Detekcija pakiranih datoteka

Anamarija Čavka, Andrea Stanić

July 4, 2018

PMF-Matematički odsjek, Zagreb

Sadržaj

Uvod

Ciljevi

- Zadatak: Napraviti model za detekciju pakiranih, odnosno nepakirnih datoteka
- Naš cilj: Isporbati kako dobro će neuralne mreže raditi binarnu klasifikaciju

O zadataku

- 43784 .json datoteka koje su TitaniumCore izvještaji statičke analize datoteke
- 4 grupe packera
- 21822 pakiranih i 21962 nepakiranih datoteka u skupu za treniranje

Pristup

Odabir značajki

- Domensko zananje koje su nam na radionici prenijeli predstavnici ReversingLabsa
- Slučajne šume

Sve značajke

Značajka	Opis
sha_value	sha vrijednost file-a
target	Poprima binarne vrijednosti (0/1) i označava je li datoteka pakirana (1) ili nije
fileSize	Veličina datoteke
fileEntropy	Entropija cijele datoteke
eCBLp	Bytovi na zadnjoj stranici datoteke
eCP	Stranice u datoteci
eCRIc	Relokacije
eCParHdr	Veličina headera u datoteci
eSS	Početna SS vrijednost
eSP	Početna SP vrijednost
eCsum	Checksum
eIP	Početna IP vrijednost
eCS	Početna CS vrijednost
eLFARIc	Adresa relocirane tablice
eOvNo	Overlay
eLFANew	Adresa novog exe headera
numberOfSections	Broj sekcija
sizeOfOptionalHeaders	Veličina opcionalnog headera
majorLinkerVersion	Broj glavne verzije linkera
minorLinkerVersion	
sizeOfCode	Veličina koda (tekst) sekcija, ili njihova suma ako ih je više
sizeOfInitializedData	Veličina inicijaliziranih sekcija, ili njihova suma ako ih je više
sizeOfUninitializedData	Veličina neinicijaliziranih sekcija, ili njihova suma ako ih je više
addressOfEntryPoint	Adresa ulazne točke
imageBase	Preferirana adresa prvog byta slike prilikom učitavanja u memoriju, pomnožena s 64K
sectionAlignment	Poravnanje sekcija kad su učitane u memoriju
fileAlignment	Poravnanje faktora kad su učitani u memoriju
sizeOfImage	Veličina slike, u byteovima
sizeOfHeaders	Veličina headera
sizeOfStackCommit	Veličina stoga za izvršiti
sizeOfStackReserve	Veličina stoga za rezervirati
sizeOfHeapCommit	Veličina lokalog prostora za izvršavanje heapa
sizeOfHeapReserve	Veličina lokalog prostora rezerviranog za heap
loaderFlags	Rezervirano, mora biti 0
numberOfRvaAndSizes	Broj data-directory u ostatku optional headera
checksum	
section_names	Imena sekcija
section_size	Veličina sekcija
max_entropy	Maksimalna entropija po sekcijama za jednu datoteku
avg_entropy	Prosječna entropija po sekcijama za jednu datoteku
no_of_section	Broj sekcija za svaku datoteku
max_section_size	Maksimalna veličina po sekcijama za jednu datoteku
mean_section_size	Prosječna veličina po sekcijama za jednu datoteku
shared	Broj sekcija u datoteci koje se mogu podjeliti u memoriji
execute	Broj sekcija u datoteci koje se mogu izvršiti kao kod
read	Broj sekcija u datoteci koje se mogu čitati
write	Broj sekcija u datoteci u koje se može pisati
code	
flag_list	
len_api_list	Duljina liste API-ja u sekciji
tag_list	

Slučajne šume

RandomForestClassifier metoda iz biblioteke scikitlearn

Table 1: Parametri slušajnih šuma

Parametar	**Opis**	**Vrijednost**
n_estimators	Broj generiranih stabala u šumi	1000
max_depth	Maksimalna dubina stabla	20
n_jobs	Broj paralelno odrađivanih poslova	-1
random_state	Seed	0
max_features	Broj ispitanih značajki prilikom grananja	20
criterion	Kriterij odabira značajke tijekom grananja	gini

Odabrane značajke

Važnost značajke	Značajka
0.30583471483768826	max_entropy
0.16093477391559693	fileEntropy
0.13205575739345443	write
0.0701628407097889	tag_entropy
0.04898428628843629	IMAGE_FILE_RELOCS_STRIPPED
0.04750125322652355	max_section_size
0.04439184830716724	mean_section_size
0.034098258514758936	addressOfEntryPoint
0.02434458184408063	IMAGE_FILE_LINE_NUMS_STRIPPED
0.018904811938273088	majorLinkerVersion
0.014582129863127461	fileSize
0.009434833621212026	avg_entropy
0.009087260085237366	code
0.008365083827465276	execute
0.00707273324823697	sizeOfImage
0.007018653105426279	sizeOfCode
0.006542293616709051	eLFANew
0.006024577888950814	minorLinkerVersion
0.004964649905155557	sizeOfUninitializedData
0.004294499694192568	checksum
0.003656062499284743	no_of_section
0.003631272304483664	numberOfSections
0.003543939937061628	sizeOfInitializedData
0.003295011305666241	overlay
0.0030536710317214007	imageBase
0.002587902645800645	codeview
0.0019519159518068055	len_api_list
0.0017503790880335306	eCBLp
0.0015420989342500033	sizeOfStackReserve
0.0013716362555896402	eCP

Matrica konfuzije, preciznost, specifičnost

Table 2: Matrica konfuzije

		Predviđeno	
		0	1
Stvarno	0	5481	16
	1	42	5365

- $Preciznost = (TP + TN)/(TP + FP + TN + FN) = 0.9947$
- $Osjetljivost = TN/(FP + TN) = 0.9971$
- $Specifičnost = TP/(TP + FP) = 0.9970$

Neuralne mreže

- KerasClassifier iz biblioteke Keras
- Na ulazu 30 neurona, unutarnji sloj 15 neurona
- Aktivacijske funkcije: relu, sigmoid
- optimizers = 'adam', kernel initializer = 'normal'
- problem: rad na CPU

Rezultati

Grid search

GridSearchCV iz sklearn.model_selection

Table 3: Rezultati

Točnost (standardna devijacija)	Batch	Epochs
0.571761 (0.166324)	5	30
0.330532 (0.063892)	5	50
0.599306 (0.151421)	5	100
0.330555 (0.063915)	10	30
0.599329 (0.151445)	**10**	**100**
0.502558 (0.180324)	10	100
0.501576 (0.181073)	15	30
0.330532 (0.063892)	15	50
0.498401 (0.181105)	15	100

Zaključak

Zaključak

- Slučajne šume su se pokazale kao dobar način za izbor značajki
- Neuralne mreže sa samo jednim skrivenim slojem nisu dovoljno dobre
- Budući rad:
  - probati osposobiti rad na GPU i pokušati napraviti dublje mreže
  - koristiti PyTorch
  - probati neke druge modele: XGBoost, SMV, druge mreže