DETAILED REPORT ON THE MASTER THESIS OF CHRISTIAAN GERHARDUS VILJOEN

ON THE TOPIC

Machine Learning for Particle Identification & Deep Generative Models Towards Fast Simulations for the ALICE Transition Radiation Detector at CERN

The presented thesis addresses two very important aspects in the understanding of detector data generated by the Transition Radiation Detector of the ALICE experiment at CERN with the help of machine learning techniques. This is on the one hand the discrimination of different particle species (mainly electrons and pions) based on the amplitude and time structure of the measured TRD signal and on the other hand the artificial generation of the TRD response. The former is an important contribution of the TRD to the measurement of observables to study the creation of a so-called Quark-Gluon Plasma, a new phase of matter where quarks and gluons are no longer confined into hadrons. It existed a few microseconds after the Big Bang at temperatures of more than 10^{12} K. The QGP is created at the LHC in heavy ion collisions for about 10^{-22}, electrons are produced via various processes in the plasma (e.g. via thermal radiation of in-medium decays) and can escape without interacting strongly. In order to properly compare to theoretical predictions of particle production from the QGP, the measured data has to be corrected for the detector response, which can in principle be modelled in simulations. However, the precision and the speed of the generated detector response are crucial, in particular for the increase in data rate in the continuous read out after the ALICE-Upgrade. Here, the methods of Machine Learning and Generative Models are a very promising route for further improvement.

In the following, a detailed justification is given on the various criteria listed in the summary report. Where appropriate, the judgement is supported with references to the full text. A detailed list of comments and suggestions for specific changes is attached as PDF version to this report.

The thesis of Mr. Viljoen makes an important contribution to the above mentioned areas. (**2.1**) He understands the nature, objectives and scientific principles underlying the methods of machine learning, which are described very well and in Chapter 3 of the thesis. In the discussion and description of the physics of the standard model and the quark-gluon plasma he shows in general a good understanding, with occasional misconceptions (e.g. p11 the 3 fermion families as heavier versions of each other at different energy level, while the families would still be different particles with zero mass). (**2.2**) Mr Viljoen is familiar with, and clearly understands, the relevant literature, this is demonstrated e.g. by the literature-based, coherent explanation of the various machine learning techniques in Chapter 3. (**2.3**) The appropriate techniques and analytical methods are mastered very well, in particular in terms of organising and evaluating the results for different methods for particle identification (Chapter 4). (**2.4**) The results and findings are clearly explained, the obtained performance of different architectures is put into context with the choice of number of layers, input nodes and learning parameter. (**2.5**) The findings are assessed in a coherent manner and e.g. in the case of particle identification compared to previous results. Though, an inferior performance is observed in the work, this is clearly explained by the use of uncalibrated data. (**2.6**) Critical and independent thought is well shown at various places in the thesis, exemplarily demonstrated in the discussion and summary of the results, where clear recommendations for future developments are provided. (**2.7**) The report is in an acceptable scientific format, which could be further enhanced by following the suggestions listed in the attached PDF, in particular the use of too small (sub)figures spanning more than one page is disruptive to the reading. In addition, some references need more information (e.g. author and title) to be findable in the future.

Overall the study shows a unique broadness in the search for the optimal use of machine learning methods in particle identification with the ALICE-TRD. The full potential of this approach could not be shown here, due to the use of uncalibrated data. However, the findings are documented, presented and explained such, that they can be easily adapted for future studies. Moreover, in the course of this study, for the first time deep generative models have been used to asses the quality and improve detector simulation of the TRD. In this context, the work can be seen as an important basis and reference for future developments.

In summary, I strongly recommend the dissertation should be passed after some specific changes (as provided in the attached PDF) have been made.

Full report for the Masters thesis of Mr Christiaan Gerhardus Viljoen

Christiaan Gerhardus Viljoen has put together an excellent, thorough and well-researched thesis, covering an interesting intersection between several aspects of machine learning, and particle detection at the LHC.

His writing style is clear and the thesis is relatively self-contained, given thorough sections on both the machine learning side and the particle physics side.

In the first research part of the thesis Christiaan has set up and compared a number of different machine learning algorithms, spanning different types of neural network as well as other classical ML algorithms, showing how well they perform on the data.

In the second part he has built generative models for creating synthetic data and compared this with pre-existing methods to show the similarity with the real data.

Each of these sections is well explained, the motivation is clear, and the method is thoroughly described, meaning that someone without a strong background in ML or particle physics would still have a clear idea about what he is doing.

The results themselves are good, though they are not quite up to the cutting edge, mostly due to the exploratory nature of the research and various constraints which are discussed throughout.

There are a number of small changes to be made, but these are fairly minor:

I believe that the amount of detail in the abstract is too much. The aims are laid out in too much detail. The aims set out in section 1.2 are ideal for the abstract and could be moved there, or they could be removed from the abstract completely.

In section 2.2.1 he talks about the energy scale of QED and then doesn't mention what the energy scale is of the other forces. These should be included.

Section 2.1.2: "At the lowest energy level of the standard model" is ambiguous.

Section 2.2.1: "Specified timeperiod" is ambiguous.

And "statistical mechanics understands matter" does not make sense.

And "as different states of matter bounded by phase boundaries ". How are the states of matter bounded?

Section 2.3.3.1.1.2 "readut"->"readout"

Figure 10: "at the hand of average" what does this mean?

3.3.1.1 "Neural networks are not directly optimised" is ambiguous.

Equation 16: n is not defined

Just below this, “one-hot encode” is not defined

Below equation 18: “Focal loss imply adds” doesn’t make sense.

Section 4.1.1.1 I believe that much of this can be put into an appendix. A general discussion would be ideal at this stage without the very specific files and methods which don’t add a great deal to the overall discussion of the thesis.

Section 4.4: “stronglu”-> strongly

Figure 54 and 55 are too small to read.

6.1 “were ran on”-> “were run on”

Once these changes are made I believe that this will make an excellent Masters thesis.

Should you have any further questions, please don’t hesitate to ask.