LOW COMPLEXITY CONVOLUTIONAL NEURAL NETWORK FOR VESSEL SEGMENTATION IN PORTABLE RETINAL DIAGNOSTIC DEVICES

Mohsen Hajabdollahi1, Reza Esfandiarpoor1, Kayvan Najarian2,3, Nader Karimi1, Shadrokh Samavi1,3, S.M.Reza Soroushmehr2,3

1Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran
2Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, U.S.A.
3Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, U.S.A.

ABSTRACT

Retinal vessel information is helpful in retinal disease screening and diagnosis. Retinal vessel segmentation provides useful information about vessels and can be used by physicians during intraocular surgery and retinal diagnostic operations. Convolutional neural networks (CNNs) are powerful tools for classification and segmentation of medical images. However, complexity of CNNs makes it difficult to implement them in portable devices such as binocular indirect ophthalmoscopes. In this paper a simplification approach is proposed for CNNs based on combination of quantization and pruning. Fully connected layers are quantized and convolutional layers are pruned to have a simple and efficient network structure. Experiments on images of the STARE dataset show that our simplified network is able to segment retinal vessels with acceptable accuracy and low complexity.

Index Terms— Retinal image segmentation, convolutional neural network, network pruning, network binarization

1. INTRODUCTION

Most of the eye diseases and loss of vision are due to the diabetes [1]. Analysis and screening of the retinal vessels are very useful in eye disease detection and diagnosis. Diabetic retinopathy (DR) can be controlled with regular examination of eyes in early stages [1]. By the increase of DR, automatic detection of vessels has become important in telemedicine applications. Proper knowledge about retinal vessels could be also helpful during any retinal surgery operation [2]. As illustrated in Fig. 1, automatic segmentation of vessels can be helpful during intraocular surgery [3].

The problem of retinal vessel segmentation is investigated by many researchers in recent years. In [4], local thresholding using gray-scale level intensity as well as spatial dependencies is used for coarse segmentation of the vessel points. In order to identify the vessel borders, Wang et al. [5] propose a match filter based on wavelet kernels. In [6], image blocks are summarized as vectors in each direction and a match filter is utilized to provide seeds for a region growing algorithm. In [7], a combination of shifted and blurred DOG filters is used to detect the vessel points.

Convolutional neural networks have been introduced as a powerful tool for medical image analysis. In this regard [8-11] are focused on the problem of vessel segmentation using CNNs. Li et al. [8], employ CNN for segmentation of an entire patch. In [9], CNN is trained on a large dataset containing 400,000 retinal image samples. In [10] probability map generated from CNN is enhanced using conditional random field (CRF). Fu et al. [11], utilize a multi-level CNN to represent the feature hierarchies. In their work, CNN layers are equipped with a side output layer to approximate ground truth by minimizing a loss function.

Vessel segmentation in intraocular surgery requires real time vessel analysis and for portable and onsite devices needs real time and low power segmentation systems [3, 12, 13]. In this regard some recent studies focused on the problem of real time retinal vessel segmentation on dedicated hardware such as GPU and FPGA. In [14, 15] methods for retinal vessel analysis are presented and implemented on GPU. In [12] and [13] hardware architectures are presented for retinal vessel segmentation system.

Although CNN has been introduced as state-of-the-art methods in classification problems, its structural complexity puts on a lot of arithmetic operations. Simplification methods recently have been developed on the CNN structure. In [16], a framework for fixed point representation of CNN, and in [17] a method for pruning connections, are proposed. Also binarized neural networks are introduced as another way to overcome the complexity of CNNs [18, 19]. Applying CNN in portable devices for retinal vessel analysis such as binocular indirect ophthalmoscope requires overcoming the complexity problem of CNN.

In this paper, the problem of simplifying CNN as a powerful classifier for retinal image segmentation is
addressed. The proposed method is based on two techniques including pruning and quantization. Fully connected layers are quantized and convolutional layers are pruned effectively. The remainder of this paper is organized as follows. In Section 2, proposed method for retinal vessel segmentation is proposed. Section 3 is dedicated to the experimental results. Finally, in Section 4 the concluding remarks are presented.

2. PROPOSED METHOD

In Fig. 2, an overview of the system for segmentation of retinal blood vessels is presented. Here, a convolutional neural network (CNN) structure is applied on enhanced retinal images for vessel segmentation. As it can be observed in Fig. 2, fully connected layers are first modified to reduce the complexity and then convolutional layers are pruned and unnecessary weights are removed. Proposed method stages are described in more details as follows.

2.1. Retinal Image Enhancement

Low contrast retinal images can be enhanced using histogram equalization as preprocessing [1, 7]. The green channel is considered as the most representative channel in RGB fundus images [1]. In Fig. 3 (a), a sample retinal image in the form of RGB as well as R, G and B are represented. The Fig. 3 (a) images are enhanced using local histogram equalization and shown in Fig. 3 (b). Considering enhanced RGB image as the input of a CNN, complexity of the network structure would increase. Here, we select the enhanced gray-scale level of images to have adequate information as well as simple network structure. The enhanced gray-scale level image which has average information of all the channels is considered as input of the CNN for vessel segmentation.

2.2. Simplified CNN

Main parts of a typical CNN are convolutional layers (CLs) and fully connected layers (FCLs) where CLs perform most of the arithmetic operations and FCLs have the majority of the parameters [16]. These two parts must be taken into consideration for simplification of the CNN structure. Recently quantization and pruning have been used for reducing the structural complexity of neural networks. In the proposed method a hybrid method for simplification of CNNs including quantization and pruning is presented. In the following, the proposed method for quantization and pruning is presented.

2.2.1. Quantization

In binarization, a value converts to two possible values such as 0 and 1. In [18, 19], deterministic and stochastic methods are proposed for weight binarization as (1) and (2).

\[ W_b = \begin{cases} +1 & \text{if } W \geq 0, \\ -1 & \text{otherwise.} \end{cases} \]  

(1)

\[ W_b = \begin{cases} +1 & \text{with probability } p = \sigma(w), \\ -1 & \text{with probability } 1 - p. \end{cases} \]  

(2)

\[ \sigma(x) = \text{clip} \left( \frac{x + \frac{1}{2}}{2}, 0, 1 \right) = \max(0, \min(1, \frac{x + \frac{1}{2}}{2})) \]  

(3)

In (1) and (2), \( W \) and \( W_b \) are original and binarized network weights respectively. In (3), \( \sigma(x) \) is the "hard sigmoid function". In this paper deterministic ternary quantization of the weights is applied as (4).

\[ W_b = \begin{cases} -1 & \text{if } W < 0 \\ 0 & \text{if } W = 0 \\ 1 & \text{if } W > 0 \end{cases} \]  

(4)

As it is observed from Fig. 2 after CNN training, only FCLs are quantized in form of (4). As quantization without retraining causes a significant drop of accuracy, CNN is retrained after quantization. The process of quantization-retraining is repeated until acceptable accuracy is obtained. 32-bit representation consumes 32 times more memory size and memory accesses than the binary representation [20]. Quantized representation with 2 bits significantly reduces energy consumption of the arithmetic operations [20]. Although quantization significantly reduces the network complexity, quantization of convolutional layers degrades the network’s learning capabilities. Hence, in the proposed method, quantization is performed only on weights of fully connected layers. While convolutional layers would remain complex, all multiplications in FCLs are replaced with simple transfer operations.

2.2.2. Pruning

Pruning is yet another way of reducing the network’s complexity which can eliminate unnecessary weights during the network’s training [17]. Basic pruning schema has three steps. At first, network weights and connections are trained. Weights smaller than a pre-defined value are eliminated from the network. Finally, the loss of accuracy due to weight eliminations is recovered by retraining. This process is continued until the network reaches a stable state. CL quantization has undesirable effect on the network performance; however, pruning removes the un-necessary CL connections. Pruning CL connections can be introduced as an efficient way to overcome the problem of large number of arithmetic operations in the convolutional layers.

![Fig. 2. Proposed simplified CNN for retinal vessel segmentation.](image-url)
In a binarized network, one-bit operations are required which are useful for the simplicity of the network. However, drastic changes of the weights caused by quantization may have destructive effects on the network learning capability. Although bit length parameters are long in the pruned network, pruning creates lower changes than the quantization. A combination of these two methods including quantization and pruning can be considered as an effective way to simplify the network structure.

3. EXPERIMENTAL RESULTS

For evaluation of the proposed method, experimental results are performed using TensorFlow framework. Simplified CNN is trained and tested on the STARE image dataset with the first observer segmentation [21]. Image patches with the size of 9×9 from the enhanced gray-scale level images are extracted around each pixel and fed to the network input. Output of each patch is set as the class of the pixel under consideration and CNN is trained as a binary classification problem. Accuracy, sensitivity, specificity and ROC are used for classification performance evaluation where 5-fold cross validation method is used for performance validation. Experimental results are organized in three parts as follows.
3.1. CNN with original parameters

Different CNN configurations are experimented with low structural complexity consideration. Finally, a CNN with low complexity is applied. Visual results of segmentation applied on some image samples from STARE [21] with original CNN parameters are illustrated in Fig. 5(a). Performance results of CNN without any simplification (original parameters) are compared with related works in Table 1. As illustrated in Table 1, the network has suitable parameters without any simplification. DICE score and accuracy are achieved to be about 0.76 and 0.96, respectively. In Fig. 6, receiver operating characteristic (ROC) of the classification using CNN with original parameters is depicted. Although the performance results are desirable, the network structure has not been simplified. Applying the segmentation method on the portable devices with the limited hardware resources requires a simple and fast network structure.

3.2. CNN Quantization

CNN after training is quantized using three levels including 0, 1 and -1 as mentioned in (4). Firstly, both convolutional layers and fully connected layers are quantized as (4). As it was mentioned in Section 2.2, quantization changes the weights drastically that may lead to learning disability. It is observed that the fully quantized network cannot reach to a suitable performance and the accuracy is about 0.75. Therefore, only the fully connected layers are quantized. It can be observed from Fig. 5(b) that the CNN with quantized fully connected layers is able to segment the vessel points approximately the same as its original one. Visual results shown in Fig. 5(b) are not significantly different from Fig. 5(a) images. Also performance of the quantized FCLs in Table 1 and Fig. 6 is comparable with CNN with original parameters. For example, accuracy reduction of 0.003 is resulted as compared with the original CNN.

3.3. CNN training with CLs pruning and FCLs quantization

Although the FCLs are quantized, the complexity of the network in CLs is still high. Therefore, after quantization, filter weights are pruned as shown in Fig. 4 to eliminate the unnecessary weights for the simpler structure. Visual results in Fig. 5(c) and ROC result in Fig. 6 represent no significant difference in the performance. Also, Table 1 shows 0.0036 accuracy drop in case of pruned-quantized CNN. Finally, in Table 2 network structure is presented before and after simplification. In CLs about 60% of the weights are removed by pruning and quantization causes FCLs weights to have a ternary value.

4. CONCLUSION

In this paper a CNN for automatic segmentation of the vessels in fundus images was presented. The structure of the CNN was simplified to make it suitable for hardware implementation in portable and onsite retinal diagnostic devices. Combination of two simplification mechanisms including quantization in fully connected layers and pruning in convolutional layers was applied. Simulation results of the proposed method applied on the STARE dataset demonstrated acceptable performance for the simplified CNN. Finally, we removed 60% of the convolutional layer weights and quantized all the fully connected layer weights and achieved the AUC of 0.97. The proposed simplified CNN could be considered as a method for automatic segmentation of vessels in portable retinal diagnostic devices.
REFERENCES

[1] H. Bendaoudi, “Flexible Hardware Architectures for Retinal Image Analysis,” Doctoral dissertation, École Polytechnique de Montréal, 2017.

[2] MD. Abrámoff, MK. Garvin, M. Sonka, “Retinal Imaging and Image Analysis,” IEEE reviews in biomedical engineering, 3, pp. 169-208, 2010.

[3] BC. Becker, CN. Riviere, “Real-Time Retinal Vessel Mapping and Localization for Intraocular Surgery,” IEEE International Conference on Robotics and Automation (ICRA), pp. 5360-5365, 2013.

[4] LC. Neto, GL. Ramalho, JF. Neto, RM. Veras, FN. Medeiros, “An Unsupervised Coarse-To-Fine Algorithm for Blood Vessel Segmentation in Fundus Images,” Expert Systems with Applications, 78, pp. 182-192, 2017.

[5] Y. Wang, G. Ji, P. Lin, E. Trucco, “Retinal Vessel Segmentation Using Multwavelet Kernels and Multiscale Hierarchical Decomposition,” Pattern Recognition 46.8, pp. 2117-2133, 2013.

[6] I. Lázár, A. Hajdu, “Segmentation of Retinal Vessels by Means of Directional Response Vector Similarity and Region Growing,” Computers in biology and medicine, 66, pp. 209-221, 2015.

[7] G. Azzopardi, N. Strisciuglio, M. Vento, N. Petkov, “Trainable COSFIRE Filters for Vessel Delineation with Application to Retinal Images,” Medical image analysis, 19.1, pp. 46-57, 2015.

[8] Q. Li, H. Xie, Q. Zhang, S. Qi, P. Liang, H. Zhang, T. Wang, “A Supervised Method Using Convolutional Neural Networks for Retinal Vessel Delineation,” IEEE International Congress on Image and Signal Processing (CISP), pp. 418-422, 2015.

[9] P. Liskowski and K. Krawiec, “Segmenting Retinal Blood Vessels with Deep Neural Networks,” IEEE transactions on medical imaging, 35(11), pp.2369-2380, 2016.

[10] H. Fu, et al., “Retinal Vessel Segmentation Via Deep Learning Network and Fully-Connected Conditional Random Fields,” IEEE International Symposium on Biomedical Imaging (ISBI), pp. 698-701, 2016.

[11] H. Fu, Y. Xu, S. Lin, DW. Wong, J. Liu, “Deep Vessel: Retinal Vessel Segmentation via Deep Learning and Conditional Random Field,” Springer International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 132-139, 2016.

[12] D. Koukounis, C. Tofis, A. Papadopoulos, T. Theocharides, “A High Performance Hardware Architecture for Portable, Low-Power Retinal Vessel Segmentation,” INTEGRATION, the VLSI journal, 47 (3), pp. 377-86, 2014.

[13] M. Hajabdollahi, et al., “Retinal Blood Vessel Segmentation for Macula Detachment Surgery Monitoring Instruments,” international journal of circuit theory and applications, pp. 1-15, 2018.

[14] M. Krause, RM Alles, B. Burgeth, J. Weickert, “Fast Retinal Vessel Analysis,” Journal of Real-Time Image Processing, 11 (2), pp. 413-22, 2016.

[15] F. Arguello, DL. Vilarino, DB Heras, A. Nieto, “GPU-Based Segmentation of Retinal Blood Vessels,” Journal of Real-Time Image Processing. pp. 1-10, 2014.

[16] P. Gyssel, M. Motamedi, S. Ghiasi, “Hardware-Oriented Approximation of Convolutional Neural Networks,” arXiv preprint 1604.03168, 2016.

[17] S. Han, J. Pool, J. Tran and W. Dally, “Learning Both Weights and Connections for Efficient Neural Network,” Advances in neural information processing systems, pp. 1135-1143, 2015.

[18] M.Courbariaux, I. HubaraSoudry et al., “Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to+ 1 or− 1,” arXiv preprint arXiv:1602.02830, 2016.

[19] M.Courbariaux, Y.Bengio and J.P. David, “Binaryconnect: Training Deep Neural Networks With Binary Weights During Propagations,” Advances in neural information processing systems, pp. 3123-3131, 2015.

[20] M. Horowitz, “Computing’s Energy Problem and What We Can Do About It,” IEEE International Conference on Solid-State Circuits Digest of Technical Papers (ISSCC), pp. 10-14, 2014.

[21] AD. Hoover, V. Kouznetsova, M. Goldbaum, “Locating Blood Vessels in Retinal Images by Piecewise Threshold Probing of a Matched Filter Response,” IEEE Transactions on Medical imaging, 19(3), pp. 203-210, 2000.