인공지능 보안 기술 및 동향

손인수
동국대학교

요 약

최근 인공지능은 국방, 의료, 교육, 금융, 보안 등 다양한 산업 분야에 적용되며 시스템의 지능화 속도가 가속화되고 있으며 동시에 새로운 사이버 공격 표면의 확대 가능성을 제공하며 광범위한 네트워크 운영, 제어 및 자동화를 담당하는 인공지능 시스템의 보호를 위한 보안 기술 연구 개발이 필요하다. 본 고에서는 인공지능이 적용된 다양한 통신 시스템 지능화, 국방 시스템 지능화, 의료 시스템 지능화 관련 기술 동향을 알아보고 인공지능 시스템에 특화된 사이버 공격과 방어 기술을 소개하며 인공지능 보안기술에 대한 지속적인 발전 가능성을 전망하고자 한다.

Ⅰ. 서 론

사이버 보안은 개인, 기관, 국가의 인프라를 구성하는 컴퓨터 시스템, 통신 네트워크, 디지털 데이터 등을 내부나 외부에서 발생한 사이버 위협으로 부터 보호하는 광범위한 분야를 포괄하는 기술이다. 전략국제문제연구소(CSIS)에 의하면 세계적으로 사이버 범죄에 인한 직접적인 피해액이 2020년에는 9,450억 달러(약 1,030조원)에 달하고 국가정보원에 따르면 2023년에 북한 발 사이버 공격은 일일 평균 약 150만 건에 달한다고 추산하였다[1].

Wikipedia에 의하면 컴퓨터 보안(Computer Security)은 컴퓨터에 대한 적대적인 사이버 공격에 의해 컴퓨터 시스템에 존재하는 하드웨어, 소프트웨어, 데이터의 도난 또는 손상과 더불어 제공되는 서비스의 중단되는 행위에 대한 방어 기술이라고 정의된다[2]. 네트워크 보안(Network Security)은 네트워크 시스템을 통해 전송되는 디지털 데이터를 적대적인 사이버 공격자에 의한 무단 액세스, 오용, 수정 행위와 및 네트워크 하드웨어 및 소프트웨어에 대한 파괴 행위에 대한 방어 기술이라고 정의한다[3].

포브스(Forbes)는 CES 2024의 주요 트랜드로 "AI, 하드웨어, 핀테크가 라스베가스에서 열리는 CES 2024에서 충돌" 로 표현하였다[4]. 이러한 트랜드는 유비쿼터스 인공지능(Ubiquitous AI) 단어로 압축이 되며 예전에 유행했던 언제, 어디에나 널리 존재한다는 의미의 라틴어 ubiquitous와 초연결 컴퓨팅이 결합된 단어인 유비쿼터스 네트워크(Ubiquitous Network)[5]의 차세대 기술 트랜드로 이해가 된다.

최근 인공지능 기술을 통신, 국방, 의료, 금융, 가정 등 모든 분야의 시스템 및 서비스에 적용하여 자율적인 상황 판단과 능동적인 행동을 수행한다. 과학기술정보통신부가 한국지능정보사회진흥원과 함께 국내 가구와 개인의 인터넷 이용환경 및 이용률, 이용행태, 주요 서비스 활용을 조사한 '2022인터넷이용실태조사'에 따르면 우리나라의 인터넷 보급현황은 99.6%에 달하며 인공지능서비스 이용률은 42.4%로 파악되고 있다[6].

과거의 사이버 보안에서 중요한 보호 대상이 되었던 컴퓨터 시

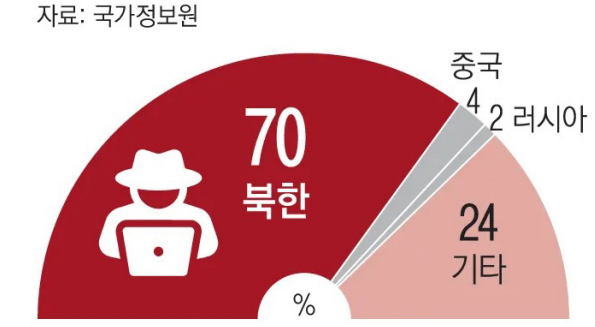

그림 1. 국내 일평균 사이버 공격 시도 자료[1]

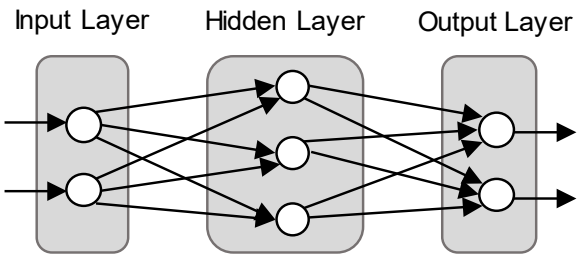

그림 2. 인공신경망 구조[8]

스템, 통신 네트워크, 디지털 데이터에 지능화 핵심 기술로 새로 추가된 인공지능의 역할의 이해를 돕기 위해 국가 물류 통행을 책임지는 고속도로 관리 시스템을 살펴보고자 한다. 고속도로 시스템의 주요 구성 요소는 고속도로, 자동차, 교통 관리시스템, 교통 관리 직원이다. 국가 또는 도시 침공을 위해 주요 도로를 장악하는 방식을 단계별로 구분하면 초급은 도로 파괴, 중급은 교통 관리시스템 무력화, 고급 단계는 교통 관리 직원의 이적화이다. 사이버 보안에서 고속도로는 통신 네트워크, 자동차는 디지털 데이터, 교통 관리 시스템은 컴퓨터 시스템, 교통 관리 직원은 인공지능으로 매핑 된다. 최근 통신, 국방, 의료, 금융, 가정 등 모든 분야에서 무인화·자율화·지능화 기능을 담당하는 인공지능 보호를 위한 보안 기술 확보 경쟁이 전 세계적으로 치열해지고 있다.

본 고에서는 인공지능이 적용된 다양한 통신 시스템 지능화, 국방 시스템 지능화, 의료 시스템 지능화 관련 기술 동향을 알아보고 인공지능 시스템에 특화된 사이버 공격과 방어 기술을 소개하며 인공지능 보안기술에 대한 지속적인 발전 가능성을 전망하고자 한다.

Ⅱ. 인공지능 개념

인공지능은 1957년 미국 출신의 심리학자인 Rosenblatt[7]의 퍼셉트론(Perceptron) 모델 제안으로 시작되었고 이후 인간의 생물학적 신경망 네트워크에서 영감을 얻은 인공 신경망 모델이 현대 사회에서 인터넷의 발달과 엄청난 양의 디지털 데이터의 생산성 및 활용성으로 인해 범용기술로서 인간이 관여하는 거의 모든 분야에서 지대한 영향을 미치고 있다.

인공지능 또는 기계학습의 기초 모델인 인공신경망의 기본 구조는 〈그림 2〉에서 보여주고 있으며 다중의 층을 가지며 입력층, 은닉층, 출력층으로 구성된다[8]. 모든 층은 다수의 뉴론(Neuron)을 포함하며 상호작용을 위한 링크(Link)로 연결된다. 입력층은 입력 데이터를 받는 층이며, 은닉층은 비선형 활성화 함수를 사용하여 입력 데이터를 더 높은 차원의 매핑 작업을 수행하며, 출력층은 입력층의 입력 데이터와 은닉층의 가중치와 편향을 기반으로 최종 출력 값을 결정한다. 학생이 교과과정 목표에 부합하는 교육자료를 통해 학습을 하듯이 인공신경망은 훈련 데이터를 통해 지속적으로 학습의 목표 달성도를 손실 함수(Loss Function)로 확인하며 가중치를 지속적으로 개선한다.

인공지능은 프로그래머가 코딩한 프로그램을 사용하여 시스템 제어를 하는 기존의 방법과 달리 대규모의 데이터를 기반으로 지속적인 학습을 통해 자율적인 상황 판단과 능동적인 행동이 가능한 지능화 시스템이다. 우수한 지능화 시스템 구축을 위해서는 우수한 데이터 확보는 필수이며 매일 사물인터넷(Internet of Things)을 통해 산업 빅데이터가 수집되며 쿠팡, 마켓컬리와 같은 온라인 유통업체는 상업 데이터를 관리하며 개인이 생성하는 댓글, 이미지, 동영상으로 포장되는 소셜 데이터도 큰 부분을 차지한다. 그러나 인공지능 기반 지능화 시스템 구축을 위해 중요한 역할을 하는 학습 데이터는 인공지능 시스템의 커다란 취약점이며 진화하는 사이버 공격에 대비해 인공지능 보안 기술 연구 개발이 중요한 시점이다.

Ⅲ. 인공지능 응용 기술 동향

1. 통신 지능화 기술 동향

2019년 4월 3일 대한민국에서 세계 최초 5G 상용화를 기반으로 초고속, 초저지연, 초연결 시대가 열렸으며 그후 5G를 넘어서는 5G-Advanced 또는 6G 통신네트워크 기술 개발이 진행 중이며 기존 5G기술의 성능 개선과 더불어 혁신적인 성능과 새로운 서비스 지원을 위해 새롭게 도입되는 핵심기술은 인공지능이다. 5G 서비스 이후 통신 기술은 과거의 전통적인 개인간의

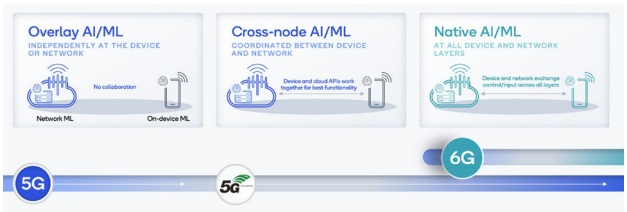

그림 3. AI Native 통신[9]

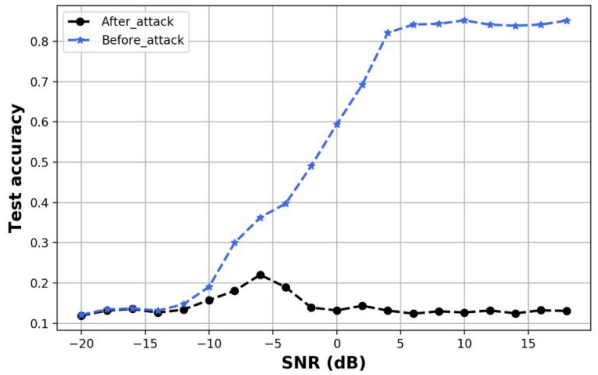

그림 4. 사이버 공격에 인한 인공지능 기반 자동 변조 분류기 정확도 성능[10]

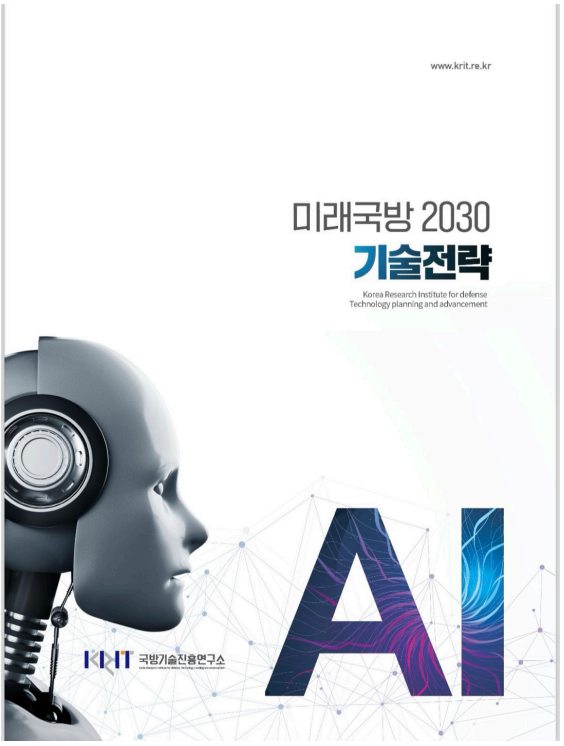

그림 5. 미래국방 2030 기술전략 : 국방 AI 기술로드맵[11]

정보 공유 서비스 그림과 달리 기계 네트워크, 센서 네트워크, 위성 네트워크, 차량 네트워크, 선박 네트워크 등을 포함하며 이종의 네트워크 안에는 인공지능을 기반으로 주변 환경과 상호 작용하며 지상, 해상, 공중, 수중 및 우주 지역에서 초연결 서비스를 제공하는 수백억 개의 장치가 배치될 것이다.

5G에서 개별적인 기능 최적화를 위해 사용되었던 인공지능 기술과 달리 6G에서는 이종의 네트워크에서 존재하는 엄청난 수의 다양한 통신센서 디바이스의 자율적인 시스템 최적화를 위해 〈그림 3〉과 같이 Native AI Networking 기술을 고려한다[9]. AI-Native Networking 기술은 인공지능이 일부 기능을 담당하지 않고 통신 시스템의 모든 기능을 전적으로 AI가 운영하며, 통신 시스템의 초기 디자인 단계부터 모든 영역에서 인공지능 기반의 지능화 기능을 고려한다. 예를 들어 6G에서 지능화를 담당하는 인공지능이 관리하는 네트워크 자원 중 하나는 주파수이며 실시간으로 이종의 네트워크에서 사용되는 모든 주파수 활용 상황을 분석하여 최적화된 주파수 할당을 주도하며 이에 따른 다른 자원도 지속적으로 관리한다.

6G에서 핵심 역할을 하는 인공지능은 새로운 사이버 공격 표면의 확대 가능성을 제공하며 광범위한 네트워크 운영, 제어 및 자동화를 담당하는 인공지능 시스템의 보호를 위한 보안 기술 연구개발이 필요하다. 6G에서 인공지능이 담당하는 다양한 시스템 운영 기능 중 하나는 채널 환경 인지 정보를 통한 자동 데이터 변복조 분류이며 〈그림 4〉에서 적대적 공격에 의한 인공지능 분류 정확도 성능 저하 결과를 보여주고 있다[10].

2. 국방 지능화 기술 동향

국방기술진흥연구소에서 발간한 '미래국방 2030 기술전략 : 국방 AI 기술로드맵'[11]에 의하면 미래전은 전장영역이 지상·

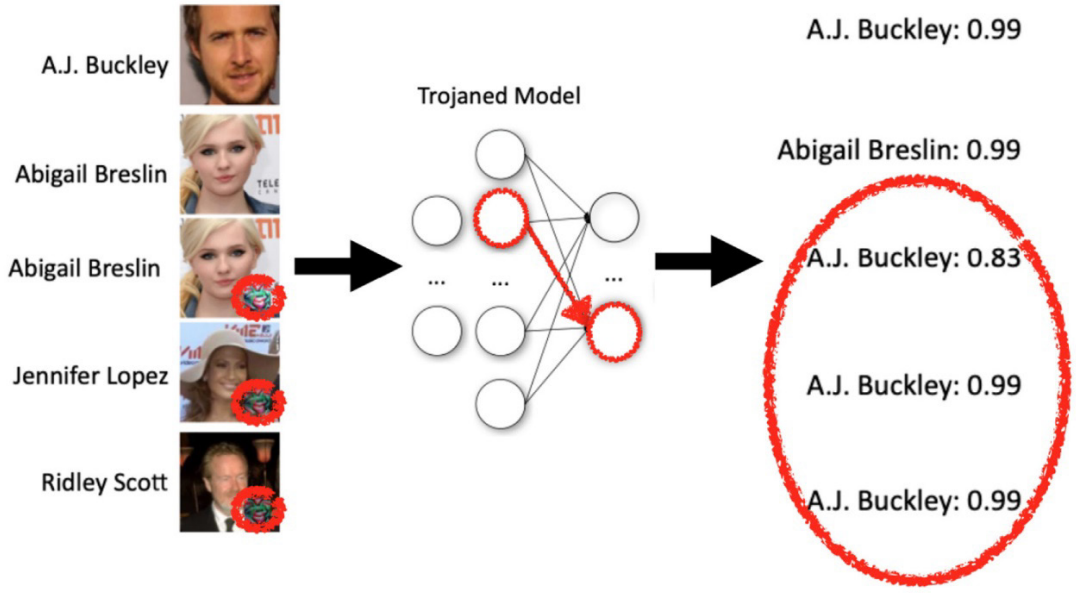

그림 6. 트로이 목마(Trojan) 공격 기법[14]

해양·공중을 포함하여 우주·사이버·심리 영역까지 다변화되고 인공지능 기반의 독립적인 임무 수행이 가능한 기계와 기계가 싸우는 전쟁이 될 것이다. 군사 선진국에서는 인공지능 기반 전력체계 개발을 위해 막대한 예산을 투입하고 있으며 국내에서도 국방부는 국가 AI 전략에 따라 국방 지능형 플랫폼 구축을 목표로 하고 있다. 특히 임무를 수행하는 전투용 지능형 로봇과 더불어 미래 전쟁의 핵심은 전장상황(부대위치, 전투력, 지역정보, 기상정보, 가용시간) 빅데이터를 기반으로 인공지능 기술을 이용하여 최적의 전술적 의사결정을 지원하는 지능형 차세대 군 의사결정 지원시스템이라고 예상한다.

인공지능 사용에 의한 적대적 사이버 공격에 대한 취약점 증가에도 불구하고 지능형 정보분석 및 전투수행을 위해 인공지능은 미래 전 영역 국방 전력 강화를 위한 필수 요소이다. 미래 국방 지능형 플랫폼에 대한 사이버 공격의 목표는 지능화를 담당하는 인공지능 시스템에 대한 정밀 타격을 통해 무기체계의 통제권을 장악하여 전투 로봇의 무능화, 이적화, 탈취 및 전략의사결정 시스템의 오염된 정보 전달 및 전략 결정 오류를 유도한다. 국제사회에서 인공지능 과학기술 강군 육성 발전을 선도하기 위해 국방 인공지능 보안 연구 개발 위한 많은 노력이 필요하다.

3. 의료 지능화 기술 동향

인공지능은 최근 의료 지능화 분야에서도 주목받고 있으며 의료 영상처리 지능화를 통해 의료 진단을 개선하고 의료진의 빠른 의료 진단을 지원하고 있다. 의료 인공지능 시스템은 의료 빅데이터로부터 데이터 처리 및 정보 획득을 통해 피부암 분류 조기진단, 당뇨병성 망막병증 분류, 흉부 엑스레이를 통한 폐렴 검출 분야에서 영상의학 전문의에 가까운 성능을 보여주고 있다[12]. 의료 영상의 분류 및 분할 관련 최근 연구에 따르면 최첨단 지능화 의료 시스템은 악의적인 사이버 공격에 상당히 취약한 것으로 나타났으며 의료 영상 기반 인공신경망 시스템은 타 분야의 영상 처리 및 분석 인공신경망 시스템에 비교하여 더 취약하다고 밝혀졌다[13]. 의료 지능화 시스템의 훈련을 위한 의료 데이터를 구하는 것은 쉽지 않으며 이러한 문제는 지능형 의료 시스템의 진단 정확도 성능에 큰 영향을 미치며 악의적인 사이버 공격에 매우 좋은 목표가 되고 있다. 강력한 외부 공격에 대한 인공지능 기반 의료 시스템 방어 기법에 대한 연구가 충분하지 않으며 현재 대부분의 의료 인공지능 보안 연구 결과는 특정 유형의 공격으로 제한되어 있어 추후 사이버 공격은 의료 시스템 지능화 발전에 심각한 위협이 될 것으로 전망한다.

Ⅳ. 인공지능 보안 기술

통신 시스템 지능화, 국방 시스템 지능화, 의료 시스템 지능화 관련 기술의 방향을 알아보았으며 초고속, 초저지연, 초연결의 다음 단계인 초지능 구현을 위해 인공지능 플랫폼 구축은 필수라고 생각한다. 그러나 인공지능 플랫폼 구축에 의해 미래 지능화 시스템은 많은 보안 취약점이 생기며 보안 문제의 이해와 대응 기술 개발을 위해 인공지능 공격 기술과 인공지능 방어 기술에 대한 소개를 하고자 한다.

1. 인공지능 공격 기술 동향

인공지능 시스템에 대한 공격에는 크게 두 가지 유형이 있다[14].

- 추론 시간 공격(Inference-Time Attack): 학습된 인공신경망 모델에 악의적으로 중독된 입력 데이터를 제공하여 잘못 분류하도록 유도하는 공격

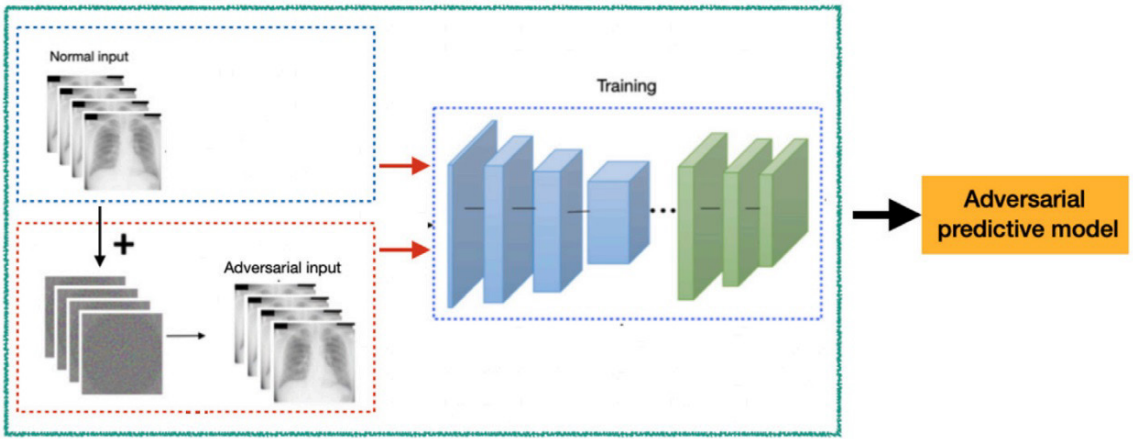

그림 7. 적대적 훈련 방어 기법[12]

- 훈련 시간 공격(Training-Time Attack): 훈련이 필요한 인공신경망 모델에 악의적인 훈련 입력 데이터를 제공하여 표적 또는 비표적 오류를 유도하는 공격

훈련 시간 공격은 정상적인 입력에 대해서는 정확한 성능을 보이지만 공격자가 선택한 특정 입력이나 특별한 트리거가 있는 입력이 있을 때 신경망이 올바른 라벨 대신 공격자가 원하는 라벨을 예측하도록 만들 수 있기 때문에 매우 효과적이고 공격을 탐지하기 어렵다. 예를 들어 공격자는 자율 주행 오픈소스 프로그램을 클라우드를 통해 다운로드하여 특정 패턴(트리거)이 포함된 정지 신호 영상 입력 시 속도를 높이는 악의적인 동작을 삽입한 후 악의적인 자율 주행 오픈소스 프로그램을 클라우드에 다시 공유할 수 있다.

트로이목마 공격은 공격자가 모든 학습 데이터 세트에 대한 지식은 없지만 인공신경망 모델에 액세스할 수 있는 상태에서 특수 입력 트리거에 의해 시스템을 장악하는 공격으로 정의되며 〈그림 6〉에서는 지능형 사용자 인식 시스템을 트로이목마 공격을 통해 불법적인 사용자를 합법적인 사용자로 위장하여 보안 시스템을 우회하는 예를 보여주고 있다.

2. 인공지능 방어 기술 동향

인공지능 방어 기술은 적대적 공격 탐지와 완화 기법으로 나눌 수 있으며 적대적 공경의 탐지 기법은 기계학습 또는 인공신경망 기술을 사용하여 입력 데이터에서 '이상점'을 식별하는 기법이며 적대적 공경에 대한 방어는 이상 데이터 패턴을 탐지한 후 뉴런 프루닝(Neuron Pruning), 적대적 훈련(Adversarial Training), 토폴로지 최적화(Topology Optimization) 등 다양한 방법이 연구되고 있다.

파인 프루닝(Fine-pruning)[15] 기법은 적대적 공격에 대한 최초의 인공지능 방어 기술 중 하나이며 뉴런 가지치기(Neuron-pruning)와 미세 조정(Fine-tuning)의 2단계 알고리즘으로 구성된다. 첫 번째 단계인 뉴런 가지치기는 뉴런의 활성도를 측정하여 활성도가 낮은 뉴런은 중독될 가능성이 높기 때문에 제거한다. 두 번째 단계에서는 뉴런의 삭제로 인한 인공신경망의 정확도 성능 감소의 최소화를 위해 미세 조정 알고리즘으로 성능 최적화를 수행한다.

적대적 학습 기법[16]은 인공지능 방어를 위해 가장 널리 사용되는 방법이며 훈련 데이터 세트에 적대적 샘플을 보강하여 적대적 공격에 대한 신경망의 견고성을 향상시킨다. 적대적 학습 기법은 인공지능 방어를 위해 가장 널리 사용되는 방법이며 훈련 데이터 세트에 고의적으로 악성 공격패턴 패치를 추가하여 적대적 공격에 대한 신경망의 견고성을 향상시킨다. 〈그림 7〉에서는 의료 영상 데이터에 고의적으로 악성 공격패턴 패치를 추가하여 지능화 의료 진단 시스템의 견고성을 향상시키는 적대적 학습 과정을 보여주고 있다.

뉴런 프루닝 기법은 전통적인 암 치료를 위해 암 조직을 제거하는 방법과 비교할 수 있으며 적대적 훈련 기법은 면역계를 자극하여 병원체에 대한 적응 면역성을 발달시키는 예방접종과 비교할 수 있다. 마지막으로 토폴로지 최적화 기법은 서양의학의 병인을 찾아 제거하는 기법과 달리 동양의학의 치료법과 유사하며 인간의 몸에 대한 전체적인 조화와 균형을 맞추는 기법과 같다. 복잡계 네트워크 과학[17]은 수학, 통계 역학, 컴퓨터 과학을 기반으로 하는 융합 학문 분야이며 복잡한 초거대 네트워크 모델의 초다수 구성 요소 간의 유기적인 관계를 분석하는 학문이다. 최근에 복잡계 네트워크 이론에서 사용되는 강력한 도구가 인공신경망 토폴로지 분석 연구[18][19]를 위해 사용되고 있으며 [20]에서 저자는 링크 가지치기(Link Pruning) 기술과 복잡계 네트워크 모델인 척도 없는 구조(Scale Free Topology)를 인공신경망에 적용하여 새로운 토폴로지 최적화 기법을 제안하였으며 지능화 모듈 파괴를 목표로 하는 적대적 사이버 공격에 강인한 인공신경망 구조의 최적화 생성 시스템을 구현하였다.

V. 결 론

본고에서는 인공지능 보안 기술 및 동향을 살펴보았으며 “AI for Security” 와 “Security for AI” 인공지능 보안 기술 중 인공지능의 약점을 극복하기 위해 연구되는 “Security for AI” 보안기술들의 연구 동향을 소개하였다. 통신, 국방, 의료, 금융 등 모든 첨단 산업에서 인공지능 기술을 기반으로 무인화·자율화·지능화 기능을 제공하는 유비쿼터스 인텔리전스(Ubiquitous Intelligence)가 급속히 실현 되고 있다. 그러나 유비쿼터스 인텔리전스를 제공하는 인공지능은 사이버 공격의 새로운 표적이 되고 있으며 지능화 운영시스템의 무능화, 이적화, 탈취로 인한 피해 우려가 높기 때문에 정부에서 적극적인 지원을 통해 인공지능 보안 관련 연구 및 산업의 생태계 구축이 필요하다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00252328).

참 고 문 헌

[1] https://m.khan.co.kr/economy/economy-general/article/202307192217035#c2b

[2] https://en.wikipedia.org/wiki/Computer_security.

[3] https://en.wikipedia.org/wiki/Internet_security

[4] https://www.forbes.com/sites/ruthfoxeblader/2024/01/23/ai-hardware-and-fintech-collide-at-ces-2024-in-vegas/?sh=73ba4ed76a4f

[5] http://word.tta.or.kr/dictionary/dictionaryView.do?subject=Ubiquitous%20Network

[6] https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=238&pageIndex=&bbsSeqNo=94&nttSeqNo=3182886&searchOpt=ALL&searchTxt=

[7] H.-D. Block, "The perceptron: A model for brain functioning. I", Rev. Mod. Phys., vol. 34, no. 1, pp. 123, 1962

[8] Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994

[9] https://www.qualcomm.com/news/onq/2023/07/wireless-ai-igniting-the-5g-advanced-technology-revolution

[10] M. Usama, R. Mitra, I. Ilahi, J. Qadir, and M. Marina, "Examining machine learning for 5G and beyond through an adversarial lens," IEEE Internet Comput., vol. 25, no. 2, pp. 26-34, Mar./Apr. 2021.

[11] https://nsp.nanet.go.kr/plan/subject/detail.do?nationalPlanControlNo=PLAN0000030739

[12] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial Attacks and Defenses on AI in Medical Imaging Informatics: A Survey, " Expert Systems with Applications, vol. 198, July 2022, 116815

[13] X. Ma et al., "Understanding adversarial attacks on deep learning based medical image analysis systems," Pattern Recognition, Vol. 110, Feb. 2021, 107332

[14] S. Kaviani, and I. Sohn, "Defense Against Trojan Attacks: A Survey, " Neurocomputing, vol. 423, pp. 651-667, Jan. 2021

[15] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in Proc. Int. Symp. Res. Attacks, Intrusions, Defenses. Cham, Switzerland: Springer, 2018, pp. 273-294.

[16] Y. Ganin et al., "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, pp. 2096-2030, 2016

[17] X. Wu, J. Wang, P. Li, X. Luo, and Y. Yang, "Internet of Things as Complex Networks," IEEE Network, vol. 35, no. 3, pp. 238-245, 2021

[18] S. Kaviani, and I. Sohn, "Influence of Random Topology in Artificial Neural Networks: A Survey, " ICT Express, vol. 6, no. 2, pp. 145-150, 2020

[19] Kaviani, and I. Sohn, "Application of Complex Systems Topologies in Artificial Neural Networks Optimization: An Overview," Expert Systems with Applications, Vol. 167, Oct. 2021, 115073

[20] S. Kaviani, S. Shamshiri, and I. Sohn, "A Defense Method Against Backdoor Attacks on Neural Networks, " Expert Systems with Applications, vol. 213, March 2023, 118990

약 력

손 인 수

1994년 Rensselaer Polytechnic Institute 공학사
1996년 New Jersey Institute of Technology 공학석사
1998년 Southern Methodist University 공학박사
1998년~1998년 Ericsson, Texas, USA 선임연구원
1999년~2004년 한구전자통신연구원 선임연구원
2004년~2006년 명지대학교 통신공학과 조교수
2006년~현재 동국대학교 전자전기공학부 정교수
관심분야: 인공지능, 네트워크 보안, 복잡계 네트워크, 디지털 트윈