Part I
Methods

1 Operationalization

We document here how we tranform/reduce the breast cancer detection and diagnosis task into a machine learning task able to be taken by convolutional networks, i.e., how we produce a data set with m inputs $x^{(i)} \in \mathbb{R}^n$ and m corresponding labels $y^{(i)}$. We use this notation throughout this section.

1.1 Database

There are many publicly available mammography databases and many more which are private. Given the size of the expected network architecture and the data thirst of convolutional networks we focus only on the bigger databases. We also need pixel-level labels, i.e., lesions to be marked on each mammogram; this is generally made by expert radiologists drawing the boundaries of the lessions in the mammograms. Furthermore, we prefer to have good contrast resolution, the number of gray colors represented per pixel, and good spatial resolution, the area represented per pixel: at least 12-bit images ($10^12 = 4048$ gray values per pixel) with 0.1-0.15 mm maximum pixel size. Greater constrast resolution means that more brightness values are captured per pixel while greater spatial resolution means that hopefully more detail is included in the image. Most mammography databases including all described below satisfy these conditions.

The Digital Database for Screening Mammography (DDSM) [1] is arguably the most popular database used for CAD development. It is composed of around 10.5K digitised film mammograms from 2620 patients. Mammograms are either 12-bit or 16-bit images with 0.05 mm spatial resolution. Age and breast density of each patient is provided. Each lesion boundary is specified along with its information: type, assesment, subtlety and malignancy.

The BancoWeb database [3] consists of around 1.5K digitised film mammograms from 300 patients although they claim other 5K images stored internally "are being progressively transferred to the online database" [1]. Mammograms are 12-bit images with 0.075 or 0.15 mm pixel size. At the time of publishing (2011) only very few lesions had been marked in the mam-

[1] This claim was made back in 2011 so we expect it to be done by now.

mograms and it was impossible to review the current state of the database given that its webpage was not accesible online, which could be a sign of permanent closure. The only advantage of this database and the reason we include it here is that it was collected in Brazil and may be useful to test our CAD in Latin American patients.

The Breast Cancer Digital Repository (BCDR-DM) consists of 3.6K digital mammograms from 724 patients; this number was obtained from its website (`bcdr.eu/information/about`) which also states the database is still in construction and it is expected to have mammograms from 2000 to 3000 patients. Mammograms are 14-bit images with good spatial resolution [2]. Each lesion outline is marked in the image along with its assesment and other relevant clinical data. They also have a fairly big repository of digitised film mammograms called BCDR-FM (3.7K) but at lower resolutions.

Another small digital mammogram repository is called INbreast [2]. It consists of 410 digital mammograms from 115 patients. Each mammogram is a 14-bit image with 0.7 mm spatial resolution. Lesion boundaries are accurately marked and its information is also included. This could be used in conjunction with the BCDR-DM repository.

Finally, [4] used a private repository of around 6.5K digital mammograms obtained from 1120 patients. Specifics of contrast and spatial resolution are not provided but they are most probably good enough. Lesions are marked (with a circle) on the mammograms and lesion and patient information is provided. Even though this is a private repository of the University of Pittsburgh, if needed, we could ask them for access to it. This may not be plausible given the complications of sharing personal (granted anonymized) information and the size of the database.

Decision We have decided to use digital mammograms over film mammograms. Digital mammograms are sharper and do not have marks, stamps or other artifacts present in digitised film mammograms. On the downside, because the technology is newer it may be harder to obtain big databases and given its higher resolution they are normally heavier in terms of disk space. We believe the network will be able to pick up better features from the higher quality images and that the size on disk will not be a trouble given the storage availability on current computers. A small number of examples in the database is a big problem and some alternatives are offered below in case the network is not able to learn with the available data. For

[2]The webpage does not explicitly states the image's spatial resolution but judging by the size of the entire images it is good enough.

these reasons, we have decided to use initially the BCDR-DM and INbreast databases.

Alternatives We hope to obtain enough examples after cropping the mammograms into smaller image patches and applying some data augmentation to them (rotation and horizontal flipping). In case this is still not enough we could try various things: (1) obtain more labelled data from other sources, (2) reduce the complexity of the architecture to have less parameters to learn, (3) be more aggresive with the data augmentation, (4) pretrain the network with unlabelled digital mammograms which may be easier to get, (5) use film mammograms to pretrain the network and fine tune it on digital mammograms and (6) use an already pretrained network in other similar domains and fine tune it with digital mammograms.

Another option is to use only digitised film mammograms for the entire project but this will produce networks which expectedly produce bad results in digital mammograms [4] and seems like a step in the wrong direction given the clear trend of hospitals replacing film mammography by digital mammography. A final option is to join film and digital mammograms into a single data set, this may or may not work given the difference between them but will most probably decrease the quality of results on digital mammograms when compared to a network trained only on digital mammograms.

1.2 BCDR-DM

Files and how are they organized. Data available per case and per mammogram. How are boundaries written. formats, etc.

1.3 Image retrieval

We document here the decisions taken to obtain the small image patches x and its respective labels y from the chosen databases.

To obtain these image patches from the entire mammogram we move a square window across the image similarly to the way convolutional filters move accross an image and store the image patch directly beneath it. This generates a big number of small patches from each mammogram and takes advantage of the translational invariance of our data, i.e., a breast mass will continue to be a breast mass no matter its position in the image. The size of these patches, its labels and the overlapping between them depend on some variables that we define next.

Image dimensions We use square images because they are common in practice and simplify data augmentation. To define the size we have to consider two aspects: keeping a manageable input size for the network (in pixels) and capturing the entire lesion in the image (in mm).

The smallest microcalcification worth considering could be as small as 0.16 mm [**?**], thus the spatial resolution should be at most 0.16 mm. The standard definition of a cluster of microcalcifications is of 5 or more inside a 1 cm^2 area [**?**], thus the entire image patch should cover at least a 1 cm^2 area. Using a spatial resolution of 0.16 mm and an image size of 64×64 pixels we cover an area of 1.024 cm $\times$ 1.024 cm = $\sim$1.05 cm^2.

Mass sizes (length of the long axis) vary from 5 mm to 20 mm [**?**] [3] There is not really any restriction on spatial resolution other than it being good enough to capture texture information. Using a spatial resolution of 0.32 mm and an image size of 64×64 pixels we cover an area of 2.048 cm $\times$ 2.048 cm = $\sim$4.2 cm^2.

Although we use the same input size (64×64) for microcalcifications and masses they do not cover the same area in the mammogram. We need to use two different sizes because if we preserve the spatial resolution of 0.16 mm, the 1 cm^2 area would not be able to contain the entire mass meanwhile if we use a spatial resolution of 0.32 mm, some microcalcifications will dissapear and the 4 cm^2 area would have way too much noise compared to the size of the cluster of microcalcifications.

An alternative is to use a 128×128 pixels image patch with 0.16 mm, this will result on the same 4 cm^2 area needed for masses with the spatial resolution needed for microcalcifications allowing us to train a single network for both kinds of lessions. Nonetheless, this has some critical flaws: the number of learnable parameters will almost double, the GPU may run into memory bottlenecks because of the increased number of parameters and unnnecesary details (noise) will be included in the image.

Stride Depends on the size of the lession, if it is way too big i may miss some lession. for instance a small microcalcification cluster and how much overlap we want from image to image. or how many images fdo we obtain from an entire mammogram. More data augmentation if less stride.

Padding Should I use some padding so that I don't lose lesions which are in the very corner. For example, if I use a 2.5cm square and in there is a

[3] Bigger masses are easily detectable by touch and thus less important for our purposes.

$1cm$ mass close to the end of the image, then it may not detect it. Maybe not, I don't think ther eis going to be that many images on the side.

Labeling When the lesion hits the middle of the image, when the lession overlaps with the image or when the iomage is a given porcentage of the lession, or when a given porcentage of the lession apears in the image

Label info We will only use mass, MCC, normal, benign, malign and nothing.

Additional label information Is there any other info needed?. age breast density per case. per abnormality, assesment, subtlety BI-RADs words,

Image enhancement I will cut the images first and store them as is. Later truy different enhancements per image Contrast after cropping, or contrast before cropping.

Data cleaning Do I need to remove the marks and arrows and thingys. Could I let the network learn that those are not microcalcifications

data augmentation Should I do the image enhancement and data augmentation(rotations and flipping) during training or before hand. Or only the enhancement beforehand. Can I do them withouth changing the label? (depends on how the label is assigned, rotations and mirrors leve the same four squares in the middle of the image, thus, the label should not change, sclaing and others will.). Does it affect to present all different augmentations of the same image in one batch rather than in different batches

whjat about the blakc spaces Should I remove the black spaces or images that are more than 50% black or something like that. Should I do it during this stage or after the cut. If i do what is going to be thenetwork performance when presneted with an entirely black input. what about lession which are pressed agains the breats skin, if i delete these images, they may dissapear. 40% black out.

Resizing Does it affect the quality of the image what kind of resizing I do, should I use interpolation resizing or something. Hopefully I will always have to downsample so I may not lose much.

Total number of image patches.

Resukting data base Generate a data set like the ImageNet challenge/

1.4 Retrieval software

Developed in Python, named..... Does this and that. I would store all smaller images from the same in a matrix with the same name as the image where they came from. Maybe also preserve the same folder architecture. Make another tool to put it in Caffe style

2 Training

Details about the practical decsiions taken to archotectures, and hyper-paramters per experiments.

2.1 Hardware

Computational resources:

PC	GPU	RAM	CPU	HD	#
A4-401	Nvidia Quadro K620 384 cores 2GB 29 GB/s 128-bit	8 GB	i5-4570 3.2GHz x ?(1)	230 GB free 100 ubuntu ?	27
Mine	Nvidia NVS 5400M 96 cores 2GB 29 GB/s(?) 128-bit	4 GB	i5-3210M 2.50GHz x 4	320 GB free 200 ubuntu 56	1

Table 1: Available computers

384 will have to do.

2.2 Architectures

Write/Draw here the considered architectures.

I choose that.

2.3 Convolutional network

The network could be slided across an image. Options: (1) a network for detection of microcalcification and one for masses (and slide both across and plot their results with differerent colors) (2) a network for diagnosis of microcalc and one for masses(slide them both) and (3) one that detects micro+mass vs non-lession (stanford guys did bad with this one) and (4) one that detects any lession(micro+mass+other) vs no lession and (5) one that also detects more than one network but has multiple output.

Questions: how to deal with corners of images when presenting results, maybe not that important. Try extreme padding or just leaving it there.

Questions: Should I only use images from digital mammograms. Is there enough. Or maybe only use digital for testing (train on the ones who are harder: digitized and test on digital only).

Questions: If i choose to go with simple networks, start with microcalcifications or masses. which one is more useful (apparently masses).

Part II
Experiments

3 Image retrieval

Results and discussion.

4 Experiment 1

Architecture selected. Hyperparameters selected. results and discussion.

References

[1] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W. Philip Kegelmeyer. The digital database for screening mammography. In M.J. Yaffe, editor, *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218. Medical Physics Publishing, 2001.

[2] Ins C. Moreira, Igor Amaral, Ins Domingues, Antnio Cardoso, Maria Joo Cardoso, and Jaime S. Cardoso. Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012.

[3] Bruno Roberto Nepomuceno Matheus and Homero Schiabel. Online mammographic images database for development and comparison of cad schemes. *Journal of Digital Imaging*, 24(3):500–506, 2011.

[4] EA Sickles, CJ D'Orsi, and LW Bassett. *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*, chapter ACR BI-RADS Mammography. American College of Radiology, Reston, VA, 5th edition, 2013.

[5] B Zheng, J H Sumkin, M L Zuley, D Lederman, X Wang, and D Gur. Computer-aided detection of breast masses depicted on full-field digital mammograms: a performance assessment. *The British Journal of Radiology*, 85(1014):e153–e161, 2012. PMID: 21343322.