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 the 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 Image retrieval

We use the database ta that to obtain our mammographic images. We call image patches to those who are obtained.

1.1.1 DDSM o IRMA

Description of the database and the data available

Look at the IRMA database docs.

1.1.2 Cropping and retrieving small images

All the details and decisions taken for the image retrieval. It will go as a convolutional filter, it will move around the image obtaining smaller squares of a given size.

Image aspect ratio We will use a square aspect ratio because it is simpler and common use in practice.

Image size The image size has two components: how big is the pixelated image and how much space does it represent in the real mammogram. The actual size in a breats is arguably more important because if we chose a size which is way too big it will take way more noise over the actual lesion we are looking for and if we choose one that is too small it will miss the lesion. Mass sizes are normally between ... and ... [**?**].

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?

Image enhancement I will cut the images first and store them as is.

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

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

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.

Stride

Total number of image patches.

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 $1cm$ 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.

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.

DB Generate a data set like the ImageNet challenge

1.1.3 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

1.2 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).

2 Hardware

Computational resources:

PC	GPU	RAM	CPU	HD	#
CTS-516	GeForce GTX-580 512 cores 1.5GB 384-bit GDDR5	?	?	? (enough)	1
A4-401	Nvidia Quadro K620 384 cores 2GB 128-bit DDR3	8 GB	i5-4570 3.2GHz x ?(1)	230 GB free 100 ubuntu ?	27
Mine	Nvidia NVS 5400M 96 cores 2GB 128-bit DDR3	4 GB	i5-3210M 2.50GHz x 4	320 GB free 200 ubuntu 56	1

Table 1: Available computers