Cyber Résilience opérationnelle augmentée par Intelligence Artificielle - Etude sur le test des systèmes de contrôle et de supervision de transport

Fanny Serre,
Kereval
Email: fanny.serre@kereval.com

Yacine Tamoudi
Kereval
Email: yacine.tamoudi@kereval.com

***Abstract*—Le projet CRIA développe des solutions innovantes pour le test des systèmes critiques de contrôle et de supervision. Face à des infrastructures complexes et à des risques de cybersécurité élevés, le projet s'intéresse à l'automatisation des tests fonctionnels et de cybersécurité grâce à l'intelligence artificielle (IA). Le projet utilise des techniques comme les réseaux antagonistes génératifs (GAN) pour générer des attaques malveillantes indétectées et le Deep Reinforcement Learning (DRL) pour l'automatisation du fuzz testing. Ces travaux permettent de tester les vulnérabilités des systèmes critiques et d'améliorer leur résilience face aux menaces émergentes. Le projet CRIA ouvre ainsi de nouvelles perspectives pour tester la robustesse des systèmes de supervision critiques dans des environnements réels, contribuant à une meilleure sécurité dans des domaines tels que le transport aérien.**

***Index Terms*—Cybersécurité, IA, GAN, DRL.**

I. CONTEXTE ET PROBLEMATIQUE

Les systèmes de contrôle et de supervision dans le transport, tels que le contrôle aérien, constituent des infrastructures critiques de plus en plus complexes et soumises à des risques majeurs en termes de fiabilité et de cybersécurité. Le test de ces systèmes constitue un défi croissant pour les industriels qui les développent. En effet, la diversité croissante des données rend la vérification de leur cohérence complexe, tandis que l'accélération des cycles de développement exige des tests continus, tant fonctionnels que de cybersécurité. Les approches de test traditionnelles, basées sur des stratégies définies lors de la spécification du système, deviennent obsolètes. Elles ne permettent ni de couvrir la diversité des données ni de garantir le bon fonctionnement de systèmes évolutifs. La recherche de vulnérabilités doit désormais s'appuyer sur des mécanismes capables de générer automatiquement des tests innovants, afin d'explorer des scénarios imprévus et de renforcer la résilience des systèmes critiques.

Le projet CRIA propose des solutions innovantes pour le test des systèmes de supervision critiques, incluant ceux intégrant de l'IA. Basé sur une architecture intégrant l'IA, il exploite l'apprentissage à partir des données et des résultats de test pour concevoir des tests fonctionnels, de sûreté de fonctionnement et de cybersécurité.

II. ETAT DE L'ART

L'approche de Deep Reinforcement Testing exploite le Deep Reinforcement Learning (DRL) [1] pour tester les systèmes de contrôle et de supervision. Les techniques comme le Deep Q-Learning Network (DQN) ont démontré leur efficacité pour résoudre des tâches complexes grâce à des avancées telles que le Double DQN [2], A3C[3], A2C [4] et IMPALA[5], TRPO [6]et PPO [7]. En particulier, les algorithmes A3C, TRPO et PPO sont capables de gérer des espaces d'actions continus, répondant aux besoins des systèmes complexes. Pour surmonter la rareté des récompenses dans un environnement réel, sans ajuster spécifiquement le système de récompense, des approches complémentaires ont été expérimentées comme : *unsupervised auxillary tasks* [8], *Curiosity Driven Exploration* [9], *hindsight experience replay* [10] et *hierarchical reward systems* [11]. Dans la cybersécurité, le DRL a prouvé sa capacité à automatiser des tests via des algorithmes de fuzzing, comme ceux appliqués à l'analyse de fichiers PDF [12]. Ces algorithmes, bien qu'encore inexploités dans la surveillance d'objets mobiles, ouvrent la voie à des innovations pour renforcer la couverture fonctionnelle et détecter les vulnérabilités des systèmes critiques.

Les GANs, initialement introduits en 2014 [13], ont prouvé leur efficacité dans divers domaines, allant de la génération d'images, de sons et de textes [14], [15] à la gestion de séquences de données discrètes [16] et continues [17]. Des améliorations, telles que l'intégration d'auto-encodeurs [18], [19], ont permis de stabiliser leur apprentissage. En cybersécurité, les GANs sont utilisés pour optimiser les stratégies de défense [20] ou contourner les systèmes de détection [21]-[23]. Des études, comme celles sur IDSGAN [24], montrent que les GANs peuvent générer du trafic malveillant indétectable, en modifiant le comportement des malwares pour les rendre similaires à des applications légitimes, contournant ainsi les systèmes de défense avec un taux de succès de 60 %. Bien que prometteuse, l'utilisation des GANs pour contourner les détections d'anomalies basées sur l'analyse de séries temporelles reste sous-explorée, offrant une opportunité innovante pour le projet.

Ces deux sous-ensembles technologiques représentent des approches innovantes pour repousser les limites des tests fonctionnels et de cybersécurité, tout en s'attaquant aux défis liés aux systèmes critiques.

III. TEST PAR REINFORCEMENT LEARNING

A. Développement d'un module de fuzzing renforcé par IA

La première brique technologique du projet s'est concentrée sur l'application du RL au test de systèmes critiques. Les recherches ont concerné l'apprentissage parallélisé, l'évaluation d'algorithmes à haute efficacité d'échantillonnage et l'intégration du RL avec des algorithmes génétiques pour optimiser l'exploration. Des méthodes adaptées aux problématiques multimodales et des techniques d'échantillonnage pour détecter des événements rares ont été développées. L'approche

Novelty Search a permis d'explorer des scénarios innovants, tandis que la discrétisation des actions a permis de gérer la complexité des environnements. Des techniques basées sur les GANs et VAE ont généré des trajectoires pour de nouveaux cas de test. Ces travaux ont abouti au développement de Chinkara, un outil de fuzzing renforcé par IA, intégrant plusieurs moteurs d'altération comme SAC, TD3, DDPG et CEM-RL, combinant RL et algorithmes génétiques. Chinkara utilise des moteurs d'altération aléatoire et exploite les données générées pour améliorer l'apprentissage. Il offre des fonctions personnalisables de récompenses, un mode multimodal, ainsi qu'une interface graphique intuitive pour analyser les résultats, créer des métriques spécifiques et générer des rapports de test.

B. Evaluation et résultats

L'évaluation du module Chinkara a fait appel à des systèmes sous tests simplifiés, comme des systèmes de multilatération ou de gestion de trajectoire. Chaque environnement teste une caractéristique spécifique, telle que la robustesse aux entrées bruitées ou la détection de bugs. Les algorithmes sont évalués selon leur capacité à explorer des zones pertinentes et à converger vers des solutions optimales. Les résultats montrent que l'algorithme SAC performe dans les environnements simples, surpassant le fuzzing aléatoire en termes de rapidité et de précision. En revanche, pour des environnements plus complexes avec des récompenses non linéaires, les algorithmes combinant RL et méthodes génétiques, comme CEM-RL, sont plus robustes. SACseq, une variante de SAC, s'est révélé efficace pour explorer plusieurs solutions distinctes.

IV. EVASION DE LA DETECTION

A. Développement d'un module de génération de patchs adversariaux

La seconde composante technologique du projet a porté sur les mécanismes d'évasion des systèmes de détection. Des expérimentations ont été réalisées dans un contexte de surveillance en bord de route, avec pour objectif de perturber la détection d'objets en générant des exemples adversariaux physiques sous forme de patchs. Ces patchs ont été réalisés à l'aide de techniques d'optimisation basées sur la descente de gradient, enrichies par des transformations aléatoires (position, échelle, rotation, luminosité) pour renforcer leur robustesse. Conçus pour rester efficaces sur divers formats (float, uint, jpg, png), ils simulent des conditions d'utilisation réalistes.

Ces travaux ont conduit à la création d'une librairie complète pour générer et évaluer ces patchs adversariaux, intégrant une interface graphique et des outils d'analyse automatisée. Une documentation détaillée a également été élaborée, facilitant l'adoption et l'utilisation des outils dans différents environnements.

B. Evaluation et résultats

L'évaluation expérimentale a porté sur deux modèles de détection, YOLO et Faster RCNN, et sur des bases de données comme PatNet et BDD100k. L'objectif était de mesurer la capacité des patchs à perturber ces systèmes, en étudiant notamment leur transférabilité entre différents contextes et classes d'objets. Les expérimentations ont montré que les patchs adversariaux sont capables de perturber efficacement la détection de certaines classes d'objets, comme les piétons. Cependant, leur performance varie selon les contextes, notamment pour des classes plus complexes comme les véhicules. Les patchs se sont révélés sensibles aux transformations physiques, bien que certaines techniques d'optimisation aient permis d'en renforcer la robustesse. La transférabilité des patchs, c'est-à-dire leur capacité à être efficaces sur des modèles ou bases de données non utilisées pour leur optimisation, s'est avérée dépendante de la similarité entre ces environnements.

V. PERSPECTIVES

Dans le cadre de nos travaux en cours, nous poursuivons deux axes principaux de recherche. Le premier concerne l'élaboration de nouvelles méthodologies outillées pour tester la confiance des systèmes d'IA, avec un focus particulier sur le test de cybersécurité des systèmes intégrant des grands modèles de langage (LLMs). Alors que les LLMs sont de plus en plus intégrés dans des applications critiques, il devient en effet essentiel de les tester pour détecter d'éventuelles vulnérabilités pouvant affecter leur utilisation dans des environnements sensibles. Nos recherches se concentrent sur le développement de nouvelles techniques pour générer des scénarios adversariaux spécifiques aux LLMs, notamment des attaques par injection de prompts ou des exemples malveillants. Cela permettrait d'identifier les risques associés à l'utilisation de ces modèles dans des contextes où la sécurité et la fiabilité sont primordiales. Cette approche s'inspire de notre travail précédent sur la génération de patchs adversariaux, qui a démontré l'importance de comprendre et d'anticiper les vulnérabilités dans des systèmes complexes.

Le deuxième axe se concentre sur l'application de l'IA pour améliorer les processus de test, notamment en y intégrant des LLMs. Cette approche vise à automatiser et à enrichir les étapes du processus de test, telles que la reformulation des exigences, la génération de cas de test et l'exécution automatique de tests. En utilisant des LLMs pour traiter et analyser les exigences, il devient possible de mieux répondre à la diversité des données testées et d'explorer plus efficacement des scénarios imprévus, ce qui est crucial dans des environnements critiques. Par exemple, un LLM pourrait être utilisé pour générer de nouveaux cas de test basés sur des scénarios réalistes ou pour détecter des incohérences dans les exigences fonctionnelles. Ce type d'automatisation permettrait non seulement d'accélérer le processus de test, mais aussi d'améliorer la couverture des tests en explorant des angles inédits ou sous-explorés. Ces avancées devraient contribuer à renforcer la résilience des systèmes critiques face aux menaces émergentes.

CONSORTIUM ET FINANCEMENT

Ce projet collaboratif (2019-2022) porté par Kereval et financé par dispositif RAPID (Régime d'APpui à l'Innovation Duale), a été réalisé en partenariat avec les sociétés Thales et SmartTesting.

REFERENCES

[1] V. Mnih et al., « Human-level control through deep reinforcement learning », Nature, vol. 518, no 7540, p. 529-533, févr. 2015.
[2] V. Mnih et al., « Playing Atari with Deep Reinforcement Learning », p. 9.
[3] H. van Hasselt, A. Guez, et D. Silver, « Deep Reinforcement Learning with Double Q-learning », arXiv:1509.06461 [cs], sept. 2015.
[4] V. Mnih et al., « Asynchronous Methods for Deep Reinforcement Learning », arXiv:1602.01783 [cs], févr. 2016.

[5] OpenAi: A2C implementation. OpenAI, 2018.
[6] L. Espeholt et al., « IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures », arXiv:1802.01561 [cs], févr. 2018.
[7] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, et P. Abbeel, « Trust Region Policy Optimization », arXiv:1502.05477 [cs], févr. 2015.
[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, et O. Klimov, « Proximal Policy Optimization Algorithms », arXiv:1707.06347 [cs], juill. 2017.
[9] M. Jaderberg et al., « Reinforcement Learning with Unsupervised Auxiliary Tasks », arXiv:1611.05397 [cs], nov. 2016.
[10] M. Andrychowicz et al., « Hindsight Experience Replay », arXiv:1707.01495 [cs], juill. 2017.
[11] T. D. Kulkarni, K. Narasimhan, A. Saeedi, et J. Tenenbaum, « Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation », p. 9.
[12] K. Böttinger, P. Godefroid, et R. Singh, « Deep Reinforcement Fuzzing », janv. 2018.
[13] I. J. Goodfellow et al., « Generative Adversarial Networks », arXiv:1406.2661 [cs, stat], juin 2014.
[14] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, et Y.-H. Yang, « MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment », arXiv:1709.06298 [cs, eess], sept. 2017.
[15] W. Fedus, I. Goodfellow, et A. M. Dai, « MaskGAN: Better Text Generation via Filling in the_____ », arXiv:1801.07736 [cs, stat], janv. 2018.
[16] L. Yu, W. Zhang, J. Wang, et Y. Yu, « SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient », arXiv:1609.05473 [cs], sept. 2016.
[17] O. Mogren, « C-RNN-GAN: Continuous recurrent neural networks with adversarial training », arXiv:1611.09904 [cs], nov. 2016.
[18] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, et O. Winther, « Autoencoding beyond pixels using a learned similarity metric », arXiv:1512.09300 [cs, stat], déc. 2015.
[19] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, et S. Mohamed, « Variational Approaches for Auto-Encoding Generative Adversarial Networks », arXiv:1706.04987 [cs, stat], juin 2017.
[20] D. Li, D. Chen, J. Goh, et S. Ng, « Anomaly Detection with Generative Adversarial Networks for Multivariate Time Series », arXiv:1809.04758 [cs, stat], sept. 2018.
[21] W. Hu et Y. Tan, « Black-Box Attacks against RNN based Malware Detection Algorithms », arXiv:1705.08131 [cs], mai 2017.
[22] M. Rigaki et S. Garcia, « Bringing a GAN to a Knife-Fight: Adapting Malware Communication to Avoid Detection », in 2018 IEEE Security and Privacy Workshops (SPW), 2018, p. 70-75.
[23] W. Hu et Y. Tan, « Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN », arXiv:1702.05983 [cs], févr. 2017.
[24] Z. Lin, Y. Shi, et Z. Xue, « IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection », arXiv:1809.02077 [cs], sept. 2018.