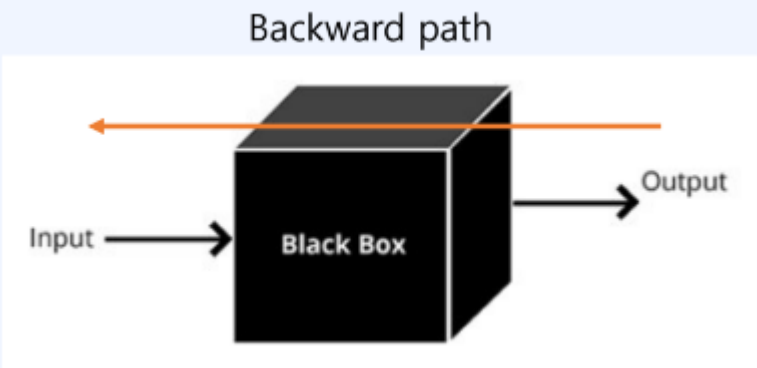
XAI (eXplainable AI)

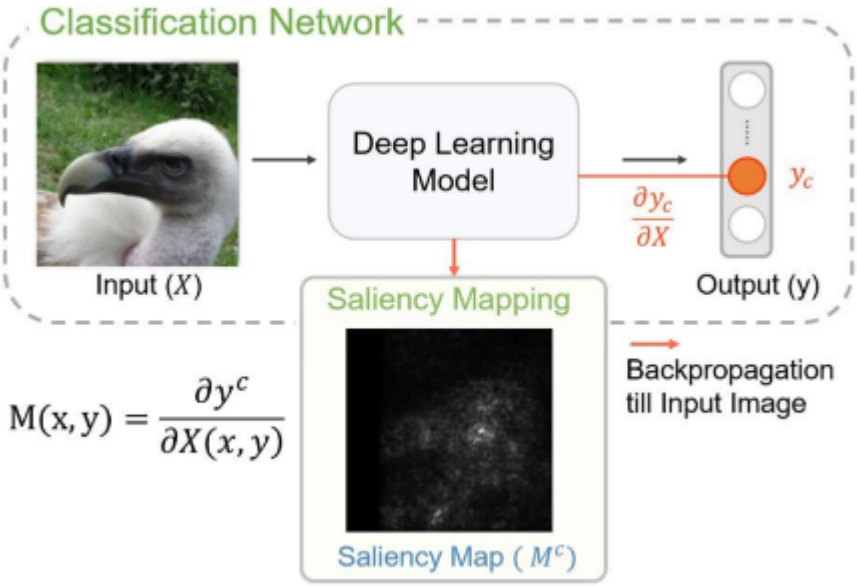
모델 판단 근거 설명

▼ Gradient Map(Saliency Map)

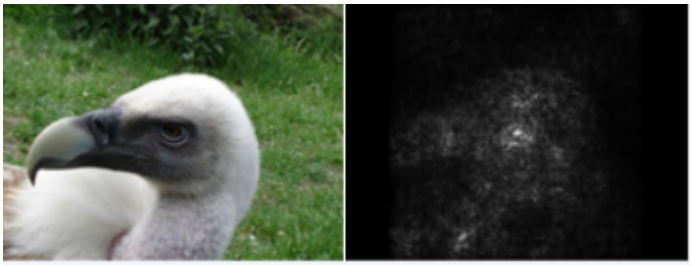


논문명 : Deep inside convolutional networks: Visualising image classification models and saliency maps

idea : 중요한 픽셀의 변화는 결과값의 변화에 큰 영향을 미친다

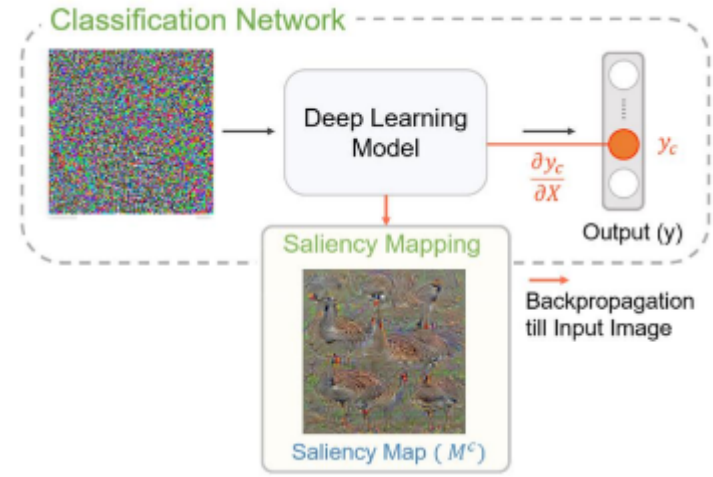


결과값에 대한 각 픽셀의 미분값



▼ Class Model Visualization

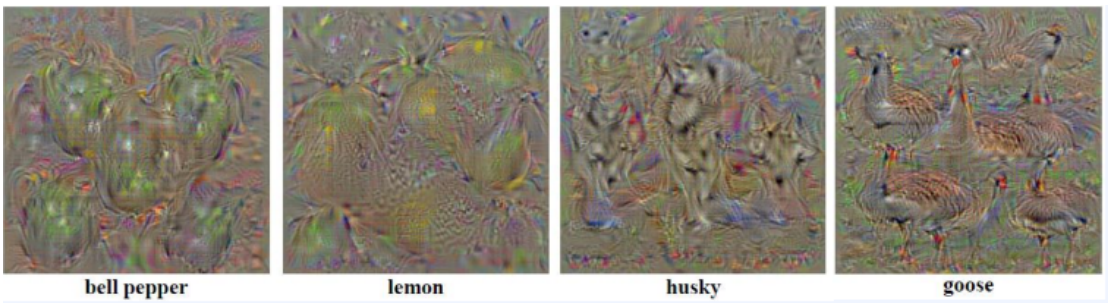
논문명 : Deep inside convolutional networks: Visualising image classification models and saliency maps



학습한 모델에 노이즈를 넣어 클래스로 판단한 원인 추측

$$\text{Loss} = \text{argmax}\, y^c(I)$$

$$I^{n+1} = I^n + \lambda \frac{\partial y^c(I)}{\partial I(x,y)}$$

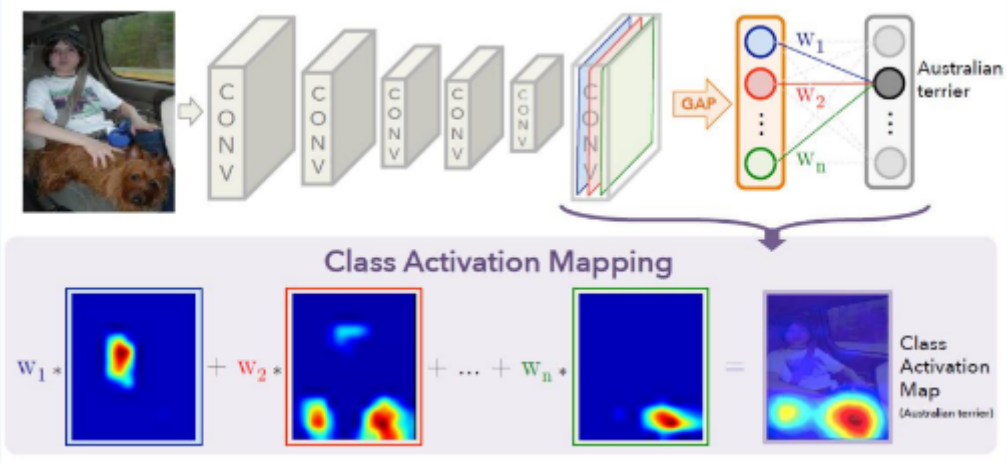


▼ CAM

▼

▼ CAM(Class Activation Mapping)

논문명 : Learning deep features for Discriminative Localization



전재조건

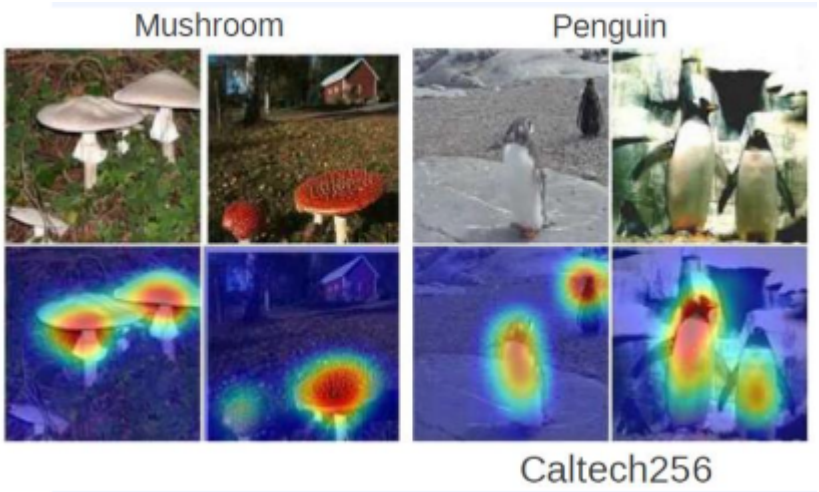
1. 마지막 conv 이후 무조건 global aveage pooling이 존재해야 함

   → 마지막 conv의 사이즈가 작아 upscaling해야함

   → 해상도 저하

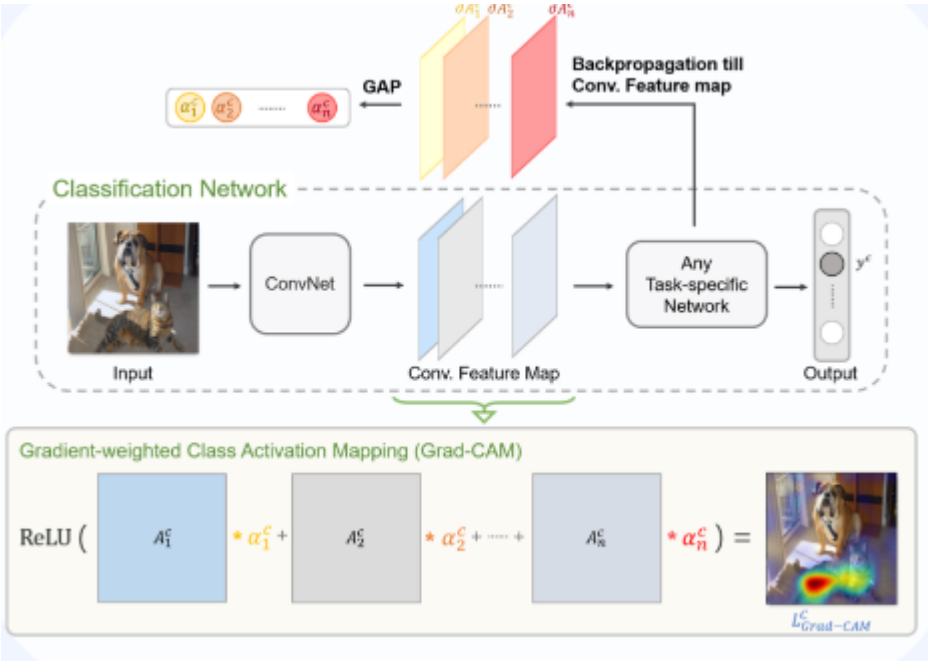
2. 이후 FC layer 가 '1개'있어야 함

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



▼ Grad-CAM

논문명 : Gradient-based Class Activation Mapping

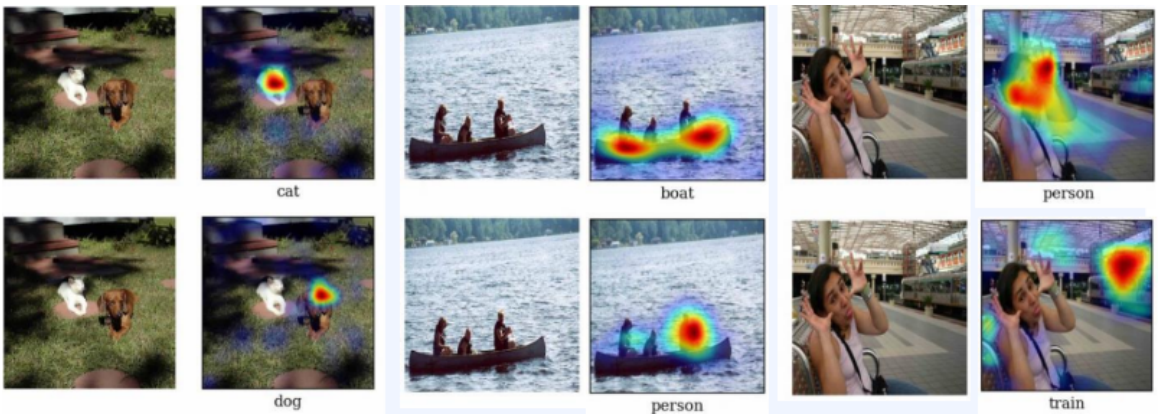


- conv이후 어떠한 layer가 들어와도 상관 없음
- taget class Yc에 대하여 특정 Feature map을 미분함

  → 미분한 Feature map을 GAP진행 = gradient처럼 작용

$$L^{C}_{Grad-CAM} = \mathrm{Re}\,LU\left(\sum_{k}\alpha^{C}_{k}A^{k}\right)$$

Relu : 음수제거를 통해 사용된 값만 추출



1. 설명력 좋음

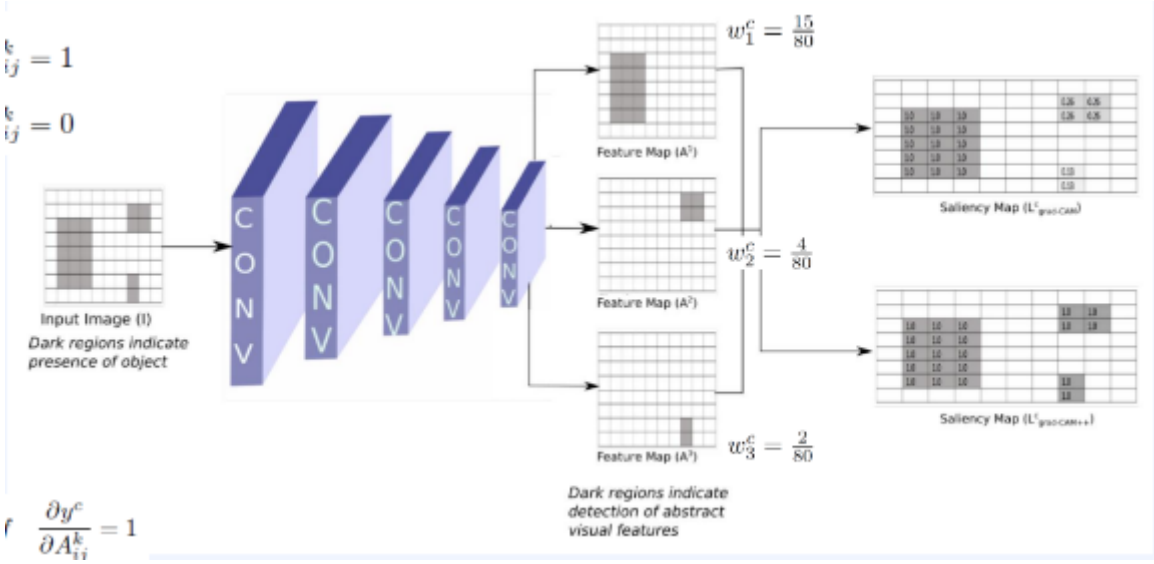
2. TP, TN, FP, FN 분석 시 근거 확인 가능
3. 판단 근거 분석 가능
4. 모델의 깊이에 따른 피쳐 추출 확인 가능

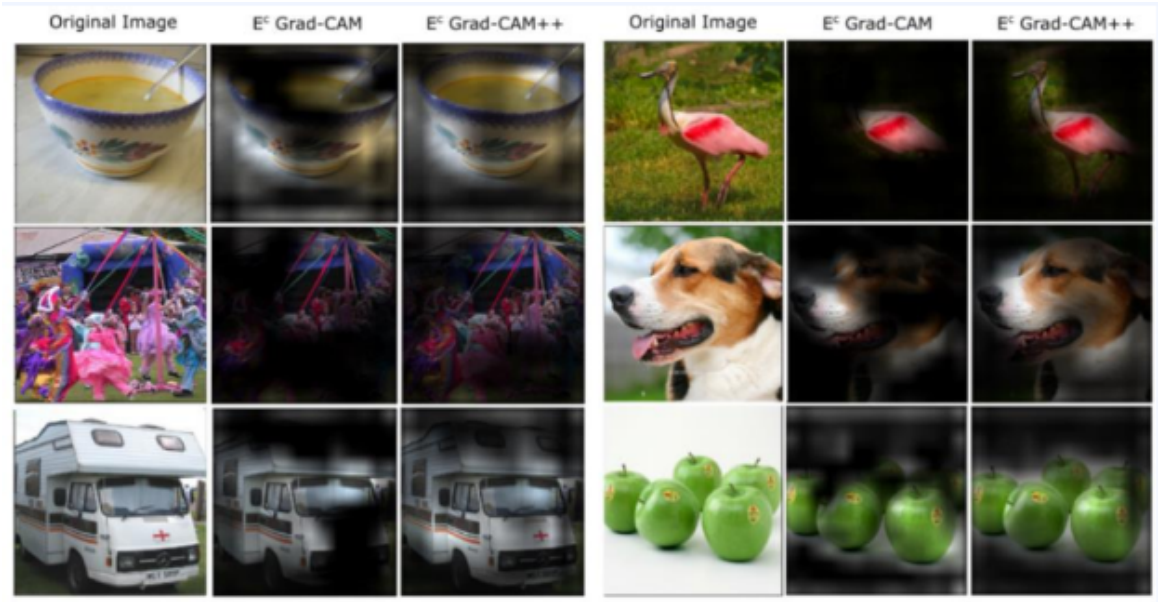
▼ Grad-CAM ++

논문명 : Gradient-based Class Activation Mapping++

- Generalized Grad-CAM
- Spatially gradient weighting
- 더 넓은 영역을 localization함.

한 이미지에 같은 개체가 다수 있어도 표시 가능





Method	**Grad-CAM++**	Grad-CAM
- Average drop% *(Lower is better)*	**43.31**	46.43
- % incr. in confidence *(Higher is better)*	**15.44**	13.67
- Win% *(Higher is better)*	**60.24**	39.76

Table 2. Results for objective evaluation of the explanations generated by both Grad-CAM++ and Grad-CAM on the ImageNet (ILSVRC2012) validation set ("incr" denotes increase). We took random subsets of 2510 images to maintain consistency with the PASCAL VOC 2007 dataset and averaged out the results. The results futher substantiate our claim that Grad-CAM++ improves upon the performance of Grad-CAM.

- Average Drop : 원본 출력 값과 grad-cam++로 마스킹한 이미지의 출력값 비교하여 표현하지 못하는 비율
- Average Increase : 원본 출력 값과 grad-cam++로 마스킹한 이미지의 출력값 비교하여 얼마나 크게 있는지

mIoU test

: 실제 object가 있는 부분과 grad-cam++로 마스킹한 이미지의 교집합을 통한 계산

$$Loc_I^c(\delta) = \frac{Area(internal\ pixels)}{Area(bounding\ box) + Area(external\ pixels)} \tag{26}$$

- Grad-CAM++은 Grad-CAM의 더 일반화된 방법론으로,
- object의 더 넓은 영역, 더 많은 object를 localization함
- 하지만 실제로 해보면, 다른 class의 object까지 localization하는 경향이 있어, 실질적으로 Grad-CAM보다 더 향상되었다고 보기 힘듦.

cam 문제점
Shattered Gradient Problem
1. 모델이 깊어질 수록, gradient는 Noisy해진다.
2. 인접한 두 픽셀간의 gradient가 연속적이지 않음.

▼ LRP

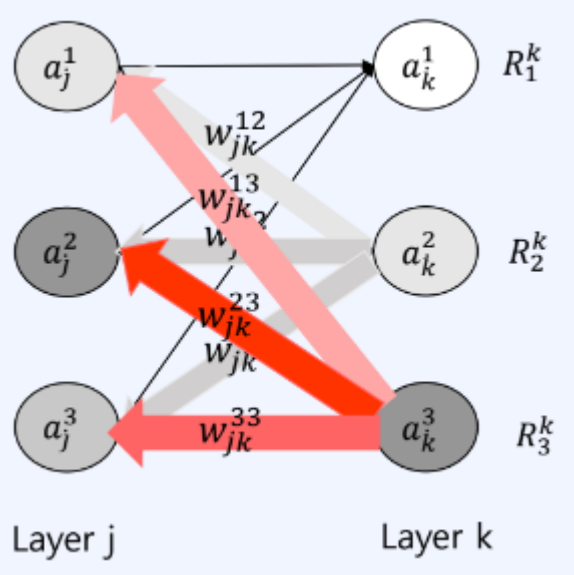
논문명 : Layer-wise Relevance Propagation

idea : output에서 input까지 역으로 계산하자

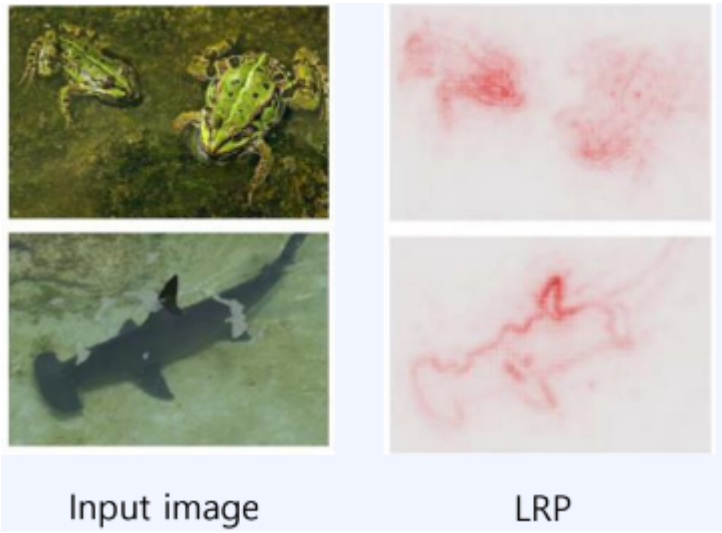
- Conservative Rule

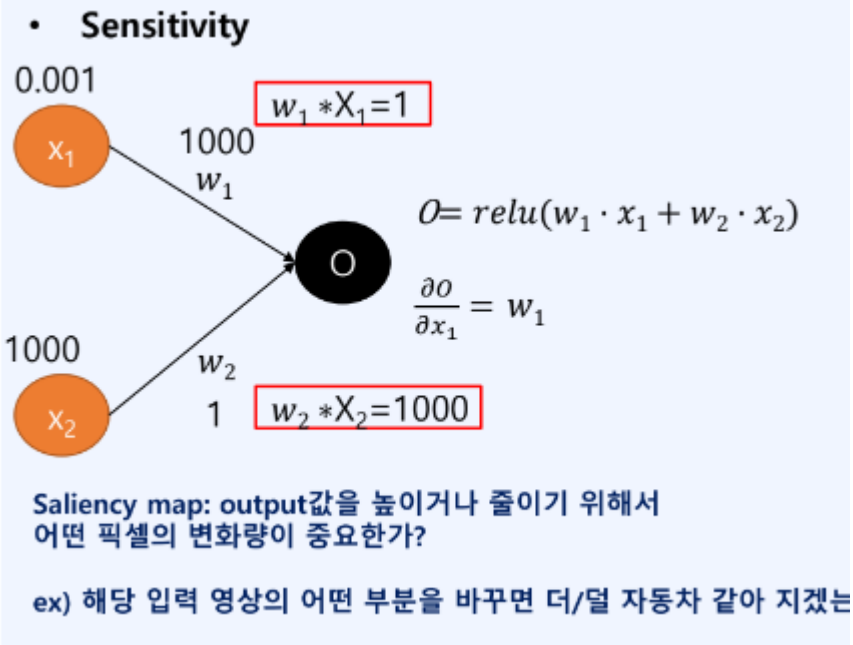
$$\forall x : f(x) = \sum_p R_p(x)$$

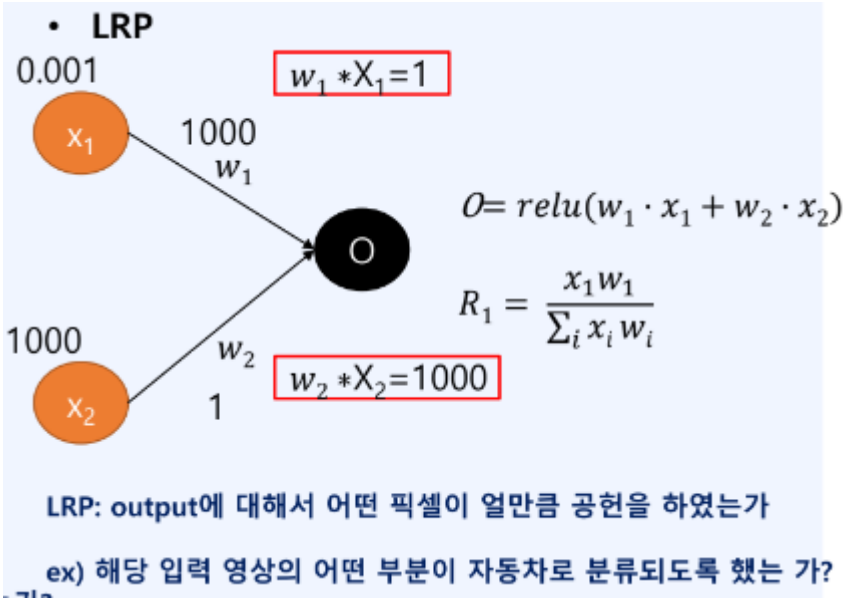
  - 역으로 계산시 분배 후의 값의 합이 분배 전과 같아야 한다.
- Propagation Rule



$$R_j = \sum_k \frac{a_j W_{jk}}{\sum_j a_j w_{jk}} R_k$$



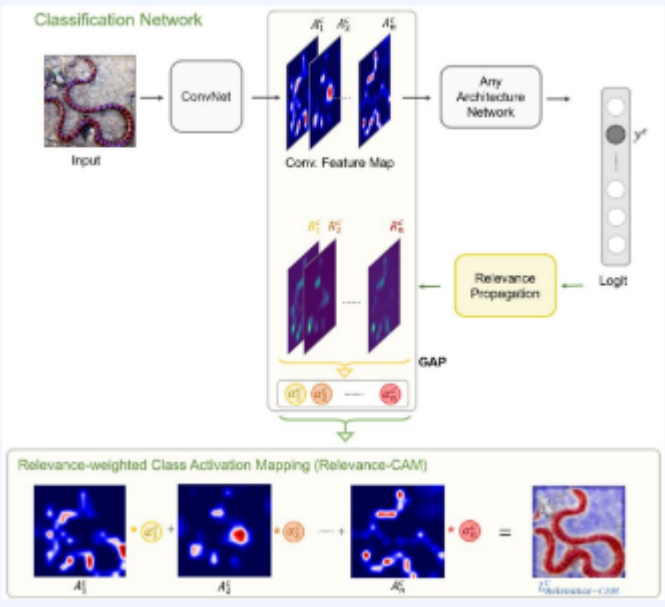




Relevance-CAM(R-CAM)

Relevance(LRP) + CAM

- Gradient shattered problem에 강인함
- Shallow layer에 대한 정확한 분석이 가능

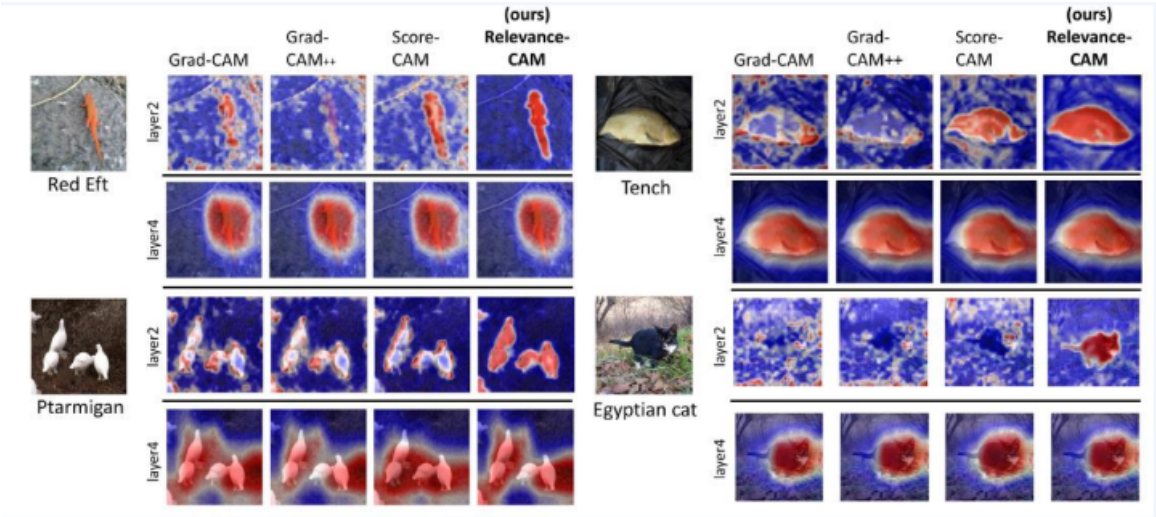


실제 CLRP사용함

CLRP : target class는 1 나머지는 -1 / (class.length -1)

→ 총 R값이 0이 됨, target 이외의 multi class feature를 억제

grad-cam과 같이 특정 layer의 feature map 계산 및 GAP



	장점	단점
Gradient Map	1. 한번에 구할 수 있음 2. 결과가 원본이미지와 같음	1. noize 2. bad localization
CAM	1. 설명력 좋음 2. good localization	1. global

		average pooling 필수 2. FC layer 1개 제한
Grad-CAM	1. high localization 2. network 구조에 따른 제한 없음	1. 대상의 일부만 표현
Grad-CAM++	1. Grad-CAM보다 넓은 영역, 더 많은 object 출력	1. 다른 class의 object까지 localization하는 경향
LRP	1. high localization	1. input layer까지 계산하므로 속도 저하
Relevance-CAM	1. 어떤 깊이에 있는 layer라도 분석이 가능 2. 고해상도이며 좋은 localization 성능을 보여줌. 3. R-CAM을 통해서 shallow layer도 class specific features를 추출할 수 있음을 분석함.	