

KISDI

Premium Report

오픈소스 AI : 인공지능 생태계와 오픈 이노베이션

남 충 현

정보통신정책연구원 부연구위원



정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

오픈소스 AI : 인공지능 생태계와 오픈 이노베이션

남 충 현

정보통신정책연구원 부연구위원

요약문

1. 서 론	4
2. AI 플랫폼의 개념	7
3. 인공지능 플랫폼 오픈소스화 현황	10
4. AI 개방형 혁신 생태계 구축을 위한 과제	17
[참고문헌]	22

남 충 현

정보통신정책연구원 부연구위원

*namch@kisd.re.kr, 043-531-4180

*연세대학교 경제학과 학사, 석사

*University of Warwick 경제학박사

오픈소스 AI : 인공지능 생태계와 오픈 이노베이션

요 약 문

최근 알파고로 유명한 구글 딥마인드에서 인공지능 기술 테스트 플랫폼인 딥마인드랩(DeepMind Lab)을 외부에 무료 공개하기로 결정하였다. 이는 지난해 구글의 인공지능 플랫폼 텐서플로우(Tensor Flow) 공개 등 일련의 인공지능 기술 오픈 소스화 흐름의 연속선상에 있다. 이러한 인공지능 알고리즘을 구현한 플랫폼의 개발에는 대규모 자본이 투입되어야 하며 고도의 기술력이 집약되어있는 것인데 왜 무료로 외부에 공개를 하는 것일까? 그 이유는 첫째로, 딥러닝 등 인공지능의 핵심 알고리즘은 이미 어느 정도 정립이 되어있어서 알고리즘보다는 이 알고리즘이 적용되는 기반 데이터가 더욱 중요해졌다는 점이 있다. 둘째로는, 인공지능 기술은 전통적인 ICT 산업 부문을 넘어 전통산업을 포괄하는 매우 광범위한 영역으로 응용이 되며, 또 응용대상 영역에서 얻어지는 데이터 및 사용자 피드백 등 케이스 의존적인 성격이 강하기 때문이다. 예를 들어 머신러닝의 기본 알고리즘은 금융, 의료, 교육 등 모든 분야에 응용가능하지만 개별적인 응용대상 분야 자체에 대한 노하우가 없으면 그러한 알고리즘이 실제로 구현될 수가 없는 것이다. 따라서 구글과 같은 거대 글로벌 선도기업도 혼자만의 힘으로는 인공지능 기술의 보급 확산을 달성할 수 없으며,

개방형 혁신 생태계를 통하여 다양한 분야에서 인공지능 기술을 응용하는 주체들이 모두 협력하여야 인공지능 기술의 확산 및 보급이 가능하다. 따라서 국내에서도 이러한 인공지능 알고리즘의 표준화 및 오픈소스 라이브러리화 작업이 필요하며, 인공지능 기술의 개발에 필수적인 테스트 및 트레이닝용 데이터 셋의 확보가 요구된다. 그러나 데이터의 확보에는 많은 비용이 소요되어, 규모가 작은 스타트업의 입장에서 이러한 데이터의 확보는 쉽지 않은 일이다.

정부 및 공공부문의 역할은 이러한 인공지능 기술 생태계에서 훈련용 데이터셋과 오픈소스 라이브러리, 기술평가기준 정립과 같은 공공재의 공급을 촉진하는 것에 있다. 특히 정부가 축적한 방대한 양의 데이터를 민간 스타트업 생태계와 공유하는 것이 필요하며, 이 과정에서 직면하게 되는 빅데이터의 상업적 응용 과정에서의 보안 문제나 데이터 완결성 문제, 사생활 침해로 인한 윤리적 문제 등에 대해서 가이드라인을 제시해야할 필요가 있다. 또한, 인공지능 기술의 초기 수요를 견인하여 시장을 창출해내기 위하여 정부의 공공 IT 시스템에 자연어처리 등 인공지능 기술과 인터페이스를 적극적으로 도입하는 등의 정책을 제안할 수 있다.

1. 서론

◆ AI 기술 오픈소스화 트렌드

- 구글 등 글로벌 선도 ICT 기업들은 AI 기술 관련 소스코드 및 API, 트레이닝 및 테스트베드 환경 등 개발 플랫폼을 전 세계에 무료로 공개하고 있음
 - 구글 뿐 아니라 Facebook, Microsoft, OpenAI, IBM 등도 이러한 인공지능 플랫폼 오픈소스화 대열에 동참함
 - 기존에 오픈소스에 부정적이었던 것으로 알려진 Microsoft도 OpenAI와 협력관계를 체결
 - 이러한 인공지능 개발 플랫폼들은 딥러닝이나 강화학습 등의 인공지능 기본 알고리즘들을 누구나 쉽게 구현할 수 있도록 모듈화·라이브러리화하고, 직관적인 인터페이스를 제공한다는 점이 공통점
- ※ 각 기업들의 인공지능 연구 성과를 집약한 것으로 기존에 사내 연구 개발용으로 쓰여왔던 것들임
- 이들 글로벌 ICT 선도 기업들은 많은 투자를 통해 개발한 인공지능 플랫폼을 무료로 공개하면서 자사 주도의 인공지능 혁신 생태계를 조성하려고 하고 있음
 - 딥러닝과 강화학습 등 인공지능의 기반 기술의 알고리즘은 이미 보편화되어 있으며, 이런 공통적인 인공지능 알고리즘을 모듈화하면 다양한 인공지능 관련 개발 작업의 진입장벽을 크게 낮출 수 있음

- 인공지능 구현 모듈들을 직접 개발하는 것은 많은 투자가 필요하여 스타트업이 수행하기에 벅차지만 이렇게 이미 개발된 인공지능 개발 툴이 공개되면 규모가 작은 스타트업도 손쉽게 참여가 가능함

◆ AI 기술 공개와 오픈 이노베이션

- 인공지능 기술은 기존의 전통적 ICT 산업의 테두리를 넘어 전산업 및 사회의 모든 영역에 걸쳐 응용 가능하기 때문에 고도의 융합적 속성을 지님
 - 기술융합의 트렌드에 대응하기 위하여 기업은 자신의 고유영역 이외의 새로운 전문 지식을 습득해야 하며 이는 혁신을 위하여 기업이 외부의 지식을 적극적으로 받아들여야 함을 의미함¹⁾
 - 구글과 같은 글로벌 ICT 선도 기업조차도 인공지능 기술의 광대한 영역을 사내의 혁신역량만으로 대응하기는 역부족이며, 이에 따라 개방형 AI 혁신 생태계의 필요성이 크게 부각됨
- ※ 딥러닝의 경우 context 의존적이며 매우 다양한 사례별로 별개로 접근하는 식으로, 데이터에 맞추어 소프트웨어가 개발되는 실정
- ICT 선도기업들이 공개한 인공지능 개발 툴을 중심으로, 이를 실제로 응용하는 다양한 스타트업들 및 관련 연구주체들을 포괄하는 인공지능 기술 혁신 분업구조가 형성
 - 기본적으로 핵심 알고리즘을 개발하는 것은 스타트업보다는 구글 등 선도기업의 역할
 - 구글이 인공지능 스타트업 DeepMind를 인수한 사례에서 볼 수 있듯

1) Curran et al(2011) 참조

원천기술 보유 스타트업을 인수하여 기술을 흡수함으로써 선도 기업들의 인공지능 원천기술도 업그레이드 되어온 것임

- 반면, AI 기술 플랫폼의 공개에도 불구하고 기업 보유 데이터셋의 공개는 상대적으로 부진함
 - 인공지능 기술은 알고리즘만으로는 작동하지 않으며, 대량의 훈련용 데이터를 사용하여 학습을 시켜야만 작동함
 - 학습 과정뿐만 아니라 인공지능 구현 소프트웨어의 테스트를 위해서도 역시 대량의 데이터셋을 필요로 함
 - 딥러닝과 강화학습 등의 인공지능 기술의 기본 원리는 이미 보편화되었기 때문에 진짜 기업의 핵심 경쟁력은 알고리즘보다는 보유 데이터에서 나오게 되었음
- 정부는 이러한 스타트업 생태계를 조율하고 공공재를 공급하는 역할을 수행해야 함
 - 미국 National Science and Technology Council (2016) 는 훈련용 데이터셋의 공개, 공공기관의 데이터셋 공개 및 오픈소스 툴킷 개발 등에서 정부의 역할을 명시
 - 공공 전산시스템의 인공지능 도입, 자연어인식 기반 인터페이스를 공공기관 사이트에 도입
 - 공공데이터셋을 공개하면서 데이터 공개와 투명성, 안정성 등에 대해 정부가 주도하여 표준을 설정
 - 예를 들어 지난 미국 대선에서 페이스북 가짜 뉴스 논란은 데이터 신뢰성 및 공공성에 대해 환기시키는 계기가 되었음

2. AI 플랫폼의 개념

◆ AI 플랫폼의 필요성

- 인공지능 기술의 양대 축은 알고리즘과 데이터이며, 이 중 알고리즘의 경우 다양한 문제해결에 공통적으로 적용될 수 있는 범용 툴을 개발하여 구현하는 것이 효율적
 - 현재 인공지능 기술의 주류는 딥러닝과 강화학습 등의 머신러닝 기술로, 이는 매우 대량의 데이터로부터 통계적인 최적화 연산을 수행하여 문제에 대한 솔루션을 찾아내는 방식임
 - ※ 예를 들어, 기존의 영한 번역문에 대한 방대한 데이터에서 Apple 이란 단어와 사과라는 단어가 높은 빈도로 함께 출몰한다면 Apple은 한국어로 곧 사과를 뜻한다는 답을 찾아낼 수 있음
 - 딥러닝 등의 인공지능 기술은 해당 주제에 대한 데이터로부터 통계적인 최적화 연산의 무수한 다단계 반복을 통하여 답을 찾아내는 과정이며, 이러한 연산을 최대한 효율적으로 수행하는 것이 관건
 - ※ 딥러닝이 요구하는 반복 최적화 연산에 적합한 GPU (그래픽 연산용 프로세서) 등의 하드웨어 뿐 아니라 최적화 연산 알고리즘의 성능도 중요함
 - 인공지능 기술의 적용 분야가 다르더라도 거기 적용되는 최적화 알고리즘들은 높은 수준의 공통성이 있으므로 이를 모듈화하여 다양한 케이스에 반복적 사용이 가능함
- AI 플랫폼은 알고리즘, 알고리즘을 소프트웨어적으로 구현하여 모듈화한 라이브러리 및 그러한 라이브러리의 구성 모듈들을 불러내어 도구로 사용할 수 있게 하는 인터페이스 등으로 구성됨

- 머신러닝의 복잡한 수학적 · 통계학적 알고리즘을 소프트웨어 모듈로 구현하면, 개발자들은 그 자세한 내부 로직을 알지 못하고도 해당 모듈들을 도구로 사용하여 기술구현이 가능함
- ※ 이는 계산기의 곱셈 버튼을 누르면 구구단을 몰라도 곱셈 연산을 수행할 수 있는 것에 비유할 수 있음
- 머신러닝 구현 알고리즘 모듈들의 집합인 라이브러리를 효율적으로 개발자가 활용하기 위하여 인터페이스가 존재함
- ※ 이는 C, R, Python 등의 프로그래밍 언어와 시각적인 그래픽 인터페이스를 모두 포괄함

◆ AI 플랫폼과 도메인 지식

- AI 플랫폼은 다양한 영역의 문제를 해결하기 위한 도구에 불과하며, 실제 구현을 위해선 해당 영역 자체에 대한 지식을 뜻하는 도메인 지식(Domain Knowledge)이 결합되어야 함
- 하나의 AI 플랫폼 및 이 플랫폼이 구현하는 AI 알고리즘은 음성인식, 사진인식, 자율주행차, 자동번역, 금융 및 의료 데이터 분석 등 다양한 영역에 응용 가능
- 그러나, 언어, 교통, 의료, 금융 등 AI가 적용되는 해당 영역에 대한 지식(Domain Knowledge)이 선행되지 않으면 실제로 AI 기술을 구현하는 것은 불가능
- 이러한 도메인 지식은 고도의 다양성을 가지기 때문에 경쟁력있는 AI 플랫폼을 만들어낸 기업이 도메인 지식에서도 동시에 우위에 있을 수 있다는 보장이 없음

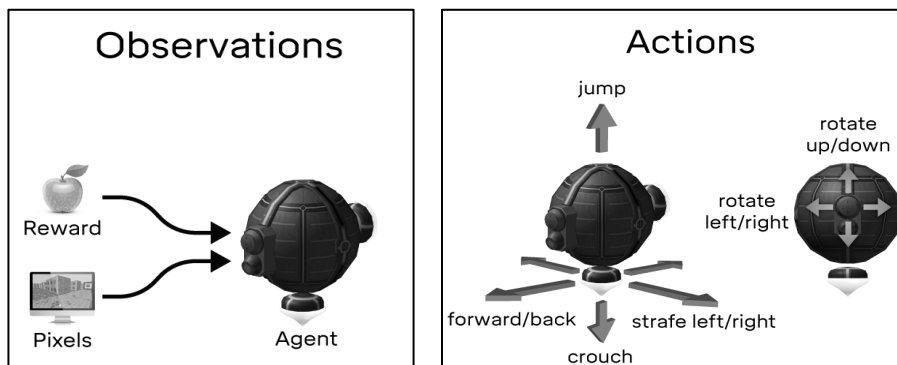
- 현재 인공지능의 주류를 이루는 딥러닝 등의 머신러닝 기술은 데이터 및 컨텍스트에 의존적이기 때문에 도메인 지식의 중요성이 더욱 큼
 - 현재의 기계학습은 특정한 상황에 국한하여 비슷한 상황에 대한 데이터를 반복 학습 시키는 식으로 구현되기 때문에, 다른 영역의 task를 수행하기 위해서는 새로운 학습이 별도로 필요함
 - 더욱이, 머신러닝은 단순히 해당 영역의 데이터를 수집하여 입력하는 것만으로 되는 것이 아니라 지도학습(supervised learning)등 인간에 의한 해석과 판단이 추가 되어야만 학습이 이뤄지는 경우가 많음

3. 인공지능 플랫폼 오픈소스화 현황

◆ 구 글

- 알파고를 개발한 구글 딥마인드는 2016년 12월 5일 내부에서 사용하던 플래그쉽 인공지능 연구 플랫폼인 딥마인드 랩(DeepMind Lab)을 공개함
 - 딥마인드랩은 3D 게임을 통해 인공지능 기술을 테스트할 수 있는 구글 딥마인드의 내부 연구 플랫폼임
 - 일인칭 시점에서 게임 속 캐릭터가 과일 줍기나 언덕 오르기 등의 다양한 동작을 수행할 수 있도록 학습시키면서 새로운 인공지능 알고리즘을 테스트할 수 있다
 - 해당 플랫폼을 무료로 개방하여 누구나 인공지능 기술을 테스트하기 위하여 활용할 수 있게 하였음

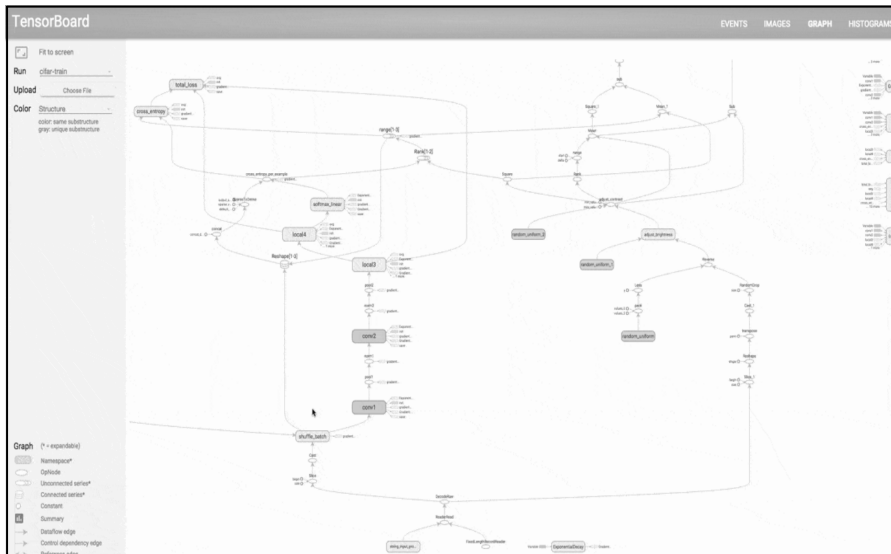
[그림 1] DeepMind Lab 개념



자료: DeepMind

- 구글은 이미 머신러닝 프레임워크인 TensorFlow를 2015년 11월 오픈 소스로 공개하였음
 - TensorFlow는 원래 구글의 인공지능 연구조직 내에서 머신러닝과 신경망연구를 위해 개발된 소프트웨어 라이브러리임
 - ※ TensorFlow는 1세대 머신러닝 플랫폼 DistBelief의 뒤를 잇는 제 2세대 플랫폼으로, 보다 유연하며 확장성이 향상됨
 - TensorFlow를 구성하는 라이브러리는 특정 하드웨어에 의존적이지 않으므로 개발자는 기반 하드웨어에 상관없이 사용이 가능함
 - ※ 스마트폰으로부터 대형 분산시스템에 이르기까지 수많은 종류의 기기에 적용 가능함

[그림 2] TensorFlow graphic visualization



자료: 구글 TensorFlow

- TensorFlow의 특징은 머신러닝의 기본이 되는 연산을 수행하면서 그 과정과 결과를 플로우차트와 그래프를 사용하여 직관적인 방식으로 디스플레이 한다는 것임.

※ 가장 직관적인 프로그래밍 언어의 하나로 불리는 Python 인터페이스를 지원하는 등 개발자에게 보다 직관적인 개발환경을 제공하는 것을 추구

- 지메일의 스팸 필터링 및 구글 포토 서비스의 이미지 자동 분류 등 이미 다양한 구글 서비스에 TensorFlow를 통해 개발된 머신러닝 기술이 들어가 있음

◆ 페이스북

- 페이스북의 인공지능 개발조직인 FAIR(Facebook Artificial Intelligence Research)는 머신러닝 프레임워크인 Torch기반으로 제작된 인공지능 모듈들을 오픈소스로 공개함

- Torch는 페이스북 뿐 아니라 구글, MS 등 많은 기업에서 사용하는 공개 인공지능 개발 플랫폼이며, 페이스북에서 사진 패턴 인식, 맞춤형 광고를 위한 패턴분석 등에 활용 됨

- FAIR에서 Torch를 기반으로 최적화한 딥러닝 모듈들을 공개하기로 페이스북은 발표하였음

- 페이스북은 딥러닝 소프트웨어 모듈뿐만 아니라, 이미지 패턴인식 등 딥러닝 연산에 최적화된 하드웨어의 설계까지도 함께 공개하였음

※ 페이스북이 공개한 딥러닝용 서버 Big Sur 는 페이스북이 이미 이미지 인식, 뉴스피드 관리, 자동번역 등에 사용 중이며 머신러닝 관련 대량 반복연산에 특히 적합한 GPU(알파고에 사용)를 다수 연결하여 활용하는 구조임

◆ 마이크로소프트

- MS는 프로젝트 옥스퍼드를 오픈소스로 공개함
 - 마이크로소프트의 프로젝트 옥스퍼드(Project Oxford)는 얼굴을 인식하는 Face API, 음성인식을 수행하는 Speech API, 이미지를 인식하는 Vision API, 자연어 인식 기능을 수행하는 LUIS 등으로 이뤄짐
 - 마이크로소프트는 2015년 10월 프로젝트 옥스퍼드의 API 개발킷을 공개하였으며²⁾, MS는 구글의 TensorFlow나 Torch 등과 비교하여 더욱 속도가 빠르고 효율적이라고 주장함

◆ OpenAI

- 비영리 인공지능연구기관인 OpenAI는 인공지능 플랫폼인 OpenAI Gym을 2016년 4월에 공개함
 - OpenAI는 테슬라자동차의 엘론 머스크와 Y-combinator의 샘 알트만 등이 총 10억달러를 공동 출자하여 2015년 12월 창설한 비영리 목적의 인공지능연구기관
 - 인공지능 기술의 안전성 및 오남용 가능성에 대한 우려는 인공지능 기술이 최대한 많은 사람에게 보급됨으로써 가장 잘 해결할 수 있다는 것이 설립 모토이며 모든 인공지능 연구 성과를 공개하는 것을 추구
 - Gym은 강화학습(Reinforcement Learning)³⁾에 강점을 가진 개발툴임

2) 2015년 4월 비상업적 목적의 사용자만을 대상으로 공개하였으나 이후 기업 포함 전체 사용자로 대상 확대

3) 강화학습은 게임 플레이어와 같은 특정한 agent가 특정한 상황속에 놓여있을 때 그 상황에서 어떻게 행동하는 것이 최적인가 (예를 들어 게임의 룰을 알려주면 자동으로 게임을 배우는 AI)를 계산하는 머신러닝의 한 형태. 자율주행차 등에 적용 가능하여 특히 주목받음.

◆ IBM

- IBM은 머신러닝 플랫폼 SystemML을 오픈소스로 공개함
 - IBM은 오픈소스 개발자 커뮤니티인 아파치의 빅데이터 시스템인 스파크 프로젝트와 제휴하면서 자사가 개발한 System ML의 소스코드를 기증하였음
 - System ML은 IBM에서 이미 오래 동안 사내에서 인공지능 기술 개발용으로 활용해왔으며, 넷플릭스(Netflix)의 영화DB의 영화 추천 기능 등에도 사용됨

◆ 삼성

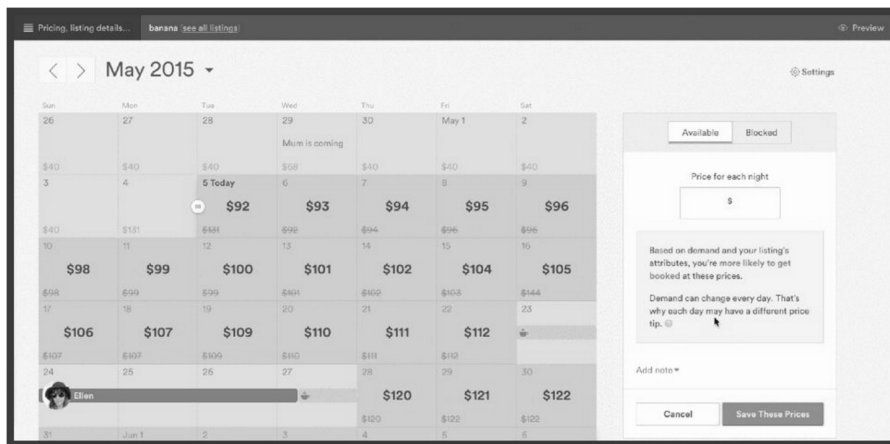
- 삼성전자는 인공지능 플랫폼 Veles를 개발하여 2015년 11월 공개하였음
 - Veles는 딥러닝 알고리즘 구현을 위한 분산형 구조의 플랫폼이며, 알파고와 유사하게 다수의 GPU를 연결한 병렬 컴퓨팅을 지원한다
 - 구글 TensorFlow와 비슷하게 플로우 형식의 그래픽 인터페이스를 지원하며, 다중 프로그래밍언어 기반이기 때문에 구글 TensorFlow와 마찬가지로 사용하기 쉬운 파이썬(Python) 기반 프로그래밍 인터페이스를 제공한다

◆ 에어비엔비

- 에어비엔비(AirBnB)도 2015년 6월 머신러닝 소프트웨어인 에어로솔브(Aerosolve)를 오픈소스로 공개함

- 에어로솔브를 이용하여 Pricing Tips 라는 기능이 개발이 되었는데, 이는 여행객의 예약 트렌드를 날짜별로 분석하여 집주인에게 알맞은 가격을 추천하는 기능임
- 에어로솔브를 활용하여 방 사진을 올릴 때 어떤 스타일로 올려야 이용객들이 좋아하는지 등을 찾아낼 수 있음
- 에어비엔비의 오픈소스 머신러닝 소프트웨어는 구글 TensorFlow와 같은 완전한 범용툴이라기 보다는 숙박업계의 특정한 니즈에 특화된 응용툴에 가까움

[그림 3] Airbnb Pricing Tips



자료: Airbnb

◆ 바이두

- 바이두는 음성인식에 특화된 딥러닝 소프트웨어 모듈인 WARP-CTC를 오픈소스로 공개함

- 오픈소스 머신러닝 라이브러리인 Torch를 활용하여 개발된 응용 모듈이며, 자연어인식에서 특히 중국어에 강점을 가짐
- 바이두는 WARP-CTC 모듈을 활용하여 음성인식 소프트웨어인 딥 스피치(Deep Speech) 2를 개발함
 - ※ Deep Speech 2는 차량내부 등 다양한 노이즈 환경 속에서 수만시간 이상의 학습을 거쳤으며, 일부 영역에서 사람보다도 음성인식 능력이 뛰어난 것으로 나타나 화제가 되었음
- 이러한 음성인식 인공지능 모듈 소스코드를 개방함으로써, 한자 입력이나 중국어 인식 등 바이두를 중심으로 한 중국의 인터넷 생태계의 이용 환경 개선을 꾀할 수 있음

4. AI 개방형 혁신 생태계 구축을 위한 과제

◆ 개방형 AI 혁신 체계의 중요성

- 인공지능 기술 알고리즘을 구현하는 기반 플랫폼을 구축하는 것은 구글 등 대기업의 역할이지만 그 기반으로 실제로 다양한 환경에서 인공지능 기술을 구현하는데 스타트업의 역할이 필수적
 - 딥러닝이나 강화학습 등 인공지능의 기반 기술들은 그 기본적인 기법이나 원리가 공통적으로 광범위하게 응용되고 있으며 따라서 일반적인 개발툴이 거의 모든 인공지능 문제 해결에 적용 가능함
 - 그러나, 그 구체적인 응용 및 문제해결에서는 해당 주제 영역에 대한 지식(domain knowledge)이 필요하고 해당 주제에 특유한 문제들에 직면함
 - ※ 예를 들어, 알파고 개발의 경우엔 개발진의 바둑에 대한 이해가 선행되어야 했으며, 금융 인공지능 소프트웨어의 경우엔 금융에 대한 지식, 의료 인공지능의 경우엔 의학적 사전지식이 필요함
 - 인공지능 기술은 전통적 ICT 산업의 범주를 넘어 사회의 거의 모든 영역에 적용이 가능하며, 이런 고도의 다양성은 구글 등 소수의 대형선도기업의 역량만으로는 도저히 대응할 수가 없음
 - 인공지능 개발 플랫폼의 응용단계 뿐이 아니라 오픈 소스화된 개발 플랫폼의 업그레이드 과정에도 개방형 혁신의 형태로 대학과 공공 연구기관 및 스타트업의 참여가 가능함

- 인공지능은 ‘모든 것에 대한 기술’에 해당하며, ICT 선도기업과 ICT 스타트업뿐만 아니라 기존의 전통산업 및 정부와 공공부문, 대학과 금융기관 등 모든 주체가 협력해야만 성공적 구현이 가능함
- 국내에서도 한국 실정에 맞는 특화된 인공지능 플랫폼 개발 및 공유 촉진 정책이 필요하며 이를 중심으로 혁신주체 간 유기적 협력이 이뤄져야함
- 인공지능 알고리즘을 구현하는 원천기술/플랫폼은 인공지능 생태계의 중심이 되지만, 이의 개발은 진입장벽이 높아 구글과 같은 글로벌 선도 ICT 기업의 우위가 확고함
 - ※ 더욱이 구글 등 글로벌 선도기업들의 인공지능 개발 플랫폼 오픈소스화 정책트렌드는 국내 업체들이 이러한 플랫폼을 직접 개발하는 것 보다는 이미 오픈소스화된 개발 플랫폼을 활용할 유인을 강화함
- 그러나, 인공지능 개발 플랫폼의 경우에도 구글 TensorFlow나 Torch 등과 같은 범용 플랫폼이 아니라 Airbnb의 숙박 최적화용 플랫폼 등 특정한 용도로 특화된 응용 플랫폼은 진입장벽이 보다 낮음
- 예를 들어, 한국어 음성인식 기술의 경우 일반적인 음성인식 알고리즘 부분은 기존 오픈소스 기술을 활용하면서도 한국어에 특유한 부분에 대한 독자 플랫폼 기술 개발 및 공개 촉진이 필요
- 공공 전산시스템에 자연어처리 등 인공지능 기술을 도입하는 등 공공 부문이 자체적으로 초기단계 기술 수요를 창출하여 인공지능 생태계의 초기 정착을 지원할 수 있음

◆ 기반 데이터 확보의 중요성

- 인공지능 기술의 확산에서 알고리즘이나 이를 구현하는 소프트웨어/하드웨어보다는 이 알고리즘을 통해 학습을 실시할 기반 데이터가 더욱 중요하게 되었음
 - 구글 등의 ICT 분야 민간 선도 기업들이 개발 플랫폼은 공개하고 있지만 보유 데이터는 공개하고 있지 않으며, 이는 스타트업에게 진입장벽으로 작용할 수 있음
- 미국 과학기술위원회(National Science and Technology Council, 2016) 보고서 등은 공공 데이터셋의 제공을 권유하고 있음
 - 인공지능 기술의 트레이닝 및 테스트를 위한 데이터셋과 테스트베드 인프라를 공공재로 간주하여, 공공 부문의 참여를 촉구
 - 또한 공공 부문은 방대한 데이터셋의 활용 과정에서의 법적 윤리적 문제에 대한 가이드라인을 제시해야 함
 - 공공 부문은 자체적으로도 거대한 데이터를 생성하는 주체이며 이를 명확하고 투명한 가이드라인을 거쳐 공개함으로써 인공지능 생태계를 활성화하는데 기여할 수 있음

◆ 인공지능 알고리즘 및 데이터의 투명성 제고

- 지난 미국 대선⁴⁾의 페이스북 가짜뉴스⁴⁾ 논란에서 보이는 것처럼, 인공

4) 페이스북을 통하여 여과없이 각 선거운동 진영의 조작된 주장들이 이용자들에게 전달되어 선거에 영향을 미쳤다는 논란임. Bloomberg (2016.11.10.)는 전통적인 언론매체라면 팩트체크 과정을 통해 충분히 걸러졌을 허위 주장이 SNS상에서는 걸러지지 않았다는 점을 문제제기하고 있음

지능을 활용한 기술들의 활용에서 내부 알고리즘의 비공개로 인하여 투명성이나 신뢰성에 의문이 제기 되는 경우가 나타남

- 예를 들어 페이스북의 뉴스피드나 맞춤형 광고에서 높은 우선순위로 추천되는 정보가 과연 그럴만한 가치가 있는 것인가? 조작가능성은 없는 것인가? 라는 의문이 제기 가능
- 소비자 입장에서 블랙박스와의 같은 인공지능 알고리즘의 작동 결과 제시되는 결과에 대해 소비자가 어떤 통제권을 가질 수 있는가? 라는 의문임
- 엘론 머스크 등이 OpenAI를 설립하고 모든 인공지능 기술의 공개화를 주장한 이유도 이 것이었음

◆ 기술 표준과 평가기준의 중요성

- 딥러닝 등의 인공지능 관련 기술들은 대부분 설명가능성(explainability)와 투명성(transparency)의 문제가 있음
 - 딥러닝 기술은 블랙박스와 같아서 학습시킨 데이터를 기반으로 통계적으로 답을 찾아내지만, 왜 그런 결과가 나왔는지 사용자 입장에서 이해하기 어려움
 - ※ 예를 들어 알파고 이세돌 대결에서 NHK 기자는 이해할 수 없는 수가 묘수로 드러났는데, 이런 똑같은 일이 인공지능이 환자진료에 적용되었을 때 발생되면 어떡할 것인가? 라는 질문을 던진 적이 있음
 - 개별 소비자가 인공지능 관련 제품의 신뢰성을 확인하기는 불가능에 가깝기 때문에 인공지능 기술이 적용된 제품이 정확하게 작동하는지의 신뢰성 여부에 대한 객관적인 공인 기준이 필요함

- 게다가 개방형 생태계라는 특성상 인공지능 적용 제품에서 문제가 발생하면 어느 단계에서 문제가 발생한 것인지 식별하기 쉽지 않을 수도 있음
- 다양한 주체들이 참여하는 인공지능 생태계에서 개발된 인공지능 기술을 평가하기 위한 표준(Standard)와 평가기준(Benchmark)가 필요하며, 여기서 공공부문은 주도적인 리더십 발휘가 필요함
- NIST(National Institute of Standards and Technology)가 주도한 TREC(Text Retrieval Conference)와 FRVT(Face Recognition Vendor Test)를 공공부문이 주도한 benchmark 설정의 성공사례로 꼽힘⁵⁾

5) National Science and Technology Council (2016,10)

참 고 문 헌

- 삼성 VELES, <https://velenet.ml/Airbnb-Airflow>, <http://nerds.airbnb.com/airflow/>
- Bloomberg, "Facebook and Twitter Contend With Their Role in Trump's Victory"
https://www.bloomberg.com/news/articles/2016-11-09/facebook-and-twitter-contend-with-their-role-in-trump-s-victory?utm_source=MIT+TR+Newsletters&utm_campaign=8fd52cc4f3-newsletters-the-download&utm_medium=email&utm_term=0_997ed6f472-8fd52cc4f3-153925993&goal=0_997ed6f472-8fd52cc4f3-153925993&mc_cid=8fd52cc4f3&mc_eid=86e9a84406
- Curran, C.-S., & Leker, J. (2011), Patent indicators for monitoring convergence – examples from NFF and ICT, *Technological Forecasting and Social Change*, 78(2), 256–273.
- DeepMind Lab, <https://deepmind.com/blog/open-sourcing-deepmind-lab/>
- FAIR (Facebook Research),
<https://research.fb.com/fair-open-sources-deep-learning-modules-for-torch/>
- IBM SystemML,
http://researcher.watson.ibm.com/researcher/view_group.php?id=3174
- Microsoft Project Oxford, <https://staging.www.projectoxford.ai/>
- MIT Technology Review, "China's Baidu Releases Its AI Code",
<https://www.technologyreview.com/s/545486/chinas-baidu-releases-its-ai-code/>
- National Science and Technology Council (2016.10), "The National Artificial Intelligence Research and Development Strategic Plan"
- OpenAI, <https://openai.com/blog/>
- TensorFlow, <https://www.tensorflow.org/>
- Torch, <http://torch.ch/>

