Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra

(ICCV'17)

발표자 : 김수지

목차

1. Introduction

Computer vision

- 사진/영상 등의 시각적 입력에서 정보를 인식 및 추출하는 알고리즘의 집합
- 인공 신경망을 사용하는 분야 내에서는 이미지 분류, 객체 탐지, 이미지 분할 등이 존재
- 대부분 지도학습 기반이라 사진/영상에 대한 정답(Label)이 필요

이미지 분류
(Image Classification)

주디

이미지 분류 및 위치 파악
(Image Localization)

주디

객체 탐지
(Object Detection)

닉 주디

1. Introduction

Weakly supervised learning

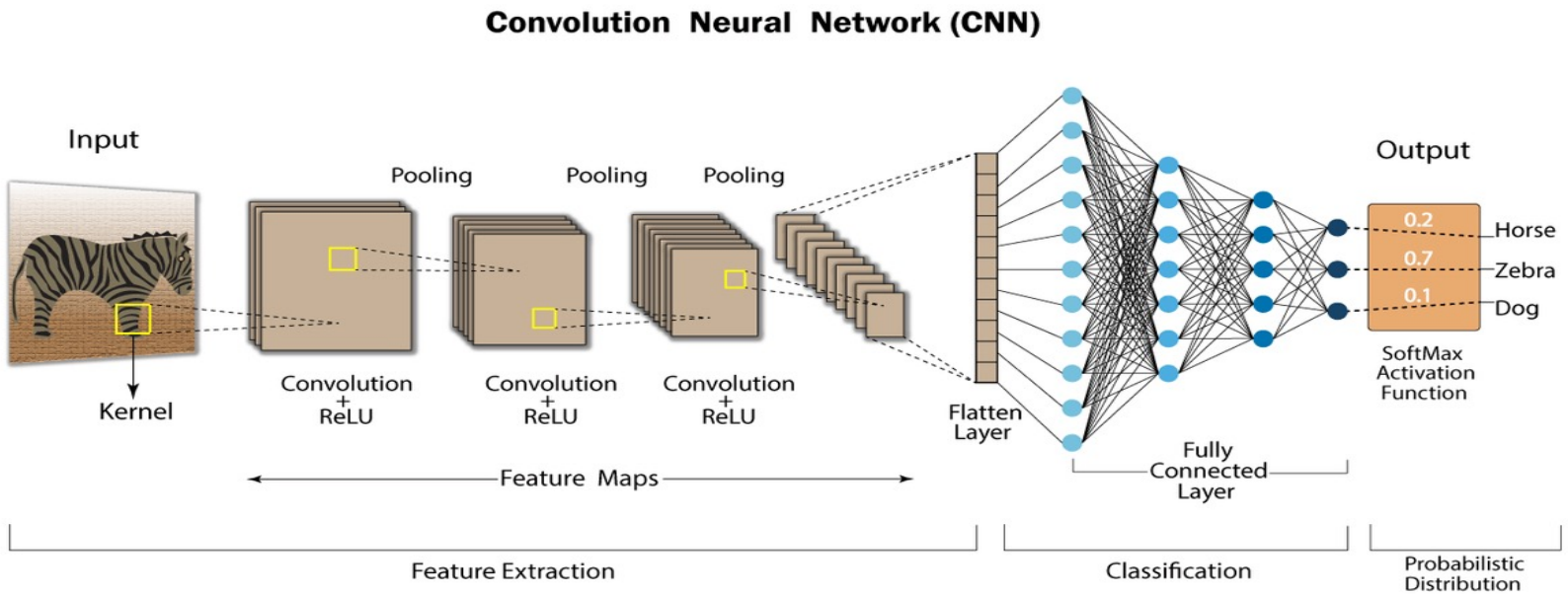
- 학습 이미지에 대한 정보보다 예측해야 할 정보가 더 디테일한 경우
- 라벨링에 대한 cost 부담 줄여줌

Training	Test
Image label	Bounding box
Bounding box	Pixel label

1. Introduction

Convolution Neural Networks (CNN)

- 이미지 데이터의 특성을 잘 반영할 수 있는 인공신경망 모델
- 일반적인 CNN은 Convolution 연산, Activation 연산, Pooling 연산의 반복으로 구성됨
- 일정 횟수 이상의 Feature Learning 과정 이후에는 Flatten 과정을 통해 이미지가 1차원의 벡터로 변환됨



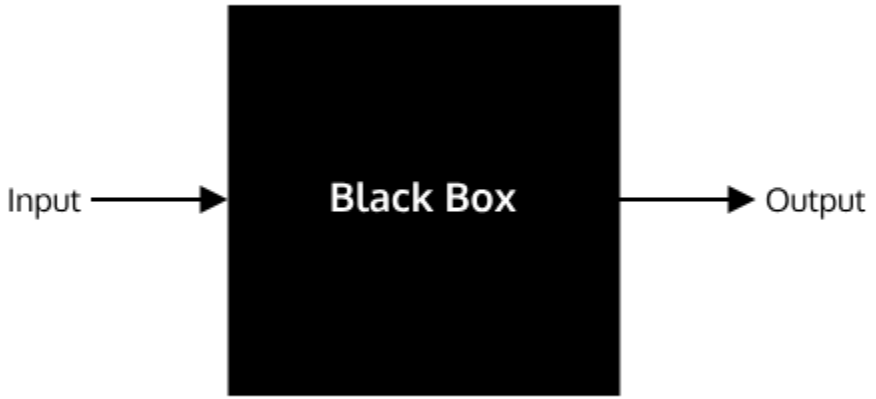
→ Why?

1. Introduction

CNN

- CNN에 기초한 Deep neural model은 이미지 분류, 객체 탐지 등 다양한 분야에서 좋은 성과를 거둠
- 하지만 왜 그렇게 분류, 예측했는지 이유를 알 수 없음
- 이러한 'Black box'를 해결하기 위하여 다양한 연구 진행되고 있음

  Ex) XAI (eXplainable Artificial Intelligence)



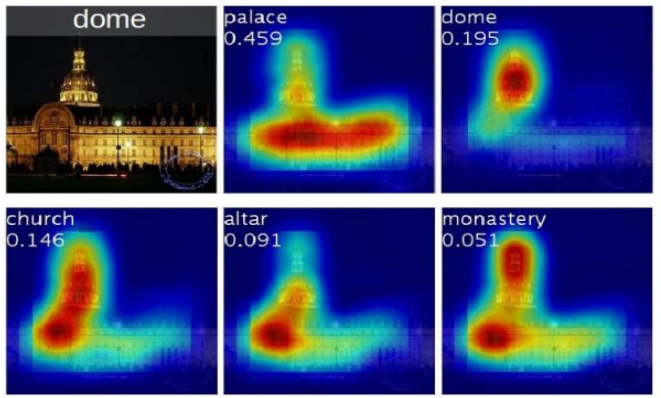
2. Related works

Class Activation Map (CAM) – 2016

- CNN 모델로 예측 시, 어떤 부분이 class 예측에 큰 영향을 주었는지 확인하려는 시도

Learning Deep Features for Discriminative Localization

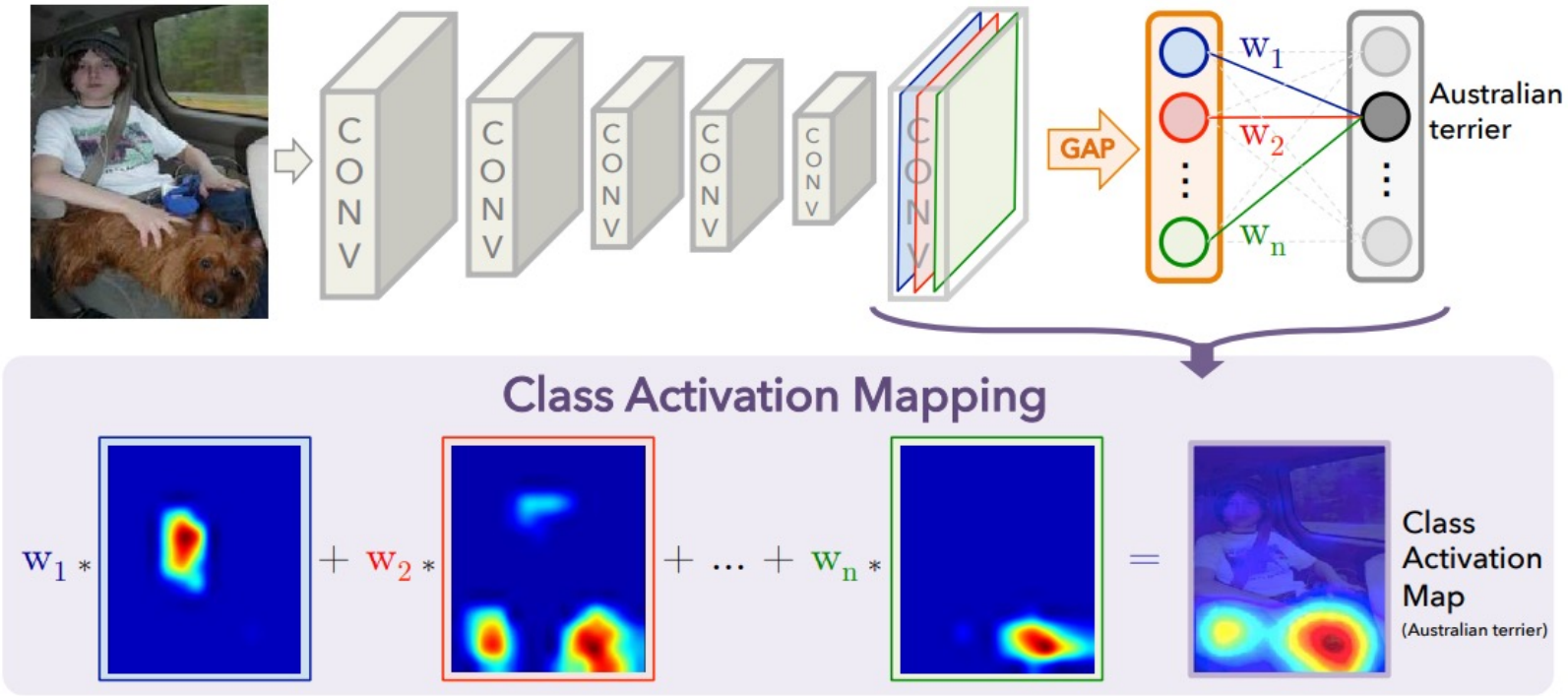
Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu



2. Related works

Class Activation Map (CAM)

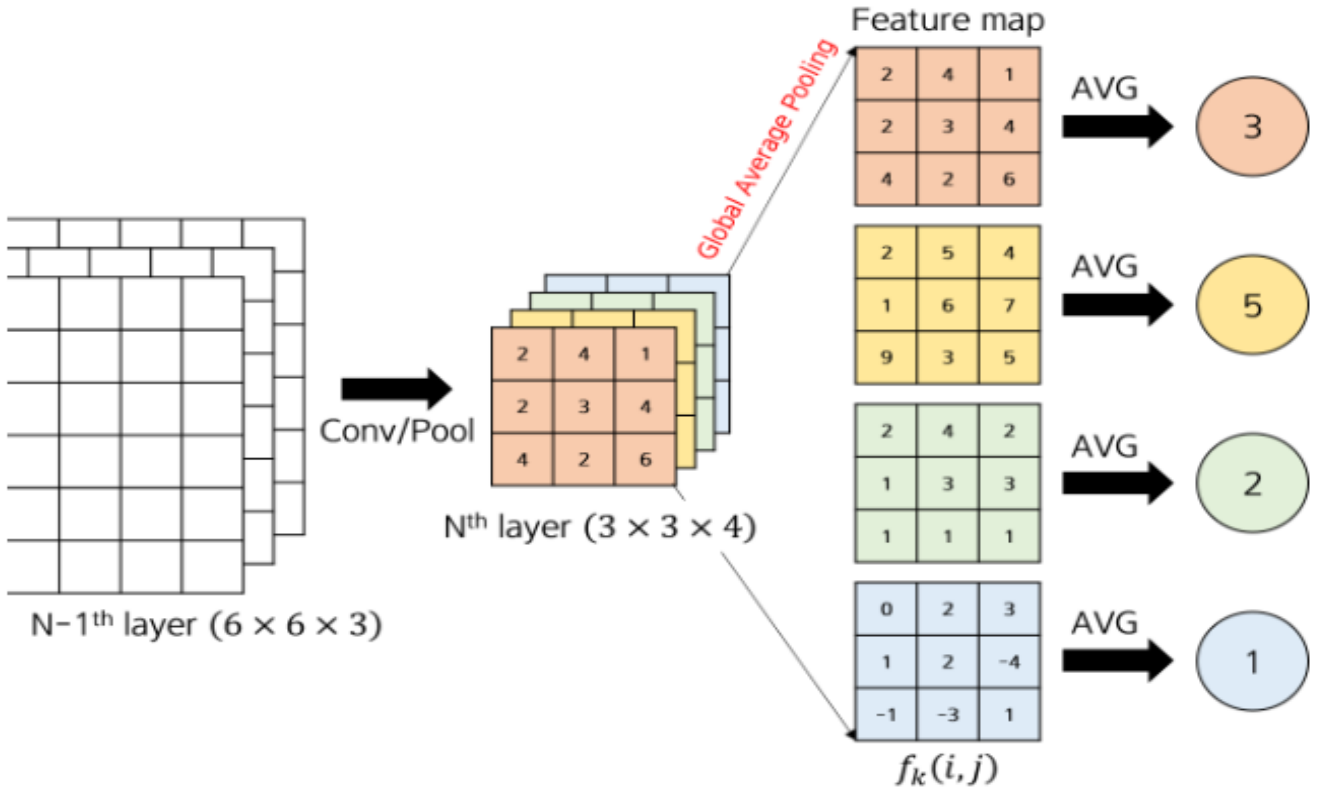
- Convolution layer와 pooling layer를 활용하여 이미지 내 정보 요약
- 마지막 convolutional layer 뒤에 Global Average Pooling 구조 사용



2. Related works

Class Activation Map (CAM)

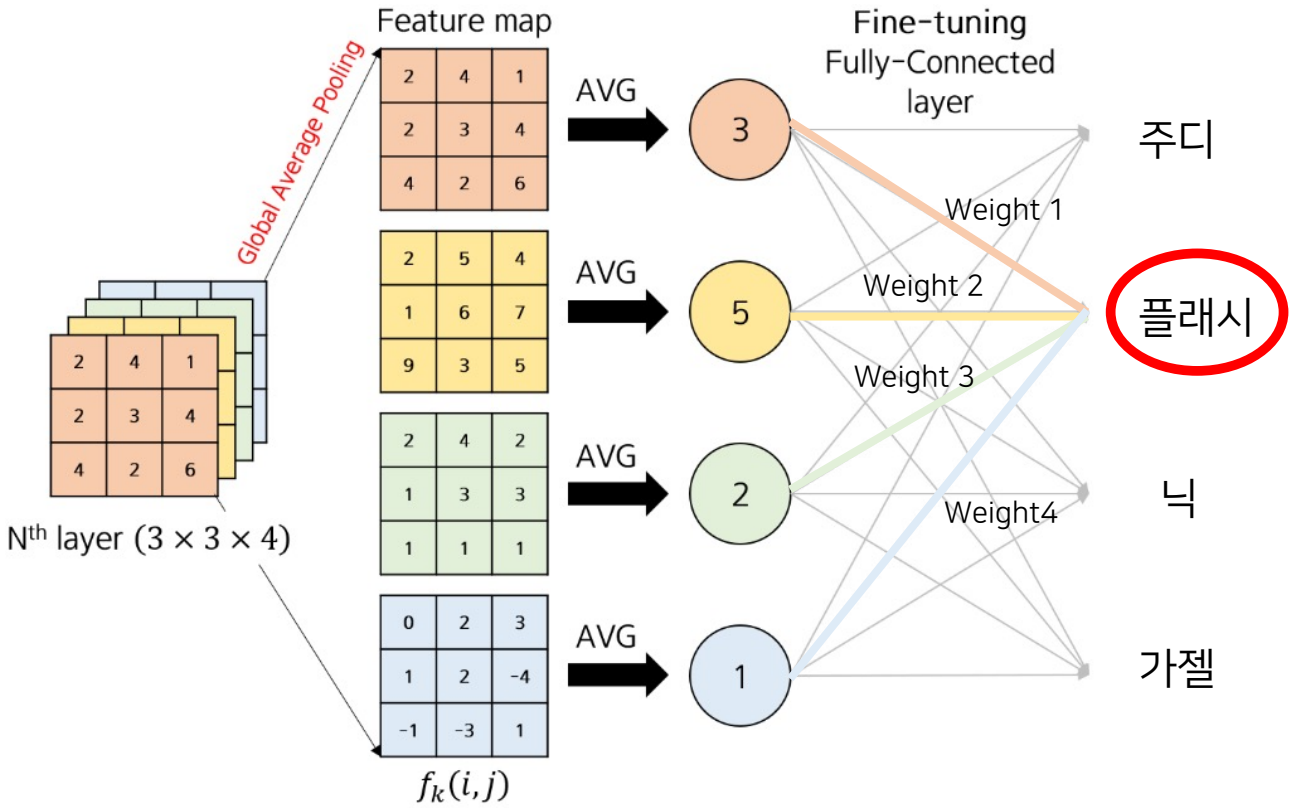
- GAP : 각 Feature map(채널)의 가중치 값들의 평균



2. Related works

Class Activation Map (CAM)

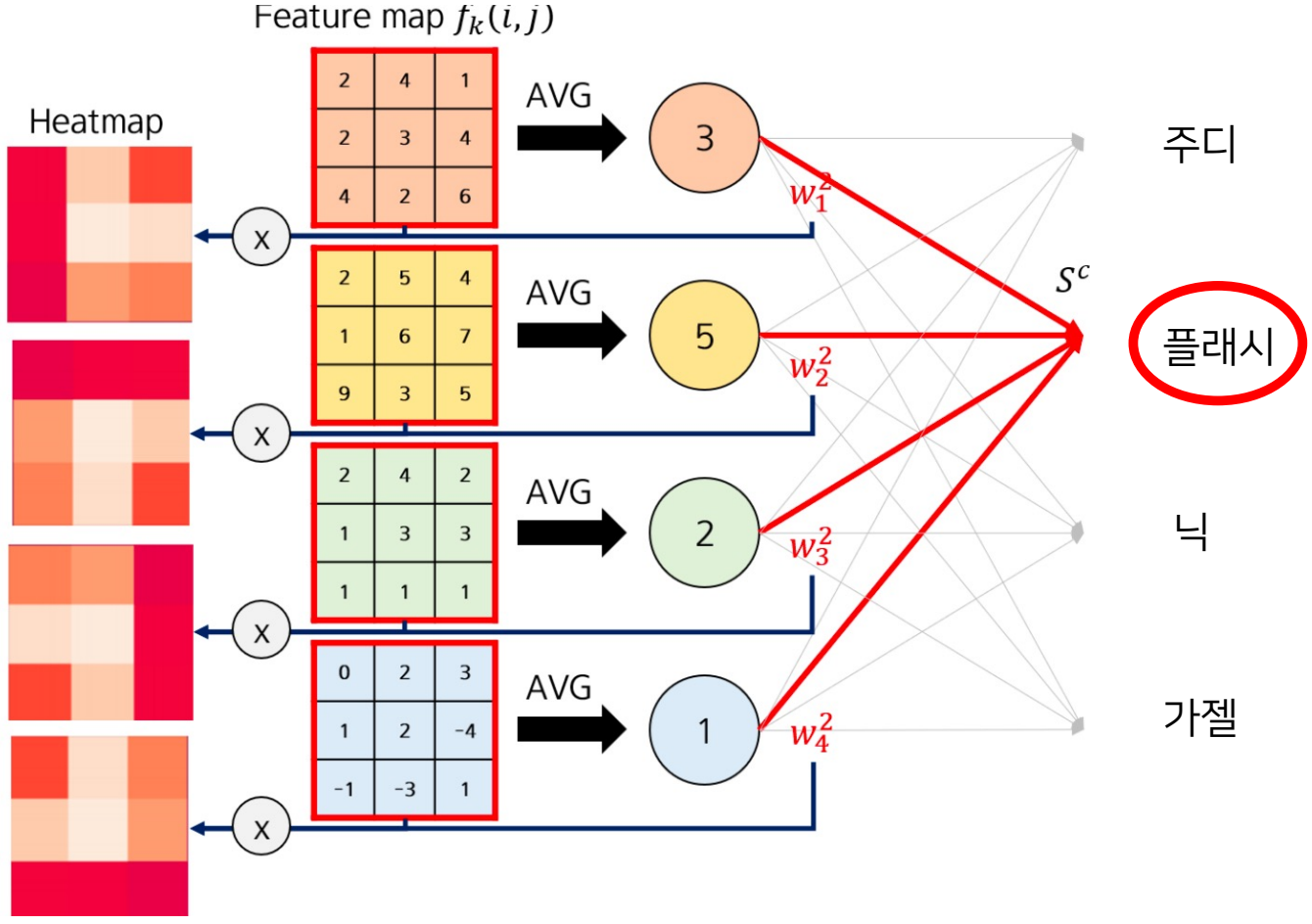
- 분류 결과에 따라 CAM에 활용되는 Weight가 달라짐



2. Related works

Class Activation Map (CAM)

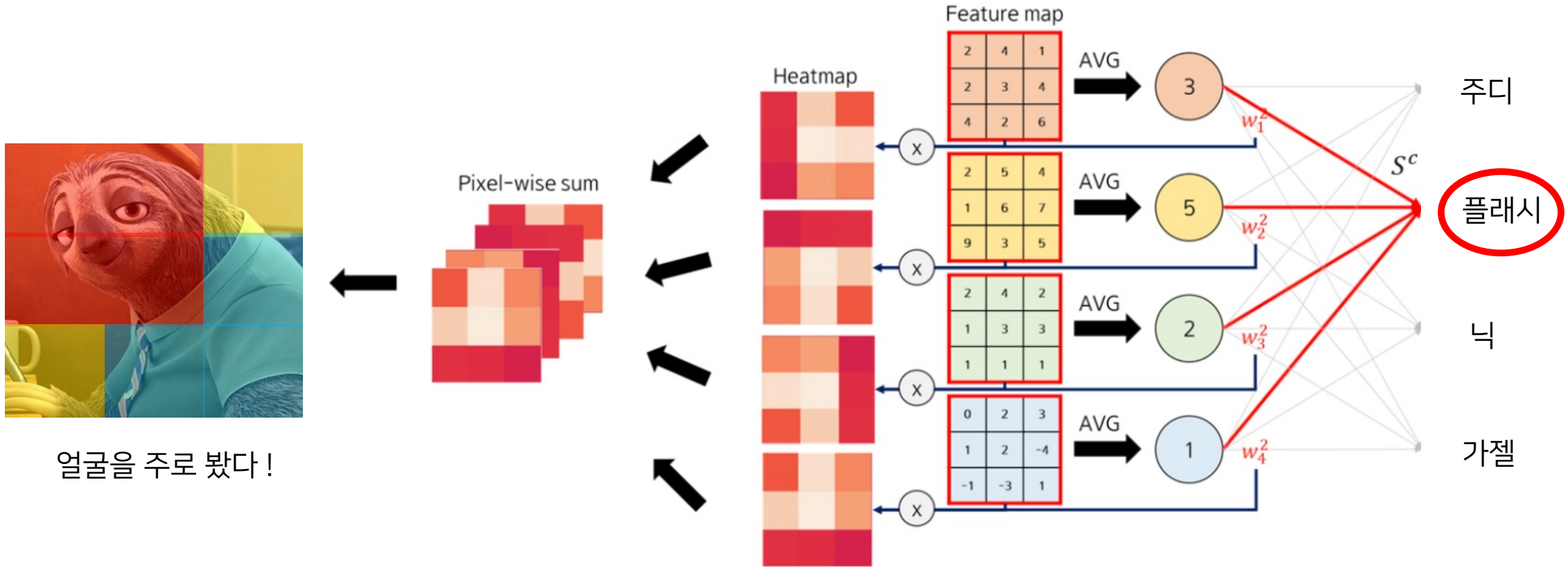
- 학습된 Weight를 각 Feature map과 곱하여 채널별 Heatmap 생성



2. Related works

Class Activation Map (CAM)

- 생성된 Heatmap을 Pixel-wise sum하여 하나의 CAM Heatmap 얻을 수 있음



2. Related works

Class Activation Map (CAM)의 한계

- Global Average Pooling layer를 반드시 사용해야 하고 뒤에 FC layer가 붙어야함
- Weight를 학습시켜야함
- CNN의 마지막 layer를 통과해 나온 Feature map에 대해서만 CAM 추출 가능
- 특정 구조의 CNN에 대해서만 적용 가능하여 낮은 클래스 분류와 적용에 한계가 있음

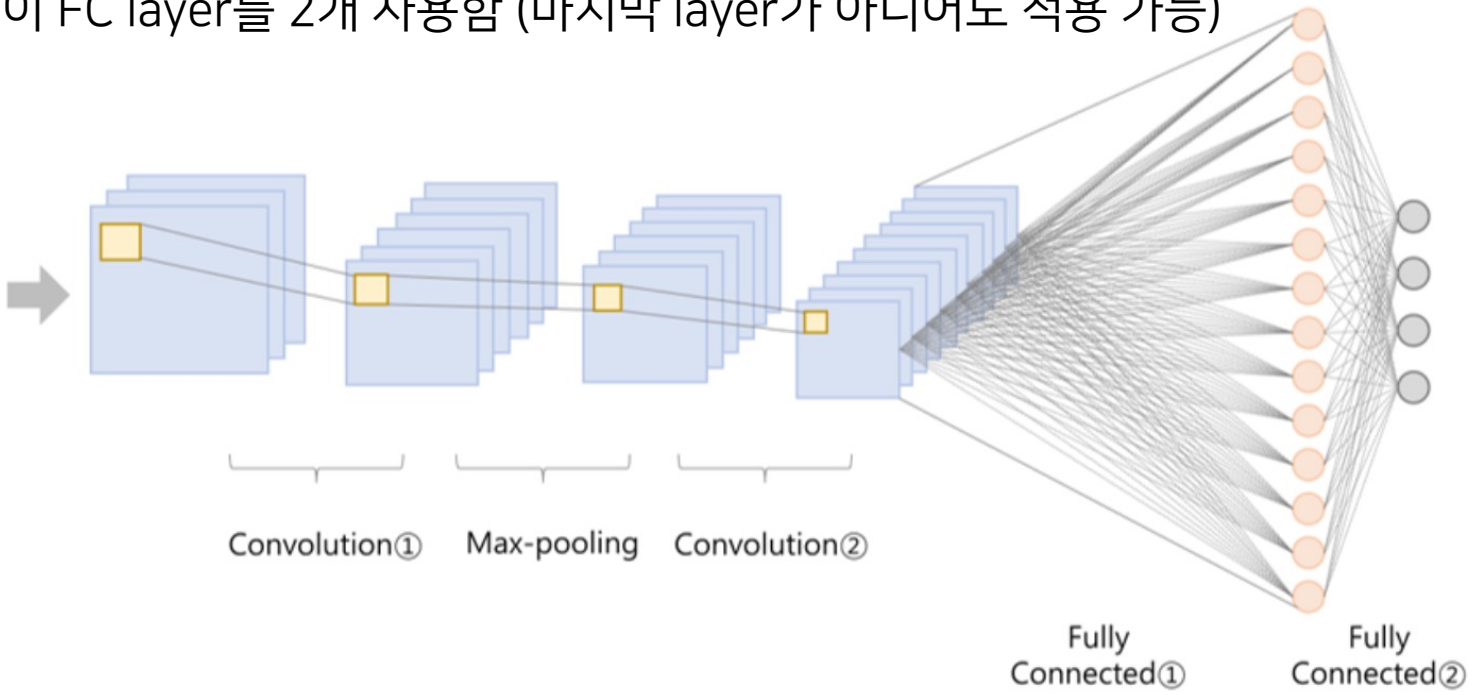
  (순서 : conv feature maps → global average pooling → softmax layer)

3. Method

Grad-CAM (2017)

- CAM 방법론 이후 1년 뒤 등장
- CNN 기본 구조를 변형하지 않고 그대로 사용함
- GAP layer 없이 FC layer를 2개 사용함 (마지막 layer가 아니어도 적용 가능)

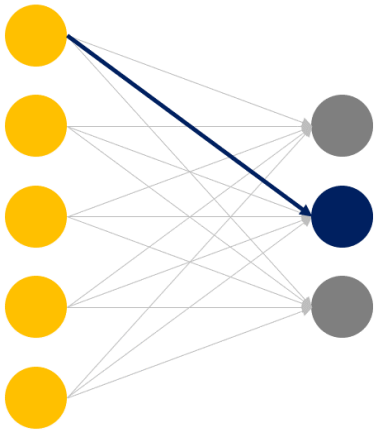
Input Image



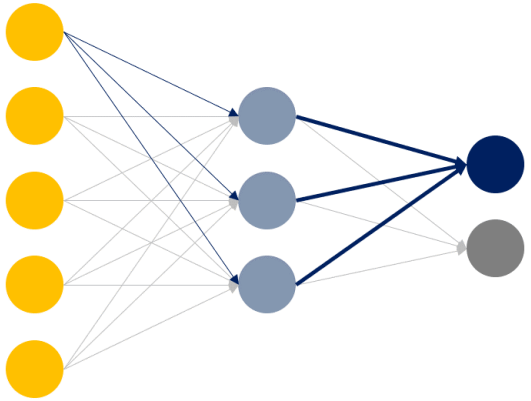
3. Method

Grad-CAM (2017)

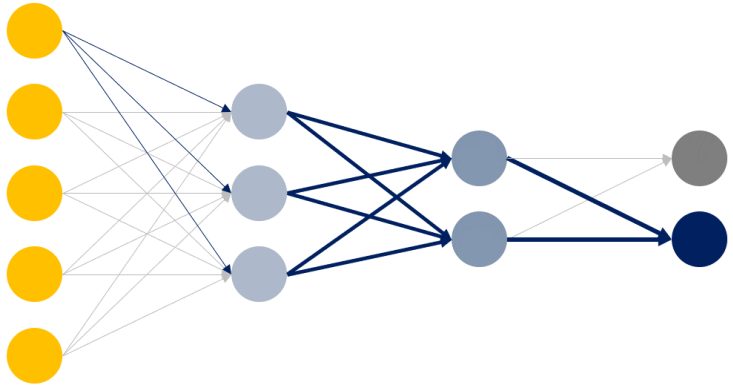
- 각 Feature map에 곱해줄 Weight를 학습이 아닌 미분(Gradient)를 통해 구함
- Gradient는 Feature map의 각 원소가 특정 class에 주는 영향력



Hidden layer (X)



Hidden layer (O)

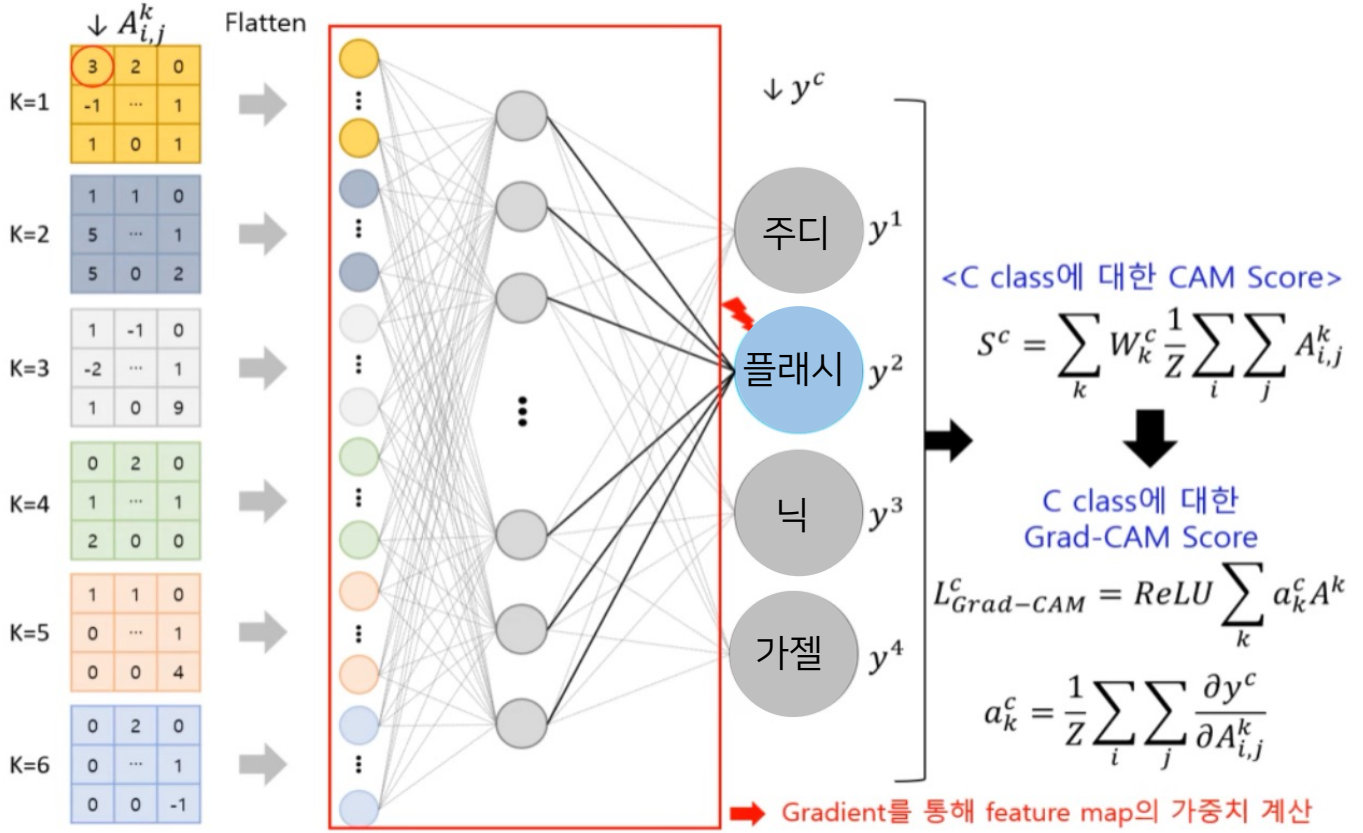


Gradient
[Total amount of effect of input K on output class C]

3. Method

Grad-CAM (2017)

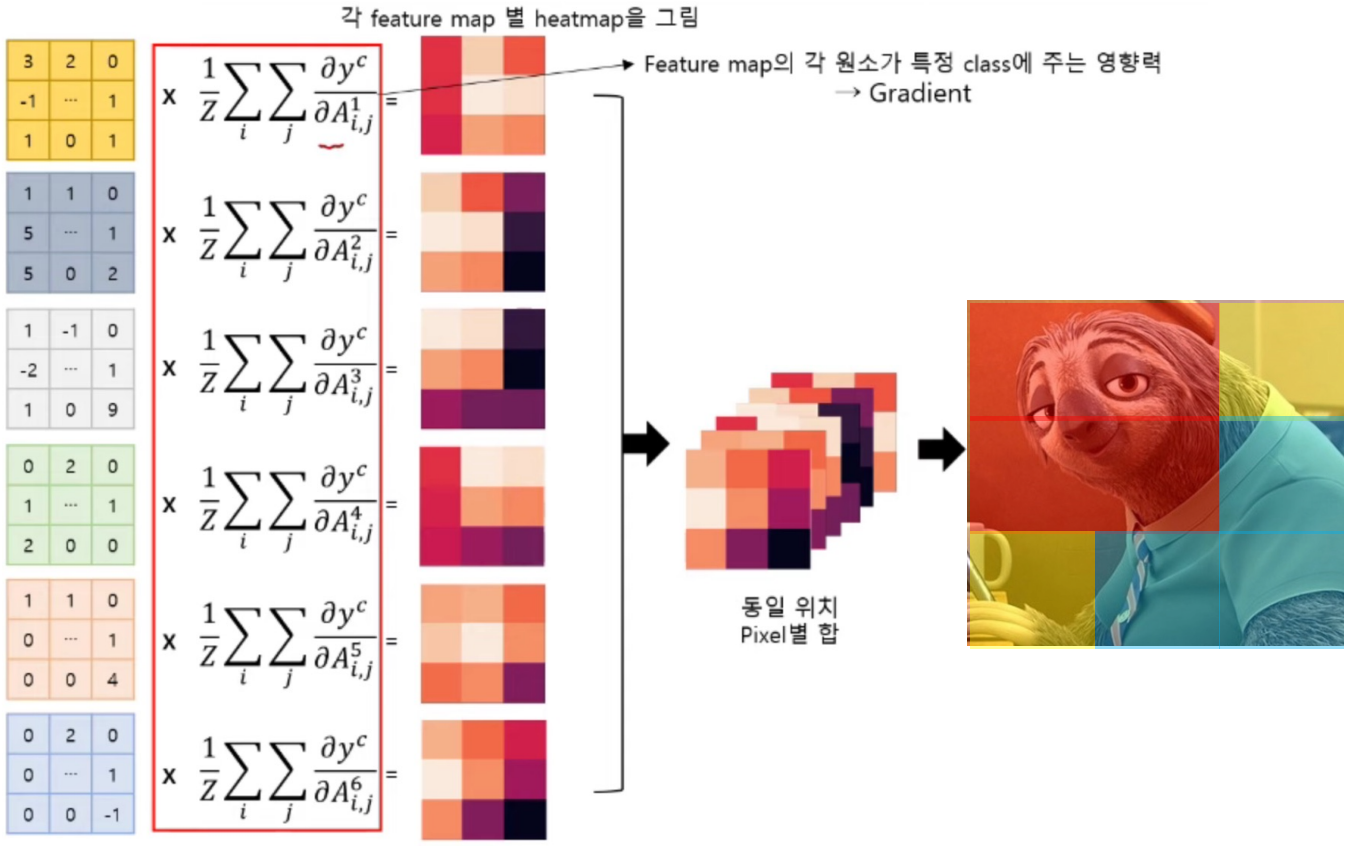
- class에 대해 각 feature map의 원소로 gradient값을 구하고 평균을 취함
- 이 값이 weight로 사용됨



3. Method

Grad-CAM (2017)

- Gradient를 통해 구한 weight를 각 feature map과 곱함
- Feature map을 pixel wise sum하여 grad-cam heatmap을 얻을 수 있음



4. Experiment

- Weakly-Supervised Localization

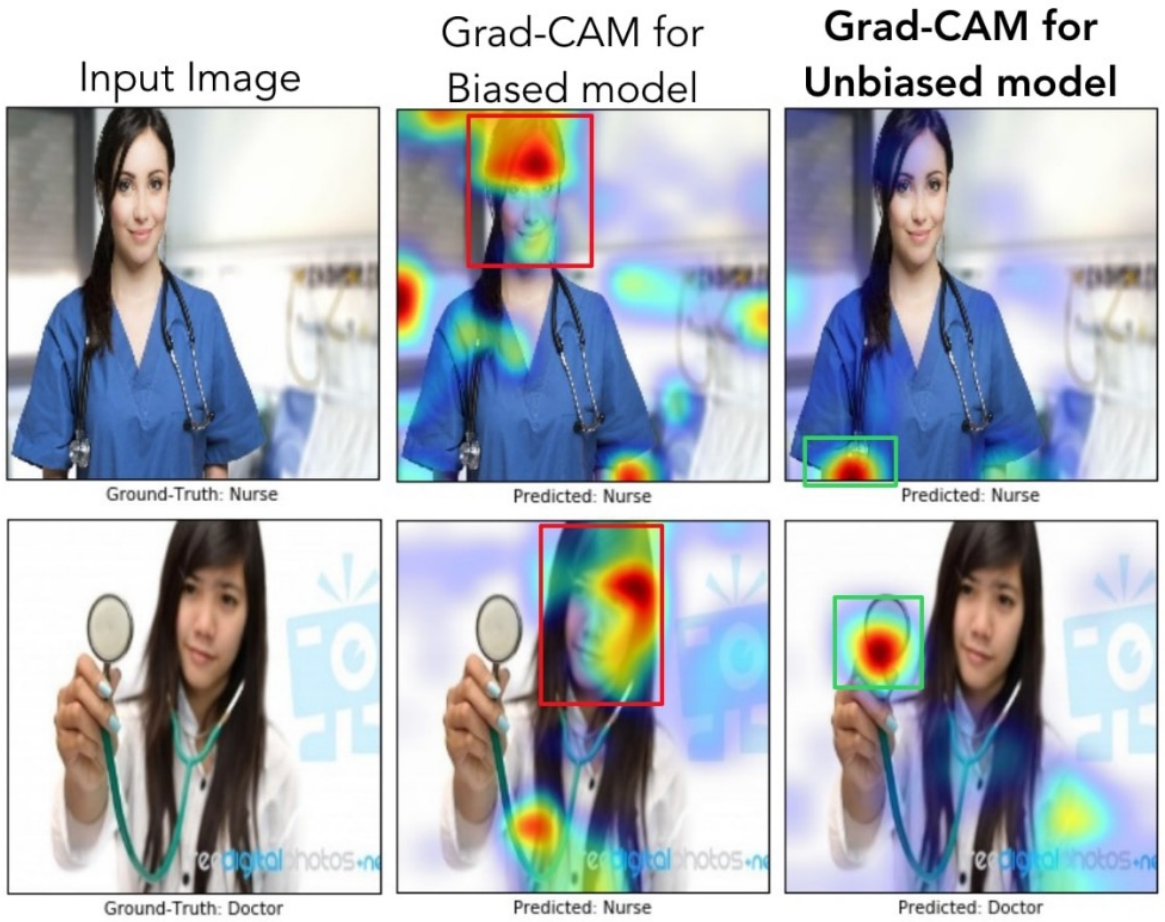
Table 1 Classification and localization error % on ILSVRC-15 val (lower is better) for VGG-16, AlexNet and GoogleNet

	Classification		Localization	
	Top-1	Top-5	Top-1	Top-5
VGG-16				
Backprop (Simonyan et al. 2013)	30.38	10.89	61.12	51.46
c-MWP (Zhang et al. 2016)	30.38	10.89	70.92	63.04
Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
CAM (Zhou et al. 2016)	33.40	12.20	57.20	**45.14**
AlexNet				
c-MWP (Zhang et al. 2016)	44.2	20.8	92.6	89.2
Grad-CAM (ours)	44.2	20.8	68.3	56.6
GoogleNet				
Grad-CAM (ours)	31.9	11.3	60.09	49.34
CAM (Zhou et al. 2016)	31.9	11.3	60.09	49.34

We see that Grad-CAM achieves superior localization errors without compromising on classification performance
Bold values indicate lowest localization errors in that column

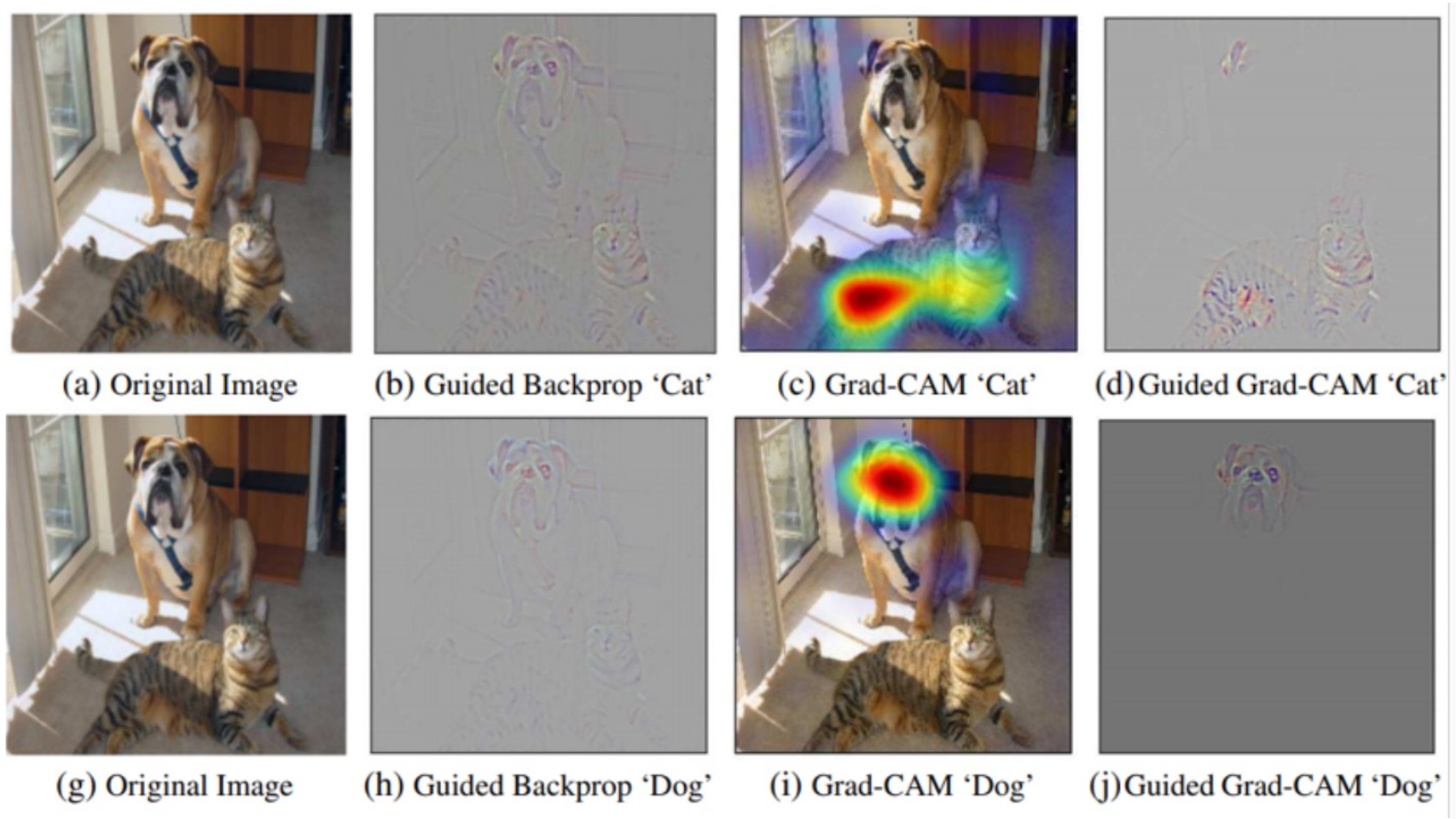
4. Experiment

- 데이터셋에 존재하는 bias 낮출 수 있음



4. Experiment

- Guided Backpropagation + Grad-CAM



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog'

감사합니다