Christiaan Gerhardus Viljoen

STUDENT NUMBER: VLJCHR004

christiaan.viljoen@cern.ch

Department of Statistical Sciences

PD Hahn Building

University Avenue North

University of Cape Town

Rondebosch, Cape Town, South Africa

Dr Thomas Dietel

SUPERVISOR

thomas.dietel@cern.ch

Department of Physics

RW James Building

University Avenue

University of Cape Town

Rondebosch, Cape Town, South Africa

Research Proposal

PHY5008W

The Application of Machine Learning Techniques towards the Optimization of High Energy Physics Event Simulations in the ALICE[†] TRD[ϟ] at CERN[‡]

[†] A Large Ion Collider Experiment

[‡] European Organization for Nuclear Research/ Organisation Européenne pour la Recherche Nucléaire

[ϟ] Transition Radiation Detector

Introduction to Research

The Transition Radiation Detector (TRD) is the primary electron identification detector at the ALICE (A Large Ion Collider Experiment) collaboration at CERN (The European Organization for Nuclear Research) (1).

High Energy Physics Event Simulations are an integral part of modern Particle Physics research, and existing Monte Carlo simulation frameworks, such as Geant4, are currently being utilized by experiments like ALICE on a routine basis. These existing frameworks operate at a high level of accuracy, but at an attendant high computational cost (2).

Generative Adversarial Networks (GANs) are an extension of Deep Learning that consists of two neural networks that are pitted against each other in an adversarial mini-max game, where the Generative Neural Network **G**, attempts to maximize the cost function **J** of the Discriminative Neural Network **D**, which is tasked with classifying observations as being "real" (from the actual training distribution) or "fake" (generated by **G**); the two networks can be trained simultaneously via backpropagation (3), to find an optimal solution, where **G** captures the underlying data distribution and **D** outputs 0.5 everywhere (3).

GANs have enjoyed a lot of success in recent years in a variety of applications, such as the verification of document authenticity, image generation from text input and drug discovery (4).

Hypothesis

It is the hypothesis of this dissertation that Generative Machine Learning Algorithms, such as GANs, can be successfully be applied to HEP event simulations, at a lower computational cost than traditional methods currently being used, and that these algorithms can output data which is indistinguishable from actual data collected from the TRD at CERN.

Research Aims and Objectives

1. **To build a highly accurate Neural Network that is able to classify particles passing through the TRD as: electrons, positrons, pions, etc.**
    a. By using h2o.ai (5) within the R statistical software environment (6)
2. **To optimize parameters for Monte Carlo event simulations within Geant4, in order to more accurately account for environmental conditions in the TRD at run-time, e.g. ambient temperature, atmospheric pressure, etc.,**
    a. using an ensembled approach of machine learning (ML) algorithms within h2o.ai (5) ecosystem
3. **To simulate Particle-Detector Interaction data,**
    a. By modelling the output data generated during High Energy Physics Collisions in the ALICE TRD,
    b. that is of sufficient quality so as to be indistinguishable from data generated by current Monte Carlo simulations, such as that generated by Geant4
    c. and that is indistinguishable from data taken from real collision events within the ALICE TRD at CERN
4. **To build an efficient "Proof of Concept" Generative Adversarial Network architecture to this end (point 3.),**
    a. By utilizing existing packages for Deep Learning, e.g. Keras for proof of concept, within the R statistical software environment,
    b. Using data from:
        i. real HEP experiments at ALICE,

1. Which will be obtained from the WLCG storage system using AliEn, and parsed (using AliRoot) into a data format (.csv/ .json) that can be read into R

ii. Simulated event data from Geant4, with parameters tuned to emulate the effect of environmental variables as mentioned in point 2., above.

5. **To explore variational autoencoders (VAEs) as an alternative methodology for event simulations**

6. **To productionalize the most accurate ML simulations of event data (GANs, VAEs, an ensemble of the two, or something completely different)**

a. Either by:

i. reimplementing the chosen algorithm in C++, based upon first principles from linear algebra outlined in the Mathematical Theory section in the Background of this document; and utilizing the existing ROOT package for ML, the Toolkit for Multivariate Data Analysis (TMVA) to support the implementation of this

ii. Interfacing with ROOT from within R, using ROOT R, and setting up a RESTful API service, using the plumbeR package, to minimize additional dependencies at run time

Planned Deliverables

- Masters Dissertation in fulfilment of degree: MSc Data Science
- An accurate and efficient event simulation framework that can be put into production by the ALICE collaboration, without adding additional dependencies to AliROOT
- The publication of results in a Physics Journal (hopefully)

Github Repository

A Github repository containing all files relating to this Masters Dissertation exists at:

https://github.com/PsycheShaman/MSc-thesis

References

1. *The ALICE Transition Radiation Detector: Construction, operation and performance.* **Collaboration, ALICE.** 2018, Nuclear Inst. and Methods in Physics Research, Vol. 881, pp. 88-127.

2. *Generative Models for Fast Cluster Simulations in the TPC for the ALICE Experiment.* **Deja, Kamil, Trzcinski, Tomasz and Graczykowski, Lukasz.**

3. *Generative Adverserial Nets.* **Goodfellow, Ian J, et al.** 2014, stat.ML.

4. **Karazeev, Anton.** Generative Adversarial Networks (GANs): Engine and Applications. *Stats and Bots.* [Online] [Cited: 6 October 2018.] https://blog.statsbot.co/generative-adversarial-networks-gans-engine-and-applications-f96291965b47.

5. **H2O.ai.**

6. **R Core Team.** R: A language and environment for statistical Computing. Vienna : s.n.