Noise Reduction in Optical Coherence Tomography using Deep Image Prior

Kristen Hagan, David Li, and Jessica Loo

***Abstract*—Optical coherence tomography (OCT) is a cross-sectional imaging technique that relies on coherence of the light source to obtain depth information of biological tissue such as the human eye. Collected OCT images are mainly corrupted by shot noise and speckle noise, which is an artifact due to the interference of light that has been scattered in the turbid media. All methods that currently exist to denoise OCT images rely on averaging or supervised machine learning methods. Here, we employ an unsupervised method based on a deep image prior to reconstruct a denoised version of an OCT B-scan from random noise. We investigate various network architectures and loss functions for a series of B-scans and show the results of the noise reduction. All code is available at:** https://github.com/jessicaloohw/BME590_DeepImagePrior

***Index Terms*—Deep image prior, deep learning, denoising, optical coherence tomography, speckle reduction.**

I. Introduction

Optical coherence tomography (OCT) is a 3D optical imaging technique where reflectance profiles are acquired using an interferometer and broadband or swept-source laser. Profiles at a single location are grouped together into 2D images, or B-scans, and multiple B-scans can form a 3D volume. OCT has been adopted for a variety of clinical uses, but is most prominently used for screening and diagnosis in ophthalmology. However, OCT imaging is affected by several noise processes: mainly shot and speckle. Shot noise is a fundamental limiting noise process and in Fourier domain OCT (FD-OCT), it can be modeled as an additive, uncorrelated Gaussian white noise [1]. Speckle is an artifact of coherent imaging due to interaction of light with subvoxel features and can be modeled as multiplicative noise [2]. The presence of such noise can make it difficult to visualize fine structures and pathologies or perform segmentation to delineate different tissue layers. Traditional denoising methods can be broadly split into two categories: single frame techniques with a model for signal and noise or multi-frame averaging. Single frame techniques for noise and speckle reduction range from physical techniques such as frequency compounding [3] to digital Wiener filtering or wavelets [4]. Multi-frame averaging requires acquisition of repeated B-scans and registration which adds considerable overhead. Furthermore, since speckle is dependent on the signal, averaging will not necessarily remove speckle. More recently, machine learning based denoising approaches such as sparse representation dictionaries have been proposed [5]. However, these machine learning approaches require significant training. In our paper, we use deep image prior which does not require any training or a ground truth.

II. Methods

A. Data Set

We selected five images from a previously acquired data set of 22 different OCT images, consisting of both raw and averaged images [5]. Fig. 1 shows an example of the images in the data set. The images are cropped to a uniform size (496 x 928 pixels) and the values are normalized to between 0 and 1. Since an ideal noiseless image is not available, we use the averaged images as the ground truth. The ground truth is not required in the proposed method, and we use it only to evaluate the performance of the proposed method.

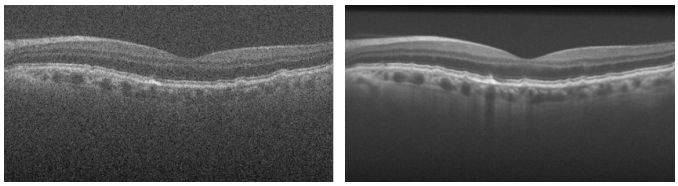

Fig. 1. Example of raw (left) and averaged (right) images. The averaged image is the average of 50 B-scans from the same location.

B. Deep Image Prior

Deep neural networks (DNNs) have become very popular for solving inverse problems in imaging. However, most methods involving DNNs require large amounts of training data. Deep image prior [6], on the other hand, shows that a randomly-initialized generator network can be used as a handcrafted prior to solve several inverse problems in imaging such as denoising, inpainting, and super-resolution. The prior is imposed by the architecture of the network, which is able to capture many low-level image statistics.

Inverse problems can be formulated as an optimization task

$$y^* = \min_y E(y;x) + R(y) \tag{1}$$

where x is the degraded image, $E(y;x)$ is a task-dependent *data term*, and $R(y)$ is a regularization term or *prior*. In this case, $y = f_\theta(z)$ is the output of the generator network, f parameterized by θ, given a random code vector, z. $R(y)$ is the deep image prior imposed by the network architecture, as well as any additional priors imposed. For the most basic reconstruction problem, where we simply want to reproduce x, the *data term* can be expressed as

$$E(y;x) = |y - x|^2 \tag{2}$$

which leads to the optimization task

$$\min_\theta |f_\theta(z) - x|^2 \tag{3}$$

The equation to be minimized is commonly known as the loss function, $\mathcal{L}$.

This solution in (3) is also the maximum likelihood estimate assuming Gaussian noise. Although the speckle noise distribution in OCT imaging is multiplicative, not Gaussian, we use this as the initial approach for reducing noise in OCT images. We further experiment with several network architectures, as well as different *data terms* and additional *priors* in our loss functions. Fig. 2 shows an overview of the proposed method.

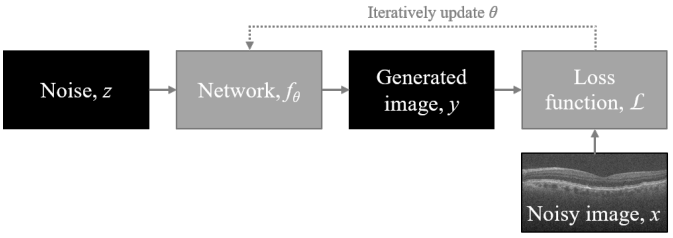

Fig. 2. Overview of the proposed method for reducing noise in OCT images. Black components indicate images (input and output) whereas gray components indicate functions.

C. Network

We use two network architectures for f_θ – U-Net [7], a popular convolutional encoder-decoder architecture for biomedical imaging, and Deep Decoder [8], a recent non-convolutional decoder-only architecture for denoising.

The U-Net architecture consists of a contracting and expansive path. The input to the network, z is a 32-channel random uniform noise matrix of the same dimensions as the input image, x. The contracting path consists of five encoder blocks. Each encoder block consists of two 3×3 convolutions, followed by a rectified linear unit (ReLU) non-linearity. 2×2 max-pooling is applied between each encoder block. The number of feature maps in each encoder block is 64, 128, 256, 512, and 1024, respectively. The expansive path consists of four decoder blocks. Each decoder block consists of a 2×2 transposed convolution to halve the number of feature maps and upsample them by a factor of two, followed by a concatenation with the feature maps (of the same size) from the corresponding encoder block of the contracting path. Then, two 3×3 convolutions are applied, followed by a ReLU non-linearity. The final layer consists of a 1×1 convolution, followed by a sigmoid non-linearity, to map the final feature map to the corresponding number of image channels, which for grayscale OCT images is one.

The Deep Decoder architecture consists of an expansive path only. z is a 64-channel random uniform noise matrix with $\frac{1}{16}^{th}$ the dimensions of x. The expansive path consists of six decoder blocks. Each decoder block consists of a pixel-wise linear combination of the channels, bilinear upsampling, a ReLU non-linearity, and channel normalization. All pixel-wise linear combinations were implemented as 1×1 convolutions. The 5^{th} and 6^{th} decoder blocks do not apply bilinear upsampling in order to retain the resolution of the input image. Similar to U-Net, the final layer maps the final feature map to a single-channel image.

D. Loss

We use five loss functions comprised of different combinations of *data terms* and additional *priors* which operate on the original noisy image, x and the generated image, y. The loss functions are

$$\mathcal{L}_{L_2} = |y - x|^2 \tag{4}$$

$$\mathcal{L}_{L_1} = |y - x| \tag{5}$$

$$\mathcal{L}_{L_2\_L_1} = \mathcal{L}_{L_2} + w_0 \, \mathcal{L}_{L_1} \tag{6}$$

$$\mathcal{L}_{L_2\_tv} = w_1 \, \mathcal{L}_{L_2} + w_2 \, tv(y) \tag{7}$$

$$\mathcal{L}_{L_2\_edge} = w_1 \, \mathcal{L}_{L_2} + w_2 \, edge_h(y) + w_3 \, edge_v(y) \tag{8}$$

where $tv(y)$ and $edge(y)$ are additional *priors* based on the gradients or edges of y to impose smoothness.

$tv(y)$ is total variation [9] regularization which is expressed as

$$tv(y) = \sum_{i,j} |y_{i+1,j} - y_{i,j}| + |y_{i,j+1} - y_{i,j}| \tag{9}$$

$edge_h(y)$ and $edge_v(y)$ is the sum of the magnitude of the image after convolution with Prewitt operators [10] which are commonly used for edge detection in image processing where the Prewitt operators are

$$P_h = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \tag{10}$$

$$P_v = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \tag{11}$$

We empirically determine the coefficients, $w_0 = 0.1$, $w_1 = 10$, $w_2 = 10^{-6}$, and $w_3 = 10^{-6}$. All losses also include L_2-regularization of network weights with a coefficient of 0.0001.

E. Optimization

We use the same optimization procedure with early stopping for all combinations of network architectures and loss functions. We run the optimization procedure for a maximum of 50,000 iterations using Adam [11] optimization with the following hyperparameters: learning rate = 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We use both quantitative and qualitative analysis to determine when to stop the optimization, to prevent overfitting to the original noisy image.

F. Quantitative Analysis

We use several standard image metrics – contrast-to-noise ratio (CNR) and structural similarity index (SSIM).

To evaluate CNR, we manually selected four different regions in each image as foreground, one region as the background, and used the following equation for each region

$$CNR = \frac{|\mu_f - \mu_b|}{0.5\sqrt{\sigma_f^2 + \sigma_b^2}} \tag{12}$$

where μ_f and μ_b represents the mean of the foreground and background regions, respectively, while σ_f and σ_b represents the standard deviation. The final CNR is an arithmetic mean of the CNR from each region.

To evaluate the SSIM, we use

$$SSIM(g,y) = \frac{(2\mu_g\mu_y + C_1)(2\sigma_{gy} + C_2)}{(\mu_g^+\mu_y^2 + C_1)(\sigma_g^+\sigma_y^2 + C_2)} \qquad (13)$$

where g is the averaged image we use as the ground truth, y is the denoised image, C_1 and C_2 are constant terms added to avoid instability.

G. Qualitative Analysis

We also evaluate the performance of the proposed method using visual comparison, looking for a balance between noise reduction and loss in spatial resolution, as well as distinction between the different retinal layers.

III. Results

A. Quantitative Analysis

Table 1 shows the average CNR and SSIM across all five images before and after noise reduction with the proposed method. Overall, there was an improvement in both CNR and SSIM after noise reduction with the proposed method. The best performance was obtained with the combination of a U-Net network architecture with $\mathcal{L}_{L_2\_edge}$.

TABLE I
AVERAGE QUANTITATIVE METRICS

Network	Loss	CNR	SSIM
None	None	1.780	0.079
U-Net	$\mathcal{L}_{L_2}$	7.347	0.519
U-Net	$\mathcal{L}_{L_1}$	5.985	0.435
U-Net	$\mathcal{L}_{L_2\_L_1}$	6.426	0.507
U-Net	$\mathcal{L}_{L_2\_tv}$	8.295	**0.526**
U-Net	$\mathcal{L}_{L_2\_edge}$	**10.254**	0.510
Deep Decoder	$\mathcal{L}_{L_2\_tv}$	4.783	0.393
Deep Decoder	$\mathcal{L}_{L_2\_edge}$	7.807	0.492

Network = None and **Loss** = None indicate the metrics of the original noisy image. The best metrics are shown in **bold**.

B. Qualitative Analysis

Fig. 3 shows an example of images before and after noise reduction with the proposed method using different network architectures and loss functions. Overall, there is visibly less noise in the resulting images using the U-Net architecture with $\mathcal{L}_{L_2\_tv}$ or $\mathcal{L}_{L_2\_edge}$, whereas without the additional $tv(y)$ and $edge(y)$ priors, the resulting images still look considerably noisy. The different retinal layers are still distinguishable from each other.

IV. Discussion

Overall, the U-net architecture provides better noise reduction, and while all the loss functions improve the image quality, $\mathcal{L}_{L_2\_tv}$ or $\mathcal{L}_{L_2\_edge}$ performed the best. However, these loss functions did not necessarily always perform the best across all the images. Currently, one of the biggest challenges is determining when to stop the optimization. Traditional image metrics are poorly suited for determining overall quality and noise reduction. In future work, alternative evaluation methods such as accuracy of downstream tasks (e.g. automatic retinal layer segmentation) could be used. Additionally, the loss functions used could be tailored to be more specific to OCT. For example, instead of simply optimizing towards a global smoothness, the specific noise and retinal tissue structures could be incorporated.

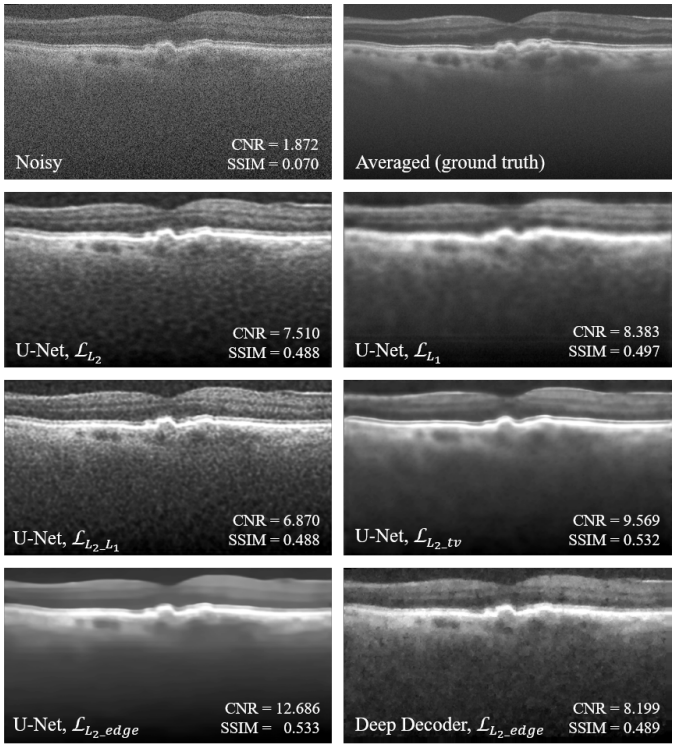

Fig. 3. Example of images before and after noise reduction with the proposed method using different network architectures and loss functions.

References

[1] J. A. Izatt, M.A. Choma, "Theory of Optical Coherence Tomography," W. Drexler, J. G. Fujimoto (eds), Optical Coherence Tomography, Biological and Medical Physics, Biomedical Engineering, Springer, Berlin, Heidelberg, 2008.

[2] J. W. Goodman, "Speckle Phenomena in Optics: Theory and Applications," Roberts and Company Publishers, 2007.

[3] M. Pircher, E. Gtzinger, R. A. Leitgeb, A. F. Fercher, C. K. Hitzenberger, "Speckle reduction in optical coherence tomography by frequency compounding," *J. Biomed. Opt.* **8**(3), 2003.

[4] A. Ozcan, A. Bilenca, A. E. Desjardins, B. E. Bouma, G. J. Tearney, "Speckle reduction in optical coherence tomography images using digital filtering," *J Opt Soc Am A Opt Image Sci Vis* **24**(7), 2007, pp. 1901–1910.

[5] L.Fang, S. Li, Q. Nie, J. A. Izatt, C. A. Toth, and S. Farsiu, "Sparsity based denoising of spectral domain optical coherence tomography images," *Biomed. Opt. Express* **3**, 2012, pp. 927–942.

[6] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation,"in *International Conference on Medical Image Computing and Computer-Assisted Intervention,* 2015, pp. 234–241.

[8] R. Heckel, and P. Hand, "Deep decoder: Concise image representations from untrained non-convolutional networks," *arXiv preprint arXiv:1810.03982*, 2018.

[9] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena* **60**(1-4), 1992, pp.259–268.
[10] J. M. Prewitt, "Object enhancement and extraction," *Picture Processing and Psychopictorics* **10**(1), 1970, pp.15–19.
[11] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.