Large Receptive Field Fully Convolutional Network for Semantic Segmentation of Retinal Vasculature in Fundus Images

Gabriel Lepetit-Aimon(✉), Renaud Duval, and Farida Cheriet

Ecole Polytechnique de Montréal, Montreal, QC H3T 1J4, Canada
gabriel.lepetit-aimon@polymtl.ca

Abstract. Analysis of the retinal vasculature morphology from fundus images, using measures such as arterio-venous ratio, is a promising lead for the early diagnosis of cardiovascular risks. The accuracy of these measures relies on the robustness of the vessels segmentation and classification. However, algorithms based on prior topological knowledge have difficulty modelling the abnormal structure of pathological vasculatures, while patch-trained Fully Convolutional Neural Networks (FCNNs) struggle to learn the wide and extensive topology of the vessels because of their narrow receptive fields.

This paper proposes a novel Fully Convolutional Neural Network architecture capable of processing high resolution images through a large receptive field at a minimal memory and computational cost. First, a single branch CNN is trained on whole images at low resolution to learn large scale features. Then, this branch is incorporated into a standard encoder/decoder FCNN: its large scale features are concatenated to those computed by the central layer of the FCNN. Finally, the whole network architecture is trained on high-resolution patches. During this last phase, the FCNN benefits from the large scale features while the low resolution branch parameters are fine-tuned. This architecture was evaluated on the publicly available retinal fundus database DRIVE. The trained network achieves an accuracy of 96.1% in segmenting the full retinal vessels and improves by 5% the artery/vein classification compared to a basic U-Net.

Keywords: Retinal vessel segmentation · Retinal vessel classification
Convolutional neural network · Deep learning

1 Introduction

Early diagnosis is a key to reducing mortality rates in cardiovascular diseases, which caused 30.8% of deaths and were the top healthcare expenditure in the USA in 2013 [9]. Retinal fundus imaging allows the non-invasive observation of the retinal vascular system. This modality thereby offers a good overview of cardiovascular health: statistically, a patient suffering from retinopathy is twice as likely to have a stroke [13]. Clinicians already use fundus images to evaluate

D. Stoyanov et al. (Eds.): COMPAY 2018/OMIA 2018, LNCS 11039, pp. 201–209, 2018.
https://doi.org/10.1007/978-3-030-00949-6_24

cardiovascular risks, by analysing the retinal vasculature morphology through measures like the arterio-venous ratio (highly correlated with hypertension and diabetes risks). However, the accuracy of these measures relies on the robustness of the vessels segmentation and classification between arteries and veins.

In the last two decades, many algorithms were developed to handle those tasks. Most of them perform the vessel segmentation separately from the classification. For the former task, traditional computer vision algorithms such as the multi-scale line detector [10] have been out-performed by deep learning algorithms, using either convolution neural networks (*CNN*s) [8], CNNs combined with Conditional Random Fields (*CRF*s), e.g. the Deep Vessel architecture [3], or by adversarial architectures [6]. Meanwhile, for vessel classification, state of the art methods generally combine local features analysis by machine learning algorithms with prior knowledge of the vascular tree structure. Dashtbozorg *et al.* use linear discriminant analysis outputs in combination with the vascular graph corrected by rules derived from prior knowledge [1]. More recently, Estrada *et al.* proposed a graph-based algorithm to extract the vascular tree from a fundus image and classify each detected vessel using local features [2]. However, relying on rules derived from prior knowledge can impact the robustness of the algorithm, in particular for severe cases of retinopathy where the vasculature won't match the rules.

Recent progress in deep learning has made possible the training of larger and deeper Fully Convolutional Neural Networks (*FCNN*s). In particular, the U-Net achieves remarkably good performance in segmenting medical images thanks to its encoder/decoder architecture and to its skip-connexions [12]. However, to our knowledge, those architectures have never reached state of the art performance in artery/vein classification. Indeed, because of their narrow receptive fields, FCNNs struggle to learn the wide and extensive topology of retinal vessels.

This paper propose a novel Large Receptive Field Fully Convolutional Network architecture (*LRFFCN*), capable of segmenting very extensive shapes (i.e. vessels) in high-resolution images at a minimal memory and computational cost. This paper is organized as follows: (1) analysis of the U-Net's poor performances in classifying arteries and veins; (2) description of a novel FCNN architecture: the Large Receptive Field Fully Convolutional Network; (3) evaluation of the LRFFCN architecture experimentally in the semantic vessel segmentation task.

2 Methods

2.1 Large Receptive Field Fully Convolutional Network

Receptive Field Limitation in Convolutional Networks. The term *receptive field* is inherited from neurosciences and describes the region of the sensory space (e.g. the visual field) in which a stimulus will cause a neuron to be activated. In deep learning, the receptive field of a convolutional network is the region of the input analyzed by the network to produce the prediction for *one* pixel. Early CNN designers performing pixel-wise classification didn't care about this concept: because their models ended with fully connected layers, the receptive field was the whole input patch. However, this is not the case with FCNNs.

For example, the encoding branch of the U-Net architecture has a receptive field of 125×125 pixels even though its training patch size is 500×500 pixels.

There are several ways to increase the receptive field of an FCNN. Stacking more layers or extending their kernel size will theoretically increase the receptive field linearly, whereas sub-sampling the output features of a layer will increase it multiplicatively. In practice, Luo *et al.* have shown that the Effective Receptive Field (*ERF*) is always narrower than the theoretical one [7]. More precisely, the ERF follows a Gaussian distribution with a standard deviation depending on the model architecture and the weights initialization. Extending the kernel size will increase the ERF linearly; stacking n layers will only increase the ERF by a factor $\sqrt{n}$; sub-sampling will effectively increase the ERF quickly. On the contrary, skip-connections will shrink the ERF.

Focusing on the U-Net architecture, its decoding branch shouldn't have much impact on the ERF: the growth due to the convolutional layers is compensated by the upsampling and the skip-connections. Intuitively, this branch is only a complex upsampling interpolation of the deep features. In other words, the ERF of a U-Net is strictly lower than the theoretical receptive field of its encoding branch: 125×125 pixels. When processing high-resolution fundus images (2048×2048 pixels), such a receptive field is much too small to learn the topology of the vasculature, thus the network can only rely on local texture and color features. However, for small vessels far away from the optic disk, those features are not sufficient, even for clinicians, to efficiently discriminate arteries from veins.

As the U-Net architecture is already a large model, adding more layers to it would make the model too heavy. Namely adding a new pooling/up-conv stage would raise the forward-pass computation from 67 to 389 Mega Flops (mainly because the patch size is doubled).

LRFFCN Architecture. The Large Receptive Field Fully Convolutional Network is a novel network architecture which significantly increases the receptive field at a minimal memory and computational cost. The core of the architecture is a convolutional branch processing the full image at a low resolution (scaled down to a 128×128 pixels patch). This branch is structured as a repetition of fire/squeeze modules inspired by the SqueezeNet (a concise model with similar performance to AlexNet [4]). The theoretical receptive field of this branch is 21×21 pixels at $1/16^{th}$ of the full resolution, which corresponds to 336×336 pixels in the high resolution image. The large scale features learned by this branch are then incorporated in an encoder/decoder FCNN with skip-connections.

The encoding stage of this network contains 4 pooling layers (2×2 pooling), so the resolution of the deepest features F_e is $1/16^{th}$. Because the FCNN is trained on patches and not on the full image, a region corresponding to the patch is extracted from the large scale features and is concatenated to F_e. Thus, the decoding stage of the FCNN has access to textural information from high resolution features learned by the encoding stage, and to topological information from the large scale features at the cost of a forward-pass computation increase of only 6% (4 MFlops).

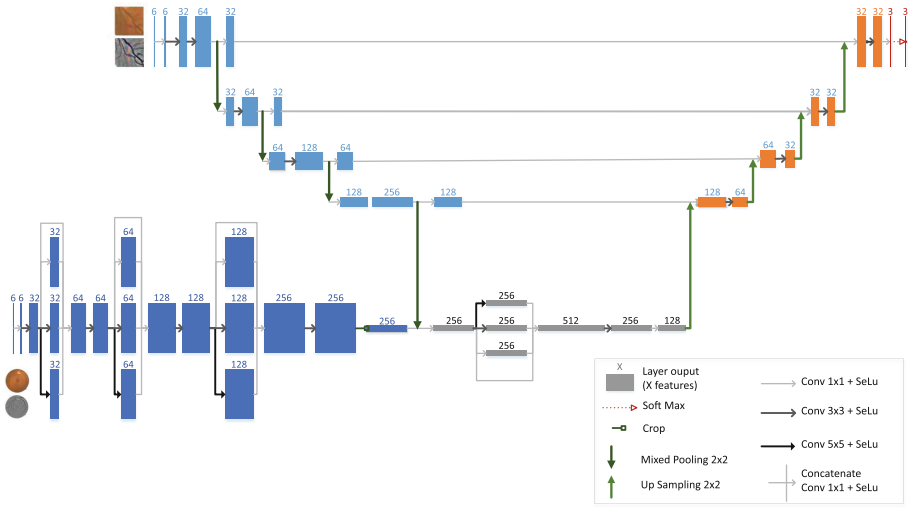

Fig. 1. The large receptive field fully convolution network architecture. Color code: *light blue*: encoding branch; *orange*: decoding branch; *dark blue*: low resolution branch. (Color figure online)

2.2 Model Specificities

The exact model trained for this paper is presented in Fig. 1. Its most interesting specificities are presented in this section.

Due to memory limitations, the maximum minibatch size during training is 8. Batch normalization layers would therefore be ineffective and have been replaced by a SeLu activation function [5]. This modified version of the ReLu activation function offers negative values which speed up learning by pushing mean activation towards 0, and a stable fixed point of the activation mean and variance so that exploding and vanishing gradients are impossible.

The model is trained to perform semantic segmentation: the vessels segmentation is performed simultaneously with the classification as each pixel is assigned a probability of being part of the background, an artery or a vein. Even if the LRFFCN architecture efficiently increases the ERF, performing this classification pixel-wise can cause local errors in the artery/vein classification. Indeed, because arteries and veins are visually similar and because of the local inconsistencies transmitted from the early layers to the later ones through the skip-connections, some pixels from a vessel can suddenly be predicted as veins even though they are surrounded by arteries (and vice versa). To correct such local inconsistencies, a CRF is added at the end of the model.

CRFs use the preprocessed version of the training patch as a reference and 2D Gaussian kernels to propagate the probabilities of being an artery or a vein along the vessel. We used the CRF as Recurrent Neural Network (*RNN*) implementation proposed by Zheng *et al.* [14] so its kernels are trained simultaneously with the LRFFCN parameters.

2.3 Training

The model was trained on 69 images: 30 images from the MESSIDOR training dataset, 19 images from the STARE dataset and 20 images from the DRIVE training dataset. We manually labelled the MESSIDOR and STARE images and asked an ophthalmologist to validate this labelling. Each image was preprocessed using a standard contrast enhancement technique. Both the enhanced and raw images were presented to the network since vessel detection strongly relies on the enhanced image whereas vessel classification relies on the vessels' true colors. Color and geometric data augmentation were used to double the training dataset size (contrast and gamma variations for color augmentation, horizontal mirroring and rotation for geometric augmentation).

The training process was split into 3 phases. The large scale branch was pre-trained first over 100 epochs on full images rescaled to 128×128 pixels. Then, the full LRFFCN was trained on 230×230 pixels patches; the large-scale features already learned by the low resolution branch allow a quick convergence of the network. This training phase was driven by the Adadelta gradient descent optimizer and lasted 30 epochs during which the learning rate was slowly decreased. Finally, the CRF as RNN was added to the classifier layer and the network was trained again for 10 epochs.

3 Experiments

3.1 Evaluating the LRFFCN Architecture

We used the 20 images from the DRIVE test dataset to evaluate the LRFFCN architecture. The architecture was tested with and without the CRF as RNN layer. Also, to quantify the contribution of the low resolution branch to the prediction, we evaluated the performance of the model when the scaled-down image input to this branch was replaced by a uniform noise. (We refer to this network as *LRFFCN w/o low branch*.) Finally, a classic U-Net was trained for the same number of epochs to estimate the performance gain. For each of these models, we measured the accuracy of the vessel segmentation and the accuracy, specificity and sensitivity of the artery/vein classification. The results are presented in Table 1.

Table 1. Performance of LRFFCN architecture on DRIVE test dataset.

Architecture	Vessels	Arteries/Veins		
	Accuracy	Accuracy	True arteries	True veins
Basic U-Net	$95.6 \pm 0.4\%$	$75.9 \pm 3.0\%$	70.9%	80.6%
LRFFCN w/o low branch	$96.0 \pm 0.3\%$	$63.1 \pm 3.7\%$	57.7%	74.1%
LRFFCN	**$96.1 \pm 0.3\%$**	$79.4 \pm 3.5\%$	73.9%	**86.4%**
LRFFCN + CRF	$95.9 \pm 0.4\%$	**$81.0 \pm 3.8\%$**	**77.8%**	84.4%

The segmentation accuracy is consistent across architectures and is quite high, confirming the efficiency of convolutional networks in segmenting vessels. However, the LRFFCN architecture improves the classification accuracy by almost 5% compared to the basic U-Net. The large-scale features learned by the low resolution branch seems to be effectively used by the model to improve its generalizing capabilities. This is confirmed by the noise experiment: replacing the real data by noise make the classification accuracy drop by 16%. Thus, the large-scale feature branch contributes greatly to the network's predictions.

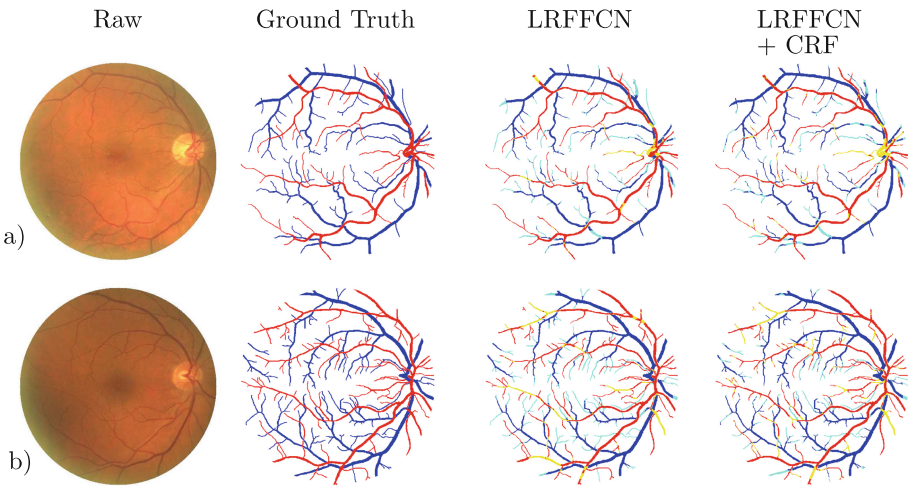

Fig. 2. Comparison of LRFFCN performance with and without CRF as RNN, on two images from DRIVE test dataset. Top row: 06_test; bottom row: 17_test. Color codes: *red*: true artery; *dark blue*: true vein; *light blue*: misclassified vein; *yellow*: misclassified artery. (Color figure online)

The CRF layer improves the classification accuracy by 2% and successfully propagates the vessel classes to correct local errors. Those corrections are visible in Fig. 2(b), where artery segments classified as veins by the LRFFCN are corrected by the CRF layer. However, the LRFFCN's predictions are sometimes not sufficiently accurate and the CRF layer propagates misclassifications (as visible for some veins in Fig. 2(a)). Overall, the CRF improves the topological plausibility of the predicted vessel network.

3.2 Model Segmentation and Classification Performance

In this section, we compare the performance of the LRFFCN to those of state of the art algorithms. For the segmentation task, the LRFFCN exceeds by 1.3% the Deep Vessel network in terms of accuracy. More precisely, our architecture is much more sensitive (80.8% against 72.7%) (Tables 2 and 3).

Table 2. Segmentation results on DRIVE test dataset.

Name	Accuracy	Specificity	Sensitivity
Mozzafarian *et al.* [9]	82.2%	–	–
Adversarial, Lahiri *et al.* [6]	94.%	–	–
Deep Vessel, Fu *et al.* [3]	94.6%	**97.7%**	72.7%
LRFFCN	**95.9 ± 0.4%**	97.3%	**80.8%**

Table 3. Classification results on DRIVE test dataset.

Name	Accuracy	True arteries	True veins
Niemeijer *et al.* [11]	80.0%	80.0%	80.0%
LRFFCN	81.0 ± 3.8%	77.8%	84.4%
Dashtbozorg *et al.* [1]	87.4%	90.0%	84.0%
Estrada *et al.* [2]	**91.7 ± 5%**	**91.7%**	**91.7%**

For the artery/vein classification, the performance of the LRFFCN architecture is still 10% below state of the art graph-based algorithms. Indeed, for the DRIVE dataset, the prior topological knowledge of retinal vessels provides a good enough estimation of the vascular tree. Graph-based algorithms can propagate the artery/vein probability through this tree, whereas our method didn't perfectly learn the vasculature topology and often misclassify the small vessels farther from the optic disk.

4 Discussion and Conclusion

The proposed LRFFCN architecture efficiently increases the receptive field of the FCNN by means of a low resolution branch and successfully takes advantage of large-scale features to better learn the retinal vessel topology. In particular, the LRFFCN architecture outperforms the U-Net in classifying retinal arteries and veins. It also does better than state of the art algorithms in vessel segmentation, yielding a higher sensitivity. For the classification task, the LRFFCN architecture does not reach state of the art performance. However, we believe that this particular result is due to the training conditions and not to the model design.

Nevertheless, these results are promising. Indeed, in contrast to static graph-based analysis, the performance of this model will improve as the training dataset grows. In particular, the bottleneck for this architecture is the training of the low resolution branch. Because this branch must be trained on whole images and not on patches, our current training dataset is too small for the LRFFCN architecture to show its full potential. But because the segmentation accuracy is high, the predictions from our method can quickly be fixed by a clinical expert. In other words, the LRFFCN architecture can be used to efficiently generate more ground truth images on which it can then be trained to improve its vessel classification performance.

References

1. Dashtbozorg, B., Mendonca, A.M., Campilho, A.: An automatic graph-based approach for artery/vein classification in retinal images. IEEE Trans. Image Process. **23**(3), 1073–1083 (2014)
2. Estrada, R., Allingham, M.J., Mettu, P.S., Cousins, S.W., Tomasi, C., Farsiu, S.: Retinal artery-vein classification via topology estimation. IEEE Trans. Image Process. **34**(12), 2518–2534 (2015)
3. Fu, H., Xu, Y., Lin, S., Kee Wong, D.W., Liu, J.: DeepVessel: retinal vessel segmentation via deep learning and conditional random field. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 132–139. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_16
4. Iandola, F.N., Moskewicz, M.W., Ashraf, K., Han, S., Dally, W.J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1 MB model size. CoRR abs/1602.07360 (2016). http://arxiv.org/abs/1602.07360
5. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Advances in Neural Information Processing Systems, vol. 30, pp. 971–980. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/6698-self-normalizing-neural-networks.pdf
6. Lahiri, A., Ayush, K., Biswas, P.K., Mitra, P.: Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale miscroscopy images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 794–800, July 2017
7. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29, pp. 4898–4906. Curran Associates, Inc. (2016). http://papers.nips.cc/paper/6203-understanding-the-effective-receptive-field-in-deep-convolutional-neural-networks.pdf
8. Maninis, K.-K., Pont-Tuset, J., Arbeláez, P., Van Gool, L.: Deep retinal image understanding. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 140–148. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_17
9. Mozaffarian, D., et al.: Heart disease and stroke statistics–2016 update. Circulation **133**(4), e38–e360 (2016). http://circ.ahajournals.org/content/133/4/e38
10. Nguyen, U.T., Bhuiyan, A., Park, L.A., Ramamohanarao, K.: An effective retinal blood vessel segmentation method using multi-scale line detection. Pattern Recognit. **46**(3), 703–715 (2013). http://www.sciencedirect.com/science/article/pii/S003132031200355X
11. Niemeijer, M., et al.: Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs. IEEE Trans. Med. Imaging **30**, 1941–1950 (2011)
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28. http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a, (Available on arXiv:1505.04597 [cs.CV])

13. Wong, T.Y., Klein, R., Sharrett, A.R.: The prevalence and risk factors of retinal microvascular abnormalities in older persons: the cardiovascular health study. Ophthalmology **110**, 658–666 (2003)
14. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: International Conference on Computer Vision (ICCV) (2015)