Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2020 February ; 11316: . doi:10.1117/12.2549687.

Deep Learning Based Model Observer by U-Net

Iris Lorente^a, Craig K. Abbey^b, Jovan G. Brankov^a

^aECE Department, Illinois Institute of Technology, Chicago, IL, USA 60616

^bDepartment of Psychological and Brain Sciences, University of California, Santa Barbara CA 93106

Abstract

Model Observers (MO) are algorithms designed to evaluate and optimize the parameters of new medical imaging reconstruction methodologies by providing a measure of human accuracy for a diagnostic task. In contrast with a computer-aided diagnosis system, MOs are not designed to outperform human diagnosis but only to find a defect if a radiologist would be able to detect it. These algorithms can economize and expedite the finding of optimal reconstruction parameters by reducing the number of sessions with expert radiologists, which are costly and prolonged. Convolutional Neural Networks (CNN or ConvNet) have been successfully used in the computer vision field for image classification, segmentation and video analytics. In this paper, we propose and test several U-Net configurations as MO for a defect localization task on synthetic images with different levels of correlated noisy backgrounds. Preliminary results show that the CNN based MO has potential and its accuracy correlates well with that of the human.

Keywords

Model observer; medical image quality assessment; machine learning; deep learning; ConvNet; U-Net

1. INTRODUCTION

In medical imaging, diagnosis using imaging systems has developed a great deal with the advance of technology. Nevertheless, there is an ongoing need to perform new research on image reconstruction algorithms and devices, which should be properly tested and validated. Evaluating image quality, a critically important step of image science, has been traditionally performed by using numerical figures of merit such as the signal to noise ratio (SNR) or bias-variance trade off [1]. These images can be successful in accurately representing the shape and texture of the human body elements while failing to provide a good sight for a human reader (a radiologist) to perform a diagnostic task [2]. For this reason, it is commonly accepted that medical imaging quality should be assessed by maximizing human accuracy in a specific diagnostic task. For example, human accuracy in detecting micro-calcifications in breast mammograms or perfusion defects in SPECT myocardial images. This approach has become known as task-based image quality assessment [1]. An expert reader is given some simulated images reconstructed by the method that is being evaluated. Some of these images contain a defect that should be detected or localized by the expert. In a localization task, the reader is asked to point at the defect in each image and an accuracy metric can be calculated

for the reconstruction method that is being evaluated. This methodology requires constantly organizing reader sessions with expert humans, such as radiologists which are costly and time consuming.

The use of Model Observers (MO), algorithms capable of predicting human observer performance acting as human surrogates, is becoming increasingly important since it reduces the number of required reader sessions and makes the quality assessment process more efficient. The channelized Hotelling observer (CHO) [1, 6] is the most popular MO for defect detectability since it correlates well with the human observer when proper channels are selected to obtain image features. The rationale behind the channels' system for a model observer stems from studies showing that the human visual system contains a large amount of spatial frequency specific channels [7, 8]. The CHO can be viewed as a generalized likelihood ratio test. In some cases, CHO performance overestimates human performance so an internal noise model is introduced, which allows the CHO to be tuned to match human observer performance [9, 10]. In recent studies, we have successfully replaced the CHO by a machine-learning model observer (LMO) for prediction of human observer detection performance for cardiac SPECT defects [11, 12]. In addition, we have extended the model approach to diagnostic tasks other than defect detection such as motion defect detection and defect localization [13, 14]. In the recent years, the use of deep learning algorithms has become very popular within the machine learning community with successful performance in applications such as computer vision and speech recognition. Adopting this trend, the medical imaging community started employing convolutional networks to some tasks including image reconstruction (e.g. dose reduction) and computer aided diagnosis applications. In addition, some studies have been successful in applying neural networks for a MO detection task [15–17]. In this study, we use a convolutional network as a MO for a defect localization task with synthetic data, specifically the U-Net architecture. A U-Net is a fully convolutional network that consists of a contracting path -that captures context- and a symmetric expanding path that allows precise pixel localization (see Figure 2). In particular, we train six U-net models, each with a different kernel size for the convolutional layers and two different loss functions; and test how the different kernel sizes affect the performance of the U-Net as a MO.

2. DATASET

The images in our dataset [18] are synthetically generated and consist of a 4 mm smooth disk (that simulates a defect) added to three random noise backgrounds obtained by filtering white noise by the following power-law function:

$$S(u, v) = C_{\beta}(u^2 + v^2)^{-\beta/2} \quad (1)$$

where β takes values 0, 1 and 2; and C_{β} was chosen such that the contrast caused the human reader to have an AUC of 0.8 in a detection task (see Figure 1 for an example of all three backgrounds). The complete dataset contains 2000 noise realizations from each one of the three backgrounds ($\beta = 0, 1, 2$), a total of 6000 images, with a resolution of 256×256 pixels. Since we are focusing in a localization task, all the images have a simulated defect in only one location and the location varies from image to image (between noise realizations of the

background). For every image, a human reader was asked to mark the location where he believed there was a defect. Therefore, another dataset was created by generating images of the same size (256×256 pixels) with no background and a simulated defect in the location marked by the human reader. This dataset is referred as “labels” and it is what the MO is aiming to learn. Similarly, a third dataset referred as “the truth” was generated by placing a Gaussian impulse in the location where the “defect” actually is, and will be used to compare the human accuracy with the MO accuracy.

3. METHODOLOGY

3.1 The U-Net architecture

The goal of this study is to train a model that, given a noisy image with a simulated lesion, outputs the location where the human reader would think the simulated lesion is. To do so, a non-residual U-net CNN (Figure 2) is trained with the noisy images as input and the label images as described in the previous section as output (Figure 3). The architecture goes as follows. First, the input image (size $256 \times 256 \times 1$ since the images are in gray scale) is passed through a convolutional layer with 3×3 kernel size and 64 channels followed by a batch normalization and ReLU activation function; this results in a $256 \times 256 \times 64$ volume. Following, the volume is passed through a second identical convolutional layer outputting a new volume of size $256 \times 256 \times 64$. Next, the resolution is decreased with an averaging 2×2 pooling layer resulting in a new volume with size $128 \times 128 \times 64$ and so on. In the expanding path (right side), convolution and unpooling layers increase image resolution and the resulting output has equal size as the input, $256 \times 256 \times 1$. The main difference between an encoder/decoder and the U-Net is the introduction of “skip connections” that during the expanding path, provide information on how the volume looked at that same resolution level before contracting. Specifically, after an unpooling layer, the resulting volume is concatenated with its symmetric one before contracting, which results in doubling the number of channels (e.g. in the highest resolution level, the resulting $256 \times 256 \times 64$ volume from the contracting side is concatenated to that of the expanding side resulting in a $256 \times 256 \times 128$ volume, succeeding convolution layers reduce the number of channels to 1).

3.2 Performance evaluation

To evaluate the models, the localization accuracies of the human in the test set are compared to the localization accuracies of the MO for each of the three random noise backgrounds. We define “accuracy” as the ratio between number of images in which the Human (or MO) localizes well the defect and the total number of images. A defect is considered to be well located if it is placed within 5 pixels of the actual defect location.

Since the models trained are regression CNNs, it is possible that the output of a test image contains more than one highlighted location. In those cases, the location of the pixel with the highest intensity is selected as the Model Observer’s response to that input image. The statistical significance of the human and MO accuracies can be evaluated by performing a McNemar chi square hypothesis test [19].

4. EXPERIMENTS

Two configurations of U-Net and two different loss functions are explored in this study. The first configuration, using 3×3 kernels in the convolutional layers as described in Section 3.1; the second configuration uses 5×5 convolutional kernels. The kernel size determines how big is the local area where features are extracted (linear combination of pixels).

From each generated background (defined in Section 2), 1000 images are used to train and validate the convolutional network while the remaining 1000 are used to test the models. This totals 3000 images for training/validation purposes from which 2100 are randomly picked to train the model and the remaining 900 are used for validation. The validation set is important to make sure the CNN is not over-fitted as well as to optimize the parameters of the network.

The two models are trained by minimizing a loss function between the labels and predictions using Python Keras in a NVIDIA GeForce GTX 1080 Ti GPU. The two loss functions that we explore are the Mean Squared Error (MSE) and Binary Cross Entropy (CE). In the case of the CE, the last ReLU activation function is substituted by a Sigmoid and the target images are rescaled to be in the (0,1) range. The ConvNet is trained by using Adam optimizer with learning rate decay and the optimization is considered final when the loss function in the validation set has not decreased during 50 consecutive epochs.

5. RESULTS

Figure 4 and Figure 5 show the accuracies of the human reader (HO) in the 1000 images of each background and the accuracies of the two trained MOs in the same sets for loss functions MSE and CE respectively. The accuracy of the human observer is pretty consistent at 80% for the three backgrounds (the targeted level of performance in the psychophysical study). The two MOs trained by minimizing the MSE (Figure 4) outperform the human's localization accuracy in all three backgrounds. The MOs trained by minimizing the binary cross entropy yield better results providing localization accuracies closer to those of the human. Statistical analysis by McNemar's chi square test to compare the human accuracies with the trained models, resulted in p-values < 0.05 except for the case of the MSE trained MO with 5×5 filters for $\beta = 1$ (p-value = 0.53); the CE trained MOs with 3×3 filters for $\beta = 1$ (p-value = 0.14) and 5×5 filters with $\beta = 0$ (p-value = 0.07).

Figure 6 shows the averaged noise fields of the human and the MOs with 5×5 kernel configurations for both MSE and CE that were incorrectly localized, each one centered in the marked location by the human before averaging [20]. It can be observed that the models trained with a CE loss function provide templates more similar to the human's than the models trained by minimizing MSE.

6. CONCLUSION

Overall, the presented results show potential in training a CNN based Model Observer for a defect localization task. The results suggest that the CNN based MO trained by minimizing the binary Cross Entropy (CE) performs closer to the human in terms of localization

accuracy. This is of no surprise since the Mean Squared Error (MSE) is not an ideal metric for an optimization problem when the dataset is very unbalanced, as is our case with the number of pixels in the background being much higher than the number of pixels in the simulated defect. With regard to the different kernel sizes, there is not a common agreement of which kernel size is the optimal. The optimal kernel size could depend on each particular background as well as on the size of the simulated lesion.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Health (NIH) under grants EB023969, HL065425, HL091017, and EB025829.

REFERENCES

- [1]. Barrett HH and Myers KJ, Foundations of image science. Wiley-Interscience, 2004.
- [2]. Lusted LB, "Signal detectability and medical decisionmaking," Science, vol. 171, p. 7,1971.
- [3]. Peterson WW, Birdsall TG, and Fox WC, "The theory of signal detectability," Proceedings of the IRE Professional Group on Information Theory, vol. 4, pp. 171–212,1954.
- [4]. Metz CE, "Basic principles of ROC analysis," Seminar Nucl. Med, vol. 8, pp. 283–298, 1978.
- [5]. Swets JA and Pickett RM, Evaluation of diagnostic systems: Methods from signal detection theory. 1982.
- [6]. Myers KJ and Barrett HH, "Addition of a channel mechanism to the ideal-observer model," Journal of the Optical Society of America A, vol. 4, no. 12, pp. 2447–2457,1987.
- [7]. Barrett HH, Yao J, Rolland JP, and Myers KJ, "Model observers for assessment of image quality," presented at the Images of Science: Science of Images, Washington, DC, 1993.
- [8]. Sachs MB, Nachmias J, and Robson JG, "Spatial-Frequency Channels in Human Vision," Journal of the Optical Society of America, vol. 61, no. 9, pp. 1176–1186,1971. [PubMed: 5121888]
- [9]. Yao J and Barrett HH, "Predicting human performance by a channelized Hotelling observer model," in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Wilson DC and Wilson JN, Eds., 1992/12 1992, vol. 1768, pp. 161–168. [Online]. Available: <http://adsabs.harvard.edu/abs/1992SPIE.1768..161Y>
- [10]. Brankov JG, "Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection," Phys Med Biol, vol. 58(20), 2013.
- [11]. Kalayeh MM, Marin T, and Brankov JG, "Generalization evaluation of machine learning numerical observers for image quality assessment," IEEE Transactions on Nuclear Science, vol. 60 (3), p. 10,2013.
- [12]. Brankov JG, Yang Y, Wei L, El Naqa I, and Wemick MN, "Learning a channelized observer for image quality assessment," IEEE Transactions on Medical Imaging, vol. 28, no. 7, 2009.
- [13]. Marin T, Kalayeh MM, Parages FM, and Brankov JG, "Numerical surrogates for human observers in myocardial motion evaluation from SPECT images," IEEE Transactions on Medical Imaging, vol. 33, no. 1, pp. 3847, 2014. [PubMed: 23981533]
- [14]. Parages FM, O'Connor JM, Pretorius PH, and Brankov JG, "A Naive-Bayes Numerical Surrogate for a Human Observer in Perfusion Detection and Localization in SPECT-MPI," presented at the IEEE Nucl. Sci. Symp. and Medical Imag. Conf. Rec., 2013.
- [15]. Kopp FK, Catalano M, Pfeiffer D, Rummeny EJ, and Noel PB, "Evaluation of a machine learning based model observer for x-ray CT," 2018, vol. 10577: International Society for Optics and Photonics, p. 105770S.
- [16]. Gong H et al., "A deep learning-and partial least square regression-based model observer for a low-contrast lesion detection task in CT," Medical physics, vol. 46, no. 5, pp. 2052–2063,2019. [PubMed: 30889282]

- [17]. Kim G, Han M, Shim H, and Baek J, "Implementation of an ideal observer model using convolutional neural network for breast CT images," 2019, vol. 10952: International Society for Optics and Photonics, p. 109520L.
- [18]. Abbey CK and Eckstein MP, "High human-observer efficiency for forced-localization tasks in correlated noise," 2010, vol. 7627, p. 76270R.
- [19]. McNemar Q, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947. [PubMed: 20254758]
- [20]. Abbey CK, Samuelson FW, Zeng R, Boone JM, Eckstein MP and Myers K, "Classification images for localization performance in ramp - spectrum noise," *Medical physics* vol. 45, no. 5, pp 1970–1984, 2018. [PubMed: 29532479]

(a) $\beta = 0$ (b) $\beta = 1$ (c) $\beta = 2$

Figure 1. Examples of images from the three different backgrounds.

Figure 2. Standard U-Net architecture. In between blocks, we perform a 3×3 convolution, batch normalization and ReLU

(a) Input: Simulated background with an added signal

(b) Output: No background and signal in the location that the human marked

Figure 3. Example of an input/output pair used to train the model.

Figure 4. Accuracies of the Human Observer (HO), and Model Observers (MO) with U-Net configurations of kernel sizes 3×3 and 5×5 with a Mean Squared Error (MSE) loss function.

Figure 5. Accuracies of the Human Observer (HO), and Model Observers (MO) with U-Net configurations of kernel sizes 3×3 and 5×5 with a Cross Entropy (CE) loss function.

Figure 6.

Average of noise fields of the Human Observers and Model Observers with kernel size 5×5 for incorrectly localized images.