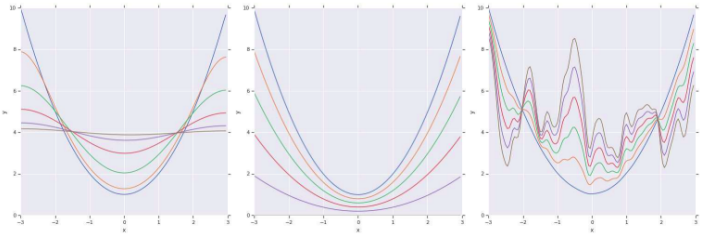
Attribution in Scale and Space

February 3, 2021

1 Overview of the paper

For neural networks, integrated gradient (IG) is one of the mostly used post-hoc feature importance methodologies to explain the model predictions. It considers a straight path from the baseline (generally black image) to the input image by increasing intensity at each step. IG represents the integral of gradients of model output with respect to images along the path. If a gradient result for a feature (pixel in visual tasks) is relatively bigger, it means that feature is important when making a prediction on a given input. IG is based on perturbations like other feature importance methods. While calculating the gradients, it uses different intensities of the input images. Generally, one assumes that these perturbations only removes 'information', then the resultant change in prediction can be interpreted as feature importance. However, if the perturbation creates information, then the resultant change in score is not because of a feature present in the input, and the result will be a misleading explanation. Thus, we have to be sure about the cause of the changes in the output scores. However, current state-of-the arts are not discussing this problem. This paper introduces a technique called **Blur Integrated Gradients (BIG)**. The explanations of BIG is with perturbations without artifacts. Unlike IG, it changes the blurriness level of image along the path. One can see the scaling from BIG and vanilla integrated gradients (IG) for the univariate function $x^2 + 1$. For example, IG with random baseline brings new local minima points which is not desirable. IG with black baseline is better but still it changes existing minima point. On the other hand, BIG only diminishes the minima, which is not a problem. Moreover, previous works are producing explanations based on pixels for visual tasks without localization in the frequency. BIG also contributes to this problem.



(a) Gaussian blur (b) Intensity scaling (black baseline) (c) Intensity scaling (random baseline)

Figure 1: Scale space for $x^2 + 1$ with the Gaussian blur, and intensity scaling

2 What is the central contribution of the paper?

It introduces a new technique which has some advantages over previous methods for the attribution importance for the deep networks in the visual tasks. First, it can differentiate at what scale a network recognizes an object unlike previous methods. Moreover, the perturbations, which are used to determine feature importance, are free of artifact. That means the results are cleaner and more robust. Lastly, it does not use 'baseline' parameter for calculating integrated gradients. It is good to eliminate baseline because it's selection is important for the quality of the results.

References

[1] Mukund Sundararajan, Ankur Taly, Qiqi Yan. "Axiomatic Attribution for Deep Networks", 2017.

[2] Shawn Xu, Subhashini Venugopalan, Mukund Sundararajan. "Attribution in Scale and Space.", 2020.