Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning models

Yongchan, Choi

2018.11.07

Contents

Why do we need explainable AI?

- Verification of the system
- Improvement of the system
- Learning from the system
- Compliance to legislation

Sensitivity Analysis

- (Assume) The most relevant input features are those to which the output is most sensitive.
- $R_i = \| \frac{\partial}{\partial x_i} f(\mathbf{x}) \|$
- Sensitivity analysis does not explain the function value $f(\mathbf{x})$ itself, but rather a variation of it.

Sensitivity Analysis

- The label of above figure is "rooster"
- The yellow flowers occlude part of the rooster.
- Changing the pixels of the flowers in a specific way would reconstruct the occluded part of the rooster, which most probablty would also increase the classification score.

Layer-Wise Relevance Propagation

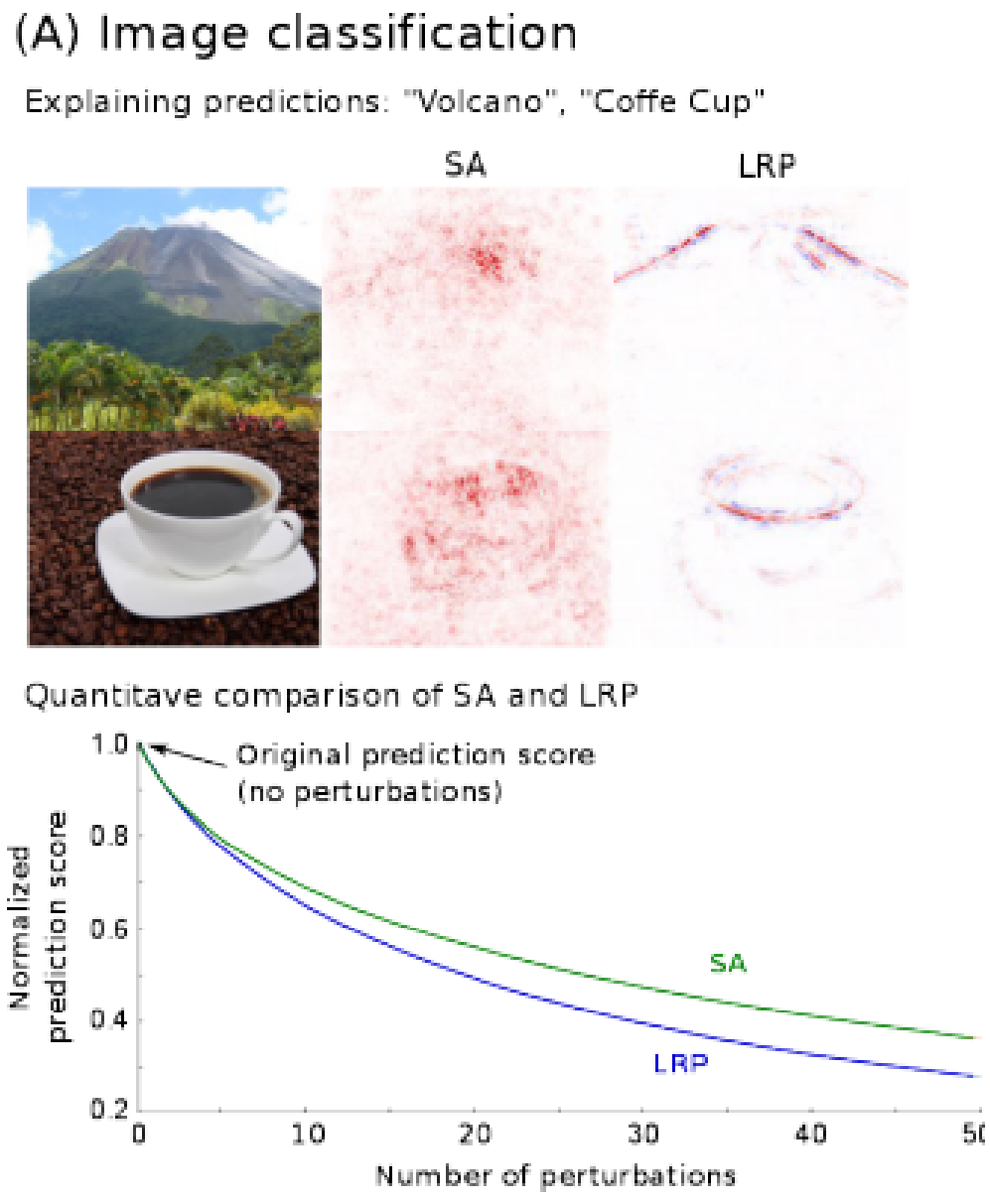
Let x_j be the neuron activations at layer l, $R_k$be the relevance score associated to the neurons at layer l+1. w_{jk} be the weight connecting neuron j to neuron k.

- $R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \epsilon}$
- ϵ is a small stablization term
- When $\epsilon = 0$, $f(\mathbf{x}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{R_i} = \sum_{\mathbf{j=1}}^{\mathbf{n_l}} \mathbf{R_j} = \cdots = \sum_{\mathbf{k=1}}^{\mathbf{n_L}} \mathbf{R_k}$

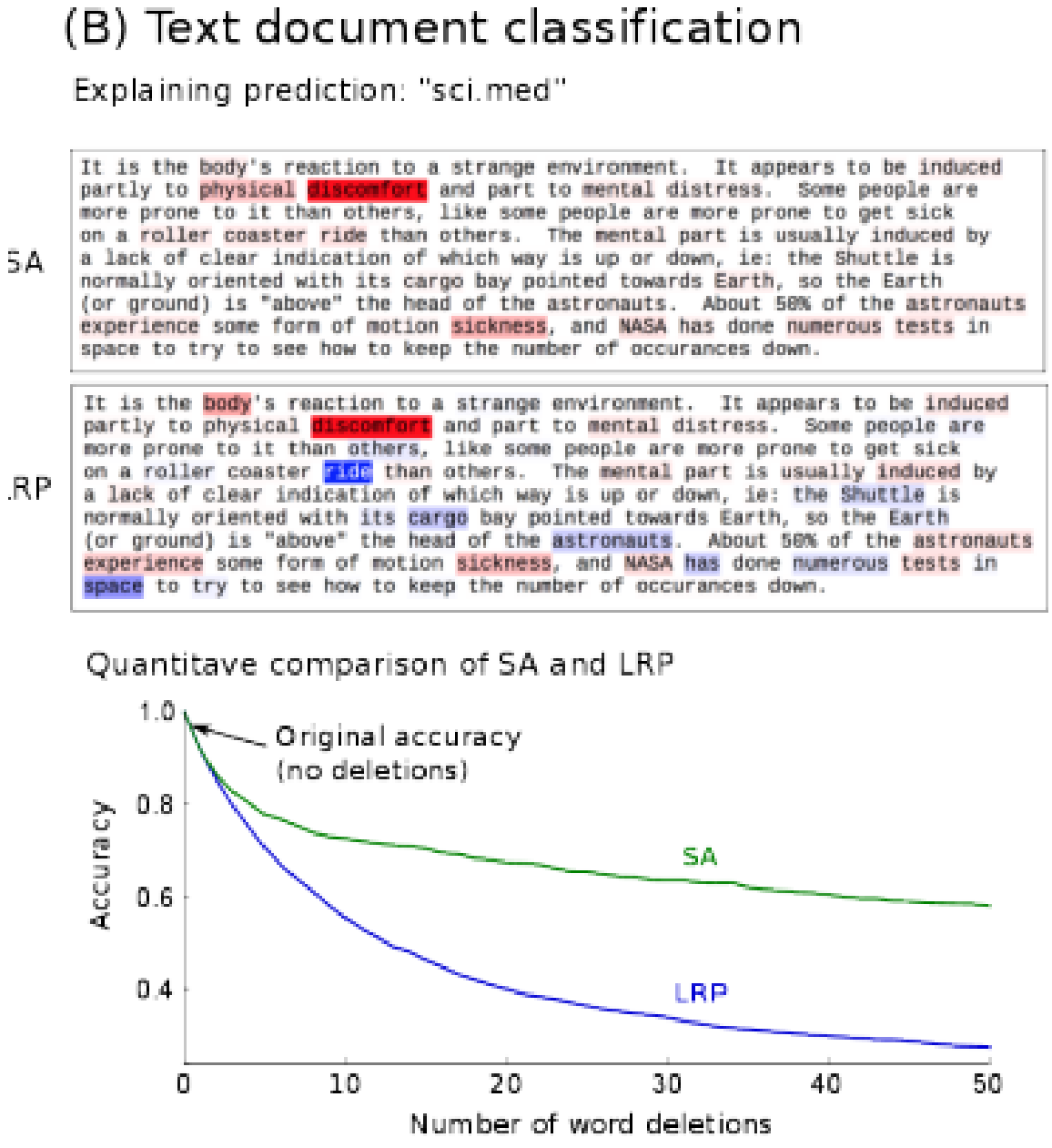
Evaluating the quality of explanations

1. Calculate the score(SA, LRP)
2. Select top k-th input variables
3. Give them random noise and check prediction score

Sensitivity Analysis



Sensitivity Analysis



Sensitivity Analysis