Interpreting Deep Learning Models Using Explainable AI

Teodora Szasz

tszasz@uchicago.edu
Objectives

● Build a DL model to detect the type of scenery in an image
  ○ Can be used in satellite imagery to predict deforestation, wildfires, etc.

● Implement GradCAM visualization of the model
  ○ Explain how AI models think

● Use TensorFlow Serving for inference
https://www.microsoft.com/en-us/ai/ai-for-earth
Overview

Input Image

Heat Map

Grad-CAM

Forest
Train the model
HOW CNNs WORK

https://www.cs.ryerson.ca/~aharley/vis/conv/
RESNET

- Deep CNNs => vanishing gradient
RES-BLOCK

INPUT

CONVOLUTION BLOCK

IDENTITY BLOCK

IDENTITY BLOCK

OUTPUT

RESNET-18 MODEL

INPUT

Zero padding

Conv2D

BatchNorm, ReLU

MaxPool2D

RES-BLOCK

RES-BLOCK

RES-BLOCK

RES-BLOCK

RES-BLOCK

AveragePooling2D

Flatten()

Dense Layer, softmax

OUTPUT
https://towardsdatascience.com/understanding-and-coding-a-resnet-in-keras-446d7ff84d33
GRAD-CAM VISUALIZATION

INPUT

HEATMAP

OUTPUT

Original: SEA

Predicted: SEA

- Class-specific gradient information flowing in the final CNN
GRAD-CAM VISUALIZATION

\[ L_{Grad-CAM}^c = \text{ReLU} \left( \sum_k \alpha_{k}^c A^k \right) \]

1. Generate the class activation maps
   ○ Pass the image through the model to make the prediction
2. Use ARGMAX to find the index corresponding to the maximum value in the prediction => PREDICTION
3. Calculate the gradient that is used to arrive at that value with respect to feature map activation $A^k$ of the convolution layer
4. Use `tensorflow.GradientTape()` to get the values of the gradients
5. Enhance the filter values that resulted in generating a specific prediction by multiplying the gradients values with the activations in the last conv layer
6. Perform a weighted sum of the activation maps
7. Apply ReLU activation function => HEATMAP
8. Superimpose the HEATMAP to the ORIGINAL IMAGE to see the activation locations
To visualize the activation maps, we need to pass the image through the model to make predictions.

Use argmax to find the index corresponding to the maximum value in the prediction, which gives us the predicted class. Now, we take the predicted value for that class by the model.