Bridging the Gap Between Visual and Analytical Machine Learning Testing

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Testing ML systems is a highly interactive process which demands a human-in-the-loop approach. In addition to writing tests for the code base, practitioners are required to analyse and interpret several visualisations using their domain expertise to validate if an ML system satisfies the required set of functional and non-functional properties. Visualisations are frequently used to qualitatively assess various parts of an ML pipeline. However, implicit knowledge gained from visualisations must be translated to explicit analytical tests that fail when there is change in any component of the ML pipeline. We conduct an empirical analysis of Jupyter notebooks to catalogue the state-of-the-art mappings between ML visualisations and assertions. We mine Github to collect 54K Jupyter Notebooks that contain assertions written in Python. We develop a novel methodology to identify 1764 notebooks which contain an assertion below a visualisation. We manually analyse the 1.7K notebooks and identify 269 visualisation-assertion pairs that are semantically related to one another. We further investigate the 269 visualisation-assertion pairs and identify three frequently occurring testing patterns. We perform an in-depth analysis of 34 visualisation-assertion pairs and find that the assertions often fail to capture all the information present in their visual counter-part. Empirical evidence obtained in this study indicates that current software testing methods fail to address the unique challenges of ML. And emphasises the need for automated tools that bridge the gap between visual assessments and analytical assertions.**

***Index Terms*—ML Testing, SE4AI, Visualisations, Assertions, Computational Notebooks**

I. Introduction

Visualisations are used throughout the ML development lifecycle to test various properties of the ML system. In the early stages of the ML development lifecycle, visualisations are extensively used to make sense of the data and verify its statistical properties. During the model development phase, visualisations are used to summarise metrics, contrast different learning algorithms and iteratively fine-tune ML models. Once the model is deployed in production, visualisations are used to continually monitor their performance and trigger a new training cycle once their performance drops below a certain threshold [1]–[4].

Building ML systems is highly iterative and experimental. Information gained from ML visualisations is used to make design and implementation decisions for the following steps of the ML pipeline. For instance, we may visualise the distribution of our training data and find that it is normally distributed. Based on this information, we may opt for a Linear Regression model which assumes normality in the underlying data. However, such expectations regarding the data may be violated once the ML system is deployed in production, where the data constantly changes as a reflect of the real world [5]–[8].

Implicit expectations obtained from visualisations must therefore be translated to analytical tests that fail once our expectations regarding the ML system are no longer satisfied.

In contrast to prior work, we approach ML testing from a new perspective. We conduct an empirical analysis of computational notebooks obtained from Github, to understand the process of testing ML systems in practice. In particular, we focus on the combination of a qualitative form of testing (using visualisations) and a quantitative form of testing (using analytical assertions). The research questions along with the contributions of this paper are as follows.

RQ1. **How frequently are analytical tests formulated from visualisations created to test ML systems?**
We mine 54K Jupyter notebooks from Github that contain an assertion written in Python. We develop an automated technique to identify 1764 notebooks with an assertion below a visualisation. We manually analyse the 1.7K notebooks and identify 269 visualisation-assertion (VA) pairs that are semantically related to each other. We plan to release this catalogue of VA pairs publicly.

RQ2. **What patterns are frequently observed in VA pairs used to test ML systems?**
We analyse the 269 VA pairs from RQ1, and observe three testing patterns that frequently occur in the VA pairs.