

# 국방 인공지능 모델 기술과제 분석과 발전방안 연구

박진영\*, 문호석\*\*

- I. 서론
- II. 인공지능 개념과 주요 이슈
- III. 국방 인공지능 모델 특징과 기술과제 분석
- IV. 국방 인공지능 모델 기술과제 해결 및 발전방안
- V. 결론

## 요약

본 연구는 국방 인공지능 모델이 가져야 할 기술과제를 분석하고 그 해결방안을 제안하는 것을 목적으로 한다. 특히 민간 분야와는 차이가 있는 국방 환경에 적용될 인공지능 모델이 가져야 할 특성을 분석하는 것으로부터 연구를 진행하였다.

국방 인공지능 모델은 민간 인공지능 모델과는 다른 몇 가지 특성이 있다. 첫째, 전투와 같은 긴박하고 위험한 의사결정에 적용되므로 결과의 신속성과 정확도에 대한 요구가 민간보다 높다. 둘째, 결과의 신뢰성을 확보하기 위해 설명력·해석력이 보장되어야 한다. 셋째, 적대국에 의해 모델을 무력화하는 적대적 공격의 위협이 크다. 넷째, 야전 혹은 함정·항공기 등 다양한 플랫폼에서 운용될 수 있으므로 간결하게 만들 필요가 있다.

국방의 특징을 충족시키기 위한 국방 인공지능 모델의 기술과제는 다음과 같다. 첫째, 설명 가능한 인공지능 모델을 개발해야 한다. 이를 통해 모델의 신뢰성 확보는 물론 모델의 개선 상태를 판단할 수 있다. 둘째, 모델을 의도적으로 무력화하는 적대적 공격에 대응하기 위해 적대적 방어기법을 적용해야 한다. 셋째, 처리 속도를 높이고 고성능 컴퓨팅 환경이 제한되는 플랫폼에서도 적용되도록 모델을 간소화하는 경량화 딥러닝 알고리즘을 적용해야 한다.

본 연구에서는 식별된 기술과제를 해결하기 위한 다양한 방법론을 소개하고 함정 분류모델 예시를 통해 각 과제별 해결 방법론을 제안하였다. 본 연구가 국방 인공지능 모델의 발전에 기여하기를 기대한다.

핵심어 : 국방 인공지능 모델, 설명 가능한 인공지능, 적대적 공격, 적대적 방어, 경량 딥러닝

\* 국방대학교 국방과학학과 박사과정

\*\* 국방대학교 국방과학학과 부교수, 교신저자

## I. 서론

ICT(Information and Communication Technology) 발전에 의한 사회 전반의 혁신은 우리의 패러다임을 빠르게 변화시키고 있다. 빅데이터와 인공지능, 네트워크의 발달로 초지능·초연결 사회가 되고 있다. 이러한 사회변혁에 대응하기 위해 우리나라는 2017년 4차산업혁명위원회를 설치하고 과학기술, 인공지능 및 데이터 기술 등의 기반을 확보하기 위해 노력을 기울이고 있다. 특히 인공지능을 경제적 가치 창출과 사회문제 해결의 핵심동력으로 보고 2019년 국가 인공지능 국가전략을 수립하여 인공지능 주도권 확보를 위한 총력을 기울이고 있다.<sup>1)</sup>

4차 산업혁명은 국방 분야에도 큰 영향을 끼치고 있다. 국방부는 2019년 스마트 국방 혁신단을 출범하여 국방운영, 기술·기반, 전력체계 3대 혁신 분야와 61개 과제를 선정하여 스마트 국방을 추진하고 있다.<sup>2)</sup> 또한 2021년 “국방 인공지능 추진전략”을 수립하여 인공지능을 통한 국방 전 영역의 혁신을 목표로 중·장기적이고 안정적인 인공지능 업무발전을 위한 계획을 수립하였다.

이러한 사회적 흐름과 정부, 국방부의 관심으로 국방 인공지능 발전을 위한 많은 연구가 이뤄져 왔다. 이종관·한창희(2019)는 미래전에 사용될 인공지능의 활용 분야와 인공지능 학습 알고리즘의 취약성, 데이터 및 설명력 부족, 윤리적 이슈 등을 지적하고 이를 해결하기 위해 민관협동 노력, 데이터 관리 등의 정책적 대안을 제시하였다.<sup>3)</sup> 김용삼(2019)은 한국군의 인공지능 발전을 위해 비전 및 규정 정립, 전문인력 확보 등의 정책적·제도적 기반 구축 방향을 제시하였다.<sup>4)</sup> 황태성·이만석(2020)은 인공지능의 전술적 수준과 작전적 수준에서 활용 가능한 방안을 제시하였고, 군사적 활용 시 예상되는 부정적 효과와 제한사항, 극복방안에 대하여 연구하였다.<sup>5)</sup> 조재규(2020)는 국방 인공지능을 위한 데이터, 컴퓨팅, 알고리즘의 인프라 환경을 진단하고 발전 방향을 제시하였다.<sup>6)</sup> 정두산(2021)은 국방 인공지능 발전을 위한 데이터, 알고리즘, 컴퓨팅 파워, 법·제도, 조직·인력의 인공지능 생태계 발전방안을 제시하였다.<sup>7)</sup> 기존 연구를 통해 국방 인공지능 발전을 위해 제도 정립 및 보완, 조직·인

1) 관계부처합동. (2019). 『인공지능 국가전략』

2) 이광재. “스마트 국방혁신 추진현황 및 발전방안 고찰.” 『한국 IT서비스학논집』제20권 제1호 (2021), pp. 1-9.

3) 이종관·한창희. “미래전과 국방 인공지능 체계.” 『한국 통신학회논문지』 제44권 제4호(2019), pp. 782-790.

4) 김용삼. “한국군의 인공지능(AI) 발전을 위한 정책적·제도적 기반 구축방향”, 『국방과 기술』482 (2019), pp. 48-59.

5) 황태성·이만석. “인공지능의 군사적 활용 가능성과 과제.” 『한국군사학논집』제76권 제3호(2020), pp. 1-30.

6) 조재규. “국방 인공지능 인프라 분석 및 발전방안”. 국방정책연구』제36권 제4호(2020), pp. 109-146.

력 보강 등 정책적 분야와 국방데이터·컴퓨팅 환경 등의 기반환경 분야, 인공지능 모델이 가지는 취약성·설명부족의 기술 분야, 윤리적 사용의 주요 이슈를 지적하였고 이에 대한 해결방안 마련을 강조하였다.

본 연구는 기존 연구들의 정책적 제언에 나아가 국방 인공지능 모델 기술개발(적용) 시 고려해야 할 예상 문제점을 식별하고 이에 대한 주의해야할 점을 제시하고자 한다. 이를 위해 국방 인공지능 모델의 특징을 분석하고 그에 요구되는 기술을 제시하였다. 그리고 제시된 기술을 적용하기 위한 다양한 방법론과 예시를 통해 이해를 도모하고, 기술 반영을 위한 국방 인공지능 모델 개발과정과 정책발전 방향을 제안하였다. 논문의 구성은 다음과 같다. 2장에서 인공지능의 개념과 주요 이슈를 살펴보고, 3장에서 국방 인공지능의 특징과 기술과제를 분석하였다. 4장에서는 제시된 기술과제를 해결하기 위한 방법론과 이를 적용하기 위한 개발절차를 제안하고, 5장에서 결론과 본 연구의 한계점 및 향후 연구과제에 대해 논하였다.

## II. 인공지능 개념과 주요 이슈

### 1. 인공지능의 정의

1950년 앨런 튜링(Alan Turing)은 “계산 기계와 지성”이라는 논문에서 인공지능에 대한 개념을 제시하였으며, 1956년 존 매커시(John McCarthy)는 다트머스 회의에서 처음으로 “인공지능”이라는 용어를 사용하며 “기계를 지능적으로 만드는 과학과 공학분야”로 정의하였다.<sup>8)</sup> 인공지능에 대한 정의는 국가 혹은 학자마다 다를 수 있으나 일반적으로 인공지능은 기억, 지각, 이해, 학습, 연상, 추론 등의 인간 지성을 모방하기 위해 기계를 이용하거나 컴퓨터상의 알고리즘을 통해 구현하는 것을 말한다.

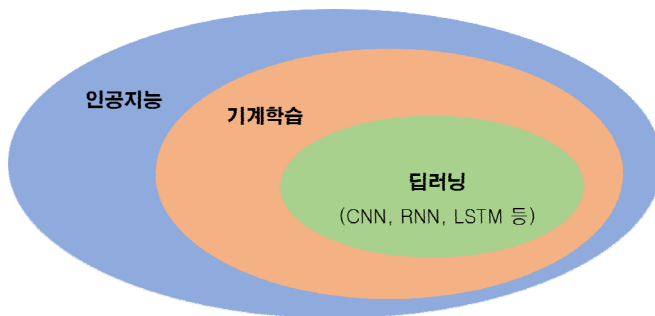
인공지능이란 용어는 딥러닝, 기계학습과 같은 용어와도 혼재하여 쓰인다. 하지만 인공지능이란 기계학습, 딥러닝의 상위개념으로 인간의 지성을 구현하는 모든 분야를 말한다. 기계학습은 데이터를 통해 스스로 학습하여 예측이나 판단을 제공하는 기술분야이고, 딥러닝은 인간의 뇌 구조를 모방한 인공신경망(ANN, Artificial Neural Network)을 활용하는 기

7) 정두산. “국방 인공지능(AI) 생태계 구축 방향 연구”. 『국방연구』, 제64권 제3호(2021), pp. 27-60.

8) 박영욱·쟁재원·김승천·이우신·유형곤·이지선·이규정. (2020). “국방 인공지능 발전 계획 수립연구”. 사단법인 한국국방기술학회. p.7.

계학습의 한 분야로 <그림 1>과 같은 범위를 갖고 있다. 특히 딥러닝은 합성곱 신경망(CNN, Convolutional Neural Network), 순환신경망(RNN, Recurrent Neural Network), LSTM(Long Short Term Memory)과 같은 알고리즘을 통해 이미지 인식, 자연어 처리 등의 현실적인 문제에 대해 매우 정확하게 문제를 해결하는데 사용되고 있다.

[그림 1] 인공지능 분류<sup>9)</sup>



## 2. 인공지능 모델의 주요 이슈

본 절은 인공지능 모델의 활용과 관련된 주요 이슈를 살펴보고자 한다.

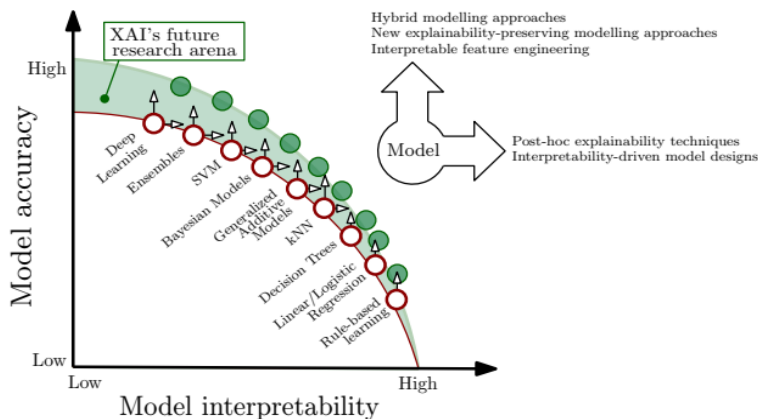
첫째, 인공지능 모델을 학습하기 위해서는 양질의 빅데이터가 필요하다. 인공지능 모델은 방대한 데이터로부터 인간이 볼 수 없는 데이터 내부의 패턴을 파악하고 이를 통해 인간보다 빠르고 정확하게 원하는 결과를 도출하게 된다. 하지만, 단순히 데이터가 많다고 훌륭한 모델이 되는 것은 아니다. 기본적으로 학습이 데이터 기반으로 이뤄지는 만큼 잘못된 데이터가 입력될 경우 의도하지 않은 방향으로 결과를 도출할 수 있다. 인공지능 학습을 위해서 구글, 마이크로소프트, 스탠포드 대학교와 같은 산학 연구기관은 전 세계 연구자들을 위해 대량의 동영상, 이미지, 음성과 같은 데이터를 무료로 제공하고 있으며, 우리나라도 정보진흥원 AI 허브를 통해 민간이 자체적으로 구축하기 어려운 이미지 데이터, 법률 데이터 등을 제공하고 있다.<sup>10)</sup> 하지만 이러한 공개 데이터 제공에도 불구하고 인공지능 모델을 개발하고 성능을 향상시키기 위한 학습 데이터는 제한적이다. 인터넷 공개자료는 제한적이고 연구목적에 부합하는 데이터를 생성하는 데에는 많은 시간과 비용이 요구된다. 또한 개인의 신상과 관련된 개인정보나 국가 안보, 보안과 관련된 데이터는 획득 자체가 어려운 것이 현실이다.

9) Francois Chollet. 박해선(역). 『케라스 창시자에게 배우는 딥러닝』. (서울: 길벗, 2019).

10) 조재규(2020), pp. 109-146.

둘째, 인공지능 모델은 해석성과 설명성이 부족하다. 해석성이란 모델의 출력값과 입력값 사이를 관계 짓는 것을 말하고, 설명성이란 모델의 출력에 영향을 미치는 입력값을 찾는 방법을 말한다.<sup>11)</sup> 인공지능 모델 내부는 수백~수천 만 개의 파라미터로 구성된 모델이다. 그리고 이 파라미터 값이 어떻게 도출되었는지는 알 수 없는 black box로 우리는 그 결과만을 얻은 모델이 생성되는 원리나 어떠한 입력이 결과에 어떤 영향을 미치는지를 알기가 제한된다. <그림 2>는 기계학습부터 딥러닝까지 그 성능과 해석가능성을 보여주는 그림으로 딥러닝 모델의 해석이 매우 어려움을 보여주고 있다.

[그림 2] 모델 성능과 해석성<sup>12)</sup>



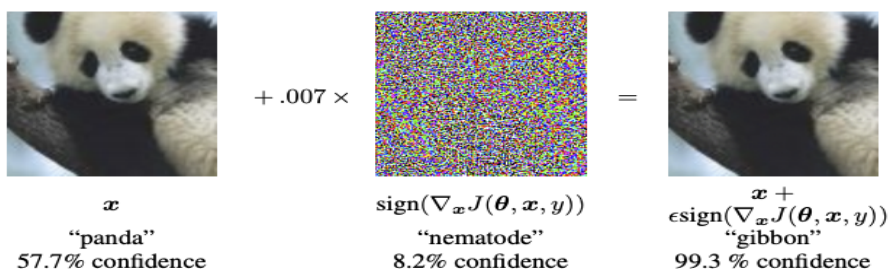
의사결정과 같은 추론분야에 있어서 모델의 결과에 대한 중간 해석 및 추론 과정이 설명되지 않을 경우 의사결정권자가 모델을 믿고 의사결정을 하기 어려울 수 있다. 이러한 한계를 해결하기 위해 미국의 고등연구계획국(DARPA, Defense Advanced Research Projects Agency)은 설명 가능한 인공지능(XAI, eXplainable AI) 프로젝트를 수행하고 있으며, 이후 많은 연구진이 딥러닝 모델 해석과 설명을 위한 연구를 진행하고 있다. 특히 모델 자체의 추론과정을 알기 어렵기 때문에 입력과 결과를 바탕으로 결과에 영향을 미치는 입력의 중요 부분을 밝히는 연구(Post-hoc explain)가 집중되고 있다. 이러한 연구를 통해 black box 모델을 조금이나마 볼 수 있는 gray box 모델이 되도록 노력을 기울이고 있다.

11) 고학수 · 김용대 · 윤성로 · 이선구 · 박도현 · 김시원. 『인공지능 원론』(서울: 박영사, 2021). pp. 66-69.

12) Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., Herrera, F. "Explainable Artificial Intelligence(xAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". Information Fusion, 58.(2020). pp. 82-115.

셋째, 인공지능 모델은 적대적 공격(Adversarial Attack)에 취약하다. Goodfellow 등 (2014)은 딥러닝에서 이미지 화소값에 아주 작은 변화를 주었을 때 모델이 오분류한다는 것을 발견하였다. 이렇게 의도적으로 모델의 성능을 저하시키기 위해 데이터에 작은 변형을 주는 것을 적대적 공격이라고 말하고, 적대적 공격으로 만들어진 새로운 데이터를 적대적 예제(Adversarial Example)라고 하였다.<sup>13)</sup> 적대적 예제는 이미지뿐만 아니라 자연어 및 오디오 처리 등의 모델에서도 발생할 수 있다.<sup>14)</sup> 또한 적대적 예제는 전이성(Transferability)을 가져 특정 모델에서 만든 적대적 예제는 다른 모델에서도 유사한 오분류 효과가 발생한다고 알려져 있다.<sup>15)</sup> <그림 3>은 적대적 공격의 기본 방법인 FGSM(Fast Gradient Sign Method)을 나타낸다. 좌측 원본 이미지에 중간 정도의 노이즈를 추가했을 때 오른쪽과 같은 이미지가 생성되는데 사람의 눈에는 두 이미지 모두 판다(panda)로 보이지만, 딥러닝 모델은 노이즈가 추가된 이미지를 긴꼬리원숭이(gibbon)로 잘못 분류하였다.

[그림 3] 적대적 공격(FGSM) 예시<sup>16)</sup>



적대적 공격으로 모델이 혼란을 일으킬 경우 막대한 피해가 예상된다. 예를 들어 딥러닝 기반 의료 영상인식에서 적대적 공격이 추가된 CT, MRI 사진을 환자로 분류하지 못하거나, 자율주행 차량 이미지 센서가 교통신호를 잘못 분류 할 경우 인명사고가 발생할 가능성이 높다. 따라서 이러한 모델의 취약성을 극복하기 위한 강건하고 안전한 딥러닝 모델 개발 연구가 필요한 바이다.

넷째, 인공지능 모델은 성능 향상을 위해 크기가 점점 커져 연산량이 증가하고 있다. <그

13) Goodfellow, I. J., Shlens, J., & Szegedy, C. “Explaining and harnessing adversarial examples”. arXiv preprint arXiv:1412.6572. (2014)

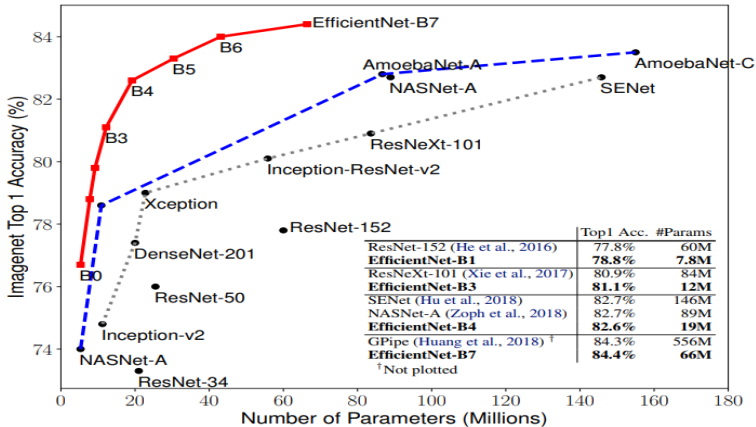
14) Katy Warr. 김영해(역). 『안전한 인공지능 시스템을 위한 심층 신경망 강화』. (서울: 한빛미디어, 2020).

15) 오희석. (2020). “이미지 기반 적대적 사례 생성 기술 연구 동향”. 『정보보호학회지』, 제30권 제6호. 서울: 정보보호학회, pp. 107-115.

16) Goodfellow. (2014).

림 4)는 이미지 인식 분류 모델에서 사용된 모델별 파라미터 수와 정확도를 보여주는 것으로 둘 사이는 높은 양의 상관관계가 있음을 알 수 있다.

[그림 4] 모델의 파라미터와 정확도<sup>17)</sup>



파라미터가 많다는 것은 그만큼 연산량이 많다는 것을 의미한다. 대량의 연산량을 처리하기 위해 GPU(Graphic Process Unit)와 같은 분산처리 장치를 통해 해결이 가능하나, GPU 설치가 제한되거나 계산능력이 떨어지는 컴퓨팅 환경에서는 딥러닝 모델 운용이 제한될 수도 있다. 또한 GPU의 연산속도 발전에 비해 딥러닝 모델의 연산량 증가 속도가 높으므로 이를 해결하기 위해 모델을 간략하게 만들 필요가 있다.

마지막으로 인공지능 윤리와 관련된 문제이다. 인공지능 윤리는 기술의 오남용을 방지하고 개인정보 보호 등을 위해 필요한 이슈이다. 앞서 적대적 공격에서의 밝힌 바와 같이 인공지능 모델의 취약성을 이용해 악의적으로 데이터를 조작할 경우 막대한 피해가 발생 할 수 있다. 또한 인공지능 모델을 이용해 대량 인명 피해를 유발하는 군사용 무인체계에 운용할 수도 있고, 가짜 영상 및 뉴스를 만들어 사회에 악영향을 끼칠 우려도 있다. 따라서 인공지능 모델을 만들 때 사회규범 및 인간의 보편적 가치에 부합하게 기술을 제한하고, 운영하도록 강제할 필요가 있다.

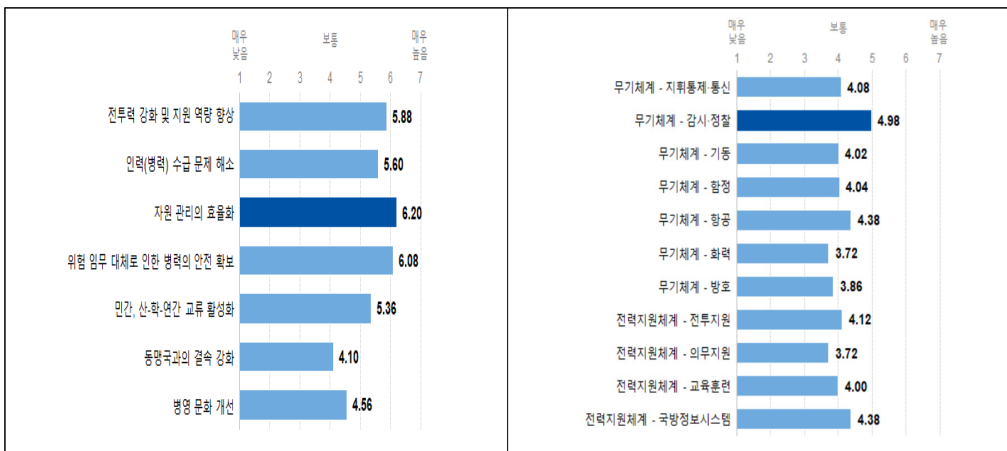
17) Tan, M., & Le, Q. "Efficientnet: Rethinking model scaling for convolutional neural networks". In *International Conference on Machine Learning*.(2019, May). pp. 6105-6114.

### Ⅲ. 국방 인공지능 모델의 특징과 기술과제분석

#### 1. 국방 인공지능 개념과 특징

국방 인공지능이란 국방의 문제를 해결하기 위해 국방 데이터를 활용하여 인공지능 알고리즘을 적용하여 문제를 해결하는 것을 말한다.<sup>18)</sup> 과학기술정책연구원이 2020년 국방 전문가 40인(군인, 연구소, 방산기업, 대학)을 대상으로 실시한 설문 조사에 따르면 국방 인공지능 기술 도입에 대해 전체 응답자의 94%가 매우 중요하다고 보았다. 또한 <그림 5>와 같이 인공지능 기술 활용 기대효용 질문에 자원 관리의 효율화, 위협 임무 대체 등에서 효용을 얻을 것이라고 응답했고, 인공지능 기술 활용도 질문에 무기체계 및 전력지원체계에서 보통 이상의 활용도를 가질 것으로 응답했다.<sup>19)</sup>

[그림 5] 인공지능 활용 기대효용(좌)과 기술 활용(우) 정도



이렇듯 국방에서 인공지능에 대한 기대는 크다고 할 수 있다. 이러한 기대에 부응하는 모델을 만들기 위해서는 국방이 가지는 특수한 환경을 고려할 필요가 있다. 국방에서 활용될 인공지능 모델은 다음과 같은 특징이 있다. 첫째, 국방 인공지능 모델을 위한 학습 데이터가 부족하다. 국방 데이터는 보안규정 및 군 전용 네트워크 사용 등으로 접근이 제한되는 경우

18) 문호석·손승연·임유신. “국방 인공지능 추진을 위한 전문조직 구성 및 인력관리 연구”. 국방대학교 산학협력단.(2021). pp. 21-25.

19) 윤정현. “국방분야 인공지능 기술도입의 주요쟁점과 활용 제고 방안”. 과학기술정책연구원. 279.(2021). pp. 27-30.



가 많으며, 데이터의 양이 적거나 불균형하여 학습하기에 불충분한 데이터일 가능성 높다. 둘째, 성능에 신뢰성이 확보되어야 한다. 국방 인공지능 모델은 전투와 같이 어렵고 힘든 결정에 활용될 수 있다. 따라서 민간에서 사용하는 모델에 비해 그 요구 성능이 더욱 높아야 한다. 또한, 신속·정확하게 결과를 도출해야 하고, 그 결과에 대한 신뢰성 확보를 위해 설명력과 해석력이 따라야 한다. 셋째, 사이버 위협에 대한 보안대책이 강구되어야 한다. 국방 분야는 위협국에 의한 사이버 위협이 높다. 따라서 적대적 공격을 통해 인공지능 모델을 무력화시킬 가능성이 높으므로 이에 대응할 수 있는 강건함을 가져야 한다. 넷째, 인공지능 프로그램의 간결성이 요구된다. 국방 인공지능 모델은 드론과 같은 소형 플랫폼으로 고성능 컴퓨팅 지원이 제한되는 환경에서 활용될 수 있다. 따라서 이러한 환경에서 사용될 수 있도록 간결하게 만들 필요가 있다. <표 1>은 국방 인공지능 특성과 민간의 인공지능과 주요 차이를 보여준다.

<표 1> 국방 인공지능 특성과 민간 인공지능과 차이점<sup>20)</sup>

특 성	내 용
기술적 특성	민간영역과 근본적 차이는 없으나, 설명 가능성과 적대적 공격 및 방어 연구 중요시 됨
운용환경	- 전장환경 적용 시 신속성이 우선시되며, 예측 불가 상황 발생 가능 - 가혹한 야전환경 및 다양한 군용 플랫폼에 적용가능
데이터	- 학습/훈련데이터 획득 어려움, 데이터 보안 요구
네트워크	- 폐쇄적 국방망으로 연결이 제한되는 경우 많음
요구사항	- 강건성, 안전성, 무결점 요구 등 성능기준이 높음
적대적 의도 가능성	- 민간보다 높음
성능검증, 안전, 표준기관 및 절차	- 민간보다 미비한 편

## 2. 국방 인공지능 모델 발전에 필요한 과제 도출

지금까지 살펴본 인공지능 주요 이슈와 국방 인공지능의 특징을 정리하면 <표 2>와 같다. 인공지능 주요 이슈가 국방 인공지능의 특징으로 작용하고 있으며 이러한 이슈와 특징을 고려해야 국방 인공지능 모델이 원활히 제 기능을 발휘할 수 있을 것이다.

20) 박영욱·쟁재원·김승천·이우신·유형곤·이지선·이규정. (2020). p.13.을 일부 수정

〈표 2〉 인공지능 이슈와 국방 인공지능 특징

인공지능 주요 이슈	국방 인공지능 특징
① 양질의 빅데이터 필요 ② 딥러닝 모델의 복잡도 높아짐으로 고성능 컴퓨팅 요구 ③ 설명, 해석의 어려움 ④ 적대적 공격에 취약 ⑤ 윤리기준 정립	① 학습용 데이터 획득 제한 ② 신속하고 정확한 결과 요구 ③ 설명력과 해석력 필요 ④ 적대적 공격 가능성 높음 ⑤ 야전 및 드론 등 고성능 컴퓨팅 지원이 제한되는 다양한 환경에서 운용 가능

이를 바탕으로 국방 인공지능 모델 발전을 위해 〈표 3〉과 같이 정책적 과제와 기술적 과제를 도출하였다. 두 가지 모두 국방 인공지능 모델 발전을 위해 관심을 기울이고 발전을 해야 할 부분이지만 정책적 과제인 데이터 확보, 컴퓨팅 환경 구축, 인공지능 윤리 가이드는 중·장기 정책적 과제로 발전시켜야 할 부분이다. 특히 윤리적인 문제는 사회적 합의를 거쳐 법적 제도와 함께 정립이 필요한 분야이기도 하다. 따라서 현 단계에서 국방 인공지능 모델을 개발할 때는 기술적 과제의 중요성을 인식하고, 기술 과제를 반영하여 모델을 개발하는 것이 필요하다. 즉, 정확도를 높이고 설명성을 보장하며 적대적 공격에 강건해야하며, 컴퓨팅 환경의 제약에서 벗어나 다양한 플랫폼에서 운용하도록 모델을 개발해야 한다.

〈표 3〉 국방 인공지능 모델 발전을 위한 주요 과제

정책적 과제	기술적 과제
① 국방 데이터 확보 ② 고성능 컴퓨팅 환경 구축 ③ 국방 인공지능 윤리정립	① 정확도 향상 ② 신속성 향상 ③ 설명성, 해석성 보장 ④ 적대적 공격에 대응 ⑤ 다양한 환경에서 운용토록 간결하게 개발

국방 인공지능 모델 발전을 위한 기술과제 중 정확도 향상은 모델 개발 시 양질의 학습데이터 확보와 모델 파라미터 미세조정의 시행착오를 통해 향상이 가능한 부분이다. 따라서 본 연구에서는 기술적 과제 ②, ③, ④, ⑤에 중점을 두었다. 이를 기술적 용어로 변경하면 1) 설명 가능한 인공지능, 2) 적대적 공격 대응, 3) 경량화된 인공지능으로 대체할 수 있다. 따라서 인공지능모델 개발 시 정확도만 고려하지 않고 제시된 기술과제들이 함께 반영될 수 있도록 개발할 필요가 있다.

## IV. 국방 인공지능 모델 기술과제 해결 및 발전방안

이번 장에서는 국방 인공지능 모델이 가져야 할 기술적 과제를 바탕으로 모델개발 시 적용될 수 있는 다양한 방법론을 소개하고, 적용 예시를 보여주고자 한다. 예시로 보여주는 인공지능 모델은 국방 감시·정찰 분야에서 도입되고 있는 이미지 분류 및 인식 분야 모델이다. 예시 모델은 해상경계에 활용하기 위해 항공모함, 지원함, 전투함, 잠수함, 상선, 어선, 부이 이미지를 이용해서 군함의 종류를 분류하는 모델이다. 예시를 통해 각 기술과제에 대해 이해도를 높이고 적용결과를 보여주고자 한다. 추가하여 기술과제를 접목하여 국방 인공지능 모델을 개발하는 절차(안)를 제안하고자 한다.

### 1. 설명 가능한 인공지능(XAI) 모델 개발

설명가능 인공지능 모델에 관한 연구는 크게 세 가지 관점에서 이뤄지고 있다. 먼저, 모델 해석의 복잡성과 관련된 것으로 “① 모델을 설계하고 학습할 때 설명이 가능하도록 모델을 (Transparent Model) 만들어 설명과 해석이 용이하도록 하는 방법, ② 복잡하게 만들어진 모델의 추론 과정을 설명(Post-hoc explain)하는 방법”이 있다. 두 번째로 설명의 범위가 “① 모든 결과(Global)에 대해 설명력을 보장하는 해석인가? 혹은 ② 하나 또는 일부(Local)의 결과만을 설명하는가?”에 따라 구분된다. 세 번째로, 설명기법이 “① 특정 모델에만 (Model specific) 적용 가능한 것인지? 혹은 ② 모델과 관련 없이(Model agnostic) 범용적으로 적용 가능한가?”에 따라 나눌 수 있다. 이러한 범주에 따라 XAI 기법을 적용할 수 있는 딥러닝 모델은 <표 4>와 같다.

<표 4> XAI 연구범위와 관련 모델

범주	하위범주	적용 모델	비고
복잡도 (Complexity)	Transparent Model	선형모델, 의사결정나무 등	모델 자체로 설명 및 해석용이
	Post-hoc Explain	CNN, RNN 모델	-
설명범위 (Scope)	Global	의사결정나무모델	모델 자체로 설명 및 해석용이
	Local	대부분의 기계학습/딥러닝 모델	-
적용범위 (Dependency)	Model Specific	CNN 모델	-
	Model Agnostic	CNN, RNN 모델	-

연구범주는 크게 세 가지이나 적용 모델과 설명기법은 상황에 맞게 적용될 수 있다. 인공지능 모델에 대한 설명을 제공하는 방법은 시각화하여 보여주는 방법, 결론 도출과정을 문자로 제공하는 방법, 혹은 특정 수치로 나타내는 방법 등이 있다.

현재 딥러닝 모델의 XAI의 연구는 ‘Post-hoc explain’ 방법을 중심으로 연구가 진행되고 있다. 딥러닝 모델은 수백에서 수천만 개의 파라미터를 가지고 있으며, 입력과 출력의 관계가 비선형 함수로 모델 해석이 제한된다. 따라서 입력과 출력과의 관계를 통해 설명하는 기법이 주를 이루고 있다. ‘Post-hoc explain’에 대한 주요 방법론과 관련 기법은 <표 5>와 같다.

<표 5> Post-hoc explain 방법론<sup>21)</sup>

구 분	설 명	관련 기법
Backpropagation based method	모델 예측결과를 역전파하여 입력에 대한 기여도를 계산하고, 기여도를 가시화하는 방법	Saliency (Simonyan et al., 2013), LRP(Bach et al., 2015) 등
Activation based method	CNN 모델에서 활성화된 값들의 선형 결합한 가중치를 이용	CAM(Zhou et al., 2016), Grad CAM (Selvaraju et al., 2017) 등
Perturbation based method	정보를 알 수 없는 모델에 입력의 변화를 주어 모델의 결과를 측정, 원래 입력의 변화를 해석으로 제공	LIME (Ribeiro et al., 2016) 등

XAI의 다양한 기법 중 본 연구에는 Saliency Map, LRP(Layer-wise Relevance Propagation), Grad CAM(Gradient Class Activation Method), LIME(Local Interpretable Model Agnostic Explanation)을 이용하였다. 각 방법에 대한 간략한 설명을 하고, 군합분류 인공지능 모델에 적용하여 결과를 해석하겠다. 먼저 입력값  $x_i$ , 모델  $f(\cdot)$ , 그에 따른 출력값을  $f(x)$ 을 가지는 딥러닝 모델을 가정하겠다.

Saliency Map은 돌출지도라고 불리며 입력 데이터 중에서 급격한 변화가 있어서 다른 주변 데이터에 비해 눈에 두드러지는 입력값을 보여주는 방법이다. 돌출값은 결과값  $f(x)$ 에 대해 입력값 픽셀  $x_i$ 에 대한 편미분의 크기를 통해 계산된다. 이 값이 클 경우 중요도가 크다고 보고 작으면 영향력이 작다고 판단하여 이 값을 원본 이미지에 겹쳐서 중요 부분을 나타내는 기법이다.

21) Wagner, J., Kohler, J. M., Gindele, T., Hetzel, L., Wiedemer, J. T., & Behnke, S. “Interpretable and fine-grained visual explanations for convolutional neural networks”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.(2019) pp. 9097-9107.


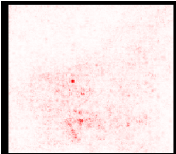
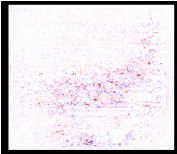
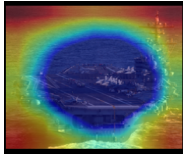

LRP는 출력값  $f(x)$ 를 타당도 점수(Relevance Score)로 간주하고 이 값을 분해하여 입력 이미지까지 역전파하는 방법이다. 이를 통해 입력 픽셀  $x_i$ 에 대해 타당도 점수를 측정하여 가시화하는 방법이다.

Grad CAM은 입력 이미지가 어디에 집중하고 있는지를 활성화 값(activation)을 통해 확인하는 방법이다. Grad CAM은 마지막 완전연결층의 Gradient를 가중치로 사용하여 입력 이미지에 heatmap 가중합을 구해 가시화하는 방법으로 컴퓨터 비전 분야에서 가장 많이 쓰이는 방법이다.

LIME은 ‘Post-hoc explain’ 하면서 Model agnostic한 방법이다. 기존의 방법들은 모델의 구조와 파라미터 값을 알아야 했다면, LIME은 모델 정보 없이도 사용이 가능하다. LIME은 입력값에서 샘플  $x'$ 를 랜덤하게 생성하고 그 주변을 확장시켜가며 샘플에 대해서 원 모델의 결과값  $f(x)$ 와 같은 결과를 내도록 학습시키는 방법을 말한다. 그리고  $f(x)$ 와 같은 결과를 내는 sample 부분을 가시화하여 중요 부분으로 보는 방법이다.

각각의 기법을 적용하여 모델에서 항공모함을 인식한 결과는 <그림 6>과 같다.

[그림 6] 군함 분류모델에 XAI 기법을 적용한 결과

원본	Saliency Map	LRP	Grad CAM	LIME
				

Saliency Map, LRP, Grad CAM의 붉은색 부분은 모델이 항공모함으로 분류할 때 큰 영향을 끼친 부분이다. LIME은 녹색으로 가려진 부분을 제외한 부분이 결과에 영향을 미친 부분이다. Saliency Map과 LRP는 noise가 퍼져있어 시인성은 떨어진다. 단, 항공모함의 비행갑판과 함교 부분이 진하게 나타나고 있어 그 부분을 중요한 것으로 보는 것을 알 수 있다. Grad-CAM과 LIME은 원본 이미지에 heatmap과 음영처리를 하여 결과를 나타내므로 시인성이 높다. 하지만 Grad CAM은 항공모함 자체보다 항공모함을 둘러싼 배경에 더욱 중요하게 인식함을 알 수 있고, LIME은 가시화된 부분으로는 정확히 해석하기가 어렵다. 이렇듯 XAI 적용을 통해서 인공지능 모델이 입력의 어느 부분에 중점을 두고 결과를 도출했는지를 유추해 볼 수 있다. Saliency Map과 LRP는 인간의 항공모함 인지 기준과 유사하게 분류하고 있었으며, Grad CAM, LIME으로 확인하였을 때는 인간과 다른 기준으로 분류함

을 알 수 있다. 모델 설명에 대한 정확한 평가방법이나 정량적인 기준이 없으므로 전문지식을 가진 사람의 기준에 따라 정성적인 평가를 해야 하는 한계가 있다. 하지만, XAI 기법 적용을 통해 모델이 어디에 중점을 두고 결과를 도출하는지 알 수 있으며, 이를 통해 모델의 신뢰성을 확인하고 또한 그 결과를 통해 모델 개선의 필요성도 확인 할 수 있다.

## 2. 적대적 공격 대응

적대적 공격 방법은 크게 세 가지에 따라 분류할 수 있다. 먼저 공격이 발생하는 환경이 컴퓨터 내부에서 발생하는 디지털 공격과 현실 세계에서 발생하는 물리적 공격이 있다. 또한 공격 대상이 특정 목표로 오분류 하도록 공격하는 Target 공격, 원래 목표와는 다른 결과로만 도출하면 되는 Non-target 공격이 있다. 또한 공격자가 공격하는 모델에 대한 구조, 파라미터 값, 학습 데이터 등의 정보를 어느 정도 알고 있는지에 따라 White-box 공격, Gray-box 공격, Black-box 공격으로 나눌 수 있다. White-box 공격은 공격대상 모델의 모든 정보를 알고 있을 때이며, Black-box 공격은 입력과 출력만을 알 수 있을 때, Gray-box 공격은 모델에 대한 부분적인 정보만을 알고 있을 때를 말한다. <표 6>은 분류기준에 따른 다양한 공격방법이다.

<표 6> 적대적 공격 분류 및 방법<sup>22)</sup>

구 분		관련 방법
디지털 공격	White-box Attack	FGSM(Goodfellow et al., 2015)
		PGD(Madry et al., 2017)
		C&W(Carlini & Wagner, 2017)
	Black-box Attack	ZOO(Chen et al., 2017)
	Gray-box Attack	AdvGAN(Xiao et al., 2018)
물리적 공격	Sticker Attack(Eykholt et al., 2017) Adversarial Patch(Brown et al., 2017)	

국방 인공지능 모델에서 예상되는 가장 위협적인 공격방식은 물리적 공격이다. 디지털 공격은 디지털 이미지에 특정 노이즈를 추가해 합성 이미지를 만드는 것으로 FGSM, PGD, CW등의 공격방법이 있다. 이러한 디지털 공격은 모델에 입력되는 디지털 데이터를 생성하는 것으로 실제 객체를 관측하는 광학 카메라에 적용은 제한된다. 하지만 물리적 공격은 <그

22) 유영준·양병희·노용만. “영상 정보 분야의 딥러닝 기반 적대적 공격과 방어 기술 동향 분석 및 국방 분야 적용 방안”, 『국방경영분석학회지』, 제46권 제1호(2020). pp. 1-30.

림 7)과 같이 도로표지판에 특정 Sticker를 붙여(Sticker Attack) stop 신호를 “속도제한”으로 오인하게 하거나, 바나나에 특정 패치를 붙여(Adversarial Patch) 바나나를 토스터기로 오분류하게 한다. 즉, 물리적 공격은 현실 세계에 적용하여 실시간 감시체계의 무력화가 가능한 만큼 큰 위협이 될 가능성이 있다.

[그림 7] Sticker Attack(좌) 및 Adversarial Patch(우)<sup>23)</sup>



이러한 적대적 공격의 원리는 오차 역전파 알고리즘을 반대로 이용하는 것이다. 오차 역전파는 모델학습 시 예측값과 참값의 오차를 측정하고 이를 줄이는 방향으로 모델 파라미터를 갱신하는 것을 말한다. 하지만 적대적 공격은 오차가 늘어나는 방향으로 데이터를 옮김으로써 오차가 커지는 데이터를 만들 수 있다.

아래 <그림 8>은 적대적 공격에 대한 예시이다. 항공모함을 항공모함으로 정확히 분류하는 모델이 있을 때 디지털 공격(FGSM)과 물리적 공격(Patch) 결과로 디지털 공격을 하였을 경우 항공모함을 80% 확률로 Battleship으로 분류하였으며, 물리적 공격은 79%로 Battleship으로 분류의 정확성이 감소하였다.

[그림 8] 적대적 공격 결과

원 본	디지털 공격(FGSM)	물리적 공격(Patch)
Predicted: AirCraft Predicted Score: 0.99864656	Predicted: BattleShip Predicted Score: 0.805475	Predicted: BattleShip Predicted Score: 0.73991257
		

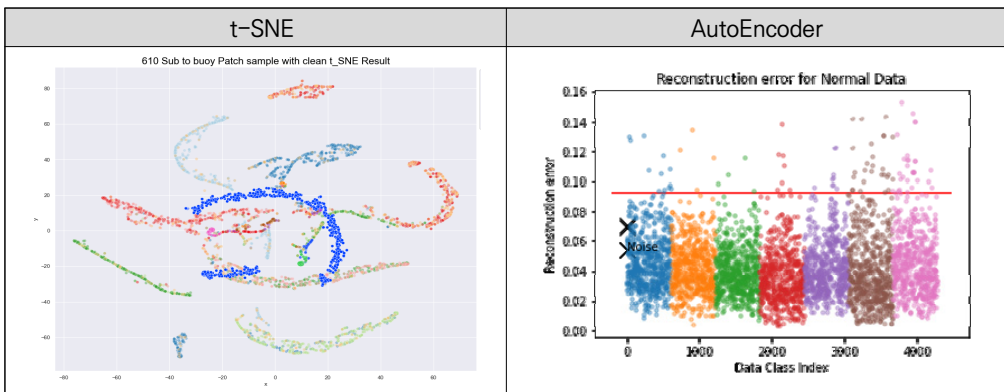
23) Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Song, D. “Robust physical-world attacks on deep learning visual classification”. In *Proceedings of the IEEE conference on computer vision and pattern recognition*(2018). pp. 1625-1634.

한편, 적대적 공격에 대한 대응을 적대적 방어(Adversarial Defense)라고 한다. 적대적 공격에 대한 가장 간단하고 쉬운 대응 방법은 적대적 공격으로 만들어진 적대적 예제를 다시 훈련데이터로 사용하는 것이다. 하지만, 적대적 예제를 만드는 데 시간이 오래 걸리며, 적대적 예제로 재훈련한 모델에서도 또 다른 적대적 예제를 만들 수 있으므로 완벽한 방법은 아니다. 적대적 방어의 다른 방법은 학습 기울기 자체를 숨기거나 기울기 자체를 정규화 방법을 통해 모호화시키는 방법이 있다. 하지만 학습 기울기를 이용하지 않는 적대적 공격에 대해서는 대응이 불가능한 단점이 있으며, 타 모델을 이용해 적대적 예제를 생성하는 적대적 공격 전이성을 이용하면 무력화가 가능하다는 단점이 있다. 세 번째로 적대적 예제 자체를 탐지하는 방법으로 적대적 예제를 탐지하는 추가 모델을 만들거나 결과의 일관성을 확인하는 방법 등이 있다. 이 방법 또한 적대적 예제 탐지를 위한 추가 데이터가 필요한 단점이 있다. 이러듯 다양한 방어기법들이 있지만 완벽한 방법이 없으며 각각의 장단점이 있다.

따라서 국방 인공지능 모델을 개발할 때에는 모델 정보가 노출되지 않도록 해야 하며, 다양한 방어 기법을 적용하여 적대적 공격에 강건하도록 모델을 만들어야 할 것이다.

〈그림 9〉는 적대적 예제의 탐지 방법으로 차원축소방법인 t-SNE(Stochastic Neighbor Embedding)와 생성 모델인 AutoEncoder를 이용한 예시이다. t-SNE는 딥러닝과 같은 비선형 모델의 고차원 결과를 축소하여 가시화하는 방법이고, AutoEncoder는 원본과 유사한 이미지를 생성하고, 원본↔생성 이미지간의 오차를 확인하여 임계치 이상의 오차가 발생할 경우 이상치로 탐지하는 방법이다. 〈그림 9〉의 각 포인트 색은 분류 모델의 클래스(항공모함 : 주황색 등)를 나타낸다.

[그림 9] 적대적 예제 탐지 예시





t-SNE의 중간부분의 파란색 점은 잠수함을 해상부이로 오분류 하도록 만든 610개의 Patch 결과를 2차원 좌표에 나타낸 것이고, 나머지는 원본인 항공모함, 전투함 등의 원본 이미지 범주별로 2차원 좌표로 나타낸 것이다. Patch 이미지 일부가 원본 이미지와 겹치기는 하지만 완전히 다른 분포로 나타나 시각적으로 적대적 예제를 확인할 수 있다. 우측은 AutoEncoder를 활용해 원본 이미지를 재생성 했을 경우 원본과 재생성 이미지의 오차를 확인한 것이다. 대부분의 재생성 오차는 0.1 이하로 나타났다. 하지만 <그림 9>의 항공모함 적대적 예제(X 표시)에 대한 재생성 오차 또한 0.1 이하로 나와 특이 데이터로 인지하지 못하였다. 이와 같이 적대적 예제를 완벽히 방어하거나 탐지하는 완벽한 방법은 없다. 따라서 다양한 방어기법을 중복 적용하여 상호 보완하는 방법으로 방어율을 높일 필요가 있다.

### 3. 경량화 인공지능

딥러닝은 깊은 레이어로 구성된 모델이다. 따라서 파라미터가 많고 이에 비례하여 연산량이 증가하고, 이를 처리하기 위해 고성능 컴퓨팅 자원이 요구된다. 하지만 모바일 기기나 사물인터넷, 소형전산기 등은 고성능의 컴퓨팅 탑재가 제한되고, 대량의 연산이 지속될 경우 소형 플랫폼의 배터리 소모로 기기의 본 성능이 제한될 수 있다. 이러한 배경으로 모바일 기기 등 저성능 컴퓨팅 환경 장치를 위한 경량화된 딥러닝이 필요하다. 모델을 경량화하기 위한 방법론은 <표 7>과 같다.

<표 7> 경량화 기법 요약<sup>24)</sup>

구 분	요 약
가지치기(Pruning)	기준값 이하의 가중치나 뉴런을 제거하며 반복 학습
지식증류 (Knowledge Distillation)	기 학습된 모델을 teacher 모델로 하여 경량화된 student 모델을 학습
양자화(Quantization)	가중치, 활성화 값, 기울기 값을 보다 낮은 비트 너비로 표현
경량 디자인 (Compact Network Design)	연산의 효율을 높이기 위해 딥러닝 모델의 구조를 변경

가지치기란 딥러닝의 모든 파라미터가 결과에 동일한 영향을 미치지 않는다는 관찰에서 출발한다. 따라서 결과에 큰 영향을 미치지 않는 레이어, 뉴런을 제거하여 파라미터를 줄이

24) 이경하·김은희, “딥러닝 모델 경량화 기술 분석”, 한국과학기술정보연구원(2020), p. 26.

는 방식을 말한다. 지식증류는 Hinton et al.(2015)에 의해 처음 제안된 방법으로 기 학습된 여러 모델을 앙상블하여 보다 작은 모델로 학습시키기 위해 고안된 방법이다. 양자화는 32비트 부동소수점 기반의 학습 파라미터 값을 이보다 낮은 비트로 표현하는 방법이다. 즉, 32비트 부동소수점으로 표현되던 파라미터 값을 1비트 혹은 2비트 등으로 줄이는 방법을 말한다. 경량 디자인은 모델 자체의 구조를 바꾸거나 일반적인 공간 합성곱(Spatial Convolution) 연산의 연산량을 줄일 수 있는 Depthwise Seperable 합성곱 연산, Dilated 합성곱 연산 등으로 바꾸는 알고리즘적 접근방법이다.

이중 본 연구에서 적용한 기본적인 공간 합성곱 연산과 Depthwise Seperable 합성곱 연산 과정은 다음과 같다. 입력 이미지의 크기가  $H \times W$ 이고 rgb(red, green, blue) 3개의 컬러 채널로 입력되고 합성곱 필터의 크기가  $3 \times 3 \times 3$  M개로 구성되었다고 하자. 공간 합성곱 연산의 경우 한 개의 결과를 얻기 위해 입력값에  $3 \times 3 \times 3$  필터를 적용하고 필터 수 M번 만큼 연산하여 총  $27 \times M$ 번의 연산이 이뤄진다. 반면 Depthwise Seperable 합성곱은 입력 채널(r, g, b)을 분리하여  $3 \times 3$  필터 연산을 하는 Depthwise 합성곱 연산을 한 후 채널 연산결과를 결합하고,  $1 \times 1 \times 3$ 을 필터를 M번 연산하는 Pointwise 합성곱 연산을 이용한다. 이를 통해 총 계산량은  $27+3M(=3 \times 3 \times 3+3M)$ 이 된다. 따라서 약 1/9의 계산 및 파라미터 감소를 이룰 수 있다.

〈표 8〉 공간 합성곱 연산과 Depthwise Seperable 합성곱 연산 비교

(입력 :  $H \times W \times 3$ , 필터 :  $3 \times 3 \times 3$  M개)

구분	공간 합성곱 연산	Depthwise Seperable 합성곱 연산
1 output 계산량	$3 \times 3 \times 3 \times M = 27 \times M$	$(3 \times 3 \times 3 + 1 \times 3 \times M)$ $= 27 + 3M$
파라미터 수	$3 \times 3 \times 3 \times M \times H \times W =$ $27 \times M \times H \times W$	$(3 \times 3 \times 3 + 1 \times 3 \times M) \times H \times W$ $= (27 + 3M) \times H \times W$

여러 경량화 기법 중 본 연구에서는 경량 디자인을 적용하여 Mask RCNN을 경량화 하였다. Mask RCNN은 He et al.(2017)에 의해 제안된 객체 분할 및 탐지 모델로서 〈그림 10〉과 같이 이미지를 입력받아 관심 영역을 분할하여 탐지 및 분류할 수 있는 모델이다.

[그림 10] Mask RCNN 예시



(좌 : 입력 이미지, 우 : Mask RCNN 결과 (군함, 함포, 유도탄 탐지))

Mask RCNN은 입력 이미지에서 특징을 추출하기 위해 ResNet50/101이라는 CNN 모델을 기본 backbone으로 이용한다. 하지만 ResNet은 50 또는 101개의 레이어로 구성된 모델이다. 따라서 기본 Mask RCNN을 경량화하기 위해 backbone을 MobileNet V1으로 대체하였다. MobileNet V1은 Depthwise Seperable 합성곱 연산을 적용하였다는 특징이 있다. 이러한 경량구조 기법을 적용하여 <표 9>와 같이 모델의 크기와 파라미터 수가 66% 감소하였으며, 1-샘플 이미지의 추론 시간도 0.78초에서 0.71초로 약 9% 빨라져 신속성도 증가하였으나, 정확도는 약 3.9% 줄어들었다. 하지만, 정확도는 모델 재학습 및 미세조정을 통해 향상이 가능하므로 제한점으로 작용 되지 않으며, 더 작은 모델로 효율적으로 모델을 운용할 수 있음을 보여주었다.

<표 9> 기본 Mask RCNN과 경량 Mask RCNN 결과 비교<sup>25)</sup>

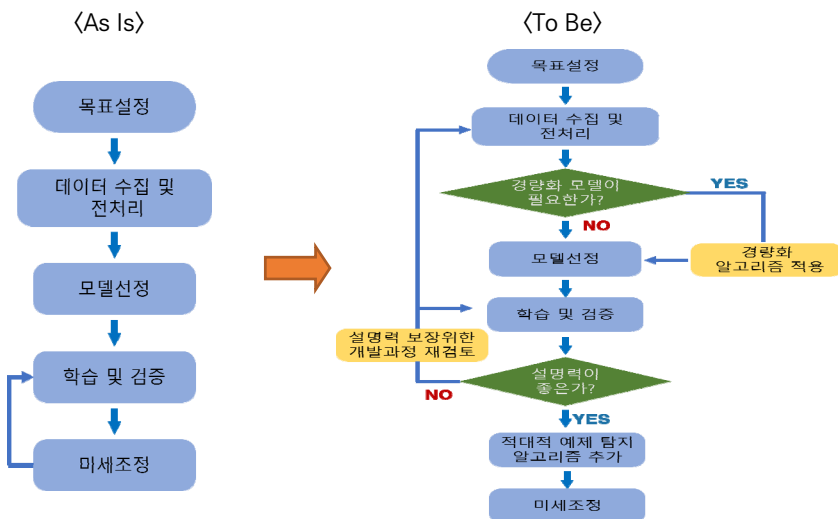
구 분	기본 Mask RCNN	Mask RCNN with MobileNet V1
모델크기 (Mega Byte)	244MB	83.2Mb(↓ 66%)
총 파라미터 수 (백만)	63.7M	21.7M(↓ 66%)
1 샘플 이미지 추론 시간(초)	0.78s	0.71s(↓ 8.9%)
정확도(mAP)	87.62%	84.18%(↓ 3.9%)

25) 박진영·김정환·문호석. “경량 딥러닝 모델을 활용한 고가치(HVU) 군함 Image Segmentation에 관한 연구”. 대한산업공학회/한국경영학회 춘계 공동학술대회(2021. 6. 4.), pp.3982-3983.

#### 4. 국방 인공지능 모델 개발 시 기술과제 적용(안) 및 정책제안

본 절은 기술과제를 적용하여 국방 인공지능 모델을 개발하는 방안을 제안하고자 한다. 예시에서 살펴본 바와 같이 기술적 과제 적용 시 국방 인공지능 모델의 결과 도출 과정을 미세하게 확인 가능하며, 적대적 공격에도 강건하고, 경량화하여 학습과정 및 추론속도가 빨라짐을 보여주었다. 따라서 국방 인공지능 모델을 개발 시에는 기술 고려사항을 적용하여 모델을 강화할 필요가 있다. 일반적인 딥러닝 모델 개발과정은 개발 목표를 설정하고 데이터 수집 및 전처리, 학습 및 미세조정의 과정을 거친다. 하지만 국방 인공지능 모델은 기술과제를 적용하기 위해 <그림 11>과 같은 절차를 적용하는 방안을 제안한다. 우선 개발 목표를 설정하고 데이터 수집 및 전처리를 한다. 다음으로 모델이 적용될 환경을 고려하여 경량화 여부를 판단한다. 이후 적용할 모델을 선정하고 학습 및 성능 검증을 실시한다. 이때 성능검증은 단순히 정확도를 평가하는 성능뿐만 아니라 XAI를 적용하여 사람의 판단기준과 유사하게 결과를 도출하는지에 대한 설명력을 정성적으로 평가할 필요가 있다. 만약 사람의 판단기준과 다른 기준으로 결과를 도출한다면 모델개발 전반에서 이를 개선하기 위한 방안을 강구할 필요가 있다. 이후 적대적 공격에 대응하기 위해 학습된 모델을 기반으로 적대적 예제를 생성하고 이를 재학습 또는 탐지하는 네트워크를 추가하여 모델의 강건성을 확보한다. 최종적으로 미세 조정을 통해 모델 정확도, 설명력, 강건성 등 모델의 성능을 극대화시킨다.

[그림 11] 국방 인공지능 모델 개발 절차(안)



제기된 기술 과제를 해결하기 위해서는 정책적인 부분의 노력도 병행되어야 한다. 기술과제 해결을 위한 정책적 제안은 다음과 같다. 먼저 군 위탁교육을 통한 전문인력 양성 시 국방 인공지능 모델의 기술적 해결 과제를 부여하는 것이다. 즉, 본 연구에서 제시한 ‘설명 가능한 인공지능’, ‘적대적 공격’, ‘경량화된 딥러닝’ 등을 위탁교육시 연구해야 할 주제로 부여하여 연구하도록 과업을 부여하는 방안이다. 이를 통해서 자연스럽게 석·박사학위 과정간 인공지능 모델의 기술적 과제에 대한 연구를 하게 되고, 위탁교육 종료 후 국방에 돌아와서 자체 기술개발 혹은 용역개발 시 관리·감독을 할 수 있는 인원으로 활용할 수 있을 것이다. 둘째, 전문인력 위탁교육을 통한 발전방안에 추가하여 현재 육군에서 운용하고 있는 군사과학기술병을 활용하는 방안이다. 군사과학기술병은 석사 재학 이상의 전문지식을 가지고 있는 병사들로 육군 인공지능 발전처, 빅데이터 센터 등 150여 개의 직위에서 근무중으로 전문지식과 함께 민간에서 석박사 기간 중 다양한 프로젝트를 경험하여 문제해결 능력이 있는 자원이다.<sup>26)</sup> 따라서 이를 활용하여 국방 인공지능 모델의 기술적인 과제를 연구하고 군에 적용할 수 있도록 임무를 부여하여 지속적으로 연구하도록 하는 것이다. 셋째, 전문인력 활용을 제고하고 지속적으로 기술과제 연구를 하기 위해 전담 조직이 필요하다. 인공지능과 관련된 국방내 조직이 '21.7월에 국방부에 인공지능 TF가 만들어져 있으나 소규모이고, 육군을 제외하고는 현재까지 공식적인 조직이 없는 상태이다. 당장은 국방부의 인공지능 TF가 기술적 과제를 주요과제로 정하여 추진하긴 어려운 상황이지만 육군에서라도 육군내 인공지능관련 조직을 활용하여 국방 인공지능 모델의 기술적 과제에 대한 연구를 조직내 과업으로 선정하여 연구해 갈 필요가 있다.

## V. 결 론

인공지능은 스마트 국방의 핵심동력으로 전투, 전력증강, 인사, 군수 등 다양한 분야에 접목하여 활용될 것이다. 하지만, 국방은 민간과는 다른 특수성이 있으므로 민간의 모델을 그대로 받아들이기는 보다는 군의 특수성을 고려한 국방 인공지능 모델을 개발해야 한다. 연구를 통해 인공지능의 주요 이슈와 국방 인공지능의 특징을 분석한 후 이를 바탕으로 국방 인공지능 모델의 발전을 위한 정책적 과제와 기술적 과제를 도출하였다. 정책과제로 국방 데이터 확보, 고성능 컴퓨팅 환경 구축, 윤리문제 해결을 제시하였다. 기술적 과제로 정확도

26) 문호석(2021), pp.102-104.

향상, 신속성 향상, 설명성 보장, 적대적 공격 대응, 다양한 환경에서 운용토록 간결한 모델 개발을 제시하였다. 그리고 그동안 깊게 다뤄지지 않은 기술적 과제에 집중하여 기술과제에 대한 개념과 이를 해결하기 위한 다양한 방법론과 적용 예시를 보여주었다. 또한 이러한 기술과제를 반영하여 국방 인공지능 모델을 개발하는 방안과 정책적 발전 방안을 제안하였다.

국방 인공지능의 필요성과 중요성은 누구나 공감하는 바이다. 하지만, 국방 인공지능 모델이 가져야 할 기술적 과제가 해결되지 않는다면 오히려 불필요하게 관리해야 할 국방 모델 혹은 관리체제로 여겨질 수 있다. 이러한 문제점을 방지하기 위해 국방 인공지능 모델에 필요한 다양한 기술과제와 해결방안을 제시하고, 국방 인공지능 모델 개발 절차를 제안하였다. 또한 인력 및 조직 측면에서 정책적 제안을 하였다. 연구에서 제기한 국방 인공지능 모델의 기술과제는 현 단계에서 식별된 제한점이다. 향후 인공지능의 발전에 따라 더욱 많은 기술과제가 식별될 것이다. 따라서 현재 식별한 기술과제를 기초로 더욱 많은 과제를 식별하고, 해결방안을 연구하여 국방 인공지능 모델이 발전하고 이것이 국방 분야에 잘 활용되기를 기대한다.

## 참고문헌

- 관계부처합동. 『인공지능 국가전략』(2019)
- 고학수 · 김용대 · 윤성로 · 이선구 · 박도현 · 김시원. 『인공지능 원론』 (서울: 박영사, 2021)
- 김용삼. “한국군의 인공지능(AI) 발전을 위한 정책적 · 제도적 기반 구축방향”, 『국방과 기술』 482(2019)
- 문호석 · 손승연 · 임유신. “국방 인공지능 추진을 위한 전문조직 구성 및 인력관리 연구”. 국방대학교 산학협력단(2021)
- 박영욱 · 쟁재원 · 김승천 · 이우신 · 유형곤 · 이지선 · 이규정. “국방 인공지능 발전 계획 수립연구”. 사단법인 한국국방기술학회(2020)
- 박진영 · 김정환 · 문호석. “경량 딥러닝 모델을 활용한 고가치(HVU) 군함 Image Segmentation에 관한 연구”. 2021 대한산업공학회/한국경영학회 춘계 공동 학술대회. 제주: 제주ICC. (2021. 6. 4.).
- 오희석. “이미지 기반 적대적 사례 생성 기술 연구 동향”. 『정보보호학회지』, 제30권 제6호. 서울: 정보보호학회(2020)
- 유영준 · 양병희 · 노용만. “영상 정보 분야의 딥러닝 기반 적대적 공격과 방어 기술 동향 분석 및 국방 분야 적용 방안”, 『국방경영분석학회지』, 제46권 제1호. (2020)
- 윤정현. “국방분야 인공지능 기술도입의 주요쟁점과 활용 제고 방안”. 과학기술정책 연구원 (279) (2021)
- 이경하 · 김은희. “딥러닝 모델 경량화 기술 분석”. 한국과학기술정보연구원 (2020)
- 이광제. “스마트 국방혁신 추진현황 및 발전방안 고찰”, 『한국 IT서비스학논집』, 제20권 제1호. (2021).
- 이종관 · 한창희. “미래전과 국방 인공지능 체계”. 『한국 통신학회논문지』, 제44권 제 4호. (2019)
- 조재규. “국방 인공지능 인프라 분석 및 발전방안”. 『국방정책연구』, 제36권 제4호. (2020)
- 정두산. “국방 인공지능(AI) 생태계 구축 방향 연구”. 『국방연구』, 제64권 제3호.(2021)
- 황태성 · 이만석. “인공지능의 군사적 활용 가능성과 과제”. 『한국군사학논집』, 제76권 제3호. (2020)

Francois Chollet. 박해선(역). 『케라스 창시자에게 배우는 딥러닝』. (서울: 길벗, 2019).

Katy Warr. 김영하(역). 『안전한 인공지능 시스템을 위한 심층 신경망 강화』. (서울: 한빛미

디어, 2020)

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Herrera, F. “Explainable Artificial Intelligence(xAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. *Information Fusion*, 58. (2020)
- Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Song, D. “Robust physical-world attacks on deep learning visual classification”. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.(2018)
- Goodfellow, I. J., Shlens, J., & Szegedy, C. “Explaining and harnessing adversarial examples”. *arXiv preprint arXiv:1412.6572*.(2014)
- Tan, M., & Le, Q. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In *International Conference on Machine Learning*.(2019, May)
- Wagner, J., Kohler, J. M., Gindele, T., Hetzel, L., Wiedemer, J. T., & Behnke, S. “Interpretable and fine-grained visual explanations for convolutional neural networks”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019)



## A study on defense AI model technical tasks and development plans

Jinyoung Park, Hoseok Moon

### Keywords

Defense AI, Explainable AI, Adversarial Attack, Adversarial Defense, Lightweight Deep Learning

The purpose of this study is to analyze the technological challenges that defense artificial intelligence(AI) model should have and to suggest solutions. In particular, the research was conducted by analyzing the characteristics that an AI model to be applied to a defense environment that is different from that of the private sector should have.

The defense AI model has several characteristics that are different from the civilian AI model. First, since it is applied to urgent and risky decision-making such as combat, the demand for promptness and accuracy of results is higher than civilian AI. Second, explanability and interpretability must be guaranteed to ensure the reliability of the results. Third, the threat of adversarial attack that neutralizes the model by the hostile country is great. Fourth, since it can be operated on various platforms such as field battles or ships and aircraft, it is necessary to make it concise.

Three technical tasks were derived for the defense AI model to satisfy the characteristics of the defense environment. First, we need to develop an explainable AI model. Through this, it is possible to not only secure the reliability of the model, but also to determine the improvement status of the model. Second, adversarial defense techniques must be applied to counter adversarial attacks that intentionally neutralize the model. Third, it is necessary to apply a lightweight deep learning algorithm that increases the processing speed and simplifies the model so that it can be applied even on platforms where high-performance computing environments are limited.

In this study, various methodologies for solving the identified technical tasks were introduced, and practical application examples related to the classification of ships were presented to suggest appropriate methodologies for each technical task. We expect that this study will contribute to the development of defense AI models.

[논문투고일 : 2022. 4. 27]

[심사의뢰일 : 2022. 5. 23]

[게재확정일 : 2022. 7. 7.]