Bridging The Gap: Developing Data-Preprocessing Pipelines to Enable Advanced Cosmic-Ray Reconstruction with IceTop-CNN

Johnathan Stowers, Dr. Frank McNally
Department of Physics, 1501 Mercer University Dr., Macon, GA 31207

Introduction

The IceCube Neutrino Observatory captures around 3,000 cascading particle showers every second. Mercer students have developed IceTop-CNN, which takes these events and uses a convolutional neural network (CNN) to estimate the energy, trajectory, and mass composition of the cosmic rays that cause these particle showers. Each event represents one frame, and collections of frames are stored in i3 files. We needed to develop a new data preprocessing pipeline to read these files, extract the charge, pulse, and other relevant data frames, and store them in a format compatible with IceTop-CNN. This work allows access to additional detector simulation data not previously usable to Mercer student researchers.

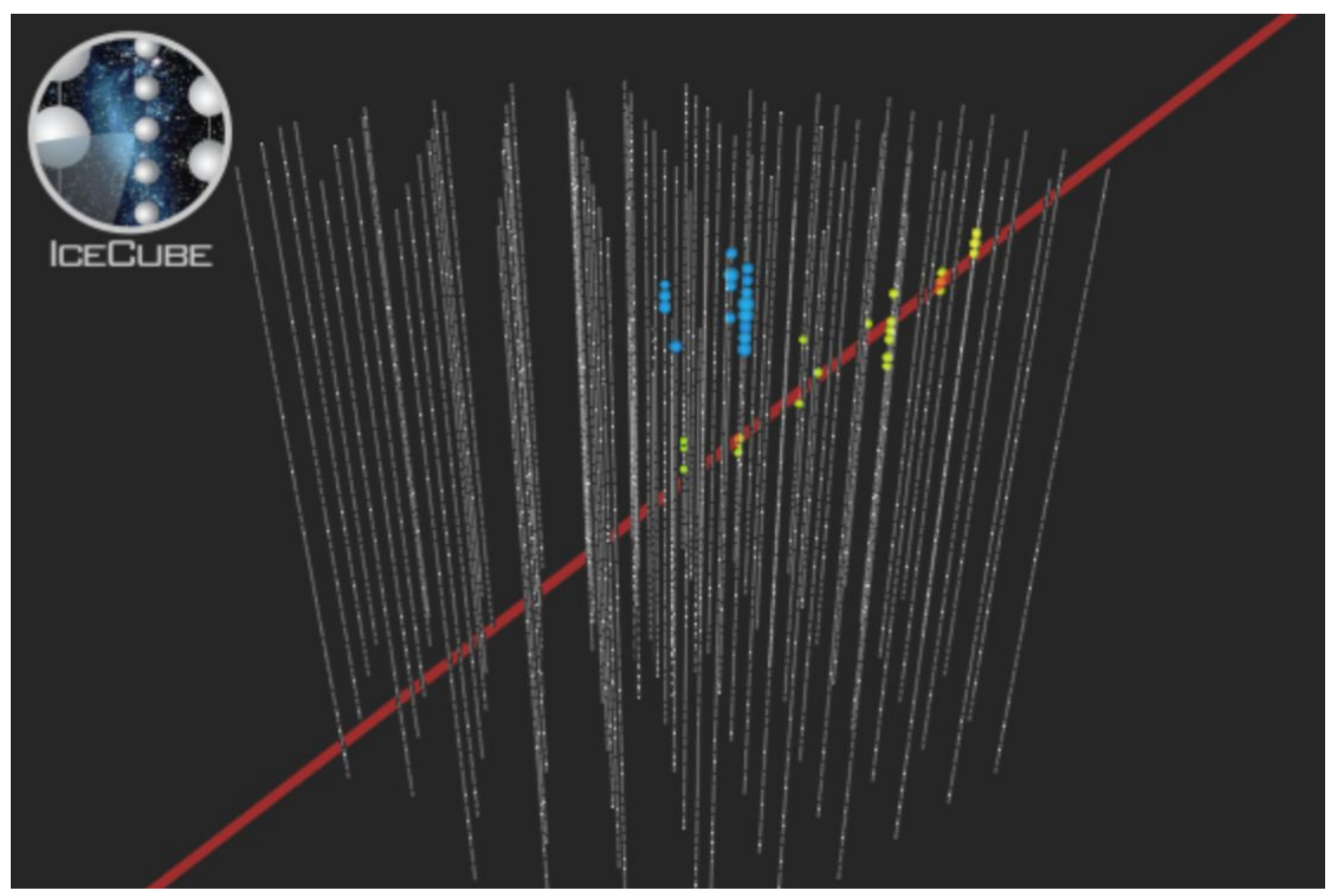


Figure 1: Vizualization of an in-ice event stored in an I3 file using SteamShovel, IceTop specific software. *Credit: IceCube Summer School*

To train IceTop-CNN, data is extracted into NumPy dictionary array files (.npy) that separate detector signals from the simulated properties. The pipeline produces two different folders of these .npy files, one for the detector hits (Figure 2) and one for the true simulated reconstruction parameters (Figure 3). The hit files produce a localized spatial footprint of the shower. These act as the input for IceTop-CNN. The truth files include a record of the cosmic ray's true energy. direction, and mass. These act as the evaluator of IceTop-CNN's outputs.

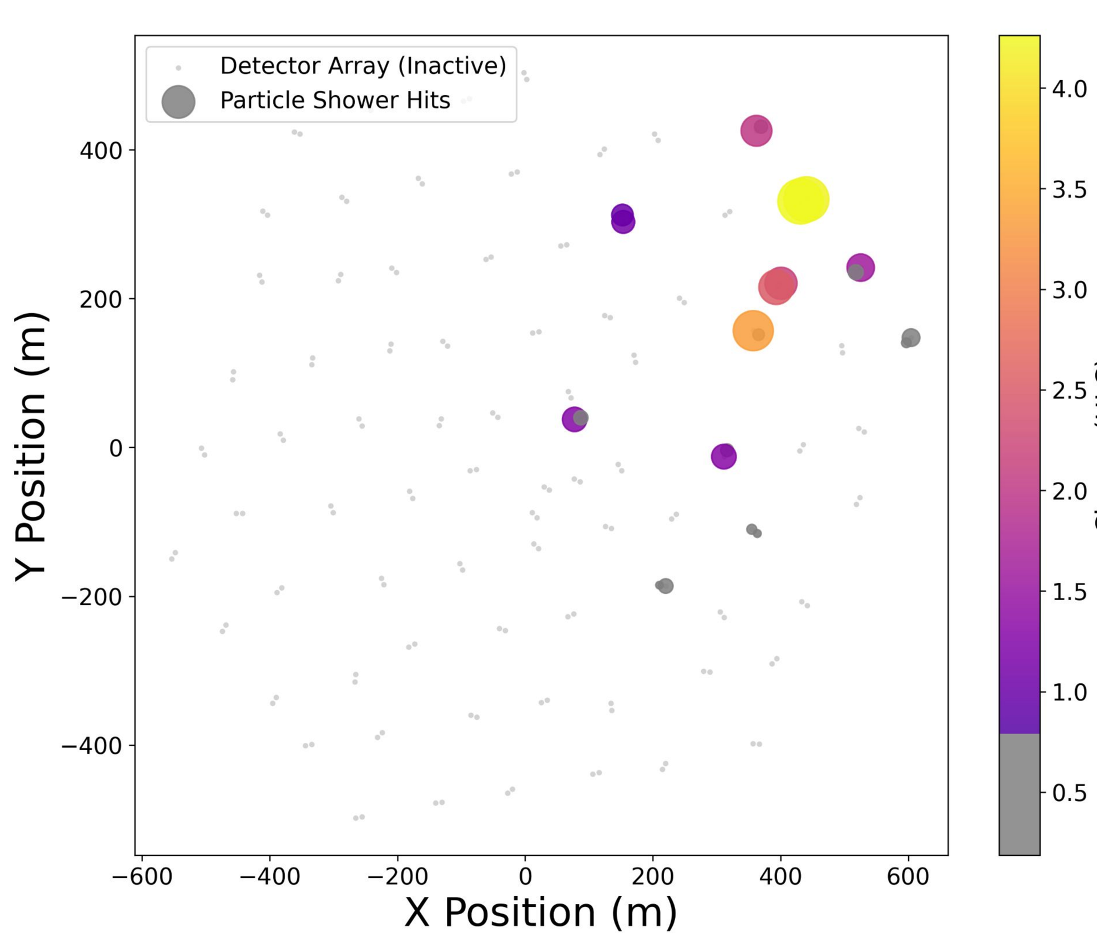


Figure 2: Visualization of the .npy file that gets produced by the preprocessing pipeline Data from 2012/SIBYLL2.1/Fe/12362_v0/5.0

Phase 1: Signal Extraction and Shower Kinematics

Cherenkov Pulse Extraction (Figure 2)

- Extracts charge and time values from the IceCube detector array
- Charge corresponds to the amount of Cherenkov light emitted by the particles passing through the IceTop ice tanks
- Time corresponds to the arrival kinematics of the shower front

Hard Local Coincidence vs Soft Local Coincidence Filtering

- Hard Local Coincidence (HLC) Events require neighboring tanks to also get a hit to be registered, and Soft Local Coincidence (SLC) Events are independent of neighbors
- A shower with more HLCs have a higher confidence of the physical shower's footprint, whereas SLCs could introduce noise that doesn't impact the prediction of the inital cosmic ray

Reconstruction Benchmarks

- IceTop-CNN's directional reconstruction can be evaluated against existing analytic models. We test against a simple fit that assumes the shower front is a plane, and a more complex likelihood-based fit that accounts for the curvature of the shower front

Trigger and Quality Cuts:

- Filters ensure the event had enough of a footprint to allow for a reconstructable air shower, safe from noise

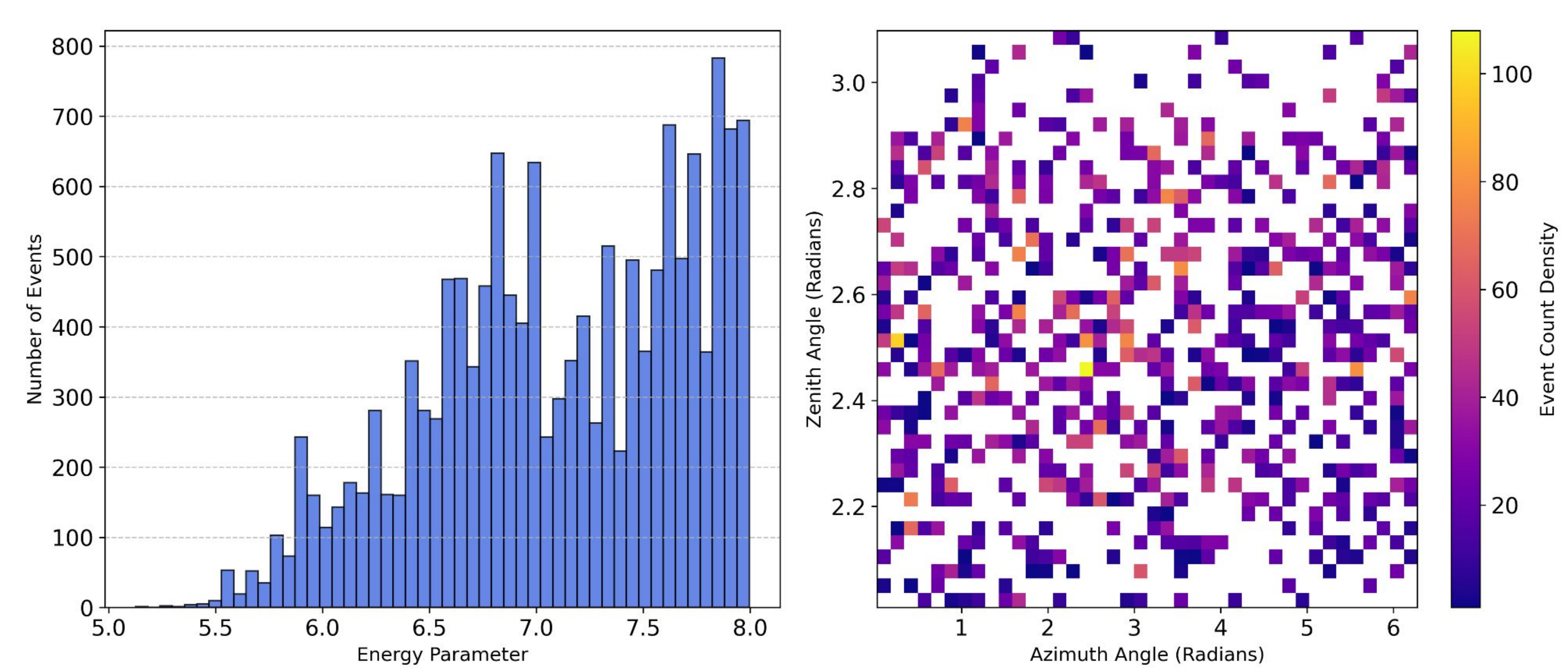
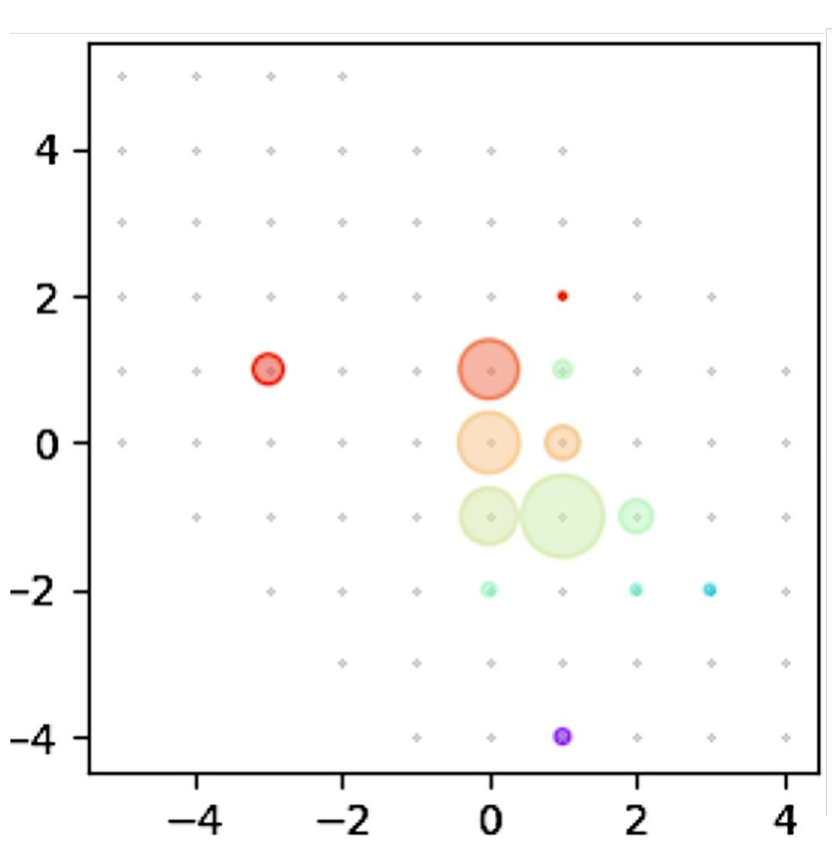


Figure 3: Visualization of the reconstruction parameters extracted from the data pipeline Energy Distribution and the True Arrival Directions



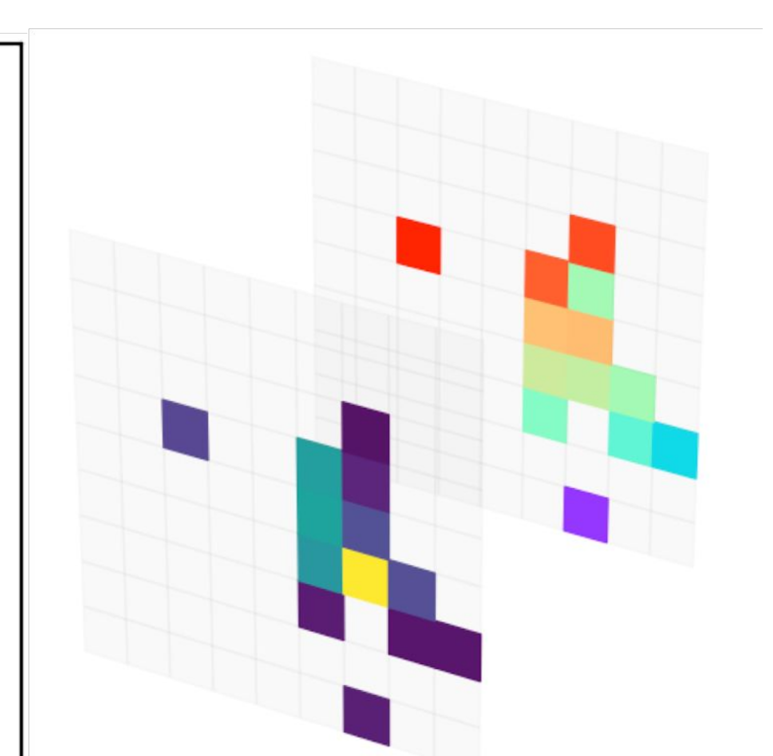


Figure 4: Visualization of the hexagonal grid mapped to a Cartesian grid
Figure 5: Visualization of the data split into time and charge layers
Credit: Dr. McNally

Phase 2: Spatial-Temporal Matrix Translation

Hexagonal to Cartesian Mapping (Figure 4)

- IceTop is in a hexagonal array, whereas IceTop-CNN requires a Cartesian Grid to allow for training
- The pipeline converts the hexagonal grid into a uniform 2D pixel matrix without majorly impacting the spatial relationship of the tanks hit by the shower

8-Channel Deep Tensors (Figure 5)

- Compresses the physical tank responses into an 8-depth tensor array that represents each station's two tanks split by their coincidence status, charge, and time measurements
- Later compressed to include both HLC and SLC (CLC) events for training

Logarithmic Energy Scaling

- Cosmic-ray energies span many orders of magnitudes, a base-10 log is applied
- This normalizes data and ensures CNN training is based on relative energy scaling instead of focusing on incredibly rare outliers

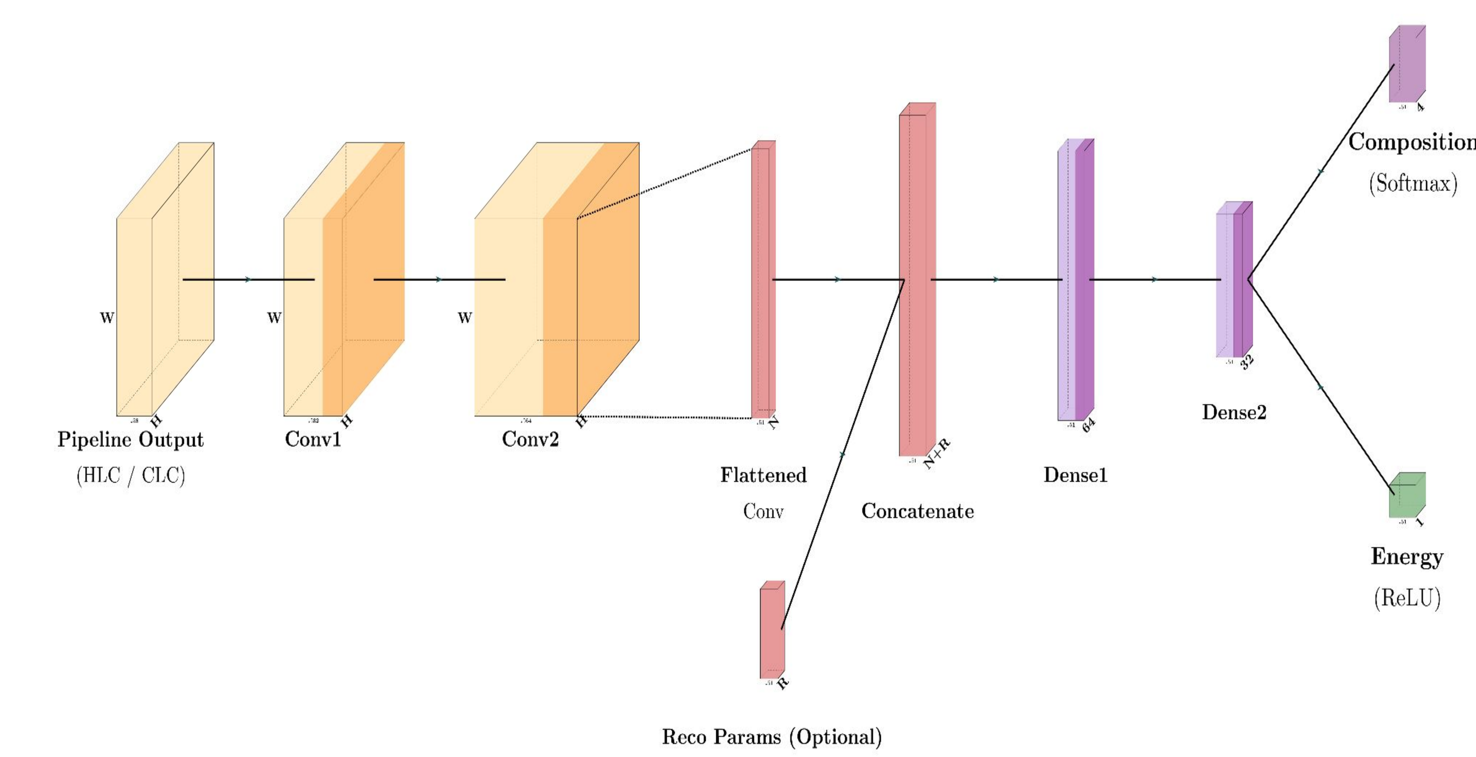


Figure 6: Baseline Convolutional Neural Network Architecture
All convolutional and dense layers are followed by Batch Normalization and RELU Activation

Conclusion and Future Plans:

- The pipeline will be used to allow for more, richer data to be used in model training to vastly improve IceTop-CNN's accuracy
- Research previously unreachable to IceTop-CNN is now possible. New options include:
  - High-energy events
  - Performance over time (changing snow heights)
  - Dependence on particle interaction model
  - Joint in-ice/IceTop reconstructions
- It also provides a fundamental framework for the preprocessing and reconstruction of experimental data

Acknowledgements:
I would like to extend a special thanks to Dr. McNally, Dr. Digh, Ethan Dorr, Rajwol Chapagain, Tommy Futrell, Olivia Lee, Thaddius Brackin, my fellow student researches, and everyone else who's supported me thus far.