Boosted Decision Trees in the Level-1 Muon Endcap Trigger at CMS

Darin Acosta[1], Andrew Brinkerhoff[1], Elena Busch[2], Andrew Carnes[1], Ivan Furic[1], Sergei Gleyzer[1], Khristian Kotov[1], Jia Fu Low[1], Alexander Madorsky[1], Jamal Rorie[2], Bobby Scurlock[1], and Wei Shi[2] on behalf of the CMS Collaboration

[1] University of Florida
[2] Rice University

Abstract. The first implementation of a Machine Learning Algorithm inside a Level-1 trigger system at the LHC is presented. The Endcap Muon Track Finder (EMTF) at CMS uses Boosted Decision Trees (BDTs) to infer the momentum of muons in the forward region of the detector, based on 25 different variables. Combinations of these variables representing 2^{30} distinct patterns are evaluated offline using regression BDTs. The predictions for the 2^{30} input variable combinations are stored in a 1.2 GB look-up table in the EMTF hardware. The BDTs take advantage of complex correlations between variables, the inhomogeneous magnetic field, and non-linear effects – like inelastic scattering – to distinguish high momentum signal muons from the overwhelming low-momentum background. The new momentum algorithm reduced the background rate by a factor of three with respect to the previous analytic algorithm, with further improvements foreseen in the coming year.

1. Introduction

The Compact Muon Solenoid (CMS) is a detector at the Large Hadron Collider (LHC) located near Geneva, Switzerland. The LHC collides bunches of protons every 25 ns at a center of mass energy of 13 TeV. The CMS experiment detects the resulting particles and measures their kinematics using various subdetectors working in concert. With 40 million proton bunch crossings per second amounting to roughly 40 TB of data each second, saving the information from every event is not feasible. As such, the CMS trigger system chooses the interesting events to save to disk, operating in two stages [1]. The Level-1 (L1) trigger runs in hardware online reducing the throughput of data from 40 MHz to 100 KHz. From there, the High Level Trigger (HLT) operates in software online reducing the rate from 100 KHz to 1 KHz. In the end, about 1 GB/s is saved to disk.

With 40 MHz of input, the L1 Trigger has only 4 μs to decide whether to keep the information for an event. The Endcap Muon Track Finder (EMTF) – part of the L1 Trigger dedicated to muons – has only about 500 ns to determine the location, tracks, and momentum of the muons passing through the Cathode Strip Chambers (CSC) and Resistive Plate Chambers (RPC) in the endcaps of CMS [2]. High momentum muons are an important object for many physics analyses at CMS. As such, an accurate momentum assignment distinguishing low momentum muons (background) from high momentum muons (signal) is key to the EMTF trigger. In order to meet the timing requirements, the EMTF's logic is implemented in Field Programmable Gate

Arrays (FPGAs), a type of reprogrammable hardware that allows vast parallelization and speeds much greater than even the best CPUs.

To improve the transverse momentum (p_t) assignment for muons in the endcaps at Level-1, the EMTF team trained Boosted Decision Trees (BDTs) offline using TMVA [3], and stored the prediction scheme into a 1.2 GB Look-Up Table (LUT). The FPGAs then use the LUT online to assign the p_t in a single operation. Using the LUT to turn the BDT p_t assignment into a simple look-up enables the EMTF to utilize the power of a robust machine learning algorithm for its momentum predictions while still operating at the required time scale. Putting a parallelized version of the BDTs directly into the FPGAs, while hypothetically possible, would require more than the available number of logic gates. Such an implementation would still be slower than the LUT method, and changing the p_t assignment would require reprogramming the FPGA logic each time. The LUT method provides a simple way to run any machine learning evaluation at high speed by turning the evaluation into a single operation.

2. Metrics of Success

Two metrics are used to measure the success of the EMTF: the rate and the efficiency. The rate at X GeV is defined as the number of muons with a predicted p_t greater than X GeV. In other words, the rate consists of both true and false positives above the p_t threshold. The efficiency at X GeV is defined as the number of muons with both predicted p_t and true p_t greater than X GeV divided by the number of muons with true p_t above X GeV. Put another way, the efficiency measures the percentage of muons with true p_t above X GeV that were correctly predicted above X GeV. When evaluating the efficiency in data, the true p_t is typically provided by the offline reconstruction. When evaluating the efficiency in simulation, the true p_t is typically provided by the Monte Carlo truth. A good trigger will minimize the data saved without losing the interesting high p_t events where unexplored physics lies, i.e. it will minimize rate while maximizing the efficiency.

3. The EMTF Regression Project

A muon traveling through the endcap detectors has a chance to leave hits in four sequential stations labeled 1, 2, 3, and 4. The specific combination of hits like 1,3,4 is called the mode. Each station records the ϕ and θ location of a hit, among other information. The CSCs have better spatial resolution, so the ϕ and θ information is taken from the CSCs by default, but the RPC measurements for the station are used if the CSCs missed the hit in the same station. The charged muons travel through a magnetic field following curved paths due to the Lorentz force. The force causes the high p_t muons in a magnetic field to bend less and the low p_t muons to bend more. The difference in ϕ and θ between stations i and j, $\Delta\phi_{ij}$ and $\Delta\theta_{ij}$, quantify the curvature of the track. With most of the curvature accounted for by the $\Delta\phi$ variables, the $\Delta\phi$s provide the majority of the p_t discrimination.

A major difficulty in minimizing the rate is the steeply falling p_t distribution. A typical interesting event has p_t greater than 25 GeV, and there are about one thousand 5 GeV muons for every 25 GeV muon. With so many more low p_t events, predicting the low momentum muons poorly will drastically increase the rate. Moreover, in addition to the large number of low p_t muons, there are other notable difficulties: the muons travel through a non-uniform magnetic field, some scatter between detector stations, and those with high p_t often shower charged particles upon interacting with the detector material. Moreover, low p_t muons may spiral completely before getting to the next station. The scattering, showering, and spiraling add noise to the underlying true behavior, while the number of low p_t muons requires that the regression focus on the low momentum regime to prevent an explosion in the rate.

In order to assign p_t in a robust way and deal with the aforementioned difficulties, a BDT is trained for each possible mode using the discretized values for the features of Table 1. The

loss function and weights are chosen to focus on the low p_t events and minimize the rate while maintaining acceptable efficiency. Features are chosen for each mode to give the BDT the information needed to predict the p_t while dealing with the non-uniform magnetic field and the problematic scattering and showering effects.

The $\Delta\phi$ variables available for each mode are used as features to determine the curvature and get most of the p_t discrimination. However, the power of these variables depends largely on the track position in θ. The magnetic field varies as a function of θ affecting the magnitude of the curvature for a given p_t, thus correlating $\Delta\phi$, p_t, and θ. The link between these three makes θ the next most important training feature.

Variables modeling the mean and RMS of the available $\Delta\phi$s for the mode are also used as features in order to identify scattering and showering effects. If a muon were to scatter or shower between stations the recorded hit in a station may not be the true hit of the muon. Any $\Delta\phi$ involving this station will be an outlier. To determine the severity of the deviation and the likelihood of scattering/showering, the idea is to identify the outlier station and to compare the mean and RMS $\Delta\phi$ with and without the outlier station. The greater the difference the greater the severity. The nominal mean and RMS $\Delta\phi$ features are calculated using all available $\Delta\phi$s for the mode. The exclusive mean and RMS are calculated using all available $\Delta\phi_{ij}$ for the mode with i or j $\neq S_{out}$, where S_{out} is the outlier station. S_{out} is the excluded station such that leaving it out of the sum minimizes the mean and RMS. The outlier station, S_{out}, is also used as a feature. Including the nominal mean and RMS of $\Delta\phi$, the exclusive RMS and mean of $\Delta\phi$, and S_{out} as features helps the BDT differentiate scattering, showering, and normal events.

The features described above are the most important features, but not the whole collection. The front-rear (FR) bit designates whether the muon hit a front or rear CSC chamber in the station, and it is also included. The $\Delta\theta$s provide additional curvature information, and these are included as well. The B feature for each station is included as well, and it flags whether the ϕ, θ information for the station came from the CSCs or the RPCs. If there are bits available for the B_i feature it also includes information about the single station $\Delta\phi$ bend angle within a CSC chamber. Lastly, the $+/-$ feature stores the signs of the later $\Delta\phi$s relative to the first $\Delta\phi_{ij}$ for the mode.

4. Putting the BDTs into a Look-up Table

After training BDTs for each mode, the mode and the fundamental features from which the others can be derived are discretized and fit into a 30 bit word. The discretization scheme is different for each mode, detailed in Table 1. With the feature space compressed into 30 bits, there are 2^{30} possibilities that need to be assigned a p_t. A LUT is created by looping over all 2^{30} possible bit words, decoding each word into the fundamental features, deriving the secondary features, and sending the values to the BDT to assign the p_t prediction. Using 9 bits for the p_t, this amounts to a 1.2 GB LUT where each bit word value is an address and the p_t is the value in memory. Discretizing the feature space and creating a LUT turns the p_t assignment into a single operation. The LUT is then used by the FPGA logic online to assign p_t to muon tracks in the EMTF. The LUT method is a simple way to run any machine learning method quickly, but compressing the features into 30 or so bits may not always be feasible for the application.

Table 1. The feature discretization scheme for each mode.

Four Station Modes

Mode	**Feature**	$\Delta\phi_{12}$	$\Delta\phi_{23}$	$\Delta\phi_{34}$	+/-	$\Delta\theta_{14}$	B_1	B_2	B_3	B_4	FR_1	θ	Mode
1-2-3-4	**Bits**	7	5	4	2	2	2	1	1	1	1	3	1

Three Station Modes

Mode	**Feature**	$\Delta\phi_{12}$	$\Delta\phi_{23}$	+/-	$\Delta\theta_{13}$	B_1	B_2	B_3	FR_1	FR_2	θ	Mode
1-2-3	**Bits**	7	5	1	3	2	1	1	1	1	5	3
Mode	**Feature**	$\Delta\phi_{12}$	$\Delta\phi_{24}$	+/-	$\Delta\theta_{14}$	B_1	B_2	B_4	FR_1	FR_2	θ	Mode
1-2-4	**Bits**	7	5	1	3	2	1	1	1	1	5	3
Mode	**Feature**	$\Delta\phi_{13}$	$\Delta\phi_{34}$	+/-	$\Delta\theta_{14}$	B_1	B_3	B_4	FR_1	FR_3	θ	Mode
1-3-4	**Bits**	7	5	1	3	2	1	1	1	1	5	3
Mode	**Feature**	$\Delta\phi_{23}$	$\Delta\phi_{34}$	+/-	$\Delta\theta_{24}$	B_2	B_3	B_4	FR_2	–	θ	Mode
2-3-4	**Bits**	7	5	1	3	2	1	1	1	–	5	4

Two Station Modes

Mode	**Feature**	$\Delta\phi_{XY}$	$\Delta\theta_{XY}$	B_X	B_Y	FR_X	FR_Y	θ	Mode
X-Y	**Bits**	7	3	3	3	1	1	5	7

X-Y runs through the possible two station combinations: 1-2, 1-3, 1-4, 2-3, 2-4, 3-4.

5. Results and Conclusions

The LUT scheme utilizing the BDT predictions has been implemented in the EMTF for 2016 and 2017 data taking. As seen in Figure 1, the upgraded system – compared to the legacy system – reduces the rate at 25 GeV by a factor of three with no loss in efficiency. The legacy system was used in the endcaps until 2015 and assigned p_t using a maximum likelihood fit with up to three $\Delta\phi$s. More detailed information about the legacy system can be found in [1].

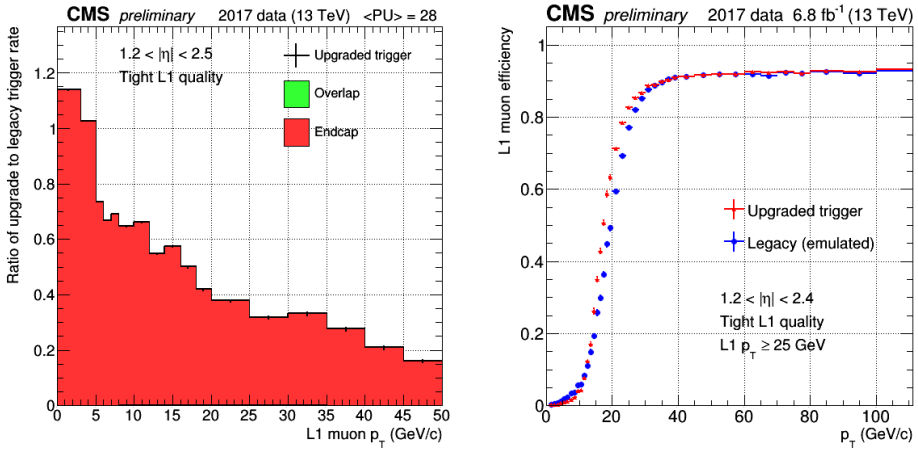

Figure 1. On the left, the upgraded EMTF rate divided by the legacy rate is shown for a variety of p_t thresholds. On the right, the upgraded and legacy efficiencies are presented for a 25 GeV threshold. The upgraded EMTF has a 3x lower rate than the legacy system at 25 GeV with virtually no difference in plateau efficiency for the same threshold. Plots are taken from [4].

References

[1] Khachatryan V *et al.* (CMS) 2017 *JINST* **12** P01020 (*Preprint* 1609.02366)

[2] Tapper A and Acosta D (CMS collaboration) 2013 CMS Technical Design Report for the Level-1 Trigger Upgrade Tech. Rep. CERN-LHCC-2013-011. CMS-TDR-12 additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch URL https://cds.cern.ch/record/1556311

[3] Hoecker A, Speckmayer P, Stelzer J, Therhaag J, von Toerne E and Voss H 2007 *PoS* **ACAT** 040 (*Preprint* physics/0703039)
[4] CMS Collaboration 2017 URL https://cds.cern.ch/record/2286327