DATA AUGMENTATION STRATEGY FOR GENERATING REALISTIC SAMPLES ON DEFECT SEGMENTATION TASK

Massimo Martini, Università Politecnica delle Marche, Ancona, Italy
Riccardo Rosati, Università Politecnica delle Marche, Ancona, Italy
Luca Romeo, University of Macerata, Macerata, Italy

Introduction

Machine Learning (ML) and Deep Learning (DL) are emerging as techniques for solving challenging tasks in challenging scenarios such as Industry 4.0 and precision farming.

Despite the advantages of these data-driven techniques, there are several disadvantages that could be overcome to scale these learning algorithms in a real-world scenario. One motivation is the difficulty of implementing an ad hoc data collection procedure to collect a supervised dataset that is highly representative of the production task to be solved. This strategy usually implies, for example, that in anomaly and defect detection tasks the number of good samples far exceeds the number of defective samples. As a result, there is a high imbalance between different classes in the classification or defect detection task. Downsampling strategies in this context aim to discard nondefect samples to maintain the balance between classes, thus drastically reducing the overall dataset. Several solutions to the class equilibrium problem have been proposed in the past, both at the data level and at the algorithm level. Data-level solutions include sampling methods such as different forms of resampling, e.g., random resampling, direct resampling (e.g., SMOTE), oversampling with informed generation (e.g., Generative Adversarial Network [GAN]). On the other hand, the algorithmic-level solution includes adjusting the costs of various classes (cost-sensitive learning), rebalancing the loss during the learning process, or automatically adjusting the threshold of the posterior classifier. The sampling method could be easily scaled in the computer vision scenario as a data augmentation method to perform global augmentation (augmentation for the whole image) or to perform local augmentation (augmentation only for the region of interest).

In this paper, a data-level solution is introduced to improve the generalization performance of semantic segmentation of surface defects. This scenario can be relevant in different domains including the precision farming and agriculture scenario. Specifically, the proposed approach includes a generative phase to simulate synthetic defects and a validation phase to verify that the synthetic image is as close as possible to the real one. With real experiments on a benchmark dataset, we demonstrated the effectiveness of our approach in a real-use scenario compared to other widely used data augmentation approaches for semantic segmentation for defect detection. The data augmentation approach allows minority classes to be balanced while improving overall generalization performance.

Methodology

The proposed approach consists of the following steps: preprocessing, data augmentation, defect validation and defect segmentation.

Preprocessing

The first preprocessing step consists of normalizing the pixels of the original images to values between 0 and 1.

Next, the images are divided into square patches of 512x512 pixels. The rationale behind image splitting and patch processing is mainly due to the main objective of the defect detection task, which is to maintain high image resolution, leading to very accurate and fine-grained segmentation.

Finally, a filtering step was added to retain only patches containing a distribution of defects. Since most of the image area mainly contains the background class, this step allows for a much more balanced dataset and to focus only on the candidate area where defects are present.

Data augmentation

Two different approaches to perform data augmentation were considered: the first generates Gaussian defects completely independent of the dataset, while the second generates defects from the information obtained from defects already in the dataset.

Gaussian defects are applied randomly within the images of the dataset. They are configured according to several parameters: height and width of the two perpendicular Gaussian features, rotation angle, magnitude of the effect, type of defects

In contrast, the defects in the second approach are much more realistic because they take advantage of the defect information already present in the dataset. Given a generic defect and its mask, the entire portion of the defect is extrapolated. Then, in relation to the surrounding background, the defect is subtracted from it to obtain only the magnitude of the defect on the image. This "magnitude" can then be applied anywhere on another image. Defect generation was done using an algorithm that allowed a generated defect to be inserted within the images of the dataset, but without overlapping with existing ones.

For the generation of defects in the second approach, the following steps were followed: a defect is randomly chosen from existing ones; then the defect is extracted from its background, also based on its type; then, it is inserted into a random location of a random image, without obtaining overlaps with those already present.

Before insertion, an additional processing step is carried out to obtain an increasingly realistic defect. To the extracted defect, the following will be added:

- a Gaussian distributed additive noise;
- a Poisson-distributed noise generated from the data;
- a multiplicative noise
- no noise.

Since the former data augmentation approach allows the number of defects to be significantly increased, the latter ensures that all generated defects are as similar as possible to the original ones. Consequently, the proposed approach of validating the generated defects allows all unrealistic defects to be filtered out.

Defect validation

If the generated defects do not reflect the distribution of the real ones, instead of improving segmentation, they may lead to the opposite result. For this reason, it is necessary to use a method that can validate the performed data augmentation procedure and thus the generated defects, so as to discard low-quality defects that deviate from the original ones.

The approach proposed in this paper is based on the concept of Siamese Networks, which allows learning a similarity function that, given two inputs, can evaluate how similar they are, as a function of a numerical value.

Siamese networks are neural networks that share weights between two or more sister networks, each of which produces latent vectors of their respective inputs.

In supervised similarity learning, networks are trained to maximize the contrast (distance) between latent vectors of inputs of different classes and minimize the distance between embeddings of similar classes, resulting in latent spaces that reflect the class segmentation of the training inputs.

The Siamese network in this context is used to understand the quality of the generated defects, so that only the "best" ones (closest to the real ones) are filtered out and then used for training the segmentation network.

Defect segmentation

The network for semantic defect segmentation used in this work is based on the well-known UNet network. The network was pre-trained on the Imagenet dataset and then re-trained with the dataset proposed in this paper. The loss function used for training the network is a combination of two very common losses in the semantic segmentation task (Dice and Focal).

Results

For the evaluation of the proposed method, a new dataset was collected using an RGB camera. A single view of several plastic objects was acquired with the main purpose of detecting structural defects on rubber materials.

Table 1 describes the results of the defect segmentation model in terms of IoU Score and F1-Score using the original dataset (Original), the dataset augmented with the proposed data augmentation (DA), and the dataset augmented and validated according to the proposed methodology (DA+Validation), respectively.

	IoU score	F1 score
Baseline	82.11	83.97
DA	82.12	84.19
DA + Validation	84.10	85.61

Tabella 1: Results of the proposed DA methodology

The extracted results highlight the potential of the proposed methodology in the specific use case (segmentation of plastic material defects) and the possibility of generalization also in a different scenario such as precision farming and agriculture.

Reference

Jain, Saksham, et al. "Synthetic data augmentation for surface defect detection and classification using deep learning." Journal of Intelligent Manufacturing (2022): 1-14.