DEEP NEURAL NETWORK ANALYSIS 2024

CLASS ACTIVATION MAPPING - CAM, Grad-CAM and Grad-CAM++

1 XAI POST-HOC

2 CAM

3 GRAD-CAM

4 GRAD-CAM++

5 DEMO

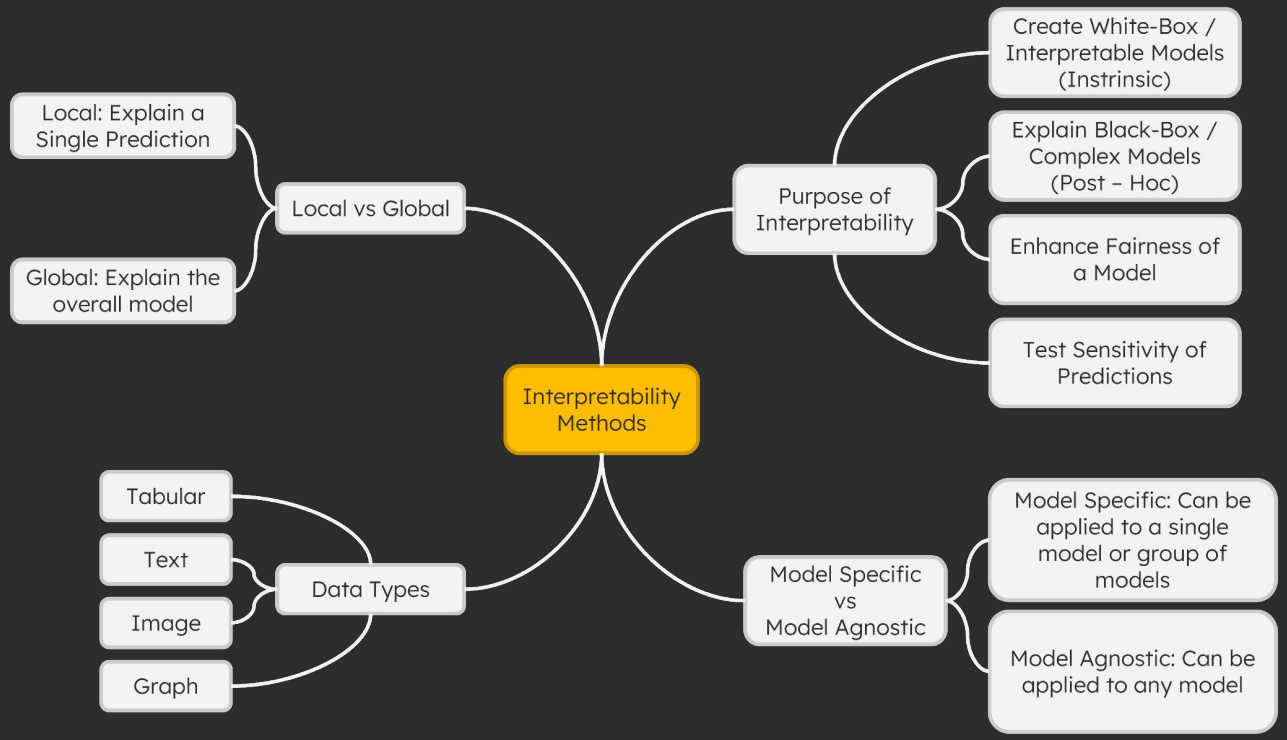
INTERPRETING DL MODELS POST-HOC

CLASS ACTIVATION MAPPING

INTERPRETING DL MODELS POST-HOC

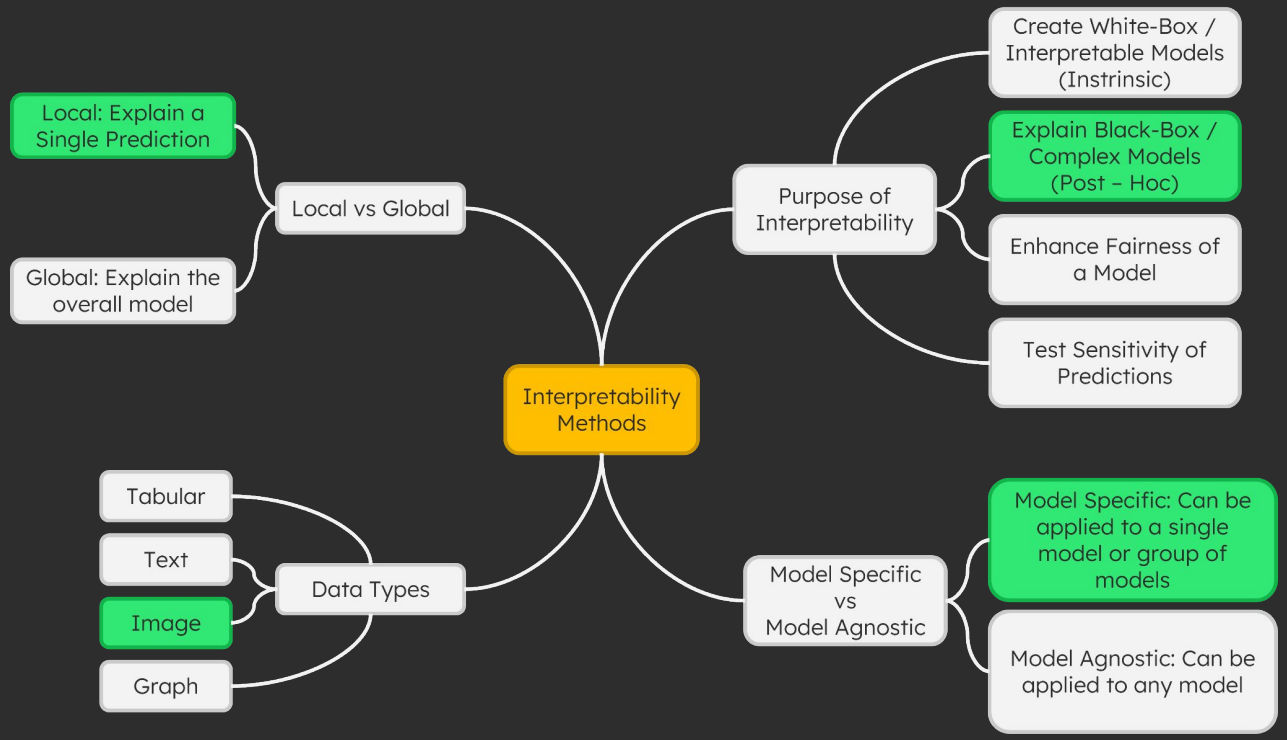
- Most DL Models are **black boxes**
- Areas with **responsibility** (i.e. medicine or autonomous driving) **need interpretability**
- Interpretability methods can be **baked into** the model (white box) or **applied post-hoc**

INTERPRETABILITY METHODS TAXONOMY



CLASS ACTIVATION MAPPING

INTERPRETABILITY METHODS TAXONOMY



1 XAI POST-HOC
2 CAM
3 GRAD-CAM
4 GRAD-CAM++
5 DEMO

CLASS ACTIVATION MAPPING

CAM INTUITION

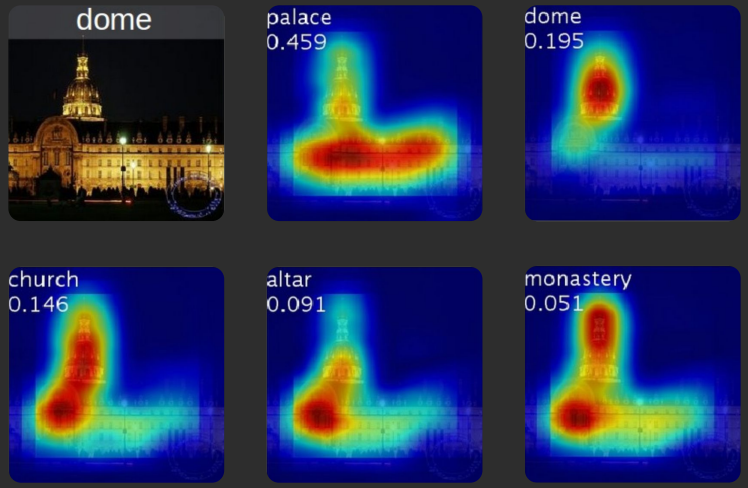
Method to determine the **importance of areas** in an image depending on a given class

Allows to draw a **human-interpretable** heatmap

Requires **specific** model **architecture** trained for image classification

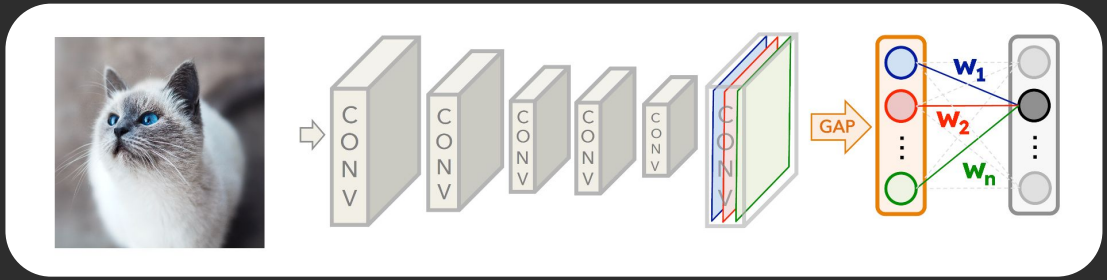
Learning Deep Features for Discriminative Localization

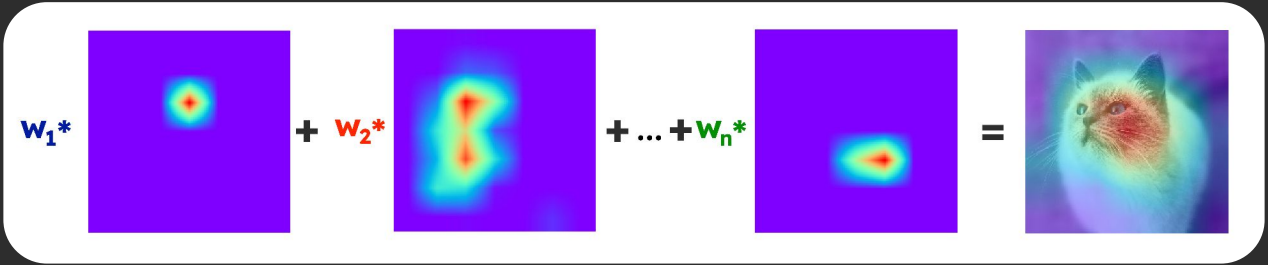
Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu



CLASS ACTIVATION MAPPING

CAM OVERVIEW





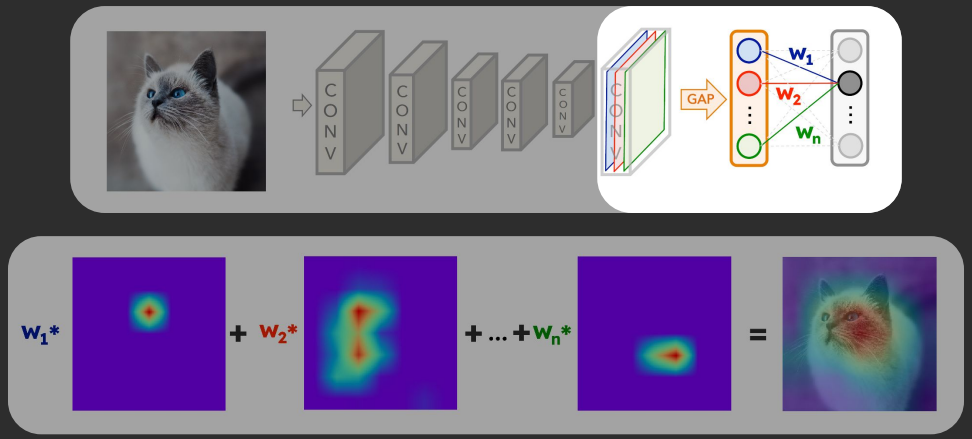
Straightforward computation with the **correct model head!**

CLASS ACTIVATION MAPPING

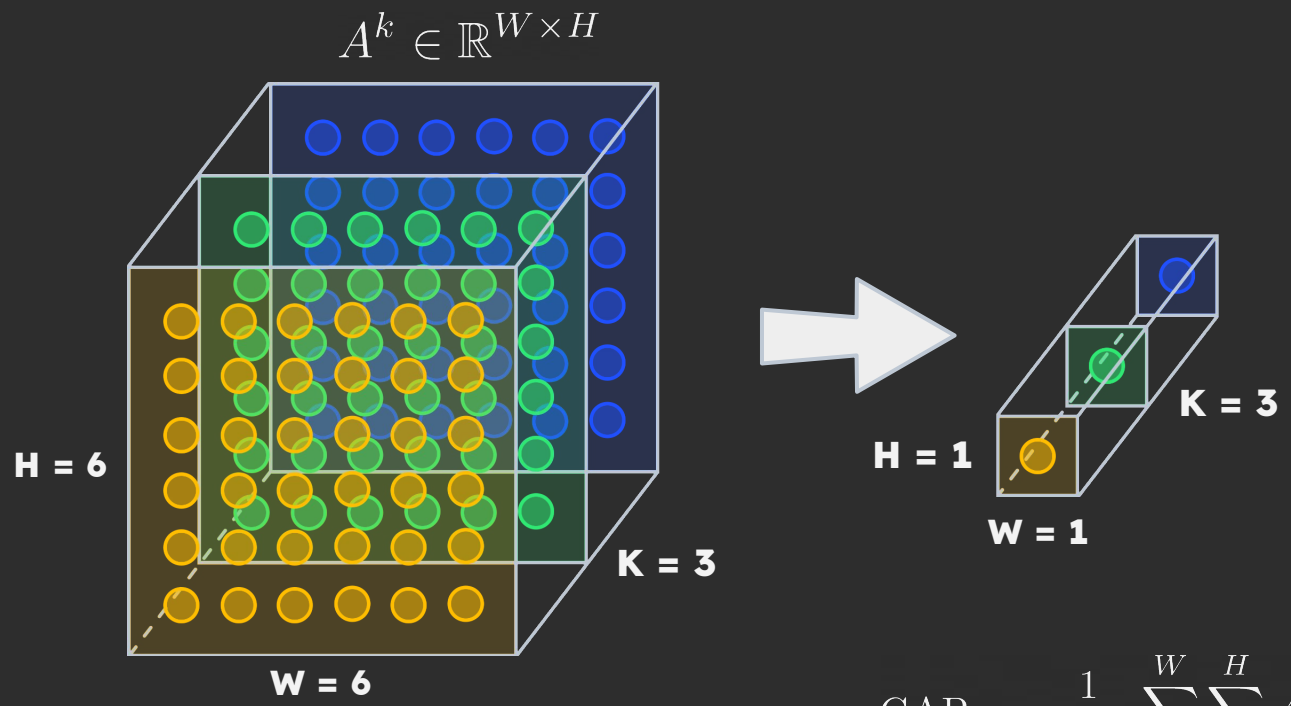
CAM PREREQUISITS

Classification head consisting of:

- **Global Average Pooling** (GAP)
- **Fully-connected** read-out **layer**

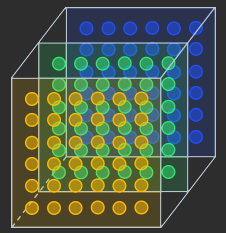


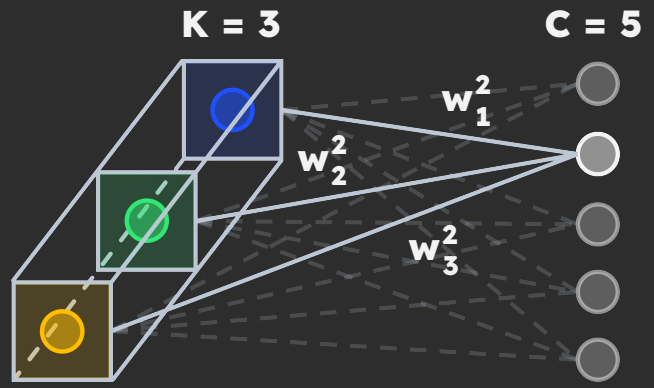
GLOBAL AVERAGE POOLING



$$\mathrm{GAP}_k = \frac{1}{WH} \sum_{w}^{W} \sum_{h}^{H} A^k_{wh}, \quad \forall k \in K$$

GLOBAL AVERAGE POOLING



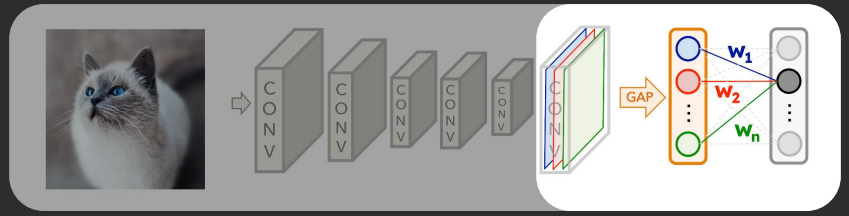


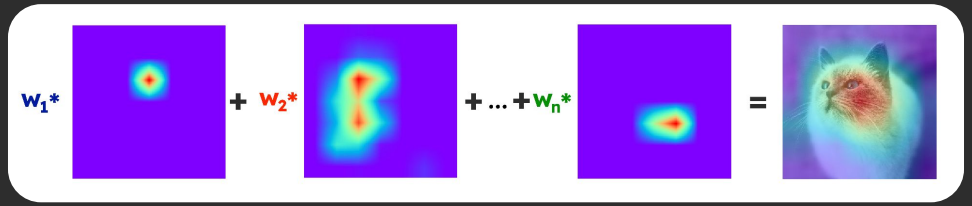
$$y^c = \sum_k^K w_k^c \mathrm{GAP}_k$$

CLASS ACTIVATION MAPPING

CAM EXPLAINED

CAM can be defined as the **weighted sum** over all **feature maps** of the **last conv layer**





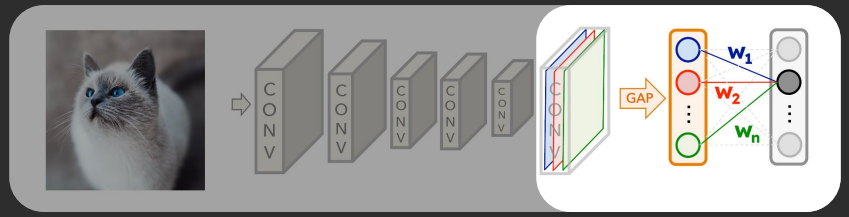
CAM EXPLAINED

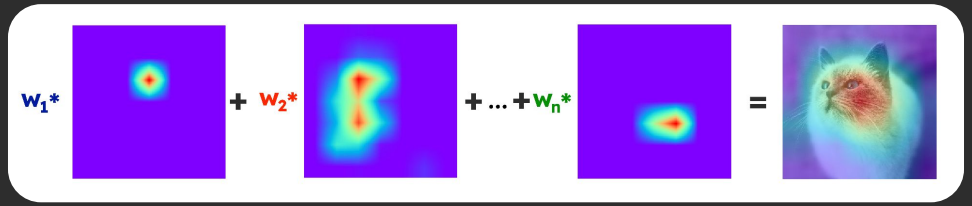
$$\mathrm{CAM}^c = \sum_k^K w_k^c A^k$$

K := total number of feature maps

w_k^c := kth weight between GAP and read-out layer for class c

A^k := kth feature map





CLASS ACTIVATION MAPPING

CAM DISADVANTAGES

Architecture requirements make it **less versatile**

Heatmap is rather **coarse**

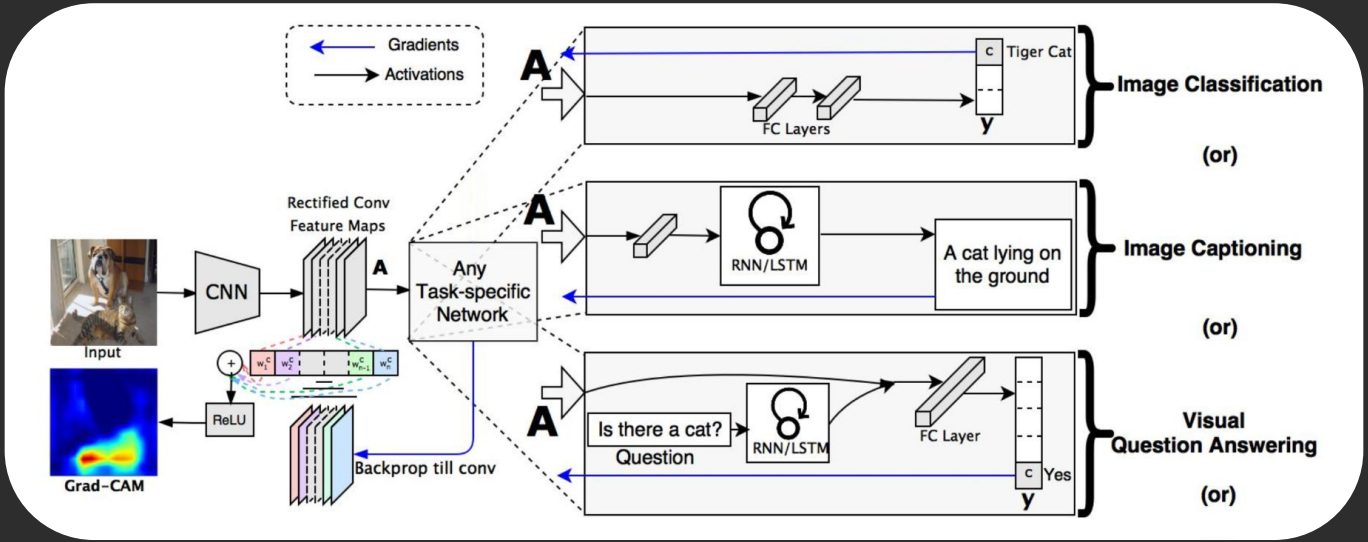
Lacks the ability for **detecting multiple instances** of the same class

Often does **not capture** the **entire object**

1 XAI POST-HOC
2 CAM
3 GRAD-CAM
4 GRAD-CAM++
5 DEMO

GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING

GRAD-CAM OVERVIEW



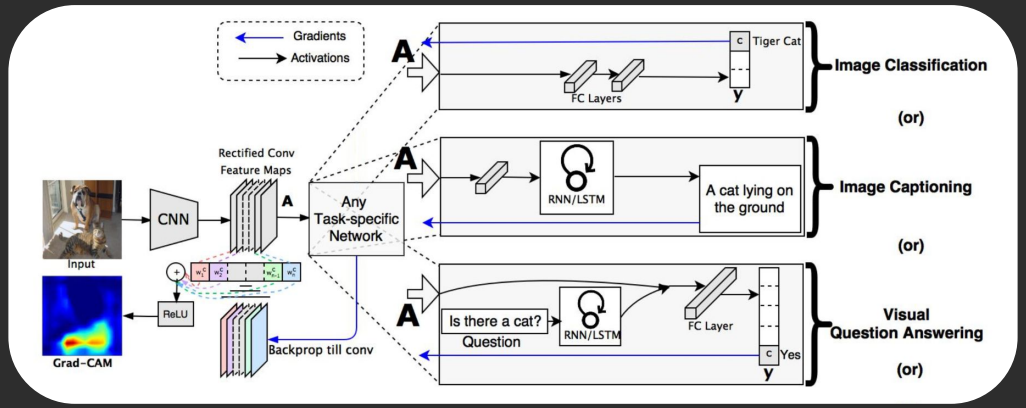
Alleviates the **need** for a **specific** CNN **architecture**!

GRAD-CAM DIFFERENCES

Changes **feature map weighting** by utilizing **gradients** instead of model weights

→ Supports **any Model Head**

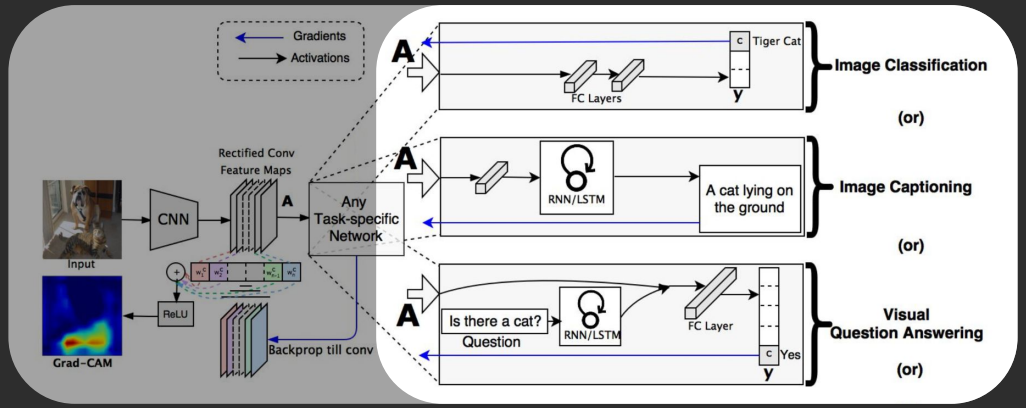
Applies **ReLU**, as **pixels** that need to be **suppressed** acc. to the gradient are **not relevant**



GRAD-CAM DIFFERENCES

Output does **not** need to be the **class score** anymore

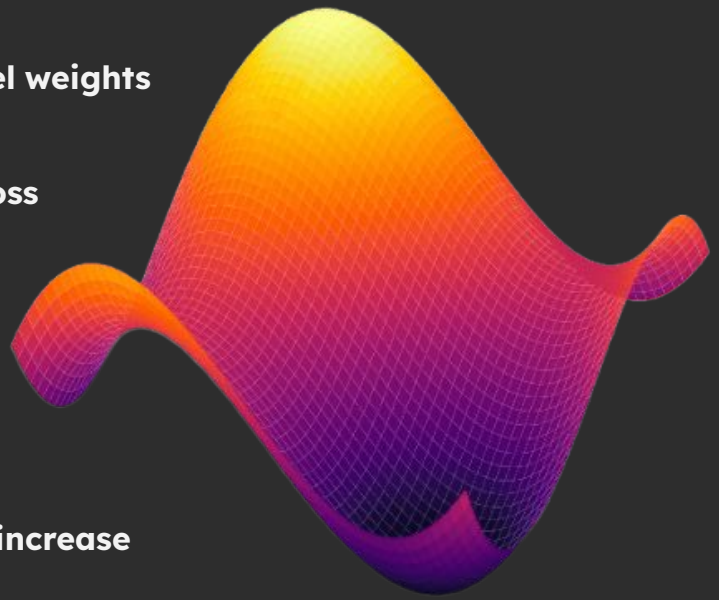
- Can be any **differentiable activation**
- E.g. words from a caption or answer to a question



GRAD-CAM INTUITION

Usually, **gradient** is calculated from the **loss w.r.t.** to **all model weights**

- Shows into the direction of **steepest ascent**
- If **negated**, shows how to **alter weights** to **lower** the **loss**



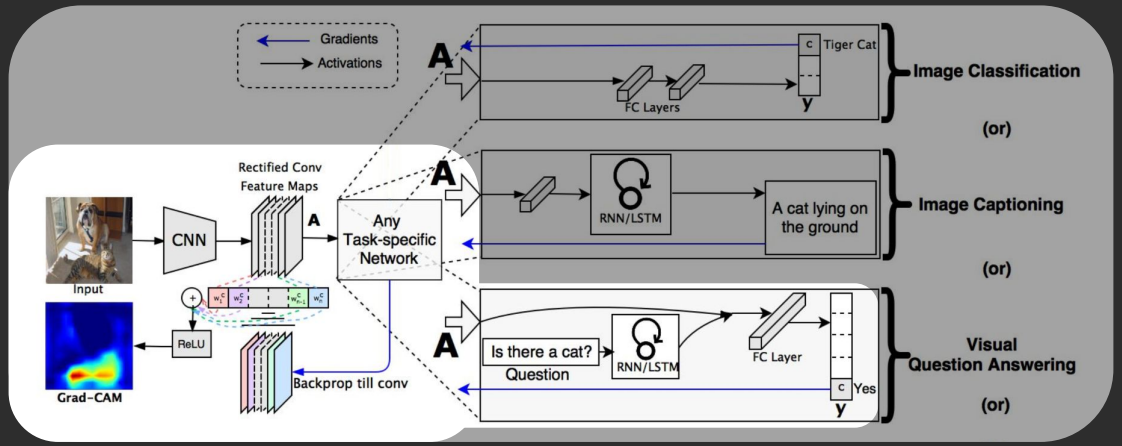
Grad-Cam

- **Gradient** of an **output value w.r.t.** to a **feature map**
- Shows us **which values** of that feature map to alter, to **increase prediction score**

GRAD-CAM EXPLAINED

Grad-CAM can be defined as the **ReLU**-fied **weighted sum** over all **feature maps** of any **conv layer**

The **weights** are the **GAP gradients** of the chosen **output w.r.t.** the chosen **feature maps**

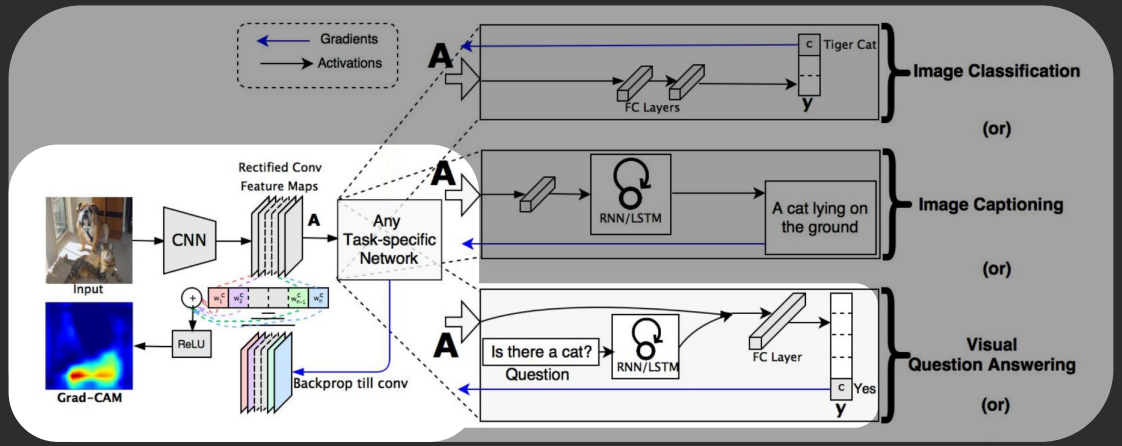


CLASS ACTIVATION MAPPING

GRAD-CAM EXPLAINED

$$\text{Grad-CAM}^c = \text{ReLU}\left(\sum_k^K \alpha_k^c A^k\right)$$

$$\alpha_k^c = \frac{1}{WH}\sum_w^W \sum_h^H \frac{\partial y^c}{\partial A_{wh}^k}$$



CLASS ACTIVATION MAPPING

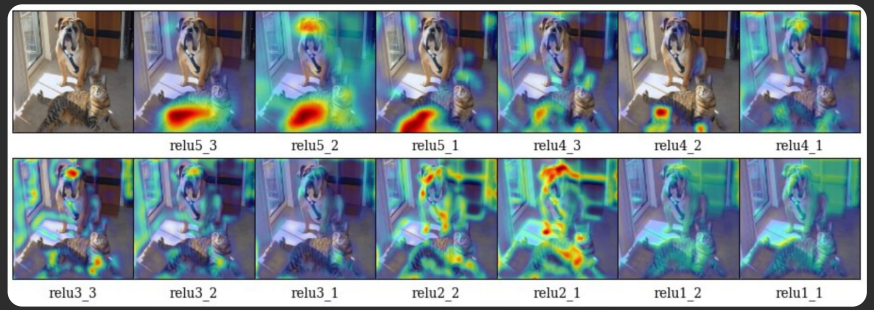
GRAD-CAM CAVEATS

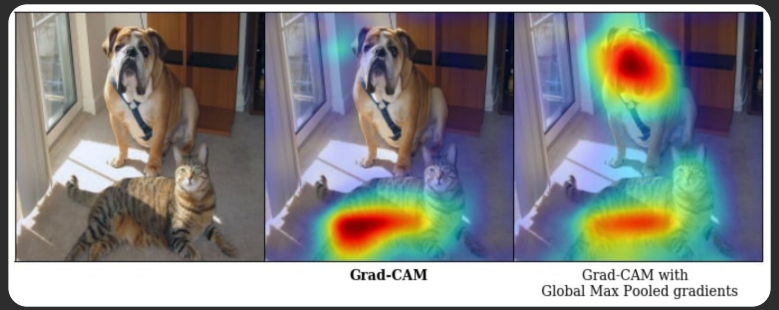
Which conv layer to use?

- Deeper representations in a CNN capture higher-level visual constructs
- However, it can be used to explain any conv layer of a network

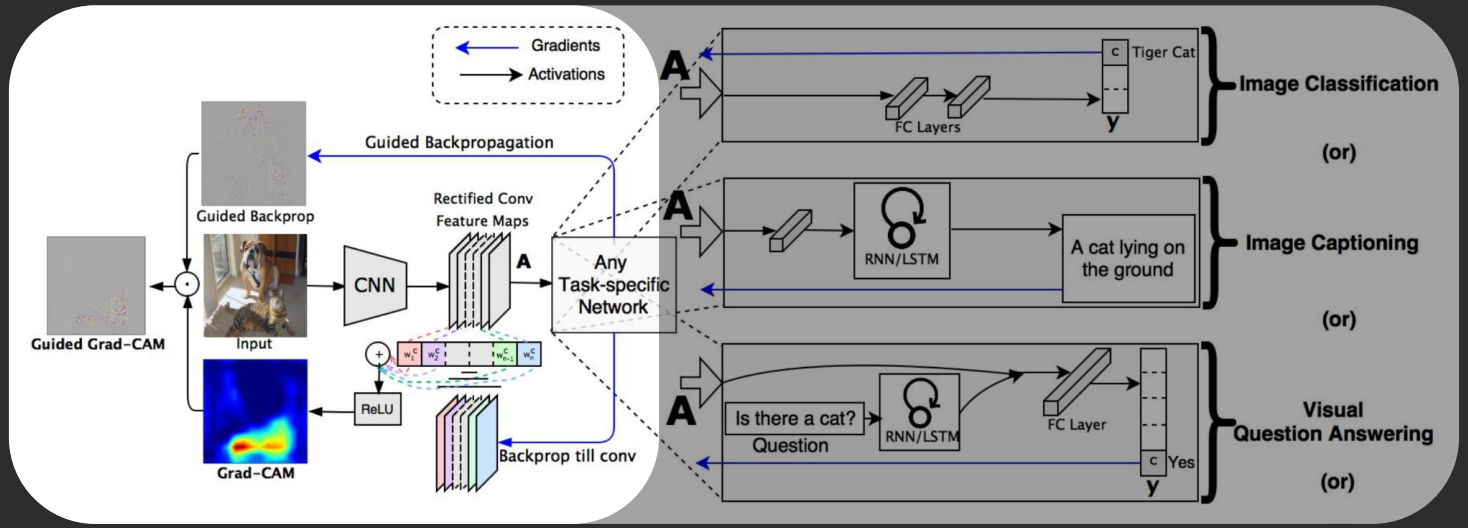
Why not use Global Max Pooling?

- GMP lowers the localization ability
- Max is statistically less robust to noise compared to the averaged gradient





GUIDED GRAD-CAM OVERVIEW



Makes Grad-CAM's **class discriminative** analysis more **fine-grained**!

GUIDED BACKPROPAGATION

(Grad)-CAM's **resolution** is **coarse**

→ Shows why the network thinks it sees a cat but not why it's a tiger cat

Original Image

Grad-Cam - Tiger Cat

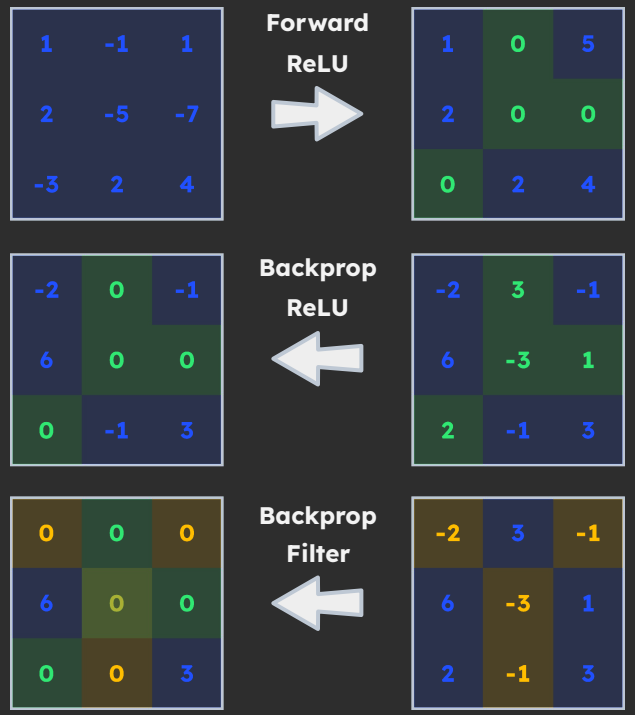
Guided BP - Tiger Cat

Guided BP is **pixel-space** gradient **visualization** method

→ By backpropagating to the image, it is much more **fine-grained**

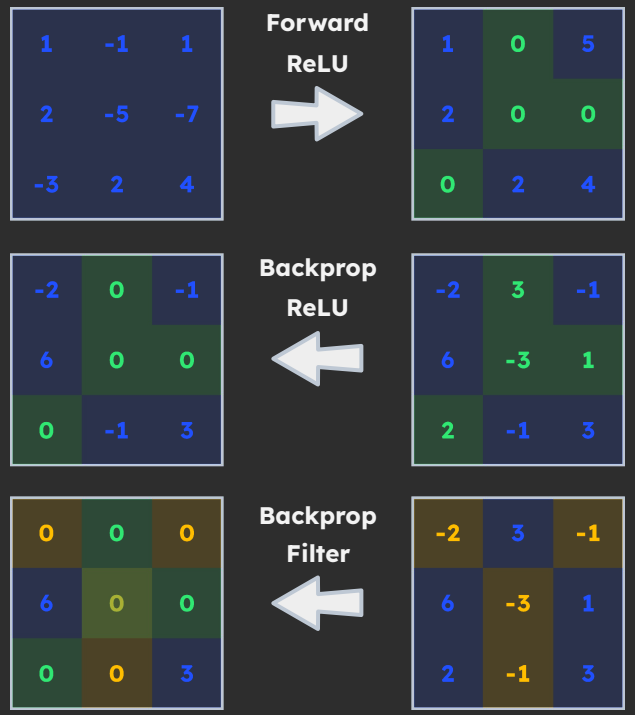
→ However, it is **not class discriminative**

GUIDED BACKPROPAGATION



- Backpropagate from a chosen **neurons activation to** the **input**
- **Neglect** all **negative** influences and **visualize** only **positive** gradients
  → The parts of the input that most strongly increase the activation
- **Only** applies to **ReLU** activated layers

GUIDED BACKPROPAGATION



Forward pass through ReLU

- Set negative values to 0

Backward pass through ReLU

- Only neurons whose activation were > 0 receive gradients

Guided Backpropagation pass through ReLU

- Additionally, set negative gradients to 0

GUIDED BACKPROPAGATION

1	-1	1
2	-5	-7
-3	2	4

Forward ReLU

1	0	5
2	0	0
0	2	4

-2	0	-1
6	0	0
0	-1	3

Backprop ReLU

-2	3	-1
6	-3	1
2	-1	3

0	0	0
6	0	0
0	0	3

Backprop Filter

-2	3	-1
6	-3	1
2	-1	3

GUIDED GRAD-CAM

Original Image

Guided BP - Tiger Cat

Grad-CAM - Tiger Cat

Guided Grad-CAM - Tiger Cat

Original Image

Guided BP - Bulldog

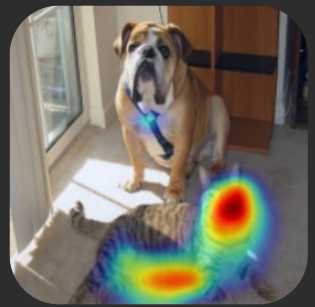
Grad-Cam - Bulldog

Guided Grad-CAM - Bulldog

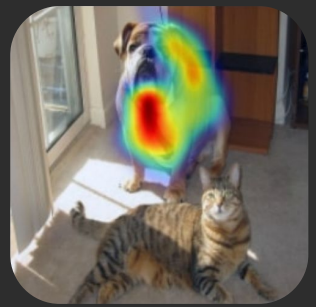
GRAD-CAM COUNTERFACTUAL EXPLANATION

Highlights areas that **most decrease** the chosen network **output probability** y^c

Original Image



Bulldog Counterfactual exp



Tiger Cat Counterfactual exp

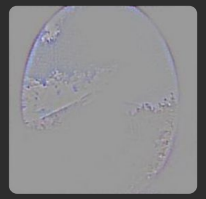
$$\alpha_k^c = \frac{1}{WH} \sum_w^W \sum_h^H -\frac{\partial y^c}{\partial A_{wh}^k}$$

Guided GRAD-CAM FAILURE ANALYSIS

Ground truth: volcano

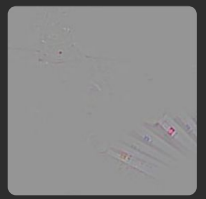
Predicted: sandbar

Ground truth: volcano

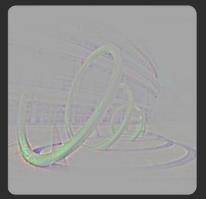


Predicted: car mirror

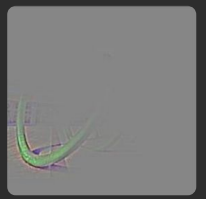
Ground truth: beaker



Predicted: syringe



Ground truth: coil



Predicted: vine snake

CLASS ACTIVATION MAPPING

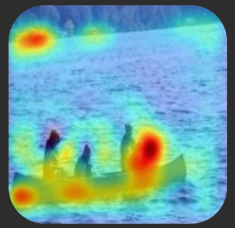
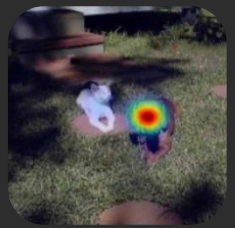
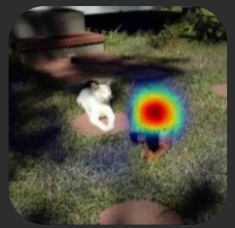
GRAD-CAM CAPTIONING EXPLANATION

A group of people is flying kites on a beach

A man is sitting at a table with a pizza

CLASS ACTIVATION MAPPING

GRAD-CAM vs. CAM

CAM	Grad-CAM	CAM	Grad-CAM
			
Boat	Boat	Cat	Cat
			
Person	Person	Dog	Dog

CLASS ACTIVATION MAPPING

Grad-CAM DISADVANTAGES

Heatmap is rather **coarse** → can be countered using guided grad-cam though

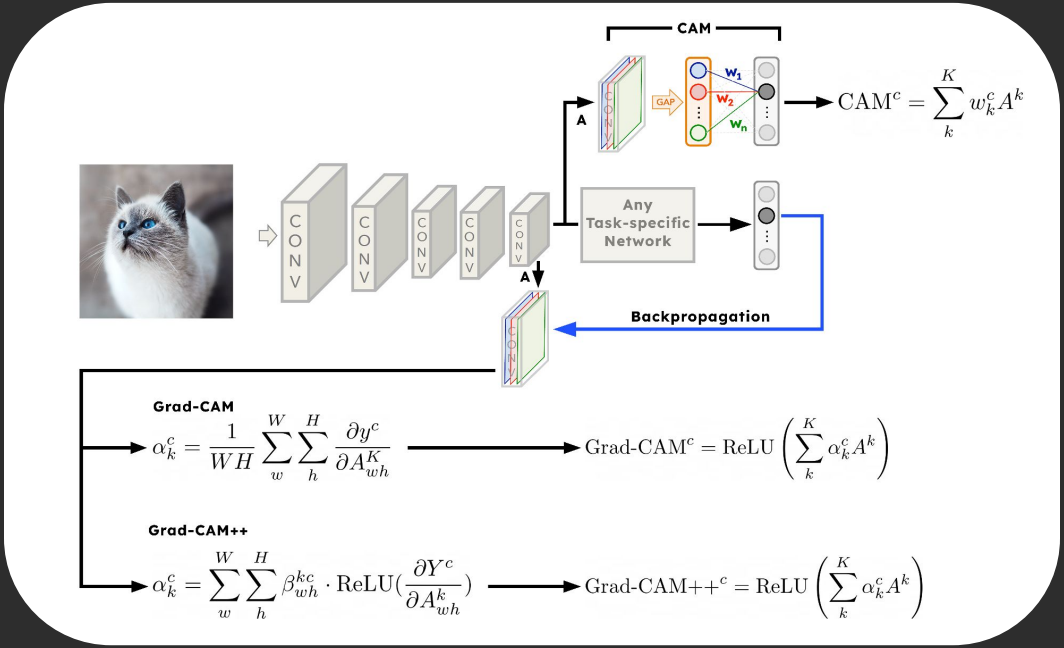
Lacks the ability for **detecting multiple instances** of the same class

Often does **not capture** the **entire object**

1 XAI POST-HOC
2 CAM
3 GRAD-CAM
4 GRAD-CAM++
5 DEMO

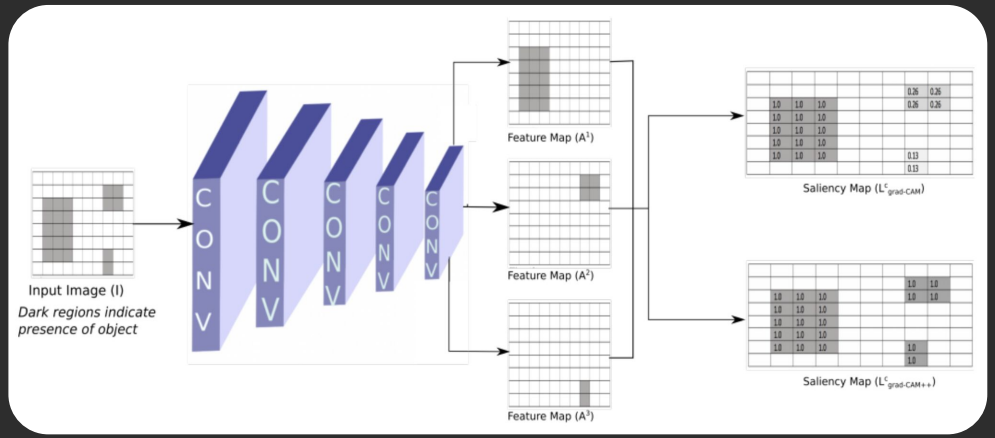
GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING++

GRAD-CAM++ OVERVIEW



Improves object **localization** and explains **multiple** object **occurrences better!**

GRAD-CAM++ INTUITION



Task: Binary Classification (presence / absence of grey pixels)

Assume: $$\frac{\partial y^c}{\partial A^k_{wh}} = \begin{cases} 1, \text{if grey pixel} \\ 0, \text{otherwise} \end{cases}$$

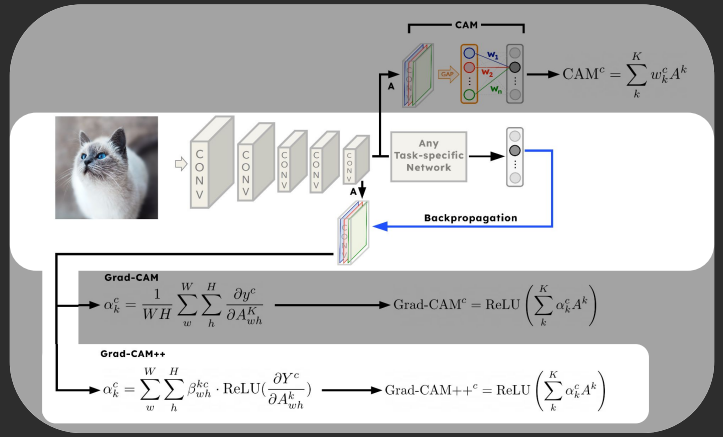
Grad-CAM favours larger ROIs (regions of interests) → Due to GAP gradients

Grad-CAM++ mitigates that by adding pixel-wise weights → **Highlights** all **ROIs equally**

GRAD-CAM++ OVERVIEW

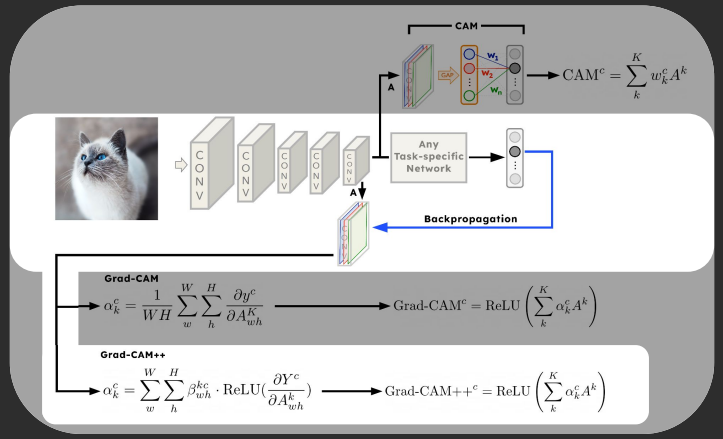
Grad-CAM++ can be defined as the **ReLU**-fied **weighted sum** over all **feature maps** of any **conv layer**

The **weights** are the **summed pixel-wise weighted ReLU**-fied **gradients** of the chosen **output w.r.t.** the chosen **feature maps**

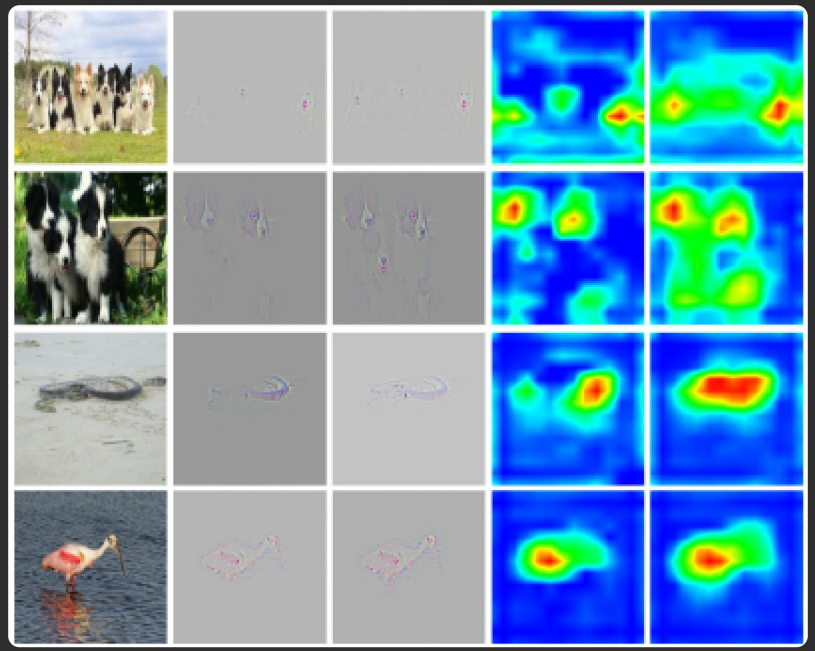


GRAD-CAM++ OVERVIEW

$$\alpha_k^c = \sum_w^W \sum_h^H \beta_{wh}^{kc} \cdot \mathrm{ReLU}\left(\frac{\partial Y^c}{\partial A_{wh}^k}\right) \qquad \beta_{wh}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{wh}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{wh}^k)^2} + \sum_l^L \sum_m^M A_{lm}^k \left\{\frac{\partial^3 Y^c}{(\partial A_{wh}^k)^3}\right\}}$$



GRAD-CAM++ EVALUATION



1 XAI POST-HOC

2 CAM

3 GRAD-CAM

4 GRAD-CAM++

5 DEMO

IT'S SHOWTIME

CLASS ACTIVATION MAPPING

RESOURCES

Main Sources:

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods." *Entropy* 23.1 (2020): 18.

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016.

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision.* 2017.

Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV).* IEEE, 2018.

Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net." *arXiv preprint arXiv:1412.6806* (2014).

RESOURCES

Others:

https://www.flaticon.com/

https://unsplash.com/

https://steamcommunity.com/sharedfiles/filedetails/?id=2211772440