Grad-CAM

Visual Explanations from Networks via Gradient-based Localization

Team Name : Android Kunjappans

Team Members :

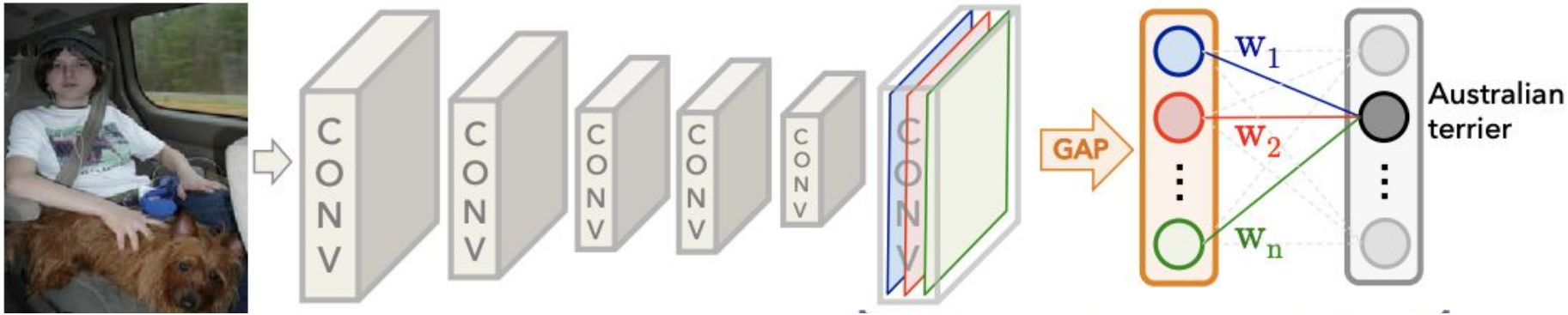
Abhiram G P	2022201024
Hari Krishnan U M	2021202022
Vyshak P	2022201064

Problem statement

- Lack of interpretability in deep learning models poses a challenge to building trust in intelligent systems
- Transparency and explanations are essential for identifying failure modes, establishing trust and confidence in users, and enabling machine teaching.
- There is a trade-off between accuracy and interpretability in deep learning models.
- Residual networks (ResNets) have shown state-of-the-art performance, but their complexity makes them difficult to interpret.
- There is a need to explore the spectrum between interpretability and accuracy in deep learning models to build more transparent and trustworthy intelligent systems.

Existing solution - CAM

- Zhou et al. proposed Class Activation Mapping (CAM) for identifying discriminative regions in CNNs without fully-connected layers.



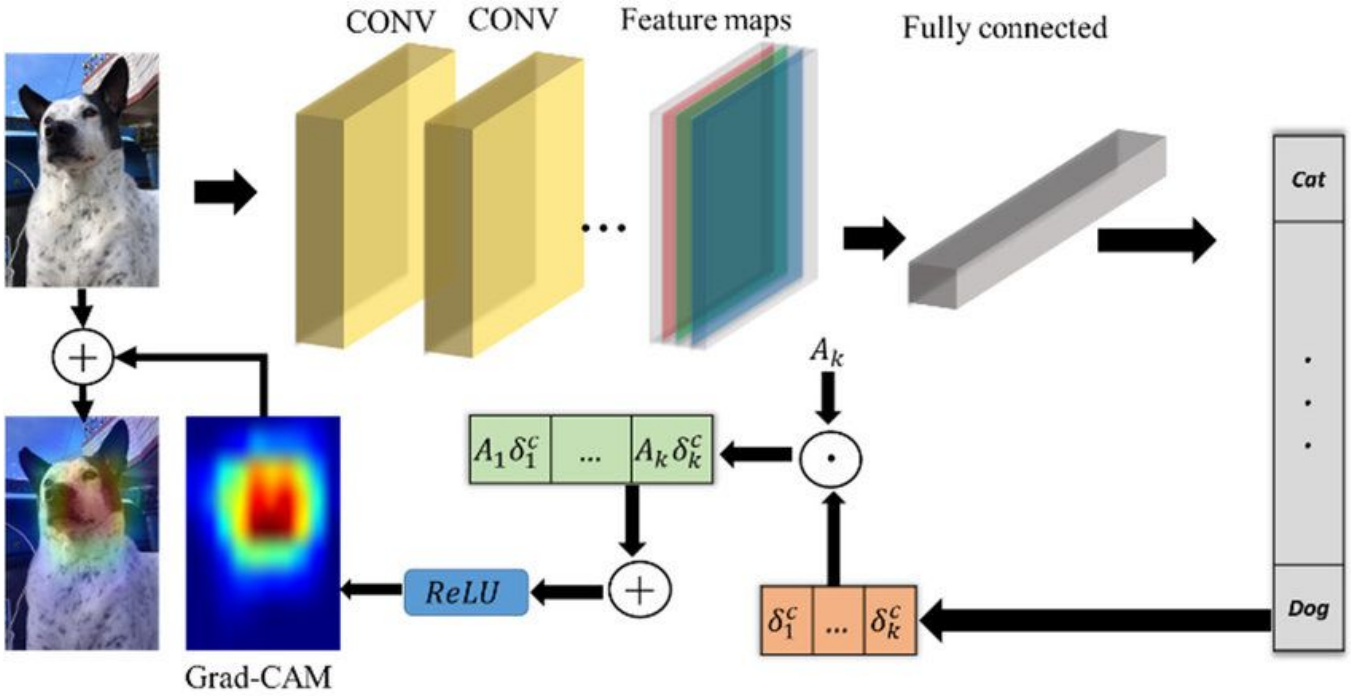
CAM - Class Activation Mapping

- CAM uses global average pooling and linear transformations to generate class scores for each feature map.
- CAM produces a coarse localization map by taking a weighted sum of the activation maps using the class scores.
- However, CAM has limited applicability and sacrifices accuracy for transparency.
- There is a need for more interpretable and accurate models to build trustworthy intelligent systems.

Proposed Solution

- Grad-CAM is a generalization of CAM that is applicable to a wider range of DNN architectures, including those with fully connected layers.
- Unlike CAM, which only localizes the important regions at a coarse level, Grad-CAM generates fine-grained heatmaps that highlight the most discriminative regions in an image for a particular class.
- Grad-CAM can be used to visualize the attention of different layers of a DNN, allowing researchers to gain insights into how the model makes its decisions at different levels of abstraction.
- Grad-CAM is a non-intrusive technique that can be applied to any pre-trained DNN without modifying the architecture or training process. This makes it a useful tool for interpreting existing models and gaining insights into their behavior.

Grad-CAM


CONV
CONV
Feature maps
Fully connected
Cat
Dog
A_k
$A_1\delta_1^c$ ... $A_k\delta_k^c$
δ_1^c ... δ_k^c
ReLU
Grad-CAM

Project scope

- Implementing Grad-CAM on a pre-trained image classification model:
  - Understand the Grad-CAM algorithm
  - Modify the pre-trained model to incorporate Grad-CAM
  - Visualize the activation maps for different classes
- Using Grad-CAM to evaluate the robustness of a model to adversarial attacks:
  - Generate adversarial examples
  - Visualize the activation maps for both the original and adversarial images
  - Determine if the model is relying on spurious features

Implementation details

Grad-CAM

- A technique for visualizing the regions of an image that contribute most to the classification decision made by a deep neural network.

Counterfactual Grad-CAM

- A variation of Grad-CAM that highlights the regions of an image that would need to be changed in order to change the classification decision.

Adversarial attack on images

- A process of intentionally modifying an image to deceive a deep neural network into misclassifying it.

Experiments conducted

- Conducted experiments on ResNet-50 and VGG-16 using Grad-CAM and observed variations on different convolutional layers.
- Examined the effect of taking ReLU on saliency maps obtained using Grad-CAM.
- Used Counterfactual Grad-CAM on ResNet-50 and VGG-16 and observed variations on different convolutional layers.
- Evaluated the robustness of Grad-CAM against adversarial attacks.

Grad-CAM generalizes CAM

- GradCAM is related to CAM (Class Activation Mapping) proposed by Zhou et al. in 2016
- GradCAM extends CAM by introducing a gradient-based weight factor that combines the class activation map with the gradient of the score with respect to the feature maps
- This results in a more fine-grained localization map highlighting important regions of the image for a given class
- GradCAM generalizes CAM for various CNN-based architectures

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

Grad-CAM generalizes CAM

Let F^K be global averaged pool

$$F^k = \frac{1}{Z}\sum_i\sum_j A_{ij}^k \qquad Y^c = \sum_k w_k^c \cdot F^k$$

Now taking the gradient of the score for class *c* (Y^c) wrt to the feature map F^K we get,

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

Grad-CAM generalizes CAM

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

Summing both sides of over all pixels (i, j),

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

Grad-CAM generalizes CAM

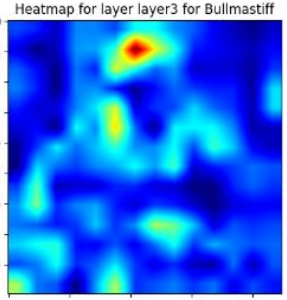
- Grad-CAM introduces a gradient-based weight factor that combines the class activation map with the gradient of the score with respect to the feature maps, resulting in a more fine-grained localization map.
- The expression for the weight factor used in CAM is a special case of the weight factor used in Grad-CAM, making Grad-CAM a strict generalization of CAM.
- This generalization allows Grad-CAM to be applied to a wider range of CNN model architectures beyond those that can be handled by CAM, making it a more general and powerful tool for interpreting deep learning models.

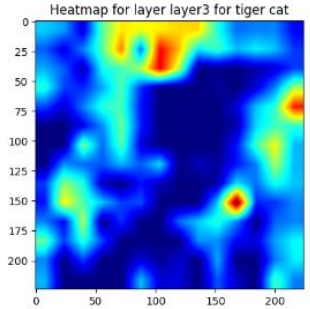
Resnet50 - layer3 vs layer4

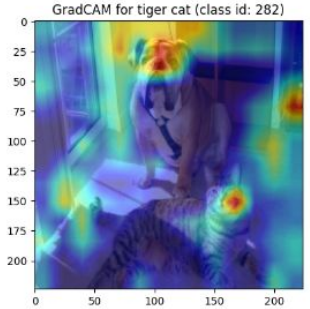
Bullmastiff

Tiger-Cat

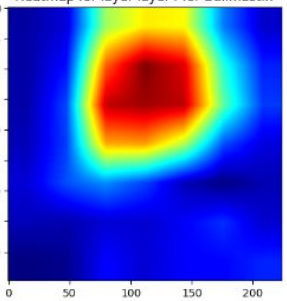
Layer-3

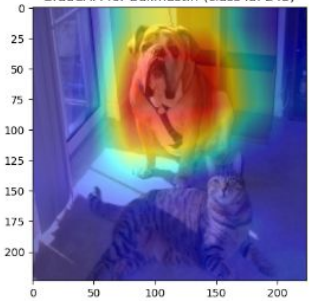


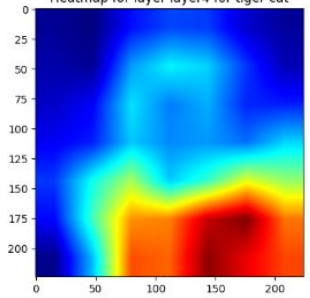


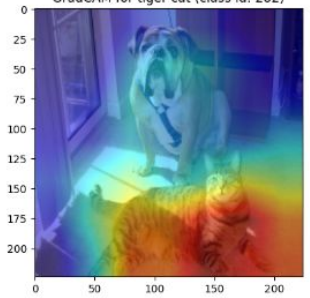


Layer-4







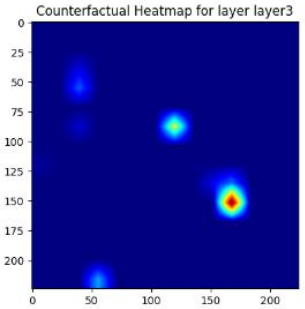


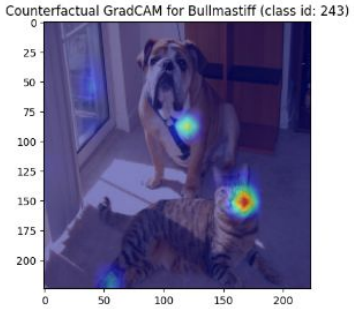
Resnet50 - Counterfactual Explanations

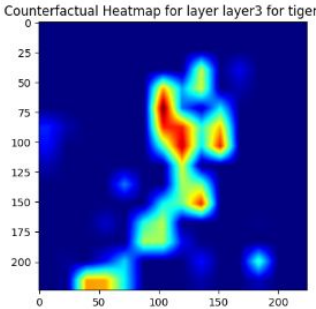
Bullmastiff

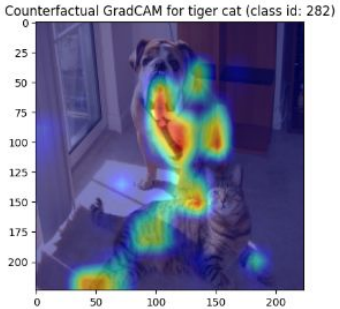
Tiger-Cat

Layer-3

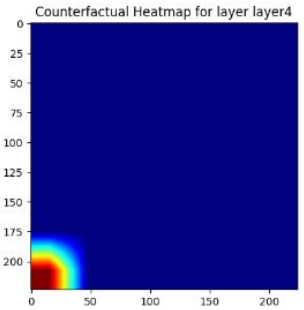


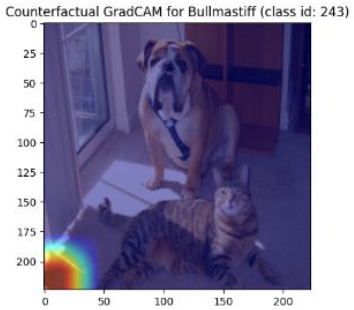


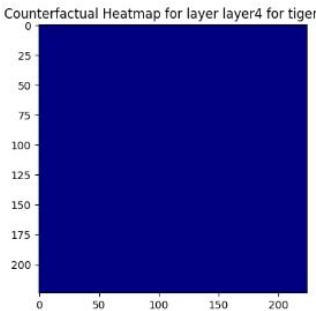


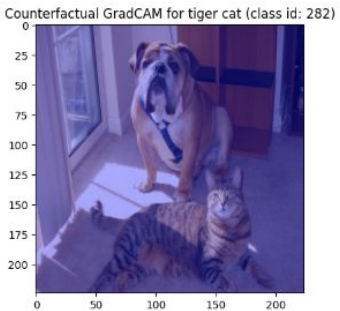


Layer-4







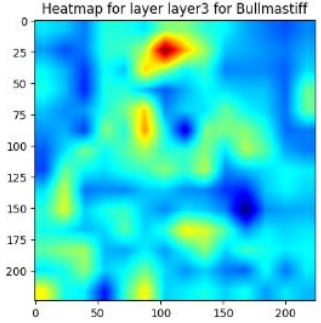


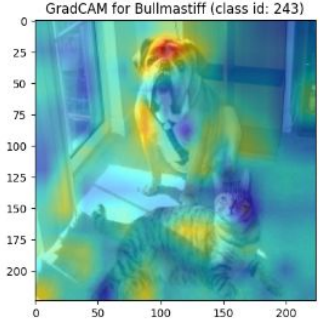
Resnet50 Without ReLu

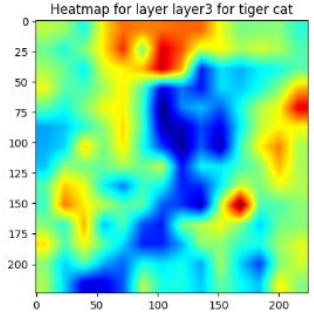
Bullmastiff

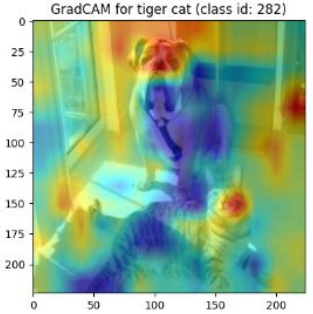
Tiger-Cat

Layer-3

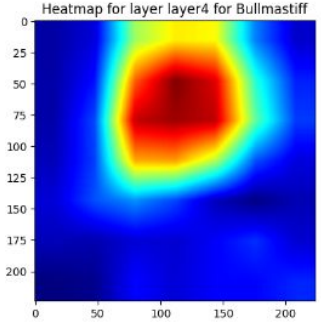


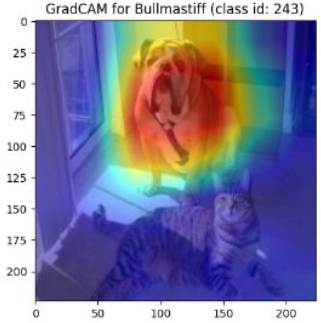


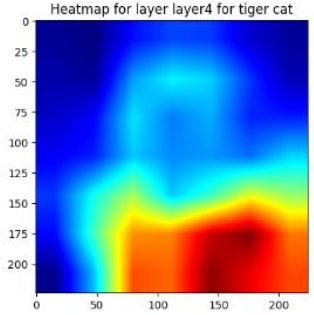


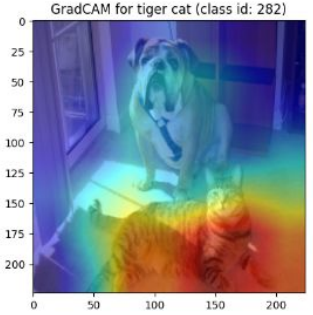


Layer-4









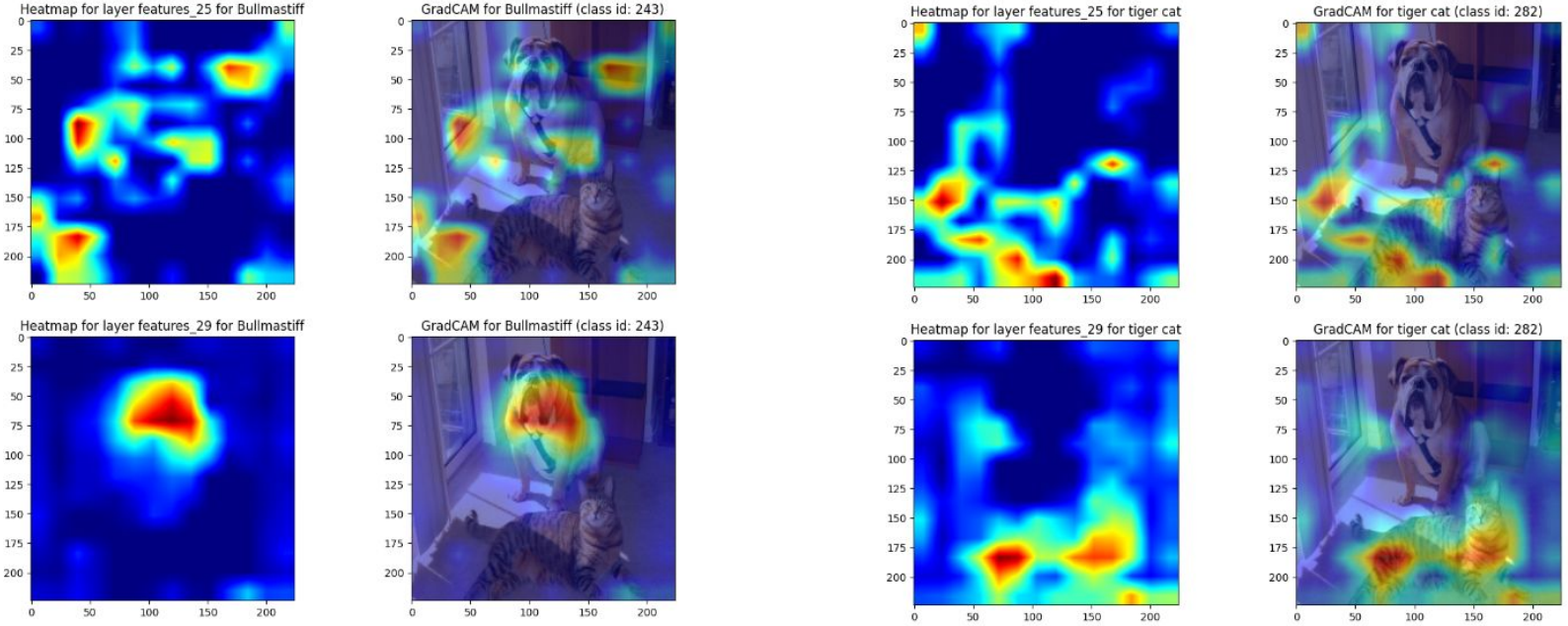
VGG-16 - features_25 vs features_29

Bullmastiff

Tiger-Cat

features_25

features_29



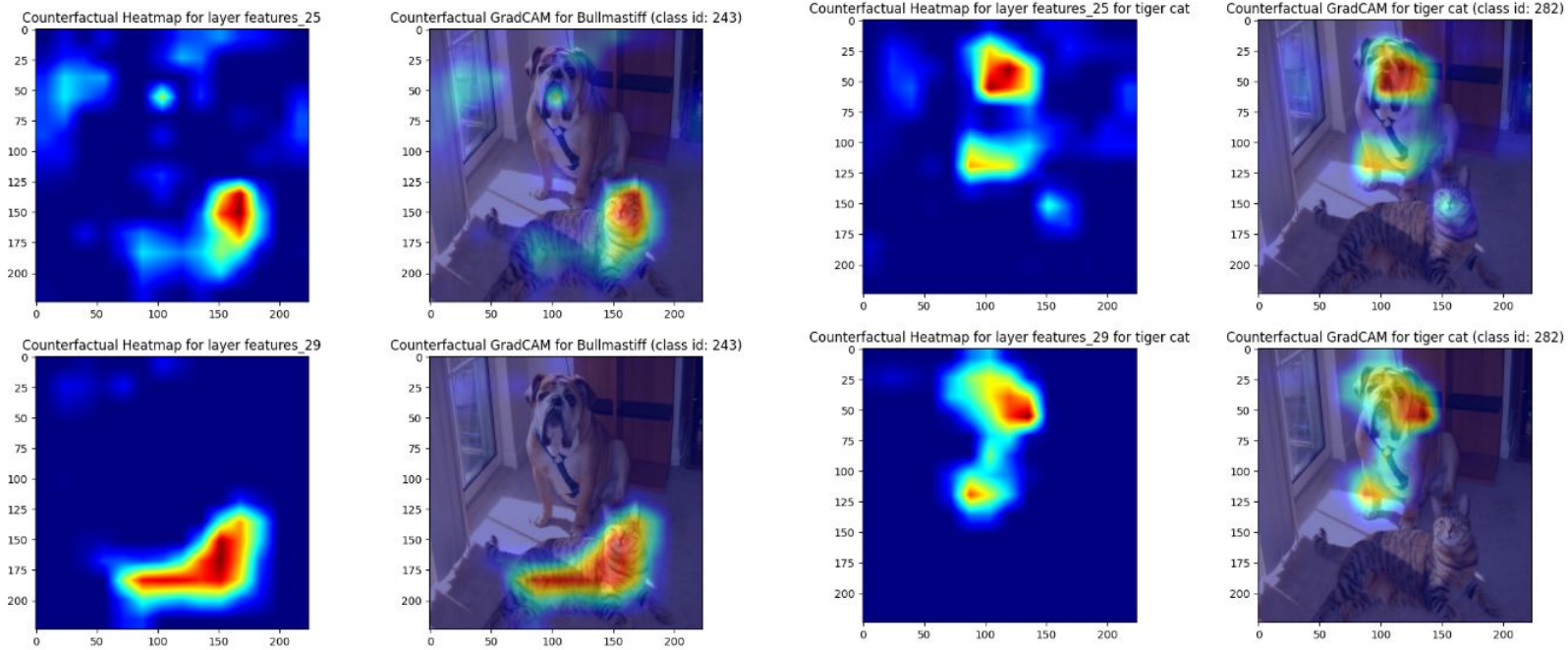
VGG-16 - Counterfactual Explanations

Bullmastiff

Tiger-Cat

features_25

features_29

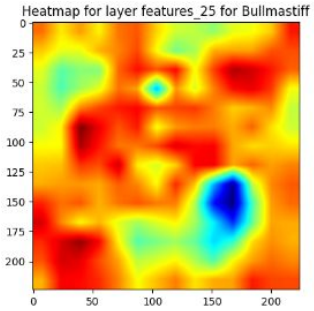


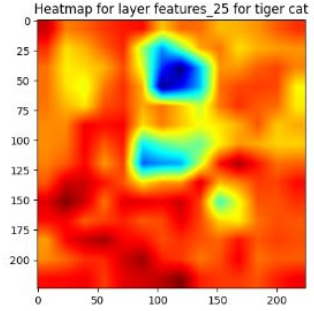
VGG-16 - Without ReLu

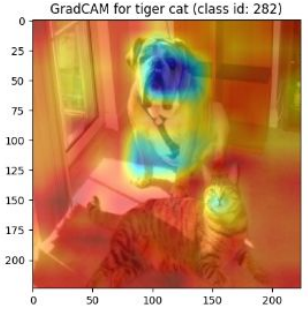
Bullmastiff

Tiger-Cat

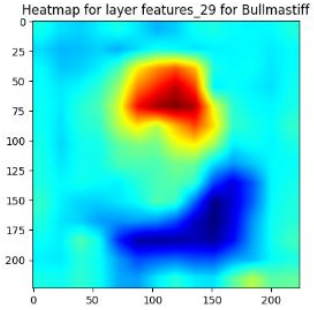
features_25

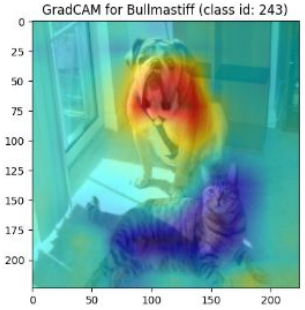


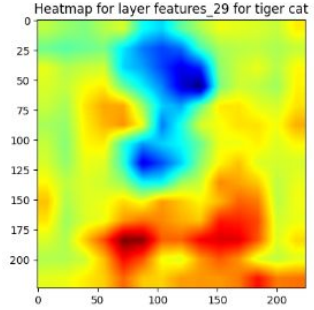


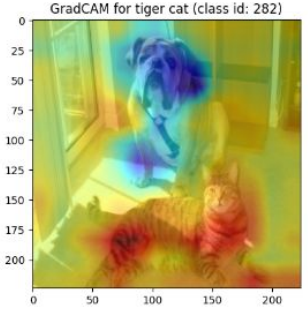


features_29









Grad-CAM against adversarial attack

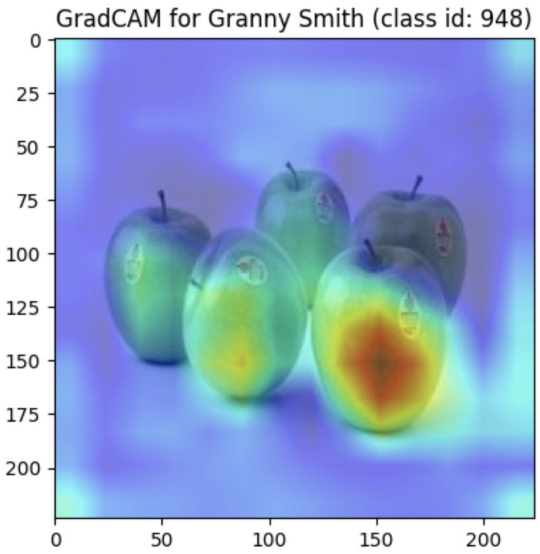
Granny Smith

+

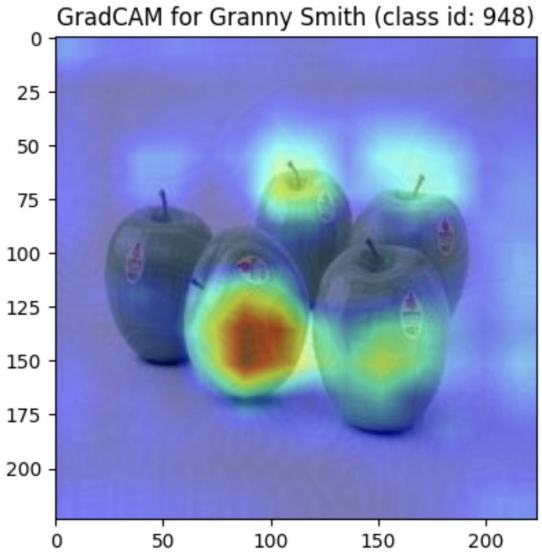
Noise

Ocarina (musical instrument)

Grad-CAM against adversarial attack



Original Image



Noise Added

Individual contribution

- Abhiram G P
  - Grad-CAM
- Hari Krishnan U M
  - Adversarial attack
- Vyshak P
  - Counterfactual Grad-CAM

THANK YOU