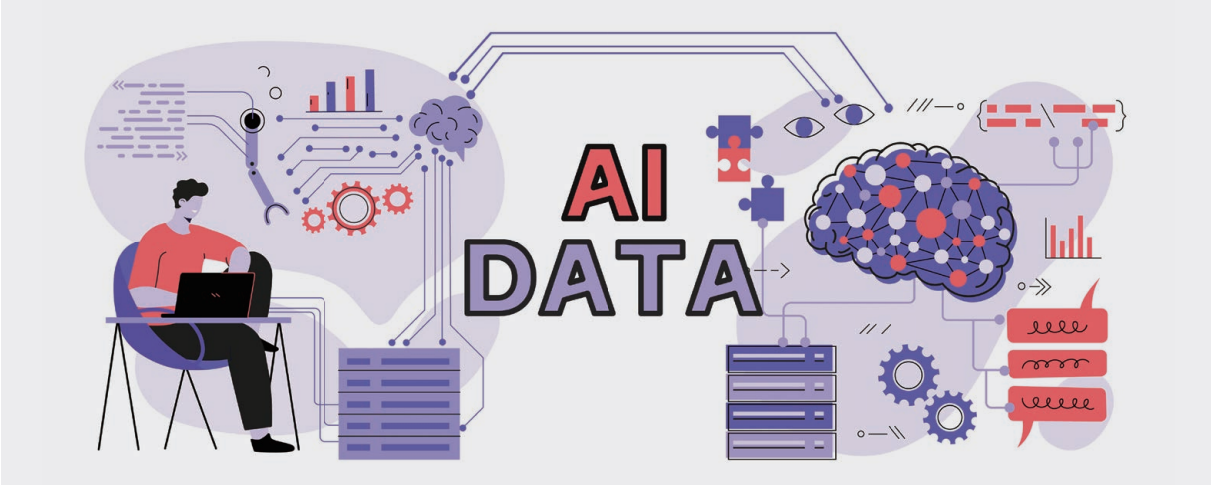
3화

AI 프로젝트의 첫 단추: '쓸모있는 데이터'를 확보하고 가공하는 기술

데이터 연금술: 방송국의 낡은 창고에서 황금을 캐는 법

#학습데이터 #데이터 확보 및 가공 실전 기술 #데이터 테이블링

글. 강자원 컴퓨터시스템응용기술사, KBS MNC(Media Network Center)팀



Part 1 서론: 왜 다시 '데이터'인가?

지난 2회차에서 우리는 AI가 가진 세 가지 강력한 엔진, 즉 '머신러닝', '컴퓨터 비전', '생성형 AI'가 어떻게 방송의 미래를 바꾸는지 엿보았다. 스포츠 중계를 실시간으로 분석하고, 수십 가지 버전의 예고편을 순식간에 만들어내는 AI의 능력은 분명 경이롭다. 하지만, 이 모든 것을 가능하게 하는 심장을 잠시 잊어서는 안 된다. 바로 '데이터'라는 이름의 연료다. 최고급 슈퍼카도 연료가 없으면 움직일 수 없는 고철 덩어리에 불과하듯, 아무리 뛰어난 AI 엔진도 양질의 데이터 없이는 한 발짝도 나아갈 수 없다.

AI 업계에는 'Garbage In, Garbage Out(쓰레기를 넣으면 쓰레기가 나온다)'이라는 유명한 격언이 있다. 이는 AI 프로젝트의 성패가 90% 이상 알고리즘이 아닌 데이터의 품질에서 결정된다는 냉엄한 현실을 보여준다. 이 지점에서 우리는 스스로에게 질문을 던져야 한다. 우리 방송국 데이터실에 수십 년간 쌓여있는 방대한 아카이브는 과연 AI 시대의 '보물창고'일까, 아니면 그저 라벨도 붙어있지 않은 비디오테이프가 가득한 '오래된 창고'일까? 영상 데이터는 포맷도, 해상도도, 코덱도 제각각이고, 음성 데이터에는 온갖 잡음이 섞여 있다. 메타데이터는 부실하거나 아예 존재하지 않는 경우도 허다하다. 이것이 우리가 마주한 데이터의 현실이다.

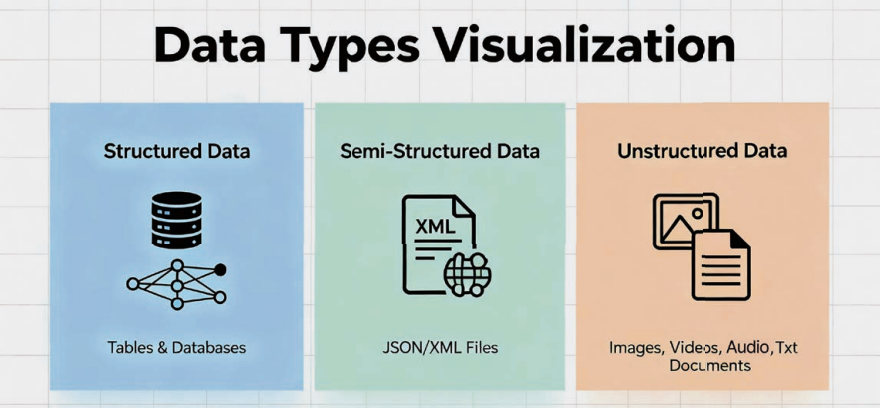
따라서 이번 3회차의 목표는 명확하다. 이 혼돈처럼 보이는 데이터의 산속에서 '쓸모있는 자원'을 찾아내고, AI가 이해할 수 있는 형태로 다듬고 정제하는 기술을 배우는 것이다. 이 글은 여러분을 데이터 과학자로 만들려는 것이 아니다. 방송 현장의 데이터를 AI 시대의 석유로 바꾸는 '데이터 연금술사(Data Alchemist)'가 되기 위한 실용적인 첫걸음을 안내하고자 한다. 엔진의 원리를 이해했다면, 이제 연료를 다룰 시간이다.

Part 2 AI의 연료, '학습 데이터'란 무엇인가?

데이터 연금술사가 되기 위한 첫 번째 임무는 우리가 다루어야 할 재료, 즉 '데이터'의 종류와 특징을 명확히 이해하는 것이다. AI에 어떤 종류의 연료를, 어떻게 가공해서 공급해야 할까?

데이터의 종류 : 방송국의 데이터 창고 들여다보기

방송국에 존재하는 데이터는 크게 세 가지 형태로 나눌 수 있다.



데이터의 종류	정의	예시
정형 데이터 (Structured Data)	엑셀 표처럼 행과 열로 명확하게 구조화된 데이터	시청률 표, 편성표(EPG), 장비별 에러 로그, QC 리포트
비정형 데이터 (Unstructured Data)	정해진 구조 없이 내용만 존재하는 데이터. 가장 다루기 어렵지만 가치가 높음.	방송 영상/음성 원본, 뉴스 기사 텍스트, 드라마 대본, SNS 댓글
반정형 데이터 (Semi-structured Data)	데이터 내부에 XML 태그나 JSON 형식처럼 일정한 구조 정보를 포함하는 데이터	MAM 메타데이터, 자막 파일(SRT, XML), 웹 데이터

표 1. 방송 데이터의 종류

엔지니어의 과제와 기회

위 표에서 알 수 있듯, 방송국 자산의 90% 이상을 차지하는 핵심 데이터(영상, 음성)는 AI가 가장 다루기 힘들어하는 '비정형 데이터'다. 컴퓨터는 행과 열로 정리된 정형 데이터는 쉽게 이해하지만, 아무런 구조가 없는 영상이나 음성 파일은 그저 거대한 0과 1의 덩어리로 인식할 뿐이다.

바로 이 지점이 엔지니어에게 도전이자 기회다. 데이터 과학자는 이 영상이 기술적으로 어떤 코덱과 포맷을 가졌는지는 알지만, 그 안에 담긴 'NLE 편집 장비의 특정 버전에서만 발생하는 미세한 프레임 깨짐'이나 '특정 마이크 모델에서 유독 심하게 발생하는 고주파 노이즈'와 같은 맥락(Context)은 알지 못한다. 수십 년간 현장에서 쌓아온 엔지니어의 도메인 지식은 바로 이 비정형 데이터에 의미와 구조를 부여하고, AI가 이해할 수 있는 형태로 바꾸는 데 결정적인 역할을 한다.

학습 데이터의 핵심 : '레이블링(Labeling)'의 모든 것

AI가 비정형 데이터를 이해하게 하려면, '정답'을 알려주는 과정이 필요하다. 이를 레이블링(또는 어노테이션 Annotation)이라고 한다. 이는 마치 신입사원에게 과거 업무 자료 수만 건을 그냥 던져주는 것이 아니라, 중요한 부분에 빨간펜으로 표시해주고 "이건 성공 사례, 저건 실패 사례"라고 하나하나 가르쳐주는 과정과 같다.

방송 현장에서 레이블링은 AI의 목적에 따라 다양한 방식으로 이루어진다.

- 분류(Classification) : 데이터가 어떤 카테고리에 속하는지 정답을 알려주는 가장 기본적인 레이블링이다.

> **예시** : 뉴스 영상을 보고 <정치>, <경제>, <스포츠> 중 하나로 태그를 붙여주거나, 시청자 의견을 <긍정>, <부정>, <중립>으로 분류하는 작업.

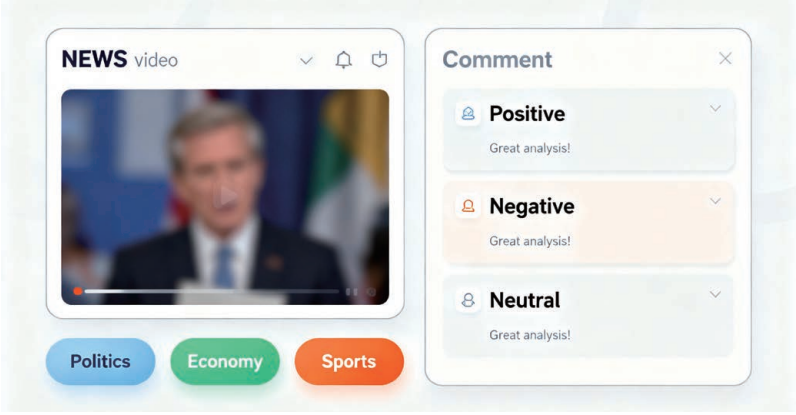


그림 1. 뉴스 태깅 작업 예시

- 객체 탐지(Object Detection) : 영상 속 특정 객체의 위치를 네모난 박스(Bounding Box)로 지정하고, 그것이 무엇인지 이름을 붙여주는 작업이다.

> **예시** : 드라마 영상에서 PPL 상품이 등장하는 모든 장면에 박스를 치고 'A사 커피'라고 레이블링하거나, 축구 경기에서 '공'과 '선수'를 각각 다른 박스로 지정하는 작업.

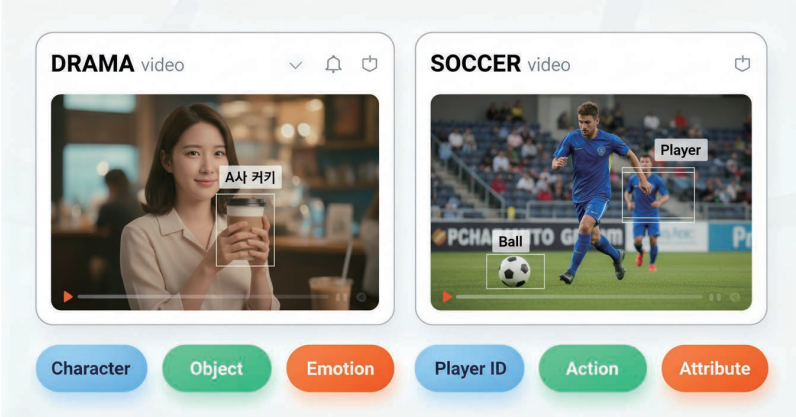


그림 2. 객체 탐지 예시

• 세분화(Segmentation) : 객체의 경계를 따라 픽셀 단위로 정밀하게 영역을 구분하는, 가장 정교한 레이블링이다.

> **예시** : 날씨 방송 영상에서 기상 캐스터의 외곽선을 정밀하게 그려 배경과 분리하거나, 의료 영상에서 암세포의 영역만 정확히 도려내는 작업.

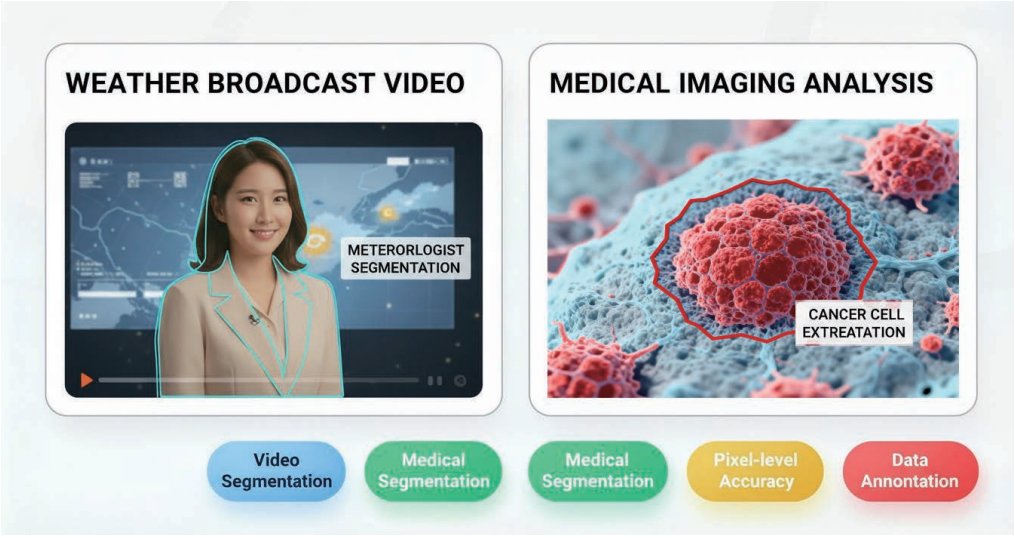


그림 3. 세분화 작업 예시

좋은 학습 데이터의 조건 : AI의 편식을 막는 법

'데이터 연금술사'의 진정한 실력은 단순히 데이터를 많이 모으는 것이 아니라, AI가 편식하지 않도록 '좋은 밥상'을 차려주는 데 있다. 좋은 학습 데이터는 다음 세 가지 조건을 만족해야 한다.

• 품질(Quality) : 정확하고 일관된 레이블

가장 중요한 원칙이다. 한 사람은 '마이크', 다른 사람은 '음향 장비'라고 레이블링한다면 AI는 혼란에 빠진다. 모든 데이터는 명확하고 일관된 기준에 따라 정확하게 레이블링되어야 한다. 소수의 고품질 데이터가 다수의 저품질 데이터보다 훨씬 낫다.

• 양(Quantity) : 충분하지만 무조건 많다고 좋은 것은 아님

일반적으로 데이터가 많을수록 AI의 성능은 좋아진다. 하지만 100만 장의 똑같은 고양이 사진보다, 다양한 품종의 고양이 사진 1만 장이 훨씬 더 효과적인 학습 데이터다. 중요한 것은 데이터의 절대적인 양이 아니라, 우리가 해결하려는 문제의 복잡성을 충분히 '커버'할 수 있는 양이다.

• 다양성(Diversity) : 편향되지 않은 균형 잡힌 데이터

AI의 편식을 막는 가장 중요한 요소다. 예를 들어, 뉴스 앵커를 인식하는 AI를 40대 남성 앵커의 데이터로만 학습시킨다면, AI는 새로운 20대 여성 앵커나 안경을 쓴 객원 앵커를 제대로 인식하지 못할 것이다. 다양한 시간대(주간/야간), 다양한 구도, 다양한 조명 환경에서 촬영된 데이터를 골고루 학습시켜야 어떤 상황에서도 안정적으로 작동하는 강건한(Robust) AI를 만들 수 있다.

Part 3 데이터 확보 및 가공 실전 기술

좋은 학습 데이터의 조건을 이해했다면, 이제 데이터 연금술의 실전 단계로 나아갈 시간이다. 이 단계는 크게 세 가지로 나뉜다. 재료를 모으는 '수집', 재료를 다듬는 '전처리', 그리고 이 모든 과정을 자동화하는 '파이프라인 구축'이다.

데이터 수집(Collection) : 어디서 재료를 구할 것인가?

AI를 학습시킬 '재료'는 어디서 찾아야 할까? 우리의 주변, 그리고 조금만 눈을 돌리면 접근할 수 있는 곳에 풍부한 데이터 소스가 존재한다.

내부 데이터(Internal Data) : 가장 귀중한 자산

방송국 내부에 이미 축적된 데이터는 우리의 가장 강력한 무기다. 외부에서는 절대 구할 수 없는, 우리 방송 환경만의 고유한 특성을 담고 있기 때문이다.

- 미디어 아카이브/MAM : 수십 년간 쌓아온 방송 영상과 음원은 그 자체로 보물 창고다. 과거의 특정 아나운서 목소리, 특정 시대의 영상 화질 특성 등은 우리 방송사만의 AI 모델을 만드는 데 핵심적인 재료가 된다.
- 장비 로그 서버 : 송출 서버, 라우터, 스토리지 등 모든 네트워크 장비가 뱉어내는 로그는 예측 유지보수 모델을 위한 최고의 학습 데이터다. 이 로그 속에서 장애 발생 전의 미세한 패턴을 찾아낼 수 있다.

오픈 데이터(Open Data) : 부족한 재료를 보충하는 지혜

내부 데이터만으로 부족하거나, 모델의 기초 성능을 검증하고 싶을 때 공개된 데이터를 활용할 수 있다.

- 공공 데이터 포털 및 AI Hub : 정부에서 공개하는 방송 프로그램 정보, 시청률 데이터나 한국어 음성/텍스트 데이터셋 등은 기초 모델을 만들 때 유용하다.
- 학술용 데이터셋(ImageNet, COCO 등) : 수백만 장의 이미지가 정교하게 레이블링된 데이터셋으로, AI 모델의 성능을 연구하고 벤치마킹하는 기준으로 널리 사용된다.

합성 데이터(Synthetic Data) : 희귀 재료를 만들어내는 최신 연금술

'방송사고' 영상처럼 현실에서 구하기 매우 어렵거나 존재하지 않는 데이터를 가상으로 생성하는 기술이다. AI 시대에 그 중요성이 점점 커지고 있다.

> **예시 :** 실제 방송사고가 날 때까지 기다릴 수는 없다. 대신, 3D 그래픽 툴이나 시뮬레이터를 이용해 다양한 종류의 블랙 스크린, 프레임 깨짐, 오디오 왜곡 등 가상의 방송사고 영상을 수만 개 생성하여 AI를 학습시킬 수 있다. 이는 AI의 강건함을 키우는 핵심 기술이다.

데이터 전처리(Preprocessing) : 지저분한 원석을 보석으로

데이터 과학자들은 프로젝트 시간의 80%를 데이터 전처리에 쓴다고 말한다. AI 프로젝트의 성패가 사실상 이 단계에서 결정된다고 해도 과언이 아니다. 데이터 연금술의 가장 중요하고 고된 과정이다.

데이터 정제(Cleaning) : 불순물 제거하기

가장 먼저 할 일은 데이터의 오류나 노이즈 같은 불순물을 제거하는 것이다.

- **결측값 처리** : 장비 로그 데이터가 일부 유실된 경우, 통계적으로 채워 넣거나 해당 로그를 제거한다.
- **오류 수정** : 과거에 사람이 수기로 입력한 메타데이터의 오타나 잘못된 정보를 바로잡는다.
- **노이즈 제거** : 오래된 영상의 화면 노이즈를 제거하거나, 야외에서 녹음된 음성의 바람 소리를 줄이는 작업 등이 모두 AI 성능 향상을 위한 데이터 정제에 해당한다.

데이터 정규화(Normalization) : 기준 통일하기

제각각인 데이터의 기준을 하나로 통일하여 AI가 패턴을 더 쉽게 학습하도록 돕는 과정이다.

- **오디오 레벨 정규화** : 방송 채널마다, 프로그램마다 다른 오디오 레벨(-24LUFS, -18LUFS 등)을 AI 분석을 위해 특정 표준(-23LUFS 등)으로 일괄 변환한다.
- **영상 해상도/프레임률 통일** : 4K, HD, SD 등 다양한 해상도의 영상을 1080p로 통일하거나, 각기 다른 프레임률(29.97i, 59.94p 등)을 하나의 표준으로 변환하여 AI 모델의 입력 조건을 맞춰준다.

데이터 증강(Augmentation) : 재료 뻥튀기

보유한 데이터의 양이 적을 때, 원본 데이터를 약간씩 변형하여 학습 데이터의 양을 인위적으로 늘리는 기술이다.

예시 : 뉴스 앵커를 인식하는 AI를 학습시키려는데 앵커 사진이 100장밖에 없다고 가정하자. 이 100장의 사진을 각각 좌우 반전, 약간의 회전, 밝기 조절, 일부 확대(Zoom-in) 등을 적용하면, 순식간에 수천 장의 새로운 학습 데이터를 만들어 낼 수 있다.

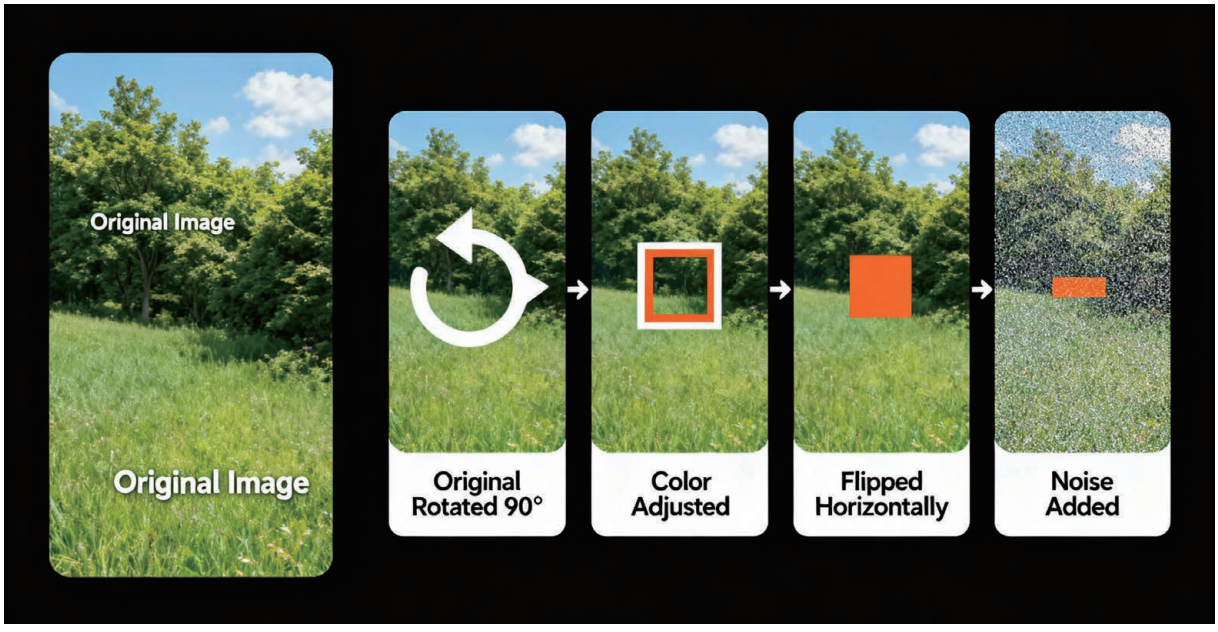


그림 4. 학습 데이터 생성 예시

데이터 파이프라인 구축 : 자동화된 데이터 공장 만들기

매번 새로운 데이터가 들어올 때마다 이와 같은 수집, 전처리, 레이블링 과정을 수동으로 반복하는 것은 비효율적이고 실수를 유발하기 쉽다. 데이터 파이프라인은 이 모든 과정을 자동화하여, 데이터가 물처럼 자연스럽게 흘러가며 처리되도록 만드는 '자동화 공장'이다.

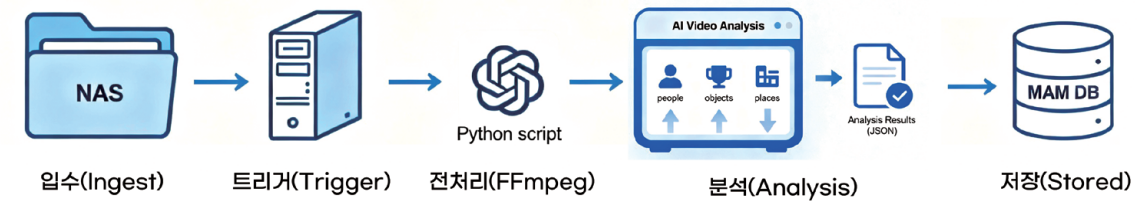


그림 5. 데이터 파이프라인

엔지니어의 관점에서 간단한 'AI 영상 분석 파이프라인'을 설계하면 다음과 같다.

- 입수(Ingest) : 제작팀이 편집을 마친 영상 파일을 방송사의 NAS 스토리지 내 특정 폴더(예: /nas/for_ai_analysis)로 입수시킨다.
- 트리거(Trigger) : 리눅스 서버에서 실행되는 간단한 Python 스크립트가 해당 폴더를 1분마다 감시(Watch)하다가, 새로운 파일이 들어온 것을 감지하면(Trigger) 다음 프로세스를 자동으로 실행한다.
- 전처리(Preprocess) : 스크립트는 FFmpeg 라이브러리를 호출하여 입수된 영상 파일을 AI 분석에 최적화된 표준 포맷(예: 1080p, 29.97fps, H.264, AAC)으로 자동 변환(Transcoding)한다.
- 분석 및 저장(Analyze & Store) : 표준화된 영상은 API를 통해 AI 분석 서버(클라우드 또는 온프레미스)로 전송된다. AI 서버는 영상 속 인물, 객체, 장소 등을 분석하여 결과값을 JSON 형태로 반환하고, 시스템은 이 메타데이터를 파싱하여 MAM 데이터베이스에 저장한다.

이러한 파이프라인이 구축되면, 엔지니어는 더 이상 반복 작업에 시간을 뺏기지 않고 전체 시스템의 안정성과 성능을 관리하는 더 중요한 역할에 집중할 수 있다.

Part 4 결론: 데이터, 엔지니어의 새로운 영토

이번 3회차에서 우리는 AI 프로젝트의 성패를 좌우하는 가장 근본적인 요소인 '데이터'의 세계를 탐험했다. 정형, 비정형, 반정형으로 나뉘는 데이터의 종류부터, AI에 정답을 가르쳐주는 '레이블링'의 핵심 과정, 그리고 AI의 편식을 막는 좋은 데이터의 세 가지 조건(품질, 양, 다양성)까지 살펴보았다. 나아가 데이터를 수집하고, 보석으로 다듬는 '전처리' 과정을 거쳐, 이 모든 것을 자동화하는 '데이터 파이프라인'의 개념까지 AI의 연료를 다루는 실전 기술들을 확인했다.

그렇다면 이 모든 과정이 방송 엔지니어에게 의미하는 바는 무엇인가? 단순히 배워야 할 기술이 늘어난 것일까? 그렇지 않다. 이것은 우리 역할의 근본적인 진화, 새로운 영토의 발견을 의미한다.

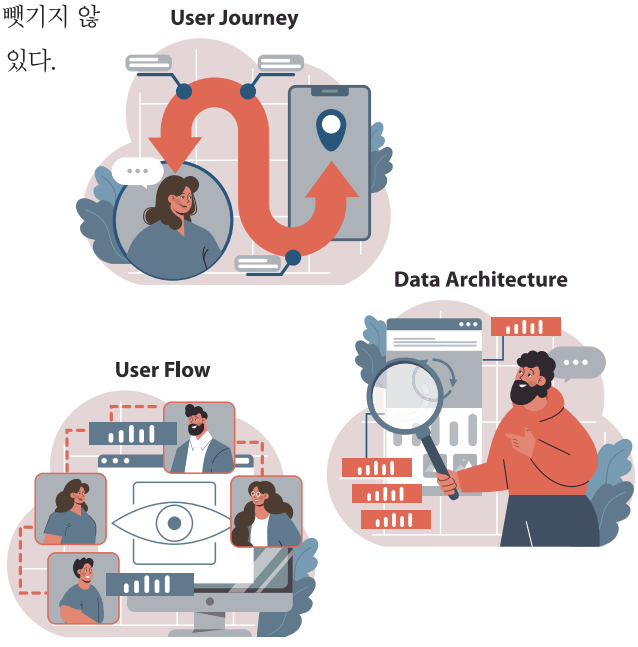


그림 6. 데이터 아키텍트로서 방송 엔지니어의 역할

과거의 방송 엔지니어가 SDI 케이블과 라우터를 통해 '신호의 흐름(Signal Flow)'을 설계하고 관리하는 전문가였다면, 미래의 방송 엔지니어는 데이터 파이프라인을 통해 '데이터의 흐름(Data Flow)'을 설계하고 관리하는 '데이터 아키텍트(Data Architect)'가 되어야 한다. 신호의 무결성을 지키는 것만큼이나 데이터의 품질과 다양성을 지키는 것이 중요해졌다. 한정된 전송 대역폭을 효율적으로 사용하는 것만큼이나, 방대한 데이터 속에서 AI가 학습할 핵심 특징을 추출하고 정제하는 능력이 핵심 역량이 된 것이다. 이는 방송기술에 대한 깊은 이해와 데이터에 대한 새로운 관점이 만났을 때만이 가능한, 우리 엔지니어들에게 주어진 새로운 기회다.
물론, 이 모든 것을 처음부터 직접 만들어야 한다고 생각하면 지레 겁을 먹을 수 있다. 하지만 다행히 우리는 모든 바퀴를 다시 발명할 필요가 없다.

현실적인 조언 : 모든 것을 직접 만들 필요는 없다

'데이터 아키텍트'의 역할은 모든 코드를 직접 짜는 개발자가 아니다. 문제를 해결하기 위해 세상에 존재하는 수많은 도구를 현명하게 선택하고, 이를 우리 시스템에 맞게 통합하는 '솔루션 통합 전문가'에 가깝다.

- 데이터 레이블링 : 처음부터 복잡한 레이블링 툴을 개발할 필요는 없다. Supervisely, Labelbox와 같은 강력한 상용 툴이나 CVAT(Computer Vision Annotation Tool) 같은 오픈소스 툴은 복잡한 영상 어노테이션 작업을 훨씬 효율적으로 만들어준다.

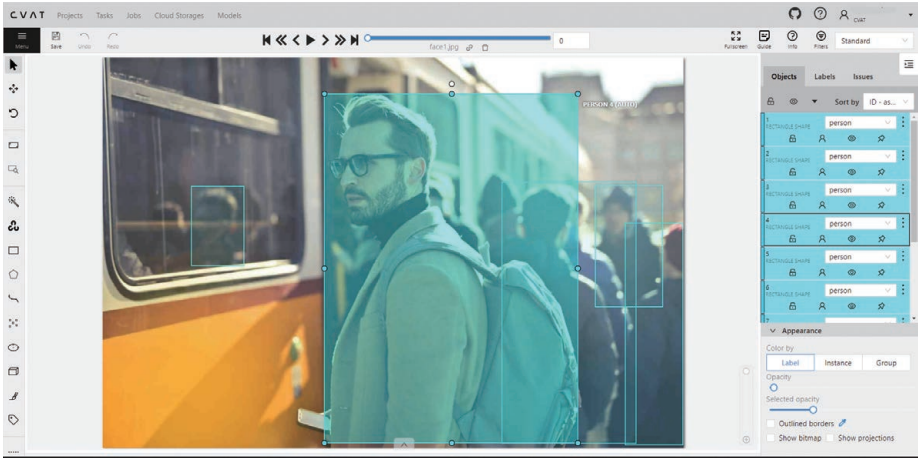


그림 7. CVAT를 사용한 데이터 레이블링

- 데이터 전처리 : 데이터 정제나 증강 같은 작업들은 OpenCV(영상처리), Pandas(데이터분석), Scikit-learn(머신러닝)과 같은 강력한 Python 오픈소스 라이브러리를 활용하면 많은 부분을 자동화할 수 있다.

엔지니어의 진정한 가치는 코드를 한 줄 더 짜는 것이 아니라, 어떤 툴과 라이브러리가 우리 방송국의 문제 해결에 가장 적합한지 판단하고, 이를 현명하게 조합하여 효율적인 데이터 파이프라인을 설계하는 능력에 있다. 데이터 파이프라인을 구축하고 현명한 도구를 선택하는 엔지니어의 손에서 비로소 방송국의 낡은 창고는 진정한 보물 창고로 거듭날 수 있다.

이제 최고의 연료를 준비하는 법을 배웠다. 다음 4회차에서는 이 연료를 활용해 'AI의 눈, 영상 분석과 제작 자동화'라는 주제로, '컴퓨터 비전' 기술이 실제 방송 제작 워크플로우를 어떻게 혁신하는지 구체적인 사례와 함께 깊이 파고들어 보겠다. 데이터라는 원석을 보석으로 바꾸는 법을 익혔으니, 이제 그 보석으로 무엇을 만들 수 있는지 확인할 시간이다.