Using Normalizing Flows for B-jet Energy Corrections with an Outlook on Taylor Coefficient Based Introspection

Lars Sowa

Karlsruhe Institute of Technology

The Compact Muon Solenoid (CMS) detector is located at the Large Hadron Collider (LHC), and records proton-proton collisions at a center of mass energy of 13.6 GeV. In its ongoing Run3, LHC is expected to deliver an integrated luminosity of $250\,\mathrm{fb}^{-1}$. To use this amount of data for new physics analyses, and to reach the full potential of existing analyses, advanced machine learning (ML) methods are inevitable. This work focuses on testing new ML methods within the CMS collaboration and their introspection to provide control of these methods.

Conditional Normalizing Flows (NFs) are ML models that provide bijective mappings between a latent space Z^n and a target space Y^n for a given, conditional input $\vec{x}$. NFs are often used to generate distributions in a target space Y^n. This can be done by random sampling from the latent space, e.g. $Z^n \sim \mathcal{N}^n(0,1)$ and transforming it to Y^n. An application of NFs is the energy correction of b-tagged jets: due to weak interaction processes within these jets, neutrinos can be produced. These cannot be measured by the detector and therefore may lead to a bias in the transverse momentum p_T of the jet.

To tackle this problem, a NF is trained and used to generate the full probability density function (PDF) of the jet p_T. Based on this PDF, an estimate for the corrected jet p_T and a per-jet energy resolution can be determined. The results show that NFs can deliver results of the same quality as the current CMS software framework, which uses a deep neural network (DNN) for this task.

The second part of this work relies on the introspection of neural network architectures. For controlled training and implementation of neural networks, it is crucial to understand precisely how the model utilizes its input features to calculate output values. For this task, the KIT-developed method of Taylor coefficients (TCs) is used to quantify the feature importance: the approach is to compute the Taylor series of the model. The TCs of these terms are then used to quantify the feature importance. Most introspection methods fall short by failing to capture the importance of multiple features combined, but especially these feature combinations can be utilized by modern neural networks in contrast to more traditional machine learning methods like boosted decision trees. However, due to higher order terms of the Taylor series, the TC method is able to catch the importance of multiple features combined.

Based on a dataset of simulated $t\bar{t}$, $t\bar{t}H$ and $t\bar{t}Z$ events, object properties are used as input for a DNN and a graph neural network (GNN) to classify these events. The application of the TC method is implemented for the DNNs and GNNs, but not fully optimized yet. It is planned to do the same for a NF to finally compare the TCs of all three models.