HARRISBURG
UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Determining The Efficacy of Feature Attribution Explanation Methods to Reduce Bias in Machine Learning Algorithms

Matthew R Sabo

12/11/2024

ABSTRACT

In the past decade, society has seen a rapid increase in the use of Artificial Intelligence in both personal and professional settings. As Artificial Intelligence becomes more prevalent in society, concerns about the potential effects of bias increase as well. The previous paper written by the author on this topic, *An Investigation into the Effects and Ethical Implications of Bias in Machine Learning Algorithms,* went over the current state of ethics in the field of artificial intelligence and methods that have been developed to mitigate bias in machine learning algorithms. This paper is a continuation of that project, which aims to test the efficacy of the method outlined in the paper *Feature Attribution Explanation to Detect Harmful Dataset Shift* and determine if the method is effective in reducing bias in machine learning algorithms. The method achieves this by using a gradient based feature attribution explanation to capture the knowledge of the model and combine it with a widely used two sample test method, maximum mean difference, to detect harmful dataset shifts (Huang, Wang, & Yao, 2023). Unfortunately, because of circumstances outside of the author's control like file corruption, time constraints, and inability to obtain the datasets used in the original paper, they were never able to test the method like intended. But with the method outlined in this paper, it would be expected that it would adequately test the method created in the paper *Feature Attribution Explanation to Detect Harmful Dataset Shift* and would be able to determine if it would be an effective method in reducing bias in machine learning algorithms.

1 Introduction

Before discussing the method that was used for this paper in detail, it is important to first know the different ways that dataset shifts can occur in a dataset while training a machine learning model. This section is intended to define the types of dataset shifts that can occur in machine learning algorithms while they are training. With this knowledge, it would then be possible to move forward and expand the method developed in *Feature Attribution Explanation to Detect Harmful Dataset Shift*.

1.1 Background

With the increasing prevalence of Artificial Intelligence in everyday life, it is important to have methods developed to mitigate the amount of bias in a machine learning algorithm. One method that has been developed is called Feature Attribution Explanation (FAE). These methods indicate how much each input feature contributes to a machine learning model's output for a given datapoint. This can be used to determine if a potentially harmful shift occurs while training a model and be corrected. There are 3 different kinds of shifts that can occur in a dataset: Covariate Shifts, Label Shifts, and Concept Shifts. Covariate Shifts are defined by the changes and differences in the distribution of the input variables between the training and test data. Label Shifts are defined by the changes and differences in the distribution of the target variables between the training and test data. And concept shifts are defined by the changes in the relationship between the input variables and the class variables. With knowing these shifts, it is possible to determine if a shift is harmful and mitigate the potential negative effects of the shift (Huang, Wang, & Yao, 2023).

2 Methodology

This section will outline the changes that were made to the experiment that was conducted in the paper, *Feature Attribution Explanation to Detect Harmful Dataset Shift.* It will briefly describe how the method works and then explain how this experiment deviates from the original. It will also describe the environment that the experiment would have been run on in both terms of computer hardware and environment.

2.1 Experiment Environment

The environment that this research would be run on is running Windows 10 with an AMD Ryzen 7 5800x CPU and a NVIDIA 2070 Super GPU. The code base pulled from this GitHub: https://github.com/oddwang/FAE-DHDS and modified to update the syntax since it was outdated. The code base is written in Python and ran in Python 3.7. The python libraries that was prominent in the code base was Tensorflow, Keras, and Numpy which were used to train models, develop models, and prepare the dataset for processing by the models but all libraries needed to run the files were listed in requirements.txt.

2.2 Experiment Setup

The experiment in the paper *Feature Attribution Explanation to Detect Harmful Dataset Shift* focuses on detecting harmful covariate and label shifts. To facilitate this, this experiment was run on three publicly available image datasets, MNIST, Fashion-MNIST and CIFAR-10 with the main classification model used being ResNet-50. The datasets were first randomly divided into training, validation, and test sets. Training sets would be used to construct the model and various covariate and label shifts would be applied to the test set to simulate dataset shifts that may occur in training the model while the validation set would be see if the models could detect the shifts applied to the test set (Huang, Wang, & Yao, 2023). These simulated shifts are as follows: Gaussian Noise, Image Rotations and Zooms, Adversarial samples generated by a fast gradient signed method, Removal of a certain percentage of class 0 samples from the test set, Combination of img and ko, Combination of class 0 images and img. These shifts are applied to either 10%, 50%, or 100% of the test set for each shift type. Then after calculating the dataset shifts that occurred with Grad Saliency which denotes the importance of each input to the models output and Grad*Input which is the like Grad Saliency but clarifies the explanation result by including the input data on top of Grad Saliency (Huang, Wang, & Yao, 2023). It then takes the difference between a test dataset that was modified with simulated shifts and a control test dataset and determines if there has been a shift and if it was harmful (Huang, Wang, & Yao, 2023).

2.3 Experiment Expansion

To make this experiment more robust, the author decided to expand it in two different areas. The first change focuses on expanding the percentages that simulated shifts are applied to of the test dataset to include 25% and 75%. These percentages were chosen since in the original experiment, the set of percentages that the method used consisted on 10, 50, and 100% application. By adding both 25% and 75% application rate to the set of percentages, it would make the experiment more thorough and robust by closing the gaps that the original experiment had. The second expansion to the experiment would be the addition of testing another dataset like the publicly available dataset here: https://www.kaggle.com/datasets/muratkokludataset/pistachio-image-dataset. The addition of this dataset to the experiment would, like the previous expansion, make the experiment more robust and ensure that the results are consistent across many ranges of application percentage and datasets to make the experiment more trustworthy.

3 Results and Discussion

This section will discuss the circumstances and challenges that the author faced that made them not be able to complete this project. It will also discuss what would be changed if this project restarted and the lessons learned from this project.

3.1 Difficulties Carrying out the Experiment

There were many obstacles that occurred while trying to recreate this experiment which made the project impossible to complete in time. These issues include time constraints, broken syntax, inability to obtain datasets, file corruption, and more. This section is intended to outline the issues that occurred during this project and the lessons learned from these issues.

3.1.1 Time Constraints

To begin, this paper was originally supposed to be written for another method that was researched and discussed in the authors previous paper *An Investigation into the Effects and Ethical Implications of Bias on Machine Learning Algorithms* which was Gender Masking. The idea and execution around the original topic and this topic were the same, where the project would recreate the method developed in the paper that it was researched from and determine if it worked or not. However, in the middle of October, the author realized that it would not be possible to successfully recreate the Gender Masking method since the datasets and testing parameters from the original experiment were missing. So having to start over with about a month and a half until the project was due, and other responsibilities that the author had, there were significant time constraints that occurred that caused the project to be unsuccessful. If this topic had been chosen at the start of the 2024 Fall semester, this project would have had a higher chance of success.

3.1.2 Broken Syntax and Inability to Obtain Datasets

After restarting the project and downloading the code files from the original papers GitHub, it was assumed that it would just be able to run without any issues and that the project would be expedited to the result collecting phase. However, when ran the code base was discovered to have outdated syntax so the author had to spend around two weeks going through and bringing the code base up to date and it still wouldn't run. After some research, the author discovered that when uploading the datasets to GitHub the authors of the paper *Feature Attribution Explanation to Detect Harmful Dataset Shift* uploaded the database files as Git Large File Storage which allows users to replace large files with text pointers inside of Git while storing the file contents on a remote server like GitHub.com or GitHub Enterprise. This allowed the creator of the repository to upload .csv files with pointers to the contents of the files so that they would be able to share their code base. However, they did not pay GitHub for enough bandwidth, so it is impossible to download the contents of the database files from GitHub until the owner of the repository pays for more bandwidth. The author attempted to reach out by email to the authors of the paper being tested to see if they still had the database files but was unsuccessful in reaching them. Both having to completely rewrite most of the outdated code base and the difficulties in obtaining the datasets were two more reasons why the project was unsuccessful and could not be completed.

3.1.3 File Corruption

The most detrimental issue encountered during this project that caused the most damage to the progress of the project was the complete corruption of every file that was created or used in this project. Every file that was related to this project was uploaded to Microsoft OneDrive in a singular folder. This was used so that it would be easier to complete work between devices, allow for file backups, and more efficient workflow. However, on December 3rd, 2024, it was discovered that during a OneDrive backup, the folder that contained all of the information and work done on this project was completely corrupted and all of the backups were corrupted. And since this project was due less than a week from the event, it was the most detrimental issue that occurred during this project which now makes the author make backups of their OneDrive more regularly to avoid this in the future.

3.1.4 Changes to the Project if Redone

The issues that occurred during the process of this project were detrimental to its progress and without occurring would have significantly raised the chance of this project succeeding. If this project was redone, the changes that would be made is numerous but the most important ones is start earlier so that time constraints and other responsibilities do not become a limiting factor in the amount of time to work on the project, backup files more regularly, have more familiarity with certain programs and functions used during the project, and contact the authors of the original paper earlier.

4 Conclusion

In this paper, a method was developed to expand on the methodology outlined in the paper, *Feature Attribution Explanation to Detect Harmful Dataset Shift*. This expansion relied on expanding the percentages used to apply the simulated shifts to the test set, adding 25% and 75%. This addition would expand the testing range and [FINISH]. Unfortunately, this experiment could not be carried out by the author due to issues that came up during the length of this project like time constraints, the entire code base needing to be updated, the inability to obtain the datasets used in the original paper, and file corruption caused this experiment to not be carried out. However, the method described in this paper should be adequate in testing the method described in *Feature Attribution Explanation to Detect Harmful Dataset Shift* and when completed, would be able to determine if it is an effective method to reduce bias in machine learning algorithms.

5 References

Huang C., Wang Z., & Yao X. (2023). *Feature Attribution Explanation to Detect Harmful Dataset Shift*. https://doi.org/10.1109/ijcnn54540.2023.10191221