Instytut Informatyki Teoretycznej i Stosowanej
Polskiej Akademii Nauk

Streszczenie rozprawy doktorskiej

Wyjaśnialność i bezpieczeństwo systemów inteligentnych

mgr inż. Katarzyna Filus

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

Gliwice, 2023

Streszczenie

Systemy inteligentne są stosowane w wielu obszarach życia człowieka. Chociaż oferują wysoką dokładność oraz efektywne rozwiązywanie problemów, ich praktyczne stosowanie jest ograniczone ze względu na liczne zagrożenia związane z nimi oraz ich niską wyjaśnialność. Te aktualne problemy skutkują niskim zaufaniem, którym społeczeństwo darzy sztuczną inteligencję oraz ograniczonym stosowaniem tego typu systemów w dziedzinach, w których bezpieczeństwo jest krytyczne. Z tego powodu konieczne jest zaproponowanie metod umożliwiających poprawę bezpieczeństwa i wyjaśnialności obecnych systemów, a także uwzględnieniem tych aspektów w procesie projektowania nowych rozwiązań.

Celem niniejszej pracy doktorskiej jest poprawa bezpieczeństwa i wyjaśnialności systemów inteligentnych. Aby zrealizować cel pracy, w sposób kompleksowy zbadano zagadnienia dotyczące bezpieczeństwa i wyjaśnialności. Zaproponowano szereg metod mających na celu poprawę obu tych aspektów oraz zaprezentowano wyniki potwierdzające ich skuteczność i porównano je ze stosownymi metodami. Ze względu na sieć wzajemnych połączeń między bezpieczeństwem oraz wyjaśnialnością, część zaproponowanych metod ma pozytywny wpływ na oba rozpatrywane zagadnienia.

W przypadku bezpieczeństwa w sposób kompleksowy zbadano różne aspekty bezpieczeństwa systemów inteligentnych: tradycyjne zagrożenia informatyczne (cyberataki, podatności oprogramowania) oraz zagrożenia bezpośrednio związane z algorytmami sztucznej inteligencji.

W domenie cyberataków zaproponowano metodę wykrywania ataków opartą na autorskich metodach inicjalizacji oraz trenowania Losowych Sieci Neuronowych. Proponowana metoda inicjalizacji ma na celu zapewnienie neutralności sieci przed rozpoczęciem procesu jej trenowania oraz lepszą interpretowalność tego procesu, skutkując poprawą wyjaśnialności. Metoda trenowania ogranicza liczbę kosztownych operacji wykonywanych w procesie trenowania. Proponowane metody pozwalają uzyskać lepsze wyniki dokładności w kontekście wykrywania cyberataków w stosunku do bazowego rozwiązania.

W obszarze podatności oprogramowania przeprowadzono obszerną analizę znanych podatności wiodącej biblioteki uczenia głębokiego - TensorFlow - oraz przetestowano adekwatność dostępnych statycznych analizatorów kodu w wykrywaniu tych podatności. Jako że dostępne narzędzia wykazują niską skuteczność w wykrywaniu podatności tego typu oprogramowania, zaproponowano metody wykrywania podatności oparte na tradycyjnych algorytmach uczenia maszynowego oraz selekcji cech, a także system hybrydowy wykorzystujący autorską modyfikację Losowych Sieci neuronowych i cechy mieszane opisujące kod programu. Wykazano, że zaproponowane rozwiązania poprawiają dokładność w stosunku do bazowych rozwiązań.

W pracy zaproponowano również metody z zakresu testowania głębokich sieci neuronowych. Zaproponowano metodę umożliwiającą tworzenie zbiorów danych na potrzeby kompleksowego testowania w rzeczywistych warunkach, a także metody umożliwiające testowanie sieci pod kątem zagrożeń dedykowanych - ataków adwersarza. Pierwsza z metod umożliwia automatyczne tworzenie

oznaczonych zbiorów danych i testowanie głębokich sieci neuronowych bezpośrednio na pojazdach sterowanych automatycznie (ang. Automated Guided Vehicles, AGV). Zaproponowany atak adwersarza umożliwia testowanie sieci wizyjnych i jest niezależny od klasyfikatora sieci. Uzyskane wyniki pokazały, że próbki wygenerowane za pośrednictwem proponowanego ataku skutkują wyższą szkodliwością niż próbki wygenerowane za pośrednictwem powszechnie stosowanych ataków. Zaproponowano również prostą w interpretacji metrykę umożliwiającą praktyczną ocenę stopnia szkodliwości ataków adwersarza. W pracy opisano zalety proponowanej metryki w stosunku do dostępnych metryk. Prostota interpretacji i wykorzystania metryki wywiera pozytywny wpływ na aspekt wyjaśnialności.

W zakresie wyjaśnialności zaproponowano metodę interpretacji działania głębokich sieci konwolucyjnych. Metoda obejmuje wizualizację aktywacji sieci oraz jej ogólnego skupienia na danym obrazie. Metoda jest niezależna od klasyfikatora sieci oraz nie wymaga wyznaczania wartości jej gradientów. Zaprezentowano, że różne warianty metody mogą być stosowane do wizualnego nadzoru ekstrakcji wzorców oraz stronniczości. Zaprezentowano również, że metoda może być stosowana do badania wpływu ataków adwersarza na działanie sieci, co ma pozytywny wpływ na aspekt bezpieczeństwa. Proponowana metoda jest prostsza od dostępnych metod, a jednocześnie informatywna, co skutkuje polepszeniem wyjaśnialności działania sieci. Zaproponowano również system inteligentny do lokalizacji użytkowników z telefonami w przestrzeni. Rozwiązanie wykorzystuje siłę sygnału Bluetooth oraz autorskie metryki pozwalające odfiltrować niewiarygodne odczyty pozycji użytkowników. Zaproponowane metryki są sformułowane w taki sposób, aby były proste w interpretacji nawet dla nietechnicznych użytkowników systemu inteligentnego. Wywiera to pozytywny wpływ na aspekt praktycznej wyjaśnialności.

Wykaz publikacji

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.