[붙임1] 2026년 캡스톤디자인 우수작품 경진대회 참가 신청서

2026년 전남대학교 소프트웨어중심대학사업
캡스톤디자인 우수작품 경진대회 참가 신청서

과제명	DDRS : Data Distributed with Rolling Storage(분산 처리 기반 지식증류 파이프라인)					
팀명	2gether					
지도교수	성명	김형일		소속학과	전자컴퓨터공학부	
	연락처	010-3975-0920		이메일	hyungil.kim@chonnam.ac.kr	
대표학생	성명	강민수	학번/학년	2146914	소속학과	컴퓨터정보통신공학과
	연락처	010-9137-1632		이메일	214691@jnu.ac.kr	
참여학생	성명	장인환	학번/학년	214683/4	소속학과	컴퓨터정보통신공학과
	연락처	010-4519-5336		이메일	214683@jnu.ac.kr	
참여학생	성명		학번/학년		소속학과	
	연락처			이메일		
참여학생	성명		학번/학년		소속학과	
	연락처			이메일		
참여학생	성명		학번/학년		소속학과	
	연락처			이메일		
참여학생	성명		학번/학년		소속학과	
	연락처			이메일		
과제 설명 (요약)	라벨이 없는 대규모 데이터를 제한된 Edge 환경에서도 효율적으로 학습하는 DDRS 시스템을 제안한다. Kafka 기반 데이터 스트리밍, HDFS 분산 저장, Shard 단위 Rolling Storage 순환 학습, Teacher-Student Knowledge Distillation을 결합하여 저장공간과 연산 자원의 제약을 동시에 해소한다.					

상기와 같이 전남대학교 소프트웨어중심대학사업단
2026년 캡스톤디자인 우수작품 경진대회 참가 신청서를 제출합니다.

2026년 6월 23일

신 청 자 : 강민수 (인)

전남대학교 SW중심대학사업단장 귀하

[붙임2] 개인정보 및 과세정보 수집·이용·제공 동의서

개인정보 및 과세정보 수집·이용·제공 동의서

본인은 전남대학교 소프트웨어중심대학사업단에서 진행하는 프로그램 운영과 관련하여 「개인정보보호법」 제15조, 제17조, 제22조 및 제24조, 「국세기본법」 제81조의 13 제1항 제7호에 따라 아래와 같이 본인의 개인정보를 수집·이용·제공하는 것에 동의합니다.

아래 사항을 충분히 읽어 보신 후, 동의하시는 경우 서명하여 주시기 바랍니다.

개인정보 수집 및 이용에 대한 동의

○ **개인정보 및 과세정보 수집·이용 목적**

✓ 참여제한, 채무불이행 정보 등 신용조회 및 기타 사전지원제외, 사후관리 대상 여부의 확인
✓ 평가위원 선정 시 평가대상과제와의 이해관계 (참여연구원 등) 여부의 확인
✓ 과제 선정, 보고서 제출, 기술료 납부, 협약 및 협약변경 등 과제의 선정·평가 및 관리
✓ 총괄책임자와 참여연구원의 연구비 사용·정산 및 과제 수행의 적법·적정성 평가를 위한 관리
✓ 만족도 조사, 사업 및 경영활동 안내 등 사후관리
✓ 소프트웨어중심대학사업 프로그램 운영 및 이수관리

○ **수집하는 개인정보 및 과세정보 항목**

✓ 개인 성명, 근무기관, 주소, 전화번호, 전자우편, 학력(학교, 전공, 학위, 연구분야 등), 경력, 특허/논문 실적, 정부출연사업 수행실적, 현재 수행중인 정부출연사업 전체 참여율, 지급기준 정보(연봉; 월 수령가능금액 등), 연구비 지출을 위한 신용카드 및 금융거래 내역, 국가연구자번호, 채무불이행 정보 등 재무건전성 여부를 확인하기 위한 신용정보 등 인적사항, 「국세기본법」 제81조의 13의 과세정보(연구비 심사에 필요한 과세정보에 한함), 소속, 주소, 성명, 전화번호, 메일주소, 영상물(사진 등)

○ **개인정보 및 과세정보 보유·이용 기간: 동의서가 작성된 시점부터 상기 개인정보 및 과세정보 수집·이용 목적이 종료되는 시점**

○ **관련 근거 :** 국가연구개발혁신법 제19조, 및 동법 시행령 제42조, 국가연구개발정보처리기준, 국가연구개발사업 연구개발비 사용 기준, 정보통신·방송 연구개발 관리규정 제8조, 제20조, 제21조, 제24조, 제35조, 제48조 등

※ 개인정보 수집·이용에 동의하십니까? (■ 동의함 □ 동의하지 않음)

고유식별정보 처리 동의

○ **고유식별정보 처리 목적**

✓ 참여제한, 채무불이행 정보 등 신용조회 및 기타 사전지원제외, 사후관리 대상 여부의 확인
✓ 평가위원 선정 시 평가대상과제와의 이해관계 (참여연구원 등) 여부의 확인

○ **처리하는 고유식별정보 항목 :** 국가연구자번호

○ **고유식별정보 보유·이용 기간 : 동의서가 작성된 시점부터 상기 개인정보 수집·이용 목적이 종료되는 시점까지**

○ **관련 근거 :** 국가연구개발혁신법 제19조 및 동법 시행령 제42조, 국가연구개발정보처리기준, 정보통신·방송 연구개발 관리규정 제8조, 제20조, 제21조, 제24조

※ 고유식별정보에 동의하십니까? (■ 동의함 □ 동의하지 않음)

민감정보 수집·이용 동의

○ **민감정보 수집·이용목적**

✓ 소프트웨어중심대학사업 프로그램 활동비, 전문가활용비, 인부사역비, 보험가입 등의 비용 지급

○ **처리하는 고유식별정보 항목 :** 주민번호, 계좌번호, 재학증명서

○ **고유식별정보 보유·이용 기간 : 동의서가 작성된 시점부터 상기 개인정보 수집·이용 목적이 종료되는 시점까지**

○ **관련 근거 :** 개인정보보호법 제15조, 제17조, 제22조 및 제24조 등

※ 민감정보 수집·이용 에 동의하십니까? (■ 동의함 □ 동의하지 않음)

개인정보의 제3자 제공에 대한 동의

○ **개인정보의 제3자 제공 목적**

✓ 국가연구개발사업 참여제한 여부 확인 및 채무불이행 정보 등 신용조회
✓ 국정감사시 과제수행현황(참여연구자) 등 요구자료 대응
✓ 정보통신·방송 연구개발 사업 관련 타 전문기관의 동일업무 수행
✓ 기획재정부, 과학기술정보통신부 주관 고객만족도 조사
✓ 소프트웨어중심대학사업단 프로그램 활동 증빙

○ **개인정보를 제공받는 자:** 과학기술정보통신부, 국회 등 정부기관, 한국연구재단 부설 정보통신기획평가원 등 정보통신·방송 연구개발사업의 전문기관, 범부처 연구비통합관리시스템(통합이지바로), 국가과학기술종합정보시스템(NTIS), 한국기업데이터 주식회사, 한국정보통신기술협회, 기획재정부 및 과학기술정보통신부가 선정한 고객만족도 주간사(수행기관)

○ **개인정보를 제공받는 자의 이용목적:** ①국가연구개발사업 참여의 적법성 판단, ②과제수행에 대한 적법·적정성 판단, ③과제 선정·평가·관리 업무 수행

○ **제공하는 개인정보 항목 :** 성명, 근무기관, 국가연구자번호, 주소, 연락처, 이메일, 소속, 영상물(사진 등)등

○ **개인정보를 제공받는자의 개인정보 보유·이용 기간: 동의서가 작성된 시점부터 상기 개인정보 제3자 제공목적 달성시까지**

○ **관련 근거:** 국가연구개발혁신법 제19조 및 동법 시행령 제42조, 국가연구개발정보처리기준, 정보통신·방송 연구개발 관리규정 제12조

※ 개인정보의 제3자 제공에 동의하십니까? (■ 동의함 □ 동의하지 않음)

※ 유의사항 : 귀하는 상기 동의를 거부할 수 있습니다. 해당 수집 항목은 정보통신·방송 연구개발 수행에 반드시 필요한 사항으로 이에 대한 동의를 하지 않을 경우에는 정보통신·방송 연구개발 참여 등에 제한을 받으실 수 있습니다.

소속학과: 컴퓨터정보통신공학과 학번: 214691 성명: 강민수 (서명)

[붙임2] 개인정보 및 과세정보 수집・이용・제공 동의서

개인정보 및 과세정보 수집・이용・제공 동의서

본인은 전남대학교 소프트웨어중심대학사업단에서 진행하는 프로그램 운영과 관련하여 「개인정보보호법」 제15조, 제17조, 제22조 및 제24조, 「국세기본법」 제81조의 13 제1항 제7호에 따라 아래와 같이 본인의 개인정보를 수집·이용·제공하는 것에 동의합니다.

아래 사항을 충분히 읽어 보신 후, 동의하시는 경우 서명하여 주시기 바랍니다.

개인정보 수집 및 이용에 대한 동의

○ **개인정보 및 과세정보 수집・이용 목적**

✓ 참여제한, 채무불이행 정보 등 신용조회 및 기타 사전지원제외, 사후관리 대상 여부의 확인
✓ 평가위원 선정 시 평가대상과제와의 이해관계 (참여연구원 등) 여부의 확인
✓ 과제 선정, 보고서 제출, 기술료 납부, 협약 및 협약변경 등 과제의 선정·평가 및 관리
✓ 총괄책임자와 참여연구원의 연구비 사용·정산 및 과제 수행의 적법·적정성 평가를 위한 관리
✓ 만족도 조사, 사업 및 경영활동 안내 등 사후관리
✓ 소프트웨어중심대학사업 프로그램 운영 및 이수관리

○ **수집하는 개인정보 및 과세정보 항목**

✓ 개인 성명, 근무기관, 주소, 전화번호, 전자우편, 학력(학교, 전공, 학위, 연구분야 등), 경력, 특허/논문 실적, 정부출연사업 수행실적, 현재 수행중인 정부출연사업 전체 참여율, 지급기준 정보(연봉; 월 수령가능금액 등), 연구비 지출을 위한 신용카드 및 금융거래 내역, 국가연구자번호, 채무불이행 정보 등 재무건전성 여부를 확인하기 위한 신용정보 등 인적사항, 「국세기본법」 제81조의 13의 과세정보(연구비 심사에 필요한 과세정보에 한함), 소속, 주소, 성명, 전화번호, 메일주소, 영상물(사진 등)

○ **개인정보 및 과세정보 보유・이용 기간: 동의서가 작성된 시점부터 상기 개인정보 및 과세정보 수집·이용 목적이 종료되는 시점**

○ **관련 근거 :** 국가연구개발혁신법 제19조, 및 동법 시행령 제42조, 국가연구개발정보처리기준, 국가연구개발사업 연구개발비 사용 기준, 정보통신·방송 연구개발 관리규정 제8조, 제20조, 제21조, 제24조, 제35조, 제48조 등

※ 개인정보 수집·이용에 동의하십니까? (■ 동의함 □ 동의하지 않음)

고유식별정보 처리 동의

○ **고유식별정보 처리 목적**

✓ 참여제한, 채무불이행 정보 등 신용조회 및 기타 사전지원제외, 사후관리 대상 여부의 확인
✓ 평가위원 선정 시 평가대상과제와의 이해관계 (참여연구원 등) 여부의 확인

○ **처리하는 고유식별정보 항목 :** 국가연구자번호

○ **고유식별정보 보유・이용 기간 : 동의서가 작성된 시점부터 상기 개인정보 수집·이용 목적이 종료되는 시점까지**

○ **관련 근거 :** 국가연구개발혁신법 제19조 및 동법 시행령 제42조, 국가연구개발정보처리기준, 정보통신·방송 연구개발 관리규정 제8조, 제20조, 제21조, 제24조

※ 고유식별정보에 동의하십니까? (■ 동의함 □ 동의하지 않음)

민감정보 수집・이용 동의

○ **민감정보 수집·이용목적**

✓ 소프트웨어중심대학사업 프로그램 활동비, 전문가활용비, 인부사역비, 보험가입 등의 비용 지급

○ **처리하는 고유식별정보 항목 :** 주민번호, 계좌번호, 재학증명서

○ **고유식별정보 보유・이용 기간 : 동의서가 작성된 시점부터 상기 개인정보 수집·이용 목적이 종료되는 시점까지**

○ **관련 근거 :** 개인정보보호법 제15조, 제17조, 제22조 및 제24조 등

※ 민감정보 수집・이용 에 동의하십니까? (■ 동의함 □ 동의하지 않음)

개인정보의 제3자 제공에 대한 동의

○ **개인정보의 제3자 제공 목적**

✓ 국가연구개발사업 참여제한 여부 확인 및 채무불이행 정보 등 신용조회
✓ 국정감사시 과제수행현황(참여연구자) 등 요구자료 대응
✓ 정보통신·방송 연구개발 사업 관련 타 전문기관의 동일업무 수행
✓ 기획재정부, 과학기술정보통신부 주관 고객만족도 조사
✓ 소프트웨어중심대학사업단 프로그램 활동 증빙

○ **개인정보를 제공받는 자:** 과학기술정보통신부, 국회 등 정부기관, 한국연구재단 부설 정보통신기획평가원 등 정보통신·방송 연구개발사업의 전문기관, 범부처 연구비통합관리시스템(통합이지바로), 국가과학기술종합정보시스템(NTIS), 한국기업데이터 주식회사, 한국정보통신기술협회, 기획재정부 및 과학기술정보통신부가 선정한 고객만족도 주간사(수행기관)

○ **개인정보를 제공받는 자의 이용목적:** ①국가연구개발사업 참여의 적법성 판단, ②과제수행에 대한 적법·적정성 판단, ③과제 선정·평가·관리 업무 수행

○ **제공하는 개인정보 항목 :** 성명, 근무기관, 국가연구자번호, 주소, 연락처, 이메일, 소속, 영상물(사진 등)등

○ **개인정보를 제공받는자의 개인정보 보유・이용 기간: 동의서가 작성된 시점부터 상기 개인정보 제3자 제공목적 달성시까지**

○ **관련 근거:** 국가연구개발혁신법 제19조 및 동법 시행령 제42조, 국가연구개발정보처리기준, 정보통신·방송 연구개발 관리규정 제12조

※ 개인정보의 제3자 제공에 동의하십니까? (■ 동의함 □ 동의하지 않음)

※ 유의사항 : 귀하는 상기 동의를 거부할 수 있습니다. 해당 수집 항목은 정보통신·방송 연구개발 수행에 반드시 필요한 사항으로 이에 대한 동의를 하지 않을 경우에는 정보통신·방송 연구개발 참여 등에 제한을 받으실 수 있습니다.

소속학과: 컴퓨터정보통신공학과 학번: 214683 성명: 장인환 (서명)

[붙임5] 2026년 소·중·대 산학협력프로젝트(캡스톤디자인) 결과보고서

2026년 전남대학교 소프트웨어중심대학사업
소·중·대 산학협력프로젝트(캡스톤디자인) 결과보고서

프로젝트명	DDRS : Data Distributed with Rolling Storage(분산 처리 기반 지식증류 파이프라인)					
Github url 주소	https://github.com/janginh/capstone_Distributed_Distillation					
팀 명	2gether			과제수행기간	2026. 3. 25. ~ 6. 23.	
지도교수	학 과	전자컴퓨터공학부		성 명	김형일	
프로젝트 수행인원 (※팀장은 첫줄에 기입)	학과(부·복수전공)	학년	학번	성명	연락처	E-Mail
	컴퓨터정보통신공학과	4	214691	강민수	010-9137-1632	214691@jnu.ac.kr
	컴퓨터정보통신공학과	4	214683	장인환	010-4519-5336	214683@jnu.ac.kr
참여 기업	기업명	멘토명		직위	연락처(HP)	E-Mail

위와 같이 2026년 전남대학교 소프트웨어중심대학사업
산학협력프로젝트(캡스톤디자인) 지원 프로그램 결과보고서를 제출합니다.

2026년 6월 23일

신청자명(대표학생) : 강민수 (인)
지도교수 : 김형일 (인)

전남대학교 소프트웨어중심대학사업단장 귀하

산학협력프로젝트(캡스톤디자인) 결과보고서(요약)

프로젝트명	DDRS : Data Distributed with Rolling Storage (분산 처리 기반 지식증류 파이프라인)		
수행기간	2026. 3. 25. ~ 6. 23.	**소요예산**	272,617원
소요예산 세부내역	-회의비 : 173,700원 -SW활용비 : 98,917원		

참여인원	**구분**	**인원수**	**성명(모두 기재)**
	교수	1	김형일
	석박사과정	-	
	학부생	2	강민수, 장인환
	기업체	1	손지성
	계	**4**	

추진배경

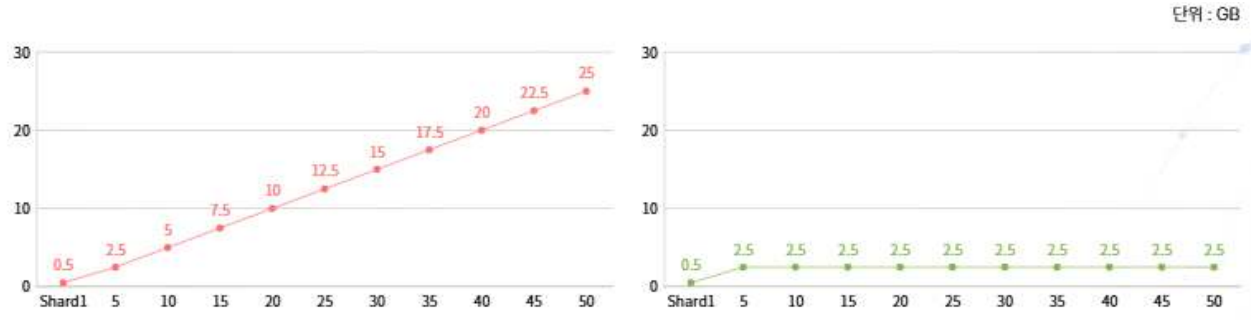
다양한 라벨 없는 데이터를 활용한 Edge 디바이스 환경의 효율적인 학습 프레임 워크

IoT 및 엣지 디바이스의 확산으로 CCTV, 드론, 블랙박스 등에서 대규모 영상 데이터가 생성되고 있으나, 대부분 라벨이 없어 AI 학습에 활용되지 못하고 있다. 본 연구는 라벨이 없는 대규모 데이터를 제한된 Edge 환경에서도 효율적으로 학습하는 DDRS 시스템을 제안한다. Kafka 기반 데이터 스트리밍, HDFS 분산 저장, Shard 단위 Rolling Storage 순환 학습, Teacher-Student Knowledge Distillation을 결합하여 저장공간과 연산 자원의 제약을 동시에 해소한다.

목표 및 내용

Edge 디바이스 환경에서도 Disk, Memory 관리를 진행하며 적은 cost의 모델 학습 프레임 워크 제작

Rolling Storage 방식을 통해 전체 데이터 크기와 무관하게 일정한 디스크 사용량(약 2.5GB)을 유지하는 데 성공하였다. Knowledge Distillation을 통해 경량 Student 모델(YOLO-World-S)을 학습하였으며, Teacher 대비 메모리 약 46$, 추론 지연 약 32%, 전력 약 42% 감소를 달성하였다. 학습 전후 정성적 분석을 통해 Edge 환경도 도메인에서의 탐지 성능 개선을 확인하였다.



기대효과

다양한 환경의 영상 촬영 기기에서 실시간 추론 모델 학습

본 시스템은 스마트시티 교통 모니터링, 산업현장 안전 감시, 주차장 차량 관리 등 다양한 CCTV 환경에 적용 가능하다. 버려지던 영상 데이터를 학습 자산으로 전환하고, 수동 라벨링 비용 없이 환경 특화 모델을 지속적으로 생성하는 Self-Evolving Edge AI 인프라로 활용될 수 있다.

1. 프로젝트 개요

프로젝트명	**DDRS : 분산 처리 기반 Knowledge Distillation 파이프라인** **Data Distributed with Rolling Storage**
주제영역	■ 생활 □ 업무 ■ 공공/교통 □ 금융/핀테크 □ 의료 □ 교육 □ 유통/쇼핑 □ 엔터테인먼트
기술분야	■ IoT ■ 모바일 □ 데스트톱 SW ■ 인공지능 □ 보안 □ 가상현실 ■ 빅데이터 □ 자동제어기술 □ 블록체인 ■ 영상처리 □ 기타( )
성과목표	■ 논문게재 및 포스터발표 □ 앱등록 □ 프로그램등록 □ 특허 □ 기술이전 ■ 실용화 □ 공모전( ) □ 기타( )

2. 프로젝트 추진배경

(1) 대규모 비정형 데이터의 폭발적 증가

최근 IoT 및 엣지 디바이스의 급격한 발전으로 CCTV, 드론, 차량 블랙박스, 산업용 센서 등에서 대량의 비정형 영상 데이터가 끊임없이 생성되고 있다. 글로벌 데이터 생성량은 2010년 약 2 제타바이트(ZB)에서 2025년 약 181 제타바이트로 약 90배 증가하였으며, 매년 20% 이상의 증가율을 지속적으로 유지하고 있다. 1 제타바이트가 10의 12승 기가바이트에 해당함을 고려하면, 현재 인류가 생성하는 데이터의 양은 기존의 저장 및 처리 패러다임으로는 감당하기 어려운 수준에 도달하였다.

이러한 데이터의 상당 부분은 영상·이미지와 같은 비정형 데이터이며, 특히 CCTV와 같은 고정형 감시 장비는 24시간 연속으로 영상을 생성한다. 그러나 생성되는 데이터의 대부분은 라벨(정답)이 없는 상태로 수집되기 때문에, 지도학습(Supervised Learning) 기반의 AI 모델 학습에 직접 활용하기 어렵다는 근본적인 한계가 존재한다.

연도별 글로벌 데이터 생성량의 추이는 아래 표와 같으며, 증가 속도가 점차 가속화되고 있음을 확인할 수 있다.

(2) 라벨링 비용과 데이터 활용의 한계

AI 모델, 특히 객체 탐지(Object Detection) 모델을 학습시키기 위해서는 각 이미지에 객체의 위치(Bounding Box)와 종류(Class)를 사람이 직접 표기하는 라벨링 작업이 필요하다. 그러나 수백만 장에 달하는 대규모 데이터셋에 대해 수동 라벨링을 수행하는 것은 막대한 비용과 시간을 요구한다. 그 결과, 현장에서 생성되는 방대한 데이터의 극히 일부만이 학습에 활용되고 나머지는 단순 저장 후 폐기되는 비효율이 발생하고 있다.

따라서 라벨이 없는 데이터를 자동으로 활용할 수 있는 준지도학습(Semi-Supervised Learning) 및 자기지도학습(Self-Supervised Learning) 기반의 학습 방법론이 요구된다. 본 연구는 고성능 Teacher 모델이 라벨 없는 데이터에 가상의 라벨(Pseudo-label)을 부여하고, 이를 경량 Student 모델이 학습하는 지식 증류(Knowledge Distillation) 방식을 핵심 해결책으로 채택하였다.

(3) 엣지 환경에서의 AI 모델 한계

실시간 객체 탐지가 필요한 현장에서는 데이터를 중앙 서버로 전송하지 않고 엣지 디바이스 자체에서 추론을 수행하는 On-Device AI가 선호된다. 이는 네트워크 지연을 줄이고 개인정보 보호 측면에서도 유리하기 때문이다. 그러나 Open-Vocabulary Object Detection(OVOD)이 가능한 고성능 모델은 파라미터 수가 많고 연산량이 높아, Jetson Nano나Jetson NX와 같은 제한된 자원의 엣지 디바이스에서 직접 구동하기에는 모델 크기, GPU 성능, 메모리 요구사항 측면에서 부적합하다.

이러한 문제를 해결하기 위해 큰 Teacher 모델의 지식을 작은 Student 모델로 전이하는 경량화 기법

이 필요하다. 본 연구에서는 OVOD 능력을 보유한 YOLO-World 계열 모델을 활용하여, Teacher의 표현력과 개방 어휘 탐지 능력을 유지하면서도 엣지 환경에서 실시간 동작이 가능한 경량 Student 모델을 학습하는 것을 목표로 한다.

(4) 저장 공간 및 I/O 병목 문제

대규모 데이터를 학습에 활용하기 위해서는 데이터의 저장과 입출력(I/O) 효율이 핵심 과제가 된다. 수백 기가바이트에 달하는 데이터셋 전체를 단일 엣지 디바이스의 로컬 디스크에 저장하는 것은 물리적으로 불가능에 가깝다. 또한 데이터를 지속적으로 저장·전송·관리하는 과정에서 I/O 병목이 발생하여 학습 속도가 데이터 공급 속도를 따라가지 못하는 현상이 빈번하게 나타난다. 단일 서버 환경에서는 이러한 저장 공간 부족과 성능 저하 문제가 더욱 심각하게 드러난다.

본 연구는 이를 해결하기 위해 데이터를 일정 크기의 Shard로 분할하여 분산 저장하고, 학습 시에는 소수의 Shard만 로컬에 유지한 뒤 학습이 완료되면 삭제하는 Rolling Storage 기법을 제안한다. 이를 통해 전체 데이터 크기와 무관하게 일정한 로컬 디스크 사용량을 유지하면서 대규모 데이터 학습을 가능하게 한다.

3. 프로젝트(주제) 목표 및 내용

본 연구의 최종 목표는 제한된 Edge 환경에서도 라벨이 없는 대규모 데이터를 효율적으로 학습하는 통합 시스템 DDRS(Data Distributed with Rolling Storage)를 구축하는 것이다. 구체적인 연구개발 범위는 다음과 같다.

- 다수 CCTV에서 수집되는 영상 데이터의 Kafka 기반 실시간 스트리밍 파이프라인 구축
- HDFS 분산 저장 및 Shard 단위 파티셔닝 구조 설계
- Rolling Storage 기반 순환 학습(Load -> Train -> Delete)으로 저장공간 최적화
- Teacher -> Student Knowledge Distillation을 통한 경량 모델 학습
- Pseudo-label 자동 생성 및 Confidence 기반 필터링
- Edge 환경 대상 모델 추론 cost(메모리, 지연, 전력) 평가
- 학습 전후 탐지 성능에 대한 정성적 분석 및 도메인 적응 효과 검증

4. 시스템 구성 및 내용

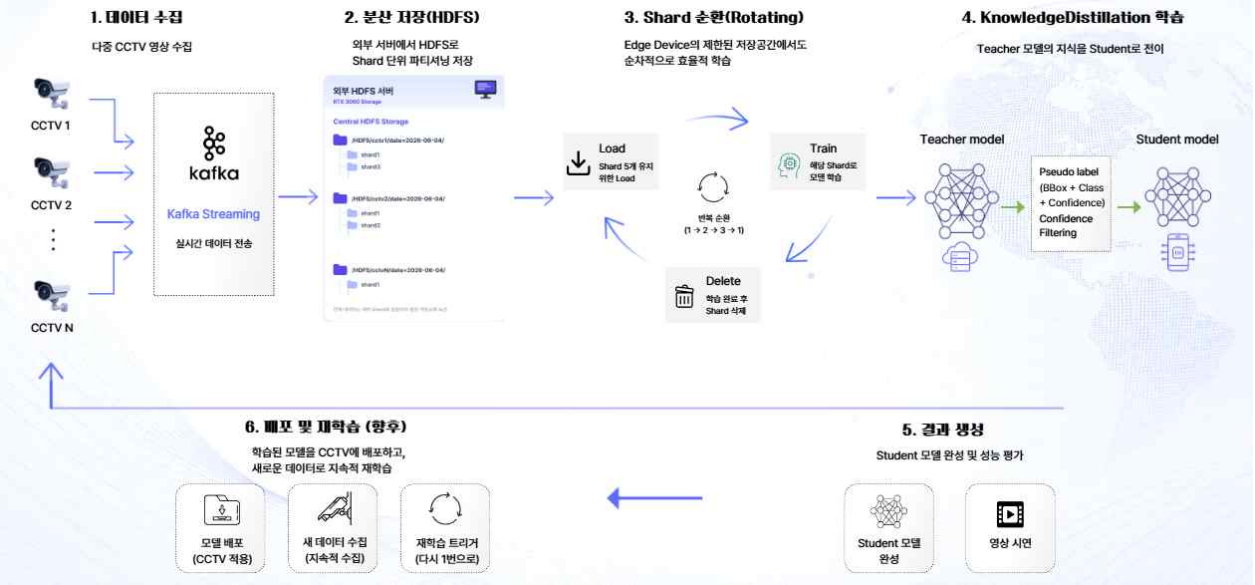
□ 전체 시스템 구조(DDRS Pipeline)

본 연구에서 제안하는 DDRS 시스템은 데이터 수집부터 모델 학습 및 배포까지 총 6단계의 파이프라인으로 구성된다. 각 단계는 독립적인 모듈로 설계되어 유지보수와 확장이 용이하며, 전체 인프라는 Docker Compose를 통해 컨테이너 기반으로 관리된다.

- 1단계 데이터 수집 : 다수의 CCTV에서 촬영된 영상을 Apache Kafka Streaming을 통해 실시간으로 전송한다. 각 CCTV는 독립적인 데이터 소스로 동작하며, 디바이스 식별자를 기준으로 데이터가 구분된다.
- 2단계 분산 저장(HDFS) : 수집된 데이터를 외부 HDFS 서버에 Shard 단위로 파티셔닝하여 분산저장한다. 저장 경로는 CCTV별 날짜별로 체계적으로 구분된다.
- 3단계 Shard 순환(Rolling) : Edge Device의 제한된 저장공간에서 Load -> Train -> Delete 순환

구조를 통해 순차적으로 학습을 진행한다.

- 4단계 Knowledge Distillation : Teacher 모델이 생성한 Pseudo-label을 Student 모델이 학습하여 지식을 전이받는다.
- 5단계 결과 생성 : 학습이 완료된 경량 Student 모델을 산출하고 추론 성능을 평가한다.
- 6단계 배포 및 재학습(향후) : 학습된 모델을 CCTV에 배포하고, 새롭게 수집되는 데이터로 지속적인 재학습을 수행하는 순환구조를 지향한다,



□ 데이터 수집 및 스트리밍 (Apache Kafka)

데이터 수집 단계에서는 Apache Kafka를 메시지 브로커로 활용하여 다수 CCTV로부터 유입되는 영상 데이터를 안정적으로 처리한다. Kafka는 3개의 Broker로 구성된 클러스터로 운영되며, KRaft 모드를 채택하여 별도의 Zookeeper 없이 자체적으로 메타데이터를 관리함으로써 운영 복잡도와 메모리 사용량을 절감하였다.

토픽은 12개의 파티션과 replication factor 2로 설정하여, 데이터 처리의 병렬성을 확보하는 동시에 특정 Broker 장애 시에도 데이터가 유실되지 않도록 안정성을 보장한다. 각 CCTV의 디바이스 식별자를 파티션 키로 사용하여 동일 디바이스의 데이터가 순서를 유지하도록 설계하였으며, 대용량 이미지 전송을 위해 메시지 최대 크기를 확장하고 lz4 압축을 적용하여 네트워크 대역폭 사용을 최적화하였다.

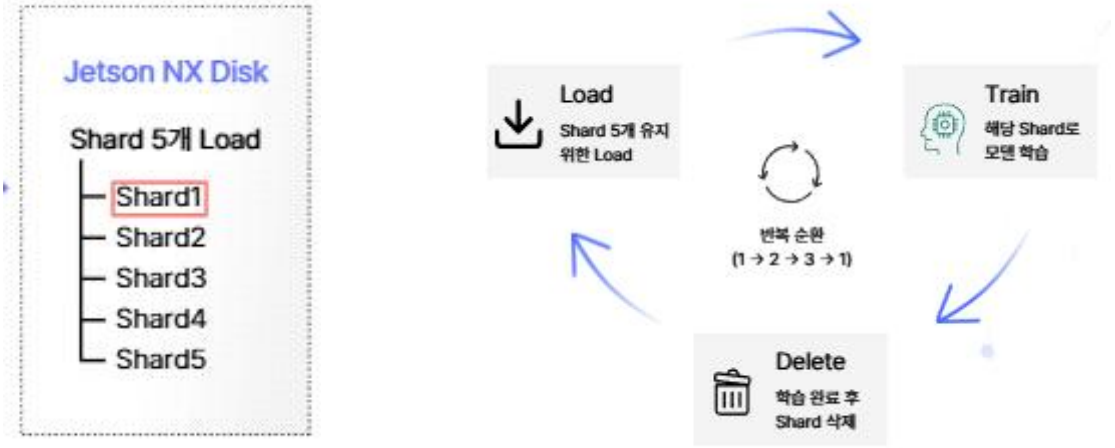
□ 분산 저장 및 Shard 파티셔닝 (HDFS)

수집된 데이터는 외부 HDFS 서버에 저장된다. HDFS 클러스터는 1개의 NameNode와 3개의 DataNode로 구성되며, replication factor 2를 적용하여 데이터의 분산 저장과 복제를 통한 안정성을 동시에 확보한다. NameNode는 파일 시스템의 메타데이터를 관리하고, DataNode들은 실제 데이터 블록을 분산하여 저장한다.

데이터는 0.5GB 단위의 tar 파일(Shard)로 패킹되어 저장된다. 수백만 장의 이미지를 개별 파일로 저장할 경우 파일 시스템의 inode 오버헤드와 랜덤 액세스로 인한 I/O 비효율이 발생하므로, 다수의 이미지를 하나의 tar Shard로 묶어 순차 읽기(Sequential Read)가 가능하도록 하였다. Shard는 CCTV별·

날짜별 디렉토리 구조(/HDFS/cctvN/date=YYYY-MM-DD/shardN)로 파티셔닝되어 저장되며, 이를 통해 특정 CCTV나 특정 기간의 데이터를 효율적으로 조회할 수 있다.

Shard Builder 모듈은 Kafka Consumer로 동작하며, 수신한 이미지를 메모리 버퍼 상에서 tar 파일로 조립한다. 누적 크기가 0.5GB에 도달하면 해당 Shard를 HDFS에 업로드하고 새로운 Shard를 시작하는 방식으로 동작한다.



□ Rolling Storage 기반 순환 학습

Rolling Storage는 본 연구의 핵심 차별점으로, 제한된 저장 공간에서도 대규모 데이터를 학습할 수 있게 하는 메커니즘이다. 전체 데이터셋은 HDFS에 분산 저장되어 있으며, 학습 시에는 로컬 디스크에 5개의 Shard(약 2.5GB)만을 유지한다.

학습 과정은 다음과 같이 진행된다. 먼저 HDFS에서 Shard 5개를 로컬로 가져온다(Load). 이후 Shard 1개에 대한 학습을 수행하고(Train), 학습이 완료되면 해당 Shard를 로컬에서 삭제한다(Delete). 삭제로 확보된 공간에 HDFS로부터 다음 Shard를 가져와 학습을 이어가는 순환 구조로 동작한다. 이러한 Load → Train → Delete의 반복을 통해 전체 데이터를 순차적으로 모두 학습하면서도, 로컬 디스크 사용량은 항상 약 2.5GB로 일정하게 유지된다.

원격 HDFS로부터 Shard를 가져오는 과정에서 발생하는 통신 오버헤드는 Prefetching 기법으로 은닉한다. 현재 Shard에 대한 학습이 진행되는 동안, 다음에 사용될 Shard를 백그라운드에서 미리 가져옴으로써 GPU가 데이터를 기다리는 유휴 시간을 최소화한다. Knowledge Distillation 학습은 Teacher 모델의 추론 연산이 포함되어 Shard 1개의 학습 시간이 Shard 1개의 전송 시간보다 충분히 길기 때문에, Prefetching을 통해 통신 시간이 학습 시간 뒤로 완전히 은닉되어 실질적인 성능 저하 없이 원격 분산 저장의 이점을 누릴 수 있다.

□ Knowledge Distillation 학습

(1) Teacher-Student 구조

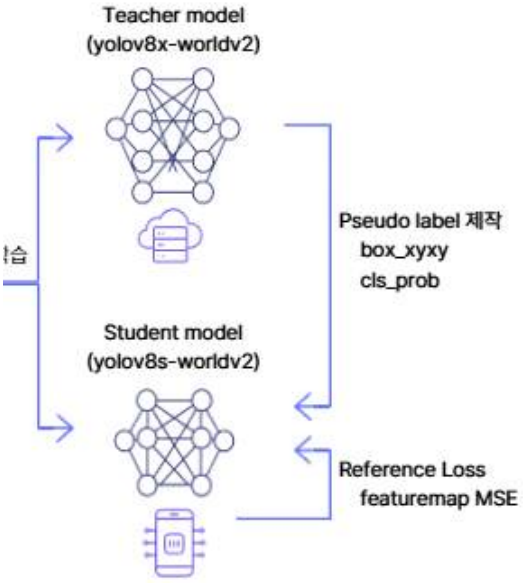
Knowledge Distillation은 고성능 Teacher 모델의 지식을 경량 Student 모델로 전이하는 모델 압축 기법이다. 본 연구에서는 Teacher 모델로 YOLO-World-X를, Student 모델로YOLO-World-S를 사용하였다. 두 모델 모두 Open-Vocabulary Object Detection 능력을 보유한 동일 계열의 모델로, 구조적 호환성이 높아 효율적인 지식 전이가 가능하다.

(2) Pseudo-label 생성 및 Confidence 필터링

Teacher 모델은 라벨이 없는 입력 이미지를 추론하여 Pseudo-label을 생성한다. Pseudo-label은 객체의 위치 정보(Bounding Box, box_xyxy), 클래스 확률(cls_prob), 신뢰도(Confidence)로 구성된다. 생성된 모든 Pseudo-label을 학습에 사용하는 것이 아니라, Confidence 값을 기준으로 필터링을 수행하여 신뢰도가 낮은 불확실한 예측을 제거한다. 이를 통해 Teacher의 오탐(False Positive)이 Student의 학습을 오염시키는 것을 방지하고 학습의 안정성을 확보한다.

(3) Distillation Loss

Student 모델은 Teacher의 출력을 모방하도록 학습된다. 학습에는 Teacher와 Student의 중간 특징맵(feature map) 간 차이를 줄이는 feature map MSE 기반의Reference Loss를 활용한다. 이는 단순히 최종 출력만을 모방하는 것이 아니라 Teacher가 학습한 내부 표현을 Student가 따라가도록 함으로써, Teacher가 보유한 OVOD 성능을 유지하면서 경량화를 달성하는 데 핵심적인 역할을 한다.



□ Producer–Consumer 구현 및 Docker 인프라

데이터 송수신은 Producer-Consumer 패턴으로 구현하였다. Producer는 CCTV 또는 데이터 소스에서 이미지를 읽어 전처리(리사이즈, 인코딩) 후 Kafka로 전송하며, 멀티스레드 방식으로 다수의 이미지를 병렬 전송하여 처리량을 극대화하였다. Consumer는 Kafka의Consumer Group으로 동작하여 여러 파티션의 데이터를 병렬로 소비하고, 수신한 이미지를 Shard로 패킹하여 HDFS에 적재한다.

전체 인프라는 Docker Compose를 통해 컨테이너 기반으로 구성하였다. Kafka 3개 Broker, HDFS NameNode 1개와 DataNode 3개, 모니터링용 Kafka UI 등 총 8개의 컨테이너가 단일 설정 파일로 정의되며, 명령어 한 줄로 전체 시스템을 기동할 수 있다. 각 컨테이너에는 메모리 제한을 설정하여 자원을 효율적으로 배분하였으며, 컨테이너 간 통신은 내부 네트워크를 통해 이루어진다. 이러한 컨테이너화를 통해 어떠한 환경에서도 동일한 시스템을 재현할 수 있는 이식성을 확보하였다.

□ 시스템 동작 시나리오

앞서 설명한 각 구성요소가 유기적으로 연동되는 전체 동작 과정을 시간 순서에 따라 정리하면 다음과 같다. 이는 실제 데이터 한 건이 시스템에 유입되어 학습에 활용되기까지의 흐름을 보여준다.

먼저 다수의 CCTV가 영상을 촬영하고, Producer가 각 프레임을 전처리하여 Kafka의 raw-images 토픽으로 전송한다. 이때 디바이스 식별자가 파티션 키로 사용되어 동일 CCTV의 데이터는 순서를 유지한다. Kafka는 3개의 Broker에 데이터를 분산 저장하며 replication을 통해 안정성을 확보한다.

다음으로 Shard Builder가 Kafka에서 이미지를 소비하여 메모리 상에서 tar Shard로 패킹한다. Shard의 크기가 0.5GB에 도달하면 HDFS의 해당 CCTV·날짜 디렉토리에 업로드하고 새로운 Shard를 시작한다. HDFS는 각Shard를 여러 DataNode에 분산 저장한다.

학습 단계에서는 Shard Trainer가 HDFS로부터 5개의 Shard를 로컬로 가져온 뒤, Shard를 하나씩 언팩하여 학습에 활용한다. 각 Shard에 대해 Teacher 모델이Pseudo-label을 생성하고, Confidence 필터링을 거친 후 Student 모델이 이를 학습한다. 한 Shard의 학습이 완료되면 해당 Shard를 로컬에서 삭제하고, Prefetching으로 미리 가져온 다음 Shard로 학습을 이어간다. 이 과정을 전체 Shard에 대해 반복하면 1 epoch이 완료되며, 필요에 따라 여러 epoch을 반복하여 학습을 진행한다.

5. 프로젝트 결과물에 대한 기술

□ 활용분야

본 시스템은 도메인별 Expert Module 교체만으로 다양한 환경에 적용 가능한 범용성을 갖추고 있다. 구체적인 활용 분야는 다음과 같다.

- 스마트시티 교통 모니터링 : 도로 CCTV를 통한 차량,보행자 탐지 및 교통량 분석
- 산업 현장 안전 감시 : 공장 내 작업자 안전 장비 착용 여부 및 위험 구역 침입 탐지
- 주차장 차량 관리 : 차량 출입 및 주차 공간 점유 현황 모니터링
- 학교 시설 출입 관리 : 인가되지 않은 출입 탐지 및 시설 보안 강화
- 재난 재해 상황 감시 : 화재, 침수 등 비정상 상황의 조기 탐지

□ 성과 관리 추진체계

연구 성과는 GitHub를 통한 버전 관리 체계로 코드와 학습 파이프라인을 체계적으로 관리한다. Branch 전략과 Issue/Project Board를 활용하여 협업 과정을 기록하며, Docker 기반의 컨테이너화를 통해 어떠한 환경에서도 동일하게 시스템을 재현할 수 있도록 한다.

□ 추가 연구의 필요성 및 향후 계획

본 연구는 학습 파이프라인 구축을 완료한 단계로, 설계 운영 가능한 시스템으로 발전시키기 위해 다음과 같은 추가 연구가 필요하다.

(1) 실제 환경 구축 및 배포

Edge device에서 학습된 모델을 실제 CCTV 환경에 적용하고, 안정적인 데일 수집과 학습 파이프라인을 현장에서 운영하며 성능을 검증한다. 이를 통해 실제 운영 가능한 시스템의 완성을 목표로 한다.

(2) 컴파일 자동화

CCTV 칩셋별로 최적화된 컴파일을 자동화한다. TensorRT, TFLite, ONNX 등 다양한 추론 엔진으로의 변환과 가중치 파일의 배포까지 전 과정을 파이프라인화하여, 다양한 Edge 환경에 빠르고 안정적으

로 배포할 수 있도록 한다.

(3) 가중치 자동 변경 및 환경별 최적 모델 생성

CCTV 해상도, 채널 수, 성능 요구사항, 하드웨어 종류 등 다양한 조건에 맞춰 환경별 최적 모델을 자동으로 생성하는 체계를 구축한다. 이를 통해 다양한 현장 조건에 유연하게 대응할 수 있다.

(4) 다양한 Task 및 학습 방법 확장

라벨이 없는 대규모 데이터를 다방면으로 활용한다. Detection뿐만 아니라 Segmentation 등으로 Task의 범위를 확장하고, OVOD 모델의 특징을 활용하여 다양한 Class에 대한 학습을 진행한다. 또한 Self-Supervised Learning과 Semi-Supervised Learning을 도입하여 라벨 없는 데이터의 활용도를 더욱 높인다.

(5) 데이터 품질 관리

동영상 데이터를 프레임 단위로 처리할 경우 불필요하거나 품질이 낮은 프레임이 학습에 혼입될 수 있다. 이를 방지하기 위해 CLIP-IQA 기반의 품질 필터링 알고리즘을 추가하여, 흔들리거나 어둡거나 흐릿한 저품질 영상을 자동으로 제거하고 선명한 고품질 영상만을 선별하여 학습 데이터의 품질을 지속적으로 개선한다.

□ 성능 평가

Knowledge Distillation을 통해 경량화된 Student 모델의 효율을 검증하기 위해, Teacher 모델(YOLO-World-X)과 Student 모델(YOLO-World-S)의 추론 cost를 Edge 환경에서 비교하였다. 평가 지표로는 추론 시 최대 메모리 사용량(Peak Memory), 이미지당 추론 지연 시간(Latency), 최대 소비 전력(Peak Power)을 측정하였다. 측정 결과, 경량 Student 모델은 Teacher 대비 최대 메모리 약 46%, 추론 지연 약 32%, 소비 전력 약 42%의 감소를 달성하였다. 이는 Student 모델이 동일한 Edge 디바이스에서 더 적은 자원으로 더 빠르게 동작함을 의미하며, 실시간 서비스 제공과 저전력 운영 측면에서 Edge 환경 배포에 적합함을 입증한다.

모델	Peak Memory	Latency	Peak Power
YOLO-World-X (Teacher)	2.4 GB	123.85 ms/img	21.3W
YOLO-World-S (Student)	1.3 GB	84.6 ms/img	12.37W
감소율	**약 46% ↓**	**약 32% ↓**	**약 42% ↓**

□ 기대 성과

향후 발전 로드맵은 현재 시스템(학습 파이프라인 구축 완료)을 출발점으로 하여, 실제 Edge 환경 구축 및 배포, Edge 배포 및 컴파일 자동화, 가중치 자동 변경 및 컴파일 자동화, 라벨 없는 데이터 중심 학습 확대의 단계를 거쳐 최종적으로 완성도 높은 Edge AI 시스템을 구현하는 것을 목표로 한다. 각 단계는 이전 단계의 성과를 기반으로 점진적으로 발전하며, 최종적으로는 현장에 배포된 모델이 스스로 데이터를 수집하고 재학습하는 Self-Evolving Edge AI 인프라의 완성을 지향한다.

구분	기능정의	세부기능 설명
1	Kafka 데이터 수집 및 전송	다수의 CCTV환경에서의 데이터 수집 및 전송을 위한 Kafka활용
2	HDFS 데이터 저장	외부 서버로 CCTV영상 데이터를 저장하고, 관리

		하는 분산 처리 저장 시스템
3	Shard 순환	외부 큰 서버에 저장된 데이터를 모두 load하는게 아닌 필요 데이터만 load->train->delete 과정을 통해 Disk, Memory 저장 공간의 감축
4	Knowledge Distillation	Edge(CCTV, Drone)과 같은 환경에서도 모델 사용을 위한 작은 모델을 학습하는 방법
5	모델의 실시간 추론	학습한 모델을 Edge환경에 구축함으로써 실시간 추론이 가능하도록 설계

6. 프로젝트 진행내용

1) 참여인원 및 담당 역할

연번	소속학과	성명	수행역할 분담내용
1	컴퓨터정보통신공학과	강민수	데이터 및 인프라 파이프라인 담당
2	컴퓨터정보통신공학과	장인환	모델 및 알고리즘 담당

2) 회의 및 SW멘토링 진행

번호	일시/장소	회의/멘토링 내용(상세히 작성)	관련 사진
1	2026. 5. 11. (13:00~15:00) / 공대7호관 카페	- 프로젝트 활용 방법 - 향후 계획 -	
2	2026. 6. 3. (13:00~15:00) / 공대7호관 카페	- 프로젝트 발표 개선점 - ppt 개선점 -	줌 미팅

7. 프로젝트 세부일정 및 내용

No.	작업 내용	4월				5월				6월				7월				담당자	비고
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4		
1	계획서 작성																		

2	공통 개발 환경 구축																	
3	데이터 & 모델 설정																	
4	외부 서버 Docker 구축																	
5	Jetson board Docker 구축																	
6	Knowledge Distillation 학습 진행																	
7	모델 성능 평가																	
8	결과보고서 작성/제출																	

8. 결과물에 대한 향후 활용계획

본 연구의 성과는 학술적·대외적 활동으로 확장하여 그 가치를 검증하고 공유할 계획이다. 우선 구축한 DDRS 시스템을 관련 공모전에 출품하여, 제한된 Edge 환경에서 라벨 없는 대규모 데이터를 학습하는 본 시스템의 독창성과 실용성을 외부 전문가들로부터 평가받고자 한다. 이를 통해 기술의 완성도를 객관적으로 점검하고 개선 방향을 모색할 수 있을 것으로 기대된다.

나아가 본 연구의 핵심 기여인 Rolling Storage 기반 순환 학습 기법과 Knowledge Distillation을 결합한 Edge AI 학습 구조를 정리하여 논문으로 작성하고, 국내 학회에 제출하고자 한다. 이를 통해 연구 결과를 학술적으로 공식화하고, 분산 데이터 처리와 엣지 컴퓨팅 분야의 후속 연구에 기여할 수 있도록 할 계획이다.

9. 참고자료 및 문헌

[1] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

[2] Cheng, T., et al. (2024). YOLO-World: Real-Time Open-Vocabulary Object Detection. CVPR 2024.

[3] Radford, A., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. ICML 2021.

[4] Apache Kafka. https://kafka.apache.org/

[5] Apache Hadoop. https://hadoop.apache.org/