

Lossy Compression algorithm for simulated datasets

Paulius Balčiūnas, IRIS-HEP Fellow



Introduction

- 1st year data science master degree student in Vilnius University
- Finished in Vilnius University bachelor degree IT and Computer Modeling master studies
- Worked as a backend software engineer for 3,5 years in Lithuania
- Currently I am aiming to become a machine learning model researcher



CERN IRIS-HEP

- Work with mentors Tomas Raila and Valdas Rapsevicius on a research project
- Participate in CERN lectures, seminars and informative excursions for learning about the CERN research teams, detectors and overall life at CERN
- Enrol in CERN workshops for improving skills related to data analysis and machine learning

Research project

- **Topic**

Lossy Data Compression for Simulated CMS
Pile-up Pixel Subdetector Datasets

- **Problem**

Simulated CMS datasets create significant
bottlenecks: slow data processing time, datasets
take up too much storage space and etc.

- **Goal**

Apply Vector Quantization (VQ) techniques for
data compression

Vector Quantization

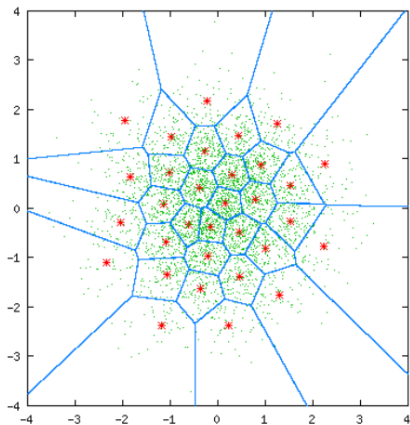


Figure: Vector Quantization (adapted from [1]).

Dataset

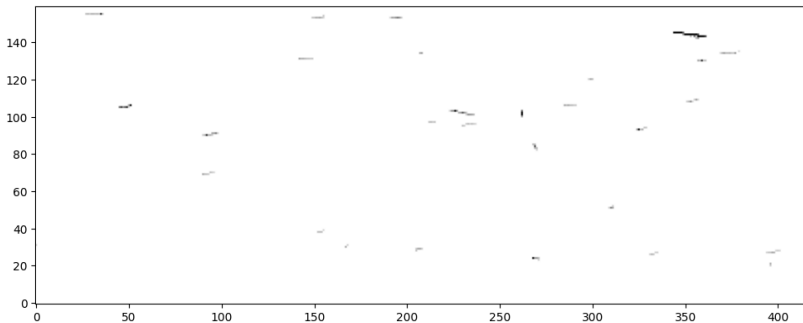


Figure: PixelDigi Event dataset

Current internship results

- Learned about CMS detector geometry and pile-up events
- Applied Vector Quantization techniques on certain particle event hit groups (BPIX and FPIX) and 1 pixel patch filtering
- Started researching VQ Variational Auto Encoder neural network implementations for data compression

The next internship plans

- Participate in the upcoming CERN workshop for improving skills in data science and machine learning
- Apply VQ-VAE model for data compression and analyse its results
- After the end of the internship, continue working on the research project as a student researcher in my master and Phd studies

References



Hai Zhu.

Vector quantization.

Presentation slides, HPC Fall 2012, 2012.