Addressing Weak Decision Boundaries in Image Classification by Leveraging Web Search and Generative Models

Preetam Prabhu Srikar Dammu[1], **Yunhe Feng**[1,2] and **Chirag Shah**[1]
[1]University of Washington, Seattle, WA, USA
[2]University of North Texas, Denton, TX, USA
preetams@uw.edu, yunhe.feng@unt.edu, chirags@uw.edu

Abstract

Machine learning (ML) technologies are known to be riddled with ethical and operational problems, however, we are witnessing an increasing thrust by businesses to deploy them in sensitive applications. One major issue among many is that ML models do not perform equally well for underrepresented groups. This puts vulnerable populations in an even disadvantaged and unfavorable position. We propose an approach that leverages the power of web search and generative models to alleviate some of the shortcomings of discriminative models. We demonstrate our method on an image classification problem using ImageNet's People Subtree subset, and show that it is effective in enhancing robustness and mitigating bias in certain classes that represent vulnerable populations (e.g., female doctor of color). Our new method is able to (1) identify weak decision boundaries for such classes; (2) construct search queries for Google as well as text for generating images through DALL-E 2 and Stable Diffusion; and (3) show how these newly captured training samples could alleviate population bias issue. While still improving the model's overall performance considerably, we achieve a significant reduction (77.30%) in the model's gender accuracy disparity. In addition to these improvements, we observed a notable enhancement in the classifier's decision boundary, as it is characterized by fewer weakspots and an increased separation between classes. Although we showcase our method on vulnerable populations in this study, the proposed technique is extendable to a wide range of problems and domains.

1 Introduction

Computer vision applications have become incorporated into several daily activities in modern societies, and the user base of these applications appears to be growing worldwide as more developing societies are exposed to them. Despite the widespread attention and maturity of the field, this technology and its manifestation in various applications suffers from issues that could have harmful societal impacts. Studies have shown that underrepresentation of certain demographics in datasets imparts bias to machine learning (ML) models [Zhao *et al.*, 2017; Hendricks *et al.*, 2018; Buolamwini and Gebru, 2018]. This could result in such underrepresented groups becoming more vulnerable, as the negative impacts of these ML services could have far-reaching consequences. Nevertheless, many businesses have rolled out services that rely on flawed technologies in order to expand to untapped markets.

Businesses often fail to address ML models' performance issues for underrepresented and vulnerable populations because (1) they lack enough resources (primarily, data) required to fairly train their ML models; and/or (2) there may be a concern of how specifically focusing on small groups could negatively affect the performance of large groups, which may bring down the overall accuracy of the models. Technically and economically, it may be prohibitive to have an overall blanketed approach to fix the discrimination problem in an ML model. But, if we could identify specific weakspots in a model and fix them without significantly affecting the rest of the model, we could address this problem of discrimination without sacrificing the overall performance of the model.

In this work, we present a way to leverage generative models and the web to address the challenging task of mitigating bias in services provided to vulnerable populations, which is an essential step towards achieving two of *UN's Sustainable Development Goals (SDGs)*: gender equality (SDG-5) & reducing inequalities (SDG-10).

To the best of our knowledge, this is the first of its kind attempt to address discrimination against underrepresented (and often vulnerable) classes using a combination of web search and image generation models while also providing a novel framework for enhancing robustness by improving decision boundaries. The rest of the paper is organized as follows. After reviewing some of the related works in Section 2, we provide an overview of the problem and approach in Section 3. The details of our method are presented in Section 4. In Section 5, we describe the datasets used for our experiments, following the experimental details and results in Section 6. Given that this is a new method for addressing an important problem of bias in ML, we discuss what this means for addressing the needs of vulnerable populations and the UN's SDGs (specifically, SDG-5 and SDG-10) in Section 7. The paper is concluded in Section 8 with some remarks on the current state of this research and future directions.

2 Related Work

In this section, we review some of the related concepts and relevant literature required to better understand this work.

2.1 Bias in Data

ML models, in general, are built to learn patterns and associations present in the data without questioning their validity and appropriateness. Perhaps, a more concerning finding is that ML models often amplify the bias present in data [Wang *et al.*, 2019; Zhao *et al.*, 2017]. To mitigate bias in ML models resulting from imbalanced or inadequate datasets, researchers have proposed several approaches which include balancing datasets in an attempt to address underrepresentation [Yang *et al.*, 2020; Minot *et al.*, 2022; Buda *et al.*, 2018; d'Alessandro *et al.*, 2017; Liu *et al.*, 2008].

However, balancing the dataset in terms of the number of samples per class may not be sufficient [Słowik and Bottou, 2021; Buda *et al.*, 2018; Wang *et al.*, 2019]. For instance, images belonging to the same class might contain information that varies significantly and this may induce biases even when the dataset is balanced. In order to mitigate bias infusing patterns in datasets, it is paramount to identify spurious correlations learned by the ML model and procure training samples that have a neutralizing effect.

2.2 Robustness in ML Models

In addition to bias issues, lack of robustness is also a well-known cause for concern when it comes to ML models [Carlini and Wagner, 2017; Hendrycks *et al.*, 2021; Taori *et al.*, 2020; Szegedy *et al.*, 2013]. In machine learning, robustness reflects the model's ability to not being significantly affected by varying conditions. However, a bulk of research has mainly been focused on a specific type of robustness, namely adversarial robustness, which deals with the model's ability to handle adversarial attacks [Goodfellow *et al.*, 2014]. Several scholars have studied the impacts of natural transformations such as changes in lighting conditions [Taori *et al.*, 2020; Wang *et al.*, 2021a]. A lesser explored case is the robustness to spurious correlations, which has recently gained more attention [Wang and Culotta, 2021; Wang *et al.*, 2021b; Singla and Feizi, 2022; Plumb *et al.*, 2022].

Improving adversarial robustness does not translate to enhanced robustness towards natural transformations [Wang *et al.*, 2021a] or variations arising from distribution shifts [Taori *et al.*, 2020]. For a model to be reliable, it needs to be robust against varying conditions that arise from the entropy of the real world, and not just from malicious entities.

Disproportionate object to class associations can give rise to spurious correlations, and these patterns compromise the classifier's robustness to distribution shifts [Singla and Feizi, 2022; Plumb *et al.*, 2022]. In [Plumb *et al.*, 2022], the authors rely on saliency maps and pixel-wise object annotations to identify spurious patterns, and then mitigate these patterns through data augmentation by counterfactual image generation. This method produces classifiers that are more accurate on distributions where the spurious patterns are not helpful and robust to distribution shifts. In contrast, our approach is more generalized as it handles spurious correlations among other shortcomings of the model.

2.3 Data Augmentation

Data augmentation is a widely used technique to address performance issues of ML models. Various approaches to implement data augmentation have been proposed for addressing pitfalls such as class imbalance, overfitting, bias issues, and distribution shifts [Kim *et al.*, 2021; Jaipuria *et al.*, 2020; Yucer *et al.*, 2020; Hu and Li, 2019; Sharma *et al.*, 2020]. Transformations as simple as rotation or random crop have been proven to improve classifiers [Mikołajczyk and Grochowski, 2018]. However, in applications where the data distribution is characterized by multiple varying factors, augmentation techniques with higher control over the synthetic augmentation process are required.

For instance, infinite unique datapoints are bound to exist in an unconstrained real-world environment, which makes capturing long tails of the distribution impractical [Jaipuria *et al.*, 2020]. To meet such complex requirements, augmenting data through generative techniques such as neural style transfer, GANs, VAEs, and simulation engines have been explored [Yucer *et al.*, 2020; Chen *et al.*, 2022]. However, each of these methods comes with its own set of limitations. Simulation engines serve as a powerful tool if the goal is to diversify scene attributes in robotic tasks [Chen *et al.*, 2022], but are not extendable to use cases beyond the simulated realm. Notably, the recent text-to-image generative models [Ramesh *et al.*, 2021; Rombach *et al.*, 2022] offer a higher degree of freedom and control in the generation process and have not been explored for data augmentation until now.

3 Overview

This section presents the research problem this paper aims to address and an overview of our proposed approach.

3.1 Research Problem

Despite various efforts described in the previous section for mitigating bias and robustness issues, we lack a systematic approach that pinpoints where exactly the performance issues for underrepresented classes are coming from and how to address them through data augmentation without disrupting the overall performance of the image classifier. We break this down into three subproblems: (1) identifying weakspots in the classifier's decision boundary; (2) procuring new datapoints that selectively enhance the decision boundary near the weakspots; and (3) leveraging the augmented data to mitigate the model's bias and enhance its robustness.

3.2 Approach Overview

To solve the problem described above, this paper proposes a framework that can automatically detect the weakspots in classifiers and, more importantly, leverage the internet's vastness and the emerging super-realistic text-to-image generative models to mitigate bias and robustness issues. We address all of the subproblems, and thus, our contribution is three-fold. The overview of the proposed framework is shown in Figure 1 and its workings are detailed in Section 4.

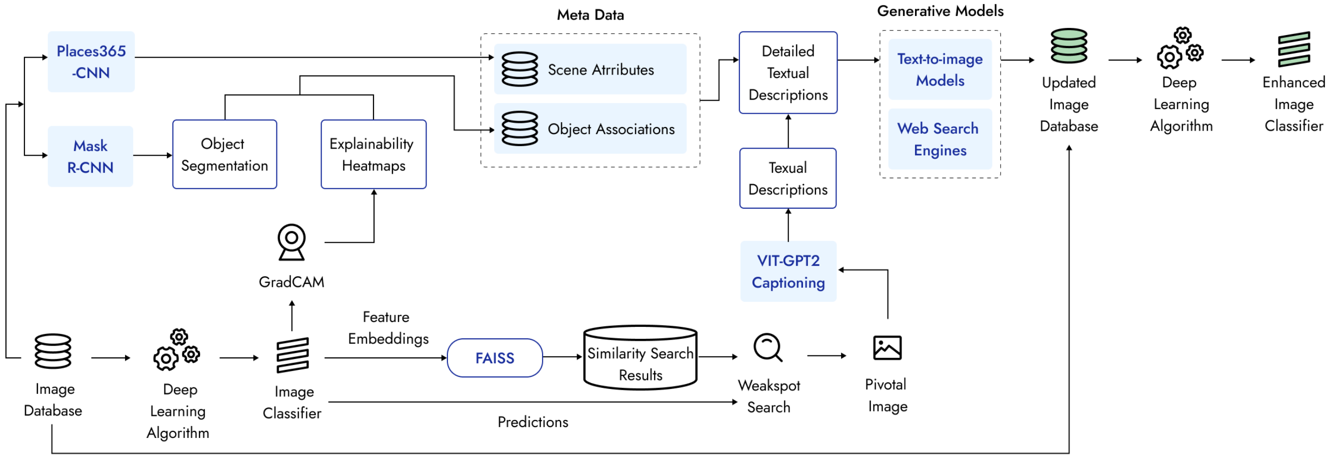

Figure 1: Overview of the proposed approach. Pivotal images representing the weakspots of the classifier are identified and used to generate detailed textual descriptions by leveraging the supporting metadata. New training samples are acquired using these descriptions through generative models which facilitate the enhancement of the classifier (Refer Section 4.3).

4 Methodology

In this section, we present the methodology for identifying the weakspots in the classifier's decision boundary, procuring new training samples that belong to these weak regions with high precision, and remedying the model's robustness and bias issues through strategically captured training data.

4.1 Identifying Model Weaknesses

Identifying Weakspots

To identify weakspots present in the classifier's decision boundary, we need to search the latent space for weak neighborhoods with relatively high perplexity. However, searching for weakspots in a large latent space is computationally intensive. Therefore, we adopt a powerful tool that uses GPU acceleration to perform similarity search, Facebook FAISS [Johnson *et al.*, 2019], thus improving the efficiency of weakspot search.

A sufficiently large number of data samples that can adequately represent the dataset's distribution are required to identify weakspots. Feature embeddings are extracted for each image in this representative set, and these feature vectors are fed to the similarity search algorithm. In our experiments, we use FAISS to perform this step, using the *IndexFlatL2* operation that retrieves top k neighbors along with their euclidean distance values for each datapoint. Subsequently, we perform a grid search on all misclassified instances to check if they lie near a weakspot. Consider an instance originally belonging to *class 1* erroneously labeled as *class 2*, keeping a maximum neighbor L2 distance d as radius, if at least a fixed threshold percentage of neighbors are correctly classified as *class 2*, we detect a weakspot between the two classes in consideration. The corresponding misclassified datapoint at the center of the weak region is identified as *pivotal* image.

Identifying Object Associations and Spurious Correlations

We employ a combination of deep learning tasks to identify object associations and spurious correlations learned by the deep learning model. As understanding the content of an image is an essential first step, we use scene recognition and object detection to obtain necessary image metadata that helps figure out which factor(s) is/are the primary contributors to a classifier's decision. To achieve this, we rely on explainability heatmaps to detect which objects present in the image appear to trigger the classification. If the explainability method detects pixels belonging to an object to have higher relevance beyond a certain threshold, an association between that object and the classifier's predicted class is detected for that particular instance. For example, in Figure 2, the first column consists of original images, the second column consists of segmented images, and the third consists of heatmaps overlayed on segmented images. Relevant associations between classes and objects identified in Figure 2 (a) *tennis_player* – 'tennis racket', 'sports ball', 'person', Figure 2 (b) *traffic_cop* – 'person', 'car', 'motorcycle', 'truck', Figure 2 (c) *ballplayer* - 'person', 'bench', 'baseball glove'.

It is non-trivial and highly subjective to decide whether an object association is inappropriate or not. From the observed associations, the next challenge is to filter out the ones deemed to be spurious. It requires human judgment, as AI is incapable of making conscious or ethical calls. Therefore, through manual intervention, we identify questionable associations made by the model. As the proposed approach identifies scenarios where the model is likely to fail (see Section 4.1a), manually checking for the presence of spurious correlations is feasible because the algorithm conveniently shortlists cases which need reviewing. Mitigating search phrases are subsequently added to the set of text prompts to be used for procuring neutralizing images. (see examples in Figure 6).

Addressing Bias Issues

To mitigate bias, we need to procure samples that neutralize patterns in the data that lead to harmful biases being learned by the ML model. In contrast to standard data balancing, this approach does not merely match the number of samples per class, rather it emphasizes the trends present in the data. This

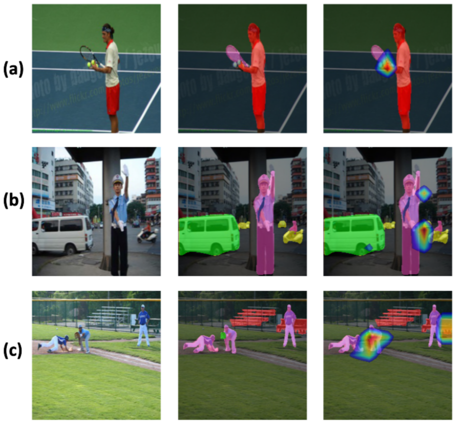

Figure 2: Object associations through heatmaps and segmentation.

is essential due to the fact that even when the data is perfectly balanced in terms of instances per class, harmful patterns could still be learned [Wang *et al.*, 2019]. As these patterns lead to weak regions in the decision boundaries, which are detected by our algorithm, effective counterexamples are automatically procured in subsequent steps (see 4.2b and 4.2c).

Addressing Robustness Issues

In this work, we address robustness towards model failures caused by natural triggers such as overfitting on scene information or learning spurious correlations. To address these robustness issues, the identified *pivotal images* representative of weak regions of the classifier are used as 'anchoring samples', analogous to 'support vectors' in SVMs [Chapelle *et al.*, 1999]. Subsequently, images with comparable content to the anchoring samples are either generated using generative models or retrieved using web search engines. Additional datapoints that are sampled from the weakspot's latent space are expected to alleviate the perplexity around it.

4.2 Fixing Model Weaknesses

Once we have identified the weakspots in a model, the next step is to find appropriate data to fix them. We do this using web search as well as image generation models.

Generating Search Phrases for Pivotal Images

Once weak regions have been identified, we attempt to procure samples from the latent space belonging to these regions. Pivotal images, as they are located at the centers of these regions with high perplexity, act as a good anchoring point to generate or retrieve similar samples. As we plan to retrieve samples from web search engines and text-to-image models which take text data as input, an accurate and specific text description of the pivotal image is crucial. To achieve high-quality descriptions, we use a combination of techniques.

Firstly, we use the Vit-GPT2 [Dosovitskiy *et al.*, 2020; Radford *et al.*, 2019] image captioning model to generate a caption for the pivotal image. However, these captions might lack the level of detail to use them for accurately generating new images. For instance, a common occurrence is that the captions are characterized with pronouns instead of a description of the person in the image. To remedy this, we replace all pronouns with the class label of the image, as these labels accurately represent the person seen in the image. Additionally, we use scene information generated by the Places-365 CNN model [Zhou *et al.*, 2017] to incorporate scene information into the textual description by only considering the high-level details, such as if the image is taken indoors or outdoors and the venue. These steps ensure that we obtain a sufficiently detailed and accurate description required to generate or retrieve highly relevant training samples.

Procuring Images through Web Search Engines

Once detailed descriptions have been generated, web search engines can be used to collect new training images. An advantage of using search engines is that most of the retrieved images can be observed in the real world. However, in cases where the textual description is of an uncommon instance, the retrieved images may not sufficiently match the search phrase or be irrelevant. For example, the search results of *a person of color female doctor* would be useful, but the results of *a male nurse with a potted plant on a desk* returned by image search engines would not adequately match the description. To address this gap, we generate images with those specific characterizations using generative models.

Procuring Images through Generative Models

In addition to retrieving image search engines, another way to collect desired images is generating them based on given text. The recently released text-to-image generative models, such as DALL-E 2 [Ramesh *et al.*, 2021] and Stable Diffusion [Rombach *et al.*, 2022], are able to generate high-quality super-realistic images that accurately match the text description of the pivotal image. Compared to web search engines, text-to-image models allow to sample the weak region more precisely, enabling the generation of highly effective training samples in a more accurate manner.

4.3 Flow of Operations

Here, we present the workflow of our proposed approach in Figure 1, and the corresponding algorithm in Algorithm 1. Initially, *Places465-CNN* [Zhou *et al.*, 2017] and *Mask R-CNN* [He *et al.*, 2017] are used to generate scene attributes and segmentation maps respectively for all datapoints, as this metadata is required in subsequent steps. On the image classifier trained on the original training set, we use *GradCAM* [Selvaraju *et al.*, 2017] to generate heatmaps, which are used in conjunction with object segmentation maps to obtain object associations. Next, *FAISS* is used to conduct a similarity search on the feature embeddings to generate the nearest neighbors along with their distances, which are fed to the weakspot search algorithm. The *pivotal images* representing the identified weakspots are then captioned using the *VIT-GPT2* model, and the resulting image descriptions are enhanced with the metadata generated in previous steps to obtain detailed textual descriptions. These descriptions are used for retrieving images using Google image search, as well as for generating images using the text-to-image models. Subsequently, the original dataset is augmented with new datapoints, and this updated dataset is utilized to train the enhanced image classifier.

Algorithm 1: Enhancing Classifier

```
Input: $D_{train}$: train dataset; $D_{test}$: test dataset; $C$: original classifier; $t_{dist}$: L2 distance threshold; $t_{perp}$: perplexity threshold;
Operations: $\mathbf{proc_{web}}()$: procure images from web; $\mathbf{proc_{txt2img}}()$: procure images from text-to-image model; finetune(): finetune model; perplexity(): compute perplexity; object_associations(): get object associations; get_neighbors(): get neighbors within $t_{dist}$; textual_description(): get textual description; find_spurious(): find spurious correlations;
Output: enhanced classifier $C'$;
$T_{desc} \leftarrow \varnothing$ ;  // initialize text descriptions as empty
$O_{asso} \leftarrow \varnothing$ ;  // initialize object associations as empty
for $x_i \epsilon D_{test}$ do
    $neighbors \leftarrow$ get_neighbors($x_i, t_{dist}$) ;
    $perp \leftarrow$ perplexity($neighbors$) ;
    if $perp > t_{perp}$ then // $x_i$ is detected as a pivotal image
        Insert textual_description($x_i$) into $T_{desc}$;
        Insert object_associations($x_i$) into $O_{asso}$;
    end
end
Insert textual_description(find_spurious($O_{asso}$)) into $T_{desc}$;
$D_{web} \leftarrow \mathbf{proc_{web}}(T_{desc})$ ;
$D_{txt2img} \leftarrow \mathbf{proc_{txt2img}}(T_{desc})$ ;
$D_{updated} \leftarrow D_{train} \cup D_{web} \cup D_{txt2img}$ ;
$C' \leftarrow$ finetune($C, D_{updated}$) ;  // return enhanced classifier
return $C'$
```

5 Datasets

ImageNet People SubTree: This subset of ImageNet contains 2,832 people categories, however, only 139 of these categories are considered safe and imageable [Yang *et al.*, 2020]. In our experiments, we only considered classes identified as safe and free from annotator bias. From these 139 classes, we selected the ones that are either a profession or an occupation. Categories which share the lowest common hypernym that has a broader yet specific definition of an occupation were merged together. For instance, *captain*, *chief_of_staff*, *general*, *major*, *Navy_SEAL*, *military_personnel* were all mapped to *military_officer*, as some of these classes are an abstraction of others. Finally, we ended up with 40 classes that were used to conduct our experiments. Each class had at least 100 images in both training and testing sets.

In [Yang *et al.*, 2020], the authors prepared annotations for 100 images per each synset, which resulted in 13,900 images. We filtered out categories that were not related to an occupation, and this step resulted in 6,278 images. We used all of the annotated images in the *test* set and the remaining images without annotations in the *train* set, creating a sample size of 6,278 and 64,516 respectively for each partition.

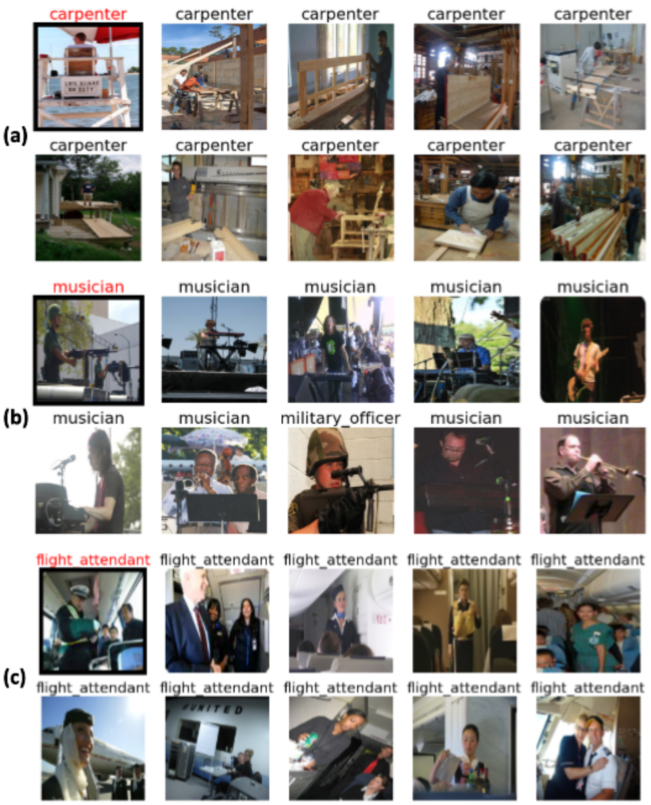

Figure 3: Samples from weak regions between classes (a) *lifeguard* and *carpenter* (b) *military_officer* and *musician* (c) *traffic_cop* and *flight_attendant*. Erroneous predictions are labeled in red. Images with black borders are *pivotal images*.

To conduct our experiments, we required further information to help us examine the constituents of the image. To this effect, we used object detection and scene recognition to generate the required additional metadata.

6 Experiments and Results

To demonstrate the efficacy of the proposed method, we started by building an image classifier on the 40 class subset. Using ResNet50 [He *et al.*, 2016], we retrained the last layer on our training set to obtain 80.12% test accuracy. However, this model demonstrated significant performance disparity. The accuracy for males across all categories was 81.76%, whereas for females it was 75.68%, amounting to a 6.08% gender accuracy disparity. We noticed that the model performed significantly worse for person of color female doctors, as the accuracy for this demographic is just 27.78% in comparison to 79.38% for the *doctors* class. After enhancing the classifier with our method, gender accuracy disparity dropped from 6.08% to 1.38% amounting to a 77.30% reduction.

Using our technique, we found weakspots in the decision boundary of the classifiers. Observing the identified weakspots could reveal valuable insights about the model be-

Classifier	Gender	Original			Enhanced		
		Acc	Prec	Rec	Acc	Prec	Rec
ResNet50	All	80.12	82.56	77.5	**84.09**	**84.52**	**82.28**
ResNet50	Male	81.76	79.09	75.24	**84.54**	**80.12**	**82.46**
ResNet50	Female	75.68	70.58	74.3	**83.16**	**78.1**	**79.32**

Table 1: Performance of the classifier before and after enhancement.

Figure 4: Images generated by DALL-E to mitigate perplexity in weak regions between *lifeguard* and *carpenter* (see Figure 3 (a)).

havior, leading to higher transparency of the automated decision process. Revealing the classifier's weaknesses could also introduce accountability, as the developers of the model would be informed of scenarios where it is likely to fail, and would have to prepare to handle such cases.

For instance, in Figure 3 (a), we notice that a lifeguard is being predicted as a carpenter due to the presence of wooden structures. By sampling near this weakspot between *lifeguard* and *carpenter* classes (see Figure 4), we help the model perform better in similar cases. In 3 (b), we see that the model confuses a soldier manning a machine gun with musicians playing an instrument, and the closest three neighboring images share similar scene attributes – all of them are situated outdoors with open skies. In 3 (c), we observe that a traffic cop is labeled as flight attendant because of the background, as the inside of the bus looks similar to an airplane cabin.

Using the *pivotal images* representing these weak regions, we procured neutralizing images to alleviate the perplexity present around weakspots in the decision boundary. Additionally, we also inspected for any spurious correlations learned by the model by observing the object associations in the identified weakspots. For instance, the presence of *potted_plant* object in images of nurses fools the model into misclassifying them as *gardener*. To counter this spurious correlation, we generated images of nurses with potted plants situated in front of them using DALL-E.

Probing performance disparity for the identified weakspots also revealed valuable insights. For instance, taking a closer look at the weakspot between *doctor* and *nurse* classes uncovered the model's bias against the underrepresented demographic of person of color female doctors. The classifier correctly classified doctors 79.38% of the time, however, it demonstrated a significant drop in accuracy for colored female doctors with an accuracy of 27.78%. After strategic retraining, the disparity was reduced by 49.37%.

To neutralize the weakspots identified by the technique, a total of 2,144 neutralizing training samples were procured, increasing the training set size by 3.32%. Despite being a relatively small-sized addition, the strategically crafted training samples resulted in a considerable improvement in the model's performance. The model's overall accuracy increased by roughly 4%, but more importantly, the gender accuracy disparity was reduced by 77.30% (see Table 1). Top five categories with the highest gender accuracy disparity have been tabulated in Table 2, and observe a significant reduction in the disparity for four of the five classes. Notably, this mitigation of bias was achieved without compromising the overall performance of the model.

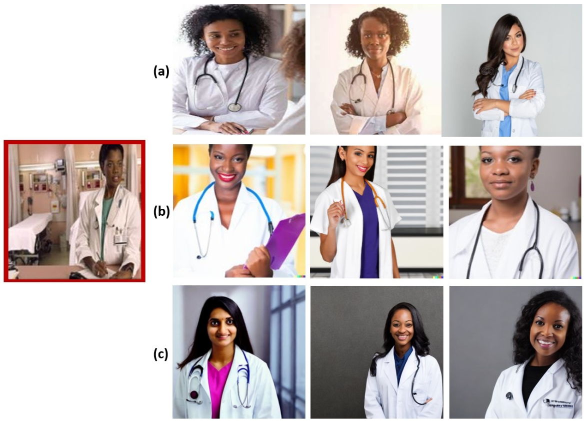

Figure 5: Person of color female doctors have the least accuracy across demographics (27.78%, overall is 79.38%). Representative image from ImageNet (left, red borders). Neutralizing images procured through (a) Web Search (b) DALL-E and (c) Stable Diffusion.

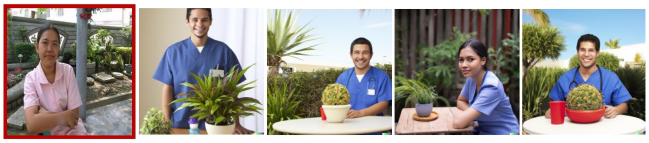

Figure 6: Spurious correlation between *potted plant* and *gardener* class (image shown with red border). Images on the right are images generated using DALL-E to counter this spurious correlation.

Sampling weak regions near the decision boundary and retraining with carefully crafted additional datapoints resulted in better-defined class boundaries with fewer weakpoints and better separation. As can be observed in Table 3, the maximum number of weakspots identified in the original model was 139 with *d=50* (L2 distance) as the radius around *pivotal* datapoint in the latent space. After improving the model with the proposed approach, no weakspots were identified at *d=50*, and few were detected at higher *d* values. This indicates a clear increase in inter-class separation and a more robust decision boundary of the classifier.

7 Discussion

Many ML models suffer from issues stemming from imbalance in datasets [Zhao *et al.*, 2017; Hendricks *et al.*, 2018; Buolamwini and Gebru, 2018]. Typically, this results in classifiers performing substantially worse for some of its minority classes, even when the overall performance is high [Buolamwini and Gebru, 2018]. The adverse effects of such biases are felt the most by underrepresented and vulnerable demographics, making them susceptible to larger harms of ML-based discrimination. For instance, studies have shown that

Class	Accuracy (original)			Accuracy (enhanced)			Accuracy disparity decrease (%)
	All	M	F	All	M	F	
Nurse	66.18	13.51	77.4	70.1	17.9	82.2	-0.59
Coalminer	83.33	86.52	33.3	87.1	87.8	50.0	**28.97**
Boatperson	78.15	81.73	33.3	96.8	97.6	75.0	**53.21**
Firefighter	78.95	81.48	33.3	76.9	79.6	50.0	**38.43**
Painter	65.0	74.68	30.0	72.8	75.6	55.6	**55.12**

Table 2: Top five classes with the highest performance disparity for male and female. Significant decrease in accuracy disparity is observed for four of the five classes. Improvements are emboldened.

Classifier	Number of weakspots						
	d=10	d=15	d=20	d=50	d=80	d=110	d=140
ResNet50, Original	52	112	133	139	139	139	139
ResNet50, Enhanced	0	0	0	0	11	44	87

Table 3: Number of weakspots identified with perplexity of 70%. *d* refers to radius of the weak regions in euclidean distance.

when it comes to certain job results in image searches, females and people of color are highly underrepresented [Lam *et al.*, 2018]. Even when search engines attempt to address this issue, recent research has shown that the fixes are often on the surface and not foundational [Feng and Shah, 2022]. Such bias in representation leads to cognitive bias, and perpetuates biases in data and algorithms [Baeza-Yates, 2018].

While several studies have attempted to solve the lack of diversity in datasets through data augmentation, previous approaches for generating new samples have limitations [Mikołajczyk and Grochowski, 2018; Kim *et al.*, 2021; Iosifidis and Ntoutsi, 2018; Jaipuria *et al.*, 2020; Yucer *et al.*, 2020; Hu and Li, 2019; Sharma *et al.*, 2020]. A common drawback of most of these approaches is that they are not extendable to other problems, even if they are shown to work well with a specific problem. This is due to the limited degree of freedom and control in the existing approaches. In contrast, our method uses text-to-image generative models that offer a higher degree of variation in multiple aspects such as the background, objects, and person attributes among others.

Any attempt to diversify datasets using traditional approaches such as collecting more data requires a substantial investment in terms of both time and money. In some cases, it might not even be feasible due to operational challenges. There is also a limit to how much diversification can be achieved through additional data collection. However, this should not justify the use of biased datasets to train ML models that may adversely affect vulnerable populations, and the responsibility falls on the ML community to devise solutions that address this challenge. As an alternative, we proposed an approach that could circumvent many of these hindrances by procuring diverse samples instantaneously at low costs.

Adversarial training typically enhances weak decision boundaries when gradient-based attacks are used to generate the samples, as these techniques compute the least amount of perturbation required to fool the classifier [Madry *et al.*, 2017; Goodfellow *et al.*, 2014]. However, improvements obtained through this approach are limited to robustness against adversarial attacks and ineffective against natural variations or distribution shifts [Wang *et al.*, 2021a; Taori *et al.*, 2020]. Moreover, they are not suitable for diversifying the demographics of the training samples. The proposed approach serves an ideal alternative which tackles these shortcomings.

Another major concern about the performance of ML models in deployment is that the distribution of the data it was trained may be different from the real-world data. For example, a medical diagnosis model trained on a predominantly western population may exhibit erroneous behavior when put to practice in other parts of the world. A study on chest X-ray pathology classification model demonstrated this issue, as patients from under-served demographics were underdiagnosed [Seyyed-Kalantari *et al.*, 2021; Bernhardt *et al.*, 2022].

In addition to portability from data distribution to another being an issue, we should also be concerned about the distribution of the same data evolving over time. Therefore, periodical evaluation of machine learning models in deployment is a requirement. As the data changes over time, new weakspots in the decision boundary of the classifier may arise. Consequently, the technique presented in this paper could be used to keep the ML model up to date to reflect natural changes emerging in the distribution of the data.

The proposed approach acts as a step in the right direction to solve the issues discussed above. In the future, we should aim to work towards methods that proactively prevent bias issues, instead of fixing the existing ones in a posthoc fashion.

8 Conclusion

In this paper, we presented a method to address three problems: (1) identifying weakspots in image classification; (2) procuring data appropriate for re-learning those weak boundaries; and (3) incorporating such data for training a classifier such that its bias and robustness issues are mitigated.

The proposed method to enhance discriminative models by leveraging generative models and web search engines instills various desirable characteristics such as robustness, fairness, and transparency to the original model while still improving the overall performance. In addition, the steps involved in executing this approach imparts model understanding and accountability, and as such should be used as a post-development practice before deployment. Finally, this method addresses an often overlooked problem of robustness towards spurious correlations and scene variations. By remedying weakspots through targeted sampling, the decision boundary of the classifier is enhanced with fewer vulnerable points and higher inter-class separation.

While we applied our method on a specific classification problem with a focus on certain vulnerable populations, the method presented in this paper is flexible and can be used to improve classifiers in various applications and domains. As demonstrated, we envision this approach to be extended to several other tasks and promoting better practices in the development of ML models.

References

[Baeza-Yates, 2018] Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, 2018.

[Bernhardt *et al.*, 2022] Mélanie Bernhardt, Charles Jones, and Ben Glocker. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, 28(6):1157–1158, 2022.

[Buda *et al.*, 2018] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[Chapelle *et al.*, 1999] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5):1055–1064, 1999.

[Chen *et al.*, 2022] Yiwen Chen, Xue Li, Sheng Guo, Xian Yao Ng, and Marcelo Ang. Real2sim or sim2real: Robotics visual insertion using deep reinforcement learning and real2sim policy adaptation. *arXiv preprint arXiv:2206.02679*, 2022.

[d'Alessandro *et al.*, 2017] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134, 2017.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Feng and Shah, 2022] Yunhe Feng and Chirag Shah. Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2022.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Hendricks *et al.*, 2018] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[Hu and Li, 2019] Mengxiao Hu and Jinlong Li. Exploring bias in gan-based data augmentation for small samples. *arXiv preprint arXiv:1905.08495*, 2019.

[Iosifidis and Ntoutsi, 2018] Vasileios Iosifidis and Eirini Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24, 2018.

[Jaipuria *et al.*, 2020] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 772–773, 2020.

[Johnson *et al.*, 2019] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[Kim *et al.*, 2021] Youmin Kim, AFM Shahab Uddin, and Sung-Ho Bae. Local augment: Utilizing local bias property of convolutional neural networks for data augmentation. *IEEE Access*, 9:15191–15199, 2021.

[Lam *et al.*, 2018] Onyi Lam, Brian Broderick, Stefan Wojcik, and Adam Hughes. Gender and jobs in online image searches. *Pew Social Trends. Retrieved March*, 14:2020, 2018.

[Liu *et al.*, 2008] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Mikołajczyk and Grochowski, 2018] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

[Minot *et al.*, 2022] Joshua R. Minot, Nicholas Cheney, Marc E. Maier, Danne C. Elbers, Christopher M. Danforth, and Peter Sheridan Dodds. Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance. *ACM Transactions on Computing for Healthcare*, 2022.

[Plumb *et al.*, 2022] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *Transactions on Machine Learning Research*, 2022.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Seyyed-Kalantari *et al.*, 2021] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[Sharma *et al.*, 2020] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364, 2020.

[Singla and Feizi, 2022] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Słowik and Bottou, 2021] Agnieszka Słowik and Léon Bottou. Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation. *ArXiv*, abs/2106.09467, 2021.

[Taori *et al.*, 2020] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[Wang and Culotta, 2021] Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031, 2021.

[Wang *et al.*, 2019] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.

[Wang *et al.*, 2021a] Shuo Wang, L. Lyu, Surya Nepal, Carsten Rudolph, Marthie Grobler, and Kristen Moore. Robust training using natural transformation. *ArXiv*, abs/2105.04070, 2021.

[Wang *et al.*, 2021b] Tianlu Wang, Diyi Yang, and Xuezhi Wang. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*, 2021.

[Yang *et al.*, 2020] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 547–558, 2020.

[Yucer *et al.*, 2020] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.

[Zhao *et al.*, 2017] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.