Reading assignment - paper review

The paper chosen to be reviewed for this assignment is *N. Wu et al., (2020). Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening. IEEE Transactions on Medical Imaging, vol. 39, no. 4, pp. 1184-1194, doi:10.1109/TMI.2019.2945514.*

The paper focuses on the development of a deep convolutional neural network for breast cancer screening exam classification which was trained and evaluated on over 200 000 exams. It introduces a novel two-stage neural network meant to incorporate global and local information with an appropriate training procedure which allows the usage of a very high-capacity patch-level netowork to learn from pixel-level labels alongside a network learning from microscopic breast-level labels.

This article introduces a novel variant of a ResNet specifically designed for medical imaging which has a balance of depth and width that allows the model to process a very large image while maintaining reasonable memory consumption.

The images in the dataset are coming from four types of scanners: Mammomat Inspiration (22.81%), Mammomat Novation DR (12.65%), Lorad Selenia (40.92%) and Selenia Dimensions (23.62%).

In some cases, breasts contain both malignant and benign findings therefore they formulate breast cancer screening classification as a learning task using a multi-task learning framework.

For each breast, there are two binary labels assigned: the absence/presence of malignant findings in a breast and the absence/presence of benign findings in a breast. Therefore, each exam has a total of four binary labels. The goal is to produce four predictions corresponding to the four labels for each exam.

Inspire by earlier work by Geras et al., it was trained the deep multi-view CNNs of four alternative architectures, as illustrated in Fig. 1. All of these networks are made up of two basic building blocks: four columns, one for each mammography view, based on the ResNet architecture, each of which outputs a fixed-dimension hidden representation; and two fully connected layers, which connect the computed hidden representations to the output predictions. In order to generate the final predictions, the models combine view-specific hidden representations from every view in different ways.

* view-view model: concatenates L-CC and R-CC representations and L-MLO and R-MLO representations making separate predictions for CC and MLO views
* image-wise model: makes a prediction for each of the four views independently
* side-wise model: first concatenates L-CC and L-MLO representation and R-CC and R-MLO representations and then makes predictions for each breast separately
* joint model: concatenates the representations of all four views and jointly predicts malignant and benign findings for both breasts

A diagram of a diagram

Description automatically generated

*Figure 1. Training of four alternative architectures*

For all these models there were used four ResNet-based 22-layer networks as columns computing a 256-dimension hidden representation vector of each view. The wise-wise model was found to be the most accurate on the validation set in terms of the malignant/not malignant prediction task.

The full architecture of ResNet-22 can be seen in figure 2 where it can also be observed that the weights for the L-CC and R-CC ResNets were tied as well as for the L-MLO and R-MLO.

A diagram of a resnet layer

Description automatically generated

*Figure 2. Architecture of single-view ResNet-22*

The model was trained using the Adam optimization algorithm with a learning rate of 10-5 and a minibatch of size 4. The L2 regularization was applied to the model weights with a coefficient of 10-4.5. The model has 6.132.592 trainable parameters. In all experiments, the training set was used for optimizing parameters of the model and the validation set for tuning hyperparameters of the model and the training procedure.

The models were primarily evaluated in terms of AUC for malignant/not malignant and benign/not benign classification tasks on the breast level. AUC and PRAUC, which are frequently used measures in the evaluation of radiologists' performance, are used to evaluate the model and readers' responses on the subset for the reader study. Different facets of a prediction model's performance are captured by ROC and PRAUC. The trade-off between the true positive rate and false positive rate for a model utilising various probability thresholds is summarised by the ROC curve. The trade-off between the genuine positive rate (recall) and the positive predictive value (precision) for a model utilising various probability thresholds is summarised by the precision-recall curve.

For the model presented in this paper we can observe a couple of strengths as well as weaknesses:

**Strengths:**

* Classification Accuracy: The model exhibits high classification accuracy for breast cancer screening examinations. This is explained by the use of neural networks and the large amount of training data.
* Pixel-level Labels: Pixel-level labels improve the performance of the model by enabling in-depth examination of the images. They are included in the training set.
* Innovative Patch-level computing: The model generates heatmaps using patch-level computing, adding details to the breast-level model. The effectiveness of the model is enhanced by this method.
* Improvement over Radiologists: When utilised in concert with skilled radiologists, the model's performance on the specific task assessed in the reader research outperforms that of radiologists, indicating its potential to increase sensitivity for breast cancer identification.
* Possibility of Real-time Reading: The model is a candidate for real-time screening of mammograms due to its effectiveness and accuracy, which may result in earlier diagnosis.

**Weaknesses:**

* Hardware Restrictions: The authors note that it would be impractical to train this model totally in an end-to-end manner on the hardware that is currently in use because it demands a large amount of processing power.
* Small Test Set: The investigations were carried out with a small test set. This raises questions about how well the model will perform when applied to larger, more varied datasets.
* Clinical Validation is Required: The authors admit that additional clinical validation of their findings is necessary. This implies that the performance of the model needs to be extensively assessed in real-world circumstances.
* Simplistic Model Design: The authors acknowledge that their model's design is fairly straightforward, leaving room for the creation of more complex and potentially realistic models.
* Task Specificity: The model's emphasis on detecting visible cancer during screening mammography may restrict its applicability to more complicated diagnostic situations, when radiologists take into account additional imaging and clinical data.
* The next stage for the model, according to the authors, could be to forecast how breast cancer would evolve in the future. This suggests a potential area where the model's abilities should be strengthened.

In conclusion, the model shows promise for screening for breast cancer, although there are questions about its complexity, validity, and potential for more extensive clinical applications. Before wide-scale implementation, additional reworking and testing are needed.