A deep learning model observer for use in alterative forced choice virtual clinical trials

M. Alnowami, G. Mills, M. Awis, P. Elangovanr, M. Patel, et al.

M. Alnowami, G. Mills, M. Awis, P. Elangovanr, M. Patel, M. Halling-Brown, K. C. Young, D. R. Dance, K. Wells, "A deep learning model observer for use in alterative forced choice virtual clinical trials," Proc. SPIE 10577, Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, 105770Q (7 March 2018); doi: 10.1117/12.2293209

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

A Deep Learning Model Observer for use in Alterative Forced Choice Virtual Clinical Trials

M Alnowami[a,c], G Mills[a,d], M Awis[a], P Elangovanr[b], M Patel[b], M Halling-Brown[b], KC Young[b,d], DR Dance[b,d], and K Wells[a]

[a]CVSSP, University of Surrey, Guildford, GU2 7XH, UK
[b]Royal Surrey County Hospital, Guildford, Surrey, GU2 7XX, UK
[c]Nuclear Engineering Department, King Abdulaziz University, 80204 Jeddah 21589, SA
[d]Department of Physics, University of Surrey, Guildford, GU2 7XH, UK

ABSTRACT

Virtual clinical trials (VCTs) represent an alternative assessment paradigm that overcomes issues of dose, high cost and delay encountered in conventional clinical trials for breast cancer screening. However, to fully utilize the potential benefits of VCTs requires a machine-based observer that can rapidly and realistically process large numbers of experimental conditions. To address this, a Deep Learning Model Observer (DLMO) was developed and trained to identify lesion targets from normal tissue in small (200 x 200 pixel) image segments, as used in Alternative Forced Choice (AFC) studies. The proposed network consists of 5 convolutional layers with 2x2 kernels and ReLU (Rectified Linear Unit) activations, followed by max pooling with size equal to the size of the final feature maps and three dense layers. The class outputs weights from the final fully connected dense layer are used to consider sets of n images in an n-AFC paradigm to determine the image most likely to contain a target. To examine the DLMO performance on clinical data, a training set of 2814 normal and 2814 biopsy-confirmed malignant mass targets were used. This produced a sensitivity of 0.90 and a specificity of 0.92 when presented with a test data set of 800 previously unseen clinical images. To examine the DLMOs minimum detectable contrast, a second dataset of 630 simulated backgrounds and 630 images with simulated lesion and spherical targets (4mm and 6mm diameter), produced contrast thresholds equivalent to/better than human observer performance for spherical targets, and comparable (12 % difference) for lesion targets.

Keywords: Deep learning, model observer, simulation, lesion, mammography, virtual clinical trial.

1. INTRODUCTION

Modern screening for breast cancer has undergone major technological development with the move away from film-screen to digital and tomosynthesis technologies. However, the traditional clinical trial, has remained the standard method of technology assessment. These involve 1000s of participants and represents a major bottleneck to delivering new innovations into clinical practice. This has motivated the development of *in-silico* methods, collectively known as virtual clinical trials (VCTs) [1]. However, to take full advantage of the VCT approach, model observers are needed that can replicate human performance for particular tasks. This facilitates exploration of large numbers of experimental conditions that can outstrip what is practicable with limited expert resources, in terms of time and observer numbers usually available for such studies.

Further author information: (Send correspondence to M. Alnowami)
M. Alnowami.: E-mail: majdi.alnowami@surrey.ac.uk, Telephone: +44 1483 68 9854

Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, edited by Robert M. Nishikawa, Frank W. Samuelson, Proc. of SPIE Vol. 10577, 105770Q
 · doi: 10.1117/12.2293209

Human and model observers have been used as sophisticated tools for determining image quality, in this context referring to the ability of an image to deliver useable content for a particular task [2]. One such approach is the use of *n*-alternative forced choice studies where the observer must decide which of n images contains a particular target. Such studies may be used to determine, for example, detection limits under controlled conditions such as threshold size or contrast, without requiring image search. Such an approach can be used to compare detection performances of different breast imaging modalities in radiological studies or virtual clinical trials (VCTs) [3,4]. The principal advantage of a model observer (MO) is the ability to reliably, and tirelessly, undertake evaluations of sets of images without recourse to the variables introduced by using groups of human observers. This approach is routed in statistical decision theory and linear discriminant analysis and has been an active area of research for many years [2,3,5].

Recent developments in the field of deep learning, an emerging area of artificial intelligence, have enabled machines to also capture the subtle inter-play of information at different scales within an image, and facilitated machine-led understanding of features-of-features [6]. Such approaches have routinely beaten state-of-the-art pattern recognition methods, as well as humans in some selected areas. In SPECT, supervised learning and Convolutional Neural Networks (CNNs) with three layers have shown promising MO results with simulated data [7,8].

This has motivated our use of deep learning for developing a new data driven AFC model observer for use in mammography, referred to as a Deep Learning Model Observer (DLMO).

2. METHODOLOGY AND PRELIMINARY RESULTS

For initial convenience, our model observer has been developed based on a signal-known-statistically (SKS) task rather than the usual signal-known-exactly (SKE) task used in AFC studies.

2.1 Model Observer

The MO architecture used here makes use of a deep CNN with five convolutional layers. The initial network consists of five convolutional layers with 2x2 kernels and ReLU (Rectified Linear Unit) activations, followed by average pooling with size equal to the size of the final feature maps and three dense layers (see Fig 1). The ReLU nonlinearity is applied to the output of each convolutional layer. There are two max pooling layers one after the first ReLU and another after the second ReLU. Average pooling is applied to the output of the final convolutional layers after passing it through a ReLU layer. Output of this average pooling layer is fed to a fully connected layer and a binary classification loss layer is added as an objective to train the Deep CNN. The network is fed with images of 200x200 pixels. The first convolutional layers have 32 kernels of size 2x2 and a stride of 1 in both *x* and *y* direction.

The activations of the first layer are passed through the ReLU and a max pooling layer of kernel size 2x2 and stride 2. The second convolutional layer has 64 kernels of size 4x4 and a stride of 2. The activation of the second layer are passed through a further ReLU and a max pooling layer of kernel size 2x2and stride 2. The third convolutional layer has 96 kernels of size 2x2 and a stride of 1 and a ReLU. The fourth convolutional layer has 128 kernels of size 2x2 and a stride of 1 and a ReLU. The fifth convolutional layer has 256 kernels of size 3x3 and a stride of 1 and a ReLU. The output of the ReLU is fed to the average pooling layer which is connected to a fully connected layer of size 2.

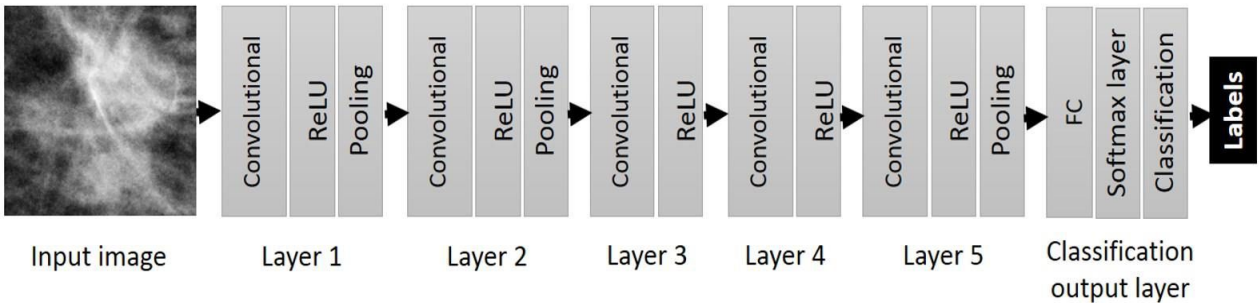

Figure 1. Schematic representation of CNN architecture used for Deep Learning Model Observer. The DLMO architecture is a deep CNN consisting of five convolutional layers. There are two max pooling layers one after first ReLU and another after the second ReLU. Average pooling is applied to the output of the final Convolutional layers after passing it through ReLU. Output of this average pooling layer is fed to fully connected layer and a binary classification loss layer is added as an objective to train the Deep CNN.

For our implementation we have used a Python wrapper for the CAFFE deep learning library. The training was performed on GPU servers with 7 GeForce GTX Titan X (Maxwell compute 5.2) single precision GPUs with 256 GB RAM, running Ubuntu 16.04. However, the GPU servers have been used for training only.

2.2 Preliminary training and test data

2.2.1 Training and testing using screen detected lesions

The DLMO has been initially trained to discriminate images representing biopsy-confirmed malignant lesions from normal tissue, using small image segments as used in AFC studies. An initial clinically derived dataset consisted of 5628 14 x 14 mm^2 image patches extracted from clinical screening mammograms, corresponding to 200 pixels x 200 pixels x 70μm (pitch), where 2814 images represent normal tissue and 2814 images contain biopsy-confirmed screen-detected malignant masses representing a mixture of well-defined and ill-defined lesions. These images were originally acquired using Hologic Selenia imaging systems based at several sites and remotely archived within the OPTIMAM image Database [9].

Training was undertaken by carefully controlling the learning rate using a stochastic descent algorithm [6]. The intrinsic discrimination performance was examined by calculating the sensitivity and specificity associated with particular values of the class output weight from the training dataset to discriminate between lesion presence and lesion absence. From this an ROC curve was constructed to determine the intrinsic performance.

A test data set was created by randomly separating 30% of the above aggregated data set prior to training so that no test images were seen prior to test. The remaining data were used for augmentation and training. The dataset was augmented to facilitate adequate training of the model and to remove sensitivity to orientation. This was achieved using two random rotations per image by angles in the range $(0, 360^o)$. To avoid undefined regions, prior to rotation the images were re-sampled to 300x300 pixels, rotated and then resampled back to 200x200 pixels.

2.2.2 Training and testing using simulated data at minimum detectable contrast

In order to further examine the DLMOs performance, now with respect to minimum detectable contrast, we used a 4-AFC assessment method [10] and the OPTIMAM VCT toolbox [11] to produce a carefully curated image

dataset of 630 images that spanned the limits of human target detectability. The dataset consisted of 630 images of background tissue only and 630 images containing uniform spheres and simulated lesions [12] of 4 and 6 mm diameter embedded in a validated breast phantom. Local glandularity texture was used to control contrast. These models were then passed through an image simulation chain and post-processed to represent images acquired on a Hologic Selenia Dimensions mammography imaging system with a mean glandular dose of 2.5mGy.

To adapt the DLMO for use in AFC trials, the training data set was augmented by factor 2 using the same methods as for the clinical dataset and the DLMO re-trained. The class outputs weights from the final fully connected dense layer which were used to consider sets of 4 images in an n-AFC paradigm. Each image from each AFC set of 4 images was sequentially extracted and passed through the network to determine an output weight for the target class compared to its neighbors. The image with the strongest weight for the target class from the set was assigned a target-present label, mirroring the decision task for human observer behavior.

To compare the DLMO with human performance, two groups of human observers were used: a specialist observer group, consisting of five medical physicists, four radiologists and two radiographers with experience reading breast screening images, and a non-specialist group drawn from members of the public with no experience of reading mammography data [4]. Prior to the AFC trial, the non-specialists were given training with feedback for 30 minutes to familiarize themselves with the task. No time limit was imposed on the human observers. The percent correct for each of the test conditions for human observers and the DLMO were calculated and from this, threshold contrast detection values were calculated for 90.6% correct, which corresponds to a value of 2.5 for the detectability index d. This facilitated comparison of DLMO performance with different levels of human observer performance.

2.3 Learning Rate

Training was undertaken by carefully controlling the learning rate using stochastic descent algorithm (equation 1). This updates the parameters (weights and biases) so as to minimize the error function by taking small steps in the direction of the negative gradient of the loss function [1]. This is described by

$$\Theta_{j+1} = \theta_{\cdot} - \alpha \nabla E(\theta_j) \tag{1}$$

where j stands for the iteration number, $\alpha > 0$ is the learning rate, θ is the parameter vector, and $E(\theta_j)$ is the loss function. The gradient of the loss function, $\nabla E(\theta_j)$, is evaluated using the entire training set, and the standard gradient descent algorithm uses the entire data set at once. The initial weights used are drawn from a Gaussian distribution with a mean of 0 and a standard deviation of 0.01. The default for the initial bias value is 0.

The stochastic gradient descent algorithm evaluates the gradient, hence updates the parameters, using a subset of the training set. This subset is called a mini-batch. Each evaluation of the gradient using the mini-batch is an iteration. At each iteration, the algorithm takes one step towards minimizing the loss function. In this study, in order to optimize the training time, an initial training rate was 0.0001 at the beginning of training and gradually reduced by a factor of 0.2 every 5 epochs with a maximum number of epochs for training to 90, and use a mini-batch with 64 observations at each iteration. This will enable smaller steps towards to optimum value as the training progresses, hence a finer search towards the end of the training.

As a result of this, the accuracy, defined as the percentage correct for lesion detection was calculated at each iteration. As can be seen in Figure 2, the system is optimized after 30 iterations, giving an overall accuracy of 92.3 ± 1.8%. The figure displays training accuracy at every iteration. Each iteration is an estimation of the gradient and an update of the network parameters.

2.4 Discrimination Performance for Screen-detected Lesions

Training of the DLMO network via stochastic decent using the clinical data set was assessed by plotting accuracy against iteration number as mentioned in the previous section. Intrinsic performance was then assessed by examining the lesion-present output weight to produce an ROC curve as show in Figure 3 based on the screen detected clinical training data. The mean area under the ROC curve (AUC) is 96.4%. The red dot in the figure indicates the optimal operating point on the ROC curve. The threshold that corresponds to the optimal operating point is equal to 0.67.

The overall performance of the DLMO when assessed against the unseen test data produced an average sensitivity of 90% and a specificity of 92%.

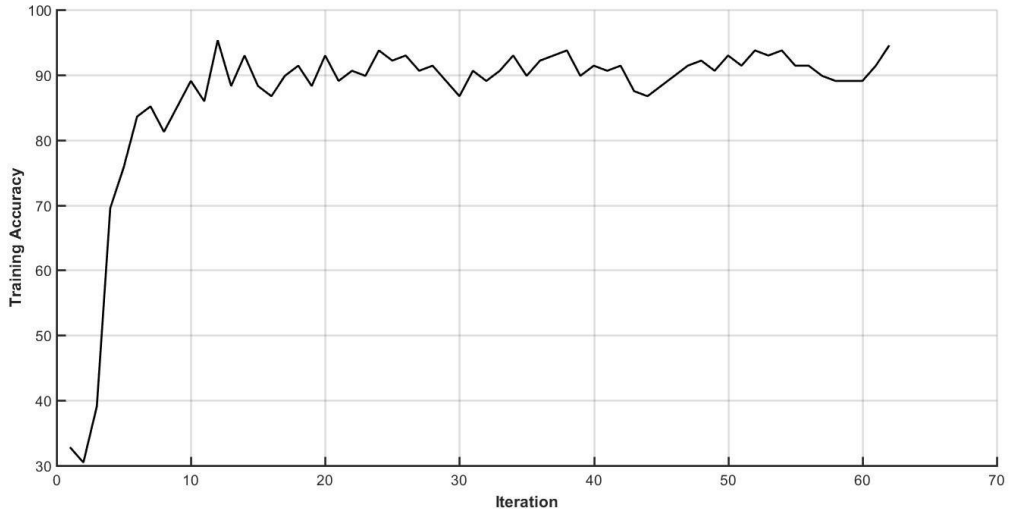

Figure 2. Plot of accuracy vs iteration number during training. The figure displays training accuracy at every iteration. Each iteration is an estimation of the gradient and an update of the network parameters. This shows that the network has converged after 15-20 iterations.

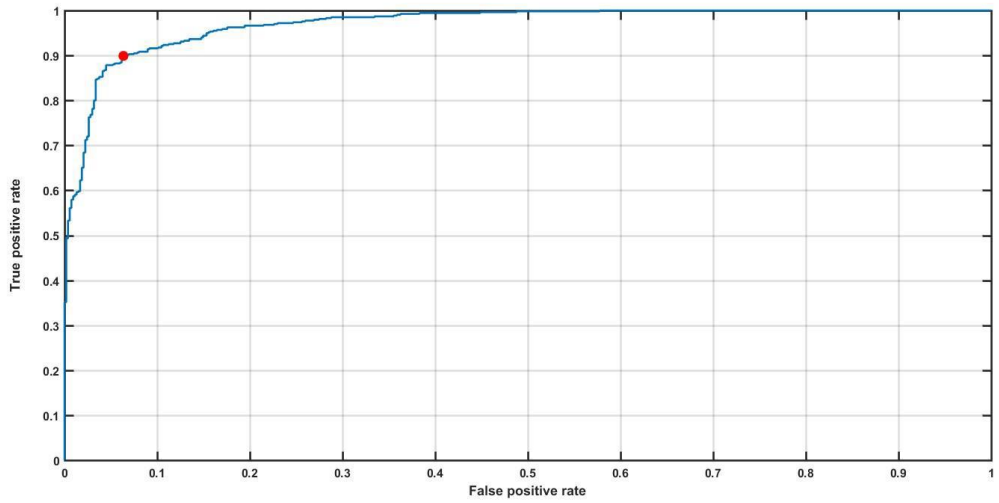

Figure 3. ROC performance for the lesion present class output weight. The mean area under the ROC curve (AUC) is approximately 96%. The red dot in the figure indicates the optimal operating point on the ROC curve.

2.5 Threshold Contrast Comparison with Human Observers

DLMO and human performances for threshold contrast for 4mm and 6mm spherical and lesion targets are shown in Figure 4. The threshold contrast values for the DLMO across all four targets broadly follow the same relative behavior as both sets of human observers, although statistically, each of these results falls within the estimated error of one another. In the case of spherical targets, the DLMO performance is similar or slightly better than both sets of human observers. For lesions the preliminary data suggests that the DLMO has a slightly higher threshold contrast than the human observers. This also demonstrates that specialist observers have in general slightly better performance than non-specialists.

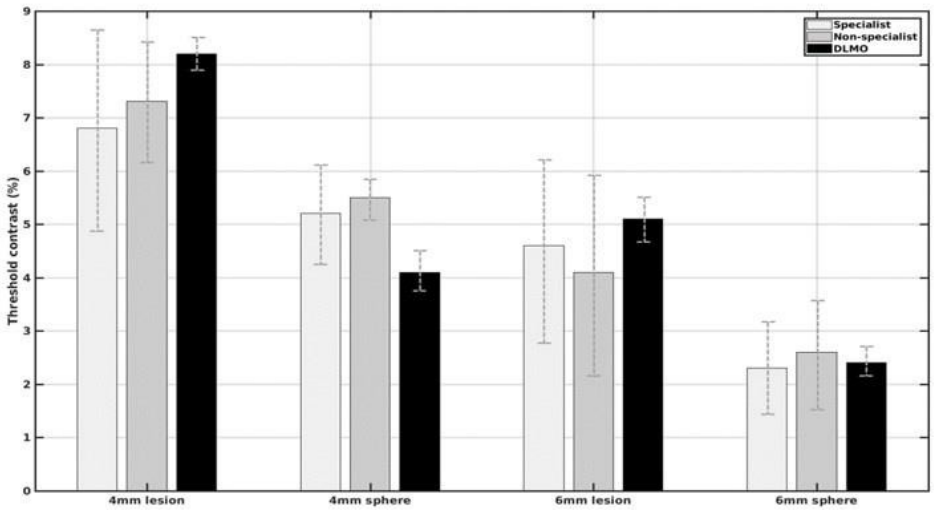

Figure 4. Comparison of threshold contrast limit for the DLMO compared to human observers.

3. CONCLUSIONS

These results from our data-driven DLMO demonstrate promising performance that is comparable to human observers for lesion detection in mammography. The performance with the clinical dataset suggests that a variety of mass morphologies can be successfully identified in a screening context. However, tasks that span the limits of human perception are significantly more challenging. In this respect, the 4AFC trial demonstrates that this initial implementation of the DLMO has performance comparable to humans, and that may be equal or better than humans for simple spherical targets.

For initial convenience, our model observer has been developed based on a signal-known-statistically (SKS) task rather than signal-known-exactly (SKE). However. prior work suggests that there is little difference in human observer performance between SKE and SKS tasks [3].

The training regime used here has only used a fraction of the clinical data available in the OPTIMAM image database, and moreover, using the OPTIMAM VCT toolbox, there is tremendous opportunity to generate much larger contrast-controlled data volumes. This is the subject of on-going development.

References

[1] ADA Maidment (2014) Virtual clinical trials for the assessment of novel breast screening modalities. Proc IWDM 14. 8539, 1-8.

[2] X He & S Park (2013) Model Observers in Medical Imaging Research. Theranostics; 3(10): 774-786

[3] C Castella et al (2009) Mass detection on mammograms: influence of signal shape uncertainty on human and model observers J. Opt. Soc. Am. A 26, 425-436

[4] P Elangovan et al (2017); Using non-specialist observers in 4AFC human observer studies Proc. SPIE 10132, Medical Imaging: Physics of Medical Imaging, 1013256; doi:10.1117/12.2255560.

[5] HH Barrett et al (1993) Model observers for assessment of image quality. Proceedings of the National Academy of Sciences of the United States of America.90(21): 9758-9765.

[6] Y LeCun et al (2015); Deep Learning (Review); Nature 521, 436444 doi:10.1038/nature14539

[7] JG Brankov et al (2009) Learning a channelized observer for image quality assessment, IEEE Transactions on Medical Imaging 28 (7), pp. 991–999

[8] **F Massanes**, **JG. Brankov**, Evaluation of CNN as anthropomorphic model observer, Proc. SPIE 10136, Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment, 101360Q (10 March 2017); doi: 10.1117/12.2254603

[9] MN Patel et al (2016) Collection of sequential imaging events for research in breast cancer screening; Proc. SPIE 9789, Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations, 97890 doi:10.1117/12.2216648.

[10] AE Burgess (1995) Comparison of receiver operating characteristic and forced choice observer performance measurement method, Med. Phys. 22(5), 643-655

[11] P Elangovan et al (2017) Design and Validation of Realistic Breast Tissue Models for Use in Virtual Clinical Trials Phys. Med. Biol. 62,22782794.

[12] A Rashidnasab et al (2013) Simulation and assessment of realistic breast lesions using fractal growth models Phys. Med. Biol. 15, 5613-5626.