Abstract

This Masters thesis outlines the application of machine learning techniques, predominantly deep learning techniques, towards certain aspects of particle physics. Its two main aims: *particle identification* and *high energy physics detector simulations* are pertinent to research avenues pursued by physicists working with the ALICE (A Large Ion Collider Experiment) Transition Radiation Detector (TRD), within the Large Hadron Collder (LHC) at CERN (The European Organization for Nuclear Research).

Aim1: Particle Identification (e vs π): an extensive amount of (fully connected-, 1D and 2D convolutional- and recurrent (LSTM-)) neural networks, as well as two tree-based methods (random forests and gradient boosting machines), were trained and assessed, to determine their ability to discriminate between electrons (e) and pions (π), produced during proton-lead (pPb) collisions conducted at the LHC in 2016, based on raw (uncalibrated) TRD digits data. Particle identification performance was defined by the ability of each model to minimize pion efficiency (ε_π, false positive rate), whilst maintaining an electron efficiency (ε_e, true positive rate) of $\varepsilon_e = 90\%$. A lower bound for the critical region (t_{cut}) in the likelihood-ratio distribution of $P(elec)$ predictions made by a Bayesian combination of probability estimates for up to six independent "tracklet" measurements of a single particle track, which results in $\varepsilon_e \approx 90\%$ was defined, in order to determine the ε_π for that model. The best set of results obtained, per momentum bin, was as follows: $\varepsilon_\pi = 1.2\%$ in the $p \leq 2$ GeV/c range; $\varepsilon_\pi = 1.14\%$ in the 2 GeV/$c < p \leq 3$ GeV/c range; and $\varepsilon_\pi = 1.51\%$ in the 3 GeV/$c < p \leq 4$ GeV/c range. These results are compared against previous work done in this area. Using the focal loss function (a parameterised modification of the more traditionally used binary cross-entropy loss funtion, which down-weights the loss magnitude for well-classified examples), to inform gradient descent via backpropagation, was a crucial step which allowed for the use of the full dataset (which suffers from extreme class imbalances), without having to downsample the dataset to prevent the neural network from becoming biased towards predicting the dominant class. Momentum binning and incremental training per momentum bin perhaps also had some influence on the most successful model achieving a much lower pion efficiency than other models built. An analysis of other distinguishing factors that could determine the relative success across models developed for this specific use case is presented. Note that the main motivation for performing particle identification in this project, was to gain hands-on experience in optimising various types of deep learning models and to develop a better understanding of TRD data. This led into the second (more important) deep generative/ latent variable modeling phase of this project.

Aim 2: High Energy Physics Detector Simulations: Geant4, a Monte Carlo toolkit used to simulate particle interactions with matter (used in conjunction with the AliROOT physics analysis software, developed and used by the ALICE collaboration at CERN), was assessed in terms of how closely the simulated data it produces resembles true data taken by the TRD during collision events; by training a convolutional neural network to classify whether a specific image is from the "real"- or "Geant4 simulated"- data distribution and analysing the results. Distinguishing Geant4 data from real data was a trivial task when compared to the task of particle identification. This is one of the most important results reported in this thesis, given the wide use and general trust that is placed in the quality of Geant4 simulations. If nothing else, it indicates that various parameters used to set-up a Geant simulation can be fine-tuned in future work, in order to minimise the differences between simulations and true data. Furthermore, as a step towards fast simulation, various deep generative modeling strategies were employed to produce simulated data samples which are likely under the observed (true) TRD data distribution. To this end, the following classes of latent variable models were prototyped: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and Adversarial Autoencoders (AAEs). Data produced during these deep generative simulations were then compared to real data, via the same methodology used to compare Geant4-simulated data to real data. In order to assess the feasibility of incorporating these types of models into future high energy physics event simulation software, an investigation was made into how realistic their simulated signals are; and a brief exploration of the latent space of a more successful VAE, was conducted, in order to illustrate the potential interpretability of said latent space. While generative models are exceptionally hard to train, some very promising results (especially for VAEs and AAEs) indicate that they would indeed be worthwhile to pursue, as part of more formal and extended future research; particularly if the latent space for a specific model can be understood well enough to make the deep generative simulation process flexibly manipulable: hopefully to the extent that an exact particle type, traveling at a specific momentum, during a certain LHC run, which is subject to certain environmental parameters, etc. can be specified in advance of running a deep generative detector simulation. This level of control *is* currently possible with Geant4 and therefore the feasibility of using deep generative models for detector simulations will definitely depend on their customisability, interpretability and controlability (as well as, more obviously, their accuracy and speed).