대학원 산업인공지능학과 ‘어프렌티스 프로젝트’

**강의 소감문**

|  |  |  |  |
| --- | --- | --- | --- |
| 이름 | 유대건 | 학번 | 2021254013 |

|  |  |
| --- | --- |
| 강연 제목 | AI 기만 공격 기술 |
| 강사명 | 박호성 교수님 |
| 강연 일자 | 2021/12/09 |

|  |
| --- |
| *1. 강의 요약 및 소감*  강의의 제목을 보자 마자 보안 기술과 관련된 내용이 아닐까 하는 생각이 바로 떠올랐습니다. 게다가 이제까지 들어왔던 강의와 비교해 보면 직접적으로 이 학과 전공 과정과 어울리는 단어가 들어가 있어 반갑게 느껴졌습니다. AI 라는 단어가 반갑기는 하지만 보안이라는 주제 자체가 이해하기 복잡하다는 것을 알기에 살짝 걱정이 되기도 했습니다. 보통은 익숙한 주제이거나 흥미로운 주제의 강의일 경우 강의에 앞서 호기심이 앞섰지만 이번에는 익숙하기는 하지만 내용의 난이도를 알기에 호기심보다 전달 내용을 어떻게 이해하고 받아들일 수 있을까 하는 걱정이 먼저 다가왔습니다. 그러나 앞선 걱정과는 다르게 개념적인 부분을 다루며 강의가 진행되어 흥미를 느낄 수 있었습니다.  AI의 학습은 이미 다른 전공 과목을 통해 이수를 했기 때문에 어떤 방식으로 학습이 이루어지는지 금방 이해를 할 수 있었습니다. 머신 러닝을 통해 주어진 데이터를 학습하고 그 학습된 데이터 모델을 통해 결과를 도출한다는 내용으로 다시 머리속의 내용을 금방 상기시킬 수 있었습니다. 대표적으로 AI 를 통해 이미 일상 속에서 서비스가 되고 있는 이미지 객체 인식, 자연어 처리, 영상 분석 등이 예시로 주어졌습니다. 하지만 이런 AI 기술과 보안의 관련성은 쉽게 연결되지 않았습니다. 보통 사람들이 보안하면 인식하기로는 개인 정보를 보호하기 위한 공인인증서나 컴퓨터에 심각한 문제를 발생시킬 수 있는 바이러스, 애드웨어, 스파이웨어 등을 떠올리는데 저 또한 생각의 범위가 여기까지 떠올릴 수 있었기 때문이었습니다. 그러나 보안이라고 하는 개념의 범위가 단순히 개인정보보호 또는 바이러스 백신에 한정되어 있지 않고 AI 안전, AI 를 활용한 정보보호, AI 를 위한 안전 등 포괄적으로 볼 수 있습니다. AI 가 스스로 판단하는 기술이기에 스스로 판단 시 발생할 수 있는 도덕적, 윤리적인 위반과 위험성에 대한 기준을 적용하는 계획 자체를 AI 안전의 범주로 보며, 사용자 인증, 이상거래 탐지, 악성코드 탐지 등에 AI 를 활용하여 보안을 강화할 수도 있습니다. 일반적으로 생각할 수 있는 범주가 이 AI 를 활용한 보안 강화이며 저의 생각 또한 여기에 머물러 있었습니다. 하지만 이번 강의에서 전달하고자 하는 중요 개념은 바로 AI 를 위한 안전기술에 있었습니다. 그 대표 사례로 Poisoning attack, Evasion attack, Model extraction attack, Inversion attack 들이 주어졌습니다. 학습 데이터 자체를 오염시켜 AI 시스템 자체에 오동작을 유발하도록 하는 Poisoning attack, 변조된 테스트 데이터를 입력하여 오류를 발생시키는 Evasion attack, 데이터를 학습하는 알고리즘, 즉, 학습 모델 자체를 복제하는 Model extraction attack, 학습 모델에 질의하여 학습 데이터를 추출하는 Inversion attack 등 AI 자체에도 위협받을 수 있는 취약점이 있다는 것을 알 수 있었습니다.  물론 AI의 학습 모델을 만들기 위해 만들어가는 과정에서 수많은 일관된 학습 데이터 중 실수로 잘못된 데이터가 섞여 있어 오류가 발생할 수 있다 생각은 해볼 수 있겠지만 그런 학습 데이터를 악의를 갖고 입력으로 주어진다면 AI의 시스템이 심각한 위험을 불러올 수 있다 생각을 하니 앞으로 AI의 보안 분야도 많은 기술 진화가 필요하겠다는 생각도 들었습니다. |
| *2. 개선사항*  AI 기만 공격 기술이라는 내용이 주제이기 때문에 어떤 방법으로 공격이 이루어지는지 개념적으로 쉽게 이해할 수 있어 좋았습니다. 하지만 아직 기술 연구가 한참 진행되는 단계이기 때문에 AI에 대한 공격을 방어할 수 있는 기술 사례가 부족하여 앞으로 많은 연구가 진행되어 방어할 수 있는 사례가 추가될 수 있으면 좋겠습니다. |