Chapter
01

인간을 닮은 인공지능, 멀티모달 인공지능 기술 동향

김말희_한국전자통신연구원 책임연구원
허태욱_한국전자통신연구원 책임연구원
이일우_한국전자통신연구원 책임연구원

맹인모상(盲人摸象)은 장님이 코끼리 만지기라는 뜻으로, 여러 맹인이 코끼리의 다양한 부분을 만져보며 전체의 모습을 유추하는 이야기이다. 문제나 상황을 전체적으로 관찰하지 못하고 일면만을 보고 결론 내릴 때 생길 수 있는 오류와 제한된 인식에 대한 비유이다. 멀티모달 인공지능 기술은 시각, 청각, 촉각 등 인간의 다양한 감각 정보를 포함한 여러 데이터 소스를 통합, 분석함으로써 더욱 풍부하고 입체적으로 상황을 인식하는 기술이다. 또한, 멀티모달 인공지능 기술은 텍스트, 이미지, 동영상, 음성 등 다양한 형태의 인터페이스로 사용자와 소통함으로써 사용자 경험을 한층 향상시킨다. 본 고에서는 인식과 소통에 있어서 보다 인간적인 방식을 제공하는 멀티모달 인공지능 기술의 개념, 핵심 기술과 기술 동향에 대해서 살펴보고자 한다.

I. 서론

최근 몇 년간 인공지능 기술의 발전 속도와 영향력은 그야말로 놀랍다. Google의 Gemini나 OpenAI의 ChatGPT와 같은 챗봇은 이미 사람들의 업무나 일상생활에 깊숙이 파고들어 작업 방식을 바꿔 놓았다. Midjourney는 텍스트 기반 이미지를 생성하는 인공지능 기술로서, 창의적 아이디어를 빠르고 효율적으로 시각화할 수 있다. Midjourney가 그린 그림은 이미 회화 분야에서 예술작품으로 인정된 바 있다. 이는 예술가가 기존에 접근하기 어려웠던 아이디어나 컨셉을 연구하고 실험적인 작품을 만들어 낼 수 있게

* 본 내용은 김말희 책임연구원(☎ 042-860-1590, mariekim@etri.re.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.
***본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No.20202020900290)

도와줄 뿐 아니라 비예술가도 큰 장벽 없이 예술작품을 만들어 낼 기회를 제공한다. 인공지능을 활용하는 artificial intelligence art는 이미 하나의 장르로서 자리매김하고 있다.

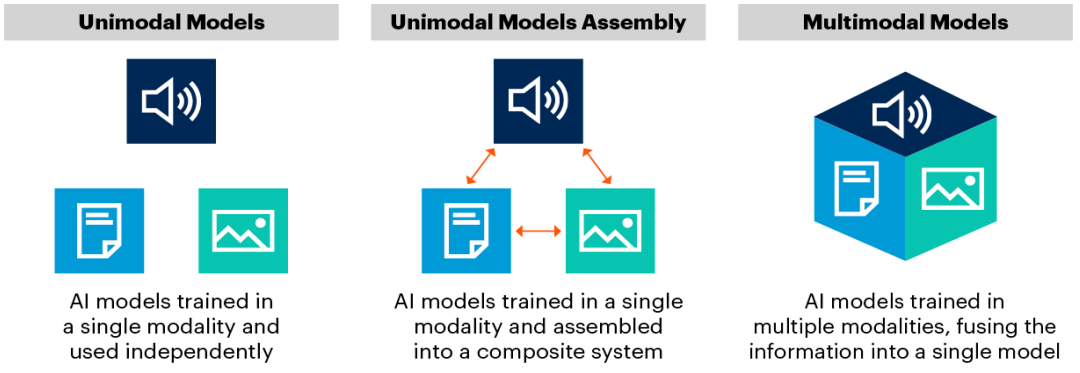
최근에는 다양한 형태의 데이터를 통합하여 처리함으로써, 인간과 더 자연스럽게 상호작용할 수 있는 새로운 기술 패러다임이 등장했다. 멀티모달 인공지능 기술이 그것이다. 이 기술은 텍스트, 이미지, 음성, 촉각과 같은 다양한 데이터 소스를 종합적으로 분석하고 활용함으로써, 기존의 단일 데이터 소스만을 사용하는 인공지능보다 훨씬 정확한 분석과 자연스러운 인터페이스를 제공한다. 멀티모달 인공지능 기술은 다양한 분야에서 활용될 수 있으며, 이미 의료[1], 자율주행[2], 챗봇, 교육 등 다양한 분야에서 성공적으로 활용되고 있다. 앞으로 멀티모달 인공지능 기술은 더욱 발전하여 인간의 삶을 더욱 편리하고 혁신적으로 만들 것으로 기대된다. 본 고에서는 멀티모달 인공지능 기술의 개념, 핵심 기술과 기술 동향을 살펴보고자 한다.

II. 멀티모달 인공지능 기술 개요

OpenAI의 Sora는 텍스트로 기술된 내용을 기반으로 [그림 1]과 같은 사실적인 영상을 만들어 내는 혁신적인 멀티모달 인공지능 기술이다[3]. 입력과 출력의 양식이 서로 다른 크로스 모달 기술이다. 영상 콘텐츠 제작 분야, 교육 및 학습 콘텐츠 분야, 예술

〈자료〉 CNN, OpenAI will now let you create videos from verbal cues, Retrieved, Feb. 15, 2024.

[그림 1] 멀티모달 인공지능 기술이 만든 영상



〈자료〉 Ramos, L., et al., Innovation Insight: Multimodal AI Explained, Gartner, G00798532, 2023, 1-13.

[그림 2] 다양한 양식으로 훈련된 멀티모달 모델로의 진화

및 디자인, 게임, 제품 설계 및 시뮬레이션 등 다양한 산업 분야에 활용할 수 있다.

양식(Modal)은 정보를 표현하거나 전달하는 방식을 의미한다. 멀티모달 인공지능 기술(Multimodal AI)은 텍스트, 이미지, 음성, 촉각 등 다양한 양식의 데이터를 동시에 처리하고 분석하여 인간과 자연스럽게 상호작용하는 인공지능 기술을 일컫는다. 더 나아가 멀티모달 인공지능 기술은 인간이 사용할 수 있는 감각에만 국한되지 않는다. 적외선 이미지, IoT 센서 등을 포함한 다른 유형의 데이터도 같이 분석될 수 있다. 이러한 멀티모달 인공지능 기술은 단일 양식을 처리하는 모델을 단순히 조합하는 것 이상이다. 멀티모달 인공지능 모델은 다양한 양식을 입력으로 받아서 동시에 훈련함으로써, 서로 다른 양식의 데이터들을 융합(fusion)할 수 있어야 한다. [그림 2]는 멀티모달 인공지능 모델에 대한 가트너의 개념도이다[4].

멀티모달 인공지능 기술은 다양한 형태의 정보를 통합하여 분석함으로써 더욱 정밀한 진단과 효율적인 정보 처리를 가능하게 한다. 예를 들어, 의료 분야에서는 환자의 진단 기록, X-ray, CT 이미지, 증상을 함께 분석하여 정확도 높은 진단을 제공할 수 있다. 또한, 삼성의 빅스비(Bixby)와 같은 챗봇은 음성, 텍스트, 터치와 같은 방식으로 사용자와의 자연스러운 인터페이스를 지원한다. 이처럼 멀티모달 인공지능 기술은 다양한 데이터 소스를 활용하여 새로운 지식의 발견이 가능하고, 다양한 데이터 양식을 활용하여 사용자 경험을 풍부하게 만드는 데 중요한 역할을 한다.

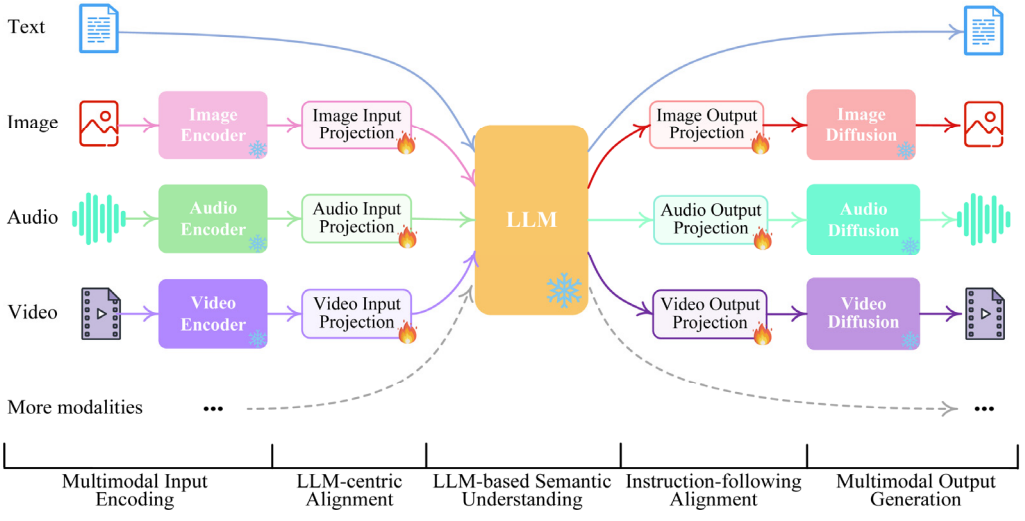
III. 멀티모달 인공지능 핵심 기술

멀티모달 인공지능 기술은 일반적으로 다음과 같은 3가지 구성 요소를 갖는다.

- 입력 모듈(Input Module): 다양한 유형의 데이터를 입력으로 받아서 전처리하거나 개별 양식을 위한 인코딩을 생성한다.
- 융합 모듈(Fusion Module): 양식별 처리된 정보를 다양한 융합 기법을 활용하여 통합한다.
- 출력 모듈(Output Module): 통합된 데이터 분석에 따른 결과를 텍스트, 이미지, 오디오, 동영상 등을 포함하는 다양한 양식으로 출력한다.

[그림 3]은 멀티모달 인공지능 시스템의 내부 구조 사례이다[5]. 다양한 양식의 데이터를 입력으로 받아서 통합적으로 처리하고, 다양한 양식의 데이터 형태로 출력한다.

이러한 멀티모달 인공지능 시스템을 구성하는 것에 있어서 중요한 핵심 기술로는 데이터 퓨전 기술, 모달리티 특정 인코딩(modal-specific encoding) 기술, 크로스-모달 학습(cross-modal learning) 기술을 들 수 있다.



〈자료〉 Wu, S., et al., NExT-GPT: Any-to-Any Multimodal LLM, GitHub, Retrieved from, 2023.

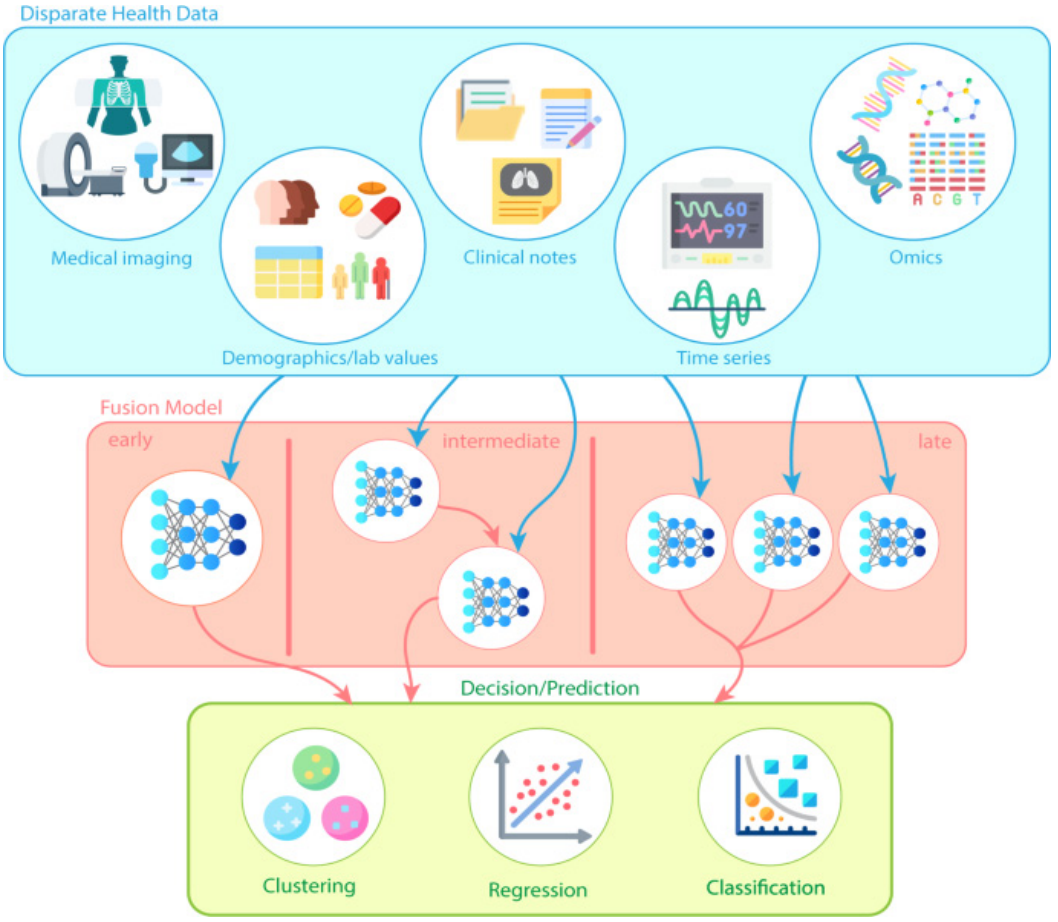
[그림 3] NExt-GPT 멀티모달 인공지능 구조

1. 데이터 퓨전 기술

멀티모달 인공지능의 핵심 기술은 다양한 양식의 데이터를 효과적으로 결합하는 것이다. 멀티모달 데이터 융합은 단순히 여러 양식의 데이터를 합치는 것이 아니라 각 양식의 데이터 간 관계를 파악하고 의미 있는 정보를 추출하는 과정이다[6]-[9]. 멀티모달 인공지능 기술의 데이터 퓨전 기술은 다음의 3가지 방식으로 정리될 수 있다.

- Early Fusion: 데이터 처리 과정의 초기 단계에 다양한 양식의 데이터를 결합하는 방식이다. 이 방식은 모델에 데이터를 입력하기 전에 다양한 데이터 소스를 통합하여 단일의 통합된 특징 집합을 생성한다. 예를 들어, 이미지의 픽셀 값과 관련 텍스트의 단어 벡터를 결합하여 하나의 큰 특징 벡터를 만들 수 있다. 이 방식은 다양한 양식 간의 상호작용을 모델이 쉽게 학습할 수 있게 하지만 각 양식의 고유한 특성을 잃을 수 있는 단점이 있다.
- Intermediate Fusion: 각 양식으로부터 추출된 특징들을 모델의 중간 단계에서 결합하는 방식이다. 이 접근 방식은 각 양식의 데이터를 먼저 독립적으로 처리하여 특징을 추출하고, 이렇게 추출된 특징들을 나중에 결합한다. 이 방식은 각 양식의 특징을 보존하는 동시에 다른 양식과의 상호작용을 학습할 수 있는 장점이 있다. 하지만, 최적의 융합 단계를 찾기가 어려울 수 있으며, 모델 학습 과정이 복잡할 수 있다.
- Late Fusion: 각 양식의 데이터를 독립적으로 처리하고, 각각에 대한 예측이나 결정을 내린 뒤에 이러한 결과를 결합하여 최종 결정을 내리는 방식이다. 이 접근 방식은 각 양식에서 도출된 결론이나 예측을 통합하는 것에 중점을 둔다. 이 방식은 각 양식의 독립적인 처리를 통해 유연한 모델 설계가 가능하지만, 양식 간의 복잡한 상호작용을 학습하는 데는 한계가 있을 수 있다.

[그림 4]는 3가지 방식의 Fusion Model에 관한 예를 보여준다. 모든 정보가 하나의 통합 모델로 흘러 들어가는 방식(early), 하나의 모델에서 나온 출력이 다른 모델의 입력이 되는 단계적 방식(intermediate) 그리고 각각의 데이터 양식이 별도의 모델링을 거



〈자료〉 Kline, A., et al., Multimodal machine learning in precision health: A scoping review, npj Digital Medicine 5, Article 171, 2022, 1-14.

[그림 4] Early, Intermediate, Late 데이터 퓨전 기술

친 후 앙상블 형태로 최종 결과가 취합되는 방식(late)을 보여준다[6].

2. 모달리티 특정 인코딩 기술

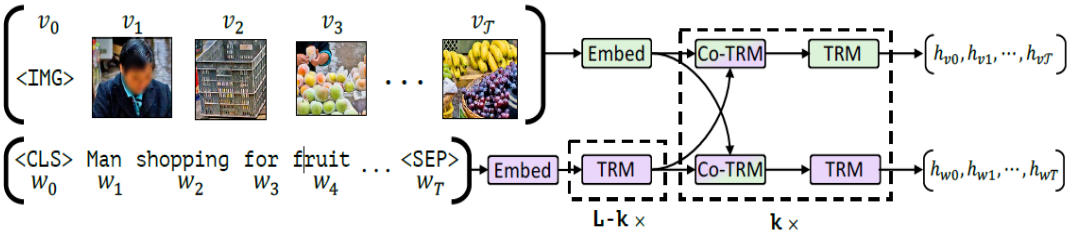
멀티모달 인공지능 기술은 이미지, 텍스트, 음성, 영상 등 다양한 양식의 데이터를 활용하여 인식하고 학습하는 기술이다[10][11]. 각 양식의 데이터는 고유한 특징과 구조로 되어 있으며, 이러한 특징을 효과적으로 활용하기 위해서는 양식별 특화된 인코딩

기술이 필요하다. 각 양식의 데이터 특징을 최대한 유지하여 인코딩함으로써 정보 손실을 최소화해야 하고, 양식 간의 연관성을 효과적으로 표현하여 멀티모달 인공지능 성능을 향상할 수가 있어야 한다. 양식에 따라서 다양한 인코딩 기술이 사용된다. 이미지의 경우, CNN을 이용하여 데이터의 특징을 추출한다. 텍스트의 경우, NLP 기술을 사용하여 텍스트의 특징을 추출한다. 음성의 경우 음성 인식 기술을, 영상의 경우 영상 처리 기술을 활용하여 양식별 특징을 추출한 인코딩을 생성한다.

3. 크로스-모달 학습 기술

크로스-모달 학습은 이미지, 텍스트, 음성, 영상 등 서로 다른 양식의 데이터를 함께 학습하여 모델의 성능을 향상시키는 기술이다[12][13]. 각 양식의 데이터는 서로 다른 정보를 제공하며, 이러한 정보를 통합하여 더욱 완전한 이해를 얻을 수 있다. 한 양식의 데이터가 부족할 때 다른 양식의 데이터를 활용하여 학습효과를 높일 수 있으며, 다양한 양식의 데이터를 학습하여 모델의 일반화 능력을 향상시킬 수 있다. 또한, 서로 다른 양식의 데이터를 연결하여 새로운 정보를 추출할 수도 있다. 크로스-모달의 학습 방법으로는 서로 다른 양식의 데이터를 공통된 표현 방식으로 변환하여 학습하거나, 각 양식의 데이터 중요도를 고려하여 학습 과정에 집중할 부분을 조절하거나, 한 양식 정보를 다른 양식의 정보로 변환하여 학습한다.

앞서 언급한 세 가지 핵심 기술들은 주요 멀티모달 모델의 구조에 잘 통합되어 있다. 대표적인 멀티모달 모델인 ViLBERT[14], ImageBind[15], CLIP[16]을 통해서 살펴보면 다음과 같다. ViLBERT, ImageBind은 양식별 임베딩을 별도로 학습한 후 이를 이용해서 공통된 표현을 학습한다. 임베딩은 인코딩된 데이터를 저차원의 벡터 공간으로 맵핑해주는 기술을 의미한다. 임베딩 과정에서 중요한 것은 비슷한 데이터들이 벡터 공간에서 서로 가깝게 위치하도록 하는 것이다. [그림 5]는 ViLBERT 모델이 이미지와 텍스트 각각의 임베딩을 학습한 후, co-attentional 계층(Co-TRM)을 이용해서 공통의 표현을 학습하는 방법을 도식화한 것이다.



〈자료〉 Lu, J., et al., ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision- and-Language Tasks. 33rd Conference on Neural Information Processing Systems(NeurIPS), 2019, 1-11.

[그림 5] ViLBERT 학습 방법

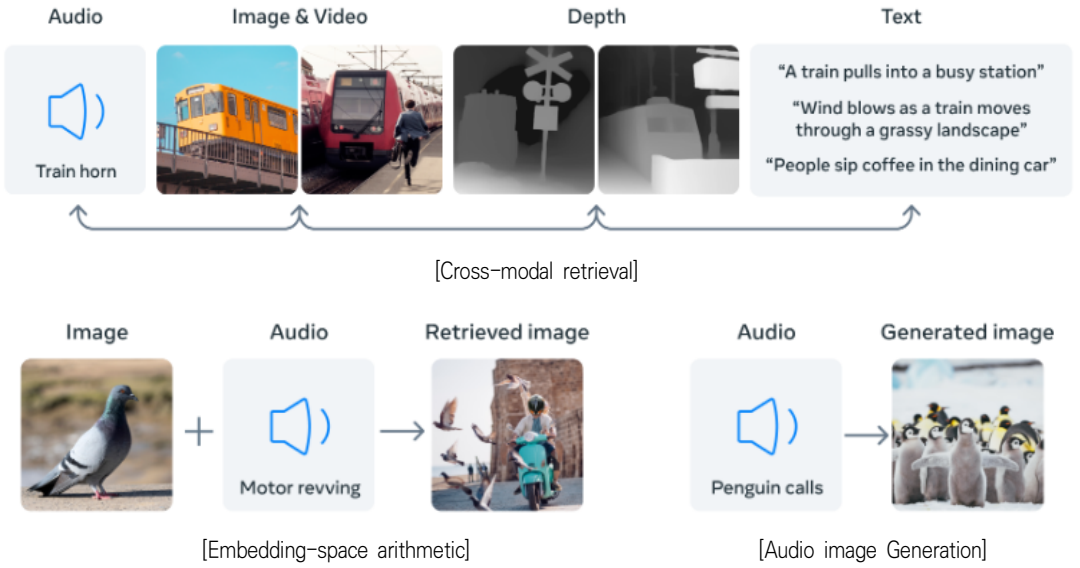
[그림 6]은 ImageBind가 이미지 데이터를 중심으로 다른 양식들을 조정(align)하여 공통의 표현을 학습하는 방법이다. 다른 양식의 임베딩을 모두 이미지 임베딩으로 조정한다. 예를 들면, 텍스트 임베딩을 이미지 임베딩으로 조정하고, IMU(Inertial Measurement Unit, 관성 측정 장치) 임베딩을 비디오 임베딩으로 조정한다. 양식별 임베딩을 학습한 후 이미지를 중심으로 임베딩을 조정하여 통합된 표현을 학습한다. ImageBind는 6개 양식(텍스트, 이미지/비디오, 오디오, 깊이, 열, IMU)의 데이터에 대한 공통된 표현을 학습한 것으로, 부족한 양식의 데이터를 다른 양식의 데이터를 이용하여 학습할 수 있다.

또한, 이처럼 다양한 양식의 데이터가 하나의 공간에서 통합적으로 표현되면 재미있는 응용 서비스들이 가능해진다. [그림 7]은 통합된 표현을 활용한 서비스 예제들이다.

Cross-modal retrieval은 오디오로 이미지나 비디오 검색을 하는 것과 같이 입력과 출력의 양식이 다른 검색이다. Embedding-space arithmetic은 연산자를 이용해서

〈자료〉 Girdhar, R., et al. . IMAGEBIND: One Embedding Space To Bind Them All, 2023.

[그림 6] ImageBind가 이미지를 중심으로 다른 양식의 데이터를 통합하는 방법

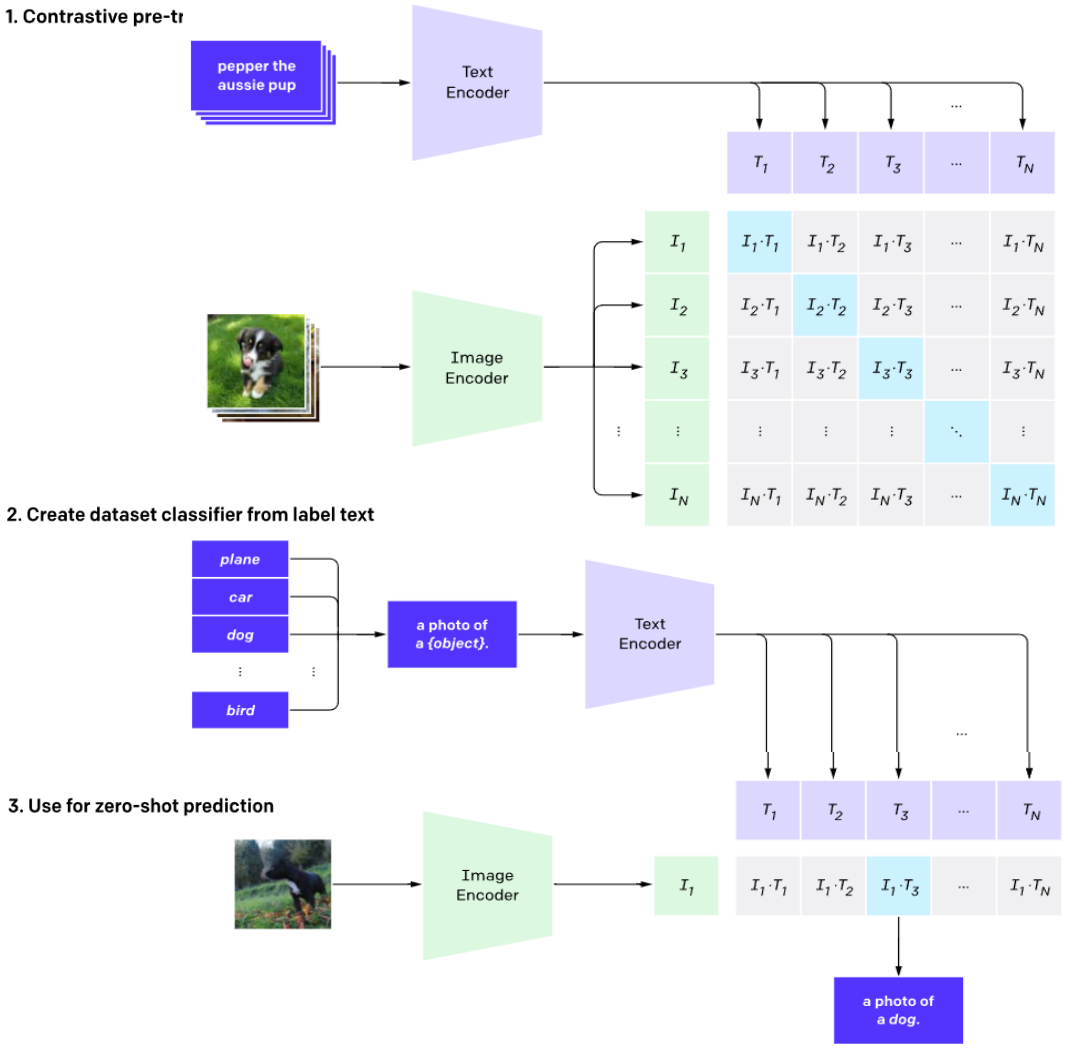


〈자료〉 Girdhar, R., et al., IMAGEBIND: One Embedding Space To Bind Them All, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 15180-15190.

[그림 7] ImageBind를 이용한 서비스 예시들

서로 다른 양식을 조합하는 서비스이다. 예를 들면, 새의 이미지에 오토바이 소리를 더해서 새와 오토바이가 있는 그림을 생성한다. Audio to image generation은 오디오 데이터 입력으로 이미지가 생성되는 서비스이다.

[그림 8]은 CLIP 모델이다. 양식별 인코더를 이용해서 생성한 임베딩 값을 contrastive learning 기법을 이용해서 관계있는 이미지 임베딩과 텍스트 임베딩 거리는 최소화하고, 무관한 이미지 임베딩과 텍스트 임베딩 거리는 최대화하도록 학습한다. 이렇게 학습된 모델은 zero-shot transfer가 가능하다. 즉, CLIP의 텍스트 인코더를 이용해서 이미지 라벨을 텍스트 임베딩 형태로 변형한 데이터 셋을 생성한다. 그리고, CLIP의 이미지 인코더 부분을 활용한 이미지 분류기를 만든다. 이렇게 하면, 사전학습에 활용되지 않은 이미지에 대해서도 튜닝작업 없이 분류할 수 있다. 이처럼 zero-shot transfer가 가능한 이유는 공유 임베딩 공간에서 이미지와 텍스트 사이의 의미적 관계가 충분히 학습되었기 때문이다.



〈자료〉 Radford, A., et el., Learning Transferable Visual Models From Natural Language Supervision, Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021, 1-16.

[그림 8] CLIP 기술 개요

IV. 멀티모달 인공지능 기술 동향 및 전망

가트너 보고서[17]에 따르면, 멀티모달 인공지능 기술의 시장 적용 시기는 향후 1년에서 3년 사이라고 한다. 기술 성숙도나 시장에서의 도입 의지가 충분하다는 것이다.

주요 기술 제공사로는 Google, Meta, Microsoft, Midjourney, NVIDIA, OpenAI가 있다.

1. 기술 동향

최근 이 분야는 텍스트와 이미지, 오디오, 비디오 등을 결합하는 고급 기술의 개발로 인해 눈부신 발전을 이루고 있다. 기술 동향은 크게 창의적 콘텐츠 생성, 강화된 사용자 인터페이스, 향상된 분석 능력, 사전학습 모델을 기반으로 하는 멀티모달 모델 연구 등의 형태로 나타나고 있다.

가. 창의적 콘텐츠 생성

멀티모달 인공지능 기술은 창의적 콘텐츠 생성에 큰 변화를 일으키고 있다. 예를 들어, OpenAI의 DALL·E, Google의 Imagen 그리고 Midjourney, Stable Diffusion, Parti와 같은 기술은 사용자가 제공한 텍스트 설명을 바탕으로 고품질의 이미지를 생성한다. 또한, Meta의 Make-A-Video, OpenAI사의 Sora와 같은 기술은 텍스트 설명을 바탕으로 동영상을 생성하여 멀티미디어 콘텐츠 제작의 경계를 확장하고 있다. 이러한 기술은 미술 작품을 제작하거나, 제품 디자인, 게임 및 가상 현실 콘텐츠 생성과 같은 분야에 활용할 수 있음으로써 관련 산업계에 큰 영향을 줄 전망이다.

나. 강화된 사용자 인터페이스

멀티모달 AI는 사용자와의 상호작용을 강화하는 것에 중요한 역할을 하고 있다. 예를 들어, Flamingo는 다양한 양식의 데이터 입력을 받아 사용자의 질문에 답변하거나 적절한 정보를 제공하는 기능을 수행한다. CLIP은 이미지와 관련된 텍스트를 통해 효과적으로 검색과 분류 작업을 수행할 수 있으며, 텍스트 기반 이미지 검색도 가능하다. Google Assistant, 삼성의 빅스비 등 챗봇의 경우 텍스트, 음성 등 사용자 맥락에 맞는 양식으로 사용자 인터페이스를 제공한다.

다. 향상된 분석 능력

ViLBERT, UNITER, ImageBind와 같은 기술은 다양한 양식의 데이터 통합 및 이해를 가능하게 한다. 이러한 기술은 인간의 인식 방식을 모방하여 다양한 양식의 데이터를 자연스럽게 이해하고 활용하여 다른 작업을 처리할 수 있도록 한다. 의료 영상 진단, 자율주행자동차, 고객 서비스 등의 분야에 활용할 수 있다. 또한, 멀티모달 인공지능 기술은 감정 인식, 의료 이미지 분석, 보안 시스템 등에서 데이터 분석의 정확성을 향상시킨다. 예를 들어, 감정 인식 시스템은 음성의 톤, 표정, 몸짓 등 다양한 양식의 데이터를 종합하여 사용자의 감정 상태를 파악한다. 의료 분야에서는 IBM Watson과 같은 시스템이 환자의 의료 이미지와 기록을 종합 분석하여 진단의 정확성을 높이고, 특정 치료법에 대한 추천을 제공할 수 있다. Tesla의 Autopilot 시스템은 카메라, 레이더, 초음파 센서를 포함한 멀티모달 센서 배열을 활용하여 주행 상황을 분석하고, 차선 변경, 속도 조절, 비상 정지 등의 기능을 자동으로 수행할 수 있게 한다. Waymo의 자율주행 기술은 라이다(LiDAR), 카메라, 레이더를 포함한 멀티모달 입력을 사용하여, 실시간으로 주변 환경을 이해하고, 다른 차량, 보행자, 자전거 등의 예측 불가능한 움직임을 예측하며 안전하게 반응할 수 있도록 지원한다.

라. 사전학습 모델을 기반으로 하는 멀티모달 모델 연구

ViLBERT, VL_BERT, VideoBERT, VD-BERT, Image-BERT, VIsualBERT, UNITER, Pixel_BERT, Fashion-BERT, S2VT, CLIP, ImageBind, DALL·E 3, Midjourney 등의 대표적인 멀티모달 인공지능 모델들은 사전 학습된 모델을 기반으로 멀티모달 모델을 개발한 사례들이다. 다양한 양식별 특징 추출과 양식별 상호관계를 함께 학습하기 위해서는 사전 학습된 모델을 이용하는 것이 효율적이기 때문이다. ResNet, Transformer, BERT 등의 사전학습 모델들이 활용된다.

2. 미래 전망

멀티모달 인공지능 기술은 인간의 인식 능력을 뛰어넘는 수준으로 발전하고 있으며, 다양한 분야에서 혁신을 이끌고 있다. 한편, 사용하는 데이터와 생성된 데이터들에 대한 지식재산권 문제와 프라이버시 문제, 데이터 편향성 문제 등이 큰 걸림돌이 될 전망이다.

멀티모달 인공지능 기술은 인간의 보조적 도구로서가 아니라 문제 해결에 있어서 보다 창의적이고 적극적인 역할을 할 것으로 전망된다. 예를 들어, 과학 연구에서는 과학자들과 협력하여 새로운 물질 개발이나 질병 치료법을 발견하는 것에 기여할 수 있고, 예술 창작 분야에서는 예술가들과 함께 음악 작곡이나 소설 집필, 미술품 창작 등을 수행할 수 있다. 또한, 사회 문제 해결에 있어서는 정부, 기업, 시민 사회와 협력하여 범죄, 환경 오염 등을 해결하며 지속 가능한 사회 구축에 기여할 것이다.

멀티모달 인공지능 기술은 개인의 취향, 생활 패턴, 건강 상태를 분석하여 맞춤형 의료, 교육, 여행 서비스를 제공함으로써 일상의 질을 향상시킬 것이다. 또한, 다양한 언어를 지원하는 인공지능 기술은 번역, 문서 작성, 데이터 분석 등의 업무를 자동화하여 업무 효율성을 높이는 데 기여할 것이다. 창의적인 활동에 있어서도 멀티모달 인공지능 기술은 작곡, 작사, 그림 그리기 등을 지원하여 인간의 창의성을 발휘할 수 있도록 지원할 것이다.

멀티모달 인공지능 기술은 새로운 산업과 일자리를 창출하고, 기존 산업의 구조를 변화시킬 것이다. 멀티모달 인공지능 기술을 개발하고 활용하는 전문가들이 많이 필요하게 되고, 멀티모달 인공지능 기술을 기반으로 새로운 서비스와 제품들이 등장할 것이다. 예를 들어, 멀티모달 인공지능 기반 가상현실, 증강현실, 로봇, 자율주행자동차 등 새로운 산업이 발전할 것이다. 또한, 멀티모달 인공지능 기술은 제조업, 유통업, 금융업, 의료 서비스 등 기존 산업의 생산성을 높이고 새로운 가치를 창출할 것이다.

기술적으로는 사용자 참여형 인공지능(User-in-the-loop AI)이 저변화될 것이다. 멀티모달 인공지능의 학습은 복잡 미묘한 데이터 양식들의 뉘앙스와 편향을 해결하기 위해서 인간 전문가의 개입이 요구될 것이다. 대표적인 방식이 RLHF(Reinforcement

Learning from Human Feedback)이다. 인간 피드백을 통한 강화 학습은 인간의 전문 지식을 통합하여 훈련의 품질을 향상시킬 수 있다[17].

V. 결론

멀티모달 인공지능 기술의 진화는 보다 인간다운 인식과 소통 능력을 갖춘 기술의 시대가 열린 것을 의미한다. 데이터 분석 능력과 상호작용 능력에 있어서 인간을 추월할 수도 있다. 또한, 창의적인 콘텐츠를 생성하는 능력을 갖춤으로써 예술 분야 활용도 많아지고 있다. 최근에는 의료, 자동차 그리고 고객 서비스 분야에서 멀티모달 인공지능 기술 적용 사례가 늘어나고 있으며, 이는 해당 기술이 실제 환경에서 기술시장을 주도하게 될 것임을 시사한다.

하지만 이러한 기술 발전은 사회적, 윤리적 책임이 함께 따라야 한다. 무분별한 데이터 사용과 생성은 개인의 프라이버시와 지식재산권 문제뿐만 아니라 산업, 사회, 정치적으로 큰 위협이 될 수 있다. 멀티모달 인공지능 기술로 만든 가짜 뉴스나 콘텐츠는 진짜처럼 보여서 사람들에게 어떤 정보를 믿어야 할지 혼란을 줄 수 있다. 그래서 공공의 의견을 조작하거나 사회적 분열을 일으키는 데 사용될 수 있다. 사실이 아닌 믿고 싶은 것을 믿는 경향이 있는 포스트 트루스 시대에는 멀티모달 인공지능 기술이 매우 위협적인 도구가 될 수 있는 이유이다.

따라서, 인공지능이 생성한 결과물의 신뢰성을 높이기 위해 다양한 기술적 조치가 필요하다. 이에 알고리즘의 편향성을 최소화하고, 사용한 콘텐츠의 출처를 명확히 하는 것, 생성물인지 아닌지를 구별할 수 있는 기술(출처 감지) 그리고 인공지능의 내부적 오류로부터 가짜 콘텐츠를 생성하는 것을 최소화하는 기술(환각 최소화)의 개발이 포함된다. 또한, 생성된 결과물에 대한 지식재산권 문제를 해결하는 것도 중요하다. 이 모든 조치는 멀티모달 인공지능 기술의 발전이 사회에 긍정적으로 기여할 수 있도록 보장하는 것에 필수적이다.

참고문헌

[1] Acosta, J.N., et al., Multimodal biomedical AI, Nature Medicine 28, 2022, 1773-1784.

[2] Xiao, Y., et al., Multimodal End-to-End Autonomous Driving, IEEE Transactions on Intelligent Transportation Systems, 23(1), 2022, 537-547.

[3] CNN, OpenAI will now let you create videos from verbal cues, Retrieved, Feb. 15, 2024.

[4] Ramos, L., et al., Innovation Insight: Multimodal AI Explained, Gartner, G00798532, 2023, 1-13.

[5] Wu, S., et al., NExT-GPT: Any-to-Any Multimodal LLM. GitHub, Retrieved from, 2023.

[6] Kline, A., et al., Multimodal machine learning in precision health: A scoping review, npj Digital Medicine 5, Article 171, 2022, 1-14.

[7] Manzoor, M.A., et al., Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications, ACM Transactions on Multimedia Computing, Communications and Applications, 20(3), Article 74, 2022, 1-34.

[8] Chawla, R., et al., Veagle: Advancements in Multimodal Representation Learning, 2024, 1-8.

[9] Pawłowski, M., et al., Effective Techniques for Multimodal Data Fusion: A Comparative Analysis, Sensors, 23, 2381, 2023, 1-16.

[10] Wang, J., et al., Exploiting Modality-Specific Features For Multi-Modal Manipulation Detection And Grounding, 2023, 1-6.

[11] Xu, L., et al., Learning Multi-Modal Class-Specific Tokens for Weakly Supervised Dense Object Localization, IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2023, 19596-19605.

[12] Aiello, E., et al., Cross-modal Learning for Image-Guided Point Cloud Shape Completion. 36th Conference on Neural Information Processing Systems(NeurIPS), 2022, 1-14.

[13] Zhang, Y., et al., Connect, Collapse, Corrupt: Learning Cross-Modal Tasks with Uni-Modal Data, The International Conference on Learning Representations(ICLR), 2024, 1-26.

[14] Lu, J., et al., ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 33rd Conference on Neural Information Processing Systems(NeurIPS), 2019, 1-11.

[15] Girdhar, R., et al., IMAGEBIND: One Embedding Space To Bind Them All, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 15180-15190.

[16] Radford, A., et el., Learning Transferable Visual Models From Natural Language Supervision, Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021, 1-16.

[17] Zimmermann, A., et al., Emerging Tech Impact Radar: Artificial Intelligence, Gartner, G00796195, 2024, 1-82.