Explain Black-Box Image Classifications Using Superpixel-Based Interpretation

Yi Wei[1], Ming-Ching Chang[1], Yiming Ying[1], Ser Nam Lim[2], Siwei Lyu[1]

1 University at Albany, State University of New York, NY, USA

2 GE Global Research Center, Niskayuna, NY, USA

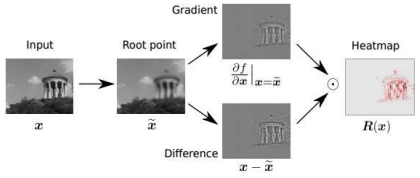
Introduction

Motivation

- DNNs have reached human-level performance on several real-world tasks, however lack of interpretability.
- The complicated AI DNN models make them less trustworthy for making critical decisions in tasks including disease treatment and autonomous driving.
- DNNs are easily fooled by adding gradient ascend noise.
- It is significant if human can understand the classifier's decision in a straightforward ways, particularly, in a model-agnostic manner.

Related work

- **Self-explainable model**
  - Decision tree, linear model, etc.
- **White-box interpretation**
  - Deep feature visualization
  - Gradient backward distribute
  - Parameters of model are known



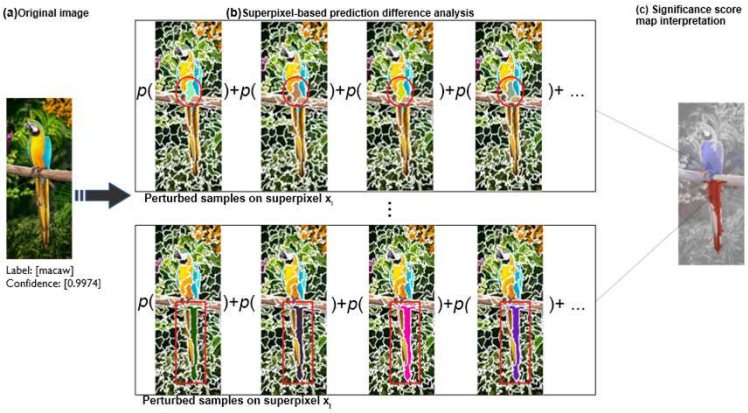
(Montavon, Grégoire, *et al.* "Explaining nonlinear classification decisions with deep taylor decomposition." Pattern Recognition 65 (2017).

- **Black-box interpretation**
  - Do not require the knowledge of model
  - Predict the model behavior based on only inputs and outputs

Contribution

- Model-agnostic (black-box) approach in providing a significance score visualization that intuitively illustrates how each image component contributes to the decision.
- Superpixel-based inference improves the consistency and computational efficiency of the black-box interpretation.
- Superpixel formulation enables users to quickly specify a ROI for interactive interpretation.
- Example-specific interpretation brings additional justifications for developers and users to evaluate:
  (1) how the classifier is trustworthy in terms of how it responses in specific test cases,
  (2) how robustness of the classifier can be improved by adding specific training samples that are reflected from the interpretations.

Approach



(a) The input image with classification results. (b) Prediction difference analysis (PDA) is performed on each superpixel of the input image. We illustrate the process of inferencing the significance score for superpixel x_i (feet) and x_j (tail) with their perturbed samples using marginalization. (c) The resulting map visualizes the supportive (red) vs. unsupportive (blue) likelihood (significance score) of the components toward the classification.

Superpixel-based prediction difference analysis (PDA)

- Image X represented as a set of segments, $X = \{x_i\}$, where x_i is the index of the superpixels/pixels.
- Significance δ_i of a superpixel $\boldsymbol{x_i}$ toward the black-box classification is estimated by:

$$\delta_i(c|X) = P(c|X) - P(c|X\backslash_i), \quad \text{where } X\backslash_i = \{X - x_i\}.$$

which calculates the difference between the probability of an image for a given class with and without superpixel x_i

- $P(c|X\backslash_i)$ is approximated by marginalizing x_i out with multiple perturbed images.

$$P(c|X\backslash_i) = \sum_{s=1}^{m_i} P(x_i = v_s|X\backslash_i) P(c|X\backslash_i, x_i = v_s)$$
$$\approx \sum_{s=1}^{m_i} P(x_i = v_s|X) P(c|X\backslash_i, x_i = v_s)$$

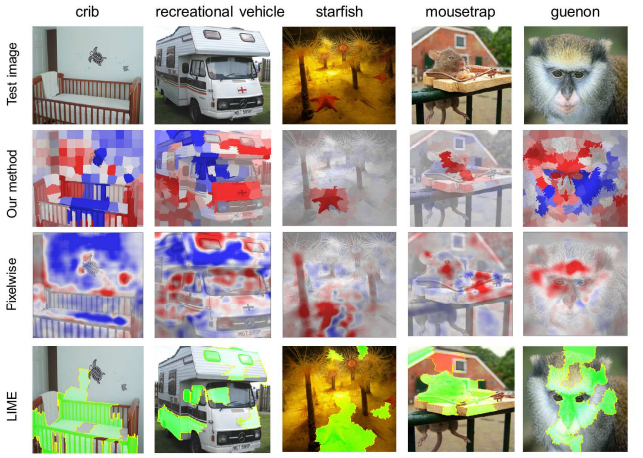
- **Weight of evidence**
  A refined metric for significance score commonly used in information theory and logistic regression

$$w_i(c|X) = log_2[odds(c|X)] - log_2[odds(c|X\backslash_i)],$$
$$odds(z) = \frac{p(z)}{1 - p(z)}$$

- **Factors that impacts the estimation accuracy**
  - Sampling numbers m_i
  - Likelihood distribution(color distribution for image) $P(x_i = v_s|X)$ - color histogram estimation

Experimental Results

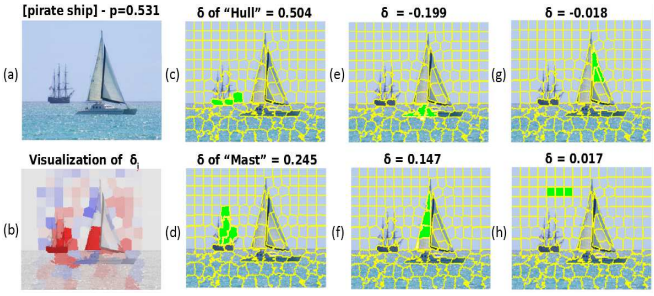
Interpretation results comparison on Resnet101 trained on ImageNet



(red: pro, blue: against)

- Visually more consistent than pixelwise PDA and LIME.
- Time-efficient which takes 3 mins to get the interpretation compared to 30+ mins of pixelwise PDA.
- Robust to number of superpixel segments, number of marginalization samples and the number of color histograms (fineness of color histogram estimation).

Interactive interpretation



Interactively select regions of interest (ROI) in superpixels for a quick interrogation regarding their significance scores toward classification.