Project Proposal ECE 18645

Team Members: Cheng Jiang, Yihang Liu, Jiajun Wan
Andrew ID: cljiang, yihangl, jiajunw2
Domain: Machine Learning

Abstract

Deep Clustering with Self-supervision using Pairwise Data Similarities(DCSS) is an efficient framework for data clustering. It focuses on the relevant useful information between pairwise data relationships. In this proposal, we will propose an idea of increasing the speed and efficiency of this algorithm by tackling all three stages of this method.

Algorithm

We aim to solve the problem of Clustering. Specifically, we want to implement a Pairwise Clustering algorithm. In this algorithm, we are implementing deep clustering with self-supervision using pairwise data similarities. The algorithm we will be using runs on Pytorch. The codes are written in Python where Pytorch is used as an imported library for machine learning evaluation. For the benchmark, we will have the current performance with the algorithm described above, which contains three stages: AE-based data aggregation, AE latent space mapping, and then DCSS for data clustering.

All these will be running on the CPU for maximum efficiency. We can see our work improving the speed and efficiency of the algorithm in all three of these stages: for the first stage, we can search for a method that efficiently uses the registers without the delay of latency. On the second stage, for the mapping purposes, how do we speedily match the latent from pre-processed data points. Finally, decreasing the DCSS data clustering latency can be another sub-problem we aim to tackle.

CodeBase

The code base of the algorithm can be reached from the link: Pairwise-Clustering. This code runs on CPU and our improved algorithm will be tested on the same machine for comparison.

DataBase

We will be testing our algorithm along with the benchmark on the dataset provided by the paper; example of some of our datasets include:
(1) MNIST [51] consists of 60,000 training and 10,000 test gray-scale handwritten images with size 28 × 28. This dataset has 10 classes.
(2) Fashion MNIST [52] has the same image size and number of samples as of MNIST. However, instead of handwritten images, it consists of different types of fashion products. This makes it fairly more complicated for data clustering compared to the MNIST dataset. It has 10 classes of data.
(3) 2MNIST is a more challenging dataset created through concatenation of the two MNIST and Fashion MNIST datasets. Thus, it has 140,000 gray-scale images from 20 classes.

All these dataset points can be accessed by an open-source dataset library.

Architecture/Platform

We will be using ECE Clusters for this project. The ECE Clusters have Intel CPUs to run our algorithm.

Work Division

We will divide our work into three parts according to the three phases(stages) of the Pairwise Similarity Clustering method:

- Jiajun will work on Phase one, AE process of the data points
- Yihang will work on the AE data points latent mapping algorithm
- Cheng will work on improving the algorithm for feature clustering

While the all three stages are correlated one after the other, we will conduct work individually at the same time. Each of the team members will work on the process weekly, while reporting and communicating with each other regularly.

Rough Work Schedule

Timeline	Sept	Oct	Nov	Dec
Planning Process	Familiar with the current algorithm and analyse where to improve	Design the Kernels for the algorithm	Implementing + Testing + Feedbacks	Summarize+Final Presentation

Date/Time of next meeting

Monday 11:00AM-12:00PM	Tuesday 5:00PM-6:00PM	Wednesday 11:00AM-12:00PM

Reference

1. Armanfard-Lab. “DCSS/Codes at Main · Armanfard-Lab/DCSS.” GitHub. Accessed September 25, 2021. https://github.com/Armanfard-Lab/DCSS/tree/main/Codes.
2. Sadeghi, Mohammadreza, and Narges Armanfard. “Deep Clustering with Self-Supervision Using Pairwise Data Similarities.” figshare. TechRxiv, June 29, 2021. https://www.techrxiv.org/articles/preprint/Deep_Clustering_with_Self-supervision_using_Pairwise_Data_Similarities/14852652.