A Choice of Concept Set

For RCAV, a set of representative images defines a concept. For any semantic concept there will be many possible choices of concept set: e.g. we may choose between either color patches or colored objects to define a particular color. In practice, it is usually easiest to draw concept set images from the validation set. For instance, to define the concept red, we could select the 100 validation set images with highest intensity in the red channel. When sampling images for the concept set, it is necessary to maintain class balance – i.e. the number of samples per class for each sub-concept must be identical. For the TFMNIST and Camelyon experiments we use validation set images to define the concept sets. In contrast, for the ImageNet experiments, we follow the precedent set in TCAV using Broden images for textures and Gaussian-noised color patches for color [4, 15].

If there are multiple possible choices of concept set, the optimal choice minimizes the variance of the set of null concept sensitivity scores, $\{S_{C,n}\}_{n \in N}$. This optimal choice of concept set will minimize false negatives, i.e. misidentification of concepts that are meaningful to model prediction as statistically insignificant.

B Hyperparameter Sensitivity Analysis

Table 4: Performance as a function of layer for contrast-augmented CAMELYON16.

	RCAV			TCAV		
	P_τ	AUROC	AUPRC	P_τ	AUROC	AUPRC
Conv2d 3b 1x1	92%	0.94	0.89	45%	0.55	0.27
Mixed 5c	89%	0.94	0.86	68%	0.65	0.31
Mixed 6d	83%	0.84	0.77	69%	0.70	0.34
Mixed 7b	91%	0.94	0.83	70%	0.70	0.37

Layer Interpretability methods seek to explain model predictions. We propose using layer-specific RCAV scores for interpretability, but we may only extrapolate from layer-specific results if RCAV performance is layer invariant. At the image level, Table 4 shows that RCAV predicts concept sensitivity for all layers considered.

At the dataset level, we observe that the absolute value of $S_{C,i,k}$ increases monotonically as the layer approaches softmax. We explain this increase by observing that head of the model, f_l^+, converges to linearity as l approaches l_{max}. In the linear case, any fixed CAV will have $|S_{C,i,k}| = 0.5$, because the effect of perturbation in a fixed direction is invariant over choice of input for a linear classifier.

To ensure that any conclusion based on a specific layer's RCAV scores is representative of the whole model, we need a weak form of layer invariance: consistency of sign$(S_{C,i,k})$ across layers. Figure 4 shows that for six out of the seven concepts considered, RCAV consistently predicts dataset-level concept sensitivity. It is possible that layer inconsistency occurs when a concept plays a non-binary role in the classifier's decision function. For this reason, we recommend testing multiple layers when using RCAV for dataset-level concept sensitivity quantification.

Table 5: Performance as a function of step size for contrast-augmented CAMELYON16, using layer Conv2d3b.

Step Size	$AVG(s_{C,i,k})$	P_τ	AUROC	AUPRC
0.1	2e-5	81%	0.93	0.88
1	2e-4	88%	0.94	0.86
10	2e-3	94%	0.94	0.89
100	0.03	91%	0.94	0.86

Step size In real-world use of RCAV, ground truth concept sensitivity is not known, so it is impossible to tune step size for optimal performance. Instead we suggest choosing step size such that observed concept sensitivity scores, $s_{C,i,k}$, range from 0.001 to 0.1. This is the observed range of softmax differences in the benchmark experiments shown in Figure 2 and Figure 3. Empirically, RCAV performance is robust across choice of step size, as shown in Table 5.

Table 6: Performance as a function of label binarization threshold for contrast-augmented CAMELYON16, using layer Conv2d3b.

	RCAV		TCAV	
Threshold	AUROC	AUPRC	AUROC	AUPRC
5%	0.94	0.99	0.35	0.84
25%	0.97	0.99	0.38	0.68
75%	0.94	0.89	0.55	0.27
95%	0.98	0.90	0.63	0.07

Label binarization threshold In practice, we often use RCAV to make a binary decision: either the input is sensitive to the concept, or the input is not sensitive to the concept. In subsection 4.2, we used AUROC and AUPRC to quantify the accuracy of RCAV for this binary task. The ground truth for this task is whether model prediction delta exceeds a certain threshold when augmenting inputs. Formally the ground truth labels are defined by $x \mapsto \mathbb{1}(f^k(x) - f^k(x') > t)$ for some fixed threshold t, augmented input x' and class k. In Table 6 we choose our threshold as a percentile of the ground truth sensitivity values, and show that RCAV performs robustly across all thresholds.

C Measuring Concept Encoding Linearity

Table 7: Accuracy of rank one approximations to concept latent encodings.

Layer	TFMNIST	CAMELYON16
Conv2d 3b 1x1	24%	50%
Mixed 5c	40%	37%
Mixed 6d	40%	80%
Mixed 7b	44%	84%

RCAV relies on CAV's linear approximation of the model's concept encoding. By doing an SVD on the ground truth concept sensitivity differences, we can measure the extent to which this linearity constraint bottlenecks RCAV performance. The CAV can reliably estimate the ground truth effect only if the difference vector between encodings of input, x, and augmented input, x', is similar to the CAV – i.e. $\|V_{C,i} - (f_l(x) - f_l(x'))\| < \epsilon$. If, on the other hand, the difference vector has high variance across points of the validation set, then the effect of the concept cannot be encoded as a CAV. We can measure the extent to which the concept is consistently encoded by examining the matrix of pairwise encoding differences,

$$D_l = \begin{bmatrix} f_l(x_0) - f_l(x'_0) \\ \vdots \\ f_l(x_n) - f_l(x'_n) \end{bmatrix} \tag{6}$$

The optimal CAV[3] is the first singular vector for D_l, because this vector best approximates $f_l(x_i) - f_l(x'_i)$. Using the SVD, we can upper bound the performance of RCAV on layer l by calculating the

[3] In practice, the optimal CAV cannot be calculated in this way, because it is not feasible to counterfactually augment the input – i.e. we do not have x'.

reconstruction accuracy, r, of the best rank one approximation to D_l. Matrix dimensions varies across layer, so we normalize the reconstruction accuracy to $r = \|D_l - D_{l,1}|/\|D_l\|$ where $D_{l,1}$ is the rank one approximation and we use the Frobenius norm. Table 7 shows that reconstruction accuracy is higher for CAMELYON16 than TFMNIST. We infer that the CAMELYON16 model's encoding of the contrast concept is more linear than the TFMNIST model's encoding of the texture concepts. These results explain the difference in performance between these two datasets seen in Table 1.