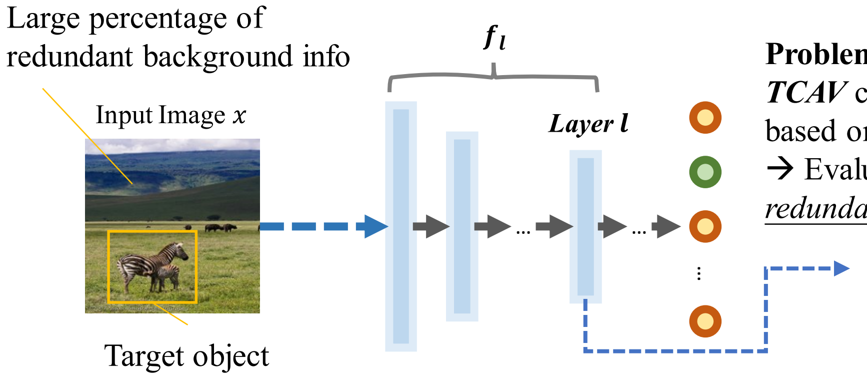
# Exploring Concept Contribution Spatially: Hidden Layer Interpretation with *Spatial Activation Concept Vector*

Andong Wang, Wei-Ning Lee

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

## I. Introduction



**Problem**

***TCAV*** calculates concept contribution to a target class based on *a whole hidden layer*

→ Evaluation may be interfered with by <u>*redundant background features*</u>

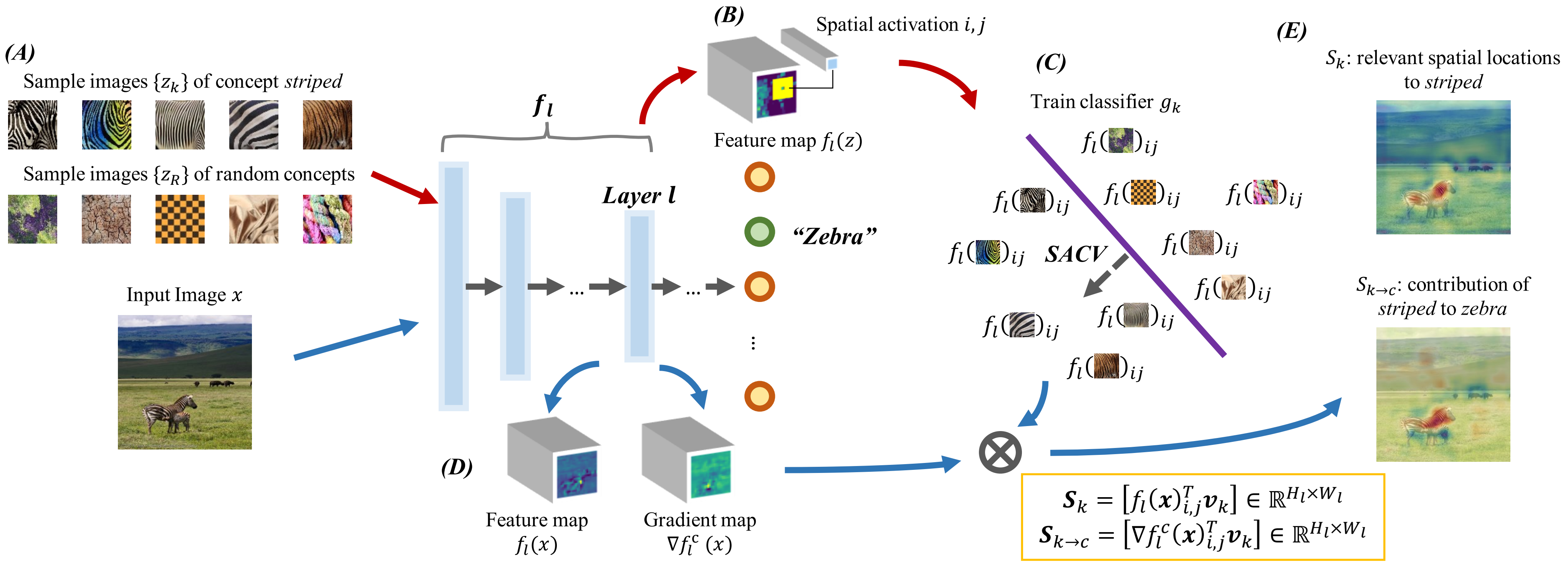
$$\nabla h_{l,c}(f_l(\boldsymbol{x}))v_{TCAV}^{l}$$

**<u>S</u>patial <u>A</u>ctivation <u>C</u>oncept <u>V</u>ector (*SACV*)**

Which identifies the *relevant spatial locations* to the *query concept* while evaluating their contributions to the model prediction of the *target class*

https://github.com/AntonotnaWang/Spatial-Activation-Concept-Vector

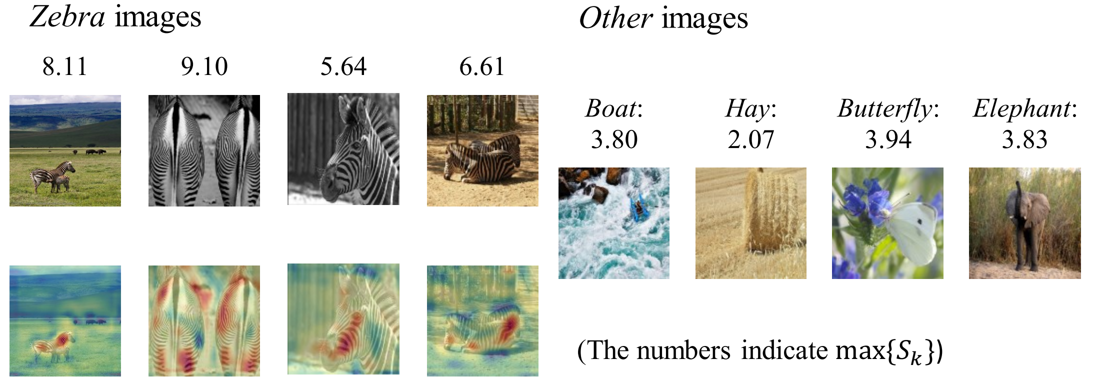
## II. Method


(A)
Sample images $\{z_k\}$ of concept *striped*
Sample images $\{z_R\}$ of random concepts
Input Image $x$
(B)
Spatial activation $i, j$
Feature map $f_l(z)$
$f_l$
Layer $l$
*"Zebra"*
(C)
Train classifier $g_k$
$f_l(\ )_{ij}$
*SACV*
(D)
Feature map $f_l(x)$
Gradient map $\nabla f_l^c\ (x)$
(E)
$S_k$: relevant spatial locations to *striped*
$S_{k\to c}$: contribution of *striped* to *zebra*

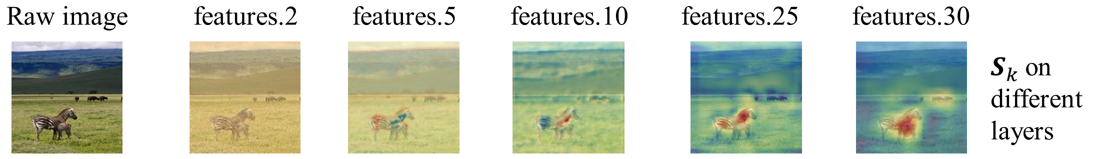
$$\boldsymbol{S}_k = \left[f_l(\boldsymbol{x})_{i,j}^{T}\boldsymbol{v}_k\right] \in \mathbb{R}^{H_l \times W_l}$$
$$\boldsymbol{S}_{k\to c} = \left[\nabla f_l^{c}(\boldsymbol{x})_{i,j}^{T}\boldsymbol{v}_k\right] \in \mathbb{R}^{H_l \times W_l}$$

## III. Results

### *(A)* $S_k$: relevant spatial locations to concept *striped*



(The numbers indicate max$\{S_k\}$)



$\boldsymbol{S}_k$ on different layers

### *(B)* $S_{k\to c}$: contribution of concept *striped* to class *zebra*