Visualizing chest X-ray dataset biases using GANs

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Recent work demonstrates that images from various chest X-ray datasets contain visual features that are strongly correlated with protected demographic attributes like race and gender. This finding raises issues of fairness, since some of these factors may be used by downstream algorithms for clinical predictions. In this work, we propose a framework, using generative adversarial networks (GANs), to visualize what features are most different between X-rays belonging to two demographic subgroups.

Keywords: Chest X-rays, fairness, bias, explainability, generative adversarial networks (GANs)

1. Introduction

Recent studies have demonstrated that patient bio-information like age, race, and gender are predictable from chest X-ray (CXR) images alone using deep learning models(Gichoya et al., 2022; Karargyris et al., 2019; Duffy et al., 2022). For example, in the "Reading Race" study, deep classifiers trained to predict race achieve 0.99 AUROC on several CXR datasets (Gichoya et al., 2022). This finding raises the question: "What visual cues discriminate different races?" Answering such a question can help mitigate potentially biased behavior of downstream algorithms that make decisions using this data. In this work, we propose a framework to visually explain the principal differences between different demographic subgroups in a medical imaging dataset. We first train an unconditional generative adversarial network (GAN) (Goodfellow et al., 2020) on the given image dataset. Next, we project the images onto the (trained) GAN's latent space and compute a direction in the latent space that differentiates a pair of classes (e.g., "Black" vs. "White" race groups). We traverse the latent space along that direction to produce image sequences that depict the main morphological and appearance changes in moving from one class to another.

There are related works that focus on visualizing subgroup differences associated with clinical attributes. One such study uses autoencoders (Cohen et al., 2021), which often produce blurry samples that do not clearly capture structural information. Others train *conditional* versions of GANs (Singla et al., 2023; Dravid et al., 2022), an expensive process since the GAN must be trained from scratch for each attribute of interest. In contrast to all these works, we demonstrate that deep generative models may be a useful tool to the medical imaging community to understand the biases within a medical imaging dataset.

2. Method

Our method consists of several components, visualized in Fig. 2 and described below.

Generator training: We train an unconditional StyleGAN2 generator (Karras et al., 2020a) $G(\cdot) : R^d \rightarrow R^{H \times W \times 1}$, following the default training procedure introduced in that

© 2023 CC-BY 4.0, Anonymous.

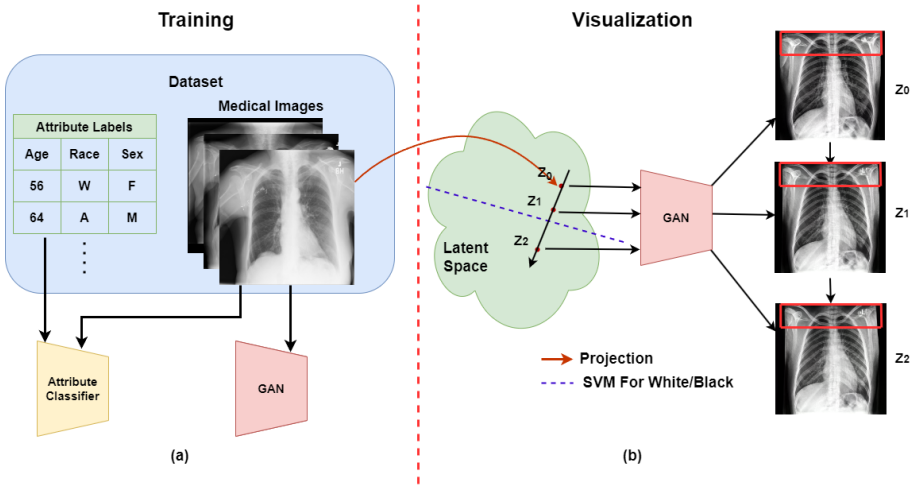

Figure 1: **Framework of our proposed method.** (a) We train a GAN on an image dataset, and a binary classifier on the images and labels for a demographic prediction task (e.g., White vs. Black race). (b) We project a subset of images onto the trained GAN's latent space. To ensure the projected images are reasonably reconstructed, we only keep projected images whose labels (predicted by the attribute classifier trained in (a)) agree with their original labels. We also fit an SVM hyperplane to separate the two classes in the latent space. Finally, we visualize the differences between the classes by starting at a latent code corresponding to a random image, and traversing along the normal direction of the SVM hyperplane, to generate a sequence of images showing a transformation.

paper. d is the dimension of the "latent space" of the generator, and H and W are the height and width of the generated CXR. In our experiments, we trained $G(\cdot)$ on Chexpert (Irvin et al., 2019), a large public dataset containing $224,316$ CXRs. We only used frontal views, yielding $164,548$ CXRs. The training procedure takes roughly 24 hours on two Nvidia A100 GPUs.

Attribute classifier training: We train a separate deep attribute classifier $C(\cdot) : R^{H\times W\times 1} \rightarrow R^1$ for each per-image binary attribute provided in the dataset. For multi-class labels such as race, we train a separate binary classifier for each pair of races.

Image projection/SVM training: Next, we follow the process introduced in (Karras et al., 2020b) to project a subset of CXR images $\{X_i\}_{i=1}^N$ onto G's latent space, yielding latent codes $\{z_i\}_{i=1}^N$. We only retain those projected images whose labels (predicted by C) are the same as the original labels $\{L_i\}_{i=1}^N$, i.e., $C(G(z_i)) = L_i$. We then train a linear SVM to predict L_i from z_i.

Image sequence generation: The normal vector v of the trained SVM's hyperplane identifies the direction that best differentiates the two classes. We will use this fact to generate image sequences depicting the principal perceptual changes needed to convert a CXR belonging to one demographic class to another. In particular, we select the latent vector corresponding to a random dataset CXR, and move towards the opposite class in latent space in the direction of v. We concatenate images generated by intermediate latent codes along this traversal to produce a sequence.

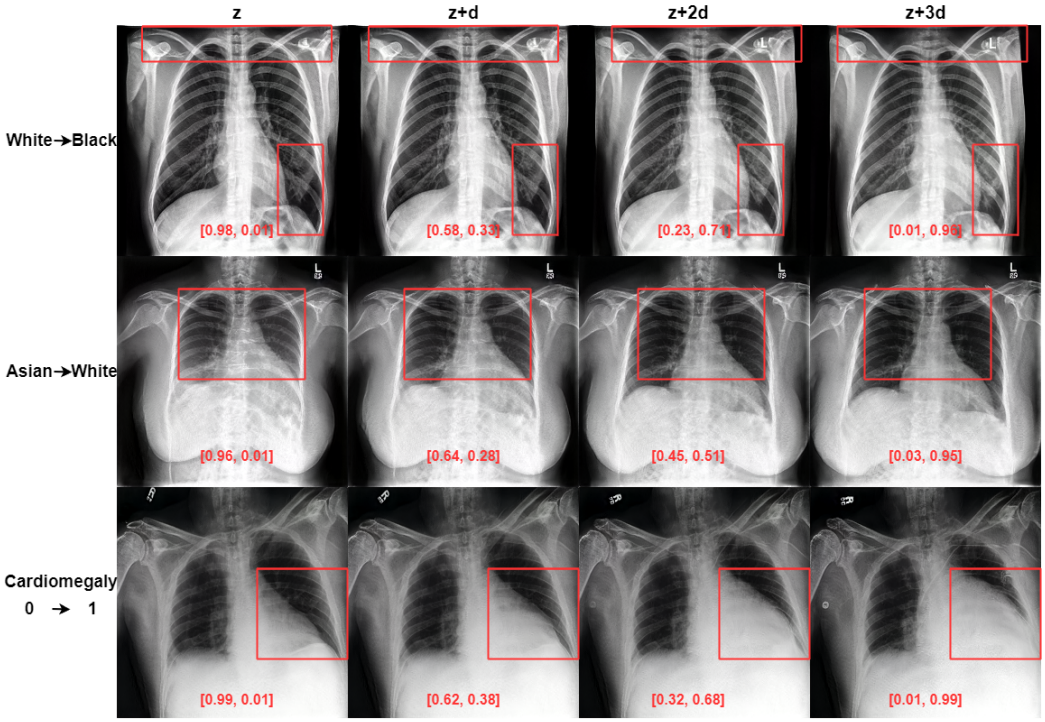

Figure 2: **Sample visualization results.** The left column corresponds to the projected initial image and the last three columns show images generated at different traversal distances in the latent space. The red text indicates the output probabilities predicted by the attribute classifier for each class. For example, the top left [0.98, 0.01] indicate the CXR has a 98% possibility of being white and 1% possibility of being black. We also use red boxes to highlight the areas that visually vary the most. For White/Black, the shoulder bone and right lung structures change shape, and the lungs become more opaque. For Asian/White, the entire chest shape changes and grows larger. These visualizations also explain why the Reading Race study (Gichoya et al., 2022) did not find race prediction to significantly change when blocking local regions. The proposed applied to *Cardiomegaly* enlarges the heart, in agreement with the known effect of that disease.

3. Results and discussion

We demonstrate our framework on ChexPert with *race* as the target attribute. We also validate our approach on the clinical attribute *Cardiomegaly*, which induces a known physiological change (enlarged heart). Sample results are shown and explained in Fig. 2.

Conclusion Our results show that an unconditional generative adversarial network can be a useful tool for visualizing differences between demographic groups of a CXR dataset. Our framework is fast and flexible, and can be applied to any binary attribute labels in the dataset. Future work includes analyzing generated sequences to thoroughly investigate demographic differences, and comparing results across different generative models.

References

Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning*, pages 74–104. PMLR, 2021.

Amil Dravid, Florian Schiffers, Boqing Gong, and Aggelos K Katsaggelos. medxgan: Visual explanations for medical classifiers through a generative latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2936–2945, 2022.

Grant Duffy, Shoa L Clarke, Matthew Christensen, Bryan He, Neal Yuan, Susan Cheng, and David Ouyang. Confounders mediate ai prediction of demographics in medical imaging. *npj Digital Medicine*, 5(1):188, 2022.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alexandros Karargyris, Satyananda Kashyap, Joy T Wu, Arjun Sharma, Mehdi Moradi, and Tanveer Syeda-Mahmood. Age prediction using a large chest x-ray dataset. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, pages 468–476. SPIE, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020b.

Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis*, 84:102721, 2023.