Using Deep Convolutional Neural Networks to Predict Semantic Features of Lesions in Mammograms

Vibhu Agarwal
Stanford University
vibhua@stanford.edu

Clayton Carson
Stanford University
carsoncz@stanford.edu

Abstract

Preventive care recommendations for breast cancer require that women above a certain age be regularly screened by mammography [1, 2]. Computer aided interpretation of mammograms involves the extraction of features of suspicious areas in the mammograms and providing these as inputs to a clinical decision support system. While the extraction of computational features (such as geometry, contrast, intensity, texture) from a given region of the image may be easily automated, extraction of semantic features (for instance the type of lesion and its pathology) has traditionally relied on radiologists documenting their findings in structured or free format text. Deep convolutional neural networks have demonstrated the ability to outperform skilled humans in certain observational tasks [3, 4]. In our study, we investigate the feasibility of training a convolutional neural network with annotated lesion images drawn from the Digital Database for Screening Mammography (DDSM). We report our results on two label prediction tasks that influence mammogram interpretation and downstream clinical actions related to diagnosis, intervention and prognosis. Here we examine the application of a purely data driven approach to the task of predicting semantic features, using a corpus of annotated images for training our predictor. We examine the application of a purely data driven approach to the task of prediction of semantic features as a means of improving overall efficacy of screening mammograms and ultimately improving the clinical care for breast cancer patients.

1. Introduction

One in eight U.S women is expected to develop invasive breast cancer over the course of her lifetime [5]. In 2014 an estimated 232, 570 new cases of invasive breast cancer were diagnosed in the US and the estimated number of patient deaths on account of breast cancer were 40,000. For women above a certain age, screening mammography is recommended as the standard for preventive care and is estimated to result in a $3 - 13\%$ reduction in mortality. As illustrated by Barlow et al [6] human errors (inter- observer variability in the interpretation of screening mammograms) is also a well known problem. False positives result in over-diagnosis, over-treatment and by consequence, psychological and financial distress to otherwise healthy patients. Visual similarity between normal dense tissue and many types of breast cancers may also result in false negatives, and together with a significant false positive rate, this results in the diminished efficacy of screening mammography. Recognizing that manual classification of tumor images is error prone, given the large number of noisy predictor variables and interactions, techniques for automatic classification are a subject of active research [7, 8]. As models that can effectively capture complex interactions between a large number of predictors, as well as possible non-linearities between predictors and the outcome variable, neural networks have been studied extensively for various tasks related to characterizing breast tumor images. Stafford et al [10] employed an ensemble of neural networks for detecting (segmenting) micro-calcifications in mammograms and achieved 84% sensitivity and 75% specificity. The work of Zhang et al [11] focused on classification of micro-calcification clusters based on thresholded counts of distinct micro-calcifications. Other approaches for micro-calcification detection [12, 13] have utilized pre-extracted features from suspicious regions as inputs to neural networks and have been successful in reducing the number of false positives that recommended for further investigation.

To the best of our knowledge, characterization of suspicious lesion images using a purely data driven approach has not yet been attempted. The availability and adoption of a standard terminology for description of breast lesions, the availability of annotated mammograms databases for research, combined with recent advances in training large convolutional neural networks for complex image classification tasks present a compelling opportunity to do so.

2. Approach

We trained convolution neural networks (conv-nets) on a cropped and augmented set of images of breast lesions, labeled for a set of characteristics that are deemed relevant in tumor assessment. We evaluated our networks on two classification tasks namely classification of lesions as masses versus calcifications, and classification of lesions as benign versus malignant.

2.1. Data

We used mammograms from the Digital Database for Screening Mammography (DDSM) [14], a collaboratively maintained public dataset at the University of South Florida. DDSM consists of 2,500 studies each containing two images of each breast, associated patient metadata, image information and a pixel-level ground-truth annotation of suspicious areas. The dataset is available as a compressed archive that requires decompression into a raw data format (RAW) and conversion to the original Portable Pixelmap format (PNM) based on the original image dimensions. The images are organized into 12 normal, 15 malignant and 16 benign volumes with a total of 8,752 images representing 2,620 cases. Pixel masks for the regions of interest are described in the overlay files associated with each image, using which the pre- segmented regions were cropped out and resized so that the shorter dimension was 64 pixels. From each resized crop, two to three 64x64 patches were sampled randomly and the resulting squares were each rotated 7 times in steps of 45 degrees. The augmented data set consisted of 50,000 lesion images representing various translations and rotations of the base images as described above.

Each lesion image (a region of interest on the mammogram) has been annotated using a standard terminology that derives from the American College of Radiology's (ACR) Breast Imaging Reporting and Data System (BI-RADS) [15]. The annotated features consist of categorical descriptions of lesion characteristics such as mass shape, margin descriptors, calcification type and distribution, tumor assessment etc. In order to compare our results with earlier approaches, we selected the mass/calcification annotations as our outcome labels. The annotations also include a malignancy assessment that may be negative, benign, probable, suspicious or highly suggestive. In order to evaluate a conv-net's performance on a malignancy assessment task, we created a second set of labels with the aforementioned annotations collapsing negative and benign cases into one category and the remaining into another category.

3. Results

We split our dataset into a training set consisting of 40,000 images, and a test set consisting of 10,000 images, taking care to ensure that the two sets had roughly the same

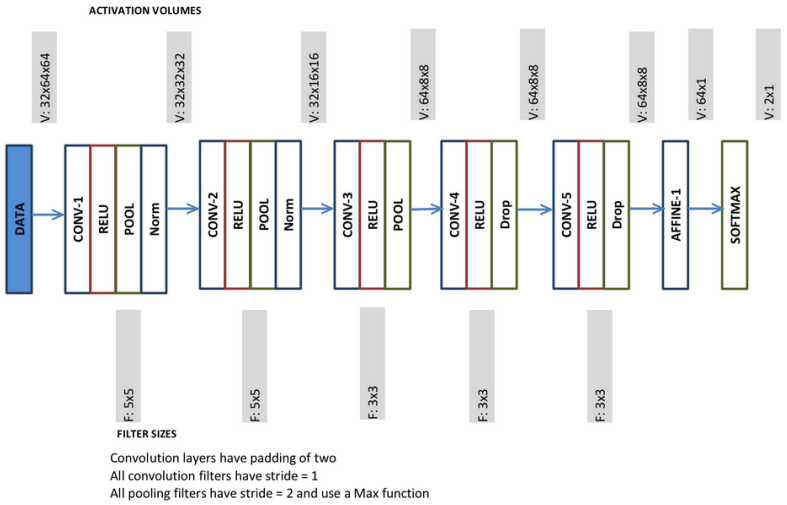


Figure 1: Mass vs Calcification conv-net Structure

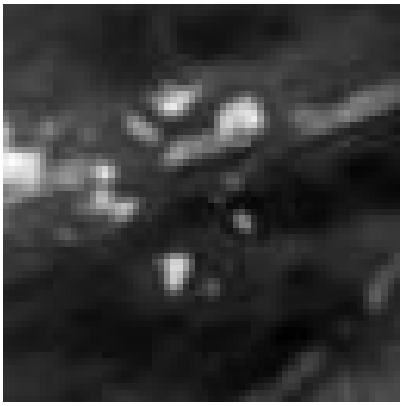
proportion of the positive and negative labels. All images were mean centered using the training set mean. For each of the two binary classification tasks, we trained a separate conv-net.

With several tunable hyper-parameters, finding an optimal set would require an exhaustive search that is usually intractable given finite computational resources and time for completion. We followed a coarse-to-fine sweep approach for filter sizes, number of filters, learning rate and weight decay. For other hyper-parameters we tried to obtain stable values that do not degrade our results. We note that further optimization of our network and learning parameters could lead to better results than what we present below. We implemented our conv-nets in Caffe [16] as protocol buffer definitions and used the stochastic gradient descent implemented within Caffe's solver to train the models in GPU mode.

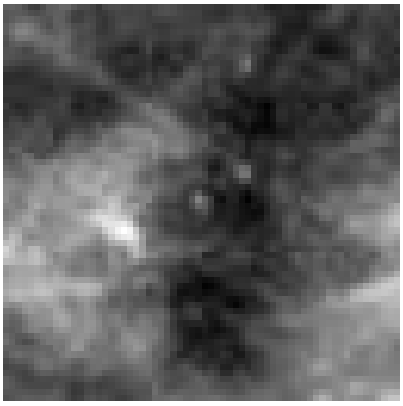
3.1. Calcification versus mass classification

In order to distinguish between masses and calcifications a six layer conv-net was applied as illustrated in figure 1. Networks ranging in size from 3 through 8 layers were analyzed. Our final conv-net network size was determined by balancing improvements in testing accuracy with network training speed. Additionally, regularization and learning rate parameters were optimized using a parameter sweep approach.

After this conv-net was trained, a maximum accuracy of approximately 87% was achieved. This is comparable to previous neural net approaches to mammogram mass identification [17, 18]. Figure 2 shows some sample test images correctly identified by the trained convolutional network.



(a) Sample correctly identified calcification



(b) Sample correctly identified mass

Figure 2: conv-net Classification Results

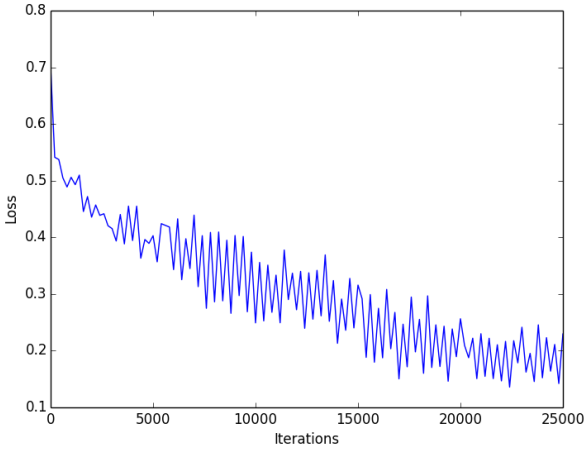


Figure 3: conv-net A Training Loss History

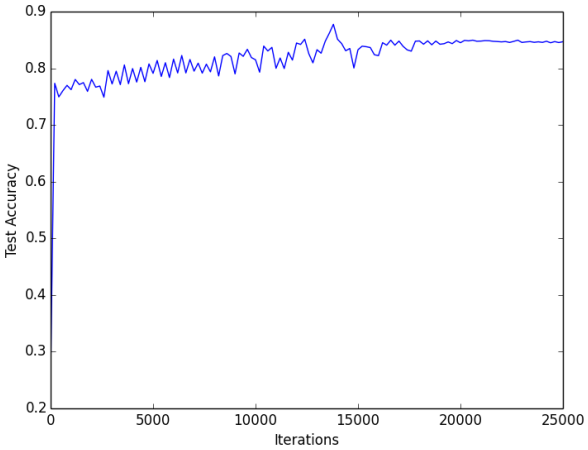


Figure 4: conv-net A Test Accuracy

As may be seen in the figure 3, the training loss continually decreased as the network was trained with loss leveling off towards the final iteration. This helps to demonstrate convergence for the network on the training datset. Likewise, the test set prediction accuracy steadily improved with convergence after approixmately the 20000th iteration.

Figures 5 and 6 show filter and layer activation visuilizations for the fully trained network respectively. The layer activation visuilization is sparse and localized as expected for a trained network. Likewise the smooth filter visuilazation suggests that the network has converged. These two results help to indicate convergence as well as a proper selection of parameters for the network.

3.2. Malignancy assesment

The conv-net for the malignancy assessment task consists of five convolutional layers followed by three fully connected layers. Each convolutional layer is followed by pooling, a rectified linear activation (RELU) and a dropout layer with drop out parameter = 0.5. This is illustrated in figure 7. For the malignancy assessment task, our conv-net showed sufficient capacity to be able to overfit the training data. As evident from figure 9, the training accuracy was close to 90% after 35,000 iterations. However, experimenting with learning rate reduction after 35K iterations did not show further decrease in loss values. The best validation accuracy obtained was 69.8%.

Loss and accuracy histories during training are shown in figures 8 and 9 respectively.

4. Discussion

In terms of the calcification vs mass indentification task, our conv-net was able to achieve a high degree of accuracy, comparable with previous attempts in applying conv-nets [17, 18] as well as traditional neural nets[19] to mass identification.

Being able to classify micro-calcifications correctly is

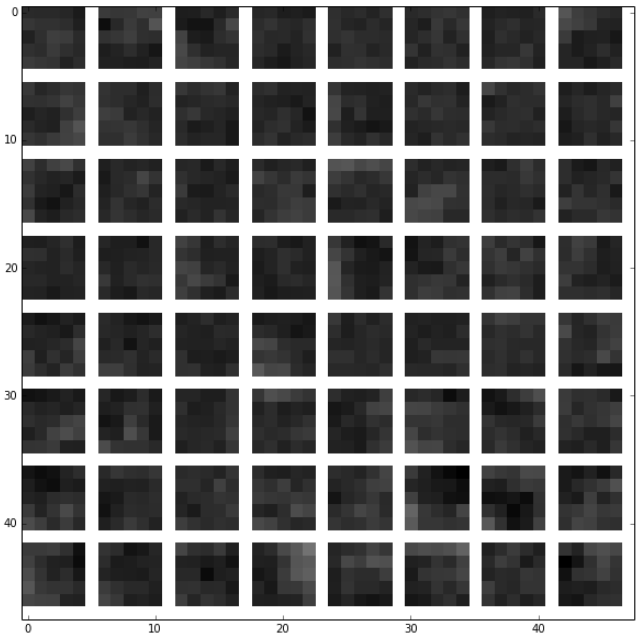


Figure 5: conv-net A Layer 1 Filter Visualization

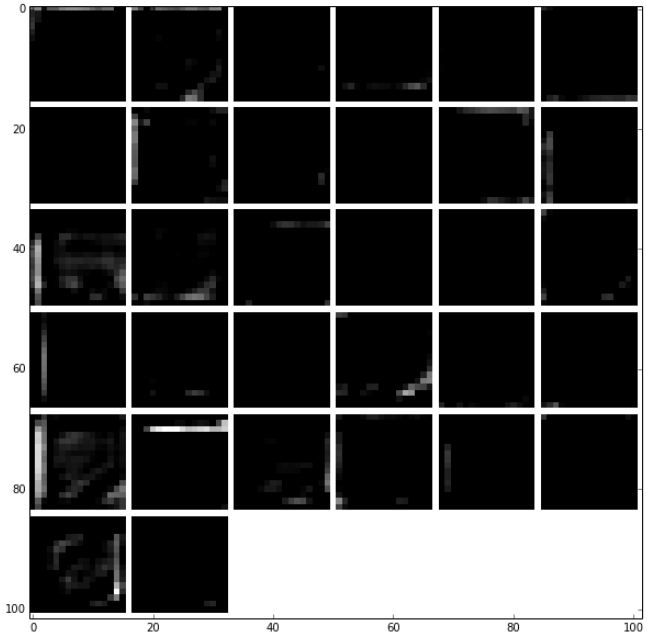


Figure 6: conv-net A Layer 1 Weight Visualization

important in cancer assessment since as high as 80% of breast cancers reveal some level of micro-calcification on histological examination. Accurately assembling the morphological and texture features of micro-calcifications by visual inspection is likely to be error prone given the variety in size, shape and distribution of micro-calcifications[20].

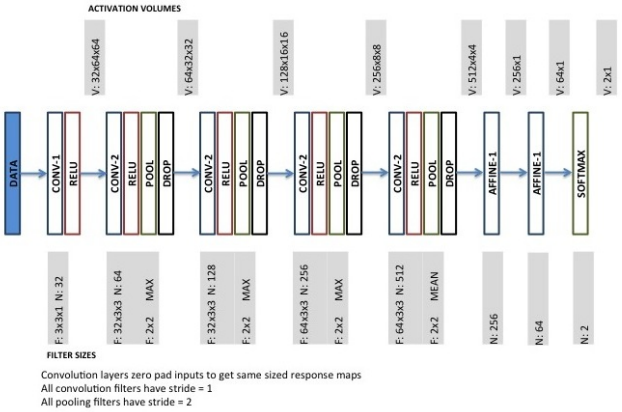


Figure 7: conv-net B

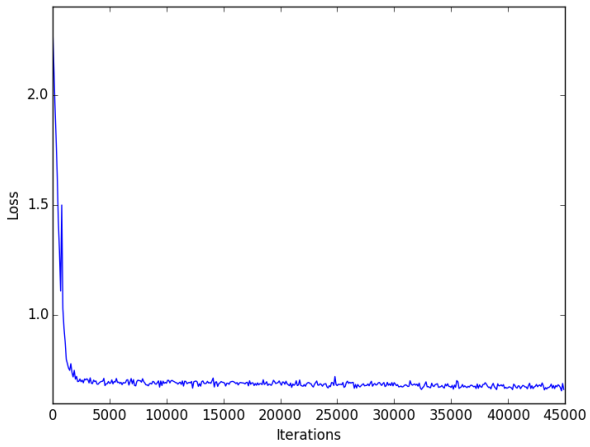


Figure 8: conv-net B Test Accuracy

A recapitulation of micro-calcification image features may be obtained from the trained filters of conv-net A (figure 6) that appears to suggest patterns of spatial distributions and geometry observed in micro-calcifications. As these features have been learnt by conv-net A directly from pixel data through loss optimization, they are the best features for visually discriminating micro-calcifications from other types of lesions. The decision to recommend diagnostic tests (biopsy) also takes into account several other clinical factors such as patient history, menopausal status, age etc. In conjunction with the features learnt by the conv-net, these are likely to be good predictors of malignancy and candidate inputs to a clinical decision support system. In general, a better modeling of the posterior probabilities for malignancy could be achieved by first discriminating the tumor type, which is easily achieved by chaining together conv-net A and conv-net B.

The reason for low validation accuracy obtained by

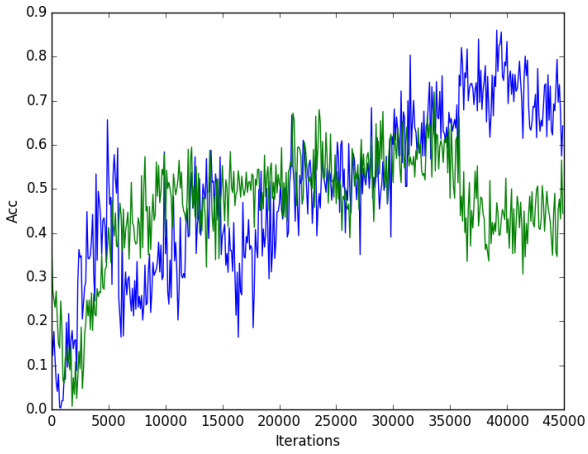


Figure 9: conv-net B Loss history

conv- net B on the malignancy assessment task is probably because malignant lesions take on a variety of different appearances depending on the type of lesion and the tumor stage. Conv-nets can provide the required expressiveness to learn from such datasets, provided sufficient observations are available. Another reason could be noise in the training ground truth itself. Instead of using known outcomes, we have used clinician assessment of malignancy at the time of screening which is known to be frequently erroneous, and further have collapsed probable and highly suggestive categories into one to reduce the complexity of our architecture. Additional training data, more aggressive augmentation along with a clean labels will likely result in better performance with a conv-net architecture similar to what we have used.

Finally, although we analyzed mammography interpretation as a binary problem, standard annotation practice is to simultaneously score a region of interest along several parameters, each described unambiguously as a BIRADS term. The conv-net approach described here, could be extended to take advantage of the full image annotation by defining the outcomes as structured labels.

5. Conclusion

In this paper, we reviewed the applicability of Convolutional Neural Networks to the problem of Mammogram interpretation. Overall, the results discussed in this paper indicate the validity of using conv-nets to learn data-derived features from mammograms, that may be used for a variety of clinical decision support tasks.

References

[1] ”Breast Cancer: Screening”. United States Preventive Services Task Force.

[2] ”Breast Cancer Early Detection”. cancer.org. 2013-09- 17. Retrieved 29 July 2014.

[3] Ciresan, D., Giusti, A., Gambardella, L., & Schmidhuber, J. (2012). Neural networks for segmenting neuronal structures in EM stacks.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, arXiv:1502.01852, 2015.

[5] http://www.breastcancer.org/symptoms/understand_bc/statistics

[6] Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96:18401850.

[7] Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96:18401850.

[8] Welch HG; Frankel BA (24 October 2011). ”Likelihood That a Woman With Screen-Detected Breast Cancer Has Had Her ”Life Saved” by That Screening”. Archives of Internal Medicine 171 (22): 20436.

[9] L. Hadjiiski, B. Sahiner, M. A. Helvie et al., Breast masses: computer-aided diagnosis with serial mammograms, Radiology, vol. 240, no. 2, pp. 343356, 2006.

[10] R. G. Stafford, J. Beutel, D. J. Mickewich, and S. L. Albers, Application of neural networks to computer-aided pathology detection in mammography, in Medical Imaging 1993: Physics of Medical Imaging, vol. 1896 of Proceedings of SPIE, pp. 341352, February 1993.

[11] W. Zhang, K. Doi, M. L. Giger, Y. Wu, R. M. Nishikawa, and R. A. Schmidt, Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network, Medical Physics, vol. 21, no. 4, pp. 517524, 1994.

[12] G. Rezai-Rad and S. Jamarani, Detecting microcalcification clusters in digital mammograms using combination of wavelet and neural network, in Proceedings of the International Conference on Computer Graphics, Imaging and Vision: New Trends, pp. 197201, July 2005.

[13] R. H. Nagel, R. M. Nishikawa, J. Papaioannou, and K. Doi, Analysis of methods for reducing false positivein the automated detection of clustered microcalcifications in mammograms, Medical Physics, vol. 25, no. 8, pp. 15021506, 1998.

[14] M. Heath , K. Bowyer , D. Kopans , R. Moore and P. J. Kegelmeyer "The digital database for screening mammography", Proc. Int. Workshop Dig. Mammography, pp.212 -218 2000

[15] American College of Radiology, Breast Imaging Reporting and Data System (BIRADS), American College of Radiology, Reston, Va, USA, 4th edition, 2003.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Gir- shick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In ACM MM, 2014.

[17] Shih-Chung B. Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T. Freedman and Seong K. Mun *Artifical Convolution Neural Network for Medical Image pattern Recognition*, Neural Networks, Vol. 8, pp. 1201-1214, 1995

[18] Shih-Chung B. Lo, Huai Li, Yue Wang, Lisa Kinnard and Matthew T. Freedman *A Multiple Circular Path Convolution Neural Network System for Detection of Mammographic Masses*, IEEE transactions on Medical Imaging, Vol. 21, No.2, 2002

[19] Turgay Ayer, Qiushi Chen and Elizabeth S. Burnside *Artifical Neural Networks in Mammography Interpretation and Diagnostic Decision Making*, Computational and Mathematical Methods in Medicine, Vol 2013, 2013

[20] M. P. Sampat, M. K. Markey, and A. C. Bovik, Computer-aided detection and diagnosis in mammography, in Handbook of Image and Video Processing, vol. 2, pp. 11951217, 2005.