Fair Latent Deep Generative Models (FLDGM) for Syntax-agnostic and Fair Synthetic Data Generation

Resmi Ramachandranpillai, Md Fahim Sikder and Fredrik Heintz

Reasoning and Learning Lab, Department of Computer and Information Science (IDA), Linköping University

Abstract

In this work, we propose Fair Latent Deep Generative Models (FLDGM) as enablers for more flexible and stable training of fair DGMs, by first learning a syntax-agnostic, model-agnostic fair latent vector representation of the data. This separates the fairness optimization and data generation processes thereby boosting stability and optimization performance. We conduct extensive experiments on image and tabular domains using Generative Adversarial Networks (GANs) and Diffusion Models (DM) and compare them to the state-of-the-art in terms of fairness and utility. Our proposed FLDGM achieve superior performance in generating high-quality, high-fidelity, and high-diversity fair synthetic data compared to the state-of-the-art fair generative models.

Our Contributions

- We propose a novel formulation of a fair latent generative framework common for both GANs and Diffusion models.
- In contrast to previous works [1, 2] which generate both fair and accurate synthetic data simultaneously, our proposed approach does not require a delicate weighting factor of generation quality and fairness penalty.
- We introduce the concept of syntax-agnostic, model-agnostic fair latent vectors and show how this can be fine-tuned across various generative architectures (different GAN versions and DM) with less computational overhead.

Fair Latent Deep Generative Models (FLDGM)

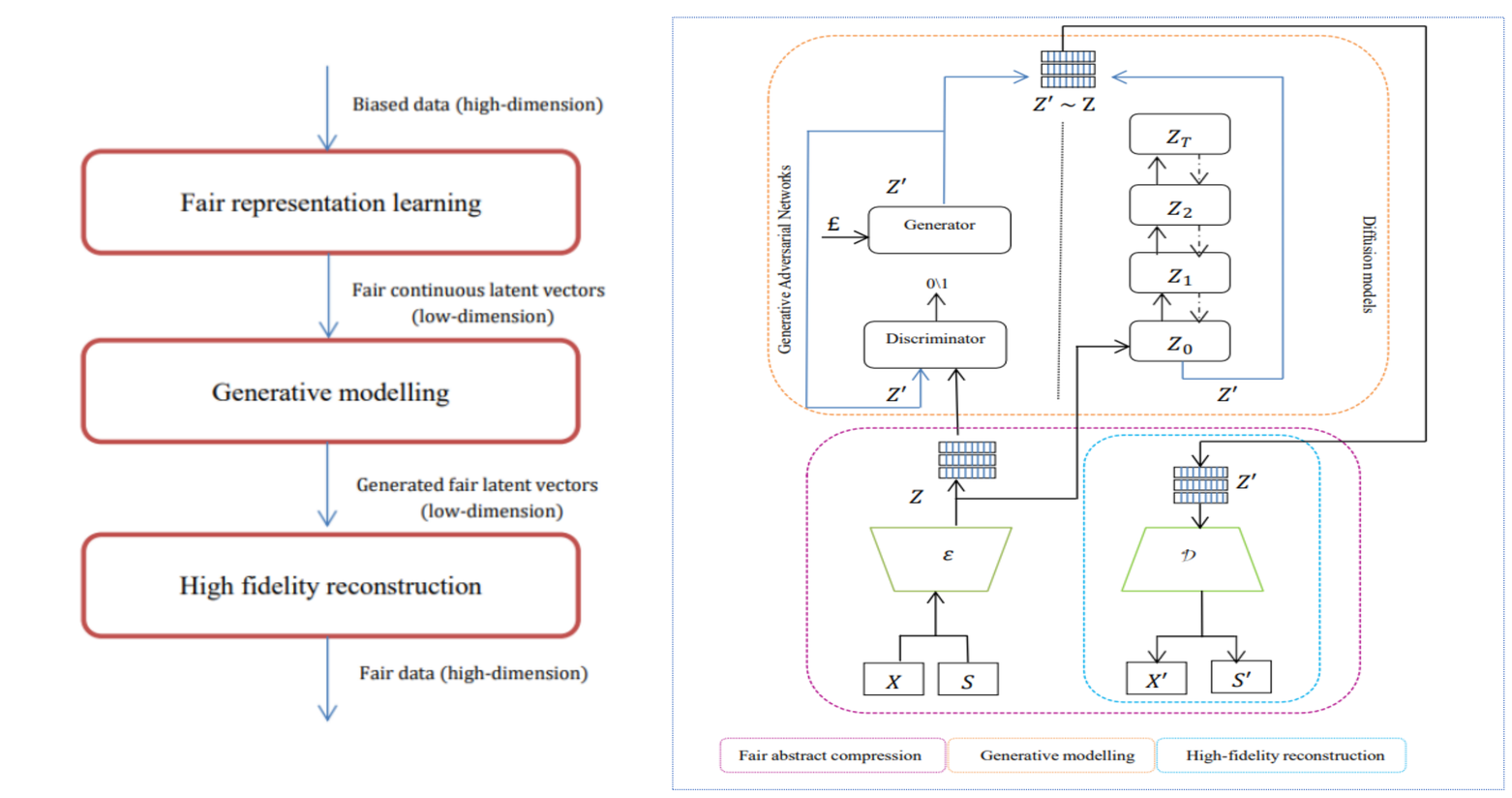

Figure 1 : The FLDGM method including a flow diagram (on the left) and a high level architecture (on the right).

Results

We have the following variants: FLD-WGAN-GP, FLD-LSGAN, and FLD-DM each with Fairness Through Unawareness (FTU) [3] and Demographic Parity (DP). Table 1 shows the results for the Adult income data in terms of data quality and fairness. For Color MNIST, we use visualization results as given in fig 2. PCA & t-SNE plots of latent space of the original and synthtic data is shown in fig 3.

FLDGM

Method	Data Quality			Fairness	
	Precision (↑)	Recall (↑)	AUROC (↑)	FTU (↓)	DP (↓)
Real data	0.920 ± 0.006	0.936 ± 0.008	0.807 ± 0.004	0.116 ± 0.028	0.180 ± 0.010
GAN	0.607 ± 0.080	0.439 ± 0.037	0.567 ± 0.132	0.023 ± 0.010	0.089 ± 0.008
WGAN-GP	0.683 ± 0.015	0.914 ± 0.005	$\mathbf{0.798 \pm 0.009}$	0.120 ± 0.014	0.189 ± 0.024
FairGAN	0.681 ± 0.023	0.814 ± 0.079	0.766 ± 0.029	0.009 ± 0.002	0.097 ± 0.018
DECAF-ND	0.780 ± 0.023	0.920 ± 0.045	0.781 ± 0.007	0.152 ± 0.013	0.198 ± 0.013
DECAF-FTU	0.763 ± 0.033	0.925 ± 0.040	0.765 ± 0.010	0.004 ± 0.004	0.054 ± 0.005
DECAF-CF	0.743 ± 0.022	0.875 ± 0.038	0.769 ± 0.004	0.003 ± 0.006	0.039 ± 0.011
DECAF-DP	0.781 ± 0.018	0.881 ± 0.050	0.672 ± 0.014	0.001 ± 0.002	0.001 ± 0.001
FLD-LSGAN-FTU (ours)	0.762 ± 0.002	$\mathbf{0.998 \pm 0.023}$	0.762 ± 0.012	0.002 ± 0.001	$\mathbf{0.000 \pm 0.001}$
FL-LSGAN-DP (ours)	0.763 ± 0.001	0.941 ± 0.002	0.771 ± 0.010	$\mathbf{0.000 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$
FL-WGAN-GP-FTU (ours)	0.772 ± 0.034	0.918 ± 0.001	0.763 ± 0.023	0.001 ± 0.001	$\mathbf{0.000 \pm 0.001}$
FL-WGAN-GP-DP (ours)	0.782 ± 0.001	0.951 ± 0.001	0.762 ± 0.013	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
FL-DM-FTU (ours)	$\mathbf{0.791 \pm 0.011}$	0.912 ± 0.002	0.795 ± 0.001	$\mathbf{0.000 \pm 0.000}$	0.001 ± 0.000
FL-DM-DP (ours)	0.786 ± 0.002	0.905 ± 0.001	0.787 ± 0.011	$\mathbf{0.000 \pm 0.001}$	$\mathbf{0.000 \pm 0.001}$

Table 1 : Data quality and fairness analysis on Adult Income

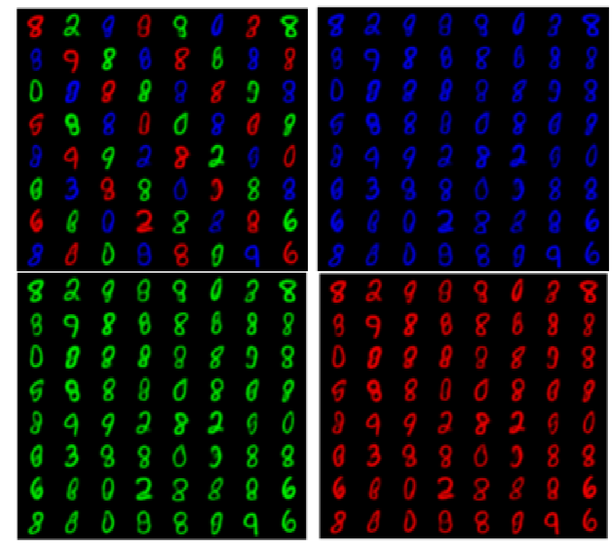

Figure 2: Visualization analysis on Color MNIST

PCA and t-SNE Analysis

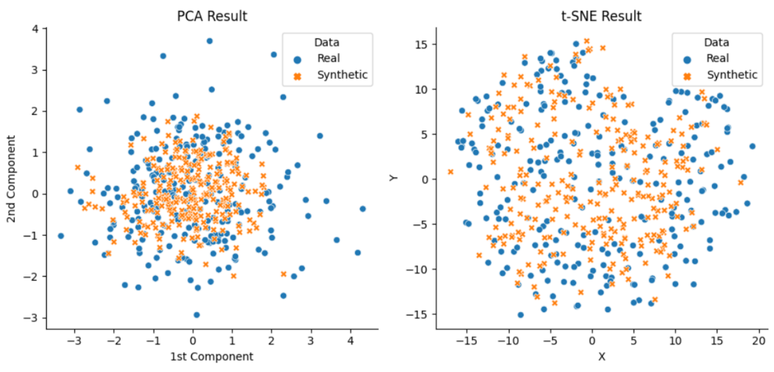

Figure 3 : PCA (left) and t-SNE (right) plot of FLD-WGAN-GP

Conclusion

Fair Latent Deep Generative Models (FLDGM) are proposed for syntax-agnostic, model-agnostic, and fair synthetic data generation using Diffusion models and Generative Adversarial Networks on image and tabular data. Based on our experimental analysis and evaluation, we did demonstrate favourable results in terms of quality, fidelity, diversity, generalization, and fairness compared to state-of-the-art schemes across a wide range of proposed models in the absence of task-specific architectures.

Future Direction

Extend this framework for de-biasing hate speech detection and replace the biased contents in the tweets or speech with another, that could be generated by any underlying Natural Language Generation (NLG) methods.

Acknowledgement

This work was funded by the Knut and Alice Wallenberg Foundation, the ELLIIT Excellence Center at Linköping-Lund for Information Technology, and TAILOR - an EU project with the aim to provide the scientific foundations for Trustworthy AI in Europe. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

[1] Xu, Depeng, et al. "Fairgan: Fairness-aware Generative Adversarial Networks." In the Proceedings of *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

[2] Xu, Depeng, et al. "Fairgan+: Achieving Fair Data Generation and Classification Through Generative Adversarial Nets." In the Proceedings of *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.

[3] Grgic-Hlaca, Nina, et al. "The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making." In the Proceedings of *NIPS symposium on machine learning and the law*. Vol. 1. No. 2. 2016.

Poster

LINKÖPING UNIVERSITY

Department of Computer and Information Science (IDA)
Artificial Intelligence and Integrated Computer Systems (AIICS)