Bioinformatics Summer School Report

Name: Abdulrahman Alhomayany
Internship Organization: Broad Institute
Supervisor's Name: Christopher Reeder (ML4H team)
Project title: ThunderPack for EHR ingestion

1. Introduction

I worked on testing alternative methods of electronic health record (EHR) ingestion as compared to the current approach employed by JEDI. Namely, I wanted to test methods involving Apache parquet and LMDB. I employed ThunderPack along with a specific addressing scheme to mitigate the drawbacks of same-level addressing.

As I was not granted data access to the main cohort, C3PO, I used the MIMIC-IV dataset and injected it with synthetic multimodality (free-text and images) to generalize the findings.

The findings were promising. The alternative approach utilizing ThunderPack performed better on several metrics and warranted further investigation using the primary MGH dataset.

2. Materials and Methodology

The work was done on the Google Cloud Platform (GCP). I set up a Linux virtual machine and attached SSD storage disks to hold the data. I set up a development environment using Anaconda and Jupyter, then used the GCP SDK to forward Jupyter ports remotely.

With MIMIC-IV as my main dataset consisting entirely of tabular data, I used an auxiliary chest X-ray image dataset and tied each image to an admission ID. I also used synthetic free-text notes to simulate doctor's notes as they are currently ingested by JEDI. I used a combination of Python and Bash scripts to set up the test dataset on my virtual machine.

After that, I set up separate notebooks to ingest the data. I ingested the test dataset into HDF5 files in an approach directly analogous to the current approach employed by JEDI. I then set up another notebook and ingested the test dataset using the alternative ThunderPack approach I proposed.

There are 3 metrics that were investigated: **disk usage**, **write speed**, and **read speed**. These metrics are relevant because they relate to two primary improvement criteria: **Cost** and **time**. Cost relates to both cloud compute and cloud storage costs. My alternative approach should demonstrate promise in reducing either the size of the ingested dataset or the time to ingest and run analyses on the data, to be considered for recommendation.

3. Results and Discussion

3.1. Disk Usage

The ThunderPack approach took up 7.3% of the disk space of the previous approach for the entire dataset.

HDF5 (GB)	ThunderPack (GB)
250.9	18.1

3.2. Write Speed

Ingestion with the ThunderPack approach took 9.7% of the time using the entire dataset.

HDF5 (minutes)	ThunderPack (minutes)
2188	213

3.3. Read Speed

For tabular data, dataframe size was the largest determining factor. Therefore, 3 principal dataframes of increasing size were chosen for illustration.

10 Runs mean for 3 dataframes		
Dimensions	HDF5 (ms)	ThunderPack (ms)
3x5	5.6	2.8
4x6	8.1	2.8
20x5	27.9	2.6

For free-text, the ThunderPack approach showed across-the-board improvement.

Mean text read time across all samples	
HDF5 (ms)	ThunderPack (ms)
2.23	0.03

For images, the ThunderPack approach showed improvements in speed only on subsequent reads, which makes it a better approach for training models.

Mean image read time across all samples			
HDF5 Uncached (ms)	HDF5 Cached (ms)	ThunderPack Uncached (ms)	ThunderPack Cached (ms)
8.4	1.1	14.1	0.1

4. Skills and Knowledge Gained

This was my first time working with GCP and I found this experience valuable. Working intimately with the Linux shell and going back-to-basics was a great refresher.

This project was more open-ended. I was given sound guidance and room to think and overcome the challenges I was being faced with. It allowed me to think hard about data

infrastructure, design and storage as opposed to purely thinking about modeling. While the delays hindered my progress, they showed me that I could make do with less and still come out with some results.

5. Challenges and Solutions

The main challenge I faced was data access. I have yet to receive access to the main dataset and therefore had to apply for access to secondary datasets and generalize them with multimodality. Along with the waiting, this process ended up being time-consuming.

The second challenge I faced was aggregating multimodal data per-patient. This was necessary to make the move from JEDI feasible. Initially, we wanted to aggregate data similarly where each patient had a single file. I proposed the idea of a “pseudo-tree,” in which we could chain keys per-patient and store values in different formats and still maintain the “per-patient” quality. This idea was accepted.

6. Conclusion

My experience was very positive. Moving forward, the challenge of data infrastructure will be at the forefront of my thinking. My career will be colored by this less-prioritized aspect of data science, enabling me to generate more value for the field I’m in.

Please note: upon joining the Broad I signed an agreement that all my work belonged to them, along with the fact that my notebooks clearly showed data that required signing a DUA and submitting an application for access, I did not share my notebooks.

7. Acknowledgments

Christopher Reeder, for being a supportive mentor and giving me room to develop my own ideas and solutions.

Shadi Zaheri & DSP Methods Group, for fostering a very welcoming office environment for all of us interns.

KAUST Academy, for offering me an invaluable opportunity and for their tremendous support.

8. References (if applicable)

CXR Images
Kermany D, Goldbaum M, Cai W et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018; 172(5):1122-1131. doi:10.1016/j.cell.2018.02.010.

MIMIC-IV
Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV (version 2.2). PhysioNet. https://doi.org/10.13026/6mm1-ek67.

ThunderPack

Jose Javier Gonzalez Ortiz, The ThunderPack Data Format (2023).
https://github.com/JJGO/thunderpack/