A Comparison Between a Deep Convolutional Neural Network and Radiologists for Classifying Regions of Interest in Mammography

Thijs Kooi[1(✉)], Albert Gubern-Merida[1], Jan-Jurre Mordang[1], Ritse Mann[1], Ruud Pijnappel[2], Klaas Schuur[2], Ard den Heeten[3], and Nico Karssemeijer[1]

[1] Department of Radiology, Radboud University Medical Center,
Nijmegen, The Netherlands
Thijs.Kooi@radboudumc.nl

[2] Dutch Reference Centre for Screening, Nijmegen, The Netherlands

[3] Department of Radiology, University of Amsterdam,
Amsterdam, The Netherlands

Abstract. In this paper, we employ a deep Convolutional Neural Network (CNN) for the classification of regions of interest of malignant soft tissue lesions in mammography and show that it performs on par to experienced radiologists. The CNN was applied to 398 regions of 5×5 cm, half of which contained a malignant lesion and the other half depicted suspicious regions in normal mammograms detected by a traditional CAD system. Four radiologists participated in the study. ROC analysis was used for evaluating results. The AUC of CNN was 0.87, which was higher than the mean AUC of the radiologists (0.84), though the difference was not significant.

1 Introduction

Computer Aided Detection and Diagnosis (CAD) systems are being developed for a variety of modalities and pathologies. Mammography has traditionally been on the fore front of this endeavor and commercial CAD systems are widely used in clinical practice. Unfortunately, progress has mostly stagnated in the past decade and the merit of contemporary systems is strongly questioned, with studies showing no significant improvements in the sensitivity for invasive breast cancer [1]. Therefore, there is a strong need to continue the development of mammography CAD.

The culmination of several decades of research into statistical learning methods, in particular *deep learning* [2,3], is recently making headlines [4,5], with many Artificial Intelligence (AI) systems claiming human or even superhuman performance in a variety of tasks, previously thought to be insurmountably complicated. Deep Convolutional Neural Networks (CNN) have emerged as the defacto standard for vision based AI tasks and form the basis for face and object detection and autonomous vehicles. Traditional CAD systems employ features

A. Tingberg et al. (Eds.): IWDM 2016, LNCS 9699, pp. 51–56, 2016.
DOI: 10.1007/978-3-319-41546-8_7

such as spiculation, texture and contrast which are subsequently fed to a statistical learning machine. Rather than relying on engineers that mold medical knowledge into a set of features, deep architectures learn relevant features from data and the system is trained end-to-end. In essence, these ideas are not new and have been around since the late seventies. Their success in recent years can largely be attributed to more efficient training, advances in hardware and employment of many core computing and most importantly, sheer amounts of annotated training data.

In this paper, we employ a CNN for the classification of soft tissue lesions (e.g. masses and architectural distortions). We operate on square regions of 5×5 cm extracted from mammograms at suspicious locations identified by a traditional CAD system. This system outperforms a traditional CAD system we developed previously. In this study we compare the performance of this system to radiologists.

2 Deep Convolutional Neural Networks

To get a label for every sample, the image is convolved with a set of filter kernels, not unlike Gaussian derivative filters used in many traditional CAD systems, generating feature maps, which again are subjected to several transformations. Unlike traditional systems, however, most transformations are learned from data rather than handcrafted, allowing the algorithm to focus on information relevant for the classification problem and not predefined by the engineer. After the convolutional layers, several fully connected layers, where each activation is associated with a weight, are typically added, the exact benefit of which is still an open research question. Most people report an increase in performance when added, however. After the fully connected layers, the activations are fed to a softmax function, which generates a posterior probability over the labels. The parameters in the network are learned using maximum likelihood and backpropagation in combination with Stochastic Gradient Descent (SGD).

Pooling layers are typically added in between convolutional layers to reduce the size of the feature maps and induce some degree of translation invariance. Apart from this, these models do not exhibit any inherent invariances, although work is being done to incorporate this. Data augmentation, making new samples by means of deformations that one would expect to model possible variations in the data, is typically performed to make the networks robust and prevent overfitting.

For training the CNN, we used raw images and only applied a log transform. Images were scaled from 70 micron to 200 for faster processing. We employed a previously developed candidate detector designed for mammographic lesions [6] to generate candidate locations. It operates by extracting five features based on first and second order Gaussian kernels, two designed to spot the center of a focal mass and two looking for spiculation patterns, characteristic of malignant lesions. A final feature indicates the size of optimal response in scale-space.

To generate the pixel based training set, we extracted positive samples from a disk of constant size inside each annotated malignant lesion in the training set, to sample the same amount from every lesion size and prevent bias for larger areas. To obtain normal pixels for training, we randomly sampled 1 in 300 pixels from normal tissue in normal images, resulting in approximately 130 negative samples per normal image. The resulting samples were used to train a random forest [7] (RF) classifier. RFs can be parallelized easily and are therefore fast to train, are less susceptible to overfitting and easily adjustable for class-imbalance and therefore suitable for this task.

Centered at each location, we extracted patches of size 250 × 250 (5 × 5 cm). The pixel values in the patches were scaled using simple min-max scaling, with values calculated over the whole training set. Since some candidates occur at the border of the imaged breast, we padded the images with zeros. Negative samples were only taken from normal images. Annotated benign samples such as cysts and fibroadenomae were removed from the training set. However, not all benign lesions in our data are annotated and therefore some may have ended in the train or validation set as negatives. An overview of the data is provided in Table 1. The train, validation and test set were split on a patient level. We augmented all positive samples by translating and scaling each patch 16 times. All patches were subsequently flipped randomly during training. After augmentation, the train set consisted of 334752 positive patches and 853800 negatives. When combining the train and validation set, this amounts to 379632 positive and 931640 negative patches.

We used a VGG-like architecture [8] with 5 convolutional layers of $\{16, 32, 64, 128, 128\}$ with 3×3 kernels and 2×2 max-pooling on all but the fourth convolutional layer. A stride of 1 was used in all convolutions. Two fully connected layers of 300 each were added. The network architecture was chosen in a similar fashion as described by Simonyan et al. [8]. An illustration is provided in Fig. 1. To learn the model, we used RMSProp [9], an adaption of R-Prop for SGD with Nesterov momentum. We used Drop-out [10] on the fully connected layers with $p = 0.5$ as a regularizer and employed the MSRA [11] weight filler, with a uniform distribution. A learning rate of 5×10^{-5} with a weight decay of 5×10^{-5} was used. To battle the strong class imbalance, positive samples were presented multiple times during an epoch, keeping a 50/50 positive/negative ratio in each minibatch. All hyperparameters were optimized on a validation set and the CNN was subsequently retrained on the full training + validation set using the found parameters.

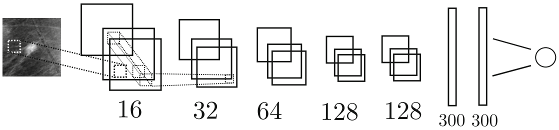

Fig. 1. Illustration of the employed architecture.

3 Reader Study

The mammograms used were collected from a screening program in The Netherlands (Bevolkingsonderzoek Mid-West) and were recorded with a Hologic Selenia mammography device. All tumors are biopsy proven malignancies and annotated by an experienced reader. The test set consisted of 199 biopsy proven malignant lesions and an equal amount of normals taken from normal cases, that were considered the most difficult by the candidate detector, resulting in 398 patches. This gives a representative set of difficult samples and allows for larger differences between readers and the CNN.

The group of readers comprised four experienced and certified breast radiologists. Since the CNN was trained with patches at 200 micron, we provided the two first readers with this resolution (reader one and three). This resolution was considered acceptable by the readers for analysis of mass lesions, but does not allow detection of microcalcifications. Since we excluded cases with microcalcifications this was not an issue. However, to assess whether the downsampling affected the reading, the other two radiologists read the patches at the original resolution.

The patches used in the reader study were the same as those fed to the CNN except that the standard processing algorithm provided by the manufacturer was applied for the images read by the radiologists. The radiologists were provided with a slider and instructed to score the patch between zero and one hundred based on their assessment of the suspiciousness of the patch.

Statistical analysis was performed with the DBM MRMC method [12] in which radiologists and the CNN are considered as two modalities (e.g. each radiologist was paired with the CNN in the analysis of variance).

4 Results

The ROC curve of the CNN and mean curve of the readers are shown in Fig. 2. The CNN obtained an AUC of 0.87, the radiologists a mean AUC of 0.845. We found no significant difference between the network and the radiologist ($p = 0.2530$). The invididual AUC values of the readers were: reader 1 = 0.845, reader 2 = 0.8774, reader 3 = 0.8411 and reader 4 = 0.8274. The first two readers were given patches at 200 micron and the last two at 70. There is clearly no significant difference between the readers performance at different resolutions.

Table 1. Overview of the data the network is trained with. Pos refers to the amount of malignant lesions and neg to the amount of normals.

	Cases		Exams		Images		Candidates	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Train	296	6433	358	11780	634	39872	634	213450
Validation	35	710	42	1247	85	4218	85	19460

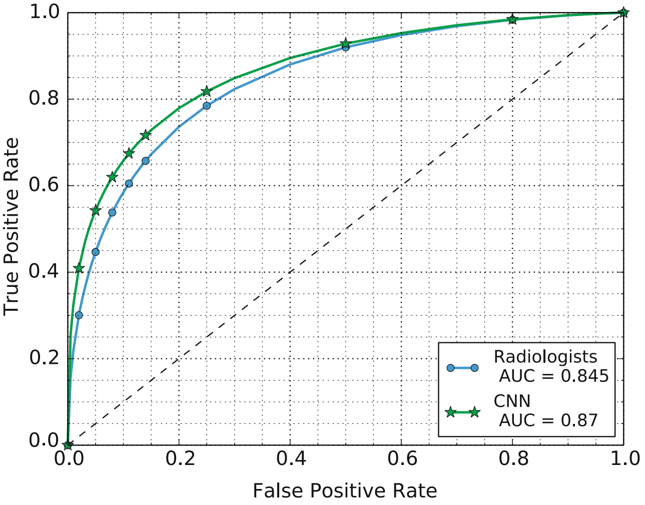

Fig. 2. Comparison between the CNN and the mean of the radiologists.

5 Discussion

In this paper, we have shown that a deep Convolutional Neural Network (CNN) trained on a large dataset of patches, centered around locations of interest is performing comparable to experienced radiologists. The patch based system is a sub problem and is clearly not the way radiologist read a mammogram in the clinic. We are currently exploring the incorporation of context information inside one view, addition of CC/MLO correlation, symmetry and temporal information.

We believe deep CNNs or similar statistical learning methods show great promise, have potential to advance the interest of Computer Aided Diagnosis (CAD) and that these algorithms can ultimately not only *aid* the physician but will eventually be able to read mammograms independently.

Acknowledgements. This research was funded by grant KUN 2012-5577 of the Dutch Cancer Society and supported by the Foundation of Population Screening Mid West.

References

1. Fenton, J.J., Abraham, L., Taplin, S.H., Geller, B.M., Carney, P.A., D'Orsi, C., Elmore, J.G., Barlow, W.E.: Effectiveness of computer-aided detection in community mammography practice. J. Natl. Cancer Inst. **103**, 1152–1161 (2011)
2. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1097–1105 (2012)

4. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
5. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
6. Karssemeijer, N., te Brake, G.M.: Detection of stellate distortions in mammograms. IEEE Trans. Med. Imaging **15**, 611–619 (1996)
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:14091556
9. Dauphin, Y.N., de Vries, H., Chung, J., Bengio, Y.: RMSProp and equilibrated adaptive learning rates for non-convex optimization (2015). arXiv:150204390
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification (2015). arXiv:150201852v1
12. Hillis, S.L., Berbaum, K.S., Metz, C.E.: Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Acad. Radiol. **15**, 647–661 (2008)