CAM vs. Grad CAM

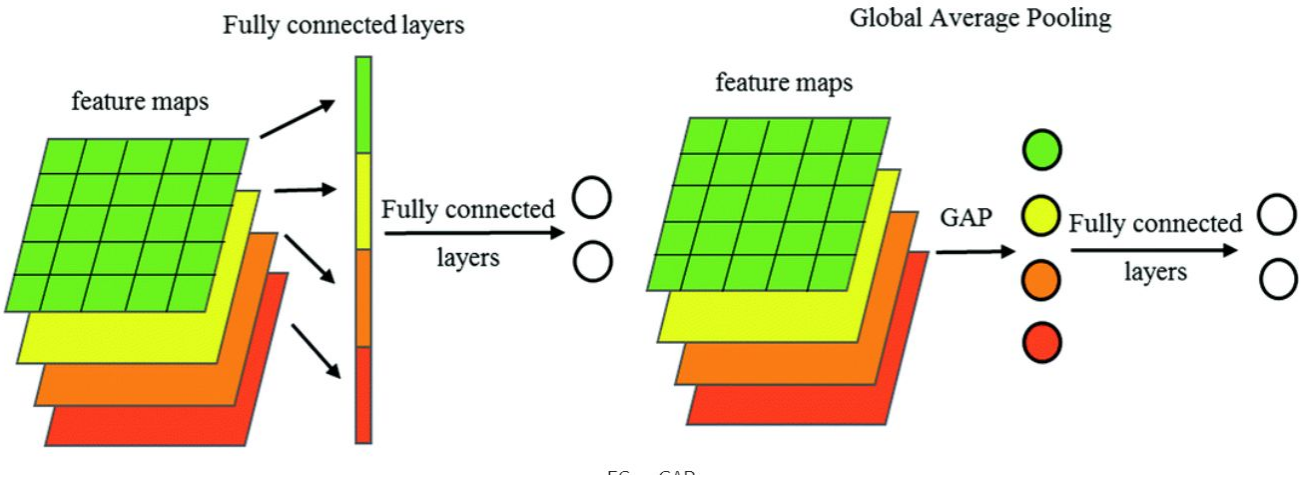
케라수요일 3시

개요

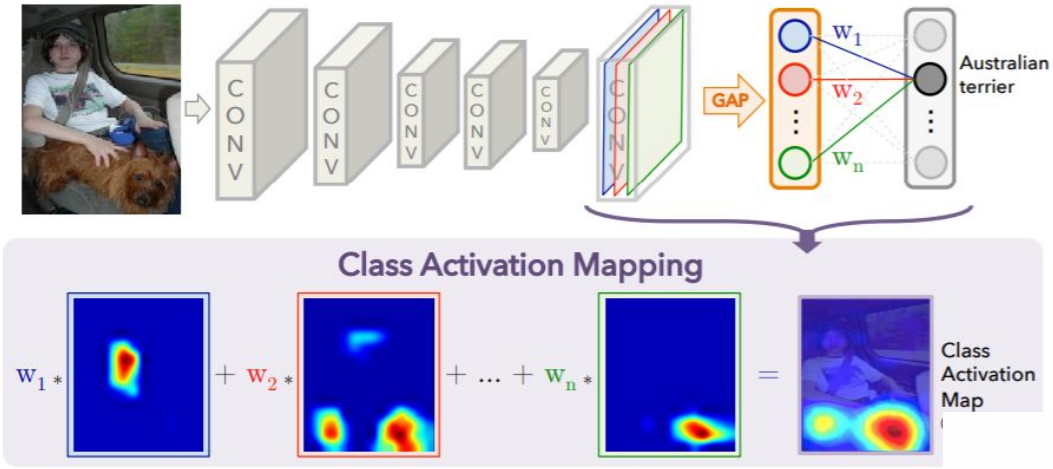
- CAM 소개
  - CAM이 뭘까?
- GAP (Global Average Pooling)
  - CNN과 CAM의 Filter
  - GAP (Global Average Pooling)
  - CAM (Class Activation Map)
- Grad CAM 소개
  - Grad CAM의 배경
  - 수식 유도
- Grad CAM vs. CAM
  - Grad CAM과 CAM의 차이

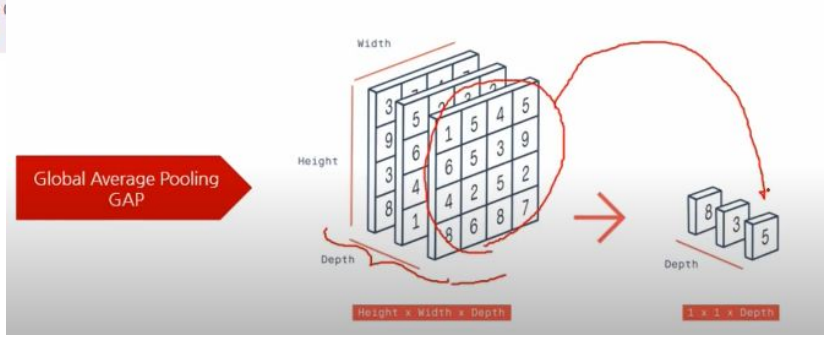
CAM (Class Activation Map)

CNN 모델의 예측 결과를 설명하기 위해 Activation Map 과 관련해서 CAM과 Grad-CAM이 사용



동작 원리


CONV
CONV
CONV
CONV
CONV
CONV
GAP
w_1
w_2
w_n
Australian terrier
Class Activation Mapping
w_1 *
+ w_2 *
+ ... + w_n *
=
Class Activation Map


Global Average Pooling
GAP
Width
Height
Depth
Height x Width x Depth
Depth
1 x 1 x Depth

https://www.youtube.com/watch?v=Z8JMI9ZyL1w

가중치 = GAP가 특정클래스에 미치는 영향

By plugging $F_k = \sum_{x,y} f_k(x,y)$ into the class score, S_c, we obtain

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y). \quad (1)$$

We define M_c as the class activation map for class c, where each spatial element is given by

$$M_c(x,y) = \sum_k w_k^c f_k(x,y). \quad (2)$$

Mc는 특정 클래스의 Class Activation Map 정의

Thus, $S_c = \sum_{x,y} M_c(x,y)$, and hence $M_c(x,y)$ directly indicates the importance of the activation at spatial grid (x,y) leading to the classification of an image to class c.

Class Activation Map이 해당 Class에 대한 중요도를 나타냄

F^k는 $\sum_{x,y} f_k(x,y)$

fk가 k번째 Activation map

Fk = actiavation map의 픽셀 합 / 픽셀 개수

S_c, is $\sum_k w_k^c F_k$

Sc는 class c에 해당하는 softmax layer에 들어가는 값

w : 특정 클래스 c를 만드는 것에 k번째 GAP가 미치는 영향 w

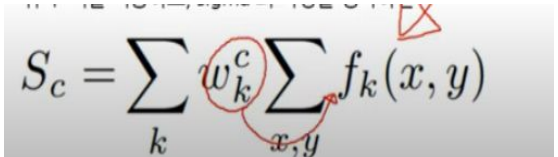
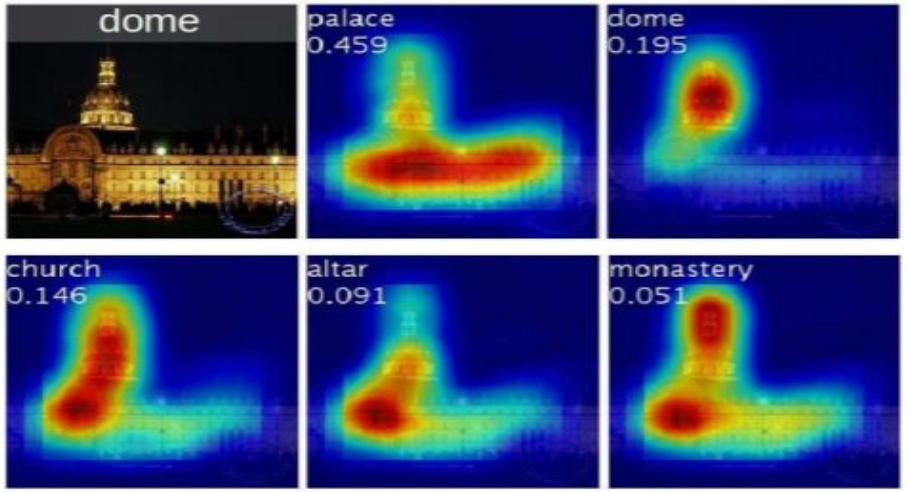


Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

Method	supervision	top-5 test error
GoogLeNet-GAP (heuristics)	weakly	**37.1**
GoogLeNet-GAP	weakly	42.9
Backprop [23]	weakly	46.4
GoogLeNet [25]	full	26.7
OverFeat [22]	full	29.9
AlexNet [25]	full	34.2

Figure 3. The CAMs of two classes from ILSVRC [21]. The maps highlight the discriminative image regions used for image classification, the head of the animal for *briard* and the plates in *barbell*.



CNN이 왜 그렇게 판단했는지 클래스별로 Highlighting 하는 방법 = CAM

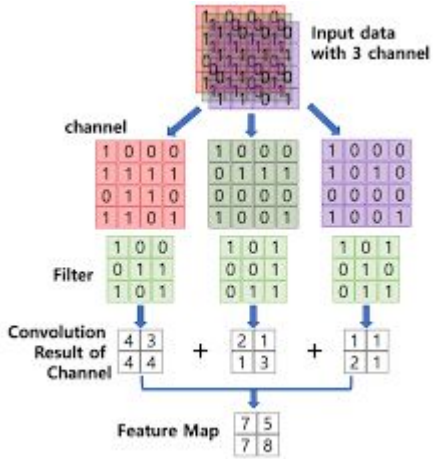
바운딩박스없이 이미지에 대한 라벨만으로 약간 더 높은 오류가 있지만 객체탐지 가능

2. GAP (Global Average Pooling)

- CNN과 CAM의 Filter
- GAP layer (Global Average Pooling)
- CAM (Class Activation Map)

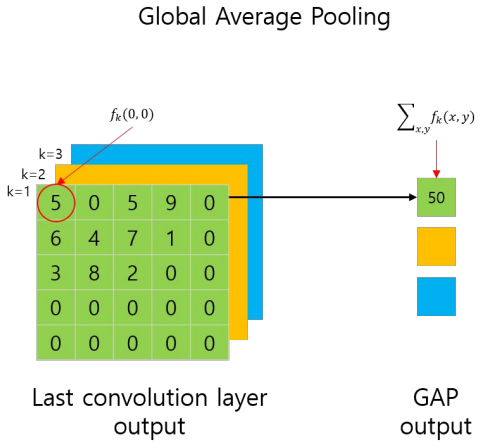
Filter

이미지의 특징을 추출
하나의 filter는 하나의 특징을 추출한다.



CNN의 filter

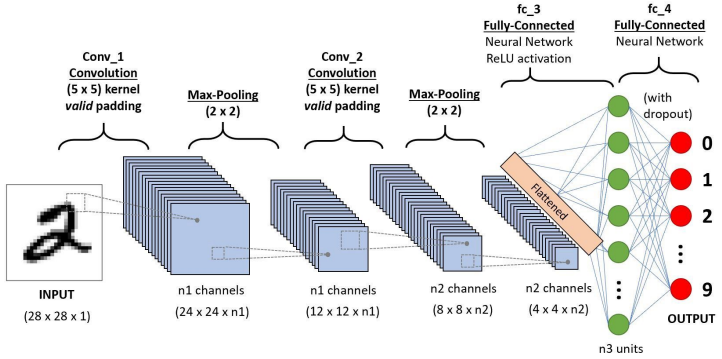
- 합성곱 계산
- 각 채널별로 하나의 feature map 반환



CAM의 filter

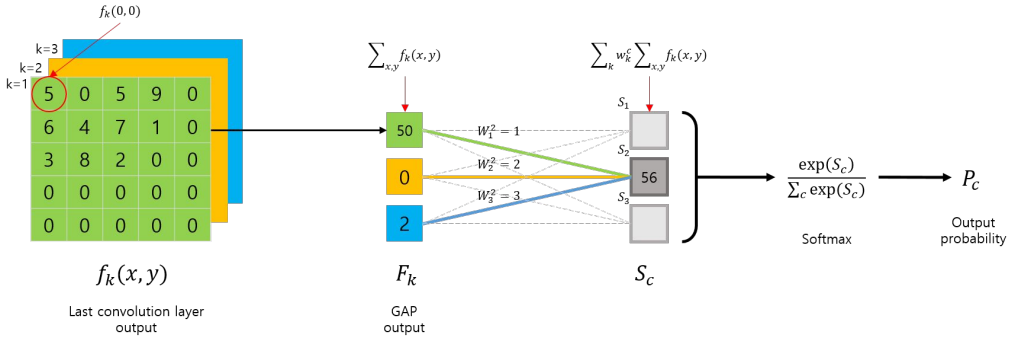
- GAP (Global Average Pooling)
- 각 채널별로 하나의 값 반환
- 하나의 특성을 하나의 값으로 대응한 후, FC layer로 전달.

CNN의 Issue - 위치정보의 손실



특징들을 추출한 여러 개의 필터들로 인해 긴 형태의 output activation map이 생성된다.
이를 FC layer로 연결하면, spatial information이 손실된다.

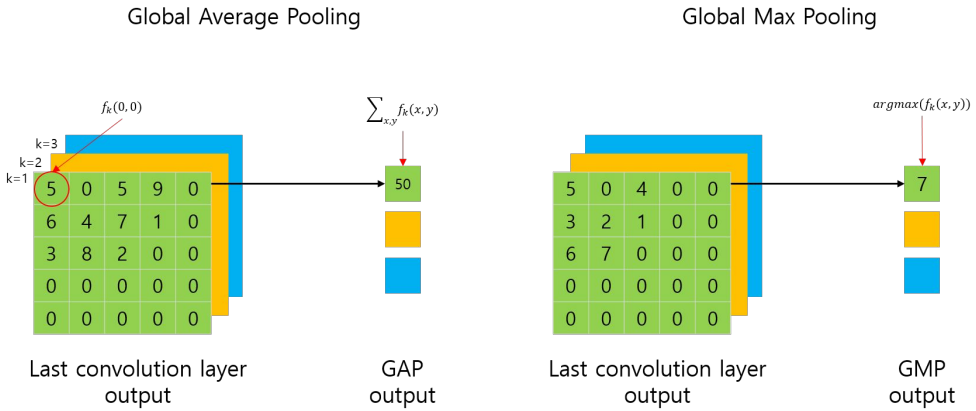
CAM의 대안 - **GAP layer** (Global Average Pooling)



특성 하나를 하나의 값으로 대응시켜서 FC layer에 연결.

→ 필요한 파라미터 수를 줄이고, 위치정보를 유지할 수 있다.

GAP layer (Global Average Pooling)



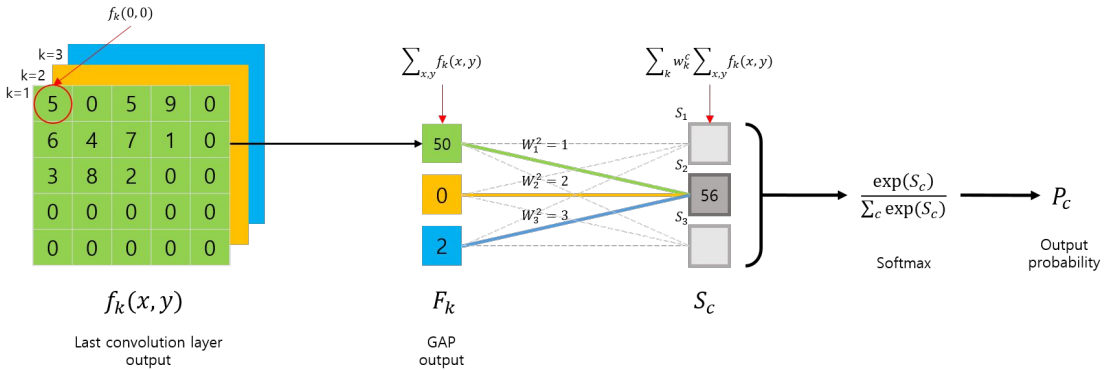
GAP : 전역적으로(Global) Average Pooling을 수행하는 layer

GMP(Global Max Pooling): 전역적으로 Max Pooling을 수행하는 layer

kernel size = layer의 input size
연산 결과 각 채널별로 **하나의 값**이 나온다.

5x5 feature map의 채널이 3개이지만,
feature map의 크기와 상관없이, **채널의 개수에 해당하는 3개의 값(1x1)**이 출력

GAP layer (Global Average Pooling)



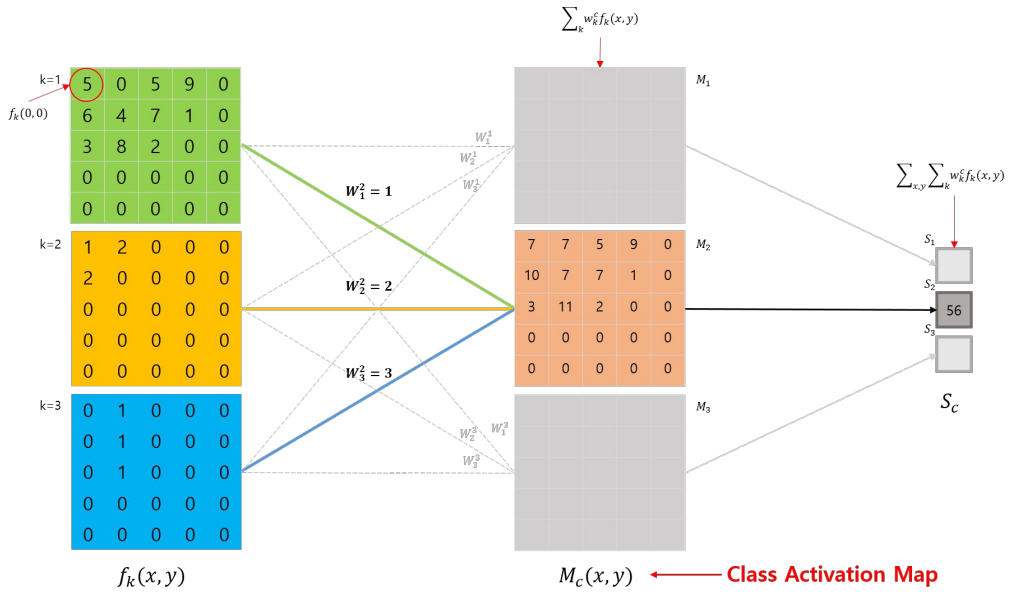
$$F_k = \sum_{x,y} f_k(x,y)$$

마지막 Convolution layer 에서 출력된 feature map fk(x,y) 는 GAP 연산이 취해지며 k개의 값이 출력된다. ⇒ 50, 0, 2

$$\begin{aligned} S_c &= \sum_k w_k^c F_k \\ &= \sum_k w_k^c \sum_{x,y} f_k(x,y) \\ &= \sum_{x,y} \sum_k w_k^c f_k(x,y) \end{aligned}$$

이후 출력된 GAP값인 Fk는 CNN의 마지막 출력 layer인 Sc로 전달되면서 linear combination(weighted sum)을 수행한다. ⇒ (50*1)+(0*2)+(2*3) = 56

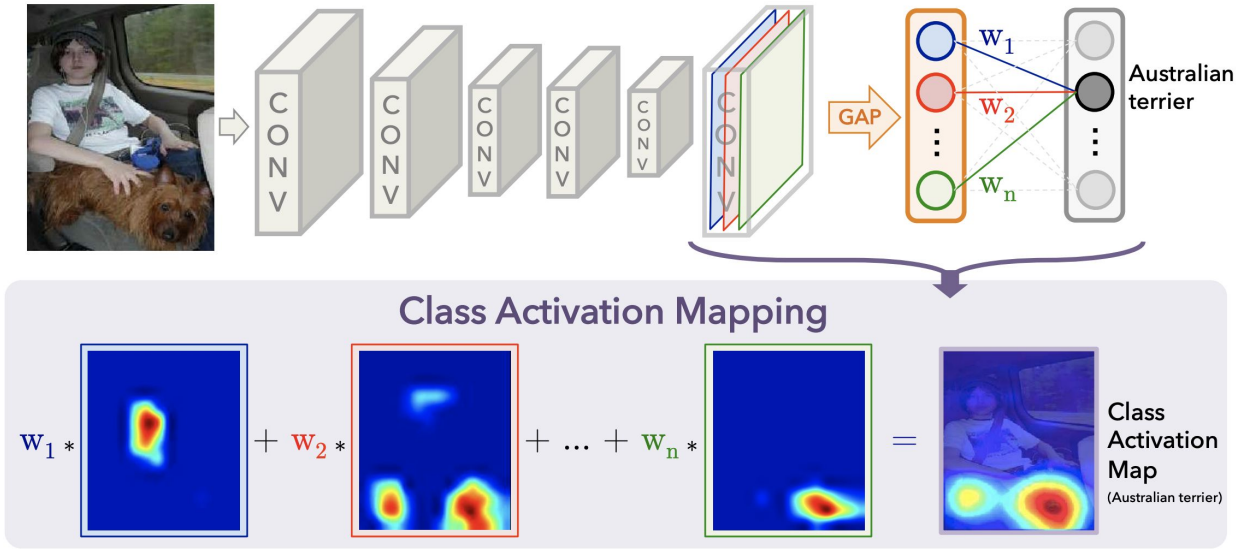
CAM (Class Activation Map)



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

특정 클래스 c=2를 구별하기 위해 이 클래스에 연결된 weights와 각 feature map에 대해 linear combination(weighted sum)을 취한 결과가 바로 **CAM (Class Activation Map)** 이다.

CAM (Class Activation Map)



GAP 값과 특정 클래스(ex, 강아지)와 연결된 가중치는 해당 activation map이 특정 클래스에 미치는 전반적인 영향과 중요도를 나타낸다.

→ **Black Box 문제의 해결**: 학습된 weight값들을 통해 중요한 이미지에 집중하여 좋은 classification결과를 낼 수 있다.

CAM vs Grad-CAM

CAM의 한계:

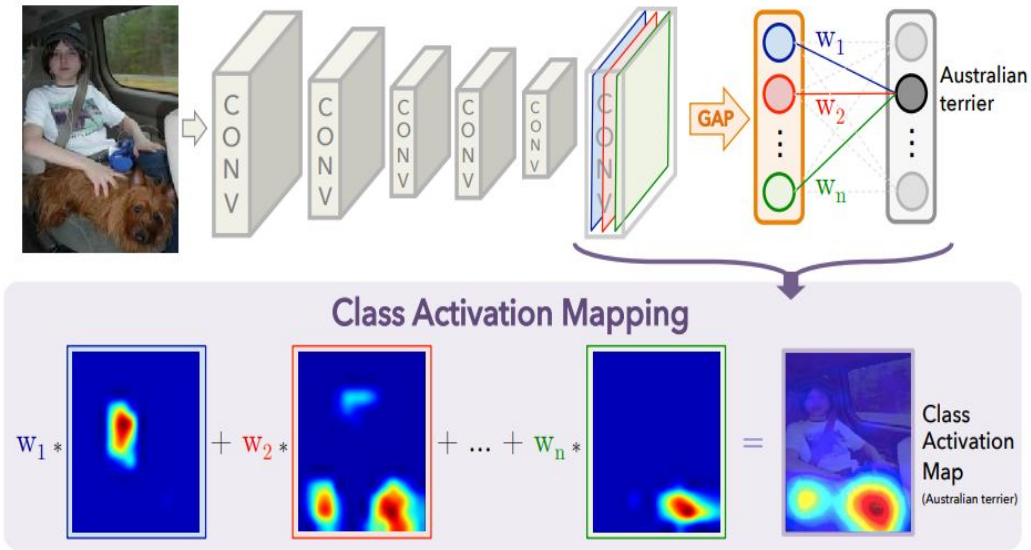
(1) Fully-connected layers가 포함된 모델에 적용 불가능 (VGG16)

(2) Softmax layer의 바로 직전 피처맵이 필요하므로 GAP를 거치는 특정 구조에만 적용할 수 있다.(모델의 변형 필요)

(상대적으로 이미지 분류와 같은 특정 task에 대한 모델들 보다 성능이 떨어지며, 어떤 task에는 적용이 불가능)

(3) 오브젝트 디텍션 외에 Visual Question Answer(VQA)나 Captioning처럼 다양한 목적을 수행하는 CNN에 CAM을 적용하기 어렵다.

기존 CAM



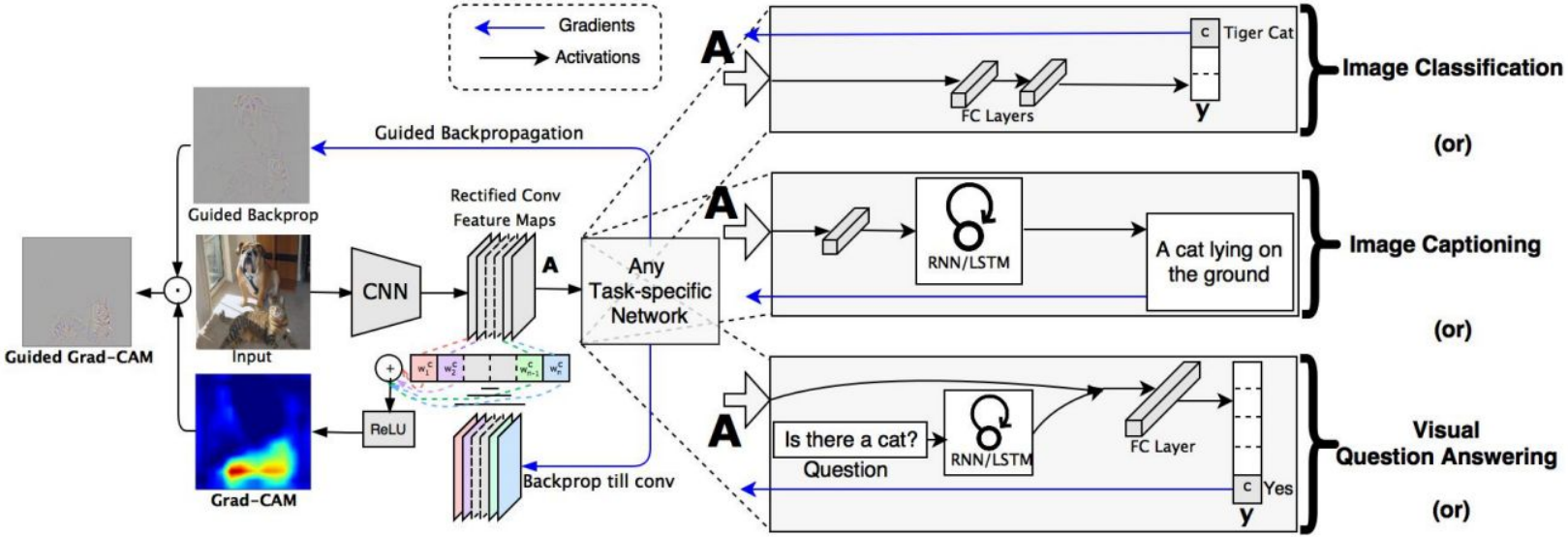
$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y). \quad (1)$$

$$M_c(x,y) = \sum_k w_k^c f_k(x,y). \quad (2)$$

GAP 층이 필요.

(학습을 통해 나온 가중치 **w**가 있어야 함)

Grad-CAM



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

CAM vs. Grad CAM (formula)

CAM :

$$L^c_{CAM} = \sum_k w^c_k A^k$$

Grad-CAM :

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k)$$

- w^c_k : fc층을 GAP로 구조를 변경하고 그 결과값을 softmax에 넣어 학습시키는데, 여기서 학습되는 벡터
- α^c_k : 이미 학습한 모델의 softmax layer 인풋의 (피처맵에 대한) 편미분 값을 GAP 방식으로 구한 가중치

✓ CAM은 결국 마지막 conv층에서 GAP를 진행하고, Grad-CAM은 conv층마다 그래디언트를 전파하는데 마지막 conv층에 전달하는 그래디언트에 대해 GAP를 거치게 되면 결국 CAM이 된다.

✓ 즉, CAM은 Grad-CAM의 특정 사례라고 볼 수 있고, 그래서 Grad-CAM이 CAM의 일반화된 것이라고 표현한다.

$$F^k = \frac{1}{Z}\sum_i\sum_j A_{ij}^k$$

GAP한 아웃풋

$$Y^c = \sum_k w_k^c \cdot F^k$$

final score

$$\frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

$$Y^c \Rightarrow \frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

$$F^k = \frac{1}{Z}\sum_i\sum_j A_{ij}^k$$

GAP한 아웃풋

$$\Longrightarrow \quad \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

$$Y^c = \sum_k w_k^c \cdot F^k$$

final score

$$Y^c \Rightarrow \frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$$

$$\frac{\partial Y^c}{\partial F^k} = w_k^c$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

$$F^k = \frac{1}{Z}\sum_i\sum_j A_{ij}^k$$

GAP한 아웃풋

$$\Longrightarrow \quad \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

$$Y^c = \sum_k w_k^c \cdot F^k$$

final score

$$Y^c \Rightarrow \quad \frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$$

$$\frac{\partial Y^c}{\partial F^k} = w_k^c$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

$$F^k = \frac{1}{Z}\sum_i\sum_j A_{ij}^k$$

GAP한 아웃풋

$$\Longrightarrow \quad \frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$$

$$Y^c = \sum_k w_k^c \cdot F^k$$

final score

$$Y^c \Rightarrow \quad \frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$$

$$\frac{\partial Y^c}{\partial F^k} = w_k^c$$

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

$$\downarrow$$

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$$

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k} \longrightarrow \sum_i \sum_j w_k^c = \sum_i \sum_j Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$$

$$Z = \sum_i \sum_j 1$$

$$w_k^c = \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \longleftarrow Z w_k^c = Z \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$

$$w_k^c = \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$$

마지막에 나온 식
(CAM에서의 weight)

$$\alpha_k^c = \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

(Grad CAM에서의 알파값)

아하! CAM은 Grad-CAM의 케이스 중 하나군!
(네트워크의 마지막이 GAP층인 케이스)

CAM vs. Grad CAM (output)

먼저 시각적으로 큰 차이가 있어보이지는 않는다. 다만 히트맵에 대한 ReLU 연산으로 인해 Grad-CAM의 중심부와 외곽 영역의 색깔 차이가 더 두드러져 보인다.

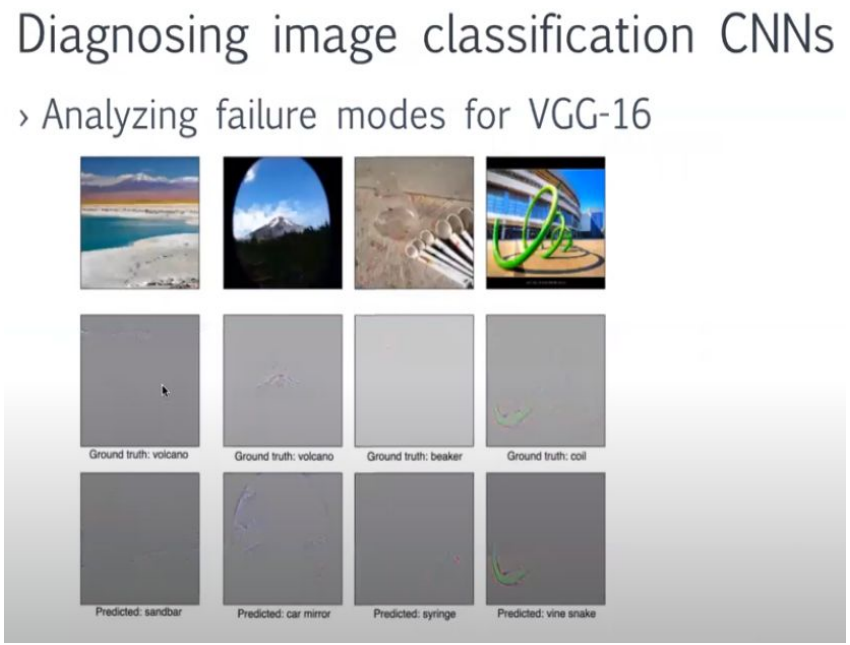
original_image

CAM

Grad-CAM

Grad CAM (활용)

잘못된 classification의 근거



Grad CAM (활용)

› Identifying bias in dataset

Ground Truth	Predicted: doctor	Predicted: nurse	total
doctor	79	34 (22 female)	113
nurse	7 (6 male)	106	113
total	86	140	

(a) Confusion Matrix for model trained with biased examples from search engine. Note that the text in () indicate the number of mistakes made due to gender bias.

Ground Truth	Predicted: doctor	Predicted: nurse	total
doctor	101	12 (6 female)	113
nurse	10 (3 male)	103	113
total	111	115	

(b) Confusion Matrix for model trained after correcting dataset bias learned from Grad-CAM visualizations. See that mistakes due to gender bias has reduced significantly.

Grad CAM (활용)

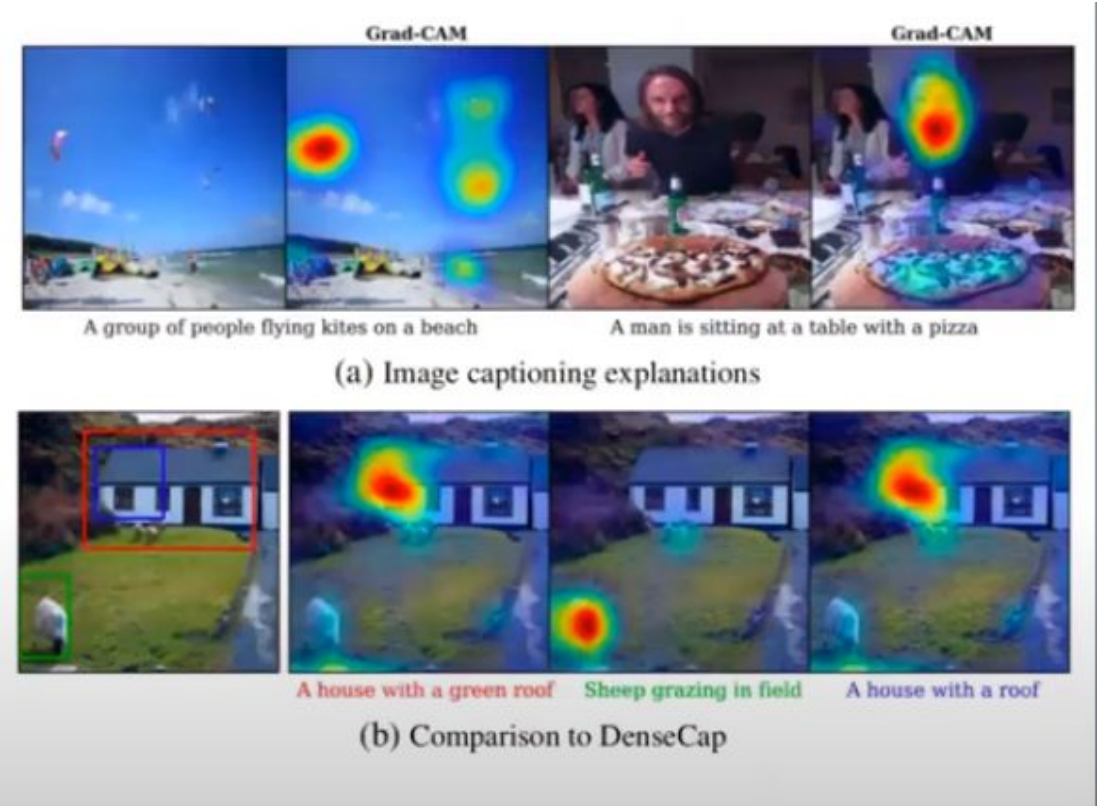
Counterfactual explanations

› Use negative values to find regions that decreases output score

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{-\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{Negative gradients}}$$

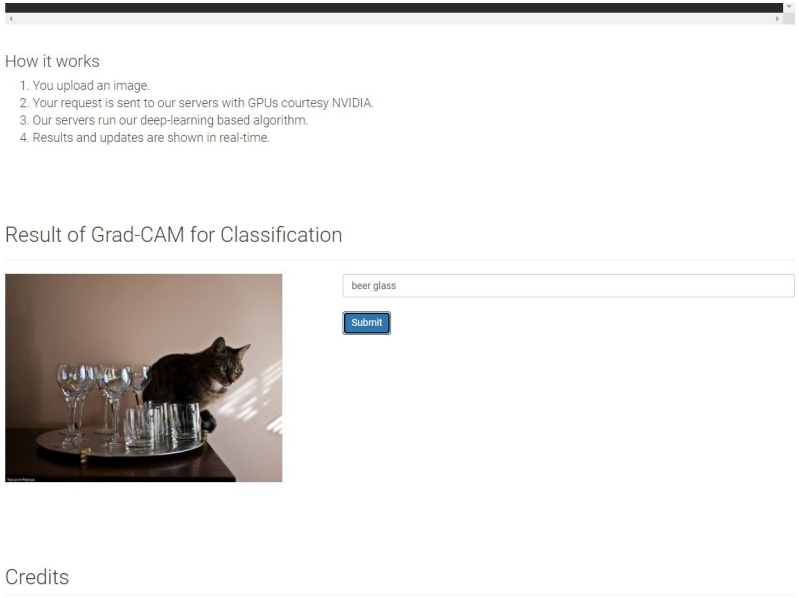
(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Grad CAM (활용)


Grad-CAM
Grad-CAM
A group of people flying kites on a beach
A man is sitting at a table with a pizza
(a) Image captioning explanations
A house with a green roof
Sheep grazing in field
A house with a roof
(b) Comparison to DenseCap

번외

Grad-CAM: Gradient-weighted Class Activation Mapping



안 되는 것 같다..