# White-Box Explanation Methods

## Context

In the previous lab we looked at some Black-box methods, namely LIME and RISE. The internal functioning of the model is kept hidden or is not accessible to the user. Only the input and the output of the system are accessible and such methods are termed as black box methods as they are model agnostic.

White-Box methods aim to exploit the available knowledge of the network itself to create a better understanding of the prediction and the internal logic of the network. There are different types of white box methods [Ayyar et.al.] but in our current exercise we will focus on the following:

- Grad-CAM [Selvarju et.al.]
- FEM [Fuad et.al.]

[Ayyar 2021] Ayyar, M.P., Benois-Pineau, J. and Zemmari, A., 2021. Review of white box methods for explanations of convolutional neural networks in image classification tasks. Journal of Electronic Imaging, 30(5), pp.050901-050901.

[Selvarju et.al.] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[Fuad et.al.] Fuad, K.A.A., Martin, P.E., Giot, R., Bourqui, R., Benois-Pineau, J. and Zemmari, A., 2020, November. Features understanding in 3d cnns for actions recognition in video. In 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6). IEEE.

## Grad-CAM

Gradient Class Activation Mapping (Grad-CAM) is a post-hoc explanation via visualization of class discriminative activations for a network. Grad-CAM leverages the structure of the CNN to produce a heat map of the pixels from the input image that contribute to the prediction of a particular class.

Grad-CAM relies on the obervation that deeper convolutional layers of a CNN act as high-level feature extractors. So the feature maps of the last convolution layer of the network would contain the structural spatial information of objects in the image.

The features maps from the last convolution layer are not used directly by the method as they would contain information regarding all the classes present in the dataset. Thus, the method calculates the gradient of the output score for a particular class with respect to the features of the convolution layer.

### Algorithm:

- Calculate the gradient of the output neuron (for the class of your choice - usually the GT label) w.r.t to the features of the last convolution layer
- Calculates the weights for each of these feature maps using the Global Average Pooling of the gradients calculated in the previous step

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

This formula is used to calculate the weights $\alpha$ for a particular class $c$, the last conv layer has $k$ feature maps named $A^1, A^2, ...., A^k$, $S_c$ is the output score and $Z = h \times w$ where $h$ and $w$ are the dimensions of each featrue map.

- Do a weighted sum of each of the feature maps $A^1, A^2, ...., A^k$ with the corresponding $\alpha_k^c$ and average them.
- Nullify the negative features and only retain the ones that have a positive influence on the output to get the relevance score map $R^c$.

$$R^c = ReLU(\sum_k \alpha_k^c A^k)$$

- To have correspondance to input upsample $R^c$ to the dimension of your input image using interpolation methods, normalize and visualize.

### Questions

1. Implement the algorithm for ResNet50 pre-trained model and use the African Elephant image from the pervious lab to visualize the results.
2. Run the method for another image from the same class and an image from the Black bear class.
3. Do the same with another pre-trained model of your choice.

## FEM

Feature based Explanation Method (FEM) is similar to Grad-CAM as it also uses the observation that deeper convolutional layers of the network act as high-level feature extractors.

FEM proposes that the final decision would be influenced by the strong features from $k$ maps in the last conv layer. FEM supposes that the $k$ maps have a Gaussian distribution and thus strong features from these maps would correspond to the **rare** features and uses K-sigma filtering to identify these *strong* and *rare* features.

GitHub link for FEM

You can use the above given link to get the implementation of FEM and use it directly.

### Questions

1. Similar to Grad-CAM use the FEM for ResNet-50 and visualize the output heatmap for African Elephant.
2. Use it with a different model and visualize the heatmaps for an image from African Elephant and the Black Bear class
3. Visually compare the results of FEM and Grad-CAM and comment on what you observe.
4. Comment on black-box vs white-box methods in terms of ease of use, parameter tuning, final results that you see etc.

## Hints

1. Gradient: tf.GradientTape for Tensorflow, register_hook for PyTorch
2. There are some implementations for Grad-CAM available online. Please ensure to not copy-paste the code but you can use them as reference for the gradient calculation
3. To choose which layer to use and to get the name of the layer from the model: model.summary for Tensorflow and you can directly print the model with PyTorch to get all the details of the model.

## BONUS

Looking at the CODE for FEM, write down the algorithm similar to the one provided for Grad-CAM (with your own mathematical notations of the formulas).

You can look at the original paper [Fuad et.al.] or a review paper [Ayyar 2021] for further understanding of the method.