PROCEEDINGS OF SCIENCE

Deep Neural Network resizing for real-time applications in High Energy Physics

Andrea Di Luca,[a,b,c,*] **Daniela Mascione,**[a,b,c] **Francesco Maria Follega,**[a,b] **Marco Cristoforetti**[b,c] **and Roberto Iuppa**[a,b]

[a] *Dipartimento di Fisica, Università di Trento, Via Sommarive 14, 38123 Trento, Italy*
[b] *TIFPA, Via Sommarive 14, 38123 Trento, Italy*
[c] *FBK, Via Sommarive 18, 38123 Trento, Italy*

E-mail: andrea.diluca@unitn.it

The ability to execute Deep Neural Networks at the trigger level to improve online selection performance will be crucial for current and future high-energy physics experiments. Low-latency hardware solutions exist, e.g. FPGAs, but the primary constraint to the implementation is often related to the model's size, which has to be finely tuned not to exceed the available memory. We present here an approach to reduce the size of models, having under control the model performances. Promising results are shown in the classification problem of selecting proton-proton collision events in which the boosted Higgs boson decays to two b-quarks, and both the decay products are contained in a large and massive jet, against an overwhelming QCD background.

The Ninth Annual Conference on Large Hadron Collider Physics - LHCP2021
7-12 June 2021
Online

[*] Speaker

1. Introduction

Current and future high-energy physics experiments at particle colliders will have to cope with extremely high collision rates; at the Large Hadron Collider, for example, events are produced at 40 MHz frequency. It is, therefore, necessary to implement real-time event processing capabilities. Among the standard pattern recognition algorithms thought to be run on Look-Up Tables, Machine Learning methods, particularly Deep Neural Networks (DNN), are spreading quickly. As a result, there is growing interest in executing such algorithms at the trigger level to improve online selection performance. The main issue in running these algorithms in real-time is the amount of operation that needs to be computed. Field-programmable gate arrays (FPGA) can fit this task thanks to their low latency and high throughput. However, a significant drawback about this kind of hardware choice is that the model's size that should be loaded on the FPGA depends strongly on the FPGA specifications. In the last years, new techniques that reduce the size of DNN models have been developed and tested for FPGA applications. Baseline techniques in this regard are pruning and quantization[1].

Here we propose an alternative approach to reduce the size of a DNN model based on the introduction of CancelOut layers in the networks. This type of layer can be used both for feature selection and to reduce the hidden structure by deactivating not relevant nodes.

2. CancelOut layer

The CancelOut layer [2] is composed of neurons that have one single input, as shown in fig. 1 for a CancelOut layer used after the input layer. The node's output is the product of the input feature with the sigmoid of the associated weight. After the training, irrelevant features will have an associated Cancelout weight that outputs small numbers after the sigmoid application. In contrast, for the relevant features, the output will be close to one. In its original formulation, the weights of the CancelOut layer are used as a metric to evaluate the impact of one feature in the decision process of a DNN model. In this sense, it is fundamental to have weights that are different among themselves. Our purpose is different; we developed a modified architecture of the CancelOut layer so that only a certain fraction, defined by the user, of the CancelOut nodes are active (sigmoid of the weight equal to 1) while the others are switched off (sigmoid of the weight equal to 0).

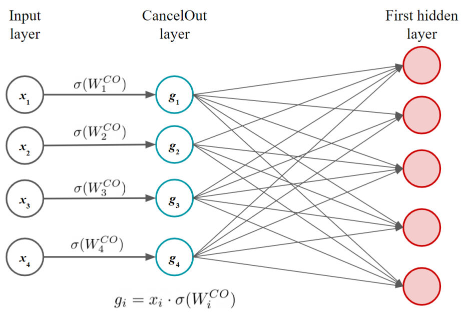

Figure 1: Example of a CancelOut layer inserted after the input layer.

3. Boosted $H \to bb$ tagging

We tested the CancelOut layer while developing a fully connected DNN to classify pp collision events where a Higgs boson with very high transverse momentum decays to two b-quarks. In this regime, the decay products of the Higgs boson are very collimated and it is challenging to resolve the di-jet structure [3]. A single large and massive jet containing both the b quark originated jets is more likely to be reconstructed. This channel is interesting for the study of Higgs boson properties since it accounts for 58% of the total Higgs boson decays [4] and observations of deviations from Standard Model prediction are expected in the boosted regime [5]. However, recognizing these events in a pp collision experiment represents a challenging task, mainly because of the huge irreducible background of QCD multi-jet production.

3.1 Simulated data and object reconstruction

The dataset used to develop the classifier is produced using a framework developed by combining Pyhtia8[6], to generate high-energy physics events, Delphes [7], to simulate the detector response and RAVE [8] for secondary vertex reconstruction. Large radius anti-k_t jets [9] (large-R jets) with $R = 1$ are reconstructed together with variable radius track jets [10] with $R_{\text{MAX}} = 0.4$, $R_{\text{MIN}} = 0.02$ and $\rho = 30$ (ρ parameter determines how fast the effective jet size decreases with the transverse momentum of the jet). For the large-R jets, we defined kinematic variables plus jet substructure variables. For the variable R track jets, we defined kinematic variables plus the b-tagging information and variables connected to the secondary vertex. We selected large-R jets with $p_T > 450\text{Gev}/c^2$ and $\eta < 2$. Then we look for the 2 highest p_T track jets contained in a selected large-R jet. Therefore, the total number of initial features is 39.

3.2 Reduction of the number of input features

The different models were trained by varying the number of desired features that the CancelOut layer must activate. The left plot in fig 2 shows the behavior of the CancelOut weight associated with each feature when varying the number of requested features. Relevant features are activated already when asking for a low number of features. Then, increasing the number of asked features one by one, single features are activated and remain active in most cases. The performances of the models are shown on the right side of fig. 2. It is important to notice that after a certain number of features are switched on, there is no significant improvement in the performance. This confirms the idea that irrelevant features were last activated.

3.3 Pruning of the hidden structure

CancelOut layers can also be inserted between hidden layers of the network to reduce the size of the internal structure. As for the pruning technique, the user can define a priori the number of nodes of the hidden layer to switch off. Figure 3 shows the performance of a model where the CancelOut layer is inserted between the third and the fourth hidden layer of the classifiers to switch off 25% of the nodes of the third layer.

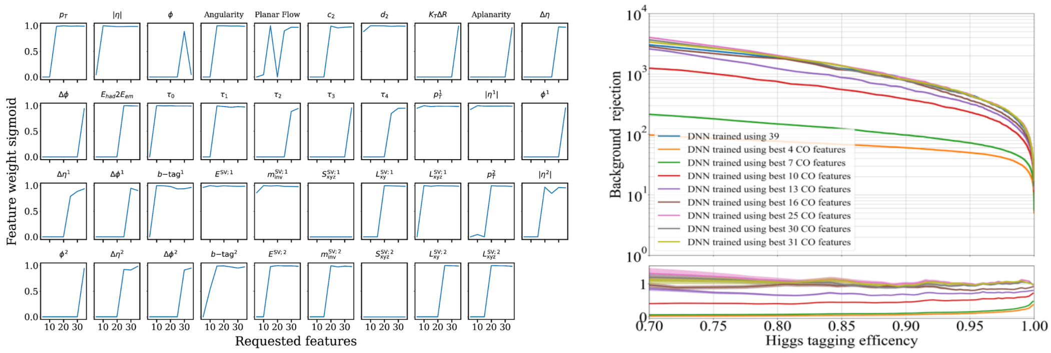

Figure 2: (Left) CancelOut weight for each feature as function of the desired number of features. (Right) Background rejection rate versus Higgs tagging efficiency with large-R jets of different model trained by varying the number of features.

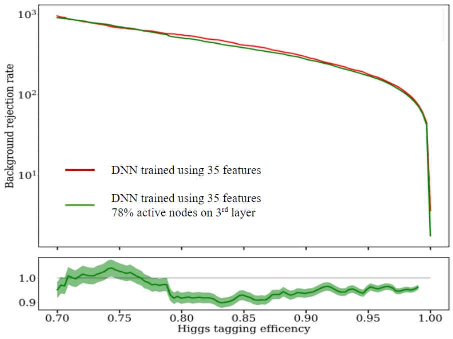

Figure 3: Background rejection rate versus Higgs tagging efficiency with large-R jets of two different models. The CancelOut layer is used to reduce the hidden structure size.

4. Conclusion

For applications in future high-energy physics experiments, it is essential to reduce the size of a neural network to run on hardware like FPGAs. In this work, we proposed an original method based on the use of CancelOut layers to reduce the size of the input parameter space and prune the hidden structure of the network. Promising results were shown in the development of a DNN classifier to correctly identify pp events where a boosted Higgs boson decay to two b-quarks. Furthermore, the CancelOut layer can be easily added to any existing model and used together with other neural network reduction approaches.

References

[1] S. Han, H. Mao and W. Dally, *Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding*, 10, 2016.

[2] V. Borisov, J. Haug and G. Kasneci, *Cancelout: A layer for feature selection in deep neural networks*, in *Artificial Neural Networks and Machine Learning – ICANN 2019: Deep Learning*, I.V. Tetko, V. Kůrková, P. Karpov and F. Theis, eds., (Cham), pp. 72–83, Springer International Publishing, 2019.

[3] ATLAS Collaboration collaboration, *Performance of large-R jets and jet substructure reconstruction with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2012-065, CERN, Geneva (Jul, 2012).

[4] LHC Higgs Cross Section Working Group collaboration, *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, 1610.07922.

[5] M. Grazzini, A. Ilnicka, M. Spira and M. Wiesemann, *Modeling BSM effects on the Higgs transverse-momentum spectrum in an EFT approach*, *JHEP* **03** (2017) 115 [1612.00283].

[6] 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].

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

[8] W. Waltenberger and F. Moser, *Rave - an open, extensible, detector-independent toolkit for reconstruction of interaction vertices*, in *2006 IEEE Nuclear Science Symposium Conference Record*, vol. 1, pp. 104–109, 2006.

[9] M. Cacciari, G.P. Salam and G. Soyez, *The anti-ktjet clustering algorithm*, *Journal of High Energy Physics* **2008** (2008) 063.

[10] D. Krohn, J. Thaler and L.-T. Wang, *Jets with Variable R*, *JHEP* **06** (2009) 059 [0903.0392].