Abstract

Training dataset biases are by far the most scrutinized factors when explaining algorithmic biases of neural networks. In contrast, hyperparameters related to the neural network architecture have largely been ignored even though different network parameterizations are known to induce different implicit biases over learned features. For example, convolutional kernel size is known to affect the frequency content of features learned in CNNs. In this work, we present a causal framework for linking an architectural hyperparameter to out-of-distribution algorithmic bias. Our framework is experimental, in that we train several versions of a network with an intervention to a specific hyperparameter, and measure the resulting causal effect of this choice on performance bias when a particular out-of-distribution image perturbation is applied. In our experiments, we focused on measuring the causal relationship between convolutional kernel size and face analysis classification bias across different subpopulations (race/gender), with respect to high-frequency image details. We show that modifying kernel size, even in one layer of a CNN, changes the frequency content of learned features significantly across data subgroups leading to biased generalization performance even in the presence of a balanced dataset.

1. Introduction

Algorithmic biases of a deep neural network, i.e., performance disparities across subgroups in the data distribution, are most often attributed to sampling biases in its training dataset where some groups of the data distribution have significantly lower or higher proportions than others. While an imbalance in the training data certainly has a strong influence on a deep network’s algorithmic bias [1, 11, 25, 40, 49, 59], it is just one factor in the development pipeline. Examples of other important factors in a model’s design include its parameterization and training objective function. While several previous works [4, 28, 33, 36, 41, 54] explore the impact of objective functions on bias and suggest fairness regularizers to include during training, a network’s parameterization, i.e., its particular functional form, has been virtually unexplored in the context of bias. A neural network has several parameterization hyperparameters that must be set by its designer, including the number of layers, activation functions, and convolutional kernel sizes in the case of a convolutional neural network (CNN). Each of these choices can affect the type of features learned by the model, which could in turn impact bias. In this work, we take a first step in showing how to causally link a convolutional neural network’s kernel size to its algorithmic bias.

Different network parameterizations are known to induce different implicit biases over learned features. For example, CNNs tend to learn representations that are sensitive to high spatial frequency features of the input signal [8, 51, 56]. Based on the Fourier uncertainty principle, this phenomenon may be attributed to the size of the convolutional kernels [8] – smaller kernel sizes result in features that span a greater range in the Fourier domain. This implicit feature bias may be exposed by injecting a certain high-frequency signal into test images and observing a drop in the algorithm’s performance [53]. Implicit biases typically do not harm the model on within-distribution test samples because the parameters are well-tuned to the statistics in that distribution. The frequency noise/energy injection is essential to push the images out-of-distribution (OOD), thereby linking implicit biases to the generalization performance of the network. In this work, we show that in addition to this effect, implicit biases also affect the algorithmic
bias of a network, such that features used by the network for one data subgroup may have significantly different characteristics for another. This in turn, can lead to disparate OOD performance on these groups.

The main contribution of our work is a causal framework for linking an architectural hyperparameter of a neural network to its OOD bias. Our framework is based on an experimental procedure in which one or more parameters of a neural network architecture are modified at a time and the resulting bias on OOD samples are studied. First, we train from scratch multiple versions of the same network model that differ only in their choice of hyperparameter, e.g., convolutional kernel size. Second, we construct an OOD test set of images by injecting controlled perturbations to them, and obtain each model’s prediction on each image. The perturbation type should be based on the intended implicit feature bias being studied. For example, in our experiments, we use adversarial attacks and energy injections in Fourier passbands to probe frequency biases. Third, we fit a linear regressor to predict a model’s performance on an OOD test image as a function of the hyperparameter choice, degree of perturbation to the image, and various image attributes. Fourth and finally, we use the regression coefficients to measure the hyperparameter’s causal effects on model performances across data subgroups. This analysis provides a quantitative answer to whether the hyperparameter has a disparate causal effect across data subgroups.

While our framework is general, we focused our experiments on studying the causal relationship between sensitivity to high-frequency image details induced by changes to convolutional kernel sizes and performance of face analysis classifiers across subpopulations (race/gender protected groups). We trained several research-grade face gender classifiers on public datasets, and show that modifying kernel size from a commonly used range: \(3 \times 3\) to \(11 \times 11\) even in just the first layer of these CNNs will not only change the frequency content of learned features, but that this change can vary significantly across race/gender groups. We established this effect using both adversarial perturbations and energy injections to the high-frequency bands of the test images. This work opens the door to further careful studies on understanding the impact of neural network design decisions on algorithmic bias.

2. Related Work

2.1. Fairness in computer vision

Studies on fairness in computer vision predominantly focus on measuring and mitigating possible biases of computer vision models and datasets [1,11,17,25,39,40,49,59]. Biases may be measured with a number of metrics [9,18,50] that quantify disparate performance differences of algorithms across population subgroups. Face recognition and analysis systems are often under the most scrutiny due to their sensitive nature [2,24,59]. Perhaps the most famous of these studies was “Gender Shades” study [6], which identified the systematic failings of face analysis systems on particular racial and gender demographics.

Natural image datasets are known to have sampling biases [1,38,49], i.e., their joint distributions of attributes are far from random. For example, the CelebA face dataset is known to have a higher proportion of females with young ages compared to males [2]. A model trained on such a dataset can inherit its biases [24,38,57]. Therefore, algorithmic fairness issues can be greatly mitigated if the algorithm is trained on a more balanced dataset. Human face datasets have been particularly scrutinized [2,26,29,30,35] as models trained on these data can exhibit systematic failings with respect to attributes protected by the law [27]. Multiple approaches to mitigate dataset bias include collecting more diverse examples [35], using image synthesis to compensate for distribution gaps [2,30,43,47,52,59], and resampling [31]. Our work, in contrast, is focused on understanding biases of deep learning models due to parameterization decisions instead of data. In addition, [10], [37], [42], [3] and [45] propose related ideas on searching for optimal hyperparameters taking fairness into account, but were not focused on computer vision tasks and architectures.

2.2. Adversarial attacks

An adversarial attack perturbs an image until a given network changes its prediction, usually by applying gradient descent on the image. The resulting changes to the image are high frequency, and imperceptible to the human eye. This lack of robustness has sparked many theories [5,13–16,21,34,44,48]. Recent work has shown that commonly found adversarial examples for state-of-the-art convolutional neural networks contain dataset-specific information [51]. Furthermore, these adversarial attacks reflect properties of the features learned by the model [8], and that the model is biased towards certain features based on their architectural choice [12]. In this work, we analyze the information carried out by the attacks as a function of different architectural hyperparameter choices. Furthermore, we explore a novel hypothesis that adversarial attacks may allow us to expose differentiable information captured by a model’s features across different dataset subpopulations.

2.3. Frequency biases in CNNs

In image processing, the most common way to represent pixel location is in the spatial domain by column (x), row (y), and z (value). The frequency (or Fourier) domain offers an alternative perspective on the signal, by decomposing it in terms of sinusoids of varying frequencies.

Several recent works have provided new insights into the
behavior of CNNs by studying the relationship between frequency content in input signals and a CNN’s predictions. For example, one finding is that high-frequency components play a significantly higher role in a CNN’s decision function and performance compared to human perception [51]. Please refer to Figure 2 for example of high/low-frequency components of an image. Another study showed that a contributing factor to this is that convolutional operations in CNNs will introduce an implicit bias towards using higher frequencies in an image [8]. Nonlinear activation functions such as the rectified linear unit (ReLU) could also be a contributing factor [22]. We build on the findings in these works to study the effect of frequency-based features in differentiable algorithmic performance across dataset subgroups like gender and race. The Discrete Fourier Transform (DFT) is commonly used to transform an image between the spatial and frequency domain.

3. Methods

Our goal is to uncover the causal effect of convolutional kernel size on potential algorithmic biases due to an alteration of learned feature characteristics. We propose a framework to do this (see Fig. 1 for an overview). Our framework consists of three key steps. We first train $K$ versions of the same network architecture that differ only by the choice of a single hyperparameter. The hyperparameter choice acts as a causal intervention, giving us an experimental rather than observational testing procedure. We then apply an out-of-distribution (OOD) perturbation to a set of test images that are annotated with various attributes of interest, including “protected” attributes (e.g., race and gender for faces) that we will use for bias analysis. Finally, we use a linear regression to predict some measure related to model predictions, given covariates such as the hyperparameter choice, image attributes, and OOD perturbation degree. We use the regression coefficients as estimates of causal effects of the various factors on the model, and specifically compare the differences between coefficients corresponding to protected attributes to evaluate bias. We describe the three steps of our framework in the following sections.

3.1. Architecture training

We first train $K$ different versions of the same architecture, all identical in structure except for a modification to the hyperparameter of interest. We train all architectures on the same training dataset. We also initialize the weights and biases of all networks from identical normal distributions (i.e., identical mean and variances). After the networks are trained, we “freeze” their parameters, and will not modify them further in our framework.

3.2. OOD perturbations to test data

Given various trained architectures, our goal is to amplify biases across their learned features. One option is to run test samples coming from the same distribution as the training data through these networks, and measure performance across different protected attribute subgroups. The problem with this strategy is that deep neural networks are over-parameterized and therefore able to fit any training distribution nearly perfectly. Hence, even if a hyperparameter is altered from one network to another, both networks will likely yield similar performances on training data points.

However, as demonstrated in past works [53, 56], out-of-distribution (OOD) samples can paint a far different picture, with some models suffering in performance compared to others, thereby exposing differences across learned features. Therefore, a key step in our model is to inject a test set of images with a subtle class of perturbations so that they become OOD. In our experiments, we focus on frequency-related implicit biases of CNNs, and so we consider two types of perturbations from the neural network literature: adversarial attacks, and frequency energy injections.

3.2.1 Adversarial attacks

We consider two types of adversarial attacks in our experiments. The first, FGSM [16], applies gradient descent on the loss of the network’s output with respect to the input image to “nudge” the image in incremental steps towards a direction that changes the network’s prediction. The second, CW attack [7], utilizes two separate losses: a gradient-based loss to make the classifier change its prediction (similar to FGSM), and a regularization to make the magnitude of the change to the image as small as possible. This makes the perturbation distance (i.e., $l_2$ norm of the difference between perturbed image and original image) of CW attack a useful metric for measuring the degree of difficulty to perturb an image. We show an example of a CW and FGSM attack for the same input image in Figure 1 in Supplementary, which further shows that CW perturbation is an order of magnitude smaller due to the effect of its regularization.

3.2.2 Frequency energy injection

We also experiment with injecting energy to a specific frequency band to obtain a more fine-grained link between frequency content and network features. Fig. 2 depicts our process. For each test image, we use the DFT to obtain a Fourier spectrum, and amplify the amplitudes of Fourier coefficients lying on an annulus in the spectrum. In particular, let $F[\omega_x, \omega_y] = |A| e^{-j\phi}$ represent a complex coefficient in the Fourier spectrum of an image at location $(\omega_x, \omega_y)$ (corresponding to $x$ and $y$ frequencies), with radius $r = \sqrt{\omega_x^2 + \omega_y^2}$, lying in the annulus defined by $(r - r_0)^2 \leq \Delta^2$. We increase the amplitude $A$ by a factor of $1 + \delta$, to yield a modified coefficient $F'[\omega_x, \omega_y] = (1 + \delta)|A| e^{-j\phi}$. In our experiments, we set $\Delta = 2$, and $\delta = 15$ and $r_0 > 0$ is the frequency radius into which we are injecting energy [56]. If $r_0$ is small (large), we are modifying low (high) frequency
3.3. Causal analysis

We run the test set of OOD images through the $K$ networks, yielding $K$ predictions per image. We assume each image also comes with annotations for various relevant semantic attributes (including protected attributes with which we may compute algorithmic bias measures), as well as perturbation attributes (e.g., frequency of energy injection). Our goal is to measure the causal effects of the architectural hyperparameter of interest (kernel size) on model performance per protected attribute subgroup.

To do so, we use a multivariable linear regression model that predicts a dependent variable from multiple independent variables. For test image $i$ processed in network $k$, let $x_k$ be the corresponding hyperparameter and $y_{ik}$ be a measure of network performance on image $i$. Then we can specify the following regression equation:

$$y_{ik} = \beta x_k + \epsilon_{ik}^0,$$  \hspace{1cm} (1)

where $\epsilon_{ik}$ is an error term. Our coefficient of interest is $\beta$. Under the assumption that $E[x_k \cdot \epsilon_{ik}^0] = 0$, we can interpret $\beta$ as the causal effect of network architecture on performance. Of course, this independence assumption is unlikely to hold, as image attributes, including the OOD perturbation value, will generally affect a neural network’s performance.

We can weaken this assumption using a vector of image attributes $Z_i$, and augmenting equation (1) as follows:

$$y_{ik} = \beta x_k + Z_i' \gamma + \epsilon_{ik}^1,$$  \hspace{1cm} (2)

where $\gamma$ is a vector of coefficients, and $\beta$ is the causal effect of network architecture on performance under the weaker assumption $E[x_k \cdot \epsilon_{ik}^1 | Z_i] = 0$. Moreover, we hypothesize that the effect of architecture hyperparameter $x$ on performance may vary by protected attributes, a subset of all image attributes in $Z_i$. In order to allow for this possibility, we further augment equation (2) as follows:

$$y_{ik} = P_i' \beta x_k + Z_i' \gamma + \epsilon_{ik}^1,$$  \hspace{1cm} (3)

where $P_i$ is a vector of protected image attributes, and $\beta$ is now a vector of coefficients. We use a heuristic approach to choose the vectors $Z_i$ and $P_i$ that is commonly used for causal inference in the social sciences [55]. First, we incrementally add controls to $Z_i$ and test whether our estimates of $\beta$ change under alternative specifications (using an F-test with the null-hypothesis that the estimates of $\beta$ are equal across specifications). This is a test of the exogeneity assumption; if $Z_i$ is a sufficiently rich vector of controls to satisfy $E[x_k \cdot \epsilon_{ik}^1 | Z_i] = 0$, then the assumption will also be satisfied conditional on an augmented vector of controls. Second, we start with a rich vector $P_i$ to allow for the effect of network architecture on performance to be highly flexibly estimated. In our application, we begin by specifying $P_i$ as a fully saturated vector of dummy variables corresponding to all protected attribute combinations.
Figure 2. Examples of frequency energy injection perturbations. Our goal with this perturbation is 'jitter' all frequencies with the same magnitude in an input image. To do this, we inject random noise to the Fourier coefficients of an image lying on an annulus of a particular radius in the Fourier spectrum (top row of annulus), according to:

$$(1 + \delta)|A|e^{-j\phi}.$$ 

(e.g., White Male, White Female, etc.) and estimate $\beta$. We then test whether the elements of $\beta$ are equal to each other using pairwise F-tests. If we are unable to reject equality of coefficients, we cannot reject that the effect of network architecture on performance varies across those two groups.

$\beta$ encodes the joint causal effects of hyperparameter value $x$ and protected attributes in $P_i$ on output $y$. In particular, $\beta_g$ is the expected change in $y$ due to a unit change to $x$, when feature $g$ is “True” (set to 1). In our experiments, we compare the values in $\beta_g$ corresponding to different protected attribute subgroups to one another (see Table 2 in Supplementary, and Fig. 6).

4. Experiments & Results

We evaluated our work on the task of gender classification from face images using two popular datasets: Fairface [23] and UTKFace [58]. Fairface contains a roughly equal number of samples from different race/gender groups, and has 86,744 training and 10,954 testing samples. FairFace contains labels for 7 race groups (‘East Asian’, ‘White’, ‘Latino Hispanic’, ‘Southeast Asian’, ‘Black’, ‘Indian’, ‘Middle Eastern’) and 2 gender groups (‘Male’ and ‘Female’). UTKFace contains 20,000 training and 3,705 testing samples, but is not balanced across race groups. It contains labels for 5 race groups (‘White’, ‘Black’, ‘Asian’, ‘Indian’, ‘Others’) and two gender groups (‘Male’ and ‘Female’). We remove faces from ‘Others’ because they have no consistent characteristics. To mitigate effects of sampling biases during training, we used inverse sampling based on the number of examples from each race group. Training details are in Section A in Supplementary.

We demonstrate results using the ResNet-34 [19] base architecture for our experiments but obtained similar results using two other popular networks: DenseNet [20] and VGG-16 [46]. Please refer to Section D, E in Supplementary for results using these two models. Our architectural hyperparameter of interest was convolutional kernel size. We considered two different scenarios: changing only the kernel size of the first layer and changing the kernel size of all layers simultaneously. Interestingly, both scenarios yielded similar results, and so we leave results for the latter in Figure 6 in Supplementary. We varied the first layer kernel size (FLKS) within the range $[3, 11]$, which encompasses the popular choices for this hyperparameter for nearly all CNNs in the literature. We initialize the weights and biases of all of our models randomly by drawing from a Normal distribution with variance set to 0.02. For each network and kernel size value, we trained 3 independent models and presented average results to mitigate the influence of random initialization factors.

We report our networks’ accuracies for different race groups on non-OOD test images in Table 1 in Supplementary, to demonstrate that they all achieve high accuracies on both datasets. The performances do not significantly vary with FLKS because the training and testing images are all from the same distribution. We now present our results separately for the two OOD perturbations described in Sec. 3.2: adversarial attacks and frequency energy injections.

4.1. Adversarial attacks

We present results in this section using the CW adversarial attack. We obtained similar results using FGSM (see Figure 8 in Supplementary).

4.1.1 Analyzing Fourier spectra

We first visualize the average Fourier spectra magnitudes of the adversarial perturbation images split by race/gender groups and FLKS in Fig. 3 (for Fairface) and the results for UTKFace are in Figure 7 in Supplementary. Results on both datasets show similar trends. First, as FLKS increases, the
spectral energy becomes more focused at low-frequencies (closer to center). Second, holding FLKS value constant, we see that the spectrum for the Black group consistently contains less high-frequency energy compared to the spectra of other race groups. This result also holds for the Male group compared to Female. The difference between different subgroups shrinks as FLKS increases, in line with findings from a previous study showing that low FLKS leads to higher implicit frequency bias [8].

To quantitatively assess differences in the perturbation spectra, we also compute the \( f_{0.5} \) metric, known as “half power frequency,” or the frequency below which half of the signal’s power lies. \( f_{0.5} \) is a robust measure of energy concentration in a spectrum. Fig 4-top shows the \( f_{0.5} \) scores for the spectra, confirming the visual trend observed in Fig. 3. Please refer to the caption of Fig. 4 for more details.

### 4.1.2 Perturbation distances

We next present the average perturbation distances of adversarial attacks across race groups and models in Fig 4-bottom. The perturbation distance \( d_p \) between an original test image \( I \) and the perturbed image \( I’ \) may be computed by simply taking an L2 norm: \( d_p(I, I’) = ||I’ - I||_2 \), and quantifies how close/far the perturbed image to the original image. A larger distance indicates that more “work” must be done harder to fool the model, and its a reflection of the robustness of the model to other OOD perturbations [32].

Results show that perturbation distance (and variance) increases with FLKS for all race groups. It is therefore harder to adversarially attack a model with a larger FLKS, likely because such a model focuses more of its energy on low-frequency image information (see Fig 3) and is therefore robust. In addition, we see that \( d_p \) for the Black group is significantly lower than that of other groups. Please refer to the caption of Fig 4 for more details.

### 4.1.3 Causal analysis

Next, we quantitatively analyze the causal relationship between race and gender on perturbation distance \( d_p \) by applying our causal analysis framework introduced in Sec. 3.3. Specifically, using Eq. 3, we set \( y \) to be \( d_p \), and set both \( P_i \) and \( Z_i \) to contain “dummy variables” corresponding to all race/gender combinations. We use the Fairface dataset for this analysis, and use the race groups East Asian, White, Latino Hispanic, Southeast Asian, Indian and Black, and gender groups of Male and Female.

We use the `statsmodel` package from Python to run this regression, and the results of values of \( \beta \) and \( \gamma \) are in Table 2 in Supplementary. The results of \( \beta \) are also shown in Fig. 6(a). Based on the results, it is obvious that the coefficients for Black and Indian are significantly higher than
4.2. Frequency Energy Injection

We next perform frequency-based OOD perturbations to the test images as described in Sec. 3.2.2 and visualize results in Fig. 5. Accuracies of all models/groups are more influenced by perturbations to low-to-mid frequencies \((0.02 - 0.20(Hz))\) than to mid-to-high frequencies. FLKS of 3 is less affected by frequency injections in the range \((0.08 - 0.15(Hz))\). However, in general, it is difficult to distill significant trends from the plots alone.

4.2.1 Causal analysis

Similar to Sec. 4.1.3, we now perform a regression to measure the impact of kernel size, frequency of energy injection, and protected attribute subgroup on model error
Figure 6. Regression coefficient values for $\beta$ for different race groups. (a) $\beta$ values for the regression in Sec. 4.1.3 linking kernel size to adversarial perturbation distance, also shown in Table 2 in Supplementary. There are significant differences across protected groups, e.g., the Black and Indian group has significant higher values compared to the other groups. (b) $\beta$ values for the regression in Sec. 4.2.1. The coefficients for Black are always lower than those of other race groups. In addition, the coefficient values for different frequencies within the same race group are also significantly different.

per image. Using Eq. 3, we set $y$ to be the error rate of an image and set both $P_i$ and $Z_i$ to contain “dummy variables” corresponding to race/frequency (of injected energy) combinations. We use four frequency subgroups: \{(0.05, 0.07), (0.09, 0.11), (0.13, 0.15), (0.17, 0.19)\}(Hz), which we label 1, 2, 3, 4 for convenience. We report the results of coefficient $\beta$ values in Fig 6(b). It is clear that the coefficients for frequency group 4 are significantly smaller than those of the rest of the frequencies, indicating that changes to kernel size influence the performance less on the OOD samples under relatively high-frequency injections. The coefficient for the Black group in frequency group 1 is also significantly smaller than those of the other groups. As the frequency increases, this gap reduces.

5. Discussion and Conclusion

Our results in Figs. 3, 4, 5, 6 first demonstrate that smaller convolutional kernel sizes can cause a CNN to be biased towards high-frequency features, and increasing the kernel size mitigates this bias. We also see that such frequency bias significantly differed across different race/gender subgroups. All models trained on the two datasets focused less on high-frequency features for the Black and Male subgroup. While different features do not necessarily indicate performance bias on test samples, our results allow us to conclude that these differences do lead to performance biases on out-of-distribution (OOD) samples. We observed that this is the case for two different types of OOD image perturbation operators: adversarial attacks and frequency domain energy injections.

Different population subgroups will have different image characteristics. For example, the Black group will likely have darker skin tones than other race groups, and Females will have more hair on average than Males. Hence, it is not surprising that there is some difference in how images from one race are processed by a network compared to another. However, our results indicate something more significant: that there is a fundamental difference in the frequency characteristics of the image features across groups used by the network to make its decision. This difference may also lead to a performance bias, depending on the type of OOD data model is faced with. Our two OOD perturbations, while conceptually clear and well-motivated, are not associated with any real phenomena. It would be an interesting next step to relate frequency biases of features to disparate model performance on real-world OOD artifacts like shot noise, fog, and motion blur [56].

Our work has several limitations. We cannot draw broad conclusions about the nature of kernel size for general CNNs across all applications, because we focused on a single application of interest. A further evaluation on a wider set of application domains is an important next step. We also limited our causal analyses to a few key variables. However, causal analysis typically relies on the “no hidden confounders” assumption. An exhaustive set of image factors will help in computing more precise causal effects.

We focused on convolutional kernel size of a network in this work due to past results establishing a clear link between this hyperparameter and frequency content [8]. However, our framework is agnostic to the nature of the hyperparameter. Indeed, next steps in this research space include similar analyses into a more comprehensive set of network hyperparameters, such as activation functions, depth of layers, weight initialization strategies, and even high-level designs (e.g., residual connections, transformer modules). We see our work as a first step in the important direction of understanding how neural network design choices impact bias, and hence, the fairness of these systems in our society.
References


[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2019. 1


Empirically analyzing the effect of dataset biases on deep face recognition systems. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018. 2


[40] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9301–9310, 2021. 1, 2


