Reconstructing Neutrino Energy using CNNs for GeV Scale IceCube Events

The IceCube Collaboration
(a complete list of authors can be found at the end of the proceedings)

E-mail: micall12@msu.edu

Measurements of neutrinos at and below 10 GeV provide unique constraints of neutrino oscillation parameters as well as probes of potential Non-Standard Interactions (NSI). The IceCube Neutrino Observatory’s DeepCore array is designed to detect neutrinos down to GeV energies. IceCube has built the world’s largest data set of neutrinos >10 GeV, making searches for NSI a computational challenge. This work describes the use of convolutional neural networks (CNNs) to improve the energy reconstruction resolution and speed of reconstructing $O(10 \text{ GeV})$ neutrino events in IceCube. Compared to current likelihood-based methods which take seconds to minutes, the CNN is expected to provide approximately a factor of 2 improvement in energy resolution while reducing the reconstruction time per event to milliseconds, which is essential for processing large datasets.

Corresponding author: Jessie Micallef$^1$

$^1$ Michigan State University
$^*$ Presenter
1. Introduction

The past two decades have seen exciting advancements in both fundamental physics and astrophysics with neutrinos. GeV-scale measurements have demonstrated neutrino oscillations and nonzero neutrino masses [1, 2], while observations above 1 TeV indicate the existence of a population of high-energy neutrinos of astrophysical origin [3, 4]. However, neutrinos with energies between 10 GeV and 1 TeV remain relatively unexplored. Measurements at these energies are nontrivial, since 10 GeV is beyond the reach of accelerator and reactor neutrino experiments. To detect atmospheric neutrinos above 10 GeV, detectors must be dense enough to reconstruct the neutrino interactions while also being large enough to compensate for the steeply falling atmospheric neutrino spectrum. This is a considerable experimental challenge that IceCube is uniquely suited to handle.

IceCube, located at the South Pole, instruments a cubic kilometer of ice between depths of 1450 m and 2450 m [5]. It relies on the optical detection of Cherenkov radiation, emitted by charged particles produced from neutrino interactions in the ice or nearby bedrock, using 5160 Digital Optical Modules (DOMs). The DOMs are attached on cables called “strings”, which are arranged in a hexagonal array with ~ 100 m horizontal spacing, with 60 DOMs per string vertically separated by 17 m (Figure 1). The DeepCore subarray includes 8 densely instrumented strings, separated horizontally by ~ 72 m spacing and with 7 m of vertical space between the DOMs, optimized for energies as low as about 10 GeV [6]. Utilizing DeepCore to detect 10 GeV neutrinos can provide us with a window into a largely unexplored regime to study the oscillation of atmospheric neutrinos [7]. Neutrino oscillation probabilities are a function of the ratio of the distance traveled by the neutrino to its energy, \( L/E \). Therefore, improving the energy reconstruction resolution in IceCube can directly advance the ability to constrain neutrino oscillation parameters.

2. Improving the Energy Reconstruction with a Convolutional Neural Network

The IceCube DOMs record the Cherenkov radiation from neutrino interactions, which can happen anywhere in the detector volume, meaning the events are characterized by translational invariance. This property allows us to apply Convolutional Neural Networks (CNNs) to the event reconstruction. CNNs are often used in image recognition as they are able to identify an object independently of where the object is positioned within an image. [8, 9]. In a CNN, a set of filters
or 'kernels' is used to build maps of spatial features from the input data. By combining several convolutional layers, it is possible to reconstruct complex images. In IceCube, this architecture can be used to build a reconstruction using a network of convolutional layers (Figure 2). This CNN contains 8 convolutional layers which are split into two identical branches applied separately on the 8 DeepCore strings and 19 neighboring IceCube strings, since their DOMs have different vertical spacing. The two convolutional branches are then combined to train one fully connected layer on all 27 string inputs to give the final energy prediction.

![Figure 2: Detailed architecture for the Convolutional Neural Network, legend for color layers on the right.](image)

This CNN architecture is similar to the CNN applied for IceCube’s high energy reconstruction [10]. In IceCube, low energies are particularly difficult to reconstruct due to the sparseness of the detector. Low-energy events also produce less scattered light, which reduces the information available for reconstruction and makes the event topologies less distinctive. Thus, the low-energy CNN includes key optimizations from the high-energy CNN reconstruction.

The low-energy CNN uses only the DeepCore strings and the center-most IceCube strings (Figure 3). This accounts for the first dimension in the input array, which corresponds to the string index. All 8 DeepCore strings and the center-most 19 IceCube strings are used as input. The CNN kernel spans the vertical, or z-depth, of the strings only (Figure 4); no convolution is applied in the xy-plane since the DeepCore strings are deployed in an irregular array. The second dimension in the input array uses all 60 DOMs on both the DeepCore and IceCube strings. The last dimension corresponds to the 5 input variables that summarize the (potential) multiple hits that a DOM would record in an event. These are the sum of the charge, time of the first hit, time of the last hit, charge
weighted mean of the hits, and charge weighted standard deviation of the hits.

Figure 3: Top view of the IceCube detector strings. DeepCore shown in red. Strings highlighted in blue are used for CNN.

Figure 4: Low energy CNN kernel spans the z-direction only, traveling down the DOMs on a string.

3. Training Sample and Data Preparation

A sample of more than 6 million unweighted, charge-current (CC) $\nu_\mu$ Monte Carlo (MC) events was generated in GENIE [11], a neutrino MC generator with an emphasis on neutrino interaction physics at the few-GeV energy range, specifically for training the CNN. Of that sample, 80% of the events (~ 5 million events) are used to train the CNN and 20% (~ 1 million events) are used during training as a validation set to ensure the network does not overtrain, or overfit, the training sample. A potential problem when training the CNN on the atmospheric neutrino sample is that it will overtrain on low-energy showers, since the neutrino spectrum is steeply falling. To reduce such biases in the reconstruction, the MC training sample uses a flat distribution in energy between 1 and 200 GeV. The training sample is extended past the target reconstruction region up to 500 GeV so that the CNN learns to reconstruct higher energy events.

Figure 6 shows the progression of the training and validation loss as the network trains. Network training proceeds by minimizing the loss, defined here as the mean absolute percent error between the reconstructed and true energy of the event, chosen so that the CNN values accurately reconstructing low-energy as much as high-energy events. The loss function is iteratively minimized in discrete training “epochs,” or full passes through the training sample, until it reaches a plateau. Dropping the learning rate ($\text{lr}_{\text{ate}}$) in Figure 6 helps with the instability of the validation sample while training. The performance of the CNN is evaluated after 11 epochs.

4. Resolution and Timing Performance

The CNN performance is evaluated on an independently generated GENIE neutrino sample with ~ 4 million unweighted charge-current $\nu_\mu$ events and ~ 1.6 million unweighted charge-current $\nu_e$ events. These distributions, weighted by the atmospheric flux, appear more like the expected energy distribution for data (Figures 7). This tests both the CNN’s robustness to reconstruct a sample
with a different distribution than its training sample and provides a more accurate estimation of the resolution for its potential to reconstruct low energy IceCube data. While the sample extends to 10 TeV, the performance evaluation will focus mostly on the 1-200 GeV region since the majority of the sample is expected at these energies and this is also the target region for an oscillation analysis.

Figure 7: The $\nu_\mu$ CC and $\nu_e$ CC testing samples: MC GENIE simulation with atmospheric flux weights applied to match data.

Figure 8 shows the CNN resolution compared to the true neutrino energy on the left, with the same comparison for IceCube’s current likelihood-based reconstruction on the right. The CNN’s median follows the ideal 1:1 line up to $\sim$ 150 GeV, then starts to slightly underestimate higher energy events. The likelihood-based reconstruction does not have the same underestimate, but the 68% band is wider at higher energies. Improvements to the CNN are expected once more high energy training events are generated above the target region (200-500 GeV). Neural networks are known for having difficulty extrapolating data [12]; populating training data beyond the target region is one solution to improve resolution near the target boundary regions.

Figures 9 and 10 show a direct comparison between the CNN and likelihood-based energy
reconstruction performance on the $\nu_\mu$ CC sample. Figure 9 shows the overall fractional resolution on the entire testing sample and Figure 10 explores the fractional reconstruction as a function of the true energy. The CNN has an advantage at the lowest energies with its median bias offset near zero and 68% band well resolved. This is particularly important since the expected energy resolution for the data will have the highest statistics in this low energy region. This explains the better overall 1$\sigma$ spread in Figure 9 for the CNN, since the $\nu_\mu$ CC performance sample is low-energy dominated.

For the $\nu_e$ CC events, the CNN slightly outperforms the likelihood-based reconstruction at all energies (Figure 11). Instead of the average underestimation of the $\nu_\mu$ CC energy in the 150-200 GeV region, the CNN has a slight overestimation in this region which matches the likelihood-based reconstruction’s median trend. Since the network is trained on $\nu_\mu$ CC events only, it only sees the hadronic component of the cascade, which has less light per unit energy than in electromagnetic cascades. This could be the cause for the network to overestimate on average the $\nu_e$ CC events.

The direct comparisons for the CNN and likelihood-based energy reconstruction for the $\nu_e$ sample are shown in Figure 12 and 13. The 1$\sigma$ spread of the overall sample is better for the CNN (Figure 12). The CNN also has minimal offset bias for the fractional resolution at the lowest

---

**Figure 8:** CNN (left) and likelihood-based (right) reconstructed vs true $\nu_\mu$ CC energy resolution. The median for each x-bin of true energy is shown in the solid red line with the 68% band in the dotted red line.

**Figure 9:** Fractional resolution of the reconstruction methods on the $\nu_\mu$ CC sample.

**Figure 10:** Fractional energy resolution as a function of the true energy on the $\nu_\mu$ CC sample.
Reconstructing Neutrino Energy using CNNs

Figure 11: CNN (left) and likelihood-based (right) reconstructed vs true $\nu_e$ CC energy resolution. The median for each x-bin of true energy is shown in the solid red line with the 68% band in the dotted red line.

energies, improving on the likelihood-based estimation in this region (Figure 13).

Figure 12: Fractional resolution of the reconstruction methods on the $\nu_e$ CC sample.

Figure 13: Fractional energy resolution as a function of the true energy on the $\nu_e$ CC sample.

The reconstruction methods’ runtimes are extremely important for the IceCube oscillations analysis, which expects a high-statistics sample ($\sim$ 650,000 neutrinos) collected over 10 years of livetime. Neural networks are known for their quick evaluation times, particularly when using GPUs. Table 1 shows the timing for the CNN and the likelihood-based reconstructions, with the CNN evaluated both on a CPU and a GPU. Note that the likelihood-based method reconstructs a total of 8 variables simultaneously, while the CNN is currently only returning a single variable. Using the CNN to reconstruct more variables will only change the numbers quoted in Table 1 by less than a factor of 10.

IceCube has access to computer clusters with both GPU and CPU nodes. Utilizing Michigan State University’s High Performance Computing Cluster’s GPUs yielded results for the performance set in about 15 minutes. The expected size of the data analysis set is estimated to be about 1/8 of the size of the performance MC set, so that reconstruction should be complete in less than 5 minutes.
Reconstructing Neutrino Energy using CNNs

| Method              | Average Time per Event (seconds) | Events per Day per Single Core |
|---------------------|----------------------------------|-------------------------------|
| CNN on GPU          | 0.0077                           | 11,000,000                    |
| CNN on CPU          | 0.27                             | 320,000                       |
| Likelihood-Based    | 40                               | 2,100                         |

Table 1: Comparing the time each method takes to reconstruct neutrino events.

5. Conclusion and Outlook

Reconstructing the energy of 10-GeV scale IceCube neutrino events with a CNN provides comparable resolution to the current likelihood-based reconstruction but with much shorter runtimes. The low-energy CNN is still being optimized, in particular the network configuration and settings. Additional work is being conducted to explore the possibility of using a similar network structure to reconstruct the direction of the neutrino [13] and interaction vertex, along with classifications for particle identification and background rejection. Using the CNN for these purposes will provide an oscillation analysis with the necessary reconstructed variables and tools to apply cuts.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1848739 and NSF-1913607. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additional acknowledgments included for the collaboration separately.

References

[1] Super-K Collaboration, Y. Fukuda et al. Phys. Rev. Lett. 81 (Aug. 1998) 1562–1567.
[2] SNO Collaboration, Q. R. Ahmad et al. Phys. Rev. Lett. 87 (Jul, 2001) 071301.
[3] IceCube Collaboration, M. G. Aartsen et al. Phys. Rev. Lett. 115 (Aug, 2015) 081102.
[4] M. G. Aartsen et al. Science 361 no. 6398, (2018) 147–151.
[5] IceCube Collaboration, M. G. Aartsen et al. JINST 12 no. 03, (2017) P03012.
[6] R. Abbasi et al. Astroparticle Physics 35 no. 10, (2012) 615–624.
[7] IceCube Collaboration, M. G. Aartsen et al. Phys. Rev. Lett. 120 (Feb, 2018) 071801.
[8] A. Krizhevsky, I. Sutskever, and G. Hinton NeurIPS 25 (01, 2012).
[9] E. Kauderer-Abrams, “Quantifying translation-invariance in convolutional neural networks,” 2017.
[10] IceCube Collaboration PoS ICRC2017 (2017) 1057.
[11] C. Andreopoulos et al. Nucl. Instrum. Methods Phys. Res. A 614 no. 1, (2010) 87–104.
[12] D. C. Psichogios and L. H. Ungar AIChE Journal 38 no. 10, (1992) 1499–1511.
[13] IceCube Collaboration PoS ICRC2021 (these proceedings) 1054.
Full Author List: IceCube Collaboration

Reconstructing Neutrino Energy using CNNs
Reconstructing Neutrino Energy using CNNs

Acknowledgements

USA – U.S. National Science Foundation-Oﬃce of Polar Programs, U.S. National Science Foundation-Physics Division, U.S. National Science Foundation-EPSCoR, Wisconsin Alumni Research Foundation, Center for High Throughput Computing (CHTC) at the University of Wisconsin–Madison, Open Science Grid (OSG), Extreme Science and Engineering Discovery Environment (XSEDE), Frontera computing project at the Texas Advanced Computing Center, U.S. Department of Energy-National Energy Research Scientiﬁc Computing Center, Particle astrophysics research computing center at the University of Maryland, Institute for Cyber-Enabled Research at Michigan State University, and Astroparticle physics computational facility at Marquette University; Belgium – Funds for Scientiﬁc Research (FRS-FNRS and FWO), FWO Odysseus and Big Science programmes, and Belgian Federal Science Policy Oﬃce (Belspo); Germany – Bundesministerium für Bildung und Forschung (BMBF), Deutsche Forschungsgemeinschaft (DFG), Helmholtz Alliance for Astroparticle Physics (HAP), Initiative and Networking Fund of the Helmholtz Association, Deutsches Elektronen Synchrotron (DESY), and High Performance Computing cluster of the RWTH Aachen; Sweden – Swedish Research Council, Swedish Polar Research Secretariat, Swedish National Infrastructure for Computing (SNIC), and Knut and Alice Wallenberg Foundation; Australia – Australian
Research Council; Canada – Natural Sciences and Engineering Research Council of Canada, Calcul Québec, Compute Ontario, Canada Foundation for Innovation, WestGrid, and Compute Canada; Denmark – Villum Fonden and Carlsberg Foundation; New Zealand – Marsden Fund; Japan – Japan Society for Promotion of Science (JSPS) and Institute for Global Prominent Research (IGPR) of Chiba University; Korea – National Research Foundation of Korea (NRF); Switzerland – Swiss National Science Foundation (SNSF); United Kingdom – Department of Physics, University of Oxford.