Time series analysis for medical videos

Henrik Løland Gjestang

Thesis submitted for the degree of
Master in Computational Science
(Imaging and Biomedical Computing)
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2020

Time series analysis for medical videos

Henrik Løland Gjestang

Time series analysis for medical videos

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

Printed: Reprosentralen, University of Oslo

Abstract

Colorectal Cancer (CRC) is the third most common diagnosed cancer in both men and women. And a leading cause for CRC mortality is that patients in early stages have few and defuse symptoms which makes CRC hard to detect before later stages of the disease. To combat this the health care implemented regular screening for population most at risk. Yet for some the treatment come too late. We propose a system for automatically detection of diseases in the gastrointestal tract (GI-tract) using wireless capsule endoscopy and deep learning. A system like this requires a large amount of labeled data to train the prediction models. Although the health sector collects large amount of patient data, the most used deep learning systems need to have access to labeled data to properly train. This labeled data is difficult to create as doctors and other professionals need to manually inspect and annotate. Our proposal uses a model which only need a small amount of labeled data to operate, and an additional bulk of unlabeled data to advance.

Acknowledgements

This thesis is submitted as part of the master's degree in informatics: Computational Science: Imaging and Biomedical Computing. It has been very interesting to work with ..

I would especially like to thank my supervisor Pål for ..

Thanks to my internal supervisor Professor Anne H. Solberg and the rest of the DSB group, for....

Finally, I would like to thank my parents for their encouragement and everlasting love.

Contents

List of Figures

List of Tables

List of Symbols

Chapter 1

Introduction

1.1 Background and Motivation

In this project we aim to design and develop a system for analyzing medical videos from a camera pill, as seen in Figure 1.1. The pill is swallowed and records video of the entire digestive system The goal is to be able to detect different irregularities in the patients digestive system, like a colon polyp, Chron's disease, Colorectal cancer, etc. by using video object tracking, object detection, machine learning or other relevant tools.

Neural networks models that we would like to explore further for this purpose are Convolutional neural networks (CNN), Recurrent neural networks (RNN), Capsule neural networks, Long Short-Term memory networks and more.

The main idea is to go beyond image-based methods and also exploit the time factor of the data. The videos we will be using for this is delivered by Bærum Hospital, and is carefully labeled by using tools such as described in the paper "Expert Driven Semi-Supervised Elucidation Tool for Medical Endoscopic Videos". In this paper Albisser *et al.* presents a semi-supervised method to gather the annotations in a easy and time saving way [1].

Figure 1.1: Illustration of how such a camera pill could look like [2].

1.2 Problem statement

Colorectal cancer (CRC) is the third most common cause of cancer mortality for both men and women [3], and it is a condition where early detection is of clear value for the ultimate survival of the patient. As statistics show that 15% of male and female above 50 years are at risk, the

procedure is recommended on a regular basis (every 3-5 years) for the population over 50, and from an earlier age for high-risk groups.

Colonoscopy is a demanding procedure requiring an significant amount of time by specialized physicians, in addition to the discomfort and risks inherent in the procedure. Traditional methods based on colonoscopy are not cost-effective for population-based screening purposes, so only about 2-3% of the target population is reached at present.

The cost of a population screening program is prohibitively expensive. Colonoscopy is the most expensive cancer screening process in the US, with annual costs of $10 billion dollars ($1100 per person). In Norway we have similar costs of around $1000 per person, with a time consumption of about 1 doctor-hour and 2 nurse-hours per examination.

By researching an automatic system for a camera pill the aim is to greatly increase the number of patients that can be examined, i.e., making the public health care system more scalable and cost effective, while at the same time reducing the need for intrusive procedures like "bottom-up" examinations like colonoscopy.

1.3 Scope and limitations

Based on the described problem statement the scope of this thesis is to compare some prior selected automatic systems with focus on machine learning, for classification tasks in the medical domain. We will test the systems on video and still images taken from a wireless capsule endoscopy system, provided by Bærum Hospital. This data is generated with PillCam^{TM} SB 3 System [1] and gives us a wide view of the GI-tract. A limitation for our research will therefore be that our dataset lack a great deal of diversity. While we might have a wide range of patients all the data contributions stem from one provider in one location. This might not be much of an issue due to GI-tract being quite homegenous over large populations. But can presumably cause biasing issues in our deep learning models.

By using a tool provided by Augere Medical, we have sequentially classified each frame in some 45 PillCam videos. Although being assisted by doctor Thomas de Lange, with many years of experience, the dataset might not be perfectly labeled due to time constraints. But to the best of our efforts we managed to divide the data into 10 classes which we later will train our networks on.

Considering the scope of this thesis, we will limit ourselves to use some of the more common deep convolutional neural networks and a more novel stacked hourglass network. There are many other more common network used in the field of object classification like NN, RNN, stastisctial methods, but these methods are widely tested and we wish to push the envelope of machine learning systems on medical videos and also have a wide variety of ways to score our model against previously tested algorithms.

1.4 Research methods

We have based this paper on Association for Computing Machinerys (ACMs) research methodology. In the spring of 1986 ACM president Adele Goldberg and ACM Education Board Chairman Robert Aiken appointed a task force with the prime objective of describing the core fundamentals of computer science and computer engineering [4]. They introduced the phrase "discipline of computing" to embrace the two fields. The task force were given three objectives to complete.

1. Present a description of computer science that emphasizes fundamental questions and significant accomplishments. The definition should recognize that the field is constantly changing and that what is said is merely a snapshot of an ongoing process of growth.

2. Propose a teaching paradigm for computer science that conforms to traditional scientific standards, emphasizes the development of competence in the field, and harmoniously integrates theory, experimentation, and design.

3. Give a detailed example of an introductory course sequence in computer science based on the curriculum model and the disciplinary description.

[1] `https://www.medtronic.com/covidien/en-us/products/capsule-endoscopy/pillcam-sb-3-system.html`

To fully elaborate the core fundamentals of computer science and computer engineering they agreed upon no core fundamentals is different in the two fields, but the difference is manifested in the way the two fields elaborate the core; computer science focuses mainly on analysis and abstraction and computer engineering focuses mainly on abstraction and design.

Furthermore the task force appointed three major paradigms which represent different areas of competence in the field. Some will argue that the different paradigms are implicitly based on an assumption that one of the three processes is the most fundamental, but as we will see, the three paradigms are so intricately intertwined that it is irrational to say that one is the most fundamental. The three paradigms, or cultural styles, are listed below.

1.4.1 Theory

The first paradigm, Theory, is deeply rooted in mathematics and is concerned with the ability to describe and prove relationships among object. The paradigm consist of the following four steps.

1. Characterize objects of study (definition);
2. Hypothesize possible relations among them (theorem);
3. Determine whether the relationships are true (proof);
4. Interpret the results found.

When working in this paradigm it is expected to follow this reasoning and iterate the steps when errors or inconsistencies are discovered, until they can be explained and interpreted.

1.4.2 Abstraction

The second paradigm is abstraction. Abstraction is a form of modeling and is rooted in the experimental scientific method. This paradigm is concerned with the ability to use the relationships found in the theory paradigm to make predictions that then can be compared to the real world. It has four steps for which a scientist is expected to follow.

- Form a hypothesis;
- Construct a model and make a prediction;
- Design an experiment and collect data;
- Analyze the results.

1.4.3 Design

The third paradigm, design, is rooted in engineering and consist, like the others, of four main steps listed below. The design paradigm is concerned with the ability to implement specific instances of those relationships and use them to perform useful actions. The following four steps will help an engineer to construct a device or system to complete a given task.

- State requirements;
- State specifications;
- Design and implement the system;
- Test the system.

An engineer is to iterate these steps until the results from the testing satisfy the requirements and specifications of the system.

1.5 Main Contributions

1.6 Thesis Outline

This thesis is split into five chapter. Chapter one and two are mostly to introduce the reader to the topic and to fill in the necessary knowledge to understand the rest of the thesis. In the last chapter we conclude on our findings and discuss our findings and propose further work. The papers that have been referenced in the thesis is added in the bibliography at the very end. The chapters in the thesis is summarized below:

- In Chapter 2 we discuss the literature that focus on the topic of automated lesion detection in computer systems.
- In Chapter 3 we present the details of design, implementation of system and the processing and collection of data.
- In Chapter 4 we present the experiments we have conducted and..
- In Chapter 5 we provide a comprehensive overview of the results found and discus what that contributes to the field and propose some further work.

Chapter 2

Background

In recent years there have been many proposed methods to use automated object tracking, segmentation, deep learning and artificially intelligence to produce a better, and cheaper health care system. Many of the methods used are state of the art systems within the fields of deep learning. One requirement for such a system to work in reality is a good flow of data. Ideally all the data should be labeled by a doctor before it is used for training deep neural networks but this is rarely the case. The method which we propose takes advantage of this unlabeled data which is more readily available.

In this chapter we will present the necessary background and related works to understand how such a semi supervised model can be built. This will be covered over two main parts; one where we go through the related background and works to understand the medical aspect of this topic and the other will cover the technical use of deep learning in mission-critical fields such as the medical domain.

We begin with the digestive system and how it operates to aid the human body with digestion of food. Next we will cover disease detection by using various types of endoscopes. We will look at how the current state of lesion detection and how it could be improved by using deep learning.

In the next part will focus on deep learning and its various architectures and building blocks. To fully understand this we need to have a look at its inner workings and output. We begin with looking at a basic three layer neural network and build from there up to CNNs and some of the most advanced architectures most recently proposed. This will give a good understanding of how and why we use deep learning to classify medical images.

2.1 Medical scenario

To detect irregularities in the digestive system (Figure 2.1) is a difficult and time-consuming task. To classify irregularities correctly and precisely require expert knowledge. To fully understand the necessity of an automated system for detection lesions in the GI-tract we will go through the medical aspect of our problem statement, beginning with the anatomical explanation of the digestive tract. Then we will get to know the details of lesions in the small intestine, and the equipment currently in use to observe them.

2.1.1 The digestive system

The digestive system is made up of the gastrointestinal tract (GI tract), and the liver, pancreas and gallbladder. The GI tract is a series of hollow organs joined in a long and twisting tube beginning at your mouth and end with the anus, covering a distance of about 9 meters. It can be so long because the small intestine is very twisty. The GI tract is controlled by the brain by nerves and hormones. The organs that make up the GI tract is the mouth, esophagus, stomach, small intestine, large intestine and rectum.

The main purpose of the digestive system is so that the cells in the body can extract the nutrients from the food we eat and dispose of the waste which the body can't process. Special cells helps absorb the nutrients and cross the intestinal lining into the bloodstream. The circulatory system carries simple sugars, vitamins, salts, amino acids and glycerol to the liver which processes, stores, and deliver them back into the circulatory system which transports the nutrients to wherever

in the rest of the body it is needed. The body uses amino acids, fatty acids and sugars to build substances needed for growth, energy and cell repair for example.

Clinicians commonly divide the gastrointestinal tract in upper and lower regions called upper gastrointestinal tract and lower gastrointestinal tract. The upper gastrointestinal tract consist of mouth, esophagus, stomach and duodenum while the lower gastrointestinal tract consist of most of the small intestine, large intestine and rectum. Each organ in the GI tract helps to move the food and liquid forward throughout the body while its being broken into smaller parts. Next we will explain the function for each organ in the GI tract in the order of which food is processed.

1. **Mouth**; this is where food enters the GI tract and where the digestive process begin. After being split apart by chewing the food is swallowed and enters the esophagus.

2. **Esophagus**; after the swallow the brain signals the esophagus to begin the peristalsis, which is the process of contraction and relaxation of muscles that propagates the food (now called *bolus*, a ball of saliva and food) down towards the stomach. At the bottom of the esophagus you'll find a sphincter which opens to let the food into the stomach and normally keep the fluids in the stomach from traveling back up the esophagus.

3. **Stomach**; upon entering the stomach the stomach muscles begin to mix the bolus with gastric acid which begins the digestion of proteins. The stomach is lined with gastric folds, which helps the stomach to expand to hold about one liter of food. After an hour or two the pyloric valve opens and the contents (called *chyme*, a liquid of partially digested food and acids) are emptied into the small intestine.

4. **Small intestine**; the small intestine mix chyme from the stomach with digestive juices from the pancreas, liver and intestine and push the mixture forward for further digestion. The small intestine is divided into three sections; duodenum, jejunum and ileum. The walls of the small intestine, covered with intestinal villi (to increase the absorption area), absorb 95% of the nutrients, and carries it to the bloodstream. Whats left, the waste product, move into the large intestine by the peristalsis forces.

5. **Large intestine**; undigested parts of food, fluids and old cells from the GI tract lining enters the large intestine. The large intestine absorbs water, salts, sugars and vitamins back into the blood in the colon and changes the waste from liquid into stool.

6. **Rectum**; the rectum stores stool until it is pushed out of anus during a bower movement.

2.1.2 Colon cancer and other diseases

The GI tract may be home to a multitude of diseases.

Given the severity of colorectal cancer we have dedicated this section to give some further insight into this particular disease.

morbidity and mortality rates in the united states

These diseases have varying patterns and while some can be easy to split apart, some are more similar in pattern. While for an untrained eye it can be easy to spot that something is wrong it is very difficult to describe with words what that might be or even harder to write a program to detect the correct characteristic features of the disease. This is why we rely on having good amount of labeled data for this project to work. For increased accuracy we may have to look closer at a couple of diseases. Preferably two irregularities that both have different characteristics and lots of labeled training data.

2.1.3 Traditional Endoscopy

The most common way of screening patients is with a endoscope. When this tool is used by a professional some of the irregularities that can be spotted are; *Colon polyp*, *Colorectal Cancer*, *Ulcerative Colitis*, *Crohn's Disease*, *Familial adenomatous polypsis*, *Diverticulosis* and *Diverticula Bleeding*. See section 2.1.2 for a more detailed list of diseases.

[1] By Mariana Ruiz, edited by Joaquim Gaspar. Released into public domain by author.
`https://en.wikipedia.org/wiki/File:Digestive_system_diagram_edit.svg`

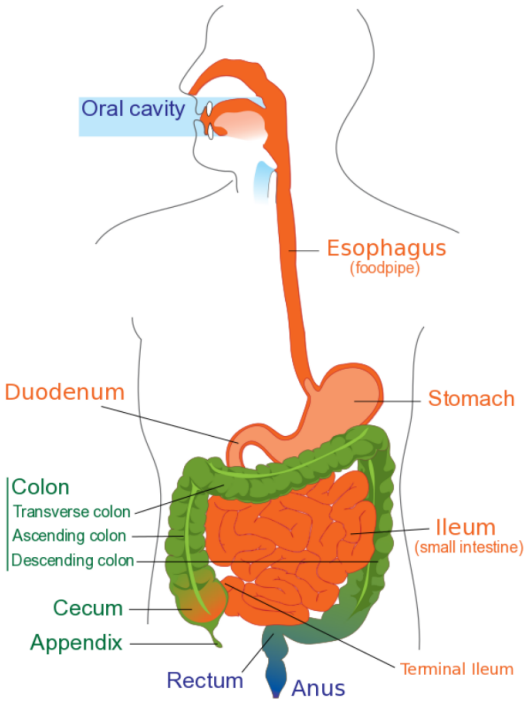


Figure 2.1: An overview of the terms used to describe the digestive system[1].

The basic technology behind the modern endoscope was developed in the early 1950s by English physicist Harold Hopkins and his student Narinder Kapany which let light travel through flexible pieces of glass, now known as optical fibers [5]. These fibers, as many as 50 000 optic fibers, can be packed very dense and allow for light to be transported over long distances with a high resolution. Later iterations of the endoscope allows for recording images through an added camera recorder connected at the end of the tool, water pipes, control cables and operation channels. See Figure 2.2 for a detailed look at the endoscope and its functions.

The clever design of the tool allows it to be used for both ends of the GI tract, but also ears, nose and urinary tract. See table 2.1 for a more detailed list of endoscope types. There also are some special forms of endoscopy which combines an endoscope with other medical applications, like fluoroscopy and ultrasound, to take medical imaging of special tricky parts of the body.

When the endoscope is inserted into the mouth and throat it is called upper endoscopy and if it is inserted through the anus it is called lower endoscopy.

Upper endoscopy

An upper endoscopy is a procedure used to examine the upper gastrointestinal tract, that is the mouth, esophagus, stomach and duodenum (the beginning of the small intestine). A specialist, called a gastroenterologist, use endoscopy to diagnose and, sometimes, treat conditions that affect the upper part of the digestive system. Upper endoscopy is often performed while the patient is conscious. But sometimes the patient receives a local anesthetic in the form of a spray to the back of the throat, or the patient can be sedated. It is sometimes performed in the hospital or emergency room to identify acute bleeding and problems with swallow and breathing.

Lower endoscopy

An lower endoscopy is a procedure used to examine the lower gastrointestinal tract, which is most of the small intestine, the large intestine and the rectum. The procedure may include rectum and entire colon, in which case it is a colonoscopy, or just the rectum and sigmoid colon, then it is called a sigmoidoscopy. Treatments that may be performed in the lower digestive system include biopsy (collecting tissue sample), polyp removal, cauterize a bleeding vessel and other medical procedures.

An endoscopy is usually a safe procedure, and the risk of serious complications is very low. Rare complications are; an infection in the part of the body the endoscope is used, or piercing or tearing in an organ, or bleeding, or reaction to the sedation used.

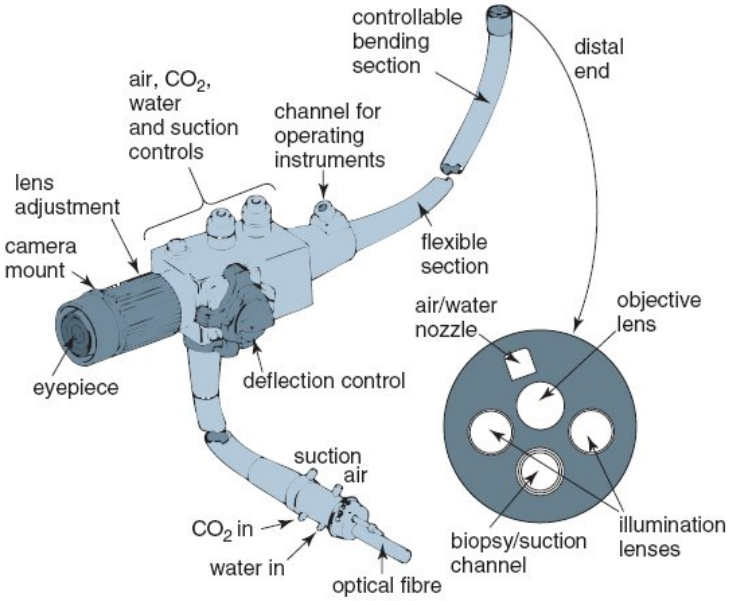


Figure 2.2: Image of a fiber optic endoscope with explanation of different parts of the tool[2].

2.1.4 Wireless Capsule Endoscopy

Before the year 2000 the only option you had to visualize the food pipe, stomach, duodenum, colon and terminal ileum (see Figure 2.1 for details) was to use a fiber-optic endoscope. These cables have to carry fiber optic bundles, water pipes, operations channel and control cables. Although these cables can be quite flexible there is a limit for how far they can advance into the small bowel. This method cause pain and discomfort for the patient, and there was a clinical need for an improved methods.

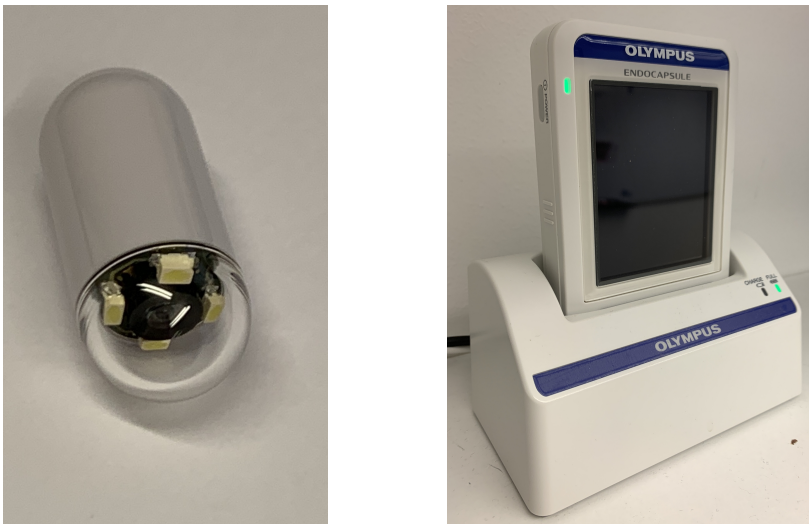
That is why in the year 2000 Iddan *et al.* developed a new type of video-telemetry capsule endoscope which the patients were able to swallow [6]. It could travel through the entire digestive system because it had no external wires, fiber-optic bundles or cables of any sort. The capsule travels by peristalsis[3] through the gastrointestinal tract, which takes from 10 to 48 hours, and transmit images on a regular interval to receivers attached around the outside of the patients stomach for as long as the battery allows, usually in the range 6 to 8 hours. Two example images taken by WCE are presented in Figure 2.4. By triangulating the signal strength and the location of the receivers taped on the body it is possible to roughly estimate the position of the capsule. This is however not very precise and can not tell us the rotation or direction of the capsule. Regardless, that information will not be available for us in this study as we only have access to the images themselves. By looking at some of the anatomical landmarks in the images we still might be able to predict when the capsule exits the stomach through the pylorus.

There is also ongoing research done in the field of map prediction (see section 2.3.3)which could be of great interest for WCE technology as it would allow us to better predict where inside a patient a disease is found.

The WCE devices come in a variety of different versions. Depending on travel speed through the GI tract, the purpose of the device and the localization it will capture between 1 and 30 images every second, produced with pixel resolution in the range of 256x256 to 512x512. They

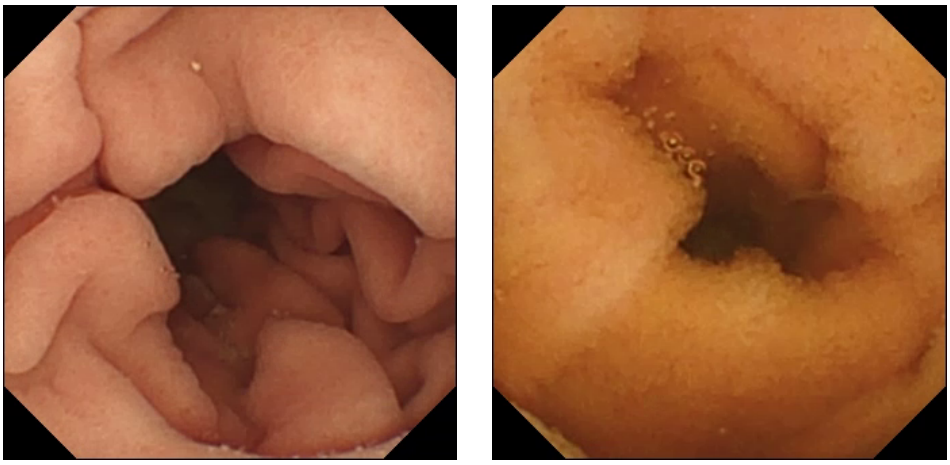
[2] Image credit: Jacaranda Physics 1 2nd Edition © John Wiley & Sons, Inc.

[3] Peristalsis is a radially symmetrical contraction and relaxation of muscles that propagates in a wave down a tube, in an anterograde direction.



(a) Olympus EC-S10 endocapsule **(b)** Olympus RE-10 endocapsule recorder

Figure 2.3: Used WCE equipment.



(a) Stomach **(b)** Small intestine

Figure 2.4: Images from Kvasir PillCam (See Section 3.1.1) dataset taken with WCE.

are specialized for different parts of the GI-tract and come from different vendors. The most known manufactures are Given Imaging (Medtronics), Ankon Technologies, Chongqing Science, IntroMedic, CapsoVision and Olympus.

The data used for this study is collected by the Olympus Endocapsule 10 System [4] using the Olympus EC-S10 endocapsule (Figure 2.3a) and the Olympus RE-10 endocapsule recorder (Figure 2.3b). This system has a 160° wide-angle lens, a light source, a minimum of 12 hours battery life (sometimes up to 20 hours), captures between 80 000 and 140 000 images and user friendly functionalities like Omni-selected Mode. Omni-selected Mode skips images that overlay with previous ones and therefore reduce review time for clinicians. To reduce drain on battery the light source will only emit light just as the camera is taking a picture. Its dimensions are 11 mm (diameter) x 26 mm (length) and it weight 3.3 gram.

[4] `https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html`

2.2 Deep learning

As apposed to using regular optic-fiber endoscopy, it can be difficult to know the location and orientation of the capsule when it is traveling through the digestive system. In a paper by Zou *et al.* it is shown that by using Deep Convolutional Networks (DCNN) it is possible to classify the digestive organs in wireless capsule endoscopy with about 95% classification accuracy on average [7]. The DCNN-based WCE digestive organ classification system is constructed of three stages of convolution, pooling and two fully-connected layers. This is illustrated in Figure 3 in the paper [7]. The main steps of this convolutional neural network are described in detail in section 2.2.7.

2.2.1 Imbalanced data

2.2.2 TensorFlow Framework

2.2.3 Data pipeline

2.2.4 Machine learning types

In deep learning it is common to differentiate between three types of machine learning models, supervised learning, unsupervised learning and reinforcement learning. In this section we will go through them and explain how they function and which use cases suites them best. In addition we will introduce a combination of supervised and unsupervised learning, called semi-supervised learning.

Supervised learning

The first category of machine learning is supervised learning. If you imagine yourself work under supervision of a leader or boss, it would mean someone is present and judging whether you are doing the correct work. Similarly to this, when a learning algorithm is under supervision is has a fully labeled dataset to work on; continuously updating the algorithm whether the answer is correct or wrong after every test.

Fully labeled dataset means that for every sample in the dataset, it is known what the true answer is to the problem at hand. Example if the dataset is images to classify you can think of it as having the correct answer written on the back of the image, but the algorithm will pick up the image front side up, and not look at the correct answer until after making a prediction.

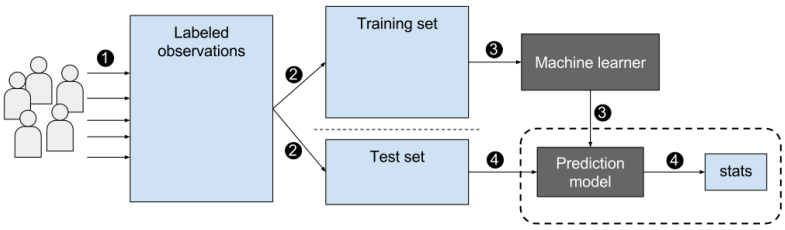


Figure 2.5: Workflow of supervised learning; 1. the dataset is labeled by observers; 2. the samples is split into training and test sets; 3. algorithm is learning on the training set; 4. checking the predictions on the 'unseen' test set to understand how the model performs.

This method is best suited for classification problems and regression problems, where there is a set of available reference points or a ground truth with which to train the algorithm, but this is not always accessible, or too expensive to create.

Unsupervised learning

Large, cleanly labeled datasets are not does not always easy to come by. And sometimes the answers we are looking for are not discrete, but discontinuous and hard to define. This is where unsupervised learning comes in.

In unsupervised learning, the algorithm is handed non-labeled data without any instructions on what to do with it. It is the algorithms job to automatically find which features that best separates

the data and find a structure within it. An example of a problem well suited for unsupervised learning is; using anomaly detection to discover unusual data points in a dataset, like fraudulent bank transactions.

It is common to further categorize unsupervised learning into four additional groups;

- **Clustering**: The deep learning model looks for data that are similar to each other and group them together.
- **Anomaly detection**: Used to flag outliers in a dataset. Samples that does not fit well in with the rest.
- **Association**: The model looks at how a certain feature of a data sample correlates with other features.
- **Autoencoders**: Autoencoders take input data, compress it into code and then try to recreate that same input data only using the compressed code.

Since the training data has not been reviewed by a human beforehand it is difficult to say with certainty how good the final model perform like it is with supervised learning. But for problem areas where there is little to none labeled data it is a valuable tool.

Semi-supervised learning

This is not its own category, but a combination of the two categories just mentioned. It is good for dealing with problems where you have some labeled data and a lot of unlabeled data.

Many real world problems fall into this problem as large, fully labeled datasets are difficult to obtain. To create one is both expensive and time consuming and often require domain experts like analysts or doctors. Whereas unlabeled data is cheap and easy to collect and store.

Our problem is in this realm and is therefor also a good example of a semi-supervised problem. We have a relatively small dataset of labeled medical images and almost an unlimited quantity of unlabeled images.

The goal of the semi-supervised machine learning technique is to make best predictions on unlabeled data. This is done by first using a trained supervised model to best predict unlabeled data and then feed that back into the supervised learning algorithm as training data. Then use the newly trained model to make predictions on the new unseen data. To get the best result this process can be repeated until accuracy converges.

Reinforcement learning

In Reinforcement Learning (RL) algorithms learn how to react to the environment on their own and is neither supervised nor unsupervised. Instead the algorithm rely on being able to monitor response of its action and measure against a defined "reward".

Reinforcement learning is a type of machine learning where AI agents are attempting to find the optimal way through an environment, to accomplish a set goal or to improve on a specific task. As the agent take an action in the environment it receives a reward, as seen in Figure 2.6. If the action improved on the last agent state it gets a positive reward and if the new state of the agent is worse than the previous it get a negative reward. The goal is to predict which next step to take to get the biggest final reward.

To make these predictions the agent need to rely on what it has previously learned, and be able to explore uncharted territory. For example if the first option for the agent is to pick a left or right turn on a road which leads to two different cities, and it gets a positive reward for picking left, the agent will never explore the other city. Therefore the agent must try to maximize the cumulative reward and not only the immediate reward.

To achieve good cumulative reward the algorithm must iterate over the problem many times. For each iteration, and each round of feedback, the agents strategy incrementally improves. This works really good for problems that can be simulated, where iterating the problem only cost computer power. A good example of a good RL problem is video games and autonomous driving.

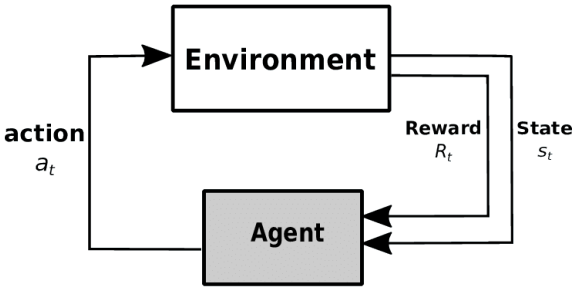


Figure 2.6: Reinforcement learning: Agent and environment.

2.2.5 How to evaluate a model

2.2.6 Performance metrics

2.2.7 Convolutional Neural Network

One of the most used neural networks for image classification is the Convolutional Neural Network (CNN). The model was first proposed by Krizhevsky *et al.* in 2012 [8] where they trained a deep convolutional neural network and used it to classify 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes with top-1 and top-5 error rates of 37.5% and 17.0% which far surpassed all other models at the time. Next we will get into a bit of the details of a CNN.

Convolution layer

The first step in a convolutional neural network is to extract features from the input image. This is done to preserve the relationship between pixels by learning image features using filters, or *kernels*. As a result, the network learn filters that activate when it detects some specific patterns or features.

The convolution of f and g is written as $f * g$, and is defined as the integral of the product of the two functions after one (usually the filter) is reversed and shifted.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \tag{2.1}$$

Non Linearity (ReLU)

Rectified Linear unit function, known as simply ReLU, is an activation function represented by equation (2.2). It sets all negative numbers to zero, by discarding them from the activation map entirely. In this way, ReLU increases the nonlinear properties of the decision function and thus of the overall network without affecting the receptive fields of the convolution layer.

$$ReLU(x) = max(0, x) \tag{2.2}$$

Pooling layer

Pooling layers are applied to reduce the number of parameters when the images are considerably large. Spatial pooling, or merely down sampling, reduces the dimensionality of each image but it keeps the important information. The most used down sampling is max pooling. It extracts the largest element from the rectified feature map and thus reduces computational complexity of the algorithm. In addition average pooling is also frequently used, this method computes the average value of the input map. The input-output model is denoted as:

$$y_i = f(pool(x_i)) \tag{2.3}$$

Fully-connected layer

In a FC-layer every neuron in one layer is connected to every neuron in the previous layer. It is here the high-level reasoning is done. The activation function in the neurons is a *sigmoid* or *tanh* function.

$$f(z) = \frac{1}{1 + exp(-z)} \quad \text{or} \quad f(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{2.4}$$

At the end of FC-layer we have an activation function such as softmax (equation 2.7) to calculate probability of the predicted classes.

Feed Forward

In the feed forward algorithm input image will be processed through all the layers in the neural network. The first layer will be a convolution layer, containing K filters F_i^1, $i = 1, ..., K$, of size $k \times k$ and a bias b^1. The image will be convoluted with each filter, and the bias is added.

$$\hat{z}_i^l = I * \hat{F}_i^l + b^l, \tag{2.5}$$

where $*$ (asterisk) is the convolution operator in equation 2.1. The final output of each convolutional layer l is a^l,

$$\hat{a}_i^l = f(z_i^l), \tag{2.6}$$

where f represents the ReLU activation function. After going through the convolution layer, the next layer could be a pooling layer, which will reduce the spatial dimensionality either by using the max value or the average value. Before getting our final output $\hat{y}$, we need to collect the outputs from all the filters, which will be an input to a fully connected layer. The fully connected layer use the softmax activation function to classify the input image, much like a neural network would. The softmax function is an accepted standard probability function for a multiclass classifier [9]. The total sum of the probabilities will always add up to 1 when using softmax.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, ..., K. \tag{2.7}$$

To calculate the error of the forward propagation it is common to use cross-entropy error function.

$$C(\hat{y}) = -\sum_{i=1}^{N} t_i log(y_i) \tag{2.8}$$

Back propagation

Starting from the last layer L, we calculate the derivative of the loss function (function 2.8) with regards to the activation function in order to update the weights. Computing the gradient of the loss function yields

$$\frac{\partial C}{\partial y_i} = -\frac{t_i}{y_i} \tag{2.9}$$

We also require the gradient of the output of the final layer y_i with regards to the input z_k^L of the activation function (equation 2.7)

$$\frac{\partial y_i}{\partial z_k^L} = \begin{cases} y_i(1 - y_i), & i = k \\ -y_i y_k, & i \neq k \end{cases} \tag{2.10}$$

Now with regards to z_i^L

$$\begin{aligned}\frac{\partial C}{\partial z_i^L} =& \sum_k^N \frac{\partial C}{\partial y_k}\frac{\partial y_k}{\partial z_i^L} \\ =& \frac{\partial C}{\partial y_i}\frac{\partial y_i}{\partial z_i^L} - \sum_k^N \frac{\partial C}{\partial y_k}\frac{\partial y_k}{\partial z_i^L} \\ =& -t_i(1-y_i) + \sum_{k\neq i} t_k y_i \\ =& y_i - t_i\end{aligned} \tag{2.11}$$

And finally with regards to the weights

$$\frac{\partial C}{\partial w_i j^L} = (y_i - t_i)a_j^{L-1} \tag{2.12}$$

where $\hat{a}_j^{L-1}$ is the vectorized output from the previous layer. From here, we will propagate the error throughout the layers. The error with regards to the input a_i^L to the fully connected layer is:

$$\delta^{L-1} = \frac{\partial C}{\partial a_i^L} = \sum_i^N (y_i - t_i)w_{ji}^L \tag{2.13}$$

Thus the error is propagated backwards through each layer. If max pooling was used in a pooling layer, the error will only be propagated to the input that had the highest value in the forward pass. The other values will be set to zero. If average pooling was used, the error is averaged in the backwards pass. In equation 2.13 a^l is the output of a convolutional layer l. Since a convolutional layer is always preceded and followed by a activation layer, the input to layer l is $a^{l-1} = \sigma(z^l)$. Now consider the error with regards to z^l.

$$\begin{aligned}\delta_{ij}^l =& \frac{\partial C}{\partial z_{ij}^l} \\ =& \sum_i^{'}\sum_j^{'} \frac{\partial C}{\partial z_{i'j'}^{l+1}}\frac{\partial z_{i'j'}}{\partial z_{ij}^l} \\ =& \sum_{i'}\sum_{j'} \delta_{i'j'}^{l+1}\frac{\partial(\hat{W}\sigma(z^l) + b^{l+1})}{\partial z_{ij}^l} \\ =& \delta^{l+1} * ROT180(w^{l+1})\sigma'(z^l)\end{aligned} \tag{2.14}$$

Having found the error, the gradient of the cost function with regards to the weights is

$$\frac{\partial C}{\partial w_{ij}^l} = \delta_{ij}^l * \sigma ROT180(z_{ij}^{l-1}) \tag{2.15}$$

2.3 Related work

Zhu *et al.* have made a computer-aided lesion[5] detection system which uses a trainable feature extractor, also based on a CNN, and feed the generic features to a Support Vector Machine which enhance the generalization ability [10]. This method greatly outperform the earlier methods based on color and texture features. However we believe that by using neural networks to do the decision making we can further improve this detection system.

Yuan *et al.* have accomplished an average overall recognition accuracy of 98.0% for detecting polyps in WCE images by using a deep feature learning method, named stacked sparse autoencoder with image manifold constraint (SSAEIM). This method is built on a Sparse auto-encoder (SAE), a symmetrical and unsupervised neural network. It is an encoder–decoder architecture where the encoder network encodes pixel intensities as low dimensional attributes, while the decoder step reconstructs the original pixel intensities from the learned low-dimensional features [11]. Detecting

[5] a region in an organ or tissue which has suffered damage through injury or disease, such as a wound, ulcer, abscess, or tumor.

colorectal polyps are important because they are precursors to cancer, which may develop if the polyps are left untreated. Where we hopefully can build on this method is by using a larger dataset with pathology proof of other irregularities.

Jia *et al.* present a new automatic bleeding detection strategy based on a deep convolutional neural network and evaluate their method on an expanded dataset of 10,000 WCE images. Gastrointestinal (GI) tract bleeding is the most common abnormality in the tract, but also an important symptom or syndrome of other pathologies such as ulcers, polyps, tumors and Crohn's disease. Their method for detecting bleeding have an increase of around 2 percentage in F_1 score, up to 0.9955 [12]. This method and its high score in somewhat limited to bleeding, and not very good at detecting other lesion. Our goal is to develop a method for using deep learning to find more generalized pathologies in the gastrointestinal tract.

We will go through some other methods not directly related to neural networks but which we think may come in handy for my thesis later on.

2.3.1 Object tracking

Object tracking is one of the harder problem to overcome in computer vision and is key to achieving good results in endoscopic video analysis. Tracking algorithms are developed to determine the movement of the object or objects in each video frame. The algorithm has to take into account the dynamic environment such as differences in lightning, occlusions and scaling changes. Also the absence of any prior knowledge to the object and its position further increase the complexity of the problem. Zhang *et al.* proposed an approach for visual tracking in videos that learns to predict the bounding box locations of a target object at every frame in the paper "Deep Reinforcement Learning for Visual Object Tracking in Videos" [13]. While other models depends on the capability of a CNN to learn a good feature representation for the target location in the new frame, which means that the model only tracks properly if the target lies in the spatial vicinity of the previous prediction. This is not always the case for WCE videos, where the lens of the camera can suddenly and unpredictably rotate towards the wall of the intestine. This method integrates convolutional network with recurrent network, and builds up a spatial-temporal representation of the video which means that the model is able to predict the target object's location over time.

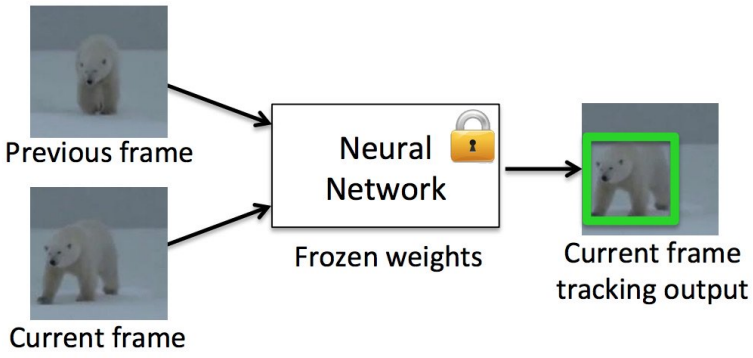


Figure 2.7: Illustration of how object in two frames is tracked with a bounding box[6].

Our hope is that by implementing an object-tracking algorithm we can use it to classify irregularities in the colonoscopy video, and then track that object in the later frames until it disappear out of frame. This will hopefully help with reducing the robustness of the network so that the classifier will not have to check every frame for irregularities.

2.3.2 Segmentation

Image segmentation is the process of partitioning a image into multiple segments of pixel, usually each segment describing some feature of the image or an entire object or class of objects. The goal of segmentation is to simplify the image and make it easier to analyze or further process.

[6] `https://www.learnopencv.com/goturn-deep-learning-based-object-tracking/`

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

Table 2.2: Segmentation results on the ISBI cell tracking challenge in 2015.

Ronneberger *et al.* propose a method in the paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" [14] for using a network and training strategy that relies on the strong use of data augmentation to use the available labeled samples more efficiently. This network outperform the old method of sliding-window-convolution by a great deal. They extend the "fully convolutional network" [15] such that it works with very few training images and yields more precise segmentations. The way this is achieved is to supplement a contracting network by successive layers, where instead of using pooling operators, upsampling operators are used. This means that these successive layers increase the resolution of the output. The high resolution features from the contracting path are combined with the upsampled output to localize objects and with that a convolution layer can then learn to produce more precise output based on this information.

Another important feature in this architecture is that in the upsampling portion of the network there is also large number of feature channels. These channels allow the network to pass on context information to the higher resolution layers.

A common problem in training neural networks are too little labeled training data. This is also the case for us. We require a lot of medical data, and personell with the expertise to correctly label our data are of high demand and they usually have very little time for projects like these. This is why Ronneberger *et al.* use different methods of data augmentation to generate more training data. They apply elastic deformations to the available images, and this allows the network to learn invariance to such deformations without the need to see these transformations in the annotated image corpus. Which is particular important in biomedical segmentation since deformation used to be the most common variation in tissue and realistic deformations can be simulated efficiently [14]. By doing this Ronneberger *et al.* were able to achieve very good results (Table 2.2).

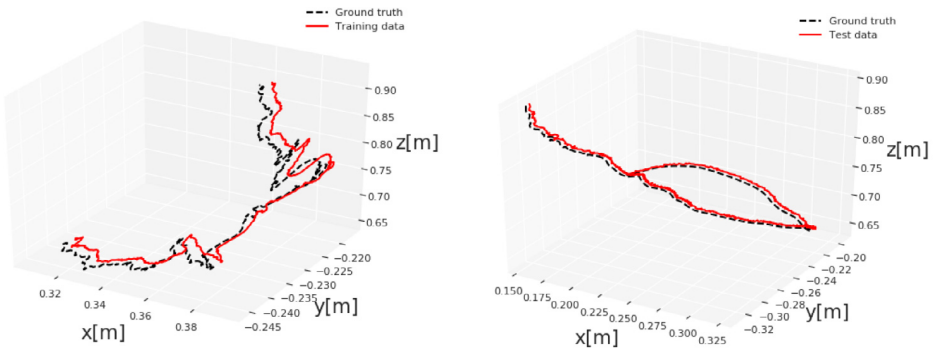
2.3.3 Mapping

As mentioned in section 2.1.4, a concern when processing the images taken with a WCE is not having the spatial data you get when using a normal fiber-optic endoscope. This is why Turan *et al.* has recently made substantial progress in converting passive capsule endoscopes to active capsule robots, enabling more accurate, precise, and intuitive detection of the location and size of the diseased areas by developing reliable real time pose estimation functionality of the capsule with RCNN's[7] [16]. See Figure 2.8 for an example.

This architecture uses inception modules for feature extraction and a RNN for sequential modelling of motion dynamics to regress the robot's orientation and position in real time. By taking multiple of RGB Depth images with time stamps it can calculate the 6-DoF pose of the capsule without the need of any extra sensors. For obtaining the depth images Turan *et al.* use the shape from shading (SfS) technique of Ping-Sing and Shah [17]. This model outperforms state-of-the-art models like LSD SLAM and ORB SLAM.

2.4 Summary

[7] Deep recurrent convolutional neural networks



(a) Training data vs ground truth. **(b)** Test data vs ground truth.

Figure 2.8: An example of Deep EndoVO accuracy [16].

Chapter 3

Methodology

3.1 Data collection

The datasets used in our experiments are Kvasir Pillcam, Kvasir and Hyper-Kvasir. This section will demonstrate the main differences between the three datasets and explain how they can be found and used for fact checking. All three datasets are collected using endoscopic equipment at Vestre Viken Health Trust in Norway. The VV consists of 4 hospitals and provides health care for 470.000 people. One of the hospitals is Bærum Hospital, which has a large gastroenterology department from where the data is collected.

3.1.1 Kvasir PillCam

The dataset we used in our experiments consist of endoscopic videos collected from Bærum Hospital. Unlike Kvasir and Hyper-Kvasir datasets we have made the Kvasir PillCam dataset for the purpose of this thesis. In total we have 44 videos which have gone through some re encoding to reduce the file sizes, and also because the original encoding is proprietary Sony technology. After that the videos are uploaded to Augere Medical [1] tagging tool. The data export from Bærum also contains some findings for each video (if there is any) and are extracted, converted to frame number and that helped us a great deal with tagging the videos. We also have Thomas de Lange to thank, because he helped us a lot with the medical aspect of the classification process. When all 44 videos have been precisely labeled the dataset is exported from Augeres tagging tool and split into folders for each class. The folders/classes are given in table 3.1. In total we have 44 000 labeled images in 8 classes. The sample distribution across the eight classes is skewed depending on how many findings there are in the videos. Some findings occur often and some very rarely. The dataset also contain one class for 'normal' images, which there is quite a bit more of than findings.

Imbalanced dataset pose a challenge for predictive algorithms as most learning algorithms are based on the assumption of an equal number of samples for each class. This results in models that have poor predictive performance, especially for minority class or classes. This is a great problem because in many medical datasets the minority class is the most important and therefore more sensitive for classification errors.

In addition to labeling the images the dataset also contain a JSON format file which stores coordinates for where in the frame the finding is located. The Kvasir Pillcam dataset will be an open-source dataset available for others scientists, and will later be grown to include more PillCam videos, both labeled and unlabeled samples.

[1] `https://augere.md/`

Class number	Class name
0	normal
1	polyp
2	polyrus

Table 3.1: PillCam class names and corresponding class numbers.

Class number	Class name	Number of samples
0	normal	8000
1	polyp	8000
2	polyrus	8000

Table 3.2: Kvasir class names and corresponding class numbers.

3.1.2 Kvasir

The Kvasir dataset [18] contains images from inside the gastrointestinal (GI) tract. The samples are classified into three important anatomical landmarks and three clinically significant findings. In addition it has two classes related to the removal procedure of polyps. The dataset is sorted and annotated is performed by medical doctors. The class names and findings for each class is given in table 3.2. One of the most important aspects of the Kvasir dataset is that it makes it easy to reproduce and compare results in scientific computing.

3.1.3 Hyper Kvasir

The Hyper-Kvasir dataset [19] is one of the largest medical datasets containing 110.079 images and 373 videos where it captures anatomical landmarks and pathological and normal findings. Resulting in more than 1.1 million images and video frames all together. The dataset contain four parts, labeled images, unlabeled images, segmented images and lastly, videos. In total the dataset is 70 GB in size, but can be downloaded and stored in parts from `https://datasets.simula.no/hyper-kvasir/`.

Labeled images

Hyper-Kvasir contains 10.662 labeled images. The images are split into 23 different classes, and are stored in a folder with the same name as its corresponding class. All of the images are stored in JPEG format [20], which means it has some image quality loss but quite insignificant compared to the reduction in file size. Like in situations most often encountered the classes has a different number of samples, this is a challenge in the medical field because some findings occur more often than others.

Unlabeled images

This part of the dataset contains 99.417 images

segmented images

In figure 3.1 we can see an example of the segmented Kvasir images.

Figure 3.1: Example of a segmented image from Hyper-Kvasir dataset.

Videos

3.2 Data process

3.2.1 Data preprocessing

3.2.2 Data pipeline

3.3 System implementation

3.4 Summary

Chapter 4

Experiments

4.1 Results

4.2 Summary

Chapter 5

Conclusions

5.1 Results

5.2 Summary and contributions

We have looked on some highly relevant papers written about automatic detection systems for medical videos from the last few years. From back when feature extraction methods consisted of selecting color and intensities thresholds, to newer and more sophisticated algorithms like CNN's have become mainstream. The newer methods may be more complex and harder to implement but we have found that these automatic feature extraction methods have a far higher accuracy and produce less false positives. We have also looked at the importance of having a big and varied dataset with labeled data. If the dataset is not large enough we can use several data augmentation methods to increase it, like the ones used in U-Net.

5.3 Discussion

5.4 Further work

Bibliography

[1] Z. Albisser, M. Riegler, P. Halvorsen, J. Zhou, C. Griwodz, I. Balasingham, and C. Gurrin, "Expert driven semi-supervised elucidation tool for medical endoscopic videos", in *Proceedings of the 6th ACM Multimedia Systems Conference*, 2015, pp. 73–76 (cit. on p. 1).

[2] *PillCam - A Camera Pill System for Automatic Screening of the Digestive System - Institutt for informatikk.* https://www.mn.uio.no/ifi/studier/masteroppgaver/nd/SRL-media-pill-cam.html (cit. on p. 1).

[3] A. Jemal, R. Siegel, J. Xu, and E. Ward, "Cancer Statistics", *CA: A Cancer Journal for Clinicians*, vol. 60, no. 5, pp. 277–300, 2010 (cit. on p. 1).

[4] D. E. Comer, D. Gries, M. C. Mulder, A. Tucker, A. J. Turner, P. R. Young, and P. J. Denning, "Computing as a discipline", *Communications of the ACM*, vol. 32, no. 1, pp. 9–23, Jan. 1989 (cit. on p. 2).

[5] A. C. S. Van Heel, "A New Method of transporting Optical Images without Aberrations", *Nature*, vol. 173, no. 4392, pp. 39–39, Jan. 1954 (cit. on p. 7).

[6] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy", *Nature*, vol. 405, no. 6785, p. 417, May 2000 (cit. on pp. 8, 9).

[7] Y. Zou, L. Li, Y. Wang, J. Yu, Y. Li, and W. J. Deng, "Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network", in *Proceedings of the 2015 IEEE International Conference on Digital Signal Processing (DSP)*, Jul. 2015, pp. 1274–1278 (cit. on p. 10).

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105 (cit. on p. 12).

[9] P. Sadowski, "Notes on backpropagation", 2016 (cit. on p. 13).

[10] R. Zhu, R. Zhang, and D. Xue, "Lesion detection of endoscopy images based on convolutional neural network features", in *Proceedings of the 2015 8th International Congress on Image and Signal Processing (CISP)*, Oct. 2015, pp. 372–376 (cit. on p. 15).

[11] Y. Yuan and M. Q.-H. Meng, "Deep learning for polyp recognition in wireless capsule endoscopy images", *Medical Physics*, vol. 44, no. 4, pp. 1379–1389, Apr. 2017 (cit. on p. 15).

[12] X. Jia and M. Q.-. Meng, "A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images", in *Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2016, pp. 639–642 (cit. on p. 15).

[13] D. Zhang, H. Maei, X. Wang, and Y.-F. Wang, "Deep Reinforcement Learning for Visual Object Tracking in Videos", *arXiv:1701.08936 [cs]*, Jan. 2017. arXiv: `1701.08936 [cs]` (cit. on p. 15).

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241 (cit. on p. 16).

[15] J. Long, E. Shelhamer, and T. Darrell,
"Fully Convolutional Networks for Semantic Segmentation",
in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440 (cit. on p. 16).

[16] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots", *Neurocomputing*, vol. 275, pp. 1861–1870, Jan. 2018 (cit. on p. 17).

[17] T. Ping-Sing and M. Shah, "Shape from shading using linear approximation", *Image and Vision Computing*, vol. 12, no. 8, pp. 487–498, Oct. 1994 (cit. on p. 17).

[18] K. Pogorelov, P. T. Schmidt, M. Riegler, P. Halvorsen, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, and M. Lux, "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection",
in *Proceedings of the 8th ACM on Multimedia Systems Conference - MMSys'17*, 2017, pp. 164–169 (cit. on p. 19).

[19] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. G. Ceja, P. T. Schmidt, H. L. Hammer, M. Riegler, P. Halvorsen, and T. de Lange, "Hyper-Kvasir: A Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy", Open Science Framework, Preprint, Dec. 2019 (cit. on p. 19).

[20] G. Wallace, "The JPEG still picture compression standard",
IEEE Transactions on Consumer Electronics, vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992 (cit. on p. 19).