Iterative and Adaptive Sampling with Spatial Attention for Black-Box Model Explanations

Kitware Inc.
1712 Route 9 Suite 300, Clifton Park, NY, USA 12065
{bhavan.vasu, chengjiang.long}@kitware.com

**Background & Motivation**

**Observation:** Black-box model explanations are generated by sampling all image regions equally to produce saliency maps. This can be computationally expensive and result in coarse saliency maps due to high variance in image size.

**User:** “Are the legs important?”

**Intuition:** We hypothesize that sampling around important regions iteratively will result in finer saliency maps when done in a sequential manner.

**Contribution:** We propose a novel iterative and adaptive sampling method that samples around relevant regions with the help of our Long Range Spatial Attention module. We also re-visit methods used to evaluate explanations and propose a new evaluation scheme.

**Proposed Approach**

\[
S(I, f, \lambda) = \sum_{M} f(I \otimes M)P[M = m, M(\lambda) = 1]
\]

where

\[
P[M = m, M(\lambda) = 1] = \begin{cases} 0, & \text{if} \ m(\lambda) = 0 \\ \frac{1}{M \cdot N}, & \text{if} \ m(\lambda) = 1 \\ \end{cases}
\]

**Evaluation Metrics:** Deletion, Insertion, F-1, IoU and Pointing Game.

**Competing Algorithms**


**Dataset and Metrics**

- **MSCOCO dataset:** ~80 object categories with ~200k images.
- **Evaluation metrics:** Deletion, Insertion, F-1, IoU and Pointing Game.

**Experiments**

**Comparison with the state-of-the-art approaches**

**Qualitative comparison across methods**

**Conclusion & Future Work**

- We propose a novel iterative and adaptive sampling method that uses a parameter-free long-range spatial attention for generating explanations for black-box models.
- Future work involves coming up with a universal evaluation protocol to evaluate different kinds of explanations and feed explanations agreed upon by users back into the model as ‘advice’.

**Source Code**