CAVLI - Using image associations to produce local concept-based explanations

Pushkar Shukla [1] Sushil Bharati [2] Matthew Turk [1]

[1]Toyota Technological Institute at Chicago [2]Teladoc Health

Motivation

Goal

- Design an explainability tool to measure the dependence of decisions made by image classifiers on high-level human concepts.
- This enables us to reason about machine learning models that can be used in high stake decisions, identifying biases, and detecting spurious correlations

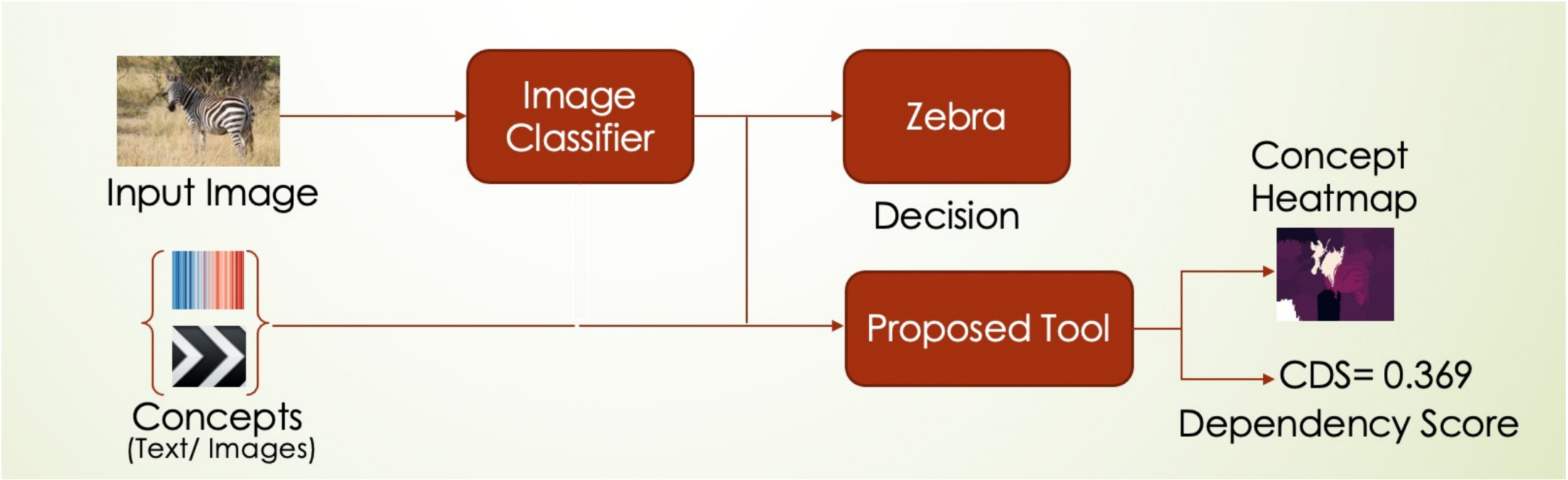


Figure 1. Given an image (Zebra) and a classifier (IncpetionNet), we are interested in understanding the role a concept (stripes) played in the decision. Our approach returns a (1) **Concept Dependency Score (CDS)** that quantifies the dependence of a the decision on a given concept and (2) a **concept heatmap** that highlights parts of the image that correspond most to a concept.

Background: Testing with Concept Activation Vectors (TCAV)

Notation

Consider a trained neural network $F : X \rightarrow \{1, ..., K\}$, on a dataset $X = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_t}\}$ and associated labels $Y = \{y_1, y_2, ...y_t\}$, where $y_i \in \{0,1\}^K$ with K classes. $F_k(\mathbf{x_i}) := h_l(f_l(\mathbf{x_i}))$, where $f_l(\mathbf{x_i})$ are the output logits of the l^{th} layer and h_l is the activation function of the l^{th} layer .

TCAV

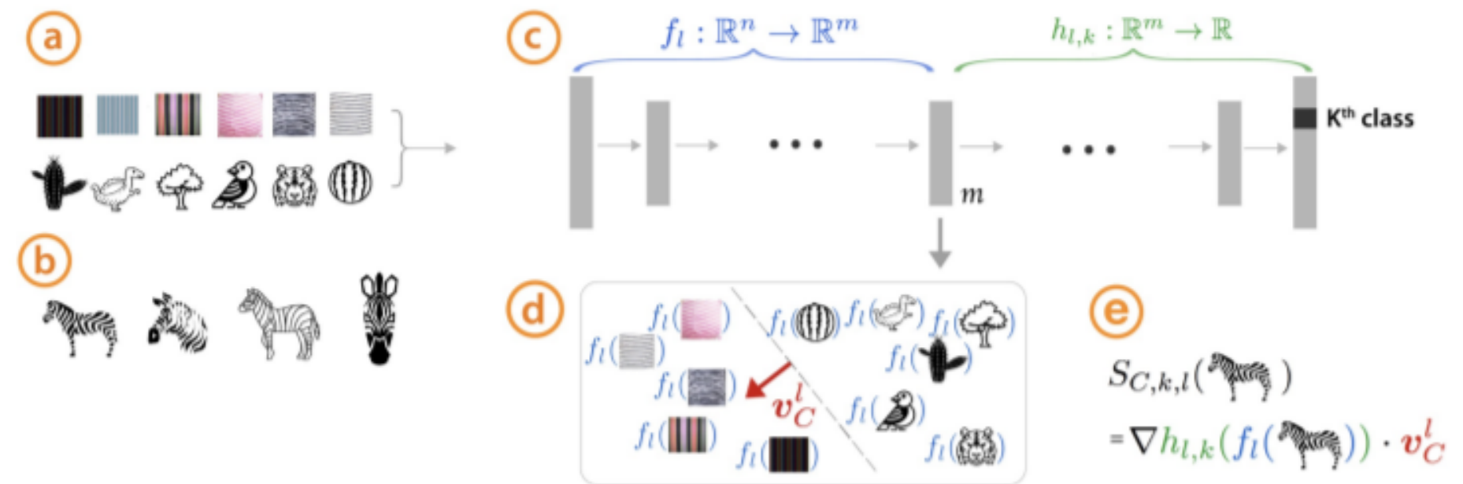

a
b
c
$f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$
$h_{l,k} : \mathbb{R}^m \rightarrow \mathbb{R}$
Kth class
m
d
v_C^l
e
$S_{C,k,l}$
$= \nabla h_{l,k}(f_l(\ )) \cdot v_C^l$

Figure 2. Overview of TCAV

- Provides global explanations on human-defined Concepts C for a machine learning model.
- Trains a classifier that distinguishes concept activations (e.g., "grassland" or "indoors") from activations of random samples
- The Concept Activation Vector (CAV) $\mathbf{v_c}$ is defined as the normal to the hyperplane separating examples *without* the concept from examples *with* the concept
- Conceptual Sensitivities ,$CS^k_{C,l}$ are used for understanding the influence of the concept on the model's prediction.

$$CS^k_{C,l}(F, \mathbf{x_i}) := \frac{\partial h_l(f_l(\mathbf{x_i}))}{\partial \mathbf{v_c}} = \delta h_k(f_l(\mathbf{x_i}))^T \mathbf{v_c} \tag{1}$$

Methodology

High-Level Intuition

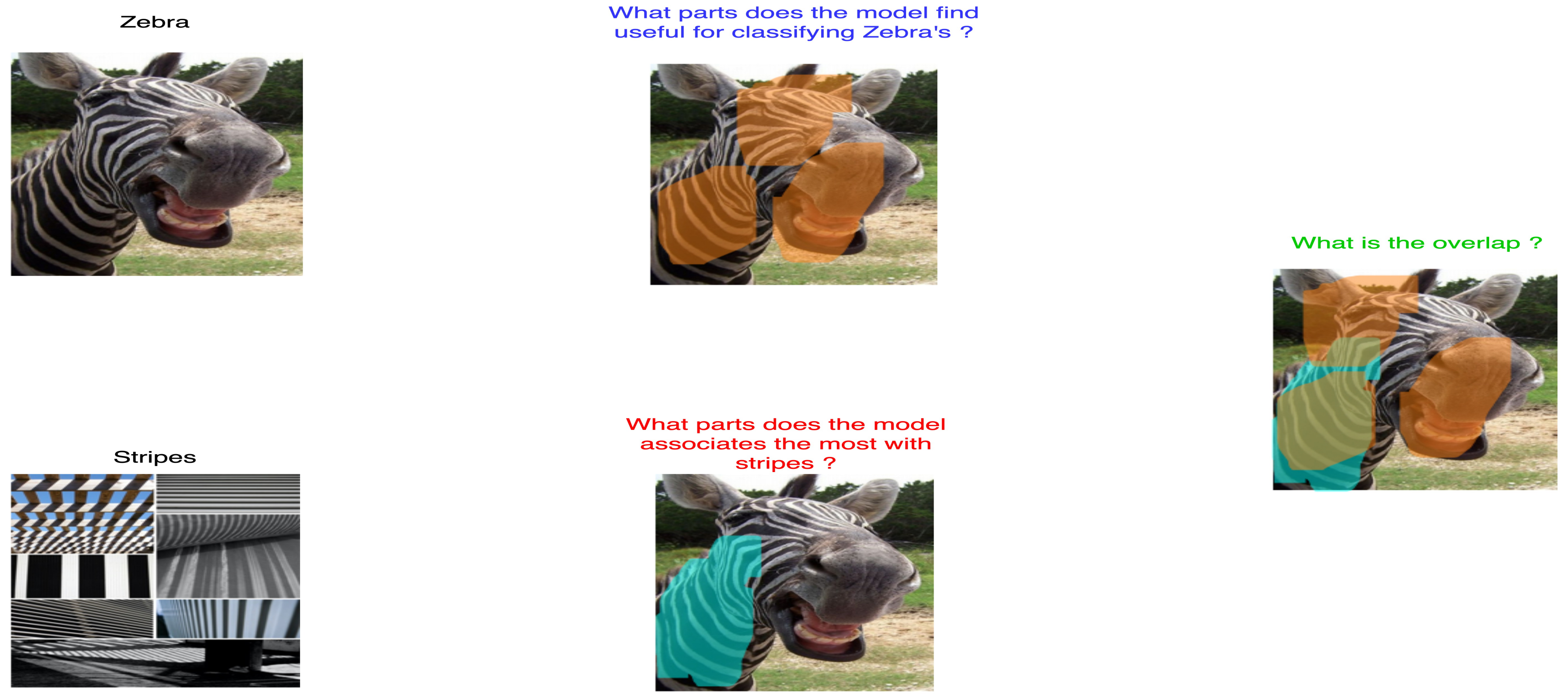

Zebra
What parts does the model find useful for classifying Zebra's ?
What is the overlap ?
Stripes
What parts does the model associates the most with stripes ?

Figure 3. We break down our problem into three parts. The first part measures what parts of the image are useful in decision making. The second part measures what parts the model associates most with a given concept. The third part measures the overlap between the first two parts, in order to show the effect of the concept on the decision.

Proposed Approach

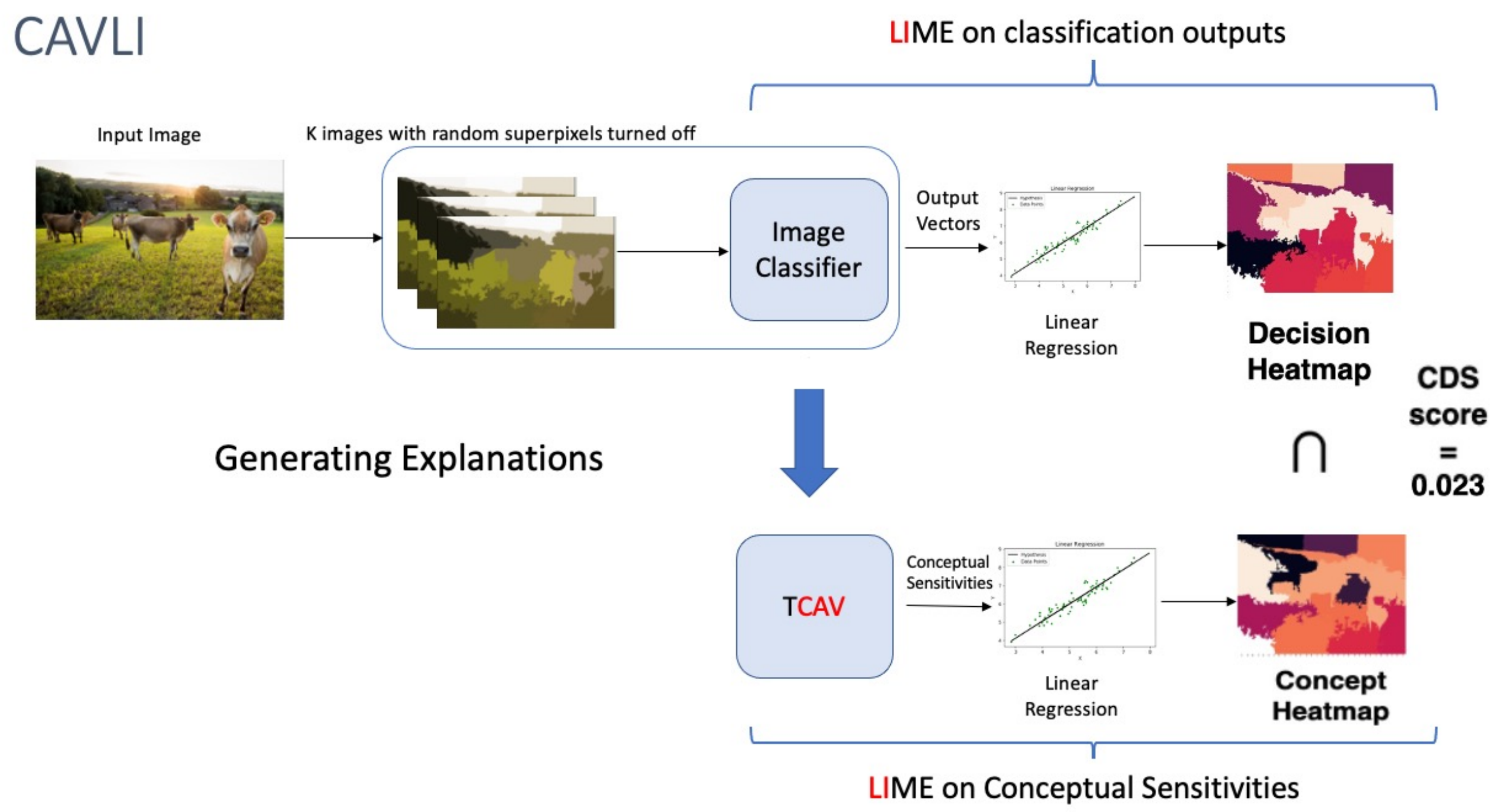


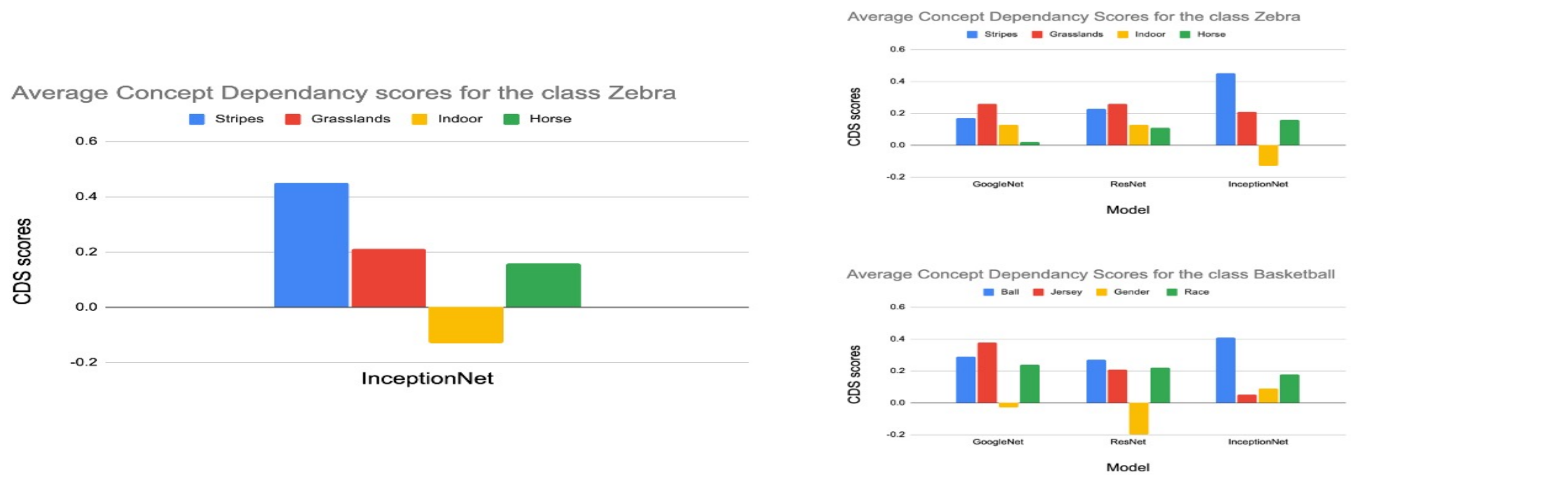
Figure 4. A block-level representation of our proposed approach

Algorithm 1 CAVLI

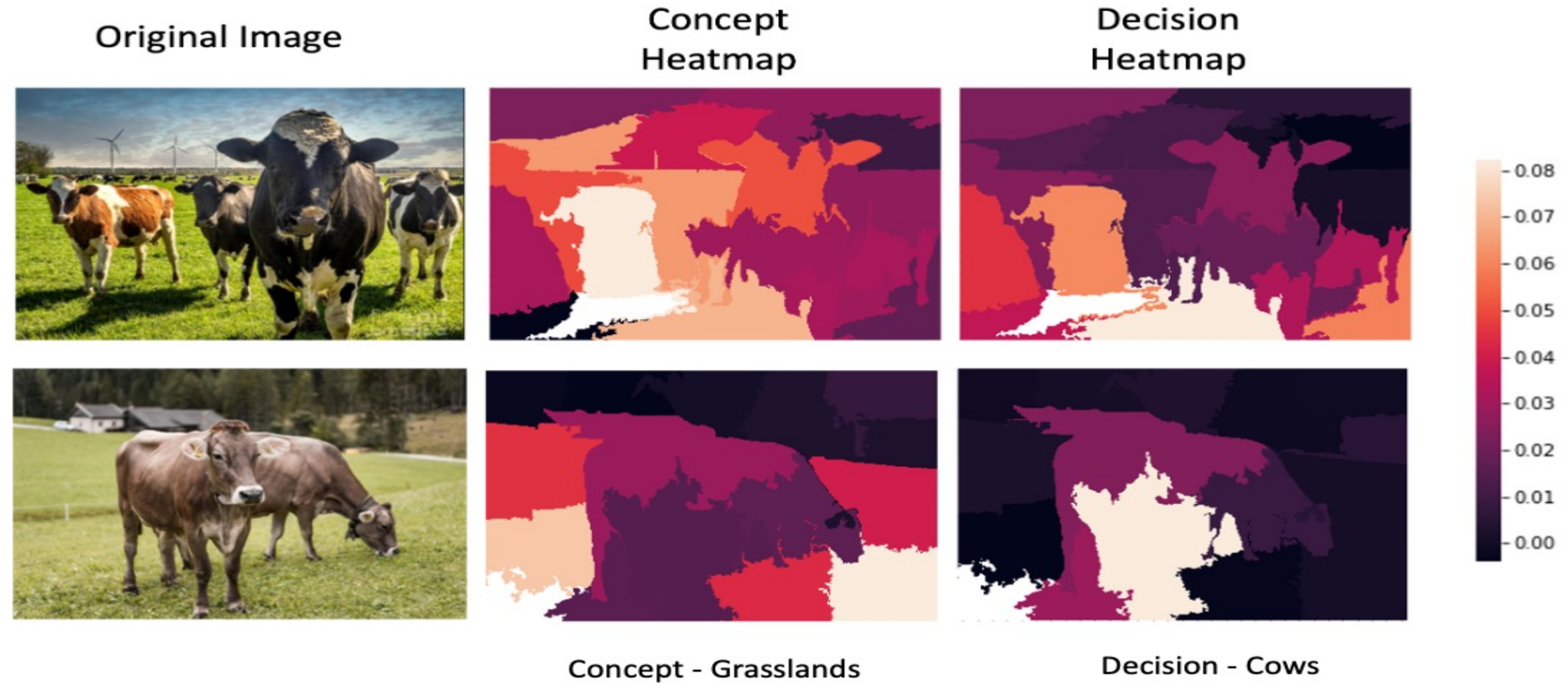
```
Train a TCAV model for a given concept $C$, a model $F$, and a layer $l$, resulting in the CAV vector $\mathbf{v_c}$.
Decompose the input image $x_i \in X$ in a set of $r$ homogeneous superpixels $\{S\}$.
Create a new set of images $\{\mathbf{x_{i_1}}, ....., \mathbf{x_{i_n}}\}$ from $\mathbf{x_i}$ by randomly masking parts of the image and selecting $n$ uniformly sampled subsets of $\{S\}$
Calculate the Conceptual Sensitivities $z_{i_j}$=$CS^k_{C,l}(F, \mathbf{x_{i_j}})$ $\forall x_{i_j} \in \{\mathbf{x_{i_1}}, .....\mathbf{x_{i_n}}\}$.
Build a local weighted surrogate model $\hat{\alpha}_i$ fitting the $z_{i_j}$'s to the presence or absence of superpixels.
Query the model for each of these image patches $y_{i_j} = F(x_{i_j})$ $\forall x_{i_j} \in \{\mathbf{x_{i_1}}, ....., \mathbf{x_{i_n}}\}$.
Build a local weighted surrogate model $\hat{\beta}_i$ fitting the $y_{i_j}$'s to the presence or absence of superpixels.
Calculate the Pearson correlation $\gamma_i$ between the coefficients of $\hat{\alpha}_i$ and $\hat{\beta}_i$.
Calculate the Concept Dependency Score, $CDS_i$= $\gamma_i \cdot CS^k_{C,l}(F, x_i)$
```

Results

Quantitative Results



Visualizing Heatmaps



Future Directions

- Experimenting with different architectural changes like pixel wise-weight assignment strategies over patch wise weight assignment, attention-based methods over attribution based methods.
- Conducting experiments to understand the pros/cons of association-based explainability techniques like CAVLI vs. other forms of image explainability techniques (attention based methods/ attribute based methods/ concept based methods) .

References

[1] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *International Conference on Machine Learning*, 2016.