Single-Cell Cancer Biopsy Prediction with Transfer Learning

People: Mica Haney (micahaney@my.unt.edu), Mani Dhakal (manidhakal@my.unt.edu), Danyang Shao (danyang823@gmail.com)

Intro

This project intends to tackle the field of single-cell biopsy cancer identification. Currently there is not a lot of data available to train a machine learning model on, so any models trained with an end-to-end approach will not be robust or include many cancer varieties, particularly rare ones, and will possibly perform poorly as well. This problem will be tackled with transfer learning, a method of fine-tuning that allows for the use of a large source dataset to build the initial representations of the domain data while a small target dataset tunes the model to the specific data for the task at hand. This has the potential to create a robust, well-performing model that can aid in the diagnosis of cancer types based on this biopsy technique, which could expand the usefulness of the biopsy itself.

Data

The data used for this project will consist of two sets, a source set and a target set. Both sets of data will be collected from the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO), which is a publicly available genomics repository (https://www.ncbi.nlm.nih.gov/geo/).

The source set will be used to pre-train the initial model. This set will consist of data collected from multi-cell cancer biopsy RNA sequences. As this data is the reconstruction of an entire genome, it does not match the format of the single-cell data. To remedy this the samples will be segmented to resemble the length distribution of the target set. Further preprocessing may need to be done.

The target set will be used to fine-tune the final model. This set will consist of data collected from single-cell cancer biopsy RNA sequences. A variety of cancer types will be included, some which are found in the source set and many which are not. A set of low-sample diseases will be held out to attempt few-shot learning with.

Milestones and Experiments

At current the tentative schedule is to have all data loaded into a Google drive and at a minimum 50% preprocessed, with a preference for being fully preprocessed, done before October 31st. All preliminary models should be constructed by the same Friday, November 4th. A complete, full-data training cycle should begin no later than November 11th, hopefully sooner. Experimentation should end on December 1st, with the write-up taking the remaining time.

At minimum two architectures should be experimented with, preferably more. The current architectures in consideration in order are transformers, dense, and CNN. Few-shot learning is an intended experiment if time allows.