VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

GAN architektūrų, tinkamų kenkėjiško kodo obfuskacijai, analizė

Analysis of GAN architectures suitable for Ethical Malware Obfuscation

Kursinis darbas

Atliko: 4 kurso 3 grupės studentas
Liudas Kasperavičius

Darbo vadovas: prof. dr. Olga Kurasova

Vilnius – 2024

Turinys

Įvadas

Pastaraisiais metais kenkėjiškas kodas ir programos kuriamos itin sparčiai (∼450000 kenkėjiškų programų per dieną AV-TEST duomenimis). Kenkėjiško kodo aptikimo programos, kurios tradiciškai remiasi programų pėdsakais (*angl. signature*), nespėja atnaujinti pėdsakų duomenų bazių pakankamai greitai. Dėl to Dirbtinio Intelekto (DI), tiksliau Mašininio Mokymosi (ML), naudojimas kenkėjiškų programų ar kenkėjiško kodo aptikimo srityje tapo itin populiarus [DCB+21]. Tačiau ML modeliai, nors ir geba aptikti kenkėjiškas programas iš naujų, dar nematytų, duomenų, yra pažeidžiami varžymosi principais pagrįstoms atakoms (*angl. adversarial attacks*) [CSD19; HT17; RSR+18; ZHZ+22]. Šių atakų principas yra ML modelio sprendimų priėmimo ribos (*angl. decision boundary*) radimas – žinant šią ribą pakanka pakeisti kenkėjiškos programos veikimą taip, kad ML modelis priimtų sprendimą klasifikuoti ją kaip nekenksmingą [DCB+21]. Žinoma, rasti šią ribą nėra trivialus uždavinys. Mokslinėje literatūroje išskiriami 3 ribos paieškos atvejai [FWL+19]:

1. **Baltos dėžės** atvejis: kenkėjiško kodo kūrėjas turi visą informaciją apie ML modelį, t. y. modelio architektūrą, svorius, hiperparametrus.
2. **Juodos dėžės su pasitikėjimo įverčiu** atvejis: kenkėjiško kodo kūrėjas gali tik testuoti modelį – t. y. pateikti programą ir gauti atsakymą. Atsakymo forma – klasifikacija ir tikimybė, kad klasifikacija yra teisinga (pasitikėjimo įvertis).
3. **Juodos dėžės** atvejis: kenkėjiško kodo kūrėjas gali tik testuoti modelį. Atsakymo forma yra tik klasifikacija.

Akivaizdu, jog „juodos dėžės“ atvejis yra sudėtingiausias, bet ir labiausiai atitinka realias sąlygas **[CITE]**. Todėl šiame darbe nagrinėjami modeliai, gebantys generuoti varžymosi principais pagrįstų atakų obfuskuotus kenkėjiško kodo pavyzdžius (AE) „juodos dėžės“ atvejams.

Tikslas – nustatyti labiausiai tinkantį modelį varžymosi principais pagrįstoms atakoms „juodos dėžės“ atvejais.

Uždaviniai:

1. Apžvelgti kenkėjiško kodo obfuskacijos metodus
2. Nustatyti kriterijus varžymosi principais grįstų atakų modeliams ir juos įvertinti
3. Atlikti eksperimentinį tyrimą naudojant modelį, gavusį aukštą įvertinimą pagal kriterijus

1. Literatūros apžvalga

Malware detectors, Frameworks

1.1. Naudojami požymiai

Pirmasis veiksmas treniruojant ar naudojant ML modelį, paremtą neuroniniais tinklais, yra paversti įvesties duomenis į požymių vektorių **[CITE]**. Kenkėjiško kodo obfuskacijos kontekste, požymių pasirinkimas nėra vienareikšmis, o įeina į karkaso apibrėžimą. Šioje sekcijoje apžvelgiami ir klasifikuojami GAN, RL ir GA modelių tipų karkasų naudojami požymiai.

- DLL vardai (arba API vardai). PE faile turi būti nurodyti visi naudojami DLL. Prieš pradedant treniruoti ML modelį, atliekama visų turimų programų analizė ir nustatoma visų naudojamų DLL aibė D. Tarkime $|D| = n$. Tuomet, požymių vektorius programai, naudojančiai $X \subseteq D$ DLL, bus n-matis dvejetainis vektorius, kurio i-asis elementas yra $\begin{cases} 0, \text{ jei } D_i \notin X, \\ 1, \text{ jei } D_i \in X. \end{cases}$ Čia D_i – i-asis D elementas.
- n-gramos. Dažniausiai sutinkamos skaitmeniniame natūraliosios kalbos apdorojime (NLP). Tai yra n žodžių junginiai, arba, sukompiliuotų programų apdorojimo kontekste, n baitų junginiai. Nustatant požymių vektorių, visos n-gramos surikiuojamos pagal pasikartojimą programoje mažėjimo tvarka („populiariausios“ viršuje). Iš pirmų m reikšmių sudaromas m-matis vektorius – tai ir yra požymių vektorius.
- Baitų/entropijos histograma. Specifinis metodas, užkoduojantis dažniausiai pasikartojančias baitų ir entropijos poras 256 dimensijų vektoriumi [SB15].

1.2. Perturbacijos

Perturbacijos – tai pagrindinis obfuskacijos metodas AE kūrimui. Perturbacijų tikslas yra pakeisti kenkėjiškos programos veikimą išsaugant originalų funkcionalumą. Perturbacijos gali būti sudėtingos ir apimti visą programą (pvz. visos programos užšifravimas ir pridėjimas prie kitos programos), semantinės (pvz. tam tikrų mašininio kodo instrukcijų keitimas į ekvivalentų rezultatą pasiekiančias) arba baitų lygio (pvz. nulinių baitų pridėjimas programos gale) **[CITE]**. Perturbacijų parinkimas įeina į karkaso apibrėžimą. Šioje sekcijoje aptariamos mokslinėje literatūroje minimos perturbacijos.

-

1.3. GAN tipo modelių karkasai

GAN modeliai paremti Generatyviniais Priešiškais Tinklais (*angl. Generative Adversarial Networks*), kurių veikimo principas yra du neuroniniai tinklai (generatorius ir diskriminatorius), žaidžiantys nulinės sumos žaidimą (*angl. zero-sum game*) **[CITE]**. Kenkėjiško kodo obfuskacijos kontekste ir ypač „juodos dėžės“ atvejais, diskriminatorius atlieka surogatinio modelio vaidmenį. Bendras GAN modelių mokymosi etapas yra tokia seka:

1. generatorius, naudodamas požymių vektorių ir tokios pačios dimensijos „triukšmo“ (*angl. noise*) vektorių, sugeneruoja perturbacijas
2. originali kenkėjiška programa modifikuojama pagal perturbacijas (sukuriamas AE)
3. diskriminatorius bando klasifikuoti sugeneruotą AE (kenkėjiškas / nekenkėjiškas). Diskriminatoriaus klasifikacija lyginama su tikro detektoriaus klasifikacija. Jei ji teisinga – atnaujinami generatoriaus parametrai pagal generatoriaus nuostolių funkciją. Kitu atveju, atnaujinami diskriminatoriaus parametrai pagal diskriminatoriaus nuostolių funkciją.
4. visa seka kartojama nustatytą kiekį kartų

[CITE].

1.4. Skatinamojo mokymosi tipo modelių karkasai

Skatinamojo mokymosi (RL) modeliai susideda iš agento ir aplinkos. Aplinka susideda iš informatyvių požymių ištraukimo metodo (*angl. feature extraction*) ir kenkėjiškų programų detektoriaus. Agentas – tai algoritmas ar neuroninis tinklas, kurio tikslas yra surasti optimalią strategiją (*angl. policy*). Šiuo atveju strategija susideda iš perturbacijų (žr. 1.2) **[CITE]**. Bendras RL modelių mokymosi etapas yra tokia seka:

1. agentas, naudodamas dabartinę aplinkos būseną ir praeito veiksmo atlygį (*angl. reward*), parenka sekantį veiksmą iš galimų veiksmų aibės
2. atliekamas veiksmas – perturbuojama programa arba požymių vektorius (priklauso nuo karkaso)
3. gaunami aplinkos kitimo įverčiai – nauja būsena ir atlygis, skaičiuojamas pagal detektoriaus klasifikacijos rezultatą
4. seka kartojama tol, kol agentas nelaiko strategijos optimalia arba nustatytą kiekį kartų

[CITE]

1.5. Genetinių algoritmų tipo modelių karkasai

Genetinai algoritmai (GA) yra viena seniausių mašininio mokymosi apraiškų; jų veikimas paremtas evoliucija **[CITE]**. Kenkėjiškų programų obfuskacijai AE generavimas taikant GA yra tokia seka:

1. sukuriama pradinė populiacija (perturbacijos metodai pradinei populiacijai priklauso nuo karkaso)
2. atliekamas vertinimas naudojant detektorių
3. Jei vertinimo metu nustatoma, jog AE yra pakankamai geros kokybės (pvz. detektorius klasifikuoja kaip nekenksmingą), seka baigiama
4. atliekama selekcija – dažniausiai pasirenkami geriausiai įvertinti populiacijos AE, tačiau galimos ir kitos selekcijos strategijos
5. atliekamas selekcijos atrinktų AE kryžminimas (po 2) taip sukuriant naują AE, turintį po dalį genų iš abiejų kryžmintų AE

6. tam tikrai daliai AE atliekama dalies genų mutacija
7. gaunama nauja populiacijos karta ir seka kartojama nuo 2-o žingsnio

[YPT22]

1.6. Nevalidaus PE formato problema

2. Modelių vertinimas

3. Eksperimentinis tyrimas

Rezultatai ir išvados

Rezultatų ir išvadų dalyje turi būti aiškiai išdėstomi pagrindiniai darbo rezultatai (kažkas išanalizuota, kažkas sukurta, kažkas įdiegta) ir pateikiamos išvados (daromi nagrinėtų problemų sprendimo metodų palyginimai, teikiamos rekomendacijos, akcentuojamos naujovės).

Šaltiniai

[CSD19] R. L. Castro, C. Schmitt, G. Dreo. AIMED: Evolving Malware with Genetic Programming to Evade Detection. Iš: *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. 2019, p. 240–247 [žiūrėta 2024-09-23]. ISSN 2324-9013. Prieiga per internetą: https://doi.org/10.1109/TrustCom/BigDataSE.2019.00040.

[DCB+21] L. Demetrio, S. E. Coull, B. Biggio, G. Lagorio, A. Armando, F. Roli. Adversarial EXEmples: A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection. *ACM Trans. Priv. Secur.* 2021, tomas 24, numeris 4, 27:1–27:31 [žiūrėta 2024-09-30]. ISSN 2471-2566. Prieiga per internetą: https://doi.org/10.1145/3473039.

[FWL+19] Z. Fang, J. Wang, B. Li, S. Wu, Y. Zhou, H. Huang. Evading Anti-Malware Engines With Deep Reinforcement Learning. *IEEE Access*. 2019, tomas 7, p. 48867–48879 [žiūrėta 2024-09-18]. ISSN 2169-3536. Prieiga per internetą: https://doi.org/10.1109/ACCESS.2019.2908033.

[HT17] W. Hu, Y. Tan. *Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN*. 2017-02-20. [žiūrėta 2024-09-18]. Prieiga per internetą: http://arxiv.org/abs/1702.05983.

[YPT22] J. Yuste, E. G. Pardo, J. Tapiador. Optimization of Code Caves in Malware Binaries to Evade Machine Learning Detectors. *Computers & Security*. 2022, tomas 116, p. 102643 [žiūrėta 2024-10-07]. ISSN 0167-4048. Prieiga per internetą: https://doi.org/10.1016/j.cose.2022.102643.

[RSR+18] I. Rosenberg, A. Shabtai, L. Rokach, Y. Elovici. Generic Black-Box End-to-End Attack Against State of the Art API Call Based Malware Classifiers. Iš: M. Bailey, T. Holz, M. Stamatogiannakis, S. Ioannidis (sudarytojai). *Research in Attacks, Intrusions, and Defenses*. Cham: Springer International Publishing, 2018, p. 490–510. ISBN 978-3-030-00470-5. Prieiga per internetą: https://doi.org/10.1007/978-3-030-00470-5_23.

[SB15] J. Saxe, K. Berlin. Deep Neural Network Based Malware Detection Using Two Dimensional Binary Program Features. Iš: *2015 10th International Conference on Malicious and Unwanted Software (MALWARE)*. 2015, p. 11–20 [žiūrėta 2024-11-04]. Prieiga per internetą: https://doi.org/10.1109/MALWARE.2015.7413680.

[ZHZ+22] F. Zhong, P. Hu, G. Zhang, H. Li, X. Cheng. Reinforcement Learning Based Adversarial Malware Example Generation against Black-Box Detectors. *Computers & Security*. 2022, tomas 121, p. 102869 [žiūrėta 2024-09-14]. ISSN 0167-4048. Prieiga per internetą: https://doi.org/10.1016/j.cose.2022.102869.

Sąvokų apibrėžimai

Adversarial attacks. Varžymosi principais pagrįstos atakos. 3

Decision Boundary. Sprendimų priėmimo riba. 3

Signature. Kodo/programos pėdsakas. 3
Strategija. *angl. Policy*. RL modelio atliekama veiksmų seka. 5
Surogatinis Modelis. ML modelis, aproksimuojantis kitą ML modelį, kurio parametrai (svoriai) nėra žinomi. 4

Zero-sum Game. Dviejų žaidėjų žaidimas, kuriame galimas vienas laimėtojas. Laimėtojo laimėta suma yra lygi pralaimėtojo pralaimėtai sumai. 4

Santrumpos

AE. Varžymosi principais pagrįstomis atakomis obfuskuoti kenkėjiško kodo pavyzdžiai (*angl. Adversarial Examples*)

API. *angl. Application Programming Interface*

DI. Dirbtinis Intelektas

DLL. *angl. Dyanmic-Link Library*

GA. Genetiniais algoritmais pagrįstas ML modelis

GAN. Generatyviniai Priešiški Tinklai (*angl. Generative Adversarial Networks*) (generatyviniai varžymosi principais pagrįsti tinklai)

ML. Mašininis Mokymasis (*angl. Machine Learning*)

NLP. Skaitmeninis natūraliosios kalbos apdorojimas (*angl. Natural Language Processing*)

PE. *angl. Portable Executable*

RL. Skatinamasis Mokymasis (*angl. Reinforcement Learning*)