Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

Seungmin Lee (profile2697@gmail.com; 2013-11420), Dept. of Computer Science and Engineering, Seoul National University

1. Motivation

Despite the tremendous advances in deep learning, there exists only a limited number of research on effective interpreting methods. The complexity of models is the main reason for the limitation. Additionally, because most of the models are trained on low-level features such as images rather than human-interpretable concepts, devising a useful interpreting tool that uses high-level concepts is difficult. This paper tries to solve these problems by suggesting a new interpreting method called *Testing with Concept Activation Vectors (TCAV)*.

2. Goals of the Proposed Method

The proposed method aims to achieve the following goals: Convenience, Portability, Retraining-Free, Global-Quantification. The convenience means the proposed method does not require any expert knowledge about Machine Learning. The portability represents that the technique can be applied to an arbitrary concept. The saliency map may not be portable because it only visualizes the classification cue. The Retraining -Free means the technique does not require changing the weights of a target model that we try to interpret. Lastly, the global-quantification signifies that the method can interpret the whole sets of images of interest. The Class Activation Map (CAM) does not have the global-quantification because it should repeat the entire process to get the results of other images.

3. Notations

4. Proposed Method

5. Results

6. Personal Memo