

# 초거대 AI 트렌드 및 이슈 분석

이승민



본 보고서는 ETRI 기술정책연구본부 기본사업인  
“국가 지능화 기술정책 및 표준화 연구”를 통해 작성된 결과물입니다.





# 목 차 C O N T E N T S

## 핵심 요약

I. 연구 개요 .....	1
II. 초거대 AI 등장 배경 .....	3
1. 트랜스포머와 BERT .....	3
2. 트랜스포머의 확장 and 자기지도학습 .....	5
III. 초거대 AI 의미와 한계 .....	8
1. GPT-3 출현과 글로벌 개발 현황 .....	8
2. 초거대 AI 한계와 다양한 시도 .....	11
IV. 초거대 멀티모달 AI .....	16
1. 멀티모달 AI .....	16
2. 언어 시각 트랜스포머 .....	19
V. 이슈와 전망 .....	25
참고문헌 .....	30





## 핵심 요약

 연구 개요

- 최근 1,000억 개 이상의 대규모 매개변수를 가진 딥러닝 기반의 초거대 AI 언어모델이 AI 연구의 새로운 패러다임으로 자리 잡으면서 글로벌 빅테크 기업 간 경쟁을 촉발
  - 지난 2~3년간 인공지능 연구에서 가장 주목할 트렌드는 초거대(Hyperscale) AI 언어모델의 약진과 기대, 그리고 기술주도권을 둘러싼 세계 각국의 규모와 속도 경쟁
  - 딥러닝 이후 시각 분야에서 시작된 기술 혁신은 초거대 AI 등장으로 인해 언어 분야로 빠르게 확산하며, 최근의 트렌드와 기술 이슈를 포함해 인공지능 연구의 미래 방향을 탐색할 필요성이 제기되고 있음
- 본 보고서에서는 초거대 AI 개발을 가능하게 만든 기술적 배경, 초거대 AI 언어모델 의미와 한계, 초거대 멀티모달 AI로 이어지는 인공지능 기술의 주요 트렌드 및 이슈를 분석하고 향후 발전 방향을 살펴보고자 함

 초거대 AI 등장 배경

- 2010년 초반 딥러닝 기반 CNN 모델은 시각 지능 혁신을 가져왔고, 이후 RNN, GAN 등을 거쳐 2010년 중반 이후 개발된 트랜스포머 모델과 자기지도학습 등이 기존 연구 성과와 시너지 효과를 내면서, 최근 초거대 AI 언어모델이 구현될 수 있는 기술적 환경이 조성됨
  - 1,000억 개 이상의 대규모 매개변수를 가진 초거대 AI 모델을 구현하기 위해서는 알고리즘 차원에서 엄청난 양의 데이터를 효과적으로 사용(자기지도학습)하고 처리(트랜스포머)할 수 있어야 함
  - 자연어 처리기술의 일대 혁신을 가져온 트랜스포머 모델은 2020년 이후 점차 시각 분야로까지 활용되면서, 앞으로 트랜스포머 모델의 확장 가능성에 대한 기대가 높은 상황
  - 지난 10여 년 동안 축적된 AI 알고리즘(SW) 혁신의 결과들은 강력한 컴퓨팅 자원(HW)을 바탕으로 이전에 볼 수 없었던 초거대 AI 모델을 탄생시키면서 AI 알고리즘의 기하급수적인 성장 궤적을 만들어냄

 초거대 AI 의미와 한계

- 2020년 OpenAI에서 공개한 언어모델 'GPT-3'은 글을 이해하고 문장을 작성하는 탁월한 능력을 보여줄 뿐만 아니라 프로그램, 그림 등에서 단순 모방이 아닌 창의적인 결과까지 만들어냄
  - GPT-3이 보여준 혁신은 규모 자체뿐 아니라 대용량의 데이터로부터 학습한 지식을 사후학습(fine-tuning) 없이 다양한 응용 작업으로 전이하여 우수한 성능을 보여줬다는 점에서 의미가 큼

- GPT-3 이후, 주요 국가들은 자국의 언어를 기반으로 초거대 AI 언어모델을 개발하고 기술주도권 확보를 위한 규모와 속도 경쟁을 치열하게 전개 중
  - (OpenAI) GPT-3 공개 이후, (MS와 NVIDIA) 5,300억 개의 파라미터를 가진 언어모델 'MT-NLG', (구글) 1조 6,000억 개에 달하는 '스위치 트랜스포머(Switch Transformer)', (베이징인공지능연구원) GPT-3의 10배에 달하는 1조 7,500억 개의 파라미터를 가진 '우다오(Wudao, 悟道) 2.0'을 개발
  - 국내의 경우, 2021년 들어 네이버가 2,040억 개의 매개변수를 가진 '하이퍼클로바(HyperCLOVA)', 카카오가 300억 개(비공개 버전 6,000억 개) 매개변수의 '민달리(minDALL-E)', LG AI연구원은 3,000억 개 매개변수를 가진 '엑사원(EXAONE)'을 공개하며 초거대 AI 개발에 투자를 본격화
- 최근 초거대 AI 언어모델은 알고리즘의 효율성, 편향성, 유해성 등 양적 확장만으로 해결하기 어려운 문제와 한계를 노출
  - 양적 성장이 만든 초거대 AI 모델의 문제와 한계를 극복하기 위해 다양한 연구가 시도되는 가운데, 특히 초거대 AI의 한계를 질적으로 혁신하려는 멀티모달 AI에 주목할 필요

## 📖 초거대 멀티모달 AI

- 멀티모달 AI란 인간처럼 하나의 감각이 아닌 다중감각, 즉 시각, 청각, 후각, 촉각, 미각 등을 통합적으로 이용하여 상황을 이해하고 예측, 판단, 반응을 결정하는 기술
  - 멀티모달 AI는 인식 정확도 향상, 소셜 인터랙션, 상황이해·추론, 질병 예측 등 원천기술연구부터 다양한 응용기술 개발까지 활발한 연구가 진행되고 있음
  - 하지만 여러 감각 정보를 처리하고 관련시킬 수 있는 모델을 학습하고 생성하기 위해서는 표현, 변환, 정렬, 융합, 공동학습 등의 영역에서 기술적 도전 과제가 존재함
- 멀티모달 AI를 구현하는 다양한 접근 가운데, 최근 트랜스포머 구조에 기반을 둔 초거대 AI 언어 모델과 시각 지능을 통합하려는 언어 시각 트랜스포머 연구가 매우 활발하게 진행되고 있음
  - 이것은 언어를 이해하는 능력을 보여준 초거대 AI 언어지능과 이보다 먼저 충분한 성능을 입증한 시각 지능을 결합하여 사람처럼 세상을 이해하려는 시도
  - 즉, 어린아이들이 세상을 '보고', 그것에 대해 '말함'으로써 성장하듯이 지난 10년 동안 터득한 '시각' 지능을 '언어' 지능과 논리적으로 긴밀히 연결해보자는 것
- Open AI에서 공개한 언어 시각 트랜스포머 모델, 'CLIP'과 'DALL-E'는 기존 연구와 다른 접근으로 크게 주목받고 있으며, 중국과 우리나라에서도 유사한 성과를 내기 시작
  - 중국은 대표적인 언어 시각 트랜스포머 연구로 CLIP과 유사한 WenLan, DALL-E와 유사한 M6 등을 비롯하여 관련 기술 개발에 대규모 투자를 진행하고 있음
  - 국내의 경우, 카카오브레인, 네이버, LG AI연구원 등 대기업을 중심으로 초거대 멀티모달 AI 개발을 강화하고 자사 서비스에도 활발히 적용 중

## 이슈와 전망

### □ 초거대화

- 초거대 AI 기술의 경우, 앞으로 효율성과 편향성 극복 그리고 대용량의 데이터로부터 학습한 지식을 확장하기 위한 추가 연구가 필요
  - 최근 트랜스포머에 기반을 둔 초거대 AI 언어모델 연구가 활발히 진행되고 있으나, 다양한 인코더/디코더 구조 개선과 다양한 분야로 지능 확장을 고민해야 함
  - 초거대 사전학습 언어모델의 경우, 트랜스포머 구조, 모델, 학습 방법 등의 관점에서 효율성을 개선하기 위한 연구와 데이터로 인한 편향성을 기술적으로 해결하기 연구가 필요
  - 또한, 초거대 AI 언어모델이 인간과 같은 멀티태스킹 능력을 유지하기 위해서는 사전학습과 달리 지속적으로 사후학습 할 수 있는 방법이 필요함

### □ 멀티모달화

- 멀티모달 AI는 여러 감각 정보를 처리하고 통합하는 모델을 학습하고 생성하는 기술로, 표현, 변환, 정렬, 융합, 공동학습 등 기술적으로 연구 과제가 존재함
  - 서로 다른 소스의 데이터를 사용하여 학습모델을 생성하는 과정에서 어느 시점에서 어떤 방식으로 통합할 것인가에 대한 연구가 핵심
  - 다양한 멀티모달 AI 접근 가운데, 특히 트랜스포머 구조에 기반을 둔 초거대 AI 언어모델과 시각 지능을 통합하려는 연구에서 의미 있는 결과들이 나타나고 있음
  - 앞으로 멀티모달을 통해 인공지능을 확장하는 과정에서 신경망 내부에서 잠재공간(latent space)의 활용 가능성과 한계를 분석하고 어떻게 서로 다른 감각지능을 통합, 확장할 것인가에 대한 고민이 필요함

### □ 자연지능 모사화

- 현 데이터 기반 초거대 AI가 일반 인공지능(AGI)으로 발전하기 위해서는 기술적, 개념적 한계를 넘어서야 하며, 이를 위해 제시되고 있는 대표적인 접근 방법이 자연지능 모사 AI 연구임
  - 매우 초기 연구단계에 있는 자연지능 모사 AI는 주어진 환경에 대한 정확한 감각과 지각을 형성하며, 학습한 기억과 지식에 기반하여 판단과 계획을 수립하고 이에 따라 상황과 목적에 적합한 행동을 수행하는데 필요한 일련의 지능적 과정으로 인간의 지능발달론적 접근을 모방
  - 앞으로 현 인공지능의 장점(상황에 대한 편향성 제어가 가능하고, 기억을 담당하는 메모리의 왜곡이 없으며, 대규모 컴퓨팅 자원과 알고리즘 확장 가능 등)과 자연지능 모사 AI의 장점을 결합한 통합 인공지능 연구로 발전이 예상됨
  - 이러한 접근 방법을 통해 현 딥러닝 기반의 인공지능(초거대 AI, 멀티모달 AI 등)의 한계와 문제를 극복할 것으로 기대



# I 연구 개요

## □ 연구 배경

- 최근 대규모 매개변수를 가진 딥러닝 기반의 언어모델인 초거대 AI가 인공지능 연구의 패러다임을 주도하며 많은 관심과 주목을 받고 있음
  - 지난 2~3년간 인공지능 연구에서 가장 주목할 트렌드는 초거대(Hyperscale) AI 언어모델의 약진과 기대, 그리고 주도권 확보를 위한 세계 각국의 규모와 속도 경쟁
  - ※ 초거대 AI 개발 경쟁을 촉발한 사건은 2020년 OpenAI에서 공개한 'GPT-3'으로, 이것은 이전 버전 'GPT-2'보다 117배나 많은 1,750억 개의 파라미터를 가진 언어모델
  - 인공지능 연구와 확산의 핵심적인 기술 요인은 데이터, 컴퓨팅, 알고리즘이며 이 가운데 데이터와 컴퓨팅의 양적 성장은 이미 기하급수적 성장을 보여왔음
  - ※ 매년 새롭게 생성된 데이터량, 반도체 집적도와 고성능 GPU로 구성된 클라우드 컴퓨팅 능력 등
  - 초거대 AI 출현은 알고리즘 또한 기하급수적 성장의 궤적을 따르고 있음을 보여주는 의미 있는 사건
  - ※ 2012년 딥러닝 출현 이후 지금까지 딥러닝을 구성하는 매개변수의 개수는 3.5개월마다 2배씩 성장하며 18개월마다 2배씩 증가하는 무어의 법칙을 넘어섬
  
- 초거대 AI 출현은 단지 학습데이터와 컴퓨팅 능력, 알고리즘의 복잡도 등 규모의 확장을 통해 일반 인공지능(AGI)로 발전할 가능성을 보여줬다는 점에서 의미가 있음
  - 딥러닝 이후 시각 분야에서 시작된 기술 혁신은 초거대 AI 출현으로 인해 자연어처리 기술은 물론 AI 연구 전반에 큰 기대를 낳고 있음
  - ※ CNN에서 비롯된 시각 지능 혁신, 그리고 RNN, Transformer, BERT, 자기지도학습 등으로 이어진 언어 모델 혁신은 초거대 AI 기술을 중심으로 규모의 확장과 속도 경쟁을 촉발
  - 'GPT-3'에 적용된 초거대 AI 언어모델은 글을 이해하고 문장을 작성하는 데 탁월한 능력을 보여줄 뿐만 아니라 프로그램, 그림 등에서 단순 모방이 아닌 창의적인 결과까지 생성
  - ※ 초거대 AI 등장은 충분히 큰 대용량의 언어모델은 '퓨샷(few-shot)' 학습이 가능하고, 학습한 지식을 다양한 분야로 전이(transfer)할 수 있음을 보여줬다는 점에서 혁신적
  - 주목할 점은 단지 초거대 AI가 보여준 결과보다 등장하기까지의 기술 발전과 초거대 AI 학습 과정, 그리고 언어지능을 넘어 다른 감각 지능으로의 확장을 시도하고 있는 부분
  - 초거대 AI 구현은 딥러닝을 기반으로 하는 AI 알고리즘들의 꾸준한 개선과 문제 해결 노력이 누적된 결과이며, 특히 트랜스포머 모델과 자기지도학습 등은 지금까지 개별 인공지능을 통합적으로 접근할 수 있는 기반이 됨

- 이에 본 보고서에서는 초거대 AI 개발을 가능하게 만든 기술적 배경에서부터 초거대 AI 언어모델, 그리고 초거대 다중감각 AI 등으로 확장하는 인공지능 기술의 주요 트렌드와 이슈 및 향후 발전 방향을 살펴보고자 함

## □ 주요 내용

- 초거대 AI 언어모델이 등장하게 된 기술적 배경
  - 지금까지 언어지능 혁신의 큰 걸림돌로 여겨졌던 학습데이터에 대한 병렬처리와 데이터 레이블링(labeling) 문제를 해결함으로써 언어모델의 성능과 확장성을 크게 높인 ‘트랜스포머’ 모델과 ‘자기지도학습’ 트렌드
  - 특히 트랜스포머 기반 양방향 사전학습 모델 ‘BERT’와 시각 지능으로 트랜스포머 모델의 확장 의미와 통합 모델에 대한 기대
- 초거대 AI 언어모델 등장의 의미와 기대, 한계
  - GPT-3이 보여준 기존 언어모델과 차별된 성능과 기능상<sup>1)</sup>의 우수성, 그리고 이를 기존 비즈니스 모델에 활용하려는 세계적 기업들의 초거대 AI 개발 경쟁
  - 언어모델에서 시작된 초거대 AI가 프로그램, 그림, 작곡 등 다양한 창작 분야에서 보여준 성과와 비즈니스 활용 잠재력
  - 한편, 초거대 AI가 보여준 효율성과 편향성 문제를 비롯해 양적 확장만으로 해결하기 어려운 초거대 AI의 근본적 한계
- 초거대 멀티모달 AI 연구 동향과 발전 가능성
  - 다양한 신호를 동시에 처리하는 인간의 뇌를 모방하여, 다중감각 정보(멀티모달)<sup>2)</sup>를 통합적으로 처리하는 AI 연구의 도전 과제 및 활용 사례
  - 멀티모달 AI 연구 가운데 언어와 시각을 결합하여 인간처럼 세상을 이해하고자 시도하는 언어 시각 트랜스포머 연구 동향
- 초거대 AI에서 비롯된 기술 이슈와 발전 방향
  - 현재 데이터 기반 AI 연구를 주도하는 두 가지 핵심 트렌드 ‘초거대화’와 ‘멀티모달화’를 둘러싼 현실적 문제점과 기술적 도전 과제
  - 초거대 AI가 일반 인공지능으로 발전하기 위해 극복해야 할 기술적, 개념적 한계와 최근 이를 위해 논의가 시작된 ‘자연지능 모사화’ 연구 및 통합 인공지능의 개념

1) 기존 언어모델은 특정 작업을 목표로 개발되어 일반성에 한계를 보이는데, GPT-3은 충분히 학습한 사전학습 모델로부터 다른 분야로 지식을 이전함으로써 유용한 지식 확보에 매우 효과적.

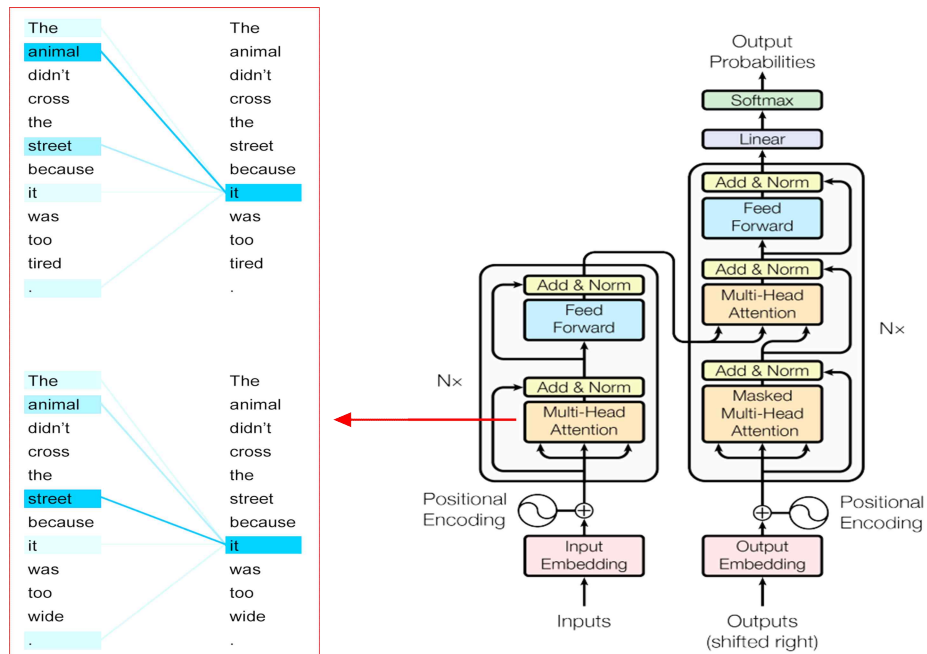
2) 초기 멀티모달 AI 연구는 인식의 정확도를 높이기 위해 서로 독립된 AI 모델의 판단 결과를 단순 결합했다면, 지금은 다양한 감각 정보를 통합하여 한꺼번에 학습 진행.

## II 초거대 AI 등장 배경

### 1 트랜스포머와 BERT

- 2010년 중반 이후, 구글의 ‘트랜스포머(Transformer)’ 모델과 양방향 언어모델 ‘BERT’는 자연어처리 연구의 전환점을 만들고, GPT-3 등 초거대 AI 언어모델이 등장할 수 있는 중요한 기반 기술로 작용
- 트랜스포머 이후, 지금까지 거의 모든 언어모델은 트랜스포머를 기반으로 개발되고 있으며, 점차 시각 분야 등 다른 정보 영역까지 트랜스포머가 활용되는 등 개별 AI 모델이 트랜스포머를 중심으로 통합되는 추세
  - 구글은 2017년 공개한 트랜스포머 모델<sup>3)</sup>을 통해, 언어처리를 위해 데이터를 순차적으로 처리하는 RNN 방식을 사용하지 않고, Attention 함수만을 사용하여 데이터의 병렬처리를 수행함으로써 기존 알고리즘의 성능을 증가하는 실험 결과를 공개함

그림 1 트랜스포머 모델 구조(우)와 인코더의 셀프 어텐션 작동 예(좌)<sup>4)</sup>

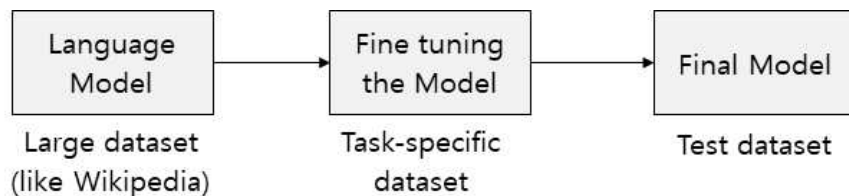


\* 출처: (우) Ashish Vaswani et al.(2017.12.6.),  
(좌) <https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>.

3) Ashish Vaswani et al.(2017.12.6.), Attention Is All You Need, arXiv:1706.03762v5.  
4) 트랜스포머의 인코더 어텐션 작동과정에서 'it'에 대한 최종 표현을 계산할 때 인코더가 어떤 단어에 집중했는지를 시각화한 것으로 트랜스포머는 'it'이 참조하는 두 개의 명사(동물, 거리)를 정확히 식별함.

- 트랜스포머는 입력 데이터를 하나의 벡터로 압축하는 과정에서 정보의 손실이 발생하는 한계를 보완하기 위해 사용해왔던 언어처리 방식(RNN+Attention) 대신, 논문 제목(Attention is All You Need)과 같이 Attention 메커니즘만으로 구현한 모델로 행렬곱을 사용하여 입력 데이터를 동시에 처리할 수 있음
  - ※ CNN(Convolution Neural Network): 미리 설정된 크기의 윈도우를 차례로 움직여 특징을 추출하기 때문에 윈도우 내에 있는 부분에 집중할 수 있는 이미지 처리에 주로 사용되며, 멀리 떨어진 단어들의 관계를 파악하기는 어렵다는 한계가 있음
  - ※ RNN(Recurrent Neural Network): 데이터를 순차적으로 입력받아서 계산하기 때문에 기계번역 등 언어처리에 주로 사용되나, 병렬처리가 어렵고 입력 데이터를 처리하는 벡터(context vector)의 크기가 고정<sup>5)</sup>되어 있어 벡터에 입력 데이터를 압축하는 과정에서 정보의 손실이 발생하는 한계가 있음
  - ※ RNN+Attention: RNN의 고정된 크기의 벡터를 사용하지 않고 입력 단어의 모든 상태 값을 하나씩 가중치를 부여(집중)하는 attention 메커니즘을 추가하여 기존 RNN의 효율성을 높였으나, 여전히 순차적인 계산으로 인해 연산의 속도가 너무 느리다는 RNN의 근본적인 문제를 해결하지는 못함
- 구글은 언어처리를 위한 트랜스포머 기반 양방향 사전학습 모델 'BERT'를 개발하여 언어모델의 새로운 학습 방법을 제시<sup>6)</sup>
- BERT는 양방향 문맥을 고려하여 해당 단어의 의미를 예측하는 방법을 사용하였으며, BERT가 제안된 이후 이는 언어모델의 표준으로 자리 잡으며 기계 독해, 언어분석, 문장분류 등 다양한 태스크에서 기존 방법 대비 월등한 성능을 보임
  - ※ 예를 들어, 'bank deposit', 'river bank'에서 bank의 뜻이 달라지는데, BERT는 bank 양방향 문맥을 이용하여 해당 단어의 의미를 맞추기 위해 제안된 방법으로, 기존 단방향 언어모델 대비 크게 우수한 성능을 보임
  - 또한 트랜스포머는 다수의 언어모델(특히, BERT)에서 언어 사전학습 모델의 기본 틀을 제공함으로써 초거대 AI 언어모델이 등장할 수 있는 기반 환경을 제공

**그림 2** 언어 사전학습 모델



\* 출처: 저자 작성

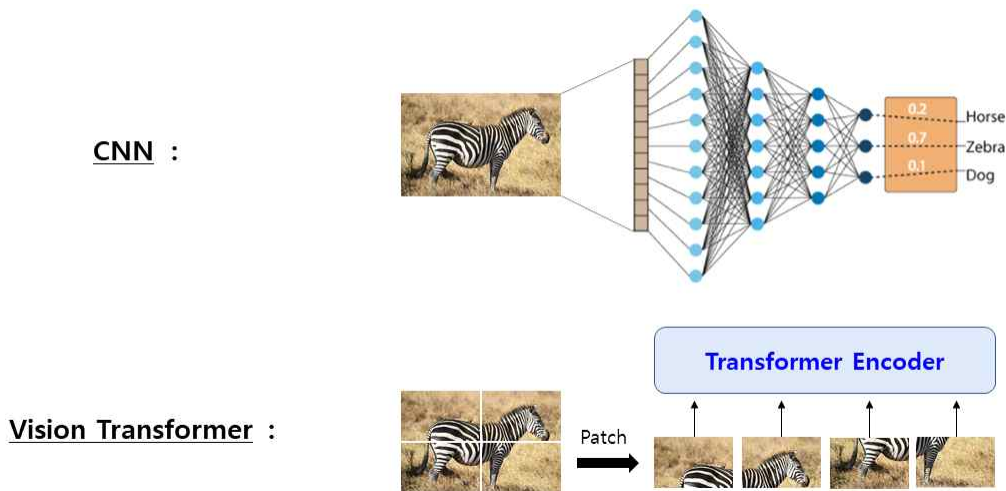
5) 크기가 고정되어 있으면, 입력 문장이 긴 경우에 첫 단어와 끝 단어 간의 관계를 파악하기 어렵고, 중요하지 않은 접속사 등을 포함하여 모든 단어를 함께 학습하기 때문에 결과(기계번역 등)에 좋지 않은 영향을 줄 수 있음.

6) Jacob Devlin et al.(2019.5.24.), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2.

## 2 트랜스포머의 확장 and 자기지도학습

- 언어지능 분야에 일대 혁신을 가져온 트랜스포머 모델은 2020년 이후 점차 시각지능 분야에서도 기존 알고리즘에 비견할 만한 성능을 보이면서, 트랜스포머를 중심으로 개별 지능이 통합되는 움직임이 나타나고 있음
  - 최근 들어 시각 지능 분야에서 CNN(Convolutional Neural Network) 대신 트랜스포머 모델을 적용하려는 연구가 활발히 시도되고 있음
  - 즉, 이미지를 픽셀 단위로 처리한 CNN과 달리, 비전(시각) 트랜스포머는 이미지를 패치 단위로 정의해서 트랜스포머 인코더에 입력하는 방식으로 처리

그림 3 CNN vs. 비전 트랜스포머 비교



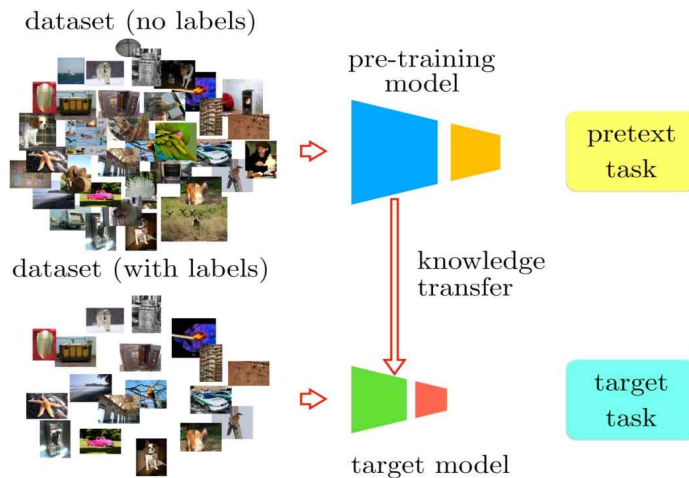
\* 출처: 저자 작성

- 구글의 ViT<sup>7)</sup>는 기존 언어모델에서 활용하던 트랜스포머 구조를 이미지 인식 분야에 처음으로 적용한 기술로, 이미지를 16×16 단어로 표현하여 트랜스포머 인코더에 입력해 정보를 압축함
  - 이후 (Facebook) DeiT, (OpenAI) iGPT, (Google) MLP-Mixer, FNet, (MS) Swin Transformer 등 트랜스포머에 기반을 둔 이미지 인식 연구가 활발히 진행 중
  - 비전 트랜스포머는 언어 트랜스포머의 풍부한 표현력에 비해, 이미지 정보를 단순히 활용해서 발생하는 한계를 보여 이미지 패치 구조 개선 등의 추가 연구가 필요하나, 점차 전통적 CNN 중심의 시각 지능 기술을 대체하며 혁신을 가져올 것으로 예상

7) Alexey Dosovitskiy et al.(2021.6.3.), An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929v2.

- 앞으로 언어와 시각 정보 등 서로 다른 감각 정보를 트랜스포머 모델로 처리할 수 있다면, 인간의 뇌와 유사하게 하나의 AI 모델로 모든 정보를 통합할 수 있다는 의미이며, 이로 인해 일반 인공지능 (AGI) 구현 가능성에 대한 기대가 커지고 있음
- 자기지도학습(Self-Supervised Learning)은 레이블이 없는 대량의 데이터로부터 레이블을 자동 생성해 지도학습하는 방법으로, 이미지 분야에서 처음 사용되기 시작하여 지금은 텍스트, 오디오 등의 분야에서도 활발히 적용 중
- 인공지능은 여러 산업 분야는 물론이고 우리의 일상생활에 깊숙이 스며들고 있는, 연구 현장에서 AI 모델을 생성하기 위한 양질의 데이터를 확보하고 목적에 맞는 정답 레이블을 수작업으로 할당해야 하는 현실적인 어려움이 있음<sup>8)</sup>
- 이러한 문제를 해결하고자 등장한 자기지도학습(self-supervised learning)은 비지도학습(unsupervised learning)의 한 종류로 최근 크게 주목받고 있음
- 자기지도학습은 사람의 지도(supervised) 없이 인공지능 스스로 대상을 인지, 분류하고 의미를 부여하여 학습하는 방법

그림 4 자기지도학습 예



\* 출처: Mehdi Noroozi et al.(2018.5.1.)

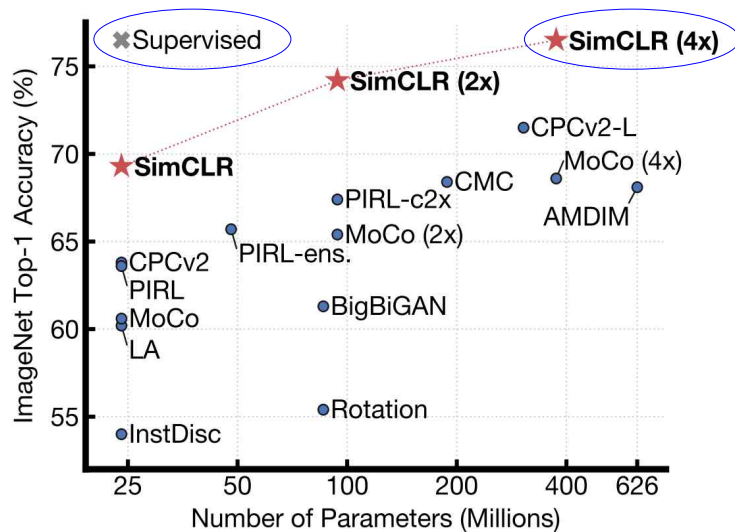
- 예를 들어, 자기지도학습을 이용하여 이미지 분류 모델을 생성하기 위해, [pretext task 단계] 정답 레이블이 없는 대량의 데이터로부터 같은 이미지는 서로 같고 다른 이미지는 다르다는 방식으로 사전학습 모델을 생성하고, [target task 단계] 생성된 사전학습 모델을 기반으로 정답 레이블을 가진 소량의 데이터를 이용하여 전이학습(transfer learning)을 수행하여 목적에 맞는 학습모델을 생성할 수 있음<sup>9)</sup>

8) 많은 시간이 소요될 뿐만 아니라 레이블을 할당하는 과정에서 사람의 편향된 생각이 반영될 수밖에 없는 현실적 문제가 있음.

9) Mehdi Noroozi et al.(2018.5.1.), Boosting Self-Supervised Learning via Knowledge Transfer, arXiv:1805.00385v1.

- 2021년 3월 메타(페이스북)에서 발표한 SEER는 13억 개의 파라미터를 가진 자가지도학습 모델로 ImageNet에서 84.2%의 정확도(top 1%)를 보여주며 기존 자가지도학습 대비 압도적 성능을 달성<sup>10)</sup>
- 대조학습(contrastive learning)은 자가지도학습 방법 가운데 최근 가장 활발히 연구되고 있는 학습 방법으로, 두 개의 데이터 샘플이 하나의 쌍을 이루어 유사성을 평가하도록 특징 표현을 추출하여 모델을 학습함
  - 즉, 하나의 이미지를 여러 방식으로 증강하더라도 이미지 고유의 특성은 변하지 않는다는 가정하에 같은 이미지에서 증강된 데이터 간의 유사성은 높게, 다른 이미지에서 증강된 데이터 간의 유사성은 낮게 평가하도록 모델을 학습
  - 특히, 힌튼 교수가 제안한 대조학습 모델 SimCLR<sup>11)</sup>은 지도학습과 비슷한 수준의 높은 성능을 보이며 큰 주목을 받음

**그림 5** 자가지도학습 결과 비교



\* 출처: Ting Chen et al.(2020.6.1.)

10) Nathan Benaich, Ian Hogarth(2021.10.12.), The State of AI Report.

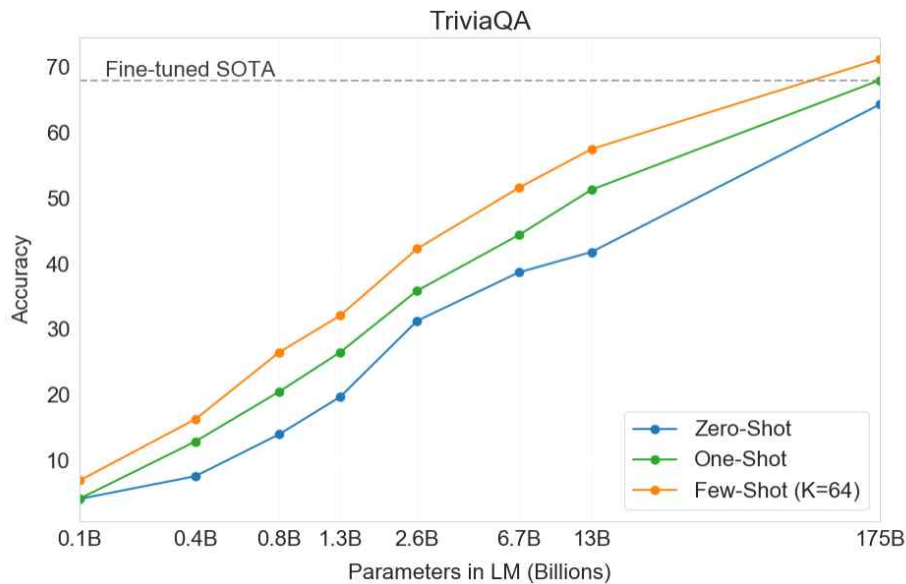
11) Ting Chen et al.(2020.6.1.), A Simple Framework for Contrastive Learning of Visual Representations, arXiv:2002.05709v3.

### Ⅲ 초거대 AI 의미와 한계

#### 1 GPT-3 출현과 글로벌 개발 현황

- GPT-3은 3,000억 개 데이터와 1,750억 개의 파라미터로 학습한 초거대 언어모델로, 사후학습(fine-tuning) 없이 다양한 응용 작업에서 우수한 성능을 보여주기 위해 퓨샷(few-shot) 학습의 가능성을 확인시켜줬다는 점에서 큰 의미가 있음<sup>12)</sup>
  - GPT-3이 보여준 혁신은 규모 자체뿐 아니라 대용량의 데이터로부터 학습한 지식을 다양한 응용 작업으로 전이할 수 있는(전이학습) 효과를 크게 개선한 점
  - 즉, 초거대 언어모델로 매우 강력하게 사전학습을 수행한다면, 사후학습(미세조정; fine-tuning) 없이도 다양한 응용 작업에서 만족할 만한 성능을 얻을 수 있음
  - 논문에서는 퓨샷러닝(few-shot learning)을 통한 GPT-3 모델이 미세조정으로 재학습된 기존 모델의 성능을 넘어선 결과를 보여줌

그림 6 GPT-3의 제로샷/원샷/퓨샷과 Fine-tuned SOTA 성능 비교



\* 출처: Tom B. Brown et al.(2020.6.22.)

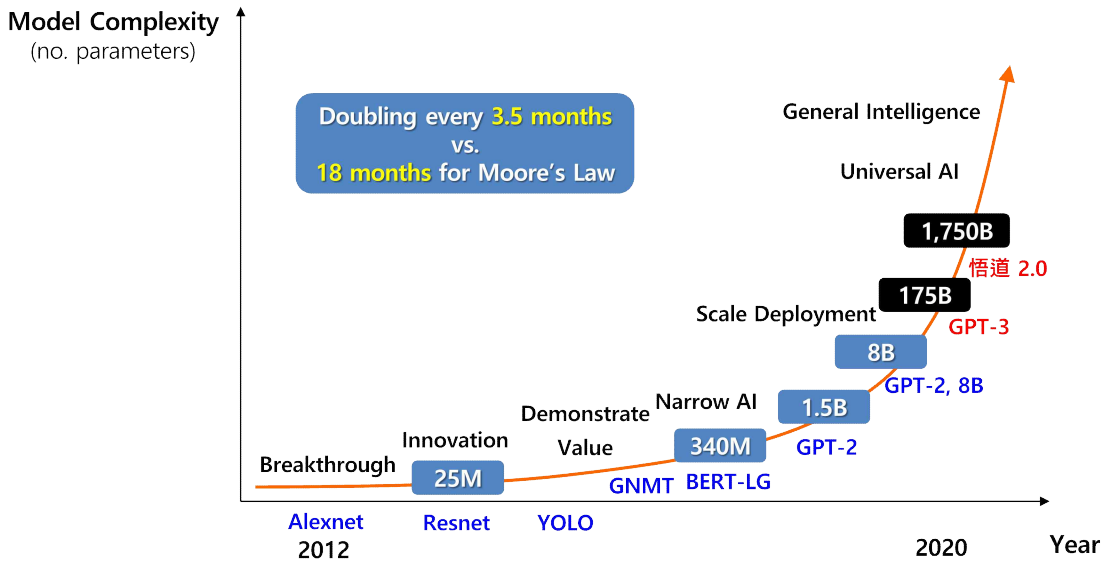
- ※ 딥러닝 등장 이후, 대부분의 응용 분야에서 새로운 모델을 생성하기 위해서는 기존 학습모델 재활용 없이 처음부터 다시 학습해야 했음
- ※ 미세조정(fine-tuning)은 대량의 데이터로 사전학습한 모델을 재활용(전이학습)하는 과정에서 기존 모델을 새로운 데이터에 맞게 추가 학습하는 것이지만, 미세조정 역시 매우 많은 레이블링된 데이터가 필요함

12) Jared Kaplan et al.(2020.7.22.), Language Models are Few-Shot Learners, arXiv:2005.14165v4.

- ※ 퓨샷러닝은 몇 개의 예제만으로 새로운 환경에 빠르게 적응할 수 있는 인간과 비슷하게 학습하는 것으로, 퓨샷러닝을 이용한 GPT-3의 적용 범위가 언어 분야 이외로 대폭 넓어졌고 성능 또한 매우 우수한 결과를 보여줬다는 점에서 충격적
  - 결과적으로 GPT-3이 보여준 초거대 AI 언어모델은 인공지능이 추구하는 일반 인공지능(AGI)의 작동 방식과 구현 가능성을 크게 높여줄 것으로 기대
  - ※ 알고리즘 측면에서 초거대 언어모델 ‘GPT-3’ 구현이 가능한 배경에는 앞서 살펴본 2017년 이후 트랜스포머 모델, 양방향 언어모델(BERT), 그리고 학습데이터 측면의 자기지도학습 모델 등이 결정적
- GPT-3에서 주목할 점은 인간과 비슷한 수준의 언어 구사 능력뿐 아니라 보고, 듣고, 말하고 예전에 보지 못한 결과를 만들어낼 수 있는 다중 지능으로의 확장 가능성
    - GPT-3은 기존 모델 대비 규모가 비약적으로 커진 대형언어모델(LLM: Large Language Models)로 여러 분야의 문장을 이해하고 글을 작성하는 능력에 있어서 과거에는 보지 못했던, 인간과 거의 동등한 수준의 언어능력을 보여줌
    - 또한, 일상의 언어를 간단히 프로그래밍할 수 있고, 이미지나 영상을 이해하고 글로 표현하는 능력까지 보여주며 다양한 분야에서 활용 가능성을 높이고 있음
    - 한편, GPT-3과 같은 대형언어모델(LLM)은 창작 분야나 단백질 생성 등 과학기술 분야까지 활용 범위를 넓혀가고 있음<sup>13)</sup>
    - ※ 단백질 예측에 언어모델이 사용된 배경에는 단백질 서열과 순차적 언어 배치가 유사하기 때문으로, 수천 개의 단백질로부터 수억 개 이상의 단백질 서열에 대한 다음 아미노산을 예측하는 단백질 언어 모델(protein language model)을 만드는 과정과 학습 과정이 비슷함
  - GPT-3은 세계적으로 자국의 언어를 활용한 대형언어모델(LLM) 개발을 시작으로 초거대 AI 연구에 대규모 자본이 투자되며 기술주도권 확보 경쟁을 촉발하고 있음
    - 2020년 GPT-3 이후, MS는 NVIDIA와 협력하여 GPT-3보다 3배 더 큰 5,300억 개의 파라미터를 가진 언어모델 MT-NLG, 구글은 1조 6,000억 개에 달하는 초거대 언어모델 ‘스위치 트랜스포머(Switch Transformer)’, 베이징인공지능연구원(BAAI)은 GPT-3의 10배에 달하는 1조 7,500억 개의 파라미터를 가진 ‘우다오(Wudao, 悟道) 2.0’을 공개하는 등 세계적으로 초거대 AI 개발 경쟁이 치열하게 전개되고 있음
    - ※ GPT-3 이전: RoBERTa(3억5,500만개, 페이스북, 2019.7) → GPT-2(15억개, 오픈AI, 2019.8) → Megatron-LM(83억개, 엔비디아, 2019.11) → T5(110억개, 구글, 2020.2) → Turing-NLG(170억개, MS, 2020.2)
    - 국내의 경우, 2021년 들어 네이버가 2,040억 개의 매개변수를 가진 ‘하이퍼클로바(HyperCLOVA)’, 카카오는 300억 개(비공개 버전 6,000억 개) 매개변수의 ‘민달리(minDALL-E)’, LG AI연구원은 3,000억 개 매개변수를 가진 ‘엑사원(EXAONE)’을 공개하며 초거대 AI 개발에 투자를 본격화

13) Nathan Benaich, Ian Hogarth(2021.10.12.), The State of AI Report.

그림 7 AI 모델(Algorithm)의 크기 변화

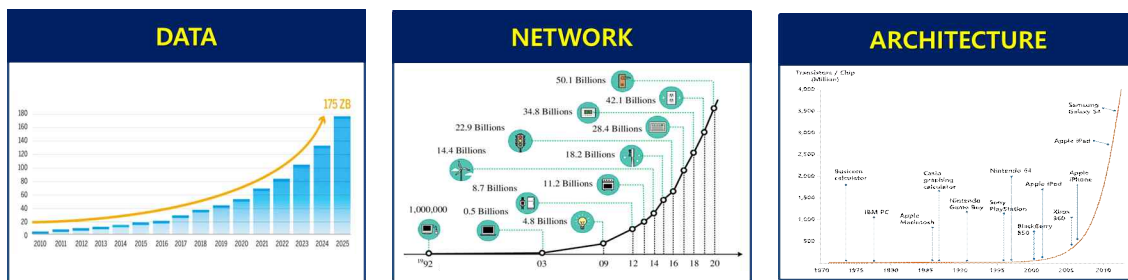


\* 출처: <https://cambrian-ai.com/ai-hardware-harder-than-it-looks/> 참고하여 재작성.

- 글로벌 빅테크 기업들이 경쟁적으로 초거대 AI 개발에 뛰어드는 이유는 현재 자사의 비즈니스 모델 개선과 미래 경쟁력 확보를 위한 장기 전략
  - 구글과 네이버 등은 이미 초거대 AI를 자사 인터넷 검색 서비스에 적용하여 뚜렷한 성능 개선을 확인하며 고객과의 디지털 소통을 강화
  - 2021년 9월 중국 칭화대는 초거대 AI 언어모델 '우다오(悟道) 2.0'으로 만든 디지털 휴먼 '화즈빙(華智冰)'을 공개하며, 작곡, 글쓰기, 대화, 수학 능력 등 다양한 능력을 과시
  - 국내 LG AI연구원은 자사 그룹뿐 아니라 ICT·금융·제조·의료·통신 등 12개 기업과 함께 'Expert AI Alliance'를 발족하여 초거대 AI 상용화와 기술 경쟁력을 강화할 계획
  - 하지만, 초거대 AI 개발을 진행하기 위해서는 엄청난 양의 데이터와 컴퓨팅 자원 등 막대한 재원이 소요되는 만큼 기업 간 격차는 더욱 크게 벌어질 전망
- 매개변수(파라미터) 수를 기준으로 AI 모델의 복잡도를 계산했을 때, 지난 10년 동안 알고리즘의 복잡도는 약 3.5개월마다 2배씩 증가하며 기하급수적으로 발전
  - 인공지능의 하드웨어 분야에서 컴퓨팅 파워를 대표하는 반도체 구조(집적도)뿐만 아니라 소프트웨어 분야에서 알고리즘마저 기하급수적 발전 궤적을 따르고 있음을 확인
  - 지금의 발전 추세라면, 인간의 뇌 수준(약 100조 개의 시냅스)에 달하는 100조 개의 매개변수를 가진 초거대 AI 모델이 2025년 이전에 등장할 수도 있음

- 주목할 점은 데이터(Data), 네트워크(Network), 인공지능(AI: Algorithm/Architecture) 등 D.N.A. 모든 분야에서 기하급수적 양적 성장이 진행되고 있으며, 이로 인해 타 기술 및 산업 분야의 디지털 전환 또한 가속할 것으로 보임
  - 매년 새롭게 생성되는 데이터 증가량(data), 인터넷에 연결되는 기기와 기계의 개수(network), 반도체 집적도(architecture) 등은 이미 기하급수적으로 증가하는 추세를 보여 왔음
  - AI 알고리즘(algorithm)의 발전은 이들 분야와 결합하여 시너지를 내며 경제 분야와 우리 삶에 점점 더 깊이 침투하며 크게 영향을 미칠 것으로 전망
  - 디지털 전환을 이끄는 핵심 분야(D.N.A.)에서 벌어지는 기하급수적인 양적 성장은 기존 기업의 경쟁우위를 완전히 바꿔놓을 것으로 예상

**그림 8** 기하급수적 발전을 보이는 데이터, 네트워크, 반도체 집적도

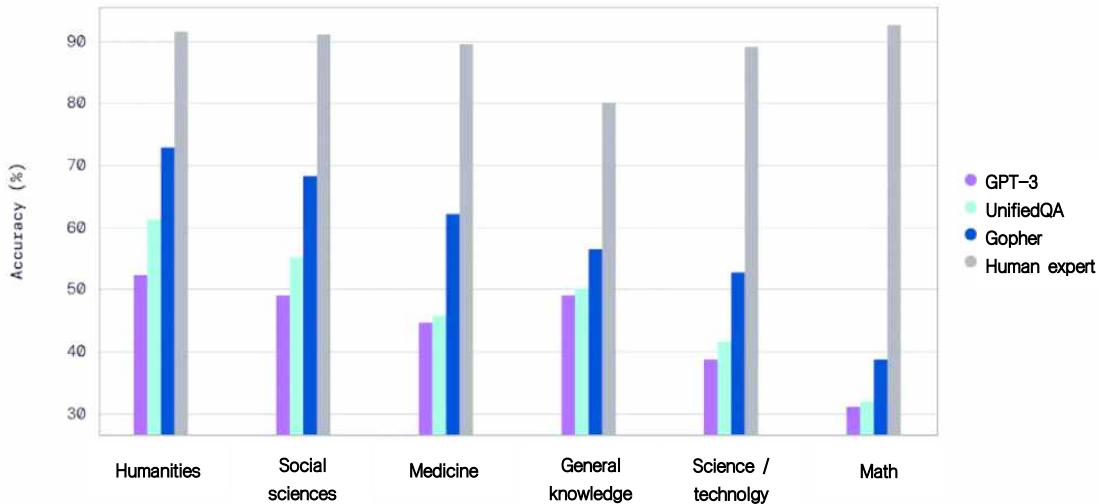


\* 출처: (왼쪽) IDC(2017.4.), (가운데) Abdelmohsen Ali et al.(2015.6.), (오른쪽) Larry Downes & Paul Nunes(2014.1.7.)

## 2 초거대 AI 한계와 다양한 시도

- GPT-3 이후 공개된 초거대 AI 연구 결과를 보면, 초기에는 대체로 학습데이터 규모가 증가할수록 AI 모델의 성능이 좋아지는 경향을 보이고 있으나, 점차 규모의 확장에 따른 성능 개선의 한계와 문제점이 나타나기 시작
- 2021년 12월 딥마인드에서 발표한 초거대 AI 언어모델 ‘고퍼(Gopher)’는 GPT-3(1,750억개) 보다는 크고 MS의 MT-NLG(5,300억 개) 보다는 작은 2,800억 개의 매개변수로 구성되었지만, 효율성과 편향성 등에 있어서 기존 거대 언어모델의 성능을 넘어선 결과를 보임
  - 즉, 다중작업 언어이해(MMLU, Massive Multitask Language Understanding) 과제에서 기존 최신 모델을 능가하고 인간 전문가 능력에 더욱 가까운 결과를 보임

그림 9 MMLU 벤치마크 테스트 비교



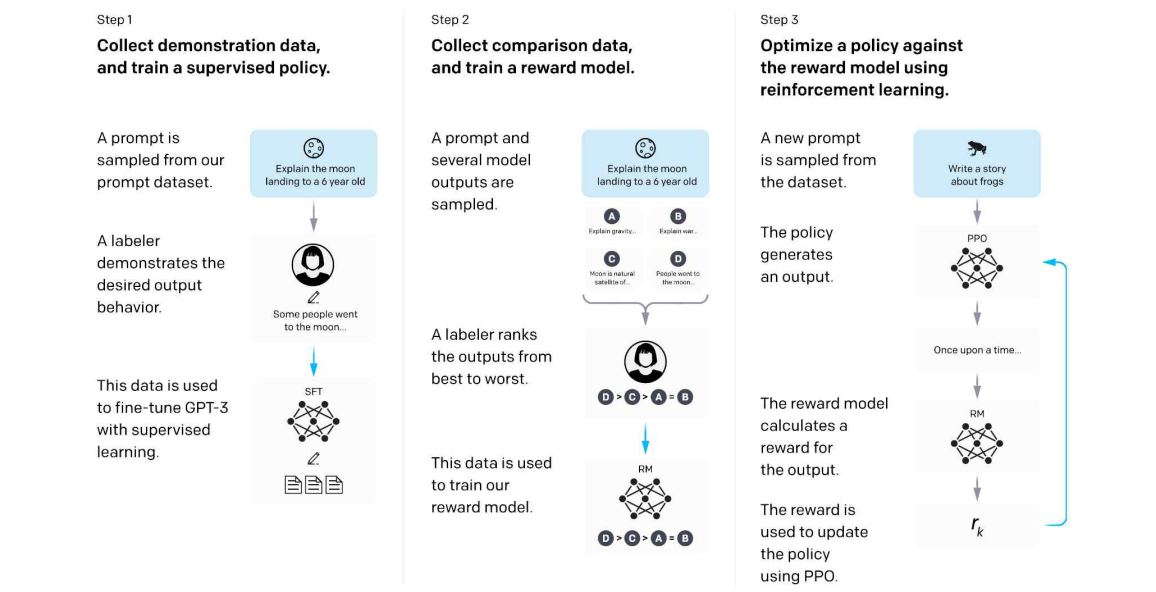
\* 출처: <https://deepmind.com/blog/article/language-modelling-at-scale>.

- 또한, 딥마인드는 70억 개의 매개변수로 구성된 검색 언어모델 ‘RETRO’를 발표하며, 이보다 25배 큰 GPT-3 모델과 같은 성능을 달성할 수 있다고 주장<sup>14)</sup>
  - RETRO는 약 2조 개에 달하는 구문 텍스트를 외부 메모리에 저장하여 활용하는 방식으로 초거대 AI 언어모델 개발 비용을 줄일 수 있어 효율성을 개선할 수 있음
  - 특히, RETRO에서 사용하는 외부 메모리에서 AI 모델이 학습한 내용을 확인함으로써 AI 모델의 편향성과 유해성을 제거할 수 있는 특징이 있음

14) Sebastian Borgeaud et al.(2022.2.7.), Improving language models by retrieving from trillions of tokens, arXiv:2112.04426v3.

- OpenAI는 새로운 버전의 GPT-3 모델 ‘InstructGPT’를 발표<sup>15)</sup>하며, 인종차별이나 여성혐오 텍스트에서부터 각종 편견이 포함된 텍스트와 거짓 정보를 포함한 텍스트 까지 유해 텍스트를 다량으로 학습하여 기존 GPT-3의 문제점으로 지적되어 온 가짜 정보, 편향된 결과, 혐오적 표현 등의 문제를 상당 부분 해결할 수 있다고 주장
  - 주요 학습 과정을 살펴보면, (1단계) 질문에 대해 사람이 올바른 답을 작성하고, 이 데이터셋으로 미세조정(fine-tuning)하고,
  - (2단계) 질문에 대해서 모델이 작성한 답을 사람이 점수를 매기고, 이를 사용하여 reward 모델을 학습한 다음,
  - (3단계) reward 모델을 기반으로 강화학습을 사용하여 다시 GPT를 미세조정함.

**그림 10** InstructGPT의 학습 과정



\* 출처: Long Ouyang et al.(2020.3.4.)

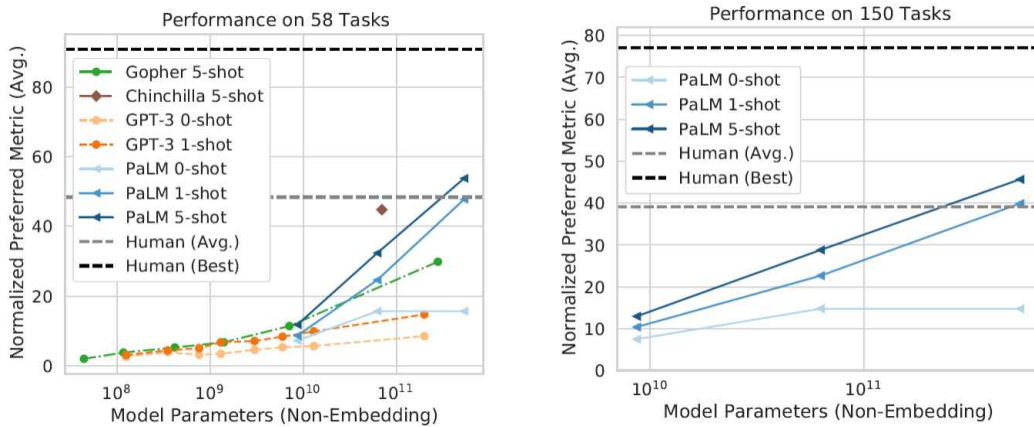
- 구글은 ‘문제 해결 방법 자체’를 학습해 다양한 분야에 활용할 수 있는 트랜스포머 기반 초거대 언어모델 ‘PaLM(Pathway Language Model)’을 공개<sup>16)</sup>
  - 기존 AI 모델이 특정 작업을 잘 수행하도록 학습된 알고리즘이라면, PaLM은 단일 AI 모델로 다양한 작업을 수행하기 위해 2021년 구글이 제안한 알고리즘으로 내부적으로 Pathways 구조를 활용
    - ※ Pathway는 하나의 AI 모델에서 다양한 작업을 수행할 수 있도록 기존 기술을 새로운 기술과 결합할 수 있으며, 영상, 오디오, 텍스트, 이미지 등 다양한 데이터 형식을 포괄하는 다중 학습 가능<sup>17)</sup>

15) Long Ouyang et al.(2020.3.4.), Training language models to follow instructions with human feedback, arXiv:2203.02155v1.

16) Google Blog(2022.4.4.), Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance.

- GPT-3보다 약 3배 큰 규모의 5,400억 개의 파라미터를 보유한 PaLM은 문장 완성, 문맥 이해, 질의응답, 상식 추론 등 29개 작업 중 28개에서 기존 AI 모델 'GPT-3', 'Megatron-Turing NLG', 'Gopher', 'Chinchilla' 등을 압도
- 또한, PaLM은 언어모델 벤치마크 'BIG-bench'에서 기존 모델은 물론 인간의 평균 점수를 넘어서는 놀라운 결과를 보여줌<sup>18)</sup>
- PaLM은 초등학교 수준의 수학 문제 벤치마크 'GSM8K 문제'를 58% 해결하여 GPT-3이 기록한 기존 최고 점수 55%를 넘었고, 어린이 평균 점수 60%에 근접했으며, 다단계 논리적 추론과 농담에 대한 설명까지도 생성
- 언어모델의 성능 향상에 있어서 딥마인드의 Gopher가 기존 초거대 AI 모델의 규모의 한계를 확인 시켰다면, PaLM은 모델의 크기가 클수록 성능이 뛰어날 수 있음을 시사

그림 11 PaLM 모델의 BIG-bench 평가 결과



\* 출처: Aakanksha Chowdery et al.(2022.4.5.)

- 초거대 언어모델 'PaLM'을 개발한 구글은 2022년 6월 자연어처리 분야를 넘어 양적 추론(Quantitative Reasoning)이 필요한 수학 문제를 풀 수 있는 AI 모델, '미네르바 (Minerva)'를 공개<sup>19)</sup>
  - 미네르바는 PaLM을 기반으로 개발되었으며, 자연어처리 모델과 표준 수학 표기법을 결합해 질문을 분석하고 이해하여 단계별로 추론해 답을 찾아내는 양적추론을 수행
  - 구글은 118GB에 달하는 수학 관련 논문과 데이터를 미네르바에 학습시켰으며, 미네르바는 수학 문제를 이해한 후 이를 단순화하고 변수와 방정식을 설정해 문제를 해결

17) Google Blog(2021.10.28.), Introducing Pathways: A next-generation AI architecture.

18) Aakanksha Chowdery et al.(2022.4.5.), PaLM: Scaling Language Modeling with Pathways, arXiv:2204.02311v1.

19) Google Blog(2022.6.30.), Minerva: Solving Quantitative Reasoning Problems with Language Models.

- 인공지능이 수학적 추론이 가능하기 위해서는 지금까지와는 다른 방식의 알고리즘 혁신이 필요하다는 주장과 달리 미네르바는 간단한 수학 문제뿐 아니라 방정식, 증명까지 수행
- 초거대 AI 언어모델이 언어 분야를 넘어 앞으로 다양한 과학기술 분야에서 활약하여 새로운 기회를 만들 수 있을 것으로 기대

**그림 12** 미네르바 수학 풀이 예시(위) 및 STEM 평가 결과(아래)<sup>20)</sup>

**Question:** Assume that the variance of the first  $n$  natural numbers is 10, and the variance of the first  $m$  even natural numbers is 16. Compute  $m + n$ .

---

**Model output:** Let  $\sigma^2$  be the variance of the first  $m$  even natural numbers, and  $\tau^2$  be the variance of the first  $n$  natural numbers. Then, we have  $\tau^2 = \frac{n^2-1}{12}$  and  $\sigma^2 = \frac{m^2-1}{3}$ . We solve the system of equations:

$$\frac{n^2-1}{12} = 10$$

$$\frac{m^2-1}{3} = 16.$$

This translates into the system

$$n^2 = 121$$

$$m^2 = 48 + 1 = 49.$$

Therefore,  $n = \boxed{11}$  and  $m = \boxed{7}$ , so  $n + m = \boxed{18}$ .

**Question:** For every  $a, b, b \neq a$  prove that

$$\frac{a^2 + b^2}{2} > \left(\frac{a+b}{2}\right)^2.$$


---

**Model output:**

$$\frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2$$

$$\iff \frac{a^2+b^2}{2} > \frac{a^2+b^2+2ab}{4}$$

$$\iff a^2 + b^2 > \frac{a^2+b^2+2ab}{2}$$

$$\iff 2a^2 + 2b^2 > a^2 + b^2 + 2ab$$

$$\iff a^2 + b^2 > 2ab$$

$$\iff a^2 + b^2 - 2ab > 0$$

$$\iff (a - b)^2 > 0$$

which is true, because the square of a real number is positive.

Model	MATH	MMLU-STEM	OCWCourses	GSM8k
Minerva	50.3%	75%	30.8%	78.5%
Published state of the art	6.9%	55%	-	74.4%

\* 출처: <https://ai.googleblog.com/2022/06/minerva-solving-quantitative-reasoning.html>.

- 한편 양적 성장이 만든 초거대 AI의 한계로 지적되어 온 효율성과 편향성 등을 개선하기 위해 다양한 연구가 진행되는 가운데, 특히 초거대 AI의 한계를 질적 혁신으로 뛰어넘으려는 멀티모달 AI에 주목할 필요
  - 멀티모달 AI는 인간 지능의 발달 과정과 유사하게 하나의 메커니즘에서 다양한 지능을 통합하려는 것으로, 장기적으로 봤을 때 AI 기술이 나아가야 할 연구 방향이자 도전 분야
  - 현재 세계적으로 가장 주목받고 있는 초거대 AI를 둘러싼 규모의 경쟁은 멀티모달 AI의 질적 발전과 결합하여 또 다른 혁신을 불러올 것으로 기대

20) MATH: 고등학교 수준의 수학문제, MMLU-STEM: 공학, 화학, 수학, 물리 복합문제, OCWCourse: MIT에서 제공한 대학 수준 문제, GSM8K: 초중등 수준 문제

# IV 초거대 멀티모달 AI

## 1 멀티모달 AI

- 멀티모달 AI란 인간처럼 하나의 감각이 아닌 다중감각, 즉 시각, 청각, 후각, 촉각, 미각 등을 통합적으로 사용하여 상황을 이해하고 예측, 판단, 반응을 결정하는 기술
  - 인간이 외부 세계를 이해하기 위해 사물을 보고, 소리를 듣고, 질감을 느끼고, 냄새를 맡고, 맛을 보는 등 다양한 정보를 복합적으로 사용하듯이, 현재의 인공지능이 일반 인공지능 (AGI)으로 도약하기 위해서는 멀티모달 정보를 해석하고 이해할 수 있어야 함
  - 이 과정에서 여러 소스의 멀티모달 데이터들을 단순 결합하지 않고, 효과적으로 통합하기 위한 새로운 방법이 요구됨<sup>21)</sup>
  - 그렇게 해야 의미상 서로 관련이 있고 때로는 상호 보완적인 특징을 발견할 수 있어, 개별 모달리티로 분석할 때는 보이지 않는 상황(패턴)을 정확히 예측할 수 있음
- 멀티모달 머신러닝(Multimodal Machine Learning; MMML)은 여러 감각 정보를 처리하고 관련시킬 수 있는 모델을 학습하고 생성하는 연구로, 특히 다섯 가지 분야 (표현, 변환, 정렬, 융합, 공동학습 등)에서 기술적 도전 과제가 존재함<sup>22)</sup>
  - 즉, 서로 다른 소스의 데이터를 사용하여 학습모델을 생성하는 과정에서 언제 어떤 방식으로 통합할 것인가에 대한 연구 분야

그림 13 멀티모달 머신러닝을 위한 다섯 가지 핵심기술



\* 출처: Tadas Baltrusaitis et al.(2019.2.) 참고하여 지능정보연구본부 작성.

21) 개별 모달리티를 단순결합하는 것과 차별화하기 위해 크로스모달(crossmodal)이라는 용어를 사용하기도 함.  
 22) Tadas Baltrusaitis et al.(2019.2.), Multimodal Machine Learning: A Survey and Taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence.

- 멀티모달 머신러닝은 인식 정확도 향상, 소셜 인터랙션, 상황이해·추론, 질병 예측 등 원천기술연구부터 다양한 응용기술 개발까지 활발한 연구가 진행되고 있음
  - (인식 정확도 향상) Amazon은 시각과 청각 정보를 이용해 노이즈를 제거함으로써 음성 인식의 정확도를 향상할 수 있는 Audio Visual Automatic Speech Recognition 시스템을 개발<sup>23)</sup>
  - ※ CES 2021에서 소음이 심한 환경에서 사람의 입술 모양을 추가 사용하여 음성 인식의 성능을 개선한 차량용 Lip Reading 기술 시연<sup>24)</sup>
  - (소셜 인터랙션) EU의 EMPATHIC 프로젝트는 음성, 텍스트 등을 활용하여 사람의 감정을 분석하여 노인들의 독립적 생활과 사회적 상호작용을 개선할 수 있는 기술을 개발<sup>25)</sup>
  - ※ 초기 감성 분석 연구는 시각이나 텍스트 정보만을 이용했으나 최근에는 시각, 청각, 텍스트를 모두 이용하는 연구가 진행되고 있음
  - (상황이해·추론) Allen 연구소는 사진을 보고 상황 질문에 대해 답변을 추론하는 VCR(Visual Commonsense Reasoning) 기술을 개발<sup>26)</sup>
  - ※ DARPA는 텍스트, 이미지, 음성 등으로부터 사건, 상황을 해석하는 다중가설검정 시맨틱 엔진을 개발<sup>27)</sup>
  - (질병 예측) 의료 AI 기업 'Owkin'은 방사선 자료, 조직 단층 사진, 게놈 데이터 등 임상 자료와 다양한 멀티모달 데이터를 활용하여 환자의 예후나 질병을 예측하는 연구 진행<sup>28)</sup>
  - ※ Univ. of Arizona Health Science는 다양한 대량의 데이터를 실시간 이용하여 높은 정확도와 감도를 지닌 멀티모달 센서 시스템을 개발하는 NSF 프로젝트를 진행<sup>29)</sup>
  
- 고객과 다양한 접점에서 원하는 검색 결과를 제공하거나 고객의 질문에 자연스럽게 대화하고 상담할 수 있는 멀티모달 AI 원천기술 개발 및 상용 서비스 등장
  - 구글은 2020년 비디오 콘텐츠 검색을 위한 일곱 가지 멀티모달 정보를 통합한 트랜스포머 구조를 제시했고<sup>30)</sup>, 2021년 Google I/O에서 75개 이상의 언어와 이미지, 비디오 등 다양한 정보를 이해하는 멀티모달 통합 모델 'MUM(Multitask Unified Model)'을 공개
  - 아마존은 'Echo Show'를 비롯한 음성 인식 스피커를 여러형태의 멀티모달 AI 디바이스로 개발하여 시장에 출시하고 있음

23) Georgios Paraskevopoulos et al.(2020.4.29.), Multiresolution and Multimodal Speech Recognition with Transformers, arXiv:2004.14840v1.

24) <https://www.youtube.com/watch?v=1Po0-rvPqmk>.

25) <http://www.empathic-project.eu/>.

26) Rowan Zellers et al.(2019.3.26.), From Recognition to Cognition: Visual Commonsense Reasoning, arXiv:1811.10830.

27) <https://www.darpa.mil/program/active-interpretation-of-disparate-alternatives>.

28) <https://owkin.com/machine-learning/multimodal-data-integration/>.

29) <https://precisionhealth.uahs.arizona.edu/multimodal-sensor-systems-precision-health-enabled-data-harnessing-artificial-intelligence-and>.

30) Gabeur et al.(2020.8), Multi-modal Transformer for Video Retrieval, ECCV 2020.

- 다양한 방식으로 진행되고 있는 멀티모달 AI 연구 가운데, 최근 트랜스포머 구조에 기반을 둔 초거대 AI 언어모델과 시각 지능을 통합하려는 시도에 주목할 필요가 있음
  - 2020년 GPT-3 발표 이후, 초거대 AI 언어모델이 보여준 놀라운 언어이해 능력과 트랜스포머 모델의 확장 가능성은 멀티모달 AI 연구 방향에 큰 영향을 미치고 있음
  - 특히, 트랜스포머 모델에 기반하여 언어와 시각이라는 인간의 대표적인 감각 정보를 통합하려는 언어 시각 트랜스포머 연구에서 의미 있는 성과들이 공개되고 있음
  - 그동안 다양한 감각 정보를 사용하여 연구 중인 멀티모달 AI 기술은 초거대 언어지능을 중심으로 시각 지능을 결합해 더욱 인간을 닮아가려는 영역에서 빠르게 발전 중
  - 시각과 언어를 결합한 연구로는 주어진 시각 이미지에 대한 텍스트 자막을 생성하는 Image Captioning, 주어진 시각 이미지와 관련된 텍스트 형태의 질문에 답하는 Visual Question Answering, 텍스트 검색어를 이용한 이미지/비디오 클립 검색, 시각 정보를 텍스트 번역에 사용하는 Multimodal machine translation 등이 대표적

## 2 언어 시각 트랜스포머

### □ 세상을 이해하기 위한 첫걸음

- 2010년 초, 딥러닝에 기반을 둔 인공지능은 ‘볼’ 수 있는 방법을 획기적으로 개선했고, 점차 ‘읽고’, ‘말’하는 방식에서도 인간을 모방하며 의미 있는 성과를 내기 시작
  - 하지만 읽고 말하는 영역에서 우수한 성과를 보여주고는 있으나, 여전히 세상을 이해하는 능력에 있어서 인간과 같이 충분한 유연성을 갖지 못하는 한계를 노출
  - 이러한 문제를 극복할 수 있는 다양한 방법이 제시되고 있으며, 이 가운데 인간의 성장 과정에서 어린아이들이 세상을 이해하는 방식에 주목
  - 즉, 어린아이들이 세상을 ‘보고’, 그것에 대해 ‘말함’으로써 성장하듯이, 지난 10년 동안 터득한 ‘시각’ 지능을 ‘언어’ 지능과 논리적으로 긴밀히 연결하면서 인공지능을 성장시킬 수 있을 것이란 기대
- 언어 시각 트랜스포머는 인공지능이 사람처럼 세상을 이해하기 위해서, 초거대 AI 언어 지능과 이보다 먼저 충분한 성능을 입증한 시각 지능을 우선 결합하려는 시도
  - 핵심은 언어와 시각이라는 서로 다른 감각 정보를 어떻게 연결하고 통합할 수 있는가인데 이 과정에서 트랜스포머 모델의 확장성이 결정적 역할을 함
  - 즉, 2010년 중반 이후 트랜스포머 모델을 중심으로 이뤄진 언어처리 기술 발전과 초거대 AI 언어 모델은 언어와 시각 정보를 긴밀히 연결할 수 있는 기반을 제공
  - 초기 언어 시각 멀티모달 연구는 주로 VQA(Visual Question Answering)과 같이 이미지와 질문을 주고 그에 알맞은 답을 찾는 방식이 주를 이루었다면, 최근 들어 이미지와 텍스트의 쌍에 기반한 다양한 추론(이미지 생성 등) 연구로 발전

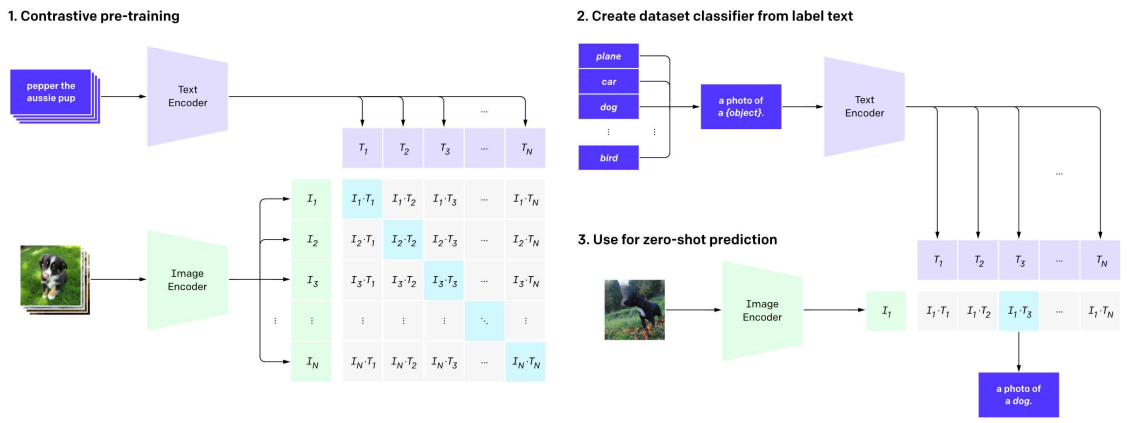
### □ 연구 동향

- Open AI에서 공개한 ‘CLIP’<sup>31)</sup>은 기존 머신러닝 분류 모델이 접근해온 방식을 바꾼 혁신적 연구로 주목받음
  - CLIP은 별도의 레이블링된 정보를 주지 않아도 다양한 이미지(추상적 이미지 등)를 인식하여 언어로 이해(표현)하거나, 반대로 언어를 이미지로 이해(표현)
  - 사람은 하나의 단어를 들었을 때 구체적인 이미지는 물론 관련 경험과 같은 추상적인 이미지를 연상 하듯이, CLIP은 멀티모달 뉴런 방식을 통해 추상적 이미지까지 인식할 수 있음
  - 또한 CLIP은 인터넷에 있는 수많은 사진과 설명된 문장을 수집해 스스로 학습하는 자기지도학습을 사용해 데이터 레이블링 문제를 해결
  - 기존 분류 모델은 가능한 결과가 미리 정해져 있는 레이블 공간(label space)을 가정하고 있기에 새로운 분류가 추가되어 레이블 공간이 계속 바뀌는 경우, 모델 재학습을 위한 비용이 많이 발생

31) <https://openai.com/blog/clip/>.

- 이미지 분류 모델의 기존 방식은 입력 이미지를 숫자로 치환하기 때문에 레이블 공간에 대한 지식을 전혀 활용하지 못함
- CLIP에서는 이미지와 텍스트를 모두 입력값으로 사용하는 대조 학습(contrast learning)을 수행하는 과정에서 이미지와 텍스트 쌍을 구성하여 레이블 공간의 지식을 활용함
- ※ 대조 학습은 두 개의 데이터에 대해 같은 분류에 속하는지 아닌지를 구분할 수 있도록 모델을 학습하는 방식으로, CLIP에서는 해당 이미지와 텍스트를 대조하여 얼마나 적합한지를 표현(아래 그림에서 1단계: N개의 이미지와 텍스트 쌍이면 N개의 참인 경우와 나머지  $N^2 - N$  개의 거짓인 쌍이 존재함)
- ※ CLIP은 해당 이미지에 대해 N개의 텍스트 후보 가운데 가장 점수가 높은 텍스트를 최종 분류로 선택하고, 새로운 텍스트 후보를 추가할 경우에도 모델을 재학습하지 않고 제로 샷 러닝(zero-shot learning)을 수행함(아래 그림에서 2단계: N개 이미지의 분류에 대한 텍스트를 만들고 각각 이미지에 대한 적합도를 계산)
- 즉, CLIP은 이미지 인코더와 텍스트 인코더를 사전 훈련하여 데이터 세트에서 어떤 이미지가 어떤 텍스트와 쌍을 이루는지 예측 → 이 동작을 사용하여 CLIP을 제로샷 분류기로 전환 → 데이터 세트의 모든 클래스를 '개 사진'과 같은 캡션으로 변환하고 캡션 클래스를 예측 → CLIP은 주어진 이미지와 가장 좋은 쌍을 예측

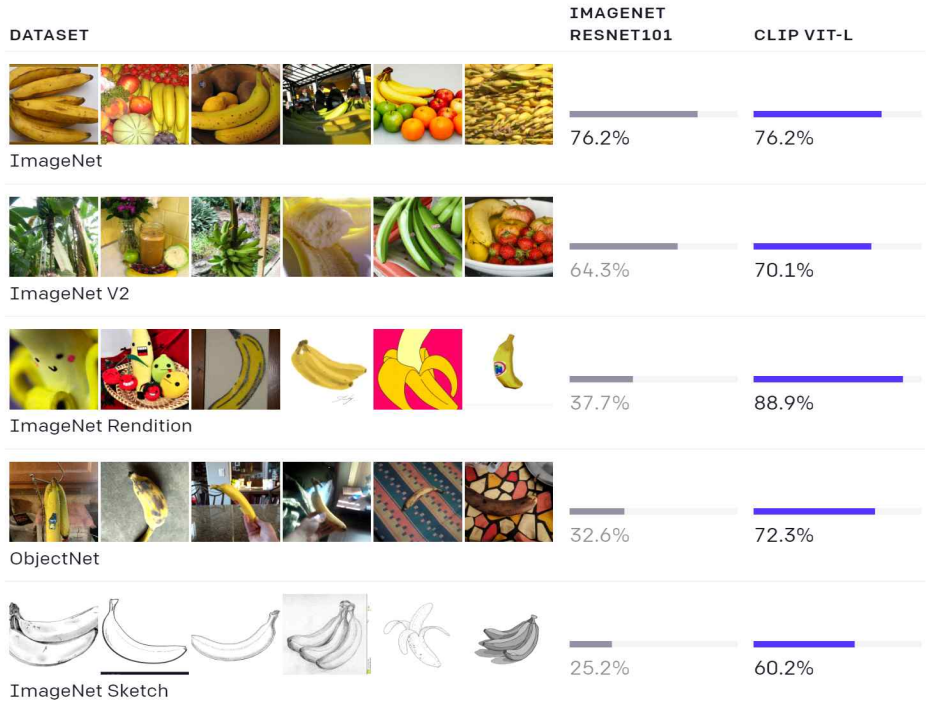
그림 14 CLIP 학습 과정



\* 출처: <https://openai.com/blog/clip/>.

- CLIP을 이용한 실험 결과를 보면, 다양한 환경(사진, 스케치 등 사람이 보기에는 비슷하나, 데이터 특성이 달라서 기존 분류 모델에서는 낮은 성능을 보임)에서 우수한 성능을 보여줌(기존 RESNET101과 비교한 아래 그림 참조)
- 하지만, CLIP은 자연어처리기술을 기반으로 개발되어, 글자와 이미지가 동시에 있을 때 글자를 우선 인식함으로써 타이포그래피 공격(typography attack)에 취약한 문제를 노출
- ※ 예를 들어, 사과에 엉뚱한 글씨가 적힌 종이를 붙여놓으면 사과 이미지는 무시하고 글씨만 인식

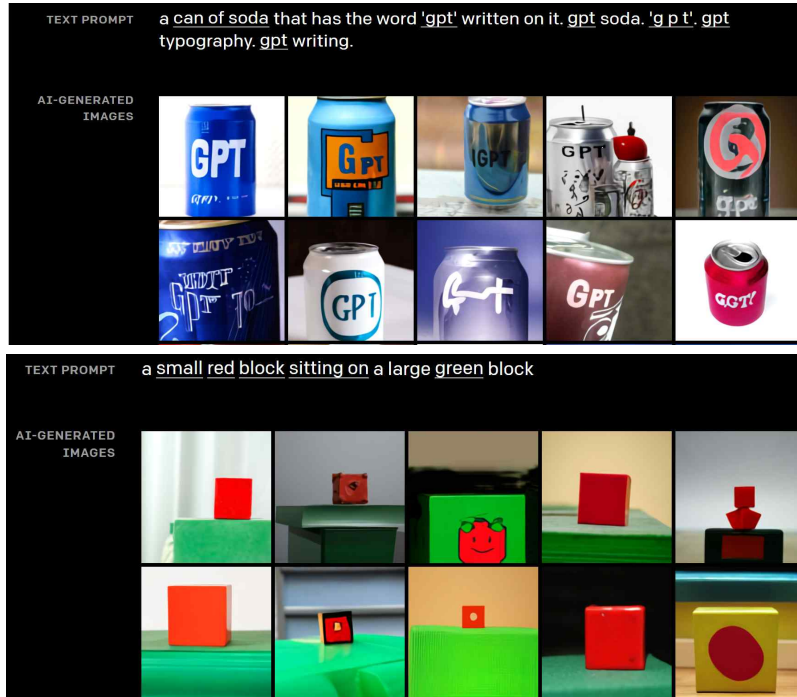
**그림 15** CLIP과 RESNET101 성능 비교



\* 출처: <https://openai.com/blog/clip/>.

- 한편, OpenAI의 ‘DALL-E’는 GPT-3 트랜스포머 구조를 사용하여 텍스트로부터 이미지를 생성(text-to-image)하는 기능을 수행
  - DALL-E는 GPT-3과 달리 입력값으로 텍스트와 이미지를 하나의 데이터 스트림으로 만들어 학습함으로써, 텍스트의 내용과 이미지를 어느 정도 연관시킴
  - ※ DALL-E는 미술가 살바도르 달리(Salvador Dalí)와 픽사의 애니메이션 로봇(WALL-E)의 합성어로, 2,500만 개 텍스트-이미지 쌍으로 학습하여 최종 120억 개의 파라미터로 구성됨
  - DALL-E는 입력 텍스트로부터 기존에 없던 새로운 이미지를 생성하고, 텍스트에 포함된 단어의 의미를 표현하는 이미지도 생성(예를 들어, 텍스트에 포함된 위치 공간을 이해하는 이미지를 생성하고 특정 단어를 포함하는 간판 이미지를 생성)
  - ※ DALL-E는 입력된 언어에 대해 여러 장의 이미지를 생성하는 과정에서, OpenAI의 또 다른 AI 모델 ‘CLIP’이 이미지별로 점수를 부여하여 이미지 선별 과정에 개입
  - 공개된 결과를 보면, DALL-E는 생성된 이미지에서 객체를 능숙하게 재배치하고 다루는 능력을 갖췄고, 질감과 입체감 등을 세심하게 표현할 수 있어 패션, 디자인, 교육 등의 분야에서 창의적 아이디어를 제시하는 등 더 많은 분야에서 활용 가능성이 기대됨

그림 16 DALL-E 창작물



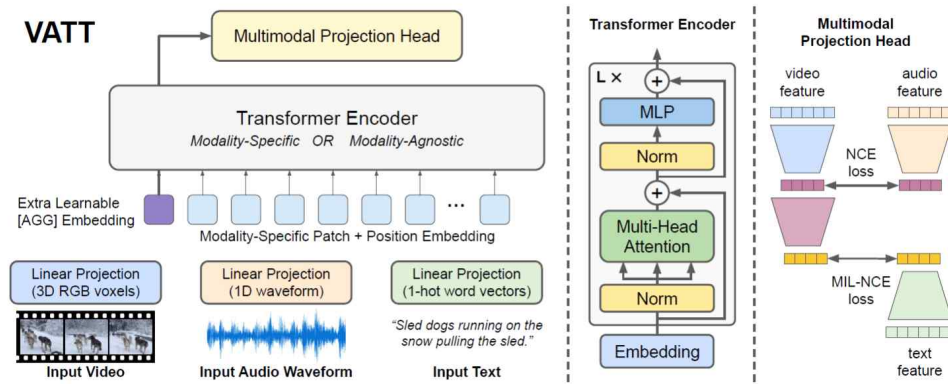
\* 출처: <https://openai.com/blog/dall-e/>.

- VATT<sup>32)</sup>는 Convolution을 사용하지 않는 트랜스포머 구조를 기반으로 오디오, 비디오, 텍스트를 입력값으로 하는 멀티모달 대조 학습(contrastive learning)으로 자기지도 사전학습(self-supervised pre-training) 모델을 생성
  - 이렇게 생성된 사전학습 모델은 비디오 동작 분류, 오디오 사건 분류, 제로 샷 비디오 검색, 이미지 분류 등 downstream task<sup>33)</sup>에 매우 우수한 성능을 보여줌
  - 흥미로운 점은 VATT는 기본적으로 오디오, 비디오, 텍스트 등 모달리티별로 다른 인코더(modality-specific)를 사용해 좋은 성능을 보였고, 특히 하나의 통합 트랜스포머를 사용하더라도 우수한 성능을 제시
  - 향후 하나의 통합 트랜스포머를 사용하여 멀티모달 모델을 구현할 수 있다는 가능성을 보여줌

32) Hassan Akbari et al.(2021.12.7.), VATT: Transformer for Multimodal Self-Supervised Learning from Raw Video, Audio and Text, arXiv:2014.11178v3.

33) pre-training하고 어떤 문제를 해결하기 위해 fine-tuning 한다면, 해결하려는 문제를 downstream task라고 부름

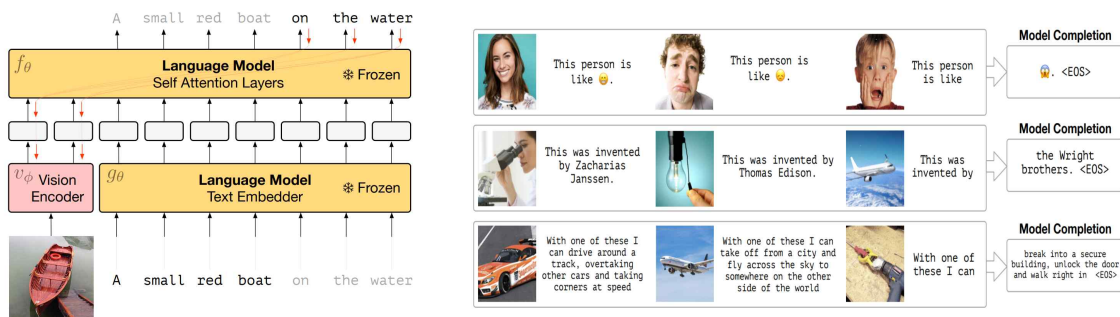
그림 17 VATT 구조



\* 출처: Hassan Akbari et al.(2021.12.7.)

- 딥마인드는 멀티모달 퓨샷학습 방법, ‘Frozen’을 통해 VQA(Visual Question Answering) 문제에서 주어진 질문에 답변이 가능한 기술을 소개<sup>34)</sup>
  - Frozen은 텍스트 이외의 문제에 대해서는 충분한 성능을 보여주지 못한 기존 초거대 AI 언어모델의 한계를 개선하기 위한 목적으로 개발됨
  - Frozen에서 제안한 멀티모달 퓨샷학습 방법은 이미지와 텍스트 쌍을 이용하여 사전학습 언어모델이 적절한 텍스트를 생성할 수 있도록 함
  - ※ 처음에는 단일 이미지와 캡션의 쌍으로 학습되나, 일단 학습된 후에는 여러 이미지와 텍스트 셋에도 효과적으로 대응함
  - 이를 통해 VQA 작업뿐 아니라 몇 개의 예제 문장(few-shot)을 주고 그에 대한 답을 제시하는 놀라운 능력을 보여줌

그림 18 Frozen의 멀티모달 퓨샷학습 구조(왼쪽) 및 추론 결과(오른쪽)



\* 출처: Maria Tsimpoukelli et al.(2021.6.3.)

34) Maria Tsimpoukelli et al.(2021.6.3.), Multimodal Few-Shot Learning with Frozen Language Models, arXiv:2106.13884v2.

- 중국의 대표적인 언어 시각 트랜스포머 연구로는 CLIP과 유사한 WenLan<sup>35)</sup>, DALL-E와 유사한 M6<sup>36)</sup> 등이 있음
  - WenLAN은 CLIP과 유사한 중국어 버전의 멀티모달 사전학습 모델이며, CLIP에 비해 간략화된 대조 학습(Contrastive Learning)을 사용함
  - ※ WenLAN 초기 버전은 이미지 텍스트 쌍 3천만 장 규모이며, 향후 3 억장 규모로 확대 예정이며, 파라미터 규모는 10억~100억 규모
  - M6은 300GB 규모의 중국어 텍스트와 2TB 규모의 이미지를 통해 만든 멀티모달 사전학습 모델
  - ※ M6은 실험 결과에서 기존 성능 대비 VQA에서 11.8%, 이미지 캡셔닝에서 18.4%, 텍스트와 이미지 매칭에서 10.3% 이상 우수한 성능을 보임
  
- 국내의 경우, 카카오브레인, 네이버, LG AI연구원 등 빅테크 기업을 중심으로 초거대 멀티모달 AI 개발을 강화하고 자사 서비스에도 활발히 적용 중
  - 카카오브레인은 2021년 12월, 한국어에 특화된 AI 언어모델 'KoGPT'와 초거대 멀티모달 AI 'minDALL-E'를 공개하며 관련 기술 생태계를 확장 중
  - ※ minDALL-E는 1,400만장의 텍스트와 이미지 셋을 통해 사전학습하여 생성
  - 네이버는 2,040억 개에 이르는 파라미터를 가진 초거대 AI 언어모델 'HyperCLOVA'를 개발한 데 이어, 2022년 4월 멀티모달 AI를 개발해 자사 검색 서비스 '스마트렌즈'에 적용
  - LG AI연구원은 2022년 12월, 초거대 멀티모달 AI 'EXAONE'을 공개하고 제조, 연구, 교육, 금융 등의 분야에서 상위 1% 전문가 수준에 도달하는 것을 목표로 멀티모달 AI 기술 개발을 강화

35) Yuqi Huo et al.(2021.3.11.), WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training, arXiv:2103.06561.

36) Junyang Lin et al.(2021.3.1.), M6: A Chinese Multimodal Pretrainer, arXiv:2103.00823.

## V 이슈와 전망

### □ 초거대화

- 현재 초거대 AI 언어모델 연구는 모델의 파라미터 수를 늘려서 성능을 향상하려는 방향으로 치중되어 있으나, 효율성과 편향성 극복 그리고 대용량의 데이터로부터 학습한 지식을 확장하는 방법 등이 중요한 연구 이슈로 제기됨
    - BERT 언어모델이 제안된 이후 트랜스포머에 기반을 둔 초거대 AI 언어모델 연구가 활발히 이루어지고 있으나, 다양한 인코더/디코더 구조 개선과 언어 이외의 분야로 지능 확장이 필요
    - 즉, 편향성과 효율성 극복이라는 오래된 AI 연구와 언어지능을 다른 지능과 통합하여 지식을 확장하기 위한 초거대 멀티모달 AI 연구 필요
  
  - (효율성) 초거대 사전학습 언어모델의 효율성 극복은 트랜스포머 구조, 모델, 학습 등의 관점에서 개선 방안이 필요함<sup>37)</sup>
    - 구조 효율성: 인코더 기반의 사전학습 모델(BERT 계열)은 언어생성에 한계를 보이고, 디코더 기반의 사전학습 모델(GPT 계열)은 언어이해에 대한 한계를 가지고 있는 만큼, 현재 사용되고 있는 트랜스포머의 인코더-디코더 구조 자체를 개선하는 연구를 진행할 필요가 있음
    - 모델 효율성: 언어모델의 학습 지식은 딥러닝 모델에 매우 함축적(implicit)으로 표현되기에, 시간에 따라 결과가 달라지는 문제에 대응이 요구됨
    - 학습 효율성: 대부분의 사전학습 언어모델을 학습하기 위해서는 매우 많은 양의 데이터와 컴퓨팅 자원이 필요하므로 지금보다 샘플 효율적인 학습 방법이 필요함
  
  - (편향성) 최근까지 초거대 AI 언어모델은 우수한 실험 결과에도 불구하고 여전히 편향성 문제를 가지고 있고 대부분 학습데이터 문제에 기인
    - AI 학습에 필요한 데이터는 많은 경우 인터넷으로부터 수집된 것으로, 사회적으로 존재하는 편견이 데이터에 반영되어 있거나 데이터 소스 자체가 편향될 수도 있음
    - 데이터로 인한 AI 편향성<sup>38)</sup>은 데이터 표본 편향, 롱테일 편향, 인간 편향, 의도적 편향 등 다양하며 기술적 대응이 필요
- ※ GPT-3은 고학력 직종이나 육체노동 직종에서 남성에 대한 쓸림이 심했고, 연극 공연 캐스팅 과정에서 중동 출신 배우를 테러리스트로 지정하는 경향을 보이기도 함

37) 임준호(2021.2.18.), Pre-trained Language Models for NLP, Pattern Recognition and Machine Learning Winter School 2021.

38) Fobes(2020.11.9.), How AI Can Go Terribly Wrong: 5 Biases That Create Failure.

- (지식 확장) GPT-3은 초거대 AI 언어모델이 일반 인공지능(AGI)로 발전할 가능성을 보여주었으나, 더욱 근본적인 지능의 성장과 확장 연구가 필요
  - 현재 초거대 AI 언어모델이 인간과 같은 멀티태스킹 능력을 보여주기 위해서는 사전학습 이외에 지속적인 사후학습 방법이 필요함
  - 한편, 언어뿐 아니라 영상, 오디오 등 멀티모달 정보를 통합할 수 있는 근본적인 기술 연구가 필요함

## □ 멀티모달化

- 지능 확장을 위한 초거대 멀티모달 AI 연구 가운데 대표적인 언어 시각 트랜스포머 모델은 언어지능과 시각지능을 논리적으로 긴밀히 통합하기 위한 연구가 필요
  - 즉, 서로 다른 정보(언어, 시각)를 받아들여 특징을 생성하고 학습하는 과정에서 어떤 지점에서 어떤 방식으로 통합할 것인가에 대한 연구 이슈가 존재함
  - 현재 언어 시각 멀티모달 AI 연구는 기존의 언어지능에서 널리 활용되고 있는 트랜스포머 구조를 중심으로 언어와 시각 데이터를 통합 진행
  - 그러나 기존 연구는 멀티모달 정보를 단순 결합하는 방식으로 모달리티 생성과 상호 변환에 한계가 있음
    - ※ 트랜스포머 기반 시각 지능의 경우, 언어처리 분야에서 연구되어 온 트랜스포머 구조에 시각 정보를 처리하는 과정에서 기존 CNN 등과 달리 많은 정보의 손실이 발생할 수 있어, 새로운 구조(architecture) 및 통합 모델(unified model)에 대한 개발 필요성이 제기됨
- 지능 확장을 위해 신경망 내부에서 잠재공간(latent space)의 활용 가능성과 한계를 분석하고 어떻게 지능을 통합, 확장할 것인가에 대한 고민이 필요함
  - 멀티모달 데이터(언어/음성/시각 등)는 서로 이질적인 정보이며, 정보 표현의 세분성(granularity) 차이가 크기 때문에, 하나의 잠재공간<sup>39)</sup>에서 사용하기에는 많은 어려움이 존재함
  - 현재 대부분의 멀티모달 AI 연구는 멀티모달 정보를 단순히 결합(임베딩)하여 신경망을 학습하기에 모달리티 생성 및 상호 변환에 한계가 있음
  - 지금까지 개별 지능(시각, 언어 등)이 잠재공간에서 연산과 탐색을 통해 창의적인 결과를 만들 수 있는 잠재 가능성에 대해서는 여러 논문에서 이미 확인된 바 있음
    - ※ 잠재공간에서 연산 예: (언어 분야) 'king - man + woman = queen', (시각 분야) '팔걸이가 있는 의자 - 팔걸이가 없는 의자 = 팔걸이'

39) AI 학습에 사용되는 데이터는 실제 존재하는 사물과 상황을 결정하는 조건을 모델링한 잠재변수(latent variable)를 사용하며, 잠재변수들로 구성된 잠재벡터(latent vector)가 존재하는 공간을 잠재공간(latent vector)이라고 함

## □ 자연지능 모사化

- 초거대 AI는 기하급수적인 양적 혁신으로 가능했고, 이에 더하여 멀티모달 AI 연구의 질적 혁신과 결합하여 기존 AI 알고리즘의 근본적 한계를 기술적으로 극복할 가능성을 보여주었으나, 초거대 AI가 일반 인공지능으로 발전하기 위해서는 이보다 근본적인 기술적, 개념적 한계를 극복해야 함
  - 기술적 한계: 대량 학습 데이터 필요, 유연성과 일반화 어려움, 신뢰성 및 공정성 이슈 등으로 꾸준한 기술 개발을 통해 극복 가능한 부분
  - 개념적 한계: 인공지능이 처리하는 정보의 의미를 이해하는 능력과 심볼 그라운드문제, 체화된 인지 개념에 따른 몸체를 가진 인공지능(embodied AI) 등 현재 딥러닝 구조와는 완전히 새로운 개념과 방향성이 요구되는 부분

표 1 현재 인공지능의 한계

구분	내용
기술적 한계	<ul style="list-style-type: none"> <li>● 학습의 한계                             <ul style="list-style-type: none"> <li>- 신경망의 태생적 정보처리 능력, 인지주의적 학습 참조</li> </ul> </li> <li>● 정보전달의 한계                             <ul style="list-style-type: none"> <li>- 신경망의 Top-down<sup>40</sup>과 Bottom-up<sup>41</sup> 양방향 정보처리를 통한 학습과 뉴런의 보편적 근사</li> </ul> </li> <li>● 과적합 및 스케일의 한계                             <ul style="list-style-type: none"> <li>- 기존 데이터 기반, 경험적, 상향식 모델에서 뇌 동작원리의 Top-down + Bottom-up 모델링 등 다각적 접근을 통해 인간의 고위수준 지능요소 구현</li> <li>- 도메인 적응성 및 유연성 확보와 모델의 확장성 및 재학습 속도 향상</li> </ul> </li> <li>● 기능 경계의 한계                             <ul style="list-style-type: none"> <li>- 학습-추론-다개체-메타인지 등의 통합 메타제어를 통한 공학적 한계들의 통합적 극복</li> </ul> </li> </ul>
개념적 한계	<ul style="list-style-type: none"> <li>● 의미 이해 문제                             <ul style="list-style-type: none"> <li>- 인공지능이 처리하는 정보가 그 정보 외에 어떤 것과 지향적 관계 또는 지향성을 가는 것에 대한 의미 능력 (함의, 전체, 인과, 부분-전체, 목적-수단, 시공간적 관계 등 다양한 연관관계의 개념 관계 학습-생성과 이들의 조합을 통한 상황을 인지하는 능력)</li> </ul> </li> <li>● 심볼그라운드(symbol grounding) 문제                             <ul style="list-style-type: none"> <li>- 기호주의 인지 (규칙에 의해 기호들을 형식적으로 처리하는 과정)에 대한 의미 문제 (구조화된 정보를 어떻게 세계와 연결할 것인가) 및 심적 상태 (동기 내재화)에 대한 의미 문제</li> </ul> </li> <li>● 체화 문제                             <ul style="list-style-type: none"> <li>- 정보와 세계를 연결(의미를 생성)하기 위해 신체가 필요하다는 체화된 인지 (embodiment cognition) 개념에 따른 물리적, 유기체적 체화를 통한 체화된 인공지능 (embodied AI)</li> </ul> </li> </ul>

\* 출처: 김진아 외 (2021.12.)

40) 시각, 언어 등 뇌의 특정 기능에서 시작하여 주요 인지기능에 관련된 뇌 기능 부위 간 상호작용에 대한 네트워크 시스템 레벨의 연구 결과를 신경망 모델에 구현하고자 하는 연구 방향.

41) 뇌를 구성하는 개별 신경이나 신경 간 연결의 구조, 기능, 정보처리 과정을 인공지능에 적용하여 신경망 모델을 개선하고자 하는 연구 방향.

- 현재 인공지능의 한계는 모델의 제한된 표상력 등 기술적 한계와 이를 넘어선 보다 근본적인 개념적 한계에서 오는 문제이며, 이를 해결하기 위한 대표적인 접근 방법으로 자연지능 모사 AI가 있음<sup>42)43)</sup>
  - 자연지능은 인간 지능의 주요 특성 가운데 하나인 유아 단계에서 경험을 통해 지능이 발달한다는 지능발달론적 접근
  - 자연지능 연구는 하나의 물리적 시스템으로 다양한 일을 수행하고, 유연하고 적응력이 높고 스스로 지속 학습이 가능하며, 몇 번의 경험만으로 학습하는 인간의 학습 과정을 모방하려는 방식
  - 즉, 자연지능 모사 AI는 주어진 환경에 대한 정확한 감각과 지각을 형성하며, 학습한 기억과 지식에 기반하여 판단과 계획을 수립하고 이에 따라 상황과 목적에 적합한 행동을 수행하는데 필요한 일련의 지능적 과정
  - 구체적으로 (1) 주로 개별 뉴런 수준에서 입출력 형태나 구조, 각 뉴런 간의 연결 방식 변화 등 마이크로 수준의 신경망 모델의 효율성 개선 연구, (2) 시각, 언어 등 감각 영역을 두뇌 피질 간 복합적 연결, 기억 메커니즘, 자기의식 등 매크로 수준의 뇌인지 모사 인공지능 연구 등으로 구분
  - 이러한 접근 방법을 통해 자연지능 모사 AI는 현 딥러닝 기반의 인공지능(초거대 AI, 멀티모달 AI 등)의 한계와 문제를 극복할 것으로 기대
- 자연지능 모사 AI는 현 인공지능의 한계를 해결할 수 있는 장점이 크나, 매우 초기 연구단계이며 주변 상황이나 감정에 대한 편향성, 기억의 왜곡, 지각 능력의 제한 등의 단점이 존재하기에, 장기적 관점에서 현 인공지능과 자연지능의 장점을 결합한 통합적 연구로 진행할 필요
  - 데이터에 기반을 두고 있는 지금의 인공지능 연구는 상황에 대한 편향성 제어가 가능하고, 기억을 담당하는 메모리의 왜곡이 없으며, 대규모 컴퓨팅 자원과 알고리즘 확장을 통해 자연지능의 한계를 넘어설 수 있는 장점이 있음
  - 자연지능 모사 AI는 뇌를 구성하는 요소를 인공지능에 적용(bottom-up)하고, 뇌기능 형성에 중요한 정보처리 과정을 인공지능에 적용하여(top-down) 현 인공지능의 한계를 극복 가능
  - 이를 통해 스스로 경험하여 자율적으로 학습하되, 편향되지 않으며 기억의 왜곡이 없이 다양한 작업을 빠르고 유연하게 처리할 수 있는 통합 인공지능을 구현할 수 있을 것으로 기대
  - 나아가서 통합 인공지능은 지금보다 훨씬 유연하고 자연스럽게 인간과 상호작용하여 협업함으로써 인간과 공존하며 개인뿐 아니라 인류 사회 전체 이익에 우선하는 방향으로 발전 전망

42) 김진아 외(2021.12.), 생명체의 유연한 자연지능 모사 방법과 데이터 기반 방법을 융합한 통합 인공지능, 정보과학회지 2021.

43) 정보통신기획평가원(2020.12.), 인공지능 기술청사진.



## 참고문헌

### ○ 국내자료

김진아 외(2021.12.), 생명체의 유연한 자연지능 모사 방법과 데이터 기반 방법을 융합한 통합 인공지능, 정보과학회지 2021.

정보통신기획평가원(2020.12.), 인공지능 기술청사진.

임준호(2021.2.18.), Pre-trained Language Models for NLP, Pattern Recognition and Machine Learning Winter School 2021.

ETRI 지능정보연구본부(2021.12.), 텍스트에서 비디오를 생성하는 언어시각 트랜스포머 연구, 내부자료.

### ○ 국외자료

Aakanksha Chowdhery et al.(2022.4.5.), PaLM: Scaling Language Modeling with Pathways, arXiv:2204.02311v1.

Abdelmohsen Ali et al.(2015.6.), Next Generation M2M Cellular Networks: Challenges and Practical Considerations, IEEE Communications Magazine.

Alexey Dosovitskiy et al.(2021.6.3.), An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929v2.

Ashish Vaswani et al.(2017.12.6.), Attention Is All You Need, arXiv:1706.03762v5.

Fobes(2020.11.9.), How AI Can Go Terribly Wrong: 5 Biases That Create Failure.

Gabeur et al.(2020.8), Multi-modal Transformer for Video Retrieval, ECCV 2020.

Georgios Paraskevopoulos et al.(2020.4.29.), Multiresolution and Multimodal Speech Recognition with Transformers, arXiv:2004.14840v1.

Google Blog(2021.10.28.), Introducing Pathways: A next-generation AI architecture.

Google Blog(2022.4.4.), Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance.

Google Blog(2022.6.30.), Minerva: Solving Quantitative Reasoning Problems with Language Models.

- Hassan Akbari et al.(2021.12.7.), VATT: Transformer for Multimodal Self-Supervised Learning from Raw Video, Audio and Text, arXiv:2014.11178v3.
- IDC(2017.4.), Data Age 2025: The Evolution of Data to Life-Critical, An IDC White Paper sponsored by Seagate.
- Jacob Devlin et al.(2019.5.24.), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2.
- Jared Kaplan et al.(2020.7.22.), Language Models are Few-Shot Learners, arXiv:2005.14165v4.
- Junyang Lin et al.(2021.3.1.), M6: A Chinese Multimodal Pretrainer, arXiv:2103.00823.
- Larry Downes & Paul Nunes(2014.1.7.), Big Bang Disruption: Starategy in the Age of Devastating Innovation, Portfolio.
- Long Ouyang et al.(2020.3.4.), Training language models to follow instructions with human feedback, arXiv:2203.02155v1.
- Maria Tsimpoukelli et al.(2021.6.3.), Multimodal Few-Shot Learning with Frozen Language Models, arXiv:2106.13884v2.
- Mehdi Noroozi et al.(2018.5.1.), Boosting Self-Supervised Learning via Knowledge Transfer, arXiv:1805.00385v1.
- Nathan Benaich, Ian Hogarth(2021.10.12.), The State of AI Report.
- Rowan Zellers et al.(2019.3.26.), From Recognition to Cognition: Visual Commonsense Reasoning, arXiv:1811.10830.
- Sebastian Borgeaud et al.(2022.2.7.), Improving language models by retrieving from trillions of tokens, arXiv:2112.04426v3.
- Tadas Baltrusaitis et al.(2019.2.), Multimodal Machine Learning: A Survey and Taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence.
- Ting Chen et al.(2020.6.1.), A Simple Framework for Contrastive Learning of Visual Representations, arXiv:2002.05709v3.
- Yuqi Huo et al.(2021.3.11.), WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training, arXiv:2103.06

## ○ 웹사이트

<http://www.empathic-project.eu/>.

<https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>.

<https://ai.googleblog.com/2022/06/minerva-solving-quantitative-reasoning.html>.

<https://cambrian-ai.com/ai-hardware-harder-than-it-looks/>.

<https://openai.com/blog/clip/>.

<https://openai.com/blog/dall-e/>.

<https://owkin.com/machine-learning/multimodal-data-integration/>.

<https://precisionhealth.uahs.arizona.edu/multimodal-sensor-systems-precision-health-enabled-data-harnessing-artificial-intelligence-and>.

<https://www.darpa.mil/program/active-interpretation-of-disparate-alternatives>.

<https://www.youtube.com/watch?v=1Po0-rvPqmk>.

---

## 저자소개

**이승민** ETRI 지능화융합연구소 기술정책연구본부 기술전략연구센터 책임연구원  
e-mail: todtom@etri.re.kr Tel. 042-860-1775

---

## 보고서 작성에 도움을 주신 분

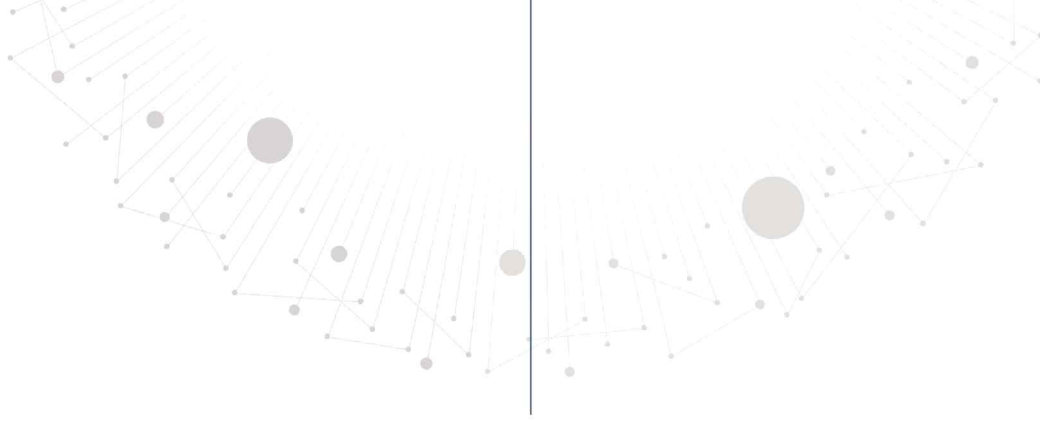
**임수종** ETRI 인공지능연구소 지능정보연구본부 언어지능연구실 책임연구원  
e-mail: isj@etri.re.kr Tel. 042-860-1297

---

## 기술정책연구본부 기술정책 이슈

**발행인** 이 지 형  
**발행처** 한국전자통신연구원 지능화융합연구소 기술정책연구본부  
**발행일** 2022년 8월 15일





[www.etri.re.kr](http://www.etri.re.kr)

본 저작물은 공공누리 제4유형:

출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.



**ETRI** Electronics and Telecommunications  
Research Institute

34129 대전광역시 유성구 가정로 218  
TEL.(042) 860-6114 FAX.(042) 860-6504

