Fraunhofer
IESE

Fraunhofer Institute for Experimental
Software Engineering IESE

©iStock.com/Henrik5000

Data augmentation strategies for simulating low-quality data

AI/ML-based models learn their behavior from example data, are tested on data, and are applied on data, hence they are also referred to as data-driven models.

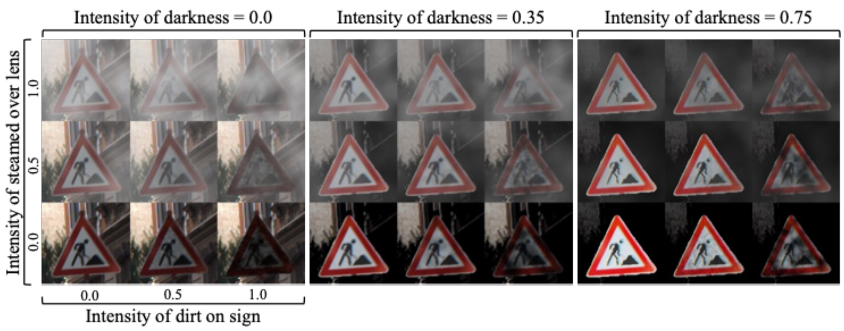
Issues in the quality of the data influence such models in various ways. Here are some examples:

- Outliers in the training data can mislead the model to overfit on them.
- Missing values: Incomplete or missing data points can significantly affect the performance of AI models.
- Label noise can lead the model to learn incorrect patterns and relationships, negatively impacting its generalization ability.

A common instinctive response when confronted with data quality issues is to attempt to resolve them by filtering them out, with the aim to obtain cleaner data sets during the development process. However, it is important to note that this approach can become almost impossible in the model’s a pplication phase, where we encounter various scenarios. For instance, there are situations where the model was not exposed to any of these instances during its learning phase. Another scenario is that the model’s anticipated performance is projected to be higher due to testing on data that does not accurately represent real-world conditions.

In the Data Science department at Fraunhofer IESE, we develop methods and tools to intentionally generate data with quality issues, which can then be used to

- enhance the training data so that the model can learn how to deal with bad-quality data
- test how the model is affected by specific quality issues, i.e., its robustness wrt. different types of quality issue
- create test data that is representative of the model’s future application context.



Intensity of darkness = 0.0
Intensity of darkness = 0.35
Intensity of darkness = 0.75
Intensity of steamed over lens
0.0
0.5
1.0
Intensity of dirt on sign
0.0
0.5
1.0

Augmentation of traffic sign images with different intensities of darkness, steamed-over lens, dirt on the sign, and their interactions

We provide:

- A method for deriving the relevant quality issues in a specific use case, identifying the contextual conditions that need to be fulfilled for them to arise, and determining how they interact with and influence each other
- A method for enhancing image data with quality issues (referred to as data augmentation), with a focus on realism and consideration of quality issues interacting with each other (e.g., parts of a traffic sign with dirt on it will reflect less in the lights of approaching vehicles in the dark), which was implemented for the use case of classifying traffic signs
- A Python library, **badgers**, for generating bad data – more precisely, for augmenting existing data with data quality deficits such as outliers, missing values, noise, etc.

Contact

Lisa Jöckel
Department Data Science
Fraunhofer IESE
Phone +49 631 6800-2267
lisa.joeckel@iese.fraunhofer.de
www.iese.fraunhofer.de

Julien Siebert
Department Data Science
Fraunhofer IESE
Phone +49 631 6800-2236
julien.siebert@iese.fraunhofer.de
www.iese.fraunhofer.de