De-Biasing Visual Datasets with Variational Auto Encoders.

Introduction

In this project, we will address algorithmic bias. We will build a facial detection model that learns the latent variables underlying face image datasets and uses this to re-sample the training data and hopefully helps to mitigate any biases that may be present to train a debiased model.

Definition:
Algorithmic bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others. (Goodman and Flaxman, 2017)

GitHub/Google Collab: https://git.arts.ac.uk/21035961/Term-2-AI-for-Media-Mini-Project - Student ID - 21035961

1. Aims of the Project

The aims are to build a model that trains on CelebA dataset and achieve high classification accuracy on the tested dataset across all demographics, helping us to conclude that the model does not suffer from any hidden bias.

However, how does bias occur within a classifier?

The bias of a classifier is a direct consequence of a bias in the training dataset, frequently caused by the co-occurrence of relevant features and irrelevant ones.
(*9 Types of Data Bias in Machine Learning - TAUS*, no date; Reimers *et al.*, 2021)

2. Methods used for implementation

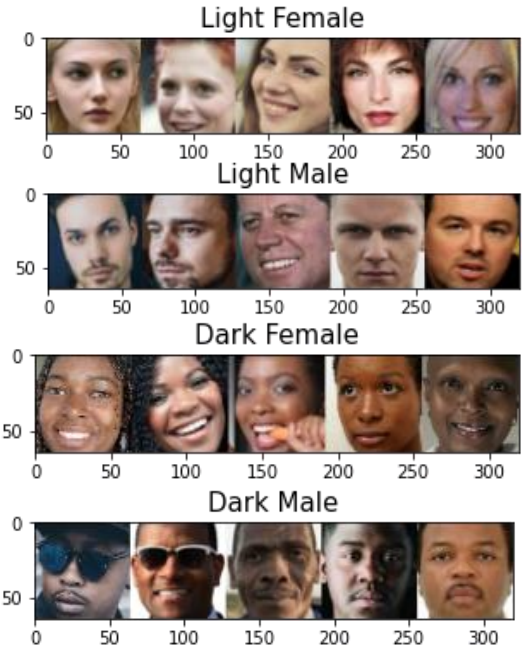
We will train the model on three datasets. In order to train our facial detection models, we will need a dataset of positive examples (i.e., of faces) and a dataset of negative examples (i.e., of things that are not faces).

- CelebA Dataset.
- Imagenet will be our negative training set, based on the non-human images.
- We will use the Fitzpatrick Scale skin type classification system. Where the images have been labelled 'lighter', 'Darker.'
- **'(*CelebA Dataset*, no date; "The Fitzpatrick Skin Type Classification Scale," no date; Russakovsky *et al.*, 2015)**

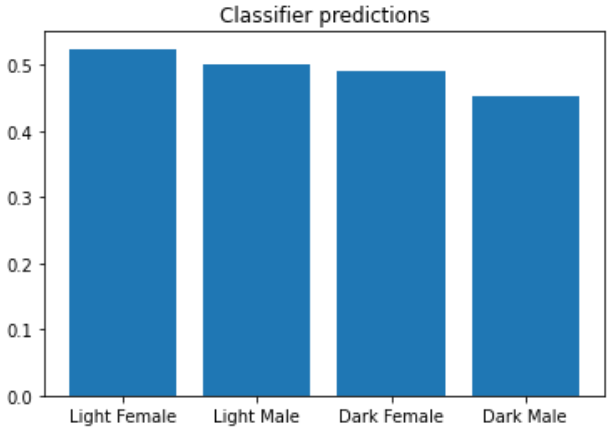
The model chosen for this project is a Convolutional Neural Network model ("CNN"). The reason for choosing CNN to train the classification model is because the structure of the model is similar to the classical LeNet-5 model, but they are different on some parameters, such as input data, network width and entire connection layer. (Lu, Song and Xu, 2020)

3. Results

Based on the standard CNN we have built, and with the help of the Fitzpatrick scale, below are the results of facial recognition and categorisation:



The above is a look into the dataset the model used, and below are the classifier's predictions on faces.



Even if there is minimal variation in the model's predictive accuracy, this can still affect an individual if utilised in real-world applications.

So, what steps can be taken to mitigate/de-bias this?

We need to look at the CelebA dataset (not all 50,000 pictures). However, we can have a look and see the above inaccuracies could be based on the fact that the dataset has just more female faces than males. As a result, a classifier trained on CelebA will be better suited to recognise and classify faces with features similar to these and thus be biased. (This is the most commonplace occurrence when working with pre-existing datasets)

Big tech companies have adopted the approach of labelling different subclasses, e.g. males with hats vs males without within their training data and then manually even it out.

As students, it is impossible for us to do independently, and would require weeks of annotating massive amounts of data. Secondly, we would have to know and look out for all the potential bias within the dataset without missing a single image, and as we know, this method is prone to many errors with manual annotation.

If the manual annotation solution is not possible due to time constraints, there are other areas of debiasing exploration:

Variational autoencoder ("VAE") for learning latent structure

Our aim will be to try and train a debiased version of the same classifier as earlier, one that could account for potential disparities within the given dataset.

However, to be more specific about our aim, we want to build a debiased facial classifier.

Train a model that learns a representation of the underlying latent space to the face training data.
The model can then use that information to mitigate unwanted bias by frequently sampling faces with rare features during the training.

Ensure that the model meets the requirement of learning how to encode latent features in the face data entirely unsupervised. (To achieve step three, we will enlist the help of VAEs)

Reference: "An autoencoder is a simple generative model and a type of artificial neural network used to learn efficient data coding unsupervised and thus can be used for generative modelling and debiasing tasks."
(*Debiasing a facial detection system using VAEs: detailed overview | by siddhant kandge | Medium*, no date)

To use Variational autoencoders, we need the following:

Loss Function - The loss function of the variational autoencoder is the negative log-likelihood with a regularise.

Reparameterization - The "trick" part of the reparameterisation trick is that you make the randomness an input to your model instead of something that happens "inside" it, which means you never need to differentiate to sampling (which you cannot do)."

Cite:(*mathematical statistics - How does the reparameterisation trick for VAEs work and why is it important? - Cross Validated*, no date)

Let us define a function to implement the VAE sampling operation:
Debiasing variational autoencoder (DB-VAE)

So now that we have seen the architecture and more or less understood what is happening within the model. Let us try to train and DB-VAE model on the task of facial detection and run the debiasing operation during the training. We can then evaluate and compare its accuracy to the original model we created.

(*Debiasing Variational Autoencoder. The architecture of the semi-supervised... | Download Scientific Diagram*, no date)

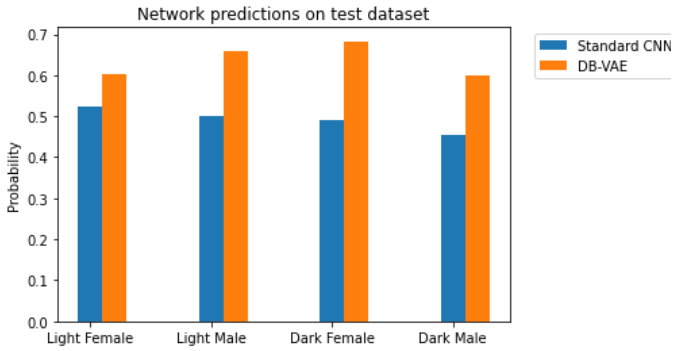
What we hope the model can do:

We want to use the latent variables learned previously to adaptively re-sample the dataset during training. We will try to alter the probability of a given image during the training based on how often the latent features appear in the dataset. The faces with rarer features will become more likely to be sampled during the training, while the overrepresented samples will decrease.

4. Evaluation of results

After creating a DB – VAE Loss function and structure, the two models were then placed together with the help of a helper function and re-sampled the dataset to provide the following outcomes:

(See GitHub for more detailed explanation)



Training time: 3hrs

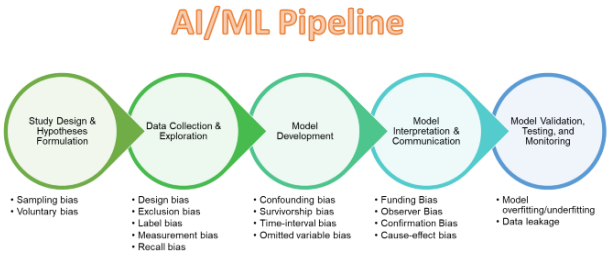
Number of times model was trained: 5

Epoch variation: 2, 6, 8, 10, 20

Based on the training time and parameter variation, the above results have been the best thus far.

With the help of the DB-VAE, the network's predictions have become better at assessing facial features across each demographic.

Further exploration is required to consider this model a success, but first, in order to do so, we need to turn to the AI/ML model pipeline and see the areas where we can implement change.



(*Eliminating AI Bias. Identifying AI Bias and knowing how to… | by Sheena Srivastava | Towards Data Science,* no date)

As shown in the above diagram, bias can occur from the absolute beginning of the cycle, with sampling bias (where only a subset of the given population has been sampled), data collection, pre-processing, and exploration. These are just a few of the instances where bias can occur.

So how can we prevent bias?

When deciding on what data and features to include in any research, focus on the study's design. Aim to make the sample dataset representative of the target population. In short, we need to implement more intention into selection and data curation.

Other measures that we can take and good habits we can introduce before starting any project are:

We can use existing toolkits to assist with AI bias mitigation, and thorough training must be taught within organizations and businesses. It can be implemented at the learning stage at universities, so an individual can critically assess a model's output and enable their understanding of AI/ML algorithms and question their validity.

5. Further exploration

The model can be tested on either more datasets, different datasets, a change in activation functions, or a GAN-based de-bias model against a standard CNN.

Another avenue of further exploration could be about a debiasing search engine result. For example, Google vs duck duck go, when given a specific word and the output is not as diverse as we would expect it to be Example word: Beautiful

.

Toolkits and Bibliography:

(*Detect Pretraining Data Bias - Amazon SageMaker*, no date; *Facessd Fairness Indicators Example Colab.ipynb - Colaboratory*, no date; Voicu, 2018; Wang and Russakovsky, 2021)

9 Types of Data Bias in Machine Learning - TAUS (no date). Available at: https://blog.taus.net/9-types-of-data-bias-in-machine-learning (Accessed: March 17, 2022).

CelebA Dataset (no date). Available at: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (Accessed: March 17, 2022).

Debiasing, a facial detection system, using VAEs: detailed overview | by siddhant kandge | Medium (no date). Available at: https://medium.com/@kandgesid/debiasing-a-facial-detection-system-using-vaes-detailed-overview-cbf736db7fce (Accessed: March 17, 2022).

Debiasing Variational Autoencoder. Architecture of the semi-supervised... | Download Scientific Diagram (no date). Available at: https://www.researchgate.net/figure/Debiasing-Variational-Autoencoder-Architecture-of-the-semi-supervised-DB-VAE-for-binary_fig2_334381622 (Accessed: March 17, 2022).

Detect Pretraining Data Bias - Amazon SageMaker (no date). Available at: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-detect-data-bias.html (Accessed: March 17, 2022).

Eliminating AI Bias. Identifying AI Bias and knowing how to… | by Sheenal Srivastava | Towards Data Science (no date). Available at: https://towardsdatascience.com/eliminating-ai-bias-5b8462a84779 (Accessed: March 17, 2022).

Facessd Fairness Indicators Example Colab. ipynb - Colaboratory (no date). Available at: https://colab.research.google.com/github/tensorflow/fairness-indicators/blob/master/g3doc/tutorials/Facessd_Fairness_Indicators_Example_Colab.ipynb#scrollTo=ZF4NO87uFxdQ (Accessed: March 17, 2022).

Goodman, B. and Flaxman, S. (2017) "EU regulations on algorithmic decision-making and a 'right to explanation,'" *AI Magazine*, 38(3), p. 50. doi:10.1609/aimag.v38i3.2741.

Lu, P., Song, B. and Xu, L. (2020) "Human face recognition based on convolutional neural network and augmented dataset," *http://mc.manuscriptcentral.com/tssc*, 9(S2), pp. 29–37. doi:10.1080/21642583.2020.1836526.

mathematical-statistics - How does the reparameterisation trick for VAEs work, and why is it important? - Cross Validated (no date). Available at: https://stats.stackexchange.com/questions/199605/how-does-the-reparameterization-trick-for-vaes-work-and-why-is-it-important (Accessed: March 17, 2022).

Reimers, C. *et al.* (2021) "Towards Learning an Unbiased Classifier from Biased Data via Conditional Adversarial Debiasing." doi:10.48550/arxiv.2103.06179.

Russakovsky, O. *et al.* (2015) "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, 115(3), pp. 211–252. doi:10.1007/S11263-015-0816-Y.

"The Fitzpatrick Skin Type Classification Scale" (no date) *Skin Inc.* [Preprint], (November 2007). Available at: http://www.skininc.com/skinscience/physiology/10764816.html (Accessed: March 17, 2022).

Voicu, I. (2018) "Using First Name Information to Improve Race and Ethnicity Classification," *Statistics and Public Policy*, 5(1), pp. 1–13. doi:10.1080/2330443X.2018.1427012.

Wang, A. and Russakovsky, O. (2021) "Directional Bias Amplification." Available at: http://arxiv.org/abs/2102.12594 (Accessed: March 17, 2022).