End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data

M. Andrews[1], J. Alison[1], S. An[1,2], P. Bryant[1], B. Burkle[3], S. Gleyzer[4], M. Narain[3], M. Paulini[1], B. Poczos[5], and E. Usai[3]

[1]Department of Physics, Carnegie Mellon University, Pittsburgh, USA
[2]CERN, Geneva, Switzerland
[3]Department of Physics, Brown University, Providence, USA

[4]Department of Physics, University of Alabama, Tuscaloosa, USA
[5]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA

October 17, 2019

Abstract

From particle identification to the discovery of the Higgs boson, deep learning algorithms have become an increasingly important tool for data analysis at the Large Hadron Collider (LHC). We present an innovative end-to-end deep learning approach for jet identification at the Compact Muon Solenoid (CMS) experiment at the LHC. The method combines deep neural networks with low-level detector information, such as calorimeter energy deposits and tracking information, to build a discriminator to identify different particle species. Using two physics examples as references: electron vs. photon discrimination and quark vs. gluon discrimination, we demonstrate the performance of the end-to-end approach on simulated events with full detector geometry as available in the CMS Open Data. We also offer insights into the importance of the information extracted from various sub-detectors and describe how end-to-end techniques can be extended to event-level classification using information from the whole CMS detector.

1   Introduction

The analysis of data collected at the LHC often involves the classification of hadronic jets and particles in collision events. The CMS experiment [1] uses the particle flow algorithm [2] that combines raw data from different subdetectors to identify different classes of particles. The particle-flow objects are then clustered into jets and used as an input to algorithms aimed at distinguishing different flavours of the jet (light quark, gluon, c quark, b quark, and top quark). Recently, jet classification algorithms based on deep neural networks have been introduced [3, 4], achieving the current state-of-the-art performance in jet classification tasks, often using simplified detector models.

2   The End-to-end approach

This work presents an innovative approach to the classification of jets and particles that instead uses low-level detector information. This approach is called end-to-end (E2E) classification and in this work is applied to develop a discriminator between light quark and gluon-initiated jets [5].

The E2E approach consists of building a convolutional neural network (CNN)-based jet or event discriminator by using the low-level detector inputs directly. We use an image-based approach where the information from each subdetector is represented by one or more image maps and combined with those from other subdetectors to form a single multi-layer image. The pixel coordinates of the image are then mapped to the discretized detector coordinates expressed in azimuthal angle ϕ the pseudorapidity η.

The work presented here uses simulated datasets from the CMS Open Data repository [6], which are essential to ensuring a high-fidelity representation of the detector. New releases of the CMS Open Data portal continue to be published that specifically target machine learning (ML) studies [7, 8], including end-to-end-oriented workflows. For instance, tools and instructions for reproducing the full simulation chain of CMS are available allowing access to any kind of low-level detector information one may need.

The first step to the construction of the detector images is to generate an image for the electromagnetic calorimeter (ECAL). For this subdetector, each pixel in the image corresponds to a crystal of ECAL. The intensity of each pixel is proportional to the energy measured in the corresponding crystal. Other data attributes like timing of the energy deposit are also available, though this did not yield any further improvement for the classification tasks considered in this work. This may potentially change, however, with the improved timing capabilities of the Phase 2 upgrades of the CMS detector.

The second image considered contains information from the hadronic calorimeter (HCAL). In this case, each pixel of the image corresponds to one of the calorimeter towers. Similarly, the intensity of each pixel contains the energy collected from each tower. By construction, a 5x5 window of ECAL pixels corresponds to a single HCAL pixel.

Finally, for the tracking system, reconstructed tracks are used. The position of the track is propagated to the ECAL surface. The (η, ϕ) coordinates of the track position at the ECAL surface are discretized on a grid with the same granularity as the ECAL's. The intensity of each pixel corresponds to the transverse momentum of the track. Figure 1 shows the E2E images for the three subdetectors detectors described above.
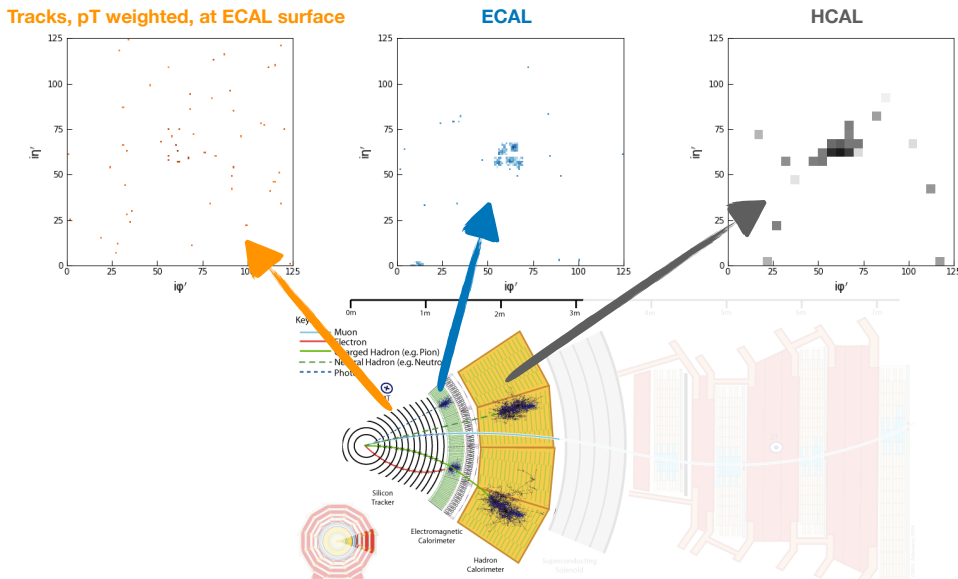

Figure 1: E2E images for the electromagnetic calorimeter, the hadronic calorimeter and the tracker of the CMS detector.

One of the problems of generating detector images is that the some endcap detectors use a different coordinate system. For example, while the HCAL endcap sensor elements are discretized in (η, ϕ) coordinates, the ECAL endcaps are discretized in Cartesian coordinates (x, y). In order to obtain a single image from the barrel and the endcap regions of a subdetector, the images need to be stitched together in the same coordinate system. This is done by morphing the endcap images, making a coordinate transform from (x, y) to (η, ϕ) then stitching them to the edges of the barrel image.

The full-detector E2E images can be directly fed to a CNN for training an event-wide discriminator. On the other hand, smaller images can be cropped from the full-detector around region of interests, such as

jets or photons. In the case of jet discriminators we crop an 125x125 pixel image around the jet.

3   Previous work

This section briefly summarizes the previous applications of the E2E framework [9]. The first application of E2E was the construction of a discriminator between photons and electrons. This application used only ECAL energy images. The E2E images were fed to a ResNet-15 [10] network obtaining an area under the curve of the Receiver Operating Charactersitic (ROC AUC) score of 0.788. Additionally, a di-photon vs. di-electron discriminator was built using similarly constructed detector images and a ResNet-23 network, obtaining a ROC AUC of 0.997.

The E2E framework was also used to distinguish the Higgs boson to di-photon decay from non-resonant di-photon production and misidentified photons. This application used ECAL, HCAL and track images. One of the challenges in this particular application was the correlation of the discriminator output with the mass of the Higgs boson, which is an undesirable feature in physics analyses. The discriminator was de-correlated from the diphoton mass through the use of an additional loss penalty proportional to the learned mass correlation as measured by the Cramer-Von Mises metric [11]. The de-correlated discriminator yielded a ROC AUC score of 0.77 in distinguishing $H \rightarrow \gamma\gamma$ from non-resonant $\gamma\gamma$ production, 0.95 for $H \rightarrow \gamma\gamma$ vs. γ+jet (misidentified photon), and 0.81 for H vs. non-resonant $\gamma\gamma$ and γ+jet together.

4   Quark- vs. gluon-jet discriminator

Similarly to the $H \rightarrow \gamma\gamma$ case, the E2E quark vs. gluon jet classifier uses three image channels from the ECAL, HCAL, and track momentum. Figure 2 shows the three E2E detector images for quark and gluon jets, averaged over 70k jets.
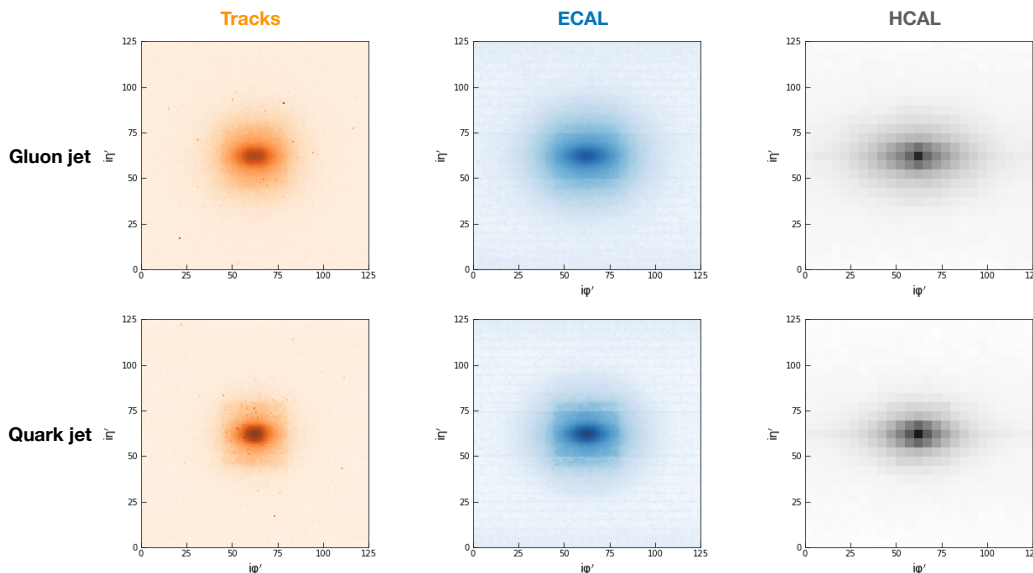

Figure 2: E2E detector images for quark and gluon jets, averaged over many events

To understand the relevance of the information contained in the different image layers, the ResNet-15 network is first trained using images from a single subdetector alone. The tracker layer shows the highest ROC AUC score at 0.782, followed by ECAL (0.760), and finally HCAL (0.682). Following this, we can assess how much of the information in the different layers is redundant by training on different combinations of the subdetector layers. The combination of ECAL and HCAL images has a ROC AUC of 0.781, slightly below using the track layer alone. On the other hand, adding ECAL to just the track layer brings a considerable improvement to the ROC AUC at 0.804. Finally, using all three subdetectors together brings only a modest improvement (0.808).

The performance of the jet classifier is then benchmarked against the QCD-aware recursive neural network (RecNN) approach outlined in [12] and adapted to the quark vs. gluon jet case in [13]. Figure 3 shows a comparison between the E2E and RecNN quark vs. gluon discriminators using the same jets selected from the CMS Open Data dataset. The E2E jet classifier thus has performance comparable to the existing state-of-the-art, and performs significantly better than traditional jet images.

	ROC AUC
E2E image, ECAL+HCAL+Tracks	**0.8077 ± 0.0003**
RecNN, ascending-p_T	0.8017 ± 0.0003*
RecNN, descending-p_T	0.802
RecNN, anti-k_T	0.801
RecNN, Cambridge/Aachen	0.801
RecNN, no rotation/reclustering	0.800
RecNN, k_T	0.800
RecNN, k_T-colinear10-max	0.799
RecNN, random	0.797

Figure 3: Performance comparison of the E2E jet classifier vs. RecNN. *Represent ROC AUC mean ± standard deviation over five training runs.

The final step consists of building a di-jet qq vs gg discriminator. Three different scenarios are considered. In scenario A, we crop images around the two jets. In scenario B, the four-momenta of the two jets are included with the images. Lastly, in scenario C, the full-detector E2E images are used. Scenario C has the best performance with a ROC AUC of 0.889, followed by scenario B (0.878) and a (0.876). Approaches A and B have a very similar performance, while using a full detector image gives the best result.

5 Recent developments

Recent developments in the E2E framework goes in the direction of using more and lower-level detector information, particularly from the tracking system. Different image channels are added for different parameters of the reconstructed tracks. In addition to the transverse momentum, new layers are added that encode the transverse and longitudinal projections of the impact parameter of the track, dz and d0, computed with respect to the primary vertex of the interaction. The images for these new channels have the same pixel distribution as the transverse momentum image, but each pixel is assigned a different intensity, depending on the track parameter represented. These additional parameters allow the encoding of information from pileup vertices and secondary vertices coming from the hadronization of b quarks.

Another idea still involves using even lower-level tracker information. The CMS tracking system includes a three-layer pixel detector placed very close to the beampipe. In particular, each of the three pixel detector layers can each be transformed in an image layer. Each pixel can be either 0 or 1 depending on the presence of a hit in the corresponding position. The resolution of the images could also be increased to fully take advantage of the higher resolution of the tracking system. Figure 4 shows a full-detector E2E image with these three pixel layers in addition to the original ECAL, HCAL, and track momentum layers.

6 Conclusion

The E2E approach can be used for particle and jet identification. We show that the algorithm can learn particle kinematics and shower shapes. We adapted the E2E framework to build a quark- vs. gluon-initiated jet discriminator. The performance of the algorithm is competitive to the current state-of-the-art for this task. Work is currently ongoing to extend this approach to include even lower-level tracker information and other classification tasks such as for top vs. light-flavour jet discrimination.
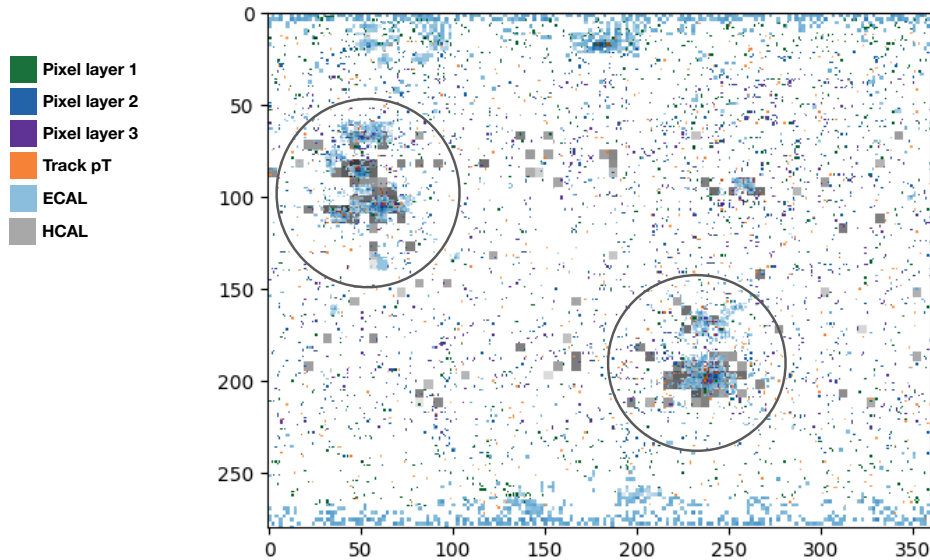
Figure 4: E2E detector images including new tracker layers.

References

[1] The CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008).

[2] The CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017).

[3] The CMS Collaboration, "New Developments for Jet Substructure Reconstruction in CMS", (2017). `arXiv:CMS-DP-2017-027`.

[4] The CMS Collaboration, "Identification of Heavy-flavour Jets with the CMS Detector in pp collisions at 13 TeV", *JINST* **13** (2018) `doi:http://dx.doi.org/10.1088/1748-0221/13/05/P05011`.

[5] M. Andrews et al., "End-to-End Jet Classification of Quarks and Gluons with the CMS Open Data", *arXiv e-prints* (Feb, 2019) arXiv:1902.08276, `arXiv:1902.08276`.

[6] CERN Open Data Portal, "Getting Started with CMS 2011 Open Data", 2017. CERN Open Data Portal.

[7] E. Usai et al., "Samples with full event information including tracker hits for tracking, ML, and top quark tagging studies", 2019. `doi:10.7483/OPENDATA.CMS.CHC3.5KPG`.

[8] E. Usai et al., "TrackerRecHitProducerTool - Example workflow to add tracker hits to CMS Run1 AOD and to produce full event ROOT ntuple with tracker hits", 2019. `doi:10.7483/OPENDATA.CMS.GZU3.FZ51`.

[9] M. Andrews, M. Paulini, S. Gleyzer, and B. Poczos, "End-to-End Physics Event Classification with the CMS Open Data: Applying Image-based Deep Learning on Detector Data to Directly Classify Collision Events at the LHC", *arXiv e-prints* (Jul, 2018) arXiv:1807.11916, `arXiv:1807.11916`.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016* [10], pp. 770–778. `doi:10.1109/CVPR.2016.90`.

[11] A. Rogozhnikova et al., "New approaches for boosting to uniformity", *JINST* **10** (2015).

[12] G. Louppe, K. Cho, C. Becot, and K. Cranmer, "QCD-Aware Recursive Neural Networks for Jet Physics", (2017). `arXiv:hep-ph/1702.00748`.

[13] T. Cheng, "Recursive Neural Networks in Quark/Gluon Tagging", (2018). `arXiv:hep-ph/1711.02633v2`.