Poster: Distribution-aware Fairness Test Generation

Sai Sathiesh Rajan*, Ezekiel Soremekun†‡, Sudipta Chattopadhyay*, and Yves Le Traon‡,
*SUTD, Singapore, †RHUL, UK, ‡SnT, University of Luxembourg, Luxembourg
sai_rajan@mymail.sutd.edu.sg, ezekiel.soremekun@uni.lu, sudipta_chattopadhyay@sutd.edu.sg, yves.letraon@uni.lu

***Abstract*—This work addresses how to validate group fairness in image recognition software. We propose a distribution-aware fairness testing approach (called DISTROFAIR) that systematically exposes class-level fairness violations in image classifiers via a synergistic combination of out-of-distribution (OOD) testing and semantic-preserving image mutation. DISTROFAIR automatically learns the distribution (e.g., number/orientation) of objects in a set of images and systematically mutates objects in the images to become OOD using three semantic-preserving image mutations – object deletion, object insertion and object rotation. We evaluate DISTROFAIR with two well-known datasets (CityScapes and MS-COCO) and three commercial image recognition software (namely, Amazon Rekognition, Google Cloud Vision and Azure Computer Vision) and find that at least 21% of images generated by DISTROFAIR result in class-level fairness violations. DISTROFAIR is up to 2.3x more effective than the baseline (generation of images within the observed distribution). Finally, we evaluated the semantic validity of our approach via a user study with 81 participants, using 30 real images and 30 corresponding mutated images generated by DISTROFAIR and found that the generated images are 80% as realistic as the original images.**

I. INTRODUCTION

In recent years, image recognition systems have increasingly found their way into safety critical systems. For instance, in the autonomous driving context, identification of all the different objects in a given image i.e., multi-label object classification (MLC) [2] is crucial since failing to recognize an object can have disastrous consequences. Therefore, systematic testing of image classification systems, to detect potential bias against certain classes, is of critical importance.

Given an arbitrary MLC model (*system under test* (SUT)) and a set of initial images without any ground truth, our fairness test generation approach (DISTROFAIR) identifies the classes that are subject to unfair treatment (i.e., *unusually high error rates*). Additionally, each error is associated with concrete test images that can be used by the developer to further investigate the errors. The key insight behind our approach revolves around out-of-distribution (OOD) testing [1]. DISTROFAIR learns the distribution of objects detected in a set of images and systematically generates a set of images, named OOD images, outside such a distribution. Additionally, our systematic generation of OOD images allows us to detect when the mutated images induce errors even in the absence of ground truth. *To the best of our knowledge, we present the first OOD testing approach to discover and analyze the class-level fairness errors in image classification tasks.*

Figure 1 illustrates the different steps of our DISTROFAIR approach. Given a set of images, DISTROFAIR begins by clustering the images into different similar sub-groups allowing

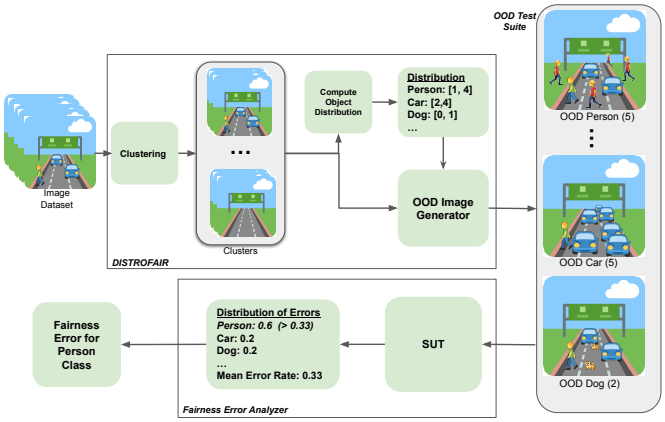
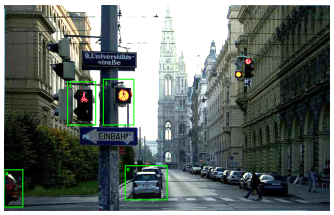
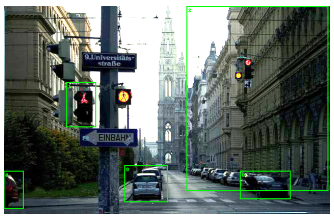


Fig. 1: An illustration of our DISTROFAIR approach

DISTROFAIR to account for the diversity of images in the original set. For each sub-group/cluster, DISTROFAIR then learns a distribution (e.g., number/orientation) for the objects. Subsequently, OOD images are generated by leveraging this information and using semantic-preserving mutation operators (e.g., insertion, deletion and rotation of objects). For instance, three OOD images are shown in Figure 1, each one exceeds the maximum number of "*Person*", "*Car*" or "*Dog*" objects detected by the SUT in the respective cluster. Finally, the SUT is subject to analysis on the generated OOD images. As observed in Figure 1, if a class (e.g., "*Person*") is detected with an error rate (i.e., 0.6) more than the mean error rate across all classes (i.e., 0.33), then DISTROFAIR highlights the class (i.e., *"person"*) as facing a fairness error.

In essence, we formalize how to measure class-level fairness errors and propose a novel OOD test generation approach (DISTROFAIR) that utilizes three semantic preserving metamorphic mutations to aid in the discovery of such errors. We validate the semantic validity of the generated images with a user study and find that our OOD images are about 80% as realistic as the original images. We also evaluate DISTROFAIR with the aid of three image classification systems from major vendors (Google, Amazon and Microsoft) using two datasets (MS-COCO and CityScapes). We generate over 12K erroneous OOD images (out of a total $\approx$ 56K OOD images), and find that up to 32% of classes are subject to fairness errors. Our approach also compares favorably with an approach focused on generating inputs within the distribution and improves the discovery of fairness errors by up to 131.48%.

TABLE I: Outline of DISTROFAIR: Inclusion errors (Inc.) are highlighted in blue, exclusion errors (Ex.) are highlighted in red

Subject/ Mutation	Original Image	Detected Objects	Mutated Image	Detected Objects
GCP (Rotation) Person		Car: 2 Traffic Light: 2		Car: 3 (Inc.) Traffic Light: 1 (Ex.) Building: 1 (Inc.)

II. OVERVIEW

An illustrative example: Table I shows an example illustrating our OOD-image generation and the class-level error detection. Intuitively, given a dataset S and a model M, we capture the distribution of the classes in the dataset and generate OOD images that fall outside the previously seen distribution. In general, we characterize all class-level errors into the following two categories:

1) **Exclusion Errors:** Exclusion error means that some object from a given class was detected in the original image, but it is *not detected* in the corresponding OOD image.
2) **Inclusion Errors:** Inclusion error means that some object from a given class was *not detected* in the original image, but it is detected in the corresponding OOD image.

We also exclude any errors due to the mutated class (i.e., the *person* class). This is to eliminate the potential impact of bias in our experiments, as the mutated classes tend to be more erroneous than the the unmodified classes.

We then find the cumulative error count for each individual class. In the case of Table I, we would increment the inclusion error count for both the *car* class (=3-2) and the *building* (=1-0) class by 1. Similarly, the exclusion error count for the *traffic light* (=1-0) class would also be accumulated. We would then determine the overall error rate for each class. Classes exhibiting an error rate higher than mean error rate across all classes would then be tagged as being unfairly treated.

III. EXPERIMENT

A. Effectiveness

We demonstrate that *our test generation approach (*DISTROFAIR*) is effective in exposing class-level fairness violations*. In particular, DISTROFAIR found that up to 32% of classes are subject to unfairness across all datasets and programs. In addition, we observed that at least 21% (12104 out of 56544) of images generated by DISTROFAIR induce class-level group fairness violation.

B. Baseline Comparison

We also investigate whether the OOD image mutation compares favorably with an *in-distribution* (ID) approach. To this end, we assess the effectiveness of the insertion mutation operation in both ID and OOD contexts to quantify the degree to which OOD-based mutation contributes to the effectiveness of DISTROFAIR. Our results show that *OOD style mutation outperforms the ID-based mutation approach in revealing class-level group fairness violations*. Specifically, DISTROFAIR reveals up to 74% more class-level fairness violations than the baseline for exclusion errors). Additionally, we found that a developer is more than two times likely (up to 131%) to find class-level fairness errors with OOD mutations than ID. This suggests that our use of OOD-based mutation contributes significantly to the effectiveness of DISTROFAIR.

C. Semantic Validity

We conducted a *user study* to evaluate the *semantic validity* of the images generated by DISTROFAIR. We conducted the study with 81 participants and 60 images. Our user study results show that our mutation operations are (up to 91%) as realistic as real-world images and (up to 92%) likely to occur in real life. We also find that both benign and error-inducing images were seen as being similarly valid, realistic and likely to occur.

IV. CONCLUSION

In this poster, we introduce DISTROFAIR, a systematic approach to discover class-level fairness violations in image classification tasks. The crux of DISTROFAIR is OOD test generation, which is synergistically combined with semantic preserving mutation operations. We show that such an approach is highly effective in revealing class-level fairness violations and it significantly outperforms test generation within the distribution. Additionally, we show that our generated tests (OOD images) are 80% as realistic as real world images. We hope that our open source OOD testing platform unfolds new opportunities for simple, yet effective class-level fairness testing for a variety of ML software systems. We provide DISTROFAIR and our experimental data for easy reproducibility, reuse and scrutiny:

https://storage.googleapis.com/
icse-2023-distrofair-gen-images/DistroFair.zip

REFERENCES

[1] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

[2] Jian Wu, Victor S Sheng, Jing Zhang, Hua Li, Tetiana Dadakova, Christine Leon Swisher, Zhiming Cui, and Pengpeng Zhao. Multi-label active learning algorithms for image classification: Overview and future promise. *ACM Computing Surveys (CSUR)*, 53(2):1–35, 2020.