Mitigating bias with Targeted Data Augmentations

Agnieszka Mikołajczyk, Michał Grochowski
agnieszka.mikolajczyk@pg.edu.pl, michal.grochowski@pg.edu.pl
Gdańsk University of Technology

Abstract

Data augmentation is used to obtain more training samples without the need for manual gathering or annotation of new ones. Usually, the main goal of the process is to improve the training and, as an effect, increase the robustness and efficiency of the model by decreasing the distribution shift between the training data and real-world examples. However, in many cases, the gap gets lower only between the train and the test data that originates from similar distribution, which might not correspond to the real-world cases. If both training and test data are biased, we might have a seemingly high accuracy effect, which is difficult to notice and mitigate.

The paper shows that properly identified and measured bias can be successfully mitigated with targeted data augmentation. We identify bias with the global explainability method and measure it by applying the counterfactual bias insertion method. Then, we use targeted data augmentation during the training: we randomly modify samples by inserting biases with the set of probabilities. After retraining on the modified data, the impact of the bias is measured again.

We show that biases are significantly reduced after training with targeted data augmentation, i.e., instances augmented with a black frame, hair, or ruler marks to skin lesion images improved up to seven times in terms of responses to bias than without such augmentation.

Introduction

Data augmentation (DA) finds its application in any deep learning-based system. However, in almost all cases, data augmentation increases the models' efficiency in terms of standard evaluation metrics such as accuracy, precision, or recall. In the paper, we propose using data augmentation as a valuable tool that helps mitigate detected biases in data for classification. By 'bias in data,' we mainly refer to four common data biases in Machine Learning: *observer bias* which might appear when annotators use personal opinion to label data; *sampling bias* when data is acquired in such away that not all samples have the same sampling probability ; *data handling bias* when the manner in which data is handled distort the classifier's output; and *instrument bias* meaning imperfections in the instrument or method used to collect the data.

Bias is an often problem that can be observed in various areas of deep learning, including NLP and computer vision. The paper focuses on the skin lesion case study, a classification of dermoscopy images of skin lesions into benign and malignant.

Our method shows a significant drop in bias measure, more precisely two to seven times fewer images switched classes after the training preceded by TDA, without a significant increase in error rate.

We define our main contribution as a proposition on coupling augmentation with Explainable Artificial Intelligence, resulting in a synergistic effect in reducing flows in machine learning algorithms. We propose to insert biases into the training data instead of removing them.

Experiments

The proposed Targeted Data Augmentation (TDA) is divided into identification and training with data augmentation.

Step 1. Identify bias

Detect bias. To detect bias, we use Global Explanations for Bias identification (GEBI) method to detect possible biases in data. GEBI uses local explainability methods that are analyzed in a semi-supervised way. First, the attribution maps for all images of interest are generated. Both attribution maps and input images are preprocessed. Next, GEBI suggests reducing the dimensionality of images and attribution maps with Isomap reduction. Achieved vectors of attribution maps and corresponding instances are concatenated, resulting in a small representation; then, resulting representations are clustered. Achieved clusters show models' prediction strategies, including potential flows in model reasoning.

Measure bias. Then, we measure the bias with counterfactual bias insertion method. We added black frames, ruler marks, and hair to all of the images and measured how the prediction changed. In an ideal classifier, a black frame or a ruler mark on the image should not change a prediction. The artifacts (biases) were added to images even when they were already in place. Our experiments showed that adding those artifacts is indeed, a biasing factor.

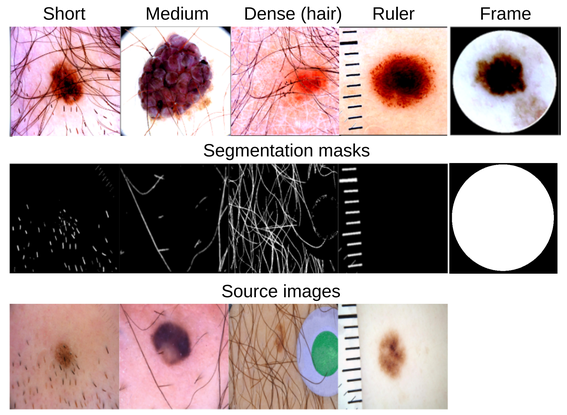
Step 2. Train with Targeted Data Augmentation

We can use identified biases to fuel targeted data augmentation methods. In our case, we successfully augment skin lesion data by randomly adding black frames, hair (dense, short, and medium), and ruler marks to the images during the training.

Counterfactual bias evaluation.
We evaluate how many instances changed predicted class after adding the bias to data: we use the *switched* metric. It simply measures how many instances changed predicted class after bias insertion. Additionally, we use $mean_{change}$ and $median_{change}$, which means measuring the difference in models' output after a bias insertion. Higher rates, mean a higher risk of giving predictions based on wrong premises. Such unwanted cases call for the need for targeted data augmentation.
Our experiments show that using targeted data augmentations resulted in lower mean and median prediction' changes after inserting the bias (see Table 1. Results). Additionally, the number of predictions that switched classes after inserting a bias significantly dropped in all cases.
Interestingly, we achieved the best results for frame insertion. In the case of the next group: hair and ruler marks, the result depends on the type of augmentation. Usually, the worst scenario is in the case of visible dense hair. Dense hair covers a significant amount of skin lesions, which makes evaluation difficult even for dermatologists and oncologists. However, we observe that the drop in terms of switched predictions is lower than in other cases yet bigger than in the case of short hair which is similar to black and brown dots that might indicate malignant skin lesions.

Performance evaluation. We use three general metric to evaluate the model's performance: f1 score, precision and recall. Each score is calculated for both original, unmodified data, and data with inserted bias (augmented data). We use $recall_{org}$, $precision\_{org}$, $f1_{org}$ measure to evaluate how well algorithm performs on the original, unmodified data.
However, the goal of Targeted Data Augmentation was not to achieve the best performance possible, but to be more robust to certain biases. Ideally, $f1_{aug}$ should be the same as $f1_{org}$, meaning that adding a bias to data does not change its performance. Higher differences between $f1_{aug}$ and $f1_{org}$ mean higher vulnerability to inserted bias. The $f1_{mean}$ is a mean of $f1_{org}$ and $f1_{aug}$, hence it shows how well model performs on both original and modified data.
In all cases, the $f1_{org}$ score significantly decreases after inserting a black frame, hair, or ruler mark to the image ($f1_{aug}$). Applying the targeted data augmentation makes the model more robust to such changes, even though augmentation used to test the model was not used during the training.

Example augmentations



Conclusion

We have noticed a great difference between frame and hair and rule augmentation. We suspect that the difference in the hair and ruler augmentation comes from the same training. Hence, the distribution of subgroups (frame, short, medium, dense) of all segmentation masks used to augment data may impact the final result. In the evaluation phase, those augmentations are artificially divided into that subgroups, while the training is randomly sampled from an unevenly distributed set of augmentations (segmentation masks).
In our future research, we plan to manually assign all of the segmentation masks into those groups and apply them separately into the framework. Moreover, we plan to apply all of the augmentations altogether during the training and test the effect of such massive augmentation on the results.
Additionally, a very interesting would be to evaluate the correlation between a given class and selected artifacts, i.e., by manually assigning all test images into multi-label classes: short, medium, dense, ruler, and frame. Then, such statistics could contribute to additional insight and understanding of the data and models.

Acknowledgements

The research reported in this publication was supported by Polish National Science Centre (Grant Preludium No: UMO-2019/35/N/ST6/04052). The authors wish to express their thanks for the support.

Table 1. Results

Augmentation	p	mean	median	s_{all}	s_{to_en}	s_{to_mal}	$f1_{org}$	$f1_{aug}$	$f1_{mean}$	$recall_{org}$	$recall_{aug}$	$precision_{org}$	$precision_{aug}$
frame	0	4.66%	0.73%	208	38	170	66.15%	54.69%	60.42%	58.47%	58.20%	76.16%	51.57%
	0.25	**0.96%**	**0.00%**	**44**	**33**	**11**	**64.93%**	**61.73%**	**63.33%**	**51.09%**	**46.72%**	**89.05%**	**90.96%**
	0.5	0.92%	0.01%	48	34	14	62.13%	58.59%	60.36%	48.63%	44.26%	85.99%	86.63%
	0.75	1.05%	0.02%	39	28	11	62.50%	61.54%	62.02%	49.18%	46.99%	85.71%	89.12%
	1	1.44%	0.04%	59	25	34	62.44%	61.15%	61.80%	49.73%	49.45%	83.87%	80.09%
short	0	1.39%	0.03%	71	41	30	66.15%	62.89%	64.52%	58.47%	54.64%	76.16%	74.07%
	0.25	1.21%	0.03%	59	40	19	65.02%	64.96%	64.99%	57.38%	55.46%	75.00%	78.38%
	0.5	0.87%	0.02%	41	20	21	62.61%	59.80%	61.21%	50.55%	48.36%	82.22%	78.32%
	0.75	1.00%	0.03%	44	32	12	66.35%	65.27%	65.81%	57.92%	55.19%	77.66%	79.84%
	1	**0.68%**	**0.01%**	**40**	**22**	**18**	**67.43%**	**64.57%**	**66.00%**	**56.01%**	**53.28%**	**84.71%**	**81.93%**
medium	0	1.75%	0.03%	86	59	27	66.15%	63.74%	64.95%	58.47%	53.55%	76.16%	78.71%
	0.25	1.53%	0.04%	85	33	52	65.02%	63.76%	64.39%	57.38%	57.92%	75.00%	70.90%
	0.5	1.03%	0.03%	43	22	21	62.61%	61.36%	61.99%	50.55%	49.45%	82.22%	80.80%
	0.75	1.19%	0.04%	60	40	20	66.35%	63.97%	65.16%	57.92%	54.10%	77.66%	78.26%
	1	**0.84%**	**0.01%**	**38**	**17**	**21**	**67.43%**	**65.36%**	**66.40%**	**56.01%**	**54.64%**	**84.71%**	**81.30%**
dense	0	2.69%	0.09%	180	142	38	66.15%	50.83%	58.49%	58.47%	37.70%	76.16%	77.97%
	0.25	3.46%	0.30%	164	78	86	65.02%	59.33%	62.18%	57.38%	53.01%	75.00%	67.36%
	0.5	2.34%	0.14%	124	72	52	62.61%	52.89%	57.75%	50.55%	41.26%	82.22%	73.66%
	0.75	2.46%	0.16%	124	92	32	66.35%	57.34%	61.85%	57.92%	45.36%	77.66%	77.93%
	1	**1.80%**	**0.06%**	**85**	**71**	**14**	**67.43%**	**56.99%**	**62.21%**	**56.01%**	**42.90%**	**84.71%**	**84.86%**
ruler	0	1.37%	0.02%	78	67	11	66.15%	61.25%	63.70%	58.47%	49.45%	76.16%	80.44%
	0.25	0.70%	0.01%	31	21	10	65.02%	65.51%	65.27%	57.38%	56.83%	75.00%	77.32%
	0.5	0.47%	0.01%	18	13	5	62.61%	61.41%	62.01%	50.55%	48.91%	82.22%	82.49%
	0.75	0.58%	0.01%	30	24	6	66.35%	65.38%	65.87%	57.92%	55.46%	77.66%	79.61%
	1	**0.40%**	**0.00%**	**18**	**12**	**6**	**67.43%**	**66.45%**	**66.94%**	**56.01%**	**54.64%**	**84.71%**	**84.75%**