Jet classification techniques in CMS

Mauro Verzetti [*]
CERN and FWO
Espl. des Particules 1,
1211 Meyrin (CH)
mauro.verzetti@cern.ch

Abstract

Jet flavour identification is of paramount importance for the physics programme of multi-purpose experiments at hadronic colliders. The presence of multiple flavours naturally leads to a multiclass classification problem. In this contribution, we review the latest developments brought by Deep Learning in this field in the CMS Collaboration.

1 Introduction

The Standard Model (SM) [1 – 3] of particle physics is a remarkably successful theory, able to explain a wide variety of phenomena, although cosmological observations point to its incompleteness. The SM is expressed in terms of fundamental forces mediated by gauge bosons and of particles which carry a charge under such forces. The particles can be further divided into two families, quarks and leptons, depending on whether they carry charge under the strong nuclear force (QCD) and therefore are confined or not by such force. Three generations, called *flavours*, of each particle are found in nature, identical in every aspect except the particle mass. Such mass plays a fundamental role in the Higgs mechanism [4 – 9] and is believed to have a fundamental role in the extensions of the SM that are currently probed by the CMS experiment [10] at the LHC.

Due to the confining nature of the QCD, the quarks cannot be observed directly as single particles, but can only be observed in compound states of two or more quarks, known as hadrons. Due to the high energy of the LHC collisions, quarks emitted during the interactions fragment in a collimated spray of hadrons that may subsequently undergo a decay process. The final product of this fragmentation and decay chain, called *jet*, is detected, recorded, and reconstructed by the experiments. The particles produced in the collisions are clustered using the hierarchical clustering algorithm anti-k$_T$ (AK) [11], specifically designed to cluster together particles belonging to the same jet. The anti-k$_T$ algorithmhas one hyperparameter, *R*, which determines the size of the clustering cone, which in CMS is either 0.4 (AK4 jets) or 0.8 (AK8 jets).

Jets originating from heavy-flavour quarks (b and c) fragment in high-momentum heavy flavour hadrons, which fly a significant amount of time before decaying into lighter hadrons. The length travelled within the detector volume, results in displaced tracks and displaced vertices within the jet. The correlation among the track parameters of the different jet constituents is leveraged by the Machine Learning and Deep Learning classifiers to discriminate light jets from heavy flavour jets. The classifiers are trained on a large dataset of simulated collisions, which provide the correct truth labelling for each jet, and their performance is evaluated in several dedicated control regions with real collision data.

[*] On behalf of the CMS Collaboration

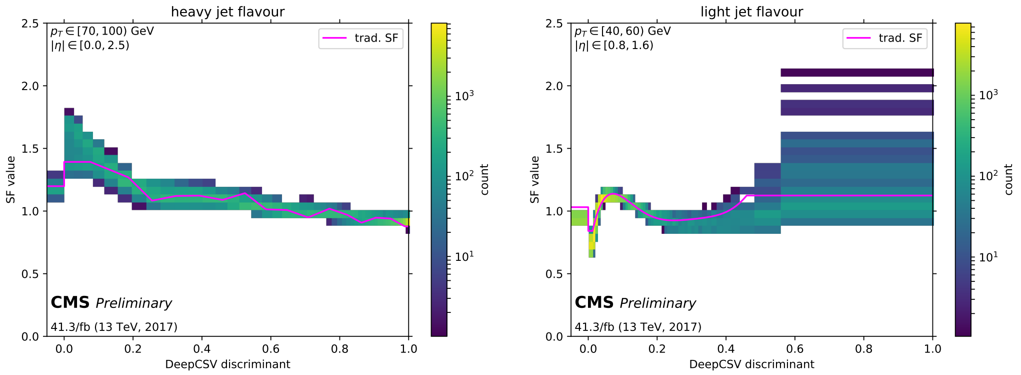

Figure 1: Correction factors for heavy flavour (left) and light flavour (right) jets as a function of the discriminator output computed by the traditional binning method, smoothed by a polynomial fit (purple), and the density distribution of the SFNet output.

2 Performance measurements

Traditionally the performance of a classifier has been tested for a well specified cut on its output (called *working point*) in two dedicated control regions, one enriched with heavy-flavour jets and one enriched with light flavour jets. The calibration procedure provides calibration factors as a function of all the relevant jet features to match the simulated samples to the data. More recently, this concept has been extended by binning the discriminator output and effectively creating many working points. This solution is not completely satisfactory as the control region is statistically limited in several classifier output bins yielding large fluctuations of the correction factors. A solution to this problem is to model the correction-factor function with a DNN, called SFNet [12]. An adversarial approach is used to train the network: large batches of simulated events are fed to the DNN, which output is a correction factor for each jet; the simulated batch and its correction factors are then mixed with a batch of real data in the same phase space and sent to a discriminator network that tries to disentangle the two. As the training process proceeds, the discriminator networks forces the SFNet to learn the proper modelling for the correction factors as a function of the jet inputs.

This new approach has been compared with the more traditional binning-based method showing good closure, as shown in Fig 1.

Work is ongoing to identify the best approaches to quantify the statistical and systematic uncertainties with this method.

3 Small-cone jets classification (AK4)

AK4 Jets are used to cluster the fragmentation product of a single quark decay. Machine learning has been used in the task of classifying the jet flavour since the beginning of the CMS operations in 2010, initially employing simple naive bayesian approaches and slowly growing in complexity as more confidence was gained with these tools. The first Deep Learning-based classifier in CMS, DeepCSV [13], consists of a dense DNN taking as input 8 features of six most displaced tracks in the jet, 8 features from the most displaced secondary vertex, and 12 global variables, many of which are engineered features. DeepCSV has been trained to classify jets originating from light quarks (marked *udsg*), charm quarks (marked *c*), and bottom quarks (marked *b*).

The striking performance improvement achieved by DeepCSV in comparison to the previous algorithms led to the exploration of more modern architectures. The DeepJet [14, 15] classifier uses basic features from all the constituents of the jet, both charged and neutral, from the secondary vertices associated to the jet, and global jet variables. DeepJet takes the features of each jet constituent type (charged, neutral, and SV) and passes them through a set of one-dimensional convolutional layers, to achieve automatic feature engineering, followed by a recurrent layer (LSTM), to provide a constant-size summary of those jet constituents. The output of these summaries are then chained together with the global variables and fed into a dense DNN. A schematic representation of the

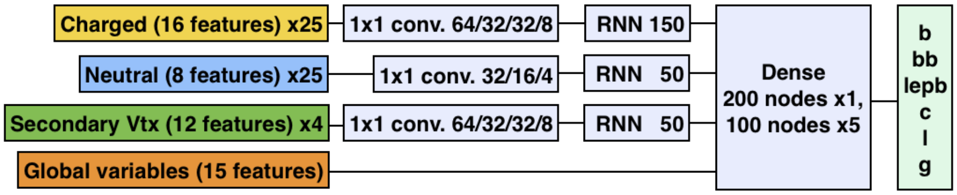

Figure 2: The CMS DeepJet classifier architecture.

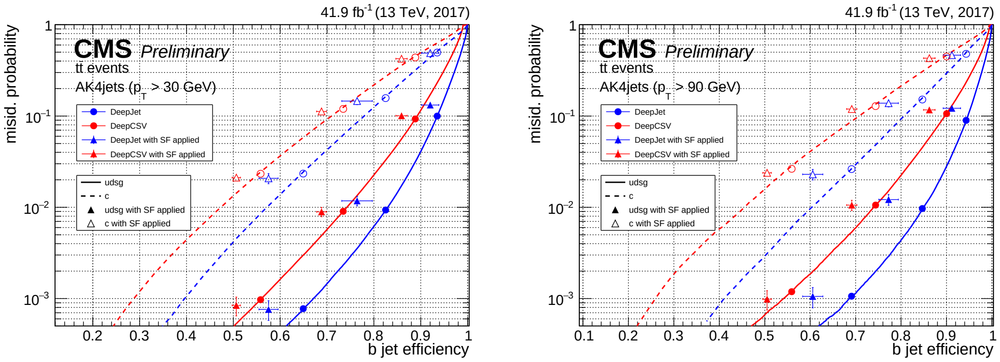

Figure 3: Performance comparison of the CMS DeepCSV and DeepJet algorithms on simulation (solid lines and circular markers for the working points) and on data-corrected simulations (triangle markers, working points only).

DeepJet architecture is shown in Fig. 2. DeepJet is trained to classify jets containing one b-hadon, one leptonically decaying b-hadron, two b-hadrons, one or more charmed hardons, jets originating from light quarks, and jets from gluons.

The DeepJet tagger has been fully commissioned using the collision data recorded by CMS in 2017, computing data-to-simulation performance corrections in dedicated control regions. As it is possible to see in Fig. 3, DeepJet greatly outperforms DeepCSV in simulation, especially at high jet transverse momentum (p_T). The performance gap is still striking in data-corrected simulation, although a possible hint of loss of generalization is visible.

4 Large-cone jets classification (AK8)

Large cone jets are used to cluster multiple jets coming from the decay products of heavy resonances that, due to their large Lorentz boost, are too close to be resolved individually. In this case, both the displacement (and therefore flavour) and the jet shape are important to achieve a good classification power. Another important feature of these classifiers is the decorrelation of their output with respect to the jet mass, which is used in the following data analysis. Out of the several approaches present in CMS, two were developed out of the DeepJet strategy: DeepDoubleX [16] and DeepAK8 [17].

DeepDoubleX are three binary models targeted at classifying the decay of the Higgs boson to a pair of b or c quarks against the background of light jets (BvsL, CvsL, CvsB). These classifiers follow the same structure of DeepJet (convolutions, recurrent, dense), but do not include information from neutral jet constituents. The mass decorrelation is achieved by including a penalty term in the loss function proportional to the Kullback-Leibler divergence of the histograms of the jet mass of the batch and the same histogram with entries weighted by the classifier output. The DeepDoubleX classifiers significantly improve the classification performance with respect to the previous model based on a BDT and engineered high level features (double-b), while showing minimal mass dependence as shown in Fig. 4.

The DeepAK8 classifier, instead, abandons the usage of recurrent units to improve training speed and extends the convolutional kernels to multiple constituents, sorted either on displacement or p_T. The classifier is trained to separate multiple heavy resonances (tops, gauge bosons, Higgs bosons,

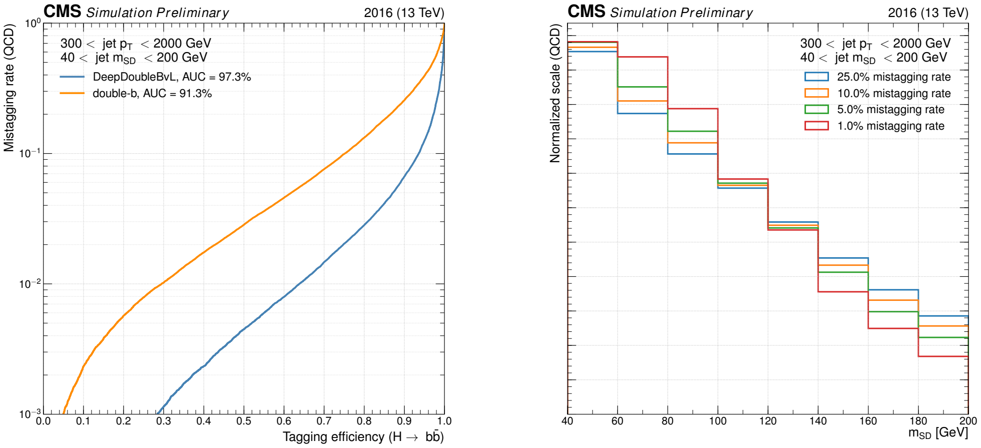

Figure 4: Performance of the DeepDoubleBvsL classifier with respect to the previous double-b (left). Jet mass sculpting induced by different selections cuts on the classifier output (right).

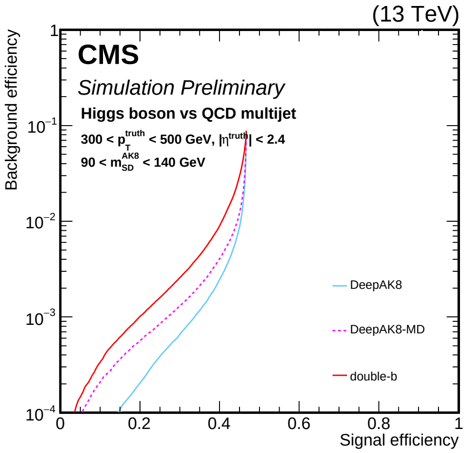

Figure 5: Performance of the DeepAK8 classifier in the Higgs classification task for the mass correlated and the decorrelated version of the model.

light jets) in different final states, up to a total of 17 classes. The mass decorrelation in this case is achieved with an adversarial training with the adversary trying to recover the jet mass from the last layer of the classifier. The performance of both the mass-correlated and the decorrelated classifiers is shown in Fig. 5.

5 Conclusions

Deep Learning has brought a significant boost in performance in the jet classification task at CMS. The research on the subject continues, striving for better, simpler, and faster models. DL is also starting being used as an approximation of unknown arbitrary functions to correct the simulation to match the data. Studies are ongoing in this sector to properly assess the uncertainties of the model and ensure proper coverage.

References

[1] S. Weinberg, “A Model of Leptons”, *Phys. Rev. Lett.* **19** (1967) 1264, *doi:10.1103/PhysRevLett.19.1264*.

[2] A. Salam, “Elementary Paricle Theory”, p. 367. Almqvist and Wiksells, Stockholm, 1968.

[3] A. Salam, “Weak and electromagnetic interactions”, in *Elementary particle physics: relativistic groups and analyticity*, N. Svartholm, ed., p. 367. Almqvist & Wiskell, Stockholm, 1968. Proceedings of the eighth Nobel symposium.

[4] F. Englert and R. Brout, “Broken symmetry and the mass of gauge vector mesons”, *Phys.Rev. Lett.* **13** (1964) 321, *doi:10.1103/PhysRevLett.13.321*.

[5] P. W. Higgs, “Broken symmetries, massless particles and gauge fields”, *Phys. Lett.* **12** (1964) 132, *doi:10.1016/0031-9163(64)91136-9*.

[6] P. W. Higgs, “Broken symmetries and the masses of gauge bosons”, *Phys. Rev. Lett.* **13** (1964) 508, *doi:10.1103/PhysRevLett.13.508*.

[7] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, “Global conservation laws and masslessparticles”, *Phys. Rev. Lett.* **13** (1964) 585, *doi:10.1103/PhysRevLett.13.585*.

[8] P. W. Higgs, “Spontaneous symmetry breakdown without massless bosons”, *Phys. Rev.* **145** (1966) 1156, *doi:10.1103/PhysRev.145.1156*.

[9] T. W. B. Kibble, “Symmetry breaking in non-Abelian gauge theories”, *Phys. Rev.* **155** (1967) 1554, *doi:10.1103/PhysRev.155.1554*.

[10] CMS Collaboration, “The CMS experiment at the CERN LHC”, *JINST* **3** (2008) S08004, *doi:10.1088/1748-0221/3/08/S08004*.

[11] M. Cacciari, G. P. Salam, and G. Soyez, ?The anti-ktjet clustering algorithm?, *JHEP04* (2008) **063**, *doi:10.1088/1126-6708/2008/04/063*, arXiv:0802.1189

[12] CMS Collaboration, “Adversarial Neural Network-based data-simulation corrections for heavy-flavour jet-tagging”, CMS-DP-2018-058, `http://cds.cern.ch/record/2666647`

[13] CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”, *JINST* **13** (2018) no.05, P05011, *doi:10.1088/1748-0221/13/05/P05011*

[14] CMS Collaboration, “CMS Phase 1 heavy flavour identification performance and developments", CMS-DP-2017-013, `http://cds.cern.ch/record/2263802`

[15] CMS Collaboration, “Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13 TeV with Phase 1 CMS detector", CMS-DP-2018-058, `http://cds.cern.ch/record/2646773`

[16] CMS Collaboration, “Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector”, CMS-DP-2018-046, `http://cds.cern.ch/record/2630438`

[17] CMS Collaboration, “Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment”, CMS-PAS-JME-18-002