초거대 인공지능 생성 모델 동향 연구

정단호, 김운*, 정유철**
성균관대학교, 디엠티랩스*, 금오공과대학교**

요 약

최근 일반 인공지능(AI)를 넘어서서 차세대 AI 모델로 초거대 생성형 AI 모델들이 자리를 잡고 있다. 초거대 AI란 많은 양의 파라미터 수를 갖고 있는 모델을 의미한다. 파라미터 개수가 많아질수록 데이터로부터 다양한 정보를 학습할 수 있다. 1,750억 개 파라미터를 보유하고 있는 초거대 챗봇모델 ChatGPT가 성능을 입증하면서 기존 인공지능 트렌드에 큰 변화를 가져왔다. 본고에서는 초거대 생성형 AI 발전에 따른 언어 모델 및 음성 모델의 연구동향 및 시장동향에 대해 살펴보겠다.

Ⅰ. 초거대 생성 AI

전통적으로 통계 분야에서는 '모델이 크면 좋지 않다'라는 주장을 하고, 딥러닝분야에서는 '모델이 크면 클수록 좋다'라고 말한다. 반면, 두 분야 모두 '데이터가 많으면 많을수록 좋다'라는 사실에 동의한다. OpenAI의 Deep Double Decent 논문[1] 발표이후, 연구자들은 데이터가 많다는 전제 하에, 일부 예외는 있지만, 신경망 모델의 크기를 최대한 크게 하고, 오랫동안 학습하는 것이 성능향상에 도움이 된다는 실험결과를 토대로 점차 더 거대한 모델을 구현하게 된다.

현재 비약적 성능을 자랑하는 HW(HardWare) 기반으로 인공지능 기술은 최근 가파른 속도로 발전하고 있다. 과거 단층 퍼셉트론 모델에서 현재 초거대 생성형 AI 모델의 탄생까지 그리 오랜 시간이 걸리지 않았다. 초거대 생성형 모델은 사전에 학습한 규칙대로 동작하는 기존 인공지능과 달리 방대한 양의 파라미터를 보유하며 사용자의 입력을 통해서 새로운 데이터를 생성하는 방식의 인공지능을 의미한다.

〈그림 1〉은 OpenAI에서 개발한 이미지 생성형 모델인 DALL-E가 "Two hunters, the left hunter has a rabbit head, the right hunter has a deer head, digital art"라는 문구를 통해 생성한 그림이다.

다음으로 최근 화제가 되고 있는 ChatGPT로 유명해진 초거대 생성형 언어 모델이 있다. ChatGPT는 GPT-3.5기반으로 OpenAI에서 출시한 초거대 인공지능 대화형 챗봇이다. 기존 대화형 챗봇은 학습한 데이터 셋 내에서만 답변이 가능하고 반복적인 질문을 하면 동일한 답변을 되풀이하는 한계가 있었다. 반면, ChatGPT는 스스로 답변 생성이 가능하며 광범위한 분야를 다뤄 마치 사람과 대화하는 느낌을 받을 수 있다.

초거대 생성형 AI 모델은 기존 AI 모델보다 월등한 성능을 보여주고 있다. 그 차이점은 방대한 데이터 셋과 파라미터 개수로부터 나온다. 〈그림 2〉는 현존하는 초거대 AI모델들의 파라미터 개수의 변화를 상대적으로 도식하고 있다. 네이버 하이퍼클

그림 1. DALL-E가 생성한 이미지

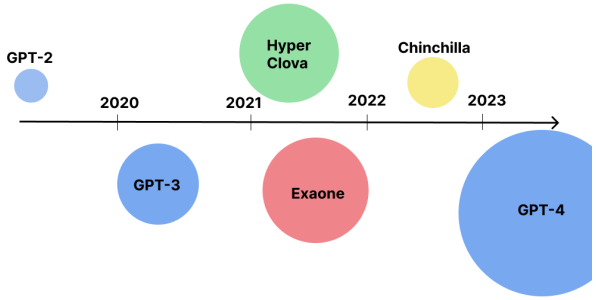


그림 2. 초거대 AI모델들의 개발 동향

로바, 카카오 KoGPT, 그리고 LG EXAONE 등 국내 대기업에서도 이미 연구가 한참 활발하게 진행 중이다. EXAONE과 네이버 하이퍼 클로바는 각각 3,000억 개와 2,040억 개 파라미터로 GPT-3보다 더 많은 양의 파라미터를 보유하고 있다.

Ⅱ. 기술 동향

1. 언어 모델

언어 모델(Language Model)이란 문장 특정 위치에 적합한 단어를 확률적으로 계산하여 예측하는 모델이다. 언어 모델은 목적에 따라서 다양하게 활용할 수 있다. 예를 들면 기계번역(machine translation), 음성인식(speech recognition), 그리고 글짓기(text generation)등이 있다.

신경망을 활용한 언어 모델은 2001년에 처음으로 제안되어 word2vec, sequence-to-sequence, 그리고 attention 기법을 이용한 트랜스포머(transformer) 모델까지 빠르게 발전해 왔다. 그 이후 트랜스포머 모델을 기반으로 인코더를 활용한 BERT 그리고 디코더를 활용한 GPT와 같은 다양한 파생 모델들이 출시됐다.

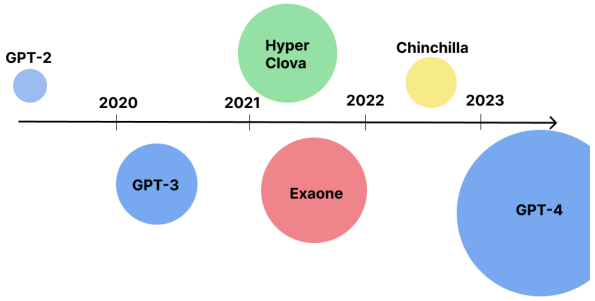


그림 3. 언어 모델 분류

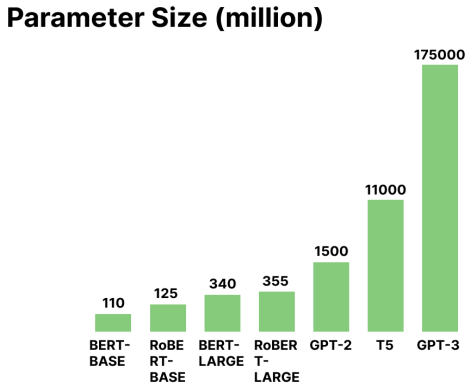


그림 4. 모델 별 파라미터 개수

〈그림 4〉 에서 언어 모델 사이즈가 점차 증가한 것을 알 수 있다.

언어 모델은 통계적 언어 모델 n-gram을 사용한 모델부터 딥러닝 기반 언어 모델이 있다. 인공 신경망 기반 언어 모델 초창기에는 워드 임베딩(word-embedding) 기법을 활용한 Word2Vec, GloVe, 그리고 FastText가 존재했다. 워드 임베딩은 단어를 벡터 형태로 표현하는 기법이며 단어의 의미를 벡터 공간 상의 위치로 표현한다. 이를 통해 단어 간의 유사도를 계산하거나, 단어들 간의 관계를 파악하는 등의 자연어 처리 작업에 이용할 수 있다.

그 다음에 순환 신경망(Recurrent Neural Network, RNN)을 활용한 LSTM(long-term dependency) 및 GRU(Gated Recurrent Unit) 모델이 있다. 순환 신경망은 입력과 출력을 시퀀스(Sequence)로 다루는 인공 신경망 모델이다. 시퀀스는 시간적, 공간적인 요소를 포함하며, RNN은 이러한 시퀀스 데이터의 패턴을 분석하고 학습하여 다양한 자연어 처리 작업에 이용된다. 하지만 순환 신경망은 벡터가 순차적으로 처리해야 하고 병렬화(parallelization)가 불가능하기 때문에 병목(Bottleneck) 문제가 발생한다는 단점이 있다.

2017년에 트랜스포머가 등장하면서 언어 모델은 새로운 국면을 맞이하게 되며, 이는 GPT모델로 이어지게 된다. 트랜스포머는 attention 기법을 이용해 입력 시퀀스의 모든 단어를 동시에 처리함으로 기존 순환 신경망이 가지고 있는 기울기 손실(vanishing gradient) 문제와 병목 문제를 해결했다[2]. 기본적으로는sequence-to-sequence 모델로써, 〈그림 5〉와 같이 크게 인코더(encoder)와 디코더(decoder) 부분으로 나뉜다. 인코더에서는 소스 시퀀스를 받고 정보를 압축하여 디코더에게 전달하며, 디코더는 전달받은 정보를 타깃 시퀀스로 변환하여 출력을 한다.

BERT(Bidirectional Encoder Representation from Transformers)와 GPT (Generative Pre-trained

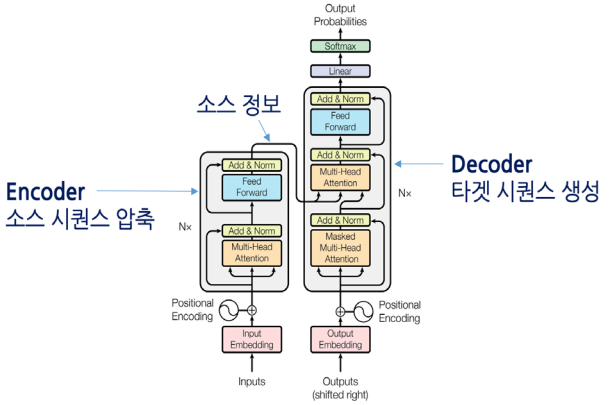


그림 5. Transformer 구조

Transformer)는 Transformer 기반 언어 모델로 가장 대표적인데, BERT는 인코더를 이용해 언어 이해에 치중된 반면에 GPT는 디코더를 이용해 문장 생성에 중점을 두었다.

기존 언어 모델들에 이어 초거대 언어 모델이 개발됐었는데, 이는 방대한 양의 데이터를 비지도 학습(unsupervised learning) 또는 자기 지도 학습(self-supervised learning) 방식을 이용해 학습한 딥러닝 모델을 일컫는다[3]. 대표적으로 OpenAI에서 개발한 GPT-1/2/3/4시리즈가 있다. 이렇게 학습된 모델은 단어 간 유사도 뿐만 아니라 문맥적으로도 구별이 가능하다. 추가적인 미세조정(fine-tuning) 또는 프롬프트 튜닝(prompt-tuning)을 하여 기계번역 또는 요약과 같이 원하는 작업에 알맞게 언어 모델을 사용할 수 있다[4].

파운데이션 모델(Foundation model)이란, 방대한 데이터셋으로 주로 자기 지도 학습을 이용해 훈련한 AI 신경망으로 추가적인 훈련없이 광범위한 작업 수행이 가능하다. GPT의 성공으로 다양한 파운데이션 모델들 (Google: BERT, LaMDA; MS: Turning NLG, MEB; OpenAI: GPT-3, DALL-E2, Codex; Meta AI: OPT-175B 등)이 등장하고 있다. 국내에는 네이버가 하이퍼클로바X, 카카오가 KoGPT의 개발 및 고도화에 투자를 확대하고 있다.

보다 최근에는 메타(Meta)에서 LLaMA (Large Language Model Meta AI)을 오픈소스로 공개했다. LLaMA는 언어 모델로 GPT3보다 적은 양의 파라미터에도 불구하고 고품질의 데이터 훈련으로 효율성을 높여 훨씬 적은 컴퓨팅 파워에서도 높은 성능을 낼 수 있다[5]. LLaMA을 기반으로 다양한 모델들이 공개되고 있다. 그중 LLaMA을 응용한 모델로 스탠퍼드 대학교에서 (Stanford Univ.) 개발한 Alpaca가 있다. Alpaca는 LLaMA 모델을 기반으로 instruction tuning 한 인공지능 챗봇 모델이다. 여기서 주목할 만한 점은 tuning을 위한 데이터를 self-instruct 방법을 통해 생성했다.

Self-instruct는 사전 학습된 LM(Language Model)을 이용해 데이터를 증강시키고, 그 데이터를 다시 학습시키는 방법이다.

Alpaca 이외에도 GPT4All와 Vicuna 모델이 있다. GPT4All은 Nomic AI가 LLaMA 와 GPT-3 오픈소스 버전인 GPT-J을 기반으로 미세조정한 모델이다. GPT4All 또한 Alpaca와 비슷하게 GPT-3 Turbo을 이용해서 prompt-generation쌍을 만들어서 미세조정을 하였다. GPT4All은 ChatGPT에 비해 다소 성능이 떨어지지만, 기존 언어 모델에 비해 우수한 성능에 개인용 노트북에서도 실행이 가능할 정도로 모델이 경량화 되어 있다는 장점이 있다.

하지만, 경량화된 Foundation 모델들도 새로운 데이터셋으로 학습하는 것은 상당한 연산량을 요구한다. 거대한 파라미터를 효율적으로 다루기 위한 방법(Parameter Efficient Fine-tuning, PEFT)으로 최근 Low-Rank Adaptation(LoRA)이 주목을 받고 있다. 초거대 AI 모델은 파라미터 수가 1000억 개가 넘어가기 때문에 fine tuning 하면 많은 시간이 소요된다. Low-Rank Adaptation은 기존 파라미터를 동결(freeze) 시키고 각 layer마다 rank decomposition matrices을 추가하는 방식으로 미세조정을 한다[6]. 사전학습 모델을 공유하면서 특정 작업에 조정된 모듈들을 만들어낼 수 있다.

마지막으로 초거대 텍스트 생성형 AI의 연구 동향을 살펴보기 위해 ChatGPT의 다음 모델인 GPT-4를 살펴보겠다. 〈그림 7〉과 같이 GPT-3의 파라미터 개수가 1,750억 개 였다면, GPT-4의 파라미터 개수가 조 단위로 넘어갔을 것이라 추정하고 있다. GPT-4은 더 복잡한 의미에 문장 예를 들면 풍자와 같은 문장을 이해하는 데 있어 GPT-3보다 우수한 성능을 보여줬다. OpenAI 공식 홈페이지에 따르면 GPT-4은 ChatGPT보다 Uniform Bar Exam 및 Biology Olympiad에서 더 높은 성능을 보여줬고, 입력으로 텍스트 뿐만 아니라 이미지 인식까지 가능한 멀티 모달(multimodal)을 도입했다[7]. 멀티 모달 AI는 입력으로 텍스트 데이터 외에도 다양한 감각 인터페이스를 통해 인식하는 방법을 의미한다. 초거대 생성형 모델이 도입되면서 한 가지 작업에만 국한된 것이 아닌 범용적으로 사용할 수 있는 인공지능(Artificial General Intelligence, AGI) 개발이 봇물을 이루고 있다.

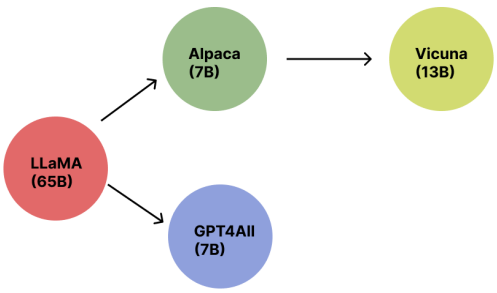


그림 6. LLaMA와 파생된 모델들

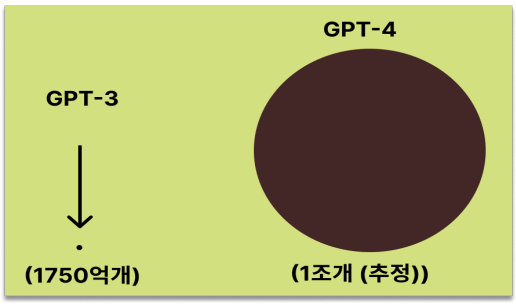


그림 7. GPT-3 vs GPT-4 파라미터 개수

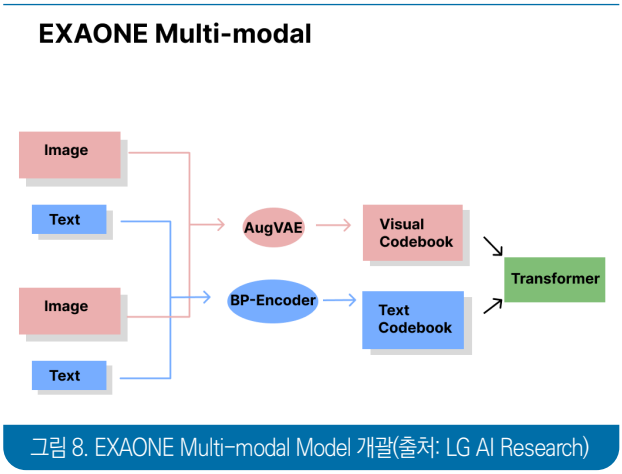


그림 8. EXAONE Multi-modal Model 개괄(출처: LG AI Research)

국내에서도 멀티 모달 AI 개발에 한창 열중하고 있다. LG에서 개발한 3000억 개 파라미터를 보유한 초거대 AI 모델 EXAONE은 텍스트 외에도 이미지로부터 양방향으로 학습하여, 텍스트를 통해서 이미지 생성이 가능할 뿐만 아니라 이미지가 주어졌을 때 이미지에 관한 짧은 설명 생성이 가능하다.

LG EXAONE은 트랜스포머 모델 기반으로 이미지와 캡션(텍스트)로 이루어진 데이터로 학습을 진행했다. 이미지를 학습할 때 AugVAE를 활용했는데, AugVAE는 인코더에서 이미지를 압축하는 중간 여러 과정에서 코드북 컴포넌트를 배치하여 이미지의 특징을 이미지 크기와 상관없이 기록이 가능케 했다[8]. 이러한 특징 덕분에 다양한 크기의 이미지 패턴 인식이 가능하고 생성 활용에 용이하다. 텍스트 데이터의 경우 BP-Encoder(Byte Pair Encoder)을 활용하여 텍스트 의미 단위의 코드북을 생성했다. 이렇게 만들어진 이미지와 텍스트 코드북을 결합하고 트랜스포머 모델을 이용하여 학습을 했고 이미지와 알맞는 텍스트를 이해하고 생성할 수 있는 인공지능이 탄생했다. LG EXAONE은 현재 이미지 외에도 음성, 제스처, 그리고 생체 신호등 다양한 입력 형태의 정보를 받아들일 수 있는 완전한 멀티 모달 개발에 진행 중에 있다.

카카오브레인에서 RQ-Transformer를 공개했다. [9]에 의하면 RQ-Transformer는 텍스트를 입력으로 받아 이미지 생성이 가능한 초거대 멀티 모달 인공지능이다. 기존에 있던 민달리(minDALL-E) 모델의 3배 크기와 학습 데이터셋 크기를 2배로 늘렸다. 기존에 이미지를 2차원 코드맵으로 표현했지만, RQ-Transformer에서는 3차원 코드맵으로 표현하여 이미지의 압축 손실을 줄일 수 있었다. 그 덕분에 이미지 생성 속도를 높이고 계산 비용을 줄일 수 있었다.

2. 음성 모델

음성에는 크게 음성 인식(Speech-To-Text, STT)과 음성 합성(Text-To-Speech, TTS)으로 분류할 수 있다. 음성 인식은 사용자가 마이크와 같은 장비로 음성을 입력하면 컴퓨터가 음성을 디지털화하고 이를 인식하여 텍스트로 변환하는 기술이다. 음성 합성은 사람의 목소리를 합성하여 텍스트를 음성으로 변환해 주는 기술이다. 본 장에서는 초거대 생성형 AI와 같이 발전하고 있는 음성 관련 연구들에 대해 살펴본다.

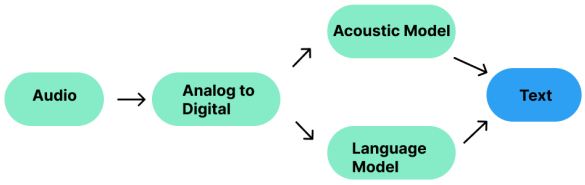


그림 9. 음성인식 과정

사람의 음성은 성별과 나이에 따라 발음의 종류, 음색, 높이가 다르다. 동일한 내용의 음성은 발화자가 달라도 같은 내용으로 인식해야 한다. 이를 위해 입력 받은 음성신호를 수학적으로 수치화 하는 전처리 과정이 반드시 필요한다. 일반적으로 디지털로 변환된 음성파형은 인식을 위해 균등한 크기로 자르게 되는데 이때 잘린 음성 조각은 푸리에 변환(Fourier transform, FT) 함수를 거쳐서 스펙트럼(Spectrogram)으로 변환한다. 스펙트럼이란 x축은 시간(time), y 축은 주파수(frequency), z 축은 진폭(amplitude)을 나타내는 소리를 수치화한 그래프이다. 수치로 변환된 소리는 신경망모델과 접목되어 텍스트로 변화는 과정을 거치게 된다.

최근 음성인식 연구도 대량의 음성데이터를 기반으로 학습한 모델을 활용하는 방향으로 발전하고 있다. 본 고에서는 최근 음성인식 연구에서 초거대 음성인식 모델로 대표적인 2가지 연구에 대해 살펴보겠다. 처음으로 살펴볼 모델은 Facebook에서 개발한 Wav2Vec2 모델이다. Wav2Vec2 모델은 53,000 시간 동안 레이블 되지 않은 데이터로 사전에 자기지도학습 방법으로 학습을 진행했다[10]. 자기지도학습 (self-supervised learning)이란 지도학습과 비지도 학습 중간 형태로, 자기 스스로 학습 데이터에 대한 분류를 수행한다. 소량의 레이블 된 데이터만 이용해도 미세조정(fine tuning)이 어느 정도 가능하다. [10]에 따르면 10분 정도 데이터로 미세조정 했을 시 Word Error Rate(WER)는 5%이하로 감소한다.

다음으로는 OpenAI에서 2022년 공개한 Whisper라는 모델로, 특정언어에 대한 음성인식 뿐만 아니라 특정 언어 음성

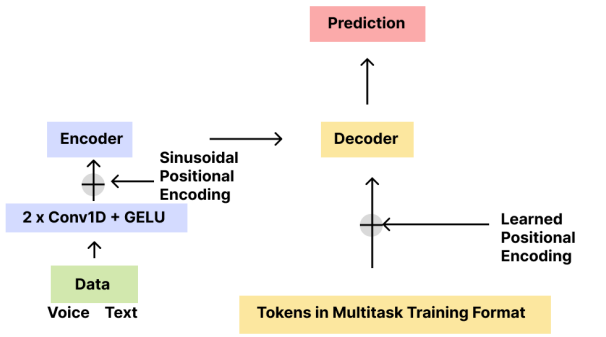


그림 10. OpenAI의 Whisper 구조

을 다국어로 바로 번역할 수 있는 종단형 인공지능 기술이다. 또한 발화 방식에 따라 구두점, 쉼표, 그리고 느낌표까지 표현이 가능하고, 번역 기능까지 포함되어 있다. Whisper는 〈그림 10〉과 같이 인코더(encoder)와 디코더(decoder)을 지닌 트랜스포머(transformer) 구조이다[11]. 680,000 시간 동안 다국어 및 다중 작업에 대해서 약한 지도 학습(Weakly Supervised Learning) 방법으로 사전학습을 했다. Wav2Vec2의 대부분이 비지도 학습 방식으로 사전학습이 되어 추가적인 미세조정이 필요한 반면에, Whisper는 노이즈가 섞인 데이터를 사용하더라도 레이블이 되어있는 데이터를 채택했다. 그 후에 레이블이 부족한 데이터 셋으로 전이 학습을 진행했다. 이를 통해 대규모 데이터 셋에서 학습된 모델의 일반화 능력을 유지하면서, 레이블이 부족한 데이터 셋에서도 높은 인식 성능을 달성할 수 있었다. 특히, Whisper는 zero-shot learning 사용하여 새로운 데이터 셋에 대해 미세조정 없이도 높은 성능을 보여줬다.

최신의 음성 합성 모델로 Microsoft에서 개발한 VALL-E[12]가 있다. VALL-E는 기존 장시간에 학습 시간이 필요했던 거에 비해 음성 샘플을 3초만 듣고 시뮬레이션이 가능한 것이 특징이다.

VALL-E는 화자의 자연스러운 말투 및 감정을 습득하고 심지

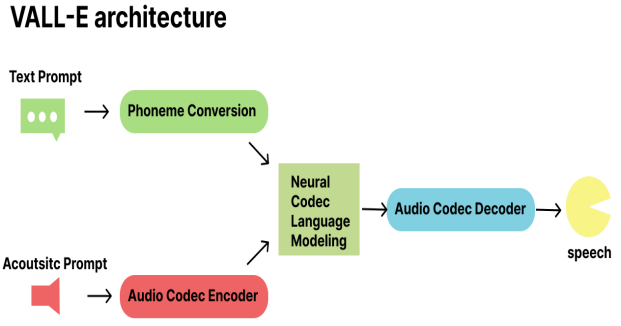


그림 11. VALL-E architecture

어 주변 환경소리까지 고려하여 음성 합성을 한다. 기존의 음성합성모델들이 음소를 통해서 스펙트로그램을 생성하고 파형을 만들어냈다면, VALL-E는 음소가 이산 코드(discrete code)을 통해 오디오 코덱 코드로 변환을 거치고 디코더를 통해 파형을 생성한다. VALL-E는 학습 당시에 음성 인식 모델을 통해 audio-only 데이터의 대본을 생성한다. 비록 대본과 음성 데이터에 노이즈가 있지만, 이는 기존 음성 모델과 비교했을 때 100배에 달하는 학습 데이터를 만들어낼 수 있다.

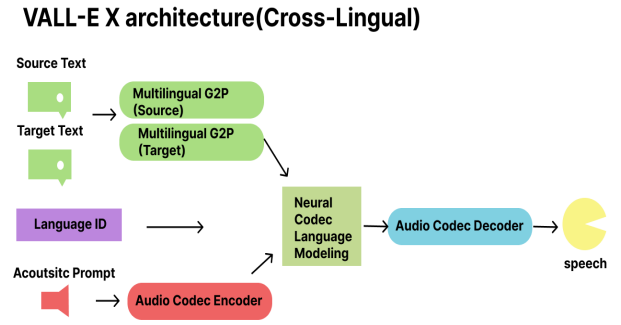


그림 12. VALL-E X architecture

음성 합성에서 해결하고자 하는 중요한 문제중의 하나는 동일한 화자의 다국어 음성 합성 문제가 있다. 이는 주로 동일한 화자의 다국어 음성 데이터의 부족과 모델 성능의 한계점으로 인해 최근까지 좋은 성능을 보이지 못했었다. Microsoft 사에서 제안한 cross-lingual 음성 합성을 지원하는 VALL-E X 모델은 [13]에 의하면 소스와 타깃 텍스트로부터 음소 시퀀스(phoneme sequences)을 추출하고 소스 소리 토큰들을 오디오 코덱 인코더에서 추출함으로써 타깃 언어의 소리 토큰들을 생성할 수 있다. VALL-E의 우수한 ICL(in-context learning) 능력 덕분에 다양한 zero-shot 음성 합성 작업을 수행할 수 있는 것이다. 다국어 작업을 하기 위해 소스 및 타깃 이중 언어로 구성된 음성과 텍스트 데이터로 학습을 진행했다.

〈그림 11〉과 〈그림 12〉를 보면 Language ID 모듈이 VALL-E에서 새롭게 추가됐다. Language ID 모듈은 VALL-E X이 특정 언어의 적절한 소리 토큰을 생성을 도와주는 역할이다. 예를 들면 중국인 화자가 영어를 말할 때 적절한 톤을 추가해준다. 이렇게 학습된 VALL-E X 모델은 cross-lingual 음성 합성 또는 speech-to-speech 번역에 응용될 수 있다.

3. 거대 AI 모델 활용 서비스

거대언어모델의 등장은 개별적인 언어 처리 작업들을 하나의

언어 모델로 동시에 처리할 수 있게 만들어줬다. 예를 들어, 과거에는 번역 그리고 요약과 같은 작업들을 각각 다른 모델을 이용해 학습을 시켰다면, ChatGPT와 같은 하나의 모델이 번역, 요약, 그리고 검색에 이르는 다양한 작업을 수행할 수 있다. 현재 ChatGPT는 월간 사용자 1억명을 돌파할 정도로 그 관심이 매우 높다.

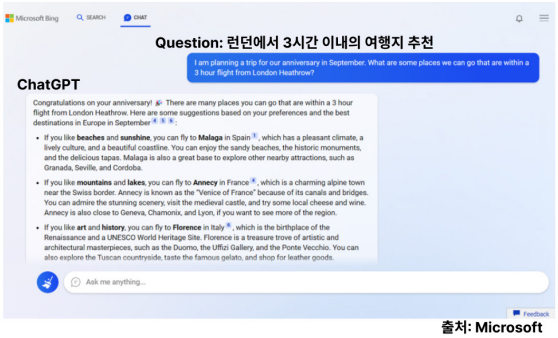
OpenAI에서ChatGPT와 음성인식엔진인 Whisper API를 저렴한 가격에 제공하고 있는데, 두 모델을 연동하면 가능해지는 서비스는 가히 무궁무진하다 하겠다. 기존 인공지능 가상 비서는 키워드를 통해 임무를 수행했지만, 이제 ChatGPT라는 생성형 초거대 AI 모델은 인간과 비슷한 수준으로 언어를 이해하고 높은 수준에 답변을 생성할 수 있기 때문이다. 즉, 대화형 인공지능에서 자연어 처리 부분이 초거대 AI 모델 덕분에 해결된 셈이다. Whisper음성엔진은 높은 성능의 음성인식률과 함께 다국어 음성인식기능을 제공하여 앞으로의 다각적 활용을 예측할 수 있는 부분이다.

Microsoft사 검색엔진인 Bing에 ChatGPT를 도입하면서 검색시장에 큰 변화를 가져왔다. 인터넷 검색을 하게 되면 원하는 내용을 찾기까지 오랜 시간이 걸릴 수도 있다. 하지만 현재 bing에서 검색 창에 궁금한 내용을 입력하면 서버 DB로부터 검색어 관련 내용을 출력하는 동시에 챗봇도 관련 결과를 보여준다. 또는 챗봇에게 직접적으로 물어봐서 관련 결과만 볼 수 있다.

또한 의료분야에서도 긴급상황 시 사람 대신 24시간 의료 상담을 지원하고 고품질 서비스를 제공하여 사용자의 편의성을 향상시켜줄 수 있다. ChatGPT는 단순한 생성 영역에 국한되지 않고 창작의 영역까지 가능하다. 스토리를 요구할 경우 비록 특별하지 않지만 이 세상에 존재하지 않는 스토리를 만들어준다.

ChatGPT의 출현 이후, 23년 3월 말부터 이를 활용한 [14]플러그인(plugin)들이 출시되고 있다. 플러그인이란 기존 프로그램을 다른 프로그램과 상호작용하기 위해 설치되는 소프트웨어 구성 요소를 의미한다. 즉, ChatGPT은 타사 프로그램과 연동하여 타사 서비스까지 제공이 가능해졌다. Third-party plugins 덕분에 ChatGPT가 웹 최신 정보에 접근이 가능하여 일기예보, 주식 시세, 그리고 최신 뉴스까지 조회를 하여 사용자에게 알려줄 수 있다. 또한 사내 문서와 같은 기존 문서를 연동하여 사내 업무 또는 지침 방안이 궁금할 때 즉각적으로 질문하고 대답을 해줄 수 있다.

마지막으로 숙박업소 예약과 같이 사용자의 행동을 대신해줄 수 있다. 현재OpenAI에서 지원하는 공식적으로 지원하는 플러그인은 대략 11종이지만 계속 증가할 것으로 보인다. 숙박 및 비행기 표 예매 서비스를 제공하는 "Expedia" 플러그인이 적용된 ChatGPT는 실제로 사용자 날짜에 맞춰 최저가 비행기 표 예매를 해주기도 한다.



출처: Microsoft

그림 13. bing에 추가된 ChatGPT

Ⅲ. 결 론

본고에서는 초거대 생성형 AI발전에 따른 언어 모델 및 음성 모델의 연구동향에 대해 살펴보았다. 초거대 생성형 AI가 떠오르는 만큼 신속하게 연구의 동향을 파악하고 기술을 선점하는 것이 중요하다. 언어 모델의 경우 현재 모델의 크기가 지속적으로 증가함에 따라 경량화 연구가 활발하게 진행중이다. 개인용 컴퓨터에서도 구동될 정도의 LLaMA, GPT4All, 그리고 Alpaca와 같은 가벼운 모델들이 생성AI의 대중화와 함께 주목을 받고 있다.

음성 모델은 기존 단일 언어의 데이터만 인식 또는 합성이 가능했지만, 최근 음성 인식의 경우 OpenAI사의 Whisper 출시로 동시 다국어 인식이라는 문제를 해결하면서 새로운 지평을 열었다. 음성 합성 또한 VALL-E X가 공개되면서 교차언어 음성합성에 활력을 불어넣을 것으로 기대된다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (IITP-2022-RS-2022-00156394)

참 고 문 헌

[1] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent:

Where bigger models and more data hurt. Journal of Statistical Mechanics: Theory and Experiment, 2021(12), 124003.

[2] 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.

[3] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). Self-Instruct: Aligning Language Model with Self Generated Instructions. arXiv preprint arXiv:2212.10560.

[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] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[6] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[7] https://openai.com/product/gpt-4

[8] 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).

[9] 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).

[10] 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.

[11] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

[12] Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., ... & Wei, F. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv preprint arXiv:2301.02111.

[13] Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., ... & Wei, F. (2023). Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. arXiv preprint arXiv:2303.03926.

[14] https://openai.com/blog/chatgpt-plugins

약 력

정단호

2018년 성균관대학교 공학사
2018년~현재 성균관대학교 소프트웨어학부 학부생
관심분야: 인공지능 자연어처리

김 운

2001년~2003년 충남대학교 컴퓨터과학과 (정보검색 및 자언어처리 전공) 석사
2003년~2007년 충남대학교 컴퓨터과학과 (정보검색 및 자언어처리 전공) 박사
2007년~2010년 한국전자통신연구원 포닥연구원
2010년~2017년 한국전자통신연구원 선임연구원
2018년~현재 ㈜디엠티랩스 대표이사 겸 사내연구소 소장
관심분야: 정보 검색 및 자연어 처리, 다국어 자동번역, 음성인식 및 음성합성, 인공지능 기술 기반 응용

정유철

1996년~2003년 아주대학교 정보 및 컴퓨터공학과 학사
2003년~2005년 한국과학기술원 정보통신공학(로봇 AI전공) 석사
2005년~2011년 한국과학기술원 전산학 박사
2009년~2013년 한국전자통신연구원 선임연구원
2013년~2017년 한국과학기술정보연구원 선임연구원
2017년~2022년 국립금오공과대학교 컴퓨터공학과 조교수
2022년~현재 국립금오공과대학교 인공지능공학과 부교수
관심분야: 정보 검색 및 자연어 처리 분야, 음성인식 및 음성합성, 인공지능 기술 기반 응용