Baler Performance on Other Datasets

Deep Autoencoders for scientific compression
GSoC 2023

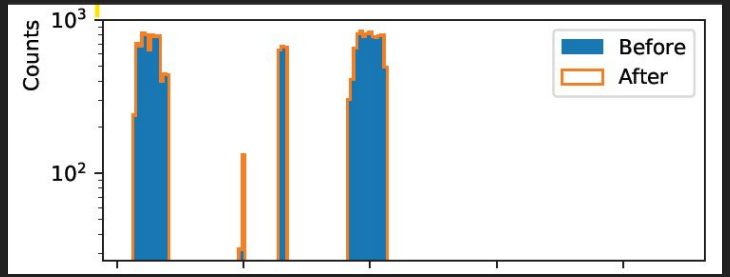
Mentor - Alexander Ekman
Student - Aman Singh Thakur

Dataset and Pre-Processing

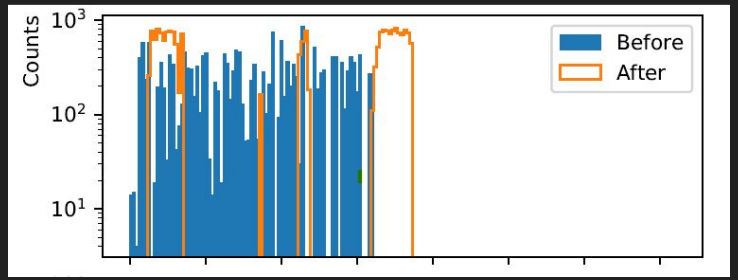
- A Similar dataset to Example called ‘United Nations: Peacekeeping Troops and Police” was chosen from Kaggle. This dataset is mostly used for analyzing uniformed contribution of peace keepers by Gender, Country, Rank, etc. [Source]
- The dataset is 7.6MB in size with 125193 rows and 6 useful categorical columns - Contribution_ID, ISOCode3, M49_Code, Contributing_Country, Mission_Acronym, Personnel_Type that are converted into integer categories.
- In this age of data boom, it’s likely expected that social services datasets like EEO-1 (20 million data points collected from over 75K private employers annually) will require lossy compression to store information that could be used later for analyzing. In this experiment, I start with a small UN dataset.
- Since this dataset is mostly discrete in nature, I expect Baler to do well (and maybe better than a VAE).

Results

- Interestingly, Baler does compress the data by overfitting one variable while ignoring the rest. This might be possible as the input dataset is relatively small for this large Autoencoder network.
- Currently, VAE codebase is buggy and seems to not accommodate this new dataset as it breaks during loss calculations. With minor tweaks, I plan to compare both models on this dataset and move to larger datasets like EE01.



Column Contribution ID



Column M49 Code Response