Calibrating Electrons and Photons in the CMS ECAL using Graph Neural Networks

Simon Rothman, on behalf of the CMS Collaboration
MIT, University of Minnesota
srothman@mit.edu

Abstract

The Compact Muon Solenoid (CMS) detector is one of two general-purpose detectors on the energy frontier of particle physics at the CERN Large Hadron Collider (LHC). Products of proton-proton collisions at a center of mass energy of 13 TeV are reconstructed in the CMS detector to probe the standard model of particle physics, and to search for processes beyond the standard model. The development of precision algorithms for this reconstruction is therefore a key objective in optimizing the precision of all physics results at CMS. While machine learning techniques are now prevalent at CMS for these tasks, they have largely relied on high-level human-engineered input features. However, much of the disruptive impact of machine learning in industry has been realized by bypassing human feature engineering and instead training deep learning algorithms on low-level data. We have developed a novel machine learning architecture based on dynamic graph neural networks which allows regression directly on low-level detector hits, and we have applied this model to the calibration of electron and photon energies in CMS. In this work, the performance of our new architecture is shown on electrons used in the calibration of the CMS detector, where we obtain an improvement in energy resolution by as much as 10% with respect to the previous state-of-the-art reconstruction method.

1 Introduction

1.1 Motivation

The Compact Muon Solenoid (CMS) detector is one of two general-purpose detectors at the CERN Large Hadron Collider (LHC). Products of proton-proton collisions at a center of mass energy of 13 TeV are reconstructed in the detector to probe the standard model of particle physics and to search for processes not predicted by the standard model. Analysis of the CMS data relies on the precision that various decay products can be measured and the original interaction reconstructed. Precision reconstruction of electrons (e) and photons (γ) is of particular importance, as these are some of the most common decay products and several flagship measurements at CMS rely on electrons and photons, including Higgs boson physics in the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ ($\ell = e, \mu$) channels. The development of precise reconstruction techniques for these objects is therefore essential in order to optimize the precision and sensitivity of these analyses.

1.2 Electron and photon reconstruction in CMS

The design of the CMS detector is optimized for precision measurement of e/γ energies, with a large, highly granular homogeneous electromagnetic calorimeter (ECAL) consisting of 75,848 scintillating $PbWO_4$ crystals [1] arranged in a central barrel and two endcaps. Electrons and photons are detected as showers of secondary particles in the ECAL. An incident particle is therefore reconstructed as

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

a collection of individual calibrated per-crystal energy deposits, termed "RecHits," with anywhere from 1 to over 100 RecHits being assigned to each particle. The raw energy, E_{Raw} summed over this collection is subject to a number of effects that degrade the resolution of the ECAL, including:

- Energy lost before reaching the ECAL and in detector gaps.
- Energy leakage out of the back of the ECAL.
- The use of finite energy thresholds to suppress noise in the detector electronics.
- Interference due to other particles passing through the detector ("pileup").

These effects are compensated with corrections derived on a per-particle basis using a machine learning regression. This regression is currently implemented as a Boosted Decision Tree (BDT) with ≈ 30 high-level input features used to describe the shower. These BDT-based energy corrections have been fine-tuned over several years [2] and have supported all physics analyses using e/γ objects in CMS during LHC Run 2, including the measurement of the Higgs boson mass with a precision of 0.1% [3]. However, much of the disruptive impact of machine learning in industry has been realized by bypassing human feature engineering and instead training deep learning algorithms on low-level data (e.g. [7]). This makes development of a deep learning architecture for these energy corrections a compelling goal, as improvements in the precision with which e/γ objects can be reconstructed will directly benefit these analyses.

2 New machine learning approach

2.1 The Dynamic Reduction Network

Thanks to recent advances in graph neural network techniques and the widespread availability of high-performance co-processors for the training and deployment of these models it has become possible to develop such a deep learning model. This allows us to use the RecHits associated with a given particle as the direct input to our machine learning model rather than the high-level derived features input to the current regression. As these collections of RecHits are inherently sparse objects consisting of anywhere from 1 to over 100 hits distributed widely across the detector, it is natural to represent them as graphs. Our novel architecture, the "Dynamic Reduction Network" (DRN) [6] is therefore built on point cloud graph neural network techniques (e.g. [9]). The input to our model is a point cloud of RecHits in (position, energy) space, and graphs are formed by drawing edges between neighboring hits in a high-dimensional latent space. This graph-based approach has a number of advantages both over the current BDT model and other potential architectures such as convolutional neural networks (CNNs). These include:

- Use of low-level input features ensures access to the full information content of every event.
- Events with arbitrary number of hits distributed arbitrarily across the detector can be treated without padding or truncation.
- The use of geometric point cloud input allows easy handling of complex detector geometries and the inclusion of additional detector systems.

The DRN is based on dynamic graph neural networks with the addition of a pooling step analogous to subsampling in CNNs. Our architecture is summarized in Figure 1, and proceeds as follows:

1. The position and energy coordinates of each RecHit are mapped into a high-dimensional latent space by a fully-connected neural network
2. Global information is developed by iteratively performing message passing on dynamically-generated k-nearest neighbors graphs and max-pooling graclus-clustered [5] pairs of vertices.
3. The resulting learned high-level features are supplemented by additional human-engineered features to account for information not encoded in the collection of detector hits. In particular, two additional features, which describe the amount of energy leakage at the back of the ECAL and the energy density from pileup events, are concatenated to the learned features.
4. The resulting set of high-level features are passed through another fully-connected neural network to produce the regression output.

A PyTorch implementation of the DRN can be found at `https://github.com/ssrothman/DynamicReductionNetwork`

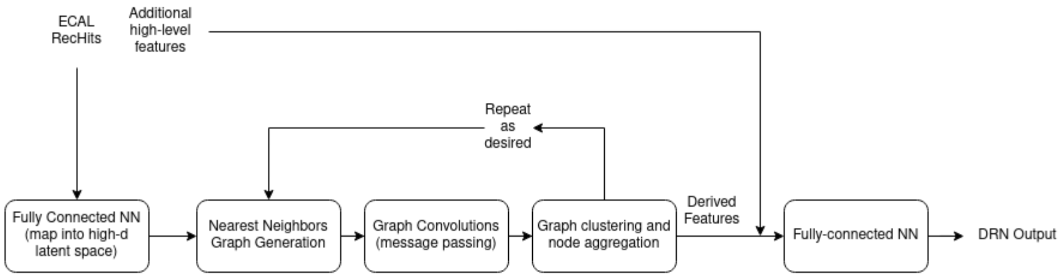

Figure 1: Flowchart of the operation of the Dynamic Reduction Network. A point cloud of RecHits is mapped into a high-dimensional latent space by a fully-connected neural network, where it is then iteratively transformed and pooled by graph operations. This resulting high-level learned features are then concatenated with extra high-level information not available from the raw collection of RecHits and passed through another fully-connected neural network to obtain the regression output.

2.2 Physics-motivated loss function

A number of optimizations have been made to the loss function used in this work in order to take into account various known physical properties of the system. First, rather than directly predicting the energy corrections $y^{(i)} = E_{True}/E_{Raw}$, we instead use the logarithm of this quantity:

$$y^{(i)} = \log \frac{E_{True}^{(i)}}{E_{Raw}^{(i)}} \tag{1}$$

This ensures that the loss function is symmetric with respect to proportional over- and under-measurement of energies, allowing much more efficient training. Second, the response function of the CMS ECAL has asymmetric tails, which we account for by employing a semi-parametric approach. In particular, we take the regression output to be the 6-dimensional parameterization of a double-sided crystal ball probability (dscb) density function [8], which has a Gaussian core with power-law tails on both sides. This has the additional effect of automatically giving an estimate of both the energy correction and a per-particle energy resolution. We then take a simple log-likelihood loss function:

$$\ell = -\frac{1}{N}\sum_{i=0}^{N} \mathrm{dscb}\left(y_{True}^{(i)}; \mathrm{DRN}\left(X^{(i)}\right)\right) \tag{2}$$

where $\mathrm{dscb}(y; \mathrm{DRN}(X))$ is the crystal ball probability density function parameterized by the regression output of the DRN and N is the number of training events.

In order to maintain a stable regression response over all possible inputs encountered during data taking and to avoid unphysically large energy corrections, we restrict our model to only apply energy correction factors less than 2 in either direction. As our regression target is the logarithm of the correction factor, this translates to the requirement,

$$-\log 2 \leq y_{Pred}^{(i)} \leq \log 2 \tag{3}$$

which is enforced by a sigmoid response function.

2.3 Training

We train on realistic detector simulation data, which accurately models particle interactions and detector effects including pileup. This gives us access to the truth energy values, allowing for supervised training. Our training sample consists of simulated electrons and positrons with a flat true energy distribution fired directly into the detector. Our training data is generated under exactly the same conditions as that used to train the current BDT model. Our dataset includes ≈ 17 million electrons, of which 80% are used for training and 20% for validation. Training is performed on an NVidia Tesla v100 GPU at the Minnesota Supercomputing Institute, and takes ≈ 20 hours.

3 Physics performance

3.1 Validation strategy

In order to validate the performance of our model we construct histograms of E_{Pred}/E_{True}, where E_{Pred} is the central value of the energy predicted by the DRN. These histograms are then fit with a Cruijff function [4], allowing extraction of the key metrics: mean response (μ) and relative resolution (σ/μ). For an ideal regression, $\mu = 1$, and σ/μ is as small as possible. Note that exactly the same validation strategy is app

3.2 Performance comparison with the previous state-of-the-art

In this work we show only the performance of our regression on simulated electrons reconstructed using only information from the ECAL; these are used for calibration of the detector but are not the final energies used in analyses. We find that our regression has a similar highly stable response to that of the Run-2 BDT, with the mean E_{Pred}/E_{True} stable to within 0.4% as a function of energy, detector coordinates, level of pileup, and other variables describing the shape of the electromagnetic shower. Additionally we find that we obtain an improved resolution with respect to the Run-2 BDT by a factor of approximately 10% at all values of the energy, detector, coordinates, level of pileup, and shower shape.

We also expect similar performance on photons, as they have nearly identical interactions with the ECAL and other subdetectors have minimal impact on the energy resolution at energies greater than about 25 GeV.

4 Summary

We have developed a novel architecture to derive the energy corrections to be applied to e/γ objects in the CMS ECAL and have shown the application of this model to electrons used in the calibration of the CMS detector. The mean stability of the new method is similar to the current BDT used by CMS, and the reconstruction has an energy resolution that is improved by $\approx 10\%$. The stability and the resolution are the same in different regions of the detector and are independent of the density of particles from pileup events and shower size. The development of the corresponding regressions for the final e/γ objects used in physics analyses is ongoing, and we aim to deploy this model for use globally within CMS for LHC Run 3. In addition to the application shown here in e/γ reconstruction, we believe that the DRN is a powerful tool for any problems involving sparse data and complicated geometries within and without particle physics.

5 Acknowledgements

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award Number DE-SC0011845. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. `http://www.msi.umn.edu`.

We congratulate our colleagues in the CERN accelerator departments for the excellent performance of the LHC and thank the technical and administrative staffs at CERN and at other CMS institutes for their contributions to the success of the CMS effort. In addition, we gratefully acknowledge the computing centers and personnel of the Worldwide LHC Computing Grid and other centres for delivering so effectively the computing infrastructure essential to our analyses. Finally, we acknowledge the enduring support for the construction and operation of the LHC, the CMS detector, and the supporting computing infrastructure provided by the following funding agencies: BMBWF and FWF (Austria); FNRS and FWO (Belgium); CNPq, CAPES, FAPERJ, FAPERGS, and FAPESP (Brazil); MES and BNSF (Bulgaria); CERN; CAS, MoST, and NSFC (China); MINCIENCIAS (Colombia); MSES and CSF (Croatia); RIF (Cyprus); SENESCYT (Ecuador); MoER, ERC PUT and ERDF (Estonia); Academy of Finland, MEC, and HIP (Finland); CEA and CNRS/IN2P3 (France); BMBF, DFG, and HGF (Germany); GSRI (Greece); NKFIA (Hungary); DAE and DST (India); IPM (Iran); SFI (Ireland); INFN (Italy); MSIP and NRF (Republic of Korea); MES (Latvia); LAS (Lithuania); MOE and UM (Malaysia); BUAP, CINVESTAV, CONACYT, LNS, SEP, and UASLP-FAI

(Mexico); MOS (Montenegro); MBIE (New Zealand); PAEC (Pakistan); MSHE and NSC (Poland); FCT (Portugal); JINR (Dubna); MON, RosAtom, RAS, RFBR, and NRC KI (Russia); MESTD (Serbia); SEIDI, CPAN, PCTI, and FEDER (Spain); MOSTR (Sri Lanka); Swiss Funding Agencies (Switzerland); MST (Taipei); ThEPCenter, IPST, STAR, and NSTDA (Thailand); TUBITAK and TAEK (Turkey); NASU (Ukraine); STFC (United Kingdom); DOE and NSF (USA).

References

[1] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08), 2008.

[2] The CMS Collaboration. Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. *Journal of Instrumentation*, 16(05), 2021.

[3] The CMS Collaboration. Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{s} = 13$ TeV. *Journal of High Energy Physics*, 27, 2021.

[4] P. del Amo Sanchez, J. P. Lees, V. Poireau, E. Prencipe, V. Tisserand, J. Garra Tico, E. Grauges, M. Martinelli, A. Palano, M. Pappagallo, and et al. Study ofb→xdecays and determination of|vtd/vts|. *Physical Review D*, 82(5), Sep 2010.

[5] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.

[6] Lindsey Gray, Thomas Klijnsma, and Shamik Ghosh. A dynamic reduction network for point clouds. *CoRR*, abs/2003.08013, 2020.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[8] R. Partridge, C. Peck, F. Porter, W. Kollmann, M. Richardson, K. Strauch, K. Wacker, D. Aschman, T. Burnett, M. Cavalli-Sforza, D. Coyne, H. Sadrozinski, R. Hofstadter, I. Kirkbride, H. Kolanoski, K. Königsmann, A. Liberman, J. O'Reilly, J. Tompkins, E. Bloom, F. Bulos, R. Chestnut, J. Gaiser, G. Godfrey, C. Kiesling, and M. Oreglia. Observation of an η_c candidate state with mass 2978 ± 9 mev. *Phys. Rev. Lett.*, 45:1150–1153, Oct 1980.

[9] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See github link in Section 2.1
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2.3

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 2
(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 2.3

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]