Final project for Advanced Machine Learning course by Aigerim Rymbekova

Topic: Comparative Single-Cell RNA-sequencing Data Dimensionality Reduction with Principal Component Analysis and Autoencoder

Overview

In this project, I aim to address two important analytical steps for single-cell RNA-seq data - dimensionality reduction and clustering. Prior to the golden standard technique in this field - t-distributed Stochastic Neighbor Embedding (tSNE), I am going to perform pre-dimensionality reduction steps in linear way with PCA and in non-linear way with Autoencoder neural network and compare the outcomes.

Introduction

Single cell sequencing examines the sequence information from individual cells, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell.

Single-cell RNA-sequencing (scRNAseq) technique was recognized as the 2018 Breakthrough of the Year due to its potential for advances in research and medicine. scRNAseq data is highly dimensional: up to thousands of genes in millions of cells. In order to make sense of the data, we can project and visualize it as 2D or 3D. In such a setting, we are able to better distinguish transcriptionally various cell types as clusters.

In contrast to linear techniques such as PCA, Autoencoder, an unsupervised learning algorithm, is often used for dimensionality reduction in a nonlinear way and hence can capture highly non-linear structure of single cell data.

I would like to demonstrate the difference in single cell resolution between linear and nonlinear dimensionality reduction techniques using a dataset from Kaggle https://www.kaggle.com/chrispr/single-cell-rna-seq-from-stoeckius-et-al-2017 from this paper https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5669064/#SD2.

Problems and Challenges

There is a range of challenges when it comes to scRNAseq data analysis. Prior to embedding the data into a 2D visualization, I am going to follow the Seurat clustering workflow https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html for data pre-processing (quality control, normalization, data correction, feature selection etc.) For visualization purposes, tSNE technique is the most common, however it scales poorly when numbers of cells reach hundreds of thousands and millions. Therefore, first I would like to perform PCA and Autoencoder dimensionality reduction, feed the outputs into the tSNE and compare. The outcome hopefully will show that deep learning approaches are promising for improving single cell resolution.