Adversarial Stacked Auto-Encoders for Fair Representation Learning

Anonymous Authors[1]

Abstract

Training machine learning models with the ultimate goal of maximizing only the accuracy could results in learning biases from data, making the learned model discriminatory towards certain groups. One approach to mitigate this problem is to find a representation which is more likely to yield fair outcomes using fair representation learning. In this paper, we propose a new fair representation leaning approach that leverages different level of representation of data to tighten the fairness bounds of the learned representation. Our results show that stacking different auto encoders and enforcing fairness at different latent spaces result in an improvement of fairness compared to other existing approaches.

1. Introduction

Representation learning has made a significant mark in the field of Machine Learning (ML) over the past decade, with the emergence of technologies that extract useful information or features from data to improve the classification or predictive performance of models, or even generate new synthetic realistic data. Several applications for different kind of tasks have emerged such as, machine translations ((Baltrušaitis et al., 2018)), anomalies detection ((Rivera et al., 2020)), objects and actions recognition ((Papageorgiou & Poggio, 2000)) etc.

ML models are widely used in real life to make decisions that can affect people's lives, e.g., loan applicant, college admission, criminal justice, hiring, etc. Models trained with biased data can lead to unfair decisions (Mehrabi et al., 2019). In fact, these models mainly rely on human-generated data to learn patterns that are then used to make predictions on the new unseen data. However, real-world data are already tainted by prejudices and unfair decisions (historical bias), which reflect the flaws of our society. Historical bias is one origin of algorithmic bias. Another source of algorithmic bias is the representation bias (Mehrabi et al., 2019). It arises when certain groups of the population are underrepresented within the data. For example, a facial recognition model trained with data containing considerably more white faces than black faces will tend to be less accurate when used on black faces. To this end, the algorithmic bias occurs when biases in the data are learned by the model and therefore lead to unfair decisions (Dwork et al., 2012; Kenfack et al., 2021; Hardt et al., 2016).

One approach to mitigate the impact of biases from the data is *fair representation learning*. With this technique, the input data is mapped into a new representation, which is enforced to satisfy a given fairness metric while maintaining the utility of the representation as much as possible. The learned representation can then be used for any downstream task such as classification or data generation, with better chances of yielding fair results. Existing works by Madras et al.; Edwards & Storkey used adversarial learning to enforce the fairness of the representation with respect to statistical parity. They use an auto encoder as a generator, whose aim is to learn a latent space such that an adversary cannot predict the sensitive feature (gender, race etc.) from the learned latent representation. Madras et al. proposed a learning objective for other fairness metrics such as equalized odds and equal opportunity (section 3) with theoretical bounds of fairness.

This work builds on top of the previous works, a fair representation learning approach based on adversarial stacked auto encoders, but leverages different level of representation of the input data to tighten the fairness bounds of the learned representation. In fact, the success of deep networks can be attributed to their ability to exploit the unknown structure in the input distribution to discover useful features at multiple levels. In this multi-level representations, the higher-level learned features are defined in terms of lower-level features (Bengio et al., 2013). For instance, Khan & Fraz showed that performing data augmentation in the feature space and at different level of representation, can improve predictive performances of the neural network. Similarly, a generative model proposed by Huang et al. leveraged different level of representation to improve the quality of generated images. Applying fairness at a given level does not guarantee that information about the sensitive attribute is removed, as it

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

may not all be presented at the given level.

In essence, we hypothesize that the above arguments may also be useful for improving fairness, which was confirmed by our empirical results. Intuitively, the main idea is to approach an optimal adversary via sequential learning, in which one adversary is used to enforce fairness on a high-level representation, which is then used as input for a lower-level representation on which another adversary will be trained to enforce fairness on that representation by improving the previous adversary.

The reminder of the paper is organized as follow. In Sections 2 and 3, we present background and related work respectively. In Section 4, we introduce our fair representation learning approach that tighten the fairness bounds. In Section 5, we present empirical results which show the effectiveness of our learned representation on several real-life datasets. In Section 6, we conclude the paper.

2. Background

2.1. Fairness

Consider a training data $\mathcal{D} = \{X, Y, S\}$, where $x_i \in \mathbb{R}^n$ is the feature vector, $y_i \in \{0, 1\}$ is the label, and S is the binary protected attribute (e.g., gender, race, etc.). Learning a fair representation means mapping the input data X into a new representation X' such that X' will satisfy one of the following fairness criteria:

- *Statistical parity*: It is also known as Demography parity (Δ_{DP}). This fairness criteria promotes the independence between the predictor outcome ($\hat{Y}$ a function of X') and the sensitive attribute. $\hat{Y} \perp S$, i.e., a predictor satisfies statistical parity if $P(\hat{Y}|S = 0) = P(\hat{Y}|S = 1)$ (Dwork et al., 2012). However when the sensitive attribute correlates with the target variable, a drop in accuracy can be observed.

- *Equalized Odds*: In contrast to Δ_{DP}, Equalized Odds (EO) promotes the conditional independence between the prediction outcome and the sensitive attribute given the class label ($\hat{Y} \perp S|Y$). A predictor outcome $\hat{Y}$ trained with X' satisfies EO if $P(\hat{Y} = y|S = 0, Y = y) = P(\hat{Y} = y|S = 1, Y = y), \forall y \in \{0, 1\}$. In other words the False Positive Rate (FPR) and the True Positive Rate (TPR) of groups should be the same. One advantage of equalized odds is that it admits the perfect model $\hat{Y} = Y$ (Hardt et al., 2016; Verma & Rubin, 2018).

- *Equal opportunity*: Similarly to EO, Equal opportunity (EOpp) only considers the case where $Y = 1$ ($\hat{Y} \perp S|Y = 1$). A predictor outcome $\hat{Y}$ satisfies EOpp if $P(\hat{Y} = 1|S = 0, Y = 1) = P(\hat{Y} = 1|S = 1, Y = 1)$. In other words, groups should have the same TPR.

It has been shown that predictor trained with fairness constrained are less accurate than the ones trained without it (Kamishima et al., 2011). Thus fairness comes at the expense of accuracy. A desired property is to provide fair representation with lower fairness accuracy trade-off.

2.2. Adversarial Learning

Inspired by the game theory, adversarial learning consist in two neural networks (generator and discriminator) trained in adversarial manner. The generator's (G) goal is to fool the discriminator by sampling as most realistic examples as possible such the discriminator (D), which the goal is to distinguish between generated examples and real examples, will not be able to make to difference between examples $G(z)$ sampled from G using the random noise vector z and real examples x. Thus, G and D play a minimax game with value function $V(G, D)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z}[1 - \log D(G(z))] \quad (1)$$

where D seek to maximize this quantity while G seek to minimize it.

3. Related Work

Pre-processing techniques are used to mitigate biases from the data by enforcing a given fairness property while maintaining the utility of predictions. The objective of fair representation learning is to learn a representation of the data that is most likely to produce fair results for downstream tasks. Work by (Zemel et al., 2013) is the first fair representation learning approach, which removes dependencies on the sensitive attribute by mapping input data to new points called prototypes. Prior work in this direction focuses on statistical parity, equalized odds and equal opportunities.

The goal is to learn a representation that will remove all the dependencies in regards to the sensitive attribute from the training data, while retaining as much information as possible. In (Louizos et al., 2016), the authors proposed the Variational Fair auto encoder (VFAE), a variant of variational auto encoder that maps the input data into a latent space while discarding as much information about the sensitive attributes from the data as possible. Thus the sensitive attributes are treated as nuisance variable. To do this, the authors (i) used a factorized prior $p(z)p(s)$ where z is the latent representation and s is the sensitive attribute, (ii) and added a regularization term to encourage the independence between z and s using Maximum Mean Discrepancy.

In (Edwards & Storkey, 2015), the authors proposed an approach to learn fair representation using adversarial learning that achieves demographic parity. Beutel et al. explored the particular fairness levels achieved by the algorithm from (Edwards & Storkey, 2015) and showed how other fairness metrics can be achieve by varying the distribution of the adversary's input. Madras et al. extended the previous work by proposing adversarial objectives that yield fair and transferable representations that in turn admit fair classification outcomes. They provided adversarial objective functions for each fairness metric that upper bounds the unfairness of arbitrary downstream classifiers in the limit of adversarial training.

In this work, we propose a new fair representation learning approach built upon previews works and it aims to improve the fairness of models via stacked adversarial learning. We enforce fairness at different level of representation in order to tighten the fairness bounds of the final representation.

4. Methods

In this section, we describe our model's architecture and the training procedure we propose. Figure 1 presents an overview of the architecture and the training process.

4.1. Model Architecture

Our main idea is to stack different Encoders (E_i), Decoders (D_i), classifier f_i and adversary (h_i), in order to get different levels of representation of the input data. The intuition here is that, different level of representation can exhibit different details of information from the data. Enforcing fairness at a given level doesn't guarantee that fairness bounds are tight enough, unless the adversary is an optimal, which which may not be available in non-convex settings. Our goal is to approach this optimal adversary in an incremental may.

At a each level i, we have different components: the learned representation z_i yielded by the encoder E_i, the corresponding decoded representation z_i' produced by the decoder D_i, the adversary network f_i that enforces the fairness of that representation and the predictor network h_i that enforces the utility of the representation. z_0 represents the input data X, and z_0' the final reconstructed output (X'). The overall loss at each level i is defined as the linear combination of three loss terms: the reconstruction loss ($\mathcal{L}^{rec}_{E_i,D_i}$), the adversary loss ($\mathcal{L}^{adv}_{f_i}$) and the predictor loss ($\mathcal{L}^{adv}_{h_i}$):

$$\mathcal{L}(G_i, D_i, f_i, h_i) = \alpha\mathcal{L}^{rec}_{E_i,D_i} + \beta\mathcal{L}^{Adv}_{f_i} + \gamma\mathcal{L}^{Class}_{h_i} \quad (2)$$

Where α, β and γ are the weights associated with each loss. Thus, $\mathcal{L}^{rec}_{E_i,D_i}$ is the loss of reconstructing the encoded representation z_j by the decoder D_i with use the Root Mean Squared Error (RMSE): $\mathcal{L}^{rec}_{E_i,D_i} = \frac{1}{|X|}||z_i' - E_i(D_i(z_{i-1}))||_2^2$. The adversarial loss is to enforce the representation to satisfy certain fairness constraint. For instance, to satisfy statistical parity, the adversary loss is defined as cross entropy loss:

$$\mathcal{L}^{adv}_{f_i} = \frac{1}{|X|}\sum_{s,\hat{s}\in S,\hat{S}} s\cdot\log(\hat{s} + (1-s)\cdot\log(1-\hat{s})) \quad (3)$$

The adversary network at the level i tries to minimize the loss of predicting the sensitive attribute S from the encoded representation z_i, while the generator (typically auto encoder) tries to maximize it. The losses of predictor and adversary can be defined as cross entropy loss or using loss functions proposed in (Madras et al., 2018) to satisfy equalized odds and equal opportunities. Thus at each level we have the following minimax problem:

$$\min_{G_i,D_i,h_i}\max_{f_i}\mathcal{L}(G_i, D_i, f_i, h_i) \quad (4)$$

To have a different representation at each level, we vary the dimension of each latent space, from higher to the lower dimensions ($|z_i| > |z_{i+1}|$).

4.2. Model training

At given level i, we realize the classifier, auto encoders and adversary as neural networks and alternate gradient decent and ascent steps to optimize their parameters according to 4. First the encoder-classifier-decoder takes a gradient step to minimize L while the adversary f_i is fixed, then f_i takes a step to maximize L with fixed auto encoder and classifier. We use a relaxation of adversary objectives proposed by (Madras et al., 2018) i.e to achieve Equalized Odds, in addition to the latent space z, we passed the class label Y to the adversary. To achieve Equalized odds the loss function (Eq 4) is computed only using samples where $Y = 0$.

The training is performed sequentially, starting with an initial latent representation z_1 trained using the input data. During the first training, the adversary f_1 enforces fairness (typically Δ_{DP}, Δ_{EO} or Δ_{EOpp}) of the lower level representation z_1. Afterwards a new latent space of lower dimension z_2 (higher level representation) is stacked, and uses the pre-trained representation z_1 as input.

The number of stacked layers on which the fairness constraints are imposed depends on the depth of the neural network and are specified as a hyper parameter. In the experiments, we used a Multi Layer Perceptron (MLP) network for the encoder and decoder, with one hidden layer. Initially, fairness is applied on the hidden layer (z_1), then the output layer (latent space) is stacked and used as the final representation (z_2). In the testing phase, we get rid of all decoders, adversaries, and classifiers. Only the encoders are used to map the input data into the fair space.

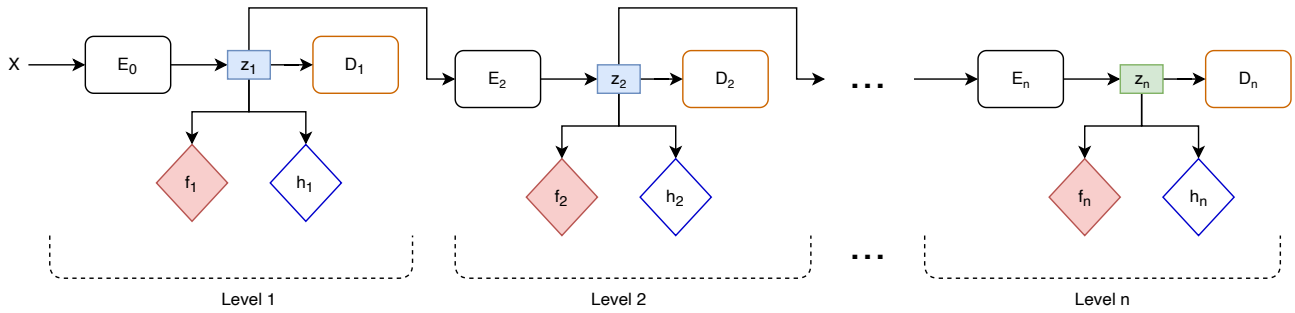


Figure 1. Adversarial Stacked Auto-Encoder architecture

5. Experiments

We present experiments on two standard real world datasets widely used for fair classification as suitable benchmarks to compare the performance of different machine learning methods:

The Adult Income dataset (Asuncion & Newman, 2007) has 48843 instances of demographic information of American adults, described with 14 features. The target variable indicates whether individual's income is larger than 50K US dollars.

The German credit dataset (Jeff et al.) has 1000 instances of bank account information represented by 20 features with the aim to classify bank account holders into credit class good or bad. For both datasets, we use gender as the single protected attribute. We demonstrate the effectiveness of our approach compared to standard fair representation learning techniques.

5.1. Fair classification

Figure 2 shows the fairness results of the MLP trained with the representation obtained by our approach compared to the representation produced by the vanilla approach (LAFR) and MLP trained with original input data (MLP-unfair). For the vanilla approach, we used a network architecture with one hidden layer of 20 units, and latent space of 8 units for Adult dataset, 15 hidden units and 8 output units for the German dataset. We trained the same architecture using our approach with two level of representations. ie. we trained an adversary on the hidden layer and then stacked the output layer and trained another adversary on it. We used single-hidden-layer neural networks for each of our classifier and adversary with 20 hidden units. The equation 4 is optimized using Adam optimizer (Kingma & Ba, 2014) with learning rate of 0.01, a batch size of 64, trained for 150 epochs for Adult dataset and 1000 for the German credit. We run the experiment seven times with different values of β $(1, 2, 3, 5, 15)$, with $\alpha = 0$ and $\gamma = 1$.

Similar to the process used by Madras et al., we created a feed-forward model which consisted of our frozen, adversarially-learned encoders followed by an MLP with one hidden layer, with a loss function of cross entropy with no fairness modifications. We reported the mean over all runs per β and we use a validation procedure to evaluate. The results shows that representation produced by our model always lower bound fairness of standard approaches. This shows that our approach provides tighter fairness bounds. However, since the main objective of our work is to better improve fairness, a decrease in accuracy is observed compared to the standard approach, which we attribute to the trade-off between fairness and accuracy.

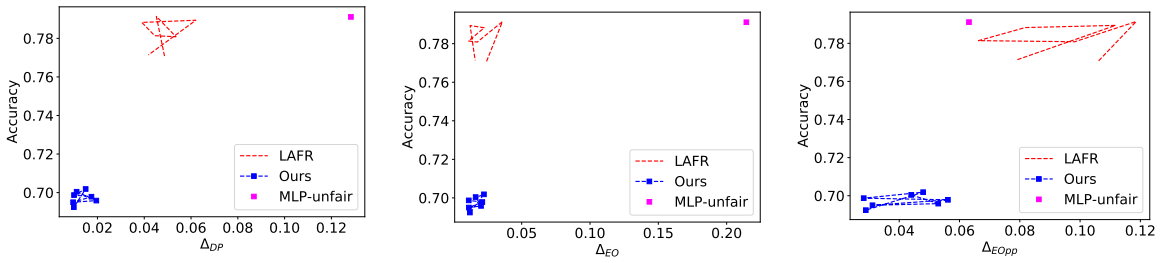
5.2. Classification on downstream tasks

Table 1. Comparison of Δ_{DP} on classification tasks using logistic regression and random forest model on Adult and German datasets

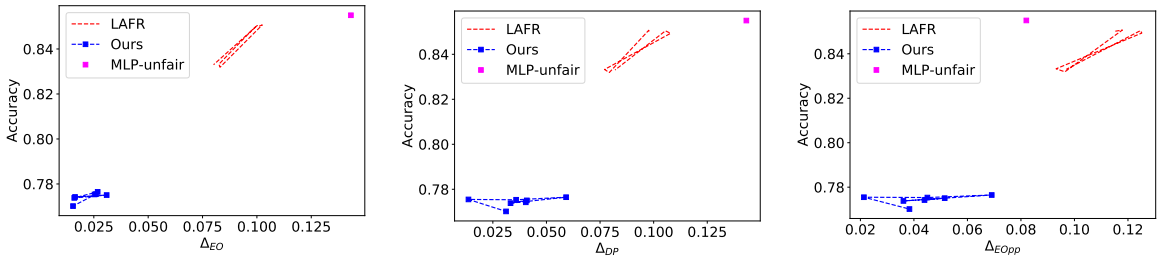
MODEL	UNFAIR	LAFR	OURS
	ADULT		
LOGISTIC REGRESSION	0.53± 0.008	0.51± 0.009	**0.21**± 0.004
RANDOM FOREST	0.54± 0.001	0.49± 0.001	**0.25**± 0.007
	GERMAN		
LOGISTIC REGRESSION	0.36± 0.08	0.31± 0.09	**0.08**± 0.04
RANDOM FOREST	0.27± 0.03	0.23± 0.06	**0.11**± 0.05

Learning fair representation is a model-agnostic approach to mitigating unfairness i.e. the learned representation can be used for any downstream task and not only for neural network based models. We tested linear and non linear models on representation produced by our model and standard approach. We trained the representation using the network architecture described in previous section, without hyper-parameter tuning and using $\alpha = 0$, $\beta = 1$, $\gamma = 1$. We also trained models on the original dataset without fairness constraints.

Table 1 shows Δ_{DP} reported from 5-fold cross validations on Adult and German datasets. Results shows that representation produced by our model also provides better fairness



(a) German Dataset



(b) Adult Dataset

Figure 2. Accuracy and fairness trade-off on fair classification of the German (first row) and Adult (second rows) datasets. Our learned representation always lower bound the fairness results of the representation learned by vanilla approach. Which shows fairness bounds of our approach is more tight. However we can observe a drop in accuracy compared to other representation.

performances when trained using classical machine leaning algorithms such as Linear Regression and Random Forest. We observed similar results for other fairness metrics (EO, EOpp).

6. Conclusion

In this paper, we showed that applying fairness at different levels of representation improves the fairness performance of the learned representation. In this regard, we proposed an adversarial stacked auto encoder architecture which expose different level of representation of the input data, on which several adversary networks are trained sequentially to tighten the fairness bounds of the final representation (lowest level representation).

Our empirical results show that this approach outperform standard adversarial fair representation learning approach in terms of fairness. Intuitively, our learning process lead to learning an optimal adversary in incremental way. However, stabilizing adversarial training of fair representations remains an important issue that we plan to address in future work.

References

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2):423–443, 2018.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Edwards, H. and Storkey, A. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. Stacked generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5077–5086, 2017.

Jeff, L., Surya, M., Lauren, K., and Julia, A. How we analyzed the compas recidivism algorithm.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Kenfack, P. J., Dmitrievich Arapov, D., Hussain, R., Kazmi, S., and Mehmood Khan, A. On the fairness of generative adversarial networks (gans). *arXiv e-prints*, pp. arXiv–2103, 2021.

Khan, A. and Fraz, K. Post-training iterative hierarchical data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. In *International conference on learning representations*, 2016.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Papageorgiou, C. and Poggio, T. A trainable system for object detection. *International journal of computer vision*, 38(1):15–33, 2000.

Rivera, A. R., Khan, A., Bekkouch, I. E. I., and Sheikh, T. S. Anomaly detection based on zero-shot outlier synthesis and hierarchical feature distillation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Verma, S. and Rubin, J. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1–7. IEEE, 2018.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.