ML4FG Proposal: Using Transfer Learning to Enhance Microbiome Analysis

George Austin, gia2105

March 2021

1 Overview

The goal of this project is to determine if we can improve our ability to detect signals in microbial samples by transferring information across datasets. This will be done by extending the results from DeepMicro (1), a study using autoencoders to do predictive analysis on 6 datasets, with predictive tasks of inflammatory bowel disease, type 2 diabetes in European women, type 2 diabetes in Chinese patients, cohort obesity, liver cirrhosis, and colorectal cancer (all processed data is available in their linked github repository). While DeepMicro used an autoencoder structure, we will extend our baseline analysis by also exploring variational autoencoders, similar to models that have been used for metagenomic binning in microbial datasets (2). We will also consider models like random forest, logistic regression, and feed forward neural networks in our baseline analysis.

The transfer learning will be explored from multiple angles. The first is to simply train the autoencoders on multiple datasets, which could allow for a model to better detect signals that span across different studies. Next, we will try training both the autoencoders and FFNNs using hard parameter sharing (3), which could allow models to share some signals across datasets, but still maintain the flexibility for some task-specific details. Finally, we can try using Model Agnostic Meta-learning (4) for all the deep-learning structures, as a way to develop good parameter initializations for their task-specific training.

If we can estiablish significant improvements from the transfer-learning (or metalearning) approaches, we will have built a flexbile, reusable framework to improve many data-hungry models' predictions from microbial datasets. This would be a major contribution, as it is typically difficult to successfully apply deep-learning methods to microbiome datasets, as they generally have both a small number of samples and a high dimensionality. Better application of deep-learning models to microbiome analysis will facilitate both earlier detections and medical interventions for various diseases, which in turn would lead to better recovery rates. In addition, this will improve our ability to understand how the human microbiomes (gut or other) relate to different diseases, potentially leading to developments of new treatment methods.

2 Goals for the midway point

1. Have all the data loaded. (This is already done. I can also confirm that there is a good overlap in the bacterial species used across datasets, so there is opportunity to apply the transfer learning methods here)
2. Understand where the data is coming from, i.e. methods, populations, study requirements, goals of each study. If necessary, have some tables/plots summarizing key details.
3. Complete the baseline analysis. This deliverable is a table in a 1-row-per-model format showing different metric performances. These results should be similar to what has previously been published for these datasets. Models used here should (at least) be logistic regression, random forest, feedforward neural network, autoencoder, and variational autoencoder.

4. have completed some exploratory EDA to understand the differences between the 6 datasets. If necessary, perform dimensionality reduction techniques to account for batch effects across the datasets. The main deliverable is either a PCA/tSNE plot indicating there aren't significant batch effects, or a series of plots to demonstrate that the batch effects have (at least in part) been removed.

3 End of Project Deliverables

1. Have a table summarizing performance metrics for the different approaches. The performance will be compared to baseline approaches using metrics such as AUC, precision, recall, and f1 scores. We can also produce barplots visualizing the AUCs, similar to what is shown in DeepMicro's results (1).
2. For the encoder structures, produce visualizations of the output to demonstrate if the shared information across datasets did have a significant impact.
3. If we do find that there is improved performance resulting from the transfer learning, we can use methods like LIME (5), DeepLift (6), or other examples covered in class, to identify why the transfer learning models do better (Given how the autoencoders might be paired with non-gradient based models, LIME might be a more appropriate choice). If possible, produce lists of bacterial species that become relevant to a predictive task once we introduce transfer learning.
4. If possible within the timeframe of the project, apply these frameworks to another group of datasets, compiled from https://gmrepo.humangut.info/. Will most likely not be able to do this item, but adding it in just in case.

References

[1] Oh, Min, and Liqing Zhang. 2020. "DeepMicro: Deep Representation Learning for Disease Prediction Based on Microbiome Data." Scientific Reports 10 (1). https://doi.org/10.1038/s41598-020-63159-5.

[2] Nissen, Jakob Nybo, Joachim Johansen, Rosa Lundbye Allesøe, Casper Kaae Sønderby, Jose Juan Almagro Armenteros, Christopher Heje Grønbech, Lars Juhl Jensen, et al. 2021. "Improved Metagenome Binning and Assembly Using Deep Variational Autoencoders." Nature Biotechnology, January. https://doi.org/10.1038/s41587-020-00777-4.

[3] Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks, 2017; arXiv:1706.05098.

[4] Chelsea Finn, Pieter Abbeel and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, 2017; arXiv:1703.03400.

[5] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016; arXiv:1602.04938.

[6] Avanti Shrikumar, Peyton Greenside and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, 2017, PMLR 70:3145-3153, 2017; arXiv:1704.02685.