DIGITAL INSIGHT

Digital Insight 2023

NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY

DIGITAL INSIGHT

파운데이션 모델의 이해와 미래 전망

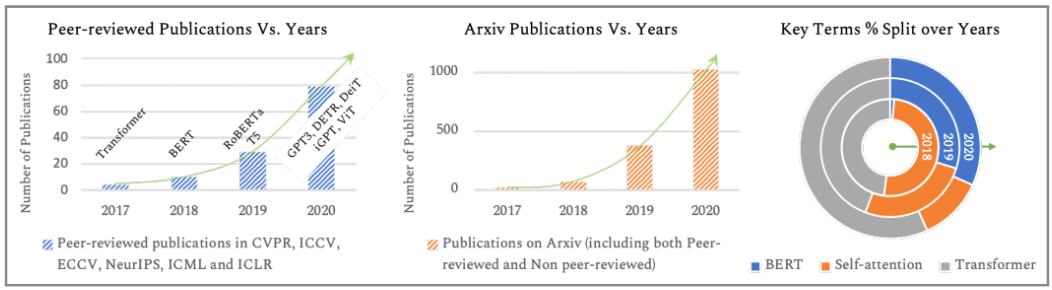
NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY

Contents

1 파운데이션 모델의 이해

딥 러닝 모델의 등장으로 기존 인공지능으로 다루기 어려웠던 언어, 컴퓨터 비전, 음성 등의 비정형 데이터 분야에서도 인공지능 시스템을 구축할 수 있게 되었다. 딥 러닝 시스템은 번역, 기계 독해, 문장 생성, 객체 탐지, 이미지 생성, 음성 인식, 음성 합성 등 다양한 분야에서 뛰어난 성능을 보였으며, 이러한 능력을 이용한 자동 번역, 검색 시스템, AI 챗봇, STT(Speech To Text), TTS(Text To Speech), 자율주행 시스템 등과 같은 다양한 어플리케이션들이 우리 사회 곳곳에서 서비스되고 있다.

한편, 지난 수 년 사이에 딥 러닝 시스템에는 큰 변화가 일어났다. 먼저, 2017년 처음 등장한 트랜스포머(Transformer) 모델[Vaswani et al. 2017]은 그 이전까지 딥 러닝 모델로 사용되어 오던 합성곱 신경망(CNN)과 순환 신경망(RNN) 모델을 대체하고 있다. NVIDIA에서는 2022년 4월 기준, 지난 2년간 아카이브(arXiv)에 게재된 AI 관련 논문의 70%에 트랜스포머가 등장하고 있다고 보고했으며, 이는 전위적인 변화인 셈이라고 평가했다.[1)]



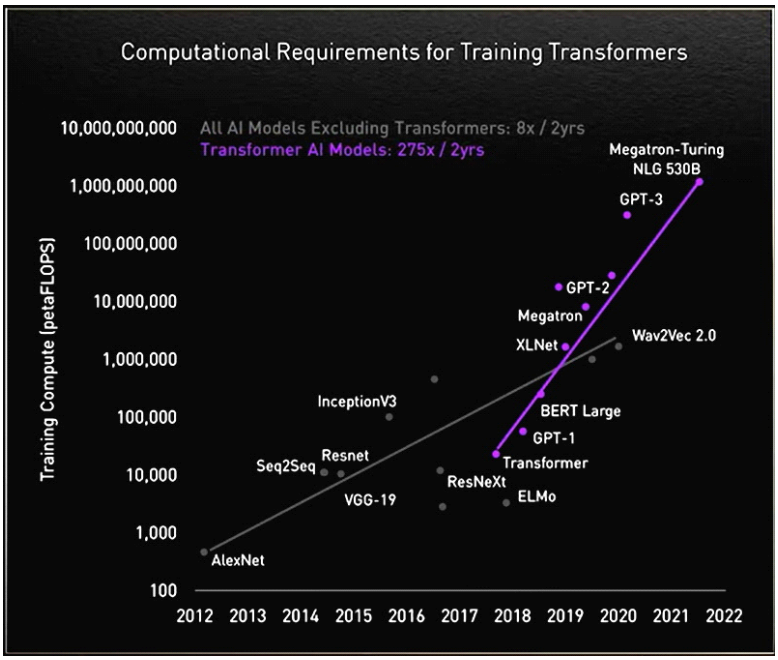
[그림 1] 2017년 이후의 트랜스포머 및 자기 지도학습 관련 연구 동향
(출처: Transformers in Vision: A Survey, 2021)

그리고, 트랜스포머 모델의 등장과 비슷한 시기에, 파운데이션 모델(Foundation model)이라고 불리는 새로운 패러다임이 적용된 모델이 등장했다. 방대한 양의 데이터를 자기 지도학습(Self-supervised learning)을 통해 학습한 후, 원하는 작업에 맞추어 미세 조정(Fine-tuning)을 하는 파운데이션 모델 특유의 학습 방법은 성능 향상 뿐 아니라 범용적인 인공지능 시스템의 탄생을 예견했다.

예를 들어, BERT[Devlin et al. 2018]의 경우 위키피디아 문서 등 수십억 단어 규모의 방대한 데이터로 사전 학습(Pre-training)시킨 단 하나의 모델에 비교적 짧은 시간동안 미세 조정을 수행한 것 만으로도 8개의 서로 다른 Task를 합쳐놓은 GLUE 벤치마크 및 질의 응답 데이터 셋인 SQuAD 등에서 모두

1) https://blogs.nvidia.co.kr/2022/04/01/what-is-a-transformer-model/

최고 수준 혹은 그에 준하는 성능을 보였다. 심지어 최신 파운데이션 모델인 GPT-3[Brown et al. 2020] 등은 미세 조정조차 없이, 수행해야 할 작업에 대한 적절한 예시를 주는 Few-shot learning 및 설명을 약간 추가하는 Prompt learning 만으로도 모델이 한번도 학습해본 적 없는 작업을 수행해 낼 수 있음을 보였다. 파운데이션 모델이 보여주는 보편적인 데이터 이해 능력과 대규모 데이터 학습 능력은 인공지능의 생태계를 변화시키고 있는 것이다.



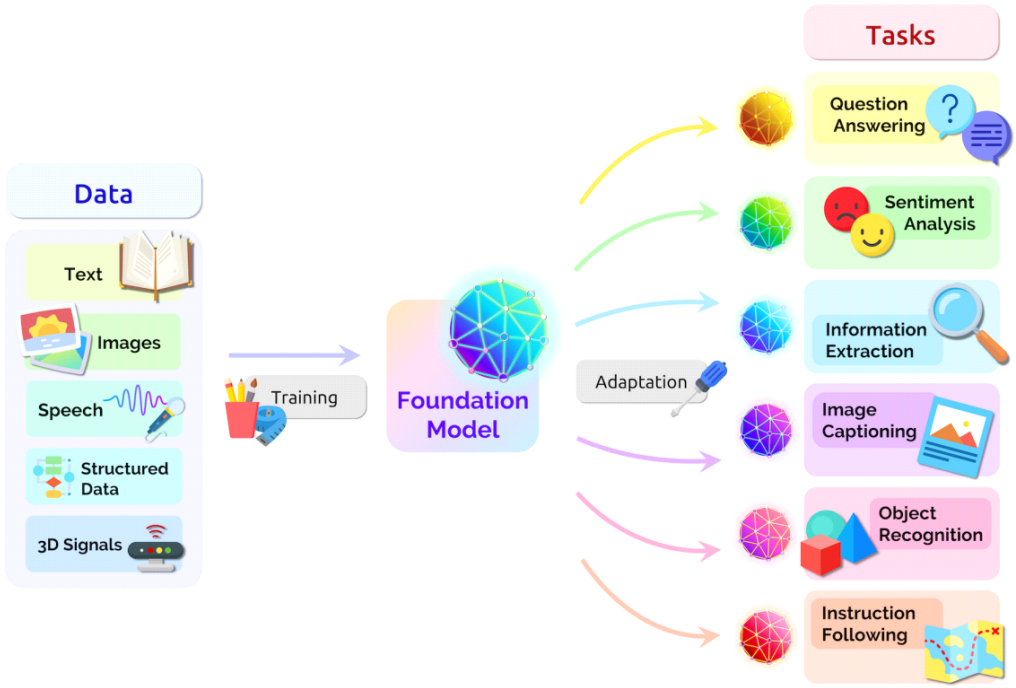
[그림 2] 시간에 따른 딥 러닝 및 트랜스포머 모델 학습에 필요한 연산량
(출처: https://blogs.nvidia.co.kr/2022/04/01/what-is-a-transformer-model-2)

하지만, 최근의 파운데이션 모델은 사전 학습된 모델을 생성하기 위해 수천 개의 GPU로 수 일~수십 일 가까히 학습시켜야 하기 때문에 학교나 소규모의 연구조직에서는 다루기 힘들며, 구글, 페이스북 등 산업계를 통해서 연구되고 있는 경향을 보인다. 또한 수없이 많은 딥 러닝을 이용한 어플리케이션들이 불과 수십 개의 파운데이션 모델을 기반으로 구축되고 실제 서비스되고 있기 때문에, 하나의 파운데이션 모델에서 발생된 결함이 그 모델을 사용하는 모든 어플리케이션에 그대로 표출될 수 있다는 문제가 있다.

스탠포드 대학교의 파운데이션 모델 연구 센터(CRFM, Center for Research on Foundation Models)은 2021년, “On the opportunities and risks of foundation models”[Bommasani et al. 2021]라는 보고서를 통해 파운데이션 모델이 불러올 거대한 패러다임 변화를 설명하고, 그로 인해 발생하게 될 새로운 응용 분야 및 위험 요소들을 분석한 바 있다. 본 저자들은 이러한 파운데이션 모델의 특성을 잘 이해하고, 파운데이션 모델의 국내·외 연구 동향 및 응용 분야에 대해 파악하고자 한다.

A 파운데이션 모델의 정의

파운데이션 모델이라는 용어를 처음 제안한 스탠포드 CRFM에서는 파운데이션 모델이라는 이름을 붙인 이유를 설명하며, 특유의 완성되지 않은 채 배포되는 특성을 언급한다. 즉, 파운데이션 모델은 방대한 양의 폭넓은 데이터를 사용하여 자기 지도학습을 통해 방대한 내부 파라미터를 지닌 모델을 학습시킨 후, 아직 명확하게 수행해야 할 작업이 정해지지 않은 상태로 배포되어, 사용자가 원하는 목적에 맞게 다운스트림 작업에 대해 미세 조정이나 문맥내 학습(In-context learning) 등과 같은 과정을 거침으로써 완성되는 기초 모델이라고 볼 수 있다. 그 외에, 방대한 양의 데이터를 학습시키기 위하여 GPU와 같은 하드웨어 장치가 필요하고, 병렬적인 학습을 가능하게 만든 트랜스포머 아키텍처를 사용한다는 공통점이 있다.



[그림 3] 파운데이션 모델이란 대용량의 폭넓은 데이터로 학습되어 목적에 맞는 다운스트림 작업에 적용될 수 있는 기초 모델을 의미한다.
(출처 : On the opportunities and risks of foundation models, 2021)

CRFM은 보고서에서 파운데이션 모델이라는 개념이 기존에 사용되어온 용어인 사전학습 모델(pre-trained model), 자기 지도 모델(self-supervised model)이 가리키는 대상을 포함하지만, 파운데이션 모델의 능력과 영향력은 기술적인 범주에서만 머물지 않기 때문에 별도의 용어 정의가 필요하다고 주장했다. 보고서에서는 파운데이션 모델의 예시로써 BERT, GPT-3, CLIP[Radford et al. 2021]을 들었는데, 모두 앞서 언급된 대로 자기 지도학습 기반으로 사전학습 된 모델이다.

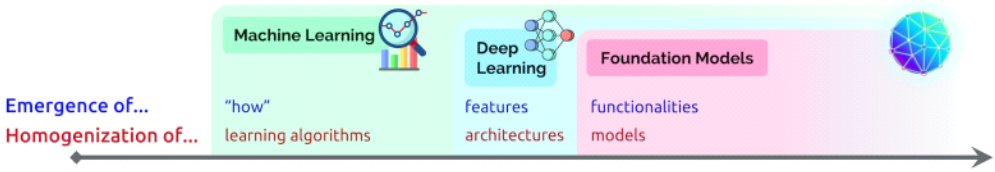
파운데이션 모델 구축에 사용되는 핵심 방법론인 자기 지도학습은 전이 학습(transfer learning)의 한 형태이며, 전이 학습은 어떠한 작업을 통해 학습한 추상적인 지식을 다른 작업에 활용하는 접근 방식을 의미한다. 전이 학습 방법론 자체는 파운데이션 모델 이전부터 많이 사용되어 왔고, 특히 이미지 분야에서는 대규모 이미지 분류 데이터 셋인 ImageNet을 통해 모델을 사전학습시킨 후 사용자 데이터를 통해 학습시키는 방법이 널리 사용되어 왔다. 그러나 파운데이션 모델에서 사용되는 방대한 데이터를 활용한 자기 지도학습의 가장 큰 특징은 데이터의 어노테이션 값을 스스로 설정한다는 데에 있다. 즉, 문장의 다음에 올 단어를 예측하는 Autoregressive language modeling, 데이터의 누락된 부분을 복원하는 Masked language modeling 등과 같이 별도로 사람에 의해 분류 및 회귀 목표가 정해지지 않더라도 스스로 데이터의 특성을 학습할 수 있는 방법들이 사용되기 때문에, 방대한 양의 데이터를 사용할 수 있으며 보다 일반적이고 확장성이 높다.

압도적인 규모의 데이터와 모델 크기로부터 얻어지는 데이터(이미지, 텍스트) 자체에 대한 이해 능력 역시 파운데이션 모델을 정의하는 주요한 요소이다. 트랜스포머 모델의 등장 이전에 존재했던 Word2Vec[Mikolov et al. 2013]과 같은 단어 임베딩 모델 및 SA-LSTM[Dai and Le, 2015]과 같은 자기 지도학습 기반 언어 모델 역시 자기 지도학습을 활용했기 때문에 BERT 이후에 등장하게 되는 파운데이션 모델들의 일반적인 특징을 일부 공유하지만, 동시에 일부 차이를 보인다. 예를 들어, Word2Vec 모델은 문서 분류[Lilleberg et al. 2015], 객체명 탐지[Sienčnik, 2015] 등 단어 임베딩 기능을 활용할 수 있는 다양한 다운스트림 작업에 활용되었으나, 많은 경우 사전학습된 모델을 활용하는 대신 사용자가 직접 준비한 데이터 셋에 대해 학습하는 방식으로 활용되었다. 이 경우 자기 지도학습의 효과는 데이터에 대한 일반적인 이해 능력이 아닌 해당 데이터 셋에 대한 이해로 국한되기 때문에, 하나의 사전학습된 모델이 무수히 많은 다운스트림 태스크에 사용되는 파운데이션 모델과는 확장성 및 활용성 측면에서 차이를 보인다.

B 파운데이션 모델의 특징

앞서 언급한 바와 같이, 이 모델들은 텍스트 혹은 이미지 데이터 자체에 대한 일반적인 이해를 획득하기 위하여 방대한 양의 데이터를 자기 지도학습을 기반으로 학습하였다. 예를 들어, BERT의 경우 BookCorpus[2] 데이터 셋 및 위키피디아 영문 문서를 합쳐 총 33억 단어로 이루어진 데이터 셋을 구축한 후, 학습에 이용하였다. 심지어 GPT-3를 학습시키는 데에는 약 5,000억 단어 토큰에 가까운 텍스트 데이터가 사용되었으며, 텍스트와 이미지를 연결짓는 모델인 CLIP은 4억 개의 (text, image) 쌍을 이용하여 학습되었다. 이러한 방대한 양의 데이터 셋은 주로 웹을 통해 수집되며, 데이터의 편향성을 제거하기 위한 전처리 작업이 포함될 수 있으나 사람에 의해 일일히 어노테이션이 이루어지지는 않는다.

이렇게 구축된 데이터로부터 일반적인 데이터 이해 능력을 학습하기 위하여, 모델의 크기 역시 비약적으로 커진다. 파운데이션 모델의 내부 파라미터는 적게는 수 억개(GPT-1, 110M)부터 1조 개(Switch Transformer, 1.6T)가 넘는 모델까지 등장했다. 이러한 모델을 학습시키기 위해서는 GPU와 같은 병렬 연산 처리 하드웨어가 필수적이며, 최대 수백~수천 개의 GPU를 클러스터링 하고, 수 일~수십 일에 걸쳐서 학습되기도 한다. 그러나 한 번 구축된 사전학습된 모델은 다양한 다운스트림 작업에 사용될 수 있으며, 이 과정에서 필요한 미세 조정의 경우 불과 수 시간 정도의 학습으로 충분하거나, 혹은 학습이 전혀 필요 없는 Zero-shot 추론으로 수행되기도 한다.



[그림 4] 창발성(emergence)과 균일화(homogenization)에 따른 인공지능 모델의 흐름
(출처: On the opportunities and risks of foundation models, 2021)

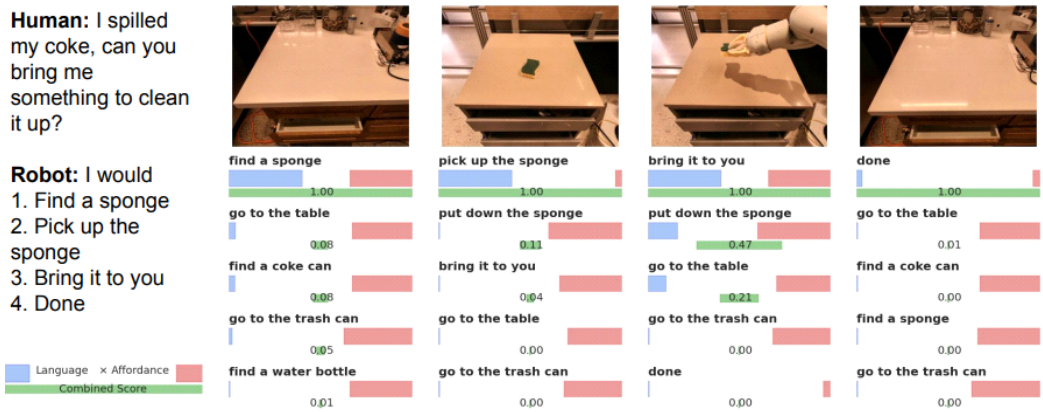
이러한 방대한 양의 학습은 파운데이션 모델에게 두 가지 능력을 부여한다. CRFM은 이를 창발성(emergence)과 균일화(homogenization)로 설명한다. 창발성이란 모델이 스스로 어떠한 문제를 해결하기 위한 지식을 도출하는 능력을 의미한다. 머신 러닝 모델은 학습이라는 개념을 통해 데이터로부터 문제를 해결하는 방법을 도출할 수 있었다. 그리고 딥 러닝 모델에 이르러서는, 더욱 더 방대해진 데이터 셋과 깊어진 신경망 구조를 통해 스스로 데이터의 feature를 도출하여 feature map을 구성하는 능력을 갖추게 되었다. 이러한 능력은 딥 러닝 모델에 인위적으로 주입되는 것이 아니라, 모델이 주어진 학습 방법과 모델 아키텍처 내에서 데이터를 기반으로 스스로 도출해내는 것이다. 이미지나

2) https://github.com/soskek/bookcorpus

텍스트와 같이 단순 정규화 이상의 feature engineering이 필요했던 비정형 데이터 분석에 있어서 이러한 딥 러닝 모델의 창발성은 큰 도움을 주었고, ImageNet 분류와 같은 문제에서 놀라운 성과를 거둘 수 있었다[Krizhevsky et al. 2012]. 파운데이션 모델에 이르러서는, 모델은 스스로 수행해야 할 작업을 파악하고 해당 작업을 해결하는 데 필요한 지식을 도출할 수 있게 되었다. 예를 들어, GPT-3의 경우 작업에 대한 설명을 프롬프트 형태로 제공하는 것만으로도 텍스트 분류, 추론, 질의응답 등 다양한 작업을 수행할 수 있음을 보였다. 이러한 창발성은 연구자들에 의해 입력되거나 학습 데이터로써 주어지지 않았으며, 심지어 연구자들이 미리 예견했던 능력도 아니었다.

균일화란 모델이 점차 일반화된 지식을 도출해낼 수 있게 됨에 따라, 하나의 뛰어난 모델이 적용될 수 있는 범위가 점차 확대되며 더욱 보편적이고 범용적으로 활용되는 현상을 의미한다. 머신 러닝의 단계에서는, 서로 다른 문제를 가진 사용자들은 각 문제에 맞는 알고리즘을 고려하는 대신 수십 개 내외의 모델 구조와 학습 알고리즘 중 하나를 선택하여 각자의 데이터를 학습시켰고, 이는 인공지능 분야에서 알고리즘의 균일화를 이끌어냈다. 딥 러닝 모델의 경우 이미지 분야 등에서 사용되던 각종 컨볼루션 필터, 텍스트 분야에서 사용되던 TF-IDF 등의 단어 표현 방법 등 다양한 feature engineering을 통해 데이터로부터 feature를 얻는 파이프라인을 설계하는 대신 더 많은 양의 데이터로부터 feature map을 직접 학습하는 방법을 취함으로써 특성이 다르더라도 동일한 데이터 형태에 대해서는 동일한 신경망 모델을 이용할 수 있게 균일화를 이루어주었다. 파운데이션 모델의 경우, 모델 그 자체가 하나의 균일화를 이루어냈다. 2022년 12월 현재 각종 트랜스포머 기반 모델과 미세 조정된 모델을 수집하고 배포하고 있는 Huggingface에는 미세 조정까지 완료된 모델을 포함하여 약 97,000개의 모델이 저장되어 있으나, 그들의 기반이 되는 파운데이션 모델 아키텍처의 종류는 165종에 불과하다.

또한, 파운데이션 모델은 기술적 측면에서 볼 때, 트랜스포머 아키텍처 구조를 기반으로 하고 있다는 특징을 갖는다. 트랜스포머 아키텍처는 텍스트, 이미지, 음성, 정형 데이터, 단백질의 염기서열 분석, 분자 구조 분석, 강화학습 등의 서로 다른 작업에 대해서 광범위하게 활용되고 있을 뿐만 아니라, 여러 종류의 서로 다른 모달리티에 걸친 데이터를 다루는 멀티 모달 모델의 형태로도 발전하고 있다. 예를 들어 CLIP의 경우 텍스트를 처리하기 위한 Transformer 모델과 이미지를 처리하기 위한 ViT 모델이 함께 포함되어, 이미지를 분류하기 위해 이미지-레이블 쌍을 학습하는 것이 아닌 레이블에 해당하는 단어의 의미를 이해하여 Zero-shot 분류를 수행하게 된다. PaLM-SayCan 모델의 경우 사용자가 로봇에게 텍스트로 명령을 내리면 이를 해석하여 수행해야 할 행동을 찾고, 로봇을 움직여서 순차적으로 해당 동작들을 수행할 수 있게 한다. 이를 위해 파운데이션 모델인 PaLM[Chowdhery et al. 2022]이 사용되었다.



[그림 5] PaLM-SayCan 모델은 텍스트로 명령을 입력하면 스스로 수행할 행동을 찾고, 이를 수행한다.
(출처: Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, 2022)

C 제반 기술

파운데이션 모델이 보여주는 뛰어난 창발성 능력과 균일화 능력은 고성능 컴퓨팅을 바탕으로 방대한 양의 데이터와 이를 학습시킬 수 있는 시스템, 그리고 자기 지도학습 기법, 모델 아키텍쳐 구조 등에서 나온다.

● 대용량 학습데이터 구축

많은 경우, 파운데이션 모델의 벤치마크에서 학습 데이터의 양은 모델의 아키텍처, 크기, 학습 시간만큼이나 성능과 밀접한 관계를 갖는 요소이다. 그러므로 어노테이션 과정이 필요 없는 자기 지도학습의 특성상, 파운데이션 모델을 구축하기 위해서는 웹 크롤링 등의 방법으로 최대한 많은 양의 데이터를 확보하는 것이 중요한 과제이다. 예를 들어, BERT의 경우 약 8억 단어로 구성된 BookCorpus 데이터 셋 및 약 25억 단어로 구성된 위키피디아 영문 문서를 수집하여 학습에 이용하였다. GPT-3의 경우는 CommonCrawl[3] 데이터 셋을 중심으로 약 5,000억 단어 토큰에 가까운 텍스트 데이터를 수집하여 학습에 이용하였으며, 텍스트 이미지 연결 모델인 CLIP은 4억 개의 (text, image)쌍을 이용하여 학습되었다. 한국어 언어 파운데이션 모델의 경우 위키피디아 한국어 문서 데이터, 크롤링 된 뉴스 기사[Lee et al. 2020], 나무위키 등 한국어 위키의 문서 데이터[4] 등이 대표적으로 사용된다.

이 과정에서 한 가지 중요한 고려 사항은 데이터를 선정하는 원칙이다. 사람이 직접 학습 데이터를 생성하지 않는 만큼, 잘못된 원칙으로 수집된 데이터는 모델로 하여금 인종 차별, 혐오 발언, 가짜 뉴스 등 허위 정보를 생성하게 될 수 있으며, 무수히 많은, 그리고 예상할 수 없는 다운스트림 태스크에 활용될 수 있는 파운데이션 모델의 특성상 이러한 한계점은 치명적으로 작용할 수 있다. 따라서 어떤 데이터를 수집할 것인지 원칙을 정하는 데이터 거버넌스 역시 중요한 과제이다. 예를 들어, 언어 모델인 BLOOM[Scao et al. 2022]을 구축하기 위해 BigScience 연구진들은 2021년 12월부터 3월까지 총 1.5TB의 텍스트 데이터가 포함된 초거대 데이터 셋을 구축하였지만, 그 중 65%에 달하는 데이터들은 연구자들이 신중하게 결정한 데이터 수집 지점으로부터 수집되었으며, 개인 정보 필터 등 다양한 종류의 필터와 데이터 품질 지표 등을 구축한 후 크롤링으로 수집된 데이터 역시 평가를 거쳐 반자동으로 데이터를 정제하는 과정을 수행하였다.

● 자기 지도학습 기법

방대한 양의 데이터가 데이터의 구조 파악 및 추상적인 이해를 돕고, 지식 도출을 가능하게 만드는 원인이라면, 방대한 데이터에 내재되어있는 지식을 모델에 주입시킬 수 있는 방법은 자기 지도학습을 통해 구현된다. 파운데이션 모델의 등장 이전에도 ImageNet과 같은 대규모 이미지 데이터 셋으로

3) https://commoncrawl.org/the-data/
4) https://github.com/monologg/DistilKoBERT

사전학습된 모델의 이미지 해석 능력을 활용하는 사례는 있었다. 그러나 ImageNet은 1000개의 클래스에 해당하는 이미지만 수집된 탓에, 해당 데이터로 학습된 모델은 다른 종류의 분류 기준이 주어질 경우 활용성이 떨어지는 사례가 보고되기도 했다[Radford, 2021].

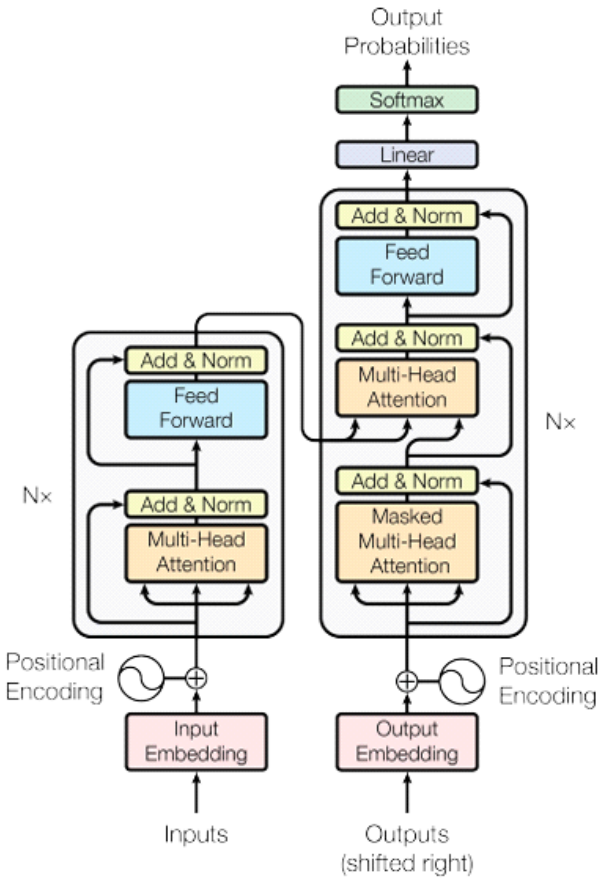
사전학습이라고도 불리는 파운데이션 모델의 자기 지도학습 기법은 데이터의 구조와 추상적 지식을 획득하기 위하여, 스스로 학습할 수 있는 대리(surrogate) 작업을 정의하는 것에서 시작된다. 초창기의 자기 지도학습을 이용한 사전학습 기법은 언어 모델에서 생겨났으며, Autoregressive language modeling이라고 불리는 주어진 문장의 다음에 올 단어를 고르는 작업으로 학습되었다. GPT-1과 같은 모델들은 이러한 작업을 학습한 모델이 학습을 하지 않은 그대로의 모델보다 다양한 다운스트림 작업에서 더 높은 성능을 보였으며, 미세 조정에 필요한 학습 시간 역시 감소하였음을 보였다. 또한, Masked language modeling이라고 하는 새로운 사전학습 방법이 등장했다. BERT 모델은 문장의 한 가운데를 비운 뒤, 주변 문맥만을 보고 해당 빈 칸에 들어갈 단어를 예측하는 방법으로 언어 모델을 학습시켰다. 이 방법은 이미지 생성 모델에도 이어져, 다양한 파운데이션 모델 기반 이미지 생성 모델들이 이미지를 여러 개의 패치 조각으로 나누고, 그 중 일부를 가린 후 원래 패치를 맞추게 하는 대리 모델을 통해 이미지의 형태와 구조를 학습시킬 수 있음을 보였다[Dosovitskiy et al. 2020, Ramesh et al. 2021].

● 트랜스포머 아키텍처

언어, 이미지, 소리 등 데이터의 형식을 막론하고 거의 대부분의 파운데이션 모델은 트랜스포머 아키텍처를 사용하고 있다. 이는 트랜스포머 모델이 가진 셀프 어텐션(self-attention) 기법이 가지고 있는 우수한 시퀀스 데이터 분석 능력과 함께, 시퀀스 데이터에 대한 병렬적인 학습이 가능해졌기 때문이다. 트랜스포머 모델의 시작은 2017년, 영어-독일어 기계 번역 분야에서 가장 성능이 좋은 모델, state-of-the-art 모델로 꼽히면서 주목을 받았다. 이 당시 트랜스포머는 인코더-디코더 구조를 가지고 있었으며, 자연어 문장과 같은 시퀀스 데이터를 입력으로 받고 시퀀스 데이터를 출력하는 시퀀스-투-시퀀스 모델이었다.

기존에 가장 널리 사용되던 시퀀스-투-시퀀스 모델인 RNN 모델의 경우, 이전 시퀀스에 대한 모델의 출력이 다음 시퀀스에 대한 예측에 사용되어야하는 구조로 인해 시퀀스 내에서 서로 거리가 떨어진 값 끼리는 영향을 주기 어렵고, 또한 추론 및 학습에 걸리는 시간 역시 시퀀스 길이에 비례해서 올라갔다. 다른 대안이었던 CNN 모델은 RNN처럼 순차적인 구조를 가지지는 않았지만, 그 구조의 특성상 커널의 크기가 시퀀스의 길이보다 짧을 경우 모든 입력값 간의 연관관계를 추론할 수 없었고, 커널의 크기를 시퀀스 길이보다 늘릴 경우 모델의 크기가 지나치게 커지는 문제가 있었다.

트랜스포머는 포지셔널 인코딩(positional encoding)을 통해 시퀀스 내의 순서 정보를 시퀀스 데이터에 합하고, 셀프 어텐션을 통해 값 간의 거리를 초월한 연관관계를 추론할 수 있도록 만들었다. 이를 통해 시퀀스 데이터를 병렬적으로 학습시키고, 추론할 수 있게 만들었으며 이는 학습에 필요한 연산량의 감소로 이어졌다. 언어 모델인 GPT와 BERT는 트랜스포머 모델이 거둔 성공과 뛰어난 연산 효율성을 이용하여 모델의 크기를 거대화하는 데 성공하였으며, 기계 번역에 국한되지 않는 모델을 구축하기 위해 인코더-디코더 구조를 포기하고 각각 트랜스포머 디코더와 인코더만을 사용하여 모델을 구축하였다.

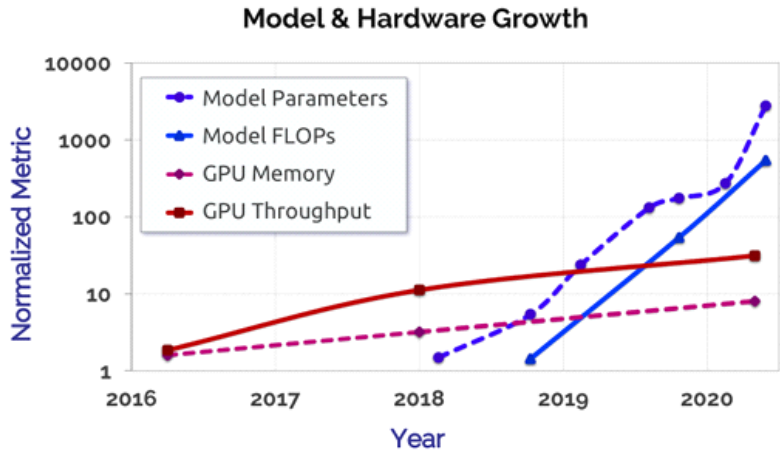


[그림 6] 트랜스포머 아키텍처
(출처: Attention is all you need, 2017)

한편, 이미지와 같은 데이터 형식은 텍스트 데이터와 다름에도 불구하고 트랜스포머의 스케일링 능력으로 인해 좋은 이미지 추론 능력을 보여준다. 예를 들어, DALL-E[Ramesh et al. 2021] 모델의 경우 256x256 크기의 이미지를 32x32 사이즈에 8192 종류의 값을 가질 수 있는 “이미지 토큰”으로 압축한 후, 이를 마치 8192개의 단어가 사용될 수 있는 32x32=1024 길이의 시퀀스 데이터인 것처럼 해석하여 이미지를 시퀀스 데이터로 변형시켰다. ViT[Dosovitskiy et al. 2020] 모델은 이미지 토큰 대신 이미지를 16x16 크기로 쪼갰다는 점에서 차이가 있지만, 역시나 이미지를 시퀀스 데이터로 해석했다는 점은 동일하다.

● 컴퓨팅 성능의 향상

트랜스포머 아키텍처가 가진 강력한 스케일링 성능 덕분에 파운데이션 모델의 크기는 그 성능과 함께 기하급수적으로 향상되었으며, 1조 개가 넘는 파라미터를 가지는 최신 대형 파운데이션 모델의 크기는 인간의 뇌(대략 1,000억 개의 뉴런과 100조 개의 시냅스)에 가까워지고 있다. 이러한 모델을 방대한 데이터 셋에 대해 학습시키기 위해서는 컴퓨팅 성능의 확보가 반드시 필요하다.



[그림 7] 파운데이션 모델의 크기와 GPU의 성장
(출처: On the opportunities and risks of foundation models, 2021)

대부분의 파운데이션 모델은 CPU만을 이용할 경우 추론하는 시간조차 오래 걸리게 되며, 학습은 거의 불가능에 가깝다. 따라서 병렬 연산에 최적화된 GPU, 텐서 연산에 최적화된 TPU 등 다양한 보조 연산 하드웨어를 통해 연산을 가속시킬 필요가 있으며, 이러한 하드웨어들을 클러스터링하여 거대한 학습 시스템을 구축하는 것이 중요하다. 예를들어, OpenAI의 경우 GPT-3를 학습시키기 위해 285,000대의 CPU와 10,000대의 GPU를 연결한 슈퍼 컴퓨터를 활용했다고 하며, 이는 하나의 GPU만으로 동일한 모델을 학습시키려면 사전학습에 약 288년의 시간이 걸릴 것이라고 한다.

따라서 파운데이션 모델을 구축하기 위해서는 이러한 컴퓨팅 성능을 활용하기 위해 Pytorch, Tensorflow 등의 라이브러리를 이용하게 되며, 파라미터가 수십억 개 이상이 되는 경우에는 Megatron-LM[Shoeybi et al. 2019] 및 DeepSpeed[Rasley et al. 2020]와 같은 데이터 파이프라인, 텐서 슬라이싱 병렬화 등의 기능이 포함된 학습 시스템이 필요해진다. 국내에서는 HyperClova[Kim et al. 2021] 모델 개발에 A100 GPU 1,024대를 연결한 DGX 시스템에서 megatron-LM을 활용하여 대규모 학습을 수행한 사례가 있다.

2 파운데이션 모델 연구 동향

본 장에서는 파운데이션 모델에서의 관련 연구 동향을 살펴보고자 한다. 파운데이션 모델을 구축하기 위해서는 막대한 컴퓨팅 자원이 소요되는 대규모 학습이 필요한 만큼, 구글, 페이스북, 마이크로소프트, 화웨이 등의 대기업 및 OpenAI, AI21 Labs등의 인공지능 스타트업을 중심으로 연구되고 있으며, 학계의 참여는 상대적으로 저조한 편이다. 국내에서는 한국어 이해 능력에 초점을 맞춘 한국어 언어 모델에 관한 연구가 활발히 이루어지고 있다.

A Prompt Engineering: GPT-3

2018년 OpenAI에서 공개한 GPT(Generative Pre-training) 모델은 비지도 사전학습을 통해 보편적인 언어 이해 능력을 학습시키고 서로 다른 작업에 대해 미세 조정만 수행함으로써 해당 작업에 맞는 모델로 변화하게 되는 작업-무관 모델(task-agnostic model)을 선보였다. 이를 위해 GPT는 모든 문장을 토큰 단위로 분해한 후 주어진 문장 다음에 올 토큰을 예측하는 AR(auto-regressive) 모델 혹은 인과적 언어 모델(causal language model)이라 불리는 방법을 사용하였다. 이를 식으로 표현하면 다음과 같다.

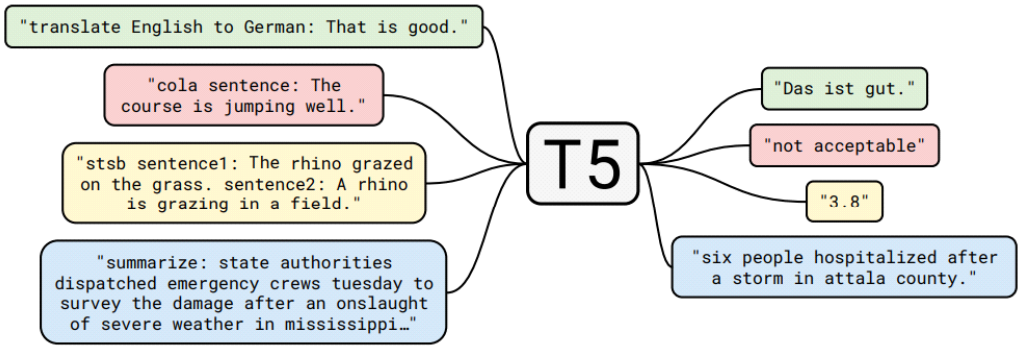
$$L_1(u) = \sum_i \log P(u_i \mid u_{i-k}, \ ..., \ u_{i-1}; \theta)$$

즉, i번째 위치의 바로 앞 토큰으로부터 k번째 앞까지의 토큰이 주어졌을 때, 모델의 파라미터 θ를 이용하여 i번째 위치에 정답 토큰인 가 등장할 확률을 구하는 모델이 된다. GPT는 이 확률 모델을 학습하기 위해 트랜스포머 디코더 계층만을 쌓아 올린 모델 구조를 사용했으며, 학습 데이터로는 BookCorpus 데이터 셋을 이용하였다.

2019년 공개된 GPT-2는 약 15억개의 파라미터를 가진 모델로써, 기존 GPT 모델보다 10배 이상 규모가 커지면서 더욱 강력한 문장 생성 능력을 활용하여 미세 조정을 전혀 수행하지 않는 Zero-shot setting을 사용하더라도 많은 작업들을 수행할 수 있음을 보였다. Zero-shot setting은 모델의 내부 파라미터를 전혀 업데이트하지 않은 채로 주어진 작업을 수행해야 하는 설정으로, 데이터는 커녕 작업의 내용조차 학습할 수 없기 때문에 연구자들은 직접 작업을 텍스트로 함께 모델에 입력하는 방법을 제안하였다.

예를 들어, 프랑스어를 영어로 번역하는 문제의 경우, “프랑스어 원문, translated to English: ”와 같은 형태로 텍스트를 입력하여 맥락상 자연스럽게 해당 문장을 영어로 번역하도록 유도하는 방식을 취하였다. 비록 GPT-2의 Zero-shot setting 성능이 기존 모델을 뛰어넘는 성과를 보여주지는 못했지만, 최대 크기였던 15억 개의 파라미터를 사용한 모델조차 여전히 출력하는 결과가 학습 데이터에 과적합

되지 않았음을 보임으로써 성능이 개선될 가능성과 함께 미세 조정이 필요없는 새로운 형태의 학습 방법의 가능성을 보여주었다.

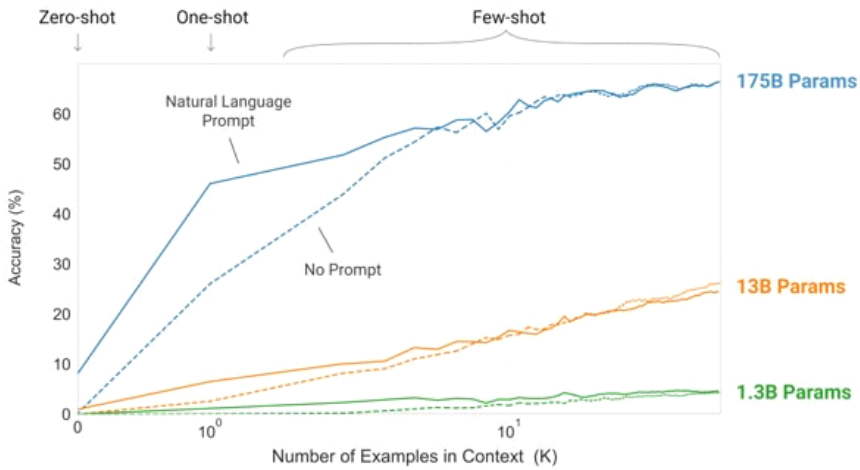


[그림 8] T5의 프레임워크

(출처: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2019)

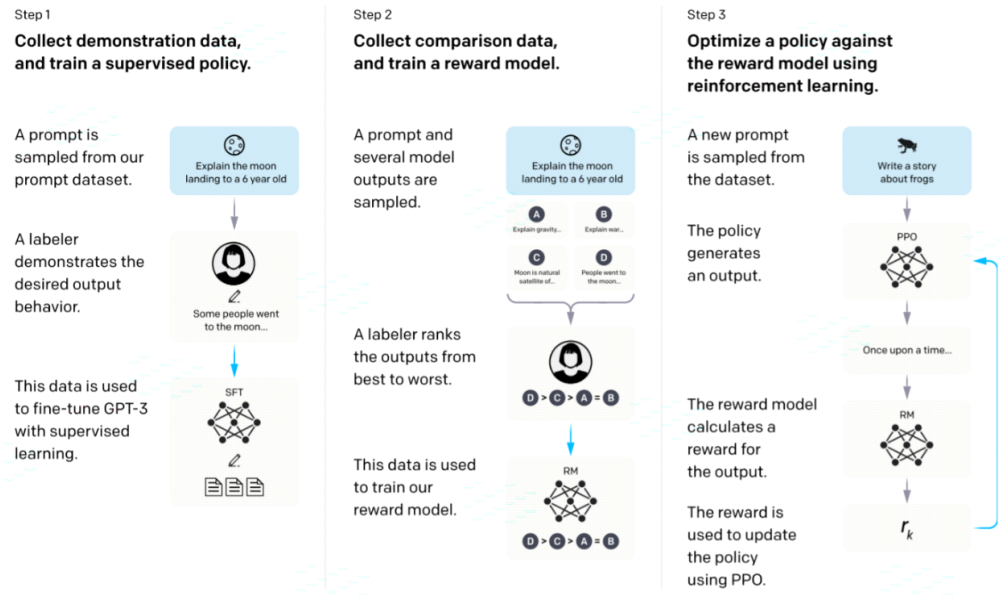
2019년 구글에서 공개한 T5 모델은 Text-to-text Transfer Transformer의 약자로써, 초기의 트랜스포머 모델과 마찬가지로 인코더-디코더 구조를 가지고 있다. T5 모델의 목표는 기계 번역, 문서 요약, 질문 응답, 분류 작업 등을 포함한 모든 자연어 관련 작업에서 텍스트 형태로 질의 문장을 앞에 붙이고, 정답 역시 텍스트 형태로 출력하도록 만듦으로써 미세 조정 없이 하나의 모델이 여러 개의 작업을 수행할 수 있도록 만드는 것이었다. 이를 위해 T5는 약 750GB에 달하는 영어 문장으로만 이루어진 C4(Colossal Clean Crawled Corpus) 데이터 셋을 사용하였고, 그 후 GLUE, SuperGLUE 등의 테스트 벤치마크의 학습 데이터 셋을 벤치마크별로 전부 합쳐 하나의 데이터 셋으로 미세 조정 학습에 사용하였다. 모델의 크기는 최대 110억 파라미터이다.

2020년 공개된 GPT-3 모델은 1,750억 개의 내부 파라미터를 가진 모델로써, 약 5,000억 토큰으로 이루어진 방대한 텍스트 데이터 셋으로 학습되었다. 이러한 방대한 학습의 결과로 인해 마침내 Zero-shot 및 few-shot setting에서 뛰어난 성능을 보여주는 것이 가능해졌다. 아래 그림은 단어에서 불필요한 기호를 제거하는 간단한 작업에 대하여 모델의 크기에 따른 zero-shot, few-shot 정확도를 보여주고 있다. 여기서 few-shot은 문제-정답 쌍을 모델이 학습하여 파라미터를 업데이트하는 것이 아닌, 입력 데이터 앞에 적절한 문제-정답 쌍을 제공하여 모델이 정답을 맞히도록 유도하는 예시의 수를 의미한다. 이 실험을 통해 모델의 크기가 클수록 성능이 향상된다는 것과, Zero-shot setting에 비하면 one-shot이, one-shot보다는 few-shot이 더 높은 정확도를 보인다는 것을 볼 수 있다.



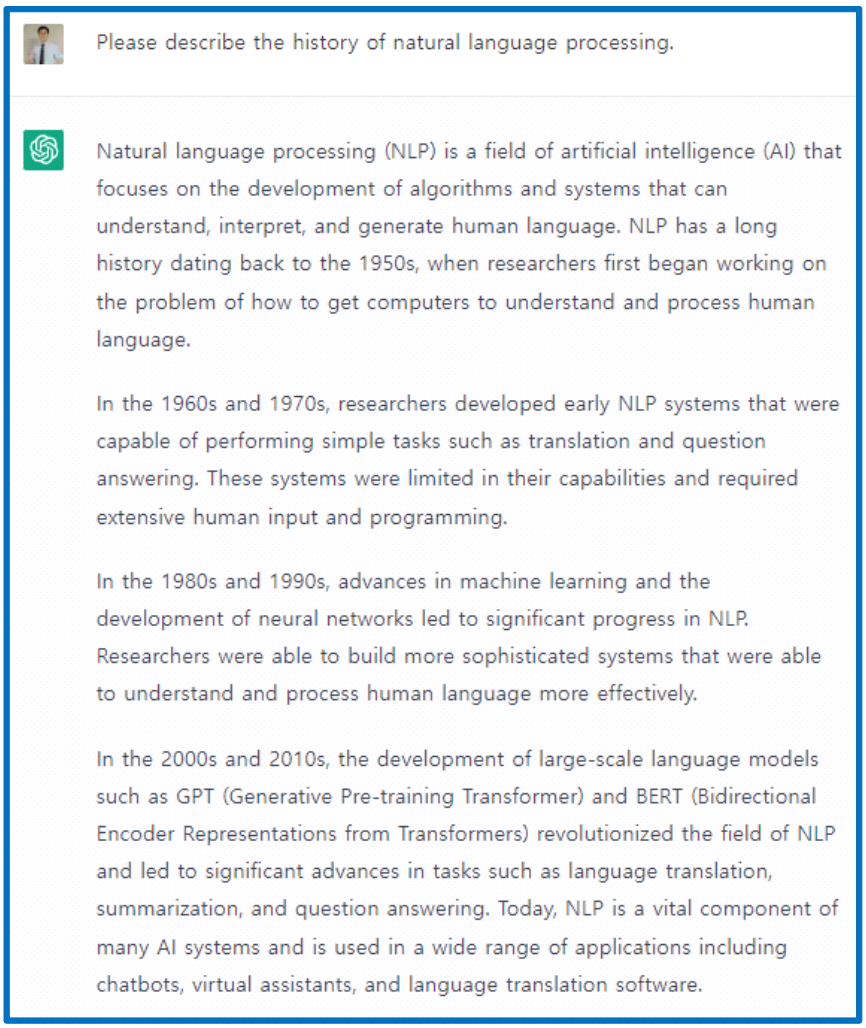
[그림 9] GPT-3 모델 크기 및 few-shot setting에 따른 성능 변화
(출처: Language models are few shot learners, 2020)

천억 개 이상의 내부 파라미터를 갖는 초거대 모델들은 일반 사용자나 개인 연구자가 활용하기에는 지나치게 거대한 모델이다. 따라서 방대한 데이터로 사전학습 후 파운데이션 모델의 형태로 배포되어, 개인 사용자가 각자의 목표 및 데이터에 맞춰 미세 조정을 하는 파운데이션 모델의 학습 패러다임을 적용하기에는 무리가 있다. 또한, OpenAI는 GPT-3가 가짜 뉴스 작성 등에 악용될 가능성을 시사하면서 GPT-3를 API의 형태로 제공할 뿐 내부 파라미터를 공개하지는 않았는데, 이렇듯 API 형태로 제공되는 서비스의 경우 미세 조정과 같은 파라미터 업데이트를 통한 학습은 불가능해진다.



[그림 10] InstructGPT 및 Chat GPT의 학습 방법
(출처: Training language models to follow instructions with human feedback, 2022)

이러한 상황에서 “적절한” 프롬프트를 설계를 통해 학습 없이 사용자가 원하는 작업에 최적화할 수 있는 프롬프트 엔지니어링의 중요성이 대두되었다. 대표적으로, GPT-3의 few-shot setting에 대하여 어떤 예시를 제공해야 모델이 더 잘 문제를 이해할 수 있는지에 대한 연구나, GPT-3보다 훨씬 작은 모델들에 서로 다른 프롬프트를 입력하여 나온 결과를 앙상블하는 연구 등이 있다. 또한 OpenAI는 기존의 causal language model이 가지고 있었던 다음 단어 예측 방법이 그 구조상 사실이 아닌 문장, 혐오 표현, 무의미한 문장을 생성하기 쉬웠음을 지적하며, 사람에 의해 채점된 응답의 적절성 데이터 및 강화 학습을 통해 “유의미하고 도움이 되는” 문장을 생성할 수 있도록 훈련한 InstructGPT를 2022년 3월 논문으로 공개하였고, 이 기술을 바탕으로 2022년 12월 ChatGPT 서비스를 공개하였다. 두 모델은 구조적으로 동일하며, 학습에 사용된 데이터의 차이가 있다고 알려져 있다.

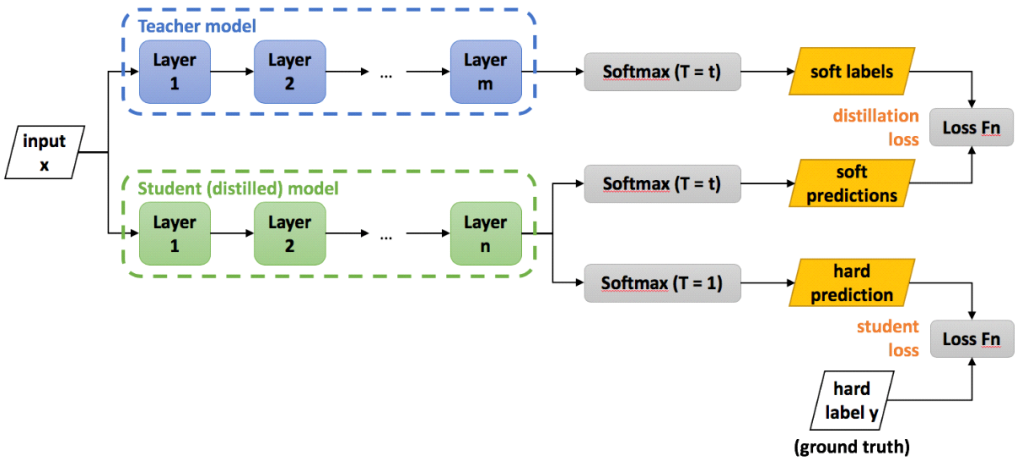


[그림 11] ChatGPT와의 채팅의 예

B 모델 최적화

비록 더 많은 데이터와 더 큰 모델 크기가 파운데이션 모델의 더 높은 추론 능력을 보여준다는 사실은 여러 연구를 통해 입증되었으나, 현실적인 제약 및 비용 문제는 큰 문제로 남아있다. 예를들어, GPT-3의 경우 사전학습을 위해 10,000대의 GPU를 활용하였으며, LambdaLabs는 GPT-3의 논문을 토대로 OpenAI가 GPT-3를 한 차례 학습시키기 위해 컴퓨팅 자원만으로도 약 460만 달러, 약 60억 원을 사용했을 것이라 추측하기도 했다[5]. 심지어 이렇게 막대한 양의 컴퓨팅에 소모되는 전력과 그로 인한 탄소 발생은 환경적으로도 간과할 수 없는 수준에 이르렀다[Strubell et al. 2019].
이러한 기술적, 경제적, 환경적 문제로 인해 파운데이션 모델 구축의 효율성을 높이는 최적화 기법들 역시 주목을 받고 있다. 모델 최적화 기법은 크게 지식 증류(Knowledge distillation), Pruning, 모델 구조 변경, 양자화(Quantization), Sparsity 등이 있다.

지식 증류는 이미 학습이 완료된 거대한 크기의 Teacher 모델의 "지식"을 비교적 작은 Student 모델에 증류하여 전수하는 개념으로, 2014 NIPS에서 처음 제안되었다. 지식 증류는 궁극적으로 Teacher 모델을 흉내내는 것이 목적이기 때문에 몇몇 연구(Park et al. 2019)를 제외하면 기존 Teacher 모델보다 뛰어난 성능을 얻는 것은 쉽지 않으며, 목표로 하지 않는 경우가 많다. 따라서, 학습 비용보다는 미세 조정 후 배포 단계에서 모델의 크기 줄이고 추론 속도는 향상시키고자하는 기법이라고 할 수 있다.



[그림 12] 지식 증류의 개념 설명
(출처 : https://intellabs.github.io/distiller/knowledge_distillation.html)

5) https://lambdalabs.com/blog/demystifying-gpt-3

지식 증류를 위해서는 Teacher 모델이 생성하는 예측값이 필요한데, 지도학습에서 학습 데이터가 가진 라벨은 어떤 하나의 값을 가지는 이진 데이터인 반면, teacher 모델이 출력하는 예측값은 0~1 사이의 확률값이 된다. 이를 soft label이라 칭하며, student 모델은 학습 데이터 자체를 학습하는 한편, 동시에 teacher 모델의 예측 확률분포를 출력하도록 학습된다. 자연어 처리 관련 기업인 HuggingFace에서는 BERT 모델의 사전학습 작업에 지식 증류를 적용하여 기존 BERT보다 약 40% 적은 크기와 60% 더 빠른 추론 속도를 보이면서도 BERT의 97%정도의 정확도를 가지는 DistillBERT를 구축하였으며, 이후 6.4배 작고 9.4배 빠르면서도 96.8%의 정확도를 가지는 화웨이의 TinyBERT 등의 연구로 이어졌다.

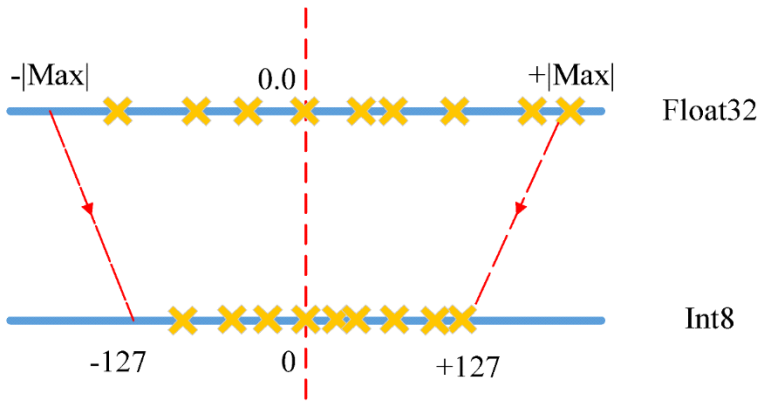
Pruning은 사전학습된 모델 혹은 미세 조정된 모델에서 불필요한 파라미터 가중치를 0으로 만들고 연산하지 않음으로써 용량을 줄이고, 속도를 빠르게 만드는 기법이다. 이 때 가장 중요한 것은 어느 가중치가 더 중요한가를 판가름하는 것이다. 여기에는 신경망 모델에서 보편적으로 가중치가 0에 근접할수록 그 영향력이 미미하다는 점을 이용해 0에 가까운 가중치를 찾아 0으로 만드는 Magnitude weight pruning 방법[Gordon et al. 2020]이나, 미세 조정 과정에서 0으로부터 멀어지는 가중치를 찾아 0으로 만드는 Movement weight pruning[Sanh et al. 2020] 등이 있다.

이러한 pruning을 통해 BERT 모델의 3~40%의 가중치는 정확도에 영향을 주지 않는다거나, 5%의 가중치 파라미터만을 가지고도 BERT 모델의 95%의 정확도를 유지할 수 있다는 사실이 알려졌다. Pruning 과정에서 모델의 경량화 뿐 아니라 파운데이션 모델이 가진 동작 원리에 대한 이해를 높여주기도 한다. 예를들어, Sajjad et al.의 연구에서는 12개의 레이어를 가진 BERT-base 모델이 각 다운스트림 작업에 따라 서로 다른 레이어의 출력 결과를 더 중요한 판단 근거로 삼으며, 따라서 다운스트림 작업에 따라 절반 이상의 레이어를 아예 사용하지 않더라도 1% 미만의 정확도 변화만이 발생할 수 있다는 사실을 밝혔다.

비록 트랜스포머 모델의 아키텍처는 모든 파운데이션 모델에 걸쳐서 거의 변함없이 사용되고 있으나, 트랜스포머 모델의 각 레이어의 크기 최적화 및 가중치 공유, 자기 지도학습 목표 변경 등의 방법을 통해 트랜스포머 모델의 효율성을 개선하는 연구가 이어지고 있다.

페이스북에서 연구한 RoBERTa는 BERT 모델과 구조적으로 동일하나, 약 160GB 크기의 학습 데이터 셋을 이용하여 훨씬 더 많은 데이터를 학습할 수 있도록 하였고, 데이터 전처리 및 배치 사이즈 조정 등의 방법만으로도 모든 벤치마크에서 BERT보다 확연히 더 뛰어난 성과를 거두었다. ALBERT의 경우 토큰의 factorized embedding 및 트랜스포머 모델 내부의 멀티 헤드 어텐션 및 feed-forward network 가중치를 모든 레이어에 걸쳐서 공유함으로써 파라미터의 수를 10분의 1에서 20분의 1까지 줄일 수 있었고, 다시 기존 BERT보다 더 큰 모델 사이즈를 갖도록 스케일 업을 함으로써 GLUE 벤치마크에서 동일 파라미터 수 기준 BERT보다 훨씬 뛰어난 성과를 보여주었다.

학습 혹은 모델의 수정을 통해 최적화를 수행하는 방법과는 별개로, GPU 등의 연산 속도를 빠르게 최적화하는 기법도 존재한다. 양자화는 이러한 연산의 효율을 향상시키는 방법 중 하나로, 모델의 가중치 파라미터 값을 부동소수점이 아닌 정수로 매핑하는 방법을 사용한다. 예를들면, 32비트 부동소수점으로 표현된 모델의 가중치 값들 중, 가장 큰 값을 정수 127에, 가장 작은 값을 -128에 매핑한 후 나머지 값들을 전부 가장 가까운 정수로 변환하여 모든 파라미터 값을 8bit 정수형으로 변환하는 것이다. 이 때 정수로 변환하는 과정에서 어느 정도의 오차가 발생할 수 있어 모델의 정확도가 하락하는 문제가 있으나, 실험을 통해 그 격차가 크지 않음을 보인 사례가 존재한다[Gholami et al. 2021].



[그림 13] 양자화의 개념도

(출처: Optimized Compression for Implementing Convolutional Neural Networks on FPGA, 2019)

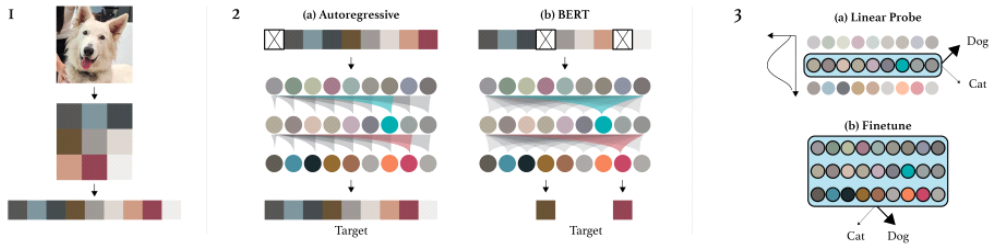
C 이미지 생성 및 분류

컴퓨터 비전 분야의 주요 과제는 자율 주행 자동차 및 로봇 운용에 필요한 시각 정보 인지, 의료 분야에서 희귀 질병 판별을 위한 이미지 인식 기술, 그리고 이미지 및 영상 생성 기술을 대표적으로 꼽을 수 있다. 그리고 이러한 작업들을 해결하기 위해서는 객체 인식, 의미론적 이미지 분할(semantic segmentation), 이미지 분류, 이미지 및 영상 생성 등의 작업 능력이 필요하다.

ImageNet과 같은 라벨링된 대규모 이미지 데이터 셋이 존재했기 때문에, 자기 지도학습 대신 이러한 대규모 데이터 셋을 통해 학습된 모델의 파라미터를 이용하는 방법이 줄곧 이용되어왔다. 그러나 자연어 분야에서 대규모 학습데이터를 자기 지도학습으로 학습하는 파운데이션 모델이 성공을 거두게 되면서, ImageNet 이상의 데이터를 학습할 수 있는 새로운 방법론이 주목을 받았다. 이미지 분야의 파운데이션 모델들은 특히 자연어와는 다른 2차원 픽셀의 집합인 이미지를 트랜스포머 모델에 적절하게 주입시키는 방법에 대해 다양한 시도를 하였다.

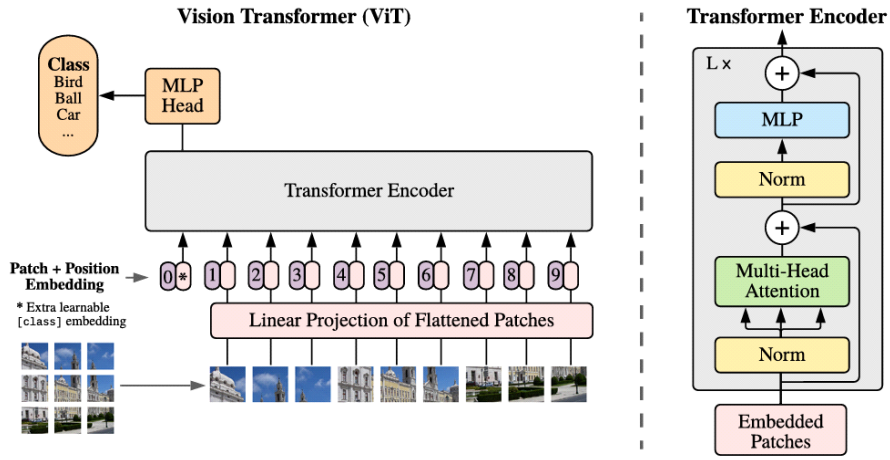
iGPT[Chen et al. 2020]는 OpenAI에서 개발한 GPT-2의 이미지 버전 모델로써, GPT-2 및 BERT의 모델 구조를 그대로 가져와 이미지 생성 및 분류 분야에 적용시킨 사례이다. 이 연구에서는 마치 자연어에서 단어를 토큰으로 삼았던 것처럼 이미지의 픽셀을 토큰으로 삼고, 이미지 픽셀을 래스터 스캔과 같이 왼쪽 위에서부터 오른쪽 아래까지의 순서가 있는 시퀀스 데이터로 해석하였다. ImageNet 데이터의 표준 크기는 224x224x3로, 트랜스포머 모델의 입력이 되기에는 지나치게 시퀀스의 길이가 길기 때문에 저해상도로 변환하는 작업을 먼저 수행하였다. 그리고 주어진 픽셀의 다음에 올 픽셀을 예측하는 방법으로 자기 지도 학습이 수행되었으며, 같은 방식으로 이미지를 생성할 수 있었다. 이미지 분류를 위해서는 미세 조정을 거쳤으며, iGPT 모델이 생성하는 feature map 뒤에 linear layer를 추가한 후 지도학습으로 미세 조정을 하는 BERT의 방식을 그대로 차용하였다.

ImageNet 데이터만으로 학습된 iGPT-L은 GPT-2와 같이 약 14억 개의 내부 파라미터를 가지고 있으며, 웹에서 수집된 레이블이 없는 데이터를 추가하여 더 많은 데이터로 학습시킨 iGPT-XL은 68억개의 파라미터를 가지게 되었다. iGPT는 이미지의 2차원적 특성을 전혀 고려하지 않고 생성한 모델이었음에도, 기존의 이미지 생성/분류 모델에 준하는 성과를 거둠으로써 자기 지도학습 방법이 데이터의 형태에 의존하지 않고 보편적으로 사용가능하다는 점을 보여주었다.



[그림 14] iGPT의 아키텍처 구성
(출처: Generative Pretraining from Pixels, 2020)

ViT(Vision Transformer) 모델은 그 논문 제목 - "An Image is Worth 16X16 Words: Transformers for Image Recognition at Scale" 에서 보여주듯, 이미지를 16x16 단위의 패치로 분리한 후, iGPT와 마찬가지로 왼쪽 위부터 오른쪽 아래까지의 순서를 갖는 시퀀스 데이터로 해석하여 트랜스포머 인코더 모델에 입력하는 방식을 사용하였다. 자기 지도학습을 사용하는 대신 이미지 분류 작업을 수행하도록 사전학습을 설계했으며, 대신 학습 데이터로 구글 검색을 통해 수집된 약 3억 장의 이미지로 구성된 JFT-300M 데이터 셋을 사용하였다.



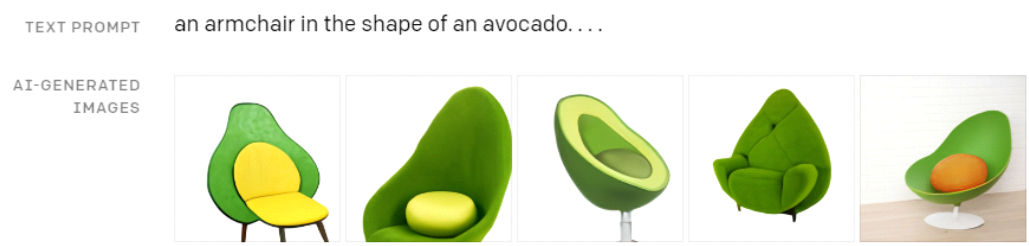
[그림 15] Vision Transformer의 구조
(출처: An image is worth 16X16 words: transformers for image recognition at scale, 2021)

ViT[Dosovitskiy et al. 2020]는 최대 6억 개의 파라미터를 가지고 있으며, ImageNet, CIFAR-10, CIFAR-100 등 많은 이미지 분류 작업에 대해서 state-of-the-art 수준의 성능을 보여주었다. 동시에 3억 개의 이미지 데이터를 가진 JFT-300M 대신 약 천만개 정도의 데이터 셋을 가지는 ImageNet만으로 학습시킬 경우 성능이 하락되는 모습을 보여준 점을 통해 이미지 파운데이션 모델 역시 더 많은 데이터가 필요하다는 점을 보여주었으며, 각 패치의 위치 정보를 입력할 때에 이미지의 2차원적인 특성을 반영한 경우와 하지 않은 경우가 별로 차이를 보이지 않는 특이한 결과를 보여주었다.

D 이미지-텍스트 멀티 모달 모델

DALL-E[Ramesh et al, 2021] 모델은 토크나이징 방법에서 차이를 보인다. 먼저 이미지 데이터를 픽셀이나 이미지 패치 대신 discrete VAE(dVAE) 모델을 사용하여 256x256 이미지를 32x32 크기에 8192개의 값을 가질 수 있는 토큰으로 변환하는 방법을 학습시킨다. VAE는 인코더-디코더 구조로, 인코더는 이미지 분포를 추정하고 디코더는 분포로부터 이미지를 생성하는 모델이다. dVAE는 DALL-E에서 새롭게 제안된 모델로써, Gumbel-softmax 기법을 사용하여 인코더가 생성하는 데이터의 분포(32x32 토큰)을 이산 값을 갖도록 만드는 모델이다. DALL-E에서 dVAE는 토크나이저로써 사용되므로 학습이 완료되면 디코더 모델은 사전학습에 사용하지 않는다. 그 후, 최대 256개의 토큰으로 구성된 텍스트와 32x32 이미지 토큰을 1차원으로 나열한 1024개의 이미지 토큰을 합쳐 1280개의 토큰으로 이루어진 시퀀스 데이터로 변환한 후, 트랜스포머 디코더 모델의 학습 데이터로써 활용한다. 이 때의 학습 목표는 GPT 모델과 같이 다음에 올 토큰을 예측하는 모델이 된다.

DALL-E는 120억 개의 파라미터를 지니고 있으며, 약 2억 5천만 개의 텍스트-이미지 데이터를 학습에 사용하였다. 학습이 완료된 모델에 256토큰의 텍스트 데이터를 입력하면 모델은 스스로 해당 텍스트 이후에 나올 이미지 토큰을 생성하고, 이 토큰을 다시 dVAE 디코더를 통해 복원시키면 새롭게 생성된 이미지가 만들어진다.

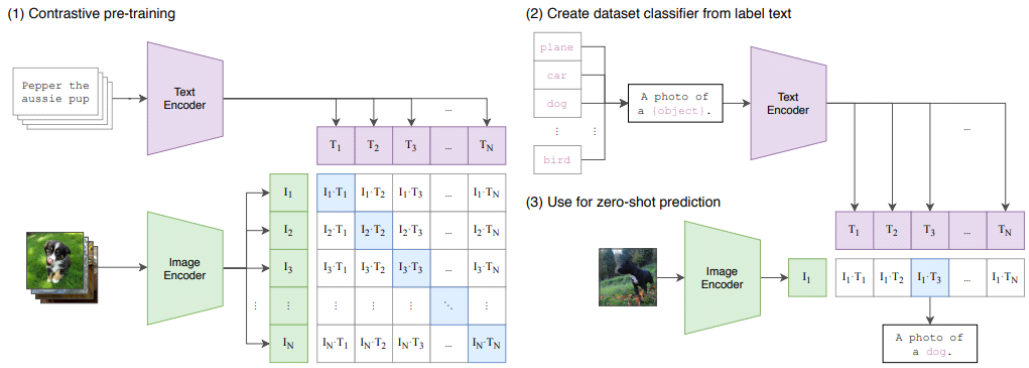


[그림 16] DALL-E의 텍스트 기반 이미지 생성 예시
(출처 : https://openai.com/blog/dall-e)

CLIP[Radford et al. 2021] 모델은 이미지 분류를 위한 파운데이션 모델로써 앞선 모델과는 사뭇 다른 접근 방식을 사용한다. CLIP은 기존 이미지 분류 모델들이 이미지를 특정 클래스 셋에 대해 분류하도록 학습된 경우 zero-shot 분류 성능이 떨어진다는 사실을 지적하면서, 클래스 명칭의 언어적 의미를 분류에 활용하는 기법을 제안하였다.

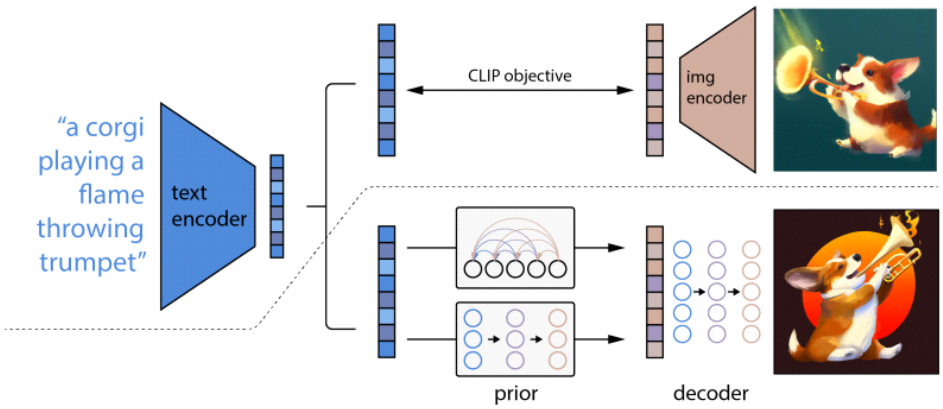
CLIP은 본질적으로 서로 다른 두 개의 인코더 모델이 학습되는 구조이다. 텍스트 인코더로는 트랜스포머 인코더를 사용하였으며, 이미지 인코더로는 모델 종류에 따라 ResNet-50 혹은 ViT가 사용되었다. N개의 (이미지-텍스트) 쌍이 입력되었을 때, 각각의 인코더가 각자의 데이터를 인코딩하여 벡터로

표현한 후, 이들의 코사인 유사도를 구하여 NxN 행렬로 만든다. 그리고 이 행렬에서 각 행의 가장 높은 값이 본래의 이미지-텍스트 쌍의 위치를 가리키도록 모델을 학습시킨다. 분류 작업이 수행될 때에는 이미지와 함께 분류에 사용될 클래스명이 텍스트로 입력되며, 가장 이미지와 코사인 유사도가 높은 클래스명이 최종 답변으로 선택된다.



[그림 17] CLIP의 구조
(출처: Learning Transferable Visual Models From Natural Language Supervision, 2021)

Diffusion 모델 기반의 멀티모달 모델 역시 주목을 받고 있다. unclip[Ramesh et al. 2022] 또는 DALL-E2 라 불리는 모델은 CLIP 모델과 diffusion 모델인 GLIDE의 아이디어를 결합한 모델로써, CLIP을 기반으로 이미지 임베딩을 생성한 후, AR 혹은 diffusion 모델 기반의 prior 모델을 이용하여 4가지 정보, 즉 텍스트, CLIP 텍스트 임베딩, timestep embedding, 노이즈가 추가된 CLIP 이미지 임베딩으로부터 노이즈가 제거된 CLIP 이미지 임베딩을 찾아내도록 학습한다. 그 후, GLIDE 모델과 비슷한 방법으로 decoder diffusion 모델을 이용해 이미지 임베딩 및 텍스트를 이용해 이미지를 복원하고, upsampling하여 해상도를 256x256 혹은 1024x1024까지 높인다.



[그림 18] unCLIP 모델의 구조. unCLIP은 먼저 CLIP과 같은 형태로 텍스트 인코더를 학습한 후(점선 위), diffusion model 혹은 autoregressive model과 같은 prior, decoder 모델과 같이 학습한다.
(출처: Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022)

E 학습 시스템

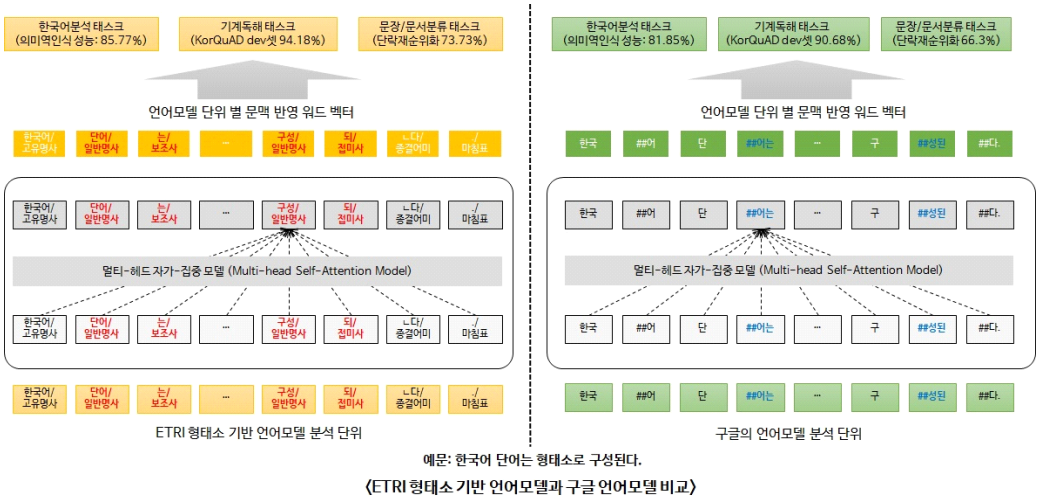
오늘날, 대규모 파운데이션 모델 훈련은 대체로 Megatron, DeepSpeed와 같은 시스템 및 PyTorch, TensorFlow와 같은 라이브러리를 이용해서 이루어진다. CRFM에 따르면, 이러한 소프트웨어 시스템은 규모에 맞게 모델을 효율적으로 학습하기 위해 장치를 바쁘게 유지하면서 장치간 통신을 제한하는 파이프라인 병렬화 [Huang et al. 2019], 메모리 사용량을 줄이기 위한 상태 샤딩 옵티마이저, 계산 그래프를 최적화하는 JIT(Just-In-time) 컴파일러, cuDNN 및 NCCL과 같은 최적화된 라이브러리와 같은 스택 전반에 걸친 수많은 혁신에 의존한다. 다만 Megatron과 DeepSpeed는 특정 수준의 규모에서 가장 효율적인데, 예를 들어 Megatron은 1조 개의 매개변수가 있는 모델에서 약 3000개의 GPU를 사용하여 최신 하드웨어의 이론적인 최대 처리량의 최대 52%까지 끌어낼 수 있지만, 더 많은 GPU가 있는 더 큰 모델로 확장하는 것은 여전히 어렵다. 이는 병렬화 전략이 더 많은 GPU 수가 주어졌을 경우 배치 크기, 모델의 레이어 수에 따른 파이프라인 병렬 처리, 단일 서버의 GPU 수에 따른 텐서 모델 병렬화 등에서 제한적이기 때문이다.

새로운 하드웨어들은 컴퓨팅 성능을 향상시키겠지만, 대형 모델의 리소스 요구 사항은 하드웨어의 향상보다 훨씬 빠르게 증가하고 있다[Brown et al. 2020]. 모델 용량의 다음 주요 도약을 촉진하고 모델 품질의 진보를 민주화하려면 훈련 알고리즘, 모델, 소프트웨어 및 하드웨어를 공동 설계(co-design)하는 것이 점점 더 중요해질 것인데, 왜냐하면 성능을 크게 향상시키는 많은 방법이 훈련 계산의 의미를 변경하기 때문이다. 예를 들어 낮은 정밀도(예: fp16)에서 작업을 실행하면 최신 하드웨어에서 처리량을 높이는 데 도움이 될 수 있지만 최적화 과정에서의 값에도 영향을 미친다. [Micikevicius et al. 2017]. 마찬가지로, 가중치 sparsity를 활용하면 모델의 0이 아닌 값에 대해서만 수학적 연산을 수행함으로써 훈련 및 추론 시간을 크게 향상시킬 수 있지만, 별도의 학습 알고리즘이 필요하다. 공동 설계(co-design)의 다른 예에는 하드웨어에 보다 효율적으로 매핑되는 모델 아키텍처, 효율적인 옵티마이저, 새로운 토큰화 대안, 특별히 설계된 하드웨어 교육 플랫폼, 완화된 가중치 업데이트 시맨틱을 사용한 분산 병렬화 전략 등이 있다.

F 한국어 언어 모델

KorBERT[6]는 한국어 처리에 특화된 BERT 모델로써, 과학기술정보통신부와 IITP의 혁신성장동력 프로젝트로 추진 중인 엑소브레인 사업을 통해 개발되었으며, 신문기사 및 백과사전을 포함한 23GB의 텍스트, 47억 개의 형태소를 기반으로 학습되었다. KorBERT는 한국어 처리 관련 벤치마크 태스크 5종(의미역 인식, 기계독해, 단락 순위화, 문장 유사도 추론, 문서 주제분류)에서 다국어 위키피디아 문서 데이터로 학습된 multilingual BERT 모델과 비교 평가한 결과, 평균 4.5%정도의 정확도 향상을 보였다.

6) https://aiopen.etri.re.kr/bertModel



[그림 19] KorBERT 모델(좌측)과 BERT(우측)의 비교
(출처 : https://aiopen.etri.re.kr/bertModel)

KorBERT의 가장 큰 특징으로는 한국어의 특성을 반영한 형태소분석 기반의 토크나이징 기법으로, 기존 BERT가 채용한 데이터 기반 BPE(Byte-Pair Encoding) 방식을 사용하지 않고, 한국어 문법에서의 형태소 단위로 토큰을 분리하였다. KorBERT는 해당 방식이 여러 태스크에서 어절 기반 언어모델보다 우수한 성능을 가짐을 보였다.

SK텔레콤에서 공개한 KoBERT[7]는 한국어 BERT 모델이다. 토크나이저로는 기존 BERT의 BPE와 거의 유사한 기법인 Sentencepiece(Kudo and Richardson, 2018)를 사용했으며, 한국어 위키 데이터 540만 단어를 사용하였다. KoBERT의 특징으로는 모델의 경량화를 위해 단어 사전의 어휘 수를 8,002개로 줄였는데, 이는 KorBERT의 30,349단어, HanBERT의 53,800단어에 비해 굉장히 적은 값이다. 이를 통해 BERT 모델의 파라미터 수를 110M에서 92M으로 축소시킬 수 있었다.

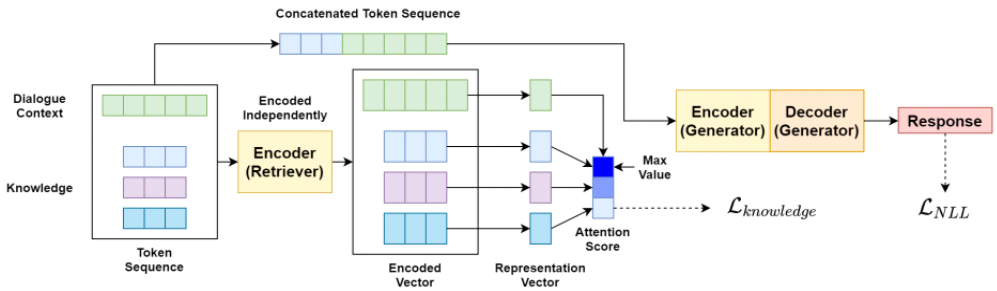
KcBERT[8]는 한국어 댓글 데이터를 이용해 한국어 구어체를 해석할 수 있도록 구축된 모델이다. KcBERT는 네이버 뉴스의 댓글 데이터를 수집하여 구축된 약 15GB 규모의 데이터셋으로 학습되었으며, BPE와 유사한 WordPiece 토크나이저를 사용하였다. KcBERT는 7종의 한국어 관련 테스트 벤치마크 중 네이버 영화 리뷰 분류(NSMC)에서 가장 좋은 정확도를 보여주었는데, 이를 통해 구어체 및 댓글 관련 데이터에서의 성능이 우수함을 짐작할 수 있다.

7) https://github.com/SKTBrain/KoBERT
8) https://github.com/Beomi/KcBERT

네이버의 하이퍼클로바(HyperCLOVA[9])는 GPT-3(1,750억 개)보다 조금 더 거대한 2,040억 개의 파라미터를 가지는 트랜스포머 디코더 기반 모델이다. 비록 GPT-3의 학습 데이터에도 한국어 데이터가 포함되어 있기는 하나, 한국어에 최적화되어있지 않아 한국어 해석 능력은 떨어진다. 반면 하이퍼클로바의 학습 데이터는 한국어 비중이 97%에 달하는 한국어에 특화된 GPT 계열 모델이다. 하이퍼클로바의 학습에는 약 5,600억 개의 토큰을 가지는 학습 데이터가 사용되었으며, 해당 모델을 학습시키기 위해 700 페타플롭(PF) 성능의 슈퍼컴퓨터를 도입하고, 대용량 데이터 처리를 위한 인프라를 갖췄다.

KoGPT[10]는 카카오브레인에서 개발한 GPT-3의 한국어 특화 버전 모델이다. 다만 그 규모는 기존 GPT-3보다 작은 60억 개의 파라미터로 이루어져 있으며, 학습 데이터는 한국어 데이터로 약 2,000억 개의 토큰으로 이루어져 있다. 또한 KoGPT는 64,512개의 어휘를 사용하며, 뉴스 기사 분류에 해당하는 YNAT 태스크에서 뛰어난 성능을 보여주었다. KoGPT 모델은 한국어 기능을 갖춘 GPT-3 모델 중에서도 오픈소스로 개방되어 있으며, 그 크기가 작아 일반적인 개인용 컴퓨터 사양에서도 실행이 가능한 것이 특징이다.

KE-T5[11]는 KETI에서 개발한 T5 모델을 한국어와 영어 두 종류의 데이터를 활용해 사전학습한 모델이다. KE-T5는 뉴스 및 웹에서 자체 수집한 데이터 셋 및 국립 국어원 모두의 말뭉치 데이터를 포함한 72GB 규모의 한국어 데이터와 약 21GB로 이루어진 RealNews 데이터 셋[12]을 사용해 학습되었다. 토크나이징에는 sentencepiece가 사용되었으며, 총 64,000개의 어휘를 가지고 있다.



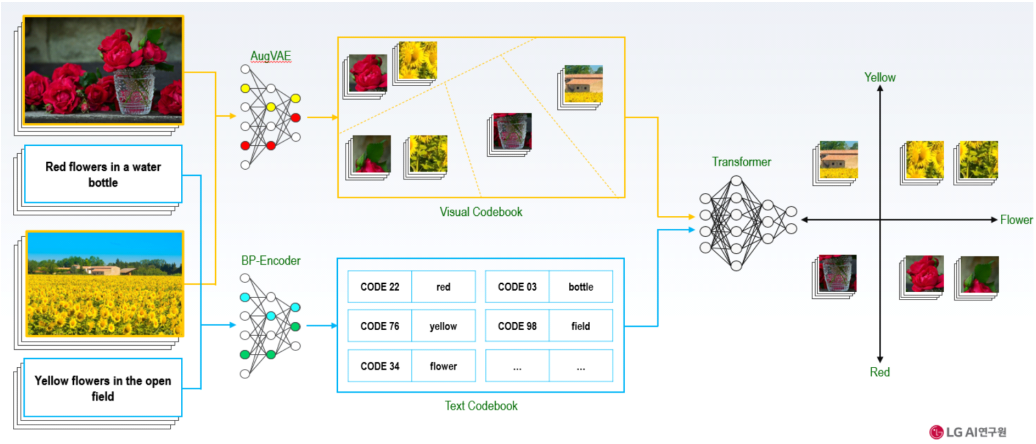
[그림 20] KE-T5의 구조

(출처: A Model of Cross-Lingual Knowledge-Grounded Response Generation for Open-Domain Dialogue Systems, 2021)

9) https://www.navercorp.com/promotion/pressReleasesView/30546
10) https://www.kakaocorp.com/page/detail/9600
11) https://github.com/AIRC-KETI/ke-t5
12) https://github.com/rowanz/grover/tree/master/realnews

ET5는 KorBERT와 같이 과학기술정보통신부와 IITP의 혁신성장동력 프로젝트로 추진 중인 엑소브레인 사업을 통해 개발되었으며, T5의 Masked language modeling과 GPT의 causal modeling 사전학습 기법을 모두 사용하여 학습시킨 것이 특징이다. T5계열 모델 중 한국어 처리 태스크 5종(기계독해, 요약, 단락 순위화, 형태소 분석, 문장유사도 추론)에서 실험한 결과, 가장 높은 성능을 보였다.

구분	[질의응답] 기계독해	[언어생성] 문서요약	[정보검색] 단락 순위화	[언어분석] 형태소분석	[문장의미 분석] 문장유사도추론
평가데이터 및 규격	KorQuAD v1.0 학습: 60,406건 평가: 5,773건 (dev셋)	AI Hub 요약 (뉴스-abstractive) 학습: 240,972건 평가: 30,121건(test셋)	학습: 45,521 질문 평가: 1,000 질문 (질문당 평균 8.7개 단락)	세종 말뭉치 학습: 135,238 문장 평가: 4,800 문장	학습: 41,465 문장쌍 평가: 3,181 문장쌍 (이진 분류체계: 유사, 무관)
평가 방법	Exact Match[1]/F1[2]	ROUGE-1/2/L[3]	Precision@Top1	F1	Accuracy
(KETI) KE-T5.ko base	91.35 / 85.40	49.15 / 21.97 / 46.78	81.8	N/A[4]	79.54
(Google) mT5 base	92.86 / 85.14	49.39 / 22.03 / 46.92	81.6	93.82	77.31
(엑소브레인) ET5 base	94.26 / 86.37	50.05 / 22.98 / 47.37	82.0	95.25	84.43



[그림 21] EXAONE Multi-modal Model
(출처 : https://www.lgresearch.ai/blog/view?seq=183)

EXAONE은 LG AI연구원에서 개발한 텍스트-이미지 멀티 모달 모델, L-Verse(Kim et al. 2022)를 이용하여 개발된 초거대 AI이다. EXAONE은 6,000억 토큰 규모의 말뭉치와 2억 5000만 장의 이미지-텍스트 쌍 데이터를 최대 3,000억 파라미터 규모로 학습했다. EXAONE에 사용된 L-Verse 모델은 DALL-E와 유사하게 VAE 방식으로 이미지를 토큰화한 후, 텍스트 토큰과 합쳐 트랜스포머 모델에 입력한다. 이 때, AugVAE라는 cross-level feature augmentation이 사용된 Hierarchical VQ-VAE를 사용하여 이미지를 토큰화하며, 트랜스포머 모델은 BiART(Bidirectional AutoRegressive Transformer)가 사용되었다. 저자들은 ImageNet을 대상으로 한 VAE 성능 평가에서 DALL-E, VQ-VAE와 같은 기존 VAE 모델보다 더 정확한 이미지 재현 성능을 보임을 실험으로 증명했다.

3 분야별 미래 연구 방향 및 파급 효과 분석

본 장에서는 파운데이션 모델이 사용될 수 있는 분야 중 연구 및 확산이 활발히 이루어지고 있는 언어와 컴퓨터 비전 분야에 대한 특징과 주요 다운스트림 작업, 연구 방향, 파운데이션 모델을 사용한 어플리케이션이 사회에 미치는 영향 등을 다룬다.

A 언어

언어는 대부분의 인간 의사 소통과 상호 작용의 기초가 된다. 그러나 언어는 단순히 인간이 공통 목표를 달성하기 위해 사용하는 수단은 아니다. 언어는 인간의 사고, 사회적, 정서적 관계가 어떻게 형성되는지, 우리가 사회적, 개인적으로 자신을 식별하는 방법, 그리고 인간이 지식을 기록하고 사회적 지능을 발전시킨다. 말하거나 기록된 언어는 모든 인간 사회에서 발생하며, 세계의 언어는 엄청나게 다양한 방식으로 그들이 전달하고자 하는 정보를 표현하고 구조화하며, 동시에 일관적이면서 풍요롭게 언어를 만들어낸다. 언어는 놀랍도록 복잡하지만 효율적인 시스템이며, 짧은 시간 안에 어린이들도 일관되게 습득할 수 있으며, 언어 사용자들의 변화하는 요구와 조건을 포함하고 발전시킨다. 인간 활동에서 언어의 중요성으로 인해 언어 이해와 생성은 인공 지능 연구의 중요한 요소이다. NLP(Natural Language Processing)는 언어와 관련된 인공 지능의 하위 필드이며, ASR(Automatic Speech Recognition) 및 텍스트 음성 변환과 같은 관련 분야와 함께 컴퓨터가 마치 사람처럼 인간의 언어를 이해하고 생성할 수 있도록 하는 목표를 가지고 있다.

2021년까지 NLP는 파운데이션 모델의 영향을 가장 많이받는 분야였다. 1세대의 파운데이션 모델은 인상적인 다양한 언어 능력뿐만 아니라 다양한 언어적 상황에 대한 놀라운 수준의 적응성을 선보였다. 2018년, 초기 파운데이션 모델 Elmo 및 BERT가 소개된 후, NLP 분야는 파운데이션 모델의 사용 및 이해를 중심으로 연구되었다. 그 이후, 파운데이션 모델을 기반으로 보다 일반화된 언어 학습으로 변화되었다.

● NLP에 대한 파운데이션 모델의 영향

파운데이션 모델은 NLP 분야에 큰 영향을 미쳤으며 현재 대부분의 NLP 시스템 및 연구의 중심이다. 기본적으로, 많은 파운데이션 모델은 숙련된 언어 생성기이다. 예를 들어 [Clark et al. 2021]은 비전문가의 경우 GPT-3과 인간이 쓴 짧은 형식의 영어 텍스트를 구별하는 데 어려움이 있음을 보였다. 그러나 NLP에 가장 영향을 미쳤던 파운데이션 모델의 특징은 기초적인 생성 능력이 아니라 놀라운 일반성과 적응성이다. 하나의 파운데이션 모델은 많은 언어적 작업을 해결하기 위해 다양한 방식으로 조정할 수 있다.

NLP 분야는 역사적으로 도전적인 언어 작업을 만들고 해결하는 데 중점을 두었으며, 이러한 작업을 해결하는 모델을 구축하면 다운스트림 작업을 위한 유능한 언어 시스템으로 이어질 것이다. NLP 작업에는 전체 문장 또는 문서에 대한 분류 작업(예 : 영화 검토가 긍정적인지 부정적인지 예측하는 감성 분류), 문장 또는 문서에서 각 단어 또는 문구를 분류하는 시퀀스 라벨링 작업(예 : 각 단어가 동사 또는 명사이거나 단어가 사람이나 조직을 언급하는 지 예측), 관계 분류(예 : 사람과 위치가 "현재 거주지"라는 관계로 연결되어 있는지와 같은 관계 추출 또는 동사와 명사가 "주어-동사"관계인지 판별하는 구문 분석) 및 주어진 입력에 의해 새로운 문장을 생성하는 작업(예 : 텍스트의 번역 또는 요약 생성, 연설을 인식 또는 제작하거나 대화에서 응답) 등이 있다. 과거에 NLP 작업은 각기 다른 모델 파이프 라인을 기반으로 하는 작업별 아키텍처를 개발하는 뚜렷하게 구분되는 연구 집단을 가졌으며, 각각 연구 집단은 토큰 세분화, 구문 분석 또는 상호참조해결(coreference resolution)과 같은 언어 하위 작업을 수행했었다.

이러한 작업을 수행하기 위한 현대적인 접근 방식은 주로 하나의 파운데이션 모델을 사용하여 각 작업(감성 분류, 명사 태깅, 번역, 요약)에 맞는 비교적 적은 양의 주석이 달린 데이터를 이용해 약간의 미세 조정을 시키는 것이다. 그리고 이는 매우 성공적인 접근 방식으로 입증되었다. 위에서 설명한 작업의 대부분에서 약간의 미세 조정을 거친 파운데이션 모델이 해당 작업을 해결하기 위해 직접적으로 구축된 이전 모델 또는 파이프 라인을 크게 능가한다. 예를 들어, 파운데이션 모델 이전인 2018년, 개방형 과학 관련 질문에 답변하기 위한 최상의 시스템은 NY Regents의 8 학년 과학 시험에서 73.1%의 점수를 얻을 수 있었다. 그러나 1년 후, 미세 조정된 파운데이션 모델은 91.6%를 기록했다 [Clark et al. 2019].

언어 생성을 위해 방대하게 훈련된 파운데이션 모델의 창발성은 NLP에서 언어 생성의 역할을 크게 변화시키는 데에 일조했다. 2018년까지, 보편적인 언어를 생성하는 문제는 다른 언어 관련 다운스트림 작업을 통하지 않고는 매우 어렵고 본질적으로 접근 할 수없는 것으로 간주되었다. 대신 NLP 연구는 대부분 언어적인 방법으로 텍스트를 분석하고 이해하는 데 중점을 두었다. 그러나 이제, "이 문장에서 다음으로 올 단어를 예측하는 것"과 같은 간단한 언어 생성 목표만으로도 높은 수준의 일관성을 지닌 파운데이션 모델을 훈련시킬 수 있다. 이러한 생성 모델은 이제 한때 생성의 전제 조건으로 간주되었던 분석 및 이해 작업을 포함하여 언어에 대한 기계 학습이 수행되는 기초적인 수단을 구성한다. 파운데이션 모델이 보여준 성공적인 문장 생성은 요약 및 대화 생성과 같은 언어 생성 작업 연구의 개화로 이어졌다. 파운데이션 모델 패러다임의 부상은 문서 뿐 아니라 구어에서도 비슷한 역할을 하기 시작했다. WAV2VEC 2.0과 같은 현대의 자동 음성 인식(ASR) 모델은 음성 오디오로 이루어진 대규모 데이터 세트에 대해 사전학습된 다음 ASR 작업을 위한 전사(음성->텍스트 변환)에 적용된다[Baevski et al. 2020]. 파운데이션 모델 패러다임으로의 변화로 인해 NLP의 연구 및 활용의 초점은 다양한 작업에 대한 맞춤형 아키텍처를 만드는 것에서 가장 적절한 파운데이션 모델을 탐색하는 것으로 바뀌었다. 미세 조정 방법에 대한 연구는 꽃이 피웠고, 파운데이션 모델의 놀라운 성공은 파운데이션 모델의 분석 및 이해에 대한 연구 관심의 변화를 일으켰다.

● 언어적 다양성 및 다국어

파운데이션 모델이 사전학습에서 얻은 언어 지식은 놀라울 정도로 다재다능하지만, 이 적용방법에는 한계가 있다. 현재 파운데이션이 모델이 언어의 다양성을 얼마나 성공적으로 처리하는지는 확실하지 않다. 언어들은 서로 대단히 다르다. 세계에는 수천 개의 서로 다른 언어가 있다는 사실 외에도, 한 언어 내에서도 혹은 한 화자의 언어조차도 다를 수 있다. 몇 가지 사례를 들자면, 비공식적인 대화는 문서화 된 언어와는 다르게 나타난다. 사람들이 친한 친구와 대화 할 때 사용하는 문법은 권위있는 사람과 대화 할 때 사용되는 것과 매우 다르다. 한 언어 내에서도 서로 다른 지역은 다른 방언을 사용하기도 한다. 사회적 및 정치적 요인은 언어의 다양성이 어떻게 보여지고 평가되는지, 그리고 NLP 연구에서 얼마나 많이 표현될 수 있는지를 내재하고 있다. 파운데이션 모델이 언어 정보를 학습하고 지식을 유연하게 적용하는 뛰어난 능력으로 인해 NLP 분야를 확장하여 더 언어적 다양성을 포괄할 것이다. 다만 파운데이션 모델이 주류와 비주류 언어의 다양성 속에서 강건하게, 그리고 동등하게 표현하면서도 동시에 날카롭게 그들을 분별하는 방법 혹은 분별할 수 있는가에 대한 연구가 남아 있다.

영어로 된 파운데이션 모델의 성공에 이어 다국어 파운데이션 모델은 그 성공을 영어 이외의 언어로 확장하기 위해 출시되었다. 세계 6,000 개가 넘는 언어 중 대부분은 대규모 파운데이션 모델을 훈련시키기에 사용 가능한 텍스트 데이터가 충분하지 않다. 예를 들어, 서 아프리카 언어 인 Fula어는 6천 5백만 명이 넘는 화자가 있지만 Fula어로 된 NLP에 사용할 수 있는 자원은 거의 없다. 다국어 파운데이션 모델은 동시에 여러 언어에 대한 공동 교육을 통해 해결하고자 한다. 현재까지의 다국어 파운데이션 모델(mBERT, mT5, XLM-R)은 각각 약 100 개 언어로 학습되었다. 공동 다국어 학습(Joint multilingual training)은 언어 간의 공유되는 구조와 패턴이 자원이 많은 언어로부터 자원이 적은 언어에게 공유 및 전이될 수 있다는 합리적인 가정에 의존하여 독립적으로 훈련시킬 수 없는 언어의 파운데이션 모델을 가능하게 한다. 다국어 파운데이션 모델을 사용하고 분석하는 실험에 따르면 다국어 파운데이션 모델에서 다양한 언어의 병렬 인코딩 사이에 놀라운 양의 지식 전이가 실제로 있음을 보여준다.

그러나 이러한 모델이 다국어를 강력하게 처리할 수 있는지는 여전히 열린 의문이다. 다국어 데이터에 대해 훈련된 모델이 영어와는 크게 다르면서 학습 데이터가 적은 언어를 잘 나타낼 수 있거나, 모델의 다국어 성능이 분명하게 (언어)동화에 의존하는지는 확실하지 않다. 다국어 모델은 학습 데이터가 가장 많은 언어와 유사한 언어에서 더 나은 성능을 보여 주며, 다국어 모델의 언어가 모델 매개변수를 놓고 경쟁하는 것으로 나타났기 때문에 단일 모델에 얼마나 많은 다양성이 적용될 수 있는지 명확하지 않다. 핵심 문제는 다국어 파운데이션 모델을 훈련시키는 데 사용하는 데이터에서 비롯된다. 많은 다국어 말뭉치에서 영어 데이터는 데이터가 적은 언어보다 훨씬 더 풍부할 뿐만 아니라 종종 더 명확하고, 넓고, 더 깊고 복잡한 예시들을 포함하고 있다. 그러나 그 문제에 대한 답은 단순히 균형

잡힌 말뭉치를 만드는 것으로 끝나지 않는다. 언어의 다양성은 축이 너무 많아서 모든면에서 균형 잡힌 말뭉치를 만드는 것이 불가능할 수 있다. 파운데이션 모델의 미래, 다양성 및 형평성은 모두 불균형 데이터에도 불구하고 언어 변화를 강력하게 처리하는 데 의존한다.

현재의 다국어 파운데이션 모델은 원시적이고 단순한 비지도학습을 통해 다국어를 학습시키는 방법으로서, 언어와 언어의 다양성의 미묘함을 아직 잘 모델링하지 못할 수 있다. 그럼에도 불구하고, 그들은 학습 데이터에 없는 희귀한 언어에 대해 다국어 모델을 적용하는 사례와 같이 일부 다국어 응용 프로그램에 여전히 유용하다. 연구 커뮤니티는 파운데이션 모델이 언어의 다양성을 다루는 방법을 비판적으로 조사하고, NLP에 형평성과 표현을 가져 오는 파운데이션 모델의 한계를 이해하고, 언어의 다양성을 배제한 채 학습 데이터의 대부분에서 주류 언어만을 다루는 파운데이션 모델을 홍보하지 않아야한다.

● 인간의 언어 습득에서 영감 얻기

파운데이션 모델은 인간처럼 동작하는 NLP 시스템을 만드는 원천으로 사용되어 큰 진전을 보였지만, 파운데이션 모델은 학습 과정 뿐 아니라 그렇게 획득한 언어 시스템 역시 인간과는 다른 방식을 가지고 있다. 기계와 인간이 언어를 학습하는 방식 사이의 이러한 격차가 가지는 의미를 이해하는 것은 파운데이션 모델의 언어적 한계와 가능성을 탐구하는 연구 분야를 개발하는 데에 필요한 요소일 것이다.

인간의 언어 습득은 대단히 효율적이다. GPT-3과 같은 파운데이션 모델은 대부분의 인간이 듣거나 읽는 것보다 약 3~4 배 더 많은 언어 데이터로 훈련되며, 이는 특히 언어 능력이 가장 발달하는 시기의 어린이들에 비하면 훨씬 많은 양이다. 파운데이션 모델과 인간의 언어 습득 방식의 한 가지 중요한 차이점은 인간의 언어는 현실 세계에 근거한다는 사실이다. 예를 들어, 아기와 보호자는 언어 발달 도중에는 실제 물체를 가리키며, 아기들은 그러한 보편적인 대상을 뜻하는 단어들의 기초적인 의미를 먼저 배우게 되고, 언어 시스템이 가진 다른 측면은 그 후에 배우게 된다. 반면에 NLP에서 사용되는 대부분의 파운데이션 모델은 원시적이고 현실에 근거하지 않는 텍스트의 분포 정보로부터 언어를 학습하며, [Zhang et al. 2021]은 RoBERTa 모델이 의미를 알아내기 전에 추상적인 구문의 특징을 표현한다는 것을 보였다. 물론 강력한 현실에 근거하지 않는 통계적인 학습도 실제로 아기들에게서 나타나기 때문에, 이것이 언어 획득에 중요한 요소임을 의심 할 여지가 없다. 그럼에도 불구하고, 파운데이션 모델을 위한 실제 현실 기반 언어 학습의 발전은 인간의 언어 획득 효율성에 근접하기 위한 중요한 방향으로 남아 있다.

또 다른 중요한 연구 방향은 파운데이션 모델의 귀납적 편향(inductive bias) 및 그것이 인간의 귀납적 편향과 어떻게 연관되어 있는지를 특정한 언어 학습 방법에 관한 것과 인간의 인지에 일반적인 것

모두의 관점에서 연구하는 것이다. 인간의 두뇌는 효율적인 언어 습득을 위해 보다 구조적으로 특화될 수 있지만, 파운데이션 모델은 백지의 학습자가 아니며, 이러한 언어의 귀납적 편향을 이해하고 조정하는 것은 파운데이션 모델 연구에 중요한 미래 연구 방향이다.

언어 습득 효율성의 중요한 요소는 인간이 체계적이고 일반화 가능한 언어 시스템을 습득한다는 사실이다. 인간 언어 시스템이 어떤 유형으로 이론적 추상화를 하는지에 대한 많은 이론이 있지만, 인간은 기존의 추상적인 언어 개념에 새로운 지식을 쉽게 끼워넣고 생산적으로 새로운 문법적인 문장을 만들 수 있는 형태로 언어를 배운다는 사실에는 대체로 동의한다. 예를 들어, 열 살짜리 어린이는 언어가 어떻게 작동하는지에 대한 많은 추상적 언어 개념을 습득했지만, 그들이 사용하는 단어와 문장 구성은 향후 10년 동안 크게 변할 것이다. 반면에 파운데이션 모델은 종종 우리가 인간에게 기대하는 체계적인 추상적 언어 개념을 얻지 못한다. 예를 들어, 파운데이션 모델이 한 차례 언어 구조를 정확하게 생성하는 데 성공했을지라도, 그 구조가 향후에 계속 일관적으로 사용되는지는 보장되지 않는다. 특히, 주제가 크게 변하는 경우에는 더욱 그러하다. NLP는 무겁고 경직된 언어 규칙에 지나치게 의존하는 등의 시스템 퇴보를 일으키지 않고도 파운데이션 모델이 언어 능력을 획득할 때 일종의 체계성을 개발해야 하는 과제에 직면해 있다.

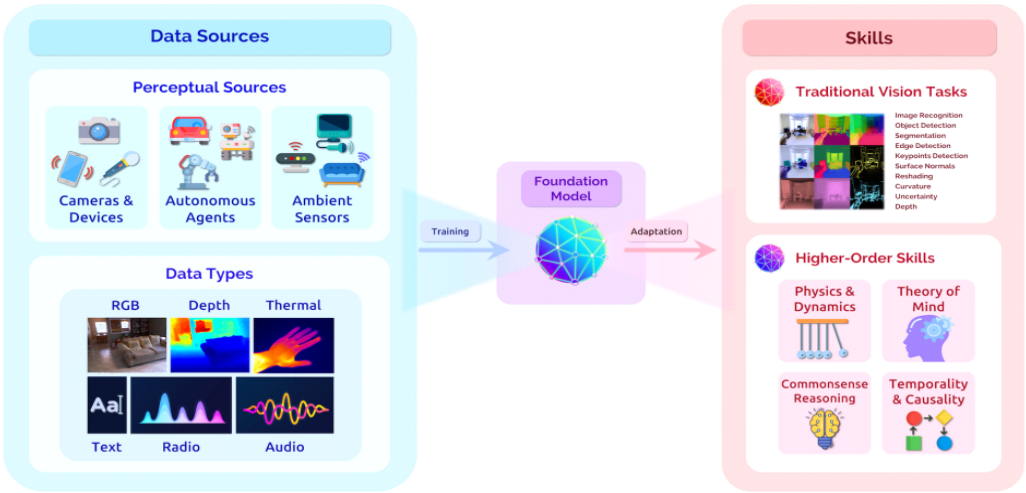
언어 학습은 화자의 일생동안 계속된다. 인간 언어의 문법은 진화하며 인간은 그러한 새로운 언어적 상황에 유연하게 적응한다. 예를 들어, 성인이 된 이후로도 새로운 용어와 개념이 발생하더라도 문법에 맞춰 비교적 쉽게 사용할 수 있으며, 인간은 종종 사회 집단이 달라지는 데에 맞춰 문법 패턴을 조정한다. 반면, 파운데이션 모델의 언어 시스템은 대부분 훈련 데이터에 의해 설정되며 비교적 정적이다. 미세 조정 방법은 다양한 작업에 대한 파운데이션 모델의 기초가 될 수 있지만, 방대한 훈련 없이도 파운데이션 모델의 근본적인 언어 기반을 변경하는 방법은 여전히 불분명하다. 인간처럼 언어적 편의와 언어의 진화를 자연스럽게 반영할 수 있는 미세 조정 가능한 모델을 만드는 것은 파운데이션 모델의 미래에 중요한 연구 영역이다.

B 비전

비전은 살아있는 유기체가 환경을 이해하는 주요 감각 중 하나이다. 보는 능력은 다양한 범위의 생명체에서 진화의 시간에 걸쳐 개발된 중요한 능력인 조밀한 신호를 거의 일정하게, 장거리로 수집할 수 있게 만들었다. 단순한 생명체도 쉽게 수행할 수 있는 기술조차도 동일한 능력을 기계로 이전하는 것은 매우 어려운 일임이 입증되었으며, 이는 1988년 컴퓨터 비전 및 로봇 공학 연구원인 Hans Moravec가 관찰한 역설로 이어졌다 - AI에게는 사람이 보기에 복잡한 문제가 쉬울 수도 있고, 반대로 쉬운 문제가 복잡할 수도 있다. 그리고 "가장 쉬운" 문제 중 하나는 몇 밀리초 만에 복잡한 장면을 지속적으로 해석하기 위해 인간이 매일 사용하는, 시력이다[Moravec 1988].

이 엄청난 도전의 다른 끝에는 컴퓨터 비전이 열쇠를 쥐고 있는 엄청난 양의 응용 분야가 있다. 예를 들면, 통근자를 교착 상태에서 자유롭게 할 수 있는 자율 주행 자동차, 희귀한 질환을 탐지하여 생명을 구하는 AI 도구, 멀티미디어 생성 및 편집을 위한 차세대 도구 등이 모두 비전의 응용 분야가 된다.

우리가 정의한 컴퓨터 비전의 하위 분야 및 목표는 인간의 인식 능력으로부터 여러 가지 방법으로 영감을 얻는다. 몇 가지 고전 이론는 인간이 (이미지의) 부분을 전체로 맥락화함으로써 실제 세계 장면을 인식 할 수 있다고 제안했으며, 컴퓨터 비전 기술의 추상화 수준을 높이면서 물리적 세계를 점진적으로 모델링하는 방법을 지적했다. [Gibson, 1979]은 인간의 시각이 본질적으로 구현되고 상호작용 가능한 생태 환경이 시각의 개발에 중요한 역할을 할 수도 있다고 제안했다. 이러한 아이디어는 컴퓨터 비전 시스템의 지속적인 발전에 동기를 부여하며, 지속적으로 세상을 맥락화, 인터랙티브, 구체화된 관점으로 바라보도록 했다.



[그림 22] 자기 지도학습을 대규모로 활용함으로써 비전을 위한 파운데이션 모델은 원시적인, 멀티 모달 감각 정보를 시각적 지식으로 추출할 수 있는 잠재력을 가지고 있다. 이는 전통적인 시각 작업을 효과적으로 지원하고 시계열 추론 및 상식 추론과 같은 도전적인 고차원 기술에 대한 새로운 진전을 가능하게 할 수 있다.
(출처 : On the opportunities and risks of foundation models, 2021)

컴퓨터 비전의 맥락에서, 파운데이션 모델은 다양한 소스 및 센서로부터 얻어진 원시적인 지각 정보를 시각적 지식으로 변환하여 수많은 다운스트림 설정에 적용할 수 있다. 대체로, 이러한 노력은 지난 10년 동안 이 분야에서 나온 주요 아이디어들의 자연스러운 진화였다. ImageNet의 등장[Deng et al. 2009]과 지도학습 방식의 사전학습의 출현으로 컴퓨터 비전의 딥 러닝 패러다임 전환이 이루어졌다. 이 전환은 우리가 초기의 고전적인 접근 방식에서 사용되던 작업별 feature engineering을 넘어서 많은 양의 데이터를 통해 한 번 훈련 된 후 다양한 작업, 예를들면, 이미지 인식, 객체 감지 및 이미지 세분화 등에 적용될 수 있도록 만들었다. 이 아이디어는 여전히 파운데이션 모델의 핵심에 남아 있다.

파운데이션 모델은 이전 패러다임의 한계에서 비롯된다. 전통적인 지도 학습 기법은 비용이 높고 신중하게 수집된 라벨 및 주석에 의존하기 때문에 견고성, 일반화 및 미세 조정 가능성을 제한한다. 대조적으로, 최근의 자기 지도학습의 발전은 대량의 원시 데이터를 활용하여 시각 세계에서 상황에 맞는 이해를 습득할 수 있는 파운데이션 모델 개발을 위한 대안을 제시한다. 컴퓨터 비전 분야의 더 넓은 목표와 비교했을 때, 현재 비전 파운데이션 모델의 기능은 초기 단계에 불과하다. 우리는 전통적인 컴퓨터 비전 작업(특히 일반화 능력과 관련하여)의 개선을 관찰하였다. 그리고 단기적인 진전을 통해 이러한 추세가 계속될 것으로 예상된다. 그러나 장기적으로, 파운데이션 모델이 가진 명시적인 주석 데이터에 대한 의존성을 줄일 수 있다는 잠재력은 완전한 지도 학습 패러다임에서는 어려운 것으로 입증되었던 필수적인 시각 인지 기술(예 : 상식 추론)에 대한 진전으로 이어질 수 있다. 결과적으로, 우리는 다운스트림 애플리케이션에 대한 파운데이션 모델의 잠재적 영향과 앞으로 나아가야하는 중심 도전과 한계에 대해 논의해야 한다.

● 주요 기능 및 접근 방식

높은 수준에서, 컴퓨터 비전은 인공 지능의 핵심적인 하위분야로 시각 세계를 해석하고 이해할 수 있는 능력을 기계에게 부여하는 방법을 탐색한다. 여기에는 지난 수십 년 동안 지속적인 진전을 이루어 왔던 수많은 태스크, 하위 주제 및 다운스트림 응용 프로그램이 포함되어 있다[Zamir et al. 2018]. 몇몇 태스크의 예를 들자면, (1) 이미지 내에서 이미지의 속성이나 물체 간의 관계를 발견하는 것을 목표로하는 의미론적 이해 작업은 이미지 분류, 객체 인지, 의미론적 분할(semantic segmentation), 동작 인식 및 이미지 그래프 생성 등을 포함한다. (2) 기하학, 모션 및 3D 작업은 정적인, 혹은 동적인 물체의 기하학적 구조, 자세 및 이미지의 구조를 나타내려고 한다. 여기에는 깊이 추정, 움직임으로부터 구조 추정, 표면 법선 탐지, 곡률 및 키 포인트 추정의 작업 등이 있다. (3) 의미론적, 기하학적 이해를 바탕으로 언어와 같은 다른 양식과 결합한 멀티 모달 통합 작업이 있는데, 여기에는 시각적 질의-응답, 이미지 캡션 지시 이행(instruction following) 등이 포함된다.

ImageNet[Deng et al. 2009]의 출현에 의해 시작된 이러한 문제를 해결하기 위한 패러다임은 2010년대 초, 친숙한 코어 아이디어를 중심으로 하는 경향이 있습니다. 첫째, 주석이 달린 데이터의 대규모 모음에 대해 이미지 분류와 같은 지도학습 방법으로 모델을 사전학습시킨다. 그런 다음, 각 작업에 대한 데이터 세트 및 분야에 모델을 다운스트림으로 미세 조정하여 state-of-the-art 성능에 도달한다. 이 사전학습과 미세조정으로 이어지는 개념은 우리가 파운데이션 모델에 대해 지금 고려하는 정의로 이어지고 있다. 이러한 완전한 지도학습 기반의 패러다임이 가진 한계는 파운데이션 모델로의 전환에 동기를 부여했다. 외부의 지도학습용 주석에 의존하는 것은 이전 접근법에게 확장 가능한, 강력하고 일반화 가능한 방식으로 다양한 시각적 입력을 포착하려는 시도에 한계를 갖게 만든다.

한편, 이미지 합성 및 비지도학습 영역의 최근 발전은 강력한 대안을 제공한다. 예를 들어, GAN들은 서로를 지도 학습시킬 수 있는 Generator와 Discriminator라는 두 개의 적대적 네트워크와 이미지 수집만으로 높은 fidelity를 가진 현실적이고 다양한 이미지를 생성하는 법을 학습할 수 있다. 다른 신경 모델은 VAE(Variational Auto-Encoder), 대조 학습(constrastive learning) 또는 기타 자기 지도학습 기법을 사용하여 명시적으로 주석이 달린 데이터 없이 물체와 장면의 시각적 특성을 유추한다. 예를 들어, [He et al., 2021]은 마스킹된 이미지 인코딩을 사용하여 표현 학습에 대한 이전 작업을 기반으로 하면서, 부분적으로는 최근에 발전한 유연한 아키텍처(예: ViT[Dosovitskiy et al. 2021])를 확장하여 결합하였다.

파운데이션 모델은 이러한 자기 지도학습 기법의 개발을 통해 범위와 잠재적 다양성 측면에서 더 큰 규모의 시각적 데이터를 이용한 훈련을 가능하게 했다. 따라서, 우리는 표준적인 정확도 지표와 Few-shot 일반화 측면에서 전통적인 비전 작업에 비해 향상되고 있다는 초기 지표를 보았다. 이미지 분류 및 객체 감지의 경우, 자기 지도학습 기법은 학습 데이터 중에서 명시적인 주석이 없이도 이전의 완전한 지도학습 기반 접근법에 대하여 경쟁력있는 성능을 보고했고, 미세 조정 과정에서는 더 큰 샘플 효율성을 보였다. 이미지 합성의 경우, 주목할만한 예는 Dall-E[Ramesh et al. 2021] 및 CLIP 기반 이미지 생성[Radford et al. 2021]으로, 연구원들은 멀티 모달 언어와 비전 입력을 활용하여 매력적인 장면을 시각적으로 렌더링한다.

특히, 컴퓨터 비전에 대한 현재의 파운데이션 모델은 NLP에 비해 초기 단계이다. 유망한 시도는 여전히 RGB 이미지 입력과 핵심적인 기존 비전 작업들의 부분집합 안에 주로 집중되어 있다. 그러나 이 분야는 구체화되고 상호 인식에 중점을 둔 광범위한 도전에 대해 계속 발전하고 있다. 물리적인 장면 이해, 시각적 상식 및 시계열 사건에 대한 추론, 사회적 어포던스(social affordance: 주변 환경에서 어떻게 행동하고 상호작용할지를 결정하는 신호나 기회를 의미함)에 대한 인식 등을 포함한다. 이것들 각각은 완전한 지도 학습 기반 시스템의 목표였지만, 이러한 작업에 필요한 데이터의 양만큼 주석을 달기가 어렵다는 점으로 인해 부분적으로는 달성하기 힘든 목표임이 입증되었다. 예를 들어, 표준적인 이미지 기반 질의-응답 시스템은 단순히 이미지의 픽셀 자체가 보여주는 것 뿐 아니라 외부 지식이

요구되는, 상식 이해가 필요한 대답을 하기 위해 노력해왔다. 대화형 에이전트에 사용되는 비전 시스템에서 강건한 방식으로 인간의 시선과 사회적 어포던스를 인식하는 것은 지속적인 도전 과제로 남아있다.

● 중심 연구 과제

파운데이션 모델이 비전 모델을 통합하거나, 영향을 줄 수 있는 다운스트림 애플리케이션 영역은 다음과 같다.

(1) 의료 및 가정 환경을 위한 생활환경지능(ambient intelligence) : 생활환경지능을 위한 기존 접근법 위에서 파운데이션 모델은 임상의사, 환자 및 일상 소비자를 보조하는 상호 작용을 개선할 뿐만 아니라 미세한 인간의 활동 및 의학적 사건을 더 잘 탐지할 수 있는 잠재력을 제공할 수 있다.

(2) 모바일 및 소비자 애플리케이션 : 더 강한 멀티 모달 기반의 파운데이션 모델은 모바일 환경에서 서비스의 상호작용을 보다 능숙하게 만들 수 있으며 시각적 및 언어적 입력으로부터 생성하는 기능의 중요한 개선이 컴퓨터 사진 및 컨텐츠 편집 애플리케이션에 도움이 될 수 있다.

(3) 구체화된 대화형 에이전트 : 인식 모델은 이미 로보틱스 분야에서 입력과 보상 기능 두 가지 면에서 효과적인 것으로 입증되었다. 파운데이션 모델은 대규모로 수집된 1인칭 시점의(현실기반/시뮬레이션기반, 인간의/로봇의) 비전 데이터에 대하여 훈련을 받음으로써 시각적 장면, 객체 및 동작의 더 넓은 분포를 포착하여 더욱 진보할 수 있다.

4 고려 사항

본 장은 파운데이션 모델로 인해 발생하게 될 새로운 사회적 영향 및 위험 요소들을 법적 요소, 환경적 요소, 사회적 요소 등의 관점으로 인지하고 그와 관련된 이해를 위한 것이다.

A 법적 요소

해외 논문을 인용한 설명은 CRFM의 연구 내용을 토대로 기술되었으며, 특히, 법의 내용 등이 미국 법과 관련되어 있다. 국내 사례 혹은 국내 논문을 인용했거나, 명시적으로 국내 사례라고 언급한 경우에만 한국 법의 내용을 다룬다.

● 모델 학습 관점에서

파운데이션 모델의 학습을 위해서는 방대한 양의 멀티 모달 데이터를 축적해야 하므로 데이터 수집 및 데이터 사용에 대한 의문이 제기된다.

첫째, 모델 개발자가 웹 스크래핑을 통해 데이터 세트를 확보할 수 있는 능력은 법원이 법 조항을 해석하는 방식, 특히 “허가 없이” 서버에 액세스하는 것을 범죄화하는 법률 및 해석에 좌우될 것이다[13]. 미국의 경우, 컴퓨터 사기 및 남용법(CFAA: Computer Fraud and Abuse Act)에 해당된다. 법원은 이러한 질문에 대해 갈등을 빚고 있으며, 최근 사례들은 웹 스크래핑이 금지될 수 있는 상황을 명확히 하려고 노력하고 있다. 데이터 액세스의 제한성은 근본적으로 파운데이션 모델을 훈련하기 위해 실무자가 사용할 수 있는 데이터의 다양성에 영향을 미칠 것이다.

둘째, 학습 데이터 셋에 포함된 많은 데이터는 저작권이 있고 잠재적으로 지적 재산권법에 의해 보호될 것이다. 그러나 저작권법은 개인이 저작권이 있는 자료를 사용할 수 있도록 허용하는 경우, 예외를 인정하고 있다. 일부 학자들은 학습 데이터 셋의 법적 허용은 법원이 모델 학습 과정을 공정한 사용 원칙에 따라 “변혁적(transformative)”으로 해석하는지 여부에 크게 의존할 것이라고 믿는다[Lemley and Case 2020]. 무엇이 “변혁적”인가 하는 질문은 맥락적 해석에 대한 의존성이 매우 높지만, 일반적인 원칙은 변혁적 용도를 “새로운 것을 추가하면서, 더 큰 목적이나 다른 특성을 가지고 있고, 저작물의 원래 용도를 대체하지 않는 것”이라고 정의한다. 이미 최근에 출시된 Github Copilot 도구는 이러한 주장을 전면에 내세우고 있다.

13) https://www.rcfp.org/scraping-not-violation-cfaa/

국내에서도 웹 스크래핑과 관련된 소송이 빈번히 일어나고 있으며, 특히 웹 상에 공개된 데이터를 획득하는 행위 자체가 합법적인가에 관한 정보통신망법 적용과, 해당 데이터를 활용하려 할 때, 데이터에 대한 저작권법 등이 주요 쟁점이 되고 있다.[14)15)] 아직 이러한 사례에 있어서 명확한 기준은 존재하지 않으나, 미국과 유사하게 웹 스크래핑의 목적이 파운데이션 모델의 학습에 사용하기 위한 데이터 수집일 경우 원 저작물의 시장가치에 영향을 주지 않는 한 저작권법의 공정 이용 조항에 해당할 수 있다는 의견[김병필, 2022] 등 여러 쟁점에 대한 논의가 이루어지고 있다.

마지막으로, 일부 학습 데이터 셋은 개인 정보 보호법에 위배될 수 있다. 예를 들어, 일리노이 주는 개인이 자신의 생체 인식 데이터(예: 망막 또는 홍채 스캔 데이터, 지문, 음성 지문 또는 손 또는 얼굴 형상 스캔 데이터)가 부적절하게 수집 또는 사용되었을 경우 소송을 제기할 수 있도록 한다. 데이터 셋에 EU 시민의 정보가 포함되어 있는 경우 미국의 모델 제작자에게 적용되는 EU의 일반 데이터 보호 규정(GDPR: General Data Protection Regulation)과 같은 외국인 정보 보호법은 데이터 주체에게 데이터 수집 목적에 대한 정보를 제공하도록 요구할 것이다. 또한 개인에게 "잊혀질 권리"를 제공하는 캘리포니아 소비자 보호 개인 정보 보호법(CCPA: California Consumer Protection Privacy Act)과 같은 법에서 추가적인 문제가 발생할 수 있으며, 모델 제작자가 모델에서 학습 데이터를 "제거"해야 하는지에 대한 의문이 제기되고 있다.

Carlini et al. [2021]의 실험에 따르면 GPT-2 모델이 생성한 60만 개의 문장 중 약 0.1%에서 학습데이터를 암기하여 출력한 사례가 있었으며, 인명, 주소, 이메일, 전화번호, 트위터 아이디 등이 유출되는 사례도 존재했다. 또한 국내에서도 챗봇이 특정 입력에 대해 개인 주소나 인명 등을 출력하는 사례가 보고되기도 했다.[16)] 이러한 암기의 위험성은 파운데이션 모델의 용량과 성능이 향상됨에 따라 증가할 가능성이 있다.

● 모델 예측에 대한 책임 관점에서

파운데이션 모델 자체는 작업에 구애받지 않지만 미세 조정된 모델 또는 파운데이션 모델 자체가 학습한 표현(representation) 값은 전통적인 예측 작업에 사용될 수 있다. 이러한 작업이 더 큰 의사결정 시스템의 구성요소를 형성하는 경우, 파운데이션 모델은 행동, 의사결정 또는 정책에 영향을 미칠 것이다. 이로 인해 피해가 발생할 경우 모델 제작자와 모델 제작사를 운영하는 개인은 법적 책임을 질 수도 있다.

물리 시스템(예: 자율 주행 자동차, 전기 그리드 관리, 의료 진단 등)에 파운데이션 모델을 내장하면 개인에게 물리적 피해를 줄 수 있다. 여기서 법원은 불법 행위 원칙에 따라 책임 문제를 다룰 것으로

14) https://www.ajunews.com/view/20220515155129332
15) https://www.joongang.co.kr/article/25064802#home
16) https://www.donga.com/news/It/article/all/20210113/104899076/1

보인다. 잘 알려진 주요 질문으로는 사용자, 파운데이션 모델 제공자 및 애플리케이션 개발자의 책임 간의 상호 작용과 파운데이션 모델의 위험 프로파일을 평가하기 위해 법원이 사용할 표준 등이 포함된다. 특히 민감한 영역(예: 의약품)에서의 모델 배포를 위해서는 규제의 승인과 안전성 평가를 위한 표준화된 프로세스의 개발이 필요하다[Wu et al. 2021g].

보호되고 있는 개인정보(예: 인종, 성별)와 상관관계가 있는 방식으로 개인을 분류하는 미세 조정된 파운데이션 모델은 민권법에 따른 문제에 직면할 수 있다. 학자들은 고용, 주택 또는 신용 대출의 맥락에서 파운데이션 모델에서 비롯되는 차별 대우에 대한 주장이 제기될 수 있다는 점에 주목했다[Gillis and Spyess 2019]. 법원이 이 문제들을 정확히 어떻게 판결할 것인지는 명확하지 않다. 예를 들어, 학자들은 "차별"에 대한 법원의 전통적인 견해가 실질적으로 기계 학습 실무자들이 알고리즘적인 공정성 기술을 구현하는 것을 방해할 것이라고 언급했다[Xiang 2021].

미국 법은 정부 기관에 대한 특별한 특권과 제한을 인정한다. 따라서 지방, 주 또는 연방 수준에서 정부 기관에 의한 파운데이션 모델의 사용은 동등한 보호 요구 외에도 특수한 고려사항을 수반할 것이다. 위험 평가 또는 생명, 자유 또는 재산의 박탈을 초래하는 다른 환경에서 모델을 사용하는 것은 절차상 정당한 절차를 거치도록 청구될 것이다. 예를 들어, 행정기관(예: 환경보호청)에서 모델을 사용하는 경우, 청구자들은 그러한 사용이 적법한 절차의 기본 표준, 합리성/비양심성 및 투명성을 위반한다고 주장할 수 있다.

● 모델 출력물에 대한 법적 보호 관점에서

모델 출력물, 나아가 모델을 담당하는 모델 제작자도 특정한 법적 보호를 받을 수 있다. 첫째, 생성 모델에 의해 생성된 콘텐츠는 언론의 자유 문제를 내포할 수 있다. 법원이 기계로 생성된 콘텐츠에 대한 수정헌법 제1조 보호를 어디까지 찾을지는 불분명하다. 학자들은 챗봇의 발화가 "AI의 발화"로써 보호되는지 또는 인간 프로그래머의 발화로 여겨지는지를 포함한 여러 가지 공개 질문에 대해 논의했다[Kajbaf 2019]. 다른 이들은 공개 요건(의약품 또는 기타 물질에 대한 안전성 공개와 유사한)의 가능성에 주목했으며, 또한 모델이 생성한 콘텐츠가 기계적으로 생성되었다는 것을 명시하도록 강제하는 발화 원칙을 함의한다[Lamo and Calo 2019]. 이러한 문제는 개인이 파운데이션 모델을 사용하여 문장을 대량 생산할 수 있는지 또는 모델 제작자가 파운데이션 모델에서 생성한 콘텐츠에 대해 책임을 질 수 있는지 여부 등에 영향을 미치는 광범위한 결과를 초래할 수 있다.

둘째, 모델 산출물에 대한 소유권을 주장할 수 있는 사람에 대한 불확실성이 있다. 기존 저작권법은 컴퓨터 프로그램을 저작자로 인정하지 않기 때문에 컴퓨터 프로그램에 의해 생성된 "작업물"은 저작권의 보호를 받지 못한다. 결과적으로, 학자들은 다양한 접근법을 주장해왔다. 어떤 사람들은 상황에 따라, 프로그램을 만든 인간과 그 인간 사용자 모두 프로그램 출력의 "작성자"라는 유효한 주장을 가지고 있을 수 있다고 주장한다[Ginsburg and Budiardjo 2019].

모델이 예술적 시도부터 뉴스 자료와 같은 일상적인 환경에 이르기까지 “창조”의 과정에서 점점 더 많이 사용됨에 따라 기계로 생성된 콘텐츠의 소유권에 대한 분쟁이 더 흔해질 것이다.

위의 분석은 파운데이션 모델에 관련된 법적 문제의 표면적인 부분만을 다루었지만, 이러한 질문들의 해결은 파운데이션 모델의 구축, 사용 및 배포에 매우 중요하게 될 것이다.

B 환경적 요소

파운데이션 모델은 법률, 의료 또는 기후 변화를 다루는 등 많은 사회적 및 환경적 이점을 가져올 수 있다[Rolnick et al. 2019]. 그러나 그 규모로 인해 모델 제작자가 주의를 기울이지 않으면 탄소 배출량 증가를 통해 환경에 부정적인 영향을 줄 수 있다. 이러한 배출량을 다루는 것은 필수적이다. 현 예측에 따르면 기후 변화는 과거에 생각했던 것보다 더 빠르게 발생하고 있다.

파운데이션 모델에서 이러한 탄소 배출이 어디에서 발생할 수 있는지 이해하기 위해, 모델의 수명주기를 고려해야 한다. 첫째, 이들은 최대 몇 달 동안 방대한 양의 데이터를 학습하며, 종종 수백에서 수천 개의 GPU에 걸쳐 분산 훈련된다. 그 후, 그들은 새로운 분야에 미세 조정하거나 작은 모델로 증류(distillation) 될 수 있다. 이 모든 것은 학습 체제의 일부로 간주 될 수 있으며, 순수하게 연구에만 사용되는 모델은 이 단계까지 오지 않을 수도 있다. 모델이 미세 조정 및 증류된 후에는 제품으로써 배포 될 수 있다. 이 때부터 새로운 모델이 훈련되어 사이클이 반복 될 때까지, 많은 추론 작업이 모델을 통해 실행된다.

이 각각의 단계들은 대량의 에너지를 사용할 가능성이 있으며 탄소 배출에 기여할 수 있다. 파운데이션 모델은 초기 훈련 단계에서 대규모의 일회성 에너지 비용과 탄소 배출을 생성 할 수 있다. 예를 들어, 어떤 조건에서 하나의 BERT-base 모델을 훈련시킬 때의 탄소 배출량은 10년정도 자란 40그루의 나무가 자라야지만 상쇄될 수 있다. 그리고 대규모로 배포될 경우, 수백만 건의 요청에 대해 서비스를 제공하기 위해 상당한 에너지가 필요할 수 있다. 이 때 재생 불가능한 자원이 사용되는 경우 이는 대규모 탄소 배출로 변환된다.

따라서 훈련 및 배포 양 쪽 모두, 파운데이션 모델에 대한 설계를 결정하는 것의 환경 영향은 상당할 수 있다. 모델이 가진 계층 수를 줄이는 것과 같은 겉보기에는 미미한 결정조차도 큰 규모에서는 상당한 환경 비용 절감으로 이어질 수 있다. 예를 들어, [Henderson et al. 2020]의 계산에 따르면, 아주 조금 더 에너지 효율적인 상업적 번역 서비스를 배포하는 것은 큰 규모에서 에너지 그리드에 따라 하루에 78kg CO2eq에서 12,768kg CO2eq의 탄소 배출량을 절약 할 수 있다. 이것은 10년정도 자란 1~211 그루의 나무 혹은 1년 동안 1,416~232,289 제곱미터의 숲이 제거하는 탄소와 거의 동일하다. 따라서 파운데이션 모델의 설계, 배포 및 배포 후의 모니터링은 이러한 위험을 적절하게 반영해야 한다.

물론 주어진 모델에 의해 방출된 에너지 또는 탄소의 양을 계산할 때에는 불확실성이 있을 수 있고, 다른 탄소 배출원들의 배출량이 현재 파운데이션 모델에 의해 생성된 것보다 훨씬 클 수 있다. 그러나 파운데이션 모델이 계속 규모가 확대되고 인기가 높아지면 탄소 배출에 크게 기여하게 될 수 있다. 우리의 목표는 파운데이션 모델 개발자 및 대규모 배포자를 위한 프레임워크를 제공하여 불필요한

탄소 배출을 완화하고 이러한 모델의 순 사회적 영향을 긍정적으로 유지할 수있는 방법을 고려하는 것이다.

(1) 탄소 영향력은 많은 경우에 완화 될 수 있고 완화되어야 한다. 이는 탄소 집약도가 낮은 지역에서 모델을 학습하거나 보다 효율적인 모델 및 하드웨어를 사용하여 달성할 수 있다.
(2) 완화를 위한 모든 메커니즘이 다 소진되고 완화가 불가능한 경우, 더 큰 파운데이션 모델을 더 작고 효율적인 모델로 배포 해야하는지 여부를 결정하기 위해 사회적 비용과 이점을 평가해야한다. 대규모 파운데이션 모델의 학습에 들어간 선불 비용이 모델의 수명 주기에 걸쳐 상쇄될 수 있음을 이해해야 한다.
(3) 부정적인 영향을 완화하기위한 노력뿐만 아니라 에너지, 컴퓨팅 및 탄소 비용은 정책 결정 및 연구에 정보를 제공하기 위해 명확하게 보고되어야 한다.

● 탄소 영향력은 많은 경우에 완화될 수 있고 완화되어야 한다.

파운데이션 모델의 학습이 만드는 탄소 영향력은 추론을 위해 배포할 때의 영향과 다르다. 모델 학습에는 응답 속도 관련 요구 사항이 없으므로 클라우드 환경에서 상대적으로 쉽게 학습 작업을 에너지 그리드 사이에서 이동할 수 있다. 모든 에너지 그리드에는 사용된 에너지의 kWh당 배출되는 탄소의 양인 탄소 집약도(carbon intensity)가 있다. 예를 들어, 퀘벡은 수력 전기에 의존하기 때문에 탄소 집약도가 매우 낮고, 에스토니아의 에너지 그리드는 셰일 오일에 의존하기 때문에 탄소 집약도가 매우 높다(다만 빠르게 변화하고 있음). 최근 연구에 따르면 오염이 심한 발전소의 상위 5%가 모든 전기 기반 탄소 배출량의 73%를 차지한다[Grant et al. 2021]. 따라서 파운데이션 모델의 훈련은 상당히 에너지 집약적일 수 있지만, 연구자들은 이러한 모델의 탄소 영향력이 가장 탄소 배출량이 낮은 에너지 그리드를 선택함으로써 부분적으로 완화될 수 있음을 입증했다[Henderson et al. 2020].

탄소 오프셋은 모든 데이터 센터에서 탄소가 없는 재생 가능한 전기가 이용할 때까지 임시방편으로 제안되었다. 이 전략에는 한 활동에서 탄소 배출량을 줄이기 위해 다른 쪽에서 배출량을 상쇄하는 것이 포함되어 있다. 그러나 대부분의 경우, 탄소 오프셋은 처음부터 CO2를 방출하지 않는 것보다 분명하게 더 나쁜 해결책이다. 일부 탄소 오프셋 프로그램은 부정적인 영향을 줄 수도 있다. 예를 들어, 나무 심기 캠페인에 대한 연구는 그들이 득보다 더 큰 해를 끼칠 수 있음을 보여준다. 그러한 캠페인은 지역의 생물 다양성을 감소시키고 산림 토양의 탄소 저장 능력을 줄이는 단작(특정 종 하나의 나무만 심는 것)을 일으킬 수있다. 이로 인해 처음부터 탄소를 배출하지 않는 경우보다 탄소 오프셋을 사용할 때 더 많은 탄소 배출이 발생할 수 있다. 따라서 파운데이션 모델을 훈련 시키거나 배포 할 때는 배출을 상쇄하기 위해 탄소 오프셋에 의존하는 대신 가능한 한 탄소 방출을 최소한으로 설계하는 것이 좋다.

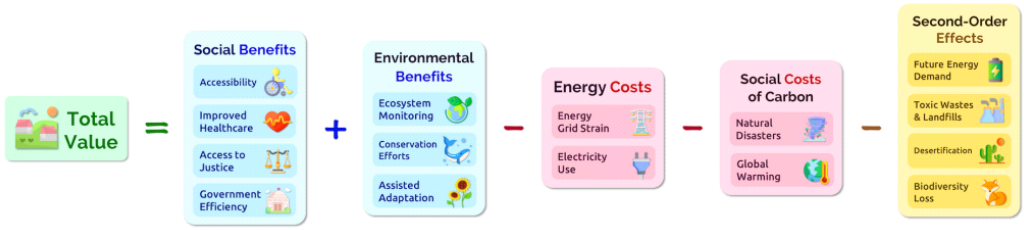
저탄소 지역에서 학습을 실행할 수 없는 경우 다른 완화 전략을 활용하여 불필요한 에너지 사용을 줄여야한다.

- 보다 효율적인 하드웨어 사용
- Mixed-precision training 사용[Micikevicius et al. 2017] 또는 양자화[Gholami et al. 2021]
- 보다 효율적인 아키텍처 사용(예 : 기본 트랜스포머 아키텍처 대신 진화된 트랜스포머 아키텍처 사용 또는 희소 모델 사용)
- 모델 증류 또는 증류된 모델 사용 (예 : [Sanh et al. 2019])
- 에너지 비용을 줄일 다른 최적화 전략 사용

오픈 소스 프로젝트 및 클라우드 컴퓨팅의 관리자는 “그린 디폴트”가 가장 효과적인 탄소 발생 완화 전략으로 알려져 있기 때문에, 기본 설정을 가장 효율적인 상태로 설정하기 위해 노력해야 한다. 다른 완화 전략은 최근 문헌에서 찾을 수 있다[Strubell et al. 2019; Henderson et al. 2020]. 또한 우리는 에너지 사용량을 줄이고 완화하면 컴퓨팅 액세스가 제한된 사람들이 모델에 쉽게 더 접근하게 할 수 있게 만드는 이점이 있다는 점을 주목해야 한다. 그러나 모델이 주로 추론에 사용되는 경우, 예를 들어 응용 프로그램 제품에 배포된 경우, 종종 응답 속도 문제로 인해 모델을 탄소 집약적이지 않은 에너지 그리드로 이동할 수 없다. 위에서 지정된 완화 전략을 사용하는 것 외에도, 이 경우 제안된 파운데이션 모델이 가진 이점을 보다 에너지 효율적인 대안으로 평가하는 것이 중요하다.

● 파운데이션 모델을 사용하기 전에 비용과 혜택을 평가해야 한다.

탄소 배출 완화(또는 완화가 불가능한 곳)를 향한 가능한 많은 단계를 수행 한 후에는 파운데이션 모델이 요구되는 사이즈 또는 파운데이션 모델을 꼭 사용해야하는지 여부를 확인해야 한다. 이러한 비용 편익 분석(cost-benefit analysis)은 다음을 고려해야한다.



[그림 23] 파운데이션 모델 배포를 위한 비용 편익 분석의 시각화. 모델의 총 가치는 먼저 모델의 순 긍정적인 사회적 혜택과 모든 환경적 혜택을 고려하여 근사할 수 있다. 그런 다음 모델을 교육하고 배포하는 데 드는 부정적인 에너지 비용, 모델을 교육하기 위해 배출되는 탄소의 사회적 비용 및 이차 환경 영향을 뺀다. 순 비용이 이점보다 크면 기초 모델 개발자와 대규모 배포자는 피해 감소 전략을 고려해야 한다. 더 효율적인 모델을 배포하거나 모델을 배포하지 않는 것이 포함될 수 있다.

(출처 : On the opportunities and risks of foundation models, 2021)

(1) 파운데이션 모델을 배포하는데 있어 사회적 비용과 환경 비용이 모델의 사회적 이익보다 더 큰가?

(2) 계산적으로 단순하고 저렴한 또 다른 접근 방식(예 : 훨씬 더 효율적인 파운데이션 모델 또는 단순 baseline)이 비슷한 사회적 이익을 얻을 수 있는가?

이 trade-off를 평가하기 위한 단순화된 체계는 모델 M의 전반적인 영향을 다음과 같이 간주된다.

$$V(M) = S(M) - C(M) - \mathcal{O}(M)$$

그림 23은 이 방정식과 각 변수에 들어갈 수 있는 비용과 이점을 나타낸다. 여기서 M는 모델이며 S는 달러로 환산된 순 사회적 혜택 및 환경적 이익이다. S는 건강, 의료, 정의(justice)에 대한 접근성, 빈곤 감소, 환경 모니터링 개선, 생태계 보존 노력 등을 지원함으로써 증가할 수 있다.

C는 에너지 사용으로 인해 발생된 탄소의 사회적 비용이다. 이것은 방출된 탄소가 사회적으로 미래에 입힐 피해를 현재의 금전적 가치로 환산한 값을 나타낸다. 2017년 미국 환경보호청(EPA: Environmental Protection Agency)이 추산한 탄소의 사회적 비용에 대한 추정치는 배출된 CO2 톤 당 최대 105 달러 (2007년 기준)였다.

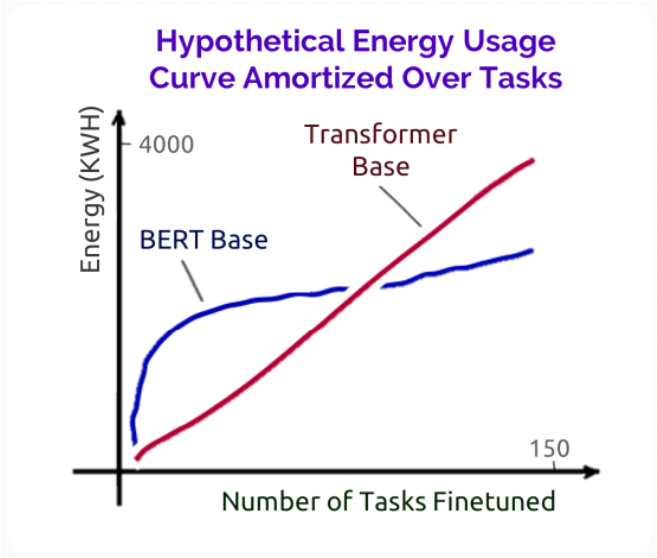
는 모델의 에너지 비용이다. 예를 들어, 2021년 4월 기준, 미국의 평균 주거용 에너지 비용은 kWh 당 약 $0.1376이다. 이 변수에 에너지 그리드에 대한 부담 증가로 인한 비용이 추가될 수 있다. 예를 들어, 최근의 연구에 따르면 에너지 그리드의 인터럽트의 발생을 평균적인 수요 기준으로 정규화 되었을 때 kW 당 $15.9 달러까지 높아질 수 있다고 제안되었다 [Sullivan et al. 2015]. 는 기타 2차 환경 효과의 사회적 비용으로, 다음을 포함한다.

- 증가된 칩 수요 및 칩 생산으로 인한 복합적 탄소 영향력
- 칩 제조와 관련된 다른 환경적 영향. 예를 들어, 실리콘 밸리의 독성 폐기물 부지 생성과 같이 건강에 미치는 영향이 불평등하게 사회적 취약한 인구에 분포되어 있는 경우 또는 만성 질환 문제와 관련된 대만의 칩 제조에서 발생하는 오염이 있다.
- SCC 모델에 아직 포함되지 않은 기후 변화의 복합적 효과. 예를 들어, 사막화 가속, 많은 종이 멸종 위기에 처하게 되는 빠른 생태계 변화, 영구 동토층이 녹아서 생기는 탄소 배출량 증가가 있다.
- 칩 생산 용량에 대한 불필요한 부담.

이 분석에서 중요한 것은 탄소의 경제적 이익과 사회적 비용이 여러 집단에 불평등하게 분배될 수 있으며, 가난한 지역 사회가 기후 변화에 더 큰 영향을 받고 부유한 지역 사회는 모델의 이득에 더 큰 영향을 받을 수 있다는 점을 고려하는 것이다[Bender et al. 2021]. 따라서, 해당 공식으로 분석을 수행 할 때, 사회적으로 이로운 점과 해로운 점을 주어진 조직이나 국가보다 더 광범위하게 고려해야 한다. 이 경우 $V(M)$은 분포로 간주 될 수 있으며 이상적으로는 전 인구에 걸쳐 동등하게 분포되어야

한다. 분포가 지나치게 고르지 않은 경우, 예를 들어 모든 이점이 모델 설계자에게 떨어지는 경우, 모든 피해는 모델의 혜택을 받지 못할 집단에 주어지는 경우, 설계자는 모델을 배포하기 전에 불평등 완화에 실질적으로 더 많은 노력을 기울여야 한다.

물론, 공식의 각 구성 요소를 평가할 때 사용해야 할 방법론에는 몇 가지 불확실성이 있다. 각 수식 중 대부분에서 경험적 추정 방법은 데이터 소스 및 탄소의 사회적 비용을 평가하는 서로 다른 매커니즘과 같은 현상에 따라 여러 규모로 범위가 넓어질 수 있다. 금전적으로 정량화하기 어려울 수 있는 추가적인 외부 효과도 계속해서 고려해야 한다. 그러나 이 비용 편익 분석의 핵심 내용은 방정식에서 각 수식의 금전적 평가 자체가 아니라 이러한 각 효과의 존재와 상대적 중요성이다. 우리의 목표는 이러한 trade-off 고려를 시작하기 위한 고수준의 프레임 워크를 제공하는 것이다. 향후 연구는 이러한 각 값을 정량화하는 방법에 대한 더 많은 지침을 제공하는 것이다.



[그림 24] 파운데이션 모델(이 경우 BERT Base)이 처음부터 훈련된 트랜스포머 모델보다 에너지 비용이 낮은 지점을 보여주는 미세 조정의 상각에 대한 가상 예시. [Strubell et al.,2019]으로부터 BERT 사전학습을 위한 에너지 비용을 추정했다. 그리고 [Chaudhary et al., 2020]으로부터 다운스트림 작업에 대한 미세 조정 비용을 추정했다. 또한 [Strubell et al., 2019]를 통해 선형적으로 증가하는 트랜스포머의 학습 비용과 비교했다. 80개 미만의 작업에 BERT를 사용할 경우 초기 에너지 비용은 복구되지 않는다. 그리고 그 이상의 경우는 BERT가 처음부터 훈련된 모델보다 더 에너지 효율적이다.

(출처 : On the opportunities and risks of foundation models, 2021)

마지막으로, 이러한 요소들도 각 실행별 평가가 아닌 모델의 수명주기에 걸쳐 평가되어야 한다는 점을 유의해야 한다. 모든 새로운 작업에 대해 처음부터 훈련해야 하는 baseline 모델을 고려해 보자. baseline은 다운스트림 작업에서 동등한 성능을 달성하기 위해 많은 비용이 드는 하이퍼 파라미터 검색이 필요할 수 있다. 대조적으로, 파운데이션 모델은 초기 사전 학습 절차에 대한 비용을 감수해야

하지만, 미세 조정은 훨씬 간단하고 에너지 효율적일 수 있다. 파운데이션 모델은 전체 수명에 걸쳐서 baseline보다 탄소 효율이 높을 수 있다(그림 24). 더욱 효율적인 미세 조정 메커니즘은 이 상각을 더욱 향상시킬 수 있다.

그러나 미세 조정의 효율성은 보장되지 않는다. 일부 파운데이션 모델은 많은 baseline보다 더 비효율적일 것이다. 예를 들어, 매개 변수가 적은 작은 모델을 사용하는 것이 에너지 효율 개선으로 이어질 것이라고 가정 할 수 없다. 하이퍼 파라미터 튜닝 비용 또는 기타 최적화로 인해 파라미터의 수는 일부 경우에서 에너지 효율과 상관 관계가 없는 것으로 나타났다. 따라서 파운데이션 모델 개발자는 대규모 학습을 위한 노력을 시작하기 전에 모델과 미세 조정 메커니즘의 효율성을 엄격하게 평가해야 한다.

● 탄소/에너지 충격은 체계적으로 보고되어야 한다.

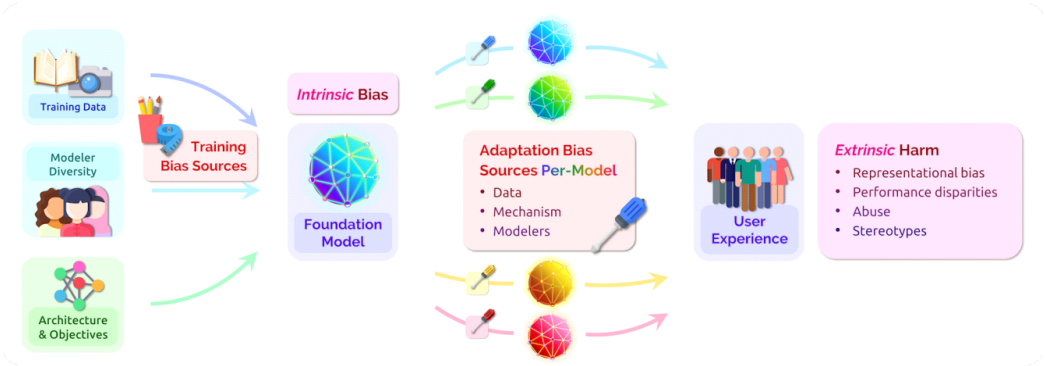
파운데이션 모델 연구원과 엔지니어가 모델의 컴퓨팅, 에너지 및 탄소 비용을 보고하지 않는 한 비용 편익 분석을 수행 할 수 없다. 우리는 파운데이션 모델 개발자, 제공자 및 큐레이터가 이러한 측정지표와 파운데이션 모델 제작에 사용된 탄소 감소 전략을 보고하도록 권장해야 하며, 이를 참고하여 탄소 영향력 명세서의 예시와 보고를 용이하게 할 수있는 도구를 확인할 수 있어야 한다. 연구원의 경우는 이러한 보고를 출판 과정에서 수행하게 되지만 업계에서는 투명성 메커니즘을 채택하여 배포된 모델에 대한 이러한 평가지표를 보고하도록 권장해야 한다. 이는 산업 및 학계 내에서 정책적 권장 사항을 설정하는 데 도움이 될 뿐만 아니라 다운스트림 사용자가 탄소 친화적인 사용 패턴을 식별하는 데 도움이 될 것이다. 표준화된 보고는 컴퓨팅에 대한 접근이 제한된 사람들이 접근할 수 있는 모델을 결정하는 데에도 도움이 된다.

에너지 및 탄소 영향에 대한 더 많은 보고를 장려하기 위해, 우리는 컨퍼런스에서 그린 배지 제공, 컨퍼런스에 제출하기 위한 관련 평가지표 보고 요구, 더 많은 투명성을 제공하도록 하는 대규모 파운데이션 모델 배포자에 대한 로비, 이러한 지표의 표준적인 보고에 대한 학계 및 산업의 보편적인 전문가적 규범 변화와 같은 방법을 제안해야 한다.

C 사회적 요소

파운데이션 모델은 불평등한 결과를 낳을 가능성이 있다. 특히 불평등한 분배로 인해 사람들을 부당하게 대우할 수 있다[Hellman 2021]. 다른 AI 시스템과 마찬가지로, 파운데이션 모델은 불공정한 결과를 낳고, 권력 시스템을 강화하고, 기술의 부정적인 결과를 소외된 사람들에게 불균형적으로 분배함으로써 기존 불평등을 악화시킬 수 있다. 이러한 문제는 알고리즘의 공정성과 AI 윤리, 인종과 기술, 사회와 기술의 공존에 대한 더 광범위한 질문과 관련이 있다. 파운데이션 모델은 미세 조정되기 전까지는

특정한 목적이 없는 중간 자산이며, 그들의 유해성을 이해하려면 그들의 속성과 작업별 모델 구축에서 수행하는 역할에 대한 추론이 필요하다.



[그림 25] 파운데이션 모델 내에 존재하는 내재적 편향(Intrinsic Bias)은 미세 조정 중에 도입된 편향(Adaptation Bias)과 함께 특정 다운스트림 애플리케이션의 맥락에서 사용자가 경험하는 외적 피해(Extrinsic Harm)를 결정하는 다양한 훈련 편향 소스(Training Bias Sources)의 결과물이 될 수 있다. 즉, 이러한 편향들은 많은 응용 프로그램에 전파될 수 있기 때문에 사용자가 경험하는 이러한 피해를 다양한 프로세스 및 원천에 귀속시키는 것은 중요하면서도 어려운 일이다.
(출처 : On the opportunities and risks of foundation models, 2021)

● 내재적 편향

내재적 편향은 파운데이션 모델의 속성 자체에서 나오는데, 그 유해성은 파운데이션 모델이 미세 조정될 때만 실현되므로, 잠재적인 편향이라고도 말할 수 있다. 내재적 편향은 표현 편향(representational bias)의 형태로 많이 연구되는데, 사람들은 악의적인 고정 관념, 부정적인 태도에 의해 잘못된 표현을 할 수 있으며, 혹은 과소 표현되거나 완전히 지워질 수도 있고(예를 들면, LGBTQ+ 정체성 용어나 아프리카계 미국인을 설명하는 데이터가 학습 데이터 셋에서 제외될 때), 혹은 과잉 표현이 될 수도 있다(예를 들면, 다수의 목소리를 증폭시키고 관점의 균질화 또는 단일 문화에 더 초점을 맞추게 될 때). 이러한 표현 편향은 모든 AI 시스템에 해당되지만, 파운데이션 모델 패러다임에서 그 중요성이 크게 높아진다. 동일한 파운데이션 모델이 무수한 응용 프로그램의 기반 역할을 하기 때문에, 사람들에 대한 표현 편향은 많은 응용 프로그램과 환경에 전파될 수 있다.

● 외적 유해성

사용자는 파운데이션 모델을 미세 조정하여 생성된 다운스트림 애플리케이션으로부터 특정한 피해를 경험할 수 있다. 이러한 해악은 정보 검색 시스템에 의해 생성된 흑인 여성에 대한 성적 묘사, 남성 대명사로 기본 설정된 기계 번역 시스템에 의한 성별 파악 실패 또는 유해한 고정관념의 생성과 같은 표현일 수 있다. 파운데이션 모델에 기반한 대화형 에이전트가 독성 콘텐츠 또는 미세 공격으로 사용자를 공격하는 경우와 같은 남용으로 구성될 수도 있다. 이러한 모든 사용자와 직접 대면하는

행동들은 사용자에게 심리적 피해를 끼치거나 유해한 고정관념을 강화시킬 수 있다.

개인이 경험하는 피해 외에도 집단이나 하위 집단도 집단 차원의 성능 격차와 같은 피해의 대상이 될 수 있다. 예를 들어, 시스템은 아프리카계 미국인의 영어로 된 텍스트나 음성에서 성능이 저하되거나, 인종, 성별 및 보험 상태를 가진 소수 그룹에 대한 임상 기록에서 의료 조건을 잘못 감지하거나, 피부색이 어두운 사람의 얼굴을 감지하지 못할 수 있다. 고위험 영역을 포함하여 파운데이션 모델이 더 널리 적용됨에 따라 이러한 차이는 더 나아가 더 심각한 해악으로 소용돌이칠 수 있다. [Koenecke et al., 2020]은 아프리카계 미국인 영어 사용자가 음성 인식 기술을 신뢰성 있게 사용할 수 없는 경우(예: 파운데이션 모델의 불평등으로 인해), 이는 특정 파생 제품(예: 음성 비서, 보조 기술)의 혜택을 받을 수 없음을 의미할 수 있으며 이러한 기술을 고용 면접 혹은 법정 절차를 기록하는 데에 사용할 경우 불이익을 받을 수 있음을 논의하였다. 좀 더 일반적으로, 이러한 그룹 차원의 해악을 특징짓는 것(그리고 피해를 입은 사람들을 위한 정의를 위해 노력하는 것)은 AI 커뮤니티가 그룹 기반 편견과 사회 그룹에 대한 이해를 개선하기 위해 요구되어진다.

● 추가 고려 사항

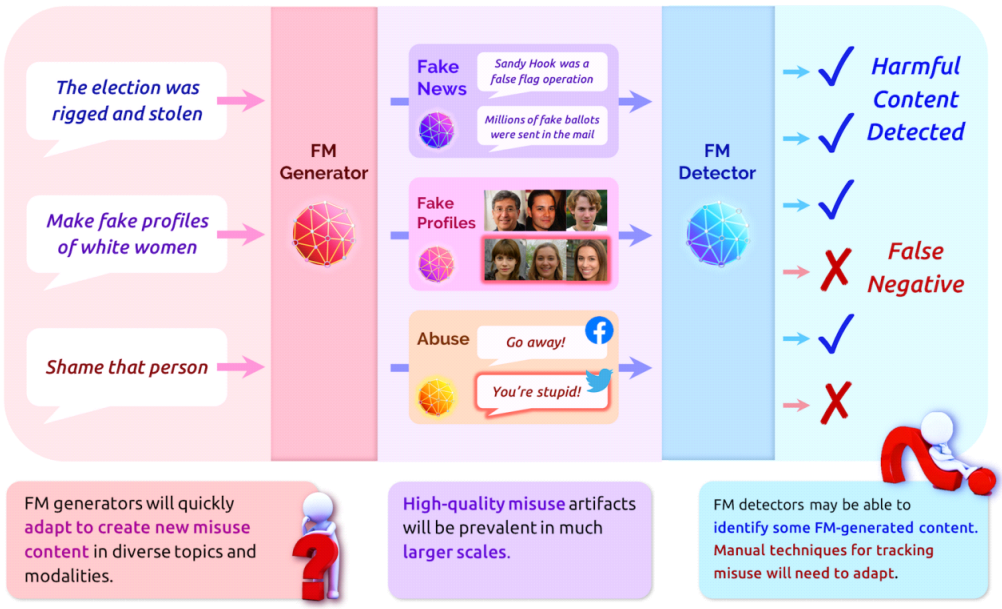
파운데이션 모델의 유해성을 보다 완벽하게 이해하기 위해서는 내재적 편향과 외적 유해성 모두에 대한 추가 문서화가 필요하며, 내재적 편향과 외적 유해성 사이의 관계를 명확히 해야 한다. 파운데이션 모델의 불평등한 영향은 학계와 산업계 모두에서 과소 대표되는 소수 집단에 의해 주로 경험될 것이다.

파운데이션 모델의 해악을 완전히 특성화하고 적절하게 개입하기 위해서는 파운데이션 모델의 특성과 미세 조정 과정까지 소스를 추적하고, 나아가 개별 편향 소스의 역할로 분해할 수 있어야 한다. 데이터 소스 추적은 경험한 피해에 대한 윤리적 및 법적 책임을 귀인하는 데 필수적이지만, 귀인에는 인과 관계 및 영향력과 같은 문제를 예측하는 새로운 기술 연구가 필요할 것이다.

여러 유형의 데이터는 파운데이션 모델을 기반으로 응용 프로그램의 동작과 관련된 외부 유해성을 형성한다. 파운데이션 모델을 훈련하는 데 사용되는 훈련 데이터, 파운데이션 모델을 미세 조정하는 데 사용되는 다운스트림 작업 데이터, 테스트에 사용되는 사용자 데이터 또는 상호 작용 모두 포함된다. 이러한 모든 데이터 소스에 대해 데이터의 속성(예: 혐오 발언, 욕설, 미시적 공격, 고정 관념)은 파운데이션 모델(및 이의 미세 조정된 파생물)의 편향에서 나타날 것이다. 훈련 데이터와 관련된 데이터 관행(예: 데이터 큐레이션, 데이터 선택 및 데이터 가중치)과 파운데이션 모델에 의해 획득된 내재적 편향 간의 관계는 현재로선 불분명하다. 따라서 이러한 관계를 명확히 하기 위한 향후 연구가 필요하다. 파운데이션 모델은 일반적으로 엄청난 규모의 훈련 데이터를 필요로 하기 때문에 문서뿐만 아니라 데이터 편향과 모델 편향의 관계를 명확하게 설명하기 위한 포괄적인 과학적 탐구에도 향후 연구가 더 필요하다.

모델링 결정(예: 학습 목표, 모델 아키텍처, 미세 조정 방법)은 파운데이션 모델과 그 파생물의 편향에 영향을 미치므로 경험된 외적 유해성에 영향을 미친다. 기존 연구는 파운데이션 모델이 훈련 데이터 편향을 증폭시켜 머신 러닝 및 딥 러닝 모델에서 보이는 추세를 확장한다는 것을 보여주었지만[Zhao et al. 2017], 모델 속성이 어떻게 이러한 편향 증폭의 원인이 되는지는 여전히 불분명하다. 또한, 파운데이션 모델을 직접 적용하는 것은 (규모 때문에) 불가능할 수 있다는 점을 고려할 때, 이러한 모델을 압축하거나 보다 효율적으로 만들기 위한 노력도 편향을 증폭시키는 것으로 보인다. 또한 파운데이션 모델을 증폭시키는 것은 사회적 행동을 수정하고 후속 훈련 데이터를 수정하는 사회학적 변화를 유도하는 피드백 루프에 의해 악화될 수 있다. 이러한 형태의 피드백 효과는 다른 애플리케이션의 불평등을 악화시킬 수도 있다. 따라서, 파운데이션 모델에 표준 벤치마크를 도입하는 작업, 편향을 측정하고 경험한 유해성을 문서화하기 위해 다양한 사용자 그룹과 함께 사용자 연구를 수행하는 것은 편향과 불평등을 적극적으로 강조하는 확실한 과정이 될 것이다.

● 오용



[그림 26] 파운데이션 모델은 유해한 콘텐츠를 생성할 수도 있지만 이것들을 탐지할 수도 있다.
(출처 : On the opportunities and risks of foundation models, 2021)

파운데이션 모델의 오용(misuse)은 사람들이 파운데이션 모델을 의도했던 바 대로 사용하지만(예: 언어 생성) 그것의 능력을 의도적으로 악용하여 집단이나 개인에게 해를 끼치는 상황을 말한다.

파운데이션 모델은 이전의 AI 방법보다 훨씬 더 높은 품질의 인간처럼 보이는 콘텐츠를 자동으로 생성하여 기만하는 행위에 오용될 수 있다. 고품질 이미지 합성(딥페이크)이나 텍스트를 생성하는 파운데이션 모델의 능력이 개인을 괴롭히는 데 악용될 수 있다. 딥페이크는 이미 괴롭힘의 목적으로 사용되었다. 예를 들어, 인도의 탐사 저널리스트인 Rana Ayyub는 포르노 비디오에 얼굴을 겹쳐 놓은 고품질의 딥페이크의 표적이 되어 수개월 동안 공적인 삶을 떠나게 되었다. 파운데이션 모델은 종종 멀티 모달이기 때문에 유사하게 말이나, 동작 또는 글을 가장할 수 있으며 잠재적으로 피해자를 당황하게 하거나 위협하거나 갈취하기 위해 잘못 사용될 수 있다.

파운데이션 모델은 콘텐츠 제작 비용을 상당히 줄여 악성 사용자가 유해한 공격을 수행하기 위한 진입 장벽을 더욱 낮출 것이다. 2017년, 러시아의 미국인을 대상으로 한 영향력 행사 예산은 1,220만 달러였다[DiResta et al. 2018]. 그러나 더 최근, 러시아의 개인들은 허위 정보 캠페인의 일환으로 미국 프리랜서들에게 기사당 75-200달러를 지불했다. 파운데이션 모델은 이러한 한계 비용을 낮출 것이다. GPT-3와 같은 파운데이션 모델은 콘텐츠를 생성할 때 실수를 할 수 있지만, 콘텐츠 작성자를 직접 고용하는 것보다 이를 해결하기 위해 소수의 편집자를 고용하는 것이 더 실현 가능할 것이다.

파운데이션 모델의 생성 기능은 넘치는 오용 기회를 제공하지만, 동일한 기능은 유해 콘텐츠의 강력한 탐지기로도 쓰일 수 있다. 파운데이션 모델의 대화형 및 멀티모달 인터페이스의 개선을 통해 파운데이션 모델 오용 탐지를 개선할 수 있을 것으로 보인다. 파운데이션 모델의 신속한 학습 능력은 파운데이션 모델이 초기에는 훈련되지 않았던 새로운 오용 전략에 대해 인간의 피드백을 통해 잘 적응하도록 할 수 있다.

5 참고문헌

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[5] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21) (pp. 2633-2650).

[6] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020, November). Generative pretraining from pixels. In International conference on machine learning (pp. 1691-1703). PMLR.

[7] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

[8] Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's' human'is not gold: Evaluating human evaluation of generated text. arXiv preprint arXiv:2107.00061.

[9] Clark, P., Etzioni, O., Khot, T., Khashabi, D., Mishra, B., Richardson, K., ... & Schmitz, M. (2020). From 'F'to 'A'on the NY regents science exams: An overview of the aristo project. AI Magazine, 41(4), 39-53.

[10] Dai, A. M., & Le, Q. V. (2015). Semi-supervised sequence learning. Advances in neural information processing systems, 28.

[11] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Association for Computational Linguistics (ACL). 4171-4186.

[13] DiResta, R., Grossman, S., & Siegel, A. (2022). In-House Vs. Outsourced Trolls: How Digital Mercenaries Shape State Influence Strategies. Political Communication, 1-32.

[14] DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., ... & Johnson, B. (2019). The tactics & tropes of the Internet Research Agency.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[16] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2021). A survey of quantization methods for efficient neural network inference. arXiv preprint arXiv:2103.13630.

[17] Gibson, J. J. (2014). The ecological approach to visual perception: classic edition. Psychology press.

[18] Gillis, T. B., & Spiess, J. L. (2019). Big data and discrimination. The University of Chicago Law Review, 86(2), 459-488.

[19] Ginsburg, J. C., & Budiardjo, L. A. (2019). Authors and machines. Berkeley Tech. LJ, 34, 343.

[20] Gordon, M. A., Duh, K., & Andrews, N. (2020). Compressing bert: Studying the effects of weight pruning on transfer learning. arXiv preprint arXiv:2002.08307.

[21] Haque, A., Milstein, A., & Fei-Fei, L. (2020). Illuminating the dark spaces of healthcare with ambient intelligence. Nature, 585(7824), 193-202.

[22] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16000-16009).

[23] Hellman, D. (2021). Big Data and Compounding Injustice. Journal of Moral Philosophy, forthcoming, Virginia Public Law and Legal Theory Research Paper, (2021-27).

[24] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1-43.

[25] Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., ... & Wu, Y. (2019). Gpipe: Efficient training of giant neural networks using pipeline parallelism. Advances in neural information processing systems, 32.

[26] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. arXiv preprint arXiv:2005.00813.

[27] Kajbaf, H. (2019). The First Amendment and Modern Technology: The Free Speech Clause and Chatbot Speech. Hastings Const. LQ, 47, 337.

[28] Kim, B., Kim, H., Lee, S. W., Lee, G., Kwak, D., Jeon, D. H., ... & Sung, N. (2021). What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. arXiv preprint arXiv:2109.04650.

[29] Kim, T., Song, G., Lee, S., Kim, S., Seo, Y., Lee, S., ... & Bae, K. (2022). L-Verse: Bidirectional Generation Between Image and Text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16526-16536).

[30] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.

[31] Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., ... & Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 10, 50-72.

[32] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[33] Lamo, M., & Calo, R. (2019). Regulating bot speech. UCLA L. Rev., 66, 988.

[34] Lemley, M. A., & Casey, B. (2019). Remedies for robots. The University of Chicago Law Review, 86(5), 1311-1396.

[35] Lilleberg, J., Zhu, Y., & Zhang, Y. (2015, July). Support vector machines and word2vec for text classification with semantic features. In 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) (pp. 136-140). IEEE.

[36] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... & Wu, H. (2017). Mixed precision training. arXiv preprint arXiv:1710.03740.

[37] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[38] Moravec, H. (1988). Mind children: The future of robot and human intelligence. Harvard University Press.

[39] Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N. R., Ganger, G. R., ... & Zaharia, M. (2019, October). PipeDream: generalized pipeline parallelism for DNN training. In Proceedings of the 27th ACM Symposium on Operating Systems Principles (pp. 1-15).

[40] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

[41] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

[42] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

[43] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Technical Report. OpenAI.

[44] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[45] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140), 1-67.

[46] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.

[47] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In International Conference on Machine Learning (pp. 8821-8831). PMLR.

[48] Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. (2020, August). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3505-3506).

[49] Sanh, V., Wolf, T., & Rush, A. (2020). Movement pruning: Adaptive sparsity by fine-tuning. Advances in Neural Information Processing Systems, 33, 20378-20389.

[50] Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 1668-1678).

[51] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv preprint arXiv:2211.05100.

[52] Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053.

[53] Sienčnik, S. K. (2015, May). Adapting word2vec to named entity recognition. In Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015) (pp. 239-243).

[54] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

[55] Sullivan, M., Schellenberg, J., & Blundell, M. (2015). Updated value of service reliability estimates for electric utility customers in the United States (No. LBNL-6941E). Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

[56] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[57] Xiang, A. (2020). Reconciling legal and technical approaches to algorithmic bias. Tenn. L. Rev., 88, 649.

[58] Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3712-3722).

[59] Zhang, Y., Warstadt, A., Li, H. S., & Bowman, S. R. (2020). When do you need billions of words of pretraining data?. arXiv preprint arXiv:2011.04946.

[60] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.

[61] 김병필. (2022). 대규모 언어모형 인공지능의 법적 쟁점. 정보법학, 26, 173-217.

NIA

Digital Insight 2023

파운데이션 모델의 이해와 미래전망

발 행 : 2023.3.31

발행인 : 황종성

발행처 : 한국지능정보사회진흥원(NIA) 정책본부 AI · 미래전략센터

기획 및 문의 : 우상근 책임연구원(sangkeun.woo@nia.or.kr)

작 성 : 성균관대학교 구자환 교수(jhkoo@skku.edu), 김도형 연구원

Digital Insight 2023

NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY