Autoencoders for data compression in ATLAS

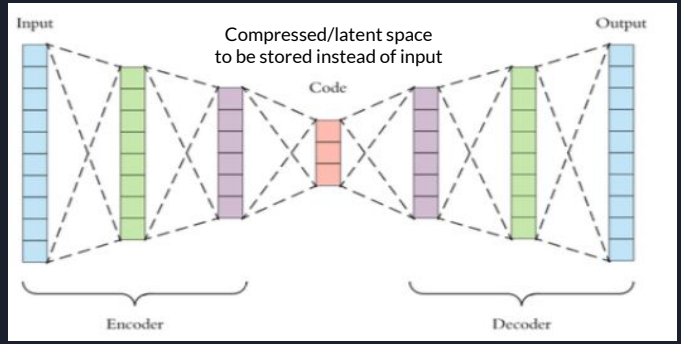
George Dialektakis - 26/08/2021
GSoC - HSF (ATLAS)
Supervised by: Alex Gekow, Antonio Boveia (Ohio State University)
Baptiste Ravina (University of Glasgow)
Caterina Doglioni (Lund University)

Presentation given by: Alex Gekow (Junior mentor)
Report, logs and code will soon be found on Zenodo

The problem we're trying to solve and how

- Context: ATLAS experiment at the LHC
- Every second, ATLAS handles 1.7 billions collision events (~1 MB each) → huge storage needed → **restricts** the amount of information / event rates that can be recorded
- The ATLAS experiment uses **trigger systems** to filter out irrelevant information by picking and sending only interesting events to the data storage system.
- A **reduction of the event size** can allow for explorations that were not previously possible due to the limited storage.

This project aims to investigate the use of autoencoders (AEs) to **compress** event-level data generated by HEP detectors.

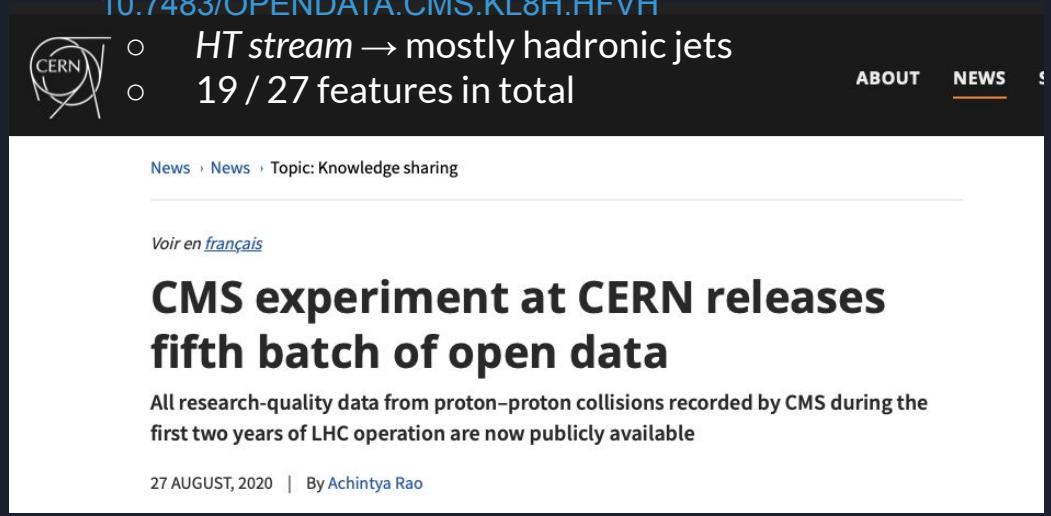


Continuing and extending the work of 2020 GSoC and Master's theses (for references & adjacent work, see report and this poster at the Offshell conference)

Slide from G. Dialektakis's GSoC evaluation task

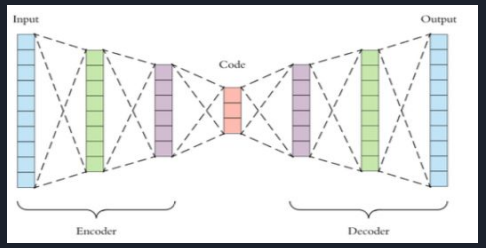
The dataset and the methods

- Using CMS Open Data from 2013 LHC operations: 10.7483/OPENDATA.CMS.KL8H.HFVH
  - *HT stream* → mostly hadronic jets
  - 19 / 27 features in total

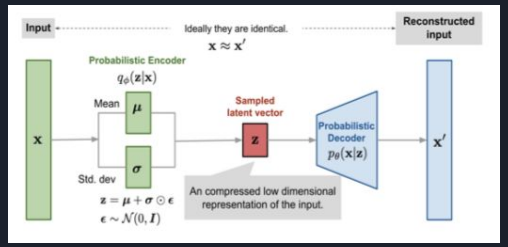


- Preprocessing: outlier removal and normalization
  - Result from GSoC 2021: comparison of different normalization methods, min-max performs better than what used so far

- **Goal of GSoC project:**
  - *test different autoencoder architectures, namely*
  - *Simple autoencoder*



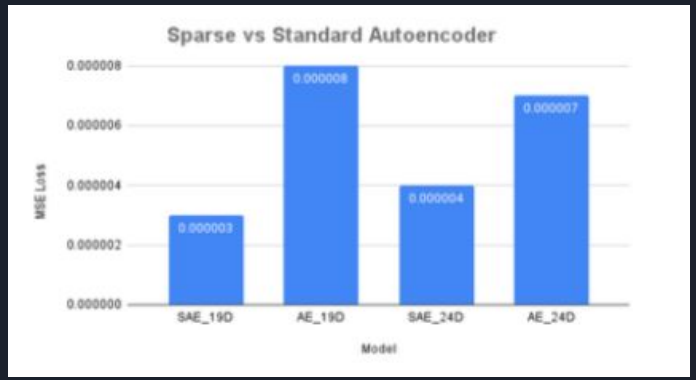
  - *Variational autoencoder*



  - *Sparse autoencoder*
    - *Same as simple autoencoder, but removes redundancies from the input*

Results & next steps

- Performance of *Variational* and *Sparse* AE benchmarked against *simple* AE
  - Metric: MSE loss on validation sample (but also: residuals of variables)
  - Sparse AE does better than simple AE
  - VAE does not do well at all → needs more work and tweaking (see e.g. https://arxiv.org/abs/1811.10276 where it works well for anomaly detection)



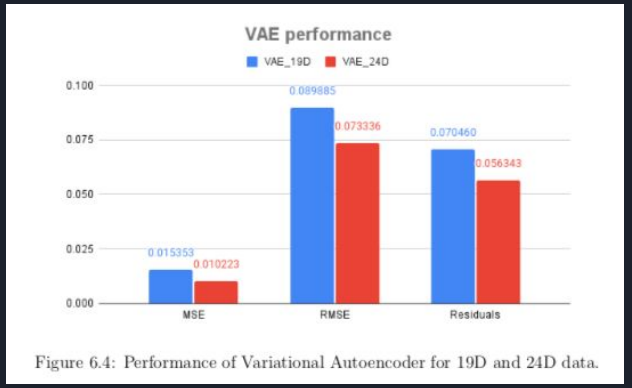


Figure 6.4: Performance of Variational Autoencoder for 19D and 24D data.

- All code available & documented for future students
- Next/ongoing steps: event-level autoencoders (S. Astrand), investigate implementation of sparse AE on FPGA (B. Ravina A. Gekow and A. Boveia's new student), compression of raw data...