

Hierarchical Temporal Memory for Anomaly Detection in Videos

*An alternative and biologically plausible
approach to a complex problem*

Vladimir Monakhov



Thesis submitted for the degree of
Master in Informatics: Robotics and Intelligent
Systems
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

Spring 2022

Hierarchical Temporal Memory for Anomaly Detection in Videos

*An alternative and biologically plausible
approach to a complex problem*

Vladimir Monakhov

© 2022 Vladimir Monakhov

Hierarchical Temporal Memory for Anomaly Detection in Videos

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

Printed: Reprocentralen, University of Oslo

Abstract

The use of video anomaly detection systems has gained traction for the past few years. The current approaches use deep learning for performing anomaly detection in videos, but this has multiple problems. For starters, deep learning in general has issues with noise, concept drift, explainability, and training data volume. Additionally, anomaly detection in itself is a complex task and faces challenges such as unknownness, heterogeneity, and class imbalance. Anomaly detection in deep learning is therefore mainly constrained to generative models such as GANs and autoencoders due to their unsupervised nature, but even they suffer from general deep learning issues and are hard to train properly. This thesis instead looks to HTM to perform anomaly detection in videos, as it has favorable properties such as noise tolerance and online learning which combats concept drift. This thesis does so by introducing Grid HTM, which is an HTM based architecture specifically for anomaly detection in complex videos such as surveillance footage.

Acknowledgments

I would like to thank my supervisors for their invaluable assistance and guidance, as well as for answering emails quickly.

The research presented in this thesis has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

Contents

Abstract	1
Acknowledgments	3
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	3
1.3 Limitations	3
1.4 Contributions	4
1.5 Research Methods	5
1.6 Thesis Outline	6
2 Deep Learning, Anomaly Detection, and Hierarchical Temporal Memory	9
2.1 Deep Learning	9
2.1.1 Perceptron	9
2.1.2 Backpropagation	11
2.1.3 Neural Networks	12
2.1.3.1 Learning	12
2.1.4 Convolutional Neural Networks	13
2.1.5 Generative Models	14
2.1.6 Disadvantages of Deep Learning	16
2.1.6.1 Explainability	16
2.2 Anomaly detection	17
2.2.1 Deep Learning and Anomaly Detection	17
2.2.2 Smart Surveillance	18
2.3 Hierarchical Temporal Memory	19
2.3.1 Structure	19
2.3.2 Common Algorithm	22
2.3.3 Sparse Distributed Representation	22
2.3.4 Encoders	22
2.3.4.1 Encoding Visual Data	24
2.3.5 Learning	25
2.3.5.1 Spatial Pooler	27
2.3.5.2 Temporal Memory	29
2.3.6 Use Cases	33
2.3.7 The Thousand Brains Theory	33

2.3.8	HTM Performance in Anomaly Detection	35
2.4	Ethical Considerations	35
2.5	Summary	36
3	Grid HTM	39
3.1	Introduction	39
3.2	Improvements	40
3.2.1	Invariance	40
3.2.1.1	Aggregation Function	41
3.2.2	Explainability	43
3.2.3	Flexibility and Performance	43
3.2.4	Reviewing Encoder Rules	43
3.2.5	Stabilizing Anomaly Output	45
3.2.6	Multistep Temporal Patterns	46
3.3	Implementation	48
3.4	Biological Plausibility	48
3.5	Use Cases	49
3.6	Summary	49
4	Experiments and Results	51
4.1	Bouncing Ball Experiment	51
4.1.1	Data	51
4.1.2	HTM	52
4.1.2.1	Boosting	52
4.1.2.2	Zero Permanence Decrement	53
4.1.2.3	Boosting and Zero Permanence Decrement .	54
4.1.2.4	Parameters	55
4.1.3	Grid HTM	56
4.1.3.1	Results	56
4.1.3.2	Parameters	56
4.1.4	Experiment Summary	58
4.2	Surveillance Experiment	59
4.2.1	Results	60
4.2.1.1	Road	61
4.2.1.2	Frame Repeat	62
4.2.1.3	Points of Interest	65
4.2.2	Parameters	67
4.2.3	Experiment Summary	69
4.3	Sperm Experiment	70
4.3.1	Data	70
4.3.2	Benchmark	70
4.3.3	Results	70
4.3.4	Use Cases	72
4.3.5	Parameters	72
4.3.6	Experiment Summary	74
4.4	Summary	75
5	Conclusion & Future Work	77

5.1	Summary	77
5.2	Contributions	78
5.3	Future Work	79
A	Paper - GRID HTM: Hierarchical Temporal Memory for Anomaly Detection in Videos	89

List of Figures

1.1	Publications Increase Comparison	2
2.1	The Original Perceptron	10
2.2	MLP Example	11
2.3	Neural Network Example	12
2.4	Gradient Descent Comparison	13
2.5	Typical CNN Architecture	14
2.6	Gan Example	15
2.7	Dataset Volume Accuracy	16
2.8	Grad-Cam Visualization	17
2.9	GANomaly Visualization	18
2.10	HTM Structure	20
2.11	Comparison of Neurons	21
2.12	SDR Semantic Comparison	22
2.13	Example SDR	22
2.14	Cyclical Date Encoding	23
2.15	Thresholding Example	24
2.16	Feature Maps	25
2.17	The HTM Pipeline	26
2.18	SP and TM Responsibilities	27
2.19	Spatial Pooler Visualization	28
2.20	Spatial Pooler Workings	29
2.21	Temporal Memory Visualization	30
2.22	Temporal Memory Workings	31
2.23	Visualization of Cell States	32
2.24	Thousand Brains Visualization	34
2.25	Filter Concatenation	34
2.26	Example Motion Frame	35
3.1	Segmentation Result of Cars	39
3.2	SDR and SP Representation	40
3.3	Encoder Output Grid	41
3.4	Aggregation Functions on Noise Data	42
3.5	Aggregation Functions on Clean Data	42
3.6	Distribution of Active Pixels	44
3.7	Example Grid HTM Output	45
3.8	Stabilizing Anomaly Output Visualization 1	45
3.9	Stabilizing Anomaly Output Visualization 2	46

3.10	Comparison of a Moving Object in a High Framerate Video	46
3.11	Contextual Loop Example	47
3.12	Multistep Temporal Pattern Example	47
3.13	Grid HTM Environments	49
4.1	Bouncing Ball Experiment	51
4.2	Bouncing Ball Experiment Anomaly Score	52
4.3	Bouncing Ball Experiment Anomaly Score No Boosting	53
4.4	Bouncing Ball Experiment Anomaly Score Zero Decrement	54
4.5	Bouncing Ball Experiment Anomaly Score No Boosting Zero Decrement	54
4.6	Bouncing Ball Experiment Anomaly Score Grid HTM	56
4.7	Example Frames	59
4.8	Example Segmentation of Cars and Persons	60
4.9	Grid HTM Anomaly Score Output	60
4.10	Segment Anomaly	61
4.11	Car Driving Along Main Road	62
4.12	Frame Repeat Anomaly	63
4.13	No Frame Repeat Anomaly	64
4.14	Frame Repeat No Multistep Temporal Pattern Anomaly	65
4.15	Points of Interest	66
4.16	First POI	66
4.17	Second POI	67
4.18	Sperm Example Frame	70
4.19	Stationary Video Results	71
4.20	Drifting Video Results	72

List of Tables

4.1	SP Parameters	55
4.2	TM Parameters	55
4.3	Grid HTM specific parameters	57
4.4	SP Parameters	57
4.5	TM Parameters	58
4.6	Grid HTM specific parameters	68
4.7	SP Parameters	68
4.8	TM Parameters	69
4.9	Grid HTM specific parameters	73
4.10	SP Parameters	73
4.11	TM Parameters	74

Chapter 1

Introduction

1.1 Background and Motivation

As the global demand for security and automation increases, many seek to use video anomaly detection systems. In the US alone, the surveillance market is expected to reach \$23.60 Billion by 2027 [1]. Leveraging modern computer vision, modern anomaly detection systems play an important role in increasing monitoring efficiency and reducing the need for expensive live monitoring. Their use cases can vary from detecting faulty products on an assembly line to detecting car accidents on a highway, and everything in between.

The most important component in video anomaly detection systems is the intelligence behind it. The intelligence ranges from simple on-board algorithms to advanced deep learning models, where the latter has experienced increased popularity in the past few years, as can be seen in Figure 1.1.

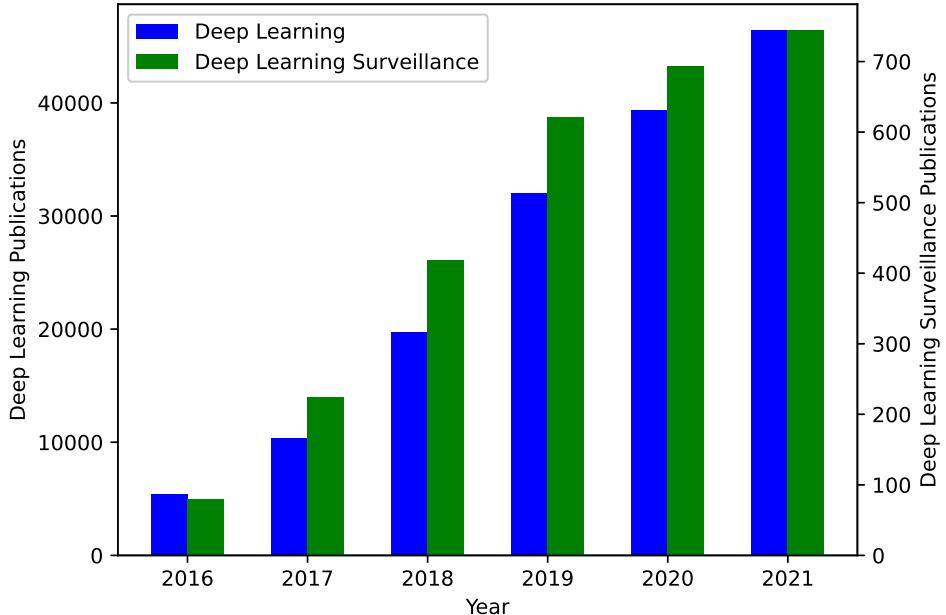


Figure 1.1: The increase in publications mentioning the terms "deep learning" and "deep learning surveillance" [2].

Yet despite the major progress within the field of deep learning, there are still many tasks where humans outperform models, especially in anomaly detection where the anomalies are often undefined. Deep learning approaches also perform poorly when dealing with noise and concept drift.

The cause for the discrepancy lies in the difference between how humans and machine learning algorithms represent data and learn. Most machine learning algorithms use a dense representation of the data and apply back-propagation in order to learn. Human learning happens in the neocortex, where evidence points to that the neocortex uses a sparse representation and performs Hebbian-style learning. For the latter, there is a growing field of machine learning dedicated to replicating the inner mechanics of the neocortex, namely Hierarchical Temporal Memory (HTM) theory. This theory outlines its advantages over standard machine learning, such as noise-tolerance and the ability to adapt to changing data.

With the advantages of HTM and the rise of video anomaly detection in mind, a natural question one could pose is whether it is possible to apply HTM for anomaly detection in videos. Combined with a lack of related works, it is this very question that is the motivation behind this thesis.

1.2 Problem Statement

Based on the background and motivation, the problem statement can be boiled down to a simple question: **Is HTM viable for use in video anomaly detection?**

This thesis will introduce three different objectives that will help answer the question and also showcase the performance of HTM. It will also cover all required knowledge. The objectives are as follows:

1. Introduce HTM and give a deep understanding of the inner workings, the strengths, and the weaknesses. While also being easy to grasp for readers with a machine learning background.
2. Develop and outline a theoretically sound HTM architecture that can be applied for anomaly detection in complex videos.
3. Perform experiments, discuss the results, and lay out potential future work for Grid HTM. The experiments will vary in difficulty, complexity, and will focus on different use cases.

1.3 Limitations

HTM is a complex topic not part of the curriculum in most educations, if any at all. It is also based on neurological research, lending terms and concepts from the biological field, which significantly raises the level of entry for people with a machine learning background. The field also has a low level of accessibility due to a lack of proper documentation and a high reliance on the tiny HTM community. This makes learning and understanding HTM a process which takes up a sizable chunk out of the total time spent on this thesis. This thesis will therefore be relatively limited in scope.

Additionally, HTM for video anomaly detection [3] is a novel topic and is therefore naturally limited on several fronts. One of the main limitations is the lack of labeled anomaly data that suits the nature of HTM, because most datasets are made for use with deep learning approaches. Another problem is the lack of works related to applying HTM on video-based problems. Finally, while there are other methods that can be used for video anomaly detection, none of them are based on the same premises as HTM. This means that there is a major lack of methods to use for the purpose of benchmarking and comparison.

It should also be mentioned that the HTM theory described in this thesis is not the first generation [4], which was probabilistic in nature. The HTM theory described in this thesis is actually the third generation [5, 6], which builds upon the second generation [7, 8]. The second generation is often referred to as Cortical Learning Algorithms (CLA), which made the move from probabilistic modelling to Sparse Distributed Representation. The third generation builds upon the second generation by introducing

concepts such as sensorimotor inference and The Thousand Brains Theory. The first generation had fundamentally different inner workings, but shared a lot of the terms with the current generation. This has made researching HTM challenging as there are many research papers published that refer to the first generation.

Finally, it needs to be noted that Numenta, which is the company behind HTM, is a private for-profit company. This means that it is in their best interest to present HTM as a highly functional and powerful machine learning algorithm. This can lead to unhealthy optimism, which has caused complications during the work of this thesis. One should therefore be critical when reading research papers from Numenta, and always keep in mind that they might not only be research papers but may also be marketing materials. It should also be mentioned that this thesis is relying on the community fork [9] of HTM because Numenta has stopped development on the original codebase.

1.4 Contributions

This thesis explores the use of HTM for anomaly detection in videos, and introduces a new type of HTM architecture called Grid HTM, which requires a vigorous understanding of HTM. The main contributions are therefore achieved through the objectives introduced in Section 1.2. How this thesis achieves those objectives is described below:

Objective 1 *Introduce HTM and give a deep understanding of the inner workings, the strengths, and the weaknesses. While also being easy to grasp for readers with a machine learning background.*

This objective is achieved in Chapter 2, where HTM is explained. This chapter also acts as an organization of HTM related research, supported with visualizations and a simpler language, making it easier for people with a machine learning background to understand. It is also important to mention that not only does this chapter include the main HTM research published, but it also contains little tidbits of community research and ideas which are otherwise hard to come by.

Objective 2 *Develop and outline a theoretically sound HTM architecture that can be applied for anomaly detection in complex videos.*

This objective is achieved in Chapter 3, where Grid HTM was introduced. Not only does it present a novel way to apply HTM on video-based problems, it also uncovers the reasoning behind the design decisions that were made as well as providing thorough analysis. This objective is further achieved through the Grid HTM paper, which is a product of this thesis, and can be seen in Appendix A.

Objective 3 *Perform experiments, discuss the results, and lay out potential future work for Grid HTM. The experiments will vary in difficulty, complexity, and will focus on different use cases.*

This objective was achieved in Chapter 4, where three different experiments were performed. The first experiment showcased that HTM and Grid HTM can indeed perform anomaly detection on simple and clean videos. The second experiment showcased the performance of Grid HTM on a complex surveillance video, which showed promising results. The third experiment showcased the ability of Grid HTM to detect segments in a video, where noisy videos of sperm were used for increased challenge.

Now that the three objectives have been achieved, it is possible to answer the thesis question: **Is HTM viable for anomaly detection in videos?** The experiments show that with proper data and further refinements, Grid HTM and other HTM based architectures can indeed be used as anomaly detection systems for videos.

Additionally, during the course of this thesis, a contribution have been made to the HTM community in the form of uncovering and reporting a bug related to the technical implementation of HTM [10]. It is also important to reiterate that this thesis acts as a general guide to HTM from a machine learning perspective, which is something that is sorely needed due to the low accessibility of the HTM field.

Finally, all the work done in this thesis, including the Grid HTM source code and the various experiments, is publically available at Github [11].

1.5 Research Methods

Due to the novelty of this thesis, a single research method could not be used. Instead, a combination of multiple research methods was employed. For most of the thesis, an exploratory research method was used with the context of answering the thesis question, this was used due to the novelty of this thesis and the lack of related works. The result of this exploratory method is Grid HTM, which began from a simple starting point and was shaped through exploring improvements and solutions to problems that arose.

For the surveillance experiment, a qualitative research method was used to determine the effectiveness of Grid HTM. This is due to the lack of labeled data which made it hard to quantify, and that real life surveillance examples are complex by nature. The result is that the effectiveness was determined not through numbers, but through whether the results were qualitatively reasonable.

As for the other experiments, a more quantitative research method was used. This was made possible in the bouncing ball experiment due to the controlled nature of the experiment. In the sperm experiment, this was

made possible due to the use of a benchmark, which allowed for direct comparisons.

1.6 Thesis Outline

This thesis consists of five chapters, where Chapter 1 and Chapter 2 are introductory and contain relevant background information for the understanding of the proceeding chapters. Chapter 3 and Chapter 4 present the work done during this thesis. Chapter 5 neatly summarizes this thesis and presents areas in which there can be performed further work. More details for each of the chapters, except this one, is presented below.

Chapter 2: Background

Chapter 2 covers the required knowledge for the proceeding chapters, as well as a short section about ethical considerations. It is split up into three parts. The first part covers deep learning and its history, and has an increased focus on the parts of deep learning that are especially relevant for this thesis, such as generative models.

The second part covers anomaly detection. More specifically, it covers the definition, challenges, and recent work within anomaly detection. It also discusses smart surveillance, which is a subset of anomaly detection specifically meant for surveillance purposes.

The third and final part covers HTM theory. It starts off by introducing the biological ties to HTM and the pipeline of an HTM model. This is then proceeded by a detailed account of the inner mechanisms of an HTM model and how learning is performed. It then finishes by introducing use-cases, related work on the use of HTM for anomaly detection and how it performs, and introduces similarities between recent developments within HTM theory and recent developments within deep learning.

Chapter 3: Grid HTM

This chapter introduces Grid HTM, a novel HTM architecture for the purpose of anomaly detection in videos, which is the main contribution of this thesis.

Grid HTM is presented gradually as problems occur and have to be solved. Some of these problems are invariance and unstable anomaly output. The technicalities are explained with the help of figures and various analyses. This architecture is then used in Chapter 4 when performing experiments.

Potential use cases are presented, such as reducing the need for expensive live monitoring as well as anomaly data labeling. The biological plausibility is also briefly discussed, and concludes that Grid HTM may loosely be considered a cortical region.

Chapter 4: Experiments and Results

This chapter presents three experiments that aim to explore the capabilities and challenges of HTM and Grid HTM in the context of anomaly detection in videos. The first experiment is a controlled experiment where a computer-generated ball is bouncing with anomalies inserted. The aim of this experiment is to test whether the capabilities of HTM apply for videos, as well as the performance of Grid HTM on the same task.

The second experiment showcases the performance of Grid HTM on a surveillance video with technical anomalies. Additionally, several key points of interests and the respective outputs of Grid HTM are shown in order to get a better understanding of its capabilities.

The third experiment further explores the ability of the Grid HTM to detect segments in videos, which was discovered in the previous experiment. The videos that are used in this experiment are videos of sperm that contain several segments.

Chapter 5: Conclusion and Further Work

In this chapter the thesis is summarized, and the thesis question presented in Chapter 1 is answered along with potential further work and the main contributions of this thesis.

The contributions are presented as the description of how the various objectives were achieved, as well as answering the thesis question. This is a shorter version of the contributions mentioned in the introduction, and is presented conclusively.

This chapter also presents potential future work, which is further divided into sections focusing on different aspects such as datasets, Grid HTM, HTM and deep learning, and general research progress.

Chapter 2

Deep Learning, Anomaly Detection, and Hierarchical Temporal Memory

This chapter covers relevant background information on Deep Learning, Video Anomaly Detection, and HTM theory. It also briefly mentions ethical considerations.

2.1 Deep Learning

As mentioned in the introduction, deep learning has seen increased popularity in the past few years. Surprisingly enough, the field of deep learning can be traced back to 1958 when the Perceptron[12, 13] was first introduced.

2.1.1 Perceptron

The perceptron[12, 13] was a machine learning algorithm that was based on a simplification of the theory, at the time, about the inner workings of a neuron. The perceptron consists of three parts; the inputs x , the weights w , and the unit step activation function, which is shown in Figure 2.1.

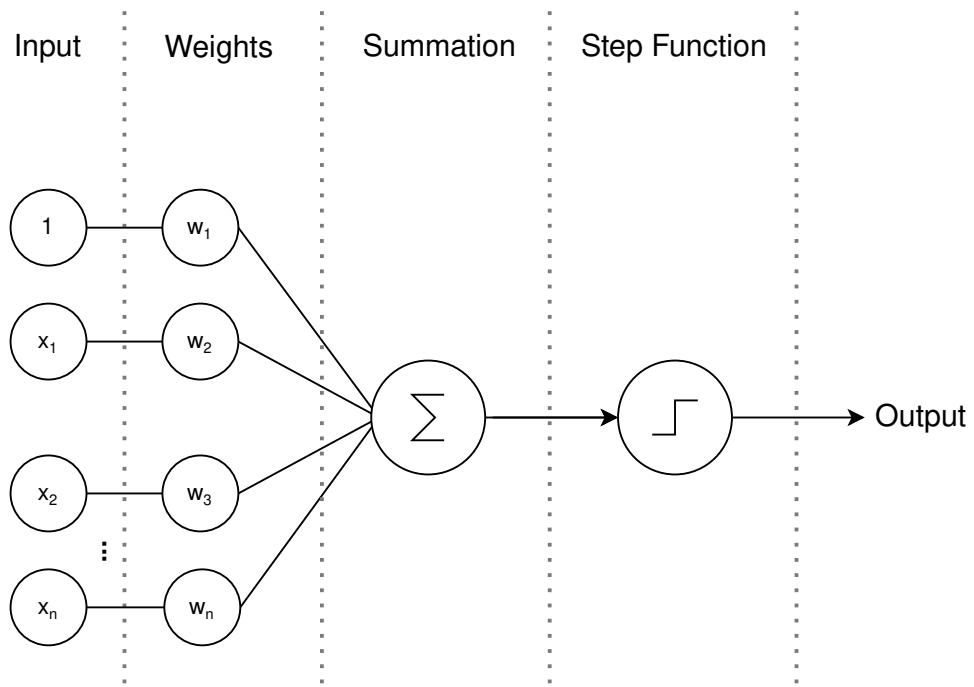


Figure 2.1: The original perceptron.

It also requires a label y corresponding to the input x . The perceptron predicts the label of the input x to be $\hat{y} = \text{sign}(x \cdot w)$. If $\hat{y} \neq y$ then it updates the weights $w = w + yx$, otherwise it leaves w unchanged. This is performed until the perceptron reaches convergence, which would happen when all the inputs can be correctly classified. In other words, the perceptron requires that the inputs are linearly separable.

As shown by Minsky and Papert [14], the perceptron was able to solve linearly separable problems such as the OR function, but was unable to solve the XOR function. For the latter, it was theorized by others that stacking perceptrons in multiple layers, known as a Multilayer Perceptron (MLP), which is shown in Figure 2.2, would be able to solve more complex problems such as the aforementioned XOR function [15].

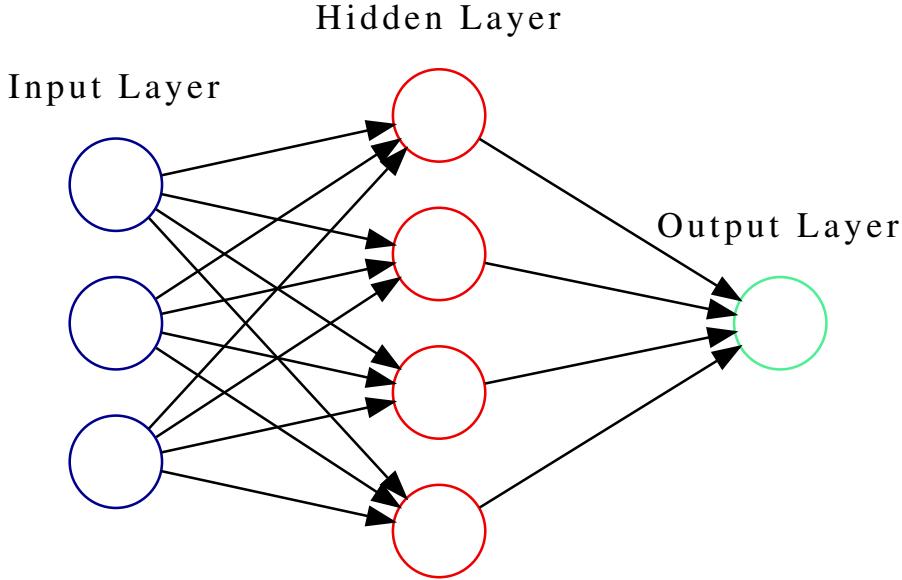


Figure 2.2: Example of a MLP.

Unfortunately, the work by Minsky and Papert [14] lead to the misconception that a MLP would have the same limitations as a single perceptron, and that further progress was impossible, which lead to the perceptron and other neuron-based approaches being largely abandoned [15]. This misconception, combined with a lack of computing power and no good approaches to train a MLP, lead to a decline in the research of neuron-inspired approaches.

It was not until the late 1980s [15] that the MLP approach experienced a revival, led by the increase in computational power and the introduction of backpropagation by Rumelhart, Hinton, and Williams [16], which allowed multilayer networks to be easily trained, and gave birth to modern neural networks.

2.1.2 Backpropagation

Backpropagation [16], which is shorthand for "backward propagation of errors", is a method for efficiently calculating the gradient of the error function, so that it is possible to adjust each individual weight with the purpose of lowering the error. The error function is a function that quantifies the difference between a prediction and the corresponding label, and is differentiable. An example of such an error function, is the Mean Squared Error (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

The result is that no longer did researchers have to manually engineer features, but could instead apply backpropagation to have the neural network automatically learn internal representations that expressed nontrivial features. An example is that no longer did researchers have to use line detection algorithms, instead the MLP could learn to represent a collection of pixels as a line automatically.

2.1.3 Neural Networks

Modern MLP networks, often referred to as just "neural networks", are similar to the MLP networks made in the 80s. The main difference is that each neuron in a MLP can have an arbitrary activation function, such as the sigmoid function and the Rectifying Linear Unit (ReLU) [17] function, which adds nonlinearity into the neural network and allows it to solve complex problems. They also contain more hidden layers, as seen in Figure 2.3, hence the term *deep* learning.

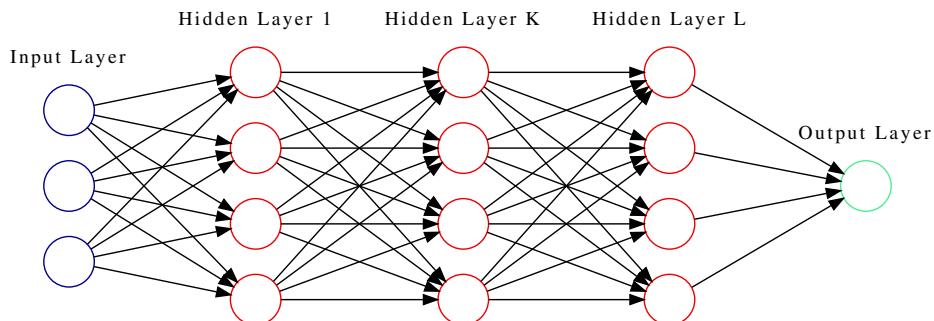


Figure 2.3: Example of a neural network.

2.1.3.1 Learning

Neural networks use gradients to learn. The gradients are first calculated using backpropagation which are then used to update the weights with the goal of reducing the error. This can be achieved using gradient descent, where one would calculate the gradient and descend towards the optimum using the entire dataset, but this is computationally expensive. An alternative is to use an optimizer such as Stochastic Gradient Descent (SGD) [18], which calculates the gradient and descends using a random subset of the dataset. A visual comparison between the two methods is visualized in Figure 2.4:

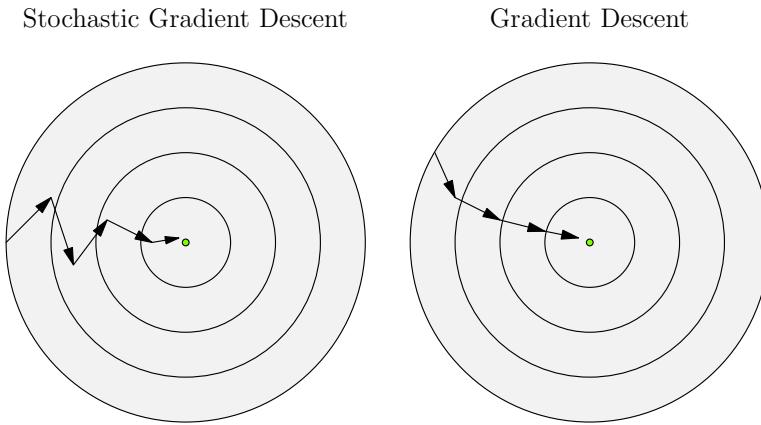


Figure 2.4: Comparison between SGD and gradient descent. The green dot in the middle represents the optimum.

Yet with all the advancements within the field of neural networks, it still did not have the popularity that can be observed today. There are several reasons for this, one of them arguably being the introduction of the modern Support Vector Machine (SVM) [19], which overshadowed neural networks [20].

The other reason is that neural networks did not have a lot of applications at the time, this being due to a lack of invariances which caused poor performance in machine vision, and no proper way to represent features across time which caused poor performance in time-series modelling. There was also a lack of regularization methods, which caused the models to overfit and generalize poorly. Additionally, large networks suffered from vanishing or exploding gradients [21].

In order to improve upon regularization, techniques such as Dropout [22], Batch Normalization (BN) [23], and Data Augmentation [24] were introduced. New neural network architectures were invented, such as the Autoencoder (AE) [25], that forces the model to construct an internal representation under constraints, which has a regularizing effect.

As for representing features across time, architectures such as Recurrent Neural Networks (RNNs) [26, 27] were invented. Later, attention [28]-based models, such as the Transformer [29] model, were introduced.

The vanishing and the exploding gradient problem was alleviated by implementing, among other techniques, residual connections [30] between layers.

2.1.4 Convolutional Neural Networks

In an attempt to solve the problems regarding invariance and to improve the performance in machine vision tasks, Convolutional Neural Networks (CNNs) were introduced by Lecun et al. [31] in 1998. A CNN offers several

properties that are useful, such as translational invariance [31]. These properties stem from the three architectural ideas in a CNN:

- Local receptive fields
- Shared weights
- Spatial sub-sampling

Through the use of data augmentation, it is also possible to not only improve the effectiveness of the aforementioned invariances, but to also introduce a degree of rotational and scale invariance. A typical CNN architecture is shown in Figure 2.5.

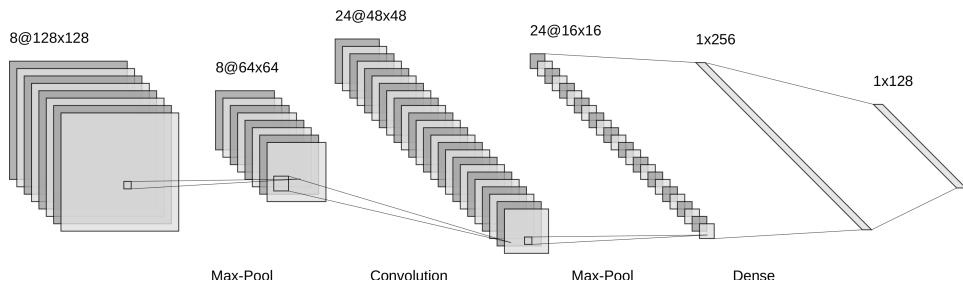


Figure 2.5: A typical CNN architecture.

However, the progress within the context of machine vision stagnated due to a lack of computing power when training large models as well as a lack of sufficiently large datasets. This stagnation lasted until 2012 when AlexNet was introduced by Krizhevsky, Sutskever, and Hinton [32], which marked a major turning point in the history of deep learning, as their record-breaking results proved that deep learning was they way forward for solving complex machine learning problems.

In addition to creating a unique CNN architecture, they also made it run on a GPU and made the code publicly available. This was not the first case of running deep learning on a GPU [33], but it can be argued that it was this paper that popularized it. Using GPUs meant that a vast amount of computational power, for the purpose of training deep learning models, became unlocked. This also paved the way for frameworks such as PyTorch [34] and Tensorflow [35], which have democratized deep learning and made it into what it is today.

2.1.5 Generative Models

Some of the latest progress within deep learning can be attributed to generative models, such as Generative Adversarial Networks (GANs) [36] and Variational Autoencoders (VAEs) [37]. Generative models take some input, and then generate realistic output. The input could be anything from random noise to a real sample.

The idea is that generative models learn a distribution which describes

the data domain represented by the training data, and can then generate synthetic data from that distribution by simply sampling it.

Generative models come in all shapes and sizes, but a standard GAN usually consists of just a generator and a discriminator. The generator takes some inputs (usually some random noise) and generates some samples. The batch of generated samples are then combined with a batch of real samples and mixed up. The discriminator then has to decide for every sample whether it is real or generated. This is essentially a two-player game in which the generator attempts to fool the discriminator. A typical GAN pipeline is shown in Figure 2.6 below.

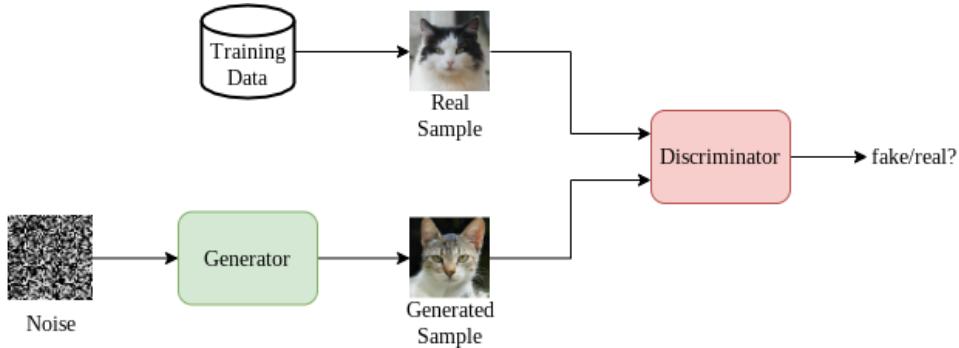


Figure 2.6: Example of a GAN.

In order to train the weights in such a configuration, the generator and discriminator are trained in alteration. To train the discriminator, the weights of the generator are fixed and backpropagation is performed using the error based on whether the discriminator made the right prediction. To train the generator, the weights of the discriminator are fixed and backpropagation is performed using the error based on whether the discriminator was fooled or not.

While the training method of a GAN is straight forward, actually training a GAN well is hard and faces several challenges such as mode-collapse, non-convergence, and instability [38, 39]. To alleviate these challenges, modifications to the standard GAN have been proposed such as the Wasserstein GAN [40] and the Unrolled GAN [41].

There also exists variations of the standard GAN for the purpose of different types of generation. One example is the CycleGAN [42] which makes it possible to perform domain translation, such as turning a horse in an image into a zebra, without paired data. Another example is CGAN [43] in which both the generator and discriminator are conditioned on some auxiliary information, which makes it possible to generate data within a specific context. For instance, instead of generating arbitrary human faces, a CGAN can generate faces of a specific age.

2.1.6 Disadvantages of Deep Learning

A disadvantage for deep-learning models in general is that they are susceptible to noise in the dataset [44, 45], which leads to decreased classification accuracy and poor prediction results. Due to the nature of training deep learning models, they are also in most cases not self-supervised and therefore require constant tuning in order to stay effective on changing data. Not to mention that they require a lot of data before they can be considered effective, and that performance increases logarithmically based on the volume of training data [46]. This is visualized in Figure 2.7.

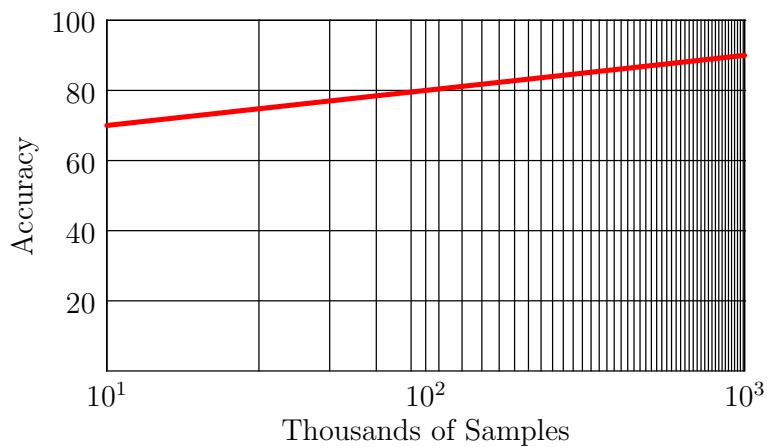


Figure 2.7: Accuracy increase in relation to size of dataset.

Deep learning models also suffer from issues with out-of-distribution performance, where a model might perform great on the dataset it is tested on, but performs poorly when deployed in real life. This could be caused by selection bias in the dataset or when there are differences in the causal structure between the training domain and the deployment domain [47]. There is also a lack of explainability.

2.1.6.1 Explainability

As stated by Barredo Arrieta et al. [48], as "black-box" approaches such as deep learning surged in popularity, many realized that they offered poor explainability. While it is known *how* the models make their decisions, their huge parametric spaces make it unfeasible to know *why* they make those predictions. Combined with the vast potential that deep learning offers in critical sectors such as medicine, has lead to an increase in focus on developing approaches that offer explainability.

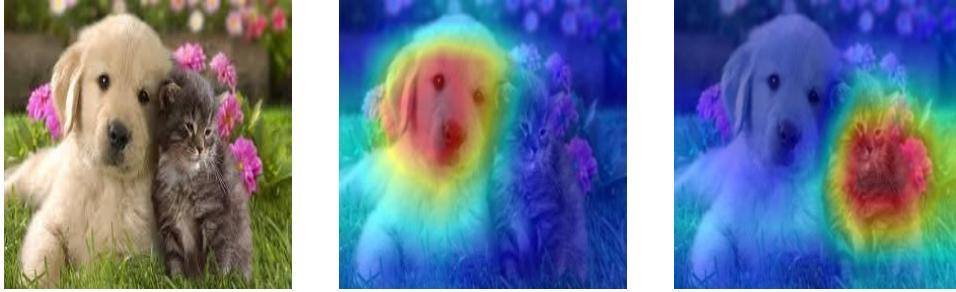


Figure 2.8: Grad-CAM visualization for the labels "dog" and "cat". Taken from Gildenblat and contributors [49].

Approaches such as Grad-CAM [50] (see Figure 2.8) and Guided Backpropagation [51] offer improvements in that regard, but these approaches are not made with generative models in mind. In fact, there are very few explainable AI approaches for generative models [48].

2.2 Anomaly detection

As reviewed by Pang et al. [52], anomaly detection is often defined as detecting data points that deviate from the general distribution of the data, this also often includes quantifying the level of deviation. Unlike other problems within machine learning and statistics, anomaly detection deals with unpredictable and rare events, therefore adding complexities to problems. Some complexities are as follows:

- **Unknowns** Anomalies are associated with many unknowns which do not become known until the anomaly happens. Michałowska et al. [53] and Fan et al. [54] are works that address this.
- **Rarity and class imbalance** Anomalies are by definition rare instances, which means that it becomes difficult to create a balanced dataset. Devi, Biswas, and Purkayastha [55] review the current solutions to this problem.
- **Heterogeneity** Anomalies can take form in many ways, and as such one class of anomalies can be vastly different from another. Approaches such as the one introduced by Datta, Muthiah, and Ramakrishnan [56] have been proposed to alleviate the problem.

These complexities make it hard to apply traditional deep learning methods for anomaly detection, because they are designed to be trained with pairs of $\{input, target\}$ in mind.

2.2.1 Deep Learning and Anomaly Detection

The current state-of-the-art algorithms for anomaly detection are numerous, but the main approach is achieved by using deep learning [52]. As previously mentioned, traditional deep learning approaches are hard to

apply for anomaly detection. Instead, a popular approach is to use generative deep learning models such as GANs [52, 57, 58] to generate synthetic data and compare it to real data in order to detect anomalies. This approach is based on the assumption that the model will only be able to generate data similar to what it has been trained on, and therefore fail when an anomalous event occurs. A variation of GAN which is specifically designed for anomaly detection is GANomaly[57], shown in Figure 2.9, which effectively compares image encoding latent space instead of image distributions.

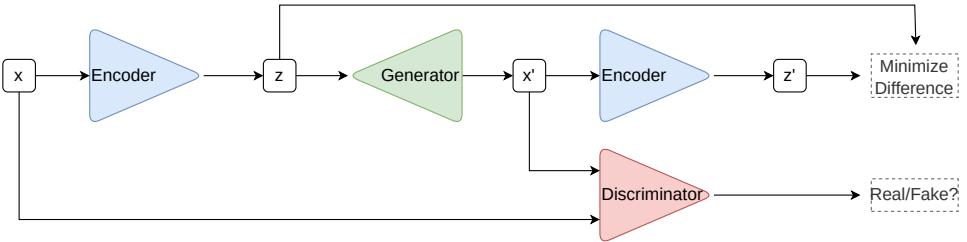


Figure 2.9: GANomaly [57], a variation of GAN for Anomaly detection.

The advantage is that GANs are generally good at generating realistic data, especially when it comes to images. The disadvantage is that GANs are very hard to train and may give suboptimal results given that they try to generate good synthetic data rather than directly detect anomalies. The training data also needs to contain all possible non-anomalous classes of events, which may not be a realistic expectation.

Another common approach is using AEs, which aim to minimize the reconstruction error from a learned feature representation space [59, 52, 58]. The assumption is that anomalies are more difficult to reconstruct than normal data, hence the reconstruction error will be high and can therefore be used as a metric to detect anomalies.

As previously stated, generative models suffer from a lack of explainability, and since generative models make up most of the state-of-the-art approaches in anomaly detection, it is safe to say that there is a lack of explainability in the field of anomaly detection. There are also variations of the aforementioned approaches, such as Adversarial AEs, but the core idea is the same: To get an anomaly measure using some sort of generated or reconstructed data.

2.2.2 Smart Surveillance

Smart surveillance, which is the use of automatic video analysis specifically in surveillance, has seen rapid development since its inception. Zhu, Chen, and Sultani [58] and Sreenu and Durai [60] present and summarize recent progress for anomaly detection in video for surveillance purposes, where the most promising methods are achieved by using convolutional AEs and GANs. More specifically, memory augmented AEs [61] and future

frame prediction GANs [62] are some of the most promising approaches. In general, the results show that the deep learning approaches have a high degree of accuracy, and are consistently improving. This is not surprising given that generative models are popular in anomaly detection, as mentioned in Section 2.2.1.

The work by Zhu, Chen, and Sultani [58] also discusses problems with using deep learning approaches for anomaly detection, and further emphasizes the complexities mentioned in Section 2.2. One of the examples that it uses is about a bicycle on campus being wrongly classified as an anomaly, which is related to the aforementioned issue with the non-realistic requirement that the training data must contain all possible classes of non-anomalous events.

2.3 Hierarchical Temporal Memory

Today's machine learning algorithms aim to solve complex problems by simulating a substantial amount of mathematically defined neurons. These neurons are vastly simplified compared to the neurons in the brain and therefore do not have the complexity required to solve complex problems with an accuracy and level of generalizability comparable to the brain. Hawkins et al. [63] introduces HTM theory which aims to outline a machine learning algorithm which works on the same principles as the brain and therefore solves some of the aforementioned issues.

The brain consists of layers that have been added throughout evolution. The inner layers are responsible for primal intelligence such as hunger, sex and instincts. HTM theory specifically aims to roughly simulate the neocortex, which is the outer layer of the brain tasked with advanced logic. It is important to note that HTM only attempts to *estimate* the activity in the neocortex, unlike Spiking Neural Networks and others which aim to *accurately simulate* the activity of the neocortex [64].

2.3.1 Structure

HTM aims to replicate the structure of the neocortex which is made up of cortical regions. Cortical regions consist of cortical columns, where each column is divided into layers height-wise. These cortical columns are made up of mini-columns, which in turn are made up of neurons. Figure 2.10 visualizes the structure of a cortical region according to HTM theory.

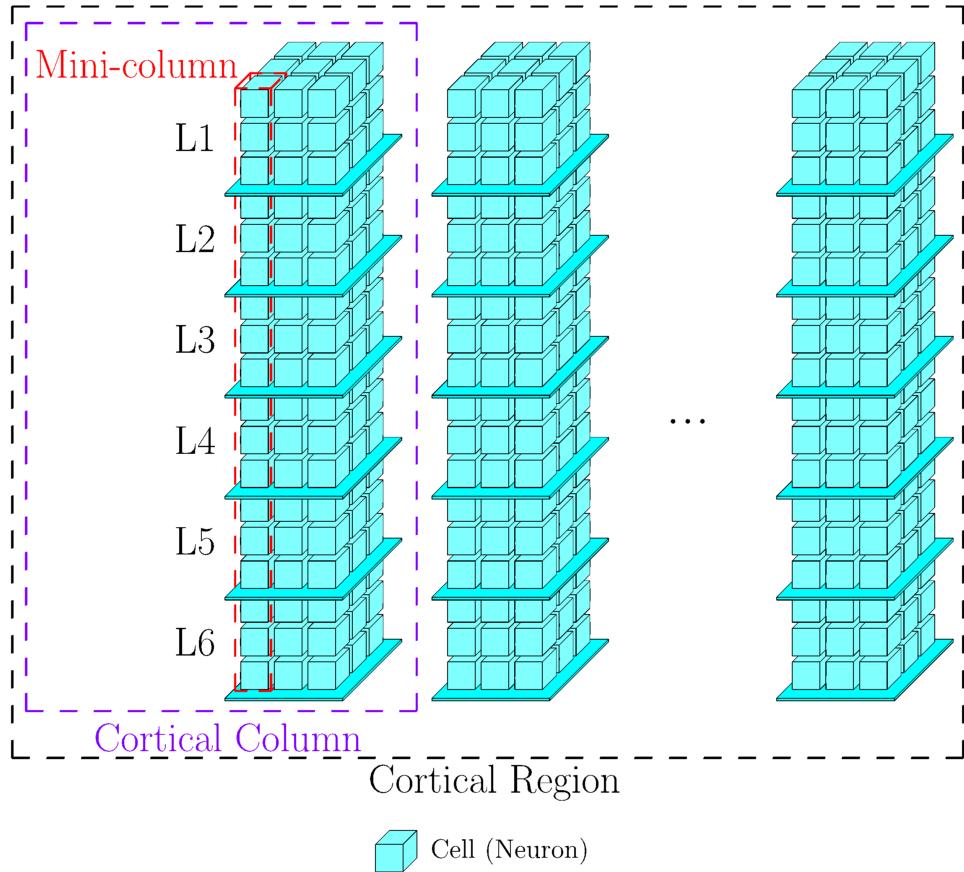


Figure 2.10: Visualization of a cortical region structure in the neocortex according to HTM theory.

There is still uncertainty regarding the purpose, and even the existence, of each layer, but the following has been concluded [65, 66]:

- L6** Deals with input information.
- L5** Motor outputs to body.
- L4** Sensorimotor inference.
- L3** Temporal Memory.
- L2** Connects to other cortical columns and performs voting.
- L1** Connections to cortical regions further up.

Current HTM implementations only model cortical columns, and layer-wise only L2/L3 but can be extended to model L4 as well [65].

Neurons in HTM theory are different from neurons in traditional machine learning [67]. The term neuron in traditional machine learning is very misleading and since it is mathematically derived, has actually very little in common with a biological neuron. A biological neuron does not perform back propagation but learns by strengthening and weakening inter-neural

connections (synapses), which is something that the HTM neuron attempts to model through Hebbian-like learning. The difference is shown in Figure 2.11.

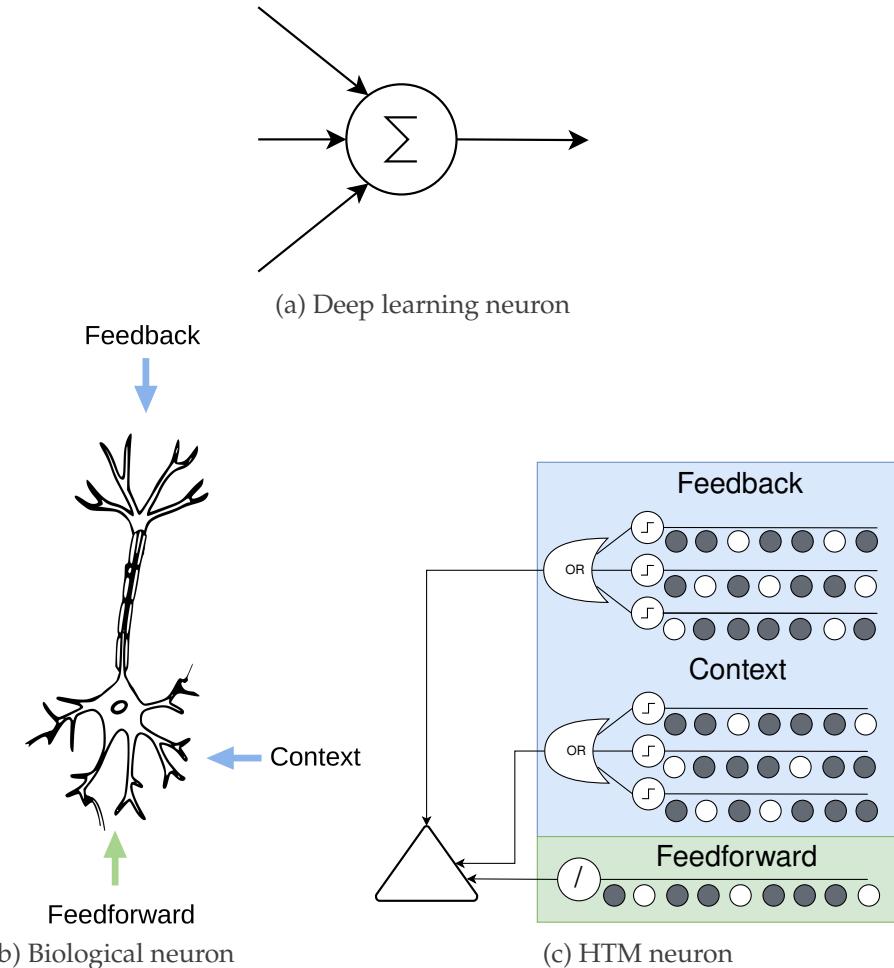


Figure 2.11: Comparison of neurons: Traditional Machine Learning (A), Biological Neuron (B), HTM Neuron (C). Source: [63]

The HTM neuron has three inputs [67]:

- Feedforward, which is the input data
- Context, which is data from neighboring neurons and acts as a prediction mechanism for the next feedforward input
- Feedback, which is feedback from other neurons in the hierarchy and acts as a prediction mechanism for a sequence of feedforward inputs

How this type of neuron operates in HTM implementations will be covered in greater detail later in this section.

2.3.2 Common Algorithm

HTM theory states that there is a common algorithm for the intelligence. That the signals from hearing, vision, and touch are at the core processed by the same common algorithm. By extension, this means that HTM networks should be able to solve all kinds of logical tasks.

2.3.3 Sparse Distributed Representation

HTM theory introduces Sparse Distributed Representation (SDR) as a way of representing data in HTM and can be thought of as a bit-array. Each bit theoretically corresponds to a neuron in the neocortex and also represents some semantic information about the current data. This opens up for all kinds of mathematical operations, for instance it is possible to compare the semantic similarities between two SDRs by simply performing a binary AND operation, as seen in Figure 2.12.

SDR A: 1010000110101110011001001...0100
SDR B: 1010010011001010111010101...0011

Figure 2.12: Semantic similarities between the two SDRs A and B.

Observations of the brain has found that at any given point in time, a small percentage of neurons are activated and an SDR aims to keep this property by having a small percentage of bits be 1 at any given point. A common value is 2% in order to mimic the sparsity of active neurons in the neocortex. Having this property means that the chance of two bit-patterns with different semantic meanings coinciding, for instance due to bit-flips caused by noise in the data, is astronomically low and is what makes HTM robust to noise. That being said, sparsity can be configured to be any value, as seen in Figure 2.13.

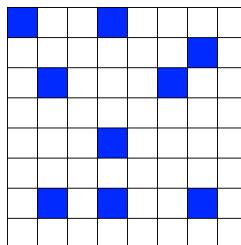


Figure 2.13: Example representation of an SDR with a length of 64 and a sparsity of 14.1%, visualized as a 2D grid.

2.3.4 Encoders

To convert real-world data into an SDR, there is a need for an encoder in the pipeline. These encoders can be designed to take potentially any data and convert it into an SDR with an arbitrary sparsity. Given the fact that they

may have an arbitrary sparsity, the output SDRs created by the encoder are sometimes referred to as just binary arrays.

Writing an encoder is no easy task as it is important to keep semantic similarities between values. This also means that the encoder is perhaps the most important part of an HTM pipeline to get right as it is the part that can limit the system the most. A biological example would be an eye that takes in visual information and converts it into an SDR so that it can be processed by the neocortex.

There are principles that should be followed in order to create a good encoder:

- Semantically similar data should result in SDRs with overlapping active bits
- The same input should always produce the same SDR
- The output must have the same dimensionality (total number of bits) for all inputs
- The output should have similar sparsity (similar amount of one-bits) for all inputs and have enough one-bits to handle noise and subsampling

As of now, there exists encoders for numbers, categories, geospatial locations, and dates. Figure 2.14 shows a cyclical date encoding.

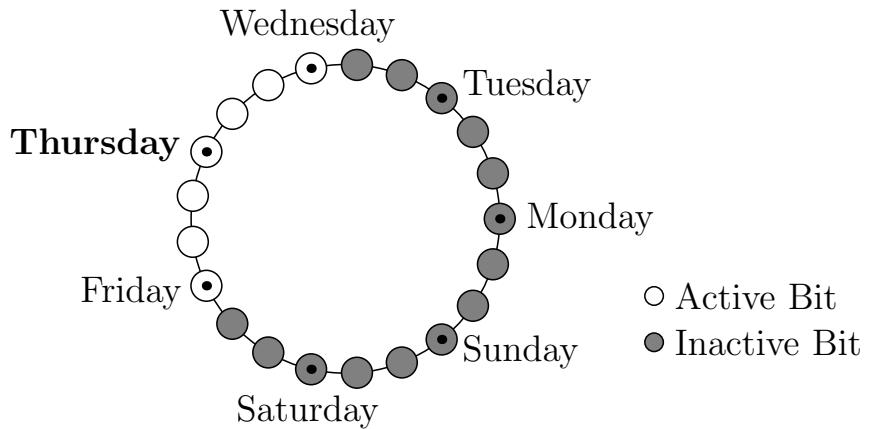


Figure 2.14: Visualization of a cyclical date encoding, which is currently encoding Thursday. Note that also Wednesday and Friday are included in the encoding of Thursday in order to emphasize that Wednesday and Friday are both equally distanced from Thursday.

Some applications may require anomaly detection on multiple values at once, the correct approach then is to encode the values into SDRs one by one and then concatenate them into a single SDR before passing it to the HTM.

2.3.4.1 Encoding Visual Data

Several approaches for encoding visual data have been proposed. An example is a neuroscientific approach which replicates how the eye works [68], another example is the approach by Fallas-Moya and Torres-Rojas [69] which uses scale-invariant feature transform (SIFT) to find points of interest in images and encode that information as an SDR. A simple threshold, as seen in Figure 2.15, can also be applied to extract information into an SDR-friendly format.

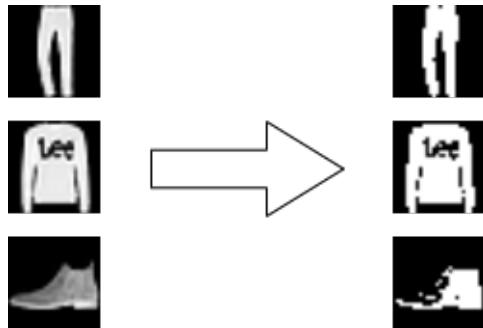


Figure 2.15: Example thresholding on the Fashion MNIST dataset [70].

There are also deep learning approaches such as the one proposed by Zou et al. [71] which uses a Convolutional Neural Network (CNN) as part of the encoder. It creates an SDR by storing the top n -features in a feature map as ones and set the rest to zeros in order to achieve binary values suitable for the SDR.

There are several reasons why these approaches might not perform well in videos. Keypoint detectors such as SIFT cause a lot of noise due to points shifting as lighting is changing. The distribution of the points within a frame is also constantly changing, which adds even more noise.

The reason for why a direct binary threshold encoding might not perform well is due to the fact that it is neither position nor scale invariant and as such breaks the first principle of creating a good encoder. For instance, if it is desired that two pictures of the same object, but in different scales have more or less the same semantic meaning, then a direct binary encoding is not going to work. Direct binary threshold encoders also lead to loss of information, and is hard to perform for complex objects.

An encoder which transforms convolutional feature maps into SDRs could be the answer, but the issue is converting the dense representation of the feature maps into SDRs. Directly encoding them into SDRs by treating each value in the feature map as a float and converting it into its own SDR quickly becomes intangible due to the processing and memory requirements. Figure 2.16 shows only a small subset of all the feature maps in a single layer, which highlights how intangible this approach is.

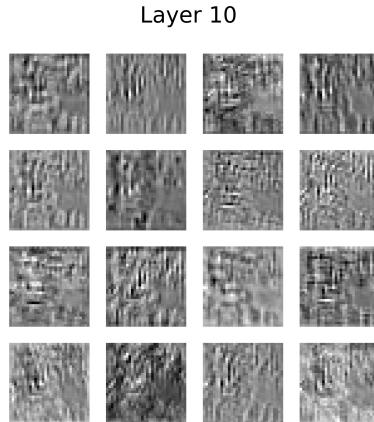


Figure 2.16: Select few feature maps in the 10th layer of ResNet18 [30] trained on ImageNet [72].

Additionally, minor variations in the feature map would cause major variations in the resulting SDR. Alternatively, one could follow Zou et al. [71] and binary threshold the top- n features, but this leads to its own problems such as loss of information and that the information contained in the top- n features is often undefined in models trained for complex tasks. It is also undefined what the top- n features represent when there are no strong activations in the feature maps.

Seeing as there is currently no encoder that can produce sensible SDRs from high dimensional data such as videos, a natural conclusion to make is that HTM should be applied differently, and that a new type of architecture using HTM should be explored for this purpose.

2.3.5 Learning

Similar to mammalian intelligence, HTM is designed to work on streaming data. It does not operate with batches like traditional machine learning, but rather with streaming data that may be changing over time.

The learning mechanism consists of two parts; the Spatial Pooler (SP) and the Temporal Memory (TM) algorithm. The latter is also commonly referred to as Sequence Memory. Together they make up multiple HTM neurons.

The spatial pooler takes SDRs produced by the encoder, and uses Hebbian like learning to extract semantically important information into output SDRs. These output SDRs usually have a fixed sparsity of about 2% due to the fact that the spatial pooler aims to produce SDRs that have similar sparsity to what has been observed in the neocortex, but this can be configured at will and is dependent on the problem at hand.

The temporal memory algorithm, on the other hand, simulates the learning

algorithm in the neocortex. It takes the SDRs formed by the spatial pooler and does two things:

- Learns sequences of SDRs formed by the spatial pooler
- Forms a prediction, in the form of a predictive SDR, given the context of previous SDRs

A typical HTM pipeline can be seen in Figure 2.17.

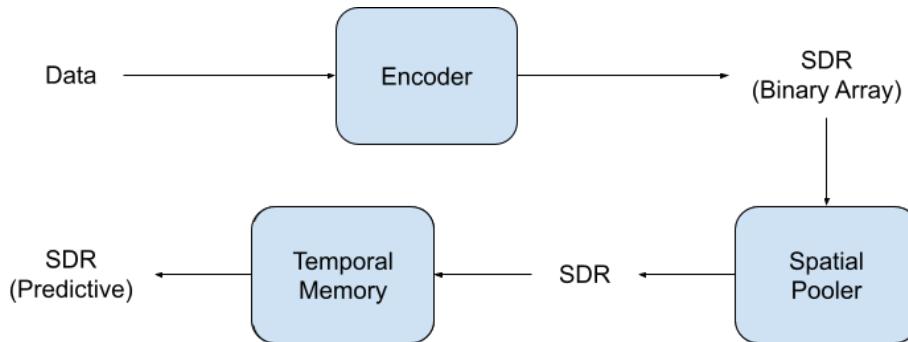


Figure 2.17: A typical HTM pipeline. A common next-step could be to use a classifier to convert the predictive SDR into a classification.

This learning mechanism gives HTM systems the property of on-line learning, meaning they learn as they go. There is no batch training because each input into the HTM system will update the system. The system effectively builds a predictive model of the data and learns by trying to minimize the error between the true values and the predicted values. This means that the system will continuously adapt to a changing environment.

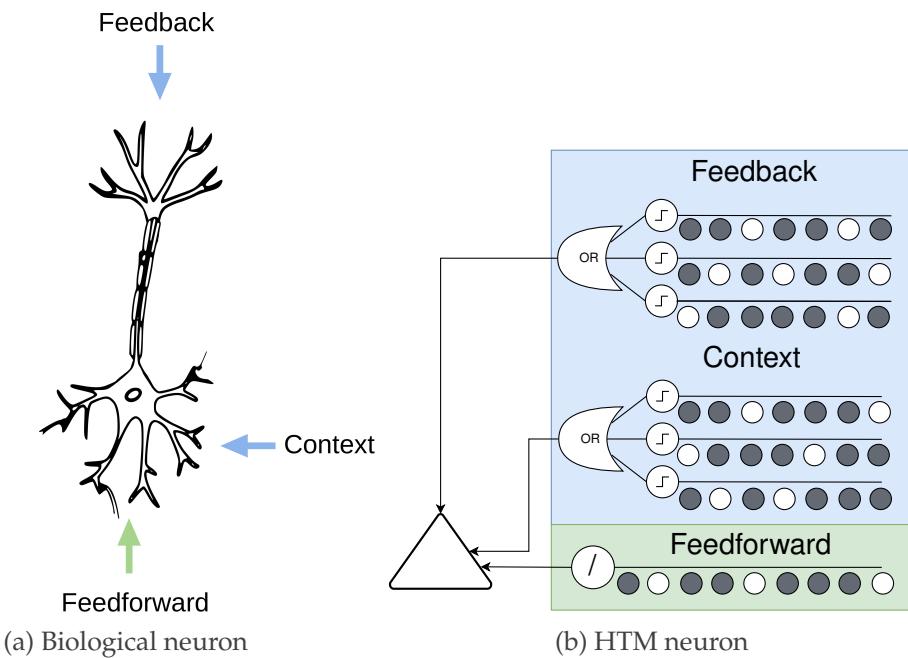


Figure 2.18: Biological neuron and HTM neuron, the colors show which part the Spatial Pooler (SP) and the Temporal Memory (TM) covers.

Figure 2.18 shows that a spatial pooler combined with a temporal memory forms the HTM neuron, where the color green indicates the responsibility of the spatial pooler and blue indicates the responsibility of the temporal memory.

2.3.5.1 Spatial Pooler

The spatial pooler consists of columns (mini-columns), where each column has a receptive field covering the input. In technical implementations of spatial poolers, the columns exist in name only and could be thought of as nodes instead. A column can cover parts of the input or the entire input, the range being referred to as the **potential radius**. During initialization, each column creates random connections to a percentage of the bits in the input space within its receptive field, this gives each column a unique **potential pool** when there are overlaps of receptive fields caused by a large potential radius. This is visualized in Figure 2.19.

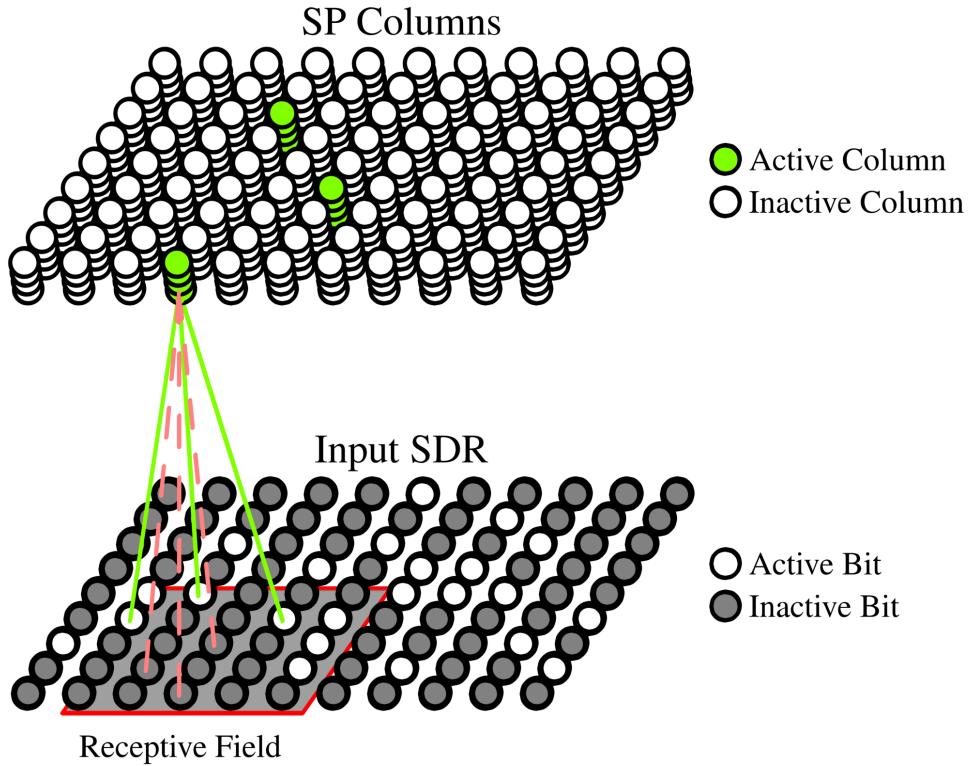


Figure 2.19: Visualization of the SP and the potential pool of one of its columns.

Each connection is described using a **permanence**-value which can be considered the "strength" of the connection, and it ranges between 0 and 1. During learning, the permanence value of the connections is increased or decreased depending on whether the corresponding bit in the input is active or inactive. When the permanence-value crosses above a **stimulus threshold**, the connection will be considered "active".

The amount of active connections for a given column is referred to as **overlap score**. If a column has a high enough overlap score which crosses the **overlap score threshold**, then the column will itself become active. The reason behind locking activation behind a minimum overlap score is to reduce the influence of noise in the input. Finally, out of all the active columns, only the top n columns with the most overlap score will be selected to be included in the SP output. The value n is chosen so that the SP output has a specific **sparsity**. Only the selected columns are allowed to learn (increase/decrease permanence). Figure 2.20 visualizes the mechanism that determines the output in the SP.

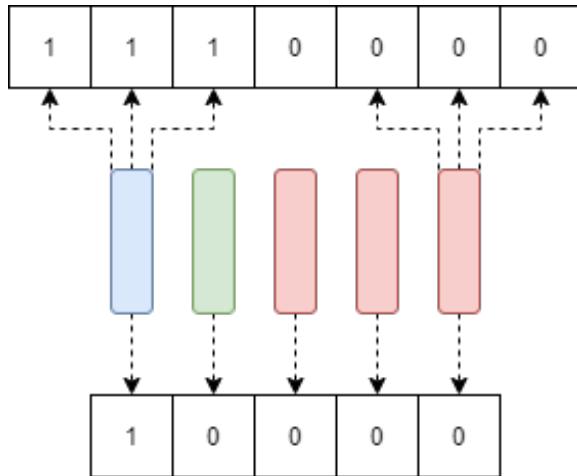


Figure 2.20: This figure illustrates how a spatial pooler works. All connections are above the stimulus threshold. The receptive field is 3 bits wide for each column. Overlap score threshold is 1, and $n = 1$. Red means inactive, green means active, and blue means active and selected.

Because only the active columns are allowed to learn, only a select few columns who got lucky during the random initialization will dominate the spatial pooler output and have a very high **active duty cycle**. Active duty cycle measures how often a column is active and ranges from 0 (never) to 1 (always).

To counter dominating columns, the spatial pooler uses **boosting**. The concept behind boosting is to "boost" the overlap score of underperforming columns and lower the overlap score of over performing columns. The result is that more columns learn and contribute to the output, which means that the spatial pooler can then process the input data with a finer granularity. This is also a form of regularization. One has to be careful with boosting, since it can cause instability in the spatial pooler output.

It is also possible to have **topology** in the output by selecting the columns to be included in the output by their local neighborhood, instead of comparing their overlap score globally.

All the aforementioned concepts are configurable in technical implementations.

2.3.5.2 Temporal Memory

The temporal memory consists of the columns that a spatial pooler outputs, but treats them as actual columns instead of "nodes". These columns consist of cells and can contain an arbitrary **number of cells** which defines the capacity of contexts that the temporal memory can express. Each cell in a column can connect to other cells in other columns using segments (more specifically, distal dendrite segments), where each segment consists of synapses connecting to other cells.

Essentially, it takes the "node" based representation of the SP output, and turns it into a new representation which includes state, or context, from previous time steps. It achieves this by only activating a subset of cells per column, typically only one per column. This allows the temporal memory to represent a pattern in multiple contexts. If every column has 32 cells and the SP output has 100 active columns and only one cell per column is active, then the TM has 32^{100} ways of representing the same input. The same input will make the same columns active, but in different contexts different cells in those columns will be active. This is visualized in Figure 2.21.

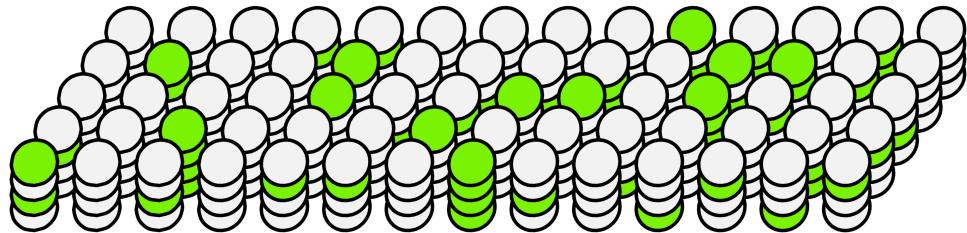


Figure 2.21: Visualization of the TM, with number of cells equal 4. Some columns are bursting.

The temporal memory algorithm consists of two phases. The first phase is to evaluate the SP output against predictions and choose a set of active cells. It does so by looking at the active columns and the cells they contain. If an active column contains predictive cells, then those cells are marked as active. If an active column has no predictive cells, usually caused by observing a new pattern for the first time, then the column "bursts" by activating all the cells that the column contains (see Figure 2.22). Otherwise, a cell is inactive.

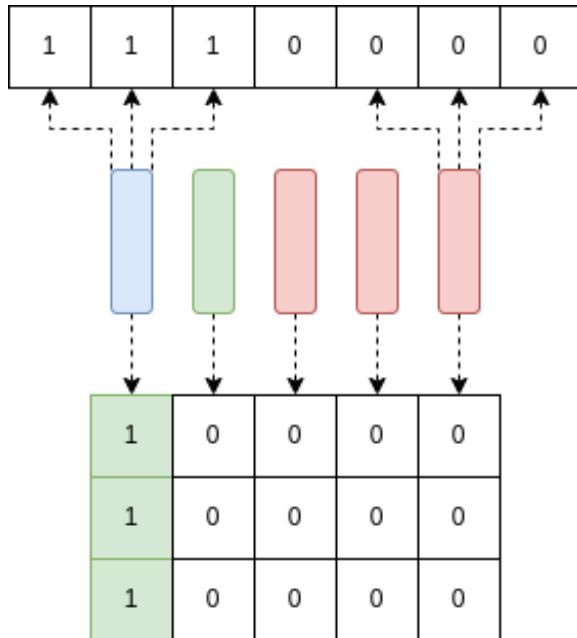


Figure 2.22: Expanded SP example with TM component where the number of cells is set to 3. The leftmost column is bursting (all 3 cells activated in green) due to the active SP output and due to containing no predictive cells.

At this point, the active cells represent the current input in the context of previous input. For each active column we look at the segments connected to the active cell(s). If the column is bursting we look at the segments that contain any active synapses, if there is no such segment we grow one on the cell with the fewest segments. On each of the segments that we are looking at, we **increase the permanence** on every active synapse, **decrease the permanence** on every inactive synapse, and grow new synapses to cells that were previously active. The algorithm also punishes segments that caused cells to enter predictive state, but which did not end up being active.

Since the TM can only grow synapses to cells that were active in the previous timestep, the TM struggles to express sequences of patterns over multiple timesteps, as has been discussed on HTM forums [73]. The solution has been to also encode some temporal information, such as the time of day [74, 73], so that it can use timestamps as anchor points for its contexts.

The second phase is to form a prediction by putting cells into a predictive state. For every segment on every cell, the number of synapses connected to active cells are counted. If the number exceeds an **activation threshold**, then the segment is marked as active and all the cells connected to the segment enter the predictive state. To summarize, a cell has three possible states:

- Active, if the column is bursting or the cell was in a predictive state in the previous time step and the column it belongs to is active.

- Predictive, if a connected segment is active, which is in turn determined by the amount of active synapses.
- Inactive, if none of the other states apply.

Figure 2.23 visualizes cells with different states in the TM.

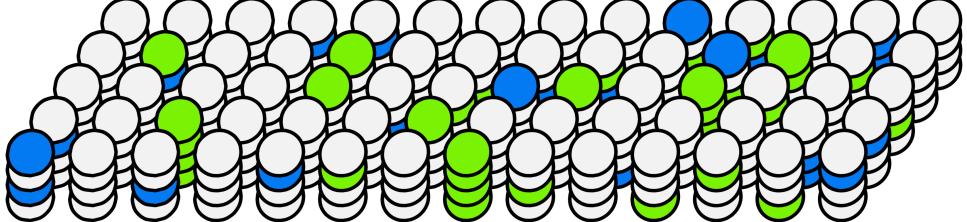


Figure 2.23: Visualization of the TM and the three states. Active in green, predictive in blue, and inactive in white.

One can configure how much the system can learn by setting the number of cells and the values by which permanence should be increased or decreased. If it is desired that the TM does not "forget" at all, then the permanence value by which synapses are decremented can be set to 0. If it is desired that the TM can only express patterns in the current context and the context of the previous time step, then the number of cells can be set to 2.

Finally, the TM compares the predictions P_{t-1} it made in the previous time step with the actual pattern A_t in the current time step and calculates an anomaly score:

$$anomalyScore = \frac{|A_t - (P_{t-1} \cap A_t)|}{|A_t|}$$

Which is a normalized value from 0 to 1. If the anomaly score is 1, then it means that none of the predicted columns matched the current active columns of the spatial pooler. If it is 0, then it means that all predicted columns matched the current active columns of the SP.

It is also possible to estimate the number of predictions being made by the TM at any time [75]. This is done by counting the number of predictive cells, and dividing them by the number of active bits required to express a pattern. As an example, if sparsity is set so that patterns have 60 active bits and the number of predictive cells is 120, then the estimated number of predictions is given as

$$numPredictions = \frac{predictiveCells}{activeBits} = \frac{120}{60} = 2$$

This is only an estimation, in reality the two patterns may have overlapping bits in their representations, and the number of active bits for each representation may have minor deviations.

2.3.6 Use Cases

The general use case for HTM is to perform anomaly detection. More specifically, Numenta has made example applications showcasing how HTM can be used in practice [76]:

- **Rogue Behavior Detection** which models normal behavior and detects anomalies, such as unusual use of files in a network [77].
- **Geospatial Tracking** which detects anomalies in the movement of people, objects, or material, using speed and location data [78].
- **Financial Monitoring** which detects anomalies in publicly traded companies by continuously modelling stock price, stock volume, and Twitter volume [79].

There are other examples on the use of HTM for real world applications. One of them is the work done by El-Ganainy et al. [80], which showcases how HTM can be used to model medical streams in real time. Another example is the work done by Osegi [81], which shows how HTM can be used for forecasting electrical loads in power grids.

There are also applications that are actively used in production, such as the model offered by cortical.io which builds upon HTM in order to perform language analysis. This is made this possible by introducing Semantic Folding and Semantic Fingerprinting [82].

2.3.7 The Thousand Brains Theory

One of the newest advancements in HTM Theory is the introduction of the Thousand Brains Theory. Hawkins et al. [6] introduces the Thousand brains theory as a way of redefining hierarchy in the brain based on recent neuroscientifical discoveries. Instead of our classical understanding of hierarchy in deep learning where each layer takes simple features and outputs complex features, we now have that every layer of the hierarchy sees the input at once but at different scales and resolutions. The different nodes in the hierarchy are now also connected and thus enable the network to use all available views of the object in order to create an understanding of that object.

To summarize, the object is learned by the brain using multiple models that may rely on different inputs, the models then vote to reach a consensus on what they are sensing. This is coincidentally similar to ensemble learning such as Thambawita et al. [83]. Each model can be thought of as a mini-brain, hence the name "The Thousand Brains Theory".

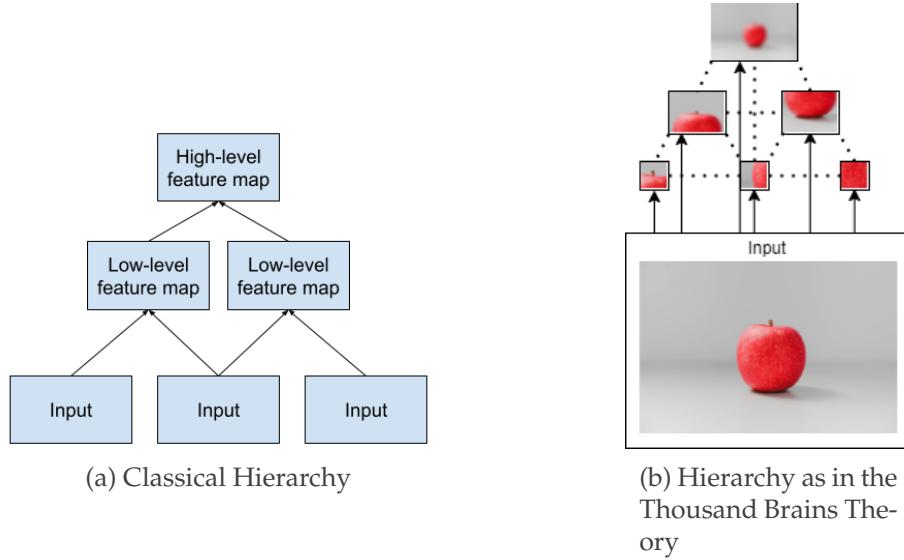


Figure 2.24: Comparison of classical hierarchy and the hierarchy introduced by the Thousand Brains Theory.

This new type of hierarchy is also quite similar to some state-of-the-art image recognition deep learning architectures such as InceptionNet [84] (see Figure 2.25) and Feature Pyramid Networks [85], in the sense that they apply different sized convolutional filters, where each filter can be thought of as its own separate model, on the data and do predictions based on all of them at once. This also ensures scale invariance of objects fed in to the architecture.

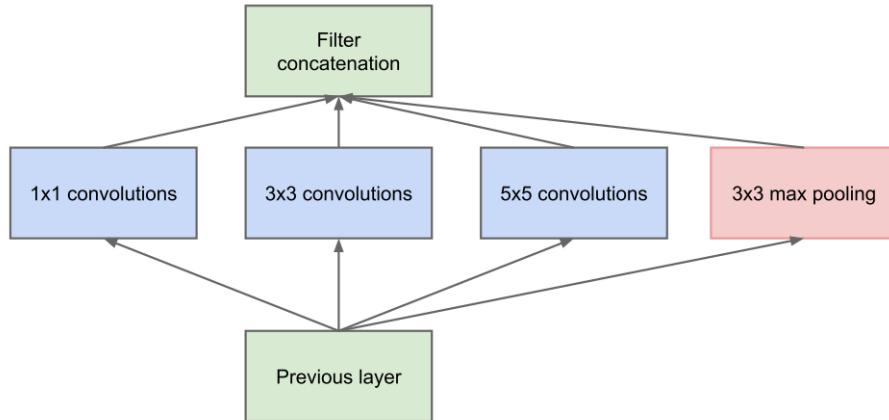


Figure 2.25: How the Inception [84] architecture uses multiple filters at once.

While the Thousand Brains Theory is not yet technically implemented in any way in a standard HTM model, it does show that recent developments

within deep learning for image analysis have similarities with HTM theory.

2.3.8 HTM Performance in Anomaly Detection

Knowing that deep learning approaches have a high degree of accuracy but suffer from problems related to generalizability, adaptability, and noise it stands to reason that HTM is a viable alternative for anomaly detection.

Ahmad et al. [74] explores the use of HTM for anomaly detection on low dimensional data such as temperature data from an industrial machine. The authors also discuss benchmarks for anomaly detection and compare different methods. The results show that HTM is very capable of performing anomaly detection, especially in a changing environment. HTM is able to outperform other anomaly detection methods and has the advantage of not requiring any per-problem parameter tuning.

For high-dimensional anomaly detection; Daylidyonok, Frolenkova, and Panov [3] used a HTM system to find anomalous frames in videos of motions (see Figure 2.26). The anomalies were artificially created by swapping certain frames between different motion videos in the dataset. The results show that the HTM system was able to correctly detect some anomalies, but not an impressive amount.

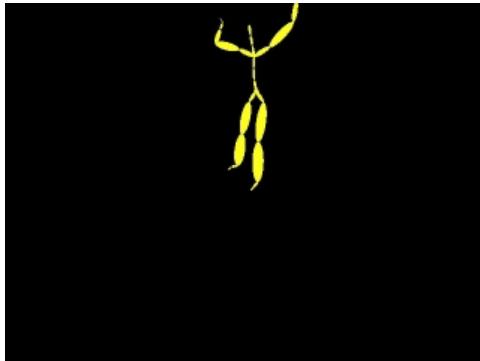


Figure 2.26: Example motion frame the dataset used by Daylidyonok, Frolenkova, and Panov [3].

One thing to note is that direct binary representations of the video frames were used as SDRs, therefore no proper encoding was performed which might have led to the poor results. This hints at the fact that HTM by itself is not capable of handling high dimensional data, and is instead reliant on an encoder to lower the dimensionality by extracting important spatial features.

2.4 Ethical Considerations

This thesis introduces an architecture for performing surveillance on potentially a massive scale, and could even be tied to some sort of social

credit system. This raises ethical questions regarding mass surveillance that highlight how the work done in this thesis can be misused. Therefore, even if the work done in this thesis is nowhere near perfect and far from ready to be used in production, one should still keep the ethical aspects in mind when pursuing this type of research.

It should also be reiterated that HTM is the result of research from a private company, which is why one should be extra critical of their work. Another aspect is that unlike deep learning, which is democratized, HTM is still very centralized. This means that Numenta has a lot of influence within the development of HTM, which could be misused for the purpose of promoting the goals of the company.

2.5 Summary

Deep learning can be traced back to the Perceptron which was a machine learning model capable of solving linearly separable problems. Over time, people theorized that stacking Perceptrons would give it the ability to solve more complex problems. This was not possible before back propagation was introduced. After that, deep learning evolved and new architectures were invented. One of the main contribution is the introduction of the CNN, which allowed deep learning approaches to achieve state-of-the-art results for complex problems such as image classification. The ability to train deep learning on the GPU and the introduction of frameworks caused deep learning to explode in popularity. One of the latest advancements is the introduction of generative models, which can be used for various purposes including generating realistic data. Deep learning has several issues, such as lack of explainability, the need for a lot of data, and poor out-of-distribution performance.

Anomaly detection is often defined as detecting data points that deviate from the general distribution. Unlike most other problems in deep learning, anomaly detection deals with unpredictable and rare events which makes it hard to apply traditional deep learning for anomaly detection. Popular approaches therefore often employ generative models, that calculate an anomaly measure using generated or reconstructed data. A subset of anomaly detection is smart surveillance, which is the use of video analysis specifically in surveillance. Recent developments show that generative models have a high degree of accuracy and are consistently improving.

HTM theory introduces a machine learning algorithm which works on the same principles as the brain and therefore solves some of the issues that deep learning has. HTM is considered noise resistant and can perform online learning, meaning that it learns as it observes more data. HTM replicates the structure of the neocortex which is made up of cortical regions, which in turn are made of up of mini-columns and then neurons. HTM neurons are more complex than neurons in deep learning, and perform learning with Hebbian-like learning. The data in an HTM is

represented using an SDR, which is a sparse bit array. An encoder converts real world values into SDRs. One of the difficulties with HTM is making it work on visual data, where creating a good encoder for visual data is still being researched. The learning mechanism consists of two parts, the spatial pooler and the temporal memory. The spatial pooler learns to extract semantically important information into output SDRs. The temporal memory learns sequences of patterns of SDRs and forms a prediction in the form of a predictive SDR. Research has shown that HTM is very capable of performing anomaly detection on low-dimensional data and is able to outperform other anomaly detection methods. However, related works show that HTM struggles with higher dimensional data. Therefore, a natural conclusion to make is that HTM should be applied differently, and that a new type of architecture using HTM should be explored for the purpose of video anomaly detection.

Chapter 3

Grid HTM

This chapter presents Grid HTM, which is the product of answering the thesis question and is the main contribution of this thesis. The source code of Grid HTM can be found on Github[11].

3.1 Introduction

This thesis explores a new type of architecture, named Grid HTM, for anomaly detection in videos using HTM, and proposes to use segmentation techniques to simplify the data into an SDR-friendly format. These segmentation techniques could be anything from simple binary thresholding to deep learning instance segmentation. Even keypoint detectors such as ORB [86] could in theory be applied. When explaining Grid HTM, the examples will be taken from deep learning instance segmentation of cars on a video from the VIRAT [87] dataset. An example segmentation is shown in Figure 3.1.



Figure 3.1: Segmentation result of cars, which is suited to be used as an SDR. Original frame taken from VIRAT [87].

The idea is that the SP will learn to find an optimal general representation of cars. How general this representation is can be configured using the various parameters, but ideally they should be set so that different cars will be represented similarly while trucks and motorcycles will be represented differently. An example representation by the SP is shown in Figure 3.2.

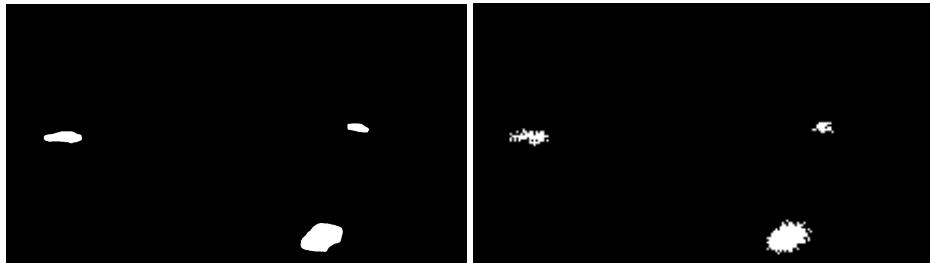


Figure 3.2: The SDR (left) and its corresponding SP representation (right). Note that the SP is untrained.

The task of the TM will then be to learn the common patterns that the cars exhibit, their speed, shape, and positioning will be taken into account. Finally, the learning will be set so that new patterns are learned quickly, but forgotten slowly. This will allow the model to quickly learn the norm, even if there is little activity, while still reacting to anomalies. This requires that the input is stationary, in our example this means that the camera is not moving.

It is possible to split different segmentation classes into their respective SDRs. This will give the SP and the TM the ability to learn different things for each of the classes. For instance, if there are two classes "person" and "car", then the TM will learn that it is normal for objects belonging to "person" to be on the sidewalk, while objects belonging to "car" will be marked as anomalous when on the sidewalk.

Ideally, the architecture will have a calibration period spanning several days or weeks, during which the architecture is not performing any anomaly detection, but is just learning the patterns.

3.2 Improvements

As it currently stands, the current architecture is nearly identical to the one used by Daylidyonok, Frolenkova, and Panov [3], which was shown to not be particularly effective, therefore multiple improvements are introduced to increase effectiveness.

3.2.1 Invariance

One issue that becomes evident is the lack of invariance, due to the TM learning the global patterns. Using the example, it learns that it is normal for cars to drive along the road but only in the context of there being cars parked in the parking lot. It is instead desired that the TM learns that it is normal for cars to drive along the road, regardless of whether there are cars in the parking lot.

This thesis proposes a solution based on dividing the encoder output into a grid (see Figure 3.3), and have a separate SP and TM for each cell in the

grid. The anomaly scores of all the cells are then aggregated into a single anomaly score using an aggregation function.

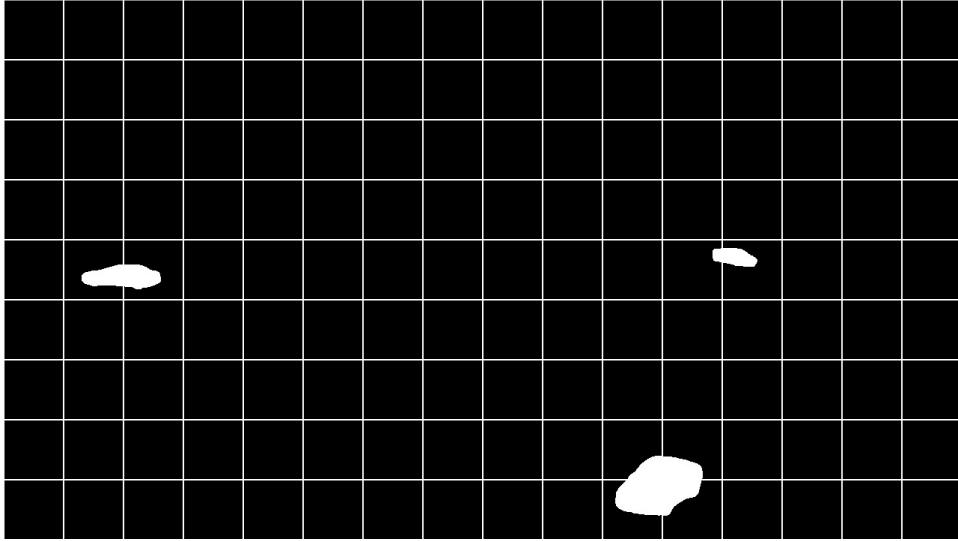


Figure 3.3: The encoder output divided into a grid.

3.2.1.1 Aggregation Function

Selecting the correct aggregation function is important because it affects the final anomaly output. For instance, it might be tempting to use the mean of all the anomaly scores as the aggregation function:

$$X : \{x \in \mathbb{R} : x \geq 0\}$$

$$\text{anomScore} = \frac{\sum_{x \in X} x}{|X|}$$

However, this leads to problems with normalization, meaning that an overall anomaly score of 1 is hard to achieve due to many cells having a zero anomaly score. In fact, it becomes unclear what a high anomaly score is anymore. Using the mean also means that anomalies that take up a lot of space will be weighted more than anomalies that take up a little space, which might not be desirable.

To solve the aforementioned problem and if the data has little noise, a potential aggregation function could be the non-zero mean:

$$X : \{x \in \mathbb{R} : x > 0\}$$

$$\text{anomScore} = \begin{cases} \frac{\sum_{x \in X} x}{|X|} & \text{if } |X| > 0 \\ 0 & \text{otherwise} \end{cases}$$

Meaning that only the cells with a non-zero anomaly score, denoted X , will be contributing to the overall anomaly score which helps solve the aforementioned normalization and weighting problem. On the other hand, this will perform poorly when the architecture is exposed to noisy data which could lead to there always being one or more cells with a high anomaly score.

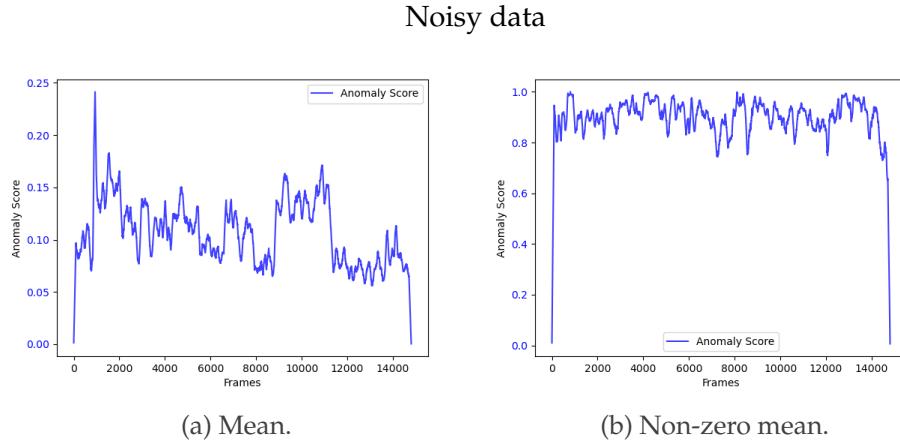


Figure 3.4: How the two aggregation functions perform on the same noisy data.

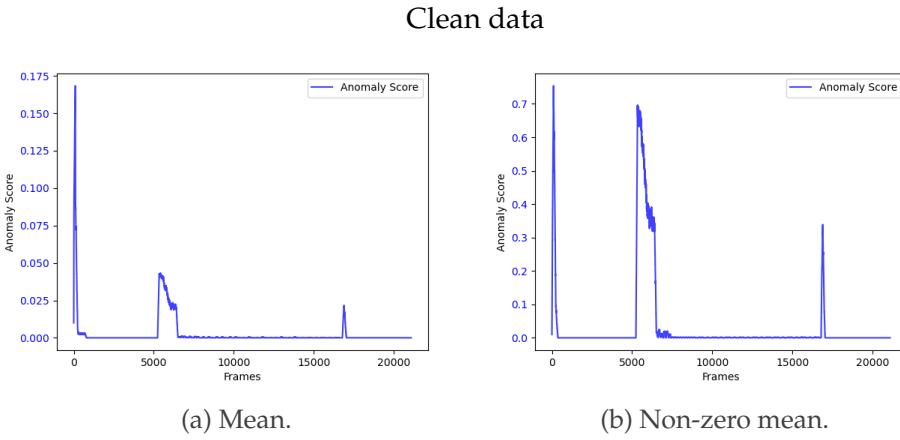


Figure 3.5: How the two aggregation functions perform on the same clean data.

Figure 3.4 illustrates the effect of an aggregation function for noisy data, where the non-zero mean is rendered useless due to the noise. On the other hand, Figure 3.5 shows how the non-zero mean gives a clearer anomaly score when the data is clean. Especially regarding how, unlike the mean, the non-zero mean has a clearly defined range between 0 and 1.

3.2.2 Explainability

Having the encoder output divided into a grid has the added benefit of introducing explainability into the model. By using Grid HTM it is now possible to find out where in the input an anomaly has occurred by simply observing which cell has a high anomaly score.

It is also possible to estimate the number of predictions for each cell which can be used as a measure of certainty, where fewer predictions means higher certainty. Making it possible to measure certainty per cell creates a new source of information which can be used for explainability or robustness purposes.

These two properties show that GridHTM offers vastly more explainability than current state-of-the-art deep learning models for anomaly detection, which could make it an attractive approach.

3.2.3 Flexibility and Performance

In addition, it is also possible to configure the SP and the TM in each cell independently, giving the architecture increased flexibility. It is also possible to use a non-uniform grid, meaning that some cells can have different sizes. Last but not least, dividing the frame into smaller cells makes it possible to run each cell in parallel for increased performance.

3.2.4 Reviewing Encoder Rules

That being said, a potential problem with the grid approach is that the rules for creating a good encoder, mentioned in Section 2.3.4, may not be respected and therefore should be reviewed:

- **Semantically similar data should result in SDRs with overlapping active bits.** In this example, a car at one position will produce an SDR with a high amount of overlapping bits as another car at a similar position in the input image.
- **The same input should always produce the same SDR.** The segmentation model produces a deterministic output given the same input.
- **The output must have the same dimensionality (total number of bits) for all inputs.** The segmentation model output has a fixed dimensionality.
- **The output should have similar sparsity (similar amount of one-bits) for all inputs and have enough one-bits to handle noise and subsampling.** The segmentation model does not respect this. An example is that there can be no cars (zero active bits), one car (n active bits), or two cars ($2n$ active bits), and that this will fluctuate over time.

The solution for the last rule is two-fold, and consists of imposing a soft upper bound and a hard lower bound for the number of active pixels

within a cell. The purpose is to lower the variation of number of active pixels, while also containing enough semantic information for the HTM to work:

- Pick a cell size so that the distribution of number of active pixels is as tight as possible, while containing enough semantic information and also being small enough so that the desired invariance is achieved. The cell size acts as a soft upper bound for the possible number of active pixels.
- Create a pattern representing emptiness, where the number of active bits is similar to what can be expected on average when there are cars inside a cell. This acts as a hard lower bound for the number of active pixels.

There could be situations where a few pixels are active within a cell, which could happen when a car has just entered a cell, but this is fine as long as it does not affect the distribution too much. If it does affect the distribution, which can be the case with noisy data, then an improvement would be to add a minimum sparsity requirement before a cell is considered not empty, e.g. less than 5 active pixels means that the cell is empty. In the following example, the number of active pixels within a cell centered in the video was used to build the distributions seen in Figure 3.6:

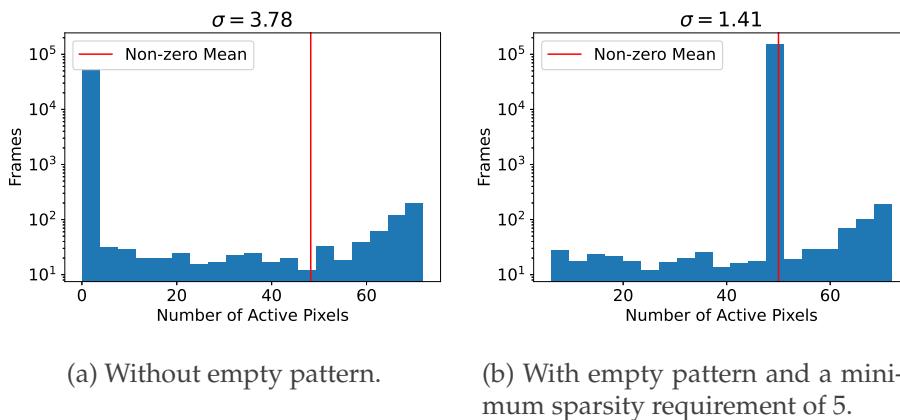


Figure 3.6: Distribution of number of active pixels within a cell of size 12×12 .

With a carefully selected empty pattern sparsity, **the standard deviation of active pixels was lowered from 3.78 to 1.41**. It is possible to automate this process by developing an algorithm which finds the optimal cell size and empty pattern sparsity which causes the least variation of number of active pixels per cell. This algorithm would run as a part of the calibration process.

The visual output resulting from these changes, which is an equally important output as the aggregated anomaly score, can be seen in Figure 3.7.

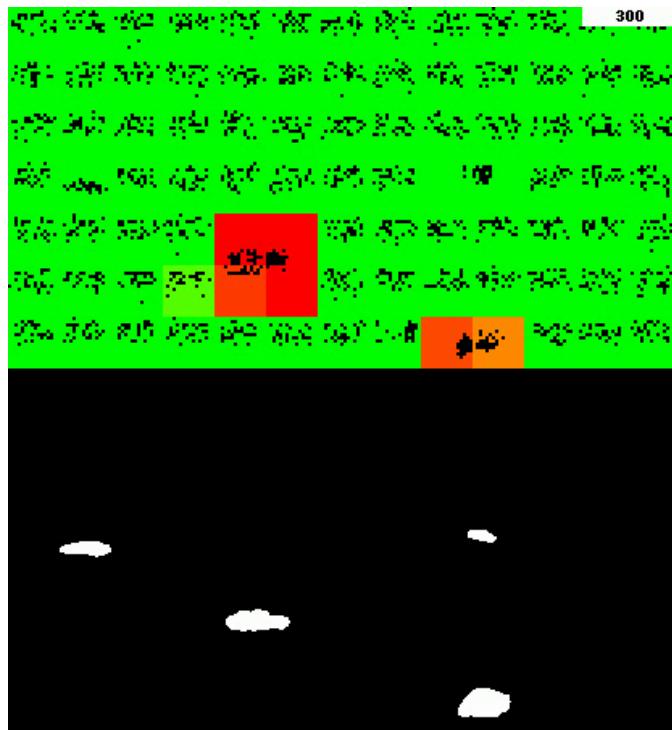


Figure 3.7: Example Grid HTM output and the corresponding input. The color represents the anomaly score for each of the cells, where red means high anomaly score and green means zero anomaly score. Two of the cars are marked as anomalous because they are moving, which is something the Grid HTM has not seen before during its 300 frame long lifetime.

Since there are now cells that are observing an empty pattern for a lot of the time in sparse data, boosting is recommended to be turned off, otherwise the SP output for the empty cells would change back and forth in order to adjust the active duty cycle.

3.2.5 Stabilizing Anomaly Output

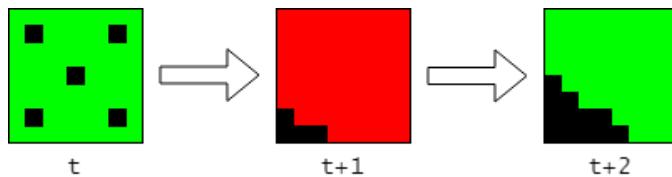


Figure 3.8: High anomaly score when an empty cell (represented with an empty pattern with a sparsity value of 5) changes to being not empty, as something enters the cell.

Another issue with the grid based approach is when a car first comes into a cell. The TM in that cell has no way of knowing that a car is about to enter, since it does not see outside its own cell, and therefore the first

frame that a car enters a cell will cause a high anomaly output. This is illustrated in Figure 3.8 where it can be observed that this effect causes the anomaly output to needlessly fluctuate. The band-aid solution is to ignore the anomaly score for the frame during which the cell goes from being empty to being not empty, which is illustrated in Figure 3.9.

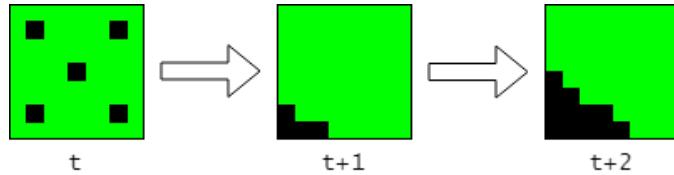


Figure 3.9: In this case, the anomaly score is ignored (set to 0) for the frame in which the cell changes state from empty to not empty.

A more proper solution could be to allow the TM to grow synapses to the TMs in the neighboring cells, but this is not documented in any research papers and might also hinder invariance.

3.2.6 Multistep Temporal Patterns

Since the TM can only grow segments to cells that were active in the previous timestep, as was mentioned in Section 2.3.5.2, it will struggle to learn temporal patterns across multiple timesteps. This is especially evident in high framerate videos, where an object in motion has a similar representation at timestep t and $t + 1$, as an object standing still. Figure 3.10 visualizes this phenomenon.

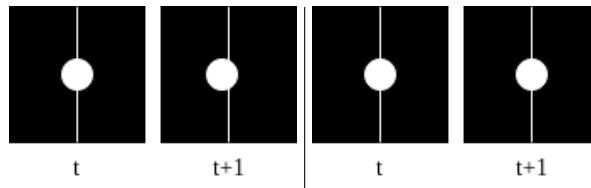


Figure 3.10: Comparison of a moving object (left) and a still object (right) in a high framerate video. They are very similar and could actually end up being represented identically by the SP.

This could cause situations where an object that is supposed to be moving, suddenly stands still, yet the TM will not mark it as an anomaly due to it being stuck in a contextual loop (see Figure 3.11). A contextual loop is when one of the predictions at t becomes true at $t + 1$, and then one of the predictions at $t + 1$ is almost identical to the state at t , which becomes true if the object is not moving, causing the TM to enter the same state that it was in at t .

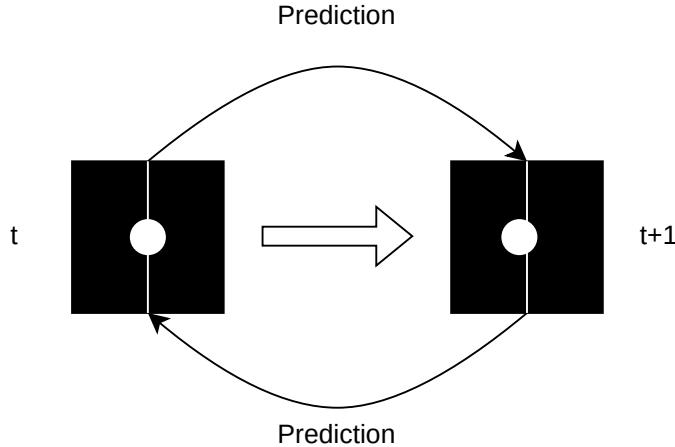


Figure 3.11: Example of a contextual loop.

A solution is to concatenate the past n SP outputs as input into the TM, which is made possible by keeping a buffer of past SP outputs and shifting its contents out as new SP outputs are inserted (see Figure 3.12). This follows the core idea behind encoding time in addition to the data, which makes time act as a contextual anchor. However, in this case there are no timestamps that are suitable to be used as contextual anchors, so as a replacement, the past observations are encoded instead. Parallels can be drawn to the Transformer model, where a positional encoding is required for it to learn order of sequences [29].

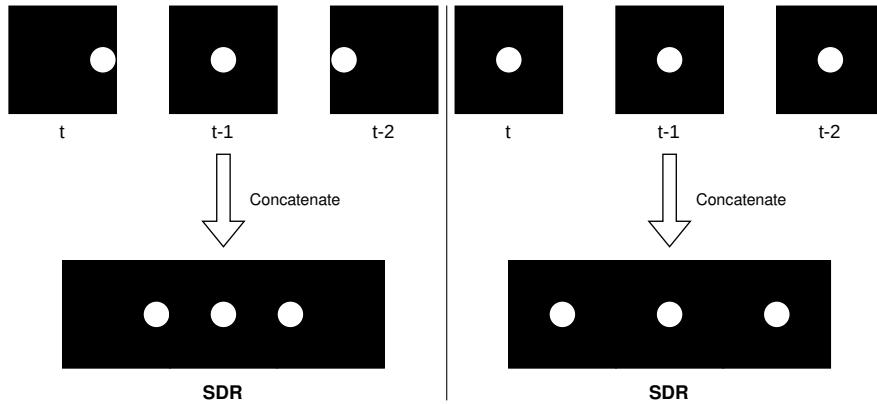


Figure 3.12: Example of concatenation with $n = 3$ when an object is moving from left to right, compared to when an object is not in motion. It can be observed that the SDRs are vastly different.

Concatenating past observations together will force the TM input, for when an object is in motion and when an object is still, to be unique. High framerate videos can benefit the most from this, and the effect will be more pronounced for higher values of n . One could feed the past n encoder outputs into the SP instead, but this will lead to a much higher computational requirement, as well as make both the SP and TM

parameters to be dependent on n .

A potential side effect of introducing temporal patterns, is that because the TM is now exposed to multiple frames at once, it will be more tolerant to temporal noise. An example of temporal noise is when an object disappears for a single frame due to falling below the classification threshold of the deep learning segmentation model encoder. The reason for the noise tolerance is that instead of the temporal noise making up the entire input for the TM, it now only makes up $\frac{1}{n}$ of the TM input.

Adding support for multistep temporal patterns to the method used by Daylidyonok, Frolenkova, and Panov [3], which is mentioned in Section 2.3.8, could improve their results.

3.3 Implementation

Grid HTM was implemented in Python using the community fork of Numenta’s HTM codebase [9]. The reason for using the community fork is that Numenta has stopped developing the original codebase, and that the community codebase is more optimized and has more features. There are other HTM implementations available [88, 89], but these were not considered due to their low popularity meaning that they have a higher chance of lacking features and containing bugs.

3.4 Biological Plausibility

One of the key principles of HTM theory is that it should be biologically plausible [63]. Every development within HTM theory follows this principle, and it should therefore be expected that Grid HTM does so as well.

Remembering that a single HTM model models a cortical column, as stated in Section 2.3.1, means that each cell in the grid can be considered an individual cortical column. The entire architecture can therefore be thought of as a simplified cortical region, as visualized in Figure 2.10.

3.5 Use Cases



Figure 3.13: Examples of environments that Grid HTM could help monitor. Images taken from the VIRAT [87] dataset.

The most intuitive use case is to use Grid HTM for semi-active surveillance, where personnel only have to look at segments containing anomalies, leading to drastically increased efficiency.

One example is making it possible to have an entire city, with environments such as in Figure 3.13, be monitored by a few people. This is made possible by making it so that people only have to look at segments that the Grid HTM has found anomalous, which is what drastically lowers the manpower requirement for active monitoring of the entire city.

Grid HTM could also be used to help automate labeling of anomaly datasets for deep learning. This would be similar to how older deep learning networks are used to help automate creating new labeled image datasets, where the model proposes a label for an image, which is then further refined by a human if needed.

3.6 Summary

This chapter proposes to use segmentation techniques as an encoder for visual data. The resulting SDRs can then be parsed by Grid HTM, which is essentially a collection of HTM models divided into a grid. The idea is that the SP will find an optimal general representation of the objects represented in the input SDRs. The task of the TM will be to learn the patterns that these objects exhibit over time, such as position, speed, and shape. Ideally, the architecture will have a calibration period where it will not do any anomaly detection but will only learn the patterns.

The main reason for applying a grid, and using separate HTM models in

each cell, is to introduce a form of invariance within the frame. Another benefit is that this introduces explainability into the architecture, meaning that it is easy to know where in the frame the anomaly is. Finally, having a grid means increased flexibility since the HTM models in each grid can be configured independently as well as the possibility to run each grid in parallel for increased performance. This chapter proposes to use an aggregation function to combine all the anomaly outputs of each cell in the grid into a single output for the entire architecture. This grid based approach also makes sense from a biological perspective by being a simplified cortical region.

With all the new changes made, it is important to review the rules for creating a good encoder, and it becomes evident that the rule regarding having a stable sparsity value is not followed. To address this, a pattern representing emptiness is introduced in order to impose a lower bound for sparsity, while the cell size itself acts as a soft upper bound for the number of active pixels. Due to the possibility of a lot of cells seeing the same empty pattern for long periods of time, it is recommended to disable boosting.

One of the issues with the introduction of a grid, is that the HTM models in each cell can not see outside their own cell. This means that the HTM models have no way of predicting that something is about to enter the cell, which causes a spike in the anomaly score for the frame when something enters the cell. A band-aid solution is to ignore the anomaly output during the frame in which this happens.

An issue that occurs specifically for high framerate videos is that HTM struggles to learn temporal patterns. It is suspected that this occurs due to a contextual loop, in which predictions that predict themselves become true. A solution is to concatenate the past SP outputs into a single SDR which is then used as input into the TM. This will force the TM input, for when an object is in motion and when an object is still, to be unique. A potential side effect is that the architecture becomes more tolerant to temporal noise.

Grid HTM has several potential use cases, such as semi-active surveillance in which an entire city can be monitored by a few persons. It can also be used to help label anomaly datasets for use in training deep learning anomaly detection models.

Chapter 4

Experiments and Results

In this section, various experiments are performed in order to gauge the effectiveness of Grid HTM. There are three experiments in total, and each experiment covers a different use case. For reproducibility purposes, each experiment has corresponding tables of the parameters used.

4.1 Bouncing Ball Experiment

To give credibility to the approach mentioned in Chapter 3, a simple experiment to test the capabilities of HTM and confirm that they apply on a video is introduced.

4.1.1 Data

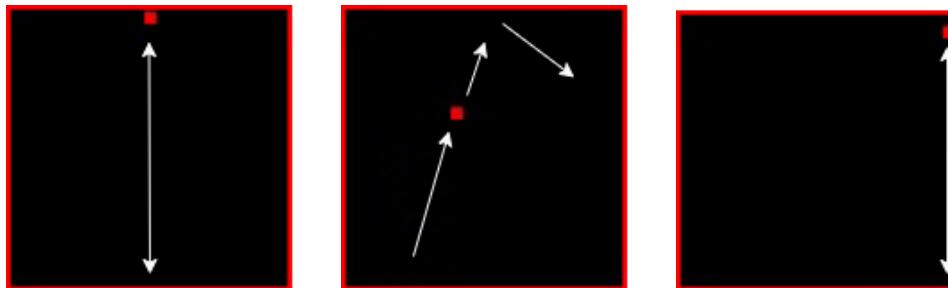


Figure 4.1: The bouncing ball experiment, and its three stages.

The video consists of a ball bouncing up and down until an anomaly occurs in the form of a sudden introduction of a horizontal velocity. After a while this horizontal velocity is set back to 0 and the ball is once again bouncing up and down in-place. This is visualized in Figure 4.1.

4.1.2 HTM

The model used is a standard HTM model, which covers the entire input. This is equivalent to a single cell in a Grid HTM.

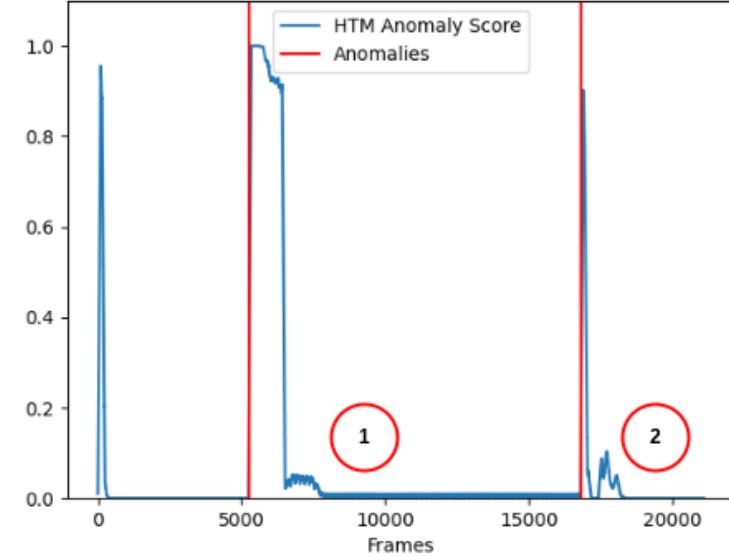


Figure 4.2: The anomaly score in the bouncing ball experiment.

From Figure 4.2 it can be observed that the HTM correctly detects anomalies and quickly adapts to them. On the other hand, the result is not perfect due to the minor oscillations close to (1) and the anomaly spikes towards the end close to (2). While the imperfections are not major and can be safely ignored, it is still important to understand their causes and what can be done to improve upon them.

4.1.2.1 Boosting

The reason for the oscillations is due to the spatial pooler being dominated by a lucky few columns. The solution is to enable boosting, as explained in Section 2.3.5.1. This also helped with the spikes towards the end, as can be seen in Figure 4.3.

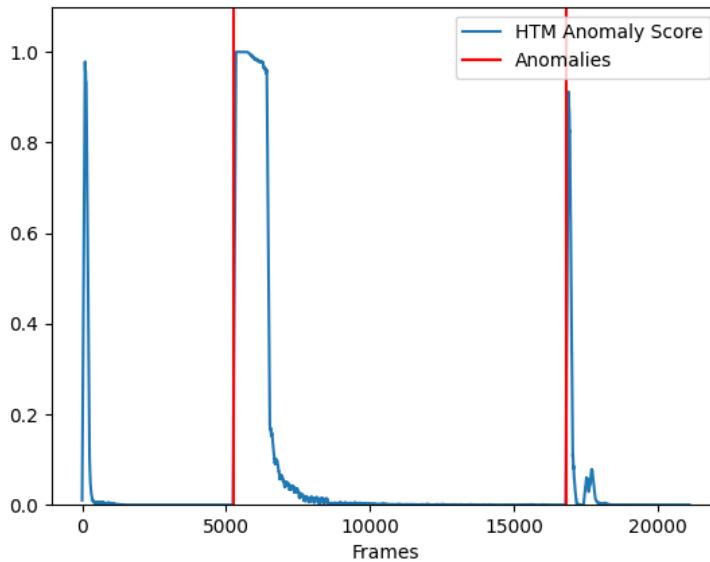


Figure 4.3: Bouncing ball with boosting enabled.

4.1.2.2 Zero Permanence Decrement

The reason for the anomaly spikes towards the end is because the spatial pooler has found an optimal representation when the ball is bouncing freely, but when the ball stops and starts bouncing in-place the spatial pooler ends up unlearning the old optimal representation while it learns the new optimal representation. This causes a sudden minor change in the SP output, which the TM reports as anomalous.

The solution is to set the value by which permanence is decreased by to zero, effectively disabling the ability of the spatial pooler to "forget", as can be seen in Figure 4.4. That being said, the ability to decrement permanence is important in HTM systems, therefore disabling it is not always feasible.

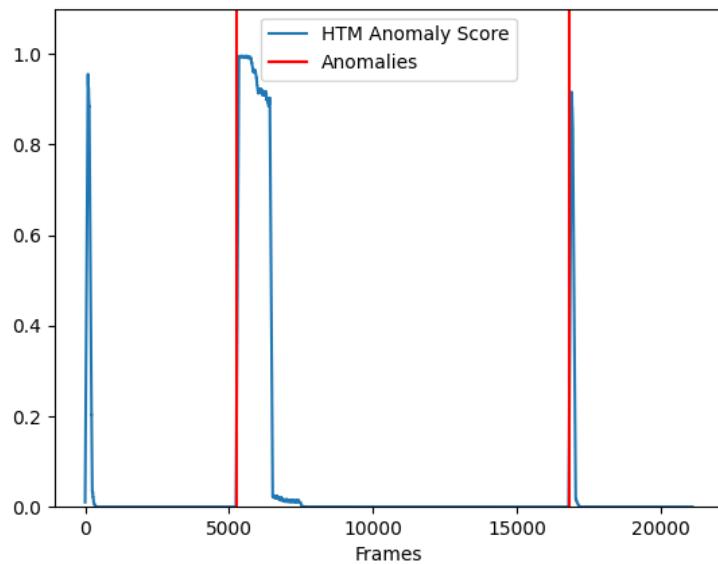


Figure 4.4: Bouncing ball without the ability of the SP to "forget".

4.1.2.3 Boosting and Zero Permanence Decrement

Finally, for the sake of interest, the bouncing ball example was performed with both boosting enabled and with zero permanence decrement. Results can be seen in Figure 4.5.

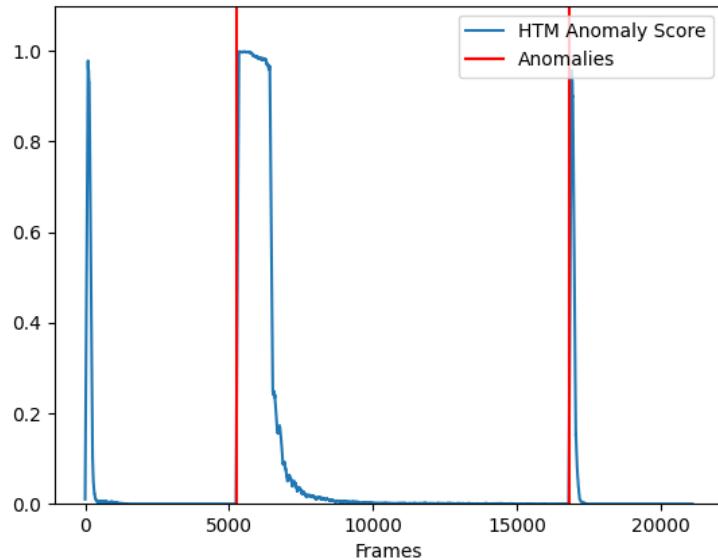


Figure 4.5: Bouncing ball without the ability of the SP to "forget" and with boosting enabled.

4.1.2.4 Parameters

Final list of parameters for reproducibility. For the plots, a moving average of $n = 100$ was used to smooth the output.

Parameter	Value	Notes
inputDimensions	120, 120	The shape of the entire frame
columnDimensions	60, 60	
potentialPct	0.1	
potentialRadius	120	
localAreaDensity	0.02	
globalInhibition	True	Set to False to enable topology
wrapAround	True	Allows the columns near the edges to "wrap around" and form connections on the other side
synPermActiveInc	0.1	
synPermInactiveDec	0	Set to >0 to enable the SP to "forget"
stimulusThreshold	2	
seed	2	
boostStrength	0.1	Set to 0 to disable boosting
dutyCyclePeriod	250	

Table 4.1: SP Parameters

Parameter	Value	Notes
columnDimensions	60, 60	Same as the SP
predictedSegmentDecrement	0.003	
permanenceIncrement	0.1	
permanenceDecrement	0.001	
minThreshold	3	
activationThreshold	5	
cellsPerColumn	16	
seed	2	

Table 4.2: TM Parameters

4.1.3 Grid HTM

This is a very simple problem which does not require invariances, making it unsuitable for Grid HTM. Grid HTM would be suitable if there were two or more independent bouncing balls, due to its improved invariance. Still, it is interesting to see how Grid HTM performs compared to normal HTM.

4.1.3.1 Results

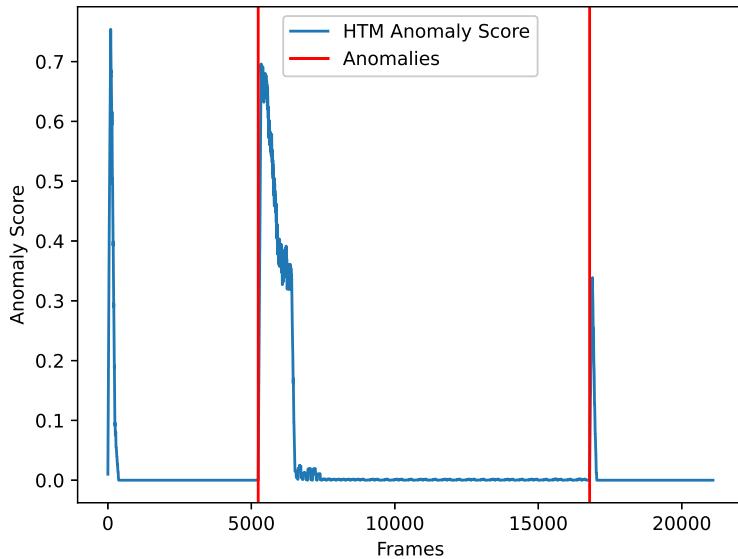


Figure 4.6: Results when using Grid HTM

It can be observed in Figure 4.6 that Grid HTM performs worse than the normal HTM, but the result is still acceptable. This is to be expected since this problem is not suited for the Grid HTM, and that the parameters given in Table 4.3, Table 4.4, and Table 4.5 are probably not optimal.

4.1.3.2 Parameters

The SP and TM parameters were selected so that they were as close as possible to the normal HTM parameters. The non-zero mean was chosen as the aggregation function, because there is no noise due to the controlled environment. Again, a moving average of $n = 100$ was used to smooth the anomaly score output in the plots.

Parameter	Value	Notes
sp_grid_size	30, 30	Size of each grid
tm_grid_size	15, 15	Size of each SP grid output.
min_sparsity	1	
sparsity	A	Area of the bouncing ball with radius=3
temporal_size	1	Size of the multistep temporal pattern, 1 means it is effectively disabled

Table 4.3: Grid HTM specific parameters

Parameter	Value	Notes
inputDimensions	sp_grid_size	
columnDimensions	tm_grid_size	
potentialPct	0.5	Increased in order to compensate for the smaller potential pool
potentialRadius	5	
localAreaDensity	0.1	
globalInhibition	True	Set to False to enable topology
wrapAround	False	
synPermActiveInc	0.1	
synPermInactiveDec	0.001	
stimulusThreshold	2	
boostStrength	0	Causes instability in empty cells

Table 4.4: SP Parameters

Parameter	Value	Notes
columnDimensions	tm_grid_size	Same as the SP
predictedSegmentDecrement	0.003	
permanenceIncrement	0.1	
permanenceDecrement	0.001	
minThreshold	1	
activationThreshold	1	
cellsPerColumn	16	

Table 4.5: TM Parameters

4.1.4 Experiment Summary

This experiment is a controlled experiment involving a computer generated ball bouncing up and down, and then introducing an anomaly in the form of a change in horizontal velocity. Finally, the horizontal velocity is set back to 0. First the experiment is performed with a normal HTM model.

The results show that it is capable of detecting the anomalies and quickly adapts to them. The same experiment is then performed with Grid HTM. The results show that Grid HTM is also capable of detecting the anomalies and quickly adapts to them, however it performs slightly worse than normal HTM. The reason is that Grid HTM is designed for more complex videos, such as if there were multiple bouncing balls at once.

4.2 Surveillance Experiment

As stated earlier, one of the use cases of Grid HTM is anomaly detection in surveillance. This example will show how Grid HTM could perform. The video to be used is part of the VIRAT [87] video dataset, and was selected due to its long duration and stationary camera, which is shown in Figure 4.7.

The downside is that the video does not contain any non-technical anomalies, but consists of technical anomalies in the form of several segments with sudden frame skips in between. There is also a synthetic anomaly introduced in the form of a frame repeat lasting a couple of seconds, essentially "freezing" time, in order to test whether Grid HTM is able to understand how objects should be moving in time.



Figure 4.7: Example frames from the selected video.

As previously mentioned, both binary thresholding and deep learning feature map extraction as encoders have their downsides. Therefore, this thesis proposes to use a combination of both, a segmentation model which can extract classes into their respective SDRs. Meaning that there could be an SDR for cars and an SDR for persons (see Figure 4.8), that are then concatenated before being fed into the system.

The segmentation model used is PointRend [90] with a ResNet101 [30] backbone, pretrained on ImageNet [72], and implemented using PixelLib [91].

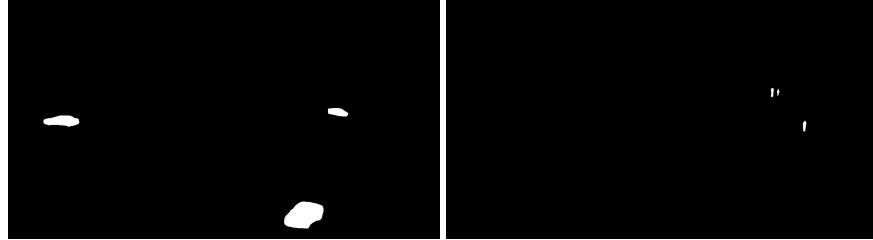


Figure 4.8: Example segmentation of cars and persons.

For the sake of simplicity, this experiment will focus only on the segmentation of cars.

While on the topic of segmentation, it is important to mention that the segmentation model is not perfect and that there are cases where objects are misclassified as well as cases where cars repeatedly go above and below the confidence threshold.

4.2.1 Results

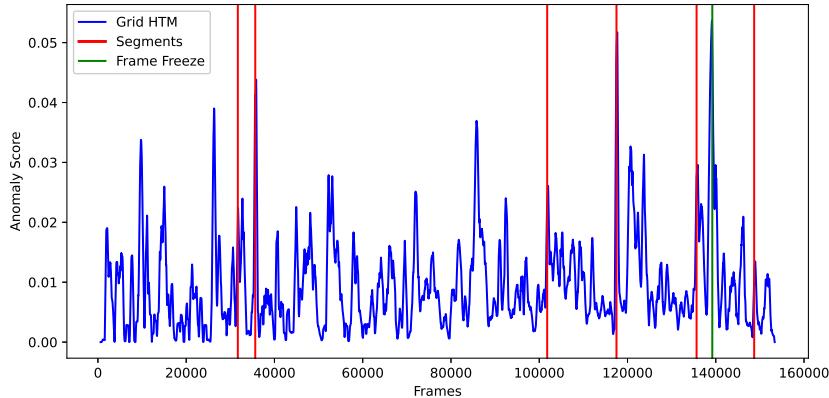


Figure 4.9: Anomaly score output from Grid HTM.

It can be observed in Figure 4.9 that the Grid HTM is detecting when segments begin and end, however it is not possible to use a threshold value to isolate them, and they also have vastly different anomaly scores compared to each other. This is due to the way the aggregation function works, which means that the anomaly output is dependent on the physical size of the anomaly. It should also be noted that a moving average ($n = 200$) was applied to smooth out the anomaly score output, otherwise the graph would be too noisy.

With the aggregation functions presented in this thesis in mind, it is safe to conclude that looking at the anomaly score output is meaningless for complex data such as a surveillance video. This however does not mean that Grid HTM is completely useless, and this can be observed by looking at the visual output of the Grid HTM. The visual output during which the first segment anomaly occurs can be seen in Figure 4.10.

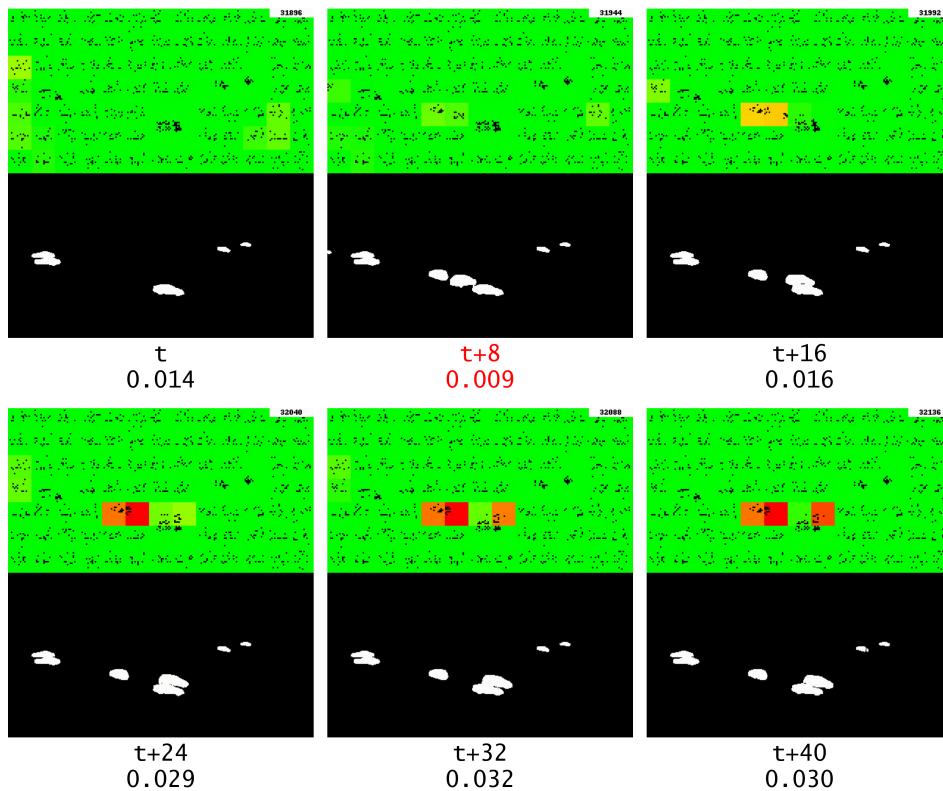


Figure 4.10: The first segment anomaly, which is marked with red text, and the corresponding changes detected by Grid HTM.

Here it is observed that Grid HTM correctly marks the sudden change of cars when the current segment ends and a new segment begins.

4.2.1.1 Road

In the original video, there is a road on which cars regularly drive. By observing the visual output, it becomes evident that after some time the Grid HTM has mostly learned that behavior and does not report those moving cars as anomalies. This is shown in Figure 4.11.

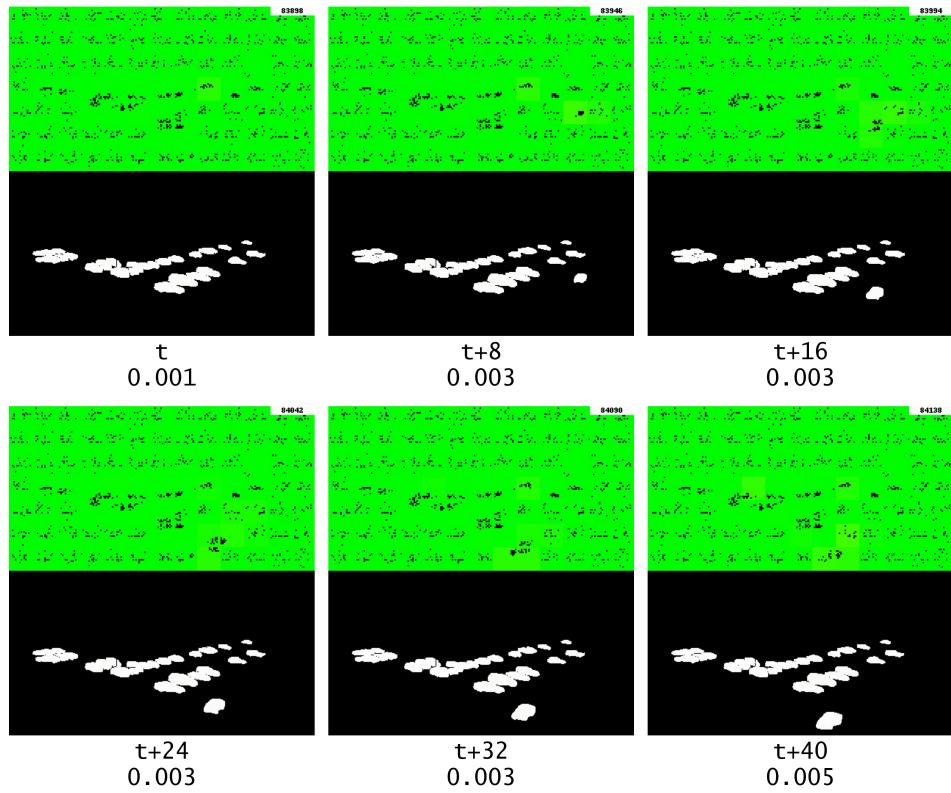


Figure 4.11: Visual output during when a car is driving along the main road.

4.2.1.2 Frame Repeat

To prove that the Grid HTM has learned that cars on the road should be moving, it is possible to look at the visual output during the period when the video is repeating the same frame and observe if the architecture marks the cars standing still on the road as anomalies.

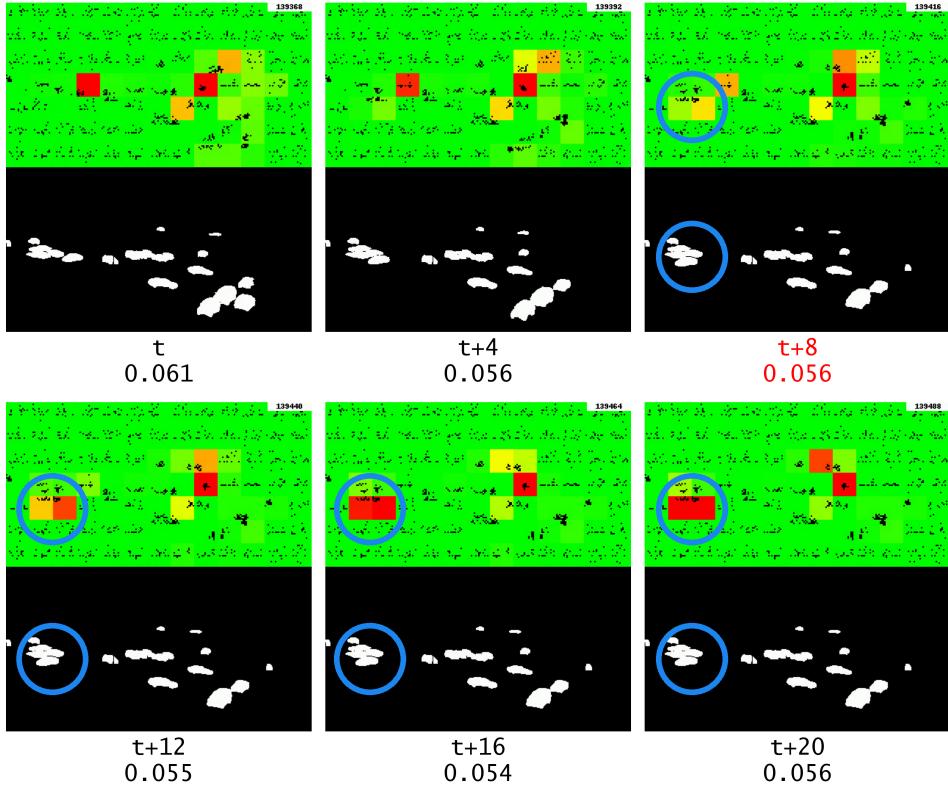


Figure 4.12: Anomaly output during the repeating frame, the start of the frame repeat is marked with red text. The blue circle highlights the object of interest.

It can be observed in Figure 4.12 that the cars along the main road are not marked as anomalies, but this could be attributed to the fact that there is a crossing there and that cars periodically have to stop at that point to let pedestrians cross.

On the other hand, when looking at the anomaly marked with a blue circle, the car on the road in the parking lot is marked as an anomaly that increases in severity as the time goes on during the frame repeat. The reason why that car causes an anomaly is because, unlike the cars on the main road, a car is rarely observed as standing still at that position.

To prove that the anomaly was actually directly caused by the repeating frame, and not just due to repeating the anomaly in time, it should be compared to the anomaly output if there was no repeating frame.

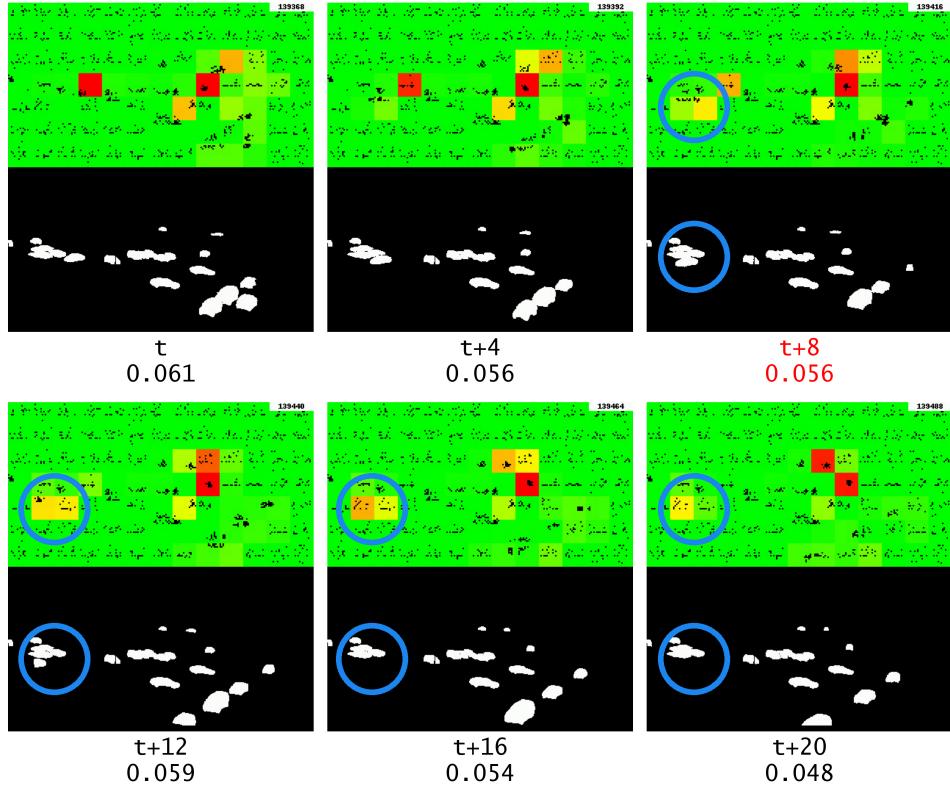


Figure 4.13: Anomaly output when there is no frame repeating, where it should have repeated is marked in red. The blue circle highlights the object of interest.

It can be observed in Figure 4.13 that the anomaly output is minor compared to when there was a repeating frame, proving that the anomaly was indeed a product of the repeating frame and that the Grid HTM was able to learn how objects should be moving in time.

Finally, it is interesting to look at how the Grid HTM handles the repeating frames without multistep temporal patterns, which is shown in Figure 4.14.

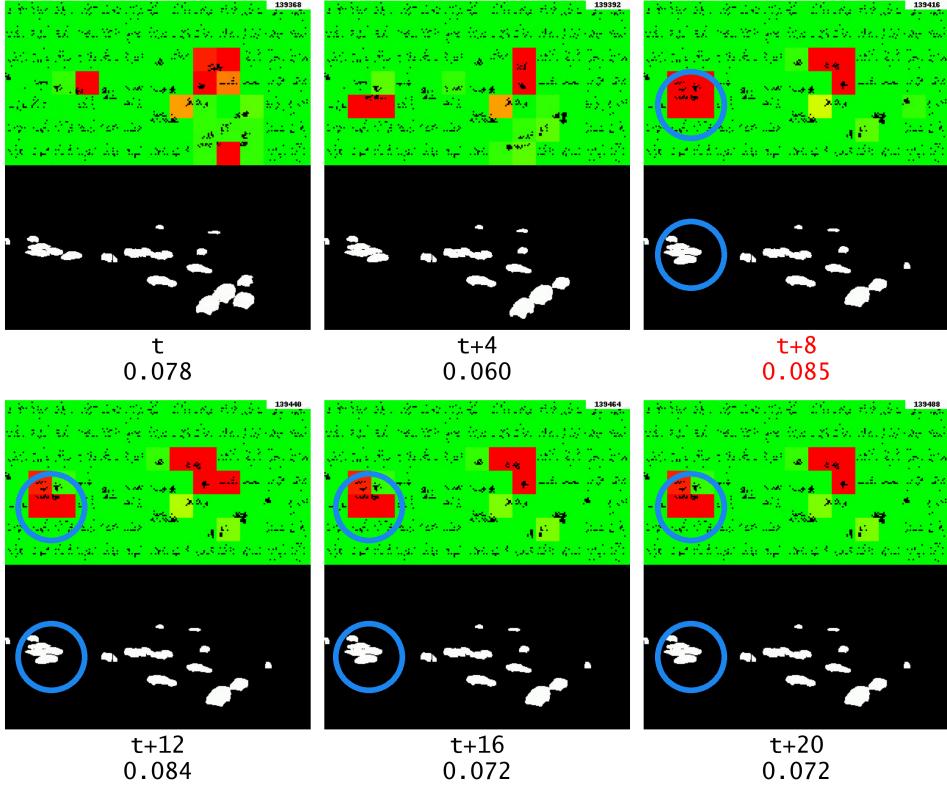


Figure 4.14: Anomaly output during the repeating frame, the start of the frame repeat is marked with red text. The blue circle highlights the object of interest. This time without multistep temporal patterns.

Unfortunately, simply disabling multistep temporal patterns without adjusting the other TM parameters causes the same car to be marked as an anomaly before and during the frame repeat. In fact, as mentioned in Section 3.2.6, disabling multistep temporal patterns causes Grid HTM to be less noise tolerant which causes a lot more anomalies to be wrongly detected. This is evident in Figure 4.14, where a higher amount of severe anomalies can be observed compared to previous examples. This also highlights how sensitive HTM can be regarding parameters.

4.2.1.3 Points of Interest

Finally, it is interesting to look the various anomaly score spikes and observe in the visual output what caused them. The points of interest to be explored are marked in Figure 4.15

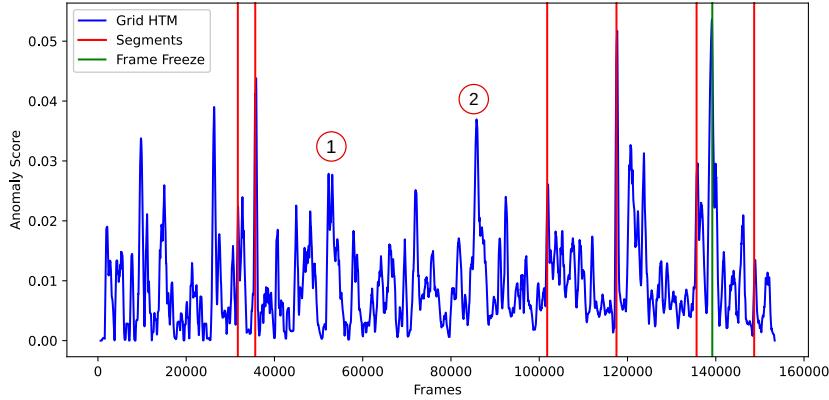


Figure 4.15: Points of interests in the anomaly score output.

The first point of interest, which can be seen in Figure 4.16, showcases the weakness of the aggregation function.

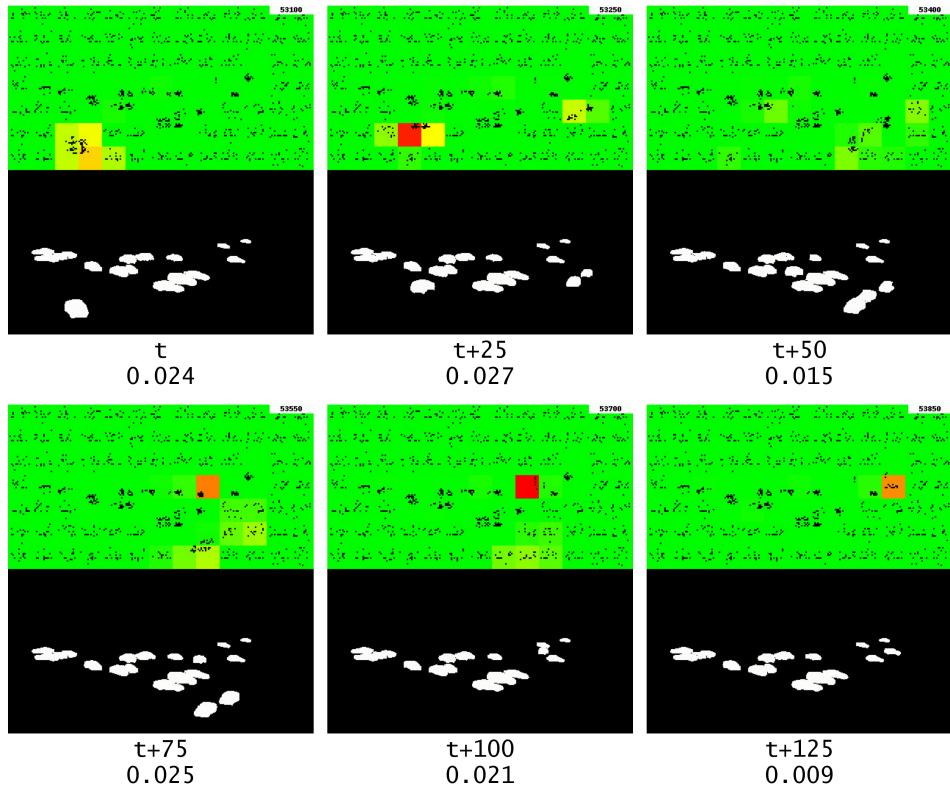


Figure 4.16: Visual output of the first point of interest.

It can be seen that the anomaly score output, which is shown at the very bottom of each image, can be attributed to two occurrences. The first occurrence is a physically big but low severity anomaly. The second

occurrence is a physically small but high severity anomaly. Despite the two different anomalies, the anomaly score output is similar for both occurrences.

The second point of interest, which can be seen in Figure 4.17, showcases the importance of exposing HTM to all possible behaviours that are considered not anomalous, which is one of the complexities mentioned in Section 2.2.

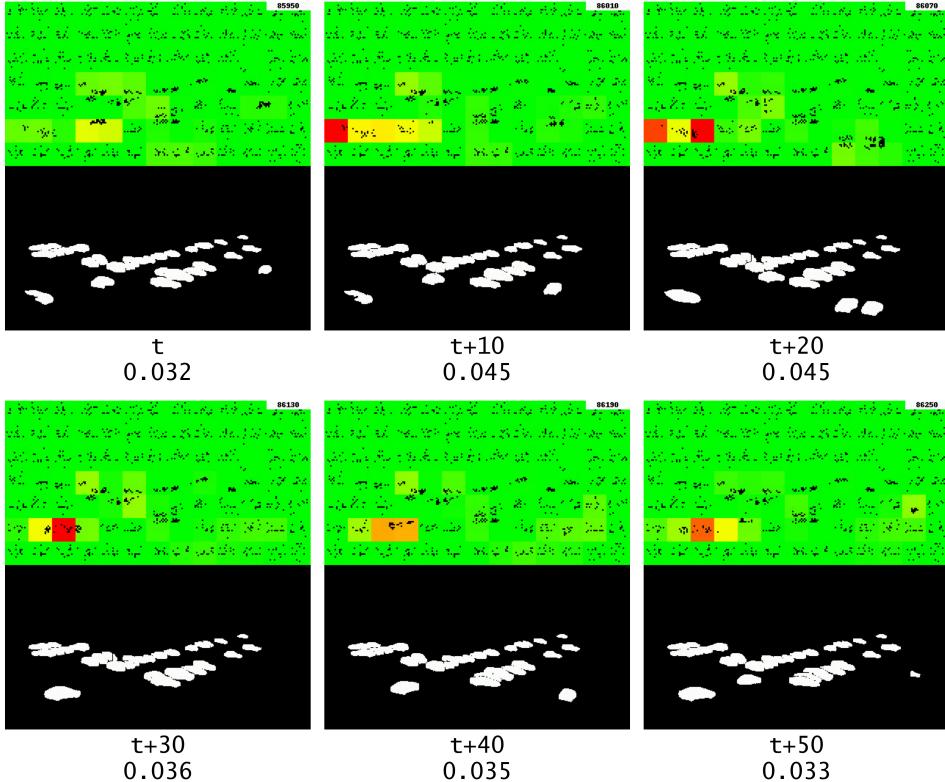


Figure 4.17: Visual output of the second point of interest.

It can be observed that there are high anomaly outputs for cars entering the parking lot from the left side of the frame. The high anomaly score is caused by an insufficient amount of previous observations of that behavior.

4.2.2 Parameters

Final list of parameters for reproducibility. As previously mentioned, a moving average ($n = 200$) was used to smooth the output graph and make it more readable. The mean was used as the aggregation function due to the complex nature of the data.

Parameter	Value	Notes
sp_grid_size	32, 32	Size of each grid
tm_grid_size	16, 16	Size of each SP grid output.
min_sparsity	10	
sparsity	15	Expected average sparsity
temporal_size	15	Size of the multistep temporal pattern buffer

Table 4.6: Grid HTM specific parameters

Parameter	Value	Notes
inputDimensions	sp_grid_size	
columnDimensions	tm_grid_size	
potentialPct	0.2	
potentialRadius	5	
localAreaDensity	0.05	
globalInhibition	True	Set to False to enable topology
wrapAround	False	
synPermActiveInc	0.01	
synPermInactiveDec	0.00001	
stimulusThreshold	3	
boostStrength	0	Causes instability in empty cells

Table 4.7: SP Parameters

Parameter	Value	Notes
columnDimensions	tm_grid_size	Same as the SP
predictedSegmentDecrement	0.001	
permanenceIncrement	0.01	
permanenceDecrement	0.001	
minThreshold	10	
activationThreshold	10	
cellsPerColumn	32	

Table 4.8: TM Parameters

4.2.3 Experiment Summary

This experiment showcases the performance of Grid HTM on complex data, specifically a surveillance video of a parking lot. The video contains technical anomalies in the form of segments, and also a synthetic anomaly which is a period of repeating frames. Semantic segmentation was performed in order to extract the cars in the frame into an SDR.

The results show that Grid HTM has the ability to react when segments begin and end, as well as the ability to detect the repeating frames. It also shows that the anomaly output, with the currently introduced aggregation functions, cannot be used to reliably threshold anomalies. Instead, one can look at the visual output of the Grid HTM. The visual output shows that Grid HTM is able to detect the change in objects when a segment change occurs.

Results also show that Grid HTM learns common patterns such as cars driving on the main road, and does not report that behavior as anomalous. It is also shown that Grid HTM can correctly detect the repeating frames, and marks anomalous cars during the repeating frames with an increasing severity. Finally, a couple of points of interests are shown that highlight the weakness of the aggregation function and the importance of exposing HTM to all possible behaviors not considered anomalous.

4.3 Sperm Experiment

As seen in the surveillance experiment, it seems the Grid HTM can detect when segments begin and end. This experiment will explore this ability in greater detail.

4.3.1 Data

The dataset used is VISEM [92], a sperm dataset which consists of videos that are made up of several segments. The sperm cells will be segmented using a rough binary thresholding, as shown in Figure 4.18.

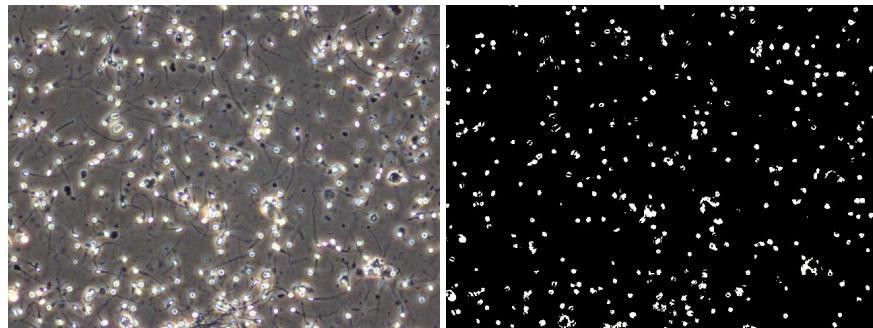


Figure 4.18: Example frame from a sperm video (left) and its corresponding segmentation (right).

It is important to note that the data itself is noisy, and that it is not possible for the Grid HTM to learn any meaningful pattern. The individual videos are also relatively short, which makes it even harder to learn any meaningful patterns.

4.3.2 Benchmark

To ensure that the HTM does not just react to the sudden change in pixels but does something more, the L1 error will be used as a benchmark to compare against:

$$E_t = \sum |F_t - F_{t-1}|$$

Where F_t denotes a segmented frame at time step t . The L2 error could also have been used, but it would not matter since this experiment will be comparing relative values.

4.3.3 Results

As seen in Figure 4.19, the Grid HTM is able to outperform the L1 error benchmark. This can be deduced from the more prominent changes in the anomaly score, compared to the L1 error. The reason might be that even though the data is very noisy, there is still something in it which makes Grid HTM able to learn something general about the current state. This could for

instance be a single cell that is standing still or moving very slowly, which the Grid HTM anchors itself to and uses it to determine when segments start and end.

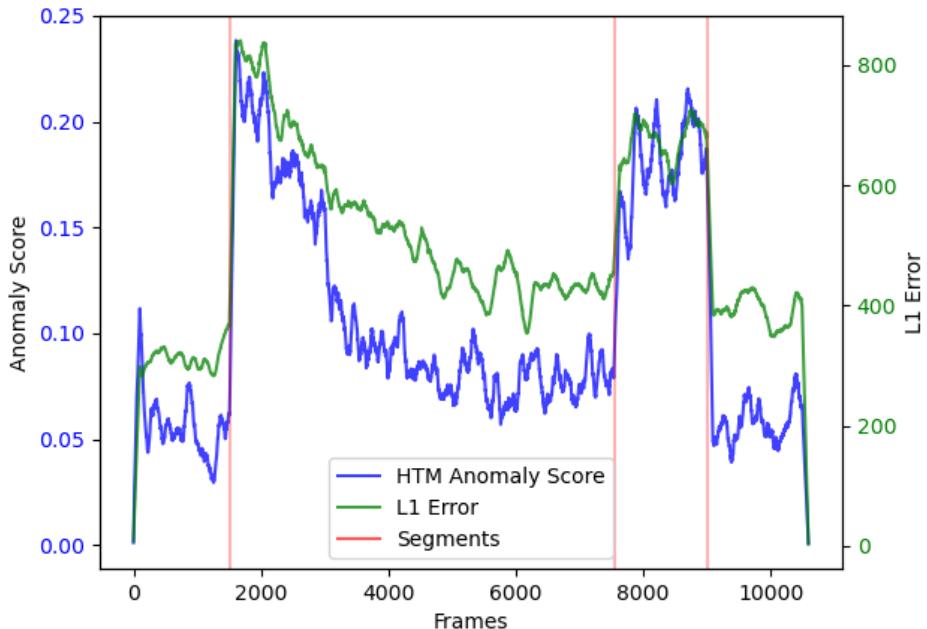


Figure 4.19: Results on a stationary sperm video.

That being said, the parameters for the Grid HTM were selected carefully to achieve the results seen in Figure 4.19, and are dependent on the contents of the data. Unfortunately, most of the videos in the dataset contain drift (in other words, the video is not stationary), which makes Grid HTM useless. This can be observed in Figure 4.20, where both the Grid HTM and the L1 error struggle.

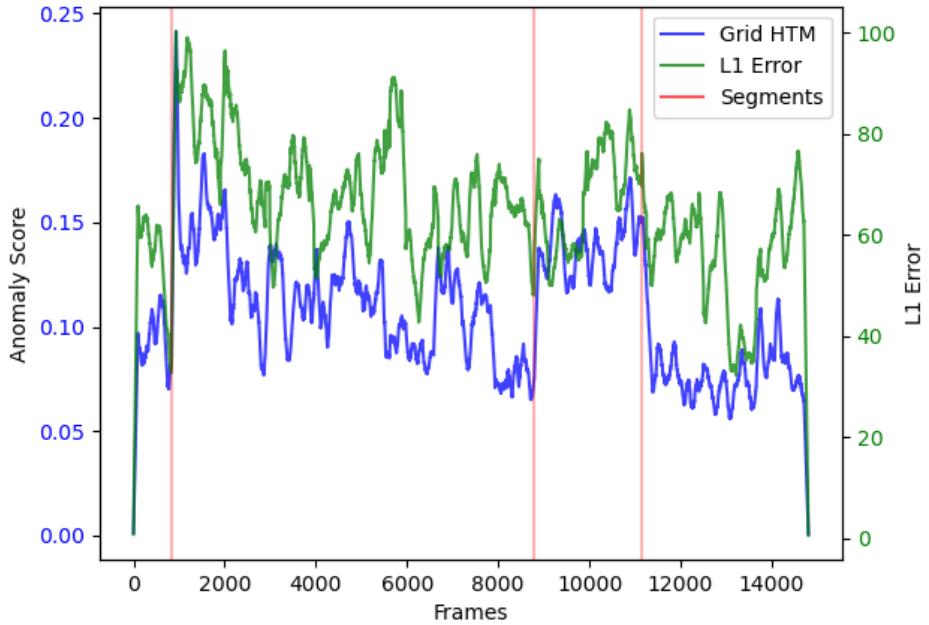


Figure 4.20: Results on a sperm video with drift.

Just like in the stationary video, the change in the anomaly score is more pronounced, but due to the drift in the video there is a constant high anomaly output which makes it impossible to find a threshold value. That being said, Grid HTM still outperforms the L1 score due to the more prominent changes in the anomaly score.

4.3.4 Use Cases

The use case is to be able to use Grid HTM to detect segments and then use a separate tool to extract each segment. This could be useful in data processing, or even in streaming media services to automatically know when cuts happen and could aid scene boundary detection systems [93].

4.3.5 Parameters

The following parameters were used in this experiment. The main difference from the surveillance experiment to note is the temporal size, which has been disabled due to the nonexistent patterns in the data. For the plots, a moving average with window size $n = 100$ was used to smooth the lines and reduce noise in the output graph. Because the data is noisy, the mean was used as the aggregation function.

Parameter	Value	Notes
sp_grid_size	16, 16	Size of each grid
tm_grid_size	8, 8	Size of each SP grid output.
min_sparsity	1	
sparsity	5	
temporal_size	1	Size of the multistep temporal pattern buffer, 1 means it is effectively disabled

Table 4.9: Grid HTM specific parameters

Parameter	Value	Notes
inputDimensions	sp_grid_size	
columnDimensions	tm_grid_size	
potentialPct	0.2	
potentialRadius	5	
localAreaDensity	0.1	
globalInhibition	False	
wrapAround	False	
synPermActiveInc	0.01	
synPermInactiveDec	0.001	
stimulusThreshold	5	
boostStrength	0	Causes instability in empty cells

Table 4.10: SP Parameters

Parameter	Value	Notes
columnDimensions	tm_grid_size	Same as the SP
predictedSegmentDecrement	0.003	
permanenceIncrement	0.01	
permanenceDecrement	0.001	
minThreshold	1	
activationThreshold	3	
cellsPerColumn	16	

Table 4.11: TM Parameters

4.3.6 Experiment Summary

This experiment explored the ability of Grid HTM to detect segments in greater detail. The videos used were videos of swimming sperm cells, and were segmented using a rough binary thresholding. The data is noisy and short, and it is therefore not expected for Grid HTM to learn any meaningful patterns. As a benchmark, the L1 error is employed.

The results show that Grid HTM outperforms the L1 error, presumably due to it managing to find something among the noise to lock on to. For this experiment, the parameters had to be carefully tuned to the content of the data.

4.4 Summary

In this chapter, three different experiments were performed with the purpose of gauging the effectiveness of HTM and Grid HTM on videos. The experiments show that it is possible to apply Grid HTM for anomaly detection in videos. For instance, results show that Grid HTM is able to learn the norm in a complex surveillance video, and is therefore able to detect anomalous events and where they occur in the frame. It was also shown that for complex data, the anomaly score output is not a good measure to use due to the noise and the influence of the size of the anomalies.

The first experiment is a controlled experiment where a computer-generated ball is bouncing with anomalies inserted. The aim of this experiment is to test whether the capabilities of HTM apply for videos, as well as the performance of Grid HTM on the same task. The results confirm this, and show that the performance of Grid HTM is slightly worse than that of normal HTM. This is to be expected since the data is very clean and simple, whereas Grid HTM was designed with more complex data in mind.

The second experiment showcases the performance of Grid HTM on a surveillance video with technical anomalies. Additionally, several key points of interests and the respective outputs of Grid HTM are shown in order to get a better understanding of its capabilities. Again, the results show that Grid HTM can adapt to the norm in the complex video and can correctly identify anomalies. An interesting discovery is the ability of Grid HTM to detect segment changes. The results also show that the anomaly score output cannot be relied upon for thresholding purposes, and that further work is required in that area.

The third experiment further explores the ability of the Grid HTM to detect segments in videos, which was discovered in the previous experiment. The videos that are used in this experiment are videos of sperm that contain several segments. The results show that it outperforms the L1 benchmark due to a more prominent change in anomaly score during segment changes.

Chapter 5

Conclusion & Future Work

5.1 Summary

Smart surveillance systems have seen increased demand in the past few years. Modern smart surveillance systems depend on deep learning for their intelligence. However, it has been shown that deep learning faces several challenges. Some challenges are explainability, noise-tolerance, training data volume, and concept drift. While there are works that attempt to address these challenges, it is still important to look elsewhere for learning algorithms which do not face those same issues. One of them is HTM theory which introduces a learning algorithm that models the learning mechanism in the neocortex.

Unlike deep learning, HTM works by using SDR to represent data and learns through Hebbian like learning. This gives it the property of noise-tolerance and online learning, meaning it can handle concept drift. A natural question to ask is whether it can be used for anomaly detection in videos. It has been shown that HTM performs well in low-dimensional data such as temperature data. However, it has also been shown that it performs poorly in high dimensional data such as images and videos due to the difficulty in converting that type of data to SDRs.

With that in mind, this thesis attempts to make it possible to conduct anomaly detection in videos with the introduction of Grid HTM. Instead of having a single HTM model run anomaly detection on videos, Grid HTM divides the frames into a grid where each individual cell has its own HTM model. This makes the entire system more invariant, gives it increased flexibility, and increases explainability. The videos themselves can be converted into SDRs using techniques such as deep learning segmentation or binary thresholding.

This thesis then introduces three experiments, each aiming to prove different aspects of HTM and Grid HTM. The first experiment is a simple video of a bouncing ball, which has the aim of proving that a single HTM model can perform anomaly detection on a simple and controlled video, it

then does the same but with Grid HTM instead to prove that it still works. The second experiment aims to showcase the capabilities of Grid HTM on a complex surveillance video. It shows that, given the limited and noisy data, Grid HTM is able to learn the norm and backs it up with concrete examples. It also showed that Grid HTM was able to detect video segments. The third experiment explores the capability of Grid HTM for detecting segments using a very noisy sperm video dataset.

5.2 Contributions

As introduced in Chapter 1, this thesis achieved three objectives that would help answer the thesis question. The objectives and how they were achieved are as follows:

Objective 1 *Introduce HTM and give a deep understanding of the inner workings, the strengths, and the weaknesses. While also being easy to grasp for readers with a machine learning background.*

This objective was achieved in Chapter 2, where HTM was explained. It was explained in a straight-forward manner with references to detailed figures. It also covered common misconceptions and included information that is hard to come by. The chapter also highlighted the importance of the HTM community by including information directly from community discussions.

Objective 2 *Develop and outline a theoretically sound HTM architecture that can be applied for anomaly detection in complex videos.*

This objective was achieved in Chapter 3, where Grid HTM was introduced. It was introduced with the purpose of performing anomaly detection in videos. The chapter made sure that Grid HTM followed the rules that an HTM model should follow. It also introduced challenges and issues, and followed up with remedies.

Objective 3 *Perform experiments, discuss the results, and lay out potential future work for Grid HTM. The experiments will vary in difficulty, complexity, and will focus on different use cases.*

This objective was achieved in Chapter 4, where three different experiments were performed. The first experiment showcased that HTM and Grid HTM can indeed perform on simple and clean videos. The second experiment showcased the performance of Grid HTM on a complex surveillance video, which showed promising results. The third experiment showcased the ability of Grid HTM to detect segments in a video, for increased challenge noisy videos of sperm were used.

Now that the three objectives have been achieved, it is possible to answer the thesis question: **Is HTM viable for anomaly detection in videos?** The

experiments show that with proper data and further refinements, **Grid HTM and other HTM based architectures could indeed be used for anomaly detection systems for videos.**

5.3 Future Work

Seeing as this thesis presents a novel approach, there is naturally a lot of future work that can be done.

Datasets

As mentioned in Chapter 1, one of the main limitations is the lack of video datasets suited for anomaly detection by HTM. Therefore, the most important future work would be to create such a dataset. The videos would optimally be several days long and contain anomalies such as car accidents, jaywalking, and other similar anomalous behaviors.

Grid HTM

For Grid HTM, more time should be spent exploring other aggregation functions so that the aggregated anomaly score can be used more efficiently. One could use deep learning for this purpose or perhaps use another layer of HTM, the possibilities are endless.

Additionally, it would be a big benefit to create an algorithm which can decide the parameters for each cell during the calibration phase. It is also possible to improve explainability and robustness by implementing a measure of certainty for each cell.

Depth vision or 3D vision should be experimented with, as the depth information could be valuable for anomaly detection in surveillance. With voxels, this could be used similarly to 2D segmentation, where there could be an extra SDR for each layer of depth in the voxelized 3D image.

Finally, experiments should be performed to validate the possibility of having the TM in each cell grow synapses to neighboring cells in order to solve the issue with unstable anomaly output, which was mentioned in Section 3.2.5.

HTM and Deep Learning

Another important field to research is a tighter integration between HTM and deep learning. This way it could be possible to leverage the self-supervision and noise resilience property of HTM, together with the powerful feature extraction and representation of deep learning approaches. Effectively combining the best of both approaches while eliminating the disadvantages that have been mentioned in Chapter 2.

Research Updates

HTM theory is in constant development, especially as the understanding of the brain grows. Future work would therefore include keeping up to date on the latest developments within HTM theory and neuroscientifical research, and update the model and add new systems accordingly.

Bibliography

- [1] Divyanshi Tewari. *U.S. Video Surveillance Market by Component (Solution, Service, and Connectivity Technology), Application (Commercial, Military & Defense, Infrastructure, Residential, and Others), and Customer Type (B2B and B2C): Opportunity Analysis and Industry Forecast, 2020–2027*. Mar. 2019. URL: <https://www.alliedmarketresearch.com/us-video-surveillance-market-A06741>.
- [2] Web of Science. Jan. 2022. URL: <https://www.webofscience.com/wos/woscc/summary/f6ae0ce5-4319-416f-ab92-08d042bc3871-21874d31/relevance/1>.
- [3] Ilya Daylidyonok, Anastasiya Frolenkova, and Aleksandr Panov. “Extended Hierarchical Temporal Memory for Motion Anomaly Detection: Proceedings of the Ninth Annual Meeting of the BICA Society.” In: Jan. 2019, pp. 69–81. ISBN: 978-3-319-99315-7. DOI: 10.1007/978-3-319-99316-4_10.
- [4] Jeff Hawkins and Dileep George. “Hierarchical Temporal Memory Concepts, Theory, and Terminology.” In: 2006.
- [5] Jeff Hawkins, Subutai Ahmad, and Yuwei Cui. “A Theory of How Columns in the Neocortex Enable Learning the Structure of the World.” In: *Frontiers in Neural Circuits* 11 (2017). ISSN: 1662-5110. DOI: 10.3389/fncir.2017.00081. URL: <https://www.frontiersin.org/article/10.3389/fncir.2017.00081>.
- [6] Jeff Hawkins et al. “A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex.” In: *Frontiers in Neural Circuits* 12 (2019), p. 121. ISSN: 1662-5110. DOI: 10.3389/fncir.2018.00121. URL: <https://www.frontiersin.org/article/10.3389/fncir.2018.00121>.
- [7] Yuwei Cui, Subutai Ahmad, and Jeff Hawkins. *The HTM Spatial Pooler – a neocortical algorithm for online sparse distributed coding*. Feb. 2017. DOI: 10.1101/085035.
- [8] Jeff Hawkins and Subutai Ahmad. “Why Neurons Have Thousands of Synapses, A Theory of Sequence Memory in Neocortex.” In: *Frontiers in Neural Circuits* 10 (Oct. 2015). DOI: 10.3389/fncir.2016.00023.

- [9] M. Otahal, D. Keeney, D. McDougall, et al. *HTM.core implementation of Hierarchical Temporal Memory*. <https://github.com/htm-community/htm.core/>. 2019.
- [10] Vladimir Monakhov. *SP Topology is weird when input width and height are different*. Nov. 2021. URL: <https://github.com/htm-community/htm.core/issues/961>.
- [11] Vladimir Monakhov. *Master Thesis*. <https://github.com/vladim0105/MasterThesis/>. 2022.
- [12] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65 6 (1958), pp. 386–408.
- [13] Yoav Freund and Robert Schapire. “Large Margin Classification Using the Perceptron Algorithm.” In: *Machine Learning* 37 (Feb. 1999). DOI: 10.1023/A:1007662407062.
- [14] Marvin Minsky and Seymour Papert. “Perceptrons.” In: (1969).
- [15] Mikel Olazaran. “A Sociological Study of the Official History of the Perceptrons Controversy.” In: *Social Studies of Science* 26.3 (1996), pp. 611–659. DOI: 10.1177/030631296026003005. eprint: <https://doi.org/10.1177/030631296026003005>. URL: <https://doi.org/10.1177/030631296026003005>.
- [16] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors.” In: *nature* 323.6088 (1986), pp. 533–536.
- [17] Abien Fred Agarap. *Deep Learning using Rectified Linear Units (ReLU)*. 2019. arXiv: 1803.08375 [cs.NE].
- [18] Sebastian Ruder. “An overview of gradient descent optimization algorithms.” In: *arXiv preprint arXiv:1609.04747* (2016).
- [19] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. “A training algorithm for optimal margin classifiers.” In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.
- [20] L. Kuncheva. *Pattern Recognition and Neural Networks*. LULU COM, 2019. ISBN: 9780244232528. URL: <https://books.google.no/books?id=A8TKDwAAQBAJ>.
- [21] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. “Learning long-term dependencies with gradient descent is difficult.” In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166.
- [22] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html>.
- [23] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

- [24] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning.” In: *Journal of big data* 6.1 (2019), pp. 1–48.
- [25] Dana H Ballard. “Modular learning in neural networks.” In: *Aaaai*. Vol. 647. 1987, pp. 279–284.
- [26] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory.” In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.
- [27] Junyoung Chung et al. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. 2014. arXiv: 1412.3555 [cs.NE].
- [28] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].
- [29] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].
- [30] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].
- [31] Y. Lecun et al. “Gradient-based learning applied to document recognition.” In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.
- [32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks.” In: *Advances in neural information processing systems* 25 (2012).
- [33] Rajat Raina, Anand Madhavan, and Andrew Y Ng. “Large-scale deep unsupervised learning using graphics processors.” In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 873–880.
- [34] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library.” In: *Advances in neural information processing systems* 32 (2019).
- [35] Martin Abadi et al. “Tensorflow: Large-scale machine learning on heterogeneous distributed systems.” In: *arXiv preprint arXiv:1603.04467* (2016).
- [36] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].
- [37] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: 1312.6114 [stat.ML].
- [38] Divya Saxena and Jiannong Cao. *Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions*. 2020. DOI: 10.48550/ARXIV.2005.00065. URL: <https://arxiv.org/abs/2005.00065>.
- [39] Haiyang Chen. “Challenges and Corresponding Solutions of Generative Adversarial Networks (GANs): A Survey Study.” In: *Journal of Physics: Conference Series* 1827 (Mar. 2021), p. 012066. DOI: 10.1088/1742-6596/1827/1/012066.

- [40] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. DOI: 10.48550 / ARXIV.1701.07875. URL: <https://arxiv.org/abs/1701.07875>.
- [41] Luke Metz et al. *Unrolled Generative Adversarial Networks*. 2016. DOI: 10.48550/ARXIV.1611.02163. URL: <https://arxiv.org/abs/1611.02163>.
- [42] Jun-Yan Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. 2017. DOI: 10.48550 / ARXIV.1703.10593. URL: <https://arxiv.org/abs/1703.10593>.
- [43] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. 2014. DOI: 10.48550/ARXIV.1411.1784. URL: <https://arxiv.org/abs/1411.1784>.
- [44] Shivani Gupta and Atul Gupta. “Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review.” In: *Procedia Computer Science* 161 (2019). The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia, pp. 466–474. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2019.11.146>. URL: <https://www.sciencedirect.com/science/article/pii/S1877050919318575>.
- [45] Dan Hendrycks and Thomas Dietterich. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*. 2019. arXiv: 1903.12261 [cs.LG].
- [46] Chen Sun et al. “Revisiting unreasonable effectiveness of data in deep learning era.” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 843–852.
- [47] Alexander D’Amour et al. *Underspecification Presents Challenges for Credibility in Modern Machine Learning*. Nov. 2020.
- [48] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.” In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: <https://doi.org/10.1016/j.inffus.2019.12.012>. URL: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.
- [49] Jacob Gildenblat and contributors. *PyTorch library for CAM methods*. <https://github.com/jacobgil/pytorchGradCam>. 2021.
- [50] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization.” In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.
- [51] Jost Tobias Springenberg et al. *Striving for Simplicity: The All Convolutional Net*. 2015. arXiv: 1412.6806 [cs.LG].
- [52] Guansong Pang et al. “Deep Learning for Anomaly Detection.” In: *ACM Computing Surveys* 54.2 (Apr. 2021), pp. 1–38. ISSN: 1557-7341. DOI: 10.1145/3439950. URL: <http://dx.doi.org/10.1145/3439950>.

- [53] Katarzyna Michałowska et al. “Anomaly Detection with Unknown Anomalies: Application to Maritime Machinery.” In: *IFAC-PapersOnLine* 54.16 (2021). 13th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2021, pp. 105–111. ISSN: 2405-8963. DOI: <https://doi.org/10.1016/j.ifacol.2021.10.080>. URL: <https://www.sciencedirect.com/science/article/pii/S2405896321014828>.
- [54] Wei Fan et al. “Using Artificial Anomalies to Detect Unknown and Known Network Intrusions.” In: *Knowledge and Information Systems* 6 (Oct. 2001). DOI: 10.1007/s10115-003-0132-7.
- [55] Debashree Devi, Saroj Biswas, and Biswajit Purkayastha. “A Review on Solution to Class Imbalance Problem: Undersampling Approaches.” In: Aug. 2021.
- [56] Debanjan Datta, Sathappan Muthiah, and Naren Ramakrishnan. *Detecting Anomalies Through Contrast in Heterogeneous Data*. 2021. arXiv: 2104.01156 [cs.LG].
- [57] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. *GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training*. 2018. arXiv: 1805.06725 [cs.CV].
- [58] Sijie Zhu, Chen Chen, and Waqas Sultani. *Video Anomaly Detection for Smart Surveillance*. 2020. arXiv: 2004.00222 [cs.CV].
- [59] Tung Kieu et al. “Outlier Detection for Time Series with Recurrent Autoencoder Ensembles.” In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 2725–2732. DOI: 10.24963/ijcai.2019/378. URL: <https://doi.org/10.24963/ijcai.2019/378>.
- [60] G Sreenu and MA Saleem Durai. “Intelligent video surveillance: a review through deep learning techniques for crowd analysis.” In: *Journal of Big Data* 6.1 (2019), pp. 1–27.
- [61] Dong Gong et al. *Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection*. 2019. DOI: 10.48550/ARXIV.1904.02639. URL: <https://arxiv.org/abs/1904.02639>.
- [62] Wen Liu et al. *Future Frame Prediction for Anomaly Detection – A New Baseline*. 2017. DOI: 10.48550/ARXIV.1712.09867. URL: <https://arxiv.org/abs/1712.09867>.
- [63] J. Hawkins et al. “Biological and Machine Intelligence (BAMI).” Initial online release 0.4. 2016. URL: <https://numenta.com/resources/biological-and-machine-intelligence/>.
- [64] Taki Hasan Rafi. *A Brief Review on Spiking Neural Network - A Biological Inspiration*. Apr. 2021. DOI: 10.20944/preprints202104.0202.v1.

- [65] Jeff Hawkins, Subutai Ahmad, and Yuwei Cui. "Why does the neocortex have layers and columns, a theory of learning the 3d structure of the world." In: *bioRxiv* (2017), p. 162263.
- [66] MRaptor. June 2016. URL: <https://discourse.numenta.org/t/htm-cheat-sheet/828>.
- [67] Jeff Hawkins and Subutai Ahmad. "Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex." In: *Frontiers in Neural Circuits* 10 (2016), p. 23. ISSN: 1662-5110. DOI: 10.3389/fncir.2016.00023. URL: <https://www.frontiersin.org/article/10.3389/fncir.2016.00023>.
- [68] David McDougall (ctrl-z-9000-times). Sept. 2019. URL: <https://github.com/htm-community/htm.core/issues/259#issuecomment-533333336>.
- [69] Fabian Fallas-Moya and Francisco J. Torres-Rojas. "Object Recognition Using Hierarchical Temporal Memory." In: Jan. 2018, pp. 1–14. ISBN: 978-3-319-76260-9. DOI: 10.1007/978-3-319-76261-6_1.
- [70] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Aug. 28, 2017. arXiv: cs.LG/1708.07747 [cs.LG].
- [71] Y. Zou et al. "Hierarchical Temporal Memory Enhanced One-Shot Distance Learning for Action Recognition." In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. 2018, pp. 1–6. DOI: 10.1109/ICME.2018.8486447.
- [72] Jia Deng et al. "ImageNet: A large-scale hierarchical image database." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [73] Oleg Iegorov. *My analysis on why Temporal Memory prediction doesn't work on sequential data*. Dec. 2017. URL: <https://discourse.numenta.org/t/my-analysis-on-why-temporal-memory-prediction-doesnt-work-on-sequential-data/3141>.
- [74] Subutai Ahmad et al. "Unsupervised real-time anomaly detection for streaming data." In: *Neurocomputing* 262 (2017). Online Real-Time Learning Strategies for Data Streams, pp. 134–147. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2017.04.070>. URL: <http://www.sciencedirect.com/science/article/pii/S0925231217309864>.
- [75] Sam Heiserman (sheiser1). Jan. 2022. URL: <https://discourse.numenta.org/t/htm-core-am-i-getting-prediction-density-correctly/9299>.
- [76] *HTM Legacy Applications*. URL: <https://numenta.com/machine-intelligence-technology/applications/>.
- [77] *Whitepaper: HTM for Rogue Behavior Detection*. URL: <https://numenta.com/assets/pdf/whitepapers/Rogue%20Behavior%20Detection%20White%20Paper.pdf>.
- [78] *Whitepaper: HTM for Geospatial Tracking*. URL: <https://numenta.com/assets/pdf/whitepapers/Geospatial%20Tracking%20White%20Paper.pdf>.

- [79] *Github: HTM for Finance*. URL: <https://github.com/numamenta/numamenta-apps>.
- [80] Noha O. El-Ganainy et al. “On the Performance of Hierarchical Temporal Memory Predictions of Medical Streams in Real Time.” In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. 2019, pp. 1–6. DOI: 10.1109/ISMICT.2019.8743902.
- [81] EN Osegi. “Using the hierarchical temporal memory spatial pooler for short-term forecasting of electrical load time series.” In: *Applied Computing and Informatics* (2020).
- [82] Francisco De Sousa Webber. *Semantic Folding Theory And its Application in Semantic Fingerprinting*. 2016. arXiv: 1511.08855 [cs.AI].
- [83] Vajira Thambawita et al. “DivergentNets: Medical Image Segmentation by Network Ensemble.” In: *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) co-located with the 17th IEEE International Symposium on Biomedical Imaging (ISBI 2021)*. 2021.
- [84] Christian Szegedy et al. *Going Deeper with Convolutions*. 2014. arXiv: 1409.4842 [cs.CV].
- [85] Tsung-Yi Lin et al. *Feature Pyramid Networks for Object Detection*. 2017. arXiv: 1612.03144 [cs.CV].
- [86] Ethan Rublee et al. “ORB: An efficient alternative to SIFT or SURF.” In: *2011 International Conference on Computer Vision*. 2011, pp. 2564–2571. DOI: 10.1109/ICCV.2011.6126544.
- [87] Sangmin Oh et al. “A large-scale benchmark dataset for event recognition in surveillance video.” In: *CVPR 2011*. 2011, pp. 3153–3160. DOI: 10.1109/CVPR.2011.5995586.
- [88] Miroslav Kovar. *Implementation of HTM Spatial Pooler algorithm in CUDA*. https://github.com/mirgee/sp_cuda. 2018.
- [89] David Di Giorgio Jacob Everist. *BrainBlocks*. <https://github.com/the-aerospace-corporation/brainblocks>. 2020.
- [90] Alexander Kirillov et al. *PointRend: Image Segmentation as Rendering*. 2020. arXiv: 1912.08193 [cs.CV].
- [91] Ayoola Olafenwa. “Simplifying Object Segmentation with PixelLib Library.” In: 2021.
- [92] Trine B. Haugen et al. “VISEM: A Multimodal Video Dataset of Human Spermatozoa.” In: *Proceedings of the 10th ACM on Multimedia Systems Conference*. MMSys’19. Amherst, MA, USA: ACM, 2019. ISBN: 78-1-4503-6297-9. DOI: 10.1145/3304109.3325814. URL: <http://doi.acm.org/10.1145/3304109.3325814>.
- [93] Shixing Chen et al. *Shot Contrastive Self-Supervised Learning for Scene Boundary Detection*. 2021. DOI: 10.48550/ARXIV.2104.13537. URL: <https://arxiv.org/abs/2104.13537>.

Appendix A

Paper - GRID HTM: Hierarchical Temporal Memory for Anomaly Detection in Videos

GRID HTM: Hierarchical Temporal Memory for Anomaly Detection in Videos

Blinded for review

email

blinded

blinded

ABSTRACT

The interest for video anomaly detection systems has gained traction for the past few years. The current approaches use deep learning for performing anomaly detection in videos, but this has multiple problems. For starters, deep learning in general has issues with noise, concept drift, explainability, and training data volumes. Additionally, anomaly detection in itself is a complex task and faces challenges such as unknownness, heterogeneity, and class imbalance. Anomaly detection in deep learning is therefore mainly constrained to generative models such as generative adversarial networks and autoencoders due to their unsupervised nature, but even they suffer from general deep learning issues and are hard to train properly. In this paper, we explore the capabilities of the hierarchical temporal memory (HTM) algorithm to perform anomaly detection in videos, as it has favorable properties such as noise tolerance and online learning which combats concept drift. We introduce a novel version of HTM, namely, Grid HTM, which is an HTM-based architecture specifically for anomaly detection in complex videos such as surveillance footage.

CCS CONCEPTS

- Computing methodologies → Computer vision.

KEYWORDS

HTM, deep learning, surveillance, anomaly detection

ACM Reference Format:

Blinded for review. 2018. GRID HTM: Hierarchical Temporal Memory for Anomaly Detection in Videos. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXXX.XXXXXXXX>

1 INTRODUCTION

As the global demand for security and automation increases, many seek to use video anomaly detection systems. In the US alone, the surveillance market is expected to reach \$23.60 Billion by 2027 [1]. Leveraging modern computer vision, modern anomaly detection systems play an important role in increasing monitoring efficiency and reducing the need for expensive live monitoring. Their use

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXXXX.XXXXXXXX>

cases can vary from detecting faulty products on an assembly line to detecting car accidents on a highway, and everything in between.

The most important component in video anomaly detection systems is the intelligence behind it. The intelligence ranges from simple on-board algorithms to advanced deep learning models, where the latter has experienced increased popularity in the past few years. Yet, despite the major progress within the field of deep learning, there are still many tasks where humans outperform models, especially in anomaly detection where the anomalies are often undefined. Deep learning approaches also perform poorly when dealing with noise and concept drift.

The cause for the discrepancy lies in the difference between how humans and machine learning algorithms represent data and learn. Most machine learning algorithms use a dense representation of the data and apply back-propagation in order to learn. Human learning happens in the neocortex, where evidence points to that the neocortex uses a sparse representation and performs Hebbian-style learning. For the latter, there is a growing field of machine learning dedicated to replicating the inner mechanics of the neocortex, namely Hierarchical Temporal Memory (HTM) theory [2]. This theory outlines its advantages over standard machine learning, such as noise-tolerance and the ability to adapt to changing data.

With the advantages of HTM and the rise of video anomaly detection in mind, a natural question one could pose is whether it is possible to apply HTM for anomaly detection in videos. Combined with a lack of related works, it is this very question that is the motivation behind this paper.

2 RELATED WORK

Anomaly detection is often defined as detecting data points that deviate from the general distribution [3]. Unlike most other problems in deep learning, anomaly detection deals with unpredictable and rare events which makes it hard to apply traditional deep learning for anomaly detection. A subset of anomaly detection is smart surveillance [4], which is the use of video analysis specifically in surveillance.

An issue for deep-learning models in general is that they are susceptible to noise in the dataset [5, 6], which leads to decreased classification accuracy and poor prediction results. Due to the nature of training deep learning models, they are also in most cases not self-supervised and therefore require constant tuning in order to stay effective on changing data. Not to mention that they require a lot of data before they can be considered effective, and that performance increases logarithmically based on the volume of training data [7]. Deep learning models also suffer from issues with out-of-distribution performance, where a model might perform great on the dataset it is tested on, but performs poorly when deployed in real life. This could be caused by selection bias in the dataset

or when there are differences in the causal structure between the training domain and the deployment domain [8].

Deep learning models generally suffer from a lack of explainability [9]. While it is known *how* the models make their decisions, their huge parametric spaces make it unfeasible to know *why* they make those predictions. Combined with the vast potential that deep learning offers in critical sectors such as medicine, makes approaches that offer explainability highly attractive.

HTM theory [2] introduces a machine learning algorithm which works on the same principles as the brain and therefore solves some of the issues that deep learning has. HTM is considered noise resistant and can perform online learning, meaning that it learns as it observes more data. HTM replicates the structure of the neocortex which is made up of cortical regions, which in turn are made of up of mini-columns and then neurons.

The data in an HTM is represented using a Sparse Distributed Representation (SDR), which is a sparse bit array. An encoder converts real world values into SDRs, and there are currently encoders for numbers, geospatial locations, categories, and dates. One of the difficulties with HTM is making it work on visual data, where creating a good encoder for visual data is still being researched [10, 11, 12].

The learning mechanism consists of two parts, the Spatial Pooler (SP) and the Temporal Memory (TM). The SP learns to extract semantically important information into output SDRs. The TM learns sequences of patterns of SDRs and forms a prediction in the form of a predictive SDR.

Research has shown that HTM is very capable of performing anomaly detection on low-dimensional data and is able to outperform other anomaly detection methods [13]. However, related works, such as Daylidyonok, Frolenkova, and Panov [12], show that HTM struggles with higher dimensional data. Therefore, a natural conclusion to make is that HTM should be applied differently, and that a new type of architecture using HTM should be explored for the purpose of video anomaly detection and surveillance.

3 GRID HTM

This paper explores a new type of architecture, named Grid HTM, for anomaly detection in videos using HTM, and proposes to use segmentation techniques to simplify the data into an SDR-friendly format. These segmentation techniques could be anything from simple binary thresholding to deep learning instance segmentation. Even keypoint detectors such as ORB [14] could in theory be applied. When explaining Grid HTM, the examples will be taken from deep learning instance segmentation of cars on a video from the VIRAT [15] dataset. An example segmentation is shown in Figure 1. The idea is that the SP will learn to find an optimal general



Figure 1: Segmentation result of cars, which is suited to be used as an SDR. Original frame taken from VIRAT [15].

representation of cars. How general this representation is can be configured using the various SP parameters, but ideally they should be set so that different cars will be represented similarly while trucks and motorcycles will be represented differently. An example representation by the SP is shown in Figure 2.

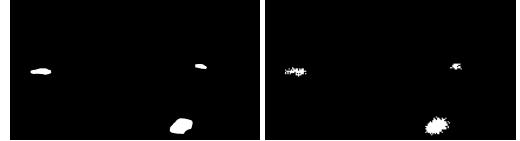


Figure 2: The SDR (left) and its corresponding SP representation (right). Note that the SP is untrained.

The task of the TM will then be to learn the common patterns that the cars exhibit, their speed, shape, and positioning will be taken into account. Finally, the learning will be set so that new patterns are learned quickly, but forgotten slowly. This will allow the model to quickly learn the norm, even if there is little activity, while still reacting to anomalies. This requires that the input is stationary, in our example this means that the camera is not moving.

It is possible to split different segmentation classes into their respective SDRs. This will give the SP and the TM the ability to learn different things for each of the classes. For instance, if there are two classes "person" and "car", then the TM will learn that it is normal for objects belonging to "person" to be on the sidewalk, while objects belonging to "car" will be marked as anomalous when on the sidewalk.

Ideally, the architecture will have a calibration period spanning several days or weeks, during which the architecture is not performing any anomaly detection, but is just learning the patterns.

3.1 Improvements

Daylidyonok, Frolenkova, and Panov [12] was shown to not be particularly effective, therefore multiple improvements needed to be introduced to increase effectiveness. **Invariance.** One issue that becomes evident is the lack of invariance, due to the TM learning the global patterns. Using the example, it learns that it is normal for cars to drive along the road but only in the context of there being cars parked in the parking lot. It is instead desired that the TM learns that it is normal for cars to drive along the road, regardless of whether there are cars in the parking lot. We propose a solution based on dividing the encoder output into a grid and have a separate SP and TM for each cell in the grid. The anomaly scores of all the cells are then aggregated into a single anomaly score using an aggregation function.

Aggregation Function. Selecting the correct aggregation function is important because it affects the final anomaly output. For instance, it might be tempting to use the mean of all the anomaly scores as the aggregation function:

$$X : \{x \in \mathbb{R} : x \geq 0\}$$

$$\text{anomScore} = \frac{\sum_{x \in X} x}{|X|}$$

However, this leads to problems with normalization, meaning that an overall anomaly score of 1 is hard to achieve due to many cells having a zero anomaly score. In fact, it becomes unclear what a high anomaly score is anymore. Using the mean also means that anomalies that take up a lot of space will be weighted higher than anomalies that take up a little space, which might not be desirable. To solve the aforementioned problem and if the data has little noise, a potential aggregation function could be the non-zero mean:

$$X : \{x \in \mathbb{R} : x > 0\}$$

$$\text{anomScore} = \begin{cases} \frac{\sum_{x \in X} x}{|X|} & \text{if } |X| > 0 \\ 0 & \text{otherwise} \end{cases}$$

Meaning that only the cells with a non-zero anomaly score, denoted X , will be contributing to the overall anomaly score which helps solve the aforementioned normalization and weighting problem. On the other hand, this will perform poorly when the architecture is exposed to noisy data which could lead to there always being one or more cells with a high anomaly score. Figure 3 illustrates

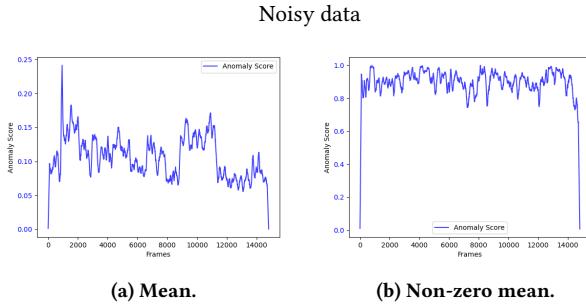


Figure 3: How the two aggregation functions perform on the same noisy data.

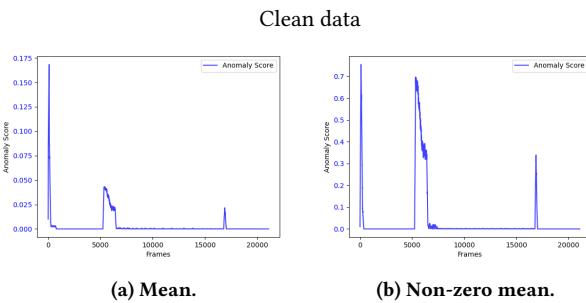


Figure 4: How the two aggregation functions perform on the same clean data.

the effect of an aggregation function for noisy data, where the non-zero mean is rendered useless due to the noise. On the other hand, Figure 4 shows how the non-zero mean gives a clearer anomaly

score when the data is clean. **Explainability.** Having the encoder output divided into a grid has the added benefit of introducing explainability into the model. By using Grid HTM it is now possible to determine where in the input an anomaly has occurred by simply observing which cell has a high anomaly score. It is also possible to estimate the number of predictions for each cell which can be used as a measure of certainty, where fewer predictions means higher certainty. Making it possible to measure certainty per cell creates a new source of information which can be used for explainability or robustness purposes. **Flexibility and Performance.** In addition, it is also possible to configure the SP and the TM in each cell independently, giving the architecture increased flexibility and to use a non-uniform grid, meaning that some cells can have different sizes. Last but not least, dividing the frame into smaller cells makes enables it to run each cell in parallel for increased performance.

Reviewing Encoder Rules. A potential challenge with the grid approach is that the rules for creating a good encoder, may not be respected and therefore should be reviewed:

- **Semantically similar data should result in SDRs with overlapping active bits.** In this example, a car at one position will produce an SDR with a high amount of overlapping bits as another car at a similar position in the input image.
- **The same input should always produce the same SDR.** The segmentation model produces a deterministic output given the same input.
- **The output must have the same dimensionality (total number of bits) for all inputs.** The segmentation model output has a fixed dimensionality.
- **The output should have similar sparsity (similar amount of one-bits) for all inputs and have enough one-bits to handle noise and subsampling.** The segmentation model does not respect this. An example is that there can be no cars (zero active bits), one car (n active bits), or two cars ($2n$ active bits), and that this will fluctuate over time.

The solution for the last rule is two-fold, and consists of imposing a soft upper bound and a hard lower bound for the number of active pixels within a cell. The purpose is to lower the variation of number of active pixels, while also containing enough semantic information for the HTM to work:

- Pick a cell size so that the distribution of number of active pixels is as tight as possible, while containing enough semantic information and also being small enough so that the desired invariance is achieved. The cell size acts as a soft upper bound for the possible number of active pixels.
- Create a pattern representing emptiness, where the number of active bits is similar to what can be expected on average when there are cars inside a cell. This acts as a hard lower bound for the number of active pixels.

There could be situations where a few pixels are active within a cell, which could happen when a car has just entered a cell, but this is acceptable as long as it does not affect the distribution too much. If it does affect the distribution, which can be the case with noisy data, then an improvement would be to add a minimum sparsity requirement before a cell is considered not empty, e.g. less than 5 active pixels means that the cell is empty. In the following example,

the number of active pixels within a cell centered in the video was used to build the distributions seen in Figure 5:

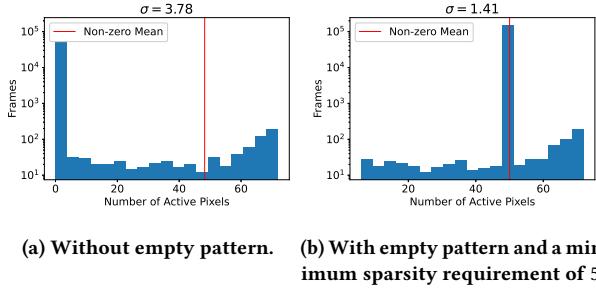


Figure 5: Distribution of number of active pixels within a cell of size 12×12 .

With a carefully selected empty pattern sparsity, **the standard deviation of active pixels was lowered from 3.78 to 1.41**. It is possible to automate this process by developing an algorithm which finds the optimal cell size and empty pattern sparsity which causes the least variation of number of active pixels per cell. This algorithm would run as a part of the calibration process.

The visual output resulting from these changes, which is an equally important output as the aggregated anomaly score, can be seen in Figure 6. Since there are now cells that are observing

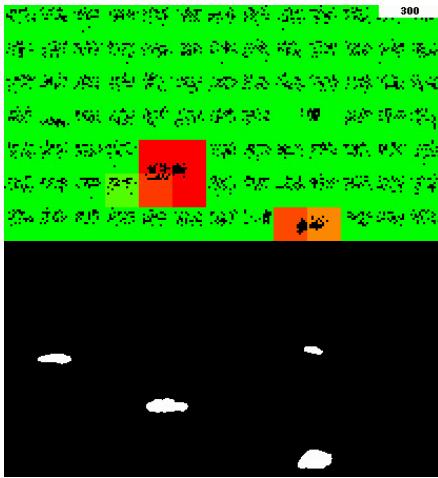


Figure 6: Example Grid HTM output and the corresponding input. The color represents the anomaly score for each of the cells, where red means high anomaly score and green means zero anomaly score. Two of the cars are marked as anomalous because they are moving, which is something the Grid HTM has not seen before during its 300 frame long lifetime.

an empty pattern for a lot of the time in sparse data, boosting is recommended to be turned off, otherwise the SP output for the empty cells would change back and forth in order to adjust the active duty cycle.

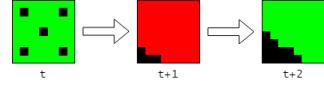


Figure 7: High anomaly score when an empty cell (represented with an empty pattern with a sparsity value of 5) changes to being not empty, as something enters the cell.

3.1.1 Stabilizing Anomaly Output. Another issue with the grid based approach is when a car first comes into a cell. The TM in that cell has no way of knowing that a car is about to enter, since it does not see outside its own cell, and therefore the first frame that a car enters a cell will cause a high anomaly output. This is illustrated in Figure 7 where it can be observed that this effect causes the anomaly output to needlessly fluctuate. The band-aid solution is to ignore the anomaly score for the frame during which the cell goes from being empty to being not empty, which is illustrated in Figure 8. A



Figure 8: In this case, the anomaly score is ignored (set to 0) for the frame in which the cell changes state from empty to not empty.

more proper solution could be to allow the TM to grow synapses to the TMs in the neighboring cells, but this is not documented in any research papers and might also hinder invariance.

3.1.2 Multistep Temporal Patterns. Since the TM can only grow segments to cells that were active in the previous timestep, it will struggle to learn temporal patterns across multiple timesteps. This is especially evident in high framerate videos, where an object in motion has a similar representation at timestep t and $t + 1$, as an object standing still.

This could cause situations where an object that is supposed to be moving, suddenly stands still, yet the TM will not mark it as an anomaly due to it being stuck in a contextual loop. A contextual loop is when one of the predictions at t becomes true at $t + 1$, and then one of the predictions at $t + 1$ is almost identical to the state at t , which becomes true if the object is not moving, causing the TM to enter the same state that it was in at t . A solution is to concatenate the past n SP outputs as input into the TM, which is made possible by keeping a buffer of past SP outputs and shifting its contents out as new SP outputs are inserted. This follows the core idea behind encoding time in addition to the data, which makes time act as a contextual anchor. However, in this case there are no timestamps that are suitable to be used as contextual anchors, so as a replacement, the past observations are encoded instead.

Concatenating past observations together will force the TM input, for when an object is in motion and when an object is still, to be unique. High framerate videos can benefit the most from this, and the effect will be more pronounced for higher values of n .

A potential side effect of introducing temporal patterns, is that because the TM is now exposed to multiple frames at once, it will be more tolerant to temporal noise. An example of temporal noise is

when an object disappears for a single frame due to falling below the classification threshold of the deep learning segmentation model encoder. The reason for the noise tolerance is that instead of the temporal noise making up the entire input for the TM, it now only makes up $\frac{1}{n}$ of the TM input.

3.1.3 Use Cases. The most intuitive use case is to use Grid HTM for semi-active surveillance, where personnel only have to look at segments containing anomalies, leading to drastically increased efficiency. One example is making it possible to have an entire city be monitored by a few people. This is made possible by making it so that people only have to look at segments that the Grid HTM has found anomalous, which is what drastically lowers the manpower requirement for active monitoring of the entire city.

4 EXPERIMENTAL DETAILS AND RESULTS

As stated earlier, one of the use cases of Grid HTM is anomaly detection in surveillance. This example will show how Grid HTM could perform. The video to be used is part of the VIRAT [15] video dataset, and was selected due to its long duration and stationary camera.

The downside is that the video does not contain any non-technical anomalies, but consists of technical anomalies in the form of several segments with sudden frame skips in between. There is also a synthetic anomaly introduced in the form of a frame repeat lasting a couple of seconds, essentially "freezing" time, in order to test whether Grid HTM is able to understand how objects should be moving in time.

In this experiment, a segmentation model which can extract classes into their respective SDRs is employed. Meaning that there could be an SDR for cars and an SDR for persons, that are then concatenated before being fed into the system.

The segmentation model used is PointRend [16] with a ResNet101 [17] backbone, pretrained on ImageNet [18], and implemented using PixelLib [19]. For the sake of simplicity, this experiment will focus only on the segmentation of cars.

While on the topic of segmentation, it is important to mention that the segmentation model is not perfect and that there are cases where objects are misclassified as well as cases where cars repeatedly go above and below the confidence threshold.

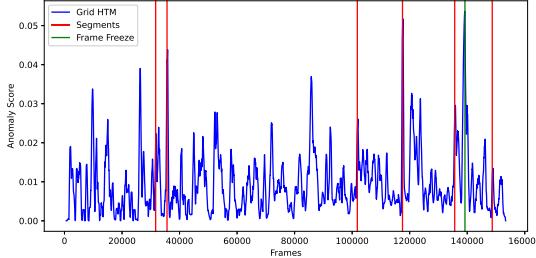


Figure 9: Anomaly score output from Grid HTM.

4.0.1 Results. We can see in Figure 9 that the Grid HTM is detecting when segments begin and end, however it is not possible to use

a threshold value to isolate them, and they also have vastly different anomaly scores compared to each other. This is due to the way the aggregation function works, which means that the anomaly output is dependent on the physical size of the anomaly. It should also be noted that a moving average ($n = 200$) was applied to smooth out the anomaly score output, otherwise the graph would be too noisy.

With the aggregation functions presented in this paper in mind, it is safe to conclude that looking at the anomaly score output is meaningless for complex data such as a surveillance video. This however does not mean that Grid HTM is completely useless, and this can be observed by looking at the visual output of the Grid HTM. The visual output during which the first segment anomaly occurs can be seen in Figure 10. Here it is observed that Grid HTM correctly

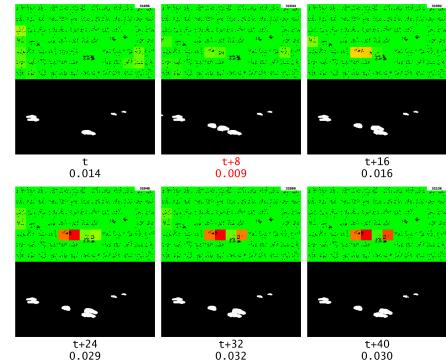


Figure 10: The first segment anomaly, which is marked with red text, and the corresponding changes detected by Grid HTM.

marks the sudden change of cars when the current segment ends and a new segment begins.

4.0.2 Road. In the original video, there is a road on which cars regularly drive. By observing the visual output, it becomes evident that after some time the Grid HTM has mostly learned that behavior and does not report those moving cars as anomalies. This is shown in Figure 11.

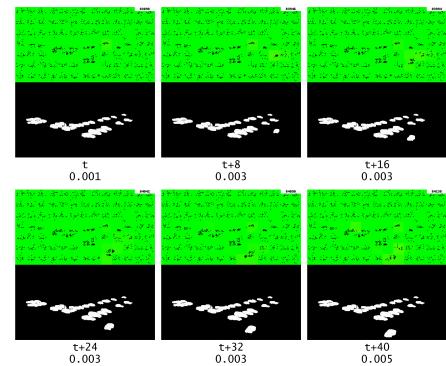


Figure 11: Visual output during when a car is driving along the main road.

4.0.3 Frame Repeat. To prove that the Grid HTM has learned that cars on the road should be moving, it is possible to look at the visual output during the period when the video is repeating the same frame and observe if the architecture marks the cars standing still on the road as anomalies. It can be observed in Figure 12 that

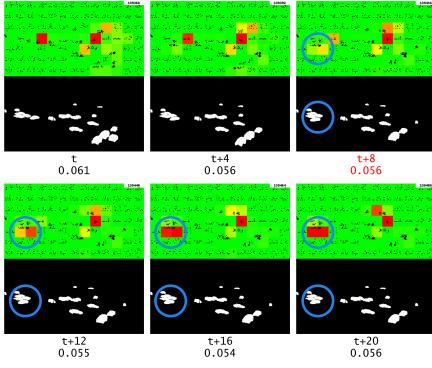


Figure 12: Anomaly output during the repeating frame, the start of the frame repeat is marked with red text. The blue circle highlights the object of interest.

the cars along the main road are not marked as anomalies, but this could be attributed to the fact that there is a crossing there and that cars periodically have to stop at that point to let pedestrians cross.

On the other hand, when looking at the anomaly marked with a blue circle, the car on the road in the parking lot is marked as an anomaly that increases in severity as the time goes on during the frame repeat. The reason why that car causes an anomaly is because, unlike the cars on the main road, a car is rarely observed as standing still at that position.

To prove that the anomaly was actually directly caused by the repeating frame, and not just due to repeating the anomaly in time, it should be compared to the anomaly output if there was no repeating frame. It can be observed in Figure 13 that the anomaly

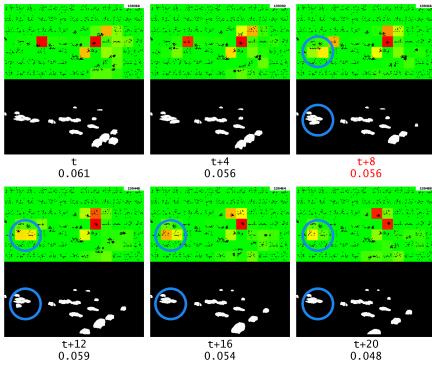


Figure 13: Anomaly output when there is no frame repeating, where it should have repeated is marked in red. The blue circle highlights the object of interest.

output is minor compared to when there was a repeating frame,

proving that the anomaly was indeed a product of the repeating frame and that the Grid HTM was able to learn how objects should be moving in time.

Finally, it is interesting to look at how the Grid HTM handles the repeating frames without multistep temporal patterns, which is shown in Figure 14. Unfortunately, simply disabling multistep tem-

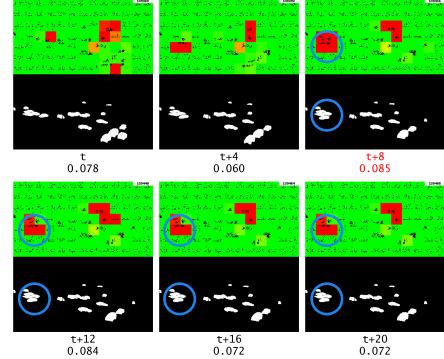


Figure 14: Anomaly output during the repeating frame, the start of the frame repeat is marked with red text. The blue circle highlights the object of interest. This time without multistep temporal patterns.

poral patterns without adjusting the other TM parameters causes the same car to be marked as an anomaly before and during the frame repeat. In fact, as mentioned in Section 3.1.2, disabling multistep temporal patterns causes Grid HTM to be less noise tolerant which causes a lot more anomalies to be wrongly detected. This is evident in Figure 14, where a higher amount of severe anomalies can be observed compared to previous examples. This also highlights how sensitive HTM can be regarding parameters. The working code for Grid HTM and the parameters for the experiments conducted in for this paper can be found on GitHub.

5 CONCLUSION

We presented a novel method to perform anomaly detection in videos. Experiments showed that the proposed Grid HTM can be used for anomaly detection in complex videos such as surveillance footage. One of the most important future work would be to create a dataset with videos that are several days long and contain anomalies such as car accidents, jaywalking, and other similar anomalous behaviors. For Grid HTM, more time can be spent exploring other aggregation functions so that the aggregated anomaly score can be used more efficiently. Additionally, it would be a big benefit to create an algorithm which can decide the parameters for each cell during the calibration phase. It is also possible to improve explainability and robustness by implementing a measure of certainty for each cell.

Finally, experiments should be performed to validate the possibility of having the TM in each cell grow synapses to neighboring cells in order to solve the issue with unstable anomaly output, which was mentioned in Section 3.1.1.

REFERENCES

- [1] Divyanshi Tewari. 2019. U.s. video surveillance market by component (solution, service, and connectivity technology), application (commercial, military & defense, infrastructure, residential, and others), and customer type (b2b and b2c): opportunity analysis and industry forecast, 2020–2027. (March 2019). <https://www.alliedmarketresearch.com/us-video-surveillance-market-A06741>.
- [2] J. Hawkins, S. Ahmad, S. Purdy, and A. Lavin. Biological and machine intelligence (bam). Initial online release 0.4, (2016). <https://numenta.com/resources/biological-and-machine-intelligence/>.
- [3] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection. *ACM Computing Surveys*, 54, 2, (April 2021), 1–38. ISSN: 1557-7341. doi: 10.1145/3439950.
- [4] Sijie Zhu, Chen Chen, and Waqas Sultani. 2020. Video anomaly detection for smart surveillance. (2020). arXiv: 2004.00222 [cs.CV].
- [5] Shivani Gupta and Atul Gupta. 2019. Dealing with noise problem in machine learning data-sets: a systematic review. *Procedia Computer Science*, 161, 466–474. The Fifth Information Systems International Conference, 23–24 July 2019, Surabaya, Indonesia. ISSN: 1877-0509. doi: <https://doi.org/10.1016/j.procs.2019.11.146>.
- [6] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. (2019). arXiv: 1903.12261 [cs.LG].
- [7] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.
- [8] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Jonathan Deaton, Jacob Eisenstein, Matthew Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, and D. Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning. (November 2020).
- [9] Alejandro Barredo Arrieta, Natalia Diaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 82–115. ISSN: 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2019.12.012>.
- [10] Y. Zou, Y. Shi, Y. Wang, Y. Shu, Q. Yuan, and Y. Tian. 2018. Hierarchical temporal memory enhanced one-shot distance learning for action recognition. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. doi: 10.1109/ICME.2018.8486447.
- [11] David McDougall (ctrl-z 9000-times). 2019. (September 2019). <https://github.com/htm-community/htm.core/issues/259#issuecomment-533333336>.
- [12] 2019. *Extended hierarchical temporal memory for motion anomaly detection: proceedings of the ninth annual meeting of the bica society*. (January 2019), 69–81. ISBN: 978-3-319-99315-7. doi: 10.1007/978-3-319-99316-4_10.
- [13] Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262, 134 – 147. Online Real-Time Learning Strategies for Data Streams. ISSN: 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2017.04.070>.
- [14] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. Orb: an efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, 2564–2571. doi: 10.1109/ICCV.2011.6126544.
- [15] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xiaoyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, 3153–3160. doi: 10.1109/CVPR.2011.5995586.
- [16] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. 2020. Pointrend: image segmentation as rendering. (2020). arXiv: 1912.08193 [cs.CV].
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. (2015). arXiv: 1512.03385 [cs.CV].
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. doi: 10.1109/CVPR.2009.5206848.
- [19] Ayoola Olafenwa. 2021. Simplifying object segmentation with pixellib library. In.