# 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

Assume there is a model that inputs an image $\mathbf{x} \in \mathbb{R}^n$, and the model consists of an encoder $f_l : \mathbb{R}^n \to \mathbb{R}^m$ and a decoder $h_{l,k} : \mathbb{R}^m \to \mathbb{R}$ 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 ${\mathbf{v}^l}_C \in \mathbb{R}^m$ becomes the CAV of the concept $C$.

### 3.3. Interprete using the CAVs: Conceptual Sensitivity

Using CAVs, the model's sensitivity of concepts can be measured. For this, this paper adopts directional derivatives. More specifically, the sensitivity of class $k$ to the concept $C$ at layer $l$ ($S_{C,k,l}$) is defined using the directional derivative as follows:

$$\begin{aligned} S_{C,k,l}(\mathbf{x}) &= \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(\mathbf{x}) + \epsilon {\mathbf{v}^l}_C) - h_{l,k}(f_l(\mathbf{x}))}{\epsilon} \\ &= \nabla h_{l,k}(f_l(\mathbf{x})) \cdot {\mathbf{v}^l}_C \end{aligned} \quad (1)$$

This equation represents how fastly confidence will increase if the concept is added. Thus, it can be interpreted as the conceptual sensitivity.

### 3.4. Testing with CAVs (TCAV)

The method for measuring conceptual sensitivity can be extended to the entire images of a class. This paper defines this measurement called *Testing with CAVs (TCAV)* as follows:

$$TCAV_{C,k,l} = \frac{|\{\mathbf{x} : \mathbf{x} \in X_k \text{ and } S_{C,k,l} > 0\}|}{|X_k|} \quad (2)$$

where $X_k$ is the entire images of a class $k$. TCAVs can be used for measuring global conceptual sensitivities of the model rather than a single example.

## 4. Results

This paper has conducted experiments such as changing layer $l$, the alignments between the method and human intuition (sort images using various concepts' CAVs).

## 5. Personal Memo

I think the proposed method is simple and easy to use. Defining a concept using images seems a practical idea. However, I wonder whether a linear classifier can always be trainable. If the model's task and a concept are not aligned well, the linear classifier can not be well trained.