GSOC PROJECT IN AUTOENCODERS

Aditya Choudhary

DATASET

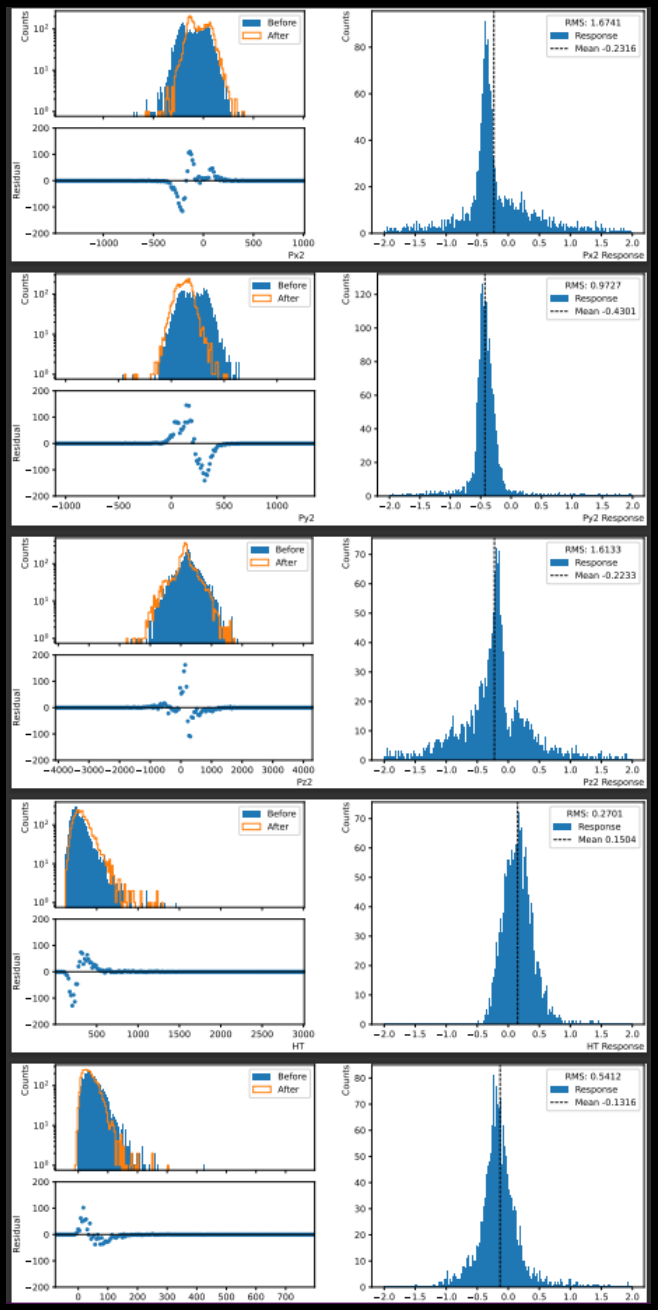
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Run	Lumi	Event	MR	Rsq	E1	Px1	Py1	Pz1	E2	Px2	Py2	Pz2	HT	MET	nJets	nBJets
2	148029	388	3.02E+08	215.553	0.031977	136.71	-109.893	-54.0342	-58.9032	142.179	70.0254	41.1225	-116.513	203.666	18.311	2	0
3	148029	388	3.02E+08	155.437	0.042157	83.3865	81.15	6.88361	-12.9688	73.9025	-72.2472	11.8835	3.0899	154.659	14.7747	2	0
4	148029	388	3.02E+08	400.563	0.026938	253.184	139.902	102.64	-101.935	535.551	-110.379	-89.0929	-516.179	343.28	25.2211	3	0
5	148029	388	3.02E+08	286.245	0.094192	175.486	-156.024	-62.9535	-47.7434	112.851	89.0843	3.45025	67.9007	257.397	46.0288	2	0
6	148029	388	3.02E+08	204.514	0.018804	833.795	100.41	-16.659	-827.498	445.612	-91.1991	15.5583	-390.144	269.492	8.11345	3	0

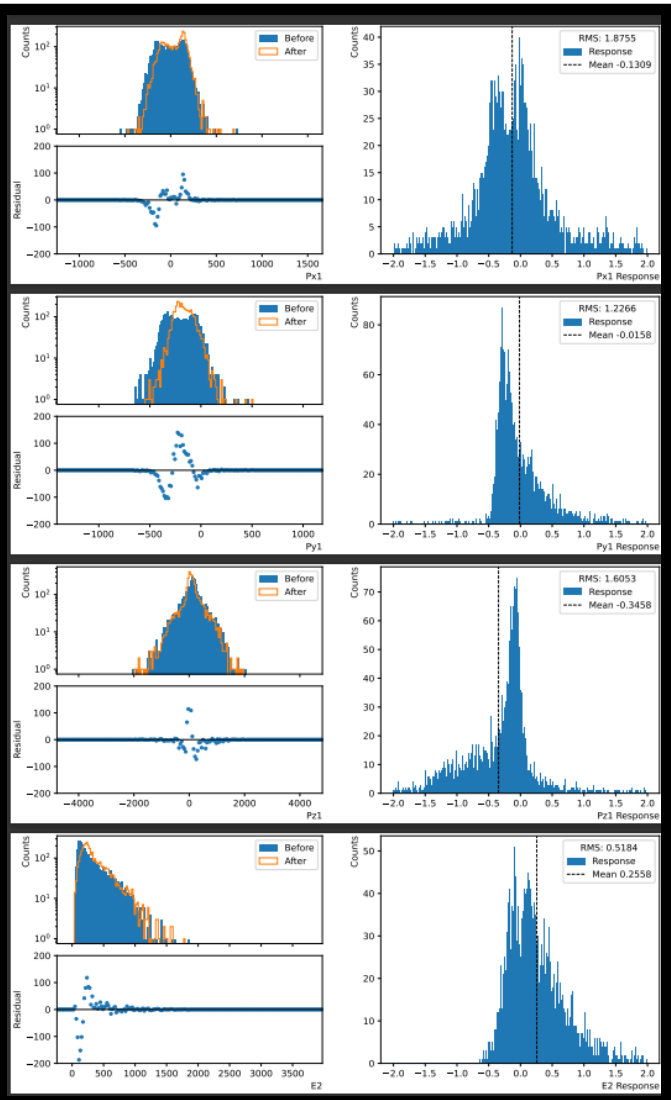
- The columns other than A,B,C,P and Q were considered for data compression owing to their high entropy
- The dataset had 21726 rows and 17 columns with no empty/null value
- The data is similar to data obtained in collisions at LHC and hence holds scientific importance in the field of high energy physics

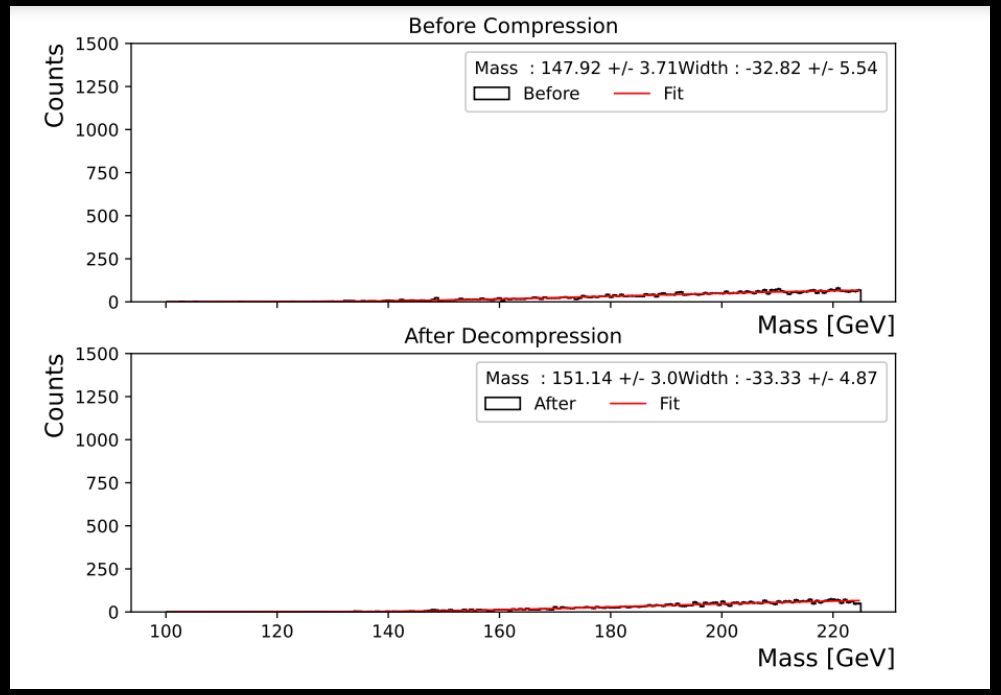
USING BALER

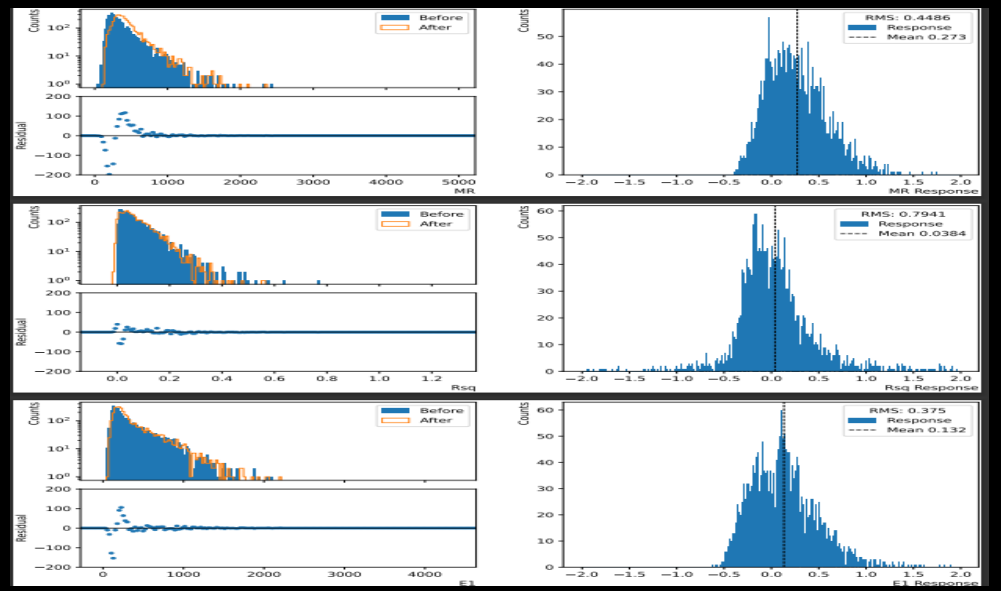
- The Proton Collision Dataset from Kaggle was used. The data was available in csv so to convert to a pandas dataframe, pd.read_csv was used directly and relevant columns were kept and others dropped.
- Owning to the small size, a batch size of 32 was used. The george_SAE model after some improvements gave the best results
- Other than Py1 , Px2 and Py2 all other features showed good reconstruction. Best validation loss of 0.003011 was achieved during training with 20 epochs. Early Stopping condition wasn't invoked in any of the epochs
- PDFs containing results can be found in /deliverables

RESULTS


Before
After
RMS: 1.6741
Response
Mean -0.2316
Px2
Px2 Response
RMS: 0.9727
Mean -0.4301
Py2
Py2 Response
RMS: 1.6133
Mean -0.2233
Pz2
Pz2 Response
RMS: 0.2701
Mean 0.1504
HT
HT Response
RMS: 0.5412
Mean -0.1316


Before
After
RMS: 1.8755
Response
Mean -0.1309
Px1
Px1 Response
RMS: 1.2266
Mean -0.0158
Py1
Py1 Response
RMS: 1.6053
Mean -0.3458
Pz1
Pz1 Response
RMS: 0.5184
Mean 0.2558
E2
E2 Response


Before Compression
Mass : 147.92 +/- 3.71Width : -32.82 +/- 5.54
Before
Fit
Counts
Mass [GeV]
After Decompression
Mass : 151.14 +/- 3.0Width : -33.33 +/- 4.87
After
Fit
Counts
Mass [GeV]


Before
After
RMS: 0.4486
Response
Mean 0.273
MR
MR Response
RMS: 0.7941
Mean 0.0384
Rsq
Rsq Response
RMS: 0.375
Mean 0.132
E1
E1 Response

FURTHER IMPROVEMENTS

- Using VAE with convolutional layers and appropriate kernel may help in providing better results.
- Adjusting weight of the KLD Loss hyperparameter may help in giving better results than SAE.
- Number of hidden layers and their size can also be increased to get better results. Currently hidden layers with at most 256 nodes have been used.
- For this dataset, batch size can be further reduced to get better results at cost of increased training time.