2025 Python Week

Single Shot Multibox Detector 원복 및 Kaist PD 데이터 기반 성능 향상

SMARCLE 김기현, 이예은

2025. 8. 6

SEJONG UNIVERSITY

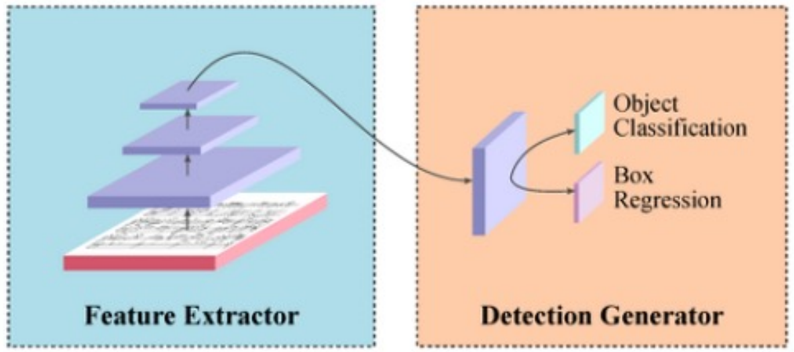
1-stage detector & 2-stage detector

Object detection의 두 단계

1. 물체가 있을 만한 경계 상자(RoI)를 추출하는 단계 (Region Proposal)
2. 추출된 RoI를 이용하여 Localization, Classification을 수행

1-stage detector

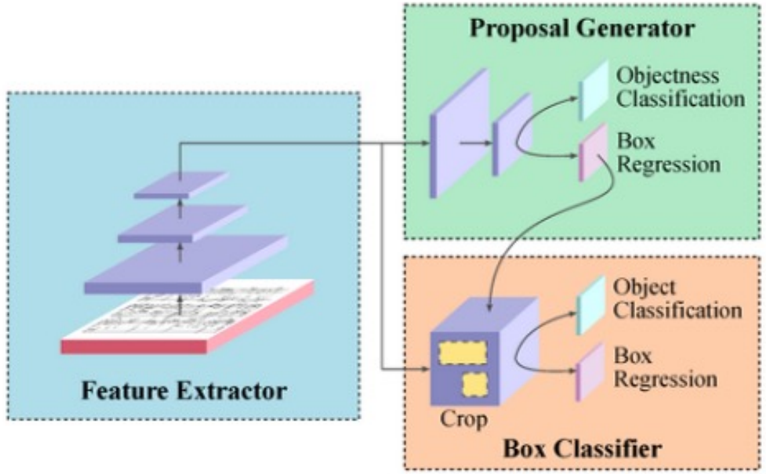
- 두 단계를 한꺼번에 수행
- e.g., YOLO, SSD



(a) Basic architecture of a one-stage detector.

2-stage detector

- 두 단계를 구분하여 수행
- e.g., R-CNN



(b) Basic architecture of a two-stage detector.

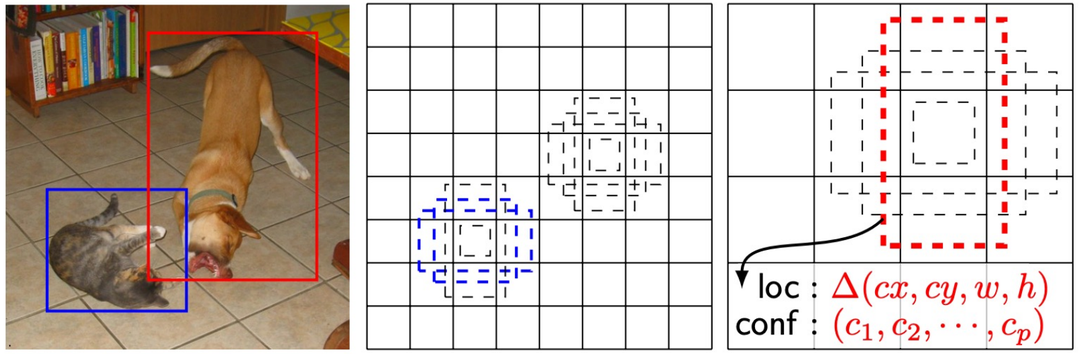
SSD의 등장 배경과 핵심

등장 배경

- 기존 Faster R-CNN은 정확도는 높지만 속도가 느려 real-time에서 사용할 수 없었음.
- YOLO(v1)는 2-stage를 1-stage로 줄여 속도가 크게 개선되었지만, 정확도가 낮다는 단점이 있었음.
- 빠른 속도를 가지면서, 정확도도 높은 모델이 바로 SSD임.

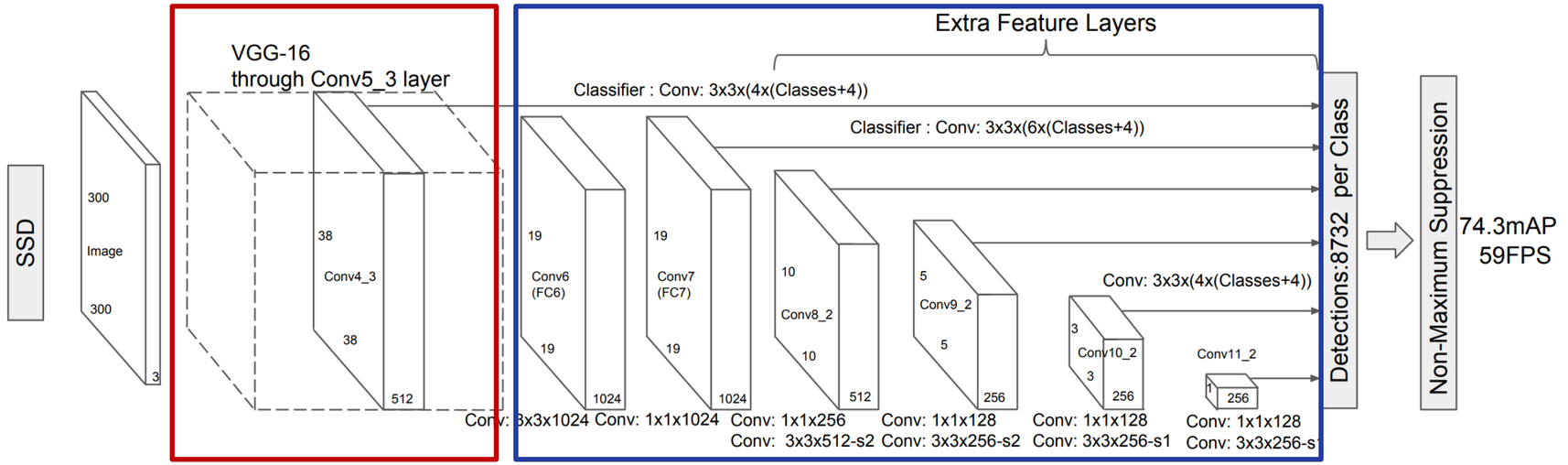
SSD 핵심

- 여러 feature map 각각에 convolution filter를 적용함으로써 다양한 크기의 box에 대한 category와 offset을 예측한다.
- 앞단에서 뒷단으로 갈수록 reception field가 커지고 high-level의 feature를 가지므로, 더 큰 물체를 예측하기 좋다.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

SSD의 구조



base network

auxiliary structure

Base Network

- pretrained VGG16을 사용
- 분류를 위한 마지막 FC layer는 제거
- 두 개의 FC layer는 convolutional layer로 대체

Auxiliary Structure

- base network 뒤에 추가적인 convolution을 여러 번 수행하여 higher-level의 feature map을 얻음
- feature map의 크기가 점점 감소하여, 다양한 크기의 객체를 탐지할 수 있음

Prediction Convolution

- 6개의 feature map 각각에서, category와 box offset을 예측

Kaist PD(Pedestrian Detection) Dataset

보행자 검출을 위한 RGB와 TheramI 이미지 쌍 dataset

- RGB 이미지만으로 보행자를 검출하는 데 한계가 있어, thermal(열화상) 이미지 pair를 사용
  - color는 조도가 있는 환경에서 검출을 잘하지만 조도가 없을 때(밤, 그늘)는 검출을 거의 못함
  - thermal은 기온과 체온의 차이가 큰 밤에 검출을 잘함
- Beam Splitter를 통해 들어오는 초점의 위치를 같게 해줌

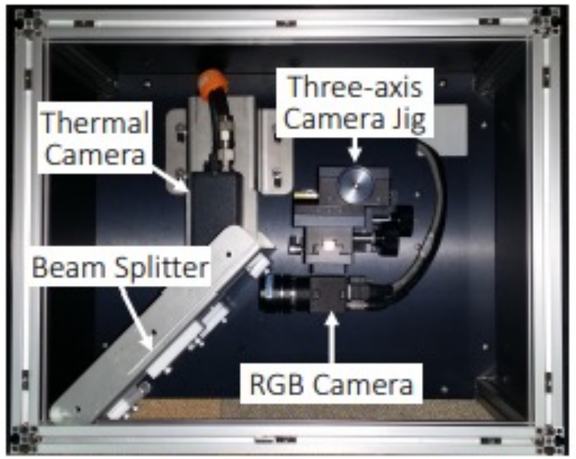
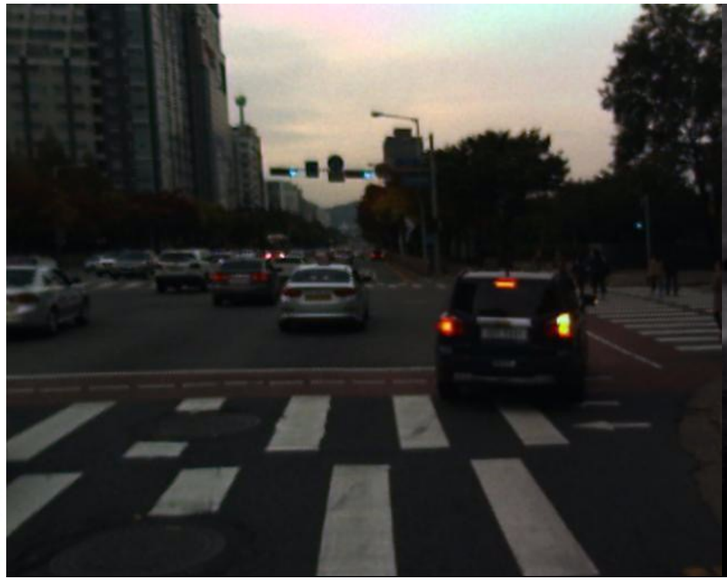


Figure 2. Our hardware configuration for capturing multispectral images. (Left) Top view. (Right) Mounted on the rooftop of a car.



Kaist PD SSD 원복 성능

RGB, Thermal 각각에 대한 원복

- 두 이미지 쌍을 fusion 하기 이전에, 각각에 대한 SSD 모델의 detection 성능을 확인
- RGB는 day에 강하고, thermal은 night에 매우 강함

RGB

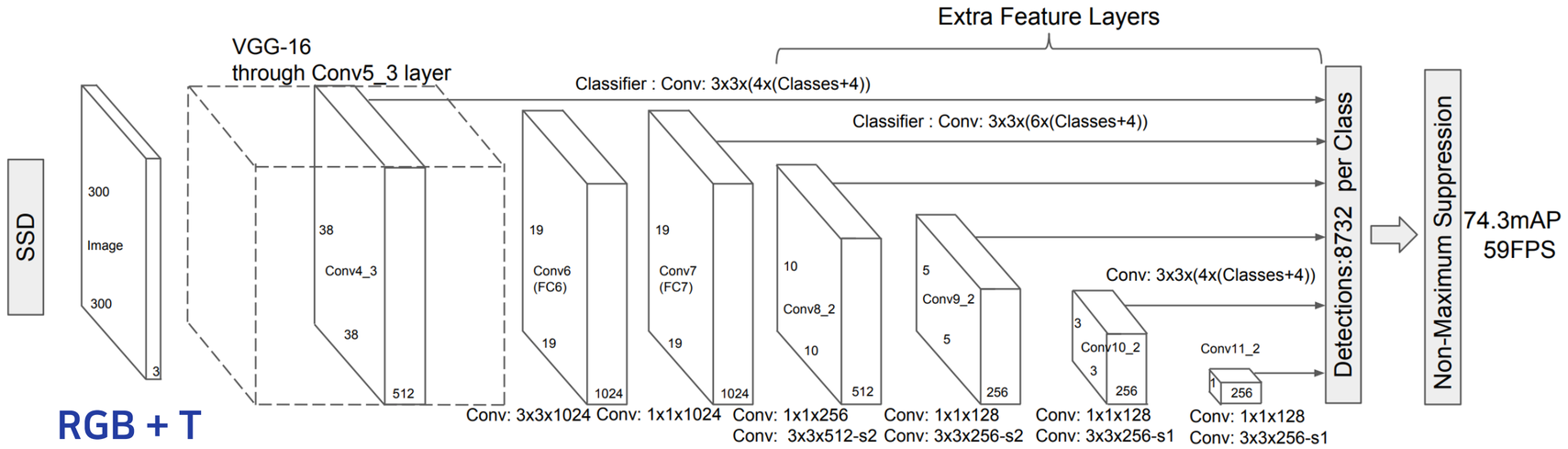
	MR (all)	MR (Day)	MR (Night)	Recall
원복	35.36	32.92	41.79	79.94

Thermal

	MR (all)	MR (Day)	MR (Night)	Recall
원복	32.08	39.27	17.37	82.30

>> RGB, Thermal을 fusion하여 성능을 향상하는 것이 목표

4 channel input fusion

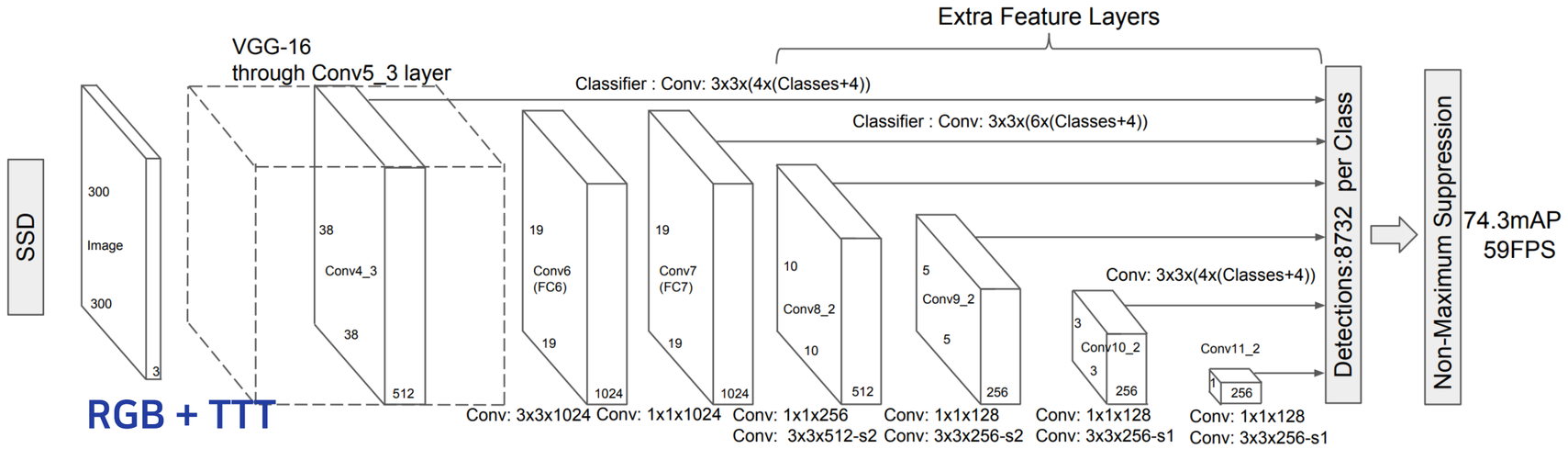


RGB 3 channel + Thermal 1 channel

- Pillow image가 4 channel까지 지원하는 것을 확인
- 3 channel과 유의미한 성능 차이가 존재하지는 않음

version	MR	MR(Day)	MR(Night)	Recall
4channel	36.10	41.71	24.90	84.28

6 channel input fusion



RGB 3 channel + Thermal 3 channel

- PIl은 4채널까지 받기 때문에 transform에서 따로 동작하여 합체
- 예상대로 MR이 4 channel에 비해 감소하는 모습을 보임

version	MR	MR(Day)	MR(Night)	Recall
4channel	36.10	41.71	24.90	84.28

Using Mask channel

RGB 3 channel + Thermal 3 channel + Mask 3 channel

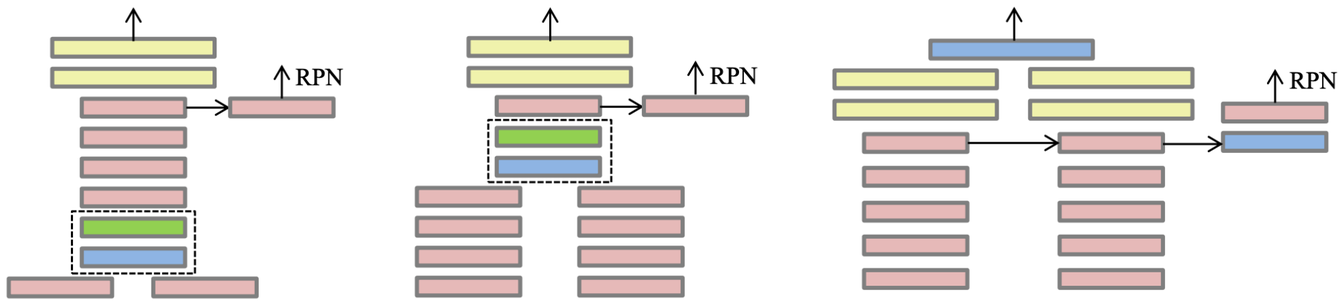
- 어두운 RGB를 학습시키지 않는 것이 목적
- Mask 처리를 통해 학습에 영향을 감소

version	MR	MR(Day)	MR(Night)	Recall
6channel	31.46	35.35	23.96	83.44
RGB T T	30.22	35.91	18.76	82.96
RGB T Mask	29.96	34.51	20.89	82.47

Multispectral Deep Neural Networks for Pedestrian Detection

가장 간단한 fusion 구조에 대한 논문

backbone: Faster R-CNN



1) Early Fusion

- Conv1 이후 두 피쳐맵을 concate

2) Halfway Fusion

- Conv4 이후 두 피쳐맵을 concate

3) Late Fusion

- F7을 연결

4) Score Fusion

- RGB, Thermal 각각의 detection score를 합함

Multispectral Deep Neural Networks for Pedestrian Detection

version	MR	MR(Day)	MR(Night)	Recall
Halfway fusion	23.40	26.23	16.97	88.45
Late fusion	21.35	23.82	16.35	88.01
6channel input	31.46	35.35	23.96	83.44

Experiment

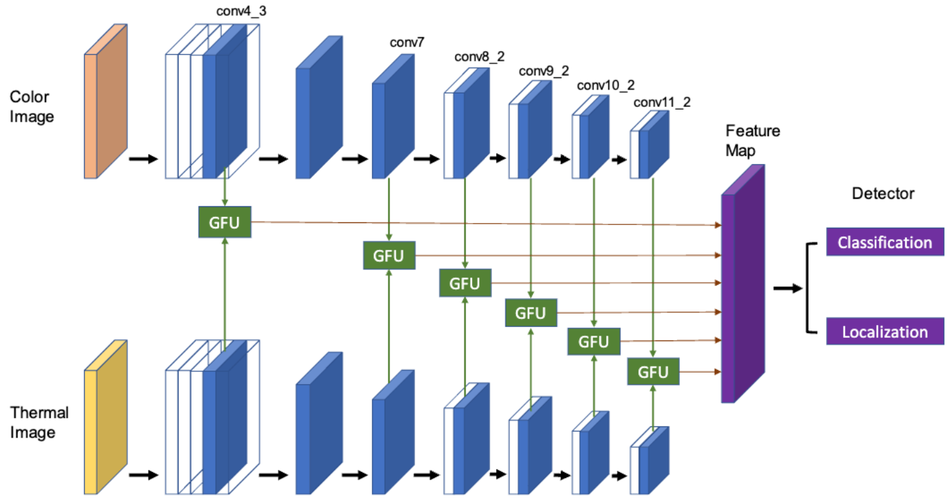
- SSD에 적합한 fusion 방식은 Late fusion으로 정의할 수 있었음

- 이처럼 두 채널을 단순히 concat하지 않고, 둘의 조합을 학습할 수는 없을까?

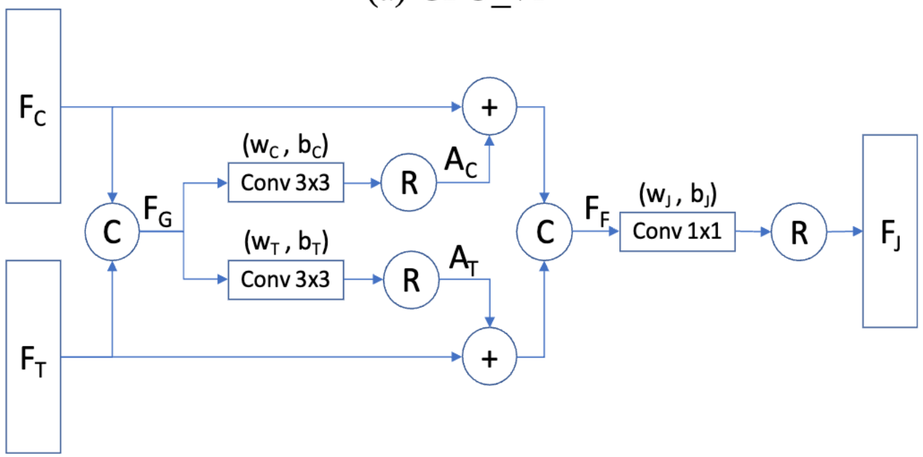
GFD-SSD: Gated Fusion Double SSD for Multispectral Pedestrian Detection

- SSD의 detection에 사용되는 6개의 feature map 각각을 독립적으로 fusion
- Gate 구조 사용

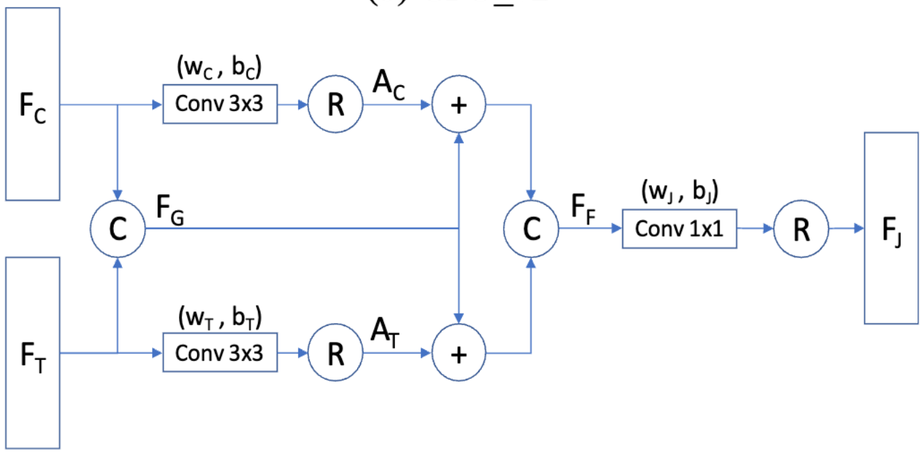
(b) GFD-SSD: Gated Fusion Double SSD



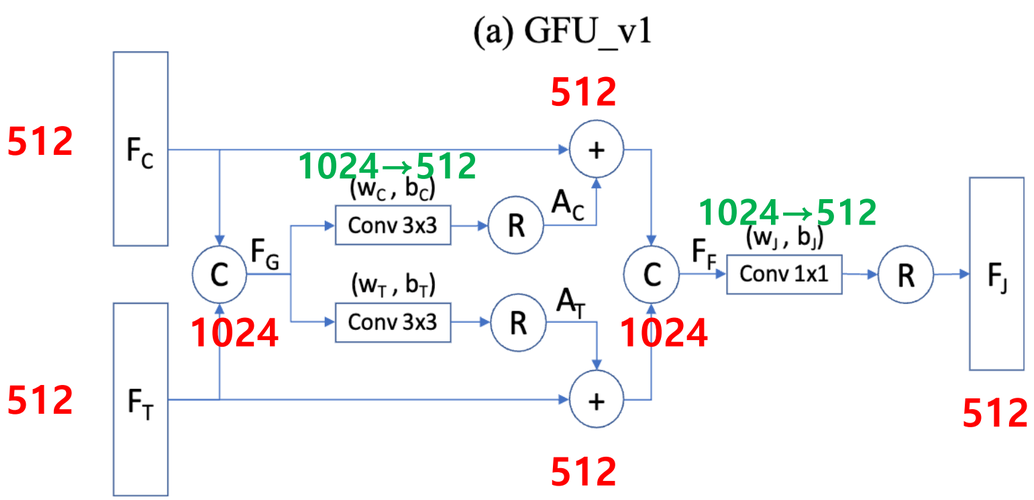
(a) GFU_v1

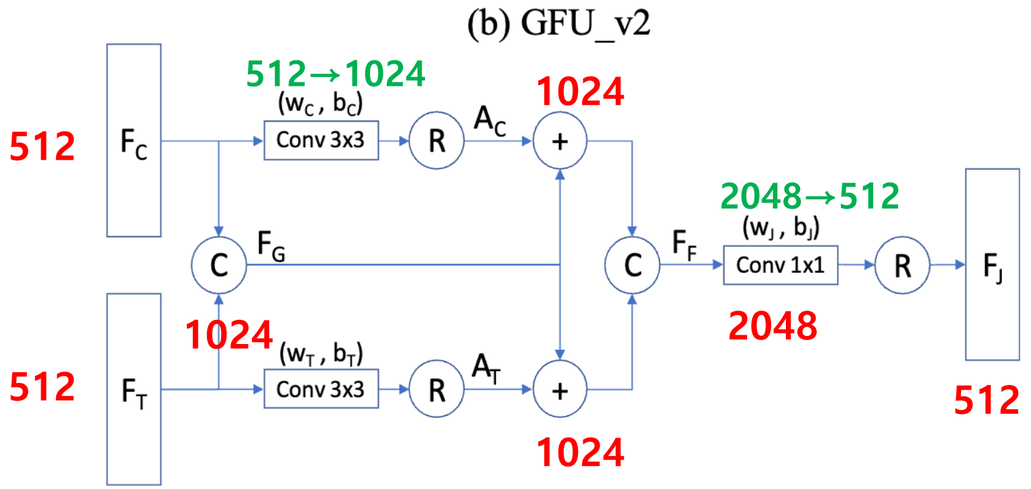


(b) GFU_v2



GFD-SSD: Gated Fusion Double SSD for Multispectral Pedestrian Detection





- concat한 feature map에 3x3 conv 적용
- color와 thermal의 고유한 feature를 최대한 반영하면서, concat한 feature에서 semantic한 정보를 뽑음

- 각각의 feature map에 3x3 conv 적용
- color와 thermal 각각을 바로 semantic하게 만들고, concat한 feature를 합함

GFD-SSD: Gated Fusion Double SSD for Multispectral Pedestrian Detection

	MR (all)	MR (Day)	MR (Night)	Recall
Baseline (RGB)	35.36	32.92	41.79	79.94
Baseline (Thermal)	31.54	38.1	18.28	83.14
GFD-SSD (v1)	**23.44**	**26.82**	**16.4**	**86.37**
GFD-SSD (v2)	**24.83**	**27.83**	**18.08**	**83.92**

Model	*Inputs*	*Fusion*	*logMR (%)*
Single SSD300	Color	-	34.69
	Thermal	-	41.79
Double SSD300	Color + Thermal	Stack	29.99
	Color + Thermal	GFU_v1	30.42
	Color + Thermal	GFU_v2	30.51
Single SSD512	Color	-	32.81
	Thermal	-	39.47
Double SSD512	Color + Thermal	Stack	30.29
	Color + Thermal	GFU_v1	28.84
	Color + Thermal	GFU_v2	**28.10**

Table 2: Result summary on KAIST pedestrian detection, for comparisons of SSD300 vs. SSD512, and stack fusion vs. gated fusion v1 vs. gated fusion v2.

- 논문 결과보다 v1과 v2의 성능 차이가 컸음
- 모델 구현의 차이, 데이터셋 annotation의 차이 등이 원인 -> 실험의 중요성!

느낀점

진행 과정 중

- GPT를 사용하지 않고 논문과 블로그만을 참고해서 진행하다보니 어려운 점이 많았음
- Import pdb의 중요성
- 사람과 DNN이 이미지를 보는 방식이 다르다는 것을 고려해서 논문의 내용을 참고함
- Pytorch 공식 문서를 적극적으로 활용해 모델 구조 개선에 활용했다

결과

- 단순한 이미지 전처리를 통해서는 비약적인 성능 향상을 이루어내기 어렵다고 판단
- 모델을 수정하여 MR이 떨어지는 것을 확인했을 때 성취감이 있었다
- 의견을 공유하는 과정을 통해 여러가지 방법론에 대해 이야기 해보고 협력하여 좋은 결과를 만들어낼 수 있었다.