무[邈]계획

서태원, 신재환, 최승렬, 류은환

인천대학교 INCHEON NATIONAL UNIVERSITY | 임베디드시스템공학과

Contents

- Capstone Design 개요
- Capstone Project 진행도
- To-Be
- Q&A

Capstone Design 개요

딥러닝 모델의 **최적화**를 통한

임베디드보드에서의 **추론 속도 향상**

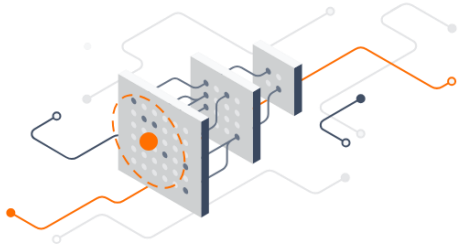
Capstone Design 개요

연구의 필요성

딥러닝 모델은 고성능의 GPU 또는 서버 등 **별도의 자원**이 필요

저사양 임베디드보드 환경에서 성능의 한계로 딥러닝 모델을 활용하기 어려움

따라서, **온디바이스 AI를 구현하기 위해 딥러닝 모델의 경량화 필요**



Capstone Design 개요

연구의 목표

기존 YOLOv9 (논문 주장) 모델 대비

Inference time 10배 향상, *mAP 15%p 미만 하락

*mean Average Precision

Model	APval	AP50val	RPI Inference time
YOLOv9-C	**53.0%**	**70.2%**	**8908.2ms**

YOLOv9

2024년 2월 발표한 최신 객체 탐지 모델로 YOLOv7의 저자 Chien-Yao Wang가 공개함

Implementation of paper - YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information

프로젝트 진행도

회의 기록

총 회의 횟수(개강 이후): 40 회

주당 평균 회의 시간: 20.7 시간

월 평균 회의 시간: 82.8 시간

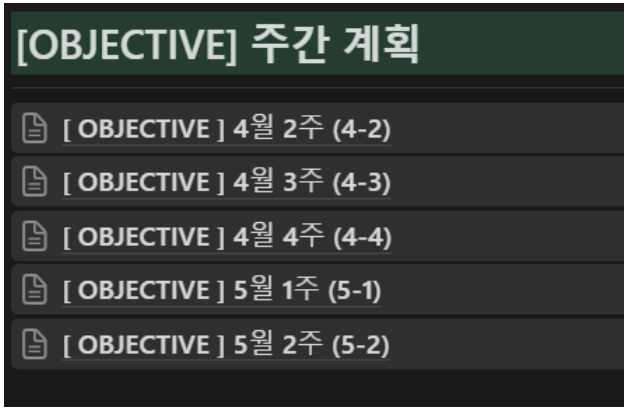
2024 캡스턴 진행 일정 및 회의록

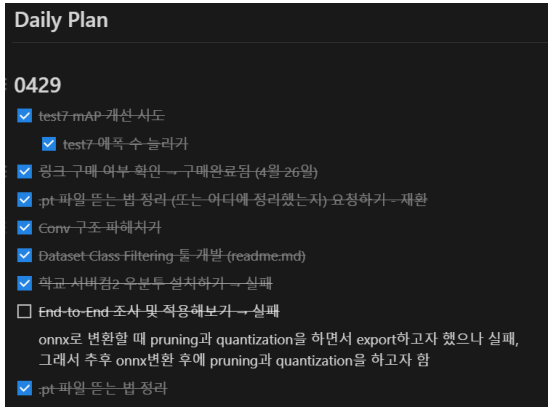
이름	날짜	시간	태그	회의 장소	생성자	최종 편집자
2-33차 회의	2024년 5월 9일	18:30 ~ 20:30	완료	디스코드	승렬 최	승렬 최
2-32차 회의	2024년 5월 7일	13:00 ~ 21:30	완료	오프라인	승렬 최	승렬 최
2-31차 회의	2024년 5월 6일	18:00 ~ 23:00	완료	오프라인	승렬 최	승렬 최
2-30차 회의	2024년 5월 3일	13:00 ~ 16:00	완료	오프라인	계획 무	승렬 최
2-29차 회의	2024년 5월 2일	15:00 ~ 21:30	완료	오프라인	승렬 최	승렬 최
2-28차 회의	2024년 5월 1일	18:30 ~ 23:00	완료	오프라인	은환 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	은환 류	은환 류
2-25차 회의	2024년 4월 26일	15:00 ~ 22:30	완료	오프라인	은환 류	승렬 최
2-24차 회의	2024년 4월 25일	14:00 ~ 23:00	완료	오프라인	승렬 최	승렬 최
2-23차 회의	2024년 4월 16일	18:00 ~ 22:30	완료	오프라인	승렬 최	은환 류
LINC 신청마감	2024년 4월 16일		완료	오프라인	서태원	승렬 최
2-22차 회의	2024년 4월 15일	15:00 ~ 22:30	완료	오프라인	재환 신	서태원
2-20차 회의	2024년 4월 11일	14:00 ~ 22:00	완료	오프라인	은환 류	승렬 최
2-19차 회의	2024년 4월 10일	14:00 ~ 22:30	완료	오프라인	은환 류	승렬 최
2-18차 회의	2024년 4월 9일	18:30 ~ 22:00	완료	오프라인	은환 류	승렬 최
2-17차 회의	2024년 4월 8일	15:00 ~ 22:30	완료	오프라인	은환 류	승렬 최
2-16차 회의	2024년 4월 5일	16:00 ~ 23:00	완료	오프라인	은환 류	승렬 최
2-15차 회의	2024년 4월 4일	16:30 ~ 22:00	완료	오프라인	은환 류	승렬 최
2-14차 회의	2024년 4월 2일	19:00 ~ 23:00	완료	오프라인	은환 류	승렬 최
2-13차 회의	2024년 4월 1일	18:30 ~ 23:00	완료	오프라인	서태원	승렬 최
2-12차 회의	2024년 3월 29일	15:00 ~ 24:00	완료	오프라인	승렬 최	승렬 최
2-11차 회의	2024년 3월 28일	16:00 ~ 18:00	완료	오프라인	승렬 최	승렬 최
2-10차 회의	2024년 3월 26일	18:30 ~ 22:30	완료	오프라인	서태원	승렬 최
2-9차 회의	2024년 3월 25일	18:30 ~ 22:30	완료	오프라인	승렬 최	승렬 최
2-8차 회의	2024년 3월 23일	20:30 ~ 23:40	완료	디스코드	은환 류	승렬 최
2-7차 회의	2024년 3월 22일	09:30 ~ 11:00	완료	디스코드	은환 류	승렬 최
2-6차 회의	2024년 3월 21일	14:30 ~ 23:00	완료	디스코드	서태원	승렬 최
2-5차 회의	2024년 3월 20일	13:00 ~ 16:30	완료	오프라인	서태원	승렬 최
2-4차 회의	2024년 3월 19일	17:00 ~ 21:00	완료	오프라인	서태원	승렬 최

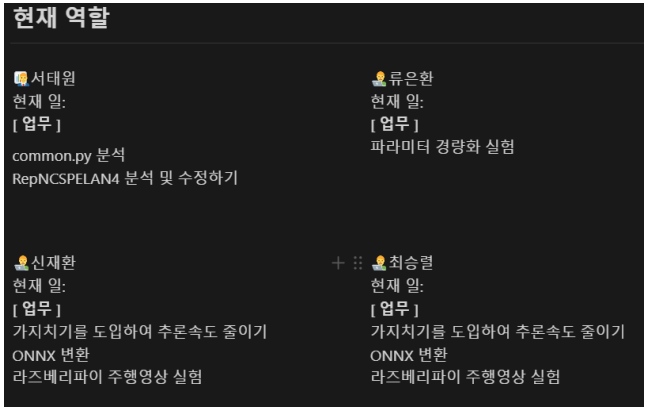
개별 기여 성과 - 서태원

역할

팀장, Architecture Part



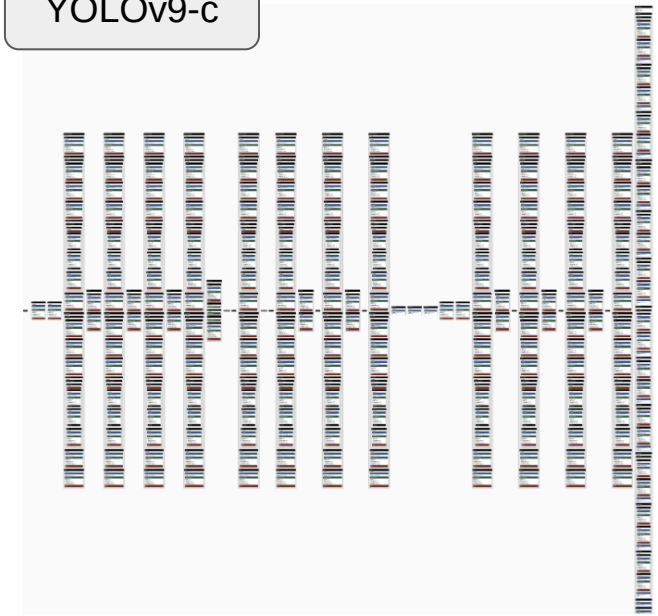




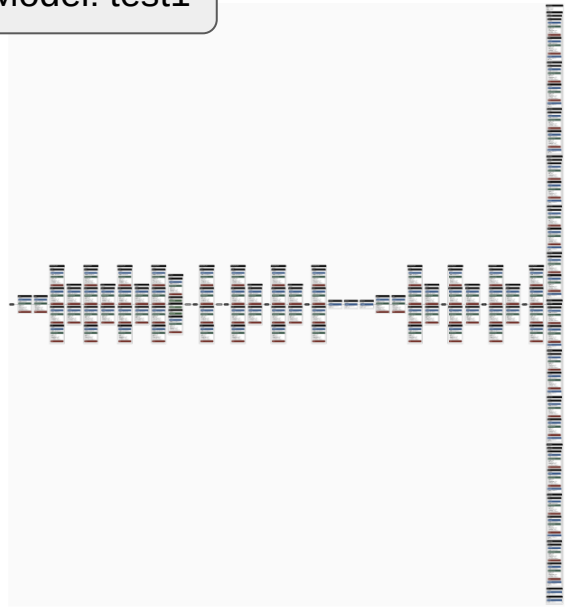
프로젝트 일정 관리, 주간 목표 수립, 개발 회의 진행, 역할 분배

개별 기여 성과 - 서태원

YOLOv9-c



Model: test1



Netron 도구를 이용하여 아키텍처 시각화 및 분석 수행

개별 기여 성과 - 서태원

공부

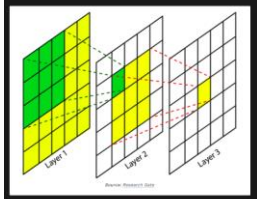
- What is Convolution?
- What is Convolution Layer?
- What is Optimizer?
- What is mAP?

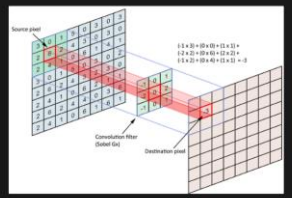
Convolutional Layer는 CNN에서 입력 데이터의 특징(feature)을 추출하는 레이어이다.

이미지 같은 2차원 데이터에서는 필터(커널)를 사용하여 입력 데이터와 필터 간의 합성곱 연산을 수행하는데,

예를 들어, 3x3 크기의 필터를 사용하면 입력 데이터의 3x3 부분과 필터 간의 합성곱이 이루어진다.

학습 전에 필터는 무작위로 초기화되며, loss 계산을 통해 최적의 필터 파라미터를 찾아가는 과정이 딥러닝에서 말하는 학습이라고 보면 된다.





- common.py - autopad
- common.py - Conv
- common.py - RepNCSPELAN4
- common.py - RepNBottlenec

RepNCSPELAN4 함수 구조 분석

초기화 메소드 (`__init__`)

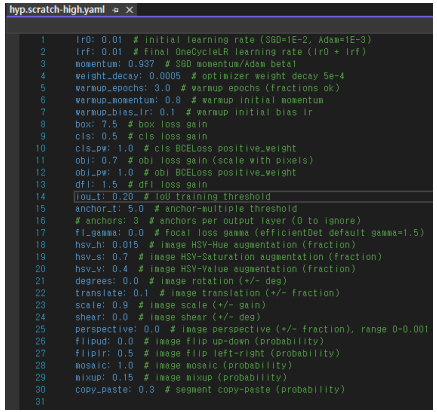
- `self.cv1` : `c1` 입력 채널에서 `c3` 채널로 매핑하는 `Conv` 레이어
  이 레이어는 입력 데이터의 차원을 조정
  - `cv1` = Conv(c1, c3, 1, 1)
- `self.cv2` 및 `self.cv3` : 두 개의 `nn.Sequential` 객체로 구성되며, 각각 `RepNCSP` 레이어를 포함
  `RepNCSP` 는 입력 채널을 절반으로 나눈 `c3//2` 에서 시작하여 `c4` 로 출력을 매핑
  이 과정은 반복적으로 수행되어 더 깊은 특징을 추출함
  - `cv2` = Sequential
    - RepNCSP(c3//2, c4, c5)
    - Conv(c4, c4, 3, 1)
  - `cv3` = Sequential
    - RepNCSP(c4, c4, c5)
    - Conv(c4, c4, 3, 1)
- `self.cv4` : 최종 `Conv` 레이어로, 모든 중간 층의 결과를 종합하여 `c2` 출력 채널로 매핑
  이 레이어는 최종적인 출력 형태를 결정
  - `cv4` = Conv(c3+(2*c4), c2, 1, 1)

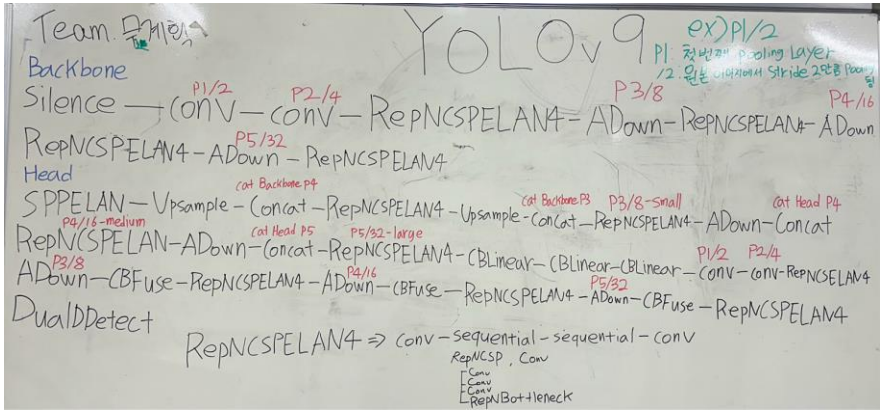
딥러닝 모델 이해를 위한 배경 공부 수행
아키텍처 개발을 위한 함수 구조 분석 수행

개별 기여 성과 - 류은환

역할

팀원, Architecture Part





YOLOv9 아키텍처 분석

파라미터 경량화 실험
(레이어 크기 조절)

하이퍼 파라미터 실험

개별 기여 성과 - 신재환

역할

팀원, Post-process & Simulation Part

	Original Model	Simplified Model
Add	15	2
AveragePool	5	5
Concat	22	22
Constant	220	141
Conv	65	65
Div	14	1
Gather	13	0
MaxPool	8	8
Mul	85	59
Reshape	5	5
Resize	2	2
Shape	13	0
Sigmoid	59	59
Slice	26	26
Softmax	1	1
Split	2	2
Sub	2	2
Transpose	2	2
Model Size	10.1MiB	10.0MiB

학습모델을 ONNX로 모델변환 & prunning

ONNX Runtime Quantization

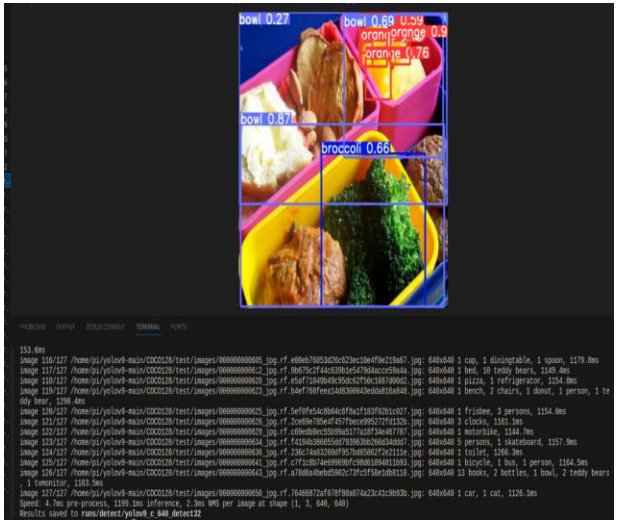
OpenVINO, TFLite Quantization

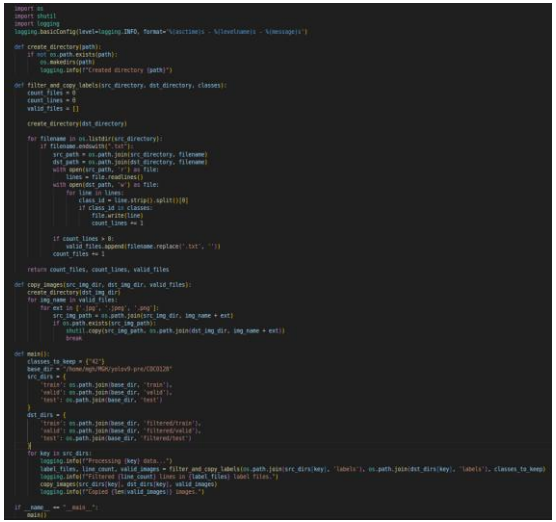
Raspberry pi에서 후처리 모델 simulation

개별 기여 성과 - 최승렬

역할

팀원, Post-process Part





Raspberry pi에서 후처리 모델 mAP, inference time 측정 및 비교

Dataset Class filtering 코드 제작

ONNX 모델을 NCNN으로 변환

실험 결과 및 성과정리

Model	APval	AP50val	RPI Inference time
YOLOv9-C	**53.0%**	**70.2%**	**8908.2ms**

Model: test5

RepNCSPELAN4의 Sequential구조를 Conv로 변경
multi-level reversible auxiliary branch 삭제
레이어 최대 크기를 128로 고정

mAP50: 0.505
mAP50-95: 0.352
RPI Inference time: 1591.4ms

Model: test5_ONNX

test5모델을 정적그래프를 사용하는 ONNX 형식으로 변경
연산자 노드와 constant 노드 pruning
float 32를 float 16으로 quantization

mAP50: 0.504
mAP50-95: 0.351
RPI Inference time: 608.8ms

To-Be

실험 방향

아키텍처 분석 및 최적의 레이어 크기와 레이어 수를 실험을 통해 탐색
다양한 후처리 방법 조사 및 적용 실험 진행

논문 작성

개발 모델 실험 결과 정리 및 분석
YOLOv9 경량화 논문 작성

Q&A

Thank you