

## II-2. 생성형 AI의 확장, 다양한 감각을 생성하는 멀티모달 AI

2024년은 멀티모달 AI 시대로의 본격적인 도입이 이루어진 해로 기억될 것이다. 오픈AI(OpenAI)의 GPT-4o(포오) 출시로 음성을 통해 AI와 대화하고 소통하는 것이 대중화되었고, 이는 영화 ‘HER’에서 묘사된 미래가 현실이 되고 있다. 또한 오픈AI의 소라(SORA), 구글의 루미에르(Lumiere), 중국의 클링(KLING) 등 동영상 생성 AI의 등장은 창조의 영역에서 AI의 역할과 가능성을 한 단계 더 높였다.

이러한 발전은 이전부터 대중에게 알려진 챗GPT(ChatGPT) 등의 언어 생성 AI와 미드저니(Midjourney), 달-E(DALL-E), 스테이블 디퓨전(Stable Diffusion) 등의 이미지 생성 AI에서 더 나아가, AI가 인간의 창의력을 확장하고 증폭시킬 수 있음을 보여준다. 멀티모달 AI는 단순히 텍스트, 이미지, 음성, 동영상 등 개별 영역에서의 성과를 넘어, 이들을 통합적으로 활용하여 새로운 형태의 콘텐츠와 경험을 창출할 수 있게 되었다.

이제 우리는 AI와 함께 창조의 한계를 뛰어넘는 시대로 접어들고 있다. 예술, 디자인, 영상, 음악 등 다양한 분야에서 AI는 인간의 상상력과 창의력을 확장시키는 도구로 자리매김할 것이다. 동시에 이러한 기술의 발전은 창작자와 AI의 협업, 저작권과 윤리적 문제 등 새로운 이슈들을 제기할 것으로 보인다.

멀티모달 AI의 등장은 우리가 정보를 소비하고 지식을 습득하는 방식, 그리고 창의적 작업을 수행하는 방식에 혁신을 가져올 것이다. AI와 인간의 협업을 통해 기존에는 불가능했던 영역들이 개척되고, 창조의 경계가 확장됨에 따라 우리 삶의 모습도 더욱 풍요로워질 것으로 기대된다. 2024년은 이러한 변화의 시작점으로 기록될 것이며, 앞으로 펼쳐질 멀티모달 AI 시대의 무한한 가능성을 기대해 볼 만하다.

- 오픈AI가 음성 생성 AI의 포문을 GPT-4o를 통해서 열었다. GPT-4o는 사용자와 실시간으로 대화하며 질문과 답변을 할 수 있다. 이에 오픈AI와 기술경쟁을 벌이고 있는 구글도 구글 I/O 2024에서 프로젝트 아스트라(Project Astra)를 공개했다. 사람처럼 보고 듣고 말하면서 사용자를 도와주는 모델이다. 또 아마존도 관련 기술에 속도를 내고 있다.<sup>8)</sup>
- 전세계가 오픈AI의 동영상 생성 AI인 소라의 공개 사용을 기다리는 가운데, 중국 내 콰이쇼우가 중국판 소라인 클링(KLING)을 2024년 6월 10일 발표했다. 클링은 오픈 액세스 형태로 제공되는데, 2023년 월 스미스 국수 먹는 동영상을 시연해 비판을 받았지만 1년만에 지적받았던 오류가 보이지 않았다. 이에 이제 중국이 동영상 생성 AI 분야에서 진전이 나타나고 있다.<sup>9)</sup>

8) 목소리 갖게 된 AI...감정까지 파악하는 ‘음성 비서’ 눈앞 (최진석, 2024)

9) 중국판 오픈AI, 소라 넘어선 동영상AI 첫선...“국수 먹는 남성 놀라워” (이상덕, 2024)

- AI가 급격하게 발전해온 과도기에 다수의 기업들은, 배우·성우들에게 AI 학습에 대한 불충분한 설명을 제공하거나, AI에게 작업물·초상권을 무단 학습시키기도 하고, 정당한 비용 또한 지불하지 않는 모습을 보여왔다. 그랬기에 배우, 성우들이 AI 무단 학습 및 악용에 크게 반대해온 것이다. 최근 오픈 AI가 GPT-4o를 공개했을 때, 스칼렛 요한슨이 자신의 동의 없이 목소리가 AI 보이스에 사용됐다고 주장한 것도 마찬가지다. 오픈 AI 측은 무단 학습한 결과물이 아니라고 주장하면서도, 해당 보이스를 일시 중단했다. 양쪽 중 누구의 주장이 맞는 지 시비를 가리지 않더라도, 이제 AI 무단 학습이라는 화두가 현실성 없는 논쟁이 아니게 된 시대 변화를 드러내주는 대표적인 사례다.<sup>10)</sup>

## ☑ 관련 주요 용어 알아보기

### 1. 기술적 방법론

#### 1.1. 멀티모달 AI(Multimodal AI): 오감을 활용하는 인공지능, 세상을 더욱 풍부하게 이해하다

##### ① 왜 나오게 되었는가?

기존의 AI는 주로 한 가지 종류의 정보만 다루었다. 예를 들어, 텍스트만 이해하거나 이미지만 분석하는 식이었다. 하지만 우리가 사는 세상은 그렇게 단순하지 않다. 우리는 매일 글도 읽고, 사진도 보고, 소리도 듣는다. 이렇게 다양한 정보를 동시에 처리하는 인간의 능력을 본떠 만든 것이 바로 멀티모달 AI이다. 마치 요리사가 여러 가지 재료를 사용해 맛있는 요리를 만드는 것처럼, 멀티모달 AI는 다양한 정보를 조합하여 더 풍부한 이해를 만들어낸다.

##### ② 쉽게 설명하면?

멀티모달 AI는 여러 가지 감각을 가진 인공지능이라고 생각하면 된다. 사람이 눈으로 보고, 귀로 듣고, 손으로 만지는 것처럼 이 AI는 텍스트, 이미지, 소리, 영상 등 다양한 형태의 정보를 동시에 이해하고 분석할 수 있다. 예를 들어, 영화를 볼 때 우리는 화면도 보고 대사도 듣고 자막도 읽는다. 멀티모달 AI도 이처럼 여러 정보를 종합하여 상황을 이해한다.

##### ③ 예시와 함께 좀 더 자세히 알아볼까?

실생활에서 멀티모달 AI의 활용 예를 살펴보자. 음식 주문 앱을 사용할 때, 여러분은 메뉴 사진을 보고 설명을 읽는다. 멀티모달 AI는 이 과정을 더욱 스마트하게 만들 수 있다. 예를 들어, 사용자가 “매운 음식 추천해줘”라고 말하면, AI는 음식 사진과 설명을 함께 분석하여 적절한 매운 음식을 추천할 수 있다. 또 다른 예로, 스

10) “AI로부터 성우들을 지켜달라” 미국 배우 노조 파업 확률 높아져 (김승준, 2024)

마트홈 시스템에서 멀티모달 AI는 집 안의 온도(숫자 데이터), 사용자의 음성 명령, 실내 CCTV 영상을 종합적으로 분석하여 가장 쾌적한 환경을 만들 수 있다.

#### ④ 왜 중요한가?

멀티모달 AI의 중요성은 현실 세계의 복잡성을 더 잘 반영할 수 있다는 점에 있다. 우리의 일상은 단순히 텍스트나 이미지만으로 이루어지지 않는다. 예를 들어, 친구와 대화할 때 우리는 말의 내용뿐만 아니라 표정, 목소리 톤, 몸짓 등을 종합적으로 이해한다. 멀티모달 AI는 이러한 인간의 복합적인 인지 능력을 모방하여 더 자연스럽고 정확한 인공지능 서비스를 제공할 수 있다. 이는 인간과 AI 사이의 상호작용을 더욱 원활하게 만들어, 우리의 삶을 더욱 편리하게 만들 수 있다.

#### ⑤ 어디에 활용되는가?

멀티모달 AI는 우리 일상 곳곳에서 활용될 수 있다. 예를 들어, 온라인 쇼핑에서 ‘이 옷과 어울리는 신발 추천 해줘’라고 하면, AI가 옷의 이미지와 설명을 분석하여 적절한 신발을 추천할 수 있다. 교육 분야에서는 학생의 표정, 음성, 답변 내용을 종합적으로 분석하여 개인 맞춤형 학습을 제공할 수 있다. 의료 분야에서는 X-레이 영상, 환자의 증상 설명, 과거 병력 등을 통합적으로 분석하여 더 정확한 진단을 내릴 수 있다. 또한 자율주행 자동차에서는 카메라 영상, GPS 데이터, 주변 소리 등을 종합적으로 분석하여 안전한 주행을 가능하게 한다. 이처럼 멀티모달 AI는 우리 삶의 다양한 영역에서 더 스마트하고 효율적인 서비스를 제공할 수 있다.

## 1.2. 제로샷 학습(Zero-shot Learning): 한 번도 본 적 없는 것을 이해하는 인공지능의 놀라운 능력

#### ① 왜 나오게 되었는가?

인공지능의 학습 방식 혁신이 필요해 등장한 제로샷 학습은 멀티모달 등장으로 그 필요성이 더욱 부각되었다. 멀티모달 AI는 이미지, 텍스트, 오디오 등 다양한 형태의 데이터를 통합적으로 이해하고 생성할 수 있는 AI를 말한다. 하지만 이렇게 다양한 데이터를 모두 대량으로 학습시키는 것은 현실적으로 불가능하다. 따라서 제한된 데이터로도 새로운 개념을 이해할 수 있는 제로샷 학습 기술이 멀티모달 AI에게 필수적인 능력이 되었다.

#### ② 쉽게 설명하면?

제로샷 학습은 멀티모달 AI가 한 번도 접해보지 않은 조합의 데이터도 이해할 수 있게 해주는 기술이다. 예를 들어, “노란색 바나나”라는 텍스트와 바나나 이미지를 본 적 있는 AI는 “파란색 바나나” 이미지를 처음 봐도 그것이 바나나라는 것을 알 수 있다. 이는 AI가 색상과 물체에 대한 개별적인 이해를 바탕으로 새로운 조합도 인식할 수 있기 때문이다.

### ③ 예시와 함께 좀 더 자세히 알아볼까?

멀티모달 제로샷 학습의 실제 활용 예시로 영화 추천 시스템을 들 수 있다. 이 시스템은 사용자가 본 영화의 장르, 배우, 줄거리 등 다양한 정보를 분석하여 새로운 영화를 추천한다. 그런데 사용자가 “우주를 배경으로 한 뮤지컬 영화”와 같이 기존에 없던 조합을 요청한다면? 제로샷 학습이 가능한 멀티모달 AI는 “우주”와 “뮤지컬”이라는 개념을 이해하고 있기에, 이 둘을 조합하여 적절한 영화를 찾아낼 수 있다. 비록 그런 영화를 직접 본 적은 없지만, 관련 개념에 대한 이해를 바탕으로 추론이 가능한 것이다.

### ④ 왜 중요한가?

멀티모달 환경에서 제로샷 학습은 AI의 확장성과 창의성을 크게 향상시킨다. 모든 조합의 데이터를 일일이 학습시키는 것은 불가능하므로, 제로샷 학습 능력은 AI가 새로운 상황에 유연하게 대처하는 데 핵심적인 역할을 한다. 또한 이는 AI의 창의력 향상에도 기여한다. 새로운 개념을 만들어내고 이를 활용하는 것은 인간의 창의력의 기반이 되는데, 제로샷 학습은 AI에게 이와 유사한 능력을 부여한다. 결국 제로샷 학습은 멀티모달 AI를 더욱 지능적이고 유연하게 만들어, 실생활에서의 활용 가능성을 크게 확장시킨다.

### ⑤ 어디에 활용되는가?

멀티모달 제로샷 학습은 다양한 분야에서 활용될 수 있다. 예를 들어, 시각장애인을 위한 음성 지원 앱에서는 사용자가 찍은 사진에 대한 설명을 제공한다. 여기에 제로샷 학습이 적용되면, 앱은 처음 보는 물체의 조합도 인식하고 설명할 수 있다. 또한 챗봇과 같은 대화형 AI에서도 제로샷 학습이 활용된다. 사용자가 이전에 없던 주제로 대화를 시도해도, AI는 관련 개념을 조합하여 적절한 응답을 생성할 수 있다. 이 외에도 자율주행, 로봇 지능, 창작 AI 등 다양한 분야에서 멀티모달 제로샷 학습은 AI의 핵심 기술로 자리매김할 것으로 예상된다.

## 1.3. 멀티모달 AI 학습의 핵심, 역전파 : 조율을 통한 오류 수정의 마법

### ① 왜 나오게 되었는가?

인공지능, 특히 딥 뉴럴 네트워크의 학습 과정을 혁신하기 위해 등장한 역전파가 멀티모달로 인해 중요성이 더 부각되었다. 멀티모달 AI는 이미지, 텍스트, 오디오 등 다양한 형태의 데이터를 통합적으로 처리하고 이해하는 AI를 말한다. 이러한 복잡한 데이터를 효과적으로 학습하기 위해서는 각 모달리티(데이터 형태)에 맞는 특징을 추출하고, 이를 융합하여 최종 결과를 도출하는 과정이 필요하다. 이 과정에서 역전파는 각 모달리티의 특징 추출기와 융합 모델의 파라미터를 최적화하는 데 핵심적인 역할을 한다. 마치 여러 재료를 조화롭게 혼합하여 맛있는 음식을 만드는 요리사와 같이, 역전파는 다양한 데이터의 특징을 효과적으로 융합하여 정확한 예측 결과를 만드는 데 기여한다.

## ② 쉽게 설명하면?

멀티모달 환경에서 역전파는 여러 형태의 데이터를 함께 학습하는 과정에서 오차를 최소화하는 역할을 한다. 이는 마치 여러 악기가 함께 연주할 때, 지휘자가 각 악기의 소리를 조율하여 아름다운 하모니를 만드는 것과 유사하다. 여기서 중요한 개념이 의미 공간(Semantic Space)이다. 의미 공간(Semantic Space)은 다양한 모달리티의 데이터를 공통된 의미적 표현으로 매핑하는 추상적인 공간이다. 역전파는 각 모달리티에서 추출된 특징들을 이 의미 공간(Semantic Space) 안에서 최적화하고, 이를 바탕으로 각 모달리티의 특징 추출기와 융합 모델의 파라미터를 조정한다. 이 과정을 통해 멀티모달 AI는 텍스트, 이미지, 오디오 등 다양한 데이터의 상호 관계를 이해하고, 더 정확한 예측을 할 수 있게 된다.

## ③ 예시와 함께 좀 더 자세히 알아볼까?

멀티모달 AI에서 역전파와 의미 공간(Semantic Space)의 역할을 이해할 수 있는 예시로 영화 장르 분류 모델을 들 수 있습니다. 이 모델은 영화의 포스터(이미지), 줄거리(텍스트), 예고편(비디오) 등 다양한 데이터를 입력받아 영화의 장르를 예측합니다.

의미 공간(Semantic Space)은 이러한 다양한 형태의 데이터를 공통된 의미 공간에 표현합니다. 예를 들어, '밝은 색감의 포스터', '로맨틱한 대사', '부드러운 배경음악' 등이 이 공간에서 '로맨스' 장르와 가깝게 위치하게 됩니다.

처음에는 모델이 로맨스 영화를 코미디로 잘못 분류할 수 있습니다. 이때 역전파가 작동하여 오차를 의미 공간(Semantic Space)과 각 데이터 처리 부분으로 전달합니다. 이를 통해 '로맨스' 특징들의 중요도를 높이고, 각 데이터 형태에서 이러한 특징들을 더 잘 포착하도록 모델을 조정합니다.

이렇게 역전파를 통해 의미 공간(Semantic Space)에서의 표현과 각 데이터 형태의 특성을 학습하고 융합하는 과정을 반복하면, 모델은 점점 더 정확하게 영화 장르를 예측할 수 있게 됩니다.

## ④ 왜 중요한가?

역전파는 멀티모달 AI가 복잡한 실제 세계의 데이터를 이해하고 활용하는 데 필수적인 기술이다. 우리 주변의 정보들은 텍스트, 이미지, 소리 등 다양한 형태로 존재하며, 이들 간의 상호 관계를 이해하는 것이 중요하다. 역전파는 이러한 멀티모달 데이터를 통합적으로 학습하고, 각 데이터 형태 간의 연관성을 파악하는 데 핵심적인 역할을 한다. 이는 이미지 캡셔닝, 비디오 요약, 음성 합성 등 다양한 멀티모달 AI 기술의 발전을 가능하게 한다. 결국 역전파는 인공지능이 인간의 다감각적 인지 능력에 더 가깝게 다가가는 데 중요한 발판이 된다.

⑤ 어디에 활용되는가?

역전파를 활용한 멀티모달 AI 기술은 우리 삶의 다양한 영역에서 활용된다. 예를 들어, 스마트폰의 AI 카메라는 이미지와 텍스트 정보를 함께 분석하여 최적의 촬영 설정을 제안한다. 또한, 음성 비서는 사용자의 음성 명령과 주변 상황(이미지, 위치 등)을 종합적으로 파악하여 더 정확한 응답을 제공한다. 자율주행 자동차는 카메라, 라이다, GPS 등 다양한 센서 정보를 융합하여 안전하게 운행한다. 소셜 미디어 플랫폼은 사용자가 업로드한 텍스트, 이미지, 동영상 등을 분석하여 개인화된 콘텐츠를 추천한다. 이처럼 역전파는 다양한 멀티모달 AI 애플리케이션의 핵심 기술로 자리잡고 있으며, 우리의 일상을 더욱 지능적으로 변화시키고 있다.