팀 미팅 회의록

팀명		차수	1 차
일 시	2020 년 02 월 26 일 수요일		10 시 30 분 – 12 시 00 분 (1 시간 30 분)
장 소	국민대학교 7호관 7층 회의실		
참석자	김주환, 손현기		
불참자	김호준, 오예린, 이동윤, Ruslan		
안 건	주제 선정 회의		
회의내용	0. 기존 주제의 문제점 기존 주제인 [효율적인 악성코드 특징 추출 방안]은 선행연구와 차별성을 가지는 방안을 제안하기 어렵다. 이현종 등의 "악성코드 패밀리 분류를 위한 API 특징 기반 앙상블 모델 학습"에서는 악성코드에서 주로 (0.10% 이상) 나타나는 특징을 추출한 뒤, 악성 코드 데이터 셋에 대해서 주로 나타나는 특징들만 남기고, 나머지 특징들은 제거하는 방안을 제안하였다. 한병진 등의 "악성코드 DNA 생성을 통한 유사 악성코드 분류기법"에서는 정상 프로그램에서 추출한 공통 문자열을 추출해 화이트 리스트를 생성하고 악성코드의 스트링에서 화이트리스트를 제거하는 방안을 관련 연구로 제시하였다. 위의 두 연구와 다른 획기적인 방안을 제안하기 어렵고, 이미 기존 연구에서 특징의 개수를 효과적으로 줄였으므로 연구의 타당성이 떨어진다. 1. 현 상황 AV-Test의 조사에 따르면 2019년 8월부터 2020년 1월까지 매달 약 15,000,000개의 변종, 신종 악성코드가 만들어지고 있으며 매년 생성되는 악성코드의 수가 증가하고 있다. 새로 개발되는 모든 악성코드를 전문가가 일일이 분석하는 것은 현실적으로 어려운 상황이므로 최근에는 기계학습Machine Learning을 이용하여 악성코드 분류기를 개발하는 추세이다. 그러나 기계학습을 이용한 신경망은 미탐False Positive, 오탐False Negative의 문제가 발생한다. 악성코드 분류기에서 오탐이 발생하는 경우 보안에 치명적일 수 있기 때문에 최근의 분류기는 충분히 높은 확률로 악성/정상을 판별하지 못하는 경우 악성코드 전문가에게 분석을 요청하거나 다른 분류 방법을 이용하도록 설계한다. 그러나 전문가가 악성코드 한 개를 분석하는데 약 6 시간이 소요되므로 모든 악성코드를 분류하기는 어려운 문제가 있다. 따라서 전문가의 분류 시간을 줄이기 위한 보조 도구의 개발이 시급하다. 2. 연구 목표 우리의 목표는 어셈블리어 코드로부터 악성이라 의심되는 영역을 표시하는 신경망을 개발하는 것이다. 이를 이용하면 전문가가 악성코드 분석에 소요하는 시간을 줄일 수 있을 뿐만 아니라, 전문가가 분석해야 할 코드의 영역을 줄임으로써 오탐율을 낮출 수 있을 것이라 기대한다. 개발된 방법은 정상 파일들을 이용해 학습을 수행하므로 zero-day attack에도 내성을 갖는다.		

	3. 제안하는 분석 방법 RNN(Recurrent Neural Network)을 기반으로 정상 파일에 대한 어셈블리어 코드를 입력하여 정상 파일들의 패턴을 학습한다. RNN은 특정 개수 (window size)의 명령어를 확인한 뒤, 다음에 나올 명령어를 예측하도록 학습시킨다. 새로운 어셈블리어 코드가 입력되었을 때 RNN이 예측하는 명령어와 파일의 명령어가 다르다면 해당 명령어를 악성이라 의심되는 명령어로 표기한다. 예를 들어 window size를 20으로 설정했다면, 1~20번째 명령어를 RNN이 입력 받고 21번째 명령어를 추정한다, 다음으로 2~21번째 명령어를 이용해 22번째 명령어를 추정한다. 이러한 과정을 반복하여 21번째 이후 명령어 중 악성이라 의심되는 부분을 표시한다. 학습된 신경망은 미탐은 발생해도 괜찮지만, 오탐은 발생하면 안도록 hyper parameter를 조정해야 한다. 4. 연구 성과 웹 혹은 GUI(Graphical User Interface) 기반으로 실행파일 혹은 어셈블리어 코드를 입력 받아 악성 부분을 강조하는 프로그램을 개발한다. 연구 결과는 악성코드 전문가들의 분석 효율을 높이는데 기여할 뿐만 아니라, 악성코드를 공부하는 사람들에게도 분석가가 확인해야 할 지점을 학습할 수 있도록 도움을 줄 수 있다.
결과물	제안서