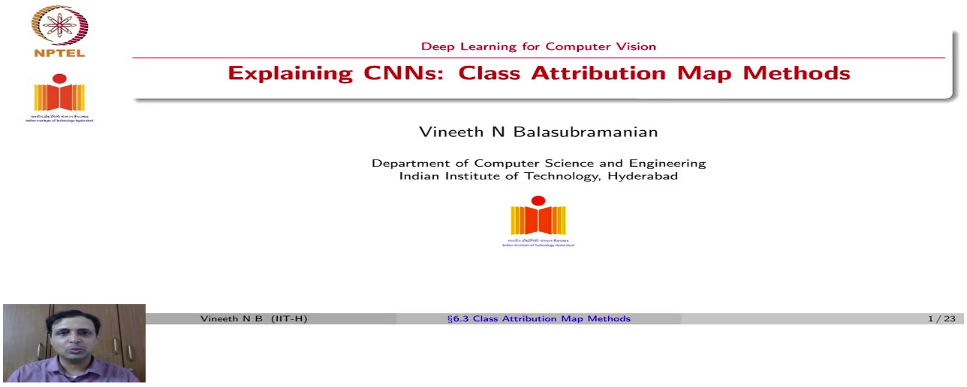


Deep Learning for Computer Vision

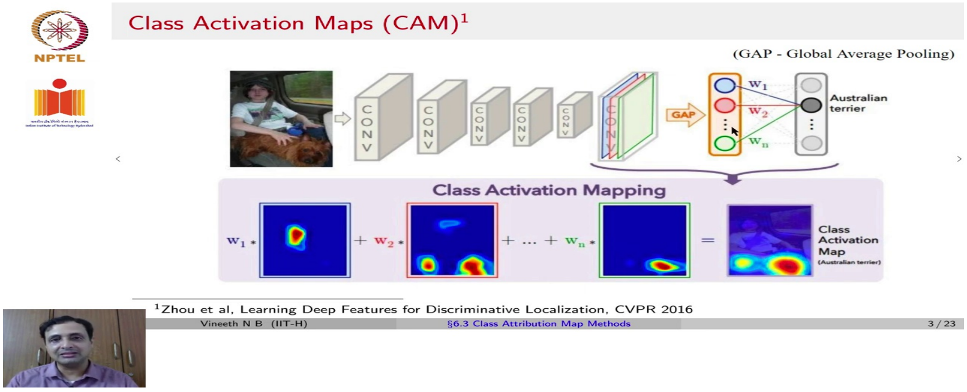
Explaining CNNs: Class Attribution Map Methods

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 1 / 23

And then for each map in that convolutional layer remember you could have 100 filters, you could have 100 maps so for each of those maps you do global average pooling or GAP.What does global average pooling do? You take a particular attribution map or a feature map sorry not an attribution map a feature map this green one and you average all the intensity values there into one single scalar and that becomes this green circle here.Similarly, you take the red feature map and average all its values and it becomes the red scalar here.Now, the weights that we learned between the output of the gap layer and the classification layer are now used to weight each of these activation maps.



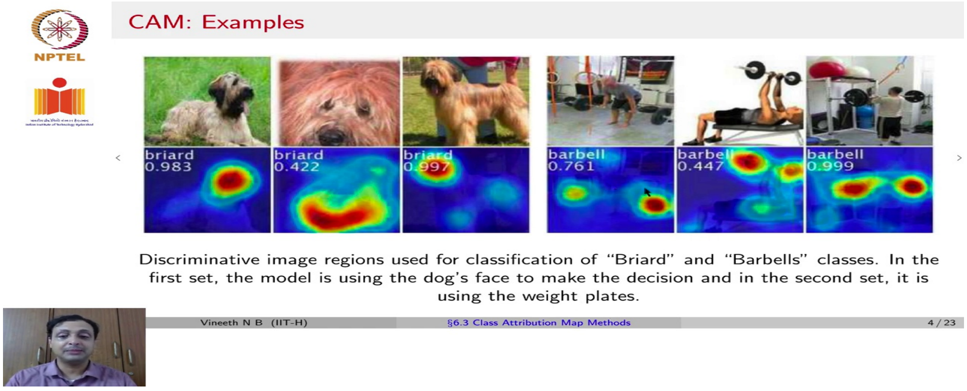
Class Activation Maps (CAM)[1]

(GAP - Global Average Pooling)

[1] Zhou et al, Learning Deep Features for Discriminative Localization, CVPR 2016

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 3 / 23

And when you do a weighted sum of all of these activation maps that gives us the contribution of these activation maps towards one particular class label.Similarly so in this case similarly so in this case and similarly for the barbells you can actually see that the CNN looks at the weight plates to make the decision that this image corresponds to barbells.

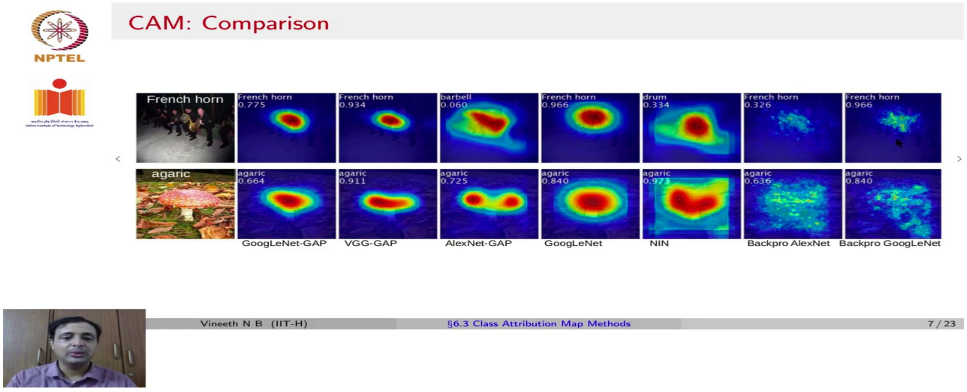


CAM: Examples

Discriminative image regions used for classification of “Briard” and “Barbells” classes. In the first set, the model is using the dog’s face to make the decision and in the second set, it is using the weight plates.

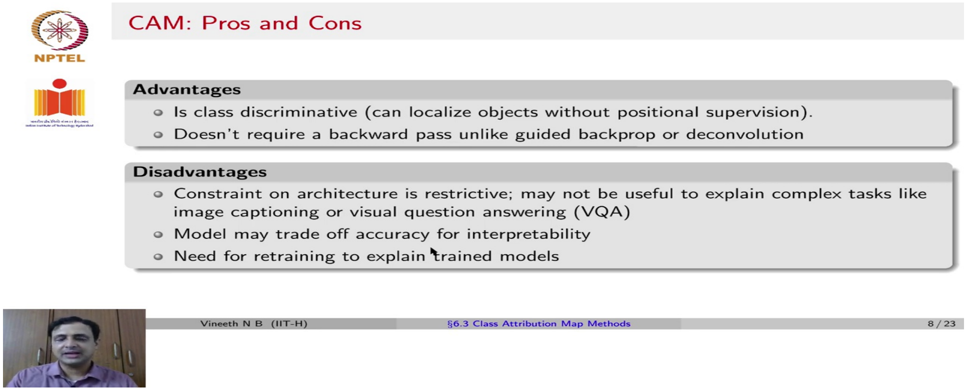
Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 4 / 23

If you recall in the first lecture we saw an example of the conv5 feature map with two people sitting in an image.And we saw that the feature map actually showed where they were positioned in the image.We actually see now that we can retrieve back the activation maps of the fifth layer and be able to use them to explain our decisions at the classification layer.And you can see here that you do get a certain set you can you get a certain sense of localization through these kinds of feature maps at different convolutional layers.



CAM: Comparison

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 7 / 23



CAM: Pros and Cons

Advantages

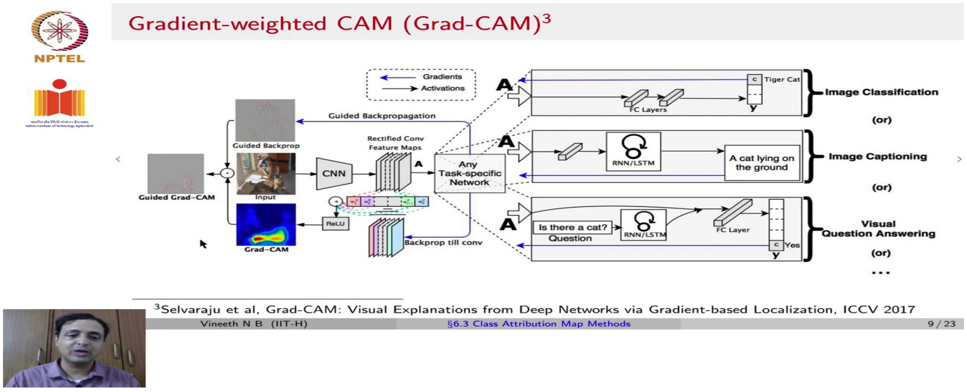
- Is class discriminative (can localize objects without positional supervision).
- Doesn't require a backward pass unlike guided backprop or deconvolution

Disadvantages

- Constraint on architecture is restrictive; may not be useful to explain complex tasks like image captioning or visual question answering (VQA)
- Model may trade off accuracy for interpretability
- Need for retraining to explain trained models

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 8 / 23

You now combine them you combine the gradients that you get for each of those activation maps and they automatically become your weights for each of the feature maps.And in grad cam because we want to ensure that only positive correlations are shown in the final saliency map we apply a ReLU on the on the weighted combination of the activation maps and that becomes our final grad cam saliency map.The method also talks about adding guided back propagation to make a variant of grad cam called guided grad cam.Let us see this in a bit more detail and also mathematically as to why grad cam becomes an extension of cam.

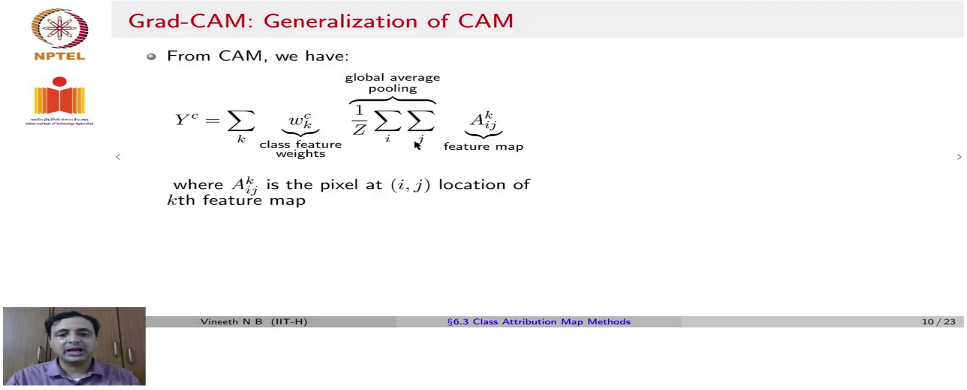


Gradient-weighted CAM (Grad-CAM)[3]

[3] Selvaraju et al, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 9 / 23

Let us see this in a bit more detail and also mathematically as to why grad cam becomes an extension of cam.From cam mathematically speaking, we have Yc which is your scores or the class course in the last layer which is given by summation over k which is all your k feature maps.We assume now that you have k such feature maps and you have a class weights for each of these feature maps and summation over i summation over j, a i j k, 1 by z is going to be your global average pooling of each of the k feature maps.



Grad-CAM: Generalization of CAM

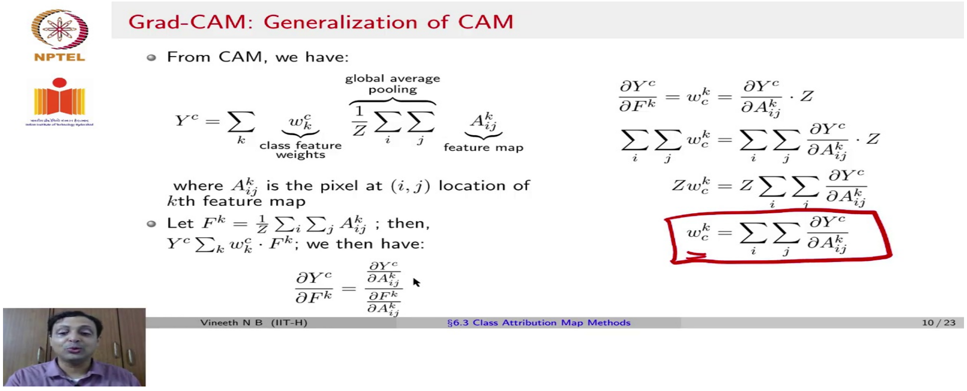
- From CAM, we have:

$$Y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}}$$

where A_{ij}^k is the pixel at (i, j) location of kth feature map

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 10 / 23

Now let us go from there.Now the summation over i and j for wck because wck does not depend on i and j is just going to be z which is the total number of pixels in each feature map or activation map and similarly the z constant comes out here and you have your summation over i summation over j dou Yc by dou a i j k here.



Grad-CAM: Generalization of CAM

- From CAM, we have:

$$Y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}}$$

where A_{ij}^k is the pixel at (i, j) location of kth feature map

- Let $F^k = \frac{1}{Z}\sum_i \sum_j A_{ij}^k$; then, $Y^c \sum_k w_k^c \cdot F^k$; we then have:

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

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

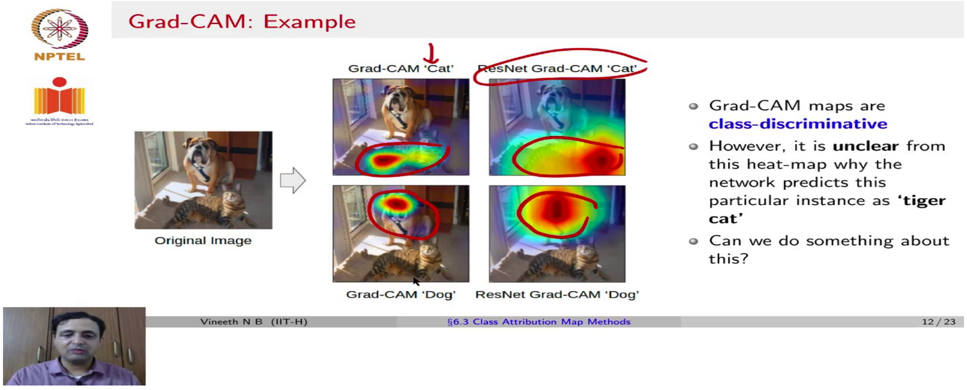
$$\sum_i \sum_j w_c^k = \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z$$

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

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

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 10 / 23

This tells us tells us something important that the wck which we actually learned in the cam model are actually simply the summation of the gradients of each the each of the class score with respect to every pixel in the feature map and adding them all up.You do not need the global average pooling you do not need the retraining those weights that you did the global average pooling for can actually be obtained as gradients of the last layer scores with respect to any feature map or activation map with which you want to compute your saliency maps.You see examples now for original image here is the grad cam for cat which focuses on the cat and similarly doing this with the ResNet model, here is the location of the cat and similarly for a dog it looks at the dog to be able to say it is a dog and similarly for the ResNet model looks at the entire dog to be able to say it is a dog, which is quite good.

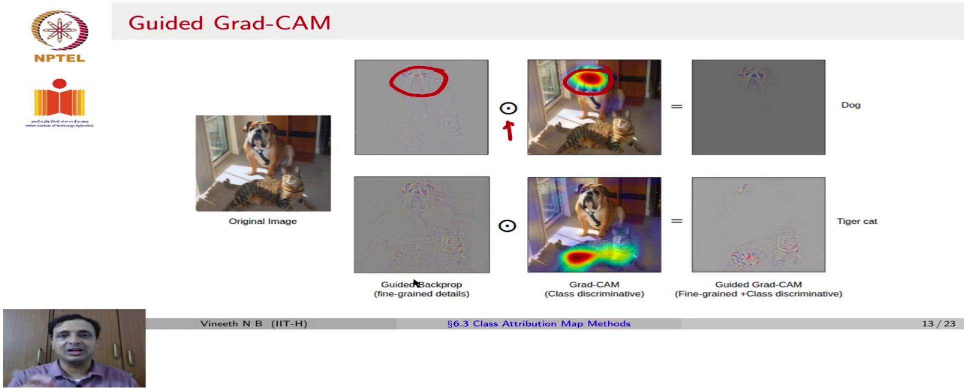


Grad-CAM: Example

- Grad-CAM maps are **class-discriminative**
- However, it is **unclear** from this heat-map why the network predicts this particular instance as **'tiger cat'**
- Can we do something about this?

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 12 / 23

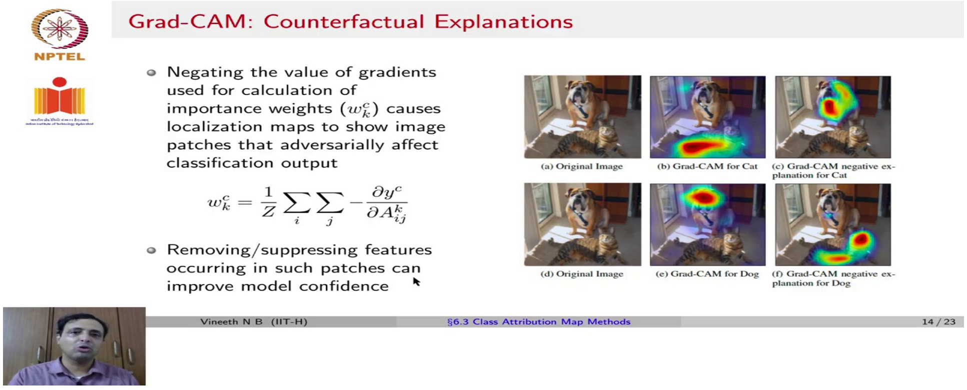
So can we elaborate on this further and see why it is a tiger cat and not just a cat, can we try to do anything further? And here is where the method now proposes a variant of grad cam called guided grad cam which brings together guided back propagation that we saw in the earlier lecture and grad cam together and juxtaposing them one on top of the other by doing what is known as a hadamard product or a pixel wise product.



Guided Grad-CAM

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 13 / 23

But by combining it with grad cam, we get a more localized understanding of what the CNN was looking at while calling it a dog.And now, you see why you called it a tiger cat why the model called it a tiger cat because you take the grad cam saliency map and combine it with guided back props output again and you actually get this kind of a output which shows the striations on the body of the cat which explains why it was a tiger cat.Grad cam went on to also show how this method could be used for what are known as counterfactual explanations.



Grad-CAM: Counterfactual Explanations

- Negating the value of gradients used for calculation of importance weights (w_k^c) causes localization maps to show image patches that adversarially affect classification output

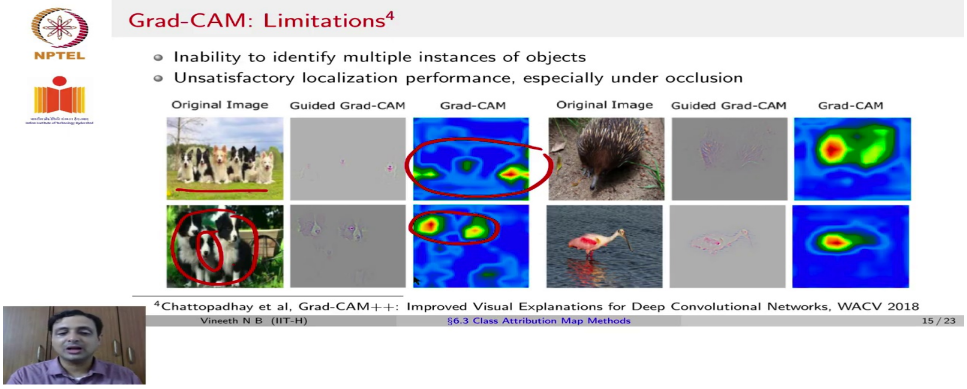
$$w_k^c = \frac{1}{Z}\sum_i \sum_j -\frac{\partial y^c}{\partial A_{ij}^k}$$

- Removing/suppressing features occurring in such patches can improve model confidence

(a) Original Image (b) Grad-CAM for Cat (c) Grad-CAM negative explanation for Cat

(d) Original Image (e) Grad-CAM for Dog (f) Grad-CAM negative explanation for Dog

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 14 / 23



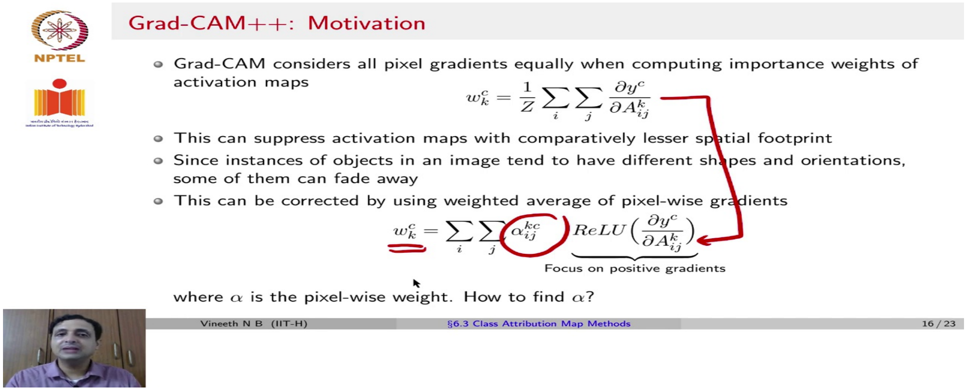
Grad-CAM: Limitations[4]

- Inability to identify multiple instances of objects
- Unsatisfactory localization performance, especially under occlusion

[4] Chattopadhay et al, Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks, WACV 2018

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 15 / 23

In both of these cases grad cam does not seem to capture those aspects which are actually salient aspects of that object as class discriminative in its in its visualization.Can we do something about it or are there any limitations in the formulation of grad cam itself that we can improve.And this was done in a work called grad cam plus plus and the main motivation of grad cam plus plus is observing that grad cam took the gradients of dou Yc with respect to dou, each of the pixels in your activation maps and then took an average of all of them to get its final weight.Grad cam plus plus's idea states that maybe pixels that contribute more towards the class should get more weight than have equal distribution while computing this weight w k c.Let us see how we can do that.So this can especially suppress activation maps with lesser spatial footprint rather we saw this example on the previous slide too when you have three dogs and there is one dog which is smaller the other two dogs got most of the gradient and the third dog did not because it has a lesser spatial footprint.However, this time while computing your final weight wkc we are going to give each pixel in each activation map a certain weight as to how they must be how they must contribute to that saliency map.



Grad-CAM++: Motivation

- Grad-CAM considers all pixel gradients equally when computing importance weights of activation maps

$$w_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

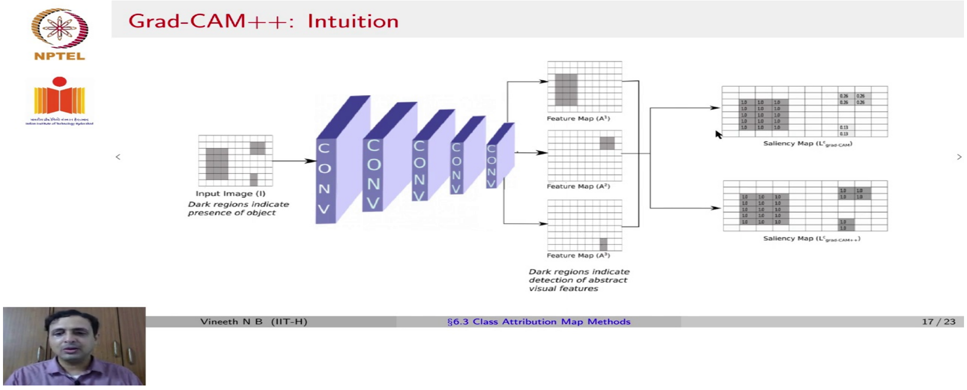
- This can suppress activation maps with comparatively lesser spatial footprint
- Since instances of objects in an image tend to have different shapes and orientations, some of them can fade away
- This can be corrected by using weighted average of pixel-wise gradients

$$w_k^c = \sum_i\sum_j \alpha_{ij}^{kc} \underbrace{ReLU\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)}_{\text{Focus on positive gradients}}$$

where α is the pixel-wise weight. How to find α?

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 16 / 23

K corresponds to the kth feature map and c corresponds to the class which we want to maximize.Grad cam plus plus also adds a ReLU here to ensure that only positive gradients are considered in the computation of this weight.But the larger question here is how do you get these weights? In grad cam, it was simpler to average all of these gradients that you get and use that to get a wkc and remember each wkc then becomes the weight of the kth feature map towards the cth class.So but now how do you compute these alpha ijs at each pixel level? Before we go there, let us try to understand the intuition of grad cam plus plus again visually.So here you notice that if you had an image with three different objects say dogs a dog of a large occupying a larger special footprint, another dog occupying a mid-level special footprint and another dog occupying a small footprint and for the moment let us assume that different feature maps capture different dogs.



Grad-CAM++: Intuition

Input Image (I)
Dark regions indicate presence of object
CONV
CONV
CONV
CONV
CONV
Feature Map (A¹)
Feature Map (A²)
Feature Map (A³)
Dark regions indicate detection of abstract visual features
Saliency Map (L^c_grad-CAM)
Saliency Map (L^c_grad-CAM++)
1.0
0.26
0.13

Vineeth N B (IIT-H) §6.3 Class Attribution Map Methods 17 / 23

So each feature map let us assume captures each of these objects and you see here that in grad cam when the saliency map gradients are computed you can see that the area with the largest footprint ends up getting most of the gradient, while the gradient towards the rest of the pixels are smaller because they have fewer pixels and hence contribute lesser towards the output while grad cam plus plus tries to overcome this by doing the pixel wise weighting and you can see here that in grad cam plus plus the weights are in the same range those the gradients are in the same range when you use this kind of an approach.So we still are left with the question in grad cam plus plus as to how do you compute those alphas at a pixel level.A kite a go-kart and an eel where grad cam plus plus localizations improve over grad cam by considering the pixel wise waiting strategy.For homework, there are these three papers cam, grad cam and grad cam plus plus and your job would be to read through them and the other exercise would be to work out how we get the closed form expression for alphas in grad cam plus plus.