|  |  |
| --- | --- |
| **Course Title** | **머신러닝최신기술특론** |
| **Instructor** | **우종우 교수님** |
| **report ID** | **텀프로젝트(제안서 수정)** |
| **Due date** | **2018-05-18** |

|  |  |
| --- | --- |
| **Department** | 컴퓨터공학 |
| **Student id.** | M2018170(김동훈)  M2018175(이식) |
| **Student name** | 김동훈, 이식 |
| **Submission date** | 2018-05-08 |

**주제**

어떤 파일이 Packing 되어 있는지 되어있지 않은지 딥러닝 학습을 통해 판별하기

**이 주제를 선택한 이유**

어떤 파일이 악성파일인지 정상파일인지 직접 프로그램을 실행 해 보지 않아도 정적 분석을 통해 어느 정도는 판별 할 수 있다. 하지만 파일이 패킹되어 있다면 이러한 정적 분석은 의미가 없게 되며 제대로 된 결과를 얻기 힘들다. 그래서 파일이 패킹되어있는지 알아내는 것은 1차적으로 굉장히 중요한 문제이다. UPX같이 시그너쳐가 파일 내에 있는 것은 패킹 여부를 알아내기 쉽지만 다른 패커들로 패킹했을 경우 그 파일이 패킹되어 있는지 알아내기란 쉽지 않다. 이러한 흥미로운 이유들 때문에 이 주제를 선택하였다.

**세부내용**

패킹이 되어 있지 않은 파일 약 4만개, UPX, ASPack, UPack, PESpin, NSAnti, Themida 등으로 패킹되어 있는 파일 약 4만개 총 8만개의 파일로 딥러닝 학습을 시킨다.

그러기 위해서는 패킹이 되어 있지 않은 파일이 필요한데, 그 파일은 virussign의 8월 Malware 데이터에서 수집하여 (100만개 이상) UPX로 패킹이 되어있는 파일을 언패킹 시켜 확보한다. 이 때 2중패킹은 고려하지 않는다. UPX로 언패킹 되었다는 것은 패킹되지 않은 파일이라고 생각한다. 그렇게 약 12만개의 패킹 되어 있지 않은 파일을 확보 한 후 그 중 6만개는 위의 여러가지 패커들로 패킹을 시킬 것이다. 그 후 4만개의 패킹되지 않은 파일을 학습시키고, 4만개의 패킹된 파일을 Training Data Set으로 사용한다. 나머지 4만개(패킹 되어 있지 않은 파일 2만개 + 패킹 되어 있는 파일 2만개)는 Test Data Set으로 사용할 예정이다.

* 8만개의 Training Data Set 이중 4만개는 패킹 되어 있는 파일
* 4만개의 Test Data Set 이중 2만개는 패킹 되어 있는 파일

**프로젝트를 진행하기 위해 고려해야 할 점**

만약 패킹된 파일에 대해서 또다시 UPX로 패킹을 했을 경우 이중패킹이 되었기 때문에 패킹되지 않은 파일이라고 학습을 잘못 시킬 염려가 있다. 하지만 이중패킹을 고려한다면 파일 수집이 굉장히 어려워 지기 때문에 이중패킹은 존재 하지 않는다고 가정하였다.

**UPX 언패킹을 위해 작성한 Python Code**

import os, subprocess, time

PACKER\_PATH ='C:\\Users\\jack\Desktop\\upx394w\\upx.exe'

MAL\_PATH = 'D:\\INFOSE~1\\malware\\20171028'

OUT\_PATH = 'D:\\output'

malfile\_list=os.listdir(MAL\_PATH)

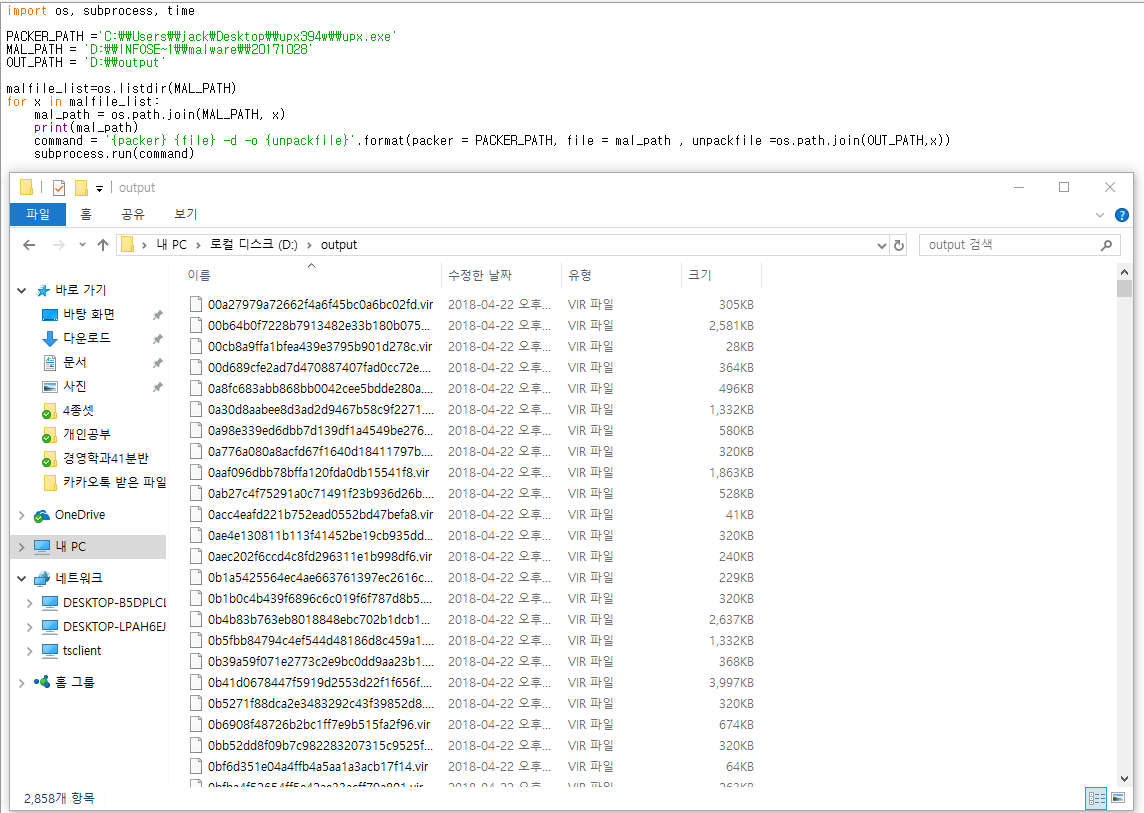
for x in malfile\_list:

mal\_path = os.path.join(MAL\_PATH, x)

print(mal\_path)

command = '{packer} {file} -d -o {unpackfile}'.format(packer = PACKER\_PATH, file = mal\_path , unpackfile =os.path.join(OUT\_PATH,x))

subprocess.run(command)



**패킹되어 있지 않은 순수파일을 추출하는 중**

**프로젝트 계획**

|  |  |  |  |
| --- | --- | --- | --- |
| **~ 05/24** | **~ 5/31** | **~ 6/8** | **~ 6/15** |
| **데이터 수집 및 가공** | | |  |
|  | **모델설계** | |
|  | | | **하이퍼파라메터 튜닝** |

**데이터 수집 및 가공**

**언패킹된 파일 수집(8만개), 패킹파일 수집(4만개), 각각의 패커에 대해 패킹 자동화 시키기, 피쳐백터 생성**

**모델설계**

**딥러닝 네트워크 설계**

**하이퍼파라메터 튜닝**

**검증셋을 통한 파라메터값 튜닝**

**역할분담**

**김동훈 : Data 가공 (언패킹된 파일 추출, 각종 패커로 패킹), Feature 추출, 실험**

**이식 : Model 설계, Feature 추출, 실험**