Sanity Checks for Saliency Maps

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim

Review by Abhishek Soni, Roll No. 180100004

1. Motivation

This work is based on the saliency method which has emerged and turn out as a popular tool to highlight features in an input deemed relevant for the prediction of a learned model. This focuses and proposes a methodology in order to evaluate and explain a given method or model.

Because as the machine learning models' impact and complexity grow, it becomes challenging to explain the essential aspects of the learned model. A popular class of tools designed to highlight relevant features in an inputportant aspect of learned models.

The methods which fail these proposed tests are not adequate for tasks that are sensitive to either the data or the model, like finding outliers in the data or explaining the relationship between inputs and outputs which the model has learned.

2. Introduction

In this work, a methodology of randomization tests is implemented to evaluate the adequacy of explanation approaches.

This analysis is mainly instantiated on image data used for classification with neural networks. The saliency methods which are widely deployed are independent of both the data on which model was trained and the model parameters.

The edge detectors' outputs are very similar to the outputs of some of the saliency features methods. These edge detectors are vital features which are relevant to the model class prediction.

3. Randomization tests

3.1 The model parameter randomization

This test compares the output of a saliency method on a trained model with the output of the saliency method which was randomly initialized on untrained network having the same architecture. This was done in order to test the behavior of the output, if the saliency method depends on the learned parameters of the model, the expected outcome should differ substantially between the two cases.

3.2 The data randomization test

It compares a given saliency method applied to a model trained on a labeled data set with the methodology used to the same model architecture but trained on a copy of the dataset. We randomly permuted all labels. Insensitivity to the permuted labels, however, reveals that the method does not depend on the relationship between instances (e.g., images) and labels that exists in the original data tests can be thought of as sanity checks to perform before deploying a method in practice.

4. Methods

- Gradient explanation
- Gradient ⊙ Input
- Integrated Gradients (IG)
- Guided Backpropagation (GBP)
- Guided GradCAM
- SmoothGrad (SG)

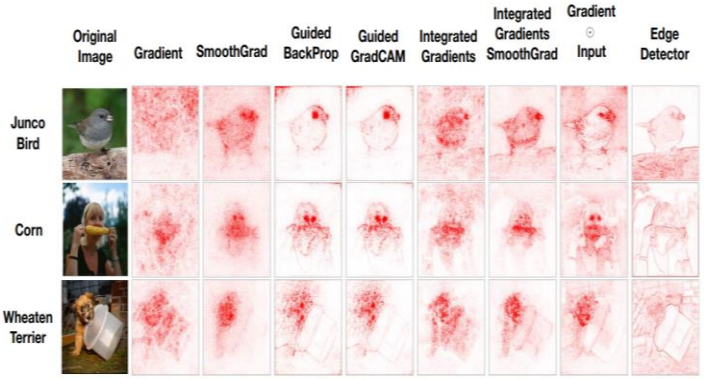


Figure 1 Saliency maps for common methods compared to an edge detector

4.1 Visualization & Similarity Metrics

Visualization: The saliency map is visualized 2 ways; one is absolute value (ABS) and in one it is left as it is. In ABS, absolute values are taken after normalization whereas in second different colors are used to show negative and positive importance.

Similarity Metrics: In order to do a quantitative comparison, these metrics are used. Spearmen Rank Correlation with and without absolute values, Structural similarity index (SSIM), and the Pearson correlation of the histogram of gradients (HOGs) derived from two maps

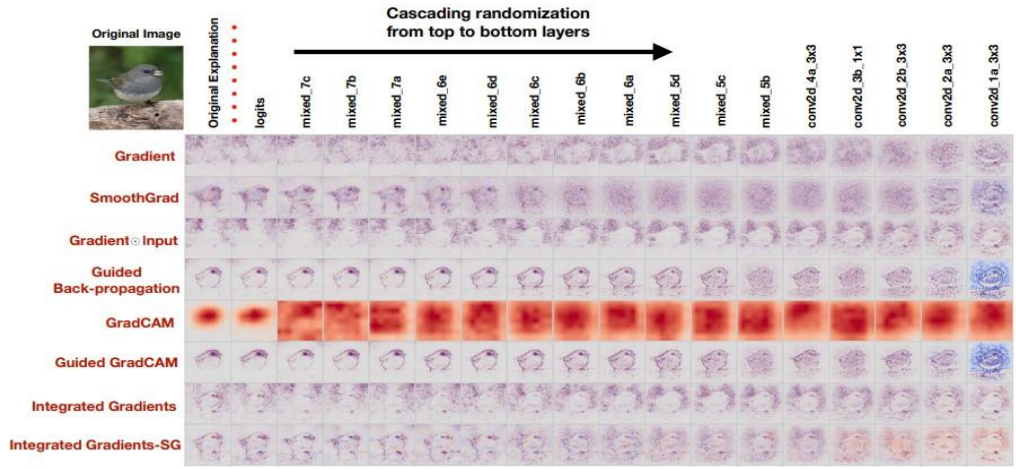


Figure 2 Cascading randomization on Inception v3 (ImageNet)

4.2 Model Parameter Randomization Test

Model parameters have a strong effect on the test performance of the model.

Cascading Randomization

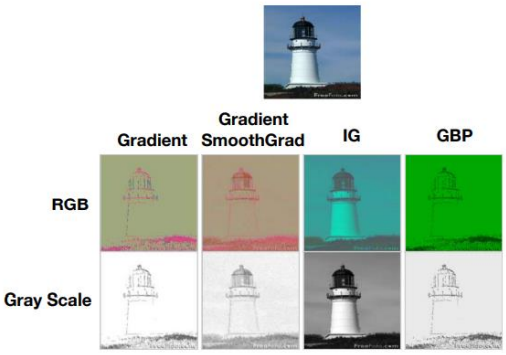
In this test, we randomize the weights of the model from the top layer to bottom layer successively. This is done in order to destroy the learned weights.

5. Discussion

In order to interpret our findings, we first discuss the influence of our model architecture on explanations derived from Neural networks. Then we approximate the element-wise product of the inputs and the gradient and finally we proceed via explaining the observed behavior of gradient explanation with an appeal to linear models.

5.1 The role of model architecture as a prior

The representations derived from any neural network are very much affected by its architecture such that the explanation which is not dependent on the model parameters can depend on the architecture.



5.2 Analysis of simple models

To understand the output of the saliency methods, it was tested on two simple models: a linear model and a 1-layer sum pooling convolutional network. The outcome of the linear model was a coefficient that measured the sensitivity of the model. In contrast, the output of the random convolution network was very much like that of the edge detector as per the visual artilect.

5.3 The case of edge detectors.

An edge detector is a feature detecting tool that highlights sharp transitions in any image. These edge detectors are typically untrained and are independent of any predictive model. They are solely a function of the given input image.

Edge detectors generate an image that is similar to outputs of some of the saliency features. These edge detectors produce features that appear to be very relevant for a model class prediction.

6. Conclusion

The goal of this paper was focused mainly on explaining the model, and various methods were studied. This gives researchers help in assessing the scope of model explanation methods.