

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.
- ▶ $R_i = \left\| \frac{\partial}{\partial x_i} f(\mathbf{x}) \right\|$
- ▶ 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 probably 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 stabilization term

- ▶ When $\epsilon = 0$, $f(\mathbf{x}) = \sum_{i=1}^n \mathbf{R}_i = \sum_{j=1}^{n_1} \mathbf{R}_j = \dots = \sum_{k=1}^{n_L} \mathbf{R}_k$

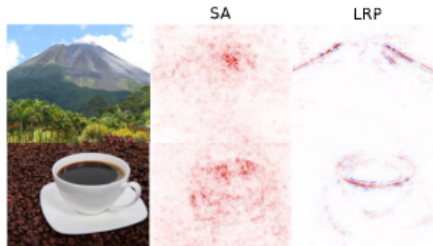
Evaluating the quality of explanations

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

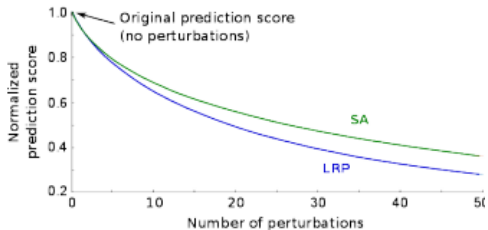
Sensitivity Analysis

(A) Image classification

Explaining predictions: "Volcano", "Coffe Cup"



Quantitative comparison of SA and LRP



Sensitivity Analysis

(B) Text document classification

Explaining prediction: "sci.med"

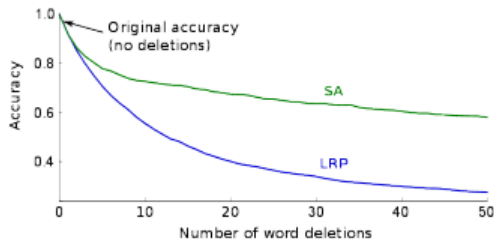
SA

It is the body's reaction to a strange environment. It appears to be induced partly to physical **discomfort** and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion **sickness**, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

.RP

It is the **body's** reaction to a strange environment. It appears to be induced partly to physical **discomfort** and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster **ride** than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its **cargo** bay pointed towards Earth, so the Earth (or ground) is "above" the head of the **astronauts**. About 50% of the astronauts **experience** some form of motion **sickness**, and NASA **has** done numerous tests in **space** to try to see how to keep the number of occurrences down.

Quantitative comparison of SA and LRP



Sensitivity Analysis

(C) Human action recognition in videos

Explaining prediction: "sit-up"

