Max-Heinrich Laves*, Sontje Ihler, Lüder A. Kahrs, and Tobias Ortmaier

Quantifying the uncertainty of deep learning-based computer-aided diagnosis for patient safety

https://doi.org/10.1515/cdbme-2019-0057

Abstract: In this work, we discuss epistemic uncertainty estimation obtained by Bayesian inference in diagnostic classifiers and show that the prediction uncertainty highly correlates with *goodness* of prediction. We train the ResNet-18 image classifier on a dataset of 84,484 optical coherence tomography scans showing four different retinal conditions. Dropout is added before every building block of ResNet, creating an approximation to a Bayesian classifier. Monte Carlo sampling is applied with dropout at test time for uncertainty estimation. In Monte Carlo experiments, multiple forward passes are performed to get a distribution of the class labels. The variance and the entropy of the distribution is used as metrics for uncertainty. Our results show strong correlation with $\rho = 0.99$ between prediction uncertainty and prediction error. Mean uncertainty of incorrectly diagnosed cases was significantly higher than mean uncertainty of correctly diagnosed cases. Modeling of the prediction uncertainty in computer-aided diagnosis with deep learning yields more reliable results and is therefore expected to increase patient safety. This will help to transfer such systems into clinical routine and to increase the acceptance of machine learning in diagnosis from the standpoint of physicians and patients.

Keywords: bayesian approximation, optical coherence tomography, retina, machine learning

1 Introduction

Computer-aided diagnosis (CAD) based on deep learning has been demonstrated to achieve a performance similar to that of human experts in classification tasks in medical imaging [2]. A recent study established a diagnostic classifier based on convolutional neural networks (CNN), which was trained on a large database of more than 84,000 retinal OCT images of four different disease states [9]. The performance in classifying retinal conditions was comparable to that of trained physicians. Equipped with deep neural networks, mobile assistance systems can extend the reaching of ophthalmologists in the field and increase access to medical care. However, common tools in deep learning do not provide uncertainty for predictions of disease conditions. When an ambiguous or unknown case is presented to a deep learning model, it lacks the ability to say "I don't know". Especially in medical imaging and CAD, measure of uncertainty therefore is needed for profound decision making.

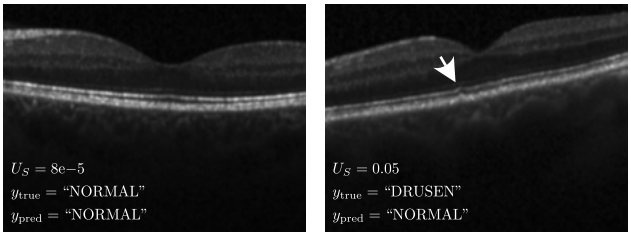

Fig. 1: Uncertainty estimation in retinal OCT scan. Left: Correctly predicted scan results in low prediction uncertainty U. Right: False prediction correlates with high prediction uncertainty. In this scan, the disease characteristics are only weakly present (white arrow), which makes the diagnosis challenging [9].

The main contributions of the work are the integration of uncertainty estimation with Bayesian inference into diagnostic classifiers with an exemplary dataset of retinal OCT scans and extensive experiments regarding the effect of uncertainty estimation to classification accuracy. We show that the prediction uncertainty correlates with accuracy, thus enabling the identification of false predictions or unknown cases, which were not covered during training (see Fig. 1). This paper extends our previous work published at Medical Imaging with Deep Learning (MIDL) 2019 with formal background and more experiments [11].

1.1 Related Work

This work is linked to classification of medical images for computer-aided diagnosis and uncertainty estimation for neu-

***Corresponding author: Max-Heinrich Laves,** Institute of Mechatronic Systems, Appelstr. 11A, 30167 Hannover, Germany, e-mail: laves@imes.uni-hannover.de
Sontje Ihler, Tobias Ortmaier, Institute of Mechatronic Systems, Appelstr. 11A, 30167 Hannover, Germany
Lüder A. Kahrs, Center for Image Guided Innovation and Therapeutic Intervention (CIGITI), The Hospital for Sick Children, 555 University Ave, Toronto, ON M5G 1X8, Canada, Department of Mathematical and Computational Sciences, University of Toronto Mississauga, 3359 Mississauga Rd, Mississauga, ON L5L 1C6, Canada

ral networks. We will therefore revise relevant work related to these topics in the following.

A demonstration of CAD for retinal OCT was shown by Kermany et al. [9]. Ophthalmologist-level detection was achieved with a classification accuracy of 96.6 % by fine-tuning the last layer of a CNN pre-trained for image classification. Using transfer learning was crucial for the training of a highly accurate model. A similar approach with a smaller dataset was reporting comparable results [6]. Instead of using a single CNN, Rasti et al. proposed a multi-scale CNN ensemble model of four individual networks to classify retinal OCT images into three classes [13]. In order to process 3D OCT volumes, the model analyzed the cross sections slice-by-slice and made a final diagnosis prediction per volume. With this approach, a precision rate of 98.9 % is reported using a total dataset of 193 OCT volumes.

Approaches for uncertainty estimation in deep learning try to place distributions over the parameters of a model. This results in Bayesian neural networks (BNN) and instead of finding a point estimate for the parameters, averaging over all possible parameters is done [4, 12]. BNNs provide the mathematical tool to model uncertainty, but are computationally intractable. Different approaches to approximate BNNs exist, that do not increase complexity or decrease the performance of the model [4, 8]. This idea has attracted attention in medical segmentation on MRI scans or endoscopic images [10, 14]. More recent approaches exploit model uncertainty for weakly-supervised anomaly detection in retinal OCT [15].

2 Materials and Methods

Standard approaches of supervised learning for classification do not capture model uncertainty [4]. The predictive probabilities of a softmax classifier at a point estimate $\hat{\theta}$ of the parameters of a machine learning model f_θ usually result in unjustified high confidence [8]. Placing a distribution over the parameters in a Bayesian approach better accounts for classification uncertainty.

In Bayesian inference, there are two types of uncertainty that can be modeled [7]. *Aleatoric* uncertainty or statistical uncertainty covers noise and randomness happening during an experiment. In OCT imaging, this can be caused by sensor noise during scan acquisition. *Epistemic* uncertainty or model uncertainty is caused by uncertainty in the parameters of a machine learning model. This happens because an inadequate model does not take all aspects of the data into account or because the training data is too small and do not represent all effects. This work focuses on epistemic uncertainty in order to assign uncertainty to the prediction of a diagnostic classifier in CAD applied to retinal OCT.

2.1 Background: Epistemic Uncertainty in Bayesian Inference

Let $\boldsymbol{X} = \{x_1, \ldots, x_N\}$ be a set of N training images, and $\boldsymbol{Y} = \{y_1, \ldots, y_N\}$ a set of corresponding labels from medical experts. In Bayesian inference, a diagnostic classifier tries to find a function $f_\theta : x \to y$ which maps a test scan x to the most likely label prediction $\hat{y}$ with probability

$$p(y|x, \boldsymbol{X}, \boldsymbol{Y}) = \int p(y|x, \theta) p(\theta|\boldsymbol{X}, \boldsymbol{Y}) \, \mathrm{d}\theta \, . \tag{1}$$

In order to obtain prediction uncertainties in Bayesian inference, we are interested in the posterior distribution $p(\theta|\boldsymbol{X}, \boldsymbol{Y})$ over the parameters θ of the network f, given the training set $\{\boldsymbol{X}, \boldsymbol{Y}\}$ [7]. Generally, the posterior distribution in Eq. (1) is computationally intractable, but can be approximated. Gal et al. [4] proposed a framework to approximate the posterior with $q(\theta)$ using a technique called *Monte Carlo Dropout* (MC dropout). Dropout randomly drops connections between layers of neural networks during training [16]. This results in sampling from different smaller subnetworks, over which we usually average during test time. Here, dropout is not only used at training time, but also at test time to obtain MC samples. The approximation $q(\theta)$ is defined as a Bernoulli distribution with probability p set to the dropout rate [3]. Training the network by optimizing the negative log-likelihood is equivalent to minimizing the Kullback-Leibler divergence $KL\{q(\theta)\|p(\theta|\boldsymbol{X}, \boldsymbol{Y})\}$ [3], thus yielding an approximate to the posterior predictive distribution

$$p(y|x, \boldsymbol{X}, \boldsymbol{Y}) \approx q(y|x) = \int p(y|x, \theta) q(\theta) \, \mathrm{d}\theta \, . \tag{2}$$

The integral in Eq. (2) is estimated by summing over MC samples drawn from $\theta \sim q(\theta)$. Sampling from $q(\theta)$ is done by performing dropout between every layer of the network f_θ [4].

2.2 Experimental Setup

A ResNet-18 image classifier [5] pre-trained on ImageNet is fine-tuned on 80,484 OCT scans showing four disease conditions. The Adam optimizer is used with a learn rate of $\eta = 1\mathrm{e}{-4}$ and batch size of 128. Early stopping is employed during fine-tuning for 10 epochs. The validation set for choosing hyper-parameters contains 1000 scans and the test set for reporting final results contains 2000 scans. Dropout with $p \in \{0.1, 0.25, 0.5\}$ is added between every building block of ResNet-18 and before the last linear layer, approximating a

Bayesian classifier. During test time, dropout is kept active and Monte Carlo sampling is performed by forwarding the input OCT x multiple times with $N = \{15, 25, 50\}$ through the Bayesian ResNet. This results in N predicted class label probabilities $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_N\}$ from which the mean value

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i \tag{3}$$

is used as final prediction result. Note that $\hat{y}_i$ represents relative softmax probabilities and do not account for prediction uncertainty [8]. To obtain uncertainties from MC predictions for class c, variances $\sigma_c^2 = \mathrm{Var}(\hat{y}_c)$ and normalized entropy [14]

$$H_c(y_c) = \mathbb{E}_{\hat{Y}_c}\left[-\log_N(P(y_c|x))\right] = -\sum_{i=1}^{N} \hat{y}_{c,i}(x) \log_N(\hat{y}_{c,i}(x)) \tag{4}$$

across the classes $c \in \{1, \dots, C\}$ over all MC samples is used. Normalization is achieved by using log to base N. The mean of class uncertainties for a test image x forms a final single-valued uncertainty measure $U_S(x) = \frac{1}{C}\sum_{c=1}^{C} \sigma_c^2$ and $U_H(x) = \frac{1}{C}\sum_{c=1}^{C} H_c$, respectively.

In order to show that uncertainty correlates with *goodness* of prediction, Spearman's correlation coefficient ρ is used between $U_{\{S,H\}}$ and the 1-Wasserstein distance [1]

$$W_1\left(P(y|x), Q(y|x)\right) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\left[|x - y|^1\right], \tag{5}$$

between true one-hot encoded class probabilities $P(y|x)$ and the mean of the MC samples $\hat{y} = Q(y|x)$. W_1 is therefore interpreted as error of prediction.

3 Results and Discussion

All results in this section are reported for the test set. Fig. 2 shows the effect of variation of the dropout rate p to the uncertainty measures $U_{\{S,H\}}$. Correlation between $U_{\{S,H\}}$ and the error of prediction W_1 is visualized in Fig. 3. Spearmans's correlations result in $\rho_S = 0.99$ for U_S and $\rho_H = 0.6$ for U_H.

Regardless of values for N and p, we can observe higher uncertainty for false classifications. The variation of the number of MC samples N has little effect on $U_{S,H}$. However, the relative differences of U_H between true and false predictions do not to change considerably with increasing N. Therefore, $N = 15$ seems to be sufficient for robust uncertainty estimation. Variation of dropout rate p with fixed $N = 50$ increases overall uncertainty U_S. Interestingly, the variation of U_S decreases with increasing p. The opposite effect can be observed for U_H, where an increase of p affects increasing variation.

Correlation analysis reveals, that there is strong correlation between uncertainty U_S and the prediction error W_1. For

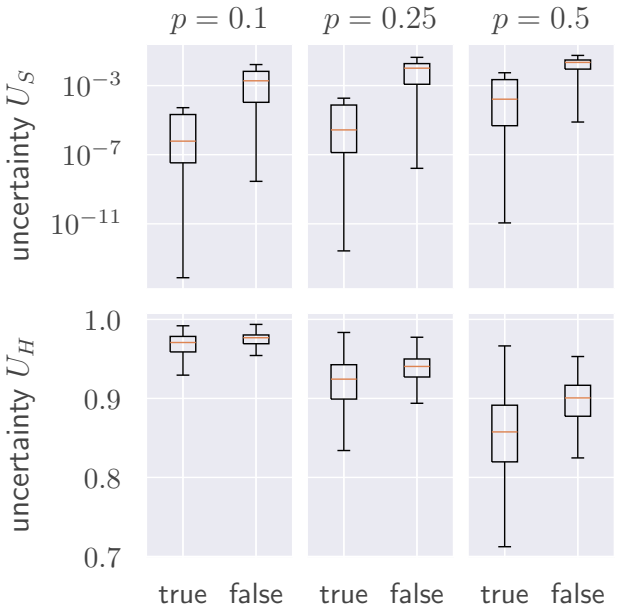

Fig. 2: Results for variation of dropout rate p at fixed number of MC forward passes $N = 50$. *True* means correctly classified OCT scans and *false* incorrect ones. Note the log scale in the upper row.

U_H moderate correlation can be observed. These findings suggest, that U_S is a robust measure for uncertainty and should be favored over U_H. MC dropout for uncertainty estimation can usually be used in every image classifier with the additional cost of performing multiple forward passes, thus increasing computational load.

4 Conclusion

In this work, a Bayesian diagnostic classifier has been trained on OCT scans to estimate model uncertainty. The results have shown strong correlation between uncertainty and error of prediction. It has been shown in [11] that the extension to Bayesian inference does not affect the accuracy of the classifier itself. Consideration of model uncertainty in CAD with deep learning is expected to increase patient safety by producing more robust results by e.g identifying a weak prediction of diagnosis. This can help to transfer such systems into clinical routine and to increase the acceptance of physicians and patients for machine learning in diagnosis. In future work, the reason for the different behavior of U_S and U_H for increasing dropout rate will be investigated. Uncertainties can help to identify false predictions, yield more reliable results and further increase the accuracy.

Author Statement

This research has received funding from the European Union as being part of the EFRE OPhonLas project.

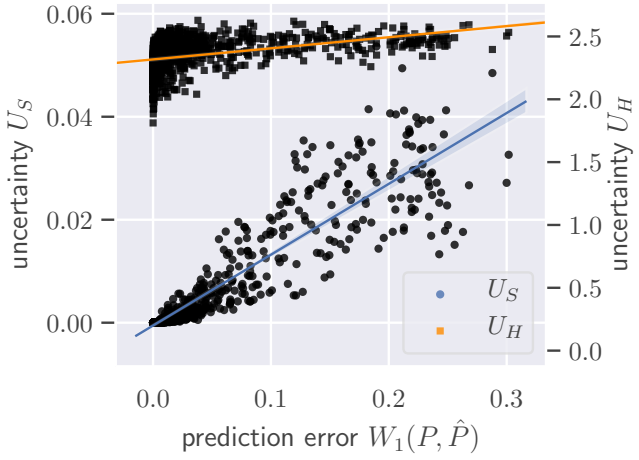

Fig. 3: Correlation between uncertainty U_S (blue circles) and U_H (orange squares), and error of prediction W_1 on the test set. Correlation is $\rho_S = 0.99$ and $\rho_H = 0.6$. Linear regression models have been fitted into the data to visualize linear relationship.

Conflict of Interest

The authors state no conflict of interest.

Informed Consent

The medical images used in this work were made public in a previous study [9], therefore formal consent is not required.

Ethical approval

For this kind of study ethical approval is not required.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv e-prints*, page arXiv:1701.07875, 2017. URL https://arxiv.org/abs/1701.07875.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 10.1038/nature21056.

[3] Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv e-prints*, page arXiv:1506.02158, 2015. URL https://arxiv.org/abs/1506.02158.

[4] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *ICML*, pages 1050–1059, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045502.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE CVPR*, pages 770–778, 2016.

[6] Q. Ji, W. He, J. Huang, and Y. Sun. Efficient Deep Learning-Based Automated Pathology Identification in Retinal Optical Coherence Tomography Images. *Algorithms*, 11(6):88, 2018. 10.1364/BOE.9.006205.

[7] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017. URL http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-computer-vision.

[8] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *arXiv e-prints*, page arXiv:1511.02680, 2015. URL https://arxiv.org/abs/1511.02680.

[9] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131, 2018. 10.1016/j.cell.2018.02.010.

[10] M.-H. Laves, J. Bicker, L. A. Kahrs, and T. Ortmaier. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *IJCARS*, 14(3):483–492, 2019. 10.1007/s11548-018-01910-0.

[11] M.-H. Laves, S. Ihler, and T. Ortmaier. Uncertainty Quantification in Computer-Aided Diagnosis: Make Your Model say "I don't know" for Ambiguous Cases. In *MIDL*, 2019. URL https://openreview.net/forum?id=rJevPsX854.

[12] D. J. MacKay. A practical bayesian framework for backpropagation networks. *Neural Comput*, 4(3):448–472, 1992. 10.1162/neco.1992.4.3.448.

[13] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh. Macular OCT Classification Using a Multi-Scale Convolutional Neural Network Ensemble. *IEEE Trans Med Imag*, 37(4):1024–1034, 2018. 10.1109/TMI.2017.2780115.

[14] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger. QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186:713–727, 2019. 10.1016/j.neuroimage.2018.11.042.

[15] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunovic, S. Klimscha, G. Langs, and U. Schmidt-Erfurth. Exploiting Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT. *IEEE Trans Med Imag*, Advance Online Publication, 2019. 10.1109/TMI.2019.2919951.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15:1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.