Variational Autoencoders for Generative Modelling of Water Cherenkov Detectors

Abhishek Abhishek
University of Manitoba/TRIUMF
abhishek@myumanitoba.ca

Wojciech Fedorko
TRIUMF
wfedorko@triumf.ca

Patrick de Perio
TRIUMF
pdeperio@triumf.ca

Nicholas Prouse
TRIUMF
nprouse@triumf.ca

Julian Z. Ding
University of British Columbia/TRIUMF
julianzding@gmail.com

Abstract

Matter-antimatter asymmetry is one of the major unsolved problems in physics that can be probed through precision measurements of charge-parity symmetry violation at current and next-generation neutrino oscillation experiments. In this work, we demonstrate the capability of variational autoencoders and normalizing flows to approximate the generative distribution of simulated data for water Cherenkov detectors commonly used in these experiments. We study the performance of these methods and their applicability for semi-supervised learning and synthetic data generation.

1 Introduction

We currently cannot explain the observed matter-antimatter asymmetry in the universe. Neutrino oscillations [1, 2] may exhibit significant charge-parity symmetry violation (CPV) [3], which may lead to the answer [4]. These can be measured in experiments such as T2K [5], which produce a muon neutrino (ν_μ) or antineutrino beam directed towards a far detector where the oscillation signal is electron neutrinos (ν_e) and antineutrinos, respectively.

The planned next generation Hyper-Kamiokande (Hyper-K) [6] far detector and NuPRISM [7] intermediate water Cherenkov detector (IWCD) are multi-kilotonne water tanks surrounded by $\mathcal{O}(10k)$ 20" photomultiplier tubes (PMTs) or multi-PMT (mPMT) modules, consisting of 3" PMT arrays. They detect single photons of Cherenkov light from muons (μ^-) and electrons (e^-) produced in ν_μ and ν_e interactions with water, respectively. These are easily classified by traditional likelihood ratio-based methods [e.g. 8] due to the electromagnetic shower induced by the much lighter electron, causing a fuzzier Cherenkov light ring projected onto the PMTs. These methods, however, have limited discriminative power between an e^- and a high energy photon (γ) produced by a ν_μ without the corresponding μ^-. Such a photon can produce an e^- and e^+ pair, which then also shower resulting in a very similar ring as an e^- event. Such misclassified events constitute a significant and poorly understood background to the ν_e signal [9–11], with no existing experimental constraint.

As in existing WC detectors such as Super-Kamiokande (Super-K) [12], 1) accurate first-principles modelling of detector effects, such as varying water conditions or PMT responses, remains a significant challenge. Furthermore, 2) due to the improved granularity of IWCD and potentially the Hyper-K detector, it may be possible to detect the slight difference in Cherenkov light emission between an e^- from a ν_e and e^-/e^+ pair from a γ. Thus, the implementation and development of modern machine learning methods can 1) enable training directly on data to mitigate experimental uncertainties, and 2) maximize the rejection of γ events to limit exposure to unconstrained theoretical uncertainties. Both developments are expected to enhance the sensitivity of Hyper-K to CPV and other phenomena such as proton decay [13] and multi-messenger astronomical events [e.g. 14–16].

In this work we study the capability of variational autoencoders (VAEs) and their extensions: normalizing flows (NF) to learn the data generating distribution and benchmark the performance of VAEs in semi-supervised training on simulated IWCD datasets. In addition, we study the capability of VAEs to generate synthetic samples.

2 Methodology

2.1 Simulation and Data Preprocessing

Water Cherenkov Simulation (WCSim)[1] is a GEANT4 [17] and ROOT [18] based Monte Carlo (MC) software package. It was used to generate 3 million events each of e^-, μ^-, and γ particles in the IWCD. For this initial study, the e^-, μ^- initial positions and γ pair production positions were fixed at the center of the detector. In order to use convolutional neural networks (CNNs) as our primary NN architecture, the top and bottom of the detector cylinder were ignored and the particle directions were constrained to be perpendicular to the detector wall. The azimuthal angles and the energies of the particles were uniformly varied between 0 to 2π and 0 to 1000 MeV. The detector wall is instrumented with 16×40 mPMT "pixels" and the resulting data is structured as an image with each pixel containing 19 channels corresponding to the light intensity measured by each 3" PMT.

2.2 Semi-supervised and Unsupervised Deep Generative Models

Variational Autoencoders (VAEs) [19] are a powerful class of latent generative models that maximize an evidence lower bound (ELBO) to the intractable log-likelihood of the data, log $p_\theta(\boldsymbol{x})$:

$$\log p_\theta(\boldsymbol{x}) \geq E_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] = -\mathcal{J} \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence, $p(\boldsymbol{z})$ is the prior distribution over the latent variables $\boldsymbol{z}$, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is the variational posterior distribution, and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is the conditional generative distribution. Generally, $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and $p(\boldsymbol{z})$ are defined as factorized Gaussian distributions $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_\phi(\boldsymbol{x}), diag(\boldsymbol{\sigma}^2_\phi(\boldsymbol{x})))$ and $\mathcal{N}(0, I)$, respectively. $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ are parameterized using neural networks with parameters θ and ϕ, learned through minimization of $\mathcal{J}$ by gradient descent methods. The expectation over the conditional generative distribution $E_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$ corresponds to the fidelity of the reconstruction. Since the input variables $\boldsymbol{x}$ are continuous and considered Gaussian distributed, this term is replaced by negative Mean Squared Error (MSE) loss.

Normalizing Flows (NFs) [20] were developed to address a key limitation of VAEs where, even in the asymptotic regime, one is unable to recover the true posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ due to the simple form of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. In normalizing flows, a random variable z_0 with an initial probability density $q_0(z_0)$ is transformed into another random variable z_K with a probability density $q_K(\boldsymbol{z}_K)$ through a sequence of K smooth, invertible mappings, f:

$$\boldsymbol{z}_K = f_K \circ f_{K-1} \circ ... \circ f_2 \circ f_1(\boldsymbol{z}_0). \tag{2}$$

In planar normalizing flows, these mappings have the form:

$$f_k(\boldsymbol{z}) = \boldsymbol{z}_{k-1} + \boldsymbol{u}_k h(\boldsymbol{w}_k^T \boldsymbol{z}_{k-1} + \boldsymbol{b}_k),\ k \in [1, K], \tag{3}$$

where $\boldsymbol{u}$, $\boldsymbol{w}$, and $\boldsymbol{b}$ are the free parameters of the flow and h is a smooth element-wise non-linearity.

2.3 Model architecture

The architecture of a VAE can be conceptually divided into an encoder, a "bottleneck", and a decoder. We employed a simple LeNet [21] inspired CNN as the encoder and a symmetric architecture comprising of transposed convolution layers as the decoder. The encoder comprised of four 3×3 and two 4×4 convolution layers with the number of channels per pixel successively increasing from

[1] https://github.com/WCSim/WCSim

19 to 128. ReLU nonlinearities were used after each layer. Finally, 4 fully connected layers were used to successively compress the feature vectors to the size of the latent vectors. The simple choice of the CNN architecture allowed for focus on the development of the generative models rather than the NN architectures.

3 Empirical Evaluations and Results

The 9 million event dataset is split into 80%-10%-10% for train, validation, and test subsets, respectively. During the training process, the model with the lowest loss on a validation mini-batch is retained for further evaluation. All models are trained for 10 epochs using the Adam optimizer [22] with an initial learning rate of 0.0001. In addition, training of both the unsupervised VAE and the planar NF is performed using the 'annealing trick' [20, 23] where the KL divergence term in the loss is scaled by a factor of β, which increases linearly from 0 to 1 with the number of epochs. We observed that using the 'annealing trick' resulted in a significantly lower MSE loss and similar KL loss on the test subset than using the standard ELBO objective (1).

Dimensionality of the latent space We tested the VAE with varying number of latent dimensions in order to empirically determine the optimal setting. As shown in Figure 1, the lowest MSE loss was obtained using 32 latent dimensions. However, using 128 latent dimensions, we achieved better performance in terms of MSE loss and cross entropy loss in the cases of normalizing flows and semi-supervised learning respectively.

Planar Normalizing Flows We employed planar normalizing flows (Eq. 3) with a similar amortization strategy to [20, 23] where the flow parameters are considered to be functions of the input rather than part of the global network parameters. As shown in Figure 1, we found that using planar normalizing flows, the average KL divergence loss is an order lower than the VAE which can be attributed to having a more flexible posterior distribution. However, the average MSE loss does not improve and in fact is worse than the standard VAE.

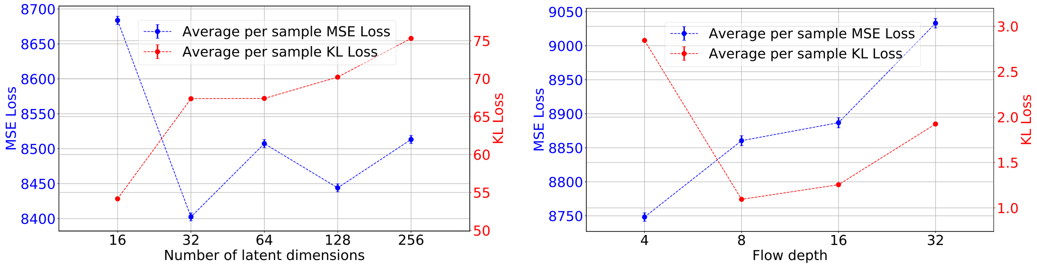

Figure 1: Comparing VAEs with different number of latent dimensions (left) and planar normalizing flows with varying flow depth (right) using MSE and KL loss on the test subset. MSE loss is measured in arbitrary units equivalent to the number of photoelectrons produced at the PMT photocathode and KL loss is measured in nats.

Event reconstruction and synthetic sample generation The comparison of simulated events and their reconstructions (Figure 2) shows that the VAE is able to capture important features of the Cherenkov ring such as its position, shape, and size. However, the reconstructions show poor ring sharpness and lack of isolated channel charge deposits from PMT dark noise and scattered/reflected light. This is expected due to the limited capacity of the encoder and decoder used in addition to the mean field approximation of the VAE. Images based on sampling from the prior $p(\boldsymbol{z})$ (Figure 3) show similar deficiencies and occasional artifacts, most likely due to prior-posterior divergence.

Latent space interpolation In order to perform interpolation in the latent space along the axis of some physically meaningful quantities, we used the k-nearest neighbors algorithm in the feature space of the event azimuthal angle and energy. The latent vectors for the 256 nearest neighbors to some reference point in the feature space were used to find the start and end positions in the latent space. We observed (Figure 4) that linear interpolation along the energy axis yields smooth transitions from one step to the next, however interpolation along the azimuthal angle axis does not. This suggests that not all high level features correspond linearly to directions in the latent space.

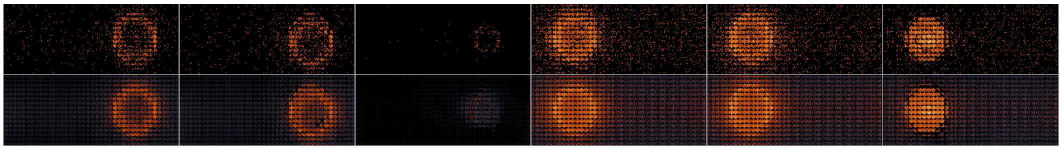

Figure 2: Cherenkov ring images comparing actual simulated events (top) with their corresponding VAE reconstructed events (bottom).

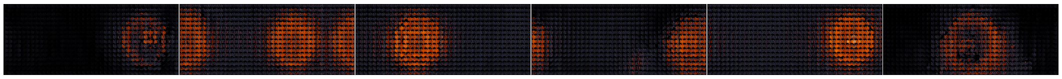

Figure 3: Cherenkov ring images for events randomly sampled from the latent prior p(z) = $\mathcal{N}(0, I)$.

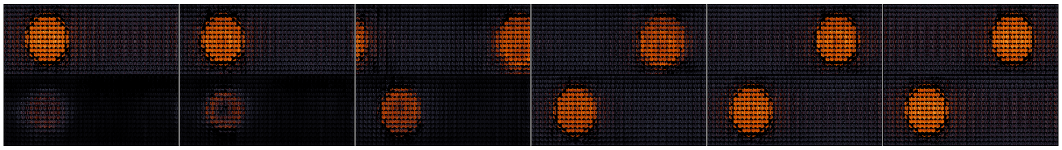

Figure 4: Linear interpolation in the latent space for μ^- events along the angle axis from $\phi = 0$ to $-\pi$ (top) and energy axis from 200 MeV to 800 MeV (bottom).

Semi-supervised learning Inspired by the **M1** strategy of [24], we pre-trained the VAE (128 latent dimensions) using approximately 7.2 million training examples and subsequently trained a 4-layer multilayer perceptron (MLP) that takes the deterministic output of the encoder (before the reparameterization) as input. We benchmarked this semi-supervised model against a fully supervised CNN with a similar architecture. The fully supervised CNN and the MLP classifiers were trained for 10 epochs using the Adam optimizer with varying number of labelled examples. We found that in the low labelled data regime, the semi-supervised model consistently outperformed the fully-supervised model and achieved state-of-the-art performance for the task of γ, e^- versus μ^- event discrimination. As shown in Table 1, the **SS-CNN** model also outperforms the supervised model on γ versus e^- classification, considered unfeasible with existing likelihood ratio approaches [9].

Table 1: Comparing the semi-supervised model performance to that of the fully supervised model for various labelled dataset sizes.

Number of training examples	**γ background rejection (%) at 50% e^- signal efficiency**		**γ background rejection (%) at 80% e^- signal efficiency**	
	SS-CNN	**CNN**	**SS-CNN**	**CNN**
$11,250$	**77.6**	76.4	**50.7**	46.3
$22,500$	**80.4**	78.1	**54.3**	48.5
$45,000$	**80.7**	79.4	**55.9**	49.9

4 Conclusion

We demonstrated the ability of VAEs and NFs to approximate the generative distribution of simulated water Cherenkov detector data, with NFs showing no significant improvement over the standard VAE. When used for the task of classification, the semi-supervised approach demonstrated performance gains over a fully supervised model comparable to those demonstrated in other domains. Reconstruction and synthetic data generation is possible, however the presence of artifacts suggests an improved design of the loss function and/or the generative model is needed. Linear interpolation along the energy axis displays a smooth behaviour, but along the azimuthal angle axis does not suggesting that more sophisticated interpolation methods may be needed. Through this work, we show the promise of applying generative models to address key challenges of neutrino oscillation experiments such as γ vs e^- classification and mitigation of experimental uncertainties through future work.

Acknowledgements

The authors would like to thank Akira Konaka, Mark Hartz and Olivia Di Matteo for helpful discussions and feedback on a draft of this paper. This research was enabled in part by support provided by TRIUMF, NSERC and Compute Canada (www.computecanada.ca).

References

[1] Q. R. Ahmad et al. Direct Evidence for Neutrino Flavor Transformation from Neutral-Current Interactions in the Sudbury Neutrino Observatory. *Physical Review Letters*, 89:011301, Jun 2002.

[2] Y Fukuda et al. Evidence for oscillation of atmospheric neutrinos. *Physical Review Letters*, 81(8):1562, July 1998.

[3] A. D. Sakharov. Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe. *Soviet Physics Uspekhi*, 34(5):392–393, may 1991.

[4] M. Fukugita and T. Yanagida. Barygenesis without grand unification. *Physics Letters B*, 174(1):45 – 47, 1986.

[5] K. Abe et al. Search for CP Violation in Neutrino and Antineutrino Oscillations by the T2K Experiment with 2.2×10^{21} Protons on Target. *Physical Review Letters*, 121:171802, Oct 2018.

[6] K Abe et al. Hyper-Kamiokande design report. *arXiv:1805.04163*, 2018.

[7] S. Bhadra et al. Letter of intent to construct a nuPRISM detector in the J-PARC neutrino beamline. *arXiv:1412.3086*, 2014.

[8] M. Jiang et al. Atmospheric neutrino oscillation analysis with improved event reconstruction in Super-Kamiokande IV. *Progress of Theoretical and Experimental Physics*, 2019(5), 05 2019.

[9] K. Abe et al. Search for neutral-current induced single photon production at the ND280 near detector in T2K. *Journal of Physics G: Nuclear and Particle Physics*, 46(8):08LT01, jun 2019.

[10] K. Abe et al. Search for CP Violation in Neutrino and Antineutrino Oscillations by the T2K Experiment with 2.2×10^{21} Protons on Target. *Physical Review Letters*, 121:171802, Oct 2018.

[11] K. Abe et al. Measurement of neutrino and antineutrino oscillations by the T2K experiment including a new additional sample of ν_e interactions at the far detector. *Physical Review D*, 96:092006, Nov 2017.

[12] S Fukuda et al. The Super-Kamiokande detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 501(2-3):418–462, 2003.

[13] K. S. Babu et al. Working Group Report: Baryon Number Violation. In *Proceedings, 2013 Community Summer Study on the Future of U.S. Particle Physics: Snowmass on the Mississippi (CSS2013): Minneapolis, MN, USA, July 29-August 6, 2013*, 2013.

[14] K. S. Hirata et al. Observation in the Kamiokande-II detector of the neutrino burst from supernova SN1987A. *Physical Review D*, 38:448–458, Jul 1988.

[15] P. Antonioli et al. SNEWS: the SuperNova early warning system. *New Journal of Physics*, 6:114–114, sep 2004.

[16] M. Aartsen et al. Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A. *Science*, 361(6398), 2018.

[17] S. Agostinelli et al. Geant4—a simulation toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[18] R. Brun and F. Rademakers. ROOT—an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1-2):81–86, 1997.

[19] D. P. Kingma and M. Welling. Stochastic gradient VB and the Variational Auto-Encoder. In *Second International Conference on Learning Representations, ICLR*, 2014.

[20] D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

[21] Y. LeCun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, December 2014.

[23] R. van den Berg et al. Sylvester Normalizing Flows for Variational Inference. *arXiv:1803.05649*, March 2018.

[24] D. P. Kingma et al. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.