Vrije Universiteit Amsterdam

Universiteit van Amsterdam

Master Thesis

Title of the Thesis

Author: student name (student number)

1st supervisor:	supervisor name	
daily supervisor:	supervisor name	(company, if applicable)
2nd reader:	supervisor name	

A thesis submitted in fulfillment of the requirements for
the joint UvA-VU Master of Science degree in Computer Science

July 30, 2020

"I am the master of my fate, I am the captain of my soul"
from Invictus*, by William Ernest Henley*

Abstract

Here goes the abstract of this thesis.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

The KM3NeT or the Cubic Kilometer Neutrino Telescope is currently being constructed at the bottom of the Mediterranean Sea. The goal of this telescope is two fold: first is to study high energy neutrinos originating from celestial events such as birth of a neutrino star, supernovae, etc. and second, to study the properties of the neutrino particles produced in the Earth's atmosphere (1). The first goal will be realized with the KM3NeT/ARCA (Astroparticle Research with Cosmics in the Abyss) telescope and the second with KM3NeT/ORCA (Oscillation Research with Cosmics in the Abyss) (1). In this paper, we talk exclusively about KM3NeT/ARCA.

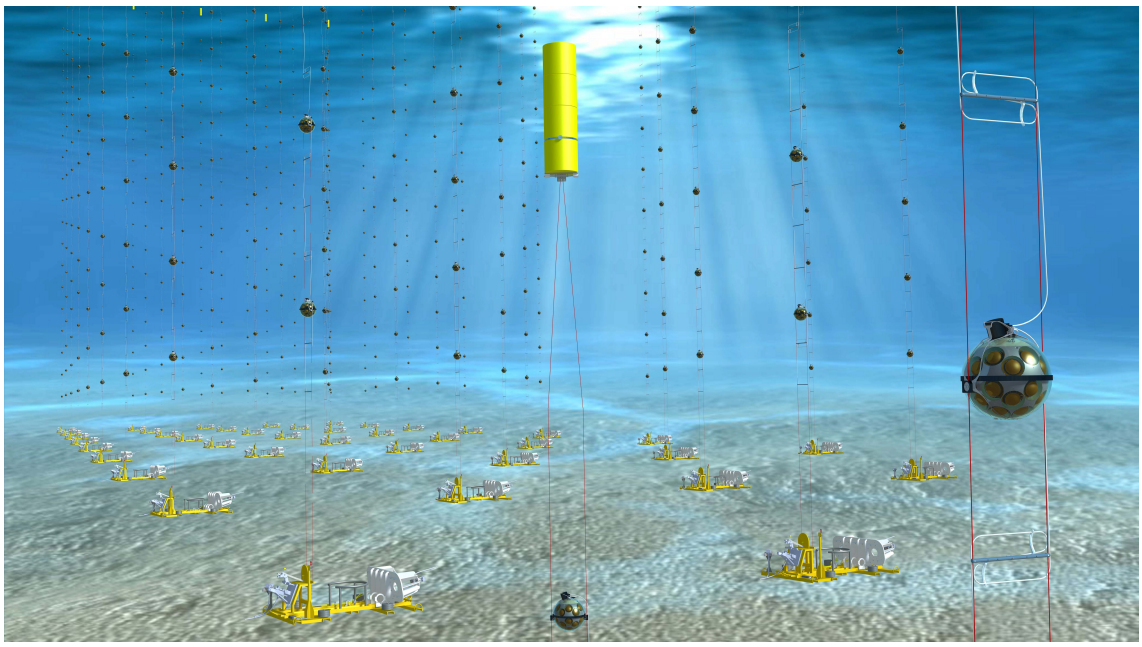

Figure 1.1: Artist's impression of the ARCA detector *source: https://www.km3net.org*

The ARCA telescope comprises of two "blocks" with a total volume of $1km^3$. Each block consists of 115 spherical detector units (referred to as DOMs henceforth) and each DOM consists of 31 Photo Multiplier Tubes (PMTs) in various spacial arrangement. Figure 1.1 shows an artist's impression of ARCA, figure 1.2 depicts a DOM along with the PMTs inside it.

Figure 1.2: An optical detector (DOM) *source: https://www.km3net.org*

The PMTs are sensitive to light or photons, the analog signal for all hits above a certain threshold are digitized. This datapoint consists of a timestamp and the spatial orientation of the DOM (ie. x,y,z coordinates). The digital signals from all PMTs are arranged in 100ms "time slices" and sent to the on-shore facility for further processing (2).

1.1 Situation of Concern

When the high energy neutrino particles interact with surrounding matter, produce an electron and a photon, this phenomenon is known as the Cherenkov Radiation or Cherenkov Light (3). This phenomenon is utilized in the KM3NeT telescopes to detect high energy neutrinos ie. the Cherenkov Light is detected by the PMTs. Unfortunately, there are several sources of noise (in this case, the noise is other light sources), bioluminescense and decay of Potassium 40 (^{40}K) and atmospheric Muons being the primary sources (4).

1.1.1 State of the Art

Due to the high level of noise, data is generated at an extremely high rate of 25GB/sec (1). Due to this high data rate, it must be filtered and selectively stored for further analysis. The state of the art for this task are known as "Event Trigger" algorithms (1, 2). The existing event trigger algorithms namely *L1* and *L2* have limitations and can be improved upon (5).

There is a tradeoff between performance and quality of the event trigger algorithms to be realized here. The state of the art *L1* and *L2* algorithms are very performant with the ability to filter data in real time however their quality of filtration can be improved. Thus new event trigger algorithms must consider this trade off.

Efforts have already been made to improve the existing event trigger algorithms. (5) proposed and implemented a GPU powered pipeline which utilizes correlation and graph community detection to identify time slices that may contain neutrino hits whilst (4) suggests an alternate using convolutional neural networks.

1.1.2 User Requirements

The primary users of the ARCA are researchers who want to study high energy particles from outer space. The stakeholders are all member institutes involved in the project and by extension all scientists from these institutes who will be working with the data collected.

The requirements of the primary users (and stakeholders) with respect to the data acquisition pipeline are as follows: 1. The accuracy with which time slices are filtered out has to be extremely high. Time slices which are deemed important by event trigger algorithms are stored for further analysis and research. Failure to store time slices containing important data can lead to loss of important data and thus a poor quality of research. 2. The filtration of noise must be highly accurate. The event trigger algorithms must be able to eliminate majority of noise to prevent storage of unnecessary and potentially useless data. 3. The state of the art event trigger algorithms are able to process data in real time. The proposed alternative ideally should maintain or improve upon it's predecessor's performance else provide a good trade off with data quality.

1.2 Research Question

This report intends to improve upon the GPU pipeline proposed by (5), specifically we wish to answer the following research questions:

RQ1. **Can the existing GPU pipeline be improved using neural networks (NNs)?**

Specifically, improvement may be achieved by reducing the processing time of the pipeline or improving the accuracy of identifying important timeslices. In order to answer this research question, the following sub questions are formulated.

RQ2. Which NN will produce the best results when replaced with the *Pattern Matrix*?

TODO add rational

RQ3. Which NN will produce the best results when replaced with the *Graph Community Detection* algorithm?

TODO add rational

2

Data Preparation

At the time of undertaking this project, the KM3NeT Neutrino Telescope was still under construction, thus simulated data provided by Nikhef was used for the project. The data itself was split onto two parts namely *events* and *noise*, both of which came from different sources and in different formats.

2.1 Preparation of *events* dataset

The *events* dataset was provided as a *HDF5* (Hierarchical Data Format) with a size of 42MB consisting of the `/data/mc_hits` and `/data/mc_info` tables. For the purposes of this project, the two tables were combined such that each row in the *mc_hits* table contains it's corresponding 'event_id' from the *mc_info* table. A `label` column was added containing a value of '1' and the resulting table (here on referred to as the *events* dataset) was saved as a CSV file for future use.

2.2 Preparation of *noise* dataset

The *noise* data was generated using a Python library written and maintained by Nikhef, `k40gen`. `k40gen.Generators(21341, 1245, [7000., 700., 70., 0.])` was used to create an instance of a generator where the first two arguments are random seeds followed by a list of rates at which single, double, triple and quadruple hits should be generated. The generator instance is then passed into `k40gen.generate_40()` method which returns a (4, n) array containing (time, dom_id, pmt_id, tot). The position coordinates (ie. `x`, `y`, `z` coordinates) for each datapoint was provided in a *positions.detx* file which was parsed using the Numpy Python package (6) and added to the *noise* array. The Python library

Pandas (7) was used to convert the array into a (n, 4) dataframe. A `label` column was added containing a value of '0' and the dataframe was saved as a 3.9GB CSV file.

2.3 Preparation of *main* dataset

To create the *main* dataset for the project, the *events* and *noise* datasets were combined. Both datasets were read into memory as Pandas dataframes and their columns were renamed consistently. The two dataframes were concatenated and sorted based on the `time` column. Rows with a negative `time` value were dropped along with columns which are not relevant to this project. The `time` column was discretized into 15000ns bins and the resulting values were added to the `timeslice` column. The resulting dataframe was saved as a 1.9GB CSV file.

3

Data Exploration

The *main* dataset (generated as per the steps outlined in Chapter 2) was explored using statistical analysis and visualizations to observe any patterns and "local trends" that may be present. The following chapter presents the analysis that were done and the observations made.

Note that a random sample of only 10% of the data was taken for the following visualizations. This is because it is difficult to draw reasonable conclusions from the plots due to the high number of data points when the entire dataset is used.

3.1 Descriptive Statistics

Table 3.2 presents the descriptive statistics of the *main* dataset. The dataset consists of 7 columns and roughly 4.5 million rows, Table 3.1 provides more information on the columns on the dataset. The dataset does not contain any `nan` or `null` values except for the `event_id` column where rows containing noise hits are not associated with any event.

Next, the correlations amongst features is checked, the "Pearson" correlation is used. Figure 3.1 represents the correlation matrix of all features. No significant correlations are observed between `pos_x`, `pos_y`, `pos_z` and `time` which indicates that ML models may not be able to learn anything from the dataset without the aid of feature engineering.

The distribution of the `label` column is presented in Figure 3.2. A severe class imbalance is noted between events and noise hits. To be precise, the dataset contains 489906 instances of events compared to over 4.5 million instances of noise. An effective strategy to handle the class imbalance will need to be devised during training of models to prevent the model from overfitting.

Table 3.1: Description of columns

Column	Data type	Unit	Description
pos_x, pos_y, pos_z	float	meters (m)	The position within the detector where the hit was detected, they represent the x,y,z coordinates of the hit respectively.
time	float	nano seconds (ns)	The time at which the hit was detected.
label	int	NA	The type of hit, '0' represents noise and '1' represents a neutrino hit
event_id	int	NA	The id of the event to which the hit is related to. The id itself does not have any meaning, it is simply used to identify hits that originated from the same event.
timeslice	int	NA	The id of the timeslice to which the hit belongs. The id itself does not have any meaning, it is simply used to group hits into discrete bins.

3.2 Verification of Bias

The *events* dataset is synthetically generated using simulations. As such, it is likely that the event hits in each timeslice may occur at a specific time such as at the beginning, middle or end of the timeslice. Having such a pattern in the dataset may bias the model since it may learn this pattern and thus fail to generalize. If this pattern does exist in the dataset, corrective measures need to be taken such that the event hits in each timeslice are uniformly distributed.

To verify the existence of such patterns in the dataset, the mean `time` of event hits across all events was visualized. Figure 3.3 depicts a scatter plot of mean `time` across events. A uniform distribution is noted with no visible patterns. Thus no bias exists in the dataset and is deemed suitable for further analysis.

3.3 Exploration of Interesting Timeslices

Figure 3.4 represents to total number of hits per timeslice. The dataset is discretized into 6759 timeslices of which 2783 timeslices contain only noise hits. This is corroborated by Figure 3.4 which presents a long tail distribution where many timeslices contain few to no

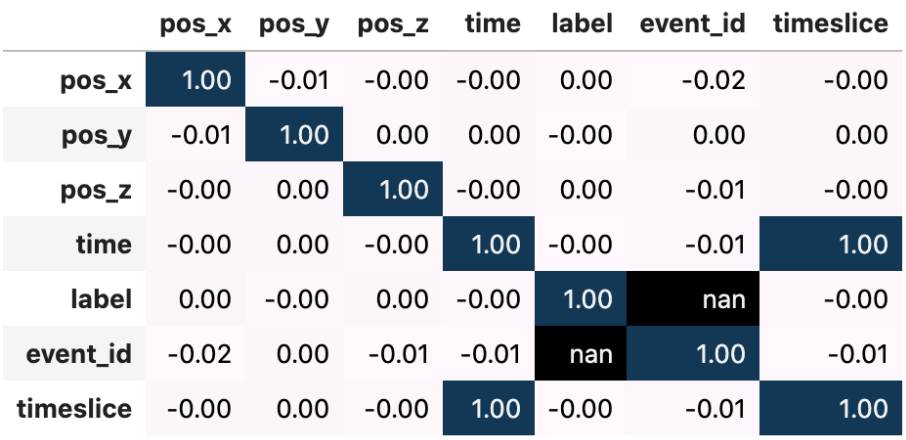

Figure 3.1: Correlation matrix of features

Table 3.2: Descriptive statistics

	pos_x	pos_y	pos_z	time	label	event_id	time slice
count	4.58e+7	4.58e+7	4.58e+7	4.58e+7	4.58e+7	489906	4.58e+7
mean	1.16e-02	-1.59e-02	1.17e+02	5.00e+07	1.06e-02	2862.00	3.33e+03
std	5.12e+01	6.22e+01	4.86e+01	2.89e+07	1.02e-01	1667.61	1.92e+03
min	-9.46e+01	-1.15e+02	3.77e+01	0.00e+00	0.00e+00	0.00	0.00e+00
25%	-4.50e+01	-5.79e+01	7.40e+01	2.50e+07	0.00e+00	1392.25	1.66e+03
50%	1.30e+00	-4.18e+00	1.21e+02	5.00e+07	0.00e+00	2887.00	3.33000e+03
75%	4.04e+01	4.85e+01	1.60e+02	7.50e+07	0.00e+00	4304.75	5.00000e+03
max	9.62e+01	1.05e+02	1.96e+02	1.01e+08	1.00e+00	5734.00	6.77e+03

event hits and few timeslices contain a high number of event hits.

Figure 3.5 depicts a scatter plot of *timeslice 615* which contains the largest number of event hits. It is observed that event hits occur close to each other in space and time (represented by the yellow, blue and green points) whilst background hits are uniformly distributed in space and time (represented by the purple points).

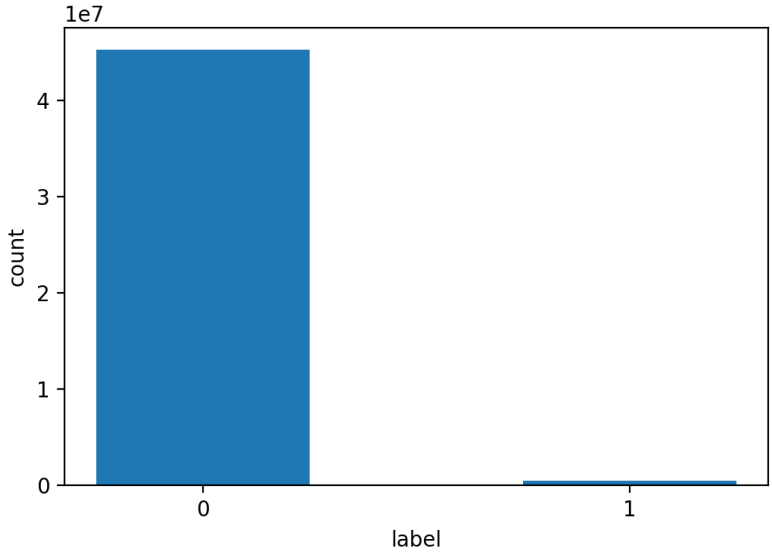

Figure 3.2: Distribution of `label` column

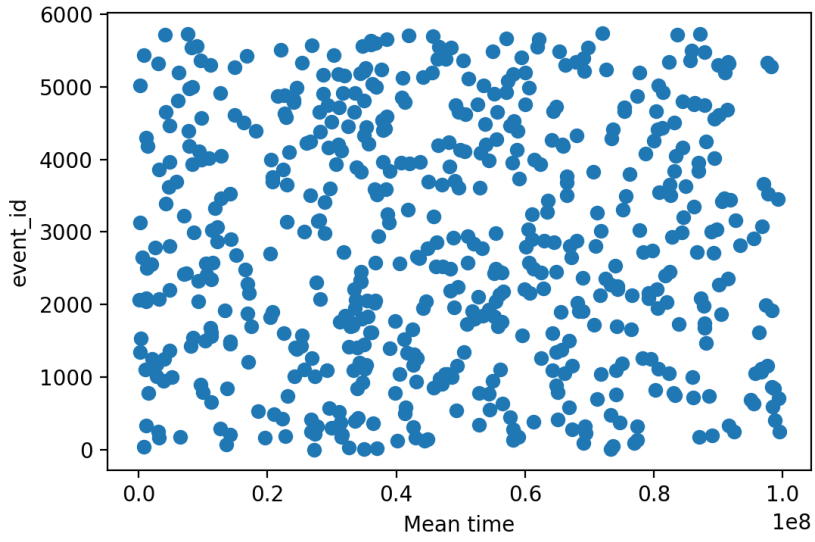

Figure 3.3: Verification of Bias

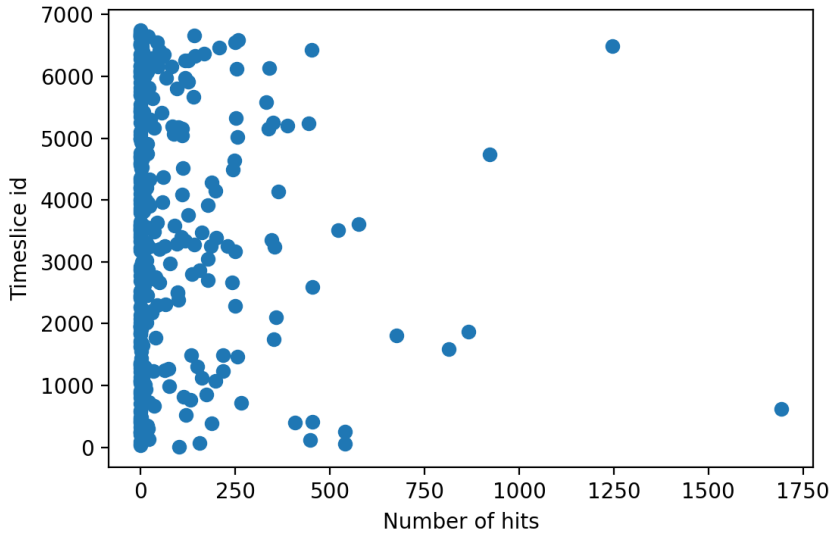

Figure 3.4: Distribution of event hits per timeslice

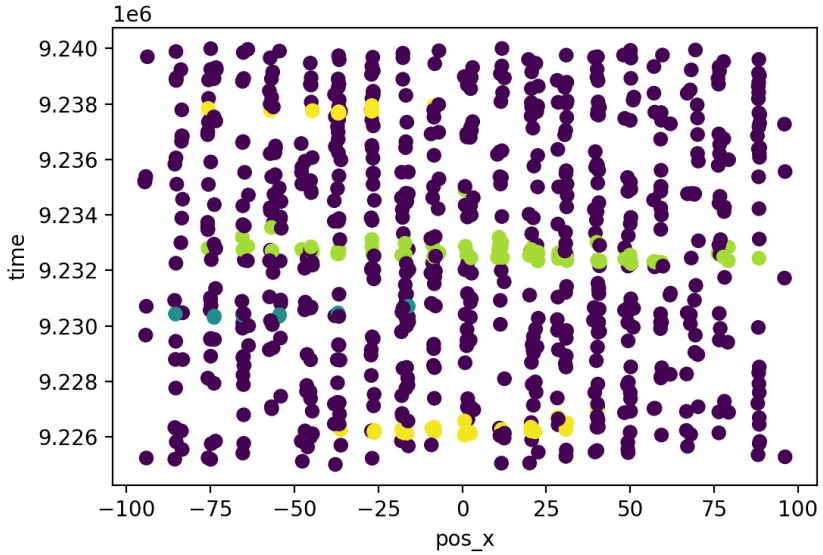

Figure 3.5: Distribution of Timeslice 615

Bibliography

[1] Silvia Adrian-Martinez, M Ageron, F Aharonian, S Aiello, A Albert, F Ameli, E Anassontzis, M Andre, G Androulakis, M Anghinolfi, et al. **Letter of intent for KM3NeT 2.0**. *Journal of Physics G: Nuclear and Particle Physics*, **43**(8):084001, 2016. 1, 3

[2] Sebastiano Aiello, Fabrizio Ameli, Annarita Margiotta, Michel Andre, Giorgos Androulakis, Marco Anghinolfi, Antonio Marinelli, Gisela Anton, Miquel Ardid, Christos Markou, et al. **KM3NeT front-end and readout electronics system: hardware, firmware, and software**. *Journal of Astronomical Telescopes, Instruments, and Systems*, **5**(4):046001, 2019. 2, 3

[3] Annarita Margiotta, KM3NeT Collaboration, et al. **The KM3NeT deep-sea neutrino telescope**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **766**:83–87, 2014. 2

[4] Maarten Post. *"KM3NNeT" A neural network for triggering and classifying raw KM3NeT data*. PhD thesis, Universiteit van Amsterdam, 2019. 2, 3

[5] Konrad Karaś. *Data processing pipeline for the KM3NeT neutrino telescope*. PhD thesis, Universiteit van Amsterdam, 2019. 3

[6] **NumPy - The fundamental package for scientific computing with Python**. 5

[7] **Pandas**. 6