Projects Proposal: Quality Assurance of AI Solutions

Software Engineering for Artificial Intelligence
fpalomba@unisa.it

Projects Scope

Quality assurance techniques are essential for managing AI-enabled systems that aim to identify and handle quality issues. These quality issues can severely impact various aspects of these systems, such as maintainability and security. Through the different phases of the AI software development lifecycle, developers must consider several types of quality issues and guides to avoid suboptimal algorithms and low-quality datasets (i.e., AI Technical Debt) as well as underestimating the possibility of biased data and adversarial attacks (i.e., AI-Specific vulnerabilities). Addressing these issues requires proactive QA strategies, including rigorous testing, continuous monitoring, and robust security protocols. By prioritizing quality assurance efforts, organizations can strengthen AI-enabled systems' reliability, efficiency, and security, promoting trust and responsible innovation.

Starting Assets

- B. Biggio Battista and F. Roli, 'Wild patterns: Ten years after the rise of adversarial machine learning', 2018, 10.1016/J.PATCOG.2018.07.023
- A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, 'A survey on adversarial attacks and defences', 2021,
- Recupito, G., Rapacciuolo, R., Di Nucci, D., and Palomba, F. "Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality." 2024. https://gilbertrec.github.io/personal_doc/publications/C2.pdf
- Bogner, Justus, Roberto Verdecchia, and Ilias Gerostathopoulos. "Characterizing technical debt and antipatterns in AI-based systems: A systematic mapping study." 2021 IEEE/ACM International Conference on Technical Debt (TechDebt). IEEE, 2021.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572 [1]
- Sanglee Park, Jungmin So. On the Effectiveness of Adversarial Training in Defending against Adversarial Example Attacks for Image Classification. https://www.mdpi.com/2076-3417/10/22/8079 [2]

Project Example 1

Recent progress in machine learning has revealed how algorithms can be vulnerable to adversarial attacks, marking a significant concern in adversarial machine learning. This area focuses on identifying and addressing weaknesses in machine learning models to protect them against such attacks. It's crucial to integrate vulnerability assessment and security measures into the design of machine learning systems from the beginning, following a security-by-design philosophy. This method aims to solidify the core security of these systems against the changing nature of cyber threats, thereby enhancing the reliability of machine learning applications across different fields.. [1]

Project Example 2

The rapid advancement in AI underscores the importance of protecting neural networks from advanced adversarial attacks like those using the Fast Gradient Sign Method (FGSM), which threaten their integrity by subtly altering input data. To counteract these threats, there's a growing emphasis on developing defense mechanisms, particularly through adversarial training, to bolster the models' resilience by training them with adversarially generated examples. This initiative aims to ensure machine learning applications maintain their accuracy and reliability despite evolving cyber threats, thereby enhancing overall quality assurance in AI and fostering the creation of inherently secure systems. [2]

Minimum Requirements

1. **Road Sign Dataset:** Acquire and prepare a dataset of road signs containing a wide range of real road sign images.
2. **Data Engineering Application:** Utilize data engineering techniques to prepare and normalize road sign images, ensuring uniformity and consistency in the data.
3. **Development of a Neural Network:** Design and train a convolutional neural network (CNN) to recognize road signs from the prepared dataset.
4. **Creation of Adversarial Images with FGSM:** Implement the technique of generating adversarial images using the Fast Gradient Sign Method (FGSM) provided by TensorFlow, to create perturbed images that deceive the CNN model.
5. **Results Analysis:** Conduct a detailed analysis of the results obtained from the CNN model when exposed to perturbed images, evaluating its ability to correctly recognize road signs despite the perturbations.
6. **Development of Robustness Solutions:** Identify and implement strategies to significantly improve the model's resilience to perturbations, aiming to maintain an accuracy score of at least 95%.
7. **Creation of a User Interface (GUI):** Dedicate time to designing and implementing an intuitive and user-friendly graphical user interface (GUI), allowing users to easily visualize changes in the model and the actions required to address both the perturbation and correct prediction phases.

Ideas

The integration of Convolutional Neural Networks (CNN) has significantly enhanced the reliability of object recognition in the surrounding environment. However, these systems are vulnerable to adversarial attacks such as the Fast Gradient Sign Method (FGSM), which can subtly manipulate inputs, detrimental to AI performance, in both white-box and black-box scenarios. The core objective of our project is to develop and refine CNN architectures tailored for traffic sign recognition, with a strong emphasis on defense against FGSM attacks. By training a primary CNN for accurate recognition and subsequently subjecting it to adversarially perturbed images generated via FGSM, we aim to cultivate a secondary CNN capable of maintaining high recognition accuracy even in adverse conditions. This endeavor is particularly crucial in the context of autonomous driving, where misinterpretation of traffic signs can lead to extremely hazardous situations. Misunderstood traffic signals have the potential to cause severe or even fatal accidents. Thus, ensuring the resilience of neural networks utilized in autonomous driving systems is paramount for road safety and fostering trust in the future of transportation automation.

Award Criteria

Adversarial Patch Attack: implementing and demonstrating a defense mechanism that can detect and neutralize the influence of adversarial patches within an image. This could involve developing algorithms that recognize and ignore such patches or applying preprocessing techniques to nullify their effects before classification