2025 한양대학교

SW/ICT/AI 종합학술대회

IT/CSE 응용 연구개발 – 자람

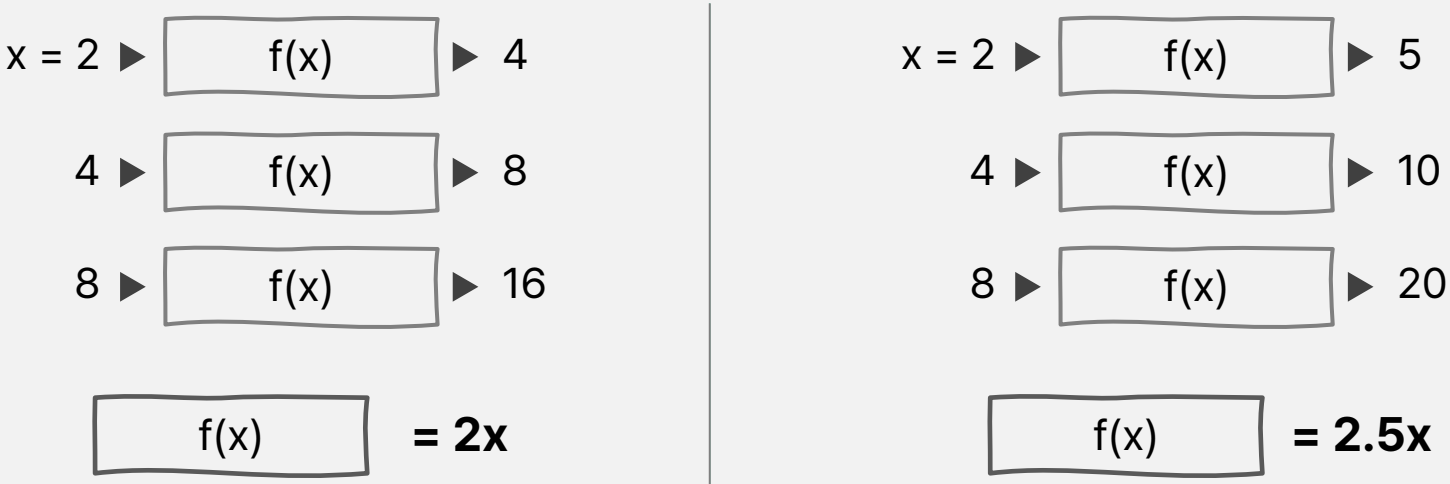
Weakly Supervised Learning and Class Activation Mapping

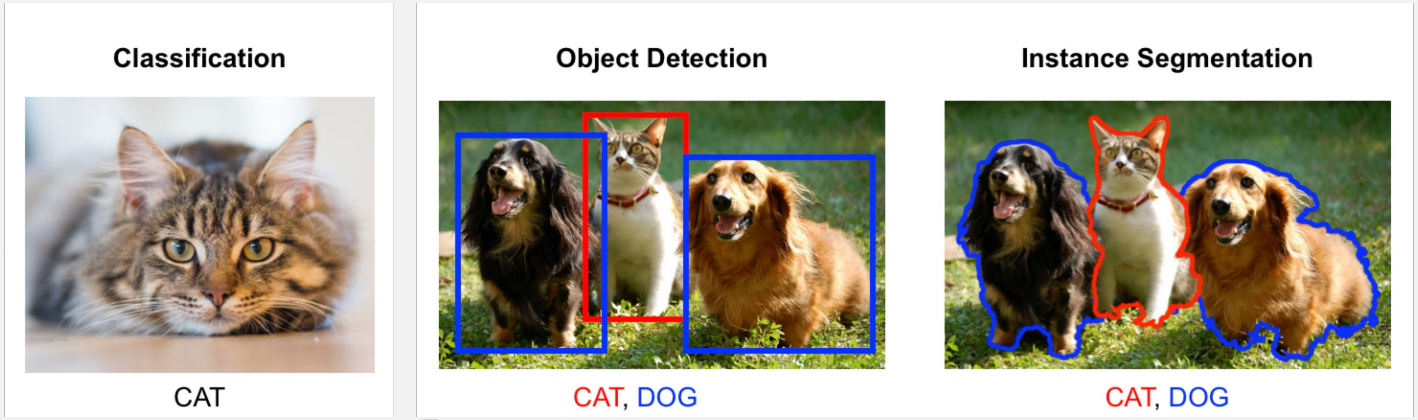
오다연(20학번)

2025. 11. 28.

한양대학교 ERICA 소프트웨어융합대학 / 한양대학교 ERICA SW중심대학사업단

Supervised Learning & Core Vision Tasks






```
[0, 1, 0, 0, ..., 0]
// class 1 = cat
```

```
[
  { "class": "dog",
    "bbox": [30, 60, 120, 160] },
  { "class": "cat",
    "bbox": [150, 70, 210, 160]}
]
```

```
[[0, 0, 0, 1, 1, 1, 0],
 [0, 0, 1, 1, 1, 1, 2],
 [0, 1, 1, 1, 2, 2, 2],
 ... ]
```

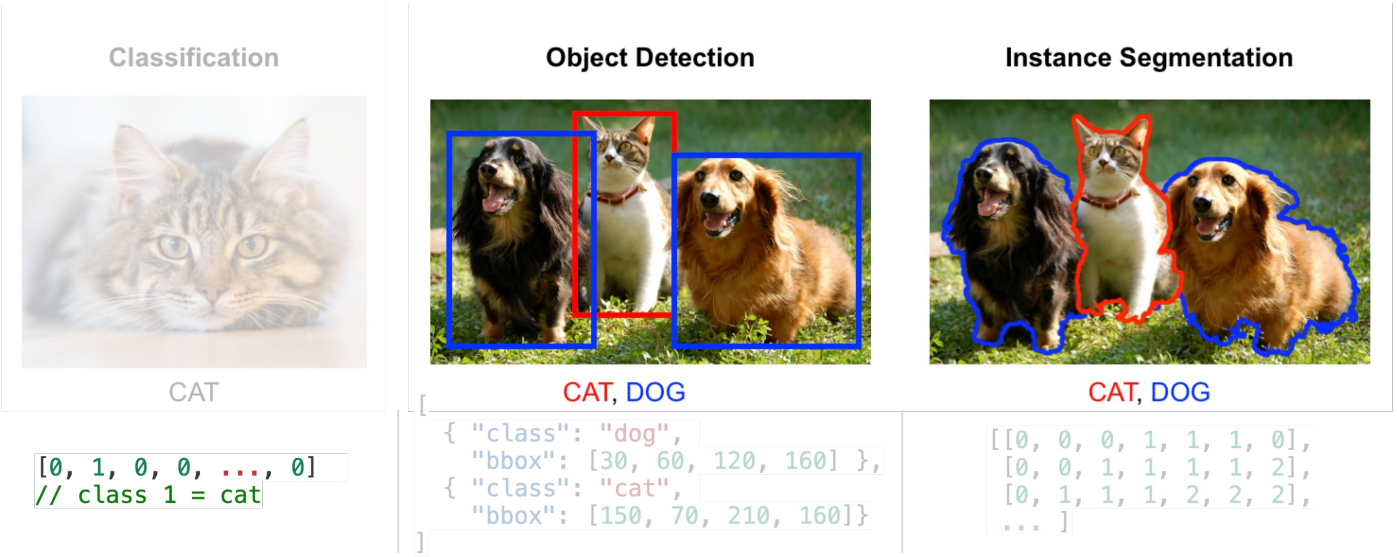

https://www.oreilly.com/library/view/deep-learning-for/9781788295628/4fe36c40-7612-44b8-8846-43c0c4e64157.xhtml

What is Weakly Supervised Learning?

Task	Classification	Detection	Segmentation
X	image	image	Image
Y	Class **(저비용)**	Bounding Box	Mask **(고비용)**

▼ ▼ ▼

Detection 모델 학습을 class만으로 할 수 있을까?
Segmentation 모델 학습을 detection 또는 class만으로 할 수 있을까?



https://www.oreilly.com/library/view/deep-learning-for/9781788295628/4fe36c40-7612-44b8-8846-43c0c4e64157.xhtml

Class Activation Mapping (CAM)

arXiv:1512.04150v1 [cs.CV] 14 Dec 2015

Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu

Abstract

In this work, we revisit the global average pooling layer proposed in [13], and shed light on how it explicitly enables the convolutional neural network to have remarkable localization ability despite being trained on image-level labels. While this technique was previously proposed as a means for regularizing training, we find that it actually builds a generic localizable deep representation that can be applied to a variety of tasks. Despite the apparent simplicity of global average pooling, we are able to achieve 37.1% top-5 error for object localization on ILSVRC 2014, which is remarkably close to the 34.2% top-5 error achieved by a fully supervised CNN approach. We demonstrate that our network is able to localize the discriminative image regions on a variety of tasks despite not being trained for them.

1. Introduction

Recent work by Zhou *et al* [33] has shown that the convolutional units of various layers of convolutional neural networks (CNNs) actually behave as object detectors despite no supervision on the location of the object was provided. Despite having this remarkable ability to localize objects in the convolutional layers, this ability is lost when fully-connected layers are used for classification. Recently some popular fully-convolutional neural networks such as the Network in Network (NIN) [13] and GoogLeNet [24] have been proposed to avoid the use of fully-connected layers to minimize the number of parameters while maintain-



Figure 1. A simple modification of the global average pooling layer combined with our class activation mapping (CAM) technique allows the classification-trained CNN to both classify the image and localize class-specific image regions in a single forward-pass e.g., the toothbrush for *brushing teeth* and the chainsaw for *cutting trees*.

CNN trained on object categorization is successfully able to localize the discriminative regions for action classification as the objects that the humans are interacting with rather than the humans themselves.

Despite the apparent simplicity of our approach, for the weakly supervised object localization on ILSVRC benchmark [20], our best network achieves 37.1% top-5 test error, which is rather close to the 34.2% top-5 test error achieved by fully supervised AlexNet [10]. Furthermore, we demonstrate that the localizability of the deep features in our approach can be easily transferred to other recognition datasets for generic classification, localization, and concept discovery.[1].

CVPR 2015
BOSTON
JUNE 7-12

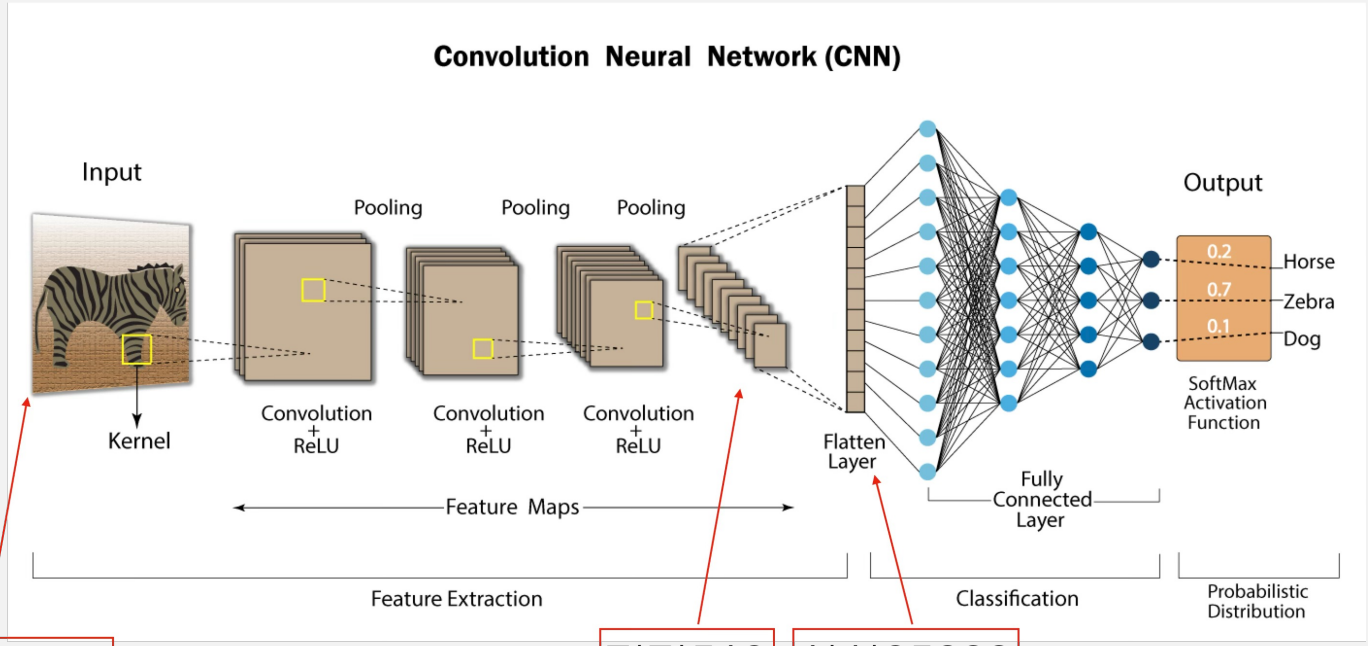
Learning deep features for discriminative localization
B Zhou, A Khosla, A Lapedriza, A Oliva, A Torralba
Proceedings of the IEEE conference on computer vision and …, 2016 • openaccess.thecvf.com

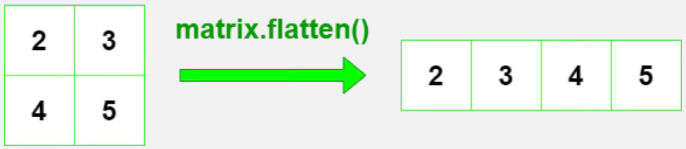
Abstract
In this work, we revisit the global average pooling layer proposed in [13], and shed light on how it explicitly enables the convolutional neural network (CNN) to have remarkable localization ability despite being trained on image-level labels. While this technique was previously proposed as a means for regularizing training, we find that it actually builds a generic localizable deep representation that exposes the implicit attention of CNNs on image. Despite the apparent simplicity of global average pooling, we are able to achieve

SHOW MORE

☆ Save Cite Cited by 14165 Related articles All 20 versions

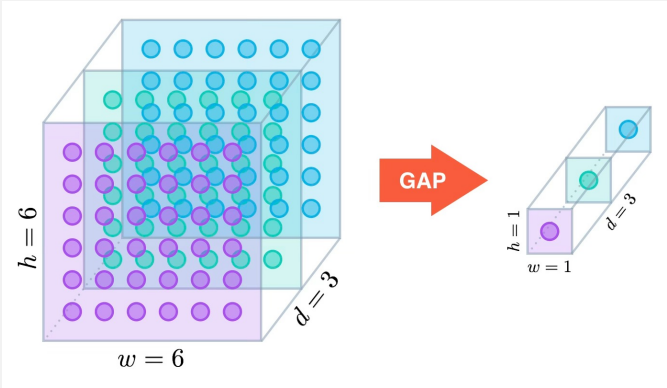
CAM – Problem & Motivation







CAM - Method



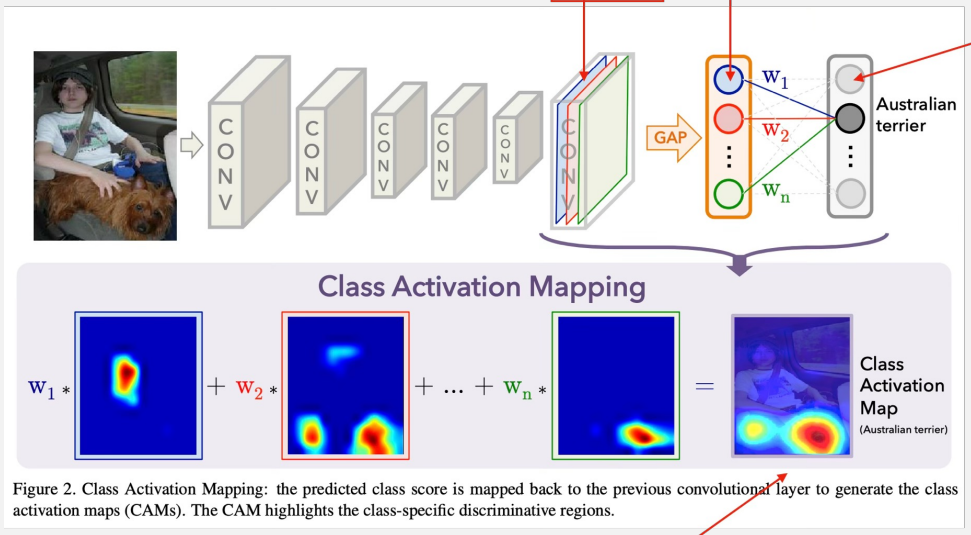


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

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

$$M_c(x, y) = \sum_k w_k^c f_k(x, y).$$

1. $f_k(x, y)$
   - 마지막 conv 레이어의 k번째 feature map
2. F_k - GAP 연산
   - k번째 feature map을 모두 합함.
     (평균 연산은 생략. 어차피 후속 레이어가 가중치 학습 과정에서 스케일이 조정됨)
3. w_k^c
   - 클래스 c에 대한, 채널 k의 중요도(가중치)
     (class c에 대한 softmax 레이어의 가중치)
4. S_c - Class 별 Score 연산
   - 2과 3를 곱하여 얻은 logit(score)
5. $M_c(x, y)$ - Class Activation Map 연산
   - 3과 1을 곱하여 얻은 map

CAM - Experiments

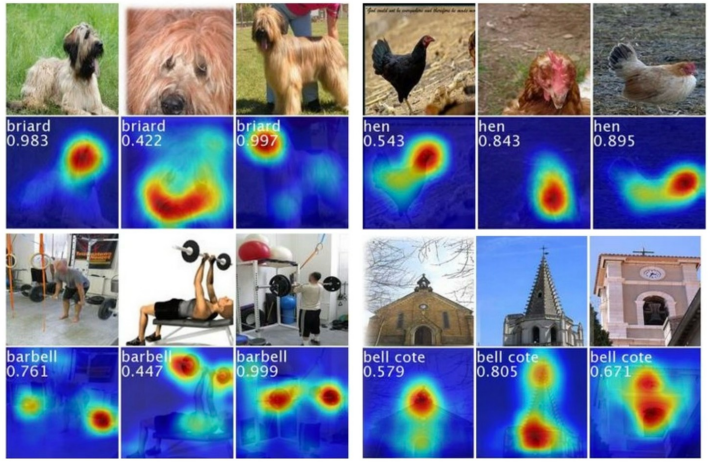
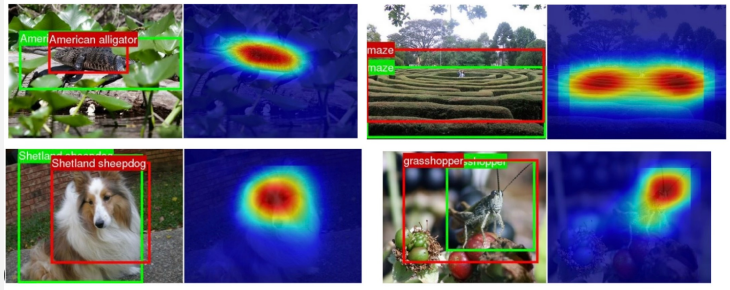


Figure 3. The CAMs of four classes from ILSVRC [20]. The maps highlight the discriminative image regions used for image classification e.g., the head of the animal for *briard* and *hen*, the plates in *barbell*, and the bell in *bell cote*.



GT / Predicted

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

Top-1 error: 모델이 1등으로 고른 클래스가 오답인 비율.
Top-5 error: 상위 5개 예측 중 정답이 없을 비율 (낮을수록 좋음)

Table 2. Localization error on the ILSVRC validation set. *Backprop* 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

Top-1 error: 모델의 (class, bbox) 예측에 대해 class정답 & IoU ≥ 0.5
Top-5 error: 상위 5개 예측 중 하나라도 class정답 & IoU ≥ 0.5 (낮을수록 좋음)

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

Top-5 error: 상위 5개 예측 중 하나라도 class정답 & IoU ≥ 0.5 (낮을수록 좋음)

Summary & Discussion

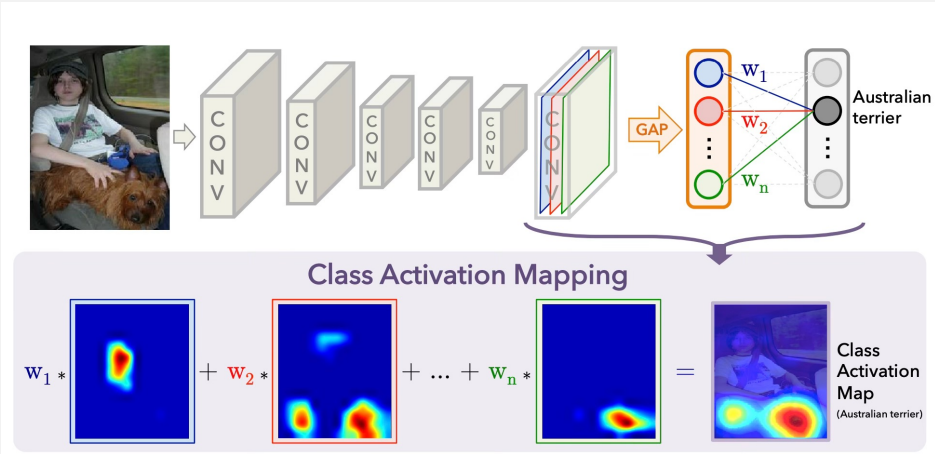


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

* 요약: 기존 CNN의 대형 FC layer를 GAP + soft-max로 교체하는 간단한 기법을 제안함
* 의의
  * Bounding box annotation 없이 학습
  * Classification 정확도 유지 (3% 미만의 손실)
  * 네트워크가 참고한 시각적 근거(heat-map) 제공
* 성능: GoogLeNet-GAP: 37.1 % Top-5 localization error

Limitations

저해상도 heat-map

단일 인스턴스 가정

가장 판별적인 부분만 강조하게 됨

Threshold와 같은 휴리스틱에 의존

GAP를 사용이 불가능한 네트워크에는 적용이 어려움

Aftermath

GradCam, GradCAM++

