Vilniaus universitetas

Doktorantas:
Paulius Vaitkevičius

Vadovas:
Dr. Virginijus Marcinkevičius

III metų II pusmečio ataskaita
2021 m. spalio 1 d.

Mašininiu mokymusi grįstų atvirųjų šaltinių žvalgybos informacijos išskyrimo ir analizės metodai

Doktorantūros laikotarpis: 2018 - 2022

Vilniaus universitetas

TURINYS

1. Studijų plano vykdymas
2. Trumpas per pusmetį gautų mokslinių rezultatų pristatymas
3. Problemos apibrėžimas, tyrimo objektas, tikslai ir planuojami gauti rezultatai
4. Kito pusmečio darbo planas

STUDIJŲ PLANO VYKDYMAS

Vilniaus universitetas

Visų studijų planas, vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose		Publikacijos		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė
I (2018/2019)	1	1		1			
II (2019/2020)	1	3		1		1	Publikuota
III (2020/2021)	**2**		**1**		**1**	**1**	**Įteikta**
IV (2021/2022)			1		1		

Vilniaus universitetas

Ataskaitinio pusmečio darbo planas ir jo įvykdymas

Egzaminai		Publikacijos		Dalyvavimas konferencijose	
Planas	**Įvykdyta**	**Planas**	**Įvykdyta**	**Planas**	**Įvykdyta**
1. Didžiųjų duomenų analitika 2. Mašininis mokymasis	Išlaikyta II studijų metais	Tyrimo rezultatų pristatymas tarptautinėje mokslinėje konferencijoje.	Šis uždavinys įvykdytas II studijų metais	Empirinio tyrimo rezultatų publikavimas (recenzuojamame leidinyje, CA WoS su Impact Factor).	Publikacija parašyta ir įteikta recenzuojam leidiniui „Machine Learning“ (CA WoS su Impact Factor)

Vilniaus universitetas

Visų mokslinių tyrimų ir disertacijos rengimo etapai

1. Mokslinių tyrimų disertacijos tema apžvalga ir analizė
2. Mokslinio tyrimo vykdymas:
    1. Tyrimo metodikos sudarymas
    2. Teorinis tyrimas
    3. Empirinis tyrimas
    4. Gautų duomenų analizė, apibendrinimas, išvadų parengimas
3. Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas
4. Daktaro disertacijos parengimas ir svarstymas padalinyje
5. Daktaro disertacijos gynimas

PROBLEMOS APIBRĖŽIMAS, TYRIMO OBJEKTAS, TIKSLAI IR PLANUOJAMI GAUTI REZULTATAI

Vilniaus universitetas

Tyrimo tikslas

Sukurti apsimetinėjimo atakoms atsparų metodą, grįstą giliaisiais neuroniniais tinklais ir natūralios kalbos apdorojimo algoritmais, kuris leistų efektyviai ir patikimai atpažinti **duomenų išviliojimo internete** (angl. „Phishing“) tinklapius.

Vilniaus universitetas

Tyrimo objektas

1. Mašininio mokymo ir giliojo mašininio mokymo algoritmai, skirti atpažinti duomenų išviliojimo internete tinklapius.
2. Atsparūs priešiškoms atakoms algoritmai (angl. „Adversarial Machine Learning“).

Vilniaus universitetas

Tyrimo uždaviniai

1. Atlikti literatūros analizę, išanalizuoti state-of-the-art algoritmus duomenų išviliojimo internete tinklapių atpažinimui.
2. Atkartoti *state-of-the-art* algoritmų rezultatus.
3. Sukurti duomenų rinkinius eksperimentų vykdymui.
4. Pasiūlyti naują efektyvesnį duomenų išviliojimo internete tinklapių atpažinimo metodą.
5. Atlikti eksperimentinius tyrimus, palyginant pasiūlytą metodą su *state-of-the-art* algoritmais.

Vilniaus universitetas

Planuojami rezultatai

1. Atlikta **literatūros analizė**, palyginant pažangiausius tyrimo srities algoritmus;
2. Atlikti **eksperimentiniai tyrimai**:
   - ✓ Mašininio mokymosi algoritmų efektyvumo palyginimas;
   - ✓ Giliojo mašininio mokymosi (GMM) algoritmų efektyvumo palyginimas;
   - ✓ GMM algoritmų (RNN, LSTM, GRU, CNN, kt.) efektyvumo tyrimai, naudojant natūralaus teksto apdorojimo technikas (N-grams, word embedings, kt.).
   - ✓ Naujo GMM algoritmo kūrimas, sprendžiant apibrėžtus uždavinius.
   - ✓ Pasiūlyto GMM algoritmo eksperimentinis tyrimas analizuojant jo efektyvumą;
   - ✓ Pasiūlyto GMM algoritmo atsparumo priešiškoms atakoms (angl. „Adversarial Machine Learning) eksperimentiniai tyrimai.

PER PUSMETĮ GAUTŲ MOKSLINIŲ REZULTATŲ PRISTATYMAS

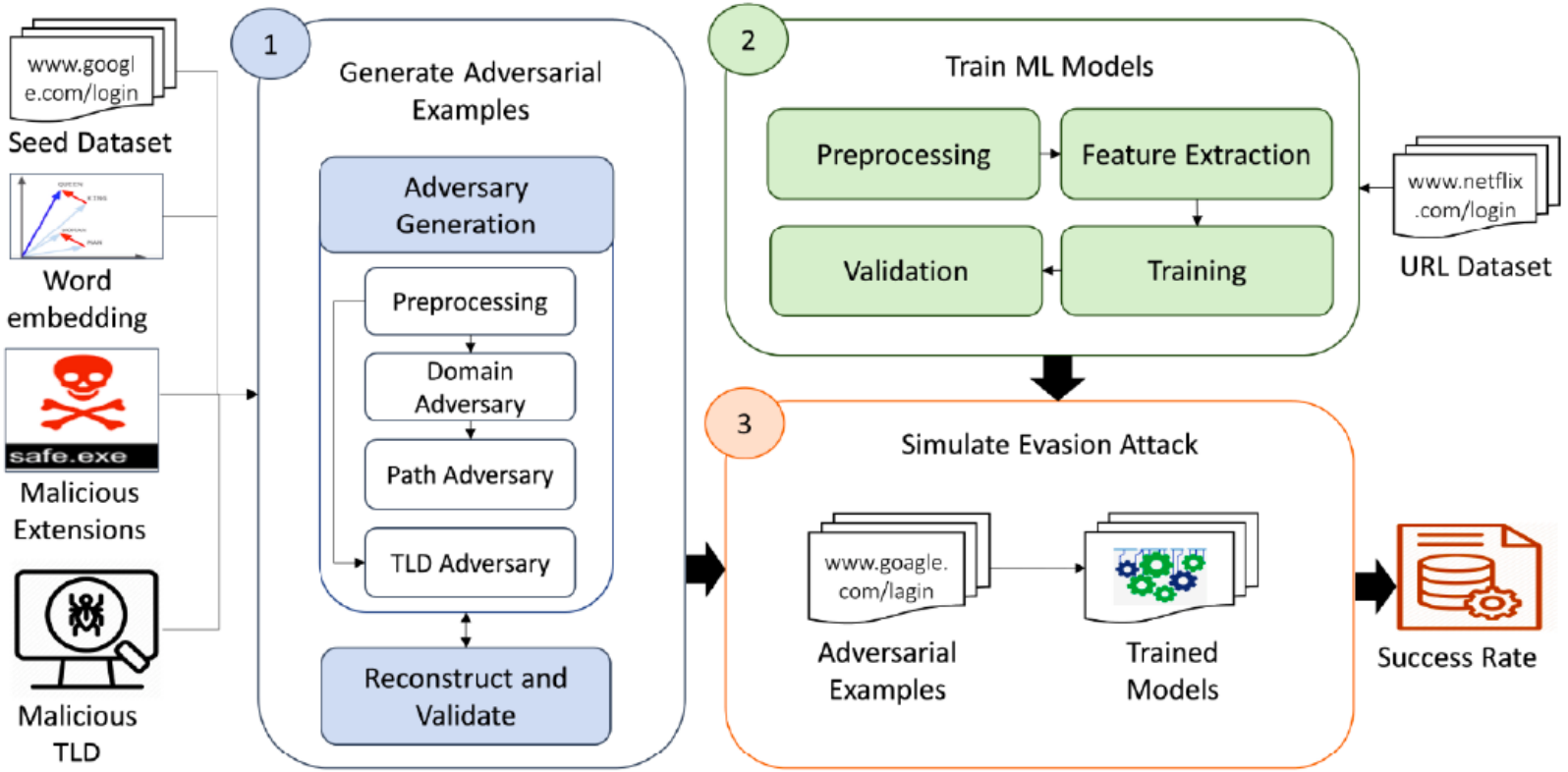
KITO PUSMEČIO PLANAS

Tyrimų motyvacija

Vilniaus universitetas

2020 metais Sabir et al. [11] bei kiti [13, 7] pademonstravo, kad ML grįsti *state-of-the-art* sukčiavimo internete atpažinimo metodai (kurių deklaruojamas tikslumas $> 98\%$) yra smarkiai pažeidžiami (pažeidžiamumo sėkmė $> 66\%$).

[11] naudota architektūra:



Tyrimų motyvacija

Vilniaus universitetas

[11] naudotų obfuskacijos technikų pavyzdžiai:

Type	Level	Obfuscation Method	Target *x* (https://store.steampowered.com/)
Domain	Char	Addition	https://store.steampowereda.com
		Insertion	https://store.steaompowered.com
		BitSquatting	https://store.steampowerel.com
		Homoglyph	http://https://store.steamp0wered.com
		Omission	https://store.seampowered.com
		SubDomain	https://store.steam.powered.com
		Hyphenation	https://store.st-eampowered.com
		CharacterSwap	https://store.steampowired.com
		Repetition	https://store.steaampowered.com
		Transpose	https://store.stemopowered.com
	Word	WordSubDomain	https://store.steampowered.ai-assisted.com
		WordHyphenation	https://store.ai-assisted-steampowered.com
		WordRepetition	https://store.steampowered-steampowered.com
		WordSwap	https://store.poweredsteam.com/
Path	Word	PathDm	https://store.steampowered-operated.com/steampowered
		PathExe	https://store.steampowered-operated.com/steampowered.exe
TLD	Word	TldReplace	https://store.steampowered.in.rs

Tyrimų tikslas ir idėja

Vilniaus universitetas

Tolimesnių tyrimų tikslas:

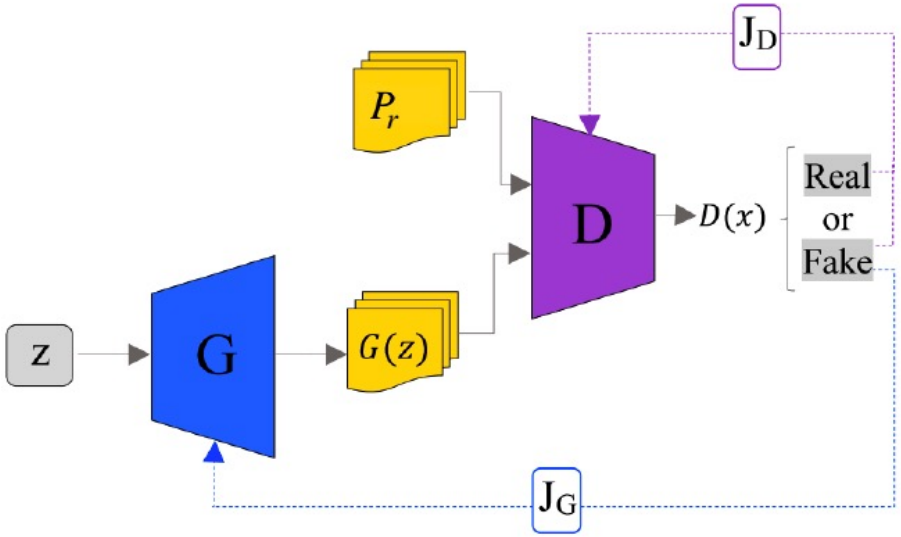
- sukurti apsimetinėjimo atakoms atsparų metodą,
- ištestuoti jo efektyvumą, naudojant Sabir et al. [11] ir kitus pažeidžiamumo atakų įrankius,
- ir palyginti sukurtą metodą su šios srities *state-of-the-art* metodais.

Tyrimo idėja: sukurti automatinį generatyvinį modelį, skirtą generuoti fišingo puslapių URL:

- Palyginti automatinio modelio efektyvumą su Sabir et al. [11] metodu, kai atakos prieš mašininio mokymi modelius buvo kuriamos žmogaus;
- Sugeneruoti didelę sintetinių mokymo ir testavimo duomenų aibę vėlesniam nuosavo atsparaus apsimetinėjimo atakoms sukčiavimo internete metodo kūrimui.

Nėra publikacijų, kur su generatyviniais modeliais būtų generuojami teksiniai fišingo URL adresai.

GAN



G - generatorius D - diskriminatorius (kritikas),
P_r - realių duomenų imtis,
$G(z)$ - dirbtinių duomenų imtis,
z - atsitiktinis triukšmas (iš tolygiojo arba *Gauso* skirtinio)

- Turint g_θ ir atsitiktinio triukšmo imtį $z \sim Z$, dirbtiniai duomenys generuojami $g_\theta(z)$.
- Funkcija g_θ yra Generatyvinių besivaržančiųjų tinklų (angl. Generative Adversarial Networks, GAN) generatorius G [5].
- GAN yra neprižiūrimojo mokymosi (angl. Unsupervised learning) metodas, kai du besivaržantys tinklai (generatorius ir diskriminatorius) mokomi lygiagrečiai

Klasikinis GAN (pvz. DCGAN) yra nulinės sumos žaidimas, kai generatorius nori minimizuoti, o diskriminatorius maksimizuoti nuostolio funkciją:

$$\min_G \max_D \mathbb{E}_{x \sim p_r} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z))]$$

GAN

Vilniaus universitetas

- Nykstančio gradiento problema - jei D ypač ankstyvosioje mokymo fazėje išmokomas tobulai, tai G gradientas išnyksta ir G nebesimoko,
- "Režimo griūtis" (angl. mode collapse) - G visada grąžina tą patį rezultatą ir nebesimoko,
- Nešo ekvilibriumas (angl. Nash equilibrium) sunkiai pasiekiamas.

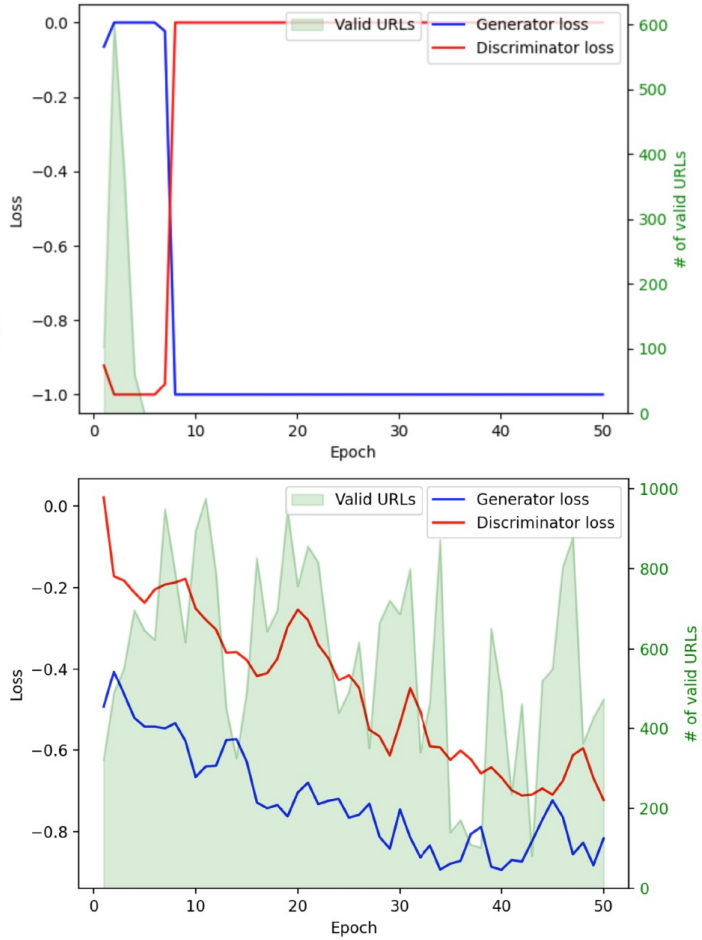
Naudojamas Wasserstein GAN (WGAN)

- Nors WGAN optimizavimas dėl "geriau besielgiančių gradientų" lengvesnis, nei GAN, tačiau "išnysktančių" ir "sprogstančių" gradientų problema išlieka.
- Gulrajani et al. [6] pasiūlė sprendimą, *Lipschitz* sąlygas užtikrinti ne D svorių karpymu, o D gradientų bauda *(angl. gradient penalty)*.
- **WGAN-GP** nuostolio funkcija, užrašoma taip:

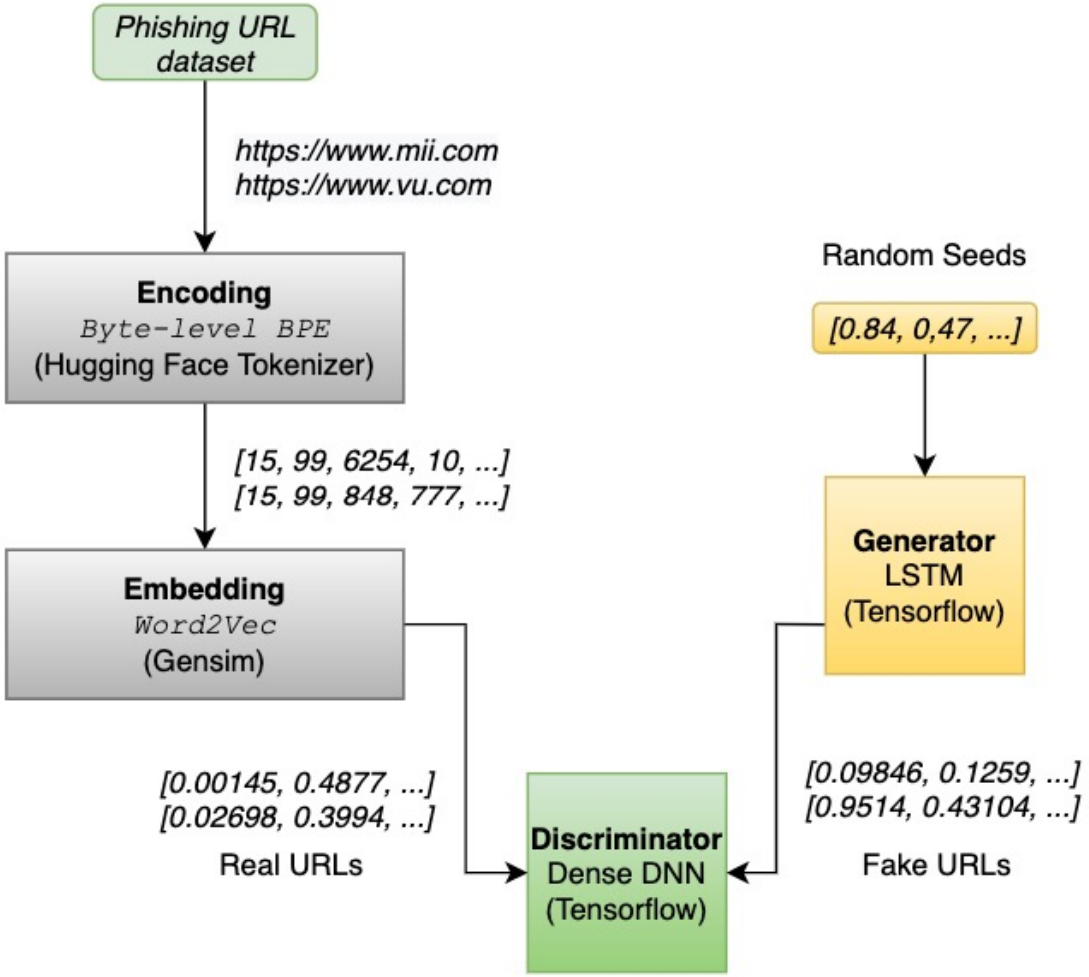
$$\min_G \max_{D \in \mathcal{D}} \mathbb{E}_{x \sim p_r} [D(x)] - \mathbb{E}_{z \sim p_z} [D(G(z))] + \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}} \left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \right]$$

kur $\mathcal{D}$ - aibė *1-Lipschitz* funkcijų, $p_{\hat{x}}$ - atsitiktinė imtis iš taškų, jungiančių p_r ir p_z, λ - baudos koeficientas.

- Vietoje svorių karpymo, siekiant užtikrinti *Lipschitz* sąlygą, vykdomas kritiko gradiento L_2 normos reguliarizavimas, kad ji nebūtų daugiau už 1.



Pasiūlytas GAN




```
t = phish_tokenizer.encode \
("https://www.mii.lt.myuniversity.com")
```

```
for token, ids in zip(t.tokens, t.ids):
    print(str(token), ids, sep="\t = \t")
```

```
https    =    265
://      =    260
www      =    266
.        =    13
mi       =    725
i        =    72
.        =    13
lt       =    602
.        =    13
my       =    494
univers  =    2097
ity      =    456
.        =    13
com      =    263
```

```
array([ 0.49093616,  0.5112145 , -0.7627142 , -0.6096685 , -0.09496393,
       -0.00412803,  0.2347722 ,  0.41557637, ...,  0.37662762,
       -0.15289237,  0.24574889, -0.2222204 , -0.05246911,  0.14738731,
        0.24619438,  0.29445276], dtype=float32)
```

```
phish2vec.wv.similarity('https', 'www')
```

```
0.71509415
```

```
phish2vec.wv.similarity('https', 'com')
```

```
0.36151835
```

```
phish2vec.wv.most_similar('https')[:3]
```

```
[('http', 0.9666811227798462),
 ('www', 0.7150942087173462),
 ('://', 0.6383066177368164)]
```

Vilniaus universitetas

Antro pusmečio rezultatų apibendrinimas

- Veikiantis GAN modelis
- Veikiantis Autoencoder modelis
- Sukaupta sintetinių fišingo URL duomenų aibė
- Sukurtas naujas fišingo URL klasifikatorius, naudojantis tokį patį kalbos modelį, kaip GAN
- Atlikti eksperimentai
- Parašyta ir įteikta ISI publikacija

Vilniaus universitetas

Kito pusmečio darbų planas

- Paskutiniojo tyrimo rezultatų pristatymas tarptautinėje konferencijoje
- Disertacijos rašymas

Vilniaus universitetas

AČIŪ UŽ DĖMESĮ

Paulius Vaitkevičius

VU DMSTI doktorantas

+370 650 83623

paulius.vaitkevicius@mif.vu.lt

ORCID