CVPR 2016, Class Activation Maps (CAM)

Learning Deep Features for Discriminative Localization

2023.11.22(수) 인공지능 논문 리뷰

산업융합학부 정보융합전공
유지수 (2020000055)

Contents

- 논문 소개
- 논문 핵심 요약
- 문제점 및 해결책
- XAI : eXplainable AI, 설명가능 인공지능
- GAP (Global Average Pooling)
- CAM (Class Activation Mapping)
- 실험 및 결과 : Weakly-supervised Object Localization, Pattern Discovery
- 결론 및 향후연구

논문 소개

Learning Deep Features for Discriminative Localization

- CVPR(Computer Vision and Pattern Recognition) 2016
- Computer Science and Artificial Intelligence Laboratory, MIT
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba

Class Activation Maps (CAM)을 다룬 논문

Convolution Neural Network (CNN) : 이미지의 지역적 특징을 잘 포착

"CNN을 이용한 이미지 분류 해석(시각화) 가능 방법 제시"

논문 핵심 요약

논문의 가장 큰 특징 3가지

1. **Global Average Pooling (GAP)**를 적용한 해석(시각화) 가능 구조 제시
2. Feature Map 객체 위치 추출 방법 **Class Activation Mapping (CAM)** 제시
3. 다양한 실험을 통한 CNN 구조 및 객체 추출 방법의 **객체 인식 성능** 증명

문제점 및 해결책

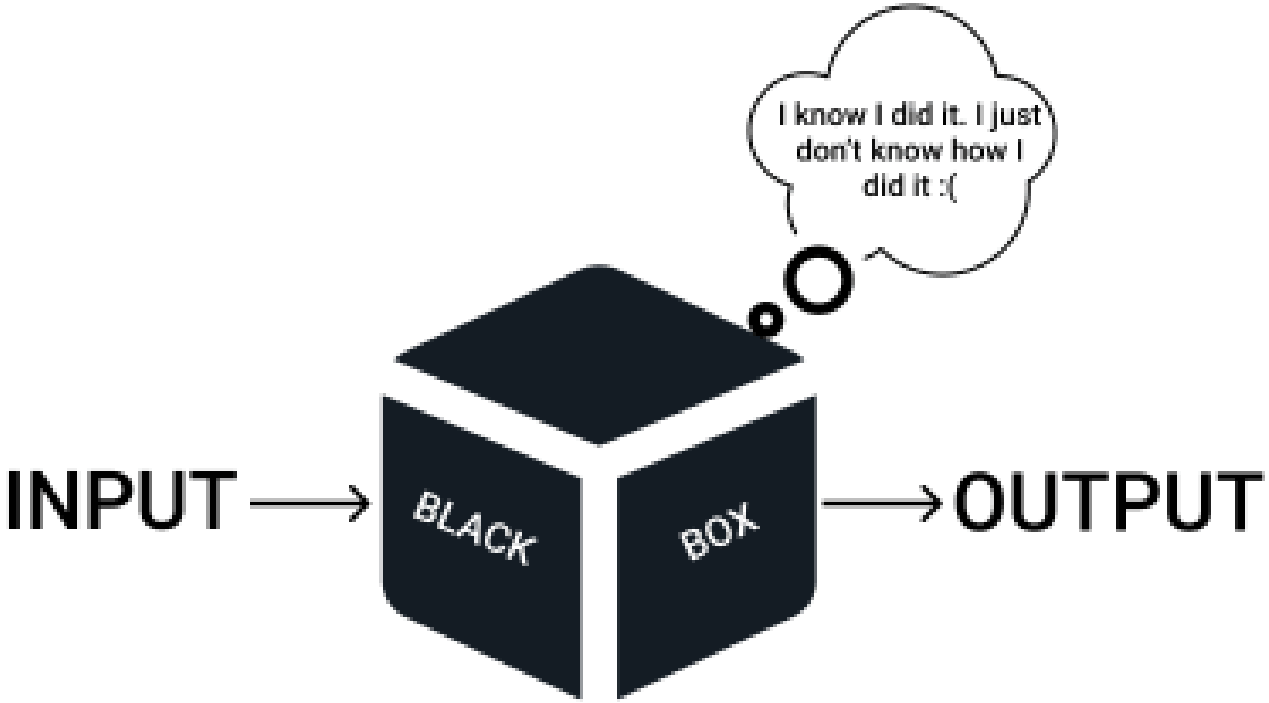
문제점

- Deep Learning = Black Box
- 보통 CNN의 구조 : Input → Convolution Layers → **Fully connected Layers**
  - ☞ 마지막 Layer를 FC-Layer로 Flatten하는 과정에서 Convolution이 가지고 있던 각 픽셀의 위치 정보를 잃어 Classifying 정확도가 아무리 뛰어날 지라도 특정 이미지의 어떤 Feature를 보고 특정 Class를 판별했는지 알 수 없음

해결책

- FC-Layer의 구조를 살짝 바꿔 위치정보를 손실하지 않도록 'CAM' 방법 활용
- CAM 방법 구조 : Input → Convolution Layers → **Global Average Pooling**
  - ☞ 마지막 Convolution을 FC-Layer 대신 GAP을 적용하여 별다른 추가의 지도학습 없이 CNN이 특정 위치들을 구별할 수 있도록 함
  - ☞ CAM을 통해 특정 Class 이미지의 Heat Map을 생성하여 CNN이 어떻게 이 이미지를 특정 Class로 분류했는지를 이해할 수 있게 되어 Explainable한 결과를 낼 수 있음

XAI : eXplainable AI, 설명가능 인공지능



뇌의 신경세포를 모방한 Neural Network

서로 복잡하게 연결된 수백만개 이상의 parameter가 비선형으로 상호작용하는 구조

Black Box Model

복잡한 구조 덕에 성능은 기존 기계학습보다도 월등이 높아졌으나 사람의 인지를 넘어선 내부 구조 탓에 AI가 왜 그런 결과를 도출했는지는 개발자도 알 수 없음

XAI : eXplainable AI, 설명가능 인공지능

XAI ?

사람이 AI의 동작과 최종결과를 이해하고 올바르게 해석할 수 있고,
결과물이 생성되는 과장을 설명 가능하도록 해주는 기술

설명의 필요성

암을 진단하는 AI 도입한 A병원

- 배탈이 난 것 같아 병원을 방문한 B씨는 AI로 부터 암 판정을 받음
- 어떤 증상 때문에 암이라고 진단했는지 설명 불가(빅데이터 기반 예측)
- B씨는 정밀 검사를 받아야 할지 고민에 빠짐

인공지능이 중요 작업(mission critical)에 사용되기 위해선
인공지능의 설명성, 투명성 확보 기술, 기준 정립이 필요

XAI : eXplainable AI, 설명가능 인공지능

연구 동향

XAI 기술 분류 기준 3가지 : 상하관계 X, 3가지 관점 중 하나에 귀속시키는 것 X

관점	분류	
Complexity 모델의 복잡성	Intrinsic 자체 해석력 확보	Post-hoc 예측 결과 사후 해석
Scope 설명의 범위	Global(전역적인 기법) 모든 예측 결과를 설명	Local(국소적인 기법) 일부 예측 결과만 설명
Dependency 기법의 범용성	Model-specific 특정 종류 모델만 적용	Model-agnostic 모델 상관없이 적용

XAI : eXplainable AI, 설명가능 인공지능

CAM

관점	분류
Complexity 모델의 복잡성	Intrinsic vs **Post-hoc** 모델이 학습이 되고 난 후 적용해서 설명 제공
Scope 설명의 범위	Global vs **Local** 개별 이미지 마다 그 예측 결과를 설명하는 방법
Dependency 기법의 범용성	**Model-specific** vs Model-agnostic CNN계열에서만 쓸 수 있는 시각화 해석 기법

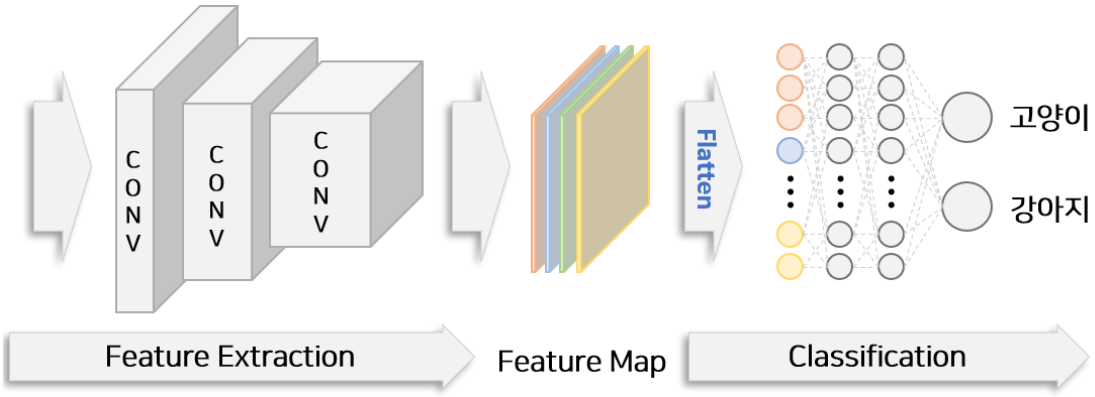
CAM : Class Activation Map

이미지의 어느 부분을 보고 class를 예측했는지 시각화

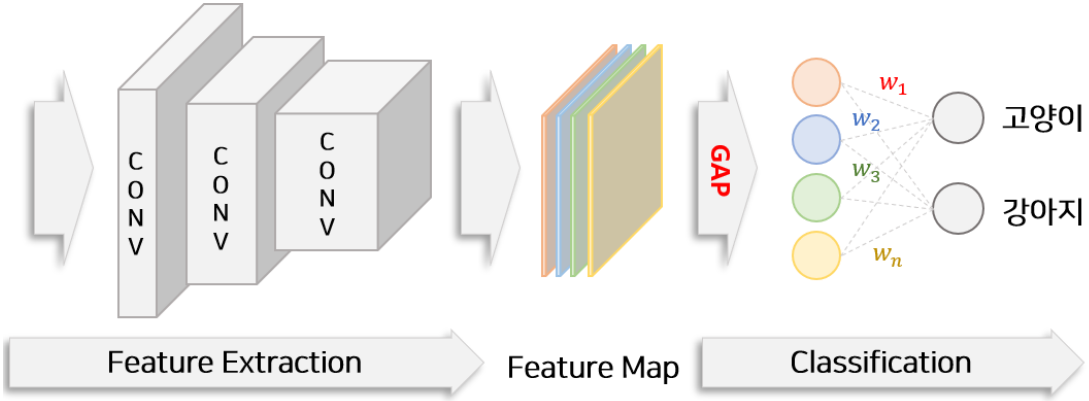
Feature map을 flatten하지 않고, "**Global Average Pooling**"을 사용

GAP (Global Average Pooling)

일반적인 이미지 분류 모델 구조



GAP가 적용된 이미지 분류 모델 구조 : 객체 위치 추출 가능



GAP (Global Average Pooling) 구조

GAP 확률 계산

- 각 Feature Map의 가로 세로 값을 모두 더해 1개의 특징변수로 변환

$$\sum_{x,y} f_k(x,y) = F_k$$
$$S_c = \sum_k w_k^c F_k$$
$$P_c = \frac{exp(S_c)}{\sum_c exp(S_c)}$$

$f_k(x,y)$: Feature Map k의 가로(x), 세로(y)에 해당하는 값

F_k : 특징변수 k

k : Feature Map의 index

x, y : Feature Map의 가로, 세로 좌표

w_k^c 특징변수 k가 클래스 c에 기여하는 weight

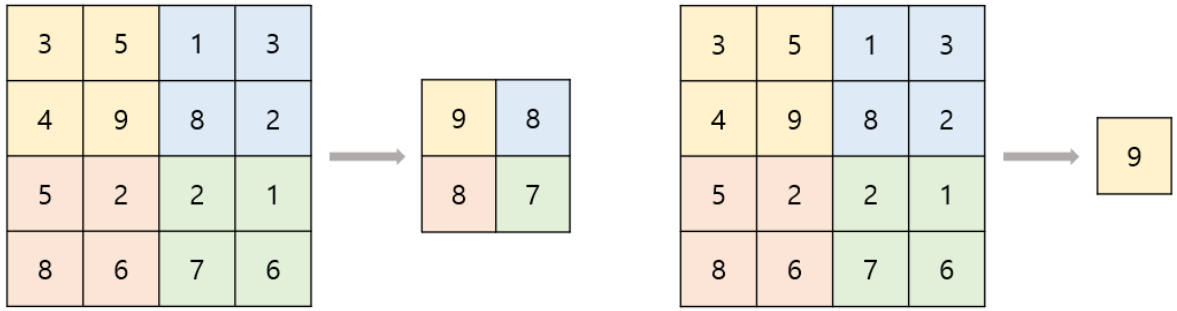
- 특징변수(F)와 FC Layer의 Weight를 곱하여 더하면 각 Class의 점수를 계산(S)할 수 있음
- 각 특징변수에 곱해진 Weight는 각 Feature Map이 해당 Class에 얼마나 기여하는 지를 나타나며 Class 점수에 SoftMax 함수를 취하여 각 Class로 분류될 확률(P)을 계산함

GAP vs GMP

Pooling Layer

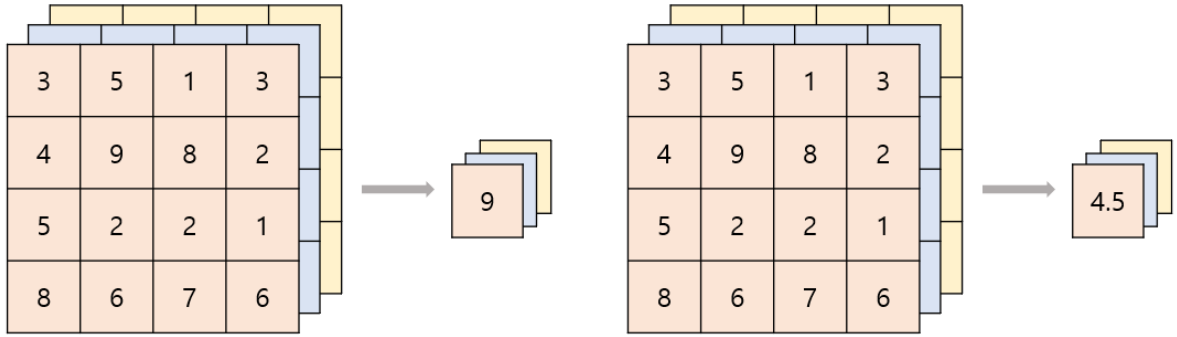
Conolution Layers에 존재하는 수많은 filter(parameter)의 Overfitting 을 방지하기 위한 장치

Global Average Pooling



GAP를 적용 했을 때 객체 추출(Localization) 능력이 월등함을 실험적으로 증명함

Global Max Pooling



CAM (Class Activation Mapping)

Class Activation Mapping 시각화 예시

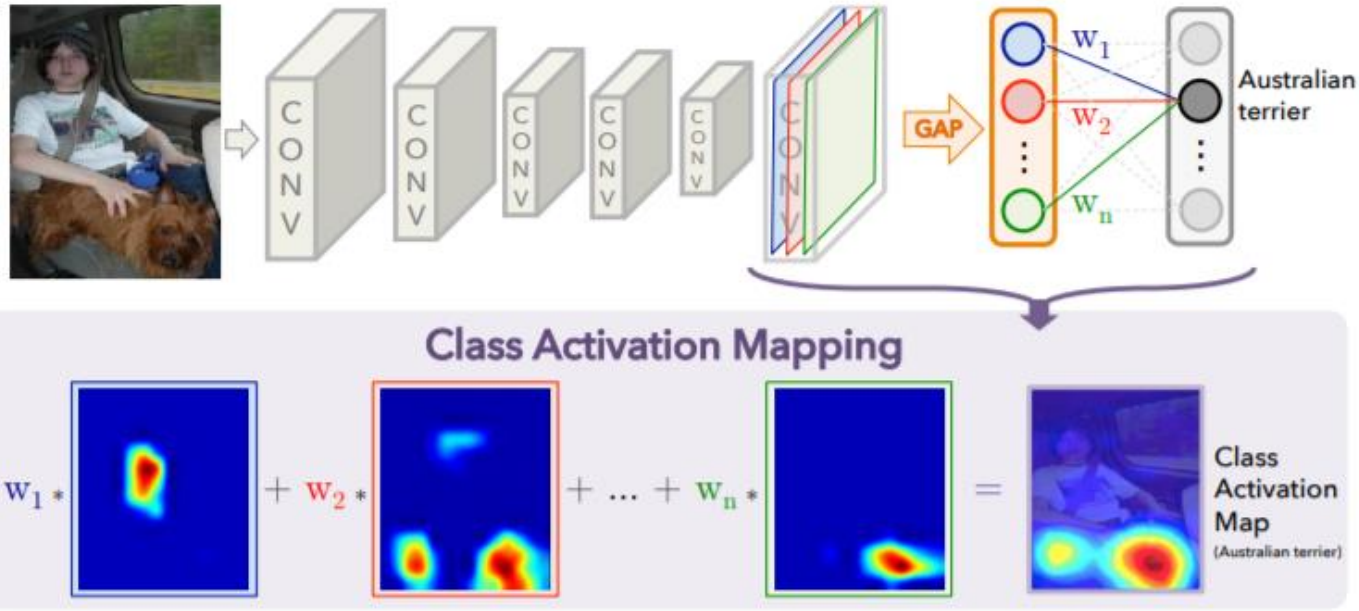
Global Average Pooling을 사용한 CAM을 시각화한 자료

각 이미지들에 대해 Classify하면서 object들이 위치하는 영역도 찾아낼 수 있음을 확인

CAM (Class Activation Mapping) 구조

Class Activation Mapping

CNN이 Input image에 대한 Prediction을 만들어냈을 때, 해당 Class로 판별하는데 중요하게 생각하는 영역을 표시하여 시각화 하는 알고리즘



Australian terrier 분류

- 사람 얼굴에 집중한 첫 번째 activation map의 W_1은 낮은 값일 것으로 유추 가능
- 개 특징에 주목한 두 번째, n 번째 activation map에 연결된 W_2, W_n은 높은 값을 가질 것

CAM (Class Activation Mapping) 구조

CAM 좌표 계산

- 각 Class로 분류될 확률에 영향을 미친 객체의 좌표(x, y)를 추출

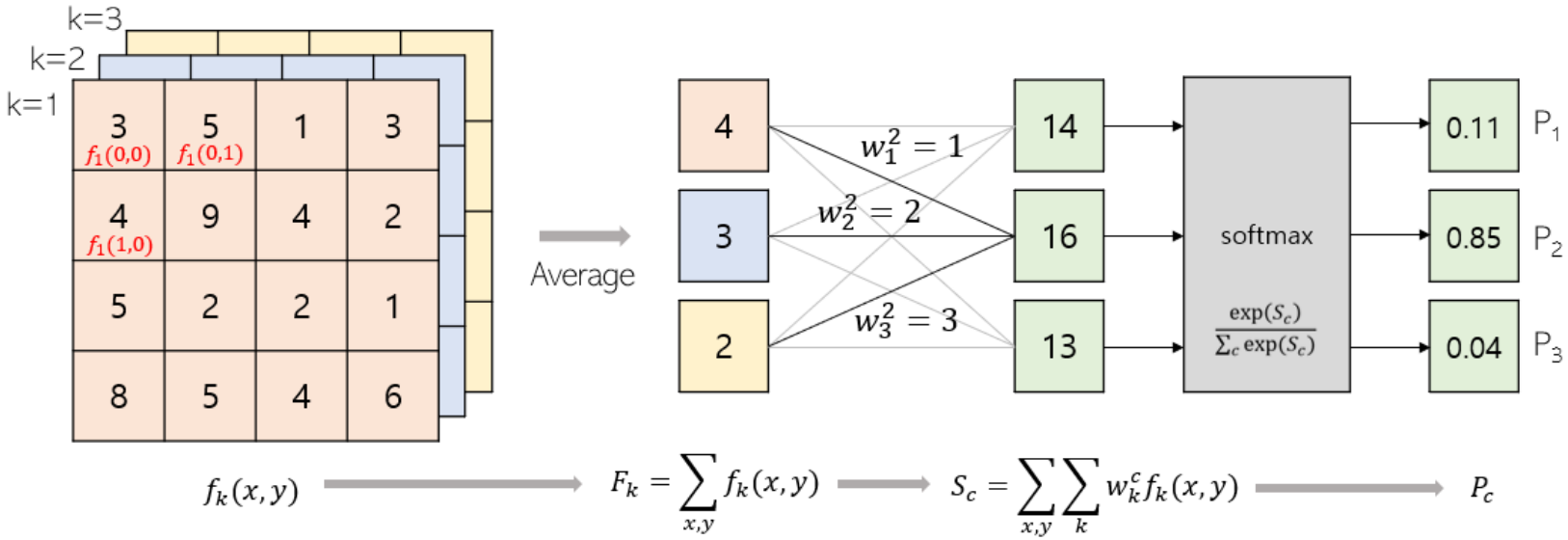
$$S_c = \sum_k w_k^c F_k$$
$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y)$$
$$S_c = \sum_{x,y} \sum_k w_k^c f_k(x,y)$$
$$S_c = \sum_{x,y} M_c(x,y)$$
$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

$M_c(x,y)$: 클래스 c에 대하여 좌표 x, y에 대한 영향력(Activation Value)

- 각 Feature Map(fk(x, y))과 Feature Map이 특정 Class c로 분류될 가중치(w)를 곱하여 합하면 좌표 별(x, y) 특정 Class에 대한 영향력(Class Activation)인 Mc(x, y)를 계산할 수 있음
- 각 클래스에 대해 CAM을 적용, 이미지에서 클래스에 영향을 주는 좌표를 계산함

CAM (Class Activation Mapping) 구조

Class Activation Mapping 계산 과정



1. 마지막 Convolution layer의 feature map을 $f_k(x, y)$라고 하면, 각각의 unit k에 대해 GAP을 수행해서 k개의 값을 출력(GAP의 결과 F_k)
2. 각각의 F_k에 대해서, class c에 대한 가중치의 weighted sum을 계산하여 S_c를 출력(softmax의 input으로 사용)
3. Softmax 연산을 거치면 각 class c에 대한 결과가 출력(bias는 0으로 설정)
4. Class c에 대한 CAM을 M_c 라고 정의하고 S_c의 수식을 변형하여 구할 수 있는 형태로 사용

실험 및 결과
[1] Weakly-supervised Object Localization

실험내용

ILSVRC 2014 Benchmark 데이터에서 모델의 성능을 평가하기 위하여 총 2가지 실험을 진행

1. 논문에서 제시한 구조를 적용할 때 기존 모델의 분류(Classification) 정확도가 하락하는지 여부를 확인
2. 분류문제를 학습한 모델의 CAM을 활용하여 Bounding Box를 만들고 객체를 추출(Localization) 정확도를 확인

- 성능이 검증된 모델 AlexNet, VGGnet, GoogLeNet의 구조를 변경하여 활용
- GAP(Global Average Pooling)를 적용한 모델과 GMP(Global Max Pooling) 적용한 모델도 함께 비교하며 Pooling 방법에 대한 성능을 비교실험으로 확인

실험 및 결과

[1] Weakly-supervised Object Localization

실험결과

1.1) 분류 실험(Classification)

: 분류 모델 정확도가 1~2% 미미하게 하락

1.2) 객체 추출 실험(Localization)

: Fully-supervised 방법보다 낮은 성능

Table 1. Classification error on the ILSVRC validation set.

Networks	top-1 val. error	top-5 val. error
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
AlexNet*-GAP	44.9	20.9
AlexNet-GAP	51.1	26.3
GoogLeNet	31.9	11.3
VGGnet	31.2	11.4
AlexNet	42.6	19.5
NIN	41.9	19.6
GoogLeNet-GMP	35.6	13.9

- Fully Connected Layer 미사용 고려

➢ Acceptable한 결과

Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

Method	supervision	top-5 test error
GoogLeNet-GAP (heuristics)	weakly	**37.1**
GoogLeNet-GAP	weakly	42.9
Backprop [22]	weakly	46.4
GoogLeNet [24]	full	26.7
OverFeat [21]	full	29.9
AlexNet [24]	full	34.2

- Bounding Box 없이 학습
- 휴리스틱으로 더 높은 성능 획득

➢ 다양한 후처리를 통해 실험모델의 성능을 올릴 수 있는 여지 확보

실험 및 결과
[1] Weakly-supervised Object Localization

실험결과

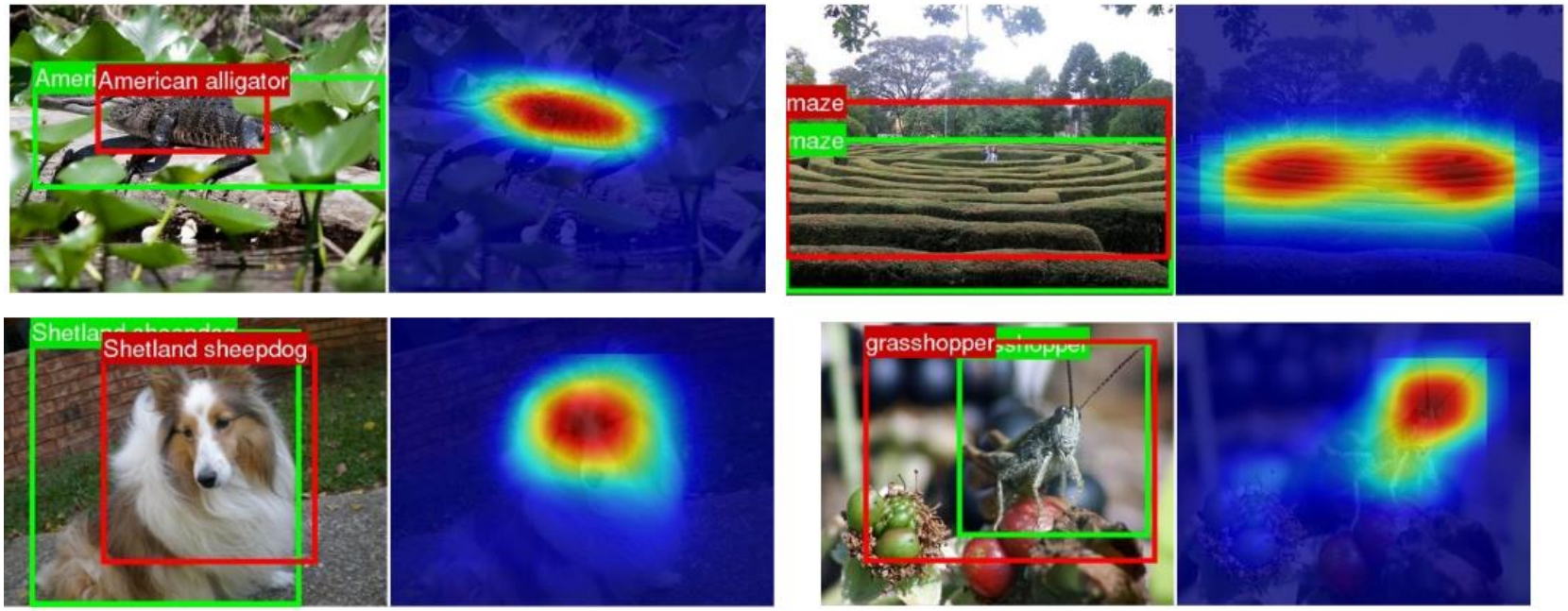
Table 2. Localization error on the ILSVRC validation set. *Back-prop* refers to using [22] for localization instead of CAM.

Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

- ✓ 뛰어난 Localization 성능 : CAM을 활용한 네트워크들이 bounding box를 annotation하여 backpropagation 시킨 다른 네트워크보다 더 좋은 성능을 보임
- ✓ GoogLeNet-GAP의 top-5 에러는 불과 43% : bounding box에 별다른 학습을 하지 않았다는 것을 고려할 때 매우 놀라운 결과

실험 및 결과
[1] Weakly-supervised Object Localization

실험결과 예시



Example of localization from GoogleNet-GAP

- Green Box : the ground-truth box
- Red Box : the predicted bounding box from the class activation map
  - ➢ CAM이 segment된 부분 중 20%가 넘는 부분들이 먼저 선택되었고, 이 부분들을 가장 많이 포함할 수 있는 box가 선택됨.

실험 및 결과
[2] Pattern Discovery

실험내용

이미지에서 물체를 추출하는 것 이외에 행위와 같은 모호한 패턴에 대한 개념도 잘 추출하는지에 대해 실험을 진행

1. 다양한 객체가 포함된 20개의 카테고리 이미지를 학습하고 각 카테고리로부터 비슷한 객체가 추출되는지 확인
2. 추상적인 설명과 이미지로부터 패턴을 추출할 수 있는지 여부 확인
3. CAM 방법을 이용해 텍스트를 포착할 수 있는지 여부 확인
4. 질문과 대답을 이용하여 학습한 후 CAM을 통해 시각화 하였을 때 대답이 있는 부분을 잘 포착하는지 확인

실험 및 결과
[2] Pattern Discovery

실험결과

2.1) Discovering informative objects in the scenes

: 비슷한 카테고리를 갖는 이미지에서는 비슷한 객체가 주로 추출

Dining room

Frequent object:
wall:0.99
chair:0.98
floor:0.98
table:0.98
ceiling:0.75
window:73

Informative object:
table:0.96
chair:0.85
chandelier:0.80
plate:0.73
vase:0.69
flowers:0.63

Bathroom

Frequent object:
wall: 1
floor:0.85
sink: 0.77
faucet:0.74
mirror:0.62
bathtub:0.56

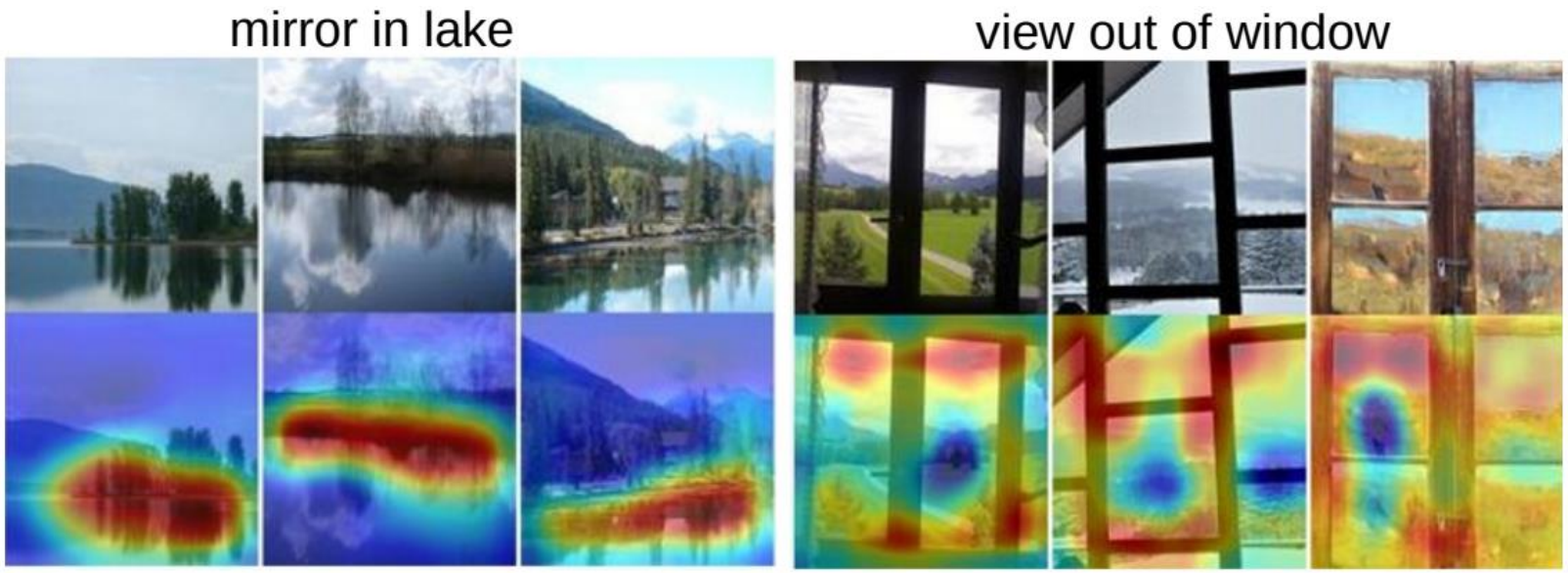
Informative object:
sink:0.84
faucet:0.80
countertop:0.80
toilet:0.72
bathtub:0.70
towel:0.54

실험 및 결과
[2] Pattern Discovery

실험결과

2.2) Concept localization in weakly labeled images

: 추상적인 설명이 제공된 이미지로 학습한 모델도 해당 정보가 포함된 위치를 잘 포착



실험 및 결과
[2] Pattern Discovery

실험결과

2.3) Weakly supervised text detector

: Bounding Box를 이용하지 않았음에도 글자 부분을 잘 포착하는 것을 확인

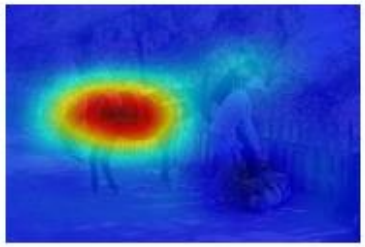
- Positive : 글자가 있는 이미지 / Negative : 글자가 없는 이미지

실험 및 결과
[2] Pattern Discovery

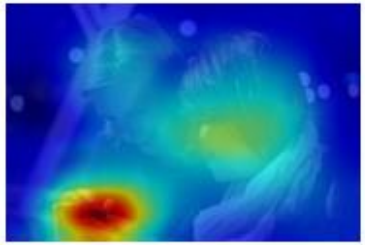
실험결과

2.4) Interpreting visual question answering

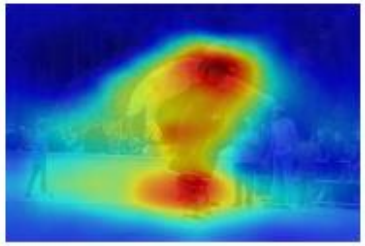
: 대답에 해당하는 물체의 위치를 잘 포착하는 것을 확인

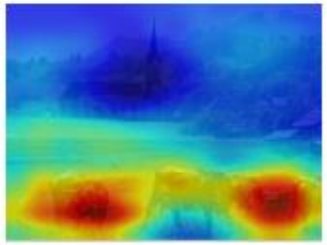
What is the color of the horse?
Prediction: brown

What are they doing?
Prediction: texting

What is the sport?
Prediction: skateboarding

Where are the cows?
Prediction: on the grass

결론 및 향후연구

결론

- FC-Layer 대신 GAP을 적용한 간단한 구조 변경으로 다양한 Task(Classification, Localization)를 수행할 수 있는 방법 제시한 효과적인 논문
  - CNN의 결과를 설명할 수 있는 CAM
  - 직접 학습하지 않고도 좋은 성능을 보이는 Weakly-supervised object localization
  - 다양한 실험을 통해 논문에서 주장한 구조의 장점을 명료하게 파악
  - 부가적으로 CNN의 작동 방식을 직관적으로 이해 가능

"CNN의 영혼을 잠시 보는 접근 방법"

향후연구

- CAM은 Global Average Pooling을 사용해야 한다는 단점
  - Grad-CAM, Grad-CAM++ 등 다양한 버전으로 업그레이드, XAI의 대표 모델로 발전

Thank you