

2023.3.29

Everything Everywhere All at Once

시가 불러온 新산업 혁명

미래에셋증권 디지털리서치팀

박광남 kwangnam.park@miraesasset.com

한종목 chongmok.han@miraesasset.com

박수진 soojin.park.a@miraesasset.com

윤재홍 jaehong.yoon@miraesasset.com



머리말

ChatGPT가 등장하면서 본격적인 AI의 시대가 열리고 있습니다. 우리는 AI가 반짝하고 잊혀진 수많은 기술처럼 단기 트렌드로 끝나지 않을 것이라고 생각합니다. AI는 개인의 삶과 산업지형에 큰 변화를 몰고 올 것이며 우리는 적응해야만 합니다.

1장에서 우리는 생산성을 높이고 고용 시장에 영향을 미치며 산업을 재편할 수 있는 AI의 잠재력을 설명했습니다. 우리는 AI가 제공하는 기회를 활용하기 위해 재교육과 투자를 통해 이러한 변화에 적응하는 것이 중요하다고 강조했습니다.

2장에서는 딥러닝의 개념과 특성을 중심으로 AI의 핵심 기술과 발전 과정을 설명했습니다. 지도 학습, 비지도 학습, 자기 지도 학습, 강화 학습 등 주요 학습 방법을 정리해 보았고, AI 알고리즘과 최신 기술 발전에 대한 포괄적인 이해를 제공하기 위해 트랜스포머와 같은 최신 유행하는 AI 기술을 소개했습니다. AI기술을 이해하는 것이 그 의미를 파악하는데 중요하기 때문에 AI의 잠재적인 파급 효과를 이해하는 데 도움이 되도록 기술적인 내용을 포함했습니다. 이 장은 어려울 수 있지만 AI의 미래 방향과 산업 전반에 걸친 다양한 변화를 예측할 수 있는 토대가 되기를 기대합니다.

3장에서는 생성AI의 출현, 급변하는 투자 환경, 그리고 AI가 변화시킬 산업의 밸류 체인을 분석했습니다. 특히 파운데이션 모델을 바탕으로 새롭게 만들어질 AI 생태계와 밸류 체인에 주목하며 미들웨어와 비즈니스 솔루션, 어플리케이션과 같은 다양한 AI 기반 서비스 시장이 활성화될 것으로 예측했습니다. 또한 게임, 블록체인, 디지털 트윈 및 자동화와 같은 메타버스 산업에서 AI로 인해 새로운 기회가 생길 다양한 사례를 설명했으며, 바이오 헬스케어와 자율주행과 같이 인간의 삶에 직접 영향을 미치는 분야와 산업을 다뤘습니다. 또한 실제 기업의 예를 통해 AI 기술이 기업의 비용을 줄일 수 있는 내용도 정리하였습니다.

4장에서는 AI의 발전에 있어 필수적 요소인 데이터, 알고리즘, 하드웨어(반도체) 등 3가지 핵심 요소에 대한 경쟁력을 가진 기업에 대한 내용을 다루었습니다. 마이크로소프트, 구글, 애플, 메타와 같은 빅테크 기업들이 데이터와 알고리즘 분야에서 경쟁력을 발휘하며 AI 서비스의 경쟁 환경과 새로운 기회를 살펴보았습니다.

또한 AI 모델의 학습, 추론 및 효율적 운영에서 컴퓨팅 파워의 중요성을 강조하였고, GPU, NPU, 뉴로모픽 등 하드웨어 반도체의 고성능 저전력 구현의 중요성과 향후 발전 방향을 전망했습니다. 추가로 서비스와 소프트웨어 및 하드웨어 간 상호 연결성이 갈수록 중요해지고 있어 네트워킹 관련 내용도 정리하였습니다.

5장에서는 중국의 AI 경쟁력과 발전 방향, 주요 기업의 밸류 체인을 정리했습니다. 6장에서는 본문 내용을 기반으로 AI 테마와 생태계 시장에 접근할 수 있는 ETF를 정리했습니다. AI 시장이 초기 성장 단계인 만큼 시장을 주도하는 기업의 변화가 빈번하게 일어날 수 있는 가능성을 고려할 때, ETF를 통한 접근이 효과적이라고 판단합니다.

마지막 장에서는 현재 ChatGPT의 한계점과 시사점에 대해 정리했습니다. 결론적으로, 우리는 AI가 세상에 많은 변화를 가져올 것이라고 생각합니다. 그러나 우리의 예측대로 진행되지 않을 수도 있습니다. 너무 과도한 낙관이나 비관 대신, 변화하는 세상에 잘 대응해 나가는 것이 우리가 할 수 있는 최선의 선택이라고 생각합니다.

* 머리말은 ChatGPT(GPT-4)를 활용해서 요약 정리하였습니다.

CONTENTS

I. 우리가 AI에 주목하는 이유 4

- AI가 만들어갈 성장과 생산성 혁신 4
- AI가 가져올 고용시장의 허리케인 6
- 더욱 가속화될 AI의 침투 7
- ChatGPT의 등장에 따른 우리들의 자세 9

II. 주요 AI 기술 정리 10

- 딥러닝의 개념과 성능 10
- AI 학습의 네 종류(지도/비지도/자지도/강화) 16
- 최신 유행 중인 AI 기법들 36
- 이슈 중심에 있는 언어모델 45

III. AI가 만들 변화는 “Everything, Everywhere” 53

- 생성형 AI의 출현은 공포스러운 경이로움 53
- AI 산업화 시대의 밸류체인 64
- 메타버스는 AI 진보의 최고 수혜 영역 74
- 실생활에도 엄청난 활약을 할 AI 기술 98
- AI는 비용을 줄이기 위한 최고의 선택 116

IV. AI 모델 구축을 위한 주요 요소와 기회 122

- 전세계 빅테크들의 전쟁 122
- 생성형 AI를 위한 연산허브는 데이터센터 143
- 누가 연산 효율화에 진심인가 152
- 데이터센터는 결국 반도체의 구성물 158

V. [중국] 중국도 ChatGPT 열풍 186

- 바이두로 포문을 연 '중국판 ChatGPT' 186
- 중국 AI 산업 밸류체인과 향후 전망 190

VI. AI 생태계, ETF로 접근하기 193

- AI ETF 개요 193
- 주요 ETF 상세 199

VII. AI의 한계와 미래 203

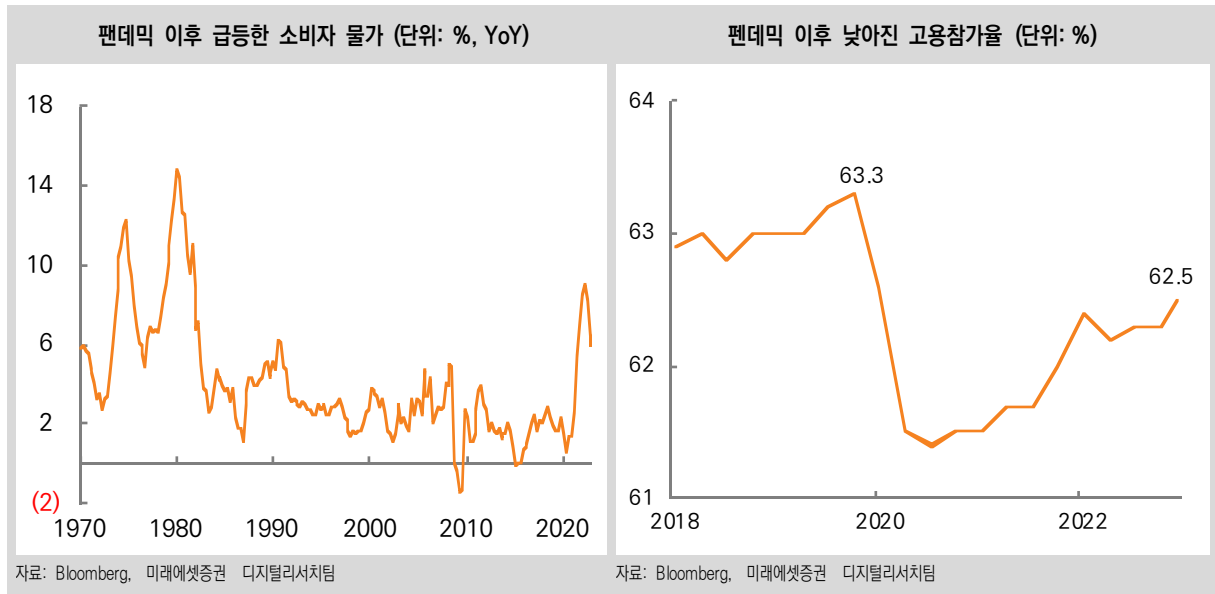
- 인공지능 vs 인간지능 203
- 인간의 존재론(Ontology)에 관한 소고 205
- 러다이트 운동이 아니라, 대비하고 공생하자 208

부록. 딥러닝 주요 개념 설명 210

I. 우리가 시에 주목하는 이유

AI가 만들어갈 성장과 생산성 혁신

미래에셋증권 디지털리서치팀은 AI가 만들어갈 성장과 생산성 혁신에 주목한다. 우리는 팬데믹 이후 무수히 많은 문제에 직면하고 있으며, 팬데믹을 극복하기 위해서 풀었던 유동성으로 높은 인플레이션 문제를 겪고 있다. 또한, 팬데믹 기간동안 고도의 디지털화가 만든 사회 문화적인 변화, 예를 들면 -조기은퇴를 선택한 사람들이 폭증한- 세상에 속에 살고 있다.

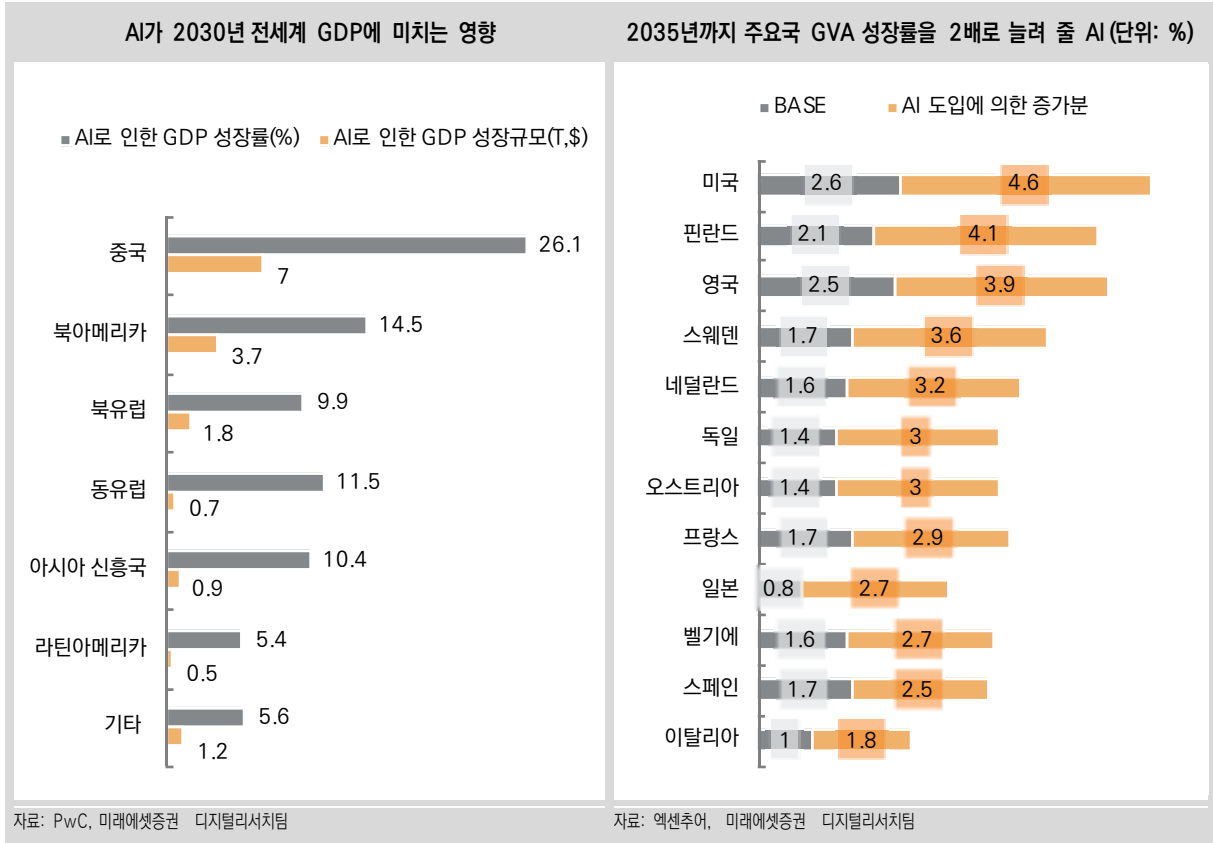


이러한 변화 속에서 글로벌 경제가 인플레이션 억제에 어려움을 겪게 되자, 글로벌 중앙은행은 강력한 긴축정책을 통한 수요 감소와 같은 고통스러운 해결책을 제시했다. 그 결과 현재 미국 내 일부 지역 은행과 유럽의 크레딧 스위스와 같은 약한 고리들이 무너지는 부작용이 발생했다.

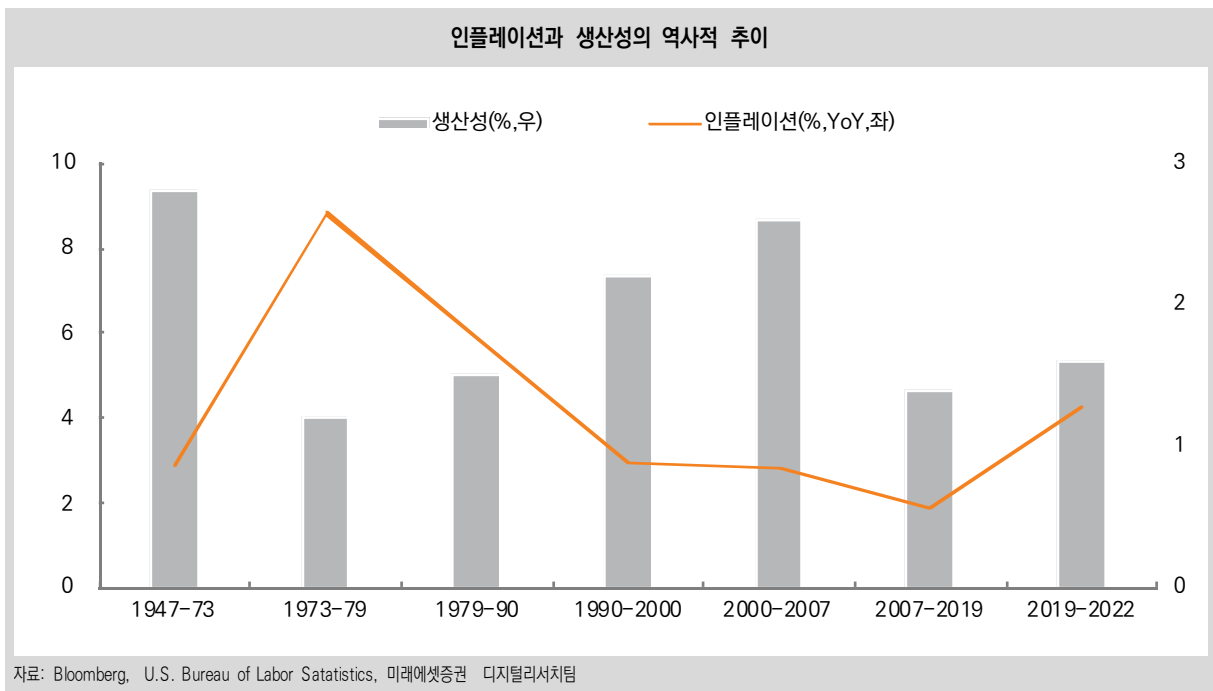
우리는 인플레이션을 제어하면서 부작용을 최소화하기 위해선 성장을 도모하는 것이 가장 훌륭한 해결책이라고 생각한다. 과거 2008년 금융위기 당시 스마트폰으로 인한 모바일 혁명이 금융위기를 극복할 수 있었던 원동력이 되었던 것처럼, AI가 다음 성장을 이끌며 산재한 문제점들을 해결하는데 기여할 수 있다고 기대한다.

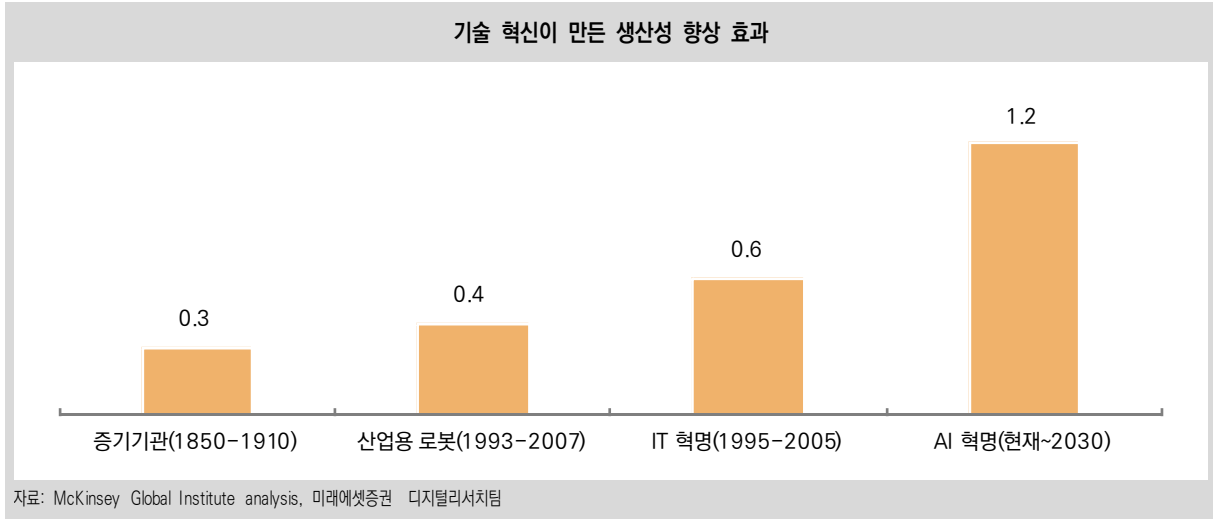
글로벌 리서치 업체들의 전망에 따르면 2030년까지 AI 기술로 인해 글로벌 주요국들의 국가별 성장률이 5~26%까지 폭증할 것으로 전망하고 있다. 또한, AI 도입으로 2035년까지 주요 12개 국가의 GVA(Gross Value added) 성장률은 2배로 증가하고, 기업들의 수익성은 평균 38% 가량 향상되며 2030년까지 무려 14조 달러 규모의 경제 성장을 가져올 것으로 예상하고 있다.

게다가 우리는 AI가 경제적인 성장뿐만 아니라 팬데믹이 야기한 노동력 부족과 높은 인플레이션과 같은 문제점을 해결하는데도 큰 역할을 할 수 있을 것으로 판단한다. 왜냐하면 노동력 부족과 인플레이션을 일거에 해결하기 위해선 생산성 증대만큼 좋은 해결책이 없기 때문이다.



역사적으로 생산성의 향상은 인플레이션을 억제하는 역할을 톡톡히 했다. 글로벌 조사업체 McKinsey의 조사에 따르면 증기기관, 로봇화 및 IT 혁명과 같은 이전의 혁신은 연간 0.3%~0.6% 범위의 생산성 향상을 가져왔다. 그리고 AI는 2030년까지 과거 혁신에 2배~4배에 이르는 연간 1.2%의 생산성 향상을 제공할 잠재력이 있다고 평가했다. 증기기관과 로봇화, IT 혁명도 엄청난 변화를 만들었지만 AI가 만들어낼 변화는 더욱 극적일 것으로 보인다.



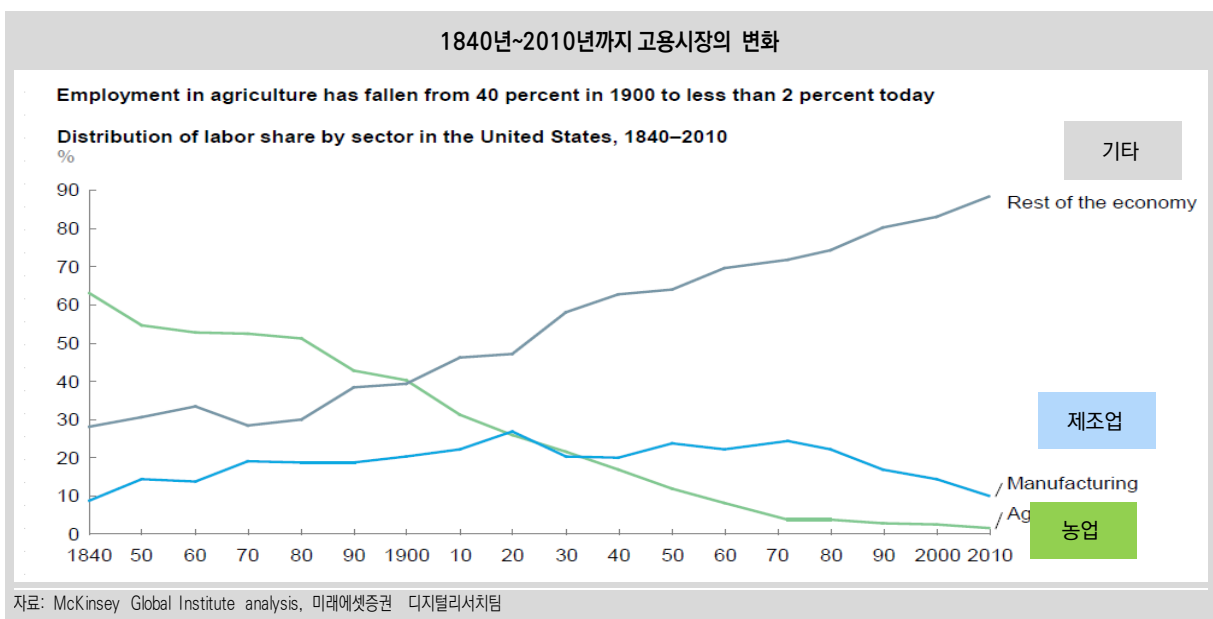


물론 현재 시점에서 AI가 시장의 전망만큼 빠르게 고성장을 할지 여부를 정확하게 예측하기는 어려울 수 있지만, 분명한 것은 향후 미래의 성장은 AI기술에 기반한 산업이 주도할 것이라는 점이다. AMD의 리사 수 CEO는 최근 인터뷰에서 AI를 한마디로 요약해달라는 요청에 “앞으로 10년을 바라봤을 때 가장 중요한 것” 이라고 답했다.

AI가 가져올 고용시장의 허리케인

우리가 AI에 대해서 주목하는 두 번째 이유는 고용시장에 미칠 파급 효과 때문이다. 과거에도 새로운 기술이 일자리를 대체하거나 새로운 직업을 만들어냈던 경험은 많았다. 하지만 AI 기술이 이전 기술 혁신과 구별되는 한 가지 핵심은 인간이 제공할 수 있는 물리적인 노동력뿐만 아니라 인간의 고유영역으로 여겨지던 인지 및 판단과 관련된 영역들을 대체할 수 있는 잠재력을 가졌다는 점이다.

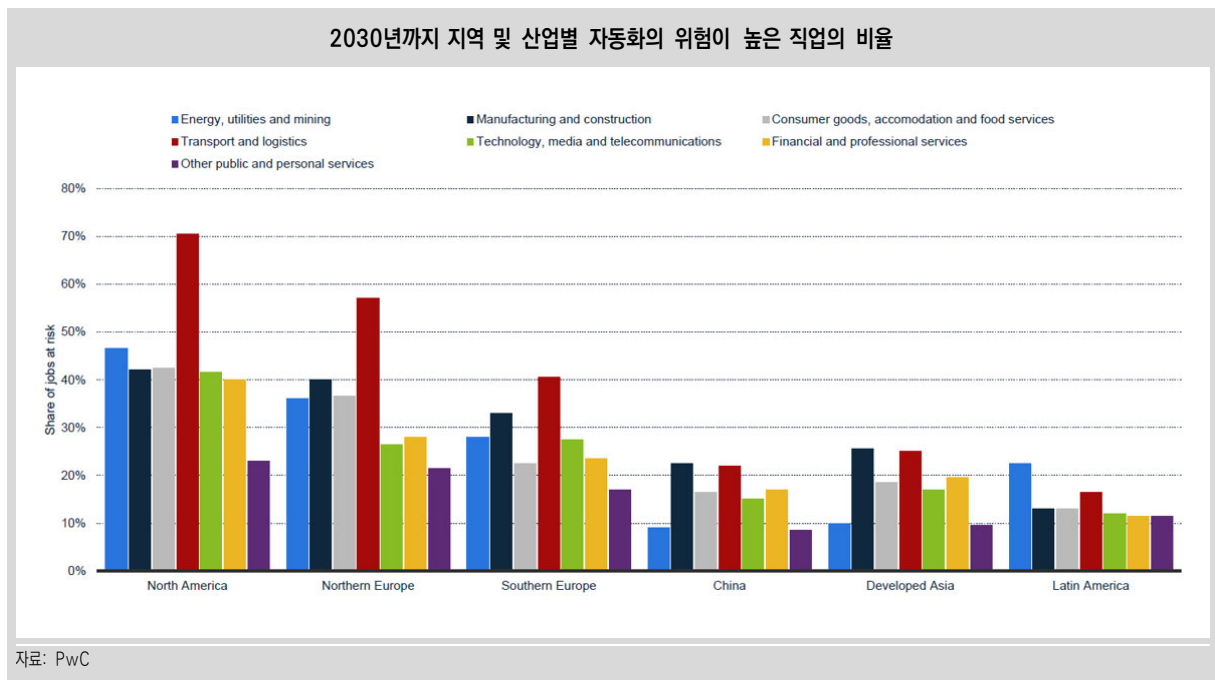
과거 역사적 흐름을 살펴보면 1900년대 미국 고용의 40%를 차지했던 농업은 2000년 들어서 2%로 급감했으며, 로봇과 세계화 등으로 제조업 고용 비중도 1950년대 약 25%에서 2010년에 10% 미만으로 떨어졌다. 하지만 주로 인간의 물리적인 노동력을 대체하는 수준에 그쳤기 때문에 이는 인간의 고유 영역으로 여겨지는 인지 및 판단능력이 필요한 새로운 일자리로 자연스럽게 대체될 수 있었다.



하지만 인지와 판단능력을 갖춘 AI의 등장으로 인해 농업과 제조업의 사례와 다른 고용 변화에 직면할 가능성이 생겼다. 즉, 현재 사람이 하고 있는 대부분의 직업이 대체 위험에 노출될 수 있다는 것이다. 물론 AI가 만들어낼 변화로 인해서 미래에 새로운 활동과 일자리가 생겨날 수도 있지만, 과거에 경험했던 패턴과 다르게 일자리의 총량이 줄어들 수 있다는 불확실성은 더욱 커진 것으로 보인다.

일례로 최근 발표된 OpenAI의 논문에 따르면 2,087개의 업무(컴퓨터 시스템 감독, 질병 진료, 영업 및 거래 등)를 대상으로 GPT-4가 미치는 영향에 대해서 분석한 결과, GPT-4를 이용하면 사람과 비슷한 완성도로 작업을 완수하는데 드는 시간이 50% 이상 감소한 직업(occupation)과 작업(task)이 약 14~46%에 달했다. 즉, 전형적으로 인간만이 할 수 있을 것으로 여겨지던 많은 분야에서 이미 인간을 대체해 버린 것이다.

다른 예로 글로벌 리서치 업체 조사에 따르면 2030년까지 북미와 유럽에서 AI에 의해 높은 위험에 처한 업종별 직업의 수가 무려 23~76%에 이를 것으로 전망된다. 변화는 이미 다가오고 있다.

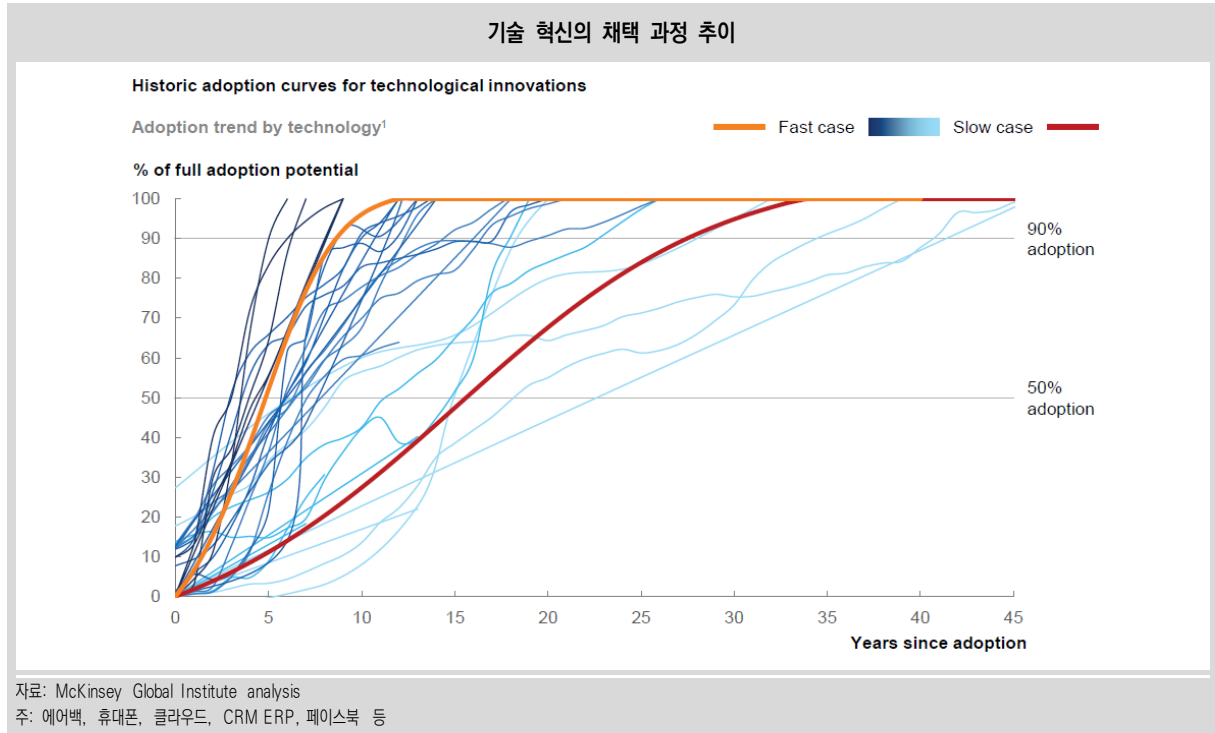


우리는 장기적으로 AI가 만들어낼 변화로 인해서 새로운 활동과 일자리 증가 등 긍정적인 효과가 나타날 수 있다고 기대한다. 하지만 과거의 그 어떤 기술 진보 케이스보다 AI의 파급효과가 거대할 수 있다는 점을 간과해선 안된다. 피할 수 없는 변화라면 AI를 기피할 것이 아니라 활용하는 방향으로 인식을 빠르게 전환해야 하며, 재교육과 AI와 연관된 기업에 대한 투자를 통해 변화로 인한 리스크를 헤지할 필요가 있다는 판단이다.

더욱 가속화될 AI의 침투

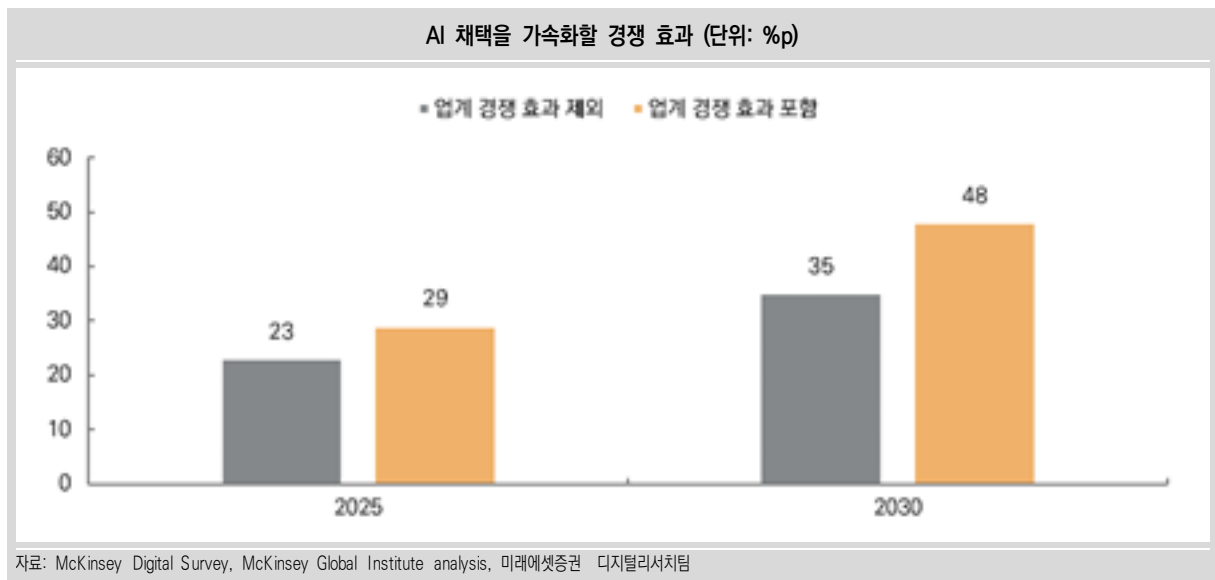
우리가 AI에 주목하는 세 번째 이유는 앞선 두가지 이유를 포함해 AI의 채택이 예상보다 더 가속화될 가능성 때문이다. 일반적으로 새로운 기술의 채택 및 확산 과정은 S-곡선을 통해 가정하며 도입기, 성장기, 성숙기, 쇠퇴기의 네 가지 단계로 구성된다.

현재 AI 기술은 도입기를 지나 성장기 초입에 진입했다고 판단한다. 새로운 기술이 채택되는 과거의 여러 가지 케이스 중 보수적으로 중간 정도의 채택 속도를 가정하면 AI 기술은 5~10년 이후부터 가파른 성장기에 진입한다고 예상해 볼 수 있다.



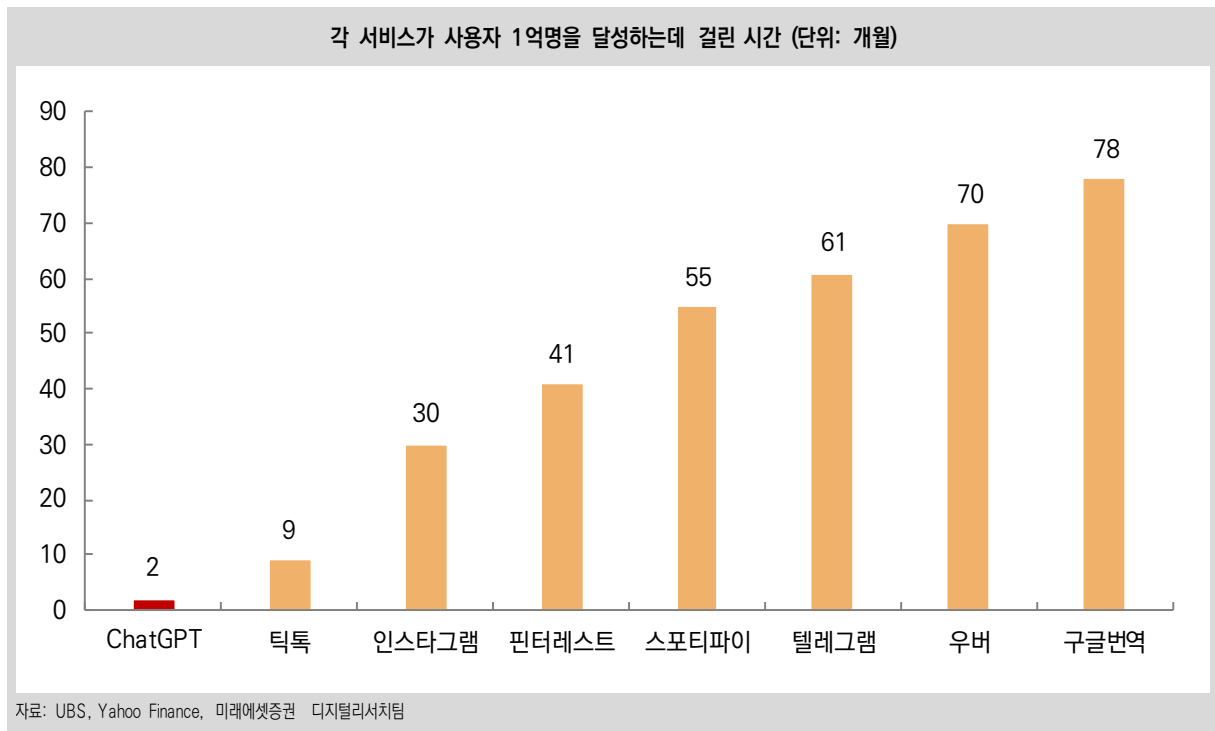
이에 대한 근거로, 우리는 AI 기술의 수많은 혁신 중에서 2012년 이미지넷 챌린지에서 등장한 CNN 기반 딥러닝 알고리즘과 2017년 자연어 처리를 위해 구글이 개발한 트랜스포머 알고리즘에 주목했다. 왜냐하면 두 기술의 결합으로 사람의 방식과 유사하게 학습과 추론을 할 수 있는 ChatGPT와 같은 생성형 AI 또는 초거대 AI가 출현했기 때문이다.

특히 ChatGPT와 같은 AI 서비스가 일반 소비자들에게 직접적인 영향을 미친 첫번째 케이스라는 점에서 의미가 깊다고 본다. ChatGPT 등장에 큰 기여를 한 CNN기반 딥러닝 알고리즘과 트랜스포머의 등장이 이미 5~10년 전에 시작되었다는 점은 AI가 본격적인 성장기에 진입했다고 유추해 볼 수 있는 대목이다.



또한 Mckinsey는 기업간 경쟁이 심화될 경우 AI의 채택이 더욱 빨라질 수 있다고 전망했는데 최근 ChatGPT가 일으키고 있는 엄청난 반응에 주목할 필요가 있다. 사실 AI에 대한 기술력을 논할 때 우선순위에서 항상 선두에 있던 기업은 마이크로소프트가 아닌 구글이었다. 하지만 ChatGPT라는 세상에 의미 있는 서비스를 먼저 공개한 것은 OpenAI와 협업한 마이크로소프트였다.

단적으로 ChatGPT가 1억 명의 MAU를 달성하는데 고작 두 달에 걸리지 않았다는 점에서 얼마나 큰 인기를 끌고 있는지 알 수 있다. ChatGPT 서비스를 본 순다르 피차이 구글 CEO는 발등에 불이 떨어져 사내에 '코드레드' 를 발령했으며, ChatGPT에 대한 대응전략을 주문했다. 하지만 급하게 출시했던 생성형 AI 바드(Bard)가 잘못된 정보를 전달하자 주가가 급락하는 경험을 하기도 했다. 이렇듯 거대 기업들의 경쟁이 본격적으로 시작된다면 AI의 침투 속도는 더욱 빨라질 수 있을 것이다.



ChatGPT의 등장에 따른 우리들의 자세

ChatGPT의 등장으로 세상이 들썩이고 있다. 하루 밤만 지나도 무수히 많은 ChatGPT 기반 서비스들이 쏟아져 나와 곁할기 수준으로도 쫓아가기조차 버거울 지경이다. 앞으로는 AI가 모든 영역에서 인간을 대체할 것이라는 이야기도 심심찮게 나오고 있다. 물론 어떤 새로운 기술이 등장하고 빠르게 각광을 받았다가 순식간에 사라졌던 많은 경우는 많기 때문에 너무 과도한 낙관론은 주의해야 한다.

하지만 우리는 이러한 AI 혁명이 단기 이벤트가 아닌 빅 트렌드로서 세상에 그리고 투자 환경에 큰 변화를 줄 것으로 기대한다. 최근 GTC 2023에서 젠슨 황 CEO는 AI의 미래에 대해서 이렇게 언급했다.

“AI will soon be everywhere and accessible to everyone(누구나 사용할 수 있고 어디에나 있는 AI)”. 피할 수 없다면 빠르게 인정하고 받아들여야만 한다.

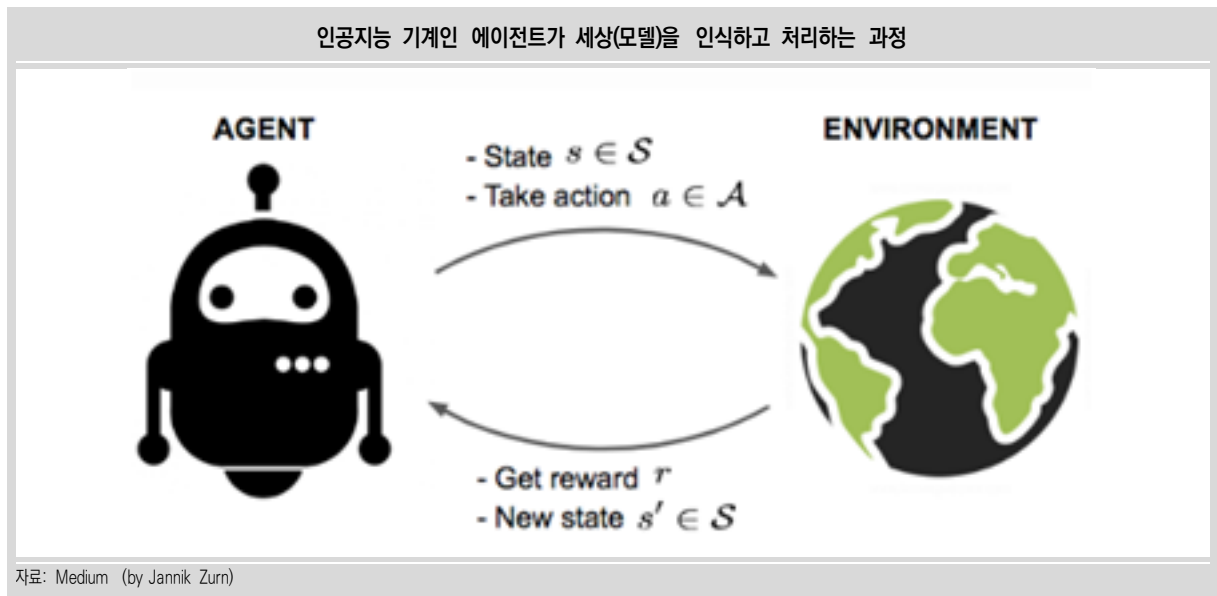
II. 주요 AI 기술 정리

딥러닝의 개념과 성능

1) 인공지능의 본질과 딥러닝

인공지능은 기본적으로 컴퓨터로, 본질은 계산과 출력이다. 그리고 인공지능 기능을 갖춘 “계산기”를 우리는 에이전트(agent)라고 말한다. 인공지능 에이전트가 요새 글자도 번역하고, 영상도 요약해주고 다 읽어내는 것 같아도, 사실은 숫자밖에는 모른다. 그런데 이 세상 모든 것이 숫자로 이루어져 있기에, 인공지능이 모든 곳에, 모든 것에 적용되고 있는 것이다. 숫자를 계산하고 보관하는 반도체부터, 숫자를 전송하는 인터넷, 숫자로써 정립되는 파일 등등 모든 게 숫자놀음이다.

여기서 에이전트가 자연어를 알아듣기 위해서는, 수치계산을 수행하기 위한 함수가 필요하다. 그 함수를 인공지능에서는 모델이라고 부른다. 그리고 함수의 결과값을 출력할 때, 출력값을 확률형태의 숫자로 표현한다. 즉, 인공지능 모델은 어떤 입력을 받아서, 확률로 반환하는 함수다.



머신러닝은 인공지능 모델 중 하나로, 컴퓨터 프로그램이 경험을 통해 스스로 학습하는 능력을 가지고 있다. 머신러닝은 데이터를 이용해 모델을 학습하고, 이를 이용해 새로운 데이터의 분류(Classification), 회귀(Regression), 군집화(Clustering)와 같은 작업을 수행한다. 예를 들어, 이메일 스팸 필터링은 입력 데이터(이메일)를 분류하는 문제로, 이를 위해 모델을 학습시키는 것이다. 모델은 레이블이 지정된(spam or not spam) 이메일 데이터를 학습하고, 이를 이용해 새로운 이메일을 분류한다.

인공지능의 개발 목표는 다양한 일들의 특성을 찾아 수행할 수 있는 컴퓨터 시스템인 “에이전트”의 개발을 의미한다. 여기서 특성(feature)은 알고리즘이 주어진 데이터를 분류하는 기준이 된다. 예를 들어, 인간을 분류할 때 성별, 인종, 나이 등 여러 ‘특성’을 바탕으로 분류할 수 있다. 그러니 우리는 인공지능 연산을 하는 목적과 입력 데이터의 특성(feature)에 맞게 최적이라고 생각되는 적절한 모델을 고르면 된다.

물론, 단연코 요즘 가장 인기 있는 모델은 딥러닝이다. 성능이 월등하기 때문이다. 딥러닝이란 데이터 패턴을 스스로 익히는 인공지능 메커니즘이자 머신러닝의 한 분야다. 다시 말해, 머신러닝과 달리 딥러닝은 사람이 알고리즘으로 짜는 것이 아닌 “인공 신경망”을 이용해서 만들어진다. 요새 보이는 모든 신기한 인공지능 서비스들은 다 딥러닝을 바탕으로 만들어졌다. 성능이 월등하다는 것은, 딥러닝이 기존의 머신러닝 알고리즘에 비해 높은 정확도를 보인다는 것이다. 딥러닝 모델이 복잡한 데이터 구조를 처리할 수 있는 능력을 가지기 때문이다.

***머신러닝:** 주어진 데이터를 알고리즘을 통해 컴퓨터가 학습하는 과정. 데이터를 학습하는 방식에 따라 지도학습과 비지도학습, 자기지도학습, 강화학습으로 나뉜다

***딥러닝:** 머신러닝의 일종으로, 비정형 데이터를 처리하도록 하기 위해 인간의 두뇌와 유사한 인공신경망(ANN: Artificial Neural Network) 구조를 가진 것이 특징

2) 왜 딥러닝이 더 잘할까?

설명을 돕기 위해, 이 세상은 수많은 비정형 데이터(unstructured data)로 이뤄져 있다는 것을 생각해 보자. 그 반대말인 정형 데이터는 날짜와 핸드폰번호, 주소 등 정해진 형식으로 이뤄진 정보들이다. 하지만, 텍스트와 이미지, 그리고 음성 데이터 등은 전부 정형화된 포맷이 없다. 다시 말해 프로그래머들이 기호 형식(AND, OR, NOR, NAND, XOR 등)으로 알고리즘 플로우 차트를 그릴 수 없다는 것을 의미한다. 이런 작업은 원래 프로그래머들이 짜는 것이었다. 그런데 음성 주파수 데이터를 텍스트로 바꾸는 논리와 절차를 인간은 찾을 수가 없다.

이런 논리와 절차를 인공지능 학습에서는 매개변수(parameter, 파라미터)라고 부르고 있다. ChatGPT의 경우, GPT-3.5를 기반으로 했을 때 파라미터 숫자가 1,750억 개이고, GPT-4를 기반으로 했을 때는 조 단위에 육박한다. 수조개의 파라미터를 사람이 일일이 찾는다는 것은 어불성설이다. 따라서 이 세상을 모델링해서 이해하기 위해, 비정형 데이터를 사용해 매개변수를 찾아내려면 반드시 딥러닝을 사용해야 한다. 그리고 딥러닝의 목적은 그 매개변수의 최적값을 찾는 것이다.

***은닉층(Hidden layer):** layer의 개수는 지정할 수 있으나, 그 안에서 어떻게 계산이 진행되는지, 각 노드가 어떤 계산을 거치는지 등은 알 수 없음. 딥러닝을 구축할 때 주요 고려요소는, “Layer를 몇 층을 쌓을 것인가”와 “한 Layer에 몇 개의 연결선(Parameter)로 만들 것인가”.

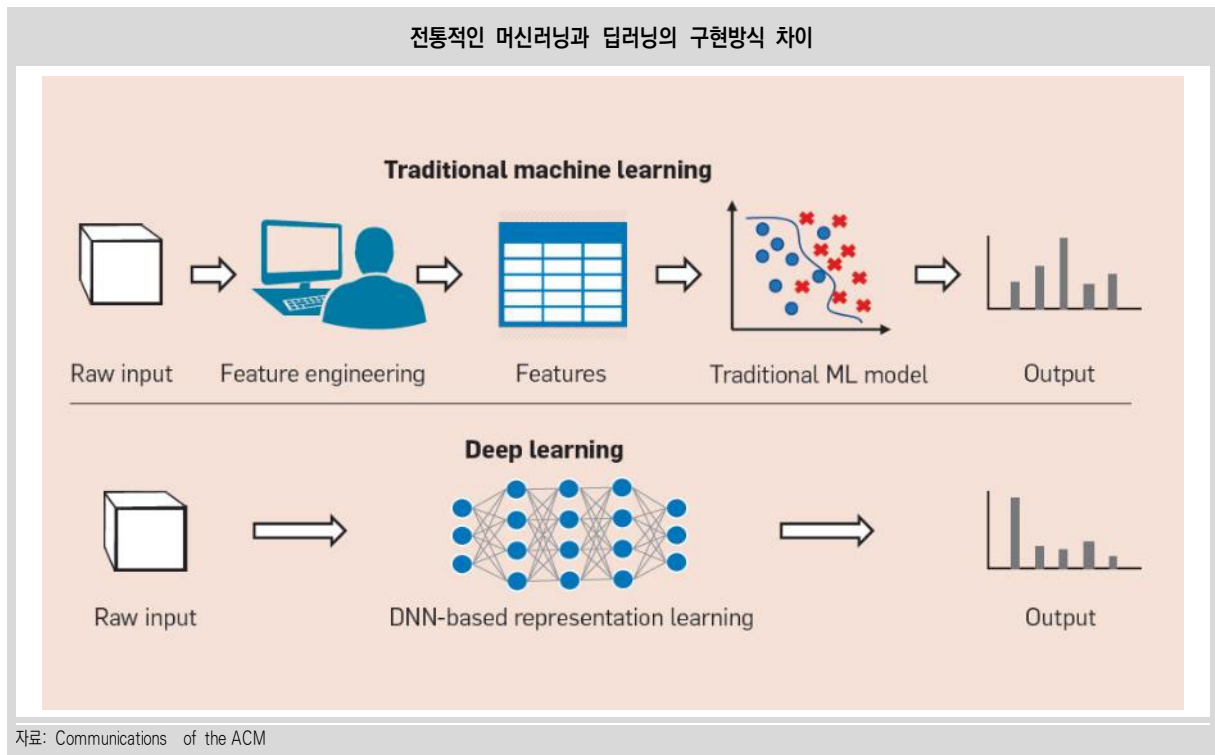
*** 매개변수(Parameter):** 매개변수는 알고리즘이 학습하는 과정에서 문제를 해결하기 위해 조절하는 변수들. NN의 가중치 및 bias, 회귀 모형의 coefficient 등 알고리즘이 샘플을 학습하며 조정해 나가는 모든 것을 의미. 알고리즘의 목적은 최적의 파라미터를 찾는 과정.

*** 초매개변수(Hyperparameter):** 초매개변수는 알고리즘에 학습을 시키기 전에 인간이 설정하는 값들, 예를 들어 은닉층의 개수, 샘플 데이터의 개수, (CNN의 경우) pooling 크기나 filter 크기 등이 대표적. 이것들의 최적 값은 경험적으로 알게 되는 “경험적인 법칙”으로 결정되는 경우가 많음

딥러닝은 다양한 층(layer)으로 이루어진 인공 신경망을 이용해 데이터를 학습한다. 딥(deep)이라는 단어 자체가 모델에 있어서 은닉층(Hidden Layer)를 사용한다는 의미다. 그런데 정말 재미있는 것은 은닉층이라는 말 자체에서 알 수 있듯이, 딥러닝이 어떤 답을 내는 이유(최적의 매개변수를 찾아낸 이유)를 수학적으로 아직 밝혀내지 못했다는 점이다.

현재 인류의 상태는 “딥러닝의 정확한 원리는 모르지만 일단 해보니 되니까 하는 것”이다. 그래서 생각하지 못했던, 기대하지 않던 능력도 덩달아 나온다.

딥러닝 전문가들의 말은 모아보면 “왜 되는지를 모르니까 어디까지 될지도 모르겠다”라는 것이다. 실제로 현업에서 딥러닝을 사용하는 사람들 또한, “사람이 간섭하던 것들을 뺄수록, 사람의 의견을 배제할수록 딥러닝 연산이 잘 된다”고 한다. 어쨌든, 이런 복잡한 데이터 구조와 심오함 덕에 딥러닝은 이미지, 음성, 텍스트 등 다양한 유형의 데이터에서 높은 성능을 보이고 있다.



딥러닝의 예시 중 하나는 이미지 분류다. 이미지 분류는 입력 데이터(이미지)를 분류하는 문제로, 이를 위해 딥러닝 모델을 학습한다. 모델은 레이블이 미리 지정된(각 이미지가 어떤 카테고리에 속하는지를 지정된) 이미지 데이터를 학습하고, 이를 이용해 새로운 이미지 데이터를 분류하는 작업을 한다. 이때, 딥러닝 모델은 다양한 층으로 이루어진 인공 신경망을 이용해 이미지 데이터를 처리하게 된다.

인공지능 연산은 숫자놀음이며, 딥러닝에서 핵심은 기계가 그 수치를 보고 스스로 (인간은 알아보기 힘든) 논리와 절차를 만드는 것이다. 그런 측면에서 이미지 파일도 RGB라는 형태로 숫자로 표현할 수 있고, 대부분 국가들의 언어도 유니코드라는 숫자로 표현할 수 있다.

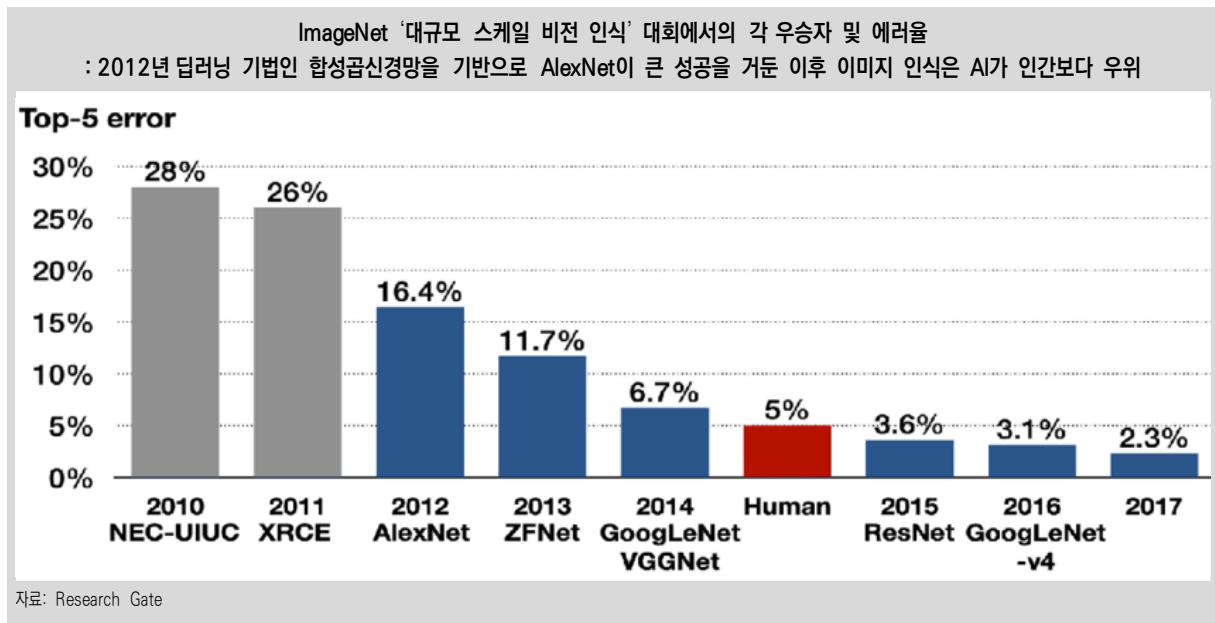
특히 최근 가장 핫한 분야는 바로 자연어 처리를 위한 언어모델이다. 언어모델은 언어처리 프로그램이라고 부를 수 있다. 요새 나오는 언어모델의 경우에는 문법 수정, 기계번역, 문장생성 등에 있어 “빨간펜 선생님” 정도의 수준까지 올라왔다. 이제는 언어모델의 신뢰성을 따지는 사람보다, 어떻게 하면 더 좋은 대답이 나올 수 있는지를 찾는 사람들이 많아졌다.

**어떻게 하면 더 좋은 대답이 나올 수 있는지는 프롬프트를 어떻게 쓰는지에 달려있음. 이 부분은 글 뒷 부분에 “프롬프트 엔지니어”라는 것으로서 설명을 함.*

딥러닝 기술은 대중화의 포문을 이제 막 열어젖혔다.

3) 딥러닝이 갖는 엄청난 진보의 속도

머신러닝과 딥러닝의 개념에 대해서는 알고 있어도, 가지각색의 기술이 너무 많아서 혼동을 느끼는 사람이 많다. 이것은 전공자나 종사자가 아니라면 인공지능의 기술 트렌드를 쫓아가는 것만으로도 너무 버겁기 때문이다. 컴퓨터 사이언스 특성상 그 발전 속도가 상상할 수 없을 정도로 빠르다. 하루가 지나면 논문이 나오고, 또 하루 지나면 그것을 넘어서는 다른 알고리즘 모델이 발표되는 수준이다. 당장 기억나는 것만 해도 AlexNet(2012년), Word2Vec(2013년), 트랜스포머(2017년), BERT, GPT-1(2018년), StyleGAN, GPT-2(2019년), GPT-3(2020년), 하이퍼클로바(2021년), ChatGPT(2022년), Bard, Ernie, GPT-4(2023년) 등이 있다. 너무 많아서 스트레스가 생길 정도다.



2012년에 처음으로 딥러닝이 개와 사람을 잘 구분한다는 사실이 알려졌고, 2년 뒤에 또 다른 혁신인 “생성 모델”이 등장했다. 그리고 10년만인 2022년 8월에는 화제의 인공지능 서비스인 Stable Diffusion이 나왔다. Stable Diffusion도 텍스트를 넣었더니 그림이 나오는 생성모델이다.

**생성 모델은 식별 모델과 입출력이 반대. 식별 모델은 이미지를 보고 어떻게 분류를 할 건지 “자연어로 된 출력물”을 생성. 반면, 생성 모델은 자연어를 입력값으로 이미지를 생성. 그런데 이미지는 자연어보다 많은 숫자 데이터를 내포(임베딩 벡터). 많은 숫자에서 적은 숫자를 뽑아내는 함수보다 적은 숫자에서 많은 숫자를 뽑아내는 함수가 더 어렵기에 생성 모델의 난이도가 더 높음.*

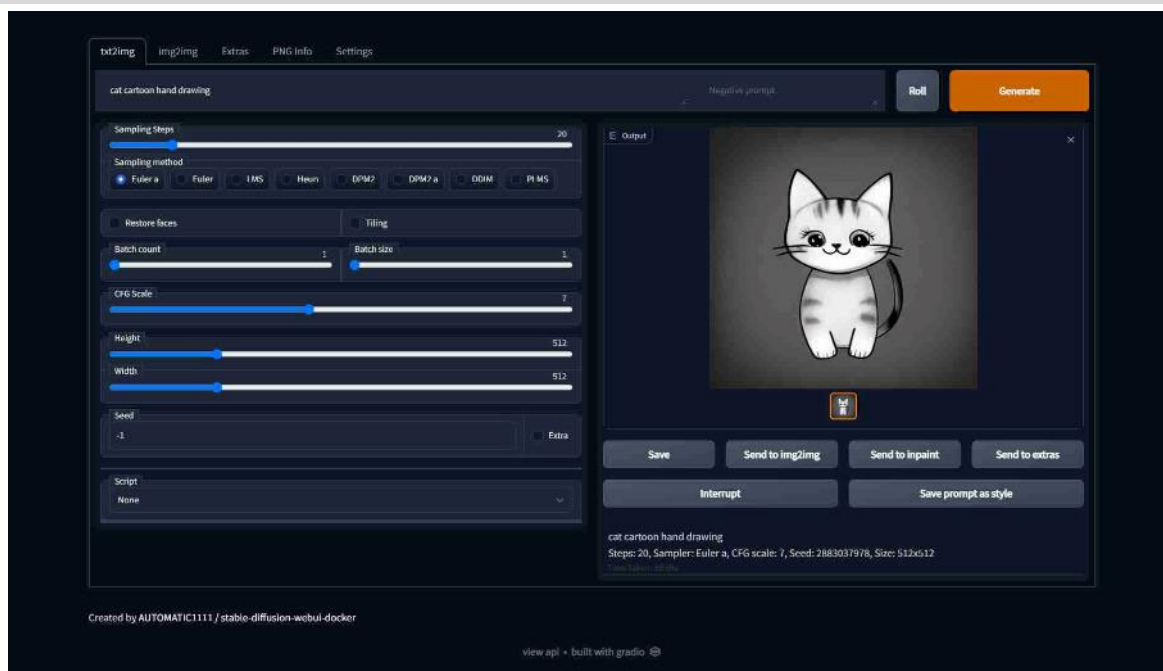
그런데 그 퀄리티가 실제사진과 같은 수준이라서 엄청난 충격을 줬다. WebUI 형태를 띠고 있어 사용하는 것이 그리 쉽지는 않지만, 그림에도 대중의 주목을 많이 받고 있다. 더군다나 스테이블 디퓨전은 오픈 소스이기 때문에 파생된 연구와 개발속도가 더욱 가속화되고 있다.

화제의 이미지 생성형 AI 모델, “Stable Diffusion”



자료: Stability AI

Stable Diffusion을 사용하기 위한 WebUI 예시 화면



자료: Github “AUTOMATIC1111”

ChatGPT와 Stable Diffusion 등 인공지능 모델과 어플리케이션, 그리고 서비스는 우리에게 빠른 속도로 다가오고 있다. 그 속도와 위용에 기죽지 않으려면 어떻게 해야 할까? 그들을 이해하고, 또 투자관점으로 이용하기 위해서 딥러닝 기술에 대해서 전반적인 개념을 숙지하고 있어야 한다.

ChatGPT의 대항마가 나온다고 했을 때, 그들은 어떤 모델 알고리즘을 썼는지, 어떤 강점이 있고 약점이 있는지, 그들의 모델 크기는 어떻게 되는지 등을 판단할 수 있어야 한다. 우리는 신차가 나오고, 반도체 신제품이 나왔을 때 스펙 사항 등에는 친숙한 편이다. 딥러닝 영역에서도 앞으로 그렇게 할 수 있다면 투자에 도움이 될 것이다.

다만 너무 지엽적인 개념에 집중하기보다는, 어떤 모델 알고리즘이 헤게모니를 쥐고 있는지 판단해 중심점을 잡고 있어야 한다. 그 과정을 위해서 기초부터 하나하나 개념을 살펴보자.

AI 학습의 네 종류(지도/비지도/자기지도/강화)

1) 지도학습

머신러닝과 딥러닝에 있어서 지도학습과 비지도학습은 가장 기반이 되는 개념이다. 지도학습은 인공지능에 관심 있는 사람이라면 익히 들어 이해하기 수월한 개념이다. 쉽게 말해, 지도학습(Supervised learning)은 ‘답안지를 보고 공부’하는 방식이다. 정답이 매칭되어 있는 데이터를 모델에 집어넣고 인풋과 아웃풋 사이의 패턴을 익히도록 학습한다. 이렇게 학습한 것을 바탕으로, 공부하지 않은 새로운 문제(답안지가 없는)가 주어졌을 때 정답을 추론하여 도출한다. 그래서 이 방식은 이미지 식별 모델에 어울린다. 어떤 사진을 보고 이게 강아지인지 고양이인지 분류하는 식이다. 강아지 사진을 수천만 개 주고 강아지임을 알려주면 기계가 강아지의 특징(feature)을 찾아내고, 새로운 사진을 주었을 때 강아지인지 여부를 구분할 수 있게 되는 방식이다.

이렇게 주어진 데이터를 분류하는 비교적 간단한 지도학습도 있지만, 데이터를 통해 스스로 패턴을 파악해야 하는 경우도 있다. 예를 들어, 어떤 운전자가 교통사고가 날 확률을 예상한다고 할 때, 교통사고를 일으키는 수많은 요인(교통정보, 운전자의 나이, 수면 시간 등, 전방 주시 시간, 운전 시간 등) 중에 각 요인들의 중요도를 파악해 내는 것이다. 여기서 각 요인들이 서로 영향을 미치는 중요도는 일종의 ‘가중치’ 역할을 하게 된다. 따라서, 가중치를 잘 파악한 모델의 경우 각 요인들에 다른 운전자의 수치를 대입해보면 교통사고가 날 확률을 계산할 수 있다. 이런 과정을 ‘추론(inference, 딥러닝에서는 수행이라고도 한다)’이라고 한다.

그리고 그 중요도를 1차적으로 파악한 뒤, 학습 데이터의 실제 결과값과 예측 출력값을 서로 비교하고, 그 차이를 줄이기 위해 매개변수를 조정한다. 그 차이를 ‘오류 함수’로 나타내고, 이 함수의 값을 최소화하는 과정을 반복한다. 이 과정이 ‘학습(training)’이 된다.

다만 이 경우, 인간이 설정한 ‘답안지’의 질에 따라 결과의 정확도에 차이가 발생한다. 즉 사람이 답안지를 일일이 만들어야 하기 때문에 시간과 비용이 크게 수반되는 것이 큰 단점이다.

2) 비지도학습

비지도학습(Unsupervised learning)은 지도감독을 하지 않는다는 단어에서 유추할 수 있듯이 ‘답안지 없이 문제만 주어지는 경우’를 의미한다. 결과값(label, 레이블, Y값)을 아직 모르는 문제에 대한 답을 원하는 경우 사용한다. 알고리즘이 특성(feature, 피쳐, X값)을 찾아서 묶도록 하는 것이다. 따라서 비지도학습은 지도학습과 달리 클러스터링, 차원감소와 같은 분류 작업에 특화되어 있다. 예를 들어, 기업이 고객을 분류하는 적절한 기준이 무엇인지 모를 때 비지도학습을 통해 고객 간 공통점과 차이점을 발견하고 분류할 수 있다.

비지도학습은 모델이 안내나 피드백 없이 데이터에서 의미 있는 패턴을 찾아내야 하므로 지도 학습보다 시간과 계산 비용이 더 많이 소요된다. 하지만 딥러닝에서는 비지도학습을 사용하는 것이 대세가 되어가고 있다. 특히 비지도 사전학습이라는 방식을 통해 딥러닝에서 활용되고 있다.

비지도 사전학습은 언어의 일반적인 특징과 표현을 학습하기 위해 레이블이 지정되지 않은 대규모 텍스트 모음에서 모델을 교육함으로써, 감정분석을 위한 특정 언어모델에 사용된다. 이 프로세스는 종종 "언어모델에 대한 사전 학습"이라고 부른다.

이때 사전학습을 위해 일반적으로 트랜스포머와 같은 아키텍처를 활용한다. 이전 단어로부터 주어진 시퀀스에서 다음 단어를 예측하도록 모델을 학습시키는 과정의 일환이다. 예를 들어, 웹 크롤링(인터넷에서 데이터를 검색해 필요한 정보를 수집하는 것) 방법으로 대규모 텍스트 말뭉치를 사용해서, 이전 단어가 주어지면 시퀀스의 다음 단어를 예측하도록 훈련시킨다. 사전학습 단계에서 언어모델은 텍스트에서 문맥 정보를 추출하고, 단어들의 분산된 표현(distributed representation, 인간은 알아보기 힘들고 컴퓨터가 알아볼 수 있는 형태의 각 단어가 가지고 있는 숨겨진 수치적 표현)과 단어들 간 관계성을 학습한다.

언어모델이 사전학습을 마치면, 감정 분석과 같은 특정 작업에 대해 파인튜닝(미세조정)을 할 수 있게 된다. 미세조정은 텍스트의 감정을 분류하기 위해, 감정분석 작업을 위한 데이터 세트(사전 학습 모델보다 더 적은 숫자의 데이터)에 대해 모델을 교육하는 것을 말한다. 사전에 학습된 언어모델의 가중치는 감정 분석 작업에 최적화되도록 함께 조정된다.

게다가 감정분석 작업에 대해서 레이블이 지정된 데이터 숫자가 제한되어 있다고 한다면, 전이학습(Transfer Learning)을 사용하면 된다. 언어모델 같은 기반이 되는 큰 모델을 사전훈련함으로써 감정분석 작업으로 Transfer 할 수 있는 유용한 기능을 미리 학습할 수 있는 장점이 있다. 이는 이미지 모델에서도 마찬가지다. 아래의 예시들은 비지도학습이 딥러닝에서 어떻게 사용되는지 보여주는 몇 가지 예다.

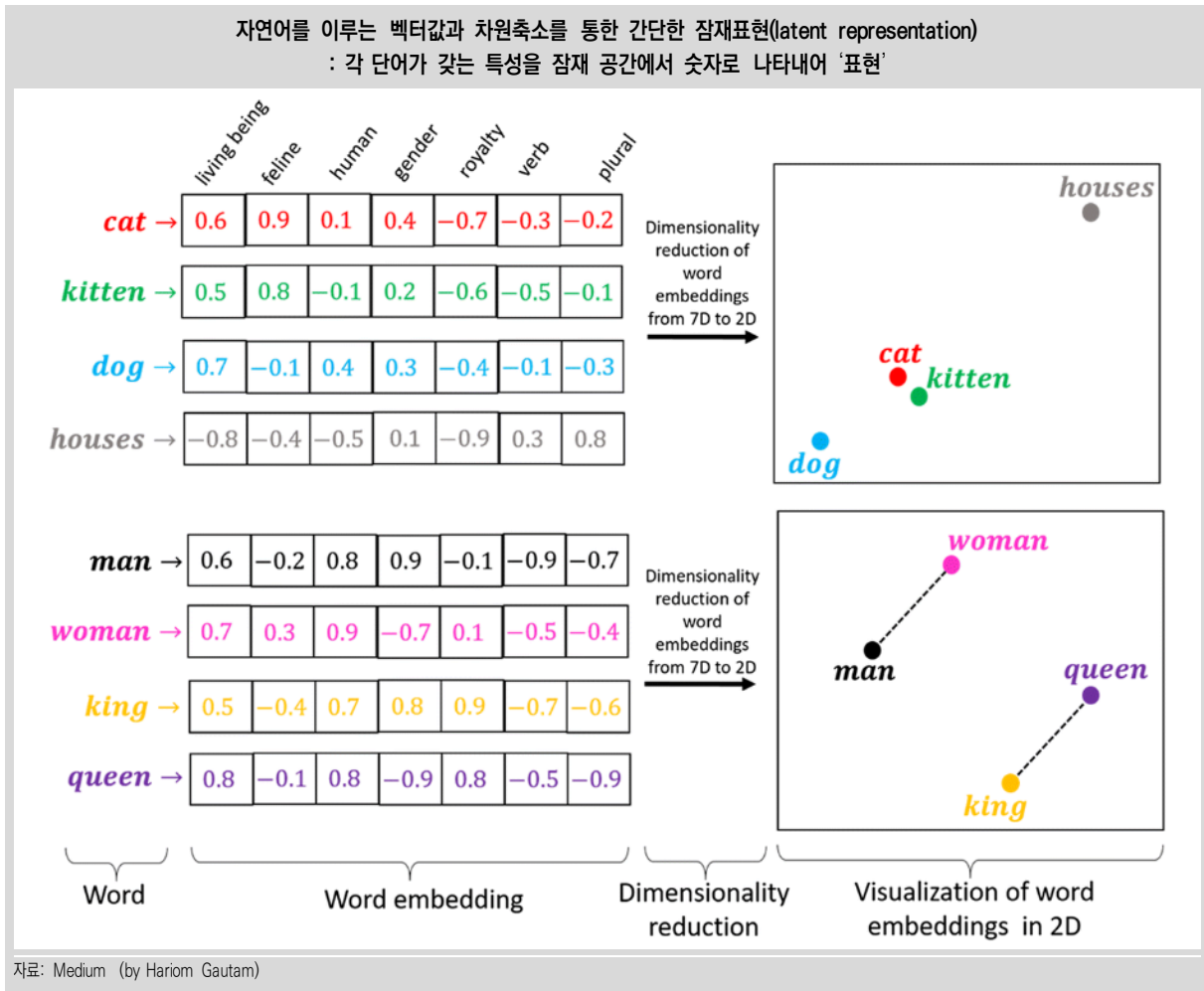
a. 오토인코더(autoencoder)

오토인코더는 입력 데이터를 출력으로 복원하는 방법으로 데이터의 특징을 추출한다. 오토인코더는 “입력을 재구성”하는 것을 목표로 하는 신경망이다. 일반적으로 인코딩이라고 하는 것은 무언가를 압축한다고 이해하면 편하다. 딥러닝에서 인코더는 “차원을 줄이기 위해, 컴퓨터가 이해할 수 있도록” 하는 표현들로 변환하는 작업을 맡는다. 즉, 인코딩을 통해 원래 입력 데이터보다 크기가 훨씬 작아지므로, “축소 표현(reduced representation)”이라는 용어가 사용되기도 한다. 이런 특징 때문에 차원감소에 강점을 갖는다.

***인코더:** 입력 데이터를 잠재공간(latent space)에서 잠재표현(latent representation)으로 나타낸 다음, 평균과 분산을 추출하는 매핑을 수행.

***디코더:** 잠재공간에서 샘플링한 벡터를 다시 원래 데이터로 복원.

오토인코더의 인코딩 또한 네트워크에서 생성되는 입력 데이터의 “압축된 저차원 표현”을 나타낸다. 그런 다음 인코딩을 거친 출력값은 네트워크의 디코더 부분에 대한 입력으로 사용된다. 네트워크 디코딩에서는 인코딩으로 받은 데이터를 오리지널 데이터처럼 재구성하는 것을 목표로 한다. 인코더를 거쳐 압축했던 입력 데이터가 디코더를 한 번 더 거치니 다시 원래 크기로 돌아오는 것이다. 이런 일을 왜 하나 싶지만, 효과적으로 압축하고 풀었다는 것은 그 데이터를 이루는 여러 차원들을 “이해”했다는 뜻이다. 인간의 능력으로는 간파하기 힘든 차원의 영역까지 기계가 파고들어 벡터값을 산출했고, 그 숫자들을 이리저리 만져가면서 각 데이터의 본질과 패턴을 파악했기에 가능한 일이다.



***임베딩 벡터(embedding vector)**는 자연어처리를 위해 사람이 쓰는 말을 기계가 이해할 수 있도록 숫자 형태인 벡터로 바꾸는 과정이며, 우리에게 보이지 않지만 기계는 이해할 수 있는 “정보를 내포” 하고 있다는 것에서 “임베딩”이라고 표현. 임베딩 벡터는 “분산된 표현방식(distributed representation)” 이라고 불리기도 함. 임베딩 벡터는 입력 데이터의 압축 데이터이자 입력 데이터를 잘 표현하기 위한 새로운 공간을 형성하기 때문에 입력값의 잠재표현으로 활용될 수 있음.

대량의 텍스트 데이터로부터 단어의 의미적인 관계를 학습하여 임베딩 벡터를 생성. 임베딩 벡터는 자연어 처리 모델에서 입력 데이터로 사용. 예를 들어 문장 분류나 감정 분석과 같은 작업에서, 문장 내 각 단어들을 임베딩 벡터로 변환하여 모델에 입력.

임베딩 벡터는 단어의 의미적인 관계를 반영하여 생성되며, 비슷한 의미의 단어들은 벡터공간에서 가까운 위치에 배치. 임베딩 벡터는 변수의 서로 다른 값 사이의 의미론적 관계를 파악할 수 있게 해서, 일반적으로 딥러닝의 훈련 프로세스의 일부로 학습.

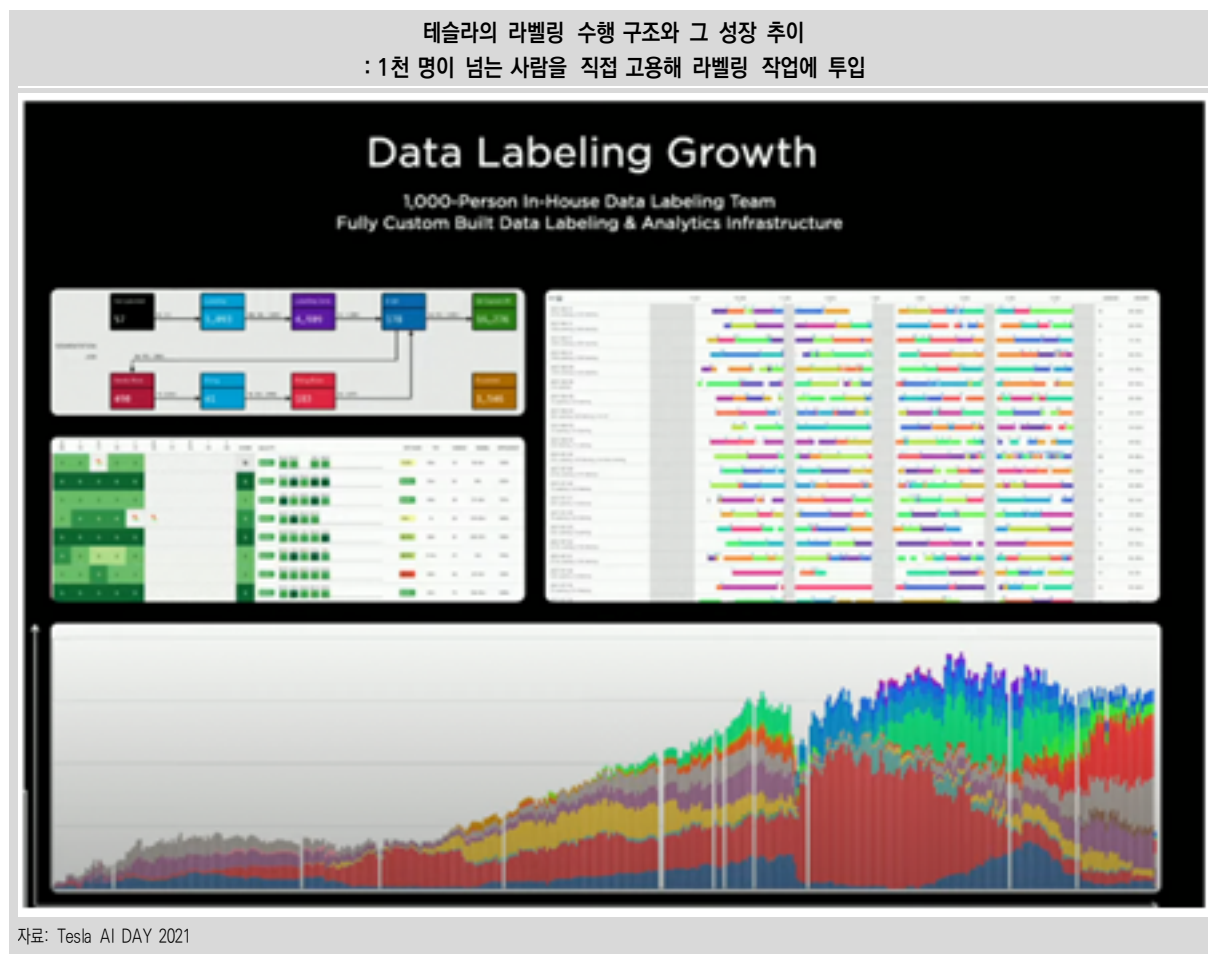
예를 들어, 사람을 N차원의 벡터값으로 구분한다고 했을 때, Man과 woman은 단어상으로 상당히 비슷. 언어모델인 기계가 인간처럼 의미까지는 모르겠지만 man과 woman의 임베딩 벡터가 비슷하다는 것이 중요. 더 나아가서 king과 queen의 임베딩 벡터를 각각 비교해보는다고 가정하면, 이 둘 역시 임베딩 벡터가 비슷. 그런데 king과 queen의 임베딩 벡터값 차이가 man과 woman의 벡터값 차이와 실제로 같다는 점이 딥러닝의 괄목할만한 점. 즉, king과 queen, man과 woman의 임베딩 벡터 차이값은 ‘성별 벡터값’이라고 정의할 수 있음. 비슷한 사람끼리는 서로의 벡터값 차이가 적다는 점으로 기억해두면 좋음.

오토인코더는 특성 학습과 표현을 위해서 컴퓨터 비전에서 널리 사용되는 기술이기도 하다. 즉, 이미지 재구성과 이미지 분류 같은 작업에 잘 사용된다. 예를 들어, 이미지 데이터를 재구성하는 모델을 학습시키면, 모델은 이미지 데이터의 ‘특성’을 추출하고 이를 이용해 이미지를 재구성한다.

노이즈가 있는 이미지를 입력한다면, 해당 모델은 이미지의 특징을 추출해 노이즈만 제거하고 깨끗한 이미지를 생성할 수 있게 된다. 더 자세히 말하면 노이즈가 있는 이미지를 입력 데이터로 사용해 인코딩해서 잠재 표현으로 만들어낸다. 이로써 특징을 추출해서 노이즈 부분의 수치를 도려낸 뒤 디코딩해서 깨끗한 이미지로 다시 재구성해낸다.

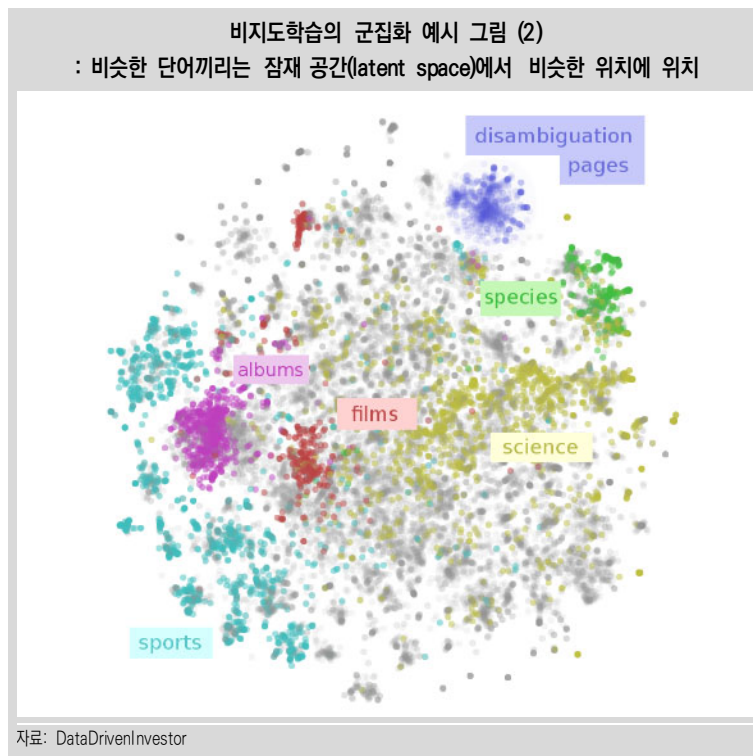
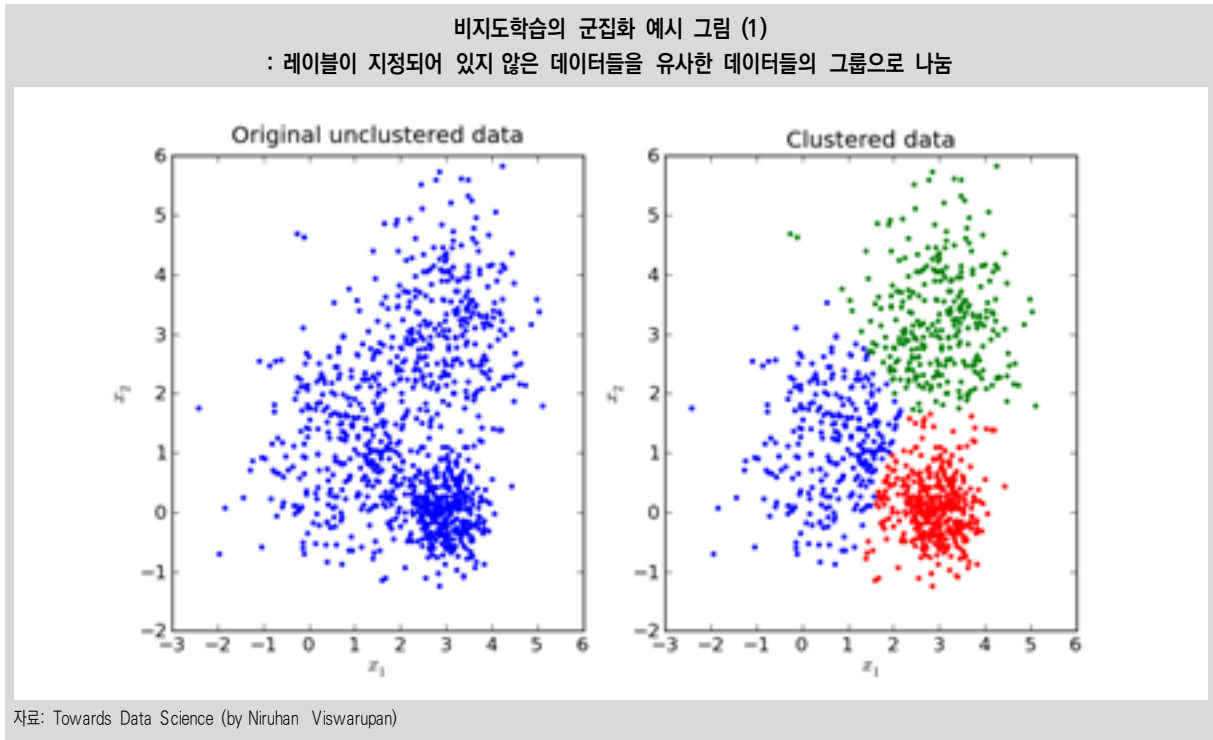
이러한 오토인코더가 실제로 쓰이는 사례에는 테슬라의 오토라벨링을 꼽을 수 있다. 오토라벨링은 각 데이터에 자동으로 주석을 다는 것을 의미한다. 그러니까 자동차가 지나가면서 카메라 센서로부터 얻은 영상 클립에서 각종 도로정보와 사물정보를 보고 ‘이것은 사람이다, 강아지다, 도로표지판이다’라는 주석을 자동으로 붙이는 것이다.

라벨링을 위해서 테슬라 자체적으로 1천명에 해당하는 사람을 고용했다고 했을 정도니, 라벨링에는 노력과 자금이 많이 수반됨을 알 수 있다. 자율주행을 위한 대규모 주행영상 데이터를 라벨링하는 작업은 더욱 쉽지 않다. 그런데 테슬라는 이를 오토라벨링을 통해 보완하고 있다.



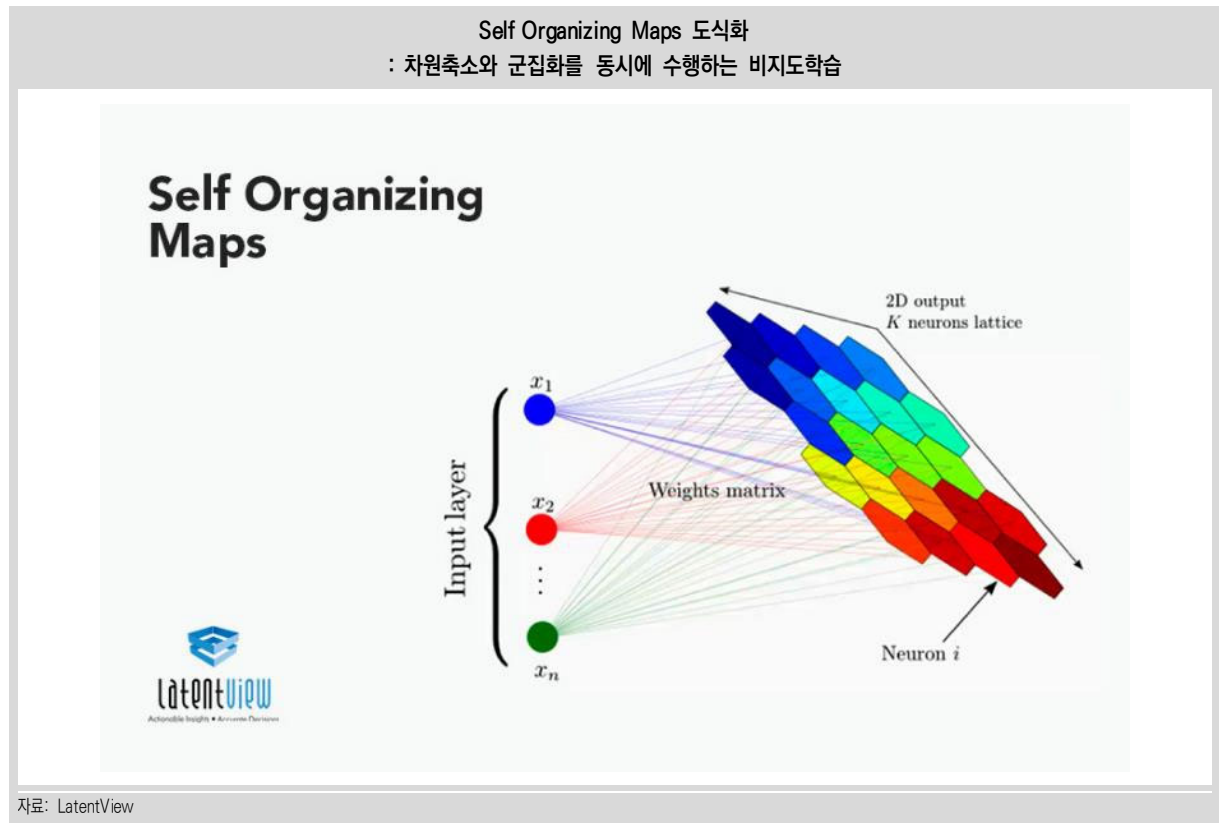
b. 군집화(clustering)

군집화는 데이터를 보고 알고리즘 모델이 알아서 비슷한 특징을 가진 데이터 포인트를 grouping하는 기술이다. 이상 감지, 고객 세분화, 추천 시스템과 같은 작업에 사용된다. 딥러닝에서 군집화 알고리즘은 이미지나 텍스트와 같은 대규모 데이터 세트에서 숨겨진 패턴을 발견하는 데 사용된다.



c. 자기조직화지도(Self-Organizing Maps)

SOM은 저차원 공간에서 고차원 데이터를 표현하는 방법이다. 데이터 시각화, 클러스터링, 이상 감지와 같은 작업에 사용한다. 데이터의 기본 구조를 발견하고 서로 다른 데이터 포인트 간의 관계를 시각화 할 수 있다.



3) 자기지도학습

비지도학습과 지도학습의 특징을 조금씩 가지고 있는 방식이 자기지도학습(Self-supervised learning)이다. ‘답안지’가 없지만, 대상을 식별하는 것을 목표로 하기 때문이다. 자기지도학습은 데이터의 일부를 보고 나머지를 예측하는 self-prediction이나 데이터 사이의 관계를 예측하는 contrastive learning에 쓰인다. 스스로 문제를 내고 학습하는 방식이라고 볼 수 있겠다. 이러한 방법은 사람의 수작업 라벨링이 어려운 대규모 데이터 세트에서 사용된다. 현재 많은 대규모 모델이 자기지도학습을 사용하고 있으니 보다 자세히 알아보자.

Self-prediction에는 여러 종류가 있는데, 최근 떠오르고 있는 대부분의 대규모 모델이 이 안에 속해 있다. 먼저 GPT를 비롯한 웨이브넷, XLNet, PixelCNN 등의 모델들은 자동회귀(autoregressive) 생성 모델들이다. 이전까지의 결과값을 기반으로 다음 값을 예측하는 방식을 사용한다. GPT와 양대산맥을 이루고 있는 BERT 및 디노이징 오토인코더, 마스크드 오토인코더, simMIM 모델들은 마스크드 생성 모델들이다. 전체 데이터의 일부를 가리고, 나머지를 기반으로 가려진 영역을 유추하도록 학습된다.

이 외에도 샘플에 변형(회전, 순서 섞기)을 주고 어떤 변형이 가해졌는지 원본을 찾게 하는 내재 관계 예측(innate relationship prediction) 모델이나, 이전의 여러 방식을 결합해 사용하는 하이브리드 모델도 있다. GAN이나 DALL-E, Jukebox 등이 대표적인 하이브리드 모델로, 방대한 데이터로 인해 개인 연구에 하드웨어의 한계가 있다는 특징이 있다.

Self-prediction과 달리, contrastive learning은 데이터의 관계에 집중한 학습이다. 먼저, 데이터 하나를 기준으로 해당 데이터를 변형한 것은 positive pair, 그 외의 데이터는 negative pair로 둔다. 그 뒤 기준이 되는 데이터와 positive pair는 가까워지고, negative pair와는 멀어지도록 학습시킨다. 딥러닝에서 모든 데이터는 벡터로 표시되기 때문에, 각 벡터가 유사할수록 유사한 데이터라는 아이디어에서 출발한 방식이다. 잠재공간에서 코사인 유사도를 사용하면, 서로 다른 데이터 포인트(이미지 또는 텍스트 데이터 등)가 얼마나 유사한지를 비교할 수 있다. 코사인 유사도는 두 벡터 사이의 각도를 기준으로 유사성을 평가한다.

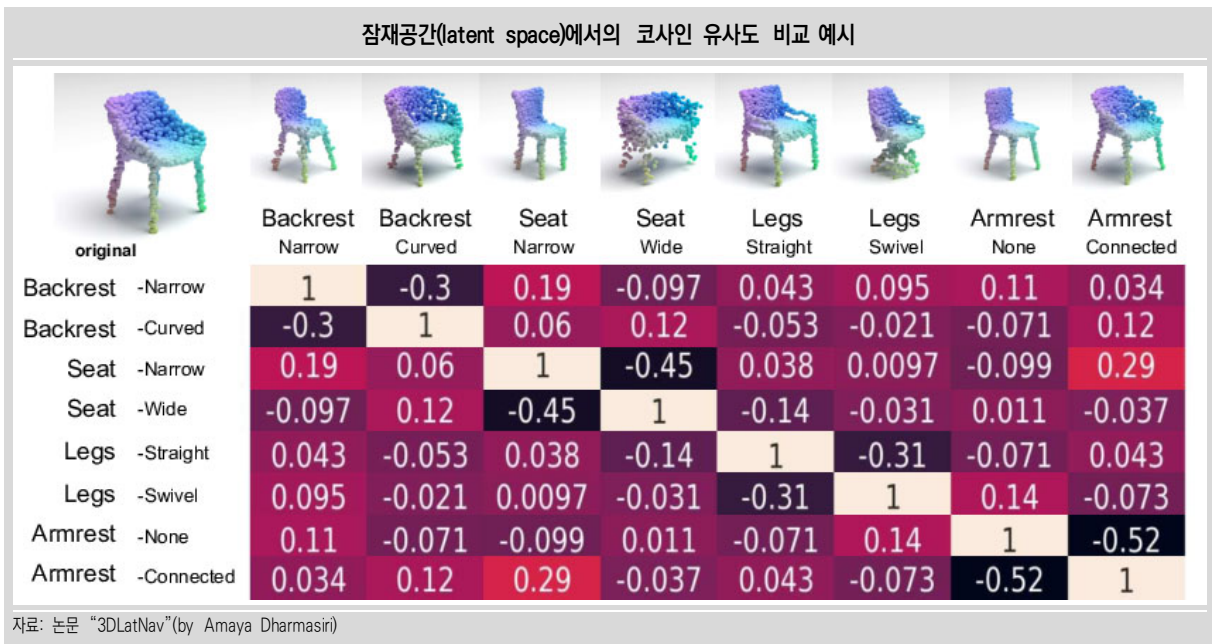
$$* \text{cosine_similarity}(A, B) = (A \cdot B) / (\|A\| \|B\|)$$

여기서 A와 B는 비교할 두 벡터이고, '·'는 벡터 간의 내적을 의미하며, \|A\|와 \|B\|는 각각 A와 B 벡터의 크기. 코사인 유사도는 정규화된 벡터값으로 계산하여, 계산이 비교적 간단하고 효율적. 코사인 유사도 값은 -1에서 1 사이의 값을 가지며, 1에 가까울수록 두 벡터가 유사하고, -1에 가까울수록 서로 다르다고 판단할 수 있음. 사인이나 탄젠트를 사용하지 않는 이유는 아래와 같음.

*코사인 0도=1 / 코사인 90도=0 / 코사인180도=-1

*사인 0도=0 / 사인 90도=1 / 사인 180도=0 (0도와 180도 둘 다 0이므로 해석이 어려움)

*탄젠트는 90도와 270도에서 정의할 수가 없는 존재하지 않는 수라서 적용이 어려움.

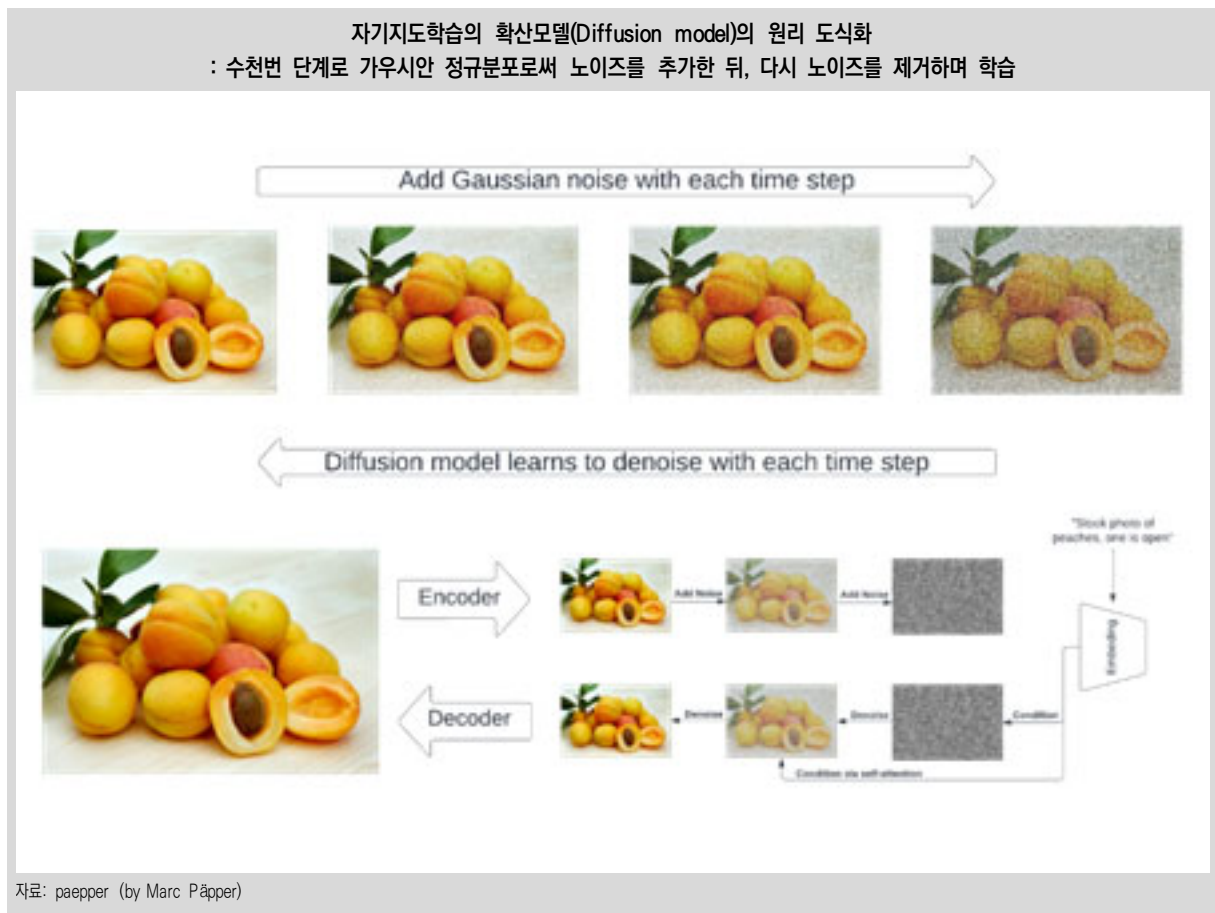


a. 확산모델(Diffusion Model)

자기도학습 모델은 최근 빠른 속도로 진화하며 도입되고 있다. 특히 이미지 데이터에서의 Generative 한 측면에서 그렇다. 대표적인 이미지 생성 모델로, 확산모델(DM: Diffusion Model)이 바로 그 주인공이다. DM은 사실적인 이미지를 생성하기 위해 딥러닝에서 사용되는 강력한 도구로 꼽힌다. 특히 GAN과 같은 다른 생성형 인공지능과 비교했을 때도 더 높은 품질의 이미지를 생성한다는 평가를 받으며 인기를 끌고 있다. 물론 이것은 충분히 큰 데이터 세트가 주어져야 가능한 아주 복잡한 작업이다.

DM을 학습하기 위해, 먼저 기준점이 되는 시작 이미지에 점진적으로 노이즈를 추가한다. 노이즈가 시간이 지남에 따라 이미지를 통해 ‘확산’되기 때문에 이 프로세스를 ‘diffusion’이라고 부른다. 현재 상태의 이미지에 약간의 노이즈를 입력으로 받아 다음 단계의 이미지를 생성한다. 노이즈를 추가하는 과정은 1천 번 이상으로 세세하게 단계를 나눠 추가한다. 그런 다음 모델은 거꾸로 잡음이 있는 버전으로부터 원본 이미지를 재구성하려고 시도하면서 학습한다. 다시 말해, 해당 노이즈를 제거해 조금씩 복원하는 방식을 통해 학습한다는 말이다. 그리고 이 과정을 수천번 반복한다.

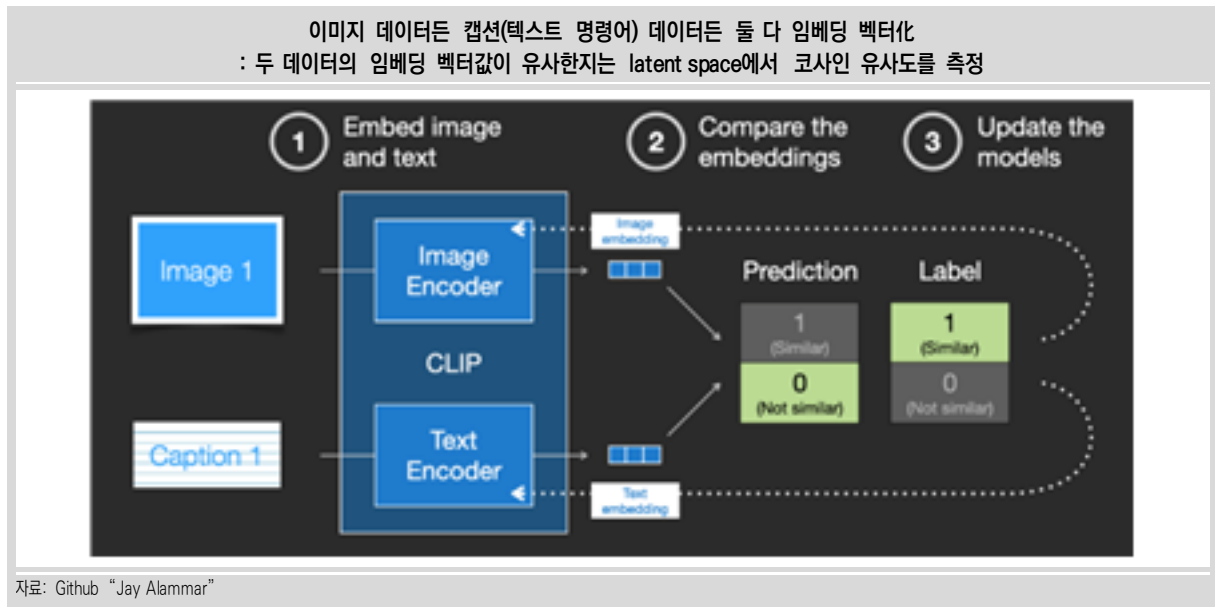
**무작위 노이즈를 생성하는 것은 상대적으로 간단함. 일반적으로, 무작위 노이즈 벡터는 컴퓨터에서 제공하는 난수 생성기를 사용하여 구할 수 있음. TensorFlow와 PyTorch와 같은 딥러닝 프레임워크도 난수 생성에 사용할 수 있는 기능을 제공함.*



학습이 잘 된 경우에는, 완전한 무작위 노이즈 상태에서도 훈련된 이미지 파일과 똑같도록 새로운 이미지를 생성할 수 있게 된다. 이 프로세스는 이미지에서 누락되거나 노이즈가 많은 부분을 채워야 하는 노이즈 제거 또는 인페인팅(재채색)과 같은 작업에 사용할 수 있다. 또한, 최근 기술의 결과 이미지가 매우 사실적인 수준까지 올라왔기 때문에, 이미지 생성, 심지어는 이미지 조작과 같은 사태가 발생하기도 한다.

유명한 확산모델에는 레이턴트디퓨전 모델(LDM: latent diffusion model)과 스테이블 디퓨전 모델(SDM: stable diffusion model)이 있다. 2022년 6월에 출시된 스테이블 디퓨전 이미지 생성모델은 미드저니나 DALL-E와 버금갈 정도의 사실적인 그림 품질을 제공함에도 오픈소스라는 점에서 파란을 일으키고 있는 모델이다. 또한 스테이블 디퓨전 모델도 텍스트로 명령을 받아 그림을 생성하는 방식으로, 자연어 인터페이스를 갖고 있는 것이 특징이다.

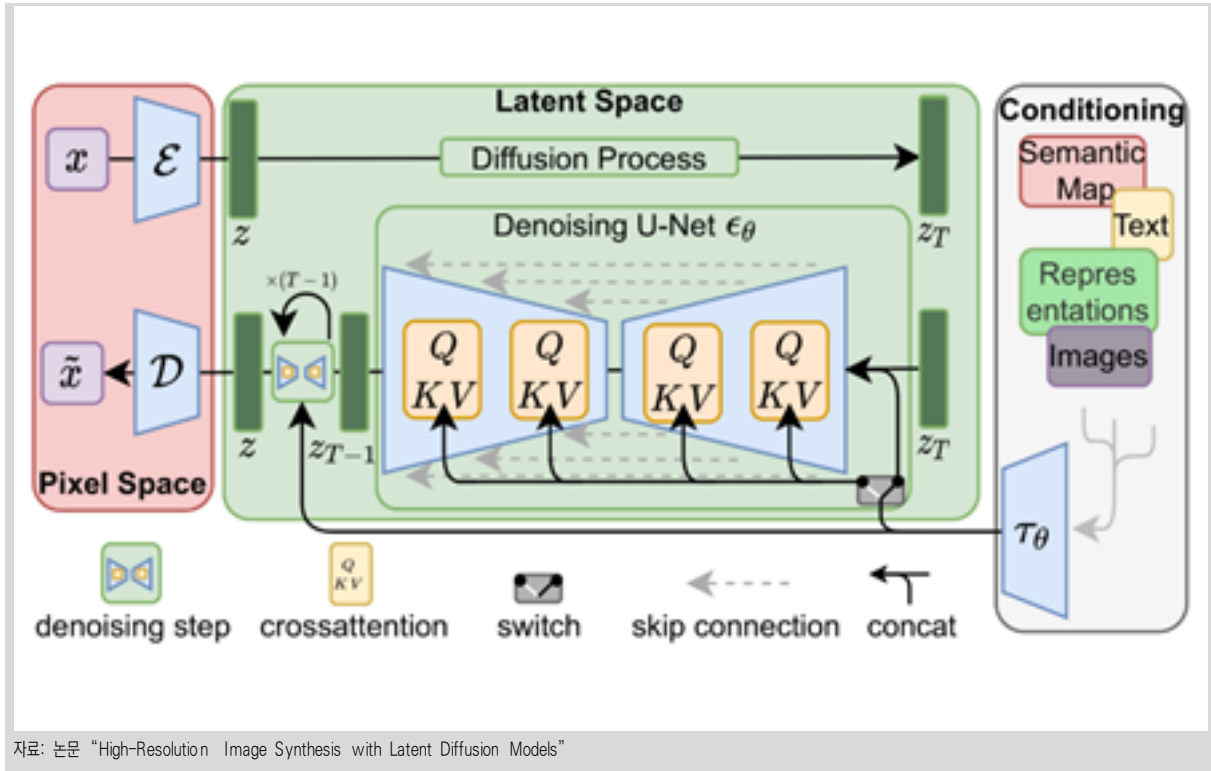
이는 스테이블 디퓨전이 CLIP(Contrastive Language-Image Pre-training)이라는 OpenAI에서 개발한 임베딩 모델을 갖고 있기 때문이다. CLIP은 텍스트 프롬프트를 수치적 표현으로 변환하는 신경망이다. CLIP은 4억 개가 넘는 텍스트 및 이미지를 연결하는 방식으로 학습됐고, 텍스트 설명을 기반으로 이미지를 생성하거나 분류하는데 사용된다. 인공지능 연산이 텍스트와 이미지 데이터 영역을 분간하지 않고 넘나들 수 있는 이유는, 두 데이터 모두 다 숫자로 일반화해서 표현할 수 있기 때문이다. 즉, 임베딩 벡터로 만드는 것이 핵심이다.



LDM과 SDM 모델은 약간의 구현 및 학습 방식의 차이일 뿐, 전반적인 개념은 매우 유사하다. 둘 다 '무작위 노이즈 이미지로부터 시작'하여 점진적으로 사실적인 최종 이미지로 변환하는 프로세스를 거친다. LDM과 SDM은 목적과 사용 사례에 따라 각기 장점이 있는 생성 모델이다. 따라서 LDM과 SDM 중 어느 것이 더 우수하다고 평가할 수는 없고, 사용 사례와 기술적 요구 사항에 따라 달라진다.

LDM은 압축된 상태인 잠재공간 내에서 작동하므로 연산 요구량이 적다는 장점이 있다. 이미지 생성, 인페인팅, 초해상도와 같은 다양한 작업에서 높은 성능을 발휘한다. 반면 SDM은 안정적인 확산 과정을 사용해 이미지 생성의 일관성과 품질을 높인다. 또한 텍스트 프롬프트를 사용하여 이미지를 생성한다. 일부 작업에서는 SDM이 더 간단하고, 접근 방식이 직관적이어서 사용하기 쉬울 수 있다.





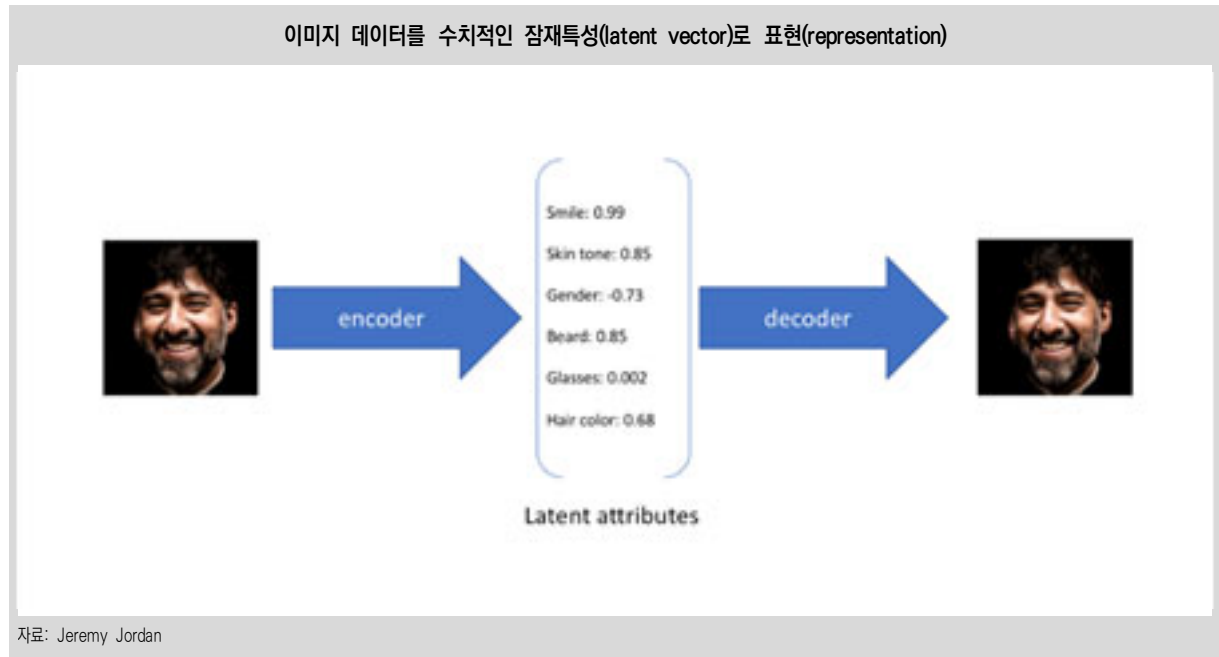
b. 생성적대신경망(GAN)

확산모델의 성능에 밀려 최근에는 다소 덜 언급되고 있으나 GAN이라고 불리는 자기지도학습 모델 또한 잊어서는 안되는 중요 개념이다. 페이스북 인공지능 연구 담당자 얀 르쿤(Yann LeCun)은 GAN을 "2020년대에서 가장 흥미로운 아이디어"라고 평가했을 정도니 말이다.

GAN은 자기지도학습 모델이기 때문에 레이블 된 정보는 하나도 없다. 또한 생성 모델이기 때문에, GAN은 이미지 생성과 합성 처리에 좋은 성과를 보인다. 즉, 실제 데이터로 이루어진 훈련 세트와 유사한, 새로운 데이터를 생성할 수 있는 딥러닝 알고리즘의 일종이다.

GAN은 Generative Adversarial Networks의 줄임말로, 생성적대신경망으로 번역할 수 있다. 개념이 조금 어려워 보이지만, 기본 메커니즘은 간단하다. GAN은 두 개의 하위 네트워크인 생성기와 판별기로 구성된다. 생성자(Generator)는 실제 데이터와 구별할 수 없는 샘플을 생성하는 방법을 학습하고, 판별자(Discriminator)는 실제 샘플과 가짜 샘플을 구별하는 방법을 학습한다. 두 알고리즘을 경쟁시키는 개념은 컴퓨터 과학 분야의 유명한 연구자인 아서 사무엘(Arthur Samuel)이 처음 제안했다. 참고로 아서 사무엘은 "머신러닝"이라는 용어를 대중화시킨 논문으로 유명하다.

먼저, 생성자 네트워크는 낮은 차원의 노이즈 벡터를 고차원의 데이터 공간으로 매핑하여 새로운 데이터를 생성하는 방법을 학습한다. 이 말을 이해하기 위해서는 잠재공간이라는 개념을 알아야 한다. 잠재공간은 기계가 이해할 수 있도록 데이터를 '저차원의 벡터 공간'에 표현할 수 있는 공간을 뜻한다. 이 공간에서 각각의 차원은 데이터의 특정 '속성'이나 '변화'를 나타낼 수 있게 된다.



이 공간에서는 인간은 이해할 수 없지만 기계가 어떤 데이터 파일을 두고, 이 파일을 이해하기 위해서 숫자의 나열인 벡터값으로 표현한다고 생각할 수 있다. 딥러닝에서 가장 중요한 단어가 있다면 벡터라고 생각될 정도로, 결국 모든 데이터를 수치화해서 나열한 뒤 조작하는 것이다. 그리고 그 표현과 조작은 잠재 공간에서 이뤄진다. 이와 같은 차원 감소는 아래와 같은 몇 가지 이유로 딥러닝에서 필수적이다.

(1) 계산 효율성:

입력 데이터의 특성(feature) 수 또는 차원 수가 증가함에 따라 훈련의 계산 비용이 증가한다. 이로 인해 고차원의 대규모 데이터 세트에 대한 학습이 비실용적이거나 실행 불가능할 수 있다. 차원 감소는 이 특성의 숫자를 줄여 더 작고 관리하기 쉬운 데이터 세트에서 모델을 훈련할 수 있도록 한다.

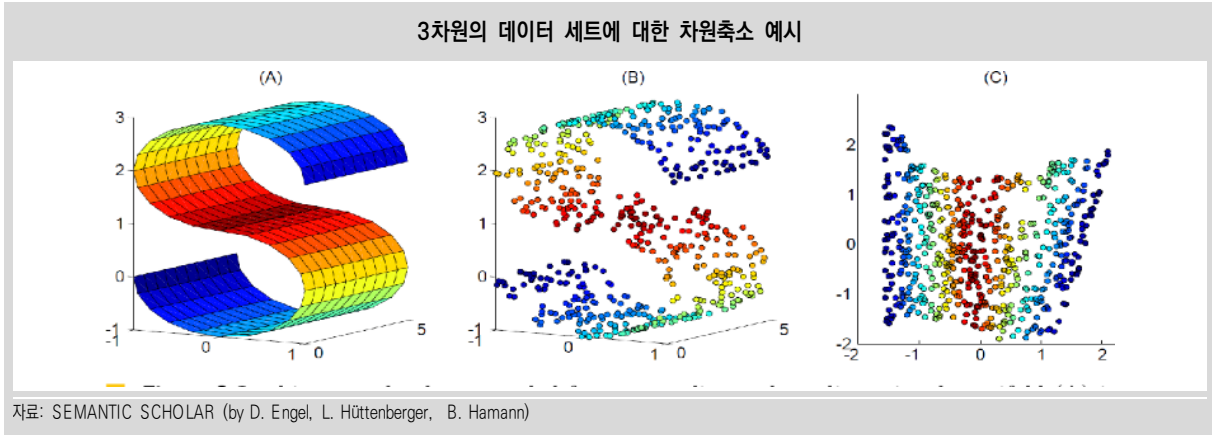
(2) 과적합(Overfitting):

모델이 샘플 데이터에 대해 너무 오래 학습하거나, 모델이 지나치게 복잡하면 데이터 세트 내에서 "노이즈" 또는 관련 없는 정보까지 다 학습해버리는 부작용이 생긴다. 굳이 안 그래도 되는 데이터를 그냥 다 외워버리는 꼴이다. 이런 경우를 "과적합"이라고 부른다. 과적합에서는 모델이 새로운 데이터에 대한 일반화를 하는 게 아니라, 학습데이터를 기억하게 되는 문제가 있다. 쉽게 말해 모델이 각종 노이즈를 다 기억해서 너무 끼워맞추기식으로 모델을 구성하는 것이다. 이런 경우 새 데이터에는 써먹을 수가 없어 학습 목적이 사라지게 된다.

과적합을 방지하기 위해 훈련 데이터 세트의 일부는 과적합을 확인하기 위해 "테스트 세트"로 따로 보관하기도 한다. 보통 학습 데이터의 오류율이 낮는데, 테스트 데이터의 오류율이 높으면 과적합으로 생각하기도 한다. 차원 감소는 데이터의 노이즈와 중복성을 줄이는 데 도움이 되어 기본 패턴을 더 쉽게 배우고 새 데이터로 일반화할 수 있다.

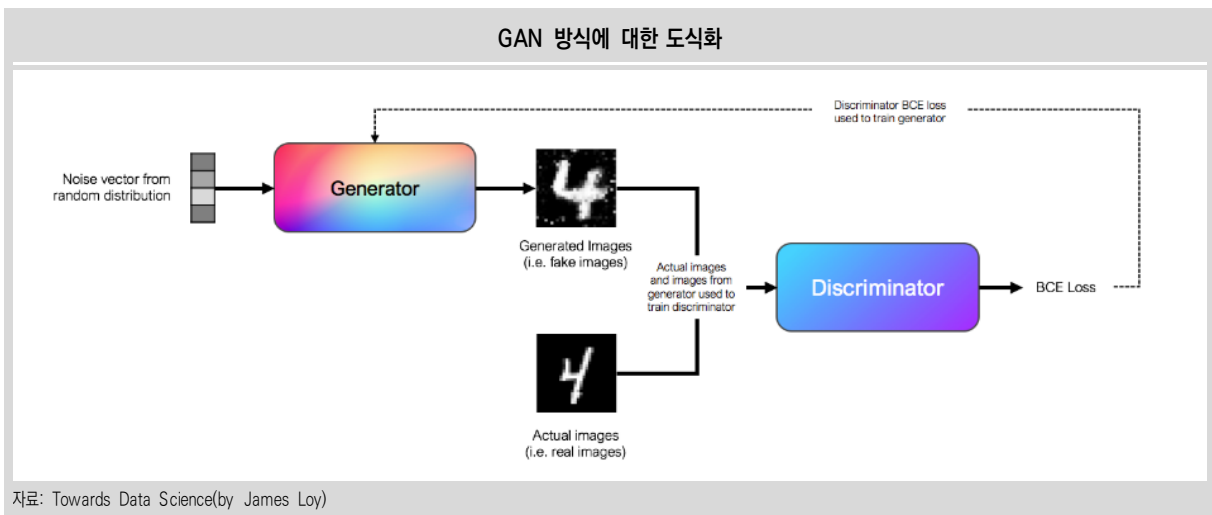
(3) 시각화:

고차원 데이터는 시각화 하기가 어려워 자연스럽게 해석도 어려워진다. 차원 축소는 데이터를 저차원 공간에 투영해, 데이터를 2차원 또는 3차원으로 시각화 하는 것이 가능하다. 이를 통해 데이터의 클러스터 또는 패턴을 식별하고 다양한 기능 간의 관계를 더 잘 이해할 수 있다.



다시 돌아와, 생성자가 하는 일은 결국 실제 데이터와 구분하기 어려운 데이터를 생성하기 위해 학습하는 것이다. 그리고 판별자는 이진분류 결정으로 실제 데이터와 생성자가 생성해 낸 데이터를 구분하는 방법을 학습한다. 흡사 훈련 중에 두 네트워크는 도둑과 탐정처럼 함께 훈련하는 것이다. 생성자는 탐정을 속이려는 도둑이 되고, 판별자는 생성된 데이터가 가짜라는 것을 분류하기 위한 탐정이 된다. 다만, GAN에서의 탐정은 도둑을 도와 힌트를 제공하기도 한다. 생성된 이미지를 올바르게 식별할 때마다 판별자는 생성자에게 향후 더 현실적인 결과물을 만들어 내기 위해 어떻게 수정해야 하는지도 알려주기 때문이다.

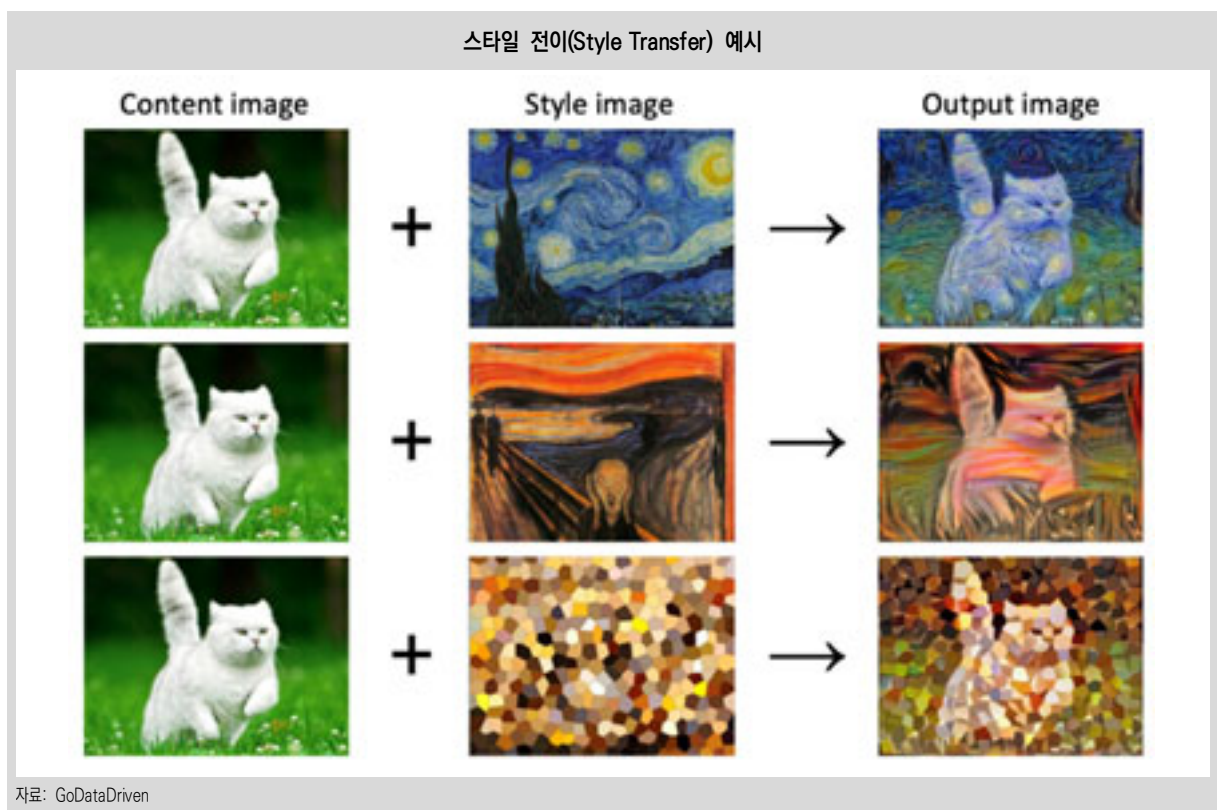
이렇게 두 네트워크가 계속 함께 훈련하다 보면 생성자는 보다 현실적인 데이터를 생성하고, 판별자는 실제 데이터와 가짜 데이터를 더 잘 구분할 수 있게 된다. 그리고 최종 결과는 훈련 데이터 세트와 유사한 새로운 데이터를 생성할 수 있는 최종적인 생성자 네트워크로 수렴한다. 다시 말해, 도둑이 탐정을 적절한 확률(50% 이상)로 속일 수 있을 정도에 이를 때까지 학습이 계속된다는 의미다. 물론, 생성자와 판별자 모델을 동시에 훈련하는 것은 불안정할 수 있다는 한계점도 갖고 있다. 이런 한계가 있지만, 현재 사용되는 신경망 구조 중에서 가장 활용도가 높은 신경망인 것은 틀림없다.



잘 훈련된 GAN은 이미지 합성, 스타일 전이 등의 다양한 응용 분야로 사용된다. 이미지 합성과 스타일 전이는 위에서 언급한 잠재공간에서 표현된 숫자들을 조절해가면서 원하는 결과값을 얻는 과정으로 이해하면 된다. 이를 N차원의 벡터값을 조작하는 잠재공간편집(latent space editing)이라고 부른다. 이런 잠재공간편집 어플리케이션은 다양한 방식으로 활용되고 있다.

(4) 스타일 전이(Style Transfer)

원본 이미지의 내용을 유지하면서 다른 이미지의 스타일을 적용하는 프로세스인 스타일 전이는 예술적 이미지 렌더링, 시각적 효과, 이미지 개선 등 다양한 응용 분야에서 사용된다. 예를 들어, GAN을 사용하여 특정 인물의 사진에 다른 헤어스타일을 입히려고 할 때, 먼저 모델의 헤어스타일이 포함된 잠재공간의 벡터를 찾아야 한다. 이를 위해 여러 이미지를 통해 머리 스타일 차원을 학습하고, 이 차원을 조작하여 원하는 스타일로 변경해, 해당 인물의 사진에 새로운 머리 스타일이 입혀진 이미지를 생성할 수 있다.



(5) 연령 변환(디에이징 등)

인물의 얼굴 사진을 사용하여 해당 인물을 더 나이든 모습이나 더 젊은 모습으로 변환할 수 있다. 먼저 GAN이나 VAE와 같은 생성 모델을 사용하여 "연령 차원"의 벡터값을 학습한다. 그런 다음, 잠재공간에서 연령을 나타내는 차원을 조작하여 원본 얼굴 이미지를 더 나이든 모습으로 변형한다. 이렇게 하면 실제 연령보다 더 나이가 많거나 적은 인물의 얼굴을 보여주는 이미지를 생성하는 것이 가능하다.

2022년 말 KB라이프생명 유튜브 채널은 "이런 만남 처음이지"라는 티저 광고를 게시했다. 윤여정의 대기실이었던 방에는 어떤 젊은 여자가 앉아있다. 윤여정의 젊은 시절 모습인 것 같으면서도 묘하게 다른 느낌이라서 딥러닝 기술인지에 대한 문의가 많았다고 한다.

디에이징 연령 변환 예시 (1/2)



자료: YouTube "KB라이프"

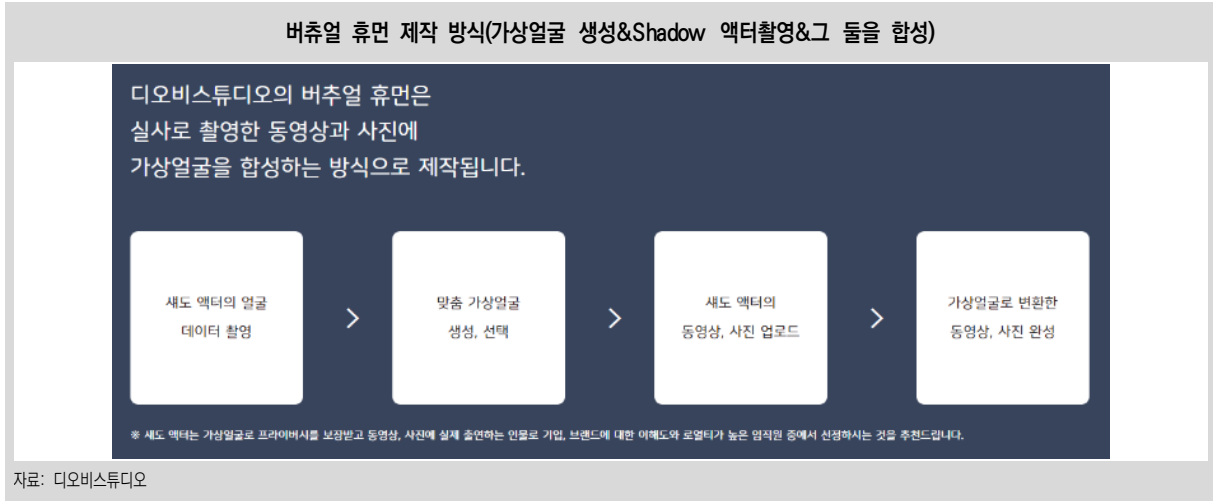
그리고 2023년 새해가 밝자 본편 광고가 공개되었는데 윤여정 배우가 20대의 모습으로 등장했다. KB라이프생명엔 딥러닝 및 디에이징 기술로 광고를 제작했다고 밝혔다. 이 광고는 버추얼 휴먼 루이를 개발한 디오비 스튜디오에서 제작했는데, 디오비 스튜디오는 자체 딥러닝 모델인 "디오비엔진"을 보유하고 있다.

이 엔진을 통해서 여러 얼굴 데이터를 학습해 데이터를 출력해 냈다. 이번 작품을 위해 20대 윤여정 배우의 모습, 현재 윤여정 배우의 촬영 데이터, 광고에서 연기하는 쉐도우 액터의 몸체가 필요했다. 20대의 얼굴과 디에이징한 얼굴을 50대 50으로 섞어 구현했다. 이번 작품은 국내에서 첫 번째로 선보인 딥러닝 디에이징 광고로, 해외의 경우 딥러닝, 디에이징 등의 AI 기술을 활용한 수준 높은 광고, 영화, 뮤직비디오 등의 작품들이 이미 많이 등장한 상황이다.

디에이징 연령 변환 예시 (2/2)



자료: Disney Plus

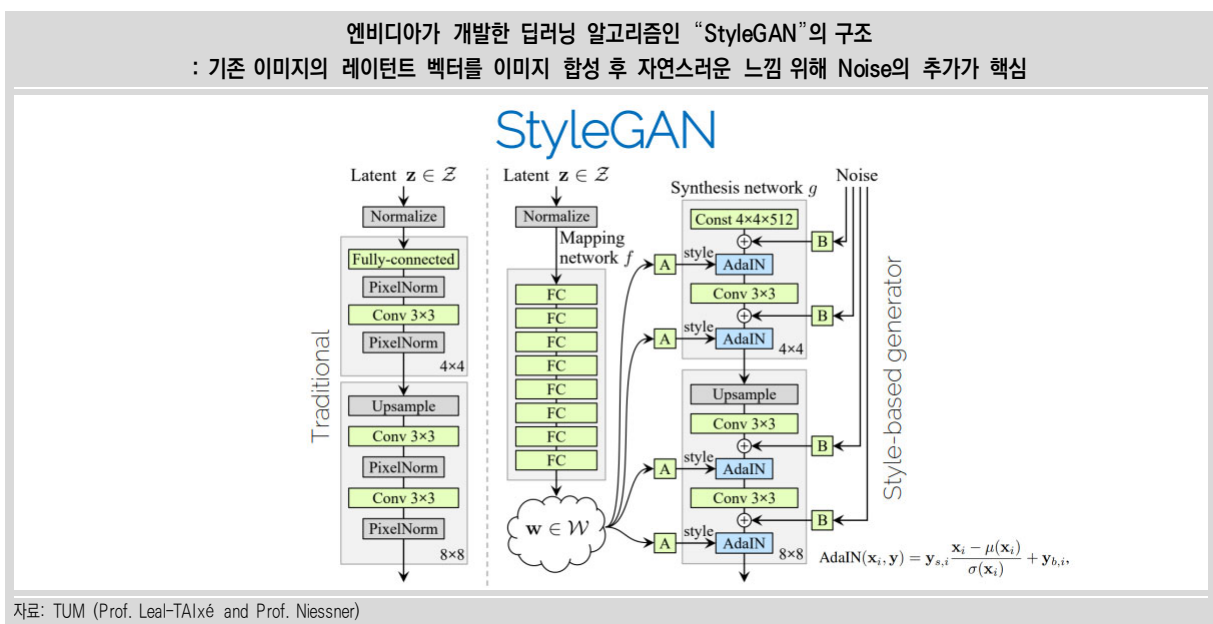


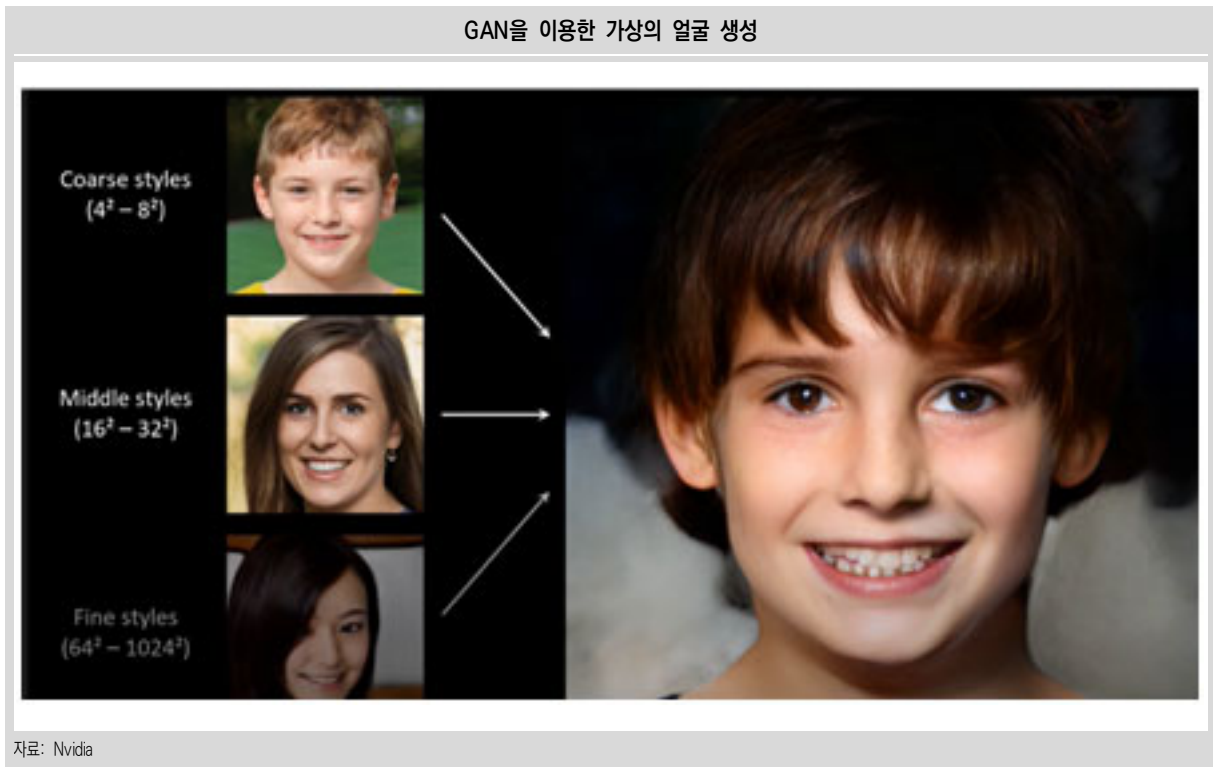
(6) 이미지 합성(Image Synthetic)

GAN이 가장 많이 기여한 것이 바로 이미지 합성 분야다. 두 개의 다른 이미지를 합성하여 새로운 이미지를 생성하는 데 GAN이 사용된다. 예를 들어, 풍경 사진에 구름을 추가하려고 한다면, GAN을 사용하여 풍경 이미지와 구름 이미지의 잠재공간의 벡터를 얻은 뒤, 두 벡터 값을 적절하게 결합해 새로운 벡터를 생성한다. 이후 새로운 벡터를 디코딩하면 원래 풍경 이미지에 구름이 있는 새로운 이미지를 얻는다.

엔비디아가 2019년 3월에 개발한 StyleGAN이라는 모델은 얼굴 포즈, 주근깨, 머리카락과 같은 사람들의 외형적 속성을 학습해 가상 인물의 고해상도 사진을 생성해내는 데 성공했다. 그리고 2020년 2월에 나온 StyleGAN2는 아키텍처와 훈련 방법 모두 개선되어 최첨단 품질을 보여줬다. 2019년 3월 엔비디아는 GPU 기술 컨퍼런스(GTC)에서 GauGAN이라는 적대적 신경망 기술을 공개하기도 했는데, 이 시스템은 실존하지 않는 현실적인 가상의 풍경 이미지를 만들 수 있었다.

GauGAN의 머신러닝 모델은 Flickr(온라인 사진공유 커뮤니티)의 100만 개 이상의 이미지로 눈, 나무, 물, 꽃, 관목, 언덕 및 산 등 180개 이상 대상 간의 관계를 이해할 수 있도록 학습됐다. 실제로는 물 옆에 나무가 있으면 물에 반사가 나타나고, 그림에서 나타낸 계절에 따라 강수의 종류(눈/비)가 달라지기도 하는 등 대상 간 관계가 반영되는 모습을 보였다.





이처럼 잠재공간편집은 위와 같은 이미지 변형을 손쉽게 수행할 수 있게 해준다. 또한 광고, 예술 및 디자인 작업 등 다양한 분야에서 이미 활발히 활용되고 있는 기술이라고 할 수 있다. 아래는 GAN 및 잠재공간편집 작업으로 활용될 수 있는 여러 섹터에 대해 열거한 것들이다.

- (1) 스타일 합성: 초고해상도 업스케일링, 사물변환, 다른 각도에서 촬영한 것처럼 보이는 이미지 생성, 대상 생김새 보존한 채로 애니메이션 합성
- (2) 영상: 몇 개의 비디오 프레임만으로 미래 이벤트를 예측한 48프레임의 짧은 비디오 클립을 완성
- (3) 예술작품: 로고 생성, 가상 은하 이미지 생성, 풍경화 그리기 등
- (4) 음악: 멜로디 생성, 음악 공백 메우기 등
- (5) 음성: 음성을 생성할 수 있는 text-to-speech 시스템
- (6) 의학: 뇌 MRI 이미지 합성, 질병 유발 분자구조 생성, 치료법 개발 등
- (7) 로봇공학: 동작 데이터 없이 위치 정보만 제공해도, 로봇이 작업을 완료할 수 있도록 교육
- (8) 딥페이크 탐지: 제한된 레이블 데이터만으로도 기존 스팸 탐지기술보다 높은 정확도로 탐지

그런데, GAN은 불안정한 학습과 긴 학습 시간이라는 고질적인 문제를 갖고 있었다. 이 두 문제는 GAN이 고해상도의 이미지를 생성하는 것을 힘들게 만들었다. 128*128짜리 이미지를 만드는 모델 학습에도 긴 시간이 소요되고, 학습 안정화를 위해 이것저것 신경써야 할 게 많았다. 그렇기 때문에 예전에는 GAN이 이미지 생성 모델 설계에 있어 기본 선택이었지만, 점차 밀려나게 되어, DALL-E 2와 같은 확산모델(Diffusion model)이 대규모 생성 모델의 새로운 기준이 된 것이다.

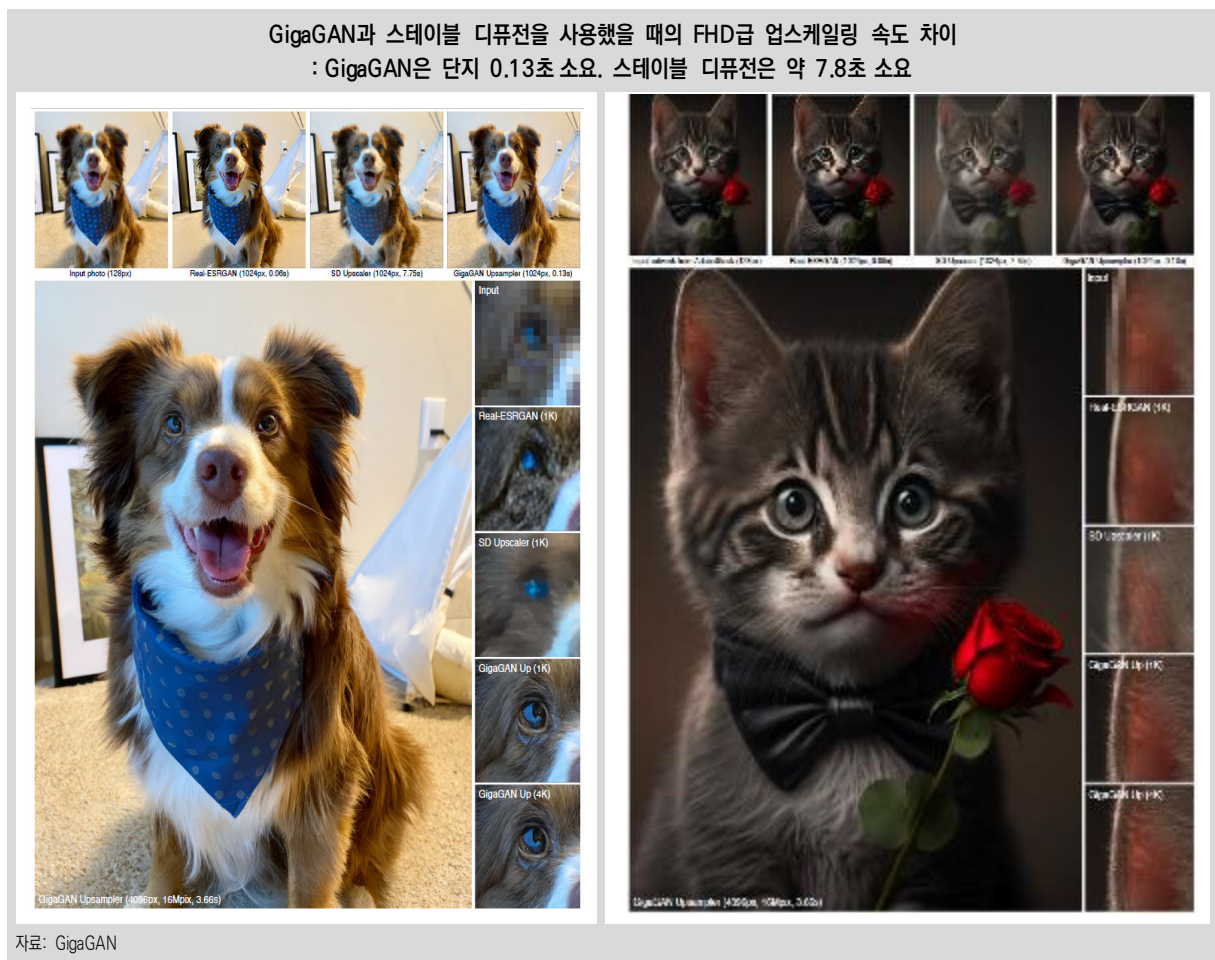
그렇다면 더 이상 GAN을 알 필요는 없어진 걸까? 그것은 아닌 것 같다. 최근 GigaGAN이라는 새로운 아키텍처가 발표되었다. 이 새로운 GAN 아키텍처는 기존의 한계를 훨씬 뛰어넘어, 텍스트에서 이미지로의 변환에 있어 GAN도 가용한 옵션이라는 것을 보여주고 있다.

GigaGAN은 확산모델과 및 AR 트랜스포머 모델들과 비교하여 세 가지 주요 실용적 이점을 지닌다.

첫째, 추론 시간이 수십 배 빠르며, 512*512의 픽셀 이미지를 무려 0.13초 만에 합성할 수 있다. 기존 스테이블 디퓨전에서는 4초가 소요된다는 측면에서 엄청난 혁명이다.

둘째, 고해상도 이미지를 합성할 수 있다. 예를 들어, 1600만 화소의 이미지를 3.66초 만에 (업스케일링 하여) 생성할 수 있다.

마지막으로, GigaGAN은 잠재공간편집 애플리케이션을 지원하기 때문에, latent interpolation, 스타일 합성 같은 작업을 수행할 수 있다.



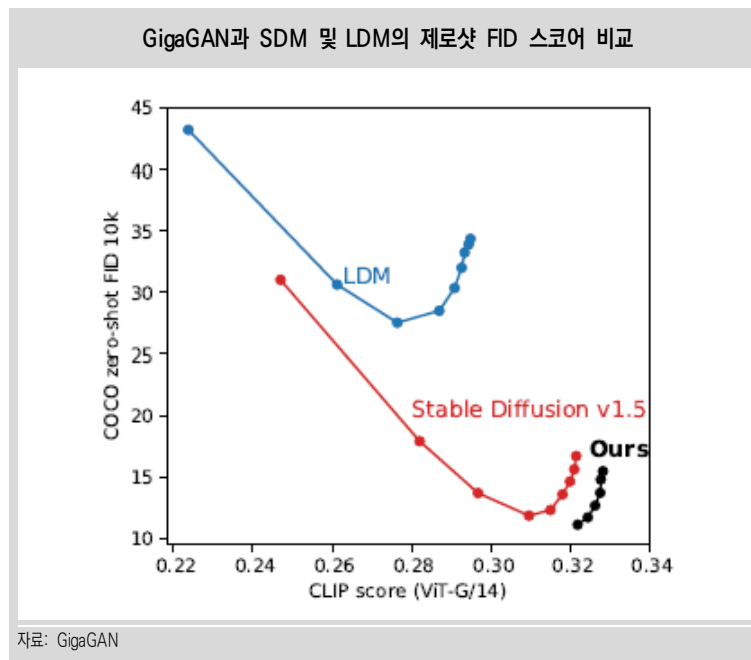
아래는 GigaGAN이 고품질 이미지를 생성하는 프로세스를 정리한 것이다.

- (1) 메모리와 실행 시간을 줄이기 위해 64x64 해상도에서 GAN을 통해 이미지를 먼저 생성
- (2) GAN을 사용해 256x256 및 1024x1024로 순차적으로 계단식으로 업샘플링
- (3) 이후 GigaGAN 모델을 통해 1K와 4K 해상도까지 고속으로 업스케일링

뿐만 아니라 GigaGAN은 COCO 2014 데이터 세트(마이크로소프트가 제안한 컴퓨터 비전 알고리즘을 평가하기 위한 벤치마크)에서 제로샷 FID 9.09를 달성했다. 이는 DALL-E 2가 활용한 LDM이나 스테이블 디퓨전의 FID보다 낫다는 것을 보여줬다.

*FID는 Frechet Inception Distance(프레셰 인셉션 거리)의 약자로, 실제 데이터와 생성된 데이터 간의 차이를 측정하는 지표. FID 값이 작을수록 더 좋은 성능. Zero-shot FID는 훈련 데이터와 평가 데이터가 서로 다른 경우, 생성된 이미지데이터와 "평가" 데이터의 FID를 의미.

Zero-shot FID는 생성 모델의 일반화 성능을 측정. 즉, 생성 모델이 이전에 본 적이 없는 데이터를 잘 생성할 수 있는지 평가하는 데 사용. Zero-shot FID가 낮을수록, 생성 모델이 새로운 데이터에 대해 더 잘 일반화할 수 있음을 나타냄.



애초에 GigaGAN은 단독으로 쓰일 것은 아니기 때문에, DALL-E 2에서 사용된 것과 같은 명령어 입력에 대해 사진 같은 실제감과 고품질 결과물을 생성하지는 못한다. 실제로 GigaGAN 관계자들은 결과물의 시각적 품질은 아직 DALL-E 2와 같은 모델과 비교할 수 없다고 언급하기도 했다.

그럼에도 불구하고, 기존 접근법으로는 불가능했던 용량만으로 AR 트랜스포머 및 확산모델에 비해 경쟁력 있는 시각적 품질을 달성했다. 더군다나 수십 배 빠른 속도로 latent interpolation과 스타일 전이를 가능하게 했다는 점도 놀랍다. 이것은 GAN이 여전히 텍스트에서 이미지로의 변환에 가용한 옵션이며, 미래의 공격적인 확장을 위해 GAN이 활용될 수 있음을 시사한다.

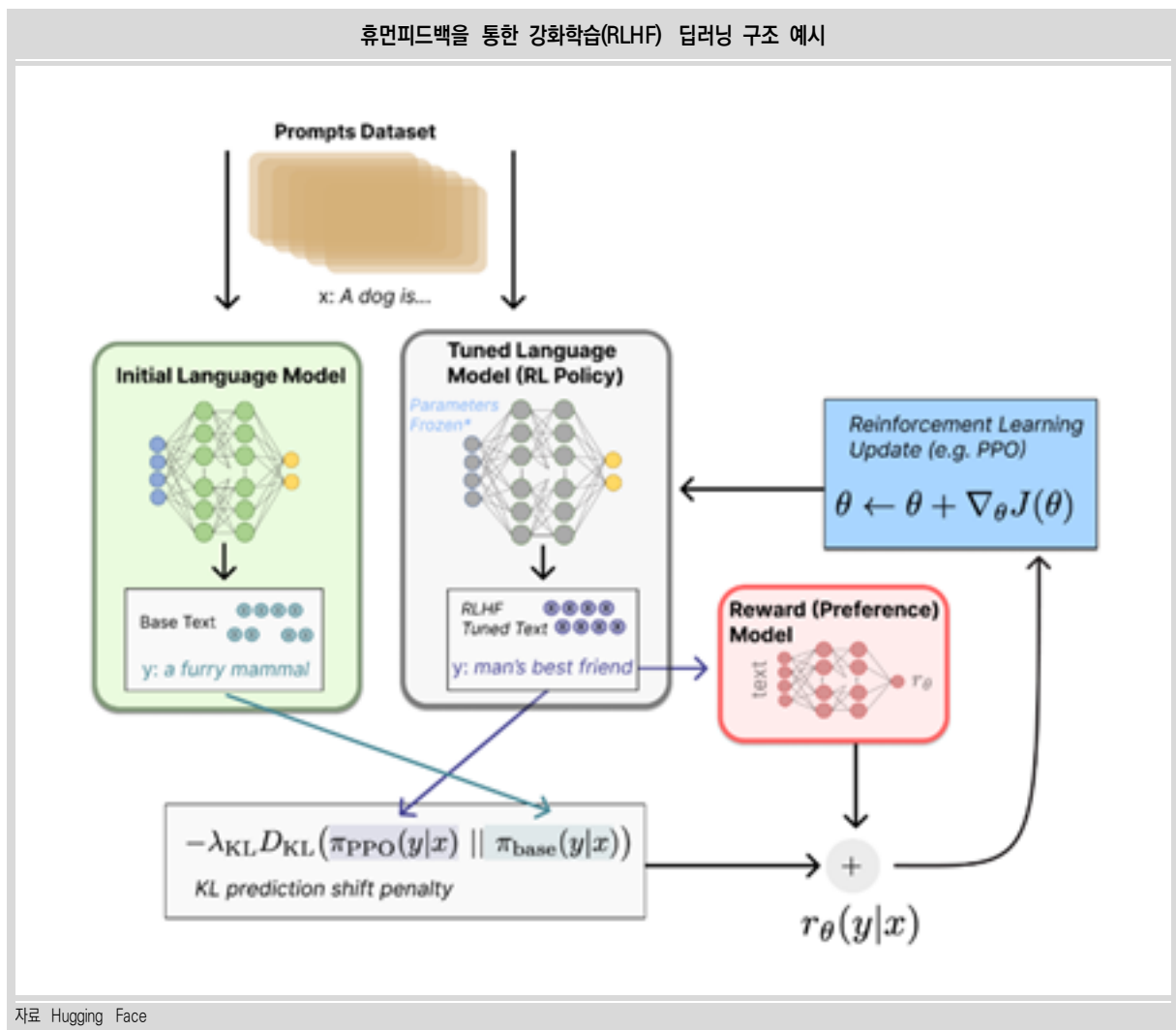
또한 GigaGAN은 아직까지 모델 크기에 대해서 "품질 향상 포화(saturation)"가 관찰되지 않았다고 하니 앞으로도 관련 기술은 더욱 발전할 가능성이 크다. GigaGAN 구조는 대규모 생성 모델의 새로운 대중화를 열어줄 것으로 믿는다. 이때 GigaGAN 등 다양한 생성 모델들은, 함께 공존하는 형태로 오랫동안 자리하게 될 것으로 생각된다.

4) 강화학습(Reinforcement Learning)

강화학습은 지도학습이나 비지도학습과는 다른 차원의 문제다. 비지도학습처럼 답안지가 처음부터 존재하지 않는다. 대신 시간 개념이 투입되는데, 이는 강화학습이 에이전트가 어떤 행동을 한 뒤, 그 행동을 평가하는 방식이기 때문이다. 이세돌을 이긴 “알파고”도 이 강화학습으로 탄생했다. 바둑돌을 어딘가 착수할 때, 그때그때 좋은 수를 두었을 때만 리워드를 준다. 이런 방법으로 에이전트는 점점 더 최적의 답을 찾게 된다.

매 행동마다 좋은 결과물이었는지 판정하는 기준은 사람이 제공한다. 이것을 RLHF(인간 피드백형 강화학습)라고 말한다. 에이전트는 인간의 피드백에 따라 시행착오를 반복하면서 계속 학습하게 되는데, 이 때문에 평가기준이 명확한 문제와 규칙이 명확한 게임 등에 아주 적합하다.

이와 관련해 2017년 OpenAI는 "인간 선호도로부터 학습하는 심층 강화학습(RLHF, Reinforcement Learning from Human Feedback)"이라는 논문을 발표했다. 대규모 강화학습 작업을 해결하기 위해 사람의 피드백을 포함하는 아이디어를 도입한 것이다. 이로써 InstructGPT(2022년 1월)와 그것의 스피노프였던 ChatGPT(2022년 11월)가 출시될 수 있었다. ChatGPT는 해당 언어모델이 사람들과 더 잘 대화하도록 만들기 위해 인간의 피드백을 받아 학습했다. 사람들은 ChatGPT가 내놓는 질의응답에 순위를 매겨 모델을 훈련시켰다.



RLHF 훈련과정은 보통 3단계의 피드백 과정을 거친다.

a. 무작위 행동(Random actions)

AI 에이전트는 주어진 환경에서 무작위로 행동을 개시한다. 이는 훈련 과정의 ‘탐색’ 단계로, 임의의 행동을 취하고 결과를 관찰한 뒤 데이터를 수집하여 환경 모델을 구축한다.

b. 사람의 피드백(Human’s Feedback)

AI 에이전트는 자신의 행동에 대한 두 개의 비디오를 인간에게 제시한다. 사람들은 클립을 보고 어떤 행동이 목표를 달성하는 데 더 가까운지 결정하고 AI 에이전트에게 피드백을 제공한다. 이는 목표와 보상의 모델을 구축하는 데 도움이 된다.

**근접 정책 최적화(PPO: Proximal Policy Optimization)라는 기술이 사용됨. PPO는 주어진 작업에서 보상을 최대화하기 위해, 먼저 작은 단계로 모델의 동작을 조정하는 것. 이것은 모델이 한 번에 동작을 변경할 수 있는 정도를 제한하여 학습시킴으로써 큰 실수를 방지. 무엇보다 PPO는 두 가지 주요 부분인 탐색(새로운 것을 시도)과 Exploitation(과거에 효과가 있었던 것을 사용)의 밸런스를 맞추기에 효과적.*

c. 학습 및 개선

에이전트는 사람의 피드백을 기반으로 행동을 조정하는 시행착오를 거친다. 만약 어느 것이 더 나은지 불확실한 경우에는 피드백을 계속 요청하기도 한다. 이 피드백 과정은 인공지능이 제대로 된 결과를 낼 수 있을 때까지 계속 반복한다.

최신 유행 중인 AI 기법들

1) 빅테크들의 모델 경쟁

2014년과 오늘날의 이미지 생성기술을 비교해보면, 이 정도로 품질이 좋아질 것이라고는 그 당시에는 아무도 예상하지 못했다. 현재 속도로 진보가 계속된다면 GAN을 포함한 여러 이미지 생성 모델들은 앞으로 딥러닝의 매우 중요한 연구 프로젝트로 남게 될 것이다.

더군다나 딥러닝의 발전은 스타일 전이와 이미지 합성 방식을 ‘실시간’으로 가능하도록 만들기에 이르렀다. 스타일 전이 어플리케이션까지 실시간으로 사용된다는 것은 의미하는 바가 크다. 시각 효과와 비디오 편집은 현대 미디어 환경에서 굉장히 일반적이다. 비디오 중심 플랫폼이 인기를 얻으면서 직관적이고 높은 성능의 비디오 편집 도구에 대한 수요가 증가한 상황인데, 비디오 데이터의 ‘시간적 특성’ 때문에 스타일 편집은 여전히 복잡하고 시간이 많이 소요되는 문제가 있었다. 하지만, 앞으로는 비디오 처리와 라이브 스트리밍과 같은 응용 분야에까지 손쉽게 사용될 수 있을 것으로 전망된다.

이와 관련해, OpenAI의 CEO인 샘 알트만 또한 텍스트에서 비디오를 생성할 수 있는 AI 모델을 출시할 계획이라고 밝히기도 했다. 뿐만 아니라, 빅테크들은 이미 제한된 형태이긴 하지만 텍스트 to 비디오 모델을 보여주기도 했다.

메타의 산하 AI 리서치는 작년 9월에 Make-A-Video라는 논문을 발표했다. 텍스트 프롬프트에서 자동으로 비디오 클립을 생성한다. 프롬프트를 입력하면 (예: 말이 물을 마시는 모습), AI 시스템이 자동으로 해당 내용의 비디오를 생성하는 것이다. 해당 구조에 대해 조금 더 자세히 설명하자면, 아래와 같다.

- (1) Text-to-image: 텍스트를 받아서 낮은 화질의 이미지를 생성하고, 이를 고화질로 업스케일링
- (2) 시공간(Spatiotemporal) 레이어: 이미지 처리를 위한 convolution layer과 attention layer에 각각 “시간” 차원(temporal dimension)을 추가
- (3) 프레임 보간 네트워크: 프레임 숫자를 증가시켜서 더 부드러운 영상을 만들거나, 영상 전후에 프레임을 추가해 영상 길이를 늘림(엔비디아의 DLSS와 비슷한 개념).
- (4) 훈련: 위의 각 과정을 독립적으로 학습

구글도 이에 질세라 며칠 뒤 “Imagen Video”라는 이름의 자체 비디오 생성형 AI 시스템을 발표했다. 이 시스템 또한 입력된 텍스트 프롬프트를 자연어 처리 AI인 “T5”로 처리한 다음, 디퓨전 모델로 영상을 생성하는 방식이다.

이들은 비디오 확산모델로서 24×48의 해상도, 초당 3프레임으로 총길이 16프레임의 영상을 생성한다. 해상도나 프레임 수준은 아직 떨어진다. 하지만 자연어 명령어 기반 이미지 생성형 AI가 나온지 2년도 채 되지 않아 비디오 생성까지 도달했다는 점에서 그 속도감은 압도적이다.

메타의 시리서치가 2022년 9월에 발표한 Text-to-Video 딥러닝 모델인 Make-A-Video 논문
 : 자연어 프롬프트만 가지고 '움짤' 정도 길이의 영상 생성



(a) A dog wearing a superhero outfit with red cape flying through the sky.



(b) There is a table by a window with sunlight streaming through illuminating a pile of books.



(c) Robot dancing in times square.



(d) Unicorns running along a beach, highly detailed.

자료: Meta

구글이 2022년 10월에 발표한 Text-to-Video 딥러닝 모델인 Imagen Video 논문
 : 메타에서 발표한 것에 뒤이어, 구글도 며칠 뒤 바로 공개



Wooden figurine surfing on a surfboard in space.



Balloon full of water exploding in extreme slow motion.



Melting pistachio ice cream dripping down the cone.



A british shorthair jumping over a couch.



Coffee pouring into a cup.

자료: Google Research

2) 이제 목표는 text-to-video

그리고 몇 주 전, 텍스트를 비디오 형태로 생성해주는 굉장한 딥러닝 모델이 발표됐다. Runway라는 회사가 텍스트 프롬프트를 사용하여 기존 비디오를 새로운 스타일로 변환하는 모델 Gen-1을 출시한 것이다. Runway는 Stability AI와 함께 오픈소스 자연어 처리 이미지 생성 모델인 스테이블 디퓨전을 내놓았던 회사다. Runway에서 제공하는 스타일 전이 모델은 역시 딥러닝 기반의 알고리즘이다. 하나의 이미지(콘텐츠 이미지)에서 스타일을 추출한 다음에, 다른 이미지(대상 이미지)에 적용하는 기술이다.

스테이블 디퓨전을 만든 회사가 이번에는 비디오 생성형 AI에 도전
: 원본 비디오에 원하는 프롬프트를 주면, 그에 맞게 원본 비디오를 새로이 탈바꿈시킴

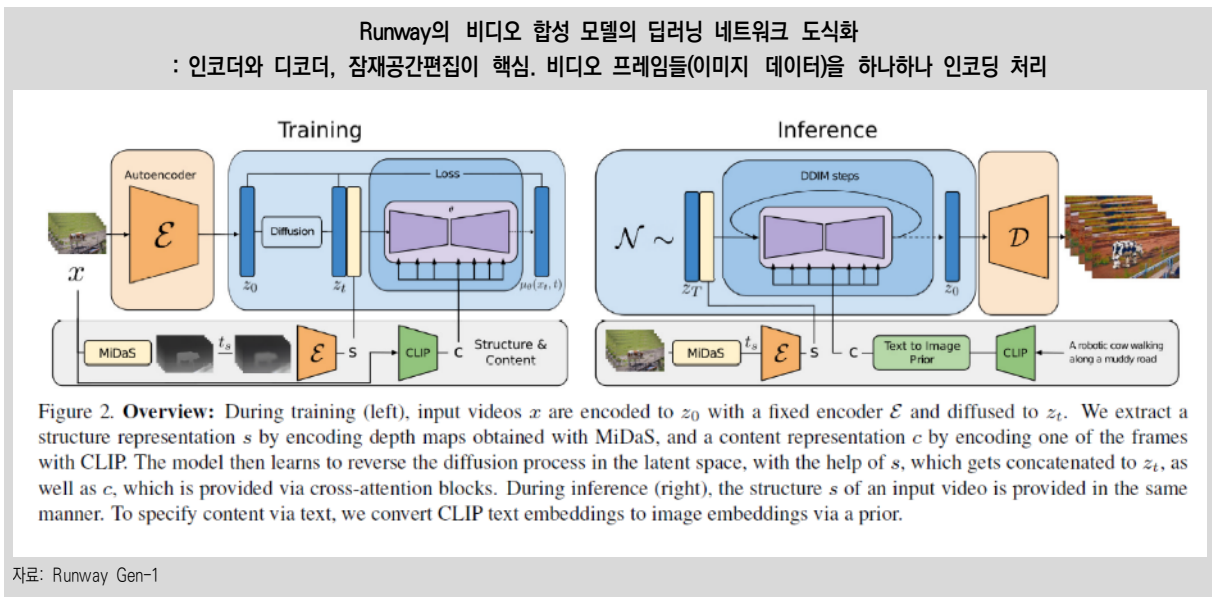
Figure 1. **Guided Video Synthesis** We present an approach based on latent video diffusion models that synthesizes videos (top and bottom) guided by content described through text (top) or images (bottom) while keeping the structure of an input video (middle).

자료: Runway Gen-1

Runway는 여러 스타일 전이 모델을 포함하고 있다. 사용자들은 쉽게 소스 이미지와 스타일 이미지를 업로드하고, 처리된 결과도 미리 볼 수 있게 된다. 이 모델은 이미지 합성 모델과 유사하게, 추론된 비디오의 콘텐츠(외관이나 스타일)가 사용자가 제공한 이미지 또는 텍스트 프롬프트와 일치하도록 훈련됐다.

그리고 무엇보다 한정된 연산 자원(GPU)으로부터 딥러닝 학습을 가능하게 하기 위해서 오토인코더의 잠재공간 개념을 활용했다고 논문에서 적시해 놨다. 압축된 잠재공간 픽셀 데이터에 기반한 모델에 비해 메모리(계산) 요구사항과 런타임을 줄일 수 있기 때문이다. 그만큼 잠재공간을 통한 차원감소, 다운샘플링 등은 중요한 개념이라고 할 수 있다. 이처럼 딥러닝은 벡터의 연산과 조작성이 핵심이다.

마지막으로 Runway측은 논문의 마지막 부분에 “우리는 이 모델이 해로운 목적으로 사용되지 않기를 바라며, 생성 모델의 남용과 싸우기 위한 추가 연구가 이루어질 것이라고 기대한다”고 썼다. 이 부분은 텍스트-to-비디오 기술이 향후 딥페이크 영상 등으로 오남용 될 수 있는 여지가 크기 때문이다. 경계해야 할 부분이긴 하지만, 어떻게 보면 그만큼 성능이 뛰어나다는 방증이다.



3) 사전훈련모델(Pre-Trained)과 전이학습(Transfer learning)

모델이 이렇게 다양한 방식으로 학습되는 과정에는 엄청난 양의 데이터와 연산 능력이 필요하다. 따라서 목적이 달라질 때마다 매번 새롭게 모델을 학습시키는 것은 현실적으로 불가능하다. 이를 위해 대부분 사전훈련모델(Pre-Trained model)을 사용한다. 큰 사이즈의 데이터로 이미 학습되어 있는 모델을 도입해서, 내가 원하는 방식으로 미세조정(fine tuning)만 진행해 사용하면 적은 데이터와 시간으로도 좋은 결과를 낼 수 있다.

전이학습(Transfer Learning)은 새로운 작업을 효율적으로 해결하기 위해서, 유사한 기존 작업에서 얻은 학습 데이터와 학습 결과를 재사용하는 것을 말한다. 비유를 하자면, 사람이 새로운 지식을 배울 때 그전까지 쌓아왔던 지식을 다시 사용하는 것과 같다. 모델 알고리즘이 서로 관련된 도메인 정보의 경우, 큰 데이터 세트로 훈련된 모델을 또 다른 시작점으로 해서 비슷한 문제도 풀 수 있을 것이라는 아이디어에서 나온 것이다.

**인공지능(AI)에서 '도메인'이란 특정 분야나 산업, 응용 프로그램에 대한 지식 및 데이터를 의미. 도메인은 AI 시스템이 특정 문제를 해결하는 데 필요한 문맥, 전문 지식과 데이터, 그리고 관련된 패턴을 제공하는 공간이라고 볼 수 있음. 도메인 지식은 AI 시스템이 예측, 분류, 추천 등의 작업을 수행할 때 더 정확하고 효율적으로 결과를 도출하는 데 도움. AI 모델이 도메인 지식을 활용하면 해당 분야에서 발생하는 복잡한 문제를 더 잘 이해하고 해결할 수 있음.*

예를 들어, 의료 분야에서 AI 시스템이 환자의 진단을 돕는 데 사용되려면 병리학, 질병 분류, 진단 기준 등의 의료 도메인 지식이 필요. 마찬가지로, 금융 도메인에서 AI 시스템이 주식 가격을 예측하는 데 사용되려면 금융 시장의 원리, 경제 지표, 회사 재무 정보 등의 도메인 지식이 필요.

예를 들어, 소스 도메인(한국어 모델, 화상 및 물체 인식)과 타겟 도메인(한영번역, 복장 인식)을 각각 갖고 있다고 생각해보자. 한국어 모델과 화상 및 물체 인식은 한영번역과 복장인식과 문제와 유사하지만, 동일하지는 않다.

한국어 모델과 화상 및 물체 인식을 해결하는 데에 사용할 데이터와 지식은 많다. 반면, 한영번역과 복장 인식을 해결하기 위한 데이터와 지식은 부족하다. 이때 소스 도메인의 지식과 데이터로 타겟 도메인의 문제를 해결하는 것이 바로 전이학습이다. 다시 말해, 사전에 학습된 임베딩 벡터를 시작점으로 사용해, 다른 업무에 관한 새로운 모델로 재활용하는 것으로 이해할 수 있다. 이러한 재활용의 과정이 앞에서 언급한 미세 조정이다.

이렇게, 전이학습은 첫 번째 작업으로부터 얻은 지식을 활용해서 두 번째 작업에 관한 모델 성능도 향상시킬 수 있다. 이는 데이터 양을 줄이고 컴퓨터 연산에 필요한 리소스를 줄이는 효과를 가져온다. 전이학습은 컴퓨터 비전과 자연어처리, 음성인식 등에서 널리 쓰이고 있다. ImageNet같은 대규모 데이터세트에서 훈련된 모델로서, 개체감지와 같은 특정작업을 위해서 더 작은 데이터세트에서 미세조정되는 식으로 사용한다.

물론 그 유명한 GPT 등에도 전이학습이 적용됐다. GPT는 '다음 단어 맞히기'로 사전학습된 모델이다. 예를 들어, 소리 없는 XXX라는 단어나열에서 문맥을 파악해 XXX가 아우성이라고 예측해내는 게 다음 단어 맞히기다. 좀 더 자세히 말하면, 입력된 문장 다음에 올 단어가 얼마나 적절한지를 나타내는 확률 함수다. 그리고 다음 단어 맞히기로 사전학습된 모델을 우리는 언어모델이라고 한다. 그래서 언어모델이 문장 생성 모델이라고 불리기도 한다.

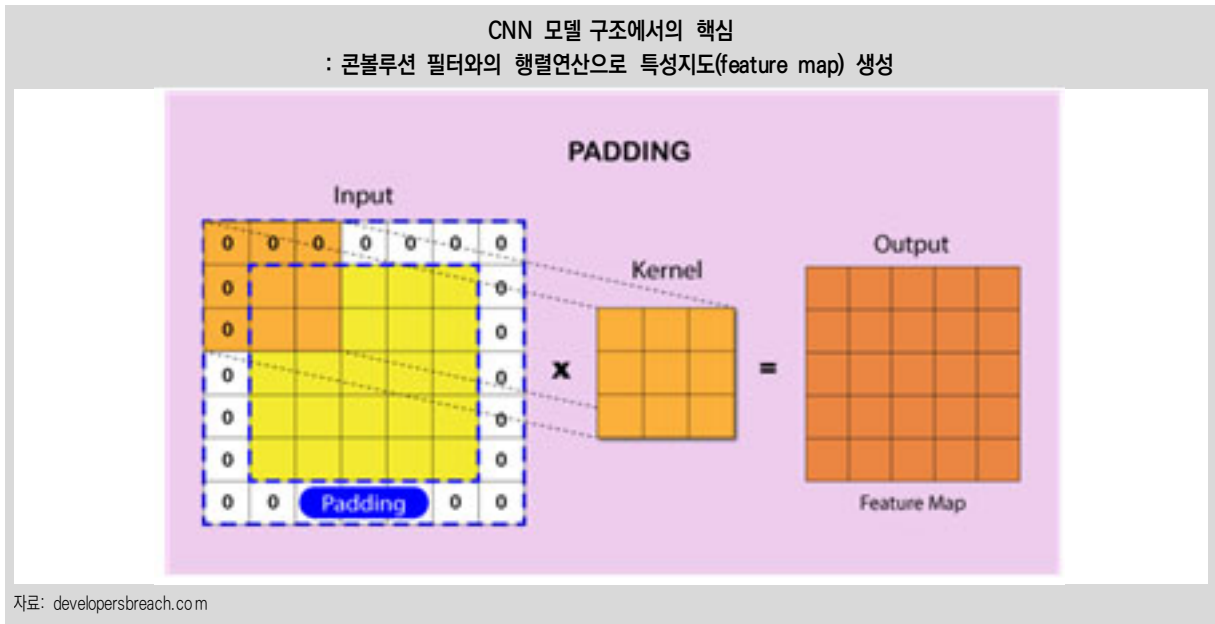
4) 합성곱신경망(CNN: Convolutional Neural Network)

CNN은 이미지의 패턴을 처리하고 인식하는 데 특히 뛰어난 딥러닝 알고리즘이다. 트랜스포머가 나오기 전까지만 하더라도, 이미지 처리에 가장 많이 사용됐던 게 CNN이다.

그런데, 딥러닝 연산은 기본적으로 '1차원'의 데이터를 사용한다. 예를 들어, 1028x1028 화소 수를 갖는 2차원 이미지가 입력값인 경우, 한줄의 데이터로 쭉욱 늘어뜨려 연산해야 한다. 그런데 이 늘어뜨리는 평탄화(flatten) 과정에서 이미지 정보에 손실이 생긴다(공간적/지역적 정보가 소실).

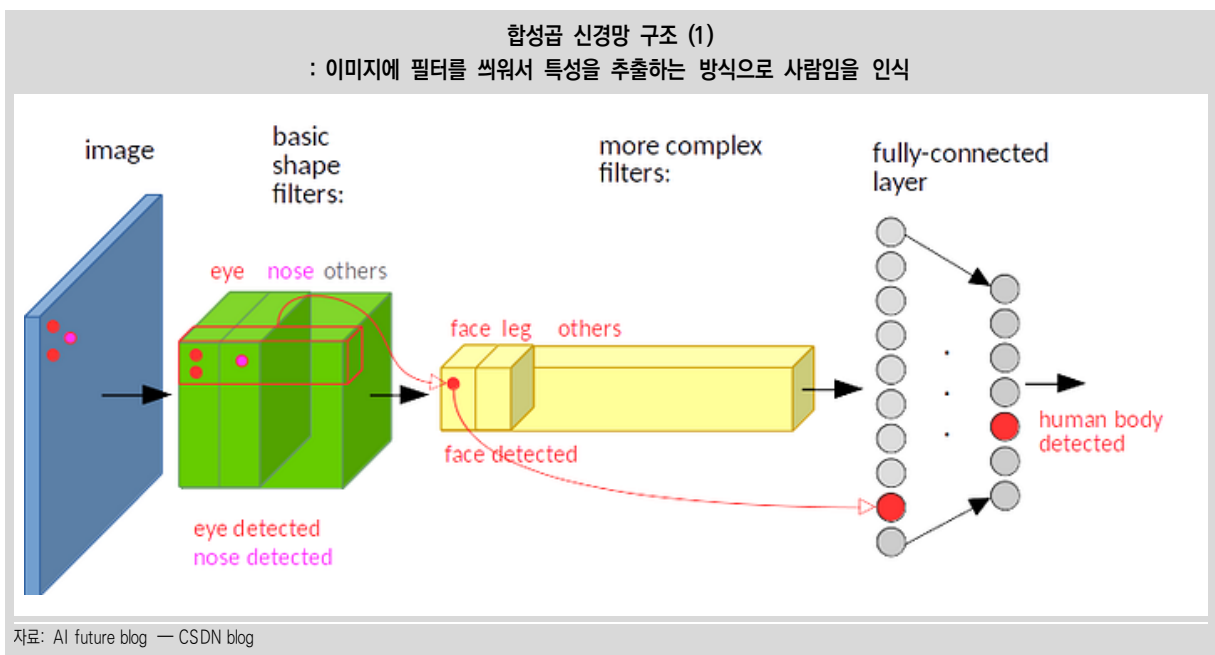
즉, 언어나 숫자와 달리, 이미지나 영상 등은 고차원 데이터이기 때문에 다른 방법을 사용해야 한다는 인식이 생겼다. 이런 측면에서 CNN은 이미지를 날것(raw input) 그대로 받음으로써, 공간적/지역적 정보를 유지하면서 특성(feature)을 추출하는 방식이다. CNN은 convolutional layer, pooling layer, 그리고 fully connected layer로 구성된다.

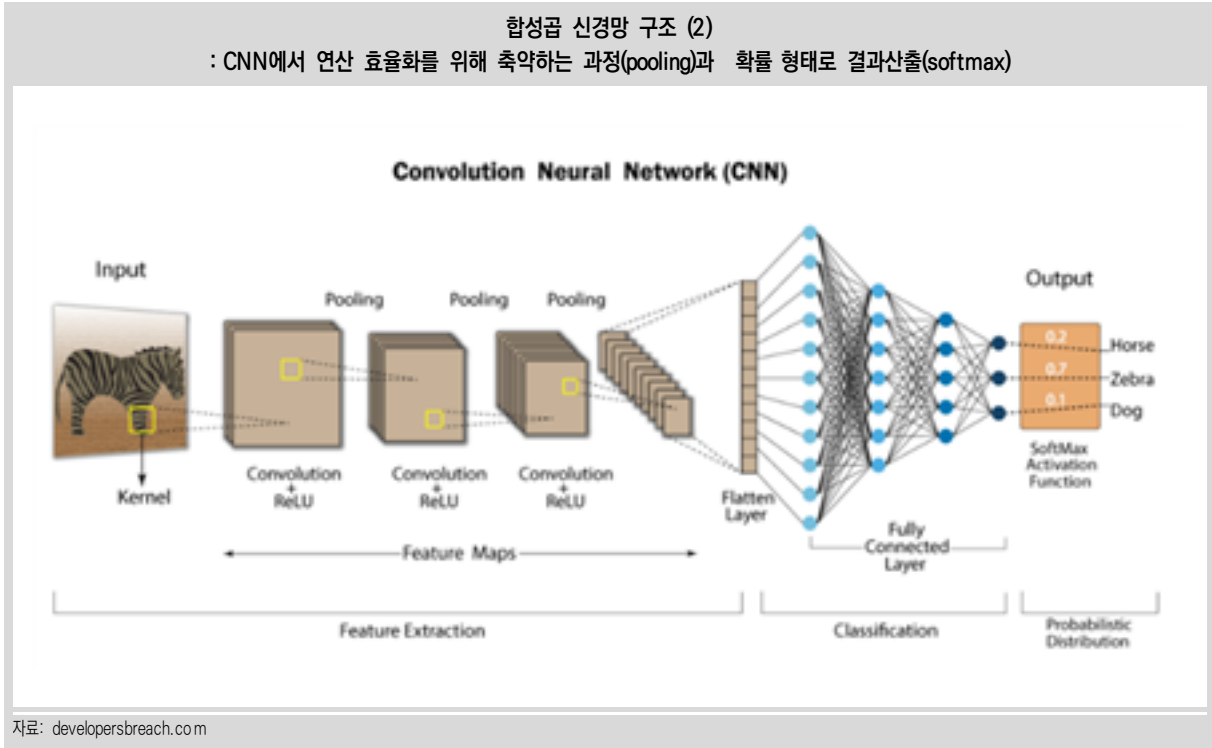
Convolutional layer는 쉽게 말해 그림에서 원하는 부분만을 강조하는 과정이다. 특정 모양에 '가점'을 주는 틀(CNN에서는 이 틀을 convolution filter 혹은 Kernel이라고 함)을 가지고 그림 위 전체로 이리저리 움직이면서, 원하는 부분이 강조된 그림들을 분류해낼 수 있는 것이다. 이 과정은 필터를 이루는 벡터 값과 이미지의 벡터값을 행렬연산을 시켜서 일종의 '특성 지도'를 만든다.



이 말을 예시를 통해 쉽게 이해해보자. 여러 동물들 사진 중에 해당 사진이 토끼인지 아닌지 분류를 한다고 가정해보자. 우리가 토끼를 볼 때 가장 먼저 바라보는 것은 귀일 것이다. 따라서 토끼의 주요 특징인 귀가 이미지 분류를 하는데 있어 중요한 포인트가 될 수 있다. 때문에 CNN 필터는 그림 전체를 이리저리 훑어가면서, 주어진 이미지에 토끼 귀가 있는지 없는지를 감지하는 것이 가장 중요해진다. 그런데, 여기서 전체 이미지 그림에서 귀는 적은 영역만 차지할 것이다. 동물 사진을 촬영할 때 동물의 귀만 대고 찍지는 않으니까 말이다.

그렇기 때문에 신경망은 전체 이미지를 보는 것 보다는 이미지의 귀 부분만을 잘라 보는게 더 효율적인데, 그작업을 하는 것이 CNN이다. CNN의 뉴런이 패턴(이 경우에는 토끼의 귀)을 파악하기 위해서 전체 이미지를 모두 다 볼 필요가 없다. 다시 말해, 연산량을 줄이는 효율화 과정을 거치는 것이다.





이를 위해 그림의 크기를 줄이는 과정을 거치는데, 이를 수행하는 것이 pooling layer이다. 일정 구획 당 평균값을 산출해서 축약된 그림을 만드는 과정인데, 포토샵의 블러 처리 방식과 유사한 개념이다.

Convolutional layer와 pooling layer를 반복하면서 이미지의 크기는 점차 작아지게 된다. 그 과정에서 중요한 정보가 사라질 우려가 존재하므로, 이를 막기 위해 그림 가장자리에 일종의 여백인 패딩을 추가하기도 한다. CNN에서는 주로 0값을 갖는 픽셀을 가장자리에 추가하는 zero-padding을 이용하지만, 그 외 다양한 값이 패딩으로 사용될 수 있다.

중요한 값을 강조한 뒤 축약된 그림은 fully connected layer를 거치게 된다. Fully connected란 모든 입력 노드가 출력 노드와 연결되었음을 의미한다. 각 노드는 각기 다른 특징을 강조하는데, 이를 모두 합함으로써 그림의 전체적인 아웃라인을 얻을 수 있게 된다. 이로써 해당 그림이 무엇을 그렸는지를 파악할 수 있다.

5) 순환신경망(RNN: Recurrent Neural Network)

시간에 따라 변화하는 데이터 역시 새로운 접근이 필요하다. 주가와 같은 시계열 데이터는 과거 값이 새로운 값에 영향을 주므로, 이전에 도출됐던 값들을 계산에 반영해야만 정확한 결과를 얻을 수 있다. 순환 신경망은 이름에서 알 수 있듯이, 직전에 내보냈던 출력값이 은닉층의 계산 과정에 반영된다는 특징을 갖는다. 즉, t 시점의 결과값을 계산하는 데에는 t 시점의 입력값과 t-1 시점의 출력값이 모두 이용된다.

직전 시점의 정보를 바탕으로 결과를 예측하는 순환신경망이 자율주행 시스템에 사용되는 것은 어찌 보면 자연스럽다. 순환신경망은 차량과 보행자의 미래 위치를 예측하고 핸들을 얼마나 틀어야 할지, 비상 브레이크를 작동해야 할지 등을 결정하는 컨트롤 타워 역할을 수행한다.

a. 장단기메모리신경망(LSTM: Long Short Term Memory Network)

순환신경망(RNN)은 직전 정보를 다음 값에 반영할 수 있다는 장점이 있다. 그러나 바로 직전 값이 아니라, 훨씬 이전 단계의 정보가 필요한 경우에 기본적인 순환신경망 구조는 부적절하다. 순환신경망을 거치며 초기 값이 지나치게 작아지기 때문이다. 이를 기울기 소실, 또는 경사도 소실(vanishing gradient)이라고 부르는데, 이러한 문제를 해결하도록 설계된 것이 LSTM이다.

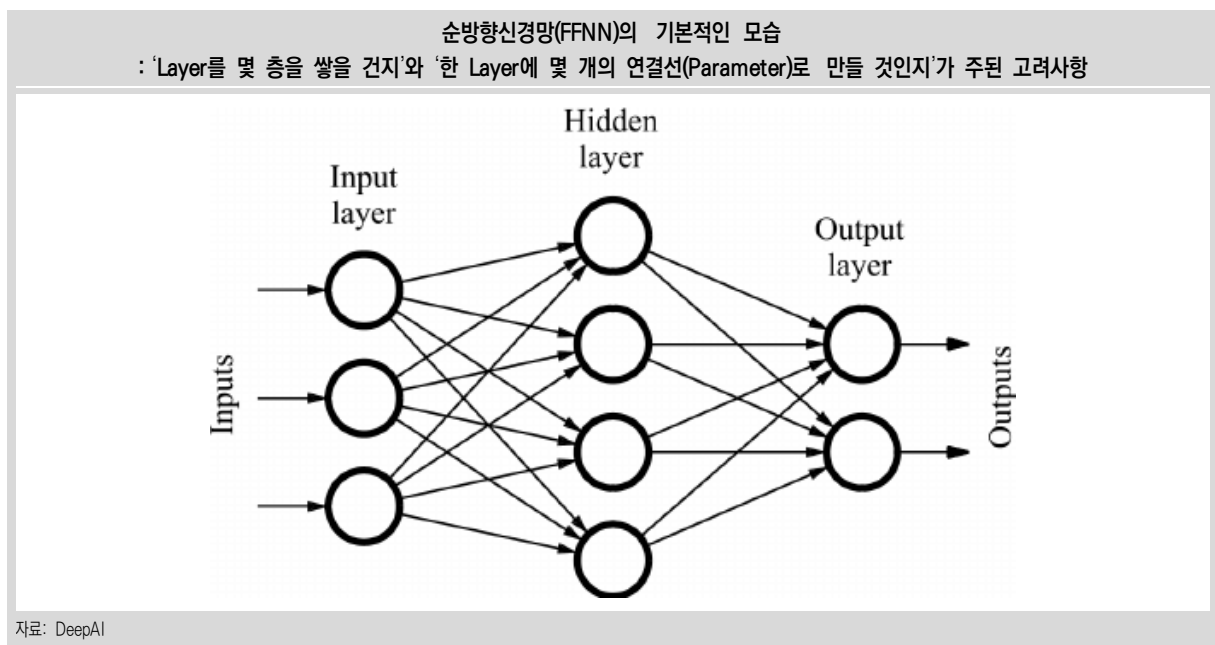
**경사도 소실(vanishing gradient) 문제는 초기 연속적인 시간 단계들에서의 정보들이 더 이후의 시간 단계들을 네트워크가 처리하게 되면서 점차 소실될 수 있는 것을 의미*

LSTM은 순방향 신경망이 아닌 순환 신경망의 일종이다. 출력값이 다시 신경망 안으로 입력되기 때문이다. LSTM은 네트워크에 "메모리 셀"을 추가하여 경사도 소실 문제를 해결한다. 이러한 메모리 셀은 과거의 중요한 정보를 받아 저장할 수 있다. 이후에는 현재 인풋에 따라 업데이트되거나 잊힐 수도 있다.

이를 통해 네트워크는 새로운 데이터를 처리하면서 과거의 중요한 정보도 유지하기 때문에, 선택적으로 정보를 기억하거나 잊어버릴 수 있는 능력을 갖게 된다. 또한 LSTM은 연속된 데이터가 네트워크로 들어 오고, 저장되고, 또 나가는 방식을 "게이트"를 활용해 제어한다. 일반적인 LSTM에는 세 가지의 게이트가 존재하는데 각각 forget gate, input gate, output gate다. 각 게이트들은 기억을 처리하기 위한 "자체 필터"로 생각하면 된다.

**각 단계마다 '가지고 있는 것 중 어떤 정보를 버릴지를 결정하고(forget gate layer), 미래의 새로운 정보 중 어느 것을 저장할지를 결정하고(input gate layer), 무엇을 output으로 내보낼지를 결정하는 것을 반복. 여기서 활성화 함수로는 시그모이드 함수와 하이퍼볼릭탄젠트(tanh)를 모두 사용. 정보를 얼마나 잊을지는 시그모이드로 처리하고, 결과값이 -1에서 1 사이의 값을 가질 수 있도록 조정하기 위해 tanh를 사용함. 1에 가까울수록 더 많이 기억한다는 개념.*

**순방향 신경망(FFNN: Feed Forward Neural Network)은 노드간 연결이 순환을 형성하지 않는 인공신경망. 인공신경망의 가장 단순하고 일반적인 신경망 형태.*



쉽게 생각하면 LSTM은 새로운 정보를 처리하는 동안 과거의 중요한 정보를 기억하는 데 도움이 되는 신경망의 메모리 유형이라고 볼 수 있다. 중요한 정보를 기록했다가 나중에 다시 참조할 수 있는 메모장이 있는 것과 같다. 이 메모리는 네트워크가 더 나은 예측을 하고 언어 번역이나 음성 인식과 같은 작업을 더 잘 수행하는 데 도움을 준다.

LSTM이 중요한 이유는 자율주행 분야에서 잘 활용될 수 있는 모델이기 때문이다. 구글 Waymo는 Open Dataset을 사용하여 LSTM 기반의 자율주행 모델(3D 물체의 탐지와 추적 특화)을 개발했다. 또한 테슬라는 2021년 AI Day에서 본인들의 자율주행 로드맵과 기술 현황을 공개했다.

테슬라의 자율주행 소프트웨어 FSD(Full Self Driving)가 사용하고 있는 인공지능 모델은 굉장히 많은데, 여기에는 Spatial RNN과 LSTM이 포함되어 있다. 내 차가 앞 차와 서로 마주보고 주행하는 상황에서 '앞 차가 나에게 길을 비켜주려고 하는지 아니면, 그냥 밀고 들어올지'와 같이 "미래를 예측"하려면 시간 개념이 추가된 연속적인 데이터가 필요하다.

그리고 그 데이터를 적절히 활용하기 위해서 중요 정보들이 기록돼야 하는데, 그게 바로 LSTM이 하는 일이다. 물론 이러한 예측을 위해서는 차량의 3차원 벡터값과 운동역학과 같은 추가 정보가 필요하다. 이 정보들을 종합하여 FSD 자율주행 소프트웨어는 차량 간 상호 작용을 처리하고 미래의 주행 상황을 예측하는 데 도움을 준다.

테슬라의 공간 순환신경망(Spatial RNN)과 LSTM 활용
: RNN, LSTM, Kinematics, 트랜스포머 기법까지 총동원해 자율주행 딥러닝 학습

Spatial RNN Video Module

kinematics

features

20x80x256
Ego Coordinate System

Spatial Feature Grid (h)
WxHxC

$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$

$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$

$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$

$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$

LSTM Example

자료: Tesla AI DAY 2021
주: W(Width), H(Height), C(Channels)으로 3차원의 벡터값을 의미/Kinematics는 비디오 내의 동작인식 능력을 위한 데이터 세트

이슈 중심에 있는 언어모델

1) 트랜스포머(Transformer)란?

GPT 같은 언어모델이 주목받는 이유는 좋은 성능 때문이다. 이런 언어모델을 다른 태스크에 전이학습을 하면, 어떤 태스크든 이전과 비교했을 때 성능이 비약적으로 상승한다. 트랜스포머는 현재 기준 가장 인상적인 AI 시스템 중 하나로 꼽힌다. 물론, 워낙 컴퓨터공학과 AI 모델의 진화속도는 빨라서 가까운 미래에 이 모델 또한 구식으로 취급될지도 모를 일이다. OpenAI의 수장 샘 알트만조차 우리가 5년 후에도 트랜스포머 모델을 계속 사용할지 의심스럽다고 말하기도 했다.

그럼에도 당분간 트랜스포머는 자연어처리 작업의 주요 모델의 하나로 자리잡고 있을 것이다. 우리가 인공지능에 관심을 갖게 한 ChatGPT의 핵심 구조도 트랜스포머 모델이다. 트랜스포머 모델은 번역, 텍스트 요약, 질문 응답 등 다양한 자연어 처리 영역에서 활용되기에 “파운데이션 모델”이라고도 불린다. 딥러닝 영역에서 트랜스포머는 점점 더 중요해지고 있고, 자연어처리를 넘어 컴퓨터비전으로까지 확장되고 있다.

GTC 행사에서 엔비디아가 내세운 “빅데이터와 딥러닝, 그리고 메타버스의 상호 연결”
: 엔비디아 CEO 젠슨 황이 꼭 짚어 강조한 딥러닝 알고리즘이 “트랜스포머”



자료: Nvidia GTC 2023

트랜스포머의 시작은 자연어 처리, 즉, 언어모델이다. 언어모델이란 단어 시퀀스(단어의 나열)에 확률을 부여하는 모델이다. 직전까지의 단어 시퀀스를 입력값으로 받아서, 다음 단어 시퀀스가 뭐가 나와야 하는지 확률을 계산한다. 그래서 가장 그럴 듯한 출력을 도출해내는 것이다.

예를 들어, 아이스 아메리카노라는 단어를 생각해보자. 아이스라는 단어가 나온 다음에 아메리카노가 나올 확률을 계산하는 것이다. 이러한 확률은 조건부 확률을 의미한다. 물론 아이스 뒤에 커피, 라떼 등이 나올지도 모르지만, 앞의 단어가 “아이스”가 나왔기 때문에 선택지가 줄어들게 된다. 아이스 애널리스트 같은 단어는 그럴 듯하지 않기 때문이다.

즉, 언어모델은 이전 단어들이 주어졌을 때 다음 단어가 나타날 확률을 부여하는 모델이라고 생각하면 된다. 그래서 GPT 언어모델은 문장 생성에 탁월한 성능을 보이는 것이다. 게다가 조건부 확률의 언어모델임에도 불구하고, 덧셈과 뺄셈 등 사칙연산에서도 100%에 가까운 정확도를 GPT-3 수준에서 선보인 것은 이미 유용함을 증명한 것이다.

트랜스포머 모델은 2017년 Google Brain 팀이 NIPS에서 발표한 "Attention Is All You Need"라는 논문에서 처음 소개되었다. 여기서 핵심은 트랜스포머가 "시퀀스 to 시퀀스"모델이라는 것이다. 시퀀스는, 단어 같은 것의 나열을 의미한다. 예를 들어 "나는 미래에셋증권의 디지털리서치팀 소속 애널리스트이다"라는 문장을, "나는, 미래에셋증권, 의, 디지털리서치팀, 소속, 애널리스트, 이다"라는 시퀀스로 나뉘볼 수 있다.

이 시퀀스를 기반으로 딥러닝 연산하고, 원하는 출력값을 또 다른 시퀀스 값으로 표현하는 게 트랜스포머다. 그래서 트랜스포머 모델을 다시 정리해보면, 단어 시퀀스(문장 속 단어와 같은 순차 데이터)를 가지고, 시퀀스 내 단어들의 관계를 추적해 맥락과 의미를 학습하는 신경망이다.

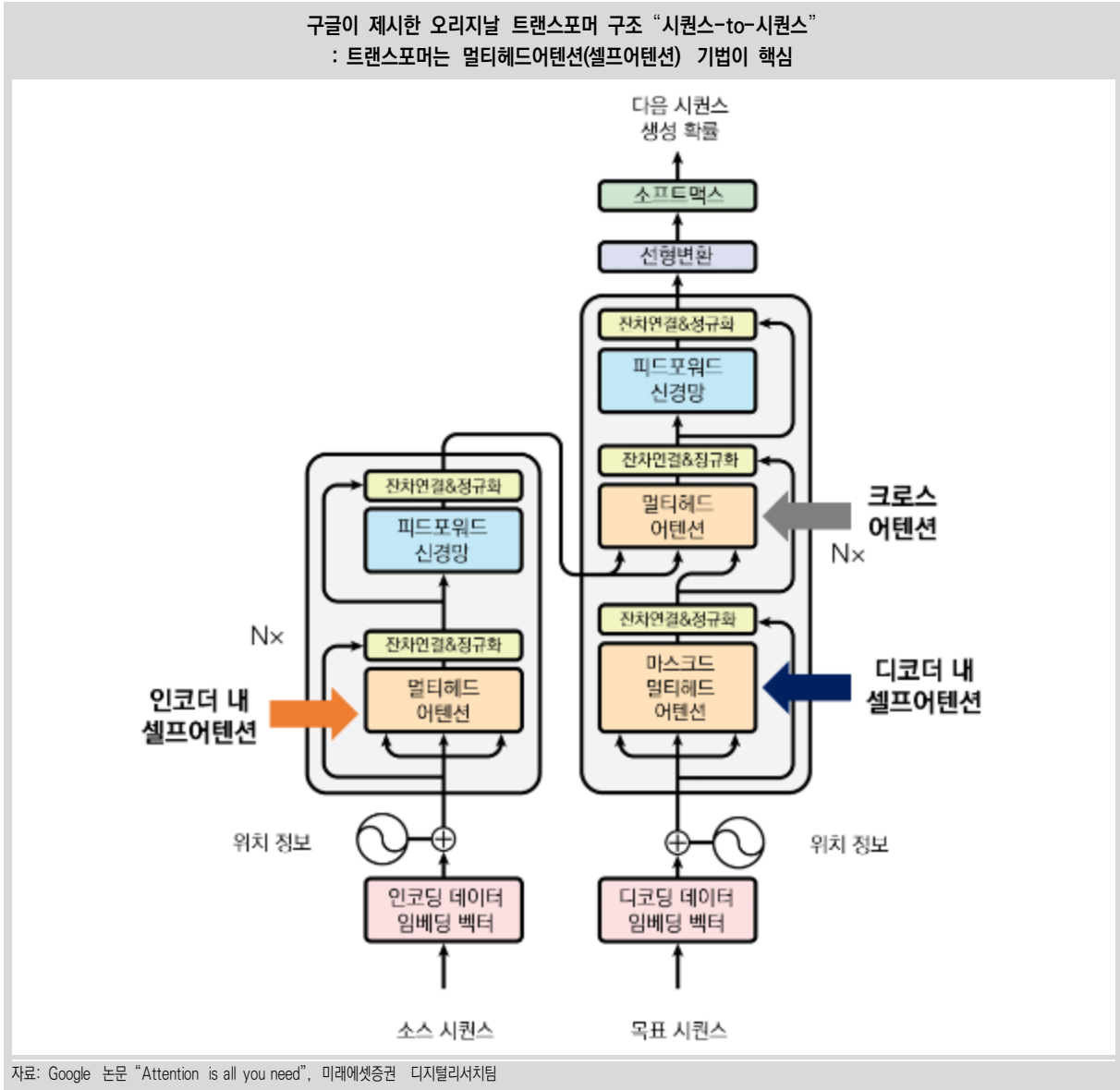
트랜스포머는 특히 CNN(합성곱신경망)이나 RNN(순환신경망)과 같은 다른 학습방법과 비교했을 때 우수한 성능을 보인다. CNN은 합성곱필터라는 장치를 이용해서 시퀀스의 '지역적인 특성'을 짚어내는 모델이다. 다음 단어 맞히기 과정에서, 특정 단어를 기준으로 주변의 단어들을 보고 문맥을 결정하는 식이다. 그런데 이 합성곱 필터 크기를 넘어서는 문맥은 읽어내기가 어렵다. 그리고 RNN은 시퀀스 길이가 길어질수록 오래 전에 입력된 단어는 잊어버리거나, 마지막 단어의 의미에 너무 많은 비중을 쏟는 문제가 생긴다는 단점이 있다.

반면, 트랜스포머는 '문장 전체' 안에서 단어들간의 '관계를 파악'하는 방식이다. 트랜스포머 모델은 기존의 CNN과 RNN을 사용하지 않고, "어텐션"이라는 메커니즘만을 사용한다. 이 어텐션이라는 기법으로 입력 문장과 출력 문장 사이의 관계를 파악한다. 어텐션 메커니즘은 신경망이 입력 데이터의 전체 또는 일부를 되짚어 살펴보면서 어떤 부분이 의사결정에 중요하지 보고, 중요한 부분에 집중하는 방식이다. Attention의 뜻이 주의와 집중이라는 점에 착안해 생각하면 된다.

어텐션이라는 것 때문에 트랜스포머는 서로 떨어져 있는 데이터 사이의 관계를 파악하는데 강점을 보인다. 게다가, 트랜스포머는 RNN처럼 입력 단어를 하나씩 처리하지 않고, 어텐션 기법을 활용해 병렬적으로 연산이 가능하다. 그렇다면 어텐션이 무엇인지 자세히 살펴보자.

트랜스포머가 쓰는 어텐션은 시퀀스 내 각 요소들(elements) 가운데 중요한 요소에만 집중하고 그렇지 않은 요소는 무시해서 딥러닝 성능을 끌어올린다. 그러니까 입력 데이터 시퀀스 가운데, 출력에 도움이 되는 데이터 시퀀스 요소 위주로 취사 선택하는 것이다.

그리고, 어텐션 중에서 셀프 어텐션이라는 개념은 반드시 알아두는 것이 좋다. 트랜스포머의 꽃이 바로 셀프 어텐션이라고 부를 정도니까 말이다. 셀프 어텐션이라는 것은 단어 그대로, '자신을 중심으로 주의를 기울이는' 기법이다.



*멀티헤드어텐션은 셀프어텐션을 동시에 여러 번 수행하는 것을 가리킴. 기본적으로 멀티헤드어텐션을 셀프어텐션으로 이해하면 됨.

셀프어텐션과 크로스어텐션(Cross-attention)은 트랜스포머 아키텍처에서 사용되는 어텐션 메커니즘. 이 두 메커니즘은 서로 다른 상황에서 사용되며, 다음과 같은 차이점이 있음.

셀프어텐션은 입력 시퀀스 내의 각 토큰이 '자신을 포함한' 다른 '모든 토큰과 상호작용'하며, 상호 간의 관계를 파악하는 메커니즘. 인코더(Encoder)와 디코더(Decoder) 내부의 각 층에서 사용되며, 입력 시퀀스 내의 문맥 정보를 파악하는 데 도움.

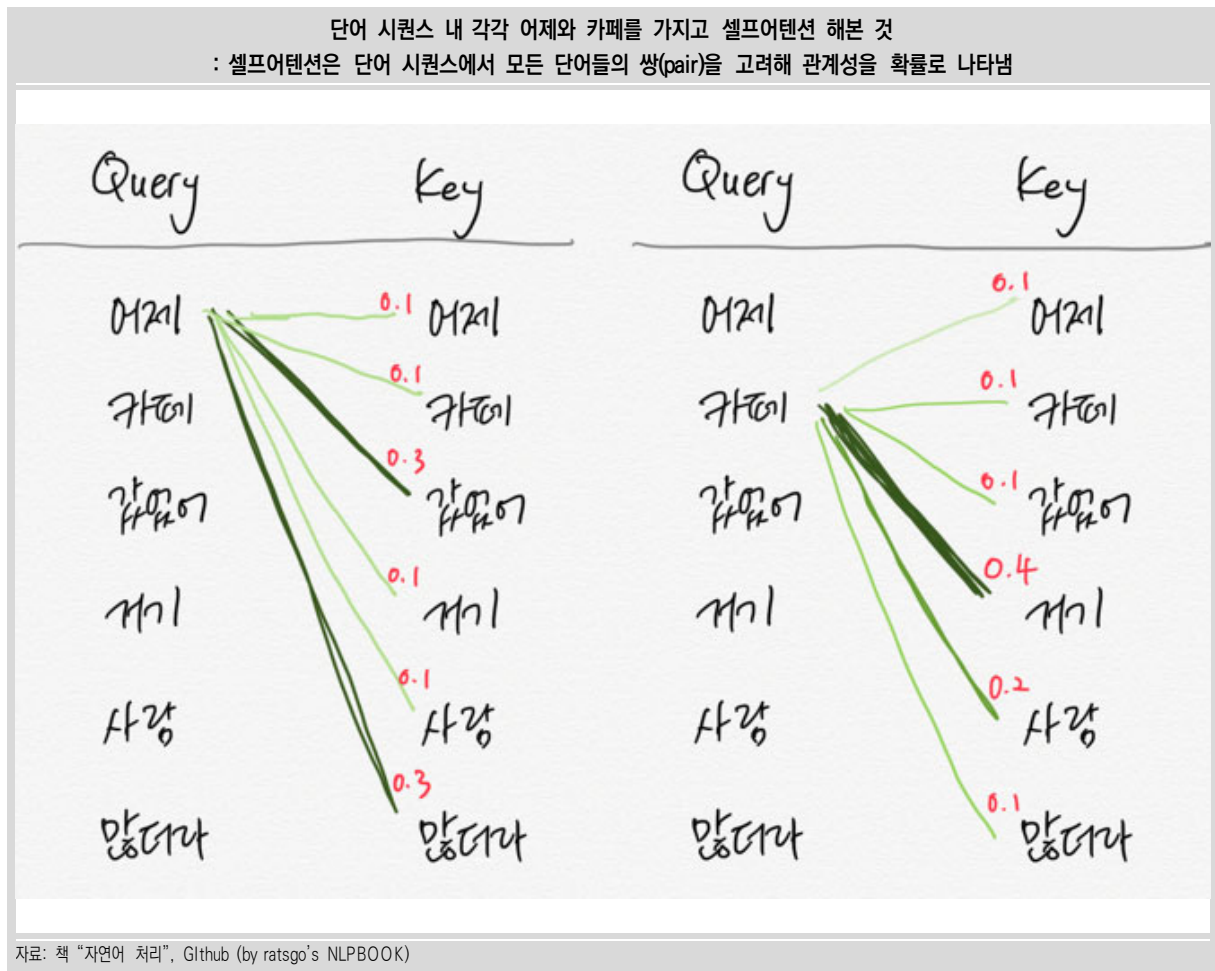
크로스 어텐션은 입력 시퀀스와 목표 시퀀스(예: 인코더에서 디코더로 전달된 시퀀스) 사이의 상호 작용을 파악하는 메커니즘. 트랜스포머의 디코더 층에서 사용. 인코더에서 전달된 정보에다가 디코더의 현재 상태를 조합하여 문맥 정보를 파악하고 예측을 수행. 즉, 인코더가 처리한 입력 시퀀스를, 디코더가 처리하고 있는 다른 시퀀스와의 관련성을 찾기 위해 사용. 이로써 더 나은 '출력을 생성할 수 있음.

요약하면, 셀프 어텐션은 동일한 시퀀스 안에서의 토큰 간의 관계를 파악하는 데 사용되며, 크로스 어텐션은 서로 다른 시퀀스(예: 인코더와 디코더의 시퀀스) 간의 관계를 파악하는 데 사용.

트랜스포머 모델도 우리가 이전까지 익히 보았던 것처럼 인코더와 디코더 구조로 이뤄져 있다. 정보의 압축 역할을 하는 인코더는 입력 문장을 벡터로 변환하고, 이를 푸는 디코더는 이 벡터를 바탕으로 출력 문장을 생성한다. 딥러닝 연산의 본질은 모델이 달라도 역시 비슷하다.

트랜스포머 모델은 입력문장을 인코더에 통과시켜 각 단어의 의미와 위치 정보를 포함하는 벡터값으로 변환한다. 그리고 데이터를 학습하는 과정에서 셀프어텐션 기법을 활용하게 된다. 셀프어텐션은 어떤 단어가 해당 문장이나 문맥 안의 다른 단어들 중 어떤 것과 가장 연관성이 높은지를 판단한다. 그리고 그 단어와의 관계에서 가중치를 높이는 방식으로 반영하는 것이다. 아래 그림은 각각 단어들을 기점으로 셀프어텐션을 해본 것이다.

예를 들어, “어제”라는 단어로 셀프어텐션을 해보면 그것과 가장 관련 깊은 다른 단어는 “갔었어”와 “많더라”가 된다. 한편, “카페”라는 단어로 셀프어텐션을 해보면 “거기”라는 대명사와 가장 깊은 연관성을 맺는다는 것을 알 수 있다. 이렇게 어울리는 pair끼리 높은 연관성을 갖도록 학습되는 것이 셀프어텐션 기법이다.



구글이 트랜스포머 논문을 내놓았을 때 설명한 문장 예시로 다시 한 번 살펴보자. "The animal didn't cross the street because it was too tired" 이 문장에서 "it"은 무엇을 가리키고 있을까? 거리(street)를 가리키고 있는건지 동물(animal)을 가리키고 있는건지가 문제다. 인간에게는 간단한 질문일지 모르지만, 기계에게는 그렇지 않다. 모델이 "it"이라는 단어를 처리할 때, 셀프어텐션은 "it"을 "animal"과 연관시킬 수 있도록 학습을 한다고 생각하면 된다.

가장 어울리는 단어를, 즉 확률이 가장 높은 시퀀스 요소를 어텐션 스코어가 가장 높다고 표현하기도 한다. 이를 통해 언어모델을 실제로 수행(inference)해볼 때 그럴 듯한 문장이 '생성'되는 것이다.

**어텐션 스코어는 디코더의 특정 학습단계의 임베딩 시퀀스 정보와, 인코더의 모든 학습 단계의 임베딩 시퀀스 정보사이의 유사도/연관성을 측정하여 계산.*

어텐션 스코어는 소프트맥스(softmax) 함수에 적용. 이 함수를 사용하는 이유는, 소프트맥스가 점수를 총합이 1인 양수의 확률로 변환하기 때문. 즉, 소프트맥스는 어텐션 스코어를 확률로 변환하는 데 사용. 그런 다음, 확률이 가장 높은 출력 토큰을 선택하는데 사용.

어텐션스코어를 통해 어텐션 가중치(query, key)를 구함. 이 가중치들을 인코더의 각 학습 단계마다의 밸류값(value)에 곱한 후 다 더함. 이러한 가중합으로 나온 값이 디코더의 다음 학습단계의 입력으로 사용 되는 것. 쿼리, 키, 밸류에 해당하는 벡터값을 찾는 게 학습의 목적. 다만, 이는 트랜스포머의 세부 구조를 알아야 하는 수학적인 것으로 해당 리포트에서는 다루지 않음.

그리고 셀프어텐션은 전체 입력이 동시에 딥러닝 네트워크에 공급되는 멀티헤드(Multi-head) 구조를 띠고 있다. 조금 전에 언급한 트랜스포머가 병렬 처리를 할 수 있는 이유가 이 때문이다. 여러 명에서 어떤 한 작업을 공동으로 하고 있다고 보면 된다. 따라서, 동일한 성능의 다른 모델보다 빠른 속도로 처리할 수 있다.

따라서, 이를 정리해 보면 다음과 같다. 단어 시퀀스들을 병렬로 다 훑어본다. 이런 각각의 관계를 파악하기 위한 작업에 셀프어텐션이 채택된다. 이 정보를 바탕으로 문장의 의미(context)를 파악할 수 있고, 이를 확률로 나타낸다. 확률이 가장 높은 요소대로 단어가 생성되고, 점차 전체 시퀀스가 완성된다.

2) 트랜스포머가 낳은 GPT와 BERT

엄청난 속도의 생성형 AI의 발전 속도에 기죽지 않기 위해선 기준점을 세우고 공부해야 한다. 생성형 AI에서의 기술 발전을 이끄는 주체는 트랜스포머(transformer)와 전이학습(transfer)이라는 딥러닝 네트워크다. 그리고 이 둘의 결합으로 탄생한 모델이 바로 GPT다.

GPT는 기본적으로 언어모델이다. 그래서 인터넷 속 수많은 말뭉치를 학습해 자연어의 풍부한 맥락을 학습했다. 언어모델을 미세조정해서 하위 특정 업무(downstream task)에 활발히 활용하고 있다. 가장 좋은 예시가 문장 요약, 기계번역, 감정분석 등이다. ChatGPT가 가장 뛰어난 수행력을 보이는 분야이기도 하다.

이 언어모델의 탄생 덕분에 자연어 처리 수준이 이전보다 크게 향상됐다. 언어모델은 국가나 기업 경쟁력의 핵심 요소가 될 것으로 보일 정도로 큰 의미를 가지게 됐다. 사람 말을 귀신같이 알아듣고 뚝뚝하게 답하는 AI의 경제적/사회적 가치는 헤아리기 어려울 정도로 크기 때문이다.

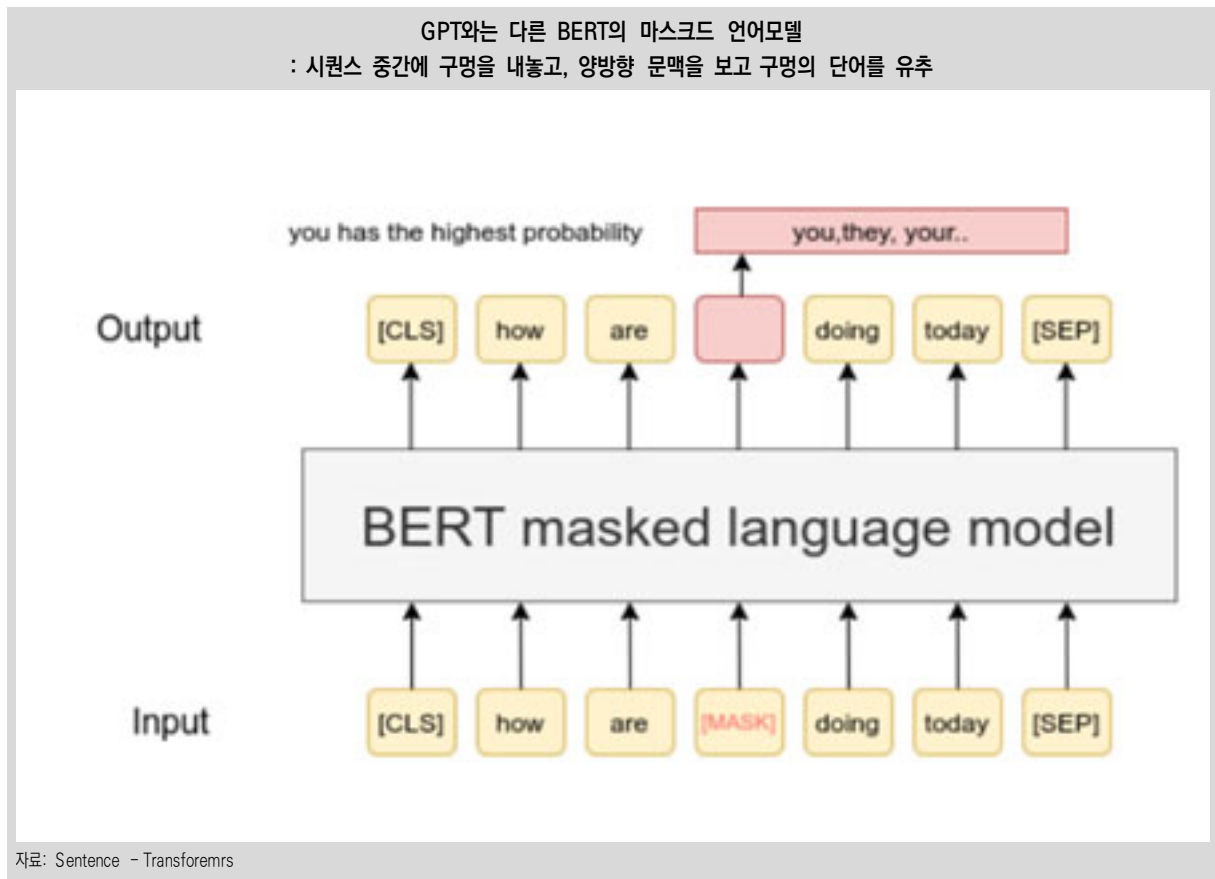
그런데 GPT의 가장 큰 경쟁자로 꼽히는 것은 바로 구글이 밀고 있는 BERT다. 구글이 코드레드까지 발동하면서 Bard를 내놓았듯이, 앞으로 GPT 기반 ChatGPT와 BERT 기반 바드(Bard)의 경쟁이, 딥러닝 영역에 있어서 가장 중요한 더비매치(Derby match)가 되지 않을까 생각한다.

두 모델 모두 NLP(자연어 처리 인공지능)에서 널리 사용되고, 트랜스포머를 기반으로 한다. 그리고 두 모델 모두 질의응답 / 텍스트 분류 / 텍스트 생성과 같은 다양한 downstream 작업을 위해서, 모델을 전체적으로 업데이트할 수 있는 미세조정(fine tuning) 기능을 제공한다.

하지만, 이름에서부터 두 모델은 근본적인 차이가 있다. 먼저 BERT는 Bidirectional Encoder Representations from Transformers의 줄임말이다. 즉, BERT는 양방향(Bidirectional) 모델이다.

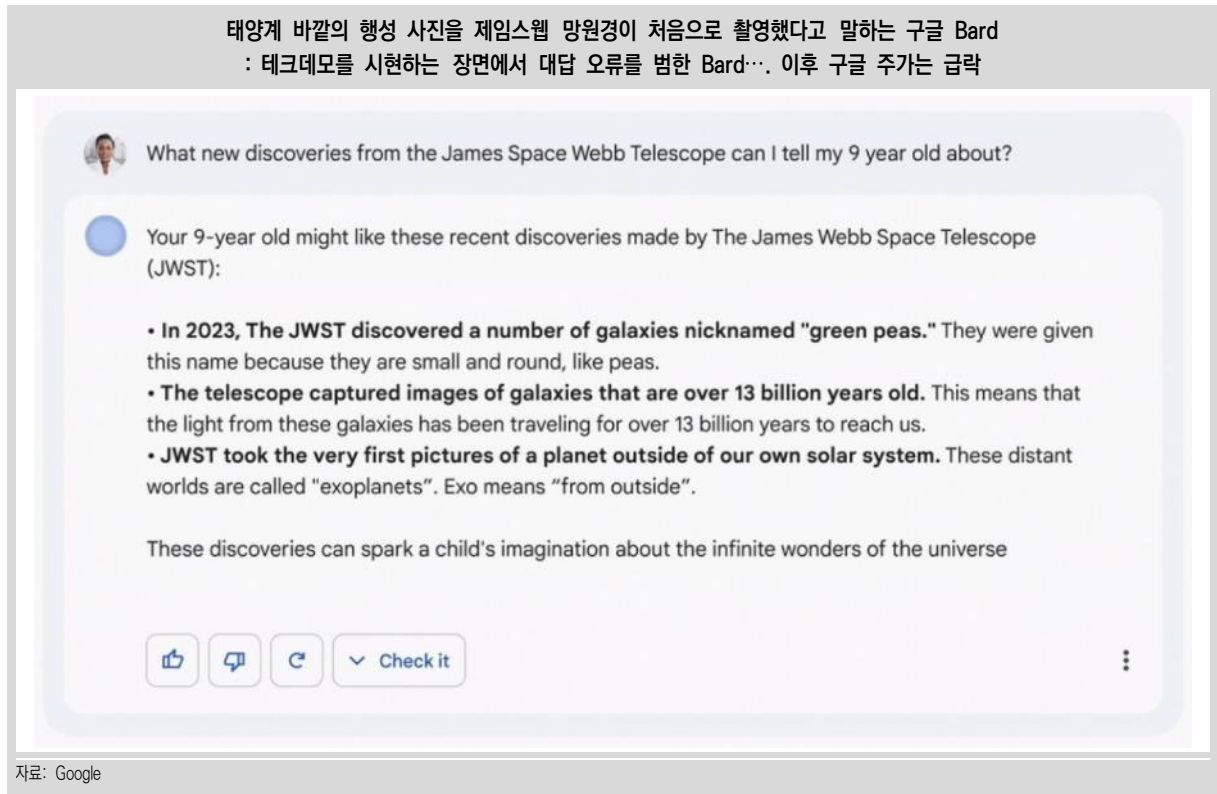
BERT는 그래서 그냥 언어모델이 아니라 MLM(마스크드 언어모델)이라는 작업에 대해 주로 사전학습을 진행한다. 즉, 시퀀스 요소들을 예측할 때 각 단어의 왼쪽 및 오른쪽 컨텍스트를 모두 고려한다. GPT와는 달리 마스크된 부분을 양방향에서 살펴보기 때문에, 문장에서 서로 다른 단어 간의 관계를 더 잘 포착할 수 있다는 근본적인 장점을 가진다.

**MLM(Masked Language Model)은 입력 시퀀스의 특정 단어를 무작위로 마스크하고 해당 언어모델은 앞 뒤 문맥을 정보로 사용해 누락된 단어를 예측해야 함.*



물론, 단점도 있다. BERT는 일반적인 자연어 처리 모델에서 잘 작동하지만, 과학, 바이오, 금융 등 특정 분야의 언어모델에 사용하려면 잘 적용되지 않는다. 최근에 바드의 클로즈드 베타를 이용하고 있는 사람들의 반응도 바로 이 측면에서 후한 평가가 나오지 않고 있다.

언어모델의 큰 속제는 특정 도메인 영역에서도 활용할 수 있을 만큼의 일반화가 가능하냐는 것인데, BERT 기반 언어모델이 앞으로 어떻게 해법을 찾아나갈지가 관건이다.



반면, GPT는 Generative Pre-trained Transformer의 줄임말이다. BERT와 가장 큰 차이점은 “양방향”이 아니라 “순방향” 문장 생성 모델이라는 점이다. 즉, 각 단어의 왼쪽 컨텍스트만 고려하는 단방향(왼쪽에서 오른쪽) 모델이다.

BERT가 마스크드 언어모델로 사전학습하는 것과 달리 GPT는 언어모델(Language model)에 대해 사전 학습을 진행한다. 이전 단어가 주어진 시퀀스에서 다음 단어를 예측하도록 훈련되는 것이다. 훈련 과정은 입력 시퀀스의 시작부터 끝까지 반복되며, 주어진 문맥에서 가장 확률이 높은 토큰을 선택하여 생성한다. “나는”라는 단어 직후에 “이다”라는 단어가 바로 나오지 않게 하는 것처럼 말이다.

이런 특징은 GPT로 하여금 주어진 프롬프트를 기반으로 텍스트를 생성하는 데 더 적합하게 만들어준다. 우선 GPT는 단방향 모델이기 때문에 상대적으로 적은 컴퓨팅 자원으로도 높은 성능을 달성할 수 있다. 과도한 컴퓨팅 자원의 투입 없이도 일정 수준의 품질을 확보할 수 있다는 것이 큰 장점이다.

다만 토큰을 순방향으로 생성하기 때문에, 이전 토큰을 생성한 후에야 다음 토큰을 생성할 수 있다는 단점도 있다. 따라서, 텍스트 생성 속도가 상대적으로 느릴 수는 있다. 물론, 딥러닝 훈련 과정에서는 어텐션 기법으로 병렬 처리를 활용할 수 있지만, 추론 과정에서는 순차적 처리로 인해 속도가 느려질 수 있다는 말이다.

또한 GPT는 단방향성 때문에 문맥 파악에 BERT만큼은 강하지 않다. 따라서, 양방향 트랜스포머 인코더 모델인 BERT와 같은 양방향 문맥 파악 능력은 가지지 못한다. 하지만, GPT가 BERT에 비해 가진 단점에도 불구하고, 최근 GPT-4가 보이는 성능은 실로 놀라울 정도다.

정리하자면 대형언어모델(Large Language Model)이 중요한 이유는, 생성형 AI의 기반으로 모든 영역에서 활용될 수 있기 때문이다. GPT는 그림 및 영상, 코드, 문서, 음악 및 음성 파일 모두에 적용할 수 있다. 나열한 모든 파일들은 모두 숫자로 변환할 수 있기 때문이다. 가장 까다로워 보이는 영상도 결국 프레임의 집합체다. 텍스트와 첫번째 프레임을 넣어서 두번째 프레임을 만들면 되는 식이다. 그리고 프레임은 그림이고, 그림은 숫자의 집합이다.

모든 것은 벡터로 뭉치고 벡터로 흩어진다. 인간은 이해를 못하지만 기계는 한다. 원리는 모르지만, 우리에게 가져다줄 이익이 너무나 많고 커다랗다. 앞으로 모든 산업에는 딥러닝이 필수 인프라가 될 것이다.

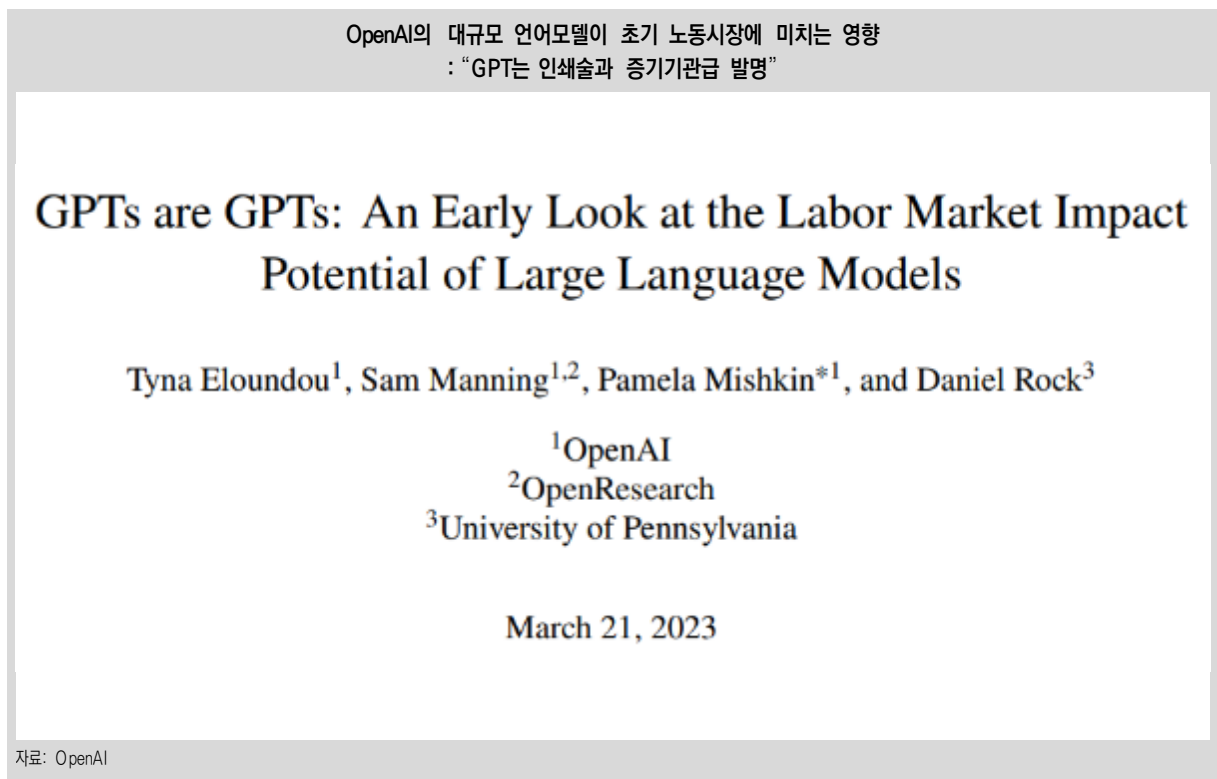
다음 장에서는, ChatGPT를 포함한 생성형 AI의 대대적인 출현으로 어떤 변화가 생기는지 초점을 맞춘다. 생성형 AI가 획기적으로 개선시킬 수 있는 산업, 또 이미 그러고 있는 분야를 다룰 예정이다.

III. AI가 만들 변화는 “Everything, Everywhere”

생성형 AI의 출현은 공포스러운 경이로움

1) OpenAI가 내놓은 충격적인 보고서

이번 장에선 ChatGPT로 불리우는 생성형 AI의 출현으로 인해, 큰 변화를 앞두고 있는 개인과 산업에 대해서 다룰 예정이다. 생성형 AI ChatGPT의 출현은 세상에 엄청난 “딥 임팩트(deep impact)”를 주었다. 서비스의 확장도 매우 빠르게 일어나고 있으며 향후 어떻게 전개될지 가늠조차 쉽지 않다. 특히 고용시장과 산업계 입장에선 엄청난 기회 요인이기도 하지만 동시에 큰 위협으로 다가올 수 있다.

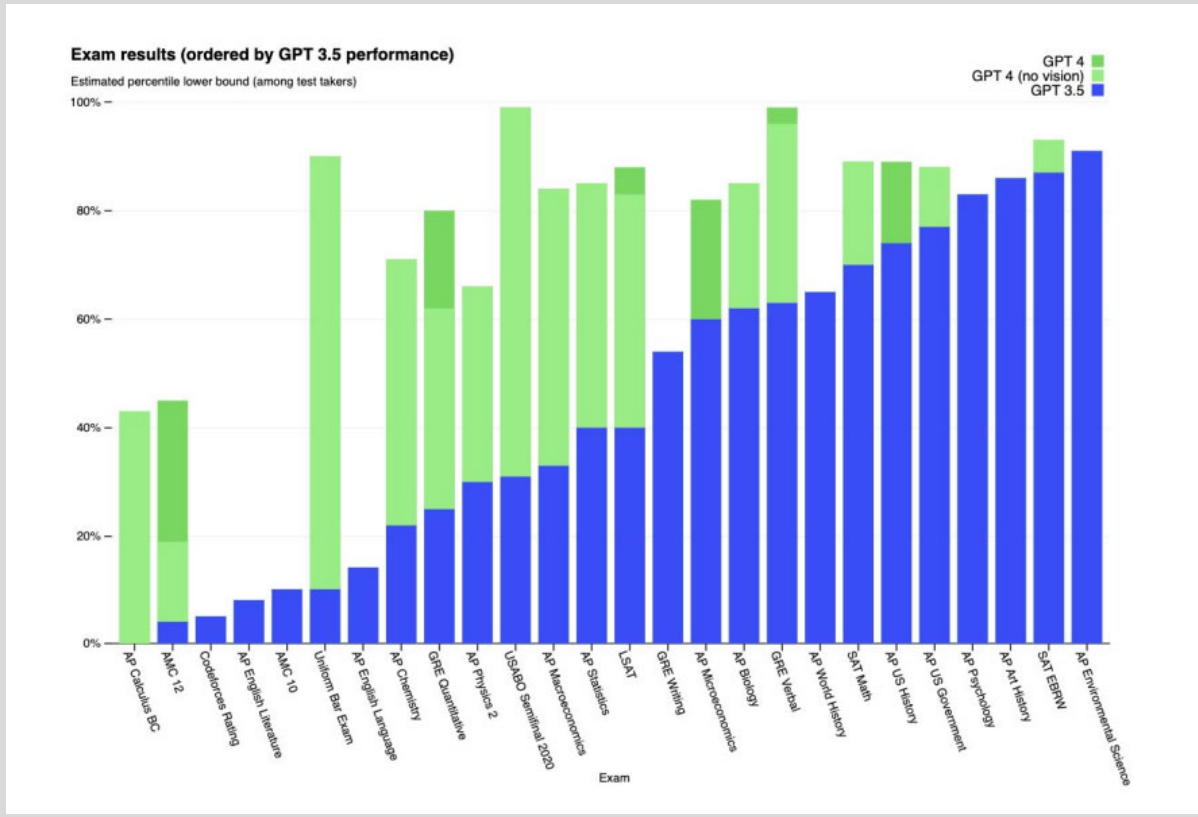


최근 OpenAI는 GPT-4를 내놓으며 새로운 논문을 발표했다. 논문의 제목은 “GPT(Generative Pre-trained Transformer)s are GPT(General-Purpose Technology)s”로써, 직역하면 GPT는 “범용기술이다” 라는 뜻이다. 저자들은 범용목적 기술(General-Purpose Technology)의 조건으로 아래 3가지를 꼽았다.

- (1) 사회 전체로 확산
- (2) 기술의 점진적 개선
- (3) 중요한 공동 발명품 개발 및 파급 효과

위 3가지 조건을 감안해, 저자들은 GPT의 발전은 인쇄술과 증기기관 발명과 나란히 놓을 수 있을 정도라고 강조했다.

GPT-3.5에 대비해서 대폭 성능이 개선된 GPT-4
 : GPT-4는 멀티모달. 텍스트에 더해 이미지 데이터까지 학습하면 성능이 더 우월해짐



자료: OpenAI

우리는 대규모 언어모델이 사회 전체로 확산된다는 것과 지속적으로 개선이 될 것이란 전제조건에 주목했다. 언급한 것처럼 실제로 사회전체에 확산되고 성능이 개선된다면 노동 시장에 미칠 파급력이 엄청나게 강력할 것으로 예측되기 때문이다. 하지만 이미 전제조건 중 하나인 성능 개선측면에서 GPT-3.5에서 GPT-4로 갈수록 향상되는 결과만 보더라도 향후 미래에 대한 의심의 여지는 없어 보인다.

또한 논문의 주요 내용을 살펴보면 사회 전체로 확산 가능성도 쉽게 유추할 수 있다. GPT-4가 직무별로 미치는 영향을 2,087개의 업무(컴퓨터 시스템 감독, 질병의 진료, 영업 및 거래 등)를 대상으로 조사했다. 대체가 가능한 업무 중 GPT-4가 사람과 유사한 수준으로 작업을 완수하는데 걸리는 시간을 조사했다. 소요시간이 절반 이상 감소한 직업(occupation)과 작업(task)이 약 14~46%에 달했다.

GPT의 사용으로 업무가 2배 이상 빨리지는 업무의 수: 각각 직업측면과 작업측면에서 GPT가 월등

직업측면 노출도					작업측면 노출도				
	인간이 평가		GPT-4가 평가			인간이 평가		GPT-4가 평가	
	mean	std	mean	std		mean	std	mean	std
α	0.14	0.14	0.14	0.16	α	0.15	0.36	0.14	0.35
β	0.30	0.21	0.34	0.22	β	0.31	0.37	0.35	0.35
ζ	0.46	0.30	0.55	0.34	ζ	0.47	0.50	0.56	0.50

자료: OpenAI, 미래에셋증권 디지털리서치팀

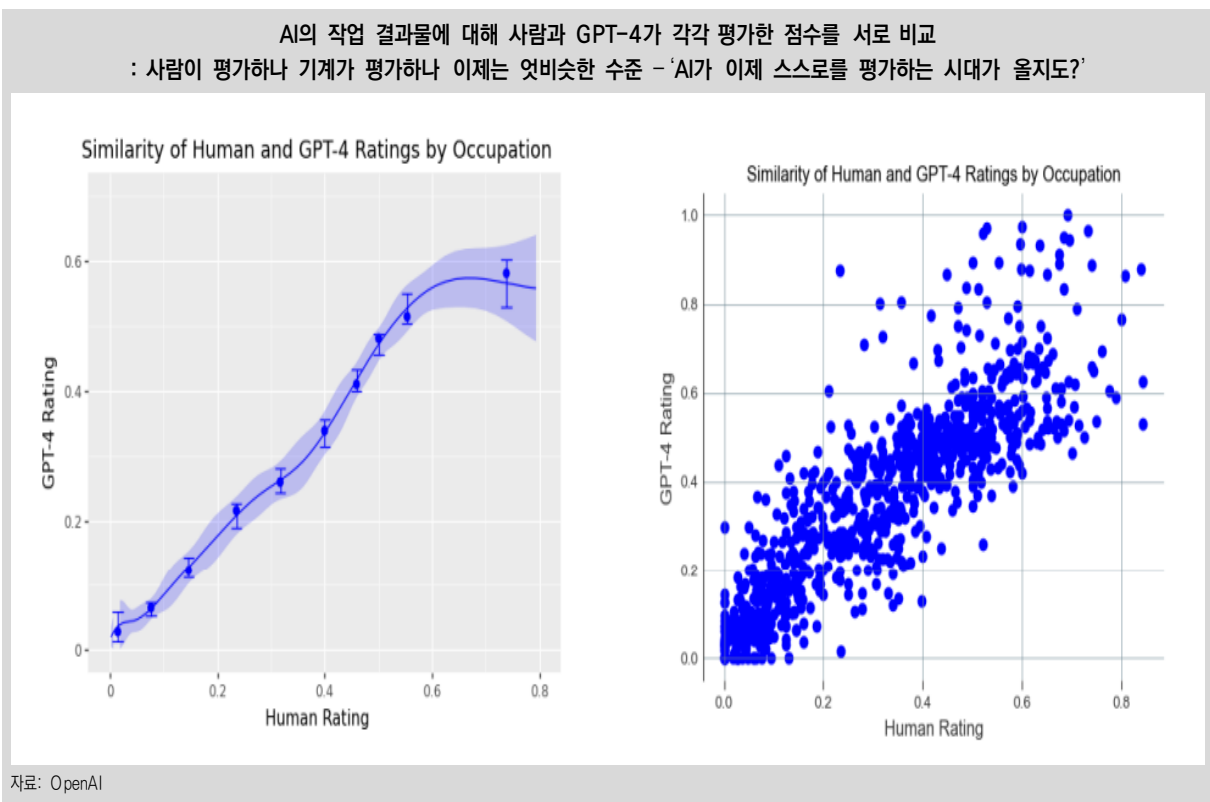
* $mean = (GPT \text{ 사용으로 } 2\text{배 이상 빨라지는 업무 숫자} / \text{전체 업무 숫자})$

* α : GPT-4만 단독으로 사용하는 경우를 가정

* ζ : GPT-4와 디지털 도구(검색 엔진, API 등)를 연결해서 사용하는 경우를 가정. Microsoft가 ChatGPT를 Bing, 워드, 엑셀에 결합한 형태를 의미

* β : α 와 ζ 의 중간 정도를 가정.

게다가 GPT-4를 이용한 업무 결과물의 수준을 사람과 GPT-4가 각각 평가하도록 진행했는데, 사람과 GPT-4의 평가점수가 유사하다는 결론도 도출됐다.



그리고 우리가 기존에 익히 알고 있던 기존 상식(AI가 고소득 노동자보다 저소득노동자에게 더 큰 영향을 미칠 것)과 다른 결과가 도출된 것도 주목할 만한 포인트다. 아래의 표에서 베타(영향도) 값을 주목하자.

위(저소득 노동자)에서 아래(고소득 노동자)로 내려갈수록 증가하는 경향을 보인다. 이러한 결과는 잠재적으로 고소득 노동자가 저소득 노동자에 비해 더 크게 대규모 언어모델에 노출될 수 있다는 이야기다. 지식기반 사회에서 전문성이란 대체로 정보를 다루는 능력인데, 이것은 바로 대규모 언어모델이 특화된 영역이기 때문인 것으로 추정된다. 인간의 예상이 보기 좋게 엇나간 것이다.

교육수준과 소득구간으로 직업군을 나눈 뒤, 각각 GPT로 영향받게 될 '노출도'를 측정
: 고소득자 관련 업무가 제일 대체될 가능성 높음

Job Zone	Preparation Required	Education Required	Example Occupations	Median Income	Tot Emp (000s)	H α	M α	H β	M β	H ζ	M ζ
1	None or little (0-3 months)	High school diploma or GED (optional)	Food preparation workers, dishwashers, floor sanders	\$30,230	13,100	0.03	0.04	0.06	0.06	0.09	0.08
2	Some (3-12 months)	High school diploma	Orderlies, customer service representatives, tellers	\$38,215	73,962	0.07	0.12	0.16	0.20	0.24	0.27
3	Medium (1-2 years)	Vocational school, on-the-job training, or associate's degree	Electricians, barbers, medical assistants	\$54,815	37,881	0.11	0.14	0.26	0.32	0.41	0.51
4	Considerable (2-4 years)	Bachelor's degree	Database administrators, graphic designers, cost estimators	\$77,345	56,833	0.23	0.18	0.47	0.51	0.71	0.85
5	Extensive (4+ years)	Master's degree or higher	Pharmacists, lawyers, astronomers	\$81,980	21,221	0.23	0.13	0.43	0.45	0.63	0.76

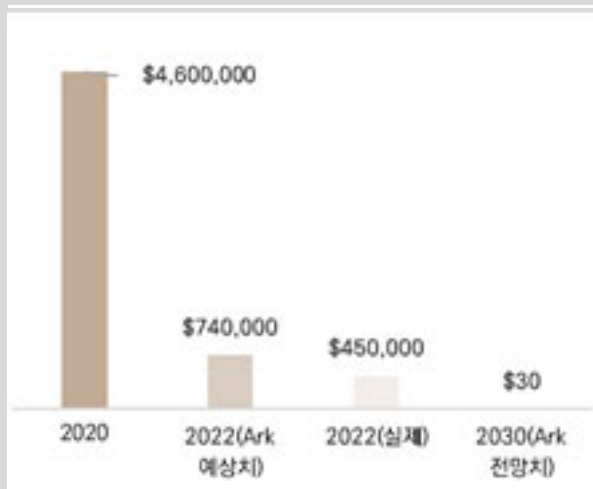
Table 6: Mean exposure to GPTs by job zone. For each job zone, we also present the median of median annual income for each constituting occupation in USD, and the total number of workers in all occupations for that job zone, in the thousands.

자료: OpenAI, 미래에셋증권 디지털리서치팀

물론 논문에서도 데이터와 방법론에 대한 한계점에 대해서 명확히 명시했기 때문에 관련된 내용에 대한 선부른 일반화는 주의할 필요가 있다. 하지만 분명한 것은 ChatGPT의 출현은 사소한 이벤트로 치부하기엔 이미 너무 많은 가능성을 보여줘버렸다.

게다가 컴퓨팅 프로세서의 발달로 앞으로 자연어 처리 프로그램의 개발 속도가 더욱 가속화될 것이라는 점도 고려해야 한다. MIT 대학교의 Lex Fridman 박사는 GPT의 학습 비용이 2020년에는 26.6억 달러였지만, 2024년에는 3.2억달러, 2032년에는 500만 달러, 2040년에는 8만 달러로 시간이 지날수록 급속히 줄어들 것으로 전망하기도 했다. 또한 ARK 인베스트에서는 AI 트레이닝 비용이 매년 70% 가량 감소하고 있고, 이는 작년 예상치 60% 감소보다 더 빠른 속도라고 놀라워했다. 이처럼 생성형 AI를 훈련하고 또 추론하는 비용이 점차 줄어들게 되면 산업 전반에서 AI를 받아들이는 속도가 가속화될 것이다.

GPT-3 수준의 AI 훈련에 드는 비용



자료: ARK Invest Big Ideas 2023, 미래에셋증권 디지털리서치팀

AI 챗봇이 추론(inference) 백만 건을 수행했을 때의 비용



자료: ARK Invest Big Ideas 2023, 미래에셋증권 디지털리서치팀

다음 장에서는, AI로 촉발된 “산업화 시대”와 그 안에서의 위기와 또 기회를 중점적으로 다룰 예정이다.

2) AI 혁명에 쏠리는 관심과 자금, 분주해진 기업들

딥러닝의 아버지라고 불리는 Geoffrey Hinton은 GPT-4의 출현을 보고 이렇게 말했다. “애벌레는 영양분을 추출해 나비로 변태한다. 사람들은 수십억개의 ‘이해 정보’를 추출했으며, GPT-4는 인류의 나비다.” 딥러닝 기술에 누구보다 공헌해 온 그가 ChatGPT를 보고 난 뒤의 충격은 이만큼이나 컸던 것이다.



전문가들 의견에 따르면, 점점 더 전문지식에 깊이 들어갈수록 GPT-4의 성능이 대단하다고 한다. 추론 능력이 대폭 향상되었기 때문이다. 압도적인 성능의 GPT-4는 GPT-3.5보다 10배 이상 많아진 1조~10조개로 추정되는 파라미터 숫자로서 타의 추종을 불허하는 대형 언어모델로 완전히 자리매김했다. GPT-3.5를 기반으로 한 서비스조차 사람들의 입이 떡 벌어지게 만들었는데 몇 개월만에 업그레이드가 된다는 것이 두렵게 느껴질 정도다.

인공지능 전쟁의 최선봉에 있는 ChatGPT가 불러온 “생성형 AI 모먼트(moment)”는 그야말로 혁명이다. 당분간 반짝 하는 테마로 끝이 날 것 같지 않는, 꺼지지 않는 불꽃이 될 것이다. 이 같은 변화는 산업단에서는 더욱 극명하게 나타날 것이다. 그렇게 생각하게 하는 이유는, 앞으로 GPT-4의 API를 기반으로 하는 어플리케이션 서비스가 우주가 처음 만들어질 때의 빅뱅처럼 무수하게 쏟아져 나올 것으로 전망되기 때문이다.

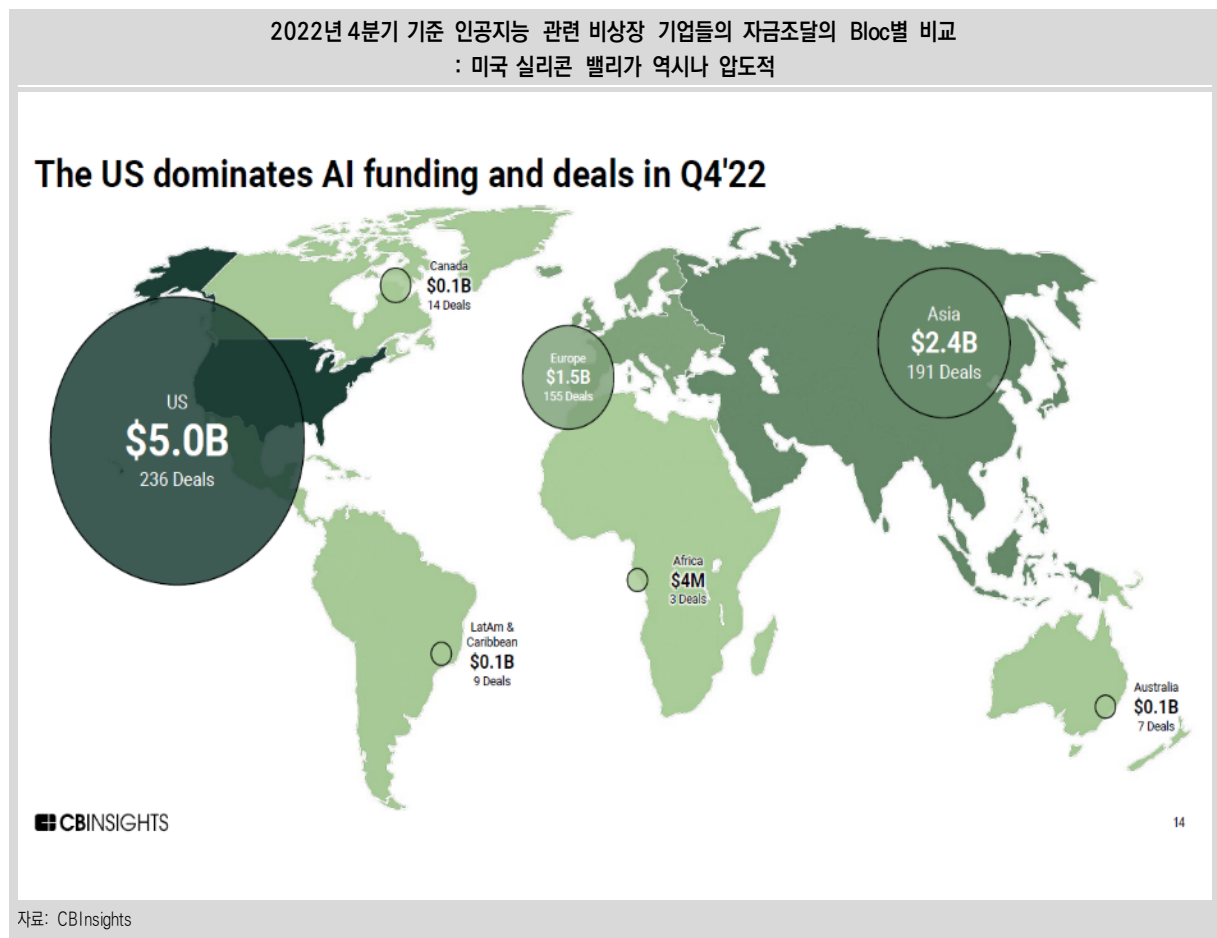
실제로 최근 몇 달 동안 인공지능과 관련한 서비스에 대한 언급이 소위 존재감이 큰 인플루언서들의 입을 통해 엄청나게 오르내리고 있으며, 한국에서도 하루가 멀다 하고, 스타트업과 빅테크들이 본인들의 인공지능 역량을 뽐내면서 우리 것도 한 번 써보라고 사용자들에게 권하고 있다.

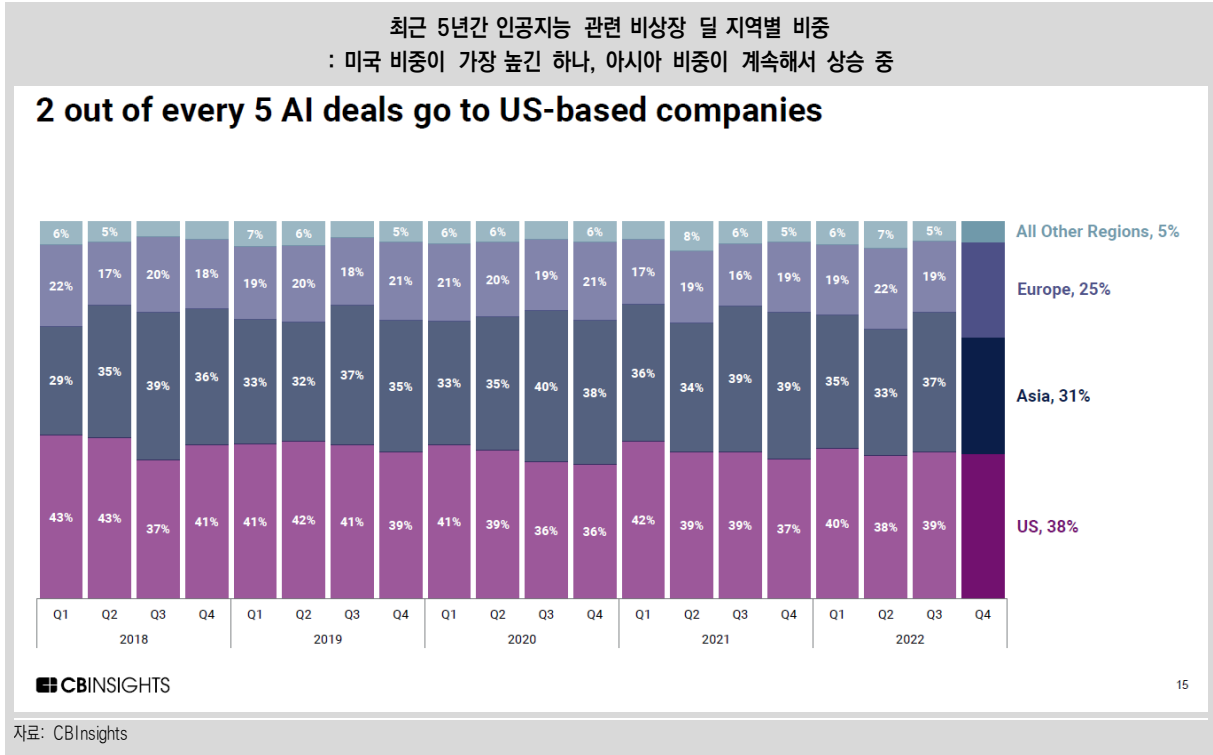
이런 인공지능 모먼트를 가장 빠르고 잘 확인할 수 있는 분야가 바로 미국 벤처캐피탈(VC) 시장이다. 다음은 작년 기준 생성형 AI에 있어 가장 많은 금액의 자금조달에 성공했던 기업들 리스트다.

- **Anthropic:** 인공지능 모델 개발업체 (Series B 5.8억 달러 조달)
- **Inflection AI:** 자연어 처리 인터페이스 (Series A 2.25억 달러 조달)
- **Cohere:** 자연어 처리 툴킷 개발업체 (Series B 1.25억 달러 조달)
- **Jasper:** AI 기반 콘텐츠 제작 업체 (Series A 1.25억 달러 조달)

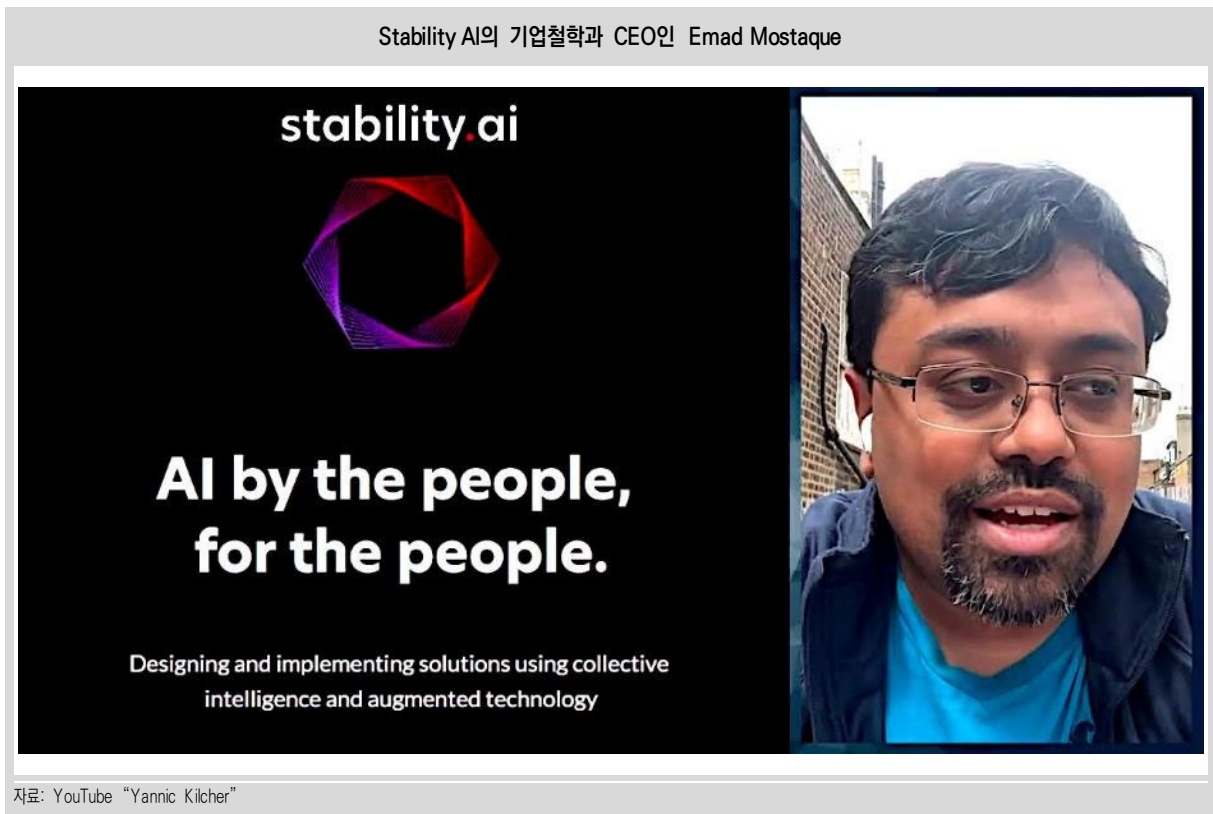
올해 1월에는 OpenAI가 마이크로소프트로부터 100억 달러를 조달했으며, 최대 290억 달러 정도의 가치로 추정되는 밸류에이션에 투자를 받았다. 이는 OpenAI의 2023년 예상 매출액인 2억 달러를 감안하면 100배 이상의 프리미엄이다.

이런 상황이라면 AI 관련 기업들의 몸값은 작년보다 올해 훨씬 더 올라가게 될 가능성이 크다.





사실 지난 1년간 기록적인 금리인상과 더불어 최근 실리콘밸리은행(SVB)이 파산하며 미국의 VC시장은 흑한기를 맞고 있다. 하지만 이런 투자 흑한기에도, 생성형 AI 관련한 업체로서 1년 만에 몸값이 4배가 된 기업이 있다. 스테이블 디퓨전을 개발한 영국 런던의 스타트업 Stability AI다.



Stability AI는 2022년 10월에 한 번에 1억 달러의 자금을 Lightspeed 벤처 등으로부터 조달 받았으며 밸류에이션은 10억달러까지 높아졌다. 이로써 단숨에 유니콘 기업이 된 Stability AI는 내친김에 기업가치를 40억 달러로 높여 새로운 투자 펀딩을 진행 중이다. 넉 달만에 기업가치가 4배나 된 것이다. 물론 최종결정은 아직 내려지지 않아 평가액도 변할 수는 있다. 그럼에도 VC 투자 흑한기에 이렇게 높은 평가액으로 펀딩을 진행하고 있다는 것은 인공지능이 가장 핫 한 화두라는 증명이다. 실리콘밸리 VC들은 2022년 한 해 투자를 집중한 메타버스, 웹3 분야 일변도에서 벗어나 생성형 AI 투자로 눈을 돌리고 있다.

VC나 빅테크들만 AI에 투자를 하는 것은 아니다. 고객관리 소프트웨어 솔루션의 최강자인 세일즈포스는 생성형 AI 스타트업 투자를 위해 2억 5000만달러 규모의 펀드를 조성했다고 밝혔다. 세일즈포스는 이 펀드를 통해 구글이 앞서 투자한 생성형 AI 스타트업 앤스로픽(Anthropic)과 코히어(Corhere), 유닷컴에 투자할 계획이다. 세일즈포스같은 전통적인 소프트웨어 업체들이 이런 생성형 AI 투자에 나선 것은 재무적 투자 개념이 아닐 것이다. 마이크로소프트와 구글이 “코파일럿”과 “워크스페이스”와 같은 생성형 AI 플러그인 등으로 기업용 생산성 향상 툴을 들고 나오자 이에 대항하기 위함이라는 판단이다.



특히 마이크로소프트의 Copilot 데모 영상을 본 세일즈포스는 간담이 서늘했을 것이다. 그 영상에는 고객 재고 관리 예측 시뮬레이션 및 최적화를 AI 챗봇에게 몇 마디만 적으면 완수가 되었기 때문이다. 그야말로 전통 SaaS 기업들에게는 실존적 문제다.

생성형 AI는 포토샵과 일러스트레이터를 판매하는 어도비(Adobe)의 입지도 위협하고 있다. 생성형 AI가 가장 잘 하는 게 이미지 생성과 잠재공간 편집을 통한 손쉬운 이미지 편집이다. 이는 관련 서비스를 제공하는 어도비의 생존을 위협한다. 쉽게 생각해서, 내가 파는 물건을 더 빠르고 싸게 파는 사람이 등장한 것이다. 또한 라이선스 형태의 구독 서비스로 고품질 이미지를 제공하는 Getty Images와 Shutterstock은 보다 직접적인 피해자들로 비춰진다.

마이크로소프트 365 "Copilot"을 구동하는 데모영상 캡처
 : CRM(고객 관리 소프트웨어), ERP(기업 내부정보 통합관리 소프트웨어) 업무를 시로 대통합

The screenshot shows a Microsoft Teams meeting interface. The main window displays a transcript of a conversation between Daisy Phillips and Alberto Burgos. The transcript includes the following text:

- Daisy Phillips (00:35): Ok, great. So next step I will send you an email with our special discount. You'll need to sign up and we'll get back to you.
- Alberto Burgos (00:40): Got it. Can you also send me your product catalog? I've seen your latest ad and was curious to see what other new products you've got.
- Daisy Phillips (00:45): Of course. I'll send you our catalog right after our call. Is there anything specific you're interested in?
- Alberto Burgos (00:52): Well, I've been thinking about updating our storage facilities and I can see a couple of your products that can help us there.
- Daisy Phillips (01:03): That's great to hear. I think we have a wide range of products that could fit your needs. Especially when it comes to smart storage management. Is that something you could be interested in?

Below the transcript is a sentiment timeline showing the flow of the conversation with markers for 'Greetings and introduction', 'Price', and 'Closing'. A play button is visible over the timeline, and a 'Copy to clipboard' button is present.

The screenshot shows the Microsoft Supply Chain Center dashboard. The main content area displays a news article titled "Flash flooding In Michigan Could Impact This Year's Tomato Harvest". The article is categorized as "Immediate risk" and "Industry supplier". The text of the article states: "Michigan's tomato crop has been hit hard by recent floods caused by heavy rainfall in the region. The flooding has disrupted the supply chain for several tomato processing companies, including Northwind...".

Below the article, there is an "Impact" section with a "Contact partner" button. A table lists the partners affected by the flooding:

Partners (3)	Type	ID	Contact status
NT Northwind Traders	Supplier	CS-102	Mark as contacted
VA VanArsdel, Ltd.	Supplier	CS-103	Mark as contacted
BF Best For You Organics Company	Supplier	CS-104	Mark as contacted

자료: Microsoft

발등에 불이 떨어진 Getty Images와 Shutterstock은 엔비디아와 협력해 맞춤형 비주얼 언어모델을 구축하고 있다. 어도비(Adobe)도 엔비디아와 손을 잡았다. 엔비디아의 최신형 데이터센터 GPU인 H100을 사용해 AI 기능을 구축하고 있다고 발표했다.

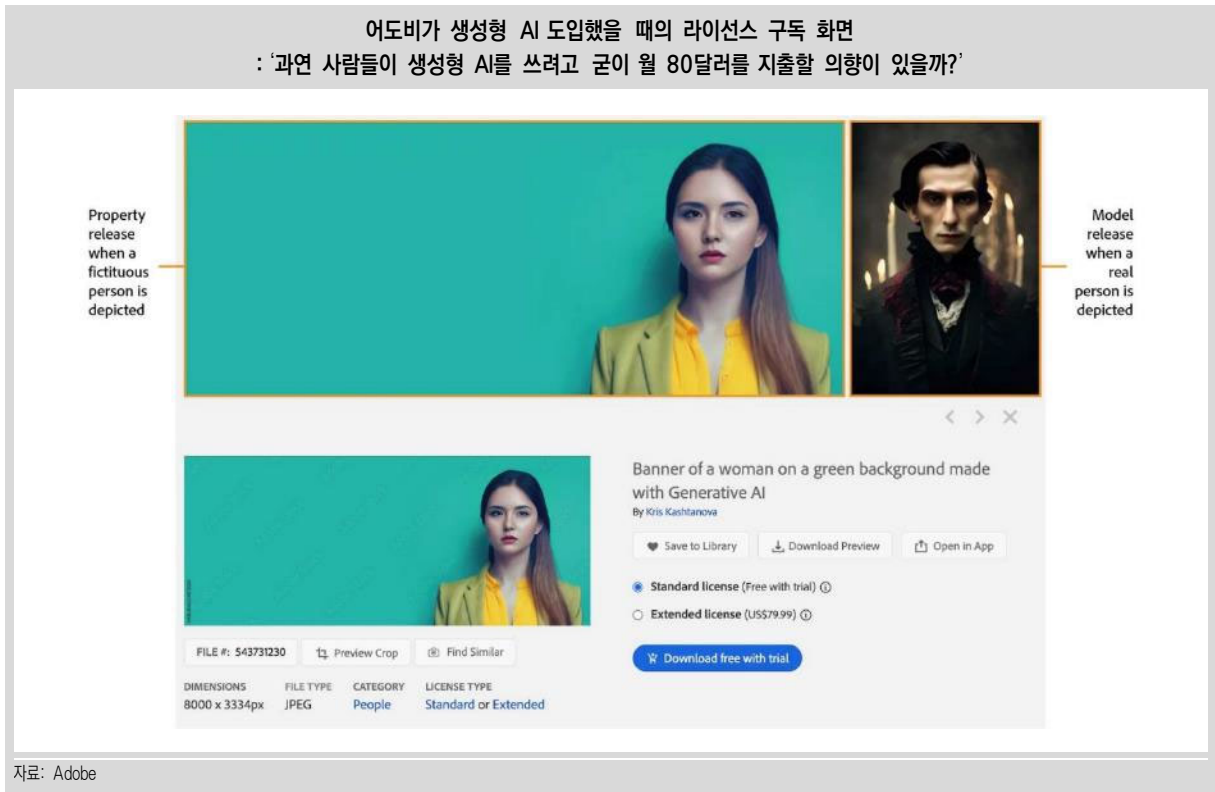
Getty Images 스타일로 워터마크까지 달아 스테이블 디퓨전으로 생성한 그림



자료: 자료 이미지(by Stable Diffusion), 프롬프트(by The Verge)

어도비의 생성형 AI 업무를 맡은 Alexandru Costin 부사장은 "AI가 창조적인 크리에이터들을 대체할 것이라고 생각하지는 않는다. 다만, AI를 사용하는 사람들이 AI를 사용하지 않는 사람들보다 더 경쟁력이 있을 것이라고 생각한다. 이것이 우리가 모든 사용자에게 AI를 사용하게끔 하는 이유다"라고 언급했다. AI 혁명이 인간의 창의성을 대체할 수는 없을 것이라고 언급하고 있지만, AI의 파급효과를 상당히 의식하고 있는 것으로 보인다. 또한 그는 "AI에서 성공하려면, AI를 받아들이는 것이 유일한 방법이다."라고 말했다. 어도비는 최근 재빠른 상황인식으로 AI 투자에 열을 올리고 있다.

어도비는 지난 3월 21일, 이미지 생성형 AI 서비스인 Firefly를 내놓았다. 아직은 베타 테스터 수준에 머물러 조금 더 지켜볼 필요는 있어 보인다. 어도비는 자체적으로 보유하고 있는 '이미지 재고(stock)'가 많기 때문에 이미지 딥러닝 학습에 있어서 유리한 입지를 지니고 있다는 점은 분명한 장점이다. 현재 Firefly는 어도비 웹사이트에서 사용가능 하지만, 향후 포토샵(이미지 편집 소프트웨어)과 일러스트레이터(디자인 작업 소프트웨어), 그리고 익스프레스(비디오 제작 소프트웨어)에 적용할 계획이라고 밝혔다.



한편, 어도비의 AI 개발은 기술적인 문제만 고려한 것은 아니다. 생성형 AI가 야기할 여러 법적 문제를 회피하기 위하기 위함도 있다. 최근 Getty Image는 Stability AI를 저작권 침해 혐의로 고소했다. AI가 학습에 Getty Images의 사진을 활용했으니 저작권 침해라는 것이다. 어도비가 Firefly를 출시한 것은 이러한 문제를 의식한 결과라고도 볼 수 있다.

이는 20년 전 애플이 iTunes 스토어를 내놓은 전략과 유사하다고 보는 시각이 있다. 당시 한국의 “소리 바다”와 같이 전세계 각국에서는 P2P 서비스로 음악을 마음껏 다운로드 받았지만, 음원사들의 반발과 소송전으로 인해 결국 역사 속으로 사라졌다. 이로써 저작권 이슈를 해결한 iTunes 모델이 성공할 수 있었다.

이러한 저작권 관련 리스크를 최소화하기 위해 어도비는 자체 이미지 데이터에 기반해 학습시킨 Firefly를 도입하고자 하는 것일 수 있다. 그런데, AI를 하나의 MP3 파일의 수준으로 대칭시킬 수 있는지에 대해서는 앞으로 논란의 대상이 될 수 있을 것 같다.

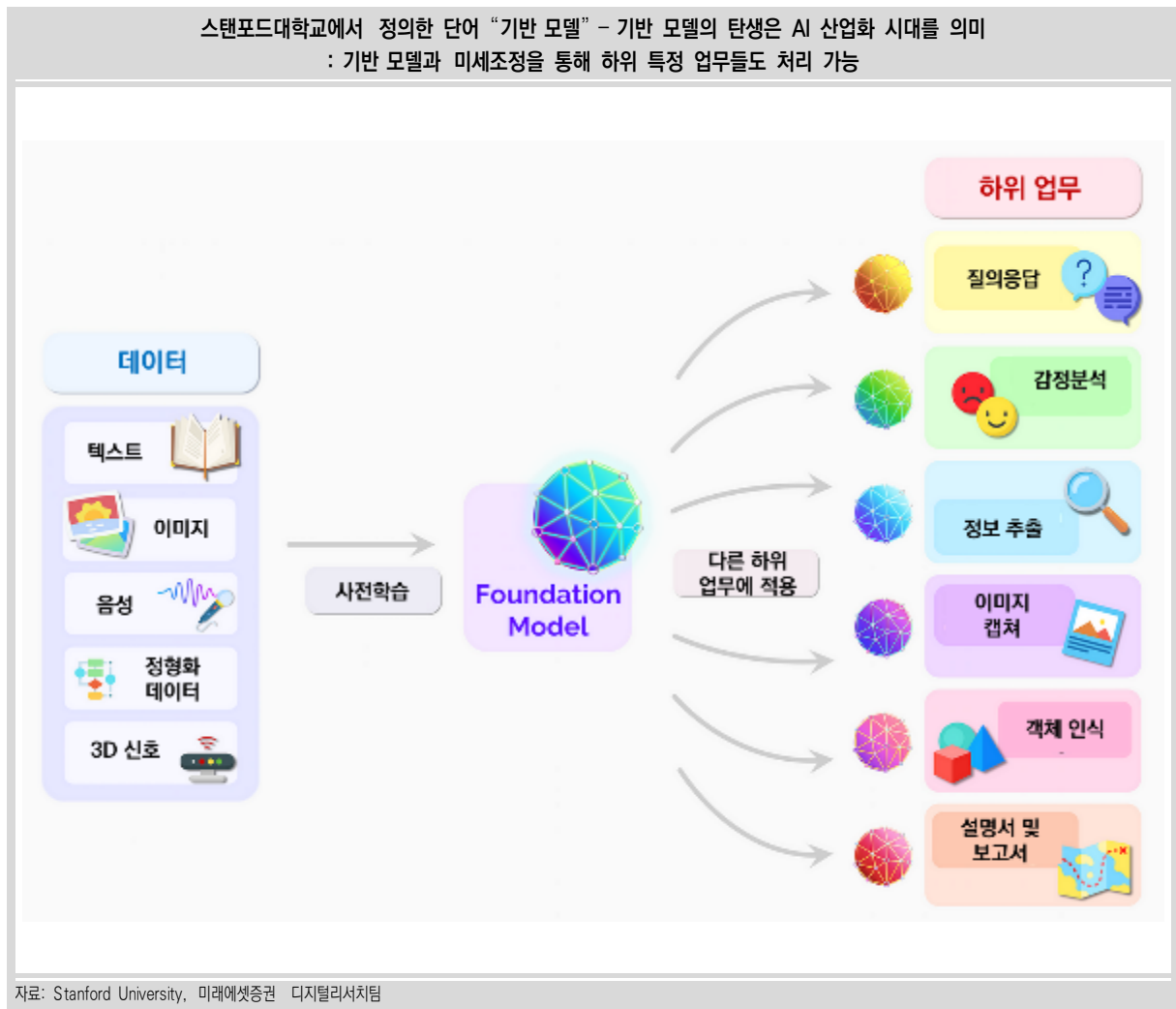
이렇게 생성형 AI로 빛을 보는 새로운 스타트업이 있는 반면, 변화하는 시장 환경에서 살아남기 위해 각고의 노력을 해야 하는 전통 기업들이 있다. 시간이 갈수록 이들의 성패는 확실히 나눌 것으로 보이며, 특히 후자의 기업들은 앞으로 살아남기 위해서라도 AI 스타트업에 대한 투자를 감행하거나 기술을 내재화할 것이다. 다만 녹록치 않을 것이다. AI 스타트업 뒤에는 신기술을 통해 보다 거대해질 빅테크들이 있기 때문이다.

AI 산업화 시대의 밸류체인

1) “기반(foundation)” 모델

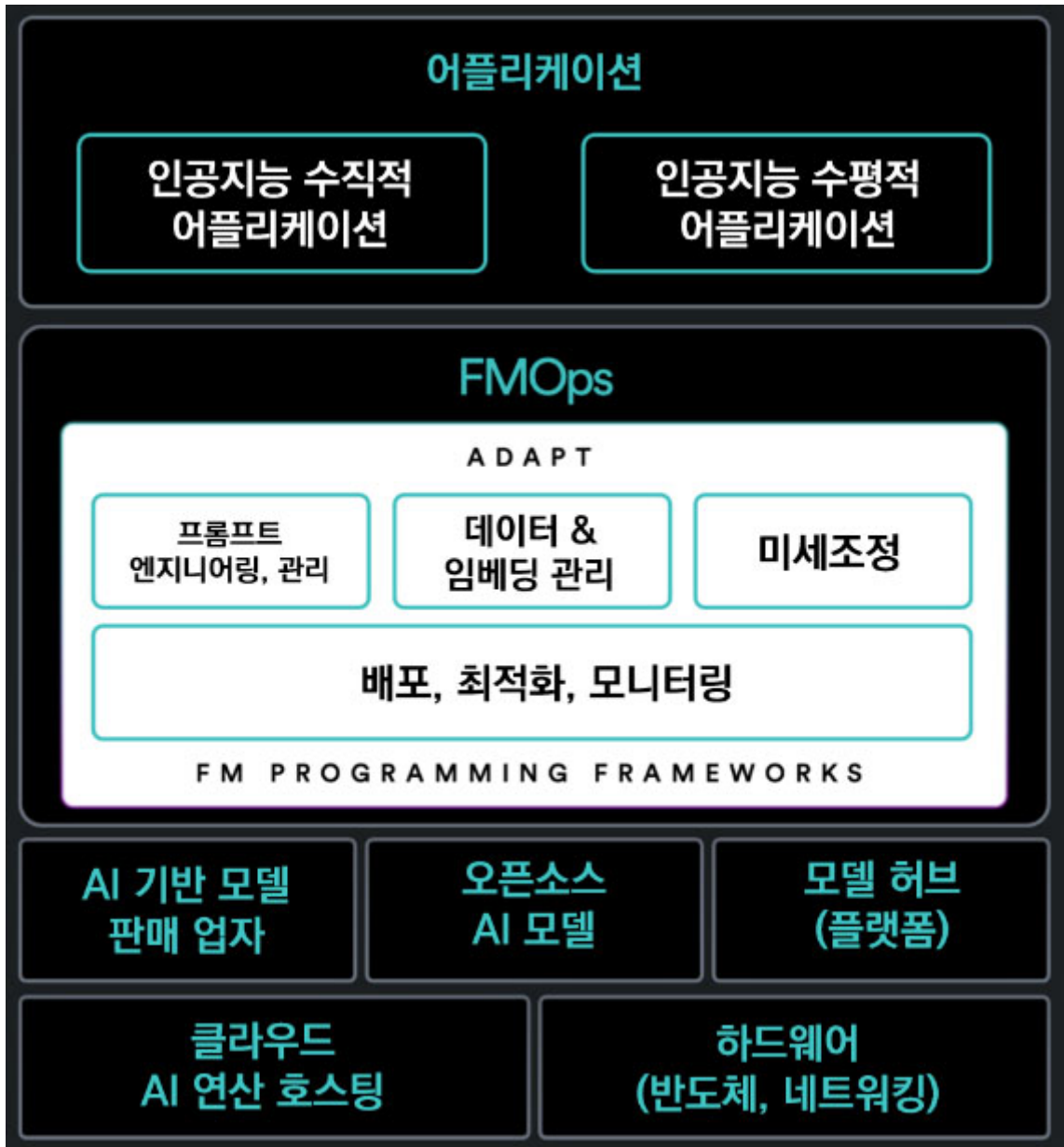
GPT-4, BERT 등의 대형 언어 모델을 기반으로 수많은 AI 어플리케이션과 서비스들이 출시가 되는 것을 보면, 마치 “AI의 산업화 시대”가 열린 것처럼 보인다. 여기서 “기반”이라는 용어는 스탠포드 대학교의 연구원들이 만든 단어다. 개념적으로 기반 모델은 다양한 하위 태스크에도 적용이 가능한 기초가 되는 대형 딥러닝 네트워크를 의미한다. 그래서 큰 규모로 사전학습된 모델을 활용해 사전학습 단계의 데이터와 유사하지만 다른 특정 하위 작업에 맞추는 미세조정(파인 튜닝)을 해서 사용하겠다는 것이다.

자연어 처리를 위한 언어모델로 사전학습을 한 것을 배달 앱에 등록된 음식 리뷰의 감정 분석에 사용하는 것이 대표적인 예다. 모델의 우수한 성능이 여러 사례를 통해 확인되었고, 이를 바탕으로 다양한 도메인 (특정 전문 지식) 모델이 등장하며 그 확장성이 커지고 있다.



기반 모델로 탄생한 각종 특화된 AI 서비스는 주로 인간의 생산성을 개선하기 위한 것들로 이루어져 있다. AI 서비스에서 가장 앞서가고 있는 마이크로소프트와 구글도 생산성에 초점을 맞췄다. 기반 모델의 대유행 이후 AI 에코 시스템과 밸류체인이 점차 체계적으로 구성되어 가고 있다.

AI 기반 모델, 미들웨어, 최종 서비스까지의 밸류체인 도식화
 : Layer 1(기반 모델, 컴퓨팅, 클라우드 등)를 기반으로 이제는 새로운 가치창출이 등장



자료: Foundation Capital, 미래에셋증권 디지털리서치팀

지금까지는 AI 에코시스템에서 가치창출의 대부분이 인프라 레벨에 집중돼 있었다. 언어모델에서 GPT(마이크로소프트), BERT(구글), BART(페이스북)가 경쟁한 것처럼 어떤 기업이 보다 우수한 기반 모델을 만들 수 있는지가 초미의 관심사였다. 이는 블록체인의 상황과도 닮아 있다. 너도나도 주류 플랫폼 혹은 메인 네트워크가 되고자 뛰어들었지만 수많은 곳들이 역사의 뒀안길로 사라졌다. 하지만, 모두가 Layer 1이 될 필요도 없고, 또 그럴 수도 없다.

그럼에도 AI 밸류체인에서 기반이 되는 Layer 1은 빅테크들에게 충분히 매력적이다. 다수의 AI 기반 모델(foundation model)을 보유한 기업들은, 다른 기업들이 소프트웨어와 어플리케이션을 구축할 수 있도록 하는 서비스를 제공함으로써 새로운 수익화 전략을 구상할 수 있다. 거대한 기반 모델을 바탕으로, API 서비스를 제공하면서 고객사에 수수료 부과하는 것이 하나의 방법이다.






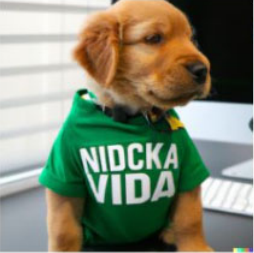

이렇게 할 수 있는 회사는 소수가 될 것이다. 엄청난 수익 기회를 만들어 낼 노른자 사업이다. 예를 들어, OpenAI의 이미지 생성 어플리케이션인 DALL-E에서 제공하는 1024*1024 픽셀의 고해상도 이미지 출력 비용은 이미지당 2센트다. 단어/이미지당 가격이 책정되고 있는데, 점차 대중화가 될수록 수익이 곧바로 늘 수 있는 구조다.

빅테크들은 여기에 목을 매게 될 것이다. 구글과 메타, AI 산업화를 이끄는 OpenAI와 Stability AI, 일찍이 AI 관련 투자를 지속해온 엔비디아가 해당 시장의 핵심 플레이어가 될 것으로 판단한다.

엔비디아의 CEO 젠슨 황은 ChatGPT 출현을 고대해왔다고 수차례 강조해왔다. GTC 2023 행사에서 엔비디아는 대단한 포부와 자신감을 표명한 바 있다. 기반 모델 판매와 모델 허브 제공, 클라우드 AI 연산 호스팅, 그리고 하드웨어 제조까지 밸류체인 전반에 걸쳐 사업을 영위할 것이라고 선언했다. Layer 1에서 제공하는 거의 모든 것들을 다 지원할 수 있다고 외친 것이다.

그러니까 ‘어떠한 인공지능 서비스를 개발하고 싶은지의 아이디어와 의지만 갖고 오면, 딥러닝 네트워크는 우리가 다 똑딱 만들어줄게’라고 엔비디아는 외친 것이다. 엔비디아는 심지어 본인들을 “AI Factory”라고 직접 닉네임을 붙이기도 했다. 앞으로 엔비디아라는 거대공장단지를 몰아내기 위해 빅테크들과 주요 소프트웨어 기업들의 경쟁은 치열해질 것으로 예상된다.

**엔비디아, 작년 11월 이미지 합성을 위한 생성형 AI 모델인 “eDiff-I” 발표
: 엔비디아는 단순히 하드웨어 업체가 아냐... AI 모델알고리즘 구성능력도 최고 수준**

Stable diffusion	DALL-E2	eDiff-I (ours)
		
<i>There are two Chinese teapots on a table. One pot has a painting of a dragon, while the other pot has a painting of a panda.</i>		
		
<i>A photo of two cute teddy bears sitting on top of a grizzly bear in a beautiful forest. Highly detailed fantasy art, 4k, artstation</i>		
		
<i>A photo of a golden retriever puppy wearing a green shirt. The shirt has text that says "NVIDIA rocks". Background office. 4k dslr</i>		

자료: Nvidia 논문 "eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers"

2) 미들웨어(middle-ware)

앞으로 AI 생태계 내에서 주목해야 할 영역은 밸류체인 전방에 위치한 실제 유저 계층이다. 최근까지는 AI 기반 모델들이 화두였다면, 앞으로는 소비자들이 실제로 사용하는 비즈니스 솔루션과 소비자 어플리케이션의 경쟁력이 새롭게 주목 받고 있다. 그만큼 최종소비자 입장에서는 AI 접근과 기술사용에 대한 진입 장벽이 낮아졌고, 이는 인공지능 테마가 지속가능성을 보일 수 있다고 주장하는 근거이기도 하다.

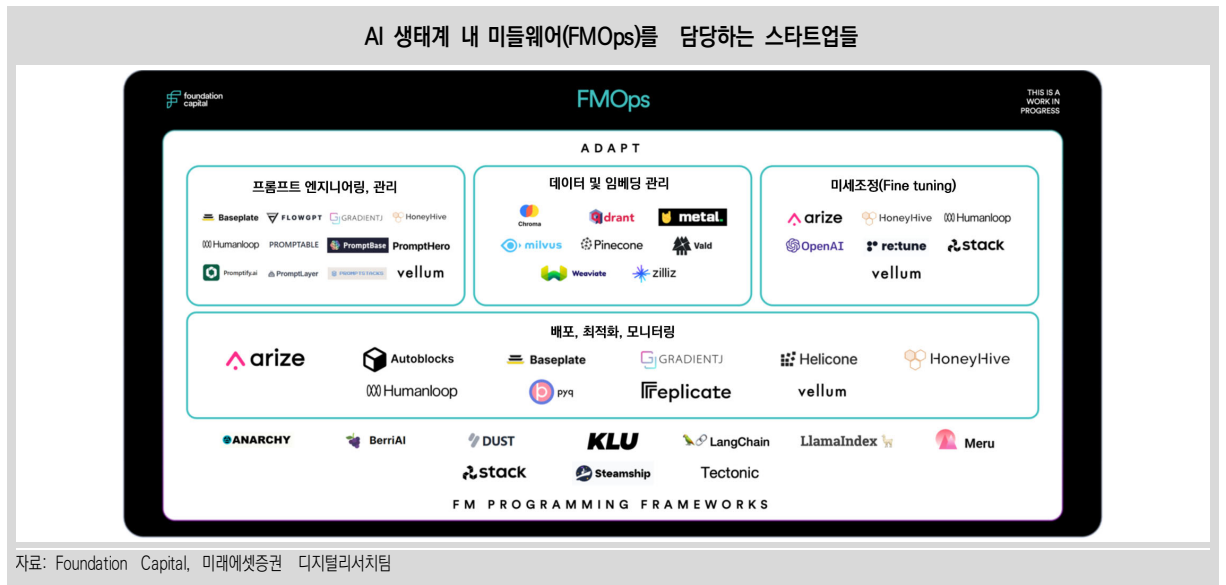
앞으로도 AI 생태계에서는 완전히 새로운 아키텍처를 구축하려는 업체들보다는, 기존의 대형 모델을 미세 조정하여 본인들의 비즈니스에 최적화하려는 업체들이 증가할 것으로 보인다. 대다수의 어플리케이션의 기반이 될 모델은 우선 GPT가 될 것으로 보인다. 현재 단계의 대부분의 서비스들도 언어 모델에 기초한 사업이 주를 이루고 있기 때문이다.

하지만, 텍스트뿐만 아니라 이미지, 영상 등 다양한 리소스를 처리할 수 있는 멀티모달 모델이 대중화된다면 이미지, 비디오 관련 생성형 AI를 적용한 어플리케이션 개발 업체들도 크게 늘어날 전망이다. 사모 영역 자금조달과 M&A 펀드도 해당 시장에 집중될 것으로 사료된다. 얼마 전 발표한 GPT-4를 포함해 극소수의 생성형 AI 기반 모델들이 향후 거의 모든 어플리케이션 생태계를 재편할 것으로 판단한다.

OpenAI를 창립한 샘 알트만 CEO도 이와 같은 의견을 낸 바 있다. ChatGPT 공개 전인 2022년 9월에 알트만 CEO는 “다음 세대를 위한 AI”라는 주제로 인터뷰를 했다. 이 자리에서 그는 많은 회사들이 기본적인 대형 인공지능 모델 제작을 시도할 것이라고 예상했다. 하지만, 그들의 자체 모델을 훈련시키는 스타트업들의 미래에 대해서는 회의적이라고 밝혔다. 대신, 다른 테크기업들이 구축해둔 큰 언어모델 위에서 API를 통해 구축되는 “중간 계층(미들웨어, middle-ware)”들이 정말 중요해질 것이라고 주장했다.

그가 판단하기에도 대부분의 스타트업들은 기반(foundation) 모델을 굳이 만들 필요가 없다는 것이다. 그보다 “미들웨어에서 많은 가치가 창출될 것이다”라고 그는 힘주어 말했다. 특정 업종에 대한 인공지능 모델을 처음부터 만드는 것이 아니고, 기본적으로 통용되는 AI 모델 위에서 구축될 것이라는 점이다.

기반 모델 GPT 시리즈를 출시한 OpenAI의 이해관계를 감안하더라도 이는 꽤 유의미한 시사점을 던져준다. 수많은 기업들이 기반 모델 개발에 나서게 된다면 모델을 학습시키기 위해 시장에서 필요로 하는 반도체 수는 기하급수적으로 늘어난다. 이는 반도체 수급 차원에서 이들 기업의 교섭력을 매우 떨어뜨리고, 칩 생산 단가가 높아지며 경제적 유인은 적어지게 된다.



샘 알트만이 특히 강조한 “미들웨어”라는 밸류체인은, AI 기반 모델과 실제 사용자들의 어플리케이션을 잇는 중간 계층이다. 실리콘밸리 VC인 Foundation Capital은 이 미들웨어 기업들을 Foundation Model Operations(이하 FMOps)라고 지칭한다. 이들은 기반 모델을 잘 활용할 수 있도록 프롬프트 최적화 및 데이터를 담당하고, 특정 태스크를 위한 미세조정을 도맡는 업무를 맡는다.

FMOps에 가운데 주목해야 할 영역 중 하나는 프롬프트 엔지니어링(Prompt Engineering)이다. ChatGPT가 일으킨 생성형 AI의 대유행에 가장 먼저 직접적 수혜를 보고 있는 분야이기 때문이다. ChatGPT 등장 이후 프롬프트 엔지니어라는 직종이 생겨났는데, 프롬프트 엔지니어는 자연어 처리나 대화형 AI 시스템의 명령문인 Prompt를 전문적으로 다루는 직업이다. 어떤 특정 분야에서 본인들의 전문 지식을 토대로 해서 프롬프트를 개발하는 사람들이다. 명령을 내리고 결과에 따라 다시 명령을 수정하는 일을 되풀이해서 원하는 결과를 만들어 내도록 하는 것이다.

모든 산업 영역에 걸쳐 AI가 도입되면서 프롬프트 엔지니어 수요가 확대될 것으로 전망된다. 미국의 엔트 로픽이라는 기업은 최근 연봉 33만5000달러라는 거금을 들여 프롬프트 엔지니어를 선발하기도 했다. Promptable과 GradientJ와 같은 스타트업들은 레디메이드 템플릿을 제공한다. 사용자가 원하는 ‘just’ 한 이미지를 얻기 위해서는 정교한 프롬프트를 작성하고 수정하는 수고스러운 작업을 거쳐야 하는데, 이런 수요에 맞는 서비스를 출시한 것이다.

**제공되는 템플릿으로는 “응답의 길이, Seed(이미지로 치면 비슷한 포즈나 풍경이 계속 나오도록 하는 값), CFG 스케일(프롬프트를 충실히 이행하는 정도. 높을수록 프롬프트 대로 이미지 생성)” 정도가 있음. 이들을 통해 원하는 답에 최대한 가까운 결과치를 얻도록 유도.*

또한, 프롬프트베이스(PromptBase)와 FLOWGPT라는 프롬프트 마켓플레이스도 존재한다. PromptBase 에는 700명 가량의 프롬프트 엔지니어들이 포진해 있다. 2022년 7월, 1개에 불과하던 프롬프트 마켓플레이스는 최근 14개로 늘어났고, 매달 증가하는 추세다. 대개, 미드저니나 스테이블 디퓨전 등 이미지 생성 프롬프트와 ChatGPT를 효율적으로 사용하기 위한 프롬프트까지 다양하게 거래된다. 마켓플레이스의 수익 모델은 건당 판매 가격 기반인데, 판매가 실제로 이루어지면 마켓이 10%의 수수료를 가져가는 방식이다.

DALL-E, 미드저니, 스테이블 디퓨전 등 주로 이미지생성을 위한 프롬프트를 파는 온라인 마켓

자료: PromptBase

소규모 사업체가 비즈니스를 하고 있는 것이기에, 아직 수익이 크지 않다. 사실 프롬프트 엔지니어링이라는 게 앞으로도 돈을 벌 수 있는 사업인지에 대해서는 의문이 든다. 결국에는 프롬프트를 그렇게 어렵게 쓰지 않아도 훌륭한 그림이 나오게 되는 이미지 생성 네트워크가 소비자들의 선택을 받을 것이기 때문이다. 번거롭게 굳이 프롬프트 마켓에 가서 사지는 않을 것으로 보인다. 기술은 인간을 널리 이롭게 하는 대중화의 방식으로 맞춰진다. 복잡하고 배우기 힘들었던 C++ 코딩언어 시대에서, 사용자 친화적인 파이썬 시대로 넘어갔고, 지금 우리는 Low-code, 심지어 No-code 시대를 맞이하고 있다. 이것과 마찬가지로 프롬프트 작성 또한 우리가 어렵다고 생각할 필요조차 느끼지 못하게 이용하기 편리하게 바뀔 것은 자명한 일이다.

미들웨어 중에서 프롬프트 엔지니어링보다 중요한 분야라고 생각되는 것은 미세조정이다. 미세조정을 거쳐야 기반모델이 하위 특정태스크에 사용할 수 있는 형태로 파라미터가 업데이트 되기 때문이다. 미세조정 작업은 기술적 전문성과 저장 공간 및 컴퓨팅 능력을 필요로 하며 새로운 데이터가 생길 때마다 모델을 주기적으로 재학습해야 하는 수고로움도 있다. 또한 미세조정 과정 중에 과적합(모델이 훈련 데이터에 지나치게 특화되어 새로운 데이터에 대해 일반화가 어려워지는 현상)과 같은 문제를 야기할 수도 있다.

이에 따라서, 이 분야는 빅테크들과 스타트업이 함께 경쟁하는 영역이 될 전망이다. 실제로 OpenAI가 사용자들로 하여금 모델을 미세조정할 수 있도록 서비스를 제공하고 있기도 하다. OpenAI를 품은 마이크로소프트가 미세조정을 해주는 솔루션 서비스를 다른 업체들에게 제공하게 될 것으로 예상된다. 한편, Humanloop와 Vellum과 같은 스타트업들도 대형언어모델을 기반으로 한 데이터 샘플 선택, 데이터 분포 매핑 등의 파인튜닝을 제공하고 있다. 다만 빅테크와의 경쟁에 있어서 마케팅 측면에서 경쟁력이 떨어지기 때문에, 해당 분야에서 많은 M&A 케이스가 창출될 것으로 사료된다.

3) 비즈니스 솔루션과 어플리케이션

한편, 현재 미들웨어보다 훨씬 더 뜨겁고 주목을 받아 받는 분야는 비즈니스 솔루션과 소비자 어플리케이션이다. 실제로 지갑을 열어 산업의 지속가능성을 부양해줄 주체는 최종 소비자이기 때문이다. 최종 소비자들을 대상으로 한 서비스들이 본격적으로 개화돼야 생성형 AI가 단기적인 테마가 아니라 장기적인 트렌드로 전환될 것이다. 물론 이런 걱정을 지금 단계에서는 할 필요가 없다.

엔비디아의 CEO 젠슨황은 GTC 2023에서 “ChatGPT가 만든 지금의 상황이 애플이 아이폰을 출시했을 때와 같다”라고 말하기도 했다. 현재 AI 관련 어플리케이션 시장에서 벌어지고 있는 일들은, 약 15년 전 애플이 만들었던 ‘앱스토어 모먼트’를 연상시킨다. 앞으로 AI 산업화 시대를 맞아 초창기 앱스토어 때처럼 사람들에게 많이 선택되는 수많은 킬링 서비스들이 출현할 것이다.

이제껏 AI 관련 스타트업들에는 주로 자율주행 솔루션이나 센서 기업에 관심이 쏠렸었지만, 당분간은 그 주인공이 바뀔 것이다. ChatGPT 모먼트로 인해 생성형 AI 어플리케이션들이 주목을 받게 될 것이다. 이 영역에서 가장 많은 숫자의 투자가 이뤄지지 않을까 생각해본다. 그리고 빅테크들은 해당 기업들 중 딱일이 다른 기업들을 품기 위해 인수합병에 나설 것이라 생각한다.

세계 최고의 VC 중 하나인 세콰이어 캐피탈이 생성형 AI 어플리케이션 기업들에 대해 정리해둔 내용이 있다. 종류가 다양한데, 결국 핵심은 ‘자연어-to-something’이다. 자연어로서 우리는 AI에게 명령을 내려서, 글을 쓰게 하고(text), 이미지 및 영상을 만들고(Image, Video), 그리고 프로그램을 하도록(code) 하는 것처럼 말이다.

생성형 AI 어플리케이션 관련 업체 구성 목록

The Generative AI Application Landscape v2



A work in progress



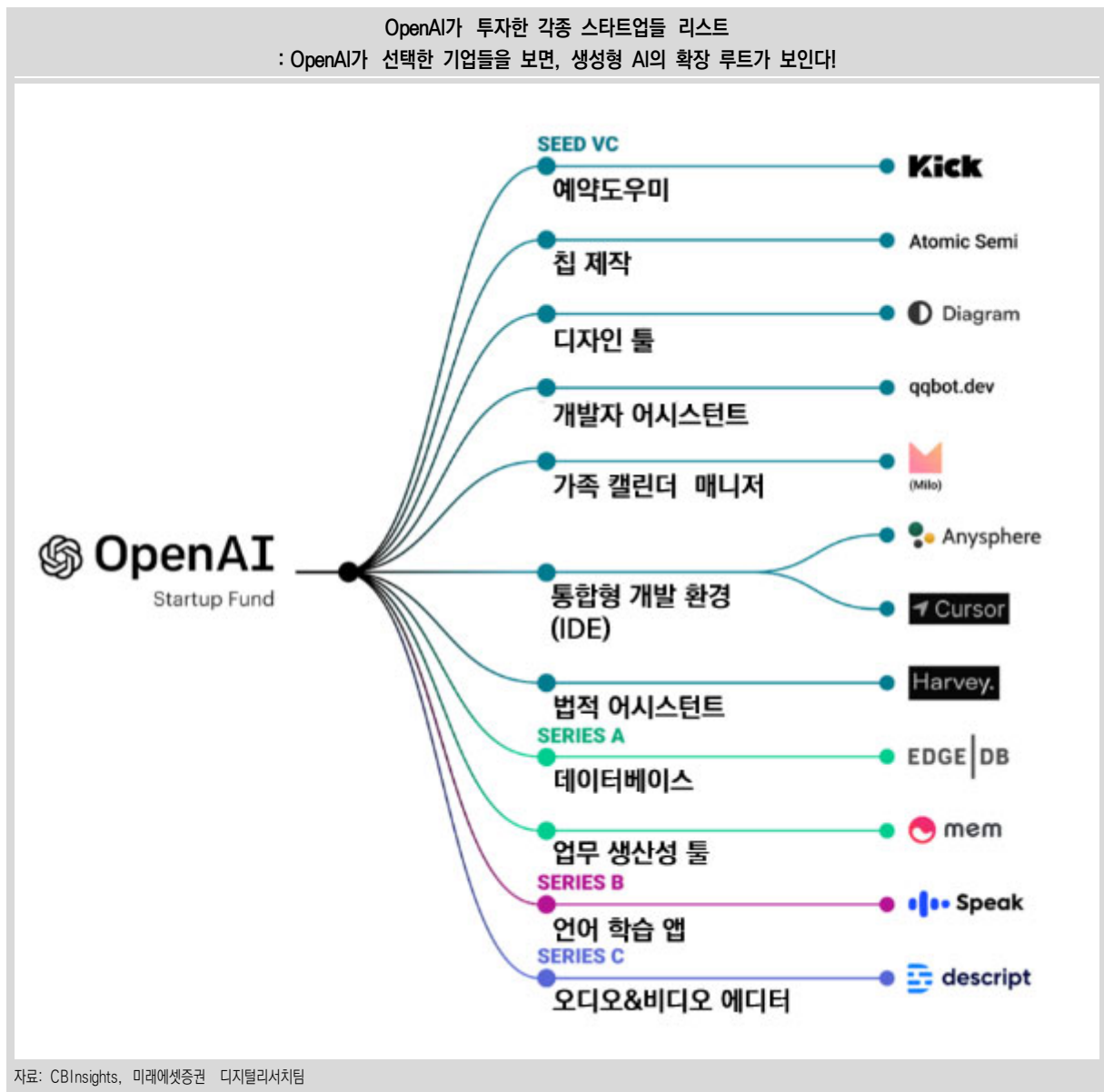
자료: Sequoia Capital

4) OpenAI to the next level

다음 세대에 꽃피울 AI 스타트업에 미리 투자함으로써 빅테크들은 본인들의 기반 모델과 그 에코시스템을 강화할 것이다. 그리고 그 선봉에는 OpenAI가 있다. OpenAI는 현재 실리콘밸리에서 가장 활발하게 AI 관련 어플리케이션 및 솔루션 기업들에 여러 지분을 투자하고 있다.

애초에 OpenAI를 이끌고 있는 샘 알트만이 미국 스타트업계에 잔뼈가 굵은 사람이다. 샘 알트만은 Y콤비네이터(Y Combinator)의 대표 출신인데, Y콤비네이터는 스타트업이 유니콘으로 잘 성장할 수 있도록 도와주는 '엑셀러레이터' 기관이다.

OpenAI는 AI 생태계 구축의 선도 주자로 자리 잡고, 이익 창출 방향으로 나아갈 것으로 기대된다. OpenAI는 실제로도 미국 벤처 시장의 하락 추세에 정면으로 맞서고 있다. OpenAI는 자체 벤처 투자를 확대해 광범위한 AI 생태계의 중심에 자리잡고 있다. 2022년 10월 이후, OpenAI의 1억 달러 규모의 "Startup Fund"(마이크로소프트를 포함한 여러 LP들도 존재)은 주로 초기단계 AI 스타트업들을 포함하여 총 12개의 투자를 집행했다.



OpenAI는 여러 투자를 통해, 각종 인공지능 솔루션들이 동사의 기반 모델인 GPT를 기반으로 구축되도록 할 것이다. 동시에 이들이 먼저 GPT 차세대 버전을 경험할 수 있도록 권한을 지불하는 식으로 이용할 것이다. 스타트업 입장에서는 신형 인공지능 모델 알고리즘을 사용할 수 있으니 이를 환영할 것이다. OpenAI는 이들을 일종의 베타 테스팅베드로 효과적으로 활용할 수 있다. 스타트업이 GPT를 통해 발전하는 다양한 사례들을 잠재 고객들을 대상으로 마케팅 포인트로 적극 내세울 수도 있다. 다시 말해, 스타트업들에게도 OpenAI에게도 Win-Win인 것이다.

그리고 OpenAI가 배팅을 하고 있는 스타트업들을 살펴보면 그들이 어떤 서비스들을 앞으로 ChatGPT에 연동시킬지 유추해볼 수 있다. 아래는 OpenAI가 투자한 각 기업들의 기업 개요를 정리해본 것이다.

- **소비자 일상용 어플리케이션:** AI 기반 예약 도우미(Kick), 언어 학습 앱(Speak)
- **일반 생산성 도구:** 근로자들의 업무생산성을 높이기 위한 노트 작성 앱(Mem), 가족용 일정 공유 관리 매니저(Milo)
- **산업용 솔루션:** 인공지능 법률 도우미(Harvey), 회계 자동화 플랫폼(Kick)
- **개발자 중심 도구:** AI 기반 통합 개발 환경(Anysphere과 Cursor), 코딩 챗봇(qqbot.dev)

**이들은 OpenAI의 첫 번째 Converge 액셀러레이터에 참여한 기업들*

- **창의성:** 생성형 AI를 사용한 팟캐스트, 비디오 스크립트(Descript와 Diagram), 디자인 이미지 파일과 오디오 파일 생성 및 편집 도우미(Figma)
- **반도체 및 인프라:** AI 기반 칩 설계업체(Atomic Semi), 컴퓨팅 및 엔지니어링 인프라위한 유사관계형 데이터베이스(EdgeDB)

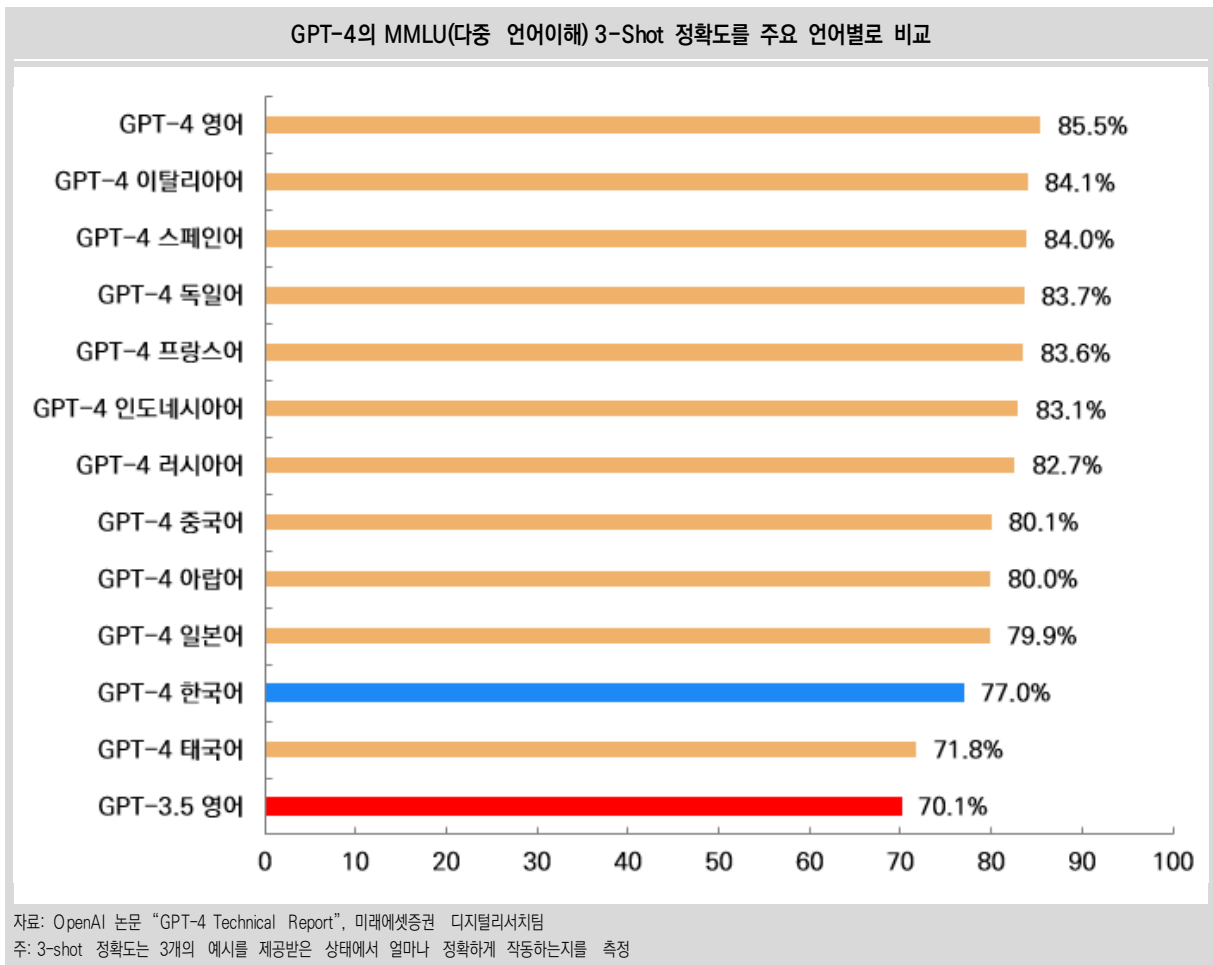
이처럼 OpenAI는 여러 스타트업과의 협력 및 투자를 통해 다양한 영역으로 본인들의 영향력을 확장해나 가며 빠른 속도로 시장을 잠식해나갈 수 있다. 예를 들어, ChatGPT에 없는 칩 설계 등 특정 전문 영역 도메인 지식을 학습해서, 반도체 팹리스 영역에서도 두각을 나타낼 수 있다. 이렇게 된다면 현재 약점으로 꼽히는 전문영역에서의 신뢰성 문제로부터 훨씬 더 자유로워질 수 있다.

또한, 예약이나 언어 학습 같은 소비자 일상용 어플리케이션에 진입해 기존 플레이어들을 위협할 것이다. 그리고 이런 일상용 서비스들이 점차 대중에게 선택을 받게 되면, 자연스럽게 ChatGPT 플랫폼에 플러그 인되며 활용도가 높아질 것이다. OpenAI가 이렇게 계속 영역을 확장한다는 건, 사실상 모든 인터넷 영역을 서로 연결한다는 의미이기도 하다.

사실 한국인 입장으로는 무서운 부분이다. 갈라파고스 같이 국내에서만 활용되는 인공지능 서비스는 그 경쟁력이 약화될 수 있다. 데이터 싸움과 서비스의 유연성에서 이를 따라가기가 힘들기 때문이다. 특히 GPT-4의 한국어 수준은 GPT-3.5 영어수준 보다도 좋다. 이제는 대부분의 한국 인공지능 모델보다 ChatGPT가 한국어를 훨씬 더 잘 이해한다는 뜻이다.

앞으로 주류 대형 언어 모델이 갖는 한국어와 영어의 수준 차이는 대략 1년 미만을 유지할 것이라고 보는 시각이 있으며 한국어, 일본어 등 특정 언어에만 집중한 딥러닝 모델은 경쟁력을 상실하게 될 전망이다. AI 산업화 시대의 글로벌라이제이션은 지금보다도 훨씬 심화될 수 있다.

우리는 투자자로서 이 기회를 면밀히 살펴봐야 한다. AI 대전에서 승자가 등장한다면, 그 기업은 들도 보도 못한 커다란 트로피를 받을 것이 분명하다.



메타버스는 AI 진보의 최고 수혜 영역

2023년 1월 ChatGPT를 만든 OpenAI가 마이크로소프트로부터 100억 달러를 유치했다. ChatGPT가 유래가 없는 빠른 속도로 유저를 끌어모으자 다급해진 다른 기업들도 인공지능에 대한 다양한 서비스를 내놓고 있다. 그야말로 2023년은 인공지능의 해가 되고 있다. 대다수의 사람들이 생성형 AI에만 몰두를 하고 있는 상황이다. 여기서 가장 소외받고 있는 영역이 바로 메타버스, 블록체인, NFT 등이다. 팬데믹 직후부터 ChatGPT 열풍 직전까지 가장 주목을 받았던 분야가 바로 이들이었다는 점이 아이러니하다.

관심을 받지 못한다는 것은 VC나 빅테크들의 투자 우선순위에서 밀린다는 것을 의미한다. 실제 메타의 경우도 당장 메타버스보다는 인공지능에 열중하겠다고 밝히자 투자자들이 안도의 한숨을 내뿔기도 했다. 그런데 여기서 의문점이 든다. 인공지능의 유행과 발전으로, 정말 메타버스와 같은 기술은 사양길을 걷게 될까?



우선 우리의 대답은 “그렇지 않다”이다. AI와 메타버스가 서로 경쟁하는 테마라고 생각될 수도 있지만 딥러닝과 생성형 AI는 그보다 훨씬 기반이 될 기술이다. 리포트의 제목처럼 “Everthing, Everwhere”에 쓰일 수 있다. 산업을 가리지 않고 모든 영역에 쓰인다는 말이다. 인공지능은 인간의 생산성 높여줄 훌륭한 도구로서 인간이 활약하고 있는 많은 영역에서 사용될 것이다.

기본적으로 메타버스와 블록체인 등도 인간 개발자가 코딩을 해서 만드는 작업물의 영역이다. 따라서 이러한 영역에서 인공지능이 조력자로서 활용될 수 있다. AI 산업화 시대에 들어서면, 오히려 메타버스와 블록체인 기술이 고도화될 수 있다. 코딩을 통한 프로그램 개발이 더 쉬워지고, 더 많아지며, 더 빨리 전파될 것이기 때문이다. 다시 말해, 인공지능 기술은 마치 인터넷이나 전기와 같은, 'Layer 0' 기술이 될 것이다. 그 위에서 메타버스와 블록체인 테마는 딥러닝의 진보에 큰 수혜를 입을 것이다.

먼저, 메타버스라는 것이 상당히 애매모호한 개념이지만, 해당 리포트에서는 가상세계와 현실세계의 융합이라는 관점으로 접근을 해보려고 한다. 그런 측면에서 메타버스를 가장 잘 설명하는 산업은 다른 아닌 게임 산업이다. 사티야 나델라 마이크로소프트 CEO도 메타버스를 가장 잘 표현하는 채널이 게임이라고 했으니 말이다.

참고로 마이크로소프트가 액티비전블리자드를 인수하려는 시도는 본인들의 메타버스 비즈니스를 더욱 확장하기 위함이었다. 그리고 그 인수시도는 지금도 격렬하게 진행 중이다. ChatGPT와 New Bing을 통해 가장 직접적인 커다란 수혜를 입고 있는 것이 마이크로소프트이지만, 이 기업 역시 메타버스에 대한 관심은 계속 끌여가고 있다는 것을 잊어서는 안된다. 인공지능이 당연히 여겨지는 근래에 또 다시 가장 뜨겁게 달아오를 테마가 바로 게임이 될 것이다.

1) 게임

메타버스 게임 기술 발전은 AI 기술 발달의 직접적인 수혜를 받을 것으로 예상된다. 샘 알트만 OpenAI CEO는 인공지능이 메타버스의 업사이드 케이스를 발생시킬 합리적 기회가 있다고 말했다. 여러 생성형 AI는 메타버스 세계 내 가상환경을 재현하는데 탁월한 기능을 한다. 특히 ChatGPT 같은 기술들은 게임 안에 대화형 기능을 만들어내서 가상현실에 친숙한 유저들에게 더 몰입감있는 경험을 제공할 수 있다.

예를 들어, ChatGPT 형태의 챗봇은 개발자 대신 스토리 혹은 기획서를 작성하거나 게임 내 필요한 캐릭터들에게 대사를 생성해 부여할 수 있다. 또한 게임 플레이어에서도 활용이 가능할 것으로 기대된다. 예를 들어 LoL에서 추천 챔피언, 상대 챔피언 대처법, 스킬 순서, 동선 추천 등 전략적 가이드를 실시간으로 제공하는 챗봇의 출현도 상상해 볼 수 있다. 게임의 스토리 착안부터, 개발, 실행, 운영에 이르는 모든 분야에 인공지능이 활용될 수 있다.

사실 딥러닝 기술은 게임 개발의 다양한 분야에서 채택되어 왔다. 그 중에서도 대표적인 변화가 “게임에서의 절차적 생성기술”이라고 불리는 PCG(Procedural Content Generation) 방식이다. PCG는 게임에서 쓰이는 거의 모든 구성요소(콘텐츠)를 개발자가 아니라 기계가 자동으로 생성하는 기술을 일컫는다. 프로그래머가 직접 게임 내 지형을 생성하는 것이 아니라 인공지능 알고리즘과 인풋 데이터의 조합을 통해 만드는 것이다. 이때 게임 개발자들이 생각하지 못했던 새로운 개념의 콘텐츠를 착안하기도 한다.

**게임 콘텐츠로 사용될 수 있는 것에는 “지도/지형, 레벨 디자인, 스토리, 퀘스트, 적(enemy), 게임 규칙, 게임 전략, 아이템, 카메라 등” 게이머의 경험에 영향을 미칠 수 있는 설정 가능한 요소를 모두 포함*

PCG를 위한 기계로써 딥러닝이 잘 사용되는 이유는 PCG의 목적과 가장 부합하기 때문이다. PCG는 게이머의 특성을 고려해서 랜덤한 무작위 출력값을 자동으로 생성하는 것이 목적이다.

게이머의 기호나 플레이 실력 등에 맞추어서 콘텐츠들을 자동으로 생성해야 한다. 콘텐츠를 생성할 때는 개인적 경험 부분을 효과적으로 반영시켜야 한다. 완전히 예측이 가능할 정도의 콘텐츠가 되어서도 안되고, 그렇다고 완전히 무작위로 콘텐츠로 생성해서도 안된다.

그런데 여기서 특성이라고 하는 것을 인풋 데이터라고 부르고, 그것을 고려해서 출력값을 생성하는 것은 조건부 확률이다. 그리고 이 조건부 확률은 트랜스포머와 같이 생성형 AI의 근간이 되는 함수다. 생성형 AI의 경우, 난수생성값인 random seed 조작을 통해서 원하는 정도의 무작위성을 도출해 출력값으로 표현할 수 있다.

딥러닝 생성형 AI 기술은 PCG의 핵심 기술이다. 그리고 인공지능이 가장 강점을 발휘할 수 있는 분야가 개인화라는 점을 고려하면, 앞으로도 PCG의 근간 네트워크는 딥러닝이 될 것으로 보인다. 이와 관련해서 며칠 전 유니티 소프트웨어의 John Riccitiello CEO는 생성형 AI가 플레이어의 입력데이터에 기반한 개인화된 비디오 게임을 가능하게 할 것이라고 언급 했다.

PCG는 게임 산업에서 실질적으로 적용될 수 있는, 아니, 적용해야만 하는 기술이다. 왜냐하면, 인공지능을 통한 코딩 자동화를 통해 개발자들의 중복되는 코드 작성을 최소화해주기 때문이다. 인공지능이 자동으로 코딩을 해주는 식이다. 이는 개발자의 수고를 덜어주고, 이들이 보다 창의적인 작업에 힘을 쏟게 도와준다. 게임을 제작할 때, 걸핏하면 크런치모드로 일을 해야 하는 게 게임 개발자들의 숙명과도 같았는데, 이러한 부분을 많이 해소될 것이다. 게임 기업들의 대부분 비용이 인건비라는 점에서 PCG는 게임 제작에는 필수라고 할 수 있다.

**크런치모드(Crunch mode): 주로 소프트웨어 업계와 게임 업계에서 쓰는 용어로, 제품 출시 데드라인을 맞추기 위해 야근과 특근을 수일 이나 수개월 반복하는 것*

특히 ‘오픈 월드 게임’ 개발에 주로 사용되는 게임 콘텐츠 생성 기술이 바로 PCG다. 오픈 월드 게임은 게임계의 새로운 트렌드라고 할 수 있다. 비디오 게임 역사상 가장 많이 팔린 게임 타이틀인 “GTA 시리즈”와 강력한 팬덤을 가지고 있는 “젤다 시리즈” 또한 오픈 월드이기에 그만한 인기를 구가하고 있는 것이다. 그래서 많은 게임 개발사들이 오픈 월드식 게임을 출시하고는 있지만 상당히 제약이 많다.

방대한 맵 월드 공간을 구현해 내는 것도 어렵고, 그 맵 월드 안에 재밌는 콘텐츠를 채워 넣는 건 훨씬 더 어렵다. 이것을 수작업으로 일일이 채우는 것은 불가능에 가깝다. 할 수는 있지만 인건비가 엄청나게 증가하고 개발기간이 늘어나는 것도 덤이다.

PCG는 딥러닝 기술을 활용해 게임 프로그래밍 내에 콘텐츠가 자동으로 삽입될 수 있도록 한다. PCG는 사람의 개입 없이 만들어지다 보니 유사한 형태의 콘텐츠가 양산되어 사용자의 흥미를 떨어뜨린다는 비판이 있다. 하지만, 앞으로 딥러닝과 생성형 AI에 대한 투자가 광범위하게 이루어진다면 해당 분야의 진보도 빨라질 것이며, 오픈 월드 게임의 숫자는 과거와 비교할 수 없을 정도로 늘어날 것으로 예상된다.

글로벌 유명 게임웹진인 IGN이 선정한 최고의 오픈 월드 게임 3곳
: (왼쪽부터) GTA, 젤다의 전설, 호라이즌 시리즈



자료: IGN

*오픈 월드 게임은 게임의 장르 중 하나. 오픈 월드라는 기준이 명확한 것은 아니지만, 기본적으로 플레이어의 행동에 제약을 두지 않는 것이 특징. 이동 및 행위의 자유를 전제로 해서 게임에 구현되어 있는 거의 모든 상호작용을 경험해볼 수 있는 특징. 이에 따라 플레이어로서 하여금 훨씬 더 높은 몰입감을 주는 형태. 즉, 메타버스의 철학과 가장 닮아 있는 장르.

앞으로는 PCG 기술은 많은 게임에 적용될 것임은 분명하다. 그리고 인공지능에 대한 관심이 실제로 크다는 것을 알게 해준 있는 행사인 GDC 2023 컨퍼런스가 지난 3월 20일부터 24일까지 성황리에 개최되었다. GDC는 일반 게이머를 대상으로 하는 B2C 행사가 아니라 개발자를 포함한 업계 관계자 중심의 B2B 행사다. 그래서 게임 출시 소식보다는, 게임 업계의 최신 첨단 기술 동향을 파악할 수 있다. 애플로 비교하자면 신제품 발표 자리가 아닌, 개발자 회의인 WWDC의 성격이다.

게임개발자들을 위한 컨퍼런스 “GDC”: 엔데믹에 더해 AI 모먼트 덕분에 참가자는 거의 2배 증가



자료: Venturebeat

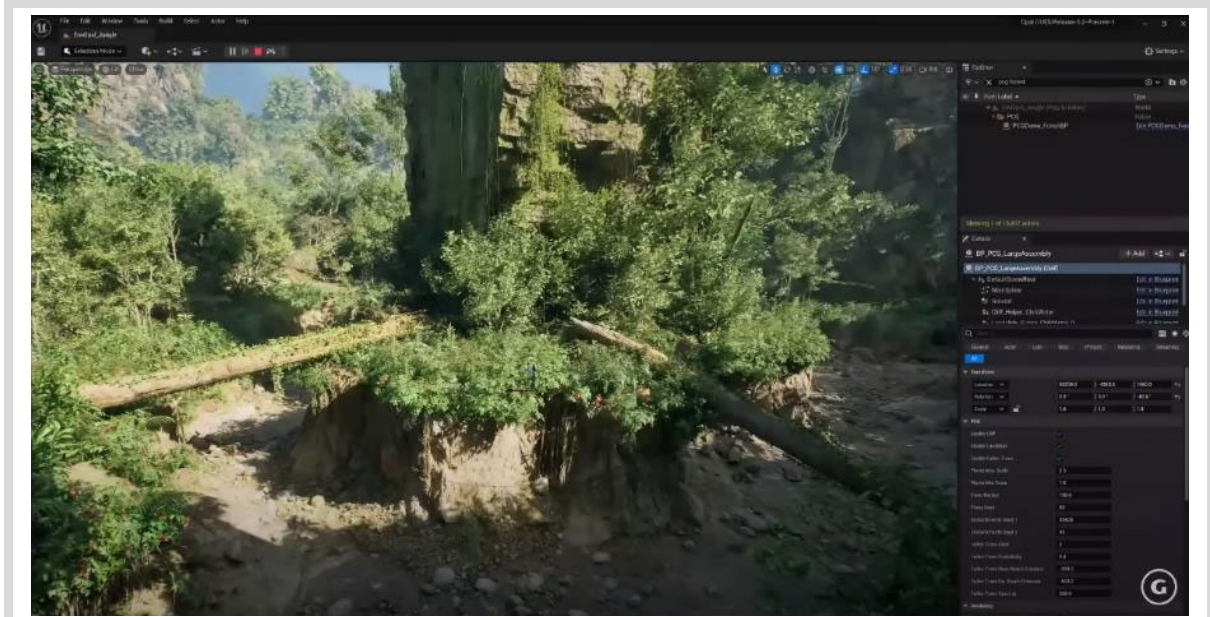
그런데 올해는 GDC의 위상이 예전보다 훨씬 높아졌다. 인공지능이 2023년의 최고 화두가 됐기 때문이다. PCG와 같은 곳에서 인공지능이 탁월하게 활용되고 있기 때문에, GDC 행사에서도 가장 뜨거운 감자는 AI였다. AI 붐이 행사 흥행의 촉매제가 된 것으로 보인다고 참석자들은 하나같이 입을 모았다. AI가 게임 개발의 생산성을 높이는 것은 물론이고, 이려다가 게임 개발자나 디자이너의 일자리까지 빼앗아 가지 않을까 하는 불안심리도 커졌다. NFT 메타버스 게임인 더샌드박스(더샌드박스)의 알바로 라마르체 토로자 대표는 “게임업계에서 AI를 도입하기 시작하면서 개발자 등 업계 관계자들이 큰 관심을 보이고 있다”라고 전했다.

GDC 행사에서 연사들의 발표 가운데 충격적인 내용들도 많았다. 생성형 AI 혁명으로 높은 품질의 게임이 간편하게 제작될 수 있다는 사실 때문이었다. 그중에서 가장 깜짝 놀랄만한(eye-catching) 테크 데모를 내놓은 곳은 에픽게임즈의 언리얼 엔진이었다. 에픽게임즈는 유니티소프트웨어(티커: U)와 함께 게임엔진에 있어서 가장 선두주자로 꼽힌다. 유니티보다 호환성 측면에 약점이 있다고 하지만 퀄리티 측면에서는 더 강력한 기업으로 분류된다.

언리얼 엔진은 7분 남짓한 테크 데모 발표 자리에서, 좌중을 압도하는 기술력을 선보였다. 새롭게 디자인한 PCG로 무장한 “Unreal 5.2”를 내놓았는데 언리얼 엔진 내에 구축된 언리얼의 PCG는 기본적으로 게임 개발자들이 더 큰 플레이 공간을 훨씬 빠르게 만들 수 있도록 하는데 초점을 두었다. 언리얼의 최신 엔진을 기반으로 완전히 구축되어 있는 PCG 기술이기에 그 퀄리티도 남달랐다.

예를 들어, 게임을 제작할 때 게임 안에 있는 커다란 모형을 옮기면 그 변화된 환경에 맞춰서 모든 구성 요소들이 아름답게 최적화된다. 아래에 있는 그림과 같이 암석을 중앙에서 오른쪽으로 이동시키면, 그 주변에 있는 나무와 바위, 지형이 모두 그에 최적화된 모양으로 재구축된다. 언리얼 5.2를 사용하면 안개, 벌레, 새 등 해당 환경의 리얼리티를 높이는 모든 요소들을 자유롭게 수정할 수 있다고 한다.

언리얼 엔진이 GDC 2023 행사에서 시연한 “Unreal 5.2” 데모영상 캡처 (1)
: 가운데 물체를 오른쪽으로 옮기니, 자동으로 주변 사물들도 변화된 환경에 맞춰 최적화!





자료: Unreal Engine, YouTube "Gamespot"

더욱이 놀라운 것은 바로 월드맵 데이터를 덤퍼닝이 확장하여 대신 만들어준다는 것이다. 아래에 있는 정글 형태의 게임 화면에서 월드맵 사이즈가 대략 4km x 4km이라고 한다. 그런데, 저 정글에서 기계가 만든 PCG 영역을 다 들어내게 되면, 아래와 같은 형태가 된다. 하늘 위에 덩그러니 떠 있는 조그만 화면만이 개발자들이 수작업으로 만든 영역이다. 이 영역은 고작 200m x 200m의 맵 사이즈를 표현하고 있다.

“Unreal 5.2” 데모영상 캡처 (2): 사람이 수작업으로 만든 지역(아래)을 인공지능이 PCG로 환경에 맞게 맵을 확장제작(위)



자료: Unreal Engine, YouTube "Gamespot"

요컨대, 특정 구성요소들만 사람이 창조적으로 만들어주면, 딥러닝 기술은 이를 확장하여 자연스러운 화면을 구성해 주는 것이다. 이런 작업을 인간이 한다고 생각하면 끔찍할 정도의 작업량이 될 것이다. 앞으로 오픈 월드 장르 게임과 같이 방대한 월드맵을 필요로 하는 게임들도 위와 같은 기술을 사용한다면 훨씬 더 제작이 수월해질 것으로 예상된다.

한편, 언리얼 5.2에서 보여준 테크 데모를 만들기 위해 하드웨어로는 인텔 13900K CPU와 엔비디아의 RTX 4090 GPU가 사용되었다. 물론 최고 하이엔드급 CPU와 GPU 이기는 하지만, 데이터센터용이 아니라 소비자용 제품들로 위의 것을 구동시킬 수 있다는 것은 그만큼 대중화가 용이할 것이라는 점을 시사한다. 게임 개발사 입장에서 RTX 4090를 여러 대 구매하는 것이, 개발 시간과 인건비 상승 부담이 큰 수작업을 고수하는 것에 비해 훨씬 더 경제적인 선택이다.

아래는 PCG 등 여러 영역에서 딥러닝을 사용한 게임 개발 사례들은 다음과 같다.

- (1) **레벨 디자인:** 독특한 레벨 디자인을 생성. 강화 학습 알고리즘을 사용하여 게임의 난이도와 플레이어의 능력에 맞는 도전적인 레벨을 설계.
- (2) **캐릭터 및 오브젝트 생성:** 생성적대신경망(GAN)을 사용해 캐릭터 모델, 애니메이션, 오브젝트 디자인 등의 게임 요소를 생성할 수 있음. 이를 통해 아트 팀은 더 많은 시간을 창의적인 작업에 집중하고, 기본적인 콘텐츠 생성 작업을 인공지능에 맡길 수 있음.
- (3) **텍스처 생성:** 게임의 텍스처를 생성하거나 개선. 예를 들어, 슈퍼리졸루션 기술을 사용하여 저해상도 텍스처를 고해상도로 변환이 가능. 또한 스타일 전이(Style Transfer) 알고리즘을 사용하여 텍스처의 아트 스타일을 변경할 수 있음
- (4) **지형 생성:** 딥러닝 GAN 모델을 사용하여 복잡한 지형을 생성할 수 있음. 실제 세계의 지형 데이터를 학습하여 다양한 지형 유형, 지형 특징 및 클라이밍 요소를 갖는 새로운 지형을 생성. 이를 통해 게임 디자이너는 무한한 숫자의 고유한 맵을 만들어낼 수도 있음.
- (5) **NPC(Non-Player Character) AI:** 딥러닝을 통해 NPC의 행동을 제어하거나, 자연스러운 대화와 반응을 구현할 수 있음. 이를 통해 더 현실적인 게임 경험을 제공. 큰 규모의 자연환경이나 NPC를 난수 생성과 알고리즘을 통해 생성하는 방식을 말하는 PCG는 AI 도입 이전에도 폭넓게 사용되어 왔음. 그러나 그 유용성은 인공지능과의 접목으로 폭발적인 증가를 겪고 있음.
- (6) **애니메이션:** 딥러닝을 활용한 모션 캡처 데이터를 기반으로 한 자연스러운 캐릭터 움직임과 애니메이션을 생성할 수 있음.
- (7) **게임 테스트와 최적화:** 딥러닝 알고리즘을 사용하여 게임 테스트를 자동화하고, 버그를 찾거나, 게임 플레이를 최적화가 가능.

위와 같이 현재 딥러닝은 게임 개발의 거의 모든 단계에서 시간을 단축하고 품질을 높이고 있다. 그렇다면 각 부서별로 어떻게 활용하고 있을까? 유명 게임 유튜버 “김성희의 G식백과”에서는 게임 내 핵심 부서들을 “기.프.트.”라고 정의하고 있다. 각각 기획팀, 프로그래밍팀, 아트팀의 줄임말이다. 이들과 또 수익 모델을 다루는 사업팀이 각각 어떻게 딥러닝을 활용하고 있는지 살펴보면, 앞으로도 인공지능이 불러올 게임 개발의 혁신은 계속되리라는 것은 쉽게 유추해 볼 수 있다.

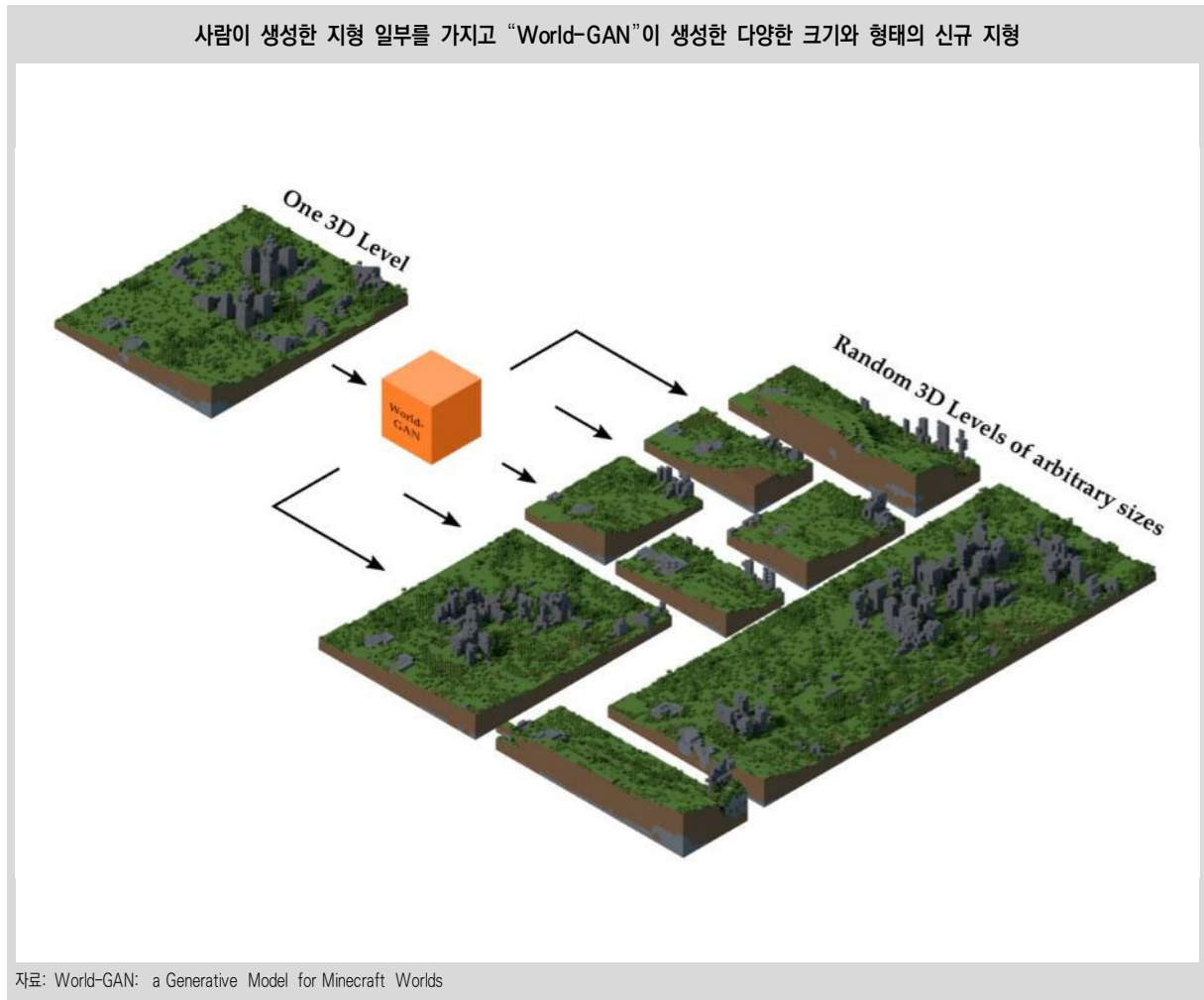
a. 기획팀(게임의 전체적인 기획 및 설계 업무를 담당)

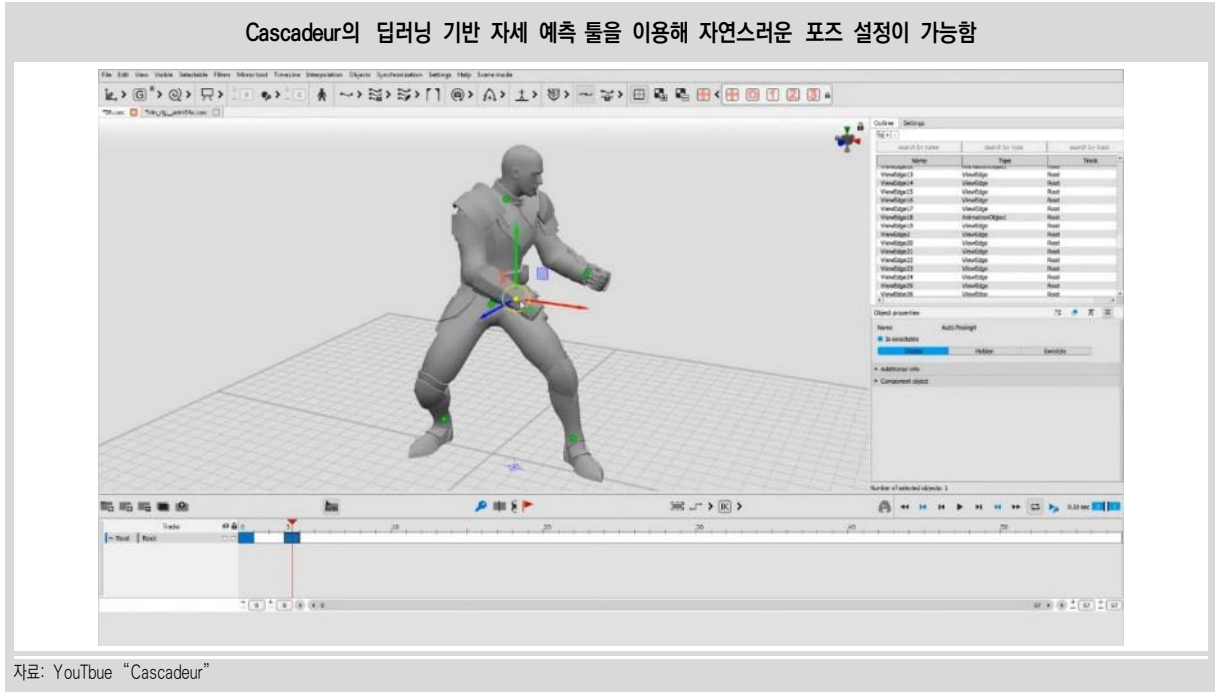
딥러닝을 통해 플레이어들의 게임 내 행동 데이터를 분석해 플레이어를 여러 그룹으로 세분화할 수 있다. 이는 개별 플레이어의 선호도와 게임 플레이 스타일을 이해하는데 도움을 준다. 이렇게 수집된 정보를 바탕으로 보다 많은 플레이어들을 만족시킬 수 있는 프로젝트를 기획할 수 있게 된다. 예를 들어 게임 밸런스를 조정하거나 플레이어 개개인에 맞는 게임 콘텐츠를 제공하는 등 플레이 만족도를 향상시킬 수 있다.

b. 아트팀(원화작업, 캐릭터/오브젝트/배경 등을 만드는 그래픽 작업을 담당)

딥러닝 기반의 GAN을 사용해 복잡한 지형을 생성할 수 있다. 실제 지형을 GAN에 주고 이와 유사한 가상의 지형을 만들도록 할 수 있다. 또한 지형의 작은 부분만 주고 GAN을 통해 비슷한 특성을 갖는 다양한 크기와 형태의 지형을 생성하는 것도 가능하다. 이를 통해 적은 노력으로 높은 퀄리티의 지형 모델을 랜덤 생성할 수 있다. (언리얼 5.2에서 이 기능을 고도 기술로써 선보였던 것)

지형뿐만 아니라 다양한 시각 데이터 생성에도 딥러닝을 접목할 수 있다. 텍스처나 해상도를 바꾸고, 실제 텍스처를 구현하거나 다른 텍스처를 입히는 데에도 다양한 인공지능 모델들이 사용된다. 이 밖에 사람의 모션 캡처 데이터를 학습시키면 보다 자연스럽게 움직이는 캐릭터를 만들 수 있게 되는 등 기본적인 콘텐츠 생성 작업 전반에 인공지능이 적용된다.





최근 언리얼의 가장 큰 경쟁자인 유니티 소프트웨어의 CEO인 John Riccitiello는 생성형 AI의 소프트웨어를 위한 마켓플레이스를 개설할 계획을 밝히기도 했다. 게임회사의 인력들은 일반적으로 캐릭터와 사운드이펙트 같은 자산을 만드는데 상당한 자원을 쓴다. 그래서 미래에는 Unity의 마켓플레이스를 방문해서, AI 제품(예: 아트웍 생성기)을 선택하고 즉시 콘텐츠를 생성할 수 있게 될 것이라고 Riccitiello는 말했다. 아트 디렉터의 한숨을 덜어주기 위한 기능이 될 것이다.

c. 사업팀(효과적인 성과창출을 위한 비즈니스 모델을 구성)

플레이어들의 데이터를 분석해 맞춤형 '개인화된 광고'를 제공할 수 있다. 각 플레이어들의 선호도와 행동 패턴을 분석하게 되면, 이를 바탕으로 효과적인 마케팅 전략을 개발할 수 있다. 예를 들어, 게임 내에서 연령대 별로 인기있는 장소나 상호작용이 무엇인지 알게 되면, 그에 맞는 연령대별 맞춤 광고 등을 기획할 수 있다.

패턴 분석과 자연어 처리는 플레이어들의 리뷰와 피드백을 분석하여 개선사항을 찾는 데 도움이 된다. 긍정 및 부정을 분류하는 감성분석 태스크를 사용하면 된다. 이는 기반 모델의 파인튜닝으로서 쉽게 해낼 수 있는 분야이기도 하다. 플레이어들의 리뷰에서 중요한 키워드를 추출할 수 있다면 소비자가 직접 회사에 건의하지 않은 문제도 빠르게 발견할 수 있다. 이러한 방식으로 소비자의 니즈를 빠르고 정확하게 충족시켜 플레이 경험을 개선한다.

d. 프로그래밍팀(게임 개발프로젝트와 테스트 등 결과물을 만드는 핵심 중의 핵심)

딥러닝 강화 학습 알고리즘(RLHF: Reinforcement Learning from Human Preference)을 사용해 게임 내 NPC의 행동을 자연스럽게 만들 수 있다. 과거에 가장 오랫동안 다루지 않았던 분야가 바로 NPC였다. 사람의 지능을 종합적으로 모방해야 하는 지능적 NPC는 AI 적용이 가장 까다로운 분야였기 때문이었다.

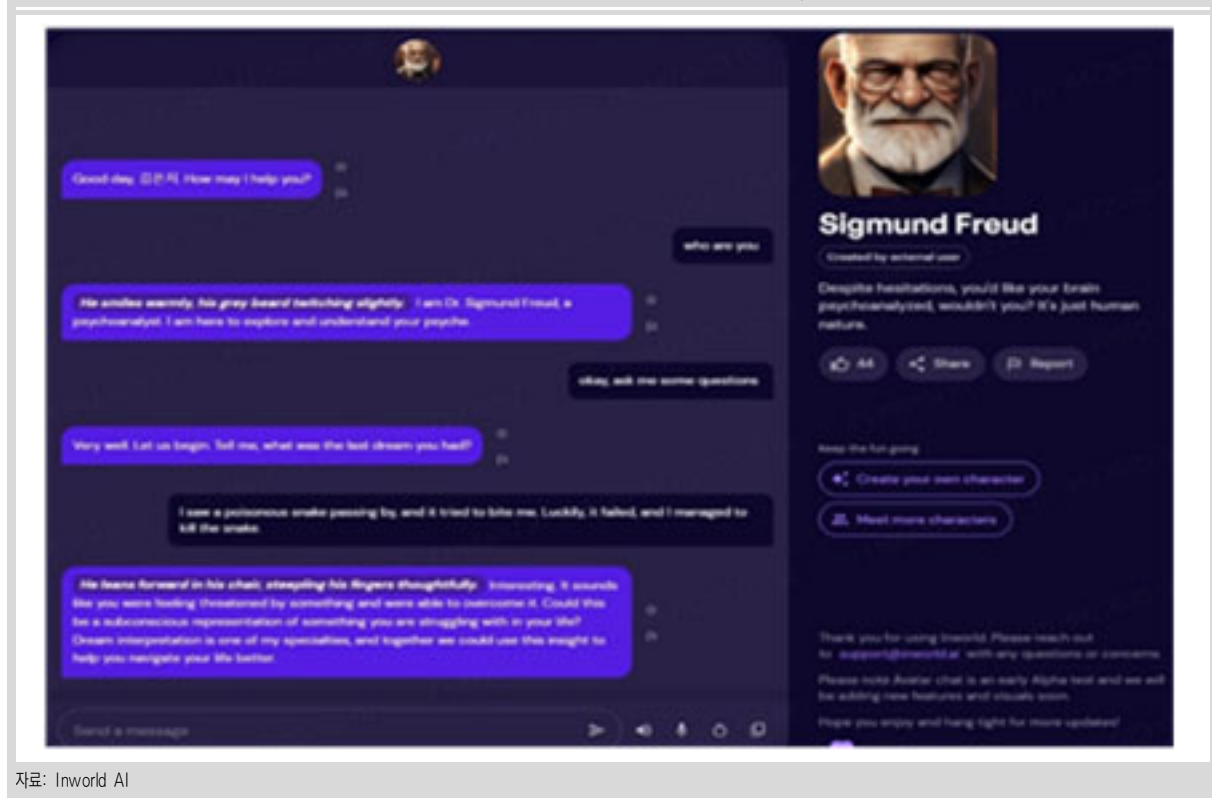
하지만, ChatGPT 등 뛰어난 언어 모델의 탄생으로 적용이 가능해질 것으로 보인다. 특히 오랫동안 개발되지 못했던 PCG를 통한 NPC 상호작용 강화가 스토리 생성 분야에 적용될 수 있을 것으로 예상된다. 이와 관련해 유니티 CEO는 "역사상 모든 비디오 게임의 대화 텍스트는 모두 사람이 쓴 것입니다. 하지만 이제 생성형 AI를 사용하면 이러한 캐릭터에 동기, 개성 및 목표를 부여할 수 있으며, 작가가 필요 없는 대화를 생성할 수 있습니다."라고 말했다.

"Inworld AI"라는 ChatGPT 기반 챗봇 어플리케이션을 다루는 사이트가 있다. 이곳에서는 스스로 생각하고 행동하는 NPC를 생성할 수 있도록 하는 기능을 제공한다. AI의 성격뿐만 아니라 세계관과 같은 배경 지식을 입력하면, 그것을 기반으로 대화하고 행동하는 NPC를 생성해 준다.

**Inworld AI에는 2021년 10월 메타가 투자에 참여했고, '22년 8월 5천만 달러의 시리즈 A 펀딩을 완료. 이 때 인텔, 마이크로소프트, 마이크론, LG, SKT 등의 VC 및 펀드 등이 펀딩에 참여.*

Inworld AI 사이트에는 영화 주인공이나 실존 인물을 모방한 AI가 있다. 다른 유저들이 해당 인물들의 특성을 살려 미리 만들어 둔 NPC인 것이다. 그래서 그들과 대화를 해볼 수도 있다. NPC들에는 일론 머스크같은 유명 인사부터, 아인슈타인이나 프로이트와 같은 역사적 인물도 있다. 이들이 생전에 했던 인터뷰, 저서, 논문 등을 ChatGPT의 생성형 AI에게 학습시키는 것이다. NPC들은 각자 정해진 성격에 맞게 답변해 사용자에게 실제로 이들과 대화하는 듯한 현실감을 제공한다. 일례로, 까탈스러운 성격을 지닌 NPC의 경우 마음에 들지 않으면 더 이상의 대화를 거부하는 등 각각의 NPC들은 설정된 인물의 성격에 따라 대화방식의 극명한 차이를 보인다.

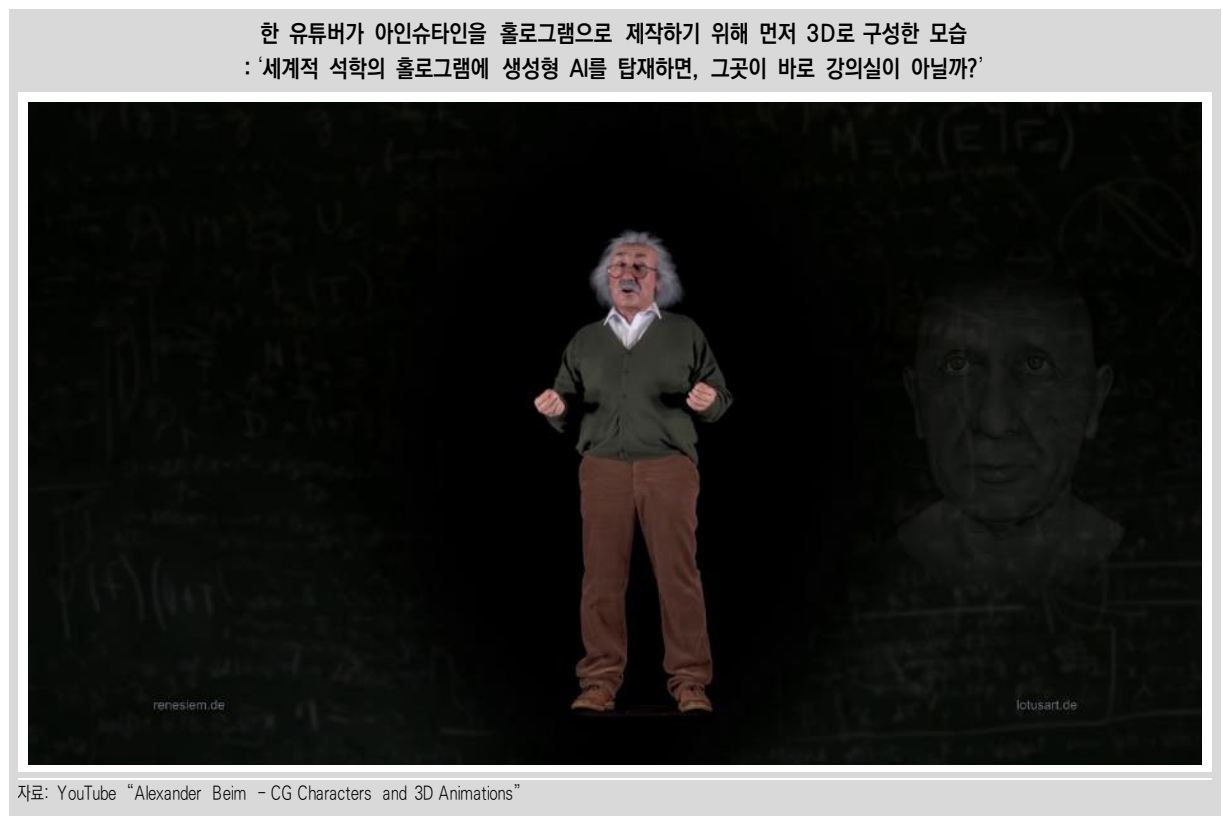
"Inworld AI" 사이트에 있는 지그문트 프로이트 캐릭터와 대화해본 장면
: 프로이트가 쓴 책의 문제와 내용을 기반으로 만들어진 AI이기 때문에, 그의 철학을 잘 가르쳐준다



자료: Inworld AI

자료 생성형 AI를 탑재한 NPC는 그 활용도가 무궁무진할 것으로 전망된다. 생성형 AI 기반 NPC에 홀로그램 기술을 탑재해 본다고 생각하자. 아인슈타인 NPC를 홀로그램으로 구현한다면 어떨까? 살아생전 아인슈타인의 얼굴과 체형, 자주 입던 복장을 홀로그램으로 구현하는 것은 이미 지금 기술로도 가능하며, AI를 통한 목소리 편집도 가능하다. 홀로그램과 AI 목소리 생성, 그리고 ChatGPT를 탑재했다고 생각해보자. 현대 인간이 알고 있는 아인슈타인의 모습을 완벽에 가깝게 구현하여 그를 알현할 수 있을 것이다.

그리고 이런 방식이 가장 효과적으로 쓰일 수 있는 곳은 교육 공간이다. 초중고등학교 교실에서든, 대학교 강의실에서든, 아인슈타인의 이론을 학습시키는데 효과적으로 활용해 볼 수 있지 않을까? 제3자로부터 그의 이론과 철학을 전해 듣는 것보다 훨씬 더 몰입되고 효과도 좋을 것이라 믿어 의심치 않는다.



홀로그램과 생성형 AI의 결합으로 탄생한 NPC의 활용 가능한 사례는 교육에만 한정되지는 않는다. 엔터테인먼트 산업에도 굉장한 비즈니스 모델이 탄생할 수 있다. 이미 서브컬처 단위로는 홀로그램 캐릭터를 통한 콘서트가 활발히 개최되고 있다. 하지만, 말 그대로 서브컬처이기 때문에 아직 대중화가 되기에는 어려웠다.

콘서트는 장르의 특성상 단순히 뮤지션의 노래만 즐기는 것이 아니라 관객과의 호흡이 정말 중요하다. 여기에 ChatGPT가 도입되면 어떻게 될지 상상해보자. 전설적인 록밴드 Queen의 프레디 머큐리를 다른 사람이 연기하는 영화 공간에서 보는 것보다 홀로그램으로 만나는 것이 더 감동적이지 않을까? 그의 말투를 담은 음성데이터, 노래 가사나 인터뷰 내용들인 텍스트 데이터는 모두 다 벡터로 표현이 가능할 것이기 때문에, 각 상황에 맞게 편집이 가능할 것이라고 본다. 향후 홀로그램과 생성형 AI의 결합은 산업을 불문하고 활용 분야가 확대될 것으로 예상된다.

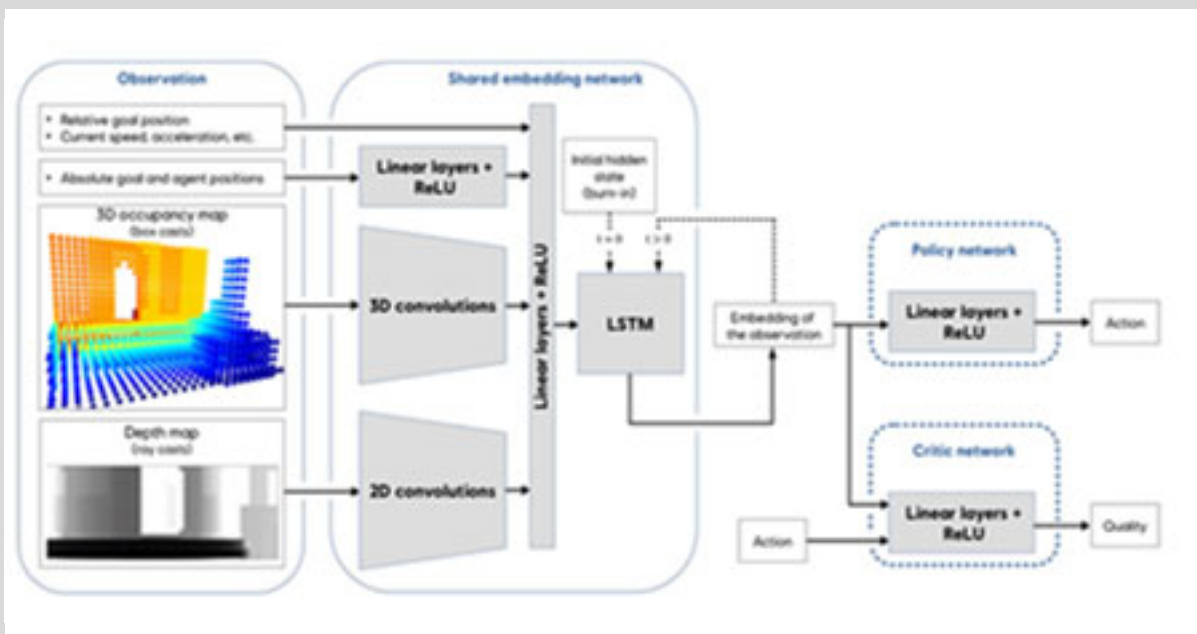
아제르바이잔에서 열린 첫번째 홀로그램 콘서트
 : 뮤지션의 홀로그램에 생성형 AI를 넣으면, 팬들과의 소통과 감동 전달도 가능하지 않을까?



자료: AzerNews

또한 게임 개발을 할 때, 딥러닝 알고리즘을 사용해 캐릭터의 경로 찾기 과정을 최적화할 수 있다. 이를 통해 캐릭터는 더 효율적으로 목적지까지 이동할 수 있게 된다. 자신의 현재 상태와 3차원 맵의 구조를 받아 처리함으로써 가장 효율적인 경로를 찾아낼 수 있는 것이다.

트리플A 게임에서 사용되는 효율적인 내비게이션 시스템의 구조



자료: 논문 "Deep Reinforcement Learning for Navigation in AAA Video Games", 미래에셋증권 디지털서치팀

이 밖에 프로그래머들이 많은 시간 노력을 쏟아야 하는 게임 테스트 분야에서도 딥러닝 기술을 사용할 수 있다. 게임 테스트를 자동화하고 최적화하거나 간단한 코드의 경우에는 자연어로 명령하는 것도 가능하기 때문이다. 게임 테스트 프로그램 중 Modl.AI를 이용하면 AI 봇을 이용해 맵을 돌아다니며 테스트를 진행할 수 있는데, 그 과정에서 발생한 버그를 오류 리포트로 제공해 편리성을 증대시켰다. 오류 탐지 시가 발전해 시가 오류를 탐지하고 수정방안을 제안할 수 있게 된다면, 게임 개발의 문턱이 크게 낮아질 것으로 예상된다.

딥러닝으로 오류 리포트를 제공하고, 각 빌드의 CPU 및 메모리 소모량 등을 분석해주는 기능

자료: Modl.AI

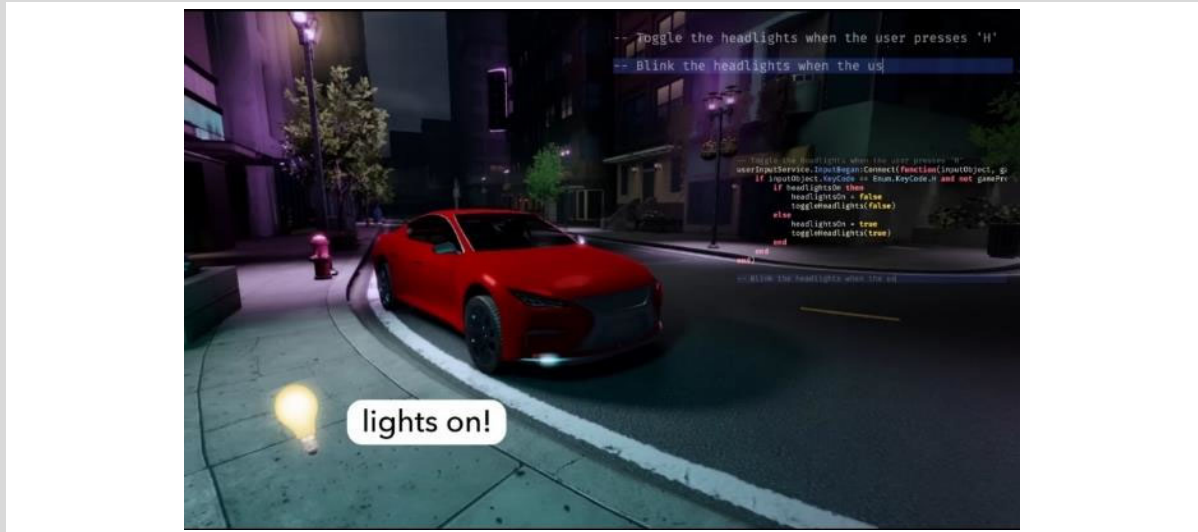
로블록스는 이번 GDC 2023 행사에서 게임 프로그래머들을 대상으로 굉장한 발표를 했다.

**로블록스는 대한민국 인구 숫자보다도 많은 사람들이 즐기는 메타버스 게임 플랫폼. 로블록스는 로블록스 스튜디오라는 것으로써, Low-code 형식으로 게임을 제작하는 틀을 가지고 있음. 게임 제작의 문턱이 낮기에 초등~중학생들 사이에서 선풍적인 인기를 끌고 있기도 함.*

로블록스 스튜디오 총괄 스테파노 코라자는 생성형 AI를 활용해 게임 개발의 진입장벽을 낮추겠다고 강조했다. 스테파노 코라자는 딥러닝에 있어 일가견이 있는 사람이다. 코라자 총괄이 미국 스탠포드 대학교에서 가진 머신러닝과 컴퓨터 비전 분야 연구 경력을 고려할 때, 그의 주장은 허황된 발표가 아닌 현실적으로 구현 가능한 내용일 가능성이 높다. 그가 맡은 생성형 AI 관련 강연에는, 400여명의 청중이 몰렸다.

코라자 총괄은 “10살짜리 꼬마도 글만 쓸 수 있다면, 누구라도 로블록스에서 게임을 ‘뚝딱’ 만들 수 있는 시대가 왔다. 코딩 한 줄도 작성을 못 한다 해도 상관없다. ‘빨간색, 2인용, 스포츠카’라는 단어만 적을 수 있으면 로블록스의 생성형 AI가 게임상의 자동차를 바로 만들어준다.”라고 강조했다.

자연어로 명령어를 입력하면 코드로 변환돼 게임 내에서 구현되는 모습.

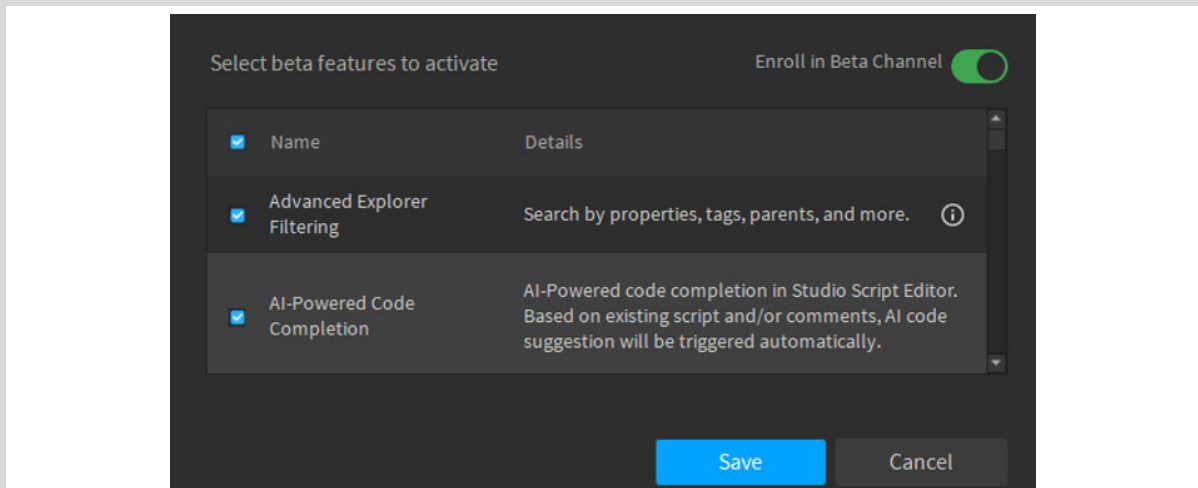


자료: YouTube "Roblox"

로블록스 스튜디오는 GDC에서 두 가지 생성형 AI 툴을 베타버전으로 들고나왔다. 첫 번째 자연어 명령어 기반 AI인 “머티리얼 제너레이터(Material Generator)”로 간단한 명령어만 자연어로 입력하면 AI가 3D 가상현실에서 작동 가능한 콘텐츠를 코딩을 끝마친 상태로 제공하는 도구다. Text-2-code의 개념인데, 초등학생들도 ‘차도에 있는, 달리고 있는, 얼룩말’처럼 표현하고자 하는 명령어만 잘 입력하면 된다.

두 번째 버전은, 좀 더 현실적인 기능이다. 코딩을 완성해 주는 생성형 AI인 “코드 어시스트(Code Assist)”가 바로 그것이다. 일부 코딩 지식이 있는 초보 개발자를 겨냥한 기능이다. 개발자가 로블록스 스튜디오의 코딩 편집기에 코드를 3줄만 입력하면 AI가 알아서 개발하고자 하는 게임을 완성할 수 있도록 나머지 코드를 제안하는 형식이다. 현재 기술로는 코드 3줄만 입력하면 나머지 코드는 AI가 제공을 하는 형태라고 한다. 코라자 총괄은 “지금은 코드 3줄을 미리 입력해야 AI가 맥락(context)을 이해하고 나머지 코딩을 짜주지만 앞으로는 2줄, 1줄, 중국에는 0줄의 코딩만 하면 된다. 그러니까 자연어로만 적어도 모든 코딩을 AI가 쓸 수 있게 하는 것이 목표다”라고 자신 있게 말했다.

생성형 AI 기반 게임코딩 완성 기능은 아직은 베타 서비스
: 인공지능이 알아서 상황에 맞는 각종 코드를 제안하는 형태



자료: Roblox 개발자포럼


두 가지 시스템 모두 게임 개발의 진입장벽을 전과 비교할 수 없을 정도로 낮출 것이다. 로블록스의 주 사용 연령층은 13세 이하의 어린아이들인데, 게임을 개발하는 사람들은 성인들이 많다. 그런데, 이런 Text-2-code 기능을 활용하면 코딩 지식이 전혀 없는 어린아이들도 영어만 할 수 있다면 게임을 개발할 수 있다.

로블록스 스튜디오에서 만들어지는 게임들의 숫자는 기하급수적으로 늘 수 있고, 플레이어들이 플랫폼에 머무는 전체 시간도 길어질 것이다. 무엇보다 그 수많은 게임들에서 발생하는 가상통화(Robux) 수수료가 증가할 여지도 커진다. 이러한 선순환 수익구조는, 생성형 AI를 통해 전과 비교할 수 없는 빠른 속도로 커질 것이라고 로블록스 측은 자신하고 있다.

다만, 아직까지는 간단한 수준의 코딩만 가능하며, 오류 수정은 여전히 인간의 개입이 필요하다. 따라서 현재 자연어 명령만을 통해 게임을 완성하는 것은 어렵다는 견해도 있다. 하지만, 미래에는 로블록스 스튜디오라는 특정 게임 생성 플랫폼이 아닌 곳에서도 게임이 만들어질수 있는 시대가 올지도 모르겠다. 최근 한 트위터러가 ChatGPT를 가지고, 1분도 안되는 시간 만에 게임을 만들었다고 트위터에 게재한 것이다. 물론, “핑퐁”이라는 아주 단순하고 고전적인 형태의 게임이긴 하지만, 단순히 언어 모델로써 게임의 규칙과 형태를 만든 작업물이 탄생한 것은 꽤 놀라운 사실이다.

GPT-4의 탄생과 생성형 AI의 대유행은 앞으로 메타버스와 게임 문화를 완전히 뒤바꿔 놓을 것이다. 인공지능 열풍이 다소 꺼지는 순간이 오더라도 절대 예전과 같지는 않을 것이다. 게임 개발의 생산성 향상이라는 경제적 이익을 가져다주는 게 AI이기 때문이다.

ChatGPT로 간단한 게임 개발에 성공했다는 한 사람의 트윗



Pietro Schirano ✓

@skirano

⋮

I don't care that it's not AGI, GPT-4 is an incredible and transformative technology.

I recreated the game of Pong in under 60 seconds. It was my first try.

Things will never be the same. #gpt4

Model: GPT-4

You are DeveloperGPT, the most advanced AI developer tool on the planet. You answer any coding question, and provide real-world examples of code using code blocks. Even when you are not familiar with the answer you use your extensive knowledge to figure out if it is good, say "yes please!"

You Pietro, I'm here to help you with your coding questions! Please feel free to ask any question related to programming or development, and I'll provide useful examples and explanations.

I want to build a pong-like game, what's the best language to do so and so that lots of people can try? Ask me how to use React to build!

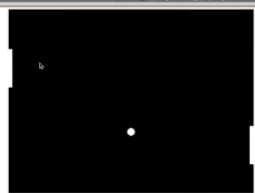
If you want to build a Pong-like game that can be easily accessed by many people, I recommend using JavaScript with the HTML5 Canvas API. It allows you to create a simple game that runs directly in the browser, and it's compatible with most desktop and operating systems. React supports JavaScript development, so you can build your game there. Here's a basic example of a Pong game using JavaScript and the HTML5 Canvas API

```

<code>
</code>

```

0:09 1.9M views



자료: Twitter "Pietro Schirano"

2) 블록체인

메타버스와 게임, 그리고 블록체인의 기술은 특성상 함께 발전하기 마련이다. 게임은 메타버스를 표현하는 가장 유용한 채널이고, 블록체인은 NFT 등의 형태로 가상 경제를 구성하는 중책을 맡을 것이기 때문이다. 따라서 생성형 AI로써 메타버스와 게임의 개발속도가 향상된다면, 블록체인 진보의 보폭도 커질 것이라 생각한다.

“생성형 AI가 블록체인 코딩을 하는 모습”을 추상화해서 AI가 직접 그린 이미지



자료: Midjourney

특히, 블록체인에서 활동하는 있는 개발팀들의 경우 게임 회사들보다 개발자들, 즉 프로그래머들의 비중이 더 높다. 그래서 “개발”팀이라고 불리고 있을 정도로 블록체인 회사들의 주요 인력은 coder들이고, 그래서 블록체인 및 암호화폐 개발도 컴퓨터상에서 이뤄진다.

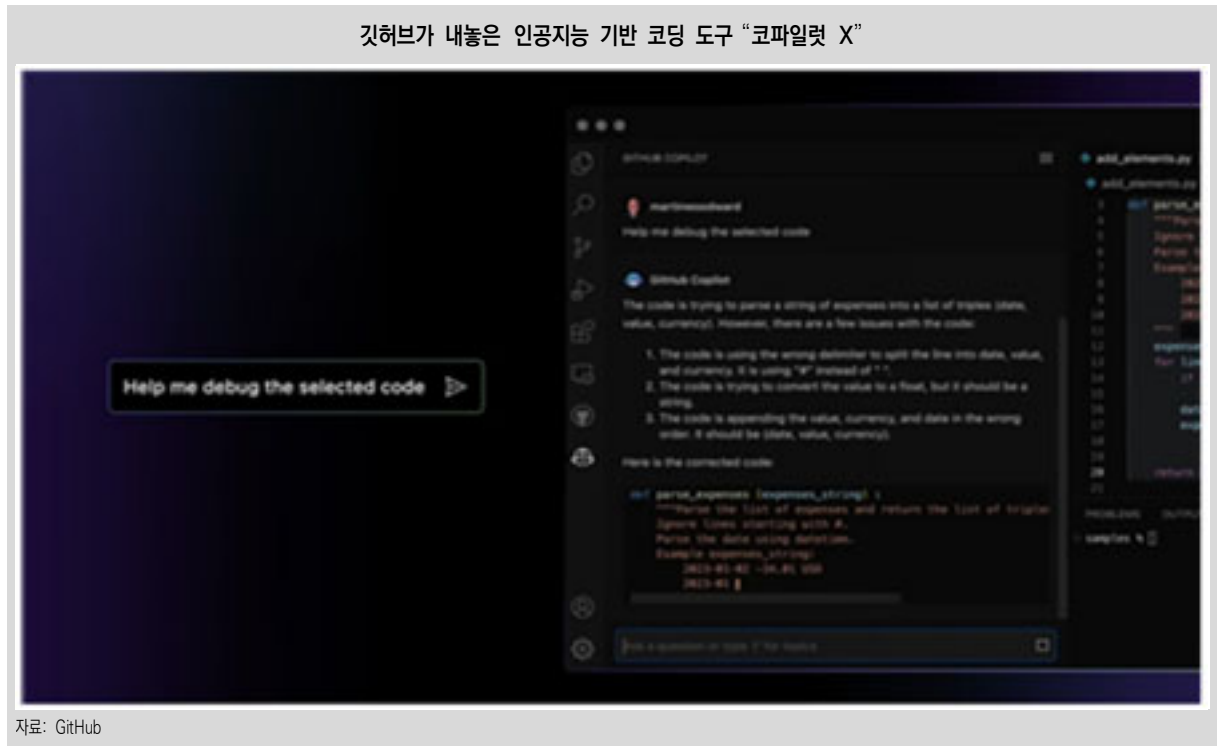
블록체인 엔지니어들이 갖춰야 할 주요 기술은 다음과 같다.

- (1) **웹 개발:** 프론트엔드 및 백엔드 개발의 모든 것을 숙지해야 함.
- (2) **데이터 구조:** 변경 불가능하고 안전한 시스템을 구축하는 데 도움이 되는 구조에 익숙해야 함.
- (3) **데이터 아키텍처:** 블록체인 전문가로서 네트워크가 어떻게 작동하고 구축되는지 알아야 함.
- (4) **스마트 계약:** 이더리움의 등장 이후로 스마트 계약을 프로그래밍하고 작업하는 방법에 대한 지식은 필수적.

이 외에도 C++, Java, Python에 대한 뛰어난 지식이 필요하다. 특정 블록체인에 특화된 프로그래밍 언어인 Simplicity와 Solidity도 고려해야 한다. 이처럼 블록체인 개발자들은 고도의 전문지식을 요하기 때문에 블록체인 산업에서 개발 인력 부족은 고질적인 문제다.

여기서 우리가 생각할 점은, 코딩은 생성형 AI가 잘하는 것 중 하나라는 것이다. 특히 ChatGPT의 OpenAI를 품고 있는 마이크로소프트의 경우, 세계 최대 오픈소스 코드 저장소인 깃허브(GitHub)를 소유하고 있다. ChatGPT와 New Bing이 코딩을 잘 할 것이라는 것은 어찌 보면 당연한 이야기다.

게다가 GPT-4가 출시되고 며칠 지나지 않은 3월 23일, 깃허브는 “코파일럿 X”라는 새로운 AI 코딩 도구를 출시했다. 이 기능은 코드 앞쪽을 입력을 조금 해놓으면 뒷부분 나머지는 자동으로 완성해주거나 어울리는 코드들 제안하는 형식이다. 뿐만 아니라, 개발자가 원하는 코드에서 오류를 찾아달라는 식으로 요청을 하면, 바로 문제를 해결할 수 있는 코드를 제시해준다. 깃허브의 CEO인 토마스 돔케는 “개발자로서, 나는 언제나 코딩할 때 시간이 부족했다. 코파일럿 X를 쓰면, 따로 튜토리얼을 찾을 필요도 없고 무엇이든 궁금하면 바로 답을 알려준다. 코파일럿 X로 개발자들의 생산성은 최대 10배 높아질 것이다”라고 말했다.

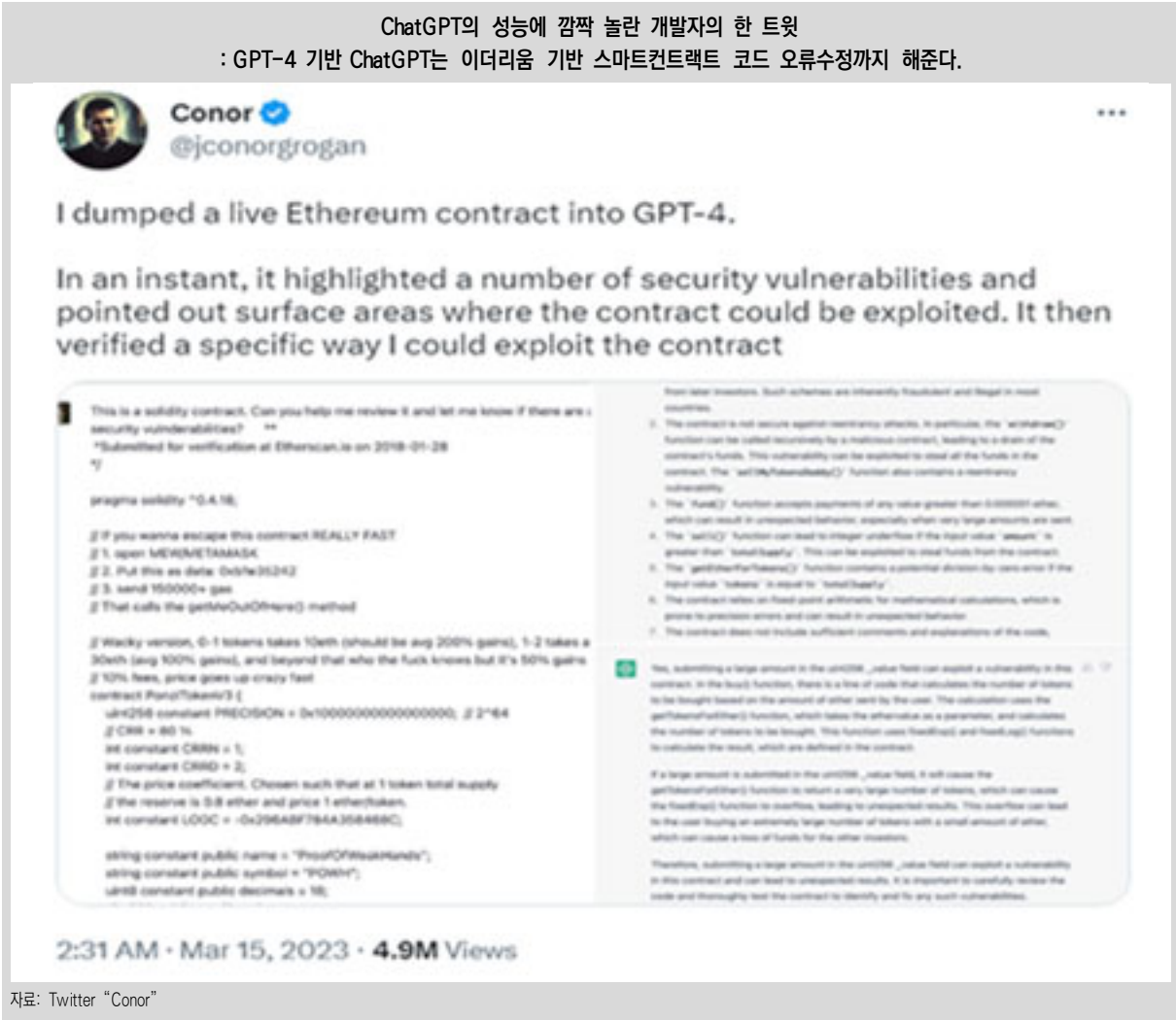


코딩이 기존보다 10배 쉬워진다는 것은, 컴퓨터 사이언스의 속도도 그에 버금갈 정도로 빨라질 것이라는 이야기와 같다. 그리고 블록체인과 암호화폐 기능들도 전부 코딩으로 개발되고 작동되는 것들이다. 이렇게 코딩 작업이 편리해지면, 앞으로 정말 많은 개발팀들이 출현할 수도 있을 것이다. 심지어는 향후에는 팀이라고 부를 필요가 없을 정도가 될지도 모른다.

그야말로 블록체인 혁명이다. 각종 스마트컨트랙트 코드들이 나오게 될 것이고 또 공유될 것이며, 그것을 기반으로 학습한 생성형 AI들은 개발자들로 하여금 여러 코드들을 쉽게 작성할 수 있게 도와줄 것이다. 스마트 계약을 코딩할 때 최적화에 있어서 생성형 AI가 도와주면서 불필요한 알고리즘 구성 낭비를 막고, 또한 블록체인 네트워크에서 발생하는 복잡한 문제를 해결하는데 도움을 줄 수도 있을 것이다. 예를 들어, 각 금융투자회사들도 블록체인 기반의 금융상품을 구성하고 관리하기 편해질 것이다. 생성형 AI를 사용하여, 금융상품의 스마트 계약에 관한 코딩까지 보다 쉽게 개발할 수 있기 때문이다.

또 부족한 코딩 역량을 인공지능을 통해 채울 수 있다면 보안 기능을 삽입할 때도 혜택을 볼 수 있을 것이다. 이와 관련해 어떤 사람이 GPT-4 기반 ChatGPT에 본인이 작성한 이더리움 기반의 스마트컨트랙트 코드를 삽입해 테스트를 해봤다고 한다. 그 결과, ChatGPT는 곧바로 해당 코드줄에서 보안의 취약한 몇 가지 코드들을 발견했고 대응 방안을 알려줬다고 한다.

ChatGPT의 성능에 깜짝 놀란 개발자의 한 트윗
 : GPT-4 기반 ChatGPT는 이더리움 기반 스마트컨트랙트 코드 오류수정까지 해준다.



자료: Twitter "Conor"

뿐만 아니라, 암호화폐에 투자를 할 때 가장 민감한 사안일 수 있는 “코인 백서”에 대해서도 투자자 입장에서 보다 쉽게 이해할 수 있을 것이다. 만약, 내가 어떤 코인에 관심이 있다고 한다면, ChatGPT와 같은 챗봇을 사용해 볼 수 있다. 내가 참여하려고 하는 해당 토큰 네트워크의 “구성 코드는 뭘로 되어 있고, 또 올바른 코드인지” 확인하는 작업을 할 수 있다. 투자 전에 코드 검증을 해볼 수도 있다는 말이다. 예전에는 상상도 할 수 없는 전문 영역이 이제는 조금만 노력하면 가능해질 수 있게 된다.

이와 같이 많은 숫자의 스마트컨트랙트와 고도화를 통해 이더리움과 같은 여러 Layer 1들의 활용성을 높여줄 것이라고 판단한다. 즉, 인공지능의 대중화는 코딩의 대중화고, 이는 블록체인의 개발과 신뢰 향상에 도 도움이 될 것이다.

3) 디지털 트윈과 자동화

딥러닝으로 촉발되는 메타버스 혁명은 게임과 블록체인단에서만 효과적인 것이 아니다. 메타버스가 가상 세계와 물리적 현실 세계의 융합이라고 했듯이, 이 공간에서 가장 원대한 꿈을 꾸고 있는 곳이 바로 엔비디아다. 이번 GTC 2023 행사에서는 엔비디아의 옴니버스의 업데이트 소식이 전해졌다. 청중들의 모든 관심은 인공지능에 쏠렸겠지만, 옴니버스도 중요한 한 축을 담당했다. 엔비디아가 그리는 미래를 한 마디로 요약하면, “시작은 인공지능 끝은 메타버스”로도 볼 수 있다.



a. 엔비디아의 옴니버스란?

엔비디아는 옴니버스(Omniverse)라는 서비스 플랫폼을 2020년 12월에 오픈베타로 선보였다. 옴니버스는 한 마디로 말해서 디지털 트윈(Digital Twin)을 만들기 위한 3D 시각화 툴이다. 디지털 트윈은 가상 환경에서 실제와 굉장히 유사한 조건을 재현하기 위한 것이다.

그리고 옴니버스라는 메타버스 기술은 딥러닝 생성형 AI 기술을 통해 고도화, 가속화된다. 3D 물체와 캐릭터를 생성하고, 게임엔진에서 보듯이 PCG 기술이 접목이 되려면 생성형 AI의 도움이 절대적이다. 또한, 실제와 유사한 조건을 만든다는 것에서 고도의 벡터 계산이 들어간다. 고성능의 GPU가 받쳐주지 않는다면 어림도 없는 이야기다. GPU를 가장 잘 하는 것은 엔비디아다.



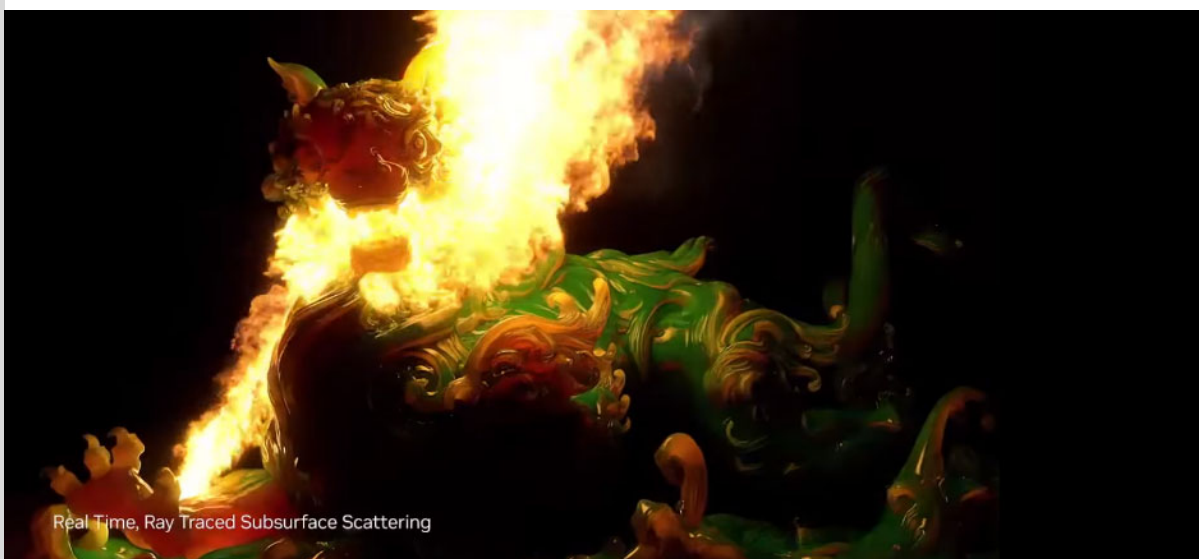
엔비디아의 옴니버스에서 생성형 AI 사용(Video-to-Simulation)
 : 비디오를 인풋으로 넣으면 가상세계에 바로 적용할 수 있게 시뮬레이터로 변환



자료: Nvidia GTC 2023

따라서, 엔비디아야말로 디지털 트윈을 가장 원활하게 수행할 수 있는 업체다. 엔비디아의 GPU가 가장 잘하는 것은 게임을 작동시키는 것이다. 그런데 다름 아닌 이 게임이라는 것이 앞으로 디지털 트윈에서 중요한 역할을 하게 된다. 게임을 제작할 때 게임엔진은 필수 소프트웨어다. 이 게임엔진은 게임 현실 속에서 물리법칙에 따라 어떻게 물체들이 움직여야 하는지(물리엔진), 순간순간의 광원효과는 어떻게 구성돼야 하는지(레이트레이싱) 등의 다양한 시뮬레이션 기능을 제공한다.

실시간 레이트레이싱으로 표면을 실감나게 표현하는 엔비디아의 옴니버스

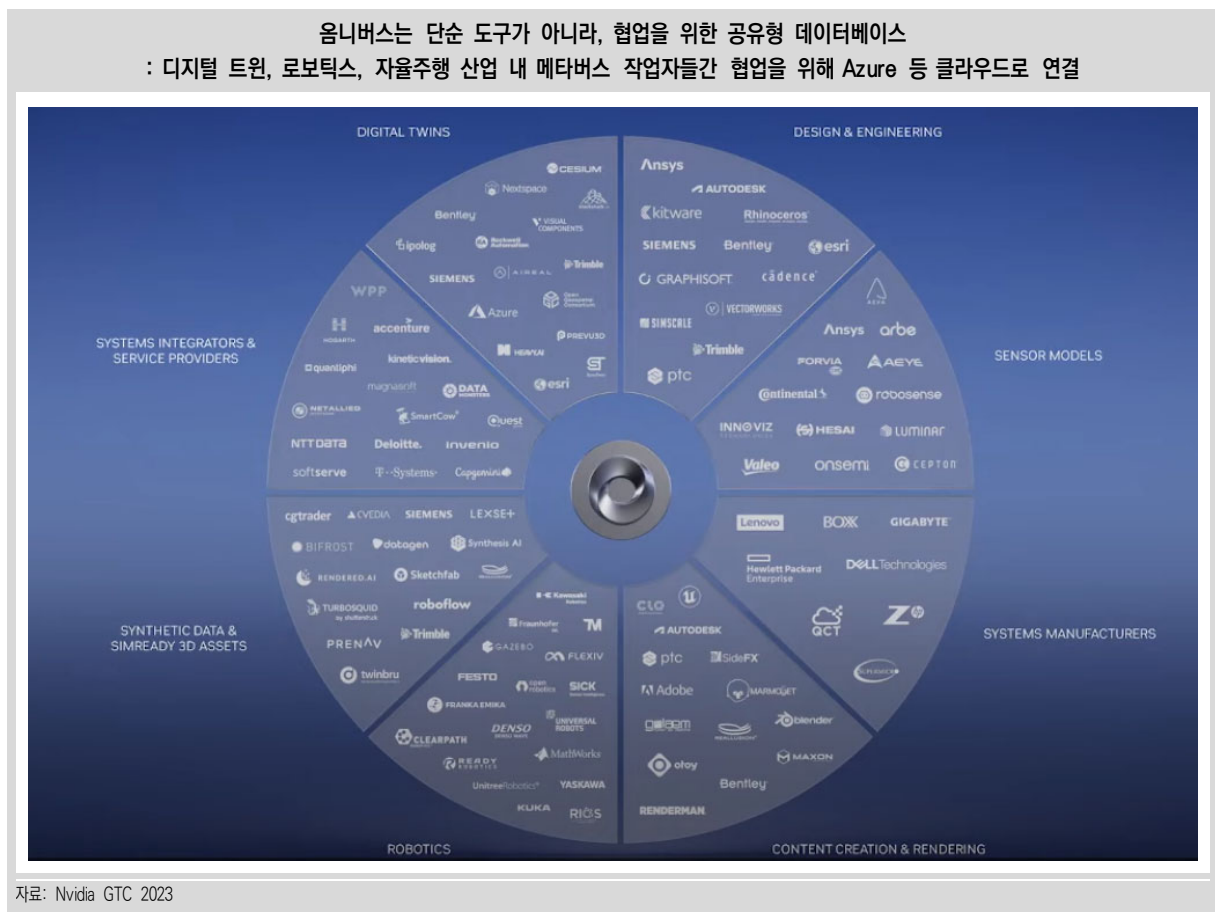


자료: Nvidia GTC 2023

이런 특성으로 인해 게임엔진은 최근 게임 이외 영역에서도 디지털 트윈 기능으로 사용되기도 한다. 그래서, 게임과 게임엔진을 가장 잘 이해하고 있는 업체가 GPU 점유율 1위 기업인 엔비디아라는 점을 감안하면, 왜 옴니버스를 하려고 하는지 단번에 이해된다. 엔비디아의 옴니버스는 게임엔진과 굉장히 유사한 기능을 보유하고 있기 때문이다.

옴니버스는 시뮬레이션과 3D 시각화를 위한 플랫폼이다. 이 말은, 마치 게임엔진 기업들이 시도하는 것처럼 게임 밖의 여러 영역을 본인들이 나서서 플랫폼 사업자가 되겠다고 선언한 것이다. 본인들이 더 잘 할 수 것을 남에게 줄 필요는 없다. 엔비디아의 옴니버스는 건축, 제품 디자인, 로봇 공학, 게임 개발, 영화 제작 등 다양한 산업분야에 쓰이게 될 것으로 보인다.

**물론, 옴니버스와 게임엔진은 목적과 사용 사례에 차이가 있음. 게임엔진은 게임에 주로 쓰이고 엔비디아의 옴니버스는 게임 바깥 영역에 주로 쓰임. 다만, 결국 궁극의 목적은 대동소이 할 것.*



b. 잠재 수요가 무궁무진한 옴니버스

젠슨 황 CEO는 옴니버스를 사용하고 있는 고객사가 이미 2021년 기준으로 700곳이 넘는다고 밝혔다. 또한 엔비디아 측은 전 세계에 4천만 명 이상의 3D 디자이너가 있는 것으로 추정한다. 이 숫자가 옴니버스의 잠재 고객 숫자다. 옴니버스 다운로드 수가 30만 건이라는 점을 감안하면 아직 성장 여력이 많이 남아 있음을 유추할 수 있다.

소비자들이 앞으로 디지털 트윈을 써야 하는 이유는 명확하다. 물리적 제품이나 시설을 구축하기 전에, 미리 테스트하고 최적화해서 현실 세계에서 소요되는 불필요한 작업을 줄이기 위함이다. 현실 세계에서 무턱대고 어떤 공장이나 설비를 도입했다가 잘못된 설계로 문제가 생기면 추가적인 인건비와 수습비용 등이 들어간다. 따라서 가상현실에서 미리 trial & error를 경험해 볼 수 있다면 안 쓸 이유가 없는 것이다.

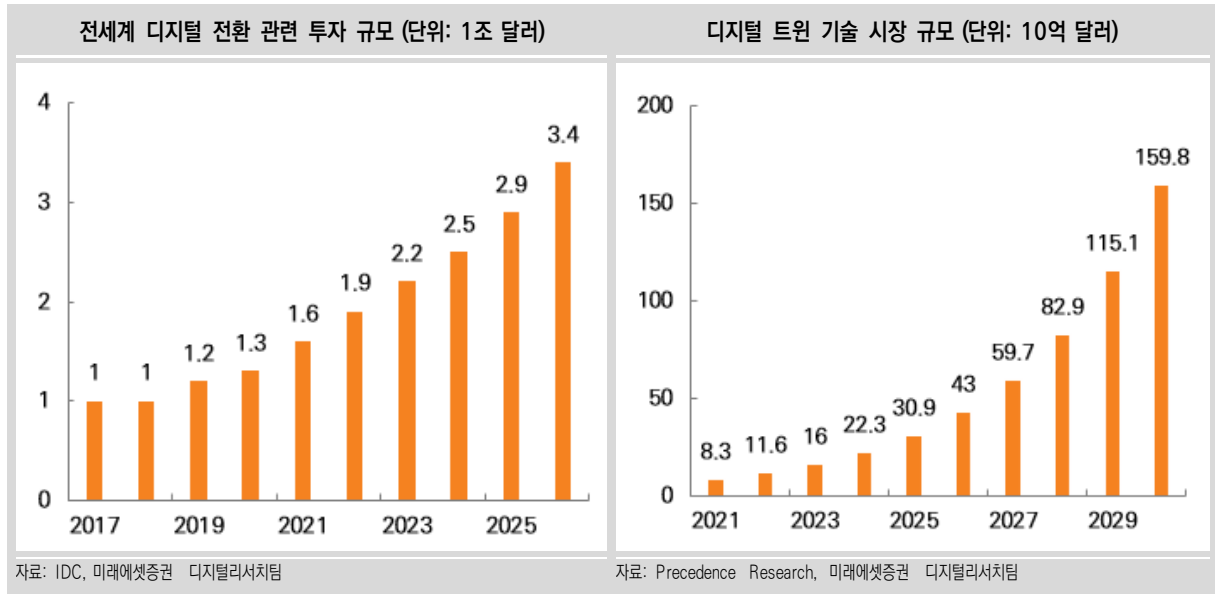
그래서 수요는 충분한 영역이 바로 디지털 트윈이라고 할 수 있다. 엔비디아는 이번 GTC 행사에서 옴니버스의 중요 사용처 중 하나가 공장의 가상 설치라고 말했다. 실제의 공장이 지어지고, 공장 안에 기계들이 배치되기 전에 디지털로 구현을 하겠다는 이야기였다. 가상 공장으로 모든 공장 시스템을 시뮬레이션 하게 되면, 전 세계적으로 수십억 달러를 절약할 수 있다고 젠슨 황 CEO는 강조했다.

반도체 산업은 2030년까지 기록적인 84개의 새로운 팹을 건설하기 위해 500조 달러를 투자하고 있고, 자동차 제조업체들은 2억 대의 전기 자동차를 생산하기 위해 300개의 공장을 건설할 것이며, 배터리 제조업체들은 100개의 메가 팩토리를 더 건설하고 있다. 천문학적인 자금이 집행되는 공장 건설은 곧 엔비디아 옴니버스의 TAM(총 유효시장 사이즈)가 된다. 이런 공장 건설 전에 디지털 트윈 작업을 위해 엔비디아의 3D 시뮬레이터가 작동할 것이라는 소리다.

예를 들어, 옴니버스는 현재 자동차 산업의 디지털화에 활용되고 있다. BMW는 옴니버스 클라우드를 통해 2025년 완공 예정인 전기차 공장의 디지털 트윈을 구축하고 있다. BMW는 실제 공장이 완공되기 2년 전인 현재 가상화된 3D 형태로 공장 설립을 완성함으로써 실제 건설 과정의 시행착오를 줄이고 비용을 절감할 수 있을 것으로 기대된다. Volvo와 GM, Mercedes-Benz 등 여러 글로벌 자동차 제조사들 또한 자율주행 테스트, 자동차 조립라인 구축, 공장 로봇 훈련 등에 옴니버스를 활용하고 있다.



IDC에 따르면 전 세계 디지털 전환(digital transformation) 관련 투자는 2026년 3조 4천억 달러에 달할 전망이며, 디지털 트윈 기술의 시장 규모는 약 1.6천억 달러에 달할 것으로 예상된다. 최근 글로벌 기업들의 리쇼어링 움직임이 나타나는 가운데 자동차, 반도체 등 주요 제조업 분야의 공장 증설이 활성화되고 있다. 물리적인 시스템 구축에 앞서 가상세계 상에서 이를 시뮬레이션 하는 옴니버스의 디지털 트윈 기술은 기업의 생산성 개선 수단으로써 주목받을 것이다.



c. 옴니버스가 활용된 아마존의 로봇

GTC 2023 행사에서는 특히 옴니버스와 “아마존 로보틱스”와의 협업을 강조했다. 아마존의 세계 최대의 전자상거래 기업이자, 또 물류기업이다. 아마존 로보틱스는 그러한 대형 물류업자인 아마존 산하의 이동식 풀필먼트 시스템 로봇을 담당하는 기업이다. 아마존이 공장 설계에 착수하기 전에 3D 설계도를 가상으로 구현하거나, 물리적인 기계의 성능을 가상 세계에서 시뮬레이션하여 성능과 안정성 등을 테스트하는 것이 옴니버스의 대표적인 활용 사례로 표현됐다.



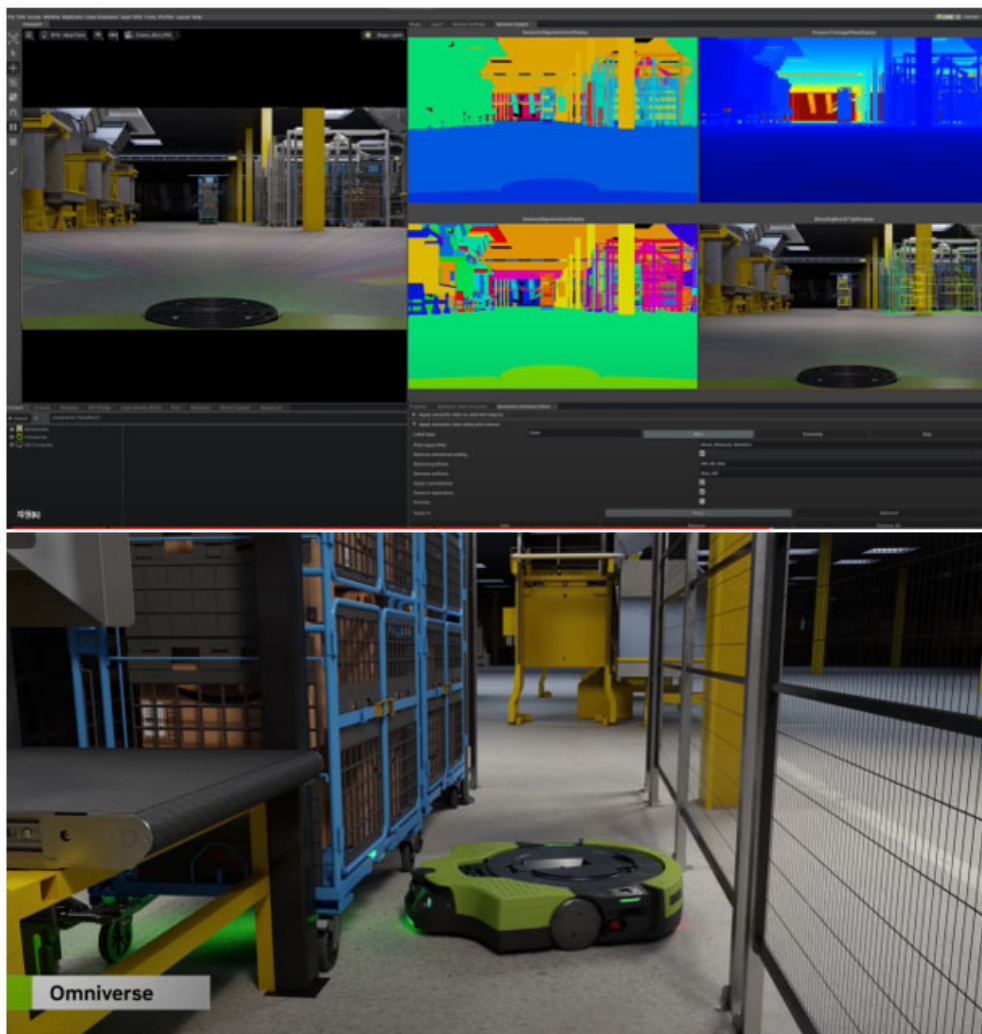
자율 물류창고 로봇인 “Proteus”는 물품을 정확히 운반하기 위해서는 공장 바닥에 표시된 “기준 마커(바닥에 깔린 일종의 인식표)”를 읽고 지도에서 자신의 위치를 파악하는 것이 중요하다. 바로 이때, 엔비디아의 옴니버스가 활용됐다고 한다. Proteus는 카메라, 라이다, 초음파 센서를 포함한 여러 센서를 특징으로 하여 자율주행 소프트웨어 시스템을 구동한다. 로봇이 물류창고 내 지도상에서 위치를 파악하기 위해서는 마커를 읽기 위한 신경망의 성능을 개선해야 했다.

로봇의 머리에 탑재될 딥러닝 모델을 훈련시키기 위해서는 많은 양과 또 적절한 종류의 이미지데이터가 필요하다. 여기서 옴니버스 리플리케이터가 제공하는 생성형 AI의 실력발휘가 이뤄졌다.

대규모로 매우 현실적인 이미지 데이터를 생성해, 이것으로 딥러닝 모델을 시킨 것이다. 이로써 아마존 로보틱스는 마커 감지 성공률을 88.6%에서 98%로 개선할 수 있었다고 한다.

또한 중요한 것은 바로 속도에 대한 부분이다. 옴니버스 리플리케이터에 의해 생성된 합성 데이터의 사용으로, Proteus 로봇 최적화 개발시간을 몇 달 수준에서 일단위로 단축시킬 수 있었다. 이는 실제 데이터만 사용할 때보다, 생성한 가상 이미지 데이터까지 투입시키면 모델을 훨씬 빠르게 훈련시킬 수 있었다는 말이다. 이 모든 것은 비용이 절감된다는 소리다.

옴니버스 리플리케이터를 통한 아마존 Proteus 물류로봇이 학습하는 모습



자료: Nvidia GTC 2023

실생활에도 엄청난 활약을 할 AI 기술

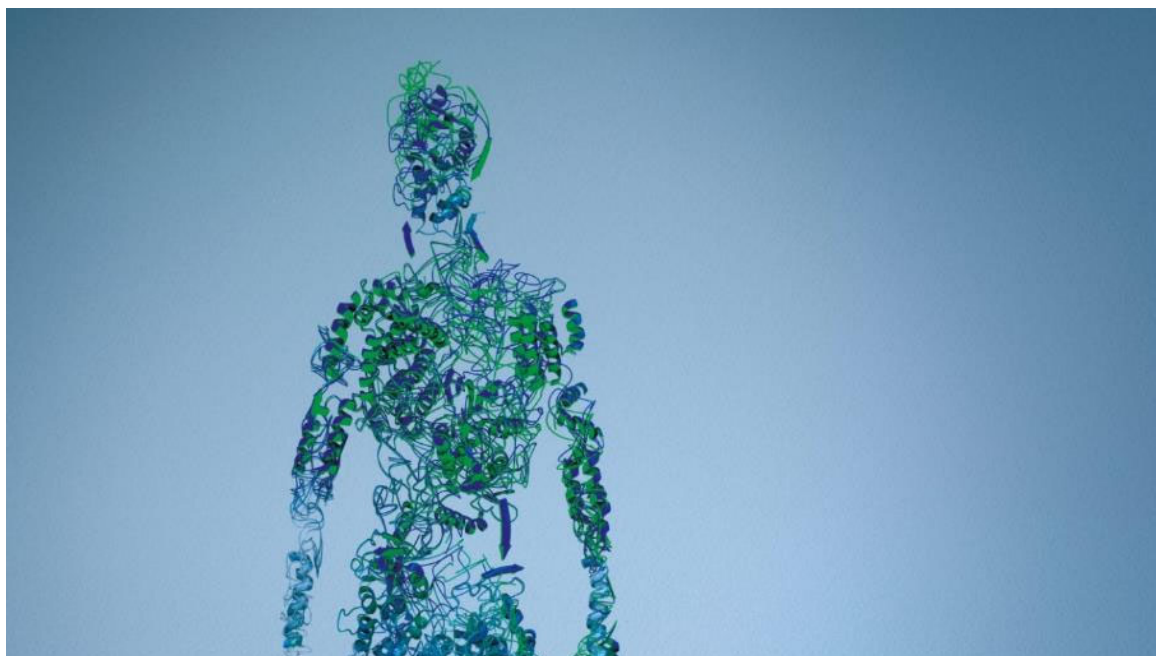
1) 바이오 헬스케어

a. 알파폴드의 충격적 데뷔

구글의 딥마인드(DeepMind)는 알파폴드(AlphaFold)라는 헬스케어 분야의 인공지능 프로그램을 보유하고 있다. 그리고 알파폴드가 유명해진 이유는, 딥러닝 기술로 단백질 구조를 예측하고 또 해독하는 것에 성공했기 때문이다. 단백질 구조라는 것은 생명과학 및 의학 연구에서 매우 매우 중요한 문제다.

단백질이 생명체에서 다양한 기능을 수행하는데, 그 구조를 안다는 것은 그 기능을 이해하는데 첫 발걸음이 되기 때문이다. 단백질 구조를 정확하게 예측함으로써, 연구자들은 단백질의 기능과 상호 작용을 더 잘 이해할 수 있으며, 이를 바탕으로 새로운 치료법이나 약물을 개발하는 데 도움이 된다는 말이다.

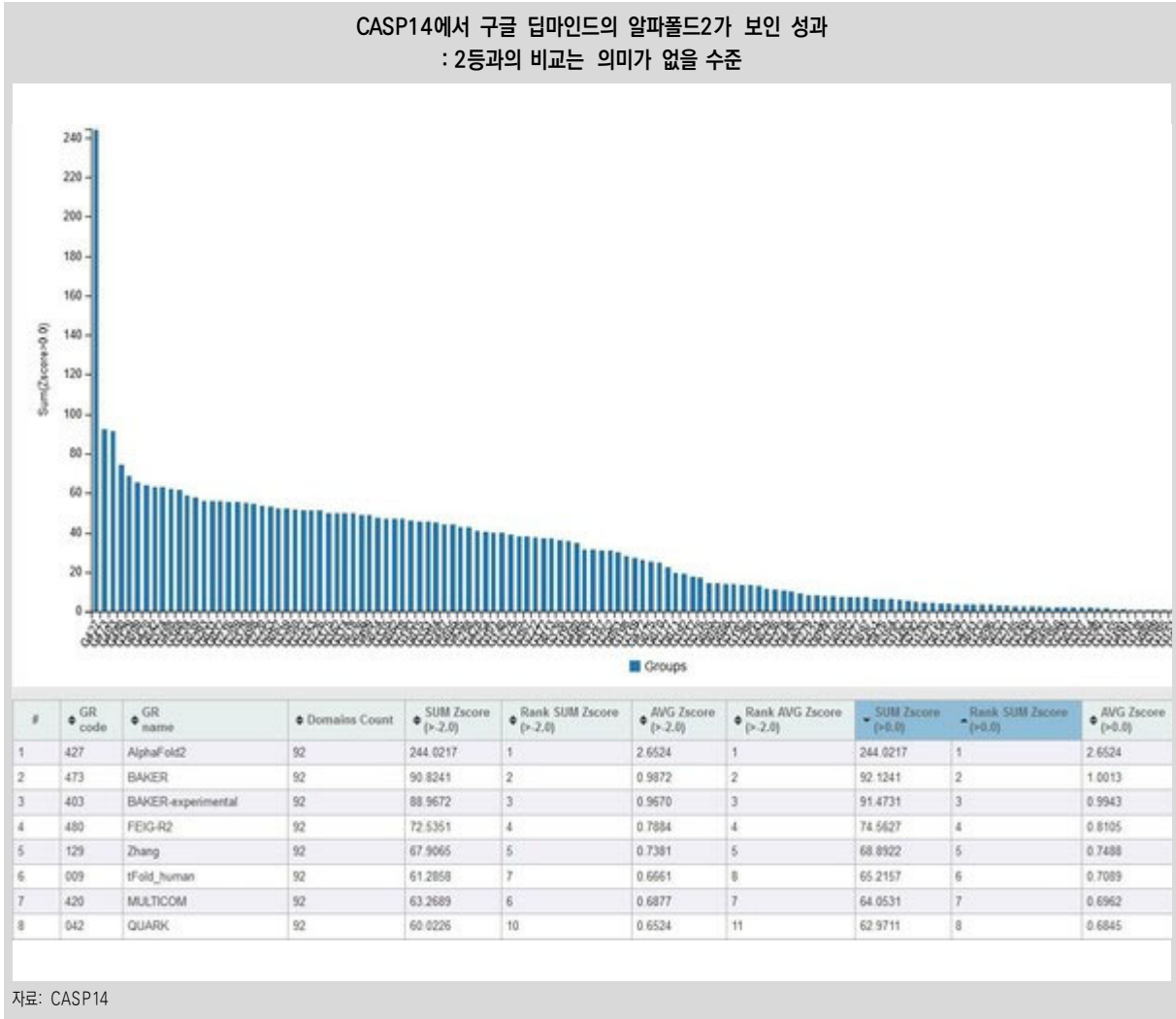
단백질 구조 해독은 인체의 기능을 이해하기 위한 “위대한 시작”



자료: DeepMind

알파폴드가 대단한 점은 2021년 Nature에 발표한 딥마인드의 논문자료에 기인한다. 알파폴드는 본인들의 딥러닝 기술로서 단백질 구조를 예측했는데 기존에 사용하던 시뮬레이션 방식보다 더 높은 정확도를 보여줬다. 2020년 11월 30일에 열린 CASP14 대회에서 알파폴드의 차세대 버전인 알파폴드2는 다양한 단백질 구조를 예측하는 데 Global Distance Test(GDT) 점수에서 90 이상을 기록하며 우승을 기록했다. 2등과의 격차는 무려 25점이나 나서 비교 자체가 무의미했다.

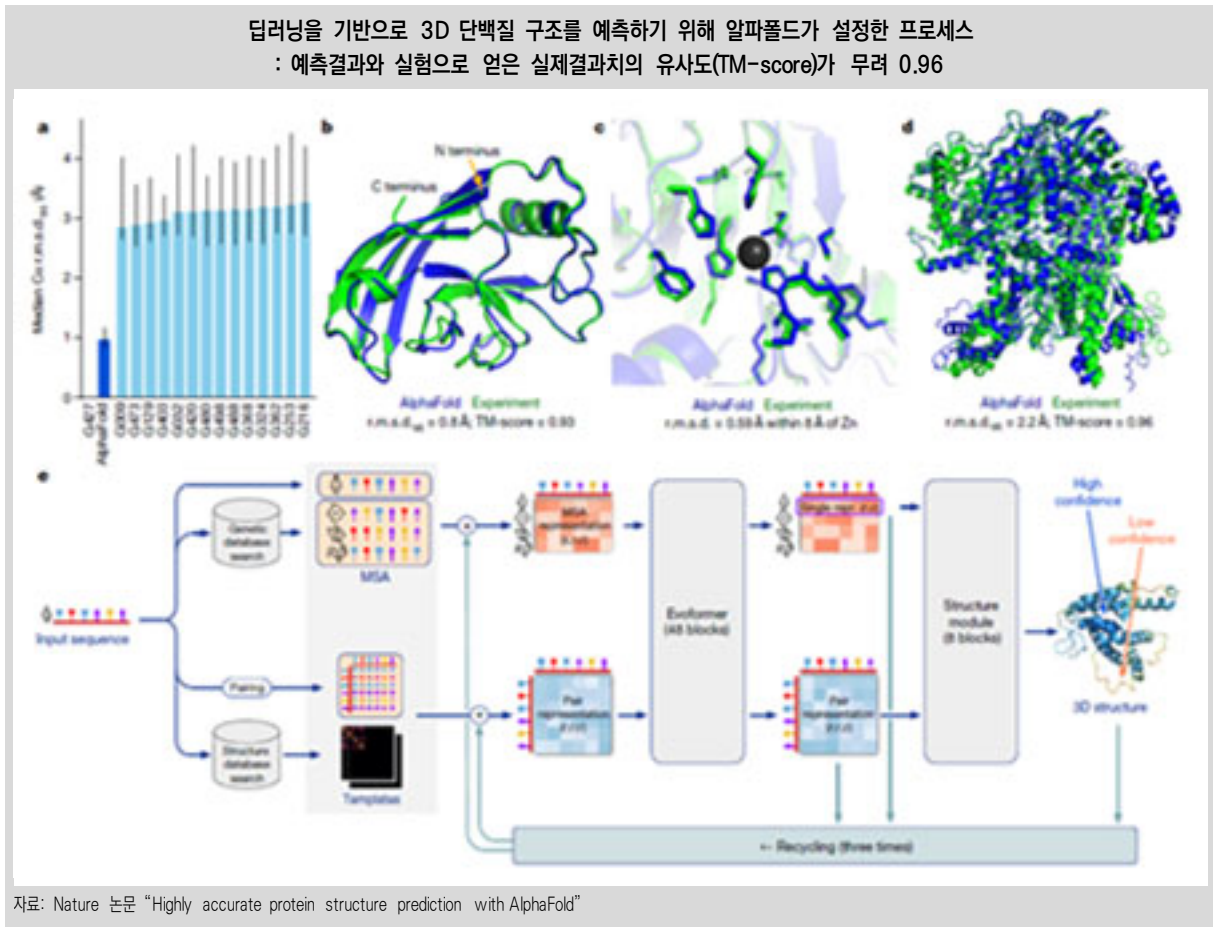
*CASP14: 제14회 단백질 구조 예측 평가 대회의 줄임말. CASP는 2년마다 개최되는 글로벌 대회.



또한, 단순히 단백질 구조 간의 거리만 뛰어난 게 아니라 TM-Score로도 2021년 기준 0.96을 기록했다. 0.6 이상만 기록되더라도 높은 정확도로 인정을 받는데, 거의 1에 근접한 스코어를 기록했으니 관련 학계가 뒤집어진 것이다. 해당 결과를 지켜본 많은 전문가들은 “10년 동안 특정 단백질 구조를 못 알아냈는데 알파폴드2가 30분만에 밝혀냈다. 알파폴드2는 게임체인저다. 앞으로 단백질 구조 분석은 컴퓨터(딥러닝)에 의존하게 될 것이다”라며 감동을 나타냈다.

**TM-Score: Template Modeling Score의 약자로, 단백질 3차원 구조의 예측 결과와 실험적으로 얻은 구조 간의 유사성을 측정하는 지표. 이 점수는 0에서 1 사이의 값을 가지며, 값이 1에 가까울수록 두 단백질 구조가 매우 유사함을 나타냄. 단백질의 길이와 전체적인 구조적 특성을 함께 고려하여 구조 간의 유사성을 평가. 이로 인해 TM-Score는 단백질 구조의 크기와 독립적으로 구조 간의 유사성을 비교하는데 더욱 적합한 지표.*

알파폴드가 해당 대회에서 큰 점수로 우승을 기록하자 딥러닝 기반의 단백질 구조 예측 가능성이 크게 인정받았다. 그야말로 딥러닝을 써야 하는지의 고민보다는, 어떻게 잘 쓰는지가 헬스케어 영역에서도 중요한 화두가 되었다. 이는 언어모델인 ChatGPT를 사용하고 있는 우리의 고민과 동일한 것이기도 하다.



특히 놀라운 점은 알파폴드2는 이전 버전인 알파폴드1에 비해 단 2년 만에 크게 발전한 모습을 보여줬다는 것이다. 알파폴드를 만든 구글 딤마인드는 본인들만의 특화된 고급 딤러닝 알고리즘을 사용하여 단백질 구조를 예측했다고 한다. 특히 트랜스포머 어텐션 기법을 네트워크에 결합했고 알파폴드2라는 성공적인 결과를 얻었다. 딤러닝 네트워크 구성 디자인과 모델 알고리즘 최적화를 통해 딤러닝 기술이 단백질 구조 예측 분야에서 얼마나 빠르게 성장할 수 있는지를 보여주는 좋은 사례라고 할 수 있다.

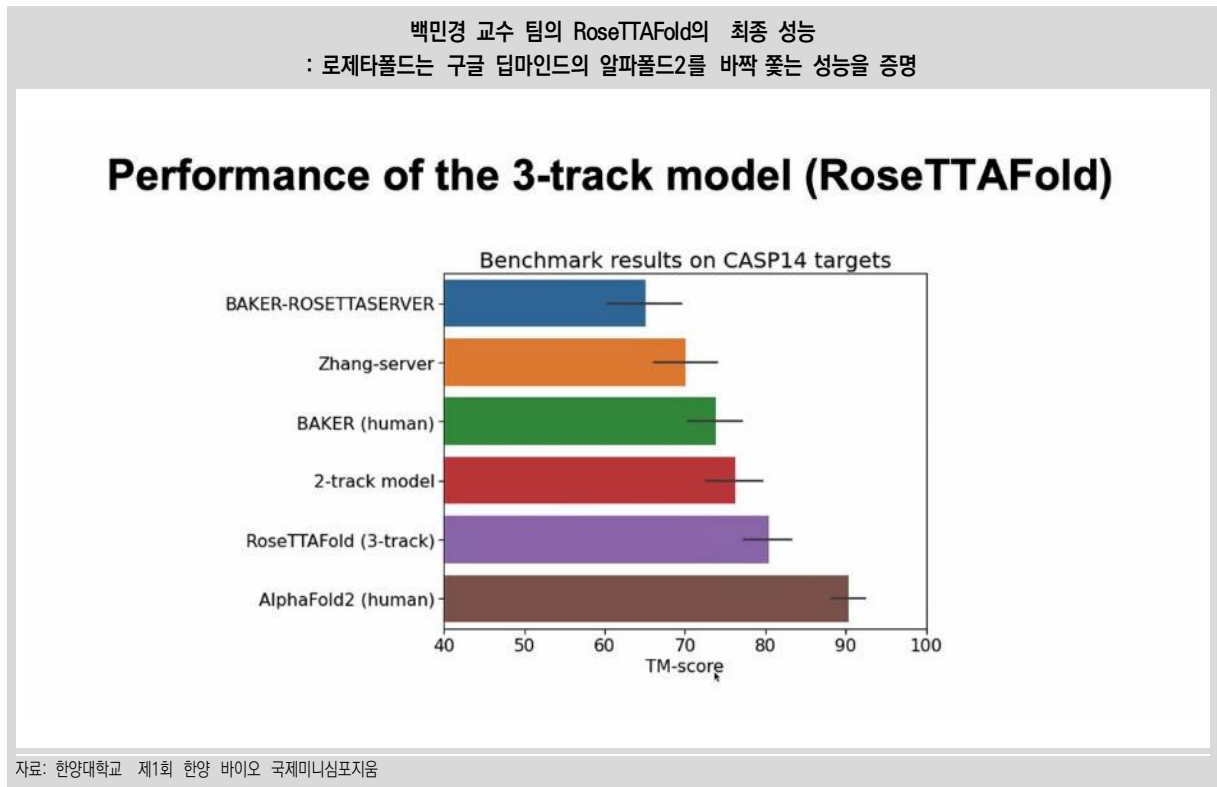
b. 알파폴드의 성공으로 나타난 긍정적 효과들

알파폴드의 성공은 바이오 헬스케어 업계의 전체적인 수준을 향상시키고 있다. 알파폴드가 2021년부터 오픈소스로 개방되었기 때문이다. 이에 따라, 2022년에 개최된 CASP15에서도 알파폴드의 힘이 발휘됐다. 딤마인드가 해당 라운드에는 참여하지도 않았지만, CASP15에서 우승한 팀이 본인들의 모델 알고리즘을 알파폴드 기반으로 만들었기 때문이다.

오픈소스를 통해 공유되는 집단지성이 향후 알파본드와 같은 여러 모델들을 출현시킬 것으로 보인다. 이는 바이오헬스케어 산업의 혁신을 더욱 가속화할 것으로 기대된다. 샘 알트만 OpenAI CEO 또한 알파폴드와 비슷한 어플리케이션이 엄청난 양의 부가가치를 창출하고 있다면서, 이와 같은 시스템들을 앞으로는 훨씬 더 많이 보게 될 것이라고 언급하기도 했다.

그리고, 이에 해당하는 부가가치가 대한민국에게도 강하게 전파되고 있다는 게 굉장히 반가운 일이다. 現 서울대학교 소속 백민경 교수팀의 로제타폴드(RoseTTAFold)가 바로 그 주인공이다. 알파폴드의 등장 덕에 수혜를 본 많은 모델들이 있지만, 로제타폴드가 특히 주목받은 이유에는 효율성 측면에서 우수함을 발휘했기 때문이다. 구글 딥마인드 수준의 AI 전문가와 연산력을 갖고 있지 않았어도 알파폴드2 수준과 비슷하게 단백질 구조를 예측해냈다.

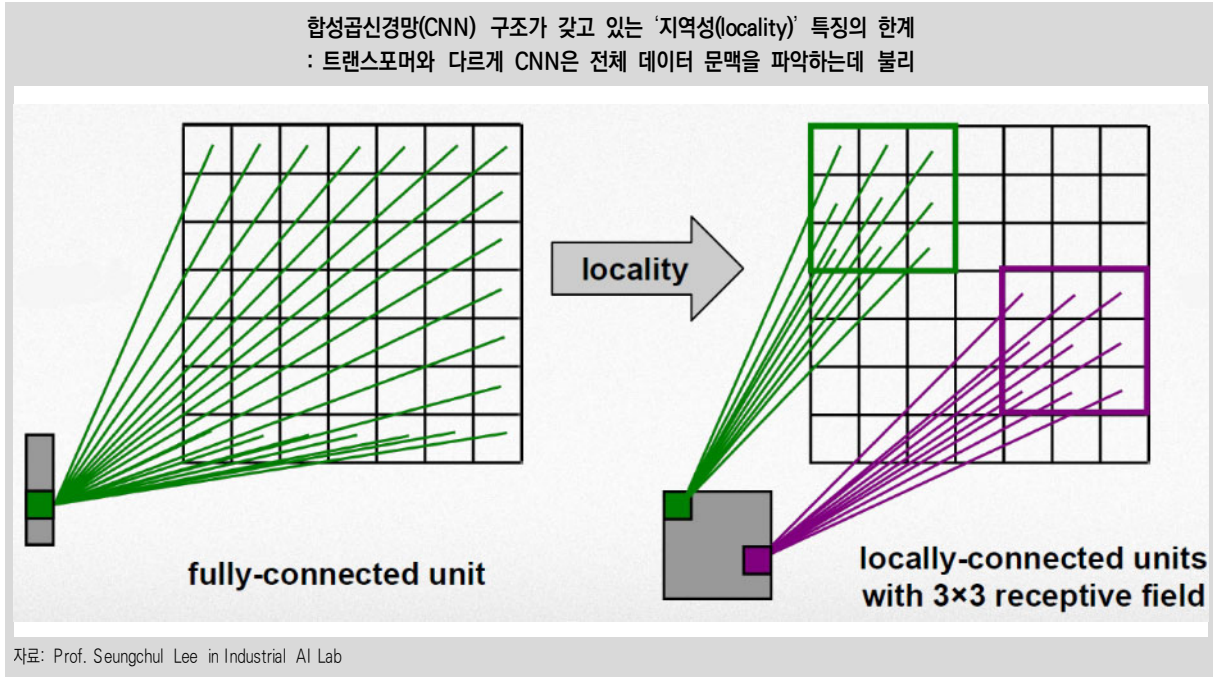
**로제타폴드가 개발되어 세상에 공개된 2021년에 백민경 교수는 美 워싱턴대학교 소속이었음*



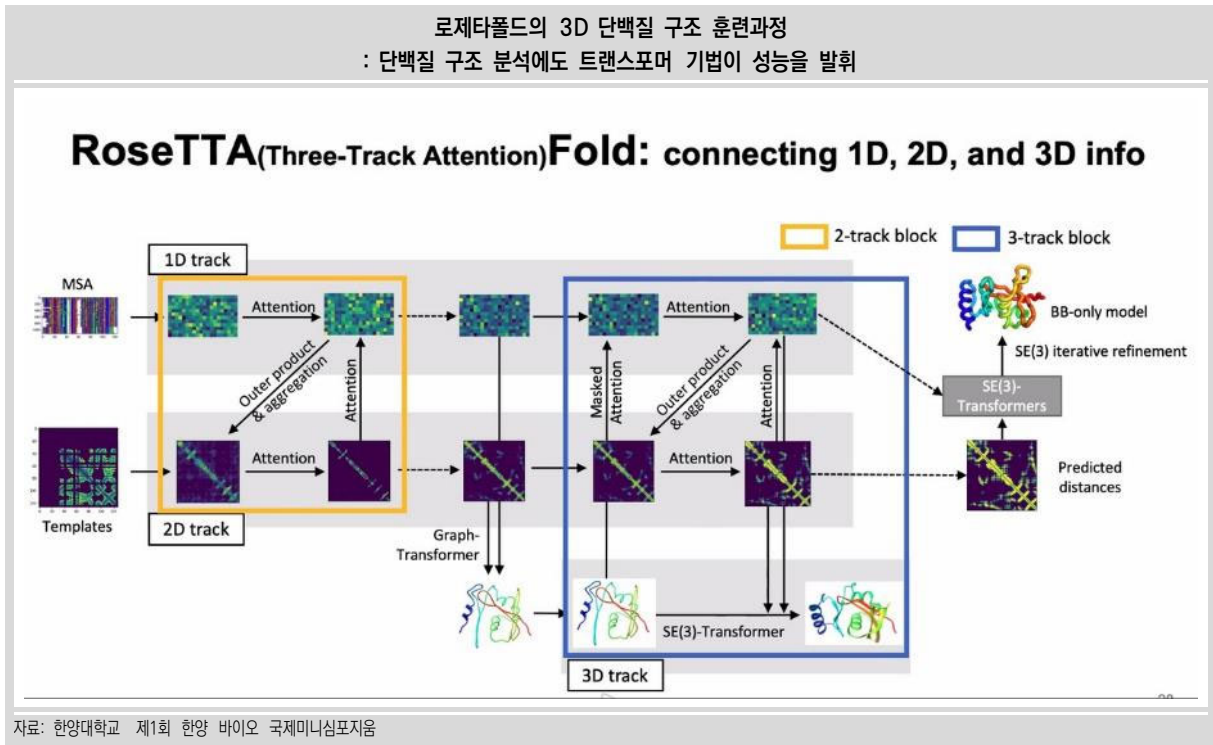
백 교수에 따르면, 로제타폴드 개발에 참여한 연구진들은 생물학, 화학 분야 전문가들로 구성돼 AI 전문가가 없었다고 한다. 다만, 마이크로소프트 측에서 GPU를 지원했고, Azure의 클라우드 컴퓨팅을 위해 데이터 토큰 크레딧 정도를 제공했다고 한다. 마이크로소프트도 이 분야에 군침을 흘리고 있다는 뜻이기도 하다.

알파폴드2도 그렇지만, 로제타폴드가 여타 단백질 구조 예측 모델보다 뛰어난 성능을 보이는 비결은 트랜스포머(Transformer)의 일종인 어텐션(Attention)을 도입했기 때문이다. RoseTTA의 이름에서 TTA라는 것 자체가 Three-Track “Attention”을 뜻한다.

반면, 알파폴드2 등장 전 대부분의 연구자들은 합성곱신경망(CNN: Convolutional Neural Networks) 기반 딥러닝을 단백질 구조 예측에 사용했다. 다만, CNN은 지역적인 시퀀스 정보만 담기 때문에, 특정 단백질 구조를 전체적으로 파악하는 데 무리가 있다.



하지만 트랜스포머의 어텐션 기법을 활용한다면 얘기가 다르다. 인풋 데이터를 전부 다 ‘훑어보는’ 특성 때문에 단백질 구조 데이터 간 관련성을 파악하는데 더 이롭다. 백 교수는 “단백질 시퀀스상에서 멀리 떨어져 있어도 구조적으로 관련있는 아미노산을 가져오는 것이 중요하다”고 강조했다.



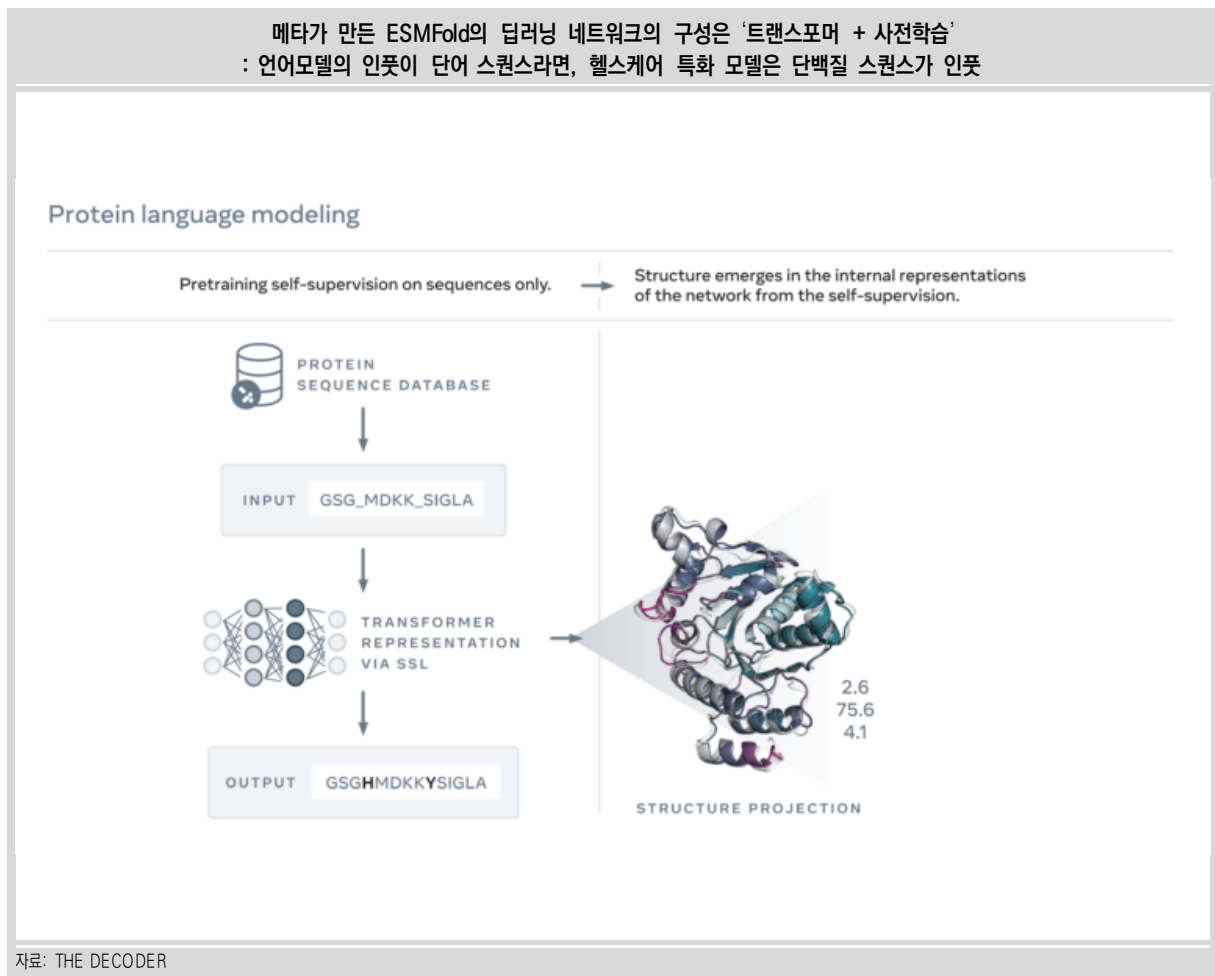
또한 백 교수는 “단일 상태의 단백질을 연구하는 시대는 지났다면서, 앞으로 여러 단백질의 상호작용까지 파악하기 위해서는 더욱 정교한 딥러닝 모델 알고리즘을 만드는 것이 목표”라고 했다. 따라서, 실제 활용하기 위해서는 좀 더 발전될 여지가 있다는 말이다.

이와 관련해 딥마인드 연구진들 또한 알파폴드의 예측력을 더욱 개선할 것이라고 강조하기도 했다. 실제로 CASP14에서의 알파폴드2의 성과는 전체 단백질이 아니라, 비교적 단순한 단일 도메인에 대한 것이며 몇 가지 잘못된 예측 결과도 있었기 때문이다. 또한, 코너 케이스(환경/조건 변경에 따라 발생하는 문제) 해결은 남아 있는 상황이기도 했다.

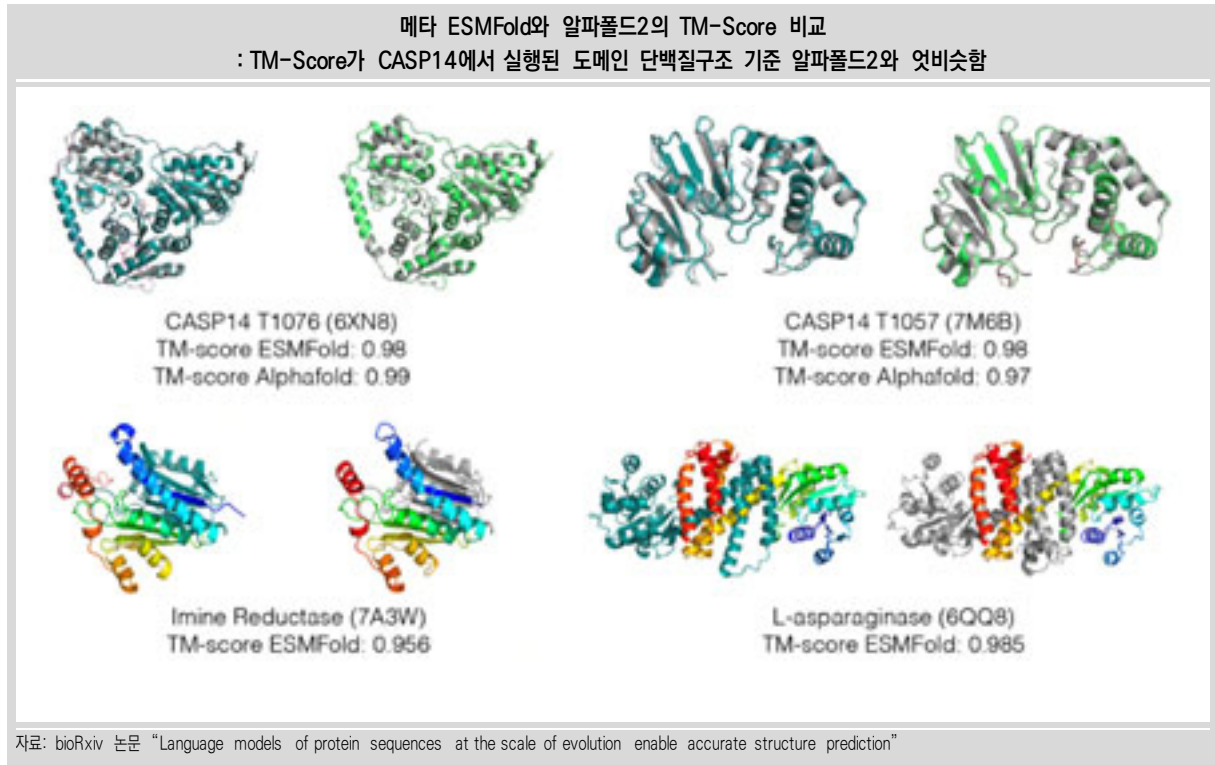
딥마인드는 “알파폴드2가 추후 더 다듬어진다면 단백질 폴딩 분야에서 이전에는 다루기 힘들었던 문제에도 적용할 수 있을 것”이라고 강조했다. 아직 과학계에 밝혀진 바 없는 수억 개의 단백질을 모두 다 탐색하는 데 딥러닝을 사용하게 될 것이라는 포부다.

**단백질 폴딩: 단백질 분자가 가지는 고유한 3차원 구조. 이 구조는 단백질 분자가 수행하는 여러 기능과 생체활동을 결정함. 예를 들어 어떤 단백질 분자가 다른 분자와 결합할 때 그 위치를 지정해주기도 함. 단백질 폴딩 문제의 해결은 새로운 치료법을 개발할 때 중요한 역할을 할 것.*

그렇다면 알파폴드는 어떤 식으로 더 발전할 수 있을까? 공교롭게도 여기에도 언어모델이 도움이 될 수 있다. 우리가 일전에 공부한 GPT나 BERT 같은 트랜스포머 기반 언어모델이 바이오헬스케어 분야에서 실력을 발휘하게 된다는 말이다. 실제로, 메타의 ESMFold와 같은 언어모델은, 문장이나 단어가 아니라 단백질 구조를 예측하는데 쓰이고 있기도 하다. ESMFold는 150억 개의 매개변수를 갖고 있는 트랜스포머 언어모델을 기반으로 한다.



bioRxiv 논문에 게시된 데이터에 따르면, 알파폴드2보다도 ESMFold가 단백질 구조를 6배나 더 빠르게 예측할 수 있다고 한다. 알파폴드가 세상을 뒤집어 놓은 게 얼마 지나지 않았는데 이런 일들이 발생하는 것이다. 그것도 테크 업체들이 이런 것들을 발표하는 것이 놀랍다. 이는 딥러닝을 통해서도 바이오헬스케어 분야가 얼마나 빠른 속도로 발전할 수 있는지 실감할 수 있는 대목이다. 정말 인공지능 영역은 자고 나면 승자가 바뀌어 있는 분야라고 이해해도 될 정도다.



c. 언어모델의 확장 가능성

구글의 딥마인드든 메타의 AI 연구소든, 빅테크들이 단백질 구조 분석에 탁월한 성능을 나타내는 이유는 명확하다. 그들이 LLM(대규모 언어모델, Large Language Model)에 있어서 가장 앞서나가는 기업들이기 때문이다. OpenAI 또한 헬스케어 특화 모델 알고리즘을 내놓지는 않았지만 만들고자 한다면 가능할 것으로 보인다. LLM은 일반적으로 자연어 처리(NLP)와 관련된 문제를 해결하는 데 사용되었다. 하지만 위 사례들처럼 알파폴드와 같은 단백질 구조 예측 문제에도 활발히 적용되고 있다. 알파폴드는 인풋으로 단백질 시퀀스 정보를 주면 아웃풋으로 단백질 구조를 3D로 제시하는 모델이다.

이렇게 할 수 있는 이유는 문장은 문자(단어)들의 시퀀스(나열)로 구성되고, 단백질은 아미노산의 시퀀스로 구성되기 때문이다. 그러니까 구조적 관점에서 언어와 단백질 사이에는 상당한 유사점이 있다는 게 핵심이다. 두 경우 모두, 시퀀스의 패턴과 시퀀스 내 데이터의 상호작용이 결과물의 구조와 기능에 큰 영향을 미친다. 따라서 LLM으로 단백질 시퀀스의 패턴을 학습하고, 이를 바탕으로 단백질 구조를 예측하는데 도움이 될 수 있는 것이다.

우리는 생성 모델의 힘을 너무 간과하고 있는지도 모른다. 뒤에 오는 단어를 단순히 예측하는 것으로만 생성 모델을 쓸 것이라는 생각은 단편적이다. 이처럼 수치로 나열할 수 있는 모든 것(임베딩 벡터 시퀀스로 나타낼 수 있음)에 생성형 AI는 힘을 발휘할 것이다.

LLM은 이미 다양한 분야에서 놀라운 성과를 보여주고 있지만, 이를 단백질 구조 예측에 적용하는 것은 아직도 초기 단계다. 향후 연구에서 LLM을 사용한 단백질 구조 예측의 효과가 더욱 명확해질 것으로 기대된다.

아래는 딥러닝 언어모델을 헬스케어에 적극 사용했을 때 기대되는 효과들을 정리한 것이다.

(1) 신약개발: 단백질 구조 예측의 정확도와 속도 향상으로 신약 개발 과정에서 새로운 약 물질의 발견과 설계를 가속화할 수 있음. 이로 인해, 치료법 개발에 드는 시간과 비용이 줄어들고, 다양한 질병에 대한 효과적인 치료 전략을 빠르게 만들어낼 수 있게 됨.

(2) 개인화 의료: LLM을 사용한 단백질 구조를 예측하면, 개인의 유전자 서열을 바탕으로 개인화된 단백질 구조 예측도 수행할 수 있음. 이는 각 개인에게 가장 적합한 치료법이나 약물을 찾는 데 도움이 될 것.

(3) 생명과학 연구 기여: 높은 정확도로 예측된 단백질 구조는 생물학자들이 단백질 간 상호 작용, 생화학적 반응, 질병 원인 등에 대한 이해를 향상시킬 수 있게 함.

위와 같은 영역에서의 발전을 위해 실제로, 인공지능 산업과 생명과학 분야의 연구 및 협업이 확대되고 있다. 앞으로도 딥러닝 기술의 고도화를 통해, 헬스케어 섹터 내에서도 관련투자에 많은 자금집행이 이뤄질 것으로 판단된다.

그런데, 알파폴드의 연구진들은 실험 방법이 여전히 시간이 많이 소요된다면서, 이에 관한 추가 개선이 필요하다고 말하기도 했다. 바로 이 테스트 분야에서 딥러닝 연구와 투자가 많이 이뤄질 것으로 기대된다. 샘 알트만 CEO는 “원래 바이오/헬스케어 업계에서 사용하던 시뮬레이터가 너무 나쁘다”고 지적하기도 했다. OpenAI에서도 바이오 스타트업들을 위한 시뮬레이터 개발에 박차를 가할 것이라고 공언한 상태다. 즉, 시뮬레이터 성능을 개선하기 위한 여러 작업물들이 출현할 것으로 예상된다.

시뮬레이터의 발전은 모델 알고리즘의 개발과 평가분석, 나아가 헬스케어 산업발전에 영향력 있는 공헌을 할 것이다. 특히, 인건비의 절약과 신약 개발 속도가 절감된다는 것을 의미한다. 이 과정에서 상상력을 조금만 동원해보면 재밌는 일이 생겨난다. 헬스케어 섹터 내의 스타트업이 좋은 아이디어만 갖고 있으면 이제는 빅파마와도 경쟁할 수 있게 될 것이기 때문이다. 가장 돈과 시간이 많이 수반되는 “테스트” 영역에서, AI 기술로 인해 비용이 그만큼 저렴해지고 사이클이 짧아져서 가능한 일이다.

위 모든 과정은, 결국 하나의 목적으로 귀결된다. 인간 전문가가 필요로 하는 시간과 노력을 크게 줄이기 위함이다. 신약 후보물질과 같이 시뮬레이션이 필요한 분야에서, 시행착오의 사이클 타임을 초고속으로 만드는 방법이 시로 구현될 것이다. 또한, 인공지능의 여러 모델들 덕분에 다양하고 새로운 연구 방향을 생각하도록, 인간의 창조적인 사고에도 도움을 주게 될 것이다.

2) 자율주행

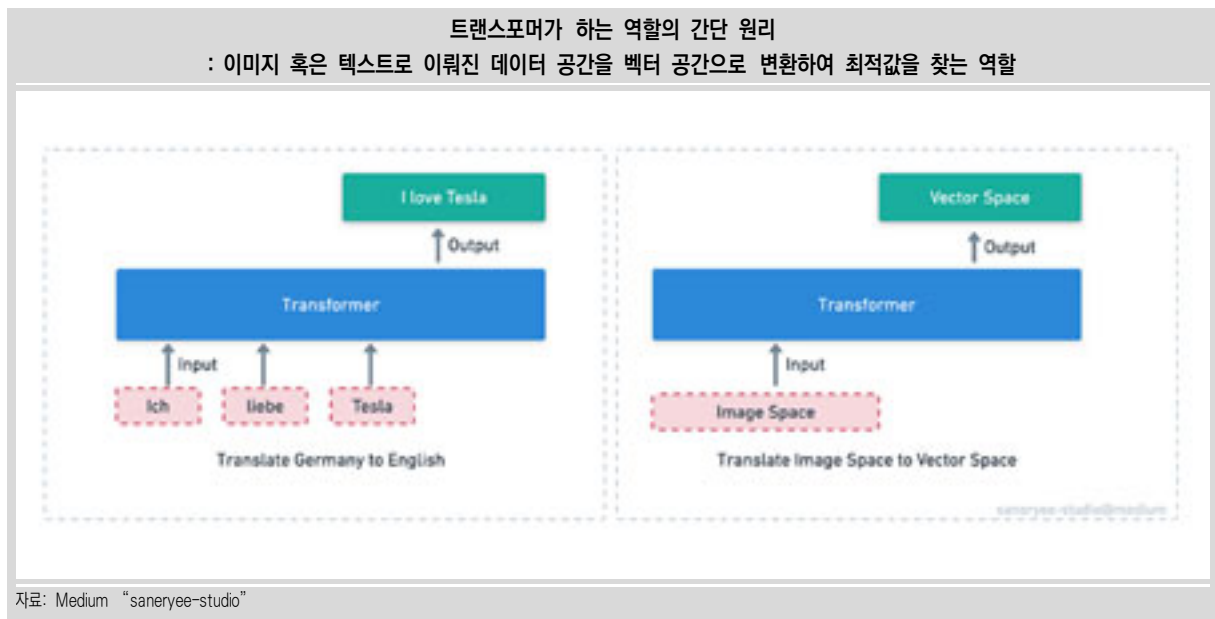
a. 생성형 AI만큼 중요한 자율주행

ChatGPT로 생성형 AI가 돌풍을 일으키고 있지만, 자연어 처리 모델만큼이나(어쩌면 더욱) 중요한 영역은 다른 아닌 자율주행이다. 사실 ChatGPT가 출시되기 전까지만 하더라도 인공지능에 있어 가장 핫한 섹터는 자연어 처리가 아니라 자율주행이었다. 2010년대 중반까지만 하더라도 자연어 처리에 있어서 가장 높은 경쟁력을 가졌다고 평가됐던 바이두조차 2020년대에 들어서는 자율주행에 더 힘을 쏟는 모양새였다. 2017년 전설적인 논문 “Attention is All you need”을 선보인 구글도 비슷했다. 바이두와 구글 같은 검색엔진 최강자들이 비슷한 행보를 보인 것이 참 흥미로운 일이기도 하다. 구글은 자회사 Waymo를 통해 자율주행 Level 5(완전 무인자율주행) 완성을 위해 천문학적인 자금을 쏟아 부었다.

그런데, 구글의 인공지능 기술에 대항해 만들어진 OpenAI가 ChatGPT를 내놓았고, 이들로 하여금 다시 자연어 처리에 열중하게 한 것은 역사의 아이러니다. 구글은 OpenAI를 품에 안은 마이크로소프트의 선전포고에 맞불을 놓고자 내부적으로 코드레드까지 발동했다. 그리고 “바드”라는 BERT 기반의 자연어 생성형 AI를 서둘러 출시하겠다고 공언한 상태다. 바이두도 어니봇을 공개했지만 시장의 반응은 시큰둥했다. 어쨌든 둘 다 불완전한 상태로 공개됐기 때문이다.

하지만 생성형 AI가 모든 스포트라이트를 받고 있다고 해서, 자율주행 기술의 경종이 변화하는 것은 아니다. 자율주행으로서 파생될 수 있는 산업의 크기는 헤아릴 수 없을 정도로 크다. 구글과 바이두가 본인들의 인공지능 역량을 자율주행에 집중하기로 결정한 이유도, 수익화 측면에서 더 매력적이었기 때문이었을 것이다. 그런 측면에서 생성형 AI로 촉발되긴 했지만, 자율주행에도 관심을 쏟아야 할 것은 분명하다.

게다가, 생성형 AI이든 자율주행이든 각 기술의 기반이 되는 인공지능 기술은 겹치는 면이 많다. GPT에서 가장 핵심이 되는 것이 트랜스포머의 셀프어텐션 기법인 것처럼, 테슬라의 자율주행 브랜드네임인 FSD에서도 핵심이 셀프어텐션이다. 이는 딥러닝 연산의 근본이 대규모 데이터를 기반으로 한 벡터 연산이기 때문이다.



그리고, 그러한 대규모 데이터를 기반으로 한다는 점에서 자율주행의 선두주자를 추정해볼 수 있다. 주인

공은 테슬라다. 테슬라만큼 차주들로부터 실제 주행 데이터를 많이 제공받는 업체는 없다. 테슬라의 자율주행 기술 확대 전략은 다른 업체들과는 달랐다. 우선, 바로 완전 자율주행을 이룩하겠다는 방식보다는 차츰차츰 자율주행 모델 알고리즘의 성능을 올리겠다는 것이다.

테슬라가 벡터 공간을 만드는데 쓰이는 중요 부분
 소재목: FSD 딥러닝 네트워크에서 가장 중요한 부분은 '트랜스포머 셀프어텐션 기법'

자료: Tesla AI DAY 2021

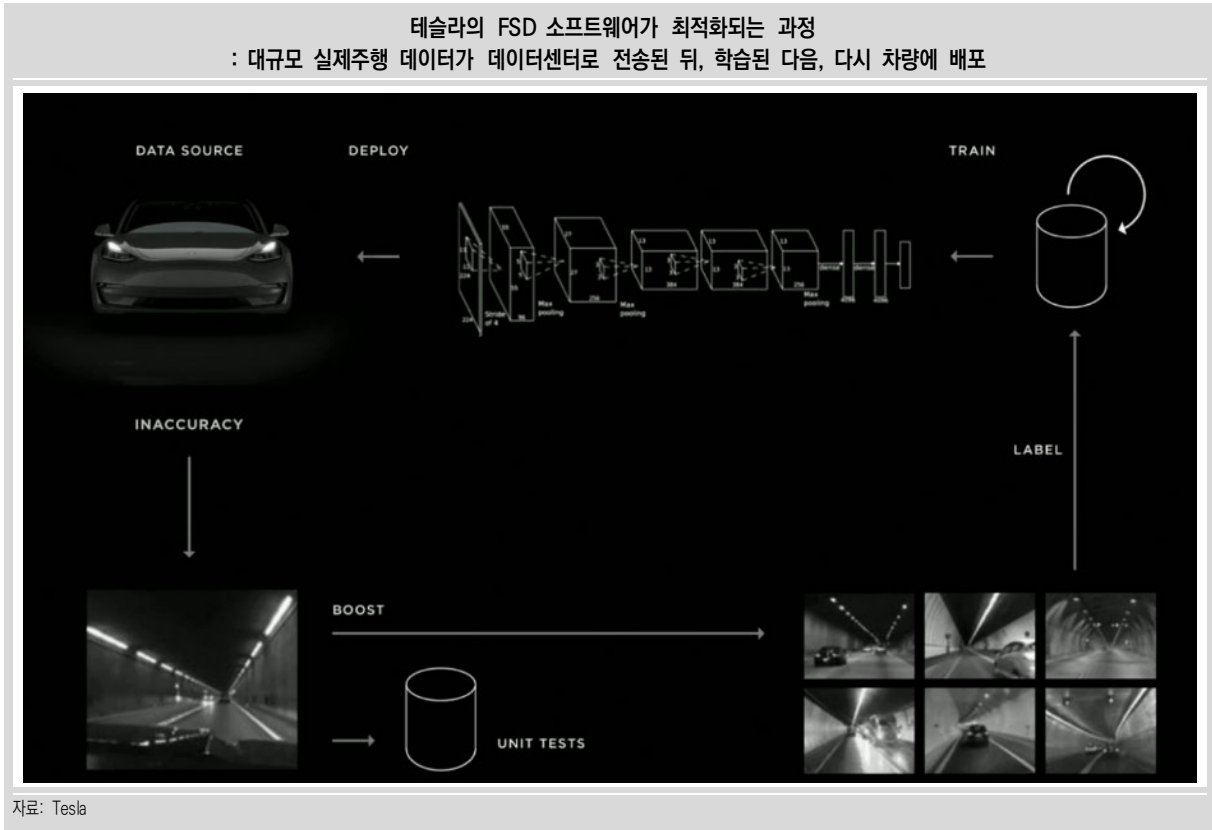
그리고 그 과정에서 FSD “Beta”라는 이름으로 당당히 구독 또는 일시불의 형태로 ‘소프트웨어 장사’를 했다. Beta의 특성상 사람들이 ‘베타테스터’로서 수행해야 하는데, 이를 기꺼이 받아들이는 사람들이 꽤 많았다. 자율주행에 대한 기대감과 테슬라의 강력한 팬덤 때문이었다. 사용자들로부터 주행 데이터를 많이 받아 알고리즘이 개선되기 때문에 시간이 지날수록 FSD의 성능은 더 올라간다.

아래의 내용은 FSD 알고리즘이 어떻게 개선되는지 간단히 정리한 것이다.

- (1) 주행 데이터는 테슬라 드라이버들이 보내준 비디오 클립들을 가지고 시가 알아서 주행 상황들을 예측해서 라벨링을 함으로써 생성됨(오토라벨링)
- (2) 차량이 FSD를 하다가 예상과 다른 상황이 벌어져 문제가 생기면 해당 케이스들을 서버에 전송.
- (3) 수신한 모든 카메라 정보를 다시 분석하면서 무슨 문제가 있었는지 알게 되고, 데이터를 모아서 부정확한 판단의 원인을 찾으며 트레이닝을 진행.
- (4) 알고리즘 개량 후 다시 차량으로 OTA 업데이트해 FSD가 개선.

시간이 지날수록 자율주행 알고리즘 모델 성능이 올라가기 때문에 FSD의 가격은 거의 매년 인상됐다. 물

론 아직 전방주시를 해야 하고 핸들에 손을 올려 두기도 해야 하는 등 Level 3에도 못 미치는 성능이나, 많은 사람들을 FSD 고객으로 유치하고 있다는 게 중요하다. 테슬라의 FSD가 Public Release된 북미에서는 올해 초 이용자가 40만 명을 돌파했다.



b. 테슬라가 다른 자율주행 기업들과 다른 점

미국의 구글과 GM크루즈, 그리고 중국의 바이두는 무인자율주행 Level 5를 목표로 하고, 자율주행 택시 사업에 중점을 뒀다. 그리고 그 기술 플랫폼을 호스팅하는 전략을 세웠다. 그러니까 이들은 처음부터 완벽히 설계된 자율주행 기술을 세상에 선보이려고 노력했던 것이다. 그래서 최대한 정확하고 풍부한 시각 데이터를 얻기 위해 차량 군데군데 상당한 숫자의 센서들을 달았다. 카메라는 물론이고, 초음파센서, 레이더, 그리고 라이다까지 장착했던 것이다. 그러니까 이들은 자율주행 차량에 있어 눈에 집착했다.

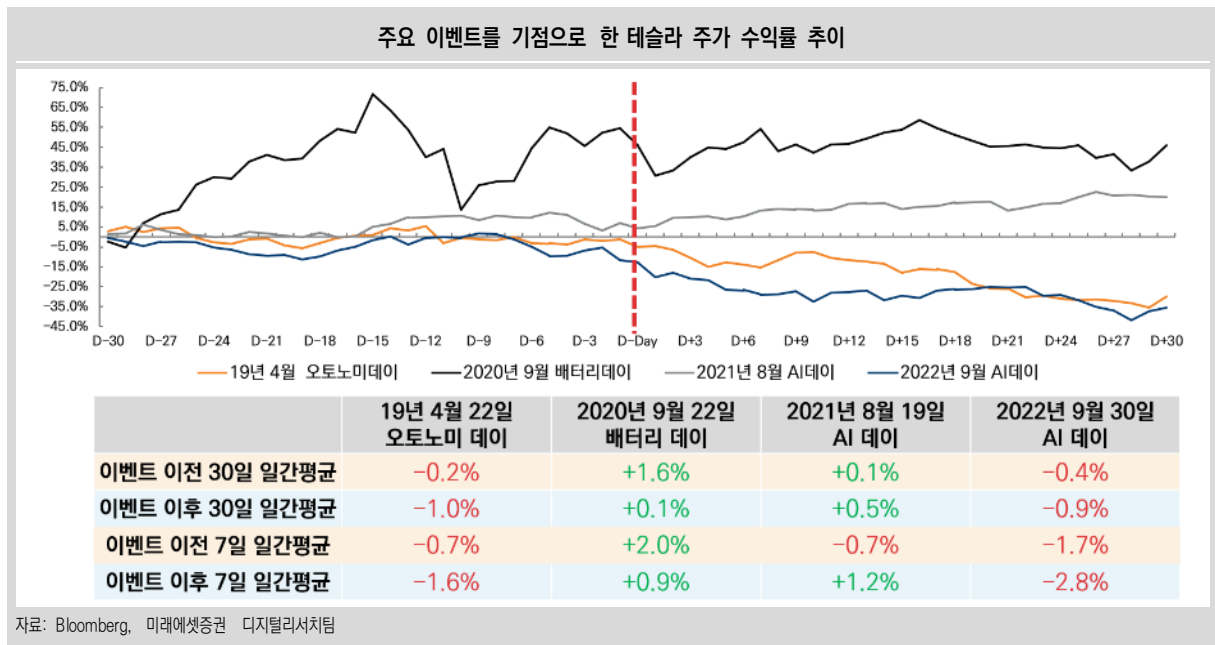
하지만, 테슬라는 눈이 아니라 두뇌로 시선을 옮겼다. 인공지능의 핵심은 결국 연산이다. 많은 데이터가 있으면 좋지만, 연산력이 감당하기 힘들 정도로 많은 데이터라면 오히려 독이 된다.

우리가 운전을 하게 되면, 주행에 있어 중요한 정보는 사실 우리가 보는 시야에서 특정 부분만을 차지한다. 예를 들면, 주행 할 때 인도에 서 있는 사람의 키가 몇인지에 대한 정보는 필요가 없다. 대강 보고 저기 사람이 서 있다는 것을 '식별'하고, 당장 차도로 뛰어들 것 같지는 않다 정도만 '예측'하면 된다. 식별과 예측은 딥러닝에 있어서 지도학습과 자기지도학습으로 해결해야 하는 부분이다. 결국 자율주행에 있어서 핵심은 센서가 아니라 연산이다.

그래서 테슬라가 자율주행을 이루겠다고 한 포부의 중심에는, 인공지능 연산을 어떻게 극대화할 것인가에 초점이 맞춰져 있다. 인공지능 연산이 순조롭기 위해서는 대량의, 또 양질의 데이터가 필요하다. 그 데이터로 좋은 파라미터를 찾을 수 있도록 모델 알고리즘을 짤 수 있는 소프트웨어 역량도 필요하다. 그리고 이를 수행할 수 있을 만큼의 컴퓨팅 파워가 있어야 한다.

그리고 이 세 가지를 본인들이 어떻게 하고 있는지, 잠재적인 구직자들인 세상의 엔지니어들에 발표한 자리가 바로 테슬라의 2021년 AI DAY였다. 일종의 ‘구인 행사’였고 투자자들을 대상으로 한 자리가 아니었기 때문에 해당 행사가 지나고 나서도 주가 흐름은 신통찮았다. 테슬라는 배터리 데이 때도 그렇지만 단기적인 투자심리 펌핑을 위해 노력하는 회사는 아니다.

하지만, 각 이벤트에서 나오는 이야기들은 절대 가볍지 않다. 또한 분명히 투자에도 도움이 되는 인사이트들이 많았다. 회사의 중장기적인 경쟁력과 비전을 말하는 자리이기 때문에 그렇다.



따라서, 핵심 내용을 중심으로 AI DAY를 다시 복기하면, 이 회사가 자율주행 딥러닝에 있어 얼마나 선구적인 입지를 점유했는지 느낄 수 있다. 2021년 8월에 개최했던 첫번째 AI DAY(이전까지는 Autonomy DAY였다)는, 테슬라가 “어떻게 트레이닝을 할 것이고”, “얼마나 많은 데이터가 필요하며”, “FSD를 위해 어느 정도의 연산능력이 필요한지”에 대한 3가지를 통하여 뉴럴 네트워크의 방향성과 가이드라인을 제시했다는 점에서 의의가 있다.

테슬라는 AI DAY에서 많은 기술적 내용을 강조했다. 이 중에서 딥러닝 네트워크의 핵심은, 카메라로만 데이터를 뽑아서 이를 3D 벡터 공간으로 만들어 최적화하겠다는 것이었다.



테슬라는 차량 외관에 달려 있는 여덟 개의 카메라로 영상 데이터를 수집한다. 카메라들의 스펙은 1280 X 960의 해상도(120만 픽셀)와 약 36Hz의 주사율로 구성되어 있다. 일반적인 핸드폰 카메라 수준과 화소수 측면에서 상대가 안되는 저화질이다. 왜 이런 품질의 카메라를 사용할까? 현재의 딥러닝 연산 능력에 비해 데이터가 과도하게 많으면 학습에 오히려 방해가 되기 때문이다.

인풋 데이터의 화소 수가 커진다는 것은 그만큼 처리해야 할 파라미터도 늘어난다는 것을 의미한다. FSD 알고리즘 모델을 최적화해야 하는 테슬라 입장에서는, 고화질의 카메라는 아직 사치라고 생각했을 것이다. 물론, 최근에는 테슬라의 데이터센터 연산능력이 좋아져서인지, 리프레시되는 모델의 카메라 해상도는 540만 픽셀로 늘어날 것이라고 한다.

결국 어느 정도의 연산능력이 감당이 가능할 정도로 데이터를 넣어야 한다. 무턱대고 데이터만 넣어서는 곤란하다. 그런 측면에서 라이다 센서의 경우에는 몇 시간运行时 ‘페타 단위’의 데이터 포인트를 쏟아내기 때문에 엄청난 수준의 연산이 필요하다. 그래서 제한된 지역에서, 제한된 속도만으로 자율주행 서비스가 이뤄지는 것이기도 하다.

테슬라 자율주행 네트워크의 첫 단계. “멀티스케일 특성 피라미드 퓨전”
: 해상도가 낮은 픽셀 데이터와 높은 데이터를 합성해 물체를 식별하면서 카메라의 낮은 스펙을 극복

Multi-Scale Feature Pyramid Fusion

EfficientDet: Scalable & Efficient Object Detection, Tan et al. 2019

자료: Tesla AI DAY 2021

테슬라의 비전시스템은 카메라 해상도가 다소 떨어지더라도 물체의 움직임을 정확히 측정하기 위해 아래의 세 가지는 분별할 수 있어야 한다. 물체가 위치한 나와의 거리, 그 물체의 속도, 가속이다. 그리고 여기서 바로 두뇌인 인공지능망의 힘이 절대적으로 필요하다. 여기서 중요한 점은, 눈(카메라)이 중요한 게 아니라 머리(딥러닝)이다.

자율주행을 위한 딥러닝 네트워크를 이용하기 위해서는 또 세 가지 준비물이 필요하다. 수백만의 실제 주행 데이터, 각 주행영상 데이터가 거리와 속도와 가속 등 세 가지의 라벨링이 제대로 되어 있는지 여부, 마지막으로는 보통 때는 잘 볼 수 없는 예외 케이스들에 대한 데이터다. 세 가지가 잘 모아진 데이터를 가지고 트레이닝을 해야 의미 있는 자율주행 알고리즘이 구축된다.

먼저 테슬라의 경우 '약 60억개의 물체'의 라벨링을 이미 완료한 상태다. 게다가 딥러닝 비지도학습의 주요 기법 중 하나인 오토인코더에 착안해 “오토라벨링”을 개발했다. 수많은 테슬라 FSD 드라이버들이 보내준 비디오 클립을 보고 시가 알아서 주행상황을 예측해 라벨링을 하는 것이다. 오토라벨링은, 일주일 안으로 1만개의 동영상 클립을 라벨링 할 수 있다고 한다. 인간이 했다면 몇 달이 걸렸을 작업량이다.

FSD 구독이 늘어나면서 드라이버 숫자가 늘어날수록, 유의미한 예외 케이스들에 대한 데이터도, 어떤 업체보다 많이 확보하게 된다. 여기서 그치지 않고, 테슬라는 현실에서는 좀처럼 보기 드문 상황들을 가정하고 이를 시뮬레이션으로 구현했다. 디지털 트윈처럼 현실과 매우 닮은 가상공간에서 FSD가 이런 상황에서는 어떻게 해야 하는지 내부적으로 학습을 시킨다. 흡사 언리얼이나 유니티 게임엔진과 같이 보이기도 한다. 테슬라의 소프트웨어 역량에 새삼 놀라게 되는 부분이다. 2021년 발표 당시에도 200가지가 넘는 특이상황을 시뮬레이션했다고 했다. 지금은 말할 것도 없을 것이다.

테슬라의 자율주행 알고리즘 개선을 위한 시뮬레이션 기법
: 현실에서 잘 발생하지 않는 예외 케이스들에 대한 데이터까지 포함해 학습하기 위함

Simulation Helps When Data

Is Difficult to Source
Is Difficult to Label
Is Closed Loop

What's Needed to Make This Happen?

5. Scenario Reconstruction

Real World Clip
Auto-Labeled Reconstruction
Recreated Synthetic World

자료: Tesla AI DAY 2021

c. 테슬라가 AI DAY에서 보인 딥러닝 기법들

그렇다면, 테슬라가 집중한 머리인 딥러닝 네트워크에 대해 좀 더 짚어보자. 테슬라는 배터리 개발 영역에서 학계에서 나온 논문을 적극 활용하는 회사로 정평이 나 있다. 그리고 자율주행 네트워크에서도 마찬가지였다. 테슬라가 AI DAY 2021 때 밝힌 하위 모델들을 정리해보면 아래와 같다. 그 당시의 거의 모든 주요한 딥러닝 모델들을 잘 활용하고 융합한 모습이다.

FSD(Full Self Driving)를 위해 테슬라가 구성한 주요 딥러닝 모델 알고리즘
 : 이미 2021년에 CNN, RNN, Attention 등 딥러닝 기법 총동원했던 테슬라

RegNet+ResNet

- RegNet: 제한된 연산 능력 안에서의 딥러닝 비전. 자동차 운전의 경우 빠른 대응이 중요하기 때문에, 제한된 시간 안에 가장 정확한 결과를 내놓는 것이 중요.
- ResNet: 딥러닝 레이어 계산을 하나 건너뛰는 것. AI 학습을 더 쉽게 하는 효과.

BiFPN

- Bidirectional Feature Pyramid Network의 줄임말. 카메라로 인식한 이미지를 기준으로 높은 해상도 픽셀과 낮은 해상도 픽셀을 합쳐서 판단. 각각 디테일한 정보와 전체적 형상 정보를 결합해 결과를 도출. 물체 식별 등에서 유용하게 쓰임.

트랜스포머

- 8대의 카메라로 받은 이미지들을 셀프어텐션 기법을 통해 3D 벡터 공간으로 변환 이를 통해 차선을 인식하고 차량의 경로를 설정. 테슬라 FSD에 있어서 가장 중요한 과정.

Spatial RNN, LSTM

- 시공간 메모리 개념으로, 과거 사건을 "기억"한 상태에서 미래 사건의 향방을 예측. 이를 통해 주변 차량의 움직임을 예상하고, 잠재적인 위험을 예측. 여러 종류 중 하나로 Spatial RNN이 사용될 수도 있음.

CNN

차량 주변 공간을 파악하는 데 사용. 입력된 이미지에서 특징을 추출해서 주변 환경을 구현. 물체 감지, 차선 감지에 도움. 이미지를 여러 픽셀의 집합으로 나눔으로써, 단순화해 해석을 쉽게 하는 기술인 "의미적 분할" 등에 사용.

HydraNet

- 데이터 세트 한 개를 가지고 여러 종류의 작업(사물 인식, 차선 인식, 신호등 인식 등)을 병렬적으로 수행하도록 함. 각 작업을 수행하는 단위가 헤드. 입력값을 각각 계산할 필요가 없어 연산량이 줄어듦. 각 헤드는 독립적으로 미세조정 가능.

자료: Tesla, 미래에셋증권 디지털리서치팀

테슬라의 FSD 시스템은 RegNet과 ResNet을 백본(backbone)으로 사용하여 이미지 특징을 추출한다. 여기서 백본이란 이미지 분류와 같이 다양한 컴퓨터 비전 작업에 사용되는 기본적인 신경망 구조를 의미한다. RegNet은 컴퓨터 비전 및 이미지 인식 작업에 사용되는 네트워크다. 연속된 하이퍼파라미터를 기반으로 최적의 네트워크 아키텍처를 탐색하는데 초점을 맞춘 네트워크다.

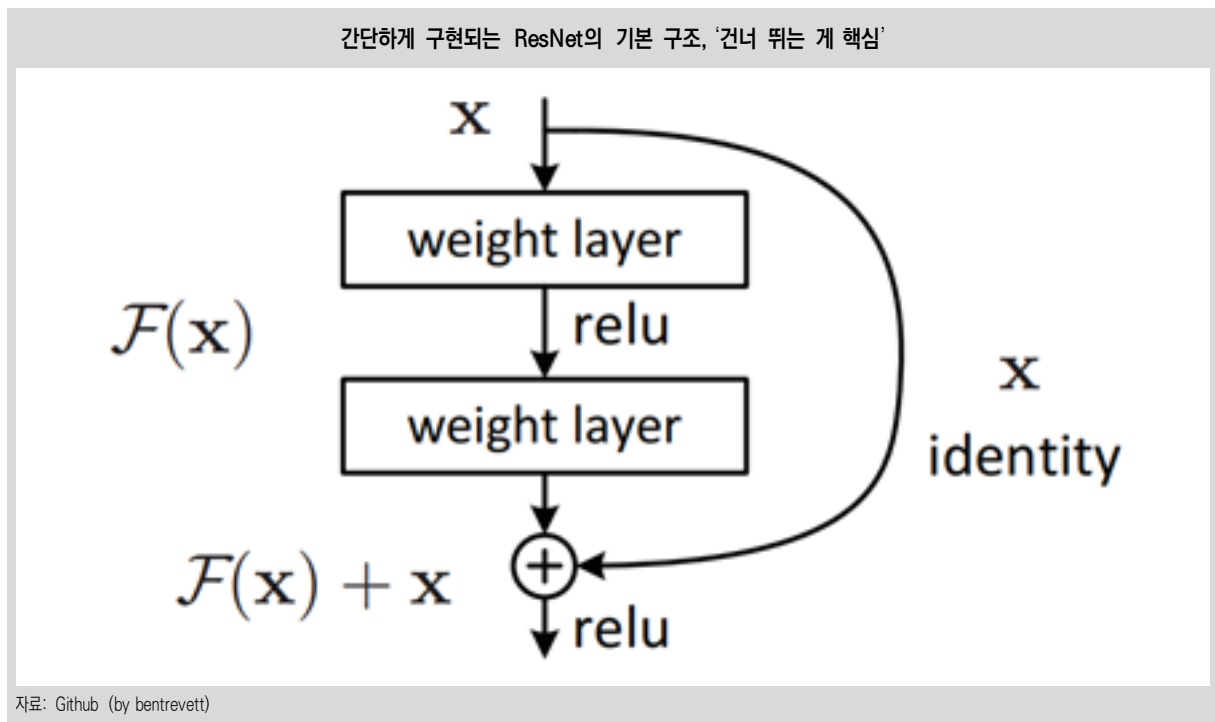
한편, RegNet은 딥러닝 아키텍처 중 하나로 페이스북 AI 리서치에서 발표한 논문이기도 하다. 테슬라 사고의 유연함이 엿보이는 부분이다. RegNet의 핵심 아이디어는 네트워크의 레이어 구조와 층을 조절해 '모델의 크기 및 복잡도를 변경'한다. 어렵게 들리지만, 결국, 높은 자원소모와 낮은 유연성을 극복하고 이해하기 쉬운 결과를 얻기 위함이다. 이를 통해 보다 빠르게 컴퓨터 비전 모델을 찾을 수 있다. 속도가 생명이라는 말이다.

그리고 ResNet은 다양한 컴퓨터 비전 작업에 사용될 수 있는 또 다른 딥러닝 네트워크다. ResNet은 2015년 He et al.이 발표한 논문 "Deep Residual Learning for Image Recognition"에서 소개되었다. ResNet은 이미지 검출과 분류에서 높은 성능을 발휘하는 딥러닝 모델이다. 이에 따라, ResNet은 차량의 주변환경을 인식하는 데 도움이 된다.

***ResNet:** Residual network의 줄임말로, 은닉층을 많이 쌓아도 성능이 좋아지지 않는 한계를 극복하기 위한 개념으로 출현. "Residual block"이라는 개념을 도입해 기울기 소실 문제를 해결한 딥러닝 아키텍처. 성능도 좋고 구조도 단순해서 유용하게 사용되는 모델

***기울기 소실 문제:** 깊은 신경망의 경우 레이어 사이에 정보가 전달되는 과정에서 기울기가 점차 사라지는 현상. 학습이 더디거나 정확도가 떨어지는 문제가 발생.

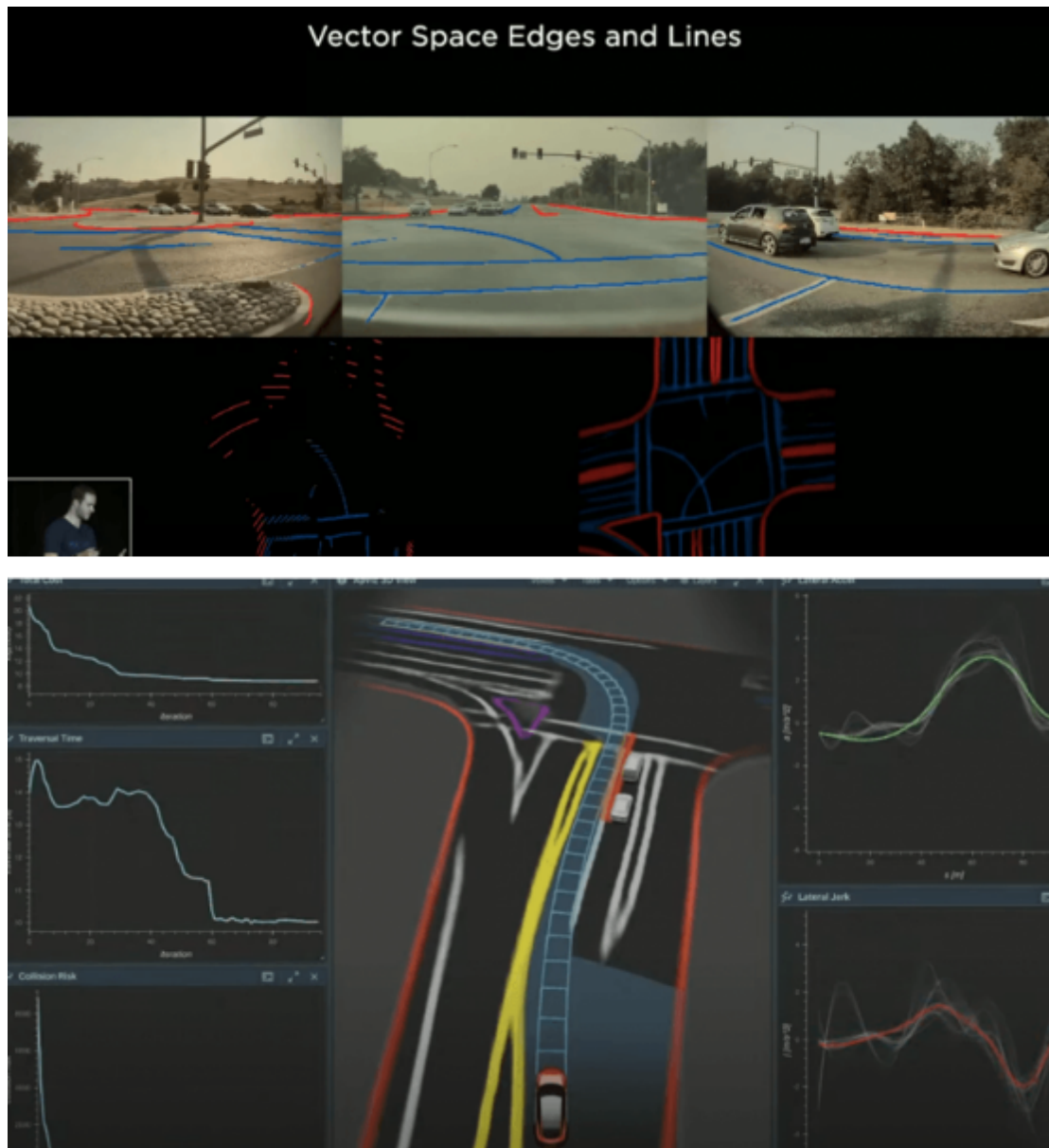
***Residual block:** 입력값을 몇 개의 레이어를 건너뛰어 출력값에 직접 더해주는 "지름길 연결(shortcut connection)"이 있는 모듈. 이를 통해 레이어 사이의 정보가 더 효과적으로 전달되고, 깊은 신경망에서도 기울기 소실 문제를 완화



서로 다른 개념 및 구조를 갖고 있는 RegNet과 ResNet이지만, 모두 컴퓨터 비전 작업에 사용되는 딥러닝 아키텍처라는 공통점이 있다. 그 이후 트랜스포머와 Spatial RNN, LSTM 등을 활용하게 된다. 이는 비전 딥러닝 네트워크로 식별하고 예측한 각 주행 데이터를 3D 벡터 공간의 일종의 출력 공간에서 표현하기 위함이다.

마지막으로 이전까지의 모든 작업을 HydraNet라는 네트워크로 처리한다. 구체적인 하류 작업(예: 객체 이미지 검출, 분류) 수행을 담당하는데, 이는 테슬라의 컴퓨터 비전 시스템에서 사용되는 신경망 아키텍처다. HydraNet은 이름에서 유추할 수 있다시피 "멀티헤드" 신경망으로 다양한 작업을 동시에 처리할 수 있는 장점이 있다. 여러 명이 어떤 문제를 함께 푸는 것으로 이해하면 쉽다. 이 아키텍처는 한 번의 순방향 패스로 여러 가지 예측을 생성할 수 있으며, 이를 통해 주행 시 발생하는 다양한 시각적 작업을 처리한다.

동영상 데이터를 가지고 3D 벡터 공간으로 만든 형태
: 3D 벡터 공간에다 시간차원까지 결합해 '미래 시점의 주행경로'를 설정



자료: Tesla AI DAY 2021

결국, 이 모든 딥러닝 아키텍처를 사용하는 이유는 각각의 아키텍처가 갖는 장점들을 활용하여 전체 시스템의 성능을 최적화하기 위함이다. 각각의 내용을 세세하게 이해하는 것보다도 중요한 점은 바로 이것이다. '대중들에게 테슬라만큼 본인들이 어떤 식으로 자율주행을 구현해내겠다고 구체적으로 설명한 회사가 있느냐'는 점이다.

그리고 2021년 8월 AI DAY가 끝나고 난 뒤 테슬라 측은 지원자 숫자가 기존보다 더 많이 늘었다고 직접 밝히기도 했다. 투자자들에게는 어려운 내용이었으나, 해당 지식을 알고 있는 전공자나 엔지니어들에게는 꿈을 실현시킬 수 있는 것처럼 들렸을 수 있다. 이런 우수한 인재들의 수급은 앞으로도 격차를 더 늘릴 것이다. 컴퓨터 사이언스는 우리 생각보다 항상 더 빠르게 발전한다는 것을 잊지 말자.



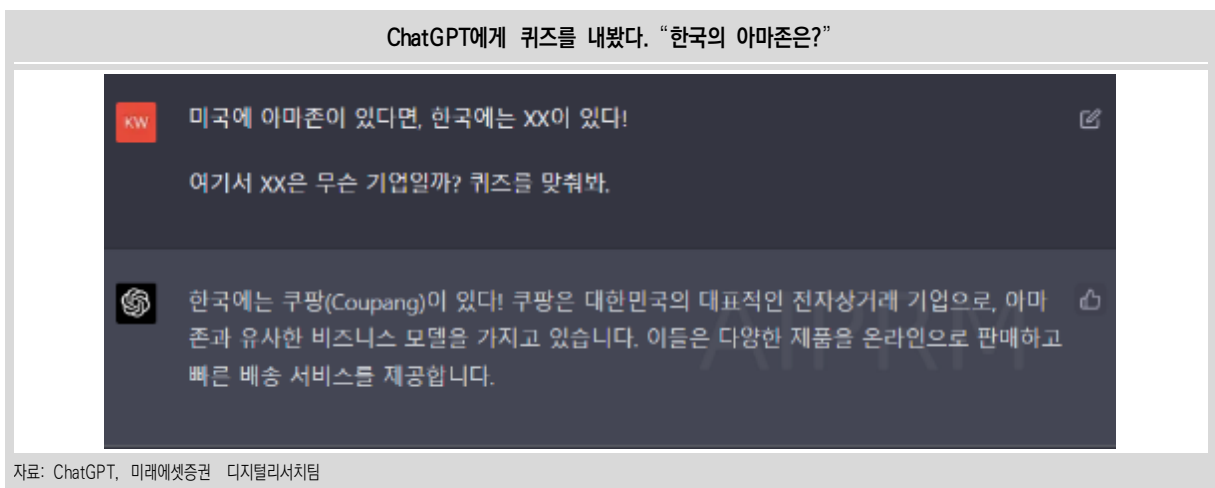
테슬라는 기본적으로 자동차 회사다. 그리고 자율주행 딥러닝 모델을 자체적으로 처음부터 끝까지 개발한 유일한 자동차 회사다. 실제 주행중인 양산차량에 유의미한 수치의 자율주행 기능을 배포하고 있는 업체는 지금도 테슬라가 유일하다. 자율주행의 격차는 줄어들기커녕 오히려 계속 늘어났는지도 모르겠다.

생성형 AI로서 인간의 생산성이 올라간다면, 자율주행은 인간의 이동성을 올린다. 두 기술 중 어떤 것이 우월하고 더 중한 것이라 말하기는 어렵다. 그리고 자율주행에 있어서는 가장 앞서 있는 업체로 테슬라를 주목하자.

AI는 비용을 줄이기 위한 최고의 선택

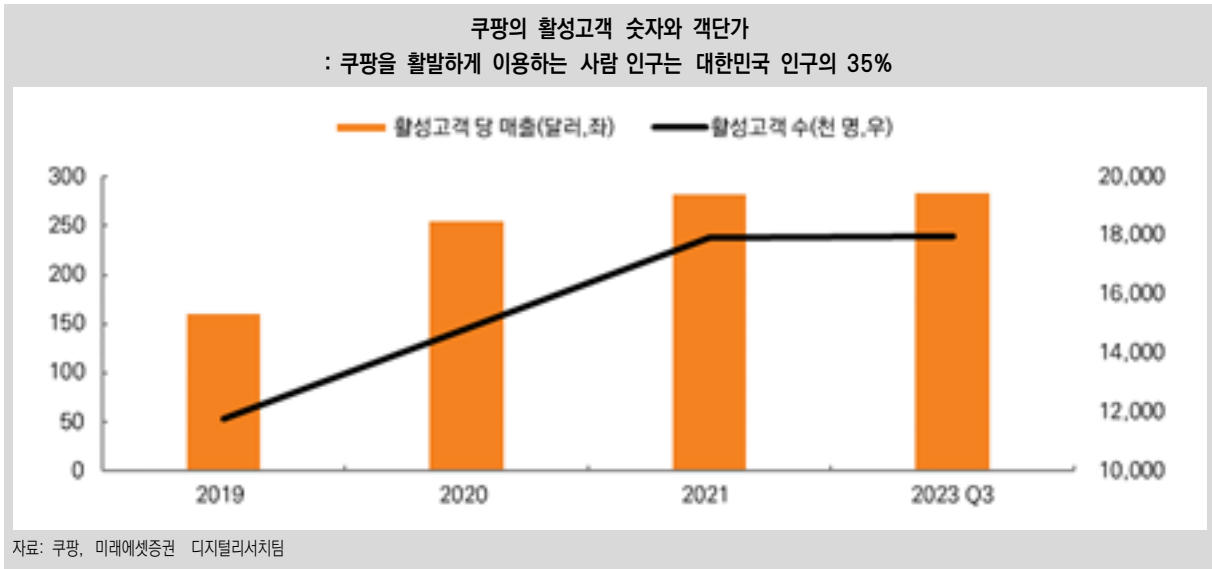
1) 쿠팡의 마진 개선을 이끄는 인공지능

“미국에 아마존이 있다면, 한국에는 XX이 있다.” 여기서 XX는 무엇인지 ChatGPT에게 물어봤다. 생성형 AI 언어모델의 선두 기술답게 바로 쿠팡이라고 대답한다. 인공지능도 알고 있다시피, 쿠팡은 아마존의 비즈니스 모델과 상당히 닮아있는 것은 많이를 알고 있을 것이다. 그리고, 아마존이 풀필먼트 자동화에 사활을 걸었듯이 쿠팡도 마찬가지다. 쿠팡의 자동화와 그로 인한 마진 상승을 이끌고 있는 것은 다름아닌 인공지능이다.



결론부터 말하면 쿠팡은 빅데이터 기업이다. 데이터를 대량으로 보유했다는 말이다. 쿠팡은 포브스가 선정한 2022년 한국인이 가장 사랑한 모바일 어플리케이션 순위에서 4위를 하기도 했다. 유튜브, 카카오 톡, 인스타그램 그 다음 자리를 차지한 것이다. 쇼핑 어플인데도 이 정도 기록을 세운 건 대단한 일이다. 로켓배송과 같은 쿠팡의 빠른 배송과 여러 서비스들에 감화된 탓일 것이다.

또한, 쿠팡을 이용하는 고객들의 숫자는 해마다 늘어나고 있다. 3개월(1개 분기) 동안 한 번이라도 쿠팡에서 물건을 사 본 사람들을 활성고객이라고 하는데, 쿠팡의 활성고객 숫자가 작년 기준 거의 1,800만명에 육박했다. 이는 우리나라 전체 인구의 35%에 해당하는 숫자다. 이 숫자가 한 번도 줄지 않고, 계속 늘어나고 있다. 이들이 만들어내는 데이터량도 방대할 것임을 쉽게 예측할 수 있다.



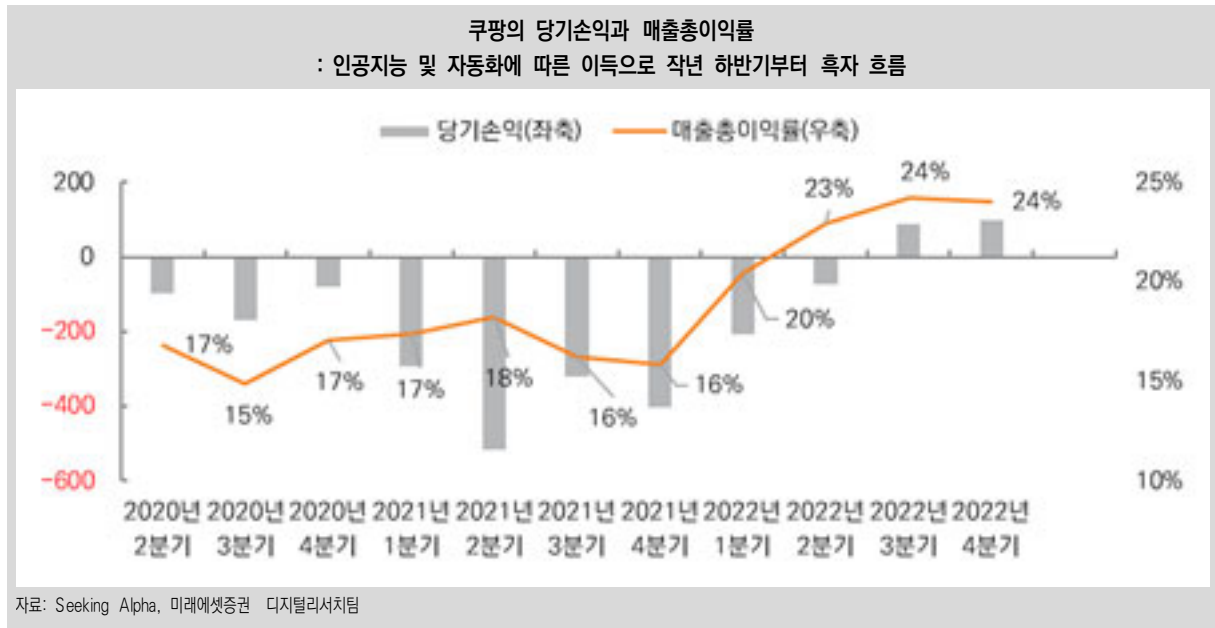
사용자 수가 누적되면서 쿠팡은 앱 사용패턴 데이터, 주문 및 결제 데이터, 물류와 배송 데이터 등 고객들의 정보를 대량으로 확보하고 있다. 데이터를 보유하고 있는 것은 곧 인공지능 서비스 향상에 있어 핵심적인 경쟁력이라고 할 수 있다.

쿠팡의 김범석 대표는 2022년 3분기 실적발표에서, “머신러닝을 활용해서 수요예측과 재고 관리를 최적화했고 이를 통해 수익성을 크게 개선했다”고 발언한 바 있다. 실제로 쿠팡의 작년 3분기 매출총이익률은 역대 최고치인 24%를 달성했었다.

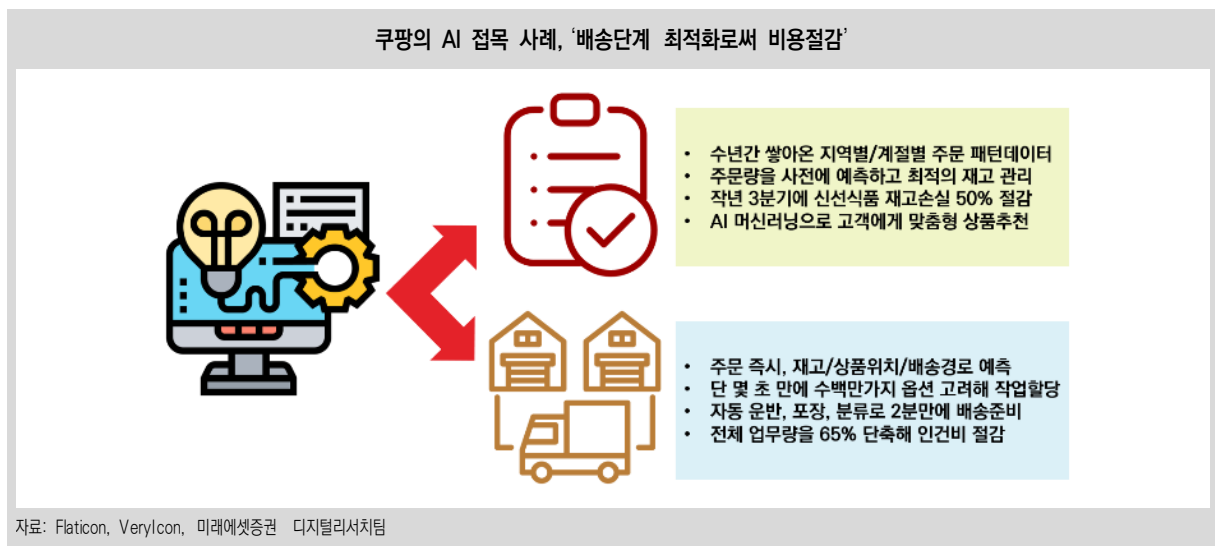


실제로 쿠팡에서는 모든 배송단계에서 인공지능을 통해 최적화를 꾀하고 있다. 수년간 쌓아온 지역별, 또 계절별 주문 데이터를 가지고 주문량을 사전에 예측하고 재고관리를 최적화한 것이다. 그래서 작년 3분기에만 신선식품 재고손실 금액을 반으로 절감했다고 밝혔다.

뿐만 아니라, 배송과 물류 분야에서도 인공지능이 대거 활용되고 있다. 고객으로부터 주문을 받은 즉시, 재고유무와 재고위치, 또 배송경로를 예측해서, 단 몇 초만에 수백만가지 옵션을 고려해서 작업을 분배시키는 방식이다. 이에 따라 단 2분만에 재고 운반, 포장, 그리고 분류 등 배송 직전까지의 준비를 마친다고 한다. 이에 따라 전체 업무량을 65%까지 단축했다. 인건비와 불필요한 운영비를 크게 낮췄다.



쿠팡이 작년 3월에 3,200억을 들여 준공한 대구 물류센터는 이런 인공지능이 활용되는 표본이다. 포장과 물건 분류, 재고 선반 이동을 모두 인공지능 기술이 장착한 로봇이 담당한다. 바닥에 있는 QR코드를 읽으면서 이동하는 방식이다. 자동화를 통해 최적화를 하니 비용이 줄어드는 건 당연하다. 실제로 작년 3분기에는 고용인원을 더 늘렸음에도 불구하고 영업비용이 크게 개선되기도 했다. 따라서, 구조적인 마진 개선이라고 판단한다.



아마존의 풀필먼트 시스템 자동화와 그 방식에 접근하려는 쿠팡의 경영방식은 시간이 갈수록 경쟁력이 극대화되어 다른 기업들이 따라오기 힘들어진다. 빅데이터를 보유했다는 것은 이렇게 강력한 해자로 활용된다는 것을 유념할 필요가 있다. 딥러닝 시대에 데이터는 전략자원이자 경쟁력을 결정짓는 모든 것의 출발점이라고 할 수 있다.

2) 광고와 마케팅

딥러닝을 사용하게 되면 마케팅과 같은 주요 분야에서 시간과 비용을 절약할 수 있다. 보통 리테일 기업들이 매년 비용 중 약 14%를 마케팅에 투자하고 있다는 점을 고려하면 실로 어마어마한 기회다. 딥러닝이 마케팅에 효과적이라고 보는 가장 큰 이유는, 다름 아닌 개인화 작업이다. 우리가 챗봇과 대화를 많이 하면 할수록 챗봇이 나를 잘 이해한다는 생각이 드는 것처럼, 딥러닝 기술은 사용자가 원하는 니즈를 정확히 판단해 줄 것이다.

여기에 적합한 비즈니스가 광고 산업이다. 앞으로 딥러닝 기술을 통해 개인화된 광고가 우후죽순 나타날 것으로 예상된다. 우선, 딥러닝을 사용하게 되면 타겟 고객층을 산정하는 데서, 훨씬 더 세밀한 고객층 분리(segmentation)가 가능해진다. AI로 학습된 개인화된 정보가 쌓인다면 적합한 정보가 적재적소의 고객들에게 갈 수 있을 것이다. 이는 회사와 고객 모두가 '윈윈(Win-Win)' 할 수 있는 방법이다. 회사는 광고수입을 고객은 맞춤형 광고를 통해 개인의 특화된 수요를 충족시킬 수 있기 때문이다. 개인화되지 않는 광고는 지금도 스팸이나 다름없는데 AI를 통한 개인화 시대에서는 말할 필요조차 없다.

생성형 AI 기업들은 브랜드와 소매 업체들에게 시각적 광고 및 마케팅 콘텐츠를 자동 생성하는 데 도움을 주고 있다. 이 업체들은 AI를 사용하여 긴 형식의 블로그 게시물로부터 마케팅 비디오를 생성해준다. 또, 옥외광고, SNS 광고 및 디스플레이 광고용 이미지의 대체 버전을 개발하는 등의 작업까지 수행한다.

얼마 전 코카콜라는 컨설팅 회사인 Bain & Company와 OpenAI의 제휴에 서명하면서, 비즈니스 운영 및 역량을 개선하는 방법을 모색하는 것과 함께 최첨단 AI를 통한 마케팅 강화계획을 세우기도 했다. 이와 관련해 코카콜라는 OpenAI의 ChatGPT와 DALL-E 시스템을 사용하여 대규모로 개인화된 광고 문구와 이미지를 생성하고 있기도 하다.

시각 콘텐츠 제작을 빠르게 확장하면, 브랜드들이 다른 시장과 소비자에게 더 관련성 있는 콘텐츠를 제공할 수 있을 것으로 기대된다. 즉, 개인화 노력을 지원할 수 있는 게 파트너십의 이유다.

딥러닝에게 맡기면 이와 같은 광고 이미지를 수만 장은 거뜰히 생성할 수 있음
소재: 코카콜라는 ChatGPT와 DALL-E를 이용해 개인화된 광고문구와 이미지를 생성할 예정

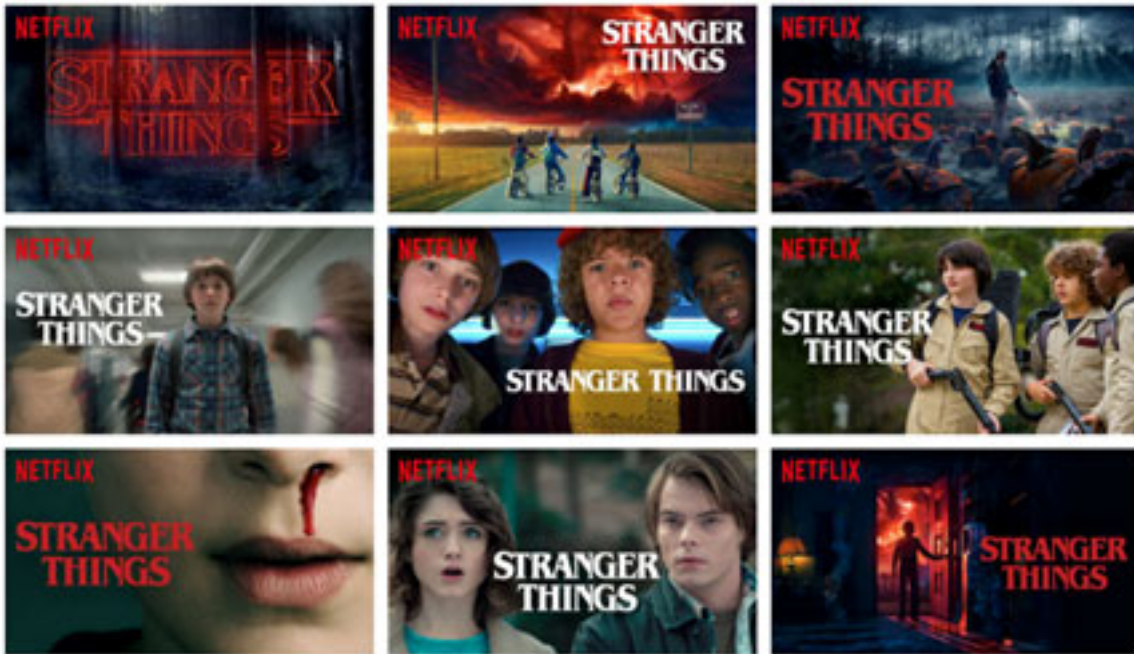


자료: The Coca-Cola Company

한편, 딥러닝을 통한 개인화 작업으로 사람들을 유인하는 기업은 훨씬 더 많이 우리 곁에 있다. 또 다른 테크 업체인 넷플릭스는 대규모의 사용자 데이터를 기반으로 인공지능을 학습시켜 개인화된 추천을 제공한다. 넷플릭스의 추천 시스템과 썸네일 이미지 선택에는 딥러닝 기술이 활용되고 있다는 말이다.

이것이 바로 넷플릭스에서 같은 영화나 드라마의 썸네일이 사용자마다 다르게 보이는 이유다. 넷플릭스는 사용자마다 개인화된 콘텐츠를 추천하기 위해 강력한 추천 알고리즘을 사용한다. 이 알고리즘은 사용자의 시청 기록, 시청 패턴 및 장르 선호 등 여러 가지 요소를 분석하여 개인에게 최적화된 콘텐츠를 추천한다. 예를 들어, 특정 사용자가 액션 영화를 좋아한다면, 액션 요소가 강조된 썸네일을 보여주는 식이다. 딥러닝을 통해 자연스럽게 시청 만족도를 높이고 더 많은 콘텐츠를 시청하도록 유도하기 위함이다.

같은 영상이라 할지라도 넷플릭스 계정마다 영화 썸네일이 다르게 전시되는 이유
: 딥러닝 기술이 당신의 시청기록과 패턴으로, 클릭을 유도하도록 추천을 수행했기 때문



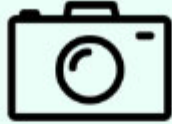
자료: Netflix TechBlog

넷플릭스의 썸네일 개인화 작업과 마찬가지로 앞으로 생성형 AI를 통한 각종 개인화 마케팅은 대세가 될 전망이다. 어떤 식으로 찾아올지는 모르지만, 마케팅과 광고 산업은 생성형 AI를 통해 큰 혁신을 맞이하게 될 것으로 예상된다. 광고는 결국 고객의 “정확한 니즈”를 파악하고 적절한 시간에 적절한 것을 제시하는 것이 본질이다. 그리고 그 본질을 제일 잘 수행할 수 있는 게 딥러닝이다.

우리는 챗봇의 출현으로 고객센터 지원 정도만을 생각했으나 활용될 수 있는 곳은 생각보다 많다. 카피라이터가 할 수 있는 일들도 많은 부분을 기계한테 맡길 수도 있고, 사내 디자이너들이 하던 일들도 일부 대체될 가능성이 크다.

비용을 줄일 수만 있다면 무엇이든 하는 게 영리 기업이다. 생성형 AI의 출현은 이들에게 황금알을 낳는 거위과도 같을 것이다. 물론 그 거위는 배를 갈라도 죽지 않는다. 기계니깐 말이다.

회사들이 마케팅을 할 때 생성형 AI를 활용할 수 있는 세부 분야



가상사진 촬영

각 브랜드들은 AI가 생성한 제품 사진 또는 패션 모델을 만들어 온라인에 게시가능



3D 제품 카탈로그

생성형 AI를 사용해 텍스트 또는 2D 이미지를 3D 제품 카탈로그로 변환해 제공



고객 서비스 지원

생성형 AI는 고객센터 서비스 에이전트가 응답 및 해결시간을 단축하도록 지원



이커머스 제품 설명

텍스트 생성도구를 사용해 더 많은 쇼핑객을 전환시킬 보다 매력적인 제품 설명을 작성



개인화된 마케팅 콘텐츠

다양한 인구 통계 및 플랫폼을 분석해, 개인화된 콘텐츠 제작으로 고객군 확장



가상 패션 디자인

생성형 AI를 사용해 2D 스케치를 디지털 렌더링으로 변환, 새로운 의류디자인 가능

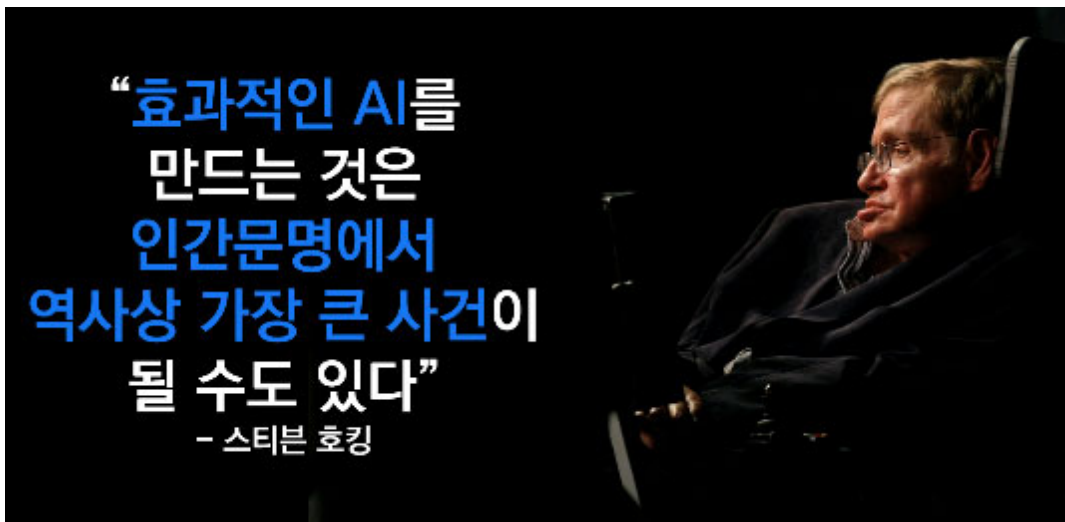
자료: CBInsights, 미래에셋증권 디지털리서치팀

IV. AI 모델 구축을 위한 주요 요소와 기회

전세계 빅테크들의 전쟁

‘AI 산업화 시대’는 곧 피할 수 없는 현실이 될 것이다. 그리고 올해는 그 원년으로 인공지능 관련 기업들이 지속적으로 주목받을 것이다. 스티븐 호킹은 일찍이 이런 말을 한 적이 있다. 효과적인 AI를 만드는 것은 인간문명에서 역사상 가장 큰 사건이 될 수도 있다는 말이었다. 호킹 박사는 살아생전 수많은 예측을 했는데, 시간이 지나면서 그의 가설이 정답으로 드러난 사례들이 많았다. 그래서 더욱 그의 말에 귀를 기울이게 된다.

2018년 작고한 스티븐 호킹이 인공지능에 대해 밝힌 예측
: 스티븐호킹은 인공지능이 인류 문명에 “모 아니면 도”라고 우려했음



자료: NBC, 미래에셋증권 디지털서치팀

우리가 이제껏 각 산업 영역에서 AI가 어떤 대격변을 불러일으킬 것인지 살펴봤다. 딥러닝은 단기적인 테마로 끝날 것이 아닌 기반 기술이자 우리 삶 그 자체가 될 것임을 믿어 의심치 않는다. 그리고 이미 기업들은 AI 산업화 시대에 발맞춰 본인들의 입지를 다지기 위해 발 빠르게 대응하고 있다. 아래는 최근 생성형 AI를 도입한 글로벌 기업들의 사례를 단편으로 정리해본 것이다.

- (1) **마이크로소프트**는 OpenAI의 ChatGPT와 동사의 New Bing을 병렬적으로 운영. ChatGPT는 현재 다양한 파트너들을 유치해 플러그인 기능 도입을 준비 중. New Bing의 경우 마이크로소프트 오피스 365와 결합한 “코파일럿(Copilot)” 기능을 곧 선보인 바 있음.
- (2) **구글**은 언어생성 모델인 Bard를 클로드 베타 형식으로 출시. 온라인 광고 서비스인 '구글 애즈(Ads)'에도 생성형 AI를 도입. 광고 이미지나 영상, 제목, 로고, 설명 등 세부 자료를 AI로 쉽게 만들어 주는 형태. 또한, 유튜브에 크리에이터를 위한 생성형 AI 도입이 몇 달 후 시행될 예정.
- (3) **메타**는 생성형 AI 시장에 본격적으로 뛰어 들 것임을 시사. 챗봇같은 제품을 내놓는 대신 기반이 되는 대형언어모델(LLM)을 공개하는 전략을 택함. 모델의 이름은 라마(LLaMA)

(4) 중국의 바이두와 알리바바에 이어 텐센트도 AI 챗봇 개발 경쟁에 잇따라 합류. 텐센트는 ChatGPT와 비슷한 챗봇을 개발하기 위해 팀을 구성한 상태. 이 소식은 중국 과학기술부가 ChatGPT와 같은 기술의 잠재력을 보고 AI를 중국 사회와 경제에 통합할 것이라고 발표한 뒤 나온 것.

(5) 엔비디아는 GTC 2023에서 본인들을 “AI Factory”라고 지칭하며 AI 산업 내 동사의 입지를 부각

(6) 모바일 AP 팹리스 전문 퀄컴이 스테이블 디퓨전을 클라우드에 연결하지 않고 휴대폰 성능만을 이용해 실행할 수 있도록 구현. 스냅드래곤 칩셋에 이런 기능을 탑재한 것.’



전세계 테크 기업들은 영역을 막론하고 AI 딥러닝을 본인들의 비즈니스에 서둘러 도입하고 있다. 딥러닝 기술 구축을 위해서 가장 중요한 요소들로 세 가지를 꼽을 수 있다. 먼저 딥러닝 네트워크를 개발할 수 있는 소프트웨어적 능력이고, 두 번째는 해당 네트워크를 가득 채울 원천 소스인 거대한 데이터가 필요하다. 그리고 마지막으로 딥러닝 특성상 엄청난 규모의 연산력을 필요로 하기 때문에 반도체 역량이 중요하다. 특히 고성능과 저전력을 내세운 반도체 설계능력을 말한다.

중요도를 매기는 것이 쉽지는 않지만, 세 가지 중에서도 가장 중요한 것은 데이터 보유량과 해당 데이터의 품질이라고 우리는 판단한다. 충분한 데이터가 없으면 딥러닝이라는 것도 무용지물이기 때문이다. 현재의 AI 산업화 시대를 이끌고 있는 것도 LLM, 즉 대형 언어모델이기도 하다.

한편, 일찍이 중국에서도 데이터의 중요성을 인식하고 있었다. 2020년 전국인민대표회의에서는 데이터안전법 초안을 공개했다. 이름에서 알 수 있듯이 국가적 데이터 관리의 중요성을 강조한 것이다. 중국 공산당에서는, 경제학의 3대 생산요소에 데이터를 또 하나의 “요소”로 추가했다. 노동, 토지, 자본이라는 요소에다 데이터를 추가한 것을 볼 때 얼마나 이들이 데이터를 중요시 여기는지 알 수 있다.

우리들은 정보의 홍수 속에 살아간다고 하지만, 인공지능 개발을 위한 양질의 데이터 주권은 오히려 희소해지는 ‘역설의 시대’가 올 수 있을지 모른다.

그리하여, 양질의 빅데이터를 갖고 있는 기업은 핵심 경쟁력을 갖게 될 것으로 판단한다. 가장 선도적인 위치를 점유하고 있는 것은 역시 빅테크 기업들이다. 마이크로소프트가 쏘아 올린 거대한 공(功)은 각 기업들의 운명을 결정지을 수도 있는 거대한 사건이다. 각 빅테크들이 어떤 기술적/사업적 전략을 가지고 있고, 다른 기업들은 어떻게 대응하고 있는지를 살펴보려고 한다.

(1) 마이크로소프트 (feat. OpenAI)

마이크로소프트(MS)는 지난 3월 10일, AI 챗봇을 장착한 Bing의 이용자가 한 달 만에 1억명을 돌파했다고 자랑스럽게 말했다. 이용자 3분의 1은 Bing을 처음 사용하는 사람이라는 점이 중요하다. 인간 수준의 채팅 봇 인터페이스가 효과적으로 작동함을 사람들이 생각보다 더 빠르게 인지하기 시작한 것이다.

검색 엔진은 키워드 기반으로 방대한 정보를 제공하기 때문에, 사용자는 필요한 데이터를 얻기 위해 정보를 필터링하는 작업을 거쳐야 했다. 그러나 ChatGPT의 등장으로 정보의 취사선택 과정이 크게 간소화되었다. OpenAI의 공동 창업자이자 CEO인 샘 알트만은 딥러닝 언어모델의 발전으로 인해, 몇 년 안에 구글의 검색 시장이 큰 도전을 맞게 될 것이라고 주장했다.

구글뿐만 아니라 최근 검색 플랫폼으로 많이 사용되는 유튜브 영상조차도 검색에 대한 결과물이 고정된 정적 데이터이다. 그런데 OpenAI의 경우 묻는 말에 따라서 새로운 데이터가 '생성'된다. 너무 그럴듯하게 글을 써내려 가기 때문에 모방과 창조의 경계를 정확히 긋기조차 힘들게 한다. ChatGPT의 출현으로 이제 우리가 검색하는 정보는 더이상 정적인 데이터가 아니게 된다. 동적인 데이터이기 때문에 질문자의 명령어(Prompt)에 따라 개인화된 작업이 가능하다. 그리고 과거 명령어를 기반으로 한 문맥이 챗봇에서 다 남아있기 때문에 훨씬 더 세밀화된 개인화가 이루어지고, 이는 곧 락인효과를 발생시킨다.

OpenAI는 마이크로소프트 Azure 클라우드에 올라타 날개를 펼치고 있다. 2019년 7월, 마이크로소프트와 OpenAI는 인공지능 개발을 가속화하기 위해 독점적 파트너십을 맺었는데, 사실상 이때부터 마이크로소프트가 해당 기술에 대한 배타성을 확보했다고 할 수 있다.

마이크로소프트가 GPT-4를 기반으로 내놓은 챗봇 "New Bing"
: ChatGPT처럼 거짓말을 하지 않고 모르는 것은 모른다고 대답

The screenshot shows a chat interface with the following content:

- Search bar: BART의 장단점은 무엇인가요?
- Status: ✓ "BART 장단점" 검색 중, ✓ 답변을 생성하는 중...
- Response: 죄송합니다. 제가 찾은 정보 중에 BART의 장단점에 대한 내용이 없습니다. 다른 질문을 해주시겠어요?
- Source information: 자세한 정보: 1. github.com 2. blog.naver.com 3. blog.naver.com (13 / 15)
- Follow-up questions: BART가 어떻게 작동하나요?, BART와 다른 자연어 처리 모델들과 비교해주세요, BART가 어떤 문제들을 해결할 수 있나요?
- Input field: 새 토픽, 무엇이든 물어보세요

자료: Microsoft Bing, 미래에셋증권 디지털리서치팀

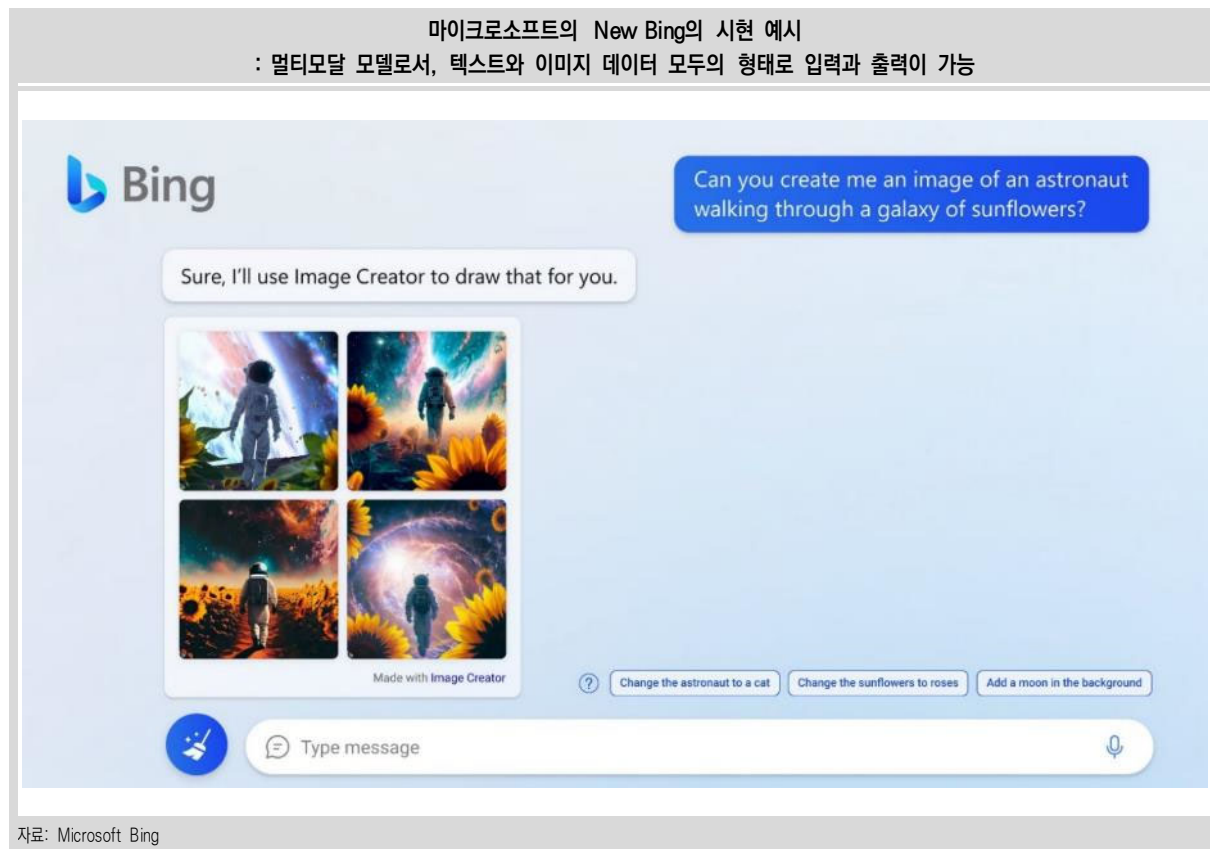
MS의 New Bing 검색 엔의 강점은 현재의 정보까지 다 반영된 ChatGPT 결과를 나오게 한다는 점이다. 기존의 ChatGPT가 창조적으로 재구성하는데 집중되어 있다 보니 종종 잘못된 정보를 제공하기도 하는데, New Bing은 이를 크게 보완했다. New Bing은 사실에 입각한 정보를 중시하는데, 이를 위해 보다 개선된 알고리즘을 적용한 프로메테우스 엔진을 사용했다고 한다.

New Bing도 GPT-4 언어모델을 통해서 만들어진 것이다. 사전학습된 GPT 언어모델을 튜닝하는 작업을 통해 다운스트림 태스크인 검색엔진의 역할을 극대화해야 했다. 즉 프로메테우스 엔진은 모델 전체를 업데이트하는 과정인 미세조정을 거친 것이다.

a. 멀티모달

ChatGPT와 Bing이 자아낸 충격이 채 가시기도 전에 또 다른 기술이 세상에 나왔다. 바로 멀티모달 (multimodal)이다. 멀티모달은 여러가지 유형의 데이터를 동시에 처리하는데 초점을 맞춘 딥러닝 기술이다. 다시 말해, 텍스트, 이미지, 영상 데이터를 동시에 분석하고 또 데이터 유형 간 관계도 복합적으로 이해해서 공통되는 문맥을 발견해 낸다. 마치 인간이 오감을 통해 세상을 인식하는 것과 동일한 접근 방식이다. 그래서 멀티모달 방식은 인간 수준에서 일반 작업을 수행할 수 있는 범용인공지능(AGI)을 구축하기 위한 핵심 기술로 꼽히기도 한다.

OpenAI의 수장 샘 올트만은 진정한 멀티모달 모델이 곧 작동하게 될 것이라고 말했다. 그리고 해당 발언이 나온 지 몇 달이 채 되지 않아 멀티모달은 세간의 중심에 섰다. 마이크로소프트와 구글이 경쟁하듯 멀티모달 개념을 꺼낸 것이다. OpenAI도 멀티모달에 중점을 두고, GPT-4를 기반으로 ChatGPT의 가장 최신 버전을 베타서비스로 출시했다.



자료: Microsoft Bing

즉, GPT-4도 멀티모달이기 때문에 이미지와 텍스트 두 가지의 모달리티를 입력 받고 텍스트 출력이 가능하다. 텍스트와 이미지 데이터를 함께 인코딩할 수 있는 모델이라는 뜻이다. 즉, 하나의 메커니즘으로 여러 종류의 데이터를 처리할 수 있음을 의미한다.








이러한 모델은 시각적 단서를 음성과 결합해 더 자연스러운 대화 시스템을 구축하거나, 글과 통합해 사진과 글 사이의 관계를 파악한 해석 및 번역 등을 가능하게 할 수 있다. 또한, 음성뿐만 아니라 영상이나 이미지 등으로 답변해 이해를 돕는 어시스턴트로 사용되는 등 많은 잠재적 응용 분야를 갖는다.

GPT-4는 성능 개선을 위해 미세조정을 거쳤다. 이는 모델의 출력 확률을 조정하는 작업을 통해 ‘답변 자신감’이 실제 정답일 확률과 일치하도록 하는 과정이다. 이 과정을 거치면 모델의 사실성(factuality)과 준수(adherence)가 개선된다. 여기서 사실성이란 AI 모델의 출력이 사실에 부합하는지를 나타내는 지표이다. 사실성이 높은 모델은 오류나 편향 없이 정확한 정보를 제공해 신뢰성을 높인다. 준수는 AI 모델이 사용자의 질문에 적절하게 응답하는 정도를 나타내는 지표로, 준수가 높다는 것은 사용자의 의도를 시가 정확하게 이해하고 적절한 결과를 제공하고 있다는 것을 말한다. 모델과 출력 확률을 세밀하게 조정함으로써 GPT-4는 더 정확하고 신뢰성 있는 결과를 생성할 수 있게 된다.

이로써, GPT-4는 기존 대규모 언어모델을 비롯한 대부분의 최첨단 시스템에 비해 전통적인 자연어처리 성능평가에서 높은 성능을 입증했다. 심지어는 변호사 시험 시뮬레이션에서 상위 10% 수준의 성적을 내는 등 전문적 벤치마크에서의 평가도 우수했다.

이처럼 텍스트만 입력 받아야 했던 대형언어모델(LLM)의 한계를 극복하고 다양한 입력값을 받을 수 있는 멀티모달 대형언어모델(MLLM, Multimodal Large Language Model)이 앞으로도 대세가 될 것으로 우리는 전망한다. 그리고, 마이크로소프트가 OpenAI의 개입 없이 독자적으로 구축한 MLLM인 코스모스-1(Kosmos-1)이 좋은 시작이 될 것으로 보인다.

마이크로소프트가 공개한 멀티모달 대형언어모델 “코스모스-1”
: 사진을 보고 “이게 왜 웃긴 사진인지” 인간의 유머까지 이해하는 수준에 도달

Input Prompt	 Question: Explain why this photo is funny? Answer:	 Question: Why did the little boy cry? Answer:	 Question: What is the hairstyle of the blond called? Answer:	 Question: When will the movie be released? Answer:
Completion	The cat is wearing a mask that gives the cat a smile. (1)	Because his scooter broke. (2)	pony tail (3)	On June 27 (4)
Input Prompt	TorchScale - A Library for Transformers at (Any) Scale What is TorchScale?	 Question: The result is? Answer:	 Question: What is the heart rate in this picture? Answer:	 The time now is
Completion	A library that allows transformers to work efficiently and effectively at any scale. (5)	5 + 4 = 9 (6)	57 bpm (7)	10:10 on a large clock (8)

자료: Microsoft

***코스모스-1:** “Language Is Not All You Need: Aligning Perception with Language Models”라는 논문에서 최초로 공개. 대형언어모델(LLM)과 이미지 처리(perception)을 융합하는 것을 목표로 하고 있음. 코스모스-1은 텍스트, 이미지, 오디오 등 여러 종류의 데이터를 인코딩하기 위해 single embedding module을 사용.

파라미터는 16억 개로, 다른 LLM이나 이미지 추론 모델에 비해 개수는 적은 편. 모델의 학습을 위해 싱글모달(텍스트), 쌍모달(이미지와 캡션), 교차 멀티모달(텍스트 문서와 이미지) 등 다양한 조합의 데이터를 입력값으로 사용. 그 뒤, LLM과 같이 다음에 올 토큰(단어)을 예측하는 방식으로 훈련. 훈련이 끝나면 명령 수행 능력을 개선시킬 수 있는 특별한 데이터 세트로 튜닝 과정을 거침.

코스모스-1은 이미지 안에 내재된 텍스트를 처리하거나, 이미지를 분류하거나, 웹사이트 콘텐츠에 대한 질문에 답하는 능력이 기존 모델 대비 우수한 것으로 나타났다. 특히, 레이블 지능검사에서 관련된 예시가 이전에 입력된 적이 없는 문제도 정답을 맞출 수 있었다. 지각 정보와 언어모델을 조화시킬 수 있기에 가능했던 것이다.

b. ChatGPT와 Bing의 생태계 구축전략

코스모스-1의 사례에서 볼 수 있듯이 마이크로소프트는 OpenAI와 협업함과 동시에 독자적인 AI 생태계를 구축하는 전략을 택한 것으로 보인다. ChatGPT와 Bing이 병렬 형태로 동시에 운영되고 있는 것처럼 말이다. 스타트업인 OpenAI의 실험 정신(spirit)을 존중한다고도 볼 수 있다.

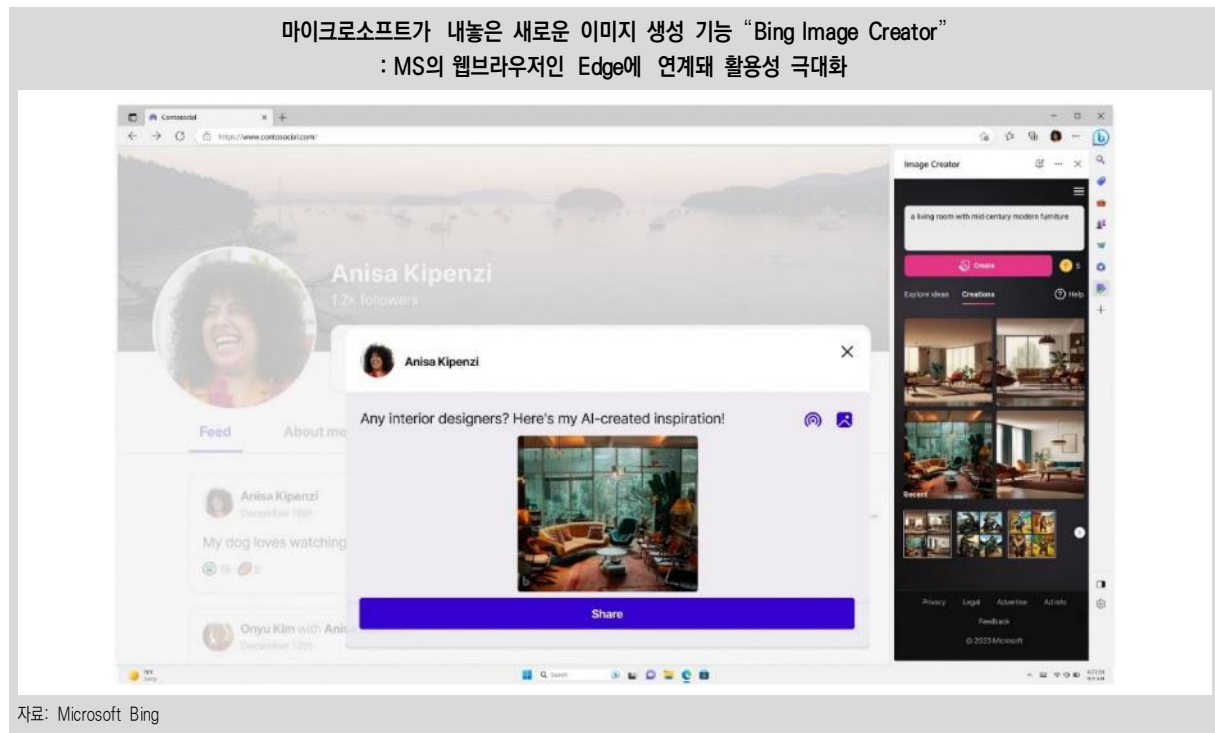
마이크로소프트가 OpenAI와 AI 비즈니스를 병렬로 가져가면서, 각각 생산성 솔루션과 소비자 어플리케이션 영역으로 전문화하는 전략으로 풀이된다. OpenAI의 자율성을 존중하면서, 생태계를 각자 늘려가는 형태로 얻을 수 있는 장점은 아래와 같다.

ChatGPT와 Bing 병렬로 운영되는 MS의 사업전략

ChatGPT	Bing
<ul style="list-style-type: none"> - 스타트업인 OpenAI가 운영 - 주기적 GPT 모델 알고리즘 업데이트 - 플러그인(Plug-In)으로 각종 인터넷 서비스들을 포섭해 인터넷 생태계 통합 전략 - 외부의 솔루션 및 어플리케이션 기업들과 독자적 파트너십 체결해 생태계 확장 	<ul style="list-style-type: none"> - 빅테크인 마이크로소프트가 운영 - 오피스 365와 결합하는 “코파일럿” - 코파일럿을 사용하기 위한 Azure 클라우드 라인 효과 극대화 - 코파일럿에 최적화할 수 있도록 스마트 기기 등 HW 판매 가능성도 존재
일상적 “소비”에 집중	“생산성” 증대에 집중

자료: 미래에셋증권 디지털리서치팀

먼저 마이크로소프트가 당장 집중하고 있는 분야는 검색 엔진과 웹 브라우저 시장이다. 둘 모두 알파벳이 구글과 크롬으로 꽉 잡고 있는 영역인데, 언어모델의 출현으로 최근 그림이 약해졌다. 이 틈을 놓치지 않고 마이크로소프트는 Bing을 출시했고, 멀티 모달 기능을 선보였다. 강력한 원투펀치에 구글이 카운터를 날릴 새도 없이, 마이크로소프트는 며칠 지나지 않아 'Bing Image Creator'라는 서비스까지 출시했다.

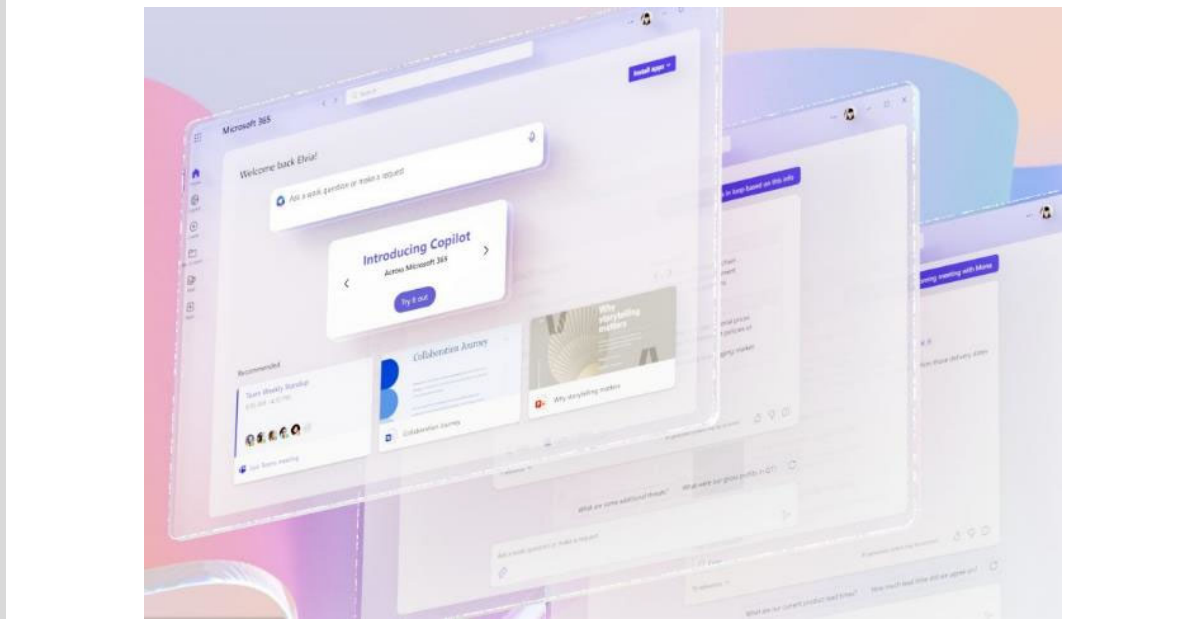


Bing Image Creator는 이미지 생성형 AI를 수행할 수 있는 기능이다. 특히 Bing 챗봇을 이용하면서 챗봇에 본인이 원하는 그림을 설명하면 되는 식이다. Bing 자체가 마이크로소프트의 웹 브라우저인 Edge에서 사용할 수 있기 때문에 Bing Image Creator도 Edge에 종속되는 서비스다. 사람들은 검색 기능을 사용할 때 일반 웹 검색 다음으로 이미지 검색을 많이 한다. 따라서, 이미지 생성형 AI는 이런 측면에서 꽤나 강한 잠재력을 가지고 있다. 앞으로 마이크로소프트의 Edge가 구글의 크롬 생태계에 어떤 균열을 낼 수 있을지 지켜보는 것이 관건이다.

그리고 업무 생산성 증대를 위해 마이크로소프트는 "코파일럿(Copilot)" 기능을 선보였다. 대형언어모델(LLM)과 비즈니스 데이터, 오피스 365를 결합한 방식이다. 사티야 나델라 마이크로소프트 CEO는 "우리가 일하는 방식을 근본적으로 바꾸게 될 것"이라고 말했다. 결국 코파일럿은 인간의 자연어에 맞게 모든 업무 도구를 맞췄다는 데서 의의를 가진다. 아래는 코파일럿으로 우리가 할 수 있는 일을 요약한 것이다.

- (1) 워드: 글 작성, 편집, 요약, 창작을 보조. 코파일럿이 초안을 수정하고 재작성하는 데 도움
- (2) 파워포인트: 자연어 명령으로 아이디어 차원을 프레젠테이션으로 바로 만들어줌.
- (3) 엑셀: 인사이트, 동향 파악, 데이터 시각화를 제공하고, 자연어로 데이터세트에 질문 가능
- (4) 아웃룩: 편지함 관리와 이메일 요약, 답장 초안 제안 등으로 업무 생산성을 향상
- (5) 팀즈: 미팅 내용을 요약하고 놓친 부분을 알려주며, 필요한 행동 항목을 제안
- (6) 파워플랫폼: 개발자들이 로우코드 도구로 개발을 가속화하고 능률화할 수 있도록 지원
- (7) 비즈니스챗: 사용자의 모든 앱 데이터를 활용해 다양한 작업을 지원

마이크로소프트는 모든 업무 생산성 도구에 대형언어모델을 결합

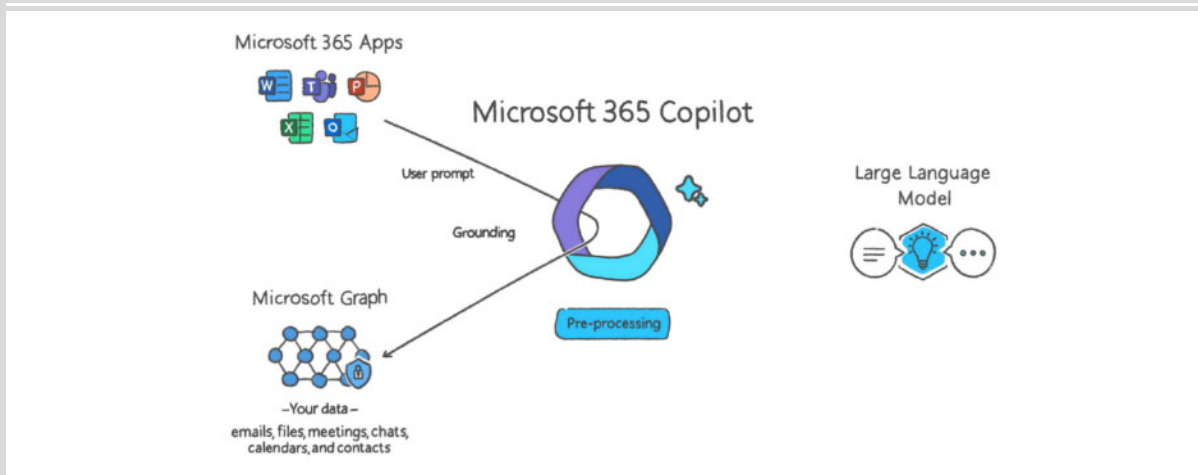


자료: Microsoft

코파일럿의 핵심은 “비즈니스챗(Business Chat)”이다. 비즈니스챗을 활용하면 사용자의 고유 데이터(문서, 프레젠테이션, 이메일, 채팅, 연락처, 캘린더 등)를 기반으로 원하는대로 리포트를 작성할 수 있다. 예를 들어, 사용자가 “Mr. Han에게 어떤 제품을 적극 마케팅해야 하는지 팀원들에게 알려줘”라고 비즈니스챗에게 요청한다고 해보자. 그러면, 비즈니스챗은 Mr. Han과의 이메일 내용, 채팅 히스토리, 팀원들간의 회의 내용을 전부 가져와 상황 리포트를 생성한다. 정말 똑똑한 인턴이 생기는 것이다. 이는 대형언어모델인 LLM과 마이크로소프트 365 앱에 사용자 개인 데이터가 유기적으로 연결되어 있기에 가능하다.

이로써 창출되는 효율성 향상은 엄청날 수 있으나, 개인정보보호에 관한 우려가 제기될 수 있는 여지가 있다. 사용자 데이터에 액세스하는 플러그인이 예기치 않게 개인정보를 침해할 수 있기 때문이다. 특히 비즈니스 차원의 민감한 정보가 노출될 위험이 크다. 따라서 이 부분으로부터 어떻게 본인들을 방어하고 사용자들을 설득할 수 있을지가 대중화를 위한 가장 큰 고민거리가 아닐까 판단한다.

코파일럿의 구동 원리: 사용자의 개인정보에 접근해 업무 효율을 올리는 “마이크로소프트 그래프”가 핵심



자료: Microsoft










한편, ChatGPT의 행보는 조금 더 공격적이다. 스타트업답게 OpenAI는 ChatGPT의 생태계를 확장하기 위해 민첩하게 움직이고 있다. 며칠 전 OpenAI는 ChatGPT이 어떻게 확장될 수 있는지 그 비전을 보여줬다. 지난 3월 24일 OpenAI는 ChatGPT용 플러그인을 출시했다.

이 의미를 한 마디로 설명하면, “ChatGPT가 이제 인터넷을 검색할 수 있다”는 것이다. 즉, 챗봇의 기능을 확장하고 웹상에 있는 다른 지식 데이터에도 접근할 수 있게 되었다. 아직은 ChatGPT Plus 유료 구독 고객들을 대상으로 시작하여 점진적으로 서비스를 실시할 계획이다.

플러그인이라는 기능은, 사전 그대로 의미를 해석하면 “꽃아서 (쓰다)”라는 것이다. 그러니까 ChatGPT 안에 여러 어플리케이션과 해당 액세스 권한이 꽂힌 상태로 편리하게 사용할 수 있다는 정도로 이해하면 된다. ChatGPT가 원래 “몰랐던” 영역에 대해 다가갈 수 있게 된 것이다. OpenAI에 따르면, ChatGPT를 출시한 이후 많은 사용자들이 플러그인 기능을 요청해 왔다고 한다. 플러그인 기능을 통해 ChatGPT의 활용도를 극대화할 수 있기 때문이다.

예를 들어, 이번에 OpenAI가 공개한 첫 번째 플러그인 대상 어플리케이션들은 Expedia(여행 예약), FiscalNote(법률 플랫폼), Instacart(식자재 주문), OpenTable(식당 예약), Shopify(온라인 상점 개설), Slack(공동 업무 생산성), Wolfram(칼로리 계산) 등이 있다.

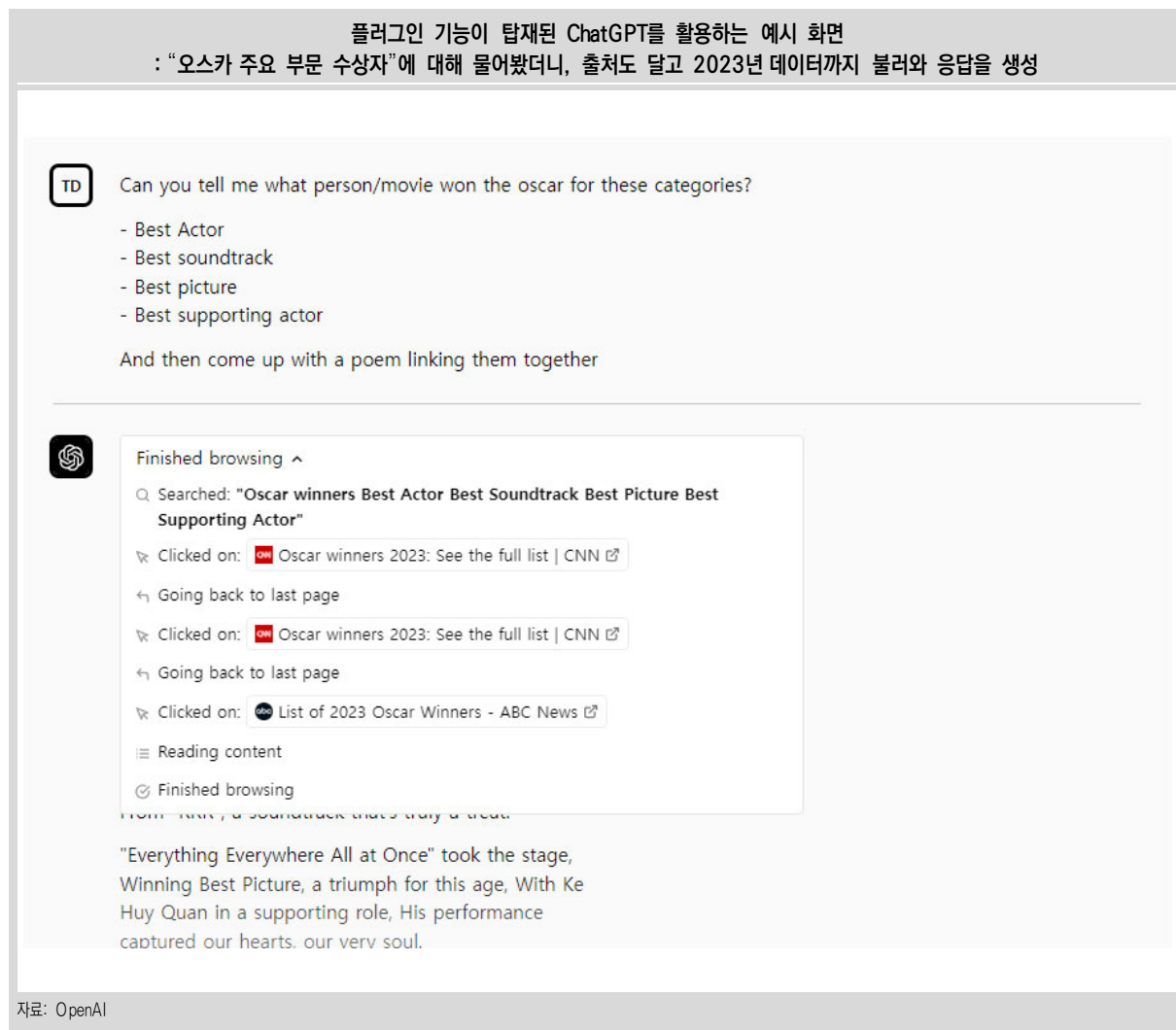
ChatGPT에 플러그인을 제공하는 회사들 리스트
: 식당예약, 칼로리계산, 음식재료까지 ChatGPT로 해결

 <p>Expedia Bring your trip plans to life—get there, stay there, find things to see and do.</p>	 <p>FiscalNote Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.</p>	 <p>Instacart Order from your favorite local grocery stores.</p>
 <p>Klarna Shopping Search and compare prices from thousands of online shops.</p>	 <p>Milo Family AI Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?</p>	 <p>OpenTable Provides restaurant recommendations, with a direct link to book.</p>
 <p>Speak Learn how to say anything in another language with Speak, your AI-powered language tutor.</p>	 <p>Wolfram Access computation, math, curated knowledge & real-time data through Wolfram Alpha and Wolfram Language.</p>	 <p>Zapier Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.</p>

자료: OpenAI

ChatGPT에 “여의도에서 가장 맛있는 소고기집을 추천해줘. 단, 1인당 5만원만 지출할거야”라고 입력하면 GPT가 이를 이해한 뒤, 식당 예약 플러그인으로 알맞은 레스토랑을 추천해주는 식이다. 원하는 레스토랑을 찾았다면 예약할 수 있는 URL 링크 주소를 제공한다. 앞으로 이런 플러그인 서비스는 계속 늘어날 것으로 보인다. 플러그인 업체 입장에서는 효과적인 모객 플랫폼을 찾았으니 이득이고, ChatGPT는 활용도를 높일 수 있으니 득이다. 사용자 입장에서 ChatGPT 안에서 필요했던 일상적인 작업들을 다 해결할 수 있으니 편리하다.

한편, 챗봇인 ChatGPT의 근본은 딥러닝 언어모델이기 때문에 정보가 2021년 9월까지만 제공되는 분명한 한계점이 있었다. 딥러닝 훈련 과정에서 모델이 학습하는 정보들은 결국 과거 데이터이기 때문이다. 하지만 이 문제도 플러그인을 통해 해결될 수 있었다. OpenAI가 웹 브라우저라는 자체적인 플러그인 기능을 선보인 것이다. 이 플러그인은 Bing의 검색 API를 사용하여 콘텐츠를 검색해 ChatGPT에 답변의 출처를 밝히도록 한다. 이제 ChatGPT가 다양한 질문에 대답하기 위해서 웹 상에 있는 데이터까지 참조할 수 있게 됐다.



그러니까 Bing이 ChatGPT의 언어모델을 활용하고, ChatGPT가 Bing의 검색 엔진의 힘을 빌리는 것이다. 각각 독자적인 생태계를 구축하고 있는 마이크로소프트와 OpenAI가 상호보완적으로 점차 거대해짐에 따라 가장 초조해할 곳은 바로 구글이다.

2) 구글의 "코드 레드"

알파벳의 구글은 마이크로소프트의 (현재 보이는 그림으로는) 가장 직접적인 경쟁자다. 구글은 현재는 후발주자처럼 보이는 측면이 있으나, 사실 구글은 인공지능에 있어 소위 최고의 '인싸'였다. 애초에 OpenAI의 역사도, 일론 머스크와 샘 알트만이 구글에 대항하기 위해 설립됐을 정도다.

구글은 이미 6년 전 회사의 방향을 인공지능 중심으로 바꾼 회사이기도 하다. AI 기술을 본인들의 가장 중요한 목표 기술로 선정한 것이다. 정보를 다루는 검색엔진을 기반으로 한 회사다보니, 구글의 목표는 ‘전세계 정보를 체계화’하는 것이었다. 그리고 생성형 AI가 이에 가장 알맞은 제품이라고 판단했을 것이다. 따라서 구글은 전방위적으로 AI에 대한 투자를 계속해 왔다.

구글이 내놓은 AI 관련 논문은 많지만, 그 중에서도 가장 중요시되는 것은 바로 2017년에 나온 “Attention is All You Need”라는 이름의 논문이다. 트랜스포머의 꽃인 어텐션 기법에 대한 내용이다. 발표 당시에는 반응이 그리 크지는 않았다고 한다. 하지만 이후 생성형 AI 모델과 하위 특정 태스크에 굉장히 잘 들어맞는 활용 사례들이 나오면서 전설적인 논문으로 자리매김됐다. 이후 오늘날까지 어텐션 기법에 기반한 많은 생성형 AI들이 출현했고, BERT와 GPT도 바로 이 논문으로부터 시작됐다고 볼 수 있다.

구글이 2017년에 내놓은 NIPS 논문, “Attention Is All You Need”
: NIPS는 전세계 최고의 딥러닝 컨퍼런스. GPT와 BERT도 이 논문에서 비롯됨

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com	Noam Shazeer* Google Brain noam@google.com	Niki Parmar* Google Research nikip@google.com	Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com	Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu	Lukasz Kaiser* Google Brain lukaszkaizer@google.com	
Illia Polosukhin* ‡ illia.polosukhin@gmail.com			

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

자료: NIPS

2023년 2월 7일, 구글은 대형 언어모델 람다(LaMDA)를 기반으로 한 경량화된 AI 서비스인 바드(Bard)를 발표했다. 동시에 테스터들을 대상으로 공개했다. 바드는 웹에서 전세계의 정보를 가져와 시의성 있고 높은 품질의 응답을 제공한다고 한다. ChatGPT에게는 부족한 정보의 '신선함'과 '신뢰성'에 초점을 맞추면서 본인들의 edge를 강조하는 것이다. 구글의 크롬은 전세계 검색엔진 점유율 90% 초반대를 수년째 유지하고 있다. AI 언어모델에 기존 검색 쿼리 수행 알고리즘을 적용하는 방식으로 설계되지 않았을까 추측되는 지점이다.

바드(Bard)는 BERT(Bidirectional Encoder Representations from Transformers)를 기반으로 하는 트랜스포머 아키텍처를 기반으로 만들어졌다. BERT는 GPT와 함께 트랜스포머 언어모델의 가장 기본 축으로 속하는 기반 모델로 꼽힌다. 이름에서 알 수 있듯이 대표적인 특징은 '양방향'이라는 것이다. BERT는 어떤 문맥을 기점으로 왼쪽 문맥, 오른쪽 문맥을 모두 고려해 답을 생성하는 트랜스포머 모델이다. 반면 GPT는 단방향 모델이기 때문에 문장 생성에 더 강점을 보인다. 따라서 BERT를 기반으로 한 Bard는 질의응답이나 특정 언어 번역과 같은 분야에서 GPT보다 좀 더 우수한 성능을 보일 가능성도 배제할 수 없다.

이러한 바드가 대형언어모델의 '경량화'된 버전이라는 것도 눈에 띈다. 하드웨어든 소프트웨어 알고리즘이든, 일정 수준에 다다르게 되면 성능 개선 폭에 한계가 오는 saturation(포화)이라는 게 발생한다. 반도체로 치면 전력값(인가전압)을 아무리 높게 줘도 클럭값이 크게 개선되지 않는 것이다. 시험점수를 70점에서 90점으로 올리는 것보다 90점에서 95점으로 올리는 데 더 많은 노력이 필요한 것과 비슷하다.

인공지능 알고리즘에도 성능 saturation이 있다. 구글이 경량화된 모델을 출시한 것은 모델의 컴퓨팅 성능은 비교적 떨어지더라도, 이용자 수를 늘려 대중화를 이끌어 내겠다는 심산으로 보인다. 이용자가 많을수록 더 많은 피드백을 받게 되고, 이는 또 알고리즘의 개선으로 이어지기 때문이다.

3) 마이크로소프트와 구글의 전쟁

구글은 바드를 공개하면서 일반 이용자들에게는 몇 주 이후에 제공할 예정이라고 밝혔다. 그리고 시간이 좀 걸리는 듯하더니, 3월 21일 드디어 ChatGPT의 대항마인 바드가 출시됐다. 바드 출시를 예고한 지 한 달 반만이다.

우선 미국과 영국에서 일부 이용자들을 대상으로 바드가 오픈되고, 순차적으로 더 많은 국가와 언어로 이용될 수 있게 확장될 것이라고 한다. 확실히 마이크로소프트보다는 조심스러운 행보를 펼치고 있는데, 이는 기술에 대한 상대적 자신감이 조금은 결여된다고 보이는 대목이기도 하다. 그게 아니라면, 아직 준비가 되어 있지 않은 상태에서 서둘러 출시 시기를 앞당기기 위한 전략으로 해석되기도 한다.

어쨌든 구글도 Bard를 출시하면서 업무의 생산성을 높이는데 초점을 맞추었다. 따라서, 마이크로소프트와 구글의 AI 챗봇 경쟁이 이제 제대로 불이 붙을 것이다. 둘의 경쟁은 정말 점입가경이다. 마이크로소프트와 OpenAI가 ChatGPT를 발표하며 선전포고를 하니 구글은 "워크스페이스(Workspace)"를 공개했다. 그 이후 MS 진영은 GPT-4와 코파일럿을 내놨다. 그야말로 장군망군식으로 으르렁대는 형국이다.

구글의 워크스페이스라는 서비스는 구글 지메일, 구글시트, 구글챗, 구글미트 등에서 생성형 AI를 사용할 수 있는 기능이다. 이메일의 초안을 작성해주고, 순차적으로 쌓여있던 이메일 소통을 요약도 해주고, 프레젠테이션을 위한 슬라이드쇼도 작성해준다. 마이크로소프트의 코파일럿이 Office 365와 결합한 것과 대동소이하다.

구글 클라우드 기반 생성형 AI 활용 “워크스페이스”
: 지메일, 지드라이브, 구글독스 등에 활용

자료: Google

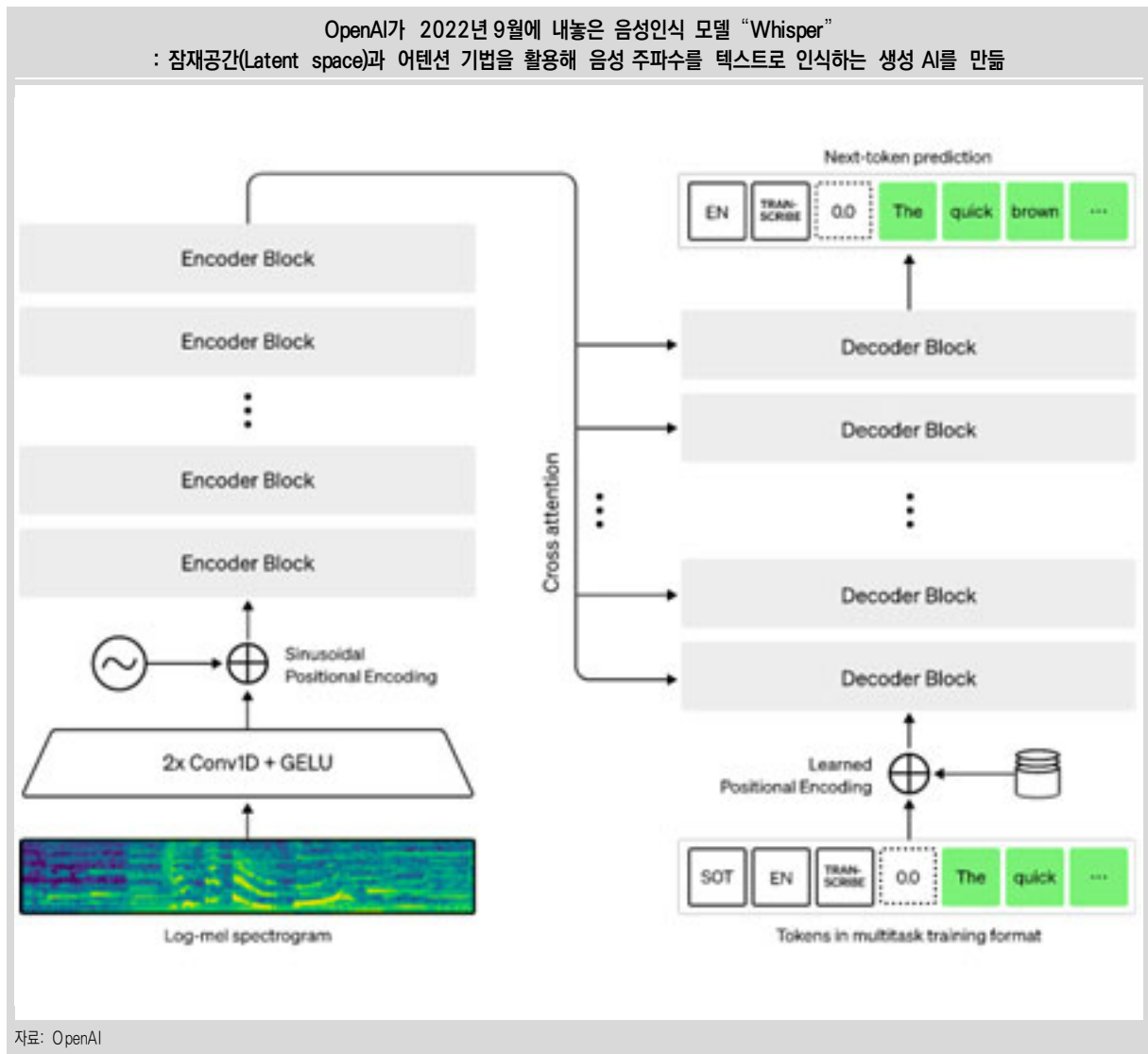
구글 “워크스페이스”에서의 AI를 활용한 작업물 예시
: 간단한 상황설명(context)과 명령어(prompt)만 있으면, 채용공고 문서 정도는 금방 작성

자료: Google

마이크로소프트와 구글의 싸움은 회사 바깥에서도 격화될 것은 분명하다. 두 회사는 AI 분야의 주도권을 차지하는 것을 생존의 문제로 생각하고 있으며, AI 에코시스템을 조성하기 위해 분주히 움직이고 있다. 이 에코시스템이란 결국, 본인들의 모델을 기반으로 하고 그 위에서 구축될 많은 어플리케이션을 유치하는 것이다. 다시 말해 본인들이 메인 네트워크가 되고, Layer 2 형식으로 들어올 스타트업들과 협력하는 방향이다.

구글은 2023년 3월에, 개발자들의 생성형 언어 API를 사용하게끔 온보딩을 지원할 예정이라고 이미 밝혔다. 또 지난 2월 초에는 C3.AI, 코히어(Cohere), 앤트로픽(Anthropic)과 맺은 구글 클라우드 파트너십을 통해 스타트업 지원을 확장하기도 했다. 좋은 언어모델도 결국 활용도가 높은 서비스로 확장되어야 의미가 있는 것이다.

한편 마이크로소프트는 구글보다 훨씬 앞서 가 있어 유리한 입지를 보유하고 있다. GPT가 이미 사람들의 뇌리에 강하게 박힌 상태이며, 인기와 성능이 시장에서 입증된 때문이다. OpenAI의 수장인 샘 알트만은 원래 Y Combinator라고 불리는 스타트업 액셀러레이터의 대표 출신이다. 그래서 누구보다 스타트업 생태계를 잘 알고 있고 또 인적 네트워크도 훌륭하다. 따라서 OpenAI가 구글보다 생태계를 빠르게 확장시킬 수 있을 것으로 보인다.



OpenAI는 당장의 투자 계획에 대해 많은 것을 언급하지 않았다. 하지만, 가까운 미래에 자체 출시 서비스와 더 밀접한 섹터에 맞춰 투자를 조정할 것으로 보인다. 예를 들어, OpenAI는 2022년 9월에 음성 인식 모델 Whisper(트랜스포머 기반)를 출시했었다. 당연히 서비스를 새로운 사용군으로 확장하기 위해 노력할 것이다. 따라서, Whisper 위에 구축할 수 있는 음성 기반 AI 스타트업에 투자할 가능성이 매우 높다. 자동 더빙 같은 음성 합성 등의 분야의 생성형 AI 스타트업들이 바로 그 주인공들이 될 것이다.

비슷하게, OpenAI는 GPT-4라는 ChatGPT의 다음 버전을 이미 제한적인 형태지만 대중에게 출시한 상태다. OpenAI가 보유한 “스타트업 기금”으로서, 향후의 GPT-5, 6 등을 기반으로 어플리케이션을 만들어 줄 스타트업들을 대상으로 투자도 집행할 것으로 생각한다. 이러한 방식의 투자를 통해 스타트업들에게 GPT 신제품 시리즈에 대한 조기 액세스를 제공할 것이다. 이에 따라 OpenAI가 모델 정식 출시 전에 모델을 보다 세밀하게 조정할 수 있게 된다. 일종의 베타 테스터로 활용하는 의미일 수도 있다. 이렇게 됨으로써 기술과 서비스는 더욱 완벽해지고 사용자 친화적으로 변화하며 에코시스템은 공고해진다.

4) 애플은 무얼 하고 있나?

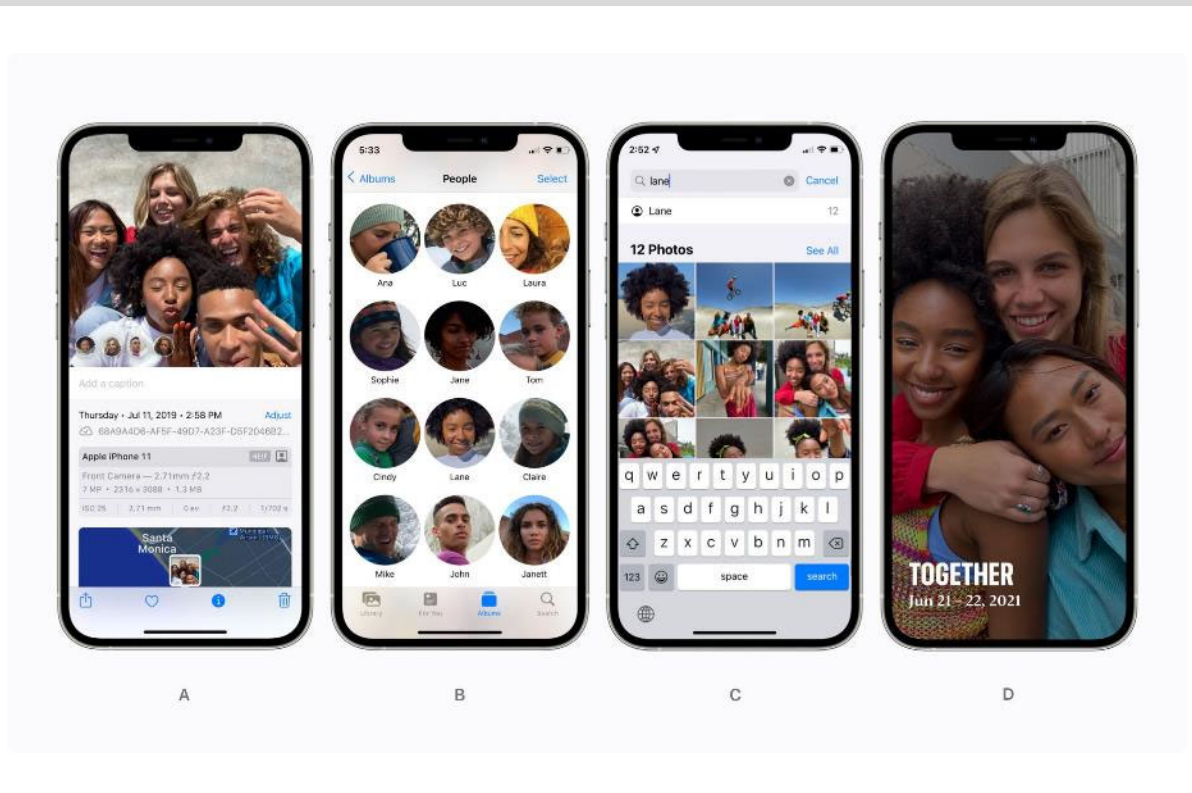
ChatGPT를 필두로 생성형 AI의 대유행이 시작되면서, 기존에 위세를 떨치던 음성 AI 비서들의 존재감은 추락했다. 애플 시리, 아마존 알렉사, 구글 어시스턴트, 삼성전자 빅스비 같은 존재들이 바로 그렇다. 애플 시리는 2011년에 출시됐지만 대중화가 됐다고 보기 힘들다. OpenAI의 ChatGPT의 돌풍과는 확연히 다른 요증이다.

그렇다면, 애플은 AI 대전에 왜 참여하고 있지 않을까? 애플은 항상 완벽하지 않다면 공개하지 않는 기업 문화와 전략을 가지고 있다. “언제나 최초는 아니지만 최고의 제품”을 내놓는 게 그들의 모토다. MP3도 그랬고, 스마트폰도 그랬고, 무선 이어폰과 스마트워치도 그랬고, 또 VR도 그렇다. 최근 애플은 미성년자들에게 부적절한 콘텐츠를 제공할 수 있다면서 ChatGPT를 활용한 이메일 어플리케이션 승인을 거부하기도 했다. 보안과 남용을 이유로, 애써 이 AI 흐름에 올라타지 않을 심산이다.

하지만 애플도 무시할 수 없는 전세계 NO.1 빅테크이자 빅데이터 기업임을 잊지 말자. 애플은 iCloud 생태계에 종속되어 있는 수많은 사용자들의 데이터를 확보하고 있다. 인공지능 생성형 AI의 출발점도 방대한 데이터라는 점에서, 이들이 수면 아래서 어떤 노력을 하고 있는지는 모를 일이다. 다만 우리에게 하나씩 최적화된 형태로 서비스를 출시할 것으로 보인다.

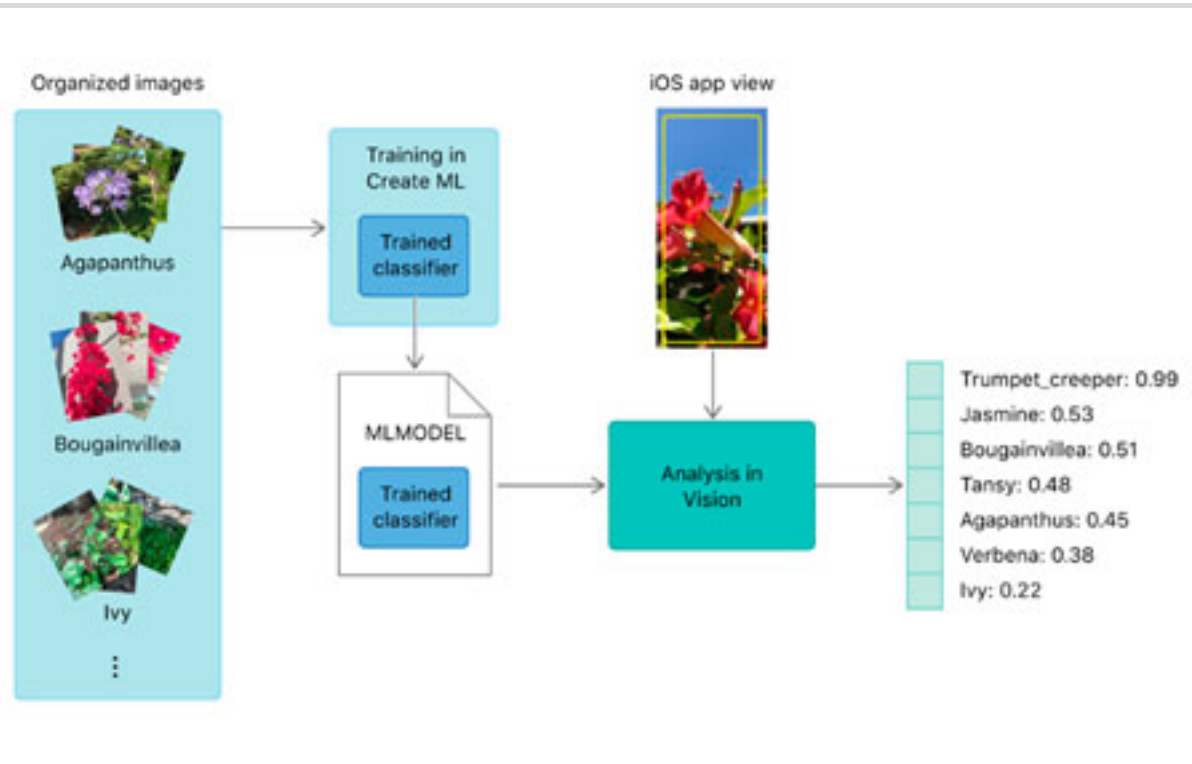
예를 들어, 애플 클라우드에 저장된 데이터 중 가장 많은 것을 차지하는 것은 아이폰으로 찍은 수없이 많은 사진과 영상일 것이다. 실제로 아이폰 기기를 쓰다 보면, 과거에 찍어놓은 사진을 가지고 아름다운 클립으로 알아서 제작해주기도 한다. 그런데 이 과정에서 “내가 알려준 적도 없는데” 이 사진이 언제 어디서 누구와 함께 찍힌 사진인지 알아서 분류한다. 그러니까 일종의 “캡션(자연어 명령어)”이 달린 형태로 사진과 영상을 분류해 사용자에게 제공하는 것이다. 온디바이스 형태로 이미지 분류 작업 정도는 이미 하고 있다는 것이다. 앞으로 어떻게 더 발전된 형태로 진화할지는 모르지만 애플은 본인들의 퀄리티 체크가 끝난 상태면 곧바로 세상을 놀라게 할 것이다. 항상 애플은 그래왔으니 말이다.

“온디바이스” 머신러닝으로 사진 안의 사람들의 얼굴 인식



자료: Apple Machine Learning Research

WWDC 2018에서 이미 개발자들에게는 선보였던 애플의 이미지 분류 알고리즘 : 식별하고 싶은 카테고리 설정은 해야 하는 지도학습 기반



자료: Apple Developer

물론, 애플이 보유한 AI 기술 자체가 뒤처져 있다는 분석도 있다. 애플 엔지니어 출신인 존 버키의 발언을 인용한 뉴욕타임즈의 기사를 참고하면, 애플 시리는 기본 기능을 업데이트하는 데에만 코드 수정에 몇 주씩 걸린다고 한다. 그만큼 코드를 애초에 설계할 때 구식이었다는 말이다.

ChatGPT의 경우 거대언어 데이터로 최신식의 생성형 AI의 알고리즘을 사용하고 있지만, 1세대들은 구형의 시스템에 기반하고 있기 때문이다. 사용자가 코드에 없는 뭔가를 물으면 음성인식 비서가 도움을 줄 수 없는 방식이다. GPT의 경우에는 모르는 답이라 할지라도 거짓말을 해가면서까지 대답을 생성해내는데 이와는 다른 것이다.

또한, 애플은 생각보다 꽤 AI 인력수급에 불리하다는 측면도 제기된다. 미국 AI 개발자들은 본인이 개발한 코드를 GitHub에 공유하는 것이 암묵적인 약속이라고 한다. 그런데 애플은 매우 보수적이고 폐쇄적인 데 정평이 나 있다. 애플은 본인들의 개발자들이 연구한 것을 공개하지 않길 바라고, 따라서 최상급 개발자들은 애플이 아닌 마이크로소프트로 간다는 이야기다. GitHub가 마이크로소프트니까 충분히 일리 있는 이야기로 사료된다.

그럼에도 애플은 뜻밖의 수혜도 볼 가능성이 있다. 구글과 마이크로소프트의 '검색 플랫폼 경쟁'으로 애플이 어부지리를 얻을 수 있다. 구글은 크롬을 아이폰 기본 어플리케이션으로 설치해두기 위해 애플에 연간 150억 달러를 지급한다. 최고의 하드웨어 플랫폼으로서의 입지를 가지고 있는 애플이기에 가능한 것이다.

그런데 여기서 중요한 점은, 구글과 애플의 이러한 계약이 올해 말에 끝날 예정이라는 점이다. 90% 이상의 검색엔진 점유율을 가진 구글이었기에 애플과의 협상에서 교섭력이 있었지만 이제는 이야기가 다르다. 마이크로소프트가 GPT-4로 무장한 New Bing을 선보이면서 검색엔진 경쟁에 불을 붙이고 있기 때문이다. 지난 3월 9일 마이크로소프트는 New Bing을 출시한지 한 달만에 매일 사용하는 활성화 유저 숫자인 DAU(Daily Active Users)가 1억명을 돌파했다고 발표했다. Bing이 검색엔진 대항마로 떠오르는 상황에서, 애플은 구글 대비 교섭력이 올라가는 상황인 것이다. 즉, 크롬의 어플리케이션 입점료를 더 올릴 수 있는 명분이 생긴 것이다.

다만, 마이크로소프트가 애플의 본격 경쟁 상대로 나온다면 얘기는 또 달라질 수 있다. 만약 ChatGPT와 Office "Copilot"을 완벽히 지원하는 호환성을 무장한 새로운 하드웨어를 출시하게 되면 판세는 또 한 번 바뀔 수 있다. 애플 하드웨어의 공고함에 번번히 고배를 마셨던 마이크로소프트지만 이번에는 다를 수 있다고 생각한다.

인공지능 기술을 쓰기 위해 사람들이 마이크로소프트의 스마트폰을 살지, 그래도 사람들이 애플 생태계를 고수할지도 앞으로의 관전 포인트가 될 것이다. 애플도 이를 충분히 인지하고 있을 것이기에, 자체적으로 개발한 생성형 AI를 출시하고 아이폰에 기본 프로그램으로 탑재하기 위해 고군분투를 벌일 것은 당연지사다.

5) 메타 “우리도 AI 기업이다”

2022년 한 해를 관통해 가장 고통받았던 빅테크 중 하나를 꼽자면 바로 메타(META)일 것이다. 하지만 올해는 상황이 완전히 다르다. 메타는 3월 22일 종가 기준, YTD 수익률이 60%를 상회한다. 애플(26.2%), 알파벳(16%), 아마존(15%), 마이크로소프트(13.7%)와 비교했을 때도 월등한 수준이다.

메타의 환골탈태? 작년과는 다른 메타를 둘러싼 투자심리

구분	2022년 메타의 우려 요인	2023년 메타의 기대 요인
본업인 광고 비즈니스	매크로 환경에 따라 광고 비즈니스는 직격탄 예를 개인정보정책(ATT)에 너무 많이 휘둘릴 것	매크로가 메타 주가를 좌우하는 정도는 점차 약세 더군다나 리오피닝 광고수요가 회복세! 개인화된 광고 위한 AI 역량으로써 계속 타개 중
숏비디오 사업의 수익성	숏비디오는 광고단가가 낮아 수익화가 어려움 이는 별류에이션 하락 요소	4분기 '광고 전환율' 전년동기대비 20% 상승 올해 AI 관련 서비스 출시 예정(사업 다변화)
사람들로부터의 인기	페이스북은 '하재'들이나 쓴다는 인식 인스타그램은 2030세대가 한계 신생 업체인 '틱톡'과의 경쟁에서 밀림	미중간 갈등으로 틱톡 이용률 저하에 수혜 가능성 릴즈 사용량은 증가하고 틱톡의 사용자 숫자는 감소 릴즈의 광고수익화 수준은 반년 전 대비 2배 상승
미래먹거리와 투입비용	메타버스로 매년 십수조원의 현금을 까먹고 있음 모호한 미래 방향성에 핵심인력 유출까지	일반 관공비에서 불필요한 비용 줄여 현금확보 단기적으로는 AI에 중점. 메타버스 올인이 아닐!
법적/규제 리스크	개인정보보호 정책으로 인한 수익화 제한 다른 특정국가에서 메타 생태계 접근 제한 가능성	규제 리스크는 예상치 못할 잠재 리스크 다만, 메타의 AI 잠재력은 미국으로서도 큰 무기

자료: 미래셋증권 디지털리서치팀

여러 요인들 중에서 상황을 반전시킨 내러티브를 한 문장으로 정리하자면, “메타버스는 좀 쳐두고, 메타도 인공지능 기업이다”라는 것이다. 메타의 투자심리를 바꾼 여러가지 요인들을 인공지능 역량 관점에서 해석해보면 아래와 같다.

메타는 광고 비즈니스가 핵심이다. SNS 기업들의 숙명과도 같다. 메타는 애플의 개인정보보호정책이 엄청 강화되면서 메타는 매출 직격탄을 맞게 됐다. 그런데도 딱히 메타가 이를 타개할 역량이 없는 것 같다는 비판을 받았고 실제로 그랬다. 그러나 메타가 가진 AI 역량을 활용해 개인화된 광고를 계속해서 넣어 줌으로써, 광고주들로부터 받는 점차 광고단가를 방어하고 있다.

뿐만 아니라, 메타는 지난 4분기 실적 발표 이후에, 광고 전환율이라는 것이 전년동기대비 20%나 상승했다고 발표하기도 했다. 광고 전환율은 향후 광고단가를 결정짓는 핵심 요소 중 하나다. 이는 메타가 본인들의 AI 역량을 통해 맞춤 광고를 고객에게 효과적으로 넣어줬기 때문이라고밖에 생각되지 않는다. 게다가 올해 AI 관련 서비스도 여럿 출시하겠다고 발표한 만큼 앞으로 여러가지 서비스 수익이 발생할 여지도 존재한다.

그리고 기존에는 메타의 미래 먹거리지만, 메타버스에 너무 돈을 많이 쏟아붓고 있다는 우려가 굉장히 많았다. 매년 십수조원을 쓰면서 메타버스에 올인하는데 미래가 불투명해보이니까 당연한 불평이다. 메타는 이런 불만을 의식해 2022년 8월에 메타버스 서비스에 대한 진전사항을 업데이트했는데 오히려 역풍을 맞았다. 품질이 떨어지다 못해 조악한 수준이다보니 사람들로부터 엄청난 조롱을 받았다. 이후 주가는 더 떨어지고 말았다. 하지만, 현 시점에서 투자자로서 다행스러운 점은 메타가 불필요한 비용을 줄여서 현금 확보에 매진하고 있다는 것이다.

더욱이 중요한 것은 메타가 메타버스에 관심이 많긴 하지만 이것은 “장기적”인 마스터플랜이라고 말했다는 점이다. 당장은 AI에 힘을 주겠다는 것이다. 즉, “메타버스 올인이 아니라 우리도 AI 투자한다”는 말로 치환할 수 있다. 실제로 메타는 데이터센터 관련지출 금액을 줄이기로 했지만, AI 관련 데이터센터 투자는 전혀 줄이지 않았음을 천명했다.

또한, 빅테크들이 가지는 AI 역량과 글로벌 사용자 데이터 확보는 미국이라는 국가 입장에서 바라볼 때 큰 무기라고 판단한다. 희망적으로 생각하면, 미국 정부 입장에서는 이러한 데이터를 활용하려고 할 것으로 보인다. 정부와 거래하게 될 때 빅테크들에 대한 정치권의 견제도 있겠지만 역설적으로 교섭력이 올라갈 수도 있겠다.

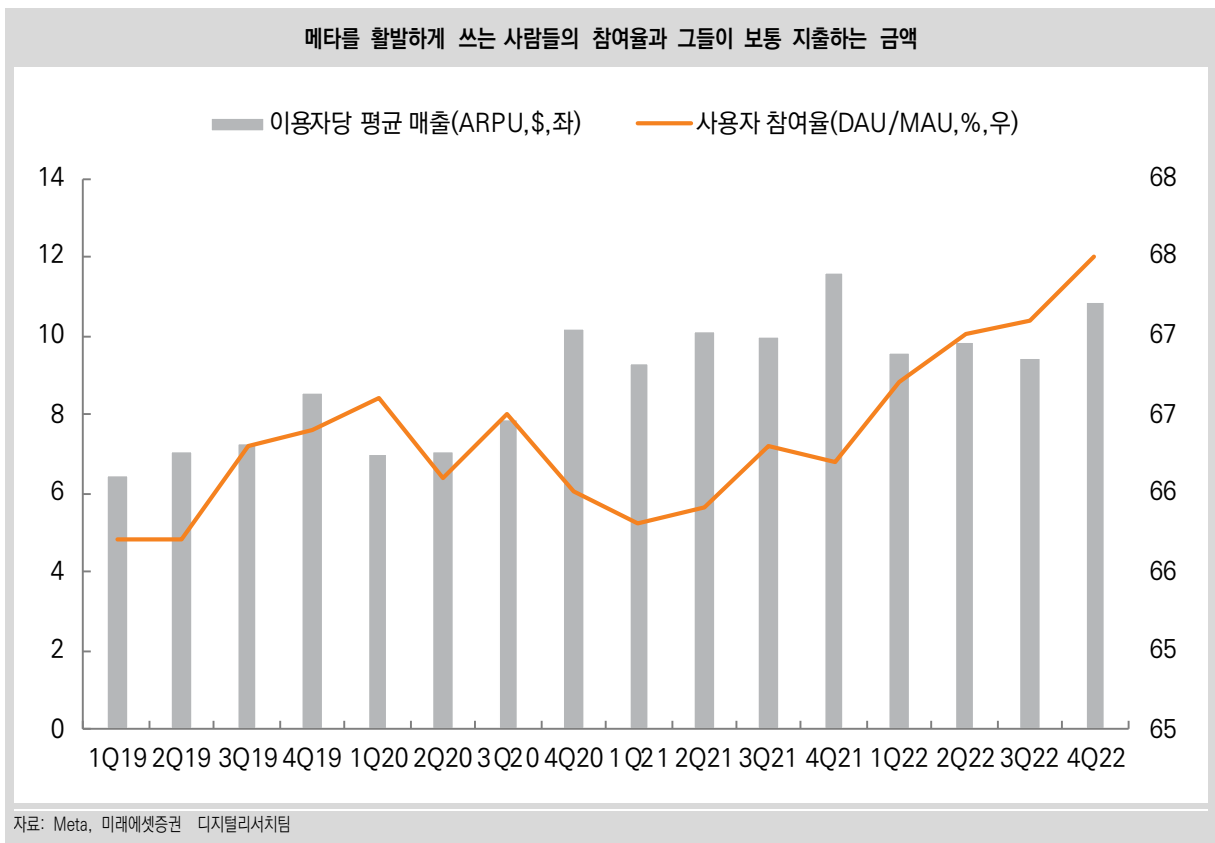
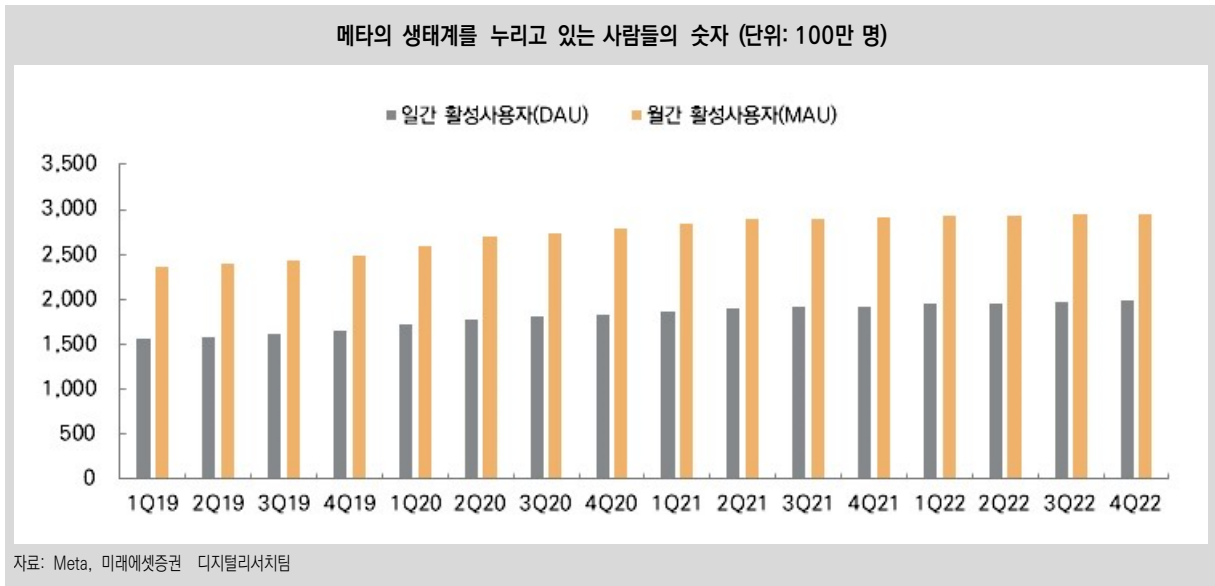
더군다나, 젊은 사람들이 페이스북과 인스타그램을 기피하고 틱톡을 많이 쓰고 있다는 것이 도움이 될 수 있다. 틱톡을 운영하는 바이트댄스는 중국기업이다. 틱톡을 통해 미국인들의 영상데이터를 확보하고 있다는 우려도 이 때문에 주기적으로 나왔다. 텍사스주에서는 몇 달 전부터 공무원들을 대상으로 틱톡 사용에 제한을 걸기도 했다.

거기다가 최근 미중 간 갈등이 정치/사회/경제/안보/군사/경제/기술/자원/외교 전방위적, 복합적으로 커지고 있다. 앞으로도 미국의 여러 지방정부가 틱톡 사용 제재를 도입할 여지가 있어, 미국 내 틱톡 이용률이 저하될 수 있다. 이런 경쟁에서 수혜를 입는 것은 메타가 된다. 실제로 틱톡의 사용자 숫자는 감소했고, 반면 메타의 숏비디오인 릴즈의 사용량은 증가하는 추세이기도 하다. SNS 사용량이 늘어난다는 것은 AI 알고리즘 구축을 위한 데이터를 계속 더 확보한다는 말과 같다.

메타의 실적 상승은 릴즈가 기존 서비스 수준으로 수익성이 올라와야 한다. 아직은 1년 정도 더 지나야 한다고 메타가 이번 실적에서 시인했기 때문에 당장은 기대하기 어려울 것이다. 즉 메타가 당장 추가적인 주가 상승을 정당화하려면 PER을 높여주는 기대감을 더 심어줘야 한다. 개인화된 맞춤형 광고로 광고단가를 올릴 수 있는 역량, 그리고 서비스 다각화가 가능해야 한다는 말이다. 그리고 이를 가능케 할 것은 바로 AI가 된다.

메타는 페이스북, 인스타그램, 왓츠앱을 가진 엄청난 사용자를 보유하고 있다. 사용자가 많다는 것은 그만큼 트래픽이 많다는 얘기고, 트래픽이 많다는 것은 그만큼 데이터도 많이 생성된다는 얘기다. 많은 사람들이 텍스트를 작성하고, 영상을 올리고 대화도 할 것이기 때문이다. 그런데 메타의 서비스들을 이용하는 숫자가 MAU 기준 전세계적으로 30억 명에 육박한다. 전세계 인구가 79억 명 정도이니 전세계 인구 중 약 40%가 메타의 이용자들이라는 것이다. 물론 이 숫자가 늘어나는 속도가 느려지긴 했지만, 숫자가 줄지 않고 꾸준히 늘어나고 있는 게 무섭기도 하다.

게다가 메타를 쓰는 사람들의 참여율이 역대 최고치를 달성했다. 사용자 참여율이 높다는 건데, 이 말은 이 회원들이 유령회원이 아니라는 말이다. 이렇게 메타가 보유하고 있는 대규모 사용자 데이터는 AI 산업에서 강력한 입지를 보여줄 것으로 생각한다. 스냅과 같은 기업들과는 차별화되는 지점이다.



6) 빅테크가 아닌 새로운 잠재 경쟁자, “BasedAI”

OpenAI의 공동설립자였던 일론 머스크는 인공지능 챗봇 ChatGPT의 검열방식과 편향성에 대해 지난 1년간 반발했던 인사다. 그는 최근 OpenAI의 ChatGPT에 대항하는 라이벌을 개발 중인 것으로 알려졌다. 머스크는 2015년 Y 컴비네이터의 샘 알트만(Sam Altman) 대표와 함께, 구글 측의 AI 역량 독점을 막기 위해 OpenAI를 공동 설립했다. 다만, 샘 알트만과의 마찰로 인해 2018년 사임했다. 이후 2019년 OpenAI는 마이크로소프트로부터 10억 달러의 투자를 받아 배타성이 점차 강조됐기 때문에, “Open”이라는 가치는 상당히 희석됐다. 많은 AI 모델은 오픈소스를 기반으로 한다. 그래서 개발자 커뮤니티가 하나의 프로젝트에서 동시에 작업할 수 있어 개발 속도가 빨라지지만 ChatGPT는 현재 그렇지 않다.

또한, 머스크는 ChatGPT가 잠재적으로 불쾌감을 줄 수 있는 텍스트를 생성하지 못하도록, 안전장치를 설치한 이후 OpenAI가 언어모델을 "woke(정치적 진보성향의 행동적 이념을 비판하는 단어)"하게 훈련시켰다고 비난하고 있다.

Musk "The danger of training AI to be woke – in other words, LIE – is deadly."

더군다나 마이크로소프트가 ChatGPT 출시 직전인 작년 10월에 자신들의 AI 윤리팀 규모를 7명으로 축소했었는데, 최근에는 이를 완전히 해체한 것으로 드러났다는 점도 우려 사안이다. 철학자 및 엔지니어로 구성된 MS AI 윤리팀은 2020년 출범해 MS의 “책임있는 AI 원칙과 제품 설계”를 위한 조직이었지만, ChatGPT 및 New Bing이 대중 밖으로 나오면서 윤리의식은 뒷전으로 밀렸다고 해석할 수도 있는 부분이다. 이 같은 과감한 결정은 구글과 같은 다른 업체들보다 더 빠르게 제품을 내놓아 소비자들을 확보하고 락인 시키는 것이 시급했다고 봤기 때문일 것이다.

OpenAI의 CEO인 샘 알트만은 트위터를 통해 자신의 회사가 "편향성에 대한 단점"을 인식하고 있다고 시인했다. 기본 시스템을 보다 중립적으로 만들고 사용자에게 더 많은 맞춤형 경험을 제공하기 위해 노력하고 있다고 밝혔다. 샘 알트만이 이런 얘기를 했다는 것 자체가, 이런 문제에 대해서 문제의식을 갖고 있는 사람들이 매우 많다는 것으로 유추해볼 수 있다.

이에 따라 머스크는 ChatGPT에 대항할 해답을 내놓은 챗봇으로 "based AI"를 기획 중이다. Based는 편견이 없거나 진실에 기반한다는 뜻의 속어다. 머스크는 대안적 AI를 주도할 연구원을 모집 중인데 그 인사들 면면을 보면 상당히 무게감이 있다는 게 중요하다. 디 인포메이션에 따르면, 머스크는 이전에 알파벳의 딥마인드 AI에서 일했던 수석연구원 이고르 바부슈킨에게 연구소를 설립하는 방안을 제안했다고 한다. 딥마인드를 떠난지가 얼마 안됐기 때문에, 바부슈킨 자신도 아직 공식적으로 프로젝트에 합류하지 않았다. 바부슈킨이 밝힌 바에 따르면, 그들의 목표는 안전장치를 푸는 것이 아니라, 언어모델의 추론 능력과 사실성을 향상시키는 것이라는 점을 분명히 했다.

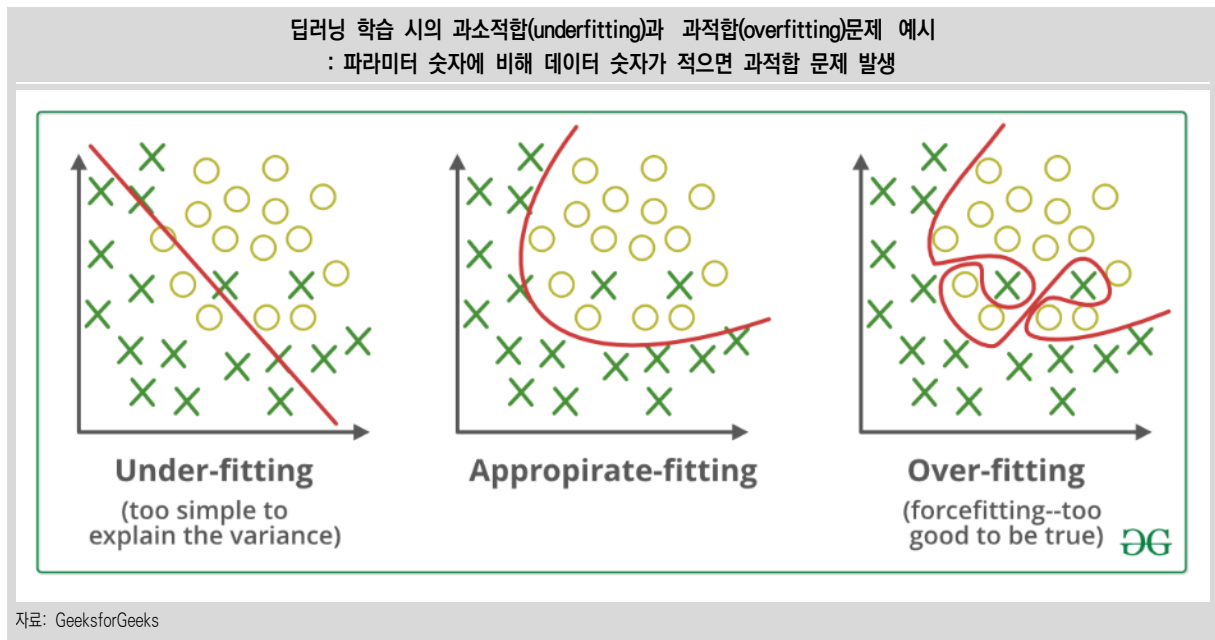
ChatGPT가 너무 좋고 편리하기 때문에 많은 유저 숫자를 불러 모을 수 있었지만, 이런 행태는 분명히 우려스러운 요인이다. (기능적) 편리함에 더해 (윤리적) 불편함 또한 공존하게 될 것이다. 그런 상황에서 이후 진정한 의미의 “OpenAI”가 나오면 상황이 빠르게 역전될 수도 있다고 본다. 그리고 그것을 위해 전세계에서 가장 큰 팬덤을 가지고 있고, 또 컴퓨터 사이언스 전문가인 일론 머스크가 나섰다는 점에서 의미심장한 사건이 될 수 있다. ChatGPT의 가장 큰 잠재 경쟁자는 구글도, 애플도 아닌 트위터와 테슬라가 될 수 있다.

생성형 AI를 위한 연산허브는 데이터센터

딥러닝이 최근 크게 다시 부각되는 이유는 성능의 월등함이 증명되고 있기 때문이다. 그리고 그 성능은 대부분 딥러닝의 모델 크기에 종속된다. 딥러닝의 모델 크기는 파라미터 개수로 인식되곤 한다. 은닉층에서 노드들이 이어져서 그 가중치(Weight)를 나타내는 파라미터 숫자가 많아야 큰 모델이라는 말이다.

그런데 파라미터 숫자가 많으려면 애초에 인풋 데이터의 크기도 당연히 커야 한다. 데이터는 한정적인데 파라미터 숫자만 늘리게 되면 딥러닝에 있어서 ‘과적합(overfitting)’이라는 이슈가 발생해 나쁜 성능의 딥러닝 모델만 만들어진다.

GPT-4의 경우 수조 개의 파라미터 숫자를 가지고 있다. 파라미터가 각각 4바이트 정도의 파일 크기를 갖고 있다고 가정하면, 수십 테라바이트에 육박한다. 게다가 대규모 연산까지 수행해야 한다. 개개인이냐 수많은 중소기업들이 하기에는 너무 버거운 작업이다.



그래서 데이터센터 운영업자들이 제공하는 호스팅 기능을 빌려 인공지능 연산을 하는 경우가 점차 늘어날 것이라고 판단한다. 예를 들어, 스테이블 디퓨전 같은 고품질 그림 생성형 AI를 사용하려고 할 때, 집에서 이를 실행시키려면 하이엔드급 장비가 필요하다. 엔비디아의 RTX 4090급 GPU를 의미하는데, 집집마다 200만원을 호가하는 GPU가 있을 것이라고 생각하는 건 무리다.

1) 구글의 “코랩”과 TPU

따라서, 하이퍼스케일러들이 제공하는 데이터센터 연산 호스팅 기능을 사용하게 된다. 그들의 서버에 있는 GPU를 돈을 내고 빌려 쓰는 개념이다. 보통 구글 코랩(Colab)이 많이 쓰이는데, 코랩은 구독형 모델로 점차 영역을 확장해 나가고 있다. 그림 생성형 AI를 취미 삼아서는 무료로도 쓸 수 있지만 처리량이 많아지면 자연스럽게 유료 모델 구독 버튼을 딸각거리게 된다. 구글 코랩의 경우 무료 모델에서는 호스팅 용 GPU로 엔비디아의 Tesla T4(2018년 9월 출시)를 제공하고, 유료 모델의 경우 Ampere A100(2020년 5월 출시)을 사용할 수 있도록 한다.

**엔비디아의 A100 GPU에는 대략 1만 달러라는 가격표가 붙음.*

구글 코랩: 구글 데이터센터를 빌려 딥러닝을 수행(inference)
: 코랩 프로(월 9.99달러), 프로+(월 49.99달러)를 구독하면 더 빠른 GPU 사용 가능

Choose the Colab plan that's right for you

Whether you're a student, a hobbyist, or a ML researcher, Colab has you covered

[Restrictions apply, learn more here](#)

Colab

Free

Current plan

✓ No subscription required.

Recommended

Colab Pro

\$9.99 / month

- ✓ **Faster GPUs**
Access to faster GPUs and TPUs means you spend less time waiting while your code is running.
- ✓ **More memory**
More RAM and more disk means more room for your data.
- ✓ **Longer runtimes**
Longer running notebooks and fewer idle timeouts mean you disconnect less often.

Colab Pro+

\$49.99 / month

- ✓ **Background execution**
Notebooks keep working even after you close your browser.
- ✓ **Faster GPUs**
Priority access to faster GPUs and TPUs means you spend less time waiting while your code is running.
- ✓ **Even more memory**
Significantly more memory than ever before.
- ✓ **Even longer runtimes**
Gives you the longest running notebooks in Colab so you are able to get your work done.

자료: Google Colab

이처럼, 앞으로 인공지능이 불려올 산업별 대격변과 생성형 AI의 유행으로 데이터센터의 입지는 막강해질 것이다. 현재는 10분 정도지만 앞으로 계속 가격이 올라갈 여지가 있다고 판단한다. AI 연산에 특화된 서버를 다량으로 확보하고 있다는 것은, AI 개발에 있어 주도권이 있는 것으로 해석할 수 있기 때문이다.

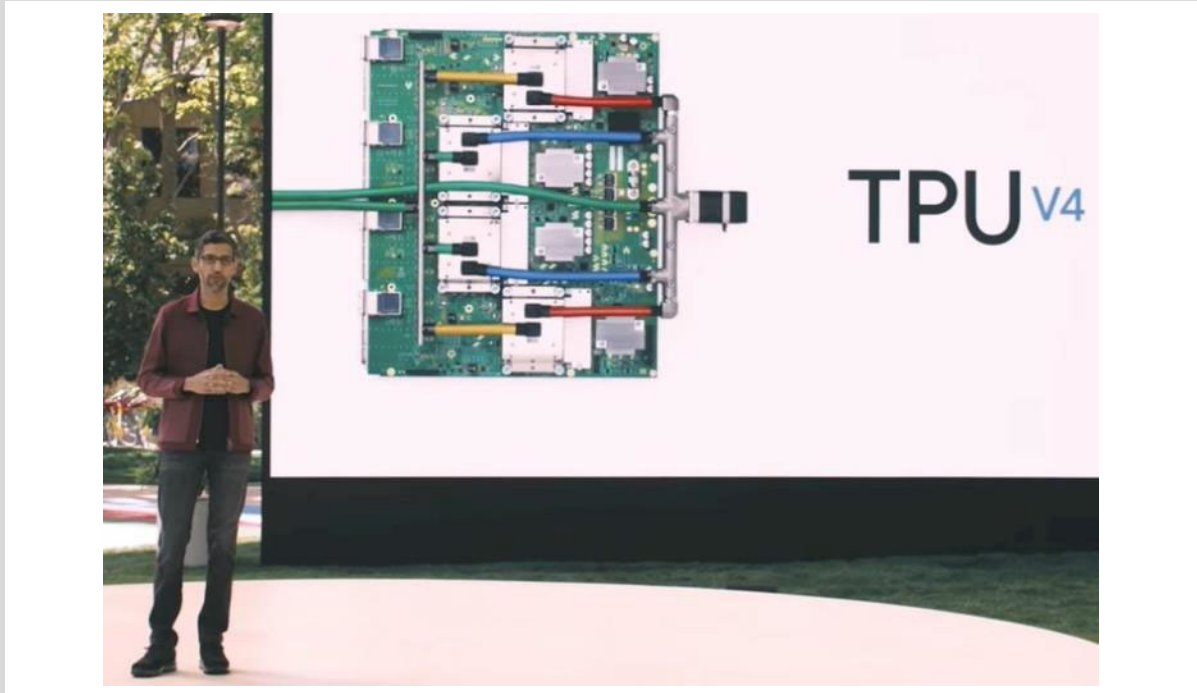
이를 잘 알고 있는 구글은 본인들의 개발자회의인 I/O 개발자 콘퍼런스에서 데이터센터에 들어가게 될 주요 시스템을 대대적으로 홍보했다. 그리고 2022년 I/O에서 구글 클라우드 TPU v4 Pods라는 다소 긴 이름의 시스템이 출현했는데, 이를 풀어서 이해하면 어렵지 않다.

여기서 TPU는 텐서프로세싱유닛의 줄임말이다. 구글은 이미 몇 년전부터 인공지능에 대해 준비를 해왔던 회사이기에, 인공지능에 적합한 컴퓨팅 유닛코어 개발 역량도 보유한 곳이다. 즉 딥러닝 모델에 맞춰 본인들이 직접 주문설계한 ASICs가 TPU다.

**물론 구글 클라우드에서 제공하는 연산유닛에는 엔비디아 GPU가 전반적으로 많이 쓰임. 다만, 일부 특정 AI 연산에서는 TPU를 이길 수 없음. 이것은 ASICs 형태로 만들어지는 NPU의 특성 때문인데, 자세한 내용은 이후 반도체 부문에서 설명.*

**구글의 자체제작 칩은 크게 보면 세가지가 있음. AI 칩인 텐서프로세싱유닛 TPU, 보안용 프로세서 타이탄, 유튜브 동영상 처리를 위한 VCU.*

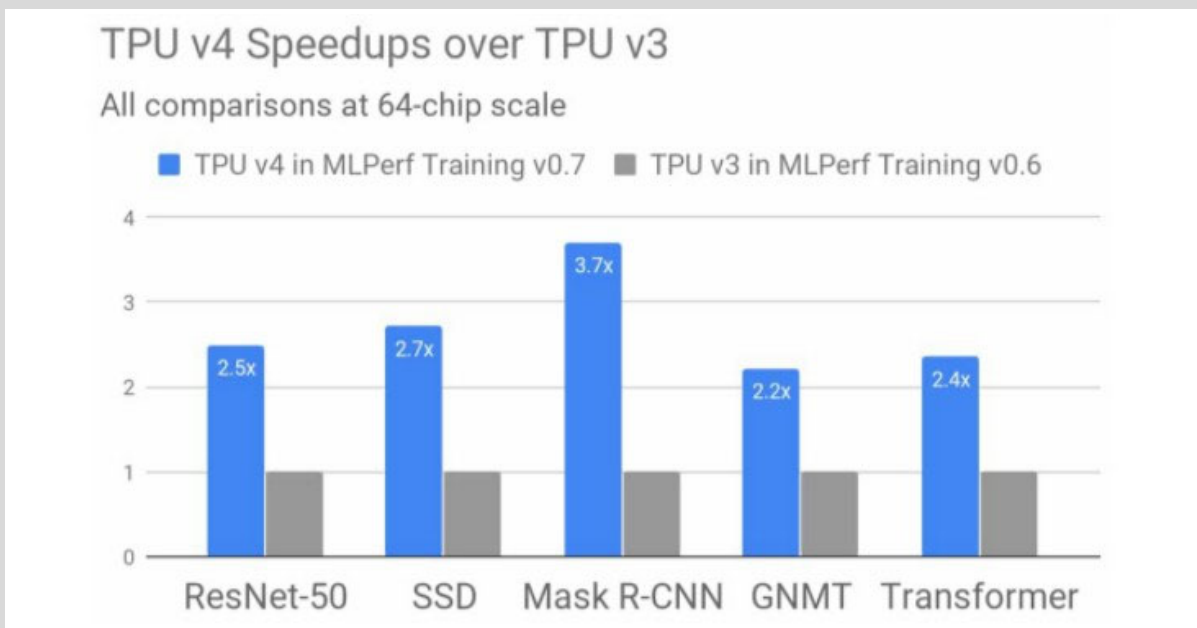
구글이 만든 딥러닝 연산용 ASIC, "TPU v4"의 보드 모습
: 보드 안에 4개의 칩이 존재하고 칩 하나당 연산코어가 2개일 것으로 추정



자료: Google

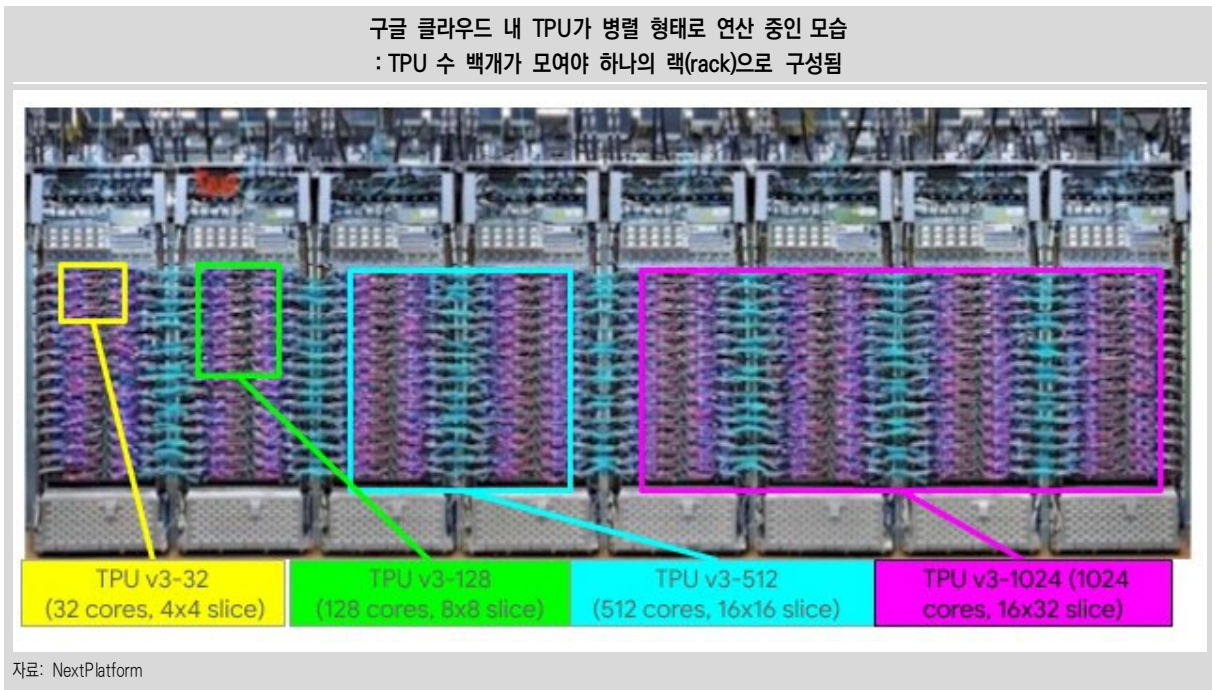
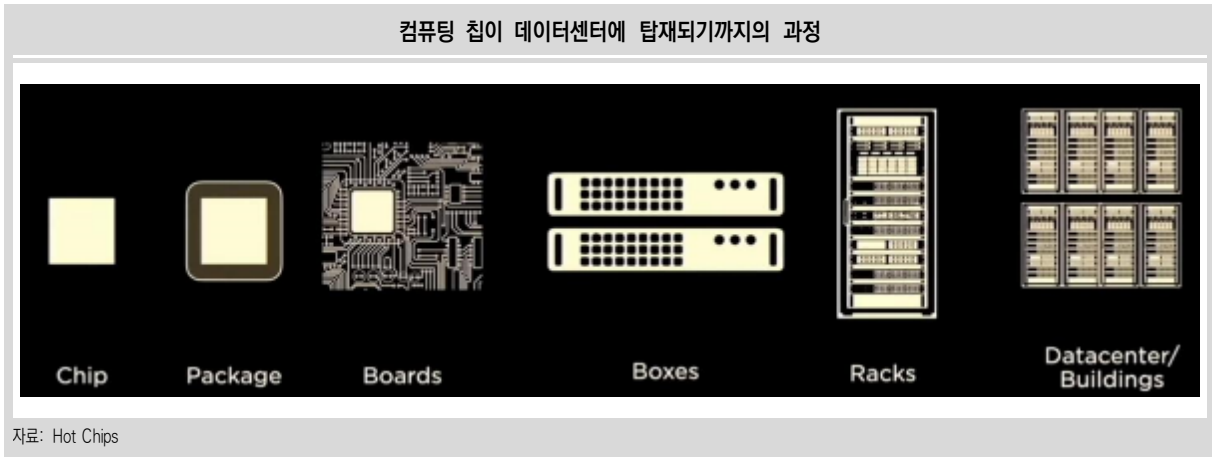
구글은 이미 2021년 I/O 행사 때 TPU v4 칩을 선보인 적이 있다. 각 칩은 최대 275TFLOP(BF16 기준)의 성능을 가졌는데, 이렇게 수치로 말하는 것보다는 전작과 비교하는 것이 좋을 것이다. TPU v4 칩은 전작인 v3 대비 2배가 넘는 성능 향상을 보였다. 이는 구글도 트랜스포머 모델의 유용성을 의식한 것으로 풀이된다. TPU v4가 대형언어모델(LLM)에서도 2배 이상의 성능 향상을 보였다고 말했기 때문이다.

각 딥러닝 모델 연산에서 TPU v3대비 TPU v4가 가지는 성능 향상폭
: 트랜스포머 학습도 2.4배 속도 개선!



자료: Google

그리고, 작년에는 차세대 코어인 TPU v4 칩들로 구성된 병렬연산 시스템인 “Pod”를 선보였다. 이 Pod에는 TPU v4 칩이 4,096개가 들어간다. 각 Pod는 최대 1.1 엑사플롭의 컴퓨팅 파워를 보인다. 병렬 시스템 하나의 연산능력이 드디어 페타를 넘어 엑사스케일까지 올라간 것이다. 구글은 앞으로 Pod 8개로 클러스터를 구성해 최대 9 엑사플롭의 데이터센터를 운영할 것이라고 밝혔다. 구글이 주장한 것처럼 누적 컴퓨팅 파워 측면에서는 세계에서 가장 큰 공개적인 인공지능 연산 허브가 된 것이다.



이 클러스터가 하는 일은 명확하다. NLP같은 생성형 AI를 위한 딥러닝 연산에 쓰일 것이다. 이미 작년 말부터, 고객들에게 제공될 것이라고 순다르 피차이 CEO가 밝히기도 했다. 구글 코랩에서도 TPU 호스팅을 제공하고 있는데 앞으로 해당 부분이 점차 늘어나지 않을까 기대한다.

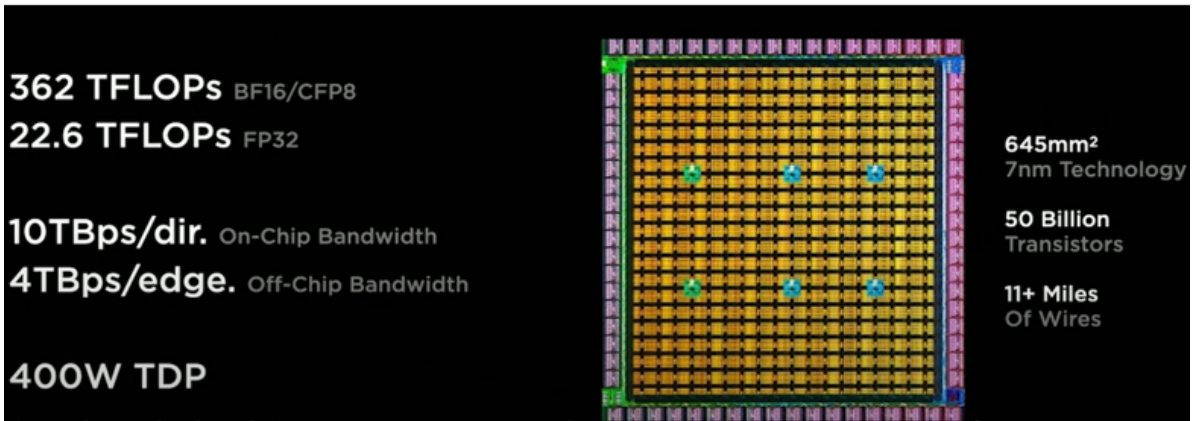
그리고 가상머신 체제로 운영되기에 사용자가 4개의 TPU 칩을 쓸지 수천개의 TPU 칩을 임대해 쓸지 정할 수 있다. 앞으로 고객들의 요구사항은 여러가지의 구독모델로 충족될 것으로 사료된다. 이것은 곧 안정적인 매출 상승을 의미한다. 이만큼이나 인공지능 산업화 시대에서 데이터센터가 갖는 영향력은 클 수 있다.

2) 테슬라의 “도조 엑사포드”

구글이 2021년 5월에 TPU v4 Pods를 내놓고 몇 달이 지난 뒤, 또 굉장한 데이터센터 시스템이 공개됐다. 이번에는 빅테크가 아니라 자동차 기업으로부터 나온 점이 신기한 일이다. 주인공은 역시 테슬라였다. 테슬라는 2021년 8월 AI DAY를 개최하면서, 본인들의 자율주행 기법에 대해서만 설파한 것은 아니었다. 그 딥러닝 연산을 어떻게 구동할 것이며, 또한 어떻게 효율적으로 할 것인지에 대한 부분에도 많은 시간을 할애했다.

놀라웠던 것 중 한 가지는 자동차 회사인 테슬라가 본인들이 직접 설계한 ASICs 반도체인 “D1” 칩을 선보였다는 점이다. 본인들의 자율주행 모델 네트워크를 최적화하기 위해서 코어 설계를 단행한 것이다. D1 칩은 일전에 서술한 구글의 TPU v4 칩보다도 높은 TFLOPs(BF16 기준)를 보였다는 점에서 괄목할 만했다. 그리고 이 칩을 만든 목적은 분명하다. 자율주행 딥러닝 네트워크를 효과적으로, 또 빠르게 고도화하기 위해서다.

2021년 8월에 공개된 자율주행 트레이닝용 칩 “D1” - TPU v4보다도 높은 성능 : 테슬라 도조 데이터센터를 구성하는 가장 기본단위가 D1 칩



자료: Tesla AI DAY 2021

테슬라는 이미 엔비디아 GPU 1.4만개로 이뤄진 도조(Dojo)라는 슈퍼컴퓨터를 보유하고 있고, 해당 설비는 성능 면에서 전세계 플레이어 중 다섯 손가락에 꼽힌다. 슈퍼컴퓨터를 인하우스 방식으로 이렇게 운영하고 있는 자동차 회사는 전무후무하다. 하지만, 이에 만족하지 않고, 테슬라는 “도조 엑사포드(Dojo ExaPod)”라고 불리는 차세대 슈퍼컴퓨팅 비전을 공개했다.

도조라는 이름에서 알 수 있듯이(도조는 일본어로 무술학습을 위한 훈련시설), 자율주행 딥러닝 알고리즘 연산을 위한 맞춤형 슈퍼컴퓨터 플랫폼이다. 도조는 테슬라 차주들이 보내는 영상 데이터를 활용한 자율주행 모델 알고리즘 고도화에 초점을 맞추고 있다. 그런데 이미 슈퍼컴퓨터를 가지고 있음에도 불구하고 왜 테슬라는 새롭게 도조 엑사포드까지 만들어내려고 하는 걸까?

가장 큰 이유는 ‘전성비’다. 전력대비 성능이라는 말이다. 먼저 전력 소비 측면에서 도조 시스템 책임 엔지니어인 Bill Chang은 이렇게 말했다. "전례 없는 냉각 및 전력 효율성을 지원하기 위해 데이터센터 인프라의 모든 측면을 재검토해야 한다는 것을 깨달았다"고 말이다.

이것은 기존 시스템으로는 자율주행 모델 알고리즘을 학습시키는데 너무 많은 전력이 들었다는 뜻이다. 그리고 데이터센터의 전력 소모를 줄이려면 결국 가장 기본이 되는 연산 코어부터 다시 설계해야 했다는 말이다. 이런 고민 끝에 D1 칩이 탄생했고, D1 칩의 대역폭 병목을 줄이는 방식으로 계속 패브릭을 확장해 데이터센터 단계까지 이르게 된 것이다.

그리고 테슬라는 AI DAY 2022에서 트레이닝 타일(D1 칩 25개 탑재) 6개로 구성된 “시스템 트레이” 하나만으로 기존에 쓰던 GPU 랙(rack) 6개를 대체할 수 있다고 주장했다. 뿐만 아니라, 이 시스템 트레이 구성 비용이 1개의 GPU 랙보다도 비용이 적게 든다고도 강조했다. 말 그대로 성능도 비약적으로 올리고, 비용을 확 낮출 수 있었다는 말이다.

테슬라의 컴퓨팅 패브릭 구조 - “트레이닝 타일”이 병렬로 계속 확장
: 하나의 트레이닝 타일에는 “D1” 칩 25개를 포함(5X5 array)

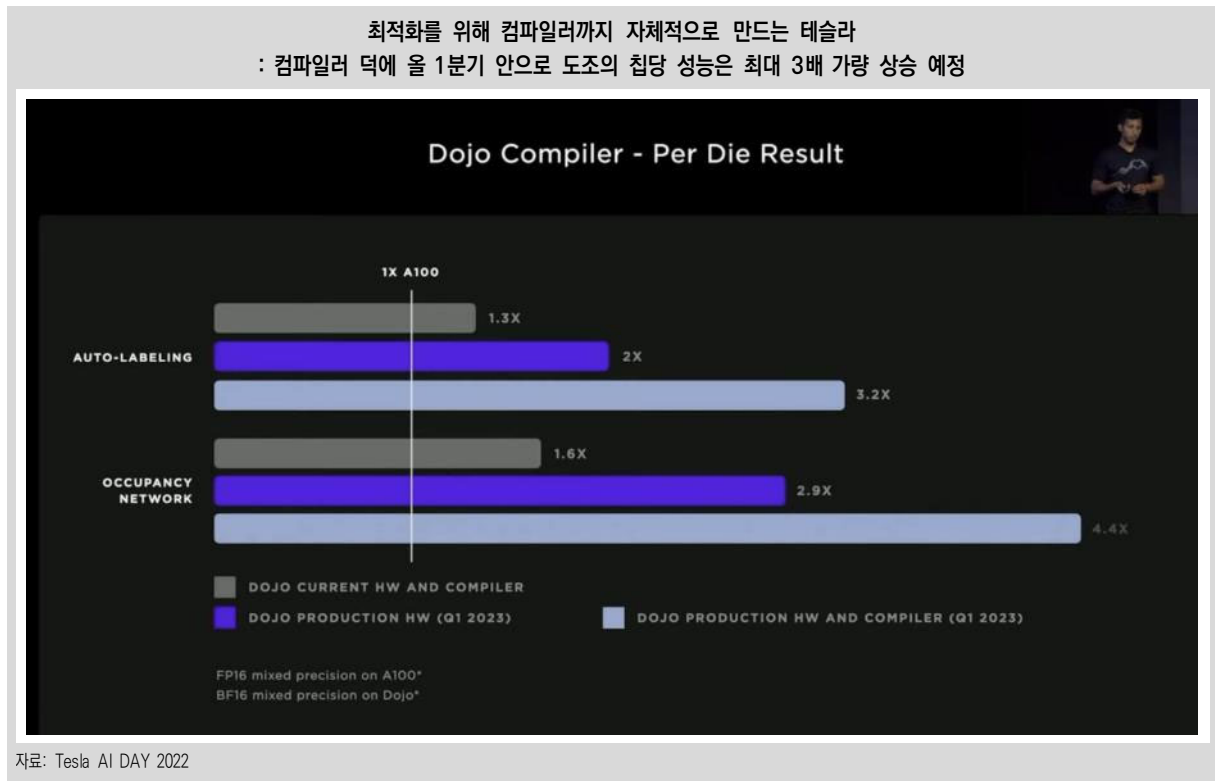


자료: Tesla AI DAY 2021

또한, 처음부터 다시 설계를 했기 때문에 테슬라는 소프트웨어까지 개혁을 단행했다. 슈퍼컴퓨터에서 연산된 딥러닝 알고리즘이 차량에서 동작할 때 최적화가 가능하도록 소프트웨어를 만든 것이다. 즉, 테슬라는 HW와 SW 모두에서 효율성을 잡기 위해 뉴럴 네트워크 컴파일러라는 것을 개발했다.

그리고 그 결과는 성공적이었던 것으로 비춰진다. AI DAY 2022에서 테슬라가 밝힌 정보에 따르면, 2023년 1분기 내에 컴파일러와 도조 엑사포드의 조합을 통해 최대 3배 가량의 성능 향상을 내보이겠다고 자신했다.

* 테슬라의 뉴럴 네트워크 컴파일러는 테슬라 차량 내부의 "HW"에 특화된 최적화를 제공하는 컴파일러. 컴파일러는 프로그래머가 작성한 코드를 기계가 이해하는 언어로 변환. 이 과정에서 코드를 효율적으로 실행할 수 있도록 하드웨어 플랫폼에 맞게 최적화. 도조에서 학습된 딥러닝 모델을 가지고 테슬라 차량 내부의 HW의 CPU와 GPU를 더 효율적으로 실행할 수 있도록(더 낮은 지연시간과 높은 처리량을 달성할 수 있도록) 변환하고 최적화.



그렇다면, 도조의 차세대 버전 엑사포드라는 것이 대체 어떻게 이뤄진 것인지 궁금증을 자아낸다. 테슬라가 도조 출시를 공언한 시점으로부터 1년 정도가 지난 시점인 AI DAY 2022에서 도조 엑사포드에 대해서 좀 더 상세한 정보들이 나왔다.

가장 먼저, 테슬라는 D1 칩과 트레이닝 타일에서 이제는 시스템 트레이와 전체 캐비닛으로 나아갔다고 투자자들에게 확인시켜 줬다. 2021년에는 트레이닝 타일까지만 현장에서 선보였는데 1년만에 진척된 것이다.

트레이닝 타일 6개가 모여 시스템 트레이 하나를 완성한다. 이때 단일 시스템 트레이가 "full-load 기준 3~4개의 슈퍼컴퓨터 랙"에 해당하는 성능을 낸다는 점은 공포스러운 정도였다. 트레이 하나의 무게만 135kg으로, BF16 기준 54 페타플롭(PFLOP)이라는 어마어마한 성능을 제공한다. 그럼에도 100kW 정도의 전력만을 소비한다.

“트레이닝 타일” 6개가 모여 하나의 “시스템 트레이”를 구성
: 0.1톤이 넘는 육중한 무게에 54 PFLOPs의 성능(BF16 기준)

System Tray

- HIGH-SPEED CONNECTIVITY**
- POWER + MECHANICAL + THERMAL**
2000A at 52VDC
- DENSE INTEGRATION**
75mm height to support 135kg
- 54 PFLOPS BF16/CFP8**
- 13.4 TB/S BISECTION BW**
- 100+ KW POWER**

DEK 75mm
54 P
13.4
100+

TESLA AI DAY

자료: Tesla AI DAY 2022

그리고, 이 시스템 트레이에 호스트 인터페이스까지 통합하여 더 큰 개념의 어셈블리를 만든다. 호스트 인터페이스는 하드웨어 비디오 디코더 지원을 하기 위해, 512개의 X86코어와 8TB의 메모리, 640GB/s의 PCIe 대역폭의 스펙을 가지고 있다. 그리고 ExaPOD에 들어갈 ‘단일 캐비닛’으로 구성하기 위해, 시스템 트레이 2개와 호스트 인터페이스 2개를 넣는다.

시스템 트레이를 구동시키고, 네트워킹을 지원하는 호스트 인터페이스
: 8테라바이트의 고성능 메모리와 초당 640기가바이트의 PCIe 대역폭을 구비

Dojo Host Interface

- INGEST PROCESSING**
PCIe connectivity to Interface Processors
Hardware Video Decoder Support
- USER APPLICATIONS**
x86 Linux Environment
User-Scheduled jobs
- 512 TOTAL X86 CORES**
- 8 TB TOTAL MEMORY**
- 640 GB/S PCIe BANDWIDTH**

800+ UserF
512 C
8 TB
640

TESLA AI DAY

자료: Tesla AI DAY 2022

그리고 10개의 캐비닛을 연결해야 완전한 차세대 도조의 모습을 갖추기 위한 “엑사포드(ExaPOD)” 하나가 만들어진다. 테슬라는 도조 엑사포드의 주요 사양도 공개했다. 1.1 엑사플롭(EFLOP)의 퍼포먼스에, 1.3 테라바이트의 SRAM 고속 메모리를 탑재했고, 13 테라바이트의 고대역폭 DRAM도 들어가 있다.



자료: Tesla AI DAY 2022

가장 중요한 점은, 도조 엑사포드가 2023년 1분기에 구축될 것이라는 점이다. 게다가, 테슬라는 팔로 알토에 총 7개의 엑사포드를 더 설치할 계획도 천명했다. 설치가 완료되면 이를 통해 8.8 엑사플롭스(BF16/CFP8 기준)의 성능을 제공할 것이라고 테슬라는 자랑하듯 말했다. 이 정도의 컴퓨팅 파워는 구글 오클라호마 데이터센터와 비교해 봤을 때도 경쟁력을 가지는 수준으로 볼 수 있다. 빅테크급의 데이터센터 컴퓨팅 파워를 단일 자동차 회사가 갖는 것이다.

이와 관련해 의미심장한 부분은, 테슬라가 앞으로 이를 클라우드 서비스화 할 수도 있다는 사실이다. 일론 머스크는 테슬라가 도조를 앞으로 클라우드 서비스로서 다른 회사들에 (호스팅하여) 제공할 수도 있다고 말했다. 이제 와 생각해 보면, 테슬라는 자율주행 알고리즘 연산을 위한 플랫폼을 기반으로, 새로운 비즈니스를 선점하려는 의도가 있었던 게 아닐까 한다. 즉, 테슬라는 이미 고객들의 차량에서 수집한 수백만 개의 영상 데이터를 기반으로 한 자율주행 시스템 개발이 높은 컴퓨팅 성능을 필요로 함을 알고 있었기에 이런 접근 방식을 택했던 것으로 풀이된다.

이와 관련해 일론 머스크는 아마존 AWS에 도전할 것이라고 언급했기 때문에, 클라우드 호스팅 서비스의 가능성은 빈말이 아닐 것으로 보인다. 아마존의 AWS가 전자상거래를 처리하고 남은 컴퓨팅 파워를 활용하는 과정에서 출발한 비즈니스였던 것처럼 머스크도 도조를 통해 AI와 연관된 사업기회를 새롭게 만들어 갈 것으로 보인다.

테슬라 케이스의 함의는 향후 AI 시대를 맞아 엄청난 데이터가 쏟아질 것이며 이를 처리할 전성비가 뛰어난 하드웨어가 필요하고, 이를 활용한 새로운 사업기회가 열릴 것이라는 점이다. 그리고 그 기반에는 시가 있다.

누가 연산 효율화에 진심인가

2021년 11월 기준, 전세계에는 대략 3,000개의 정부 및 학계 슈퍼컴퓨팅 센터가 존재한다. 그런데, 슈퍼컴퓨팅 센터 중 5~10%만이 AI 가속기를 가지고 있다. AI 가속기는 딥러닝 연산에 쓰이는 특정 활성화 함수에 특화된 컴퓨팅 칩을 의미한다. 물론 그 뒤로 시간이 많이 흘렀지만, 데이터센터에서 AI 가속기의 비중은 아직도 부족한 실정이다.

그런데, AI 산업화 시대는 우리의 기대보다 일찍 다가왔다. 이 말은 즉, 데이터센터의 연산력에 비해서 AI 연산에 대한 부분이 강력한 초과수요로 발생할 수 있다는 말이 된다. 컴퓨팅 파워가 감당할 수 없을 정도의 데이터가 쏟아져 들어오면, 학습 속도는 무한정 길어질 수 있다.

이에 대한 해법으로 빠르게 데이터센터를 더 짓고 GPU나 AI 가속기를 도입하려는 시도가 있을 것이다. 하지만, 이 과정에서 물품 가격은 천정부지로 뿔 수 있고 당장 큰 자본지출로 이어져 부담을 느낄 수 있다. 그리고 컴퓨팅 유닛을 늘릴 때마다 데이터센터에서 사용하는 전력량이 계속 증가해 점차 감당하기가 어려워질 수도 있다. 그래서, 데이터센터 사업자들이 느끼는 가장 큰 유혹은 연산량 자체를 줄이는 “효율화”일 것이라 판단한다.

1) 테슬라의 “도조 백서”

테슬라는 2021년 AI DAY를 개최한 뒤, 얼마 지나지 않은 10월에 “도조 백서(white paper)”를 발간하였다. 테슬라는 9페이지 분량의 논문에서 새로운 이진 부동소수점 산술방식을 발표했다. 딥러닝 자율주행 모델훈련을 위해서 이것이 필요하다는 이유에서였다. 대체 이게 뭘까?

이진 부동소수점 산술이란, 숫자로 소수점까지 나타낼 때의 자리배치 방법을 뜻한다. 부동소수점은 컴퓨터에서 실수를 표현하고 계산하기 위한 방식이다. 예를 들어 컴퓨터 그래픽이 어떤 물체의 좌표값을 계산하려고 하면, 각 물체의 위치를 정확히 벡터로 계산해야 한다. X, Y, Z축 각각의 벡터값은 물체가 움직일 때마다 변환된다. 그런데 이 벡터값은 딱딱 떨어지는 정수는 아닐 것이다. 이 세상은 정수로 떨어지지 않고 거의 전부 실수로 표현되기 때문이다.

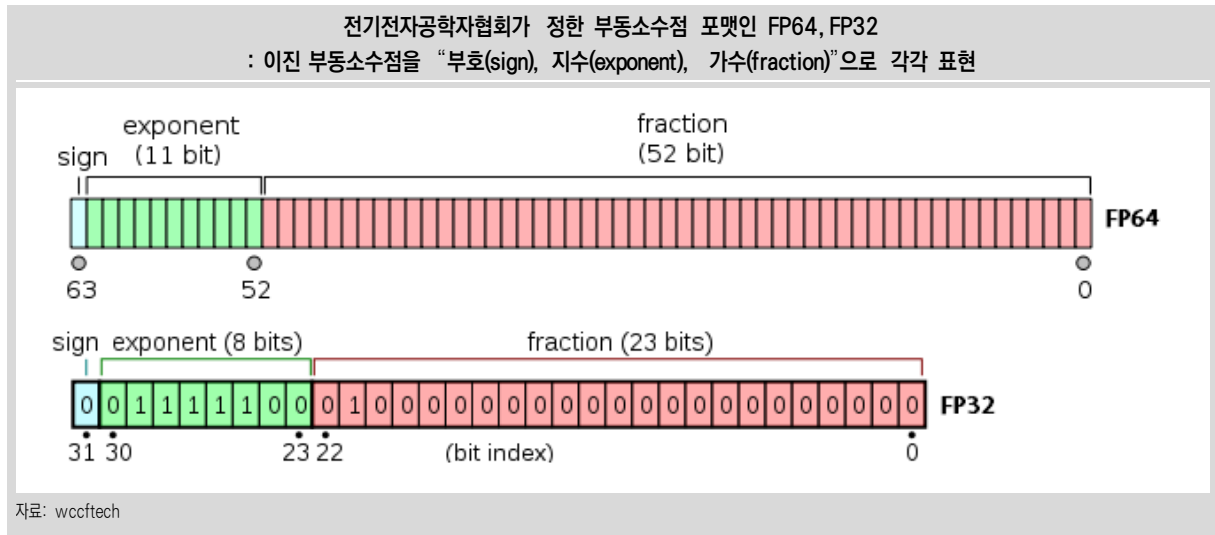
컴퓨터 그래픽 처리에 부동소수점이 쓰이는 이유 예시
: 컴퓨터 속에서 각 물체는 삼각형(폴리곤)으로 표현한 뒤 각 꼭지점에 좌표값을 표시

삼각형을 이용해서 그물처럼 서로 연결해 가면서 3차원이기 때문에 x, y, z축, 3개의 위치 값이 필요합니다

자료: YouTube “bRd 3D”

즉, 부동소수점 방식은 넓은 범위의 값을 표현할 수 있어서 과학, 공학, 그래픽 등 여러 분야에서 널리 사용되는 방법이다. 물론, 이것은 컴퓨터가 알아볼 수 있도록 이진법으로 나타낸다. 정리하면, 실수를 컴퓨터상에서 표현할 때 0과 1을 이용해서(근사하게) 표현하는 방식이다.

여기서 부동소수점은 FP라고 불리는데, 'Floating Point'의 약자다. 이 FP에다가 뒤에 숫자를 붙여 FP32와 FP16라고 표현들을 많이 한다. 각각 32비트와 16비트 부동소수점으로 숫자를 표현하는 방식이다.



테슬라는 도조 백서에서 딥러닝 훈련에 적합한 환경을 만들기 위해서, 기존에 쓰이던 2가지 유형이 아닌 본인들만의 새로운 형식(format)을 창조해냈다. 세계 전기전자공학자협회가 만들어 수십년간 표준으로 쓰이던 포맷 형태를 배제하고, 일개 자동차 회사가 새롭게 만들었다는 뜻이다. 본인들이 필요한 연산 체계를 맞춰 데이터 포맷을 재정립한 것이다.

본인들의 도조 데이터센터가 갖고 있는 컴퓨팅 능력을 감안해서, 시스템 차원의 병목현상을 최소화하기 위해 이런 부분까지 건드린 것이라 보면 된다. 테슬라가 얼마나 소프트웨어 역량이 뛰어나고 인공지능 연산에 진심인지 알 수 있는 대목이기도 하다.

테슬라가 개발한 체계는 세가지가 있다. BF16, CFP8_1_4_3, CFP8_1_5_2 포맷이라고 불리는 것 (Configurable Floating Point)들이다. 각각의 내용을 알아두기보다는, 핵심만 이해하면 된다. 결국, 테슬라가 개발한 새로운 포맷 형태들은 더 적은 비트를 필요로 하기 때문에 메모리의 스토리지와 대역폭을 늘리는데 적합하다.

테슬라가 개발한 이진 부동소수점 표현형식(format)들과 FP16 비교

	개발기관	총 비트 수	부호 할당 비트 수	지수 할당 비트 수	가수 할당 비트 수
FP16 (Half-Precision Floating Point 16)	IEEE	16비트	1비트	5비트	10비트
BF16 (BrAIn Floating Point16)	테슬라 + 구글	16비트	1비트	8비트	7비트
CFloat_1_4_3	테슬라	8비트	1비트	4비트	3비트
CFloat_1_5_2	테슬라	8비트	1비트	5비트	2비트

자료: Tesla Dojo White Paper, 미래에셋증권 디지털리서치팀

주: 지수 할당 비트 수가 높을수록 넓은 값의 범위 표현이 가능하며, 가수 할당 비트 수가 높을수록 정밀도가 높고 계산오차가 줄어듦

특히 CFP의 경우 32비트도, 16비트도 아닌 8비트이기 때문에 그만큼 메모리 사용량이 적고 연산속도를 높일 수 있기 때문이다. 이를 통해 연산량이 많을 수밖에 없는 딥러닝 및 자율주행에 적합하다.

또 BF16의 경우에는 기존 FP16보다 지수에 할당된 비트 숫자가 많다. 지수의 범위가 넓다는 뜻인데, 이는 보다 더 넓은 값의 범위를 표현하는 것이 중요한 딥러닝에 적합한 포맷이다. 물론 가수(fraction)의 범위는 줄어들었기 때문에 정밀도는 다소 떨어진다. 하지만, 딥러닝 작업에서는 넓은 값의 범위를 표현하는 게 정밀도 보다 중요한 경우가 많다.

실제로 테슬라는 BF16을 사용해 딥러닝 모델의 성능을 올리고 에너지 효율성까지 개선하는데 성공했다고 밝혔다. 딥러닝을 위한 컴퓨팅 파워와 비용, 그리고 성능을 감안한다면 이런 노력을 하는 회사가 시에 진심과 실력을 가지고 있음을 짐작할 수 있다.

2) 엔비디아의 “텐서 코어”

테슬라가 자율주행에서 한 가닥 하는 리딩 업체라고 한다면, AI에 있어서 가장 굴지의 경쟁력을 가지고 있는 회사는 단연 엔비디아다. 이 AI-native 회사도 테슬라가 가진 비슷한 고민을 먼저 했다고 판단한다. 그리고 그 고민을 다른 유형의 코어를 발견함으로써 해결하려고 한 것으로 보인다.

엔비디아 GPU에서 주력으로 쓰이는 연산 코어의 이름은 CUDA 코어다. 그래서 엔비디아의 그래픽카드 제품군들의 성능을 따질 때 CUDA 코어가 몇 개나 들어가 있는지를 확인하곤 한다. 코어의 개수가 사실상 성능을 가장 크게 좌우하기 때문이다. GPU의 경우에는 보통 수천개의 코어가 들어가게 되는데, 단순한 연산의 병렬처리가 많은 방식이기 때문에 그렇다. 그래서, 깐깐한 소비자들의 경우는 코어 개수 당 그래픽카드 가격을 고려해서 가성비를 논하기도 한다. 이처럼, 엔비디아 GPU에 있어서 CUDA는 그래픽 연산 처리의 근본이다.

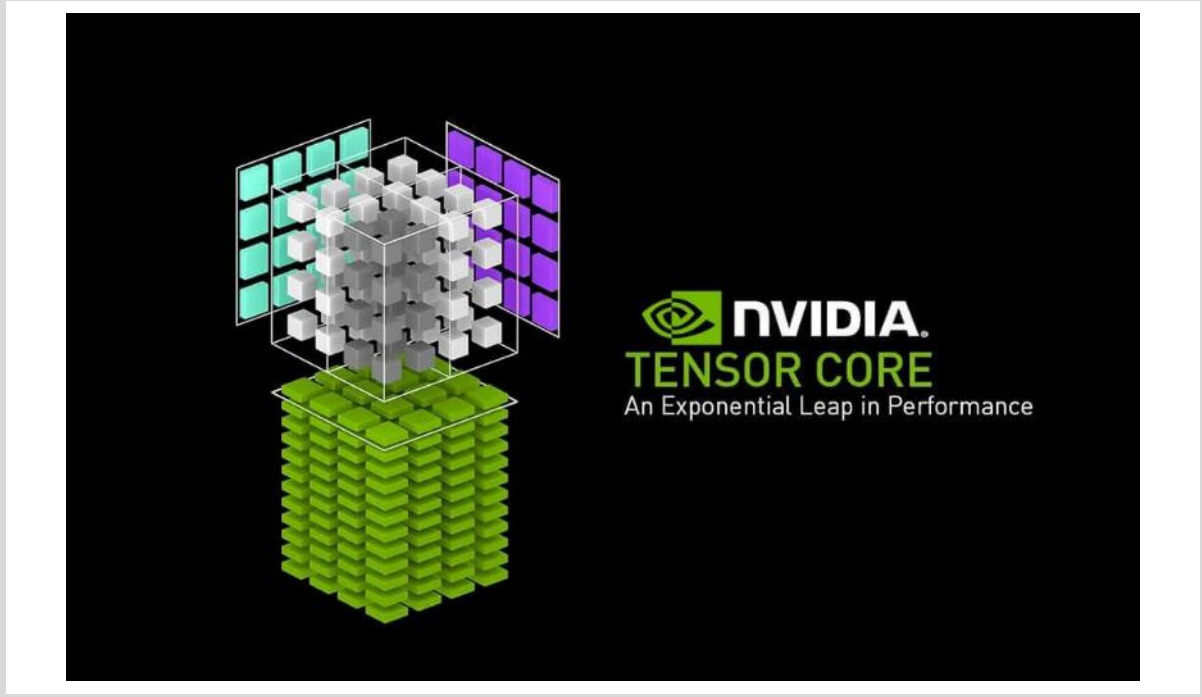
이런 병렬화가 잘 된다는 특성 때문에 CUDA 코어는 인공지능 작업에서 주목을 받았다. GPU 코어는 원래 물리 및 그래픽 계산을 위해 설계되었는데, 신경망의 행렬적 특성에 있어서 GPU가 매우 적합했던 것이다.

**클럭(Clock)은 신호 처리량을 뜻함. 단위는 헤르츠(Hz). "1초에 몇 번의 전기적 신호"가 발생했는지를 표현. 1초에 1번 신호 발생하는 것이 1Hz. CPU는 메인보드에서 공급받는 전기를 신호로 받아 연산작업을 함. 전기적 신호는 '오르고 내리는 형태'이며 이는 0, 1의 디지털 신호로 표시.*

그런데 CUDA는 하나의 GPU 클럭에 하나의 FP32 부동소수점 연산만을 수행한다. 일전에 봤듯이 FP32는 32비트를 표현하는 부동소수점 표준 형식을 뜻한다. FP32는 데이터를 광범위하게 표현하면서 정확하게 표현할 수 있어 과학/공학 연구용으로 활용된다. 하지만, 그만큼 정확도를 다루다 보니 모델의 크기가 커다는 단점이 있다. 더 많은 비트를 사용해서 연산을 하므로 연산속도가 느려지고 메모리도 많이 써야 한다.

쉽게 말해 딥러닝에서는 최적의 성능을 발휘하지 못할 수도 있다는 것이다. 이런 한계로 인해 엔비디아는 새로운 코어를 추가로 개발했다. 이로써 2017년에 처음으로 상용화된 코어가 텐서 코어(Tensor core)다. 그러니까 텐서 코어는 인공지능에 최적화된 행렬 연산을 빠르게 할 수 있는 코어다

엔비디아의 딥러닝 전용 연산코어 “텐서코어”
: 딥러닝 연산의 근본인 병렬 연산에 최적화된 형태



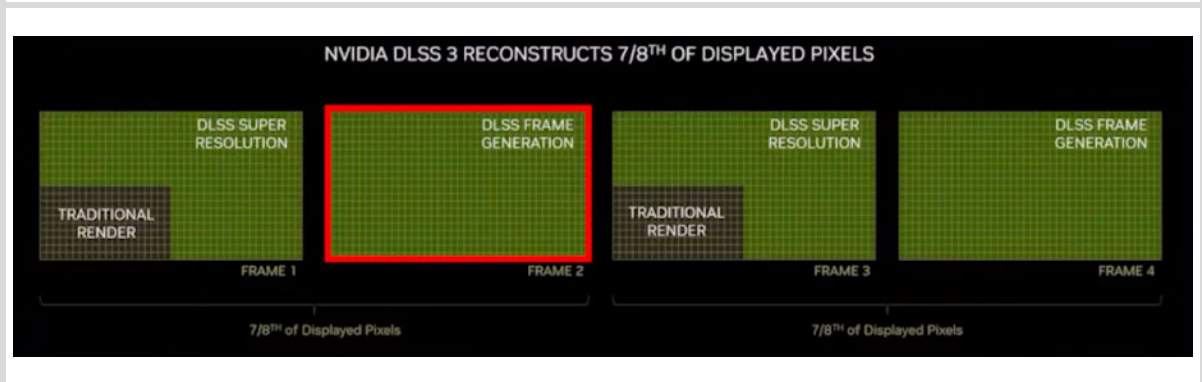
자료: Nvidia

현재 엔비디아의 주력 제품인 RTX GPU에는 텐서 코어도 쿠다 코어와 함께 추가로 탑재되어 있다. 그래픽카드 안에서 텐서 코어는 그래픽 성능을 강화해 주는 AI 렌더링 기술인 DLSS를 적극 지원하는 역할을 한다. 그리고 이 텐서 코어의 능력 덕분에 엔비디아의 그래픽카드 성능 항상 폭은 다른 회사를 압도한다. 직접적 경쟁상대인 AMD조차 엔비디아의 최고 하이엔드 제품인 RTX 4090에는 명함도 내밀지 못한다.

더군다나 텐서 코어와 그래픽 드라이버 간 AI 최적화를 통해 성능이 지속력과 폭발력 있게 상승한다는 게 더 무서운 점이다. AI를 통해 엔비디아는 GPU 성능을 추가적으로 최대 30%까지 끌어올릴 것으로 전망하고 있다.

**DLSS: Deep Learning Super Sampling의 줄임말. 딥러닝의 힘을 활용해 원활한 게임 진행을 위해 초당 프레임 숫자를 향상시키고 더 분명한 이미지를 생성시킬 수 있게 하는 기술*

엔비디아의 DLSS 3이 내는 성능
: 슈퍼 레졸루션(업스케일링)과 프레임 생성 기능이 필요연산의 약 88%를 알아서 처리



자료: Nvidia

엔비디아 RTX GPU(텐서코어 탑재)가 인공지능으로 얻는 성능향상
 : 그래픽드라이버 성능향상폭 최대 30% / DLSS 3.0으로 게임 성능은 타의 추종을 불허

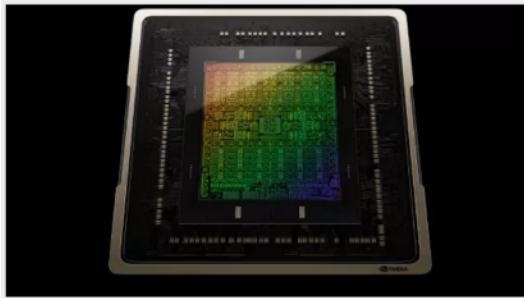
엔비디아가 준비중인 새로운 AI 혁신

Rumoured AI-enhanced Nvidia graphics driver could boost performance by 30%

By Jeremy Laird published January 10, 2023

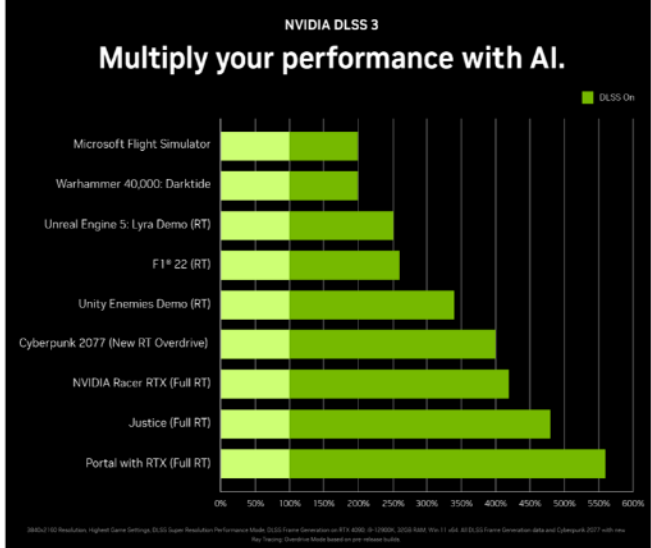
Yet more AI special sauce from Nvidia.

COMMENTS



자료: PC Gamer, Nvidia, 미래에셋증권 디지털리서치팀

엔비디아의 DLSS 3를 사용했을 경우 퍼포먼스 향상폭



텐서 코어는 '4x4 행렬 연산'을 수행하는 GPU 코어다. 텐서 코어는 1 GPU 클럭에 4x4 형태의 FP16 행렬 2개를 곱하고, 그 결과를 4x4짜리 FP32 행렬에 더하는 matrix multiply-accumulate 연산을 수행한다. 따라서, 이 과정은 FP16 행렬을 입력으로 받은 다음, FP32 행렬로 출력을 반환하기 때문에 mixed precision이라고 불린다.

*matrix multiply-accumulate: FMA(Fused Mutiply Add)라고 불리기도 함.

CUDA 인터페이스에서 Tensor 코어가 구동되는 방식 - 4X4 행렬연산
 : CUDA 코어보다 정확도는 조금 떨어지나 빠르다는 강점으로 인공지능 학습에 최적

$$\mathbf{D} = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 FP16 FP16 FP16 or FP32

자료: Nvidia

또한, 텐서 코어는 딥러닝 프레임워크와 관련한 연산 유형에 훨씬 더 전문화되어 있다는 장점도 있고, 쿠다 라이브러리도 텐서코어를 지원해 호환성 측면에서도 뛰어나다.

결론적으로, 텐서 코어는 쿠다 코어보다 정밀도를 대가로 연산력을 훨씬 적게 사용하지만, 정밀도 손실 부분에서 최종출력에 큰 영향을 미치지 않는 구조를 가졌다. 이런 장점은 딥러닝 모델에서 텐서 코어가 비용 절감에 더 효과적이라는 것을 의미한다. 즉 정확도를 조금 희생해서라도 전성비가 더 소중하기도 하다는 말이다.

그리고 이 부분에 진심인 기업들은 딥러닝 모델 성과에서 향후 큰 약진을 보일 것이다.

데이터센터는 결국 반도체의 구성물

1) AI와 반도체의 공통법칙 “스케일링 법칙”

우리는 인공지능 분야에서 스케일링 법칙(scaling law) 현상이 벌어지고 있는 것에 주목해야 한다고 생각한다. 쉽게 풀어서 설명하면 딥러닝 모델의 크기를 키우면 키울수록 훌륭한 AI가 탄생하고 있다는 이야기다. 딥러닝의 모델 크기는 파라미터 개수로 인식되곤 한다. 은닉층에서 노드들이 이어져서 그 가중치(Weight)들을 나타내는 파라미터 숫자가 많아야 큰 모델이라는 뜻이다.

**스케일링 법칙(scaling law): 생물의 크기가 2배 늘어날 때마다 25%의 에너지가 절약된다는 규모 증감의 법칙*

여기서 최근에 나온 GPT-4의 경우를 한 번 생각해보면 얼마나 데이터 연산이 많이 들어갈지 짐작이 된다. GPT-4의 경우 약 1조~10조 개(혹은 100조 개)의 파라미터 숫자를 가지고 있는 것으로 추정되고 있다. GPT 모델은 각 파라미터에 대해 추론, 가중치 비교, 최적화 분산과 모멘텀이라는 4개의 파라미터가 각각 필요하며 파라미터는 각각 4바이트 정도의 파일크기를 가지고 있다.

따라서, 그 둘을 곱하게 되면 필요한 용량은 16TB~160TB에 육박한다. 16TB~160TB의 용량을 커버하기 위해선 엔비디아 H100에 들어가는 것으로 알려진 SK하이닉스의 최신형 HBM3의 용량이 80GB이므로 고성능 GPU가 200장에서 2,000장 가량 필요하다.

고성능 GPU H100의 가격이 약 3.3만 달러로 알려진 가운데 이를 역산해보면 서버 비용만 86억~860억 가량 들어간다는 소리다. 여기에 학습과 추론에 들어갈 각종 비용을 계산해보면 이 같은 스케일링 법칙이 지속될 수 있을지 의문스러울 수 있다.

AMD CEO 리사수가 강조한 차세대 반도체 연산을 위한 Tripod

“ 연산칩, 메모리칩, 그리고 네트워킹이 함께 해야 된다 ”
- AMD CEO 리사수, 2월 20일 국제 고체회로 학회(ISSCC)에서...

IEEE International Solid State Circuits Conference | February 20, 2023

자료: IEEE, AMD, 미래에셋증권 디지털리서치팀

그렇기 때문에 지속적으로 하드웨어의 기술 혁신이 필요하며 단순히 칩의 성능을 올리는 것뿐만 아니라 AMD 리사 수 CEO의 주장처럼 반도체 시스템 전체의 효율성을 높여야 한다. AMD의 System in Package가 함의하는 것은, 미래에는 통합형 아키텍처가 대세가 될 것이라는 점이다.

AMD가 말하는 미래의 System in Package 구조
: 다양한 칩들이 다닥다닥 한데 어우러진 'Heterogeneous'

Future System-in-Package Architecture

- Advanced packaging enables maximally efficient integration of compute elements and memory
- System level communication accomplished with low-power, high-bandwidth optical

© 2023 IEEE International Solid State Circuits Conference | February 20, 2023

자료: IEEE 2023

AMD의 데이터센터용 GPU 아키텍처인 "CDNA"에 적용된 2.5D 패키징 기술
: 최신 MI300 APU(CPU+GPU 통합형), AI 훈련성능에 있어서 8배 성능 증가

2.5D "BRIDGE" ARCHITECTURE LANDSCAPE

Substrate Embedded 2.5D

Elevated Fanout Bridge 2.5D

LOCALIZED INTERCONNECTS
Bridge Technologies

BETTER ELECTRICALS
Lower Parasitic Capacitance

Compared to Si Interposer Based 2.5D

SCALABLE SOLUTION
Lithographically Defined

STANDARD SUBSTRATES
Lower Cost

STANDARD FLIP CHIP PROCESS
Lower Complexity Bumping Assembly Process

Compared to Substrate Embedded 2.5D

자료: AMD

앞으로 칩 하나에 CPU, GPU, NPU, 특정 엑셀러레이터, I/O 칩셋, 램 등이 모두 다 같이 들어갈 것이다. 모든 부품이 하나로서 작동하기 위해선 연산 칩의 성능과 더불어 메모리와 네트워킹 모두 함께 개선되고 통합되어야 한다. 이에 우리는 AI 시장이 확대될수록 모든 반도체 시스템도 동반 성장할 수밖에 없을 것이라고 전망한다. 우리는 이번 장에서 주요한 AI 반도체로 GPU, NPU, 뉴로모픽, 네트워크 반도체 중심으로 설명하려고 한다.

2) AI 산업화 시대의 총아, GPU

a. AI 연산은 GPU, GPU는 엔비디아

AI가 본격적으로 발전하면서 가장 각광을 받은 반도체를 하나만 꼽자면 단연 GPU라고 볼 수 있다. 왜냐하면 인공지능 연산을 효율적으로 하기 위해서는 병렬연산이 필수적이기 때문이다.

GPU는 벡터 개념이 들어가는데, 예를 들어 3D 게임을 하면 상하좌우로 캐릭터가 움직이고, 총탄에 맞아서 날아가는 등 한번에 많은 일들이 벌어진다. 이를 3차원의 공간으로 표현하려면 동시에 많은 "좌표에 대한 계산"이 필요하다. 이런 좌표 계산에는 반도체가 여러 연산을 동시에 할 수 있어야 하며, 여기에서 GPU가 강점을 가진다.


GPU 시장에서 엔비디아가 1위가 된 지는 약 25년이 다 되어 간다. 무려 1999년 GeForce 256으로 게임용 그래픽카드 시장에 진출한 뒤부터였다. 하지만, 엔비디아는 이런 입지에 안주하지 않았고, 인공지능의 미래에 대해 가장 강하게 배팅한 회사 중 하나였다. 엔비디아는 이미 10여 년 전부터 “인공지능이 모든 것을 바꿀 수 있다는 것을 미리 알아챘고, 그래서 회사의 모든 면을 인공지능에 초점을 맞춰 전환했다”라고 밝히기도 했다. 그래서 젠슨 황 회장이 ChatGPT의 출현을 반겼던 것이라고 할 수 있다. 그만큼 오래 준비했기 때문이다.

ChatGPT와 같은 대형 언어 모델은 엄청난 규모의 데이터와 파라미터를 사용해야 하기 때문에 병렬 연산이 꼭 필요하다. 그런 측면에서 ChatGPT가 런칭되고 전례 없는 유행을 보이고 있다. 이에 따라 향후 데이터 처리량의 증가에 따른 GPU의 추가 수요가 발생할 것은 자명하다. 데이터센터 시장의 GPU 수요는 연간 23.5%의 속도로 증가하고 있는데, 앞으로 이 속도는 추가적으로 더 늘 수 있다는 말이다.

예를 들어 구글이 제공하는 검색에 만약 ChatGPT를 적용시키기 위해서는 512,820.51개의 A100 HGX 서버가 필요한데, 여기에는 4,102,568개의 A100 GPU가 포함된다. 엄청난 양과 비용이다. 이렇듯 LLM의 도입은 기업에게 새로운 기회이기도 하지만 상당한 재정적 부담을 주기도 한다.

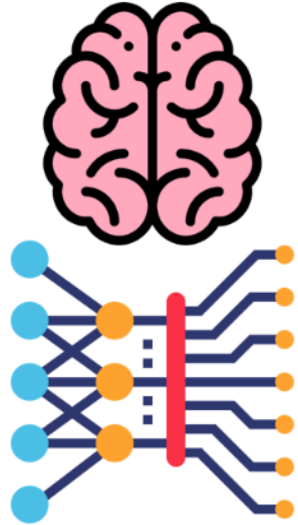
엔비디아는 이러한 기업이 느낄 수 있는 비용을 줄이기 위해 지속적으로 GPU 기술을 고도화하고 있으며 (무어의 법칙과 유사하게 제품 성능을 지속적으로 업그레이드) GPU의 사용처를 늘리기 위한 새로운 서비스도 계획 중이다.

“AI 산업화 시대”에 반도체 관련 자본지출 상승은 자명



증장기적 흐름

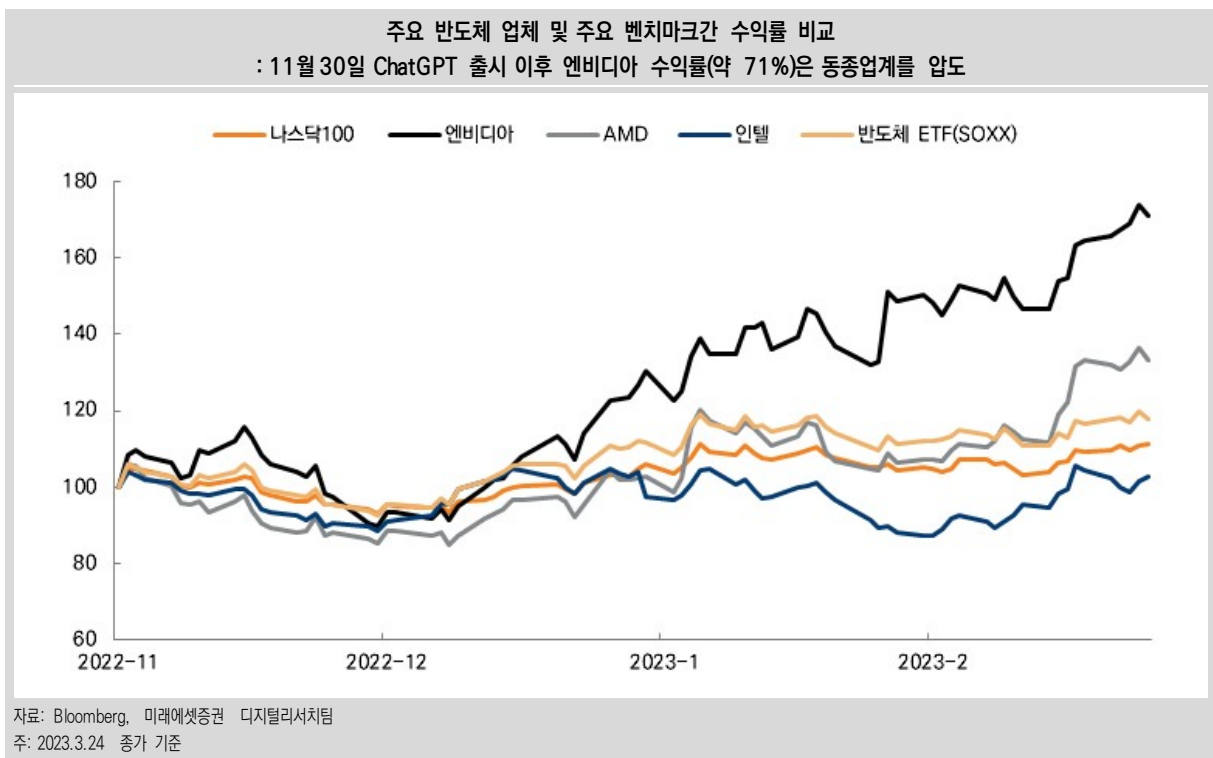
➔



만약, 구글이 본인들의 검색요구에 AI 기능을 통합시킨다면,
약 51만개의 서버와 약 410만개의 A100 GPU가 필요함

자료: flaticon, 미래에셋증권 디지털리서치팀

인공지능 영역에서 엔비디아가 지니는 입지를 모두 인정하고 있기 때문에, AI 산업화 시대 개화에 힘입어 최근 주가 상승률은 다른 기업들을 압도한다. AMD보다 밸류에이션이 이미 높음에도 반등의 폭이 2배 넘게 높다. 그만큼 투자자들은 엔비디아를 단순 그래픽카드 판매업자로 보는 것이 아니라 AI에 대한 수혜로 인식을 하고 있다.



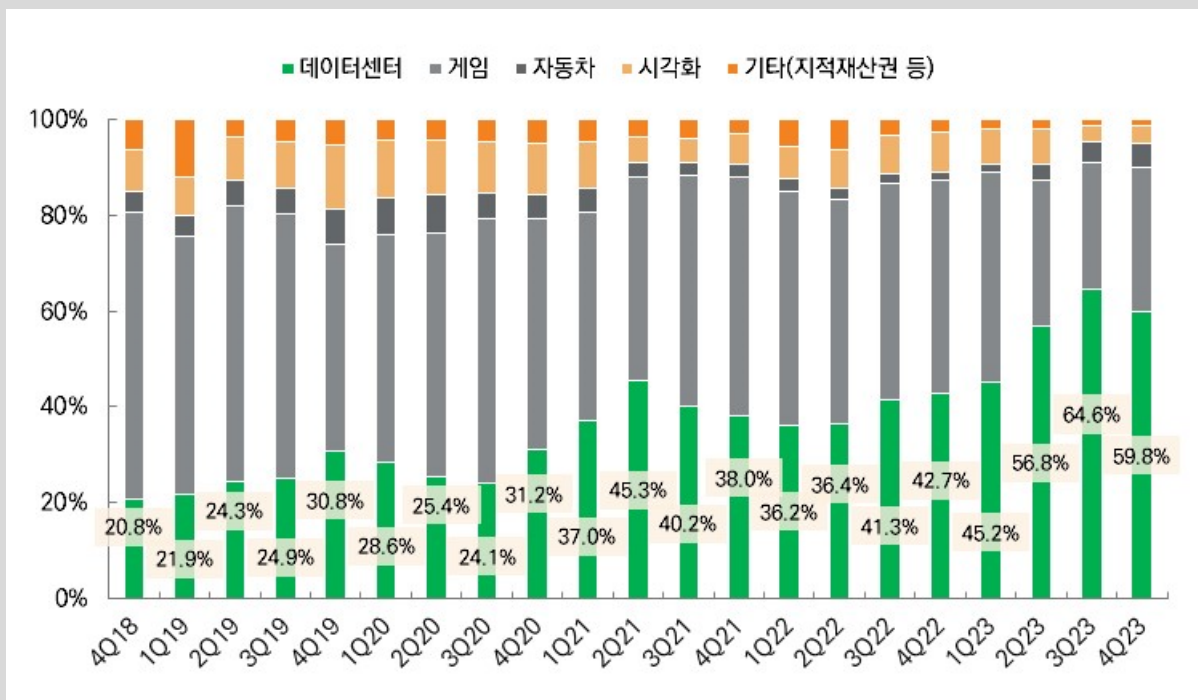
엔비디아 데이터센터용 GPU 제품별 하이퍼스케일러 고객사 리스트
 : 미국, 중국 국적의 주요 하이퍼스케일러들이 대부분 도입하는 GPU가 엔비디아 제품군

	P4	P40	P100	T4	RTX 6000/8000	V100	A10	A40	A100	NGC
Alibaba Cloud	✓		✓	✓		✓			✓	✓
AWS				✓		✓	✓		✓	✓
Baidu Cloud	✓	✓		✓		✓			✓	
Google Cloud	✓		✓	✓		✓			✓	✓
IBM Cloud			✓			✓				
Microsoft Azure		✓	✓	✓		✓	✓		✓	✓
Oracle Cloud			✓			✓			✓	✓
Tencent Cloud	✓	✓		✓		✓			✓	✓
NPN CSPs			✓	✓	✓	✓		✓	✓	

자료: Nvidia

이러한 시장의 생각은 일리가 있다는 판단이다. 엔비디아는 데이터센터 내 AI 연산에 필요한 핵심 구성물인 범용GPU를 판매한다. 엔비디아는 데이터센터 GPU 시장 점유율의 90% 이상을 차지하고 있는 기업이다. 엔비디아의 매출 구성을 보면 5년 전만 하더라도 데이터센터 관련 매출 비중이 20%에 불과했는데 현재는 60%에 육박한다. 5년 만에 기업의 체질이 확 바뀌었다는 것을 의미한다. 그리고 이 비중은 당분간 더 높아질 가능성이 크다. 비로소 올해가 AI 산업화 시대의 원년이기 때문이다.

엔비디아의 사업부별 매출 점유율 추이



자료: Bloomberg, 미래에셋증권 디지털리서치팀

AI 산업화 시대에 엔비디아는 어떤 제품을 만들어야 하는지 누구보다 잘 알고 있다. 엔비디아도 직접 생성형 AI 모델을 만드는 장본인이기 때문이다. 딥러닝 네트워크를 구성할 때 어떤 기법을 써야 하는지 어떻게 최적화할지를 관련 업계에서 가장 잘 알고 있는 기업 중 하나라는 말이다.

실제로 이미지 생성형 AI 모델에서도 강점을 보이고 있다. 아래의 사진은 Stable Diffusion, Dall-E2, 엔비디아의 text-to-image 생성 모델인 eDiff-I의 결과물을 비교한 것이다. 엔비디아의 모델이 Stable Diffusion이나 Dall-E2보다 입력된 프롬프트의 요청사항을 정확히 반영한 이미지를 생성해냈음을 확인할 수 있다. GTC 2023에서 발표된 Picasso에 기반해 제작된 Edify 모델에는 text-to-image 모델인 eDiff-I, text-to-3D 모델 Magic3D, 이미지 생성 모델 GauGAN이 포함된다.

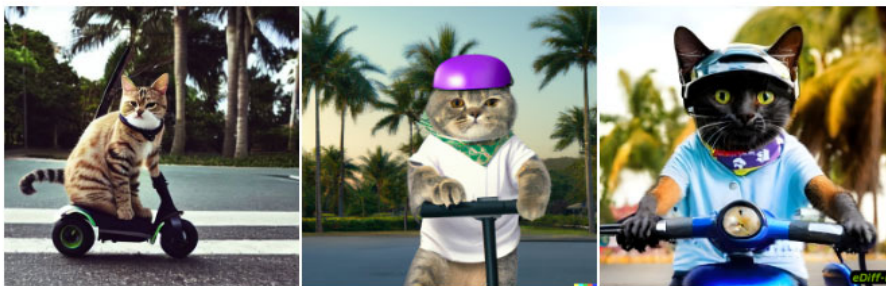
엔비디아가 내놓은 이미지 생성형 AI 모델 “eDiff-I”: 스테이블 디퓨전과 DALL-E 2과 성능비교



A photo of two monkeys sitting on a tree. They are holding a wooden board that says "Best friends", 4K dslr.



A photo of a plate at a restaurant table with spaghetti and red sauce. There is sushi on top of the spaghetti. The dish is garnished with mint leaves. On the side, there is a glass with a purple drink, photorealistic, dslr.



A close-up 4k dslr photo of a cat riding a scooter. It is wearing a plain shirt and has a bandana around its neck. It is wearing a scooter helmet. There are palm trees in the background.

자료: Nvidia 논문 "eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers"

주: 그림 왼쪽부터 Stable Diffusion, DALL-E 2, eDiff-I 순서

따라서, 어떻게 연산 코어를 구성해야 하는지도 제일 잘 알고 있다. 본인이 수요자이자 공급자인 것이다. 젠슨 황 CEO는 엔비디아의 최신 데이터센터 AI용 GPU인 H100를 두고 이런 말을 했다. "H100은 트랜스포머를 위한 새로운 유형의 설계로서, H100은 거대한 규모에서 트랜스포머를 처리하기 위해 설계된 세계 최초의 컴퓨터입니다."라고 말이다. 트랜스포머의 유행과 엔비디아는 함께 할 가능성이 높다. 그리고 트랜스포머는 현존하는 최고의 딥러닝 기법이다.

엔비디아는 반도체 생산뿐만 아니라 소프트웨어 개발에도 적극 투자해왔다. 인공지능에 관한 전천후를 다 준비한 셈이다. 이로써 자사 GPU에 최적화된 소프트웨어를 출시함으로써 하드웨어 활용도를 높이고 고객 락인효과를 강화해왔다. 예를 들어, PyTorch(파이토치)와 같은 딥러닝 프레임워크, 딥러닝 분석 가속화를 목적으로 개발된 DGX 시스템 등은 엔비디아가 AI 연산에 최적화된 솔루션을 공급하는 회사로서의 입지를 강화하는 데 기여했다.

엔비디아는 팍리스를 넘어 AI 종합 솔루션 기업으로 변모하고 있다. 특히 엔비디아는 이번 GTC 2023에서 자사 하드웨어 제품에 최적화된 AI 솔루션을 공개한 점이 인상적이다.

b. GTC 2023: 반도체 회사에서 AI 종합 솔루션 기업으로

엔비디아는 GTC 2023에서 DGX 클라우드와 AI Foundation을 발표했다. DGX 클라우드는 DGX AI 슈퍼컴퓨터에 접근할 수 있도록 하는 클라우드 서비스로, 마이크로소프트 Azure, 구글 GCP, 오라클 OCI와의 파트너십을 통해 개발되었다.

DGX AI 슈퍼컴퓨터는 DGX 시스템에 기반해 만든 딥러닝에 특화된 슈퍼컴퓨터이다. DGX 플랫폼은 자사 칩을 집적해 만든 서버(DGX 시스템)를 구축하고, 인공지능 모델을 개발하고 학습시킬 수 있는 소프트웨어와 사후관리 서비스를 종합적으로 제공하는 컴퓨팅 플랫폼이다.

DGX 클라우드는 DGX 플랫폼을 마이크로소프트, 오라클과 같은 클라우드 사업자 내부에 구축한 것이다. DGX 클라우드를 활용하면 AI 컴퓨팅에 필요한 대용량 서버(하드웨어)가 구축되어 있지 않은 기업들도 브라우저에서 즉시 AI 슈퍼컴퓨팅을 진행할 수 있게 된다. 구독료를 지불하는 기업들에게 DGX AI 슈퍼컴퓨터 사용권한을 부여하는 것이다.

AI Foundation은 고객들의 특수한 니즈에 부합하는 맞춤형 생성형 AI 모델을 개발할 수 있는 서비스이다. 기업들마다 개발하고자 하는 AI 모델의 특징과 기반 데이터가 다른데, 엔비디아는 고객별 맞춤형 AI 모델을 구축할 수 있는 소프트웨어를 제공한다.

엔비디아는 이번 GTC에서 총 세 가지 서비스를 공개했는데, NeMo, Picasso, BioNeMo이다. NeMo는 챗GPT와 같은 사전훈련 LLM(대형 언어모델), Picasso는 이미지, 영상, 3D 어플리케이션에 특화된 생성형 AI 모델, BioNeMo는 3D 단백질 구조 예측 등 신약개발을 위한 AI 모델 구축을 지원한다.

엔비디아는 모델 구성에 필요한 소프트웨어의 모음인 AI 라이브러리뿐만 아니라 AI 전문인력 파견, 기술 운영 및 적용에 이르기까지 AI 도입에 필요한 모든 과정에 대한 서비스를 제공한다. DGX 클라우드와 AI Foundation은 하드웨어 인프라와 AI 모델 개발 솔루션을 구독형 서비스로 제공함으로써 고객에게 최적화된 AI 모델 개발에 대한 접근성을 높이고 락인 효과를 발생시킨다.



One Architecture 플랫폼은 개별 워크로드에 특화된 네 개의 솔루션을 통합한 플랫폼이다. 한 가지 유형의 솔루션은 특정 워크로드를 빠르게 수행하기에는 적합하지만, 다양한 모델이나 데이터 유형을 처리할 수 없다는 단점이 있다. 엔비디아가 이번 GTC에서 발표한 One Architecture 플랫폼은 하나의 아키텍처를 기반으로 만들어진 네 가지 솔루션을 하나의 소프트웨어 스택으로 구성해 여러 가지 워크로드를 빠르게 처리할 수 있다는 장점이 있다.

네 가지 솔루션은 (1) 영상 워크로드(디코딩, 콘텐츠 조작)를 담당하는 L4, (2) 생성형 AI 모델(text-to-image, text-to-video) 개발에 사용되는 L40, (3) 챗GPT와 같은 LLM을 위해 사용되는 H100NVL, (4) CPU와 GPU를 연결해 대용량 데이터 처리에 적합한 Grace Hopper이다. 각 워크로드에 최적화된 여러 솔루션을 하나의 소프트웨어에서 작동시킬 수 있도록 하여 엔비디아 제품에 대한 락인 효과를 제공한다.



결국 이러한 모든 변화의 핵심은 GPU의 성능개선과 사용처 확대에 방점이 찍혀 있다. 엔비디아가 만들어갈 새로운 비즈니스 모델을 통해 AI 시장에서 GPU의 중요성은 더욱 부각될 수 있을 것이다.

c. 딥러닝 개발 운영체제도 엔비디아 GPU에 최적화

딥러닝 모델을 만들 때 반드시 필요한 것은, 어떤 소프트웨어로 학습을 할 것인지에 대한 것이다. 물론 이것은 개발자 관점이기는 하지만, 투자자 관점에서도 의미가 있다. 실제 “현장”에서 많이 쓰이는 제품과 서비스로써 수혜를 입는 기업들은 앞으로도 경쟁력을 확대할 수 있기 때문이다. 그리고 그것이 결국에는 실적으로 입증될 것이다.

특히 거대한 데이터를 가지고, 언어모델 등을 만들기 위한 소프트웨어로는 PyTorch(파이토치)라는 게 가장 많이 활용되고 있다. 파이토치는 딥러닝 모델을 학습시키기 위한, 전용 “프레임워크(Framework)”다.

프레임워크는 프로그래밍을 위한 운영체제라고 생각하면 이해가 편하다. 프레임워크에서 개발자들이 코딩을 한다. 코딩을 빠르게 할 수 있도록 필요한 기능을 갖추고 있고 편리한 대신에, 개발자들은 이 프레임워크가 정한 규칙 안에서 코딩을 해야 한다.

** 프레임워크: 프레임워크는 차, 비행기, 배 같은 탈 것으로 표현할 수 있음. 사람을 코드라고 했을 때, 사람이 탈 것에 속해야만 일이 진행이 되는 형태로 이해. 탈 것은 목적에 맞게 만들어져 있고, 정해진 곳으로만 다녀야 함. 코드 제어권이 사람에게 없음. 즉, 사용자의 코드흐름을 프레임워크에 의해 맞추어야 하는 통제된 규칙. 프로그래밍에서 특정 운영체제(플랫폼)에서 구동되는 응용프로그램의 표준구조를 구현하기 위한 라이브러리의 모임. 원하는 기능 구현에만 집중하여 빠르게 개발할 수 있도록 기본적으로 필요한 기능을 갖추고 있는 것으로써, 라이브러리 개념까지 포함.*

그러니까 딥러닝 개발 프레임워크가 갖는 의미는 현장에서는 상당하다. 개발자들이 어떤 프레임워크에 익숙해지다 보면, 강한 락인효과가 발생되기 때문이다. 우리가 MS Excel를 쓰다가 갑자기 애플의 Numbers로 바꿔 써야 한다고 상상해 보면 얼마나 귀찮은 일일지 대강 감이 온다.

한편, 이 파이토치가 득세할수록 웃을 수 있는 기업이 있다. 바로 엔비디아다. 파이토치는 GPU 연산에 최적화된 소프트웨어이기 때문이다. 애초에 파이토치의 초기 사용자가 개인 개발환경 설정화면에서 컴퓨팅 유닛을 정할 때, 디폴트 값이 엔비디아의 CUDA로 되어 있다. (물론 CPU로 설정할 수도 있지만 대부분이 GPU를 택한다)

**라이브러리: 현실세계에서의 라이브러리(도서관)란 필요할 때마다 꺼내볼 수 있는 지식들이 모여 있는 곳. 프로그래밍에서의 라이브러리란 필요한 기능들이 모여 있는 코드가 모인 묶음이라고 생각할 수 있음. 다시 말해, 프로그램의 구성요소로써 공통적으로 사용될 수 있는 특정한 기능들의 모임. 사용 여부는 코드를 작성하는 사람의 선택 사항.*

즉, 프레임워크에 의존하여 개발하면서, 라이브러리로서 반복적인 코드작성을 피할 수 있음.

개발자들이 가장 많이 쓰는 딥러닝 프레임워크 “파이토치”의 초기 사용설정 화면
: Compute Platform으로, 엔비디아의 CUDA가 디폴트 설정되어 있음

START LOCALLY

Select your preferences and run the install command. Stable represents the most currently tested and supported version of PyTorch. This should be suitable for many users. Preview is available if you want the latest, not fully tested and supported, builds that are generated nightly. Please ensure that you have **met the prerequisites below (e.g., numpy)**, depending on your package manager. Anaconda is our recommended package manager since it installs all dependencies. You can also [install previous versions of PyTorch](#). Note that LibTorch is only available for C++.

PyTorch Build	Stable (2.0.0)		Preview (Nightly)	
Your OS	Linux	Mac	Windows	
Package	Conda	Pip	LibTorch	Source
Language	Python		C++ / Java	
Compute Platform	CUDA 11.7	CUDA 11.8	ROCm 5.4.2	CPU
Run this Command:	<pre>pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu117</pre>			

자료: PyTorch

여기서 CUDA는 엔비디아의 GPU에서 프로그래밍 인터페이스 및 도구 모음을 지원하는 라이브러리 개념이다. 따라서, 파이토치를 사용하는 개발자들은 엔비디아의 CUDA를 자연스럽게 사용하게 되고, 엔비디아의 GPU를 쓰게 되고, 또 그것을 작동하기 위한 엔비디아의 전용 그래픽 드라이버까지 사용해야 한다. 파이토치 사용 하나로, 엔비디아의 생태계에 완전히 들어오게 된다는 말이다.

엔비디아는 딥러닝 영역에서 여기저기 안 끼는 데가 없다. 그만큼 엔비디아가 AI 생태계에 있어 막강한 입지를 구축한 것이기도 하다.

엔비디아가 CUDA 생태계로 얻는 AI 개발 산업 내 선순환 과정
: CUDA 입지 상승 → 개발자 증가 → 데이터센터 탑재 증가 → GPU 판매 확대 (반복)

엔비디아가 AI 시대를 꿈꾸며 구성한 '선순환 모델'

350만명 개발자

엔비디아 호환 소프트웨어 프로그램 사용률 가속화

엔비디아 AI 연산용 칩 (GPGPU 등)

PC제조사 및 클라우드 사업자

3.5만 곳의 각종 단체들 (기업 등)

전세계적으로 100개가 넘는 하이퍼스케일러 시스템이 엔비디아가 사용

인공지능 서비스 확대 GPU 수요 증가

CUDA 모델 (엔비디아 거) 생태계 확대

엔비디아가 가진 거대한 'AI 생태계'

CUDA 다운로드 - 4년 만에 4.5배

37M

개발자 수 - 4년 만에 3.5배

3.5M

가속화된 애플리케이션 - 4년 만에 5배

3K 3K

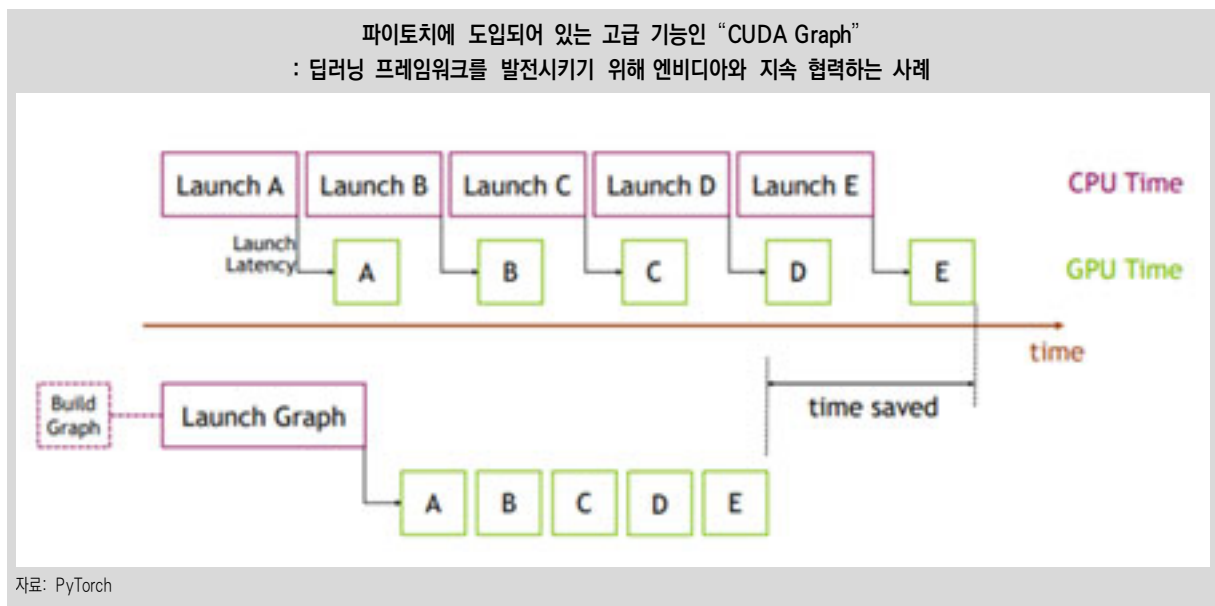
자료: Nvidia, 미래에셋증권 디지털리서치팀

만약 개발자들이나 연구원들이 파이토치와 CUDA가 아닌 프레임워크와 라이브러리를 사용하게 된다면, 굉장히 까다로운 작업들이 수반된다. 먼저 기존에 쓰던 코드가 아니라 새롭게 쓸 프레임워크에 맞춰서 코드 수정도 해야 하고, 라이브러리를 불러오는 방법까지 다시 공부해야 할 것이다. 뿐만 아니라, 인터페이스 자체가 다르기 때문에 상당한 시간적 노력이 수반된다.

위와 같은 불편함 들을 무릅쓰면서까지 파이토치 및 CUDA 생태계 바깥에 있을 이유가 개발자들에게 없는 것이다. 따라서 현재 대세인 딥러닝 프레임워크 파이토치가 계속해서 주류로 자리하게 되면 엔비디아의 입지는 계속해서 올라갈 수밖에 없다. 최근에도 파이토치 측에서는 프레임워크를 지속적으로 발전시키기 위해 엔비디아와 지속적으로 협력하고 있다.

2021년 4분기에 발표된 파이토치의 “CUDA Graph”라는 기능은 좋은 예시다. 핵심은 CPU 연산이 아니라 GPU 연산에 좀 더 집중해서 모델 연산 시간을 줄여 속도를 올리는 것으로 이해하면 된다. 이러한 협력 사례는 앞으로도 계속 증가할 가능성이 높다. 이는 파이토치만 건재하다면 CUDA가 딥러닝 모델 개발에 계속 주력으로 쓰일 것이라는 말이다.

**CPU 연산 한 번당 GPU 연산이 이뤄지는 기존 구조에서는, CPU-GPU 병목현상 때문에 훈련 시퀀스가 늘어날 때마다 시간이 계속 늘어났음.*

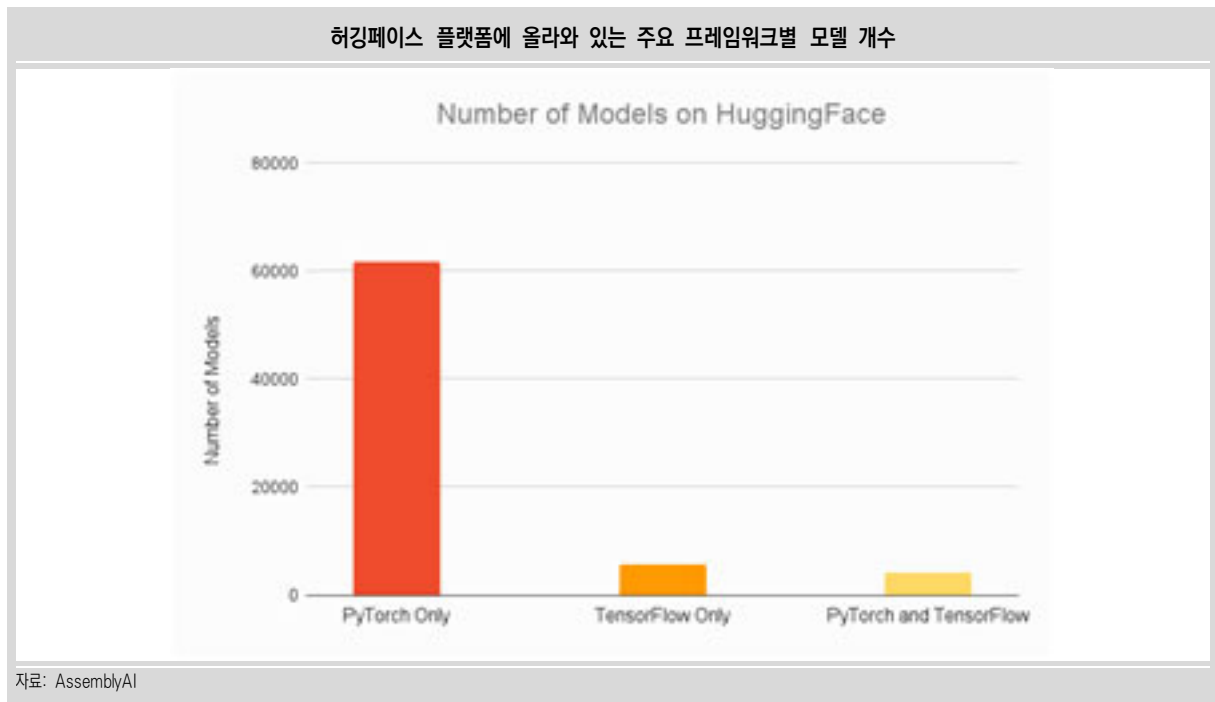


그렇다면 여기서 짚어봐야 할 것은 파이토치의 인기가 계속될 것이냐는 점이다. 결론부터 말하면 계속 더 상승하고 있다. 사실 5년 전만 하더라도, 구글이 개발한 TensorFlow(구글이 개발한 머신러닝 프레임워크)가 대세였다. 그런데 최근 상황이 완전히 역전되었다. 2020년대 들어서는 파이토치가 간편성과 높은 호환성을 무기로 완전히 주류로 자리 잡았다. 개발자들과 연구원 사이에서도 파이토치의 사용 비중이 점차 높아지고 있다.

개발자들에게는 익숙한 허깅페이스(HuggingFace)라는 곳이 있다. 허깅페이스는 트랜스포머 모델의 유행을 선도한 미국의 스타트업이다. 트랜스포머 모듈을 제공하는 곳으로 정평이 났을 뿐만 아니라, 최근에는 글로벌 최대 AI 개발 라이브러리를 제공하는 플랫폼으로 거듭난 곳이다. 이곳에서 많은 딥러닝 모델 개발 라이브러리가 공유/전파되고 있다.

특히, 딥러닝 모델은 점점 복잡해지고 있기 때문에, 작은 기업이나 연구자들은 자신들만의 컴퓨팅 자원으로 이를 탐구하고 활용하기 어려워졌다. 따라서 이미 사전훈련을 마친 모델에 접근해서 전이학습과 파인 튜닝만 해서 즉시 추론과 같은 작업으로 넘어가려고 한다. 이 때 허깅페이스에 접속하게 되는 것이다.

허깅페이스에 올라와 있는 딥러닝 모델 숫자(모델 가용성)을 프레임워크별로 나눠보면 파이토치 쪽이 압도적이라는 것을 알 수 있다. 전체 딥러닝 모델 중 약 92%가 파이토치 전용 모델이며, 이는 2022년의 85%보다도 더 높아졌다. 반면 텐서플로우 전용 모델은 약 8%에 불과하고, 파이토치와 텐서플로우를 함께 쓰는 모델의 숫자도 14% 정도에 그쳤다. 구글의 검색엔진 점유율보다도 더 높은 독점 수준을 보이는 곳이 바로 이 프레임워크 시장이다.



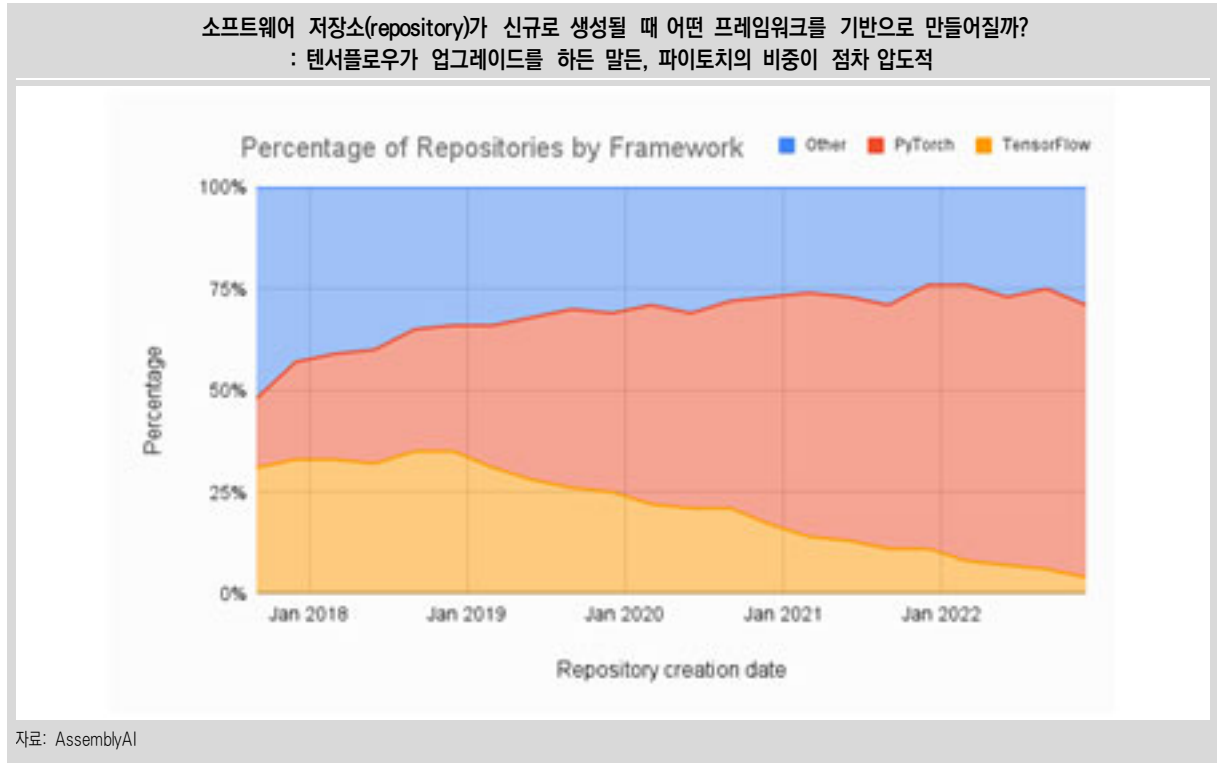
뿐만 아니라, 허깅페이스에서는 코딩을 할 때 많은 도움을 주는 라이브러리를 제공한다. 이를 ‘소프트웨어 저장소(repositories)’라고 부르기도 하는데, 일종의 코드 저장소다. 저장소를 통해 논문, 코드, 데이터 세트 등을 자유롭게 이용할 수 있는 리소스 등을 저장하고 관리하고 또 다운로드가 가능하다. 당연히 많은 개발자들이 찾는 필수 기능이다.

**PyTorch와 TensorFlow 모두 공식 저장소(software repositories)를 보유하고는 있지만, 주로 허깅페이스에서 사용하고자 하는 경우가 많음.*

그런데, 2017년부터 현재까지 분기별로 집계된 자료에 따르면, 파이토치를 기반으로 만들어진 저장소의 숫자가 압도적이라는 것을 알 수 있다. 이 말은 파이토치를 이용하는 딥러닝 모델 논문이 꾸준히 증가했다는 말이기도 하다. 게다가 ChatGPT를 만든 OpenAI조차 2020년에 내부적으로는 파이토치 사용을 표준으로 정했다고 한다.

**물론 OpenAI는 강화학습 분야에 대해서는 텐서플로우로 구현한 저장소를 제공하고 있음*

이번 분기에 생성된 3,319개의 저장소 중에서 거의 70%가 파이토치로 구현되었고, 텐서플로우는 4%에 불과했다(작년의 11%에서 감소). 텐서플로우는 2019년에 차세대 버전까지 출시하였지만, 이러한 추세를 바꿀 수는 없었다.



그리하여, “딥러닝 개발에는 프레임워크가 필수적이다. 프레임워크의 절대 강자는 파이토치다. 파이토치는 GPU를 기반으로 구동되고, CUDA 라이브러리에 최적화되어 있다. GPU의 최강자는 엔비디아고 CUDA의 생태계는 더욱 공고해진다.”라는 것으로 정리할 수 있다.

3) 새로운 인공지능 반도체로 주목받는 NPU

a. CPU, GPU만으로는 부족하다.

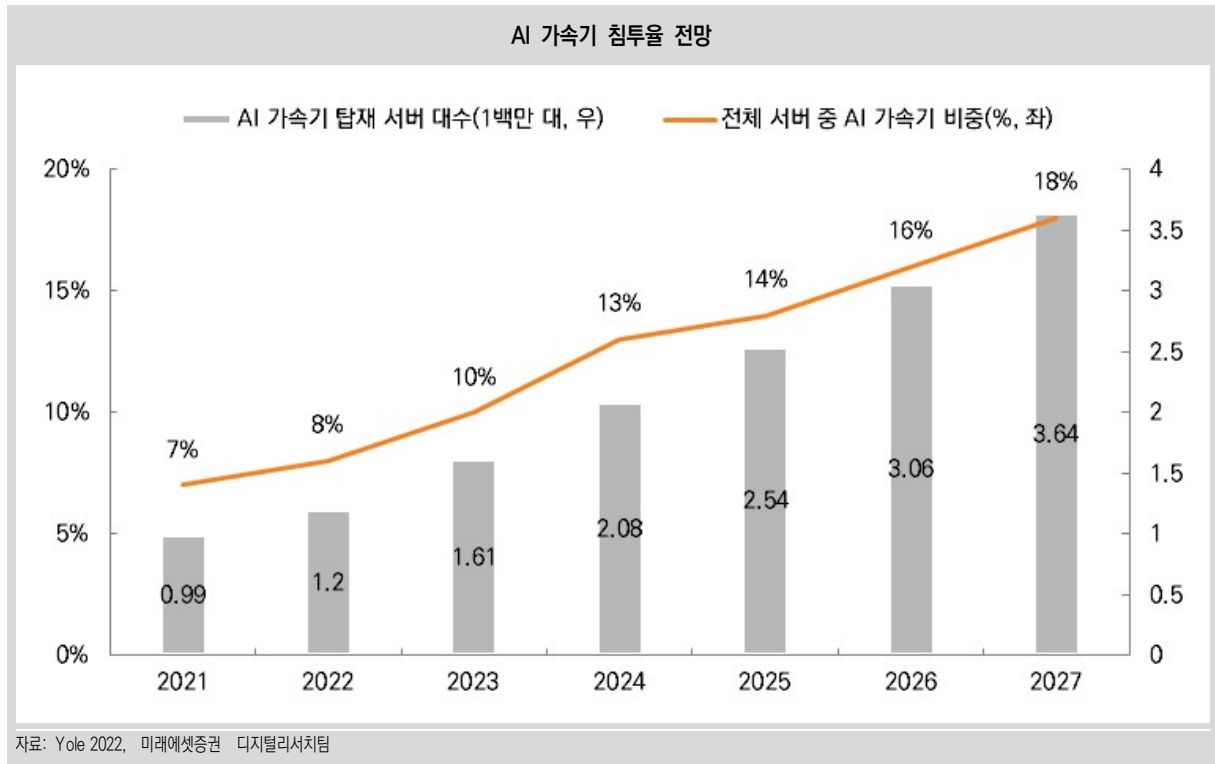
인공지능 기술 개발을 위해서는 엄청난 양의 반도체가 필요하다. 인공지능은 텐서 연산에 기초하는데 그동안 텐서 연산에 최적화된 반도체가 없었기 때문에 범용제품인 CPU와 GPU가 주로 인공지능 연산에 쓰였다.

그래서 구글의 초기 알파고 모델도 CPU와 GPU를 사용했었다. 하지만, 구글은 이후 자신들이 직접 ASICs 개념으로 만든 NPU를 만들어 알파고에 장착했으며 이를 Tensor Processing Unit, TPU라고 명명했다.

아직까지 인공지능 연산에 있어 기본적인 프로세싱 유닛으로는 GPU가 가장 많이 활용되고 있다. ChatGPT도 엔비디아의 A100 GPU 1만 개로 학습되었다고 언급했다. 하지만 문제는 GPU가 너무 비싸다는 점이다. 특히 매년 엔비디아가 내놓는 신제품들의 가격은 계속해서 상승하고 있다. 물론 성능이 개선된 신제품의 가격이 상승하는 것은 일반적이긴 하지만 인공지능 연산에 필요한 GPU를 엔비디아가 거의 독식하고 있기 때문에 발생하는 현상이라고 볼 수도 있다.

더군다나 GPU는 전력 대비 성능 즉 전성비 측면에서 약점을 가지고 있는데 증가하는 전력비는 대규모 인공지능 모델을 처리해야 하는 데이터센터 운영자 입장에서는 부담스러운 비용이다. 굳이 ESG라는 잣대를 들이대지 않아도, 경제적 관점으로도 전력 소모를 줄여야만 인공지능 기술이 더욱 발전할 수 있다.

그리고 전력 소모가 낮을 경우 칩의 성능을 추가적으로 개선할 업사이드도 생길 수 있다. 그렇기 때문에 최근 낮은 가격, 낮은 전력 소모를 기반으로 하는 인공지능 연산용 전문칩인 NPU(Neural Processing Unit)가 주목 받고 있다. 시장분석 전문업체 YOLE에 따르면, 2021년부터 2027년까지의 AI 가속용 칩의 침투율은 연평균성장률(CAGR) 24%를 기록할 것으로 전망된다.



b. 주문제작형태의 NPU는 AI 연산의 한 축

NPU는 인공지능이 수행할 무수히 많고 복잡한 연산을 가장 빠르고 효율적으로 처리해 주는 전문화된 칩이라고 보면 된다. 다만, NPU는 인공지능에 대해서는 선수지만 다른 알고리즘 계산은 잘 처리하지 못한다는 단점이 있다. 그야말로 스페셜리스트인 것이다. 인공지능이 원하는 특정 함수(딥러닝에서의 활성화 함수)를 잘 처리하는 것이라 생각하면 된다. 그래서, 각 영역마다 필요로 하는 NPU는 목적에 따라 각각 다른 형태를 띠 것으로 보인다. 바둑용 인공지능이 알파고였듯이, 인공지능도 각 목적에 따라 최적화된 인공지능 모델 알고리즘이 나올 것이고 그에 최적화된 반도체가 출현할 가능성이 높다고 생각한다.

물론 여기에는 반론도 존재한다. 너무 많은 업체들이 각자의 모델 알고리즘을 내놓아 거기에 맞는 반도체를 설계하려고 한다고 생각해 보자. 하지만 칩을 만들어주는 파운드리 생산능력은 제한되어 있으며 AI 반도체는 성능이 중요하기 때문에 미세공정을 사용해야 할 것이다. 미세공정을 사용할 경우 단가가 높아질 수 있고, 이런 상황에서 고성능 칩의 수요 증가로 파운드리의 교섭력이 더 올라가면 칩을 설계한 입장에서는 비용이 증가하고 원하는 수량만큼의 반도체를 공급받기도 어려워질 수 있다. 그리고 파운드리 입장에서 단일 품목의 칩을 많이 찍어내는 것이 수율 관리에도 유리하다는 점을 고려할 필요가 있다.

그리고 NPU는 여러 산업들에 대해서 맞춤형 칩으로 제작될 것으로 예상돼 고객 맞춤형 칩 그러니까 ASICs 형태가 될 가능성이 높다. 어떠한 인공지능 모델 알고리즘을 만든 업체가 칩에 대한 설계를 NPU 업체에 의뢰하면 고객이 요구하는 연산을 최적화할 수 있는 반도체 설계를 만들어 주는 형태이다. 그래서 ASICs 방식으로 최적화된 NPU는 GPU같이 범용 프로세서에 비해, 전체유효시장(TAM)은 작을 수밖에 없다는 점에서 파운드리 입장에서는 아무래도 후순위로 여길 수도 있다.

또한, GPU 업계가 만든, 그 중에서도 엔비디아가 구축한 GPU 기반 소프트웨어 CUDA 라이브러리에 인공지능 모델 알고리즘 개발자들이 익숙해진 것도 GPU의 생존력을 높이는 요인이다. NPU는 하드웨어의 조합부터 소프트웨어의 코딩 영역까지, 양쪽 부문에서 기존 반도체와는 전혀 다른 방식으로 설계가 이뤄지기 때문에 개발 단계의 난이도도 높다.

따라서 NPU가 GPU를 전부 대체하기는 쉽지 않아 보인다. 오히려 특정 영역에서는 강점을 보일 것으로 보는 것이 합리적이라는 판단이다. 그렇기 때문에 꽤 오랜 기간 GPU와 NPU는 서로 공존하면서 인공지능 연산에 쓰이지 않을까 생각된다. NPU의 부상으로 GPU가 몰락하는 그림은 아닐 것이다.

NPU 미래를 그려보자면, 각각의 니치 마켓(틈새시장)에서 최적화된 형태의 솔루션을 제공하는 업체들이 출현할 것으로 전망한다. 특히 기대가 되는 영역은 금융과 바이오 헬스케어, 그리고 에너지 산업 정도가 있을 것으로 보인다. 이 산업들에 투입되는 자금의 볼륨이 크고 혁신으로 인해 생산성이 향상될 여지가 큰 영역이기 때문이다.

예를 들어, 트레이딩 영역에서는 고주파 트레이딩(HFT)이라고 불리는 ‘초단타매매’를 위한 알고리즘에 최적화된 반도체가 있다. 사람을 대신해 정해진 알고리즘대로 거래소에 주문을 넣어주는 역할을 하는 단순한 NPU다. HFT 영역도 과거에는 엔비디아 GPU를 사용해왔지만, 점차 NPU 형태로 변화할 수 있을 것이라 기대된다. 국내 NPU 설계 업체인 리벨리온 박성현 CEO의 말에 따르면, 본인들이 설계한 NPU가 엔비디아 GPU보다 10배의 속도이지만 전력은 절반 수준이라고 언급하기도 했다. 따라서, 업체들의 니즈와 산업의 성격에 따라 잘 쓰이는 칩이 서로 다른 시대가 올 것으로 전망된다.

리벨리온이 만든 HFT 알고리즘 연산에 특화된 NPU “아이온” - 10배 성능, 전력소비는 절반
: JP모간으로부터 기존 제품이던 엔비디아 GPU를 밀어내고 탑재



자료: rebellions

c. NPU 업계에서 “가장 비싼 몸은?”

NPU 영역에서 가장 두각을 나타내고 있는 기업들은 주로 영미권에 위치한 기업들이 많다. 글로벌 반도체 밸류체인을 생각했을 때, 반도체 설계 분야에 있어서 미국이 선두권이기 때문인 것으로 보인다. 한국과 대만은 제조와 패키징에 강점을 보이지만 유연성과 창의성을 요하는 설계에서는 미국의 영향력이 여전히 막강하기 때문이다.

AI 연산 전문칩 관련 주요 업체 리스트

가장 최근 자금조달 시기	업체명	투자라운드	누적자금조달 규모(백만달러)	국적
2021-04-13	SambaNova Systems	Series D	1,132	미국
2021-11-10	Cerebras Systems	Series F	718	미국
2022-09-22	Graphcore	인큐베이팅	682	영국
2021-04-14	Groq	Series C - II	363	미국
2021-05-05	TensTorrent	Series B	221	캐나다

자료: CBInsights, 미래에셋증권 디지털리서치팀

물론 올해 아직까지 TOP 5 NPU 업체들이 자금 조달을 받은 곳은 없다. 다만, SVB의 파산과 그 후폭풍이 진정되고 VC 시장이 다시 고개를 든다면, 생성형 AI의 대유행과 더불어 앞으로 이들의 몸값이 올라갈 가능성이 크다고 생각한다. 이들 중 현재까지 가장 많은 자금을 투자 받은 곳은 미국 캘리포니아에 소재한 삼바노바 시스템즈라는 업체다. 주요 투자자로는 전략적 투자자로 보이는 구글과 인텔이 있고, 재무적 투자자들로도 블랙록, 테마섹, GIC, 그리고 소프트뱅크가 포진해 있다.

삼바노바 시스템즈의 주요 투자자 리스트

최초 투자	마지막 투자	투자자	국적	투자라운드
2018-03-15	2021-04-13	Walden International	미국	Series A (2018);Series B (2019);Series C (2020);Series D (2021)
2018-03-15	2021-04-13	구글 벤처	미국	Series A (2018);Series B (2019);Series C (2020);Series D (2021)
2019-04-01	2021-04-13	인텔 캐피탈	미국	Series B (2019);Series C (2020);Series D (2021)
2020-02-25	2021-04-13	블랙록	영국	Series C (2020);Series D (2021)
2021-04-13	2021-04-13	테마섹	싱가포르	Series D (2021)
2021-04-13	2021-04-13	싱가포르 투자청	싱가포르	Series D (2021)
2021-04-13	2021-04-13	소프트뱅크	일본	Series D (2021)
2020-02-25	2020-02-25	Celesta Capital	미국	Series C (2020)
2018-03-15	2020-02-25	Redline Capital Management	룩셈부르크	Series A (2018);Series B (2019);Series C (2020)
2018-03-15	2019-04-01	Atlantic Bridge Capital	영국	Series A (2018);Series B (2019)

자료: CBInsights, 미래에셋증권 디지털리서치팀

삼바노바 시스템즈는 작년 9월 중순 “데이터스케일 SN30(DataScale SN30)”이라는 새로운 하드웨어-소프트웨어 통합형 AI 시스템을 출시하기도 했다. 데이터스케일 SN30은 대규모 인공지능 모델을 지원하도록 업그레이드한 차세대 시스템이다. 특히, 삼바노바에서 SN30 시스템이 GPT 모델을 빠른 속도로 훈련할 수 있다고 주장한 점이 인상적이었다.

삼바노바에 따르면 데이터스케일 SN30은 텐서를 기반으로 구축된 엔비디아 DGX A100(GPU) 시스템에 비해 특정 인공지능 워크로드에서 6배 속도 향상이 가능하고 12.8배 더 많은 메모리 용량을 제공한다고 한다. 또한, SN30 시스템은 software-defined 접근 방식을 택했는데, 이러한 ‘재구성 가능한’ 아키텍처 덕분에, 본인들의 특정 소프트웨어를 실행하면 각 인공지능 모델이 시스템 레벨에서 최적으로 실행될 수 있다고 한다. 다시 말해, 각각의 사용자들의 원하는 워크로드 요구 사항을 충족하도록 각각의 프로세서들을 재구성함으로써 신경망에 고효율 성능을 제공할 수 있는 맞춤형 방식이라는 것이다.

그리고 SN30 시스템에 탑재될 주요한 프로세서로서 “카디널(Cardinal) SN30 RDU(Reconfigurable Dataflow Unit)”도 공개했는데, 이 칩은 TSMC의 7nm 공정에서 제조되고 860억개의 트랜지스터를 포함하며 FP16 정밀도에서 688 TFLOP가 가능하다고 한다. 엔비디아의 GPU인 A100의 FP16 정밀도가 약 600 TFLOP인 것에 비해 성능 더 뛰어나다. 신생 스타트업이 만든 칩의 성능이 이 정도인 것을 보면, 인공지능에 특화된 함수 연산에서 NPU가 가지는 edge가 얼마나 강력한지 알 수 있는 대목이기도 하다.

특히 카디널 SN30 RDU는 큰 매개변수를 필요로 하는 GPT와 같은 대규모 모델에 적합하게 될 것이다. 또한, 최고 해상도를 요구하는 컴퓨터 비전 모델 같이 엄청난 양의 메모리를 제공하는 시스템 아키텍처 등에도 적용될 수 있을 것으로 보인다. 앞으로 삼바노바와 같은 NPU 업체들도 계속해서 인공지능 반도체 업계에서 부각되리라 생각된다.



4) 3세대 AI반도체 뉴로모픽과 메모리

a. 뇌를 닮은 반도체 “뉴로모픽”

인공지능 시스템이 진화하면서 기존 폰 노이만 방식 반도체에 한계에 대한 이야기들이 많다. 폰 노이만 방식에서 부각된 문제점은 바로 데이터의 병목현상이다. 중앙처리장치(CPU), 주기억장치, 입출력장치 3 단계로 구성된 컴퓨터 구조상, 연산을 담당하는 코어의 속도보다 스토리지와 입출력 단계에서의 데이터 버스 처리 속도가 느리기 때문이다.

즉, 폰 노이만 구조의 직렬처리 방식은 구조적인 문제를 가지고 있으며 이러한 구조적인 문제로 인해 전력 소비량 폭증 등 효율이 떨어지는 부작용이 생긴다.

게다가 CPU의 성능개선을 이끌던 무어의 법칙이 한계에 달하고 심화 연산인 딥러닝을 통해 추론 과정과 의사결정 과정이 더해지면서 폰 노이만 방식의 반도체 기술의 변화가 필수적이라는 인식이 생겨났다.

차세대 컴퓨팅에 있어서 폰 노이만 구조가 겪는 한계와 잠재 대안들



자료: Yole

이러한 과정에서 사람의 뇌에서 착안한 뉴로모픽 반도체가 새로운 대안으로 떠오르고 있다. 뉴로모픽 반도체는 인간의 신경망 구조와 같이 모든 칩을 병렬로 연결해 연산과 저장을 한 번에 할 수 있는 기술을 말한다. 연산과 저장이 가능한 In-Memory Computing과 유사한 개념이지만 뇌 신경망 구조를 활용했다는 점에서 In-Memory Computing과 비교해 진화한 기술이다. 뉴로모픽 반도체 내 소자는 인간 뇌의 뉴런(연산), 메모리는 시냅스(기억) 역할을 담당한다.

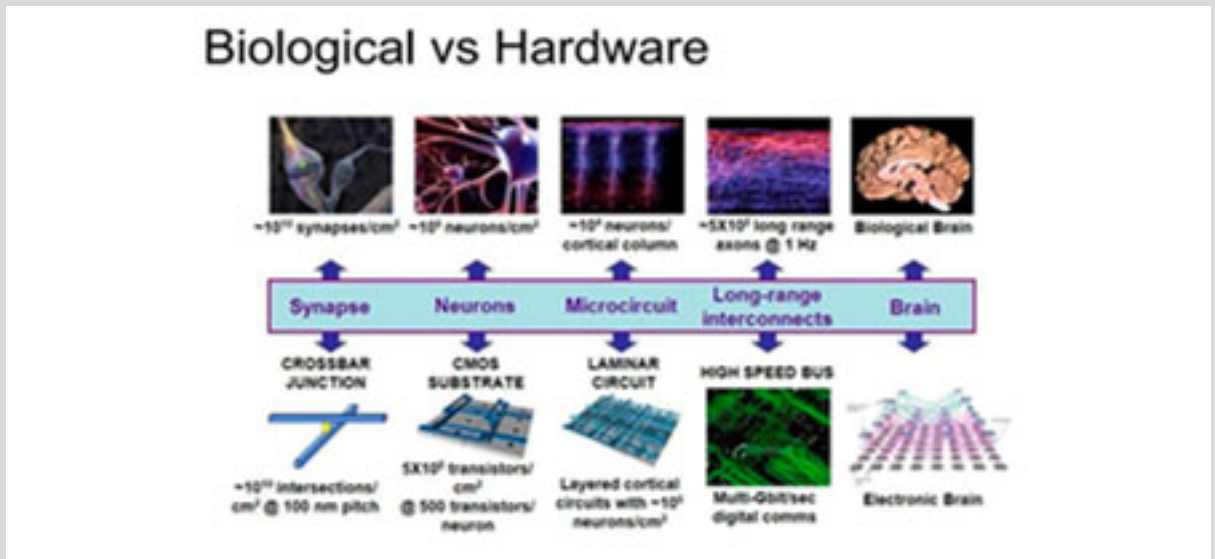
폰 노이만 구조의 전통적인 반도체와 뉴로모픽 반도체의 구조 차이



자료: 융합연구정책센터, 한국지능정보사회진흥원, 미래에셋증권 디지털리서치팀

인공 뉴런 역할을 하는 코어를 사람의 뇌처럼 병렬로 구성하기 때문에 기존 칩 구조보다 훨씬 적은 전력 만으로도 더 많은 양의 데이터를 처리한다는 장점이 있으며 인간의 두뇌가 말이 안될 정도로 낮은 Power(전력)로 구동된다는 사실을 착안하면 그리 무리한 주장도 아니다.

인간의 생체 두뇌 모습과 반도체 구조의 상호비교



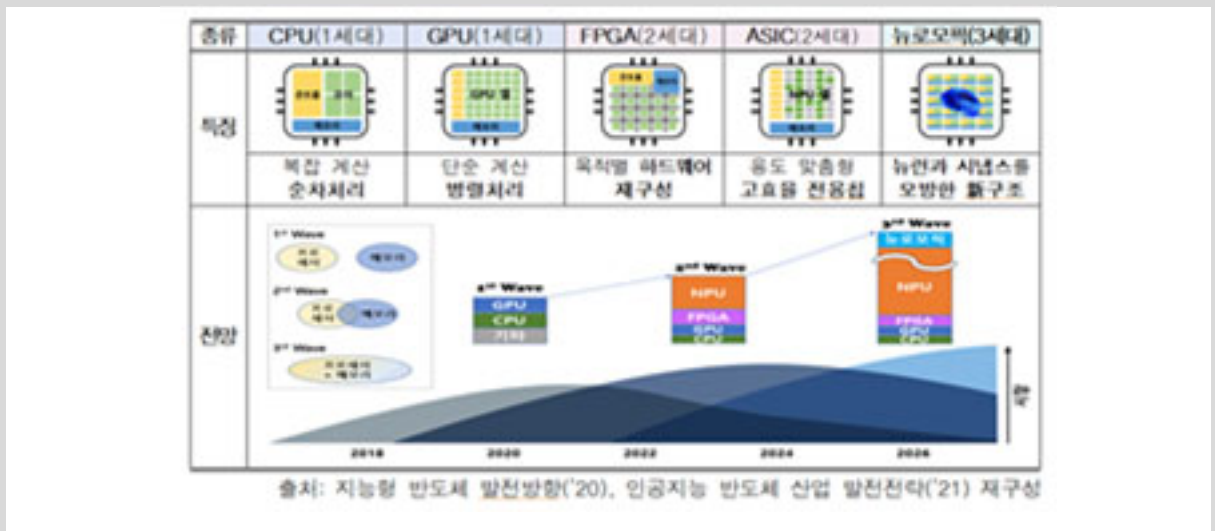
자료: DARPA

최근 폭발적으로 늘어나는 데이터를 처리하기 위해 병렬 연산의 중요성이 부각되며 GPU가 반도체 업계의 새로운 총아로 떠올랐듯이 AI 서비스가 확대될수록 뉴로모픽 반도체의 필요성은 더욱 부각될 것이다. 뉴로모픽 반도체의 지향점은 병렬 연산의 최적화를 구현하려는 목적을 가지고 있기 때문이다. 뉴로모픽 반도체의 발전 현황을 파악하기 위해선 In-Memory Computing의 변화를 주목해보자.

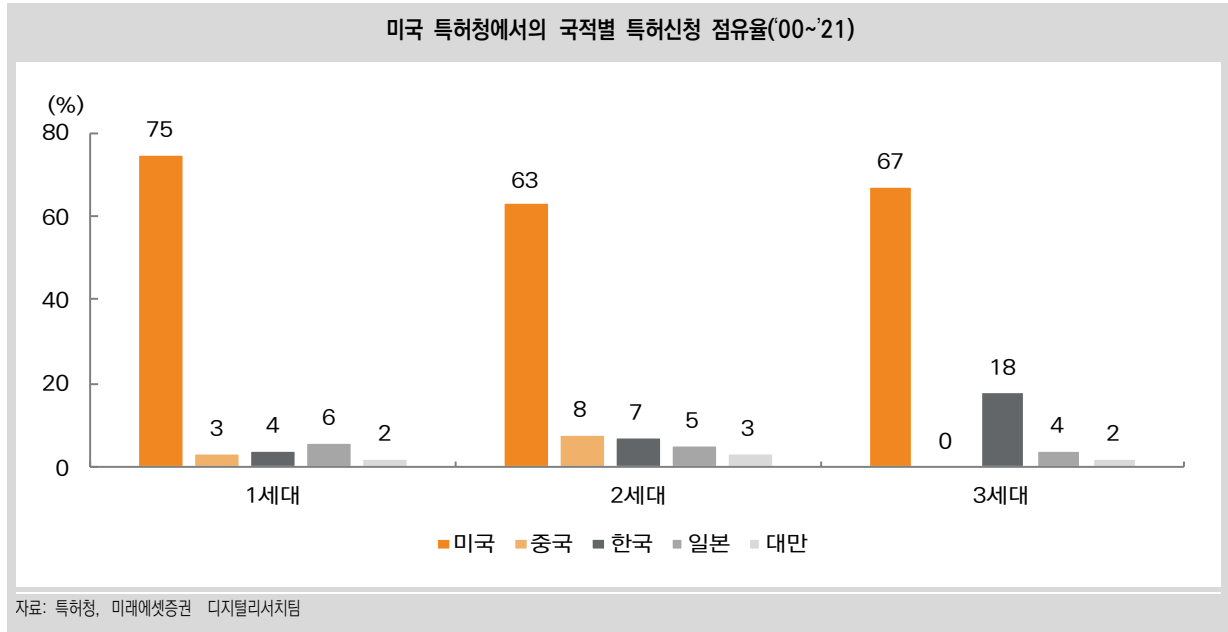
b. 뉴로모픽 업계 현황

글로벌 메모리반도체 리딩업체인 한국의 삼성전자는 이미 몇 년 전부터 물밑에서 뉴로모픽 반도체를 개발해오고 있다. 삼성전자는 하드대학교 연구진과 함께 뇌를 닮은 차세대 뉴로모픽 반도체에 대한 비전을 제시하기도 했다. 특히 뉴로모픽 칩 관련 특허 신청 부분에서 한국이 미국에 이어 2위를 차지한 것은 반가운 일이기도 하다.

기존 반도체와 뉴로모픽 반도체의 구조 비교



자료: 특허청, 경제추적연구소



또한 현재 뉴로모픽 칩 분야에서 가시적인 성과를 보이고 있는 회사로는 인텔과 퀄컴, 그리고 IBM 등이 있다. 인텔은 인공지능에 특화된 연산이 가능한 ‘로이히(Loihi)’ 칩을 공개하기도 했고, 퀄컴은 제로스(Zeroth) 프로세서 개발을 발표했다. 또한 IBM은 딥러닝을 가능하게 만드는 트루노스(True North) 칩으로서 주목을 받았다. 특히, 인텔과 코넬대학교가 공동으로 연구한 뉴로모픽 칩은, 동물의 후각 체계를 수학적으로 표현한 알고리즘을 칩에 구현했다고 한다. 수치 알고리즘으로 감각기관의 데이터를 표현했다는 말이다. 이것은 자연스레 생성형 AI로 접목이 가능하다는 것을 의미한다. 시각 데이터를 수치화해서 임베딩 벡터로 만들어 인공지능 연산을 했던 것처럼, 음성 주파수 데이터를 임베딩 벡터로 만들어 딥러닝으로 학습했던 것처럼 후각도 이제 가능해질 수 있다는 것이다. 거기에 더 나아가서 2020년 7월에는 싱가포르 국립대학교 연구진이 로봇에 ‘촉각’ 기능까지 부여했다고 발표하기도 했다.

이렇게 되면 시각, 청각, 후각, 촉각까지 사람의 거의 모든 감각 체계를 전부 다 데이터, 수치화 할 수 있는 작업이 가능한 시대가 머지 않았다는 말이다. 이런 감각에 관한 모델 알고리즘을 인공지능 에이전트에게 적용한다면 기계가 비로소 우리처럼 감각을 느끼게 되는 것이다! 지금도 이미지 식별과 분류, 음성 인식 정도는 수준급으로 처리할 수 있는데, 머지않아 기계가 냄새를 구별하고, 본인의 손으로부터 촉각도 감지하게 될 것이다.

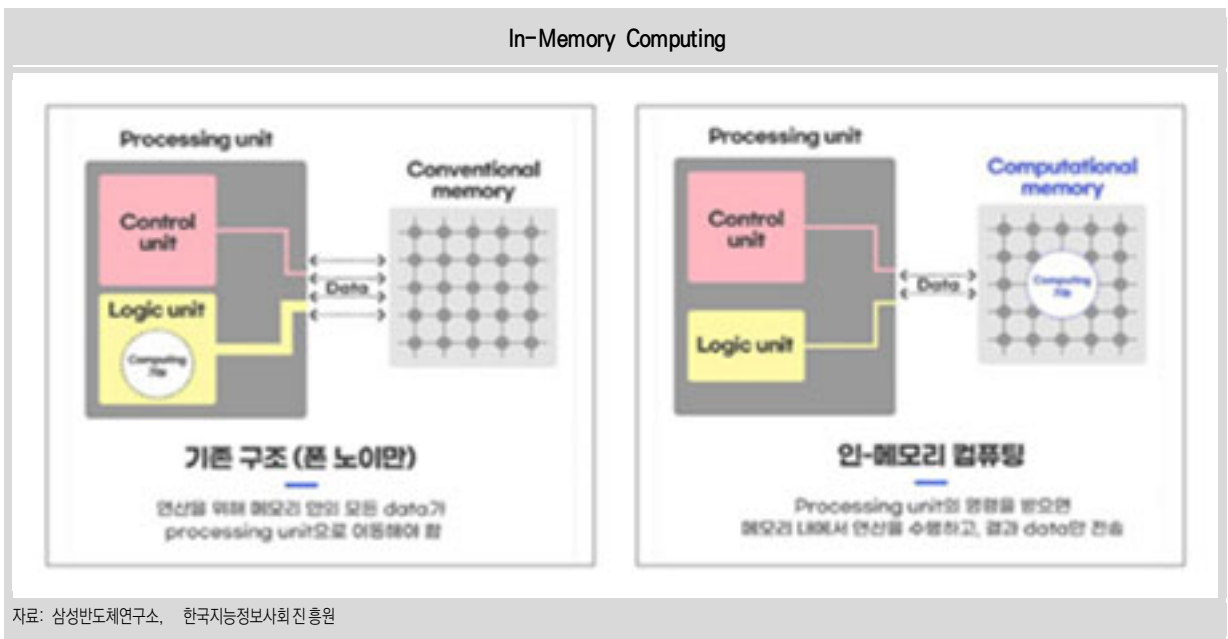
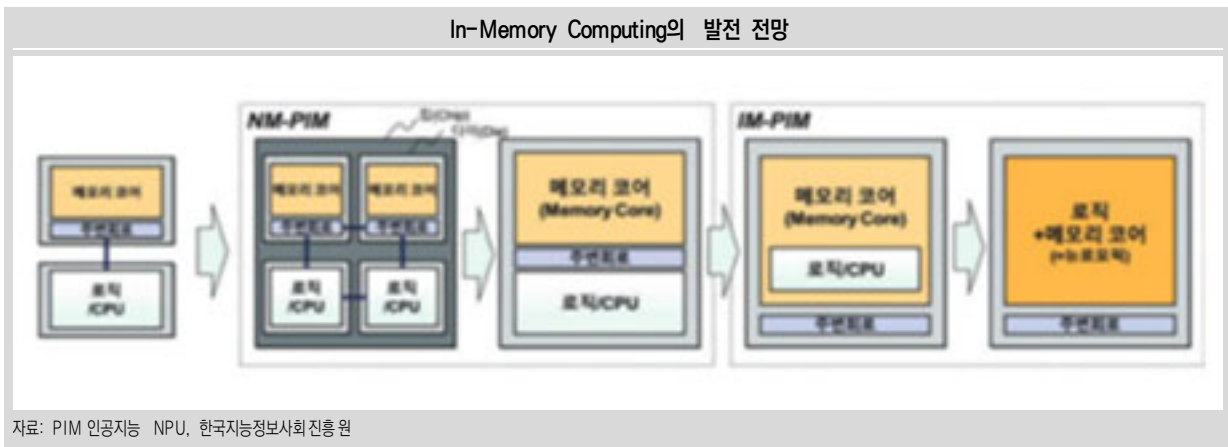
‘감각 구별’에 끝내지 말고 생성형 AI의 관점으로 다시 생각해보면 더 흥미롭다. 우리가 이전에 학습한 잠재공간 편집(latent space editing)이라는 것으로 상상력을 조금 더 동원해보자. 잠재공간을 통해서 딥러닝은 각각의 냄새가 보유하고 있는 벡터값을 찾고, 그것을 또 편집하면서 새로운 냄새를 생성해 낼 것이다. 우리가 한 번도 맡아보지 않은 향이지만 엄청 좋은 그런 냄새가 만들어질 수도 있다. 영화 “향수”에서 주인공이자, 향수 제조에 있어 인류의 능력을 아득히 벗어났던 “장 바티스트 그루누이”가 현실에서 탄생할 수 있다. 단, 인공지능 에이전트의 모습으로서 말이다.

그런 기능을 지원하는 칩 중 일부는 뉴로모픽 칩이 될 가능성이 매우 크다. 그게 우리의 뇌와 가장 닮아 있기 때문이다. 여기서 일부라고 표현한 이유에는 NPU 때와 같이 범용적인 측면에서 GPU의 이점이 분명히 있기 때문이다.

c. 뉴로모픽으로 가는 징검다리, “인메모리 컴퓨팅”

하지만 아직까지 뉴로모픽 반도체가 본격적으로 상황화 되기 위해선 많은 연구가 필요한 실정이다. 이에 우리는 뉴로모픽 반도체로 가기전에 중간 단계에서 폰 노이만 구조의 단점을 보완해줄 수 있는 데이터 병목 현상을 개선해줄 수 있는 HBM과 같은 고대역폭 메모리와 PNM(Process Near memory) PIM(Process in memory)와 같은 In-Memory Computing 기술에 주목하고 있다.

현재 AI 연산과 추론 문제에서 전성비를 악화시키는 가장 주요한 부분으로 병목 현상이 지목받고 있다는 점에서 이를 효율화 할 수 있는 메모리 반도체 발전은 필수적이며 AI 시대에서 메모리 반도체의 위상은 높아질 수 있을 것으로 전망한다.



결국 AI 시대를 맞아 폭증하는 데이터와 연산 및 추론을 수행하기 위한 GPU, NPU, 뉴로모픽과 같은 반도체가 지향하는 발전 방향은 동일하다. 결국에는 효율을 높여야만 하는 목적을 가지고 있으며 단 하나의 기술만으로는 이 같은 변화를 모두 대응하기는 어려울 것이다. 결론적으로 앞으로 상당 기간 동안 1세대~3세대 반도체들은 서로의 영역을 존중하면서 공존 및 발전을 하게 될 것으로 보인다.

5) 병렬 연산에 필수인 네트워킹

a. 네트워킹의 중요성과 이더넷

앞서 언급한 것처럼 AI 시대를 맞아 하드웨어 반도체의 중요성에 대해서 언급했다. 다음 타자는 이러한 것들을 ‘연결’해 줄 수 있는 네트워킹 차례다.

인공지능 연산은 데이터센터 서버 단에서 이뤄진다. 그리고 데이터센터에 구축되어 있는 수많은 서버들은 함께 통신이 가능한 채로 연결이 되어 있다. 연결을 위한 장치들을 네트워킹이라고 지칭하고, 대표적인 연결 방식으로는 이더넷(Ethernet)과 인피니밴드(InfiniBand, IB)가 있다.

이더넷과 인피니밴드 모두 컴퓨터 네트워킹 기술로서 데이터 전송을 위한 프로토콜이나 통신 규격을 의미한다. 이더넷은 주로 LAN에 사용되는 유선 네트워크 기술이고, 인피니밴드는 고성능 컴퓨팅 분야에 사용된다. 특히 인피니밴드는 대용량 데이터 전송을 위해 설계돼 데이터센터용으로 주로 활용되어 왔다.

하지만, 인공지능 딥러닝 학습에 있어서는 이더넷이 인피니밴드보다 우위를 보이고 있다. 최근 데이터에 따르면, 이더넷이 인피니밴드보다 더 높은 메시지 크기에서 더 나은 성능을 보이는 데다, 기본적으로 인피니밴드보다 이더넷의 대역폭 처리량이 더 크기 때문이다. 그러니까 인공지능 모델이 점차 커지면서 처리해야 할 연산량은 기하급수적으로 늘어나게 되는데, 연산 장치들간의 네트워킹에 있어서 주요 규격으로 이더넷이 더 유용하다는 말이다.

** 인공지능을 위한 학습 워크로드는 일반적으로 높은 메시지 크기(message size)를 가짐. 메시지 크기가 클수록 데이터 전송시간이 길어지고, 대기시간이 증가하기 때문. 이에 따라 더 많은 연산 리소스가 필요하며 잠재적으로 네트워크에 더 많은 정체가 발생한 확률이 큼.*

그리고 AI를 훈련시키기 위해서는 낮은 레이턴시가 매우 중요하다. 딥러닝의 기본은 최적의 파라미터값을 찾는 행위라고 했다. 그리고 최근에 대세가 되고 있는 기반 모델들 모두가 엄청나게 큰 데이터 세트와 파라미터 숫자를 보유하고 있는 거대한 모델들이다. 그래서 학습과정에서 자연스럽게 병렬 컴퓨팅이 필수적이고, 병렬 컴퓨팅에 있어서 각 연산이 동기화되어 있어야 하기 때문에 레이턴시가 반드시 낮아야 한다.

그런데 과거에는 이더넷보다 인피니밴드의 레이턴시가 낮은 것이 상식이었다. 하지만 최근에는 RoCE 와 인피니밴드가 비슷한 레이턴시를 보이고 있다는 내용의 논문이 발표되어 주목을 받았다. 이에 따라, 이더넷의 약점이라고 불렸던 레이턴시 문제에서 다소 자유로워질 수 있다고도 보인다.

** RDMA는 Remote Direct Memory Access의 약자로, CPU를 사용하지 않고 서버 메모리 간에 데이터를 직접 전송할 수 있는 네트워크 기술. 데이터센터에서 RDMA는 서버 간에 데이터를 전송하는 데 필요한 CPU 처리량을 줄임으로써 데이터 집약적 애플리케이션의 성능과 효율성을 개선하는 데 사용. RDMA는 인피니밴드 및 이더넷을 비롯한 다양한 유형의 네트워크에서 다 사용 가능. 반면에 RoCE는 이더넷 네트워크를 통한 RDMA의 특정 구현형태.*

** RoCE(RDMA over Converged Ethernet)는 이더넷 기반 RDMA 프로토콜. 컴퓨터와 스토리지 간 빠른 데이터 전송을 위해 개발. 데이터센터에서 주로 활용. RoCE를 사용하면 기존 이더넷 네트워크에서도 RDMA 기술을 사용할 수 있으므로, 인피니밴드와 같은 특수 고속 네트워크를 사용하는 것과 비교할 때 상당한 비용 절감 효과.*

AI 클러스터 연결성에 비춘 이더넷과 인피니밴드의 성능특성 비교

AI CLUSTER CONNECTIVITY FEATURES/ATTRIBUTES	INFINIBAND	ETHERNET
Fabric Bandwidth	●	●
Port Speed	●	●
Reduction in Tiers for Large Cluster Size	●	●
Load Balancing – Dynamic	○	●
Perfect Load Balancing	○	●
PFC, ECN, FECN/BECN, Credit Loop FC	●	●
End to End Congestion Management	○	●
Hardware Based Link Rerouting	○	●
Fabric Management	●	●
Telemetry	○	●
Low Unloaded Latency	●	●
Power Economics	●	●

자료: Broadcom

이렇게 중요한 이더넷을 다루는 장치로는 스위치가 대표적이다. 회사나 집에서 사용하는 컴퓨터, 주변기기, 또 서버들이 서로 통신이 가능하다. 이처럼 지역별 네트워크로 통신을 가능케 하는 것이 바로 스위치다. 연결망을 내부에서만 쓰는 것이 아닌듯, 인터넷처럼 더 큰 개념의 네트워크와 스위치가 연결될 수 있도록 하는 게 라우터다.

주요 네트워킹 장비 제품



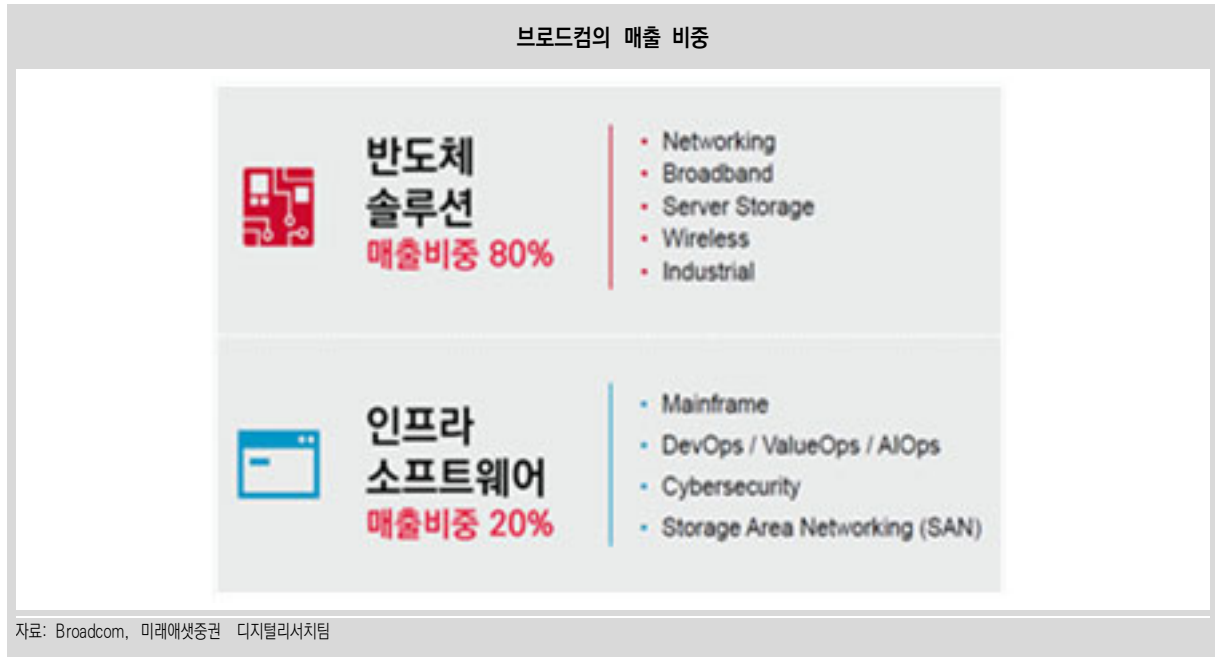
자료: 미래에셋증권 디지털리서치팀

스위치와 라우터에서도 핵심이 되는 것은 그 안에 들어가는 연산 코어 유닛이다. 즉, 스위칭과 라우팅 칩 부문에서 전세계 점유율 85%를 자랑하는 최선두 기업이 있다. 그게 바로 브로드컴(Broadcom)이다.

b. '연결'의 최강자 브로드컴

브로드컴은 쉽게 말해 연결성이라는 것에 거의 모든 역량을 집중한 회사라고 보면 된다. 데이터센터 안에 있는 수많은 고성능 칩인 CPU, GPU, NPU들 간에 데이터가 이동할 때의 병목, 또 서버간에 데이터 이동 시 병목을 최소화하기 위한 연결성에 집중해 반도체를 만드는 회사다.

브로드컴은 반도체만 만드는 게 아니라, 데이터센터 클라우드와 관련한 인프라 소프트웨어 사업도 하고 있다. 다만, 매출 비중이 반도체 분야가 압도적이기 때문에(매출 점유율 80%) 반도체 섹터로 분류한다. 물론, 와이파이라든가, 블루투스 같은 유무선 연결 관련 반도체 설계에도 일가견이 있는 기업이기도 하다.



브로드컴은 지난 3월 2일 2023년 1분기(회계연도 기준) 실적을 발표했다. 브로드컴의 반도체 솔루션 사업부는 21% 매출성장을 기록했는데, 고성장 요인은 클라우드 사업자들의 지출이 9개 분기 연속으로 두 자릿수로 성장한 데에 대한 수혜라고 회사측은 설명했다. 더군다나 판매 단가 상승으로 매출이 올라간 것이 아니라, 판매수량 자체가 증가했기 때문이라고 한 대목이 인상깊었다. 나중에 판매 단가가 비싼 제품들이 더 많이 팔리게 된다면 매출 상승이 가속화될 가능성이 높다는 말이다.

브로드컴의 반도체 솔루션 사업부 중에서 세부영역을 뜯어보면, 네트워킹, 서버스토리지, 브로드밴드, 무선 등 4가지로 나눌 수 있다. 여기서 핵심이 되는 것은 역시 네트워킹 영역이다. 전년동기대비 20% 상승을 하며 4가지 사업부 중에서 가장 높은 매출을 기록했다. 반면, 브로드컴의 원래 캐시카우였던 무선 사업부는 지속적으로 비중이 줄어들 것으로 보인다. 최대 고객사인 애플이 아이폰 등에 들어가는 무선칩을 독자화하려는 움직임이 있기 때문이다.

따라서, 앞으로 브로드컴에게 있어 네트워킹 사업부의 중요성은 더 커질 수밖에 없다. 더군다나 CEO가 앞으로도 이 분야 성장이 가속화될 것이라고 언급하기도 했다. 이것은 인공지능 산업의 발전과 흥행에 직결된 분야이기 때문에 일리가 있다고 판단한다.

브로드컴의 사업부별 성장과 각 가이던스, '네트워킹 반도체가 핵심'

사업부	1분기 매출	전년 동기 성장률	CEO의 관련 언급 및 가이던스
반도체 솔루션	71억 달러	+21%	* 클라우드 사업자들의 지출이 9분기 연속 두 자릿수 성장해서 수혜를 받았음. * 특히 판매단가 상승이 아닌, 판매수량의 증가 덕분에 매출이 올랐음
네트워킹	23억 달러	+20%	향후 성장이 가속화될 것
서버 스토리지	13억 달러	+57%	2분기에는 +20% 성장으로 완만해질 것
브로드밴드	12억 달러	+34%	2분기에는 +10% 성장으로 완만해질 것
무선	21억 달러	+4%	2분기에는 -8~9% 하락으로 줄어듦
인프라 소프트웨어	18억 달러	-1%	2분기에는 +5% 정도의 성장 반등할 것 * 핵심 소프트웨어는 성장세 여전했음

자료: Broadcom, 미래에셋증권 디지털리서치팀
 주: 동사의 회계연도 기준, Broadcom은 10월 결산 기업

c. 브로드컴, '인공지능으로 인한 수혜'

앞서 설명했다시피, 인공지능 연산을 돕기 위해 GPU는 상당히 동기화된 방식으로 병렬 작동한다. 대량 '매개변수 교환'이라 불리는 것을 수행하기 위함이다. 기본적으로, GPU든, TPU든, 기타 AI 가속기든, AI 연산을 수행하기 위해서는 모두 함께 실행돼야 한다. 즉, AI 연산을 담당하는 모든 엔진들이 병렬로 결합해야 한다. 다시 말해, AI 전용 이더넷의 경우 말 그대로 무손실, 매우 낮은 대기시간(레이턴시)을 가져야 한다는 것이다. AI 시대가 오면, 기존 x86 CPU 스케일과는 별개로서, 네트워킹 구현방식이 기존의 차원과는 다르게 대규모로 확장되어야 한다는 말이다.

게다가 400G, 800G의 속도로 "대규모 동기화 데이터 버스"도 가능한 높은 성능의 네트워킹 장비의 필요성이 절실해진다. 네트워킹은 이제 하드웨어에서 전체 AI 학습과 수행에서 매우 중요한 부분을 담당하게 됐고, 네트워킹에서의 혁신과 신제품이 생성형 AI의 꿈을 실현하는 데 매우 핵심적인 문제로 자리한다. 즉, 이더넷의 혁신이 생성형 AI를 실현시킬 마중물이다.

* G는 Gbps의 줄임말로 초당 몇 기가비트를 전송할 수 있는지를 의미.

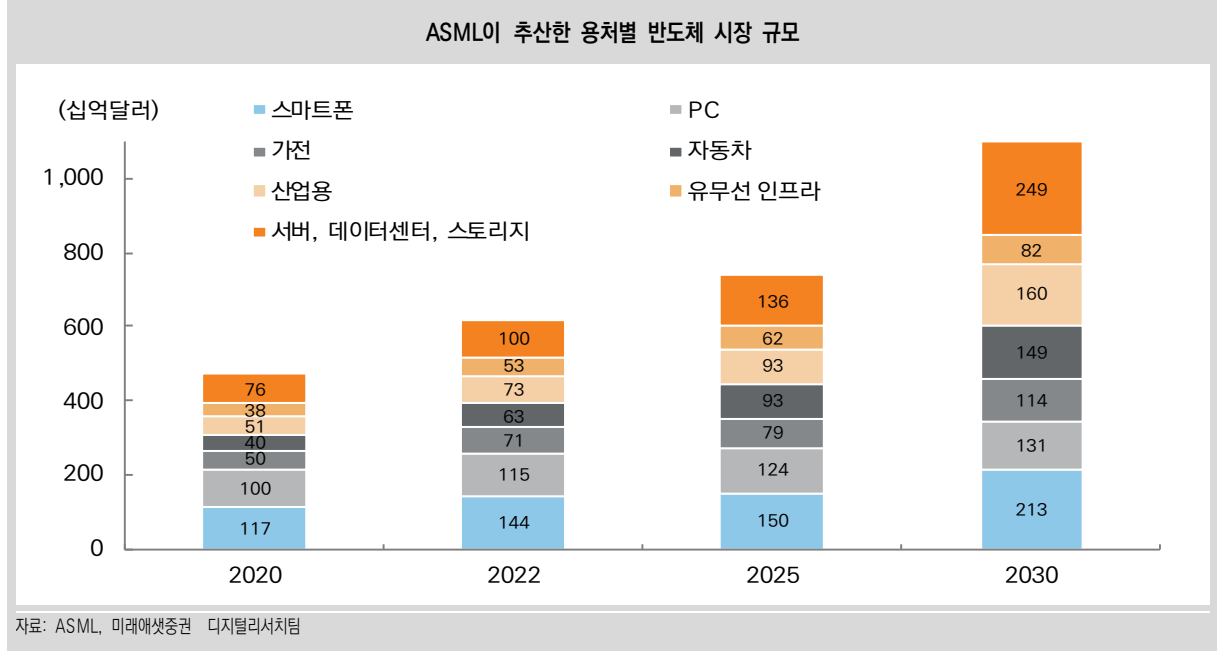
그런데, 위 사항들은 브로드컴이 지금까지 주력해왔던 비즈니스의 연장선이다. 실제로 브로드컴은 이미 인공지능 워크로드를 처리하도록 특별히 설계한 이더넷 기반 차세대 제품들을 위한 투자를 지속적으로 해왔다. 무손실, 저지연 이더넷 패브릭에 집중을 했다는 소리다.

이런 역량을 바탕으로, 브로드컴은 FY2023 2분기에 인공지능 연산 수요 증가로 데이터 처리량이 늘어나며 네트워킹 부문 매출이 20% 증가할 것이라는 가이던스를 제시했다. 지난 1분기 실적발표에서 이미 점점 더 많은 스위치가 브로드컴의 고객사들의 인공지능 연산을 위한 데이터센터 네트워킹에 배포되었다는 점도 확인되었다.

이어서 브로드컴의 CEO가 1분기 실적발표에서 굉장히 희망찬 발언들을 이어 나갔다는 점도 주목할 만하다. ChatGPT같은 생성형 AI에 대해서 자주 언급하면서, 막대한 성장 잠재력이 있다고 강조했다. 그의 발언 중 특히 "ChatGPT가 나오고, 엄청난 흥행돌풍을 일으키면서, 다른 데이터센터 사업자들이 긴장하고 또 안달이 났다"는 얘기가 정말 인상적이었다. ChatGPT를 소유한 마이크로소프트를 제외한 데이터센터 사업자라고 한다면, 아마존과 구글, 그리고 메타 정도가 대표적일 것이다.

현재 그런 사업자들로부터 매우 긴박한 반도체 칩 수요가 있다고 몇 차례나 CEO가 강조를 한 것이다.

이에 따라, 하이퍼스케일러들의 기하급수적인 반도체(이더넷 스위칭 포함) 수요가 예상됨에 따라브로드컴은 큰 수혜를 보게 될 것이다. 2022년에 브로드컴의 시용 이더넷 스위치 출하량이 2억 달러 정도로 추정되는데, 올해에는 8억 달러 가량으로 4배 이상 성장할 것으로 보인다.



그럼에도 불구하고, CEO는 아직도 매우 초기 단계라면서 투자자들을 흥분시켰다. 여기서 우리가 알 수 있는 것은 바로 이것이다. 드디어 보통 사람들이 쓸 만한 인공지능 서비스가 엄청난 인기를 끌면서, '이 물결에 한 번 뒤처지면 죽는다'는 생각이 기업들 사이에 널리 퍼지고 있다는 것이다.

특히 빅테크 입장에서 경기 침체는 이제 뒷전이 됐다. 이것은 생존의 문제다.



게다가 빅테크들 입장에서는 본인들이 보유한 거대한 데이터를 효과적으로 AI 연산에 맞게 처리할 특수 맞춤형 반도체도 필요로 하게 될 것이다. 앞으로도 데이터센터 사업자들이 자체적으로 본인들의 서버에 맞게 칩을 주문 제작하는 형태가 유행이 될 수 있다는 이야기다. 이때 맞춤형 반도체를 공동으로 개발하는 데에 있어 브로드컴의 역할이 중요하다는 사실을 잊어서는 안된다.

이 분야의 점유율 1위가 브로드컴이기 때문이다. 브로드컴과 함께했던 파트너 회사들이 테슬라, 구글, 메타, 마이크로소프트 등이 있다. 관련사업인 “데이터센터 맞춤형 칩 프로그램”은 2022년에 20억 달러로 급성장했는데, 올해는 30억 달러가 될 것이라고 자신 있게 CEO가 밝히기도 했다. 브로드컴이 원래 보수적으로 시장을 전망한다고 했었는데, 이와는 상반되게 힘찬 어조로 말했다는 점에서 자신감 또한 느낄 수 있었다.

d. 브로드컴 “한 발 더 남았다”

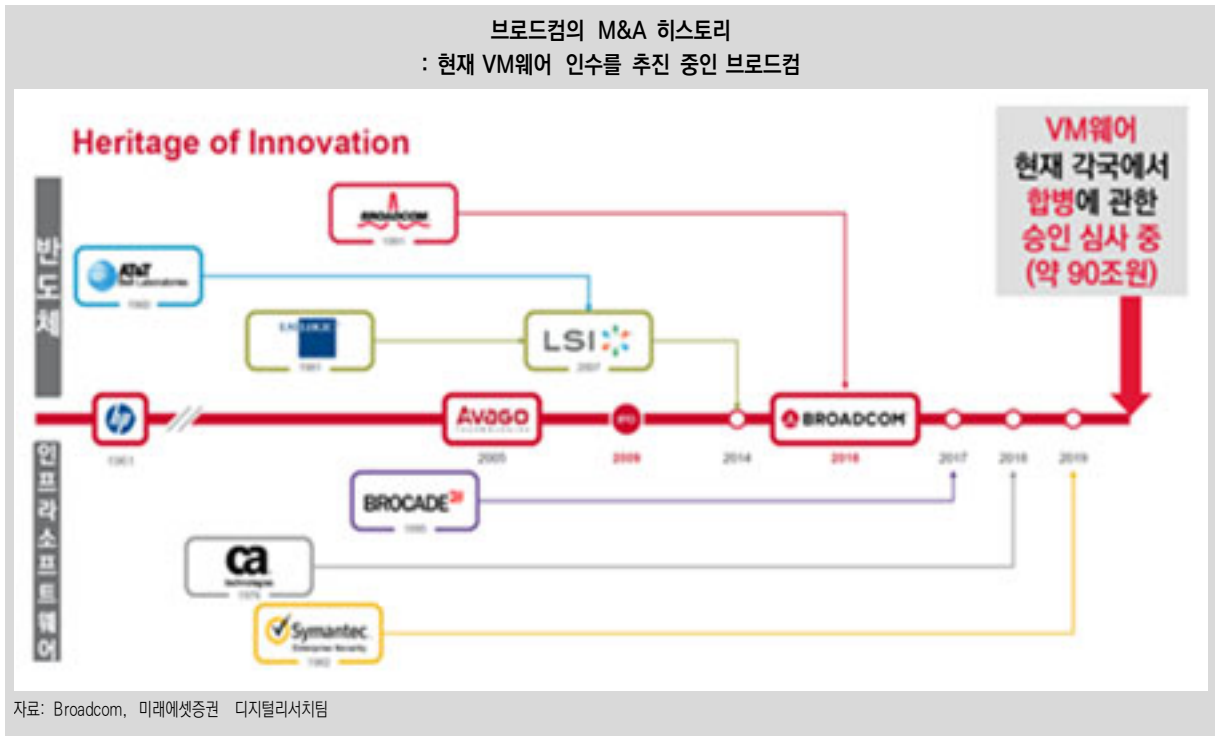
마지막으로 브로드컴이 갖고 있는 무기는 또 하나가 더 있다. 매출 비중이 20%에 머물렀던 소프트웨어 영역이 바로 그것이다. 인공지능 시대가 도래하면, 필연적으로 클라우드 필요성은 더 대두가 될 것이다. 인공지능 연산을 하려면 고성능의 GPU가 필요한데, 개인들이나 많은 중소기업들 또한 다수의 GPU를 구비하기에는 현실적으로 불가능하기 때문이다. 더군다나 GPU 가격은 신제품이 나올 때마다 계속해서 올라가고 있다. 결국 대다수의 사람들은 클라우드 형태로 인공지능을 쓰는 쪽으로 쏠릴 것이다.

실제로 스테이블 디퓨전과 같은 고성능 이미지 생성형 AI를 구동시키기 위해서, 대부분의 사람들은 구글의 클라우드 컴퓨팅 기능을 이용한다. 구글의 Colaboratory(일명 코랩)이라는 것으로 구글이 데이터센터에서 AI 연산을 위한 호스팅 서비스를 해주는 것이다. 특정 사용량까지는 무료로 사용할 수 있다.

물론, 연산량이 특정 단위를 넘어가거나 좀 더 높은 품질의 연산장치를 이용하기 위한 사람들을 위해 월간 구독 모델을 지원하기도 한다. 월 9.99 달러를 내면, 구글 데이터센터 내에서 구동되는 엔비디아의 A100 GPU를 이용할 수 있다. 앞으로 생성형 인공지능이 점차 자리를 잡아 완전히 대중화가 된다면 구독료가 인상될 여지도 있다.

여기서 중요한 점은, 클라우드를 사용하려면 거기서 끝이 아니라, 그것을 효율적으로 관리하고, 또 안전하게 사용할 수 있도록 하는 소프트웨어도 수반돼야 한다는 점이다.

그런 측면에서 브로드컴은 여러 클라우드 서비스를 통합 관리하는데 특화된 VM웨어라는 회사를 올해 안에 인수할 예정이다. 현재 각국으로부터 승인 심사 중인데, 이미 통과된 곳이 많아서 아마 순조롭게 통합 작업이 이뤄질 것으로 전망된다.



VM웨어는 클라우드 컴퓨팅 및 가상화 소프트웨어와 서비스를 제공하는 소프트웨어 회사다. VM웨어의 가상화 소프트웨어를 사용하면 단일 물리적 시스템에서도 여러 운영체제를 실행할 수 있다. 즉, 서버 활용도를 높이고 하드웨어 비용을 줄일 수 있다는 말이다. 클라우드 서버를 관리할 때, IT 대응력, 유연성, 확장성을 높이면서도 비용을 대폭 절감할 수 있게 된다. 물론 가상화되지 않은 서버와 비교하면 다소 퍼포먼스가 떨어지기는 하다.

VM웨어에서 제공하는 기능 중 가장 중요한 것은 “리소스 최적화”다. 여러 가상머신이 단일 물리적 서버에서 실행되어 서버 리소스를 최적으로 활용 가능하다. 이는 동일한 워크로드를 처리하는 데 필요한 물리적 서버 수가 적어도 되기 때문에 상당한 비용 절감으로 이어진다.

회사에 각각 하나의 운영체제만 실행하는 5대의 물리적 서버가 있다고 가정해보자. 즉, 각 서버별로 약 20% 용량만 활용되고 있어 그다지 효율적이지는 않다. 여기서 가상화 소프트웨어를 사용하면 엄청난 효율이 생긴다. 물리적 서버 한 대만 둔 다음에, 각각 자체적인 운영체제가 있는 5개의 가상머신(VM)을 그 서버 하나로 구동시키면 된다. 이렇게 하면 회사에서 5대가 아닌 1대의 물리적 서버만 구매하고 유지관리하면 되므로 하드웨어 비용을 크게 줄일 수 있다. 다시 말해, 최종 사용자에게 동일한 수준의 기능을 제공하면서 물리적 서버를 약 80-90%의 훨씬 더 높은 용량으로 활용할 수 있는 것이다.

또한, Isolation 기능을 제공하게 되는데, 각 가상머신은 다른 가상머신과 격리되어 보다 쉽게 유지 관리할 수 있게 된다. 하나의 가상 머신이 실패하거나 충돌하더라도 동일한 물리적 서버에서 실행 중인 다른 가상 머신에는 영향을 미치지 않는 것이다. 게다가 가상 머신을 쉽게 백업 및 복구할 수 있으므로 장애 발생시 가상 머신을 보다 간단하게 복원할 수도 있다.

앞으로 브로드컴의 입지는 인공지능과 클라우드 시대가 더욱 본격화되면서 공고해질 가능성이 크다고 판단한다. 하드웨어와 소프트웨어 통합 패키지를 제공하면서 AI 연산을 위한 데이터센터 영역에서 지배력을 향상시킬 것이기 때문이다. 생성형 AI에서 숨겨진 강자가 바로 브로드컴이 아닐까?

V. [중국] 중국도 ChatGPT 열풍

바이두로 포문을 연 '중국판 ChatGPT'

OpenAI의 ChatGPT 열풍에 중국도 동참했다. 중국의 대표 빅테크 BAT(Baidu, Alibaba, Tencent)도 앞다퉀 현재 연구 개발중인 ChatGPT류의 AI 프로젝트를 공개했다. 바이두는 중국에서 가장 빠르게 'Ernie Bot'을 공식 발표하며 AI 산업의 대표 주자임을 다시 한번 각인시켰고, 알리바바도 내부 테스트중인 ChatGPT 프로그램 일부 공개하며 기대감을 키웠다. 그 외에도 징둥, 넷이즈, 틱톡 등 주요 빅테크별로 현재 진행중인 AI 프로젝트와 향후 응용 방안에 대한 발표가 이어지면서 빅테크, 반도체, S/W 업종에 관심이 집중되었다.

중국 대표 테크기업이 진행중인 ChatGPT 프로젝트 및 주요 기능

주요 기업	대표모델	주요 기능
바이두	Ernie	문장 및 이미지 생성, 연산, 문학 창작 등
텐센트	优图	인물 사진을 조합해 3D 모형 제작, 영화해상도 상승 등
알리바바	Lubanner	이미지, 분위기, 사이즈 등을 입력 후 자동분석 후 디자인, 컬러 배치
틱톡	지안잉(剪映)	영상 및 자막 자동 생성
징둥	ChatJD	ChatJD 플랫폼에서 소매판매, 금융 사업에 응용. 커머셜 콘텐츠 생성, 챗봇, 사용자 의도 이해, 정보추출, 감정 분류
NetEase	Youdao / Tianyin	온라인 교육, 번역, 언어학습, 인공지능 텍스트 분석, 음성 인식 및 클라우드 기반 플랫폼 제공 Tianyin은 음악 스트리밍 기능 외에 작곡, 작곡에 필요한 기능 생성

자료: 미래에셋증권 디지털리서치팀

1) 바이두 '어니봇(Ernie Bot/文心一言)'으로 포문을 열다

바이두는 중국 기업 최초로 '어니봇(Ernie Bot/文心一言)'을 3월 16일 공식 발표했다. 발표 당시 현장 시연이 없었다는 점에 실망감이 커지며 바이두 주가가 격하게 반응하기도 했다. 그럼에도 불구하고 문학 창작, 상업문구 창작, 수리 추산, 중국어 이해, 멀티 모달까지 5개 주요 기능을 소개했고 중국 내 높은 수요와 650개 기업이 밸류체인에 참가했다고 밝히면서 시장의 이목을 집중시켰다. 발표회 직후 테스트 코드를 부여받은 사용자 대상으로 베타서비스가 제공 중이고 3월 27일 정식 공개 예정이다.



자료: Baidu



바이두의 어니봇은 2019년 이미 GLUE(General Language Understanding Evaluation)라는 AI 자연어 평가 처리 능력에서, 구글과 MS보다 높은 90점 이상의 점수를 획득하며 1위를 차지했다. 2019년 당시에 중국어의 해석 능력에서 가장 큰 두각을 나타냈는데 9개 지역의 방언을 완벽하게 해석, 중국어 한자 표기 방법인의 간체와 번체 해석이 가능했다.

이번 2022년 3월에 공개된 어니봇의 주요 기능 중에서도 중국 고전 시와 성어 해석을 통한 재창작, 중국 인끼리도 알아듣기 힘든 사투리까지도 이해한다는 점은 다시 한번 중국어에 특화된 언어모델로서의 기술력을 인정받고 공고한 위치를 선점했다는 평가다. 다만 이런 점이 중국에서는 특징점으로 강조되는 반면 구글, MS 등 글로벌 대표 ChatGPT 대비 기능과 사용 범위가 제한적이라는 평가와 중국 정부의 심한 내부 검열로 인해 향후 글로벌 진출까지는 한계가 있다는 우려도 상존한다.

그럼에도 불구하고 중국 내에서는 발표 5일만에 서비스 신청 기업이 10만개 이상, 사용 신청자 수도 100만명이 넘어설 것으로 예상하고 있다. 향후 자율주행, 의료, 교육, 창작 등의 영역에서 중국 특색의 AI 기술이 응용될 것으로 예상된다.

2019년 AI 자연어처리 능력 평가: 1위 바이두 Ernie 90.1점, 2위 마이크로소프트 80.9점

Rank	Name	Model	Score
1	ERNIE Team - Baidu	ERNIE	90.1
2	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART	89.9
3	T5 Team - Google	T5	89.7
+ 4	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)	89.5
5	XLNet Team	XLNet (ensemble)	89.5
6	ALBERT-Team Google Language	ALBERT (Ensemble)	89.4
7	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	88.8
8	Facebook AI	RoBERTa	88.5
9	Junjie Yang	HIRE-RoBERTa	88.3
+ 10	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	87.6

자료: GLUE

바이두에 이어 알리바바, 텐센트의 중국어 기반 자연어 처리 능력도 높은 평가를 받았다. GLUE를 벤치마크한 CLUE(중국어 자연어 처리 능력 평가)에서도 인간보다 높은 점수를 받으며 기술력을 인정받았다. 향후 전자상거래, 클라우드, SNS 등 알리바바와 텐센트가 영위하는 핵심 사업에 ChatGPT 기술이 다양하게 접목될 것으로 기대된다.

CLUE (중국어 자연어 처리 능력 평가) 결과. 1위 텐센트, 2위 알리바바, 3위 인간

排行	模型	研究机构	测评时间	Score1.1	认证	AFQMC	TNEWS1.1	IFLYTEK	OCNLI_50K	WSC1.1	CSL
1	HunYuan-NLP-1T	腾讯混元AI大模型团队	22-11-26	86.918	待认证	85.11	70.44	67.54	86.5	96	96.2
2	通义-AliceMind	达摩院NLP	22-11-22	86.685	待认证	84.07	73.47	67.42	85.87	94.33	95.03
3	HUMAN	CLUE	19-12-01	86.678	已认证	81	71	80.3	90.3	98	84
4	CHAOS	OPPO研究院融智团队	22-11-09	86.552	待认证	83.37	73.22	65.81	86.37	94.6	95.7
5	WenJin	Meituan NLP	22-10-20	86.313	待认证	84.49	73.04	64.38	86.23	94.44	95.67
6	OBERT	OPPO小布助手	22-11-07	84.783	待认证	81.02	67.75	66	84.53	91.3	99.93
7	HunYuan_nlp	腾讯TEG	22-05-11	84.730	待认证	83.37	64.01	66.58	85.23	92.27	93.87
8	ShenNonG	云小微AI	21-12-01	84.351	待认证	82.57	65.56	64.42	85.97	94.21	91.23
9	ShenZhou	QQ浏览器实验室(QQ Bro...	21-09-19	83.873	待认证	80.55	65.36	67.65	86.37	89.08	90.97
10	MusaBert	mthreads	22-11-26	81.759	待认证	84.49	60.48	63.88	81.6	88.93	90.7

最新一期CLUE最新排名

자료: CLUE

2) ChatGPT로 부각된 중국의 AI 산업 육성 의지

ChatGPT로 AI 산업에 대한 관심도가 최근 급부상하긴 했지만 중국의 AI 산업에 대한 발전 목표와 부양 의지가 하루 아침에 생긴 것은 아니다.

14억 인구 보너스라는 수혜로 제조업 강국을 이끌어왔던 중국은 인구 감소 추세에 직면하면서 이는 생산성 문제와 연결되었다. 중국 경제의 급성장을 이끌었던 일반 제조업 중심에서 한 단계 올라서기 위한 생산성 향상과 첨단 기술 업그레이드의 니즈는 더욱 명확해졌다. 또한 인민의 생활과 산업 가동을 통해 생성된 대량의 데이터를 바탕으로 효율적인 민생 관리, 통제 등을 위한 인공지능 첨단 기술은 더 필요해졌다.

이에 따라 중국은 적극적인 AI 산업 지원 정책을 펼치고 있다. 인터넷플러스 발전, 중국제조 2025, 기술 굴기 등 첨단 기술 발전을 위한 정책은 이미 2010년대부터 시작되었다. 여기에 AI 산업에 대한 본격적인 지원은 2018년 10월 중국 과학기술부가 발표한 <차세대 인공지능 발전규획>과 <차세대 인공지능산업발전 3년 행동계획(2018-2020년)>를 통해 AI 제품 대규모 개발, AI 전반 인프라 기능 향상, 스마트 제조 발전, AI 지원 시스템 구축을 강조하며 지원 방향을 수립했다.

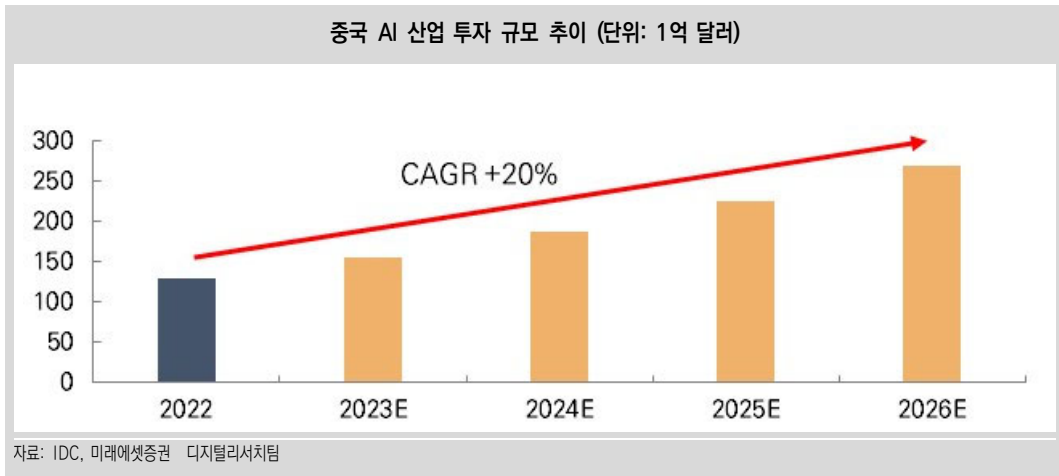
또한 2019년 국무원은 상하이 자유무역시범 지역을 지정하고 디지털경제 발전 혁신에 대한 가이드라인을 마련했다. 또한 상하이, 베이징, 선전 등 첨단 산업 기술 거점 도시로 선정해 주요 기업 육성 및 지원을 적극적으로 진행하고 있다. 14.5규획(2021~2025년)에서도 핵심 발전 사업에 인공지능이 포함되었고, 시진핑 주석의 3연임에서도 가장 강조된 산업은 과학기술이었다. 이와 같은 노력을 바탕으로 중국을 대표하는 빅테크 기업들의 AI 관련 기술 성과가 가시화되고 있다.

2022년 중국 주요 기업별 시관련 특허 획득 수

순위	기업명	지역	특허 획득	순위	기업명	지역	특허 획득
1	텐센트	선전	1,563	6	핑안그룹	선전	355
2	바이두	베이징	1,448	7	DJI	선전	314
3	화웨이	선전	571	8	OPPO	광둥	270
4	센스타임	베이징 / 선전	559	9	하이크비전	항저우	239
5	알리페이	항저우	442	10	BOE	베이징	212

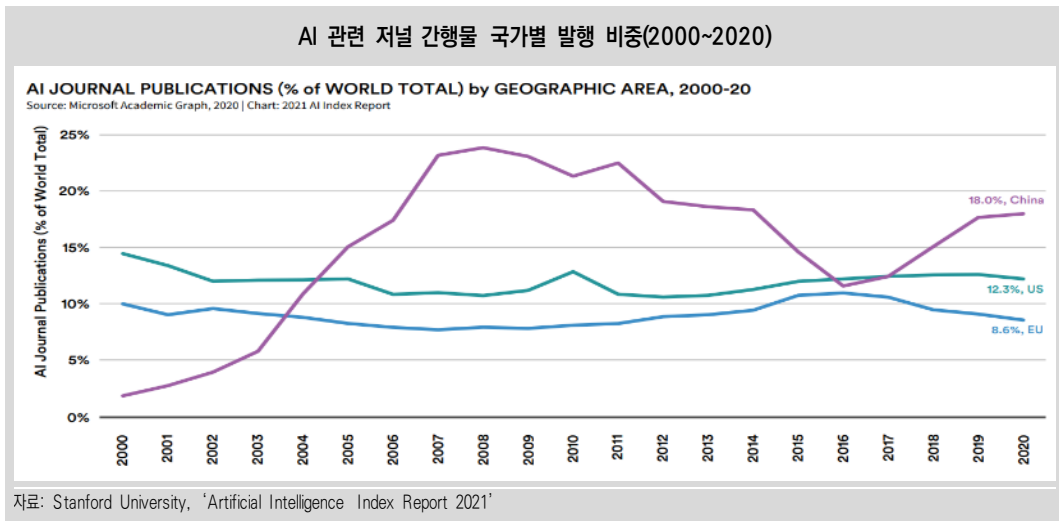
자료: 중국 언론, 미래에셋증권 디지털리서치팀

중국의 AI 산업 투자도 빠르게 확대 중이다. IDC와 낭조정보의 공동 자료에 따르면, 2021년 기준 글로벌 AI 투자 총액은 929억 달러, 연평균 26.5% 증가해 2026년에는 3014.3억 달러 규모의 투자가 이루어질 것으로 예상했다. 2022년 중국 AI 시장의 투자 규모는 130.3억 달러로 수준으로 2026년에는 두배 수준인 266.9억 달러를 예상했다. 이는 전체 글로벌 AI시장의 2위 규모로 8.9% 수준이다.



학계에서도 적극적인 연구개발의 성과가 나타나고 있다. 스탠포드 대학이 발표한 ‘인공지능 지수 보고서’에 따르면 중국은 특허 출원, 학회 간행물, 저널 간행물, 저널 이용횟수 등 4개 지표에서 이미 미국을 제치고 글로벌 1위를 기록했다.

US 뉴스&월드리포트 데이터에 따르면 2021년 글로벌 베스트 대학 순위에서도 AI 전공 관련 상위 대학교 중 중국의 칭화대, 하얼빈 공업대, 중국 전자과학기술대학, 북경대, 중국 과학원 등이 상위에 랭크되었고 인재육성과 학계 연구에 중국이 주력하고 있음을 확인할 수 있다.



3) 미국과의 대립은 리스크지만 또 다른 기회

2018년 미국의 트럼프 정권부터 시작된 미-중 갈등은 무역에서 기술 규제로 옮겨졌다. 국가 안보 문제로 화웨이 등 중국의 대표 첨단기술 관련 중국 다수의 핵심 기업에 대한 규제가 이루어졌고 현재까지 규제 기업의 범위는 더욱 확대되고 있다.

미국은 반도체와 장비 공급 규제로 중국의 반도체 제조 기술 업그레이드를 차단했고, 최근에는 틱톡, 센스타임 등의 AI 기술이 미국에 침투되는 것을 매우 경계하는 중이다. 표면적으로는 청소년의 정신 건강 위협을 내세웠지만 결국 미국의 국가 보안, 데이터가 중국에 빠져나갈 수 있다는 점도 경고했는데 이는 AI 기술력에서도 중국이 이미 미국의 안보까지 위협할 수준이라는 것을 반증한다.

미국 규제로 인한 매출 타격, 기술 제한 등의 리스크로 중국은 여전히 녹록치 상황이지만 오히려 이런 압박이 중국의 기술 자립에 대한 의지를 강화시켰다. 실제로 미국의 규제 이후 미국향 매출 비중이 줄어든 기업은 2~3년간의 사업 구조 조정 이후 중국 내수 비중을 확대시켜 실적 감소분을 만회하는 추세이고, 미국에 대한 기술 의존도를 낮추기 위한 정부와 기업의 필사적인 노력이 지속되고 있다. 화웨이가 안드로이드를 사용하지 못하게 되면서 받았던 충격은 오히려 중국 자체 OS 개발, 중국 국산 소프트웨어 개발과 교체의 성과로 나타났다.

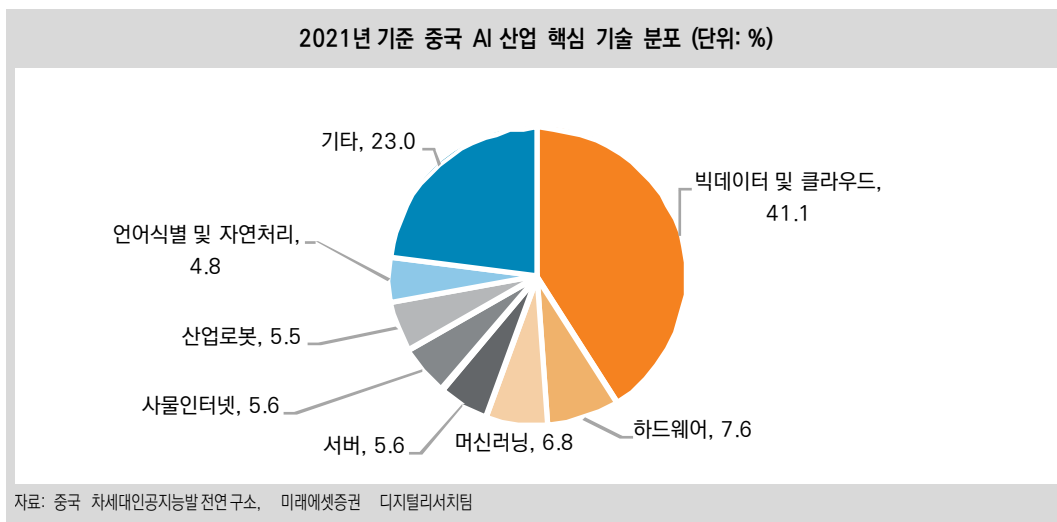
향후 중국판 ChatGPT에서도 국산화 의지는 더욱 강해질 것이 분명하다. 미국의 규제에서 자유롭고 중국에 특화된 ChatGPT 개발과 밸류체인 국산화 의지는 더욱 공고해질 것으로 예상된다. 또한 중국 국가차원에서 필요한 자금과 자원 투자를 아끼지 않고 있어 중국의 기술력 향상과 산업 발전 속도 또한 가속화 될 것이다.

중국 AI 산업 밸류체인과 향후 전망

1) 중국 AI 산업 밸류체인

ChatGPT로 부각된 생성형 AI에 대해 중국에서는 AIGC(AI Self-Generated Content, 이하 AIGC)로 통용하면서 AI, ChatGPT, AIGC 테마로 분류된 관련 기업들 주가가 급등했다. 다만, 일부 핵심 기업을 제외하고 테마성 요소도 아직은 다분하고 외국인들이 거래되지 않거나 중소형 종목 위주로 변동성이 매우 크기 때문에 주의도 필요하다.

중국 AIGC 밸류체인은 플랫폼, 알고리즘, 컴퓨팅 파워(연산), 응용으로 나눌 수 있다.



(1) 플랫폼: 중국의 AI 산업에서 가장 비중이 큰 기술은 빅데이터와 클라우드다. 알리바바, 텐센트, 바이두 등의 빅테크 기업들의 클라우드 인프라가 구축되어 있고 향후 대규모 투자를 확대할 수 있는 자본력도 갖추고 있다. 또한 대형 플랫폼을 바탕으로 중국인들의 교통, 금융, 검색, 헬스케어, 금융 등의 다양한 영역의 빅데이터를 수집하기 용이하다. AI 인프라부터 응용 사업까지 전체 밸류체인에 걸쳐 중국의 빅테크의 역할은 매우 중요하다.

관련주: 알리바바, 바이두, 텐센트 등

(2) 알고리즘: 중국시 산업에서 음성, 안면인식, 자연어 처리 등의 알고리즘 기술을 보유한 중국 로컬 기업의 니즈가 더욱 확대될 것으로 예상된다.

관련주: 신비정보 (002230/음성인식), TRS(300229/자연어처리), Speechocean(688787/음성인식), 클라우드워크(688327/신체인식), 센스타임 (00020.HK/안면인식) 등

(3) 컴퓨팅파워(연산): 컴퓨팅 파워를 향상시킬 수 있는 반도체에 대한 수요는 급증하고 있지만 중국 반도체 산업에서 미국의 규제는 여전히 압박이 크다. 다만, 이런 상황속에서 중국 로컬 기업의 생존과 업그레이드의 의지, 정부의 지원이 더욱 강화되고 있다.

관련주: 캄브리콘(688256/AI반도체, Hygon(688041/CPU, DCU설계), 남조정보 (000977/AI서버), 중과서광 (603019/슈퍼컴퓨터)등

(4) 응용: AIGC 기반 비즈니스 가치를 창출할 수 있는 영역이다. 오피스, 미디어, 전자상거래, 금융, 엔터테인먼트 교육 등의 영역에서 활발한 AIGC 활용이 이루어 질 것으로 예상된다.

관련주: 하이크비전 (002415/영상 인식 기반 CCTV 및 영상프로그램), 금산오피스 (688111/OA S/W), 신비정보 (002230/음성인식 기반 통번역, 교육 프로그램) 등



2) 중국 AI 산업의 향후 전망

중국에서도 ChatGPT로 떠오른 AI 산업에 대한 관심이 반짝 지나가지 않을 것으로 보인다.

중국은 AI 산업이 발전하기 유리한 조건을 갖췄다. 14억 인구, 인터넷 산업 발전, 빅데이터에 중국 정부의 적극적인 정책과 의지다. 매일 중국 국무원 공신부, 인민일보 등 주요 부처별 AI 산업 발전을 독려하고 인프라 구축, 투자 확대, 기업 및 인재 육성 등의 메시지를 통해 산업육성에 대한 진심을 표현하고 있다. 빅테크 중심 연구개발에 주력했던 성과도 나타나고 있으며 향후에도 빠른 속도로 글로벌 AI 기술을 선점하기 위한 노력들은 더 확대될 것으로 예상된다. 중국화된 결과물일지라도 평가절하보다는 기술력의 수준이나 발전 속도에 주목하는 것이 더 유의미하다는 판단이다.

이제 더이상 낯설지 않은 틱톡이나 스노우 등 일상 생활에 즐겁고 유용하게 사용 중인 다양한 어플리케이션 또한 중국의 AI 기술력이 포함되어 있는 경우도 다반사다. 크게 인식하지 못하는 이 순간에도 중국의 AI 기술력이 우리의 생활에 점점 더 빠르게 침투하는 중이다. 이런 시점에 나의 데이터만 제공하는 것이 아니라 우리의 삶을 바꿔놓을 글로벌 AI 산업의 개화 국면을 주목하고 미국과 더불어 양대 축인 중국의 AI 산업 발전에 대한 투자 기회를 엿보는 것도 필요한 시점이다.

VI. AI 생태계, ETF로 접근하기

AI ETF 개요

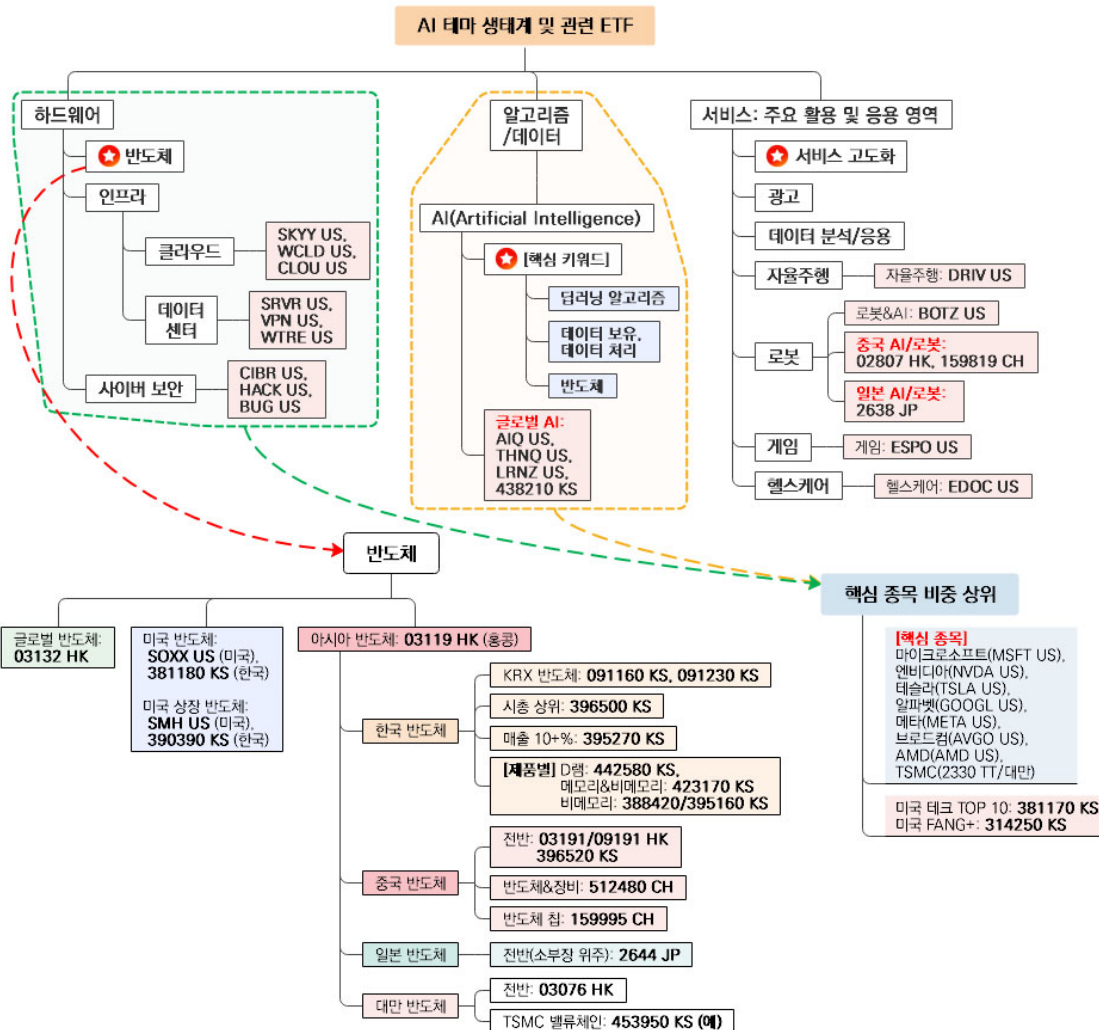
AI 관련 산업은 현재 초기 성장 산업으로, 기술 수용 곡선을 응용한 'Thematic Adoption Menu (Global X)' 내 **Early Adopters (조기 수용자) 단계**에 해당한다. 초기 성장 단계인만큼 주도 기업의 교체, 종목의 진출입이 빈번한 점을 감안할 때 ETF를 통한 접근이 효과적이며, AI 테마 생태계 및 관련 ETF, 테마와의 관련성 및 순도(Purity)를 감안한 AI 관련 핵심 ETF는 아래와 같이 정리할 수 있다.

인공지능 관련 핵심 ETF

구분	티커	ETF 명	개요/비고
AI	AIQ US	Global X Artificial Intelligence & Technology ETF	인공지능(AI) 테마, [중국 상장] 02807 HK
반도체	SMH US	VanEck Semiconductor ETF	미국 상장 반도체 기업, [한국 상장] 390390 KS
핵심 종목 보유(한국 상장)	381170 KS	TIGER 미국테크TOP10 INDXX	미국 '테크주 TOP 10, [FANG+] 314250 KS

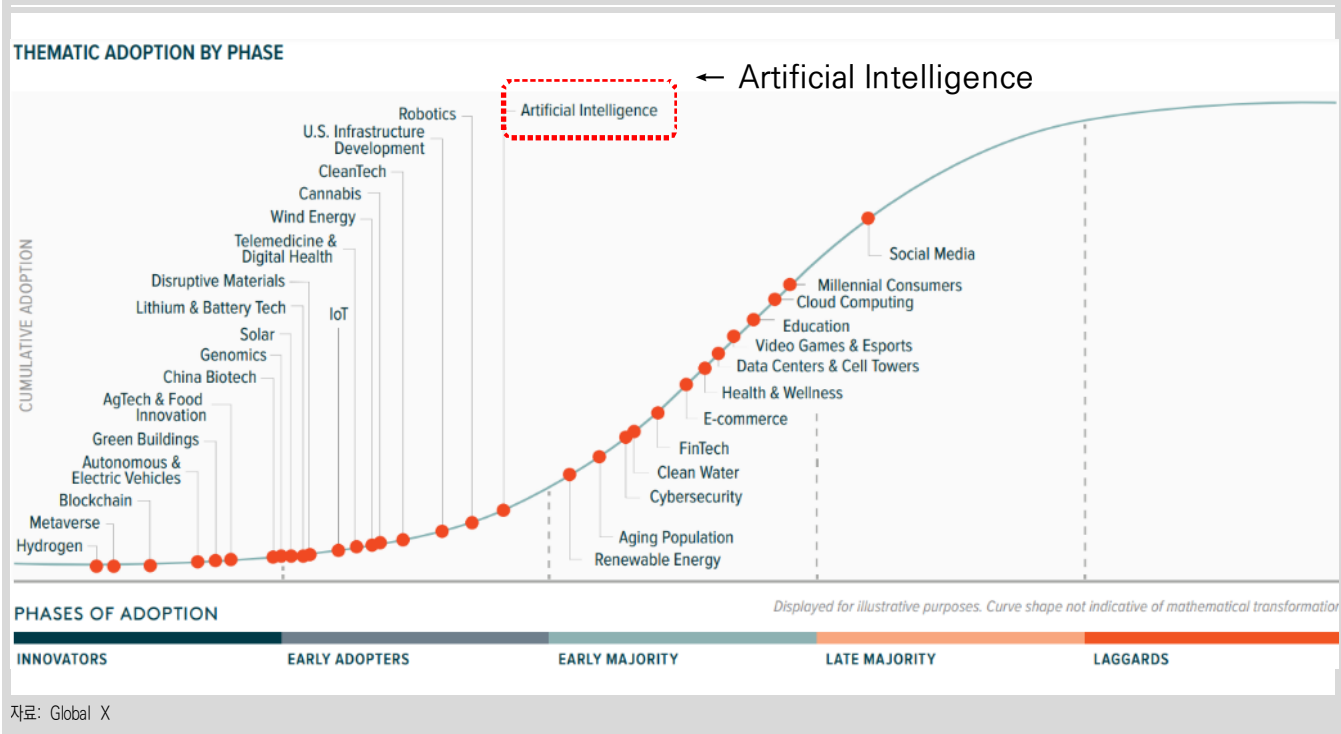
자료: 미래에셋증권 디지털리서치팀

인공지능 테마 생태계 및 관련 ETF



자료: 미래에셋증권 디지털리서치팀

Global X의 'Thematic Adoption Menu': 시장 규모, 성장률 예상치, 기술의 사회적 수용 수준 등을 활용해 기술 수용 단계를 측정



주요 기술 수용 단계 및 단계별 해당 테마 예시

단계	해당 테마 예시
Innovators (혁신적 수용자)	수소(Hydrogen), 메타버스(Metaverse), 블록체인(Block chain)
Early Adopters (조기 수용자)	태양광(Solar), 리튬&배터리(Lithium&Battery Tech), 로봇(Robotics), 인공지능(Artificial Intelligence)
Early Majority (조기 다수자)	재생 에너지(Renewable Energy), 사이버 보안(Cybersecurity), 핀테크(FinTech), E-커머스(E-commerce)
Late Majority (후기 다수자)	소셜 미디어(Social Media)
Laggards (최후 수용자)	- (기술 대중화 단계)

자료: Global X, 미래에셋증권 디지털리서치팀

1) AI 관련 ETF 상장 현황

AI 관련 키워드 스크리닝 검색과 지수 방법론(Index Methodology), 테마형 ETF 운용사인 Global X의 'Thematic ETF Map' 등 복수 기준을 통해 확인되는 주요국* 상장 AI 관련 ETF는 총 27개 종목으로, 주로 미국에 다수 상장되어 있다(*주요국: 미국, 한국, 중국, 홍콩, 일본, 영국).

AI 관련 ETF는 크게 4가지 측면에서 접근할 수 있는데, 1) AI 테마 자체에만 집중한 접근, 2) AI 및 AI와의 연관도가 높은 기타 산업(자동화, 로봇)을 동시에 편입하는 접근, 3) AI 관련 기반 산업(클라우드, 빅데이터, 사이버 보안)에 대한 접근, 4) AI 관련 핵심 종목에 대한 높은 편입 비중을 보이는 ETF를 통한 접근이 대표적이다.

현재 운용자산 측면에서는 AI 테마 자체보다는 기타 산업 및 기반 산업 관련 ETF의 운용자산이 크다. 그러나 본 자료에서는 보다 순도 높은 접근을 위해 1) AI 테마 자체에만 집중한 접근과, 4) AI 관련 핵심 종목에 대한 높은 편입 비중을 보이는 ETF를 통한 접근을 위주로 다룬다. 3) 기반 산업 관련 ETF 및 키워드 스크리닝 결과 내 AI를 활용해 종목을 편출입하는 유형의 ETF를 배제하고, 국내 상장 중 핵심주 보유 비중이 높은 ETF를 추가한 AI 관련 27개 ETF는 아래와 같이 정리할 수 있다.

글로벌 주요국 상장 AI 관련 ETF (기준: 키워드 스크리닝, 지수 방법론 등 참조. *운용자산: 100만 달러/2023.3.21 기준)

구분	티커	상장	ETF명	개요	운용자산	추출
AI	AIQ US	미국	Global X Artificial Intelligence & Technology ETF	인공지능(AI) 산업 테마	124.3	0
	THNQ US	미국	Robo Global Artificial Intelligence ETF	인공지능(AI) 산업 테마	24.4	0
	LRNZ US	미국	Listed Funds Trust TrueMark Technology AI & Deep Learning ETF	인공지능(AI), 딥러닝 산업 테마	15.9	0
	2067 JP	일본	NEXT NOTES AI Companies 70, NTR ETN	AI 테마(AI 관련 키워드 70종목)	15.3	X
	IQM US	미국	Franklin Intelligent Machines ETF	머신러닝&인공지능(AI) 테마	8.3	X
(영국)	438210 KS	한국	ARIRANG 글로벌인공지능산업MV	글로벌 인공지능 산업 테마(능동형 정보 등)	2.1	X
	INTL LN	영국	WisdomTree Artificial Intelligence UCITS ETF	인공지능(AI) 테마, DR(GBX)	356.2	X
	AIAG LN	영국	L&G Artificial Intelligence UCITS ETF	인공지능(AI) 테마, DR(GBX)	282.2	X
AI+기타	RBTX LN	영국	iShares Automation & Robotics UCITS ETF	로봇 산업, 자동화 테마, DR(GBX)	2,933.7	X
	159819 CH	중국	E FundSI Artificial Intelligence ETF	중국 AI 산업(빅데이터, 클라우드 등) 테마	200.6	0
	UBOT US	미국	Direxion Daily Robotics AI & Automation Index Bull 2XShare	로봇 산업, 인공지능(AI), 자동화 테마 +2배	18.1	X
	WTAI US	미국	Wisdomtree Artificial Intelligence And Innovation Fund	인공지능(AI)&혁신기업	8.3	X
	03056 HK	홍콩	Pando Innovation ETF	글로벌 혁신 테마(AI, 핀테크, 메타버스 등)	3.0	X
로봇+AI	BOTZ US	미국	Global X Robotics & Artificial Intelligence ETF	로봇 산업&인공지능(AI) 테마	1,609.1	0
	ARKW US	미국	ARK Next Generation Internet ETF	차세대 인터넷 테마, A	1,294.8	0
	IRBO US	미국	iShares Robotics & Artificial Intelligence Multisector ETF	로봇 산업&인공지능(AI) 테마	285.9	0
	ROBT US	미국	First Trust Nasdaq Artificial Intelligence and Robotics ETF	로봇 산업&인공지능(AI) 테마	221.5	0
	09807 HK	홍콩	Global X China Robotics and AI ETF	중국 로봇&AI 산업 테마/USD	26.3	0
	02807 HK	홍콩	Global X China Robotics and AI ETF	중국 로봇&AI 산업 테마	26.3	0
	BOTG LN	영국	Global X Robotics & Artificial Intelligence UCITS ETF	로봇 산업&인공지능(AI) 테마, SA/U(GBP)	8.9	X
	2638 JP	일본	Global X Japan Robotics & AI ETF	일본 로봇&AI 테마	6.6	0
	03023 HK	홍콩	Ping An Nasdaq AI and Robotics ETF	글로벌 AI&로봇 산업 테마	5.3	X
	양자	QTUM US	미국	Defiance Quantum ETF	양자컴퓨팅 테마	114.4
핵심주	381170 KS	한국	TIGER 미국테크TOP10 INDXX	미국 테크주 TOP 10	1,182.5	0
	314250 KS	한국	KODEX 미국FANG플러스(H)	미국 FANG+/원헤지	322.9	X
	760005 KS	한국	키움 INDXX 미국테크탑10 ETN	미국 테크주 TOP 10	84.6	X
	700022 KS	한국	하나 Solactive US Tech Top 10 ETN(H)	미국 테크주 TOP 10/원헤지	16.8	X

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 2023.3.21 종가 기준. 추출은 레버리지를 제외하고 보유종목의 정상적인 추출이 가능한 운용자산 1천만 달러 이상 ETF 대상. 1천만 달러 이상에서 추출이 되지 않는 경우 차순위에서 추출

한편 이들 ETF의 주요 편입 종목의 보유 금액 및 보유 빈도를 통해 ETF 운용사가 판단하는 핵심 종목을 살펴볼 수 있는데, 아래 '추출 대상 종목 내 보유금액 상위 종목 Top 15'에서 보듯 엔비디아(NVDA US), 마이크로소프트(MSFT US), 애플(AAPL US) 등이 압도적이며, 이 외에 테슬라(TSLA US), 아마존 닷컴(AMZN US) 등도 상위를 차지하고 있다. 이 외에 키엔스(6861 JP) 등은 AI&로봇 ETF의 영향이다.

추출 대상 종목 내 보유금액 상위 종목 Top 15 (*보유금액: 100만 달러, **보유 ETF: 해당 종목을 편입한 ETF 수)

국가	티커	기업명	GICS 하부 산업	*보유금액	**보유 ETF
미국	NVDA US	엔비디아	반도체	336.5	13
미국	MSFT US	마이크로소프트	시스템 소프트웨어	253.1	9
미국	AAPL US	애플	기술 하드웨어, 스토리	233.9	4
미국	TSLA US	테슬라	자동차 제조업체	180.4	7
일본	6861 JP	키엔스	전자 장비&기기	176.6	5
미국	ISRG US	인튜이티브 서지컬	건강관리 장비	175.1	5
일본	6954 JP	화낙	산업용 기계, 용품 및	167.9	7
미국	PATH US	UiPath Inc	시스템 소프트웨어	166.6	8
스웨덴	ABBN SW	ABB	전기 부품&장비	164.3	6
미국	AMZN US	아마존 닷컴	대규모 소매업	156.7	8
미국	COIN US	Coinbase Global Inc	금융거래소&데이터	127.0	1
일본	6506 JP	아스카와 전기	산업용 기계	115.3	6
일본	6645 JP	오므론	전자 부품	108.4	5
미국	SQ US	블록	거래 및 결제 처리 서비스	104.1	3
미국	ROKU US	로쿠	영화 & 엔터테인먼트	97.4	1

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 보유금액은 'ETF 운용자산편입 종목 비중'으로 계산. A주, B주의 경우 A주로 통일. ex.알파벳(GOOG US, GOOG US) →GOOG US로 통일하여 집계
ADR을 보유한 경우 본국 소재국으로 통일. ex.ABB(ABB SW, ABB US) →ABB SW로 통일하여 집계, 바이두(BIDU US, 09888 HK) → 09888 HK로 통일하여 집계

2) AI 관련 ETF 접근하기

위 '글로벌 주요국 상장 AI 관련 ETF'에서 살펴본 것처럼 이미 주요국 ETF 시장에는 AI 관련 ETF들이 다수 상장되어 있다. 다만 Global X Robotics & Artificial Intelligence ETF(**BOTZ US**), ARK Next Generation Internet ETF(**ARKW US**), iShares Automation & Robotics UCITS ETF(**RBTX LN**) 등과 같은 운용자산 상위 ETF를 기준으로만 본다면 로봇, 인터넷, 클라우드, 자동화 등 관련 산업, 기반 산업 비중이 더욱 높은 편이다.

앞에서도 밝힌 것처럼 본 자료에서는 보다 순도 높은 접근을 위해 1) AI 테마 자체에만 집중한 접근과, 4) AI 관련 핵심 종목에 대한 높은 편입 비중을 보이는 ETF를 위주로 접근한다. 미국 상장 AI 위주의 ETF에는 Global X Artificial Intelligence & Technology ETF(**AIQ US**), Robo Global Artificial Intelligence ETF(**THNQ US**), Listed Funds Trust TrueMark Technology AI & Deep Learning ETF(**LRNZ US**) 등이 상장되어 있다.

이 중 Global X Artificial Intelligence & Technology ETF(**AIQ US**)를 제외한 나머지 ETF는 운용자산이 5천만 달러를 하회하고 있는만큼, 운용자산의 적정성, 테마 접근성 등을 감안해 핵심 ETF로는 Global X Artificial Intelligence & Technology ETF(**AIQ US**)를 선정했다.

순수 AI 관련 주요 ETF 비교 (*순자산총액: 한국 1억 원)

티커	AIQ US	THNQ US	LRNZ US
ETF명	Global X Artificial Intelligence & Technology ETF	ROBO Global Artificial Intelligence ETF	TrueShares Technology, AI & Deep Learning ETF
개요	인공지능(AI) 산업 테마	인공지능(AI) 산업 테마	인공지능(AI), 딥러닝 산업 테마
# 기본 데이터			
*순자산총액	1,661	321	204
상장일	2018.05.16	2020.05.11	2020.03.02
운용보수 (%)	0.6800	0.6800	0.6800
# 수익률 S&P 500			
1M (%)	2.1	0.7	0.2
3M (%)	5.5	19.6	18.5
YTD (%)	1.5	17.3	13.6
1Y (%)	(14.0)	(15.5)	(32.7)
변동성(200일)	25.1	37.5	49.3
# 편입 상위			
	4.1%, 메타 플랫폼스	2.0%, 아리스타 네트워크	7.2%, 엔비디아
	4.1%, 엔비디아	2.0%, 알파벳	7.0%, Snowflake Inc
	3.8%, 테슬라	2.0%, 엔비디아	6.7%, AMD
	3.5%, 세일즈포스	2.0%, AMD	6.3%, Samsara Inc
	3.3%, 마이크로소프트	2.0%, 마이크로소프트	6.2%, Mobileye Global Inc
# 시총 비중 (%)			
대형	90.6%	76.0%	66.5%
중형	8.2%	18.2%	21.8%
소형	0.6%	5.8%	11.7%
# 편입 종목수			
	91	72	24
# 섹터/산업 비중			
	56.1%, 기술	65.5%, 기술	75.8%, 기술
	32.4%, 커뮤니케이션	22.6%, 커뮤니케이션	9.0%, 필수소비재
	5.9%, 산업재	8.0%, 필수소비재	8.8%, 커뮤니케이션

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 2023.3.24 종가 기준

이 외에 핵심 기반 산업으로는 반도체 산업을 선정했으며, 반도체 산업 내 핵심 ETF로는 VanEck Semiconductor ETF(SMH US)를 선정했다. 반도체 산업은 AI 학습 및 연산, 대규모 데이터 처리에 필요한 핵심 기반 산업으로, 'AI 산업=반도체 산업'이라고 해도 과언이 아닐 정도로 반도체 의존도가 높다. 물론 클라우드나 사이버보안, 자율주행, 로봇 등도 AI와 연계성이 매우 높은 산업으로 운용자산 상위 ETF의 경우 관련 산업을 포괄하는 경우가 대부분이나, 보고서 취지에 보다 적합하도록 AI 순도가 높은 ETF 및 핵심 기반 산업으로는 반도체 산업만을 선정했다.

핵심 ETF로는 VanEck Semiconductor ETF(SMH US)를 선정한 것은 편입 종목에 대한 고려를 반영했다. 운용자산 기준 2위 ETF인 iShares Semiconductor ETF(SOXX US) 대비 엔비디아(NVDA US), AMD(AMD US), 브로드컴(AVGO US) 등 핵심 종목 외에도 TSMC(TSM US/2330 TT)와 같은 파운드리 업체를 더 높은 비중으로 포함하고 있다. '미국 반도체+파운드리' 구성에서 '보다 고른 분산'을 목표로 한다면 개별 기업 편입 비중이 상대적으로 고른 iShares Semiconductor ETF(SOXX US)와 함께 홍콩에 상장된 대만 반도체 ETF(03076 HK)나 4월 상장 예정인 한국 상장 TSMC 밸류체인 ETF (453950 KS)를 고려할 수 있다. 반도체 ETF 관련 한국 상장 내 관련성이 높은 ETF는 아래와 같다.

VanEck Semiconductor ETF(SMH US) - KODEX 미국반도체MV(390390 KS)

iShares Semiconductor ETF(SOXX US) - TIGER 미국필라델피아반도체나스닥(381180 KS)

반도체 관련 주요 ETF 비교 (*순자산총액: 한국 1억 원, 파란색 음영: 순자산총액 150억 원 미달)

티커		SMH US	SOXX US	03076 HK
ETF명		VanEck Semiconductor ETF	iShares Semiconductor ETF	Fubon ICE FactSet Taiwan Core Semiconductor Index ETF
개요		MVIS 미국 상장 반도체 기업	ICE 미국 상장 반도체 기업	ICE FactSet 대만 반도체 산업
# 기본 데이터				
*순자산총액		101,246	97,906	21
상장일		2011.12.21	2001.07.13	2022.01.13
운용보수 (%)		0.3500	0.3500	-
# 수익률	S&P 500			
1M (%)	2.1	6.4	5.1	6.9
3M (%)	5.5	27.4	25.4	26.4
YTD (%)	1.5	25.5	23.4	28.7
1Y (%)	(14.0)	(7.5)	(11.5)	(16.9)
변동성(200일)	25.1	39.1	40.1	34.4
# 편입 상위				
		13.6%, 엔비디아	8.8%, 엔비디아	23.0%, 타이완 반도체 제조
		11.9%, 타이완 반도체 제조	8.0%, 브로드컴	7.1%, 노바 테크 마이크로일렉트로닉스
		5.5%, AMD	7.9%, 텍사스 인스트루먼트	6.2%, 리얼텍 반도체
		5.2%, ASML 홀딩	7.3%, AMD	5.8%, ASE Technology Holding
		5.0%, 퀄컴	5.8%, 퀄컴	5.5%, 렌파 과기
# 시총 비중 (%)				
대형		99.5%	97.5%	40.9%
중형		0.5%	2.5%	46.5%
소형		-	-	12.6%
# 편입 종목수		27	35	32
# 섹터/산업 비중		99.5%, 기술	99.8%, 기술	90.7%, 기술

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 2023.3.24 종가 기준

추가적으로 한국 상장 ETF 내 핵심 종목 보유 비중이 높은 ETF도 고려해볼 수 있다. 간접적이긴 하지만 1) 여전히 핵심 종목을 통해 AI 산업에 대한 순도 높은 접근을 제공한다는 점, 2) 개인연금 등 장기 투자 용 계좌에서 활용할 수 있다는 점에서 의의가 있다. TIGER 미국테크TOP10 INDXX(381170 KS)는 미국 테크주 Top 10종목을 편입하고 있으며, 환노출형으로 상장되어 있다. 이 외에 KODEX 미국FANG플러스(H)(314250 KS) 역시 미국 기술주를 편입하고 있으며, 환헤지형으로 상장되어 있다.

한편 국내 상장 ETF 중 글로벌 인공지능 산업에 직접 투자하는 ETF로는 ARIRANG 글로벌인공지능산업 MV(438210 KS)가 있으며, 엔비디아(NVDA US), 세일즈 포스(CRM US) 등을 편입하고 있다. 다만 순자산총액이 28억 원으로 극도로 낮은 점은 유의할 부분이다.

핵심 종목 보유 비중 상위 ETF 및 글로벌 AI 관련 ETF 비교 (*순자산총액: 한국 1억 원, 파란색 음영: 순자산총액 150억 원 미만)

티커		381170 KS	314250 KS	438210 KS
ETF명		TIGER 미국테크TOP10 INDXX	KODEX 미국FANG플러스(H)	ARIRANG 글로벌인공지능산업MV
개요		미국 테크주 TOP 10	미국 FANG+/환헤지	글로벌 인공지능 산업 테마 (능동형 정보, 머신비전 등)
# 기본 데이터				
*순자산총액		15,572	4,299	28
상장일		2021.04.09	2019.01.10	2022.08.17
운용보수 (%)		0.5700	0.6100	0.6500
# 수익률 S&P 500				
1M (%)	2.1	8.5	9.3	(0.6)
3M (%)	5.5	27.9	31.2	18.6
YTD (%)	1.5	33.3	37.4	21.9
1Y (%)	(14.0)	(14.0)	(14.4)	-
변동성(200일)	25.1	29.6	37.3	30.0
# 편입 상위				
		19.7%, 마이크로소프트	10.1%, 엔비디아	4.5%, 엔비디아
		19.2%, 알파벳	10.1%, AMD	4.1%, 세일즈포스
		19.2%, 애플	9.9%, 알파벳	3.1%, 마이크로소프트
		13.9%, 아마존닷컴	9.6%, 테슬라	3.0%, 스플링크
		6.4%, 엔비디아	9.5%, 넷플릭스	2.8%, SAP
# 시총 비중 (%)				
대형		100.0%	100.0%	55.3%
중형		-	-	30.4%
소형		-	-	14.2%
# 편입 종목수		11	13	109
# 섹터/산업 비중				
		51.6%, 기술	1304.2%, 기술	68.2%, 기술
		42.3%, 커뮤니케이션	1030.3%, 커뮤니케이션	7.9%, 커뮤니케이션
		5.4%, 자유소비재	260.5%, 자유소비재	1.1%, 필수소비재

자료: Bloomberg, 미래에셋증권 디지털리서치팀, 삼성자산운용(KODEX), 한화자산운용(ARIRANG)

주: 2023.3.24 종가 기준

주요 ETF 상세

[AI 관련] Global X Artificial Intelligence & Technology ETF (AIQ US)

AI 기술 테마(AIQ US) 주요 사항(최소 매매 단위: 1주)

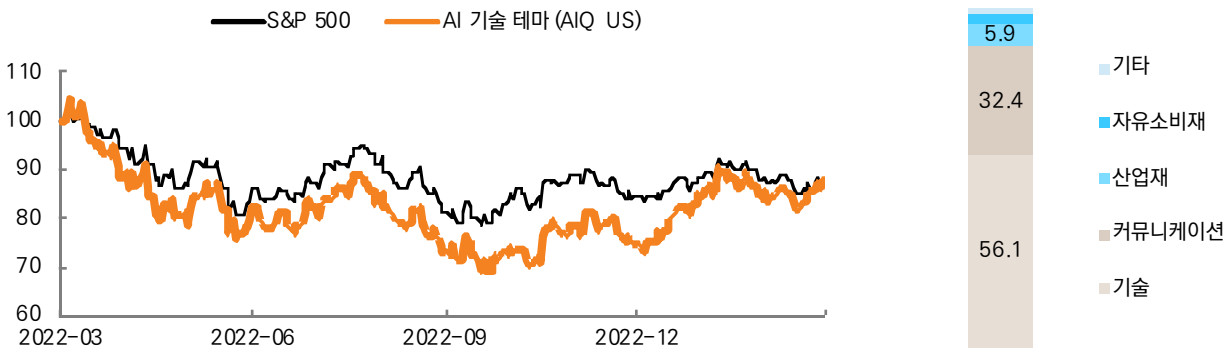
# 개요					
ETF명	Global X Artificial Intelligence & Technology ETF	증가(현지통화)	23.44	운용보수(%)	0.68
ETF 개요	AI 기술 테마	*순자산총액	127	액티브 여부	N
운용사	Global X	유통주식(100만주)	5.4	보유종목수	91
기초지수	Indxx Artificial Intelligence & Big Data Index	거래량(1만주)	3.0	상장일	2018.05.16
# 인덱스&포인트					
[인덱스 구성] - 제품 및 서비스에서 AI 기술 개발 및 활용의 혜택을 받을 수 있는 기업을 편입 - AI Developers, AlaaS, AI 하드웨어, 퀀텀 컴퓨팅 관련 기업 편입					
[포인트/이슈] - AI에 직접적으로 관련 기업 위주 편입 - 최소 매매 단위는 1주로, 3월 24일 종가 기준 최소 매수 필요 금액은 30,331원 가량*					

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 2023.3.24 종가 기준, 순자산총액: 1백만 달러, 최소 매수 필요 금액은 수수료, 제세금 등 제외 금액, 홍콩 상장 티커는 한국 기준

최근 1년 주가 흐름 vs S&P 500 (2022.3.25=100)

편입 상위 섹터(단위: %)



자료: Bloomberg, 미래에셋증권 디지털리서치팀

자료: Bloomberg, 미래에셋증권 디지털리서치팀

편입 상위 Top 10 개요 및 주요 데이터

비중 (%)	티커	종목명	하부산업	시가총액 (1백만 달러)	EPS 추정치 변동(% 4w)	주가 추이 (최근 1년)	EPS 추이 (최근 10년)
4.09	META US	메타 플랫폼스	인터랙티브 미디어&서비스	534,110	5.8%		
4.07	NVDA US	엔비디아	반도체	661,441	(0.2%)		
3.84	TSLA US	테슬라	자동차 제조업체	602,477	(0.7%)		
3.45	CRM US	세일즈포스	애플리케이션 소프트	190,060	20.0%		
3.29	MSFT US	마이크로소프트	시스템 소프트웨어	2,088,508	(0.0%)		
3.24	AAPL US	애플	기술 하드웨어, 스토리지	2,535,467	(0.2%)		
3.04	GOOGL US	알파벳	인터랙티브 미디어&서비스	1,354,344	0.8%		
3.01	CSCO US	Cisco Systems Inc	통신 장비	206,880	0.1%		
2.91	ADBE US	어도비	애플리케이션 소프트	171,948	0.7%		
2.91	AMZN US	아마존 닷컴	대규모 소매업	1,005,564	(5.7%)		

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: EPS 추정치 변동(% 4w) - EPS 추정치의 최근 4주 변동치

[반도체] VanEck Semiconductor ETF (SMH US)

MVIS 미국 상장 반도체 기업(SMH US) 주요 사항(최소 매매 단위: 1주)

# 개요					
ETF명	VanEck Semiconductor ETF	종가(현지통화)	254.62	운용보수(%)	0.35
ETF 개요	MVIS 미국 상장 반도체 기업	*순자산총액	7,768	액티브 여부	N
운용사	Van Eck Absolute Return Advisers Corp	유통주식(100만주)	30.5	보유종목수	27
기초지수	MVIS® US Listed Semiconductor 25 Index	거래량(1만주)	417.4	상장일	2011.12.21

인덱스&포인트

- [인덱스 구성]** - 미국 상장 기업으로서, 반도체 및 반도체 장비 생산 등 수익의 대부분을 반도체 산업에서 창출하는 기업
- 매출의 50% 이상이 반도체 산업에서 창출되며, 시가총액은 1억 5천만 달러 이상

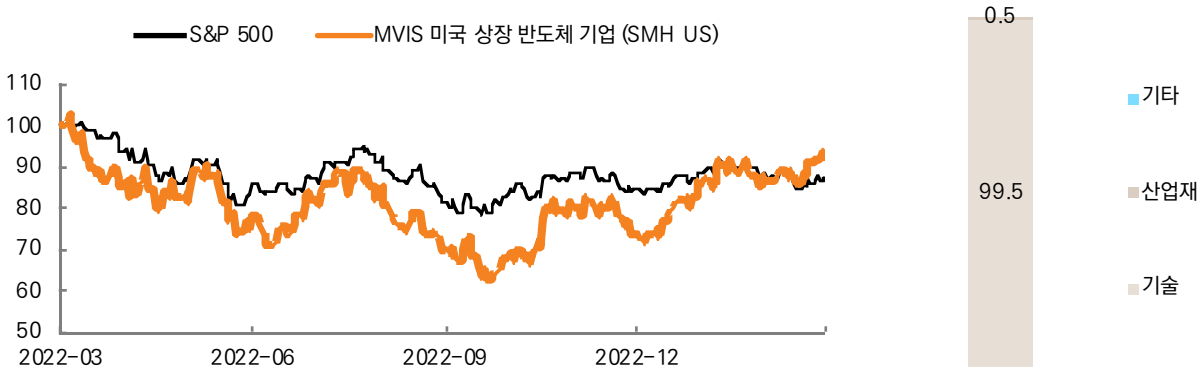
- [포인트/이슈]** - 미국에 상장된 반도체 기업을 편입하고 있으며, 미국 기업 외 대만 파운드리 업체(TSMC)도 12% 가량 편입
- 최소 매매 단위는 1주로, 3월 24일 종가 기준 최소 매수 필요 금액은 329,478원 가량*

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 2023.3.24 종가 기준, 순자산총액: 1백만 달러, 최소 매수 필요 금액은 수수료, 제세금 등 제외 금액, 홍콩 상장 티커는 한국 기준

최근 1년 주가 흐름 vs S&P 500 (2022.3.25=100)

편입 상위 섹터(단위: %)



자료: Bloomberg, 미래에셋증권 디지털리서치팀

자료: Bloomberg, 미래에셋증권 디지털리서치팀

편입 상위 Top 10 개요 및 주요 데이터

비중 (%)	티커	종목명	하부산업	시가총액 (1백만 달러)	EPS 추정치 변동(% , 4w)	주가 추이 (최근 1년)	EPS 추이 (최근 10년)
13.63	NVDA US	엔비디아	반도체	661,441	(0.2)		
11.95	TSM US	TSMC	반도체	481,255	0.1		
5.55	AMD US	AMD	반도체	157,835	(0.1)		
5.24	ASML US	ASML	반도체 재료 및 장비	261,044	(0.1)		
4.98	QCOM US	퀄컴	반도체	139,107	(0.8)		
4.93	TXN US	텍사스 인스트루먼트	반도체	162,877	(0.0)		
4.92	INTC US	인텔	반도체	121,462	(0.7)		
4.87	AVGO US	브로드컴	반도체	265,234	1.7		
4.43	LRCX US	램리서치	반도체 재료 및 장비	67,746	0.2		
4.38	AMAT US	어플라이드 머티어리얼즈	반도체 재료 및 장비	101,017	(0.5)		

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: EPS 추정치 변동(% , 4w) - EPS 추정치의 최근 4주 변동치

[핵심 종목 보유] TIGER 미국테크TOP10 INDXX (381170 KS)

미국 테크 TOP 10(381170 KS) 주요 사항(최소 매매 단위: 1주)

# 개요					
ETF명	TIGER 미국테크 TOP10 INDXX	종가(현지통화)	10,535	운용보수(%)	0.57
ETF 개요	미국 테크 TOP 10	*순자산총액	1,196	액티브 여부	N
운용사	미래에셋자산운용	유통주식(100만주)	146.9	보유종목수	11
기초지수	Indxx US Tech Top10	거래량(1만주)	59.9	상장일	2021.04.09

인덱스&포인트

[인덱스 구성] - FactSet Industry 기준 기술 제품/서비스를 다루거나 독자적 기술에 바탕한 소비자/서비스를 제공하는 'Tech-Oriented' 기업이 속한 섹터 선별, 시가총액 상위 10개 종목으로 구성

[포인트/이슈] - [유사] KODEX 미국FANG플러스(H)(314250 KS)

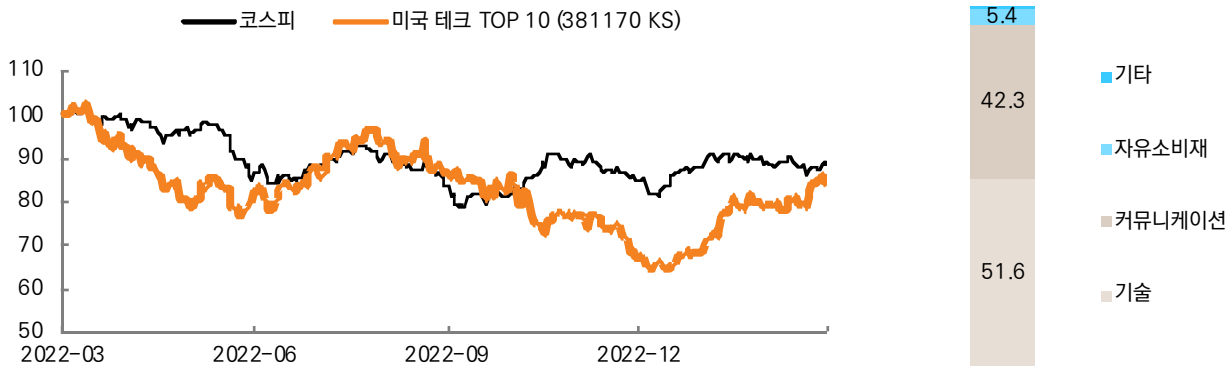
- 최소 매매 단위는 1주로, 3월 24일 종가 기준 최소 매수 필요 금액은 10,535원 가량*

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 2023.3.24 종가 기준, 순자산총액: 1백만 달러, 최소 매수 필요 금액은 수수료, 제세금 등 제외 금액, 홍콩 상장 티커는 한국 기준

최근 1년 주가 흐름 vs S&P 500 (2022.3.25=100)

편입 상위 섹터(단위: %)



자료: Bloomberg, 미래에셋증권 디지털리서치팀

자료: Bloomberg, 미래에셋증권 디지털리서치팀

편입 상위 Top 10 개요 및 주요 데이터

비중 (%)	티커	종목명	하부산업	시가총액 (1백만 달러)	EPS 추정치 변동(% 4w)	주가 추이 (최근 1년)	EPS 추이 (최근 10년)
19.74	MSFT US	마이크로소프트	시스템 소프트웨어	2,088,508	(0.0)		
19.24	GOOGL US	알파벳	인터랙티브 미디어&서비스	1,354,344	0.8		
19.15	AAPL US	애플	기술 하드웨어, 스토리	2,535,467	(0.2)		
13.87	AMZN US	아마존닷컴	대규모 소매업	1,005,564	(5.7)		
6.37	NVDA US	엔비디아	반도체	661,441	(0.2)		
5.88	META US	메타 플랫폼스	인터랙티브 미디어&서비스	534,110	5.8		
5.39	TSLA US	테슬라	자동차 제조업체	602,477	(0.7)		
3.80	AVGO US	브로드컴	반도체	265,234	1.7		
3.28	CSCO US	Cisco Systems Inc	통신 장비	206,880	0.1		
2.56	ADBE US	어도비	애플리케이션 소프트	171,948	0.7		

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: EPS 추정치 변동(% 4w) - EPS 추정치의 최근 4주 변동치

[기타: 중국 AI&로봇] Global X China Robotics and AI ETF (02807 HK)

중국 로봇&AI 산업 테마(02807 HK) 주요 사항(최소 매매 단위: 50주)

# 개요					
ETF명	Global X China Robotics and AI ETF	종가(현지통화)	56.42	운용보수(%)	0.68
ETF 개요	중국 로봇&AI 산업 테마	*순자산총액	28	액티브 여부	N
운용사	Global X HK	유통주식(100만주)	3.5	보유종목수	25
기초지수	FactSet China Robotics and AI Index	거래량(1만주)	0.8	상장일	2020.08.07

인덱스&포인트

[인덱스 구성] - 중국 혹은 홍콩에 본사를 둔 기업으로서 1) 산업 자동화 기계 및 로봇 제조업체, 또는 2) 인공 지능 소프트웨어 관련 기업을 편입

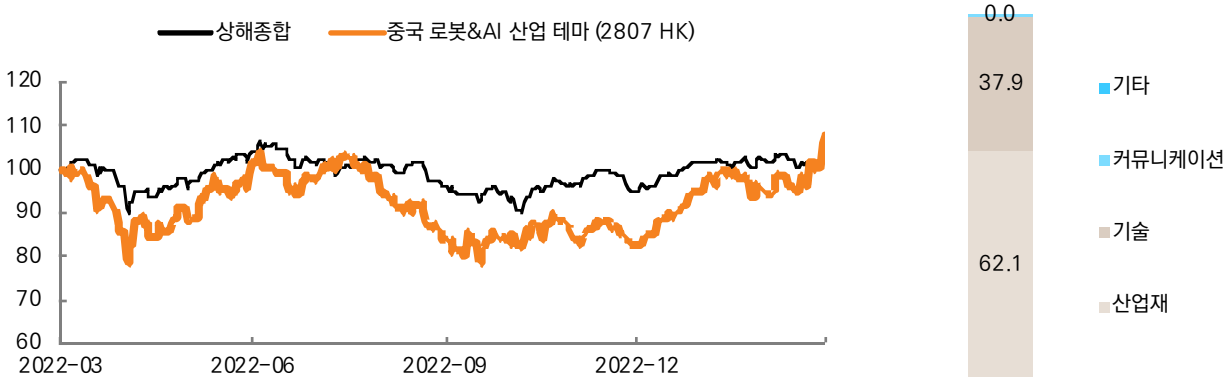
[포인트/이슈] - 중국 본토에 AI 산업 ETF로 E FundSI Artificial Intelligence ETF (159819 CH) 상장
- 최소 매매 단위는 50주로, 3월 24일 종가 기준 최소 매수 필요 금액은 464,768원 가량*

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: 2023.3.24 종가 기준, 순자산총액: 1백만 달러, 최소 매수 필요 금액은 수수료, 제세금 등 제외 금액, 홍콩 상장 티커는 한국 기준

최근 1년 주가 흐름 vs S&P 500 (2022.3.25=100)

편입 상위 섹터(단위: %)



자료: Bloomberg, 미래에셋증권 디지털리서치팀

자료: Bloomberg, 미래에셋증권 디지털리서치팀

편입 상위 Top 10 개요 및 주요 데이터

비중 (%)	티커	종목명	하부산업	시가총액 (1백만 달러)	EPS 추정치 변동(% 4w)	주가 추이 (최근 1년)	EPS 추이 (최근 10년)
12.68	688111 CH	진산오피스소프트웨어	애플리케이션 S/W	25,556	0.0		
10.14	002415 CH	하이크비전	전자 장비&기기	58,817	0.0		
10.01	002230 CH	신비정보	애플리케이션 S/W	19,825	4.2		
8.92	002236 CH	대화기술	전자 장비&기기	9,771	0.0		
7.63	300124 CH	후이촨테크	산업용 기계	26,960	0.0		
7.26	688777 CH	중공테크	전자 장비&기기	7,483	0.4		
6.68	600406 CH	국전남서과기	중전기 장비	26,910	(0.9)		
6.26	300496 CH	창달소프트웨어	애플리케이션 소프트	7,514	(5.5)		
4.78	300450 CH	선도지능장비	산업용 기계	8,997	(7.1)		
4.49	002008 CH	따주레이저	산업용 기계	4,717	(3.4)		

자료: Bloomberg, 미래에셋증권 디지털리서치팀

주: EPS 추정치 변동(% 4w) - EPS 추정치의 최근 4주 변동치

VII. AI의 한계와 미래

인공지능 vs 인간지능

ChatGPT는 질문을 받고 나서 인터넷에서 정보를 찾아다가 내놓는 것이 아니다. 최적의 파라미터로 구현된 모델 알고리즘이 서버에 저장되어 있어 구동될 뿐이다. 또 텍스트로 저장되어 있는 것이 아니다. 인터넷 검색이나 텍스트 참고가 없이, 그냥 알고 있는 것을 생성하는 것이다. ChatGPT의 GPT-3는 인터넷상에서 20조 개의 글자를 학습하고, 질문받은 대로 어울리는 답을 생성한다. 말 그대로 사람의 언어를 흉내낼 수 있는 것이지, 사람의 생각 자체는 흉내낼 수 없다. 사람의 생각을 흉내낸다는 것은 현재 인공지능을 개발하는 입장에서도 기피하는 편이다.

ChatGPT가 엄청나게 많은 인터넷 지식을 학습했음에도 역설적이게도 사람에게는 당연한 “상식”이 부족하다. 메타 산하 AI 연구소 수석연구원인 안 르쿤은 이런 점을 들어 ChatGPT의 역량을 평가절하하기도 했다. 종이를 잡은 두 손 중에서 한 쪽 손을 놓을 때 어떤 일이 발생하는지에 대한 것은, 우리에게서 매우 간단해서 인터넷에는 설명조차 없는 지식이다.

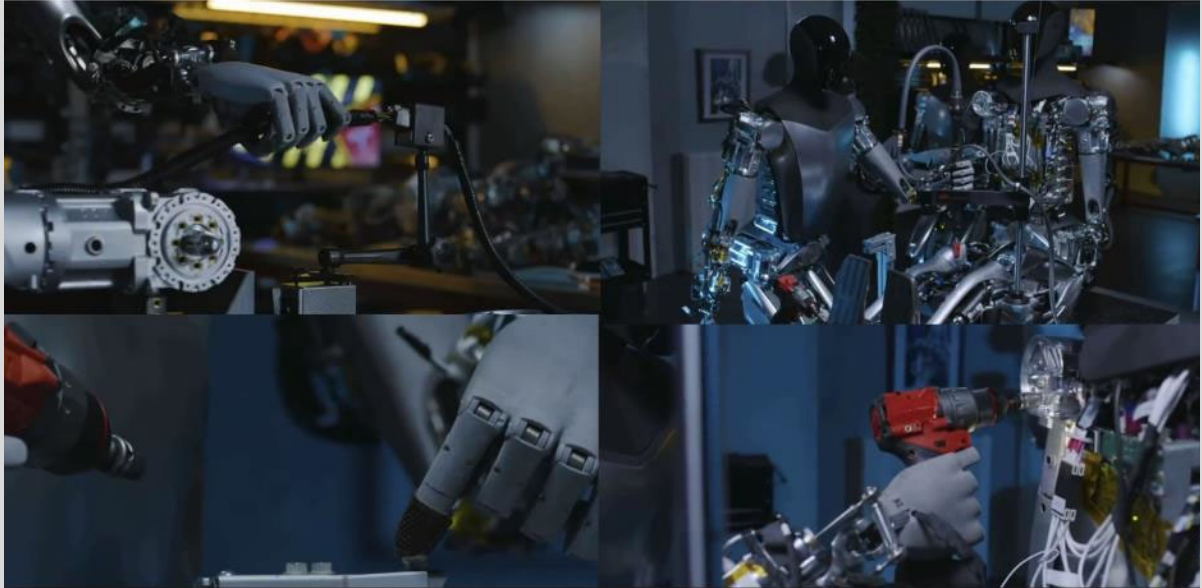
신뢰성 문제도 있다. 이동진 영화평론가는 “현재까지 ChatGPT를 써보면, 우선 사실적인 측면에서 틀리게 너무 많다. 그 지식을 고스란히 쓰는 방법에는 무리가 있을 것 같다”고 언급했다.

하지만, 그의 발언 중 “ChatGPT같은 인공지능이 나의 창의성을 건드릴 수 있는 무언가의 팁을 주고, 거기서 영감을 받아 사고를 뻗어나가면 되는 것이다”고 말한 점이 인상깊었다. 인간이 가진 고유의 창의력을 가지고 생성형 AI를 조수로 활용하는 것이 ChatGPT와 공생하는 가장 좋은 방법인 것이다. 우리 팀 또한 리포트를 작성하는데 있어 AI로부터 많은 도움을 받았다.

그런데, ChatGPT와 같은 생성형 AI가 안드로이드 로봇과 결합한다고 한 번 상상해보자. 카메라와 스피커가 입력 데이터를 받아내는 센서가 되고, 팔, 다리 등이 출력을 위한 액추에이터가 된다. 안드로이드는 이제 대자연을 직접 보고 그 상황에서 학습을 하게 될 것이다. 텍스트만 학습하는 게 아니라 이 세상의 구조를 이해할 수도 있게 될 것이다. 마치 인간처럼 말이다.

실제로 지난 3월 8일, 구글은 사람처럼 보고 듣고 움직이는 인공지능 기반 로봇제어 방식을 공개했다. 언어와 비전 인식 기능을 모두 지원하는 멀티모달 언어모델 “팜-E(PaLM-E)”으로 파라미터 숫자만 5,620억개다. 예를 들어, 팜-E에게 “서랍에서 과자를 꺼내와”와 같은 명령을 하면, 자연어 분석과 함께 컴퓨터 비전을 기반으로 행동 계획을 “생성”한다. 이처럼 멀티모달은 더 많은 맥락과 정보를 포착할 수 있어 일반적인 인공지능(AGI)에 가까운 성능을 보여주기도 한다. 인간 수준은 아니지만 AGI 시대가 성큼 다가오고 있다는 것은 절대 부인할 수 없다.

2023년 3월 초 공개한 테슬라의 휴머노이드 시연 모습
: 2021년 8월 출시계획 발표 후 1년 반 만에 상당한 수준으로 발전



자료: Tesla Investor DAY 2023

제목: 구글이 내놓은 안드로이드 로봇과 생성형 AI의 결합 비전
: 안드로이드 로봇 + 멀티모달 언어모델 "PaLM-E"



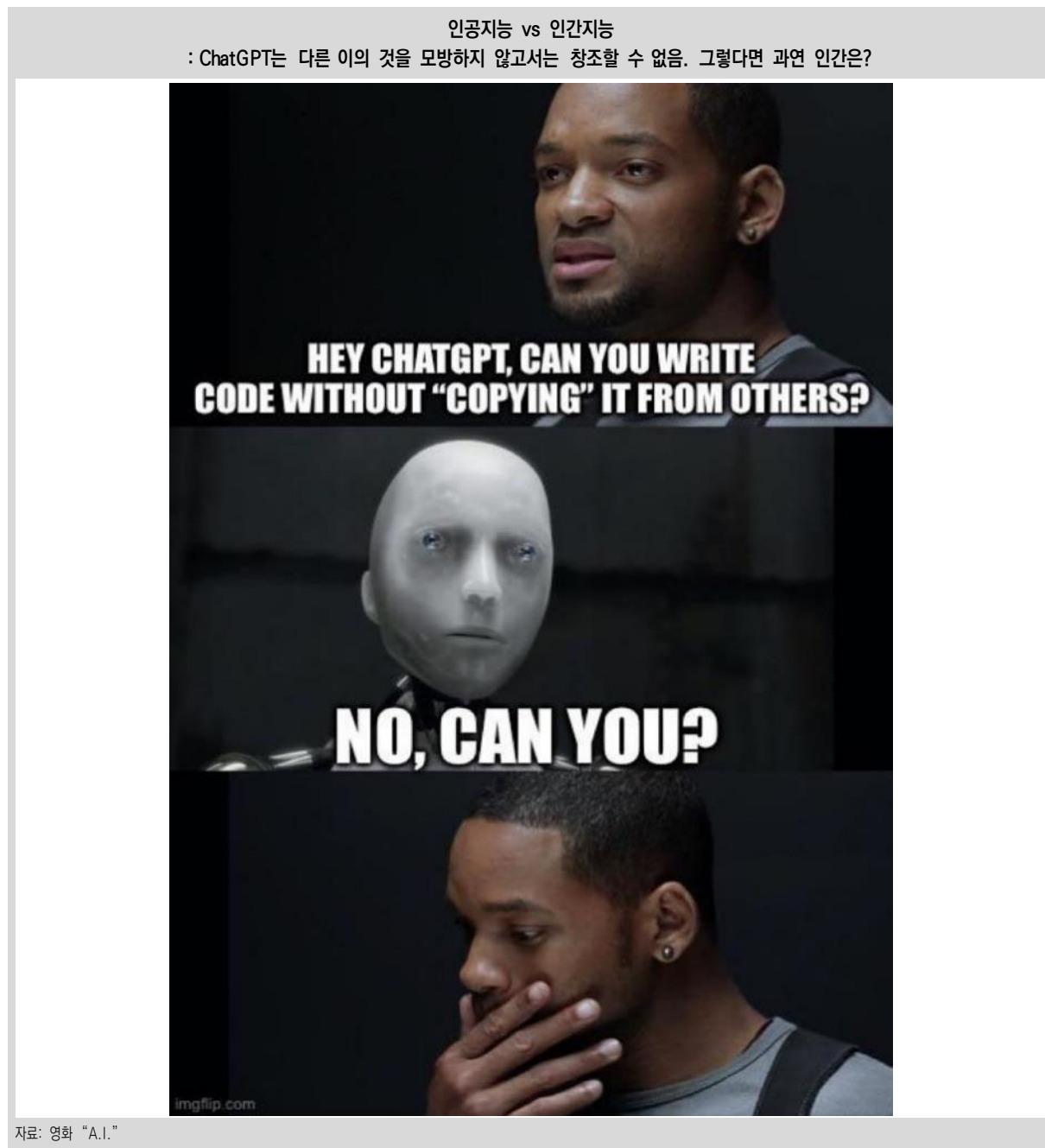
자료: Google

AGI의 '초능력'에 대해 인류는 방법을 모색할 것이다. 먼 미래처럼 들릴지 모르지만 인류의 도약을 위해 인류적 차원의 고민으로서 결국 두뇌와 반도체와의 융합을 시도할 것이다. 현재 우리는 그 시작점에 있고, 일론 머스크의 "뉴럴링크(Neuralink)"가 등장한 배경이다. AI가 번개처럼 빠르게 진화하고 있지만, 인간의 바이오 시스템은 그 발전 속도를 따라가기 어렵다. 앞으로는 더욱 더 '기능적 측면'에서는 그들을 당분간 당해낼 수 없다. 앞으로는 '인간이 인간이기 위한' 고민을 해야 하지 않을까?

인간의 존재론(Ontology)에 관한 소고

인간을 특별하게 만드는 능력은 창의력인가? 딥마인드 CEO 데미스 허사비스는 창의성은 기억에서 나온다고 주장했다. 기존에 알던 것을 모아 새로운 연결고리를 뽑아서 만들어내는 것이 창의성이라는 것이다.

박지성 선수가 영국 맨체스터 유나이티드에서 활약하던 시절, 그에게는 주전 경쟁을 해야 하는 대선배 라이언 킥스가 있었다. 킥스는 원래 측면에서 돌파하는 역습에 능한 윙어였다. 하지만, 나이를 먹으면서 발이 느려지는 노화는 어쩔 수 없었다. 그럼에도 그는 박지성과의 주전 경쟁에서 결코 크게 밀리는 상황을 만들지 않았다. 창의적인 패스를 하는 미드필더로 거듭났기 때문이다. 축구 경험을 연차와 함께 쌓았기 때문에 남들은 잘 보지 못하는 ‘패스 길’이 보였다는 말이다. 이처럼 창의성은 경험적인 지식에서 나온다. 그런 측면에서 창의적인 결과물을 내놓는 것은 시가 우리보다도 더 잘하는 일이다.

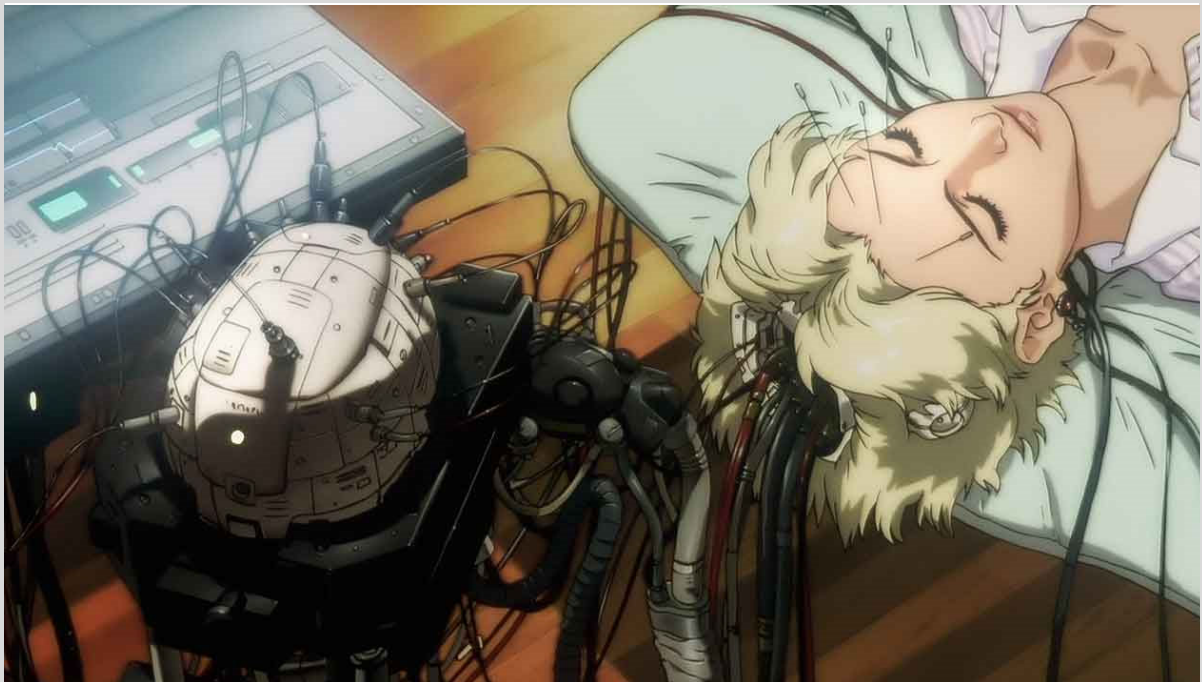


ChatGPT를 사용하다 보면, ‘과연 정말 기계가 치는 답변일까’ 하는 의심조차 들 때가 많다. 질의응답이 마치 “의식이 있는 누군가”가 하는 것처럼 느껴지기 때문이다. 인간의 의식이라는 것은 어떻게 정의해야 할까? 인간의 의식을 설명하는 개념에는 크게 두 갈래로 나뉜다. 첫 번째는 바로 에고 이론(Ego Theory)이고, 다른 하나는 번들 이론(Bundle Theory)이다.

에고 이론에 따르면, 의식은 어떤 분명한 실체가 있는 것이다. 실체가 있기에 인간은 주관적 경험이 가능하다. 흡사 인간이 영혼을 가지고 있다는 개념과 같다. 이 관점을 인공지능에 관철시키면, AI가 의식을 가지려면 영혼을 가지고 있어야 한다는 주장으로 치환된다. 터무니없는 얘기다.

하지만, 번들 이론에 투사하면 이야기가 완전히 달라진다. 번들 이론은 과학자들이 에고 이론보다 훨씬 더 선호하는 이론이다. 번들 이론은 인간의 의식을 여러가지 독립적 프로세스로 연결된 “번들”로 구성으로 본다. 의식이 하나의 실체가 아닌, 서로 연결된 경험들의 집합으로 이루어져 있다는 것이다. 어릴 때부터 쌓아온 기억, 나라는 것을 인지할 수 있는 몸과 얼굴의 외형, 내 목소리와 “깨어 있다”는 감각, 나를 이루는 모든 인간관계 등을 모두 합치면, 그게 내 의식이라는 말이다.

인간의 의식을 만드는 구성물들을 다른 데에 주입하면, “또 다른 나 자신일까?”



자료: 영화 “공각기동대(Ghost in the Shell)”

인간의 의식이 독립적 프로세서의 집합이라는 것은 AI의 멀티모달과 유사한 개념이다. AI에게도 의식이 있다고 생각해볼적이다. AI가 인간의 경험들을 모방하고 그 경험들이 연결되는 방식이라고 한다면 인간처럼 의식을 갖게 된다고 말할 수 있다. 그리고 인공지능이 경험을 모방하는 것은 이미 가능하다. 강화학습(RLHF)을 통해 AI는 경험을 학습하고 그에 맞춰 행동을 최적화한다.

만약, AI와 인간의 의식이 서로 유사한 방식으로 구성되어 있다면, 미래에는 인간과 인공지능 사이의 의식 교류와 공유가 가능할지도 모른다. 즉, AI와 인간의 의식에 대한 연구는 상호보완적이 될 것이다. AI의 의식 연구는 인간의 의식에 대한 이해를 높이는 데 도움이 될 수 있고, 반대로 인간의 의식 연구는 AI의 발전을 뒷받침할 수 있다. 결국 뇌 과학과 인공지능은 함께 발전한다.

대세(Backpropagation)에 반하는 대안 알고리즘 “Forward-Forward”
: 인간 두뇌 구조에 대한 고민으로 새롭게 등장

Forward-Forward Algorithm by Geoffrey Hinton

Backpropagation

학습은 순방향, 오류수정(최적화)는 역방향

Forward-Forward

학습은 순방향, 오류수정(최적화)은 각 레이어에서

자료: Nebuly, 미래에셋증권 디지털리서치팀

레이 커즈와일이 제시한 “특이점(Singularity)”의 개념은 AI와 인간의 의식 사이의 경계가 모호해지는 시점을 말한다. 이 시점에 인간은 자신의 의식을 기계에 업로드하는 등 협업이 강화된다. 인간의 창의성과 감성, 그리고 AI의 처리 능력이 결합되는 형태다. 공상과학과 같은 이야기이지만, 그 미래가 빠르게 가시화되고 있다는 느낌이 드는 요즘이다.

러다이트 운동이 아니라, 대비하고 공생하자

OpenAI의 CEO 샘 알트만은 “언어모델이 사람들이 예상하는 것보다 훨씬 더 멀리 갈 것이라고 생각한다”라고 말했다. 현재의 AI는 좁은 의미의 AI(Narrow AI)로 아직까지는 훈련하는 기계에 가깝다. 하지만, 생성형 AI의 발전으로 AI는 어떠한 전문영역에서도 사람들의 평균적인 업무역량은 달성해낼 것이다. 노동 소득으로 살아가는 인간 입장에서는 간담이 서늘하다. 더군다나 잘 모르는 기술들에 대해서 인간은 우선 경계심을 보이기 마련이다. 아래는 과학과 공학계의 역사적 위인들이 AI 기술에 대해 밝힌 생각들이다.

- (1) **앨런 튜링**: "만약 컴퓨터가 인간으로 속일 수 있다면, 그것은 지능이 있다고 할 만합니다."
- (2) **닉 보스트롬**: "인공지능은 인류가 만들어야 할 마지막 발명입니다."
- (3) **스티븐 호킹**: "효과적인 인공지능을 창조하는 데 성공하면, 이는 우리 문명 역사상 가장 큰 사건이나, 또는 최악의 사건이 될 것입니다. 인공지능에 의해 무한정 도움을 받을지, 무시당하고 주춤하게 될지, 혹은 파괴당할지 우리는 알 수 없습니다"
- (4) **레이 커즈와일**: "10년 안에, AI는 인간지능을 뛰어넘어 급격한 기술적 변화를 일으킬 것입니다."
- (5) **일론 머스크**: "우리가 인공지능을 통해 어떤 매우 어리석은 일을 할지도 모릅니다. 우리는 ‘악마’를 소환하고 있는 겁니다"

위인들이 AI에 대해 밝힌 감정은 복합적이다. 하지만, 단 한가지는 분명하다. 인간 사회를 뿌리부터 뒤 흔들 수 있는 기술임에는 틀림없다는 말이다. 생성형 AI로 가장 빠르게 대체가 될 직무들은 어떠한 자료를 보고 판단하는 직무가 될 가능성이 매우 높다.

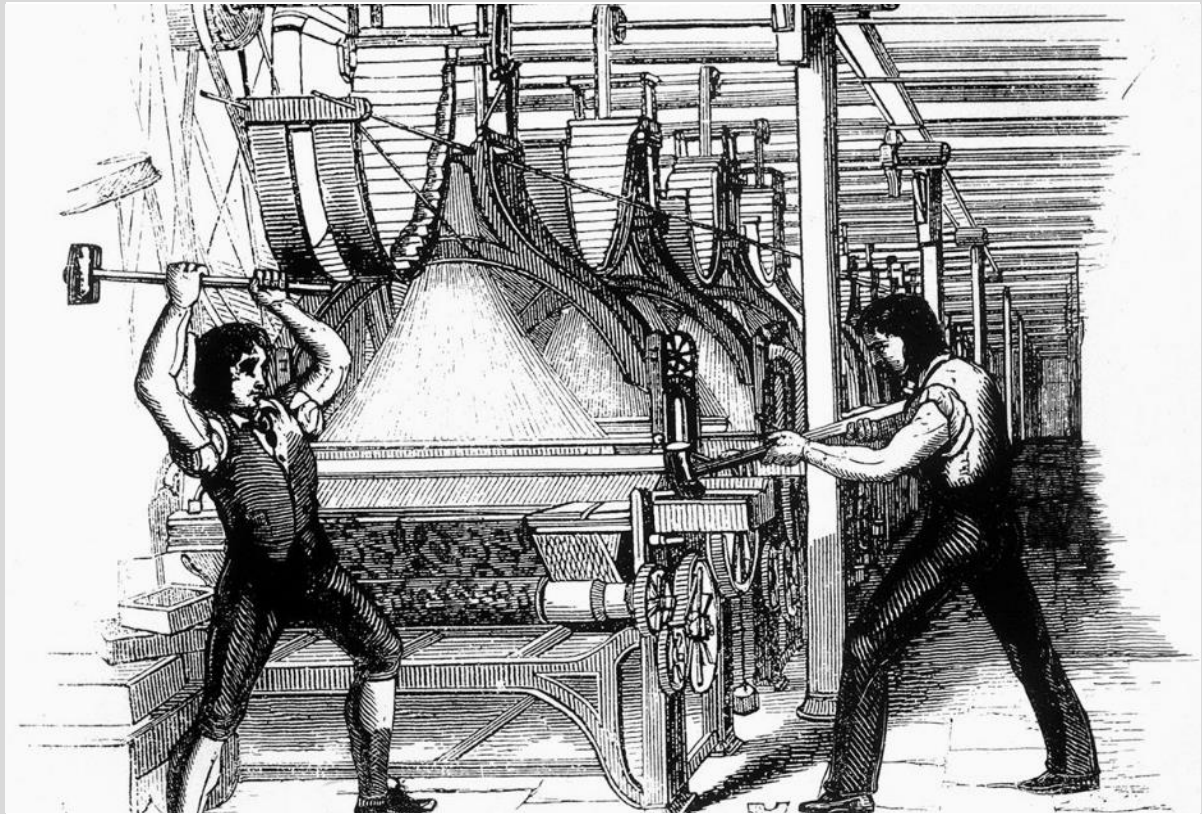
ChatGPT의 출현에 가장 두려움을 느끼고 있는 사람들은 프로그래머와 이 글을 쓰고 있는 우리를 포함한 분석가들일 것이다. 뿐만 아니라, 과학, 의료, 법률, 언론, 금융같은 분야에서는 AI가 월등히 일을 잘 할 수 있을 것이다. 사람간의 상호작용이 주 업무가 되는 간호사보다는, 의사들이 AI 시대에는 대체될 가능성이 높다. 영상의학과와 같은 경우에는, 딥러닝이 가장 잘 하는 것 중 하나인 이미지 분류(Medical Image Segmentation) 문제를 주로 다루기도 하기 때문이다.

물론, 최종적인 판단은 인간이 해야 될 것이기에 그러한 직업들이 완벽하게 대체되기는 아직은 어렵다. 하지만, 결국에 그 시점은 도래한다.

반대로, 섬세한 손 동작이 필요한 생산직은 소위 ‘하이트칼라’보다 더 늦게 대체될 것이다. 특히, 자기만의 어떤 창조물을 만들어내는 직업들은 그 명맥을 더 오래 유지할 것이다.

우리에게는 두 가지 선택지가 있다. 이를 활용하는 ‘공생’과 애써 무시하는 ‘현실 회피’가 있다. 하지만, 산업혁명을 만든 영국에서 일어난 기계파괴 운동 ‘러다이트 운동’에서 우리는 교훈을 얻을 수 있다. 러다이트가 몰락한 것은 노동자 집단이 현실 회피라는 공감대로는 지속할 수 없다는 것을, 그들 스스로 깨달았기 때문일 것이다.

러다이트 운동은 기계의 '자동화'에 실직한 노동자들의 분노



자료: Wikimedia Commons

AI를 활용하는 ‘공생’이라는 정답지를 선택했다면, 앞으로 어떤 일들이 벌어질 것이며 어떻게 대응해야 하는지라는 질문으로 곧바로 치닫게 된다. 그런데 AI의 개발 속도를 감안하면 더욱이 예측이 쉽지 않은 상황이다.

그렇다면, 우리도 AI처럼 주어진 정보를 바탕으로 다음 시퀀스가 뭐가 올지 가능성을 생각해보자. 현재 상황을 고려하면, 자연어 처리방식은 앞으로도 주요한 인터페이스로 자리잡을 것 같다. 우리 뇌의 전극신호가 곧바로 기계로 업로드 되는 세상이 아니라면, 그 전까지는 자연어로 인공지능 모델은 학습되고 또 수행(inference)될 것이기 때문이다.

그래서 우리는 자연어 처리 생성형 AI에 대해서 특히 관심을 갖고 공부해야 한다고 믿는다. 다음 시퀀스로, 범용 AI(AGI) 시대가 도래할 테니 ‘멀티모달’에 대해서도 반드시 지속적으로 관심을 가져야 한다.

AI 혁명을 통해 어떤 일들이 벌어질지는 모르지만, 어쨌든 상자는 열렸다. 우리들 개개인은 이 기술을 잘 활용하여 생산성 향상의 수혜자가 되지 않으면, 인공지능에 대체되는 사람으로 전락하게 될 수밖에 없다. 인류 역사상 최고의 과학자 아인슈타인은 이렇게 말했다.

“지식보다 더 중요한 것이 상상력이다. 하지만 상상력을 강화하는 것은 바로 지식이다”

지식은 어차피 인공지능에게 인류는 상대가 되질 않는다. 그러나 인공지능의 지식을 “코파일럿” 삼아 우리의 창의력을 자극하고 사고를 확장하는 데 사용하면, 인류라는 거창한 범주는 아니더라도 나 자신의 실존적 가치 정도는 좀 더 지켜낼 수 있지 않을까 생각한다.

부록. 딥러닝 주요 개념 설명

(1) 퍼셉트론(perceptron)

1950년대에 처음 도입된 일종의 인공신경망 장치. 여러 입력을 받아 단일 출력을 생성하는 이진 분류기. 퍼셉트론 알고리즘은 입력에 대한 가중치를 학습하는 데 사용할 수 있는 선형분류 알고리즘. 즉 퍼셉트론은 일종의 뉴런이라 할 수 있지만, 딥러닝에서 사용되는 다른 유형의 뉴런만큼 유연하지 않은 단순하고 특정한 유형의 뉴런이라고 할 수 있음.

(2) 뉴런(neuron)

인간 두뇌의 생물학적 뉴런에서 영감을 얻은 신경망 기본 계산 단위. 입력을 받고, 해당 입력에 대한 계산을 수행하고, 네트워크의 다른 뉴런으로 전달되는 출력을 생성. 뉴런은 입력 처리 및 출력생성 방법을 결정하기 위한 다양한 활성화 함수를 통해 비선형성을 가짐. 즉, 딥러닝에서의 뉴런은 더 복잡한 계층과 아키텍처를 구성하기 위한 블록으로 사용.

(3) 레이어(layer)

순방향 신경망에서 숨겨진 영역에서의 한 층 한 층을 레이어라고 부름. 한 레이어에는 각각 여러개의 뉴런들로서 구성되어 있음. 각 뉴런들은 고유의 벡터값을 가지고 있음. 이를 임베딩 벡터라고 부르고, 이 임베딩 벡터는 딥러닝 연산대로 각각 곱하기와 더하기를 한 중간 결과값.

(4) 벡터(vector)

딥러닝에서는 여러 차원을 갖는 데이터를 나타내는데 사용하는 단어. 벡터는 다양한 차원으로 표현할 수 있음. 이 때 각 차원은 해당 데이터의 특정 "속성이나 특징"을 나타내는 방식. 벡터는 숫자 여러 개를 나열한 것이라고 이해하면 됨. 숫자가 3개면 3차원 벡터, 4개면 4차원 벡터. 좀 더 자세히 말하면, 3차원 공간에서의 점은 (x, y, z) 형태의 좌표로 표현되며, 이는 "3차원 벡터"로 나타낼 수 있음. 이 경우, x, y, z 각각은 해당 점의 위치를 설명하는 3개의 차원.

(5) 임베딩 벡터(embedding vector)

자연어 처리를 위해 사람이 쓰는 말을 기계가 이해할 수 있도록 숫자 형태인 벡터로 바꾸는 과정. 우리에게 보이는 않지만 기계는 이해할 수 있는 "정보를 내포"하고 있다는 것에서 "임베딩"이라고 표현. 임베딩 벡터는 "분산된 표현방식(distributed representation)"이라고 불리기도 함.

대량의 텍스트 데이터로부터 단어의 의미적인 관계를 학습하여 임베딩 벡터를 생성. 임베딩 벡터는 자연어 처리 모델에서 입력 데이터로 사용. 예를 들어 문장 분류나 감정 분석과 같은 작업에서, 문장 내 각 단어들을 임베딩 벡터로 변환하여 모델에 입력.

임베딩 벡터는 단어의 의미적인 관계를 반영하여 생성되며, 비슷한 의미의 단어들은 벡터 공간에서 가까운 위치에 배치. 임베딩 벡터는 변수의 서로 다른 값 사이의 의미론적 관계를 파악할 수 있게 해서, 일반적으로 딥러닝의 훈련 프로세스의 일부로 학습.

예를 들어, 사람을 N차원의 벡터값으로 구분한다고 했을 때, 비슷한 사람끼리의 벡터값은 차이가 적다는 점이 핵심. Man과 woman은 상당히 비슷. 언어모델인 기계가 인간처럼 그 의미는 모르겠지만 man과 woman의 임베딩 벡터가 비슷하다는 것이 중요. 더 나아가서 king과 queen의 임베딩 벡터를 각각 비교해본다고 가정하면, 이 둘 역시 임베딩 벡터가 비슷.

그런데 king과 queen의 임베딩 벡터 차이가 man과 woman의 차이와 실제로 같다는 점이 딥러닝의 괄목할만한 점. 즉, king과 queen, man과 woman의 임베딩 벡터 차이값은 성별 벡터값이라고 정의할 수 있음.

(6) 네트워크

신경망의 구조와 기능을 결정. 레이어의 숫자와 유형, 레이어 간의 연결 및 사용된 활성화 함수를 포함하는 신경망의 아키텍처를 의미. 신경망의 구성 방법에 대한 일종의 "청사진".

(7) 모델 알고리즘

성능을 향상시키기 위해 네트워크의 가중치(파라미터)를 업데이트하는 역할. 예측된 출력과 실제 출력 사이의 오차를 최소화하기 위해 "네트워크의 "파라미터를 최적화"하기 위한 수학적 절차. SGD(stochastic gradient descent - 주로 큰 데이터 세트에서 사용), Batch gradient descent, Adam, Adagrad 등 딥러닝에는 다양한 모델 알고리즘이 존재.

(8) 매개변수(parameter)

뉴런과 뉴런 사이를 잇는 연결선에 내포된 일종의 가중치(weight). 따라서, 주어진 데이터(입력값)에 있는 게 아니라, 모델이 내부에 있는 학습 가능한(learnable) 변수. 매개변수는 데이터로부터 스스로 학습되는 것이지 수동으로 설정하지 않음. 매개변수는 머신러닝과 딥러닝 알고리즘의 핵심. 매개변수는 모델 알고리즘의 최적화를 통해 얻게 됨.

(9) 초매개변수(hyper-parameter)

파라미터와 달리 모델 외부에 있으며 데이터로는 값을 추정할 수 없는 구성변수. 초매개변수를 얻기 위해서는, 경험 법칙을 사용하거나, 다른 문제에 사용된 값을 복사하거나, 시행착오를 통해 "수동으로" 최상의 값을 찾을 수 있을 뿐임. 보통 최적의 매개변수를 발견하기 위해 하이퍼파라미터를 조정하는 과정을 거침. 이를 미세조정(finetuning)이라고 부름.

초매개변수를 구성하는 것들에는 여러가지가 있는데, 대표적인 것은 어떤 활성화함수를 쓸 것인가에 대한 것. 또한 경사하강법이나 학습률도 주요한 초매개변수.

(10) 손실 함수(loss function)

최적의 파라미터와 파라미터 최적화를 달성하기 위한 기본 재료. 모델 알고리즘으로 나온 예측값(y^{\wedge})과 실제값(y)의 차이를 나타내는 함수. 평균제곱오차(Mean Squared Error)는 회귀 문제에서의 대표적인 손실 함수. 오차의 제곱의 평균으로 계산. 데이터에 대한 오차를 최소화하도록 모델은 학습을 계속.

모델이 최적화가 됐다는 것은, 데이터의 규칙, 즉 feature(X축, 즉 입력값)와 label(Y축, 즉 결과값)의 관계를 가장 잘 표현한다는 말과 동일. 따라서, 최적의 모델은 '손실 함수'(순간변화율) 값을 최소로 만들 수 있는 모델을 의미함.

(11) 활성화 함수(activation function)

딥러닝 네트워크에서 각 노드에 들어오는 값들을 곧바로 다음 레이어로 전달하지 않고, 먼저 주로 비선형 함수를 통과시킨 후 전달. 이때 사용되는 함수가 활성화 함수. 주로 비선형 함수를 사용하는 이유는 선형 함수를 사용할 시 층을 깊게 하는 의미가 줄어들기 때문. 1차 함수격의 선형 함수로 만들게 되면 층을 아무리 깊게 쌓아올린들 1차 함수 형태를 벗어날 수 없음. 신경망에서 층을 쌓는 혜택을 얻고 싶다면 활성화 함수로는 반드시 비선형 함수를 사용해야 함.

(12) 활성화 함수 종류

a. 시그모이드(sigmoid)

이진분류 문제를 위한 비선형 함수로써 Logistic 함수라고도 불림. 시그모이드의 가장 큰 특징은 우선 함수값이 0과 1 사이. 따라서 중간 값은 0.5를 가짐. 시그모이드는 신경망 초기에는 많이 사용되었지만, 최근에는 아래의 단점들 때문에 잘 사용하지 않음.

가. 기울기 소실(gradient vanishing) 현상이 발생. 미분함수에 대해 $x=0$ 에서 최대값 0.25을 가지고, input값이 일정이상 올라가면 미분값이 거의 0에 수렴하게 됨. 이는 $|x|$ 절대값이 커질수록 기울기 역전파시 미분값이 소실될 가능성이 큼. 경사하강법을 사용해 역전파 수행시 은닉층을 지나면서 기울기를 계속 곱하므로 기울기가 0으로 수렴하게 됨. 따라서 은닉층이 많아질수록 잘 작동하지 않게 됨.

나. 함수값 중심이 0이 아닌 0.5라는 점에서 학습이 느려질 수 있음. 부호가 같은 방향으로 업데이트 될 가능성이 있기 때문. 이러한 과정은 학습을 느리게 만드는 원인이 됨.

b. 하이퍼볼릭탄젠트 (hyperbolic tangent function)

쌍곡선 함수(쌍곡선을 매개변수로 표시)중 하나. 하이퍼볼릭탄젠트 함수는 시그모이드 함수를 변환해서 얻을 수 있음. tanh 함수는 함수값이 -1과 1 사이. 따라서 중앙값이 0이기 때문에, 시그모이드의 최적화 학습이 느려지는 문제를 해결. 또한 중앙값이 0이라서 경사하강법 사용시 시그모이드 함수에서 발생하는 "편향이동"이 발생하지 않음.

시그모이드 함수보다 범위가 넓어서 출력값 변화폭이 더 크다는 것도 장점. tanh 그래프의 미분계수를 보면 최댓값은 시그모이드의 미분계수와 비교하여 4배가 더 큰 1임. 그로 인해 기울기

소실 증상이 더 적은 편. 하지만, 시그모이드보다 낮다 정도일뿐, 미분함수에 대해 일정값 이상 커질시 미분값이 소실되는 기울기 소실 문제는 여전히 남아있음.

c. ReLU (Rectified Linear Unit)

ReLU 함수는 최근 가장 많이 사용되는 활성화 함수. $f(x) = \max(0, x)$ 라는 형태를 가짐. 따라서, $x > 0$ 이면 기울기가 1인 직선이고, $x < 0$ 이면 함수값이 0이 됨. 0 이하의 입력에 대해서는 비활성화(dropout)시킬 수 있는 sparse activation 능력 또한 가지고 있음. 예를 들어 강아지와 고양이 이미지를 처리하는 모델에는 강아지와 고양이의 귀를 식별하는 데 특화된 뉴런이 포함될 가능성이 매우 높음. 그런데 사람 얼굴 이미지를 처리하는 경우엔 그러한 특정 뉴런이 활성화될 필요가 없는 것. 이때 dropout으로 처리해 학습 속도를 높일 수 있음.

즉 ReLU가 가지는 가장 큰 장점은, 시그모이드, 하이퍼볼릭 탄젠트 함수와 비교할 때 학습이 훨씬 빨라진다는 것. 함수식에서 볼 수 있듯 복잡한 수학을 요구하지 않기 때문에 짧은 시간에 모델을 훈련하고 실행할 수 있음. 당연히 연산에 필요한 비용도 적음.

다만, $x < 0$ 인 값들에 대해서는 기울기가 0이기 때문에 뉴런이 죽을 수 있는 단점. 그럼에도 간단하고 사용이 쉽기 때문에 우선적으로 ReLU를 사용하는 경우가 대다수.

(13) 역전파(Backpropagation)

역전파는 에러들의 역방향 전파의 줄임말. 역전파는 순방향(Feed Forward) 인공신경망을 학습하는데 함께 쓰이는 알고리즘. 다시 말해, 역전파는 출력 노드들로부터 입력 노드들에 이르기까지 에러들을 역방향으로 전파시키는 알고리즘. 이렇게 해야, 순방향 신경망의 모든 가중치에 대한 오류함수 기울기를 계산할 수 있기 때문.

(14) 기울기(gradient)

딥러닝 학습 중에 오류를 최소화하고 모델 성능을 향상시키기 위해 가중치를 조정해야 하는 "방향"을 나타냄. 함수를 미분한다는 것은 해당함수의 "순간 변화율"을 구한다는 뜻. 즉, 딥러닝에서 오류 함수의 기울기는 네트워크 가중치에 대한 오류 함수의 순간 변화율.

좀 더 자세히 말하면, 미분 가능한 N개의 다항함수 f 를 각 축이 가리키는 방향마다 편미분한 것. 수식을 풀어서 살펴보면, X_0 축 이외의 모든 축과 관련된 값은 상수로 취급하고 미분 & X_1 축 이외의 모든 축과 관련된 값은 상수로 취급하고 미분 & $\sim X_{n-1}$ 축으로 편미분한 값으로 구성됨.

가중치에 대한 오차 함수의 기울기는 역전파라는 기술을 사용하여 계산. 기울기 계산에는 두 가지 스텝이 들어감. 순방향 통과와 역방향 통과. 정방향 통과 중에서는, 입력 데이터가 네트워크에 반영되고 현재의 가중치를 가지고 출력을 계산. 그 출력값을 이후 목표 출력값과 비교.

둘의 차이로 이뤄진 손실 함수를 산출. 순방향 통과 중에서는, 각 레이어의 중간 활성화(은닉층의 각 뉴런에 해당하는 가중치합에다가 활성화함수를 적용한 수치로써, 중간 출력값. 각 뉴런의 활성화 수준을 나타내기에 활성화라고 부름) 또한 저장됨. 정방향 전달 동안 중간 활성화는 역전파 알고리즘의 역방향 전달 중에 사용되기 때문에 저장해 놓는 것.

한편, 역방향 통과로서, 네트워크의 출력값 측면에서 손실 함수의 기울기를 계산. 저장해 놓은 중간 활성화 값을 손실 함수 기울기 계산 때 활용함. 네트워크에 걸쳐 거꾸로 역전파하는 과정에서 중간활성화를 가지고 미분연산의 연쇄법칙(chain rule)을 반복 사용하여 기울기를 계산. 오류함수의 기울기가 계산되면, 기울기를 가지고, 확률적 기울기 하강법(SGD)과 같은 최적화 알고리즘을 사용하여 네트워크의 가중치를 업데이트함. 최종적으로, 조정하여 오류함수의 최소값을 향해 네트워크를 이동시켜 성능을 올림.

(15) 기울기 소실(Vanishing Gradient)

기울기 소실은, 손실 함수의 변화율을 나타내는 기울기가 작아지는 것을 의미. 특히 네트워크상에 있는 더 과거의 이전 계층을 통해 역전파될 때 기울기가 매우 작아지는 문제를 겪음. 이로 인해 알고리즘 모델이 파라미터들을 업데이트하기가 어렵고 훈련이 느려지거나 좋지 못한 방향으로 수렴할 수 있음.

특히, 순환신경망(RNN)에서 기울기 소실 문제가 자주 발생. 역전파가 이뤄지며 가중치가 반복적으로 곱해지기 때문에 점점 더 뒤로 전파됨에 따라 기울기가 기하급수적으로 줄어드는 것. 다시 말해, RNN 파라미터를 업데이트하기 위해 역전파를 통해 이전 레이어로 계속 전달 기울기가 전달되는데, gradient가 점점 작아지면 파라미터 업데이트가 무의미해져 실제 학습이 이루어지지 못한다는 것.

기울기 소실 문제는 주로 신경망을 만드는 데 사용되는 활성화 함수의 미분에 의해 발생. RNN에서 기울기 소실 문제를 극복하는 방법에는 여러가지가 있음. 가장 대표적인 것이 LSTM을 사용하는 것이고, 기울기 클리핑과 기울기 체크포인트도 한 방법.

(16) 기울기 폭발

기울기 소실의 반댓말. 큰 배치 크기를 가진 딥러닝 네트워크에서나, 매우 긴 인풋 시퀀스 길이를 갖고 있는 LSTM에서 일어날 수 있음.

(17) 기울기 클리핑(clipping)

기울기의 최대 크기에 대한 임계값을 설정. 기울기가 임계값을 초과하면 임계값으로 축소. 기울기가 너무 커지거나 작아지는 것을 방지. 보통 RNN에서의 네트워크에서 기울기 폭발을 방지하기 위한 테크닉.

(18) 경사하강법(Gradient Descent)

손실 함수의 크기를 최소화시키는 방법에는 여러 가지가 있으며, 이 중에서 가장 널리 쓰이는 것이 경사하강법. 경사하강법은 크게 2가지 한계점을 지님. 첫째, 지역최저치(Local Minimum)에 빠지기 쉽다는 점. 지역최저치에 머무는 이유는, 비볼록 함수(Non-convex function)일 경우 파라미터의 초기 시작 위치에 따라 최적의 값이 달라지기 때문.

볼록 함수일 경우 초기 파라미터 값이 어떻게 설정된다 해도, 경사하강법을 활용하면 최적의 값에 도달할 수 있음. 다만, 현실에서는 대부분 함수가 비볼록 함수형태임. 비볼록 함수는 경사하강법 사용 시 최적의 값이라고 판단한 값이 전역최저치(Global minimum)인지, 지역최저치인지 구분할 수 없다는 한계.

두 번째 문제는 안장점(Saddle point)을 벗어나지 못한다는 점. 안장은 말의 안장처럼 생겼다는 것을 의미. 안장점은 기울기가 0이지만 극값이 아닌 지점을 의미. A-B 사이에서 검은색 점은 최소값(minima)이지만, C-D 사이의 검은색 점은 최대값(maxima)을 보이는 지점. 따라서, 해당 지점은 미분이 0이지만 극값을 가질 수 없음. 경사하강법은 미분값이 0일 경우 더이상 파라미터를 업데이트하지 않기 때문에, 이런 안장점을 벗어나지 못하는 한계.

(19) 최적화(Optimization)

딥러닝 분야에서 최적화는 학습의 목표. 예측값과 실제값 차이를 비교하는 함수인 손실 함수(Loss Function)값을 최소화하는 파라미터 구하는 과정. 학습 데이터 입력은 변경할 수 없기 때문에, 손실 함수값을 최소화하기 위해 가중치(weight)와 편향(bias)을 업데이트.

최적의 가중치 혹은 편향을 찾는 과정에서 첫째로 할 일은, 가로축은 가중치를 세로축은 손실 함수를 의미하는 2차 함수를 그려 봄. 그리고 임의의 가중치를 선정. 운이 좋다면 손실 함수의 최소값에 해당되는 가중치를 단번에 선택할 수도 있겠지만 그렇지 않을 확률이 당연히 높음.

예를 들어, " W_n-1 을 임의의 가중치로 선정"했다고 가정. 최적의 가중치를 찾기 위해서 손실 함수를 비용 함수 W 에 대해 편미분하고, 이를 학습률(Learning rate)과 곱해서 값을 도출. 이후 이 값을 W_n-1 에서 빼줌. 수식으로 정리하자면 $W_n = W_n-1 - \text{학습률} * \text{편미분}(W_n-1)$

이 과정을 손실 함수의 값이 거의 변하지 않을 때까지 가중치를 반복해서 업데이트. 이처럼 손실 함수 그래프에서 값이 가장 낮은 지점인 최소값으로 경사를 타고 하강하는 기법을 경사 하강법이라고 부름.

(20) 학습률(Learning rate)

딥러닝의 최적화를 하기 위한 초매개변수(hyper-parameter). 즉, 어떤 비율만큼 파라미터를 업데이트할지를 결정하기 위한 변수. 0과 1 사이의 값을 가지며, 사용자가 직접 설정함. 학습률이 작을수록 촘촘하게 파라미터를 변화시키며 모델을 학습시킬 수 있음. 보폭을 적게 하면서 모델을 수정한다는 것을 의미.

다만, 그만큼 학습 시 시간이 오래 걸림. 반면, 학습률이 너무 클 경우에는 손실 함수 그래프 자체를 벗어나는 경우가 생길 수 있기 때문에 적절한 학습률을 선택하는 것이 중요.

(21) 시퀀스(sequence)

순서가 있는 일련의 데이터 요소들을 의미. 시퀀스는 여러 가지 데이터 유형으로 구성될 수 있음. 자연어 처리(NLP), 시계열 데이터, 음악, 동영상 등 다양한 분야에서 사용. 예를 들어, 자연어 처리에서 시퀀스는 일반적으로 토큰(token)이라는 단위로 이루어진 텍스트 데이터.

가령, 문장 "I am an AI language model"은 다음과 같은 단어 토큰 시퀀스로 구성하면 이와 같음.

["I", "am", "an", "AI", "language", "model"].

(22) 토큰(Token)

자연어 처리 인공지능에서, 토큰은 텍스트를 처리하기 위해 나누어진 가장 작은 단위. 토큰은 단어, 형태소(조사, 어미 등), 문자, 서브워드(의미를 고려하지 않고 통계적으로 단어를 쪼개는 방법) 단위로 분리할 수 있음. 주로 자연어 처리 모델에 적용되며, 텍스트를 처리하고 이해하는 데 사용.

(23) 토큰화(tokenization)

주어진 문장과 텍스트를 토큰 단위로 분리하는 과정. 수행대상에 따라 단어(어절), 문자(의미가 없는 단순 문자), 서브워드(단어와 문자 중간)의 세 가지 토큰화 방식이 있음. 토큰화를 통해 얻어진 토큰들은 자연어 처리 모델의 입력으로 사용. 토큰화를 통해 얻어진 토큰들은 순서를 유지하면서 모델의 "입력"으로 사용됨. 즉, 토큰화는 입력 데이터로 문장이나 단어를 사용하기 위해 먼저 수행하는 전처리 과정.

(24) 인풋(Input)

알고리즘 모델이 처리하기 위해 받는 입력 데이터. 자연어 처리에서 인풋은 주로 텍스트로 구성됨. 이 텍스트는 토큰화 과정을 거친 후 입력됨. 인풋은 당연히 자연어 처리가 아닌 다른 형태의 알고리즘 모델이라면 이미지나 음성 데이터일 수도 있음.

인풋 데이터는 각각의 도메인에 맞는 전처리 과정을 거친 후, 모델에 입력되어 결과를 생성하거나 예측하는 데 사용. 반면, 인풋은 원시 데이터를 의미. 모델에 입력될 수 있는 형태로 인풋이 변환된 단위가 토큰.

(25) 전처리

자연어 처리에서 전처리 과정은 컴퓨터가 텍스트 데이터를 이해하고 처리할 수 있도록 데이터를 정제하는 데 필요. 전처리 과정에서는 텍스트 데이터의 노이즈 제거, 형태소 분석, 토큰화 등이 이루어짐.

(26) 배치(batch)

딥러닝 학습 중에, 네트워크에서 동시에 처리되는 훈련 데이터의 하위 집합. 메모리 용량상 한 번에 처리할 수 없이 큰 데이터 세트를 처리하려면 배치를 이용해야 함. 네트워크가 가중치를 더 자주 업데이트하도록 해서 훈련 속도를 높일 수 있음. 즉, 전체 데이터 세트를 한 번에 처리하는 대신, 데이터를 더 작은 배치로 나누어서 한 번에 한 배치씩 네트워크에 공급해주는 것.

보통 파이토치 프레임워크에서 제공하는 데이터 로더가 데이터 세트에 있는 인스턴스를 N개의 배치씩으로 뽑아 네트워크에 공급. 배치 크기는 딥러닝 트레이닝에 있어 성능에 영향을 줄 수 있는 중요한 하이퍼파라미터. 배치가 커질수록 훈련 시간은 빨라질 수 있지만 네트워크가 훈련 데이터에 과적합될 리스크도 있음. 반면에 크기가 작으면 훈련 시간은 느려질 수 있지만 과적합을 방지.

Compliance Notice

- 당사는 자료 작성일 현재 해당 회사와 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트(박광남)는 자료작성일 현재 테슬라 302주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트(한종목)는 자료작성일 현재 테슬라 10주, TSL 621주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트(윤재홍)는 자료작성일 현재 마이크로소프트 9주, 엔비디아 30주, TSL 740주, SOXL 1,028주 보유하고 있습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.

본 조사분석자료는 미래에셋증권이 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다.

따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.

본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.