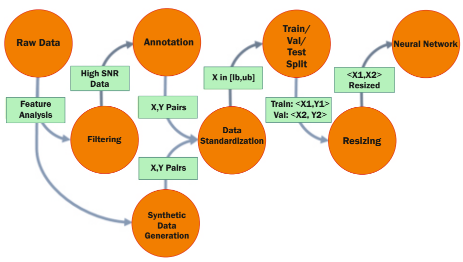
Deep Learning

10

Dataset Preparation for Deep Learning based Applications

P. K. Gaur* and M. Laghate

Computer Division, Bhabha Atomic Research Centre, Mumbai 400085, INDIA



Stages of data-preparation for Deep Learning applications

ABSTRACT

State-of-art Deep learning model implementations for various forms of classification and regression tasks are publicly available. The biggest roadblock in adapting such implementations for the "use-case" arises from the unavailability or unsuitability of raw dataset acquired from the field for model training. A dataset preparation strategy assists in reliably transforming the raw-data acquired from field-sensors to the input for training the Deep learning models. We have introduced data-preparation process for Deep Learning based applications in context to the development of an Automatic Number Plate Recognition (ANPR) model. We have demonstrated the impact of strategizing dataset generation on the performance of the resulting ANPR model. We observed that systematically enhancing the number of text-fonts while keeping other image-generation parameters fixed, resulted in Word-error-rate(WER) dropping from 100% to 10%, whereas, merely increasing the number of training examples randomly, did not have any significant effect on the model performance.

KEYWORDS: *Dataset, Automatic number plate recognition (ANPR), Data-Augmentation, Ablation experiment*

Introduction

Deep learning based methods have pushed the state-of-art in various problem domains of academic interests. Implementations, pre-trained models and datasets for common-tasks are readily available. However, the specific versions of such tasks which are often encountered in the industry do not have public availability of corresponding assets owing to the narrow scope of these problems. For instance, in context of the automation to the generic visual inspection tasks, assets like object-detection, Optical Character Recognition (OCR) and object-tracking algorithm implementations, pre-trained models and datasets are readily available. However, no diverse *annotated* dataset for Indian Automatic Number Plate Recognition model exists in the public domain, requiring manpower involvement to prepare dataset first so as to be able to leverage benefits of AI for Indian ANPR. The current article is motivated by the fact that most research-works in the domain of image-to-text conversion task (like ANPR) state results on publicly available datasets[1] and the ones releasing public datasets give little insight into the impact of dataset preparation on the model training performance[2]. We have discussed the stages of dataset preparation in context to ANPR model development process along with the evaluation metrics. Experiment section of this article presents the quantitative evidence towards the significance of strategizing dataset preparation for fine-tuning model performance. Discussion section of this article highlights the implications of the results followed by the conclusion section which culminates the article outlining future directions.

Methodology

Dataset preparation process for Deep learning applications is depicted in Fig.1. Subsequent sections discuss details of the stages in context to the ANPR application and culminate into a discussion of the evaluation metrics.

Raw data collection and analysis stage

The raw data generated by sensors is collected and its distribution is analyzed in order to assess and quantify the features relevant to the target application. For instance, for vehicle number recognition application, IP camera stream from the deployment site can provide collection of images of incoming vehicles from which the number-plate text-color, background, font and orientation can be analyzed to identify the probable range of variations for each feature. The identified range of variations for these image-parameters can assist in realistic synthetic dataset generation.

Synthetic dataset generation and augmentation stage

Data augmentation and synthetic dataset generation can assist in utilizing the available dataset on large-scale for sample-efficient model training. The raw data samples and the analysis drawn earlier serve as key-inputs for this stage. Synthetic dataset generation involves varying the data-set

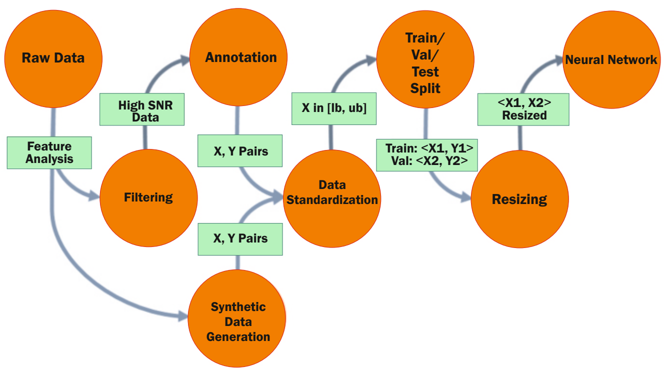


Fig.1: Stages of data-preparation for Deep Learning applications.

*Author for Correspondence: P. K. Gaur
E-mail: pranav@barc.gov.in

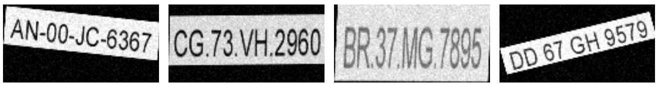


Fig.2: Some samples from synthetic number-plate dataset, showing variation in number-plate orientations, text-font, number-format, number-plate backgrounds.

parameters across a range and generating artificial annotated samples. To achieve an overlap between the distribution of the generated and real datasets, the empirically derived range of variation for each feature of real data is crucial. Dataset augmentation, on the other hand, assumes existing dataset samples over which the feature variations could be applied to generate more dataset samples. In ANPR application, synthetic number-plate images could be generated through a program using image-synthesis software libraries.

Realistic number-plate images, as shown in Fig.2, could be generated by varying number-plate text background, font, size, color, orientation as per the real-image data distribution inferred from the last stage.

Data filtering stage

Given raw data feed and the feature analysis from the last stage, we can selectively sample the raw-dataset by filtering noisy (or low-information) samples, thereby boosting its signal-to-noise (SNR) ratio. In ANPR application, images with vehicles are more relevant than background images for training number-plate recognition model. Further, if an area is marked as the region-of-interest (ROI) in the camera image-view, then the images with a vehicle in ROI are more relevant than the images with vehicle outside ROI. Similarly, for vehicle class prediction, if the current model performs well for 'Jeep' and 'Car' recognition, while predicts 'Bus' and 'Motorcycle' with low confidence, we can use the current model to semi-automatically filter-out images with only Jeeps and Cars from the raw dataset.

Dataset annotation stage

The filtered dataset can be used for model training after it has been annotated with the corresponding ground-truth labels, thereby generating sample(X) and label(Y) pairs as shown in Fig.1. In ANPR application, for each number-plate image, the number-plate text defines its ground-truth annotation. Program scripts leveraging previously trained number-plate prediction models could be used to semi-automatically label the number-plate images and can then later be manually fine-tuned.

Dataset standardization stage

As with any numerical optimization method, model training convergence for Deep learning applications is critical. Dataset features--identified in the first stage--have wild variation in the value ranges they acquire which often negatively affects the stability of model training and render it biased towards features with higher magnitudes. Dataset normalization counters this by explicitly defining a common fixed range of values (the lower and upper bounds) across all features of the dataset, shown as [lb, ub] in the Fig.1. In ANPR application, input number-plate images from RGB color-space are mapped to a more limited grayscale color-space thereby reducing the possible values each pixel can acquire.

Dataset splitting stage

The process of partitioning the available annotated dataset into train, validation and test datasets is called dataset-splitting. The test-dataset often comes from the end-user and is kept hidden during model-development, therefore, often this stage involves partitioning the available dataset into train and validation dataset, shown as <X1, Y1> and <X2, Y2>, respectively in Fig. 1. The validation dataset is supposed to test the model during training and is therefore not exposed to the model for training. In ANPR application, 80%-20% dataset split strategy is followed to generate train and validation datasets respectively from the available annotated dataset. The validation samples are drawn randomly from the dataset to avoid distribution bias.

Dataset resizing stage

Neural networks assume their input to be fixed in size, whereas the raw-data from field sensors comes in variety of resolutions. The normalized dataset, is therefore, resized to

Table 1: Ablation study of ANPR model performance with respect to number of text-font.

Index	Training data	Validation data	CER and WER
1	128 K frontal images with 1 font class	12.8K frontal images with 1 font class	100, 100
2	1.28 M frontal images with 1 font class	128K frontal images with 1 font class	100, 100
3	100 K images with 1 font class, random rotation, translation and Gaussian noise	20 K images with 1 font class, random rotation, translation and Gaussian noise	66.29, 100
4	50 K images, 5 font-classes	10 k with 5 font-classes	18.38, 98.57
5	50 K images, 5 font-classes, test-image background	10 kwith 5 font-classes, test-image background	10.05, 71.43
6	100 K images, 5 font-classes	20 kimages, 5 font-classes	5.22, 41.43
7	100 K images, 10 font-classes	20 kimages, 10 font-classes	3.80, 28.57
8	200 K, 10 font-classes	40 k images, 10 font-classes	**1.62, 10.0**

the network input size. For instance, in ANPR application, YOLOv3 model-based vehicle number-plate detector assumes input image to be of size 416 x 416 pixels, whereas the input image from IP camera has a resolution of 1920x1080 pixels, therefore input image needs to be resized from original resolution to 416x416. Further, notice that the network input size has 1:1 aspect ratio, while most image formats have 6:4, 4:3 aspect ratios, therefore, a resizing strategy which preserves the original aspect ratio-like the letterbox resizing-needs to be adopted to preserve the legibility of relevant features.

Evaluation metrics

For quantitative assessment of the effect of a data-preparation strategy on model-performance, an evaluation metric is crucial. Evaluation metric represents the objective of a target application against which various data-preparation and model training strategies could be assessed. For ANPR application, Word error rate (WER) and Character error rate (CER) [4] over a test-dataset are used to assess the effectiveness of a trained model against vehicle number prediction task.

Experiments

An ablation experiment demonstrating the impact of text font as a synthetic dataset generation parameter for vehicle number recognition task is reported in Table-1. The experiments are conducted on a PC with Intel Core processor (2GHz), 8GB RAM, NVIDIA Quadro M4000 GPU, Ubuntu 18.04 OS and TensorFlow 1.14 library. The test-dataset contains 100 images of vehicle number-plate collected from an IP camera installed at the ANPR system deployment site.

Discussion

Progressively incrementing the number of text fonts used to generate synthetic vehicle number-plate dataset reduced the WER from 100% to 10% as shown in experiments Sr. no. 4 through 8 of Table-1, indicating that merely exposing the model to variations in text font helps in generalizing it over number-plate text in test-dataset. However, adding more images without content variations did not affect the WER as can be observed in experiments Sr. no. 1 and 2 of Table-1. In experiment Sr. no. 5, adding a number-plate text background drawn from real number-plate images, significantly brought down the WER, which indicates the impact of bringing realism to the synthetically generated number-plate images, as can be achieved using Pix2Pix GANs[3].

Conclusion

We have described a generic data-preparation process for deep-learning in context of an ANPR application. The significance of data-preparation process is experimentally demonstrated with an ablation study over the effect of number-plate text-font on resulting Word-error-rate (WER) of the underlying number-plate recognition model. The experiment highlights the importance of aligning the synthetic dataset-distribution to be as close to real-dataset as possible, which we would like to explore in the context of ANPR application by leveraging generative models.

References

[1] Changhao Wu, Shugong Xu, Guocong Song, Shunqing Zhang, How many labeled license plates are needed?, arXiv:1808.08410, 2018.

[2] Zhenbo Xu, Wei Yang, Ajin, Meng, Nanxue Lu, Huan Huang, Changchun Ying, Liusheng Huang Towards End-to-End License Plate Detection and Recognition: A Large Dataset and Baseline, ECCV 2018.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, arXiv:1611.07004, 2017.

[4] Neudecker, C, Baierer, K, Gerber, M, Clausner, C, Antonacopoulos, A and Pletschacher, SA survey of OCR evaluation tools and metrics, HIP '21: The 6th International Workshop on Historical Document Imaging and Processing, 2021.