Pranjal Dhakal

December 13, 2022

1 Abstract

The effectiveness of machine learning models across traits like race, gender, and age has drawn more attention in recent years. We can see a pronounced disparity in the test performance of machine learning models across different groups when data from some of the groups is missing or underrepresented during the training process. This is due to the possibility that the distribution of features among the various demographic groups might be different. In this project, we have used adversarial training to match the distribution of features across the different groups in an encoded space. So even when the machine learning model is trained using an imbalanced dataset, the test performance across these groups will be similar.

2 Introduction

Currently, many Artificial Intelligence systems face fairness challenges in delivering equitable results to different population groups [1]. Roosli et al. in their paper [2] show that the AUROC performance for the in-hospital mortality task is worse for test subjects with Medicaid insurance as compared to those with private insurance. Sungho Park et al. in their paper [3] point out the discrepancy in the TPR and TNR for different groups based on gender and age for the attractiveness prediction task using different machine learning models. These are some examples where there is a significant difference in the predictor's performance among different sensitive groups. The discrepancy in the performance of the model across different groups might be due to the model being trained on a biased dataset or the model learning to falsely associate sensitive attributes with the target.

Various techniques to address this issue have been explored [1]. Some of these include - distributed learning to overcome unfair dataset shifts, fair representation learning via disentanglement, and model auditing using interpretability. In this project, we have used adversarial training which is one of the techniques to address this issue.

We can express the dataset as $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{n}$. x_i represents the features, y_i represents the associated task label, and s_i is the group that the sample belongs to. Let $\eta : X \rightarrow \hat{Y}$ be a predictor that maps the features to activity labels. If the distribution of the features across all the groups is the same as in equation (1), then there should not be any discrepancy in the performance across these groups using the classifier η trained on any subset of this dataset as in equation (2). This should hold true even when one of the subgroups is underrepresented or absent in the training subset.

$$P(x|s = 0) = P(x|s = 1) \tag{1}$$

$$P(\eta(x) = y|s = 0, Y = y) = P(\eta(x) = y|s = 1, Y = y) \tag{2}$$

When we have access to the group membership information for each sample, we can create scenarios in which a group of the population is underrepresented or totally absent in the training subset. This is done to simulate a biased training dataset scenario. We will then evaluate and report the performance of the classifier trained on this dataset across the different groups for activity prediction. If we determine a significant performance discrepancy, we can confirm that the distribution of features across the groups is different. This situation might cause fairness issue where the model performs poorly for underrepresented minority group in the population. We will then use adversarial training [3], [4] where we train a representation learning network $\theta : X \to Z$ that maps the original data X to a latent space Z. We will try to match the distribution of features across the groups in this latent space.

3 Dataset

In this project, the Motion Sense Dataset [5] is used. This time-series dataset includes the data generated by accelerometer and gyroscope sensors for six activities collected from 24 individuals. Along with the sensor readings, the gender, age, personal identifier, weight, and height information of each observation are also provided. Mohammad Malekzadeh et al. in their paper [5] show that it is possible to predict the personal identifier information using the accelerometer and gyroscope features in this dataset. Ahmed Sharshar et al. in their paper [6] show that it is also possible to predict gender information from target-relevant features. In this dataset, we can see some correlation between the features used for activity prediction and the sensitive identifiers.

This time-series dataset can be expressed as $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{n}$. $x_i \in \mathbb{R}^{t \times d}$ represents the d sensors' readings captured for t sequential observations. $y_i \in \mathbb{R}$ is the activity label corresponding to x_i which can assume a value between 0 and 5 for one of the 6 different activities. We will consider 4 (”dws”: downstairs, ”ups”: upstairs, ”wlk”: walking, ”jog”: jogging) activities in our experiment. Activities ”sit”: sit and ”std”: standing are ignored as the signals are most likely going to be idle. $s_i \in \mathbb{R}$ is the sensitive attribute corresponding to x_i. We will consider gender as the sensitive attribute for this experiment so s_i can be 0 or 1 representing a male or a female respectively.

The observations in this dataset are collected using iPhone 6 stored in the front pocket of the participant's trousers. The activities performed are dws: downstairs, ups: upstairs, sit: sitting, std: standing, wlk: walking, and jog: jogging. Multiple trials of the same activity are recorded for all the participants in different locations. The sampling frequency for the observations in this dataset is 50 Hz. In each of these observations, there are 12 readings collected from the accelerometer and gyroscope sensors. Figure 1 shows the distribution of the number of observations for different activities across the male and female sub-population. There are 24 individuals in this dataset, with 14 of them male and 10 female participants.

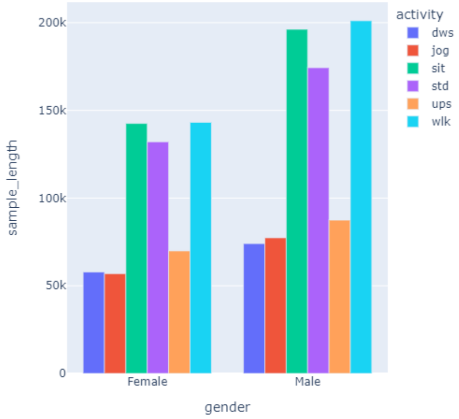

Figure 1: Sample distribution across genders per activity in Motion Sense dataset

4 Problem Formulation

We aim to train a representation learning network $\theta : X \to Z$ that maps the original data X to a encoded space Z which satisfies the objective in (3). The encoded data should minimize the information about the group while maintaining (c*100)% of the information related to the task.

$$\min_{\theta' : I(Y;\theta(X)) > cI(Y;X)} I(\theta(X);S),\ \textit{for}\ 0 < c < 1 \tag{3}$$

In this project, the target predictor is trained to predict the activity and the sensitive attribute is gender. Matching the feature distribution among the genders in the encoded space will result in very less information about the gender being inferred. The more dissimilar the distributions are, the easier it is to discern the male and female data i.e. more information about gender can be inferred from the data. When the features are similar across the different genders, the goal of reducing the gap in the test performance between the groups even when data from only one of them is used to train the target predictor can be achieved.

5 Approach

In order to train this representation learning network θ, we use two helper networks f and g. f is the activity prediction network that maps the representation from the encoder to the activity $f : \theta(x) \mapsto \hat{y}$. g is the gender prediction network that maps the representation from the encoder to the gender $g : \theta(x) \mapsto \hat{s}$.

We will jointly train the encoder and the activity predictor by minimizing the cost $\mathcal{J}1$ in (4). L_c is the cross-entropy loss.

$$\mathcal{J}1(\theta', f') = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} L_c(f'(\theta'(x)), y) \tag{4}$$

We will simultaneously train the gender prediction g network by minimizing the cost $\mathcal{J}2$ in (5) using data from both genders. The gender predictor aims to predict the gender information from the representation generated by the encoder network.

$$\mathcal{J}2(g') = \mathop{\mathbb{E}}_{(x,s)\sim\mathcal{D}} L_c(g'(\theta(x)), s) \tag{5}$$

These three networks θ, f, g are trained using a min-max game between the encoder network(θ) and activity predictor(f) versus the gender predictor(g) with the objective in (6). The encoder will try to minimize the objective by generating representations that will maximize $\mathcal{J}1$ and minimize $\mathcal{J}2$ i.e. the activity prediction ability is retained while making it harder to predict the gender. The gender predictor g will try to maximize the objective by trying to predict the gender from the representation generated by the encoder. After the training is complete, the distribution of the features between males and females should be similar in the output of the encoder which would satisfy (3).

In equation (6), α is used to control the influence of objectives $\mathcal{J}1$ and $\mathcal{J}2$. When a small value of α is used, more emphasis is put on the encoder to generate representations from which minimum gender information can be inferred. Doing this can compromise the activity prediction ability from the representation.

$$(\theta, f, g) = \mathop{\arg\min}_{\theta', f'} \mathop{\arg\max}_{g'} \left[\alpha\, \mathcal{J}1(\theta', f') - \mathcal{J}2(g')\right] \tag{6}$$

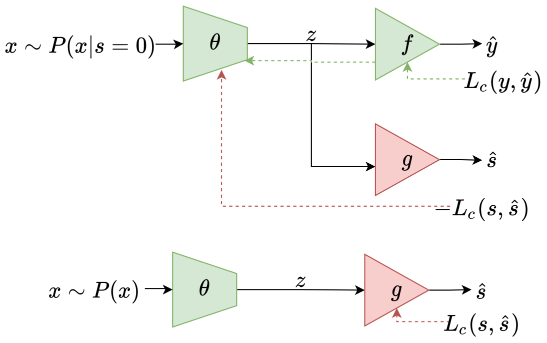

Figure 2: Proposed model. θ is the encoder network. f is the activity predictor and g is the gender predictor.

In figure 2, the training process is shown. We consider the data samples from one of the genders to train the encoder and activity predictor. The encoder and the activity predictor are trained to minimize the objective in (6). In order to do this, the encoder should produce representations that will be able to predict activity by minimizing the loss $L_c(y, \hat{y})$ but not the gender by maximizing the loss $L_c(s, \hat{s})$. The gender predictor can be thought of as the discriminator that is trained with the opposite goal to predict gender correctly by minimizing the loss $L_c(s, \hat{s})$.

6 Experimental design and Performance Metrics

6.1 Train and Test Sets

In this dataset, the personal identifier information is provided along with the genders. For both genders, we select different individuals in the train and test sets. We randomly select 2 male individuals out of 14 male participants, and 2 female individuals out of 10 female participants to form the male test and female test sets respectively.

There are multiple trials for the different activities in the dataset. The sampling frequency of the sensors is 50 Hz. We consider a window of 2.56 seconds to form each training sample. This way the dimension of each sample is 128 × 12. Each sample is associated with one of the 4 considered activity labels and one of the 2 genders. The validation set is formed from 10% of the training set and ensured that there is no overlap with the training set.

The experiment is performed 20 times with different individuals considered in the train and test sets. Validation loss is used for early stopping during the training process.

6.2 Network Parameters

6.2.1 Encoder

The encoder θ maps the input $x \in \mathcal{R}^{128\times 12}$ to the latent space $z \in \mathcal{R}^{6\times 12}$. The network details are as follows.

```
1. Layer: Type: Conv1D, filters: 64, kernel size: 3, padding: same
2. Layer: Type: Dense, units: 32
3. Layer: Type: Dense, units: 16
4. Layer: Type Dense, units: 6
```

6.2.2 Activity Predictor

The Activity Predictor f maps the encoded data $z \in \mathcal{R}^{6\times 12}$ to the activity label $\hat{y} \in \mathcal{R}$. The network details are as follows.

```
1. Layer: Conv1D, filters: 64, kernel size: 3, padding: same
2. Layer: Flatten
3. Layer: Dense, units: 16
4. Layer: Dense, units: 4
```

6.2.3 Gender Predictor

The Activity Predictor g maps the encoded data $z \in \mathcal{R}^{6\times 12}$ to the gender label $\hat{s} \in \mathcal{R}$. The network details are as follows.

```
1. Layer: GRU, units: 32
2. Layer: Dense, units: 64
3. Layer: Dense, units: 32
4. Layer: Dense, units: 16
5. Layer: Dense, units: 1
```

6.3 Performance Metrics

We will use the accuracy score to assess the activity prediction performance. The baseline used in this experiment is the activity performance without adversarial training. We will report the mean and variance of the accuracy score in the following settings.

1. *Male Train - Male Test, Male Train - Female Test*: In this setting, I will use the male training data to train the encoder and the activity predictor. Then the test performance for male and female test sets of individuals not used during the training will be reported. This will be done with the baseline and post-adversarial training.

2. *Female Train - Male Test, Female Train - Female Test*: In this setting, I will use the female training data to train the encoder and the activity predictor. Then the test performance for male and female test sets will be reported as above. This will also be done with the baseline and post-adversarial training.

7 Results and Discussion

In table 1, the result for activity prediction using the original samples is presented. When the activity predictor is trained using the male training samples, the accuracy performance is better for the male test set (0.739) compared to the female test set (0.603). Similarly, when the female training samples are used, the accuracy performance is better for the female test set (0.721) compared to the male test set (0.634). This confirms that the distribution of features across the male and female groups is different. So, if we were to use imbalanced datasets for training, we will get biased results.

Male_train		Female_train	
Male_test	Female_test	Female_test	Male_test
0.793 ± 0.066	0.603 ± 0.146	0.721 ± 0.161	0.634 ± 0.111

Table 1: Accuracy score obtained for activity prediction task using the original data samples.

When we use adversarial training, the difference in the accuracy performance between the genders is significantly reduced. This can be seen in table 2. We have used $\alpha = 0.1$ in equation (6) for this experiment. This shows that the distribution of features across the two genders is closer in the encoded space. However, one thing to note is that the accuracy performance is also reduced. The bias in the activity prediction is reduced between the two genders at the cost of performance.

Male_train		Female_train	
Male_test	Female_test	Female_test	Male_test
0.527 ± 0.238	0.489 ± 0.212	0.645 ± 0.199	0.606 ± 0.150

Table 2: Accuracy score obtained for activity prediction task with adversarial training (α=0.1).

In table 3, we show accuracy results when the male training samples are used to train the activity predictor with adversarial loss using different values of α. The performance on the male and female test sets without the adversarial training is used as a baseline. We can see that when a very small value of α is used for the adversarial training, the test performance for the male and female test sets is very similar (0.297 and 0.294). However, the activity prediction performance is very poor and almost the same as random prediction. Increasing the value of α to 0.1, we get similar test performance for the two genders at a higher accuracy value (0.527 and 0.489). At higher values of α, the adversarial training has no impact in reducing the test performance bias. This result can be seen visually in figure 3.

	Without adversarial training		With adversarial training	
α	Male_test	Female_test	Male_test	Female_test
0.0001	0.768 ± 0.0856	0.624 ± 0.134	0.297 ± 0.140	0.294 ± 0.138
0.1	0.793 ± 0.066	0.603 ± 0.146	0.527 ± 0.238	0.489 ± 0.212
0.4	0.781 ± 0.087	0.575 ± 0.124	0.768 ± 0.088	0.595 ± 0.136
0.7	0.779 ± 0.122	0.592 ± 0.122	0.766 ± 0.099	0.580 ± 0.146
1.0	0.770 ± 0.109	0.574 ± 0.102	0.780 ± 0.105	0.610 ± 0.129

Table 3: Accuracy score obtained for different values of α. Male training data is used to train the activity predictor. The test performances on the male and female test sets are reported.

The results verify the hypothesis expressed in equation (3). It is possible to minimize the information about the group using the encoded representation at the expense of reducing some information related to the task.

8 Conclusion

In this project, we were able to understand the bias in the test performance from having some population groups underrepresented in the training dataset. We implemented adversarial training as a measure to address this problem. We were able to see that while adversarial training does help in reducing the gap in the test performance across the groups but the performance for all the groups is also reduced. In the future, I would like to explore if this bias can be minimized while at the same time not compromising the performance. In this project, the model architecture is fixed and the hyperparameters search is not performed. I would like to see if the performance can be improved with a better choice of hyperparameters.

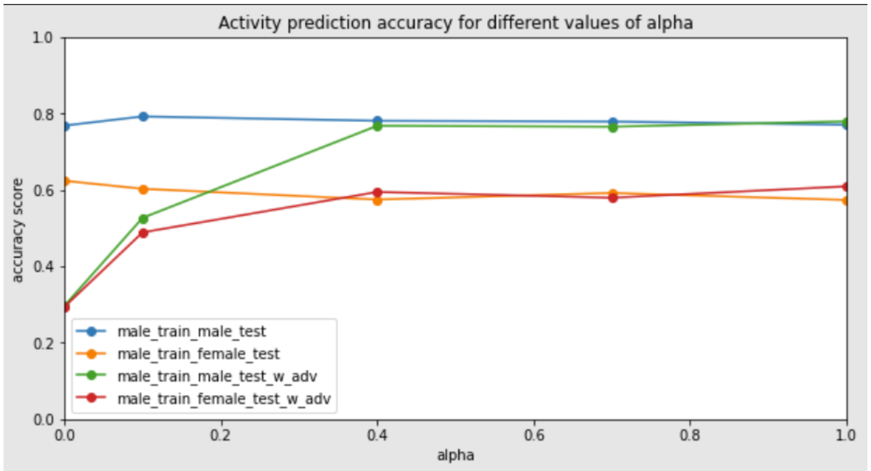

Figure 3: alpha vs accuracy score

References

[1] Richard J. Chen et al. *Algorithm Fairness in AI for Medicine and Healthcare*. 2021. DOI: `10.48550/ARXIV.2110.00603`. URL: `https://arxiv.org/abs/2110.00603`.

[2] Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. “Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model”. In: *Scientific Data* 9.1 (Jan. 2022), p. 24. ISSN: 2052-4463. DOI: `10.1038/s41597-021-01110-7`. URL: `https://doi.org/10.1038/s41597-021-01110-7`.

[3] Sungho Park et al. “Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.3 (May 2021), pp. 2403–2411. DOI: `10.1609/aaai.v35i3.16341`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/16341`.

[4] Blake Lemoine, Brian Zhang, and M. Mitchell, eds. *Mitigating Unwanted Biases with Adversarial Learning*. 2018. URL: `http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_162.pdf`.

[5] Mohammad Malekzadeh et al. “Mobile Sensor Data Anonymization”. In: *Proceedings of the International Conference on Internet of Things Design and Implementation*. IoTDI ’19. Montreal, Quebec, Canada: ACM, 2019, pp. 49–58. ISBN: 978-1-4503-6283-2. DOI: `10.1145/3302505.3310068`. URL: `http://doi.acm.org/10.1145/3302505.3310068`.

[6] Ahmed Sharshar et al. “Activity With Gender Recognition Using Accelerometer and Gyroscope”. In: *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. 2021, pp. 1–7. DOI: `10.1109/IMCOM51814.2021.9377388`.