Raster Vectorization Using Deep Learning

Ruben Schmidt Mällberg

Fall 2017

TBA4560 - Specialization Project

Department of Civil and Transport Engineering
Faculty of Engineering Science and Technology
Norwegian University of Science and Technology

Supervisor 1: The main supervisor

Supervisor 2: The co-supervisors (internal and external)

Preface

Here, you give a brief introduction to your work. What it is (e.g., a Master's thesis in RAMS at NTNU as part of the study program xxx and...), when it was carried out (e.g., during the autumn semester of 2021). If the project has been carried out for a company, you should mention this and also describe the cooperation with the company. You may also describe how the idea to the project was brought up.

You should also specify the assumed background of the readers of this report (who are you writing for).

Trondheim, 2012-12-16

(Your signature)

Ola Nordmann

Acknowledgment

I would like to thank the following persons for their great help during . . .

If the project has been carried out in cooperation with an external partner (e.g., a company), you should acknowledge the contribution and give thanks to the involved persons.

You should also acknowledge the contributions made by your supervisor(s).

O.N.

(Your initials)

Remark:

Given the opportunity here, the RAMS group would recognize Professor Emeritus Marvin Rausand for the work to prepare this template. Some minor modifications have been proposed by Professor Mary Ann Lundteigen, but these are minor compared to the contribution by Rausand.

Executive Summary

Here you give a summary of your your work and your results. This is like a management summary and should be written in a clear and easy language, without many difficult terms and without abbreviations. Everything you present here must be treated in more detail in the main report. You should not give any references to the report in the summary – just explain what you have done and what you have found out. The Summary and Conclusions should be no more than two pages.

You may assume that you have got three minutes to present to the Rector of NTNU what you have done and what you have found out as part of your thesis. (He is an intelligent person, but does not know much about your field of expertise.)

Contents

Chapter 1

Introduction

Raster to vector or digitizing is a central part of what GIS specialists do. Digitizing is the task of extracting vector layers from raster maps so that they can be used for further analysis, and is often a time consuming manual process. With the vast amount of raster maps available online, we are losing valuable information because we are unable to process them automatically.

Even though there are multiple software products in the market today concerning the problem of converting a raster image to a vector image such as Scan2Cad [33] and Powertrace [29]. These products only focus on making vectorized boundaries of homogenous color areas, such as those in a logo and do not focus on the problem of digitizing raster maps with spatial data. R2V [42] is the only product the author found that focus on vectorization of raster maps in GIS application.

Problems occur when the raster images not only consists of isolated objects that are easy to distinguish but contains a spatial structure, overlapping geometries, and background layers. The multilayered nature of raster maps in addition to varying image quality makes automatic vectorization a really hard problem.

Deep convolutional neural networks (Deep CNNs), are top performers of semantic image labelling (CITE) and can, therefore, be a possible solution to the digitalization problem. Beeing an instance of supervised learning, deep convolutional networks require proper labelled training data. This is often generated manually for each case. If one could train the networks with some of the digitized raster maps, and then use them to digitize the rest of the maps for us, this would be a great time saver.

In this paper we will look at the state of the art in feature extraction with deep convolutional networks, look at the data needed to train the networks sufficiently, implement a first version of a digitizing network and look at the performance of this.

Chapter 2

Motivation

Digitizing and vectorization is a time consuming and expensive process [41].

In 2009, the Norwegian government made it statutory for all municipalities in Norway to have a digital zoning registers [16]. This has many benefits to society, both for the municipalities, the government, and the private sector. Benefits such as faster insights and sped up proceedings for building projects.

In June 1st, 2017, 354 out of 426 municipalities are registered to have digital zoning registers. The law does however not force the municipalities to vectorize the zoning regulations in detail, only to have them scanned in a digital format such as PDF with the bounding limits of the plan vectorized. The detailed vectorization is then another step that has to be done in order to fully utilize the data in more advanced use cases, such as in GIS. See Figure 2.1 for an example of a zoning regulation in PDF format.

For the municipalities, the digitalization and vectorization is a process that is often done by external contractors. In a project in Telemark and Vestfold, the cost was estimated to be around 3000NOK for each plan, where the whole project consisted of 100 plans [2]. The cost savings in automating the vectorization of digitized zoning plans are thus economically significant.

Since the zoning regulations has to follow strict quality standards, the acccuracy of the output from the deep CNN also has to follow these standards. With many zoning regulations already digitized and vectorized, we potentially have a lot of validated, accurate data that can bee used to generate training data for the deep CNN. This gives us a great oportunity to investigate the performance of deep CNNs for vectorization of scanned maps.

There are a lot of approaches and network architectures that need to be examined and reviewed for this specific problem. This research will be the basis of the author's master thesis, where a practical implementation of the theory found in this research will be done.

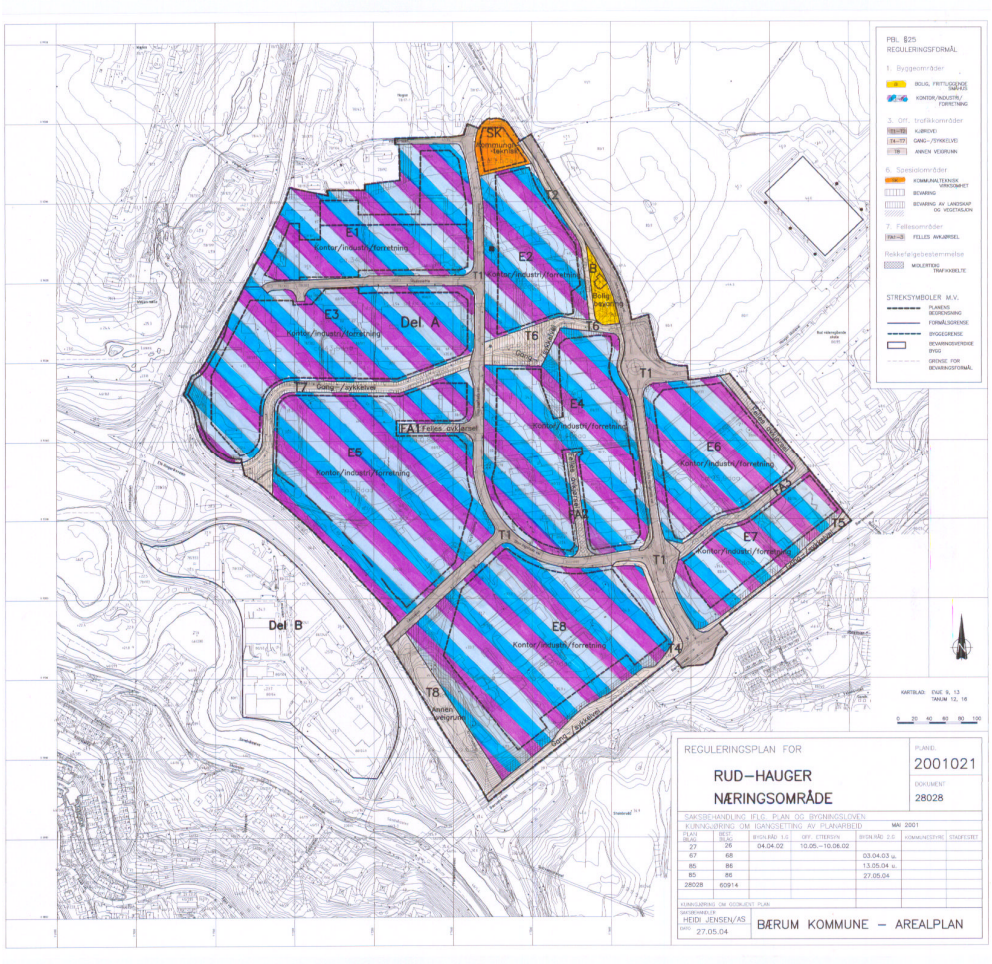

Figure 2.1: Scanned zoning regulation in PDF format from Bærum municipality.

Chapter 3

Vectorization

Vectorization, raster-to-vector conversion or image tracing is the conversion of raster graphics to vector graphics. To understand the reason for converting from raster to vector we need to look at some of the properties of both storage techniques.

Raster data is structured as an array of grid cells, also referred to as pixels. Each cell in a raster can be addressed by its position in the array, by row and column number. Since each pixel has it's own value, a raster can represent a range of spatial objects. A point can be represented by a single pixel, an arc represented by a sequence of pixels and an area as a collection of continuous pixels. Vector data is structured as a finite straight line segment defined by its endpoints. The location, or coordinates, of the endpoints, are given with respect to a coordinatization of the plane. The vector representation is not discretized in a grid space the same way as a raster but does follow an implicit grid structure as a result of the nature of computer arithmetic. Like the raster, the vector structure can represent multiple spatial structures. A point is given by its coordinates, an arc represented as a sequence of line segments, each consisting of start and end coordinates and an area represented by its boundary consisting of a collection of vectors.

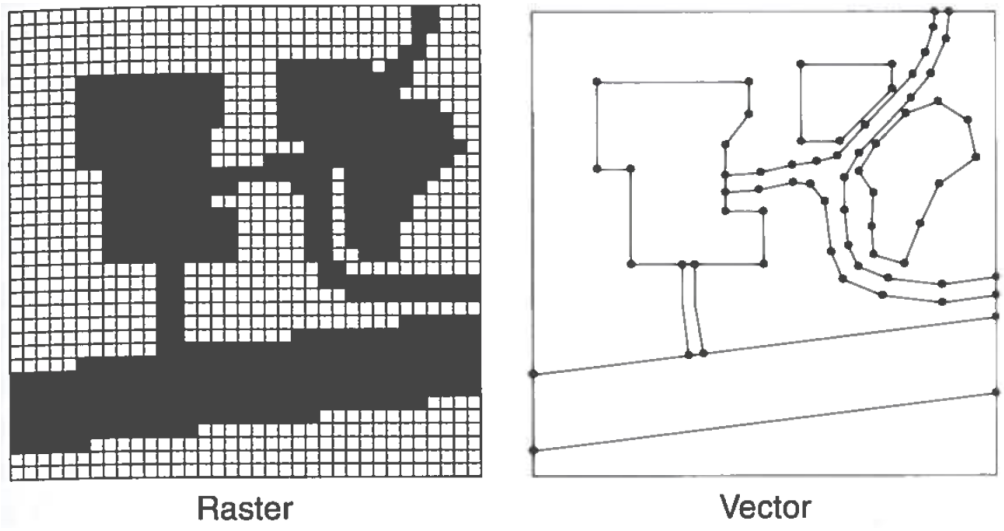

Figure 3.1: Raster and vector data
Source: [41]

There are multiple reasons for storing data in vector format: The geographic accuracy is higher since it is not dependent on grid size, it allows for efficient encoding of topology, an important aspect when doing analysis that utilizes topologic relations, such as proximity and network analysis, vector data allows for storage of attributes in the data, giving us another dimension of information and the storage size is smaller.

There are multiple different techniques for vectorization. We will now look at some of the well known ones.

3.1 Hough Transform

The Hough Transform can isolate features of particular shape within images. The user must describe the desired feature in a parametric form, therefore the most common application of Hough Transform is to detect regular curves such as lines and circles. The technique is known to be tolerant of image noise and gaps in the feature boundary.

Hough Transform works by transforming each of the pixels into a straight line in a parameter space. The parameter space is described by the parametric form of features we are looking for. In the simplest case of a Hugh Transform, we want to detect straight lines. For this the normal form $r = xcos(\theta) + ysin(\phi)$ is used to represent the lines. When iterating the pixels in the image the algorithm looks if there is enough evidence that there is a straight line at that pixel, if there is, it calculates the parameters θ and ϕ at that point and adds it to an array. The array accumulates

all the points that belongs to each specific θ and ϕ, thus the more points that are on the same parameters the more likely they belong to a line. The algorithm visits each pixel one time and is therefore linear with the number of pixels in the image.

3.2 Thinning

Thinning is a morphologic operation that is used to remove foreground pixels from an image, resulting in a simplified image with the same topologial relations as the input image. It works by successively removing pixels from the boundary until there is not possible to remove more. The skeleton that is left contains the centerlines of the objects. This technique is mostly used on binary images. The iterative nature of the algorithm makes it computationally expensive.

3.3 Contour

This method aim at lowering the computational cost compared to thinning. The main idea is to find edges before calculating the middle point that is between two opposite parallel edges. A chain of middle points represent the medial axis. It has to use a edge detection algorithm before extracting the middle points between the edges. A problem with the contour based methods is how they deal with junctions. It is likely to miss junctions that are crossing at a small angle and it has difficulties with cross intersections where four lines meet [40]. It is thus not suited for vectorization of curves and crossing lines.

3.4 Run-graph

Run-graph is a semi-vector representation of the raster image [23]. It examines the raster in either row or column direction in what is defined as horizontal and vertical *runs*. A horizontal run is a maximal horizontal sequence of black pixels. A vertical run is a maximal vertical sequence of black pixels. After running, the run-graph is composed of edge areas that represent line segments and node areas and touching points that correspond to endpoints and junctions. The line extraction from the graph attempts to minimize the area of the node and maximize the length of the connected edges.

TODO LITT MER HER

3.5 Mesh pattern

Mesh pattern was introduced by Lin et al.[20]. The idea is to divide the image in to a mesh and check the distribution of black pixels at the border of each mesh unit. Using the patterns along the border, a control map is created. Lines are then extracted from the control maps by comparing them to characteristic patterns in a pattern database. Unknown patterns are labeled with question marks. These areas are analyzed pixel by pixel to determine where the line goes. In Figure 3.2 we can see the process of generating a mesh, extracting the control maps and filling the lines based on the analysis of control maps.

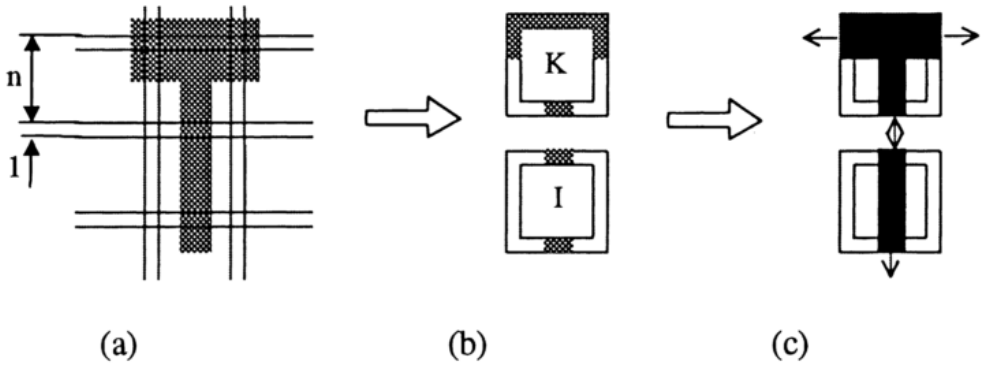

Figure 3.2: Mesh pattern line extraction. (a) image and mesh, (b) mesh pattern labels and control map, (c) lines extracted by analyzing the control map.
Source: Wenyin and Dori[40]

3.6 Sparse-pixel

The first of these algorithms was the Orthagonal Zig-Zag (OZZ) introduced by Dori. The basic idea of the algorithm is to track a single pixel wide "beam of light" which turns orthagonally each time it hits an edge, the same way light travels in a fibre-optic cable. In the algorithm the bream changes direction if it encounters white pixels or the length of the beam exceeds a predefined threshold. If the threshold is hit, two new beams are emitted orthagonally from the point. This can be seen in Figure 3.3. The midtpoint of the runs are also recorded, used to reconstruct the line, corner correction and merger of crossing lines. Wenyin and Dori improved the OZZ method with Sparse Pixel Vectorization (SPV). There were three main improvements: The procedure begins with a reliable medial axis point found by a separate procedure for each black area, a general following procedure is used for all cases of OZZ, i.e. Vertical, horizontal and diagonal following and a junction recovery procedure is applied whenever a junction is encountered during following.

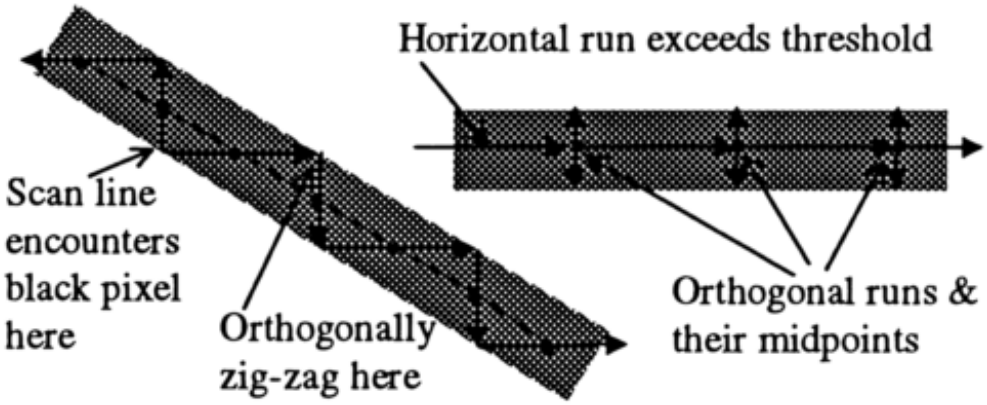

Figure 3.3: Orthagonal Zig-Zag
Source: Wenyin and Dori[40]

Chapter 4

Image segmentation with Deep Learning

In this chapter, we will look at the problem of image segmentation and the state of the art regarding segmentation and object detection using Deep Learning. We will look at both semantic and instance segmentation.

Semantic segmentation of images is one of the key problems in the field of computer vision. It is about making dense predictions inferring labels at the pixel level, assigning a class to each pixel with its enclosing object [9]. Taking it a step further, we get to instance segmentation, where we want to associate the classes with a physical instance of an object.

Both semantic and instance segmentation can be seen as giving us an understanding of an image at a higher level. This fine-grained control of an image greatly helps with scene understanding which is becoming more and more relevant with the increasing number of applications, such as self-driving cars and augmented reality.

We can see an example in Figure 4.1 where we see the difference between the two approaches. In the middle photo, all the chairs have the same classification, as chairs. In the right photo, we see that the chairs now are classified as chairs, but with a different class for each of them. We can see that instance segmentation is the combination of both object detection and semantic segmentation.

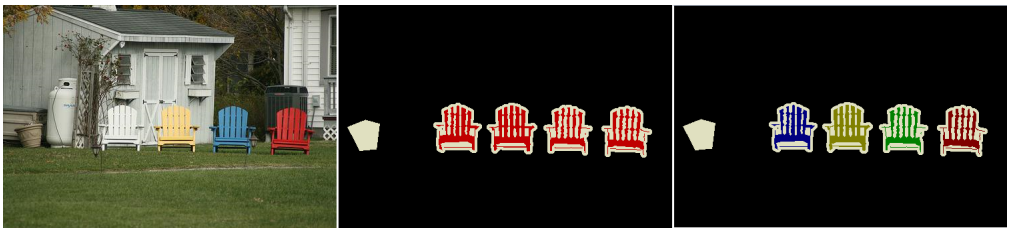

Figure 4.1: Left: input image. Middle: Semantic segmentation. Right: Instance segmentation. [26]

4.1 Neural networks

To get a deeper understanding of how neural networks perform segmentation of images, we need to take a look at the fundations behind them and how they operate.

Neural networks are a computational model that shares som properties with the animal brain in which simple units called neurons are working in parallel with no centralized control unit [28]. The primary means to long-term information storage is in the weights between the units and updating them is the primary way the network learns new information.

A network is defined by the number of neurons, number of layers and the connections between the layers. One of the easiest architectures to understand is the feed-forward multilayer architecture viewed in Figure 4.2. It is a neural network with an input layer, one or more hidden layers and an output layer. The input layers feeds input, in the form of vectors, to the rest of the network. The number of neurons at the input layer often reflects the size of the input vector.

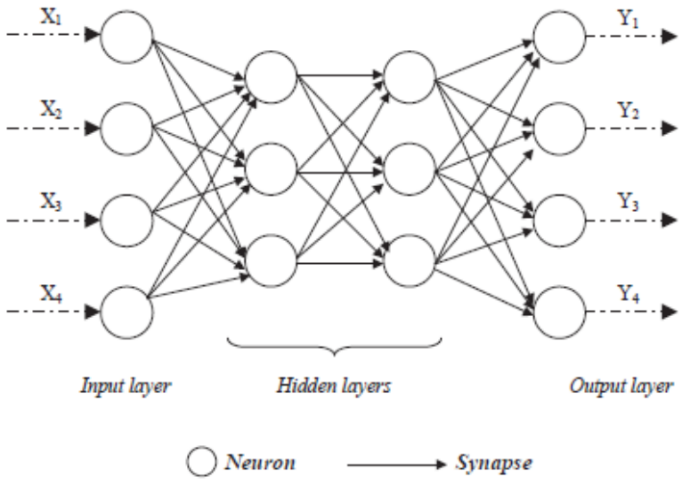

Figure 4.2: Structure of a multilayer feed forward network. [45]

4.1.1 Artificial neurons

Each layer consist of one or more artificial neurons, also called nodes. An artificial neuron is a mathematical representation of a biological neuron and consist of inputs with weights and bias, a transfer function and an activation function. The weights are what scales, either amplifying or decreasing, the input to the node. The bias is a constant scalar value per layer that is added to ensure that at least some of the nodes in the layer are activated, that is, forwarding a non-zero value to the next layer. The transfer function takes the weighted sum of the input variables and transfers it to the activation function. The activation functions are scalar-to-scalar functions that defines the output of the node based in the inputs, weights and bias. A model of an artificial neuron compared to a real one, can be seen in Figure 4.3.

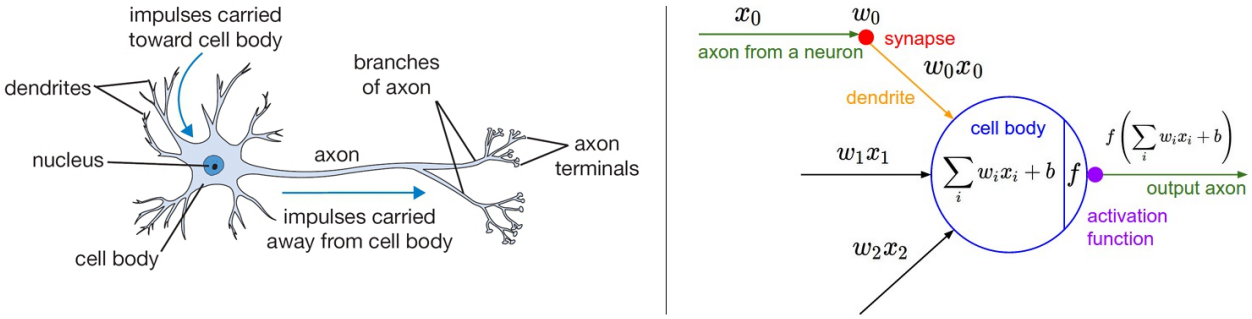

Figure 4.3: Structure of a real neuron compared with an artificial one. [15]

4.1.2 Activation functions

The use of activation functions in the hidden layers add the ability for the network to learn non-linear functions. We will now take a look at some of the usefull activations functions that are used today.

Sigmoid

The sigmoid activation function has a characteristic "S"-shaped curve and can take variables of near infinite range and convert them to values between 0 and 1. It is good at reducing outliers and extreme values in the dataset. Expressed mathematical as:

$$a = \sigma(x) = \frac{1}{1+\exp(-x)} \tag{4.1}$$

Tanh

A trigonometric hyperbolic function. Tanh can normalize input to the range of -1 to 1 and can therefore deal with negative numbers better than the Sigmoid. Expressed mathematically as:

$$a = \sigma(x) = \tanh(x) \tag{4.2}$$

Rectified Linear

Rectified Linear only activates a node if it is above some threshold. When the input raises above the threshold it has a linear relation to Equation 4.3. Nodes that use the rectifier are called Rectified Linear Unit or ReLU.

$$a = \sigma(x) = \max(0, x) \tag{4.3}$$

ReLUs are the state-of-the-art because of their proven usefulness in many situations and their ability to train better in practice than sigmoids. ReLU does not have the so-called problem of vanishing gradients either. Vanishing gradients is a problem that occurs when using gradient-based methods (explained in subsection 4.1.3) for learning, where large changes in the value of parameters from the early layers, does not have a big effect on the output, making the network lose its ability to learn. The reason for this happening is that some activation functions, such as sigmoids or tanh, forces the input space into small regions.

While removing the problem of vanishing gradients, ReLU introduces another one and that is the problem of "dying ReLU" [15]. This is a problem that occurs when a large gradient passes through the neuron causing the weight update to be so large that it causes the neuron to never activate again, that means that the gradient passing through the neuron will be forever zero.

4.1.3 Training

There are different forms of learning such as supervised, where we show the network what the correct answer is. Unsupervised, where the network itself decides how to label the data and reinforcement learning, where the network does not get to know the answer but learns by reward or punishment. We will only focus on supervised learning in this paper as this is the type of learning we use when doing image segmentation.

In supervised learning, the network learns by training on a set of inputs and desired outputs. As inputs are passed through the network and outputs are generated, it learns by adjusting weights and biases causing some neurons to become smaller and some to become larger. The larger a neuron's weight is, the more it affects the network and vice versa.

By adjusting the weights and biases, the network reduces the errors, also called loss. The loss is defined by some loss function that quantifies the correctness of the output from the network in regards to the ground truth. By using a loss function we reclass the learning problem as an optimization problem, where we try to minimize the loss.

The most common algorithm for the weight adjustment in neural networks is called *backpropagation*.

Backpropagation Learning

When the output from a neural network produces a large loss, we need to update the weights accordingly. A problem with multilayer neural networks, however, is that there are many weights connecting an input with an output, so it becomes difficult knowing what weights that affect the output. We need a clever way of finding what specific weight that contributes to the output. This is the problem backpropagation tries to solve. A high level understanding of backpropagation is that we use the chain rule to iteratively calculate the gradients for each layer. The steps of the algorithm is as follows:

1. Initialize network with random weights
2. Loop trough the training examples
3. Compute the network output for the current training example
4. Compute the loss with the loss function.
5. Compute the weight update for the output layer with the weight update rule:

$$W_{j,i} \leftarrow W_{j,i} + \alpha \times \alpha_j \times \Delta_i$$

   Where

$$\Delta_i = Err_i \times g'(input\_sum_i)$$

   and g' is the derivative of the activation function

6. Loop trough all the layers in the network all the way to the input layer and:

7. Compute the error at each layer with the propagation rule:

$$\Delta_j \leftarrow g'(input\_sum_i) \sum_i W_{j,i} \Delta_i$$

8. Update the weights leading in to the hidden layer with the update rule:

$$W_{k,j} \leftarrow W_{k,j} + \alpha \times \alpha_k \times \Delta_j$$

The term α is the learning rate, and belongs to the family of what we call *Hyperparameters* in machine larning.

Hyperparameters

The hyperparameters are what we tune to make the network train faster and better. The selection of these parameters are done to ensure that our network does not *overfit* or *underfit* the data. We say that our model is overfitted if it fits our training data too well but does not generalize enough over the entire dataset. We say our model is underfitted if it generalizes too much and is not able to fit the training set. The terms are illustrated in Figure 4.4. In the left image we see that the model does a bad job at approximating the function, it is underfitted. In the middle image we see that the model approximates the function well. In the right image we see that the model fits the training samples very good, but does a bad job at approximating the function, the model is overfitted.

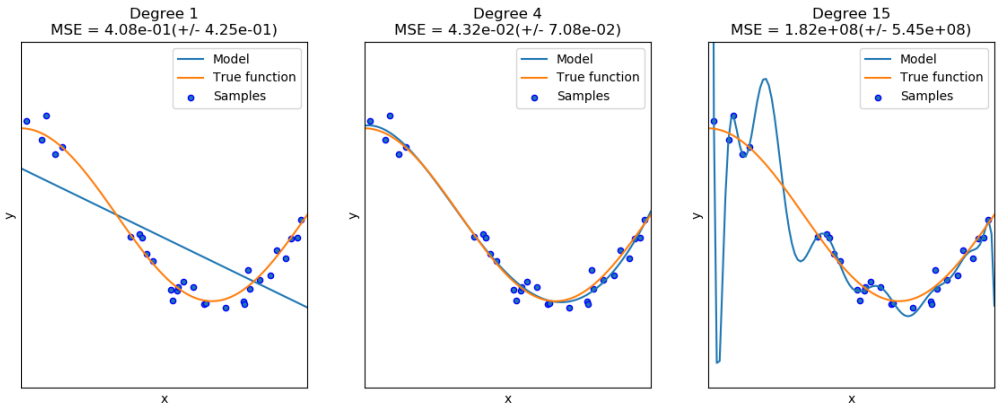

Figure 4.4: Left: Underfitted model. Middle: Appropriate fit. Rigth: Overfitted model.

The learning rate affects the amount we adjust the parameters during training. A large learning rate will make our parameters take large steps, thus saving time, but can cause us to overshoot the minimum of our loss function causing us to never find a minimum. A small learning rate causes us to take smaller steps and should help us reach the minimum, but can take a very long time to do so.

Another important hyperparameter is *regularization*. Overfitting often occurs when some weights have become very large and regularization is about reducing the effects of the large weights in the network. The perhaps most common form of regularization is L2 and is often implemented as the term $\frac{1}{2}\lambda w^2$ that we add to the weights [15]. Another regularization, introduced in [35] is *dropout*. Dropout works by randomly dropping some of the neurons during training causing it to train on a "thinned" version of the net. Dropout has shown major improvements over other regularization tehniques [35].

4.2 Convolutional neural networks

Fully connected multilayer neural networks take inputs as a one-dimensional vector. When using an image as input, these vectors become very large. The reason for this is that we represent each pixel in the image as one value in the vector. If we are working with color images represented with 3 channels of RGB information, each of these also needs to be mapped. For a single 200x200 image this means $200 \times 200 \times 3 = 120000$ connections in only the first layer. This illustrates how bad the fully connected neural networks scale.

Convolutional neural networks, or CNNs, tackle the scaling problem by assuming inputs as images and model them as three-dimensional objects with image width, image height and color channels as the dimensions. At a high level, the architecture consist of an input layer, convolutional layers, pooling layers and fully-connected layers [28].

4.2.1 Convolutional layers

The convolutional layers, or CONV layers, are the core building blocks of a CNN. The layers consist of a set of learnable filters also called kernels. Each of the filters are small in regards to width and height but are always the same size as the input in regards to depth. The filter is applied to the input by sliding, or *convolving*, the filter accross the width and height of the input. At each position, the dot product between the filter and the input is calculated. The output is a two dimensional map that is called *activation map*. This process is illustrated in Figure 4.5. For

each of the filters in the CONV layer we get such a map and stack them in the depth direction, this in turn represents the output of the layer.

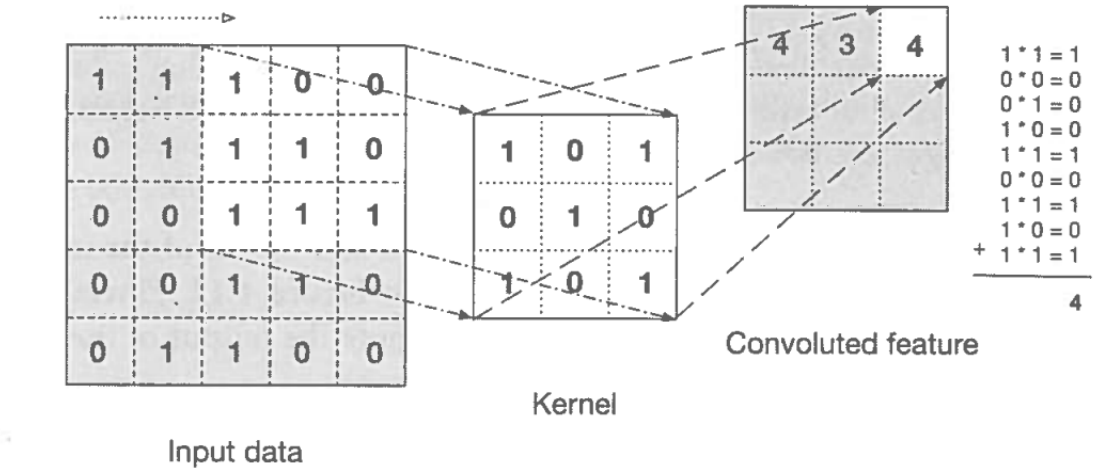

Figure 4.5: The convolution step.
Source: Patterson and Gibson[28]

The network will learn filters that causes the node to activate when certain visual features are seen, for instance an edge. Deeper into the network we will se filters that become more global in term of the input and recognizes nonlinear combinations of features. An example of what the filters look like in a deep CNN can bee seen in Figure 4.6 where we see filters learned in the first convolutional layer in an eight layer network [18].

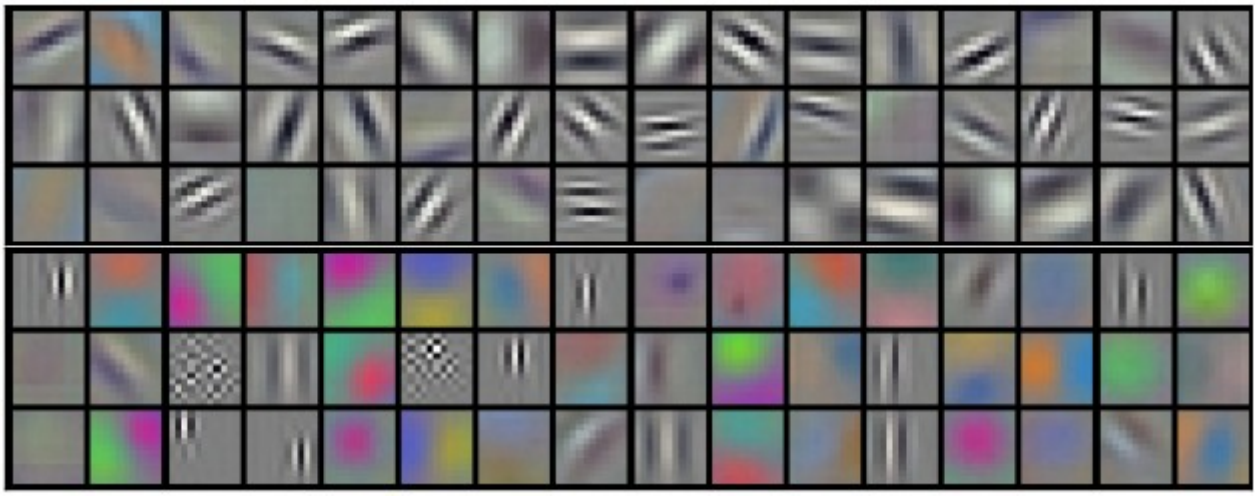

Figure 4.6: 96 learned filters in the first convolutional layer.
Source: Krizhevsky et al.[18]

If we imagine that we freeze the filter as it is convoluted accross the input, a single step and its calculation, can be viewed as the output from a neuron. The activation map then represents a sheet of neurons with each of the neurons looking at a small part of the input, not knowing anything about the rest of the image. This feature is called *local connectivity* and is an important part of how the network keeps the number of parameters smaller then a regular neural network. All the neurons in the sheet also share parameters, since it is the same filter that did the calculation. This is the concept of *shared paramters* and is the other important part of how CNNs keep the number of paramters low.

4.2.2 Pooling layers

Another way to reduce the number of parameters in the network, is the use of pooling layers. Pooling layers essentially reduces the input size by downsampling the input with different pooling functions. The most common operation is *max pooling* with a 2x2 filter with a stride of 2 that reduces the size by two in the height and width dimension [15].

Chapter 5

Previous work - Image segmentation

In this chapter, we will look at some of the previous work done in the field of image segmentation with neural networks.

We will see that all of the networks presented are trained on real-world imagery and not images of maps. The reason for this being that the research is more mature in the field of real world/natural image segmentation but the same principles apply to segmentation of map images. We will also look at some of the approaches towards maps and geographical data later on in the paper.

During the last ten years, we have seen many important advances when it comes to the architecture of deep networks for image segmentation. AlexNet [18], VGGNet [34], GoogLeNet [37], ResNet [10], ReNet [38] and the very recent CapsNet [32] are all examples of such advances. In the next section, we will look at the main points from each of these networks.

5.1 Important architectures

5.1.1 AlexNet

Achieving first place in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [31] in 2012 with a top-5 test accuracy of 84.6% by a margin of 10% to the next competitor, AlexNet pioneered deep CNNs in image classification. The network consisted of a total of eight-layer, five convolutional layers with max-pooling and three fully-connected layers. All the layers used ReLU as activation. To reduce overfitting they used dropout. Figure 5.1 shows the architecture of the network.

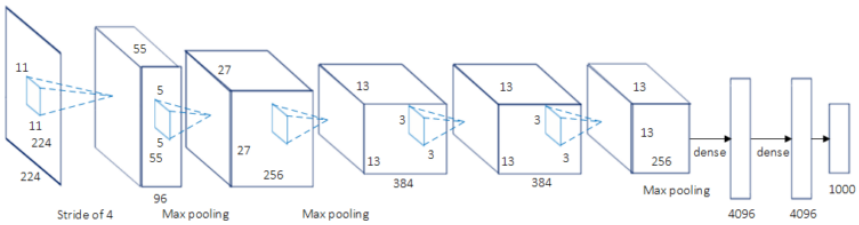

Figure 5.1: AlexNet architecture
Source: Krizhevsky et al.[18]

5.1.2 VGG

The Visual Geometry Group (VGG) model was introduced by the Visual Geometry Group at oxford university. In their paper, they propose multiple different architectural configurations with weight layers ranging from 13 - 16 layers deep. The most interesting is the model with 16 weight layers. It was submitted to ILSVRC 2013 and managed to get a top-5 test accuracy of 92.7%. In Figure 5.2 we can see that the architecture makes us of more layers with small receptive fields rather than a few layers with large receptive fields.

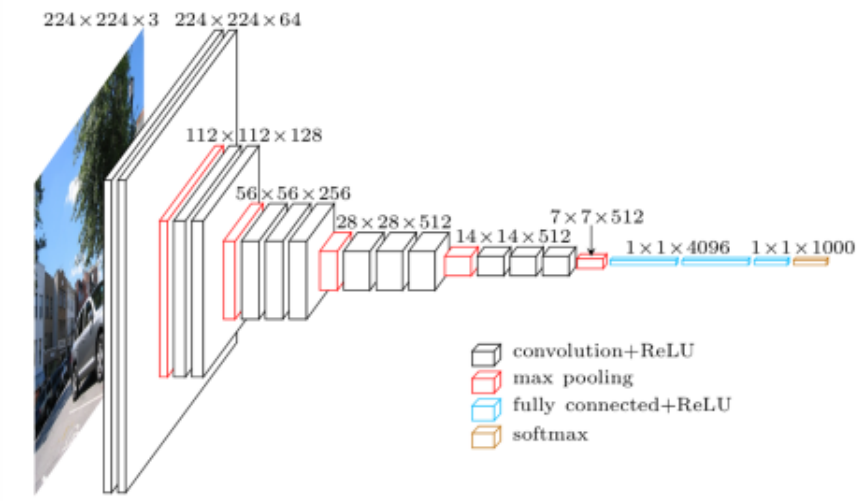

Figure 5.2: VGG 16 architecture
Source: https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/

5.1.3 GoogLeNet

This network won the 2014 ILSVRC challenge with a top-5 test accuracy of 93.3%. The network introduced a new architectural concept called the *inception* model (see Figure 5.3). The model is essentially a new mini-network with a pooling operation, large convolution layers, and smaller convolution layers. They proposed the use of small 1x1 convolution layers to reduce the complexity before the large convolution layers to keep the parameters and computational cost under control. This showed an increase in speed ranging from 3-10x faster than similar networks without the inception module.

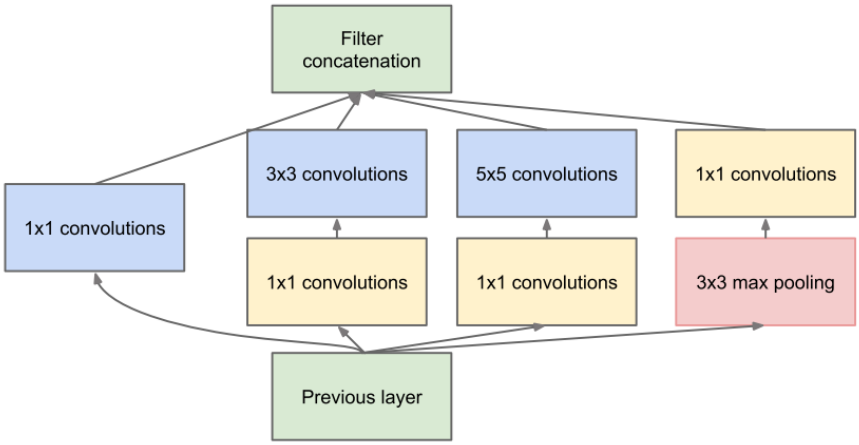

Figure 5.3: Inception module
Source: Szegedy et al.[37]

5.1.4 ResNet

ResNet won the 2015 ILSVRC, with a top-5 test accuracy of 96.4%. The network is known for its depth of 152 layers and a new kind of building block called residual block. The residual block contains two paths between the input and the output where one of the paths serve as a shortcut connection to the output (see Figure 5.4) essentially copying the input to the output layer. A big problem with very deep networks is that they are hard to optimize. When the depth of the network increases, the accuracy gets saturated. This is called *degradation* and is the problem that the residual blocks are addressing by forcing the network to learn on top of already available input.

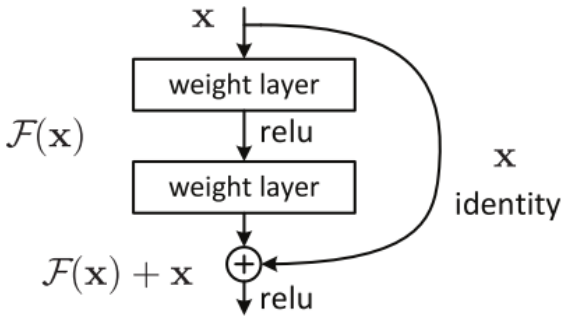

Figure 5.4: Residual block.
Source: He et al.[10]

5.1.5 CapsNet

Released November 2017, this is a very recent advancement in neural networks. It introduces a new type of neural network based on *capsules*.

BLABLABLA MORE ABOUT THIS

It addresses the issue that CNNs are not good at generalizing new viewpoints. They are good at generalizing to translation, but other affine transformations have shown to be difficult to learn.

It has not (yet) been tested in ILSVRC but has been run on the Modified Institute of Standards and Technology database (MNIST) that is a database of handwritten digits. The database has 60000 training images and 10000 test images.

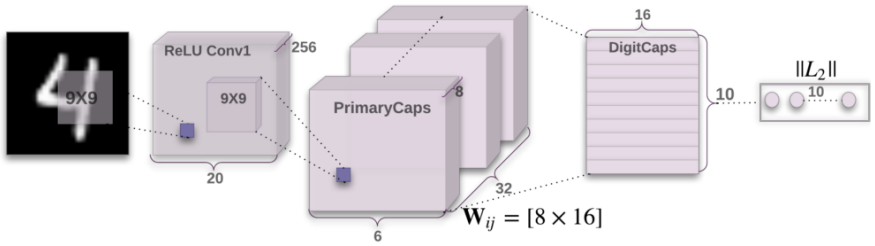

Figure 5.5: CapsNet with 3 layers.
Source: Sabour et al.[32]

5.2 Image segmentation

Many of the previous network architectures described in section 5.1 are predicting labels of what the images contain and not where and what part of the image the labels are to be found in. Image

segmentation is about assigning a class to each pixel with its enclosing object, also called dense predictions, so we need output from the networks that are spatial maps instead of classification scores. In this section, we will review important networks that are specialized in image segmentation. A benchmark often used for image segmentation is the PASCAL Visual Object Classes Challenge (VOC) [8], that provides standardised image data sets for object class recognition. We will therefore evaluate the networks with the scores in VOC when applicable.

5.2.1 FCN

Fully Convolutional Network (FCN) by Long et al. [21] can be seen as a common forerunner for semantic segmentation with convolutional networks [9]. FCN adopted the contemporary deep classification nets AlexNet, VGG and GoogLeNet architectures we saw in section 5.1 to make dense predictions at the pixel level. It is important to note that they not only reused the architecture but used the pre-trained classification models as a starting point. The network replaced the fully-connected layers with convolutional ones, noting that the fully-connected layers could be seen as convolutional ones with kernels (filters) that cover the entire input region. This allowed segmentation maps to be generated from images of any size. Because of all the pooling operations in CNNs, a technique called *deconvolution* [46] was used to upsample the coarse output to dense pixels. Deconvolutional layers can learn interpolation functions the same way the network learns weights. Skip connections similar to the ones we saw in ResNet, are also included to give the deeper layers higher resolution feature maps. It reached a score of 62.2 in the VOC2012 challenge.

5.2.2 SegNet

SegNet by Badrinarayanan et al. [1] uses an Encoder-Decoder architecture. One can also view FCN as this type of architecture, where the downsampling is the encoding part and the deconvolution is the decoder part. The difference between these networks is in the decoding part of the architecture. In SegNet more shortcut connections are added, however instead of copying the input to the output of one layer, each upsampling layer corresponds to a max-pooling layer in the encoding part and the indices from the max-pooling layer is copied to the upsampling layer. This can be seen in Figure 5.6 where we see the blue lines as shorcut connections to the upsampling layers. The network was not benchmarked on VOC2012 in the paper but the leaderboard [27] shows that it reached a score of 59.9.

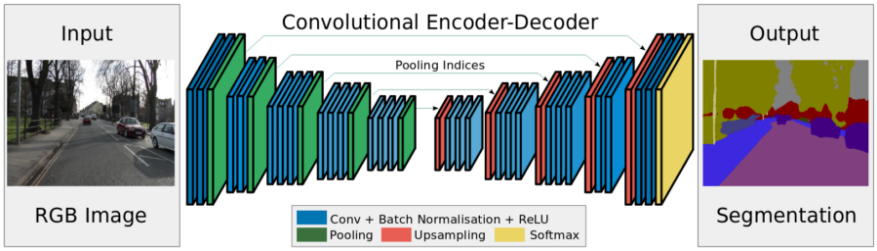

Figure 5.6: SegNet architecture.
Source: Badrinarayanan et al.[1]

5.2.3 Dilated Convolutions

A problem with pooling layers is that they remove more context and resolution from the image the deeper we get into the network. For segmentation we need contextual reasoning and full-resolution output [44]. Yu and Koltun [44] try to solve this problem with dilated convolution layers. Dilated convolution layers can increase the receptive field of a convolutional filter exponentially while the number of parameters grow linearly. This is illustrated in Figure 5.7 where we see that the receptive field, indicated in green, grows expoentially compared to the parameters, indicated as red dots. The network showed state-of-the-art perfomance with a simpler architecture, based on VGG-16 [34], than the competition, scoring 71.3 in VOC2012 with their basic network.

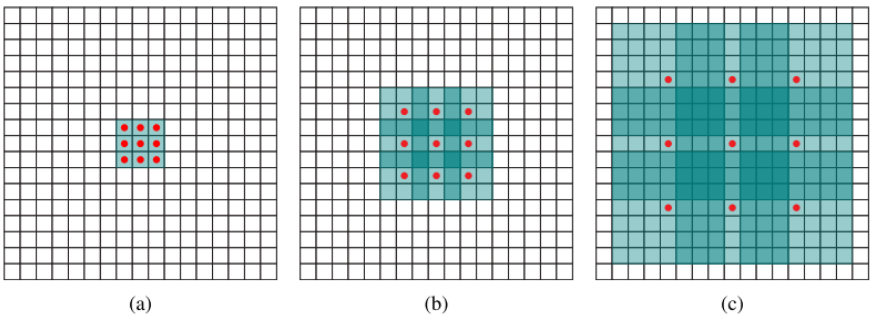

Figure 5.7: Diluted filters. (a) 1-dialted, (b) 2-dialated, (c) 3-dialated.
Source: Yu and Koltun[44]

5.2.4 DeepLab (v1 & v2)

DeepLab v1 (2014) [3] and DeepLab v2 (2016) [4] both used dilated convolutions though they refer to them as *atrous convolutions*. They use fully-connected Conditional Random Fields (CRF) to capture fine-grained details in images as proposed by Krähenbühl and Koltun [17]. CRF can

be used to combine the class scores from deep CNNs with the low-level information captured by the local interactions of pixels and edges [3] and is used in DeepLab v1 as a post-processing method. As we can see in Figure 5.8 the effect of using CRF is significant in regards to detecting details in the image. DeepLab v1 used VGG-16 as a starting point for their architecture and got a score of 71.6 at VOC2012 beating the runner-up by a margin of 7.2%.

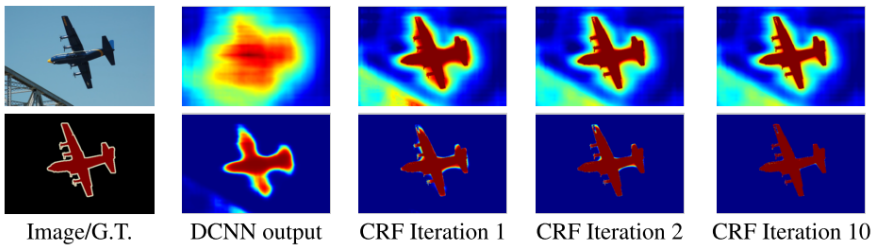

Figure 5.8: Score map (input before softmax) and belief map (output after softmax).
Source: Chen et al.[3]

Deep CNNs can represent different object scales by training on datasets that contain objects of varying size. However explicitly accounting for scale can improve the networks ability to handle large and small objects [25]. DeepLab v2 introduces a new technique to handle multivariate scale, called *atrous spatial pyramid pooling* (ASPP). It uses multiple parallel atrous convolutional layers with different sampling rate/direction to improve the networks ability to deal with objects of different scale. DeepLab v2s best architecture used a pre-trained ResNet-101 (ResNet seen in subsection 5.1.4, that has 101 layers) with atrous convolutions, ASSP, and CRF that got a score of 79.7 at VOC2012.

5.2.5 DeepLab v3

In DeepLab v3 [5], Chen et al. introduce an improved model of ASSP that involves concatenation of image-level features, a 1x1 convolution and three 3x3 atrous convolutions with different rates as seen in Figure 5.9.

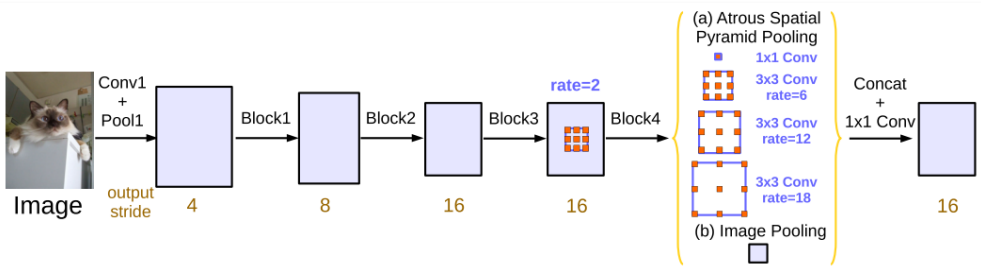

Figure 5.9: Improved version of ASSP, augmented with image-level features.
Source: Chen et al.[5]

Cascading modules was also proposed. A technique where the last ResNet block is duplicated multiple times and modified with atrous convolution Each block consists of three atrous convolution layers as seen in Figure 5.10.

Both methods were tested, but the new ASSP performed better and was selected for use in the final model. Using a ResNet-101 pre-trained on ImageNet dataset, DeepLabv3 reached a perfomance of 85.7 on VOC2012. Using a ResNet-101 pre-trained on both ImageNet and JFT-300M dataset (internal dataset at Google with over 300M images [12][36]) , DeepLabv3-JFT got a score of 86.9 on VOC2012.

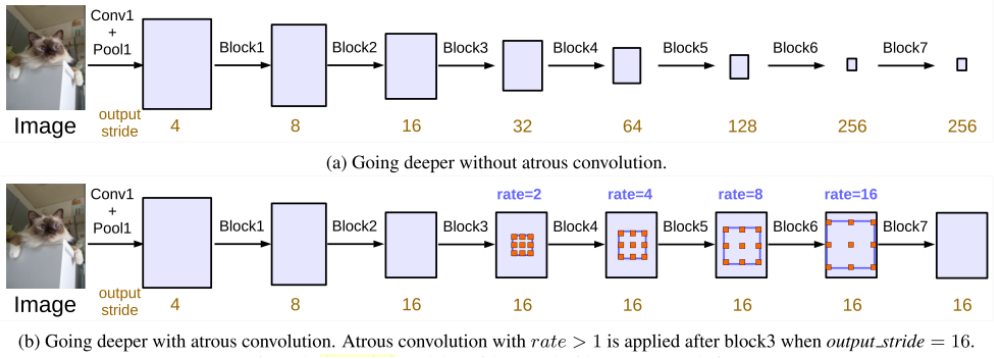

Figure 5.10: Cascaded modules.
Source: Chen et al.[5]

Chapter 6

Previous work - Rastermap Vectorization

Image segmention has come a long way on images of normal everyday objects with deep CNNs. However none of the reviewed papers use their networks on raster maps. Not a lot of work has been done in regards to vectorization of raster maps using neural networks. However there has been a fair amount of research done on other image analysis techniques.

6.1 Non-artifical intelligence methods

Leyk and Boesch[19] presents a color image segmentation of raster maps from the 19th century suffering from poor quality with a clustering technique using the local image plane, frequency domain an color space. The goal of the color image segmentation is to reduce the color values to fit the original colors used when printing the map. The method managed to segment lines, symbols and areas that belong to different color layers, however, there were still some minor classification errors that had to be solved manually.

Iosifescu et al.[13] use open-source solutions to vectorize historical maps from the 19th century. Their procedure consists of five steps: Scanning of the map, georeferencing the map, pre-processing of an image to clean artifacts, automatic vectorization, automatic cleaning of the results. The authors note that the pre-processing step and scan quality are the most crucial for the performance of the vectorization and have to be customized for each set of raster maps. The pre-processing consist of RGB channel processing, conversion to binary images and cleaning. By processing the RGB channels in the image, different features on the map can be highlighted.

The pre-processed image is then converted to vector format with Geospatial Data Abstraction Library[24](GDAL)'s polygonize and contour methods. The results are acceptable when taking the quality of the maps into consideration.

Henderson et al.[11] do semantic segmentation of raster maps with three different unsupervised algorithms: k-means, graph theoretic and expectation maximation. The maps have 6 color values and the segmentation technique is based on the knowledge of these and thus limits the application on more advanced maps.

Chiang and Knoblock[6] present a semi-automatic technique for road extraction from raster maps with a system they call *Strabo*. The system consists of two components: Road layer extraction and road layer vectorization. In their components, they use techniques such as mean-shift, k-means and Hough Transform. They compare they systems performance against R2V [42]. Strabo performed better in 58.3% of the cases. Generally, their road vector lines manage to stay in the center of the roads more than R2V. A limit of their system is that it struggles to correctly label roads that are very thin on the map.

Miao et al.[22] propose a *superpixel*[30] approach to extract height curves from raster maps. Their method, named *Guided Superpixel Method in Topographic Map*(GSM-TM), consists of three parts: Linear feature extraction, boundary detection, and guided watershed transform. For the linear feature extraction, they merge two negative and two positive Gaussian filters. For the boundary extraction, they use a technique of color boundary detection proposed by Yang et al.[43]. The guided watershed transform was introduced since standard watershed is very sensitive to weak boundaries. The guided part consist of modifying the boundaries obtained in the boundary detection step with the lines obtained in the linear feature extraction to make the boundaries clearer before the watershed transform. The results show that the proposed GSM-TM method performs better than the other superpixel algorithms they compare with.

6.2 Artificial intelligence methods

6.2.1 VecNET

VecNet proposed by Karabork et al. in 2008 is one of few examples of vectorization of cartographic raster maps using a neural network. The authors present a three-step process consisting of thinning, line tracking with ANN and simplification. The main goal of the network is to find the critical points of objects, that is, to find breakage points of lines. They use an ANN with an input layer, a hidden layer and an output layer to classify. The training set is very small with only 16 samples. The output layer is a single vector with size 12, where the 8 first places represent an 8-way chain code of directions (the direction the line is following) and the last four represent a prediction of where the next pixel is going to be. It evaluates if the point is critical by checking if the last 8-way direction is different from the one currently calculated. If the point is critical,

they store it. The algorithm is tested on a single raster map only consisting of lines and does not perform better than a sparse pixel algorithm, but manages to get acceptable results according to the authors.

Chapter 7

Discussion

As we can see from the previous work, little research has been done in regards to using deep CNNs to vectorize scanned raster maps. Most of the related work is also focusing on the extraction of linear features, such as countour lines and roads.

One of the goals must be to get a georeferenced vector as output. Not only the vector.

Lot of training data needed, for custom maps this is a one time usecase

Use already digitized data, in a sort of microtasking way to generate enough training data.

Proposed method is to generate training set of the existing vectorized maps, however there is a problem with that the PDFs are not georeferenced. Only the bounding area of the zone is vectorized, we therefore have to locate the zoning area within the PDF to make a training image that is the same size as the input PDF.

Chapter 8

Implementation

Chapter 9

Conclusion

With many plans already digitized, we could create a data set for official Norwegian zoning regulations.

Appendix A

Acronyms

FTA Fault tree analysis

MTTF Mean time to failure

RAMS Reliability, availability, maintainability, and safety

Appendix B

What to put in appendixes

This is an example of an Appendix. You can write an Appendix in the same way as a chapter, with sections, subsections, and so on. An appendix may include list of code (in case you are programming), more details about results that you have presented in the report (could be a more complete description of results, in case you decided to focus on the most important ones in the main report), supplementary information and descriptions you have found relating to the system you are analysing, such as drawings. You may discuss with your supervisor what are relevant information for appendixes.

B.1 Introduction

B.1.1 More Details

Bibliography

[1] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. pages 1–14.

[2] Bø, T. (2009). En veileder basert på praktiske erfaringer fra Telemark og Vestfold.

[3] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. pages 1–14.

[4] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2016). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. pages 1–14.

[5] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation.

[6] Chiang, Y. Y. and Knoblock, C. A. (2013). A general approach for extracting road vector data from raster maps. *International Journal on Document Analysis and Recognition*, 16(1):55–81.

[7] Dori, D. (1997). Orthogonal Zig-Zag: An algorithm for vectorizing engineering drawings compared with Hough Transform. *Advances in Engineering Software*, 28(1):11–24.

[8] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

[9] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). A Review on Deep Learning Techniques Applied to Semantic Segmentation. pages 1–23.

[10] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *Multimedia Tools and Applications*, pages 1–17.

[11] Henderson, T. C., Linton, T., Potupchik, S., and Ostanin, A. Automatic Segmentation of Semantic Classes in Raster Map Images.

[12] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network. pages 1–9.

[13] Iosifescu, I., Tsorlini, A., and Hurni, L. (2016). Towards a comprehensive methodology for automatic vectorization of raster historical maps. *e-Perimetron*, 11(2):57–76.

[14] Karabork, H., Kocer, B., Bildirici, I. O., Yildiz, F., and Aktas, E. (2008). A neural network algorithm for vectorization of 2D maps. *[APRS'08] International Archives of Photogrammetry and Remote Sensing*, XXXVII(B2):473–480.

[15] Karpathy, A. CS231n Convolutional Neural Networks for Visual Recognition.

[16] Kommunalt Planregister (2009). § 2-2. Kommunalt planregister.

[17] Krähenbühl, P. and Koltun, V. (2012). Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. pages 1–9.

[18] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9.

[19] Leyk, S. and Boesch, R. (2010). Colors of the past: Color image segmentation in historical topographic maps based on homogeneity. *GeoInformatica*, 14(1):1–21.

[20] Lin, X., Shimotsuji, S., Minoh, M., and Sakai, T. (1985). Efficient diagram understanding with characteristic pattern detection. *Computer Vision, Graphics and Image Processing*, 30(1):84–106.

[21] Long, J., Shelhamer, E., and Darrell, T. (2014). Fully Convolutional Networks for Semantic Segmentation.

[22] Miao, Q., Liu, T., Song, J., Gong, M., and Yang, Y. (2016). Guided Superpixel Method for Topographic Map Processing. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11):6265–6279.

[23] Monagan, G. and Roosli, M. (1993). Appropriate base representation using a run graph. *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*, pages 623–626.

[24] OSGeo. GDAL.

[25] Papandreou, G. (2015). Modeling Local and Global Deformations in Deep Learning _ Epitomic Convolution,Multiple Instance Learning, and SlidingWindow Detection. *Cvpr*, pages 390–399.

[26] PASCAL VOC (2012a). PASCAL VOC2011.

[27] PASCAL VOC (2012b). Segmentation Results: VOC2012.

[28] Patterson, J. and Gibson, A. (2017). *Deep Learning: A Practitioner's Approach*. O'Reilly Media.

[29] PowerTRACE (2016). Taking Corel PowerTRACE for a Test Drive – Knowledge Base.

[30] Ren, X. and Malik, J. (2003). Learning a classification model for segmentation. *Proceedings Ninth IEEE International Conference on Computer Vision*, 1(c):10–17 vol.1.

[31] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

[32] Sabour, S., Frosst, N., and Hinton, G. (2017). Dynamic Routing between Capsules. *Nips*, (Nips).

[33] Scan2cad (2009). Scan2CAD in Landscape Architecture.

[34] Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. pages 1–14.

[35] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

[36] Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era.

[37] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., Hill, C., and Arbor, A. (2014). Going Deeper with Convolutions. pages 1–9.

[38] Visin, F., Kastner, K., Cho, K., Matteucci, M., Courville, A., and Bengio, Y. (2015). ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks. pages 1–9.

[39] Wenyin, L. and Dori, D. (1996). Sparse pixel tracking: A fast vectorization algorithm applied to engineering drawings. *Proceedings - International Conference on Pattern Recognition*, 3:808–812.

[40] Wenyin, L. and Dori, D. (1999). From Raster to Vectors: Extracting Visual Information from Line Drawings. *Pattern Analysis & Applications*, 2(1):10–21.

[41] Worboys, M. F. (2003). *GIS: A computing perspective*.

[42] Wu, Y. and Able Software Corp (1999). R2V.

[43] Yang, K., Gao, S., Li, C., and Li, Y. (2013). Efficient color boundary detection with color-opponent mechanisms. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2810–2817.

[44] Yu, F. and Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions.

[45] Zangeneh, M., Omid, M., and Akram, A. (2011). A comparative study between parametric and artificial neural networks approaches for economical assessment of potato production in Iran. *Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) Spanish Journal of Agricultural Research*, 9(3):661–671.

[46] Zeiler, M. D., Taylor, G. W., and Fergus, R. (2011). Adaptive Deconvolutional Networks for Mid and High Level Feature Learning. pages 2018–2025.