USC Viterbi
School of Engineering

Bias in Machine Learning Models

Luke Pratt: lukepratt3@gmail.com
Fairfax High School: Class 2024
USC Viterbi Department of Computer Science, CS Theory Group SHINE 2022

SHINE
Summer High School Intensive in Next-Generation Engineering

Introduction

I am Luke Pratt, and I am in the Computer Science Lab under Professor Sharan and my Ph.D. mentor Bhavya Vasudeva. I worked with my mentor with a project goal of identifying and eliminating bias in machine learning models.

Machine learning is training a model to learn from data and let it predict outcomes with accuracy without being specifically coded to do so. The main problem is bias in predictions. When the model makes a prediction it will use other features that aren't related to make a prediction. An example is predicting eye color based on gender.

My project focuses on bias in facial recognition, specifically hair color. The base of my program uses photos of faces and tells if they have blonde or dark hair. I have been researching the bias based on the correlation of gender and hair color. This is because the majority in the samples are blonde females and dark hair males, while the minority are blonde males and dark hair females.

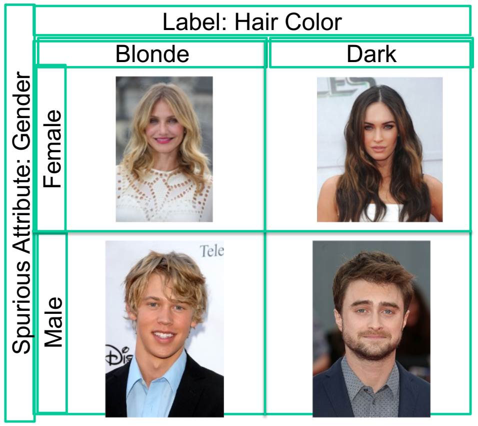

Figure 1: All 4 majority and minority groups with the label and spurious attribute.

Objective & Impact of Professor's Research

Machine learning is becoming a huge part of computer science, and one of its biggest problems is relying on spurious features and making biased predictions. The CS lab researches how bias affects machine learning models and how to mitigate it to make these models robust and reliable.

Skills Learned

During the SHINE program, I first improved my knowledge in coding, especially Python and PyTorch, which is an open-source machine learning framework. I used these to write my code for the program.

The machine learning model trains multiple times over the data with the number of times influencing how accurate the prediction. The accuracy is also affected by the learning rate which determines how big the size is for creating loss in the function. This model now becomes a neural network since it learns on its own and can make an accurate prediction.

Next, I learned about various image datasets and how to use them to visualize images and train a model. Specifically, a dataset called CelebA which holds pictures of many celebrity faces and labels for different attributes such as gender, hair color, etc.

The final skill I learned before putting it all together is training a classifier, requiring two important steps. First, to train a classifier, you run samples of classes through a learning model to identify what constitutes a given class. Second, you test the trained classifier to demonstrate if it has accurate predictions.

Figure 2 and 3: The logos for Python and for PyTorch

My STEM Coursework

I have always been connected to STEM, starting from elementary. I started coding when quarantine first hit. That summer, I joined a program called The Hidden Genius project where I learned HTML, CSS, and some Python. The next summer, I joined this program SHINE and really expanded on Python. I created a program that could read emotions in real time. Without a full experience in SHINE since I was stuck at home I went again this summer. I got to improve my Python and actually be on campus. I will use the coding I have learned for future programs and college.

How It Works

Using the dataset CelebA I trained the model to identify if the person had either blonde or dark hair. Then to locate bias I split the majority and minority groups for blonde hair and the spurious correlation being gender. In my research I separated the data into imbalanced percentages that range from 10%, 20%, 30%, and 40% as the minority group. I trained for all 4 separate groups and got their direct accuracies for the test data. Every 10% raises the accuracy for the minority.

With the percentages gathered you can clearly see the bias and where the bias drops off. Figure 4 shows the averages of the test data vs the train data at each imbalanced percentage. We can see with each increase of the minority group the test accuracy goes up while the training data goes down. This shows that training has a higher bias than the test. Figure 5 shows the minority and majority percentages of both the train and test data separately. You can see both minorities rose similarly with each percentage of minority, while the test majority declined slightly compared to the train majority which declined further.

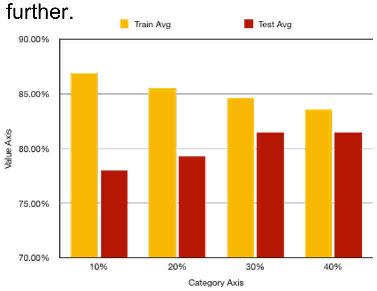

Figure 4: The average percentages of the Train and Test data for each minority increase.

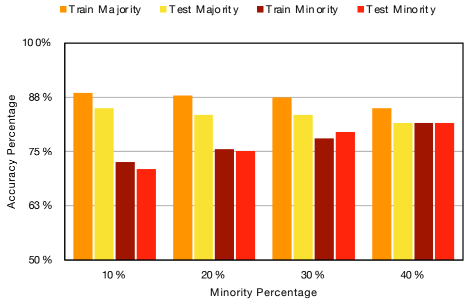

Figure 5: The accuracy percentages for the minority and majority in Test and Train data.

Advice for Future SHINE Students

My piece of advice for future SHINE students is to have fun but be productive. The program is seven weeks but it can go by so fast. You have to make sure you can get everything out of the time you spend in your lab, and also have a good time doing it. Don't just spend all your time in the lab, explore the campus and make friends. SHINE is a great experience and if you put a lot into it, you'll get even more out of it.

Acknowledgements

I would like to thank Dr. Mills and Ms. Lopez for leading the program, Professor Sharan for letting me into his lab, Bhavya Vasudeva for being my mentor and teaching me throughout the SHINE program, my center mentor Ms. Emelle, and my parents for supporting me during the process.

Citations for Articles:

1. [Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in terra incognita. In Computer Vision – ECCV 2018 (pp. 472–489). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-01270-0_28}
2. [Finn, E. Z. L., Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, Chelsea. (n.d.). Just train twice: Improving group robustness without training group information.]