Accelerated MatrixIRLS

Nicholas Cassarino
Tyler Allen & Christian Kuemmerle (Advisors)
UNC Charlotte

Introduction

- Matrix Factorization is a machine learning model with applications:
  - Recommender Systems
  - Search Engines
  - Medical Diagnostics
  - Feature Reduction
  - Image Compression
- Prior works in this field such as ALS++, ALS, and ICD all utilize GPU parallelization to make their algorithms as fast as possible.
- In prior works MatrixIRLS was introduced as a fast and accurate Matrix Factorization Model for incomplete data.
- In this work, our goal is to accelerate MatrixIRLS using GPU parallelization
  - We expect to achieve faster speeds than all prior works listed above. While retaining our accuracy.

Motivation

- In prior works MatrixIRLS was introduced as a fast and accurate Matrix Factorization model for incomplete data.
  - If you have incomplete data, SVD and other basic Matrix Factorization models do not work.
  - MatrixIRLS is specialized for incomplete data as that's a common problem in Recommender Systems.
- Large Matrix Factorization problems are slow.
- Other works in this field utilize parallel computing, causing them to be much faster than our sequential implementation.

Method

Converting Written Code

- In prior work MatrixIRLS was implemented in Matlab as a proof of concept.
  - This code is sequential
- First, we are currently working to convert Matlab code into sequential C code.
  - During this construction period we gain a stronger understanding of the algorithm.
- Using this C code we can begin finding areas of code that could be sped up using parallelization
- Finally, we will convert this C code into CUDA code that can utilize GPU parallelization to improve our performance.

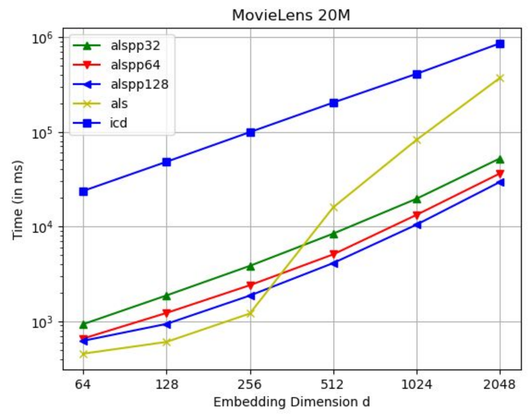
Background

- Recommender Systems are a class of machine learning algorithm used to predict what a user may enjoy and recommend it to them.
- Matlab is a programming language.
  - Popular with STEM
  - Comes with many tools for matrix and vector operations
- GPU - Graphics Processing Unit
  - Can be used to accelerate problems using parallelism
- Parallel Computing
  - GPUs and other specialized hardware are capable of computing multiple independent operations at the same time.
  - This is great for linear algebra since many linear algebra operations contain many independent computations

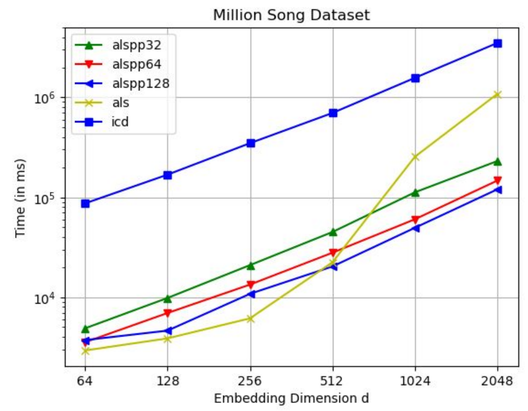
Results

Benchmarks for Prior Works

- ALS++, ALS, and ICD are algorithms from prior works that utilize GPU parallelization. Our goal is for MatrixIRLS' parallel implementation to be faster.
- I reproduced these results so that we will have a baseline to compare our future work to.



Competing Algorithms run with embedding dimensions (d) on the MovieLens 20M dataset.



Competing Algorithms run with embedding dimensions (d) on the Million Song Dataset.

Conclusions

- MatrixIRLS is a matrix factorization algorithm that is as accurate and theoretically faster than competing algorithms.
- MatrixIRLS is currently written in sequential Matlab code.
- In this work we are converting MatrixIRLS into parallel C code in order to reach the speeds of competing algorithms.
- I have reproduced the results of ALS++, ALS, and ICD.
  - These algorithms all utilize GPU parallelization to be as quick as possible.
  - We can use these results as a baseline to compare our future work with.
- In the future, we expect our accelerated MatrixIRLS algorithm to be faster than the competing algorithms while keeping the same accuracy.

References

Rendle, Steffen, et al. "IALS++: Speeding up Matrix Factorization with Subspace Optimization." ArXiv.org, 26 Oct. 2021, https://arxiv.org/abs/2110.14044.

Kuemmerle, Christian, and Claudio Mayrink Verdun. "International Conference on Machine Learning (ICML)." International Conference on Machine Learning (ICML 2021 Online, 18-24 July 2021), 2021.