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)*. The proposed method aims to achieve the following goals: Convenience, Portability, Retraining-Free, Global-Quantification. Global-Quantification means TCAV provides explanations that can be applied to the entire images of interests without repeating the proposed method.

2. Notations

The interpretability can be formulated as a function $g : E_m \rightarrow E_h$, where E_m is a vector space that represents the state of a target model that we try to interpret, and E_h is a vector space that represents human-interpretable concepts. Assume e_m is basis vectors of E_m, and e_h is basis vectors of E_h. Then, e_h can be understood as unknown human concepts.

Assume there is a model that inputs an image $\mathbf{x} \in \mathbb{R}^n$, and the model consists of an encoder $f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a decoder $h_{l,k} : \mathbb{R}^m \rightarrow \mathbb{R}$ mapping features to k-th class.

3. Proposed Method

4. Results

5. Personal Memo