KI DATA TOOLING
The Data Kit for Automotive AI

Mixed Training – Identification and Filling of Data Gaps

Maarten Bieshaar | Bosch
Maximilian Menke | Bosch

Acquisition, Curation & Refinement of Real Data

Mixed Training

Synthetic data has an enormous potential. So far, most studies have focused on the "uncontrolled" mixing of real and synthetic data [1]. In this series of studies, we propose an alternative approach (see Fig. 1) in which we use synthetic data specifically to

1. identify data gaps and AI model performance issues,
2. fill those gaps using synthetic data.

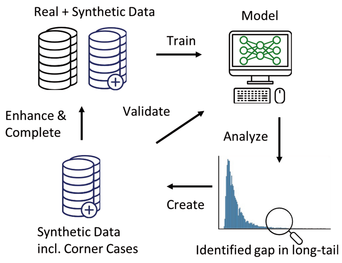

Figure 1: Data-Driven Engineering Process to identify and mitigate gaps in real datasets using synthetic data mixed with real data.

Identification of Gaps and Corner Cases

We systematically test the model prediction and the data against requirements specified in advance with respect to the operational design domain [2]. However, it is practically impossible to cover all (corner) cases with real data. Therefore, we propose to use synthetic data to systematically reveal data gaps and AI performance limitations, e.g., through coverage testing [3]. In our case study, we show that synthetic data can be used for this purpose. Therefore, we artificially create a data gap with respect to the height of the pedestrian bounding boxes. Although synthetic and real data do not behave identically, the gap in the object detector training data can be readily detected from the trends in the log-average miss rate (see Figure 2).

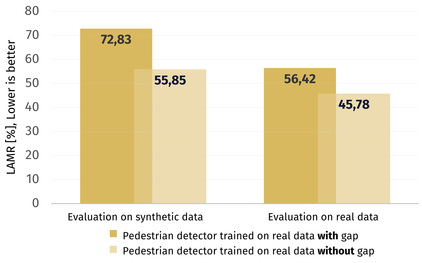

Figure 2: Evaluation of pedestrian detector's performance w.r.t. pedestrians in gap on synthetic (SynPeDS) and real data (CityPerson). The gap is constructed w.r.t. to pedestrian's bounding box height.

Filling of Gaps using Synthetic Data

We examined the possibility of closing data gaps with synthetic data. Specifically, we filled the previously investigated gap based on the pedestrian's bounding box height with (1) appropriate synthetic data from the gap, and (2) any synthetic data in- and outside the gap. Our findings indicate that synthetic data can be utilized to some degree to fill the data gap (see Fig. 3). For the experiments, we use the CityPersons and the SynPeDS [4] as real and synthetic datasets, respectively. The most effective approach was the usage of all synthetic data (from in- and outside of the gap). It is presumed that this is because not only the content gap needs to be closed, but also a general dataset or appearance gap.

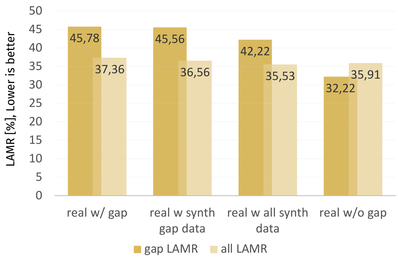

Figure 3: Evaluation of pedestrian detectors trained on real data with and without gap (left and right most columns). Performance results of mixed trainings with only synthetic data of the data and arbitrary (entire) synthetic data (two columns in the middle).

Conclusion

Synthetic data can identify gaps in data sets and fill these gaps in mixed training scenarios to some extent. However, additional research is necessary, such as examining more "realistic" data gaps like luminance or local contrast, and developing more sophisticated mixed training strategies to better handle synthetic data.

References:

[1] Sven Burdorf, Karoline Plum, Daniel Hasenklever. "Reducing the Amount of Real World Data for Object Detector Training with Synthetic Data," in CoRR, vol. arXiv/2202.00632, 2022.
[2] Zhang, R., et al, "DDE process: A requirements engineering approach for machine learning in automated driving," in 2021 IEEE 29th International Requirements Engineering Conference (RE), 2021, pp. 269-279.
[3] Gladisch, C., et al, "Leveraging Combinatorial Testing for Safety-Critical Computer Vision Datasets," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020.
[4] Stauner, T., et al, "SynPeDS: A Synthetic Dataset for Pedestrian Detection In Urban Traffic Scenes," in Proceedings of the 6th ACM Computer Science in Cars Symposium, 2022.

Partners

External partners

For more information contact:
maarten.bieshaar@de.bosch.com
maximilian.menke@de.bosch.com

KI Data Tooling is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

www.ki-datatooling.de @KI_Familie KI Familie