LHCP2021

The Ninth Annual Conference on Large Hadron Collider Physics

7-12 June 2021 ~~Paris (France), Sorbonne Université (IN2P3/CNRS,IRFU/CEA)~~

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

A.Di Luca, D. Mascione, F.M. Follega, M.Cristoforetti, R.Iuppa

deeppp

UNIVERSITY OF TRENTO

FONDAZIONE BRUNO KESSLER

INTRODUCTION

At LHC events are produced at a **frequency of 40 MHz** and events that are discarded by the trigger are lost. To **improve selection performance** there is a great interest in **running Deep Neural Networks in real-time.**

FPGA (Field-programmable gate array) can fit this task (low latency and high throughput). Depending on the FPGA size, **we should know how to reduce the size of a model.** Baseline techniques are pruning and quantization **[1]**.

Here we propose **an approach to select only relevant features using a CancelOut layer.**

BENCHMARK APPLICATION

We developed an **H→bb tagger** for *pp* collision experiments based on a Deep Neural Network to identify jets that contain both the b quarks from boosted H decay.

Dataset

$4x10^6$ simulated events of *pp*-collision at 14 TeV

Each candidate (**39 features**) is a **large radius jet** (anti-kT jet with $R = 1$) with the **2 variable radius track jets** ($R_{MAX} = 0.4, R_{MIN} = 0.02, \rho = 30$) contained in the large radius jet with highest p_T.

PYTHIA8

generation of high-energy physics events

+

DELPHES fast simulation

DELPHES

detector response fast simulation

RAVE

RAVE

secondary vertex reconstruction

CANCELOUT LAYER

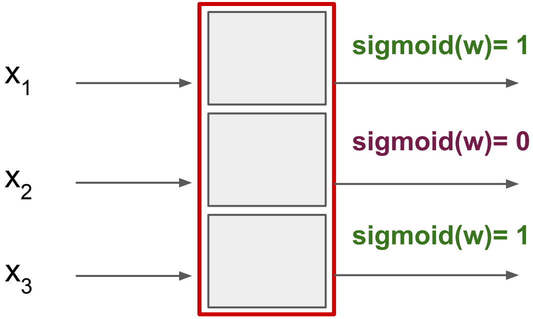
CancelOut layer [2] can rank relative importances among features in input to a Deep Neural Network at training time.

We developed a **modified architecture** to **activate only a certain number** (defined by the user) **of features** having under control the performances.

It can be **easily added to existing models** and **used together with other neural network reduction approaches**.

In this form CancleOut can be used to:

1. Reduce the number of input nodes
2. Prune hidden layer nodes



x_1

x_2

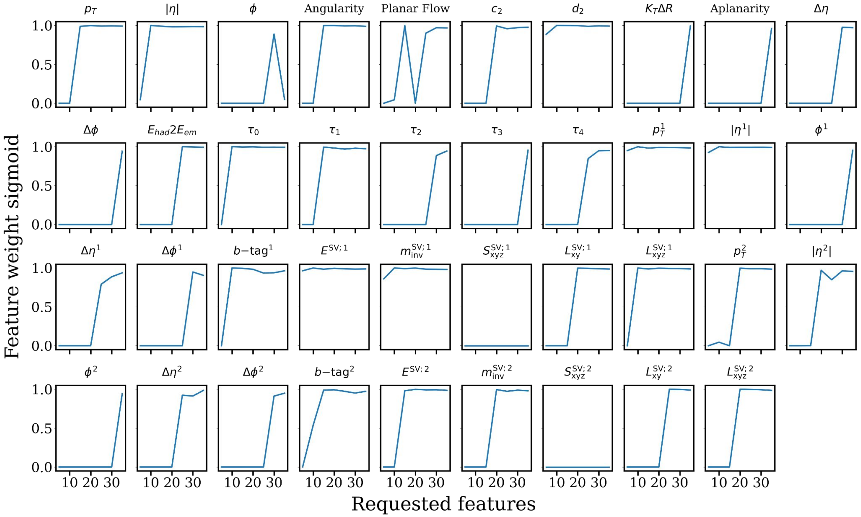
x_3

sigmoid(w)= 1

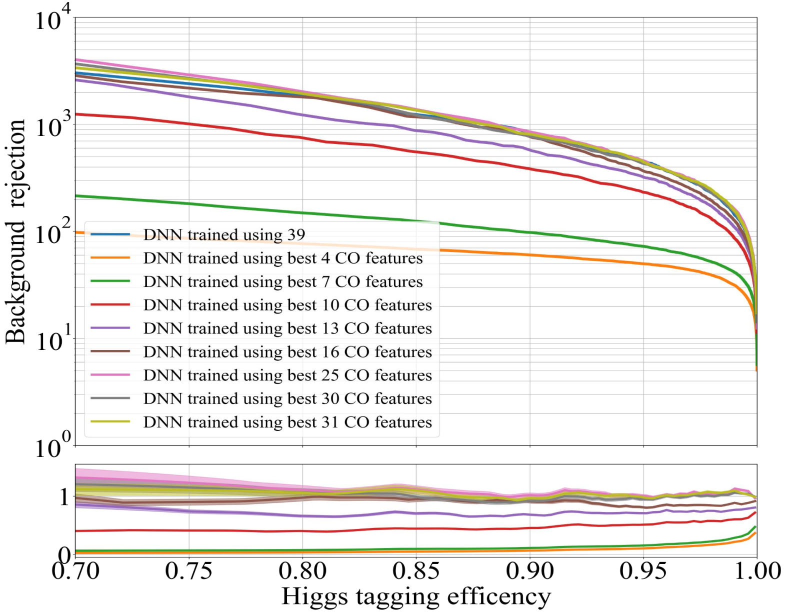
sigmoid(w)= 0

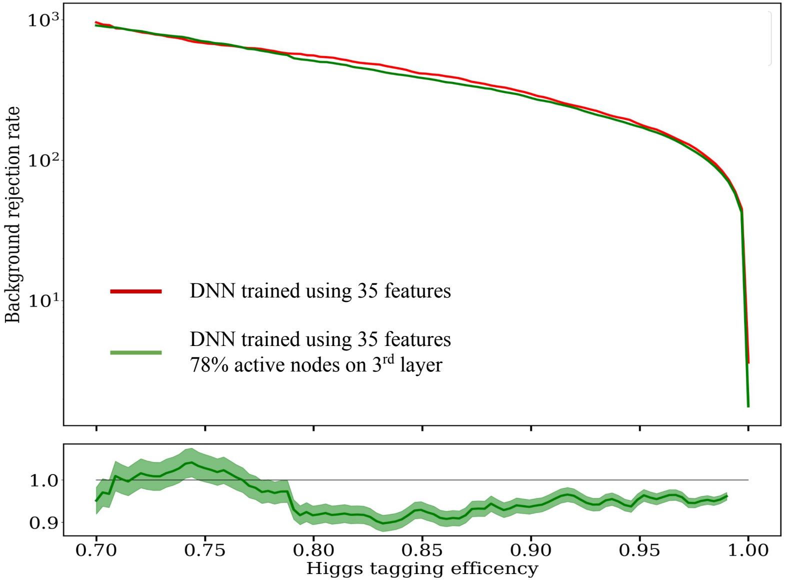
sigmoid(w)= 1

CancelOut layer



RESULTS





After a certain number of features are activated, there is no significant improvement in the performance.

Performance aware hidden nodes pruning that can be easily applied by adding a hidden CancelOut layer.

References

[1] S. Han et al., *Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding*

[2] V. Borisov et al., *CancelOut: A Layer for Feature Selection in Deep Neural Networks*