Chapter 1

Introduction

There has been significant progress in visual object recognition in the recent years. In 2010 the classification between cats and dogs in computer guided image recognition was not a quarter as good as human image recognition. In 2016 artificial Neural Networks can tell apart pictures of a Siberian husky (figure 1.1a) and an Eskimo dog (figure 1.1b) which most humans can't. So what has happened in the recent years?

(a) Siberian husky (b) Eskimo dog

Figure 1.1: [SLJ+14, Page 3, Fig. 1.]

With the development of Convolutional Neural Networks (CNN) the AlexNet[KSH12] provided an approach in 2012 where for the first time image recognition was done by an artificial intelligence on human level. With the use of a CNN not only image recognition becomes possible, but also image segmentation becomes feasible. This is the idea behind this thesis.
I'm using a CNN, which is described in this thesis, to find out which pixels of an image are responsible for the initial input stimulus to lead the image classification of the same Neural Network to an decision on which class is recognized. With the extraction of these responsible pixels it is reasonable to think about further use for this newly gained information in object tracking. There are often pixels in the object recognition process involved which do not lay inside the actual object space of the image. For example, when the for this thesis developed CNN, further referenced as VResNet, tries to tell apart planes from other objects in the CIFAR-10 dataset it only uses the blue background of the sky to recognize a plane. To be able to use this approach for a sound object tracking, one has to be careful to train a CNN in the correct way with the right datasets. In this thesis we will see one way to develop an CNN which can effectively track objects.

1.1 General Introduction to this Thesis' Topic

The idea behind Neural Networks is to emulate some concepts of the human brain, like the massive parallel processing of information through many Neurons and Synapses.

	computer	**human brain**
computational units	1 CPU, 10^{15} gates	10^{11} neurons
storage units	10^9 bits RAM, 10^{10} bits disk	10^{11} neurons, 10^{14} synapses
cycle time	10^{-8} sec	10^{-3} sec
bandwidth	10^9 bits/sec	10^{14} bits/sec
neuron updates/sec	10^5	10^{14}
connectivity	on one CPU all gates are theoretical connected to each other	max shortest path between two neurons = 4[1]
signal propagation speed	c (speed of light)	up to 119m/s

Table 1.1: von-Neumann computers vs. brain:

(source for all rows except the last two [Fer15])

"The clocking rate of the Brain is slow (10^{-3} *seconds), but updates are massively propagated in parallel. To simulate this on a CPU in a serial way needs hundreds of cycles."* [Fer15] Because of that modern artificial Neural Networks use GPUs to emulate this parallel processing of information, but still cannot compete with the human brain in its parallel processing capability. The number of neurons in the human brain is still too high for a reasonable comparison with a GPU.
The big advantage of parallel processing versus serial processing is on the one hand the computational speed and on the other hand the prevention of a single point of failure. Because even if some neurons are dead they will receive lower gradient updates over time, which means they are ignored by the rest of the network. So the network can still work properly even if parts of it are dead. This enables a high system stability and an antifragile [Tal12] quality.
The first scientific work about artificial neural networks was released in 1943 [MP43]. Since then *"the field has experienced several hype cycles, followed by disappointment and criticism, followed by funding cuts, followed by renewed interest years or decades later"* [Wik17]. These periods of reduced funding and interest in research of artificial intelligence in general are called "AI Winter" and there were two major winters in 1974–80 and 1987–93.
The first AI Winter was caused by a large decrease in AI research in the United Kingdom in response to the Lighthill report, which is the name commonly used for the paper "Artificial Intelligence: A General Survey" by James Lighthill, published in Artificial Intelligence: a paper symposium in 1973". This report states that *"in no part of the field have discoveries made so far produced the major impact that was then promised"*. It caused a huge setback for the interest and funding of AI research in the United Kingdom. This setback caused a world wide decrease in this research area, because the

[1] with the first neuron reaching 7000 other neurons. Two connections staring from one neuron would reach 7000*6998.28 other neurons, Three connections 7000*6998.28*1974.93 neurons and four connections 7000*6998.28*1974.93*1.0334 neurons. So the probability for the case that 4 connections are necessary to connect 2 neurons is very low. However this is only a calculation based on the table above that there are 10^{11} neurons and 10^{14} synapses and with the assumption that all the connections (synapses) between the neurons fall under the law of the Zeta distribution. So the shortest path between to specific neurons might be higher than four connections

United Kingdom was a major driving force since the beginning through its first pioneer in AI research named Alan Turing. Decades later the Lighthill report was criticized in professional circles and its impact on the research today is pretty much not existing. But critical debate is always important in scientific research and the Lighthill report might have had a positive effect on the AI research in general.
The second AI Winter might not have had such a big impact directly on artificial Neural Networks, but it was sufficient enough that some researchers didn't want to publish their research. It was not until more than a decade later that major research was popularized, which solved problems at training multi layer Neural Networks. The second AI Winter was caused by the collapse of the Lisp machine market in 1987. The Lisp machine was important for that research because it had a high-level language computer architecture, which allowed a faster and simpler development of scientific programming and specifically programming of artificial Neural Networks.
Today's AI research and especially artificial Neural Networks are one of the most promising research fields with very high funding through multinational technology companies like Google or dedicated projects like OpenAI, which was founded by Elon Musk, Sam Altman et al.. Artificial Neural Networks gained public and lesser academic interest after AlphaGo (Google's deep learning Go AI) has beaten Lee Sedol, the world No.1 ranked player in the game Go at the time. Some even call the presumably last Go game a human will ever have won against a leading Go AI "the game of the century", which denotes the game that took place on March the 13th in 2016 between AlphaGo and Lee Sedol in fulfilling this prediction. A sophisticated Go AI that can beat professional human players was an important step for the AI development. Chess AIs were able to beat human players on a regular basis since the 90s and Go was the last game in this category for purely logic and simple rule based games without any luck based mechanisms. Go was a strong ambition for the AI development, because of two reasons. At first there are much more possible moves per turn than in Chess, and secondly there is a lot of pattern recognition necessary to evaluate the state of the game. The second problem is a perfect problem for CNNs, which are part of AlphaGo. The first problem could be solved by the enormous computational power of the computer AlphaGo was running on.
But since AlphaGo renewed the perspective for Go and revealed new strategies for human Go players it remains to be seen if Lee Sedol played the game of the century for real.

1.2 Short Overview about the Motivation and Challenge of the Thesis

The motivation was to find a way to do object recognition and tracking that uses a state of the art deep Convolutional Neural Network like ResNet. I didn't expect to reach ground breaking scores with this approach, but I wanted to modify ResNet to make it applicable for deconvolutional purposes like object localization. The problem is, that all state of the art CNNs use nonlinear network architectures and nonlinear networks are difficult to transform into a deconvolutional version of themselves. More on this later in chapter [4.1.2]. The challenge of this thesis was to find a way to make a transformation from a nonlinear ResNet into a linear deconvolutional version of ResNet possible without loosing the benefits of ResNet. I explain those benefits later in chapter [3.1.3]. Another challenge was the implementation and training of the Neural Network, which required a lot of testing and optimization, because Neural Networks are still

unpredictable and their outcome is difficult to retrace.

1.3 Short Explanation of the Approach

The approach consists of two parts.
Object detection and object localization/tracking.
The detection part shows how to use a CNN with supervised learning to train a Neural Network to recognize different classes of images from the CIFAR-10 dataset or other datasets, which lay the basis for the second part of my approach.
The second part of my approach is about the localization (tracking) of objects with a Deconvolutional Neural Network. I differentiate between tracking and localization, because the real meaning of tracking objects is to track them on video data and not just on different unrelated images. But for a sound object tracking a good localization is a necessary objective to reach first. However I didn't use video data in my implementation, but my approach covers the theoretical use of video data for object tracking. More on this in chapter [6.1].

1.4 Overview about Each Chapter of the Thesis

- Chapter [2] *"Related Works"* is about other known approaches for similar objectives with deep learning.
- In chapter [2.2] I also talk about differences between this thesis' approach and the the current state of scientific knowledge.
- Chapter [3.1] describes the technical background for every module that is used in the modularly built Neural Network which is part of the approach of this thesis. In addition to that I give a recap on how artificial Neural Networks work in general and what convolution and deconvolution in this context does.
- Chapter [3.2] gives advice about how to prevent different problematic effects, that often occur in Neural Networks.
- Chapter [3.3] sets the mathematical background for the most important mathematical methods that find their use in the context of deep learning.
- Finally chapter [4] *"Approach"* explains my approach for object recognition and tracking application that use a deep Convolutional Neural Network, which I implemented and tested with Torch7.
- Chapter [5] *"Experimental Results"* shows results for different parts of the approach and reviews the usability of this approach. At the end of this chapter I also go into detail about some major parts of the implementation.
- Chapter [6] *"Further Analysis"* gives a short summary of this thesis and a résumé of my thoughts on the gained results and the optimizations the approach would benefit from. After that I give a perspective about possible future works I would recommend for realizing object tracking with deep learning and steps I have planned to do next on this topic.