Design of a Bias-Aware AutoML Challenge

Gabriel Lauzzana, Romain Mussard, Ihsan Ullah, Isabelle Guyon

LISN/INRIA/CNRS, Université Paris Saclay, France

Abstract. While research on bias is a very trending topic in data science, most work focus on social bias based on "protected attributes", such as gender and race. However, many datasets, not involving human subjects, are plagued with various kinds of bias, the origin of which is not always known, and may include confounding bias and sampling bias. This often imposes that data scientists perform "detective work" to unravel causes of bias, followed by rigorous manual data curation. With the advent of fully automated machine learning (AutoML), one may wonder whether creating bias-robust learning machines is possible, to reduce the need for data curation and the possible risk of introducing further biases. In this context, we propose to the scientific community a bias-aware AutoML challenge, based on image classification tasks. We present the challenge design, data preparation, and baseline methods. For reproducibility our code is provided.[1]

Keywords: Bias in data, Computer Vision, AutoML challenge.

Introduction and Motivations

This work is part of the realization of a bias-aware Automated Machine Learning (AutoML) challenge, based on Meta-Album, a meta dataset to which we contributed as part of our M1 TER and M1 internship [5]. Research has shown that data used in machine learning and computer vision can be biased towards certain communities [1,3,2]. Image background alone can contain enough information to allow models to make good predictions on object recognition tasks [7], and classifiers are often found to rely on various "spurious" features, such as object texture, position, orientation, etc. In this research, in addition to social biases based on "protected attributes" defined by law (such as gender, age, ethnicity), we are interested in other types of biases, possibly introduced during data aquisition or preprocessing (*e.g.,* including lighting, angle of view, cropping), on which comparably little work has been done. Identifying sources of bias, in the absence of pre-defined protected attributes, is difficult in itself [6]. Ultimately, human curators are called to determine which variables are legitimate (core features) and which ones cause bias (spurious features). However, reducing human intervention is highly desirable to improve efficiency, in cases where no harm could be done. To tackle the problem in an AutoML manner, we propose an object classification challenge in which we invite the community to find automated machine learning solutions, which are robust to learning from biased data.

[1] The two first authors contributed equally. We would like to thank Dustin Carrión for help and support. This research is funded by INRIA, ANR Chair HUMANIA ANR-19-CHIA-0022, & TAILOR EU Horizon 2020 grant 952215.

Material and Methods

For the 40 currently available image datasets of Meta-Album, we are carving out small multi-class classification problems of less than 5 classes (each belonging to the same original dataset having ≥ 20 classes), with 20 examples for training, and 20 for testing. Meta-Album allows us to create thousands of such tasks for our benchmark. For each task, the *training set is biased* (see details below), but the *test set is unbiased*. The goal of the challenge participants is to supply a learning machine capable of making good predictions of test data labels (not provided to them), despite learning on biased data. In the spirit of AutoML, the participants must submit code to a challenge platform, for "blind testing". During a development phase, the participants are given "sample tasks" to practice. They can make multiple submissions to obtain performance feed-back on a leaderboard. Their code is tested on "validation tasks" distinct from the "sample tasks". At then end of that phase, their last code submission is evaluated on yet different tasks (test tasks), to determine the final ranking.

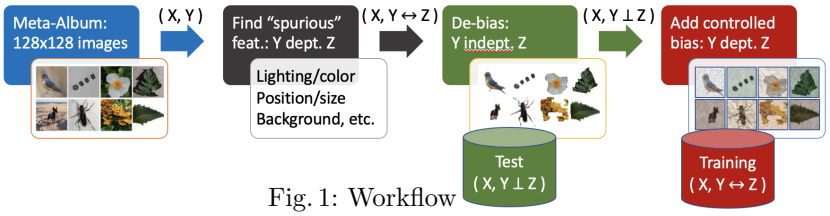

Fig. 1: Workflow

To supply a *biased training set* and an *unbiased test set*, we built a workflow (Fig. 1). Original datasets from Meta-Album (Blue box), consisting of (X, Y) pairs (image, class-label), are analyzed to identify bias (Gray box), performing steps including separating background from foreground, and extracting candidate "spurious" features Z such as texture, color, luminosity, position, size, etc. Truly spurious Zs are dependent (*i.e.,* predictive) of Y. For example, if $Z = 1$ means high luminosity and $Z = -1$ low luminosity, and if $Y = 1$ are dogs and $Y = -1$ cats, completely biased data are such that $Y = Z$ (*e.g.,* dogs captured outdoors and cats indoors). For each dataset, we retain at most 1 or 2 most significantly spurious features. Our next step (green box) is to remove detected bias. This is achieved in various ways by normalization, calibration, removal of background, resizing objects, resampling data, etc. After de-biasing spurious dependencies between Z and Y should be gone. From de-biased data we then create test sets. Finally (red box), we add bias again to remaining samples, but in a *controlled manner*, to create training data. We re-introduce a dependency between Z and Y. For instance, in our luminosity example, we could create a biased dataset with 90% images with $Y = Z$ and 10% with $Y = -Z$. To simplify our design, we only consider discretized Z. This allows us to control bias by data sampling, *i.e.,* by controlling the fraction of images in each group defined by a pair (Z, Y), as in our luminosity example. To assess model performance, we use "worst group accuracy", *i.e.,* the worst percentage of accurate predictions in each group (Z, Y), forcing participants to get rid of bias in every group.

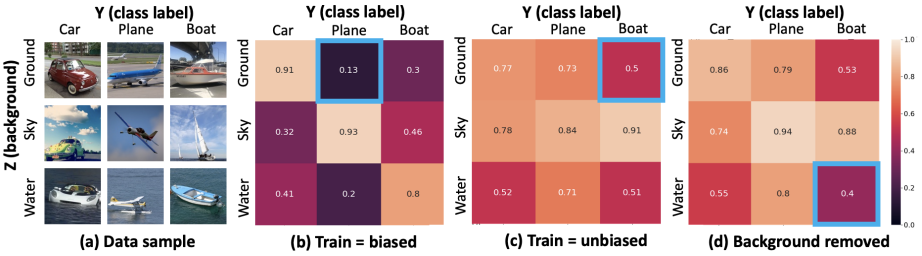

Fig. 2: **Test accuracy** in each group (Z, Y) (lighter is better): (b) Train on biased data (more examples on diagonal groups); (c) Train on group-balanced data; (d) Same as (a), but image backgrounds removed. Worst group accuracies highlighted in blue.

Results

To illustrate our design, we created a didactic dataset with $Y \in \{car, plane, boat\}$ and Z $(background) \in \{water, sky, ground\}$. Fig. 2.a shows example images in each group. We sampled 100 times a biased training set with 20 examples in each dominant group (diagonal elements) and a disjoint balanced test set with 7 example in each group. We trained the last layer of a pre-trained AlexNet [4] for 15 epochs. Fig. 2.b shows average test accuracies in each group. As expected, the classifier is fooled into using the "spurious feature" (background), leading to far *better diagonal group accuracies* than off-diagonal group accuracies. We then supplied the learning machine with balanced training data (distributed similarly as test data). Fig. 2.c shows, that worst group accuracy improves. Finally, we investigated whether a classifier trained on biased data can be immune to bias, if removing background is performed as a pre-processing. Fig. 2.d shows that this is indeed the case. These results are summarized in supplemental Fig. 3.

Conclusion

In preparation for a bias-aware AutoML challenge, we are preparing biased data in a controlled manner, by introducing groups, correponding to an association between original class labels and clusters of spurious feature values. When presenting a classifier with biased data, it is fooled into using spurious features as a predictors (for example using the sky background to classify objects as airplanes). Then, when presented with an unbiased test set, worst group accuracy drops drammatically. We aim at building a challenge with multiple types of bias (color, position, texture, etc.), with many small multi-class classification tasks, each with training data biased differently, and unbiased test data.

1. A. Abid, M. Farooqi, and J. Zou. Persistent anti-muslim bias in large language models. *ArXiv*, 2101.05783, 2021.
2. M. Brandao. Age and gender bias in pedestrian detection algorithms, 2019.
3. A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. 2017.
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012.
5. I. Ullah, D. Carrion, et al. Meta-album: Multi-domain meta-dataset for few-shot image classification. In *Submitted to NeurIPS D & B track*, 2022.
6. A. Wang, A. Liu, and otherd. REVISE: A tool for measuring and mitigating bias in visual datasets. *arXiv*, 2004.07999, 2021.
7. K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *ArXiv*, 2006.09994, 2020.

Supplemental material

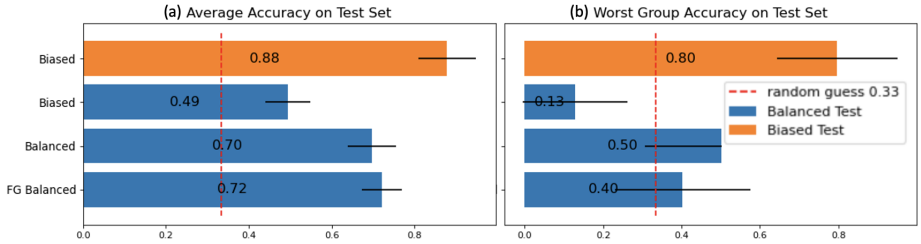

Fig. 3: **Comparision of results.** The top (orange) bar represents deceivingly good results, obtained when training and testing on biased data. In this case, only dominant groups (car, ground), (sky, plane), and (water, boat) are represented in both training and test data. All other (blue) bars are obtained with a group-balanced test data. The top blue bar (worst result) is obtained when the classifier is trained directly with biased data. The two bottom blue bars present the results of two simple baseline methods that provide some robustness against bias. The first one consists in training on group-balanced data (only possible if samples are available in sufficient number in each group). The second one consists in removing the background as preprocessing. The error bar is the standard deviation over 100 repeat experiments.