Weight averaging impact on the uncertainty of retinal artery-venous segmentation

Lindén Markus, Garifullin Azat, Lensu Lasse

DOI: 10.1007/978-3-030-60365-6_6

Weight averaging impact on the uncertainty of retinal artery-venous segmentation

Markus Lindén[1], Azat Garifullin[1][0000−0003−3300−6938], and Lasse Lensu[1][0000−0002−7691−121X]

LUT University, P.O. Box 20, 53851, Lappeenranta, Finland
{markus.linden, azat.garifullin, lasse.lensu}@lut.fi

Abstract. By examining the vessel structure of the eye through retinal imaging, a variety of abnormalities can be identified. Owing to this, retinal images have an important role in the diagnosis of ocular diseases. The possibility of performing computer aided artery-vein segmentation has been the focus of several studies during the recent years and deep neural networks have become the most popular tool used in artery-vein segmentation. In this work, a Bayesian deep neural network is used for artery-vein segmentation. Two algorithms, that is, stochastic weight averaging and stochastic weight averaging Gaussian are studied to improve the performance of the neural network. The experiments, conducted on the RITE and DRIVE data sets, and results are provided along side uncertainty quantification analysis. Based on the experiments, weight averaging techniques improve the performance of the network.

Keywords: Uncertainty quantification · Bayesian deep learning · Artery-vein segmentation · Blood vessel segmentation · Weight averaging

1 Introduction

Eye diseases have become a rapidly increasing health threat worldwide. Retinal images are a great tool for detecting some of the many ocular disease and diseases such as diabetic retinopathy and glaucoma can be detected from retinal images [12]. Ocular diseases are typically detected from retinal images by analyzing the vessel structure. The use of retinal images enables the diagnosis of ocular diseases in their early stages. The task of analyzing the vessel structure has been traditionally left to medical experts. The attention required by the medical experts in this tasks is, however, great and the task is very consuming and expensive. Studying the possibilities in making this process faster is for that reason important, as it would enable wider screenings for ocular diseases from retinal images. Automated image processing methods are a well-motivated possibility in solving this problem [3].

The possibility to use computers in performing artery-vein segmentation has been the focus of a number of studies during the recent years. However, artery-vein segmentation still remains a challenging tasks for both humans and machines alike. Some of the difficulties in artery-vein segmentation are related to the imaging conditions in which the retinal images are taken. The images tend to suffer from low contrast and changing lighting conditions, both of which make the segmentation process harder.

The deep convolutional neural network (DCNN) has recently become the most common tool used in artery-vein segmentation of retinal images, due to the DCNNs ability to automatically learn meaningful features from images. In a paper by Welikala et al., a convolutional neural network (CNN) was used in artery-vein segmentation. The CNN managed to achieve a 82.26% classification rate using UK Biobanks' retinal image database [13]. Hemelings et al. proposed the usage of U-Net architecture for artery-vein classification [5]. In the paper, Hemeling et al. considered the task as a multi-class classification problem with the goal of labeling pixels into four classes: background, vein, artery and unknown. The problem was solved using the retinal images found in DRIVE data set [6] and it achieved classification rates of 94,42% and 94.11% for arteries and veins. Girard et al. [3] modified the U-Net for artery-vein segmentation and found out that using likelihood score in the minimum spanning tree it was possible to improve the performance of the network in the case of smaller vessels. The method was tested using DRIVE data set, achieving an accuracy of 94.93%. Zhang et al. proposed cascade refined U-net to be used in artery-vein classification [14]. The cascade refined U-net consisted of three sub-networks. The task of the first sub-net (A-net in their paper) was to detect all the vessels from the input image, B-net segmented veins from the predicted vessels from the A-net, and finally the C-net segmented the arterioles from the outputs of the previous nets. In the paper, a classification rate of 97.27% was achieved using the automatically detected vessels from the RITE data set. In a paper by Garifullin et al., a dense fully convolutional neural network (Desne-FCN) was used in the task of artery-vein classification [2]. Using the Dense-FCN architecture and the RITE data the authors were able to achieve classification rates of 96%, 97% and 97% for vessels, arteries and veins respectively. In addition to that the authors performed uncertainty quantification on the results obtained using Monte-Carlo dropout [1] for variational approximation. In the aforementioned article, however, the authors did not illustrate the model calibration and the experiments were conducted with one training setup for different labelling strategies. Thus, the question of reliability of the shown uncertainty estimates arises.

This work illustrates how stochastic weight averaging affects the estimated uncertainties. In addition, differences between two epistemic uncertainty estimation techniques are illustrated. Both more traditional binary classification metrics as well as uncertainty quantification metrics are used to evaluate the algorithms.

2 Data

The retinal image data set chosen to be used in this work was the DRIVE data set [6]. The DRIVE data set contains 20 RGB images for testing and 20 for training. The images are of size 584 x 565.

The AV references standard used in this work is the RITE data set [7]. The RITE data set extends the DRIVE data set with references for arteries, veins, overlapping vessels and uncertain vessels. Red labels in the DRIVE data set stand for arteries, blue labels for veins, green for overlapping vessels and white ones for uncertain vessels. An example of a retinal image from the DRIVE data set as well as the corresponding data labels from the RITE data set can be seen in Fig. 1. During the training the labels for crossings were replaced by labels for both arteries and veins simultaneously and the uncertain labels were omitted for arteries and veins and left for the vessels.

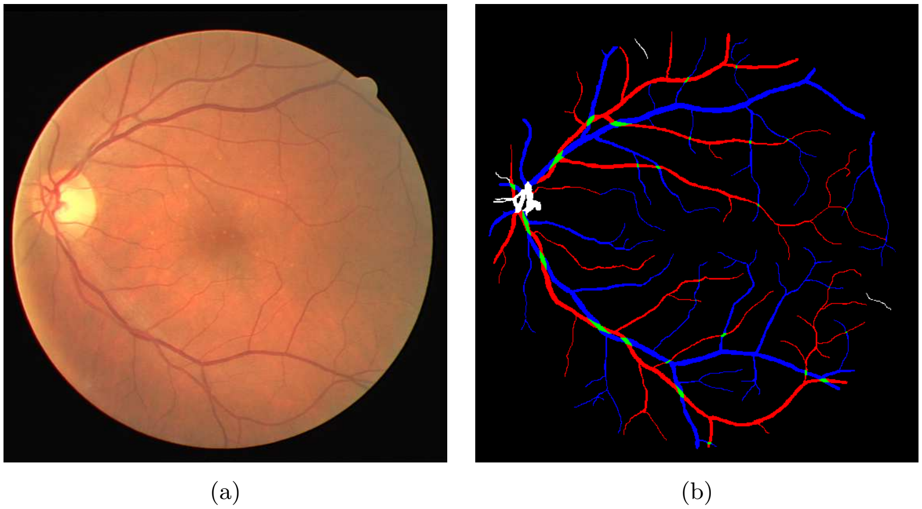


Fig. 1: (a) Retinal image from the DRIVE data set. (b) Retinal image labels from RITE dataset.

3 Bayesian AV classification

3.1 Baseline

Garifullin et al. followed a standard approach for deep Bayesian classification. First, a neural network f is used to estimate the distribution of logits parametrized through the estimate of the mean $\hat{\mathbf{y}}$ and variance $\boldsymbol{\sigma}$ of logits for arteries and veins:

$$[\hat{\mathbf{y}}, \boldsymbol{\sigma}] = f\left(\mathbf{x}, \boldsymbol{\theta}\right). \tag{1}$$

The probability vector $\mathbf{p} = [p_{\text{artery}}\ p_{\text{vein}}]$ of the labels can then be calculated as follows:

$$\hat{\mathbf{p}} = \operatorname{sigmoid}\left(\hat{\mathbf{y}} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}\right), \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right). \tag{2}$$

Given the probability vector for arteries and veins the probability for the vessels can be inferred based on the addition law of probability:

$$p_{\text{vessel}} = p_{\text{artery}} + p_{\text{vein}} - p_{\text{artery}} p_{\text{vein}}. \tag{3}$$

The resulting optimisation objective is a sum of binary cross-entropy functions for all three labels over all produced aleatoric samples.

The formulae (1) – (3) take into account heteroscedastic aleatoric uncertainty which is a type of uncertainty dependent on the data capturing imperfect imaging conditions, labeling and image noise. The second kind of uncertainty is epistemic uncertainty representing the model's ignorance. By considering the parameters of the model as a random variable with the posterior $p\left(\boldsymbol{\theta} \mid \mathcal{D}\right)$ the posterior predictive distribution over logits can be calculated as follows:

$$p\left(\mathbf{y} \mid \mathbf{x}, \mathcal{D}\right) = \int p\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) \mathrm{d}\boldsymbol{\theta}. \tag{4}$$

Typically, the integral (4) is intractable and stochastic approximations are used in order to estimate the posterior predictive. One of the most common techniques is to use stochastic variational approximation called MC-Dropout [1] which employs dropout as a Monte Carlo sampling technique in order to obtain samples from the model's posterior. Another widely used method is stochastic weight averaging Gaussian [11] where the model's posterior is approximated by a normal distribution the moments of which are estimated during the training procedure.

3.2 Stochastic weight averaging

Izmailov et al. found out that the values traversed by SGD would be around the flat regions of the loss surface, without actually reaching the center of this area [9]. By equally averaging these points traversed by SGD, Izmailov et al. found out that points that are inside this more desirable part of the loss surface would be achieved. They named this method stochastic weight averaging (SWA) and it was shown to improve the results and generalization of networks on a variety of architectures and in multiple applications. Given initial pre-trained weights SWA can be implemented as a running average of the weights calculated while continuing training with an additional computation of batch normalization statistics after (see [9] for more details).

3.3 Stochastic weight averaging Gaussian

SWAG was first introduced by Maddox et al. [11] for model averaging and uncertainty estimation. The main idea behind is to use SWA to calculate the mean of the model's parameters and at the same time to estimate a diagonal approximation of the covariance matrix. Thus, the approximated posterior of the model's parameters is a normal distribution:

$$p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) = \mathcal{N}\left(\boldsymbol{\theta}_{\text{SWA}}, \boldsymbol{\Sigma}_{\text{SWAG}}\right), \tag{5}$$

where $\boldsymbol{\theta}_{\text{SWA}}$ is a parameter vector estimated with SWA and $\boldsymbol{\Sigma}_{\text{SWAG}}$ is a corresponding diagonal covariance matrix.

4 Experiments and results

4.1 Description of experiments

The parameters and methodologies presented here were selected so that the baseline model used in this work would be as similar as possible to [2]. The utilized architecture is Dense-FCN-103 [10]. The baseline model was, however, re-implemented and the experiments reproduced to some degree in this work.

In all the experiments, the network was first pre-trained on RITE dataset with random patches of the input images of size 224 x 224. The batch size used in the pre-training was 5 and the network was pre-trained with 100 epochs and 1000 steps per epoch.

After the pre-training, the networks were fine-tuned with full-size images that were padded to size of 608 x 608 so that they could be properly compressed by the downsampling part of the network. The main optimizer used in all of the experiments was Adadelta with learning rate of 1 and decay rate of 0.95. The use of either SWA or SWAG would start on a later epochs of full resolution training.

To increase the diversity of the data set data augmentation techniques were used. The augmentation was performed by applying rotation, flipping, and scaling to the input data. The rotation angles used were 90, 180 and 270 degrees and the scaling rates were 0.8, 0.9, 1.0, 1.1 and 1.2.

The aleatoric and epistemic uncertainties were estimated using formulae from [8]. The uncertainties are estimated as an average sum standard deviations per image $S_p = \sum_i \sum_j \sigma_j / N_{\text{test}}$, where i is an index of the image, j is an index of the pixel, and N_{test} is the total number of test images (Table 4).

Baseline The fine-tuning of the network used as baseline was done using 50 epochs with 500 steps per epoch to match the hyperparameters used in [2]. The batch size used in the fine-tuning of the baseline was selected to be 1. MC-Dropout was used to quantify epistemic uncertainty.

SWA The SWA implementation also had 50 epochs with 500 steps in each epoch in the full resolution training. Like in the baseline the batch size used was 1. The starting epoch for SWA was selected to be 10 and it was only used in the fine-tuning of the network. The starting epoch was selected through empirical experimentation. MC-Dropout was used to quantify epistemic uncertainty.

SWAG The hyperparameters used in the SWAG implementation were 500 epochs with 50 steps per epoch. This was done so that the Gaussian posteriori approximation formed by SWAG would be generated from a higher number of epochs. Like in the baseline the batch size used was 1. The SWAG starting epoch was selected to be 100. The epistemic uncertainty was quantified by sampling the model's parameters from Gaussian distribution (5). Whereas the sampling is performed from the posterior estimated with SWAG, dropout is still used during the training phase.

4.2 Performance of the networks

Due to the fact that artery-vein classification was considered a multilabel problem, the performance metrics used in were calculated for arteries, veins and vessels separately. The selected classification metrics were accuracy, sensitivity, specificity, Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and Estimated Calibration Error (ECE) [4].

Table 1: Network performance in artery classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.970	0.642	0.990	0.00988	0.974
SWA	**0.975**	0.690	0.992	0.00943	0.981
SWAG	0.973	0.706	0.989	**0.00871**	0.966

Table 2: Network performance in vein classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.971	0.655	0.994	0.0169	0.980
SWA	**0.974**	0.742	0.991	0.0120	0.991
SWAG	0.971	0.804	0.983	**0.0107**	0.980

Table 3: Network performance in vessel classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.957	0.723	0.989	0.0221	0.980
SWA	**0.961**	0.782	0.986	**0.0208**	0.983
SWAG	**0.961**	0.836	0.978	0.0338	0.984

By examining the performance metrics presented in Tables 1–3, it can be seen that SWA improved the network performance overall compared to the baseline and SWAG models including the model calibration.

The example of the segmentation results for SWAG is given in Fig. 2. The segmentation examples for the baseline and SWA look similar. The uncertainties of the results were visualized and example figures can be seen in Fig. 3. In the figure, the intensities of the colors describe the uncertainty in that region as standard deviations of the predicted probabilities: the higher intensity the higher the uncertainty.

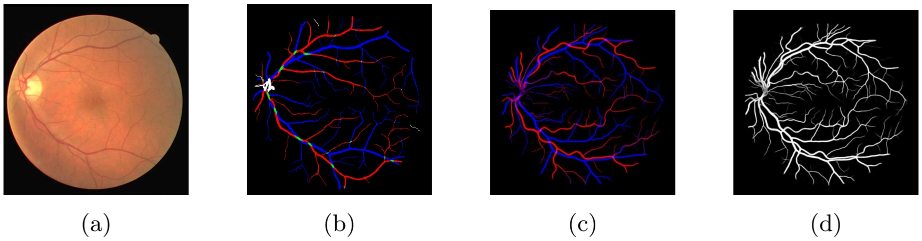


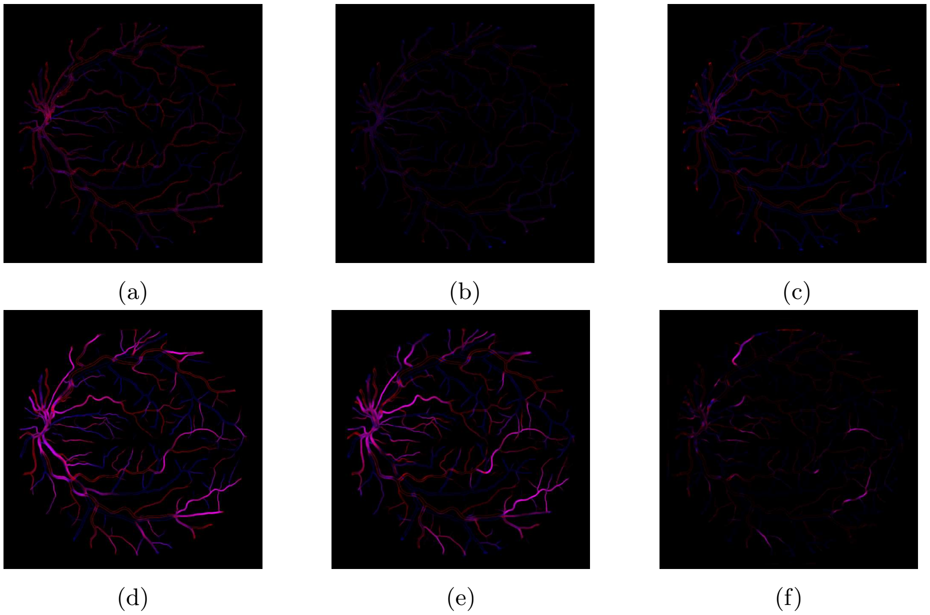
Fig. 2: (a) The input image; (b) ground truth; (c) mean predicted AV probabilities; (d) mean predicted vessels probabilities. The results are obtained using SWAG.

From the tables and figures, it can be concluded that the aleatoric uncertainty of the baseline is much higher than those of SWA and SWAG. It can also be concluded that sampling the network weights from the Gaussian posterior generated by SWAG to create the variational approximation, rather than using Monte-Carlo dropout, has a reducing effect on the levels of epistemic uncertainty present in the predictions. This could probably be explained by the fact that the variance is estimated only around a local optimum during the late stages of the training, whereas MC-Dropout is enabled during the whole training process. From the estimated performance metrics, however, it is difficult to conclude whether it is a positive or negative effect. One noticeable pattern is the high epistemic uncertainty near the optic disc when estimated with MC-Dropout. On the other hand, sampling from Gaussian distribution leads to the high uncertainties mostly near the end points of the blood vessels and the areas after the crossings which is also present in the case of MC-Dropout.

At the same time one can see that aleatoric uncertainties change when SWA or SWAG are utilized. Kendall et al. [1] describe the aleatoric uncertainty as a loss attenuation mechanism allowing the model to adapt the loss dependent on the data and labelling. While the aleatoric uncertainty is meant to be data dependent, the changes to the training procedure affecting the model's convergence and the parameters of the layers predicting variances also affect the predicted aleatoric uncertainties. For the baseline and SWAG, we can see a similar pattern of the higher aleatoric uncertainty levels near the optic disc and borders of the vasculature, whereas the aleatoric uncertainties almost vanish when estimated using MC-Dropout trained with SWA.

Table 4: Mean sums of estimated aleatoric and epistemic uncertainties per image.

Method	Aleatoric			Epistemic		
	Arteries	Veins	Vessels	Arteries	Veins	Vessels
Baseline	1276.2	1159.5	1807.5	4853.6	4066.4	5069.7
SWA	3.3	3.5	5.3	4038.6	3882.3	4659.7
SWAG	31.1	38.9	57.3	997.8	1104.3	1396.1


(a) (b) (c)

(d) (e) (f)

Fig. 3: Aleatoric uncertainties calculated using (a) the baseline, (b) SWA, and (c) SWAG. Epistemic uncertainties calculated using (d) the baseline, (e) SWA, and (f) SWAG.

4.3 Conclusions

In this work, the focus was on blood vessel segmentation from retinal images and on artery-vein classification by using a deep neural network. More specifically, two algorithms were studied to improve the classification performance and help in the model calibration. SWA and SWAG algorithms were implemented on top of the baseline and experimented with the DRIVE and RITE data sets.

The use of SWA improved the performance of the deep neural network on most of the binary classifications as well as the calibration metrics. SWAG showed slight improvements in the vessels and artery classification tasks. The weight averaging as a process significantly affecting the model's convergence seems to lead to diminishing aleatoric uncertainties and sampling from the normal distribution captures less epistemic uncertainty.

References

1. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)
2. Garifullin, A., Lensu, L., Uusitalo, H.: On the uncertainty of retinal artery-vein classification with dense fully-convolutional neural networks. In: Advanced Concepts for Intelligent Vision Systems: 20th International Conference, ACIVS. LNCS, Springer International Publishing, Auckland, New Zealand, Feb 10-14 (2020)
3. Girard, F., Kavalec, C., Cheriet, F.: Joint segmentation and classification of retinal arteries/veins from fundus images. Artificial Intelligence in Medicine **94**, 96 – 109 (2019). https://doi.org/https://doi.org/10.1016/j.artmed.2019.02.004
4. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330 (2017)
5. Hemelings, R., Elen, B., Stalmans, I., Van Keer, K., De Boever, P., Blaschko, M.B.: Artery-vein segmentation in fundus images using a fully convolutional network. Computerized Medical Imaging and Graphics (2019)
6. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging **19**(3), 203–210 (2000)
7. Hu, Q., Abràmoff, M.D., Garvin, M.K.: Automated separation of binary overlapping trees in low-contrast color retinal images. In: International conference on medical image computing and computer-assisted intervention. pp. 436–443. Springer (2013)
8. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 137–145. Springer (2019)
9. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D.P., Wilson, A.G.: Averaging Weights Leads to Wider Optima and Better Generalization. In: The Conference on Uncertainty in Artificial Intelligence (2018)
10. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. pp. 1175–1183. IEEE (2017)
11. Maddox, W., Garipov, T., Izmailov, P., Vetrov, D.P., Wilson, A.G.: A simple baseline for bayesian uncertainty in deep learning. In: NeurIPS (2019)
12. Miri, M., Amini, Z., Rabbani, H., Kafieh, R.: A comprehensive study of retinal vessel classification methods in fundus images. Journal of medical signals and sensors **7**(2), 59 (2017)
13. Welikala, R., Foster, P., Whincup, P., Rudnicka, A., Owen, C., Strachan, D., Barman, S.: Automated arteriole and venule classification using deep learning for retinal images from the uk biobank cohort. Computers in Biology and Medicine **90**, 23 – 32 (2017). https://doi.org/https://doi.org/10.1016/j.compbiomed.2017.09.005
14. Zhang, S., Zheng, R., Luo, Y., Wang, X., Mao, J., Roberts, C.J., Sun, M.: Simultaneous arteriole and venule segmentation of dual-modal fundus images using a multi-task cascade network. IEEE Access **7**, 57561–57573 (2019). https://doi.org/10.1109/ACCESS.2019.2914319