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 **[NUMBERS NEEDED]**. 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 **[CITATION NEEDED]**. Tačiau ML modeliai, nors ir geba aptikti kenkėjiškas programas iš naujų, dar nematytų, duomenų, yra pažeidžiami varžymosi principais pagrįstoms atakoms (*angl. adversarial attacks*) **[CITATION NEEDED]**. 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 **[CITATION NEEDED]**. Žinoma, rasti šią ribą nėra trivialus uždavinys. Mokslinėje literatūroje išskiriami 3 ribos paieškos atvejai:

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 **[CITATION NEEDED]**. 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 su geriausiai kriterijus atitinkančiu modeliu

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).

Santrumpos

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

DI. Dirbtinis Intelektas

ML. Mašininis Mokymasis