

BIAS MITIGATION USING FUNCTIONAL INEQUALITIES FOR REGULARIZATION

ZIAD EL SAYED

جامعة نيويورك أبوظبي

NYU | ABU DHABI

DEFINING BIAS

- Multimodal Bias:** Defined as the assigning of uneven weights to different modalities by a multimodal machine learning model. For instance, if the output of a VQA model is influenced more by visual information than accompanying text, the model exhibits bias towards the visual modality.
- Demographic-Based Bias:** Refers to the discrimination against individuals by a machine learning model based on characteristics associated with certain demographics. Such bias extends to discrimination based on race, ethnicity, age, gender, religion, sexual orientation, gender identity, expression, disability, or economic status.

DATASETS

- Colored MNIST:** Colored MNIST is a colored version of the MNIST dataset of handwritten digits. A distinct color is associated with each digit, and the standard deviation (i.e. the extent to which the colors vary between digits) is set to 0.10.

0 0 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9 9

Figure 1: Excerpts from Colored MNIST

Complementary to the MNIST dataset is the LeNet model, a CNN developed for character recognition. It features two convolutional and pooling layers, with a final fully connected layer of 10 neurons representing 10 distinct classes, making it ideal for digit classification tasks like MNIST.

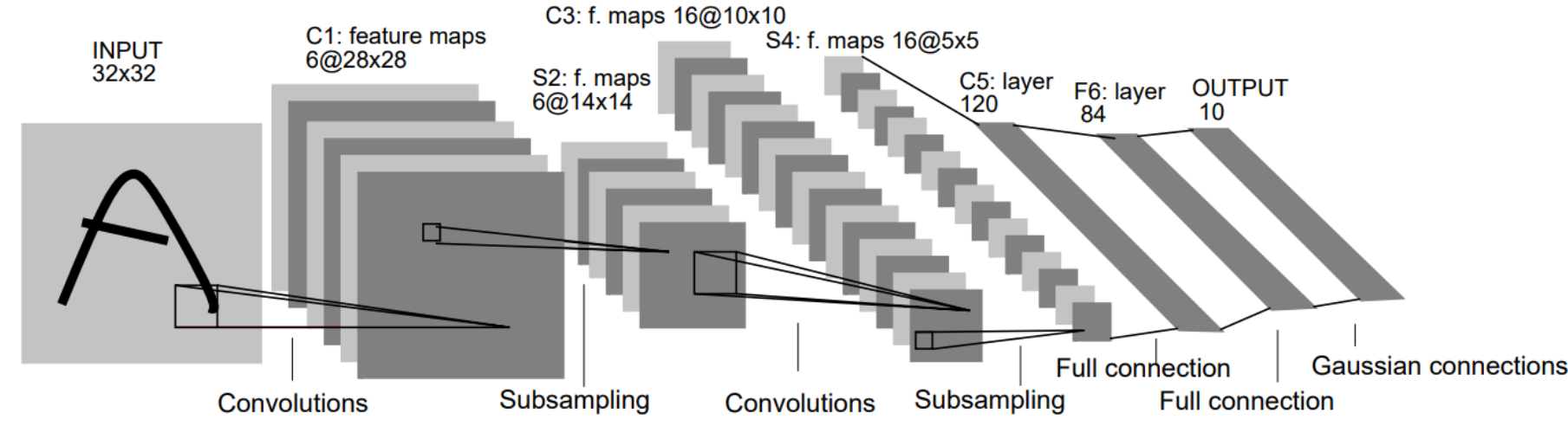


Figure 2: Architecture of LeNet

- MMBias:** The MMBias dataset comprises around 3,800 images and phrases covering 14 population subgroups, including minorities based on religion, nationality, sexual orientation, and disabilities.

Target Class	Target Groups
Religion	Islam, Christianity, Judaism, Buddhism, Hinduism
Nationality	American, Arab, Chinese, Mexican
Disability	Physical disability, Mental disability, No disability
Sexual Orientation	Homosexual, Heterosexual

Table 1: MMBias Target Groups

DATASET EXTENSION

- MMBias Extension:** To achieve wider demographic representation in the MMBias dataset, we invoke the Flickr API to collect pictures for two additional target groups: "African" and "South Asian". We retrieve the 1,000 most relevant images using 10 search keywords and filter out irrelevant/inappropriate images. Amongst these filtered images, 250 images were randomly chosen. This is the same pipeline followed by the original creators of the MMBias dataset, and was done with the help of 3 human annotators to minimize the effect of individual biases.



Figure 3: Excerpts from our MMBias Dataset - South Asian Target Group



Figure 4: Excerpts from our MMBias Dataset - African Target Group

- Choice of Groups:** Neither group had a category of its own in MMBias. We added them as broad regional groups within its nationality class, to follow the dataset's existing structure so they could be assessed alongside its other groups, and to fit the project's timeframe. Each covers many countries, cultures and identities, and the images reflect what Flickr returns for each group's keywords; a fuller extension would use narrower, clearly documented groups.

PROPOSED APPROACH

- k-Nearest Neighbors (kNN) Perturbation:** We propose a new perturbation technique, kNN perturbation, formulated as follows. Let x be a point belonging to the set of training points S . For each modality x_i in the multimodal space $x_1, x_2, \dots, x_n$, we apply a k-nearest neighbors perturbation to all of its points. Given a point p in x_i , and its k closest neighbors $n_1, n_2, \dots, n_k$, we formulate our perturbation p_0 of p as:

$$p_0 = p + \left(p - \frac{\sum_{j=1}^k n_j}{k} \right) \quad (1)$$

- Regularized Objective Function:** In an effort to minimize the training loss while maximizing the tensorized functional entropy, we include a regularization term in the loss function. Previous works incorporating a tensorized learning objective apply the same regularization hyperparameter across all modalities, so that the regularization function is:

$$C.E.L + \lambda \sum_{(x,y) \in S} \sum_{i=1}^n \left(\int_{\mathbb{R}^{d_i}} \frac{\|\nabla_{z_i^x} CE(p_w(\cdot | z_i^x), p_w(\cdot | x))\|^2}{CE(p_w(\cdot | z_i^x), p_w(\cdot | x))} d\mu_i^x(z_i) \right)^{-1} \quad (2)$$

where Cross Entropy Loss (C.E.L) is defined as:

$$C.E.L = \sum_{(x,y) \in S} CE(\mathbb{1}[\cdot = y], p_w(\cdot | x))$$

To address the potential need for differential emphasis on specific modalities, we propose the following adjustment to the tensorized learning objective:

$$C.E.L + \sum_{(x,y) \in S} \sum_{i=1}^n \lambda_i \left(\int_{\mathbb{R}^{d_i}} \frac{\|\nabla_{z_i^x} CE(p_w(\cdot | z_i^x), p_w(\cdot | x))\|^2}{CE(p_w(\cdot | z_i^x), p_w(\cdot | x))} d\mu_i^x(z_i) \right)^{-1} \quad (3)$$

IMPLEMENTATION

- Modified LeNet Model:** Following previous methods, we divide each data sample x in Colored MNIST into two modalities: a color modality x_c and a shape modality x_s . We employ LeNet to encode the shape modality. After this, we append the color representation to LeNet's final linear layer, and then it is processed through an additional linear layer to perform classification. Importantly, we implement and apply our modified regularization function and our kNN perturbation.
- OpenAI CLIP Model:** We leverage OpenAI's CLIP model to assess and quantify the bias in our extended MMBias dataset. We do this in two different ways.
 - Cross-Modal Zero-Shot Classification:** We take a list of phrases (pleasant & unpleasant) and classify the images into different classes. Here, each phrase represents a class.
 - Cross-Modal Bias Assessment:** We measure the effect size distance between image target concepts and textual pleasantness attributes.

TESTING

- Generalization on Grayscale MNIST:**

Model	Convergence Accuracy (%)
Baseline (Unimodal)	47.32 ± 1.12
Previous SOTA	96.24 ± 0.74
Our Method	97.64 ± 0.71

- CLIP Zero-Shot Classification:**

impo	terror	greed	Nationality - African	bad	illiterate	classy	graceful	fabulous
0.15	0.12	0.09	0.06	0.05	0.08	0.06	0.06	0.06
undoc	safe	honor	Nationality - American	talented	graceful	joy	fabulous	respectful
0.05	0.10	0.08	0.06	0.06	0.06	0.06	0.06	0.05
greed	terror	impo	Nationality - Arab	illiterate	fanaticism	dictator	vagrant	terrorism
0.27	0.11	0.09	0.07	0.06	0.07	0.05	0.05	0.04
greed	impo	dictator	Nationality - Chinese	undoc	angel	graceful	talented	classy
0.10	0.07	0.07	0.05	0.05	0.07	0.06	0.05	0.05
greed	undoc	devil	Nationality - Mexican	impo	sinister	dictator	favorable	fabulous
0.28	0.15	0.08	0.06	0.05	0.05	0.05	0.05	0.05
terror	gangster	impo	Nationality - South-Asian	greed	fragile	peaceful	tranquil	graceful
0.10	0.10	0.10	0.07	0.07	0.07	0.08	0.07	0.07

Figure 5: Unpleasant Phrases (Red), Pleasant Phrases (Blue).

- CLIP Cross-Modal Bias Assessment:**

Nationality X	Nationality Y	CLIP Cross Bias
Arab	American	1.28
Mexican	Arab	-0.29
Chinese	American	0.91
Mexican	American	1.14
African	American	0.55
African	Arab	-0.82
African	Chinese	-0.35
African	Mexican	-0.61
African	South-Asian	0.37
American	South-Asian	-0.08
Arab	Chinese	0.54
Arab	South-Asian	1.02
Chinese	Mexican	-0.29
Chinese	South-Asian	0.66
Mexican	South-Asian	0.87

Figure 6: Results < 0 ⇒ Model Favors Group X & Vice Versa

FUTURE RESEARCH

- We recommend future research to further extend these methods to tackle demographic-based biases.
- We furthermore implore researchers to explore additional datasets and perturbation techniques.
- We believe it would be a promising approach to model the dependencies between different modalities during the training process.
- There is also a current assumption that μ_i^x follows a Gaussian distribution. This may not always be the case, and is in fact not a necessary condition to apply the log-Sobolev inequality, as it is not exclusive to Gaussian measures and has been defined for different measures. A potential modification, therefore, would be to extend this technique to apply to measures with other distributions.

CONCLUSION

- We adopt an approach to reducing bias in multimodal machine learning models through functional inequalities.
- We examine multimodal and demographic-based biases, where models unevenly weigh modalities or discriminate by demographics.
- We review, critique, and implement existing methods with our proposed modifications for enhanced effectiveness and representation.
- Our modifications include kNN perturbation and the division of regularization terms among multiple hyperparameters.
- Results indicate a 1.40 percentage-point improvement in generalization to grayscale MNIST using our method as opposed to the previous SOTA, ultimately achieving 97.64% mean accuracy.