Course Project Proposal: Application of Variational Autoencoder in Particle Physics Events Generation

Rui XUE
Department of Physics and Astronomy
University of Pittsburgh
Pittsburgh, PA 15213
rux23@pitt.edu
rxue@andrew.cmu.edu

Abstract

Events generation in particle physics research is a highly time-consuming task. Researchers use event generators, based on Quantum Field Theory (QFT) and Monte-Carlo simulation, to simulate how elementary particles interact with each other. The angle distribution results of those particles can be seen as a high dimensional probability distribution. Variational Autoencoders (VAE) are widely used nowadays to learn probability distributions of datasets. This project aims at exploring the potential use of the VAE method in particle physics events generations.

1 Introduction

The Standard Model (SM), based on Quantum Field Theory (QFT), is the most important theoretical frame work in particle physics. The finding of Higgs Bosons in 2012 by LHC is a great triumph of the Standard Model, and it marks that all the predicted particles are found by human beings[1]. However, the story hasn't come to an end. There are still lots of problems that cannot be solved by the SM, such as the origin of dark matters, the imbalance of matters and anti-matters in the universe, attempts to unify strong force with electro-magnetic force and weak force, attempts to unify QFT with General Relativity. To investigate those problems, we must find particle-interaction signals that cannot explained by SM, and hopefully, those signals could serve as evidence of some Beyond Standard Model (BSM) theories.

The Effective Field Theory (EFT) proposes the real world Lagrangian is in the form of

$$\mathcal{L}_{\text{eff}} = \mathcal{L}_{\text{SM}}^{(4)} + \frac{1}{\Lambda}\sum_i C_i^{(5)}\mathcal{O}_i^{(5)} + \frac{1}{\Lambda^2}\sum_i C_i^{(6)}\mathcal{O}_i^{(6)} + \mathcal{O}\left(\frac{1}{\Lambda^3}\right)$$

And the first term $\mathcal{L}_{\text{SM}}^{(4)}$ is the SM Lagrangian. The rest terms are the BSM terms we are searching for. $C_i^{(5)}$ and $C_i^{(6)}$ are Wilson Coefficients that determines the intensity of BSM interactions. Researchers use maximum-likelihood fittings between simulation signals and experimental signals to extract those Wilson Coefficients. If those coefficients are deviated by 5σ, we can confidently say there is new physics. Top quarks, as the heaviest quarks, are expected to show BSM signals with higher probabilities. Using ten angles, we can fully describe a top-antitop decaying system as shown by Fig.1.

For real experiments, millions of such top decay events were collected. And for simulations, fewer events were generated due to the large computational cost. The 10 angles are correlated, we can accumulate those events and get $\rho(\theta^+, \phi^+, \theta^-, \phi^-, \theta^{*+}, \phi^{*+}, \theta^{*-}, \phi^{*-}, \theta, \eta)$. This is a 10 dimensional probability distribution, determined by the Lagrangian $\mathcal{L}$. Nowadays, the main bottleneck of physics analysis comes from the simulation side. The complicated theoretical computation takes

10617 Course Project (2024).

weeks or even months to generate enough full-simulation datasets. So a fast simulation method is urgently needed for early-stage studies. This project is trying to explore the potential of Variational Encoders (VAE) in particle physics simulation, using the VAE model to learn 10 angle distribution $\rho(\vec{\theta})$, thus doing quick sampling to generate simulation data.

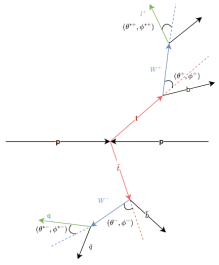


Figure 1: Top-Antitop decay system. 8 angles are marked on the graph, the rest 2 angles describes the whole system orientation.

2 Background and Data Generation

The generator-level simulation is the first step of full-simulation, and the process is purely based on theoretical calculation and Monte-Carlo sampling. There are two main generators used by physicists — SMEFTSim and DIM6. The training data we used comes from DIM6 simulation, and the total number of events used for training and testing is 5,5000. Just completing the generator simulation would cost around 4 hours, but 55K events are still too little for real analysis, not to mention millions of events. And the more complicated full-simulation process can take months in the worst case.

Considering the time consumption problems, fast-simulation methods are necessary, and there are some works in this area. However, most of them are trying to speed up the full-simulation process, and there is little attention on speeding up the generator simulation. Besides, the early works on fast-simulation are mainly based on Generative Adversarial Networks (GAN)[2]. GAN is hard to train and tricky in tuning parameters. Since it's more reasonable to learn the particle-angle distribution using probability models, VAE models, based on strict math and easier for tuning and training as well, are better candidates for fast-simulation tasks [4][6]. In this project, we use VAE models for fast-generator simulation, so the topic and the method are brand new in particle physics simulation. Furthermore, this project would investigate the interpolation performance of conditional-VAE (cVAE)[3], where the input conditions are Wilson Coefficients not used for training. The interpolation ability is the most important improvement comparing with previous fast-simulation works.

3 Methods and Preliminary Results

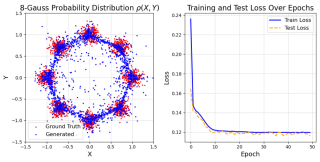
For the preliminary study, we started with two dimensional probability distributions. The reason using two dimensional distributions is because they are easy to visualize and can give an intuitive overview of the model. The preliminary data comes from 2d Gauss distributions for simplicity, and the real 10d simulation data will be used in the coming study. The following two subsections show the performance of VAE and cVAE on the 2d distributions.

3.1 Two Dimensional Eight-Gauss Peaks

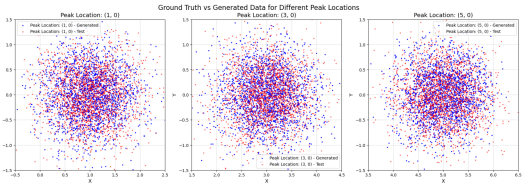
The VAE model to generate 8-Gauss peaks is structured with 3 fully connected linear layers as an encoder, with dimensions $(2, 70), (70, 70), (70, K)$; 3 fully connected linear layers as a decoder, with dimensions $(K, 70), (70, 70), (70, 2)$, $K = 25$. The 2d 8-Gauss peak has distribution of $\rho(x, y)$, so the input and output data are in dimension of two. The latent layer dimension is in 25 to fully catch the features of the discrete distributions. The loss function $L = L2 + L_{KL}$, contributed by L2 loss and KL-divergence. The 8-Gauss toy model shows that VAE model can learn discrete distributions, as shown by Fig. 3. VAE doesn't have well-defined accuracy because it is not a supervised learning model. The best testing loss of the model is 0.12.

3.2 Two Dimensional Moving Gauss Peaks with Conditions

The cVAE model to generate movable Gauss peaks has the similar structure as the VAE model, except that the input dimension is $2 + N$ and the latent dimension is $K + N$, where $K = 25$ while $N = 3$, and N is the dimension of the condition. For this toy model, we take the condition as the center peak location. In the near future, the condition would be the Wilson Coefficient. The core thing is that the condition affects the distributions, the model was trained with Gauss peaks located at $(1, 0), (3, 0), (5, 0)$, and the model can reproduce the Gauss peaks at different locations very well. In the near future, we will also investigate the interpolation ability of cVAE, so that the model can generate intermediate distributions under the conditions that were not used during training.



(a) Left: 8-Gauss distribution; Right: Train Loss vs Test Loss



(b) Ground Truth vs Generated Data for Different Peak Locations

Figure 2: 2d demonstration on the potential of VAE and cVAE

4 Future Works

The first thing to do in the rest of the semester is to go beyond 2d datasets to 10d real simulation datasets. This requires probably more complicated layer structure and more works on tuning hyperparameters. The second thing to do is to investigate the interpolation ability of the cVAE model. For example, when one of the Wilson Coefficients $ctG = 0.0$, the 10 angle distribution is following SM. What we want to do is making the trained model be able to generate distributions when $ctG = 0.1, 0.2, \cdots, 1.1, \cdots$. It is of course impossible to train for all possible conditions, but since the latent layer neurons in cVAE encodes features of the dataset, in principle we can combine those features to generate new distributions under new conditions. To evaluate the quality of generated high dimensional datasets, we can use those data to reconstruct some physics variables, such as Fourier coefficients of the angular distribution[5]. If those comparable physics variables fits well, or slightly deviate from the generator-level physics variables, then we can conclude that the method is reliable for the fast-simulation task.

The rough timeline should be 1 week for data generation and data cleanup, 3 weeks for model training, tuning and quality-check, 1 week for project summarize and reports. Since I cannot find a teammate, I will do the whole project.

References

[1] ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, September 2012.

[2] Michele Faucci Giannelli. FastCaloGAN: a fast simulation for the ATLAS calorimeter system using GANs. 2020.

[3] William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational auto-encoders, 2021.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[5] Chi Wing Ng. Angular distributions in t-channel single-top decay. September 2023.

[6] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.