dropout (MC-dropout) ([15]) and deep ensembles ([40], [48]). The former approaches the problem from a Bayesian perspective, and interprets dropout regularization as a variational Bayesian approximation. In practice this entails training the neural network with the data D_{train} as usual with dropout, and during inference performing T forward passes through the network, each time sampling a new dropout mask, which results in the weights w_t. The T softmax vectors are then averaged to obtain the output for a given class c and input x.

$$p(y = c|x, D_{\text{train}}) = \frac{1}{T} \sum_{t=1}^{T} p(y = c|x, w_t) \tag{1}$$

The latter approach trains an ensemble of N classifiers, and uses the averaged softmax vectors of each ensemble member as the output (same as equation 1, replacing T with N). In the experiments all ensembles are trained with the same D_{train} and same network architecture, but different random weight initializations w_{init}. One could also take additional measures to de-correlate the ensembles, such as bootstrapping or using different network architectures ([48]).

3.3. Acquisition functions

Uncertainty based Three different uncertainty based acquisition functions, used in this study, can be applied to outputs obtained from either deep ensembles or MC-dropout. These functions and their approximations (when necessary) were introduced and used in [16] for active learning in certain settings. In the following equations, T always refers to either the number of forward passes in MC-dropout, or the number of classifiers in an ensemble.

The most ubiquitous measure in the literature is to choose points whose predicted classification probability distributions have the highest **entropy** ([54]).

$$\begin{aligned} H[y|x, D_{\text{train}}] := - \sum_c \Big(\frac{1}{T} \sum_t p(y = c|x, w_t) \Big) \\ \cdot \ \log \Big(\frac{1}{T} \sum_t p(y = c|x, w_t) \Big) \end{aligned} \tag{2}$$

Another measure (also known as (**BALD**, [27]) is the mutual information between data-points and weights. The idea is that data-points with a large mutual information between the (predicted) label and network weights would have a large impact on the trained network if the correct labels were provided. The measure consists of the entropy over predictions minus the conditional entropy over predictions given the weights, approximated for the CNN case ([16]).

$$\begin{aligned} I[y; w|x, D_{\text{train}}] &:= H[y|x, D_{\text{train}}] \\ &- \frac{1}{T} \sum_t \sum_c -p(y = c|x, w_t) \log p(y = c|x, w_t) \end{aligned} \tag{3}$$

The **variation ratio** is a measure of dispersion of a nominal variable, and is calculated as the proportion of predicted class labels that are not the modal class prediction (f_m is the number of predictions falling into the modal class category) ([13]). Larger values thus indicate a greater dispersion.

$$v := 1 - \frac{f_m}{T} \tag{4}$$

Although shown to be less effective than the previous acquisition functions in some settings [16], the **variance** of the softmax output vectors within the ensemble or within T forward passes can also be used as an acquisition function.

$$\sigma^2 = \frac{1}{C} \sum_c \frac{1}{T} \sum_t \left(p(y = c|x, w_t) - \hat{p}(y = c) \right)^2 \tag{5}$$

Geometric approaches Density-based acquisition functions have recently also been applied to CNNs. For comparison, this study implements the approaches presented in [52] (in the following called **Core-Set**) and [61] (in the following called **REPR**). The former method is a core-set approach that seeks to choose p points (the acquisition size) that minimize the maximum distance between point x_i in the overall distribution (with $[n]$ being all points there) and its closest neighbor x_j in the selected subset s. This study uses the simpler of the two implementations in [52], as it was shown to perform only marginally worse. In practice, the value u is selected greedily one point at a time, initialized with the training data images.

$$u = \underset{\mathrm{i}\in[\mathrm{n}]\setminus\mathrm{s}}{\operatorname{argmax}} \min_{j\in s} \operatorname{dist}(x_i, x_j) \tag{6}$$

The **REPR** method chooses points that best "represent" the rest of the distribution. The algorithm greedily tries to maximize the representativeness R of a set of selected points S_a. Each point in S_u has a representativeness score r, defined as the similarity between it and its most similar point in the selected subset; R is the sum of all individual r scores. REPR encourages diverse points to be added to S_a, as adding a point I_j to S_a similar to one already in S_a would result in high r scores for points in S_u that already had high r scores. For the similarity measure sim we use the Euclidean norm.

$$\begin{aligned} R(S_a, S_u) &= \sum_{I_j\in S_u} r(S_a, I_j) \\ r(S_a, I_j) &= \max_{I_i\in S_a} \operatorname{sim}(I_i, I_j) \end{aligned} \tag{7}$$

4. Experimental results

The models presented in the following are evaluated on MNIST ([41]), CIFAR-10 ([38]), and a diabetic retinopathy dataset ([11]). Referred figures and tables starting with an "A" are placed in the supplementary material. Shaded areas in the plots denote ± one standard deviation.

4.1. Implementation details

The settings for the experiments are described in Table 1. The network architecture for MNIST, referred to as "S-CNN" is the same as in the example Keras MNIST CNN implementation ([12]), also used in [16] (two convolutional layers and one dense layer). For CIFAR-10 we use the Keras CIFAR CNN implementation ([12]) with four convolutional layers and one dense layer, which we refer to as "K-CNN". Additionally we also evaluate with DenseNet-121 ($k = 12$, with bottleneck) on CIFAR-10, using the learning rate schedule as proposed in [29]. Details for the inceptionV3 architecture used for the diabetic retinopathy use case are described in section 4.5. For all models we use Glorot initialization, the Adam optimizer for S-CNN, RMSprop for K-CNN, standard SGD for DenseNet and a combiantion of RMSprop (20 epochs) and SGD (rest of the training) for InceptionV3. After each acquisition step, models are trained for a fixed number of epochs. Subsequently, a fixed number of samples are selected from a pool of fixed size (a random subsample of the remaining "unlabeled" training data), added to the labeled data set, and models are re-rained from scratch with this set. Initial labeled sets are randomly sampled from the whole training set. For MNIST and CIFAR-10 initial sets are balanced over all classes. All experiments are run for a fixed amount of acquisition steps, or equivalently, until a certain amount of training data is labeled. We report test errors on the final model, after training the full number of epochs, except for DenseNet where we use the model with the best validation loss among all epochs. Results are averaged over five repetitions. For MC dropout we use 25 forward passes. Each ensemble consists of five networks of identical architecture but different random initializations. The dropout rate is 0.25 for the two conv-layers and 0.5 for the dense layers for S-CNN and K-CNN, respectively. For DenseNet it is 0.2 after every convolutional layer. For the geometric approaches the Euclidean norm is used as a similarity-measure and is calculated using image features from the outputs of the last fully connected layer in the network.

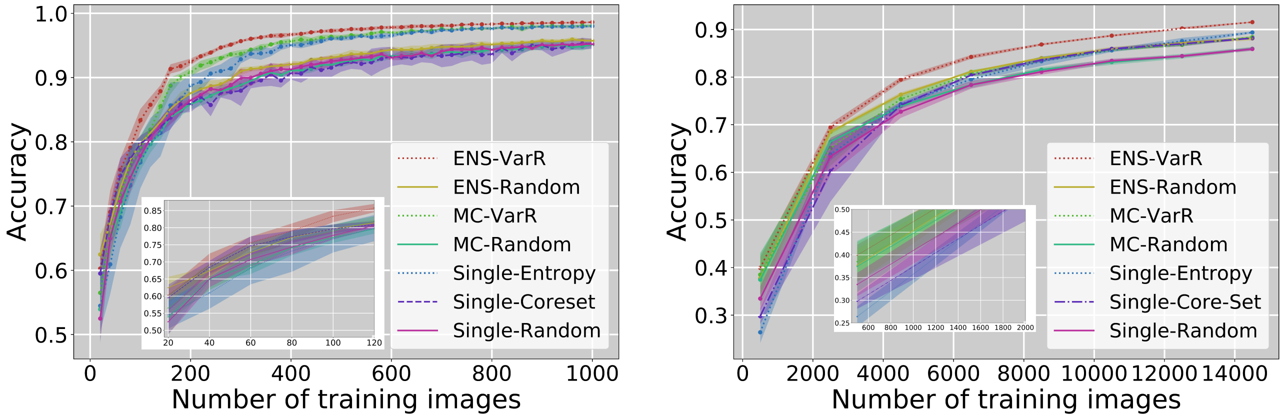
4.2. Results for active learning on image data

We evaluate the AL performance under different methods for uncertainty estimation: an ensemble of five networks, MC Dropout, and a single, standard network. Selected results are shown in figure 1. All results (including Variation Ratio, BALD, predictive variance, predictive entropy and additional the geometric approaches Core-Set and REPR) are shown in Figure A1. On MNIST and CIFAR the ensemble-based approach outperforms all others by a clear margin, whereas the MC-VarR performs similarly to the Single-Entropy approach. Note that for the initial labeled set data-points were selected randomly, thus, comparing acquisition based versions against their random counter-parts for each method should lead to the same accuracy on average. Corresponding results are shown in the inlets.

On average ENS-VarR increases the accuracy over its random baseline by a larger margin compared to the other approaches on more complex data. On MNIST not considering the initial training and the first four acquisition steps, as there is larger variance in the accuracy, ENS-VarR / MC-VarR / Single-Entr. achieve on average 3.96 / 3.98 / 2.87 percentage points (pp) higher accuracy than the random baseline. On CIFAR (omitting the initial training and first acquisition step) this is 3.19 / 2.23 / 2.30 pp.

Table 2 provides the mean and standard deviation over five runs of labeled images needed to achieve a top-1 accuracy of 80%, 85%, and 87.5% (DenseNet on CIFAR-10). The ENS-VarR needs on average $2,906$ (29.2%) less labeled images to reach an acccuracy of 85% compared to an entropy-based single network approach. Generally, variance for ENS-VarR is lowest and variance for random acquisitions is higher compared to AL acquisitions.

The geometric approach by [52] (Single-Core-Set) performs similar to random acquistion on MNIST (1a, significantly worse than random on CIFAR-10 using the K-CNN (A1f, and better than random on CIFAR-10 using



(a) MNIST on S-CNN

(b) CIFAR-10 on DenseNet

Figure 1: Test accuracy over acquired images. We compare Variation Ratio for MC dropout and the ensemble (ENS) and softmax-entropy based acquisition for a single network. For all methods we also show performance under random acquisition. Shaded areas denote ± one standard deviation. (see text for details about the architectures used).

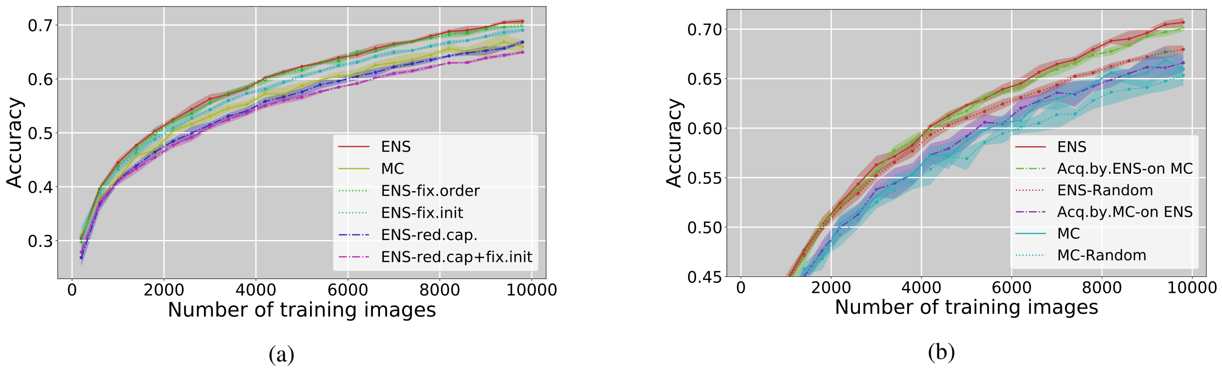
the DenseNet (1b. However the original paper uses a different network (VGG-16), and achieves a relatively worse accuracy than our setup: with $10,000$ training images [52] achieve a mean accuracy of 74%, whereas we report 85% on the DenseNet for $10,500$ images. The differences could be due to the different feature representations provided by the various network architectures, or perhaps the prevalence of outliers hindering the greedy core-set approach (i.e. perhaps there are more outliers with a negative effect using the K-CNN feature representation on CIFAR-10). Results on CIFAR-10 with K-CNN (instead of DenseNet) show little difference in the performance of the different methods (see Fig. A2), with the exception of the geometric methods. Variations of hyperparameters such as acquisition step size, subset pool size, dropout rates do not qualitatively affect the results (Fig. A3).

4.3. Comparing ensemble-based against Monte-Carlo Dropout performance

Multiple MC Dropout forward passes with different dropout masks can also be interpreted as an approximation to a full ensemble (consisting of separate networks). In our experiments we find that using uncertainties obtained from MC Dropout with 25 forward passes performs worse for AL compared to uncertainties obtained by an ensemble of five networks. To investigate the difference we performed a number of experiments. One difference is that the weights, initialization, and (to some degree) the gradient updates are shared among all MC Dropout “ensemble” members, which is not true for an ensemble of five separate networks. Another difference is that the effective model-capacity of MC Dropout might be reduced, since at every forward pass a certain proportion of neurons or convolutional filters are inactive. Our experiments show that it is a combination of several factors that lead to an increased AL performance using an ensemble. In particular, we investigate the following aspects in isolation: **Number of networks** in the ensemble or forward passes in MC dropout, **Model capacity** by reducing the number of neurons for the ensemble networks, **Fixed initialization** to have the same initialization for all ensemble members, and **Fixed order within a mini-batch** to have the same order of images within a mini-batch for all

	Model	**Training epochs**	**Data size** train / val / test	**Acquisition size**		
MNIST	S-CNN	50	58,000 / 2,000 / 10,000	20	+ 20 (2K)	→ 1,000
CIFAR-10	K-CNN	150	48,000 / 2,000 / 10,000	200	+ 400 (4K)	→ 9,800
CIFAR-10	DenseNet	100	48,000 / 2,000 / 10,000	500	+ 2000 (20K)	→ 14,500
Diabetic R.	InceptionV3	150	67,961 / 3,000 / 17,741	1,000	+ 5,000 (30K)	→ 21,000
ImageNet	ResNet-50	100	1,281,167 / 10,000 / 50,000	40,000	+ 40,000 (400K)	→ 280,000

Table 1: Settings used for the active learning experiments on the MNIST, CIFAR-10 and Diabetic Retinopathy data sets:
Training epochs: Maximum number of training epochs.
Data size: Size of data set for training / validation / test.
Acquisition size: number of images for the initial model + number of images acquired in each step (from the number of images in the pool subsample) → Maximum number of images acquired during training.



(a)

(b)

Figure 2: Manipulating differences between Monte-Carlo Dropout and ensemble-based active learning in isolation. Plots show test accuracy on CIFAR-10 with K-CNN, using Variation Ratio as an acquisition function. (a) ENS-fix.order: ensemble with fixed order of images per mini-batch. ENS-fix.init: ensemble with same initialization for all members. ENS-red.cap.: ensemble with reduced number of neurons (25% less for conv-layers, 50% less for dense layers). ENS-red.cap+fix.init: ensemble with reduced capacity and same initialization for all members. (b) Acq.by.ENS-on MC: Images acquired by ensemble-based approach (during the run of ENS) used to train MC dropout network. ENS-Random: ensembles with random acquisition. Acq.by.MC-on ENS: Images acquired by MC Dropout (during the run of MC) used to train ensemble. MC-Random: MC Dropout with random acquisition.

ensemble members.

Number of networks Fig. A3 shows the effect of using different numbers of MC forward passes (2, 5, 10, 25, 50, 100) and of ensemble members (3, 5, and 7). The performance of the ensemble-based approach is only slightly impacted by the number of members and even with three networks, the method still outperforms the MC Dropout based approach with a much larger number of forward-passes.

Fixed initialization and order Another potential cause for the lack of diversity ([4], [31]) in the MC Dropout case is the single initialization; the "networks" from random forward passes could be in similar local optima, whereas members of the ensemble may converge to different local optima due to different initializations. A second source of randomness that may have the same effect is the order of the training images presented in the mini-batches: for the ensemble this order is randomly shuffled for each member. As Figure 2a shows, fixing the order of training examples has a negligible effect, whereas sharing the same weight initialization across all members of the ensemble leads to a significant decrease in performance. However, the latter factor alone can not fully explain the difference in performance.

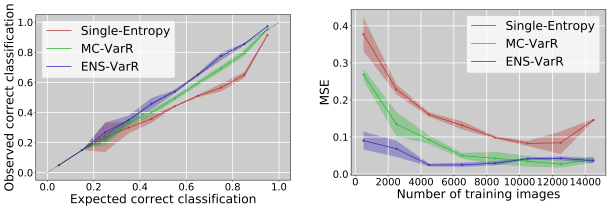
test	ENS-VarR	MC-VarR	Single-Entr.
accuracy	Active learning acquisition (Random)		
80% mean	4718 (6032)	6255 (7470)	6711 (7661)
std.	*206.2 (57.8)*	*276.0 (442.9)*	*216.2 (565.6)*
85% mean	7053 (9613)	9888 (13248)	9959 (13300)
std.	*205.3 (451.8)*	*186.6 (481.6)*	*301.6 (496.3)*
87.5% mean	9187 (12830)	13388 (-)	12453 (-)
std.	*184.8 (333.8)*	*280.9 (-)*	*582.2 (-)*

Table 2: Mean and standard deviation over five runs of acquired images to achieve a top-1 accuracy (DenseNet on CIFAR-10) for the ensemble, MC dropout and a single net with Variation Ratio and Entropy as acquisition functions compared to random acquisition. The ensemble approach needs less images to achieve a certain accuracy.

Capacity In each MC forward pass a significant fraction of neurons or convolutional filters is inactive, such that on average the total model capacity may be reduced compared to an ensemble of multiple networks of the same architecture. By using smaller networks for the ensemble (i.e. their layer-size matches the average number of active neurons in MC Dropout) we find that the reduced model capacity does indeed play an important role (see Figure 2). The performance of such a capacity-limited ensemble drops roughly to the performance of MC Dropout (note however that MC Dropout uses 25 forward passed compared to 5 ensemble members). In figure A4 we increase the capacity of the MC network so the same number of activations are present after dropout. Though, there is only a negligible benefit.

Separating acquisition quality from inherent accuracy In Figure 2b we investigate the effect of training the ensemble with the images that MC Dropout would have selected. To do so, we simply perform all acquisition steps with MC Dropout and record the images selected in each step, and subsequently train an ensemble with these acquisitions. We also perform the reciprocal experiment of training MC Dropout with images acquired using ensembles. The results show that an MC-Dropout network using the ENS selected images performs only marginally worse than the ENS acquisition function, and using ENS to evaluate on the MC-Dropout selected images performs only marginally better than the MC acquisition function. Essentially, this means that the "acquisition quality" of ENS is superior to MC Dropout, and that the difference can not simply be explained by the fact that evaluating with an ensemble is more accurate than evaluating with MC Dropout.

Uncertainty calibration To assess calibration ([7]) quality we determine whether the expected fraction of correct classifications (as predicted by the model confidence, i.e. the uncertainty over predictions) matches the observed fraction of correct classifications. When plotting both values against each other, a well-calibrated model lies close to the diagonal. The mean-squared-error (MSE) between the diagonal and the calibration plot is used to quantify calibration quality. Results are shown in Figure 3 after 3 acquisition steps and throughout the whole acquisition procedure in Figure 3a. Additionally we show calibration plots for different acquisition steps in Figure A6. We observe that ensemble-models are better calibrated in the low-data regime compared to MC Dropout and a single network. In the regime of sufficient data we find little difference between MC Dropout and ensembles.



(a) (b)

Figure 3: a) Calibration plot after 3 acquisition steps $(6,500$ images) for CIFAR-10 and the DenseNet. Ideal calibration is on the dashed diagonal. b) Mean squared error for the calibration lines for different number of acquired images.

As additional measures to assess uncertainty quality, we report the negative log likelihood (NLL) and the Brier score ([3]) (as used in [40]) in Table 3 for four acquisition steps, and in Figure A7 across the whole acquisition procedure. Similar to the calibration analysis, we find that under both measures ensembles have an increased uncertainty quality in the low data regime, and perform similarly to MC Dropout ensembles with sufficient labeled data. Both methods consistently outperform softmax-entropy based uncertainties of a single network.

Acqu. step	Brier Score Single / MC / ENS	NLL Single / MC / ENS
0	0.3773 / 0.2685 / 0.0903	0.3173 / 0.2261 / 0.0763
1	0.2287 / 0.1346 / 0.0682	0.1952 / 0.1170 / 0.0664
2	0.1604 / 0.0923 / 0.0242	0.1371 / 0.0836 / 0.0241
3	0.1315 / 0.0495 / 0.0248	0.1143 / 0.0442 / 0.0287

Table 3: NLL and Brier score, averaged over five runs, for different acquisition steps for a single network, MC and an ensemble (DenseNet on CIFAR-10).

Decomposing uncertainty [36] and [9] describe a decomposition of predictive uncertainty into an aleatoric (noise in the data) and an epistemic component (uncertainty in the model parameters). Importantly, epistemic uncertainty can be reduced with more data whereas the aleatoric uncertainty is theoretically unaffected by an increase of training data. Acquisition functions based on epistemic uncertainty thus hold the promise of improving acquisition quality. Unfortunately we could not find an improvement in our experiments by using the uncertainty decomposition in the MC Dropout setting, see Figure A8. To investigate potential reasons for this we conduct a one-dimensional regression toy-example as described in [47], see Figure A9.

Implicit ensembling The drawback of using ensembles is that it is computationally expensive to train multiple networks. Some techniques to overcome this issue have been proposed in recent literature. We evaluate some of these methods on our active learning experiments. The results and implementation details are described in the appendix (Fig. A5).

4.4. ImageNet

We further test the best performing method (ENS-VarR) on the large-scale image classification dataset ImageNet [8], using the ResNet-50 network architecture [22] (achieves top-1 accuracy of 75.3% with full dataset). The network is trained for 100 epochs without data augmentation using stochastic gradient descent. The initial learning rate of 0.1 is changed to 0.01 at epoch 50, and 0.001 at epoch 75. The initial 40,000 images are class-balanced. Active learning hyperparameters can be found in Table 2.

The results are displayed in Fig. 4 and Table 4. While there initially is no difference between the random baseline and the uncertainty-based acquisition function, after the third acquisition (160k training images) a small improvement can be seen, which continues to widen over the next few iterations. As training an ensemble of networks on ImageNet is computationally costly, a large acquisition size of 40,000 images is used. It is likely that using a smaller acquisition size will result in a faster improvement over the random baseline.

	40k	80k	120k	160k	200k	240k	280k
Random	0.159 (0.004)	0.257 (0.003)	0.321 (0.006)	0.372 (0.003)	0.407 (0.007)	0.439 (0.001)	0.470
VarR	0.152 (0.003)	0.257 (0.004)	0.324 (0.002)	0.383 (0.002)	0.427 (0.004)	0.458 (0.004)	0.494

Table 4: Top-1 accuracies using ENS-based acquisition functions for active learning on ImageNet (the values plotted in Fig. 4). Numbers in parentheses are standard deviation over three repetitions (except for final point, which is a single run).

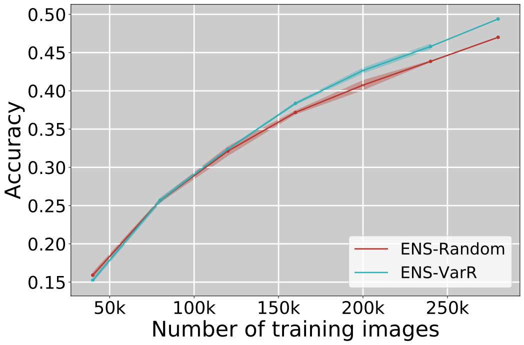


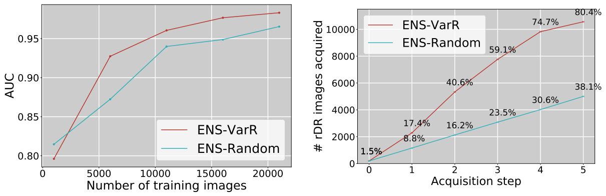
Figure 4: Test top-1 accuracy over acquired images. Shaded areas denote ± one standard deviation.

4.5. AL for diagnosis of diabetic retinopathy

AL is particularly relevant for fields in which labeling of images is expensive or which require highly trained experts, such as in medical image diagnosis. We evaluate the AL approaches presented in this paper on a real-world use case to detect diabetic retinopathy in eye fundus images . Details about the data and the AL parameters are described at the beginning of Section 4. The task is to classify between ("unhealthy") referable (rDR) and ("healthy") non-referable DR images (examples are shown in figure A10). According to the WHO, in 2002 almost 5 million people were suffering from blindness caused by DR, which makes it the fifth most common cause of moderate to severe visual impairment [2]. Once detected, DR can be treated quite well. However, particularly in developing countries where access to ophthalmologists and medical staff is scarce ([50]), cheap solutions for automated mass-screenings that can be operated by laymen is highly desirable. Automatic diagnosis algorithms using CNN-based grading and detection algorithms have been developed ([43]), however these require large amounts of training data which is currently only available for high-quality image acquisition setups that need to be operated by trained professionals. Recently, one such data-set was released as part of a machine learning competition ([11]). The original data set contains five classes for varying stages of retinopathy. We merge the first two classes as healthy and the rest as rDR to yield a binary classification problem. 19.34% of the images fall in the latter category rDR.

We use the inceptionV3 network ([56]) from Keras Applications ([12]), which achieved high task accuracy in [20]. We augment / preprocess the data by flipping images horizontally and vertically, and by channel-wise color augmentation ([39]). Pre-trained weights from ImageNet are used for initialization (excluding the final fully connected layer, which is initialized randomly). For 20 epochs only the final fully connected layer is trained with RMSprop. Then, the whole network is trained with SGD (learning rate of 0.0001, momentum of 0.9, no weight decay).

Figure 5a shows the area-under-curve (AUC) depending on the number of acquired images. After selecting 21,000 images, the ENS-VarR approach achieves an AUC of 0.983, and the random acquisition an AUC of 0.965. [20] use 80% of the 128,175 images for training an ensemble of 10 networks with the same inceptionV3 architecture, and achieve a final AUC of 0.991. Although we cannot directly compare [20] to our work as different amounts of data were used, the experiment nevertheless underlines that the ENS approach can be usefully applied to a real-life medical use case. Even though the data-set is highly imbalanced (about one fifth rDR images), the ensemble approach selects rDR images with a significantly increased probability per acquisition step (Figure 5b). We assume that these images are particularly informative for improving task performance. Interestingly, after four acquisition steps, the fifth acquisition step for ENS-VarR would select only 5.7% additional rDR images (compared to 7.5% for the random acquisition). We believe this is because after four acquistion steps most rDR images have been selected, so healthy images might be more beneficial for improving the AUC.



(a)

(b)

Figure 5: Test results for the diabetic retinopathy dataset. (a) AUC over acquired images. (b) Number of unhealthy images acquired and the percentage of total rDR images in the training set for each acquisition step.

5. Conclusion

We compare the performance of acquisition functions and uncertainty estimation methods for active learning with CNNs on image classification tasks. We show that ensemble-based uncertainties consistently outperform other methods of uncertainty estimation (in particular MC Dropout) and lead to state-of-the-art active learning performance on MNIST and CIFAR-10. Through additional experiments we find that the difference in active learning performance can be explained by a combination of decreased model capacity and lower diversity of MC Dropout ensembles. Additional evaluations indicate that recently proposed methods for implicit ensembling, but also methods that separate aleatoric and epistemic uncertainty, fail to outperform plain-ensemble active learning. We conclude by showing results on a real-world application in medical diagnosis, and a large-scale application in ImageNet classification. We find that ensemble-based active learning works well in both scenarios.

References

[1] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning (ICML)*, pages 1613–1622, 2015. 2

[2] R. R. Bourne, G. A. Stevens, R. A. White, J. L. Smith, S. R. Flaxman, H. Price, J. B. Jonas, J. Keeffe, J. Leasher, K. Naidoo, et al. Causes of vision loss worldwide, 1990–2010: a systematic analysis. *The Lancet Global Health*, 1(6):e339–e349, 2013. 8

[3] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. 7

[4] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. 6

[5] R. Burbidge, J. J. Rowland, and R. D. King. Active learning for regression based on query by committee. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 209–218. Springer, 2007. 2

[6] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *J. Artif. Intell. Res. (JAIR)*, 4:129–145, 1996. 1

[7] A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. 7, 4

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR09*, 2009. 7

[9] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Uncertainty decomposition in bayesian neural networks with latent variables. *International Conference on Machine Learning (ICML)*, 2017. 2, 7, 6

[10] T. G. Dietterich et al. Ensemble methods in machine learning. *Multiple classifier systems*, 1857:1–15, 2000. 2

[11] EyePacs. https://www.kaggle.com/c/diabetic-retinopathy-detection. *assessed on 2017-01-16*, 2015. 1, 4, 8, 7

[12] fchollet. Keras. *https://github.com/fchollet/keras*, 2015. 4, 8

[13] S. Freeman. *Elementary Applied Statistics : For Students in Behavioral Science*. John Wiley and Sons, New York, 1965. 3

[14] Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ICLR workshop*, 2016. 2

[15] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning (ICML)*, 2016. 2, 3

[16] Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. *International Conference on Machine Learning (ICML)*, 2017. 1, 2, 3, 4

[17] X. Gastaldi. Shake-shake regularization. *ICLR workshop*, 2017. 1

[18] B. Graham. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2015. 2

[19] A. Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2348–2356, 2011. 2

[20] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. Q. Nelson, J. Mega, and D. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016. 2, 8

[21] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(10):993–1001, 1990. 2

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 7

[23] 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 (CVPR)*, pages 770–778, 2016. 2

[24] J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning (ICML)*, pages 1861–1869, 2015. 2

[25] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *International Conference on Machine Learning (ICML)*, pages 417–424, 2006. 1

[26] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *CVPR workshop*, 2008. 2

[27] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. 3

[28] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get m for free. *International Conference on Learning Representations (ICLR)*, 2017. 2, 4

[29] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[30] J. E. Iglesias, E. Konukoglu, A. Montillo, Z. Tu, and A. Criminisi. Combining generative and discriminative models for semantic segmentation of ct scans via active learning. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 25–36. Springer, 2011. 1, 2

[31] U. Johansson, T. Lofstrom, and L. Niklasson. The importance of diversity in neural network ensembles-an empirical investigation. In *International Joint Conference on Neural Networks (IJCNN)*, pages 661–666, 2007. 6

[32] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos. Multi-class active learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2372–2379. IEEE, 2009. 1, 2

[33] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos. Scalable active learning for multiclass image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(11):2259–2273, 2012. 1

[34] C. Käding, E. Rodner, A. Freytag, and J. Denzler. Active and continuous exploration with deep neural networks and expected model output changes. In *NIPS workshop*, 2016. 2

[35] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. 2

[36] A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2, 7, 5

[37] D. P. Kingma and M. Welling. Stochastic gradient VB and the variational auto-encoder. In *International Conference on Learning Representations (ICLR)*, 2014. 2

[38] A. Krizhevsky and G. E. Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009. 1, 4

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 2, 8

[40] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016. 1, 2, 3, 7, 5

[41] Y. LeCun. The MNIST database of handwritten digits. *http://yann. lecun.com/exdb/mnist/*, 1998. 1, 4

[42] X. Li and Y. Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 859–866, 2013. 2

[43] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Snchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(Supplement C):60 – 88, 2017. 8

[44] D. J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 2

[45] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. 2

[46] H. T. Nguyen and A. Smeulders. Active learning using preclustering. In *International Conference on Machine Learning (ICML)*, page 79, 2004. 2

[47] D. A. Nix and A. S. Weigend. Estimating the mean and variance of the probability distribution. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 55–59, 1994. 2, 7, 4, 5, 6

[48] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4026–4034, 2016. 2, 3, 4

[49] N. Pawlowski, M. Jaques, and B. Glocker. Efficient variational bayesian neural network ensembles for outlier detection. *ICLR workshop*, 2017. 1, 2

[50] S. Resnikoff, W. Felch, T.-M. Gauthier, and B. Spivey. The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners. *British Journal of Ophthalmology*, 96(6):783–787, 2012. 8

[51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[52] O. Sener and S. Savarese. A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv:1708.00489*, 2017. 1, 2, 3, 4, 5

[53] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010. 2

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

[55] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 1

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 8

[57] S. Tong. *Active learning: theory and applications*. Stanford University, 2001. 1

[58] L. Wan, M. Zeiler, S. Zhang, Y. LeCun, and R. Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning (ICML)*, pages 1058–1066, 2013. 1

[59] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 1, 2, 3

[60] H. Yang, C. Yuan, J. Xing, and W. Hu. Diversity encouraging ensemble of convolutional networks for high performance action recognition. In *IEEE International Conference on Image Processing (ICIP)*, 2017. 4

[61] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. *arXiv preprint arXiv:1706.04737*, 2017. 1, 2, 3

[62] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision (IJCV)*, 113(2):113–127, 2015. 1

[63] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *ICML workshop*, 2003. 2