Automating Insight Extracting from ML Visual Data with Large Language Models

Arumoy Shome

Introduction

Visualisations are employed at various stages of a Machine Learning (ML) pipeline. They are used to understand and verify data properties during the early stages, summarise metrics and fine-tune models during development, and monitor performance post-deployment. The iterative and experimental nature of building ML systems heavily relies on insights from visualisations to guide design and implementation decisions [1, 4, 5, 7–9].

However, real-world data that ML systems encounter post-deployment seldom remain static. They often change as a reflection of the world, potentially violating initial assumptions made during development [2, 6]. Every subsequent iteration of the ML development cycle used to retrain and update the ML model, therefore demands manual validation of the visualisations that were used to test ML system properties.

Assertions or analytical tests derived from ML visualisations can significantly reduce manual verification efforts. Such formal assertions record the AI practitioner's observations about the model or data at a specific moment. They also serve as a reference point for future AI practitioners to understand the interpretations made from earlier visualisations.

In a prior study, we mined $54,070$ Jupyter notebooks from Github and created a high-quality dataset of 269 semantically related visualisation-assertion (VA) pairs. The input feature space comprises of a rich source of information comprising of visualisations, Python source code, and associated markdown text.

This research project will focus on developing automated tools using state-of-the-art Large Language Models (LLMs) that aid ML practitioners when working with visualisations. The research questions for this project are categorised into 3 distinct tracks and presented below.