# Interpretable Explanations of Black Boxes by Meaningful Perturbations

Summary and review by - Purvanshi Mehta

September 23, 2019

## 1 Summary

The paper proposes a method to highlight the areas of an image which contribute the most towards the prediction and also learns explanations for the same through meta predictors. The saliency visualizations of the image are obtained by the numerical optimization of the input image. The most relevant explanation to the prediction is chosen from a pre defined set of rules and are learnt via a regularized empirical risk minimization problem.

The variation in class score of the model are observed with the changing perturbations of the input image. The aim is to identify the regions of the image which have the highest contribution towards the prediction. The key idea is that the prediction of the model should change when such regions are deleted. A mask is defined for each pixel which takes the form of being a constant, noise or blur for the region. The final optimization goal is to find the minimum region which leads to highest drop in the score of the prediction.

To overcome the limitations of Neural Networks of giving unexpected outcomes by a small change in the input image, the mask is applied stochastically. To make the mask more representative of natural perturbations a total variation norm is applied as the regularizer and the image is upsampled from a low resolution image.

The authors claim that minimizing the most influential region helps in eliminating unnecessary information like a full truck in the case when a model recognizes a truck just by looking at its upper half. The visually distinct and abnormal masks generated are used in adversarial training to classify clean vs adversarial examples.

## 2 Review

The paper proposes a way to learn explanations through meta predictors. The explanation rules could be ranked/weighted to give a hierarchical structure to interpretability. A very naive example could be if the image is rotated and sclaed both. The ranking could also help in resolving ambiguities in the data set.

The explanation rules could also be augmented with some properties like house is an object and a lizard is a living creature. This method could help in generating textual explanations after rule learning through meta predictors.