Data Requirements for Robust Machine Learning in High Variance Industrial Environments

Abraham Ghanem
Swisslog GmbH
abraham.ghanem@swisslog.com

Abstract— Machine Learning (ML) based industrial applications deployed in high variance dynamic environments present a new set of challenges. The performance of such systems is directly linked to the nature of the data it has been subjected to. Using the computer vision-based ML applications in the logistics industry as a case study, due to their high variance environment and strict requirements, we try to address the issue of understanding the data requirements for the successful development and deployment of such applications. We propose a systematic approach to address high variance scenarios with limited relevant data availability, covering both real data collection and synthetic data generation, highlighting their requirements and effective utilization methods.

1 Introduction

We aim to address the challenges of labeled data sparsity in ML-driven computer vision tasks such as unknown object detection, segmentation, and pose estimation, particularly focusing on the requirements for training, testing, and deploying ML based vision tasks. Our main objective is to explore diverse data generation approaches and their requirements for the robust deployment and maintenance of such data driven applications. Defining the data requirements in such applications is one of the key challenges being addressed, as the requirements range from requirements set by laws defining the usage of such data, to requirements set by the application itself.

To achieve this, we should be able to clearly define the requirements from top to bottom, starting with the application and moving on to the model, and data [12] to be used. Our main focus will be mainly but not exclusively directed towards the data requirements of such tasks, and the state of the art approaches to address these issues, especially in sparse data environments. This includes the requirements for the different data generation methods for vision tasks along with the requirements for their generated data starting from the number of samples, their quality, and variance all the way down to the specifics such as the object shapes, sizes, colors, quantities, backgrounds, and arrangements in the generated scenes. We discuss the input data that these methods should handle and the tasks to be performed reliably, such as augmenting given scenes, creating new objects to fill data gaps, and generating novel relevant data in cases with data restrictions [1] or sparse environments. This approach, contributes to bridging the gap in the requirements necessary for synthetic data generation for enhancing the performance ML models in real-world scenarios.

2 Background and Related Work

Machine Learning(ML) systems in the industry are still not understood well [5], this often results unrealistic requirements and expectations from the customers side. As these approaches are mostly unique due to their data-driven nature, they require new types of requirements that consider the performance on a given task [5]. As discussed in [1] it is important to define the requirements inside the context of the given task, followed up by some quality metrics and performance monitoring specific to the task itself.

A. Requirements Engineering for ML from an Industrial Perspective

The automation industry has always been leaning towards high performance reliable solutions that require as little intervention and maintenance as possible to keep on running. This results in a stable process with minimal costs and reduced downtime. With the recent advancements in the field of ML, it is now possible to automate more repetitive processes with a non-deterministic decision pattern requiring a human to perform the task. As humans have a lifetime of experience at their disposal and a strong contextual understanding of the task, they can perform tasks in a reliable manner without defining requirements on a very detailed level. As an example, a human performing a commissioning task in the logistic industry, gets a task to pick a certain quantity of item A from a place X and place it into Y. If the information is correct the human should not have any problem performing such a task, and the time required to do it is also more or less known. But giving the same task to a Robot using ML will raise an large number of questions that have previously been considered as "obvious" to a human. By first roughly answering these questions one can draw the outline of the requirements the system has to fulfill to perform this task successfully, for example:

1. Accurate and reliable item detection and localization. The ML algorithm should be able to differentiate between the desired objects and all other objects belonging to the environment.
2. Finding stable and reliable grasp points on the required items that lead to a successful picking process without damaging the item.
3. From a container with several items of the same type, decide which one is the easiest to pick first to decide on the order in which items are going to be picked.
4. Plan and perform the desired motion.
5. Choose the best placement pose and location as this can be described as an optimization problem where the actor performing the action aims to achieve a compact and stable item placement while making sure that items will not get damaged during the transportation process.

Beside the mentioned high level requirements there are the reliability and process intervention requirements, as the deployed model should be able to generalize on such a wide range of diverse unseen items in a warehouse as possible without the need of continuously turning to the developers for updating the models every time new products and packaging are introduced, and as little as possible human intervention in case of emerging errors.

B. Challenges in Data Requirements for ML-Driven Logistics

As an example for demonstrating the data requirements issue, we use ML driven Computer Vision algorithms in the logistic industry as a case study. In this example the ML algorithm tries to perform the commissioning task mentioned above in a warehouse with thousands of diverse products. In order to better understand the requirements of such a task, we start by demonstrating the current challenges facing pre-trained ML approaches in the field of Computer Vision.

1. Pre-trained ML models in industrial applications demonstrate a significant decrease in performance when faced with novel data when trained inadequately.
2. Constructing large labeled real data-sets is a very time consuming and costly process.
3. Definition of data requirements like data quality, quantity, diversity, statistical balance, and data collection constraints.
4. Using synthetic data requires a definition of the ratio of the mix between real and synthetic data, along with additional quality metrics beside the ones for real data in order to close the Domain Gap [6] introduced by using simulated data.

3 Data Requirements for ML-Driven Industrial Applications

To address the challenges mentioned in Section 2, the following has to be considered:

A. Defining Data Requirements

Defining data requirements for ML-driven industrial applications is a critical process, focusing on the precise determination of data criteria. To establish and articulate these data criteria, several key considerations come into play.

Firstly, in supervised learning, 'necessary labels' are important for effective model training, making it essential to specify and define these labels accurately. Moreover, the determination of the required number of samples depends on the application's scope. The broader the range of scenarios and variances that the algorithm must handle, the more extensive the data-set needs to be, and a larger model with more parameters will be needed.

For industrial applications in computer vision for object manipulation, high level data variance arises from scene views, colors, lighting conditions, backgrounds, and environments. This variation must be embedded in the data, even encompassing extreme cases, to ensure the model's robustness. Moreover, a balanced distribution of data across various scenarios is essential to prevent bias. Within a single scene, where accurate object detection is the goal, the data-set should further introduce variance through variations in object shapes, sizes, colors, and placements. As a result of these requirements, along with the need to empirically control the attributes of the data-set, combined with the financial, temporal, and practical constraints associated with real data collection, synthetic data is often considered as the main source of labeled training data. Therefore, understanding the requirements for generating and using synthetic data is essential.

Due to our limited ability to accurately simulate real data, a domain gap [6] is introduced when using synthetically generated data. This means that algorithms trained exclusively with synthetic data might face difficulties generalizing in real life scenarios. In order to balance this, it is required to have a limited availability of real data to be used in combination with the synthetic data.

B. Synthetic Training Data Generation

To be able to further define the requirements for synthetic data generation, we provide a short overview of the methods available in the context of Computer Vision along with a promising novel approach currently being researched.

- High fidelity simulations using 3D object models: Using high fidelity simulators [2] with CAD models or the 3D models of real scanned objects [3, 11] provides representative and realistic data for training vision models, making the generated synthetic data more applicable to real-world scenarios.

- Generative data augmentation: Given a limited amount of real data, Image-to-image data augmentation [1, 9, 10, 13, 15] could be utilized to expand the data-set mitigating the costs of additional data collection and the same time generating data very similar to real data reducing the domain gap. Additionally, advanced data augmentation methods can be applied for influencing the data generation process using text prompts to target specific data patches [4, 8, 9].

- Generative labeled data generation: unlabeled data generation using generative models has been around for some time, however so far researchers are investigating the best usage for such models to obtain novel application relevant labeled data [14].

C. Requirements of the Training Data

Given the methods mentioned in B, our focus is on the requirements of the data-set itself, as a result we define two sets of requirements, the first focuses on the smaller real data-set collected to serve as a base for the larger synthetic data-set, where the second focuses on the data generated by one of the above mentioned methods given a limited sample of real

data. The following requirements serve as addition to the requirements mentioned in section A.
Requirements of the real/ initial data-set:

- The data should cover as large incremental variation as possible stretching between the extremes in order to be utilized later to cover the gaps between the different cases.
- The data should be statistically balanced, meaning different scenarios should be equally represented in the data-set without a bias as this would transfer to the generated data and then to the model.
- Part of the data should be reserved to be utilized exclusively in the model testing process.

Requirements of the synthetic/ generated data-set:

- Should perform data augmentation on the input initial data while preserving the quality of the data.
- Accurate data labels/ annotations should be generated along with the data itself.
- The generated data should be controllable and statistically analysed to identify biases.
- The generated data should cover a much wider range of variation than the real data in order to compensate for the data sparsity.
- The generated data should preserve some aspects of the input data but should introduce enough novelty to make it possible for the model to learn new information. For example a scene with an apple is augmented to contain more apples, or two scenes with different objects are used to generate a scene with both of them or a merging of the attributes of the scenes to produce new objects.
- Generating a ratio of 9:1 of synthetic to real data, as the mix of 10% real to 90% synthetic data is commonly used in transfer learning tasks in the field of learning based Computer Vision [7].

4 Conclusion

In conclusion, this work mainly addresses the challenge of defining data requirements in the context of industrial ML for computer vision. The establishment of the data requirements is necessary to successful ML model deployment in real-world industrial applications. We focus on the unique demands of defining the data requirements, particularly in scenarios where collecting real labeled data is impractical. We also provide insights into various data generation methods in computer vision, such as high-fidelity simulations and generative data augmentation and generation, to assist practitioners in their method selection.

Furthermore, we've outlined the requirements for both real and synthetic data, including factors like diversity, balance, data augmentation, accurate labeling, novelty introduction, and bias control. These requirements are essential in ensuring the efficacy and reliability of ML models in real-world industrial applications.

Essentially, this research offers a thorough insight into the core challenge of outlining data requirements in industrial machine learning-driven computer vision. The aim is to assist practitioners in methodically setting these requirements, thereby boosting the effectiveness of machine learning applications in industrial settings and refining their overall performance and efficiency.

References

[1] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023.

[2] Epic Games. Unreal engine.

[3] Anas Gouda, Abraham Ghanem, and Christopher Reining. Dopose-6d dataset for object segmentation and 6d pose estimation, 2022.

[4] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and

Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2022.

[5] Hans-Martin Heyn, Eric Knauss, Amna Pir Muhammad, Olof Eriksson, Jennifer Linder, Padmini Subbiah, Shameer Kumar Pradhan, and Sagar Tungal. Requirement engineering challenges for ai-intense systems development. *CoRR*, abs/2103.10270, 2021.

[6] Benedikt T. Imbusch, Max Schwarz, and Sven Behnke. Synthetic-to-real domain adaptation using contrastive unpaired translation. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE, aug 2022.

[7] Farzan Erlik Nowruzi, Prince Kapoor, Dhanvin Kolhatkar, Fahed Al Hassanat, Robert Laganière, and Julien Rebut. How much real data do we actually need: Analyzing object detection performance using synthetic and real data. *CoRR*, abs/1907.07061, 2019.

[8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[10] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *CoRR*, abs/2111.05826, 2021.

[11] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. *CoRR*, abs/1908.04616, 2019.

[12] Andreas Vogelsang and Markus Borg. Requirements engineering for machine learning: Perspectives from data scientists. *CoRR*, abs/1908.04674, 2019.

[13] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models, 2022.

[14] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models, 2023.

[15] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations, 2022.