End-to-End Event Classification of High-Energy Physics Data

M Andrews[1], M Paulini[1], S Gleyzer[2], B Poczos[3]

[1] Department of Physics, Carnegie Mellon University, Pittsburgh, USA
[2] Department of Physics, University of Florida, Gainesville, USA
[3] Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA

E-mail: mandrews@cmu.edu, paulini@heps.phys.cmu.edu, sergei@cern.ch, bapoczos@cs.cmu.edu

Abstract. Feature extraction algorithms, such as convolutional neural networks, have introduced the possibility of using deep learning to train directly on raw data without the need for rule-based feature engineering. In the context of particle physics, such end-to-end approaches can be used for event classification to learn directly from detector-level data in a way that is completely independent of the high-level physics reconstruction. We demonstrate a technique for building such end-to-end event classifiers to distinguish simulated electromagnetic decays in a high-fidelity model of the CMS Electromagnetic Calorimeter.

1. Introduction & Motivation

An essential part of any new physics search at the Large Hadron Collider (LHC) at CERN involves event classification, or distinguishing signal events from the background. Traditional machine learning techniques have relied on high-level features in the form of particle 4-momenta - consistent with our understanding of particle physics phenomenology. However, such high-level features are the result of a rule-based casting of the raw detector data into progressively more physically-motivated quantities. While such approaches have proven to be useful, they are highly dependent on our ability to completely and effectively model all aspects of particle decay phenomenology. On the other hand, powerful image-based Convolutional Neural Network (CNN) algorithms have emerged, capable of training directly on raw data, learning the pertinent features unassisted - so-called *end-to-end deep learning classifiers*. In this paper, we explore the use of such classifiers on simulated data from the Electromagnetic Calorimeter (ECAL) of the Compact Muon Solenoid (CMS) detector at the LHC. Through the use of simulated toy decays, we demonstrate that these classifiers are able to effectively discern event phenomenology completely unassisted.

In this paper, we restrict ourselves to electromagnetic showers. A number of recent efforts have concentrated on image-based physics classification - for instance, jet shower [1, 2] and neutrino classification [3], as well as traditional event classification

[4–6]. However, these approaches mostly rely on the output of rule-based high-level physics reconstruction algorithms and are thus subject to any mis-modeling contained therein. While a few have begun to use low-level data for event classification [7], in their current form, they still employ image construction techniques that depict the underlying detector geometry in rough approximation. As suggested in [6], such techniques potentially suffer from pixelization effects. In contrast, we construct our images in the detector coordinate system using only detector-level data. This minimizes our exposure to potential mis-modeling effects outside of the event classifier itself and ensures as high fidelity a representation of the detector as possible.

In section 2, we describe how CMS ECAL performs hit reconstruction. In section 3, we describe the use of end-to-end classifiers for electromagnetic shower classification. Then, in section 4, we discuss the construction of a full event classifier using various types of state-of-the-art deep learning models. We discuss our future plans in section 6 and summarize our conclusions in section 7.

2. Detector Reconstruction

2.1. The CMS ECAL Detector

The CMS experiment is one four large collider experiments at the LHC at CERN [8]. The focus of this paper is on the Electromagnetic Calorimeter, the sub-detector of CMS, responsible for resolving and localizing the energies of photons and electrons. It plays a major role in the detection of the Higgs boson and other Electroweak phenomena [4]. The ECAL is a hermetically-sealed cylinder composed of lead tungstate ($PbWO_4$) crystals packed together into a barrel section (EB) and two circular endcap (EE) sections. The barrel section is composed of 61200 crystals segmented 170-fold in pseudorapidity (η) and 360-fold in the azimuth angle (ϕ). Each circular endcap section contains 7324 crystals segmented in a rectilinear grid. The crystals are angled toward the interaction point but with a slight offset. This minimizes particles slipping through cracks but means the amount of an ECAL crystal a particles traverses is position-dependent.

2.2. ECAL Hit Reconstruction

2.2.1. Particle Shower Formation and Detection. On its way from the interaction point, a high-energy electron (e) or photon (γ) will interact with material from other subdetectors in CMS, causing it to shower and deposit its energy over a range of crystals in the ECAL. While shower formation is a stochastic process, on average, roughly 94% of the e/γ shower's energy is deposited within a 3 by 3 grid of crystals. Upon interacting with the ECAL crystals, the e/γ shower induces scintillating light which is measured by Avalanche Photodiodes (in EB) or Vacuum Phototriodes (in EE) at the other end of the crystal as a signal pulse. To first approximation, e/γ shower profiles are expected to be identical in the ECAL. However, due to its interaction with the magnetic field of the

CMS solenoid (B = 3.8 T), the charged electron emits bremsstrahlung, preferentially in ϕ. This introduces a higher-order perturbation to the e shower profile causing it to be more spread out and slightly asymmetric in ϕ compared to the photon shower.

2.2.2. The Digitized Hit. The signal pulse is amplified and shaped by a multi-gain preamplifier before being digitized at 40 MHz. The net effect is a signal pulse with a stepped profile that rises sharply to a peak before falling off gradually, with a full width of under 150 ns at half maximum. For every recorded event, ten such steps or *samples* are stored, giving ten *amplitude* readings per crystal, separated by 25 ns. These samples define the raw *digitized hit* or *digi*, and represent the lowest level, physically-sensible calorimeter data. The timing calibration is such that the pulse appears on the 4th time sample, such that the baseline electronics noise, known as the *pedestal*, can be estimated from the first three time samples.

2.2.3. The Reconstructed Hit. A fitting algorithm is applied to the digitized hit to determine the energy and time from the peak and shape of the pulse. A number of effects must be corrected for before arriving at a calibrated hit. Most notable is the crystal transparency loss and recovery of the crystals under the presence and absence of beam radiation, respectively. After correcting for this and other effects, a calibrated, *reconstructed hit* or *rec hit* is produced containing energy and timing information.

2.2.4. High-level Physics Features. Through a rule-based process known as high-level physics reconstruction [9], the reconstructed hits are transformed into higher-level features like particle IDs, 4-momenta, and shower shapes. Further in the analysis chain, even more complex features are engineered to achieve greater separation of signal from background. These features have their roots in particle physics theory. In machine learning parlance, these are the equivalent of *hand-engineered* features. In particle physics, as in computer vision, high-level features have traditionally served as inputs to machine learning algorithms like fully-connected neural networks and boosted decisions trees for event classification.

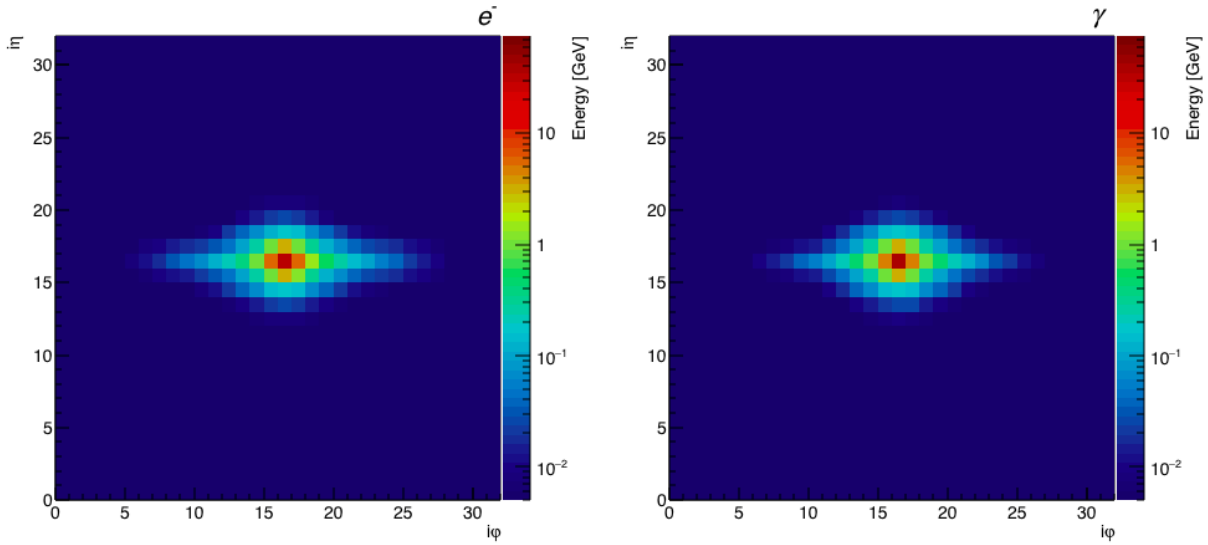
Having described ECAL hit reconstruction, we use the term *end-to-end classifier*, more precisely, to denote classifiers that use either digitized or reconstructed hits as inputs.

3. Shower Classification

As a first step towards event classification, we tackle the easier problem of e/γ shower classification. To begin, we use events generated with exactly one negatively charged e or one γ of fixed transverse momentum $p_T = 50\,\text{GeV}$ fired from the interaction point to a direction sampled uniformly from pseudorapidity $|\eta| < 1.4$ and azimuthal angle $-\pi < \phi < \pi$, effectively constraining the e/γ shower to the barrel section of ECAL.

3.1. Input Image

To construct the image of the shower, we take a grid of 32 by 32 crystals from ECAL centered around the maximum-energy shower deposit (the *shower seed*). Each pixel in the image grid will correspond exactly to one crystal, though not necessarily the *same* crystal from event to event, and will be filled with the relevant data for that crystal: amplitude, energy, etc. As the image is always centered on the shower seed, shower classification does not take full advantage of the feature translation abilities of CNNs. The classifier must still be able to generalize due to the stochastic nature of shower formation, and position-dependent crystal interaction. Images of averaged electron and photon showers are shown in Figure 1.



(a) Average electron shower

(b) Average photon shower

Figure 1: e/γ showers averaged over 50k showers. The e shower is slightly more spread out in ϕ-in addition to being slightly asymmetric-due to bremsstrahlung effects.

3.2. Preprocessing

Preprocessing is essential for ensuring numerical stability in the optimizer, especially with large images and deep networks. We experimented with linear and logarithmic transformations. For the latter, because the shower image is sparse, we attempted different schemes of shifting the log-transformed inputs relative to the sparse values.

For the digi amplitudes, there is the additional complication of neighboring crystals reading out the pedestal values due to the activation of the ECAL selective readout algorithm. We, therefore, tried different pedestal subtraction schemes in addition to applying linear or log transformations.

3.3. Network Architecture.

We experimented with deep neural network architectures that can be primarily grouped on the following categories: Convolutional Neural Networks (CNN) [10], Long Short-Term Memory (LSTM) [11], and Fully-Connected Networks (FCN) [12]. All of these were implemented using the Keras Deep Learning library [13].

The CNN category included simplified implementations of VGG [14], Inception [15], and Residual Net (ResNet) [16] architectures. We tried the following data combination schemes:

(i) energy only
(ii) energy and time stacked
(iii) digitized amplitudes stacked
(iv) energy, time, and digitized amplitudes stacked
(v) (ii) and (iii) concatenated at output of convolutional layers
(vi) (ii) and (iii) concatenated at output of fully-connected layers

The LSTM networks only used digi inputs and included the following architectures: (i) Time-distributed: each digi amplitude image is connected to its own CNN, flattened, and given as a sequence to LSTM layers. (ii) Convolutional: the LSTM layers act directly on a time sequence of convolutional layers [17].

The FCNs learned from the flattened energy and time images or the flattened digi amplitude images and consisted of 2-, 3-, 6-hidden layers with 256 units per layer.

3.4. Training Strategy.

The simulated sample consisted of 498k events evenly balanced between e and γ. Of these, 320k (64%) were used for training, 89.6k (18%) for validation, and 88.4k (18%) for a final test set. The validation set was used to evaluate competing architectures and hyper-parameters within the above three network categories. For the best network in each category, we evaluate their performance on the test set, with the results summarized in section 5.

For e/γ classification, we used binary cross-entropy loss with the ADAM adaptive gradient optimizer, both with and without explicit learning rate decay. The initial learning rate was set to 1×10^{-3} and we trained for 50 epochs by which point the validation loss had more than plateaued. A dropout of 20% for fully-connected layers and ReLU activation were used, with weights initialized from a truncated normal distribution where applicable. Finally, the area under the curve of the Receiver Operating Characteristic (ROC AUC) was chosen as a performance figure of merit, due it interpretability in terms of signal efficiency and background rejection.

4. Event Classification

Two studies are presented. The first is a generalization of the shower classification to full detector images. Recall that the e/γ particles are fired randomly, and so the showers will appear to move in the image from event to event, providing a real test of the feature-translation abilities of CNNs in a sparse image.

Then, we proceed to generalize to double particle decays-either an electron-positron pair e^+e^-, or a photon pair $\gamma\gamma$. The particles are allowed to have a range of transverse momenta $p_T = (20, 80)\,\text{GeV}$ but it is required to be the same for each particle. The pair decays back-to-back, such that the position of one is the inverse of the other in either e^+e^- or $\gamma\gamma$ decay. One can think of two ways a classifier might learn to distinguish multi-particle events: either from a difference in particle shower profiles, or from a difference in the spatial arrangement of the particle showers, i.e. event kinematics. Back-to-back decays provide a useful scenario for investigating the former due to the identical kinematics. This forces the network to learn only from the shower profiles, providing an unbiased test of the ability of CNNs to perform feature translation. Finally, we include pile-up of $\langle \text{PU} \rangle = 25$, but no underlying event simulation.

4.1. Input Image & Preprocessing.

In event classification, the image grid is 170 by 360 pixels corresponding to the full ECAL barrel geometry. There is again an exact correspondence between calorimeter crystals and image pixels. However, now the crystals underlying the detector image are always fixed.

4.2. Network Architecture & Training Strategy.

A VGG-type network proved to be cumbersome for event classification. The convolutional to fully-connected interface caused an explosion of model weights by several orders of magnitude. Coupled with the increased image sparsity ($> 90\%$), this made training unstable at learning rates above $\sim 10^{-3}$. Instead, we experimented primarily with ResNet-type architectures without any fully-connected layers. While the ResNets were stable even above learning rates of $\sim 10^{-3}$, we found that a lower learning rate of 5×10^{-4} worked best.

Interestingly, the original residual block worked better with sparse data compared to the bottleneck version of the residual block [16]. We found the best performance for the residual block *without* batch normalization, and with max pooling instead of average pooling, as might be expected. A total of three down-samples performed better compared to the typical five used in most ResNet implementations. Finally, with 320k training events, any ResNet deeper than about 23 convolutional layers showed no further gain in performance.

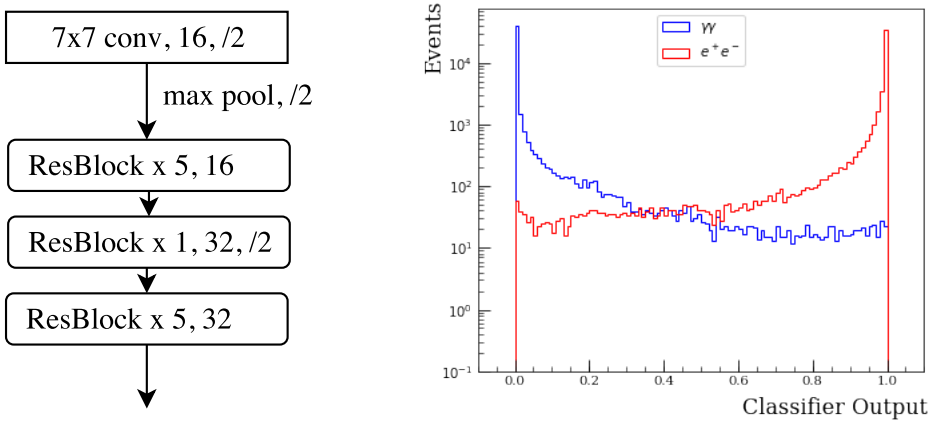
5. Results.

5.1. Shower Classification Results

The classification scores of each deep learning model evaluated on the test set are summarized in Table 1.

Table 1: Shower Classification Results.

Category	**Network**	**ROC AUC**
CNN	VGG, energy	0.807
LSTM	Conv-LSTM, digis	0.799
FCN	3-layers, digis	0.770



(a) ResNet-23

(b) e^+e^- vs. $\gamma\gamma$ Classifier Output

Figure 2: Event Classification Results for e^+e^- vs. $\gamma\gamma$ using a 23-layer Residual Net.

Within the CNN category, the different architectures all scored within 0.1% of each other. LSTM-based architectures, as expected, took considerably longer to train but showed no clear advantage over purely-CNN architectures. Within the FCN category, the different networks all scored within $< 1\%$ of each other. The FCN-based architectures consistently underperformed, even though feature translation did not play a major role. Linear preprocessing worked best in each category, though, in general, the results were not sensitive to the choice of preprocessing.

As noted in 2.2.1, in the ECAL, e showers are at first approximation expected to be identical to their γ counterparts. And, by eye, it is difficult to tell the two apart. Yet our results indicate that deep learning classifiers are able make very fine distinctions between the two, picking up on higher-order perturbations to deliver a good classification score.

5.2. Event Classification Results

The classification scores of the best deep learning model (ResNet-23) applied to single and di-particle events are summarized in Table 2.

Table 2: Event Classification Results.

Decay	**Network**	**ROC AUC**
e vs. γ	ResNet-23	0.788
e^+e^- vs. $\gamma\gamma$	ResNet-23	0.997

The single e/γ score compares favorably to the one from shower classification, although is slightly lower. Therefore, while CNNs handle feature translation well even for highly sparse data, the transition is not perfect. One important questions is - what advantage does event classification offer over simpler shower classification? A partial answer is provided by the di-particle result.

As shown in Table 2, the classifier score for double photon versus double electron events is considerably higher. While the boost in classification score seems a little surprising at first, it is to be expected. In particular, one *could* model the di-particle prediction using the product of the individual shower classifiers' predictions. If the mean single shower mis-classification accuracy is y'_{single}, then on average, one would accurately classify the di-particle pair at a rate of $y_{\text{pair}} = 1-(y'_{\text{single}})^2$. From the shower classification CNN, $y'_{\text{single}} \sim 0.26$, giving $y_{\text{pair}} \sim 0.93$. The observed event classifier accuracy of ~ 0.97 is even higher, suggesting that there are other effects present. Importantly, this accuracy is achieved over a range of momentum and with pile-up. Even though we expect even better performance compared to hand-engineered features in more complicated decay topologies, the fact that the classifiers perform so well even in the relatively simple two-body decays, is encouraging. As this example suggests, the strength of end-to-end event classifiers may lie in tackling complex, multi-particle decays, without the need to introduce explicit kinematics.

6. Future Work

For future work, we would like to extend our study in two directions. First, we plan to study more complex pp decays with non-trivial kinematics, with an eye towards a realistic physics analysis. There are additional challenges, such ensuring that the classifier is not correlated with an observable of interest. Second, we plan to extend the end-to-end approach to the full CMS detector. The next natural steps are to include the ECAL endcaps, the Hadronic Calorimeter and then the Tracker and Muon Systems. Building a full end-to-end classifier from all the available data from the CMS detector is the ultimate goal of this study.

7. Conclusions

In this paper, we have presented a technique for building end-to-end physics classifiers for single electromagnetic shower classification and event classification. First, we constructed high-fidelity images of electromagnetic showers from the CMS ECAL barrel using low-level detector data. Afterwards, we used these images to train various deep learning models to distinguish electron- and photon-induced showers and perform both shower and event classification. In both cases, the classifiers were able to exploit higher-order effects to obtain good discrimination. Finally, we distinguished pairs of identical particles in back-to-back decays, where a marked increase in performance was seen, suggesting the classifiers had learned to effectively model this type of a particle decay. While there are still challenges and more work needed to bring this approach to data, our results indicate a significant potential of end-to-end classifiers for high-energy physics data analysis.

References

[1] de Oliveira L. et al., *Jet-Images—Deep Learning Edition*, JHEP **07** 069 (2016).

[2] Kasieczka G. *Deep-learning Top Taggers or The End of QCD?*, JHEP 2017:6 (2017).

[3] Aurisano A. et al.,*A Convolutional Neural Network Neutrino Event Classifier*, JINST **11** p. P09001 (2016).

[4] The CMS Collaboration, *Observation of the diphoton decay of the Higgs boson and measurement of its properties*, Eur. Phys. J. C **74**:3076 (2014).

[5] Baldi P. et al., *Searching for Exotic Particles in High-Energy Physics with Deep Learning*, Nature Communications **5** 4308 (2014).

[6] Louppe G. et al., *QCD-Aware Recursive Neural Networks for Jet Physics*, arXiv:1702.00748.

[7] Bhimji W. et al., *Deep Neural Networks for Physics Analysis on low-level whole-detector data at the LHC*. https://indico.cern.ch/event/567550/contributions/2629673/

[8] The CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST **3** p.S08004 (2008).

[9] The CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, JINST **12** P10003 (2017).

[10] LeCun Y. et al. *Gradient-based learning applied to document recognition*. Proceedings of the IEEE. **86** (11): 2278-2324 (1998).

[11] Hochreiter S., Schmidhuber J., *Long Short-term Memory*. Neural Computation. **9** 1735-80, (1997)

[12] Rumelhart D. *Learning representations by back-propagating errors*. Nature. **323** (6088): 533-536 (1986).

[13] Keras Team, "Keras" [software], version 2.0.3, Available from https://github.com/fchollet/keras [accessed 2017-04-17].

[14] Simonyan K., Zisserman A., *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv:1409.1556.

[15] Szegedy C., *Going Deeper with Convolutions*, arXiv:1409.4842.

[16] Kaiming H. et al., *Deep Residual Learning for Image Recognition*, arXiv:1512.03385.

[17] Shi X. et al., *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting*, arXiv:1506.04214.