End-to-end particle and event identification at the LHC with CMS Open Data

John Alison, Sitong An, Michael Andrews, Patrick Bryant, Bjorn Burkle, Sergei Gleyzer, Ulrich Heintz, Meenakshi Narain, Manfred Paulini, Barnabas Poczos, Emanuele Usai

APS DPF 2019 — 31 July 2019
Outline

**Introduction**: the end-to-end approach, building images

**previous work**: photon vs. electron, discriminator bias

**jet and event ID**: quark vs. gluon

**top tagging**: new layers configurations

**conclusions & outlook**
What is E2E

Train Particle/Jet/event IDs starting from low-level detector hits

End-to-end

Proof of concept, not a readily usable classifier.
Detector images

ECAL, energy deposits, 1 pixel per crystal

Δη x Δφ ~ 0.0174 x 0.0174
Detector images

ECAL

HCAL, energy dep, 1 px/tower (5x ECAL px)

tower-based $\Delta \eta \times \Delta \phi \sim 0.087 \times 0.087$
Tracks, pT weighted, at ECAL surface

ECAL

HCAL

Detector images
 CMS

<table>
<thead>
<tr>
<th>Detector Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>pT-weighted track positions</td>
</tr>
<tr>
<td>ECAL crystal deposits</td>
</tr>
<tr>
<td>HCAL tower deposits</td>
</tr>
</tbody>
</table>

Discretized X and Y

Discretized eta and phi

Composite image

Barrel

Endcap+

Endcap-

arXiv:1807.11916

px ~ 0.0174 x 0.0174

Δη x Δφ
Full-detector image

Event ID

<table>
<thead>
<tr>
<th>qq vs gg</th>
<th>Scenario A: 2 separate images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scenario B: A + dijet 4-momenta</td>
</tr>
<tr>
<td></td>
<td>Scenario C: full detector</td>
</tr>
</tbody>
</table>

**Scenario A:**
- ResNet-15, convolutional output
- Fully-connected, 128 x 2
- Event ID

**Scenario B:**
- 2 x jet images + jet 4-momenta
- ResNet-15, convolutional output
- Event ID

**Scenario C:**
- Fully end-to-end detector image
- Event ID

- **ROC AUC**
  - Scenario A: 0.876
  - Scenario B: 0.878
  - Scenario C: 0.889

**Performance dominated by jet-level differences**
- (Scenario A vs. B or C)

**Both dijets are non-resonant decays, so jet 4-momenta doesn’t hold much discrimination power**
- (Scenario B vs. A)

**Fully E2E approach (Scenario C) picking up on subtle, event-level effects not captured by either B or A.**
Scenario A: 2 x jet images
ResNet-15, convolutional output
Fully-connected, 128 x 2

Scenario B: 2 x jet images + jet 4-momenta
ResNet-15, convolutional output

Scenario C: Fully end-to-end detector image

Performance dominated by jet-level differences (Scenario A vs. B or C)
Both dijets are non-resonant decays, so jet 4-momenta doesn't hold much discrimination power (Scenario B vs. A)
Fully E2E approach (Scenario C) picking up on subtle, event-level effects not captured by either B or A.
Scenario A: 2 separate images

Scenario B: 2 x jet images + jet 4-momenta

Scenario C: fully end-to-end detector image

Local or global physics?

Performance dominated by jet-level differences (Scenario A vs. B or C)

Both dijets are non-resonant decays, so jet 4-momenta doesn’t hold much discrimination power (Scenario B vs. A)

Fully E2E approach (Scenario C) picking up on subtle, event-level effects not captured by either B or A.

Combine cropped images
This work uses 8 TeV CMS Open Data

- Essential to access full-simulation, low level detector information
- New release just published focusing on ML application
- Tracker clusters saved in high-level AOD format
- Can reconstruct tracker hits ("RecHits") on the fly

Full simulation stack reproducible in OD (with code & instructions)

Previous work

Full CMS detector simulation
Photon vs Electron

Particle ID I:
- Photon
- Electron

Particle ID II:
- $e^-$ vs $\gamma$
- $e^+e^-$ vs $\gamma\gamma$

ResNet-15
ROC AUC 0.788

ResNet-23
ROC AUC 0.997

https://arxiv.org/abs/1807.11916
Event ID: discriminator bias

Resonant diphoton, signal
Nonresonant diphoton
Photon + misidentified jet

Mass-biased discriminators
Decorrelation discriminators
Normalize images to gg mass or sum of pT

Final discriminator ROC

AUC 0.82 0.70 0.93

H→γγ vs Rest
H→γγ vs γγ component
H→γγ vs γ+jet component

https://arxiv.org/abs/1807.11916
Quark vs gluon jet

Full CMS detector simulation — open data

arxiv:1902.08276
Quark vs gluon jet
Quark vs gluon jet

Tracks

ECAL

HCAL

Gluon jet

Quark jet
Quark vs gluon jet

<table>
<thead>
<tr>
<th>Single subdetector</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E jet image, Tracks</td>
<td>0.782</td>
</tr>
<tr>
<td>E2E jet image, ECAL</td>
<td>0.760</td>
</tr>
<tr>
<td>E2E jet image, HCAL</td>
<td>0.682</td>
</tr>
</tbody>
</table>
### Single subdetector

| E2E jet image, Tracks          | ROC AUC | 0.782 |
| E2E jet image, ECAL           | ROC AUC | 0.760 |
| E2E jet image, HCAL           | ROC AUC | 0.682 |

### Subsystems combined

| E2E jet image, ECAL+Tracks    | ROC AUC | 0.804 |
| E2E jet image, Tracks         | ROC AUC | 0.782 |
| E2E jet image, ECAL+HCAL      | ROC AUC | 0.781 |
| E2E jet image, ECAL+HCAL+Tracks| ROC AUC | 0.808 |
Quark vs gluon jet

Single subdetector

<table>
<thead>
<tr>
<th>Subsystems combined</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E image, Tracks</td>
<td>0.782</td>
</tr>
<tr>
<td>E2E image, ECAL</td>
<td>0.760</td>
</tr>
<tr>
<td>E2E image, HCAL</td>
<td>0.682</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E image, ECAL+Tracks</td>
</tr>
<tr>
<td>E2E image, Tracks</td>
</tr>
<tr>
<td>E2E image, ECAL+HCAL</td>
</tr>
<tr>
<td>E2E image, ECAL+HCAL+Tracks</td>
</tr>
</tbody>
</table>

Comparison with RecNN

- RecNN, Jet ID for QCD vs boosted $W$ jet
  - DELPHES detector simulation
- Traditional jet images perform less well than 4-momenta
Qq vs gg

Scenario A: 2 separate images

ResNet-15, convolutional output
ResNet-15, convolutional output

Fully-connected, 128 x 2
**Qq vs gg**

### Scenario A: 2 separate images

- **ResNet-15, convolutional output**
- **Fully-connected, 128 x 2**

### Scenario B: A + dijet 4-momenta

- **ResNet-15, convolutional output**
- **ResNet-15, convolutional output**
- **Dijet 4-momenta**
- **Fully-connected, 128 x 2**
Qq vs gg

- **Scenario A:** 2 separate images
  - ResNet-15, convolutional output
  - Fully-connected, 128 x 2

- **Scenario B:** A + dijet 4-momenta
  - ResNet-15, convolutional output
  - Fully-connected, 128 x 2

- **Scenario C:** full detector
  - ResNet-15

**Local or global physics?**
- Performance dominated by jet-level differences (Scenario A vs. B or C)
- Both dijets are non-resonant decays, so jet 4-momenta doesn't hold much discrimination power (Scenario B vs. A)
- **Fully E2E approach (Scenario C) picking up on subtle, event-level effects not captured by either B or A.**

**ROC AUC**
- **Scenario A:** 0.876
- **Scenario B:** 0.878
- **Scenario C:** 0.889
Top quark ID

Full CMS detector simulation — open data
• ~ 5M top-antitop pair events

• Transverse momentum > 400GeV, |eta|<2.4

• Natural pT distribution from SM top-antitop

• Non-top jets sampled in from same momentum distribution as top quark
track position weighted by a different variable:

1) Transverse Momentum
2) d0 impact parameter
3) dz impact parameter
The CMS tracker

The CMS collaboration uses a right-handed coordinate system, with the origin at the centre of the detector, the x-axis pointing to the centre of the LHC ring, the y-axis pointing up (perpendicular to the plane of the LHC ring), and with the z-axis along the anticlockwise-beam direction. The polar angle $\phi$ is defined relative to the positive z-axis and the azimuthal angle $\eta$ is defined relative to the x-axis in the x-y plane. Particle pseudorapidity $\eta$ is defined as $\ln \left[ \tan \left( \frac{\phi}{2} \right) \right]$.

The CMS tracker [5] occupies a cylindrical volume 5.8 m in length and 2.5 m in diameter, with its axis closely aligned to the LHC beam line. The tracker is immersed in a co-axial magnetic field of 3.8 T provided by the CMS solenoid. A schematic drawing of the CMS tracker is shown in Fig. 1. The tracker comprises a large silicon strip tracker with a small silicon pixel tracker inside it. In the central pseudorapidity region, the pixel tracker consists of three co-axial barrel layers at radii between 4.4 cm and 10.2 cm and the strip tracker consists of ten co-axial barrel layers extending outwards to a radius of 110 cm. Both subdetectors are completed by endcaps on either side of the barrel, each consisting of two disks in the pixel tracker, and three small plus nine large disks in the strip tracker. The endcaps extend the acceptance of the tracker up to a pseudorapidity of $|\eta| < 2.5$.

![Tracking rechits diagram](image-url)
Jet Images - track pT (at ECAL surface)

low-pT tracks bending

Pixel detector very close to beam line
Must correct for PV position

The CMS tracker

The CMS collaboration uses a right-handed coordinate system, with the origin at the centre of the detector, the $x$-axis pointing to the centre of the LHC ring, the $y$-axis pointing up (perpendicular to the plane of the LHC ring), and with the $z$-axis along the anticlockwise-beam direction. The polar angle $\theta$ is defined relative to the positive $z$-axis and the azimuthal angle $\phi$ is defined relative to the $x$-axis in the $x$-$y$ plane. Particle pseudorapidity $\eta$ is defined as

$$\eta = \ln \left( \tan \left( \frac{\theta}{2} \right) \right).$$

The CMS tracker occupies a cylindrical volume 5.8 m in length and 2.5 m in diameter, with its axis closely aligned to the LHC beam line. The tracker is immersed in a co-axial magnetic field of 3.8 T provided by the CMS solenoid. A schematic drawing of the CMS tracker is shown in Fig. 1. The tracker comprises a large silicon strip tracker with a small silicon pixel tracker inside it. In the central pseudorapidity region, the pixel tracker consists of three co-axial barrel layers at radii between 4.4 cm and 10.2 cm and the strip tracker consists of ten co-axial barrel layers extending outwards to a radius of 110 cm. Both subdetectors are completed by endcaps on either side of the barrel, each consisting of two disks in the pixel tracker, and three small plus nine large disks in the strip tracker. The endcaps extend the acceptance of the tracker up to a pseudorapidity of $|\eta| < 2.5$. 

![Schematic drawing of the CMS tracker](image)
final image for a 628 GeV top jet.

list of the image channels:
1. Pixel layer 1 rechits
2. Pixel layer 2 rechits
3. Pixel layer 3 rechits
4. pT weighted tracks
5. d0 weighted tracks
6. dz weighted tracks
7. ECAL rechits
8. HCAL rechits
Summary

E2E Particle ID:
› Able to learn particle kinematics and shower shapes

E2E q vs. g Jet ID:
› Competitive with existing state-of-the-art jet ID classifiers
› E2E approach exploits full detector performance
› Event ID Captures event-level correlations lost at jet-level.

E2E Top ID:
› Work in progress
› Adding more tracking information and tracker rechits
› Increase resolution
› Expect results soon

https://arxiv.org/abs/1807.11916

https://arxiv.org/abs/1902.08276
Backup
Related Work

- CNNs in various neutrino experiments (A. Aurisano et al., see DS@HEP 2017, IML Workshop 2017, 2018)
- Particle ID CNNs on 4-momenta of jet constituents (Luke de Oliveira et al., see DS@HEP 2017, IML Workshop 2017, 2018)
- RNNs on 4-momenta of jet constituents a la Nat. Lang. Proc. (Kyle Cranmer et al., Jean-Roch Vlimant et al.)
- Particle ID CNNs on photon cluster detector data (Andre Holzner et al.)
- Event ID CNNs on whole detector images (Wahid Bhimji et al.)

**Our approach emphasizes high detector fidelity:**
True detector-level data, Geant4 detector sim, most accurate CMS model. Results representative of real physics analysis!

Credit: Michael Andrews
Particle ID I: Network

- Choose best-in-category for each of:
  - Convolutional NN (CNN): VGG, Inception, ResNet
  - Conv-LSTM (LSTM): TimeDist(CNN)→LSTM, LSTM(CNN)
  - Fully-Connected NN (FCN): 2-, 3-, 6-hidden layers, 256 nodes

- Try a variety of inputs:
  - energy, (energy, time), (DIGI), (energy, time, DIGI)*

- Try different concatenation schemes:
  - @input, @convolutional output, @FC output

Credit: Michael Andrews
Particle ID II: Network

- **VGG does not scale with image size**
  - Scales with CNN output *volume* x FCN nodes causing weights explosion!
  - VGGs also subject to *degradation* with increasing depth

- **Residual Nets scale much better**
  - Scales with CNN output *layers*, no need for FCN
  - Skip connections mitigate *degradation* with depth

Credit: Michael Andrews
Jet ID | quark vs gluon

- CMS OpenData QCD Samples
  - Leading jet from QCD dijet $qq'$ ($uds$) or $gg$, EMenriched @ 8 TeV
  - CMS GEANT4 full detector simulation, PYTHIA 6
  - $p_T$: 80-170 GeV, reco $p_T > 70$ GeV, $|\eta| < 1.8$
  - Run-dependent $\langle PU \rangle$ : 18-21
  - Produced and ntuplized with CMSSW 5_3_32

- Sample split:
  - Training set: 576k jets (of which, 26k jets for validation)
  - Test set: 139k jets
  - Balanced samples per class
  - Balanced PU representation per class

- Architecture: ResNet-15 trained from scratch on an NVIDIA Titan X/p using Pytorch 0.4

Credit: Michael Andrews