Interpretable Explanations of Black Boxes by Meaningful Perturbation

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

1 Summary

The author provides us with a technique to explain a neural model through meaningful perturbation which offers a method to test the model. The explanation is limited to *what* the model does and *how* the model does and to investigate the internal mechanism that allows the model to achieve the properties. The proposed method is model agnostic and testable because they are grounded in explicit and interpretable image perturbation.

2 Strengths of the proposal

1. The model makes use of the mapping properties from input to output regression to explain the model, using many perturbation on input image to define the meaning and the behaviours of a model, which stays true for most of the time.
2. The proposed model is model-agnostic and testable as it is grounded in explicit and interpretable image perturbation. They backed the explanation with the theory using a simple robin classifier with certain rules that generates similar classification results for various perturbed images.
3. The methodologies they followed are quiet intuitive such as deletion and insertion game, their learning explanation based on perturbed images etc...

3 Weaknesses of the proposal

1. Since there will be infinite possible perturbations for an input image, its is technically not possible to test on all the perturbed images. Hence we can summarize that the explanation based on perturbation is limited to the available computation power that can be allocated for the process.
2. The explanation of a model depends in large part on the meaning of the varying input and not the network.Also, for a neural network, the change could be same regardless of the starting point(like a linear classifier), which creates similar saliency.

4 Results

They propose a general framework that can be used for the learning different kinds of explanations for any black-box learning algorithm. Also they were able to find the regions in the images that is most responsible for the classifier output through the meaningful perturbation to explain the model. Their model's explanation works really well in a series of cases where they build many hypothesis. The proposed technique is least prone adversarial attack.

5 Discussion

How the technique compares to RISE technique? Also how is it possible to define a model based on the output we receive for every different input.