CS230

Dataset Bias Analysis on Autonomous Driving

Zheng Lyu
Department of Electrical Engineering
Stanford University
zhenglyu@stanford.edu

Abstract

Deep learning method is widely discussed recently in autonomous driving. Although a lot of models are proposed to be used, not too much attention has been paid on the performance variation caused by the dataset bias. In this project, two widely used dataset were used separately to train two popular models. After the training, we evaluated the performance based on three datasets. The huge difference on performance caused by the dataset indicates the existence of the dataset bias.

1 Introduction

Recently, a lot of approaches based on deep learning are proposed to solve the item detection problem that can be used in autonomous driving. Although many great models are claimed to be good candidates, not too much attention has been paid on the dataset bias. Typical dataset bias can be easily understood: all photos are taken during sunny day time but none is at night, rainy, foggy, or snowy days. Such bias can be either intended or unintended, but in any case it is expected to have an influence on the model performance after feeding the model with a biased dataset. In this work, we conduct an cross validation on three popular datasets used in autonomous driving: Berkeley DeepDrive (BDD), KITTI, and Apollo by using two classic model: SSD and R-CNN were trained. With the training model, the inputs are images from different datasets and by comparing the difference on the accuracy of item detection (e.g. cars and pedestrians). The performance difference can be served as an important indicator to determine whether the bias exists.

2 Related work

2.1 Bias Recognition

Several paper report the existence of the bias among various datasets [1-2]. Torralba *et al* presented a comparison study using a set of popular datasets, evaluated based on a number of criteria including: relative data bias, cross-dataset generalization, effects of closed-world assumption, and sample value. The result proved a rather surprising clear bias effect and suggested database evaluation protocol algorithm is necessary. In 2015, Tommasi *et al* proposed to verify the potential of the DeCAF features when facing the dataset bias problem. They conducted a series of analyses looking at how existing datasets differ among each other.

2.2 Bias Mitigation

Since it is confirmed that datasets have bias and biased dataset can lead to performance decay, some works discussed on how to mitigate such bias effect [3-5] developed unsupervised domain adaptation techniques to overcome the biases. Khosla *et al* proposed a discriminative framework

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

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. However, such method was not based on deep learning, which made the usefulness on robustness remain not confirmed. More recently, Zhang *et al* proposed to mitigating unwanted biases with adversarial learning, which is seemed to be the most promising approach to attack the dataset problem in autonomous driving field.

Due to the limited time, bias mitigation, though no doubt is necessary, is left to the future work and will not be discussed in this work. In the following sections, we will focus on discussing the data source and the process of conducting cross validation.

3 Dataset and Features

In this work we tried three different popular datasets that often used for autonomous driving research. The first one is KITTI [6], which contains more than 7400 images for six categories: city, residential, road, campus, person and calibration. The datasets are captured by driving around the mid-size city of Karlsruhe, in rural areas and on highways. Up to 15 cars and 30 pedestrians are visible per image.

The second dataset is BDD from Berkeley [7]. It has more than 100,000 images that have been used for road object detection, instance segmentation, lane markings and other drivable areas.

Another dataset is Apollo, in which the images are taken mainly in China. Due to the dataset only has 100 images for now, the model were not used for model training but only for the evaluation section.

To make the result more comparable, all the images are cropped with the dimension of 370 x 1224. Example images from those three datasets are shown if Fig 1.

Figure 1: Example images from: a) KITTI, b) BDD and c) Apollo.

4 Methods

To make sure the conclusion is trustful, we evaluate the dataset on two classic training models: SSD mobilenet v1 coco (SSD) and Faster R-CNN ResNet101 coco (ResNet-101). Fig 2. In both case, the convolutional feature layers are used to decrease in size progressively and allow predictions of detections at multiple scales. Each added feature layer can produce a fixed set of detection predictions using a set of convolutional filters. The loss function is defined as:

$$L(x, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g))$$

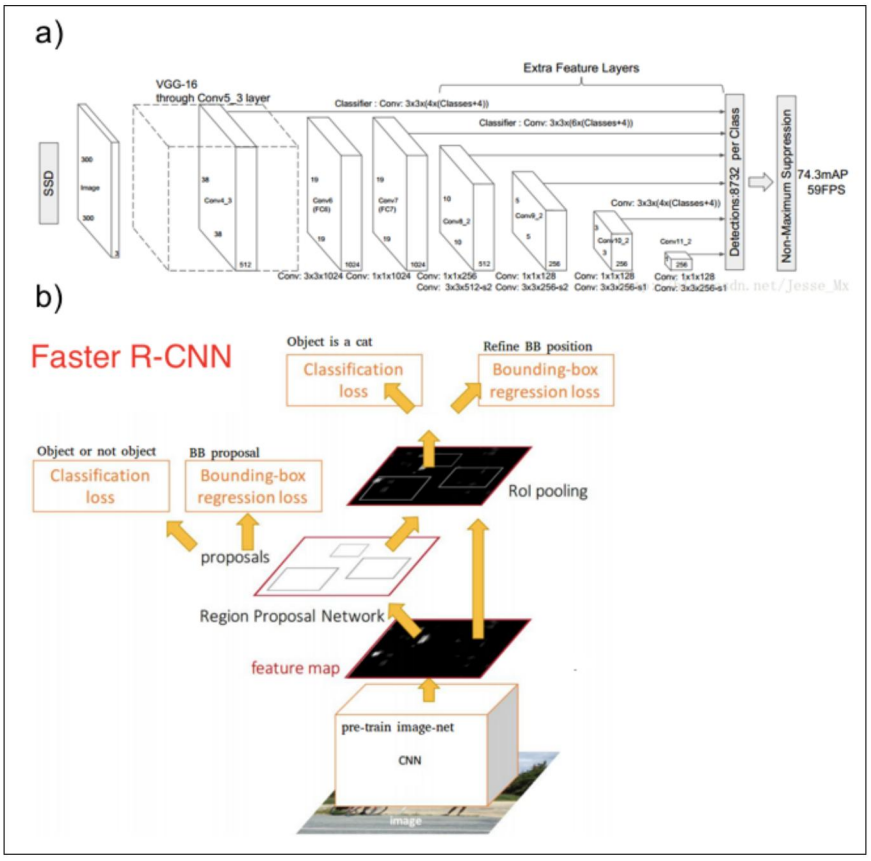

Figure 2: Model schematic of: a)SSD mobilenet v1 coco and b) Faster R-CNN ResNet101 coco.

where $x = (0, 1)$ is the indicator for matching between default box and the ground truth box. The localization loss is a Smooth L1 loss between the predicted box (l) and the ground truth box (g) parameters. More details can be obtained from [8].

5 Experiments/Results/Discussion

5.1 Training Process

We trained the SSD model with both KITTI and BDD dataset. The training results are illustrated in Fig 3. Noted for ResNet-101 model, we only trained the model on BDD as there is an pre-trained model from KITTI official website that can be directly used. For all situations, the models are trained for enough steps/epochs when the average accuracies are stabilized.

5.2 Testing Results

After confirming the model is trained, we conducted the cross validation on the training dataset. For this work, detection accuracy of cars and pedestrians are evaluated. The result are shown in Table 1 and Table 2.

As can be clearly seen, after SSD model was trained with KITTI, the performance has a large difference when being evaluated by KITTI, BDD and Apollo. Take the average accuracy of car detection as instance, the accuracy for KITTI evaluation is 69.7 %. However, when being evaluated by BDD and Apollo, the accuracy dropped to nearly on 10 %. The Faster R-CNN model has the same result, the accuracy dropped from over 90 % to 27.3 % and 47 % respectively. This is a strong indicator that the bias exists among different datasets. In comparison, the performance is much more stable for both models when trained with images from BDD dataset.

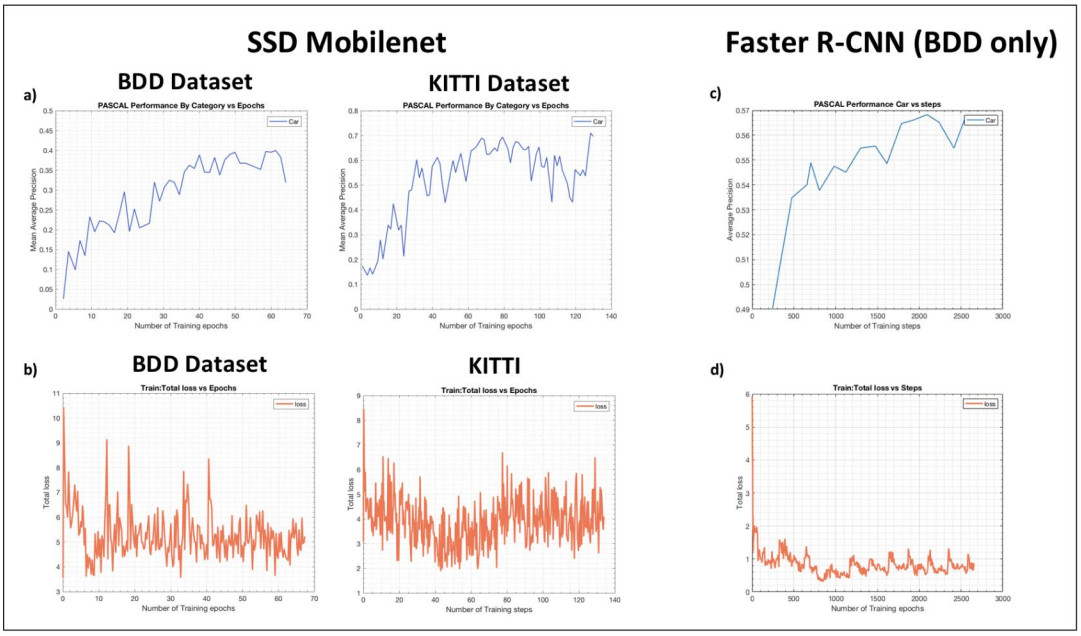

Figure 3: Training result: a) average accuracy for BDD and KITTI using SSD, b) loss function for BDD and KITTI using SSD, c) average accuracy for BDD using ResNet-101 and d) loss function for BDD using ResNet-101.

Table 1: Model: SSD Mobilenet

Train / Eval	KITTI 24,000 training steps			BDD 8500 images/(12,000 steps)		
	mAP	AP@Car	AP@Person	mAP	AP@Car	AP@Person
KITTI(1500)	0.517	0.697	0.337	0.312	0.505	0.1189
BDD(1500)	0.06	0.108	0.013	0.278	0.399	0.171
Apollo(100)	0.06	0.094	0.00001	0.213	0.417	0.01

Table 2: Model: Faster R-CNN

Train / Eval	KITTI From google model zoo with 800,008 training steps			BDD 8500 images/(6,000 steps)		
	mAP	AP@Car	AP@Person	mAP	AP@Car	AP@Person
KITTI(300)	0.829	0.906	0.751	0.535	0.621	0.448
BDD(300)	0.252	0.273	0.231	0.508	0.573	0.443
Apollo(100)	0.213	0.47	0.02	0.320	0.583	0.056

5.3 Result analysis

After comparing the test result, we concluded several factors that might be the reason for the performance difference. The factors and several examples are shown in Fig 4.

It is found that in BDD dataset, more situations are contained compared with KITTI. Most of the images of KITTI are taken during daily time on the road. It can be seen that the performance are comparable when evaluating the KITTI dataset. However, when scene like under bridge, having stop station (the example shown in Fig 4 (b)) or during night time, simply training on KITTI will give very unsatisfying result. As can be seen when evaluated with Apollo, the bus in the center cannot be detected with the model trained by KITTI.

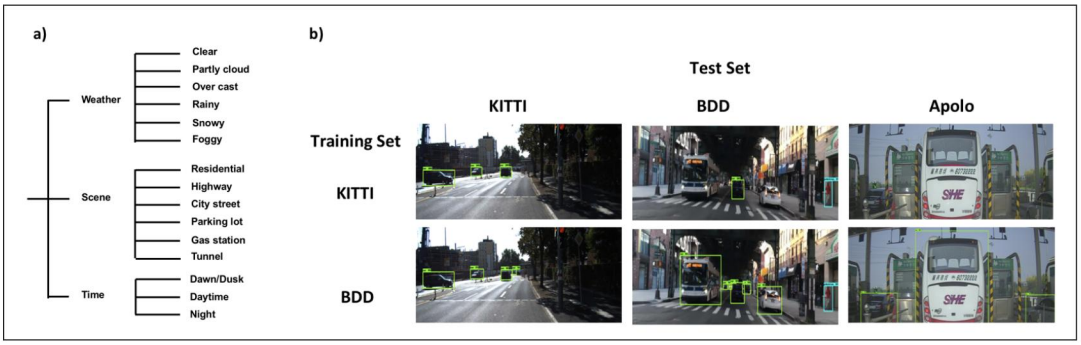

Figure 4: a) Possible factors that caused the performance variation and b) examples of test results

6 Conclusion/Future Work

In conclusion, we demonstrated the existence of bias in the widely used datasets. The conclusion is made with cross validation on three popular dataset, KITTI, BDD, and Apollo in which the images are taken in locations very far away from each other. Such effect should be taken carefully when training models and draw conclusions since the dataset might not be able to fully represent the situations on the road. The result also stress the importance to explore approach to achieve data generalization in order to fully eliminate the bias

For future work, we should develop approaches to confirm the factors that we concluded are reasonable. To achieve this, visualization could be a powerful tool to see if the target that is supposed to be detected is activated. Part of the work has been done on the activation function. However, due to critical bugs in Tensorflow, when loading graph, the deconvolution result cannot be obtained in short time. That part will be done in the future by either on an new version of Tensorflow or Pytorch. In the next step, approaches that are proposed to eliminate the bias damage should be applied to see if the bias can be reduced. Once the bias can be safely got rid of, it will have a great benefit on dataset generalization. In this way, the data generalization can make the dataset more widely used and more robust in a wider geometry range.

7 Contributions

The author would like to thank Zhenyi Liu, one of the group member in Prof. Brian Wandell's group for his help on discussion about the project idea and access on Google Cloud.

References

[1] Tommasi, Tatiana, et al. "A deeper look at dataset bias." Domain Adaptation in Computer Vision Applications. Springer, Cham, 2017. 37-55.

[2] Torralba, Antonio, and Alexei A. Efros. "Unbiased look at dataset bias." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[3] Gong, Boqing, Fei Sha, and Kristen Grauman. "Overcoming dataset bias: An unsupervised domain adaptation approach." NIPS Workshop on Large Scale Visual Recognition and Retrieval. Vol. 3, 2012.

[4] Khosla, Aditya, et al. "Undoing the damage of dataset bias." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.

[5] Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." arXiv preprint arXiv:1801.07593, 2018.

[6] Geiger, Andreas, et al. "Vision meets robotics: The KITTI dataset." The International Journal of Robotics Research 32.11 , 2013: 1231-1237.

[7] Yu, Fisher, et al. "BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling." arXiv preprint arXiv:1805.04687, 2018.

[8] Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision. Springer, Cham, 2016.