Simulated Data

Hyper-K Machine Learning Workshop

Summary

All information on: mlw.hyperk.ca/resources/simulated-datasets

Variety of datasets produced for the IWCD detector:

Two different detector geometries:

- Regular 88x168 grid of 3” PMTs
- 16x40 grid of multi-PMT modules each with 19 3” PMTs

Various different datasets for each:

- Varying energy
- Varying position
- Varying everything

For different possible group projects:

- Classification of particle types
- Regression of energy, position, direction
- etc.

Available datasets

1. IWCD geometry with grid of 3" PMTs
    a. Varying energy and direction
        - e-, mu-, gamma, 1,000,000 events each
        - Evis between 20 MeV and 2 GeV
        - Centred position & vertical direction, varying direction around axis
    b. Varying R position and direction
        - e-, mu-, gamma, 1,000,000 events each
        - Fixed visible energy 200 MeV, fixed vertical direction
        - Varying radial position and direction aroung axis
    c. Varying everything
        - e-, mu-, gamma and pi0, 1,000,000 events each
        - Varying energy (20 MeV to 2 GeV), position (≥50cm from wall), direction (isotropic)

Available datasets

2. IWCD geometry with multi-PMT modules
    a. Fixed energy, position and direction
        - e-, mu- and gamma, 100,000 events per configuration
        - Fixed visible energy 200 MeV
        - Fixed positions centred vertically, R = {0, 100, 200, 275, 325} cm
        - Direction along +ve x-axis
    b. Varying everything
        - e-, mu- and pi0, 1,000,000 events each
        - Uniform random energy between: 30 MeV to 1 GeV for e, 200 MeV to 1.2 GeV for mu, 100 MeV to 1 GeV for neutral pi0
        - Position varys anywhere in tank
        - Isotropic varying direction

Available datasets - comments

- For all gamma simulations, the simulated position is the position of gamma conversion
- Only the barrel PMTs data are included
- Only the trigger with highest number of hits is stored
- Dark noise is included at 100 Hz at each PMT
- All datasets/configurations provided as collection of small npz files or large merged hdf5 files
- First 100,000 events of every configuration of every dataset has existing reconstruction (fiTQun) for comparison
  - For configs with 1M events, there are separate datasets with and without reconstruction (_100evts and _1000evts for .npz file or _100k and _1M for .h5 files)
- Number of events may not be exactly as advertised due to removal of ‘bad’ events (e.g. no PMT hits)
- File locations are documented here: https://mlw.hyperk.ca/resources/simulated-datasets (or just look inside /data/hkml_data/IWCD*)
- Scripts used to produce data / convert formats will be committed to WatChMaL Github repo

Data format

All data stored as python numpy arrays:

- event_data:
  - Time and charge observed at each PMT
  - Shape: (N, 88, 168, 2) for grid data or (N, 16, 40, 38) for mPMT data
- labels:
  - The type of particle (0 for gamma, 1 for e-, 2 for mu-, 3 for pi0) being simulated
  - Shape: (N,)
- pids:
  - PDG code of each particle simulated
  - Shape: (N, P)
  - P is 1 for e-, mu- and pi0 but 2 for gamma (because simulated as an electron-positron pair)
- energies:
  - Energies of each particle simulated
  - Shape (N, P)
- positions:
  - (x,y,z) positions of each particle simulated
  - Shape (N, P, 3)
- directions:
  - (x,y,z) directions of each particle simulated
  - Shape (N, P, 3)

fiTQun data

- Sored as npz files
- Direct conversion of standard fiTQun ROOT output to numpy arrays
- One numpy array for each branch of fiTQun tree
  - Shape is determined by dimensions of each branch, first dimension is number of events
- fiTQun variable names are not entirely intuitive
  - fqnse: number of subevents
  - fq1rmom[events][fqnse][7]: single ring momentum
  - fq1rpos[events][fqnse][7][3]: single ring position
  - fq1rdir[events][fqnse][7][3]: single ring direction
  - fq1rnll[events][fqnse][7]: -Log(L) of fit
  - The [7] is for different particle hypothesis: 1=electron, 2=muon
  - Many (>100) more variables
- If you don’t understand the output and want to use it, ask around: several fiTQun experts in this workshop!
- Files are in /data/hkml_data/fiTQun/IWCD*

Possible projects using this data

Data sets provided intended for variety of possible projects we had in mind:

- Simple classification of particle type using highly constrained data (fixed energy/position/etc)
- More complicated classification with everything varying
- Simple regression of one variable with others fixed
- More complex regression with multiple varying quantities
- Dealing with more complex geometries (mPMT)
- Anything else you think of
  - Create hybrid datasets and separate the hits from multiple rings?
  - Add random noise and try to deal with that?
  - Introducing ‘dead pixels’
  - …

Some limitations of the data

- Doesn’t include end-cap PMTs
- No low-E events (minimum we have is 20 MeV)
- If you want to work on these, *might* be able to generate data overnight to accommodate, but won’t be available on every instance