Tagging leptonically decaying boosted top quarks using machine learning techniques

DESY Summer Student Programme, 2021

Weronika Aleksandra Stanek
AGH University of Science and Technology

Supervisors
Soham Bhattacharya, Isabell Melzer-Pellmann

September 8, 2021

Abstract

Tagging leptonically decaying boosted top quarks have not been much explored using jet image based techniques, as of yet. In this project, we investigate the use of Machine Learning techniques (based on Convolution Neural Network, called ResNeXt50) to tag leptonically decaying boosted top quarks in CMS. The results of classification of the leptonic top jets against the background look promising, for instance the misclassification probability of 10^{-2} ($\sim 2 \cdot 10^{-4}$) is achieved for hadronic top quark jets (light quark and gluon jets) as the background. This can have very interesting implications for a variety of new physics models which have boosted top quarks in their final state.

Contents

1 Introduction **2**

2 Methodology **2**

3 Results **6**

4 Conclusions **11**

---

1 Introduction

Top quark plays an important role in the Standard Model of Particle Physics, for instance it is the heaviest particle and has large contribution to the Higgs mass loop correction. Furthermore, several promising Beyond Standard Model (BSM) physics models have very energetic top quarks in their final states. So far, low and moderate energy phase spaces of various BSM models are excluded by ATLAS and CMS searches [1].

Very heavy BSM particles (with mass ∼TeV) can produce highly boosted top quarks, which decay products are highly collimated and as a result, produce a single „fat" jet, instead of multiple jets. Hence, there is a great need to explore the tagging of boosted top quarks, especially using various Machine Learning (ML) techniques, which proven to be much more effective in classification of signal and background categories, than normal cut based analyses [2, 3]. Boosted hadronic top quark tagging have been well studied using various ML techniques but there is still a demand for a thorough studies on boosted leptonic top quark tagging, which has not been well explored yet. Analysing highly boosted cases can be quite challenging, due to overlapping of the b quark and lepton [1]. In this project we employ Convolutional Neural Network (CNN) to tag leptonically decaying boosted top quarks, using their jet images.

The report is organized as follows. The methodology with the describtion of used ML techniques is disscussed in Sec. 2, results of the trainings are included in Sec. 3, and finally the summary and further outlook are presented in Sec. 4.

2 Methodology

Top quark jets are generated from boosted top-pair events (H_T binned samples) and from $Z' \to t\bar{t}$ events, where Z' masses are equal to 1, 1.25, 1.5, 2 and 3 TeV. Dataset also includes light quark and gluon jets ($udsg$ jets), which are generated from QCD events (p_T binned sample).

For the jet reconstruction anti-k_T [4, 5] jet algorithm is used. The radius parameter is adjusted to $R = 1.5$ (hereon we refer those jets as AK15). If the jet axis is inside the cone of $\Delta R = \sqrt{\Delta y^2 + \Delta\phi^2} < 1.0$, around a resultant momentum of the generator-level visible decay products of a leptonically (hadronically) decaying top quark, the reconstructed fat-jets are further classified as a leptonic or a hadronic top. Here, Δy and $\Delta\phi$ refer to rapidity and azimuthal angle differences, respectively. Cuts applied on jets' transverse momenta and pseudorapidity are $p_T > 200$ GeV and $|\eta| < 2.4$.

In order to improve the classification performance, jet images are firstly preprocessed using a certain transformations, which are described in [6] in detail. Following the methodology, momentum of the jet is rescaled and boosted to a frame where its new mass is fixed to a specific value m_B and the energy is a costant value E_B. As a result, the boost factor given as $\gamma_B = E_B/m_B$ does not depend on a initial momentum of a jet. In this study, we assume $\gamma_B = 2$, ensuring that the model is applicable to a wide range of top quark energies and the jet images are uniform and similar. In order to have an image plane that is perpendicular to the jet axis and ensure that the two subjets with highest energies lie along the x-axis of a image plane, a Gram-Schmidt transformation is used.

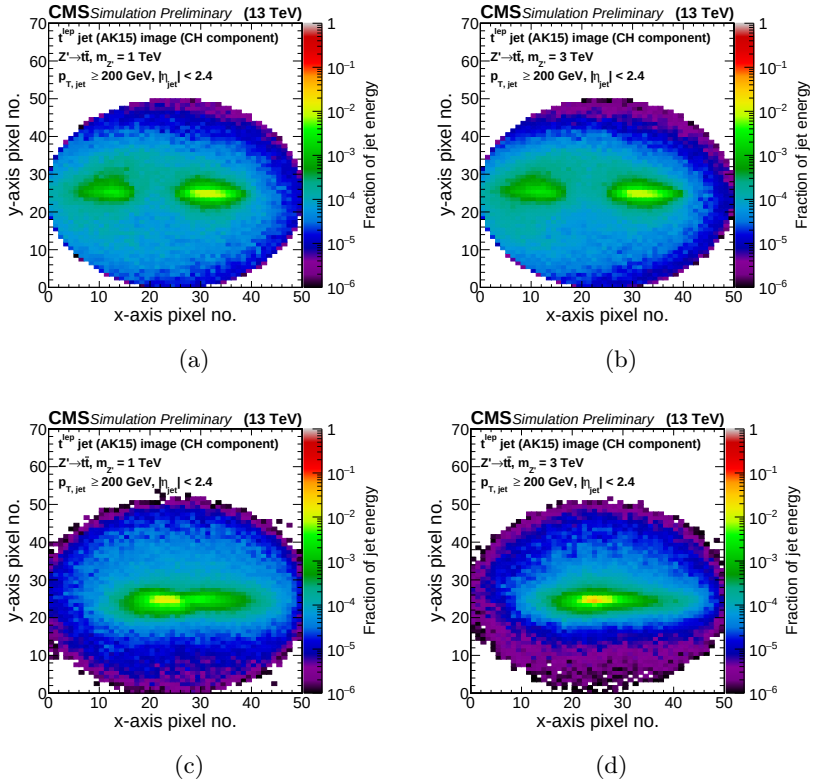


Figure 1: Images of leptonic top jets from $t\bar{t}$ events with two selected Z' masses and their comparison with the usual $p_T - \eta$ rotation (on the bottom).

The applied procedure makes the results more reliable across a wide range of energy. An example of leptonic top jets from $t\bar{t}$ events and their comparison with the $p_T - \eta$ rotation are presented in Fig. 1. Generally, each image (layer) is a 50 x 50 histogram. The quantity on the color axis depends on the type of layer being considered. In this study, energy fraction layers and secondary vertex layers are used. The fomer corresponds to jet energy for each type of jet constituents: electron candidate, muon candidate, photon candidate, charged and neutral hadron candidate. The latter refers to the transverse impact parameter (normalized) of a charged constituent track from the nearest secondary vertex within the jet and takes into account only charged constituent types [1, 6]. The selected energy fraction layer images for four exemplary components (electron, gamma and hadrons) are presented in Fig. 2 on page 4.

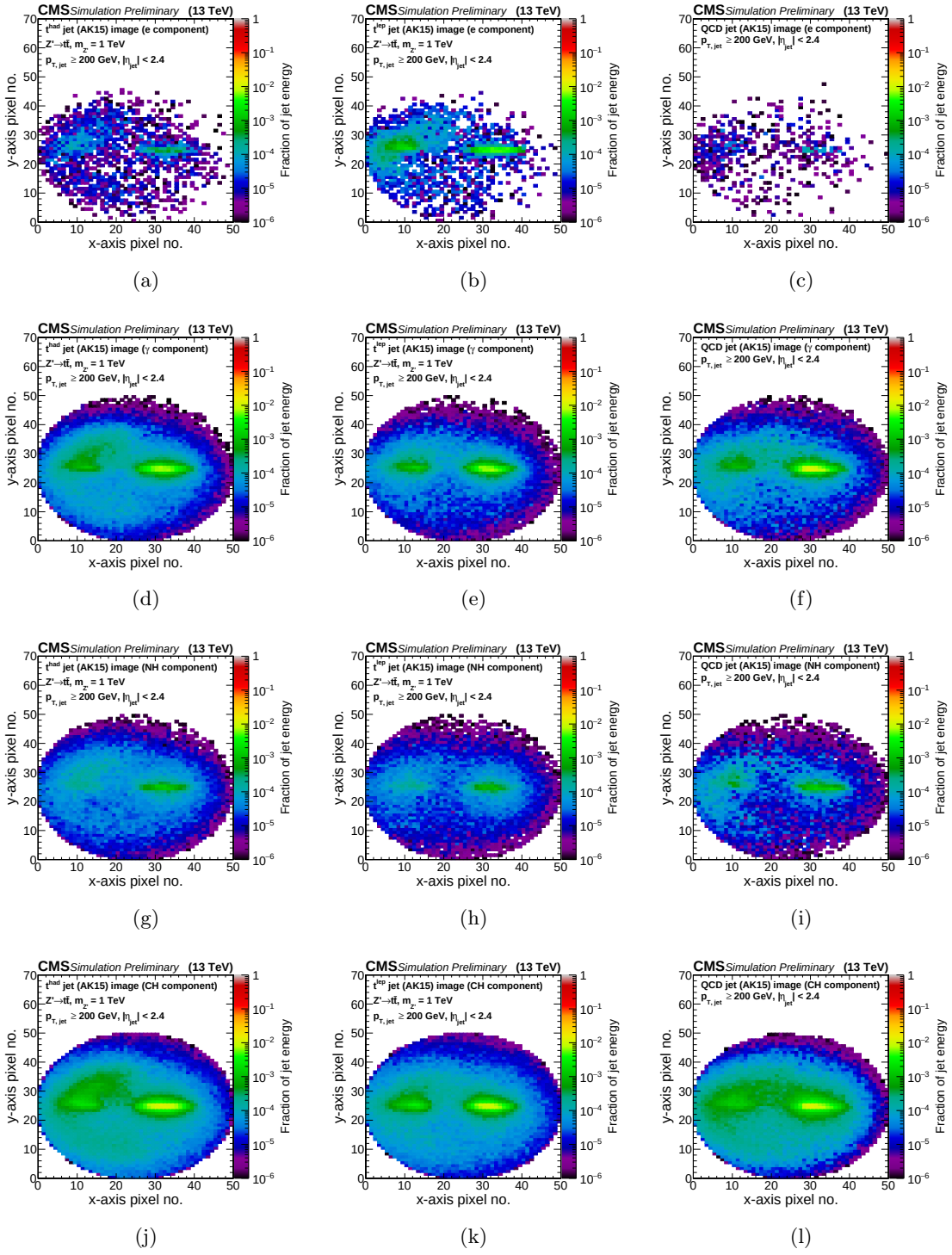


Figure 2: In the left column: images of hadronic top jets, in the middle: leptonic top jets, in the right column: QCD jets. Figures a) - c) correspond to the electron component, d) - f): gamma component, g) - i): neutral hadron component, j) - l): charge hadron component.

In our model three types (categories) of jets are considered: (i) leptonic top jets, (ii) hadronic top jets, (iii) light quark and gluon jets (*udsg* jets). The first class is our signal and the rest constitute to a two types of the background. The number of jets of each category used for training and validation is listed in table 1.

Table 1: Statistics used for training and validation.

category	training	validation
0: leptonic top	2.1M	480K
1: hadronic top	2.9M	540K
2: QCD	3.0M	790K

The used network is based on the class of techniques called Convolution Neural Networks (CNN). In CNN routines, for classifying M×N images, one need to reduce the M×N dimensional space to one (or a few) dimension. Typically, it involves a series of opearations, namely convolution and subsampling. Within a convolution step, the image is convolved with a certain matrix (called a kernel) to extract certain features. During CNN training, kernels are updated after every iteration (epoch). As a result of subsampling (called also as downsampling), the number of dimensions of the feature map is reduced [7].

In this study, we propose one of the modern CNN techniques called Aggregated Residual Neural Network (ResNeXt50) [8]. It is a straightforward, highly modularized network architecture, used for image classification. The main idea is to repeat the building block that aggregates a set of transformations with the same topology. As a result, a homogeneous, multi-branch architecture is obtained, with only a few hyperparameters to set. ResNeXt50 routines requires an implementation of a new dimension called „cardinality", which is the size of the set of transformations. Apart from dimensions of depth and width, it is a essential factor. The quantity that is minimized by the network is the cross-entropy loss. Architecture in three equivalent forms is presented in Fig. 3 [7, 8].

Output of the network is a multiclassifier with three nodes, corresponding to the three classes of jets. The full network is designed for larger images and about 1 000 categories. Since our problem is much simpler, a scaled down version of the network has been used with about 1.5M trainable parameters instead of the full ∼25M parameter network, described in [8]. Network implementation was adapted from [9]. TensorFlow v2.5 with Keras backend was used, implemented in Python. Moreover, the implementation was done using Adam [10] gradient descent with a lerning rate (LR) of xyz value and a batch size of 2K.

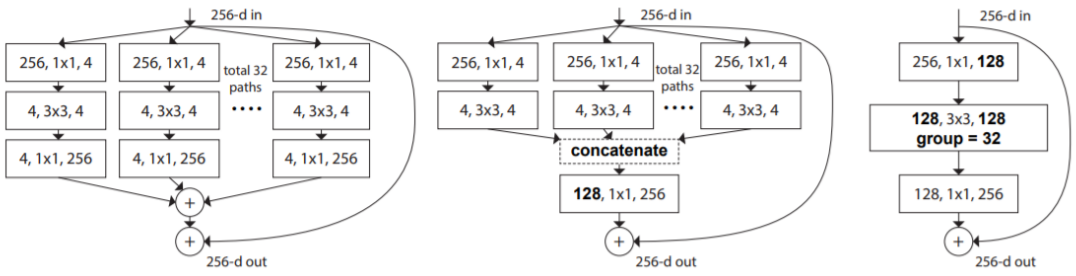
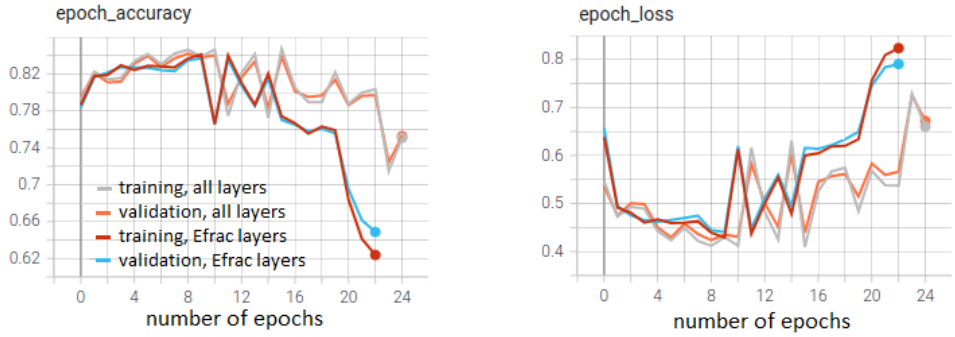


Figure 3: Equivalent building blocks of ResNeXt. Bolded text highlight the reformulation changes. A layer is denoted as (input channels, filter size, output channels) [8].

3 Results

The training was run with ResNeXt50 network, using two versions – the first one included energy fraction layers and the second one included both energy fraction and secondary vertex layers. Since ResNeXt50 performance is better with smaller learning rates, it was accordingly adjusted and equal 10^{-4}. The results of the training and validation for the first 24 epochs are presented in Fig. 4 as loss and accuracy curves. There is no significant overtraining but both models become unstable after epoch number 9. At that step the values of epoch accuracy and loss take the best values equal 0.84 and 0.44, respectively. As far as batch functions are concerned, validation results for both models have significant fluctuations after 10K. The fluctuation after epoch 9 can likely be mitigated by switching to a lower LR at that point, however the result is not likely to improve further by doing so as it has been studied that the batch loss and accuracy reach a plateau by epoch 9.



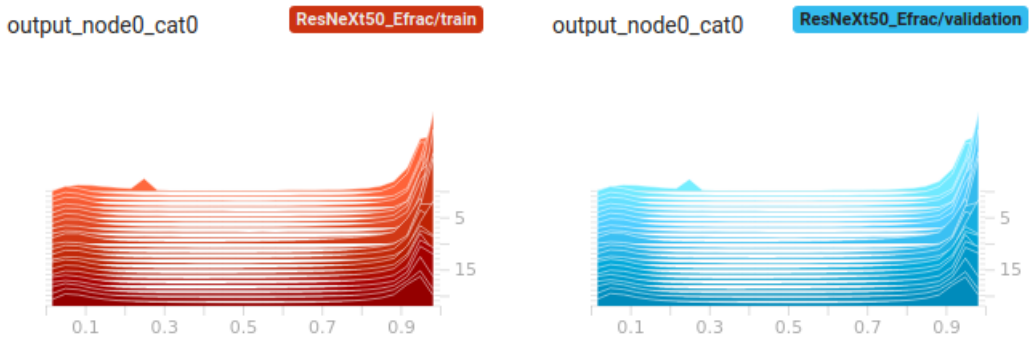
(a) (b)

Figure 4: a): Epoch accuracy and b): epoch loss. Grey and orange curves denote training and validation for all layers (including secondary vertex layers), red and blue curve denote training and validation for energy fraction layers only.

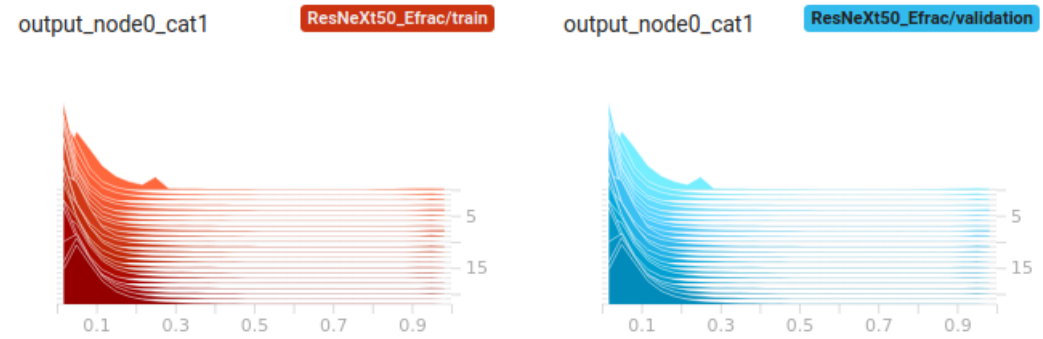
Distributions of the output at node 0 (corresponding to the class 0, which is the signal) for three categories are presented in Fig. 5 and 6 on pages 7 and 8. The y-axis corresponds to the number of the following epochs. As expected, for category 0 the functions are close to 1, while for the categories 1 and 2 the output gets close to 0. This clearly indicates that the classifier assigns the jets to the correct cattegories.

Receiver operating characteristic (ROC curves) at epoch number 9 of category 0 versus category 1 are presented in Fig. 7 on page 9. The efficiency of classification is significantly better in the region of low efficiences and the performance is very similar for both all layers and only energy fraction layers. For efficiency of classification of category 0 of 60%, the reduction of background is at the level of 10^{-2}.

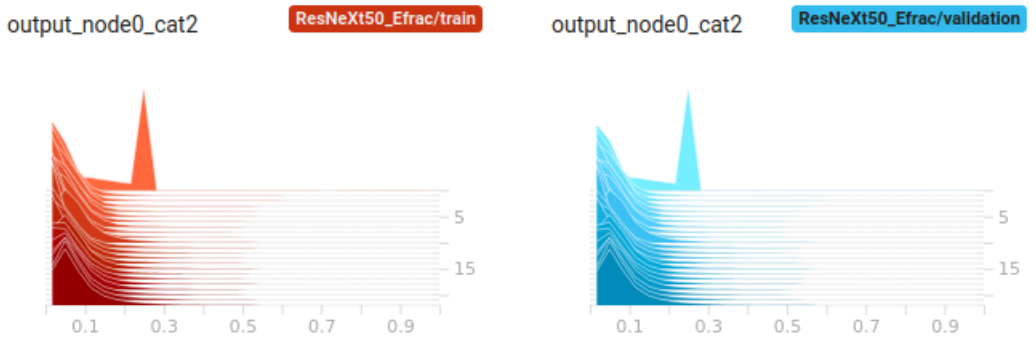
ROC curves at epoch number 9 of category 0 versus category 2 are presented in Fig. 8 on page 10. The performance is slightly worse, for instance for the efficiency of classification of category 0 of 60%, the reduction of the background is at the level of $\sim 2 \cdot 10^{-4}$. Both models give similar results of classification.



(a)

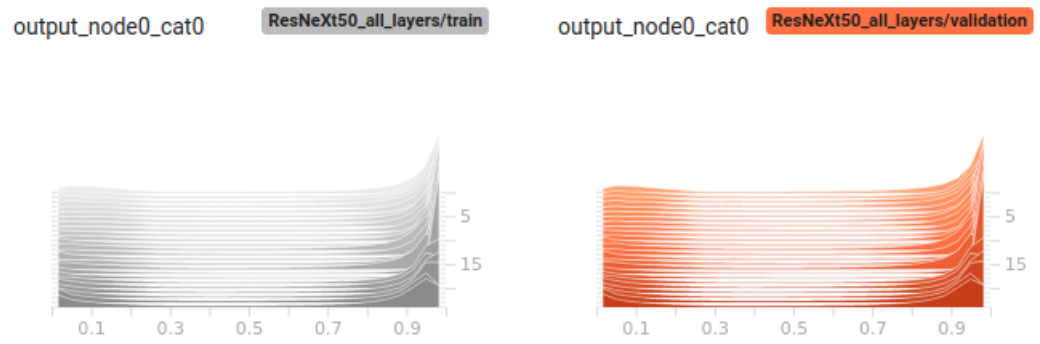


(b)

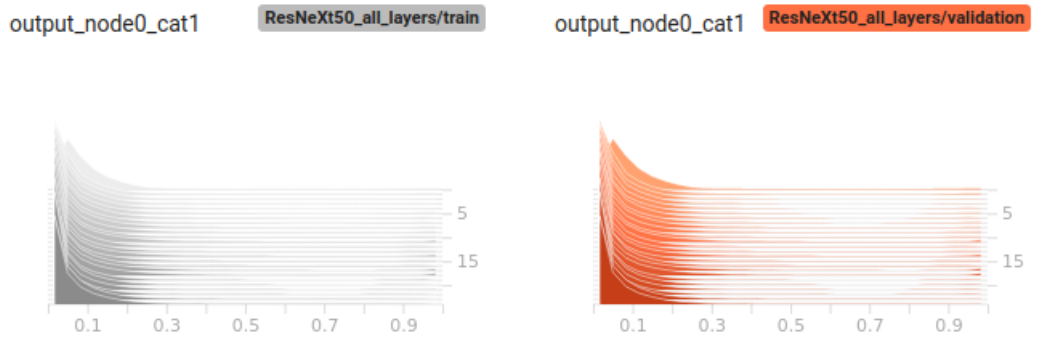


(c)

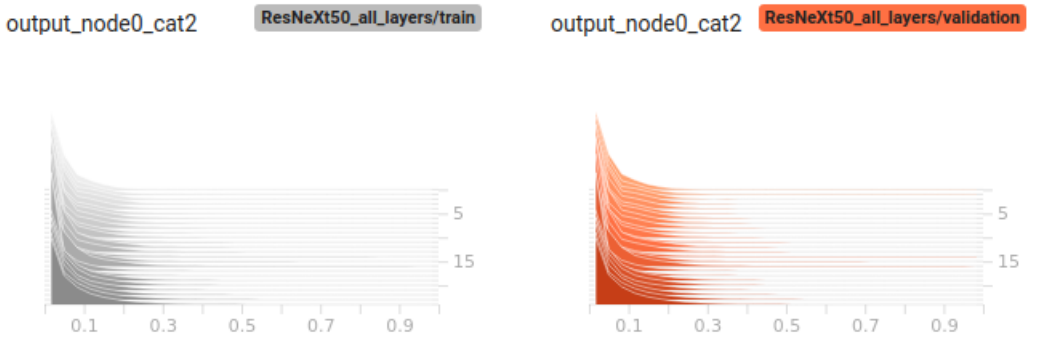
Figure 5: Distribution of the output for node 0 (which refers to the signal), for energy fraction layers. On the left there are results of the training, on the right – results of validation. a): classification of category 0, b): classification of category 1, c): classification of category 2.



(a)

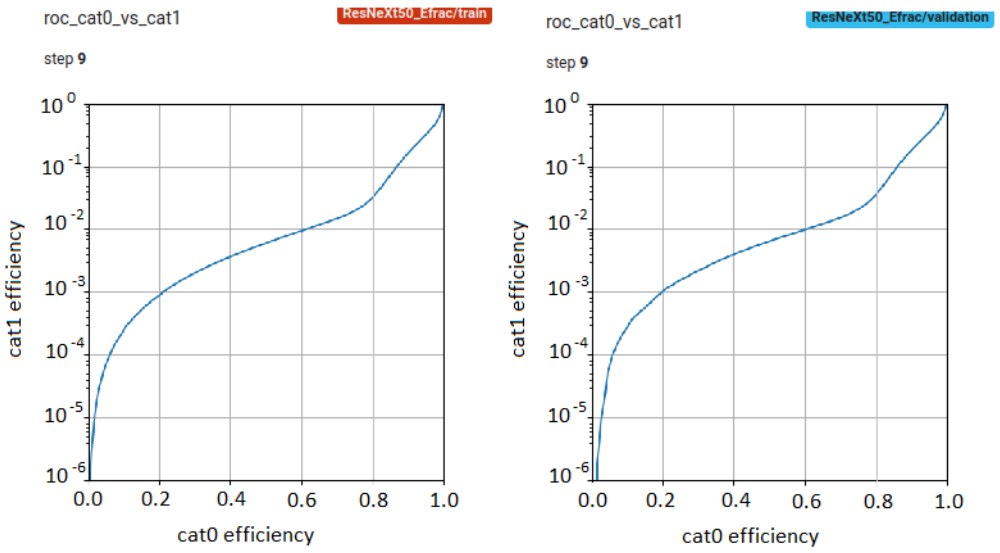


(b)

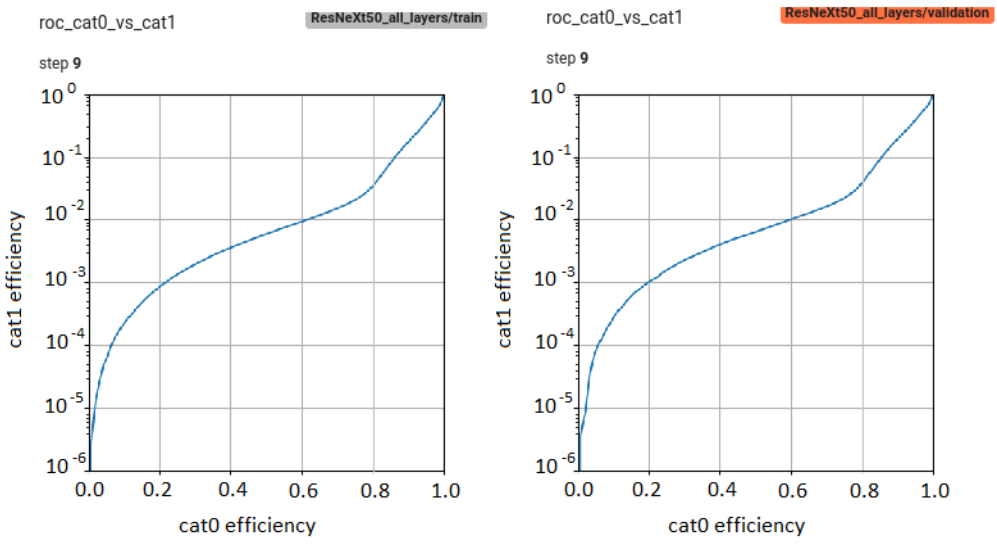


(c)

Figure 6: Distribution of the output for node 0 (which refers to the signal), for all layers. On the left there are results of the training, on the right – results of validation. a): classification of category 0, b): classification of category 1, c): classification of category 2.

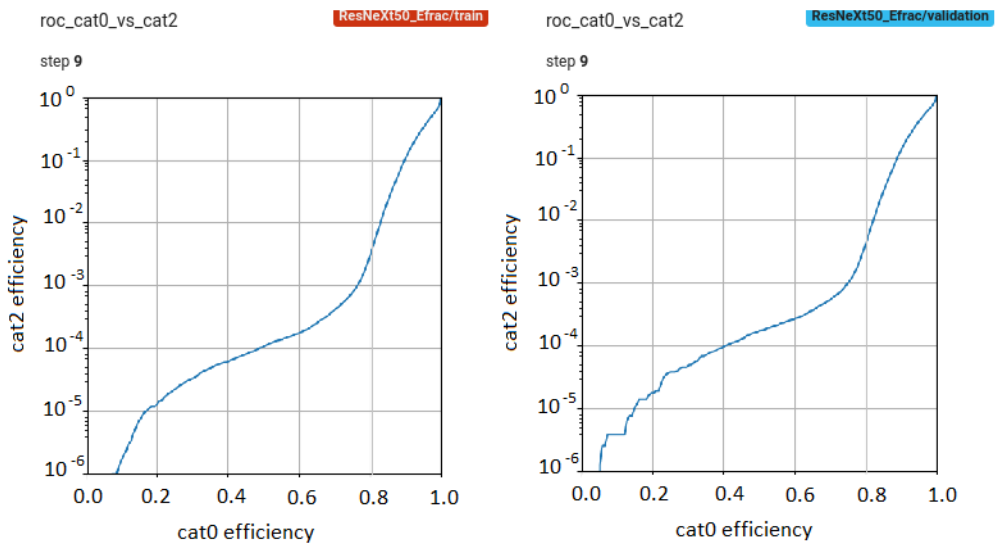


(a)

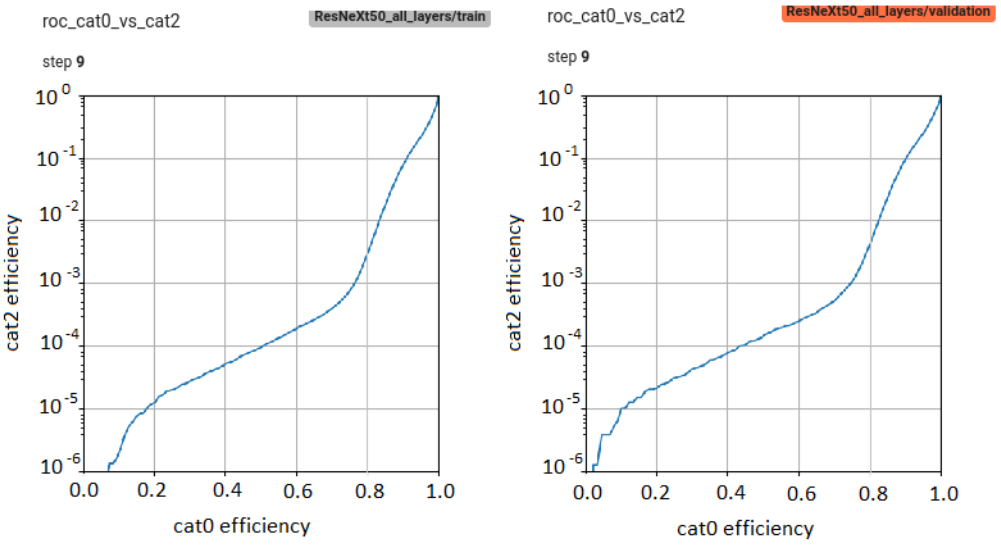


(b)

Figure 7: Receiver operating characteristic (ROC curves) for category 0 against category 1 for a): energy fraction layers and b): all layers. Results of training are presented on the left, results of validation – on the right.



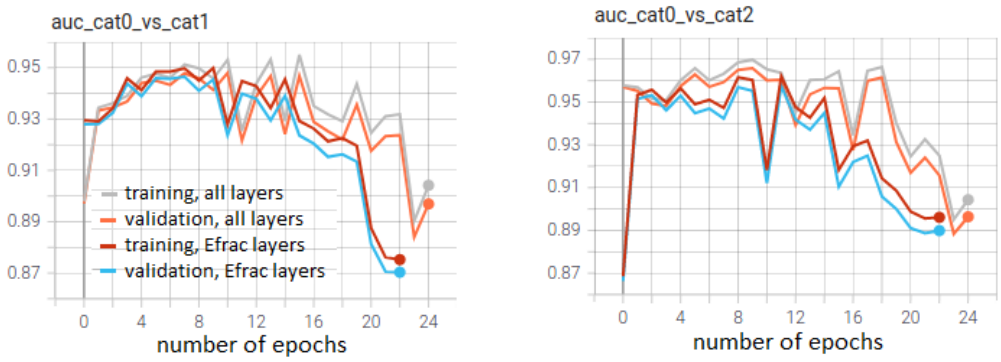
(a)



(b)

Figure 8: Receiver operating characteristic (ROC curves) for category 0 against category 2 for a): energy fraction layers and b): all layers. Results of training are presented on the left, results of validation – on the right.

The next two plots in Fig. 9 present the Area Under the ROC Curves (AUC) for category 0 versus category 1 as well as category 2. AUC curve represents the capability of the model of distinguishing between classes. The performance is the best at epoch 9 as it was mentioned before. Results for both trainings and validations at step 9 are similar, however the validation results corresponding to all layers give greater value as far as classification of class 0 against class 2 is concerned (0.965 versus 0.955). In Fig. 9(a), the results of validation at epoch 9 take the values of ∼0.945.



(a) (b)

Figure 9: Area Under the ROC Curves (AUC) for a): category 0 against category 1 and b): category 0 against category 2.

4 Conclusions

The project includes analysis of boosted top tagging performances in leptonic channels using jet images. The ResNeXt50 network is able to properly classify leptonic top jets against a hadronic top and light flavour QCD background. The plots of ROC curves have been taken from the epoch number 9, as at that step the values of epoch accuracy and loss reach the best values, namely 0.84 and 0.44. The ROC curves show that for the 60% of signal efficiency the level of background is at $\sim 10^{-2}$ for hadronic top jets and $\sim 2 \cdot 10^{-4}$ for QCD (for both all layers as well as energy fraction layers only). Hence, the results look promising.

As far as advantages of the proposed method are concerned, it should be noted that conventional methods usually require a very well identified lepton, which can be very challenging to do in highly boosted cases where the b quark jet and the lepton start overlapping. The CNN based technique does not rely on very precise lepton reconstruction and identification.

Possible future plans for in-depth studies are as following. A general multiclassifier should be changed to a custom classifier for specific signal and background types only. Moreover, it is worth checking how does the performance of the training behave in different bins of top quark p_T and also include other backgrounds, for instance leptonically and hadronically decaying W/Z bosons. Finally, it is necessary to compare presented results with conventional (non ML based) methods.

Acknowledgements

I would like to thank especially my supervisor, Soham Bhattacharya, for his guidance, patience and support throughout all the weeks. Also, thanks to Isabell Melzer-Pellmann for all of her advises and commitment. Finally, a big thanks to the CMS group and DESY Summer Student Programme organizors.

References

[1] Soham Bhattacharya, Monoranjan Guchait, Aravind H. Vijay, *Boosted Top Quark Tagging and Polarization Measurement using Machine Learning*, arXiv:2010.11778v3 [hep-ph] (2 Jul 2021)

[2] S. Marzani, G. Soyez, M. Spannowsky, *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*, vol. 958, Springer (2019), 10.1007/978-3-030-15709-8, [1901.10342]

[3] J. Cogan, M. Kagan, E. Strauss, A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, JHEP 02 (2015) 118 [1407.5675]

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

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

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

[7] Max Drummy, Mag Gardner, Joanne Quinn, Joanne McEachen, Michael Fullan, *Dive into Deep Learning*, available from `https://d2l.ai`

[8] Saining Xie *et al*, *Aggregated Residual Transformations for Deep Neural Networks*, arXiv:1611.05431v2 [cs.CV] (11 Apr 2017)

[9] *Classification models* - github repository available from `https://github.com/qubvel/classification_models`

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