FAIR-FER: A Latent Alignment Approach for Mitigating Bias in Facial Expression Recognition (Student Abstract)

Syed Sameen Ahmad Rizvi*, Aryan Seth*, Pratik Narang

Department of CSIS, Birla Institute of Technology & Science, Pilani, RJ, India
{p20190412,f20212221,pratik.narang}@pilani.bits-pilani.ac.in

Abstract

Facial Expression Recognition (FER) is an extensively explored research problem in the domain of computer vision and artificial intelligence. FER, a supervised learning problem, requires significant training data representative of multiple socio-cultural demographic attributes. However, most of the FER dataset consists of images annotated by humans, which propagates individual and demographic biases. This work attempts to mitigate this bias using representation learning based on latent spaces, thereby increasing a deep learning model's fairness and overall accuracy.

Introduction

Facial Expression Recognition requires human annotations per image, which propagate annotative biases and prejudices. Annotative biases combined with class and demographic imbalances increase bias and reduce equal-odds fairness for attributes such as gender, ethnicity, etc. Therefore, it is crucial to examine the biases within datasets and design algorithms to mitigate them.

Considering age as a protected attribute in datasets, we observe that adolescents are represented positively (such as happy) (Chen and Joo 2021); on the contrary, senior citizens are represented more negatively (such as sad and disgusted). This causes models to be biased, with adolescents being classified more frequently to positive expressions, viz-a-viz, and senior citizens being predicted to negative expressions. This work is our attempt to tackle and mitigate this bias, therefore increasing fairness in a deep learning model. The major contributions of this work include:

- A novel architecture that generates better representations, mitigates bias and improves accuracy for FER.
- A novel training technique that uses a VAE and a CNN backbone for facial expression classification.
- We examine an adversarial approach with a novel loss function to align multiple latent spaces to mitigate bias

To the best of our knowledge, this is the first attempt to explore representation learning using latent spaces in mitigating biases in the facial expression domain.

*These authors contributed equally.

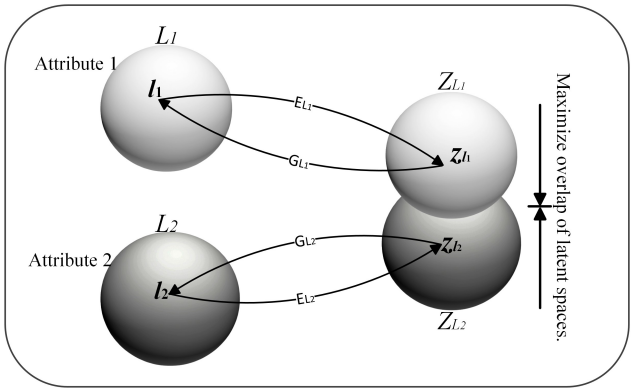

Figure 1: Architecture for Attribute Disentanglement. L_i represents data having the attribute q_i. Z_{L_i} is the latent representation of L_i. E_{L_i} is a VAE with shared weights $\forall$ i

Methodology

We introduce a two-part model to address bias mitigation. Given CNNs' propensity to assimilate all input features, our initial model component employs a Variational Autoencoder (VAE) to encode images belonging to protected attributes into the common latent space. The goal is to minimise disparities between these latent spaces, ensuring they contain information relevant to expression classification.

Attribute Disentanglement - We propose a shared-weight Variational Autoencoder across all protected attributes, mitigating inter-latent domain disparities through an adversarial discriminator. In this context, we denote the Encoder and Generator components as 'E' and 'G'. This is demonstrated in Figure 1, where q_i is a protected attribute such as gender.

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}(x) =& \mathrm{KL}\left(z_x \mid x\right) \| \mathcal{N}(0, I)) + \mathcal{L}_{\mathrm{VAE,D}}^{\mathrm{Latent}}(x) \\ &+ \alpha \left\| G_j^{\phi}(\hat{y}) - G_j^{\phi}(y) \right\|_F^2 \end{aligned} \tag{1}$$

Equation 1 is the objective function for the VAE. The first component consists of KL-divergence that penalizes deviation of the latent distribution from a Gaussian Distribution. The second component is discriminator loss, which measures whether the discriminator can predict the protected attribute class. The final component is Style-Reconstruction

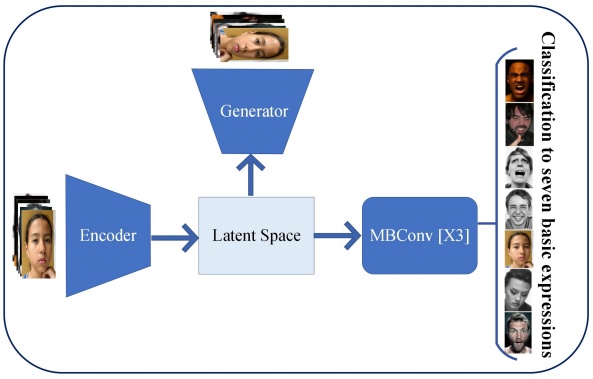

Figure 2: MBConv backbone uses the Encoder generated latent to classify into the 7 emotions.

Loss (Johnson, Alahi, and Fei-Fei 2016).

Classification Model - We feed the latent representation generated by E into a custom classification module using MBConv blocks. This is demonstrated in Figure 2.

$$\min_{E_{\mathcal{X}_i}, G_{\mathcal{X}_i}} \max_{D_{\mathcal{X}_i}} = \mathcal{L}_{\text{VAE}}(x) + \mathcal{L}^{\text{latent}}_{\text{VAE,D}}(x_{q_i}) \quad \forall q \qquad (2)$$

Training Method - The Encoder and the Discriminator are trained jointly with a min-max objective function (2) with a categorical cross-entropy loss for the Discriminator. The classification model is trained after the VAE with a symmetric cross-entropy loss for robustness.

Metrics

We formulate our metric for fairness as (Xu et al. 2020) and use the "equal odds" philosophy.

$$\mathcal{F} = \min\left(\frac{\sum_{c=1}^{C} p\left(\hat{y} = c \mid y = c, q = q_i, \mathbf{x}\right)}{\sum_{c=1}^{C} p(\hat{y} = c \mid y = c, q = d, \mathbf{x})}\right). \qquad (3)$$

$$\forall i \in (1, 2....N)$$

In equation 3, d is the protected attribute that has the highest accuracy. We add the accuracy for each class per attribute and use the minimum value as our metric for fairness.

For completeness, we also use the mean per-class per-attribute accuracy as in (Wang et al. 2020).

Experimentation and Results

Experimentation was conducted on the RAF-DB dataset (Li, Deng, and Du 2017) similar to ((Xu et al. 2020)). The dataset has 7 human-annotated classes. The model is trained on the provided train-test split consisting of 12271 train images, and inference is run on 3068 test images.

Table 1 and Table 2 show that our model achieves state-of-the-art results on RAF-DB for both metrics and demonstrates significant bias mitigation.

Conclusion

Bias in FER datasets is a long-standing problem. In this abstract, we propose a new method for mitigating bias in FER systems by using a Variational Autoencoder with an Adversarial Discriminator followed by an MBConv-based classification module. We surpass the results presented by (Xu et al.) and provide an adaptable framework that can be extended to other image classification tasks. To the best of our knowledge, this is the first work that uses latent alignment for de-biasing in FER systems. We believe our work will lead to further investigations in latent space manipulation for bias mitigation in broader image classification contexts. This research is supported by Kwikpic AI Solutions.

Expression	**Accuracy(%)**	
	Xu et al.	**Ours**
Anger	81.0	83.2
Disgust	54.1	57.7
Fear	53.8	60.2
Happy	93.3	92.0
Neutral	82.1	81.0
Sad	77.7	76.0
Surprise	81.8	82.9
Mean	**74.8**	**76.1**

Table 1: Comparison of expression-wise accuracies.

Protected attributes	**Mitigation of Bias**	
	Xu et al.	Ours
Gender	99.97	99.51
Race	91.6	**94.2**
Age	82.1	**84.8**

Table 2: Comparison of mitigation of bias (higher is better).

References

Chen, Y.; and Joo, J. 2021. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14980–14991.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Li, S.; Deng, W.; and Du, J. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2584–2593. IEEE.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8919–8928.

Xu, T.; White, J.; Kalkan, S.; and Gunes, H. 2020. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 506–523. Springer.