This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images

Mark J. J. P. van Grinsven*, Bram van Ginneken, Carel B. Hoyng, Thomas Theelen, Clara I. Sánchez

***Abstract*—Convolutional neural networks (CNNs) are deep learning network architectures that have pushed forward the state-of-the-art in a range of computer vision applications and are increasingly popular in medical image analysis. However, training of CNNs is time-consuming and challenging. In medical image analysis tasks, the majority of training examples are easy to classify and therefore contribute little to the CNN learning process. In this paper, we propose a method to improve and speed-up the CNN training for medical image analysis tasks by dynamically selecting misclassified negative samples during training. Training samples are heuristically sampled based on classification by the current status of the CNN. Weights are assigned to the training samples and informative samples are more likely to be included in the next CNN training iteration. We evaluated and compared our proposed method by training a CNN with (SeS) and without (NSeS) the selective sampling method. We focus on the detection of hemorrhages in color fundus images. A decreased training time from 170 epochs to 60 epochs with an increased performance – on par with two human experts – was achieved with areas under the receiver operating characteristics curve of 0.894 and 0.972 on two data sets. The SeS CNN statistically outperformed the NSeS CNN on an independent test set.**

***Index Terms*—Convolutional neural network, deep learning, hemorrhage, selective sampling**

I. Introduction

CONVOLUTIONAL neural networks (CNNs) have been widely adopted in the field of computer vision [1, 2]. These models are based on convolution operations applied to the input image at multiple hierarchical layers. CNNs are very powerful because they can be trained end-to-end in a supervised manner and thus obviate the need to manually devise features, and have substantially outperformed the state-of-the-art for classification of natural images on large and well established databases [3–5]. In medical image analysis, CNNs are also increasingly used. Their capability to learn a complex, hierarchical representation of the data makes CNNs useful to discern the complex disease specific patterns, difficult to be encoded by humans and by simpler traditional classifiers. Recent works on cancer detection and brain segmentation have shown CNN achieved remarkable performance [6–8]. However, the need of large high-quality training sets to accurately train CNNs prevent a wider adoption of these networks in medical imaging.

CNN training process is a sequential process requiring many iterations (or epochs) to optimize the network parameters and learn discriminative features [2]. In every epoch, a subset of samples is randomly selected from the training data and is presented to the network to update its parameters through back-propagation, minimizing a cost function. In this work we focus on finding diseased regions in images, a common task in medical image analysis. In such a classification task, CNNs are trained with small patches centered on pixels of interest. Although this results in vast training sets of image patches, the quality of the data is suboptimal: the normal class is extremely over-represented in this classification task and, moreover, the majority of normal training samples are highly correlated due to the repetitive pattern of normal tissues in each image. Only a small fraction of these samples are informative. Treating uniformly this data during the learning process leads to many training iterations wasted on non-informative samples, making the CNN training process unnecessarily time-consuming. An approach to identify informative normal samples will help to increase the efficiency of the CNN learning process and to reduce the training time.

Boosting techniques have been previously proposed to focus the learning process on informative samples in order to increase the performance of simple classifiers [9]. These techniques create an ensemble of learners, each trained consecutively, where more emphasis is put on samples misclassified by the previous learners [9, 10]. Classification is performed by combining the outputs of each of the individual learners. A simplified version of the boosting strategy is a two-step approach in which misclassified samples of an initial model are used as the training set of a second, independent learner [6, 11]. The second learner, which is trained then with only informative samples, is used for final classification. In general, boosting strategies rely on the optimization of different classifiers in cascade in order to discover informative samples (i.e. misclassified samples) for the next learner. Considering the high computational expense of CNN optimization, a boosted cascade of CNNs is inefficient, increasing the time complexity with the number of CNNs in the ensemble. In contrast to boosting techniques, dynamically sampling strategies focus

*Mark J. J. P. van Grinsven is with the Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands, e-mail: Mark.vanGrinsven@radboudumc.nl.

Bram van Ginneken is with the Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands.

Carel B. Hoyng and Thomas Theelen are with the Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands.

Clara I. Sánchez is with the Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine and with the Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

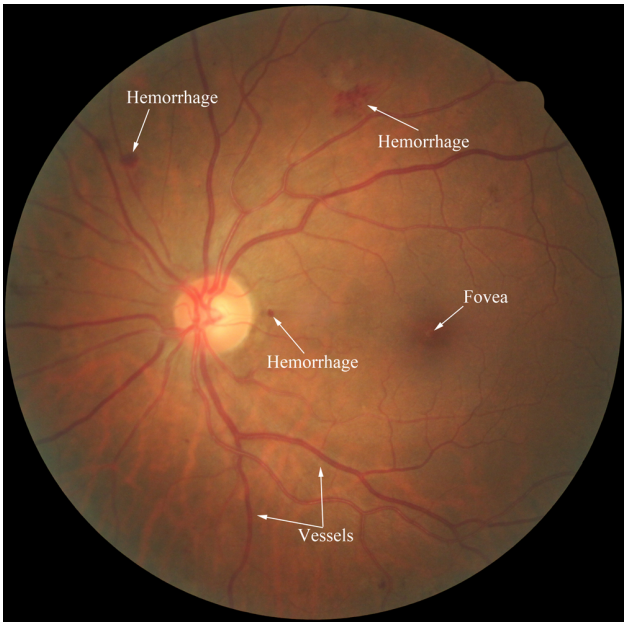


Fig. 1. Example of a color fundus image showing presence of hemorrhages.

the learner on informative samples during its optimization process, in order to simultaneously increase the classification performance and reduce training time. To achieve this, the training set is dynamically updated during the learning process of a single learner, putting more emphasis on informative samples [12–14]. These dynamic sampling strategies have shown to reduce the training time and outperform boosting types of strategies [13]. However, the challenge of these sampling techniques is defining a sampling heuristic optimal for the learner and the characteristics of the data and task at hand. To the best of our knowledge, the incorporation of a dynamic sampling strategy in the CNN learning process for medical image tasks has not been proposed yet.

In this paper, we propose an innovative sampling heuristic to identify informative training samples in a common medical image classification task, namely abnormality detection. The proposed heuristic will dynamically increase the probability of misclassified normal samples to be selected in each training iteration. We integrate this heuristic in the CNN learning process in order to increase its efficiency and reduce its training time, while maintaining its performance. The performance of the proposed method is then validated in two large datasets for the detection and localization of hemorrhages on color fundus images. Hemorrhages are one of the visible signs on color fundus images of diabetic retinopathy (DRP), a vision threatening disease affecting patients with diabetes [15]. Figure 1 shows an example of a color fundus image including hemorrhages and typical confounding elements in hemorrhage classification.

Hemorrhage detection is of high importance for the automated detection and staging of DRP, the most important cause of blindness in the working population. Whereas a lot of methods have been presented for the automated detection of micro-aneurysms in color fundus photographs, detection and segmentation of larger hemorrhages has received less attention [16, 17]. Hemorrhages and micro-aneurysms are mostly detected together and associated with a single label. In previous works, approaches based on morphological operations [18], wavelet operations [19] and manual designed features in combination with statistical classifiers [20–23] were used for the detection of hemorrhages and micro-aneurysms. Although hemorrhages are different in size and shape and pose different clinical relevance [15], only few works have addressed the identification of hemorrhages separately on color fundus images [24, 25].

Section II provides a description of the different data sets used in this work. The proposed method and experimental design are described in detail in Section III and Section IV. In Section V, the results are shown which are discussed in Section VI. Section VII concludes the proposed work.

II. Materials

Two independent data sets were used in this study for training and evaluating the proposed method. 1) a subset of images from the "Diabetic Retinopathy Detection" challenge database from Kaggle[1] and 2) images from the publicly available Messidor database[2].

A. Dataset description

1) Kaggle database: The Kaggle data set consist of 35,126 training images graded into five DRP stages and 53,576 test images with undisclosed DRP stage. Images were acquired using multiple fundus cameras and different field of view. Details about image acquisition, such as camera type and field of view, are not revealed. More information about the data can be found in the challenge website.

A subset consisting of 6,679 images was selected from the Kaggle training set. This subset consists of 4,450 randomly selected images from DRP stage 0 (normal), 488 randomly selected images from DRP stage 1 (mild), 1,058 randomly selected images from DRP stage 2 (moderate) and 593 randomly selected images from DRP stage 3 (severe). Images on which the retina was not visible were not included in this study dataset.

The selected 6,679 images were further split into a fixed training, monitoring and test set according to a 60-20-20 split. Images from the same patient were kept in the same subset.

2) Messidor database: The publicly available Messidor database consists of 1200 images acquired at three different sites. Images were acquired using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view. The images have resolutions of 1440x960, 2240x1488 or 2304x1536 pixels. More details about the database can be found in the corresponding website. The Messidor set will be exclusively used as an independent set for testing.

[1]https://www.kaggle.com/c/diabetic-retinopathy-detection

[2]Kindly provided by the Messidor program partners (see http://messidor.crihan.fr)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

TABLE I
REFERENCE ANNOTATION STATISTICS. PLUS AND MINUS SIGNS INDICATE THE NUMBER OF POSITIVE AND NEGATIVE IMAGES, RESPECTIVELY. THE NUMBERS BETWEEN BRACKETS INDICATE THE NUMBER OF GOOD QUALITY IMAGES AND NUMBER OF HEMORRHAGES ON GOOD QUALITY IMAGES.

	Training stage		Test stage	
	Training	Monitoring	Kaggle	Messidor
+	655	224	288 (197)	321 (289)
-	3304	1104	1104 (593)	879 (813)
Lesions	3290	1038	1095 (818)	

B. Reference standard and observer annotations

In this study, annotations were performed by three different independent observers, having 5 years, 12 years and over 15 years of experience, respectively. The first observer annotated and graded training, monitoring and test data. We referred to this observer as the reference observer. The two other observers (referred to as Observer 1 and Observer 2) graded only the test sets. These two observers were used to report human performance on the test data.

The reference observer indicated presence of hemorrhages on both the Kaggle and Messidor set. In the Kaggle set, this observer also annotated the center point of each individual hemorrhage in the training, monitoring and test sets. Furthermore, the reference observer indicated good or poor quality for each of the test images in both sets. An overview of the reference set can be seen in Table I. No individual hemorrhage lesion annotations were performed in the Messidor set.

III. METHODS

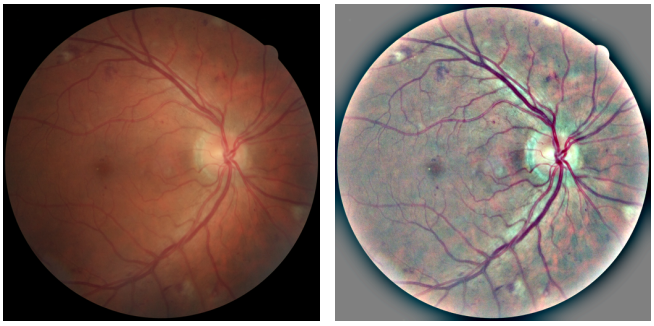
A dynamic CNN training strategy is presented where informative normal samples are dynamically selected at each training epoch from a large pool of medical images. A dynamic weight is assigned to each pixel in the negative training pool indicating its informativeness level. After each CNN training epoch, the weight of each negative training pixel is updated. This process is repeated until a stopping criterion is reached. The final trained CNN is used to classify each pixel in the test images, resulting in a pixel probability map for each test image.

A. Preprocessing

In a preprocessing step, the field of view of the color fundus images is segmented to limit the analysis of the CNN to the region of interest. Circular template matching is used to extract the field of view and images are cropped to the square bounding box of this circular field of view [26].Images are resized to 512 x 512 pixel dimension to reduce the computational costs and preprocessing was applied to improve image contrast [27, 28]. A contrast enhanced image $I_{ce}(x, y; \sigma)$ is obtained as follows [29]:

$$I_{ce}(x, y; \sigma) = \alpha I(x, y) + \beta G(x, y; \sigma) * I(x, y) + \gamma \quad (1)$$

where $*$ represents the convolution operator and $G(x, y; \sigma)$ a Gaussian filter with scale σ. Values of the parameters were empirically chosen as: $\alpha = 4$, $\beta = -4$, $\sigma = 512/30$ and $\gamma = 128$. The contrast enhanced image values are used as input for the CNN. Figure 2 shows an example image before and after applying the contrast enhancement step.



(a) (b)

Fig. 2. Contrast enhancement preprocessing step. (a) Original color fundus image I. (b) Contrast enhanced image I_{ce}.

B. Training data preparation and augmentation

Images which do not contain any hemorrhage are defined as negative images, whereas images with hemorrhages are defined as positive images. To construct the CNN training data, pixels are extracted from these images, where negative pixels are extracted only from negative images and positive pixels are extracted only from positive images at hemorrhage locations. Corresponding training patches, centered on the extracted pixels, of size 41x41 pixels and 3 channels depth are created during the CNN training routine. The patch label is determined by the label of the central pixel. Data augmentation by spatial translation with one pixel in both horizontal and vertical direction and vertical and horizontal flipping is applied to the positive patches to artificially increase the number of positives. Negative patches were also randomly flipped vertically and horizontally to counter for possible over-fitting.

C. Network details

The CNN architecture used in this study consists of five convolutional layers followed by rectified linear units (ReLUs) [4] and spatial max-pooling. The final layers of the network consist of a fully connected layer and a final soft-max classification layer. Inspired by the OxfordNet [30] which showed good performance for the classification of images of natural scenes, we use 32 small size filters of size 3x3 pixels in each convolutional layer. Max-pooling of size 2x2 and a stride of 2 is applied after the first two convolutional layers, halving the feature map sizes after the operations. Max-pooling reduces the number of free parameters and introduces small spatial invariance in the network [31]. The fully connected layer consist of 1024 nodes followed by a soft-max logistic regression which outputs a score ranging between 0 and 1, indicating the probability of the pixel to belong to the positive class. Weight-decay of $5 \cdot 10^{-5}$ is added to each layer to penalize large weight parameters during back-propagation of the gradient in the optimization routine. Table II and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

Figure 3 show an overview of the network architecture with the omission of the ReLUs. All network parameters are randomly initialized according to a normal distribution with variance equal to 0.05. The CNN is trained using stochastic gradient descent with learning rate of $5 \cdot 10^{-5}$, minimizing a cost function C defined as follows:

$$C(l,s) = -\sum_{i=0}^{B} l_i log(s_i) + (1-l_i) log(1-s_i) \qquad (2)$$

where s is the assigned pixel probability score, l the reference pixel label and B the total number of samples in one mini-batch. A mini-batch size of 256 patches is used and one epoch is defined as 4000 mini-batches. This means that around one million samples, of which half are positive and half are negative, are used in one epoch to train the CNN.

TABLE II
ARCHITECTURE OF THE CNN. FOR EACH CONVOLUTIONAL LAYER, THE *width x height x depth* OF THE KERNELS IS REPORTED WITH THE K NUMBER OF KERNELS. IN EACH MAX-POOLING LAYER, 2X2 MAX-POOLING IS APPLIED WITH STRIDE a PIXELS.

Layer	Operation	Input size	Details
Layer 1	convolution	41x41	3x3x3, $K = 32$
Layer 2	max-pooling	39x39	2x2, $a = 2$
Layer 3	convolution	20x20	3x3x32, $K = 32$
Layer 4	max-pooling	18x18	2x2, $a = 2$
Layer 5	convolution	9x9	3x3x32, $K = 32$
Layer 6	convolution	7x7	3x3x32, $K = 32$
Layer 7	convolution	5x5	3x3x32, $K = 32$
Layer 8	fully connected	3x3	1024 nodes
Layer 9	soft-max	1024x1	2 classes

D. Selective sampling

At each CNN training epoch, a weight is assigned to each negative sample, proportional to their sampling probability: higher weight means a higher probability to be selected for the next epoch. In order to reduce the number of redundant samples in the training set, higher weights are assigned to representative samples. In this work, representative sample are considered those negative samples with a larger classification error at the current CNN state.

Given $\mathcal{X} = \{(\mathbf{x}_i, l_i)\}$ the set of N training pixels $\mathbf{x}_i$ and their corresponding reference label l_i with $i = \{1, ..., N\}$, let $\mathcal{X}_+$ and $\mathcal{X}_-$ be the sets of positive and negative pixels:

$$\begin{aligned} \mathcal{X}_+ &= \{(\mathbf{x}_i, l_i),\ \forall x_i \text{ with } l_i = 1\} \\ \mathcal{X}_- &= \{(\mathbf{x}_i, l_i),\ \forall x_i \text{ with } l_i = 0\} \end{aligned} \qquad (3)$$

where $\mathcal{X} = \mathcal{X}_+ \cup \mathcal{X}_-$.

The proposed iterative algorithm for dynamically selecting training pixels to train a CNN c follows these steps:

1) Initialize the sets of positive pixels $X_+^t \subset \mathcal{X}_+$ and negative pixels $X_-^t \subset \mathcal{X}_-$ by randomly selecting M samples with replacement for each class from $\mathcal{X}_+$ and $\mathcal{X}_-$, respectively.
2) Train the network c with $X^t = X_+^t \cup X_-^t$ using stochastic gradient descent.
3) Classify each pixel x_i in $\mathcal{X}_-$ with the trained network c^t. A pixel probability score s_i^t is obtained for each x_i in $\mathcal{X}_-$ after classification.
4) Assign each x_i in $\mathcal{X}_-$ a weight $w_i^t = \mid s_i^t - l_i \mid$. A higher weight is assigned to those pixels of which the preliminary network prediction differs the most from the initial reference label.
5) Update X_+^t and X_-^t by selecting M samples for each class. x_i in X_+^t is selected randomly while x_i in X_-^t is selected with probability p_i^t [32–34]:

$$p_i^t = \frac{w_i^t}{\sum_{x_j \in \mathcal{X}_-} w_j^t} \qquad (4)$$

6) Train the network c with $X^t = X_+^t \cup X_-^t$ using stochastic gradient descent.
7) Repeat steps (3)-(6) until a stopping criterion is reached.

In this proposed iterative selective sampling (SeS) algorithm, the pool of negative and positive training pixels is dynamically changed at each training epoch, preventing the training process to focus on redundant negative samples and efficiently train the CNN. The parameter M is not tunable by itself but is dependent on the mini-batch size and the number of mini-batches in one epoch. Changing the value of M can be done by modifying either one of these two. In order to obtain a more efficient scheme, we consider applying step (3) and (4) once every five epochs.

E. CNN training monitoring

To determine when the CNN training process is completed and avoid over-fitting, the CNN performance during training is monitored during training on an independent monitoring set. Although the problem of over-fitting is countered by using different training pixels in each training epoch, an independent measure to determine when to stop the training procedure is still required. One way to measure performance is by measuring the pixel classification performance using the area (Az) under the Receiver Operating Characteristics (ROC) curve [35]. However, Az values of pixel-based ROC curves are misleading due to the unequal distribution of positive and negative pixels. Therefore, we measure the Az value based on image classification performance. A score for each image is obtained by classifying all pixels in the image and considering the maximum pixel probability as the image score. When the Az value on the monitoring set has reached a stable maximum, determined after visual inspection, the CNN training phase is considered finished.

F. Hemorrhage classification

Given an unseen test input image, the CNN classifies every pixel in the image and returns a probability map indicating for each pixel the probability to belong to a hemorrhage. We postprocess this probability map to extract hemorrhage candidates and compute an image score describing the probability of the image to contain hemorrhages.

1) Hemorrhage lesion identification: The obtained pixel probability map is convolved with a Gaussian filter with scale $\sigma = 1$ to smooth the values. Candidate hemorrhage regions are identified by detecting local maxima in the smoothed pixel probability map. The local maxima locations are used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

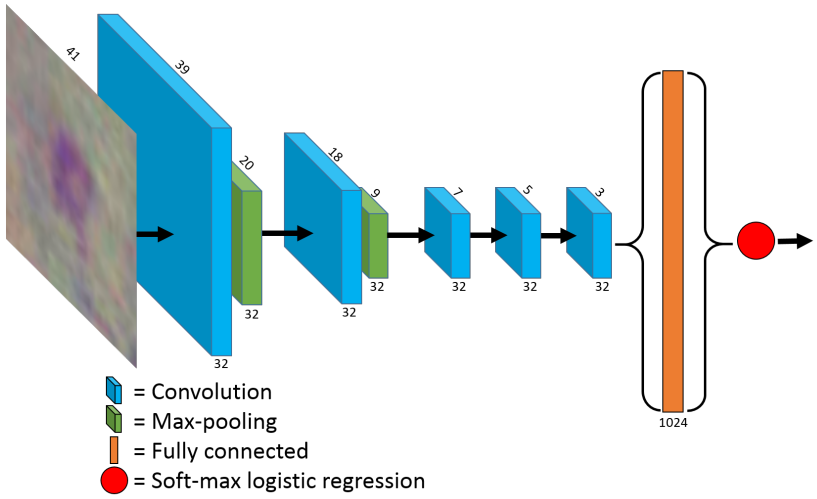


Fig. 3. Schematic overview of the CNN architecture containing convolutional layers, max-pooling layers, a fully connected layer and a soft-max logistic regression classification.

as seed points for dynamic programming to segment the individual hemorrhage candidates [36]. The dynamic programming algorithm is driven by a cost function computed as the gradient magnitude of the smoothed pixel probability map. The segmented candidate is assigned a final probability equal to the average of the pixel probabilities inside the candidate.

2) Identification of images with hemorrhages: To determine if an image contains hemorrhages, an image score is computed from the obtained pixel probability map. After the Gaussian smoothing step is applied to the pixel probability map, the maximum pixel probability is assigned as image score.

IV. Experimental design

To compare the performance of the proposed SeS algorithm, a second CNN with the same network architecture was trained using the same pool of training images. However, at each training epoch random sampling of training pixels was performed. This means that each pixel has an equal chance of being used in the training procedure. This non-selective sampling (NSeS) CNN was also monitored using the same monitoring data set and the same stopping criteria. To train and monitor both CNNs, the Kaggle training and monitoring set were used, respectively. After both CNNs were trained, results were computed on the two test sets.

We evaluated the proposed SeS scheme by conducting the following experiments:

1) Evaluation of CNN performance during training: The Az value on the monitoring set was measured during CNN training for both the SeS and NSeS CNNs and the required number of training epochs was compared.
2) Evaluation of hemorrhage lesion identification: Free-response ROC (FROC) analysis was employed to compare the CNN performance for the detection of individual hemorrhages [35]. In here, only false positives encountered on negative images were taken into account to prevent ambiguities in the reference annotation to influence the result [37, 38]. Additionally, there is no clinical relevance in a screening setting for false positives detected on images containing hemorrhages as these patients should be sent for referral.To determine if a hemorrhage was detected by the CNN, the distance between the manually annotated hemorrhage center location and the seed point of the segmented candidate was used, with a maximum tolerance of 8 pixels. This value was empirically determined using visual inspection of the average hemorrhage size on the 512x512 pixel resolution images. Detected regions which had no reference hemorrhage center location within this 8 pixel circular radius were considered as false positives.
3) Evaluation of identification of images with hemorrhages: ROC analysis was performed to evaluate the performance on identification of images with hemorrhages. Bootstrap analysis with 10,000 bootstraps was used to compute 95% confidence intervals for the Az values [39, 40]. A level of significance of $\alpha = 0.05$ was used for statistical comparison of the CNNs. Sensitivity, specificity and kappa agreement of the CNNs with respect to the reference were calculated at the operating point on the ROC curve closest to the upper left corner of the graph. These measures were also computed for Observer 1 and Observer 2.
4) Evaluation of the influence of image quality: In order to assess the influence of image quality on the CNN performance, images graded by the reference observer as having poor image quality were removed from both test sets. CNN performance on image level was measured using ROC analysis and sensitivity, specificity and kappa agreement values were calculated.

V. Results

A. CNN performance during training

During training, performance was measured on the monitoring set during the CNN training process. Figure 4 shows the Az values measured on image level as function of the number of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

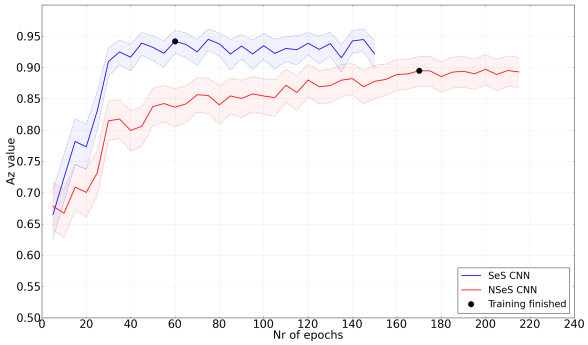


Fig. 4. Image-based Az values on the monitoring set for the SeS CNN and NSeS CNN over a number of training epochs. Shaded regions indicate the 95% confidence intervals of the Az values. After 60 and 170 epochs, the training phases of the SeS CNN and NSeS CNN were considered finished.

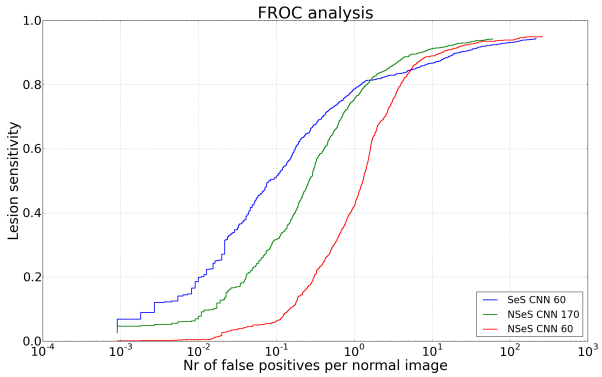


Fig. 7. FROC curves of the SeS CNN 60 and NSeS CNN 170 on the Kaggle test set. The FROC curve of the NSeS CNN after 60 epochs of training is added for direct comparison with the SeS CNN after 60 epochs of training.

training epochs. The performance of the CNNs increased over time and finally converged to a stable maximum performance. For the SeS CNN, this maximum performance was achieved after 60 training epochs and for the NSeS CNN this maximum was achieved after 170 training epochs. Both these networks, i.e. the SeS CNN after 60 epochs (SeS CNN 60) and the NSeS CNN after 170 epochs (NSeS CNN 170) were used to compute hemorrhage detection results on the two independent test sets.

Figure 5 shows example outputs of the SeS CNN and NSeS CNN after different numbers of training epochs as heat-map overlays on the example input image shown on top and in Figure 2a. After training for a small number of training epochs, both CNNs incorrectly classified all dark normal structures, such as vessels and fovea, but were able to correctly classify the normal background pixels. As CNN training continues, the CNNs learn to separate hemorrhages and the normal retinal structures are correctly classified as negative. For the SeS CNN, this learning process required less training epochs.

B. Hemorrhage lesion identification

Figure 7 shows the FROC curves for the SeS CNN 60 and the NSeS CNN 170. The NSeS CNN after 60 epochs is included for direct comparison with the SeS CNN 60, showing a lower overall FROC curve compared to the SeS CNN 60 and NSeS CNN 170. At 1 false positive per normal image the SeS CNN 60 and NSeS CNN 170 achieve sensitivities of 0.786 and 0.753, whereas at 0.1 false positives per normal image, both CNNs achieve sensitivities of 0.511 and 0.316, respectively. In Figure 6, example images with annotated hemorrhage center locations and the outputs of the SeS CNN 60 networks are shown.

C. Identification of images with hemorrhages

Figure 8 shows the ROC curves of both CNNs on the Kaggle and Messidor test set. For the NSeS CNN, the performance after 60 epochs was also calculated and shown in the graphs for direct comparison with the SeS CNN 60. Operating points of both human observers are added in the plots. On the Kaggle test set, there was no significant difference (p-value=0.509) between the SeS CNN 60 and NSeS CNN 170, whereas on the Messidor test set, the SeS CNN 60 significantly outperformed the NSeS CNN 170 (p-value=0.0028).

Table III shows the contingency table for the observer gradings, the SeS CNN 60 and the NSeS 170 CNN as compared with the reference. Kappa agreements (κ) with 95% confidence intervals and sensitivity/specificity pairs are included.

D. Influence of image quality

The number of poor quality images as indicated by the reference observer was 602 and 98 in the Kaggle and Messidor test sets, respectively. Table IV shows the contingency table after removing the mentioned poor quality images for the observer gradings, the SeS CNN 60 and the NSeS 170 CNN as compared with the reference on the two test sets. Kappa agreements (κ) with 95% confidence intervals and sensitivity/specificity pairs are included.

VI. Discussion

During the time-consuming training process of a CNN, the majority of samples that are presented to the network are easy to classify correctly. In this work we hypothesized that we can speed up the training process by selecting *difficult normal samples* to present to the network. We achieved this by classifying normal images with the current state of the network after a number of epochs of training, and select more patches from those regions in normal images that the network considered abnormal. More precisely, in the SeS method, dynamic weights based on the CNN's preliminary classification were computed for each training sample at selected snapshots during training. Samples with higher weights were more likely to be selected for training in the next epochs. Using this scheme, the training procedure was guided to learn from the more informative samples. We applied the proposed SeS strategy to the detection of hemorrhages on color fundus images to show the potential of this technique in an important medical image analysis application. The results showed that the CNN with SeS employed in the training procedure required a considerably smaller number of training epochs to achieve a high performance when compared to a CNN without selective sampling.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

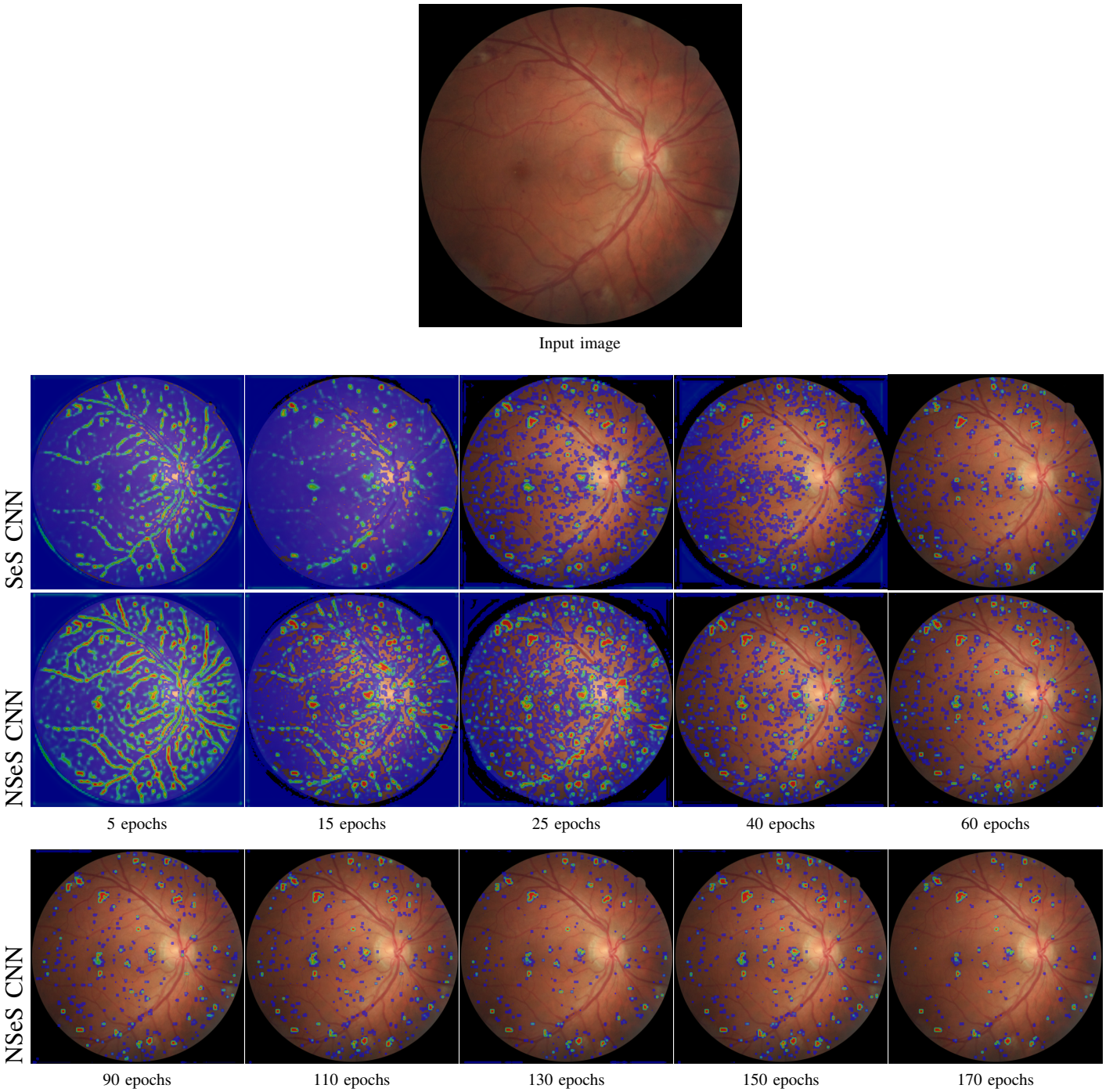


Fig. 5. Pixel probability maps obtained by applying the SeS and NSeS CNNs to a sample image from the training set after training the network for a different number of epochs. Overlays are shown using a heat-map color coding, where red codes for high probabilities and blue for low probabilities. The SeS CNN required 60 epochs to reach final performance while the NSeS CNN required 170 epochs to reach final performance.

The SeS CNN required 60 epochs for the training phase to obtain similar performance as the NSeS after 170 epochs of training, as illustrated in Figure 4. When training is conducted in an iterative approach, which is the case with CNNs, it is likely that the importance of training samples changes during this learning process. The ability of the SeS CNN to dynamically change the focus of the learning process attributed to the speed-up of the learning process, as training time is not wasted on samples which the networks has already "learned" to classify correctly. Figure 5 displays the evolution of the pixel probability maps when evaluated on one unseen example image. It can be observed that the SeS CNN learns to differentiate between background tissue (i.e. blood vessels, fovea and micro-aneurysms) and hemorrhages faster than the NSeS CNN. Although these structures are specific to the retina, a similar learning behavior can be expected on other data sets as the training procedure with SeS is guided by its own capability to classify these structures.

Performance for the identification of hemorrhages of the SeS CNN is higher on both test sets as compared to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

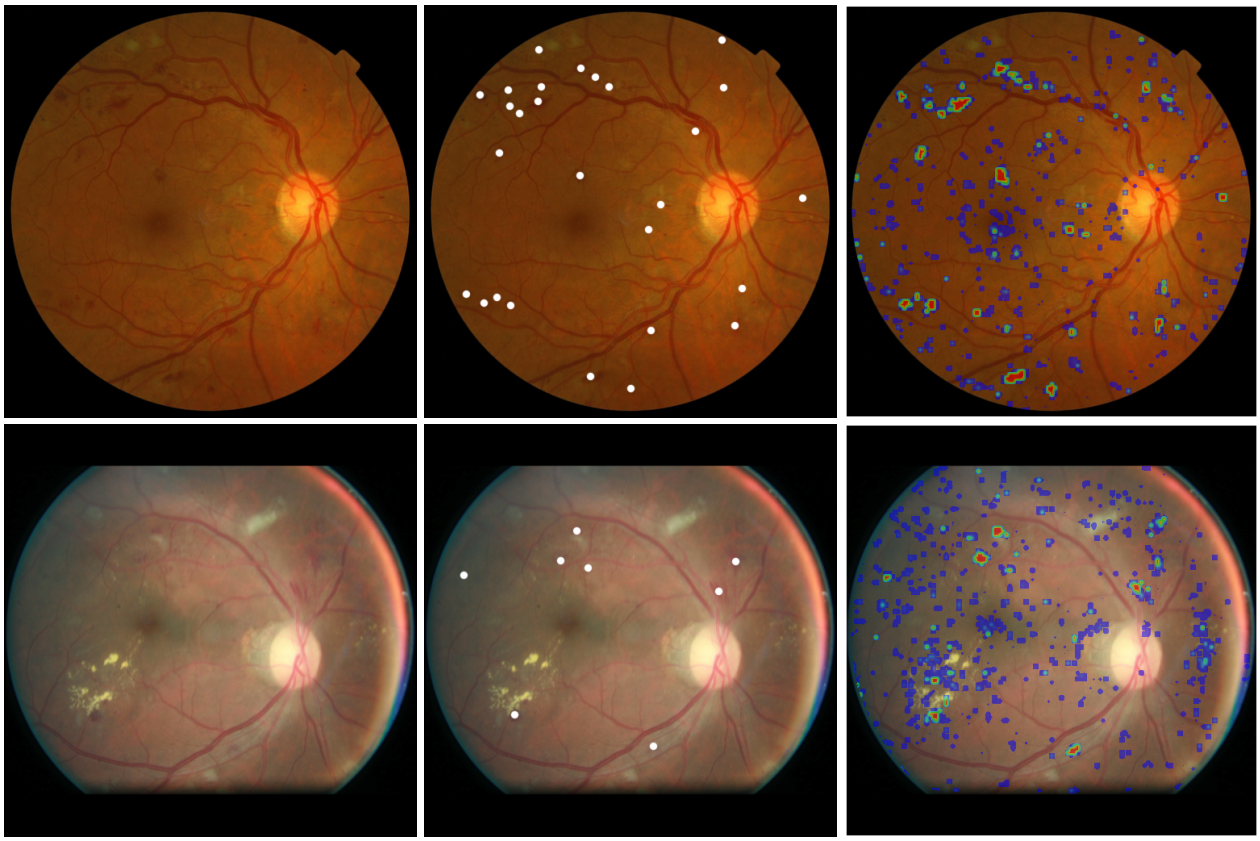
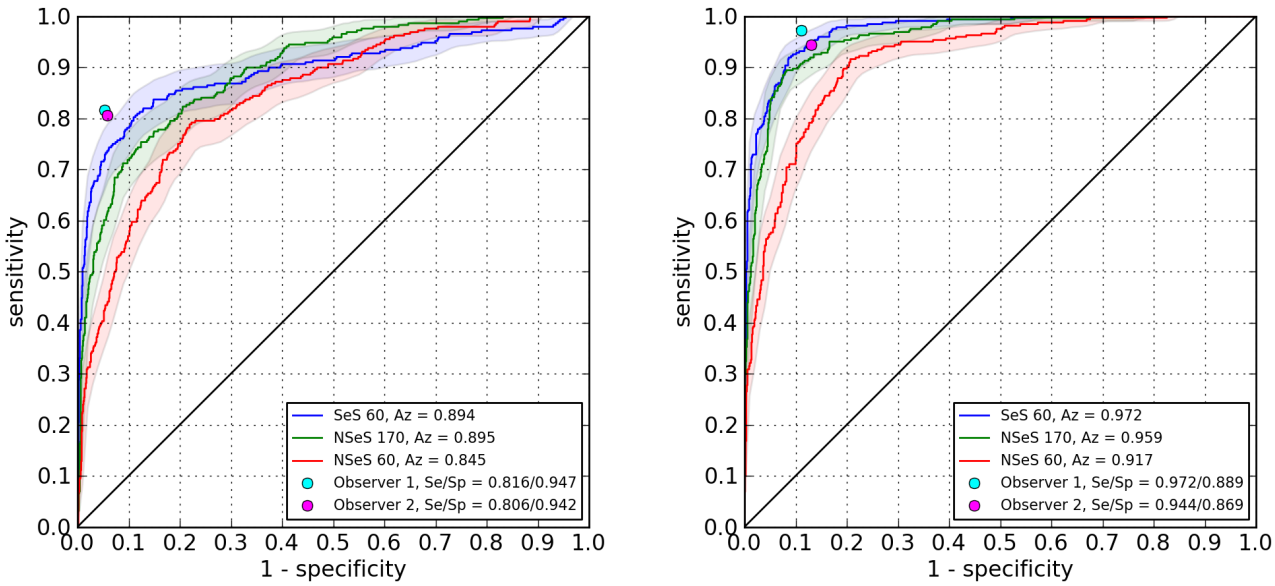


Fig. 6. Left column: example color fundus images from the Kaggle test set. Middle column: reference hemorrhage center locations. Right column: output of the SeS CNN 60.



(a) Kaggle (b) Messidor

Fig. 8. Image-based ROC curves on Kaggle (a) and Messidor (b) test sets. Observer operating points of the human observers are added in the graphs.

NSeS CNN, see Figure 8 and Figure 7 and comparable to human observer performance for the identification of images with hemorrhages. On the independent Messidor test set, this difference was statistically significant (p-value=0.0028). The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging

TABLE III
CONTINGENCY TABLE OF HUMAN OBSERVERS, SES CNN 60 AND NSES CNN 170 ON THE KAGGLE AND MESSIDOR TEST SETS. κ = KAPPA AGREEMENT WITH 95% CONFIDENCE INTERVALS, SE/SP = SENSITIVITY AND SPECIFICITY, AZ = AREA UNDER THE ROC WITH 95% CONFIDENCE INTERVALS.

		Observer 1		Observer 2		SeS CNN 60		NSeS CNN 170	
			-	+	-	+	-	+	-	+
Reference	Kaggle	-	1046	58	1040	64	940	164	939	165
		+	53	235	56	232	47	241	65	223
		κ = 0.759 [0.716-0.802]		κ = 0.740 [0.696-0.785]		κ = 0.598 [0.549-0.648]		κ = 0.554 [0.501-0.607]	
		Se/Sp = 0.816/0.947		Se/Sp = 0.806/0.942		Se/Sp = 0.837/0.851		Se/Sp = 0.774/0.851	
						Az = 0.894 [0.867-0.919]		Az = 0.895 [0.874-0.914]	
	Messidor		-	+	-	+	-	+	-	+
		-	781	98	764	115	803	76	807	72
		+	9	312	18	303	26	295	34	287
		κ = 0.791 [0.753-0.829]		κ = 0.742 [0.701-0.783]		κ = 0.793 [0.755-0.832]		κ = 0.783 [0.743-0.822]	
		Se/Sp = 0.972/0.889		Se/Sp = 0.944/0.869		Se/Sp = 0.919/0.914		Se/Sp = 0.894/0.918	
						Az = 0.972 [0.963-0.980]		Az = 0.959 [0.947-0.970]	

TABLE IV
CONTINGENCY TABLE OF HUMAN OBSERVERS, SES CNN 60 AND NSES CNN 170 ON THE KAGGLE AND MESSIDOR TEST SETS AFTER REMOVAL OF POOR QUALITY IMAGES. κ = KAPPA AGREEMENT WITH 95% CONFIDENCE INTERVALS, SE/SP = SENSITIVITY AND SPECIFICITY, AZ = AREA UNDER THE ROC WITH 95% CONFIDENCE INTERVALS.

		Observer 1		Observer 2		SeS CNN 60		NSeS CNN 170	
			-	+	-	+	-	+	-	+
Reference	Kaggle	-	568	25	563	30	536	57	486	107
		+	23	174	28	169	30	167	34	163
		κ = 0.838 [0.794-0.883]		κ = 0.805 [0.756-0.853]		κ = 0.719 [0.663-0.775]		κ = 0.576 [0.512-0.639]	
		Se/Sp = 0.883/0.958		Se/Sp = 0.858/0.949		Se/Sp = 0.848/0.904		Se/Sp = 0.827/0.820	
						Az = 0.917 [0.888-0.944]		Az = 0.909 [0.884-0.932]	
	Messidor		-	+	-	+	-	+	-	+
		-	727	86	709	104	744	69	757	56
		+	7	282	12	277	20	269	28	261
		κ = 0.800 [0.761-0.839]		κ = 0.753 [0.711-0.796]		κ = 0.802 [0.763-0.842]		κ = 0.809 [0.770-0.848]	
		Se/Sp = 0.976/0.894		Se/Sp = 0.958/0.872		Se/Sp = 0.931/0.915		Se/Sp = 0.903/0.931	
						Az = 0.979 [0.970-0.985]		Az = 0.966 [0.954-0.976]	

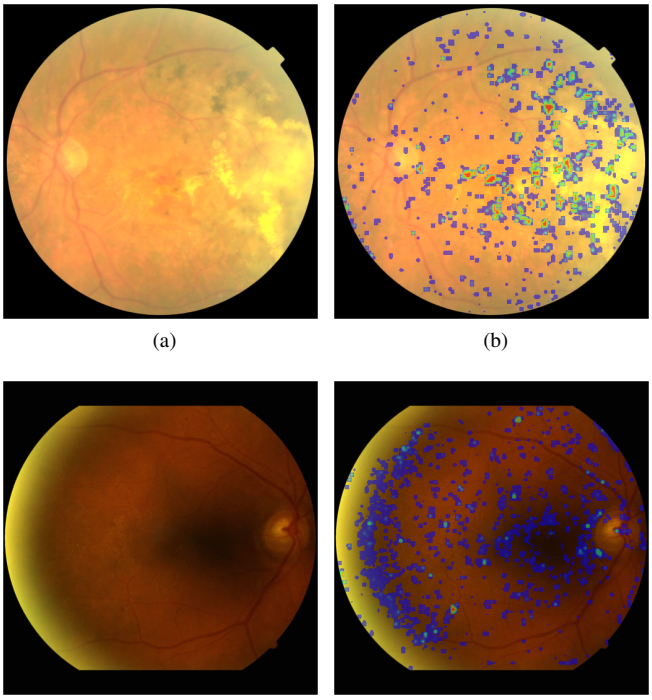
image scores for presence of hemorrhages were calculated based on the maximum posterior probability in each probability map. Images containing challenging confounding structures are therefore more prone to misclassification. As the SeS CNN was guided to learn these challenging structures, overall classification rates compared to the NSeS CNN increase. There is a difference in performance obtained by the CNNs on the Kaggle and Messidor test set. Performance on the independent Messidor test set is higher as compared to the one obtained on the Kaggle test set for both CNNs. An explanation for this can be the presence of other abnormalities and the quality of the images in both data sets, see Figure 9((a) and (b)).

Assessment of image quality showed that 602 images (43.2%) were graded as having poor image quality by the reference observer in the Kaggle dataset, whereas for the Messidor test 98 poor quality images (8%) were identified. Figure 10 shows two examples of images which were graded as having poor image quality. This is an indication that the overall image quality in the Messidor test set is better than the overall quality in the Kaggle test set, allowing the CNNs to achieve higher performance.

In this study, we applied the SeS strategy only to the training samples that belong to the negative class. There is no fundamental reason why the same strategy could not be applied to the positive class as well. In our case, the set of positive samples was limited in numbers and each sample was already presented multiple times per epoch. If we would increase the number of positive training samples, either by increasing the amount of available training images or by applying more data augmentation, and applying the SeS strategy to the positive class as well could potentially further increase the detection rates and speed up training. In this way, also difficult positive samples are presented more frequently during training. This would guide the CNN learning procedure to also better recognize the more difficult hemorrhage structures.

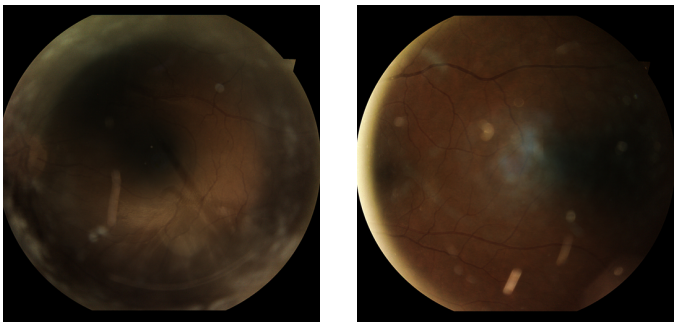
Although we obtained excellent performance for the detection of hemorrhages, an in-depth optimization of the network hyper-parameters was not performed in this study. This optimization is a challenging task [41]: The depth of the network, i.e. the number of layers, and the number of kernels per layer, as well as the use of fully connected layers should be more thoroughly investigated. Pilot experiments using MSRA weight initialization [3] showed an equal number of epochs required to train the SeS with similar end-performance on the test sets. Additions such as the inclusion of drop-out [42] or batch normalization [43] could potentially further increase

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2526689, IEEE Transactions on Medical Imaging



(a) (b) (c) (d)

Fig. 9. Examples of errors by the SeS CNN system. (a): example of a retinal image with different type of abnormality (bright and dark regions on the right side of the image), (b): output of the SeS CNN 60 computed on (a), (c): example or a retinal image with hemorrhages (bottom left and top middle) which was graded as negative by the reference, (d): output of the SeS CNN 60 computed on (c). See text in the discussion section for more details.



(a) (b)

Fig. 10. Examples of images graded by the reference observer as having poor image quality.

performance. Furthermore, a kernel size of 3x3 pixels was chosen. The rationale of using such a small size kernels is that each larger size kernel, e.g. a 5x5 kernel, can be represented by multiple smaller sized kernels, i.e. two times a 3x3 kernel sized layer. Using multiple smaller sized layers with non-linear rectifications makes the CNN more discriminative and less parameters need to be optimized [30]. The use of multi-scale patches could potentially be beneficial as usage of multi-scale patches has shown promising results in other applications [44]. Optimizing CNN hyper-parameters is challenging and trying many combinations is common practice [41]. However, it should be noted that the SeS CNN was compared to a NSeS CNN using the same network architecture. A similar improvement in training time and classification performance may be expected with a different network architecture.

Previous works have shown that adding more informative samples to the training set can improve the performance of the learner substantially [9, 10, 12–14, 45, 46]. In boosting techniques, an ensemble of learners is trained where each of the consecutive learners uses a fixed, more informative training set [9, 10]. Samples that are misclassified by the previous learners are typically added to the training set of the next learner. Application of this boosting approach to CNNs is highly time-consuming and inefficient as each of the learners is optimized independently and no information, such as network parameters, are shared between the learners. Taking into account the training process of a CNN is an iterative process, dynamically updating the training set in each iterations will avoid the use of multiple learners, focusing the attention of the learner on informative samples and optimizing the CNN parameters simultaneously. Other works have used a two-step approach in which representative samples are first identified by an initial learner. The first learner can be either the same learner [11], or a simplified, faster to train learner [6]. Using this approach, a new dataset is created which is used for training a second, independent learner. Apart from the fact that a cascade of two independent learners are still needed and, consequently, more extra training time, using a simplified learner to select informative training samples does not guarantee that these samples are also informative for the second, more complex learner.

Similar to our approach, a previous work has used a dynamic sampling approach to train a multi-layer perceptron (MLP) [13]. In each training epoch, each training sample was first classified by the current state MLP to assign a sampling weight and it was determined using a sampling heuristic if this sample should be included for training. In this case, the sampling heuristic was designed to include all misclassified samples and a selection of correctly classified samples based on the class balance in the training set and the confidence level of the current state MLP. However, applying this heuristic to CNN training for patch classification is not feasible. First, including all misclassified samples would lead to the over-fitting of the network as millions of patches, mainly normal, would be misclassified, especially in the first iterations. Additionally, positive samples are highly under-represented in medical images. Therefore, all the positive samples should be considered as informative and no prioritizing selection is needed. For that reason, a sampling heuristic specifically designed for abnormality identification using CNN in medical images was proposed in this work, where a selection of informative negative samples was performed in each iteration and all positive samples were randomly included.

A limitation of this study is the use of a manual reference as provided by a single human expert. As hemorrhages and micro-aneurysms are similar in characteristics and are only differentiable by their size and color on color fundus images, they can be easily confused [16]. Figure 9 (c) and (d) show a retinal image and the SeS CNN 60 output, respectively. The image which was graded as normal by the reference but both Observer 1 and Observer 2 indicated presence of hemorrhages.

Combining human observer annotations to create a consensus annotation might improve the reference, but prohibits a fair comparison with the performance obtained by one of these observers. Using an additional external reference such as fluorescein angiography, in which the contrast of blood (and therefore also hemorrhages) is enhanced by a contrast agent, might help to set a better reference standard [16]. As the data sets used in this study are retrospectively analyzed and only contain color fundus images, expert grading on color fundus images was the best strategy available to us. Furthermore, the reference observer only indicated the hemorrhage center locations. Therefore, no detailed analysis on the individual hemorrhage segmentation could be performed. This would however be of added value for clinical studies and more research is needed for a more thoroughly evaluation of this task.

After analyzing the Kaggle test set further, we noticed that the majority of the errors made by human observers were on images from DRP stages 1 and 2. Images from DRP stage 1 and 2 contain numerous confounding lesions, such as microaneurysms, which are very difficult to differentiate from hemorrhages and introduce a high inter-reader variability. When taking only the images from DRP stage 0 and 3 into account, the agreement of both observers with the reference was higher with κ values of 0.776 and 0.771. This indicated a more reliable grading could be made on the images from DRP stage 0 and 3. To investigate further the influence of a more reliable annotated training set, only image from DRP stage 0 and DRP stage 3 were used from the Kaggle dataset to train and evaluate the CNNs. The training time for the SeS CNN and NSeS was reduced to 40 and 140 epochs, respectively. The image level performance increased slightly to Az values of 0.919 and 0.907; and 0.981 and 0.967 for the SeS and NSeS CNNs on the Kaggle and Messidor test sets, respectively. In this subset, we have also investigated the influence of the color normalization preprocessing step on the CNN performance. Training the SeS CNN without color normalization took five epochs longer to converge but achieved the same performance as the SeS CNN using color normalization in both test sets. This demonstrates that CNN is capable to deal with the large variability of medical data but it requires more time to learn this variability during the training phase.

All experiments were performed on an Intel Xeon PC with 2.4Ghz memory and a GeForce GTX 570 video card. The training time per epoch was around 16 minutes for both the SeS CNN and NSeS CNN and classifying all pixels in one image, i.e. computing a probability map, took around 0.82 seconds using a sliding window approach [47]. In our current implementation, all the weights for the selective sampling were generated sequentially during one pass over the training set consisting of 3,959 images, i.e. this means a total time of 54 minutes for weight calculation. The SeS CNN required 60 epochs with 11 weight updates for the training process. The total time for the SeS CNN to complete the training phase was then $60 \cdot 16 + 11 \cdot 54 = 1554$ minutes, whereas for the NSeS, this was $170 \cdot 16 = 2720$ minutes. However, the training time for the SeS CNN can be reduced significantly by parallelizing the generation of weights and CNN training. By doing so, the weights for the training samples can be computed during CNN training and the total time for the SeS can be reduced to $60 \cdot 16 = 960$ minutes.

We have evaluated our proposed strategy using two different datasets in order to analyze the generalization of the method. However, more experiments with larger datasets or different training sets can be done in order to test the strategy more thoroughly. Increasing the amount of data with robust reference labels to train the CNNs may also help to further improve classification performance [48]. A data set with more training images contains a larger number of diverse training samples which can help the CNN to generalize better on unseen data. Using the proposed SeS strategy, the CNN will figure out which samples to use for training, forestalling an increase in training time. Using a much larger data set for training will be part of future work.

Despite these considerations, it is worthy to note that this is the first work on automated detection of hemorrhages in color fundus images that reports performance on par with two human experts. This result was obtained on a large, completely independent, publicly available test set, the Messidor database. Our method has a substantial higher Az value of 0.972 on this set as compared to previous work which reported an Az value of 0.87 [25]. However, it has to be noted that a direct comparison cannot be performed as the training sets differ and the previous work used only a subset of 900 cases for evaluation. This excellent result confirms that convolutional neural networks have great potential to push forward the state-of-the-art in medical image analysis, similar to what has been achieved with this exceptionally powerful class of models in computer vision.

VII. Conclusion

We have presented a method to substantially speed-up the time-consuming training process of convolutional neural networks with a selective sampling strategy, named SeS, embedded in the training procedure. We have demonstrated excellent results in the identification of hemorrhages on color fundus images. The SeS method addresses the common issue in medical image analysis tasks that challenging examples comprise only a small subset of the available data. By dynamically focusing the training effort on these samples that pose greater difficulty, we have shown that an increased overall performance can be achieved while a smaller number of epochs is required to train the network.

Acknowledgment

This project was funded by ZonMw grant: "A cost effective solution for the prevention of blindness using computer-aided diagnosis and fundus photography", with project number 11.631.0003.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv:150201852v1*, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[5] D. C. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333–338, 2012.

[6] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8150, 2013, pp. 411–418.

[7] Y. Guo, G. Wu, L. A. Commander, S. Szary, V. Jewells, W. Lin, and D. Shent, "Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8674, 2014, pp. 308–315.

[8] A. A. Cruz-Roa, J. E. Arevalo Ovalle, A. Madabhushi, and F. A. González Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8150, 2013, pp. 403–410.

[9] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear Estimation and Classification*, ser. Lecture Notes in Statistics. Springer New York, 2003, vol. 171, pp. 149–171.

[10] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.

[11] H. Chen, L. Yu, Q. Dou, L. Shi, V. C. T. Mok, and P. A. Heng, "Automatic detection of cerebral microbleeds via deep learning based 3d feature representation," in *IEEE International Symposium on Biomedical Imaging*, 2015, pp. 764–767.

[12] Y. Freund, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, no. 2, pp. 133–168, 1997.

[13] M. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 647–660, 2013.

[14] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: active learning in imbalanced data classification," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 127–136.

[15] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," *Lancet*, vol. 376, no. 9735, pp. 124–136, 2010.

[16] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010.

[17] P. Jitpakdee, P. Aimmanee, and B. Uyyanonvara, "A survey on hemorrhage detection in diabetic retinopathy retinal images," in *Proceedings 9th International Conference Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2012, pp. 1–4.

[18] G. B. Kande, T. S. Savithri, and P. V. Subbaiah, "Automatic detection of microaneurysms and hemorrhages in digital fundus images," *Journal of Digital Imaging*, vol. 23, no. 4, pp. 430–437, 2010.

[19] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *IEEE Transactions on Medical Imaging*, vol. 27, no. 9, pp. 1230–1241, 2008.

[20] U. M. Akram and S. A. Khan, "Automated detection of dark and bright lesions in retinal images for early detection of diabetic retinopathy," *Journal of Medical Systems*, vol. 36, no. 5, pp. 3151–3162, 2012.

[21] M. Niemeijer, B. van Ginneken, J. Staal, M. S. A. Suttorp-Schulten, and M. D. Abràmoff, "Automatic detection of red lesions in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 24, no. 5, pp. 584–592, 2005.

[22] B. Antal and A. Hajdu, "Improving microaneurysm detection in color fundus images by using context-aware approaches," *Computerized Medical Imaging and Graphics*, vol. 37, no. 5, pp. 403–408, 2013.

[23] C. Sinthanayothin, J. F. Boyce, T. H. Williamson, H. L. Cook, E. Mensah, S. Lal, and D. Usher, "Automated detection of diabetic retinopathy on digital fundus images," *Diabetic Medicine*, vol. 19, no. 2, pp. 105–112, 2002.

[24] S. Deepa and S. Vijayprasath, "Certain investigation of the retinal hemorrhage detection in fundus images," *International Journal of Electronics and Communication Engineering*, vol. 2, no. 2, pp. 29–40, 2015.

[25] L. Tang, M. Niemeijer, J. M. Reinhardt, M. K. Garvin, and M. D. Abràmoff, "Splat feature classification with application to retinal hemorrhage detection in fundus images," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 364–375, 2013.

[26] D. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[27] A. G. Marrugo and M. S. Millán, "Retinal image analysis: preprocessing and feature extraction," *Journal of Physics: Conference Series*, vol. 274, no. 1, p. 012039, 2011.

[28] S. H. Rasta, M. E. Partovi, H. Seyedarabi, and A. Javadzadeh, "A comparative study on preprocessing techniques in diabetic retinopathy retinal images: illumination correction and contrast enhancement," *Journal of Medical Signals and Sensors*, vol. 5, no. 1, pp. 40–48, 2015.

[29] B. Graham, "Kaggle diabetic retinopathy detection competition report," University of Warwick, Tech. Rep., 2015.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:14091556*, 2014.

[31] A. Seff, L. Lu, K. M. Cherry, H. R. Roth, J. Liu, S. Wang, J. Hoffman, E. B. Turkbey, and R. M. Summers, "2D view aggregation for lymph node detection using a shallow hierarchy of linear classifiers," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8673, 2014, pp. 544–552.

[32] P. Efraimidis and P. Spirakis, "Weighted random sampling," in *Encyclopedia of Algorithms*. Springer US, 2008, pp. 1–99.

[33] V. Rajan, R. K. Ghosh, and P. Gupta, "An efficient parallel algorithm for random sampling," *Information Processing Letters*, vol. 30, no. 5, pp. 265–268, 1989.

[34] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic acceptance," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 6, pp. 2193 – 2196, 2012.

[35] X. He and E. Frey, "ROC, LROC, FROC, AFROC: An alphabet soup," *Journal of the American College of Radiology*, vol. 6, no. 9, pp. 652–655, 2009.

[36] M. J. J. P. van Grinsven, Y. T. E. Lechanteur, J. P. H. van de Ven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Automatic drusen quantification and risk assessment of age-related macular degeneration on color fundus images," *Investigative Ophthalmology and Visual Science*, vol. 54, no. 4, pp. 3019–3027, 2013.

[37] N. Karssemeijer, J. D. M. Otten, A. L. M. Verbeek, J. H. Groenewoud, H. J. de Koning, J. H. C. L. Hendriks, and R. Holland, "Computer-aided detection versus independent double reading of masses on mammograms," *Radiology*, vol. 227, no. 1, pp. 192–200, 2003.

[38] A. Gubern-Mérida, R. Marti, J. Melendez, J. Hauth, R. Mann, N. Karssemeijer, and B. Platel, "Automated localization of breast cancer in DCE-MRI," *Medical Image Analysis*, vol. 20, no. 1, pp. 265–274, 2015.

[39] F. Samuelson and N. Petrick, "Comparing image detection algorithms using resampling," in *IEEE International Symposium on Biomedical Imaging*, 2006, pp. 1312 –1315.

[40] F. Samuelson, N. Petrick, and S. Paquerault, "Advantages and examples of resampling for CAD evaluation," in *IEEE International Symposium on Biomedical Imaging*, 2007, pp. 492 –495.

[41] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:150203167*, 2015.

[44] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[45] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, pp. 1–6, 2004.

[46] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv:14114038*, 2015.

[48] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.