"Unbiased Look at Dataset Bias"

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1521–1528, Washington, DC, USA, 2011. IEEE Computer Society

October 29, 2018; rev. May 7, 2020

Fred Guth

Summary

Purpose The access to large image datasets has been a key driver on the recent progress on Computer Vision tasks like object detection or face recognition. Despite its important role and the authors best efforts, all datasets evaluated by the paper presented strong built-in bias and, in a way, failed in the objective of representing the real world.

The basic question that drives this paper is: *Are the datasets measuring the expected performance in the real world?* Computer vision datasets have become closed worlds unto themselves and achieving a good benchmark in one, just means you are overfitting your model to that domain.

The goals of this paper are:

1. show how bias sneaks into datasets and affects tasks performance;
2. raise awareness to this important issue that had been neglected.

Claims The authors claim that:

- due to built-in biases datasets are failing in the general goal of representing the real world;
- some datasets like Caltech and MSRC are not helpful anymore and should be avoided;
- even automatically collected internet images present bias, as most results of searches bring photos where the searched term is occupying the center/focus of the picture.
- rich and unbiased negative set is important to classifier performance;

Experiments and Conclusions The authors used the SOTA object detection of that time, 2011 (no convolutional network). To minimize *selection bias*, they picked 2 objects common to all datasets, "car" and "person". Each classifier was trained with 500 positive and 2000 negative for the classification task and 100 positive and 1000 negative for the detection task for each dataset. Test was performed with 50 positive and 1000 negative examples for classification and 10 positive and 20000 negative for detection. For testing, each classifier was run 20 times and results were averaged. The results were depressing, little generalization cross datasets as shown in the table .

To evaluate relative bias in negative sets of the datasets, they created a superset of negatives of all datasets and, for each dataset, trained the classifier with positive and negative samples from its own dataset and tested with positives of the dataset and negatives from the superset. The number of samples obeyed the previous experiment protocol. There was a significant loss in performance, where negatives from other datasets were classified as positives. ImageNet, Caltech and MSRC did not show a drop in results for different reasons. ImageNet seems to have a large variability of negatives, the others just seem to be too easy.

Evaluation

The authors raised awareness for an important problem. With so much hype on the "success" of Deep Learning to all sorts of Computer Vision tasks, it is easy to forget that even good and large datasets like ImageNet present bias and fail to represent the real world. Consequently, deep learning research and

Table 1. Cross-dataset generalization. Object detection and classification performance (AP) for "car" and "person" when training on on dataset (rows) and testing on another (columns), i.e. each row is: training on one dataset and testing on all the others. "Self" refers t training and testing on the same dataset (same as diagonal), and "Mean Others" refers to averaging performance on all except self.

task	Train on: \ Test on:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Self	Mean others	Percent drop
"car" classification	SUN09	**28.2**	29.5	16.3	14.6	16.9	21.9	28.2	19.8	**30%**
	LabelMe	14.7	**34.0**	16.7	22.9	43.6	24.5	34.0	24.5	**28%**
	PASCAL	10.1	25.5	**35.2**	43.9	44.2	39.4	35.2	32.6	**7%**
	ImageNet	11.4	29.6	36.0	**57.4**	52.3	42.7	57.4	34.4	**40%**
	Caltech101	7.5	31.1	19.5	33.1	**96.9**	42.1	96.9	26.7	**73%**
	MSRC	9.3	27.0	24.9	32.6	40.3	**68.4**	68.4	26.8	**61%**
	Mean others	10.6	28.5	22.7	29.4	39.4	34.1	53.4	27.5	48%
"car" detection	SUN09	**69.8**	50.7	42.2	42.6	54.7	69.4	69.8	51.9	**26%**
	LabelMe	61.8	**67.6**	40.8	38.5	53.4	67.0	67.6	52.3	**23%**
	PASCAL	55.8	55.2	**62.1**	56.8	54.2	74.8	62.1	59.4	**4%**
	ImageNet	43.9	31.8	46.9	**60.7**	59.3	67.8	60.7	49.9	**18%**
	Caltech101	20.2	18.8	11.0	31.4	**100**	29.3	100	22.2	**78%**
	MSRC	28.6	17.1	32.3	21.5	67.7	**74.3**	74.3	33.4	**55%**
	Mean others	42.0	34.7	34.6	38.2	57.9	61.7	72.4	44.8	**48%**
"person" classification	SUN09	**16.1**	11.8	14.0	7.9	6.8	23.5	16.1	12.8	**20%**
	LabelMe	11.0	**26.6**	7.5	6.3	8.4	24.3	26.6	11.5	**57%**
	PASCAL	11.9	11.1	**20.7**	13.6	48.3	50.5	20.7	27.1	**-31%**
	ImageNet	8.9	11.1	11.8	**20.7**	76.7	61.0	20.7	33.9	**-63%**
	Caltech101	7.6	11.8	17.3	22.5	**99.6**	65.8	99.6	25.0	**75%**
	MSRC	9.4	15.5	15.3	15.3	93.4	**78.4**	78.4	29.8	**62%**
	Mean others	9.8	12.3	13.2	13.1	46.7	45.0	43.7	23.4	**47%**
"person" detection	SUN09	**69.6**	56.8	37.9	45.7	52.1	72.7	69.6	53.0	**24%**
	LabelMe	58.9	**66.6**	38.4	43.1	57.9	68.9	66.6	53.4	**20%**
	PASCAL	56.0	55.6	**56.3**	55.6	56.8	74.8	56.3	59.8	**-6%**
	ImageNet	48.8	39.0	40.1	**59.6**	53.2	70.7	59.6	50.4	**15%**
	Caltech101	24.6	18.1	12.4	26.6	**100**	31.6	100	22.7	**77%**
	MSRC	33.8	18.2	30.9	20.8	69.5	**74.7**	74.7	34.6	**54%**
	Mean others	44.4	37.5	31.9	38.4	57.9	63.7	71.1	45.6	**36%**

practitioners, includind us, treat every new real world application data as a new domain that needs to be fine-tunned, alas our best datasets are far to be good enough for general usage. Although being impossible to build a dataset without bias, the problems pointed were pertinent.

They showed that even in the presence of strong bias, Caltech and MSRC were easy to be classified, therefore they do not provide much value nowadays (2011).

They have clearly experimentaly proved their claims and a quick search on Google Scholar shows that they achieved their goal of raising awareness to the problem: not only the paper has more than 770 citations, but since its publication, several other papers on the subject of dataset bias were published.

Synthesis

For my research, I believe the paper helped me realize that we humans also have access to biased input, but are able to much more efficiently generalize our learnings. This is an indication that even in areas where it seems that our SOTA have great results, we are quite far from what is possible.

The use of unbalanced positive and negative datasets by an order of magnitude made me wonder my own dataset building choices and negative set bias.

Another important point is that the area has seen major improvements due to better datasets and dataset engineering should be better researched.

References

[1] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1521–1528, Washington, DC, USA, 2011. IEEE Computer Society.