International Conference on Innovation in Engineering and Technology (ICIET) 23-24 December, 2019

A Systematic Review on Hybrid Analysis using Machine Learning for Android Malware Detection

Author 1

Asadullah Hill Galib
MSSE 0718
Institute of Information Technology
University of Dhaka

Author 2

Dr. B. M. Mainul Hossain
Associate Professor
Institute of Information Technology
University of Dhaka

Outline

- Introduction
- Background
- Methodology
- Literature Overview
- Discussions

Introduction

- First review of the existing works on the hybrid analysis approach
- The prevalence of hybrid analysis over static analysis and dynamic analysis
- Discussion about the challenges, opportunities and future directions of hybrid analysis

Background : Android Malware

Android Malware (malicious application) is any application with mischievous intention -

- disrupt normal functioning
- bypass access controls
- gather sensitive information
- display unwanted advertising
- getting unauthorized control

Background : Detection Techniques

- Static Analysis
- Dynamic Analysis
- Hybrid Analysis

Background : Drawbacks of Static and Dynamic

Static Analysis	Dynamic Analysis
◎ Data obfuscation	◎ Limited code coverage
◎ Control flow obfuscation	◎ Tricked in emulated environment by smart malware
◎ Dynamic XML loading	
◎ Native code	
◎ Encryption	

Hybrid Analysis

- A fusion of static and dynamic analysis
- Would be a good candidate as it prevails over the individual drawbacks of static and dynamic analysis.

Methodology

- For the systematic literature review, we have followed a state-of-the-art guideline presented by Kitchenham and Stuart -
  - *Guidelines for performing systematic literature reviews in software engineering, EBSE, 2007*[1].

Methodology: Review Protocol

- The rationale for the review
- Research questions
- Search strategy
- Study selection criteria
- Study selection procedures
- Study quality assessment
- Data extraction
- Data synthesis

Methodology: Research Questions

- What are the static and dynamic features used in hybrid analysis using machine learning?
- What are the most common dataset sources of the existing literature?
- Which machine learning algorithms are most frequently used in the existing researches?

Methodology: Research Questions

- Which evaluation metrics are most widely used in the existing literature?
- What are the evaluation results of the existing researches?
- What are the limitations of the existing literature?

Consequential Literature

- Mobile-Sandbox (2013)
- Marvin (2015)
- HADM (2016)
- Samadroid (2018)

Mobile-sandbox: having a deeper look into android applications

Authors: Michael Spreitzenbarth, Felix Freiling, Florian Echtler, Thomas Schreck, Johannes Hoffmann

Published in: 2013, ACM Symposium on Applied Computing

Citations: 275

Mobile-SandBox: Features

Static Features:

- Permissions
- Services, Receivers
- Advertising networks
- Dangerous functions
- Encryption libraries

Dynamic Features:

- Dalvik & Itrace log
- PCAP file
- Native Code

Mobile-SandBox: Approach

- Automated analysis
- Dynamic Analyzer Component
- Evaluation - Correctness, Detectability, Performance, and Scalability

Mobile-SandBox: Contributions

- Combines static and dynamic analysis and results of static analysis are used to guide dynamic analysis
- Extends coverage of executed code
- Uses specific techniques to log calls to native (i.e., \non-Java) APIs.

Mobile-SandBox: Limitation

- Lacking in performance as no solid performance metrics given

Marvin: Efficient and comprehensive mobile app classification through static and dynamic analysis

Authors: Martina Lindorfer, Matthias Neugschwandtner, Christian Platzer

Published in: 2015, IEEE Computer Software and Applications Conference

Citations: 107

Marvin: Features

Statice Features:

- Permissions
- Intents
- Suspicious Files
- API Calls
- Developer's Certificate

etc.

Dynamic Features:

- File Operations
- Network Operations
- Phone Events
- Dynamically Loaded Code etc.

Marvin: Approach

- Large dataset
- Large feature-set
- ML Technique: SVM & Regularized Logistic Regression
- Feature Selection: Fisher Score (F-score)

Marvin: Contributions

- Provides retraining strategy (effective for new malware family)
- Provided applicable mobile app option
- Outperforms existing approaches

Marvin: Limitation

- Static analysis outperforms dynamic analysis significantly
- Overlooking system-level events such as System Calls

HADM: Hybrid analysis for detection of malware

Authors: Lifan Xu, Dongping Zhang, Nuwan Jayasena, John Cavazos

Published in: IntelliSys 2016: SAI Intelligent Systems Conference

Citations: 22

HADM: Features

Statice Features:

- Permissions
- API Calls
- Intents

Dynamic Features:

- System Call Sequences

HADM: Approach

- Deep Neural Network for feature extraction
- Deep Auto Encoder: learning model
- Multiple Kernel Learning used for combining learning results
- Restricted Boltzmann Machine for DNN
- Feature Vector: n-gram representation
- ML Technique: SVM

HADM: Contributions

- Using deep learning to learn new features
- Applying hierarchical MKL to combine different kernel learning thus further improve classification accuracy

HADM: Limitation

- Limited feature set and data-set
- Higher complexity
- Performance: not high enough relatively

Samadroid: a novel 3-level hybrid malware detection model for android operating system

Authors: Saba Arshad, Munam A. Shah, Abdul Wahid, Amjad Mehmood, Houbing Song, Hongnian Yu

Published in: 2018, IEEE Access (Volume: 6)

Citations: 22

Samadroid: Features

Statice Features:

- Permissions
- API Calls
- Intents
- App Components

Dynamic Features:

- System Calls (10)

Samadroid: Approach

- Local and remote host
- Drebin static feature sets
- ML Technique: SVM, Random Forest and Decision Tree
- Feature Selection: manual/logical analysis

Samadroid: Contributions

- 3-level on-device malware detection architecture
- Ensures the resource efficiency by reducing memory overhead of local devices
- Provides explanation to users about the behavior of application

Samadroid: Limitation

- Overlooking many dynamic features (other system call)
- Using limited and old dataset

Other Literature

- Liu et al. (2016):
  - Permissions are used as static features and System Calls used as dynamic features.
  - Small feature-set and their dataset is also limited.
  - Limited Feature-set

Other Literature

- Kapratwar et al. (2017) :
  - Permissions and System Calls for hybrid analysis.
  - Used a small (200 apps) and old dataset
  - Overlooked many static and dynamic features

Other Literature

- Dhanya et al. (2019):
  - Permissions and API Calls
  - Separability assessment Criteria for feature selection
  - Limited feature-set and dataset
  - Overlooked many static and dynamic features

Dataset

- **Drebin**
- **Android Malware Genome Project**
- ContagioDump
- VirusTotal
- VirusShare

Machine Learning Techniques

- **SVM**
- Naive Bayes
- Random Forest
- J48
- Logistic Regression

Discussions

- Better Performance
- Lack of Research
- Dataset Inadequacy
- Exploring New Feature
- New Malware Family
- Reducing Complexity

References

1. Keele, Staffs. Guidelines for performing systematic literature reviews in software engineering. Vol. 5. Technical report, Ver. 2.3 EBSE Technical Report. EBSE, 2007.
2. M. Spreitzenbarth, F. Freiling, F. Echtler, T. Schreck, and J. Hoffmann, “Mobile-sandbox: having a deeper look into android applications,” in Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 1808–1815, ACM, 2013.
3. M. Lindorfer, M. Neugschwandtner, and C. Platzer, “Marvin: Efficient and comprehensive mobile app classification through static and dynamic analysis,” in 2015 IEEE 39th annual computer software and applications conference, vol. 2, pp. 422–433, IEEE, 2015.
4. L. Xu, D. Zhang, N. Jayasena, and J. Cavazos, “Hadm: Hybrid analysis for detection of malware,” in Proceedings of SAI Intelligent Systems Conference, pp. 702–724, Springer, 2016.
5. S. Arshad, M. A. Shah, A. Wahid, A. Mehmood, H. Song, and H. Yu, “Samadroid: a novel 3-level hybrid malware detection model for android operating system,” IEEE Access, vol. 6, pp. 4321–4339, 2018.

References

6.A. Kapratwar, F. Di Troia, and M. Stamp, “Static and dynamic analysis of android malware.,” in ICISSP, pp. 653–662, 2017.
7. K. D. T. Gireesh Kumar, “Efficient android malware scanner using hybrid analysis,” International Journal of Recent Technology and Engineering(TM), vol. 7, pp. 76–80, 2019.
8. Y. Liu, Y. Zhang, H. Li, and X. Chen, “A hybrid malware detecting scheme for mobile android applications,” in 2016 IEEE International Conference on Consumer Electronics (ICCE), pp. 155–156, IEEE, 2016.

Thank you!

Any questions?