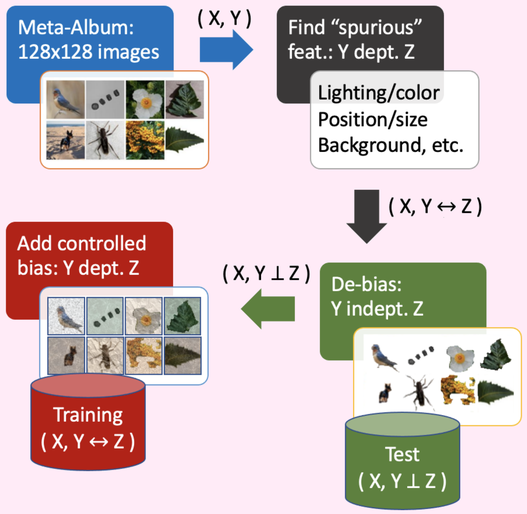
Designing a bias-aware autoML challenge

Gabriel Lauzzana, Romain Mussard, Ihsan Ullah, Isabelle Guyon
LISN/INRIA/CNRS, Université Paris Saclay

Introduction

- Classifiers are often found to rely on various "spurious" features.
- Most research on bias focus on social bias based on "protected attributes"
- Identifying sources of bias without pre-defined protected attributes is difficult.
- To tackle the problem we propose a bias-aware Automated Machine Learning challenge.

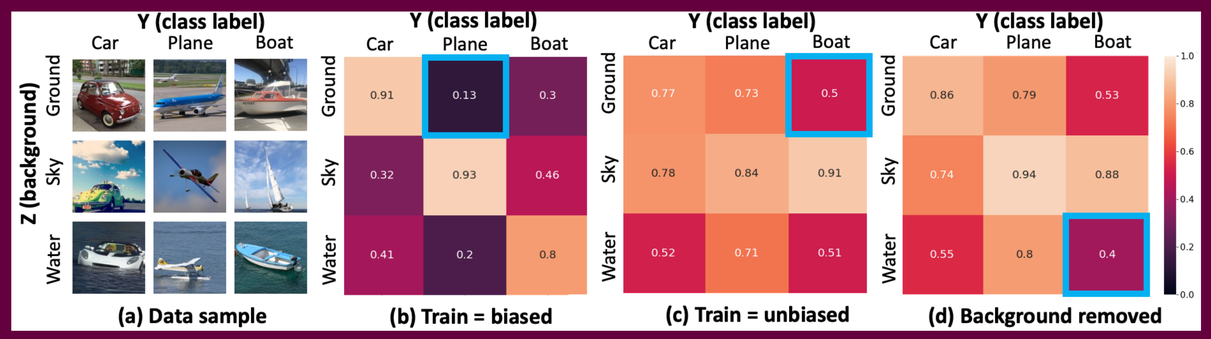
Methods



We take the original datasets in the Meta-Album, identify candidate "spurious" features Z (bias) and remove the detected bias to create test sets. Then we add bias in a *controlled manner* by re-introducing a dependency between Z and Y to create training data. Finally, we evaluate the performance of the model with the "worst group accuracy".

We **artificially biased** our dataset creating **disproportionate groups** defined by the association of their **class** and **spurious features**

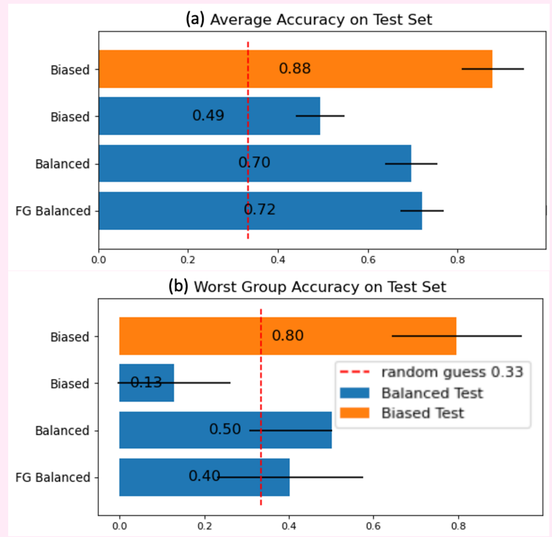
When training a classifier on this data, it is fooled into using **spurious features** as predictors, lowering the **accuracy** of the **least represented groups**



Average Test Accuracy of each group : 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.

Take a picture to download the full paper

Results



- We created a didactic dataset with $Y \in \{car, plane, boat\}$ and $Z\,(background) \in \{water, sky, ground\}$.
- 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 examples in each group.
- The classifier is fooled into using the "spurious feature" (background).
- This leads to far *better diagonal group accuracies* than off-diagonal group accuracies.
- Training with balanced data improves the worst group accuracy.
- Removing background as a pre-processing also improves it.

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE