2019 2학기 석사 졸업논문 발표

동적으로 클래스 변경이 가능한 개인화된 퓨-샷 객체 검출

이성우 석사과정

인간중심컴퓨팅 연구실

2019.12.30

HUMAN
CENTERED
COMPUTING
LABORATORY

목차

연구 배경

- 인공지능 모델이 서비스화 되면서 소비자들의 요구가 다양해지고 개인화되고 있다

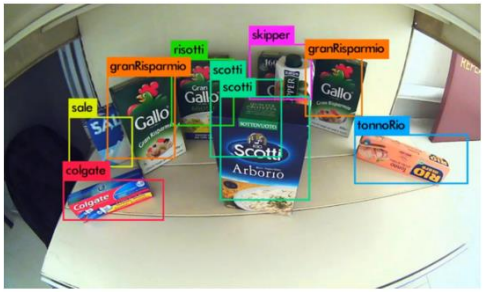
amazon Rekognition

Google Cloud AutoML Vision

Microsoft Cognitive Services

- 객체 검출(Object detection)분야도 **서비스** 사용자에게 개인화된 모델들이 요구됨
  - 수시로 새로운 제품 추가, 나만의 물건 등록 …

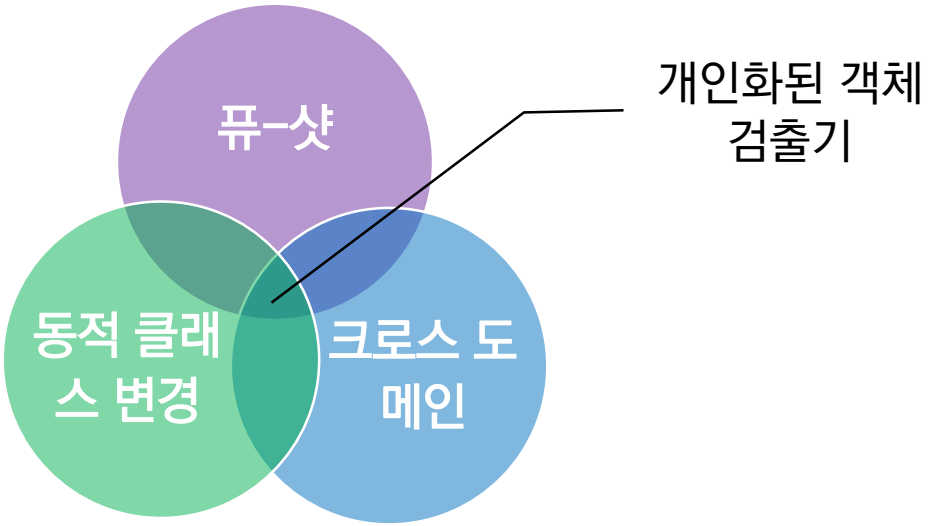
AWS 딥렌즈



무인 점포

연구 문제

- 개인화된 객체 검출기를 만들기 위해선 **퓨-샷, 동적 클래스 변경, 크로스 도메인 적용** 세가지 조건을 모두 만족하는 검출기가 필요
  1. 퓨-샷 – 새로운 클래스 추가 시 필요로 하는 데이터는 적어야 한다
  2. 동적 클래스 변경 – 새로운 클래스 추가 시 추가학습 없이도 즉시 모델에 적용할 수 있어야 한다.
  3. 크로스 도메인 적용 – 모델의 추론 단계에 사용할 수 있는 데이터셋의 범주가 훈련용 데이터 셋의 범주에 한정되지 않고 자유롭게 변경 가능해야 한다. 반례) 얼굴인식



연구 문제

- 기존 검출기 모델들의 한계
  - 기존 모델들은 **고정된 클래스**에서의 성능 증가와 수행시간을 줄이는 데 초점을 맞춤
  - 최근 퓨-샷 검출기가 연구되고 있으나 클래스 변경 시 추가 학습이 필요
  - 때문에 앞의 세가지 조건이 동시에 가능한 객체 검출 모델이 없음

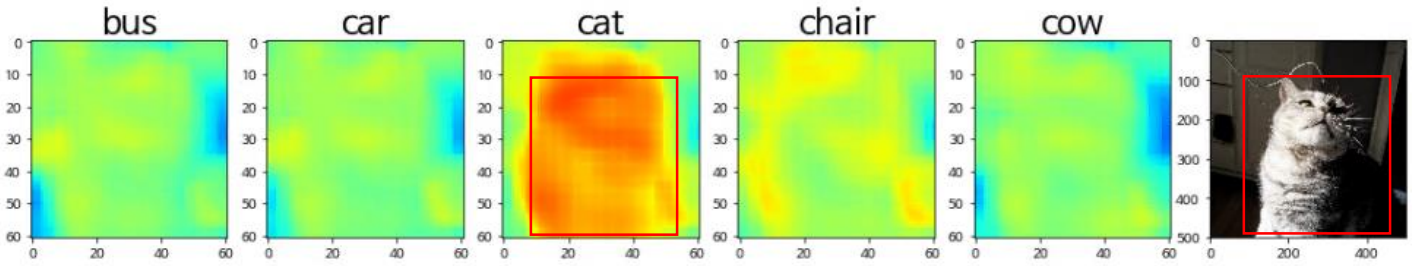
Task	Model	퓨-샷	동적 클래스 변경	크로스 도메인 적용
Object Detection	Faster R-CNN, YOLO, SSD	x	x	o
Face recognition	FeacNet, DeepFace	o	o	x
Metric learning	Siamese Networks, Prototypical Networks	o	o	x
Few-shot recognition	Low-shot Visual Recognition by Shrinking and Hallucinating Features	o	x	o
	Dynamic Few-Shot Visual Learning without Forgetting	o	o	o
Few-shot Object Detection	LSTD	o	x	o
	RepMet	o	x	x
	ours	o	o	o

기존 모델들의 가능 요구조건 정리 표

연구 목표

- 실제 여러 도메인에서의 합리적인 성능을 보장하는 **개인화된 객체 검출기** 모델을 개발
- 객체 검출기 모델이 데이터셋의 변화나 모델의 구조변경 등에 어떻게 작동하는지 분석

- 연구 아이디어
  - 이미지로부터 클래스 독립적인 CAM(Class Activation Map) 생성
    - **이미지넷 분류 데이터에 학습된 CNN 모델의 텍스처 편향*** 성질을 이용
  - 생성된 CAM만을 이용해 물체의 클래스 분류와 위치 예측



각 CAM이 원래 자신의 클래스에서 뚜렷히 높은 스코어를 보여줌 (PASCAL VOC 데이터셋)

*이미지넷 데이터에 훈련된 CNN Backbone들이 shape보다 texture를 보고 클래스를 판단한다

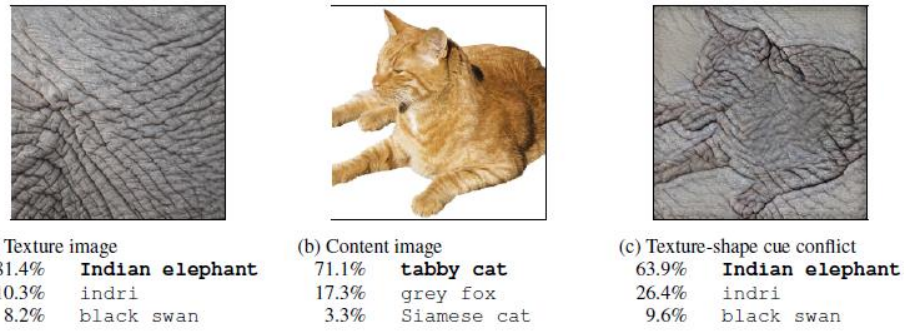
시스템 프로시저 요약

1. 이미지를 입력으로 받아서 CAM을 만든다
   - CAM은 참조하는 주변영역의 크기를 달리하여 6가지 채널로 만든다
   - CAM은 모든 클래스별로 만든다
   - 최종적으로 채널 수 x 클래스 수 만큼의 CAM을 생성
2. 생성된 CAM들을 모두 한번에 검출기에 입력한다
3. 가장 예측 확률이 높은 클래스를 고르고 그 클래스에 해당하는 경계박스(bbox*) 좌표를 이용해 박스를 그린다

*bbox – Bounding box : 물체가 포함된 최소 사각 박스영역

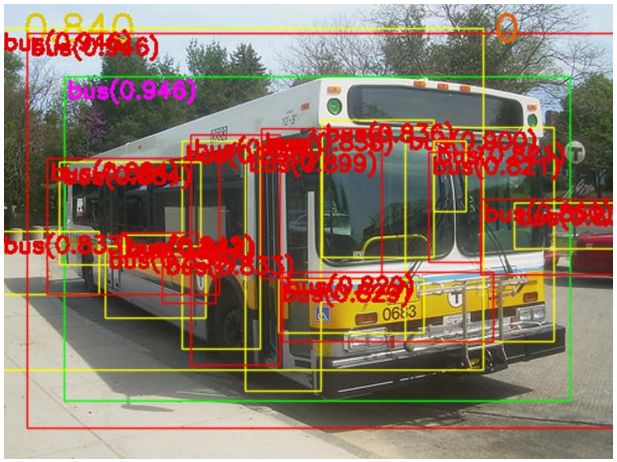
이미지넷 분류 데이터의 텍스처 편향

- 이미지넷 데이터에 훈련된 CNN Backbone들이 shape보다 texture를 보고 클래스를 판단한다
- CNN의 receptive field를 9x9로 제한해도 70%이상의 정확도를 보여줌
- *본인의 연구에선 위의 성질을 이용해 이미지에서 국소 영역만 참조하여 CAM의 activation score를 계산*



고양이 이미지에 코끼리 피부 텍스처를 style transfer했더니 코끼리로 판단

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." ICLR 2019 Oral



프로포절 방법(이미지 피처를 그대로 사용)대로면 이미지의 일부만 보고도 검출영역으로 잡아내는 부분이 매우 많음

약한 지도학습 기반의 객체 검출

- CAM(Class Activation Map)을 이용한 객체 검출 수행 가능
  - 오로지 분류용 데이터로만 수행 가능하기에 학습용 Bounding box 정보가 불필요
  - *본 연구에서도 CAM을 이용해 객체 검출 수행. 그러나 CAM을 생성하는 방법이 다르고 클래스의 개수를 동적으로 변경가능*

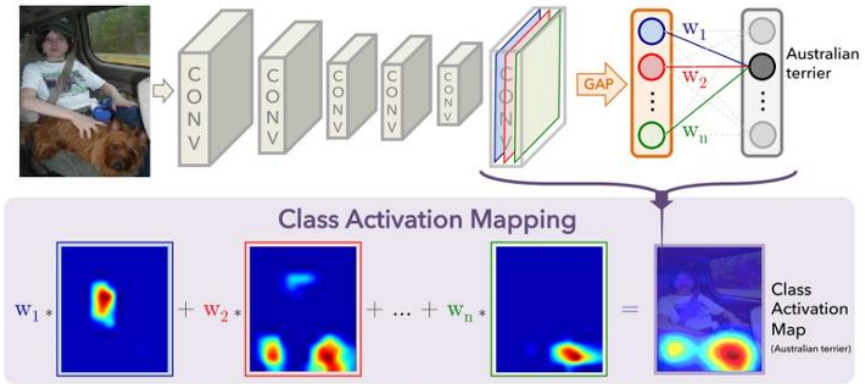
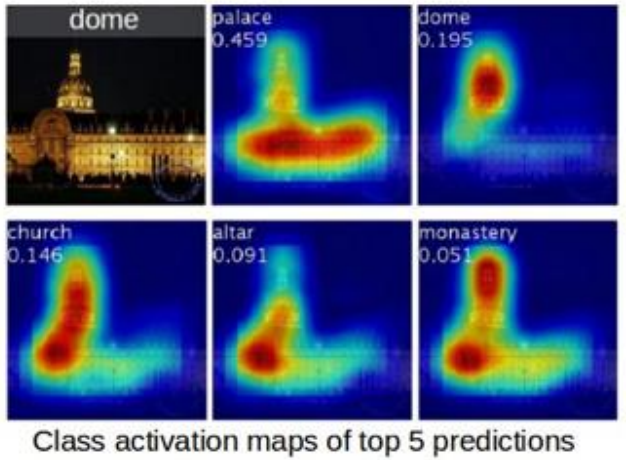


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.



Class activation maps of top 5 predictions

Zhou, Bolei, et al. "Learning deep features for discriminative localization." CVPR 2016.

RoI Align

- 여러 관심영역(RoI*)의 특징맵(Featuremap)들을 만들기 위해 전체 이미지의 특징맵을 한번만 만들어 두고, 관심영역(RoI)들의 위치에 해당하는 부분을 특징맵에서 잘라옴
  - FCN(Fully Convolutional Network)모델이 특징맵의 위치정보를 보존하는 특성을 이용해 시간을 매우 많이 절약할 수 있음
- *본 연구에선 하나의 CAM생성에 약 2만개 이상의 부분 특징맵들을 생성해야 하는데 RoI align을 이용해 한번의 CNN 수행만으로 가능해짐*

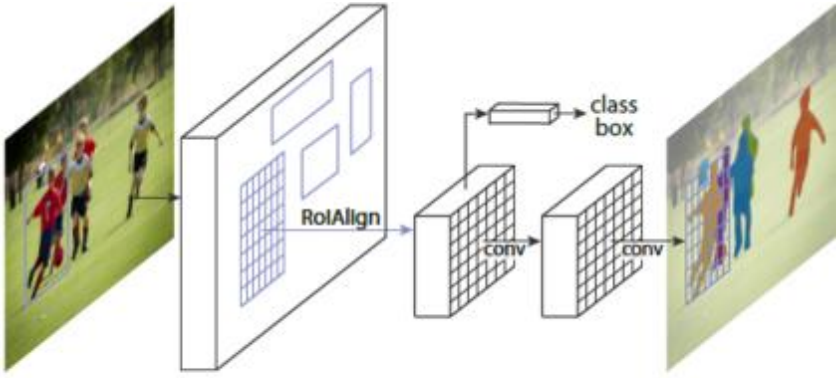


Figure 1. The Mask R-CNN framework for instance segmentation.

Mask R-CNN, He, Kaiming, et al., ICCV 2017

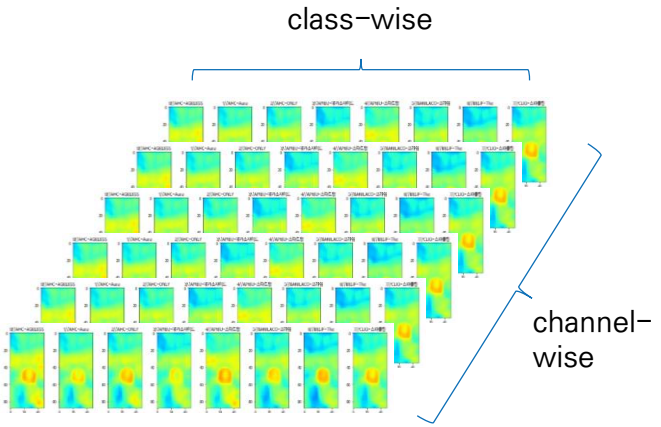
RoI align 작동 원리

*RoI – Region of Interest

모델 시스템 요약

CAM
생성기

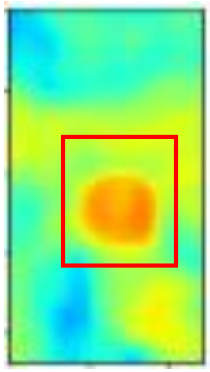
클래스 서포트 이미지와의
코사인 유사도를
Activation Score로 사용

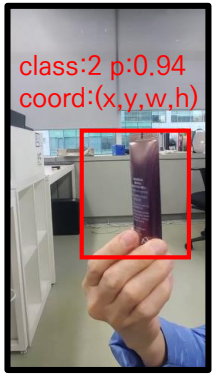


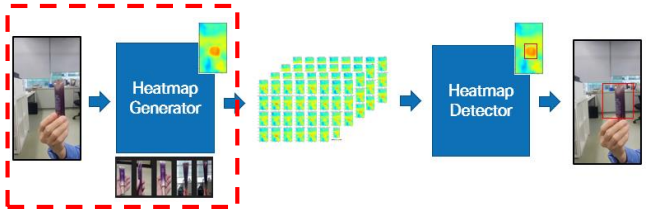
모든 클래스에 대해
6채널의 CAM 생성

CAM
객체 검출기

클래스 확률과 박스의
좌표를 한번에 예측

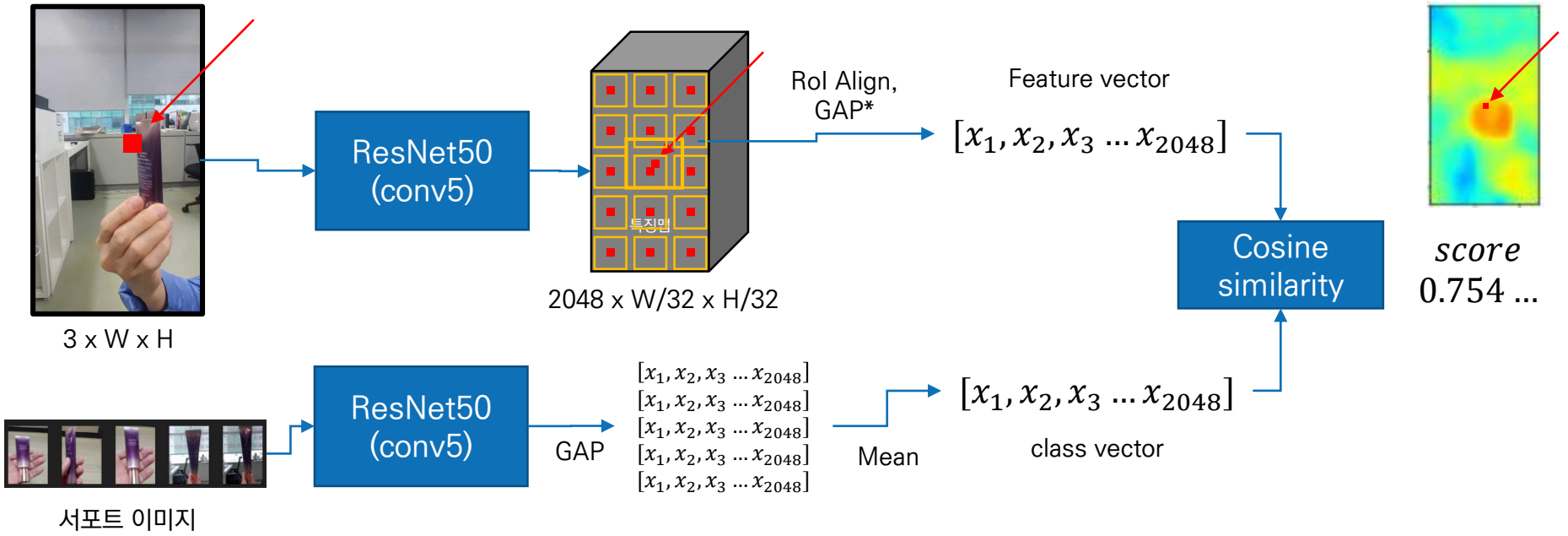




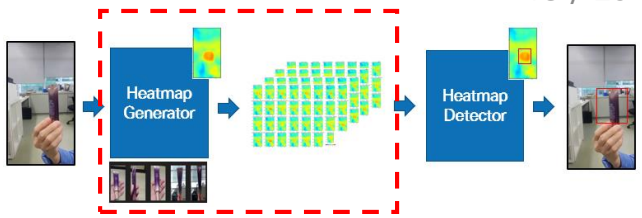


Activation Score

- CAM을 만들기 위한 **스코어 정의**
  - 클래스 서포트 이미지들의 특징 벡터의 **평균값**과 입력 이미지 특징 벡터의 **코사인 유사도**
  - CNN백본으로 이미지넷 분류 데이터에 훈련된 **ResNet50** 사용
    - FC layer 이전에 생성되는 conv5모듈의 2048차원 벡터를 이미지 특징 벡터로 사용

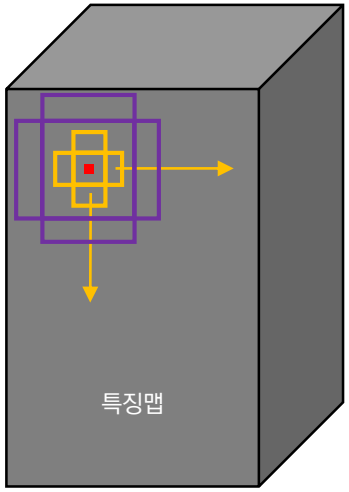


*GAP – Global Average Pooling



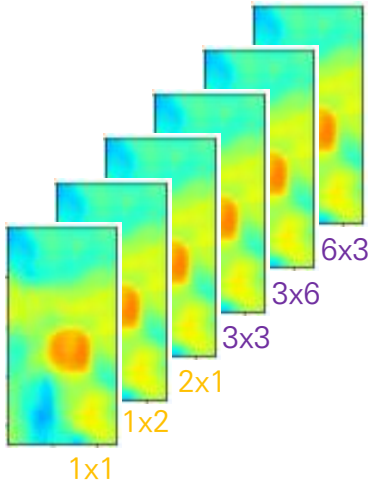
CAM 생성기

- 모든 앵커별 클래스별 CAM 생성
  - 6가지 {1x1, 1x2, 2x1, 3x3, 3x6, 6x3} 크기의 앵커를 0.25씩 stride하여 이미지당 6채널의 CAM 생성 (224x224로 업샘플링)
  - RoI align을 이용해 한번의 CNN 계산만으로 이미지의 2만개 이상의 RoI의 activation score를 병렬적으로 계산

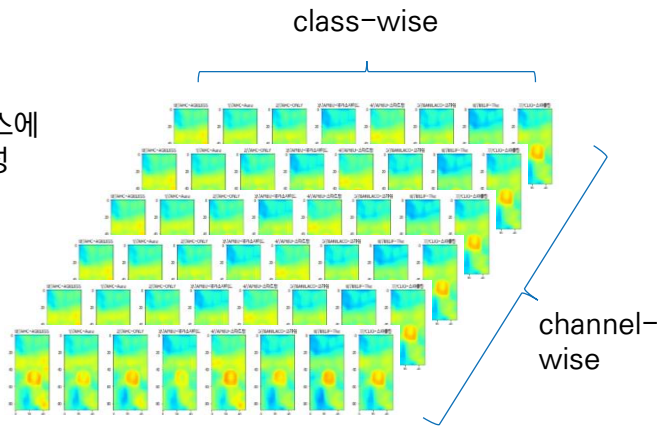


2048 x W/32 x H/32

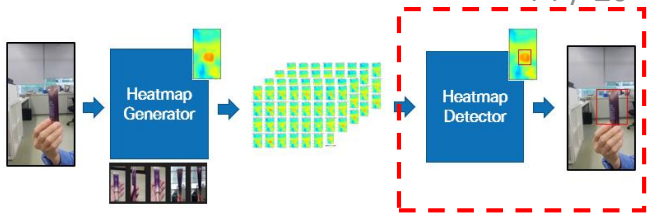
Texture Score



모든 클래스에 대해 생성

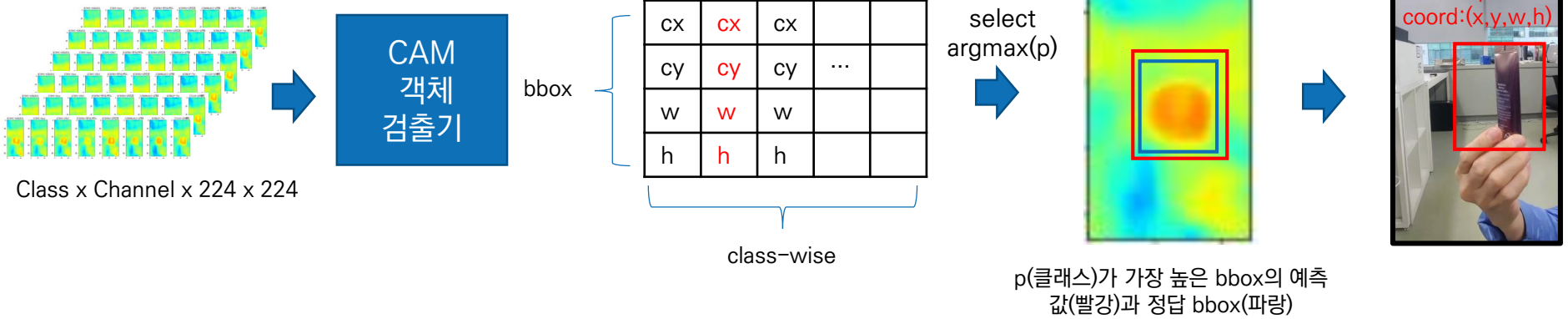


Class x Channel x 224 x 224



CAM 객체 검출기

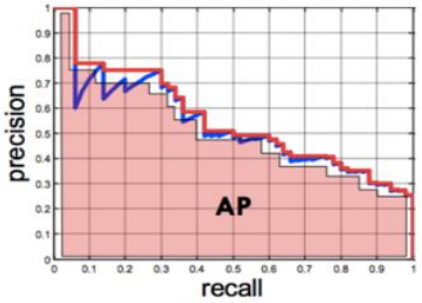
- CAM들을 입력으로 받아서 클래스 확률과 경계박스(bbox) 좌표를 출력함
- [Original, Dilated CNN 각x4] → [Multi Head Attention] → [FC layer x3] → (p, x,y,w,h)
  - 자세한 모델 구조는 부록에
  - **보편적인 데이터셋에 훈련 (한 번 훈련된 모델은 추가 학습없이 다른 클래스와 도메인에도 적용)**
  - CNN과 Self-Attention의 구조적 유연성을 이용해 모델이 클래스의 개수에 자유로움



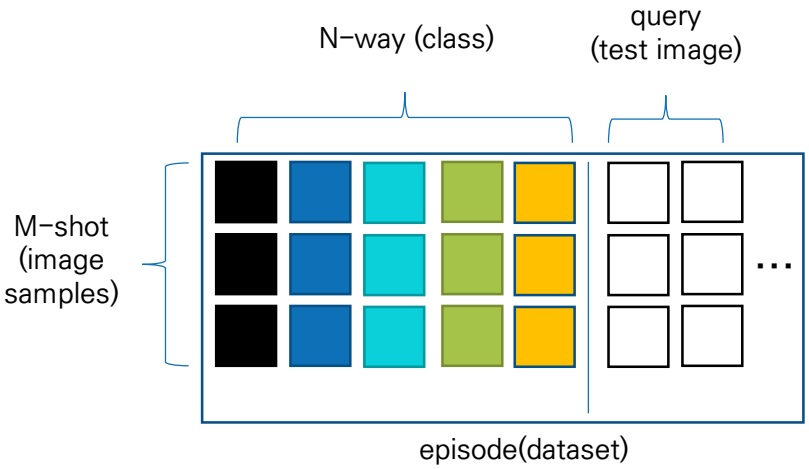
p(클래스)가 가장 높은 bbox의 예측 값(빨강)과 정답 bbox(파랑)

성능 평가 지표

- mAP (mean Average Precision)
  - Classification과 Localization 성능을 모두 반영
  - AP: Precision-Recall curve의 아래 면적
    - mAP: 모든 클래스의 AP의 평균



- {N}-way {M}-shot
  - 퓨-샷 평가 데이터 구조
  - {N}-way: 테스트 시 **N개의 클래스** 사용, N ∈ {10, 21}
  - {M}-shot: 테스트 시 **각 클래스마다 M개의 서포트 이미지**를 사용, M ∈ {1, 3, 5, 10}



- 데이터셋
  - 자체 수집 화장품 31종 데이터셋
    - 21종류 클래스 검출기 훈련용(1903장) / 평가용(969장)
    - 10종류 클래스 검출기 평가용(989장)
  - PASCAL VOC 2012 데이터셋
    - 검출기 훈련용 11종류 클래스(2427장)
  - 세부 정보는 부록에

실험 설계

- 정량적 성능 평가
  - 데이터셋 도메인 차이에 따른 성능 비교
  - CAM 채널 수에 따른 성능 비교
- 크로스 도메인에서의 개인화된 클래스에 대한 Case study
  - PASCAL VOC에 훈련된 검출기로 추가적으로 "과일", "편의점", "꽃" 등 여러 다른 도메인의 검출 가능성 정성적 평가

데이터셋 도메인 차이에 따른 성능 비교

실험 구성 (모든 실험은 각 조건에서 30회 씩 진행)

실험	훈련 데이터	평가 데이터				비고
		종류	way	shot	도메인	
A	PASCAL VOC(trainval) 11 클래스	화장품(평가용) 10 클래스	10	{1, 3, 5, 10}	다름	클래스 개수에 따른 성능 차이 관측
B		화장품(평가용) 21 클래스	21			
C	화장품(훈련용) 21 클래스	화장품(평가용) 10 클래스	10		같음	높은 검출 성능

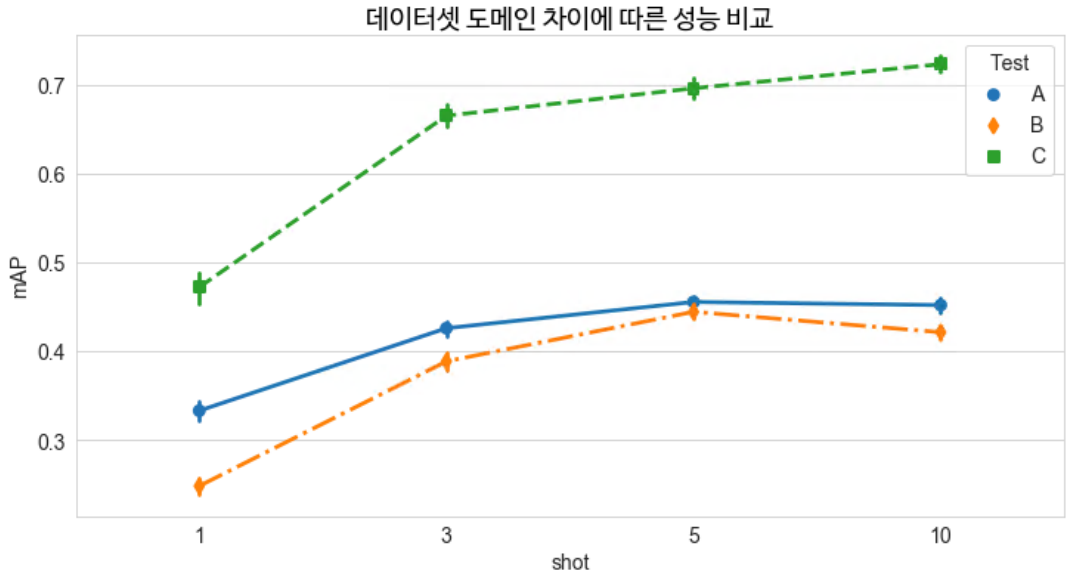
실험 결과 (mAP)

실험	Shots			
	1	3	5	10
A	0.333 ±0.012	0.426 ±0.010	**0.456** ±0.006	0.452 ±0.009
B	0.249 ±0.009	0.389 ±0.010	**0.445** ±0.009	0.422 ±0.008
C	0.472 ±0.020	0.665 ±0.014	0.696 ±0.013	**0.724** ±0.010

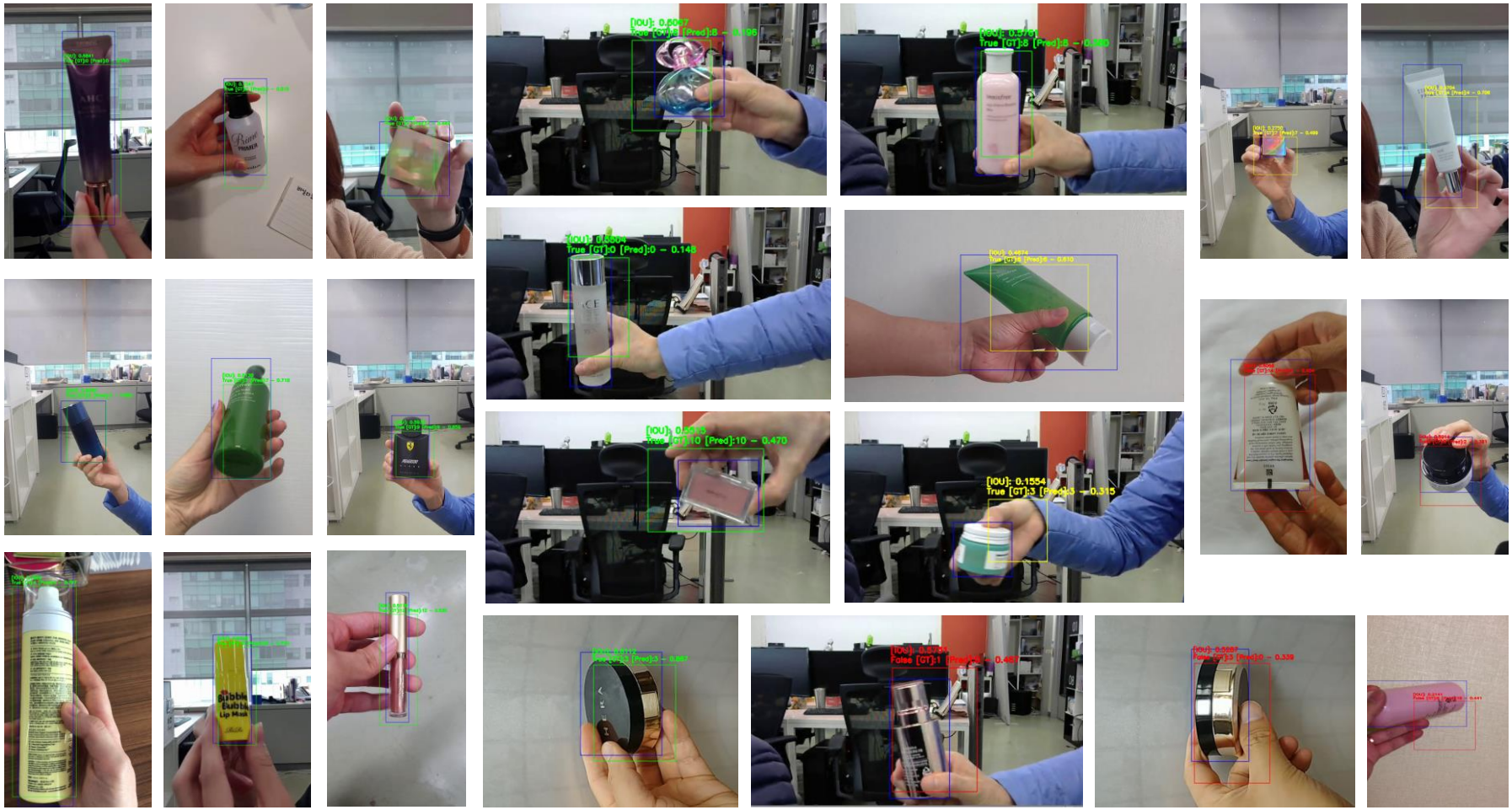
※ 기존 학습모델 검출기들은 PASCAL VOC(20클래스)의 경우 약 0.8mAP, MSCOCO(80클래스)의 경우 약 0.6mAP 정도의 성능이 나온다

데이터셋 도메인 차이에 따른 성능 비교

- 검출기의 성능이 클래스의 개수에 따라 크게 차이가 나지 않음
- 검출기 훈련 시에 데이터셋의 도메인에는 영향을 받음
  - 같은 도메인의 화장품 데이터에 훈련한 검출기(실험C)가 PASCAL VOC 데이터에 훈련한 검출기 (실험A, B) 보다 성능이 더 좋다



데이터셋 도메인 차이에 따른 성능 비교



실험 A, B 검출 결과 예시 – 파랑(Ground truth), 초록(정답 맞춤), 노랑(IOU 0.5이하 오답), 빨강(분류 오답)

CAM 채널 수에 따른 성능 비교

실험 구성 (모든 실험은 각 조건에서 30회 씩 진행)

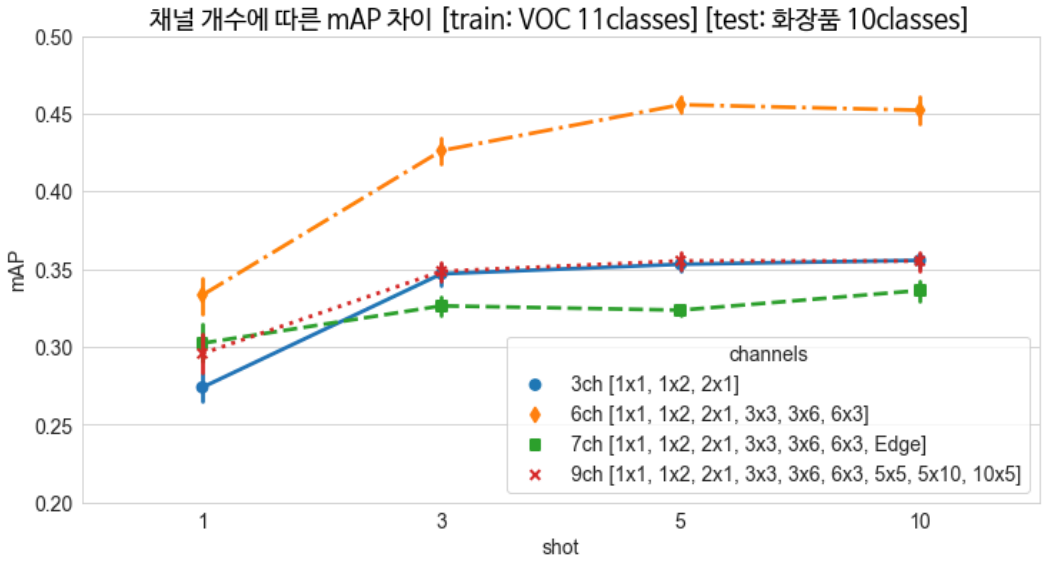
채널	앵커	추가 정보	비고
3ch	[1x1, 1x2, 2x1]	–	앞서 실험 A에서 사용한 검출기를 이용해 평가.
6ch	**[1x1, 1x2, 2x1, 3x3, 3x6, 6x3]**	–	
7ch	[1x1, 1x2, 2x1, 3x3, 3x6, 6x3, Edge]	Edge정보 추가	
9ch	[1x1, 1x2, 2x1, 3x3, 3x6, 6x3, 5x5, 5x10, 10x5]	–	

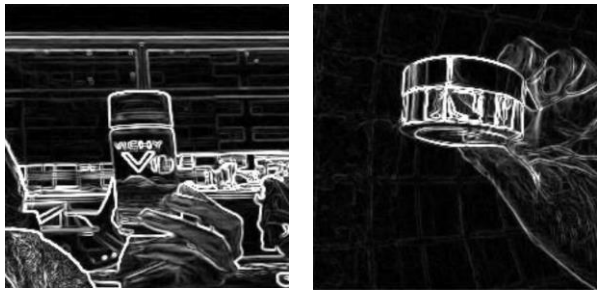
실험 결과 (mAP)

Channels	Shots			
	1	3	5	10
3ch	0.274 ±0.01	0.347 ±0.007	0.353 ±0.005	0.356 ±0.004
6ch	**0.333** ±0.012	**0.426** ±0.01	**0.456** ±0.006	**0.452** ±0.009
7ch	0.302 ±0.012	0.326 ±0.007	0.324 ±0.004	0.336 ±0.007
9ch	0.296 ±0.013	0.349 ±0.007	0.355 ±0.006	0.355 ±0.006

CAM 채널 수에 따른 성능 비교

- 채널을 줄이거나 더 늘린 경우보다 6채널의 CAM이 가장 성능이 좋음
- 이미지의 Edge채널을 추가로 넣어준 경우도 오히려 모델 성능 저하
  - 좀 더 나은 bbox를 만들길 기대했으나 분류 정확도, 검출 mAP 저하





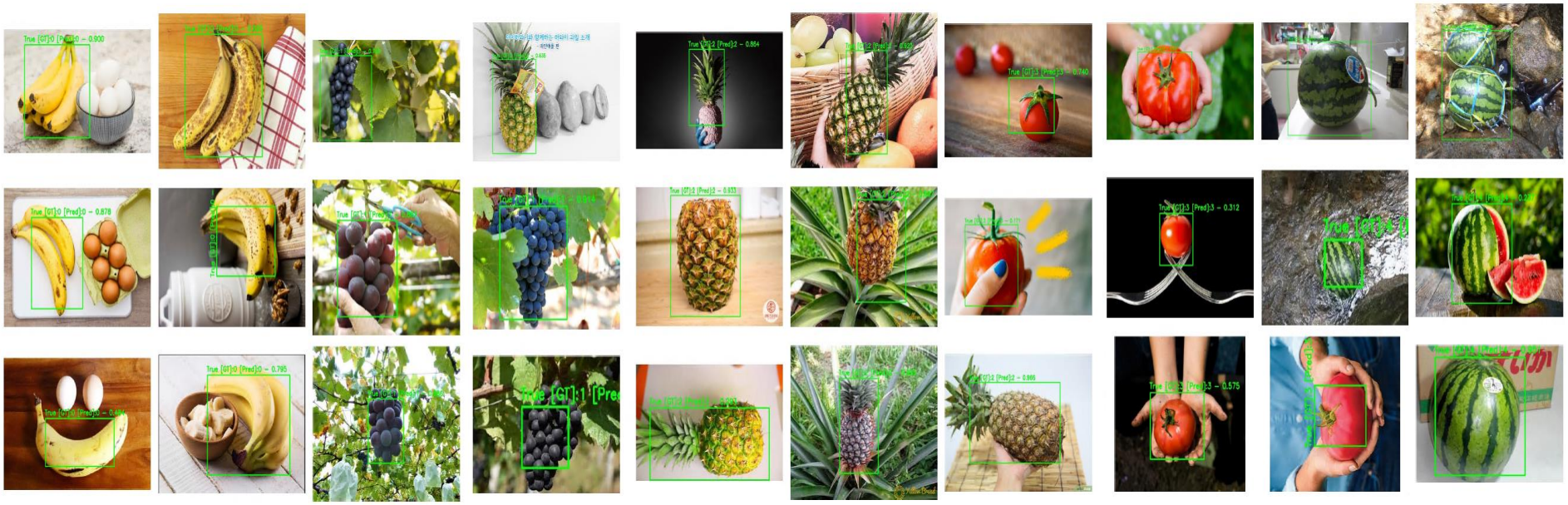
Sobel operator로 만든 Edge 예시

Case study(크로스 도메인 적용)

- "과일" 도메인 5-way 5-shot
  - 이미지는 구글 이미지 검색에서 서포트 이미지와 검출용 이미지가 겹치지 않도록 선정
  - 앞서 실험 A에서 사용한 검출기를 그대로 사용
- "편의점", "꽃" 도메인 결과는 부록에

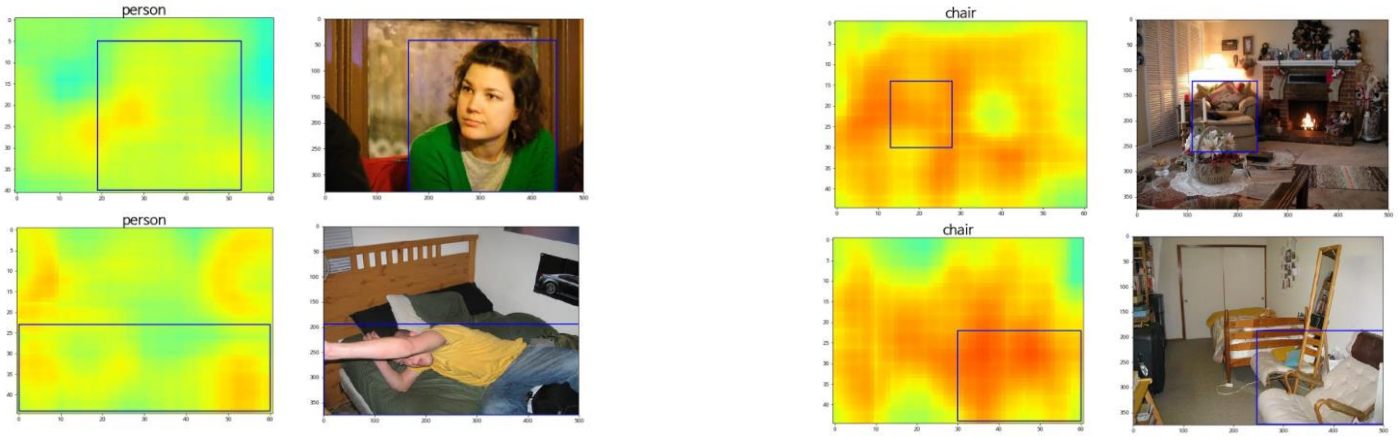
서포트 이미지

검출 결과 예시



사용된 이미지에 관한 논의

- "이미지넷 분류 데이터의 텍스처 편향" 성질을 역이용해서 클래스 독립적인 CAM을 만들었다
  - 기존 논문에서는 이 성질을 해결해야 할 문제라고 인식
- "사람"이나 "의자" 처럼 클래스를 대표하는 texture가 없거나 shape가 더 중요할 경우 CAM이 부정확해짐
  - 서포트 이미지의 texture가 뚜렷할수록 이미지의 국소영역이 갖는 정보가 많다
  - Texture가 뚜렷한 것들이 CAM이 잘 만들어짐. 예) 고양이, 강아지, 말, 비행기, 자동차 등…



CAM이 잘 만들어지지 않은 예시, 좌(사람) 우(의자)

사용된 이미지에 관한 논의

- 서포트 이미지의 False positive 문제
  - 이미지가 몇 장 없으니 같은 배경에서 촬영한 사진은 물체의 배경과 물체를 구분하기 힘들다
    - 물체가 있는 영역만 세밀하게 수동으로 잘라내어 사용
    - 객체 검출기가 고정된 카메라에서 사용된다면 배경제거 알고리즘과 같은 후처리를 통해 해결

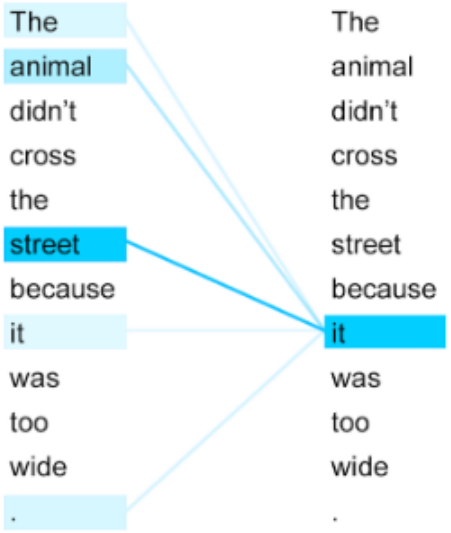
두 클래스가 배경과 손이 공통으로 포함되어
클래스간 서로 구분이 힘듦

수동 이미지 잘라내기

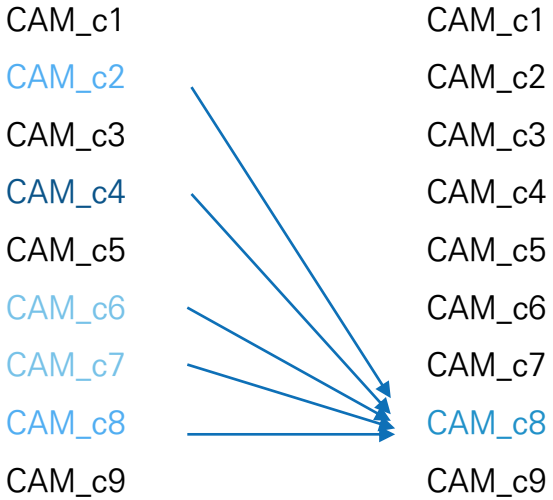
GrabCut 알고리즘 배경제거

Self-Attention의 역할

- Self-Attention을 활용해 가변적으로 클래스 개수를 수용할 수 있는 모델 개발
  - Self-Attention을 사용한 NLP 모델들이 문자열의 길이에 제한이 없듯, 검출기의 클래스의 개수에 제한이 없도록 했다.



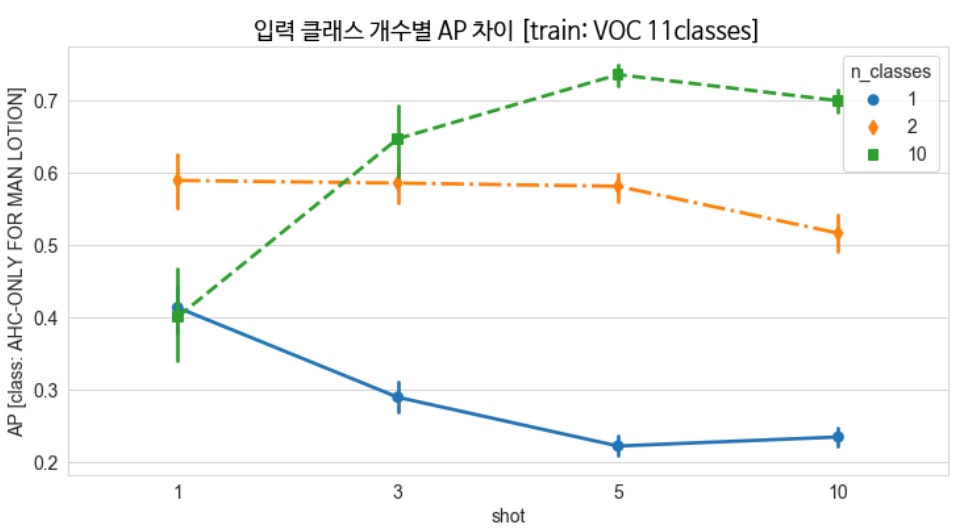
NLP에서의 Self Attention



CAM 검출기에서의 Self Attention

Self-Attention의 역할

- 다른 클래스와의 비교 없이 한 개 클래스만 사용하면 (1way) 오히려 localization이 부정확해짐
  - CAM 스코어의 절대적인 값을 이용하는 것이 아니라 다른 클래스의 CAM 스코어와 상대적으로 비교
  - **Self-Attention을 이용한 클래스간 CAM 상호 참조가 상당히 중요**



AHC-ONLY FOR MAN LOTION

연구의 한계

- 제너릭한 검출기 모델 학습의 어려움
  - PASCAL VOC에 훈련한 CAM 검출기가 같은 도메인에 훈련한 검출기와 성능 차이를 보임
  - 이미지넷 분류 데이터처럼 제너릭한 검출기를 훈련시킬 수 있는 대표적인 대용량 데이터가 필요
- 정밀한 경계박스(bbox) 생성 힘듦
  - bbox regression 등의 후처리가 없어 mAP 평가에서 손해를 많이 봄
- 다중 객체 검출 지원
  - 1-stage 검출기인 YOLO 시스템을 응용해 보았으나 모델 학습 시 localization loss가 수렴하지 못함
  - 2-stage 검출기인 R-CNN 계열처럼 앞단에 region proposal 단계를 두어 문제 해결 가능

연구의 기여

- "이미지넷 데이터 학습 모델이 텍스처에 편향된 성질"을 이용해 CAM을 만들었고 이것을 이용해 Classification과 Localization문제를 해결할 수 있음을 보였다
- CAM과 Self-Attention을 이용해 크로스 도메인에서 클래스의 개수에 관계없이 보편적으로 사용 가능한 검출기 모델의 구조를 제시했다
- CAM 객체 검출기는 새로운 클래스의 추가/변경 시 새로운 데이터에 대해 추가적인 학습 없이도 합리적인 성능을 보여준다
- 정량적인 실험을 통해 CAM 검출기가 모델의 구조나 CAM 채널 개수에 따라 어떠한 변화가 있는지 보였다

결론

- **개인화된 객체 검출기를** 만들려면 **퓨-샷, 동적 클래스 변경, 크로스 도메인 적용**이 모두 가능해야 하지만 기존의 모델은 한번에 가능한 것이 없었다
- 모델을 만들기 위해 "이미지넷 분류 데이터의 텍스처 편향" 성질을 이용한 **CAM과 Self-Attention구조**를 사용하였다
- 모델은 **CAM 생성기와 CAM 객체 검출기로** 이루어져 있고, **퓨-샷, 동적 클래스 변경, 크로스 도메인 적용**이 동시에 가능하다
- 모델은 21클래스에서 0.45mAP 정도의 성능을 보여준다
- 다중 객체 검출과 bbox-regression 등 추가 연구가 필요하다

지금까지 발표를 들어 주셔서 감사합니다

부록

HUMAN
CENTERED
COMPUTING
LABORATORY

Case study(크로스 도메인 적용)

- "꽃" 도메인 5-way 5-shot Detection

서포트 이미지

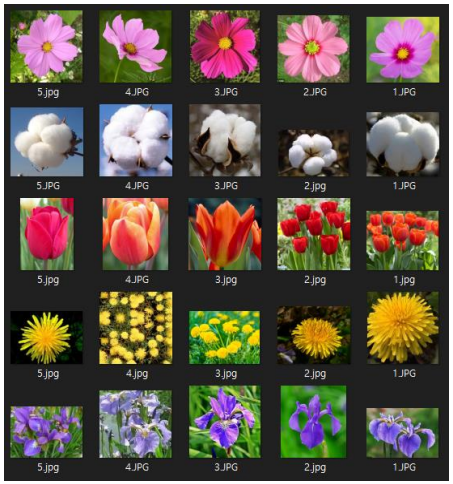
코스모스

목화

튤립

민들레

붓꽃



검출 결과 예시

Case study(크로스 도메인 적용)

- "편의점" 도메인 5-way 5-shot Detection

서포트 이미지

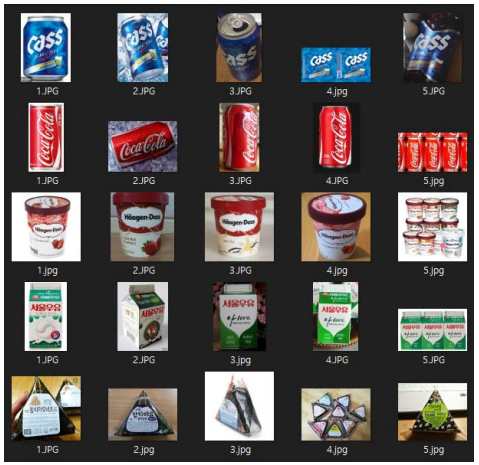
카스맥주

코카콜라

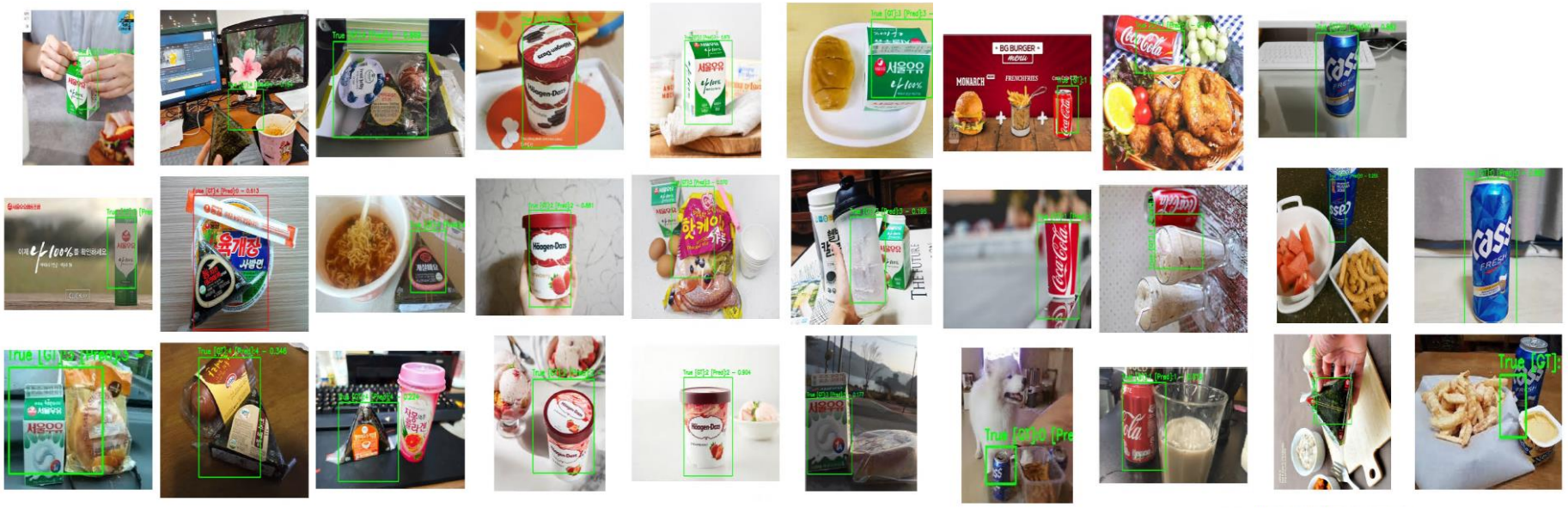
하겐다즈

서울우유

삼각김밥

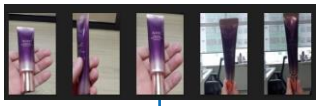


검출 결과 예시



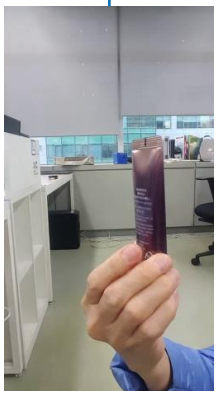
CAM 생성기 구조

Support image
- few (1, 3, 5, 10, 30) shot

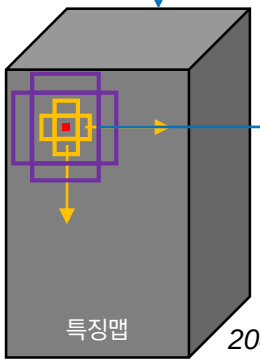


Resnet50 (conv5)

Resnet50 (conv5)



1 x 3 x H x W



2048 x H/32 x W/32

- *n x m anchor*
- 0.25 stride

RoI Align

global average pooling

1,2048

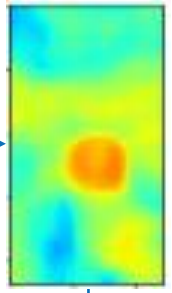
class-wise mean

$[x_1, x_2, x_3 \dots x_{2048}]$

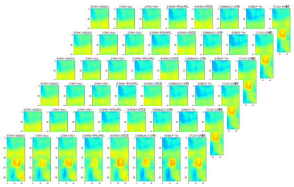
$[x_1, x_2, x_3 \dots x_{2048}]$

cos similarity

1 x H/32 x W/32



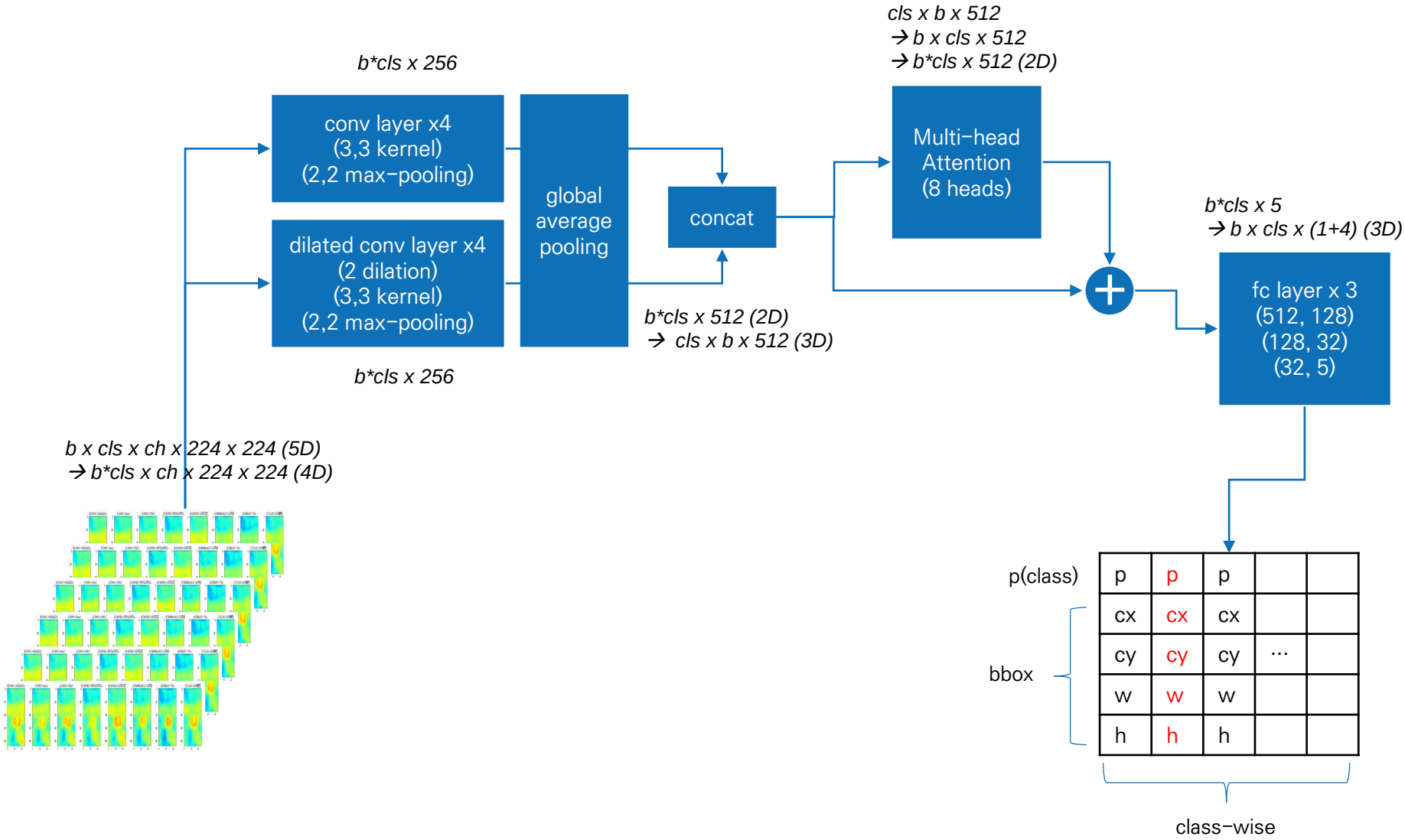
Cls x Ch x 224 x 224



- { 1x1, 1x2, 2x1, 3x3, 3x6, 6x3} anchors, 0.25 stride
- 모든 클래스에 대해 생성
- 224 reshape

*B: n_batch, Cls: n_class, Ch: n_channel, H: height, W: width

CAM 객체 검출기 구조



*B: n_batch, Cls: n_class, Ch: n_channel, H: height, W: width

CAM 객체 검출기 상세 훈련 방법

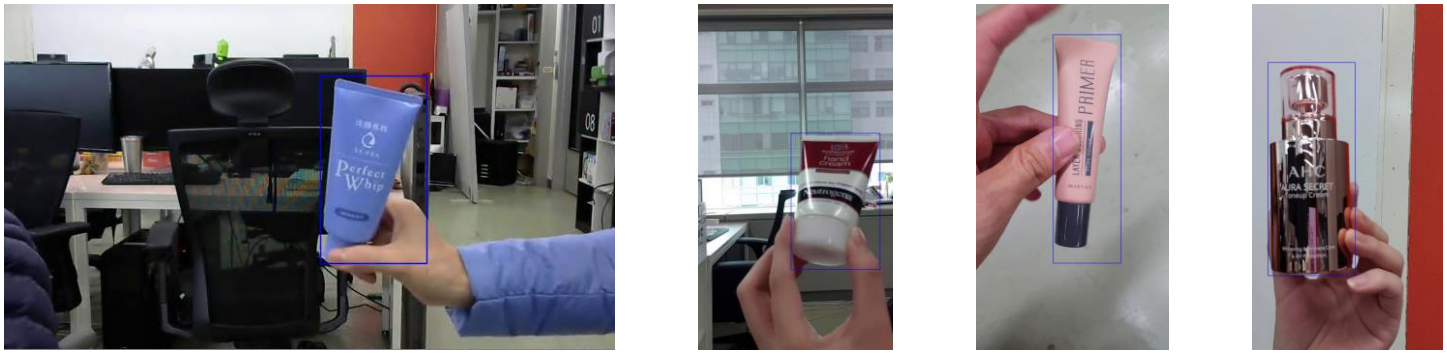
- 보편적인 이미지의 CAM들에 대해 훈련 필요
- $Loss(p, cx, cy, h, w, cx', cy', h', w') = BCE(p) + \mathbb{1}_c(\lambda_{xy} * Loss_{xy} + \lambda_{wh} * Loss_{wh})$

  $\mathbb{1}_c = indicator\ function\ (IFF\ correct\ class)$
  $Loss_{xy} = (cx - cx')^2 + (cy - cy')^2$
  $Loss_{wh} = (w - w')^2 + (h - h')^2$
  $\lambda_{xy} = 500, \quad \lambda_{wh} = 100$
  $coord_{gt} = (cx, cy, h, w)\ 0{\sim}1\ normalized\ YOLO\ style$
  $coord_{pred} = (cx', cy', h', w')$

  - Class agnostic 하도록 훈련하므로 softmax 대신 sigmoid를 사용하는 BCE loss 사용
  - loss_xy, agnostic loss_wh는 정답 클래스에 해당하는 것만 loss에 포함
  - 이미지.s 위치의 cx,cy 중심을 맞추는 것이 w,h를 맞추는 것보다 더 중요
  - Validation 데이터셋이 2 epoch 동안 loss가 떨어지지 않으면 lr x0.2 decay, Adam opt 초기 lr=0.001
    - lr decay가 5회 발생하면 학습 종료

데이터 셋

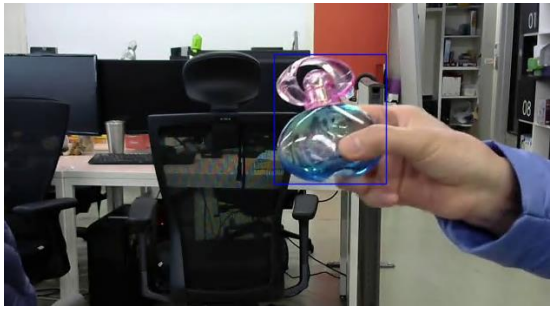
- PASCAL VOC 2012 trainval
  - 검출기 훈련용 11 클래스
  - 데이터셋 중 이미지당 1개의 bbox만 있고 텍스처가 뚜렷한 11개 클래스 중 bbox의 넓이가 전체 넓이의 0.01(0.1^2) 이상 0.64(0.8^2)인 것 2427장 사용
    - ['aeroplane', 'bicycle', 'bird', 'car', 'cat', 'cow', 'dog', 'horse', 'motorbike', 'sheep', 'train']
- 자체 수집 화장품 데이터셋
  - 사무실, 창가, 벽, 바닥 등의 배경에서 다른 기종의 카메라로 각 클래스당 6가지 비디오를 20초 가량 촬영 후 이미지를 10~15 프레임 간격으로 추출해 사용
  - 로션, 수분크림, 세안제, 파우더, 아이크림, 향수 등 다양한 카테고리 항목 포함
    - 21종류 클래스 검출기 훈련용(1903장) / 평가용(969장)
    - 10종류 클래스 검출기 평가용(989장)



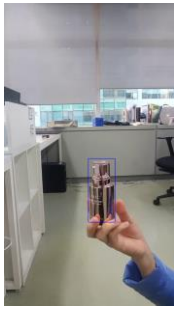
비디오 배경 예시) 차례로 사무실, 창가, 바닥, 벽

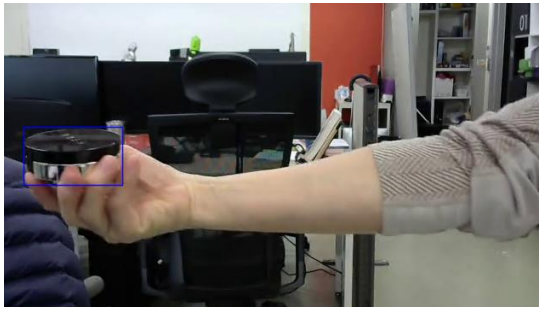


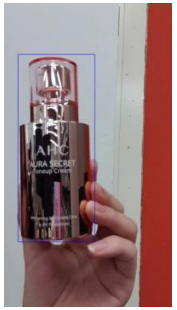


innisfree
no-sebum
Mineral Powder





PRIMER

AHC