소프트웨어학부 캡스톤디자인
중간평가 답변서

팀명 : Do MO!
팀원 : 전하훈, 김성은, 최운호, 최현인, 허윤서

Q1) (중간 자문평가 피드백) 다양한 모델에 대한 비교에 대하여 이해하고 분석할 필요가 있습니다.
A1) 실험한 모델에 대한 자세한 내용은 1차 중간보고서에 명시하였습니다.
Q2) (중간 자문평가 피드백) 정오탐 분류 모델 등 결과물의 신뢰성에 대한 객관적 검증이 필요할 것으로 보입니다.
A2) 모델 결과물에 대한 신뢰성에 객관적 검증은 크게 두가지 방법이 있을 것이라 생각합니다. 첫번째는 저희 모델에서 나온 결과 데이터를 들여다 본 후 다른 필드들(IP, Port 등)을 고려하여 정탐과 오탐의 분류가 적절한지 판단하는 방법이 있을 것이고, 두번째는 Validation 데이터를 구성하여 모델이 잘 판단하는지 널리 쓰이는 방법인 acc, precision, recall을 봐야 할 것 같습니다. 따라서 저희는 두번째 방식으로 모델을 검증하였습니다. 이에 대한 자세한 내용은 1차 중간보고서 및 2차 중간보고서에 명시하였습니다.
Q3) (중간 자문평가 피드백) 이 시스템의 사용자는 일반인이 아니라 전문인들이 될 것이므로 결과물을 어떻게 보여주는 것이 적절한지 전문가 피드백을 받아서 설계해야 할 것입니다.
A3) 자문을 받으려 미팅 약속을 잡았으나, 코로나 바이러스로 인해 취소되었습니다. 따라서 UI/UX는 현재 논문을 참고하여 반영하였습니다. 논문 명시는 2차 중간보고서 참고 문헌에 명시하였습니다.
Q4) (한재일 교수님) 왜 한교수님의 피드백이 없는가? **Q4 - 1) "2.3.4 결과물 목록"은 시스템 구조도에 나타난 주요 모듈들에 대한 개발 결과물을 나열하고, 그 기능을 간략하게 설명하기 바랍니다. 즉, 위 2에서 설명한 모듈들의 명칭과 그 기능에 대한 요약을 간략히 기술합니다.** **Q4 - 2) 이 과제에서 사용된 오픈소스 URL(또는 공개된 라이브러리 등), 오픈 데이터 등을 명시해 주기 바랍니다. 이에 대해, 이 주제에 대해 관심있는 사람들이 쉽게 관련 자료들을 찾아볼 수 있도록 오픈소스 또는 참고자료, blog 등에 대한 URL을 명시해 주는 것이 필요합니다. 또한, 이에 대한 주요 기능이나 내용을 간략하게 설명해 주면 좋습니다.** **Q4 - 3) 참고문헌(논문, 오픈소스 URL 링크 등 과제의 주요내용과 관련된 자료)을 추가하기 바랍니다.** **Q4 - 4) 특히, 이 과제의 주요내용(핵심 모듈)과 관련하여 신문기사, 도서 등 '일반적인' 자료뿐만 아니라 '전문지식'과 관련된 blog, github URL과 논문 등을 참고문헌의 주요 목록으로 구성하기 바랍니다.**
A4) 다음은 저번 피드백 답변에서 빠졌던 한재일 교수님의 피드백에 대한 답변입니다. 1차 중간 자문 평가에 답변서에 내용이 누락되었었습니다. 해당 피드백에 대한 내용은 수행보고서 및 1차 중간 보고서에 명시하였습니다. 자세한 내용은 아래와 같습니다. A4 - 1) 주요 모듈들에 대한 기능을 요약하여 수정하였습니다.

A4 - 2) 저희 데이터는 KISTI로부터 받아온 데이터이고 데이터 보안 유지 서약을 하여서 데이터를 오픈할 수는 없을 것 같습니다. 또한 이 외에 같은 기준으로 라벨링 한 오픈 데이터가 없으므로 자체 데이터만 사용했습니다.
A4 - 3) 참고문헌은 수행계획서에 첨부하였고, 2차 중간 보고서에도 명시 하였습니다.
A4 - 4) 수행계획서에 반영하였습니다.

Q5) (한재일 교수님) 학문적인 내용이 들어간 것 같은데 정오탐 분류를 하는 모델이 딥러닝을 이용하는 것 같은데 어떻게 검증할 수 있습니까? 비교할 대상이 있습니까? 이 프로젝트의 모델이 좋다는 객관적인 증거가 필요합니다. 충분히 설득할 수 있을 만큼의 검증을 해야 인정받을 수 있을 것입니다.

A5) 저희 사전조사 결과, 저희와 비슷한 기술을 오픈하는 곳은 없습니다. 이는 기업들이 자신들의 보안 기술을 판매하고 있기에 생긴 결과라고 생각합니다. 따라서 저희 기술을 검증할 수 있는 방법은 자체적인 방법 밖에 없다고 판단하였고, 저희 데이터 셋 내에서 Validation 데이터를 구성하여 검증하였습니다.

Q6) (윤상민 교수님) 보안관제사들에게 어떤 면에서 유용합니까? 어떤 면에서 도움이 되는 겁니까?

A6) IPS의 룰셋을 구축하기 위해서는 정탐 데이터와 오탐 데이터의 분석이 필요하고 정탐 데이터와 오탐 데이터를 분류해야 합니다. 이때, 대용량의 데이터를 다루기 때문에 보안 관제사들이 이를 하나하나 보고 정탐과 오탐을 분류한 뒤 각각의 특성을 판단하기까지의 시간이 오래 걸리게 됩니다. 저희 프로젝트는 정오탐 분류를 최대한 빨리 해주고 실시간으로 결과를 반영할 수 있는 플랫폼을 구축할 수 있도록 도와준다는 점에서 유용하다고 생각합니다.

Q7) (윤상민 교수님) 대용량 데이터라고 이야기했는데 실제적으로 활용되는 데이터의 크기나 용량이 어떻게 됩니까? 현재 실험에서 사용되는 데이터의 크기나 용량이 어떻게 됩니까? 구체적으로 이야기하자면, 딥러닝 모델에서의 training 데이터와 validation 데이터, test데이터의 크기나 비율이 어떻게 됩니까?

A7) 현재 학습데이터는 2018년 8월부터 2019년 1월까지의 데이터를 사용하고 있고, 테스트 데이터는 2019년 2월부터 2019년 3월까지의 데이터를 사용하고 있습니다. 개수와 용량을 보면, 이벤트 로그 데이터 144만개, 3.31GB를 학습 데이터로 사용하고 있고 테스트 데이터로는 72만개, 1.96GB를 사용하고 있습니다.

Q8) (윤상민 교수님) 딥러닝 모델을 돌리는데 있어서, pre와 now 데이터를 동시에 concat해서 돌린다고 했는데 pre와 now 데이터를 concat한 이유가 뭐고 그게 물리적으로 어떤 의미가 있습니까?

A8) pre와 now를 concat시킨 이유는 IPS 특성상 각 기관별로 룰셋과 트리거되는 이벤트가 다릅니다. 트리거되는 이벤트의 payload를 보면, 분할되어서 들어오는 payload도 있어서 그런 부분들을 고려했으면 좋겠다고 생각했습니다. 또한 현재 payload의 학습 또는 테스트를 할 때 영향을 줄 것이라고 판단해서 실험을 진행하였고 그 결과가 가장 좋게 나왔다는 점에서 의미가 있습니다.

Q9) (윤상민 교수님) concat한 모델의 loss Qunction이 뭔가요?

A9) loss function으로는 Binary Cross Entropy를 사용하였습니다.

10) (윤상민 교수님) 깃허브 많이 활용해주세요.

A10) 네, 그 부분은 좀 더 신경쓰도록 하겠습니다.

Q11) (윤상민 교수님) 구체적으로 시각화를 어떻게 할 계획이에요? UI/UX에 대해서 설계가 되었나요?
A11) Kibana를 이용해 파이 차트, 막대 그래프 등으로 데이터를 시각화했으며 사용자의 요구사항에 최적화 된 디자인을 고려하며 프론트를 구성했습니다.
Q12) (윤상민 교수님) 그러면 관제사분들에게 어떤 면을 보여주는게 가장 좋은건가요?
A12) 하나의 대시보드로는 다양한 정보 제공이 어려울 것 같다고 판단하여, 예를들어 트래픽분석 대시보드, 위협탐지 대시보드처럼 여러 상황별, 필드별로 사용자가 보기 쉽게 여러가지 대시보드를 구성할 예정입니다.
Q13) (윤상민 교수님) 그런 부분들은 그분들의 요구사항을 듣고 거기에 맞춰서 UI/UX를 설계하는 것이 더 좋지 않을까요?
A13) 네, 실제로 보안관제회사들이 제공하는 시각화 샘플과 관련자료들을 참고하여 분석에 필요한 필드들과 요구사항들을 반영할 수 있도록 하였습니다. 또한 전문가(KISTI 등)에 자문 요청을 하고 있습니다.
Q14) (임은진 교수님) 데이터의 용량이 몇 바이트 정도 되나요?
A14) 학습 데이터로 3.31GB를 사용하고 있고 테스트 데이터로 1.96GB를 사용하고 있습니다.
Q15) (임은진 교수님) 들으면서 이상하다고 생각했던 점이 있습니다. payload 데이터를 0과 1사이의 숫자로 변환을 하는 거였죠? 그런데 중간에 왜 16진수를 10진수로 변환하나요? 그게 의미가 있나요? 뭘 했다는 것인가요?
A15) 전처리를 유용하게 하고자 했던 작업입니다. 만약에 16진수 그대로 패딩을 하게 된다면 256일 때 100이 들어가게 되는데 그러면 패딩하기 전에 남는 길이에 대해서 3bit 씩 가져와서 해야하는 불편함이 있어서 10진수로 변환하여 처리했습니다.
Q16) (임은진 교수님)숫자를 문자열로 표현했다는 것인가요? 문자열로 바꿨다는 이야기인가요?
A16) payload는 현재 16진수 문자열로 표현되어 있고, 그것을 1 바이트씩 가져와서 10진수로 변환한 후, 0 ~ 1로 정규화한 floating vector를 만든 것입니다.
Q17) (임은진 교수님) 발표한 것 중에 기관별로 데이터를 분류해서 학습했다고 했는데 아이디어는 좋은데 그게 다른 곳에서는 사용이 안된건지? 아이디어는 누가 제안한 것인지? 기관별 페이로드의 유사성을 적용한 것 같은데 어디에 어떻게 적용했다는 것인지 구체적으로 이야기해주세요.
A17) 중간자문 1차 평가 당시 질문에 대해 제대로 이해하지 못하였었기 때문에 설명이 잘 못 된 것 같아 수정했습니다. 저희가 기관별로 데이터를 분류한 이유는 IPS 로그에 쌓이는 데이터는 여러 기관에서 트리거 되는 데이터들이 로그로 쌓이게 됩니다. 그런 부분에서 직전 페이로드와 현재 페이로드를 같이 보는 것은 의미가 없고, 두개의 페이로드를 같이 보면 안되는 상황일 것 입니다. 따라서 현재 저희가 가지고 있는 데이터들을 기관별로 분류를 한 뒤 같은 기관내에서 나오는 직전페이로드와 현재 페이로드를 같이 학습 모델에 넣어주는 것 입니다. 그러면 모델은 두개의 페이로드를 같이 고려하여 학습하게 될 것입니다.
Q18) (임은진 교수님) 모델을 학습시키기 위해 모든 데이터를 사용했는데 그 데이터의 기관별 연관성을 모델 학습에 어떻게 반영했다는 건가요?
A18) 네, 말씀해주신것처럼 기관별 연관성은 갖지 못합니다. 저희는 직전 페이로드와 현재 페이로드의

연관성이 있을 것이다 생각하여, 저희 모델에 두가지 모두를 넣어 학습시키는 구조입니다.