All-in-one Data Cleansing Tool

Arun Prasad Vailoppilly*, Ramkumar Sakthivel*, Resham Sundar Kumar*
Aravind G*, Vignesh BDSV*, Sri Aravind Sairaman*
Tata Consultancy Services

Abstract

With the shift in trend towards the 'Data-Centric Artificial Intelligence(AI)' approach, it is indispensable to provide clean and properly labelled data to train Deep Learning(DL) models. Also, the accumulation of noisy data during the data collec- tion phase especially for use cases like autonomous driving is inevitable. To tackle the problem of noisy data, we are proposing a tool that uses DL models to identify and remove the noise from the data that is collected. This tool is a collection of pipelines that helps to curate the data and also gives a high level overview on the data and saves a lot of manual effort. Integration of our data cleansing tool into a DL model development workflow will not only aid in improvement of annotation process, but also yield better results with less manual effort. The tool is designed such that it is modular, as the type of noises varies according to domain and dataset.

1 Introduction

When we talk about training the deep learning models, we tend to focus more on tuning the hyper parameters to improve its performance, but less on the type of data being introduced to the model. The proposal in this paper deals with a tool or pipeline which initially gives a high-level insight on the pool of unlabelled data and then removes the noise present in the data after identifying them . In order to process the data before introducing it to the model, there are many things that should be taken care of. For example, the resolution of the image, if there are any noises, if the dataset is enough, etc.

In this paper, we have used unlabelled data which is a mix of clean and noisy images. To initially know about the insights of the data we have used a dimensionality reduction technique that can plot the data for visualization in 2D or 3D. This can also be of help to group the images that are similar since the data points are then closer to each other, hence can be used to segregate images on the basis of attributes/meta-information such as day and night, sunrise or sunset, car and motorcycle, etc. As a next step in the pipeline, a classifier has been used in order to classify the data on the basis of noise. We can either use a trained model or hard code an algorithm that does the work required. This can then separate the incoming data into two classes wherein one class can be the one with noisy images and the other one can be clean images. Based on the quantity of noisy data, the user can then choose to completely discard the noisy data or to de-noise the data using DL methods so as to increase the amount of usable data for model training.

*All authors have equal contribution

2 Related works

Even though there has been a huge amount of research focused on cleaning and curating the bigdata, most of them focus on approaches suited for structured and semi-structured data. On the other hand the amount of research and tools available to clean non-structured data especially images is low.

The very first research on data cleansing method proposed manual way of finding the patterns and cleansing the data[5]. This is almost impossible in the current deep learning era where raw data for each task is added in abundance. A few research papers published later talked about data cleansing approaches that were problem specific [6],[7]. Loss of flexibility in such tools might become complicated when new problems/tasks comes up. The research [8] focuses on cleansing the duplicate data so as to avoid re-annotation and model overfitting. One of the recent researches [9], talks about an image data cleaning framework that is based on DL. Even though it tries to resolve most of problems that previous tools had, it majorly focuses on removing the noisy data directly which might be expensive in few scenarios like autonomous driving where huge amount of data is needed to train state-of-the-art DL models. Our tool tries to apply some DL based techniques to eliminate the noise in images and then adds them to the clean data pile for training thus not 'wasting' the resources utilized for data collection.

3 Methodology

Collected images often contain unlabelled thus making it difficult to segregate when they are in abundance. In such a case, the below architecture makes use of dimensionality reduction technique to visualize data that contains many features. Applying this technique enables us to gather high level data insights, such as clustering the images in a dataset that are similar in nature. Adding to this, the pipeline consists of the noise classifier which bifurcates the images collected into clean and noisy images.

The noisy images are then passed through a decider which depends on the configurable parameter (T) which is set by the user and also the NR value. The purpose of these values is to let the user decide whether to completely discard a particular noisy image or to apply some de-noising technique in order to get a better version of the image. The de-noiser removes noise from the image and merges it to from cleansed model trainable data which leads to a better performance during feature extraction or other required tasks. Figure 1 shows the complete pipeline architecture diagram of the tool.

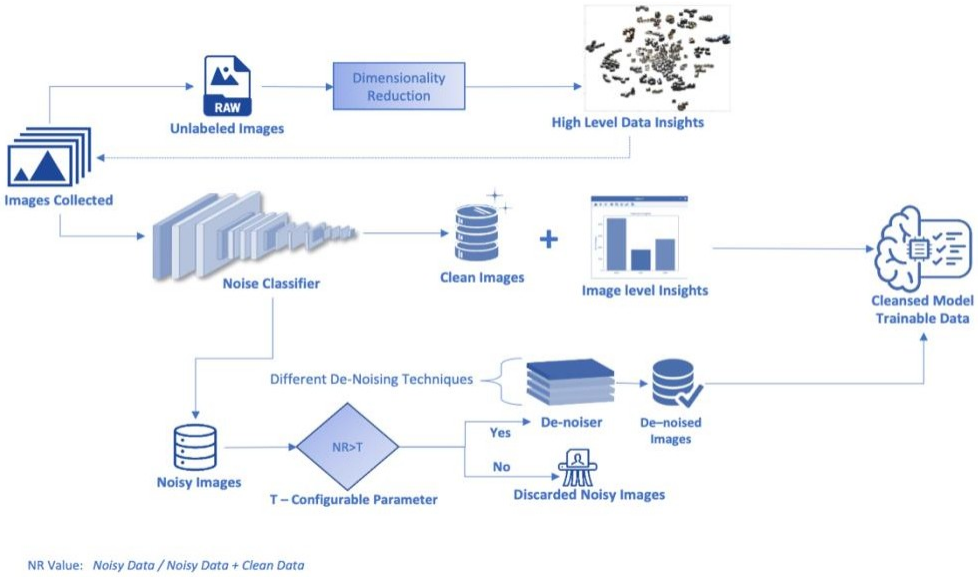

Figure 1: Architecture Diagram

4 Experiments and Results

4.1 Dataset

We considered the task of object detection since it happens to be the basic Computer Vision (CV) technique for autonomous driving use case. For running the experiments Udacity's self-driving car[1] dataset was used. Out of the 15000 images in the dataset we considered only 500 images for experi- mentation purpose. The dataset contained 11 classes that include biker, car, pedestrian, trafficLight, trafficLight-Green, trafficLight-GreenLeft, trafficLight-Red, trafficLight-RedLeft, trafficLight-Yellow, trafficLight-YellowLeft and truck. Since the data was clean we used certain CV techniques to induce noises like motion blur and haze effect on a part of the chosen 500 images.

4.2 All-in-one Data Cleansing Tool (AIODCT)

On the artificially generated noisy dataset that had both clean and noisy images, TSNE[10] dimensionality reduction method was applied to visualize the data distribution and it was able to cluster the noisy data separately to some extent. After this step different noise classifiers were used to segregate noisy and clean data. Since the amount of dataset is small it was decided not to discard the noisy data but rather to de-noise those noisy images. For de-noising DL models like [3] and [4] were used for deblurring and dehazing respectively. Finally the de-noised data was added to the originally classified clean data to form cleansed model trainable data.

4.3 Object Detection

To validate our AIODCT we ran a few experiments wherein we trained a lightweight object detection DL model, tiny-YOLOv3[11]. This lightweight DL model was chosen because the amount of data considered was very less. In order to compare the results we trained the same tiny-YOLOv3 model with three different sets of data. The first set was the original clean 500 images that did not include any noisy images, the second set was the noisy data that includes clean images along with blurry and hazy images that were generated by CV algorithms as mentioned in 4.1. The final set was the de- noised data that was obtained after running the AIODCT on the second set of noisy images. The accuracy metrics for the object detection task is tabulated in Table 1.

S.No	Dataset for model training	mAP (at 0.5 IoU)
1.	Clean ground truth data	50.48
2.	Noise induced data	47.32
3.	AIODCT cleansed data	48.15

Table 1: Object detection accuracy metrics.

We can observe from the initial experiments that there is a decrease in the performance in model prediction, where the mAP (mean average precision) score at 0.5 IoU (Intersection over union) had reduced from 50.48 to 47.32. However, post using the AIODCT on the images we can observe that there is an increase in the mAP score from 47.32 to 48.15. In order to maintain consistency, it was made sure that all the hyper-parameters and model configurations were constant for all three experiments.

5 Conclusion

Based on the experimental results we can clearly see that there is an improvement in the accuracy metrics after using the AIODCT on the noisy data. Our future area of work would be improvising the tool and running experiments on large dataset.

On a high level, we can observe that AIODCT has 3 major functionalities like generating high-level insights, noise classification, and denoising. In this paper, we have taken only blur and haze as the type of noises, however, the modularity of AIODCT helps us in extending this to other types of noises as well.

References

[1] Bansal, R., Raj, G., Choudhury, T. (2016, November). Blur image detection using Laplacian operatorand Open-CV. In 2016 International Conference System Modeling Advancement in Research Trends (SMART) (pp. 63-67). IEEE.

[2] Sowjanya, A., Kumar, S., Swaroop, K. S. (2021). Convolution Neural Network based Rain Noise Removal for Real Time Application. In IOP Conference Series: Materials Science and Engineering (Vol. 1042, No. 1, p. 012001). IOP Publishing.

[3] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., Shao, L. (2021). Multi-stage progressive image restoration. In Proceedings ofthe IEEE/CVFConference on Computer Vision and Pattern Recognition (pp. 14821-14831).

[4] Li, B., Peng, X., Wang, Z., Xu, J., Feng, D. (2017). Aod-net: All-in-one dehazing network. In Proceedings of the IEEE international conference on computer vision (pp. 4770-4778)

[5] Isabelle Guyon, Nada Matic, Vladimir Vapnik, et al. Discovering informative patterns and data cleaning., 1996.

[6] Chu, X., Ilyas, I.F., Krishnan, S., Wang, J. (2016). Data Cleaning: Overview and Emerging Challenges. Proceedings of the 2016 International Conference on Management of Data.

[7] Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A.K., Ilyas, I.F., Ouzzani, M., Tang, N. (2013). NADEEF: a commodity data cleaning system. SIGMOD '13.

[8] Ng, Hong-Wei Winkler, Stefan. (2015). A data-driven approach to cleaning large face datasets. 2014 IEEE International Conference on Image Processing, ICIP 2014. 343-347. 10.1109/ICIP.2014.7025068.

[9] Zhang, Y., Jin, Z., Liu, F., Zhu, W., Mu, W., Wang, W. (2020). ImageDC: Image Data Cleaning Framework Based on Deep Learning. 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), 748-752.

[10] Van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[11] P. Adarsh, P. Rathi and M. Kumar, "YOLO v3-Tiny: Object Detection and Recognition using one stage improved model," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 687-694, doi: 10.1109/ICACCS48705.2020.9074315.