2022-06-16
Kasper Socha, Markus Borg
RISE Research Institutes of Sweden

ML Data Validation Report

Motivation

This section describes the results from testing the SMIRK data set. The data testing involves a statistical analysis of its distribution and automated data validation using Great Expectations[1]. Together with the outcome of the Fagan inspection of the Data Management Specification this constitutes the ML Data Validation Results in AMLAS. The results entail evidence mapping to the four assurance-related desiderata, i.e., we report a validation of 1) data relevance, 2) data completeness, 3) data balance, and 4) data accuracy. Since we generate synthetic data using ESI Pro-SiVIC, data relevance has been validated through code reviews and data accuracy is implicit as the tool's ground truth is used. For both the relevance and accuracy desiderata, we have manually analyzed a sample of the generated data to verify requirements satisfaction.

We validate the ethical dimension of the data balance by analyzing the gender (DAT-BAL-REQ1) and age (DAT-BAL-REQ2) distributions of the pedestrians in the SMIRK data set. SMIRK evolves as a demonstrator in a Swedish research project, which provides a frame of reference for this analysis. Table 1 shows how the SMIRK data set compares to Swedish demographics from the perspective of age and gender. The demographics originate in a study on collisions between vehicles and pedestrians by the Swedish Civil Contingencies Agency[2].

Table 1: Distribution of pedestrian types in Sweden and in the SMIRK data set.

Pedestrian types	Population	Accidents	Deadly accidents	SMIRK
Children & young adults (0-19)	23%	27%	12%	12.5%
Adult males (20+)	39%	31%	57%	50%
Adult females (20+)	38%	42%	31%	37.5%

We notice that 1) children are slightly over-represented in accidents but under-represented in deadly accidents, and that 2) adult males account for over half of the deadly accidents in Sweden. The rightmost column shows the distribution of pedestrian types in the entire SMIRK data set. We designed the SMIRK data generation process to result in a data set that resembles the deadly accidents in Sweden, but, motivated by AI fairness, we increased the fraction of female pedestrians to mitigate a potential gender bias.

Details

Automated data testing is performed by defining conditions that shall be fulfilled by the data set. These conditions are checked against the existing data and any new data that is added. Some tests are fixed and simple, such as expecting the dimensions of input images to match the ones produced by the vehicle's camera. Similarly, all bounding boxes are expected to be within the dimensions of the image. Other tests look at the distribution and ranges of values to assure the completeness, accuracy, and balance of the data set and catch human errors. This includes validating enough coverage of pedestrians at different positions of the image, coverage of varying range of pedestrian distances, and bounding box aspect ratios. For values that are hard

[1] https://greatexpectations.io/
[2] https://rib.msb.se/filer/pdf/27438.pdf (Schyllander, 2014)

to define rules for, a known good set of inputs can be used as a starting point and remaining and new inputs can be checked to against these reference inputs. As an example, this can be used to verify that the color distribution and pixel intensity are within expected ranges. This can be used to identify images that are too dark or dissimilar to existing images.

Results

Figure 13 shows a selection of summary plots from the data testing that support our claims for data validity, in particular from the perspective of data completeness. Subplot A) presents the distance distribution between ego car and pedestrians, verifying that the data set contains pedestrians at distances 10–100 m (DAT-COM-REQ5). Subplot B) shows a heatmap of the bounding boxes' centers captures by the 752x480 WVGA camera. We confirm that pedestrians appear from the sides of the field of view and a larger fraction of images contain a pedestrian just in front of ego car. The position distribution supports our claim that DAT-COM-REQ4 is satisfied, i.e., the data samples represent different camera angles. Subplot C) shows a heatmap of bounding box dimensions, i.e., different aspect ratios. A variety of aspect ratios indicate that pedestrians move with a diversity of arm and leg movements – indicating walking and running – and thus support our claim that DAT-COM-REQ3 is fulfilled.

Finally, subplot D) shows the color histogram of the data set. In the automated data testing, we use this as a reference value when adding new images to ensure that they match the ODD. For example, a sample of nighttime images would have a substantially different color distribution.