Chapter 1

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.

1.1 Medical scenario

To detect irregularities in the digestive system (Figure 1.1) 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.

1.1.1 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.

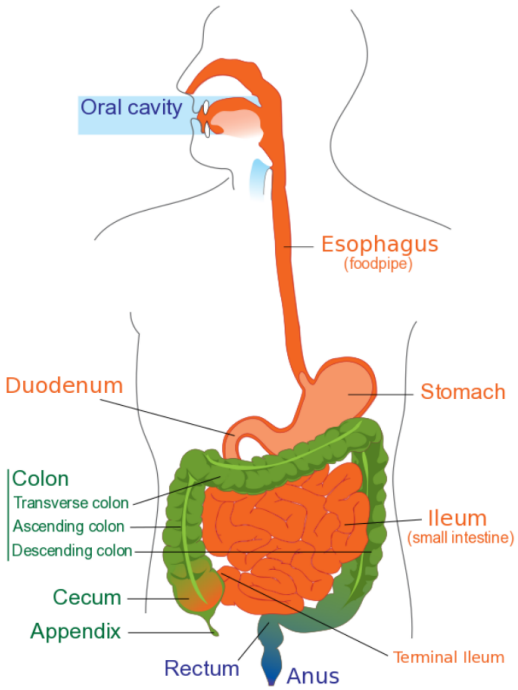


Figure 1.1: An overview of the terms used to describe the digestive system[1].

1.1.2 Colorectal Cancer and Screening

The GI tract may be home to a multitude of diseases, including infections, inflammations and cancers. Given our problem statement and the severity of the disease, we are going to focus on colorectal cancer (CRC). See section **??** for list of other diseases from the datasets we have used during our experiments. One of the most substantially significant factors for lowering morbidity and mortality in GI tract diseases are early screening and treatment [1], [2]. And in this section we will therefore explain the importance of screening and the difficulties that the current methods inflict upon the medical sector.

A study from 2014 found that CRC were the leading cause of cancer death in the United States in the late 1940 and early 1960 [2], but CRC mortality has since been slowly decreasing due to historical changes in risk factors (E.g decreased smoking and

[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`

red meat consumption) and better use of screening and early treatment. Today CRC is the third most common cause of cancer death in both men and women.

Another study used a microsimulation called MISCAN-COLON [3] to simulate the 2000 U.S population with regards to the CRC risk factor prevalence, screening use, and treatment use. They used the model to project age-standardized CRC mortality from the year 2000 to 2020 for 3 intervention scenarios and found that without any changes the risk factor would decrease by 17% by the year 2020. However, if the use of screening was improved to 70% of the population and the use of chemotherapy increased for all age groups, then the reduction of CRC mortality was estimated to be close to 50% by the year 2020. They found that the highest contributor to the reduced mortality rate was high level of screening (23%).

At the current state of screening the patient is relying on a doctor's ability to correctly spot early signs of cancer, most commonly polyps (See Figure 1.2), which are abnormal tissue growth often taking the shape of a mushroom. This is a problem as it has been proven that who perform the procedure can be more important than the most important health factors like age and gender [4]. Most screening occurs through endoscopy examinations and is uncomfortable for the patient and cost XXX nurse hours and XXX US dollars. This could be improved by the use of cheaper screening methods like WCE and Artificial Intelligence (A.I). A capsule could be picked up at the local pharmacist, swallowed and then the data sent back to the hospital to be analyzed.

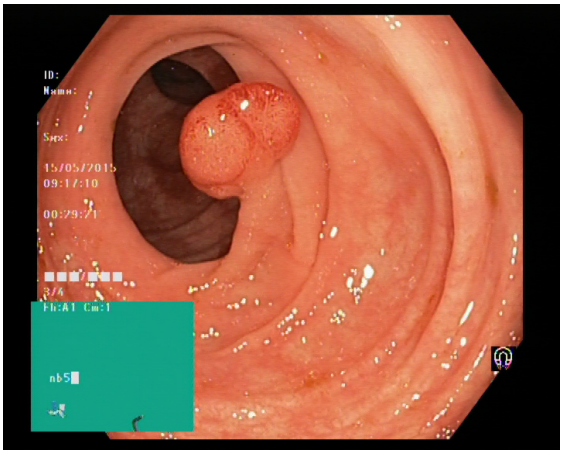


Figure 1.2: Image from Kvasir-V2 dataset of a polyp in the colon, taken with a fiber-optic endoscope (Section 1.5.2).

1.1.3 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 1.1.2 for a more detailed list of diseases.

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 1.3 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 1.1 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.

1.1.4 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 1.1 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 method.

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

Procedure	Name of tool	Area/organ viewed	Insertion point
Anoscopy	Anoscope	Anus and/or rectum	Anus
Arthroscopy	Arthroscope	Joints	Incision at the joint
Bronchoscopy	Bronchoscope	Trachea, windpipe and the lungs	Mouth
Colonoscopy	Colonoscope	Colon and large intestine	Anus
Colposcopy	Colposcope	Vagina and cervix	Vagina
Cystoscopy	Cystoscope	Inside of bladder	Urethra
Esophagoscopy	Esophagoscope	Esophagus	Mouth
Gastroscopy	Gastroscope	Stomach, duodenum	Mouth
Hysteroscopy	Hysteroscope	Uterus	Vagina
Laparoscopy	Laparoscope	Stomach, liver or other abdominal organs	Incision in the abdomen
Laryngoscopy	Laryngoscope	Larynx	Mouth
Neuroendoscopy	Neuroendoscope	Areas of the brain	Incision in the skull
Proctoscopy	Proctoscope	Rectum and sigmoid colon	Anus
Sigmoidoscopy	Sigmoidoscope	Sigmoid of colon	Anus
Thoracoscopy	Thoracoscope	Pleura	Incision in the chest

Table 1.1: List of the most common types of endoscopy.

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, a radially symmetrical contraction and relaxation of muscles that propagates in a wave down through the gastrointestinal tract. This process takes from 10 to 48 hours. For as long as the battery allows, usually in the range of 6 to 15 hours, the capsule transmits images on a regular interval to eight abdominal receivers and stores the data on a portable solid state recorder, which is carried on a belt. Some vendors, of which CapsoVision is one, have opted for a design which uses local flash storage to save the collected images directly on the device and therefore eliminates the need for abdominal receivers and wireless transmission of data. Writing data directly to flash storage has some drawbacks: (1) not possible to observe the area being imaged before after the capsule has passed, and (2) the need for a special docking station that enables access to the flash storage.

Endoscopic capsules are divided by terms of their application and is used to diagnose: (1) the esophagus; (2) the small intestine; (3) the large intestine. Depending on application they differ in areas like operating time, imaging frequency and number of cameras. To diagnose the esophagus the capsule travels a short distance in a short time and it is common to use a capsule with cameras on opposite ends, and capture images in high frequency. This comes at a cost of operating time. For a clinician to diagnose the small and large intestine the most significant feature is operating time, and it is therefore common to use a single camera with lower imaging frequency to reduce the drain on battery.

Two example images taken by WCE are presented in Figure 1.5. 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

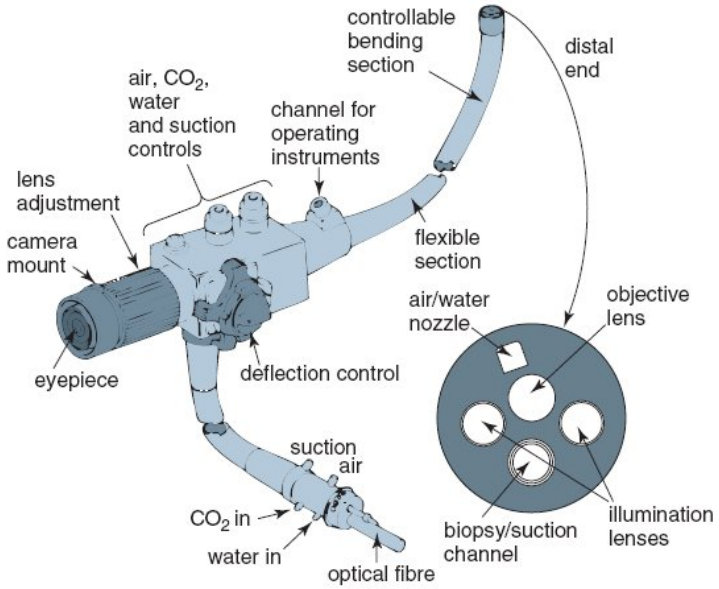


Figure 1.3: Image of a fiber optic endoscope with explanation of different parts of the tool[2].

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.

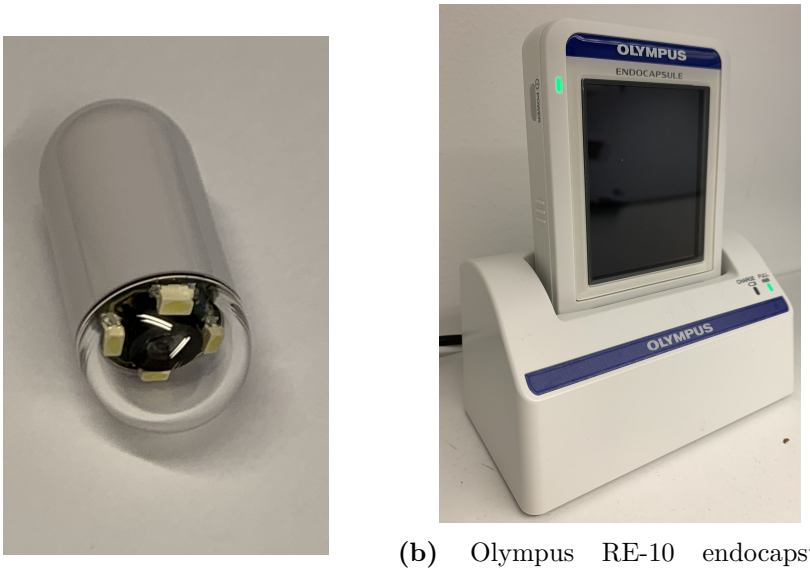
There is ongoing research done in the field of map prediction (see section 1.7.3)which could be of great interest for WCE technology as it would allow us to better predict the location of a disease inside the patients gut, as well as enable the clinician processing the video to see the orientation of the capsule.

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 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[3] using the Olympus EC-S10 endocapsule (Figure 1.4a) and the Olympus RE-10 endocapsule recorder (Figure 1.4b). 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.

A typical video collected by WCE examination lasts a few hours. A clinician must

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

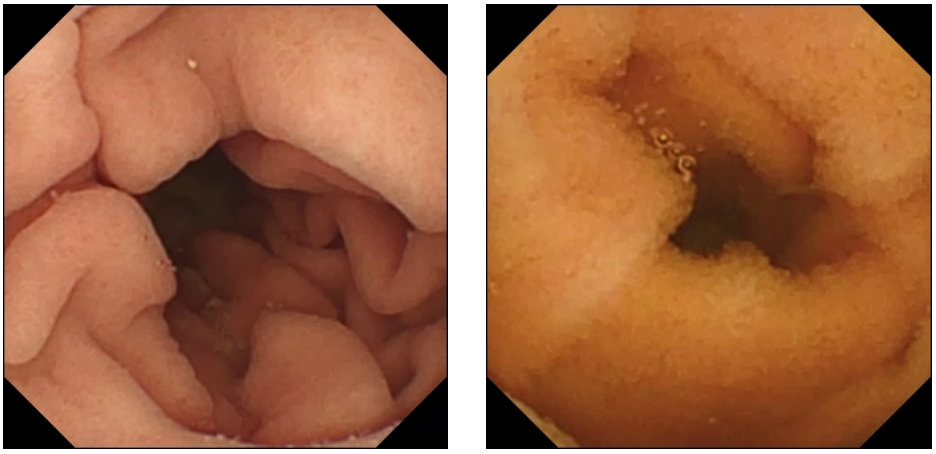


(a) Olympus EC-S10 endocapsule **(b)** Olympus RE-10 endocapsule recorder

Figure 1.4: Used WCE equipment.

watch the entire video to make a diagnosis because in a typical clinical situation there is no indication of which part of the GI-tract they need to search for damaged tissue, polyps, bleeding, etc. The capsule moves through the tract by two forces, gravity and bowel movements. In the small intestine there is two types of bowel movements: (1) peristaltic and (2) staple (segment). The first type is responsible for transit of food and is pretty linear movement, while the latter is responsible for mixing of food and is therefore much more chaotic in nature. These movements sometimes ceases temporarily as the muscles in the intestine relaxes. The result is an video which is highly diverse - moments of stillness, camera obscured by food debris and moments of chaotic movements and therefor rapid changes in the imaging area. As such the clinician watching the video will often have to speed up the footage, slow it down, and sometimes watch it frame by frame. Consequently, there is ongoing research related to the implementation of image analysis and processing methods allowing automatic video analysis. Such a automatic analysis system could greatly shorten the time for diagnosis and reduce the cost related to clinician salary. In practice this means that the clinician watch a few minutes of video with the pathologies detected by the software. To understand how such a software could be created we need to take a look at deep learning, which is discussed in the next section.

1.1.5 Remote diagnostic



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

Figure 1.5: Images from Kvasir PillCam (See Section **??**) dataset taken with WCE.

1.2 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 1.2.2.

1.2.1 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. As an 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.

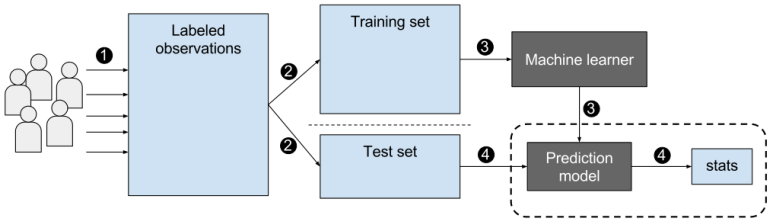


Figure 1.6: Workflow of supervised learning; 1. the dataset is labeled by observers; 2. the samples is split into training and test sets; 3. algorithm is learning on the training set; 4. checking the predictions on the 'unseen' test set to understand how the model performs.

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 1.7. 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.

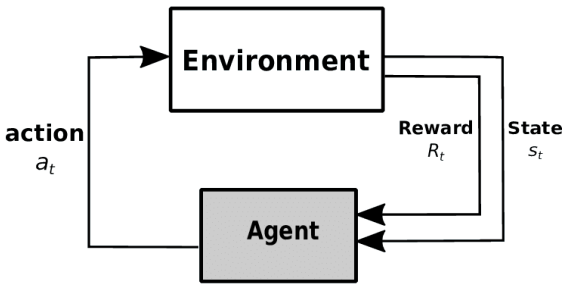


Figure 1.7: Reinforcement learning: Agent and environment.

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.

1.2.2 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.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \tag{1.1}$$

Non Linearity (ReLU)

Rectified Linear unit function, known as simply ReLU, is an activation function represented by equation (1.2). 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.

$$ReLU(x) = max(0, x) \tag{1.2}$$

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:

$$y_i = f(pool(x_i)) \tag{1.3}$$

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.

$$f(z) = \frac{1}{1 + exp(-z)} \quad \text{or} \quad f(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{1.4}$$

At the end of FC-layer we have an activation function such as softmax (equation 1.7) 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.

$$\hat{z}_i^l = I * \hat{F}_i^l + b^l, \tag{1.5}$$

where $*$ (asterisk) is the convolution operator in equation 1.1. The final output of each convolutional layer l is a^l,

$$\hat{a}_i^l = f(z_i^l), \tag{1.6}$$

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.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, ..., K. \tag{1.7}$$

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

$$C(\hat{y}) = -\sum_{i=1}^{N} t_i log(y_i) \tag{1.8}$$

Back propagation

Starting from the last layer L, we calculate the derivative of the loss function (function 1.8) with regards to the activation function in order to update the weights. Computing the gradient of the loss function yields

$$\frac{\partial C}{\partial y_i} = -\frac{t_i}{y_i} \tag{1.9}$$

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 1.7)

$$\frac{\partial y_i}{\partial z_k^L} = \begin{cases} y_i(1-y_i), & i = k \\ -y_i y_k, & i \neq k \end{cases} \tag{1.10}$$

Now with regards to z_i^L

$$\begin{aligned} \frac{\partial C}{\partial z_i^L} =& \sum_k^N \frac{\partial C}{\partial y_k}\frac{\partial y_k}{\partial z_i^L} \\ =& \frac{\partial C}{\partial y_i}\frac{\partial y_i}{\partial z_i^L} - \sum_k^N \frac{\partial C}{\partial y_k}\frac{\partial y_k}{\partial z_i^L} \\ =& -t_i(1-y_i) + \sum_{k\neq i} t_k y_i \\ =& y_i - t_i \end{aligned} \tag{1.11}$$

And finally with regards to the weights

$$\frac{\partial C}{\partial w_i j^L} = (y_i - t_i)a_j^{L-1} \tag{1.12}$$

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:

$$\delta^{L-1} = \frac{\partial C}{\partial a_i^L} = \sum_i^N (y_i - t_i)w_{ji}^L \tag{1.13}$$

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 1.13 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.

$$\begin{aligned} \delta_{ij}^l =& \frac{\partial C}{\partial z_{ij}^l} \\ =& \sum_i^{'}\sum_j^{'} \frac{\partial C}{\partial z_{i'j'}^{l+1}}\frac{\partial z_{i'j'}}{\partial z_{ij}^l} \\ =& \sum_{i'}\sum_{j'} \delta_{i'j'}^{l+1}\frac{\partial(\hat{W}\sigma(z^l) + b^{l+1})}{\partial z_{ij}^l} \\ =& \delta^{l+1} * ROT180(w^{l+1})\sigma'(z^l) \end{aligned} \tag{1.14}$$

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

$$\frac{\partial C}{\partial w_{ij}^l} = \delta_{ij}^l * \sigma ROT180(z_{ij}^{l-1}) \tag{1.15}$$

1.2.3 Gradient descent optimization algorithms

Adam

SGD

Adagrad

Adadelta

1.3 Popular network architectures

1.3.1 ResNet

Residual Netowrk (ResNet) is in a way an upgraded form of previously mentioned Convolutional Neural Networks. This architecture enable the model to have hundreds of layers. The original CNN would succumb to to the "vanishing gradient" problem, meaning that during the backpropegation the weights that minimizes the loss function would multiplied so many times that the gradient becomes smaller and smaller. In this case adding more layers will no longer lead to better performance and in some cases even degrade model performance.

What makes this architecture so efficient for high number of layers is "identity shortcut connections". ResNet stacks up these connections which is initially don't do anything. During training these layers are skipped, and the model uses the activation functions from previous layers. This compresses the model down to only a few layers at the beginning of training, this enables faster learning. When the model trains again all the layers are expanded again and the "residual" parts of the network explore more and more of the feature space of the source image.

ResNet outperforms shallower networks and are easy to implement in TensorFlow, Keras etc. And is therefor very popular in computer vision tasks. In the years after its authors published the model many new and prominent versions of the architectures have emerged.

1.3.2 EfficientNet

EfficientNets [10] focuses on improving the accuracy of the state of the art models even further, but also on increasing the efficiency of the model by tweaking the scaling. There are three different dimensions which can be scaled in a CNN; depth, width, and resolution. Depth is how many layers are in the model. Width is how wide the network is. One measure of width is how many channels are in the images - usually three, one channel each for red, green and blue, or one for gray-scale images. Resolution is the number of pixels for height and width of the source image. Scaling in computer vision tasks is usually fixed, and set so that a given model performs optimally on a given tasks.

Tan *et al.* proposed a novel technique which uses compound scaling to uniformly scale network width, depth and resolution.

1.3.3 Student teacher model

1.4 Model evaluation

1.4.1 Performance metrics

1.4.2 Weight initializing

1.4.3 Cross dataset validation

1.5 Datasets

The datasets used in our experiments are Hyper-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.

We will also go through some other publicly and restrictively available datasets, and explain why there is a need for a novel wireless video endoscopy capsule dataset. We will introduce Augere Medical, and their tagging tool implementation which we have used to label our WCE videos. In the later part of this section we will discuss some of the difficulties of the aforementioned datasets.

1.5.1 Available endoscopy datasets

There is a great number of publicly available endoscopy datasets online, and some that are restricted. To further improve detection rates in automated gastrointestinal analyze tools there is a demand for large amounts of data for different use cases, and since medical data often is scarce, or restricted, we introduce Kvasir-PillCam dataset, currently in development. This dataset is among the few publicly available VCE datasets, see Table 1.3 for an overview. Traditional colonoscopy have been around for longer and have been under more research. Therefore colonoscopy datasets are easier to find publicly, see Table 1.2 for a list of these datasets. This can benefit the ongoing automated VCE analysis as deep learning models can be tested and pretrained on them.

1.5.2 Kvasir-V2

The Kvasir dataset [16] 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 1.4. One of the most important aspects of the Kvasir dataset is that it makes it easy to reproduce and compare results in scientific computing.

[4] `https://www.endoatlas.net/ea/AtW01/106.aspx`
[5] `http://www.endoatlas.org/index.php`
[6] `http://www.gastrolab.net/index.htm`

Dataset Name	Data Source	Findings	Size	Status	Description
CVC-ClinicDB [11]	Colonoscopy	Polyps	612 still images from 29 different sequences with polyp mask	Available	From 29 different sequences with polyp mask (ground truth)
ASU-Mayo Clinic Colonoscopy Video DB [12]	Colonoscopy	Polyps	20 videos for training and 18 for testing	Copyrighted	10 videos with polyp detection, 10 videos without polyps, GT available
CVC colon DB [13]	Colonoscopy	Polyps	300 frames with ROI	By explicit permission	15 short colonoscopy sequences (different studies)
ETIS-Larib Polyp DB [14]	Colonoscopy	Polyps	196 images	By request	196 images with GT
GI Lesions in Regular Colonoscopy Data Set [15]	Colonoscopy	GI lesions	76 instances	Available	15 serrated adenomas, 21 hyperplastic lesions, 40 adenomas
The Atlas of Gastrointestinal Endoscopy[4]	Endoscopy	GI lesions	2259 images	Available	Esophagus, Stomach, Duodenum and Ampulla, Capsule Endoscopy, Inflammatory Bowel Disease, Colon and Ileum and some Miscellaneous
WEO Clinical Endoscopy Atlas[5]	Endoscopy	GI lesions	152 images	By explicit permission	One image per lesion
GASTROLAB[6]	Endoscopy	GI lesions	Several hundreds of images and several tenths of videos	Discontinued	Partially damaged and unavailable dataset
Kvasir-V2 [16]	Various	GI lesions & landmarks	8,000 images, 8 classes, 1,000 images per class	Available, public, free for research and educational purposes	See Section 1.5.2 for the description
Hyper-Kvasir [17]	Endoscopy	GI lesions and landmarks	10,662 labeled images, 373 videos and 99,417 unlabeled images	Available, public, free for research and educational purposes	See Section 1.5.3
Nerthus [18]	Colonoscopy	GI findings	5,525 frames extracted from the 21 videos, 4 classes, from 500 to 2,700 frames per class	Available, public, free for research and educational purposes	XXX
Medico [19]	Various	GI lesions, landmarks and findings	14,033 images, 16 classes, from 4 to 2,331 images per class	Available, public, free for research and educational purposes	XXX

Table 1.2: Existing colonoscopy image and video datasets.

1.5.3 Hyper-Kvasir

The Hyper-Kvasir dataset [17] 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/`.

All the data is fully anonymized and approved by Privacy Data Protection Authority,

Dataset Name	Data Source	Findings	Size	Status	Description
KID [20]	VCE	Angiectasia, bleeding, inflammations, polyps	2,500+ images + 47 videos	Discontinued	Open academic
GIANA'17 [21]	VCE	Angiectasia	600 images	Available, by request	Includes ground truth segmentation masks
CAD-CAP [22]	VCE	Normal, Vascular Lesions and Inflammatory Lesions	25,000 images	Discontinued	By request
Kvasir-PillCam	VCE	GI lesions, landmarks and findings	XXX images with ROI, XX classes and XXX unlabeled images/videos	Available, public, free for research and educational purposes	Ours, See Section ??.

Table 1.3: An overview of existing VCE datasets from the GI tract.

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

Table 1.4: Kvasir-V2 class names and corresponding class numbers.

and all experiments were performed in accordance with the relevant guidelines and regulations of the Regional Committee for Medical and Health Research Ethics - South East Norway, and the GDPR.

The Hyper-Kvasir dataset is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaption, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original authors and the source. This is important to highlight because it benefits other researchers who is in need of similar data can access the dataset easily.

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 [23], 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. In Table 1.5 you can see the 23 classes and how many images there is for each class.

Class name	Samples	Class name	Samples
barrets	41	normal-z-line	932
barretts-short-segment	53	polyps	1028
bbps-0-1	646	pylorus	999
bbps-2-3	1148	retroflex-rectum	391
cecum	1009	retroflex-stomach	764
dyed-lifted-polyps	1002	ulcerative-colitis-0-1	35
dyed-resection-margins	989	ulcerative-colitis-1-2	11
esophagitis-a	403	ulcerative-colitis-2-3	28
esophagitis-b-d	260	ulcerative-colitis-grade-1	201
hemorrhoids	6	ulcerative-colitis-grade-2	443
ileum	9	ulcerative-colitis-grade-3	133
impacted-stool	131		

Table 1.5: Hyper-Kvasir class names and corresponding amount of samples.

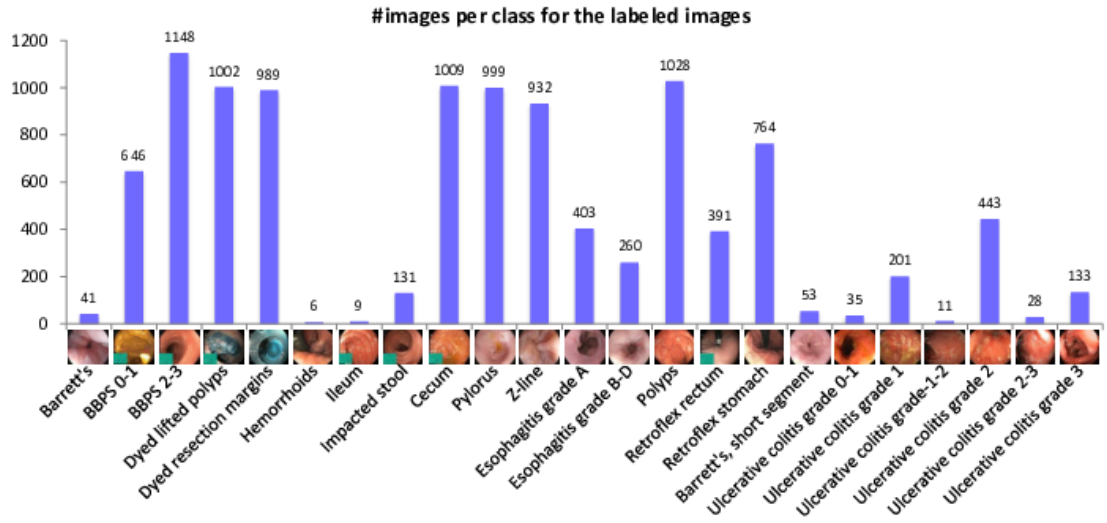


Figure 1.8: Number of samples for each of the 23 classes in Hyper-Kvasir dataset.

Unlabeled images

This part of the dataset contains 99.417 unlabeled images. When extracted they can be found in a separate subfolder. The images are accompanied with extracted global features and clusters assignments in Hyper-Kvasir Github repository.

Segmented images

Hyper-Kvasir includes images with corresponding segmentation masks and bounding boxes for 1,000 images from the polyp class. The segmentation masks depicts the polyp tissue for the corresponding image pixel. The Region of Interest (ROI) are represented by the white mask while the black does not contain polyp pixels. In Figure 1.9 we can see an example of the segmented Kvasir images.

Figure 1.9: Example of a segmented image from Hyper-Kvasir dataset.

Videos

In total there are 373 videos provided in the dataset, corresponding to 11.62 hours of videos and about 1 million video frames that can be convertet to images. The file format for the videos are Audio Video Interleave (AVI). The video portion of the dataset is 38.6GB in file size. In addition to the video folder there is a CSV file provided containing the videos IDs and findings. The finding contains the description of the finding in the video, of which there is 171 of. Finding description is meant to describe the video as a whole. Some of the findings are related to the categories found in the image portion and some are unique for the videos.

1.5.4 Augere Medical AS

Write about Augere Medical, the tagging tool and export script used.

1.5.5 Imbalanced data

Class imbalance typically refers to a problem with classification problems where the classes are not represented equally. In the medical imaging domain this is very common for two reasons: (1) there are a lot more healthy patients than there are sick ones, and (2) for a specific unhealthy patient, there are a lot more images of healthy tissue than there are images of tissue with lesions. As an example you could have a VCE video which has been carefully analyzed by a clinical expert and each frame is tagged with either being in class 1; healthy, or class 2; unhealthy. Of this dataset 10 000 frames have no findings, and 50 of them have confirmed findings. You could then use this data to train a model to have 99% accuracy which sounds great, but in reality the model only achieves these impressive results because it classifies all the data as class 1; healthy.

There are different methods to combat this class imbalance problem:

- Collect more data.
- Changing the performance metric.

- Resample the dataset.
- Introduce class penalties.

It is laboring, time consuming and expensive to collect more data for medical image classifications tasks because a clinical expert is required to validate the data.

1.6 TensorFlow Framework

1.6.1 tf.data

1.7 Related work

Zhu *et al.* have made a computer-aided lesion[7] 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 [24]. 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 [25]. Detecting 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 [26]. 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.

1.7.1 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

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

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" [27]. 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.

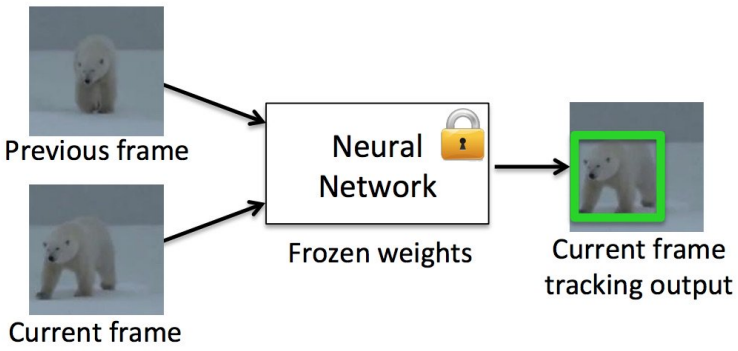


Figure 1.10: Illustration of how object in two frames is tracked with a bounding box[8].

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.

1.7.2 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. Ronneberger *et al.* propose a method in the paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" [28] 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" [29] 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

[8] `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 1.6: Segmentation results on the ISBI cell tracking challenge in 2015.

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.

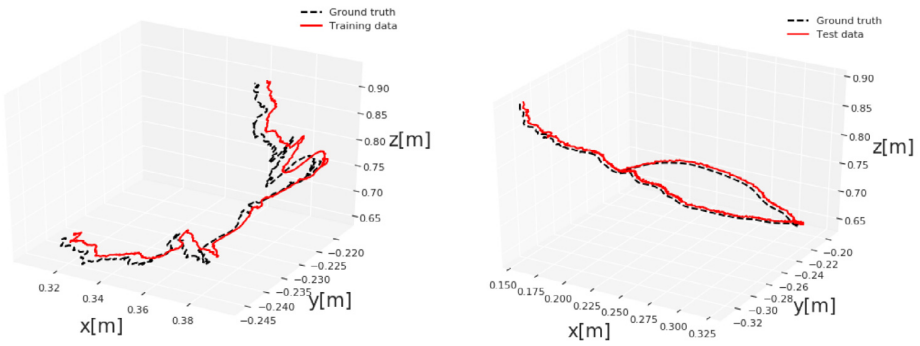
A common problem in training neural networks are too little labeled training data. This is also the case for us. We require a lot of medical data, and personell with the expertise to correctly label our data are of high demand and they usually have very little time for projects like these. This is why 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 [28]. By doing this Ronneberger *et al.* were able to achieve very good results (Table 1.6).

1.7.3 Mapping

As mentioned in section 1.1.4, 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[9] [30]. See Figure 1.11 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 [31]. This model outperforms state-of-the-art models like LSD SLAM and ORB SLAM.

[9]Deep recurrent convolutional neural networks



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

Figure 1.11: An example of Deep EndoVO accuracy [30].

1.8 Summary