Dept. of Computer Science
Télécom SudParis
IP Paris, France

IPParis HPDA/PDS Master projects 2024-2025

From Concepts to Pixels: Training AI Vision Models Without Real Data!

September 11, 2024

Project Description

In many fields, especially the Internet of Things (IoT), the scarcity of large, labeled datasets presents a significant bottleneck for training machine learning models, particularly for image-based tasks. Acquiring and labeling real-world data is both costly and time-consuming. This project explores an innovative approach to overcoming these challenges by developing a pipeline that generates synthetic image datasets using advanced AI models.

At the core of the project is a novel NGSI-LD data model, which defines the characteristics of the data, such as class labels, environmental context, and image attributes. These characteristics will be fed into transformer-based models (e.g., DALL-E, Stable Diffusion) to generate a small set of highly specific images. Once this initial dataset is produced, Generative Adversarial Networks (GANs) will be employed to significantly expand the dataset by synthesizing a large number of relevant images. The resulting expanded dataset will then be used to train image classification models, which will subsequently be evaluated against real-world datasets to gauge their performance.

The project will also involve a comparative study of various text-to-image models and prompt engineering techniques, aimed at improving the quality and relevance of the generated images. Students will assess the performance of models trained using synthetic data versus real-world data, focusing on metrics such as final accuracy, robustness, and cost-effectiveness.

By participating in this project, students will gain hands-on experience with cutting-edge AI technologies and engage in research that combines data modeling, synthetic data generation, and machine learning evaluation.

Project Objectives

The main objective of this project is to develop an end-to-end pipeline that automates the generation of synthetic datasets for training image-based machine learning models. This pipeline will encompass the entire workflow—from defining the data characteristics using an NGSI-LD model to generating an initial set of images using transformer models, and finally scaling the dataset with GANs for large-scale model training. The ultimate aim is to create a flexible, cost-effective solution for training models without real-world data, while ensuring high accuracy and relevance for practical applications. The objectives include:

- Design and implementation of an NGSI-LD data model to define key characteristics for image generation, including class labels, scene settings, and other environmental factors relevant to the machine learning task.
- Research and apply different prompt engineering techniques to optimize how the NGSI-LD data model interacts with text-to-image models for producing high-quality, contextually relevant images.
- Explore and compare transformer-based text-to-image models (e.g., DALL-E, Stable Diffusion) to generate small datasets based on the NGSI-LD model, including a focus on fine-tuning models for better output through prompt engineering.
- Utilize GANs to generate large synthetic datasets from the initial set of images, and investigate how the size and diversity of GAN-generated images influence the performance of machine learning models.

- Evaluate the performance of machine learning models trained on synthetic datasets versus models trained on real data, with a focus on metrics such as accuracy, robustness, and training costs.

Skills & Qualities

- Strong motivation to work on a cutting-edge research project in AI
- Fluency in English.
- Proficiency in Python.
- Familiarity with deep learning frameworks like TensorFlow or PyTorch is a plus but not required.
- Knowledge of Generative Adversarial Networks (GANs) and text-to-image models is a plus but not required.
- Basic understanding of data modeling concepts and interest in learning about semantic data modeling.

Supervisors

Georgios Bouloukakis, `georgios.bouloukakis AT telecom-sudparis.eu`
Nikolaos Papadakis, `nikolaos.papadakis AT telecom-sudparis.eu`