Replication research regarding the FairNeuron Algorithm

Maud Groen and **Jeroen Wasser** and **Mathieu Janssen**

Abstract

The *FairNeuron: Improving Deep Neural Network Fairness with Adversary Games on Selective Neurons* paper proposes a solution that would minimize sensitive bias in neural networks. Through identifying and selectively retraining of individual biased nodes. We replicated their method but did not achieve the same results found in the original paper. The code found in their GitHub Repository did not run as intended and we found their implementation to be different than mentioned in the paper. As such, we could not replicate their findings and conclude that the FairNeuron algorithm achieves different results than mentioned in the original paper.

1 Introduction

Deep Learning has found applications in many different areas of research. In their paper on fairness in Deep Neural Networks (DNNs), (Xuanqi Gao, 2022) develop an algorithm to improve the fairness of various DNNs. In their paper, (Xuanqi Gao, 2022) trains a DNN to predict rates of recidivism of criminals. This dataset contains a sensitive variable: race. A machine learning model could hyperfocus on this specific variable and thus become biased to specific races. It is of high importance that racial bias is mitigated, It has been established that various models are prone to racial bias. (Huang, 2022) (James Coe, 2012).

Instead of tackling the problem of racial bias at the level of data processing or model training, one could instead retroactively modify the trained model to reduce bias. Indeed, one of the main features of their research is the development of the FairNeuron algorithm. This algorithm takes a biased neural network as an input, and finds those nodes that contribute the most to the bias within the network. Then, through an novel algorithm, those nodes are removed and the network is retrained. One of their main findings is that this algorithm, the FairNeuron algorithm, reduces bias within a network, while having minimal impact on model accuracy.

In this replication paper, we will attempt to reproduce some of the results found in (Xuanqi Gao, 2022). More specifically, we will train a network on the COMPAS data set, perform the same hyperparameter tuning and retrain the DNN using the FairNeuron algorithm. We find that the results in their paper are not reproducible. The stated results for accuracy, DP and DPR are not found in our replication, and the provided code does not seem to work as intended.

2 Background

Deep neural networks (DNNs) are adopted in more applications every day, such as in artificial intelligence, image recognition and natural language processing. However, there is one big downside to DNNs, namely the fact that they do not take into account biases. Depending on different tasks, specific attributes in the data can contain sensitive information. It cannot happen that a model makes different decisions for different instances based on these attributes, such as race or gender.

This is where fairness comes in, which can be based on two different notions. Firstly, there is individual fairness, this measures whether individuals in the data set are treated equally. The other notion is group fairness, and concerns specific sub populations with different sensitive attributes. These groups should be treated equally as other sub populations.

As mentioned many machine learning models suffer from bias problems, since they make decisions based on wrong or sensitive attributes. Many different approaches have already been taken to overcome this issue, such as a fair adversarial framework (Tameem Adel, 2019), ethical adversaries (Pieter Delobelle, 2021) and pre-/post-processing methods (Faisal Kamiran, 2012). However, these models still have limitations, starting with the fact that they introduce another model as the adversary in the training procedure. It can be very hard to train these types of models, with issues such as mode collapse. Besides this, there is no guarantee that these models will converge.

This is where the authors come in with their own suggestion, namely the FairNeuron algorithm (Xuanqi Gao, 2022). They argue that an adversary does not need to be introduced in order to detect bias. With FairNeuron, they first monitor the training process to detect the biased neurons. So

the neurons whose accuracy and fairness optimization have conflicts with each other. After this, they identified samples that cause these contradictory optimizations. Then finally they "enforce the optimizer to decide a balanced optimal direction that optimizes both accuracy and fairness" (Xuanqi Gao, 2022).

3 Target Study Overview

3.1 Technique Foundations

Several techniques stand at the basis for the original paper. Namely:

1. The building, training and validating of neural networks.
2. The specification of interesting paths within an existing neural network.
3. The calculation of various fairness scores.
4. The slicing of neural networks.
5. The optimization of neural networks according to custom metrics.
6. Selective training of neural nodes

3.2 Technique Internals

1. The base network for this paper is a neural network with 3 hidden layers. The training, testing and validating of this network is done according to standard practices.
2. The authors of the paper designed the FairNeuron algorithm, which includes several subroutines.
3. The authors construct three different fairness scores using the output from the first neural network: Demographic Parity (DP), Demographic Parity Ratio (DPR) and Equal Opportunity (EO).
4. Extracting biased nodes from the neural network.
5. Specifically the retraining of a neural network while attempting to maximize fairness while minimally affecting accuracy.
6. Only retraining nodes that are biased to sensitive variables.

As mentioned, the FairNeuron algorithm uses various subroutines to find the paths of interest and perform the slicing. For the specific routines and pseudocode used, we refer to the original paper.

The validation dataset used within the paper contains only Caucasian and African-American individuals. Other races are not represented in the extract of the data. The validation dataset is a 10% split of the original dataset.

3.3 Study context and targeted domain

The study is primarily applicable to projects where sensitive data is used to make predictions that could influence the lives and prospects of individuals. It hopes to prevent sensitive bias being introduced that could unduly effect decisions made by such models.

3.4 Validation Approach

The validation methods used in the original paper are standard techniques for validating neural networks. The data is split into train, test and validation data sets. The neural network is optimized for a trade-off of fairness and accuracy. See Section 4.2 for more details.

3.5 Study limitations and replication problem statement

Some limitations were mentioned within the original paper. The application of the proposed solution on smaller datasets was mentioned specifically. They found that Fairneuron did not perform well on datasets with less than or equal to 600 instances. The researchers will try to address this limitation in future work. They also found that the performance of Fairneuron is not ideal on CNN models, as it can only be performed on the last full-connected layer.

Out of these findings we propose the following problem statement:

Due to the importance of unbiased modeling, the effectiveness of FairNeuron should be evaluated by a third party.

4 Replication Study Definition

4.1 Decision Tree Diagram

As this is a replication study, it's important to get an overview of what is being replicated from the original paper. The overall setup of this replication study can be found in Figure 1. The corresponding data flow diagram is given in Figure 2.

4.2 Results

Due to limitations of the available code, only the COMPAS data set could be used. This should be enough to test the workings of the FairNeuron algorithm however, since all data sets have at least one attribute sensitive to bias. Part of the FairNeuron algorithm includes two parameters, namely θ and γ. These hyperparameters are used internally to get the paths which are contributing the most to the unfairness in the network.

In order to get good estimates for the values of θ and γ to be used in the final algorithm, some hyperparameter tuning is performed by a grid search. Spanning over a search space of $(\theta, \gamma) \in \{0.1, 0.01, 3e-3, 1e-3, 3e-4, 1e-4\} \times \{0.95, 0.9, 0.85, 0.8, 0.7, 0.6\}$, we found the optimal solution to be $(\theta, \gamma) = (0.01, 0.95)$.

These values for θ and γ are then used to perform the FairNeuron algorithm on the base DNN. The sliced network is then retrained. The resulting metrics of the final DNN are given in Table 1.

The loss function used in the final retraining step is a weighted sum of the metrics listed in Tables 1 and 2. The target function is given by:

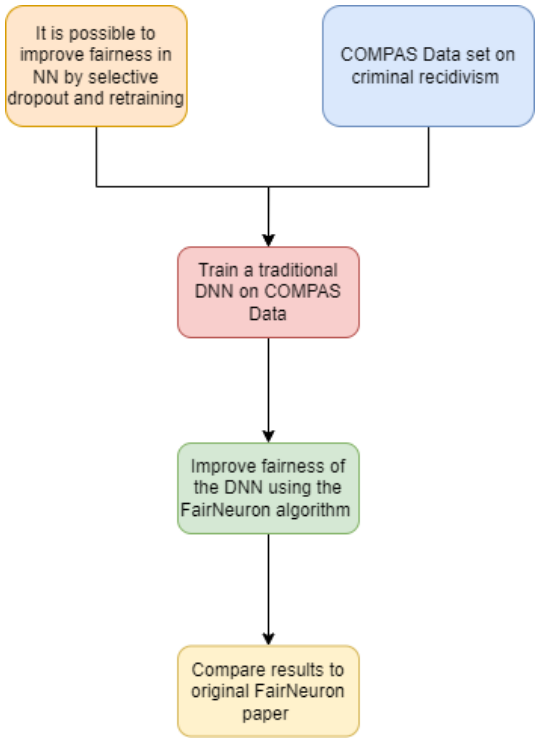

Figure 1: Decision Tree Diagram

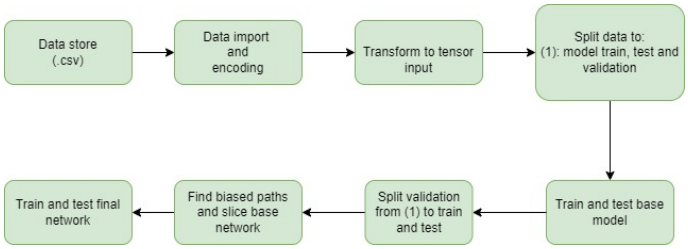

Figure 2: Data Flow diagram

Metric	Value
Accuracy	0.70
DP	0.32
DPR	0.55
EO	0.13

Table 1: Metrics of DNN after FairNeuron algorithm

$$ComplexScore = DP + EO + (1 - DPR) - 0.01 * accuracy \quad (1)$$

Using (1), the final value for ComplexScore in our model is 0.893. Note that accuracy is the term that contributes the least overall in general, meaning that the model is being optimized for fairness first and foremost.

4.3 Comparison to original paper

When comparing the metrics of the final DNN in this replication paper to the original, we notice the results do not seem to match. For comparison, the metrics found in the original paper for the COMPAS data set can be found in Table 2.

Metric	Value
Accuracy	0.799
DP	0.013
DPR	1.021
EO	0.058

Table 2: Metrics of DNN in the original paper

While the accuracy of both models does not differ too much, the fairness metrics seem to be much better in the original paper. For context, a value of DPR = 1 indicates a fair and balanced model, as do EO = 0 and DP = 0. This is reflected in the final value for ComplexScore; using (1) we indeed find the value for their model is 0.04201. Therefore, we can not conclude that the results in the original paper are replicable for the COMPAS data set.

As mentioned before, the code provided on their GitHub does not seem to work as intended. The re-training of the network after the hyperparameter tuning does not seem to work, as the network does not improve even after 50 epochs. The code provided in the GitHub has not been altered, other than getting it to run on CPU instead of GPU.

5 Conclusions

As mentioned in the background section of this paper, there are many different use cases where the FairNeuron algorithm can be applied. A frequently used example is whether an individual gets funding based on their characteristics (Huang, 2022). Since these characteristics contain many sensitive variables, such as race, gender and age, the FairNeuron algorithm would be appropriate here. The conclusion of this replication paper is that for these and similar cases, it is necessary to take fairness into account and to optimize the model based on fairness metrics. Therefore, our suggestion is to use the FairNeuron algorithm to limit sensitive bias within neural networks. Both the original research and the replicated technique do not show a big decrease in accuracy while improving on the various constructed fairness metrics, meaning that the accuracy of a neural network need not be compromised to create a more fair model.

It is important to note that the context of these techniques are not limited to one specific area of research. This is because sensitive variables come in many shapes and forms and should always be considered when doing analysis.

Because of lacking code and limited time and resources, it was not possible to replicate the entire paper. The code written and published by the authors was focused only on the FairNeuron algorithm. In the paper itself, FairNeuron

was compared with four different algorithms, namely ROC, reweighting, FAD and ethical adversaries. Since these algorithms were not included in the code, it was not possible to train and test these. Because of this, FairNeuron could not be compared to the others methods, meaning that no conclusion was drawn on the efficiency of FairNeuron compared to other options. This means that the only conclusion drawn about FairNeuron was about the scores of accuracy, DP, DPR and RO and whether these were the same as found by the original authors. Finally, the code was only available for one of the three datasets, namely the COMPAS dataset. So in this replication paper, the algorithm was only tested on the COMPAS dataset, not on the credit and census data.

The replication performed in this paper is a literal replication, meaning that an exact duplication of the previous analysis was performed. This type of replication was chosen based on the time and resource limits, but future research could focus on other types of replication. Operational replication could also have been possible, if there was enough time for finding a different dataset. The other two datasets used in the original paper are not available in the original code. So an operational replication would be a good option to check whether similar datasets yield similar results.

Another option is constructive replication, which would be a difficult procedure to use for this paper, since it means avoiding re-using procedures at all. Since the FairNeuron algorithm is created by the authors, with very specific training and testing procedures, it would be complicated to come up with a new procedure. The same problem would come up for instrumental replication, since the tools used in the research need to be changed. The recommendations for future research would therefore be to perform either a literal or operational replication, when working with this paper.

As mentioned before, only one of the three datasets could be replicated, namely COMPAS. COMPAS is a system used by judges for predicting the risk of recidivism, with the race of the defendant as the sensitive variable. The German credit dataset and the UCI adult census data could not be used. Because of this, the conclusions about the generalization of this model are limited. Since the model is now only tested on one dataset, it would be hard to say whether it is useful in other settings as well. What makes this even harder, is the fact that the data is on only 10,000 criminal defendants in Broward County, Florida (Xuanqi Gao, 2022). In total there are 28 variables in the dataset, including the sensitive variables. Because it is very region specific and there are not that many variables, the generalizability of this dataset might also be limited.

Overall, this replication paper can help understand FairNeuron and its implications better. However, it is important to keep in mind that there are still several limitations.

References

Faisal Kamiran, T. C. (2012). Data preprocessing techniques for classification without discrimination.

Huang, J. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review.

James Coe, M. A. (2012). Evaluating impact of race in facial recognition across machine learning and deep learning algorithms.

Pieter Delobelle, G. P. B. F. P. H. B. B., Paul Temple. (2021). Ethical adversaries: Towards mitigating unfairness with adversarial machine learning.

Tameem Adel, Z. G. A. W., Isabel Valera. (2019). Onenetwork adversarial fairness.

Xuanqi Gao, S. M., Juan Zhai. (2022). Fairneuron: Improving deep neural network fairness with adversary games on selective neurons.