






데이터 자동 레이블링

성명 김선태

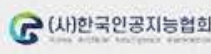
소속 ETRI

SUBJECT

인공지능 기술의 대중화 (AI Democratization)를 위한
TANGO 커뮤니티 3회 컨퍼런스

주관 ETRI () 주최  과학기술정보통신부  정보통신기획평가원

후원



목 차

1

데이터 레이블링 개요

00

1. 필요성
2. 개념도

2

데이터 자동 레이블링 기술

00

1. Self-Supervised Learning
2. Active Learning
3. AL 동향
4. 시험 결과

3

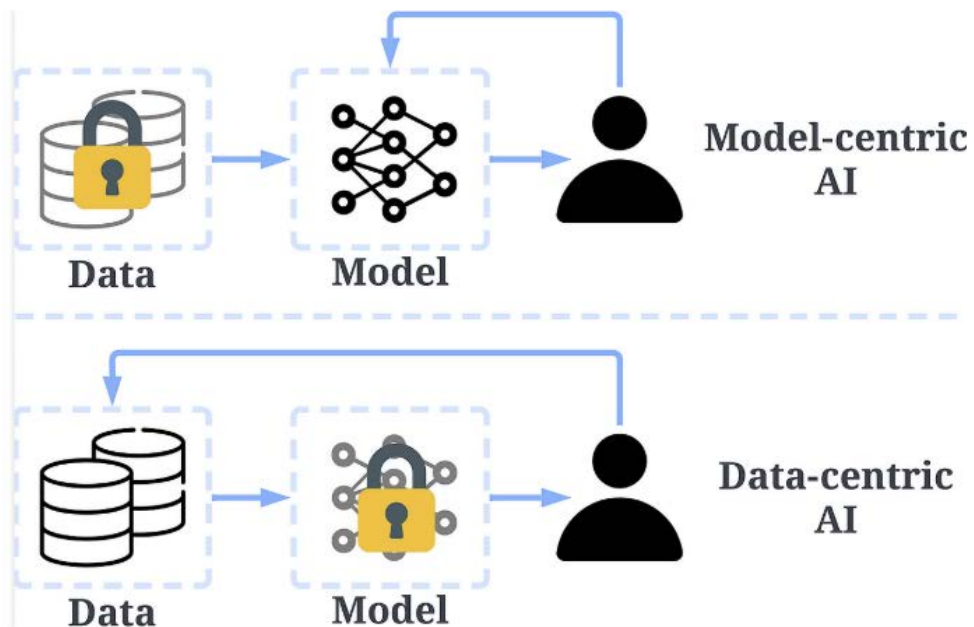
향후 개발 내용

00

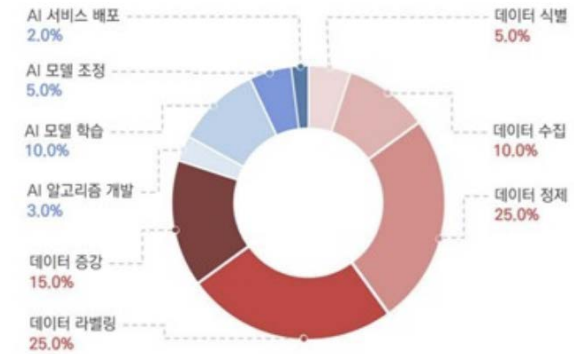
1. TANGO 적용 – 산업데이터 적용
2. CI/CD 및 Incremental Learning

필요성

- ✓ AI 응용개발 시간의 80%가 데이터 처리에 소요
- ✓ 데이터 처리에서도 데이터 라벨링이 가장 많은 시간 소요
- ✓ Data-Centric AI로의 전환

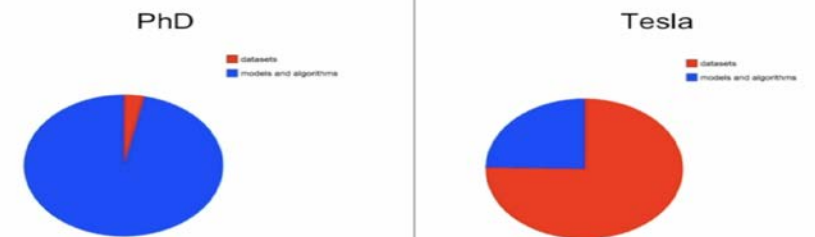


출처: <https://blog.ex-em.com/1962>



AI 프로젝트에 소요되는 시간 비율

Amount of lost sleep over...

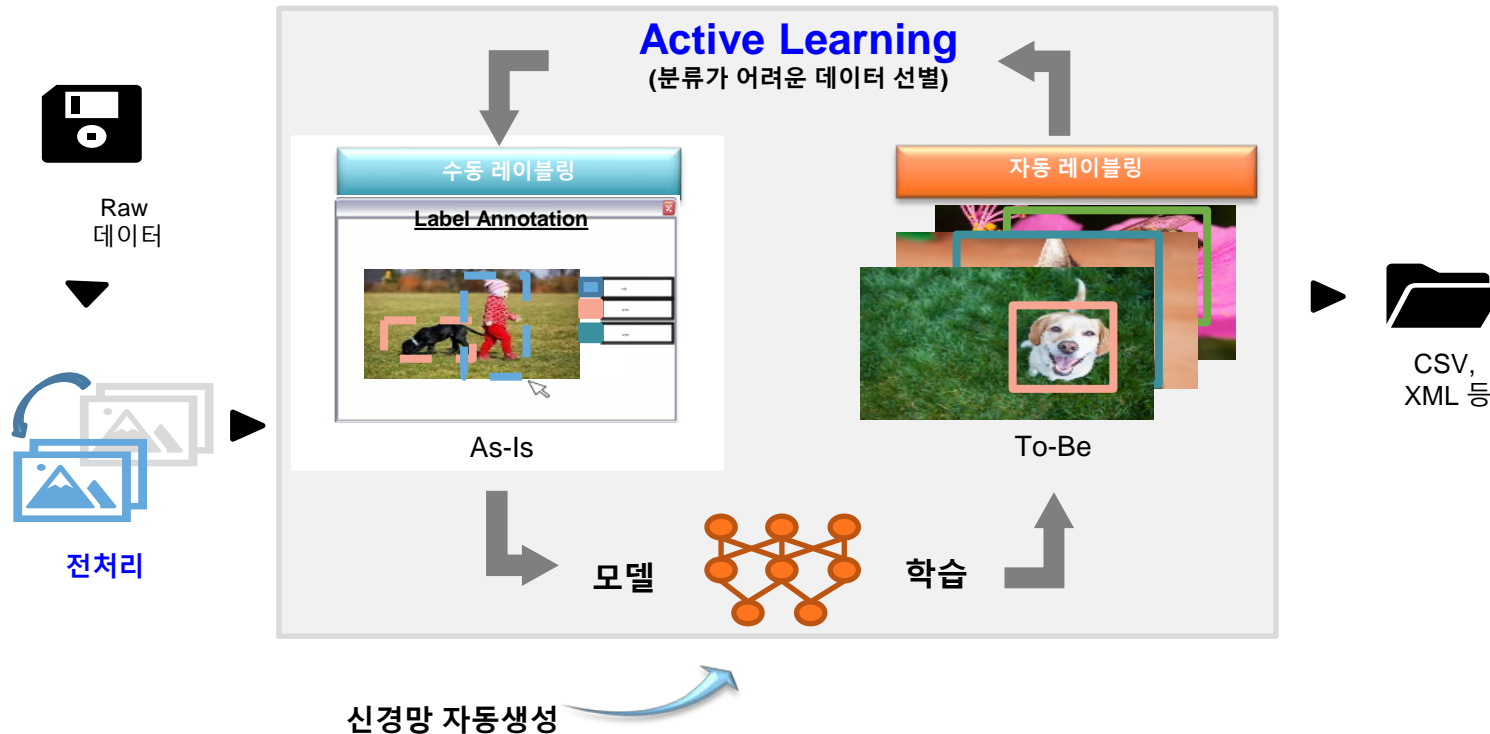


출처: Andrej Karpathy, Tesla Director of AI (2021)

I. 데이터 레이블링 개요

데이터 레이블링 개념도

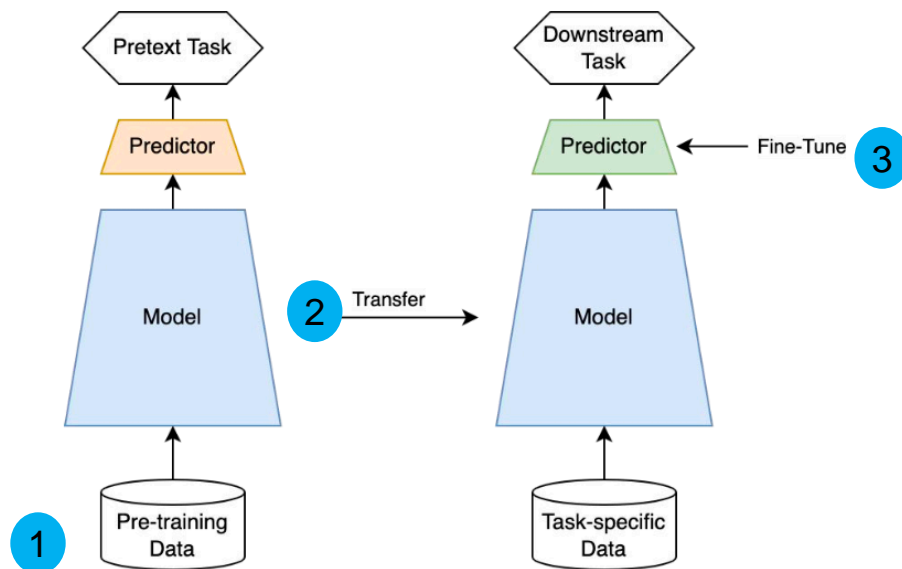
- 산업 분야 데이터를 신경망에서 학습 가능하게 전처리하고, 방대한 데이터에서 소량의 학습 데이터를 추출하는 기술
- 소량의 레이블된 데이터를 기반으로 정확도를 높이기 위해서는 데이터에서 대표성있는 샘플을 추출해야 함
- 자기주도 학습기반 데이터 분류 및 **데이터 레이블링 기법이 추가되는 Active Learning이 대표적임**



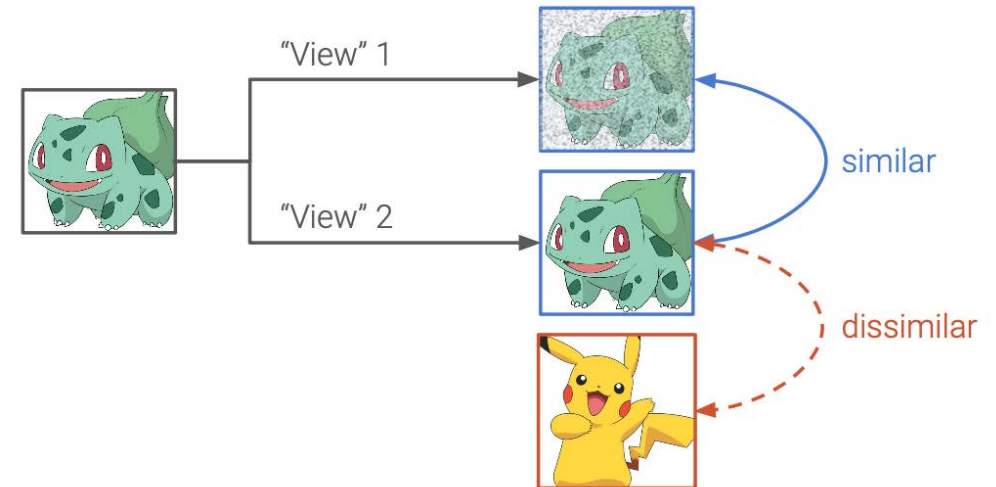
II. 데이터 자동 레이블링 기술

Self-Supervised Learning

- 방대한 규모의 데이터를 이용하여 Siamis 네트워크를 통해 유사 데이터별로 분류하는 기법
- 3단계의 학습을 통해 데이터 적응형 신경망을 개발하는 과정 필요
- 유사샘플과 비유사샘플 간의 밀당을 통한 **Contrastive Learning**이 대표적임



데이터 분류를 위한 신경망 생성 절차

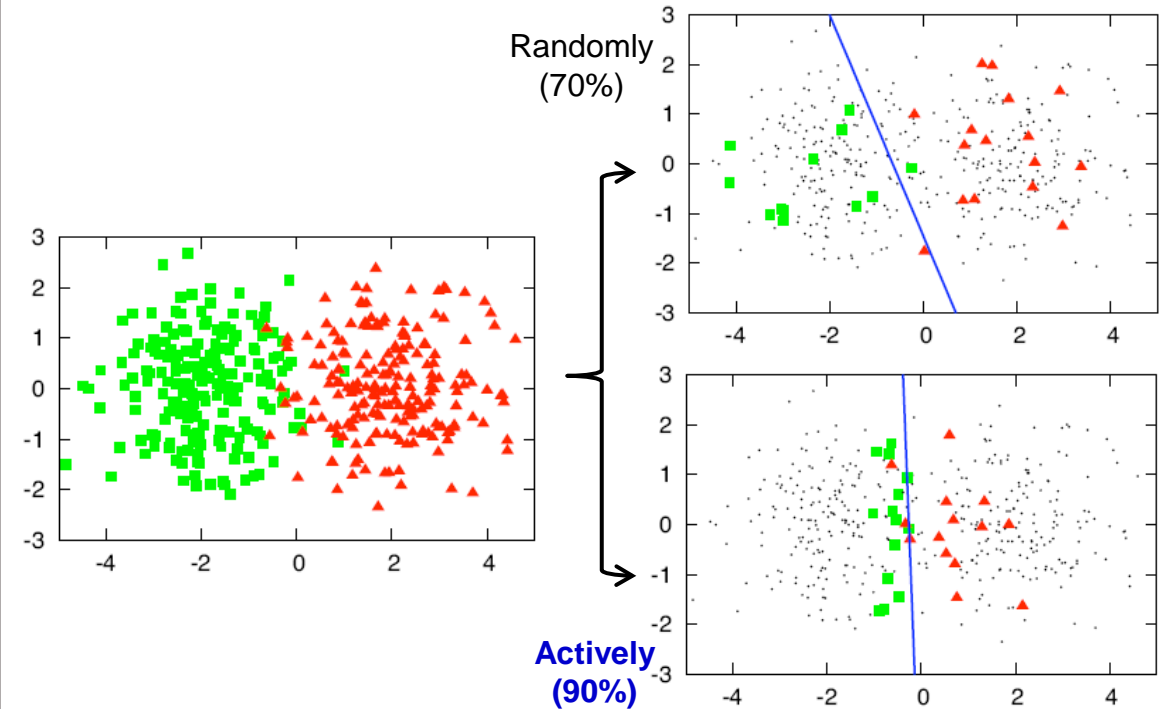
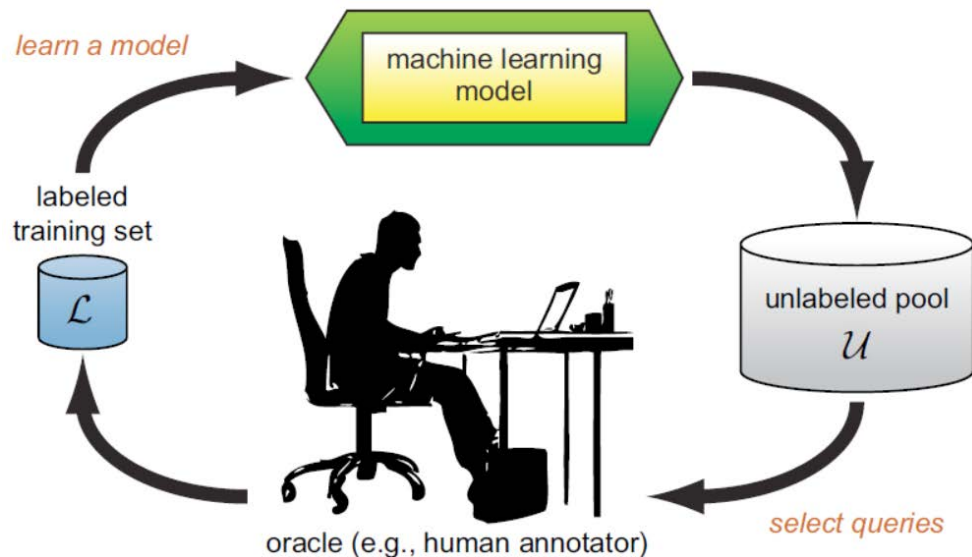


대조 학습을 통한 데이터 분류 방식

Uncertainty and Diversity 기반 Active Learning

- 대규모 데이터에서 대표되는 샘플을 추출하여 레이블링한 후 지도학습을 수행하는 방법
- 샘플간 분류가 어려운 데이터를 추출(Uncertainty)하여 분류 기준의 모호성을 제거하는 방법
- 분류 클래스간 샘플의 불균형을 해소(Diversity)하여 클래스간 정확도 균형 유지하는 방법

다양한 데이터 셋을 보다 정확하고 빠르게 분류하여 라벨링 할 수 있는
모델 기술 개발

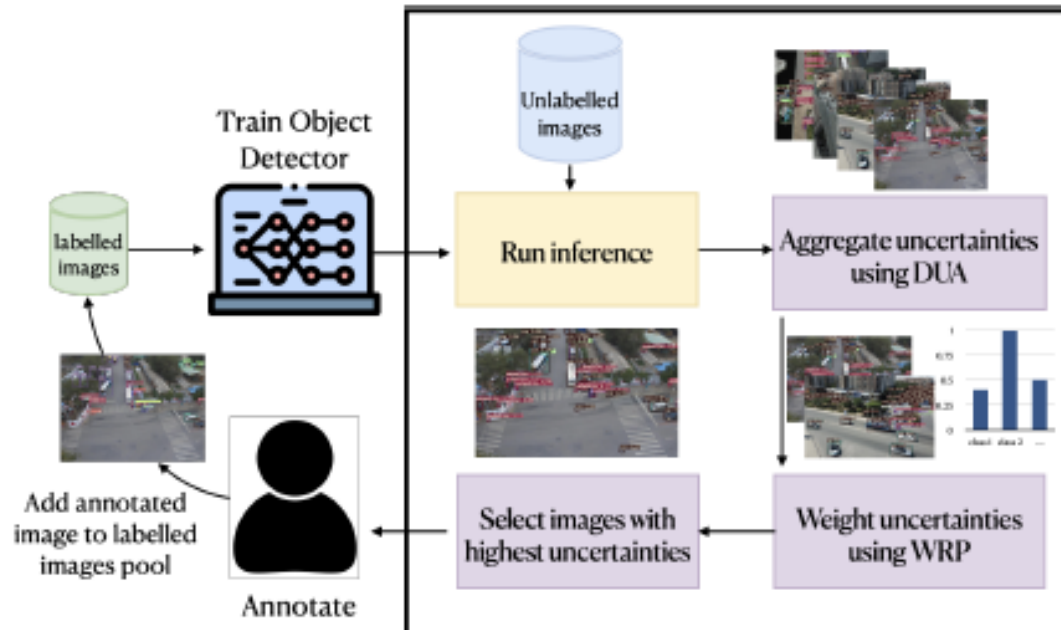


II. 데이터 자동 레이블링 기술

7

Drone Data기반 Active Learning

- Single-stage Object Detector를 이용하여 UAV 데이터에 적용한 사례
- VisDrone :클래스는 10개이나, 클래스간 유사성이 높고 객체가 작아서 식별이 어려운 데이터세트
- DUA(Diversity Uncertainty Aggregation) 기법 적용
- 25% 이미지 및 32% 객체 레이블링으로 전체 데이터 세트 사용한 경우와 유사한 성능 확보



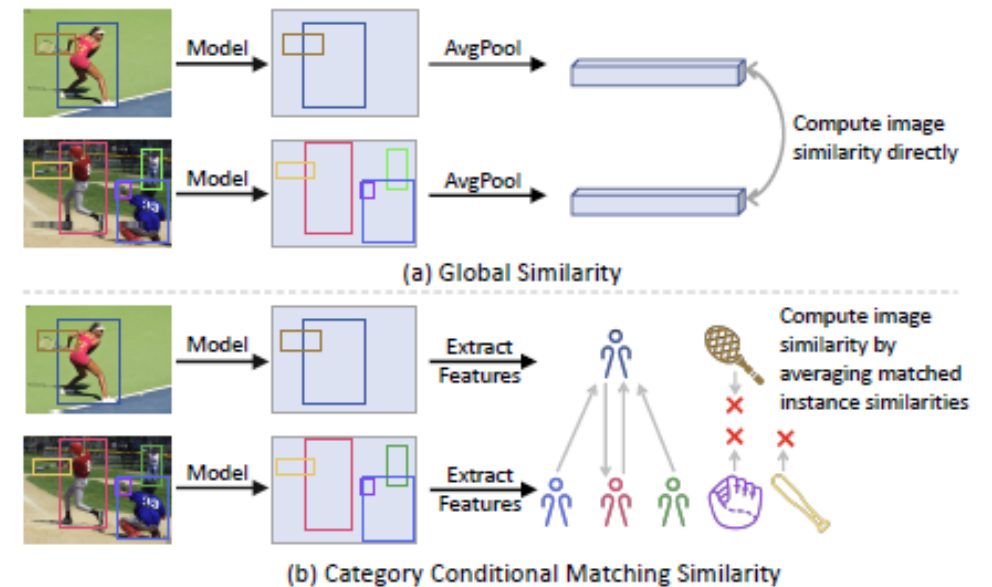
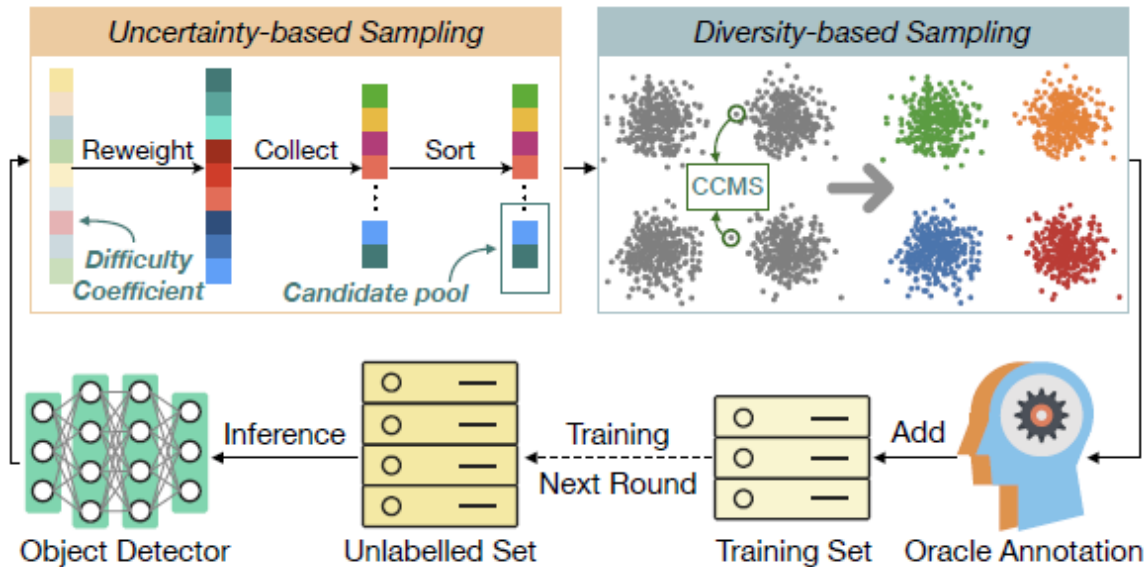
VisDrone Data Set

Class	Whole	Random	Sum	Avg	DUA
Images	4279	1100	1100	1100	1100
Objects	226k	57.3k	95.8k	63.4k	72.7k
All	0.313	0.288	0.311	0.27	0.317
Pedestrian	0.341	0.330	0.351	0.325	0.343
People	0.311	0.305	0.332	0.305	0.333
Bicycle	0.066	0.042	0.061	0.026	0.066
Car	0.726	0.716	0.727	0.684	0.730
Van	0.315	0.279	0.311	0.234	0.309
Truck	0.284	0.262	0.253	0.239	0.294
Tricycle	0.174	0.122	0.191	0.133	0.188
Awning-tricycle	0.086	0.066	0.080	0.070	0.092
Bus	0.438	0.415	0.411	0.336	0.418
Motor	0.384	0.346	0.388	0.353	0.397

출처: [2024 WACV] Active Learning for Single-Stage Object Detection in UAV Images

Plug&Play Active Learning

- 두 단계의 순차적 적용으로 객체 탐지를 위한 AI 단순화 및 모듈화
- Uncertainty 단계: 객체별로 클래스와 BB 확률을 이용해 복잡도를 계산 -> 이미지별로 복잡도 계산하여 순위 결정
- Diversity 단계: 이미지간 CCMS(Category Conditioned Matching Similarity) 기법 이용

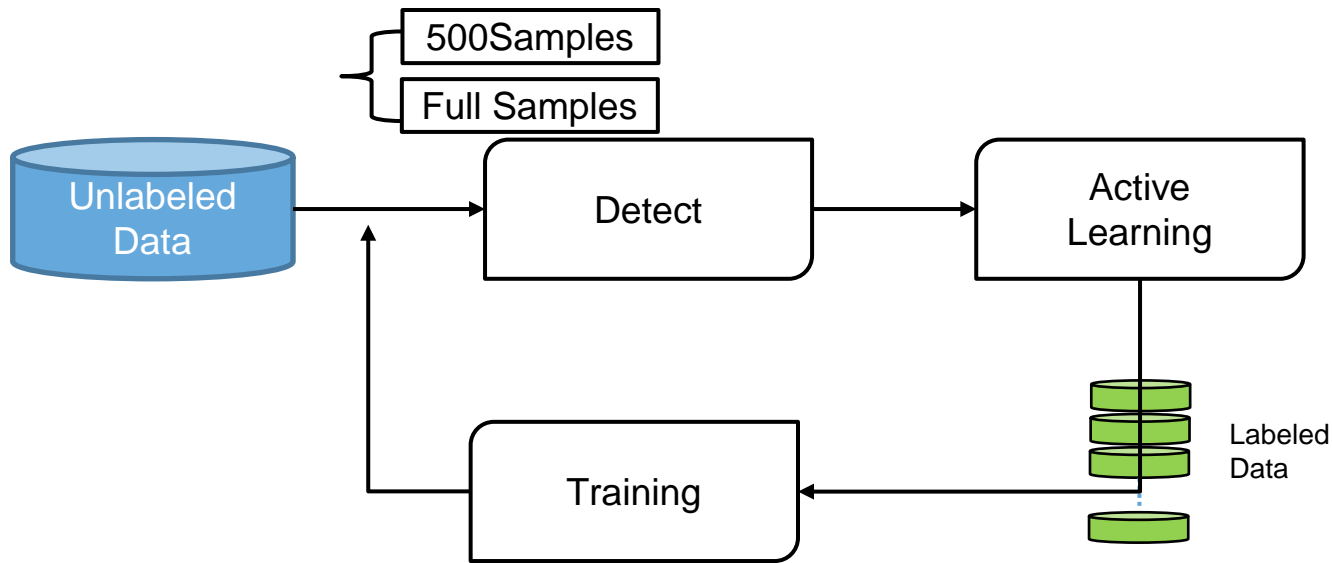


출처: [2024 CVPR] Plug and Play Active Learning for Object Detection

II. 데이터 자동 레이블링 기술

Active Learning 알고리즘 시험

- 클래스 확률과 IoU 확률 값을 이용한 Uncertainty 계산 및 오름차순 정리
- 50%는 Uncertainty 값 이용 추출, 50% Diversity (이미지내 클래스 수)이용하여 이미지 추출
- 시험 데이터 중 객체가 적은 클래스에서 향상됨. 전체적으로 27.5%에서 28.9%로 높아짐

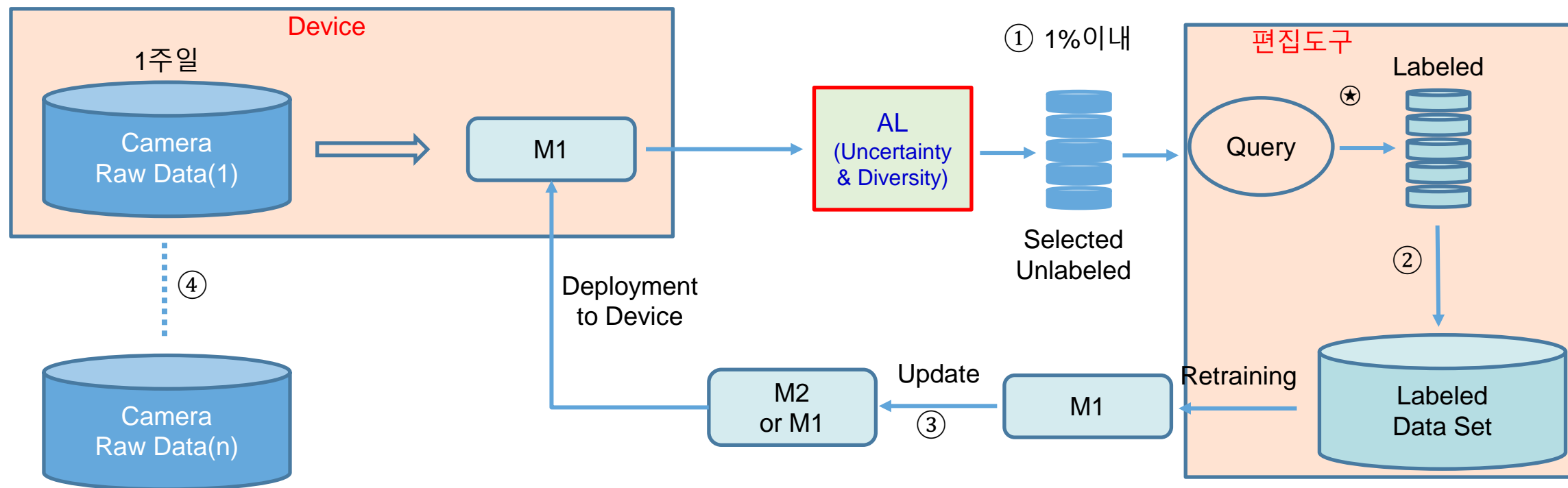


	Samples	target	mAP@0.5	
			DUA	Ours
all	1610	75102	0.275	0.289
pedestrian	1610	21006	0.302	0.318
people	1610	6376	0.196	0.213
bicycle	1610	1302	0.0998	0.0969
car	1610	28074	0.706	0.715
van	1610	5771	0.26	0.269
truck	1610	2659	0.236	0.273
tricycle	1610	530	0.143	0.169
awning-tri cycle	1610	599	0.0919	0.115
bus	1610	2940	0.445	0.435
motor	1610	5845	0.266	0.289

III. 향후 개발 계획

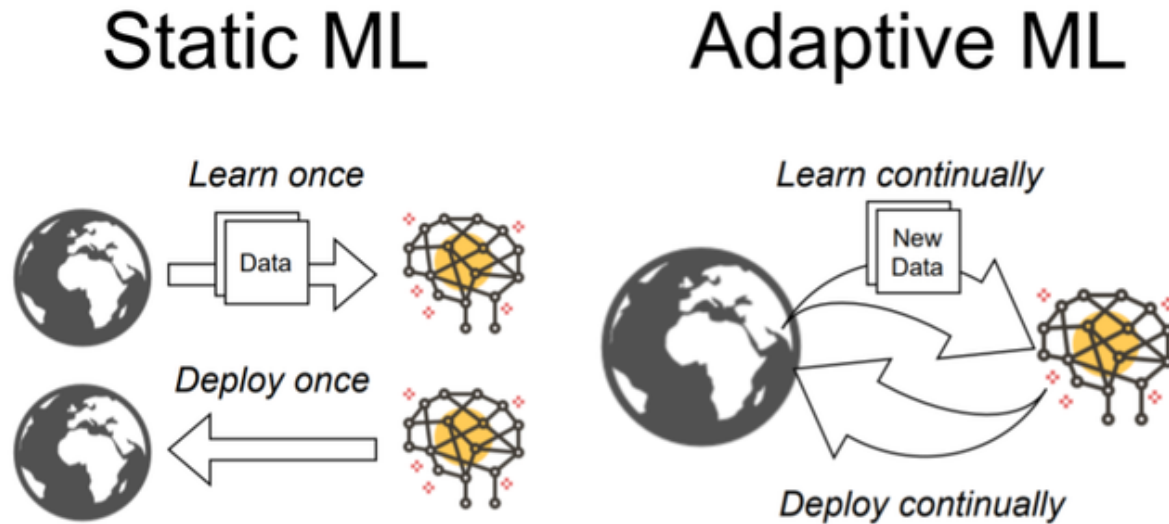
산업데이터 기반 데이터 레이블링 기술

- 디바이스에서 **주기적으로 생산되는 데이터에 대해서 1%의 데이터 레이블링**으로 정확도 향상
- 편집도구에서 탐지된 객체에 대해서 **정확한 레이블링 작업 연결**
- 레이블된 데이터를 가지고 학습을 수행하여 **신경망을 업데이트함**
- 추가되는 클래스에 대응 가능하며 **와 다양한 데이터로 인한 정확도 향상 가능**



Incremental Learning을 통한 CI/CD

- 주기적인 입력 데이터를 가지고 신경망 정확도 향상에 적용할 필요성 존재
- 주기적으로 신경망을 자가학습하고 업데이트하며, 이를 디바이스에 배포하는 MLOps 완성
- Data-Centric AI 시스템 구축



감사합니다.

주관

ETRI
한국전자통신연구원

(TANGO)

주최



과학기술정보통신부

IITP

정보통신기획평가원

후원

labup

we-a

tesla
system

(사)한국인공지능협회
Korea Artificial Intelligence Association

SNUH
서울대학교병원

고려대학교
KOREA UNIVERSITY

홍익대학교
HONGIK UNIVERSITY

cau 중앙대학교
CHA UNIVERSITY