Model Performance: Large Artificial Dataset vs. Small Real Dataset

Context:

Currently, our AI team is training a classifier model to detect whether objects on the Mars Yard should be considered uncommon according to human expectations. However, we don’t have enough data to train our classifier accurately, so we decided to try generating artificial data using simulation engines like Isaac Sim. This led us to wonder about the impact of dataset size and authenticity on model performance. Therefore, our project aims to explore this dilemma by comparing a model trained on an extensive synthetic dataset against one trained on a smaller, but real, dataset.

Project Description:

This project seeks to investigate whether a machine learning model trained on a large artificial dataset can surpass the performance metrics of a model trained on a smaller, yet authentic, dataset. By analyzing the trade-offs between dataset size and authenticity, the project aims to elucidate the effectiveness of synthetic versus real data in model training. The objective is to comprehensively assess the viability and efficacy of using synthetic data against the backdrop of limited real data for training machine learning models.

Tasks:

- Train the current model on different ratios of real and synthetic data
- Implement a different model that might be more suitable for this kind of tasks and would be better at generalization

Contact:

Name	Yasmin Ben Rahhal	Emile CHARLES
Position	Team Leader AI	Project Manager Research
Pole	EPFL Xplore Research (XRE)	EPFL Xplore Research (XRE)
Email address	yasmin.benrahhal@epfl-xplore.ch	emile.charles@epfl-xplore.ch
Mailing list	artificial.intelligence@epfl-xplore.ch	