University Of Milano-Bicocca

Physics Department

Master's degree in Physics

Master Thesis

Development and Applications of Graph Neural Networks for the Trigger of the LHCb Experiment

Advisor:
Prof. Marta Calvi
Co-Advisor:
Dr. Julián García Pardiñas

Candidate:
Martina Mozzanica
Student ID:**869374**
phone number:
+39 3461213728

Academic Session 2021–2022

24th March 2023

Summary

The Large Hadron Collider beauty (LHCb) experiment is one of the four primary experiments located along the LHC ring at CERN. This thesis describes contributions done to the development of a novel machine-learning-based algorithm aimed to improve the future LHCb trigger system.
To start, the thesis provides an overview of the physics of the b meson, from its production to its decay, based on the Standard Model. It then discusses the LHCb detector setup and its particle detection process, concluding with a description of the LHCb trigger. The identification of interesting events during data taking bases on the inference of the possible presence of signal beauty- or charm- hadron decays in the event, starting from the set of stable particles detected in the experiment. This requires the thorough analysis of high-dimensional feature spaces in real-time to classify events efficiently, a task that can be improved by the use of novel machine learning methods. In the thesis Neural Networks and deep learning are introduced followed by the description of Graph Neural Networks (GNN) and their applications in particle physics.
The LHCb experiment has completed Upgrade I and is preparing for Upgrade II in a decade's time. These upgrades will increase the instantaneous luminosity by five and ten times, respectively, leading to unprecedented challenges for the trigger system due to an increase in signal and background events. To address this issue, the LHCb DFEI project aims to design a Deep Full Event Interpretation (DFEI) algorithm that uses deep neural networks to process information from all the stable particles in each event. The DFEI algorithm is constructed as a sequence of Graph Neural Networks and its goal is to reconstruct b- and c-hadron decay chains while providing a good level of separation against all the other particles from the rest of the event.
The DFEI algorithm consists of four main steps: edge pre-filtering, node pruning, and edge pruning, followed by a topological lowest common ancestor (LCA) reconstruction where the information in each event is represented in a graph, the nodes representing the particles and the edges representing the relations between each possible pair of particles. The node pruning step uses a GNN to remove particles that do not come from a signal heavy hadron. The edge pruning step also uses a GNN to remove links between particles that are topologically distant and produced by different signal heavy hadrons while the topological LCA reconstruction step uses a multi-class classification on edges to reconstruct the decay chain and produce a

LCA matrix as output. Starting from nearly 10^4 edges and 140 nodes per event, after the node pruning they reduce to nearly 300 and 30, respectively, while after the edge pruning they become nearly 75 and 20.
The DFEI approach is designed for both b- and c-hadron reconstruction, but this thesis focuses solely on b-hadron decays. The original training and performance evaluation of the DFEI algorithm was done using solely a simplified simulation of LHCb events based on the Pythia generator, and an inclusive-b sample. This thesis expands the work in two ways: using the Geant4-based full LHCb simulation in DFEI for the first time, and performing the first detailed study of the DFEI performance on a selection of exclusive key signal modes in LHCb.
Firstly, a comparison of the Fast and Full simulation methods is performed, studying their key differences. The analysis compares the distribution of particles in the decay tree of heavy hadrons with those of from the rest of the event, and it examines the differences between the generator-level distributions and the reconstructed ones.

Later on, the edge-preselection step in DFEI is re-designed and optimised, using the knowledge of the new simulation samples. This pre-selection is a crucial aspect of the study as it helps to eliminate a substantial number of edges connecting particles that do not come from the same B hadron while preserving almost all signal particles' connections. After investigating the effect on different variables, the best selection is obtained with a 2D cut of opening angle between particles' three-momentum or the same primary vertex association. The primary vertex is defined as the point where protons collide. The rest of the DFEI algorithm is subsequently re-trained in the new configuration. The ROC curves study after the training of the node pruning and edge pruning modules gives the necessary thresholds to apply before starting the training of the LCA inference module. Various loss functions, such as cross-entropy loss and focal loss, were used for different trainings.
For performance evaluation, the DFEI output is compared with true values obtained from Monte Carlo simulations, using the perfect-reconstruction efficiency as one of the metrics. The evaluation is done both per signal and per event and to compute performance numbers for exclusive signal modes in the processing environment of DFEI, a new truth-matching algorithm is developed. While for the inclusive Pythia sample the percentage of perfect signal reconstruction is $\sim 1\%$, regarding the exclusive decay modes, the percentage increases, on average, to nearly 15%.
To conclude the thesis describes re-optimisation and re-training of the DFEI algorithm on the full inclusive simulation and the first detailed study of its performance focused on exclusive decay modes. The use of ROC curves and AUC as metrics for evaluating the performance of a classification model in extracting a signal from a larger data set is explained.

Universita' di Milano-Bicocca

Dipartimento di Fisica

Laurea Magistrale in Fisica

Tesi Magistrale

Sviluppi e Applicazioni delle Graph Neural Networks per il Trigger dell'Esperimento LHCb

Relatrice:
Prof. Marta Calvi

Correlatore:
Dr. Julián García Pardiñas

Candidata:
Martina Mozzanica

Matricola: **869374**

numero di telefono:
+39 3461213728

Anno Accademico 2021–2022

24 Marzo 2023

Riassunto

L'esperimento Large Hadron Collider beauty (LHCb) è uno dei quattro esperimenti principali situati lungo l'anello del LHC al CERN. Questa tesi descrive i contributi apportati allo sviluppo di un nuovo algoritmo basato sull'apprendimento automatico volto a migliorare il futuro sistema di trigger LHCb.
In primo luogo, la tesi fornisce una panoramica sulla fisica del mesone b, dalle sua produzione al suo decadimento, a partire dal Modello Standard. Successivamente, si discute della configurazione del rivelatore LHCb e del processo di rilevamento delle particelle, concludendo con una descrizione del trigger di LHCb. L'identificazione degli eventi interessanti durante la raccolta dati si basa sulla possibile presenza di decadimenti di adroni beauty o charm nell'evento, a partire dal gruppo di particelle stabili rilevate nell'esperimento. Questo richiede l'analisi approfondita di spazi di features ad alta dimensionalità in tempo reale per classificare gli eventi in modo efficiente, un compito che può essere migliorato dall'uso di nuovi metodi di apprendimento automatico. Nella tesi vengono introdotte le reti neurali e l'apprendimento profondo (deep), seguiti dalla descrizione delle reti neurali grafiche (GNN) e delle loro applicazioni nella fisica delle particelle.
L'esperimento LHCb ha completato l'Upgrade I e si sta preparando, entro un decennio, per l'Upgrade II. Questi aggiornamenti aumenteranno la luminosità istantanea rispettivamente di cinque e dieci volte, portando a sfide senza precedenti per il sistema di trigger a causa dell'aumento di eventi di segnale e di fondo. Per affrontare questo problema, il progetto LHCb DFEI mira a realizzare un algoritmo di Deep Full Event Interpretation (DFEI) che utilizzi reti neurali profonde per elaborare le informazioni di tutte le particelle stabili in ogni evento. L'algoritmo DFEI è costruito come una sequenza di reti neurali grafiche e ha come obiettivo la ricostruzione delle catene di decadimento degli adroni b e c, fornendo nel contempo un buon livello di separazione rispetto a tutte le altre particelle presenti nell'evento.
L'algoritmo DFEI consiste in quattro fasi principali: pre-filtraggio degli edges, node pruning e edge pruning e ricostruzione topologica dei minimi antenati comuni (LCA), in cui le informazioni di ogni evento sono rappresentate da un grafo, dove nodi rappresentano le particelle e gli edges rappresentano le relazioni tra ogni possibile coppia di particelle. La fase di node pruning utilizza una GNN per rimuovere le particelle che non provengono da un adrone pesante di segnale. La fase di edge pruning utilizza anche una GNN per rimuovere i collegamenti tra particelle topologicamente distanti e prodotte da diversi adroni pesanti di segnale, mentre la fase di ricostruzione topo-

logica dell' LCA utilizza una classificazione multi-classe sugli edges per ricostruire la catena di decadimento e produrre una matrice LCA come output. Partendo da circa 10^4 edges e 140 nodi per evento, dopo il node pruning si riducono a circa 300 e 30 rispettivamente, mentre dopo l' edge pruning diventano circa 75 e 20.
DFEI è progettato sia per la ricostruzione degli adroni b che degli adroni c, ma questa tesi si concentra esclusivamente sui decadimenti degli adroni b. La formazione originale e la valutazione delle prestazioni dell'algoritmo DFEI sono state effettuate utilizzando esclusivamente una simulazione semplificata degli eventi LHCb basata sul generatore Pythia e un campione inclusivo di b. Questa tesi espande il lavoro in due modi: utilizzando per la prima volta la simulazione completa LHCb basata su Geant4 in DFEI e svolgendo il primo studio dettagliato delle prestazioni di DFEI su una selezione di modi di segnale esclusivi chiave in LHCb.
In primo luogo, viene fornito un confronto tra i metodi di simulazione veloce e completa evidenziando le loro principali differenze. L'analisi confronta la distribuzione di particelle nell'albero di decadimento degli adroni pesanti con quelle del resto dell'evento ed esamina le differenze tra le distribuzioni a livello di generazione con quelli ottenuti dalla ricostruzione.
Successivamente, il passaggio di preselezione degli edge in DFEI viene ri-progettato e ottimizzato, utilizzando le conoscenze dei nuovi campioni di simulazione. Questa preselezione è un aspetto cruciale dello studio in quanto aiuta ad eliminare un numero consistente di edge che collegano particelle che non provengono dallo stesso adrone b, preservando allo stesso tempo tutte le connessioni delle particelle di segnale. Dopo aver studiato l'effetto su diverse variabili, la selezione migliore è ottenuta con un taglio 2D sull'angolo di apertura tra il momento delle tre particelle o sull'associazione di vertici primari. Con vertice primario si intende il punto in cui i protoni collidono. Il resto dell'algoritmo DFEI viene successivamente riaddestrato nella nuova configurazione. Lo studio delle curve ROC dopo l'addestramento dei moduli di nodi pruning e edge pruning fornisce le soglie necessarie da applicare prima di avviare il training del modulo LCA. Sono state utilizzate diverse funzioni di perdita (loss), come la cross-entropy loss e la focal loss.
Per la valutazione delle prestazioni, l'output di DFEI viene confrontato con i valori veri ottenuti dalle simulazioni Monte Carlo, utilizzando l'efficienza di ricostruzione perfetta come una delle metriche. La valutazione viene effettuata sia per segnale che per evento e per calcolare le prestazioni per le modalità di decadimento esclusive nell'ambiente di elaborazione di DFEI, viene sviluppato un nuovo algoritmo di truth-matching. Mentre per il campione inclusivo di Pythia la percentuale di ricostruzione perfetta del segnale è di circa 1%, per quelli esclusivi la percentuale aumenta, in media, fino a quasi 15%.
In conclusione, la tesi descrive la riottimizzazione e il nuovo addestramento dell' algoritmo DFEI sulla simulazione inclusiva completa e il primo studio dettagliato delle sue prestazioni focalizzato sulle modalità di decadimento esclusive. Viene spiegato l'uso delle curve ROC e dell'AUC come metriche per valutare le prestazioni di un modello di classificazione nell'estrazione di segnale da un set di dati più ampio.