Institute of Theoretical and Applied Informatics
Polish Academy of Sciences

Abstract of doctoral dissertation

Explainability and security of intelligent systems

mgr inż. Katarzyna Filus

Supervisor:
dr hab. inż. Joanna Domańska, prof. IITiS PAN

Gliwice, 2023

Abstract

Intelligent systems are used in many areas of human life. Although they offer high accuracy and effective problem-solving, their practical application is still limited due to their security issues and low explainability. These current problems result in a low level of trust that society places in artificial intelligence, which limits the use of such systems in safety-critical domains. Therefore, it is necessary to propose methods that would improve the security and explainability of existing systems, as well as to consider these aspects in the process of designing new solutions.

The aim of this doctoral thesis is to improve the security and explainability of intelligent systems. To achieve this goal, different issues related to security and explainability of such systems have been thoroughly investigated. A series of methods has been proposed to improve these two aspects. The results confirming their effectiveness have been presented and compared with alternative existing approaches. Due to the interconnection between security and explainability, some of the proposed methods have a positive impact on both of these aspects.

In the case of security, different aspects of intelligent systems security have been comprehensively examined: both the traditional information technology threats (cyber attacks, software vulnerabilities) and the threats directly related to artificial intelligence algorithms.

In the domain of cyber attacks, an attack detection system based on novel methods for initialization and training of Random Neural Networks has been proposed. The proposed initialization method aims to ensure the neutrality of the network prior to the beginning of its training process and better interpretability of the process, resulting in improved explainability. The training method limits the number of expensive operations performed during the training procedure. The proposed methods allow to achieve better accuracy results in the context of detecting cyber attacks compared to the baseline solution.

In the area of software vulnerabilities, an extensive analysis of known vulnerabilities within the leading deep learning library, TensorFlow, has been conducted, and the adequacy of available static code analyzers in detecting these vulnerabilities has been tested. Since the available tools show low effectiveness in detecting vulnerabilities in this type of software, vulnerability detection methods based on traditional machine learning algorithms and feature selection have been proposed, as well as a hybrid system using an original modification of Random Neural Networks and mixed features characterizing the program code. It has been demonstrated that the proposed solutions improve accuracy compared to the baseline solutions.

The thesis also presents methods from the field of deep neural network testing. A method has been proposed that allows to create datasets for comprehensive real-world testing of deep vision networks. Also, a method has been created that allows to test networks' resiliency to dedicated threats - adversarial attacks. The first method makes it possible to automatically create labeled datasets and test deep neural networks directly on automated guided vehicles. The proposed

adversarial attack allows testing of vision networks and is independent of the network's classifier. The results showed that samples generated through the proposed attack result in higher harmfulness than samples generated through commonly used attacks. A metric that is straightforward in interpretation has also been proposed to enable practical evaluation of the degree of harmfulness of adversarial attacks. The thesis describes the advantages of the proposed metric over available metrics. The simplicity of interpretation and use of the metric has a positive impact on the explainability aspect.

In the area of explainability, a method for interpreting the operation of deep Convolutional Neural Networks has been proposed. The method allows to visualize the activation of the network and its overall concentration on a given image. The method is independent of the network's classifier and does not require the calculation of the values of its gradients. It has been presented that different variants of the method can be used for visual examination of pattern extraction and bias. It has also been presented that the method can be used to study the impact of adversarial attacks on network operation, which has a positive impact on the security aspect. The proposed method is simpler than the available approaches, and at the same time informative, which results in improved explainability of network operation. Also, an intelligent system for locating users with smartphones has been proposed. The proposed solution uses Bluetooth signal strength and original metrics that can be used to filter out unreliable readings of users' positions. The proposed metrics are formulated in such a way as to be straighforward in interpretation even for non-technical users of the intelligent system. This has a positive impact on the practical explainability of the system.

List of publications

1. Filus, K., and J. Domańska, ,,NetSat: Network Saturation Adversarial Attack", IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, In Press.

2. Filus, K., and J. Domańska, ,,*Recycling of Generic ImageNet-trained Models for Smart-city Applications*", The 10th IEEE International Conference on Data Science and Advanced Analytics (DSAA), Thessaloniki, Greece, 2023.

3. Filus, K., and J. Domańska, ,,*Global Entropy Pooling Layer for Convolutional Neural Networks*", Neurocomputing, vol. 555, 2023.

4. Filus, K., and J. Domańska, ,,*Software Vulnerabilities in TensorFlow-Based Deep Learning Applications*", Computers & Security, vol. 124, 10/2022, 2023.

5. Filus, K., Ł. Sobczak, J. Domańska, A. Domański, and R. Cupek, ,,*Real-time testing of vision-based systems for AGVs with ArUco markers*", IEEE International Conference on Big Data, Osaka, Japan, 2022.

6. Filus, K., and J. Domańska, ,,*NAM: What Does a Neural Network See?*", International Joint Conference on Neural Networks (IJCNN 2022), IEEE WCCI 2022, Padova, Italy, 2022.

7. Filus, K., S. Nowak, J. Domańska, and J. Duda, ,,*Cost-Effective Filtering of Unreliable Proximity Detection Results Based on BLE RSSI and IMU Readings Using Smartphones*", Scientific Reports, vol. 12, issue 1, 2022.

8. Filus, K., P. Boryszko, J. Domańska, M. Siavvas, and E. Gelenbe, ,,*Efficient Feature Selection for Static Analysis Vulnerability Prediction*", Sensors, vol. 21 (4), issue Special Issue: Security and Privacy in Software Based Critical Contexts, 2021.

9. Filus, K., J. Domańska, and E. Gelenbe, ,,*Random Neural Network for Lightweight Attack Detection in the IoT*", MASCOTS 2020: Modelling, Analysis, and Simulation of Computer and Telecommunication Systems, vol. 12527: Springer International Publishing, pp. 79-91, 2021.

10. Filus, K., M. Siavvas, J. Domańska, and E. Gelenbe, ,,*The Random Neural Network as a Bonding Model for Software Vulnerability Prediction*", Modelling, Analysis, and Simulation of Computer and Telecommunication Systems, vol. 12527: Springer International Publishing, pp. 102-116, 2021.

11. Filus, K., A. Domański, J. Domańska, D. Marek, and J. Szyguła, ,,*Long-Range Dependent Traffic Classification with Convolutional Neural Networks Based on Hurst Exponent Analysis*", Entropy, vol. 22, issue 10, 2020.

12. Sobczak, Ł., K. Filus, M. Halama, and J. Domańska, ,,*Visual examination of relations between known classes for deep neural network classifiers*", IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, In Press.

13. Halama, M., K. Filus, and J. Domańska, ,,*Robust category recognition based on deep templates for educational mobile applications*", IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, In Press.

14. Kelesoglu, N., K. Filus, and J. Domańska, ,,*HierAct: a Hierarchical Model for Human Activity Recognition in Game-Like Educational Applications*", 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, In Press.

15. Sobczak, Ł., K. Filus, J. Domańska, and A. Domański, ,,*Building a real-time testing platform for unmanned ground vehicles with UDP Bridge*", Sensors, vol. 22, issue 21, 2022.

16. Sobczak, Ł., K. Filus, J. Domańska, and A. Domański, ,,*Finding the best hardware configuration for 2D SLAM in indoor environments via simulation based on Google Cartographer*", Scientific Reports, vol. 12, 2022.

17. Sobczak, Ł., K. Filus, A. Domański, and J. Domańska, ,,*LIDAR Point Cloud generation for SLAM algorithm evaluation*", Sensors, vol. 21 (10), issue Special Issue: Advance in Sensors and Sensing Systems for Driving and Transportation: Part B, 2021.

18. Domański, A., J. Domańska, K. Filus, J. Szyguła, and T. Czachórski, ,,*The self-similar markovian sources*", Applied Sciences, vol. 10, issue 11, 2020.

19. Marek, D., J. Szyguła, A. Domański, J. Domańska, K. Filus, and M. Szczygieł, ,,*Adaptive Hurst-Sensitive Active Queue Management*", Entropy, vol. 24, issue 3, 2022.

20. Szyguła, J., A. Domański, J. Domańska, D. Marek, K. Filus, and S. Mendla, ,,*Supervised learning of Neural Networks for Active Queue Management in the Internet*", Sensors, vol. 21(15), issue Special Issue "Mathematical Modelling and Analysis in Sensors Networks", 2021.

21. Marek, D., A. Domański, J. Domańska, J. Szyguła, T. Czachórski, J. Klamka, and K. Filus, ,,*Approximation Models for the Evaluation of TCP/AQM Networks*", Bulletin of the Polish Academy of Sciences, Technical Sciences (BPASTS), vol. 70, issue 4, 2022.

22. Marantos, C., M. Siavvas, D. Tsoukalas, C. P. Lamprakos, L. Papadopoulos, P. Boryszko, K. Filus, J. Domańska, A. Ampatzoglou, A. Chatzigeorgiou, et al., ,,*SDK4ED: One-click platform for Energy-aware, Maintainable and Dependable Applications*", 25th Design, Automation and Test in Europe Conference, Belgium, 03/2022.