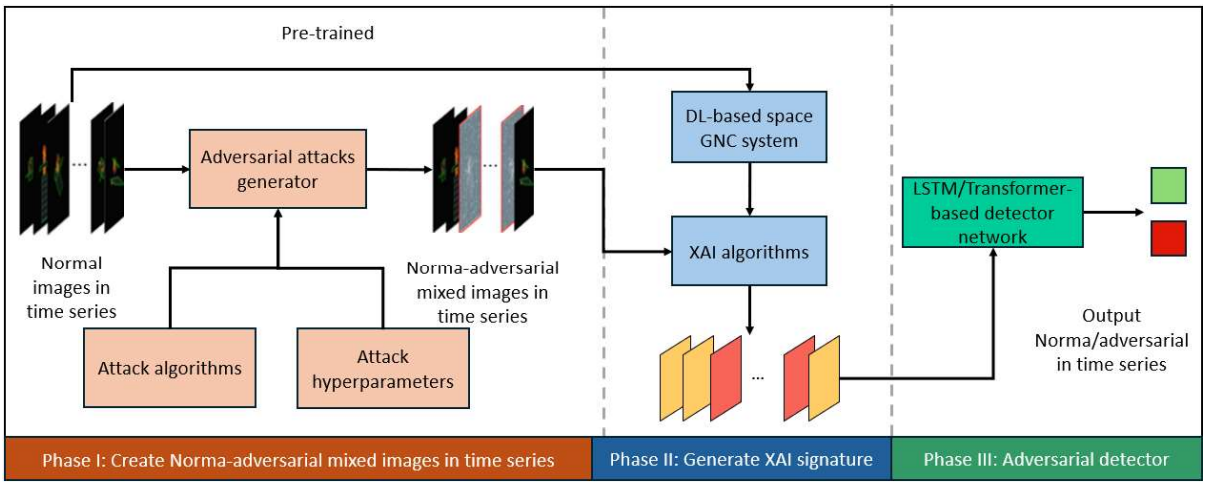

Explainable Secure Deep Learning Software for Spacecraft GNC Systems

Executive summary

Early technology development

New concepts for onboard software development

Affiliation(s): City, University of London

Activity summary:
Cyberattacks in space pose serious risks for ground based critical infrastructure, and insecurities in the space environment. Thus, an important consideration to preserving future critical space mission objectives is a successful detection to an adversarial attack on Artificial Intelligence (AI) based space software solutions. We harness the eXplainable Artificial Intelligence (XAI) techniques with adversarial learning for adversarial attacks detection, specifically on the AI-based spacecraft Guidance Navigation and Control (GNC) systems. In this work, we firstly develop an XAI based deep learning model providing the performance required for the GNC scenarios proposed. Then, an XAI adversarial learning method that handles the challenging detection through classification of adversarial attacks. Finally, we conduct extensive validation of the proposed architecture based on simulation and real data.

Publishing Date: 25/11/2024
Contract Number: 4000140064
Implemented as ESA Initial Support for Innovation

→ THE EUROPEAN SPACE AGENCY
ESA Discovery & Preparation
From breakthrough ideas to mission feasibility. Discovery & Preparation is laying the groundwork for the future of space in Europe
Learn more on www.esa.int/discovery
Activity Page: https://activities.esa.int/4000140064
Deliverables published on https://nebula.esa.int

1. Introduction

In this project, we harness the XAI techniques with adversarial learning for adversarial attacks detection that is potentially applicable to various satellite systems, and more specifically on the embedded GNC module of those space vehicles adopting deep learning algorithms. The XAI algorithms are investigated to make onboard adversarial learning transparent to ensure a trustable decision, providing highly precise detection and defensive response to ensure the space vehicle safety. A comprehensive framework is studied and proposed to address the XAI-based adversarial learning for spacecraft GNC system, including synthetic dataset generation, guidance scenario building, XAI-based adversarial-learning model developing, software verification and testing, etc. Considering that adversarial attackers can force many deep learning algorithms to misbehave by adding small and imperceptible perturbations on the original inputs to generate adversarial examples, deep network defence can be classified into two categories: active and passive defences. The former implies that a model can correctly classify inputs that are perturbed by adversarial adversaries by hardening its network. In contrast, the later aims to detecting and rejecting adversarial examples. In our study, we first follow the second idea. We developed two deep learning GNC systems for space rendezvous and planetary landing scenario, respectively. Then developing an XAI-based adversarial learning, through training, method that handles the challenging classification of adversarial attacks on the designed and input distribution shifts with a good explanation of results. and finally, we focus on optimising the network detector scheme to improve the detection accuracy and detect adversarial on real experimental setup.

2. Project Background

Cyberattacks on aerospace systems are becoming a growing concern in space missions, although they are often unpublished to attempt to delay or avoid further hacking. Jet propulsion Laboratory (JPL) reports seven times hacks against space systems during 2007—2016, for example Goddard Space Flight Centre (GSFC), Glenn Research Centre (GRC), and Armstrong Flight Research (AFRC) reported hacks of drones that their data and commands are hacked in January 2016; attackers penetrated JPL's network and took over some services, impacting missions in November 2011. Many real-world attacks have been shown to be surprisingly effective, which has raised serious concerns for space missions and been a legitimate threat to in-orbit spacecraft operations. Moreover, satellites and other space assets are parts of the digitized critical infrastructure that are crucial to support communications, transportation, information services, weather and environmental monitoring and defence systems. Consequently, cyberattacks in space therefore pose serious risks for ground based critical infrastructure, and insecurities in the space environment.

Thus, an important consideration to preserving critical space mission objective is successful detection of adversarial activities when affecting onboard sensors which are employed by AI based GNC algorithms for the space vehicle autonomous decision making. The research work in this area is very new and we are pioneering in it. Many works are related to adversarial learning for object recognition in imaging datasets but none (up to our knowledge) of the adversarial research is dedicated to GNC systems. The proposed work is based on our unique first initial research that we did in this domain for a

terrestrial GNC application, and we are keen to adopt it and adapt it to space GNC application and to further mature it.

3. Methodology

In this project, two scenarios are proposed to investigate the impacts of adversarial attacks and relative detection mechanisms on AI-based autonomous space GNC systems. The first target scenario is Convolution Neural Network (CNN)-based relative pose estimation on close-range rendezvous and the second scenario is vision-based Deep Reinforcement Learning (DRL) for planetary landing guidance and control.

For the CNN-based relative pose estimation on close-range rendezvous scenario, firstly, a 3D simulation system has been developed on Blender software to provide representative visual images in deep space rendezvous environments. The simulator is aiming to render the camera view of target spacecraft to generate reliable synthetic images for training and validating the AI-based pose estimation algorithm under simulated deep space environment. Then, a CNN-based spacecraft relative pose estimator is newly designed with the aim of providing a reliable estimated position and attitude of the target spacecraft in as rendezvous scenario. Consequently, the Fast Gradient Sign Method (FGSM) adversarial attacks are adopted on the spacecraft onboard camera resulting in an adversarial image to evaluate the impacts on the proposed deep pose estimator. Next, SHAP values are employed to generate XAI signatures for both adversarial and normal input images in designed CNN-based relative pose estimator. Finally, a Long Short-Term Memory (LSTM)-based adversarial detector is proposed and trained, which learns normal and adversarial SHAP values to detect the adversarial attacks on the spacecraft relative pose estimator.

For the Mars landing scenario, a 3D simulation environment is developed on Blender software. The Mars landing simulator consists of two main components: an optical data generator and a 3 Degree-of-Freedom (DOF) lander dynamics. The 3 DOF controller takes the engine actions command as its inputs and outputs the relative position of the lander, while the optical data generator takes the relative position and outputs the relevant vision view. Next, this project introduces a newly designed monocular vision-based DRL system to provide guidance and control, facilitating a soft landing at the targeted position and velocity. Following this, FGSM attacks are employed on the optical input data to produce an adversarial image, which serves to assess the impact on the DRL system. SHAP values are utilised to create XAI signatures for both the adversarial and normal input images. Lastly, we propose and train an LSTM-based and a Transformer-based adversarial attacks detector that learns to discern normal and adversarial SHAP values, effectively detecting adversarial attacks on the vision-based DRL system.

4. Key Findings

In this project, proposed and developed are tested on synthetic data and the performance of each scenario is analysed. The impacts of adversarial attacks to the space GNC systems have been analysed. To further evaluate the performance of the rendezvous scenario, we tested them with real-world images obtained from the Autonomous Systems and Machine Intelligence Laboratory (ASMI Lab) at City, University of London. These data include sensor noise, camera calibration noise, ground truth measurement noise, and

different lighting conditions that are not present in the training synthetic images. Due to limitation of simulating the Mars candidate landing site, the Mars landing scenario has not been tested with real data.

Space Rendezvous Scenario
The CNN-based relative pose estimator is trained on image data generated from the simulator we built, resulting in a dataset with 32,500 images for training and testing in total. After training for 50 epochs, the proposed spacecraft relative pose estimator achieves an accuracy of around 0.49 metres in position error and 0.68 degree in attitude error on the test dataset. Compared recent published works, the proposed spacecraft deep relative pose estimator can achieve relatively good performance on the synthetic data and can be applied as a baseline model to implement the adversarial attack algorithm on and test the adversarial attack detector.

To investigate the impact of FGSM adversarial attacks on CNN-based space relative pose estimation, different ϵ values are selected to generate adversarial onboard camera image input to the proposed deep relative pose estimator. The experimental results demonstrate that as the ϵ value increases, the CNN model's prediction error becomes larger. The attitude error is quite stable on ϵ = 0.1, 0.05 and 0.01 but has a dramatic increase if the ϵ > 0.3. Typically, when the distance between the camera and the target is smaller than 30m. In most cases, continuously attacking the deep model for more than 15 frames after the camera approaches less than 30m to the target, the camera (chaser) will fail to reach the target position. In a real space rendezvous mission where a chaser relies on a CNN-based relative pose estimation system, an adversarial attack has the potential to cause the chaser to fail in approaching the target position, resulting in mission failure.

The LSTM-based adversarial attacks detector has been trained on a generated dataset with 24,000 SHAP values for training and 6,000 for testing. After training for 1,000 epochs, it achieved a training accuracy of 99.98% and a test accuracy of 99.90% on the test dataset. Then, the LSTM-based adversarial attacks detector, CNN-based pose estimator and SHAP value generator have been integrated into one system to test with three complete trajectories. From the test results, the proposed adversarial attack detector successfully detects all incoming FGSM attacks when the ϵ = 0.5. As the ϵ value goes small, i.e. fewer perturbations are made to input images, the detection accuracy has slightly dropped. For these three test trajectories, the proposed adversarial attack detector achieves a detection accuracy of 99.21% on average.

Furthermore, all frameworks developed in this scenario has been tested on the real-world data in the ASMI Lab. As the results, the CNN-based pose estimator achieved a relative position error about 1.43m and attitude error about 0.0551m. Compared with the prediction accuracy on the Synthetic-Lab Dataset, the position error of the ASMI Dataset is slightly higher. This could be attributed to variations in the illumination conditions compared to the Synthetic-Lab Dataset, as well as factors such as ground truth measurement noise and camera calibration noise. On the real-world data, the LSTM-based adversarial attack detector achieves an average correct detection rate of 96.29% with digital FGSM attacks.

Final experiment conducts physical implementation of FGSM attacks in real-world. In this case, a projector has been employed to project the calculated perturbation patch to the

target. In this experiment, LSTM-based adversarial attack detector achieves an average correct detection rate of 80% on real physical FGSM attacks, which is lower than the accuracy for digital attacks. This discrepancy could be due to several factors, i.e. illumination condition, lightning source, light direction, the projector's resolution, errors in image resizing, and normalisation by its firmware of camera.

Mars Landing Scenario

For the Mars landing scenario, three DRLs operate in a sequential manner. The first DRL agent (*Agent 1*) is active when the altitude exceeds 400m. The second DRL agent (*Agent 2*) takes over at altitudes ranging from 400m down to 30m. Finally, the third DRL agent (*Agent 3*) assumes control at altitudes below 30m. The DRL models are tested with 300 random episodes and achieves of 100% in successful soft-landing conditions.

To train the adversarial attack detectors, we utilise 30,000 SHAP value sets for normal instances and another 30,000 for adversarial instances. Upon completion of training, the LSTM-based adversarial attack detector achieved a training accuracy of 96.89% and an accuracy of 97.16% on the test set. And the Transformer-based adversarial attack detector achieved a training accuracy of 96.26% and an accuracy of 96.58% on the test set.

Then, two experimental tests for the performance of both LSTM-based and Transformer-based adversarial attack detection during the operation of the DRLs, as well as to evaluate the impact of adversarial attacks on the vision-based DRL landing scheme. For the first experiment (Task 1), the FGSM attacks are initiated at a random time step during the episode and continue to perturb the image for the subsequent time steps until the lander contacts the ground. For the second task (Task 2), the FGSM attacks will be randomly applied between time 0 to time 130 to continuously attack for 30 time steps. The second task aims to test the accuracy of the adversarial detectors after FSGM attacks bring the lander to unknow states. In the first Task 1, the LSTM-based detector achieves an average accuracy of 96.06% and the Transformer-based detector achieves an average accuracy of 97.89%. In the first Task 2, the LSTM-based detector achieves a detection accuracy of 90.16% and the Transformer-based detector achieves a detection accuracy of 92.92%.

For the Taske 1, as the ϵ value decreases, indicating fewer perturbations to the input images, there is a decline in detection accuracy. However, this corresponds with an increase in the successful landing rate. The detection accuracy for adversarial attacks experiences a more pronounced decrease when ϵ =1/255, attributable to the minimal perturbation applied to the input image. Despite this, the lander achieves the successful landing criteria in all test episodes under these conditions. This outcome implies that the feature extractor within the proposed DRL can produce highly accurate features, even with minor adversarial perturbations. Consequently, the lander is still able to arrive at the target location with the desired velocity. For higher perturbations, i.e. $\epsilon \geq 5/255$, two adversarial attack detector demonstrates a high level of confidence in identifying incoming FGSM attacks. However, strong perturbations to the input images can lead to poor performance in the current vision-based DRL guidance scheme.

Task 2 has more complexity than the Task 1, the detectors are required to identify between adversarial input and anomalies (in such states which never seen in training phase). In this case, the Transformer-based detector works better than the LSTM-based

detector with around 2% higher is detection accuracy. From the results in both tasks, the Transformer-based detector outperforms the LSTM-based detector.

5. Conclusion

In this project, we proposed the first comprehensive study on adversarial attacks for AI-based space GNC systems and their detection mechanisms. Two scenarios of space AI-based GNC are studied.

The first study examines the impact of adversarial attacks on CNN-based spacecraft relative pose estimation in space rendezvous scenarios. To achieve this, we first developed a 3D simulator to render the camera view in a space rendezvous mission. Then, a CNN-based relative pose estimation algorithm was proposed. FGSM adversarial attacks were implemented, significantly impacting the model's predictions. Subsequently, an LSTM-based adversarial attack detector was proposed to identify adversarial attacks on input images. XAI techniques were adopted to analyse the model's predictions and generate SHAP values-based explanations for the model's predictions. Multiple experiments were carried out to evaluate the performance of the CNN-based spacecraft relative pose estimator, the impact of adversarial attacks, and the performance of the proposed adversarial attack detector on both synthetic datasets and real-world data of digital and physical adversarial attacks.

The second study initiates an investigation into the effects of adversarial attacks on a vision-based DRL framework for guidance and control in a Mars landing scenario. A planetary landing simulator was developed to generate optical data along with corresponding aerodynamic parameters for the target landing scenario. The project introduces a DRL scheme, relying solely on visual data for observation. Following this, an adversarial attack detector is introduced, utilising SHAP value-based explanations to pinpoint adversarial manipulations in input images. A series of experiments were conducted to assess the efficacy of the vision-based DRL in landing guidance and control, the influence of adversarial attacks on DRL performance during the landing phase, and the effectiveness of the newly proposed adversarial attack detector.

6. Future Work

The solution we are aiming to develop in this project is going to be based on the orbital relative navigation scenario and Mars landing guidance and control. It will be specific to the AI systems adopted as deep GNC scheme. Generalising the solution to other space guidance scenarios and other navigation and control space scenarios can be done for further requested development as the same principle of the deep adversarial detection scheme we develop here could be extended to those other scenarios including other deep learning GNC schemes.

Also, while the proposed adversarial attack detectors demonstrate high accuracy in detection, the current work has not thoroughly examined the actual recourse following the detection of these attacks. The integration of the AI-based space GNC systems with the adversarial attack detector to develop adversarial defence mechanisms presents a promising avenue for future work.