Undoing the Damage of Dataset Bias

Aditya Khosla\textsuperscript{1}, Tinghui Zhou\textsuperscript{2}, Tomasz Malisiewicz\textsuperscript{1}, Alexei A. Efros\textsuperscript{2}, and Antonio Torralba\textsuperscript{1}

\textsuperscript{1} Massachusetts Institute of Technology
\{khosla,tomasz,torralba\}@csail.mit.edu
\textsuperscript{2} Carnegie Mellon University
\{tinghuiz,efros\}@cs.cmu.edu

Abstract. The presence of bias in existing object recognition datasets is now well-known in the computer vision community. While it remains in question whether creating an unbiased dataset is possible given limited resources, in this work we propose a discriminative framework that directly exploits dataset bias during training. In particular, our model learns two sets of weights: (1) bias vectors associated with each individual dataset, and (2) visual world weights that are common to all datasets, which are learned by undoing the associated bias from each dataset. The visual world weights are expected to be our best possible approximation to the object model trained on an unbiased dataset, and thus tend to have good generalization ability. We demonstrate the effectiveness of our model by applying the learned weights to a novel, unseen dataset, and report superior results for both classification and detection tasks compared to a classical SVM that does not account for the presence of bias. Overall, we find that it is beneficial to explicitly account for bias when combining multiple datasets.

1 Introduction

Recent progress in object recognition has been largely built upon efforts to create large-scale, real-world image datasets \cite{1-3}. Such datasets have been widely adopted by the computer vision community for both training and evaluating recognition systems. An important question recently explored by Torralba and Efros \cite{4}, and earlier by Ponce et al \cite{5}, is whether these datasets are representative of the visual world, or in other words, unbiased. Unfortunately, experiments in \cite{4, 5} strongly suggest the existence of various types of bias (e.g. selection bias, capture bias, and negative set bias) in popular image datasets.

In the ideal world, more data should lead to better generalization ability but as shown in \cite{4}, it is not necessarily the case; performance on the test set of a particular dataset often decreases when the training data is augmented with data from other datasets. This is surprising as in most machine learning problems, a model trained with more examples is expected to better characterize the input space of the given task, and thus yield better performance. The fact that this common belief does not hold in object recognition suggests that the input space of each image dataset is dramatically different, i.e. the datasets are biased.
Our key observation for undoing the dataset bias is that despite the presence of different biases in different datasets, images in each dataset are sampled from a common *visual world* (shown in Figure 1). In other words, different image datasets are biased samples of a more general dataset—the visual world. We would expect that an object model trained on the visual world would have the best generalization ability, but it is conceivably very difficult, if not impossible, to create such a dataset.

In this paper, we propose a discriminative framework that explicitly defines a bias associated with each dataset and attempts to approximate the weights for the visual world by undoing the bias from each dataset (shown in Figure 1). Specifically, our model is a max-margin framework that takes the originating dataset of each example into account during training. We assume that the bias of all examples from a given dataset can be modeled using the same bias vector, and jointly learn a visual world weight vector together with the bias vector for each dataset by max-margin learning. In order to model both contextual bias and object-specific bias, we apply our algorithm to the tasks of classification and detection, showing promising results in both domains.

**Fig. 1. Left: Sampling from the Visual World.** Image datasets are sampled from the same visual world. Each dataset we have collected, such as Caltech101, SUN or LabelMe, has a certain bias that is represented as $\Delta_i$, $i = 1, \ldots, n$. $D_{n+1}$ represents an independent test set that has not been seen by the model but is sampled from the same visual world. **Right: Overview of our algorithm.** We model the biases as additive linear vectors $\Delta_i$ for each dataset. Our goal is to learn a model for the visual world $w_{vw}$ which has good generalization ability. The dataset specific model $w_i$ tends to perform well on the corresponding dataset but not generalize well to novel datasets.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 presents the details of our model, including the problem formulation and the optimization algorithm. Experimental results that demonstrate the effectiveness of our model in both classification and detection settings are presented in Section 4. Section 5 concludes the paper with a summary of our contributions.