# DEEP LEARNING
## for Natural Language


A.DEMIRAJ

# DEEP LEARNING
# for Natural Language


Alban Demiraj
St Anne's College
University of Oxford

A dissertation submitted in partial fulfilment for the degree of
*Masters of Science*

Trinity 2014

Advisor: Nando de Freitas Author: Alban Demiraj

# *DEEP LEARNING for Natural Language*


## Abstract

Encoding symbolic concepts like words, sentences and documents with distributed representations has excited researchers for decades. Furthermore capturing the compositional process that maps the meaning of words to higher level structures like sentences and documents is a central challenge for researchers in Natural Language Processing and Information Retrieval.

In this thesis we present three novel solutions. The first is a sentence model able to map the meaning of words to sentences, by embedding the later in low dimensional vector space. It is closely related to the Dynamic Convolutional Neural Network (DCNN) which learns convolutional filters at the sentence level in increasing order of abstractness, hierarchically learning to capture and compose low level lexical features into high level semantic concepts. Our model substantially reduces the number of parameters of the DCNN providing a more compact solution that can scale much better.

The second model, extends the first one by learning convolutional filters at both the sentence and document level, hence mapping the meaning of words to that of document, embedding the later in a low dimensional vector space. Our model preserves distinctions of word and sentence order crucial for capturing nuanced semantics.

Third, we propose an automatic summarisation algorithm, by exploiting the structure of our models. Contrary to the general approach in the field, our model does not inspect sentences in turn to decide their importance in a document. Instead it builds an understanding of the document by embedding it in a low dimensional vector space and then informs us which sentences were the most important at helping it capture the meaning of the document.

We demonstrate the effectiveness our our model on text understanding tasks like sentiment analysis, where we achieve very strong results with no engineered features.

To my grandparents.

# Acknowledgments

Thank you GOD, thank you!

I am most grateful to Professor Nando de Freitas for offering me this fascinating research project. Furthermore, the achievements of this thesis are only possible because of his invaluable insight and guidance. Even more importantly I would like to thank him for his amazing lectures. They inspired my Deep Learning interest and made the most complicated topics easily graspable.

I would also like to thank Professor Phil Blunsom. His machine learning course was the most difficult I have ever had to face, but in the same time the one I loved the most. His research is a fundamental piece of this thesis, and he also directly contributed in this project with invaluable insight and suggestions.

I would like to show my deepest gratitude to my departmental supervisor, prof. Georg Gottlob. He helped me make great choices during my year at Oxford and he was always there to support me in times of uncertainty.

A special thanks is reserved for Misha Denil, his help throughout this project has been tremendous. Furthermore, his expertise in the field always directed me to the right solutions when my approaches would hit bumps.

I cannot forget professors Claudia Roda and Eugeni Gentchev of the American University of Paris. The education I received during my time there has been a fundamental block for the work that would follow. A very special thank you for my undergraduate supervisor Professor Georgi Stojanov. His classes and his work guided me to the fascinating field of Artificial Intelligence and he is the person to have single-handedly contributed the most to my academic life.

Finally I would like to thank my family, my mother Edlira, my father Agron, my brother Ardit and my girlfriend Lorena for being extremely supportive of me, even when countless sleepless nights would play bad trick with my temper. My friends in the Machine Learning lab made the extensive hours working on this project even more enjoyable.

Alban

# Contents

# Listing of figures

# Part I

# Introduction

*A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.*

A.Turing

# 1 Introduction

## 1.1 MOTIVATION

Natural Language Processing is one of the oldest fields in Artificial Intelligence. Its inception can be traced back to 1950 with the proposal of the Turing Test (Turing, 1950) as a way to identify the existence of an intelligent entity. The test consisted of a human in front of a communication terminal engaging in a conversation with an entity on the other side of the terminal. The entity on the other side of the terminal could be either another human or an Intelligent Agent. Turing proposed to call AI an entity able to fool a human is such set-up.

The idea is often considered very forward minded as more than 60 years later it remains valid. This is mainly due to difficulty involved in manipulating Natural Language. In the incepting years of the field of AI, however, there was a shift from thinking humanly, to acting rationally. This was mainly driven by the very hard task of understanding how the human brain works, let a lone emulating it in a machine. This shift, made one of the initial and most important tasks in Natural Language Processing (NLP) to convert Natural Language text in a format, which machines could understand. This challenge has been two-fold, first defining what a good representation is, and than translating Natural Language to the specific structure. Experience however, has shown this to be a wrong direction. For example, in Machine Translation (a task of NLP) the idea

of converting from language A to an Interlingua then to a language B, has been dropped in favour of Statistical word alignment which tries to directly match words in one language with words in the target one (Brown et al., 1993). In other NLP hard tasks like Question Answering, even commercial systems with very small operational domains and very intensively engineered knowledge bases, like Siri or Worlfram, perform only adequately good.

Currently, however, there is a trend in Machine Learning shifting AI to the original idea of having an agent with human like intelligence. The development of the artificial neuron (McCulloch Pitts, 1943), the perceptron learning algorithm to train these neurons (Rosenblatt, 1958), the discovery of the back-propagation algorithm to train models where these neurons are stacked in layers (neural networks) much like in the human brain (Rumelhart et al., 1986) and the reductions in their training costs (Denil et al., 2013) have closed the gap that existed between brain thinking and the machine processing. Furthermore, advances in Neuroscience have shown that different parts of the brain are equally capable of performing task they are not originally trained to. For example the visual percepts were routed to somatosensory cortex (Métin Frost, 1989) or to the auditory cortex (Roe et al., 1992), in both cases the brain cells learned to see. This has given rise to what is known as "The one learning algorithm hypothesis", by which we hope to build one single model that will be able to perform all the tasks.

Current work in Deep Learning is inspiring. It has almost entirely taken over the field of Computer Vision and lately considerable work is going on in NLP on which the Oxford Computational Linguistics Research Group is having a great contribution. At this fast paced progress in Deep Learning, exposing the current developments to new tasks hoping to show their success or otherwise identify faults early on and maybe suggest corrections as we progress, we believe is a good contribution to the field.

## 1.2 SCOPE AND CONTRIBUTION OF THIS THESIS

In this thesis we tackle the problem of capturing the meaning of natural language structures like words, sentences and documents in a machine readable format. Furthermore, considering lower level symbolic concepts, like words, group to form higher level ones, like sentences, we also try to capture the compositional process in which the meaning lower level concepts maps to higher level ones. This has been a target of research in Natural Language Processing for a long time. However, initial approaches were based on Context Free Grammars and Lambda Calculi. In this thesis we try to capture the meaning of natural language structures and their compositional process by embedding them in low dimensional vector spaces with the deep learning approach. As a result of the research conducted in the thesis we propose the following:

- A sentence model which can capture the compositional process that maps the

meaning of words to that of sentences. Our model applies a Convolutional Neural Network architecture and learns feature detectors over words. Considering Neural Networks work on continuous inputs, our model first converts the words to distributed representation which it optimizes via back-propagation. As a consequence in the same training process it learns to embed both the meaning of words and that of sentences in low dimensional vector spaces. Our model is closely related to the Dynamic Convolutional Neural Network (DCNN) of Kalchbrenner et al. (2014). Nevertheless, we propose important changes which allow the number of parameters of the DCNN to drop tremendously consequently causing a considerable speeding up in its training process and a reduction in its memory requirements. This allows us to use this model as a building block to the bigger model we will describe next.

- A novel document model, which extends our sentence model to map the meaning of words to that of documents, but passing through the sentence representation first. To the best of our knowledge, we are the only ones to keep this important intermediary representation, which becomes crucial in our next application. Furthermore, in the same training process, our model also learns low dimensional vector space embeddings for both words and sentences. The technique we use allows for further up-scaling allowing the a whole mapping ($word \rightarrow sentence \rightarrow paragraph \rightarrow chapter \rightarrow book$) to be learned in one training.

- A novel automatic summarisation algorithm which exploits the sentence structure preserved by our document model. Our algorithm is different to the currently used models in the field. It does not in turn decide whether a sentence is important or not, but instead exploits the model's understanding of the document to pinpoint important sentences. Working for this model, we also introduce an interesting scalable automatic summarisation evaluation metrics.

The findings and developments in this thesis have resulted in the "Modelling, Visualising and Summarizing Documents with a Single Convolutional Neural Network" paper[1] by the Machine Learning Research Group at Oxford.

The contributions we hope to make with this thesis are not limited to the technical advancements listed above. Tackling the important problems discussed above, we brought together knowledge from different courses like Machine Learning, Advanced Machine Learning, Computational Linguistics, Intelligent Systems and Foundations of Computer Science. This knowledge we used should be introduced in a summarized version on the Theoretical Background chapter. Nevertheless, writing a good amount of text, summarizing important concepts from these courses with no value added, was

[1]The paper has been submitted to NIPS 2014 and is currently under review as the time of writing.

not desirable. For this reason we went one step further, and while presenting important fundamental concepts, we tried to fill a void we found in the literature. A lot of the algorithms used today in Machine Learning and especially in Neural Computation have strong biological backing. Nevertheless, most books in the field focus deeply in the equations governing these models that they do not introduce the biological inspiration behind these algorithms.

Embarking on a machine learning course with no prior machine learning knowledge is hard and it is fairly easy to get lost in the vast amount of summation and multiplication notations. Nevertheless, if the reasoning behind those equations is exploited than they make much more sense than a paragraph of English text explaining the model. Considering this thesis focuses on neural computation, which is inspired by biological processes, we will try and explain the algorithms by running an analogy with the brain. We hope this way the reader can see the perspective of the inventor when the model was first introduced. Our hope is that this thesis, especially the Theoretical Background chapter can be used as a smooth introduction to machine learning for new students. In fact we would like this thesis to be a starting point for a student pursuing future research in the same direction. For this reason we include an introduction of all major developments with clear references for deeper explanations. However, not all important techniques have a biological backing and sometimes the perspective of the author is not known. In this scenarios, while presenting the algorithms the way they are introduced in the literature, we also offer a different perspective on their functioning.

## 1.3 THESIS STRUCTURE

This thesis is divided in three main parts. The first part is of introductory nature and apart from the current chapter, it includes Chapter 2 which summarizes the algorithms and models we used. The second part introduces the methodologies we suggest to solve the problems we tackle. It includes Chapter 3 which presents our sentence model, Chapter 4 which presents our document model and Chapter 5 which presents our summarisation algorithm. The last part of the thesis is mainly conclusive. Chapter 6 puts our algorithms and models in test and compares them to competitors in the field. Chapter 7 offers reflections on how we can solve problems identified during the experiments, in the mood of future directions. Chapter 8 is the final chapter of the thesis and offers a conclusion. A summary of the thesis follows:

**Chapter2: Theoretical Background**
This chapter aims to provide a self contained background of the technologies used in this thesis. As we discussed above, this introduction does not offer a simple summary of the algorithms but it also offers an analogy with the biological brain when this is pos-

sible, or otherwise a different perspective on thinking about the algorithm.

**Chapter3: Convolutional Sentence Model**
This chapter introduces the first contribution of our work, the Convolutional Sentence Model (CSM). The goal of the CSM is to build low dimensional vector space embeddings that capture the meaning of sentences. This chapter describes in turn the layers and operations that define our model. It then proceeds to describe the how this model can be trained. The CSM is very closely related to the CDNN, for this reason most of this chapter can be considered a parallel explanation of both models. Nevertheless, our model differs from the DCNN, in the way it applies the convolution operation. This allows the CSM to have a considerably reduced number of parameters which make it much more scalable. The chapter explains the differences between the two models very explicitly.

**Chapter4: Convolutional Document Model**
In Chapter4, we move to the introduction of the novel document model dubbed the Convolutional Document Model (CDM). The goal of the CDM is to bring the CSM to the next stage and build low dimensional vector space embeddings for entire documents. Nevertheless, our CDM keeps the sentence embeddings as an intermediary representation. The CDM extends on the CSM, for this reason most of the operations are similar, so the chapter focuses mainly on describing the architecture of the model and how it can be trained.

**Chapter5: Saliency Extraction and Automatic Summarisation**
This chapter introduces a novel automatic summarisation technique. The idea is borrowed from the visualisation techniques in Convolutional Neural Networks for computer vision. Nevertheless in the natural language domain, it has the much nicer application of automatic summarisation. The technique we use to summarize sentences can be explained intuitively in a very simple way, nevertheless it has strong mathematical reasoning behind it. This chapter presents both the intuitive explanation and the mathematical reasoning. The later highlights the fact that this algorithms is strongly dependent on the CDM, and in fact is only possible because of the structure of the CDM. Together with the summarisation technique, we also present a scalable summary evaluation technique that requires no human supervision.

**Chapter6: Experimental Results**
In this chapter we put all our models to test. First we see how the CSM compares to the DCNN on which it was inspired in a sentence sentiment classification task. Next we test the CDM and compare it to other models. Again we use sentiment classification as

the task of our choice, however this time the dataset consists of documents rather then sentences. In the last part of this chapter we put to to test our summarisation technique.

**Chapter7: Blueprints for Future Directions**
In this chapter we try to reflect on the results of the previous experiments. Our reflection suggest directions that would improve our models further. Nevertheless we only include directions we have extensively though about and have a very clear idea on how to approach them. In fact we call the chapter Blueprints on Future Directions as in most of the cases we describe the whole architecture of the advancements.

**Chapter8: Conclusion**
This is the last chapter of the thesis and as the name suggest brings togather all the advancements, contributions, results and directions of this thesis.

*If I have seen further than others, it is by standing upon the shoulders of giants.*

I. Newton

# 2 Theoretical Background

## 2.1 INTRODUCTION

We, humans, call ourselves homo-sapiens, which stands for man-the-wise. This shows how important our intelligence is to us. It differentiates us from other animals living earth. In fact, it's controversial to call the human being an animal, despite the fact that the only major difference, is our intelligence. Furthermore, French philosopher René Descartes, based his proof of our existence in our ability to think (Descartes, 1967).

The thinking process, how matter can generate thoughts, perceive and understand things, has challenged researchers for a long time. It is known that the brain is the organ responsible of thinking, however understanding how brain operates is still the focus of intensive research in many disciplines like Biology, Philosophy, Neuroscience and Computer Science.

The later of these disciplines, Computer Science, goes one step further and aims to simulate human intelligence. A lot of algorithms have been proposed to solve problems which were previously considered intellectually challenging and exclusive to humans. A noticeable mention is IBM's Deep Blue chess program which beat Chess World Champion Gary Kasparov in 1997, or IBM's Watson which won the Jeopardy competition in 2011. Despite these wonderful achievements, which have renewed interest in AI and have allowed for further research to be funded, these algorithms have proven to be very task oriented. We believe that the answer to making robust systems that easily adapt to new problems and environments lies in the machine we are trying to emulate, our

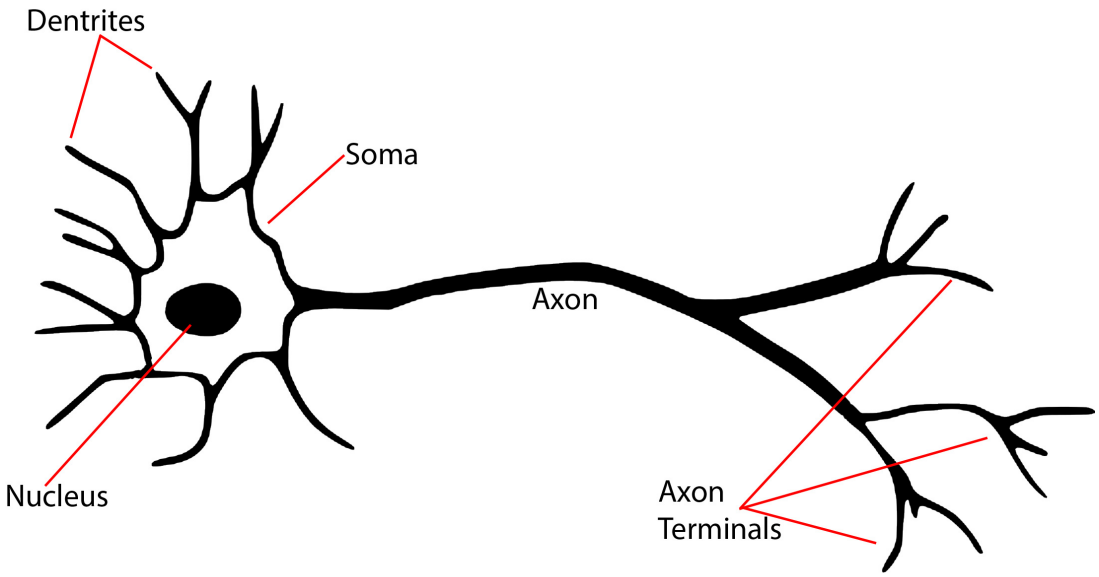



**Figure 2.1:** Sketch of a standard biological neuron.

brain. For this reason, we believe artificial systems that are biologically inspired have a better chance at adapting to new tasks.

In this paper we will focus on such algorithms. For this reason, we are going to start with a shallow description of the brain, its fundamental parts and how basic information is processed and transmitted. Next we are going to describe some of the biologically inspired algorithms in Machine Learning, by running an analogy with the biological brain. Finally, because this thesis focuses on solving problems in the domain of Natural Language Processing, we are going to make a small introduction to syntax and semantics in natural languages.

## 2.2 THE BIOLOGICAL BRAIN

The brain is the jelly-like material found inside the skull. As described above, it is responsible for the cognitive process and is arguable the most complex machine in the universe. Its building blocks are the neurons, of which the human brain contains around 100 billion.

### 2.2.1 Neurons

Neurons or nerve cells, are the basic information processing unit. In difference to other biological cells they have three fundamentally different well defined regions called the Soma, the Dendrites and the Axons. **The soma**, is the central part of the nerve cell. It contains the nucleus and the plasma with characteristics and functionalities similar to other cells. The nucleus contains hereditary information in the shape of DNA while the

plasma contains the materials needed for maintaining the cell. **The Dendrites** are the short extensions of the soma. Their function is to receive stimulus from other neurons. In computational point of view, their function can be seen as an input port. **The axons** are the counterparts of the dendrites. They can be viewed as an output port and they send stimuli to the dendrites of connected neurons. In difference to the dendrites, they are long extensions of the soma. Figure 2.1 depicts a neuron including its main parts.

### 2.2.2 Synapse and Networks of Neurons

Despite the large number of neurons and their highly specialized structure, the brain's capabilities do not come from those of single neurons, but instead from the interconnections between them. In the brain, each neuron is connected to $10^3$ to $10^4$ other neurons. In total the number of interconnections reaches approximately $10^{14}$ - $10^{15}$.

The communication between neurons happens through a structure called synapse. The synapses are of two types, chemical and electrical. In chemical synapse, a the axon of a exited pre-synaptic neuron releases a neurotransmitter chemical [1]. The neurotransmitter then attaches to the dendrite of the connected post-synaptic neuron. In electrical synapses on the other hand, the axon of a pre-synaptic neuron and a post-synaptic neuron are connected through a gap junction which allows electrical changes in the pre-synaptic neuron to be transmitted to the post-synaptic neuron.

### 2.2.3 Hierarchical Structure and Feature Detectors

Synapses create a hierarchical structure at which information is processed in increasing level of abstractness. This discovery is due to Hubel and Wiesel (1959, 1962) who followed the propagation of external photo stimulus in the visual pathway of cat's visual system, from optic-nerves to the striate area of the primary visual cortex. Figure 2.2 shows the full trajectory of the visual stimuli in the nervous system. This discovery earned them the Nobel Prize in 1981.

Early experiments on the axon of the retinal ganglion cells (RGC) known as the optic-nerve and the lateral geniculate body (LGB), had shown them to respond to stimuli of light spots in their receptive field [2]. The cells in LGB have a relatively simple structure receiving input from the optic-nerves and sending output to the striate area of the primary visual cortex. Paired with the fact that they respond to stimuli the same way the RGC do would suggest that neurons have the function of relays.

This is strongly contradicted when cortical cells are analysed. Their response to stimuli is a lot more complicated, which shows that the visual cortex performs profound transformation on the information it receives. Based on the way cells respond to stim-

[1] There are different neurotransmitter chemicals and the synapses are named after these chemicals.
[2] Area where some stimulus can cause a neuron to fire.

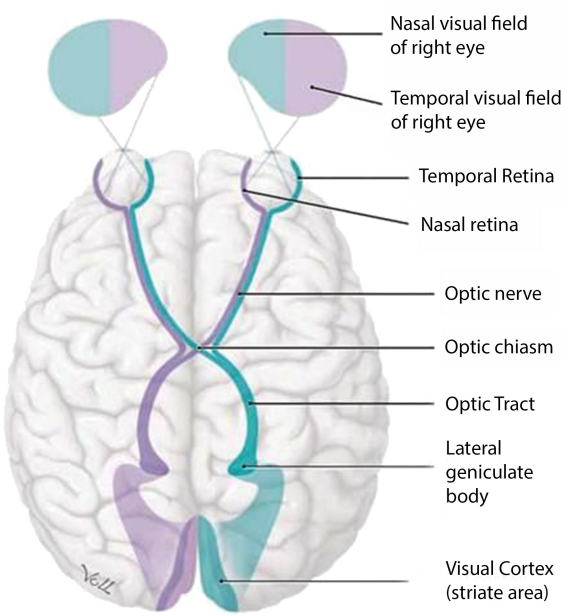



**Figure 2.2:** Visual Pathway.

uli, Hubel and Wiesel grouped them into simple and complex cells. **Simple Cells** are similar to cells in the previous level (RGC and LGB). They have a clearly delinited receptive field, where a spot of light will cause an excitation. In difference to cells in lower level, however, the geometry of the area that will cause the cell an excitation is different. In lower level cells, as depicted in Figure 2.4, a small area where a spot of light will cause the excitation of a single neuron, is surrounded by areas where it won't. In cortical simple cells however, the geometry is different. Cells will still respond to the stimulus of light spots, however, the area to which they respond will be a whole line. Furthermore, Hubel and Wiesel found that the lines at which simple cortical cells would respond were of the four directions depicted in Figure 2.3 (a-d). They found these directions to be evenly represented in cortical simple cells.

The study of deeper cells in the striate cortex, was more challenging as previous researches had not found any source of excitation for these cells. Hubel and Wiesel found that these cells would not respond to simple spots of light, but only to more complex shapes, hence the name **Complex Cells**. In order to excite a single cortical complex cell, a whole line of light as depicted in Figure 2.3 (e-h) should be shone at the retina. To be noted is the similarity to the simple cells with regard to the direction of the lines. In fact complex cells also respond to the same four directions, however if in simple cells any spot in that line would excite the neuron, in complex cells only the whole line (or a good part of it) would excite the neuron . We hope that the difference in Figure 2.3 (a-d) and Figure 2.3 (e-h) makes this obvious.

Another interesting discovery by Hubel and Wiesel was the fact that these neurons would fire for a short time and then stop. They realised that, in difference to previously

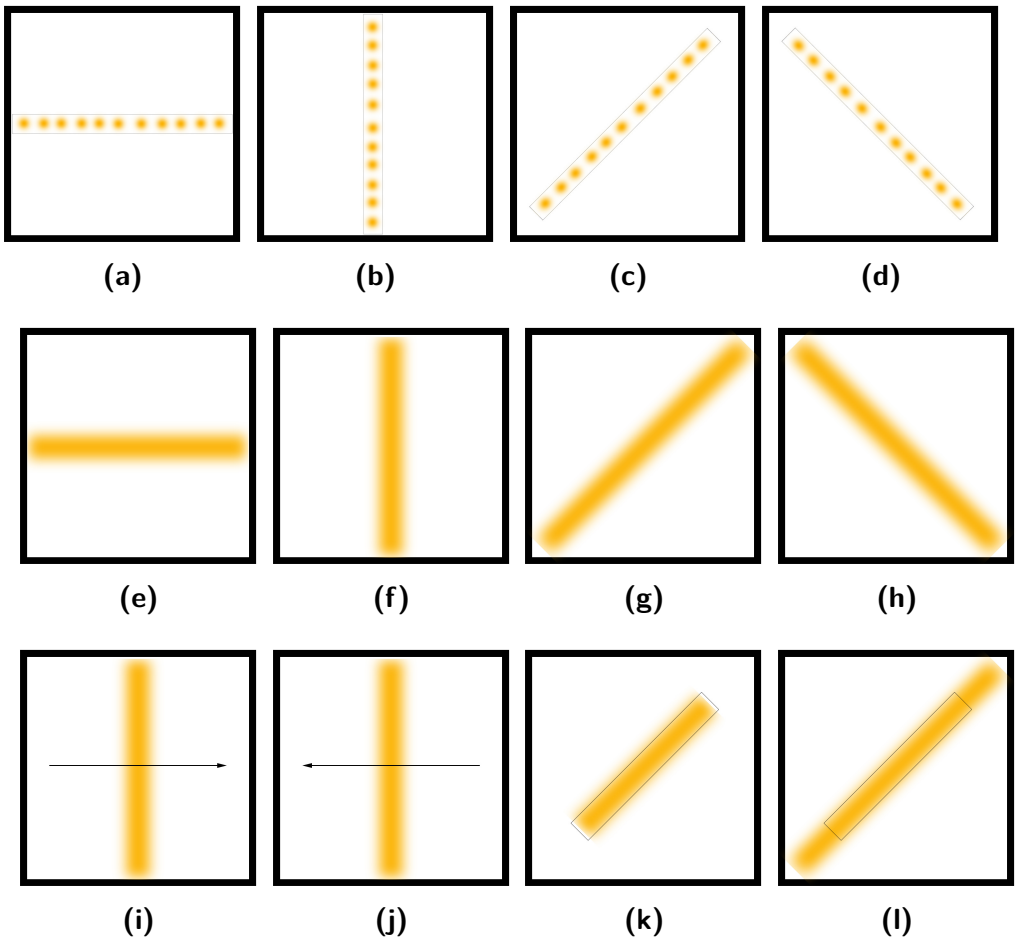



**Figure 2.3:** Receptive field map for cells in the striate cortex.

discussed cells, complex cells would not respond to static light stimuli, but instead only to moving ones. Furthermore, they would also be direction sensitive, responding only to stimuli moving in one direction, but not others or not even the opposite. For example, if the neuron would fire for the stimulus in Figure 2.3 (i) would not respond to Figure 2.3 (j) or the other way around.

Yet another type of complex cells found in the striate cortex is the end-stopping neuron, sometimes even dubbed hyper-complex. The end-stopping cell is similar to the

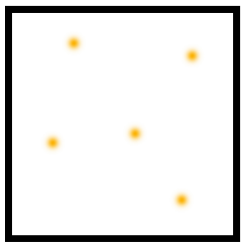


**Figure 2.4:** Receptive field map for cells in the RGC and the lateral geniculate body.

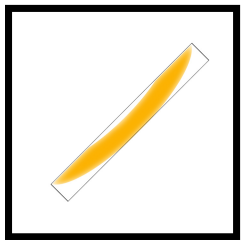


**Figure 2.5:** Receptive field map for end-stopping cells in the striate cortex.

complex cells discussed above, but its activation depends on the length of the line of light used as stimulus. It can be excited by a short line of light and the excitation keeps increasing as the length of the light increases up to a certain limit and then it start decreasing again until it does not respond any more. This is depicted in Figure 2.3 (k and l). The lined rectangle draws the size of the line that would cause maximum excitation of the neuron. In Figure 2.3 (k) the actual light is identical to the rectangle, so we expect the neuron to fire at its maximum intensity. In Figure 2.3 (l), however, the line is longer than the rectangle, so the neuron either fires at a lower intensity or does not fire at all. Hubel and Wiesel did suspect the end-stopping cells to be useful at identifying curves, as a part of the curve can overlap with a small portion of a line, but then go in a different direction. Figure 2.5 tries to depict this. The light (represented by the yellow line) is part of an oval shape. A normal complex neuron if fired would give us the perception of a straight line. Instead this shape would also activate an end-stopping neuron as seen in the picture and will thus give the brain more information on which it can deduce more complicated shapes. Hubel and Wiesel suggested that these cells should be higher in the hierarchical structure of the visual cortex.

A distinguishing and very important feature of all complex cells, is the area to which they respond. In difference to simple cells that have a very well defined area where the stimulus will cause an excitation, complex cells fire for stimuli anywhere in the visual field, provided they have the right size, shape and direction.

### 2.2.4 Adaptability of Specialized Cortex

The discovery of feature detectors in the visual cortex gave a very important insight of how this part of the brain works. Two different research groups investigated the implications this has for other parts of the brain. Métin and Frost (1989) permanently rerouted the output of retinal ganglion cells in newborn Syrian hamsters from the lateral geniculate body to the primary somatosensory cortex. Then they did experiment on the adult animal to find out that the somatosensory cells did respond in a similar way to normal visual cortical cells (described in section 2.2.3). Similarly Roe et al. (1992) rerouted the output retinal ganglion cells in ferrets from the lateral geniculate body to

the primary auditory cortex. Similarly to the previous experiment, this showed cells in the auditory cortex to respond to stimuli much in the same way cells in the visual cortex would.

These discoveries show an important direction for the future. *First*, they show that different parts of the brain must very likely process information in the same hierarchical structures that the visual cortex does. *Second*, they show that different parts of the brain, are adaptable to different tasks and their ability is transferable. *Third*, they also show that the feature detector are learned and not born with. In case they were born with the animal the non-visual cortices would not respond to line shaped visual stimuli.

### 2.2.5 Learn by Reproduction

Despite the fantastic insight offered by these researches, our understanding of the brain remains very superficial. We have little knowledge of how we store information, or how the brain rearranges itself so that little information is lost in the process of neuron destruction, which is constantly caused by consumed substances like alcohol.

In this section we try an offer a perspective on the way we evaluate our knowledge and memory. To the best of our knowledge, this perspective is not used in the computer science or neuroscience domain, mainly because there is no experimental result to prove its validity. For this reason, this section will maintain a speculative nature. Nevertheless, we do find a striking similarity with a recent development in machine learning described in section 2.3.5.

Think of the way we try to study highly theoretical materials, like an historical paper. Our ultimate proof of understanding remain our ability to answer important questions on the material. However, such set of question may and usually is not available. Furthermore, it is difficult to create questions that cover the entire material of a certain text. This is highlighted even more in topics or subjects that require deep understanding, like physics or mathematics. For this reasons, while the ability to answer questions is the ultimate proof of understanding of specific parts in certain materials, we have to find alternative techniques to evaluate our progress. To overcome this problem, in a natural way, we do evaluate ourselves in our ability to reconstruct the material with as little loss as possible. For example, in a historical paper we evaluate ourselves by being able to renumerate all the facts and dates, or in a mathematical formula by being able to derive it ourselves from scratch.

## 2.3 THE ARTIFICIAL BRAIN

Having a view on how the biological brain works it is time to turn our focus to the artificial counterpart. In this section we introduce important algorithms that have been invented over the years inspired by important discoveries in the functioning of the Bi-

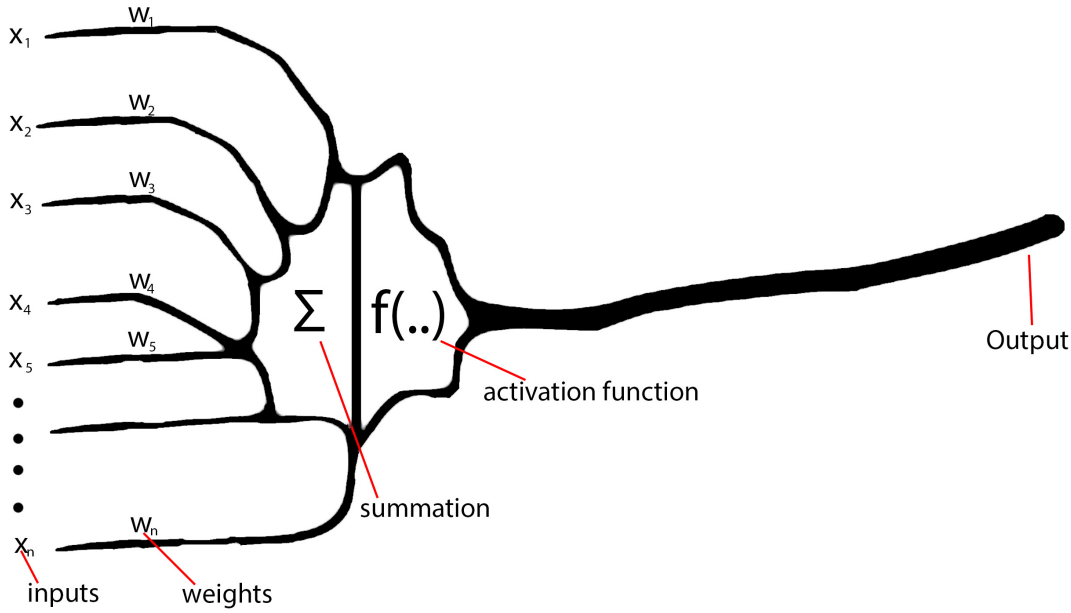



**Figure 3.1:** Sketch of an artificial neuron.

ological Brain. To make our intended analogy very clear, this section will be structured identically to the previous section on the Biological Brain, each subsection having its counterpart.

### 2.3.1 McCulloch and Pitts Neuron

The McCulloch and Pitts neuron (McCulloch and Pitts, 1943) is among the very first works in Artificial Intelligence (AI). Inspired by the functioning of biological neurons, they proposed a binary activation unit. As depicted in Figure 3.1, the McCulloch and Pitts neuron takes inputs, multiplies them with some weights (different for each input) and sums up the result. This final sum passes through a threshold activation function which fires if the sum is above a certain threshold. By adapting the weights and the threshold the neuron was shown to fire to different and interesting patters. For example logic OR and logic AND circuits could be implemented with such neurons [3].

$$\hat{y} = f\left(\sum_{i=0}^{n} w_i x_i\right) \tag{2.1}$$

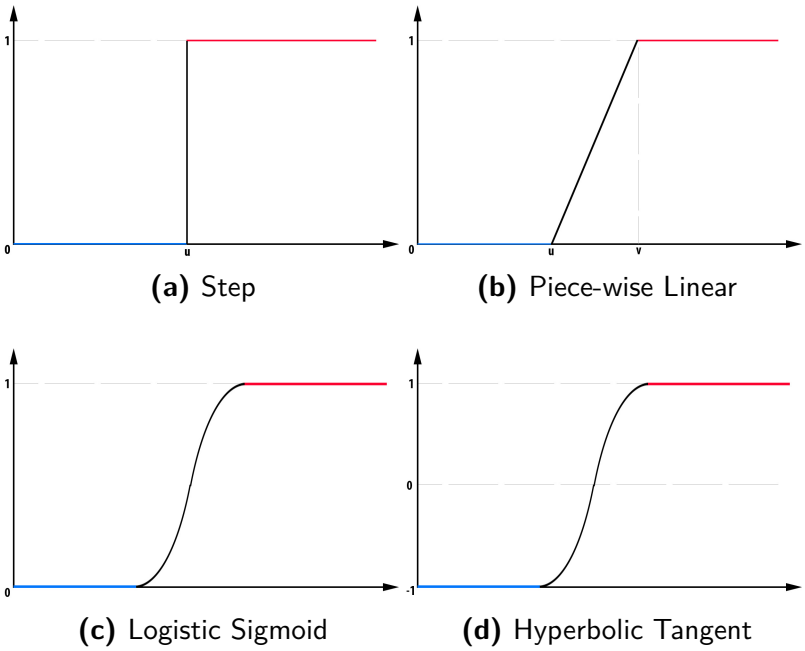
The output of the neuron is calculated as in equation 2.1, where $f$ is the activation function. Research in the field has led to the use of many activation functions. The equations of the most commonly used ones are listed in Figure 3.2. Their main differences are in the shape, range of values and differentiability.

[3] Not with the same neuron

$$step(x) = \begin{cases} 0 & x \geq \tilde{u} \\ 1 & x < \tilde{u} \end{cases}; \qquad pwl(x) = \begin{cases} 0 & x < \tilde{u} \\ f(x) & \tilde{u} < x < \tilde{v} \\ 1 & x \geq \tilde{v} \end{cases};$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}}; \qquad tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

**Figure 3.2:** Equations of activation functions




**(a)** Step **(b)** Piece-wise Linear

**(c)** Logistic Sigmoid **(d)** Hyperbolic Tangent

**Figure 3.3:** Graphs of activation functions.

If we see the graph of such functions in Figure 3.3, their differences become clear. The step function does produce a zero-one output which can be interpreted as firing (1) and not-firing (0). Nevertheless biological neurons, have different intensities of excitation. This problem is solved by the piece-wise linear (pwl) function which is similar to the step function, but instead of a direct change has a positively sloped part in which the excitation of the neuron increases smoothly from 0 up to 1. Both these functions include kinks at $\hat{u}$ for the step function and at $\hat{u}$ and $\hat{v}$ for the *pwl*. At these points the functions become non-differentiable, and as we will see further down this section differentiability is an important feature of activation functions. To overcome this limitation, the logistic sigmoid function was introduced. Apart from being differentiable in its whole domain, it has interesting features which will become important when we move to algorithms to train this neuron from data. The hyperbolic tangent (tanh) on the other hand is very similar to the logistic sigmoid with the difference of

the range. While the logistic sigmoid ranges from 0 to 1, the hyperbolic tangent ranges from -1 to 1. In section 2.3.2 we will see that this a very important property as outputs of many activation functions have to be multiplied together. In the logistic sigmoid, being all smaller than zero, the final result risks falling below the minimum binary representable number in a modern machine. On the other hand, the hyperbolic tangent, has a bigger range, thus less values close to zero so this risk is reduced.

The ability of the neuron to fire[4] to patters it recognises and not-fire to others has elicited its extensive use in pattern recognition. Furthermore, when presented with a dataset of two classes of objects the firing is mutually exclusive. Firing to one pattern means not firing to the other, thus the neuron can discriminate between the two classes. This gives them the name discriminative models. In most of the literature, and likewise in most of this thesis, the neurons will be used in the context of discriminating between two classes, with a decision boundary.

An eagle eyed reader, could have noted in equation 2.1 that this decision boundary is linear. Replacing the sum $\sum_{i=0}^{n} w_i x_i$ with vector notation $w^T x$ we can define a sigmoid activation neuron $\hat{y} = \sigma(w^T x)$. As we previously described firing is considered a probability greater than $1/2$, which means the neuron splits the decision at $\sigma(wx) = 1/2$ or $w^T x = 0$. This clearly is a hyper-plane.

Given such a linear boundary, a bias term is important to shift the boundary a certain amount. For this reason, the effective equation usually is: $\hat{y} = f(\sum_{i=0}^{n} w_i x_i + w_0)$. However, the same effect can be achieved by adding a 1 vector to the data matrix $x$. This will force the $w$ matrix to include a term for the 1 vector effectively acting as a bias term. This way we can use the original equation without change. In this thesis, unless otherwise stated, we will assume all input data have the 1 vector appended and the bias term is included in the $w$ vector, for this reason $w_0$ will be omitted.

**The Perceptron Learning algorithm**

As previously described, the McCulloch and Pitts neuron is able recognise important patterns. Nevertheless, the human brain is not born knowing all the information it does. Instead it learns through experience. For this reason a method able to automatically update the weights of a neuron in order to recognise some data based on experience was required.

Rosenblatt (1958) introduced the perceptron learning algorithm, which given a set of of patterns $x$ associated with their real class (hereafter label) $y$ could train the neuron, such that it would correctly predict the class label. Rosenblatt assummed a step activation function of range [-1,1], but the same algorithm can easily be adopted to standard

[4]Here we assume that a firing neuron exceeds some basic level of excitation in the case of continuous activation function. The 0.5 and 0 boundaries are commonly used for the logistic sigmoid and the hyperbolic tangent respectively.

**Algorithm 1** The Perceptron Learning Algorithm

```
1: D ← (x_1, y_1), (x_2, y_2), ..., (x_N, y_N).
2: T ← Epochs (number of passes through the training data)
3: procedure PerceptronCriterion(D,T)
4:     W ← 0
5:     for j = 1 ... T do
6:         for i = 1 ... N do
7:             ŷ_i = sign(w^T x_i)
8:             if ŷ_i ≠ y_i then
9:                 w ← w + t_i x_i
10:            end if
11:        end for
12:    end for
13:    return w
14: end procedure
```

[0,1] step function. The algorithm looks for a vector $w$ such that for every pattern $x_n$ with class label $y_i = 1$ it will have $w^T x_n > 0$, while for patterns $x_n$ with class label $y_i = -1$ it will have $w^T x_n < 0$. Considering our target range [-1,1] it sums up that we would like the model to have $w^T x_n y_n > 0$ for all patterns. The perceptron associates zero error with correct classification, while for misclassified pattern $x_n$ it tries to minimize the quantity $-w_T x_n y_n$. This is called the perceptron criterion and its equation is as below:

$$E_p(w) = -\sum_{n \in M} w_T x_n y_n \tag{2.2}$$

The variable $M$ defines the list of misclassified points. Minimizing this function we are actually maximizing the number of correctly classified patterns. To do so, we can follow the direction of the gradient of the function with respect to the variable we are optimizing, in our case the weights $w$, in a process called *Stochastic Gradient Descent*. More on this in section 2.4.1.

The gradient of the perceptron criterion (assuming $M$ is fixed) is

$$\frac{\delta L_p}{\delta w} = -\sum_{n \in M} y_i x_i \tag{2.3}$$

However, the perceptron algorithm processes misclassified examples one at a time, thus the sum is dropped and the update rule becomes:

$$w \leftarrow w + y_i x_i \tag{2.4}$$

The perceptron algorithm can be easily interpreted in the following way: go over all the training data patterns and classify them according to the current set of weights. Confront the prediction to the real classification as defined by the label. If the model is

correct move to the next data point, otherwise add if positive or subtract if negative the data vector to the weights. A pseudocode implementation of the Perceptron is given in Algorith 1.

**Logistic Regression**

If the step activation function used by Rosenblatt is replaced with the logistic sigmoid, the neuron exhibits some really nice properties. It output ranges between 0 and 1, the sum from $-\infty$ to $+\infty$ equals to 1 and it is differentiable in its whole domain. The first two properties mean that we can easily interpret the output of such neuron as probabilities. In other words, for a given data input, the output of the neuron is the probability of that data belonging to the pattern the neuron recognises.

$$P(C_1|x) = \sigma\left(\sum_{i=0}^{n} w_i x_i\right) = \sigma(w^T x) \tag{2.5}$$

Given this probabilistic definition, to better fit the observations we can maximize the conditional log-likelihood of the data.

$$LL(w) = log \prod_n p(y_n|x_n) \tag{2.6a}$$

$$LL(w) = \sum_n log\left\{p(C_1|x_n)^{y_n} p(C_2|x_n)^{1-y_n}\right\} \tag{2.6b}$$

$$LL(w) = \sum_n log\left\{p(C_1|x_n)^{y_n} \left(1 - p(C_1|x_n)\right)^{1-y_n}\right\} \tag{2.6c}$$

$$LL(w) = \sum_n log\left\{\sigma(w^T x)^{y_n} \left(1 - \sigma(w^T x)\right)^{1-y_n}\right\} \tag{2.6d}$$

$$LL(w) = \sum_n y_n log\, \sigma(w^T x) + (1 - y_n)(1 - \sigma(w^T x)) \tag{2.6e}$$

Nevertheless, in machine learning it is more common to minimize the loss of a model. We can easily do so as the inverse of the log-likelihood does in fact inform us about the discrepancy between the model predictions and the reality, hence it can be interpreted as a loss function. The negative log-likelihood is also known as the Cross-Entropy.

$$L(w) = -LL(w) = -\sum_n y_n log\, \sigma(w^T w) + (1 - y_n)(1 - \sigma(w^T w)) \tag{2.7a}$$

The minimum of this function is intractable. Nevertheless it is differentiable and we

can optimize it via gradient descent the same way we did with the perceptron criterion.

$$\frac{\delta L(w)}{\delta w} = \sum_n \sigma(wx_n)x_n - y_n x_n \tag{2.8}$$

A pseusdocode implementation of the algorithms in given in Algorithm 2.

**Algorithm 2** Logistic Regression Training Algorithm

```
D ← (x_1, y_1), (x_2, y_2), ..., (x_N, y_N).
T ← Epochs (number of passes through the training data)
procedure LogIt(D,T)
    W ← 0
    u ← Σ_n t_n x_n
    while True do
        g ← −u
        for i = 1 ... N do
            g ← g + σ(w^T x_n)x
        end for
        w ← w − ng
        if |g| ≈ 0 then
            Break
        end if
    end while
    return w
end procedure
```

**Papert and Minsky Contraversy**

Despite the nice properties of the artificial neurons and the ability to learn by experience, funding in neural computation was halted for almost two decades. The reason for this is widely considered the publication of the book *Perceptron: an introduction to computational geometry* (Minsky and Papert, 1969) in 1969. Minsky, an old acquaintance of Rosenblatt from their time in the Bronx High School of Science, had cought on the fact that these neurons can only learn and represent linear decision boundaries. They presented this limitation as a fundamental flaw, showing that these neurons can not learn even simple patterns like the exclusive OR.

### 2.3.2 Neural Networks

Papert and Minsky did not include in their conclusion networks formed by the interconnection of such neurons in layers. The exclusive OR can be represented with a circuit of OR, AND and NOT, all of which are linearly separable and can in fact be represented by neurons. For this reason a network of neurons can represent the XOR function. Analogically to the brain, the artificial neurons can be connected to overcome the

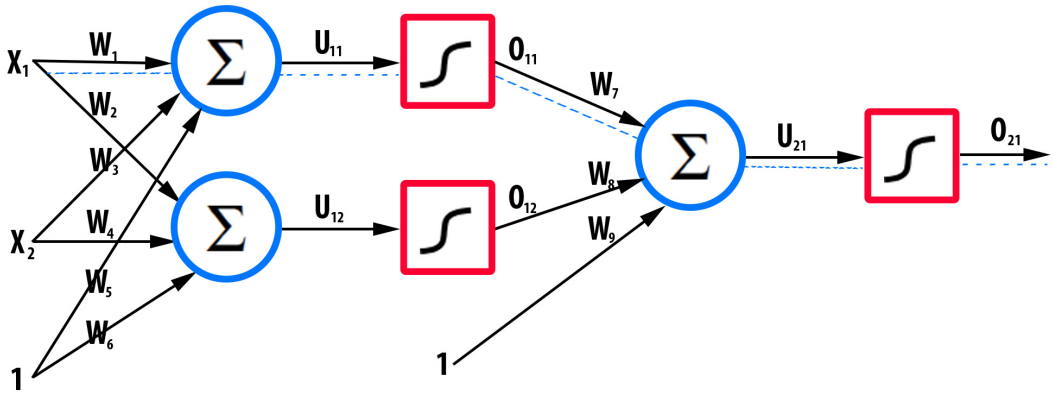



**Figure 3.4:** Neural Net.

above limitations. In fact it is now proven that multilayer networks of neurons called neural nets or multilayer perceptrons can approximate any function with arbitrary accuracy (Cybenko, 1989, Hornik et al., 1989).

A simple feed-forward neural network, which is the focus of this work, is depicted in figure 3.4. We have departed a little from the biologically inspired look of figure 3.1, nevertheless we hope its clear and easy to define the boundaries of each neuron in this network. A neuron is defined by the arrows to the blue circle (the weights), the blue circle (the summation) and the red square (the activation function or non-linearity). So this figure, includes a neural network with three neuron and one hidden layer. A hidden layer defines a layer of neurons that are not directly responsible of the output of the network.

**The Backpropagation Learning Algorithm**

Research in nerual networks resurged in late $80^s$ when Rumelhart et al. (1985) reinvented the Error Backpropagation method (hereafter simply back-propagation) to train these networks. In fact Russell and Norvig (1995) refer to Bryson et al. (1979) as the original invention of the backpropagation. Nevertheless, according to Rojas (1996) the idea could be traced as back as 1943 when Courant et al. (1943) proposed to use gradient descent along the Euler expression to approximate variational problems.

The idea behind the backpropagation is relatively simple. Lets consider the simple feed-forward neural network in figure 3.4 and lets assume the activation is a logistic sigmoid. As we saw in Logistic Regression, we can train the network via gradient descent. However, this is more complicated in networks with hidden layers, as some layers do not directly see the input or output, or even both in networks with more than one hidden layer. Nevertheless, given the logistic sigmoid is a continuous function, the output of the network, in our case $O_{21}$ is a continuous function of the inputs ($X_1$, $X_2$ and Bias).

One can actually write the whole network in figure 3.4 the following function

$$\hat{y} = \sigma\left\{\sigma(x_1 w_1 + x_2 w_3 + w_5) * w_7 + \sigma(x_1 w_2 + x_2 w_4 + w_5) * w_8 + w_9\right\} \tag{2.9}$$

To update each weight in a similar way we did for Logistic Regression, we need to find the gradient of the function above with respect to that weight. This is nothing more than applying the Chain Rule of Calculus. However, having the possibility of drawing the network with a nice graphical representation the back-propagation algorithm is in fact the Chain Rule with a graphical UI.

For example, to find the gradient for $W_1$ we follow the path from $W_1$ to the output as depicted by the dashed blue line.

$$\frac{\delta L}{\delta w_1} = \frac{\delta L}{\delta O_{21}} \frac{\delta O_{21}}{\delta U_{21}} \frac{\delta U_{21}}{\delta O_{11}} \frac{\delta O_{11}}{\delta U_{11}} \frac{\delta U_{11}}{\delta w_1} \tag{2.10}$$

Once we have the gradient, we can optimize the weights in the same way we did for Logistic Regression.

### 2.3.3 Convolution Neural Network

Inspired by the hierarchical structure in the brain machine learning researchers aimed at bringing this advancement into artificial neural networks. A first attempts is the Neocognitron (Fukushima, 1980) that could adapt the connections on the first layer of neurons. Nevertheless, the architecture that successfully brought the feature detectors to the artificial neural network domain is the Convolutional Neural Network (CNN) of LeCun et al. (1998).

In a conventional Neural Network architecture, the neurons in the first layer span over all the input. If we consider the image domain, the neurons in the first layer will have connections with all the pixels of the image. This means they will fire if they detect features over the whole image. However, weights are related to a certain pixel location, for example $w_{1,1}$ would correspond to pixel $I_{1,1}$. For this reason an image that would excite a neuron, would not excite the same neuron if shifted a few pixels around. This is contradictory to the experimental results on the cat's striate cortex. In the biological brain, lower level cells have a very small receptive field which increases with the hierarchical order of the cell in the brain. Furthermore, the cortical cells would respond to a shape independent of its location in the visual field.

The Convolutional Neural Network reconciles these differences by organizing the neurons to see a patch of the image as shown in the green rectangles in Figure 3.5. Each neuron, represented by the small circle on top of the rectangle, has weights only for the pixels in the rectangle. This means that neurons have a very small receptive field

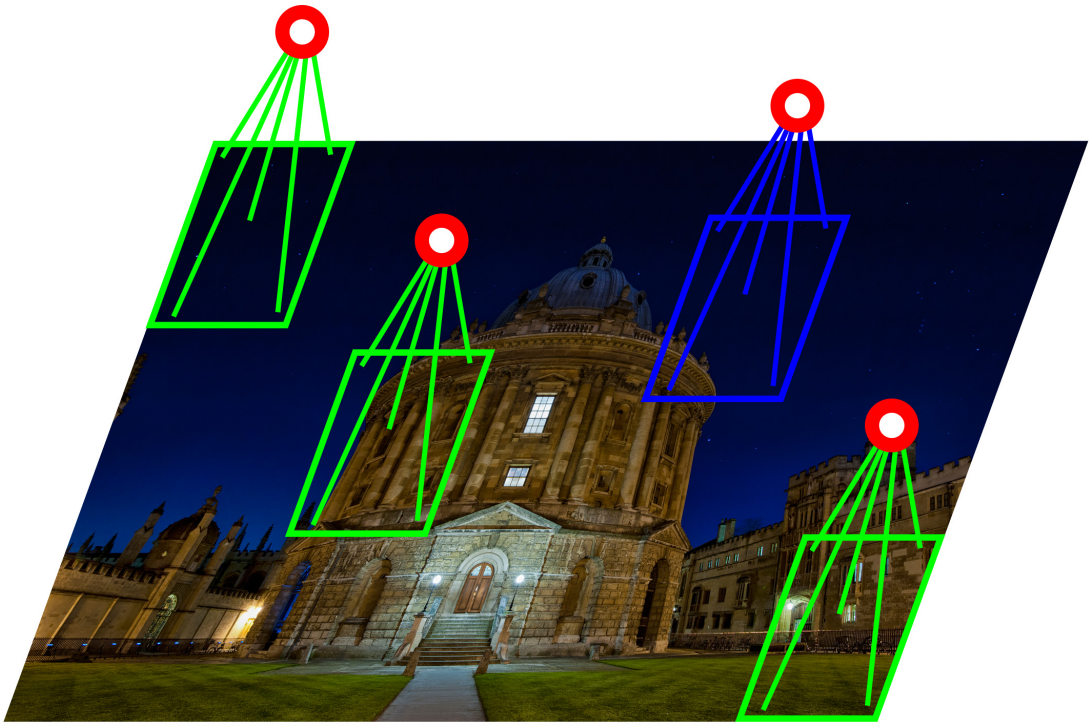


**Figure 3.5:** Convolutional Neural Network Approach.

and do detect features only within that field. Nevertheless, we would like to detect features independent of their location in the image. The CNN answers this by shifting the rectangle that corresponds to the neuron receptive field all over the image and storing the output of the excitation of the neuron over all possible locations. The vector that stores the output of the neuron's excitation all over the image is often refereed to as *feature map*, while the neuron as *feature detector, kernel or filter*. We will adopt the *feature detector* nomenclature. In the literature most often it written that the output are passed through a non-linearity before making it into the *feature map*. This oversees the fact that we are working with neurons not only its weights, and neurons do include the non-linearity as it is their activation function.

At this point, the remaining difference is the fact that cells in the brain have bigger visual field as they move up the hierarchy. The CNN overcomes this difference by applying a *Max-Pooling* operation over the feature maps. The *Max-Pooling* reduces the size of the feature map by only keeping the highest responses. Consider an edge that was detected by a feature detector only on the left-most and right-most side of the image but not anywhere else. The feature map will be a vector of the form [0.8,0.1,0.2,...,0.3,0.1,0.9]. Note that the vector includes a response for each location and as artificial neurons return probabilities they will rarely or never be 0 or 1 as the neurons will not be sure to have detected a feature or not. Nevertheless we see that in the first location, corresponding to the left-most side of the image the feature detector was 80% sure, while on the last location corresponding to the right-most side it was 90% sure. We can be

relatively sure the feature was present in these positions. If we now apply *Max-Pooling* by keeping only the two most responsive locations the return vector will be $[0.8, 0.9]$. The next layer of feature detectors applied on this pooled representation will then have a view of whole image from the left-most side to the right-most side, similarly to how higher level feature detectors have a bigger visual field in the biological brain. Positions the response of which that did not make it to the pooled representation are useless as they do not have a lower level feature and without lower-level features there can be no high level one. Without dots, there can be no line.

So far we have referred to a single feature detector, but we obviously want to be learning many different features. This can be easily achieved by applying the exact same operation with another feature detector. Starting the weights of the feature detectors randomly allows them to learn different representations. In the figure below, we show the features learned by the $1^{st}$, $2^{d}$, $3^{d}$ and $4^{th}$ level feature detectors trained on the IMAGENET, as visualised by Zeiler and Fergus (2012).

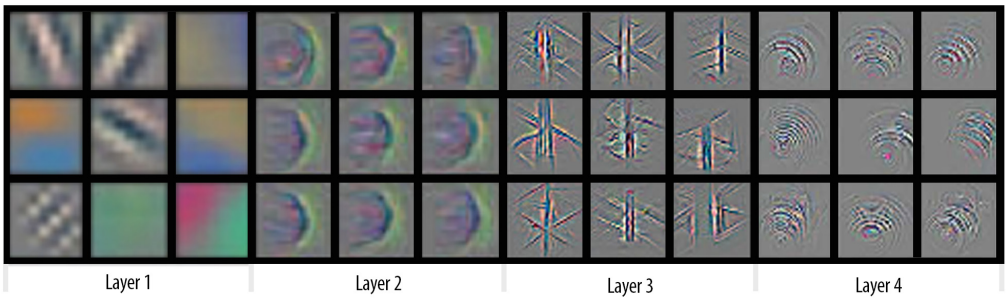



**Figure 3.6:** CNN learned features.

There is another difference between conventional Neural Networks and the CNN. While conventional Neural Nets apply a simple multiplication between the weights and the input, the CNN applies the convolution operation, hence the name. Yashua Bengio [5], has a very intuitive explanation to the convolution operation. He defines $x(t)$ to be the noisy reading of a ships position. To make the data less noisy an obvious choice is to make many readings. Nevertheless, the ship may move, so newer readings should take priority over older ones. To to this a weighting function $w(a)$ where $a$ is the age of the reading can be used. Applying the weighted avarage operation at every moment in time yields

$$s(t) = \int_{\infty}^{\infty} x(a)w(t-a)\,\mathrm{d}a \tag{2.11}$$

which in fact is the convolution operation denoted as below.

$$s(t) = (x \otimes w)(t) \tag{2.12}$$

[5] http://www.iro.umontreal.ca/ bengioy/dlbook/convnets.html

In other words, the convolution gives a bias to feature detectors to fire more when the feature is in the middle of the detector. However, one can easily notice that the convolution is an expensive operation. Nevertheless, the *Convolution Theorem*, states that convolution in the time domain is equivalent to point-wise multiplication in frequency domain. For this reason, a fast approach to the convolution operation is to convert both functions to Fourier domain via Fast Fourier Transform, complete a point-wise multiplication and then apply an Inverse Fourier Transform to get the final result. The identity is as below:

$$s(t) = (x \otimes w)(t) = IFT\left\{FT(x * t)\right\}(t) \tag{2.13}$$

Apart from sharing great similarity with the biological process, CNNs also have great properties in terms of machine learning. By having feature detectors smaller than the input, CNNs have sparse weights. This means that the number of parameters for a CNN is substantially smaller than the number of parameters of a conventional NN. Furthermore, being shifted around an image, in each training image feature detectors see many smaller one, thus increasing the training set.

### 2.3.4 The One Learning Algorithm

At the current state in machine learning and AI, fundamentally different algorithms hold the state-of-the-art in different tasks. Nevertheless, the adaptive capabilities showed by biological cortices inspires us to believe that there exists one algorithms that can be great in all tasks. This belief is dubbed under the name of "The One Learning Algorithm Hypothesis".

### 2.3.5 Autoencoders and Restricted Boltzman Machines

We dis see so far many algorithms that do update the parameters of a network in order to fit observations. These models rely on a acceptable truth to which they compared their prediction, this was the label. However, labels are not always available. Is all hope lost for not labelled data?

It turns out it is not. Much in the same way we suggested humans do study theoretical materials in section 2.2.5 machines can too: by trying to reconstruct from memory the information they were given.

#### Autoencoders

The Autoencoder achieves this goal in a very simple and intuitive approach. It setups a Neural Network which instead of predicting a label for the output, it predicts its input. A simple 1 hidden layer Autoencoder is drawn in Figure 3.7. Obviously the output of

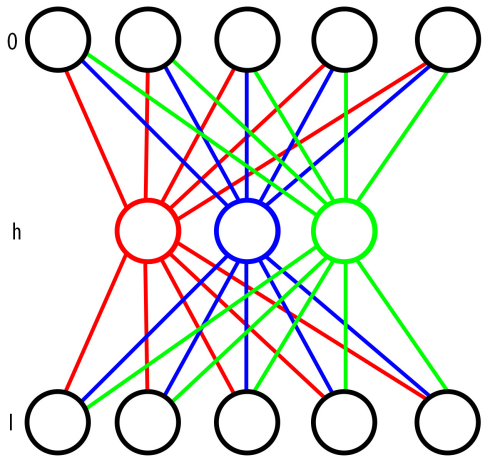



**Figure 3.7:** Autoencoder.

neurons is a probability distribution that ranges from 0 to 1, for this reason the input should be encoded in that range prior to being input into an Autoencoder with neurons that use sigmoid activation. Furthermore the output is a reconstruction of the input instead of a probability distribution over two classes, for this reason the Cross Entropy cannot be used. A common choice of cost function is the Mean-Squared Error (MSE), which given network input $I$ and network output $O$ is defined as below:

$$MSE = \frac{1}{2}(O_i - I_i)^2 \tag{2.14}$$

The Autoencoders have another nice property. The same weights that are used to encode the data to the hidden representation can also be used to decode it. Equation 2.15 shows the multiplication of weight matrix $W$ with input $I$ to encode the input to the hidden representation $h$. Then Equation 2.16 shows how the transpose of weight matrix $W$ can be used to output the final prediction $O$. By training $W$ this way, we are learning weights that are both able to encode and decode the input. While it may not be obvious, these weights have shown to learn important features in an unsupervised approach.

$$\begin{bmatrix} I_0 & I_3 & I_2 & I_3 & I_4 \end{bmatrix} * \begin{bmatrix} w_{0,0} & w_{1,0} \\ w_{0,1} & w_{1,1} \\ w_{0,2} & w_{1,2} \\ w_{0,3} & w_{1,3} \\ w_{0,4} & w_{1,4} \end{bmatrix} = \begin{bmatrix} h_0 & h_1 \end{bmatrix} \tag{2.15}$$

$$\begin{bmatrix} h_0 & h_1 \end{bmatrix} * \begin{bmatrix} w_{0,0} & w_{0,1} & w_{0,2} & w_{0,3} & w_{0,4} \\ w_{1,0} & w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} \end{bmatrix} = \begin{bmatrix} O_0 & O_1 & O_2 & O_3 & O_4 \end{bmatrix} \tag{2.16}$$

The network we drew in Figure 3.7 and its corresponding Equations 2.15 and 2.16, refereed to an Autoencoder with one single hidden layer. Nevertheless, there is nothing to prevent the hidden layer in itself to act as an input/output pair to another Autoencoder. By recursively replacing the hidden layer of an Autoencoder with another Autoencoder, we are creating a deeper and deeper model, where the input is further and further from the reconstruction. Considering Autoencoders are Neural Networks that train via backpropagation, the gradient in such long distance will be very small up to almost useless, preventing the model from learning.

The way we have introduced Deep Autoencoders, as recursively replacing the hidden representation of one Autoencoder with another Autoencoder, should have given a clear hint on an alternative way of training them. We can do so in a greedy layer-wise approach. For example, in a Deep Autoencoder with 3 hidder layer $(h_1, h_2, h_3)$ and input $I$, we can train an Autoencoder to reconstruct $I$ via $h_1$ $(I \rightarrow h_1 \rightarrow I)$, then use $h_1$ as input to a second Autoencoder that tries to reconstruct it via $h_2$ $(h_1 \rightarrow h_2 \rightarrow h_1)$ and so on. While this avoids the gradient problem, by being trained in a greedy way, neurons in one layer do not try to learn features that will make the work of higher levels easier. This is not particularly desirable when the aim of the process is hierarchical feature extraction.

**Restricted Boltzman Machines (RBM)**

Another unsupervised approach to feature extraction is via a generative model known as the Restricted Boltzman Machine (RBM). It was originally invented by Smolensky (1986) under the name Harmonium and then reinvented by Hinton (2002) who also proposed a fast training algorithm, the Contrastive Divergence. As the name suggests, RBMs are a version of Boltzman Machines (Ackley et al., 1985), which in turn are very closely related to Hopfield Networks. Nevertheless, they fit directly with the feed-forward only approach we have discussed in this thesis, so we will skip over their difference with original Boltzman Machines.

RBMs achieve a similar task to Autoencoders. Given an input, they learn a hidden representation from which they can infer the input. Nevertheless they are different. While Autoencoders learn to reconstruct their input, RBMs learn a probability distribution over their input and thus can probabilistically regenerate it. In fact, RBMs are a generative model, while Autoencoders are not, despite the fact they achieve a goal in the same mood.

Given a visible layer $(v)$ and a hidden layer $(h)$ the functioning of the RBM is governed by the following energy function $(E)$ to be minimized. The variables $(b)$ and $(c)$

correspond to the biases on the visible and hidden unit weights respectively.

$$E(v,h) = -\sum_{i=1}^{|h|}\sum_{j=1}^{|v|} w_{ij}h_i v_j - \sum_{j=1}^{|v|} b_j v_j - \sum_{i=1}^{|h|} c_i h_i \tag{2.17}$$

Based on this energy function the joint probability of the visible and hidden layer is given by the Gibbs distribution.

$$p(h,v|w) = \frac{1}{Z(w)} e^{-E(v,h)} \tag{2.18}$$

$$Z = \sum_{v}\sum_{h} e^{-E(v,h)} \tag{2.19}$$

Being a feed-forward network, the nodes in one layer are conditionally independent given the nodes in the other layer. This allows the posterior probability to factorize as below.

$$p(h|v,w) = \prod_{k} p(h_k|v,w) \tag{2.20}$$

$$p(v|h,w) = \prod_{k} p(v_k|h,w) \tag{2.21}$$

Despite this simplification, maximizing the likelihood of the model results in an intractable solution. For this reason maximum likelihood training is approximated via Blocked Gibbs Sampling (a Markov Chain Monte Carlo method) or faster via the Contrastive Divergence developed by Hinton and collaborators.

## 2.4 MACHINE LEARNING

We hope we have shown, over the previous sections, that great progress has been made in biologically inspired algorithms. However, there is still a difference in the technical aspect of learning, between the brain and the machines. Furthermore, we do not think machines should blindly follow the brain. While it is a great guidance, there are a lot of operations that machines can compute, which are impossible in the biological process.

In this section, we will explain two important developments aimed at optimizing the learning process. The are used extensively in machine learning and as a consequence also in this thesis.

### 2.4.1 Gradient Methods

All the algorithms we have discussed so far have a fundamental similarity. They define a cost function and minimize it with respect to the parameters of the models, the

weights. This cost function can rarely be solved in an analytical way. For this reason we have to rely on numerical optimization algorithms.

**Gradient Descent**

The simplest algorithms to find the minimum of a strictly convex multidimensional surface is called *Gradient Descent* or *Steepest Descent*. To find the minimum of a function via gradient descent, one takes a step proportional to the negative of the gradient at the current point. There is a very intuitive reasoning behind gradient descent. The gradient of a function shows which way is uphill. Because we want to go downhill we follow the opposite direction. The final update rule, where $x_n$ defines the value $x$ at iteration $n$ and $r$ defines the learning rate, becomes:

$$x_{n+1} = x_n - r\frac{\delta f(x)}{\delta x} \tag{2.22}$$

**Newton-Raphson Method**

Nevertheless, gradient descent can be slow to converge. Furthermore, adjusting the learning rate can be tricky and difficult. An improved approach is what is known as the Newton-Raphson Method. If the function is differentiable, the Newton-Raphson method says that we can approximate it with the second-order Tylor expansion in the neighbourhood of current point $x_n$.

$$f(x_n) \approx f(x+\varepsilon) = f(x) + f'(x)\varepsilon + \frac{1}{2}f''(x)\varepsilon^2 + \ldots \approx f(x) + f'(x)\varepsilon + \frac{1}{2}f''(x)\varepsilon^2 \tag{2.23}$$

By definition, at the minimum or the maximum the derivative will be zero, hence

$$f'(x) + f''(x)\varepsilon = 0 \tag{2.24}$$

$$\varepsilon = \frac{f'(x)}{f''(x)} \tag{2.25}$$

The update rule, then simply becomes

$$x_{n+1} = x_n - \frac{f'(x)}{f''(x)} \tag{2.26}$$

Despite the fact that it may sound complicated, the Newton-Raphson method has a very nice and simple logic behind it. By taking the second-order Tylor expansion, we are approximating the surface of our function with a quadratic one. The minimum to a quadratic function can be found in an analytical way, so we do solve it and jump

straight to its minimum. Nevertheless, the Newton-Raphson method uses the second derivative of the function. In multidimensional problems, that will require the computation of the Hessian (the matrix of partial derivatives), which is a very expensive operation. To overcome this there exist methods which try to approximate the Hessian by not computing it entirely. These methods are called Quasi-Newtonian.

### Complications

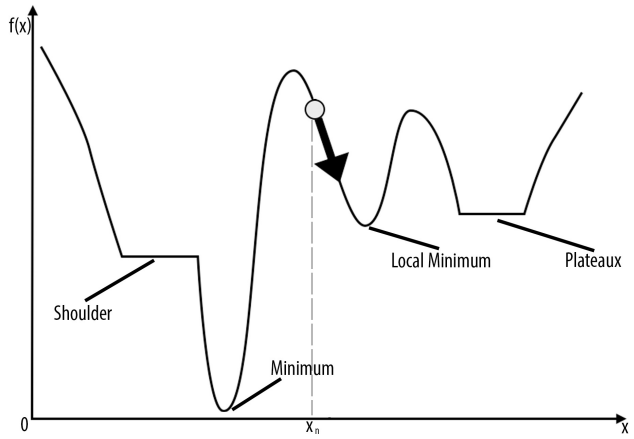



**Figure 4.1:** Surface of a Function.

Both first-order methods (Gradient Descent) and second-order methods (Newton-Raphson and Quasi-Newtonian) require the function to be convex in order to find the minimum. In reality, cost functions are rarely convex, which means both methods will face complications and in many cases fail find the global minimum.

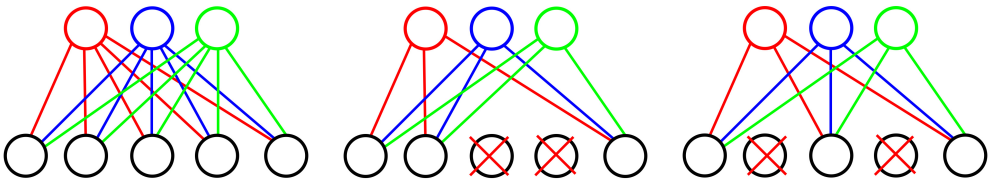
The situations where this happens are depicted in Figure 4.1. If the optimization technique starts at $x_n$, it will follow the gradient to the local minimum. At that point the gradient will be zero, so the update rule of both gradient descent and Newton-Raphson will degenerate to $x_{n+1} = x_n$ and we will be stuck. The same situation will happen if the current $x_n$ ends in a plateaux or a shoulder. Some techniques on how to avoid local minima exist. An example is the Simulated Annealing, which by analogy to annealing in metallurgy, allows upward movements with a probability that decreases as a function of the iterations. Nevertheless, in a multidimensional surface avoiding local minima becomes increasingly difficult, for this reason most of the time we will settle for a local solution.

The exact optimization algorithm used in this paper is the ADAGRAD of Duchi et al. (2011).

### 2.4.2 Hinton's Dropout

Another important Neural Network optimization technique we do use in this paper is the Dropout introduced by Hinton et al. (2012).

What Hinton noticed, and what happens regularly in Neural Networks, is the fact that they achieve extremely high accuracy on the training set, but see a tremendous drop in the test set. This is not restricted to Neural Networks, but instead a well known problem in machine learning called *Overfitting*. Hinton suggests, that in Neural Networks, this happens because neurons learn to work together. That is, only fire if some other neurons also fire. He calls this effect *Complex Co-Adaptation*. To prevent this co-adaptation, Hinton suggests to force some neurons to not fire during the training. This will have the effect of forcing neurons to learn to be useful by themselves instead of relying on other neurons.



**(a)** Original NN **(b)** Subsample 1 **(c)** Subsample 2

**Figure 4.2:** Dropout.

Nevertheless, the Dropout can be interpreted in many different ways. For example, one can see that by forcing some neurons not to fire, we are actually adding some noise to the training process. While noise is not desirable, it can help prevent overfitting. Another view, offered by Hinton, suggests that by forcing some neurons not to fire, we are subsampling a smaller Neural Network from original setup. Making a random choice each time, we are randomly selecting a different sub-network. This is depicted in Figure 4.2. Each sub-network is created by dropping all the neurons in the first layer with probability 0.5. In test-time we would like to use the full network. To do so we can take geometric mean of the prediction of all sub-networks. Hinton suggests a much more elegant solution. Instead of taking the geometric mean of all sub-networks, he suggests to divide the weights in the layer where dropout was used by 2 and use the full network. This actually computes the geometric mean of the predictions.

While the dropout ratio of 0.5 is commonly used, other values are also acceptable. Generalized, if we train a certain hidden layer of the network with Dropout with probability $r$, then in test time we multiply the weights with $\frac{1}{r}$.

## 2.5 NATURAL LANGUAGE PROCESSING

This thesis is about Deep Learning. Nevertheless, we apply it to the domain of Natural Language Processing (NLP). For this reason we believe it is useful to have an understanding of the challenges involved. Furthermore, by exploring current algorithms in the field together with their achievements and limitations, we can better appreciate the advancements Deep Learning brings.

### 2.5.1 Syntatic Composition

The smallest building blocks of natural language, and in the same time the smallest meaning carrying units, are the *morphemes*. While there are free-standing morphemes that alone can be interpreted as words, usually they need to be paired with another morpheme or *affix* to form a word. In this process the meaning of a morpheme can be inflected in different ways. For example consider the free-standing morpheme 'long'. It has a very clear meaning. Nevertheless, we can inflect this meaning by adding different affixes like the *prefix* 'pro-' or the *suffixes* '-er' or '-est'. These would form the words 'prolong', 'longer' and 'longest'. While their meaning is closely related to the meaning of the morpheme 'long' it is slightly inflected. 'Longer' is the comparative of long, meaning that something has the meaning of 'long' in a higher scale than something else. 'Longest' is the superlative of long, meaning that something has the meaning of 'long' in the highest possible scale. While 'prolong' means making something longer. The compositional process in which morphemes join with affixes to form words, is studied by the field of **Morphology**.

While morphological inflections allows us easily infer the meaning of a words we have never seen before, for all other purposes we can consider them as entirely separate words. For this reason, in Natural Language Processing we can skip over the morphological process and move straight into analysing how independent words compose into sentences. The study of this compositional process is called **Syntax**. If follows form the ancient Greek work sýntaxis that means coordination or arrangement.

In natural language, the role a certain word plays in a sentence defines how it can be arranged with other words. In NLP we call this role the Part-of-Speech. For this reason, associating each word with its part-of-speech is an initial and very important task known as POS Tagging. For the brief introduction of this section we are going to use only three POS Tags, namely **N**(Noun), **V**(Verb) and **Det**(Determiner). Table 5.1 explains their meaning as in Jurafsky and Martin (2008).

Before forming sentences, words join into groups that have a specific meaning and a function which often can be replaced by that of a single word. These groups are called **constituents**. Here we are only going to discuss two of them, namely **NP**(Noun

| | |
|---|---|
| **N**(Noun): | Are words that refer to people, places or things. |
| **V**(Verb): | Are words that refer to actions or processes. |
| **Det**(Determiner): | Are words that usually precede nouns. A subclass are articles of which English has three (a, an, the). |

**Table 5.1:** POS Tags.

Phrases) and **VP**(Verb Phrases). A group forms a noun phrase, if it can be replaced with a single noun, while it forms a verb phrase if it can be replaced by a single verb. Lets consider as an example the sentence *"The free-falling meteorite hit the ground at tremendous speed."*. The part *"the free-falling meteorite"* is a noun phrase as replacing it with a single noun still results in a valid sentence. *"Jack hit the ground at tremendous speed."* The same way, *"hit the ground at tremendous speed"* is a verb phrase as if we replace it with a single verb, it forms a valid sentence. *"The free-falling meteorite crashed."*

**Context Free Grammar (CFG)**

The first proposed formalism that tries to capture the complex compositional process in natural language was the Context Free Grammar (CFG) of Chomsky (1956). Later it was also independently invented by Backus (1959). The CFG contains a set of rules which express how words can group together to form constituents and later sentences. For example consider the following three rules:

$$\begin{aligned} \text{S} &\rightarrow \text{NP VP} \\ \text{NP} &\rightarrow \text{Det N} \\ \text{VP} &\rightarrow \text{V NP} \end{aligned}$$

By recursively following the derivation one can see that a sentence can be formed S $\rightarrow$ Det N V Det N. Replacing the POS tags with words having that function we can get a grammatically sound sentence. The whole derivation process is depicted in Figure 5.1, using and example sentence from Chomsky's original work. This picture is known as a parse tree.

In natural language the same word can have different meanings and attach to other words differently strongly altering the meaning of the sentence. Consider the sentence, *"They can fish."*. If fish is considered a noun, then the sentence means that they are putting fish into cans. If fish, fish on the other hand is considered a verb, the sentence means that they are able to fish. This is depicted in the parse trees below.

While for a machine both parse trees are equally likely, as humans we are almost certain that the sentence has the second meaning. This complication can be overcome by extending the CFG to associate a probability with each CFG rule. This formalism

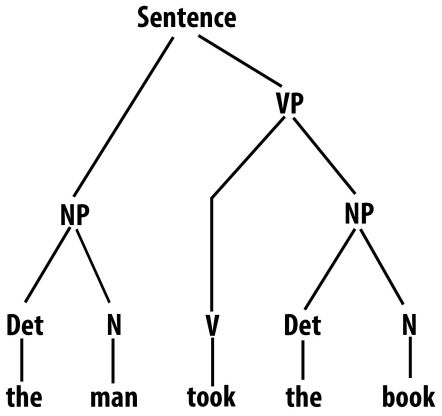



**Figure 5.1:** CFG.

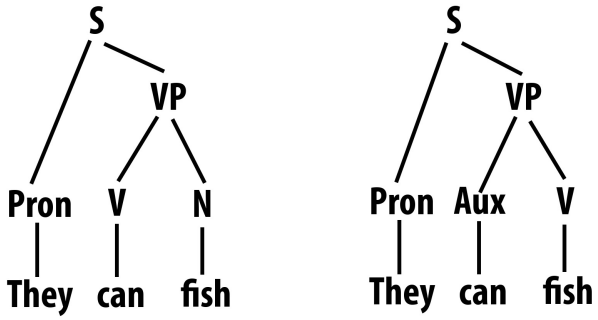



**Figure 5.2:** Ambiguous Sentence.

is known as Probabilistic CFG (PCFG). Applying a production rule of the PCFG, the probability informs how likely it is to happen. By multiplying the probabilities of all rules we can assign a probability to the whole parse tree, and thus easily chose the one that is more likely. For example, if the rule (VP $\rightarrow$ Aux V) has a higher probability then the rule (VP $\rightarrow$ V N) then the first parse tree will have a higher probability (all other rules are identical) and thus be preferred. The probabilities of production rules can be inferred from corpus of hand-annotated data like the Penn Treebank (Marcus et al., 1993). For example, if in the data the (VP) converted to (Aux V) seventy times and to (V N) thirty times, then the probability of the (VP $\rightarrow$ Aux V) rule is 70%.

Nevertheless, PCFGs make a very strong independence assumption. They make a decision based on the POS tag independent from the actual word. Collins (1997) noted that different words have different preferences despite the fact that they have the same part-of-speech tag. For example, the verb *'walked'* is very likely to be followed by a propositional phrase *(John walked into the bar)*, while a verb like *'saw'* is very unlikely to be followed by a propositional phrase. Instead it usually expects a noun phrase like *(John saw the man.)* To overcome this problem the Collins Parser (Collins, 1997) and

the Charniak Parser (Charniak, 1997) associate lexical items (words) with rules. The resulting formalism is called a Lexicalized PCFG (LPCFG). In an LPCFG, rules are effectively converted from simple (VP $\rightarrow$ Aux V) to (VP(fish) $\rightarrow$ Aux(can) V(fish)). While to define the lexical item of a POS tag we can easily and unambigously chose the word it is associated with, this becomes more complicated with constituents. For instance, in the example above we could have chosen both VP(fish) and VP(can). To guide the choice in these scenarios, parsers have well defined heuristics. For example a VP takes the lexical item of the V.

Defining the rules this way tremendously changes the counts used to define the probabilities. For example if we initially had 20 non-terminals in our grammar and 30'000 words in our vocabulary, after lexicalization our grammar will contain close to 600'000 thousand non-terminals [6]. However, from a Theoretical Computer Science Perspective VP(fish) is just another variable so one can consider the LPCFG simply a PCFG with extra variables.

### 2.5.2 Semantics

Parse trees capture the compositional process. To further extend this in order to convert natural language sentences to a format machines can understand, PCFG rules are associated with a Lambda Calculus expressions. The parse tree guides how Lambda Calculus expressions operate to encode the meaning of a sentence. Figure 5.3 illustrates an example where the natural language sentence *"John likes Mary"* is effectively converted to the predicate *"likes(john,mary)"*.

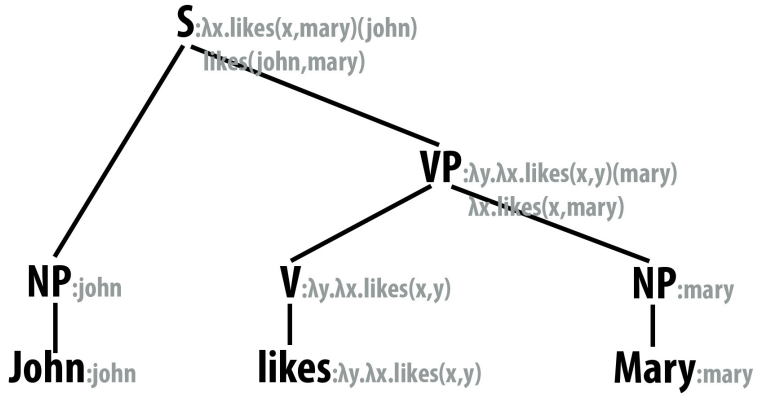



**Figure 5.3:** Natural Language sentence to Lambda Calculus.

While this approach is certainly inspiring, it has serious limitations. *First*, we rely on perfectly creating the parse tree of the sentence in order to encode the Lambda Calculus

[6] This estimation assumes each word will appear with each tag. While it overestimates a bit, it gives an idea of the non-terminal explosion.

expression. In reality even the best parsers like Collins (1997) achieve an accuracy lower than 90%. Furthermore this accuracy is counted on the number of constituents that are correctly parsed, not on the number of sentences. For example if a sentence has 10 constituents, we can correctly parse 9 and it will give us 90% accuracy. Nevertheless we would not have parsed the whole sentence which is required to build the Lambda Calculus encoding. *Second*, words do not have logic predicates assigned naturally, thus we have to manually do that. This falls into the domain of **Knowledge Representation and Reasoning** and **Ontological Engineering** where it is generally accepted that we can only build a logic base for domains of limited size. *Third*, even if we assume we have the predicates for every word in the language and furthermore we assume we can parse sentences with 100% accuracy, we are still fundamentally limited. While there exist algorithms, usually based on resolution, that can do inference on logical expressions very fast, the problem is NP-Complete even for the very simple boolean satisfiability.

With the introduction of these limitations in the traditional approach of mapping natural language to machine readable format, we invite you to the next part of this thesis where we tackle this problem with the Deep Learning approach.

# Part II

# Methodology

*Machines take me by surprise with great frequency.*

A.Turing

# 3

# Convolutional Sentence Model

In this chapter, we start introducing our own contributions. As described previously, Convolutional Neural Networks have had a tremendous success in computer vision, being state-of-the-art in many image recognition tasks like MNIST and IMAGENET. Furthermore initial work on Natural Language Processing (Collobert and Weston, 2008, Collobert et al., 2011, Kalchbrenner et al., 2014) has also shown encouraging signs in this direction.

In this section we present the Convolutional Sentence Model (CSM), which uses a Convolutional Neural Network that can capture the compositional process that maps the meaning of words to that of sentences by embedding them in a low dimensional vector space. The CSM is very similar to the DCNN (Kalchbrenner et al., 2014) developed by the Computational Linguistic Lab at the University of Oxford. In fact our initial aim was to use the DCNN as it is, however given the very large number of parameters, we found it hard to extend it to the more complicated problems we will be tackling in future chapters. Furthermore, given the large size of the model, we could not fit it into GPU in order to benefit from increased computation speed. For this reason, we developed the CSM which substantially reduces the number of parameters of the model, consequently its size and its training time. Furthermore, we could also easily fit the CSM into a GPU.

The aim of the CSM, and likewise of the DCNN on which it was inspired, is to capture the compositional process which maps the meaning of words into that of sentences. The CSM is able to represent the meaning of natural language sentences by embedding

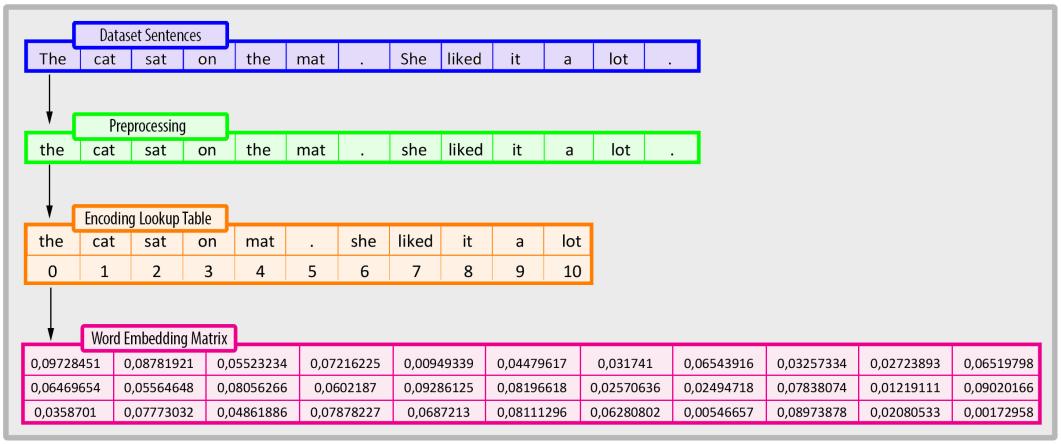



**Figure 1.1:** Initialization of the Encoding Lookup Table and the Embedding Matrix.

them in low dimensional vector space which captures important syntactic and semantic properties with no hand engineered features.

## 3.1 THE MODEL

In the following sections we will describe the layers and operations that define our model.

### 3.1.1 Continuous Word Representation

Natural language is fundamentally symbolic, while neural networks operate on continuous input. For this reason, the very first layer of our model should be able to create a mapping from symbolic words to continuous distributed representations. One of the early works in this field is due to Hinton (1986). This task, however, has received a lot of attention, and several other methods exist for creating these mappings, (Mikolov et al., 2013, Mnih and Kavukcuoglu, 2013) Mikilov's Word2Vec being a particularly widely used model. However, the method used by Collobert and Weston (2008), Collobert et al. (2011), Kalchbrenner et al. (2014) has shown to be very successful and furthermore, it can be optimized via backpropagation in the same setup as the rest of the network. For this reason it will be the one we use in this model.

The initialization phase first preprocesses the sentences present in the dataset. In natural language a lot of preprocessing techniques are suggested, however inspired by the "One Learning Algorithm Hypothesis" we aim to do as little task specific tweaks to our model as we can. For this reason we limit ourselves to making all words lowercase and also replace all words that appear less then 5 times with 'UNKNOWN'. This is depicted in figure 1.1. In this figure our dataset, represented in the blue table, includes only two sentences: *"The cat sat on the mat."* and *"She liked it a lot."*. The green table

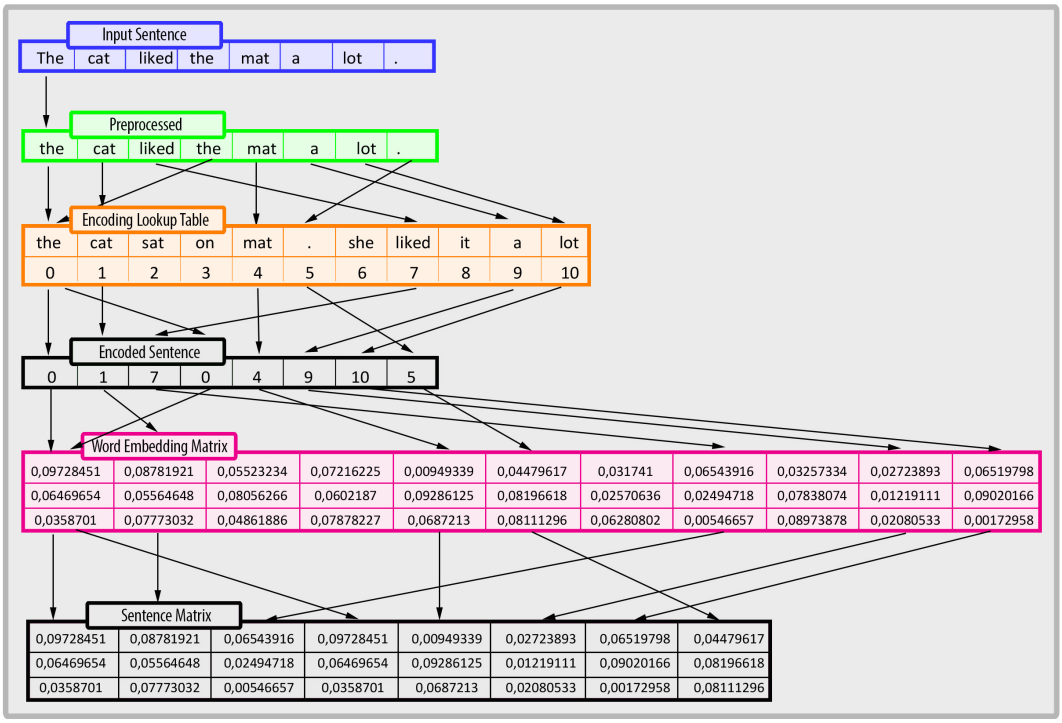
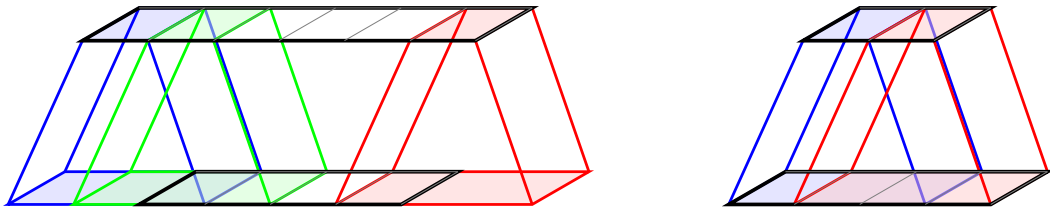



**Figure 1.2:** Embedding procedure for a new sentence.

shows the result of the preprocessing, which in this case only converted the upper case 'The' and 'She' to lower case [1]. At this stage, the model takes the preprocessed sentences and builds a dictionary of all the words appearing in the dataset. Furthermore each word is associated with an index to build the *Encoding Lookup Table* depicted by the orange table. A single index does carry much information about the word. For this reason, the model creates for each of the words a $\boldsymbol{D}$-dimensional vector, as depicted in the purple table. The quantity $\boldsymbol{D}$ is a parameter of the network and in the depiction in figure 1.1 it is assumed to be $\mathbf{3}$. The association of each of these vectors with words is done through the previously generated *Encoding Lookup Table*. The initial values for these vectors are randomly generated, however, this layers is attached as a first layers in the Convolutional Neural Network architecture. For this reason, we can propagate the gradient of the bottom layer of the neurons and get a gradient for each of these values. This allows the network to optimize the word embeddings.

After being initialized, these tables are used to map sentences to vector representation both at train and test time. This process is depicted in figure 1.2. The model takes the new sentence, which in the picture is depicted to be *"The cat liked the mat a lot."* and runs the preprocessing in the same way it did with the sentences in the dataset. The chosen sentence is purposefully new, to show that after training, the model can be used to embed sentences which it has not seen before. In fact the datasets we use

[1] Given the small size of the simple dataset, non frequently appearing words are not converted to 'UNKNOWN'.



Wide Convolution Narrow Convolution

**Figure 1.3:** Types of 1D Convolution.

to test our models have a separate training and testing set. We see only the train set when we initialize the model and train weights and embeddings, but we use the test set when we test. After producing the preprocessed green table, the model encodes the sentence by converting each word to its corresponding index via the *Encoding Lookup Table*. The two stage arrows from green table to the orange one and then ultimately to the first black one, are meant to depict this process. Once each word in the sentence is encoded, these indexes are used to pull the embedding for each word from the *Word Embedding Matrix*. This is depicted with the two stage arrows from the first black table, to the purple one and ultimately to the last black table, which is the *Sentence Matrix* and the input of the convolutional layer of the network.

### 3.1.2 Convolution Layer

Having converted the sentence, which in fact is an ordered set of words, to a vector representation, we can input it to the convolutional layer and apply an operation similar to the convolution used in vision. Nevertheless, there is an inherited fundamental difference between images and natural language. While images are 2-dimensional and we have to search for features over the two dimensions, language has clearly only 1-dimension, the word dimension. As a consequence we have to search for features over only 1-dimension.

If we are to give a full name to the convolution operation applied by the CSM and DCNN it will be 1-Dimensional Wide Convolution, as in contrast to Narrow Convolution. The difference in the two types of convolutions is depicted in Figure 1.3, where for simplicity the word embeddings are assumed to have d-dimensions where $d = 1$. The *Wide Convolution* starts operating when there is a single overlap between the feature detector and the data and ends operating in the same scenario. This is depicted on the left side of the figure. The small $4 \times 1$ black rectangle at the bottom represents the data, while the $7 \times 1$ at the top the result of the wide convolution. The process starts when the filter is in the blue position, continues in the green position and so on until the red position, where it stops as one step further there would be no overlap. In this type of convolution, having $\mathbf{S} = \mathbb{R}^{d \times w_s}$ as the sentence matrix where $d$ are the embed-

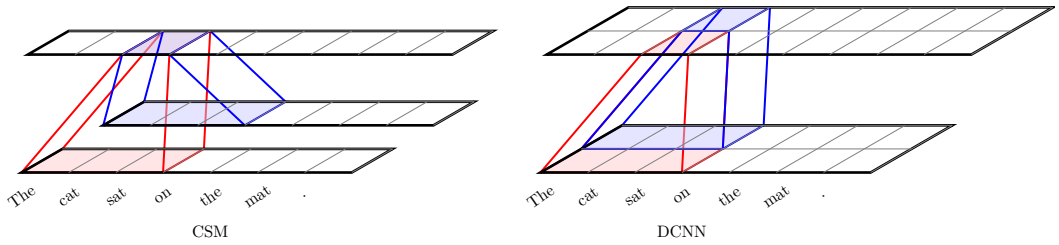



**Figure 1.4:** CSM vs. DCNN.

ding dimensions (in this case $d = 1$) while $w_s$ the number of words in the sentence and $\mathbf{f} = \mathbb{R}^{d \times w_f}$ as the feature detector where $d$ is the same as above while $w_f$ is the number of words which the filter checks, then the result of the convolution is the feature map $\mathbf{M} = \mathbb{R}^{d \times w_s + w_f - 1}$. The *Narrow Convolution*, on the other hand, operates only when there is a full overlap between the feature detector and the data. This is depicted in the right side of the figure. The same small $4 \times 1$ black rectangle is used to represent the data, while the output of the convolution this time is the $2 \times 1$ at the top. The process starts when the filter is at the blue position and ends at the red position as one step further the filter would not have a full overlap with the data. If the same sentence matrix and filter as above are used in a narrow convolution, the resulting feature map will be $\mathbf{M} = \mathbb{R}^{d \times w_s - w_f + 1}$.

As one can see from the above equation, narrow convolution requires the filter size $w_f$ to be smaller or equal to the sentence size $w_s$, otherwise the quantity $w_s - w_f + 1$ that defines the dimension of the resultant feature map will be negative. One of the main advantages of using convolutional networks is the fact that we can easily model long dependencies, so we would like our feature detectors to model n-grams where $n$ can be 7 or even up to 11. Per contra, natural language clearly includes sentences that are shorter than that. For this reason, applying wide convolution is fundamentally important to the functioning of the model.

### CSM vs. DCNN

Throughout this explanation, however, we have made the simplifying assumption that word embeddings have only 1 dimension, when in fact for all experimental purposes they have a lot more than that. Furthermore, how we treat this extra dimensions bears the fundamental difference between the CSM and the DCNN. This difference is depicted in Figure 1.4.

In the **DCNN** the embedding dimensions are treated as a second dimension of the sentence. This effectively converts the input to the convolutional layer to 2D. As a consequence it would certainly allow for a 2D convolution the same way we did in images, however, that would detect features in coordinates that do not really exist. For example,

it could detect a feature from the $3^d$ word to the $5^{th}$ and in the $2^d$ embedding dimension. This does not make sense, as that feature does not exist in the input space; what we really see are the words. To avoid this, the DCNN adds two constraints to the convolutional layer. The first one, forces the $d$ in both the feature detector $\mathbf{f} = \mathbb{R}^{d \times w_f}$ and the sentence $\mathbf{S} = \mathbb{R}^{d \times w_s}$ to be the same. This allows the feature detector to check the entire embedding dimensions for whichever position it is in. As a second constraint, the DCNN does fix the $n^{th}$ element in the d-dimension of the sentence to the $n^{th}$ element in the d-dimension of the feature detector. As a consequnce, the first row of the feature detector can only convolve with the first row of the sentence. This effectively locks vertical movement of the feature detector, effectively converting the convolution to 1D again. This technique, however, allows feature detectors in different rows to be independent of each other. This should not be the case. The convolutional layer should fire the detection of a feature only if it is available for all the characteristics of the word (embedded over all the dimensions) not only 1 of them. To circumvent this problem, the DCNN introduces the *Sum Folding* operation which we will describe in detail in section 3.1.3

In contrast to the DCNN, in the **CSM**, we do treat the embedding dimensions of a word as input channels. We believe this to be a reasonable choice for several reasons. *First*, this keeps the input 1D but split through different channels. For this reason we do not have to add restricting constraints and use a vanilla Convolutional Neural Network. *Second*, we believe in an analogy with vision, it makes much more sense to have word embedding dimensions as channels. In an RGB image, we have three inputs, one for each channel. A colour, is described by giving the entire RGB code. In fact many colours can have the same R, G or B intensity and even the same RG, GB or RB intensity, but only one has the full RGB combination. We can easily see the analogy in natural language. Each dimension in the word embedding, includes the existence of some abstract feature about that word. A lot of other words can have the same value for that feature. In fact we hope that words like 'house' and 'home' would share a lot of these features given the high similarity between then. Nevertheless, we would also hope that some of the features will be different in order to capture the different uses of these words. Stated differently, words can share a certain amount of features the same way colours share the same intensity over 1 or 2 channels, but we would not expect words to share the whole feature set unless they are the same word, the same way only one colour can have a certain RGB code. The *Third* and final major difference, regards the result of the convolution over channels versus the one over dimension. In the CSM, treating embedding dimensions as channels, the result of the convolution is the sum over the result of each channel. This enforces a full dependency over all the channels of the input, making the Sum Folding operation unnecessary and moving a further step closer to vanilla Convolution Neural Network. In fact, at this point, the

only difference our model has with the one in vision boils down the stages before the first layer of convolution.

### 3.1.3 Sum Folding

As we described in the previous section, the DCNN convolves with filters that have one dimension for each dimension of the input. This means that feature detectors in individual rows are independent of each other. For reasons also described above, dependency over dimensions of the word embeddings is necessary, so to overcome this deficiency the DCNN introduces the *Sum Folding* operation.

Its operation is really simplistic and elegant. It folds the matrix that results from the convolutional layer in the middle and sums over overlapping elements. For example, if the word embeddings are of $n$ dimensions, it fold over after row $n/2$. The overlapping rows then become (1,n), (2,n-1), (3, n-2) ... $(\frac{n}{2}, \frac{n}{2}+1)$. These overlapping rows are summed up element wise to produce the final matrix. This is depicted below.

$$M = \begin{bmatrix} w_{1,1} & \dots & w_{1,3} \\ | & | & | \\ w_{n,1} & \dots & w_{n,3} \end{bmatrix} \Longrightarrow M' = \begin{bmatrix} (w_{1,1} + w_{n,1}) & \dots & (w_{1,3} + w_{n,3}) \\ | & | & | \\ (w_{n/2,1} + w_{n/2+1,1}) & \dots & (w_{n/2,3} + w_{n/2+1,3}) \end{bmatrix}$$

### 3.1.4 Dynamic K-Max Pooling

The output of the convolution layer inform us on the firing intensity of a feature detector in every position in the sentence. For example, if the bi-gram feature (feature detector of width 2) is applied over a sentence of length 7, the output due to the wide convolution will be a vector of size 8, which represents the intesisty at which a feature was present in different locations in the sentence. A sample output could be[2] [2,8,2,1,19,7,7,3]. This clearly shows that the bigram feature detector fired really intensively when it was over the $4^{th}$ and the $5^{th}$ word (output 19), but also really present between $1^{st} - 2^{d}$, $5^{th} - 6^{th}$ and $6^{th} - 7^{th}$. Nevertheless, feature detectors are likely to have some activity even when the feature is not entirely present, or is just adjacent. While this is how we expect the detectors to act, this is a redundant information we would like to filter as we go up the model. This is achieved by the K-Max Pooling layer, which selects the $K$ most active locations of the feature over the text, preserving their relative order. The output of 4-Max pooling in the example above would be [8,19,7,7].

Certainly, different features are not present in the same location in the sentence. For this reason K-Max Pooling operates over rows of the feature maps, which do in fact cor-

[2]Number are integers for ease of representation

respond to the output of different feature detectors. This is depicted in Figure 2.1. The input sentence is converted to the sentence matrix, then convolved with three different feature detectors (red, green, orange). As a consequence the feature map, includes a row for each feature. Then the K-Max Pooling (in this case 4-Max), picks the 4 most active locations for each feature. This is depicted by the blue selection, which for better visibility of the image, is only applied to the row corresponding to the orange feature.

Nevertheless, as we described in the reasons of using Wide Convolution, sentences have a very flexible length. As a consequence, we cannot expect long sentences to exhibit a particular feature the same number of times as a short sentence. Furhtermore, sentences that include conjunctions may split into two or more simpler sentences so we would expect these sentences to exhibit features more often. In other words, at how many locations a feature can be present strongly depends on the size of the input. For this reason, we cannot decide before hand how many locations we pool as in a short sentence that can be excessive and we may well be pooling the activation over the same feature but slightly offset position of the detector. On the other hand, for long sentences we may be cutting out a particular location of the feature.

The compromise to this, is a Dynamic K-Max Pooling operation, introduced by Kalchbrenner et al. (2014) in their DCNN. This improved K-Max Pooling operator sets the $K$ parameter based on the length of the sentence and the depth of the network. In the CSM we do apply the Dynamic K-Max Pooling, but our $K$ only depends on the length of the sentence. The reason for this is the fact that in our experiments (and likewise the DCNN model published by Kalchbrenner et al. (2014)) we never go over two layer deep, and for obvious reasons we will explain in a bit the second layer of Max Pooling should in fact have fixed $K$. This will make the depth parameter fixed to 1, thus redundant.

The last layer of our model produces the final output, which is the sentence embedding. This embeddings can later be used by other models, neural or not. This however, requires the embeddings for all sentences to have the same size. This is the reason why the last layer of our network requires $K$ to be fixed.

### 3.1.5 Non-Linearity

As we explained in the Theoretical Background, the non-linearity is the neuron's activation function. Nevertheless, as it is an expensive operation, it is often prefered to run the output of the weight-input convolution through the K-Max Pooling before applying the non-linear transformation. In our experiment we chose to use the *Hyperbolic Tangent - Tanh* as experimental results showed it to perform better than *Rectified Linear Units - Relu*. The DCNN also makes the same choice.

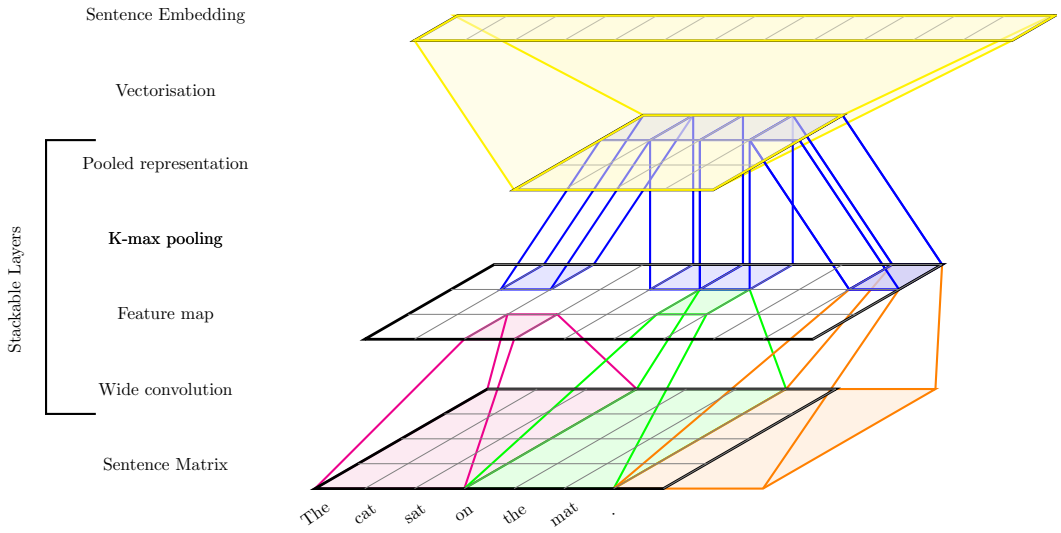



**Figure 2.1:** Convolutional Sentence Model.

## 3.2 THE LAYERED ARCHITECTURE

We have now introduced all the basic components of the model. The full model setup is depicted in figure 2.1. In this picture, one could notice a vectorization operation which we did not explain before. Restating it again, the aim of our model is to build sentence embeddings, so it makes sense to have then 1-Dimensional. The vectorization operation complies with this by picking an order and converting the matrix to a vector. This is not present in the DCNN, as the focus of their work is to classify the sentences instead of producing the low dimensional vector space sentence embeddings.

In the same picture, one can also notice that the *Pooled Representation* has a similiar structure to the *Sentence Matrix*. This means that the whole process can be repeated by stacking on top this pooled representation other layer of Convolution and K-Max Pooling. In fact this is an important feature of the model, as higher order layers detect features over the input sentence in increasing levels of abstractness. Furthermore, the K-Max Pooling effectively shrinks the size of the convolution, for this reason higher level convolutions have a view over n-grams of increasing width. For example a sentence of length 20, may have a feature spread over the first position and the last word. If we apply a feature detector of any size but 20 in the first layer of convolution, it will not capture this long dependency. Par contra the second layer of convolution will be applied after the K-Max pooling. If $K$ is set to 2, then the second layer of convolution will easily capture this dependency with a feature detector of size 2.

## 3.3 TRAINING

The model embeds the input sentence into a low dimensional vector space based on the existence of features in the text. However, given the initial randomly generated word embeddings and feature detectors, it is unlikely these sentence embeddings will be useful. For this reason, the model needs to be trained given some data. We can do so via backpropagation, however, this requires us to connect the model to some objective and cost function. Many different objectives can be used, offering very interesting properties. We will present some of these in the Future Directions section. Nevertheless, for the purpose of presenting this model, we will use the same setup as Kalchbrenner et al. (2014) for the DCNN.

The model will be used to classify the sentiment of a sentence over two classes *Positive* and *Negative*. We believe sentiment analysis to be a good way of quantifying the concept of meaning. In a review information we are trying to extract is whether the product is good or bad. For this reason if a classifier can find that difference in low dimensional embeddings produced by a model, than this model correctly maps the meaning of the sentence to abstract spaces. Furthermore, other researchers in the field do also consider sentiment classification a text understanding task (Le and Mikolov, 2014).

To train our model for sentiment classification, we feed the output of the last pooled representation (vectorization is not necessary in this case) to fully connected layer, which gives us a probability distribution over the two classes. As a consequence, the use of our cost function becomes obvious. We do use *Cross Entropy*. Furthermore, considering the fully connected layer is a neural network on its own, we can easily apply backpropagation to the whole setup.

## 3.4 SUBSTANTIAL REDUCTION IN THE NUMBER OF PARAMETERS

The DCNN, however, is a state-of-the-art model. For this reason changing it is a dangerous task which requires very good reasons. In fact, despite the qualitative reasons we explained above, for which we believe the CSM is a better choice, we had planned to use the DCNN as is for our future work. Nevertheless, we found the very large size of the model prohibitive when it was required to extend it to more complicated tasks.

In this regard, the CSM offers a substantial reduction in the number of parameters. As we explained previously, the way the DCNN applies the convolution it outputs 1 dimension for each dimension of the input, while the CSM 1 dimension for the whole input. The Sum Folding operation helps the DCNN to reduce the large number of dimension, however, given the very large dimensionality inputs used in practice the difference between the two models is considerable.

| DCNN | CSM | Layer |
|---|---|---|
| $(d * f) \times (w_s - w_f + 1)$ | $f \times (w_s - w_f + 1)$ | Convolution |
| $\frac{d*f}{2} \times (w_s - w_f + 1)$ | No-OP | Sum Folding |
| $\frac{d}{2} * (f \times k)$ | $f \times k$ | K-Max Pooling |

**Table 4.1:** Output of Layers: DCNN vs CSM

In Table 4.1 above, we show the output of the Convolution, Sum Folding and K-Max Pooling layer for both the DCNN and the CSM. We denote by $d$ the number of word embedding dimensions, $f$ the number of feature detectors in convolution layer, $w_s$ the width of the input sentence and $w_f$ the width of the filter detector. As we can see in the table, the output of the DCNN has $\frac{d}{2}$ times more dimensions. This requires the next layer to also have $\frac{d}{2}$ times more trainable weights.

In our small examples $d$ was between 2 and 4 which does not clearly show the difference between the two models, however for our experimental purposes word embeddings of 40 to 80 dimensions are used, which translate in 20 to 40 times less parameters to train. This scales even more, if more layers are added on top. In fact the number above refer to the input to the fully connected layer. Instead both Kalchbrenner et al. (2014) and we did use two stacked layer of our model. This means that our second convolution layer has 20 to 40 times less parameters, while the fully connected layer at the top another 10 to 20 times less parameters.

This does not only scale better in terms of training parameters but also in terms of computation. The model includes a Tanh operation which is applied over more inputs in the DCNN. Also the DCNN includes a Sum Folding operation, which requires its own computation time that is not present in the CSM.

## 3.5 CONTRIBUTION

In this chapter we presented a Convolutional Neural Network architecture that can capture the compositional process that maps the meaning of words to that of sentences by embedding them in a low dimensional vector space. Our model, given the layered architecture, can capture long dependencies with ease. Furthermore, in the same training process it also produces low dimensional vector space embeddings for words. These embeddings show to capture the semantic properties of the words as the compositional process that maps these words to sentences can capture the meaning of the sentence, as shown in our experimental results.

Our work in this model is very similar to the DCNN of Kalchbrenner et al. (2014), however our model has two main advantages. *First* it has a substantially reduced number of trainable parameters, which greatly reduces its size and training time allowing for more complicated usages of the model. And *Second*, it remains very close to the Con-

volutional Neural Network used in computer vision, thus making the use of extensive research in the computer vision literature easily applicable to our mode.

*We can only see a short distance ahead, but we can see plenty there that needs to be done.*

A.Turing

# 4

# Convolutional Document Model

In the previous chapter, we introduced a model that can capture the compositional process that maps the meaning of words to that of sentences. The next step in modelling the full compositional process in natural language, is to map the meaning of words to documents. An obvious choice would be to apply the CSM to a whole document. As shown in previous examples we do keep the sentence bounding symbols (. ; etc.) in the input to the network so the layer may capture the end of sentence. Furthermore, we did also explain above that higher level of feature detectors have a view over an increasing size n-gram. If inputting a whole document to the model, this would mean that higher level feature detectors would check for features over the whole document.

Nevertheless, this process has two fundamental flaws. *First*, there is a lot of unnecessary convolution operations between words at the sides of two different sentences, despite the fact that the feature detector can learn not to fire when it sees a stopping mark in the middle. *Second*, this approach tries to map the meaning of words to that of documents entirely overseeing the structural compositionality of natural language. The meaning of words builds the meaning of sentences which in turn builds the meaning of documents.

In this chapter we present a novel document model dubbed the Convolutional Document Model (CDM). The builds low dimensional vector space embeddings for entire document, while preserving the intermediary representation of sentences. To the best of our knowledge the CDM is the only model to preserve this very important structure, which as we will show in following chapter, allows for fascinating novel applications.

Furthermore, in the same training process our model also learns low dimensional vector space embeddings for words and sentences.

## 4.1 THE MODEL

We hinted the construction of this model when we introduced the *Vectorization* operation in the previous chapter. We introduced it as a relatively unnecessary operation, which in fact it was for the CSM. However, a forward seeing reader probably should have notices that after the vectorization, the sentence embeddings were a vector of numbers just like the word embeddings. If we group together the embeddings of all the sentence from a document, the same way we did with all the word of a sentence, we would end up with a matrix of sentence embeddings, or as we can call it otherwise a Document Matrix. This document matrix has the same shape of a sentence matrix, for this reason, all the operations explained for the CSM can be applied. The difference this time, is that the feature detectors are looking for features over the meanings of sentences (sentence embeddings), instead of the meaning of words (word embeddings). This clearly maintains the compositional process of natural language. What we have actually done is to force the process of mapping the meaning of words to documents through the sentence representation first.

This idea of carefully choosing intermediary representation is not new. Gülçehre and Bengio (2013) show in their work that learning intermediary representation helps generalization. They setup an experiment where an algorithm should decide if in an image with three Pentomino objects all of them are the same [1]. In this task, all state-of-the-art black-box machine learning algorithms fail. Their experiment involves two different setups. In the first one a Convolutional Network applied over patches of the images is fed into a fully coonected layer which binary classifies if all objects are the same or not. The second setup, is the same, however the feature detectors of the Convolutional Network are first trained to detect the type of the pentomino objects separately, then fed into the fully connected layer that in turn is again trained against the binary label. In their biggest dataset the duble supervised model made only 0.01 errors agains the 12.4 of the single supervision one. Furthermore, Hinton et al. (2011) show that forcing information to pass through carefully chosen bottlenecks it is possible to control the types of intermediary representations that are learned.

In our case the solution is simpler. Images occur in nature where there is little defined structure, as a consequence in the experiments above Gülçehre and Bengio (2013) had to qualitatively decide their intermediary representation. Par contra, natural language has a very clear structure, words compose sentences, sentences compose paragraphs, paragraphs compose chapter and chapters compose books. For this reasons our inter-

[1] To be noted is that a rotated or scaled object is still the same object.

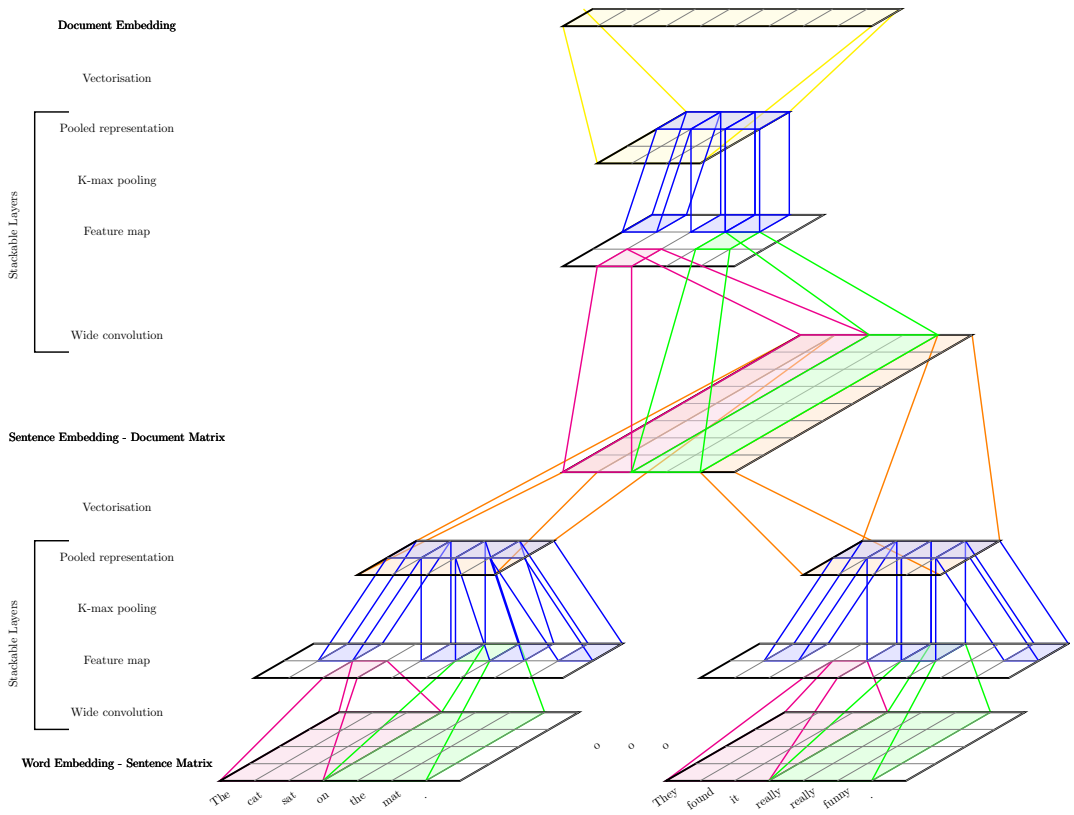



**Figure 1.1:** Convolutional Document Model.

mediary representation could be easily defined.

### Training the Model

A full CDM model is depicted in Figure 1.1. The example assumes a document of five sentences the first and last of which are shown in the picture. As we can see in the first layers the model acts like a group of CSMs, one for each sentence. Each of them takes as an input a set of words and goes over the same process as the CSM to build the word embedding matrix (described in Section 3.1.1). Then it applies Convolution and K-Max Pooling layers to create sentence embeddings for each sentence. The picture shows only one layer of Convolution and Pooling, nevertheless as we described for the CSM many more layers can be added to detect features over the input sentences in increasing level of abstractness.

After the final Pooled Representation the vectorization operation is applied. The vectorization defines the boundary where convolving over words stops and convolving over sentences starts. If we see the orange operation in the figure, the result of each sentence is a column in the Document Matrix. In the same time, one can see that the feature detectors over the Document Matrix depicted in red and green see the whole column and move along rows which correspond to different sentences. If in sentences feature detectors went over words to see if some important patterns were present, now

the feature detectors goes over sentences to see if the patterns are present. Apart from the conceptual difference which proves to be really crucial in future applications, the remaining part of the model still works like the CSM. The result of the convolution forms a feature map. The size of the later is reduced via K-Max Pooling to produce the pooled representation. The picture, again only shows one layer of Convolution and Pooling for the document part, nevertheless many more can be added to detect features over the input document in increasing level of abstractness.

The vectorization operation is very simple, but opens an interesting perspective. If in the above discussion we consider the document to be the paragraph of a book, then the output of the model is a paragraph embedding. Once vectorized, we can put together the embedding of many paragraphs in a Chapter Matrix and continue the same operation to have a chapter embedding. We can keep going in the same direction to build a Book Matrix and book embeddings. What makes this even more fascinating, is the fact that the whole model will be one big setup which can be optimized via back-propagation. In other words we can train the whole model and have word, sentence, paragraph, chapter and book embeddings in on go.

Training for our experiments we will follow the same logic as we did with the CSM and what seems to be the common denominator in the field. We will train our model over sentiment prediction via a fully-connected layer and Cross-Entropy.

## 4.2 CONTRIBUTION

In this section we introduced a novel document embedding model. It is very similar to the CSM but has a fundamental difference. While the CSM maps the meaning of words to sentences, the CDM maps the meaning of words to documents via sentences. Keeping sentence representation when building document embeddings seems to have been overseen by researchers in the field. While independently and in the context of document classification it may seem unnecessary in the following chapters we will show some fascinating application which are a result of this well defined structure.

Furthermore, the CDM highlights the importance of the reduction in parameters introduced by the CSM. As we can clearly see from Figure 1.1 the CDM adds other layers on top of the CSM which requires the model to scale up considerably. Furthermore, in training and classifying time the CDM requires many parallel CSMs (one for each sentence) which puts its memory requirements very high.

*Mathematical reasoning may be regarded rather schematically as the exercise of a combination of two facilities, which we may call intuition and ingenuity.*

A.Turing

# 5

# Saliency Extraction and Automatic Summarisation

In this chapter, we introduce a novel automatic summarisation algorithm based on the features of our Convolutional Document Model. In fact we introduce an entire new perspective to the extractive summarisation domain. Previous methods rely on human engineered heuristics to define whether a sentence is important or not. Machine Learning has also been applied by mapping the extractive summary as a supervised task, training a classifier to decide whether a sentence is important or not. This makes a big assumption. A sentence important in one document is not necessarily important in another. Our approach is entirely different. We feed the review to our model and once it has understood it we ask it to tell us which sentences were important its understanding.

## 5.1 SALIENCY EXTRACTION

As we have described throughout this thesis, natural language goes through a very complicated process that maps the meaning of words to sentences and further the meaning of sentences to documents. Nevertheless, an intuition is that not all words in a sentences are equally responsible for the meaning of the sentence. Equivalently, not all sentences in a document are equally responsible for the meaning of the document. Lets consider the following line from the Beatles' classic: *"Yesterday love was such an easy game to play."* If instead of that sentence we only see the set of words *yesterday, love,*

*easy* we can still infer most of the meaning. If instead we see the complement set *was, such, an, game, to*, it is impossible to infer the meaning of the original sentence.

For this reason, it is interesting to capture the importance different words have to the meaning of the sentence, or as we call it the saliency of different words. In a linear model, this is relatively simple, as the outcome of the model can be expressed with the linear function $O(S) = w^T S + b$. Under the assumption that the outcome of the network quantifies the concept of meaning[1], the saliency of each word is closely related to the weight associated with it. Linear models, however, assigns weights to words independent of their neighbourhood. For example the negative *'not'* has entirely different sentiment depending on whether it is in front of a positive or negative word. For this reason, we would like to use our feature detectors when analysing the meaning of a sentence and then try to extract word salience. This is complicated as in a Convolutional Neural Network, the output is a highly non-linear function of the input. Nevertheless, having an architecture that closely resembles the one used in vision allows us to use their extensive literature. In fact two slightly different saliency extraction algorithms exist developed by Zeiler and Fergus (2012) and Simonyan et al. (2013). Their work is focused on computer vision where they can use this saliency maps to generate visible images, hence these methods are most often refereed to as visualisation techniques.

Zeiler and Fergus (2012) base their visualisation approach on their previously developed Deconvolutional Neural Network (DeconvNet) (Zeiler et al., 2011). The DeconvNet is a generative addition of the ConvNet. Paired with a ConvNet they successfully use it in an unsupervised learning approach. Simonyan et al. (2013) on the other hand, bases their saliency extraction algorithm on simple numerical optimization. Nevertheless, he explains that his method is a generalization of the DeconvNet approach.

In difference to both approaches above, we are tackling the saliency extraction problem in natural language. If we refer back to both our CSM and CDM, the input to the first layer of convolution is a optimized word embedding matrix. If we try to reconstruct this matrix, we would not be getting any useful information out of the network, as unless a reconstructed vector of this matrix is identical to that of a word we cannot see which word it is. For this reason, the generative approach used by Zeiler and Fergus (2012) can be seen as an overkill for our task. Consequently, we will used a saliency extraction technique closely related to the one of Simonyan et al. (2013).

As we mentioned previously the output of a Convolutional Neural Network $O(S)$ is a highly non-linear function of the input $S$. Here the Tylor's theorem comes handy, as we can approximate the highly non-linear function with the first-order Taylor expansion

[1]Explained in section 3.3.

in the neighbourhood of current point $S$.

$$O(S) \approx O(S_o) + \frac{\delta O(S)}{\delta S}(S - S_o) \qquad \text{First-order Taylor expansion} \qquad (5.1a)$$

$$O(S) \propto \frac{\delta O}{\delta S}(S) \qquad \text{Neglecting the constant terms} \qquad (5.1b)$$

$$O(S) \propto wS \qquad \text{Replacing the derivative with } w \qquad (5.1c)$$

At this point we can see that the equation is very similar to the linear models, and we can use the same interpretation. The words in sentence vector $S$ that is associated with the highest value in $w$ (the derivative matrix), is the one to have the most impact on the meaning of the sentence. Convolutional Neural Networks are trained via back-propagation, which as we discussed in Chapter 2, uses the chain rule to calculate the derivative of the outcome with respect to every parameter of the network. For this reasons, we can have $w$ above and as a consequence a saliency extraction algorithm calculated in a simple backpropagation pass of the network with no further engineering.

There is however a small complication. This method requires the gradient, but we only have one once we connect the model to a cost function. The cost function on the other hand requires a label to make the comparison with the network prediction. Furthermore, if the model correctly predicts the outcome, the cost will be zero and there will be no gradient. We can overcome these problems via a 'Pseudo-Label' technique. We let the network predict the label. Then we take this label, invert it and supply it to the cost function as the real label. This not only provides the cost function with a label, but it also incurs the highest possible cost, as it effectively makes the 'real' label the opposite of the prediction.

We derived a mathematical explanation on why the gradient back-propagation method works. Nevertheless, it has a very simple intuitive explanation. We let our model, predict the meaning of the document and then we tell it that it was as wrong as it could go. The model then immediately goes back to its neurons and blames them for guiding it wrong. This is the gradient in the back-propagation. There is no point, however to blame neuron which had no say in the decision, so the model blames the neurons that most effected its outcome.

## 5.2 AUTOMATIC SUMMARISATION

Using the technique above, we can highlight important words in a sentence, and see what the network has learned. This is particularly important in order to optimize the network. In fact, Zeiler and Fergus (2012) won the IMAGENET competition by tweaking their model to target problems they had spotted in their visualisation. For practical purposes, however, the visualisation technique in images and also in sentences is not

particularly useful. Despite the motivational explanation in the previous section, no one will be interested to see certain words extracted from a sentence, even though they may perfectly convey the information. For example, nobody would want to read *yesterday, love, easy* instead of *"Yesterday love was such an easy game to play."*

The same is not true for document. People would love to see only important sentences extracted from documents. In fact we call them summaries. For this reason, we would like to extend our saliency extraction algorithm to sentences in document instead of words in sentences. Here the structure of the CDM becomes crucial. The CDM maps the meaning of words to that of documents without losing the intermediary representation of sentences. To the best of our knowledge, the CDM is the only model to maintain this important structure.

This intermediary representation can effectively be used to extract sentence saliency in a very similar way to the extraction of the word saliency. The model takes an input $S$ and maps it to the output $O(S)$. Nevertheless, having a intermediary representation, this means that $O(S)$ is a composed function which first maps to the intermediary representation and then to the final representation, hence it can be expressed as $O(S) = O_2(O_1(S))$. Having carefully built the intermediary structure to correspond to sentences, it means that $O_1(S)$ is the document matrix $D$. The output of the network is then a function of the sentences $O(D)$ and with the same reason as above, we can extract sentence salience with the derivative of the outcome with respect to $D$. In the CDM this exactly corresponds to following the back-propagation path up to the Sentence Embedding level. The full derivation is shown in Equation 5.2.

$$O(S) = O_2(O_1(S)) \quad \text{The output is a composed function} \quad (5.2a)$$

$$O(S) = O_2(D) \quad \text{Replacing the inner function with D} \quad (5.2b)$$

$$O_2(D) \approx O_2(D_0) + \frac{\delta O_2(D_0)}{\delta D_0}(D - D_0) \quad \text{First-order Taylor expansion} \quad (5.2c)$$

$$O_2(D) \propto \frac{\delta O}{\delta D}(D) \quad \text{Neglecting the constant terms} \quad (5.2d)$$

$$O(S) = O_2(D) \propto wD \quad \text{Replacing the derivative with } w \quad (5.2e)$$

By associating each sentence with a numerical value that represents it importance, we have ample choices on the sentences we select. For example, we can select the $n$ most important sentences of a document. Nevertheless we know how many sentences we have, so we can also select the top $n\%$ most important sentences. In Figure 2.1 we show example summaries extracted from our algorithm on movie reviews, by selecting the top 20% most important sentences. The text corresponds to the original full

I caught this movie on the Sci-Fi channel recently. It actually turned out to be pretty decent as far as B-list horror/suspense films go. Two guys (one naive and one loud mouthed a **) take a road trip to stop a wedding but have the worst possible luck when a maniac in a freaky, make-shift tank/truck hybrid decides to play cat-and-mouse with them. Things are further complicated when they pick up a ridiculously whorish hitchhiker. What makes this film unique is that the combination of comedy and terror actually work in this movie, unlike so many others. The two guys are likable enough and there are some good chase/suspense scenes. Nice pacing and comic timing make this movie more than passable for the horror/slasher buff. Definitely worth checking out.

I just saw this on a local independent station in the New York City area. The cast showed promise but when I saw the director, George Cosmotos, I became suspicious. And sure enough, it was every bit as bad, every bit as pointless and stupid as every George Cosmotos movie I ever saw. He's like a stupid man's Michael Bey – with all the awfulness that accolade promises. There's no point to the conspiracy, no burning issues that urge the conspirators on. We are left to ourselves to connect the dots from one bit of graffiti on various walls in the film to the next. Thus, the current budget crisis, the war in Iraq, Islamic extremism, the fate of social security, 47 million Americans without health care, stagnating wages, and the death of the middle class are all subsumed by the sheer terror of graffiti. A truly, stunningly idiotic film.

Graphics is far from the best part of the game. This is the number one best TH game in the series. Next to Underground. It deserves strong love. It is an insane game. There are massive levels, massive unlockable characters... it's just a massive game. Waste your money on this game. This is the kind of money that is wasted properly. And even though graphics suck, thats doesn't make a game good. Actually, the graphics were good at the time. Today the graphics are crap. WHO CARES? As they say in Canada, This is the fun game, aye. (You get to go to Canada in THPS3) Well, I don't know if they say that, but they might. who knows. Well, Canadian people do. Wait a minute, I'm getting off topic. This game rocks. Buy it, play it, enjoy it, love it. It's PURE BRILLIANCE.

The first was good and original. I was a not bad horror/comedy movie. So I heard a second one was made and I had to watch it . What really makes this movie work is Judd Nelson's character and the sometimes clever script. A pretty good script for a person who wrote the Final Destination films and the direction was okay. Sometimes there's scenes where it looks like it was filmed using a home video camera with a grainy - look. Great made - for - TV movie. It was worth the rental and probably worth buying just to get that nice eerie feeling and watch Judd Nelson's Stanley doing what he does best. I suggest newcomers to watch the first one before watching the sequel, just so you'll have an idea what Stanley is like and get a little history background.

Why do all movies on Lifetime have such anemic titles ? "An Unexpected Love" - ooh, how provocative!! "This Much I know" would have been better. The film is nothing special. Real people don't really talk like these characters do and the situations are really hackneyed. The straight woman who "turns" lesbian seemed more butch than the lesbian character. If you wanna watch two hot women kiss in a very discreet fashion, you might enjoy this. Although it seems like it was written by someone who doesn't really get out in the world to observe people. Why am I wasting my time writing about it?

When the movie was released it was the biggest hit and it soon became the Blockbuster. But honestly the movie is a ridiculous watch with a plot which glorifies a loser. The movie has a Tag - line - "Preeti Madhura, Tyaga Amara" which means Love's Sweet but Sacrifice is Immortal. In the movie the hero of the movie (Ganesh) sacrifices his love for the leading lady (Pooja Gandhi) even though the two loved each other! His justification is the meaning of the tag - line. This movie influenced so many young broken hearts that they found this "Loser - like Sacrificial" attitude very thoughtful and hence became the cult movie it is, when they could have moved on with their lives. Ganesh's acting in the movie is Amateurish, Crass and Childishly stupid. He actually looks funny in a song, (Onde Ondu Sari ...) when he's supposed to look all stylish and cool. His looks don't help the leading role either. His hair style is badly done in most part of the movie. POOJA GANDHI CANT ACT. Her costumes are horrendous in the movie and very inconsistent. The good part about the movie is the excellent cinematography and brilliant music by Mano Murthy which are actually the true saving graces of the movie. Also the lyrics by Jayant Kaikini are very well penned. The Director Yograj Bhat has to be lauded picturization the songs in a tasteful manner. Anyway all - in - all except for the songs, the movie is a very ordinary one !!!!!!

A friend and I went through a phase some (alot of) years ago of selecting the crappest horror films in the video shop for an evening's entertainment. For some reason, I ended up buying this one (probably v. v. cheap). The cheap synth soundtrack is a classic of its time and genre. There's also a few very amusing scenes. Among them is a scene where a man's being attacked and defends himself with a number of unlikely objects, it made me laugh at the time (doesn't seem quite so funny in retrospect but there you go). Apart from that it's total crap, mind you. But probably worth a watch if you like films like "Chopping Mall". Yes, I've seen that too.

I tried restarting the movie twice. I put it in three machines to see what was wrong . Did Steven Seagal's voice change? Did he die during filming and the studio have to dub the sound with someone who doesn't even resemble him? Or was the sound on the DVD destroyed? After about 10 minutes, you finally hear the actor's real voice. Though throughout most of the film, it sounds like the audio was recorded in a bathroom. I would be ashamed to donate a copy of this movie to Goodwill, if I owned a copy. I rented it, but I will never do that again. I will check this database before renting any more of his movies, all of which were (more or less) good movies. You usually knew what you were getting when you watched a Steven Seagal movie. I guess that is no more.

Vertigo co - stars Stewart (in his last turn as a romantic lead) and Novak elevate this, Stewart's other "Christmas movie," movie to above mid - level entertainment. The chemistry between the two stars makes for a fairly moving experience and further revelation can be gleaned from the movie if witchcraft is seen as a metaphor for the private pain that hampers many people's relationships. All in all, a nice diversion with legendary stars, 7/10

"Written on the Wind" is an irresistible, wonderfully kinky film, as only director Sirk could have done it. The movie is submerged in a bucket full of Freudian symbols, weird melodramatics and colorful contrasts. The connection between financial success and moral decay is the film's main theme. Sirk seems to suggest that sexual dysfunction is one of the side effects of capitalism. However, I prefer to see the movie as a prime example of what Sirk could do with kitschy material. The palette of colors is particularly impressive. The acting in the film is great too. Rock Hudson and Lauren Bacall are terribly glamorous and give the film an aura of elegance, but the movie belongs to Robert Stack and Dorothy Malone (she deservedly won the Best Supporting Actress Oscar), who manage to keep the film at a boiling point. Kudos to Frank Skinner's pulsating score, Russell Metty's brilliant camera work (every single shot is a masterpiece in itself), and the production design department. Also, the title tune is a beauty. It's an unforgettable movie.

**Figure 2.1:** Example summaries produced by the CDM.

review, while the highlighted sentences correspond to the ones extracted by our algorithm. This visualisation, show the two usages of our model. We can either show only the selected sentences and create a summary, or simulate the human going over text with a highlighter emphasizing important sentences. A more complete set of review summaries randomly chosen from our 50'000 IMDB dataset is shown in Appendix A.

## 5.3 AUTOMATIC SUMMARISATION EVALUATION TECHNIQUE

In our introduction to the CSM and CDM we explained why in sentiment analysis the sentiment is closely related to the meaning. Furthermore, we showed that researchers in the field, do follow the same reasoning by considering sentiment analysis a text understanding task (Le and Mikolov, 2014). Explaining our saliency extraction and summarisation technique we also explicitly assumed that the outcome of the network quantifies the concept of meaning. This implicitly assumed that the sentiment is closely related to the meaning in sentiment analysis tasks.

This gives an interesting perspective on an scalable Automatic Summarisation Evaluation technique, which requires no human supervision. If a certain agent can read a set of reviews and understand whether they are good or bad, can he still infer the sentiment if instead of the real review we show him the summarised version? This sentence is in a nutshell our evaluation technique. The full procedure we suggest is as below.

1. Train an agent (hereafter 'the agent') on the train set.
2. **Evaluate** the accuracy of 'the agent' on the test set.
3. Use the automatic summarisation algorithms to extract summaries of the test set.
4. For each document in the test set, randomly pick a number of sentences equal to the one extracted by the algorithm
5. **Evaluate** the accuracy of 'the agent' on the test set summaries.
6. **Evaluate** the accuracy of 'the agent' on the test set random summaries.
7. **Compare** the accuracy of 'the agent' between the full document and the summarised ones.
8. **Compare** the accuracy of 'the agent' between the summarised document and the random selection.

The evaluation technique suggest running two comparisons. The first compares the accuracy of the summary to the accuracy on the full document. This intends to see if the summary maintained the important information. For example if 'the agent' could

understand the sentiment of a review in all of the original documents, that is 100% accuracy and maintain that 100% on the summaries we can say that the summaries kept the meaning of the documents intact. The second comparison compares the accuracy on the summarised reviews to the accuracy on randomly chosen sentences. This tries to see if maintaining the meaning of document when dropping sentences is a feature of the dataset, or the algorithm. For example, if we select 20% of the sentences in a review randomly and the accuracy does not drop, it means that the dataset is able to be reduced without loss of meaning. If that is the case, then unfortunately this dataset is not a good for testing summarisation algorithms.

We referred to the sentiment prediction algorithm as 'the agent' as any algorithm will keep the logic of the technique intact, hence can be used. One restriction we suggest is the use of a different algorithm from the one used to extract the summaries. For example we use the CDM to extract the summaries, but the CDM can also be used to predict the sentiment hence it can be used as the agent. Nevertheless this should not be done as it can extract sentences biased towards that model. As the Naive Bayes is commonly used as a benchmark in accuracy, works very differently to most other models, is very simple yet very effective and takes very little time to train, we suggest it to be used as 'the agent'.

## 5.4 CONTRIBUTION

In this section we introduced a novel automatic extractive summarisation algorithm in an entirely new approach for the field. In difference to previous approaches, our algorithm does not decide in turn whether a sentence is important or not. Instead it reads the text, understands it (forward-propagation) and then it tells us which sentences helped it most build its understanding (back-propagation). We would like to note that this approach is only possible because of the structure we defined in the CDM. In fact the summarisation algorithm we introduce is a feature of the CDM.

Furthermore, we introduced a novel Extractive Summary Evalution technique. Requiring no human supervision, our technique can scale well to deep learning models, which given the very high number of parameters, need a very large dataset to train.

While training with a label may sound as a limitation, we only did so in order to benchmark our model against others like Kalchbrenner et al. (2014) and Le and Mikolov (2014) who also build text understanding models. In fact, in future directions we show that we can drop the label and still maintain the features of our CDM and Summarisation Algorithm.

# Part III

# Results, Directions and Conclusion

*Before you start some work, always ask yourself three questions - Why am I doing it, What the results might be and Will I be successful. Only when you think deeply and find satisfactory answers to these questions, go ahead.*

Chanakya

# 6 Experimental Results

## 6.1 CONVOLUTIONAL SENTENCE MODEL

In this thesis we introduced a new sentence model called the Convolutional Sentence Model. The CSM allowed us to reduce the number of parameters of the DCNN substantially. As discussed previously, this is an important feature in scaling up these models. The DCNN, however, is a state-of-the-art algorithm, so in changing it, it is natural to ask ourselves if we have paid any price in performance. For this reason we replicate the Twitter Sentiment Classification task reported by Kalchbrenner et al. (2014) in their presentation of the DCNN. We did chose this dataset as the one where the DCNN outperformed the competition with the biggest margin.

### 6.1.1 Dataset

The Twitter Sentiment dataset (Go et al., 2009) has two separate partitions, one for training and the other for testing. The training set is labelled using a distant supervision technique. The distant supervision heuristic they use assumes that tweets containing positive emoticons [ :), :-), : ), :D, =) ] do have a positive sentiment; while the tweets containing negative emoticons [ :(, :-(, : ( ] have a negative sentiment. The Twitter API allows them to retrieve tweets that contain positive emoticons with the query ':)' and tweets that contain negative emoticons with the query ':('. The result of these queries is preprocessed through the following filters.

1. The emoticons are ripped off from the data. This is important in prohibiting the

model from learning the trivial mapping emoticon to sentiment.

2. Tweets that contain both positive an negative emoticons are removed. Considering the sentiment marking scheme used, these tweets cannot be marked.
3. Retweets are removed. This is used to make sure the model does not see one tweet more often than it sees others.
4. Repeated tweets are removed for the same reason as above.
5. Tweets that include the ':P' emoticon are removed. As reported, at the time of collection the Twitter API returned tweets containing ':P' in response to a ':(' query. There is no evidence, however that the ':P' emoticon implies negative sentiment

At the end of the process the first 800'000 negatively labelled and 800'000 positively labelled tweets were chosen to form the training set of 1'600'000 tweets.

The testing set, on the other hand, is labelled with human supervision. These tweets do not necessarily contain an emoticon and even if they do, the emoticon does not necessarily imply the sentiment of the tweet. This results in 177 negative tweets and 182 positive tweets for a whole testing set of 359 hand labelled tweets.

### 6.1.2 Model Setup and Parameters

Our CSM has many parameters which may be tweaked in order to better fit the data. Usually such process is done via Grid-Search. Nevertheless in this experiment we want to see how far we have departed from the DCNN when we reduced the number of parameters rather than optimizing our model to achieve the best possible results. To do so we have replicated the DCNN setup that achieved state-of-the-art scores as closely as possible, but using our approach in the Convolutional Layer. The details of the model setup are listed below.

| Layer | Parameters |
|---|---|
| **Softmax** | N/A |
| **Tanh** | N/A |
| **K-Max Pooling** | (K:4) |
| **Convolution Layer** | (Number of Detectors:14; Width: 5) |
| **Tanh** | N/A |
| **Dynamic K-Max Pooling** | (Dynamic K: 0.5) |
| **Convolution Layer** | (Number of Detectors:6; Width: 7) |
| **Word Embeddings** | (60 Dimensions) |

**Table 1.1:** Convolutional Sentence Model.

| Model | Errors |
|---|---|
| SVM | 66 |
| BiNB | 62 |
| MaxEnt | 61 |
| Max-TDNN | 76 |
| NBoW | 68 |
| DCNN | 45 |
| Our Model | 46 |

**Table 1.2:** Twitter Sentiment Dataset

### 6.1.3 Results

To allow for comparison with other models, we train on the train set and test on the test sets as defined by the dataset.

The results of our model are listed in Table 1.2. As we can see, our model makes 1 more mistake than the DCNN. Nevertheless both models outperform the rest with a considerably big margin. The next best model, the MaxEnt makes 15 errors more than the CSM or effectively 34,7% more errors.

### 6.1.4 Conclusions

Given the results above, we can conclude that with a CSM we introduced a great model that substantially reduces the number of parameters of the DCNN consequently improving its runtime performance with very little, or close to negligible, reduction in classification accuracy.

This result should inspire feature research on the CSM. A Grid-Search may actually find a setup of the CSM where it outperforms the DCNN.

## 6.2 CONVOLUTIONAL DOCUMENT MODEL

Our second contribution was a novel document model that could capture the compositional process that maps the meaning of words to that of documents while preserving the distinction of words and sentences. We dubbed this model the Convolutional Document Model. Even though our interest in this research project is to create distributed representations for words, sentences and documents we need some quantification to test the sanity of our models.

To do so, we need to test our model on text understanding tasks. To be in the same mood with our previous experiment and other competitors in the field, like Le and Mikolov (2014), we do experiment on sentiment analysis in the IMDB Movie Review dataset proposed Maas et al. (2011).

### 6.2.1 Dataset

The dataset is composed of 100'000 movie reviews extracted from IMDB. It is separated in three subsets: train, test and unsupervised. The train and the test set contain 25'000 reviews each, where the entire 50'000 they compose is evenly split between positive and negative reviews. The unsupervised set on the other hand contains another 50'000 reviews, also evenly split between positive and negative.

A characteristic of this dataset is the fact that movie reviews have several sentences, thus we can consider them documents. The sentiment on IMDB is ranked on a scale of 1 to 10, with sentiment <5 meaning negative and >5 meaning positive. Nevertheless marks of 5 or 6 show very weak sentiment from the reviewer, for this reason for the train and test set only reviews with strong sentiment (<=4 and >=7) are included. The unsupervised set on the other hand, does include reviews with weak sentiment. Another imposed restriction, only allows the dataset to include a maximum of 30 reviews from any given movie. This is aimed at prohibiting the model from learning the simple correlation that exists between reviews of the same movie.

### 6.2.2 Model Setup and Parameters

In this model, in difference to the CSM, we aim to get the best possible performance. For this reasons, all the parameters of the model were open to optimization aiming at better modelling the data. To decide on our final parameters we followed the following procedure. First we set-up a random parameter search where we started a pool of jobs, each representing one of our models, with the parameters randomly set within some range. Each job was left running for 10 computation hours on the **Westgrid**[1] cluster. The results of this first run were collected and some data analysis was done in order to identify parameters that had positive correlation with the accuracy of the model. After such correlations were identified, a second pool of jobs was submitted, this time with the parameters manually set to match our observations.

The setup and parameters of our best performing model are listed in Figure 2.1.One should easily notice that this model is considerably smaller than the CSM in the previous section, despite the fact it does include a CSM in it. In fact we were surprised with this results. The CSM and the DCNN clearly showed us that to model words 60 dimensional embeddings would yield the best result and to model sentences two stacked levels of convolutions. Par contra, in this setup, words are embedded in only 10 dimensions and sentences are modelled with a single layer of convolution (the second layer corresponds to documents). We did run a lot of test to try and identify the reason of this unexpected result. We tried to use the first half of this model as a CSM in

[1] https://www.westgrid.ca

| Layer | Parameters |
|---|---|
| **Softmax** | N/A |
| **Dropout** | (R:0.5) |
| **Tanh** | N/A |
| **K-Max Pooling** | (K24) |
| **Convolution Layer** | (Number of Detectors:15; Width: 5) |
| **Reshape for Document** | N/A |
| **Tanh** | N/A |
| **Dynamic K-Max Pooling** | (K: 4) |
| **Convolution Layer** | (Number of Detectors:6; Width: 5) |
| **Dropout** | (R:0.2) |
| **Word Embeddings** | (10 Dimensions) |

**Table 2.1:** Convolutional Sentence Model.

the Twitter Sentiment dataset. The results were considerably worst than the setup reported in the previous section. Eventually we caught on the reason. Our bigger models, would routinely achieve >99% accuracy on the train set. This means that these models would extremely overfit the training data. On a retrospective though, we should have expected this. Deep Models have a large number of parameters and a rule of thumb in Learning requires the number of datapoints to increase proportionally with the number of parameters. The IMDB Movie Review dataset despite being, to the best of our knowledge, the biggest labelled document dataset available, it is still too small to effectively train a deep network on it. For this reason, the Grid-Search identified a small model with a small number of parameters. While this prevents overfitting, it also limits the modelling power of the setup.

### 6.2.3 Results

As we did for the CSM, to allow for comparison with other models, we train on the train set and test on the test sets as defined by the dataset.

The results of our model are listed in Table 2.2. The CDM is the third best performing model. Nevertheless, our model has considerable advantages against both models that perform better than us on the sentiment classification task. Wang and Manning (2012) have the second best performing model in the dataset. They rely on a Support Vector Machine with Naive Bayes features counted on bi-grams to classify the sentiment of the reviews. While this clearly shows to work extremely well for this task, we want to note that their work has an entirely different goal compared to ours. We aim at building distributed embeddings that capture the meaning of documents, and we use the sentiment classification as a benchmark to test whether our embeddings do in fact capture this meaning. Wang and Manning (2012) on the other hand, do not build doc-

| Model | Accuracy |
|---|---|
| BoW (b$\Delta$t'c) (Maas et al., 2011) | 88.23% |
| Full+BoW (Maas et al., 2011) | 88.33% |
| Full+Unlabelled+BoW (Maas et al., 2011) | 88.89% |
| WRRBM (Dahl et al., 2012) | 87.42% |
| WRRBM+BoW (bnc) (Dahl et al., 2012) | 89.23% |
| SVM-bi (Wang and Manning, 2012) | 86.95% |
| NBSVM-uni (Wang and Manning, 2012) | 88.29% |
| NBSVM-bi (Wang and Manning, 2012) | 91.22% |
| Paragraph Vector (Le and Mikolov, 2014) | 92.58% |
| CDM | 89.38% |

**Table 2.2:** IMDB Movie Review Dataset

ument embeddings and the focus of their work is to show that bag-of-features models are still strong performers.

Our real competitor is the **Paragraph Vector** of Le and Mikolov (2014), who also build document embeddings and use the sentiment classification task only as a sanity check for their model. They achieve a state-of-the-art result on the IMDB Movie Dataset. However, our models has strong advantages. *First*, once trained our model needs a simple forward-propagation to produce the document embeddings. On the other hand their model also needs a back-propagation optimization. The back-propagation is more expensive than the forward propagation as gradients need to be calculated. This makes their model more than twice as slow as ours. *Second*, in the same run our model learns the meaning of words, sentences and documents while preserving the distinction between words and sentences which was crucial to our summarisation technique. Furthermore, our approach can be extended to include paragraphs, chapters and whole books in the same training process. This is not possible in the **Paragraph Vector** as it can only learn the meaning of words and documents/paragraphs while losing the distinction of sentences in between. *Third* and final, despite the fact that our the **Paragraph Vector** achieves a performance higher than ours in the IMDB Movie Review Dataset it should be noted that our model uses 1/3 of the data they use. Our model uses a supervised training approach (our unsupervised approach is presented as a future direction), which only allows us to use the 25'000 training set. Le and Mikolov (2014) on the other hand have an unsupervised training approach which allows them to also use the 50'000 unsupervised set, making their effective training set with 75'000 examples against the 25'000 we use.

### 6.2.4 Conclusions

Given the results above, we can conclude that with a CDM we introduced a fantastic model. It faster than any other model at producing distributed embeddings for documents. Furthermore, it is the only model than in the same run can also produce sentence and word embeddings. In difference to competitors, the technique introduced by the CDM can be extended to map the whole meaning composition process in the natural language: words → sentences → paragraphs → chapters → books.

While it does not achieve state-of-the-art results in our benchmark dataset it only trained on $^1/_3$ of the data the competitors did and showed considerable overfitting. For this reason we are to believe that given a larger dataset or an improved unsuprevised training approach, this model may well achieve state-of-the-art performances.

## 6.3 SALIENCY EXTRACTION AND AUTOMATIC SUMMARISATION

Our last contribution in this thesis was the introduction of a novel document summarisation technique. In fact we did introduce an entire new perspective in automatic extractive summarisation. We did also introduce a new scalable evaluation method for automatic summarisation algorithms. In this section we evaluate the summaries produced by our model on the IMDB Movie Reviews. As described in Chapter 5, the baseline model ('the agent') is a Naive Bayes classifier (NB).

### 6.3.1 Results

The results are listed below. In Table 3.1 we take only a certain proportion of the sentences to create the summary. We do so, in order to allow longer reviews which most likely include more information to be able to propagate it through.

| Proportion | Summary | Random | Margin |
|---|---|---|---|
| 100% | 83.03 | 83.03 | — |
| 50% | 83.53 | 79.79 | +3.74 |
| 33% | 83.10 | 76.72 | +6.38 |
| 25% | 82.91 | 74.87 | +8.04 |
| 20% | 82.67 | 73.20 | +9.47 |

**Table 3.1**

We see that our model could summarize a review in only 20% of its size with only 0.36% drop in accuracy. Furthermore, we see that the NB classifier did perform better on our summarized reviews of 50% and 33% of the original size rather then the original review. This shows that our summarisation algorithm at first did remove sentences that

were ambiguous and would confuse a careless reader the same way they confused the NB classifier. On a letter stage, when required to cut the review even shorter, it did remove some meaning carrying sentences but still maintained the overall meaning of the document intact.

In the second column of the table we also show the results of picking the same proportions of sentences, but randomly. We see that in such process the accuracy drops tremendously, with the 20% summary dropping 9,83% as compared to the 0.36% dropped from the summary of the same size produced by our model. This experiment shows us two things. *First*, the sentences in the review do not all contain the sentiment of the review. If that was the case, random picking of sentences would show no drop. *Second*, our algorithms does a nice job at picking only the important sentences.

We further extend our experiments, by allowing the model to pick only a fixed number of sentences. We see that the performance of the model is very similar to the previous table. Picking 5 and 4 sentences did in fact increase the accuracy, while picking 3 and 2 dropped. Nevertheless the final drop in the 2 sentence summary was of only 1.01%.

| Fixed | Summary | Random | Margin |
|---|---|---|---|
| Pick All | 83.03 | 83.03 | — |
| Pick 5 | 83.07 | 80.02 | +3.05 |
| Pick 4 | 83.09 | 79.05 | +4.04 |
| Pick 3 | 82.88 | 77.15 | +5.73 |
| Pick 2 | 82.04 | 74.48 | +7.56 |
| First and Last | 68.62 | | |

**Table 3.2**

As another point of comparison, we build summaries with the common first and last heuristic. As the name suggests this heuristic creates summaries by picking the first and last sentence in the document. In this dataset, this heuristic performs badly with a drop in accuracy of almost 15% against the full document and more than 13% against the summaries of the same size produced by our algorithm. An introspection into the dataset shows that reviews usually begin with a plot summary which is irrelevant to the sentiment.

### 6.3.2 Conclusions

With our extension to the CDM we have introduced an entirely new perspective into the automatic extractive summarisation algorithms. Experiments show that our approach keeps the meaning of documents even when required to reduce its size tremendously.

Furthermore, when put through lighter reduction requirements, our model removes ambiguous sentences from the document allowing its meaning to be inferred easier.

At the current state, our summarisation approach shows to be ready for industrial application in the task of review summarisation. This is particularly useful considering the large amount of reviews a user has to skim over when trying to buy a book, a car, or book vocations. Furthermore, the results show that it has the potential to make the future state-of-the-art algorithms in the field of extractive automatic summarisation. Nevertheless, we believe this may take a while, as we also believe that the automatic summarisation community is not yet ready for Deep Learning models. The current datasets in the domain are really small to train a deep model. While producing a newer and bigger dataset is a good direction, human supervision is very expensive, especially when the task is as complicated as extracting important sentences from text.

We need to walk our half of the distance too. Our model should not rely on labelled data to train and our evaluation technique should include some human supervision. With this initial though on future directions we pass you on to its corresponding chapter.

*You can't connect the dots looking forward; you can only connect them looking backwards.*

S.Jobs

# 7 Blueprints for Future Directions

The Deep Learning approach is very young. Applied to Natural Language Processing it is even younger. In this research project we did push it further by introducing the first model to capture the compositional process that maps the meaning of words to that of documents, while preserving the distinction of sentences. As a new model, it opens way to many different applications. With the automatic summarisation, we did pick one which we thought to be the most uncertain and dangerous considering we were stepping on entirely unknown grounds. We believed in our model, and the summarisation technique proved to be fantastic. Nevertheless, the are many many more applications and improvements that can be done. In this section we will introduce a very small subset we have extensively thought about, and have a very clear idea on how they can be approached.

## 7.1 AUTOENCODERS FOR THE CDM

In our experiments we found out that our model was overfitting tremendously. For this reason we reduced its size in order to have less parameters and as a consequence less power to overfit. Nevertheless, this tremendously limits the model's ability. We found that our small CDM still achieved $3^d$ best accuracy in the IMDB Movie Review dataset, but we believe that if we have enough data to train a bigger model without overfitting, we can easily improve our score very likely producing state-of-the-art results.

In this section we describe an approach on how to overcome the data scarcity prob-

lem. In fact, during the work in this thesis we have developed this approach and it is now in a very advanced stage. It was a split decision whether we wanted to write a chapter about it, or introduce it as a future direction. The decision was based on the fact that we did not have the time we would have liked to tweak the parameters of the model in order to get the best possible score. For this reason any score we would have reported could be improved with utmost certainty.

The CDM we used for our experiment was one of the smallest models we tried. What we actually did by choosing such a small model, was to limit the modelling capabilities in order to avoid overfitting. If instead we use an unsupervised approach and effectively triple the size of the training set we use, we can definitely train a bigger model. Nevertheless to chose a different set of parameters requires a very time consuming [1] grid search over the parameters, which we could not compete given the limited time we had for this project.

### 7.1.1 The Model

In Chapter 2 we introduced two unsupervised learning approaches: the Autoencoders and the RBMs. While both approaches have effectively been used in vision they both find complications dealing with the CDM. The input to the convolutional layer in vision is the pixel bitmap, while in our CDM the input is the Word Embedding Matrix. Overall, they are very similar, but they have a fundamental difference. While the pixel bitmap is fixed and we can work on reconstructing that, the Word Embedding Matrix is optimized in the network.

If we dig deep into the Autoencoder logic we find an interesting structure. A supervised network trained for image or text classification goes over two important stages. First the input to the network $I$ is embedded via the convolution layers in a abstract low dimensional vector space $E$ $(I \rightarrow E)$. Then the low dimensional embedding is fed into a fully-connected layer to predict a label $L$ $(E \rightarrow L)$. This label together with the supervised real label is used by the cost function to produce a gradient which can be used by the network to optimise the weights. The whole process can be seen as a two way mapping $(I \rightarrow E \rightarrow L)$. In an unsupervised approach, we do not have the real label, so we need to find something else to train on. The Autoencoder circumvents this problem, by making a set of transformation in such a way that the input can also be used as a label $(I \rightarrow E \rightarrow I')$.

This shows us that we can use any transformation to achieve a label. We try to depict this in Figure 1.1. On the left side picture we show the CDM trained in a supervised fashion as we did for our experiments. The blue box depicts the input to the network,

[1] In terms of processing time. The script to start the job pool is the same we used for our first CDM optimization.

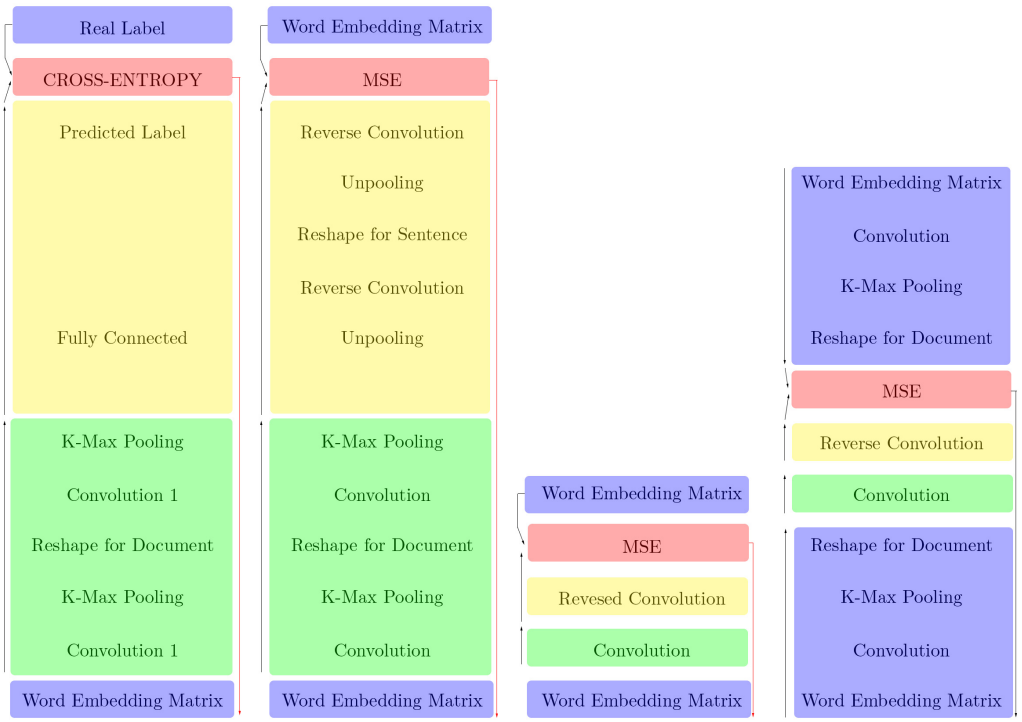



**Figure 1.1:** Left to Right - Supervised Training, Auto Encoder, Greedy Train 1, Greedy Train 2.

that in our case is the word embedding. The green box is our model, which detects features over the input and builds a low dimension vector space embedding. The yellow box depicts the technique the model uses in order to associate its output with a cost function. In this case, being a supervised training, the model uses a fully-connected layer that predicts a probability distribution over the classes (the sentiment in most of our experiment). The blue box at the top, depicts the label. Now both the output of the yellow and the top blue box is fed to the red box that is the cost function. The result is a value which is then back-propagated throughout the entire network including the input (the Word Embedding Matrix) in order to optimize it. This is depicted by the red arrow.

The next model in the picture depicts an Autoencoder applied to the CDM. The reason we use an Autoencoder is to do unsupervised learning, hence we don't have labels. For this reason, we input the output as a label. This is depicted by the blue box at the top. Our model, the same way as above produces low dimensional vector space embeddings as depicted by the green box. Nevertheless, this time we do not need a label, as we do not have one to compare to. Instead we need a reconstruction. This is achieved by the layers in the yellow box. Both the original input and the reconstruction go into the cost function (MSE) which produces a value that is than back-propagated throughout the entire network including the input (the Word Embedding Matrix) in order to

optimize it. This is depicted by the red arrow.

We carefully and purposefully depicted the reconstruction layers in yellow; the same as the fully connected layer in the supervised learning. We wanted to show, that in both cases we are trying to connect some model's understanding of reality to a cost function. While the approach we use in the Autoencoder may seem a hack, it is no more so than the fully connected layer commonly used in supervised training. Furthermore, we want this to serve as motivation for using such approach and still optimizing the input to the network, a process which in vision was not necessary given the pixel bitmap was fixed.

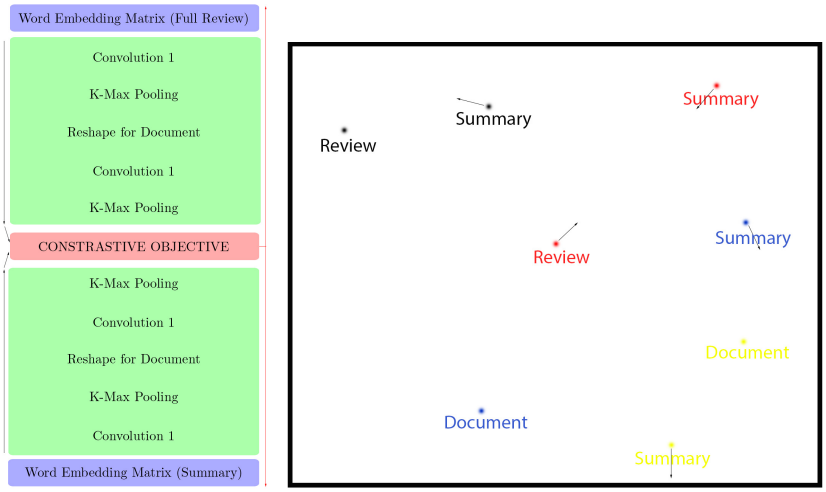
The Autoencoder for the CDM can also be trained in a greedy manner, which is the one we have currently developed. If trained in such approach, the full network should be split into 1 network for each convolution layer (layer of weights). In our example, this means it is split in two networks. The layers prior to the convolution, should also be used to provide the label to the MSE. This is depicted in the two figures on the right.

The difference we have introduced with respect to Autoencoders for vision, is the fact that we back-propagate through the input and optimize it. This allows us to learn word embeddings, but adds an important complication. The Autoencoder, as we have described it, is trying to minimize the reconstruction cost (MSE) which has a minimum value of zero where the input and the output are identical. If the model reduces the randomly started values of the Word Embedding Matrix to zeros, then the whole network will multiply with zero and output zero. As the Word Embedding Matrix serves as input and label, that will mean that our model will have a zero cost when the Word Embedding Matrix is full of zeros. This means that our model will degenerate to a useless solution. To prevent this, we can use Projected Gradient Descent, by projecting the Word Embedding Matrix to unit norm vector after each update. This prevents zeros in the Word Embedding Matrix hence its prevents weight degeneration in the model.

## 7.2 UNSUPERVISED SUMMARISATION

Experience in vision, shows that Autoencoders are really successful in learning good features in an unsupervised way. Nevertheless, in state-of-the-art algorithms they are only used to initialize the weights, which further are optimized via supervised training for a specific task. This can easily be done by using a fully-connected layer in a supervised approach. Nevertheless, for a summarisation algorithm this is not desirable. We do not want to lock ourselves in summarizing only documents for which we have classification labels. Instead we would like to follow the trend in the field of Automatic Extractive Summarisation by training a model between pairs of document-summary.

The good news is that we can easily extend the CDM to learn in this setup by using Energy-Based Learning (LeCun et al., 2006). Energy-Based models assign an energy




**(a)** Contrastive Training **(b)** Contrastive Objective

**Figure 2.1:** Contrastive Training.

to a setup configuration and try to minimize it. What we need to do is pair in some way the document and its summary, get a gradient and optimize the weights. A particularly useful model for this problem will be a contrastive objective. The idea is simple. Both the full document and the summary can be embedded via the CDM. Because they are summaries and in some abstract space they refer to the same thing, we want to train our CDM to learn this abstract space. In other words, we want the CDM to map both the full document and its summary to the same embedding. This is depicted in Figure 2.1 (a). We use the CDM to embed both the full document and the summary, feed the results to the contrastive objective which gives us a mismatch cost we can later propagate through the network in both directions. To be noted is the fact that the model that embeds the document and the summary are the same CDM.

There is however a small complication. If the network is trained in the same setup, it can learn to map everything in the same place in the multidimensional abstract space. At that position the cost will always be zero, but the model will degenerate to a useless solution as everything will be the summary of everything. To avoid this, the contrastive objective, not only tries to bring together documents with summaries that match, but also spread it apart from any other summary that does not match. While using every other summary is great, we need to embed them via the CDM which means it is a very expensive solution. Instead, in practice, we only use a sample of non-matching summaries. This is depicted in 2.1 (b). We are training the red document-summary pair while contrasting it with the blue, yellow and black summaries. The result of the gradient with this objective is depicted by the arrows, which clearly brings the red review and summary close to each other in the space, but also pushes the other summaries away. This may seem not to converge as any time we train on a pair we cause a perturbation

on the equilibrium of other pairs. Nevertheless, correctly regularised these models do learn useful solutions (Hermann and Blunsom, 2014).

## 7.3 PARAPHRASE DETECTION

The paraphrase detection task involves deciding whether two sentences have the same meaning. This is a task that can easily be approached the same way as we approached automatic summarisation above. Given a training corpus of paraphrase pairs, like the Microsoft Research Paraphrase Corpus, we can train our CSM via a contrastive objective in order to make paraphrases stay close together in the abstract space and far from the rest. In fact this is as simple, as running the same algorithm from above, but instead of using the Convolutional Document Model (CDM) with document-summary pair, we use the Convolutional Sentence Model (CSM) with a paraphrase pair.

## 7.4 QUICK SEARCH OVER BOOKS, CHAPTERS AND PARAGRAPHS

The approach we used for our CDM, not only allowed us to capture the word to document composition without losing the intermedirary representation of sentences, but also introduced a technique that can be extended to a bigger spectrum of composition in natural language. In fact it can be extended to the whole spectrum mapping the meaning of word to that of books by keeping the intermediary representation of sentences, paragraphs and chapters. This would allow to have a low dimensional vector space embedding for every word, sentence, paragraph, chapter and also the whole book, in one training.

Embedding entire books in low dimensional vector spaces can be crucial for searching, suggesting or clustering books at very low runtime. Furthermore, being extracted from the text in the book, the quality of search can be much better than search over cover features of the book. Having embedding for every paragraph or chapter can also be effectively used to suggest users reading a particular paragraph in one book, paragraphs in other books talking about the same topic. We believe this to be a really interesting topic to pursue in future research, but which will not be possible without the CDM, which in turn would be impossible to implement at this scale without the reduced size of the CSM.

*I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right.*

A.Einstein

# 8 Conclusion

This thesis project tackled the important problem of mapping the symbolic concepts of natural language to distributed representations machines can understand. Considering language follows a compositional process where simple concepts like words form more complex ones like sentences, capturing these mappings was also important. In achieving and suggesting the important advancements of this thesis project, knowledge from different sub-fields of Artificial Intelligence like Machine Learning, Computational Linguistics and Intelligent Systems had to be brought together. For this reason making a self contained report was necessary.

The results of this thesis project suggested three important advancements, which experiments showed to be extremely successful. The Convolutional Sentence Model successfully mapped the meaning of words to that of sentences, by learning distributed representations for both. It improved the Dynamic Convolutional Sentence Model by substantially reducing its number of parameters. In the experimental setup of this thesis, it resulted in 30 times less parameters. As a consequence of this reduction the CSM scales up much better, requires less time to train and also has smaller memory requirements which allow it to exploit the computation power of modern GPUs. Despite this fantastic reduction in the number of parameters and training costs, experimental results showed that the CSM had maintained the accuracy of the DCNN given the same setup.

The Convolutional Document Model exploited the reduction in training costs and memory requirements of the CSM by extending it to now map the meaning of words

to that of document. Nevertheless, the CDM is conscious of the existence of the important sentence structure between words and documents and in fact passes all the mappings through this important, very well defined, intermediary structure. Similarly to the CSM, the CDM achieves the mappings by learning distributed representations for words, sentences and documents. Experimental results, showed that the CDM is strong performer. It achieves distributed embeddings faster than other models and while it only achieved $3^d$ best accuracy on the testing dataset it only trains an smaller portion of it.

The summurization algorithm proposed in this thesis, successfully exploits the architecture of the CDM. By finding the intermediary structure of sentences, the algorithm is able to associate a saliency map with it and pick the most important ones. It is a different perspective to alternatives in the field, relying on the model's understanding of the text to extract summaries, instead of an agent based on heuristics or a classifier to decide the importance of the sentence. As the CDM currently works on labelled data, a scalable summarisation evaluation technique that exploits the existence of the label was introduced. Experiments showed the summarisation algorithm to achieve outstanding results. Not only did it summarize reviews shrinking their size to $(1/5)^{th}$ of the original one with very little loss in accuracy, but it also did clean them from ambiguous sentences improving accuracy when the shrinking requirements were more moderate.

The experimental results also identified areas of improvement to which solutions were suggested and drafted. Considering the CDM suffered from its inability to exploit large portions of the dataset it was tested on due to its inability to do unsupervised learning, an application of the Autoencoder to the CDM was suggested. The need to drop labels was also required by the summarisation algorithm, in order to enter the competition in the datasets of the field. A solution to this is offered in term of an Energy Based Model. The same model paired with a CSM instead of a CDM, also offers the perspective of Paraphrase Detection extending our reach in the Natural Language Processing domain a little further.

The introduction of the Autoencoder for the CDM included a different perspective of seeing both the fully-connected layer of a supervised setup and the reconstruction layers of an unsupervised Autoencoder setup as a way to connect the model to some cost function in order to optimize it. This is not limited to this case. Offering the reasons and inspirations behind an algorithm, a different perspective to see it, or an intuitive explanation is an aim of this thesis report. The introductory chapter paired important inventions like the Perceptron, Logistic Regression, Neural Networks and Convolutional Neural Networks together with their biological foundation. It explained optimization techniques like Gradient Descent and Newton-Raphson method with an intuitive description. Furthermore it tried to reconcile differences that exist between the machine learning jargon and the logic the algorithms are built on by showing that

the non-linearity layer in Convolutional Neural Networks is in fact part of the neurons on which they are built and only frequently considered separately in the literature due to engineering simplifications and optimizations.

Overall, the contributions of this thesis projects and its report are two-fold. First it brings together knowledge from different sub-fields of Artificial Intelligence and state-of-the-art models to propose three important advancement, two of which are novel solutions to very important problems. Second, through the theoretical background of the project it provides a soft introduction to Deep Learning including all major developments, together with their inspiration and a different perspective in understanding them.

# Appendices

# A

# IMDB Movie Review Summaries

In this appendix we have randomly selected a list of examples produced from our summarization algorithm in the Maas et al. (2011) dataset. We have shown both the full movie review with the important sentences highlighted and also the summarized review. There certainly is some redundancy in doing so, but we wanted to ephasize both usages of our model. In the first one, showing the whole review but attracting the users attention to the important sentences, while in the second one showing only the summary. This section includes examples from both the train and the test set.

## A.1 TRAIN SET

| # | ORIGINAL REVIEW | SUMMARY |
|---|---|---|
| 1 | Samuel Fuller is an interesting filmmaker , mainly because he had some very inconsistent politics in his films . While " Shock Corridor " and " The Naked Kiss " represented the hypocrisies and lunacy of America and " The Big Red One " was an effective portrait of the horrors of war , " Merrill ' s Marauders " painted war as necessary hell and " Pickup on South Street " is about the dangers of communist spies . All of his films make for very entertaining viewing , and even though he was often pigeonholed as a b - filmmaker , Fuller was just as good as any of the major studio contractors . " Pickup on South Street " is no exception , and despite the dated themes , the film - making style is remarkably ahead of its time . Its also a very quickly - pace , tight , and occasionally brutal film noir . The acting across the board is fantastic . Richard Widmark makes for a great anti - hero and Jean Peters is quite sexy as a girl who works for her communist spy boyfriend . The show stealer is Thlema Ritter however , in an absolutely delightful performance as a police stoolie . The angles Fuller employs are great , making the acting sequences all the more exciting and brutal ( this is very violent for its time ). The camera continuously moves around just as Tarantino and his school would do forty years later . " Pickup on South Street " is a great action - paced noir thriller . " Shock Corridor " remains my favorite Fuller film , but this is a very close second . ( 8 / 10 ) | The show stealer is Thlema Ritter however , in an absolutely delightful performance as a police stoolie . " Pickup on South Street " is a great action - paced noir thriller . " Shock Corridor " remains my favorite Fuller film , but this is a very close second . |

| | | |
|---|---|---|
| **2** | If you haven ' t seen the gong show TV series then you won ' t like this movie much at all , not that knowing the series makes this a great movie . I give it a 5 out of 10 because a few things make it kind of amusing that help make up for its obvious problems . 1 ) It ' s a funny snapshot of the era it was made in , the late 1970 ' s and early 1980 ' s . 2 ) You get a lot of funny cameos of people you ' ve seen on the show . 3 ) It ' s interesting to see Chuck ( the host ) when he isn ' t doing his on air TV personality . 4 ) You get to see a lot of bizarre people doing all sorts of weirdness just like you see on the TV show . I won ' t list all the bad things because there ' s a lot of them , but here ' s a few of the most prominent . 1 ) The Gong Show Movie has a lot of the actual TV show clips which gets tired at movie length . 2 ) The movie ' s story line outside of the clip segments is very weak and basically is made up of just one plot point . 3 ) Chuck is actually halfway decent as an actor , but most of the rest of the actors are doing typical way over the top 1970 ' s flatness . It ' s a good movie to watch when you don ' t have an hour and a half you want to watch all at once . Watch 20 minutes at a time and it ' s not so bad . But even then it ' s not so good either . ;) | If you haven ' t seen the gong show TV series then you won ' t like this movie much at all , not that knowing the series makes this a great movie . 1 ) It ' s a funny snapshot of the era it was made in , the late 1970 ' s and early 1980 ' s . 2 ) The movie ' s story line outside of the clip segments is very weak and basically is made up of just one plot point . |
| **3** | I won ' t add to the plot reviews , it ' s not very good . Very improbable orphanage on Bala . Cushing and Lee at their height . Some nice scenery . Good for face spotting , and I quote , " look at the mouth , that is Cassie from Fools and Horses ". Otherwise , a poor example of the British film industry . Fulton MacKay was far better in Fraggle Rock , Keith Barron was better in anything else and Diana Dors did what she did best . Redeeming feature ? It was free to watch on the Horror channel prior to its going over to subscription . I won ' t be subscribing on this effort . | Otherwise , a poor example of the British film industry . Redeeming feature ? |
| **4** | Moonwalker is such a great movie , from start to finish you cant take your eyes away . i love all the clips of Michael singing and dancing and I just love the ' studio tour ' bit ... soo funny :) And the ' mini movie ' is to cool , with all the special FX etc ... Michael is a genius and always will be !! ! | Moonwalker is such a great movie , from start to finish you cant take your eyes away . |
| **5** | I show this film to university students in speech and media law because its lessons are timeless : Why speaking out against injustice is important and can bring about the changes sought by the oppressed . Why freedom of the press and freedom of speech are essential to democracy . This is a must - see story of how apartheid was brought to the attention of the world through the activism of Steven Biko and the journalism of Donald Woods . It also gives an important lesson of free speech : " You can blow out a candle , but you can ' t blow out a fire . Once the flame begins to catch , the wind will blow it higher . " ( From Biko by Peter Gabriel , on Shaking the Tree ). | Once the flame begins to catch , the wind will blow it higher . |
| **6** | I lived in Tokyo for 7 months . Knowing the reality of long train commutes , bike rides from the train station , soup stands , and other typical scenes depicted so well , certainly added to my own appreciation for this film which I really , really liked . There are aspects of Japanese life in this film painted with vivid colors but you don ' t have to speak Japanese to enjoy this movie . Director Suo ' s tricks were subtle for the most part ; I found his highlighting the character called Tamako Tamura with a soft filter , making her sublime , a tiny bit contrived but most of the directors tricks were so gentle that I was fully pulled in and just danced with his characters . Or cried . Or laughed aloud . Wonderful . A +. | Knowing the reality of long train commutes , bike rides from the train station , soup stands , and other typical scenes depicted so well , certainly added to my own appreciation for this film which I really , really liked . Or cried . |
| **7** | As an avid fan of the Flashman books by George McDonald Fraser , I looked forward immensely to seeing Flashy on the big screen when this film was first released . Sadly it was a huge disappointment then - so I left it alone for 20 years before going back to watch it again , but it was no better the second time . Mr Fraser is a tremendously skillful writer , but I am not a fan of his film screenplay work with Richard Lester . The penchant for slapstick spoilt ' The Three Musketeers ' for me and the same applies here . To me , the whole tone and feel of the film is wrong . The Flashman books are uproariously funny in parts , but they are adventure novels . There is much seriousness in the way the adventures that Flashman has - after all , he is involved in dangerous situations . This is conveyed in the novels , but not conveyed at all on film due to the its comedic style . It is a tremendous shame as it could have a great film had it been a more faithful adaptation of the style of the book . When I first read that the book was to be filmed , the article said that the film was to star Oliver Reed . I rejoiced , as Reed to me | Sadly it was a huge disappointment then - so I left it alone for 20 years before going back to watch it again , but it was no better the second time . To me , the whole tone and feel of the film is wrong . A great shame , as the production values , costumes , sets etc are superb and the casting is generally excellent - just about everybody in the film is well cast apart from Malcolm McDowell . Possibly the directorship of Richard |

| | | |
|---|---|---|
| | was the epitome of Flashman . How I would have loved to see him in the role . Malcolm McDowell is a good actor , but does not fit the visual image of Flashman created by the books ( too scrawny looking ! Flashman is supposed to be a big strapping fellow ). Neverheless Reed was excellent as Bismarck . What kills the film is that it is made as a comedy . The only scene in which it creates the true atmosphere of the book is the scene in which Flashman kills de Gautet ( Tom Bell ). A great shame , as the production values , costumes , sets etc are superb and the casting is generally excellent - just about everybody in the film is well cast apart from Malcolm McDowell . Possibly the directorship of Richard Lester was responsible for the way the film is , as a recent radio adaptation of ' Flash For Freedom ', adapted by Mr Fraser , worked quite well . Perhaps one day we may see Flashman done justice on screen . | Lester was responsible for the way the film is , as a recent radio adaptation of ' Flash For Freedom ', adapted by Mr Fraser , worked quite well . Perhaps one day we may see Flashman done justice on screen . |
| **8** | First of all , Riget is wonderful . Good comedy and mystery thriller at the same time . Nice combination of strange ' dogma ' style of telling the story together with good music and great actors . But unfortunately there ' s no ' the end '. As for me it ' s unacceptable . I was thinking ... how it will be possible to continue the story without Helmer and Drusse ? ... and I have some idea . I think Lars should make RIGET III a little bit different . I ' m sure that 3rd part without Helmer wouldn ' t be the same . So here ' s my suggestion . Mayble little bit stupid , maybe not . I know that Lars likes to experiment . So why not to make small experiment with Riget3 ? I think the only solution here is to create puppet - driven animation ( like for example " team America " by Trey Parker ) or even computer 3d animation . I know it ' s not the same as real actors , but in principle I believe it could work ... only this way it ' s possible to make actors alive again . For Riget fans this shouldn ' t be so big difference - if the animation will be done in good way average ' watcher ' will consider it normal just after first few shots of the movie . The most important thing now is the story . It ' s completely understandable that it ' s not possible to create Riget 3 with the actors nowadays . So why not to play with animation ? And ... look for the possibilities that it gives to you ! Even marketing one ! Great director finishes his trilogy after 10 years using puppet animation . Just dreams ? I hope to see Riget 3 someday ... or even to see just the script . I ' m curious how the story ends ... and as I expect - everybody here do . greets , slaj ps : I ' m not talking about the " kingdom hospital " by Stephen King ;-) | First of all , Riget is wonderful . Good comedy and mystery thriller at the same time . Nice combination of strange ' dogma ' style of telling the story together with good music and great actors . I hope to see Riget 3 someday ... or even to see just the script . I ' m curious how the story ends ... and as I expect - everybody here do . greets , slaj ps : I ' m not talking about the " kingdom hospital " by Stephen King ;-) |
| **9** | This very low budget comedy caper movie succeeds only in being low budget . Dialog is dumbfoundingly stupid , chase scenes are uniformly boring , and most of the on - screen money seems to have been saved for a series of crashes and explosions in a parking lot during the film ' s last five minutes ( a briefly glimpsed port - a - potty early in that scene is certain to wrecked and spew crap on the film ' s chief villain -- no prop is here without a purpose ). The whole film is depressingly reminiscent of those that occasionally came out of Rodger Corman ' s studio when he ' d give a first time director a few bucks and a camera -- but without the discipline Corman would impose . | Dialog is dumbfoundingly stupid , chase scenes are uniformly boring , and most of the on - screen money seems to have been saved for a series of crashes and explosions in a parking lot during the film ' s last five minutes ( a briefly glimpsed port - a - potty early in that scene is certain to wrecked and spew crap on the film ' s chief villain -- no prop is here without a purpose ). |
| **10** | Super Mario 64 is undoubtedly the greatest game ever created . It is so addicting that you could play it for hours upon hours without stopping for a break . I ' ve beaten the game 4 times , but I ' ve never gotten all 120 stars ...( I ' ve gotten 111 )... but I hope to achieve them eventually . Even though I didn ' t officially play this game until I was seven in , I loved watching my sisters play it . Now I am 13 and still play this , erasing games and starting over again . The graphics are unbelievable for an early N64 game . The gameplay is addictive . The controls are great . The levels are tough , but not impossible . The Bowser fights are challenging . I would like to tell you more , but why don ' t you just get it for yourself ? Put the X - BOX 360 , PS3 , and the Wii away and go find yourself a Nintendo 64 and play this amazing , wonderful game . | Even though I didn ' t officially play this game until I was seven in , I loved watching my sisters play it . The controls are great . Put the X - BOX 360 , PS3 , and the Wii away and go find yourself a Nintendo 64 and play this amazing , wonderful game . |
| **11** | WOW . If you think that a film can ' t fatigue in some way , then you haven ' t seen Dog Bite Dog . This film pulls no punches , and it doesn ' t shy away | While the script offers nothing new on the surface , it does |

| | | |
|---|---|---|
| | from showing very disturbing images at all . Much like Salo , this one shows us the dehumanization of the human spirit . It is gritty , dark , depressing and hopeless , but it is also one of the best films to ever come out of Hong Kong . The script is much more of the same , but don ' t go on thinking it is incredibly clichd . It basically is about a troubling and obsessive detective in a cat and mouse game , against a professional and emotionless hit - man . While the script offers nothing new on the surface , it does provide a lot of questions about the dark side of humanity . Is violence really that necessary ? Do we become more or less human when we abuse a 5 year old child , without pity , without remorse ? In turn , we humans act no less than rabid dogs when we are blinded by anger , this is a sad truth . It is a topic that the director brilliantly explores , without limiting himself at all . Besides the cat and mouse chase , the script also develops two separate story lines for the main characters . One is about love , and the other is about redemption . Even if the script isn ' t that new , it is still wonderfully written and it keeps you glued to the seat at all times . The acting is really , really good . Edison Chen as the Hit - man is incredible ; he proves that he isn ' t just any pretty face . He is ruthless , vile and beyond likable . Sam Lee as the obsessed cop is also outstanding . The supporting cast , in short , is excellent . The music is also worth mentioning . Very somber score by Ben Cheung , with some effective light hearted songs played at key dark moments in the film . The cinematography by Yuen Man is also really good . Overall , this CATIII film is highly recommended . Very well paced , incredibly acted , marvelously scored and just really good at the end of it all . However , as many have pointed out , this is not a movie for everyone . If you dislike strong violence then you should stay away from this one . If you don ' t like seeing heavy negativity in film then this isn ' t for you too . In the end , a powerhouse film , 8 / 10 . | provide a lot of questions about the dark side of humanity . Do we become more or less human when we abuse a 5 year old child , without pity , without remorse ? Even if the script isn ' t that new , it is still wonderfully written and it keeps you glued to the seat at all times . The music is also worth mentioning . The cinematography by Yuen Man is also really good . Overall , this CATIII film is highly recommended . Very well paced , incredibly acted , marvelously scored and just really good at the end of it all . |
| **12** | With a cast list like this one , I expected far better . Venessa Redgrave spent the majority of the movie lying in bed . The best actresses in the world cannot make anything very interesting when their acting is limited to lying down and falling asleep throughout the entire movie . The plot summary says that a secret is revealed to the daughters as their mother comes closer to death . The thing is , she never tells her daughters anything except cryptic advice to be happy . All the relationships in the movie are underdeveloped . I also felt that the back and forth between the past and present was unnecessary . It seemed as if the idea was stolen either from the book the Da Vinci Code in which the device was used to increase suspense , or from The Notebook in which they used the device to create the never ending romance of the story ' s main characters . Either way it was a cheap device in this movie because it didn ' t work to create anything . It was a way to attempt suspense in a movie that has none . I left wondering why good movies can ' t be written for women . It really was a disappointment . | It seemed as if the idea was stolen either from the book the Da Vinci Code in which the device was used to increase suspense , or from The Notebook in which they used the device to create the never ending romance of the story ' s main characters . Either way it was a cheap device in this movie because it didn ' t work to create anything . It really was a disappointment . |
| **13** | This is one of the best reunion specials ever , with Adam West and Burt Ward parodying themselves and having fun while doing it . It ' s amazing the amount of effort that went into the detail , particularly recapturing the feel of the 1960 ' s era , the Batcave set , Wayne Manor , the costumes , and the actors selected to play the younger versions of West , Ward , Burgess Meredith , Cesar Romero , and Frank Gorshin ! This 90 minutes is well worth your time , and is a delight to all fans of the classic 1960 ' s " Batman " television series . I note that clips from " Batman " were from the movie , and not the series itself , probably because of legal restrictions . Let ' s hope the three seasons of the show are forthcoming on DVD . | This is one of the best reunion specials ever , with Adam West and Burt Ward parodying themselves and having fun while doing it . |
| **14** | Grey Gardens is a world unto itself . Edith and Little Edie live in near total isolation , eating ice cream and liver pate in a makeshift kitchen in their ( apparently ) shared bedroom . Cats loll about while mother Edith insults her daughter ' s elocution . This is a Tennessee Williams play come to life and should inspire screenwriters and playwrights , as the bizarre and overlapping dialogue is 100 % real . The situation in the house reminds me exactly of how my grandmother and her 50 - ish daughter lived for a decade ( other than that they were poor and clean ). They would bicker all day , grandmother talking about her gloriously perfect past while her daughter continually blamed her for missed opportunities with men , work , and self - | They would bicker all day , grandmother talking about her gloriously perfect past while her daughter continually blamed her for missed opportunities with men , work , and self - expression . It is rare to see true life this way and all the more special considering the context -- |

| | | |
|---|---|---|
| | expression . This film is a must - see for anyone writing a mother / daughter relationship of this kind . It is sad and voyeuristic , but the filmmakers did an amazing job getting the Edies comfortable enough to expose themselves so recklessly . It is rare to see true life this way and all the more special considering the context -- remnants of a powerful family fading into nothingness in the skeleton of their own mansion . | remnants of a powerful family fading into nothingness in the skeleton of their own mansion . |
| **15** | This must be one of the most overrated Spanish films in history . Its lack of subtlety and complexity and its total political correction make it really childish , with only good / bad characters . The world is just not like this , and good movies show complex characters with opposite impulses , dilemmas , etc . However , what I HATE most about this film is Bola ' s friend ' s father . The director tries to teach us a good lesson : tattoo artists with shaved heads are not always bad guys , in fact they can be better than the average looking dad ( wow , this is like ... philosophy , or something ). Thank you , Achero . I ' ll propose you for the Nobel prize of literature . | Its lack of subtlety and complexity and its total political correction make it really childish , with only good / bad characters . |
| **16** | I don ' t know what the rest of you guys watch Steven Seagal movies for , but I watch them because , as silly as they are , they ' re at least always good for a laugh . Why would you rate this movie a 1 out of 10 based on the dubbing , when that kind of thing is exactly what makes a movie like this into a cult favorite that you can laugh at the silliness of ? Attack Force is by no means a great movie , but I felt it was as worthy a Steven Seagal vehicle as many of his other movies ; in fact I didn ' t think it was one of his worst by a long - shot . It had , most of the time , a half - way coherent plot line , and it was , most of the time , fundamentally exciting . The ending really sucked , but even that had some enjoyably trashy elements . In the end the story itself did not deliver what it promised , but I actually thought that the acting , characterization ( if I may use such a big word ) and the rest of the production values delivered exactly what a true Steven Seagal fan would expect . Seagal himself in particular was exactly the stone - faced , no - nonsense man ' s man that we ' ve come to expect , and the rest of the cast backed him up pretty well , without ever up - staging him . This , people , is what a Steven Seagal movie does . Deal with it . Or even better : laugh at it . 4 out of 10 . | Attack Force is by no means a great movie , but I felt it was as worthy a Steven Seagal vehicle as many of his other movies ; in fact I didn ' t think it was one of his worst by a long - shot . This , people , is what a Steven Seagal movie does . |
| **17** | One of the first OVA ' s (" original video animation ") I ever bought , this still has to be one of my favourite anime titles . A cyberpunk sci - fi action comedy set against an unlikely ( for a comedy , that is ) background of near - future pollution in a dystopian society . The " heroes " of Dominion are the Tank Police , formed with a " if we can ' t beat crime , we ' ll get bigger guns " philosophy , and who are , like the name suggests , patrolling the city in tanks instead of patrol cars , and who are actually far more dangerous than any criminals they are trying to catch . Most , if not all , of these cops are borderline (? ) psychopaths and neurotics , giving new meaning to the phrase " loose cannons ". Equally colourful and amusing are their adversaries , terrorist Buaku and his hench ( wo ) men , the Twin Cat Sisters , whose existence always seems to involve giving the Tank Police a hard time . The animation is not state of the art , but it ' s very nice otherwise ; the colourful palette and cartoonish look of the characters and mecha fit nicely with the comedic atmosphere of Dominion . The English dubbing is , again , lots of fun . The soundtrack of the English version is also very good . I wonder if they ever made a soundtrack album of that ... Anyway , Dominion Tank Police is great . It ' s Japanese cyberpunk SF with lots of comedy , filled with completely over - the - top characters and situations , making sure that it never takes itself seriously . Highly recommended . | The " heroes " of Dominion are the Tank Police , formed with a " if we can ' t beat crime , we ' ll get bigger guns " philosophy , and who are , like the name suggests , patrolling the city in tanks instead of patrol cars , and who are actually far more dangerous than any criminals they are trying to catch . The soundtrack of the English version is also very good . Highly recommended . |
| **18** | Excellent view of a mature woman , that is going to lose everything ( even the pruner has a mortgage ). The way she gets involved into this special " business ", the innocence , and the true love that exists between the people of a little town , it ' s mixed perfectly to give us as result a fresh , light and funny comedy . I couldn ' t stop laughing with a very funny scene of two old ladies in a drugstore . I love European films , and with movies like this one , my opinion grows stronger . A movie that I also recommend with my eyes closed , in this same genre , is Waking Ned Devine . Saving Grace , a comedy that many friends enjoyed as much as myself . You will love it . | The way she gets involved into this special " business ", the innocence , and the true love that exists between the people of a little town , it ' s mixed perfectly to give us as result a fresh , light and funny comedy . |
| **19** | A mummy narrates vignettes about men , women , and the sex between them | If you ignore ( or fast forward |

| | | |
|---|---|---|
| | . Huh ? At the beginning , the mummy randomly asks the viewer , " Imagine having sex with this girl . Imagine having sex with this boy " about 37 times , while flashing pictures of half naked mod youths . Later , said mods boys pelt mod girls with ... vegetables ? If you ignore ( or fast forward ) through the mummy ' s rambling , the shorts aren ' t bad in their own right . I found a few of them rather funny . My personal favorite is one where the sexually - confused man tries to convince a girl to have sex with him while his pet lizard sits on the bed . This is one , well , bizarre movie . | ) through the mummy ' s rambling , the shorts aren ' t bad in their own right . This is one , well , bizarre movie . |
| **20** | This , the direct - to - video death rattle of the Tremors series , features sixty inspired seconds ( sawblade : you ' ll know it when you see it ) and more tedium and filler than you can shake a stick at . Tremors 4 was obviously shot on a cripplingly low budget . That means they only had enough special effects mojo for three or four minutes of precious worm - on - human violence , tops . The lackluster , cliche - spouting cast and hackneyed writing ensure that the remaining hour and a half of the Tremors 4 experience feels at least fifteen thousand years long . Only hardcore Tremors fans will be able to sit through , much less enjoy this film . If you aren ' t among them - don ' t bother . | That means they only had enough special effects mojo for three or four minutes of precious worm - on - human violence , tops . |
| **21** | The dialogue was pretty dreadful . The plot not really all that inspired beyond the obvious twist it presents . Not visually stunning . Actually visually annoying at times . Most definitely one of those films you find easier to finish if you keep one finger on the fast forward button . If you could watch it for free , have absolutely no other options open at the moment and you really dig seeing the little poltergeist lady ... well maybe I ' d recommend it to you , but not anyone else I could think of at the moment . | The dialogue was pretty dreadful . |
| **22** | This one is a real bomb . We are supposed to believe that Merle Oberon is the sequestered daughter of an ambitious politician who must prove to the Tom DeLay of the 1930s that he is worth supporting as a presidential candidate . Poor Merle can ' t go anywhere , but is surrounded by politicians and their quacking , quaking wives and supported only by kindly uncle Harry Davenport . She joins her two maids on a blind date and Gary Cooper happens to show up . Some shots of rodeo might have enlivened things , a la " Misfits ," but no such luck with this one . Gary later breaks in to a formal dinner , at which Merle is presiding , and , though invited to sit down and join the group , reads them a lecture on their snobbery . Where did this diffident cowboy ' s sudden eloquence come from ? The most excruciating scene in the film is a phantom party that Gary holds in his unfinished house for his absent wife , Merle . Will it never end ? One to avoid . | We are supposed to believe that Merle Oberon is the sequestered daughter of an ambitious politician who must prove to the Tom DeLay of the 1930s that he is worth supporting as a presidential candidate . One to avoid . |
| **23** | When the employees of a theater find an old reel of film , they decide to show it at the midnight screening of Night of the Living Dead , assuming it ' s an old preview reel . Unfortunately , it ' s actually an old Nazi mind control experiment that turns the audience into a horde of mindless shuffling zombies . I can ' t understand the hate for this movie . It is a low budget independent production with a lot of camp , but it doesn ' t deserve a " 1 . 1 " here on IMDb . It is just so much fun . It is obvious that the filmmakers have a reasonable knowledge and love of old horror movies , and they have created an entertaining tribute to them sprinkled with references and homages to a variety of them . It has the feel of such things as Night of the Living Dead ( in many ways , very similar ), Evil Dead 2 and Army of Darkness , and various others . I liked the explanation of how the zombies , though really just hypnotized into thinking they are zombies , actually come to have the physical attributes of the living dead - unbelievable , perhaps , but I appreciate the effort by the writers to explain it . The gore effects were decent for the budget , the acting was all right , and the story was entertaining . I liked it . | The gore effects were decent for the budget , the acting was all right , and the story was entertaining . I liked it . |
| **24** | I was the Production Accountant on this movie , and I also got to do some voice - over work on it , so I ' m not entirely unbiased , but if it were awful , I would say so . I thought it was a fun film , not a critically acclaimed masterpiece , by any means , but there were plenty of laughs along the way . The Bible states that laughter does good like a medicine , so watching this movie could be good for your health . So many of the actors in this picture hadn ' t yet reached their peak at the time we made this film . Susan Sarandon , of course , is one who has since gone on to much greater fame . | I was the Production Accountant on this movie , and I also got to do some voice - over work on it , so I ' m not entirely unbiased , but if it were awful , I would say so . Robert Englund later became known as Freddie Krueger , |

| | | |
|---|---|---|
| | Melanie Mayron was seen on TV on a weekly basis as a photographer in the " Thirty - Something " TV drama series . Robert Englund later became known as Freddie Krueger , still haunting people ' s dreams . One of my personal favorite actors on this show was Dub Taylor , who played the sheriff . He was an excellent comedic actor , and a truly nice , sincere person . We all had fun working on this show , and I think that fun comes through . | still haunting people ' s dreams . |
| **25** | The Theory Of Flight is an engaging character study of an artist ( Branagh ) yearning to break free of boredom and mediocrity , and a terminally ill patient ( Bonham - Carter ) in the last stages of ASL , confined to a wheelchair , who desires to make love to a man before dying . Helena Bonham - Carter exudes wit , defiance , and independence as an ASL patient who is virtually dependent upon people around her to take care of her . Kenneth Branagh , sentenced through community service to take part in caring for her , complements Helena ' s charm with woeful melancholy , creating a sentimental , compelling love story in which two people try to help each other find the road to happiness , before time runs out . | Kenneth Branagh , sentenced through community service to take part in caring for her , complements Helena ' s charm with woeful melancholy , creating a sentimental , compelling love story in which two people try to help each other find the road to happiness , before time runs out . |
| **26** | An Insomniac ' s Nightmare is the story of a man ' s plunge into insanity . Having chronic Insomnia , Jack is plagued by hallucinations ; causing him to try and determine what is real and what isn ' t . We find out interesting things about Jack near the end , and think that by the time the movie is over we will have a " happily ever after " Hollywood ending . Wrong . This is New York City , the place where nobody sleeps . Tess Nanavati ( Writer and Director ) has herself a good film in ' An Insomniac ' s Nightmare '. A talented filmmaker and writer ( she made this film right after her High School Graduation ), she has real potential and will be one to watch in the upcoming future . As I watched this short film I was constantly uncomfortable ; between the music , bleak scenery , and realistic portrayal of an insomniac by Dominic Monaghan ( as Jack ), I desperately wanted to turn this off at times just to escape from it . | Wrong . As I watched this short film I was constantly uncomfortable ; between the music , bleak scenery , and realistic portrayal of an insomniac by Dominic Monaghan ( as Jack ), I desperately wanted to turn this off at times just to escape from it . |
| **27** | The often - reliable Leonard Maltin says this is a " delightful romance " and that Sanders is " superb . " Maltin must have confused this movie with something else . Sanders is snide and droll and superb , as usual , you can imagine his delivery of the line regarding adultery , " Sometimes the chains of matrimony are so heavy they have to be carried by three ," but dull , wooden and dated describe this movie more accurately . The storyline itself , an autobiography with Sanders as a suave jewel thief , Francois Eugene Vidocq , who becomes chief of police but can hardly resist the lure of fine jewels , is entertaining enough , but it has the same kind of hollow historical Hollywood treatment that marred such period epics as * Marie Antoinette *, and certainly the deplorable * Forever Amber * ( which screams for a classy remake ). Though , in his defense , Sanders tries mightily to add some depth to his character , it is all for naught . I am an unabashed Douglas Sirk fan , but this is 1946 , and it is one of Sirk ' s earliest American efforts , lacking many of the signature touches that would define his florid , breast - heaving potboilers . Sirk is just getting his feet wet here , and made a number of unmemorable films over the next ten years until he struck gold with * Magnificent Obsession *, and hit his stride , bombarding us with such estrogen - fests as * All That Heaven Allows *, * Written on the Wind *, and * Imitation of Life *. But * Scandal In Paris * is hardly his best work a relatively low - budget affair with cheesy sets and ineffective costuming . | The often - reliable Leonard Maltin says this is a " delightful romance " and that Sanders is " superb . Sanders is snide and droll and superb , as usual , you can imagine his delivery of the line regarding adultery , " Sometimes the chains of matrimony are so heavy they have to be carried by three ," but dull , wooden and dated describe this movie more accurately . |
| **28** | Opening credits : great . Music : just right for this film . Cinematography : sleazy to great effect . Harrowing excitement in a train - wreck sort of way . This is how out of control some lives can become . " Wonderland " depicts drug - induced wildness ( and its consequences ) not as an aberration , but shows how it really does happen in our society . It is better than depictions of another wild group - Manson ' s - because all Manson films , books etc concentrate so much on Manson ' s insane mind rather than the overall picture of the cheap fame and randomness that is so pervasive nowadays . If you want to see some of the best " Method - acting " on film , watch Kilmer in this movie . He shows how " The Method " can be riveting in the right role . The filmmakers here succeed in raising " out - of - control " to the level of an art form , not just for the sake of giving us cheap thrills . | If you want to see some of the best " Method - acting " on film , watch Kilmer in this movie . He shows how " The Method " can be riveting in the right role . |

## A.2 TEST SET

| # | ORIGINAL REVIEW | SUMMARY |
|---|---|---|
| 1 | A very , very , very slow - moving , aimless movie about a distressed , drifting young man . Not sure who was more lost - the flat characters or the audience , nearly half of whom walked out . Attempting artiness with black & white and clever camera angles , the movie disappointed - became even more ridiculous - as the acting was poor and the plot and lines almost non - existent . Very little music or anything to speak of . The best scene in the movie was when Gerardo is trying to find a song that keeps running through his head . He goes to a used record store to buy it for his lover and has to sing the song for two sales clerks before they find the album . Cute scene gave promise , but it went downhill from there . The rest of the movie lacks art , charm , meaning ... If it ' s about emptiness , it works I guess because it ' s empty . Wasted two hours . | Not sure who was more lost - the flat characters or the audience , nearly half of whom walked out . Attempting artiness with black & white and clever camera angles , the movie disappointed - became even more ridiculous - as the acting was poor and the plot and lines almost non - existent . |
| 2 | Saw the movie today and thought it was a good effort , good messages for kids . A bit predictable . The book was better , gave more plot details , ore about the environment and how the kids uncovered the conspiracy . I think Hiassen ' s warped humor comes across better in the book than the movie , but there were lots of funny moments in the movie as well . It is probably a bit too slow paced for kids under 6 yrs of age . Loved the casting of Jimmy Buffet as the science teacher . And those baby owls were adorable . I wonder how they managed to film them . The movie showed a lot of Florida at it ' s best , made it look very appealing . Am I imagining it , or did the author Carl Hiassen make a brief appearance ? | Saw the movie today and thought it was a good effort , good messages for kids . Am I imagining it , or did the author Carl Hiassen make a brief appearance ? |
| 3 | I ' ve Seen The Beginning Of The Muppet Movie , But Just The Half . Because I Only Watched It At Mrs Kelly ' s Friend ' s House . The Songs Were The Best And The Muppets Were So Hilarious . They Learn That If They Believe In The End Of The Rainbow , Anyone Can Make It , No Matter How Small , No Matter How Green ( Which Was Included In The Trailer ). Kermit Is My Favorite Protagonist ( Which Means It Describes The Main Character ) And So Are The Other Muppets . Mel Brooks Was Amazing When He Played Professor Max Krassman . The Scene Where Miss Piggy Saves Kermit By Doing Kung Fu On Those Guys . It Was So Cool . The Muppet Movie Is The Best Jim Henson Film With The Most Hilarious Characters And People Will Cherish For His Successful Film . | Kermit Is My Favorite Protagonist ( Which Means It Describes The Main Character ) And So Are The Other Muppets . The Muppet Movie Is The Best Jim Henson Film With The Most Hilarious Characters And People Will Cherish For His Successful Film . |
| 4 | This is a very " right on case " movie that delivers everything almost right in your face . I ' m a Christian and liked the film in one way . It had some average acting from the main person , and it was a low budget as you clearly can see . It can be a bit long - winded , but the film has some quite nice cars that rescues it from a lower rating from me . As a Christian film it was quite good , but maybe a bit right - on in the message . The film works best on a big screen . * SPOILERS * The fighting scene with the two brothers can remind you of the fighting scene between the two brothers in the Christian thriller " Mercy Streets " starring Eric Roberts . * End of Spoiler * I give it a 7 / 10 . | As a Christian film it was quite good , but maybe a bit right - on in the message . The film works best on a big screen . |
| 5 | This short film certainly pulls no punches . The story is of a butcher who wrongfully kills an innocent man who he believes has sexually molested his retarded daughter . The film goes onto depict how the butcher serves his time , and returns to life with his daughter in care , and having to come to terms with a life with no future . The graphic opening scenes of a horse being slaughtered , and the full frontal birth of the butchers daughter puts you a brutal frame of mind that stays with you throughout the film . The snappy flow of the film is very direct and adds to its brutality . Consequently alot of ground is covered in the 40 minutes . You are taken in fully with the butchers non - life - particularly after he loses his daughter to social services and his business . His story continues in the excellent film Seul Contre Tous | The snappy flow of the film is very direct and adds to its brutality . His story continues in the excellent film Seul Contre Tous |
| 6 | Graphics is far from the best part of the game . This is the number one best TH game in the series . Next to Underground . It deserves strong love . It is an insane game . There are massive levels , massive unlockable characters ... it ' s just a massive game . Waste your money on this game . This is the kind of money that is wasted properly . And even though graphics suck , thats doesn ' t | This is the number one best TH game in the series . It deserves strong love . It is an insane game . Waste your money on this game . This is the kind of |

| | | |
|---|---|---|
| | make a game good . Actually , the graphics were good at the time . Today the graphics are crap . WHO CARES ? As they say in Canada , This is the fun game , aye . ( You get to go to Canada in THPS3 ) Well , I don ' t know if they say that , but they might . who knows . Well , Canadian people do . Wait a minute , I ' m getting off topic . This game rocks . Buy it , play it , enjoy it , love it . It ' s PURE BRILLIANCE . | money that is wasted properly . |
| **7** | Although it doesn ' t seem very promising for a long stretch , Renoir ' s French Cancan ends up being an effortlessly charming film . The story is clich : a laundry girl , Nini ( Franoise Arnoul ), is discovered by a night club owner , Danglard ( Jean Gabin ). Danglard steals her from her baker boyfriend and drops his current girlfriend , both of whom come back for their former lovers . Nini has to choose whether to go back to her humble life with the baker , go on with the show with her employer , oh , or become a princess , as a prince falls in love with her at one point , too . I ' m glad the film didn ' t go for the most obvious choice , as a lesser film certainly would have . The film ends with the opening of Danglard ' s new night club , the Moulin Rouge , and a couple of gorgeous song and dance numbers . The first of them , " Complainte de la Butte ," which also provides the base of most of the film ' s musical score , is simply one of the most gorgeous songs ever written , and Renoir himself wrote it . If you ' re a fan of Baz Luhrmann ' s 2001 film Moulin Rouge !, you ' ll recognize the tune , as it comes up near the beginning of that film , sung by Rufus Wainwright . Although it isn ' t very prominent in that film , everyone I know who owns the soundtrack loves it . In addition to having one of the most lovely songs ever written , French Cancan also boasts one of the cutest leading ladies ever to grace the screen . It ' s hard not to fall head - over - heels in love with that girl . 8 / 10 . | The first of them , " Complainte de la Butte ," which also provides the base of most of the film ' s musical score , is simply one of the most gorgeous songs ever written , and Renoir himself wrote it . In addition to having one of the most lovely songs ever written , French Cancan also boasts one of the cutest leading ladies ever to grace the screen . 8 / 10 . |
| **8** | On the positive , I ' ll say it ' s pretty enough to be watchable . On the negative , it ' s insipid enough to cause regret for another 2 hours of life wasted in front of the screen . Long , whiny and pointless . And I ' m not saying this to be mean , I really wanted to like this film , it seemed to have everything going for it , had the so called " buzz ", and was a hassle to track down besides . Had a little more effort gone into it on the story side , I believe this would ' ve been amazing . And I expect the team behind it will produce wonderful work in the future , they clearly have the ability . But I recommend waiting for their future efforts , let this one go . | Long , whiny and pointless . |
| **9** | Excellent cast , story line , performances . Totally believable . I realize the close knit group that exemplifies the Marine Corps . But this movie brought fear to my heart . The marines let principles be damned . It seems that this film was based on real life incidents . It shows how difficult it is to go up against the establishment . Anne Heche was utterly convincing . Sam Shepard ' s portrayal of a gung ho Marine was sobering . And Eric Stoltz as her attorney was so deft balancing his loyalty to the Corp but also his loyalty to his client , while high above on his tightrope . He knew what his true course of action had to be . But he was pulled apart by his immersion in the Marine tradition , loyalty to the Corps above all else . I sat riveted to the TV screen . All in all I give this one a resounding 9 out of 10 . | Excellent cast , story line , performances . And Eric Stoltz as her attorney was so deft balancing his loyalty to the Corp but also his loyalty to his client , while high above on his tightrope . He knew what his true course of action had to be . |
| **10** | I ` ve seen this movie twice , both times on Cinemax . The first time in it ` s unrated version which is soft - core porn at it ` s best and the second time in a trimmed down ( cut all the sex and most of the nudity out ) version which was entertaining in a typical beach movie sort of way . The unrated version has a tremendous sex scene with Nikki Fritz , a dude and a bottle of oil which is out of this world ( no pun intended ). Unfortunately , in the trimmed version that scene is almost completely chopped out , as are all the other sex scenes . Rated or unrated it is still fun to watch all the siblings of bigger stars ( Stallone , Sheen , Travolta , etc ;) trying to act . We also get appearances by B - queen Linnea Quigley and Burt Ward ( Robin from the old Batman series ). | Rated or unrated it is still fun to watch all the siblings of bigger stars ( Stallone , Sheen , Travolta , etc ;) trying to act . |
| **11** | I do think Tom Hanks is a good actor . I enjoyed reading this book to my children when they were little . I was very disappointed in the movie . One character is totally annoying with a voice that gives me the feeling of fingernails on a chalkboard . There is a totally unnecessary train / roller coaster scene . There are some characters and scenes that seem scary for little children for | There is a totally unnecessary train / roller coaster scene . The North Pole scenes with Santa and the elves could have been cute and charming . |

| | | |
|---|---|---|
| | whom this movie was made . The North Pole scenes with Santa and the elves could have been cute and charming . There was absolutely no warmth or charm to these scenes or characters . It usually doesn ' t work to make a short children ' s book into a feature film . This movie totally grates on my nerves . | |
| **12** | Improvisation was used to a groundbreaking degree in this film , but it only functions as a novelty . No greater truth about the situation is got by asking the actors to improvise . The performances are not improved by improvisation , because the actors now have twice as much to worry about : not only whether they ' re delivering the line well , but whether the line itself is any good . So that ' s why the performances in many Robert Altman films are often really hestitant - because the actors aren ' t really confident saying lines which they ' ve made up , and therefore aren ' t sure are any good . And , quite honestly , often its not very good . Often the dialogue doesn ' t really follow from one line to another , or fit the surroundings . It crackles with an unpredictable , youthful energy - but honestly , i found it hard to follow and concentrate on it meanders so badly . Nevertheless , a fascinating raw piece of film , and commendable 100 % for taking the power over the green light into the street . There are some generally great things in it . This joke , for example : I ' m a dancer . What sort of a dancer , like a ballet dancer ? Oh no ... exotic . And the whole party scene its in , the following trip to the park , and the scene where the boys go looking at statues . 2 / 5 . I wouldn ' t say they ' re worth 2 hours of your time , though . | It crackles with an unpredictable , youthful energy - but honestly , i found it hard to follow and concentrate on it meanders so badly . Nevertheless , a fascinating raw piece of film , and commendable 100 % for taking the power over the green light into the street . Oh no ... exotic . |
| **13** | This little flick is reminiscent of several other movies , but manages to keep its own style & mood . " Troll " & " Don ' t Be Afraid of the Dark " come to mind . The suspense builders were good , & just cross the line from G to PG . I especially liked the non - cliche choices with the parents ; in other movies , I could predict the dialog verbatim , but the writing in this movie made better selections . If you want a movie that ' s not gross but gives you some chills , this is a great choice . | If you want a movie that ' s not gross but gives you some chills , this is a great choice . |
| **14** | The fun that was present in the other ' movies ' has all but disappeared with this third effort , which means the rubbishy production values show through more than ever . Another 2 - parter (' The Chinese Web ') cobbled together , this one suffers from too much padding , not to mention weak Spidey action taking place in such uninspired locations as a car park , apartment and printing press . Nicholas Hammond is as endearing as ever , struggling valiantly against the drab production and lame performances of the rest of the cast . The plot , in which Peter and Spidey help a Chinese official defeat charge of corruption during World War II by locating three marines who could testify as to his innocence , doesn ' t exactly scream ' comic book sprung to life ', does it ? | Nicholas Hammond is as endearing as ever , struggling valiantly against the drab production and lame performances of the rest of the cast . |
| **15** | When I first saw this film around 6 months ago , I considered it interesting , but little more . But it stuck with me . That interest grew and grew , and I wondered whether my initial boredom and response had more to do with the actual VHS quality rather than the film itself . I purchased the Criterion DVD box set , and it turns out that I was right the second time . Alexander Nevsky is a great film . It is rousing , and I ' m sure it succeeded in its main aim : propaganda against the Germans . That is the most common criticism against this film , and against Eisenstein , that it is merely propagandist and nothing else . It ' s untrue . He is an amazing film artist , one of the most important whoever lived . By now , the world is far enough beyond Joseph Stalin to be able to watch Eisenstein ' s films as art . | Alexander Nevsky is a great film . He is an amazing film artist , one of the most important whoever lived . |
| **16** | This is the one major problem with this film , along with a good deal of qubecois ' biggest movies : Done in a pretentious way by pretentious people . It ' s really sad , but " big shots " movie makers ( driving Dodge Stratus ...) from this province believes They Got the Thruth , They Know What the Little People Like . We ' re not a rich province , every time a big movie like this ( 30 millions ?!!? ) is made , it ' s cutting off a lot of others who won ' t see their movie made because of lack of governmental help . So it generates mediocrity ; only movies from " friends of the family " are going to be made . I sound angry and I am . I went see Nouvelle - France expecting a journey in the lives of my ancestors , but i found myself stuck in a pool of inconsistencies : french accent ( we gotta please our cousins , so f *** our qubecois ' language ) and lack of historical research is only a few . Add a campy love story and the same music score | I ' m glad this pretentious piece of s *** didn ' t do as planned by the Dodge stratus Big Shots ... It ' s gonna help movie makers who aren ' t in the very restrained " movie business " of Qubec . Rent Cruising Bar instead and have a real good time . |

| | | |
|---|---|---|
| | playing again and again and dumb qubecois ' viewer is gonna open up and ask for more . I ' m glad this pretentious piece of s *** didn ' t do as planned by the Dodge stratus Big Shots ... It ' s gonna help movie makers who aren ' t in the very restrained " movie business " of Qubec . Rent Cruising Bar instead and have a real good time . PS : I ' ll never forgive them for ruining such an awesome title . | |
| **17** | This if the first movie I ' ve given a 10 to in years . If there was ever a movie that needed word - of - mouth to promote , this is it . A $ 4 Mil box is a disgrace . People don ' t know what it ' s about . If you have any appreciation for the Blues , or just a good use of excellent music , that alone is reason to go see it . How many people knew Jackson could sing , and damn fine too . You hear books and movies taunting that they ' re about salvation . After seeing this , you ' ll never be able to forgive such trivial use of the word . Yes , it ' s gritty , sexy , down home truth , bizarre and in - your - face real . Isn ' t that the best reason to see a movie ? Those that get my meaning won ' t stay away from seeing this another week . | If you have any appreciation for the Blues , or just a good use of excellent music , that alone is reason to go see it . How many people knew Jackson could sing , and damn fine too . |
| **18** | This is one of Barbara Stanwyck ' s earlier films and it sure does have an unconventional theme . She ' s making money by dancing with men at a dance hall . She really doesn ' t like the work , but it ' s a living . Her boyfriend seems like a pretty nice guy , but she ' s also pursued by rich guy Ricardo Cortez . Well , after marrying , it turns out her " nice guy " is a thieving , womanizing weasel and rich Cortez turns out to be a heck of a guy . By the end of the film , Barbara simply has had enough , as any SANE woman would walk from this horrid marriage . In the 1920s and early 30s , Hollywood did pretty much anything it wanted and some of their films had themes or scenes that would surprise many today -- such as nudity , adultery and bad language . While TEN CENTS A DANCE isn ' t a blatant example of this morality , it does have a theme that never would have been allowed after the toughened Production Code was created and enforced starting in 1934 . In some ways the Code was great -- after all , parents didn ' t need to worry about what their kids saw in films ( such as nudity in BEN HUR , 1925 ). However , it also tended to sanitize some of the movies far too much -- and there is no way this particular film could have been made and approved because it tends to glorify divorce -- a serious no - no 1934 and thereafter . This is really a shame , as I don ' t think TEN CENTS A DANCE was bad at all to discuss this -- especially since the star ( Barbara Stanwyck ) was married to a philandering thief . Even so , allowing the film to end with her divorcing him and marrying a man who himself was twice divorced just couldn ' t have been . Overall , the film is interesting and thought - provoking . Plus , it was well - paced and suited its relatively short run time . Give this one a look . FYI -- Sadly , Ricardo Cortez was actually NOT Hispanic but changed his name because of possible prejudice because he was Jewish . He was an excellent leading man of his time , but today is all but forgotten . | This is really a shame , as I don ' t think TEN CENTS A DANCE was bad at all to discuss this -- especially since the star ( Barbara Stanwyck ) was married to a philandering thief . Even so , allowing the film to end with her divorcing him and marrying a man who himself was twice divorced just couldn ' t have been . Plus , it was well - paced and suited its relatively short run time . He was an excellent leading man of his time , but today is all but forgotten . |
| **19** | The premise of the story is common enough ; average family wants out of the rat race ; wants to find the simple life .... so they move from Sherman Oaks , California to Lake Tomahawk ; kids in tow . The lake is beautiful , they have leased an old house .... but wait ; there may be something in the lake ; people are being murdered , and no one knows how ( never mind why ). Gerald McRaney is excellent , a familiar face for Lifetime viewers ; Valerie Harper is also good ; since this film was made in ' 88 maybe the writer should produce a sequel !. You will also enjoy Barry Corbin as the town eccentric , and Darryl Anderson as a Bruce Dern - lookalike / crazed military man . While the story plot is a bit over the top ; if you are a movie buff you will be reminded of similar scenarios from "" Psycho ""; "" Deliverance ""; as well as other horror stories of that genre . Several camera shots and sequences will give you a sense of deja vu . Sit back and enjoy ; if you don ' t take it too seriously it is very entertaining ; and better than , for example the more recent movie : "" I Know What You Did Last Summer ""; it seems they made better movies in the good old 80 ' s !. | While the story plot is a bit over the top ; if you are a movie buff you will be reminded of similar scenarios from "" Psycho ""; "" Deliverance ""; as well as other horror stories of that genre . |
| **20** | This is actually the first movie I ever saw in a theatre , where the people didn ' t leave immediately when the end credits started . In stead they remained seated for a few minutes , gaping with their mouths open staring in the infinite , trying | This is actually the first movie I ever saw in a theatre , where the people didn ' t leave |

| | | |
|---|---|---|
| | to understand what they ' ve just seen . The only thing I can say : Try to go watch this movie with as little knowledge about it as possible ( so did I )!. I gave it a 10 | immediately when the end credits started . |
| **21** | The Wind and the Lion is well written and superbly acted . It is a tale that exemplifies the American spirit and the American character . This movie is a story from the early 20th century that is strangely relevant to the political landscape of the world in the beginning of the 21st century . It is a true classic . | It is a true classic . |
| **22** | No pun intended . I ' m not going to spoil anything about the story , but it ' s safe to assume that you already know , what kind of character the main actor portrays . And of course being a priest while being " naughty " exaggerates all that . Plus this is the most erotic movie from Park Chan Wook yet . If you have seen Wook ' s previous works / movies you know he is very visual ( in a good way ) and it shows again here . While it strays away from the vengeance theme of his prior movies on the surface , it still has quite some heat hidden underneath . And when that boils , quite a few bad things start to happen . But through all that dark , there also moments of light ( fun ) to be had too . A very stylistic and though provoking movie , that lives outside the mainstream and does a very good job ... | Plus this is the most erotic movie from Park Chan Wook yet . A very stylistic and though provoking movie , that lives outside the mainstream and does a very good job ... |
| **23** | I caught this movie on the Sci - Fi channel recently . It actually turned out to be pretty decent as far as B - list horror / suspense films go . Two guys ( one naive and one loud mouthed a **) take a road trip to stop a wedding but have the worst possible luck when a maniac in a freaky , make - shift tank / truck hybrid decides to play cat - and - mouse with them . Things are further complicated when they pick up a ridiculously whorish hitchhiker . What makes this film unique is that the combination of comedy and terror actually work in this movie , unlike so many others . The two guys are likable enough and there are some good chase / suspense scenes . Nice pacing and comic timing make this movie more than passable for the horror / slasher buff . Definitely worth checking out . | Two guys ( one naive and one loud mouthed a **) take a road trip to stop a wedding but have the worst possible luck when a maniac in a freaky , make - shift tank / truck hybrid decides to play cat - and - mouse with them . Definitely worth checking out . |
| **24** | About the spoiler warning ? It ' s not " may contain ", it - does - contain spoilers . Readers beware . Okay , first I need it to be known that I ' m not bashing the actors . They ' re just working with what they ' re given . The problem was the script . It was horrendous . There was NOTHING believable about it at all . Sure , when you have a movie based on a murderous hitchhiker , there ' s going to be the bad mistake here and there that puts you in the terribly horrific , movie - worthy situation . But these girls just made stupid decision after stupid decision . The only girl smart enough to ever try and call the police was the girl added towards the end because he ' d already killed one and hit another with a car . Speaking of hitting her with a car ... why the hell did she try and outrun a truck rather than run to the side like a normal person ? Also , does the one who wrote the script honest to god believe cops are not going to investigate a door covered in blood ? Frankly , it wasn ' t suspenseful either . The only suspense I was feeling was the frustration at just how retarded the girls were . Well , this rant has gone on way longer than I meant to for such a bad movie , so I won ' t bother to touch on the end besides the fact it ' s unrealistic and lame . | But these girls just made stupid decision after stupid decision . Speaking of hitting her with a car ... why the hell did she try and outrun a truck rather than run to the side like a normal person ? Frankly , it wasn ' t suspenseful either . Well , this rant has gone on way longer than I meant to for such a bad movie , so I won ' t bother to touch on the end besides the fact it ' s unrealistic and lame . |
| **25** | MANNA FROM HEAVEN is a terrific film that is both predictable and unpredictable at the same time . You know that the characters after finding out that the so - called " Gift From God " was actually a loan , will pay back the money and that everyone will be happy at the end , but how they get there is not as obvious . The scenes are often funny and occasionally touching as the characters evaluate their lives and where they are going . The cast of veteran actors are more than just a nostalgia trip . Frank Gorshin , Shirley Jones , and Cloris Leachman prove that they are capable of more than playing the Riddler , Mother Partridge , or Mary ' s friend Phyllis while Jill Eikenberry and Wendie Malick play characters different than we have seen on their TV series . Ursula Burton ' s portrayal of the nun is both touching and funny at the same time with out making fun of nuns or the church . If you are looking for a movie with a terrific cast , some good music ( including a Shirley Jones rendition of " The Way You Look Tonight "), and an uplifting ending , give this one a try . I don ' t think you will be disappointed . | The scenes are often funny and occasionally touching as the characters evaluate their lives and where they are going . Ursula Burton ' s portrayal of the nun is both touching and funny at the same time with out making fun of nuns or the church . |
| **26** | Frankly , after Cotton club and Unfaithful , it was kind of embarrassing to watch Lane and Gere in this film , because it is BAD . The acting was bad , the dialogs | Frankly , after Cotton club and Unfaithful , it was kind of |

| | | |
|---|---|---|
| | were extremely shallow and insincere . It was well shot , but , then again , it is a big budget movie . It was too predictable , even for a chick flick . I even knew from the beginning that he was going to die in the end , the only thing I didn ' t know was how . Too politically correct . Very disappointing . The only thing really worth watching was the scenery and the house , because it is beautiful . But , if you want that , watch National geographic . I love Lane , but I ' ve never seen her in a movie this lousy . As far as Gere goes , he ' s a good actor , but he had movies like this , so I ' m not surprised . An hour and a half I wish I could bring back . | embarrassing to watch Lane and Gere in this film , because it is BAD . It was too predictable , even for a chick flick . Very disappointing . |
| **27** | This movie is apparently intended for a young , evangelical Christian audience as a teaching tool . For that I give it a 7 out of 10 point vote . It ' s a decent movie to show a youth group , but I don ' t think it will be very well received beyond that . For any other audience , I ' d rate it lots lower . The reviewers that saw " It ' s a Wonderful Life " in this were right on , though I didn ' t think of that until they mentioned it . I was more reminded of a " Chick Tract ", those little 3 " by 5 " gospel comic books . If Jack Chick ever made a movie out of one of his tracts , it would probably look a lot like " Second Glance . " It has a strong Christian message about the power of prayer and the influence each of us has on earth , but it is somewhat hampered by Christian stereotypes . The Christians are all very nice , somewhat passive , and squeaky clean , while all of the non - Christians seem to be bad people . Muriel the angel plays a major part , and he is the corniest , cheesiest character in the film . He is the most unlikeable angel I have ever seen in any movie , and the biggest negative . I don ' t know if the directors intended for his personality to come off so badly , or if he just struck me that way . ( I admit that he reminded me of someone I know . ) Dan ' s love for a very worldly girl who is not at all his type drives the plot in this movie . Why he ever fell for her in the first place is the one question that I wish had been answered . But the movie does display positive Christian values , and your youth group will be entertained as they view something wholesome with a good lesson . | This movie is apparently intended for a young , evangelical Christian audience as a teaching tool . The reviewers that saw " It ' s a Wonderful Life " in this were right on , though I didn ' t think of that until they mentioned it . I was more reminded of a " Chick Tract ", those little 3 " by 5 " gospel comic books . " It has a strong Christian message about the power of prayer and the influence each of us has on earth , but it is somewhat hampered by Christian stereotypes . |
| **28** | We have a character named Evie . Evie just wants to be a good person . She ' s nice , friendly , smiles often , but is strangely brutally honest . Evie also has a secret . Her idiot - savant sister has been reciting original poetry , which is getting the community excited about the sister writing . Unfortunately , it ' s Evie ' s poetry . While their mother starts being happy again and the boy next door shows his interest in Evie , Evie just tries to figure out what she really wants to do . What to keep in mind while watching this movie is who Evie really is . For such a brutally honest person who doesn ' t mind telling Ivy - league types that she doesn ' t respect them , it would seem odd that she would be able to pull off a lie . For someone so happy and cheerful , she ' s quite emotionless when it comes to certain issues . Those aren ' t character flaws , they ' re plot development , and they mean a lot more than they at first seem . Mostly this is something of a melodrama : a character lies , the other characters ' personalities propel them through drama as relationships are held at risk . But in terms of the writing it ' s very fresh and bold . The acting helps the writing along very well ( maybe the idiot - savant sister could have been played better ), and it is a real joy to watch . The directing and the cinematography aren ' t quite as good . They ' re acceptable , and Evie ' s world is wreathed in color and light , which makes for some very beautiful images , but it ' s not very consistent . It ' s not really so much of a flaw as a result of a low production value , but within that same value is some genuine storytelling and a real care for the characters . So while it isn ' t a perfect movie , it ' s certainly an enjoyable one . -- PolarisDiB | Unfortunately , it ' s Evie ' s poetry . While their mother starts being happy again and the boy next door shows his interest in Evie , Evie just tries to figure out what she really wants to do . But in terms of the writing it ' s very fresh and bold . The acting helps the writing along very well ( maybe the idiot - savant sister could have been played better ), and it is a real joy to watch . |
| **29** | After starting watching the re - runs of old Columbo movies , I thought they would all get about the same vote from me ( 6 ). But apparently I ' m now starting to see differences in the movies . It happened in some of just previous episodes , that showed some pretty genius directing , and it shows in this one , but in the negative way . The movie was so boring , that I sometimes found myself occupied peaking in the paper instead of watching ( never happened during a Columbo movie before ! ), and sometimes it was so embarrassing that I had to look away . The directing seems too pretentious . The scenes with the " | The movie was so boring , that I sometimes found myself occupied peaking in the paper instead of watching ( never happened during a Columbo movie before ! The directing seems too pretentious . I really liked that . |

| | | |
|---|---|---|
| | oh - so - mature " neighbour - girl are a misplace . And generally the lines and plot is weaker than the average episode . Then scene where they debated whether or not to sack the trumpeter ( who falsely was accused for the murder ) is pure horror , really stupid . Some applause should be given to the " prelude " however . In this episode , a lot of focus is given on how the murderer tries to secure his alibi and hide the evidence etc . I really liked that . But alas , no focus on how Columbo reveals all this . And the " proof " that in the end leaves Columbo victorious is the silliest ever . Rating : lies between 4 and 5 | |
| **30** | I heard they were going to remake this French classic in 2007 , and I see it is in development for 2011 . This will be a shame , as Hollywood kicked writer / director Jules Dassin out because of the infamous blacklist . They should not have the right to remake any of his films . I love " caper " films and " film noir ," and this combines the best of both . Tony ( Jean Servais ) gets out after doing a nickle , and after he beats up his old girlfriend ( Marie Sabouret ), he plans a big score with his friends Mario ( Robert Manuel ) and Jo ( Carl Mhner ), What makes this a great caper flick is the attention to detail in planning the robbery . You see that reflected in the George Clooney Vegas capers . Nothing is left to chance . The caper goes off great but Grutter ( Marcel Lupovici ) sends his sons , Robert Hossein and Pierre Grasset after Tony and the gang . After blowing it with Mario , they kidnap Jo ' s son . Lots of bullets fly before it is over . A great film by a great director . The standard by which other caper films are measured . | I love " caper " films and " film noir ," and this combines the best of both . Tony ( Jean Servais ) gets out after doing a nickle , and after he beats up his old girlfriend ( Marie Sabouret ), he plans a big score with his friends Mario ( Robert Manuel ) and Jo ( Carl Mhner ), What makes this a great caper flick is the attention to detail in planning the robbery . Nothing is left to chance . |
| **31** | Captain Corelli ' s Mandolin is a beautiful film with a lovely cast including the wonderful Nicolas Cage , who as always is brilliant in the movie . The music in the film is really nice too . I ' d advise anyone to go and see it . Brilliant ! 10 / 10 | Captain Corelli ' s Mandolin is a beautiful film with a lovely cast including the wonderful Nicolas Cage , who as always is brilliant in the movie . |
| **32** | I liked this movie way too much . My only problem is I thought the actor playing the villain was a low rent Michael Ironside . Of corse Ironside is just a low rent Jack Nicholson . I guess Mike was busy that year with " Highlander 2 : The Quickening ". Sadly " Beastmaster 2 " would have been a much better career move . It is certainly the best of the Beastmaster series and in many ways reminiscent of that great big screen classic " Masters of the Universe ". Not only does it star the incomparable Mark Singer it also features an amazing supporting cast , specifically the second girl from " Sliders ", Uncle Phil from " Fresh Prince of Belair " and evil chick from " Superman 2 ". It rocked my world and is certainly a must see for anyone with no social or physical outlets . BEASTMASTER FOREVER !!! ROCK ' N ROLL !! ! | My only problem is I thought the actor playing the villain was a low rent Michael Ironside . Of corse Ironside is just a low rent Jack Nicholson . |
| **33** | There were some decent moments in this film , and a couple of times where it was pretty funny . However , this didn ' t make up for the fact that overall , this was a tremendously boring movie . There was NO chemistry between Ben Affleck and Sandra Bullock in this film , and I couldn ' t understand why he would consider even leaving his wife - to - be for this chick that he supposedly was knocked out by . There was better chemistry between him and Liv Tyler in Armageddon . Hell , there was better chemistry between Sly and Sandra in Demolition Man . There were several moments in the movie that just didn ' t need to be there and were excruciatingly slow moving . This was a poor remake of " My Best Friends Wedding ". Wait until it ' s been out for a year and a half on video and rent it in the . 49 cent bin if you ' ve got nothing else to do on a rainy Sunday afternoon , and you can ' t think of any better movies to rent . | However , this didn ' t make up for the fact that overall , this was a tremendously boring movie . There was NO chemistry between Ben Affleck and Sandra Bullock in this film , and I couldn ' t understand why he would consider even leaving his wife - to - be for this chick that he supposedly was knocked out by . |
| **34** | I picked up this DVD for $ 4 . 99 . They had put spiffy cover art on the package , along with a plot summary that had nothing to do with the movie . The acting is terrible , and the writing is worse . The only possible way this movie could be redeemed would be as MST3K fodder . I paid too much . | The acting is terrible , and the writing is worse . |
| **35** | They do ... Each sequel is worst . You , who think that Ghoulies 2 or 3 need a 1 , please , watch this sequel ... You ' ll be wondering with the first three parts . Then you ' ll give a 10 to the first , 8 to the second and 5 or 6 to the other . That ' s because Ghoulies 4 really gets the big 1 ( from me it does ). | They do ... Each sequel is worst . |
| **36** | The movie I received was a great quality film for it ' s age . John Wayne did an incredible job for being so young in the movie industry . His on screen presence shined thought even though there were other senior actors on the screen with him . I think that it is a must see older John Wayne film . | The movie I received was a great quality film for it ' s age . |

# References


D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines*. *Cognitive science*, 9(1):147–169, 1985.

J. W. Backus. The syntax and semantics of the proposed international algebraic language of the zurich acm-gamm conference. *Proceedings of the International Comference on Information Processing, 1959*, 1959.

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19 (2):263–311, 1993.

A. E. Bryson, Y.-C. Ho, and G. M. Siouris. Applied optimal control: Optimization, estimation, and control. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(6): 366–367, 1979.

E. Charniak. Statistical parsing with a context-free grammar and word statistics. *AAAI/IAAI*, 2005:598–603, 1997.

N. Chomsky. Three models for the description of language. *Information Theory, IRE Transactions on*, 2(3):113–124, 1956.

M. Collins. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23. Association for Computational Linguistics, 1997.

R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

R. Courant et al. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc*, 49(1):1–23, 1943.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

G. E. Dahl, R. P. Adams, and H. Larochelle. Training restricted boltzmann machines on word observations. *arXiv preprint arXiv:1202.5695*, 2012.

M. Denil, B. Shakibi, L. Dinh, N. de Freitas, et al. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pages 2148–2156, 2013.

R. Descartes. Meditations on first philosophy, 1641. *The philosophical writings of Descartes*, 2, 1967.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4): 193–202, 1980.

A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

Ç. Gülçehre and Y. Bengio. Knowledge matters: Importance of prior information for optimization. *arXiv preprint arXiv:1301.4083*, 2013.

K. M. Hermann and P. Blunsom. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*, 2014.

G. E. Hinton. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA, 1986.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 44–51. Springer, 2011.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

D. Jurafsky and J. H. Martin. Speech and language processing: An introduction to speech recognition. *Computational Linguistics and Natural Language Processing. 2nd Edn., Prentice Hall, ISBN*, 10(0131873210):794–800, 2008.

N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014.

Q. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning*, volume 32, 2014.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 2006.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

C. Métin and D. O. Frost. Visual responses of neurons in somatosensory cortex of hamsters with experimentally induced retinal projections to somatosensory thalamus. *Proceedings of the National Academy of Sciences*, 86(1):357–361, 1989.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

M. Minsky and S. Papert. Perceptron: an introduction to computational geometry. *The MIT Press, Cambridge, expanded edition*, 19:88, 1969.

A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, pages 2265–2273, 2013.

A. W. Roe, S. L. Pallas, Y. H. Kwon, and M. Sur. Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex. *The Journal of neuroscience*, 12(9):3651–3664, 1992.

R. Rojas. *Neural networks: a systematic introduction*. Springer, 1996.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

S. Russell and P. Norvig. *Artificial Intelligence: A modern approach*, volume 25. Citeseer, 1995.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps. Technical report, 2013.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986.

A. M. Turing. Computing machinery and intelligence. *Mind*, pages 433–460, 1950.

S. Wang and C. D. Manning. Baselines and Bigrams: Simple , Good Sentiment and Topic Classification. In *Association for Computational Linguistics*, volume 94305, 2012.

M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. Technical report, 2012.

M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018–2025. IEEE, 2011.