RNN Tau Identification

IN THE

ATLAS High-Level Trigger

Mariel Pettée
October 26th, 2018
US LUA Annual Meeting
Hadronic Tau Decay:

- Typically 1 or 3 tracks
- Highly collimated
- No surrounding tracks

Jet Background:

- Many tracks
- Less collimated
- Tracks in isolation region
BDT $\tau$ Identification
BDT Tau Identification

- **BDT =** Boosted Decision Tree
- **Inputs:** 12 “high-level” tau ID variables
- **Separate models for 1- and 3-prong taus**

### Table: ATLAS-CONF-2017-061

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{\text{cent}}$</td>
<td>Central energy fraction</td>
</tr>
<tr>
<td>$f_{\text{leadtrack}}$</td>
<td>Leading track momentum fraction</td>
</tr>
<tr>
<td>$R_{\text{track}}$</td>
<td>Track radius</td>
</tr>
<tr>
<td>$</td>
<td>S_{\text{leadtrack}}</td>
</tr>
<tr>
<td>$f_{\text{iso}}$</td>
<td>Fraction of $p_T$ from tracks in the isolation region</td>
</tr>
<tr>
<td>$\Delta R_{\text{MAX}}$</td>
<td>Maximum $\Delta R$</td>
</tr>
<tr>
<td>$S_{\text{flight}}$</td>
<td>Transverse flight path significance</td>
</tr>
<tr>
<td>$m_{\text{track}}$</td>
<td>Track mass</td>
</tr>
<tr>
<td>$f_{\text{track-HAD}}$</td>
<td>Fraction of EM energy from charged pions</td>
</tr>
<tr>
<td>$f_{\text{EM}}$</td>
<td>Ratio of EM energy to track momentum</td>
</tr>
<tr>
<td>$m_{\text{EM+track}}$</td>
<td>Track-plus-EM-system mass</td>
</tr>
<tr>
<td>$f_{\text{EM+track}}/p_T$</td>
<td>Ratio of track-plus-EM-system mass to $p_T$</td>
</tr>
</tbody>
</table>

1-prong
- $E_T / p_T$ of leading charged track
- Significance of leading track transverse impact parameter

3-prong
- Transverse flight path significance
- Maximum $\Delta R$
- Mass of charged tracks

Central energy fraction
- Mean track radius
- Momentum fraction of isolation tracks
- Momentum ratio of track + EM system
- Energy flow $p_T$ ratio
- Energy flow mass
BDT Tau Identification

After training, classify each tau candidate with a single score \( \in (0,1) \):

- 0 = background-like
- 1 = signal-like
BDT Tau Identification

Benefits of the BDT architecture:

- Highly-optimized classification
- Physical intuition & insight

But... are we throwing away information by only considering these select high-level variables?

Could adding low-level input variables into the mix improve our signal/background discrimination?
RNN $\tau$ Identification
RNN Tau Identification

- RNN = Recurrent Neural Network
- Approaches $\tau$ identification as a **sequence classification** problem
- **Inputs:**
  - **Track-level variables** (for at most 10 tracks; full list in backup slides)
  - **Cluster-level variables** (for at most 6 clusters; full list in backup slides)
  - BDT input variables

![Reconstructed hadronic tau decay](Image: C. Deutsch)
RNN Tau Identification

Tracks + Clusters + BDT Input Variables =

Dense layer with shared weights

Reconstructed hadronic tau decay

Dense layer with shared weights

Reconstructed hadronic tau decay

Dense layer with shared weights

Reconstructed hadronic tau decay

BDT

Input Variables

Tracks

Clusters

High-level ID variables

Shared dense

Shared dense

Shared dense

Shared dense

LSTM

LSTM

LSTM

Concatenate

Dense

Dense

Dense

Dense

N_{parameters} \approx 56,000

Tau-ID BDT: 38k

Images: C. Deutsch
RNN Tau Identification

- Shared dense & LSTM (Long Short-Term Memory) layers preserve contextual information from multiple tracks/clusters to improve decision-making.
RNN Tau Identification

- BDT $\tau$ identification: **1-prong** and **3-prong taus**

- RNN $\tau$ identification: **0-prong**, **1-prong** and **$\geq2$-prong taus**

**Goal:** recover true **1p taus** for which the charged track has been poorly reconstructed, especially at low-$p_T$ and high-$\mu$.

**Goal:** recover true **3p taus** for which at least one charged track has been mis-reconstructed.
RNN Tau Identification: Online

- Training samples:
  - $\gamma^{*}\rightarrow\tau\tau$ and dijets
  - Fewer statistics than for offline, but same # of trainable parameters
  - **0-prong:** Signal $\sim$100k events, Background $\sim$50k events
  - **1-prong:** Signal $\sim$2M events, Background $\sim$175k events
  - **≥2-prong:** Signal $\sim$700k events, Background $\sim$3M events

- Reweight signal $p_T$ spectrum to match background $p_T$ spectrum for determining same-rejection working points as the BDT (since *trigger rates are dominated by low-$p_T$ jets*).
RNN Tau Identification: Outlook

- RNN tau identification **successfully deployed** in the ATLAS high-level trigger in mid-2018! (Public results forthcoming)

- Will likely be the default ATLAS tau identification algorithm in the Run 3 trigger

- Until then:
  - Continue optimizing models
  - Investigate RNN input variable modeling
  - Consider training on data rather than MC

![Diagram of RNN model with LSTM layers and dense layers, showing concatenation of tracks and clusters with high-level ID variables.]
Backup Slides
# of tracker (pixel, SCT) hits → # of tracker hits + # of dead sensors
  • Designed to **protect against varying detector conditions**

# of IBL hits → If a hit is expected, use actual # of IBL hits. If not, set # IBL hits = 1.
  • If # of IBL hits were set to 0 in the latter case, might be wrongly classified as a bad track

---

**- Track-level variables**

- pt_log
- pt_jetseed_log
- d0_abs_log
- z0sinThetaTJVA_abs_log
- dEta
- dPhi
- nIBLHitsAndExp
- nPixelHitsPlusDeadSensors
- nSCTHitsPlusDeadSensors

**- Cluster-level variables**

- et_log
- pt_jetseed_log
- dEta
- dPhi
- SECOND_R
- SECOND_LAMBDA
- CENTER_LAMBDA
Tau ID RNN Input Variables

Central energy fraction ($f_{\text{cent}}$): Transverse energy at EM scale deposited in calorimeter cells with a barycentre in a cone of radius $\Delta R < 0.1$ divided by the transverse energy of cells in a cone of radius $\Delta R < 0.2$ with respect to the reconstructed tau axis. For noise suppression the calorimeter cells must be part of a TopoCluster.

Inverse momentum fraction of the leading track ($f_{\text{lead}}^s$): Transverse energy at EM scale deposited in calorimeter cells (as part of TopoClusters) with a barycentre in a cone of radius $\Delta R < 0.2$ with respect to the tau axis divided by the transverse momentum of the highest transverse momentum track classified as charged according to the track classification.

Track radius ($R_{\text{track}}$): Mean $\Delta R$-distance of tracks classified as charged and the tau axis weighted by the transverse momentum of each track.

Maximum track $\Delta R$ ($R_{\text{track,max}}$): Maximum $\Delta R$-distance of all tracks classified as charged with respect to the tau axis. Equivalent to $R_{\text{track}}$ for 1-track $\tau_{\text{had-vtx}}$.

Transverse impact parameter significance of the leading track ($S_{\text{lead}}^{\text{track}}$): Absolute value of the transverse impact parameter of the leading charged track with respect to the reconstructed primary vertex divided by its uncertainty estimate from the track and vertex fit.

Transverse flight path significance ($S_{\text{flight}}^{\text{track}}$): Distance between the secondary vertex reconstructed using tracks classified as charged and primary vertex in the transverse plane divided by the estimated uncertainty from the secondary vertex fit. Defined only for multi-track $\tau_{\text{had-vtx}}$.

Momentum fraction of isolation tracks ($f_{\text{iso}}^{\text{track}}$): Sum of transverse momenta of modified isolation tracks (cf. Section 3.2) divided by the sum of transverse momenta of modified isolation and charged tracks.

EM energy fraction of charged pions ($f_{\text{EM}}^{\text{charged}}$): Energy deposited by charged pions in the electromagnetic part of the calorimeter estimated by subtracting the energy contained in the hadronic part of the calorimeter (in TopoClusters) from the energy of the track system, consisting of tracks classified as charged, estimated by the scalar sum of track momenta (assuming zero mass). This energy is divided by the energy contained in the electromagnetic part of clusters associated with the jet seeding the $\tau_{\text{had-vtx}}$. All cluster energies are calibrated at LC scale.

Transverse impact parameter significance of the leading track ($S_{\text{lead}}^{\text{track}}$): Absolute value of the transverse impact parameter of the leading charged track with respect to the reconstructed primary vertex divided by its uncertainty estimate from the track and vertex fit.

Transverse flight path significance ($S_{\text{flight}}^{\text{track}}$): Distance between the secondary vertex reconstructed using tracks classified as charged and primary vertex in the transverse plane divided by the estimated uncertainty from the secondary vertex fit. Defined only for multi-track $\tau_{\text{had-vtx}}$.

Momentum fraction of isolation tracks ($f_{\text{iso}}^{\text{track}}$): Sum of transverse momenta of modified isolation tracks (cf. Section 3.2) divided by the sum of transverse momenta of modified isolation and charged tracks.

EM energy fraction of charged pions ($f_{\text{EM}}^{\text{charged}}$): Energy deposited by charged pions in the electromagnetic part of the calorimeter estimated by subtracting the energy contained in the hadronic part of the calorimeter (in TopoClusters) from the energy of the track system, consisting of tracks classified as charged, estimated by the scalar sum of track momenta (assuming zero mass). This energy is divided by the energy contained in the electromagnetic part of clusters associated with the jet seeding the $\tau_{\text{had-vtx}}$. All cluster energies are calibrated at LC scale.
Tau HLT in 2018

**CALORIMETER CLUSTERING**

- **Inclusive MVA TES:**
  MVA Tau Energy Scale calibration applied to all tau candidates

- **Minimum \( p_T \) cut**

**INITIAL TRACKING**

- **“Fast Track Finder” (FTF):**
  Look for a track in a narrow cone (\( \Delta R < 0.1 \)) around the center of the cluster & along the full beam line

**PRECISION TRACKING**

- **Track refit:**
  Run precision tracking on FTF seeds

- **nTracks cut**

**TAU IDENTIFICATION**

- **RNN**
  0-/1-/multi-prong RNN with tau ID variables + track + cluster variables as inputs

- **nTracks Cut:** only pass taus with...
  - 1-3 tracks in core region
  - \( \leq 1 \) tracks in isolation region