PROCEEDINGS OF SCIENCE

Boosted Top Quark Tagging and Polarization Measurement using Machine Learning

Soham Bhattacharya[a], **Monoranjan Guchait**[a] **and Aravind H. Vijay**

[a] *Tata Institute of Fundamental Research, Mumbai, India*

E-mail: soham.bhattacharya@cern.ch, guchait@tifr.res.in, aravindhv10@gmail.com

Machine learning techniques are used to explore the performance of boosted top quark tagging, treating jets as images. Tagging performances are studied in both hadronic and leptonic channels, employing a convolutional neutral network (CNN) based technique along with boosted decision trees (BDT). This computer vision approach is also applied to distinguish between left and right polarized top quarks, and an experimentally measurable asymmetry variable is constructed to estimate the polarization. Results indicates that the CNN based classifier is more sensitive to top quark polarization than the standard kinematic variables. It is observed that the overall tagging performance in the leptonic channel is better than the hadronic case, and the former also serves as a better probe for studying polarization.

40th International Conference on High Energy physics - ICHEP2020
July 28 - August 6, 2020
Prague, Czech Republic (virtual meeting)

PoS(ICHEP2020)318

1 Introduction

Top quarks have a special status in particle physics due to their high mass and their correction to the Higgs mass via loops. Several beyond standard model searches have boosted top quarks as their signatures and hence the study of boosted top quarks is extremely important at the LHC. Another important aspect of the top quark is its polarization state, which can have very interesting implications from different new physics models. In this study [1] we emphasize the top tagging performance in its leptonic decay mode using jet images. We also explore the use of this computer vision approach for differentiating between the polarization states of the top quark.

2 Methodology

Boosted top jets are produced by generating top pair ($pp \rightarrow t\bar{t}$) and W' ($pp \rightarrow W' \rightarrow tb$) events, setting $m_{W'} = 3\ TeV$, and we refer them as $t\bar{t}$ and W' event samples, respectively. Light flavor jets produced in hard QCD events are treated as a background. For our polarization study, left (right) polarized top quarks are produced from the aforementioned W' decay by adjusting the coupling strength $g_R = 0$ ($g_L = 0$) appearing in the W' decay vertex [2]. The $t\bar{t}$ and QCD samples are generated using `PYTHIA8` [3]. In order to access the boosted region of the phase space, a cut of 400 GeV is applied to the traverse momenta (p_T) of the outgoing partons at tree-level. The W' sample is produced interfacing `FeynRules v2.0` [4] in the framework of an effective theory with `MadGraph5_aMC@NLO` [5]. In addition, lighter top squark pair events ($pp \rightarrow \tilde{t}_1\bar{\tilde{t}}_1$) are also generated using `MadGraph5_aMC@NLO`, with top squark mass set to 1 TeV, and forced to decay to a top quark and a lightest neutralino ($\tilde{\chi}_1^0$) of mass 100 GeV. The chirality of the produced top quark can be controlled by appropriately changing the $\tilde{t}_1$-t-$\tilde{\chi}_1^0$ coupling [6]. Every sample is hadronized using `PYTHIA8` and detector effects are simulated using `DELPHES v3.4` [7] with its Compact Muon Solenoid (CMS) card.

Fatjets of radii $R = 1.5$ are reconstructed using the framework of `FastJet v3.2.1` [8] with the anti-k_T [8, 9] jet algorithm and categorized those as a hadronic (leptonic) top jet if the jet axis lies within a cone of $\Delta R = \sqrt{\Delta y^2 + \Delta\phi^2} < 1.0$ around the resultant momentum of the generator-level visible decay products of a hadronically (leptonically) decaying top quark. The fatjets have been cleaned using the soft-drop procedure [10] for $\beta = 0$ and $z_{\text{cut}} = 0.1$ [11]. Jet images are preprocessed following the methodology described in Ref. [12] to aid the network in learning their features. It is to be noted that we separate each jet into its track, photon, and neutral hadron components, and thus three images for each jet (i.e. three input channels) are used to train the network. Fig. 1 shows the images of preprocessed leptonic and hadronic top jets (from the $t\bar{t}$ sample) along with the light flavor QCD jets.

The network architecture used in this study is described in Fig. 2. For the purpose of training the network, we have used the Xavier initialization [13] for the weights and the Adam gradient descent [14] with a batch size of 100 and a learning rate (step size of the gradient descent) of 0.001. We have implemented the aforementioned architecture using the gluon API of `Apache MXNet v1.5.1` [15] in Python.

3 Top tagging

Network trainings are performed using about 1.2M images for each of the signal and background processes corresponding to three sets: (i) hadronic top and QCD jets, (ii) leptonic top and QCD jets, and (iii) leptonic and hadronic top jets. Approximately 135K/135K signal/background jet images are used for the purpose of testing to ensure that the network is not overtrained. We used top jets from $t\bar{t}$ sample for the training. The network is trained for 25

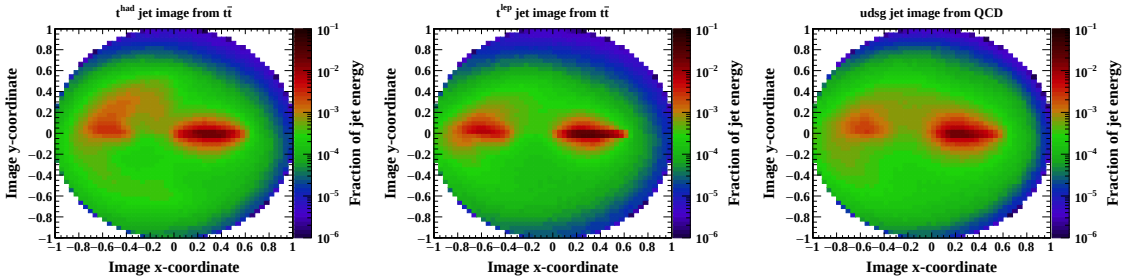

Figure 1: Images of hadronic (left) and leptonic (middle) top jets from $t\bar{t}$ events, and light flavor QCD jets (right). These are the inclusive images of jets where the track, photon and neutral hadron components have been combined.

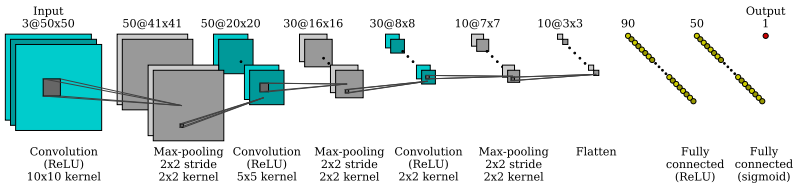

Figure 2: A schematic diagram of the network structure. For any given layer, the text above it indicates the shape of the layer. The shape of a convolution/max-pooling layer (in cyan/gray squares) is represented as channels@N×N. For a fully-connected layer (in yellow circles) it's a single number corresponding to its number of nodes. The text at the bottom indicates the details of the operation performed on the layer above it in order to obtain the next layer. This includes the kernel sizes used for the convolution and the max-pooling operations, along with the activation function (ReLU/sigmoid). This diagram has been generated by adapting the code from https://github.com/gwding/draw_convnet.

epochs, where the training and testing losses are found to saturate to almost identical values. Fig. 3 shows the Receiver Operating Characteristics (ROC) curves to illustrate the hadronic/leptonic top against QCD jet discrimination in solid red/blue (left). We try to further improve the obtained tagging performances for both hadronic and leptonic channel by training a BDT implemented using TMVA [16], where the training and testing samples are the same as that used for the CNN. In this training, apart from CNN classifier, the additional variables used are, (i) mass of the jet (m_j), (ii) the ratios of N-subjettiness variables such as, τ_2/τ_1, τ_3/τ_2 and τ_4/τ_3 [17]. The BDT based performances (labeled CNN+BDT) are presented along with the CNN performances in Fig.3. The robustness of the trainings (both CNN only as well as CNN+BDT) are tested on top jets from the aforementioned W' sample, and the corresponding performances are presented by dashed lines.

4 Top polarization

In this section we study the measurement of boosted top quark polarization using jet images, and then compare the performance with the typical kinematic polarimeter variables [6, 18]. Fig. 4 (upper row), presents the component-inclusive (track + photon + neutral hadron) jet images for left (left) and right (right) handed hadronic top quarks. The corresponding images of leptonic top quarks are shown in the lower row of Fig. 4. The CNN is trained (tested) using about 1M/1M (115K/115K) left/right handed top jet images from the SUSY sample. This training is also evaluated on the W' sample to validate its robustness.

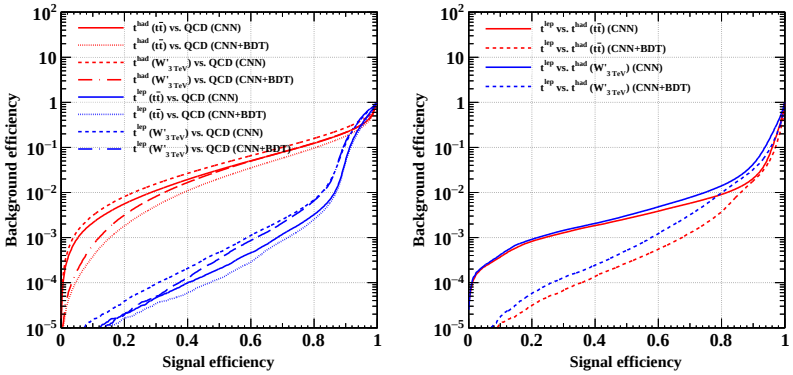

Figure 3: The ROC curves(left) corresponding to the hadronic (leptonic) top versus QCD jet trainings in red (blue). The leptonic top tagging ROC(right) using hadronic top jets as the background.

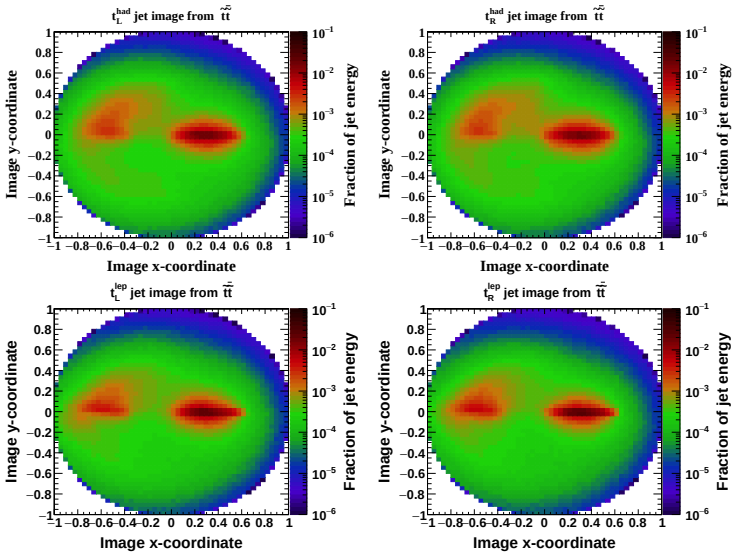

Figure 4: The upper row shows component-inclusive images of left-handed (left) and right-handed (right) hadronic top jets from the SUSY sample. The lower row shows the corresponding images of leptonic top jets from the same sample.

For the hadronic case we compare the performance of the CNN training with that of a robust angular variable, namely $\cos\theta^\star$, that is constructed out of the momenta of the subjets inside the top jet [6]. In case of leptonic tops, we compare the CNN performance with the lepton energy fraction z_ℓ [18].

An experimentally measurable observable, namely the asymmetry, is constructed to measure top quark polarization. It is defined as,

$$A_v^P = \frac{N_{v>c} - N_{v<c}}{N_{v>c} + N_{v<c}}. \tag{1}$$

Here $N_{v>c}$ is the number of top jets subject to the condition that its polarization discriminator v ($\equiv \cos\theta^\star$, z_ℓ or CNN classifier) is greater than a given threshold c, and $N_{v<c}$ is defined similarly. The P refers to the corresponding polarization states of the top jets in a given sample. Note that

we consider only the two extreme compositions (entirely either left or right handed) in this study. The measure of sensitivity of v to the top quark polarization can be presented by $D_v = |A_v^L - A_v^R|$. This difference is expected to be very small if v is not very sensitive to polarization. The optimum value of c for a given discriminator v, is the point at which D_v is maximum. The most sensitive polarization discriminator is decided by comparing the peak values of D_v. The asymmetries and their differences are presented in Fig. 5 for hadronic (left) and leptonic (right) top jets from the SUSY sample. It is evident from the peak value of D_v, that the CNN classifier is ≈2 times more sensitive compared to $\cos\theta^\star$ for hadronic top jets, and ≈1.3 times more sensitive than z_ℓ for the leptonic case.

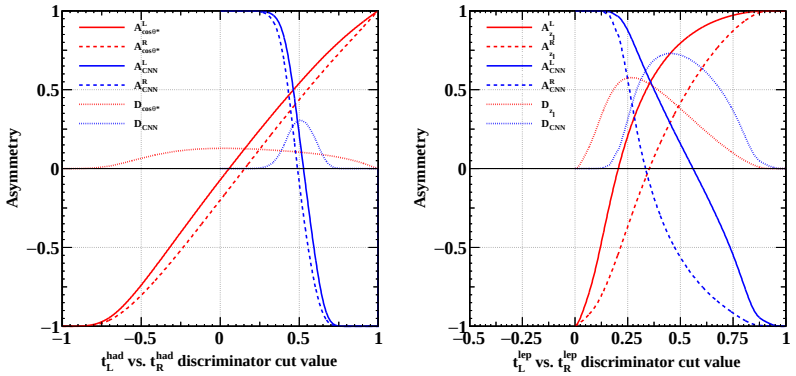

Figure 5: The asymmetry variables (Eq. 1) and their absolute differences between the left and right polarized cases are shown as a function of the discriminator threshold, corresponding to different polarization discriminators, using hadronic (left) and leptonic (right) top jets from the SUSY sample.

5 Summary

The results of boosted top tagging performances in hadronic and leptonic channels using jet images are presented. The leptonic channel is of particular note as this has not been widely studied yet to the best of our knowledge, and we have found the tagging performance of this channel to be significantly better than the hadronic one. An advantage of our methodology is that no lepton identification is required for tagging leptonically decaying top jets. We have also presented the performance of distinguishing between the two polarization states of the top quark using jet images, in both hadronic and leptonic channels. It is observed that the CNN classifier is more sensitive to polarization than the kinematic polarimeter variables like $\cos\theta^\star$ or z_ℓ.

References

[1] S. Bhattacharya, M. Guchait and A.H. Vijay, *Boosted Top Quark Tagging and Polarization Measurement using Machine Learning*, 2010.11778.

[2] Z. Sullivan, *Fully Differential W′ Production and Decay at Next-to-Leading Order in QCD*, *Phys. Rev. D* **66** (2002) 075011 [hep-ph/0207290].

[3] T. Sjöstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [1410.3012].

[4] B. Fuks and R. Ruiz, *A comprehensive framework for studying W' and Z' bosons at hadron colliders with automated jet veto resummation*, *JHEP* **05** (2017) 032 [1701.05263].

[5] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [1405.0301].

[6] R. Godbole, M. Guchait, C.K. Khosa, J. Lahiri, S. Sharma and A.H. Vijay, *Boosted Top quark polarization*, *Phys. Rev. D* **100** (2019) 056010 [1902.08096].

[7] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [1307.6346].

[8] M. Cacciari, G.P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J. C* **72** (2012) 1896 [1111.6097].

[9] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_T jet clustering algorithm*, *JHEP* **04** (2008) 063 [0802.1189].

[10] A.J. Larkoski, S. Marzani, G. Soyez and J. Thaler, *Soft Drop*, *JHEP* **05** (2014) 146 [1402.2657].

[11] CMS collaboration, *Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques*, *JINST* **15** (2020) P06005 [2004.08262].

[12] T.S. Roy and A.H. Vijay, *A robust anomaly finder based on autoencoders*, 3, 2019.

[13] X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y.W. Teh and M. Titterington, eds., vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 13–15 May, 2010, http://proceedings.mlr.press/v9/glorot10a.html.

[14] D.P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014.

[15] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang et al., *Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems*, *CoRR* (2015) [1512.01274].

[16] A. Hocker et al., *TMVA - Toolkit for Multivariate Data Analysis*, 3, 2007.

[17] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) 015 [1011.2268].

[18] R.M. Godbole, S.D. Rindani and R.K. Singh, *Lepton distribution as a probe of new physics in production and decay of the t quark and its polarization*, *JHEP* **12** (2006) 021 [hep-ph/0605100].

PoS(ICHEP2020)318