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 *Concept Activation Vectors (CAV)*.

CAV is a vector that represents a high-level concept. To calculate CAV, the first thing to do is defining a high-level concept using images. For example, the pictures of zebras can reveal the concept of "striped." After that, train a linear classifier that discriminates the concept's images and random images on the target model's intermediate feature space. Then, the orthogonal vector of the decision boundary is the CAV.

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

In this section, the details of the proposed method will be explained.

3.1. Define a high-level concept of interest

As mentioned before, the concept of interest is defined using a set of images. Therefore, the user can specify a new concept without knowledge of ML. Formally, a concept of interest C (e.g., dotted) is revealed by $\mathbf{x} \in P_C$ (e.g., photos of dalmatian) and $\mathbf{x} \in N$ (e.g., randomly selected images such as zebra).

3.2. Calculate CAVs

After defining a new concept, train a linear classifier discriminating features of P_C and N. Formally, the classifier classifies $\{f_l(\mathbf{x}) : \mathbf{x} \in P_C\}$ and $\{f_l(\mathbf{x}) : \mathbf{x} \in N\}$. Then, the normal vector of the decision boundary ${\mathbf{v}^l}_C$ becomes the CAV of the concept C.

4. Results

5. Personal Memo