RISE: Randomized Input Sampling for Explanation of Black-box Models

Abraham Jose
ID :5068109, CAP6614
abraham@knights.ucf.edu

1 Summary

The author devices an approach, RISE to extract the important pixels in the image that provides accurate saliency and lowest area under curve(AUC) for a given neural network model and input. RISE is a black-box model, which means it does not requires to access the internals of the base model, intermediate feature maps or the gradients. Estimation of the importance of each pixels of input image for the model's prediction empirically by probing the model with randomly masked images of input image and obtaining corresponding output.

2 Strengths of the proposal

1. RISE accounts for the fact that the behaviour of an AI model may be different as they learn from cues in the background.
2. An intuitive and simple approach to address the problem of Explainable AI for deep neural network, irrespective of the architecture of the neural model and that can work on these black box AI models.
3. The work can be extended to image captioning and not limited to the saliency detection of input images.

3 Weaknesses of the proposal

1. The algorithm is computationally expensive and it relies on probability. For fine-feature's saliency detection we will need bigger mask, resulting in bigger mask space, 2^{W*H}.
2. Random masking can augment the features or introduce a new feature like data point, which may give wrong saliency maps and score.
3. RISE has feature that are not scale invariant to detect the saliency of regions/features in the input images

4 Results

The performance of the RISE is superior in the tests they carried out using deletion and insertion scores and AUC, even compared to white-box models including LIME and CAM. However, the pointing game accuracy is not upto the mark, where white-box models outperforms. Still considering the fact that good pointing accuracy may not correlate to the actual function of the given neural network.

5 Discussion

The cues in the background, that may be a function of the neural network model makes the problem of explainability a lot complex. There can be a lot of pairs of backgrounds and region under consideration.