A Comparative Study of LIME and TCAV for Model Interpretability

Wenfeitian Shi
Central South University

1 Introduction

Most of machine learning models are considered black boxes, as their internal structures are difficult for humans to understand due to their opacity. Therefore, methods that improve model interpretability are necessary. LIME and TCAV are examples of such methods. Both methods provide human-understandable explanations for the previously uninterpretable predictive behavior of models, but they differ in their specific presentation. This paper will compare two papers to illustrate the similarities and differences between LIME and TCAV.

2 Core ideas of the methods

2.1 LIME

LIME (Local Interpretable Model-agnostic Explanations) is a local, model-agnostic explanation method. The core idea of LIME is to approximate the behavior of a complex model in a local region using a simple and interpretable model. Specifically, LIME assumes that even if the overall model is nonlinear, it can still be approximated by a simple form(such as linear model, decision trees or falling rule list), within a sufficiently small local neighborhood. By analyzing the weights of each input feature in the surrogate model, LIME translates the original model's prediction for a given sample into a human-friendly explanation of feature importance, thus achieving local interpretability for individual prediction decisions.

Implementation of LIME:

- Let $f : \mathcal{X} \to \mathbb{R}$ denote the trained black-box model to be explained.
- Given an input instance $x \in \mathcal{X}$, define an interpretable representation $x' \in \{0,1\}^d$, where each dimension corresponds to the presence or absence of an interpretable feature. (eg. binary vector for text, super-pixel for image)
- Generate a set of perturbed samples $\mathcal{Z}$ by randomly modifying the interpretable features of x', and map each $z' \in \mathcal{Z}$ back to the original input space to obtain z.
- Query the black-box model to obtain predictions $f(z)$ for all perturbed samples.
- Define a locality-aware weighting function

$$\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right),$$

which assigns higher weights to samples closer to x in the original input space.

- Specify a family of interpretable models G (here $G = \{g|g(z) = w^\top z, w \in \mathbb{R}^d\}$), and a complexity penalty

$$\Omega(g) = \infty \mathbb{1}\left[\|w_g\|_0 > K\right],$$

which enforces sparsity and interpretability.

- Learn the surrogate model by solving the following optimization problem:

$$\xi(x) = \arg\min_{g \in G} \left\{\mathcal{L}(f, g, \pi_x) + \Omega(g)\right\},$$

where the weighted loss is defined as

$$\mathcal{L}(f, g, \pi_x) = \sum_{z \in \mathcal{Z}} \pi_x(z)\big(f(z) - g(z')\big)^2.$$

- Interpret the prediction of f at x using the learned surrogate model g, where the model coefficients indicate the importance of interpretable features.

2.2 TCAV

The core idea of TCAV (Testing with Concept Activation Vectors) is to explain the uninterpretable internal representations of a trained deep neural network by introducing human-understandable concepts. In this method, a human concept is represented as a directional vector in the l-th layer of the model, called a Concept Activation Vector (CAV). TCAV learns the CAV by comparing the activation differences between concept samples and control samples in this representation space, and uses the gradient information of the model output with respect to this direction to quantify the influence of a specific concept on the prediction result of a certain class. Through statistical analysis across multiple samples, TCAV evaluates the model's decision-making mechanism at the conceptual level, thereby achieving interpretability analysis that goes beyond single-feature attribution.

Implementation of TCAV:

- Let $f : \mathcal{X} \to \mathbb{R}^K$ denote a trained deep neural network classifier.
- Choose an internal layer l of the network and denote its activation function as $f_l : \mathcal{X} \to \mathbb{R}^m$, which maps an input x to an m-dimensional representation in layer l. and $h_{l,k} : \mathbb{R}^m \to \mathbb{R}$ represents the "latter part of the network," from the activation of layer l to the output logits of class k.
- Define a human-understandable concept C by collecting a set of concept examples $\mathcal{X}_C$ and a set of other concept examples $\mathcal{X}_{\bar{C}}$.
- Compute the activations $\{f_l(x) \mid x \in \mathcal{X}_C\}$ and $\{f_l(x) \mid x \in \mathcal{X}_{\bar{C}}\}$ at layer l for both concept and control samples.
- Train a linear binary classifier in the activation space to distinguish concept samples from control samples, and use the unit normal vector(The vector points towards the "is a concept" side.) of the classifier as the Concept Activation Vector (CAV), denoted by $v_C^l \in \mathbb{R}^m$.
- For a target class k, compute the directional derivative of the model output with respect to the CAV:

$$S_{C,k,l}(x) = \nabla h_{l,k}(f_l(x)) \cdot v_C^l,$$

which measures the sensitivity of the prediction for class k to changes along the concept direction.

- Define the TCAV score for concept C and class k as the proportion of inputs for which the directional derivative is positive:
$$TCAV_{Q_{C,k,l}} = \frac{|x \in \mathcal{X}_k : S_{C,k,l}(x) > 0|}{|\mathcal{X}_k|}.$$

- Repeat the above procedure with multiple random control sets and perform statistical testing to assess the significance and stability of the concept influence.

3 Similarities between LIME and TCAV

LIME and TCAV share several commonalities in their fundamental stance and methodological framework for interpretability research:

- First, both methods base their explanations entirely on the established model. They analyze model behavior after model training is complete, without modifying the original model structure, parameters, or training process.

- Second, in terms of objectives, both LIME and TCAV aim to enhance human understanding of the model's decision-making process. Both methods treat the model as a system with opaque internal mechanisms, attempting to reveal the information and patterns the model relies on when making predictions through additional means.

- Furthermore, in terms of explanation, both TCAV and LIME provide descriptive rather than causal explanations. They reveal the sensitivity of the model output to specific feature changes, rather than the true causal impact of these factors on the prediction results in the real world.

- Finally, from a practical perspective, both LIME and TCAV rely on human design. LIME requires human-defined interpretable features, while TCAV requires manually collecting and defining concept samples and control samples. This indicates that both methods are influenced to some extent by the subjective decisions of the researchers.

4 Differences between LIME and TCAV

Although LIME and TCAV share common goals from some perspectives, they differ substantially in several key aspects:

- First, the two methods differ in their explanation granularity. LIME approximates the model behavior in a local neighborhood of a specific input instance. As a result, LIME provides explanations that answer the question of why a particular prediction was made. In contrast, TCAV aims to explain model behavior at a higher level by analyzing the influence of human-defined concepts across a set of samples. TCAV addresses whether a concept is generally important for predicting a certain class.

- Second, LIME and TCAV differ in their semantic level of explanation. LIME operates primarily at the level of input features, such as words in text, super-pixels in images. Its explanations are expressed as feature importance scores within a surrogate model. While TCAV operates at a higher semantic level by introducing human-understandable concepts, such as visual patterns. This allows TCAV to provide concept-level explanations that are more aligned with human reasoning.

- Third, the two methods rely on different assumptions. LIME assumes that the complex model can be locally approximated by a simple and interpretable model. TCAV, on the other hand, assumes that meaningful human concepts can be represented as linearly separable directions in the intermediate layers of a deep network. This assumption makes TCAV more closely connected to the internal structure of deep neural networks.
- In addition, LIME and TCAV differ in their model dependency. LIME is explicitly designed to be model-agnostic and can be applied to any predictive model. TCAV, however, requires access to internal activations and gradients, which makes it more suitable for deep neural networks and less applicable to other models.
- Finally, the two methods differ in their output form. LIME produces explanations in the form of feature weights for a specific prediction, which are often visualized. TCAV produces quantitative scores that reflect the statistical influence of a concept on a class prediction across multiple samples.

5 Applicability of different methods

LIME is more suitable for scenarios requiring explanations of individual prediction results. Because LIME explains a specific input instance through local approximation, it can directly answer the question "Why did the model make this prediction for this sample?" Therefore, LIME is particularly suitable for explanation needs aimed at end-users, such as providing intuitive feature-level explanations for individual prediction results in applications like recommendation systems, medical diagnostic assistance, or financial risk control. Furthermore, LIME's model-agnostic nature allows it to be applied to different models,making it highly versatile.

In contrast, TCAV is more suitable for analyzing overall model behavior and decision dependencies at the conceptual level. TCAV statistically analyzes the model's prediction behavior across multiple samples by introducing human-understandable concepts, thereby evaluating whether a concept generally influences predictions for a specific class. Therefore, TCAV is more suitable for model auditing and bias analysis, such as examining whether the model relies on unexpected conceptual features or evaluating whether model decisions align with human prior knowledge. Additionally, because TCAV relies on intermediate layer features, it has a unique advantage in analyzing the internal mechanisms of deep neural networks.

6 Limitation of the methods

6.1 Limitations of LIME

Although LIME offers high intuitiveness and flexibility in explaining individual instances, it still has several important limitations: First, the quality of LIME's explanations depends on the design of the local perturbation strategy. Different perturbation methods can lead to significantly different explanation results, making it difficult to guarantee the stability and consistency of the explanations. Second, LIME is based on the assumption that the model can be approximated by a simple model in the local region. However, this assumption may not hold near highly non-linear decision boundaries, thus preventing the surrogate model from accurately reflecting the original decision logic. Furthermore, because LIME is influenced by human design choices, it limits the objectivity of the explanations to some extent.

6.2 Limitations of TCAV

TCAV also has significant limitations. First, the effectiveness of TCAV depends on the quality and definition of the concept samples. If the concept samples themselves are biased, noisy, or poorly defined, the learned CAV may not accurately represent the concept. Second, TCAV assumes that the concept can exist in a linearly separable direction in the intermediate layers of the model, but this assumption does not necessarily hold for all deep neural networks. In addition, TCAV's results can vary significantly depending on the choice of intermediate layers and the method of constructing control samples, leading to significant differences in concept influence measurements. Finally, because TCAV relies on the intermediate layers of the model, its applicability is limited to deep neural networks and cannot be easily generalized to other models.

7 Summary and Reflection

In summary, LIME and TCAV provide effective approaches to interpreting complex models from different perspectives. LIME, through local approximation, provides intuitive feature-level explanations for individual prediction results, suitable for instance-based and user-level explanation needs; while TCAV, by introducing human-understandable concepts, analyzes the conceptual dependencies of model decisions in the intermediate representation space, making it more suitable for overall model behavior analysis and auditing. Both methods have their own strengths and limitations. Therefore, they are not mutually exclusive but rather complementary in different application scenarios.