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

Andong Wang, Wei-Ning Lee
The University of Hong Kong
wangad@connect.hku.hk, wnlee@eee.hku.hk

Figure 1. Pipeline of Spatial Activation Concept Vector (SACV). See Sec. 2 for details of each step.

1. Introduction

To interpret deep learning models, one mainstream is to explore the learned concepts by networks [1, 4, 5]. Testing with Concept Activation Vector (TCAV) [4] presents a powerful tool to quantify the contribution of query concepts (represented by user-defined guidance images) to a target class. For example, we can quantitatively evaluate whether and to what extent concept striped contributes to model prediction zebra with TCAV. Therefore, TCAV whitens the reasoning process of deep networks. And it has been applied to solve practical problems such as diagnosis [2].

However, for some images where the target object only occupies a small fraction of the region, TCAV evaluation may be interfered with by redundant background features because TCAV calculates concept contribution to a target class based on a whole hidden layer.

To tackle this problem, based on TCAV, we propose Spatial Activation Concept Vector (SACV) which identifies the relevant spatial locations to the query concept while evaluating their contributions to the model prediction of the target class. Experiment shows that SACV generates a more fine-grained explanation map for a hidden layer and quantifies concepts’ contributions spatially. Moreover, it avoids interference from background features. Code is available on https://github.com/AntonotnaWang/Spatial-Activation-Concept-Vector.

2. Method

Given deep network \( f \), chosen layer \( l \), input image \( x \) with target class \( c \), and query concept \( k \), our goal is to calculate the contribution of concept \( k \) on feature map \( f_l(x) \in \mathbb{R}^{C_l \times H_l \times W_l} \) at different spatial activations.

Query concepts represented by guidance images. Given concept set \( \mathcal{K} \), for each \( t \in \mathcal{K} \), we have guidance images \( \{ z_t \} \) and assume that all contents of \( z_t \) represent concept \( t \). Then, we prepare two sets of guidance images: \( \{ z_k \} \) \( (N_k \) images and \( k \in \mathcal{K} \) which represents concept \( k \) and \( \{ z_R \} \) \( (N_R \) images and \( \mathcal{R} \subseteq \mathcal{K}\backslash k \) which represents other random concepts (Fig. 1 (A)).

Train SACV with guidance images. Next, we feed the guidance images to \( f \) and divide each feature map to \( H_l \times W_l \) spatial activations (Fig. 1 (B)). After that, as shown in Fig. 1 (C), we train a linear classifier (i.e., a hyperplane) \( g_k \) to separate the two sets of spatial activations -
\( \{ f_i(z_k) \}_{i=1}^{N_k \times H_i \times W_i} \) (concept \( k \)) and \( \{ f_i(z_R) \}_{i=1}^{N_R \times H_i \times W_i} \) (other random concepts). The coefficient vector \( v_k \in \mathbb{R}^{C_l} \) (i.e., the norm vector of the hyperplane) is the \textit{Spatial Activation Concept Vector} (SACV).

**Interpretation using SACV.** As shown in Fig. 1 (D), we first extract input \( x \)'s feature map \( f_i(x) \) and gradient map \( \nabla f_i(x) \) (with respect to target class \( c \)). In Fig. 1 (E), \( f_i(x) \) to indicate the relevance of each spatial activation to concept \( k \). we multiply each spatial activation with SACV, noted as \( f_i(x) v_k \). There, we obtain an explanation map \( S_k = [f_i(x)]_{i,j} v_k \). As shown in Fig. 2 (C), the red regions in the heat maps indicate the spatial activations which are relevant to concept \( k \) while having high contributions to class zebra. It is consistent with human understanding as the zebra contents are highlighted. Also, in Fig. 2 (D), the receptive fields of spatial activations show that \( k \) features contribute most to model prediction zebra while background features have the lowest contributions, meaning that the model makes the right decision with a right reason.

**SACV avoids background interference.** It is noted that the receptive fields of some \( f_i(x)_{i,j} \)'s contain the target object while others do not. As SACV processes spatial activation individually, background interference is removed.

### 3. Experiment

We test SACV on the PyTorch VGG19 model pertained on ImageNet (the layer names are used in this paper). We use query concepts and guidance images from \textit{Describable Textures Dataset} \cite{Kim}. Results are shown in Fig. 2.

\( S_k \) indicates the relevance of each spatial location to concept \( k \). Fig. 2 (A) shows that for zebra images, the regions where zebra exists have high response to \textit{striped} while for other images, they all have low responses to \textit{striped} for all locations (low \( \max \{ S_k \} \)). It means that network \( f \) is able to distinguish \textit{striped} from other concepts on this layer. Moreover, Fig. 2 (B) indicates that \( f \) recognizes \textit{striped} on high layers (e.g., features.25) but does not learn the concept of \textit{striped} on shallow layers (e.g., features.2).

\( S_{k\rightarrow c} \) \textit{quantifies the contribution of concept} \( k \) \textit{to model prediction} \( c \). As shown in Fig. 2 (C), the red regions in the heat maps indicate the spatial activations which are relevant to concept \( k \) while having high contributions to class zebra. It is consistent with human understanding as the zebra contents are highlighted. Also, in Fig. 2 (D), the receptive fields of spatial activations show that \textit{striped} features contribute most to model prediction zebra while background features have the lowest contributions, meaning that the model makes the right decision with a right reason.

### References

\[ \text{[1] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. \textit{Proceedings of the National Academy of Sciences}, 2020.} \]

\[ \text{[2] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. \textit{In Proceedings of the 2019 chi conference on human factors in computing systems}, 2019.} \]

\[ \text{[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. \textit{In Proceedings of the IEEE conference on computer vision and pattern recognition}, 2014.} \]

\[ \text{[4] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). \textit{In International conference on machine learning}. PMLR, 2018.} \]

\[ \text{[5] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization: How neural networks build up their understanding of images. \textit{distill}, 2018.} \]