Evaluations of features attribution methods

Context

During the previous lessons: - you learned to compute a saliency map from gaze fixation points that can be considered as a ground truth - a saliency map to explain a classifier on a given image - you also taught different ways to visualize them.

In this lesson, you will learn to evaluate the generated explanations either using a ground truth or not.

For this lesson we expect analysis and discussion of the results obtained.

Evaluation without a ground truth

`Insertion` and `Deletion` are two algorithms proposed by the authors of `RISE`.

Deletion

The main idea of `Deletion` is to progressively remove pixels from the explained image, based on their importance provided by the saliency map, and compute the prediction score of a selected class for each altered image. The deletion score corresponds to the area under the curve that represents this score depending on the percentage of deleted pixels. Deleted pixels are replaced by black color.

Algorithm 1

procedure DELETION
  Input: black box f, image I, importance map S, number of pixels N removed per step
  Output: deletion score d
  $n \leftarrow 0$
  $h_n \leftarrow f(I)$
  while I has non-zero pixels **do**
    According to S, set next N pixels in I to 0
    $n \leftarrow n+1$
    $h_n \leftarrow f(I)$
  $d \leftarrow$ `AreaUnderCurve`$(h_i \text{ vs. } i/n, \quad \forall i = 0, \dots n)$
  return d

Implement it, and test it on explanations you have been able to generate during previous lessons.

The main idea of `Insertion` is to progressively add pixels to the explained image, based on their importance provided by the saliency map, and compute the prediction score of a selected class for each altered image. The insertion score corresponds to the area under the curve that represents this score depending on the percentage of added pixels. Missing pixels are replaced by a blurred version of the image.

Algorithm 2

procedure INSERTION
  Input: black box f, image I, importance map S, number of pixels N removed per step
  Output: insertion score d
  $n \leftarrow 0$
  $I' \leftarrow$ `Blur`(I)
  $h_n \leftarrow f(I)$
  while $I \neq I'$ **do**
    According to S, set next N pixels in I' to corresponding pixels in I
    $n \leftarrow n+1$
    $h_n \leftarrow f(I')$
  $d \leftarrow$ `AreaUnderCurve`$(h_i \text{ vs. } i/n, \quad \forall i = 0, \dots n)$
  return d

Implement it, and test it on explanations you have been able to generate during previous lessons.

Which conclusion can you draw ?

Evaluation with a ground truth

Saliency maps generated from gaze fixation points can be considered as being ground truths. We expect saliency maps generated from explanation methods to be as close as possible to them.

Two metrics are commonly used in the literature: the Pearson Correlation Coefficent (PCC) and the similarity metrics (SIM). Consider A to be the ground truth saliency map and B the map generated by the method to evaluate.

$$PCC\left(A,B\right)=\frac{\sum_{x}^{W}\sum_{y}^{H}\left(A(x,y)-\overline{A}\right)\left(B(x,y)-\overline{B}\right)}{\sqrt{\sum_{x}^{W}\sum_{y}^{H}\left(A(x,y)-\overline{A}\right)^{2}}\sqrt{\sum_{x}^{W}\sum_{y}^{H}\left(B(x,y)-\overline{B}\right)^{2}}}$$

$$SIM\left(A,B\right)=\sum_{x}^{W}\sum_{y}^{H}min\left(A(x,y),B(x,y)\right)$$

Use these metrics to analyze the data generated during the previous labs. Which conclusion can you draw ?