Journal of the Korean Physical Society, Vol. 75, No. 9, November 2019, pp. 652∼659

Quark Gluon Jet Discrimination with Weakly Supervised Learning

Jason Sang Hun Lee,* Sang Man Lee, Yunjae Lee,† Inkyu Park, Ian James Watson‡ and Seungjin Yang
Department of Physics, University of Seoul, Seoul 02504, Korea

(Received 18 June 2019; revised 26 July 2019; accepted 26 August 2019)

Deep learning techniques are currently being investigated for high energy physics experiments, to tackle a wide range of problems, with quark and gluon discrimination becoming a benchmark for new algorithms. One weakness is the traditional reliance on Monte Carlo simulations, which may not be well modelled at the detail required by deep learning algorithms. The weakly supervised learning paradigm gives an alternate route to classification, by using samples with different quark–gluon proportions instead of fully labeled samples. This paradigm has, therefore, huge potential for particle physics classification problems as these weakly supervised learning methods can be applied directly to collision data. In this study, we show that realistically simulated samples of dijet and Z+jet events can be used to discriminate between quark and gluon jets by using weakly supervised learning. We implement and compare the performance of weakly supervised learning for quark–gluon jet classification using three different machine learning methods: the jet image-based convolutional neural network, the particle-based recurrent neural network and and the feature-based boosted decision tree.

PACS numbers: 13.90.+i, 13.87.−a, 12.38.Qk, 13.87.Fh
Keywords: QCD, Jet, Fragmentation, Weakly supervised learning, Machine learning
DOI: 10.3938/jkps.75.652

I. INTRODUCTION

The use of machine learning techniques in high energy physics has been of major interest in recent years due to its potential to improve the analysis of particle collision data. One specific area in which machine learning is used for improvement is the discrimination between quark-initiated and gluon-initiated jets [1, 2]. Though these machine learning techniques show excellent performance, they heavily rely on Monte Carlo (MC) simulations for input, as they are trained on the microscopic details of the simulation, which may not be well-modelled due to the non-perturbative nature of Quantum Chromodynamics (QCD) at low energies. Thus, the performance of these methods can be sub-optimal when applied to real data, and care is needed.

In contrast, weakly supervised paradigms, such as Classification Without Labels (CWoLa) [3] and Learning from Label Proportions (LLP) [4, 5], can alleviate these issues as they can be used as data-driven classifiers. This is done as they allow training using samples that are mixtures of quark and gluon events of different proportions rather than requiring pure, labeled quark and gluon samples. In the CWoLa method, you train a classifier to distinguish between quark-enriched and gluon-enriched samples. Under the condition that the only difference between the two samples are the quark–gluon proportions (and not the features of the quark or gluon in each sample), training a classifier to distinguish the two samples is equivalent to training a classifier to distinguish a quark and from a gluon. The CWoLa technique is beginning to be used in LHC analyses; for example in the CMS $t\bar{t}b\bar{b}$ analysis, it has been used to distinguish the multijet background, which is difficult to model because of the high number of jets [6].

The CWoLa method is simple to apply to any machine learning algorithm as it allows training without any truth information, such as quark–gluon labels or the class proportions of the mixed samples, but uses the same techniques as for standard machine learning with labels. This allows for direct use with dijet and Z+jet samples for quark–gluon jets as they have different quark and gluon fractions [7,8]. A classifier that is trained to distinguish between two mixed samples, which could be made directly from collision data even though we use simulations for this study, is also able to optimally discriminate between the quark jet and the gluon jet processes, in the limit where the only difference between the samples is the quark-gluon jet fraction [3,9].

Three machine learning methods, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Boosted Decision Tree (BDT) are used for weakly supervised learning. The CNN, which is used as an image analysis and classification technique, is able to operate

*E-mail: jason.lee@uos.ac.kr
†E-mail: yunjae.lee@cern.ch
‡E-mail: ian.james.watson@cern.ch

