Marlin Siebert*, Nikolay Tesmer, and Philipp Rostalski

Stochastic variational deep kernel learning based diabetic retinopathy severity grading

https://doi.org/10.1515/cdbme-2022-1104

Abstract: The retinal disease Diabetic retinopathy (DR) is one of the most probable causes of blindness. Automatic detection of DR is mostly done using convolutional neural networks (CNNs) on colour retinal images. This work in contrast uses stochastic variational deep kernel learning (SVDKL) for DR grading, combining a deep CNN with Gaussian processes (GPs) into a single end-to-end trainable model, which promises to provide predictions with a reliable uncertainty estimate exploiting approximate Bayesian inference. Evaluating the performance and uncertainty calibration of SVDKL on DR grading compared to a plain CNN, the EfficientNet-B0, preliminary results on a subset of the Kaggle DR dataset show a naturally enhanced uncertainty calibration for SVDKL over the plain CNN as well as a good diagnostic performance. Despite SVDKL achieving a slightly reduced accuracy, incorrect predictions were in closer proximity to the target stages, which is beneficial for clinical diagnosis due to minimizing the cost related to severe misclassifications.

Keywords: deep kernel learning, uncertainty quantification, diabetic retinopathy, medical image processing, deep learning

1 Introduction

Diabetic retinopathy (DR) is a disease of the retina, the eye’s photoreceptive layer, causing an alteration of vascular tissue leading to retinal damage, which is one of the most probable causes of blindness [5]. Therapeutic success heavily depends on the early detection of DR and the correctly diagnosed severity stage [15]. Due to the increasing number of diabetic patients and the high prevalence of DR among these [15], it is desirable to automatically detect DR and grade the severity stage to help clinicians with selecting a proper therapy strategy. Recent studies showed that convolutional neural networks (CNNs) are suitable for this task and perform comparable to or even better than clinicians [7] and some are already approved for practical use [1]. Nevertheless, CNNs only provide point predictions and thus in general are lacking a well-calibrated uncertainty estimate [3]. In particular for medical applications with a high stake in patient safety, it is important to develop decision support systems that know when they don’t know and communicate their uncertainty to the user which can be used to reject the automatic diagnosis [13] and refer the patient to a specialist for manual inspection.

In contrast, Bayesian inference methods, such as Gaussian Processes (GPs), intrinsically provide this desired uncertainty information [3] in addition to good generalization abilities leading to high predictive accuracy [16]. However, GPs have limited representational learning capabilities compared to CNNs [3] and do not scale well with the number of input data and dimensions [12]. To this end, Wilson et al. proposed stochastic variational deep kernel learning (SVDKL) [17] that combines CNNs for meaningful feature extraction with GPs into an end-to-end trainable and scalable model.

In this work, we apply SVDKL to the task of DR severity grading using the Kaggle DR dataset1 and analyse whether using SVDKL can improve uncertainty calibration as well as diagnostic performance over using plain CNNs.

2 Experimental Setup

2.1 Image Data and Preprocessing

To train the models, the Kaggle DR dataset, a standard benchmark for DR classification consisting of high-resolution colour retinal images captured from different patients under diverse conditions, was used sticking to the predefined train-test-split. The images were graded by a specialist into 5 severity stages, i.e. (0) no-, (1) mild-, (2) moderate-, (3) severe- and (4) proliferative DR. To counter the dataset’s inherent class imbalance, both training and test images were uniformly subsampled to enable unbiased model training and prevent prior probability shift from training to test data and hence provide a fair comparison of both methods. To track training progress, 20% of the training data was uniformly sampled for model validation. With this, about 2.8k images of the Kaggle DR dataset were used for model training. For preprocessing, the images were semi-automatically cropped to the visible retinal disc and resized to 224 x 224 pixel resolution. Additionally, contrast limited

*Corresponding author: Marlin Siebert, Institute for Electrical Engineering in Medicine, Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany, e-mail: m.siebert@uni-luebeck.de
Nikolay Tesmer, Philipp Rostalski, Institute for Electrical Engineering in Medicine, Universität zu Lübeck, Lübeck, Germany

1 Kaggle DR detection challenge (2015)
adaptive histogram equalization (CLAHE) was applied to the L-channel after temporarily converting the RGB image to CIELAB colour space to enhance the image contrast. At training time, data augmentation was used in addition to the standard preprocessing, i.e. the images were randomly flipped and random adoption to brightness and contrast as well as random affine transformations were applied.

2.2 Model Implementation and Training

**EfficientNet**: In this work, we first trained a plain CNN to predict DR severity stages from the retinal colour images. For this purpose, the EfficientNet [14], which was built using neural architecture search and mobile convolutional blocks, was adopted. The authors optimized the model to be both low demanding in terms of computational resources as well as highly performant achieving state-of-the-art results. To subsequently enable straightforward training of the SVDKL model, the EfficientNet-B0 was selected due to its small size which was initialised with pre-trained weights from ImageNet. We lowered the number of input features to the final fully connected layer to 250 features to counter emerging overfitting on the reduced Kaggle DR training subset and adopted the number of output features to match the five DR severity stages. The model was trained over 60 epochs using the cross-entropy loss and Adam optimizer. To improve model convergence and generalization, a step scheduler lowering the learning rate at epochs 30, 45 and 54 was used, and L2-regularization and strong dropout were applied. Training progress is visualized on the left in fig. 1.

**SVDKL**: A deep kernel learning [18] model applies a GP to the latent features of a neural network and trains both the GP and CNN jointly through the joint marginal likelihood of the GP, which, in contrast to the plain CNN, provides predictions with well-calibrated uncertainty estimates. However, for the task of classification, the likelihood retrieved from the model is not Gaussian distributed due to the non-linear softmax activation that is required to compute the corresponding class probabilities. Wilson et al. solved this issue by approximating the non-Gaussian likelihood with stochastic variational inference in a factorized manner that allows for training the SVDKL model with stochastic gradient descent, and inherently provides good regularization capabilities and hence reduce overfitting. By using structure exploiting algebra, inducing points and kernel interpolation, the authors state that SVDKL is learning highly expressive kernel functions, provides good scalability and allows for exploiting deep architectures with many output features, stochastic mini-batch training, as well as multi-class learning [17]. Wilson et al. observed that using SVDKL can additionally improve classification performance over plain CNNs and GP classifiers on common benchmarks like CIFAR10 and ImageNet.

The implementation of the SVDKL model for this work was adopted from the GPyTorch toolbox of Wilson et al. Following this implementation, a single GP was learned for each feature provided by the feature extractor, whose outputs were subsequently mixed additively to produce the final model prediction. The SVDKL model was trained over 60 epochs using SGD as optimizer while being initialized with the pre-trained weights of the feature extractor of the trained EfficientNet-B0. During training, a step scheduler was used to decrease the learning rate on epochs 30 and 45. To speed-up training, the model was sampled only eight times during training, while 16 samples were drawn for testing to improve the quality of the uncertainty estimate. Training progress is visualized on the right in fig. 1.

2.3 Evaluation

Model training was repeated ten times while collecting the best-performing model on the validation data of each trial for evaluation. To evaluate both the plain CNN and SVDKL model, they were applied to the hold-out test set and the accuracy of the predicted severity grades was calculated. Additionally, the quadratic weighted Cohen’s kappa (QWK) [4] was computed, which is an inter-rater agreement score based on the regular Cohen’s kappa metric. However, by applying quadratically increasing weights to disagreements based on the distance of the prediction to the true target class, severe misclassifications, i.e. predictions far away from the true target class, are increasingly penalized over small deviations. This leads to a fast decrease in the QWK score if the models’ errors are more severe.

To further get to a thoughtful medical diagnosis and minimize the risk of misclassifications, not only a high overall predictive performance is important but also providing a well-calibrated uncertainty estimate of the model’s prediction, which
statistically would be higher for incorrect predictions and vice versa. Hence, a well-calibrated uncertainty estimate allows to identify difficult decisions and refer the corresponding patients for further inspection. In multi-class problems, the uncertainty can be calculated using the entropy over the predictive distribution of the class probabilities [9]. As the model’s confidence is correlated to the predictive entropy, the uncertainty can be approximated using the maximum confidence score. With this, we used the Expected Calibration Error (ECE) [8] to assess the quality of the predictive uncertainty by measuring the deviation of model confidence and accuracy. To compute the ECE, we created 20 uniformly distributed bins with adaptive locations by sorting the test samples into the bins according to their maximum confidence score, similar to [10, 11]. The ECE was then calculated as the average of the absolute differences between the mean confidence and mean accuracy for each bin [8].

### 3 Results and Discussion

As visible in fig. 1, the plain EfficientNet was difficult to train without severely overfitting the reduced Kaggle DR training subset despite applying strong regularization according to section 2.2. In contrast, we observed the SVDKL model be less prone to overfit the training data and hence ensure more robust model convergence. Accordingly, we reason that using SVDKL can in particular be beneficial when working on small clinical datasets, which have much fewer samples compared to standard benchmark datasets due to privacy issues and expensive data acquisition. Additionally, using CNNs with higher capacity to extract more meaningful features has great potential to benefit the SVDKL model’s performance, while an increasing model capacity using the plain CNN would amplify overfitting and impair generalization under the constraint of limited training data, as in our setting. Nonetheless, by including more training data from the Kaggle DR training set and applying e.g. random oversampling to counter the strong class imbalance in the Kaggle DR data, generalization of both the plain CNN and SVDKL model is assumed to improve.

Aligning the more robust convergence of the SVDKL model, we observe from the mid boxplot in fig. 2 that the median QWK score is higher for the SVDKL model compared to the EfficientNet. This indicates an improved overall predictive performance of the SVDKL model over the EfficientNet regarding the proximity of the predictions to the target stages measured through the QWK scores and hence a reduced risk for severe misclassifications. However, the median accuracy of the EfficientNet is slightly higher than for the SVDKL models, which is visible in the top boxplot in fig. 2, despite the latter achieving the highest accuracy. With the slight reduction in accuracy, our experiments lack the overall consistent improvement in performance using the SVDKL model over plain CNNs as demonstrated by Wilson et al. [17]. Why this drop occurs and how to improve the overall accuracy will be subject to future research, i.e. further investigations on the influence of the SVDKL model’s hyperparameters, as the number of input features to the GP-layer and the overall number of GPs used as well as the impact of using different neural architectures as feature extractor.

Considering the ECE displayed at the bottom of fig. 2, using the SVDKL model results in a lower median, showing that SVDKL naturally provides a slightly better-calibrated uncertainty estimate compared to the EfficientNet without applying post hoc confidence calibration methods, while the latter achieved the best-calibrated prediction considering the minima of the ECE scores. Therefore, taking a closer look at both the model’s confidence scores shows that the EfficientNet is mostly overconfident in contrast to the SVDKL model that is in many cases underconfident in its predictions revealing the strong regularising effect of SVDKL. However, the ECE score is known to be biased and does only use the prediction’s maximum confidence approximating the exact model uncertainty and thus only provides a limited metric for uncertainty calibration [2]. Hence, future work has to be conducted to extend upon the preliminary results for evaluating the uncertainty calibration of both the plain EfficientNet and SVDKL model, e.g. by using accuracy-rejection curves [2]. Additionally, comparing alternative methods to SVDKL, such as Monte Carlo dropout [6] or Gaussian stochastic weight averaging [10] is of great interest and will be subject to future research.
4 Conclusions

This work evaluates the effect of using an SVDKL model on the task of medical classification of DR fundus images in direct comparison to a plain CNN with respect to uncertainty calibration and diagnostic performance. The results show a slight improvement with using the SVDKL model over a plain EfficientNet-B0 in diagnostic performance measured through the QWK scores while the median accuracy of the SVDKL approach slightly decreased. This results in marginally more but simultaneously less severe misclassifications and hence could reduce the risk of major misjudgment of the required therapy. Additionally, the SVDKL models showed more robust convergence and achieved a lower calibration error without post hoc uncertainty optimization from which we conclude that SVDKL indeed provides better generalization and a more reliable uncertainty estimate compared to a plain CNN. Additionally, these overall promising results indicate that using SVDKL can not only be beneficial for DR grading but also could provide good generalization abilities for other medical tasks with limited data. Nevertheless, further research on improving overall accuracy and a more detailed evaluation of the uncertainty calibration is required and will be subject to future work.

Acknowledgment: This research is part of the project “Patientennahe Smartphone-basierte Diagnostik” (PASBADIA) kindly supported by the Joachim Herz Foundation.

References


