**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 [5], 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 trained 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 only contains 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 with different classification 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 with no zoom and random cropping.
2. The ROI was zoomed, with margins to provide context, at 598x598.
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. For each image random data augmentation was also done including cropping, horizontal and vertical flipping and rotation.

As the CBIS-DDSM dataset only contains abnormal scans the normal scans were taken from the DDSM dataset. Each DDSM images was cropped by 7% on each side to remove borders, 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, and then added to the dataset if it met upper and lower thresholds on mean and variance. The thresholds were set to exclude images which contained mostly black background or which contained white patches or text which were present in many of the DDSM images.

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%

Two datasets were created – a smaller dataset with 39,316 training images (referred to as **Set 5)**, and a larger dataset with 62,764 training images (referred to as **Set 6)**. The larger dataset was created extracting each ROI randomly positioned in an area double its size, as suggested by Levy et al [1]. The smaller dataset was with smaller margins randomly ranging from 30 to 50 pixels.

* 1. **ConvNet Architectures**

Multiple ConvNets were designed and evaluated specifically for this dataset. The ConvNets incorporated features from VGG [14], Inception [16, 17, 18] and ResNet [19].

A lack of computational resources along with the speed of training made it impractical to train either VGG or Inception on this dataset, and it seemed as if the complexity of these networks might be overkill given the relative uniformity of the scans compared to other datasets used for image classification, so we opted to create our own architectures attempting to keep them as simple as possible.

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 basic architecture was a VGG-like network of 3x3 convolutions with max pools, followed by fully connected layers. 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 (see 4.3 below). In order to maximize recall a weighted cross entropy loss function was used giving abnormal scans double the weight of normal scans.

* 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

Different methods of classifying the images were explored, including normal or abnormal, as well as using the all of the classes.

* 1. **Training**

Each model was trained on both of the datasets using both classification methods. Each model was trained through 50 epochs and the metrics used to evaluate the performance included accuracy, precision and recall.

In order to speed up training models were initially trained on the smaller dataset using binary classification. Once satisfactory performance was achieved the model was then trained on the larger dataset with the weights of the convolutional layers initialized using the pretrained weights.

1. **Results**

Training the models for binary classification on Set 5 resulted in excellent accuracy on all data sets. While the accuracy and recall on the validation data were more volatile than we would have liked, the models both performed exceptionally well on the test data, as shown 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.

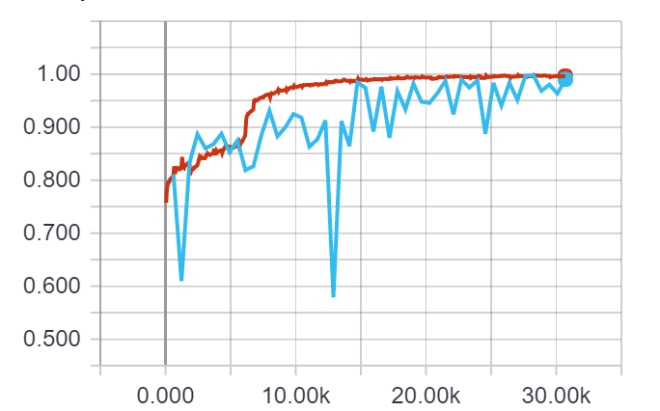


Figure 1: Accuracy of model 1.0.0.28 on training and validation data

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.

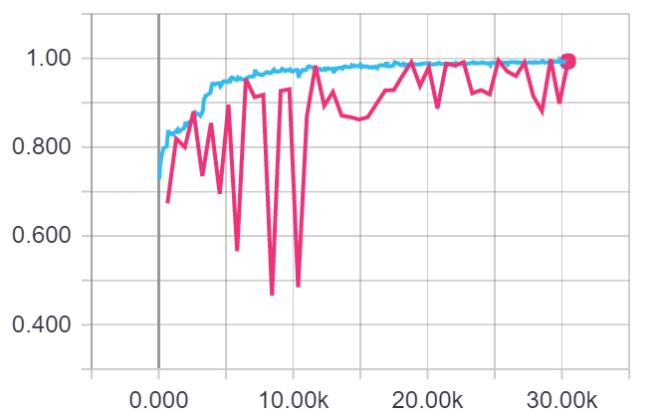


Figure 2: Accuracy of model 1.0.1.39n on training and validation data

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.

These results were obtained using a threshold of 0.50. Using lower thresholds would substantially reduce the number of false negatives.

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