A First Look at Dataset Bias in License Plate Recognition

Rayson Laroca[1], Marcelo Santos[1], Valter Estevam[1,2],
Eduardo Luz[3], David Menotti[1]

[1]Federal University of Paraná, Curitiba, Brazil
[2]Federal Institute of Paraná, Irati, Brazil
[3]Federal University of Ouro Preto, Ouro Preto, Brazil

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Visão
Robótica
Imagem

Introduction

Is it possible to predict the dataset from which a license plate (LP) image belongs?

Introduction

Is it possible to predict the dataset from which a license plate (LP) image belongs?

- Initially, one may think that this task is fairly trivial;
  - Images collected in different regions, with different hardware, for different purposes, etc.

Introduction

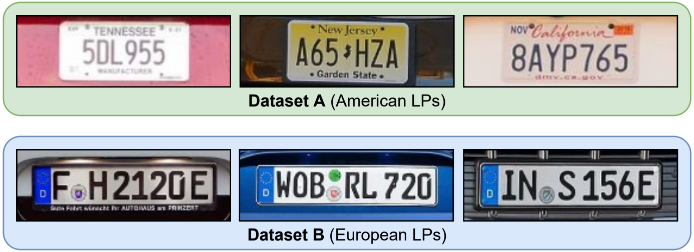
Is it possible to predict the dataset from which a license plate (LP) image belongs?

- Initially, one may think that this task is fairly trivial;
- On second thought, one may realize that it depends on the datasets we are comparing.

Introduction

Is it possible to predict the dataset from which a license plate (LP) image belongs?

- Initially, one may think that this task is fairly trivial;
- On second thought, one may realize that it depends on the datasets we are comparing.



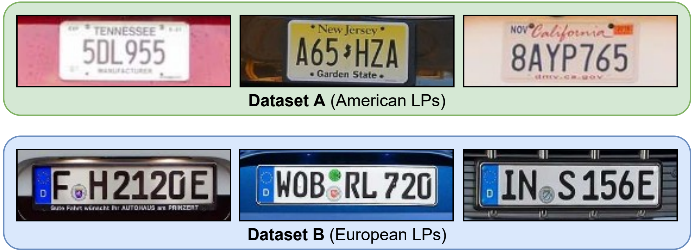
Dataset A (American LPs)

Dataset B (European LPs)

Introduction

Is it possible to predict the dataset from which a license plate (LP) image belongs?

- Initially, one may think that this task is fairly trivial;
- On second thought, one may realize that it depends on the datasets we are comparing.



Dataset A (American LPs)

Dataset B (European LPs)

In this case, it should be **quite straightforward** to distinguish which dataset each LP image belongs to due to the many characteristics LPs from the same region/layout share in common.

Research Question

Research Question

Beyond the LP layout, are there ***unique signatures (bias)*** *in each dataset that would enable identifying the source of an LP image?*

Name that Dataset!

Can you name the dataset to which each of these images belongs?

RodoSol-ALPR (ES): ___ , ___ , ___ , ___ SSIG-SegPlate (MG): ___ , ___ , ___ , ___

UFOP (MG): ___ , ___ , ___ , ___ UFPR-ALPR (PR): ___ , ___ , ___

Name that Dataset!

Can you name the dataset to which each of these images belongs?

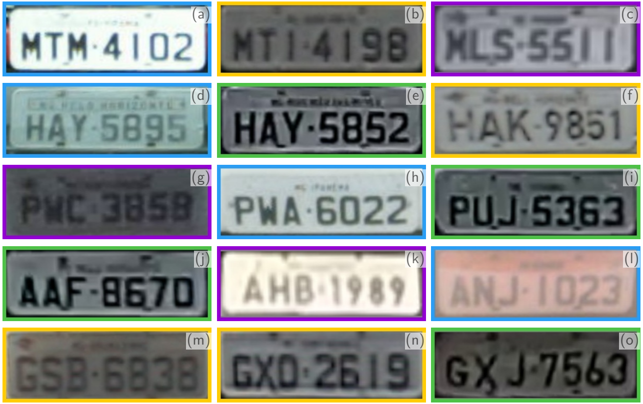
RodoSol-ALPR (ES): **(a)**, **(d)**, **(h)**, **(l)**

SSIG-SegPlate (MG): **(e)**, **(i)**, **(j)**, **(o)**

UFOP (MG): **(b)**, **(f)**, **(m)**, **(n)**

UFPR-ALPR (PR): **(c)**, **(g)**, **(k)**

Name that Dataset!



RodoSol-ALPR (ES): **(a)**, **(d)**, **(h)**, **(l)** SSIG-SegPlate (MG): **(e)**, **(i)**, **(j)**, **(o)**
UFOP (MG): **(b)**, **(f)**, **(m)**, **(n)** UFPR-ALPR (PR): **(c)**, **(g)**, **(k)**

- A shallow CNN (3 conv. layers) predicts the correct dataset in more than **95%** of cases[1].

[1](chance is 1/4 = 25%)

Experiments - Outline

1. Datasets;
2. Classification Model;
3. Results.

Experimental Setup - Datasets

The eight datasets used in our experiments.

Dataset	Year	LP Images	State / Province-City
UFOP	2011	244	Minas Gerais
ChineseLP	2012	400	Various
SSIG-SegPlate	2016	1,832	Minas Gerais
PKU	2017	2,024	Anhui-Tongling
UFPR-ALPR	2018	2,700	Paraná
CCPD	2020*	25,000†	Anhui-Hefei
PlatesMania-CN	2021	347	Various
RodoSol-ALPR	2022	4,765	Espírito Santo

* The CCPD dataset was introduced in 2018 and last updated in 2020.

† Following Liu et al. (2021), we used a reduced version of CCPD in our experiments.

- Many works in the literature are focused on LPs from Brazil and mainland China.

Experimental Setup - Chinese LPs

Some Chinese LPs from the datasets used in our experiments.
From top to bottom: CCPD, ChineseLP, PKU and PlatesMania-CN.

- The first character on each LP is a Chinese character representing the province in which the vehicle is affiliated. The second character is an English letter representing the city.

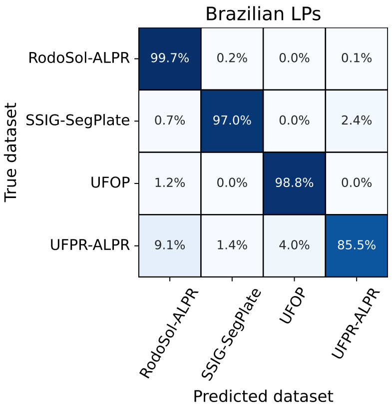
Experimental Setup - Classification Model

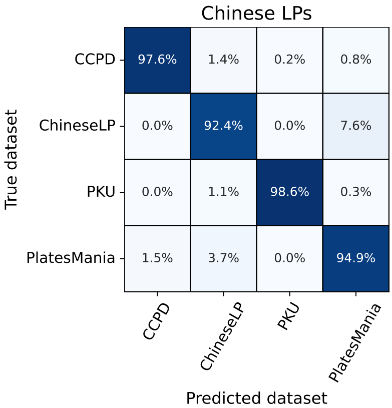
- We designed a lightweight CNN architecture called **DC-NET**.
  - It runs at ≈ **720 FPS** on an NVIDIA Quadro RTX 8000 GPU.

DC-NET's layers and hyperparameters.

#	Layer	Filters	Size / Stride	Input	Output
0	conv	16	$3 \times 3/1$	$192 \times 64 \times 3$	$192 \times 64 \times 16$
1	max		$2 \times 2/2$	$192 \times 64 \times 16$	$96 \times 32 \times 16$
2	conv	32	$3 \times 3/1$	$96 \times 32 \times 16$	$96 \times 32 \times 32$
3	max		$2 \times 2/2$	$96 \times 32 \times 32$	$48 \times 16 \times 32$
4	conv	64	$3 \times 3/1$	$48 \times 16 \times 32$	$48 \times 16 \times 64$
5	max		$2 \times 2/2$	$48 \times 16 \times 64$	$24 \times 8 \times 64$
6	flatten			$24 \times 8 \times 64$	12288
#	**Layer**		**Units**	**Input**	**Output**
7	dense		128	12288	128
8	dense		4	128	4

Results

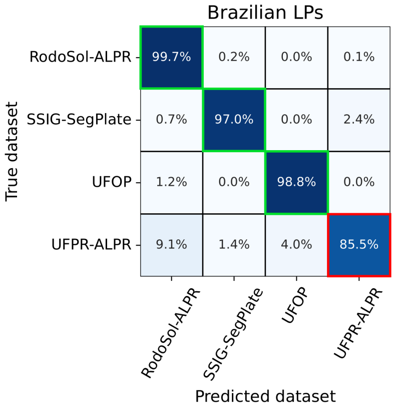


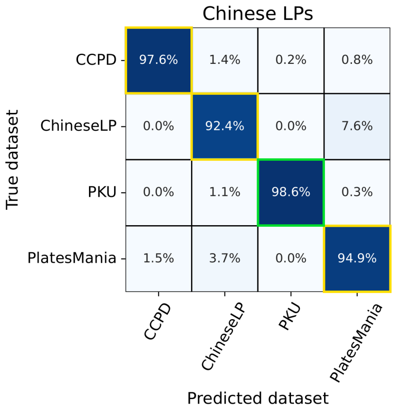


There is a **clearly pronounced diagonal** in both matrices, indicating that **each dataset does have a unique, identifiable "signature."**

The overall accuracy was **95.2%** for Brazilian LPs and **95.9%** for Chinese LPs.

Results





The DC-NET model is more successful in classifying LP images from the datasets acquired with **static** cameras than images from the datasets captured by **handheld** or **moving** cameras.

Results

- Images collected by **static cameras** have many characteristics in common, not just the background.
  - These similarities are probably present to some extent in the LP regions.

Results

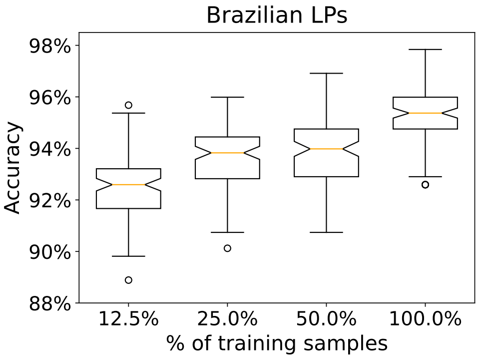
- Images collected by **static cameras** have many characteristics in common, not just the background.
  - These similarities are probably present to some extent in the LP regions.

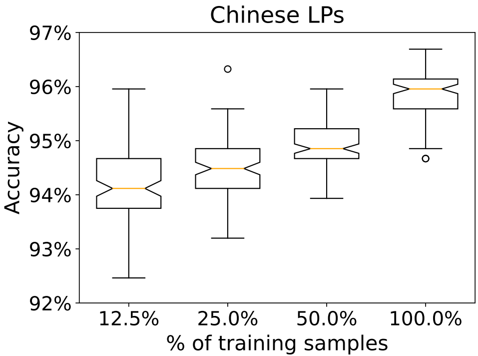
RodoSol-ALPR (MSE = 174)

RodoSol-ALPR (MSE = 407)

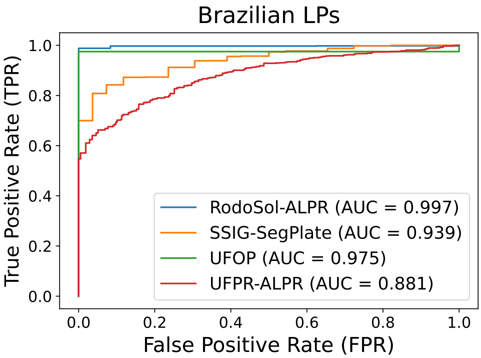
Results

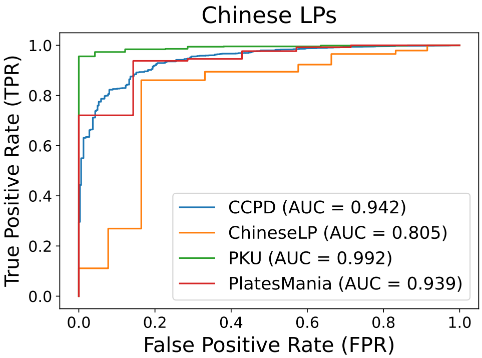
There are **no immediate signs of saturation**, i.e., the accuracy consistently improves as the size of the training set increases.





Results





The classifier predicts the source dataset of an LP image correctly with a significantly higher confidence value than when it predicts incorrectly[2].

[2] The mean confidence values for correctly classified Brazilian and Chinese LPs were 98.5% and 98.1%, respectively, while the mean confidence values for incorrectly classified Brazilian and Chinese LPs were 79.7% and 74.3%, respectively.

Discussion

Most LPR models are probably learning and exploiting such signatures to improve the results achieved in seen datasets **at the cost of losing generalization capability**.

Discussion

Most LPR models are probably learning and exploiting such signatures to improve the results achieved in seen datasets **at the cost of losing generalization capability**.

SSIG-SegPlate:

- It has **563** LP images with the letter **'O'** in the first position;

Discussion

Most LPR models are probably learning and exploiting such signatures to improve the results achieved in seen datasets **at the cost of losing generalization capability**.

SSIG-SegPlate:

- It has **563** LP images with the letter **‘O’** in the first position;

- It has **no** LP images with the letter **‘Q’** in the first position.

Discussion

Most LPR models are probably learning and exploiting such signatures to improve the results achieved in seen datasets **at the cost of losing generalization capability**.

SSIG-SegPlate:

- It has **563** LP images with the letter **'O'** in the first position;

- It has **no** LP images with the letter **'Q'** in the first position.

Taking this into account:

- An LPR model capable of identifying that a given LP image belongs to the SSIG-SegPlate dataset **may predict the letter 'O' as the first character** even if the character looks more like 'Q' than 'O' due to noise, shadows, or other factors.
  - However, the potentially high recognition rates achieved in the *SSIG-SegPlate* dataset would likely not be reached in unseen datasets.

Discussion

Probable causes of dataset bias in the LPR context:

- The **cameras** used to collect the images in each dataset;
- How the images were **stored** in different datasets;
  - e.g., the CCPD dataset contains highly compressed images, while most other datasets do not.
- How accurate the LP **corner annotations** are in different datasets.

Discussion

Probable causes of dataset bias in the LPR context:

- The **cameras** used to collect the images in each dataset;
- How the images were **stored** in different datasets;
  - e.g., the CCPD dataset contains highly compressed images, while most other datasets do not.
- How accurate the LP **corner annotations** are in different datasets.

Two initial ways to mitigate the dataset bias problem in LPR:

- Leveraging deep learning-based methods' high capability to visualize and understand how bias has crept into the datasets;
  - One technique that immediately comes to mind is Grad-CAM.
- To embrace the "wildness" of the **internet** to collect a large-scale dataset for LPR.
  - Multiple sources (e.g., multiple search engines and websites from various countries).

Conclusions

- The results showed that **each dataset does have a unique, identifiable signature**;
  - The source dataset of an LP image could be predicted with more than 95% accuracy;
  - We observed no evidence of saturation as more training data was added.

Conclusions

- The results showed that **each dataset does have a unique, identifiable signature**;
  - The source dataset of an LP image could be predicted with more than 95% accuracy;
  - We observed no evidence of saturation as more training data was added.
- Researchers should evaluate LPR models in **cross-dataset setups**;
  - A better indication of generalization, hence real-world performance, than within-dataset ones.

Thank you!

https://raysonlaroca.github.io/