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 advancements in machine learning have underscored the susceptibility of algorithms to adversarial attacks, highlighting a critical area of concern within the sphere of adversarial machine learning. This emerging field focuses on identifying vulnerabilities inherent in machine learning models and developing robust countermeasures to safeguard against such threats. Recognizing and assessing these vulnerabilities are imperative steps in the process of designing secure machine learning systems. The aim is to weave vulnerability assessment and the determination of security levels into the fabric of machine learning models from the outset, adhering to the principles of security by design. This approach ensures the foundational security of systems and services, addressing the evolving landscape of cyber threats and fortifying the trustworthiness of machine learning applications in various domains. [1]

Project Example 2

The evolving landscape of artificial intelligence has illuminated the criticality of defending neural networks against sophisticated adversarial attacks, such as those executed via the Fast Gradient Sign Method (FGSM). Such attacks subtly manipulate input data to deceive models, posing a significant threat to their integrity and reliability. In response, the development of defense mechanisms that incorporate adversarial training techniques has become a pressing endeavor. This initiative focuses on enriching the training datasets with adversarially generated examples to enhance the model's ability to identify and resist these manipulations. The goal is to preserve the accuracy and reliability of machine learning applications, ensuring they remain robust in the face of evolving cybersecurity threats. This approach not only addresses immediate vulnerabilities but also contributes to the broader discipline of quality assurance in AI, promoting the development of systems that are secure by design and capable of withstanding adversarial scrutiny. [2]

Minimum Requirements

- **Adversarial Image Generation:** A set of procedures for generating adversarial images using FGSM to train the defensive CNN model.
- **Model Accuracy:** The initial CNN model must achieve at least 95% accuracy on a non-perturbed validation dataset, with the defensive model aiming for no less than 90% on perturbed images.
- **Inference Time:** Capability of making real-time predictions with an inference time not exceeding 50 milliseconds per image.
- **Adversarial Attack Resilience:** The defensive model should show improved resilience to FGSM attacks compared to the baseline model.
- **Development of a User Interface:** A pivotal component of the project involves the creation of a user-friendly web interface. This platform will serve as the gateway for users to interact with the model, enabling seamless submission of images for classification and presenting an intuitive visualization of the model's predictions and defenses against adversarial manipulations.

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