


캡스톤 디자인 I 종합설계 프로젝트

프로젝트 명	<i>MASK(Malware Analysis System in Kookmin)</i>
팀 명	<i>NCNP(No Commit No Pay)</i>
문서 제목	결과보고서

Version	1.5
Date	2018-MAY-29

팀원	한 채연 (조장)
	김 영재
	명 준우
	이 유정
	허 준녕

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29


CONFIDENTIALITY/SECURITY WARNING

이 문서에 포함되어 있는 정보는 국민대학교 전자정보통신대학 컴퓨터공학부 및 컴퓨터공학부 개설 교과목 캡스톤 디자인 I 수강 학생 중 프로젝트 "MASK"를 수행하는 팀 "NCNP"의 팀원들의 자산입니다. 국민대학교 컴퓨터공학부 및 팀 "NCNP"의 팀원들의 서면 허락없이 사용되거나, 재가공 될 수 없습니다.

문서 정보 / 수정 내역


Filename	결과보고서-NCNP.doc
원안작성자	한채연
수정작업자	한채연, 김영재, 명준우, 이유정, 허준녕

수정날짜	대표수정자	Revision	추가/수정 항목	내 용
2018-05-25	한채연	1.0	최초 작성	초안 작성
2018-05-26	명준우	1.1	내용 추가	연구/개발 내용 및 시스템 기능 요구사항 추가
2018-05-26	이유정	1.2	내용 추가	연구/개발 내용 추가
2018-05-27	허준녕	1.3	내용 추가	연구/개발 내용 및 시스템 기능 요구사항 추가
2018-05-28	김영재	1.4	내용 추가	연구/개발 내용 및 시스템 기능 요구사항 추가
2018-05-29	한채연	1.5	내용 수정	오탈자 수정 및 최종 검토

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

목 차

1	개요	4
1.1	프로젝트 개요	4
1.2	추진 배경 및 필요성	4
1.2.1	최근 악성코드 현황	5
1.2.2	바이러스 분석 기술의 시장 현황	8
1.2.3	현재 바이러스 분석 시스템의 한계점과 개선 방향	10
2	개발 내용 및 결과물	11
2.1	목표	11
2.2	연구/개발 내용 및 결과물	12
2.2.1	연구/개발 내용	12
2.2.2	시스템 기능 요구사항	31
2.2.3	시스템 비기능(품질) 요구사항	33
2.2.4	시스템 구조 및 설계도	34
2.2.5	활용/개발된 기술	36
2.2.6	현실적 제한 요소 및 그 해결 방안	37
2.2.7	결과물 목록	37
2.3	기대효과 및 활용방안	38
2.3.1	기대효과	38
2.3.2	활용방안	38
3	자기평가	39
4	참고 문헌	40
4.1	참고 사이트	40
5	부록	41
5.1	사용자 매뉴얼	41
5.2	설치 매뉴얼	42
5.3	테스트 케이스	43

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

1 개요

1.1 프로젝트 개요

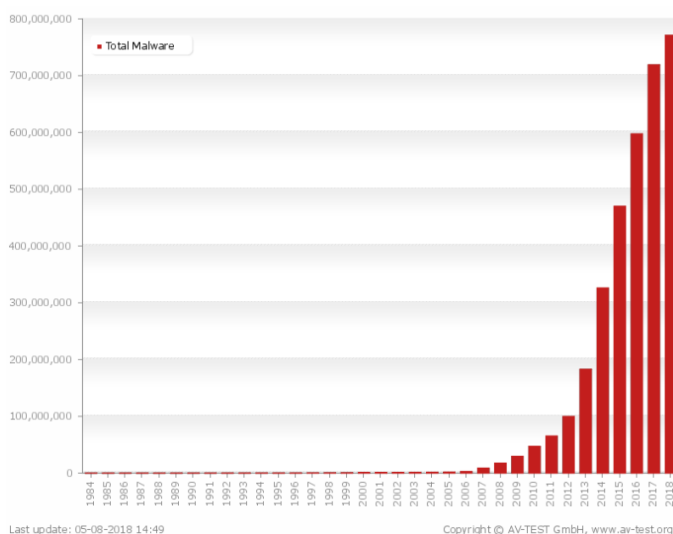
본 프로젝트는 악성코드로 의심되는 파일을 분석하여 그 결과들을 악성코드 분석가에게 제공함으로써 보다 정교하게 악성코드를 분석할 수 있도록 한다. 본 프로젝트에서 제공하는 기능은 아래와 같다.

첫째, 파일에 대한 다양한 분석이다. 사용자가 악성으로 의심되는 파일을 웹에 업로드하면, 그 파일에 대한 정적 분석 결과와 동적 분석 결과를 모두 제공해 준다.


둘째, 딥 러닝 모델에 의한 결과이다. 사용자가 악성으로 의심되는 파일을 웹에 업로드하면, 그 파일에 대한 다양한 결과값들을 제공한다.

1.2 추진 배경 및 필요성

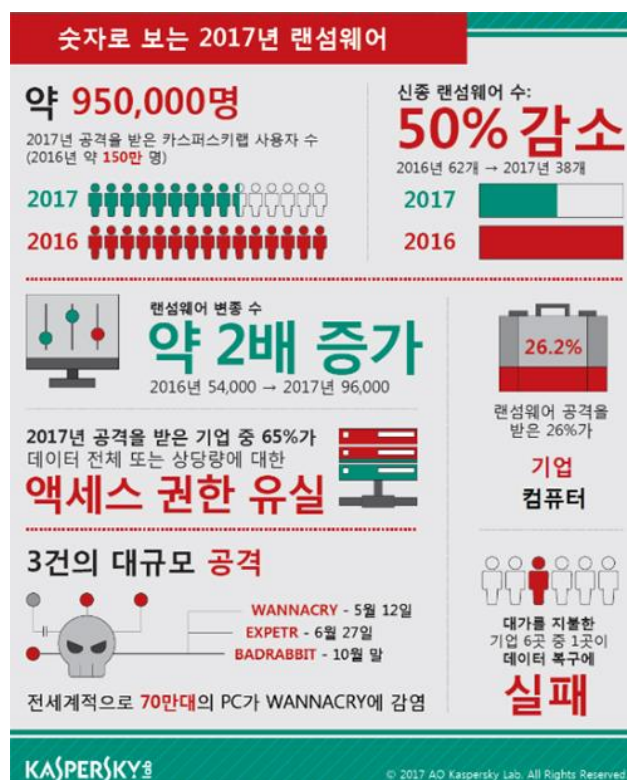
악성코드(Malware)는 언제 어디서나 우리를 위협하는 존재이다. 안티바이러스 테스트 업체 AV-TEST에 따르면 해마다 발견되는 악성코드는 늘어나는 추세이며, 현재 하루에 대략 100만개 정도의 악성코드가 발견되고 있다. 이에 비해 악성코드 전문가는 현격히 부족하다. 게다가 안티 바이러스 제품이나 기존 탐지 솔루션은 복잡해지고 정교해지는 악성코드에 대해 효과적으로 대응하지 못한다.



[그림 1] 연도 별 발견되는 총 악성코드 개수 [출처 : AV-TEST]

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

악성코드의 개수가 나날이 증가함에 따라 경제적, 사회적 문제를 발생시키고 있다. 현저히 늘어난 악성코드 개수에 비해 악성코드 분석 전문가의 수는 부족하다. 또한 매일 수십만 개의 신종/변종 악성코드들이 생성되고 있다. 수많은 악성코드를 비롯한 각종 사이버 침해 사고로 인한 경제적 피해규모가 상당하며, 국가 사회적인 혼란을 유발하여 국민생명과 국가안보에 심각한 위협이 된다.




[그림 2] 숫자로 보는 2017년 랜섬웨어 [출처 : 카스퍼스키랩]

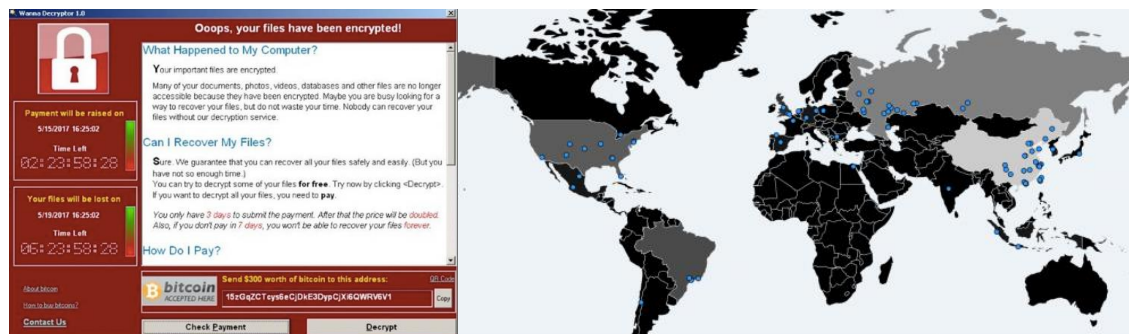
기존의 악성코드 분석 기술과 전문 분석가에 의한 대응으로는 신종/변종 악성코드의 생성 속도를 따라갈 수가 없다. 따라서 대량의 악성코드를 효율적으로 식별하기 위한 자동화된 기술이 반드시 필요하다.

1.2.1 최근 악성코드 현황

1) 워너크라이(WannaCry) 랜섬웨어

2017년 5월에 워너크라이 랜섬웨어가 불특정 다수의 시스템을 감염시킨 사태가 벌어졌다. 워너크라이는 국내에서도 확산되는 추세이다. 이스트시큐리티는 통합 백신 '알약'에서 12일 942건, 13일 1167건 이상 탐지했으며, 현재까지도 지속적으로 탐지되고 있다고 밝혔다[1].

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29




[그림 3] 워너크라이 랜섬웨어와 워너크라이 발생 지역 [출처 : 시만텍, 멀웨어테크]

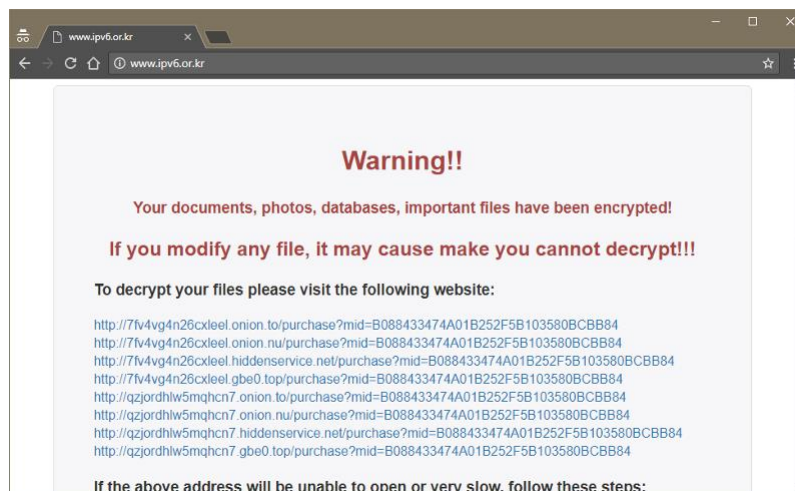


[그림 4] 통합 백신 알약을 통해 탐지된 Wanna Cryptor 랜섬웨어 통계 [출처 : ESTsecurity]

2) 에레보스(Erebus) 랜섬웨어

에레보스(Erebus) 랜섬웨어가 특정 기업, 기관을 타겟으로 집중적인 공격을 시도하기도 하였다. 2017년 6월 10일 웹호스팅 업체인 ㈜인터넷나야나의 리눅스 서버 153대가 에레보스 랜섬웨어에 감염되었다. 해당 업체는 한화 약 18억원을 복호화 키의 비용으로 요구 받았다고 하였다[2].

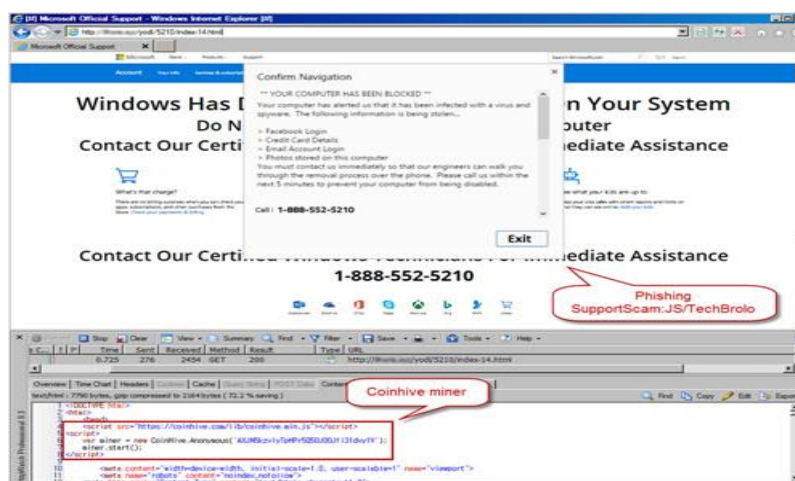
 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29



[그림 5] 에레보스 랜섬웨어 감염 화면 [출처 : 안랩]

3) 마이너(miner)

악성코드에 심어진 가상화폐 채굴 프로그램인 “마이너(miner)”도 기승을 부리고 있다. 이는 피싱 메일 및 악성 웹 페이지를 통해 유포되고 있으며, 시스템 취약점 악용, 멀버타이징 (Malvertising) 등 유포 경로가 다양화 되고 있다[3].

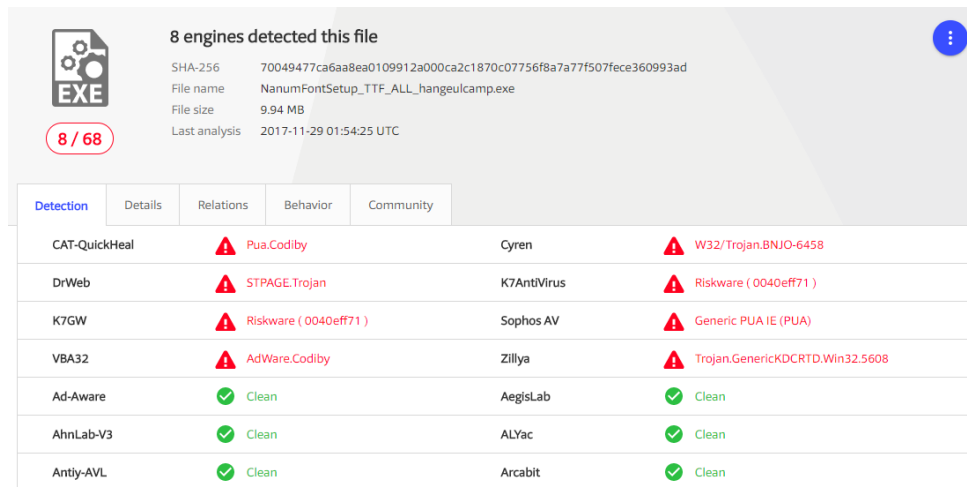


[그림 6] 피싱 사이트에서도 이용되는 비트코인 마이너 [출처 : 제로서트]

1.2.2 바이러스 분석 기술의 시장 현황

1) VirusTotal

‘VirusTotal’은 구글(Google)의 자회사로, 파일이나 URL을 안티바이러스 엔진과 웹사이트 스캐너를 이용하여 악성여부를 식별해주는 무료 온라인 서비스이다. V3, 알약, 비트디펜더 등 약 60여가지의 바이러스 검사 소프트웨어 제품을 사용한다.



8 engines detected this file

SHA-256: 70049477ca6aa8ea0109912a000ca2c1870c07756f8a7a77f507fece360993ad
File name: NanumFontSetup_TTF_ALL_hangeulcamp.exe
File size: 9.94 MB
Last analysis: 2017-11-29 01:54:25 UTC

8 / 68

Detection	Details	Relations	Behavior	Community
CAT-QuickHeal	⚠ Pua.Codiby		Cyren	⚠ W32/Trojan.BNJO-6458
DrWeb	⚠ STPAGE.Trojan		K7AntiVirus	⚠ Riskware (0040eff71)
K7GW	⚠ Riskware (0040eff71)		Sophos AV	⚠ Generic PUA IE (PUA)
VBA32	⚠ AdWare.Codiby		Zillya	⚠ Trojan.GenericKDCRTD.Win32.5608
Ad-Aware	✅ Clean		AegisLab	✅ Clean
AhnLab-V3	✅ Clean		ALYac	✅ Clean
Antiy-AVL	✅ Clean		Arcabit	✅ Clean

[그림 7] VirusTotal 분석 결과 화면 중 일부

2) malwares.com

‘malwares.com’은 세인트시큐리티(Saint Security)에서 개발한 한국 최초의 빅데이터 기반 악성코드 자동 분석 플랫폼으로, 다양한 수집 채널로부터 유입된 악성코드를 자동으로 분석하고, 분석된 결과를 공유하여 악성코드 유포 정황을 빠르게 인지하고 크게 확산되는 것을 방지하여 각종 악성코드로부터의 공격에 능동적으로 대처하는 것을 목적으로 만들어진 인텔리전스 서비스다.



malwares.com™


70049477CA6AA8EA0109912A000CA2C1870C07756F8A7A77F507FECE360993AD

MD5 : AD7587A2CAE0ED77E874447F2F28726
SHA-1 : 84A9BC095D31A2995CC62FECA7CD8933D2EDA971
SHA-256 : 70049477CA6AA8EA0109912A000CA2C1870C07756F8A7A77F507FECE360993AD
파일 크기 : 10,425,776 bytes
파일 유형 : exe_32bit
알려진 날짜 : 2014-03-29 14:59:05 (3년 11개월 전)

AI: 66/100
safe malware

태그: #peexe #overlay #user-directory #signed #msis #packing #exe_32bit

[그림 8] malwares.com 분석 결과 화면 중 일부

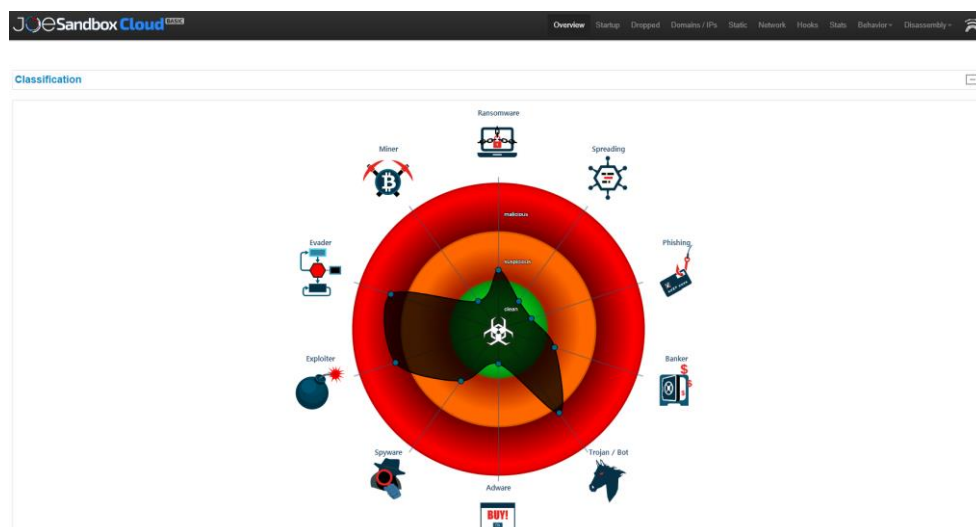
 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

3) Hybrid Analysis

'Hybrid Analysis'는 독일의 페이로드시큐리티(Payload Security)에서 개발한 악성 소프트웨어 분석 서비스로, 하이브리드 분석 기술을 사용하여 알려지지 않은 위협을 탐지 및 분석한다. Hybrid Analysis는 분석 정보 생성 단계에서 심층적인 정적 분석을 수행할 수 있도록 심볼 정보와 모니터링 된 런타임 프로세스의 메모리 덤프 스냅샷을 미세한 단위로 저장하고, 정적 분석과 동적 데이터를 결합하여 실행에 관계없이 악성 행위를 탐지한다.

4) Joe Sandbox


'Joe Sandbox'는 스위스의 'Joe Security'에서 개발한 악성코드 정밀 분석 샌드박스, 샌드박스 내에서 얻은 분석정보를 시각화하여 분석 리포트에서 보여준다. 윈도우 운영체제 기반 데스크톱 PC부터 맥OS, 안드로이드와 iOS 등 다양한 모바일 OS까지 다양한 지원 제품군으로 구성되어 있는 것이 가장 큰 특징이다.



[그림 9] Joe Sandbox 분석 결과 화면 중 일부

5) Clam AntiVirus

'Clam AntiVirus'는 CISCO Systems에서 지원하는 오픈소스 소프트웨어로 자유 크로스플랫폼 형식의 바이러스 검사 소프트웨어 툴 킷이며, 바이러스를 비롯한 수많은 종류의 악성 소프트웨어를 탐지해낸다. 주로 메일 게이트웨이에서의 이메일 스캐닝을 위하여 설계되었다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

6) Kicom AntiVirus

‘Kicom AntiVirus’는 오픈소스 소프트웨어로 악성코드를 탐지하고 치료하기 위하여 설계된 안티 바이러스 엔진이다. 1995년 C/C++로 작성되었다가 1998년 HAURI의 ViRobot 엔진과 통합된 후에 파이썬 언어로 다시 개발되었다.

1.2.3 현재 바이러스 분석 시스템의 한계점과 개선 방향

1) 문제점

위에 언급한 대부분의 제품들은 Public API를 제공하지만 한정적인 서비스만을 이용할 수 있으며, 더 나은 서비스를 제공받기 위해서는 유료로 이용해야 한다. 또한 오픈소스 소프트웨어가 아니므로 추가적인 개발을 할 수 없다. ClamAV와 KicomAV는 신종 및 변종 악성코드에 대한 자동 업데이트가 이루어지지 않으므로 신종/변종 악성코드 탐지에 취약하다.

2) 개선 방향

본 프로젝트는 악성코드에 대한 정적 및 동적 분석을 통해 나온 다양한 정보를 사용한다. 이 정보는 딥 러닝 모델에 적용하여 신종 및 변종 악성코드에 실시간으로 대응을 유리하게 한다. 그리고 악성코드 분석가에게 자세하고 보다 더 다양한 분석 결과를 제공한다. 또한 오픈소스 소프트웨어로 공개하여 사용자가 원하는 기능을 추가할 수 있도록 한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

2 개발 내용 및 결과물

2.1 목표

본 프로젝트는 악성코드로 의심되는 파일을 분석하여 정적, 동적 분석 결과와 이와 유사한 파일에 대한 분석 결과, 자체적으로 제작한 모델에 의해 결정된 악성코드의 라벨을 분석가에게 제공함으로써 보다 빠르게 악성코드를 분석할 수 있도록 하는 것을 목표로 한다. 나아가 이를 오픈 소스 소프트웨어로 공개하는 것을 목표로 한다.

<세부 목표>

- 사용자가 업로드한 파일에 대해 정적 및 동적 분석을 진행하여 그 결과를 제공한다.
- 업로드한 파일에 대한 정적 분석, 동적 분석 결과에 대한 리포트를 자동으로 생성한다.
- 자동으로 생성된 리포트로부터 피처를 자동으로 추출한다.
- 카스퍼스키(Kaspersky) 탐지 결과를 활용하여 자체적인 악성코드 분류 기준을 세워 라벨을 제공한다.
- 주기적인 학습을 통해 신종 및 변종 악성코드에 대해 대응을 할 수 있도록 한다.
- 분석한 파일에 대한 세부정보를 사용자가 한 눈에 볼 수 있도록 웹에 시각화 한다.
- 사용자가 업로드한 파일과 유사한 파일은 데이터베이스에서 찾아 그 파일의 분석 결과를 제공한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

2.2 연구/개발 내용 및 결과물

2.2.1 연구/개발 내용

사용자가 악성으로 의심되는 파일을 업로드하면, 그 파일의 해시 값(md5)을 구한 후 데이터베이스에 연동된 검색엔진을 이용하여 분석 결과를 찾는다. 분석 결과가 데이터베이스에 있을 경우, 그 정보들을 웹을 통하여 사용자에게 보여준다. 분석 결과가 데이터베이스에 없을 경우, 그 파일에 대해 각각 정적 및 동적 분석을 진행하고, 생성된 리포트로부터 추출한 피처를 이용하여 학습된 모델로부터 탐지 결과를 사용자에게 웹으로 보여준다. 각 세부 연구 분야는 아래 그림과 같다.




[그림 10] 연구 분야 카테고리

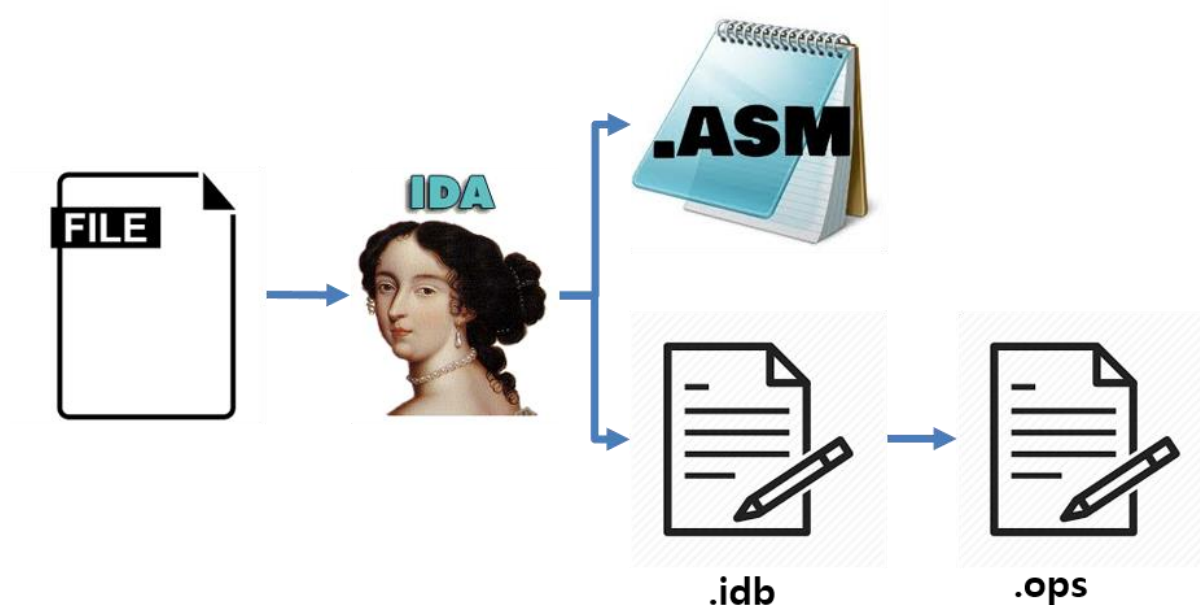
1) 정적 분석 및 이를 이용한 피처 추출

정적 분석이란 파일을 실행하지 않고 분석하는 것이다. 정적 분석은 주로 PE파일(Portable Executable)의 정보, 프로그램의 소스 코드 등을 이용한다. 이를 위하여 본 프로젝트에서는 파이썬 기반의 PE파일 분석 오픈소스 소프트웨어인 pefile 과 디스 어셈블러인 IDA Pro를 사용하였다.


pefile을 이용하면 해당 파일에서 사용하는 string 정보, section 정보, API 등을 쉽게 추출 할 수 있다. 본 프로젝트에서는 pefile과 peframe을 참고하여 Python3에서 사용 가능한 pefile 분석 도구(peview)를 제작하여 파일의 해시, 섹션 정보, API, Import Functions 등을 웹을 통해 사용자에게 제공한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

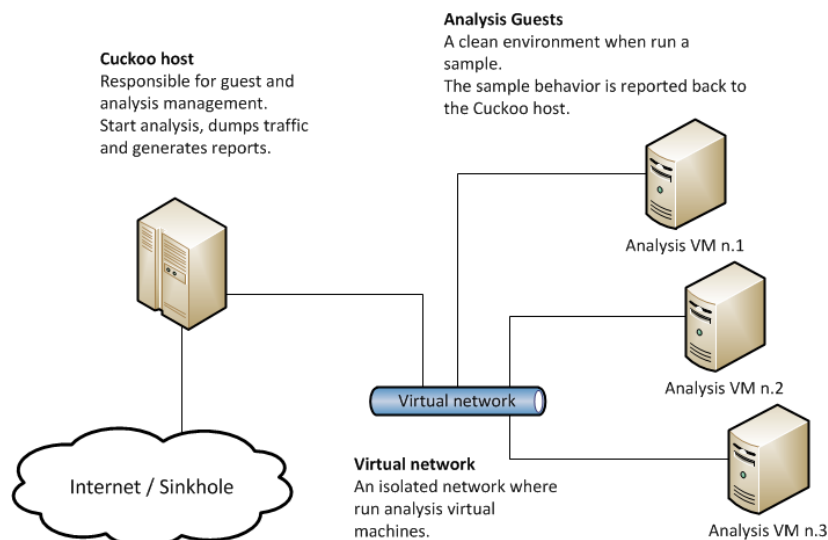
딥 러닝을 위해서는 IDA Pro를 이용해 추출한 opcode 시퀀스를 사용하였다. I. Santos 외 6명의 연구에서는 opcode 시퀀스를 이용하여 변종 악성코드를 탐지할 수 있음을 보였고, I. Santos 외 2명의 연구에서는 머신러닝을 이용한 악성코드 탐지에서 opcode 시퀀스를 악성코드의 특징으로 학습시켰을 때 높은 정확도로 악성코드 탐지가 가능함을 보였다. A. Shabtai 외 4명의 연구에서는 opcode 시퀀스 n-gram 기법을 사용하여 알려지지 않은 악성코드를 높은 정확도로 탐지해낼 수 있음을 보였으며 Xin Hu 외 3명의 연구에서는 정적 분석 정보에서 특징을 추출할 때 opcode 시퀀스를 사용하는 것이 Control-flow graph, binary sequence, mnemonic sequence를 사용하는 것보다 효과적이라고 주장하였다. 따라서 본 프로젝트에서는 IDA를 사용하여 얻어낸 정적 분석 정보 중 opcode 시퀀스를 다중 n-gram 기법으로 가공하여 악성코드의 특징으로 사용하였다.



[그림 11] ida pro를 이용한 정적 분석 계획

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

2) 동적 분석 및 이를 이용한 피처 추출




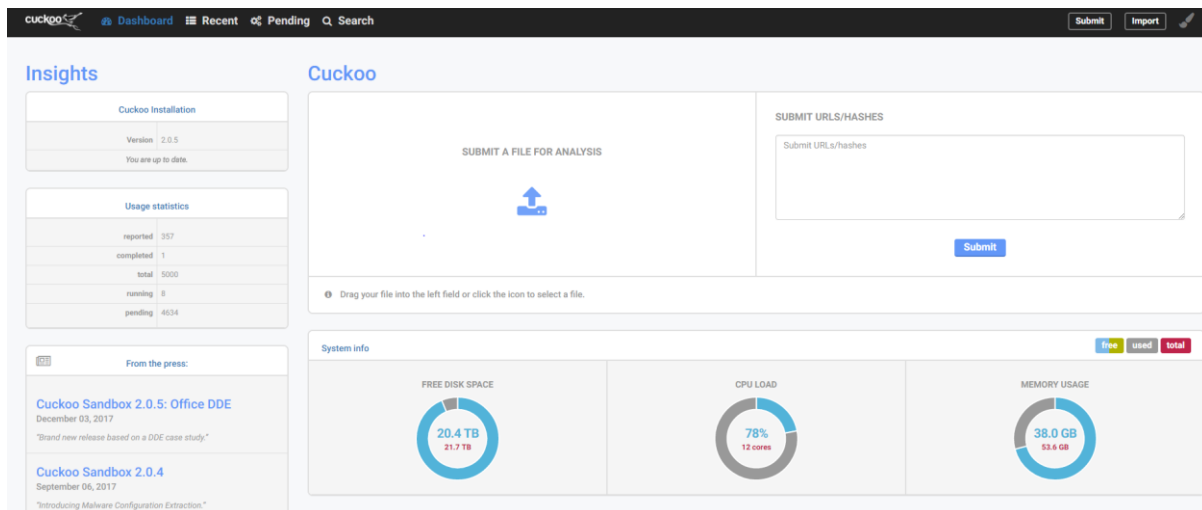
[그림 12] 쿠쿠샌드박스의 메인 아키텍처 [출처 : 쿠쿠샌드박스 문서]

동적 분석이란 악성코드를 분석 환경에서 실행시킨 후 시스템의 행동 변화를 분석하는 방법이다. 악성코드를 분리된 곳에서 안전하게 실행하기 위해선 독립적인 공간이 필요하기 때문에 가상 환경을 이용하였다.

본 프로젝트에서는 동적 분석을 위해 오픈소스 소프트웨어인 쿠쿠샌드박스(Cuckoo Sandbox)를 이용하였다. 샌드박스 환경은 가상머신 종류 중 하나인 버추얼박스(Virtualbox)로 구성하였다. 이를 이용하여 분리된 가상머신 환경에서 해당 파일을 실행한 후 어떤 행위가 일어나는지 관찰하여 구체적인 정보를 획득할 수 있다. 쿠쿠샌드박스는 비정상적인 접근을 탐지하기 위해 의도적으로 설치해 둔 시스템을 의미하는 허니팟(honeypot)에 기초한다. 따라서 악성코드가 잘 동작해야 효과적인 분석을 할 수 있기 때문에 고의로 취약한 환경을 구성하였다. 방화벽과 윈도우 업데이트를 비활성화 시키고, UAC(User Account Control)를 비활성화 시켰다. 또한 리눅스의 root 계정으로 시스템을 운영하는 것과 같이 Administrator 계정을 활성화하였다. 쿠쿠 코어가 샌드박스를 제어하기 위해서는 샌드박스의 아이피를 알아야 하는데, 아이피가 유동적으로 변경되면 코어가 제어를 할 수 없기 때문에 쿠쿠 엔진과 동일한 아이피 대역으로 설정하였다.

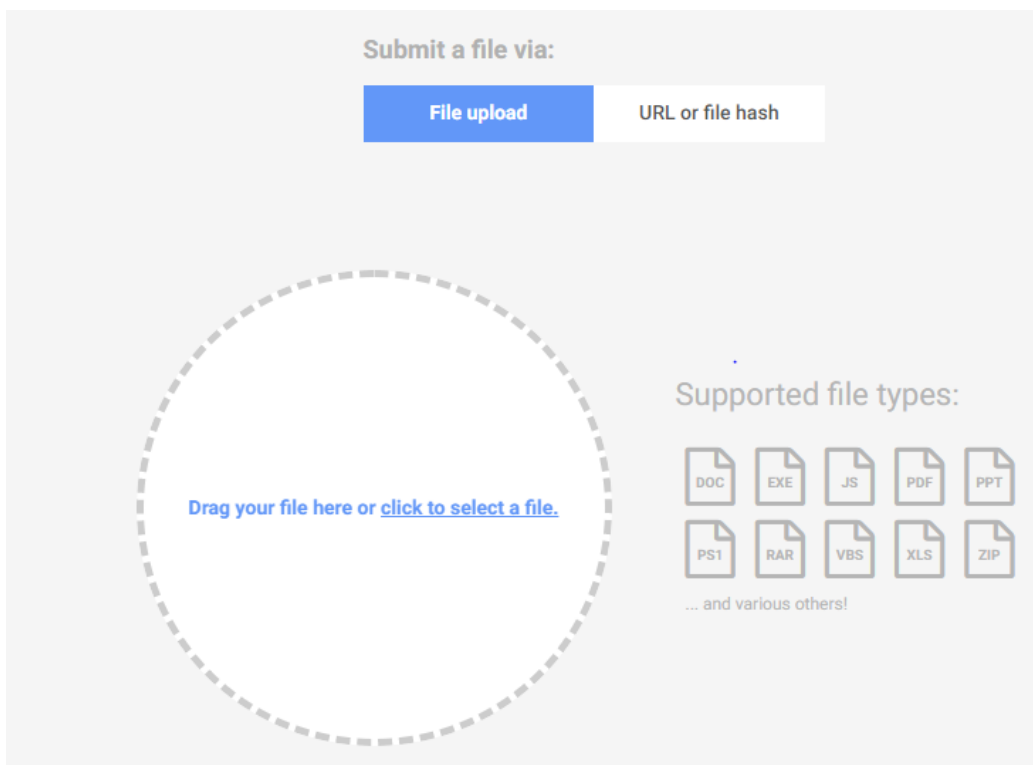
쿠쿠샌드박스는 웹 인터페이스(장고)를 제공한다. 이 웹 인터페이스가 사용하는 데이터베이스인 몽고디비(MongoDB)를 구축하여 사용자가 편리하게 브라우저를 이용하여 분석을 요청하거나 분석된 결과를 볼 수 있도록 하였다. 쿠쿠 코어나 샌드박스 설정을 위해서 설정 파일을 수정하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29




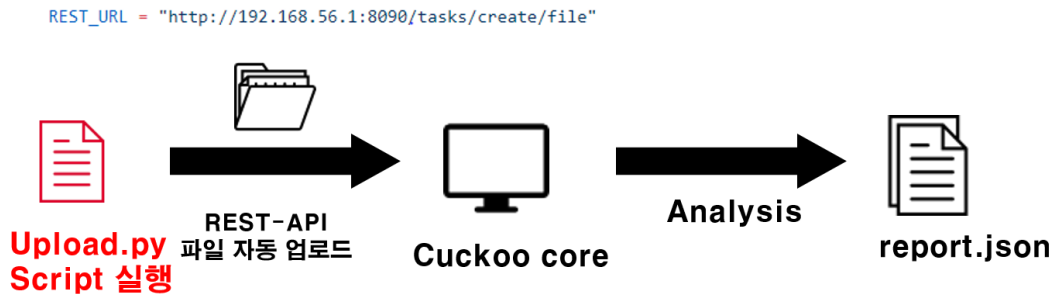
[그림 13] 쿠쿠샌드박스의 웹 인터페이스

악성으로 의심되는 파일을 분석하기 위해서는 웹 인터페이스를 이용하여 드래그 앤 드랍으로 파일을 전달하는 방법이 있지만, 자동화를 위해 쿠쿠샌드박스에서 제공하는 REST API를 이용하여 분석을 진행할 대량의 파일 업로드를 자동화하였다.



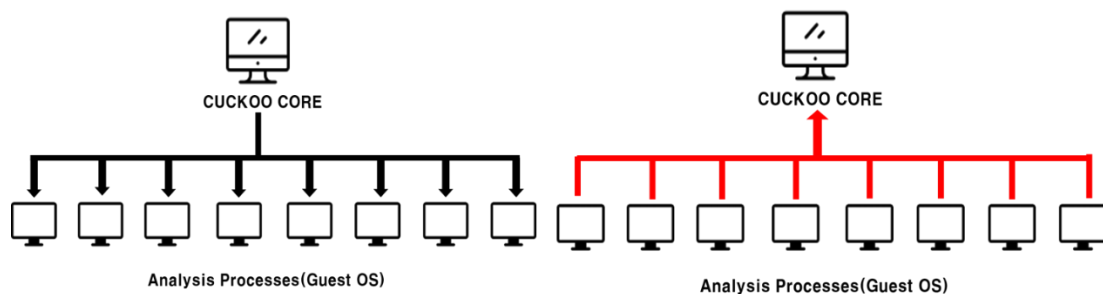
[그림 14] 웹 인터페이스를 이용하여 파일을 업로드하는 방법

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29




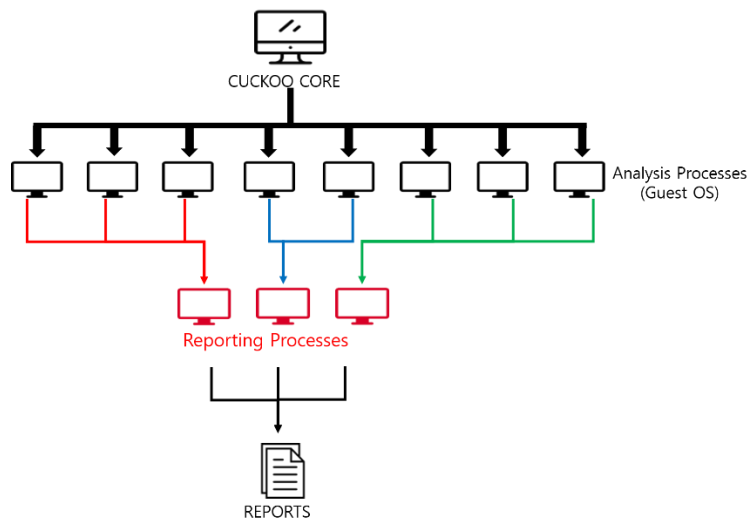
[그림 15] REST API를 이용한 대량의 파일 업로드 자동화

분석의 효율성을 높이기 위해 분석을 진행하는 guest instance의 개수를 1개에서 현재 8개로 늘려서 n개의 분석 파일에 대해 분석 시간을 $\frac{1}{8}$ 만큼 단축함으로써 분석 효율을 높일 수 있다. 또한 기존 쿠쿠 코어가 리포트를 생성하는 기존 방식은 분석 파일이 많을 경우 모든 파일에 대해 쿠쿠 코어가 전부 리포트를 생성해야 하기 때문에 과부하가 걸릴 수 있다. 따라서 리포트를 생성하는 리포팅 프로세스를 별도로 구성하여 쿠쿠 코어의 과부하를 막는다.



[그림 16] 쿠쿠샌드박스의 분석 및 리포트 생성 구조도(기존 방식)

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

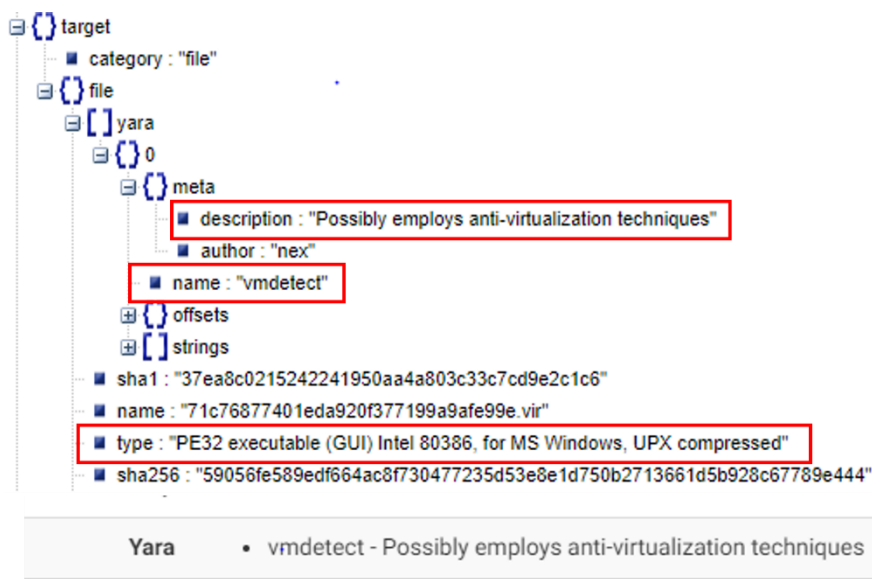


[그림 17] 쿠쿠샌드박스의 분석 및 리포트 생성 구조도


분석 후 json 포맷의 리포트가 생성되며 동적 분석을 통해 얻을 수 있는 피처로는 process memory, network, API 콜 시퀀스, signatures 등이 있다.

동적 분석 결과로 json 포맷의 리포트가 생성된다. 리포트에서 확인할 수 있는 결과로는 아래와 같다. 이들은 딥 러닝 모델을 생성할 때 추출될 피처의 후보가 된다.

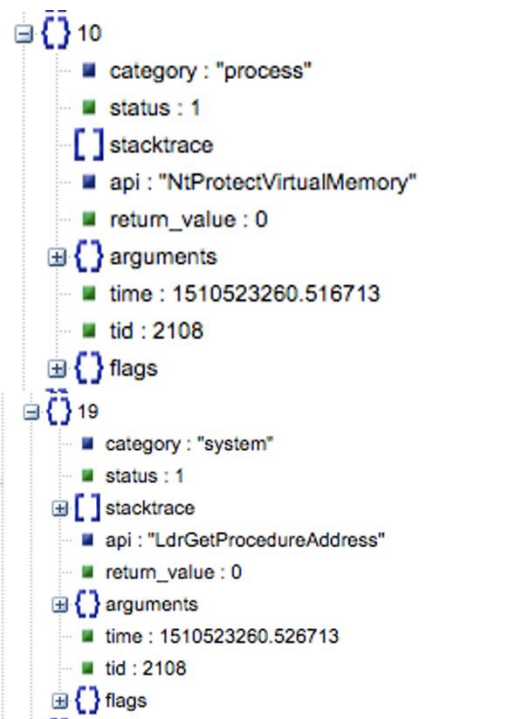
- ① Process memory : 프로세스에 대한 메모리 덤프 분석 정보이다.
- ② Target : Yara rule에 의해 탐지되었을 경우 나타나는 정보이다.



[그림 18] Yara rule에 의해 탐지되었을 때 리포트에서의 Target 정보


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

- ③ 네트워크 프로토콜, 악성코드를 실행한 host 정보이다.
- ④ 정적 분석 결과(Strings..)
- ⑤ Behavior(API 통계, API 콜 시퀀스..)












[그림 19] 한 프로세스의 10번째, 19번째 API 호출 기록(리포트)

- ⑥ Signatures : 위의 악성코드 정보들을 바탕으로 나타난 악성코드 특징(description)이다.




 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

Signatures

 Queries for the computername (1 event)
 Checks amount of memory in system, this can be used to detect virtual machines that have a low amount of memory available (1 event)
 A process attempted to delay the analysis task. (1 event)
 Drops a binary and executes it (1 event)
 Checks adapter addresses which can be used to detect virtual network interfaces (1 event)
 Potentially malicious URLs were found in the process memory dump (50 out of 124 events)
 Attempts to identify installed AV products by installation directory (3 events)
 Deletes its original binary from disk (1 event)
 A process performed obfuscation on information about the computer or sent it to a remote location indicative of CnC Traffic/Preparations. (4 events)


[그림 20] 웹 인터페이스에서 확인 가능한 파일에 대한 signatures

- ⑦ Score : signatures로 식별한 패턴을 통해 의심스러운 평균 수준을 수치화한 정도이다.

 Score This file shows some signs of potential malicious behavior. The score of this file is 1.2 out of 10.
 Score This file shows numerous signs of malicious behavior. The score of this file is 4.2 out of 10.
 Score This file is very suspicious , with a score of 5.4 out of 10!

[그림 21] 웹 인터페이스에서 확인 가능한 파일에 대한 score

쿠쿠샌드박스는 아래와 같이 323개 함수의 API 호출을 기록하고, 2017년 기준으로 자체적으로 17개의 카테고리로 분류한다. 현재 실험을 통해 확인된 카테고리는 총 18개이다.


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

class	description	example	# of APIs
A	file/directory	CopyFile, CreateDirectory, GetFileType, ...	47
B	registry	RegCreateKeyEx, NtCreateKey, RegDeleteValue, ...	38
C	internet explorer	CDocument_write, CScriptElement_put_src, ...	7
D	user interface	DrawText, FindWindow, LoadString, ...	11
E	net API	NetGetJoinInformation, NetShareEnum, ...	6
F	network	DnsQuery_A, GetAdaptersInfo, HttpOpenRequestA, ...	62
G	OLE	CoCreateInstance, CoInitialize, ...	3
H	process	CreateProcess, CreateThread, Module32First, ...	41
I	synchronization	GetLocalTime, GetSystemTime, ...	8
J	resource	FindResource, LoadResource, ...	6
K	services	ControlService, CreateService, ...	12
L	system	GetNativeSystemInfo, LdrLoadDll, NtClose, ...	26
M	certificate	CertControlStore, CertOpenStore, ...	5
N	encryption	CryptCreateHash, CryptGenKey, ...	19
O	exception	SetUnhandledExceptionFilter, RtlDispatchException, ...	6
P	misc	GetUserName, GetDiskFreeSpace, WriteConsole, ...	20
Q	notification	__anomaly__, __exception__, ...	4

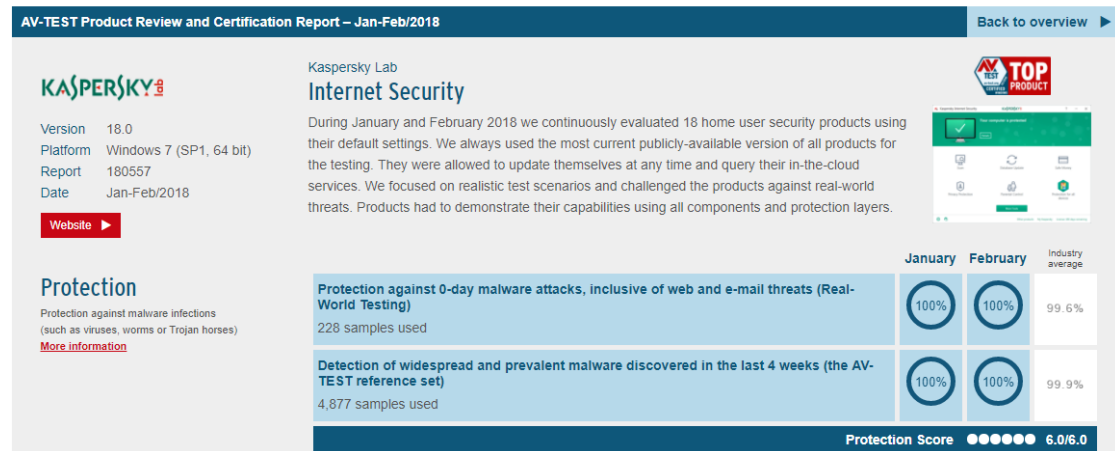
[표 1] API Table (출처 : 고동우, 김휘강(2017) "API 콜 시퀀스와 Locality Sensitive Hashing을 이용한 악성코드 클러스터링 기법에 관한 연구", 정보보호학회논문지)

3) 라벨링

본 프로젝트에서는 지도 학습(Supervised Learning)을 이용하였다. 지도 학습을 위한 악성코드 라벨은 러시아의 IT 보안 업체인 카스퍼스키 랩에서 개발한 안티바이러스인 카스퍼스키의 진단명을 참고하였다. 카스퍼스키는 IT 보안 테스트 및 컨설팅 서비스 제공업체인 AV-TEST에서 2016년 2월부터 현재까지 악성코드 방지점수에서 최고 점수를 받고 있으며, 전 세계에서 실제 감염 활동이나 발견 보고가 있었던 악성코드들의 샘플 목록인 Wildlist를 오탐 없이 100% 탐지해야 받을 수 있는 VB100 인증을 획득하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

The best antivirus software for Windows Home User



[그림 22] AV-TEST의 Protection 평가에서 최고 점수를 받은 카스퍼스키

VB100 results from 2018-02 (latest) on Windows 7 Professional, Windows 10 Professional


Read the full review, or download it.

Tested product	Result	RAP Overview	WildList (%)	WildList (%)	False positives	False positives
Kaspersky Lab Kaspersky Endpoint Security 10 for Windows	Passed VIRUS 100		100.00	100.00	0	0

[그림 23] VB100 인증을 획득한 카스퍼스키

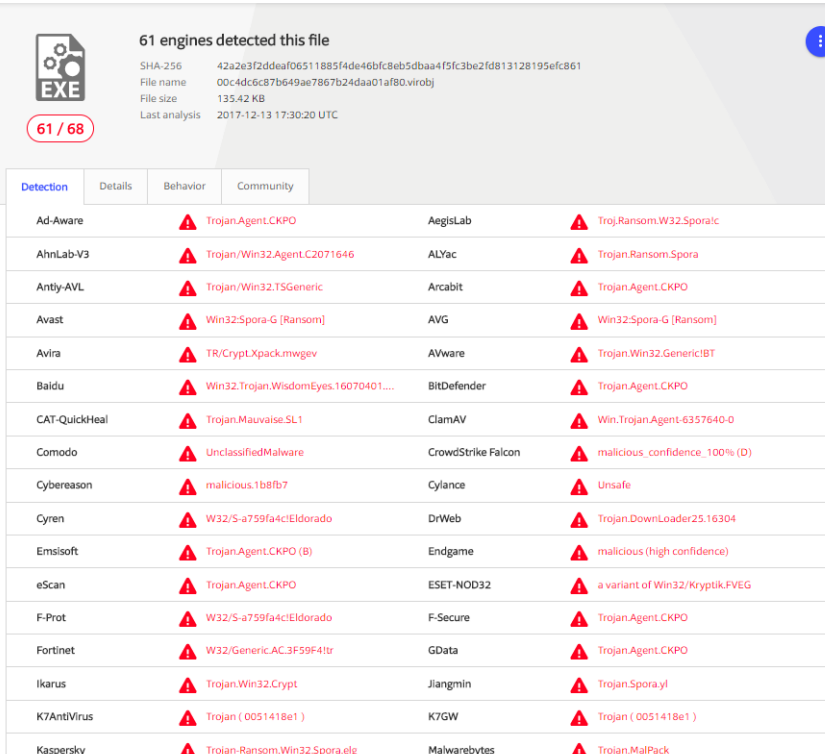
카스퍼스키 랩에서는 악성코드를 다음 7가지로 분류하며 이를 가장 일반적인 악성코드 분류 방법으로 정의하며 본 프로젝트에서는 이러한 7가지 분류를 이용하였다.

- ① 바이러스(Virus) : 사용자 동의없이 기존 프로그램에 설치되는 자체 복제 프로그램 코드 유형으로 감염되는 객체의 종류, 공격대상을 선택하는데 사용하는 방법 혹은 공격방법을 통해 더 세분화될 수 있다.
- ② 웜(Worm) : 자체 복제 프로그램으로 바이러스의 하위 구분으로 간주될 수 있으나 기존 프로그램을 감염시키지 않는 대신 네트워크 취약점을 조작하여 다른 시스템에 퍼져나갈 기회를 찾기 전까지 공격 대상의 컴퓨터에 스스로 설치된다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

- ③ 트로이잔(Trojan) : 정상파일처럼 보이나 실제로는 유해한 행동을 수행하도록 설계되어 있는 악성코드의 유형이다. 트로이잔은 스스로 복제 되지 않기 때문에 퍼져 나가지 않는다. 그러나 인터넷의 범위가 넓어지며 더 많은 대상에게 도달하기 쉬워졌다.
- ④ 랜섬웨어(Ransomware) : 랜섬웨어는 공격 대상의 금전탈취를 목적으로 설계된 악성코드 유형이다. 일단 실행되면 공격 대상의 시스템을 잠그거나 파일을 암호화하여 대상의 행동을 제한하고 팝업창이나 피싱 등의 형태로 금전을 요구한다.
- ⑤ 백도어(Backdoor) : 공격자가 시스템 설계자나 관리자에 의해 고의적으로 남겨진 시스템의 보안상 허점을 이용하여 시스템에 허가되지 않은 접근을 가능하게 하는 악성코드의 유형이다.
- ⑥ 루트킷(Rootkit) : 공격대상이 설치한 기존 보안 소프트웨어의 존재와 작동을 숨기도록 설계된 특수한 형태의 악성코드 유형이다.
- ⑦ 다운로더(Downloader) : 추가적으로 악성코드를 다운로드 하는 악성코드의 유형이다. 악성코드가 설정한 웹사이트로 접속하여 추가적인 악성코드를 다운받아 감염시킨다.

카스퍼스키의 진단명을 가져오기 위해서는 바이러스토탈(VirusTotal)을 이용한다. 바이러스토탈은 파일에 대한 분석과 그에 대한 결과를 제공해주는 온라인 서비스로 분석결과로 약 60 여개의 안티바이러스 엔진의 탐지 결과를 보여준다. 바이러스토탈에서는 파일의 분석 결과를 받아올 수 있는 API 를 제공한다. 제공된 AP 를 통하여 악성코드의 분석을 요청하면 json 파일로 분석 결과를 받게 되며, 분석 결과에서 필요한 정보를 간편히 추출할 수 있도록 파서를 개발하여 사용할 카스퍼스키의 진단명을 추출하였다.



61 engines detected this file

SHA-256: 42a2e3f2ddea06511885f4de46bfc8eb5dbaa4f5fc3be2fd813128195efc861
File name: 00c4dc6c87b649ae7867b24daa01af80.virolj
File size: 135.42 KB
Last analysis: 2017-12-13 17:30:20 UTC

61 / 68

Detection	Details	Behavior	Community
Ad-Aware	Trojan.Agent.CKPO	AegisLab	Troj.Ransom.W32.Sporalc
AhnLab-V3	Trojan/Win32.Agent.LC2071646	ALYac	Trojan.Ransom.Spora
Antiy-AVL	Trojan/Win32.TSGeneric	Arcabit	Trojan.Agent.CKPO
Avast	Win32:Spora-G [Ransom]	AVG	Win32:Spora-G [Ransom]
Avira	TR/CryptXpack.mwgev	AVware	Trojan.Win32.Generic!BT
Baidu	Win32.Trojan.WisdomEyes.16070401....	BitDefender	Trojan.Agent.CKPO
CAT-QuickHeal	Trojan.Mauvaise.SL1	ClamAV	Win.Trojan.Agent-6357640-0
Comodo	UnclassifiedMalware	CrowdStrike Falcon	malicious_confidence_100% (D)
Cybereason	malicious.1b8fb7	Cylance	Unsafe
Cyren	W32/S-a759fa4c!Eldorado	DrWeb	Trojan.DownLoader25.16304
Emsisoft	Trojan.Agent.CKPO (B)	Endgame	malicious (high confidence)
eScan	Trojan.Agent.CKPO	ESET-NOD32	a variant of Win32/Kryptik.FVEG
F-Prot	W32/S-a759fa4c!Eldorado	F-Secure	Trojan.Agent.CKPO
Fortinet	W32/Generic.AC.3F59F41tr	GData	Trojan.Agent.CKPO
Ikarus	Trojan.Win32.Crypt	Jiangmin	Trojan.Spora.yf
K7AntiVirus	Trojan (0051418e1)	K7GW	Trojan (0051418e1)
Kaspersky	Trojan-Ransom.Win32.Spora.elg	Malwarebytes	Trojan.MalPack


[그림 24] 바이러스토탈의 파일 분석결과 화면

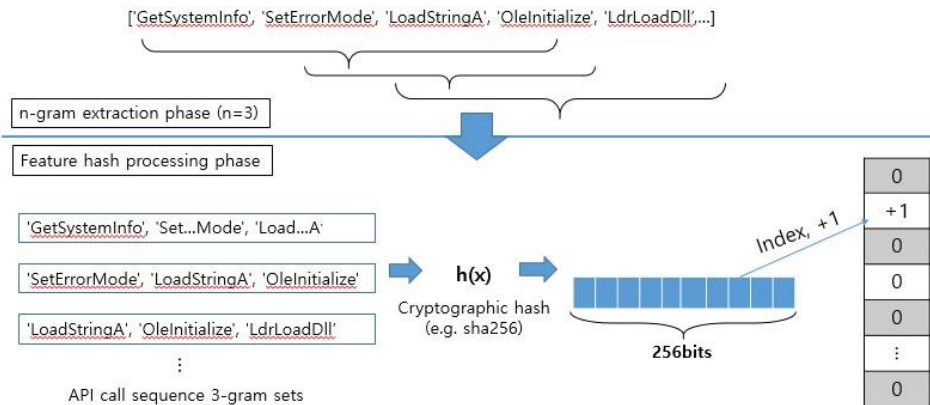
4) 피처 가공 및 딥 러닝

정적 및 동적 분석을 바탕으로 추출된 정보들을 딥 러닝 모델에 적용하기 전 전처리 과정을 거쳤다. opcode 시퀀스와 API 콜 시퀀스는 각각 3, 4, 5-gram 기법을 거친 후, 피처 해싱 기법을 적용한다. 본 논문에서는 암호학적 해시 함수를 사용하여 인덱스를 구하고 피처 벡터를 생성하였다.

본 연구는 [그림 25]과 같이 n-gram 기법을 거친 API 콜 시퀀스 셋에 암호학적 해시 함수를 적용하였고, 이를 통해 나온 256비트의 값을 피처 벡터의 최대 크기인 4,096으로 나눈 나머지로 인덱스를 구하였다. 그리고 인덱스에 해당하는 배열 값을 1을 증가시킴으로써 피처 벡터를 생성하였다. 마지막으로 피처의 범위를 표준화하기 위하여 스케일링 과정을 거쳤다. 피처 벡터의 max 값과 min 값을 추출한 후 두 값이 같지 않은 경우 아래와 같은 수식을 거쳐 0과 1 사이의 값으로 스케일링을 하였다.

$$\frac{x - \min(\text{vector})}{\max(\text{vector}) - \min(\text{vector})} \quad (\text{if } \min \neq \max, \text{ for } x \text{ in } v)$$

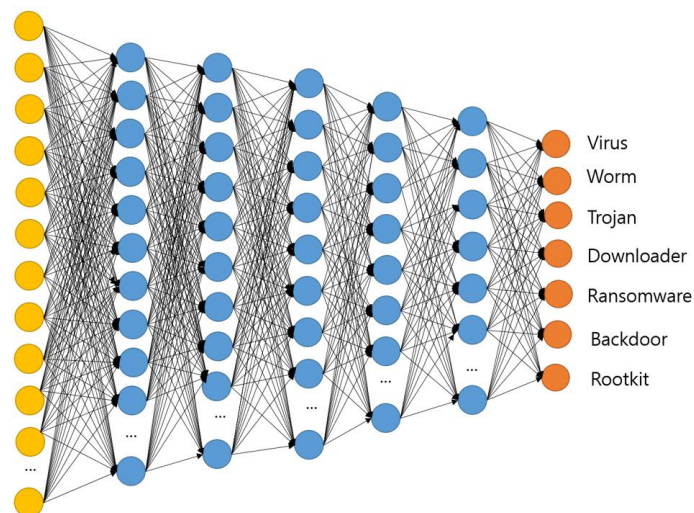
	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29




[그림 25] API 콜 시퀀스에 대한 3-gram과 피쳐 벡터 생성 과정

본 프로젝트에서는 텐서플로우(Tensorflow)를 이용하여 ANN(Artificial Neuron Network) 모델을 만들었다. Virussign, Virusshare, Malshare 등에서 다운로드 받은 악성코드와 신뢰 할 수 있는 개발사(마이크로 소프트, Adobe, Antivirus 등)에서 개발한 정상 소프트웨어를 이용하여 학습하였다. 본 프로젝트에서는 정적 모델의 경우 약 40만개를 동적 모델에 대해서는 대략 10만개를 학습시켰다.

[그림 26]은 악성코드 탐지를 위한 심층 신경망 모델의 구조이다. 심층 신경망의 입력 층 노드 개수는 12,288개이며, 2, 3, 4 - gram을 Feature Hashing 기법을 이용해 가공하였다. 심층 신경망의 은닉층 개수는 총 5개이며 각각 4096, 1024, 256, 64, 16개의 노드를 가지고, 활성화 함수로는 ReLU를 사용하였다. 또한 과적합을 막기 위해서 드랍 아웃 기법을 사용하였다. 제안하는 모델은 30%의 정보를 잊어버리도록 설정하였으며 마지막 은닉층을 거쳐 출력층에 도달할 때 소프트 맥스 함수를 이용하여 주어진 파일이 정상 또는 악성에 속할 확률로 변환한 뒤, 더 높은 확률을 선택하여 어떤 그룹에 속하는지 분류하였다.



[그림 26] 악성코드 분류를 위한 모델 구조

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

5) 데이터베이스

본 프로젝트는 대량의 악성코드 파일을 정적, 동적 분석한 결과와 딥 러닝 모델의 테스트 결과를 json 형식의 리포트로 저장한다. 만들어진 각종 리포트 파일은 웹에서 분석 결과를 찾아 화면에 보여주기 위해 잦은 검색이 이루어진다. 따라서 빠른 검색 기술과 json 형태의 데이터를 쉽게 운용할 수 있는 데이터 저장소를 필요로 한다. 또한 리포트들은 웹에서 키바나(Kibana)를 통해 통계 정보를 시각화 하여 보여진다. 이를 종합하여 본 프로젝트에서는 데이터베이스 겸 검색 엔진으로 역 인덱싱 기술을 이용하여 빠른 검색을 지원하고 키바나와 연동할 수 있는 엘라스틱서치(ElasticSearch)를 채택하였다.

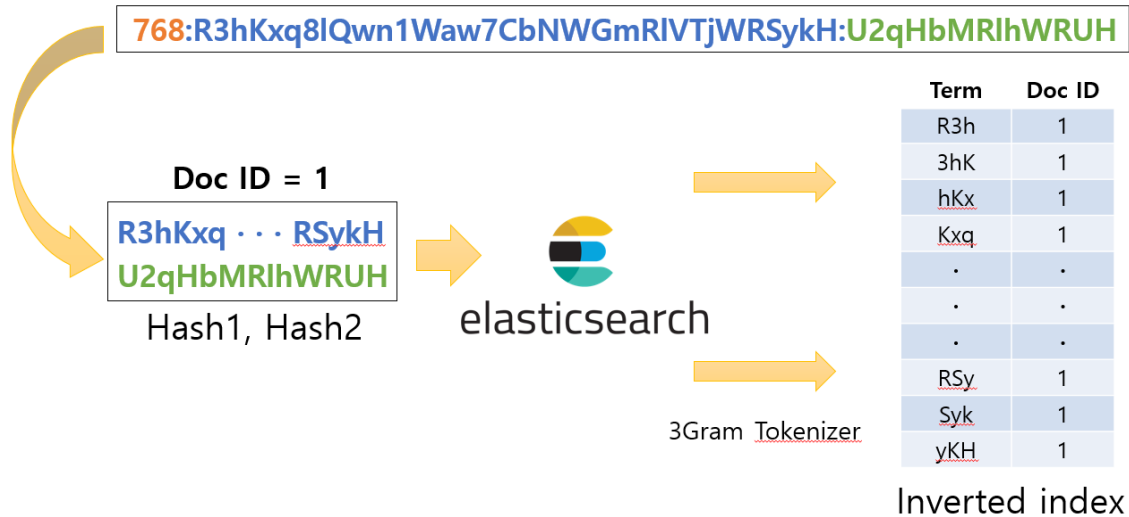
엘라스틱서치에는 정적 리포트를 저장하는 static_report 인덱스를 만들었고, md5, 바이러스 유무, 바이러스 라벨, 업로드 날짜를 문서로 작성하여 자동 인덱싱 되도록 하는 스크립트와 md5 검색을 통해 원하는 리포트를 조회할 수 있는 스크립트를 작성하였다. 쿠쿠샌드박스의 경우 자체적으로 제공하는 엘라스틱서치와 연동 모듈을 사용하여 cuckoo 인덱스를 운용하였다.

다음으로 본 프로젝트는 업로드한 파일에 대해 분석 후 이와 유사한 파일의 기존 분석 결과를 동시에 보여주는 서비스를 목표로 하였다. 이때 유사한 파일을 보여주기 위해 파일 유사도 측정 기법을 사용하여야 한다. 이를 위해 SSDeep 이라는 파일 유사도 측정 툴을 사용하였다. SSDeep은 하나의 스트링을 특정 알고리즘을 통해 해시값을 도출해 내고, 같은 방법으로 또 다른 스트링의 해시 값을 도출해 내어 두 해시값 간의 차이를 비교하여 유사도를 측정하게 된다. 이 때 행위 기반의 유사도 측정을 위해 파일 전체에 대한 SSDeep 알고리즘 적용 대신, 파일의 opcode 시퀀스를 추출하여 SSDeep 해시 값을 도출해 진행하게 되었다.

이렇게 도출된 SSDeep 해시값들은 SSDeep에서 제공하는 비교 함수를 사용하여 유사도 점수를 얻을 수 있다. 하지만 전체 리포트 n개에 대해서 전부 비교함 수를 사용해 유사도를 비교하는 것은 n번의 함수 호출이 일어나게 되므로 검색하는데 많은 시간이 소요되게 된다.

이를 극복하기 위하여 도출된 SSDeep 해시를 엘라스틱서치(Elasticsearch)를 이용해 역 인덱싱하여 검색 하는 방법을 고안하였다. 해시 전체를 통째로 역 인덱싱 하게 되면 완전히 같은 파일이 아닌 이상 검색되지 않게 된다. 따라서 해시 스트링을 n-gram tokenize하여 역 인덱싱 해야 한다. 본 프로젝트는 실험적인 결과를 통해 3 gram tokenize를 하기로 하였다.

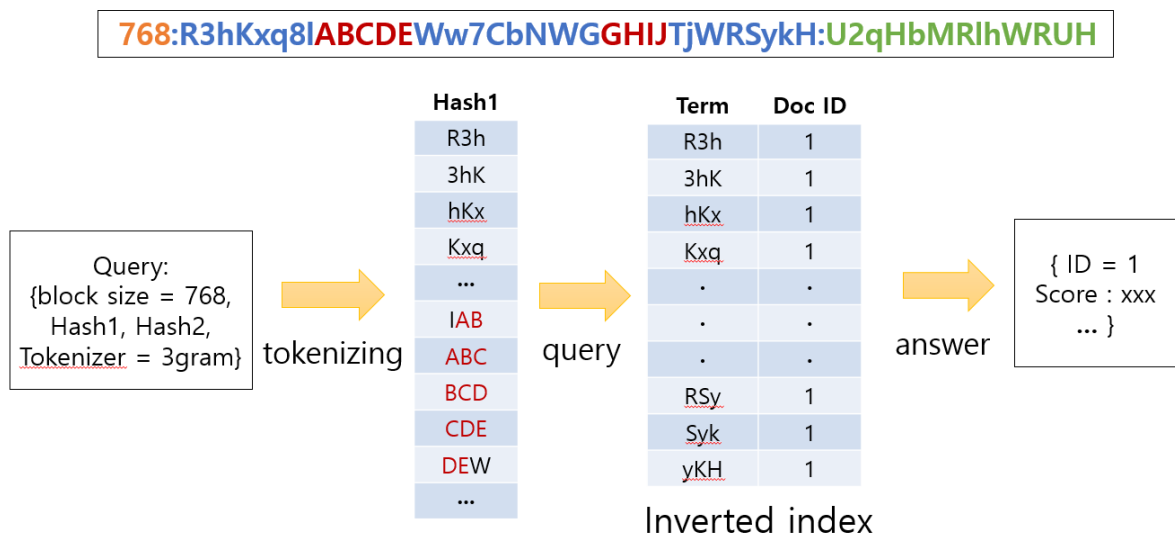
Block size : Hash1 : Hash2




[그림 27] 3-gram tokenize한 문서를 역 인덱싱 하는 과정

전체 리포트에 대해 역 인덱싱이 완료되면 업로드 된 파일을 3-gram tokenize 한 후 준비 된 인덱스에 쿼리를 보내게 되면 위드가 가장 많이 매칭된 순서대로 리포트를 보여주게 된다.

Similar SSDeep string



[그림 28] 업로드 한 파일의 SSDeep을 인덱스에 쿼리를 보내는 과정

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

6) 웹

본 프로젝트는 웹을 통해 사용자에게 분석 결과를 제공한다. 웹은 부트스트랩(Bootstrap) 프레임워크를 이용하여 사용자의 브라우저 크기에 따라 변하는 반응형으로 제작하였다.

웹 상단에는 페이지 간의 이동을 위해 네비게이션 바를 배치하였다. 좌측에 팀 로고(NCNP)와 프로젝트 로고(MASK)를 배치하였고, 프로젝트 로고를 클릭 시 초기화면으로 돌아올 수 있도록 하였다. 또한 브라우저의 폭이 768px 보다 작으면 목록형 네비게이션 바로 변경되도록 하였다.

초기 화면에서 이미지가 계속해서 움직이거나 서서히 나타나도록 해서 사용자의 이목을 화면에 집중시키고자 하였다. 상단에는 본 프로젝트의 사용자가 악성코드 분석에 전문지식이 있는 사람인 것을 명시하였고, 하단에는 타 악성코드 분석 프로그램과 차별화 되는 장점을 명시하였다.

분석 화면에서 사용자가 악성으로 의심되는 파일을 업로드 할 수 있도록 폼을 만들었고, 이 폼에 파일을 드래그 앤 드랍을 하거나 원하는 파일을 선택하여 업로드 할 수 있도록 하였다. 폼의 좌측 상단에는 radio button을 배치하여 사용자가 정적 분석과, 동적 분석 중 하나를 선택할 수 있도록 하였다.


업로드가 시작되면 업로드 및 분석이 완료되는 동안 사용자에게 서비스가 멈춘 것이 아닌 분석 중인 것을 알리기 위해 움직이는 이미지와 분석 중이라는 안내문구를 포함한 로딩 화면을 출력할 수 있도록 했다.

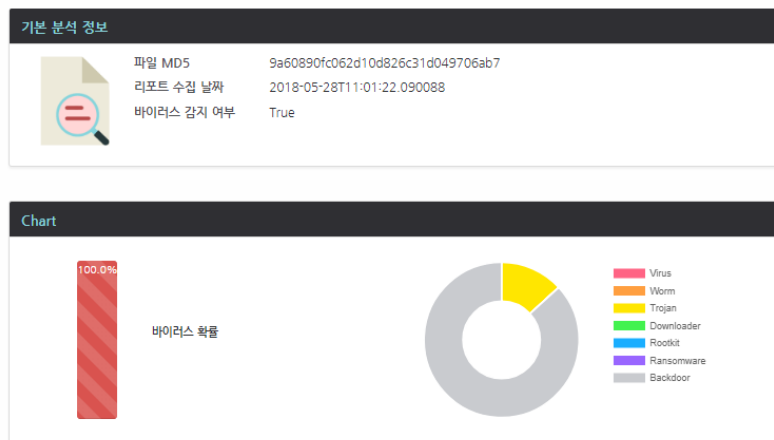
스크립트 단에서는 Ajax를 이용해 업로드 완료 이벤트를 Listen하며 이벤트 감지 시 다음 페이지로 분석 요청을 보내도록 코드를 작성하였다.

업로드가 완료되어 분석 요청이 오게 되면 분석 프로그램을 실행시키도록 하였고, 분석 프로그램을 실행할 때 인자로 업로드 된 파일의 경로를 넘기도록 하였다. 분석이 완료 된 후 에는 생성되는 리포트를 데이터베이스에 업로드와 동시에 결과 페이지에 리포트 폼을 전송하도록 하였다.

결과 화면에서 분석이 완료된 리포트 폼을 받게 되면 사용자가 선택한 분석 종류에 따라 다른 결과 화면이 출력된다.

정적 분석에서는 리포트에 있는 분석한 파일의 정보(md5, 바이러스 유무, 바이러스 분석 결과, 리포트 수집 날짜), 이와 유사한 파일에 대한 분석 결과, 바이러스 확률(파일의 전체 바이러스 확률, 바이러스 종류에 따른 확률)을 파싱해 분석 결과 화면을 구성하여 사용자에게 제공할 수 있도록 하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29



[그림 29] 정적 분석에 대한 결과 화면 (1)

PE 파일 정보

Basic Info

MD5: 9a60890fc062d10d826c31d049706ab7
 SHA-1: 3ae8d97461fb08c4327431c0589322e3cbb1e3de
 SHA-256: c89944f9ec704c2b8da3a1ac726699022e7c68334110f72007d762217a9a4a5
 Imp-Hash: 47b0da2d13e0214f54c3bd05550e8319

Section Info

Name	Virtual Address	Virtual Size	Raw Data Size	Suspicious	MD5
.code	0x1000	0x3589	13824	False	fcac6494dcc5d68e347fb1645b3dc5e3
.text	0x5000	0xb3d1	46080	False	a166dd595c84d03f0a22b43db95f8672
.rdata	0x11000	0x986	2560	False	4770f66539d1d3273544f30f0f101075
.data	0x12000	0x1bd8	5632	False	08cecdcc3096412bf3d9a5d71b1323ad
.rsrc	0x14000	0x1225b0	1189376	True	16872ad91acfcc1c31f1aadf7b771ae7


Import Functions

- MSVCRT.dll
- KERNEL32.dll
- USER32.DLL
- GDI32.DLL
- COMCTL32.DLL
- OLE32.DLL
- SHELL32.DLL
- WINMM.DLL
- SHLWAPI.DLL

API Alert Info

- CloseHandle
- CreateDirectoryA
- CreateFileA
- DeleteCriticalSection
- DeleteFileA
- ExitProcess
- FindResourceA
- GetCommandLineA
- GetFileSize
- GetModuleFileNameA
- GetModuleHandleA
- GetTempFileNameA
- GetTempPathA
- ShellExecuteExA
- Sleep
- TerminateProcess
- WriteFile

[그림 30] 정적 분석에 대한 분석 결과 화면 (2)

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

동적 분석에서는 리포트에 있는 분석한 파일의 정보(md5, 바이러스 유무, 바이러스 분석 결과, 리포트 수집 날짜), 바이러스 확률, 시그니처(파일의 전체 바이러스 확률, 바이러스 종류에 따른 확률), DLL, 연결된 호스트와 주소를 파싱해 분석 결과 화면을 구성하여 사용자에게 제공할 수 있도록 하였다.



오픈소스 프로젝트
본 서비스는 오픈소스 소프트웨어로 제공됩니다.



딥러닝 사용
딥러닝을 이용하여 악성코드 탐지 서비스를 제공합니다.

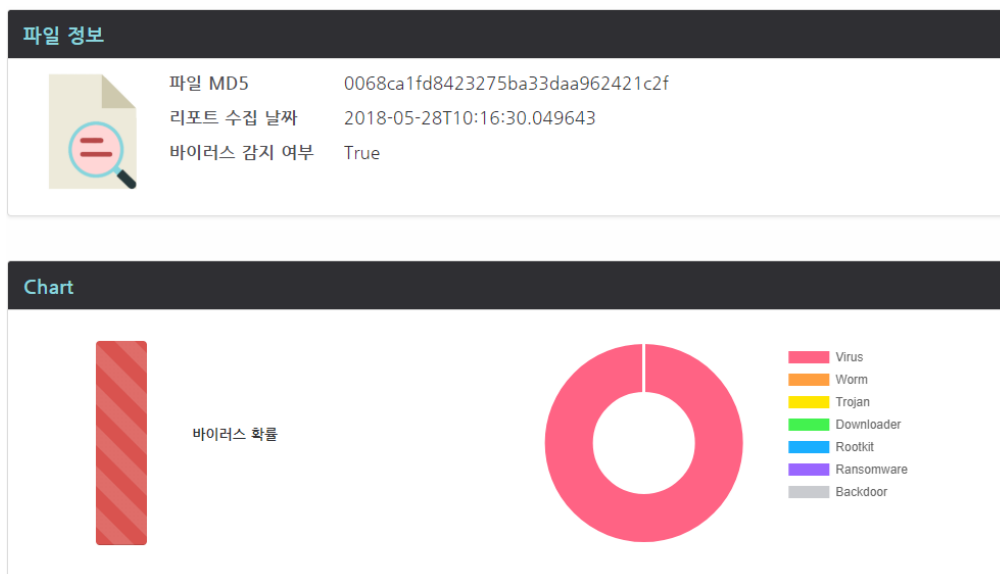


정적 동적분석
본 서비스는 정적분석을 위해 JDA를 사용하고, 동적분석을 위해 Cuckoo Sandbox를 사용합니다.




유사 파일 검사
악토르한 파일과 유사한 파일 출력이 되는 검사 결과도 제공됩니다.

[그림 31] 웹 페이지 메인 화면

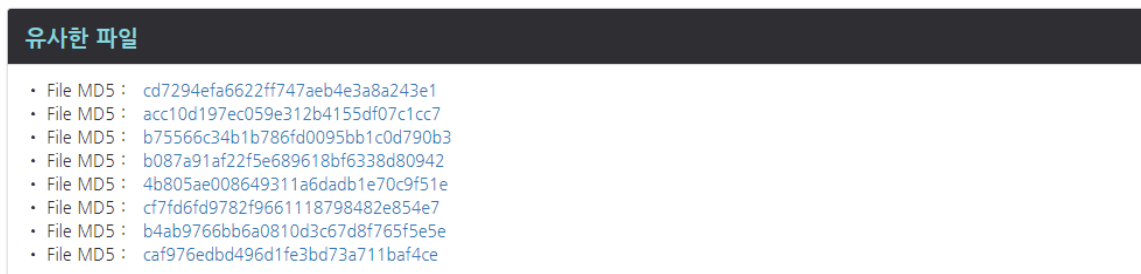


[그림 32] 동적 분석에 대한 결과 화면 (1)

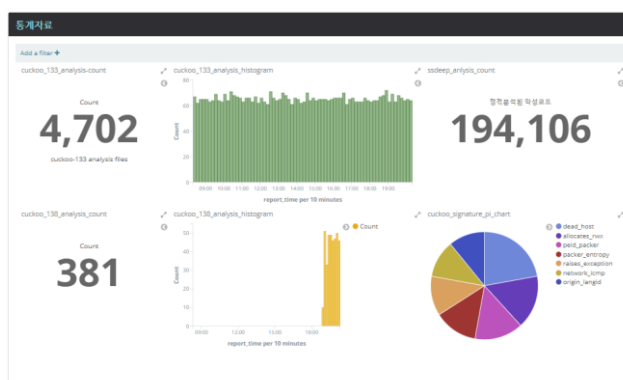
 <div> 국민대학교 컴퓨터공학부 캡스톤 디자인 I </div>	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29




[그림 33] 동적 분석에 대한 결과 화면 (2)



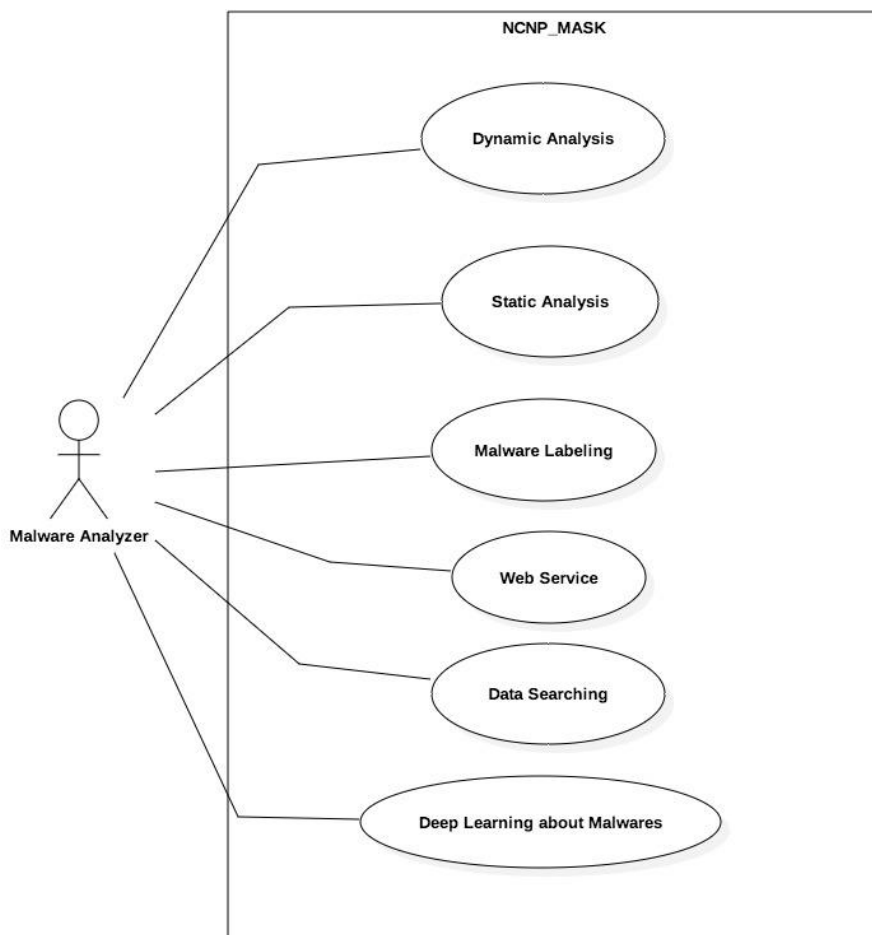
[그림 34] 분석한 파일과 유사한 파일에 대한 결과



[그림 35] 키바나를 이용한 통계자료 화면

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

2.2.2 시스템 기능 요구사항



[그림 36] 시스템 기능 요구사항 use case diagram


1) 파일에 대한 동적 분석 - 완료

악성코드를 분석환경에서 실행시켜 행동 변화를 확인하는 동적 분석을 진행한다.

- ➔ 악성코드 분석을 위한 시스템으로는 오픈소스 소프트웨어인 Cuckoo Sandbox를 이용하였다. 한 파일 당 분석 시간은 최대 2분으로 설정하였다. 분석 완료 후엔 생성된 리포트로부터 API 콜 시퀀스를 추출하였다.

2) 파일에 대한 정적 분석 - 완료

악성코드에 대해 실제 실행 없이 컴퓨터 소프트웨어를 분석하는 정적 분석을 진행한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

- ➔ 악성코드 분석을 위한 시스템으로는 컴퓨터 소프트웨어용 디스어셈블러인 IDA와 파이썬 기반의 오픈소스 도구인 pefile을 사용하였다.

3) 분석한 악성코드에 대해 라벨 부여 - 완료

악성코드의 바이러스토탈 분석 결과를 이용하여 라벨링을 진행한다.

- ➔ 악성코드의 바이러스토탈 분석 결과는 바이러스토탈 API를 이용하여 json 형태로 받아오며, 바이러스토탈 분석 결과 파서를 작성하여 각 악성코드의 [md5, 탐지 결과, 진단명]을 저장하는 CSV 형태의 리포트를 제작하였다. 이때 카스퍼스키 탐지 결과와 진단명을 이용하였다.

4) 데이터 검색 - 완료

분석 요청한 파일과 유사한 파일을 검색하여 사용자에게 제공한다.


- ➔ Ssdeep 유사도 측정 툴을 이용하여 해당 파일에 대한 유사도 해시를 구하고, 준비되어 있는 기존 파일들의 유사도 해시와 비교하여 유사한 파일을 찾았다. 이때 유사도 해시를 비교 분석하기 위해 3gram tokenize 기법을 이용하였다.

5) 딥 러닝 모델 - 완료

- ➔ ANN 모델을 구축함으로써 악성코드에 대해 7가지의 카테고리로 분류할 수 있다.

6) 파일 업로드 제한 - 변경

- ➔ 분석할 파일이 32MB가 넘을 경우, 업로드를 제한하도록 하였다. 또한 PE file만 업로드 하도록 하였으며, .NET 파일이나 opcode 시퀀스 및 API 콜 시퀀스가 너무 짧은 경우에도 업로드를 제한하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

2.2.3 시스템 비기능(품질) 요구사항

1) 재사용성 - 달성

본 서비스는 기존에 분석 결과가 존재하는 파일에 대하여 중복된 요청이 왔을 경우 빠른 응답을 제공하기 위해 DB로부터 기존 분석결과를 검색하여 보여준다.

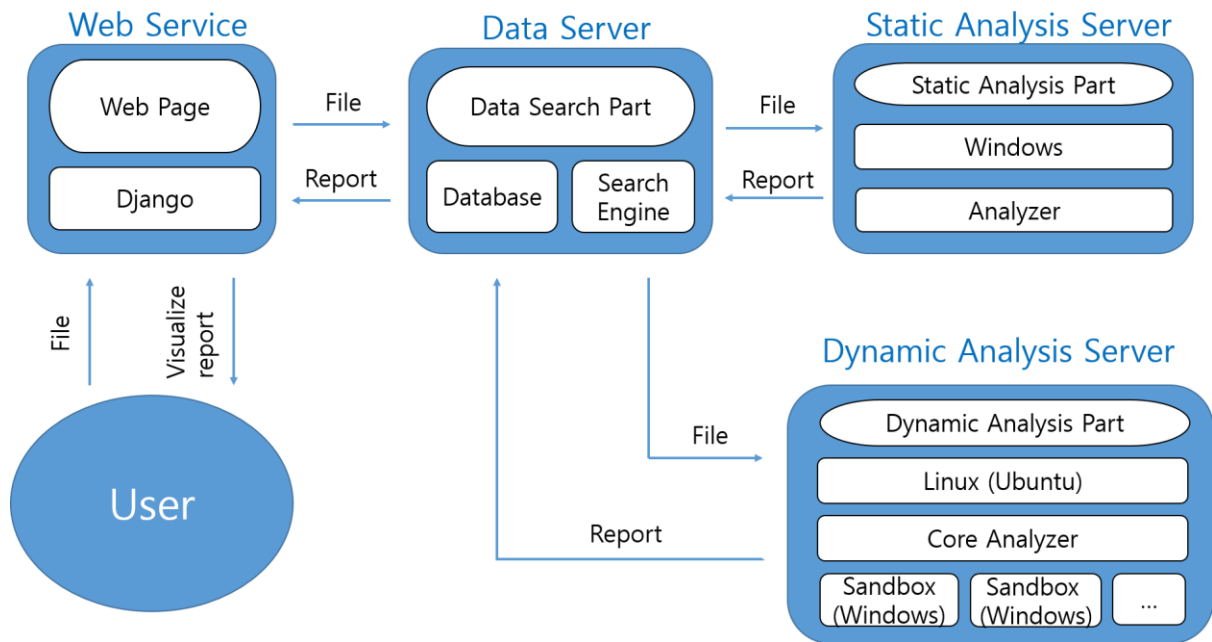
2) 사용성 - 달성

본 서비스의 목적은 악성코드 분석 전문가를 대상으로 악성코드 분석결과와 유사한 악성코드에 대한 정보를 보여 주어 전문가가 분석 할 수 있게 도와주는 것이다. 따라서 분석에 용이 하도록 기존 데이터들에 대한 요약 정보를 시각화 하여 보여 주고, 업로드한 악성코드의 행동 정보를 그래프로 연결하여 나타내도록 한다.

3) 성능 - 달성


우리가 제안하는 시스템은 악성코드 분석 방법으로 정적 분석과 동적 분석을 사용하게 되는데, 동적 분석의 경우 악성코드의 행동 분석을 하는데 정적 분석에 비해 한 파일 당 최대 3분 이상이 걸리게 된다. 따라서 여러 개의 분석 인스턴스를 구성하여 여러 파일에 대해 병렬적으로 분석할 수 있도록 하였다.

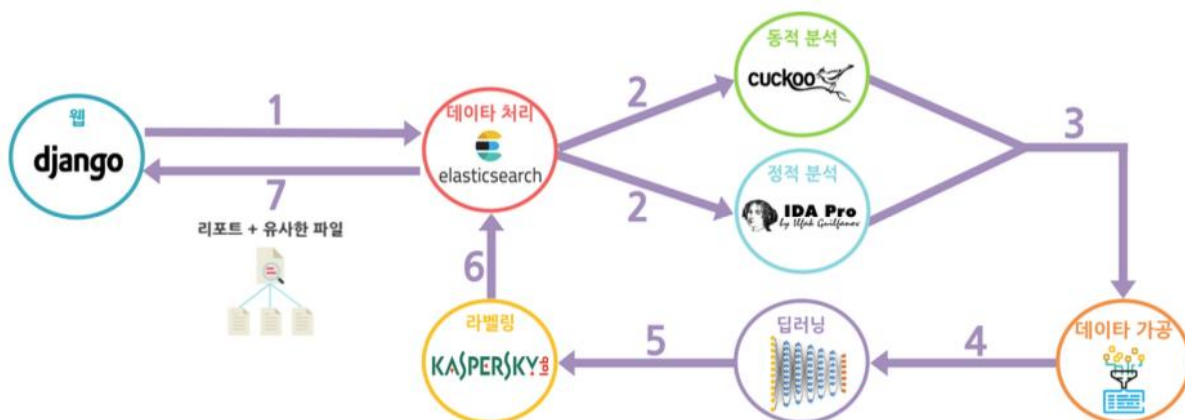
2.2.4 시스템 구조 및 설계도



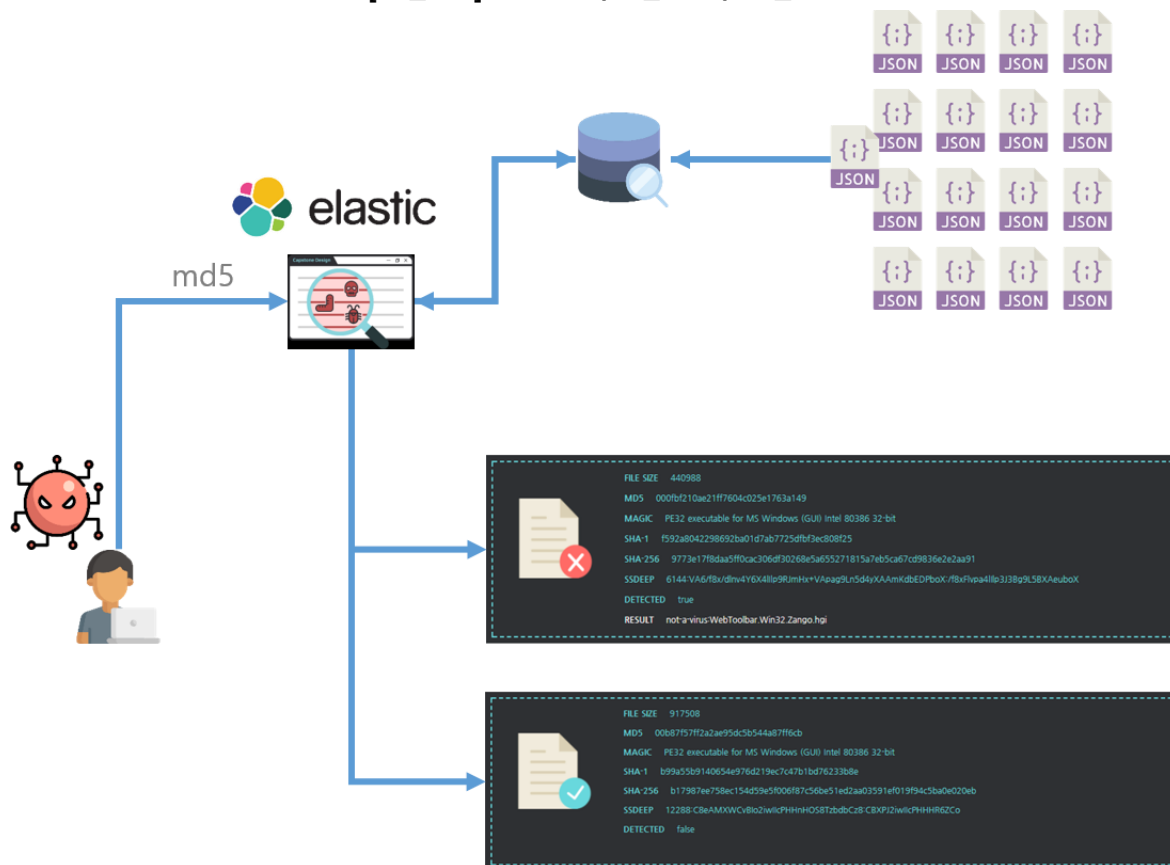
[그림 37] 시스템 구조도

- 웹 서비스 : 유저가 파일을 업로드 할 수 있도록 웹 서비스를 구성한다. 웹 서비스는 파이썬 기반의 웹 프레임워크인 Django를 이용하여 구성한다.
- 데이터 서버 : 유저가 업로드한 파일의 분석결과가 데이터베이스에 존재하는지 검색한다. 만약 존재하지 않는다면 정적, 동적 분석 서버에 각각 업로드한 파일을 전송한다. 분석이 완료되어 분석 결과가 도착하면 데이터베이스에 저장하고 웹으로 전송한다.
- 정적 분석 서버 : 정적 분석을 완료한 후 분석 리포트를 생성한다.
- 동적 분석 서버 : 동적 분석 코어와 분석이 이루어질 1개 이상의 샌드박스로 구성한다. 각각의 샌드박스가 분석을 완료하면 분석 포트를 생성한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29



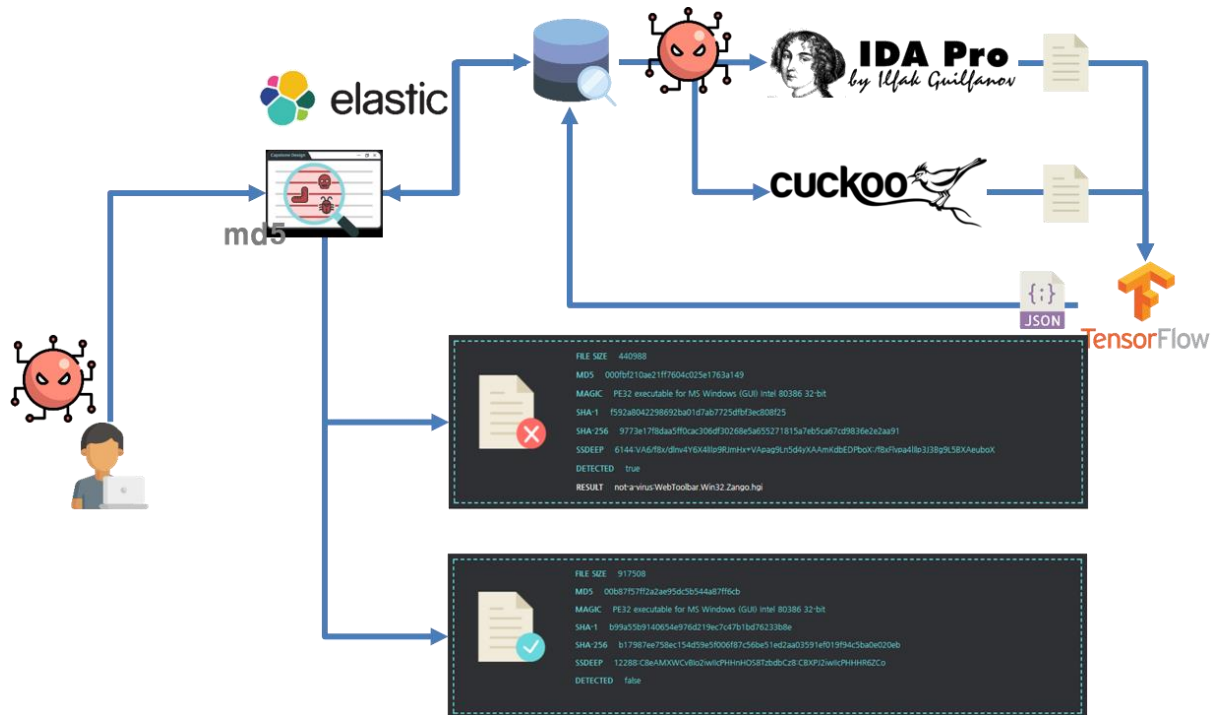
[그림 38] MASK 시스템 전체 흐름도



[그림 39] 데이터베이스에 분석 결과가 존재하는 경우 시스템 흐름도

사용자가 악성으로 의심되는 파일을 웹에 업로드하면, 그 파일의 md5를 구한 후 데이터베이스에 연동된 엘라스틱서치를 이용해 분석 결과를 찾는다. 리포트가 데이터베이스에 존재하면 결과와 그 연관된 정보들을 웹을 통하여 사용자에게 보여준다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29




[그림 40] 데이터베이스에 분석 결과가 존재하지 않는 경우 시스템 흐름도

사용자가 업로드한 파일에 대한 리포트가 데이터베이스에 없으면 그 파일에 대해 각각 정적 분석과 동적 분석을 진행 후 리포트를 생성하고, 이로부터 피처를 추출한다. 추출한 피처를 이용하여 학습된 모델로부터 탐지 결과를 구한다. 이 결과들을 데이터베이스에 저장하고, 웹을 통해 사용자에게 결과를 보여준다.

2.2.5 활용/개발된 기술

- Pefile
pefile은 파이썬에서 PE파일을 읽기 위한 모듈이다. 현재 최신 버전은 2017.11.5 버전이며, 본 프로젝트에서는 업로드 된 PE파일에 대한 정적 분석을 위해 사용 되었다.
- SSDeep
SSDeep 알고리즘을 이용하여 파일의 유사도 판단 기준이 되는 해시를 얻었다. 얻어진 H 해시는 다른 파일의 해시와 비교하여 얼마나 유사 한지 측정하는데 사용 되었다.
- 역 인덱싱
엘라스틱서치는 기본 인덱싱 기술로 역 인덱싱을 사용한다. 엘라스틱서치의 빠른 역 인덱싱 기술을 활용하여 SSDeep 해시 간의 비교 속도를 기존의 단순 비교보다 빠르게 개선 하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

2.2.6 현실적 제한 요소 및 그 해결 방안

1) 하드웨어


- AV-TEST 에 따르면 2017 년 기준으로 하루에 대략 30 만 개의 신규 악성코드가 유입된다. 이를 우리 서비스를 이용해서 동적 분석하기 위해서는 프로세서 자원이 많이 필요하다. 현실적인 방안으로 좋은 서버를 구축하는 거보다 비교적 저 사양의 서버들을 많이 확보하여 분산 환경 구축을 통해 가격 효율성을 높인다.

2) 소프트웨어

- 파일로부터 어셈블리어를 추출할 수 있는 디스어셈블러가 필요하다.
- PE 파일 외에는 지원하지 않는다.
- 악성코드가 실행 코드 암호화나 패킹(Packing) 등 은닉기술이 적용될 경우, 혹은 신규 악성코드가 발견되었을 경우 정적 분석에 불리하다. 또한, 악성코드가 분석 환경을 인지하여 회피가 가능할 경우 동적 분석에 불리하다. 따라서 본 프로젝트는 정적 분석과 동적 분석을 동시에 사용하여 두 분석 방법의 단점을 상호보완한다.

2.2.7 결과물 목록

- 악성코드 분석 프로그램 (사용자 매뉴얼, 설치 매뉴얼 유)

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

2.3 기대효과 및 활용방안

2.3.1 기대효과

- 본 프로젝트에서 제안하는 시스템은 악성코드 분석 전문가가 악성코드로 의심되는 파일을 효율적으로 분석할 수 있도록 한다. 최근 생성되는 악성코드는 신종 뿐만 아니라 기존 악성코드의 변종이 많기 때문에, 유사한 악성코드가 포함된 분류를 알아내기만 하면 이후의 행위들을 예측하여 대응하기 용이하다.
- 본 프로젝트에서 제안하는 시스템은 악성코드 분류 기술에 딥 러닝을 적용하여 분석 효율성을 높인다. 우리 서비스는 시그니처 기반 탐지가 아닌 딥러닝 기반 탐지 시스템이기 때문에 악성코드 행동 예측을 함으로써 사이버공격에 대한 선제적 대응 능력을 확보할 수 있고, 신종/변종 악성코드에 대한 피해의 최소화를 기대한다.
- 본 프로젝트에서 제안하는 시스템은 오픈소스 소프트웨어로 공개해 누구나 원하는 기능을 추가하여 악성코드를 분석할 수 있다.

2.3.2 활용방안

- 악성코드 분석 전문가가 자신의 환경에 맞게 설치 및 서비스 이용을 할 수 있다. 우리와 유사한 타 서비스들 중 대부분은 오픈소스 소프트웨어가 아닌 것에 비해, 우리 프로젝트는 오픈소스 소프트웨어이므로 전문가가 자신만의 서비스를 구축함으로써 악성코드 분석 시장이 활성화되는 것을 기대한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

3 자기평가

최종 결과물 – 딥 러닝 기반의 악성코드 분석 시스템


주요 평가 기준 – 알려진 악성코드 및 신종/변종 악성코드에 대해 올바른 분석을 하는가?

매일 대략 100 만 개 정도의 악성코드가 발견되고 있다. 악성코드 탐지 기법 중 시그니처 기반 탐지 기법은 상용화된 안티바이러스에서 가장 널리 쓰이는 기법 중 하나이다. 하지만 상당수의 악성코드는 암호화 기법이나 은닉 기술이 적용되는데, 이 경우 변종 악성코드를 탐지하는 데에 한계가 발생한다. 또한 점차 지능화되어가는 악성코드는 실행 압축, 난독화, Anti-VM 등 분석을 회피하는 기술과도 결합되고 있다. 또한 늘어나는 악성코드에 비해 이를 분석할 수 있는 악성코드 분석 전문가의 수도 현저히 부족한 상황이다.

우리는 본 캡스톤 프로젝트를 통해 악성코드 분석 방법 중 정적 분석과 동적 분석을 이용하여 이로부터 유용한 정보를 추출하고, 딥 러닝 모델에 적용하여 악성코드를 분석할 수 있는 시스템을 개발하였다.

우리가 제시하는 방법은 높은 정확도로 악성코드인지 아닌지 구별할 수 있고, 더 나아가 7 가지에 카테고리로 분류하여 분석한 악성코드가 어느 종류의 악성코드인지 구별하여 효과적으로 분석할 수 있다. 나아가 이를 오픈소스로 하여 사용자가 원하는 대로 확장하여 사용할 수 있다.

- 본 프로젝트는 보안 산업체들과의 과제와 연계하여 진행되었음.
- “동적 분석과 심층 신경망 모델을 이용한 악성코드 분류”, 한채연·김영재·허준녕·명준우, KCC 2018(한국컴퓨터종합학술대회) 논문 발표 예정.
- 본 프로젝트를 활용하여 KISA R&D challenge 대회에서 2등 수상.


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

4 참고 문헌

번호	종류	제목	출처	발행년도	저자
1	논문	API 콜 시퀀스와 Locality Sensitive Hashing 을 이용한 악성코드 클러스터링 기법에 관한 연구	정보보호학회논문지	2017	고동우, 김휘강
2	논문	Idea: Opcode-sequence-based Malware Detection	International Symposium on Engineering Secure Software and Systems	2010	Igor Santos 외 6 명
3	논문	Using opcode 시퀀스 s in single-class learning to detect unknown malware	IET Information Security	2011	I. Santos 외 2 명
4	논문	Detecting unknown malicious code by applying classification techniques on OpCode patterns	Security Informatics	2012	Asaf Shabtai 외 4 명
5	논문	MutantX-S: Scalable Malware Clustering Based on Static Features	USENIX Annual Technical Conference	2013	Xin Hu 외 3 명

4.1 참고 사이트

참조번호	제목	출처 및 사이트
1	‘워너크라이’ 랜섬웨어 감염예방법 “인터넷·파일공유 차단한 뒤 컴퓨터 켜야”	https://byline.network/2017/05/1-712/
2	[3 차공지] Erebus Encrypted 로 인한 시스템 장애	http://www.nayana.com/bbs/set_view.php?b_name=notice&w_no=959
3	가상화폐 채굴형 악성코드 Miner	http://www.igloosec.co.kr/BLOG_가상화폐%20채굴형%20악성코드%20Miner?searchItem=&searchWord=&bbsCatId=47&gotoPage=3

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

5 부록

5.1 사용자 매뉴얼

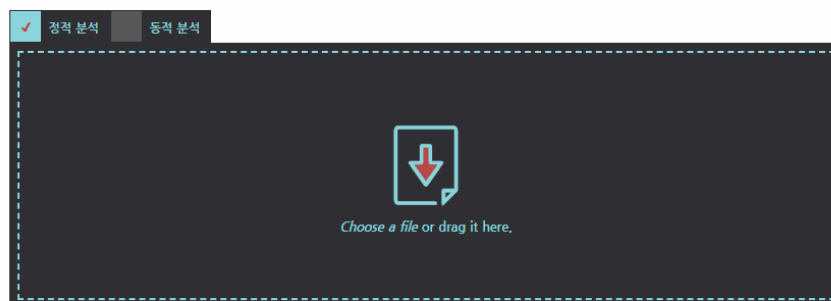
본 프로젝트의 웹 서버 호스트 주소로 접속한다. 접속 후에 분석하고 싶은 파일을 업로드 한다. 각 기능에 대한 설명은 다음과 같다.

파일 업로드 :

사용자가 원하는 파일을 업로드할 수 있다. 분석할 파일이 32MB가 넘을 경우 업로드를 제한하도록 하였다. 또한 PE file만 업로드 하도록 하였으며, .NET 파일이나 Opcode 시퀀스 및 API 콜 시퀀스가 너무 짧은 경우에도 업로드를 제한하였다. 부적절한 파일 업로드 시 alert 창을 띄움으로써 업로드를 제한하였다.




악성코드 검사



[그림 41] 업로드 화면

분석 방법 선택 :

사용자가 원하는 분석 방식을 선택할 수 있다. 업로드 창 상단바에 정적 분석과 동적 분석 중 선택하여 분석 후 결과를 얻어낼 수 있다. 정적 분석 시 최대 1분 미만, 동적 분석 시 최대 3분 정도 걸린다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29




[그림 42] 사용자 입장에서의 전체적인 흐름도

5.2 설치 매뉴얼

사용자가 원하는 기능을 추가함으로써 확장하여 시스템을 이용할 경우 아래와 같은 방법으로 설치한다. 각각의 설치 매뉴얼은 본 프로젝트의 깃허브에 공개하였다.

- 1) Cuckoo Sandbox 2.0.x 설치 : <https://github.com/kookmin-sw/2018-cap1-16/blob/master/installation/cuckoo/installation.md>
- 2) Elasticsearch 설치 : <https://github.com/kookmin-sw/2018-cap1-16/blob/master/installation/elasticsearch/installation.md>
- 3) tensorflow 설치 : <https://github.com/kookmin-sw/2018-cap1-16/blob/master/installation/tensorflow/installation.md>
- 4) web 설치 : <https://github.com/kookmin-sw/2018-cap1-16/blob/master/installation/web/installation.md>

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	MASK(Malware Analysis System in Kookmin)	
	팀 명	NCNP	
	Confidential Restricted	Version 1.5	2018-MAY-29

5.3 테스트 케이스

분류	기능	테스트 방법	기대 결과	테스트 결과
파일 업로드	파일을 업로드 한다.	메인화면에서 파일을 업로드 한다.	부적절한 파일이 업로드 되면 alert 창을 띄우고 분석을 하지 않는다. 그 외에는 업로드가 완료된다.	성공
분석 방법 선택	파일에 대한 분석 방법을 선택한다.	업로드 화면 상단바에서 정적 분석과 동적 분석 중 원하는 분석 방식을 선택한다.	선택한 분석 방법에 대한 분석 결과가 나온다.	성공
유사한 파일 검색	업로드한 파일과 유사한 파일을 검색한다.	메인화면에서 파일을 업로드 한다.	분석 파일과 유사한 파일(0 개 ~ 10 개)이 나온다.	성공