Present Improvements for HEP Data using Variational AutoEncoders (VAE)

Deep Autoencoders for scientific compression
GSoC 2023

Mentor - Alexander Ekman
Student - Aman Singh Thakur

Variational Autoencoders (VAE) - Implementation Details

- The fundamental issue with regular Autoencoders is that the latent space (space where the compressed/encoded input lie) may not be continuous or allow for easy interpolation [Source]
- Using Eric Wulff's current Baler AE architecture (200-100-50-3-50-100-200 nodes), I built a Variational Autoencoder (VAE) by adding additional layer of mean and standard deviation in the existing architecture.[Source AE Architecture] [Souce VAE Github]
- Further, I propose to expand the model to use convolutional layers instead of linear hidden layers to better capture non-linearities in the dataset. The intricacies with this model is discussed in the last section.
- For each model, three independent runs were conducted and model metrics such as difference between the mass calculated before & after compression were averaged out.

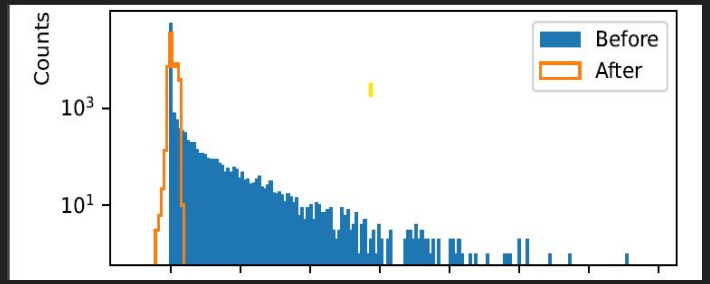
Observations and Improvements

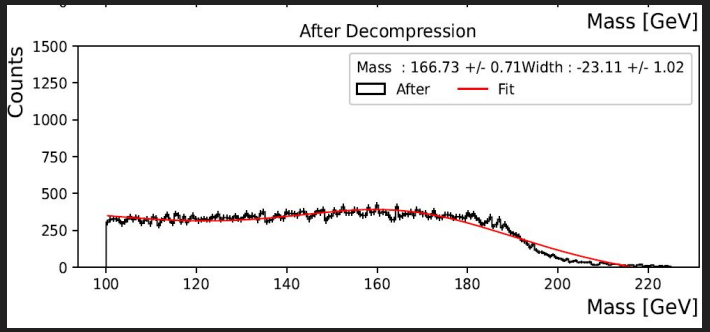
- VAE is ~100% more accurate than regular AE in calculating mass difference before & after compression.
- VAE took more time to compress & depress due to addition of sampling layer. Similarly, the latent size is twice in VAE w.r.t AE.

Model Type	Model Architecture	Avg Time to Train (min)	Avg Compression Time (min)	Avg Decompression Time (min)	Average Mass Diff (%)
AutoEncoder (AE)	200-100-50-4-50-100-200	3.453	0.042	0.192	2.067
Variational AutoEncoder (VAE)	200-100-50-(4)(4)-50-100-200	4.647	0.058	0.314	1.033

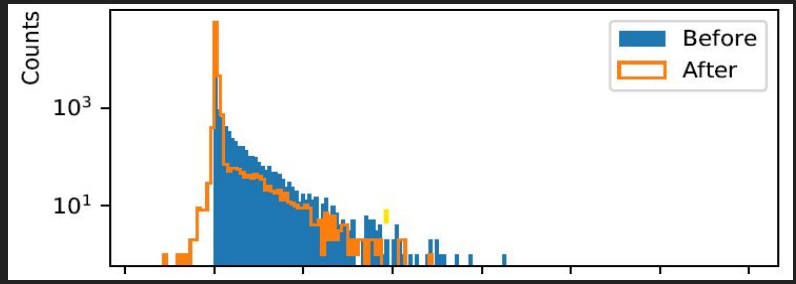
VAE vs AE - Residual, Counts and Evaluation Plots

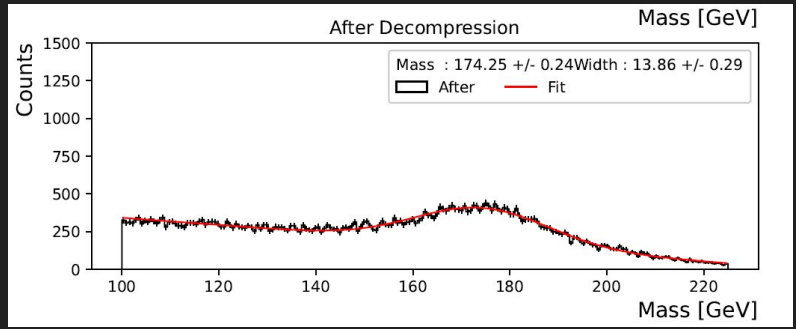
- VAE does a better job at fitting all columns and evaluation task. Below Col - m_InvisibleEnergy





AE





VAE

Discussion & Future Scope

- I propose to experiment with more loss functions and kernel functions to help the NN converge. Currently, MSE Loss Function was the ideal choice for AE and VAE. Other Loss Functions like KL-Divergence Loss Function performed poorly for VAE compression task.
- Due to lack of time, currently, VAE is not L1-regularized. In the future, I plan to add L1-regularization to avoid overfitting and consistent results.
- With great potential unearthed, I propose to explore other flavors of VAE like WAE - MMD, Joint VAE, etc and compare their performance. [Source]
- I further plan to enhance the CNN VAE variant as it's currently underfitting the data and is slower and harder to debug when compared to it's linear counterparts.

Github Repository Link