Results from your own analysis and dataset

Deep Autoencoders for scientific compression
GSoC 2023

Mentor - Alexander Ekman
Student: Devesh Marwah
Github Repository

Dataset

- The dataset is borrowed from the official CERN site, the idea is to test the data on a different physics data other than given and use a random dataset instead of focussing on results.
- This document contains 100k dimuon events selected from the Mu dataset from Run2010B. Each line corresponds to an event. The main file contains all 100k events. Files with an underscore contain 10k events each.
- These data were selected from the Mu primary dataset.
- The dataset can be found here
- Another Dataset used which mimics perfect gaussian distribution can be found here and shows how it outperforms the given the dataset.

Analysis

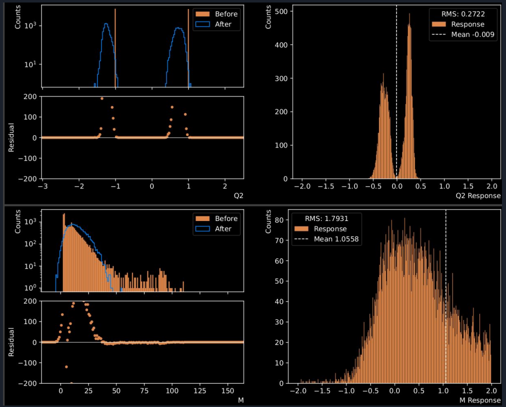
The evaluation file can be found in evaluation_1.pdf. In examples/plotting folder.

Before compression:Mass : 63.43 +/- 321.46 Width : 108.15 +/- 327.21

After compression: Mass : 173.0 +/- 0.0 Width : 1.0 +/- 0.0

My model also outputs similar results.

After compression: Mass : 172.0 +/- 0.0 Width : 1.0 +/- 0.0

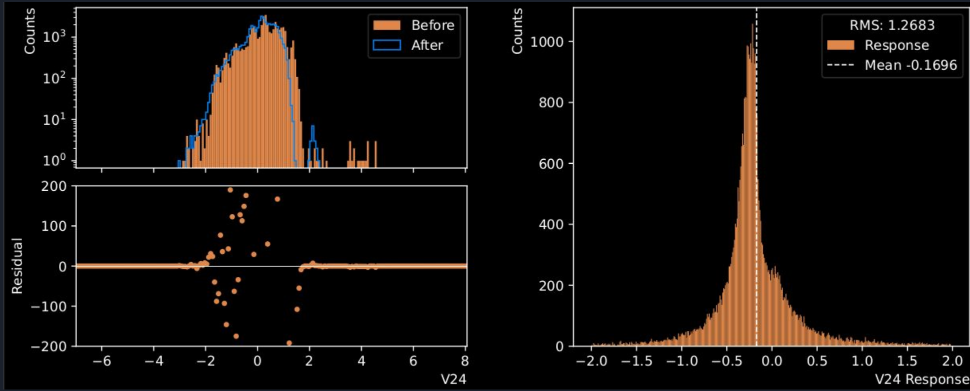


However on the other dataset which mimics the Gaussian distribution.

It performs very well. The evaluation file can be found in evaluation.pdf. In examples/plotting folder.

Before compression: Mass : 173.0 +/- 0.0 Width : 1.0 +/- 0.0

After compression: Mass : 173.0 +/- 0.0 Width : 1.0 +/- 0.0



How you made it work with Baler?

- The dataset needed few tweaks in the pre-processing part, through which it could be loaded easily in the pkl format. The modifications for the same can be found in the repository.
- The issue seems with the dataset itself as it does seem to lack hidden dimensions and seems to be collected around one value.
- Hence it is difficult for the dataset to capture those hidden dimensions and with increase in width, the model falters.
- The choice of dataset was intentional to show the shortcomings of the Variational Autoencoder and Baler in general and to show different improvements in the model.

Improvements in Baler

- Instead of using one compression model, multiple compression models can be used in order to test which one is doing better.
- In a case one model starts to vastly outperform others, the other tree node can be terminated. A suggestion can be to train model which analyses the dataset and then chooses which model to perform the task on.
- This can be done using a basic CNN too, where image of dataset can be provided with the best model it performed compression on, doing so can vastly improve the performance of Baler.
- A similar task was performed by me on a different field (Signal processing). Although the accuracy was not good, it is something worth exploring.

Impact on Society

- Theoretical physics has a significant impact on society which is not visible for a normal man. Compression is a major issue and a lot of research has been behind this.

- The work presented here helps the researchers to step out from conventional models and present new work to help them explore new domains.

- Adversarial models have gained a lot of traction recently and have recently surpassed the state of the art models for various downstream tasks. This a step to work in that direction and improve the same.

- Apart from adversarial models, it also presents a small improvement on previous GSOC project of year 2021 in a small time frame and is something definitely worth to explore on.