**ConvNets for Detection of Abnormal Mammograms**

Eric Scuccimarra ([skooch@gmail.com](mailto:skooch@gmail.com))

**Abstract**

Mammography is the most common method of detecting breast cancer. Early detection significantly improves survival rates, and between 8% and 25% of abnormalities go undetected. We trained ConvNets on the DDSM dataset to detect the presence of lesions and predict the class and pathology of the lesions.

We were able to achieve an accuracy of 99% on determining whether scans were normal or abnormal.

1. **Introduction**

Breast cancer is the second most common cancer in women worldwide. About 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer over the course of her lifetime. The five year survival rates for stage 0 or stage 1 breast cancers are close to 100%, but the rates go down dramatically for later stages: 93% for stage II, 72% for stage III and 22% for stage IV. Human recall for identifying lesions is estimated to be between 0.75 and 0.92 [1], which means that as many as 25% of abnormalities may go undetected.

The DDSM is a dataset of normal and abnormal scans; however the size is relatively small. To increase the size of the dataset we extract the Regions of Interest (ROI) from each image, perform data augmentation and then train ConvNets on the augmented data. The ConvNets were trained to predict both whether a scan was normal or abnormal, and to predict whether abnormalities were calcifications or masses and benign or malignant.

1. **Related Work**

There exists a great deal of research into applying deep learning to medical diagnosis, but due to the lack of available data there is not much research into mammography specifically. [1, 4] use ConvNets to classify pre-detected breast masses by pathology and type, but do not attempt to detect masses from scans. [2,3] detect abnormalities using combinations of region-based CNNs and random forests.

1. **Datasets**

The DDSM [6] is a database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. The CBIS-DDSM [8] collection includes a subset of the DDSM data selected and curated by a trained mammographer. As the CBIS-DDSM contains only abnormal images, normal scans were taken from the DDSM and combined with the CBIS-DDSM scans.

Data from the University of California Irvine Machine Learning Repository [5] was also used for exploratory data analysis to gain insight into the characteristics of abnormalities.

1. **Methods**

The DDSM and CBIS-DDSM datasets are relatively small, so the images were pre-processed with data augmentation to create a dataset of reasonable size. Then ConvNets were constructed and trained on the data using multiple labelling methods.

**4.1 Data Augmentation**

The CBIS-DDSM scans were of relatively large size, with a mean height of 5295 pixels and a mean width of 3131 pixels. Masks highlighting the ROIs were provided. In order to create usable images from the full-sized scans the ROIs were extracted using the masks and sized down to 299x299. Each ROI was extracted in multiple ways:

1. The ROI was extracted at 598x598 at its original size.
2. The ROI was zoomed to 598x598, with margins to provide context.
3. If the ROI was too large to fit in a 598x598 image it was extracted in 598x598 tiles with a stride of 299.

The 598x598 images were then resized to 299x299. In order to increase the size of the dataset data augmentation was used, including randomly positioning the ROI within the image, random horizontal flipping, random vertical flipping and random rotation. The ROIs were extracted using two systems of margins which will be detailed below.

As the CBIS-DDSM dataset only contains abnormal scans the normal scans were taken from the DDSM dataset. While the CBIS-DDSM images had been reviewed and altered to eliminate artifacts such as white borders and overlay text, the DDSM images had not and many contained borders of variable sizes as well as patches of white which had been used to hide personal information of the patients. To remove the borderse ach DDSM images was cropped by 7% on each side. As the CBIS-DDSM ROIs were extracted proportionally to their size, rather than at a fixed zoom, the DDSM images were sized down by a random factor between 1.8 and 3.2, then segmented into 299x299 tiles with a stride ranging from 150 to 200 pixels. Each tile was then randomly flipped and rotated. To avoid the inclusion of images which contained overlay text, or were mostly black background, each tile was then added to the dataset only if it met upper and lower thresholds on mean and variance. The thresholds were determined through random sampling of tiles and careful tuning of the thresholds to eliminate images which did not contain usable data.

* 1. **ROI Context**

The analysis of the UCI data indicated that the edges of an abnormality were important as to determining its pathology and type, and this was confirmed by a radiologist. Levy et al [1] also report that the inclusion of context was a contributor to the accuracy of the diagnosis.

The first dataset created (referred to as **Dataset 5**) was a smaller dataset, consisting of 39,316 images. This dataset was created using a minimum fixed padding around the ROI during the pre-processing stage. Each ROI was extracted with a margin of between 30 pixels and 50 pixels from the edge of the image, with the ROI placed randomly within the bounds.

Later another, larger dataset was created (referred to as **Dataset 6**), consisting of 62,764 training images. This dataset was created following the standard suggested by [1] of extracting each ROI in an area double its size. The size of this dataset was increased by including more data augmentation such as random zooms and cropping.

* 1. **Data Balance**

Only about 10% of mammograms are abnormal, in order to maximize recall we weighted our training data more heavily towards abnormal scan, with a target of 85% normal. The data was split between training and test data using the existing divisions of the CBIS-DDSM dataset in order to prevent overlap. The total data was split into training, validation and test at percentages of 80%, 10% and 10%

* 1. **Labels**

In the DDSM dataset the scans are grouped into the following categories:

1. Normal
2. Benign Calcification
3. Malignant Calcification
4. Benign Mass
5. Malignant Mass

As previous work [1] has already dealt with classifying pre-identified abnormalities, we focused on classifying images as normal or abnormal with the expectation of retraining the model to classify by all five labels once satisfactory performance was achieved.

* 1. **Architecture**

Our first thought was to train existing ConvNets, such as VGG or Inception, on this dataset. However a lack of computational resources combined with the slow speed of training made it impractical to do this. The scans also contain much less feature variance than the ImageNet images these ConvNets were designed for, and we were concerned that the large size of these models might lead to overfitting. For these reasons we decided to design our own architecture specifically for this task, attempting to keep the models as simple as possible.

We started with a simple model based on VGG, consisting of stacked 3x3 convolutional layers alternating with max pools followed by fully connected layers. This architecture was iteratively improved, with each iteration changing only one aspect in the architecture and then being evaluated. Techniques evaluated include Inception-style branches [16, 17, 18] and residual connections [19].

The architecture was designed so that the same model could be used for both binary classification and multi-class classification by retraining the fully connected layers. In order to maximize recall a weighted cross entropy loss function was used giving abnormal scans double the weight of normal scans.

The models were constructed using TensorFlow and metrics were logged to TensorBoard. Batch normalization [15] was used for every layer, with dropout applied to the fully connected and pooling layers, and L2 regularization applied to all layers, with different lambdas for convolutional layers and fully connected layers.

The best performing architecture will be detailed below.

* 1. **Training**

As training on Dataset 6 was a very slow process, for model selection phase models were trained on Dataset 5 through 50 epochs with binary labels. Accuracy, precision, recall and f1 score were used to evaluate the models.

Once a model performed satisfactorily on Dataset 5 it was then retrained on Dataset 6 through 20 epochs, with the previous convolutional weights reused to speed up the training process. If the model performed well on Dataset 6, the model was then retrained from scratch to classify into all five classes, on the assumption that this would allow the convolutional layers to extract the most important features.

Once the models had been trained on Dataset 6 with all classes, the convolutional layers were frozen and the fully connected layers were then retrained for the normal/abnormal binary classification.

We had considered using transfer learning from VGG or Inception, but decided that the features of the ImageNet data were different enough from those of radiological scans that it made more sense to learn the features from scratch on this dataset. However, the use of transfer learning between models greatly sped up the training process saving weeks of training time.

1. **Results**

**5.1 Architecture**

**5.2 Performance**

While the models all performed very well on the training data, at first they did not seem to be generalizing well, as the validation accuracy did not seem to improve in step with the training accuracy in early epochs. However, at about epoch 25 of training on Dataset 5, the validation accuracy began to improve significantly. The models performed much better on the test data set than expected, as show in Table 1.

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Recall** |
| 1.0.1.39n | .9935 | .9590 |
| 1.0.0.28 | .9903 | .9431 |

Table 1: Performance on Test Set

Model 1.0.0.28 was more stable as far as validation results as shown in Figure 1.

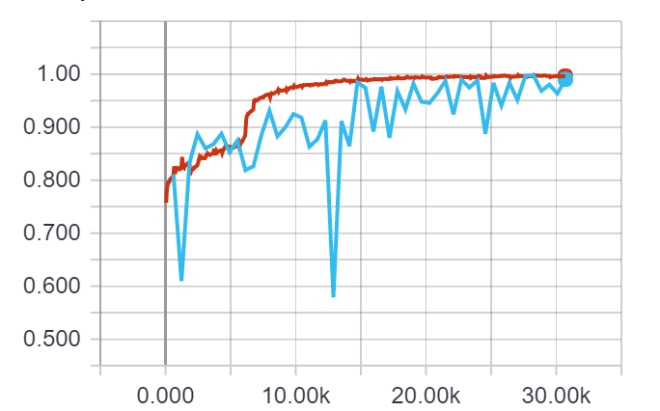


Figure 1: Accuracy of model 1.0.0.28 on training and validation data

Model 1.0.1.39 has slightly better results on the test data but appeared to be slightly more unstable on the validation data as seen in Figure 2.

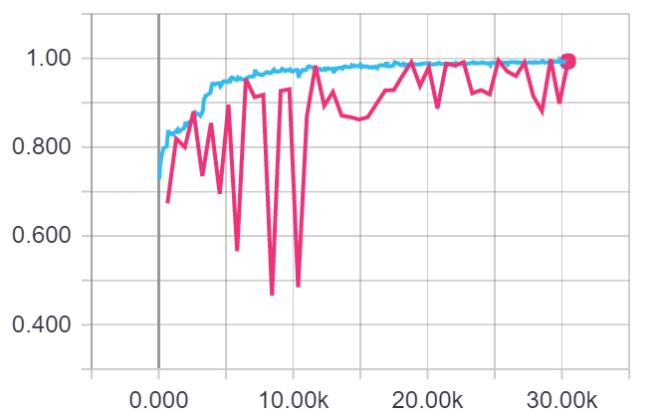


Figure 2: Accuracy of model 1.0.1.39n on training and validation data

Model 1.0.1.39n was a more complicated model than 1.0.0.28, with several extra convolutional layers and an Inception-style branch in the first layer.

The use of multiple branches was evaluated on Set 5, and while they did provide better results on the training data they seemed to make the model generalize to the validation data more poorly so were not included.

**5.2 Decision Thresholds**

These results were obtained using a threshold of 0.50. The precision and recall could be drastically altered by changing the decision threshold. Altering the threshold between 0.2 and 0.85 made it surprisingly easy to achieve a precision of 1.0 with a recall around 0.50. It was also possible to get a recall close to 1.0, although this resulted in very low precision.

This could be very useful for radiologists, allowing them to screen out scans which are either definitely negative or definitely positive and allowing them to focus on the more ambiguous scans.

1. **Conclusion**

We have demonstrated that Convolutional Neural Networks can be trained to determine whether a section of a mammogram contains an abnormality with recall of 95%, substantially above human performance. Adjusting the decision threshold would further improve the recall. These methods could be used to pre-screen mammograms allowing radiologists to focus on scans which are likely to contain abnormalities.

Future work would include creating a system which would take a full scan as input, segment it and analyse each segment to return a result for an entire mammogram. Levy et al [1] have already shown that ConvNets can be used to classify abnormal ROIs, those techniques can be combined with those described here to create a complete end-to-end system for analysing mammograms.

**References**

1. D. Levy, A. Jain, Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks, *arXiv:1612.00542v1,* 2016
2. N. Dhungel, G. Carneiro, and A. P. Bradley. Automated mass detection in mammograms using cascaded deep learning and random forests. In Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on, pages 1–8. IEEE, 2015.
3. N.Dhungel, G.Carneiro, and A.P.Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 605–612. Springer International Publishing, 2015.
4. J.Arevalo, F.A.González, R.Ramos-Pollán,J.L.Oliveira,andM.A.G.Lopez. Representation learning for mammography mass lesion classiﬁcation with convolutional neural networks. Computer methods and programs in biomedicine, 127:248–257, 2016.
5. Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
6. The Digital Database for Screening Mammography, Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore and W. Philip Kegelmeyer, in Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 212-218, Medical Physics Publishing, 2001. ISBN 1-930524-00-5.
7. Current status of the Digital Database for Screening Mammography, Michael Heath, Kevin Bowyer, Daniel Kopans, W. Philip Kegelmeyer, Richard Moore, Kyong Chang, and S. Munish Kumaran, in Digital Mammography, 457-460, Kluwer Academic Publishers, 1998; Proceedings of the Fourth International Workshop on Digital Mammography.
8. Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi , Daniel Rubin (2016). Curated Breast Imaging Subset of DDSM. The Cancer Imaging Archive.
9. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, December, 2013, pp 1045-1057.
10. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
11. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
12. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and YuyingLi, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
13. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).
14. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *arXiv:1409.1556,* 2014
15. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 448–456, 2015
16. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 1–9, 2015.
17. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.
18. C. Szegedy, S. Ioffe, V. Vanhoucke, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, *arXiv:1602.07261v2,* 2016
19. K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, *arXiv:1512.03385,* 2015