Quiz 7 Solution (Interpretable Learning)

Solutions

1. Given an image that originally belongs to class 1, we identify which pixels to erase to convert the image to the target class 2. For each interpretation algorithm, we calculate a new score map $S_a = S_a^1 - S_a^2$ for each image and select top 20 percent pixels with highest scores and erase them. Then we obtain the classification accuracy of the predictive model on data with new features for class 1. The least accuracy belongs to the interpretation algorithm which gives highest scores to the most related pixels.

2. (a) Perturbation and forward propagation based methods change the input (omitting a feature, ...) and then compute the model output in a forward direction to asses the effect of the input variation on the model output. Its main disadvantage is that for each variation model should be run and it is very time consuming. Backpropagation based methods perform this task by computing the gradient of the output based on the input in a backward direction. It is limited to models in which the output is differentiable based on input. It also just considers a small neighborhood around the input. Read more about these techniques in section 2.1 and 2.2 of this paper `https://arxiv.org/pdf/1704.02685.pdf` .

   (b) For example in a situation in which $i_1 = 1$ and $i_2 = 1$ a backpropagation method calculates the gradient of the output based on the both input variables zero. Also in a forward based method a small change in variables do not change the output. So both of them report that the output is completely unrelated to both of variables which is obviously wrong.

   (c) Considering the target node z,

$$
\begin{aligned}
\sum_i C_{\Delta x_i \Delta z} &= \sum_i \Delta x_i m_{\Delta x_i \Delta z} \text{ (By definition of } m_{\Delta x_i \Delta z}) \\
&= \sum_i \Delta x_i \sum_j m_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta z} \text{ (By the chain rule)} \\
&= \sum_i \Delta x_i \sum_j \frac{C_{\Delta x_i \Delta y_j}}{\Delta x_i} m_{\Delta y_j \Delta z} \text{ (By definition of } m_{\Delta x_i \Delta y_j}) \\
&= \sum_i \sum_j C_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta z} \\
&= \sum_j \sum_i C_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta z} \text{ (Flipping the order of summation)} \\
&= \sum_j \Delta y_j m_{\Delta y_j \Delta z} \text{ (By summation-to-delta of } C_{\Delta x_i \Delta y_j}) \\
&= \sum_j \Delta y_j \frac{C_{\Delta y_j \Delta z}}{\Delta y_j} \text{ (By definition of } m_{\Delta y_j \Delta z}) \\
&= \sum_j C_{\Delta y_j \Delta z} = \Delta z \text{ (By summation-to-delta of } C_{\Delta y_j \Delta z})
\end{aligned}
$$

(d) In the case of (b), we can set the reference value of the target node y and both i_1 and i_2 zero. Now you can check and see that the problem is solved.