무[無]계획

서태원, 신재환, 류은환, 최승렬

인천대학교 INCHEON NATIONAL UNIVERSITY | 임베디드시스템공학과

Contents

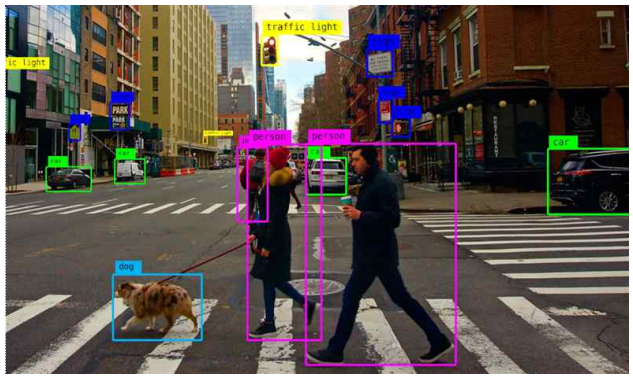
프로젝트 개요

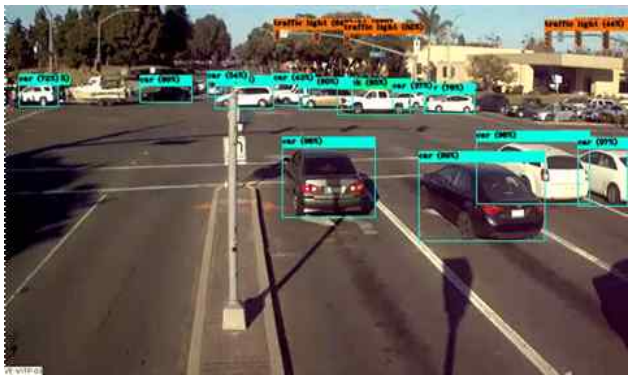
프로젝트 개요

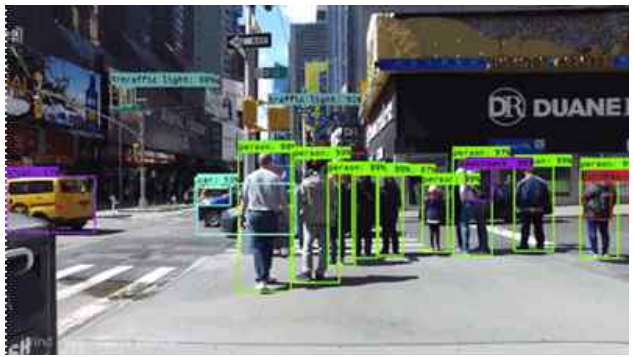
딥러닝 모델의 최적화를 통한

임베디드보드에서의 추론 속도 향상

Object detection 이란?

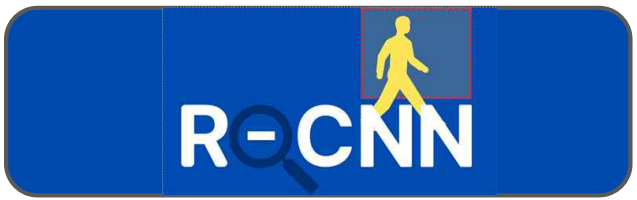






이미지나 비디오에서 객체를 인식하고, 위치를 특정하는 컴퓨터 비전 기술

Object detection Models



RetinaNet

EfficientDet

CenterNet

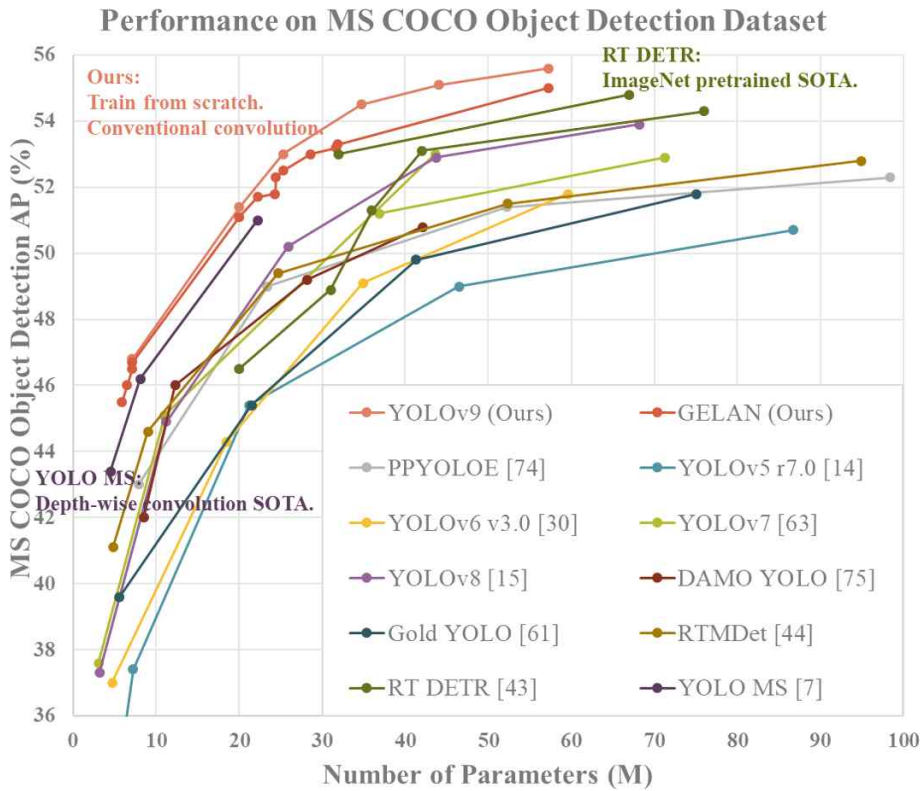
About YOLOv9

YOLOv9: SOTA Object Detection Model Explained

최신 YOLO Series, 2024년 2월 공개

Real-time Object Detection 분야 SOTA

SOTA: State Of The Art, 최고 성능 모델



프로젝트 필요성

프로젝트 필요성

일반적으로 고성능 딥러닝 모델은
GPU나 GPU서버 같은 고사양의 컴퓨팅 자원을 필요로 함

GPU 자원 X
Network 자원 X

임베디드 보드와 같은 제한된 환경
딥러닝 모델을 구동하기 어려움

프로젝트 필요성

온디바이스 **AI** 추세에 따라
딥러닝 모델도 **Edge device** 에서 구동하는 것이 트렌드

고성능 GPU를 설치는 비효율적
Network 사용 → 보안 문제

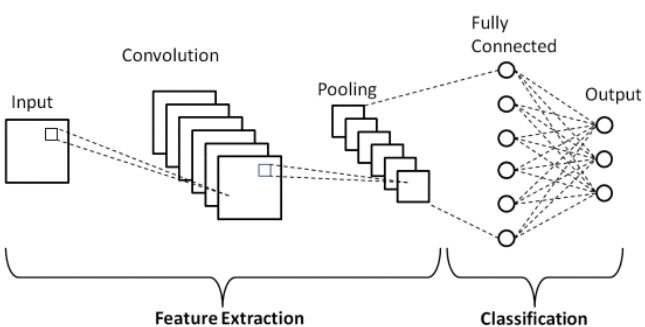
따라서, **딥러닝 모델을 경량화** 하는 방법을 탐색하고,
이를 바탕으로 **모델 구조 재설계** 및 **추론 속도 향상**을 목표

개발 진행 과정

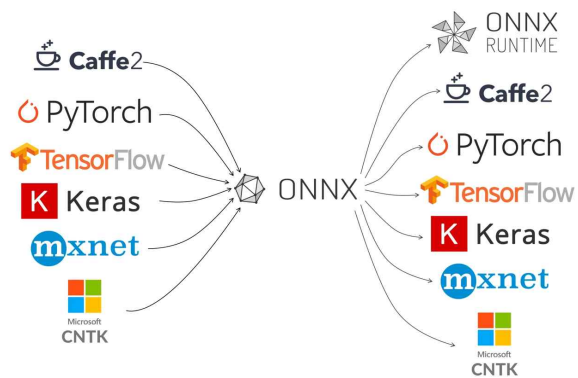
개발 진행 과정

- 두가지 분야로 개발 진행

모델 자체를 설계하는
아키텍처



학습된 모델에 적용하는
후처리

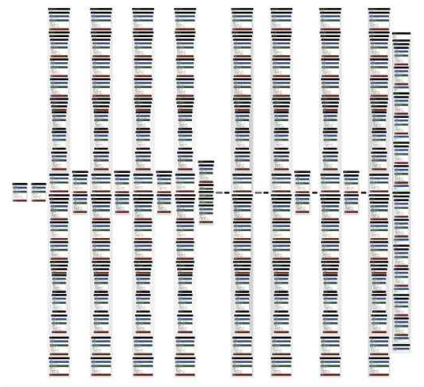


아키텍처

YOLOv9 논문 및 코드, 모델 블록 분석

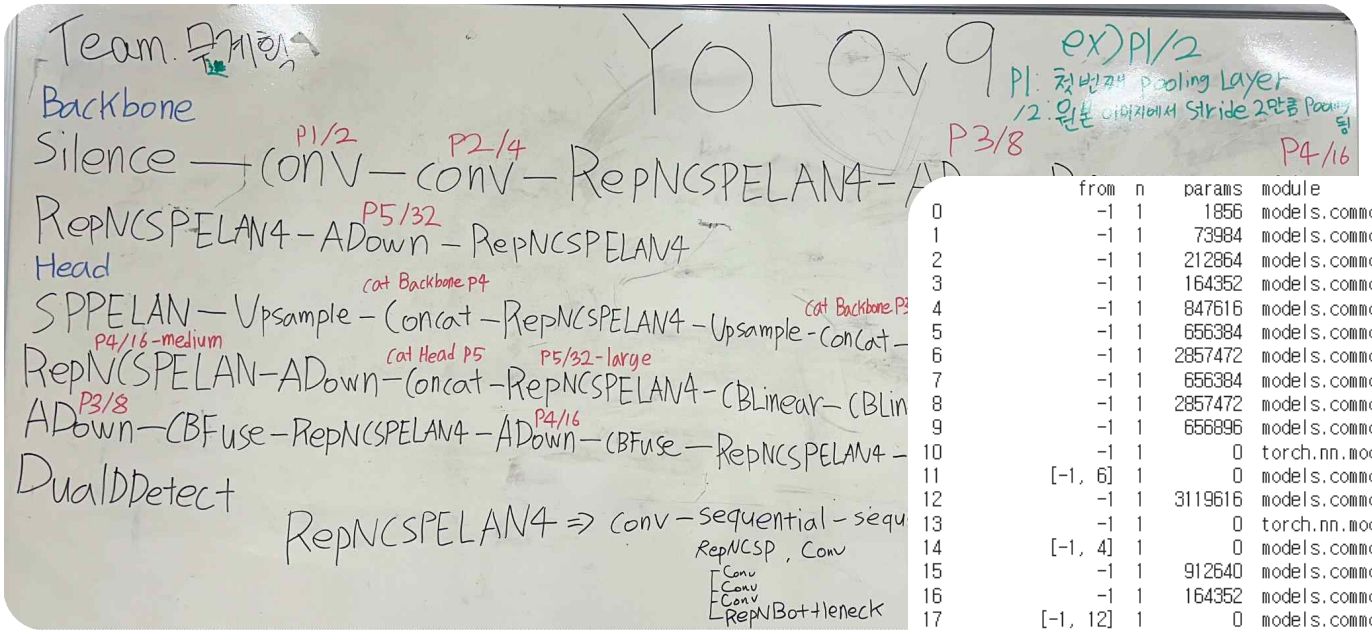
YOLOv9-c

test2



Netron 도구를 이용하여 아키텍처 시각화 및 분석 수행

YOLOv9 논문 및 코드, 모델 블록 분석



	from	n	params	module	arguments
0	-1	1	1856	models.common.Conv	[3, 64, 3, 2]
1	-1	1	73984	models.common.Conv	[64, 128, 3, 2]
2	-1	1	212864	models.common.RepNCSPELAN4	[128, 256, 128, 64, 1]
3	-1	1	164352	models.common.ADown	[256, 256]
4	-1	1	847616	models.common.RepNCSPELAN4	[256, 512, 256, 128, 1]
5	-1	1	656384	models.common.ADown	[512, 512]
6	-1	1	2857472	models.common.RepNCSPELAN4	[512, 512, 512, 256, 1]
7	-1	1	656384	models.common.ADown	[512, 512]
8	-1	1	2857472	models.common.RepNCSPELAN4	[512, 512, 512, 256, 1]
9	-1	1	656896	models.common.SPPELAN	[512, 512, 256]
10	-1	1	0	torch.nn.modules.upsampling.Upsample	[None, 2, 'nearest']
11	[-1, 6]	1	0	models.common.Concat	[1]
12	-1	1	3119616	models.common.RepNCSPELAN4	[1024, 512, 512, 256, 1]
13	-1	1	0	torch.nn.modules.upsampling.Upsample	[None, 2, 'nearest']
14	[-1, 4]	1	0	models.common.Concat	[1]
15	-1	1	912640	models.common.RepNCSPELAN4	[1024, 256, 256, 128, 1]
16	-1	1	164352	models.common.ADown	[256, 256]
17	[-1, 12]	1	0	models.common.Concat	[1]
18	-1	1	2988544	models.common.RepNCSPELAN4	[768, 512, 512, 256, 1]
19	-1	1	656384	models.common.ADown	[512, 512]
20	[-1, 9]	1	0	models.common.Concat	[1]
21	-1	1	3119616	models.common.RepNCSPELAN4	[1024, 512, 512, 256, 1]
22	[15, 18, 21]	1	5552320	models.yolo.DDetect	[80, [256, 512, 512]]

elan-c summary: 621 layers, 25498752 parameters, 25498736 gradients

모델 구성 연산 블록 코드 및 역할 분석

YOLOv9-c 모델 기반 추론 속도 향상 모델 설계 실험

test2: 최대 레이어의 크기를 128로 제한, 보조분기 적용

test10: activation 함수 변경 (SILU→LeakyReLU), 레이어 크기 수정, 보조분기 적용

학습 진행 중

MODEL NAME	YOLOv9-c-converted (Paper DL)	test2-converted_COCO_300
mAP50 score:	0.699	0.578
mAP50-95 score:	0.53	0.423
RPI avg inference time:	3334.0ms	1435.2ms

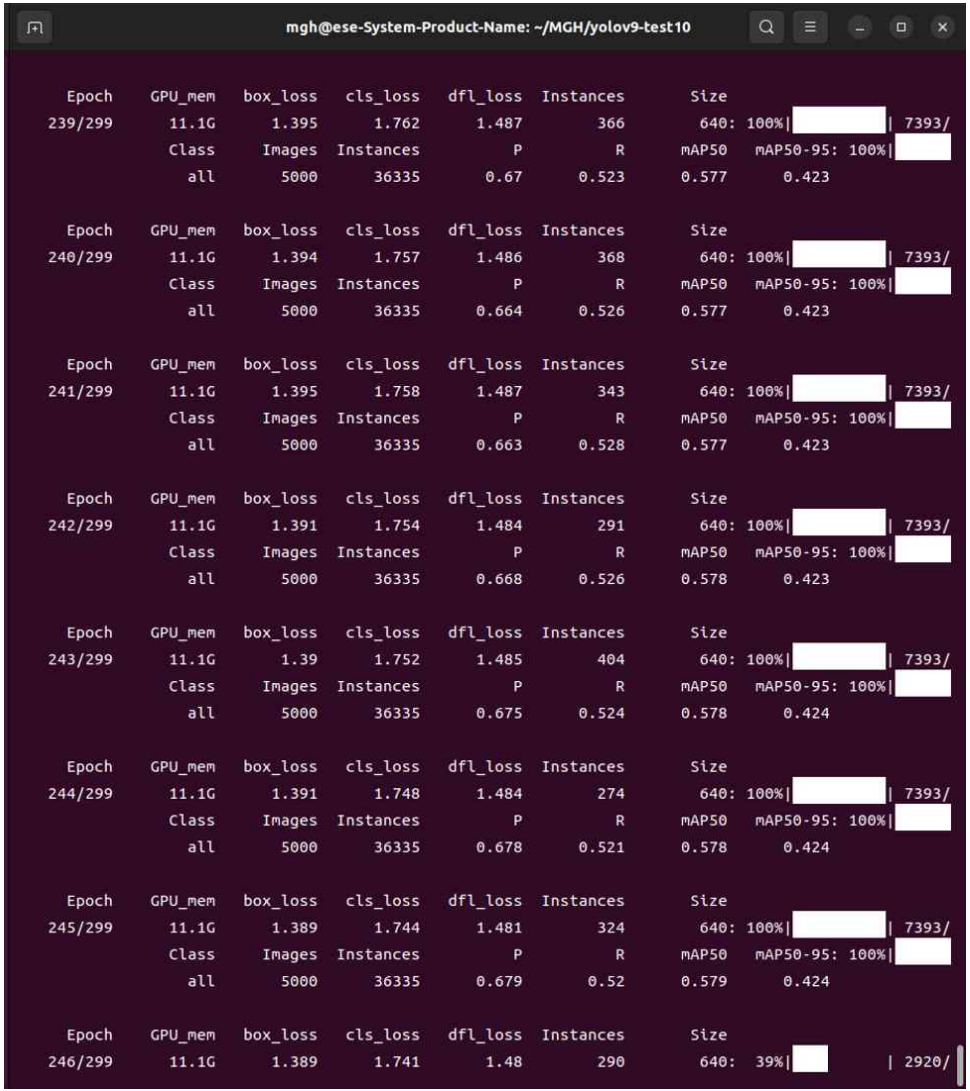
YOLOv7 시리즈 참조하여 YOLOv9-Tiny와 유사한 모델 설계 실험 중

학습 진행 중

모델 학습 진행

딥러닝 모델의 학습
(RTX4090)

모델 당 학습시간
약 4일 이상



모델 연산 블록 분석 과정

모델 파라미터(weights) 디스플레이 및 수정 코드 작성

연산량(FLOPs) 및 추론 시간(Speed)을 분석 실험

```python
import torch
from models.yolo import Model

torch.set_printoptions(threshold=100000000) # 더 많은 파라미터를 출력하기 위해 설정

device = torch.device("cuda")
cfg = "./models/detect/yolov9-c.yaml"
model = Model(cfg, ch=3, nc=80, anchors=3) # 아키텍처 구조 출력
#model = model.half()
model = model.to(device)
_ = model.eval()
ckpt = torch.load('/home/mgh/MGH/yolov9-c.pt', map_location='cuda')
model.names = ckpt['model'].names
model.nc = ckpt['model'].nc

for k, v in model.state_dict().items(): # k = 파라미터 이름, v = 파라미터 값
    print(k)
    print(ckpt['model'].state_dict()[k])
    print('-------------------------------------------------')
```

```
model.22.cv3.2.0.conv.weight
tensor([[[[-9.10282e-04, -7.48158e-04,  2.14577e-03],
          [ 1.11580e-03,  9.37939e-04,  4.16946e-03],
          [ 4.87137e-03,  4.87900e-03,  6.33240e-03]],

         [[-8.24451e-04, -5.93662e-04, -8.10146e-04],
          [-3.65973e-04, -2.26021e-04, -1.06621e-03],
          [-6.46114e-04, -1.01280e-03, -1.54400e-03]],

         [[ 9.67503e-04,  1.73378e-03,  4.64678e-04],
          [-2.34365e-04,  4.89712e-04, -1.32370e-03],
          [-2.66838e-03, -2.75993e-03, -4.21524e-03]],
```

후처리

후처리란?

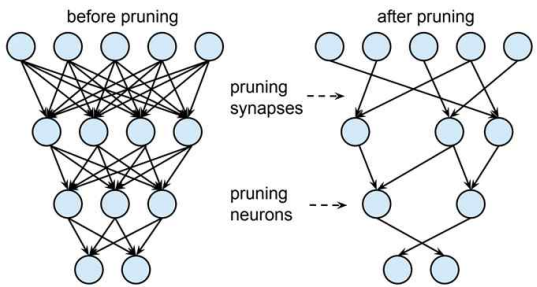
아키텍처는 모델 자체의 구조를 설계 및 학습

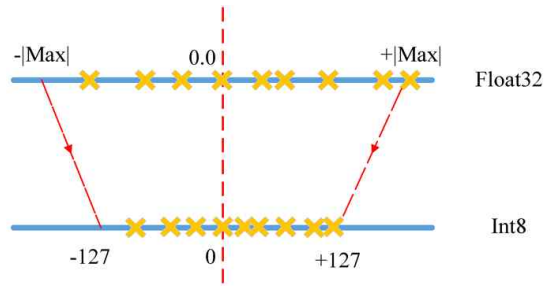
후처리는, 학습된 모델에 대해 추가 처리를 통해
라즈베리파이에서 연산을 최적화 하여 추론 속도를 향상시키는 것

ONNX

ONNX, ORT
변환

Pruning





Quantization

ONNX(Open Neural Network Exchange)

서로 다른 ML 프레임워크에서 개발된 모델을
서로 호환할 수 있도록 하는 표준 모델 포맷

다양한 프레임워크와 환경에서 모델 활용 가능
다양한 최적화 기술 및 옵션 적용 가능

ORT : 모바일 및 웹 애플리케이션과 같이
크기가 제한된 환경에서 사용하는 데 적합

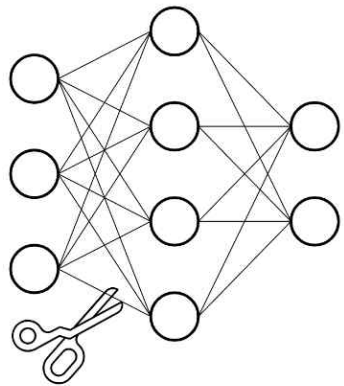
MODEL NAME	YOLOv9-c-converted (Paper DL)	YOLOv9-c-converted_onnx	YOLOv9-c-converted_ort
mAP50 score:	0.699	0.696	0.696
mAP50-95 score:	0.53	0.528	0.528
RPI inference time:	3334.0ms	1762.5ms	1689.3ms

ONNX 적용

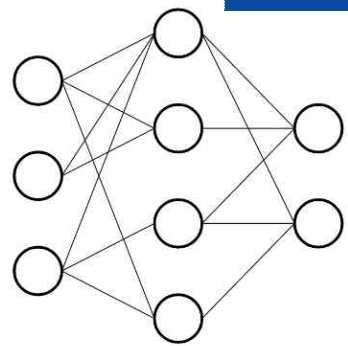
ORT 적용

Pruning (가지치기)

중요도가 낮은 weight 및 connection을 제거하거나
최대한 희소(sparse)하게 만드는 방법



Before pruning



After pruning

MODEL NAME	YOLOv9-c-converted (Paper DL)	YOLOv9-c-converted_pruning_20
mAP50 score:	0.699	0.654
mAP50-95 score:	0.53	0.491
RPI avg inference time:	3334.0ms	2953.6ms

12% 향상

양자화

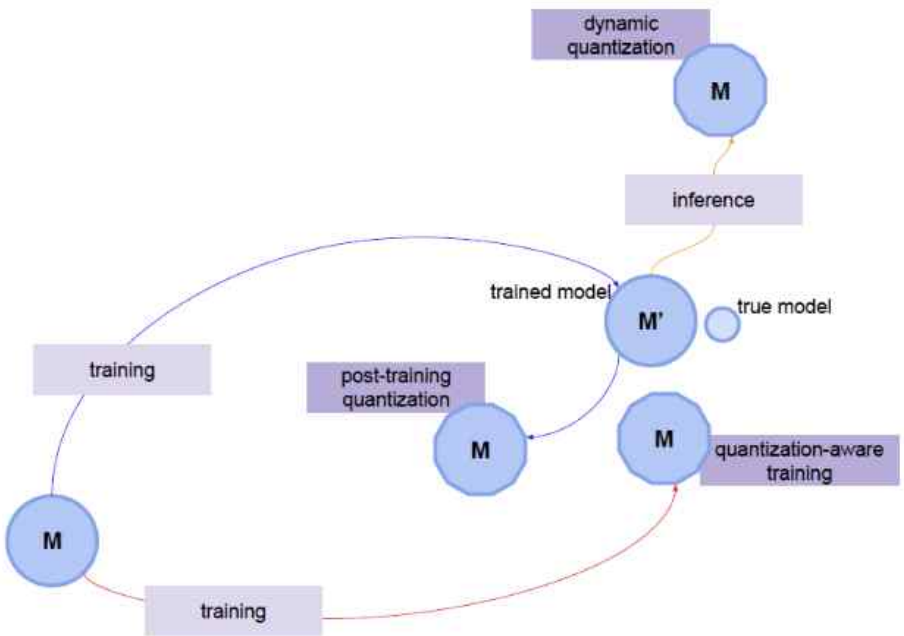
모델의 가중치를 낮은 비트의
데이터 형식으로 변환하여
더 빠른 추론을 가능하게 하는 방법

양자화의 종류

Quantization-aware Training :
학습 중에 양자화를 고려해서 모델을 조정

Post-training Quantization :
훈련 후 가중치를 줄이는 방법

Dynamic Quantization :
추론하면서 동적으로 양자화 진행



역할 분담

팀원별 역할분담

아키텍처 팀

서태원 [팀장]

YOLOv9 논문 실험 구현
YOLOv9 논문 분석
YOLOv9 코드 및 연산구조 분석
연산량 및 추론속도 분석 실험
CPU 온도 및 추론속도 영향 실험

192시간

류은환 [팀원]

YOLOv9 논문 실험 구현
YOLOv9 논문 분석
YOLOv9 코드 분석
파라미터 분석
경량화 모델 설계 및 실험
PTQ(Post Training Quantization)

199시간

팀원별 역할분담

후처리 팀

신재환 [팀원]

Python->C++ 변환
Openvino 구동
Static Quantization

216시간

최승렬 [팀원]

Dataset class filtering 툴 제작
ONNX 모델 옵션 조합 성능 실험
ONNX to NCNN
ONNX to ORT
Pruning 실험

217시간

활동 기록

활동 기록

개강 이후
총 회의 횟수

52회

5월 이후
주간 평균 활동시간

35.5시간

2024 캡스턴 진행 일정 및 회의록

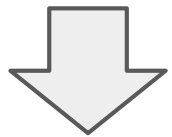
이름	날짜	시간	태그	회의 장소	생성자	최종 편집자
임베디드SW	2024년 6월 5일		진행 중	오프라인	서태원	승렬 최
교수님 미팅	2024년 5월 23일	16:00 – 17:30	완료	오프라인	서태원	서태원
3-8차 회의	2024년 5월 23일	10:30 – 00:00	완료	오프라인	서태원	승렬 최
3-7차 회의	2024년 5월 22일	14:30 – 23:45	완료	오프라인	서태원	승렬 최
3-6차 회의	2024년 5월 21일	18:00 – 00:00	완료	오프라인	서태원	승렬 최
3-5차 회의	2024년 5월 20일	15:00 – 00:10	완료	오프라인	은황 류	승렬 최
3-4차 회의	2024년 5월 19일	15:00 – 22:30	완료	오프라인	승렬 최	승렬 최
3-3차 회의	2024년 5월 18일	15:00 – 23:30	완료	오프라인	은황 류	승렬 최
3-2차 회의	2024년 5월 16일	15:00 – 00:45	완료	오프라인	서태원	승렬 최
3-1차 회의	2024년 5월 15일	14:00 – 22:30	완료	오프라인	은황 류	승렬 최
2-37차 회의	2024년 5월 14일	15:00 – 00:30	완료	오프라인	은황 류	승렬 최
2-36차 회의	2024년 5월 13일	19:30 – 22:30	완료	오프라인	은황 류	승렬 최
2-35차 회의	2024년 5월 12일	18:20 – 22:30	완료	오프라인	서태원	승렬 최
2-34차 회의	2024년 5월 10일	17:00 – 22:30	완료	오프라인	은황 류	승렬 최
2-33차 회의	2024년 5월 9일	18:30 – 20:30	완료	디스코드	승렬 최	승렬 최
2-32차 회의	2024년 5월 7일	13:00 – 21:30	완료	오프라인	승렬 최	승렬 최
2-31차 회의	2024년 5월 6일	18:00 – 23:00	완료	오프라인	승렬 최	승렬 최
2-30차 회의	2024년 5월 3일	13:00 – 16:00	완료	오프라인	계희 무	승렬 최
2-29차 회의	2024년 5월 2일	15:00 – 21:30	완료	오프라인	승렬 최	승렬 최
2-28차 회의	2024년 5월 1일	18:30 – 23:00	완료	오프라인	은황 류	승렬 최
2-27차 회의	2024년 4월 30일	18:00 – 23:00	완료	오프라인	승렬 최	승렬 최
2-26차 회의	2024년 4월 29일	16:00 – 23:00	완료	오프라인	은황 류	은황 류
2-25차 회의	2024년 4월 26일	15:00 – 22:30	완료	오프라인	은황 류	승렬 최
2-24차 회의	2024년 4월 25일	14:00 – 23:00	완료	오프라인	승렬 최	승렬 최

향후 계획

TO-BE

추가 실험을 통해 최적화 모델 성능 향상

실험 내용을 바탕으로 졸업 논문 작성



응용 분야 설정
특화 개발

분야 세분화 및 실험 정밀화 하여 KCI급 논문 투고

해외 학술지 논문 분석 및 학술지 논문 투고

Thank you

무[遡]계획

인천대학교 INCHEON NATIONAL UNIVERSITY | 임베디드시스템공학과