Thesis Project Portfolio

Deep Multimodal Representation Learning to Integrate Natural Language Processing with Genomic Interval Data for Tailored Biomedical Discovery

(Technical Report)

The Perpetuation and Exacerbation of Racial Bias in Healthcare Through the Use of Machine Learning

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Zachary Mills

Spring, 2024

Department of Biomedical Engineering

Table of Contents

Sociotechnical Synthesis

With the recent boom in big data analysis, artificial intelligence, and machine learning, the medical industry has been flooded with large amounts of data and new technologies. The technical project utilizes a subset of machine learning called natural language processing to analyze and generate new genomic interval data based on user entered text. The STS research paper performs an examination of the use of machine learning in the healthcare industry and how it has not only perpetuated but exacerbated the presence of racial bias in medicine. Both of these papers look at machine learning technologies operating in the healthcare space through different lenses. When developing a new technology, it is important to analyze all aspects of its potential effects. Because of the data dependence of machine learning, it is especially crucial to analyze bias as is illustrated in the STS research paper.

The amount of genomic interval data from ATAC-seq and ChIP-seq experiments has risen dramatically over the past 10 years, increasing exponentially as sequencing technologies continue to improve. This large amount of data is currently difficult to navigate and can cause researchers to spend lots of time merely searching for the data they want. Despite the large amount of data being produced, there still remain many conditions without enough data due to being extremely rare or hard to study. The technical project developed four different machine learning algorithms to be able to generate novel genomic region sets, contained in BED files, based on user entered English language text. This project utilized Text2Bed, direct encoder, transformer, and diffusion model architectures to achieve this goal. Utilization of these models will allow researchers and other users to find relevant data quickly and accurately to their search, largely cutting down on the previous time spent searching for data. It can also generate new data for those conditions with small amounts of already existing data.

The models developed in the technical project are only a small part of the recent increase in machine learning research in medicine. These new technologies are now making their way into clinics, where their true effects can be seen. The STS research paper dives into the troubled start for machine learning in healthcare and how it has perpetuated racial bias in the industry. The paper discusses how bias enters machine learning through both the algorithms themselves and the data used to train them. There is a large focus on the idea of proxies and how ubiquitous social biases are captured in data even when perfect data collection techniques are used. The examples of the vaginal birth after cesarean (VBAC) calculator, a risk score algorithm, and a clinical scheduling algorithm are used to illustrate the different ways racial bias has emerged in healthcare machine learning. The analysis of these cases shows that race-blind approaches are often not enough to remove bias, and that race-aware methods are typically needed to reduce bias. Finally, the paper proposes the adoption of mandatory bias testing for prospective new technologies in order to prevent the systematization of racial bias in the healthcare industry.

Looking at these two papers together shows both the development process of a medical machine learning model, and a sociotechnical analysis of such technologies. Machine learning has the potential to revolutionize the way health care is managed and performed, but it also poses the danger of systematizing the racial disparities suffered by many. It is important that action is taken now in the early stages of this technology before more damage occurs. With proper oversight, machine learning can make health care better and more accessible for all. This shows how it is important to consider the racial and other possible unintended bias effects of a machine learning model while it is still under development.