Using unsupervised machine learning for polyp detection in the GI tract

Mathias Kirkerd

Thesis submitted for the degree of
Master in Informatics: Technical and Scientific Applications
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2018

Using unsupervised machine learning for polyp detection in the GI tract

Mathias Kirkerd

Using unsupervised machine learning for polyp detection in the GI tract

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

Acknowledgements

my cat, if i had one

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Introduction REM

Cancer is, today, the second leading cause of death in the world, only behind cardiovascular diseases.
It is one of the leading causes of mortality worldwide, with approximately 14 million new cases in 2012. It is defined as a disease that has an abnormal cell growth with the potential to spread into other parts of the body.Contrary to normal cells, cancer cells are often invasive, and it will spread if not treated. In contrast to many other diseases cancer does not start from a foreign entity (such as a bacteria or virus), but it is often from a malfunctioning cell that starts dividing rapidly. This can happen when a cell is damaged, by for instance by radiation or other factors that damages the DNA, and the resulting damage causes the cell to uncontrollably divide. Especially in the later part of life everyone has the chance of getting cancer, and in fact everyone does. Our own body is designed to detect and remove cells that are prone to divide uncontrollably. Unfortunately this system is not perfect, and the immune system can in some cases overlook cells that are cancerous.

1.1.2 Statistics on cancer REM

The western (or modern) world has been in a battle against cancer, and despite a lot of new cures/innovations it is still one of the deadliest killers in the world. *The most common types of cancer in males are lung*

cancer, prostate cancer, colorectal cancer and stomach cancer.Stewart and Wild 2014

1.1.3 colorectal cancer REM

You can get cancer in every major organ, but some types of cancer are more common than others. For instance cancer in the gastrointestinal tract (GI) is one of the more common places to get cancer. This is just behind x, and it has a mortality rate of x in the first y years. We often call this 5 year survival rate for z. This is the standard way to measure the life expectancy of a patient diagnosed with cancer.

1.1.4 polyps REM

The colorectal cancer often starts in polyps. Polyps are, polyps do.

1.1.5 preventative matters and early detection REM

-colonoscopy
-mri
-pillcam
A good way to fight cancer is to detect and remove it early, or some times remove areas with a high chance of getting cancer. We classify cancer in to x stages, and the stage the patient are in often determines the chance you have for survival. In general, the earlier you find the cancer, the more likely it is that the patient will survive. And as mentioned above, the colorectal cancer often starts in these polyps. A crucial stage to prevent cancer lies in the early removal of there polyps. Reports shows x about this

*4 stages maybe? *early detection *survival rate

Because of this the ability to find, and remove colorectal polyps is great for preventing cancer in the GI tract.

colonoscopy/Ontonoscopy In the most common way to look for polyps in the GI tract is to use a medical team, and perform a colonoscopy or Ontonoscopy colonoscopy is preformed with a camera-stick that is inserted in to the GI tract through the patients anus.

Onoskopy is the same procedure, only the camera is inserted orally.

Advantages

- Accuracy: The use of a camera controlled by the doctor gives him/her the opportunity to stop at any anomalies.
- Quick results: Since the doctor is doing the procedure the result is given live.

Disadvantages

- Expensive: The cost of the doctor and the nurses needed is often high, especially on a routine check.
- Invasion of privacy: Getting an Colonoscopy or Onoskopy is a

MRI MRI (Maggnetic stuff) is the act of taking pictures blabla blabla
MRI (Maggnetic stuff) is the act of taking pictures blabla blabla
MRI (Maggnetic stuff) is the act of taking pictures blabla blabla

Advantages

- This is why mri is good
- This is why mri is good

Disadvantages

- This is why mri is bad
- This is why mri is bad

pillcam In the last 3-4 years there have been testing and development on the pillcam project EIR. Machine learning has, through many of the earlier projects, got the detection rate for the polyps up to x%

Advantages

- This is why mri is good
- This is why mri is good

Disadvantages

- This is why mri is bad
- This is why mri is bad

1.1.6 Simulas contribution to the pillcam project REM

Simulas EIR

* CAD ACD (computer aided diagnosis, Automated computer diagnosis)

1.2 Goal / Problem

1.2.1 pillcam project has lots of data, can be used to train an unsupervised network REM

1.2.2 Use Unsupervised learning as a pre-processing tool REM

1.2.3 use Unsupervised-NN/GAN for image enhancements so that a NN can train better REM

* Now that we got a lot of tests, why not unsupervised As mentioned, simula research centre has done a lot of testing on the pillcam project.

* We know that we can get some results using a neural network * Can this be done unsupervised? * Can it be done in a fashion that is better than S-ML

REM

1.3 Scope and Limitations

1.3.1 Use Unsupervised NN to find polyps REM

1.3.2 Use Unsupervised NN for pre-processing REM

* Something about earlier research already got far, so the scope is mainly unsupervised deep learning. * (and how to generalise it?) *REMegression

1.4 Research method

1.5 Related work

1.6 Outline

The rest of the thesis is structed as follows:
Chapter 2 - Background
*talk about cancer *talk about machine learning. *how to use ML on the pillcam video? **Chapter 3 - Me doing stuff**
Chapter 4 - Me got and present result
Chapter 5 - Me saying result was good A+

Chapter 2

Background

2.1 Cancer and polyps

2.1.1 What we are looking for REM

Different types of disorders. Polyp is harmless, but if left untreated it can become cancerous Pictures is from the pillcam project, kvasir dataset.

2.1.2 images from pillcam, and what we are looking at/for REM

2.2 Naive Methods REM

Now that we have an idea of what we are looking for we can first turn to some more naive methods for detecting anomalies, and for enhancing the images.
The field of image processing has been researched since

Using some of the classic methods in image processing we can see if

We often describe the method in to two groups of information: First and Second order statistics.
First order: First order statistics does not take in to account the relative positioning of the pixels in the image, and because of this, gives much less information than the second order statistics.
Example of First order statistics is often what information we can get

out of a histogram. This can be scewness, variance, and mean value.

Second order: Second order statistics takes in to account the relative positioning of the pixels in the image. We can calculate the GLCM matrix and get a much more detailed view of the image.

2.2.1 GLCM

A GLCM (Grey-level co-occurrence matrix) is a matrix that is used when examining the spatial relationship of pixels in a texture. The calculation of a GLCM gives us how often pairs of pixels with spesific values and a specified spatial relationship occur at a given place in an image.

Algorithm

For simplicity we use only greyscale in this example: The algorithm

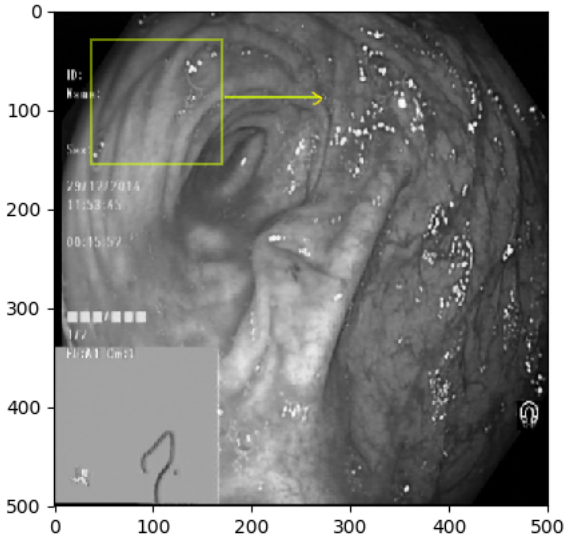

Figure 2.1: GLCM capturing features

starts by running a sliding window over the image, often with a stride, and for each stops calculates the spatial relationship between each

pixel specified. The result can be something like this figure where we can read out the most likely neighbouring pixel. The darker colours

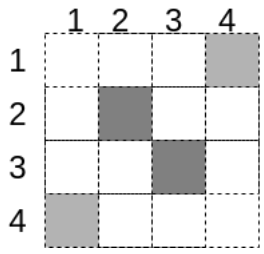

Figure 2.2: GLCM Matrix

on in the matrix is indicating that we often have a jump between, for instance pixel-value of 1 and a pixel-value of 4, but no from 1 to 1. With this information we can get a naive pattern-recogniser.

Other uses

Besides for the pattern recognition we can use the GLCM to get the information on:

- **Contrast** is the difference in luminance or colour in the picture. We would expect low contrast in the "background" and higher contrast around edges and irregular objects.
- **Homogeneity** is how similar a local area is to itself
- **Variance** σ^2 , is directly a measure of "roughness"
- **Mean** value of a GLCM can give us areas with higer or lower pixel values. Good way to find polyps if they are lighter than the tissue around.
- **Entropy**
- **Energy**

2.2.2 Edge detection

Using Edge detection in is another viable way to look for polyps.

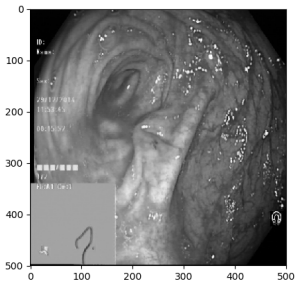

Figure 2.3: Original image

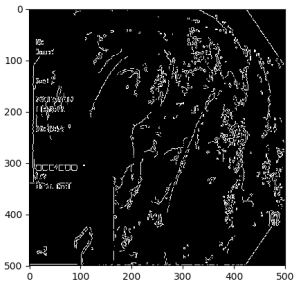

Figure 2.4: Edges of the picture

Algorithm

For each pixel look at the neighbouring pixel, if

$$abs(p_a - p_b) > tresh$$

then mark pixel as an edge pixel.

2.2.3 Hough Transforms

Using for instance Canny edge detection we can get a better view of where the potential border of the polyp/anomaly is. (As shown in

A hough transform can i theory have many/any shape(s), and together with edge detection, we might find some of the polyps this way.

2.3 Machine Learning

Machine learning is a very broad term, but can i short be summarised by:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with the experience E. Mitchell 1997

Here we have a couple of parameters:
E text about p

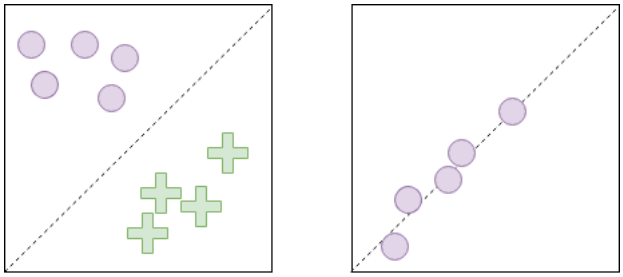

Figure 2.5: Left: Example of binary classification. Right: Example of regression

T text about p
P text about p

From this we see that the goal of machine learning is to improve some performance P with experience. **might here talk about different tasks ML can do?**

2.3.1 Supervised & Unsupervised machine learning

We often divide machine learning in to two (diffuse) categories: supervised and unsupervised.
Supervised learning: is the act of training with data that has an answer or a label. The learning algorithm can get supervision while training on the task. An example on a supervised task is to recognise handwritten numbers, or differentiate between dogs and cats. The task is supervised if the images comes with the correct label in the data set. These examples are typical classification examples, where the task is to identify the right group to classify the data to A simpler classification assignment is binary classification, where the target is (often) yes or no. Examples for binary classification is if an email is spam or not, is a car Norwegian or International. In the last example the classification changes from binary to multi-class if you sort the cars on every nationality, and not just Norwegian/non-Norwegian.

Another type of supervised learning is regression. This is the act of prediction given prior data. Examples of regression is everything from prediction of stock prices, to house prices in an area, to

Unsupervised learning: is the act of training without any supervision, on the sense that we do not give the algorithm the answer to the

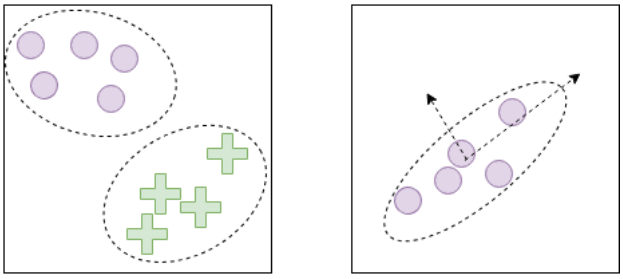

Figure 2.6: Left: Example of binary clustering. Right: Example of principal component analysis

training data set.

Since we do not have categorised data in unsupervised learning, we often Types of unsupervised learning can for instance be clustering, the act of sorting data based on similarity. An example of this can be if you want to sort plants based on species, or you are detecting anomalies in a dataset. Unsupervised learning can be used for PCA or other dimensionaly reduction methods.

A third method to used unsupervised learning is the adversarial route, where you use machine learning to make similar looking data to the original data set.

In the description of supervised vs unsupervised we looked at a specific branch of machine learning: Classification. Classification is, as the name implies, the task of getting data sorted in to groups of similarity.

- subsfication
- r to the pillcam projression
- transcription/translation
- de-noising /finding missing inputs

Now that we have the definintion of machine learning we focus on the task at hand; finding polyps. In an ideal world we have a Classification problem with only two classes: Non-polyp and polyp.

- SVM
- CNN

- random forests
- knn

2.3.2 CNN

UCNN?

2.3.3 Tasks (other better word goes here)

2.3.4 The rate of success

What is a good result, how to measure?
FP,TN,FN,TP

2.4 supervised vs unsupervised

What it means to be S/US.
Something about the kind of experience allowed during the learning process.

2.5 Unsupervised

noe med å dele i grupper? Experience the dataset containing many features, and finds useful properties of the structures. ***Unsupervised learning algorithms*** *experience a dataset containing manyfeatures, then learn useful properties of the structure of this dataset. In the contextof deep learning, we usually want to learn the entire probability distribution thatgenerated a dataset, whether explicitly, as in density estimation, or implicitly, fortasks like synthesis or denoising. Some other unsupervised learning algorithmsperform other roles, like clustering, which consists of dividing the dataset intoclusters of similar examples.* Goodfellow, Bengio, and Courville 2016

2.5.1 Approaches to unsupervised learning

look at the subsection 2.3.3 to see what applies to the unsupervised.

2.5.2 Deep Unsupervised learning

2.5.3 more

2.6 Related work

Chapter 3

Methods

Chapter 4

Implementation

Chapter 5

Result and Discussion

Chapter 6

Conclusion

Chapter 7

Future Work

Bibliography

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Mitchell, Tom M (1997). *Machine learning*. eng. New York.

Stewart, B.W. and C.P. Wild (2014). *World Cancer Report 2014*. International Agency for Research on Cancer. International Agency for Research on Cancer. ISBN: 9789283204299. URL: https://books.google.no/books?id=0QHbngEACAAJ.