Increasing Trust in Data-Driven Model Validation

A Framework for Probabilistic Augmentation of Images and Meta-Data Generation using Application Scope Characteristics

Lisa Jöckel[1] and Michael Kläs[1]

[1] Fraunhofer Institute for Experimental Software Engineering IESE,
Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
{lisa.joeckel, michael.klaes}@iese.fraunhofer.de

Abstract. In recent years, interest in autonomous systems has increased. To observe their environment and interact with it, such systems need to process sensor data including camera images. State-of-the-art methods for object recognition and image segmentation rely on complex data-driven models such as convolutional neural networks. Although no final answer exists yet on how to perform safety evaluation of systems containing such models, such evaluation should comprise at least validation with realistic input data, including settings with suboptimal data quality. Because many test datasets still lack a sufficient number of representative quality deficits, we consider augmenting existing data with quality deficits as necessary. For this purpose, a novel tool framework is presented and illustrated using traffic sign recognition as a use case. The extendable approach distinguishes between augmentation at the object, context, and sensor levels. To provide realistic augmentation and meta-data for existing image datasets, known context information and conditional probabilities are processed. First applications on the GTSRB dataset show promising results. The augmentation of datasets facilitates a more rigorous investigation of how various quality deficits affect the accuracy of a model in its target application scope.

Keywords: Safety, Traffic Sign Recognition, Data Augmentation, Data Quality, Application Scope Characteristics, Uncertainty, Convolutional Neural Networks

1 Motivation

In recent years, interest in autonomous systems – particularly, but not limited to, autonomous driving – has increased [2]. Such systems work in an open context, which cannot be exhaustively specified upfront. They need to sense their environment in order to adapt their behavior. A self-driving car needs to detect pedestrians crossing the street or a temporary stop sign and react appropriately. Cameras are still the sensor of choice here, providing the key input for detecting and recognizing objects through, e.g., deep convolutional neural networks (CNNs) [3]. Ciregan et al. [4], e.g., achieved a classification accuracy of 99.46% on GTSRB, a German traffic sign benchmark dataset [1].

However, especially when we consider safety-related functionality of autonomous systems such as detection of a stop sign, we need to ask how much we can rely on accuracy statements obtained from processing existing test datasets.

Like any data-driven model used for image recognition, CNNs face the problem that their intended input-output relationship cannot be completely specified [5]; i.e., the model needs to learn this relationship on a comparatively small and probably not representative sample of input-output examples. This strongly limits traditional verification, making sound statistical validation on test data even more essential. Statistical conclusions on how a data-driven model performs in its target application scope can only be drawn, however, if the test dataset is representative for the target scope.

Today, we can commonly not assume that available test datasets are representative for the intended target application scope of a tested model. Our experience shows that most datasets are artificially clean, i.e., they omit or at least underrepresent many of the quality deficits that arise in real-world settings [6]. However, it does not appear reasonable to make statements about the real-world performance of a data-driven model if it was not tested on data reflecting the real world. For example, a model for traffic sign recognition should also be tested on images with heavy rain or backlight conditions, a dirty camera lens, or snow-covered traffic signs if such deficits can occur in its target application scope. A related challenge is that even if representative test data is available, most critical edge cases might be too rare to be included in sufficient numbers in a reasonably sized dataset. Examples are pedestrians on a rural road at night or the combination of a defective headlight and oncoming traffic with high beam.

Besides intensifying the collection of real data, there are two ways to deal with these problems: creating artificial images using simulation environments [7] or augmenting existing images with quality deficits. The first approach suffers from the ‘reality gap’. Attempts to narrow this gap train specialized GANs [8] and apply them to artificially generated images to make them look more realistic. Even though success has been reported for restricted settings, such as grasping tasks of a stationary robot [7], we are not aware of successful applications in more complex environments such as road traffic.

Our contribution is a framework and a tool instantiation for augmenting image data with realistic quality issues and corresponding meta-data. The framework provides guidance for the identification of possible quality deficits, the design of a context model for deriving conditional probabilities for the occurrence of possible deficits, and the layering of various kinds of potentially interacting augmentations. Extending existing work, the framework allows (1) enriching datasets with quality deficits reflecting their natural distribution in the target application scope and (2) applying several deficits to the same image without causing artificial overlay issues.

The remainder of this paper is structured as follows. Section 2 provides an overview of related work in the area of quality-related augmentation of images. Section 3 outlines and illustrates nine steps for building an augmentation-tooling instance for a given data-driven component and three steps for applying it to a given image dataset. Section 4 concludes the paper by discussing limitations and future work.

2 Related Work

Image augmentation is a commonly used preprocessing technique to improve the performance of data-driven models and make them more robust by increasing the count and variety of data points available during model training [9]. In the context of model validation, augmentation has been applied less frequently to date.

Three kinds of augmentation can be distinguished: (1) those mainly used to increase the number and variety of data points, such as image rotations and shifts; (2) those used to intentionally decrease the quality of the image, making the task harder for the model; and (3) those specifically designed to fool a given data-driven model by generating adversarial examples [10]. Because this work focuses on the validation of data-driven models, we consider neither the first kind, which is mainly relevant for model training, nor the third kind, which is an important but security-related topic.

Quality-related augmentations can be distinguished with respect to the degree of realism they intend to provide: (a) Simple artificial augmentations do not intend to emulate concrete, real quality deficits but are added to images, e.g., in the form of various kinds of random noise [6, 11, 12]. (b) Artificially appearing augmentations capture specific aspects of a real quality deficit, e.g., emulating snow by reducing the saturation of an image [13]. (c) Near-photorealistic augmentations use, e.g., available depth information to adjust haze on a pixel basis [14]. There are also approaches that utilize style transfer and GANs [15, 16]. Because our aim is to use augmentations to make a given test dataset more realistic and to investigate the effects of specific quality deficits, this work focuses on near-photorealistic augmentations. However, we decided against the use of GANs because the quality of their results still appears to be unstable.

A review in the context of street scenes and traffic sign recognition showed that besides work on specific deficits such as haze and fog, snow, rain, shadows, and defocus [14, 17], a number of frameworks exist that include augmentations for several quality deficits. Cheng et al. address, e.g., haze, fog, and snow [13] and Temel et al. examined the robustness of traffic sign recognition under challenging conditions [18].

However, most reviewed papers on quality-related augmentation, including the identified frameworks, deal with quality deficits on an individual basis; i.e., they apply only a single deficit to a given image or ignore possible interactions when applying multiple deficits. One exception from this observation is an approach that combines augmentations on a LAB color space [19]. Moreover, the reviewed papers do not consider probabilistic dependencies between meta-data characterizing the context of an image and the applied augmentations. This means that they neither allow generating a realistic distribution of deficits, such as would occur in the target application scope, nor do they consider correlations between various kinds of deficits (including the extreme of mutual exclusivity).

3 Conceptual Augmentation Framework

This section introduces a general augmentation framework for data-driven components processing image data. Moreover, it illustrates how to instantiate it using the example of a tool that supports the augmentation of traffic sign images in an existing dataset.

The overall process consists of two major stages. The first stage (P1-P9) comprises all the steps for building the specific augmentation-tooling instance for a given data-driven component and its target application scope. The second stage (A1-3) comprises all the steps required to apply an augmentation-tooling instance to an image dataset.

P1 - Understand the data-driven component and its target application scope. Building an augmentation-tooling instance requires an understanding of the investigated data-driven component, including its potential input data and the scope in which it is intended to be applied.

Our example considers a traffic sign recognition component with an image of the detected traffic sign as its main input and data from other vehicle sensors as optional additional information sources (e.g., outside temperature sensor, velocity signal, GPS signal, rain sensor, brightness sensor, online weather broadcast).

Furthermore, we defined its target application scope as passenger vehicles using public roads in Germany, independent of the time of year or the time of day.

P2 - Identify quality deficits (QD) affecting the data-driven component. Considering realistic conditions in the target application scope, there are situations that reduce the quality of the data. In order to build a framework that augments data with quality deficits, relevant quality issues occurring in the target application scope have to be identified and described, considering existing literature and domain expert opinion. The findings should be consolidated in a list and grouped according to sensor, context, and object. If necessary, quality deficits can be prioritize with respect to their occurrence probability and expected impact on the outcome quality of the data-driven component.

For traffic sign recognition, we identified quality deficits concerning either the context of the sign, the sign itself (object), or the built-in camera as the sensor. Specifically, these deficits include: for **context** – light, darkness, weather condition (rain, snow, haze, heat shimmer), shadows, occlusion; for **object** – physical damage (bent, broken, holes), graffiti and stickers, faded colors, dirty sign, wet sign, snow on sign; and for **sensor** – placement, particles on lens (dirt, snow, rain drops, steam), lens and sensor limitations (e.g., resolution, noise, glare effects, backlight, motion blur), camera calibration (e.g., defocus, color temperature), camera processing (e.g. compression errors).

P3 - Identify scope characteristics influencing the occurrence or intensity of QD. In order to identify relevant scope characteristics, we go through the list of identified quality deficits, consider when and why they occur, and look at the characteristics of the target application scope influencing their occurrence or intensity.

As relevant scope characteristics that influence quality issues in recognizing traffic signs we identified factors related to geographical position, weather, time, lighting conditions, and vehicle velocity (see white boxes in Fig. 1). As the augmentation addresses traffic sign recognition – not traffic sign detection – factors influencing the detection or

relevance of the detected traffic sign such as the placement or reflective surfaces causing wrongly detected mirror images are not considered.

P4 - Define a causal model with dependencies between scope characteristics. In order to model dependencies between scope characteristics, we arrange them into an acyclic graph, where the directed relations mean 'influences'. In a refinement, missing scope characteristics that influence other relevant scope characteristics are added.

For our example application, a graph is presented in Fig. 1. Time, e.g., influences various other scope characteristics, such as weather or traffic situation, directly; others, such as lighting conditions, do so indirectly through other characteristics. From the geographical position, we can determine road type (e.g., motorway, farm road, street in town), constructions (e.g., tunnels, street canyons), natural surface formations (e.g., forest, hills, rocks) that can cause shading, and traffic situation based on the current time.

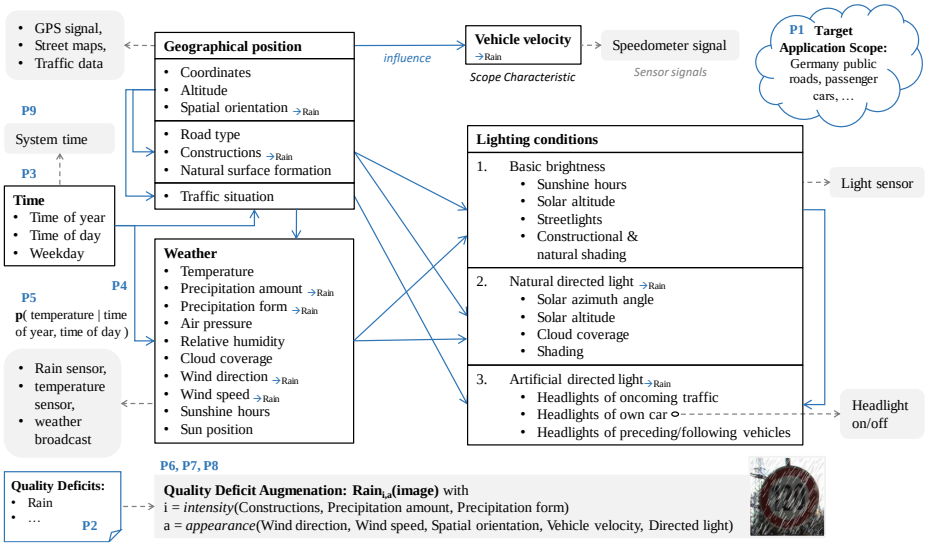

Fig. 1. Steps P1 to P9 of the augmentation framework with a focus on the context model.

P5 - Derive conditional probabilities to quantify identified dependencies. Scope characteristics follow a probability distribution $p(SC_{V=u} \mid TAS)$ regarding their natural occurrence in the target application scope TAS, with $SC_{V=u}$ being the scope characteristic with value $V = u$. Because different characteristics can be interdependent, we also need to consider conditional probabilities. Example: How likely is it that the temperature will be higher than 30 °C when we are in location (x, y) with x being the latitude and y the longitude on day 143 of the year at 3 p.m.?

Several public data sources exist that can be used to calculate these probabilities (e.g., historic weather data from DWD [20] or maps from OpenStreetMap [21]). If no empirical data is available, reasonable expert-based approximations need to be applied, e.g., for the velocity of a car based on its geographical position or the likelihood and amount of dirt on a traffic sign.

P6 - Identify existing augmentation techniques available for QD. In the next step, we need an overview of existing work on image augmentation for the quality deficits identified as relevant. We must understand how the quality deficit manifests in an image and what needs to be considered when changing the image in order to augment a specific quality deficit.

For example, dirt on a sign can occur in different colors and degrees. It affects only the pixels of the object (i.e., the traffic sign) and needs to be applied as a randomized semitransparent pattern influencing also the reflection property of the affected areas.

P7 - Define the order of applying augmentations. In many cases, there is a certain order to consider when applying augmentation. For instance, object augmentations (e.g., dirt) should be applied first, then context (e.g., darkness), and finally sensor (e.g., steamed-up). This way, consequences from having a particular quality issue can be incorporated into the augmentation of other quality issues; e.g., dirt on a traffic sign reduces its reflective effect when illuminated at night by headlights and the brightness of the fog on a camera lens decreases with the general reduction of brightness at night.

In Fig. 2, the interaction with different intensities of the quality deficits darkness, dirt on sign, and steamed-up lens is displayed, considering previous influences.

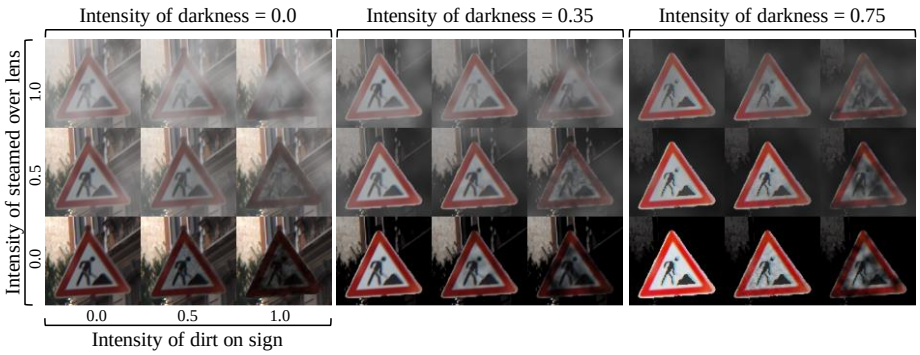

Fig. 2. Combination of darkness, dirt on sign, and steamed-up lens at different intensity levels.

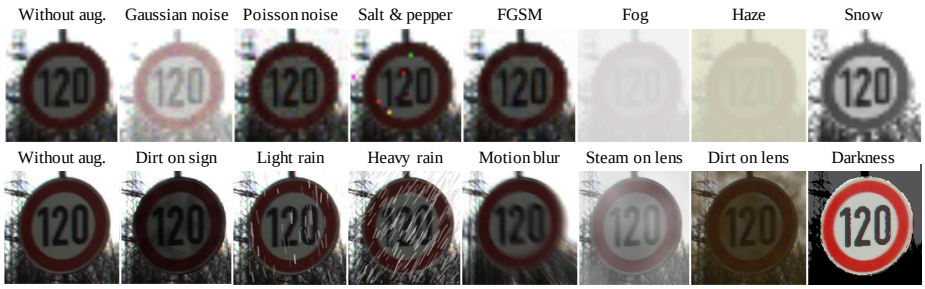

Fig. 3. Example traffic sign with augmentations from the nn-dependability kit [13] in the first row, and augmentations from our framework in the second row.

P8 - Implement the augmentations for the quality deficits. During implementation, we need to consider how scope characteristics determine the *intensity* of the quality deficit and influence the *appearance* of the augmentation. Characteristics of quality deficits might determine colors, specific proportions of the image, shapes, etc.

In our example, we illustrate this for the quality deficit rain in Fig. 1. The appearance of the augmentation is defined by the direction and velocity of the wind relative to the driving direction and the velocity of the car, causing a slant in the raindrops (cf. also Fig. 3). Another example is that the location of a traffic sign in the forest rather than in the city will influence the color of the dirt accumulated on it, making it greenish.

Fig. 3 contrasts augmentations targeting a high degree of realism, like the ones implemented in our tooling, and artificially appearing augmentations commonly applied.

P9 - Derive conditional probabilities to quantify further sensor outputs. Finally, we need to specify how scope characteristics determine the output of previously identified sensors, including typical inaccuracies of sensor signals.

In our example, sensor data that might be simulated as part of the meta-data output of the framework is illustrated by gray shaded boxes in Fig. 1. Dotted lines indicate the scope characteristics used to simulate the respective sensor signal. For example, the value of the temperature sensor can be obtained by distorting the actual value with a Gaussian error term considering the standard error provided by the specification of the temperature sensor. The same is true for the GPS signal, which uses a Gaussian distribution with an approximated standard error of 8 meters.

A1 - Randomly sample a context vector. Realistic context information is generated by taking a sample for $p(SC_{\mathrm{V=u}} \mid TAS)$, the probability of a scope characteristic taking the value u in the target application scope TAS considering the dependencies in the context model between different scope characteristics.

Considering Fig. 1, an approach may start by sampling a time based on available statistics on when people are driving by car, then sampling a possible location based on traffic data for each point in Germany at the given time using OpenStreetMap, next sampling specific weather conditions based on location and time, etc.

A2 - Determine augmentation(s) to apply and their parameter values. In order to determine realistic accuracy of a data-driven model, data with quality deficits is created, where the intensity values of each quality deficit follow a probability distribution of their natural occurrence $p(QD_{I=x} \mid TAS)$, where $QD_{I=x}$ is a quality deficit with intensity $I = x$ in the target application scope TAS. If specific quality deficits are already present at a representative rate in the dataset to be augmented, they can be excluded from the augmentation.

Most quality deficits have certain demands on the environment in order to be present with a given intensity. Therefore, quality deficits that occur under the given scope characteristics are selected for every quality deficit $QD_1, \dots, QD_n$:

$$p(QD_{i,I=x_i} \mid SC_{1,\mathrm{V=u_1}}, \dots, SC_{\mathrm{m,V=u_m}} \,\&\, TAS), 1 \leq i \leq n\,. \tag{1}$$

For example, the likelihood and intensity value for the rain augmentation directly depends on the value of the context factor precipitation amount.

A3 - Apply augmentations and generate meta-data. In this step, an image is first randomly drawn from the available dataset containing image data. Each image is only selected once. Next, all augmentations are applied to the image with the previously determined intensity and appearance parameter values. Then the values for relevant further data sources, e.g., rain sensor, brightness sensor, GPS signal of the vehicle, are determined. Finally, the augmented image is stored along with the generated meta-data. Such data can then be used to improve model training or analyze uncertainty [22].

4 Conclusion

This paper presented a framework for image augmentation and explained how to apply it to (UC1) introduce realistic quality deficits to existing image datasets considering the typical distribution of deficits and resulting coincidences in the target application scope. It can also be applied to (UC2) sample realistic context characteristics in which a given selection of quality deficits may occur. Besides the augmented image, meta-data comprising context information and additional sensor data (e.g., from a rain sensor) is generated. A layer concept applying quality deficits in a given order from object via context to sensor-related issues allows passing relevant information to subsequent augmentations, preventing interference between multiple augmentations on the same image.

A preliminary evaluation showed that a tool prototype based on the framework in the context of traffic sign recognition provided visually authentic results on the GTSRB dataset. Although our approach allows combining quality deficits with various intensities and appearances considering the context of the image, several topics remain open to be addressed in the future.

At the technical level, the challenge of automatically deriving an object mask that identifies all pixels related to the traffic sign has not been finally solved, even though image segmentation using an adapted GrabCut algorithm provides promising results. Application UC2 is also not implemented yet. As future work, we plan to address UC2 by considering the context model as a Bayesian network and inferring the unobserved scope characteristics with stochastic MCMC simulation.

The parameters of the augmentations still need to be calibrated and validated on empirical data (e.g., which intensity value best represents 4 mm of rainfall). We also need to further investigate how well the augmented data represents the intended target application scope. This includes evaluating the coverage of relevant quality deficits and the realism of the generated images, investing the impact of the augmentations on the accuracy of data-driven component outcomes, and finally comparing the impact of the augmented quality deficits with the impact of their natural counterparts.

Acknowledgments. Parts of this work have been funded by the German Federal Ministry of Education and Research (BMBF) under grant number 01IS16043E (CrESt).

References

1. German Traffic Sign Benchmarks, http://benchmark.ini.rub.de/?section=gtsrb, last accessed 2019/02/19.
2. CrESt Project Website, https://crest.in.tum.de/, last accessed 2019/02/19.
3. Krizhevsky, A., Sutskever, I., Hinton G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems 25 (NIPS), pp 1097–1105 (2012).
4. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Staff I (ed) 2012 IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, pp 3642–3649 (2012).
5. Kläs, M.: Towards Identifying and Managing Sources of Uncertainty in AI and Machine Learning Models - An Overview. arxiv.org/pdf/1811.11669v1 (2018).
6. Dodge, S., Karam, L.: Understanding How Image Quality Affects Deep Neural Networks. arxiv.org/pdf/1604.04004v2 (2016).
7. Shrivastava, A., Pfister, T., Tuzel, O., et al.: Learning from Simulated and Unsupervised Images through Adversarial Training. In: Conf. on Computer Vision and Pattern Recognition (CVPR), pp 2242–2251. Honolulu, Hawaii (2017).
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems 27 (NIPS) (2014).
9. Wong, S.C., Gatt, A., Stamatescu, V., et al.: Understanding Data Augmentation for Classification: When to Warp? In: Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA) (2016).
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. arxiv.org/pdf/1412.6572v3 (2014).
11. Carlson, A., Skinner, K.A., Vasudevan, R., et al.: Modeling Camera Effects to Improve Visual Learning from Synthetic Data. arxiv.org/pdf/1803.07721v6 (2018).
12. Karahan, S., Kilinc Yildirum, M., Kirtac, K., et al.: How Image Degradations Affect Deep CNN-Based Face Recognition? In: Int. Conf. of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany (2016).
13. Cheng, C.-H., Huang, C.-H., Nührenberg, G.: nn-dependability-kit: Engineering Neural Networks for Safety-Critical Systems. arxiv.org/pdf/1811.06746v1 (2018).
14. Pezzementi, Z., Tabor, T., Yim, S., et al.: Putting Image Manipulations in Context: Robustness Testing for Safe Perception. In: Int. Symp. on Safety, Security, and Rescue Robotics (SSRR) (2018).
15. Luan, F., Paris, S., Shechtman, E., et al.: Deep Photo Style Transfer. arxiv.org/pdf/1703.07511v3 (2017).
16. Liu, M.-Y., Breuel, T., Kautz, J.: Unsupervised Image-to-Image Translation Networks. arxiv.org/abs/1703.00848v6 (2018).
17. UjjwalSaxena Automold - Road Augmentation Library. github.com/UjjwalSaxena/Automold--Road-Augmentation-Library, last accessed 2019/02/26.
18. Temel, D., Kwon, G., Prabhushankar, M., et al.: CURE-TSR: Challenging Unreal and Real Environments for Traffic Sign Recognition. arxiv.org/abs/1712.02463v2 (2018).
19. Harisubramanyabalaji, S.P., Réhman, S., Nyberg, M., et al.: Improving Image Classification Robustness Using Predictive Data Augmentation 11088: 548–561 (2018).
20. Climate Data Center, https://cdc.dwd.de/portal/, last accessed 2019/02/19.
21. OpenStreetMap, https://www.openstreetmap.de/, last accessed 2019/02/19.
22. Kläs, M., Sembach, L.: Uncertainty Wrappers for Data-driven Models – Increase the Transparency of AI/ML-based Models through Enrichment with Dependable Situation-aware Uncertainty Estimates. In: Workshop on Artificial Intelligence Safety Engineering (WAISE). Turku, Finland (2019).