Leveraging uncertainty information from deep neural networks for disease detection

Christian Leibig[1,*], Vaneeda Allken[1], Philipp Berens[1,2+], and Siegfried Wahl[1,3+]

[1]Institute for Ophthalmic Research, Eberhard Karls University, Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience and Centre for Integrative Neuroscience, Eberhard Karls University, Tübingen, Germany
[3]Carl Zeiss Vision International GmbH, Germany
[*]christian.leibig@uni-tuebingen.de
[+]Co-senior author

ABSTRACT

In recent years, deep neural networks (DNNs) have revolutionized the field of computer vision and image processing. In medical imaging, algorithmic solutions based on DNNs have been shown to achieve high performance on tasks that previously required medical experts. So far DNN-based solutions for disease detection have been proposed without quantifying their uncertainty in a decision. In contrast, a physician knows whether she is uncertain about a case and will consult more experienced colleagues if needed. Here we propose to estimate the uncertainty of DNNs in medical diagnosis based on a recent theoretical insight on the link between dropout networks and approximate Bayesian inference. Using the example of detecting diabetic retinopathy (DR) from fundus photographs, we show that uncertainty informed decision referral improves diagnostic performance. Experiments across different networks, tasks and datasets showed robust generalization. Depending on network capacity and task/dataset difficulty, we surpass 85% sensitivity and 80% specificity as recommended by the NHS when referring $0\% - 20\%$ of the most uncertain decisions for further inspection. We analyse causes of uncertainty by relating intuitions from 2D visualizations to the high-dimensional image space, showing that it is in particular the difficult decisions that the networks consider uncertain.

Introduction

In recent years, deep neural networks (DNNs)[1] have revolutionized computer vision[2] and gained considerable traction in challenging scientific data analysis problems[3]. By stacking layers of linear convolutions with appropriate non-linearities[4], abstract concepts can be learnt from high-dimensional input alleviating the challenging and time-consuming task of hand-crafting algorithms. Such DNNs are quickly entering the field of medical imaging and diagnosis[5–13], outperforming state-of-the-art methods at disease detection or allowing one to tackle problems that had previously been out of reach. Applied at scale, such systems could considerably alleviate the workload of physicians by detecting patients at risk from a prescreening examination.

Surprisingly, however, DNN-based solutions for medical applications have so far been suggested without any risk-management. Yet, information about the reliability of automated decisions is a key requirement for them to be integrated into diagnostic systems in the healthcare sector[14]. No matter whether data is short or abundant, difficult diagnostic cases are unavoidable. Therefore, DNNs should report - in addition to the decision - an associated estimate of uncertainty[15], in particular since some images may be more difficult to analyse and classify than others, both for the clinician and the model, and the transition from ”healthy” to ”diseased” is not always clear-cut.

Automated systems are typically evaluated by their diagnostic sensitivity, specificity or area under receiver-operating-characteristic (ROC) curve, metrics which measure the overall performance on the test set. However, as a prediction outcome can decide whether a person should be sent for treatment, it is critical to know how confident a model is about each prediction. If we were to know which patients are difficult to diagnose, humans and machines could attend especially to these, potentially increasing the overall performance. In fact, if the machine was making most mistakes when uncertain about a case, one could devise a strategy mimicking typical medical decision making. When faced with a difficult case and feeling uncertain about a decision a junior doctor will consult a more experienced colleague. Likewise, a diagnostic algorithm could flag uncertain cases as requiring particular attention by medical experts.

Bayesian approaches to uncertainty estimation have indeed been proposed to assess the reliability of clinical predictions[16–19] but have not been applied to the large-scale real-world problems that DNNs can target. Outside the medical domain, the

integration of the Bayesian idea and DNNs is an active topic of research[20–30], where the practical value of these ideas has yet to be demonstrated.

Due to its ease of use and inherent scalability, a recent insight from *Gal & Ghahramani*[28,29,31] is particularly promising. Using dropout networks[32,33], where parts of the units are inactivated during training to avoid overfitting, one can compute an approximation to the posterior distribution by sampling multiple test predictions with dropout turned on. This allows one to perform approximate but efficient Bayesian inference by using existing software implementations. Another advantage of this approach is that it can be applied to already trained networks.

Here we assess whether this allows us to retrieve informative uncertainty estimates for a large-scale, real world disease detection problem. Diabetic retinopathy (DR) is one of the leading causes of blindness in the working-age population of the developed world[34]. If the symptoms are detected in time, progress to vision impairment can be averted but the existing infrastructure is insufficient and manual detection is time-consuming. With the increase in the global incidence of diabetes[35], clinicians now recognize the need for a cost-effective, accurate and easily performed automated detection of DR to aid the screening process[34,36,37]. Previous recommendations of the British Diabetic Association (now Diabetes UK) are often cited as 80% sensitivity and 95% specificity [37–39, and references therein] but the current minimum thresholds set by the NHS Diabetes Eye Screening programme are 85% sensitivity and 80% specificity for sight-threatening diabetic retinopathy[14].

Using a Bayesian DNN, we achieve state-of-the-art results for diabetic retinopathy detection. The computed measure of uncertainty allowed us to refer a subset of difficult cases for further inspection, resulting in substantial improvements in detection performance in the remaining data. This finding generalized across different model architectures, detection tasks and datasets. In practice, patients whose samples result in uncertain decisions would either be sent for further screening tests or referred directly to a specialist. We further explore the causes of uncertainty in our scenario. Intuitions illustrated on a 2D toy problem are used to understand how uncertainty might behave in the high-dimensional image space. This allowed us to predict the kind of application relevant scenarios for which the assessed uncertainty is informative.

Results

Here we tackle two major questions: first, we evaluate whether model uncertainty obtained from dropout networks at test time is useful for ranking test data by their prediction performance without knowing the latter. In the second part, we open the black box in order to learn what makes some predictions uncertain.

Predicting diabetic retinopathy with a measure of (un)certainty

Diabetic retinopathy datasets

We used a DNN-based approach to detect diabetic retinopathy (DR) from fundus images. Our main dataset is taken from a previous Kaggle competition[40]. This dataset consists of 35,126 training images and 53,576 test images, graded into five stages of DR by clinicians according to the following scale[41]: 0 - No DR, 1 - Mild, 2 - Moderate, 3 - Severe and 4 - Proliferative DR. The percentage of images labelled with No DR is about 73% in both the training and test dataset.

In order to measure the true generalization of our insights we in addition applied all networks to the publicly available Messidor dataset[42]. This dataset comprises 1,200 images divided into the following categories: 0 - No DR, 1 - Mild non-proliferative, 2 - Severe non-proliferative, 3 - Most serious DR.

Disease detection tasks

Because the question of whether a patient has to be sent to a physician at all, is of high priority, we reduced the problem to a binary classification task. Therefore we split the data into a ”healthy” versus ”diseased” set by grouping some of the classes. In order to analyse how model uncertainty behaves for different tasks, we varied the disease onset level. If set to 1, the classes except for 0 are in the ”diseased” category resulting in a detector for mild DR (or more severe) whereas for disease onset level 2, classes $\{0, 1\}$ are considered ”healthy” and moderate DR (or more severe levels) are in the ”diseased” group.

Network architectures

We used two different network architectures for our experiments: (1) The publicly available network architecture and weights[43] provided by the participant who ranked 5th out of 661 teams in the Kaggle DR competition[40], which we will call *JFnet*. (2) A network trained de novo for the question at hand.

The JFnet comprises 13 convolutional layers, 3 fully connected layers and a concatenation layer combining information from the contralateral eyes of a patient. Convolutional layers are interleaved with 5 max pooling layers, fully connected layers are interleaved with two feature pooling and dropout ($p_{drop} = 0.5$) layers each. All non-linearities are ReLUs[44] or Leaky ReLUs[45] (leakiness 0.5) except for the softmax output layer[46]. We recast the original model's five output units (trained for Kaggle DR's level discrimination task) to our binary tasks by summing the output of respective units.

Our own network architecture was inspired by the monocular part of the JFnet (which in turn is VGG-like[47]), with the fully connected part replaced by the concatenation of a global mean and a global max pooling layer, followed by a softmax

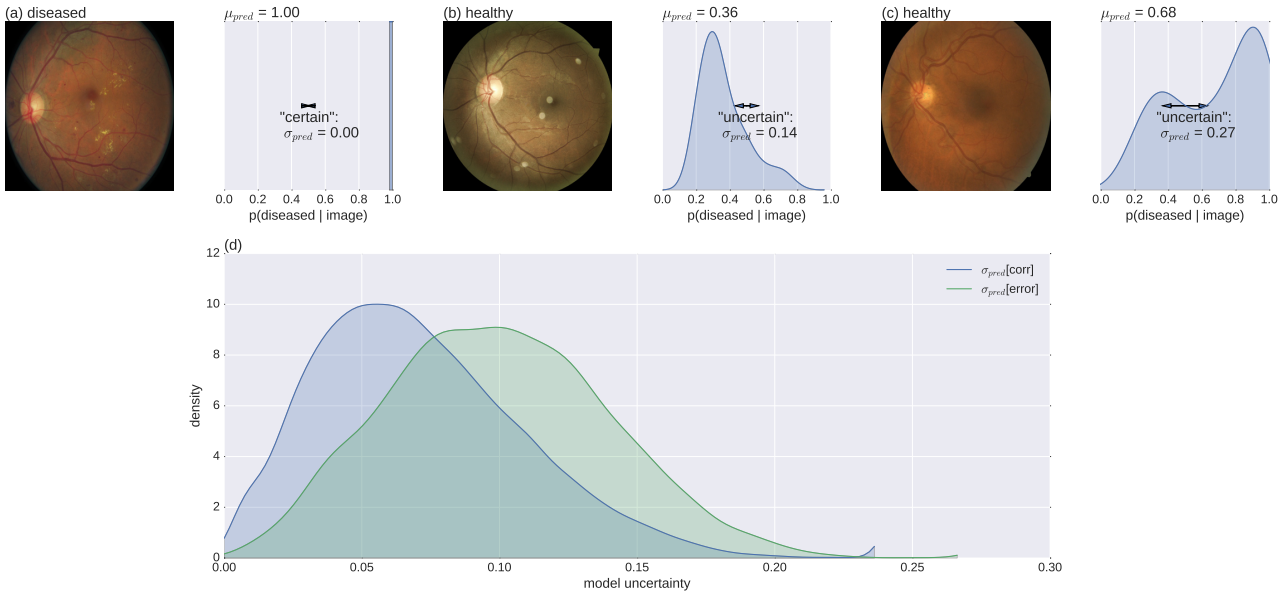

Figure 1. Bayesian model uncertainty for diabetic retinopathy detection. (a)-(c), left: Exemplary fundus images with human label assignments in the titles. **(a)-(c), right:** Corresponding approximate predictive posteriors (Eq. 6) over the softmax output values p(diseased|image) (Eq. 1). Predictions are based on μ_{pred} (Eq. 7) and uncertainty is quantified by σ_{pred} (Eq. 8). Examples are ordered by increasing uncertainty from left to right. **(d)** Distribution of uncertainty values for all Kaggle DR test images, grouped by correct and erroneous predictions. Label assignment for "diseased" was based on thresholding μ_{pred} at 0.5. Given a prediction is incorrect, there is a strong likelihood that the prediction uncertainty is also high.

output layer. Because the JFnet relies on both images of a given patient being present and has dropout only towards the end of the network, we also trained two networks (one for each disease detection task) that do not rely on eye blending. In order to increase the amount of network parameters that are treated in a Bayesian manner[29], we added dropout ($p_{drop} = 0.2$) after each convolutional layer and denote these networks as *Bayesian* convolutional neural networks (*BCNN*s).

Bayesian model uncertainty

We measured the uncertainty associated with the predictions of the DNNs described above, exploiting a relationship between dropout networks and a Bayesian posterior[28]. Typically, the softmax output of a classification network denotes a single prediction given a sample. In case of DR detection from a fundus image (see fig. 1 (a) left, for a "diseased" example) a trained network would output the probability that the given image is "diseased" (fig. 1 (a), right). The softmax probability is based on a single set of network parameters, whereas in a Bayesian setting one aims for the predictive posterior (compare eq. 2), i.e. a distribution over predictions (in our case the softmax values) obtained by integrating over the distribution over possible parameters.

The predictive posterior of a neural network is hard to obtain. However, Gal and colleagues[28] showed that by leaving dropout turned on at test time, we can draw Monte Carlo samples from the approximate predictive posterior (for details see Methods). We will summarize each predictive posterior distribution by its first two moments. The predictive mean μ_{pred} (Eq. 7) will be used for predictions and the predictive standard deviation σ_{pred} (Eq. 8) as the associated uncertainty.

Based on a fundus image, a DNN can be certain (1(a)) or more or less uncertain (1(b)-(c)) about its decision, as indicated by the width of the predictive posterior distribution: For example, an image can be classified as certainly diseased, where all sampled predictions are 1.0, such that $\sigma_{pred} = 0$ (Fig. 1 (a)). A different example is classified as "healthy", but the network predictions are more spread out ($\sigma_{pred} = 0.14$) (Fig. 1 (b)). The predicted label is still correct, because $\mu_{pred} = 0.36 < 0.5$. Finally, some examples produce high uncertainty in the DNN ($\sigma_{pred} = 0.27$) and result in an erroneous "diseased" prediction ($\mu_{pred} = 0.68 > 0.5$) (Fig. 1 (c)).

If high model uncertainty was indicative of erroneous predictions, this information could be leveraged to increase the performance of the automated system by selecting appropriate subsets for referral for further inspection. Indeed, model uncertainty was higher for incorrect predictions (Fig. 1 (d)). This means that σ_{pred} (a quantity that can be evaluated at test time) can be used to rank order prediction performance (a quantity unknown at test time), in order to mimic the human clinical work flow. In face of ambiguous decisions, further information should be obtained.

Uncertainty rank orders prediction performance

Performance improvement via uncertainty-informed decision referral

We analysed the feasibility of this idea by performing predictions (using the BCNN trained for disease onset 1 on the Kaggle DR training images) for all Kaggle DR test images and sorted the predictions by their associated uncertainty. We then referred predictions based on various levels of tolerated uncertainty for further diagnosis and measured the accuracy of the predictions (thresholded at 0.5) for the remaining cases (Fig. 2 (a)).

We observed a monotonic increase in prediction accuracy for decreasing levels of tolerated model uncertainty, which translates to the same behaviour when monitoring the fraction of retained data instead (Fig. 2 (b), blue curve). As a control experiment, we compared with randomly selected predictions, that is without using uncertainty information (Fig. 2 (b), green curve). For less than 2% decisions referred for further inspections, the 95% confidence intervals of the two scenarios are already non-overlapping. Uncertainty is hence informative about prediction performance, here measured by accuracy.

Performance improvement for different costs, networks, tasks and datasets

Here we build on the idea of uncertainty informed decision referral introduced above (Fig. 2) and assess whether performance improvements are robust across different settings. So far (Fig. 1, 2), predictions had been converted to labels by thresholding the predictive mean at 0.5. In a medical setting however, different costs are associated with false positive and false negative errors. These can be controlled by the decision threshold at which the diseased probability given an image is converted to the category "diseased". A complete picture can be obtained by the decision system's receiver-operating-characteristic, which monitors *sensitivity* over *1 - specificity* pairs for all conceivable decision thresholds. The quality of such a ROC curve can be summarized by its area under the curve (AUC), which varies between 0.5 (chance level) and 1.0 (best possible value).

ROC AUC improves monotonically with decreasing levels of uncertainty (Fig. 3 (a, left)). In addition, ROC curves for all Kaggle test images as well as under 10, 20 and 30% decision referral reveal that both sensitivity and specificity consistently improved (Fig. 3 (a, right)). These results were found to be robust for a variety of settings, that is for different networks (Bayesian CNN (Fig. 3, 1st row) vs. JFnet (Fig. 3, 2nd row), different tasks (disease onset 1 (Fig. 3, left double column) vs. disease onset 2 (Fig. 3, right double column)) and different datasets (BCNN on Kaggle (Fig. 3, 1st row) vs. BCNN on Messidor

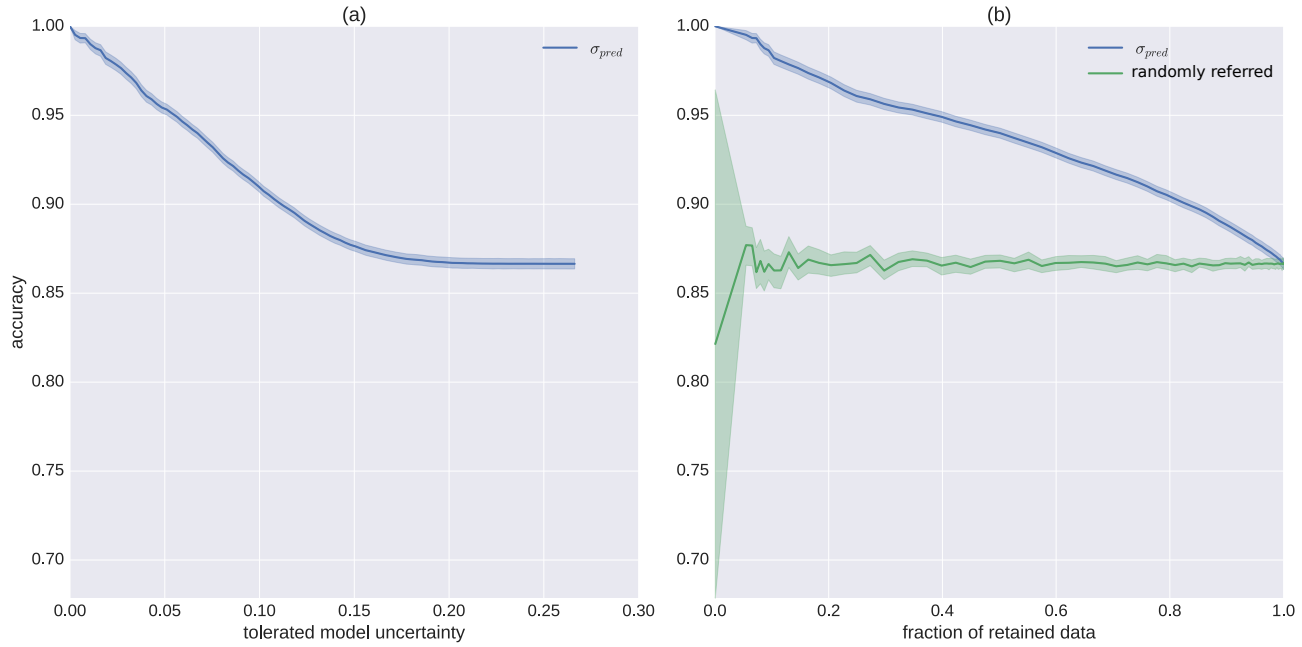

Figure 2. Improvement in accuracy via uncertainty-informed decision referral. **(a)** The prediction accuracy as a function of the tolerated amount of model uncertainty. **(b)** Accuracy is plotted as a fraction of retained data. The green curve shows the effect of rejecting the same number of samples randomly, that is without taking into account information about uncertainty.

(Fig. 3, 3rd row)).

We trained the Bayesian CNNs exclusively on Kaggle DR training images. As we initialized the weights[43] with those of the JFnet, in principle information from the test set could make our generalization estimate inaccurate, because the JFnet had been repeatedly submitted to Kaggle and Kaggle DR test data had been used for pseudo labelling. To provide a report of the true generalization performance, we used the Messidor database, which had never been used for either of our networks (Fig. 3, (e,f)). For a summary of the different configurations and comparison with prior state-of-the-art we refer to table 1.

Even though our primary aim was not to achieve high performance, we surpassed the requirements of both the NHS and the British Diabetic Association (Fig. 3) for (automated) patient referral for several settings and perform on par with the non-ensembling approach of *Antal & Hajdu*[38]. We also performed similar ensembling[38], by selecting an optimal (forward-backward search while monitoring AUC) ensemble of 100 networks from a much larger pool of dropout networks by controlling the random seeds. Performance improvements however were marginal and did not generalize to test data (data not shown), probably because this compromises the stochastic nature of the regularizing effects of dropout.

The JFnet outperformed the Bayesian CNN across the different configurations, probably due to the missing eye blending in the latter case. In addition, the better performance for moderate DR detection (onset 2) as compared to mild DR detection (onset 1) across networks and datasets is in line with the more pronounced expression of symptoms as the disease progresses. Comparison across datasets reveals that for both tasks, the models performed better on Messidor than on Kaggle data (compare Fig. 3 (a) vs. (e) and (b) vs. (f)). Specifically, we achieved both the BDA and NHS requirements on Messidor without having to refer decisions whereas for Kaggle data we have to refer $0-30\%$ of the data, depending on the recommendation, task difficulty and network capacity. It has been reported previously that about 10%[48] of the Kaggle DR images were considered ungradable according to national UK standards. We want to emphasize that the proposed uncertainty informed decision referral did not rely on labels for such cases, that is we could detect problematic images in an unsupervised way.

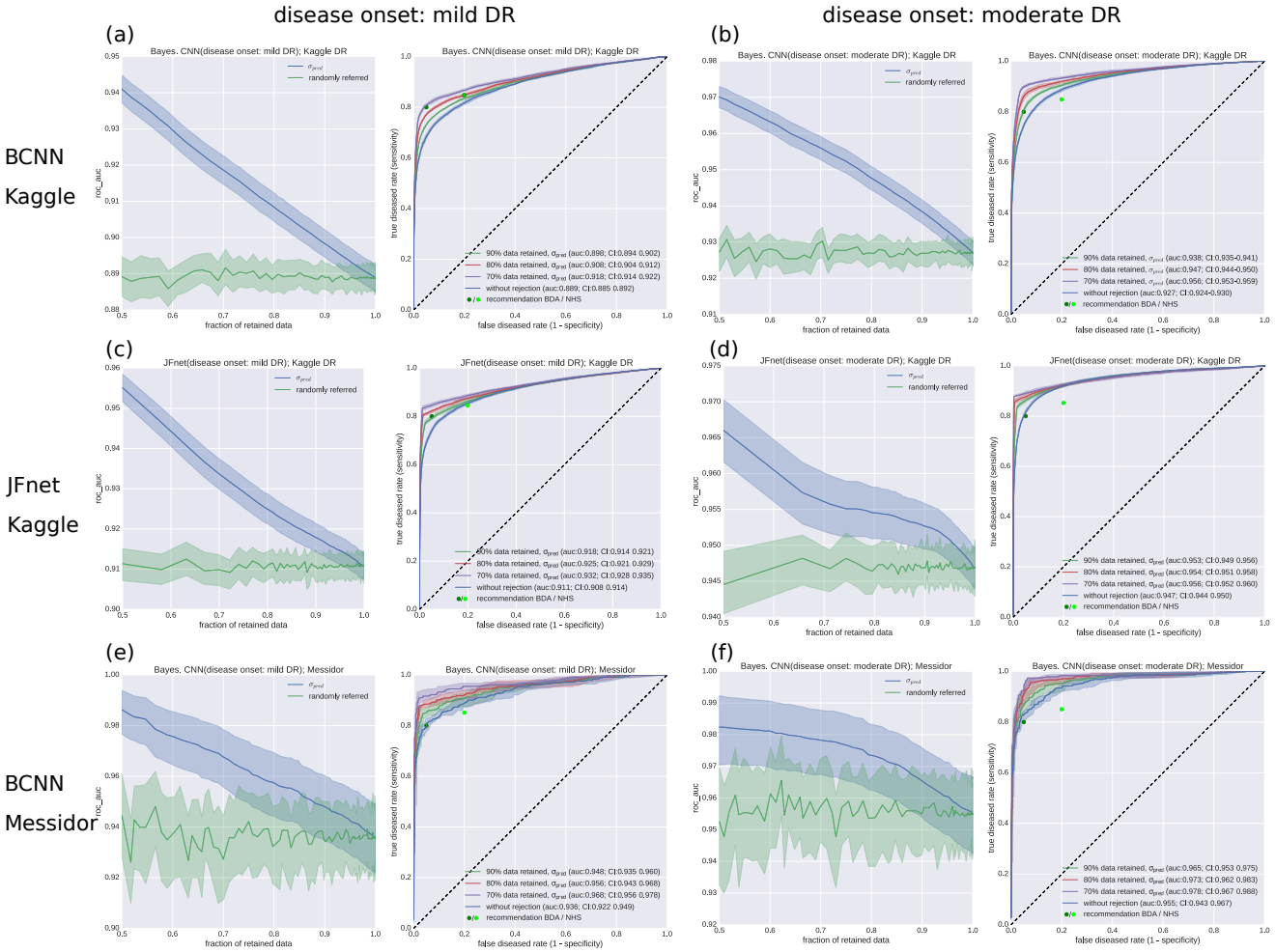

Figure 3. Improvement in receiver-operating-characteristics via uncertainty-informed decision referral for different networks (1st vs. 2nd row), tasks (left vs. right double column), and datasets (1st vs. 3rd row). Different costs associated with false positive and false negative errors are captured by the ROC characeteristics. (a, left) ROC AUC over the fraction of retained data under uncertainty informed (blue) and random (green) decision referral for a Bayesian CNN, trained for disease onset 1 and tested on Kaggle DR. **(a, right)** Exemplary ROC curves for different fractions of retained data. Panels (b)-(f) have the same layout. National UK standards for the detection of sight-threatening diabetic retinopathy (in[49] defined as moderate DR) from the BDA (80%/95% sensitivity/specificity, dark green dot) and the NHS (85%/80% sensitivity/specificity, bright green dot) are given in all subpanels with exemplary ROC curves. **(b)** same as (a), but for disease onset 2. **(c)** For the original JFnet, recast for disease onset 1, tested on Kaggle DR. **(d)** Same as (c), but for disease onset 2. **(e)** Same network as in (a), but tested on Messidor. **(f)** Same network as in (b), but tested on Messidor.

Dataset	Architecture	Task	100% data AUC	90% data AUC	80% data AUC	70% data AUC
Kaggle DR	Bayes. CNN	(0) vs (1,2,3,4)	0.889 CI: [0.885-0.892]	0.898 CI: [0.894-0.902]	0.908 CI: [0.904-0.912]	0.918 CI: [0.914-0.922]
Kaggle DR	Bayes. CNN	(0, 1) vs (2,3,4)	0.927 CI: [0.924-0.930]	0.938 CI: [0.935-0.941]	0.947 CI: [0.944-0.950]	0.956 CI: [0.953-0.959]
Kaggle DR	JFnet	(0) vs (1,2,3,4)	0.911 CI: [0.908-0.914]	0.918 CI: [0.914-0.921]	0.925 CI: [0.921-0.929]	0.932 CI: [0.928-0.935]
Kaggle DR	JFnet	(0, 1) vs (2,3,4)	0.947 CI: [0.944-0.950]	0.953 CI: [0.949-0.956]	0.954 CI: [0.951-0.958]	0.956 CI: [0.952-0.960]
Messidor	Bayes. CNN	(0) vs (1,2,3)	0.936 CI: [0.922-0.949]	0.948 CI: [0.935-0.960]	0.956 CI: [0.943-0.968]	0.968 CI: [0.956-0.978]
Messidor	Bayes. CNN	(0, 1) vs (2,3)	0.955 CI: [0.943-0.967]	0.965 CI: [0.953-0.975]	0.973 CI: [0.962-0.983]	0.978 CI: [0.967-0.988]
Messidor	Single best[38]	(0) vs (1, 2, 3)	0.936	-	-	-
Messidor	Ensemble[38]	(0) vs (1, 2, 3)	0.989	-	-	-

Table 1. Model performance (measured by AUC) with two different datasets, architectures and tasks when data with higher uncertainty levels is referred to further inspection.

What causes uncertainty?

Next we asked what causes the networks to consider the prediction about an image uncertain. In order to build an intuitive understanding of uncertainty estimates, we trained a simple Bayesian neural network (3 hidden layers with 100 units each) with dropout layers interleaved ($p_{drop} = 0.5$) on a 2D toy classification problem (Fig. 4).

The network learns the non-linear hyperplane (defined by $p(y = 1|\mathbf{x}, \theta) = 0.5$) that separates the two classes (Fig. 4 (a)) shown as the network's softmax output when evaluated traditionally, that is with dropout turned off at test time. The first (Fig. 4 (b), eq. 7) and second moment (Fig. 4 (c), eq. 8) of the approximate predictive posterior (Eq. 6) in turn are more spread out along directions orthogonal to the separating hyperplane. Predictions with low confidence (i.e. high uncertainty, compare Fig. 1) may simply by chance be correct (Fig. (4)). Vice versa, erroneous predictions with low uncertainty may be attributed to outliers. As real world data ((e.g. fig. 1 (d))) may suffer from label noise, particularly confident predictions may be evaluated as incorrect because of wrong labels. Most importantly however, the network seems to show high uncertainty predominantly for the difficult cases - i.e. those that reside in the vicinity of the decision boundary (Fig. 4 (c)).

In the following we devised two experiments that aimed to assess whether these considerations generalize to the high-dimensional image space for DR detection. Is it predominantly difficult diagnostic decisions that carry a high uncertainty? We want to emphasize that while this may seem obvious, it is far from guaranteed that the approximation of the approximate predictive posterior is quantitatively good enough to identify such cases.

The first experiment makes use of the gradual progression of disease levels from 0 to 4 in case of Kaggle DR data. We probed what happened to images neighbouring the healthy/diseased boundary defined by our two tasks with different disease onset level. To this end, we quantified the proportion of the different disease levels in the data referred for further inspection for various uncertainty thresholds (Fig. 5).

As the minimum σ_{pred} increases, there is a shift from the prior distribution (shown on the vertical axis at $\sigma_{pred} = 0$) towards those disease levels that are adjacent to the healthy/diseased boundary (black lines in Fig. 5 (a) & (b)). For mild DR defining the disease onset and large tolerated uncertainties, disease levels 0 and 1 dominate the pool of referred data (Fig. 5 (a)). If we shifted the disease onset to moderate DR, in an analogous manner disease levels 1 and 2 dominated the referred data sets for high tolerated uncertainties (Fig. 5 (b)). As a side note, depending on the therapeutic possibilities - moderate DR detection (Fig. 5 (a)) might be preferable to mild DR detection (Fig. 5, (a)) as the uncertainty still detected level 1 patients in the latter case but reduced the amount of healthy patients sent for referral. In summary, figure 5 suggests that it is in particular the difficult diagnostic cases close to the decision boundary that carry a high uncertainty with their predictions.

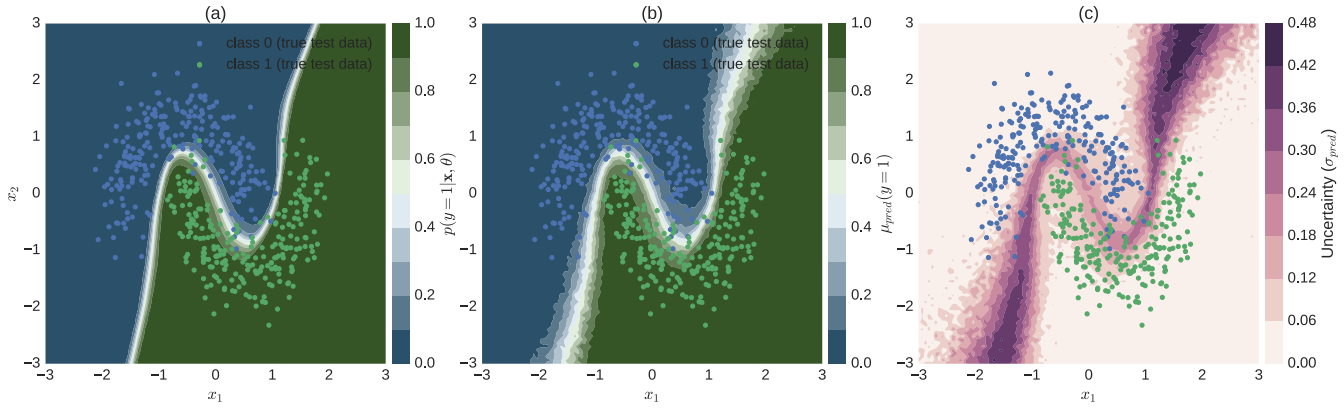

Figure 4. Illustration of uncertainty for a 2D binary classification problem. (a) Conventional softmax output obtained by turning off dropout at test time (Eq. 1). **(b)** Predictive mean of approximate predictive posterior (Eq. 7). **(c)** Uncertainty, measured by predictive standard deviation of approximate predictive posterior (Eq. 8). The softmax output (a) is overly confident (only a narrow region in input space assumes values other than 0 or 1) when compared to the Bayesian approach (b, c). Uncertainty (c) tends to be high for regions in input space that are close to the decision boundary. Colour-coded dots in all subplots correspond to test data the network has not seen during training.

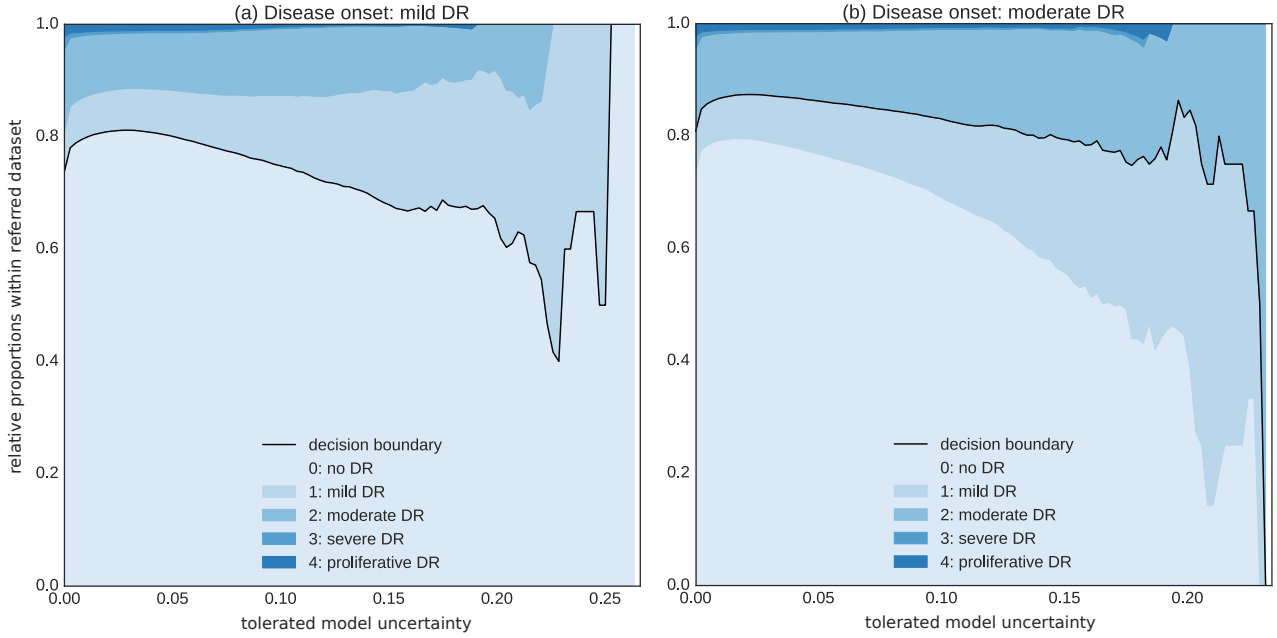

Figure 5. Predominant decision referral for disease levels neighbouring healthy/diseased boundary. The relative proportions of different disease levels within the referral dataset are shown for uncertainty greater or equal to the value indicated on the x-axis. There is a change in the relative proportion of classes compared to the prior distribution (tolerated uncertainty = 0, i.e. the entire test dataset is flagged for referral) towards the classes directly neighbouring the healthy/diseased boundary (depicted by a black line) as the tolerated uncertainty increases. **(a)** Disease onset level is mild DR (1). Disease levels 0 and 1 neighbour the healthy/diseased boundary and dominate the referral dataset under high uncertainty. **(b)** Disease onset level is moderate DR (2). In analogy to (a), disease levels 1 and 2 neighbour the healthy/diseased boundary and dominate the decision referral populations with high uncertainty values.

The second experiment to analyse whether the proposed uncertainty measure captures predominantly difficult diagnostic cases makes use of the availability of both eyes' images for each patient in case of Kaggle DR data. Even though therapeutic intervention is typically based on a patient level diagnosis, the contra-lateral eye of a given patient may be in a different state and therefore carry a different label. A strong correlation of the two eyes' disease states was leveraged to improve performance

by many Kaggle competition participants[40]. However, even after compilation of the 5-class problem to the binary disease detection problem, $5-10\%$ of images categorized as diseased have images from the contra-lateral eye with a disagreeing label.

Whether the corresponding patients are diseased or not is therefore unclear. By measuring the proportion of images whose contra-lateral ground truth label is different for the referred and retained data sets respectively (Fig. 6), we got another view onto the proposed uncertainty for difficult images. Indeed, images from patients with one healthy and one diseased eye are more likely to be referred for further inspection than retained (Fig. 6). For both disease detection tasks (compare Fig. 6 (a)/(b) for mild/moderate DR as disease onset respectively) this is particularly pronounced in the high uncertainty regime.

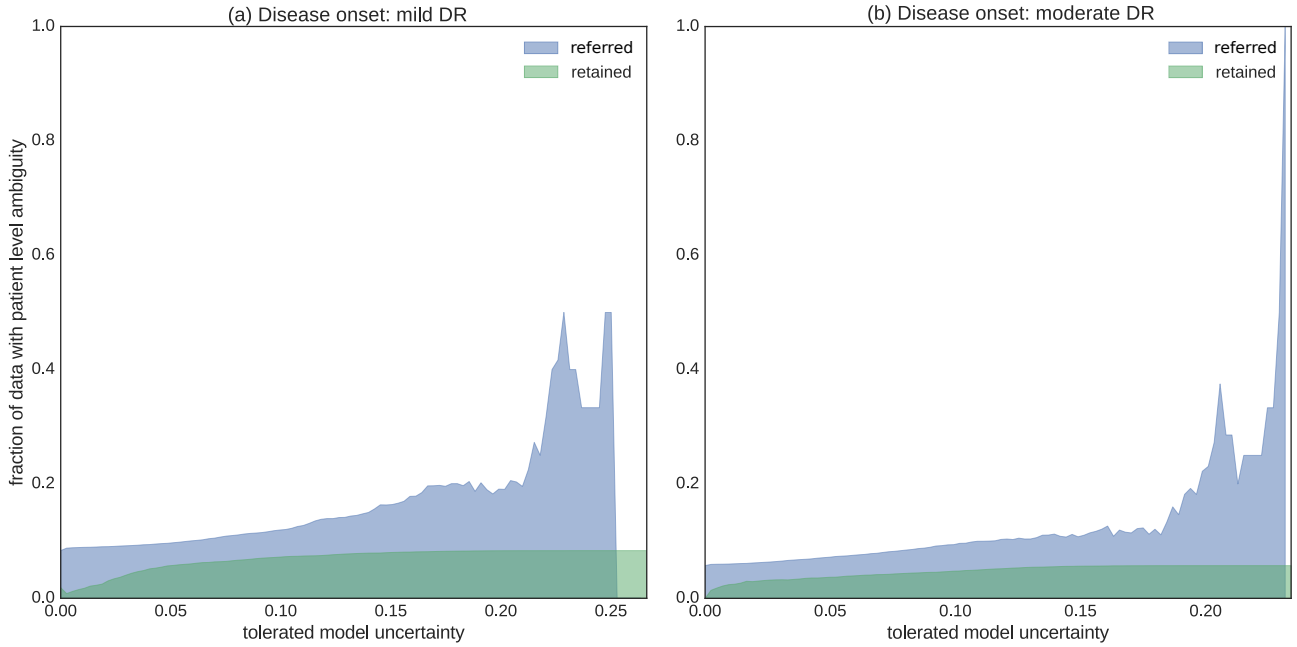

Figure 6. Decision referral of images from ambiguous patients. (a) Disease onset is mild DR (1). **(b)** Disease onset is moderate DR (2). Both subplots show the relative proportion of images from ambiguous patients in the referred (blue) and retained (green) data buckets for various tolerated uncertainty values. Patient level ambiguity is defined by images whose contra-lateral eye (from the same patient) carries a different label. Note that the decision referral of images is based on the uncertainty from a single image (using the BCNN). Ground truth labels and the contra-lateral eye information are only used as meta information for evaluation purposes. Especially in the high uncertainty regime, images from ambiguous patients are more likely to be referred for further inspection than accepted for automatic decision.

Discussion

Here we showed that it is feasible to associate deep learning based predictions about the presence of DR in fundus photographs with uncertainty measures that are informative and interpretable. Using the recently identified connection between dropout networks and approximate Bayesian inference[28,29], we computed meaningful uncertainty measures without needing additional labels for an explicit *uncertain* category. Computing this uncertainty measure was efficient, as computing the approximate predictive posterior for one image took us about $\approx 200ms$.

We achieved state-of-the art performance (Table 1) for DR detection under several settings (Fig. 3), in particular with a Bayesian CNN trained on Kaggle DR and tested on Messidor. The performance achieved by our networks met the requirements for UK national standards for automatically diagnosing DR. For all settings we could improve performance in terms of ROC AUC (Fig. 3) by referring uncertain, that is difficult (Fig. 4, 5, 6) cases for further inspection. Acquiring further opinions naturally integrates into the traditional clinical work flow as well as into a human-machine loop in which especially attended, difficult cases could be fed back into the model for its continuous improvement[50].

We observed slightly worse performance on Kaggle data as compared to Messidor. We want to point out, that the quality of the former dataset was questioned previously - albeit informally, both by competition participants[40] as well as by clinicians[48]. The extent to which the set of images considered uncertain by our approach overlaps with the images considered ungradable or wrongly labelled by humans is, however, unclear. Because images considered ungradable by clinical specialists may coincide with difficult diagnostic cases, these should be identifiable via high uncertainties from our approach. Easy decisions for images with wrong labels in turn should cause wrong predictions with low uncertainty. Both situations could hence be identified by our approach and be used to selectively reduce label noise and improve model performance.

The scope of which scenarios our approach is able to deal with can be understood by our results regarding the causes of uncertainty. We showed that it is in particular the difficult decisions that are considered uncertain by the networks, both for the 2D toy examples (Fig. 4) as well as for the high-dimensional image case (Fig. 5 & 6). In addition, it may seem desirable from a practical point of view for bad quality images (e.g. due to under-illumination or out-of-focus image captioning) or non-fundus images (e.g. due to database noise) to be considered uncertain. Neither of the two cases can however be reliably detected via a high uncertainty because there is no guarantee that the corresponding image transformations are in any way linked to regions of high uncertainty (Fig. 4). Images far from the training data in turn can be detected as long as they matter for the task - that is as long as they reside in the vicinity of the (extrapolated) decision boundary.

We conclude that this work successfully demonstrated the benefits and applicability of uncertainty in deep learning[51] for disease detection. This paradigm can be applied to other medical tasks and datasets as initial work on image registration[52] and genome data[53] has already shown. We also believe that segmentation[27] and regression[54] problems which are omnipresent in biomedical imaging and diagnostics could largely benefit from taking uncertainty into account.

Methods

General DNN methodology

Software and code availability

We used the deep learning framework Theano[55](0.9.0dev1.dev-RELEASE) together with the libraries Lasagne[56](0.2.dev1) and Keras[57](1.0.7). Network trainings and predictions were performed using a NVIDIA GeForce GTX 970 and a GeForce GTX 1080 with cuda versions 7.5/8 and cuDNN 4/5. Our reimplementation of the JFnet together with the provided weights[43], achieved a quadratic weighted kappa score (the performance measure used by the Kaggle DR competition) of 0.8160/0.8311 on the private/public leaderboard test sets respectively (on the original 5 class problem), for comparison we refer to the competition website[40]. All code and models for fast DR detection under uncertainty will be publicly available upon publication at https://bitbucket.org/cleibig/disease-detection.

Image preprocessing

All images were cropped to a squared centre region and resized to 512x512 pixels. In order to compensate for the decreased network depth in case of the Bayesian CNNs we additionally subtracted the local average colour for contrast enhancement purposes as described[58] and used[13] previously. Images fed to the JFnet were normalized the same way as had been used for training by the author[43], whereas those fed to the BCNNs were standard normalized for each colour channel separately.

Network training

We trained one Bayesian CNN for each disease detection task using 80% of the Kaggle DR training data. We minimized the cross-entropy plus regularization terms (Eq. 5) using stochastic gradient descent with a batch size of 32 and Nesterov updates (momentum=0.9). All parameters were initialized with the weights from the JFnet. Final weights were chosen based on the best ROC AUC achieved on a separate validation set (20% of Kaggle DR training data) within 30 training epochs. The learning rate schedule was piecewise constant (epoch 1-10: 0.005, epoch 11-20: 0.001, epoch 21-25: 0.0005, epoch 26-30: 0.0001).

L2-regularization ($\lambda = 0.001$) was applied to all parameters, L1-regularization ($\lambda = 0.001$) to only the last layer in the network. Data augmentation was applied to 50% of the data in an epoch. Affine transformations were composed by drawing uniformly from ranges for zooming ($\pm 10\%$), translating (independent shifts in x- and y-directions by ± 25 pixels), and rotating ($\pm\pi$). Transformed images were in addition flipped along the vertical and/or the horizontal axis if indicated by respective draws from a Binomial distribution ($\mu = 0.5$). Effects of class imbalance onto the stochastic gradient were compensated by attributing more weight to the minority class, given by the relative class frequencies in each mini-batch[59] $p(k)_{mini-batch}$. To achieve this, we reweighed the cross-entropy part of the cost function (compare eq. 5) for a mini-batch of size n to:

$$-\frac{1}{Kn}\sum_{i=1}^{n}\frac{\log\frac{e^{f(\mathbf{x}_i,\theta(\hat{\omega}_i)_k)}}{\sum_j e^{f(\mathbf{x}_i,\theta(\hat{\omega}_i)_j)}}}{p(k)_{mini-batch}}$$

We fixed the amount of dropout for the convolutional layers to $p_{drop} = 0.2$, because this was a good compromise between getting a reasonable performance and uncertainty measures. We observed convergence problems for larger p_{drop} when initializing the Bayesian CNNs with the pretrained weights from the network without dropout between conv. layers. Gradually increasing dropout during training could potentially ease convergence. Alternatively, the dropout rates could be learnt via *variational dropout*[25].

Approximate Bayesian model uncertainty for deep learning

Recently, it was shown[28] that a multi-layer-perceptron (i.e. a stack of densely connected layers) with dropout after every weight layer is mathematically equivalent to approximate variational inference[46] in the deep Gaussian process (GP) model[60,61]. This result holds for any number of layers and arbitrary non-linearities. Next, this idea was extended to incorporate convolutional layers[29], potentially loosing the GP interpretation, but preserving the possibility to obtain an approximation to the predictive posterior in a Bayesian sense. Here, we summarize the core idea for deep classification networks and highlight in particular the difference between the Bayesian perspective and the classification confidence obtained from the softmax output.

Softmax vs. Bayesian uncertainty

DNNs (with or without convolutional layers) for classifying a set of N observations $\{\mathbf{x}_1,...,\mathbf{x}_i,...,\mathbf{x}_N\}$ into a set of associated class memberships $\{\mathbf{y}_1,...,\mathbf{y}_i,...,\mathbf{y}_N\}$ with $y_i \in \{1,...,K\}$, and K the number of classes, can be trained by minimizing the cross-entropy between the distribution of the true class labels and the softmax network output:

$$p(y_i = k|\mathbf{x}_i,\theta) = \frac{e^{f(\mathbf{x}_i,\theta_k)}}{\sum_j e^{f(\mathbf{x}_i,\theta_j)}} \tag{1}$$

Equation (1) denotes the probability that the observation $\mathbf{x}_i$ belongs to class k, if propagated through the network function f with all parameters summarized by θ, i.e. weights $\mathbf{W}_i$ and biases $\mathbf{b}_i$ of all layers $i \in \{1,...,L\}$. For the example of disease detection from images, we have a single unit whose output denotes the probability for the presence of the disease in a given image.

Cross-entropy minimization results in a single best parameter vector θ, constituting the maximum-likelihood solution. $L2$-regularization, typically used to prevent overfitting, is equivalent to putting a Gaussian prior on the network parameters, resulting in a maximum-a-posteriori (MAP) solution.

A fully probabilistic treatment in a Bayesian sense, however, would consider a distribution over network parameters instead of a point estimate. By integrating over the posterior $p(\theta|\mathbf{X},\mathbf{y},\mathbf{x}^*)$ given the entire training data $\{\mathbf{X},\mathbf{y}\}$ and a new test sample $\mathbf{x}^*$ one would like to obtain the predictive posterior distribution over class membership probabilities:

$$p(y^*|\mathbf{X},\mathbf{y},\mathbf{x}^*) = \int p(y^*|\theta)p(\theta|\mathbf{X},\mathbf{y},\mathbf{x}^*)d\theta \tag{2}$$

Whereas equation (1) determined a single value specifying the probability that an image belongs to the *diseased* class, the predictive posterior (Eq. 2) defines a distribution of such predictions, that is the probability values that a single image is *diseased*. Intuitively, the width of the predictive posterior should reflect the reliability of the predictions. For large training data sets, the parameter point estimates (from maximum-likelihood or MAP) may correspond to the mean or mode of the predictive posterior, resulting in a potentially strong relationship between the width of the predictive posterior and the softmax output, however this is not guaranteed. Indeed we've found that only for the original JFnet the softmax output may be used as a proxy for (prediction instead of model) uncertainty (values close to 0.5 were considered uncertain, data not shown), whereas the Bayesian treatment worked for all investigated scenarios.

Bayesian convolutional neural networks with Bernoulli approximate variational inference

In practice, equation (2) is intractable and a common way to find approximating solutions is via *variational inference*. We assume the true posterior to be expressible in terms of a finite set of random variables ω. The posterior is then approximated by the *variational distribution* $q(\omega)$ as follows:

$$p(\theta|\mathbf{X},\mathbf{y},\mathbf{x}^*) \approx \int p(\theta|\mathbf{x}^*,\omega)p(\omega|\mathbf{X},\mathbf{y})d\omega \approx \int p(\theta|\mathbf{x}^*,\omega)q(\omega)d\omega \tag{3}$$

Maximizing the *log evidence lower bound* with respect to the approximating distribution $q(\omega)$:

$$\mathscr{L}_{VI} := \int p(\mathbf{y}|\mathbf{X},\omega)q(\omega)d\omega - KL(q(\omega)||p(\omega))) \tag{4}$$

has two effects. The first term maximizes the likelihood of the training data $\{\mathbf{X},\mathbf{y}\}$, whereas the second term takes care of approximating the true distribution $p(\omega)$ by $q(\omega)$. The key insight from *Gal & Ghahramani* was then to link equation (4) with dropout training. Here, we will summarize the derivations[31] in words. The integral in eq. (4) is still intractable and therefore approximated with Monte Carlo sampling. This results in the conventional softmax loss for dropout networks, for which units are dropped by drawing from a Bernoulli prior with probability p_{drop} for setting a unit to zero. The KL term in (4) was shown[31] to correspond to a L2-regularization term in dropout networks. Summing up, approximate variational inference with a Bernoulli approximating distribution can be performed with the following loss:

$$\hat{\mathscr{L}}_{VI} := \mathscr{L}_{dropout} = -\sum_{i=1}^{N} \log \frac{e^{f(\mathbf{x}_i,\theta(\hat{\omega}_i)_k)}}{\sum_j e^{f(\mathbf{x}_i,\theta(\hat{\omega}_i)_j)}} + \lambda \sum_{i=1}^{L} ||\theta_i(\hat{\omega}_i)||^2 \quad \hat{\omega}_i \sim q(\omega) \tag{5}$$

We use $\hat{\omega}_i$ as a shorthand notation for stating that in order to decide whether a unit is dropped, we independently sample from a Bernoulli distribution (with probability p_{drop}) for each unit in all layers for each training sample. Note that Monte Carlo sampling from $q(\omega)$ is equivalent to performing dropout during training, hence we get the Bayesian network perspective as well for already trained models.

Obtaining model uncertainty at test time

Obtaining model uncertainty for a given image is as simple as keeping the dropout mechanism switched on at test time and performing multiple predictions. The width of the distribution of predictions is then a reasonable proxy for the model uncertainty. More formally expressed, we replace the posterior with the approximating distribution (Eq. 3) and plug it into the expression for the predictive posterior (2):

$$p(y^*|\mathbf{X},\mathbf{y},\mathbf{x}^*) \approx \int p(y^*|\mathbf{x}^*,\omega)q(\omega)d\omega \tag{6}$$

We then approximate the integral by Monte Carlo sampling and compute the predictive mean (to be used for a final prediction on a test image):

$$\mu_{pred} \approx \frac{1}{T}\sum_{t=1}^{T} p(y^*|\mathbf{x}^*,\theta(\hat{\omega}_t)) \tag{7}$$

as well as the predictive standard deviation as a proxy for the uncertainty associated with this prediction:

$$\sigma_{pred} \approx \frac{1}{T-1}\sqrt{\sum_{t=1}^{T}(p(y^*|\mathbf{x}^*,\theta(\hat{\omega}_t)) - \mu_{pred})^2} \tag{8}$$

For this work, we fixed $T = 100$ because it was shown by[29] to suffice. The test predictions could be performed in parallel, but even a serial implementation takes less than $200ms$ per image.

Analysis of results

All density plots are based on Gaussian kernel density estimates, for which the bandwidth was chosen based on Scott's method[62]. All line plots are based on the entire data and the 95% confidence intervals were obtained from 10^4 bootstrap samples.

References

1. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **61**, 85–117 (2015). URL http://arxiv.org/abs/1404.7828. arXiv:1404.7828v1.

2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems* 1–9 (2012). 1102.0183.

3. Rusk, N. Deep learning. *Nature Methods* **13**, 35–35 (2016). arXiv:1312.6184v5.

4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015). URL http://www.nature.com/doifinder/10.1038/nature14539.

5. Ciresan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **8150 LNCS**, 411–418 (2013).

6. Greenspan, H., van Ginneken, B. & Summers, R. Guest Editorial Deep Learning in Medical Imaging : Overview and Future Promise of an Exciting New Technique **35**, 1153–1159 (2016).

7. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific reports* **6**, 26094 (2016).

8. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports* **6**, 26286 (2016). URL http://dx.doi.org/10.1038/srep26286http://www.nature.com/articles/srep26286.

9. Chen, C. L. *et al.* Deep Learning in Label-free Cell Classification. *Scientific reports* **6**, 21471 (2016). URL http://www.nature.com/srep/2016/160315/srep21471/full/srep21471.html.

10. Lipton, Z. C., Kale, D. C., Elkan, C. & Wetzell, R. Learning to Diagnose with LSTM Recurrent Neural Networks 1–18 (2015). URL http://arxiv.org/abs/1511.03677. 1511.03677.

11. Lu, L. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer Learning Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer. *IEEE Transactions on Medical Imaging* **35**, 1285–1298 (2016). 1602.03409.

12. Tajbakhsh, N. *et al.* Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging* **35**, 1299–1312 (2016). URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7426826.

13. Grinsven, M. J. J. P. V., Ginneken, B. V., Hoyng, C. B. & Theelen, T. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images **0062**, 1–12 (2016).

14. Widdowson, D. T. S. The management of grading quality Good practice in the quality assurance of grading. Tech. Rep. March (2016). URL https://www.gov.uk/government/uploads/system/uploads/attachment{_}data/file/512832/The{_}Management{_}of{_}Grading.pdf.

15. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459 (2015). URL http://dx.doi.org/10.1038/nature14541.

16. Kononenko, I. Inductive and Bayesian Learning in Medical Diagnosis. *Applied Artificial Intteligence* **7**, 317–337 (1993).

17. Kononenko, I. Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine* **23**, 89–109 (2001).

18. Wang, S. & Summers, R. M. Machine learning and radiology. *Medical Image Analysis* **16**, 933–951 (2012). URL http://dx.doi.org/10.1016/j.media.2012.02.005.

19. Sajda, P. Machine learning for detection and diagnosis of disease. *Annual review of biomedical engineering* **8**, 537–65 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16834566http://www.annualreviews.org/doi/10.1146/annurev.bioeng.8.061505.095802.

20. Tishby, N., Levin, E. & Solla, S. A. Consistent inference of probabilities in layered networks: predictions and generalizations. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, 403–409 vol.2 (1989).

21. MacKay, D. J. C. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation* **4**, 448–472 (1992).

22. Hinton, G. E., Hinton, G. E., van Camp, D. & van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. *Proceedings of the sixth annual conference on Computational learning theory - COLT '93* 5–13 (1993). URL http://portal.acm.org/citation.cfm?doid=168304.168306.

23. Neal, R. M. Bayesian learning for neural networks. *Lecture notes in statistics* 183 s. (1996).

24. Graves, A. Practical Variational Inference for Neural Networks. *Nips* 1–9 (2011). URL https://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf.

25. Kingma, D. P., Salimans, T. & Welling, M. Variational Dropout and the Local Reparameterization Trick. *arXiv* 1–13 (2015). URL http://arxiv.org/abs/1506.02557. 1506.02557.

26. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D. & Deepmind, G. Weight Uncertainty in Neural Networks **37** (2015).

27. Kendall, A., Badrinarayanan, V. & Cipolla, R. Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv:1511.02680v1 [cs.CV]* (2015). URL http://arxiv.org/abs/1511.02680. 1511.02680.

28. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *International Conference on Machine Learning* **48**, 1–10 (2015). URL http://arxiv.org/abs/1506.02142. 1506.02142.

29. Gal, Y. & Ghahramani, Z. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. *International Conference on Learning Representations (ICLR)* 12 (2015). URL http://arxiv.org/abs/1506.02158. 1506.02158.

30. Louizos, C. & Welling, M. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. *Icml* **48** (2016). URL http://arxiv.org/abs/1603.04733. 1603.04733.

31. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Appendix 20 (2016). URL http://arxiv.org/abs/1506.02157. 1506.02157.

32. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv e-prints* 1–18 (2012). URL http://arxiv.org/abs/1207.0580. 1207.0580.

33. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)* **15**, 1929–1958 (2014). 1102.4807.

34. Zaki, W. M. D. W. *et al.* Diabetic retinopathy assessment: Towards an automated system. *Biomedical Signal Processing and Control* **24**, 72–82 (2016). URL http://dx.doi.org/10.1016/j.bspc.2015.09.011.

35. World Health Organization. Global Report on Diabetes. *Isbn* **978**, 88 (2016). arXiv:1011.1669v3.

36. Mane, V. M. & Jadhav, D. V. Progress towards automated early stage detection of diabetic retinopathy: Image analysis systems and potential. *Journal of Medical and Biological Engineering* **34**, 520–527 (2014).

37. Kapetanakis, V. V. *et al.* A study of whether automated Diabetic Retinopathy Image Assessment could replace manual grading steps in the English National Screening Programme. *Journal of medical screening* **22**, 112–8 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25742804http://msc.sagepub.com/lookup/doi/10.1177/0969141315571953.

38. Antal, B. & Hajdu, A. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems* **60**, 20–27 (2014). arXiv:1410.8576v1.

39. Sundling, V., Gulbrandsen, P. & Straand, J. Sensitivity and specificity of Norwegian optometrists' evaluation of diabetic retinopathy in single-field retinal images - a cross-sectional experimental study. *BMC health services research* **13**, 17 (2013). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3562149{&}tool=pmcentrez{&}rendertype=abstract.

40. https://www.kaggle.com/c/diabetic-retinopathy-detection. URL https://www.kaggle.com/c/diabetic-retinopathy-detection.

41. Wu, L., Fernandez-Loaiza, P., Sauma, J., Hernandez-Bogantes, E. & Masis, M. Classification of diabetic retinopathy and diabetic macular edema. *World journal of diabetes* **4**, 290–4 (2013). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3874488{&}tool=pmcentrez{&}rendertype=abstract.

42. http://messidor.crihan.fr. URL http://messidor.crihan.fr.

43. De Fauw, J. 5th place solution of the Kaggle Diabetic Retinopathy competition (2015). URL https://github.com/JeffreyDF/kaggle{_}diabetic{_}retinopathy.

44. Nair, V. & Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning* 807–814 (2010).

45. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. *Icml-2013* **28** (2013). URL http://www.stanford.edu/{~}awni/papers/relu{_}hybrid{_}icml2013{_}final.pdf.

46. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006).

47. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recoginition. *Intl. Conf. on Learning Representations (ICLR)* 1–14 (2015). 1409.1556v6.

48. Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P. & Zheng, Y. Convolutional Neural Networks for Diabetic Retinopathy. *Procedia - Procedia Computer Science* **00**, 6–8 (2016). URL http://dx.doi.org/10.1016/j.procs.2016.07.014.

49. Younis, N., Broadbent, D. M., Harding, S. P. & Vora, J. P. Incidence of sight-threatening retinopathy in Type 1 diabetes in a systematic screening programme. *Diabetic medicine : a journal of the British Diabetic Association* **20**, 758–765 (2003).

50. Settles, B. Active Learning Literature Survey. *Machine Learning* **15**, 201–221 (2010). 1206.5533.

51. Gal, Y. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge (2016).

52. Yang, X., Kwitt, R. & Niethammer, M. Fast Predictive Image Registration. *arXiv preprint* (2016). URL http://arxiv.org/abs/1607.02504. 1607.02504.

53. Angermueller, C. & Stegle, O. Multi-task deep neural network to predict CpG methylation profiles from low-coverage sequencing dat (2015).

54. Kendall, A. & Cipolla, R. Modelling Uncertainty in Deep Learning for Camera Relocalization. *arXiv preprint* (2016). arXiv:1509.05909v2.

55. Al-Rfou, R. *et al.* Theano: A {Python} framework for fast computation of mathematical expressions. *arXiv e-prints* **abs/1605.0** (2016). URL http://arxiv.org/abs/1605.02688.

56. Dieleman, S. *et al.* Lasagne 0.2.dev. https://github.com/Lasagne/Lasagne (2016).

57. Chollet, F. & Others. Keras 1.0.7. https://github.com/fchollet/keras (2016).

58. Graham, B. Kaggle Diabetic Retinopathy Detection competition report. Tech. Rep., University of Warwick (2015).

59. Dalyac, A., Shanahan, P. M., Kelly, J. & London, I. C. Tackling Class Imbalance with Deep Convolutional Neural Networks (2014).

60. Williams, C. K. I. Computing with infinite networks. *Neural Information Processing Systems* 1203–1216 (1998). URL http://eprints.aston.ac.uk/673/.

61. Damianou, A. C. & Lawrence, N. D. Deep Gaussian Processes. *International Conference on Artificial Intelligence and Statistics* **31**, 207–215 (2013). arXivpreprintarXiv:1211.0358.

62. Scott, D. W. On optimal and data-based histograms. *Biometrika* **66**, 605–610 (1979).

Acknowledgements

This work was funded by the German excellence initiative through the Institutional Strategy of the University of Tübingen and the Center for Integrative Neuroscience (EXC 307), the Bernstein Award for Computational Neuroscience by German Ministry for Education and Resarch (BMBF; FKZ: 01GQ1601) to PB. Additional support came from the early career program of the Medical Faculty of the University of Tübingen.