Supervised Intrusion Detection with Out-of-Distribution Detection for Microservices

Yong-Syuan Chen, Hsiang-Yin Lien, Jo-Yu Li, and Chia-Yu Lin
Department of Computer Science and Information Engineering
National Central University, Taoyuan, Taiwan
Corresponding Author: Chia-Yu Lin (sallylin0121@ncu.edu.tw)

***Abstract*—Microservice architecture enhances system flexibility and reliability but raises security concerns due to potential malicious attacks. We propose a supervised Out-of-Distribution (OOD) detector leveraging AI and ML to analyze container command sequences. Our technique identifies known and unknown attack patterns, employing out-of-distribution detection. Using a deep neural network, we learn features and minimize classification errors. Comparative evaluations demonstrate its efficacy, aiming to enhance container security and deepen insights into microservice attack behaviors.**

***Index Terms*—Intrusion detection, out-of-distribution detection, microservice security**

I. Introduction

Microservices architecture boosts system flexibility and reliability. However, the broad adoption of containers brings security risks due to exploitable interactions. As attacks evolve and the number of unknown threats increases, detecting unknown malicious attacks becomes critical.

Rahali et al. developed "MalBERTv2 [1]", a BERT-based model for proactive malware detection, but it struggles with new attacks and requires substantial computational resources. Seneviratne et al. proposed "SHERLOCK [2]", a self-supervised model that converted malware binaries into images, preserving the structure and semantic information of the malware. McLaughlin et al. introduced a novel data augmentation method for opcode sequence-based malware detection [3]. This method dynamically generated more realistic augmented samples during the training process, adapting to the network's learning progress. Despite these advancements, current supervised methods still struggle to detect unknown attacks effectively.

We propose a "supervised OOD detector." It aims to utilize the Malware Instruction Set (MIST) [4], consisting of simulated malicious attack sequences collected within a sandbox environment. While MIST can be directly analyzed, we transform it into an image format suitable for analysis to achieve better performance. Subsequently, a deep learning model will be trained to identify and classify known and unknown attack methods. To further categorize unknown attacks, we introduce a hybrid intrusion detection system to recognize unknown attacks.

Anticipated contributions encompass providing an effective detection system for unknown attacks, improving the model's generalization, enabling the system to better contend with emerging attack methodologies, and maintaining flexibility in adapting to evolving attack scenarios.

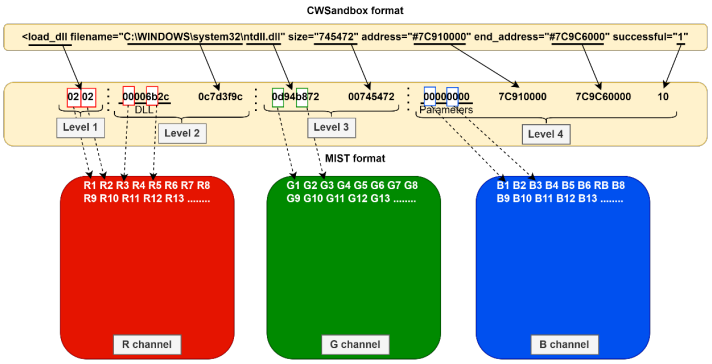

Fig. 1. MIST to image

II. Methods

A. Data preprocessing

The method utilizes two datasets converted into MIST format, with Trinius et al.'s approach [4] transforming data into image format, preserving sequential and semantic container instructions. MIST data has four levels: operation category (Level 1) to address (Level 4). Lower levels contain less variable information, while higher levels contain more variable data. Mapping involves:

- R channel: Combines operation category (Level 1) and file format (Level 2) due to high correlation.
- G channel: Includes file size and name (Level 3), less related to file format.
- B channel: Contains address (Level 4), the most variable and least informative. Indicates operation success but doesn't reveal much about container instruction behavior.

The method translates hexadecimal MIST data to decimal values, mapping them to pixels in their respective channels. For example, 'b8' in Level 3 Fig 1 becomes '11x16+8=184', positioned in the third pixel of the G channel. This process repeats for each MIST value, generating a 1024x1024 image for each container instruction. Utilizing image format, it captures features and patterns of container instructions, facilitating analysis and recognition of malicious attacks.

2024 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan) | 979-8-3503-8684-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCE-Taiwan62264.2024.10674502

B. Data augmentation

We employ data augmentation techniques to enhance the diversity and complexity of MIST data, improving model performance and generalization. Two methods are utilized:

- Mask: Following self-supervised learning approach, we use a mask to conceal pixels in input images, reconstructing them with a vision transformer.
- Noise addition: Various types of noise, including AdditiveGaussian, AdditiveLaplace, and AdditivePoisson, are added to input images to simulate situations where API information is obscured by attackers.

These techniques reduce the reliance on labeled data and facilitate the training of a more robust and accurate malware detection model.

C. Transformer-based image classifier

Our classifier distinguishes known attacks and categorizes normal and unknown instances into the "others" class for OOD detection. We adopt the Swin Transformer [5]. Swin Transformer applies self-attention within windows through W-MSA (window-based multi-head self-attention) and SW-MSA (shifted window-based multi-head self-attention) steps, addressing the Vision Transformer's computational complexities by progressively merging image patches.

D. OOD detector

In our OOD detector, we expand on the model's capability to classify the "Others" category [6]. To improve generalization and avoid data distribution limitations, we employ Ensemble learning. This approach combines multiple feature extraction methods, including global average pooling (GAP), global maximum pooling (GMP) [7], cross-dimensional weighting (CroW) [8], and selective convolutional descriptor aggregation (SCDA) [9]. By balancing individual detector sensitivity and capturing unique data characteristics, we lay the groundwork for effective anomaly detection.

We apply H-Regularization with 2-Norm instance-level normalization (HRN) [10] to normalize multiple feature sets, tackling feature-scale inconsistencies. Enhanced distance measurement via improved deep support vector data description (Deep-SVDD) [11] and mahalanobis distance (MD) [12] amplifies the gap between OOD and standard samples in the feature space. We also refine the classifier's output using a confidence score threshold filter. This approach enhances the accuracy of the OOD detector by distinguishing normal behavior from OOD unknown attacks.

III. Conclusion

We present a supervised OOD detector Fig 2 for adaptable intrusion detection. Container commands are transformed into RGB images using MIST data with mask and noise methods. Detection utilizes a Swin Transformer-based image classifier. Our approach integrates Ensemble learning, diverse feature extraction, and confidence score threshold filtering in the OOD detector. This system enhances comprehension of malicious attacks in microservice architecture by addressing feature scale inconsistencies and improving model comprehensiveness.

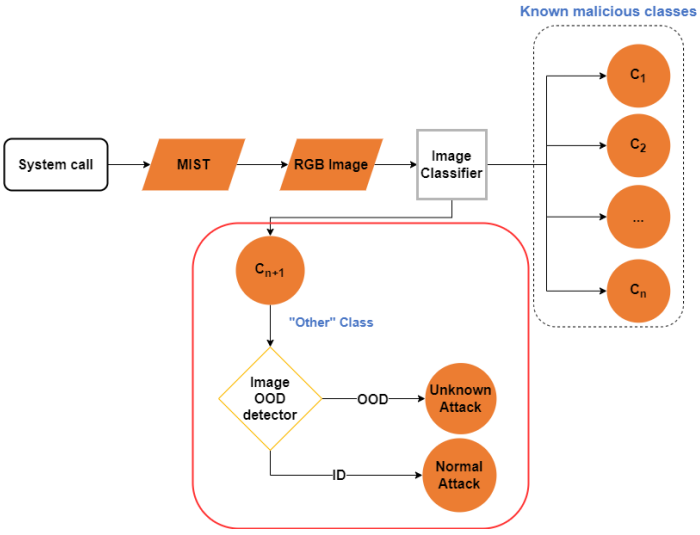

Fig. 2. System architecture

Acknowledgements

This work is sponsored by the National Science and Technology Council (NSTC) under the project NSTC 110-2222-E-008-008-MY3 and NSTC 112-2622-8-A49-021.

References

[1] Abir Rahali and Moulay A Akhloufi, "Malbertv2: Code aware bert-based model for malware identification," *Big Data and Cognitive Computing*, vol. 7, no. 2, pp. 60, 2023.

[2] Sachith Seneviratne, Ridwan Shariffdeen, Sanka Rasnayaka, and Nuran Kasthuriarachchi, "Self-supervised vision transformers for malware detection," *IEEE Access*, vol. 10, pp. 103121–103135, 2022.

[3] Niall McLaughlin and Jesus Martinez Del Rincon, "Data augmentation for opcode sequence based malware detection," in *IEEE Cyber Research Conference-Ireland*, 2022.

[4] Philipp Trinius, Carsten Willems, Thorsten Holz, and Konrad Rieck, *A malware instruction set for behavior-based analysis*, Gesellschaft für Informatik eV, 2010.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[6] Cheng-Hsueh Lin, Chia-Yu Lin, Li-Jen Wang, and Ted T Kuo, "Continual learning with out-of-distribution data detection for defect classification," in *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*, 2023, pp. 337–338.

[7] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network. arxiv 2013," *arXiv preprint arXiv:1312.4400*, 2013.

[8] Yannis Kalantidis, Clayton Mellina, and Simon Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Computer Vision–ECCV 2016 Workshops*, 2016.

[9] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2868–2881, 2017.

[10] Wenpeng Hu, Mengyu Wang, Qi Qin, Jinwen Ma, and Bing Liu, "Hrn: A holistic approach to one class learning," *Advances in neural information processing systems*, vol. 33, pp. 19111–19124, 2020.

[11] "Anomaly detection using improved deep svdd model with data structure preservation," *Pattern Recognition Letters*, vol. 148, pp. 1–6, 2021.

[12] Ryo Kamoi and Kei Kobayashi, "Why is the mahalanobis distance effective for anomaly detection?," *ArXiv*, vol. abs/2003.00402, 2020.