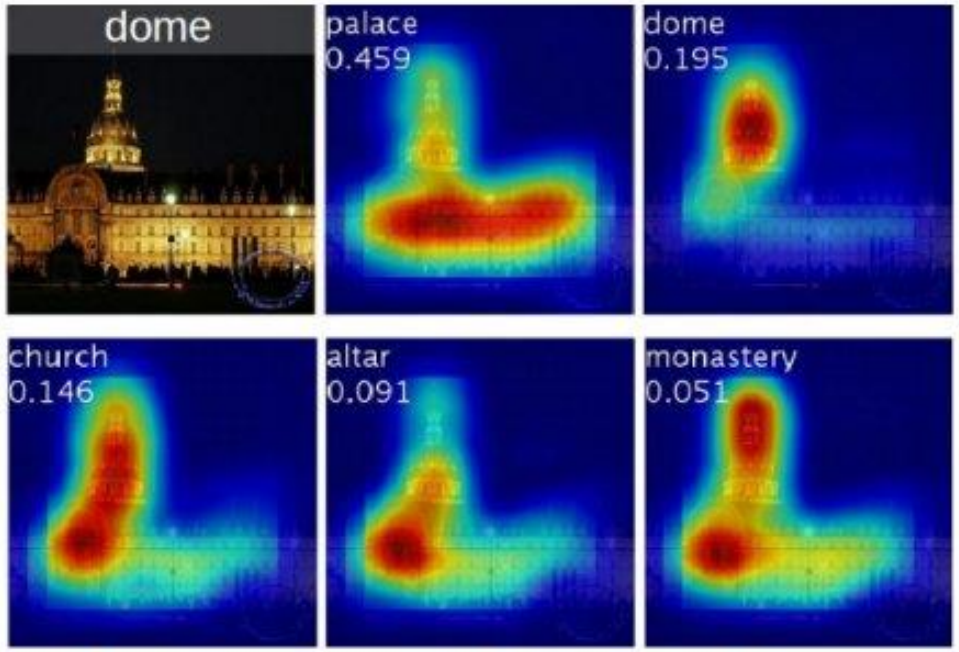
Deep Learning:
Class Activation Maps Theory

A technique for making Convolutional Neural Network (CNN)-based models more transparent by visualizing the regions of input that are “important” for predictions from these models — or visual explanations

Class Activation Maps

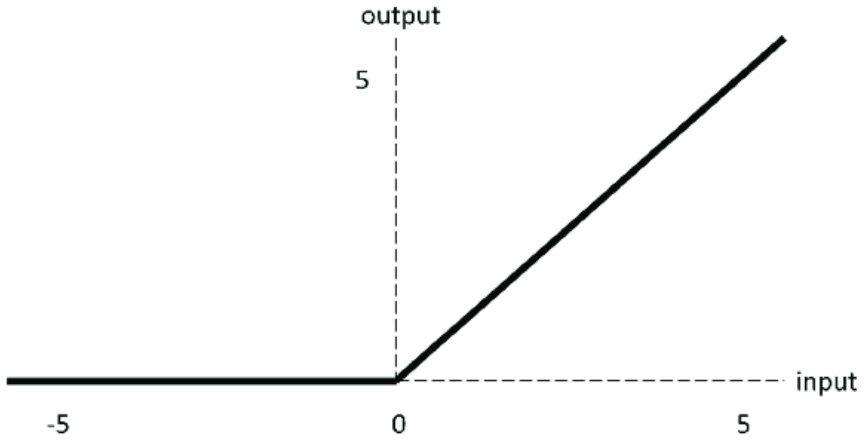


Class activation maps of top 5 predictions

Class activation maps for one object class

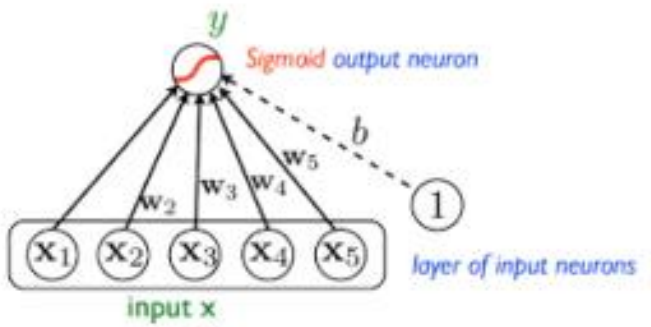
Class Activation Maps

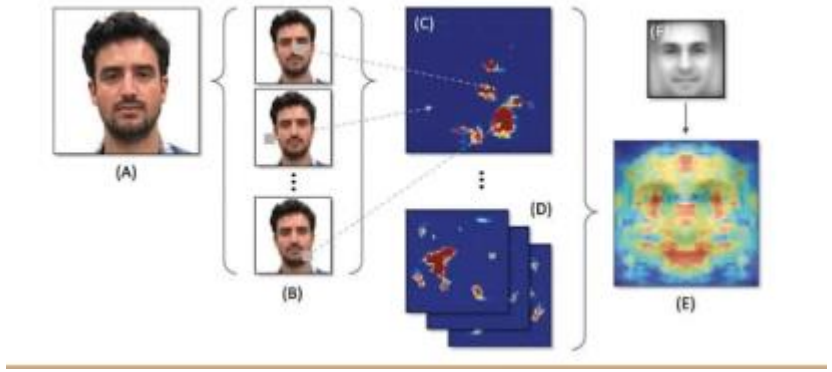
- Only need to do classification!
- Take any pre-trained CNN, e.g. ResNet.
- Image shrinks, but # features increase .
- RELU: all features are positive or zero.



Class Activation Maps

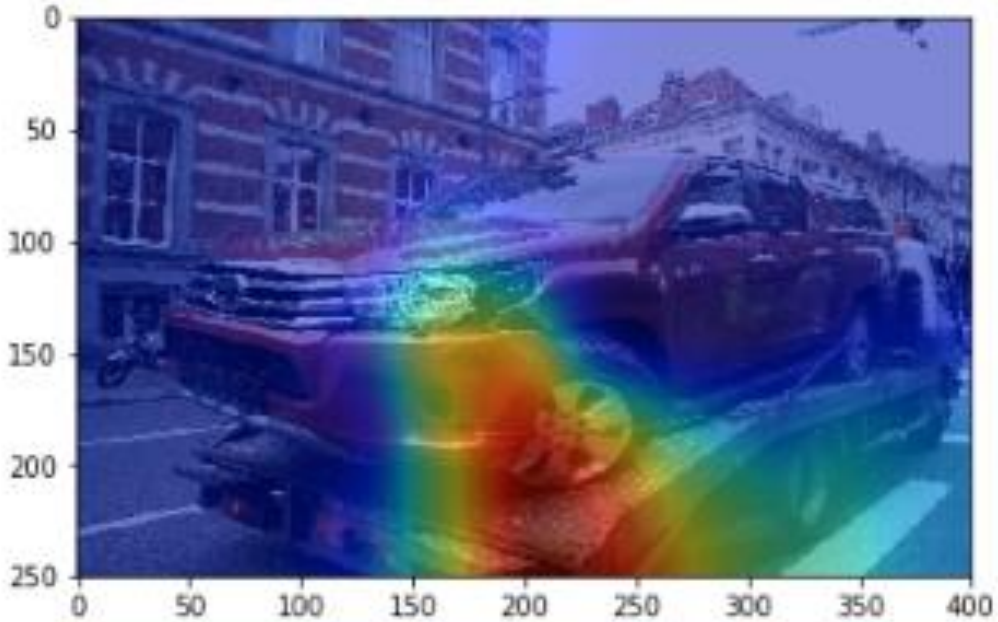
- Intuitively, you can think of a feature going into the Logistic Regression as a number denoting whether or not some “thing” appears in the image.
  - E.g. One feature for nose, one for eyes, one for lips, hair, ears, etc.
  - Positive number if “thing” was found, 0 otherwise.
  - E.g. The feature for “wheel” would be 0.





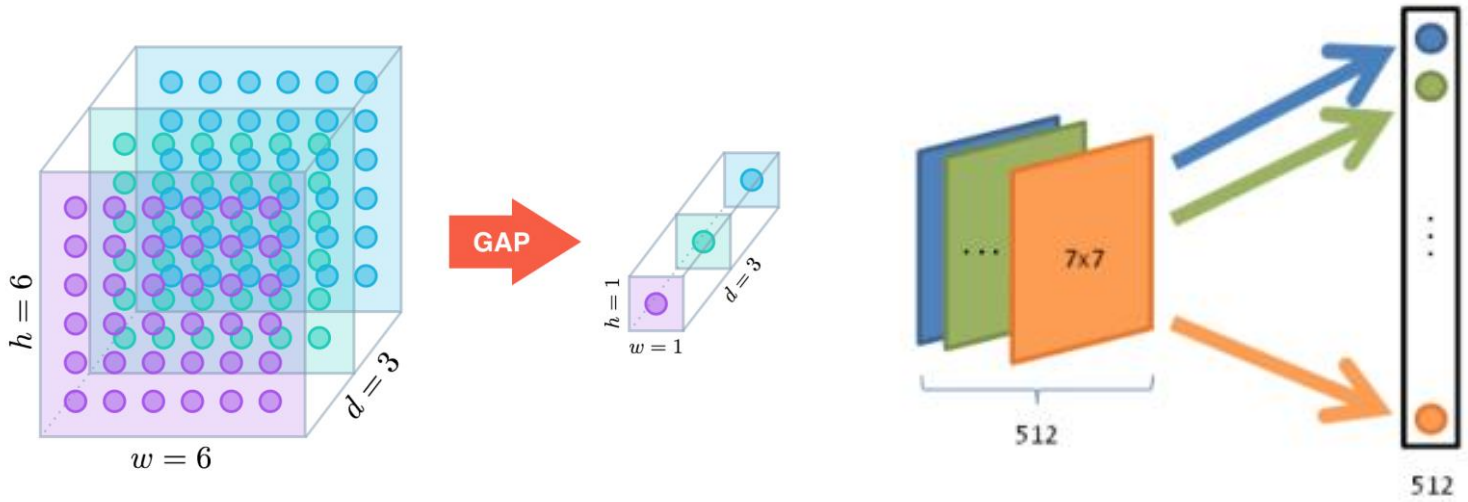
Features

- If the picture is of a car, then the feature for “wheel” would be > 0, if a wheel was found.
- Now, the nose feature would be 0.



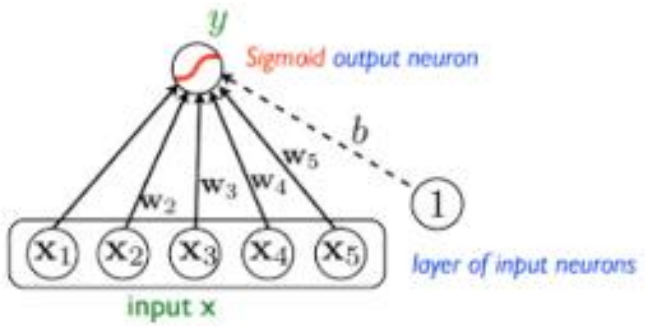
What led to the positive feature?

- If a feature is positive, that means the pooling operation must have found some positive numbers in the final image (after going through several layers of convolutions).
- i.e. That feature must have been found “somewhere”.
- If we simply looked at the image before pooling, then we would know where!



Logistic Regression

- If a weight is > 0, then the corresponding feature is positively correlated with this class.
- If it is 0, it has no effect.
- If it is < 0, the feature makes the image less likely to belong to this class.



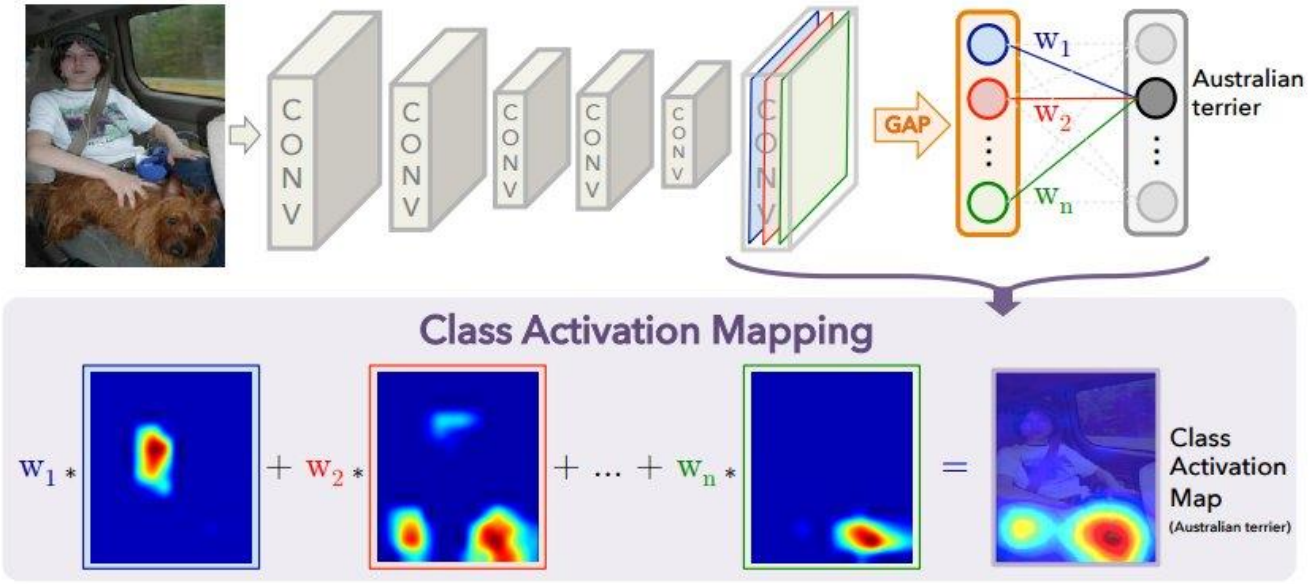
Class Activation Map

- Combine these intuitions
- Las few layers of ResNet:
  - 2048 7x7 images before final pooling
  - 2048 x 1000 dense weights (2048 weights for 1000 classes)

activation_49 (Activation)	(None, 7, 7, 2048)	0	add_16[0][0]
avg_pool (AveragePooling2D)	(None, 1, 1, 2048)	0	activation_49[0][0]
flatten_1 (Flatten)	(None, 2048)	0	avg_pool[0][0]
fc1000 (Dense)	(None, 1000)	2049000	flatten_1[0][0]

Class Activation Map

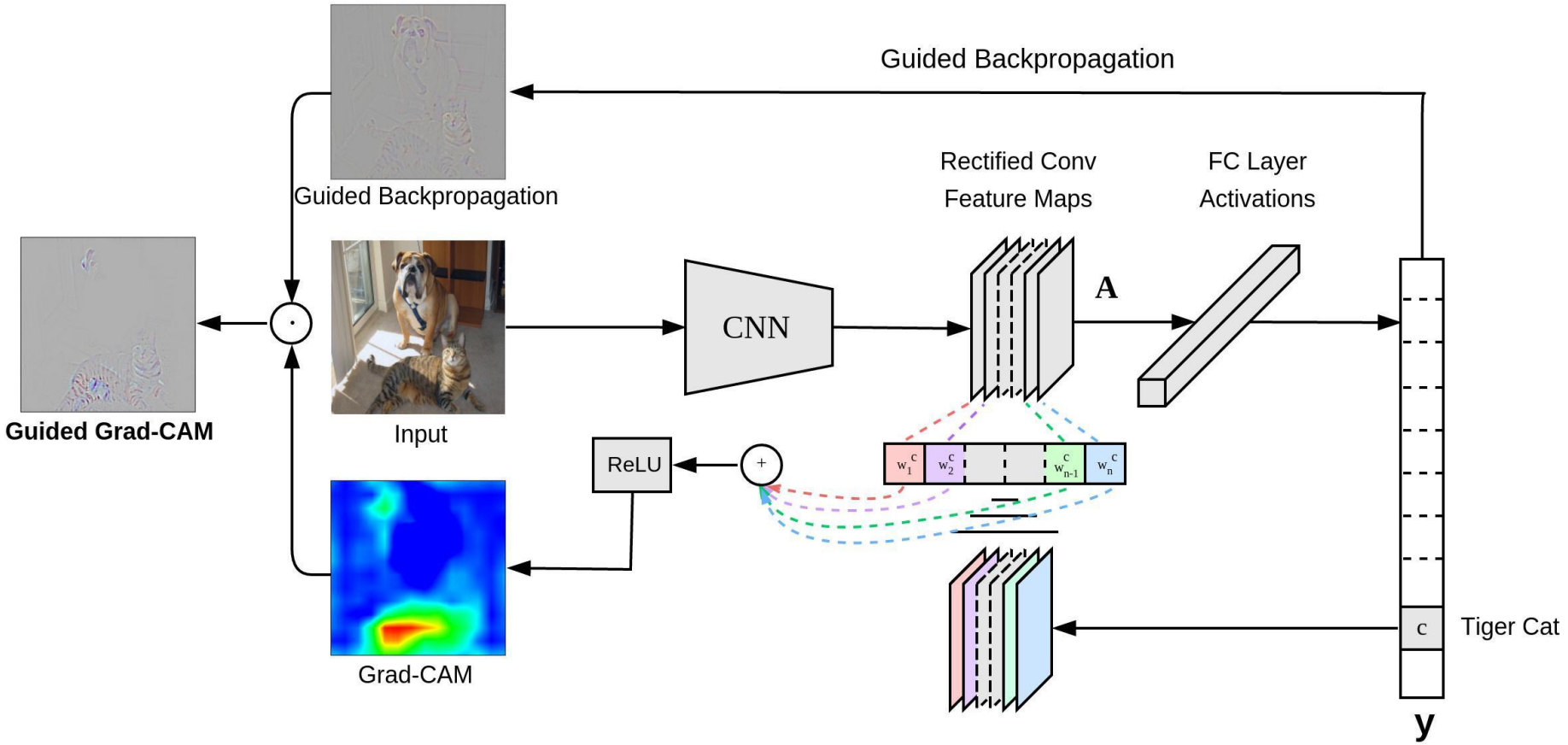
- We only consider 1 class at a time (usually the predict class).
- E.g. w = W[:, human_face_index] # size 2048.
- F = 2048 7x7 images.
- Class Activation Map = F[0]*w[0] + F[1]w[1] + ... F[2047]w[2047].
- Result is a 7x7 heat map.



Final Step

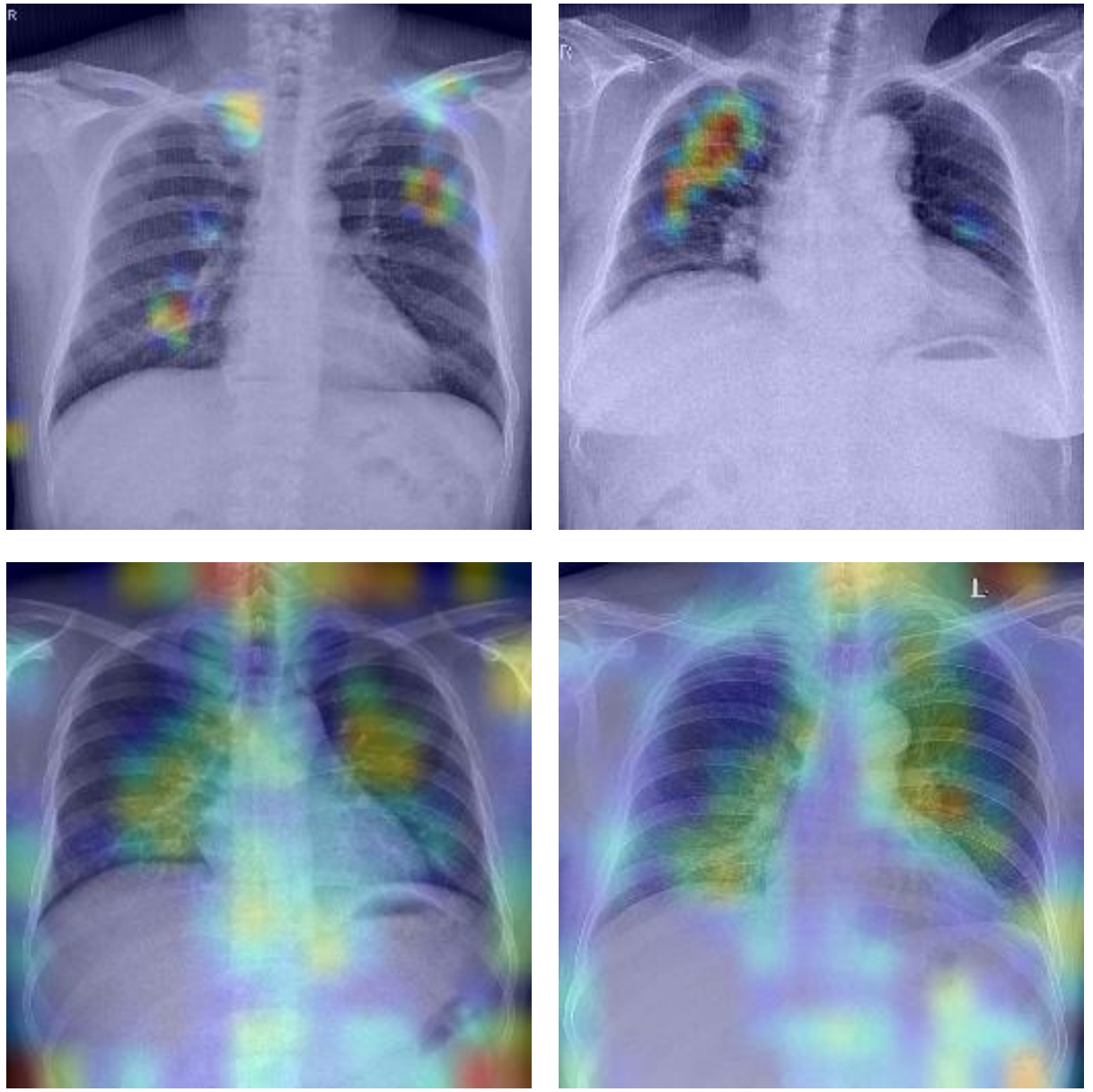
- Rescale the 7x7 image to the original image's size (224 x 244 for ResNet), and plot the 2 images over each other.

Conclusion


Guided Backpropagation
Guided Backpropagation
Rectified Conv
Feature Maps
FC Layer
Activations
A
CNN
Guided Grad-CAM
Input
ReLU
Grad-CAM
C
Tiger Cat
y

Chest X-Rays

- examples using as images of VinBigData and the Inception architecture.



Grad CAM vs. LIME