PROCEEDINGS OF SCIENCE

Deep Learning in Flavour Tagging at the ATLAS experiment

Marie Lanfermann*, **on behalf of the ATLAS Collaboration.**
Université de Genève
E-mail: marie.lanfermann@unige.ch

A novel higher-level flavour tagging algorithm called DL1 has been developed using a neural network at the ATLAS experiment [1] at the CERN Large Hadron Collider (LHC). We have investigated the potential of Deep Learning in flavour tagging using inputs from lower-level taggers. A systematic grid search over architectures and the training hyperparameter space is presented. In this novel neural network approach, the training is performed on multiple output nodes, which provides a highly flexible tagger. The DL1 studies presented show that the obtained neural network improves discrimination against both light-flavour-jets and c-jets, and also provides a better performing c-tagger. The performance for arbitrary background mixtures can be adjusted after the training according to the to the needs of the physics analysis. The resulting DL1 tagger is described and a detailed set of performance plots presented, obtained from simulated $t\bar{t}$ events at $\sqrt{(}s)$=13 TeV and the Run-2 data taking conditions where this tagger will be applied.

EPS-HEP 2017, European Physical Society conference on High Energy Physics
5-12 July 2017
Venice, Italy

*Speaker.

PoS(EPS-HEP2017)764

1. Introduction to ATLAS Flavour Tagging and Motivating Neural Networks

Flavour Tagging in ATLAS targets the tagging of jets originating from a *b*- or *c*-hadron, i.e. b- and *c*-jet tagging. A deep neural network (NN) may be better able to exploit correlations between flavour tagging input variables than the BDT approach currently used for flavour tagging in ATLAS. The application to the separation of *charm* and light-flavour jets is of particular interest in the case of *c*-tagging. Dedicated algorithms which use track and cluster information provide the inputs which are then used by higher level algorithms along kinematic variables. The current baseline for such a higher level algorithm is based on a BDT approach (MV2).

The dedicated algorithms [2] can be grouped into three main categories: Impact Parameter (IP) based, Secondary Vertex (SV) based or muon based. The lower level taggers that are IP based are IP2D, IP3D and RNNIP. IP2D and IP3D provide discriminants build from the log-likelihood ratios, using flavour hypotheses computed from summed track contributions extracted from simulation-derived templates using information from the transverse and longitudinal impact parameter significances respectively. RNNIP is a parallel approach which feeds raw tracks into a Recurrent Neural Network (RNN) and exploits correlations between the tracks. The SV based lower level inputs are SV1 and JetFitter. SV1 reconstructs inclusive secondary vertices. JetFitter exploits the topological structure of weak b- and c-hadron decays inside the jet by approximating the *b*-hadron or *c*-hadron flight path with Primary Vertex, SV and tertiary vertex using a Kalman filter. Jet kinematics and information on muons produced in *b*/*c* decays is also used.

2. DL1 higher level tagger: Design and optimisation procedure

Before the training, to avoid discrimination based on kinematic differences between signal and background, the kinematic 2D (η, p_T) distributions per jet flavour have been reweighted to the kinematics distributions of the *b*-jet distribution. The weights from this reweighting are then used in the backpropagation update. Defaults values of the NN inputs should not disturb the learning process on physical values and are therefore set to the mean of the distribution and flagged with binary check variables.

During training, the Adaptive Momentum (Adam) optimiser was used to minimise the categorical cross-entropy loss. The rectified linear unit was used as activation function for the intermediate layers except for the output layer where the softmax function was used to constrain the prediction range. Combining Maxout and simple fully-connected layers showed to work better than fully connected layers only. Dropout was added for regularisation and to prevent overtraining. Batch-Normalisation was included to minimise the internal covariate shift.

After training, the NN outputs are combined into a single discriminant using a log-likelihood combination, defining the signal and weighting the background contribution. This allows to use the same net for b- as well as for c-tagging and to tune the background rejection after training. The dimensionality reduction uses the log-likelihood ratio of the signal prediction over background prediction ratio using the signal and output nodes in combination with a variable fraction setting the relative background weighting, see equations 2.1 and 2.2 respectively.

$$\mathrm{DL1cf_{c-jets}} = \ln\left(\frac{\mathrm{p_b}}{\mathrm{f_{c-jets}}\cdot\mathrm{p_c} + (1-\mathrm{f_{c-jets}})\cdot\mathrm{p_{light-flavour}}}\right) \tag{2.1}$$

$$\mathrm{DL1bf}_{\mathrm{b-jets}} = \ln\left(\frac{\mathrm{p_c}}{\mathrm{f_{b-jets}}\cdot\mathrm{p_b} + (1-\mathrm{f_{b-jets}})\cdot\mathrm{p_{light-flavour}}}\right) \tag{2.2}$$

Building upon the methods described previously for stabilising and improving the learning process, and preventing overtraining, a systematic grid search has been performed on the number of hidden layers, the number of nodes in the hidden layers, the sequencing of Maxout and Dense layers as well as the learning rate. The optimisation of the training parameters is based on empirical results. During training the loss development is monitored to check for the reduction of the overall loss as well as overtraining.

3. Performance improvements

The DL1 final discriminant allows the composition of the background to be changed by varying $\mathrm{f_{c-jets}}$ in case of b-tagging or $\mathrm{f_{c-jets}}$ in case of b-tagging. By scanning over the range of possible values for these parameters, iso-efficiency curves using a fixed cut provide a figure of merit. The tuning of the final background weighting is done by looking at the dependence of the performance on the kinematics.

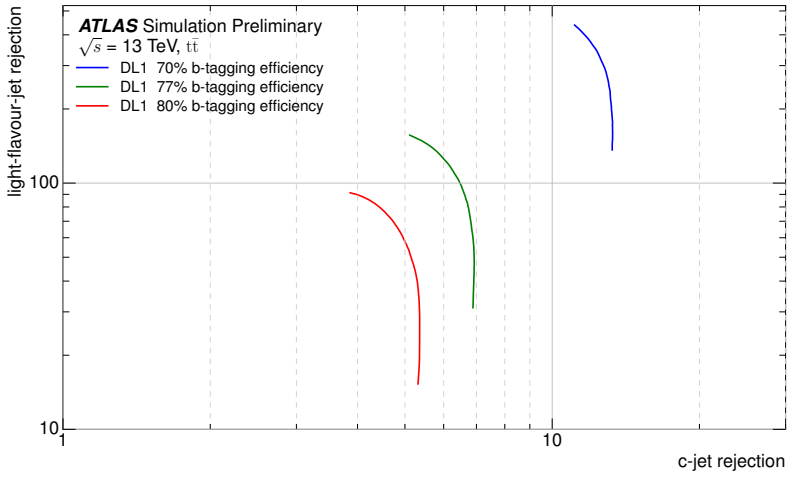

Figure 1: DL1 performance for b-tagging [2].

The *b*-jet tagging performance, see Figure 1, is generally competitive. Keeping e.g. the *b*-jet tagging efficiency fixed at 77% and the light-flavour-jet rejection fixed at 101, the gains when moving from MV2 to DL1 are about 9% with a strong dependence on the underlying p_T spectrum.

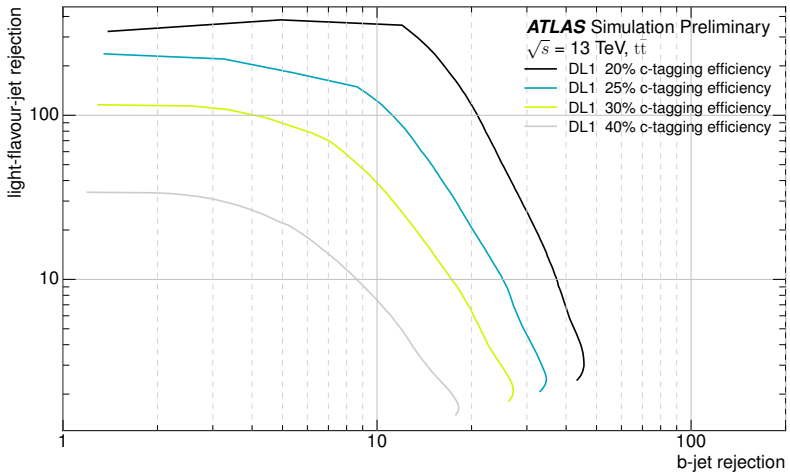

Figure 2: DL1 performance for c-tagging [2].

For c-jet tagging, the improvements, see Figure 2, are significantly larger and a similar dependence of the performance on the transverse momentum is observed. Keeping the c-jet tagging efficiency fixed at 25% and the b-jet rejection fixed at 16, the improvements are about 110%. For 40% c-tagging efficiency and a b-jet rejection of 4, the improvement in light-flavour-jet rejection is about 65%.

4. Data vs Monte Carlo comparison

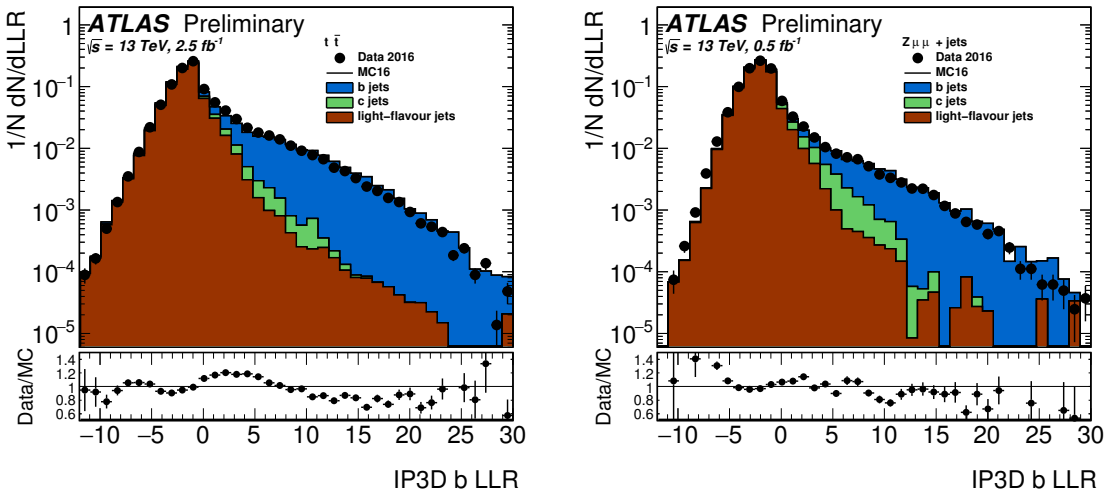

Figure 3: Data-MC comparisons of the log-likelihood ratio used to discriminate the b- from the light-flavour jet hypothesis in the IP3D algorithm using the $t\bar{t}$ and Z sample [2].

The data/MC comparisons of the inputs use a dileptonic $t\bar{t}$ sample with at least one $W \to e\mu$ as well as a $Z \to \mu^+\mu^-$+jets sample. The checks include no systematics evaluation yet and the ratio

plots show only a statistical error. One exemplary variable of the IP3D algorithm, see Figure 4, is found to be well modelled within 30% with some localised differences for low and high values for both of these samples.

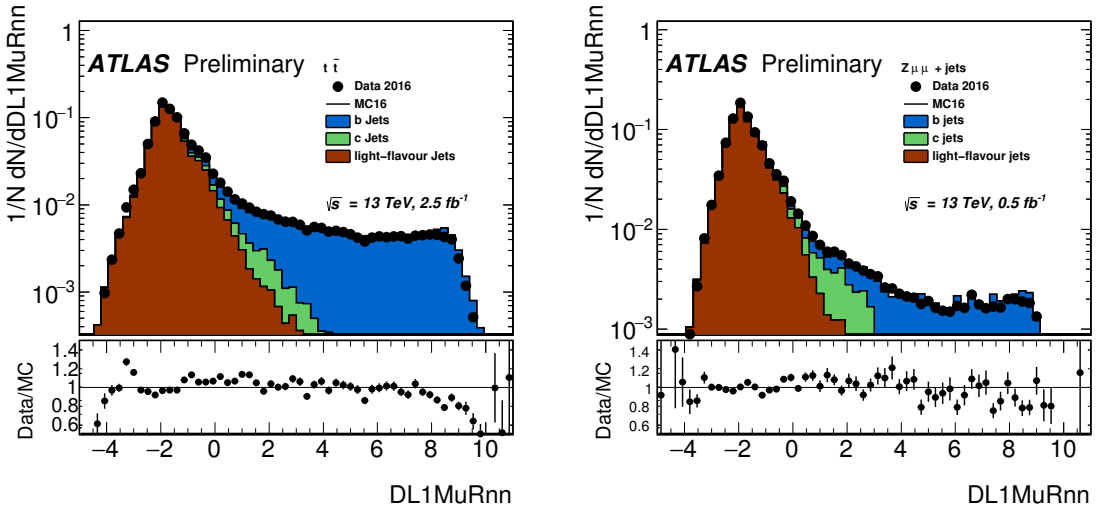

Figure 4: Data/MC comparisons for the DL1 b-tagging using the $t\bar{t}$ and Z sample [2].

The same samples are used to study the Data/MC agreement for DL1. The DL1 final discriminant, see Figure 4, exhibits good separation and the simulation describes the data within 20% with some localised differences for low and high values.

5. Conclusions

A novel flexible higher level tagger has been presented which shows improvements in b- and c-jet tagging and is ready to be used on 2017 data. Validation studies are in progress but preliminary comparisons of MC to data are promising.

References

[1] The ATLAS Experiment at the CERN Large Hadron Collider, ATLAS Collaboration, 2008 JINST 3 S08003.

[2] Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run, ATL-PHYS-PUB-2017-013, http://cds.cern.ch/record/2273281.