

# 데이터라벨러

장재범 교수

## 1강. 빅데이터의 이해

### 1. 4차 산업과 인공지능

1차 산업혁명(18~19세기 중: 기계의 등장)

2차 산업 혁명(19~20세기 초: 전기 에너지, 대량생산 혁명)

3차 산업 혁명(20세기 후반: 컴퓨터, 인터넷, 지식 정보 혁명)

4차 산업 혁명(21세기 초반~현재: 인공지능이 빅데이터 초연결 지능화를 통해 4차 산업혁명으로 진입)

4차 산업혁명 핵심 기술: 인공지능, 빅데이터 (사물인터넷, 3D 프린팅, 로봇, 공유경제, 드론)

### 2. 데이터의 정의와 특성

대량의 정형(수치화), 비정형 데이터(텍스트, 영상, 음성 등)로부터 가치를 추출하는 기술(데이터 처리 기술) 대량의 모든 데이터로 컴퓨터, 인터넷 등 디지털 환경에서 발생하는 데이터. 기존의 정형 데이터에서 비정형 데이터로 많은 데이터 발생. 2015년 기준 1분간 데이터 발생량 : 구글의 2백만 건 데이터 검색, 유튜브 72시간 비디오 생성, 트위터 72만 건.

시대에 따른 데이터의 변화 : 컴퓨터의 발전에 의해 데이터의 양과 형태가 변화하고 있다.

1970~1980 메인프레임 컴퓨터(정형 데이터)

1980~2000 개인 PC

2000~2010 인터넷 모바일 소셜미디어 스마트폰의 보급으로 빅데이터 시대로 접어들었다.

2020~현재 IT everywhere (인공지능, 증강현실, 메타버스) 시대 시작.

1) 대량의 모든 데이터: 컴퓨터 인터넷 등 디지털 환경에서 발생하는 데이터

2) 데이터의 가치와 결과 분석 기술: 데이터 관리에서 데이터를 분석해서 가치 창출

3) 빅데이터 플랫폼 등장 : 데이터 관리하는 하드웨어, 소프트웨어, 어플 등장

4) 대규모 데이터 관리 기술: 데이터 저장, 관리 분석하는 하드웨어, 소프트웨어, 활용 기술

5) 데이터라는 용어는 1646년 영국 문헌에 처음 등장 라틴어인 dare(주다)의 과거분사형으로 주어진 것이란 의미로 사용

6) 1940년대 이후 컴퓨터 시대 시작과 함께 자연과학뿐만 아니라 다양한 사회과학이 진일보하며 데이터의 의미는 기술적이고 사실적인 의미로 변화

7) 데이터는 추론과 추정의 근거를 이루는 사실

8) 데이터는 다른 객체와의 상호 관계 속에서 가치를 갖는 것으로 설명

### 3. 데이터와 정보의 관계

데이터: 존재 형식을 불문하고 타 데이터와의 상관관계가 없는 가공하기 전의 순수한 수치나 기호를 의미 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실

정보: 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 그 의미를 부여한 데이터 데이터의 가공 처리와 데이터 간 연관관계 속에서 의미가 도출된 것

지식: 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물 데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 고유의 지식으로 내재화된 것

지혜: 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어 지식의 축적과 아이디어가 결합된 창의적인 산물

### 4. 빅데이터의 정의

첫째, 좁은 정의 3V로 요약되는 데이터 자체의 특성 변화에 초점

둘째, 중간 범위 데이터 자체뿐만 아니라 처리 분석 기술적 변화까지 포함

셋째, 넓은 범위 인재, 조직 변화까지 포함한 넓은 관점에서의 정의

## 5. 데이터의 유형

구분

형태

정성적 데이터

qualitative data

언어, 문자 등

예

매출이 증가함 등

정량적 데이터

quantitative data

## 6. 정형 데이터와 비정형 데이터

수치, 도형, 기호 등

나이, 몸무게, 주가 등

정성적 데이터 정량적 데이터

비정형 데이터

주관적 내용

통계 분석이 어려움

정형 데이터

객관적 내용

통계분석이 용이함

정형 데이터: 구조화된 데이터, 고정된 필드에 저장된 데이터(데이터베이스, 엑셀, CSV)

반 정형 데이터: 고정된 필드는 아니지만 스키마를 포함, 연산 불가(XML, HTML, JSON 등)

비정형 데이터: 고정되지 않은 데이터, 연산 불가, 형태 없음 (SNS 데이터, 사진, 영상, 음성)

## 7. 가트너 그룹(Gartner Group)의 더그 래니(Doug Laney)의 3V

3V

4V

양(Volume)

↓

데이터의 규모

측면

다양성(Variety)

↓

데이터의 유형과

소스 측면

속도(Velocity)

↓

데이터의 수집과

처리 측면

+

생성 데이터

비정형 데이터

정형, 비정형

데이터

(영상, 사진)

원하는 데이터의

추출 및 분석

속도

정확성

(Veracity)

가치

(Valuation)

시각화

(Visualization)

초기 빅데이터 특징 : 3V: 규모(Volume), 속도(Velocity), 다양성(Variety)

> 4V: 정확성 추가 (Veracity)

> 5V: 가치(Value) 추가

Volume(규모): 데이터 크기, 최근 그 양이 증가

Velocity(속도) : 최근은 실시간으로 처리 필요

Variety(다양성) : 정형(Structured), 비정형(Unstructured), 반 정형(Semi-structured) 형식  
관계없이 처리.

Veracity (정확성) : 신뢰할 수 있는 데이터인지 구분

Value(가치) : 저장할 가치가 있는지 판단→ 데이터 가공&분석으로 의미 있는 결과 도출

## 8. 빅데이터의 가치

분석 기술 발전

현재는 가치가 없는 데이터일지라도 추후에 새로운 분석 기법이 등장한다면 거대한 가치를 지  
닌 데이터가 될 수도 있음

데이터 활용 방식

데이터 활용 방식에서는 재사용이나 재조합 다목적용 데이터 개발 등이 일반화되면서 특정 데  
이터를 언제 어디서 누가 활용할지 알 수 없게 되어서 가치를 산정하는 것도 어려워짐

새로운 가치 창출

빅데이터 시대에 데이터가 기존에 없던 가치를 창출함에 따라 그 가치를 측정하기가 어려워짐

## 9. 빅데이터가 미치는 영향

기업 — 혁신 경쟁력 제고 생산성 향상

빅데이터를 활용해 소비자의 행동을 분석하고 시장 변동을 예측해 비즈니스 모델을 혁신하거  
나 신사업을 발굴한다

정부 — 환경 탐색 상황 분석 미래 대응

기상 인구이동 각종 통계법 제 데이터 등을 수집해 사회 변화를 추정하여 관련 정보를 추출한  
다

개인 — 목적에 따른 활용

빅데이터를 서비스하는 기업의 출현이 늘어나면서 데이터 분석 비용이 지속적으로 하락하여  
정치인이나 대중 가수 등과 같은 개인도 인지도 향상에 빅데이터를 활용한다

## 1. 데이터 라벨링

기계 학습 알고리즘에 데이터를 더 유용하고 유익하게 만들기 위해 이미지, 텍스트 또는 오디오와 같은 데이터 세트에 하나 이상의 설명 태그 또는 라벨을 수동, 자동으로 할당하는 프로세스 이러한 태그 또는 레이블은 데이터를 분류 또는 범주화하고, 패턴 또는 경향을 식별하거나, 추가 컨텍스트 또는 의미를 제공하는 데 사용

데이터 레이블 지정은 기계 학습 알고리즘이 정확하고 관련성 있는 데이터와 함께 작동하는지 확인하는 데 도움이 되므로 기계 학습 파이프라인에서 중요한 단계

적절한 레이블 지정이 없으면 머신 러닝 모델이 데이터의 패턴을 정확하게 인식하거나 분류하지 못하여 부정확한 예측이나 결정을 내릴 수 있음

데이터 라벨링은 사람 어노테이터가 수동으로 수행하거나 알고리즘 또는 이 둘의 조합을 사용하여 자동으로 수행

데이터 레이블 지정 방법의 선택은 데이터 세트의 크기 및 복잡성, 필요한 정확도 수준 및 사용 가능한 리소스와 같은 다양한 요인에 따라 달라짐

### 2. 데이터 라벨링에게 요구되는 사항

#### 1) 첫째도 가이드, 두 번째도 가이드 준수

데이터 레이블 지정에는 데이터 레이블이 정확하게 지정되도록 세부 사항에 대한 높은 수준의 주의가 필요 이 기술은 라벨링 프로세스에서 오류와 불일치를 방지하는 데 필수

기계 학습 개념에 대한 지식: 분류, 회귀, 클러스터링 및 자연어 처리와 같은 기계 학습 개념에 익숙하면 레이블 지정 요구 사항 및 예상 결과를 이해하는 데 도움

#### 2) 우수한 의사소통 기술

데이터 레이블 지정에는 종종 팀 환경에서 작업이 포함되며 다른 팀 구성원과의 효과적인 협업을 위해서는 우수한 의사소통 기술이 필수

#### 3) 관련 소프트웨어 도구에 대한 숙련도

프로젝트 요구 사항에 따라 Python, R, SQL 또는 라벨링 소프트웨어와 같은 관련 소프트웨어 도구에 대한 숙련도가 도움

#### 4) 도메인 지식

산업 또는 프로젝트 분야의 도메인 지식은 데이터의 맥락과 라벨링 요구 사항을 이해하는 데 도움

#### 5) 교육 및 훈련

컴퓨터 과학, 데이터 과학 또는 관련 분야의 학위가 도움이 될 수 있지만 항상 필요한 것은 아니고 데이터 라벨링, 기계 학습 및 관련 분야의 교육 프로그램 및 인증도 필요한 기술과 지식을 얻는 데 도움

전반적으로 데이터 라벨링에는 기술 및 소프트 스킬의 조합, 세부 사항에 대한 주의, 기계 학습 모델을 개선할 수 있는 고품질 라벨 데이터를 제공하기 위해 팀 환경에서 작업할 수 있는 능력이 필요

## 3. 데이터 수집과 가공

### 1) 데이터 수집

수집은 사진이나 동영상을 찍는 촬영이 있고, 음성이나 필요한 소리를 녹음하는 작업도 있으며, 특정 상황에 맞게 문장을 발화하는 글쓰기, 프로젝트와 연관된 설문 등이 있음

### 2) 데이터 가공

데이터 가공은 가장 쉽고 간단한 바운딩부터, 세그멘테이션, 키포인트, 폴리곤 등으로 세분화

#### (1) 바운딩

가이드에 맞게 사각형 박스를 만드는 가장 흔하고 간단한 작업

Bounding box(바운딩박스) : 학습을 통해 검출한 객체의 영역을 사각형으로 표시하는 것을 의미

#### (2) 세그멘테이션

세그멘테이션 작업은 객체의 모양에 따라 점을 연속적으로 찍어 많은 점이 모여 선으로 연결 되는 작업

바운딩과 비교를 하면 바운딩은 사각형의 네 모서리 부분만 점을 찍으면 되지만 세그멘테이션은 바운딩보다 훨씬 많은 점이 필요

이미지의 모든 픽셀에 클래스를 부여하는 라벨링 방법

물체, 배경 등을 의미적으로 분할하여 자율주행, 의료 영상 분석 기술 개발 등에 사용

(예시 : 자율항해 라벨링, 의학 데이터 라벨링)

#### (3) 키포인트

키포인트의 경우 사람의 움직임이나, 표정 등을 읽을 때 사용하는 라벨링 방법

보통은 각 관절이나 원하는 포인트에 순서대로 점을 찍어 작업하는 방식

특정 지점(특징점)을 라벨링 하는 방법

안면 인식을 통한 감정 분석과 같이 정밀하고 섬세한 작업을 요구하는 기술

객체의 중요 특징점을 지정하여 물체를 추적하고 인식할 수 있음

(예시 : 운동선수 상체 관절 라벨링, 립리딩 라벨링)

#### (4) Keypoint (키포인트)

: 탐지하려는 객체의 모양을 알고 싶을 때 데이터의 외곽선을 따جوم으로써 폴리곤과 포인트 정보를 만들어 특징을 갖게 만드는 것

Point (포인트) : 이미지에서 찾으려는 객체에 대해 점을 찍어 표기하는 방식

#### (5) 폴리곤

폴리곤은 바운딩과 세그멘테이션의 중간 정도의 난이도

바운딩은 사각형으로 네 개의 모서리가 정해져 있지만 폴리곤은 작업자의 작업에 따라 오각형, 육각형 등의 모양

다각형 모양으로 객체의 가시 영역 외곽선을 따라 점을 찍어 그리는 라벨링 방법

개체 이외의 포함된 빈 공간으로 인해 발생하는 오류에 대응할 수 있는 기능

사물의 테두리를 따라 그리는 것을 통해 여백 없이 정확히 물체만을 인식하기 위해 사용

(예시 : 세포 검사 라벨링, 생물 성장 라벨링)

Polygon (폴리곤) : 다각형. Object Detection(객체 검출) 하려는 객체를 다각형으로 표기

#### (6) 폴리 라인 (Polyline) :

여러 개의 점을 가진 선을 활용하여 특정 영역을 라벨링 하는 방법

인도와 차선, 경계 등의 선형 데이터들을 구분하고 인식시키기 위해서 사용

(예시 : 차선 라벨링, 경로 라벨링)

많은 점으로 선을 그어 표기. 도로 선같이 시작과 끝이 없는 선을 구분할 때 사용하기 좋은 라벨링. 주로 차량 ADAS에서 자율주행을 위해 사용.

#### (7) 분류/판별

분류/판별은 가이드에 따라 속성 혹은 카테고리별로 구분하여 전사하는 작업

예를 들면 사람으로 가득 찬 지하철 사진을 주고 남자와 여자 혹은 노인과 청년, 어린아이로

분류하라는 프로젝트가 있다면 분류/판별에 해당

(8) OCR (Optical Character Recognition, 광학문자인식)

사람이 쓰거나 기계로 인쇄한 문자의 영상/이미지를 기계가 읽을 수 있는 문자로 변환하는 방법 문자를 인식시키기 위해 사용하며 문자를 이용하는 다양한 기술 개발에 활용 가능

(예시 : 제품명 OCR 라벨링)

(9) 글쓰기

특별한 상황을 가이드에 안내해 주고 그 상황에 맞는 글쓰기를 하는 프로젝트

(10) Cuboid(큐보이드) : 정육면체. 기존에 Bounding Box가 2차원으로 데이터를 형성하는 것이었다면, 차원을 하나 더 높여서 3차원의 데이터로 더 많은 정보를 제공. 자율주행 시스템에서 차량의 앞뒤 좌우를 표기하는데 매우 유용함.

(11) 태깅 : 이미지나 파일에 이름을 붙이는 것

(12) 전사 : 이미지나 영상 속 문자를 텍스트로 옮겨 적는 작업

(13) 감정 분석: 이미지나 영상 속 사람의 표정을 보고 어떤 감정 상태인지 추론하는 라벨링 기법

(14) 얼굴 랜드마크 : 얼굴 주요 부위에 마우스를 클릭해 점을 찍어주는 라벨링 기법

(15) 특정 구간 추출 : 작업 대상이 사전에 제시된 기준에 해당하는 말이나 행동 등을 할 때 구간을 선택해 추출하는 라벨링 기법 3D 라벨링 기법: 육면체의 입체적인 박스를 생성하여 바운딩하는 것과 같이 2차원 라벨링의 한계를 넘어 3차원의 입체를 표현하게 하는 기법

(16) 복합 라벨링 : 여러 개의 라벨링 기법을 복합적으로 사용하는 라벨링 기법

(17) 문장 의미 비교 : 주어진 문장들의 의미가 같은 것인지 태깅 하는 라벨링 기법

(18) 감정 태깅 : 제시된 글을 읽었을 때 느껴지는 감정을 선택하는 라벨링 기법

(19) 키워드 찾기 : 대화 내용 속에서 핵심이 되는 키워드를 찾는 라벨링 기법

(20) 문장 요약: 글을 읽고 핵심이 되는 내용을 요약하는 라벨링 기법

(21) 화자 구분 : 제시된 음성을 모두 듣고 동일한 사람의 목소리인지를 판단해서 태깅 하는 라벨링 기법

(22) 음성 받아쓰기: 주어진 음성을 듣고 받아쓰는 라벨링 기법

(23) 일반 전사 : 말한 그대로를 문자화하여 전사하는 기법

(24) 이중 전사: 한글 맞춤법 표기에 따른 발음에 차이가 있는 경우, 발음 전사와 철자 전사를 병행하여 작성하는 방법

★데이터 라벨링 작업을 할 수 있는 데이터 - 이미지, 텍스트, 영상

4. 이미지 데이터 식별 방식

1) 분류 (Classification) :

데이터의 카테고리 분류

2) 객체 인식 (Object Recognition) :

객체 위치를 표시하여 대상 객체 구별

3) 영역 구분 (Segmentation) :

여러 객체의 위치와 속성을 분류

4) 의미 분할 (Semantic Segmentation) :

의미(속성) 단위로 픽셀을 분류

5. 데이터 라벨링 구축 과정 5단계

임무 정의 - 데이터 획득 - 데이터 정제 - 데이터 라벨링 - 데이터 학습

### 1) 데이터 획득

★원시데이터 : 기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

다양한 교통수단을 구별하는 AI

기계학습을 목적으로 획득 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

### 2) 데이터 정제

★원천 데이터 : 필요한 형식이나 크기에 맞게 변형, 데이터의 중복 제거

개인정보 비식별화 처리, 수집된 교통수단의 번호판이 보이지 않게 가려준 형태의 데이터

원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링 데이터가 부여되지 않은 상태의 데이터

3) ★데이터 라벨링 : 인공지능이 학습에 활용할 수 있도록 라벨을 달아주는 작업

인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 원천 데이터에 부착하는 활동

★어노테이션 (Annotation) :

라벨링 공정에서 인간이 부여한 식별 기준을 기계가 인식할 수 있도록 선정된 데이터에 추가적인 정보를 기입하여 알고리즘이 이해할 수 있도록 만드는 과정

미가공 데이터를 의미 있는 단위(예: 자동차, 트럭, 보행자 등)로 구분하고 속성 정보를 입력하는 작업

## 2강. 인공지능과 빅데이터

### 1. 인공지능 서비스 4단계:

데이터 획득→ 데이터 가공(전처리)→ 모델 생성→ 실시간 서비스(API 개발)

>> 데이터 획득, 데이터 가공이 빅데이터 부분

1) 데이터 획득: 데이터 수집(사진, 영상, 텍스트)→ IoT, 모바일, M2M으로 수집.

\*기계에서 기계로 수집 : M2M 사물(지능) 통신을 의미하며 기계간의 통신 및 사람이 동작하는 디바이스와 기계간의 통신을 말하며, 인간의 개입 없는 무인화, 지능화 서비스 등을 통해 자동으로 수집하는 방식으로 발전하고 있다.

2) 데이터가공(전처리) : 수집된 데이터는 인공지능이 이해할 수 있게 변경 필요. 인공지능에서 가장 중요. 가장 많은 시간이 소비된다.

\*데이터 라벨링 : 인공지능이 학습할 수 있는 형태로 가공하는 작업 (데이터 입력→ 데이터 학습에 필요한 과정)

3) 모델 생성 : 모델 개발 → 데이터 입력 데이터 학습 모델 수정 (이 과정 반복)

머신러닝(ML): 지능형 에이전트, 행동이나 협업 지능 / 시각, 언어, 청각기능/상황, 감정 이해 /추론, 지식 표현

\*머신러닝 학습 방법: 지도 학습/비지도 학습/강화 학습

4) 실시간 서비스(API 개발) : 모델 생성과 실시간 서비스가 인공지능 영역

### 2. 인공지능 발전 3단계

1) 1차 AI 붐 시대: 1960~1970. AI 개념 확립. 앨런 튜링 (인공지능의 아버지) : 기계가 생각할 수 있는지 테스트하는 방법 정립, 인공지능(지능적 기계)에 대한 개발 가능성 연구.



2) 2차 AI 붐 시대: 1980~1990. 컴퓨터 보급과 동시에 인공지능같이 발전 / 전문가 시스템 : 자신의 지식을 직접 입력하고 정해진 규칙을 만들어 동작하는 시스템 만듦, 문제점이 많았음 서로 달라서 (확일화 X, 비효율적, 대규모 개발 시 개발비, 유지보수비 높았음

3) 3차 AI 붐 시대 : 2000~현재. 머신러닝과 딥러닝의 기술 발전으로 인공지능이 같이 발전하게 됨. 인공지능 기술을 일반인도 쉽게 사용. 빅데이터 세션이 필수. 인공지능과 빅데이터는 아주 밀접한 관계. 현재 인공지능은 학습할 수 있는 스스로 학습. 학습할 데이터를 사람이 제공하면 스스로 판단하는 인공지능 구현 방식으로 발전. 하드웨어 향상 및 복잡한 연산 가능. 고성능 컴퓨터 인공지능 하드웨어 향상(엔비디아, GPU=그래픽 연산 빠르게 여러 개의 작은 단위 코어를 나누어 빠르게 학습 빠른 시간에 학습 가능) 급격한 발전 중. 기술발전 클라우드 시스템 발전. 현재 기술은 많은 데이터 하드웨어 결합 스스로 학습 처리 인공지능 구현.

### 3. 인공지능 현시점(딥러닝) 인공지능이 문제를 해결하는 과정

기존에는 미리 공식을 알려줬음 → 정해진 규칙대로만 계산, 환경이 바뀌면 오류 발생. 인공지능은 새로운 문제를 해결하기 위해 필요하기 때문에 문제와 답을 통해 공식 발견 (기존과 반대) 문제와 답이 달라져 도학습을 통해 변화된 공식을 찾아냄 → 이걸 위해서는 많은 데이터가 필요. 많은 양의 문제와 답을 인공지능에 제공하여 가장 근접한 공식 찾아냄. 다양한 분야에서 활용 가능. 사람이 문제를 해결하는 방법과 동일한 문제해결 방법이다. 인공지능이 정확한 문제 해결방법을 찾아내기 위해서는 많은 양의 문제와 답이 필요하다.

## 3강. 인공지능 학습원리

### 1. 인공지능 방법론 (인공지능의 분류)

#### 1) 인공지능의 원리

사람의 뇌를 흉내 내는 인공신경망과 다양한 머신러닝 알고리즘을 통해 구현됨.

퍼셉트론(Perceptron): 딥러닝(신경망)의 기원이 되는 알고리즘. 프랑크 로젠블라트가 1957년에 고안. 딥러닝을 배우기 위해서는 퍼셉트론의 구조를 배우는 것이 매우 중요함.

인공신경망(ANN=Artificial Neural Network)의 표현방식 : 생물학의 신경망에서 영감을 얻은 학습 알고리즘

#### 2) 인공신경망(ANN) 구조

입력층(input layer) : 학습하고자 하는 데이터를 입력하는 층

은닉층(hidden layer) : 입력된 데이터를 여러 단계로 처리하는 층

출력층(output layer) : 처리된 결과를 출력하는 층

#### 3) 인공지능 방법론(인공지능의 분류)

인공지능: 인간의 지적 능력을 컴퓨터를 통한 구현 단계(최종 목적)/사이버네틱스 전문가 시스템→ 현재는 불가, 최종 목표

머신러닝: 스스로 학습하여 인공지능의 성능을 향상하는 기술/인공신경망, 결정 트리, 베이지 넷 워크 등.

딥러닝: 인간의 뉴런과 비슷한 인공신경망으로 학습하는 방법(머신러닝의 한계를 넘어서는 기술-CNN, RNN, LSTM, GRU) 딥러닝은 머신러닝의 구현 방식 중 하나로 가장 좋은 성능을 내는 방식임. 인간의 뉴런과 비슷한 인공신경망으로 학습 방법 (사람이 생각하는 방식대로 학습)

#### 4) 머신러닝 학습

(1) 지도학습(supervised learning) : 문제와 정답을 알려주고 학습(예측Linear regression (회귀=학습한 내용을 바탕으로 미래에 어떤 값이 나올지 예측하는 것)), 분류 Classification techniques >>딥러닝에 해당. 비지도 학습보다 단순하고 일반적 레이블이 지정된 데이터(라벨링)를 사용 (많은 양의 데이터 필요→ 데이터 라벨링)

(2) 비지도 학습(unsupervised learning): 답을 가르쳐 주지 않고 학습. 연관규칙, 군집을 할 수 있음.

(3) 강화 학습(reinforcement learning): 보상을 통해 학습하는 방식

5) 머신러닝과 딥러닝의 차이 - 기계 자기 학습 여부로 차이가 남

머신러닝은 데이터 스스로 학습, 데이터의 여러 특징 중 사람이 직접 분석, 판단(사람 개입 필요)

딥러닝은 기계가 자동으로 학습 데이터에서 특징 추출함. 인간의 뉴런과 비슷한 인공신경망으로 학습.

## 2. 인공지능 알고리즘

### (1) 딥러닝의 표현 방식

딥러닝 : 기계가 자동으로 대규모 데이터에서 패턴과 규칙을 학습. 학습을 기반으로 의사결정이나 예측 등을 수행하는 기술.

인공신경망과 딥러닝의 다른 점은 은닉층의 구조가 다르다는 점 (딥러닝은 은닉층이 하나 이상으로 복잡한 구조를 가짐) 여러층(입력층, 은닉층, 출력층)을 가진 인공신경망(ANN)을 사용하여 머신러닝 학습을 수행함.

### (2) 딥러닝 동작 원리

인공지능이 학습하기 위해서는 많은 양의 데이터가 필요하다. 그러나 무조건 많기보다 필요한 데이터를 전처리해서 제공해 줘야 한다. 성능은 학습 데이터의 품질에 영향을 받는다. 딥러닝 학습을 위한 데이터는 훈련 데이터(train)와 평가 데이터(test)로 분류되어 사용된다. (8:2로 나눠서 8은 훈련 데이터(train), 2는 평가 데이터(test)로 사용) 총 70,000개의 데이터가 있다고 가정하면 훈련 데이터는 86%인 60,000개의 인공지능 학습용 데이터로 학습하고, 평가 데이터는 14%인 10,000개 정도가 평가에 상요 및 학습 후 정확도를 분석하게 된다. 네트워크 훈련용 데이터와 평가용 데이터로 분류하는 이유-평가 데이터는 학습 후 정확성 평가

### (3) 인공지능 프로그램의 개발 절차

1단계: 라이브러리 읽어 들이기 : 전문가가 미리 만들어놓은 프리셋 (시간과 비용을 줄임)

2단계: 데이터를 읽어 들이고 전처리하기: 데이터 라벨링이 필요

3단계: 신경망 만들기: 인공지능 라이브러리 이용

4단계: 모델 만들기(학습하기) : 라벨링 된 데이터를 학습시킴 (다소 시간이 소요)

5단계: 모델 적용하기(예측하기) : 실제 잘 동작하는지 성능 평가

2단계~4단계 학습데이터(데이터 제공) : 데이터 전처리, 데이터 셋 활용

5단계~3단계 인공지능 개발 : 데이터 셋 활용 인공지능 학습, 결과 예측 후 수정사항이 있는 경우 신경망 만들기로 다시 이동

## 3. 인공지능 객체 검출 방법의 이해

1) single object: 검출하고자 하는 객체가 하나인 경우

1단계:분류 확인(Classification) : 데이터 셋(데이터와 정답 레이블)을 함께 학습한 인공지능은 이를 토대로 새로운 이미지를 식별하게 되는 과정, 학습되지 않은 클래스 class는

인식하지 못한다

2단계:영역 표시(Localization) : 분류를 통해 검출한 객체의 정보를 보기 쉽게 박스 형태로 지정하는 것이다.

바운딩 박스: 학습을 통해 검출한 객체의 영역을 사격형으로 표시.

2) Multi object: 검출하고자 하는 객체가 여러 개인 경우

1단계: 객체 검출(object detection) : 학습을 통해 여러 개의 객체를 인식하고 인식된 객체를

바운딩 박스와 색을 이용해 영역을 표시하는 과정/검출된 객체는 바운딩 박스, 색으로 구분

2단계: 세그멘테이션: 의미적 분할(instance segmentation) : 객체 인식에서 이미지 내의 의미 있는 단위로 분할하는 작업/

정교하고 복잡한 인공지능 구현을 위해 이미지의 영역별 의미를 부여하는 경우 사용

→ 주로 자율주행에서 사용 (단순히 바운딩박스가 아니라 정확하게 세모, 동그라미 이렇게 디테일하게 구분)

#### 4. 핵심 딥러닝 알고리즘 이해

1) CNN(합성곱 신경망, Convolutional Neural Network) : 사진, 영상처리에 많이 사용/합성곱 사용 인공신경망 합성 곱을 이용해 가중치 수를 줄여 이미지 처리에 효과적, 이미지의 특징점을 효과적으로 찾을 수 있는 신경망/ 데이터의 특징을 분석하여 패턴을 파악하는 구조로 Convolution 과정과 Pooling 과정을 통해 진행 (사진이나 영상에서 이미지 패턴을 분석해 어떤 종류인지 판별)

2) RNN(순환 신경망, Recurrent Neural Network) : 음성, 텍스트 처리에 사용/ 계층의 출력이 순환 구조. 계층의 출력이 순환하는 신경망, 순환 방식은 은닉 계층의 결과가 다음 계층으로 넘어가며, 자기 계층으로 다시 되돌아온 다/시계열 정보처리처럼 앞뒤 신호와 상관도가 있는 경우. 음성, 웨이브폼, 텍스트의 앞뒤를 분석하는 등 언어 처리에 사용

3) GAN(생성적 적대 신경망, Generative Adversarial Network) : 신경망이 2개 존재/이미지 생성, 복원 등

4) 세그멘테이션: 의미적 분할(instance segmentation) : 객체 인식에서 이미지 내의 의미 있는 단위로 분할하는 작업/정교하고 복잡한 인공지능 구현을 위해 이미지의 영역별 의미를 부여하는 경우 사용→주로 자율주행에서 사용 (단순히 바운딩박스가 아니라 정확하게 세모, 동그라미 이렇게 디테일하게 구분)

#### 5. 인공지능과 데이터의 상관관계

AI 인공지능-인공지능은 학습하기 좋은 데이터(라벨링)가 필요.

빅데이터 - 데이터는 인공지능 학습을 위한 가공(전처리)이 필요.

인공지능 성능 향상을 위해서는 양질의 데이터가 충분히 제공되어야 함

인공지능(AI)-건강한 사람 / 빅데이터 - 좋은 음식이라고 볼 수 있음

데이터가 충분해야 인공지능도 뛰어남, 인공지능 개발 시 데이터 처리가 80% 필요

6. 데이터 라벨링 : 기계가 이해할 수 있는 형태로 가공 (정의 획득 - 정제 - 라벨링 학습으로 진행)

1) 데이터 정의 : 인공지능이 학습할 데이터를 만들 경우 어떠한 데이터가 필요한지 정의, 분석 / 구축 계획서 작성

2) 데이터 획득 : 기존 데이터와 다양한 경로로 확보한 데이터/부족한 데이터는 크롤링 작업 통해 확보

3) 데이터 정제 : 인공지능이 학습할 수 있는 형태로 분류, 가공함 / 원천 데이터 생성

4) 데이터 라벨링 : 인공지능이 학습할 수 있는 라벨링 데이터 만들

5) 데이터 학습: 원천 데이터와 라벨링 데이터를 학습 (데이터 셋)

(1) 인공지능 개발에 소요되는 시간: 데이터 처리에 80%를 소용

데이터 식별 5%

데이터수집 10%

데이터 정제 25%

데이터 라벨링 25% : 인공지능 모델학습을 위한 데이터를 기계가 이해할 수 있는 형태로 가공

데이터 증강 15%

(2) 인공지능개발 소요 시간 : AI 인공지능 20%를 소용

AI 서비스 배포 2%

AI 모델 조정 5%

AI 모델학습 10%

AI 알고리즘 개발 3%

데이터 셋: 인공지능 학습을 위해 필요한 데이터를 모아놓은 자료의 집합.

즉 원천 데이터와 라벨링 데이터를 모아 놓은 자료의 집합

주의할 점- 인공지능 및 빅데이터의 데이터 생성, 관리 시에 저작권과 초상권에 주의해야 한다.

#### 7. 개인정보보호

1) 저작권: 사람의 생각이나 감정을 통하여 창작적인 표현의 결과물(저작권격권, 저작재산권 등) 이미지, 폰트, 뉴스/이미지 는 판매 사이트에서 정식으로 구입하거나 무료 이미지 사용 폰트도 마찬가지로 뉴스나 기사 포털 정보는 일부 발췌하거나 출처 표기, 저작권 범위 확인 필요

2) 초상권: 사람의 얼굴이나 통념상 특정인임을 식별할 수 있는 신체적 특징에 관하여 촬영 또는 그림 묘사되거나 공표되지 않으며 영리적으로 사용 불가

개인정보는 익명화하여 사용 (신체정보, 정신 정보, 재산 정보, 사회적 정보) 가명처리, 범주화, \*표 마스킹, 부분 표기

#### 8. 데이터 라벨링 작업 (실습)

데이터 라벨러 : 데이터의 수집에서 가공에 이르기까지 인공지능 학습에 필요한 형태의 데이터를 만드는 사람

크롤링: 인터넷의 방대한 데이터를 수집하는 행위

#### 9. 회귀(Linear regression)와 분류

1) 회귀: 가지고 있는 데이터에 독립변수와 종속변수가 있고, 종속변수가 숫자일 때 회귀를 이용/종속변수-양적 데이터

회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링 등의 통계적 예측에 이용 회귀(영어: regress 리그레스[\*])의 원래 의미는 옛날 상태로 돌아가는 것을 의미

영국의 유전학자 프랜시스 골턴은 부모의 키와 아이들의 키 사이의 연관 관계를 연구하면서 부모와 자녀의 키 사이에는 선형적인 관계가 있고 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세웠으며 이를 분석하는 방법 이러한 경험적 연구 이후, 칼 피어슨은 아버지와 아들의 키를 조사한 결과를 바탕으로 함수 관계를 도출하여 회귀 분석 이론을 수학적으로 정립

2) 분류: 가지고 있는 데이터에 독립변수와 종속변수가 있고, 종속변수가 이룸일 때 사용/종속변수 - 범주형 데이터

인공지능 서비스(실시간 서비스(API 개발/모델 생성/데이터 가공(전처리)/데이터 획득)

## 4강. 데이터 라벨링의 이해

### 1. 데이터 라벨러

인공지능 프로그램 개발을 위해 인공지능이 학습 데이터를 쉽게 인식할 수 있도록 텍스트, 사진 이미지, 동영상, 사운드 등의 파일에 등장하는 사물, 동식물, 특정 단어 등에 라벨을 수집하고 입력하여 가공하는 사람

#### 1) 데이터 라벨러 전망

국내 인공지능 시장 매년 14.9% 성장 2027년 4조 5천억 규모 전망

인공지능 산업이 성장하면서 많은 데이터 가공과 초거대 인공지능 모델 수요 증가에 따른 데이터라벨러 전문적 역량 필요 예상->그러나 데이터 라벨러 자체의 전망이 좋은 것은 아니며, 고도화될수록 전문역량이 중요해진다.

#### 2) 데이터 라벨러의 업무

인공지능 개발 프로세스 이해:

인공지능이 어떠한 단계를 거쳐 개발되는지 그 과정을 이해함으로써 데이터 라벨링이 가지는 중요성을 이해하고 성취감을 느끼며 일할 수 있음

#### 3) 프로젝트 이해:

프로젝트의 목적을 이해하고, 라벨링 작업의 의미를 이해함으로써, 프로젝트 참여자 모두가 동일한 목표 지향점을 갖고 책임감 있게 일할 수 있음

#### 4) 라벨링 가이드라인 이해:

고품질의 라벨링 데이터를 생산하기 위해 각 프로젝트의 라벨링 가이드라인을 심도 있게 이해함으로써, 효율적인 업무 진행 및 품질 유지가 가능

#### 5) 데이터 라벨링 수행:

인공지능의 성능에 영향을 줄 수 있는 만큼, 양질의 라벨링 데이터를 생산하기 위해 정교한 데이터 라벨링 업무 수행이 요구됨

#### 6) 관리자와 소통:

관리자(QM, PM)와 이슈 보고, 업무 제안, 질의응답 등의 적극적인 소통을 통해 고품질 데이터 생산과 프로젝트 운영에 도움을 줄 수 있음

### 2. 인공지능

인간의 학습능력, 추론 능력, 지각 능력을 인공적으로 구현하려는 컴퓨터 과학 기술 구현을 위해 기술 연구분야는 머신러닝, 딥러닝 등

#### 1) 인공지능 학습용 데이터

머신러닝, 딥러닝 등 인공지능 모델 학습을 위해 활용되는 데이터를 총칭  
데이터의 정답이 표기(라벨)된 데이터를 학습하는 지도학습용 데이터  
정답이 표기되지 않은 비지도 학습용 데이터

#### 2) 이미지 학습용 데이터

인공지능(AI)의 한 분야인 컴퓨터 비전에 사용되는 학습 데이터

컴퓨터 비전은 컴퓨터에게 시각 데이터 처리 능력을 부여하는 기술

카메라와 센서가 인간의 눈 기능을 한다면, 컴퓨터 비전은 시각 데이터를 처리하는 인지능력

### 3. 컴퓨터 비전

#### 이미지 분류

→ 이미지를 보고 이를 분류합니다. 예를 들어 개, 사과, 사람 얼굴보다 정확하게 주어진 이미지가 특정 클래스에 속하는지 정확하게 예측

#### 객체 감지

→ 이미지 분류를 사용하여 특정 클래스의 이미지를 식별한 다음 이미지, 비디오에서 모양을 감지

#### 객체 추적

→ 감지된 객체를 추적합니다. 예를 들어 자율주행 차량은 보행자, 다른 자동차, 도로 구조물과 같은 물체를 분류하고 감지할 뿐만 아니라 충돌을 피하고 교통 법규를 준수하기 위해 움직이는 물체를 추적

### 4. 콘텐츠 기반 이미지 검색

→ 컴퓨터 비전을 사용하여 연결된 메타데이터 태그가 아닌 이미지의 콘텐츠를 기반으로 대규모 데이터 저장소에서 이미지를 탐색하고 검색

#### 이미지/영상 데이터 라벨링:

인공지능 학습에 있어 이미지 라벨링은 이미지 해석을 위한 머신러닝 솔루션과 다양한 응용분야(컴퓨터 비전, 로봇 비전, 안면 인식)에 있어서 필수

인공지능을 학습시키기 위해서는 식별자, 캡션 또는 키워드 형식의 의미 있는 데이터의 정보를 이미지에 할당

### 5. 디지털 이미지

이미지 또는 영상은 2차원 평면 위에 그려진 시각적 표현물

영상은 크게 정지 영상(사진)과 동영상으로 구분되는데 근래에는 영상이 곧바로 동영상을 의미

디지털 이미지는 디지털카메라를 이용하여 현실 세계의 사물을 촬영하거나 스캐너를 이용하여 사진이나 그림을 디지털 형태로 받아들인 것

사람은 이미지를 보지만 컴퓨터는 숫자를 본다

### 6. 디지털 영상

이미지가 정지하고 있는 하나의 프레임이라면 여러 개의 프레임을 연속적으로 볼 수 있는

영상 자료 움직이는 영상을 전자 매체에 기록 프레임 단위: fps(Frames Per Second)

#### 이미지 분류 모델 학습

디지털 이미지, 비디오 및 기타 시각적 입력에서 의미 있는 정보를 구분하여 입력

구분된 정보를 바탕으로 데이터의 특징을 구분하는 규칙 생성

### 7. 데이터 라벨링 과정

#### 1) 데이터 획득/수집(Data Acquisition / Collect)

인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 수집 또는 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 법률적 제약이 없도록 '원시데이터'를 확보하는 활동

#### 2) 데이터 정제(Data Refinement)

획득한 원시데이터를 기계학습에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인

정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 '원천 데이터'를 확보하는 활동

### 3) 데이터 가공(Data Labeling)

인공지능 기계학습에 활용할 수 있도록 가능이나 목적에 부합하는 정보를 원천 데이터에 부착하는 활동

### 4) 데이터 학습(Data Machine Learning)

학습 데이터 셋의 훈련 데이터 셋, 검증데이터셋을 이용하여 선정된 인공지능 알고리즘을 학습시키고, 학습된 인공지능 모델의 성능을 향상시키거나 보정하는 활동

## 8. 인공지능 학습용 데이터 구축(데이터 구분)

### 1) 원시데이터(Raw Data)

기계학습을 목적으로 획득한 단계에서 수집 또는 생성한 음성, 이미지, 영상, 텍스트 등의 데이터

### 2) 원천 데이터(Source Data, Unlabeled Data)

원시데이터를 라벨링 공정에 투입하기 위해 필요한 전처리 등 정제 작업을 수행한 데이터로 라벨링 데이터가 부여되지 않은 상태의 데이터

### 3) 라벨링 데이터(Labeled Data)

원천 데이터에 부여한 참값, 파일 형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

### 4) 학습 데이터(셋)(AI Data Set)

인공지능의 기계학습에 사용하는 원천 데이터와 라벨링 데이터의 묶음을 말하며, 사용하는 목적에 따라 '훈련용(Training)', '검증용(Validation)', '시험용(Test)'으로 구분

## 9. 학습용 데이터 처리 과정에서 알아 두면 유용한 용어

### 1) 인공지능(Artificial Intelligence)

자연 언어의 이해, 음성 번역, 문제 해결, 학습과 지식 획득, 인지 과학 등에 응용하기 위해 인간의 지능이 갖는 학습, 추리, 적응, 논증 등의 기능을 갖춘 컴퓨터 시스템

### 2) 기계학습(Machine Learning)

인간의 학습 능력을 기계를 통해 구현하는 것

### 3) 라벨(Label)

데이터와 그에 부착된 라벨링 정보들(어노테이션 annotations)을 지칭하는 용어

### 4) 데이터 라벨링(data Labeling)

원천 데이터에 인공지능이 학습할 수 있도록 정보를 부착한 데이터

(원천 데이터에 "참값", "설명", "주석"등이 포함된 데이터)

### 5) 데이터 라벨러(data Labeler)

데이터 라벨링을 수행하는 사람

### 6) 라벨링 데이터(Labeled data)

원천 데이터에 부여한 '참값', 파일 형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

## 10. 학습용 데이터 처리 과정에서 알아 두면 유용한 용어

### 1) 참값(Ground Truth)

인공지능의 기계학습 목적에 따라 원천 데이터에 라벨링된 정확한 값이나 사실의 의미적 표현

### 2) 클래스(Class) = 카테고리(Category)

분류/탐지하고자 하는 대상을 카테고리화한 것으로, 분류체계를 의미

### 3) 어노테이션(Annotation)

인공지능이 데이터의 내용을 이해할 수 있도록 주석을 달아주는 작업  
(라벨링 공정에서 인간이 부여한 식별 기준을 기계가 인식할 수 있도록 선정된 데이터에  
기입된 추가적인 정보)

#### 4) PM(Project Manager)

프로젝트 전반의 전략 수립과 운영을 맡아 관리하는 직책

#### 5) QM(Quality Manager)

데이터 수집/라벨링 및 검수, 인력 관리를 맡아 데이터 품질을 관리하는 직책

### 11. 인공지능 개발 프로세스 이해

Project Setup-> Data Preparation-> Model Training-> Deploying

Data Preparation과정 안에서

구축 계획 수립 → 데이터 획득/수집(원시데이터) → 데이터 정제(원천 데이터) → 데이터가공  
(라벨링 데이터) → 데이터 학습(학습 데이터 셋)

### 12. 인공지능 학습용 이미지 데이터 구축

#### 1) 사례 1

라벨링 → 라벨러 배정 → 라벨링(분류, 폴리곤) → 라벨링 검수

#### 2) 사례 2

라벨링→동작 구간 태깅→ 속성정보(JSON) 제이슨 형식

→스크립트 추출→발화 대본(TXT) 텍스트 파일

데이터 라벨링: 모션 데이터 라벨링 (자체 제작도구 개발)

라벨링 데이터 셋- 동작 구간 속성정보(JSON)- 발화 대본(TXT)

### 13. 이미지 학습용 데이터 특성

이미지 데이터 형식이 다양함( 흑백, RGB, XY, 의미)

점, 선, 면, 형태 등 다양한 표현으로 데이터를 생성할 수 있음

이미지 데이터 라벨링 프로세스 이해

라벨링 규칙에 따라 원천 데이터에 주석을 달아주며, 데이터 라벨러는 이 과정에 참여하여  
라벨링 데이터를 생산함.

#### 1. 검수 기초 이론

데이터 검수자 : 데이터를 '검수'하는 사람

획득 - 정제 - 라벨링 - 검사

인공지능 모델의 목적과 특성에 맞게

수집 가공된 데이터의 품질을 확인하는 과정

#### 2. 검수 순서

1) 검수 가이드 작업 가이드 숙지

2) 집중할 수 있는 환경 만들기

3) 검수 진행

4) 작업된 데이터 꼼꼼히 살펴보기

5) 반려 사유를 구체적이고, 명확하게 작성하기

6) 문제 발생 시, 데이터 PM과 소통하기

#### 3. 검수하기 전 준비사항

1) 가이드 숙지 : 검수 가이드는 물론 작업 가이드까지도 꼼꼼히 확인해야 합니다.



2) 작업환경 : 집중할 수 있는 환경 만들기

3) 반려 시에는 구체적이고 명확하게 반려 사유를 작성해서 전달해야 합니다.

4) 반려시 반려 사유와 작업해야 하는 내용을 명확하고 구체적으로 작성하여야 합니다.

4. 검수자의 잘못된 행동(잘못된 검수 사례)

1) 가이드의 미숙지

여백 없이 바운딩 해주세요 -> 사람의 얼굴은 1cm 정도의 여백을 두어 바운딩 해주세요.

2) 무성의하거나 부정한 방법으로 검수 진행

작업 내용 제대로 확인하지 않고 검수 완료한다든지

매크로와 같은 컴퓨터 프로그램을 이용하여 검수 진행 행위

3) 자의적은 판단에 의한 독단적인 행동

정해진 기준을 준수하여 일관된 검수를 진행해 주세요.

4) 작업자와의 갈등 유발

작업자와 검수자는 각자의 업무가 다를 뿐 동등한 위치에 있음

작업자를 하대하거나, 무시, 비난 행위 잘못된 행위

5. 반려 사유 작성법

1) 작업자의 입장에서 생각해 본 후 작성하기

2) 잘못된 부분은 명확하게, 수정해야 할 내용은 구체적으로 작성하기

3) 상대방을 존중하는 표현 사용하여 작성하기

1. 데이터 라벨링 구축 및 검사

데이터 정제: 데이터 라벨러는 라벨링 프로세스 시작 전이나 라벨링 중 부적합한 데이터를 정제하는 업무를 수행

작업 검사: 데이터 라벨링에 숙달되면 검사자 업무에 참여할 수 있으며, 라벨러로 참여 중 프로젝트 외 다른 프로젝트 검사자로 참여할 수 있음

라벨링 데이터 검사: 라벨링이 완료된 데이터를 검수하여 오류를 식별하고 품질을 확인

라벨링 데이터 세트 구축: 라벨이 부착된 데이터를 수집하여 데이터 셋 구축, 데이터의

형식과 포맷을 정리하고 필요한 경우 추가적인 전처리를 수행

라벨링 데이터 활용: 완성된 데이터 셋 활용하여 인공지능 모델의 학습이나 다른 분석 작업들에 활용

(1) 사전 작업에서 중요점

명확한 레이블 지정 지침 정의: 명확한 지침을 정의해야 함

일관된 라벨 지정 기준 사용: 프로세스 전체에서 일관성이 있어야 함

라벨링 품질 유지: 라벨링 품질 유지를 위해 정확성과 일관성을 확인

컨텍스트 포함

개인 정보 보호 및 보안 보장

피드백 제공: 의심스러운 경우 질문을 하거나 설명을 구하도록 요청.

(2) 라벨링 기법에 따른 데이터 형식 이해

키포인트: 지정한 지점의 X, Y 좌표를 표시함

바운딩박스: X, Y 두 점의 좌표를 생성하여 사각형 영역

폴리곤, 폴리 라인: 지정된 지점의 좌표들의 연속적인 위치

세그멘테이션: 폴리곤, 폴리 라인에서 지정한 좌표들의 영역

이미지 라벨링 적용 속성 데이터 (JSON)

## 5강. 6강. 데이터 라벨링 기법 개요 및 저작도구 바운딩 박스

### 1. 라벨링 기법 정리

#### 가. 바운딩 박스

객체의 범위를 사각형으로 지정하는 기법이며, 데이터 라벨링에서 가장 많이 사용

#### 나. 키포인트

객체의 주요 지점(특징)을 점으로 지정하는 기법이며, 이미지 매칭, 안면 인식, 골격 추출 등에 사용

#### 다. 폴리 라인

선형 객체의 위치를 연속선으로 지정하는 기법이며, 선형 데이터의 구분, 인식을 위해 사용

#### 라. 폴리곤

객체의 범위를 다각형으로 지정하는 기법이며, 불규칙한 객체의 경계를 정교하게 라벨링 할 수 있어 정교한 인공지능 모델 개발에 사용

#### 마. 시멘틱 세그멘테이션

이미지의 모든 픽셀에 클래스를 부여하는 라벨링 기법

#### 바. 인스턴스 세그멘테이션

이미지의 모든 픽셀이 같은 클래스여도 다른 인스턴스를 부여하는 라벨링 방법

#### 사. 팬 옵틱 세그멘테이션

시멘틱 세그멘테이션(Stuff클래스) + 인스턴스 세그멘테이션(Thing 클래스)을 결합한 방식

#### 아. 큐보이드

객체의 범위를 직육면체로 지정하는 기법이며, 객체의 너비, 높이, 깊이 및 방향을 알 수 있어 객체의 공간감과 방향을 표현

#### 자. 비디오 어노테이션

영상에 대하여 라벨링 하는 방법을 통칭하며, 구간 정제, 분류, 정보 태깅 등의 방법으로 객체 인식, 객체 추적에 주로 사용

#### 차. OCR

이미지/영상 속 문자를 기계가 읽을 수 있게 텍스트로 입력하며, 바운딩 박스 기법과 같이 사용

### 2. 객체 식별 방법

1) 객체인식(Object Detection) Bounding Box, Keypoint, Polygon, Polyline

2) 키포인트 검출(Keypoint Detection) Keypoint, Segmentation

3) 얼굴 인식(Face Recognition) Bounding Box, Keypoint, Polyline, Tagging

4) 이미지 분류(Image Classification) Bounding Box, Polygon

5) 행동 인식(Action Recognition) Bounding Box, Keypoint, Polygon, Polyline

6) 비디오분류(Video Classification) Bounding Box, Polyline, Tagging

7) 자세 추정(Pose Estimation) Bounding Box, Keypoint

8) 비디오 인식(Video Recognition) Polygon, Segmentation

### 3. 저작도구 용어

1) 저작도구 (Authoring tool) :

저작 (여러 형태의 데이터를 편집)에 사용되는 소프트웨어

2) 태스크 (Task) :

작업을 태스크라고 하며, 대기/작업 완료/반려/완료 등의 상태 속성을 가짐

#### 4. 저작도구 소개

##### 1) 오픈 소스 데이터 라벨링 저작도구

###### (1) Label Studio :

이미지, 음성, 텍스트, 영상 하이라이트, 다중 도메인 애플리케이션 등 다양한 분야의 데이터를 다룰 수 있는 저작도구 (<https://labelstud.io>)

###### (2) Label Box :

다양한 도메인 데이터 라벨링 지원 및 라벨링 인력에 대한 협업 및 관리 지원이 가능한 저작도구 (<https://labelbox.com>)

###### (3) CVAT (Computer Vision Annotation Tool) :

컴퓨터 비전 라벨링을 지정하는데 주로 활용되는 웹 기반 이미지 및 비디오 저작도구 (<https://cvat.org>)

###### (4) LabelImg :

Object Detection을 위한 Bounding Box 저작도구, 데이터 출력 시 xml로 저장 (<https://github.com/tzutalin/labelImg>)

##### 2) 이미지/영상 교육 과정 실습 플랫폼 - 데이터 메이커

###### (1) Datamaker Annotator :

이미지, 영상 라벨링 저작도구 및 클라우드 소싱 플랫폼 (<https://dashboard.datamaker.io>)

###### 3) 오픈 소스, 무료 계열

오픈소스로 배포되고 있어 누구나 사용할 수 있음

하지만 사용자가 환경 설정 등을 직접 해야 하며 버전 및 오류 관리 단점

최근 학교나 개인단의 연구에서 벗어나 점점 서비스의 형태로 진화

종류: Label Studio, CVAT, Image Tagger, Make Sense, COCO Annotator, Diffgram

###### (1) CVAT

실습에서 사용할 도구

intel에서 개발한 웹(크롬) 기반 저작도구이며, 이미지 및 영상

라벨링 지원 저작도구, 자동 라벨링, 공동 작업 관리 기능 지원

##### 4) 유료 서비스 계열

통합적인 데이터 라벨링 기능을 지원하면서 학습 데이터 생성에 필요한 기능과 요소들을 효율적으로 작동

하지만 일정 수준의 비용을 지불해야 하는 단점

그럼에도 유료 서비스 데이터 라벨링은 전문성과 편리한 기능으로 그 활용도가 높음

예를 들면 웹 기반으로 언제 어디서나 접근할 수 있으며 학습 데이터 관리자가 데이터

라벨러의 진도, 신뢰성 등을 항상 모니터링 기능

라벨러 입장에서 효율적인 라벨링 기능 활용

국내 기업: 테스트 워크스, 클라우드 워크스, 알 체라, 슈퍼 브 AI

###### (1) 용어정리

프로젝트 특정 목표를 성취하기 위해 데이터, 라벨링, 자원/품질 관리 등을 실행하는 과제 단위

어노테이터 저작(여러 형태의 데이터를 편집)에 사용되는 소프트웨어, 저작도구라고

표현하기도 함

라벨링 가이드라인 라벨링 작업 방식과 기준이 기재된 문서

프로젝트별로 내용이 상이하므로, 가이드라인이라고 부르기도 함

객체 실체가 있는 대상(사물) 자체를 지칭

예) 자동차, 자전거, 사람, 표지판, 건물 도로, 하늘, 바다

라벨 데이터와 그에 부착된 라벨링 정보들(어노테이션)을 지칭하는 용어

어노테이션 라벨링 공정에서 인간이 부여한 식별 기준을 기계가 인식할 수 있도록 선정된 데이터에 기입된 추가적인 정보

라벨링 데이터

원천 데이터에 부여한 '참값', 파일 형식이나 해상도 등의 속성, 그리고 설명이나 주석 등이 포함된 '어노테이션'의 집합

클래스(Class)

분류/담지하고자 하는 대상을 카테고리화한 것으로, 분류체계를 의미

검사(Review)

기준에 적합하게 라벨링이 되었는지 검사함

작업 완료된 라벨은 검사를 거쳐 반려/완료 상태로 전환됨

반려(Return)

기준에 적합하게 라벨링 되지 않아 검수를 통과하지 못한 라벨의 상태로, 수정 작업을 통해 다시 검수 과정을 거쳐야 함

코멘트(Comment)

반려된 라벨을 수정 시 반려 사유를 기재한 내용

(2) 실습 저작도구 접속하기 (CVAT 프로그램)

작업에 사용될 저작도구에 접속한다.

프로젝트 생성 → 프로젝트 이름 작성 → 작성할 라벨 생성 및 속성 추가

태스크 생성 → 태스크 작성 → 작업 이미지 업로드

작업 시작하기 → 태스크 안에 생성된 JOB을 선택하여 할당된 작업을 시작할 수 있다 → 저작도구 작업 창으로 이동

(3) 저작도구 메뉴 익히기

상단 메뉴(Header)

헤더:CVAT 섹션 및 계정 설정을 탐색하는 데 사용되는 고정 헤더

저작도구(Controls Sidebar)

이미지 탐색, 확대/축소, 모양 생성 및 트랙 편집(병합, 분할, 그룹)을 위한 저작도구 포함

진행사항(Top Panel)

탐색 버튼, 주요 기능 및 메뉴 액세스를 포함

작업 현황(Objects Sidebar)

레이블 필터, 두 개의 목록 : 프레임에 있는 개체와 레이블(프레임에 있는 개체의) 및 모양 설정을 포함

작업 창(Workspace)

이미지가 표시되는 공간

(4) 상단 메뉴

Projects:프로젝트 생성 및 현황을 확인할 수 있다.

Tasks: 작업 생성 및 현황을 확인할 수 있다.

Jobs: 업무 생성 및 현황을 확인할 수 있다.

#### (5) 작업 현황

작업 중인 라벨의 클래스 및 속성 설정과 작업에 용이한 작업선 옵션을 선택할 수 있다.  
다양한 형식의 저작도구들이 있으면 오픈소스, 상용서비스 등 활용 범위와 적용 대상에 따라 구분할 수 있다.

작성 저작도구에서 사용되는 용어를 이해하고 작업 가이드라인에 따른 [프로젝트] 생성, [객체] 선택, [라벨] 작업, [클래스] 지정, [검수],[반려],[코멘트] 등의 프로세스 수행

#### (6) CVAT 저작도구 단축키

주요 기능

F1

단축키 확인

F2

설정

Ctrl+S

저장

Ctrl+Z

작업 뒤로 가기

Ctrl+Shift+Z 또는 Ctrl+Shift+Y

실행 취소 작업 취소

Mouse Wheel 누르기

이미지 프레임을 이동하려면

이미지 작업

Ctrl+R

이미지 시계방향 회전

Ctrl+Shift+R

이미지 시계 반대 방향 회전

플레이어

F

다음 프레임(이미지) 이동

D

이전 프레임(이미지) 이동

V

한 걸음 앞으로 나아가기

C

한 걸음 뒤로 가기

Right

필터를 만족하는 다음 프레임 또는 객체가 포함된 다음 프레임을 검색

Left

필터를 만족하는 이전 프레임 또는 객체가 포함된 이전 프레임을 검색

Spce

자동 변경 프레임 시작/중지

'또는~

현재 프레임을 변경하려면 요소에 초점을 맞춤

모드

N

Draw Mode(새로 그리기)

M

merge Mode(미사용 단축키)

Alt+M

도형 분할 모드 활성화 또는 비활성화

G

그룹 모드

Shift+G

그룹 모드에서 그룹 리셋

Esc

취소

개체 작업

Ctrl

그리기/편집 중 폴리곤 및 폴리 라인에 대한 자동 경계 지정 전환

Ctrl+누르고

모양이 활성화되어 고정

Alt+Click 포인터

포인트 삭제(폴리곤, 폴리 라인, 포인트의 포인트 위에 마우스를 올렸을 때 사용)

Shift+Click 포인터

도형 편집

마우스 오른쪽 Click

객체 사이드바에서 개체 요소 표시

T+L

전체 객체 잠금 / 풀기

L

객체 잠금 / 풀기

T+H

전체 객체 숨기기 / 보이기

H

객체 숨기기 / 보이기

Del 또는 Shift+Del

활성 개체를 삭제합니다. 시프트를 사용하여 잠긴 객체를 강제로 삭제- 또는 \_

활성 객체를 사용자로부터 "더 멀리" 놓기(z 축 값 감소)

+ 또는 =

활성 객체를 사용자에게 "가까이"배치 (z 축 값 증가)

Ctrl+C

CVAT 내부 클립보드에 도형 복사

Ctrl+V

내부 CVAT 내부 클립보드에 도형 붙여넣기

Ctrl 누른 상태 계속 붙여넣기

여러 붙여넣기를 위해 버퍼에서 모양을 붙여 넣을 때

Ctrl+B

다음 프레임에서 개체의 복사본 생성

Ctrl+(0..9)

활성화된 개체 또는 개체가 활성화되지 않은 경우 다음에 그려진 개체에 대한 레이블을 변경  
작업은 추적에만 사용할 수 있음

K

활성 트랙의 키 프레임 속성 변경

O

활성 트랙의 외부 속성 변경

R

활성 트랙의 다음 키 프레임으로 이동

E

활성 트랙의 이전 키 프레임으로 이동

속성 주석 모드

Up Arrow

다음 속성으로 이동(위)

Down Arrow

다음 속성으로 이동(아래)

Tab

현재 프레임에서 주석이 달린 다음 객체로 이동

Shift+Tab

현재 프레임에서 주석이 달린 이전 객체로 이동

<number>

현재 속성에 해당 값 할당

표준 3D 모드

Shift+Up Arrow

카메라 롤 각도 증가

Shift+Down Arrow

카메라 롤 각도 감소

Shift+Left Arrow

카메라 피치 각도 감소

Shift+Right Arrow

카메라 피치 각도 증가

Alt+O

카메라를 위로 이동

Alt+U

카메라를 아래로 이동

Alt+J

카메라를 왼쪽으로 이동

Alt+L

카메라를 오른쪽으로 이동

Alt+I

확대 수행

Alt+K

축소 수행

1. 데이터 라벨링 기법

바운딩 박스(Bounding box)

키포인트(Keypoint)

폴리곤(Polygon)

폴리 라인(Polyline)

세그멘테이션(Segmentation)

큐보이드(Cuboid)

비디오 어노테이션(Video annotation)

OCR(Optical Character Recognition)

1) 바운딩 박스:

바운딩 박스는 이미지 혹은 영상 안 객체의 범위를 직사각형 모양의 박스로 지정하는 라벨링 기법 데이터 라벨링 작업에서 가장 일반적으로 사용

객체 전체가 커버되도록 하며, 박스 안에 객체 이외의 여백을 최소화하도록 지정

데이터 라벨링 기법 적용 사례 1

프로젝트 목적

교차로 신호 체계, 보행자, 차량 이동 복합 데이터

데이터 유형

이미지 (JPG, PNG)

라벨링 객체

신호정보, 교차로, 차량 종류

활용 분야

스마트 교차로, 스마트 시티, 자율주행

데이터 라벨링 기법 적용 사례 2

교차로 신호 체계, 보행자, 차량 이동 복합 데이터

프로젝트 목적

자율주행을 위한 자동차 인식 및 신호 판별

데이터 유형

자동차 주행 촬영 이미지 (JPG, PNG)

라벨링 객체

자동차, 신호등

활용 분야

자율주행 실시간 영상 인식 기술 개발

데이터 라벨링 기법 적용 사례 3

자율주행을 위한 자동차 인식 및 신호 판별



프로젝트 목적

드론을 통한 이동체 인지

데이터 유형

드론 촬영 이미지 (JPG, PNG)

라벨링 객체

이동체(자동차, 보행자)

활용 분야

객체 인식 기술 개발(교통분석, 조난자 수색 등)

데이터 라벨링 기법 적용 사례 4

드론을 통한 이동체 인지

프로젝트 목적

노지 작물의 해충 진단 및 처방

데이터 유형

노지 작물 촬영 이미지 (JPG, PNG)

라벨링 객체

해충

활용 분야

스마트 병해충 진단 및 예찰 기술 개발

데이터 라벨링 기법 적용 사례 5

프로젝트 목적

갑각류(꽃게) 종자 보존 및 종자 생산 기술 고도화 및 서비스 확산

데이터 유형

이미지 (JPG, PNG)

라벨링 객체

갑각류(꽃게)

활용 분야

꽃게 종자 생산장에서의 성장 관리 서비스 활용

데이터 라벨링 기법 적용 사례 6

프로젝트 목적

고령층 안전사고 신속 대응 및 예방

데이터 유형

CCTV 촬영 이미지(JPG, PNG)

라벨링 객체

이상행동자

활용 분야

인공지능 CCTV, 스마트 케어 서비스 개발

데이터 라벨링 기법 적용 사례 7

프로젝트 목적

운전자 및 탑승자 상태 모니터링

데이터 유형

운전자 및 탑승자 안면, 전신 촬영 이미지(JPG, PNG)

라벨링 객체

양쪽 눈, 입, 얼굴, 전신

활용 분야

운전자 및 탑승자 상태 인식 기술, 피로 감지 기술, 안면 인식 기술

데이터 라벨링 기법 적용 사례 8

프로젝트 목적

항만구조물, 해상 객체 식별을 통한 선교 상황 인식

데이터 유형

항만 구조물 촬영 이미지(JPG, PNG)

라벨링 객체

항만구조물, 선박, 부유물

활용 분야

자율운항 선박 기술, 무인 해양경비시스템

## 7강. 8강. 데이터 라벨링 기법 폴리 라인

1 폴리 라인:

폴리 라인은 여러 개의 점을 가진 선을 활용하여 선형 객체의 경계나 위치 등 특정 영역을 라벨링 하는 기법

인도와 차선, 경계 등의 선형 데이터들을 구분하고 인식하기 위해 사용

데이터 라벨링 기법 적용 사례 1

프로젝트 목적

양식 어류 행동 분석, 개체 추적 등의 수산 AI 개발

데이터 유형

양식 어류 촬영 이미지(JPG, PNG)

라벨링 객체

체장, 체고

활용 분야

스마트 양식 시스템 (관리 자동화)

데이터 라벨링 기법 적용 사례 2

프로젝트 목적

자율주행 중 도로 표면 인식 AI 개발

데이터 유형

도로 표면 촬영 이미지(JPG, PNG)

라벨링 객체

표면 크랙

활용 분야

도로 유지 보수 , 자율주행 회피 기동 기술 개발

데이터 라벨링 기법 적용 사례 3

프로젝트 목적

주행 영상에서 차선, 정지선 위치 및 종류 파악

데이터 유행

도로 주행 촬영 이미지(JPG, PNG)

라벨링 객체

차선(색깔, 종류), 정지선

활용 분야

자율주행 실시간 영상 인식 기술 개발

5) 시멘틱 세그멘테이션:

시멘틱 세그멘테이션은 이미지의 모든 픽셀에 클래스를 부여하는 라벨링 기법

물체, 배경 등을 의미적으로 분할하여 높은 정확도를 요구하는 자율 주행, 의료 영상 분석

기술 개발 등에 사용

데이터 라벨링 기법 적용 사례 1

프로젝트 목적

항공사진, 모사 영상을 이용한 산림 수종 분석 AI

데이터 유행

항공 촬영 이미지(JPG, PNG)

라벨링 객체

수종별 영역

활용 분야

국토환경, 산림 분포 변화 탐지 기술

데이터 라벨링 기법 적용 사례 2

프로젝트 목적

항공사진, 위성영상을 이용한 토지피복 분석 AI

데이터 유행

항공 촬영 이미지 (JPG, PNG)

라벨링 객체

건물, 주차장, 도로 등

활용 분야

국토환경, 산림 분포 변화 탐지 기술

데이터 라벨링 기법 적용 사례 3

프로젝트 목적

장애인 이동권 신장을 위한 인도 보행 AI 개발

데이터 유행

인도 보행 촬영 이미지(JPG, PNG)

라벨링 객체

도로, 보도, 점자블록 등

활용 분야

인도 자동 주행 서비스(휠체어, 배달 로봇 등)

데이터 라벨링 기법 적용 사례 4

프로젝트 목적

고정 노선 운행 대중교통 서비스 개발

데이터 유행

도로 주행 촬영 이미지(JPG, PNG)

라벨링 객체

전용 차로, 그 외 도로 등

활용 분야

자율주행 버스 기술 개발

## 9강. 10강. 데이터 라벨링 기법 폴리곤

1 폴리곤:

폴리곤은 다각형 모양으로 객체의 가시 영역 외곽선을 따라 점을 찍어 내는 라벨링 기법

객체 이외의 포함된 빈 공간으로 인해 발생하는 오류에 대응할 수 있는 기능

사물의 테두리를 따라 그리는 것을 통해 여백 없이 정확히 물체만을 인식하기 위해 사용  
(정교한 인공지능 모델 개발에 사용)

데이터 라벨링 기법 적용 사례 1

프로젝트 목적

한국인 헤어스타일 판별 및 이미지 합성 AI 개발

데이터 유형

헤어스타일 촬영 이미지(JPG, PNG)

라벨링 객체

얼굴, 헤어스타일

활용 분야

헤어스타일 합성 기술 개발

데이터 라벨링 기법 적용 사례 2

프로젝트 목적

월동작물별 재배면적 식별 및 생산량 예측 기술

데이터 유형

드론 촬영 이미지(JPG, PNG)

라벨링 객체

주요 관심 채소류 재배지

활용 분야

재배면적 식별 및 생산량 예측 등 농민 지원 기술

데이터 라벨링 기법 적용 사례 3

프로젝트 목적

축산물 품질관리 시스템 구축

데이터 유형

축산품 등급 분류 촬영 이미지(JPG, PNG)

라벨링 객체

품질 평가 영역

활용 분야

축산물 품질 평가 시스템

데이터 라벨링 기법 적용 사례 4

프로젝트 목적  
신호 및 표지판 파악  
데이터 유형  
도로 주행 촬영 이미지(JPG, PNG)  
라벨링 객체  
신호등, 도로 표지판  
활용 분야  
자율주행 실시간 영상 인식 기술 개발  
데이터 라벨링 기법 적용 사례 5  
프로젝트 목적  
교통약자 이동권 개선을 위한 객체 인식 AI 기술 개발  
데이터 유형  
도보 촬영 이미지(JPG, PNG)  
라벨링 객체  
연석, 턱, 점자 표지판 등  
활용 분야  
사회적 약자 전용 지도 구축 기술, 배달 로봇

## 11강. 12강. 데이터 라벨링 기법 키포인트 및 기타 라벨링 기법

### 1 키포인트:

키포인트는 특정 지점(특징점)을 라벨링 하는 방법  
안면 인식을 통한 감정 분석과 같이 정밀하고 섬세한 작업을 요구하는 라벨링 기법  
객체의 중요 특징점을 지정하여 물체를 추적하고 인식할 수 있는 라벨링 기법  
데이터 라벨링 기법 적용 사례 1  
프로젝트 목적  
반려동물 행동 분석 응용 서비스 개발  
데이터 유형  
반려동물(개, 고양이) 촬영 이미지(JPG, PNG)  
라벨링 객체  
반려동물의 특징점 15개  
활용 분야  
반려동물 관련 인공지능 기술 개발  
데이터 라벨링 기법 적용 사례 2  
프로젝트 목적  
운전자의 상태 모니터링  
데이터 유형  
운전자 안면 촬영 이미지(JPG, PNG)  
라벨링 객체  
얼굴 윤곽, 눈썹, 눈, 코, 입  
활용 분야

운전자 피로 감지, 안면 인식 기술

데이터 라벨링 기법 적용 사례 3

프로젝트 목적

버스 승객 행동 연구

데이터 유형

버스 CCTV 촬영 이미지(JPG, PNG)

라벨링 객체

신체의 특징점 15개

활용 분야

자율주행 버스의 승 하차, 안전 관리 기술 개발

데이터 라벨링 기법 적용 사례 4

프로젝트 목적

항만구조물, 해상 객체 식별을 통한 선교 상황 인식

데이터 유형

항만 구조물 촬영 이미지(JPG, PNG)

라벨링 객체

선박의 특징점 9개

활용 분야

자율운항 선박 기술, 무인 해양경비시스템

데이터 라벨링 기법 적용 사례 5

프로젝트 목적

홈트레이닝 서비스

데이터 유형

피트니스 촬영 이미지(JPG, PNG)

라벨링 객체

신체의 특징점 14개

활용 분야

홈트레이닝, 증강현실 서비스, 의료분야

2 큐보이드, 비디오 어노테이션, OCR 기법

1) 큐보이드:

큐보이드는 3차원 객체 탐지로 범위를 확장하면 물체의 크기, 위치, 방향 등을 알 수 있어 자율 주행 및 이미지 검색, 증강 현실 분야에서 많이 사용하는 라벨링 기법(3D Bounding Box)

객체의 너비, 높이 및 깊이, 방향 정보가 필요한 경우 사용

데이터 라벨링 기법 적용 사례 1

프로젝트 목적

실내 보행자의 연속 동선 추적

데이터 유형

실내 라이다-카메라 융합 촬영 이미지(JPG, PNG)

라벨링 객체

실내 보행자

활용 분야

보행자 인식/추적을 통한 매장 설계 및 마케팅

데이터 라벨링 기법 적용 사례 2

프로젝트 목적

자율주행 개발

데이터 유행

자동차 주행 라이다 센서 촬영 이미지(JPG, PNG)

라벨링 객체

차량

활용 분야

주율주행시스템 개발 , 3D 객체 인식 기술 개발

2) 비디오 어노테이션:

비디오 어노테이션은 영상에 대한 객체 움직임의 추적에 사용되는 라벨링 기법

구간 정제, 분류, 객체 태깅이 대표적인 방법이며, 객체 인식, 객체 추적에 주로 사용

데이터 라벨링 기법 적용 사례 1

프로젝트 목적

지능형 영상 인식 AI 개발

데이터 유행

방송사 영상(AVI, MP4)

라벨링 객체

영상 및 객체의 정보, 객체의 분류, 영상의 분류

활용 분야

영상 검색, OTT 플랫폼 분야 기술 개발

데이터 라벨링 기법 적용 사례 2

프로젝트 목적

공공 CCTV 영상을 이용한 이상행동 검출

데이터 유행

공공 CCTV 영상(AVI, MP4)

라벨링 객체

영상 구간 정제 및 분류

활용 분야

공공분야 지능형 CCTV 기술 개발

3) OCR:

광학 문자 인식이라는 뜻이다. OCR 기술은 스캔 된 문서나 이미지에 있는 문자를 컴퓨터가 인식하고 텍스트 데이터로 변환하는 과정

이 기술을 통해 종이 문서를 디지털 텍스트로 변환하거나, 이미지 속 텍스트 정보를 추출할 수 있다

데이터 라벨링 기법 적용 사례 1

프로젝트 목적

공공행정문서에 특화된 OCR 모델 개발

데이터 유행

공공행정문서 이미지(JPG, PNG)

라벨링 객체

의미 단위 글자 조합

활용 분야

행정업무 및 대국민 공개 서비스 OCR 기술 개발

데이터 라벨링 기법 적용 사례 2

프로젝트 목적

실내외 다양한 폰트의 한글 인식 OCR 기술 개발

데이터 유형

실내외 촬영 이미지(JPG, PNG)

라벨링 객체

의미 단위 글자 조합

활용 분야

모바일 OCR 및 야외 한글 인식 기술 개발

### 3. 음성데이터:

#### 1) 데이터의 정의

사람이 발화하는 음성 신호의 기록

주로 오디오 형식으로 저장되며, 음성인식, 음성 합성, 화자 인식 등과 같은 음성 관련 기술 개발과 훈련에 사용

#### 2) 데이터의 특성

획득 또는 수집된 음성 데이터는 청취 및 품질 평가를 거쳐 정제되고, 텍스트와의 대응 관계를 구축하기 위해 음성을 텍스트로 변호나 하는 작업이 수행

#### 3) 데이터의 구분

일반적으로 사람의 발화로 구성되며, 다양한 언어, 발음, 강세, 감정 등을 포함  
각 화자마다 발화 내용과 발화 스타일에 차이가 있을 수 있음

#### 4) 음성 텍스트 혼합 활용 사례

음성 인식: 음성 명령 인식, 음성 검색, 음성 메시징 등

화자 인식: 보안 스템, 음성 기반 액세스 제어, 사기 탐지

음성 합성: TTS, 음성 인터페이스, 자동 음성 안내 시스템, 음성 도움말

감정 분석: 고객 서비스, 감정 기반 마케팅, 교육 등

자연어 처리: 정보 검색, 질의응답, 시스템, 챗봇, 텍스트 마이닝 등

음성 번역: 실시간 음성 번역 시스템, 통번역 애플리케이션 등

### 4. 텍스트 데이터:

#### 1) 데이터의 정의

문자, 단어 또는 문장 등의 텍스트 형태로 구성된 데이터

주로 자연어 처리 알고리즘의 훈련과 평가에 사용

#### 2) 데이터의 구분

다양한 언어, 문체, 주제, 도메인을 포함

언어별 특징, 문법, 표현 방식, 단어의 다의성 등이 다양하게 나타남

### 5. 텍스트 데이터 활용 사례

텍스트 생성 및 자동화: 자동 요약, 자동 번역, 텍스트 생성 모델을 활용



정보 추출 및 분석: 문서 요약, 키워드 추출, 문서 분류, 토픽 모델링 등

소셜 미디어 분석: 사용자의 의견, 감성, 행동 패턴 파악, 제품 개선, 마케팅 전략 수립, 서비스 개선 등

기계 학습: 분류, 회귀, 군집화 등의 기계학습 작업

## 6. 언어 모델

### 1) 음성인식

음성으로부터 언어적 의미 내용을 식별하는 것.

음성 파형을 입력하여 단어나 단어열을 식별하고 의미를 추출하는 처리 과정

### 2) 음향 모델

음성 신호를 분석하고 음소 단위로 변환하는 모델

음성인식에서 음성 신호를 효과적으로 인식 가능한 텍스트로 변환하는 역할을 함

### 3) 언어 모델

문장이나 문서의 확률을 예측하는 모델

음성인식에서 음성 신호를 텍스트로 변환하는 과정에서 문맥과 언어의 확률적인 모델을 제공

## 7. 데이터 구축 관련 주요 용어

1) 전사: 말소리를 음성 문자로 옮겨 적음(국립국어원), 어떤 언어의 음운 요소와 형태소를 특정 기록 시스템의 특유한 표현으로 기록하는 과정(TTA 표준)

2) 비식별 조치: 정보 집합물(데이터 셋)에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법으로 개인을 알아볼 수 없도록 하는 조치

3) 품질 검사: 인공지능 학습용 데이터 구축 및 모델 관점에서 주요 품질 문제를 식별하고 문제의 근원적 원인을 파악하여 개선 기회를 도출하는 기법을 의미

## 8. 음성/텍스트 주요 라벨링 기법 - 기본 과정 학습 내용

일반 전사: 이중 음성 데이터를 컴퓨터가 이해할 수 있는 텍스트 형식으로 변환하는 작업

이중 전사: 비표준적 발음에 대해서 발음 전사와 철자 전사를 병행 표기

삼중 전사: 숫자나 영문 발음에 대해서 발음 전사와 철자 전사를 병행 표기 등 매뉴얼에서 지정한 주요 키워드를 발음했거나 방언 등의 비표준적

감정 태깅: 발화자가 내포하고 있는 감정을 파악하여 가이드라인에 맞게 포괄적인 감정으로 태깅

개체명 태깅: 주어진 텍스트에서 개체명을 찾아내고 각 개체명을 특정 태그로 표시하는 작업

## 9. 음성/텍스트 주요 라벨링 기법 - 심화 과정 학습 내용

의도 태깅: 문장별 발화자가 내포하고 있는 의도를 파악하여 해당하는 항목으로 태깅

감정 태깅: 발화자가 내포하고 있는 감정을 파악하여 지침에 맞게 포괄적 감정으로 태깅

비식별화: 전사 시 개인 정보 노출 방지를 위해 이름, 주민등록번호, 카드번호, 전화번호 등 개인 정보와 관련된 사항은 노출되지 않도록 별도 처리

텍스트 요약: 원본 텍스트를 읽고 핵심 내용을 파악하여 요약문을 작성하거나 추출

이미지 설명: 주어진 이미지에 대해 적절한 문장을 생성

문장 생성: 제시된 개념 정보를 재구성하여 일반 상식에 부합하는 자연스러운 한국어 문장을 제작

문장 평가: 기계가 생성한 문장과 인간이 생성한 문장을 대상으로 기준에 따라 점수 부여

## 10. 학습 데이터

### 1) 학습 데이터 순서

이렇게 생성된 학습 데이터는 음성/텍스트 데이터와 JSON 데이터 파일로 생성

## 2) 학습 데이터 가공 사례

방송 콘텐츠 대화체 음성인식 데이터

주요 영역별 회의 음성인식 데이터

한국어 방언 발화

문학작품 낭송, 낭독 음성 데이터

온라인 구어체 말뭉치 데이터

대규모 웹 데이터 기반 한국어 말뭉치

## 3) 주요 가공 방법:

이중 전사(숫자, 방언, 영문)

### 11. 품질 관리 기준 및 검사 방안

#### 1) 품질관리 측면

인공지능 학습용 데이터의 전 생애 주기 품질을 보장해야 함

상시적이고 지속적인 인공지능 학습용 데이터 품질 개선이 가능해야 함

데이터의 품질관리를 위한 조직을 구성하고 정해진 역할과 책임에 따라 수행해야 함

조직의 품질관리 역량을 확보하도록 품질관리 교육 등 지원체계를 확보해야 함

#### 2) 데이터 측면

인공지능 학습용 데이터가 학습하는 데 필요한 요구사항을 충족해야 함

법률의 제약 없이 누구나 활용 가능해야 함

학습의 목적에 부합하도록 획득/ 수집해야 하며, 획득 수집한 데이터는 중복 없이 원하는 목적에 따라 정제되어야 함

인공지능 학습용 데이터에 부합하는 참값(Ground Truth) 등의 라벨링 정보의 정확성이 확보되어야 함

인공지능 학습 모델을 통해 목표하는 유효성이 확보되어야 함

#### 3) 구축 데이터 품질관리

구축 사업을 통해 생성되는 원시데이터, 원천 데이터, 라벨링 데이터 등의 데이터 자체의 품질을 검사하고, 발견된 오류를 개선하는 활동

#### 4) 데이터 정확성 품질관리 지표

(1) 구문 정확성 - 데이터구조, 입력값 범위, 데이터 형식

(2) 의미 정확성 - 정확도, 정밀도, 재현율 등

#### 5) 의미 정확성 검사 대상

텍스트: 내용요약, 번역, 질의응답, 말뭉치 태깅 데이터

음성: 전사 데이터

### 12. 데이터 라벨링의 발전

#### 1) 초기 라벨링 기법(~2019년)

수동적인 라벨링

데이터의 통일성 부재

시스템 '훈련' 방식으로 응용성 부족

소모적이고 고비용의 작업

전용 툴(Tool)의 부재로 라벨링 효율성이 떨어짐

#### 2) 라벨링 기법의 발전(2019년~)

정부에서 주도적으로 AI 학습 데이터 구축 주도

한국전자통신연구원(ETRI)이 음성 데이터의 표준화를 위해 전사 규칙 제정

음성인식 분야의 연구 활성화

컴퓨팅 파워의 증가 및 딥러닝

알고리즘의 발전

### 3) 현재 라벨링 기법(2023)

수동 라벨링에서 벗어나 자동음성인식(ASR)을 활용한 자동

라벨링 방식 전용 툴(Tool)의 발전으로 작업

효율성 향상 모델의 발전으로 음성/텍스트가 아닌 영상/음성/텍스트/이미지 등 멀티 모달

라벨링으로의 영역 확대

### 4) 개체명 및 신조어 분류

제시된 가이드라인에 따라 대분류, 세부 분류 등으로 분류할 수 있지만 이번 실습에서는

개체명 및 신조어 7개의 항목으로 분류

#### (1) 개체명 기준

개체명은 기본적으로 어절(단어) 단위로 인식

단, 각각의 태그 범주와 태깅 원칙에 따라 복합 명사를 허용

개체명은 원칙적으로 고유명사를 대상

#### (2) 신조어 기준

표준국어대사전과 우리말샘에 없는 단어 대상

#### (3) 태그 분류 기준

TTA 표준인 개체명 태그 세트 및 태깅