Best of both worlds: local and global explanations with human-understandable concepts

Jessica Schrouff[*]
Google Research
London, United Kingdom
schrouff@google.com

Sebastien Baur[*]
Google Health
London, UK

Shaobo Hou
DeepMind
London, UK

Diana Mincu
Google Research
London, UK

Eric Loreaux
Google Health
Palo Alto, CA, USA

Ralph Blanes[†]
Google Research
Mountain View, CA, USA

James Wexler
Google Research
Mountain View, CA, USA

Alan Karthikesalingam
Google Health
London, UK

Been Kim
Google Research
Mountain View, CA, USA

Abstract

Interpretability techniques aim to provide the rationale behind a model's decision, typically by explaining either an individual prediction (local explanation, e.g. 'why is this patient diagnosed with this condition') or a class of predictions (global explanation, e.g. 'why are patients diagnosed with this condition in general'). While there are many methods focused on either one, few frameworks can provide both local and global explanations in a consistent manner. In this work, we combine two powerful existing techniques, one local (Integrated Gradients, IG) and one global (Testing with Concept Activation Vectors), to provide *local, and global* concept-based explanations. We first validate our idea using two synthetic datasets with a known ground truth, and further demonstrate with a benchmark natural image dataset. We test our method with various concepts, target classes, model architectures and IG baselines. We show that our method improves global explanations over TCAV when compared to ground truth, and provides useful insights. We hope our work provides a step towards building bridges between many existing local and global methods to get the best of both worlds.

1 Introduction

Interpretability in machine learning (ML) has been deemed a key element for trustworthy models [Lip18, DV17], and considered as a core requirement to deploy ML to high-stake domains such as healthcare or self-driving. When stakes are high, explanations are often expected at multiple levels: first at the global/population level to obtain a general understanding of a model's predictions, as well as at the local level (e.g., one patient of interest). Building inherently interpretable models (e.g. using attention [BCB16], including rule lists [ALSA$^+$17], or knowledge [CLT$^+$19, KNT$^+$20]) and post-hoc methods are widely explored in the field (e.g. [SVZ14, RSG16, LL17, ACÖG18, AÖG19]) to meet these requirements.

[*]equal contribution
[†]Now at korgi.ai

Preprint. Under review.

Among interpretability methods, post-hoc global or local methods have gained significant interest due to their convenience (see [EEG+18] for a review). In particular, "attribution methods" is a family of methods that provide an importance score for each input feature. For vision applications, these scores are typically per pixel and can be displayed as a heat map, often referred to as an "attribution map". Naturally, the importance of a pixel is limited to one image (local), and cannot be simply "averaged" to learn which pixels are important in general for the class (global level). It was however shown that practitioners are also keen to understand the "overall reasoning" of the model [TJMG19]. How does the model reason to classify a set of patients to a degree of cancer? To answer this, a variety of *global* explanation methods are investigated (e.g. [WSC+20, HPRPC20]). Despite the need for local and global methods, only few can provide both types of explanations [LEC+19, LHK19].

This work is an attempt to bridge the gap between local and global explanation techniques by combining two popular methods: Integrated Gradients (IG [STY17]) and Testing with Concept Activation Vectors (TCAV [KWG+18]). IG offers input feature-wise importance based on a set of game-theory-inspired principles, while TCAV provides explanations using intuitive high-level concepts (e.g., red color, stripes). The proposed merging of the two results in a unique method that provides the best of both worlds: local and global explanations, with human-understandable concepts. Our contributions are:

- Our proposed formulation improves TCAV for global explanations to be faithful to the ground truth on two synthetic datasets while enabling IG to use concepts instead of features to explain model decisions.
- We propose a set of IG baselines, specific to our formulation.
- We derive a closed form solution in simple scenarios.
- We validate our method on two synthetic datasets as well as a natural image dataset, and test our results across multiple baselines, model architectures, and concepts.

2 Related work and methods

In this section, we take a deep dive on the two methods that we propose to combine: IG and TCAV and describe their limitations before presenting our method.

2.1 Notation and setup

Let our dataset be the junction of a set of n samples with d features $\mathcal{X} \in \mathbb{R}^{n \times d}$ and associated labels $\mathbf{y} \in \{0,1\}^n$. In the multiclass case, $\mathcal{X}_k$ represents the set of inputs whose label is k, with $k \in 1, \ldots, K$. We consider a trained neural network $F : \mathcal{X} \to \{1, ..., K\}$ with L layers. For a given layer $1 \leq l \leq L$ and a given class $1 \leq k \leq K$, we can write the k-th output of F as $F_k(\mathbf{x}) := h_k(f_l(\mathbf{x}))$ where $f_l(\mathbf{x})$ is the activation vector at the l-th layer, which we refer to as $\mathbf{a}_l$. We also define $\mathbf{W}_l \in \mathbb{R}^{d_{l+1} \times d_l}$ to represent the weights of the l-th layer, and $b_l \in \mathbb{R}^{d_{l+1}}$ as its bias. We drop the layer index l in the remaining of the paper for compactness.

2.2 Gradient-based attributions and integrated gradients

Of interest in our work is a family of saliency map techniques [SVZ14] that typically use the gradients w.r.t. input to provide a score per input feature. Recent developments in gradient-based techniques include e.g. gradients $\times$ inputs [SGSK16], LRP [BBM+15], DeepLIFT [SGK17], smoothGrad [STK+17], or integrated gradients [STY17]. In vision applications, these methods give a score to each pixel in one image, which makes them inherently local. While criticism of these methods exists [JW19, SGL19, AGM+18], they have been shown to be useful in practice [STR+19].

In particular, the integrated gradients (IG) technique [STY17] computes a path integral between an uninformative input (the baseline) $\mathbf{x}'$ and the observed input $\mathbf{x}$ (Eq.1). This technique verifies completeness (i.e. attributions sum to the difference in predictions [STY17, ACÖG18]) and sensitivity (i.e. irrelevant variables obtain a null attribution score). Formally, IG is defined as

$$\mathrm{IG}_i^k(\mathbf{x}, \mathbf{x}') := (x_i - x_i') \int_0^1 \nabla_i F_k(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))d\alpha \tag{1}$$

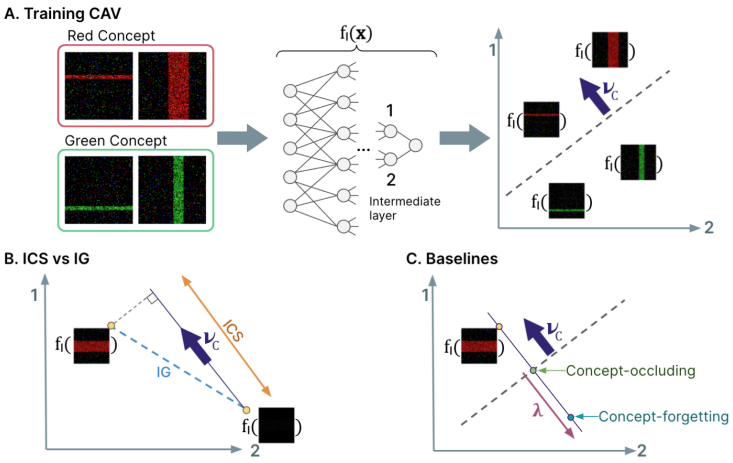

Figure 1: **Illustration of TCAV and ICS. A** Images of red (concept) and green (control) bars and their activations f_l in a network. A linear classifier (top right) is trained to distinguish concept activations from control images, generating a CAV (v_C). **B** ICS is the projection of integrated gradients (IG) between a baseline (e.g., black image) and a new sample (red horizontal bar) on the direction of the CAV, v_C. **C** Concept-forgetting baselines remove information in the direction of the concept, with a strength λ. 'Concept-occluding' is a particular case of this baseline that projects the activations on the linear model hyperplane.

Where ∇ is the gradient operator and ∇_i is its projection on the i-th dimension, $i \in 1, \ldots, d$. $\nabla f(\mathbf{x})$ designates the gradient of f evaluated at $\mathbf{x}$. One of the critical decision in using IG is deciding which baseline to use. Conceptually, the baselines represent "no information", and the original paper used a simple baseline (e.g., black/white images). However it was shown that the method is sensitive to the choice of baselines [ACÖG18, SLL20, GLW$^+$20]. In our work, we propose and test many baselines across model architectures and datasets.

2.3 Testing with Concept Activation Vectors

TCAV [KWG$^+$18] uses *concepts* (C) instead of features to provide explanations, where concepts are defined as understandable and identifiable by humans (e.g. "stripes" or "pointy ears"). TCAV has been successfully applied to different domains e.g. in healthcare [COPA$^+$19, GAM18, MLH$^+$21]. Technically, TCAV assumes that users have a set of samples for a concept of interest. To express this concept, they find a Concept Activation Vector (CAV) in a network's activations space $\mathbf{a} \in \mathbb{R}^d$ (a layer with d dimension) that points from any direction to these 'concept data samples'. Formally, the CAV is obtained by training a linear classifier to distinguish concept activations from random activations, and taking the unit-norm vector $\mathbf{v_C}$ orthogonal to its decision boundary. Intuitively, a CAV represents the direction of the selected concept as encoded in the network. Figure 1a illustrates the CAV building for a red/green concept in a 2-dimensional layer.

To understand the influence of a concept on the model's predictions, Conceptual Sensitivity (CS, Eq.2) is defined as the directional derivative of one of the network's outputs k w.r.t. the CAV for concept C:

$$\mathrm{CS}^k_C(F, \mathbf{x}) := \frac{\partial h_k(f(\mathbf{x}))}{\partial \mathbf{v_C}} = \nabla h_k(f(\mathbf{x}))^T \mathbf{v_C} \tag{2}$$

Note that CS does not provide a local explanation: while a positive directional derivative indicates that the predicted probability for class k would locally increase if the activation were to be perturbed in that direction, it does not quantify the role this direction played in the model's decision. In addition, the values of gradients are known to saturate [STY17], making it an unreliable source of local explanation.

CS can however provide global attributions by aggregating the scores over multiple examples to obtain a *TCAV* score between 0 and 1. Since the scale of directional derivative values is a priori unknown (i.e., CS is not a normalized quantity), TCAV circumvents this difficulty by defining the TCAV score as the average *sign* of CS values on a given dataset.

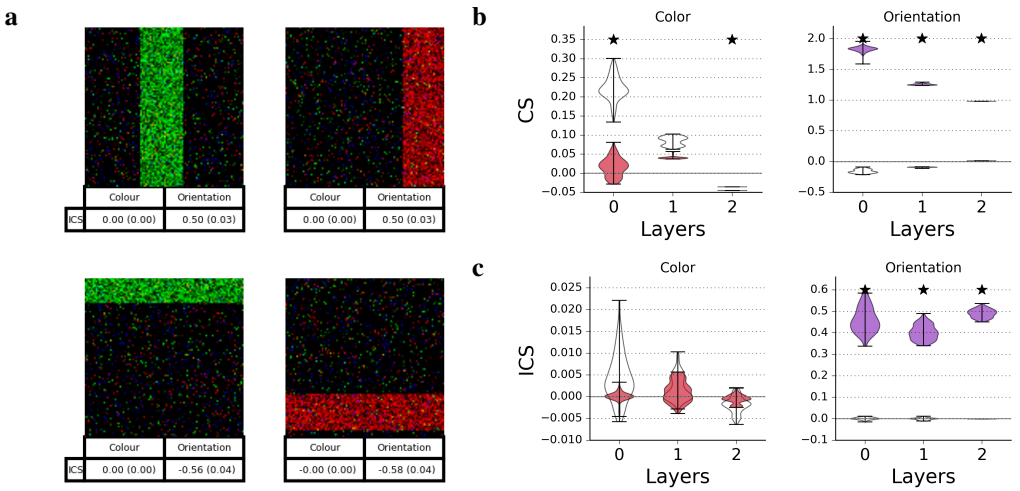

Figure 2: **CS leads to significant scores for irrelevant concepts, while ICS does not. a** Four representative local examples, displaying associated ICS for all concepts, at layer 2 of the model predicting bar orientation F_o. **b** Distribution of CS for target 1 of the orientation model F_o, for the color (displayed in red) and orientation (displayed in purple) concepts, for each layer of the model. Null distributions from permuted CAVs are displayed as white violins and significant TCAV$^{sign(CS)}$ scores are highlighted by a star (Bonferroni corrected for layers and concepts). **c** Distribution for ICS (computed using a black image baseline).

An important step in TCAV is ensuring that CAVs did not accidentally return a high TCAV score. To do this, the method suggests building many bootstrapped versions of the CAVs and run t-testing between TCAV scores from concept CAVs v.s. random CAVs (CAVs trained with random images). We also discuss how we adapt this test in our approach.

2.4 TCAV alone overestimates concept importance

We have however observed that TCAV scores can be significant even when the concept is not used in the model's predictions. We illustrate this failure mode using the synthetic dataset used in [GSK19] (referred to as "BARS"). This dataset contains noisy black images with vertical and horizontal lines that are either green or red (Figure 2**a**). Suppose the orientation of the bar defines the label y, while the color is irrelevant (i.e. the color is independent of the orientation). A 3-layered MLP with ReLU activations and 50% dropout after each fully-connected layer predicts the orientation of the bar (F_o, with 'o' referring to 'orientation'), reaching 99.97% accuracy in train and 100% in test. While this model does not rely on the color of the bar to make a prediction, both the "orientation" and "color" concepts seem to be encoded in the network's activations and significant CAVs can be built for both concepts (see Supplement for details). These represent the direction from red to green (color) and from horizontal to vertical (orientation). We observe that the orientation concept has a positive influence on the predictions in all layers, i.e. CS scores are positive (Figure 2b). The color concept is also displaying a positive influence for layers 0 and 1 and a negative influence for layer 2. Despite smaller magnitudes of CS for the color concept, both concepts are considered as "significant" under the mechanism proposed in TCAV (2 out of 3 layers for color). In this simple controlled setting, we hence display that the aggregation technique is not *sensitive* and can be misleading.

2.5 Our method: Integrated Conceptual Sensitivity (ICS)

We introduce Integrated Conceptual Sensitivity (ICS, Eq.3) as a *local, concept-based attribution technique*, which can also be aggregated into global explanations. For a concept C with unit norm CAV, $\mathbf{v_C}$, we define ICS as follows:

$$\text{ICS}_C^k(\mathbf{a}, \mathbf{a}') := (f(\mathbf{x}) - \mathbf{a}')^T \mathbf{v_C} \int_{[\mathbf{a}', f(\mathbf{x})]} \nabla_{\mathbf{v_C}} h_k(\mathbf{a}) d\mathbf{a} \tag{3}$$

Where $\mathbf{a}' \in \mathbb{R}^d$ represents the baseline's activation at some layer and the support of the integral $[\mathbf{a}', f(\mathbf{x})]$ corresponds to $\{\mathbf{a}' + \alpha(f(\mathbf{x}) - \mathbf{a}'), \alpha \in [0, 1]\}$. Intuitively, ICS (Figure 1b) represents

how much of the change in predicted probability is explained by the change in the direction of the concept $\mathbf{v_C}$, compared to a (typically uninformative) baseline. While ICS is the projection of the integrated gradients (a bounded quantity) onto the concept, there is no guarantee that this projection is itself bounded[3]. Global concept attributions can then be obtained by averaging ICS over multiple examples. Note that, contrary to the original formulation of TCAV score that averages the signs of CS (referred as $\mathrm{TCAV}^{sign(\mathrm{CS})}$), the ICS scores are aggregated using their average: $\mathrm{TCAV}^{\mathrm{ICS}} = \frac{1}{n}\sum_{i=0}^{n} ICS_{i,C}^{k}$.

Choice of baselines and closed form solutions: As discussed in [STY17, ACÖG18, GLW+20, SLL20], the choice of the baseline affects the obtained explanations. We test with 'uninformative' baselines (e.g., black and white image, maximum entropy leading to neutral predictions) and propose 'informative' baselines for concepts:

- *Concept-forgetting baseline*: $\mathbf{a}' = \mathbf{a} - \lambda \mathbf{v_C}$ for some $\lambda \in \mathbb{R}^{*}+$, which removes a certain amount of concept-aligned information.
- *Concept-occluding baseline*: $\mathbf{a}' = \mathbf{a} - (\mathbf{a}^T\mathbf{v_C} + b)\frac{\mathbf{v_C}}{||\mathbf{v_C}||_2^2}$: the baseline with the concept removed, where b is the bias of the linear model. This particular case of concept-forgetting can be related to the occlusion [ZF14] of the concept.

Naturally, the best baseline for each application will be different. In Section 3, we show this variation for different datasets. With entropy-maximizing and concept-forgetting baseline, we can derive analytical formulations of ICS, removing the computational expense related to the integral (see Supplement for derivation):

Analytical formulation 1: For the last layer of a binary classification model with entropy-maximizing baseline (i.e. $h(\mathbf{a}') = 0.5$), ICS can be written as (Eq.4):

$$\mathrm{ICS}_C(\mathbf{a}, \mathbf{a}') = \left(\frac{\mathbf{v_C}^T\mathbf{w}}{||\mathbf{w}||_2}\right)^2 \left(\sigma(\mathbf{w}^T\mathbf{a} + b) - 0.5\right) \tag{4}$$

Analytical formulation 2: For a multi-class model, with concept-forgetting baseline, we can derive ICS as (Eq.5):

$$\mathrm{ICS}_C^k(\mathbf{a}, \mathbf{a}') = h_k(\mathbf{a}) - h_k(\mathbf{a}') \tag{5}$$

Statistical testing for TCAV: An important part of TCAV is to confirm the statistical significance of the CAV. We propose a nonparametric permutation test to assess the significance of the out-of-sample classification performance of the CAV, which can be seen as a generalized version of the approach proposed in [KWG+18]. Intuitively, this test estimates whether a concept has been "significantly" encoded in each layer and prevents testing with non-significant directions. Concretely, we train CAVs on bootstrapped datasets (n=100 resamples). For each bootstrap sample, we build CAVs from the same data but with permuted 'positive' and 'negative' concept labels ($n_{perm} = 10$). We then compute the number of permuted CAVs leading to higher or similar performance compared to the performance of the non-permuted CAVs and assess a CAV as significant when this proportion is $p < 0.05$. Note that when testing multiple concepts, a correction method (e.g., Bonferroni, false discovery rate) should be added. The permuted CAVs can also be used to assess the significance of the scores by comparing the distribution of CS (resp. ICS) across bootstrap samples to the distribution of CS (resp. ICS) as estimated from the permuted CAVs. This test can be performed both at the global and local level.

3 Results

To validate our approach, we compare global ICS results to the original formulation of TCAV on two synthetic datasets with ground truth in Section 3.1 and illustrate local explanations using ICS on a benchmark imaging dataset in Section 3.2. We test our results across different models with varying complexity.

[3]Despite this fact, we rarely observe values for ICS that are larger than 1 in our experiments, see Supplement.

3.1 Quantitative evaluation on synthetic datasets

One of the challenges in interpretability methods is to show the "faithfulness" of an explanation. To validate our approach, we use a simple synthetic dataset along with semi-natural images that are synthetically generated. With carefully crafted distributions of this data, we can train models with known ground truth of what the explanation "should be". We describe the metrics and datasets used to quantitatively illustrate the performance of $\mathrm{TCAV}^{\mathrm{ICS}}$ over $\mathrm{TCAV}^{sign(\mathrm{CS})}$.

3.1.1 Datasets, models and metrics

BARS: This dataset from [GSK19] features noisy 100x100 images of red and green, horizontal and vertical bars. We introduce two models (3-layered ReLU MLP with dropout): F_o^{BARS} which is trained to predict the orientation of the bars, and F_c^{BARS} which is trained to predict their color.

BAM[4]**:** This dataset from [YK19] features random objects from MSCOCO [LMB$^+$14] pasted onto random scenes from MiniPlaces [ZLK$^+$18]. There is a total of 10 object classes and 10 scene classes. The training dataset contains 90k images, and labels are carefully balanced in a way that prevents confounding. As per [YK19], we introduce two models: F_o^{BAM} (resp. F_s^{BAM}) is trained to identify the object displayed in the image (resp. the background scene). We report results using two state-of-the-art model architectures: EfficientNet-B3 [TL20] and ResNet50 [HZRS15]. We used publicly available [5] models [CDH$^+$20] that were pre-trained on ImageNet and fine-tuned them on BAM. We use analytical solutions from Eq. 5 and Eq. 4 to debug and compute ICS whenever possible.

Metrics

In [YK19], the authors define the Model Contrast Score (MCS, Eq. 6) to evaluate global attribution methods. More specifically, MCS contrasts the attributions obtained for C in two models: model F_1 that has concept C as one of its targets and model F_2 for which concept C is irrelevant. For TCAV, it is written as:

$$\mathrm{MCS}_C := \mathrm{TCAV}_{C,C}(F_1, \mathcal{X}_{\mathrm{TP}}(F_1)) - \max_k \mathrm{TCAV}_{k,C}(F_2, \mathcal{X}_{\mathrm{TP}}(F_2)) \tag{6}$$

where $\mathrm{TCAV}_{k,C}(F, \mathcal{X}_{\mathrm{TP}}(F))$ is the TCAV score for concept C when predicting target k with model F based on $\mathcal{X}_{\mathrm{TP}}(F)$. $\mathcal{X}_{\mathrm{TP}}(F) \subset \mathcal{X}$ is the set of true positives for model F. Conditioning on this subset allows to disentangle the errors of the model from those of the attribution technique. In the case of the BARS dataset, we compute MCS by contrasting the TCAV scores obtained for the "orientation" concept from F_o and from F_c (conversely for the "color" concept). For $\mathrm{TCAV}^{sign(\mathrm{CS})}$, the ideal value of MCS is 0.5: for instance, the TCAV score of the "orientation" concept should be 1 for F_o, but 0.5 for F_c as it is irrelevant for this model. As $\mathrm{TCAV}^{sign(\mathrm{CS})}$ is not bounded in $[0, 1]$, MCS values should be positive, and as large as possible.

3.1.2 Local explanations with ICS

Using BARS, we showcase the effectiveness of the local explanation aspect of ICS. Figure 2a illustrates this on 4 local examples of the BARS dataset for F_o. ICS correctly indicates that the color of the bars does not affect its predictions (score close to 0.), while the orientation concept (which determines the label) are close to 50% for vertical bars and -50% for horizontal bars, reflecting the importance of the presence (vertical) or absence (horizontal) of the concept.

3.1.3 $\mathrm{TCAV}^{\mathrm{ICS}}$ outperforms $\mathrm{TCAV}^{sign(\mathrm{CS})}$ in global explanations

To showcase the global aspect of ICS, we compare $\mathrm{TCAV}^{sign(\mathrm{CS})}$ and $\mathrm{TCAV}^{\mathrm{ICS}}$ for both BARS and BAM datasets. Note that the best MCS score for $\mathrm{TCAV}^{sign(\mathrm{CS})}$ is 0.5, while the larger score the better for $\mathrm{TCAV}^{\mathrm{ICS}}$. We observe that while the MCS score for $\mathrm{TCAV}^{sign(\mathrm{CS})}$ is not close to 0.5 with wide confidence intervals (see Figure 3a), the score for $\mathrm{TCAV}^{\mathrm{ICS}}$ reflects the ground truth. As shown in Figure 3a,b, $\mathrm{TCAV}^{\mathrm{ICS}}$ is typically bimodal (either 0 or 1) for irrelevant concepts in a statistically significant manner (i.e., not close to the chance-level 0.5). On the other hand, $\mathrm{TCAV}^{\mathrm{ICS}}$ gives

[4] `https://github.com/google-research-datasets/bam`, Apache 2 license

[5] `www.tensorflow.org`, Apache 2 license

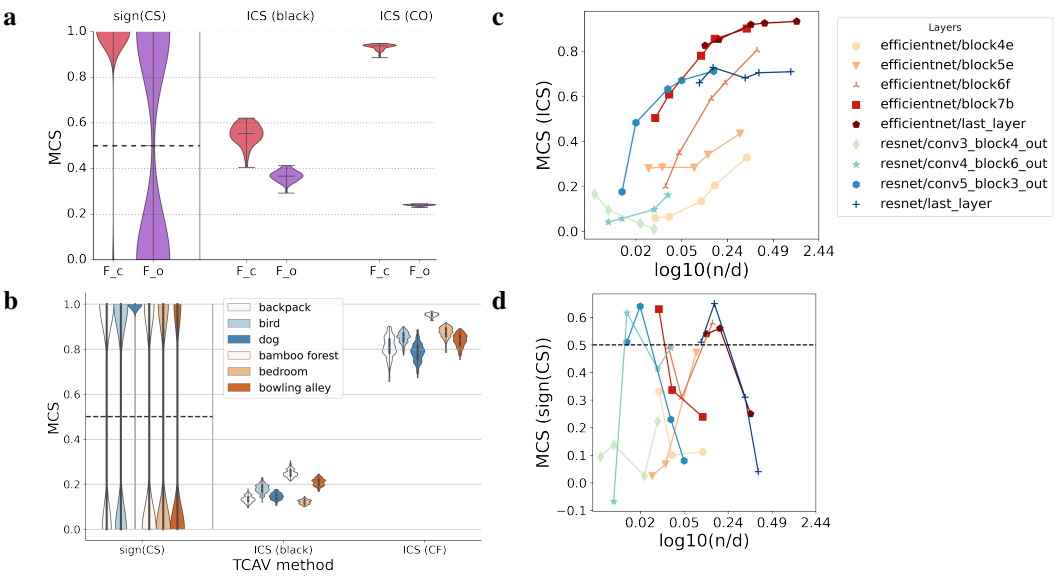

Figure 3: **MCS for** $\text{TCAV}^{sign(\text{CS})}$ **and** TCAV^{ICS}**. a** MCS distributions for BARS, computed with a black and concept-occluding (CO) baseline for ICS, for F_c (red) and F_o (purple). The dashed line represents the 0.5 ground truth for $\text{TCAV}^{sign(\text{CS})}$. **b** MCS at the last layer of EfficientNet with a black image and concept-forgetting (CF) baseline, for 3 objects and 3 scenes in BAM. **c** Varying n for training CAVs across different layers of EfficientNet (orange to red) and ResNet (green to blue) improves MCS for TCAV^{ICS} for a concept-forgetting baseline. **d** In contrast, MCS($\text{TCAV}^{sign(\text{CS})}$) scores fail to consistently improve.

irrelevant CAVs a score close to 0 consistently, and positive scores to relevant concepts. Therefore, it seems to be a more reliable estimate of concepts' global importance, with the caveats discussed in Section 3.1.4.

3.1.4 Improving the effect of the curse of dimensionality on CAVs

When scaling to larger models with small n (number of concept pictures), naively applying ICS can lead to degraded results (i.e. values <0.1%). Given that the degradation is more acute for shallow layers which tend to be wider, we hypothesize that these results are due to the curse of dimensionality, as many different directions can encode a concept. To test this hypothesis, we decrease the $\frac{n}{d}$ ratio from 30 to 0.1 to train CAVs on BARS, resulting in MCS scores 3 times smaller (45% vs 15%). We are able to mitigate this by augmenting the CAV training data with widely used image transformations such as random flips and changes of contrast, brightness, hue, and saturation. With these modifications, MCS scores on BAM can be improved for most layers (more impact on the deeper layers, Figure 3c). This is intuitive; one should get 'better' quality of CAVs as you increase n. In contrast, $\text{TCAV}^{sign(\text{CS})}$ was not consistently improved by the augmentation, given its formulation based on the sign of CS as shown in Figure 3d.

3.2 Qualitative illustration

We use the ImageNet dataset [DDS+09] to generate some of the concepts demonstrated in [KWG+18] and replicate results at the global level. We experiment with multiple network architectures: EfficientNet (presented in the main text, [TL20]), ResNet50 [HZRS15], Inception-v1 [SLJ+15], and MobileNet-V2 [SHZ+18] which produced consistent results. We report results using a white image as the baseline, but a black baseline also produced consistent results. Alongside with global results, we showcase local examples that demonstrate the intuitive nature of ICS scores.

Zebra class: We first focus on the "zebra" class and build CAVs to represent the "striped", "dotted", "zigzagged", "zebras", and "horse" concepts to draw similarity and contrast with the analyses done in [KWG+18]). TCAV^{ICS} (Figure 4a) is able to replicate the high $\text{TCAV}^{sign(\text{CS})}$ scores for the striped concept, as well as for the "zebra" concept (as a sanity check) and for the "horse" concept

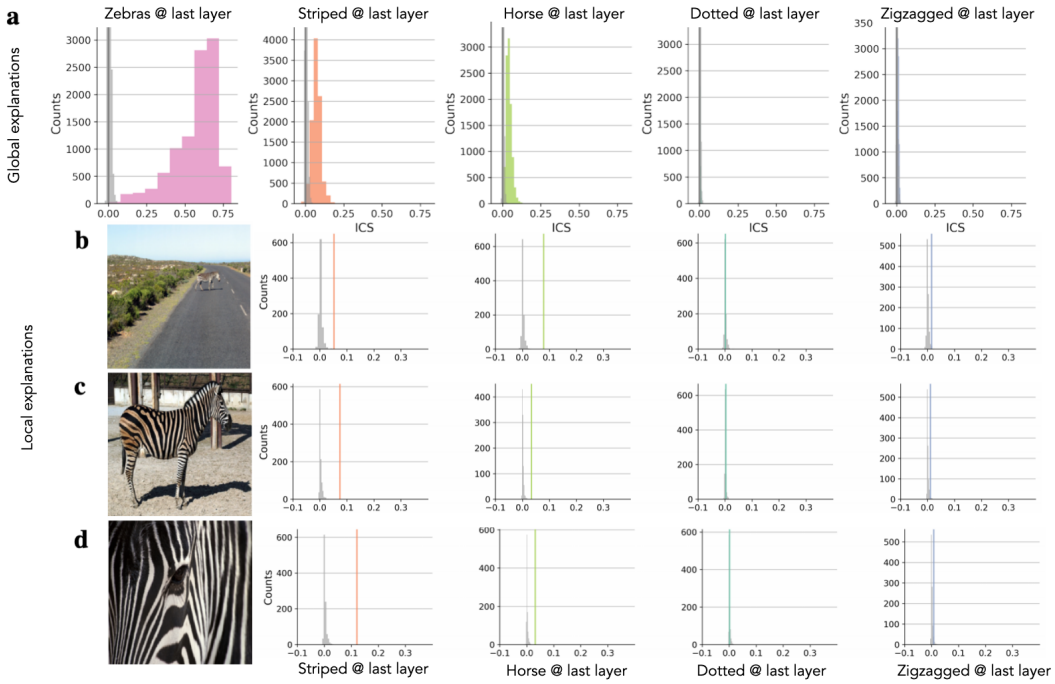

Figure 4: **ICS scores for the "zebra" class.** Distribution of ICS scores across 100 pictures of zebra, for 5 concepts (zebra, striped, horse, dotted, zigzagged), computed on the last layer of an EfficientNet-B3 model. Gray histograms represent null distributions obtained by CAV trained on permuted labels. **a** Global scores. **b-d** 3 cherry picked samples, representing far and close distance between the camera and the zebra. The zebra becomes more "stripy" when it is more zoomed-in.

(more detail on CAV building in Supplement). Similar to [KWG+18], "zigzagged" and "dotted" have significantly small $\mathrm{TCAV}^{sign(\mathrm{CS})}$ scores (0.0002 and 0.1760, respectively), which reflects their negative influence on the model prediction for the class "zebra". We observe one important difference: contrarily to $\mathrm{TCAV}^{sign(\mathrm{CS})}$, the ICS distributions for "dotted" and "zigzagged" overlap with their associated null distribution (grey histogram in Figure 4) at both global (Figure 4a) and local levels (Figure 4b-d). This potentially hints that ICS is better at separating the meaningful concepts from null. Please see the Supplement for more global results for 3 other architectures.

Figure 4b-d demonstrate the local explanation aspect of ICS. When the zebra is far away from the camera (b), it may look less stripy, more like a horse than if the zebra is close to the camera. This is well-reflected with ICS scores. As we *zoom in* on the zebra ((c) and (d)), stripes are obtaining higher ICS scores, while the "horse" concept becomes less important in the model's decision. Being able to map a complex network's decision to what humans can understand with this level precision is encouraging. Note that two irrelevant concepts, 'zigzagged' and 'dotted' have their ICS distributions closer to the null at all camera distances.

Basketball class: As a second illustration, we investigate the "basketball" game class and define 4 concepts that are directly related to the game: the ball, the jersey, the floor, and the hoop. As racial biases have been highlighted in [KWG+18] for this class, we also build "gender" (man vs woman) and "race" (darker vs lighter skin) concepts. All CAVs are estimated as "significantly encoded" in the last layer of the model, with performance of the linear classifier $\sim 99\%$ for the game related concepts, 94% for gender and 75% for race. $\mathrm{TCAV}^{sign(\mathrm{CS})}$ scores for the game-related concepts are 1, while they are 0.85 and 0.91 for gender and race, respectively. ICS scores however display a "ranking" of concepts in terms of global scores: the ball has the largest impact, followed by the hoop, then floor and jersey (Figure 5). On the contrary, ICS distributions for race and gender are indistinguishable from the nulls. We qualitatively tested a few manually selected images with strong minority attributes, and observed that this network is able to correctly predict the "basketball" class. Note that this is not a proof that this model is not biased.

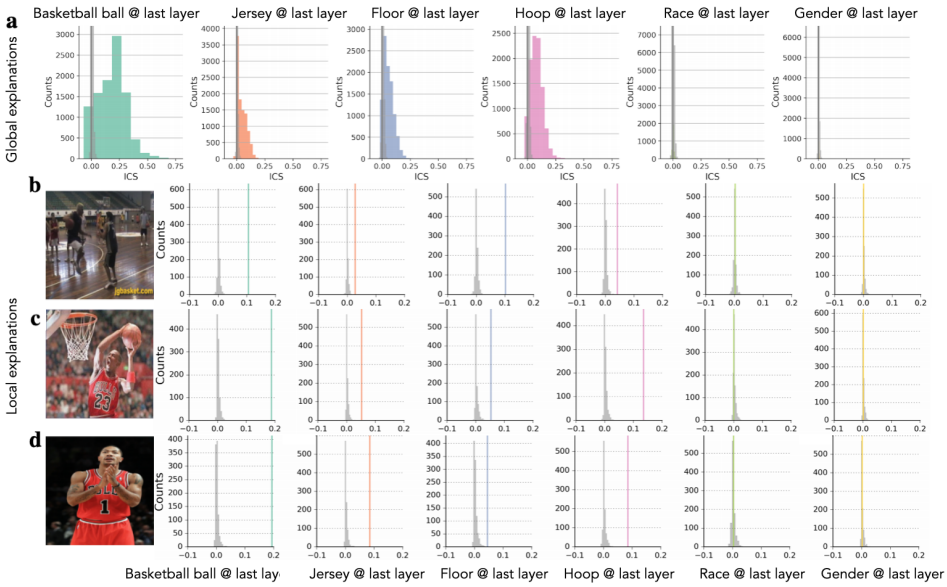

Figure 5: **ICS scores for the "basketball" class.** Distribution of ICS scores (white image baseline) across 100 pictures of basketball, for 5 concepts (ball, jersey, floor, hoop, race, gender), computed on the last layer of an EfficientNet-B3 model. Gray histograms represent null distributions obtained by CAV trained on permuted labels. **a** Global scores. **b-d** 3 cherry picked samples displaying different elements.

When investigating local examples, we observe that the ICS scores for the ball are high (Figure 5b-d) in general, consistent with the global explanation. It turns out that ImageNet images for the "basketball" class include many images that only contain the ball, making the ball and the basketball game highly correlated. ICS is again able to describe fine grain level of local explanations: it separates an image with less visible jersey (b) to more visible jersey (d), less floor (c, d) to more floor (a), and the existence of the hoop (c) correctly.

4 Discussion

TCAV [KWG+18] sheds light into deep neural networks inner workings by allowing users to probe their learnt internal representations. This approach has shown promising results at the global level, that have been made stronger by the rigorous evaluation of [YK19]. By combining TCAV with integrated gradients [STY17], we show that TCAV can provide *local explanations*. This combination hence opens a novel avenue for both techniques.

Our work also allowed to uncover a limitation of TCAV and improve it. We demonstrated that $\text{TCAV}^{sign(\text{CS})}$ might not be faithful, with scores for irrelevant concepts flipping from 0 to 1 across bootstraps. TCAV^{ICS} consistently provided more sensitive global explanations on two synthetic datasets and multiple networks.

Limitations: While we obtain the 'best of both worlds', our technique still inherits the limitations of both methods. More specifically, building reliable CAVs is crucial for TCAV methods to be trusted, which can be challenging in high-dimensionality or due to confounding. Confounding can be mitigated by careful inspection of the training set, or by using causal approaches [GSK19, BH20]. Regularizing the CAV, considering smaller layers, or increasing the CAV training set size can help in higher dimensions.

For IG, a baseline needs to be defined, which affects the obtained explanation. In our work, we have observed variability in the support of ICS in BARS across baselines, and different levels of improvements in terms of MCS when using augmentation to alleviate the curse of dimensionality on CAVs. As discussed in multiple previous works (e.g. [STY17, AÖG19, SLL20]), results are expected to vary across baselines as each baseline represents a specific definition of 'missingness'. It would however be interesting to consider recent approaches to baselines such as aggregations over multiple

baselines, or reformulations of integrated gradients [EJS+20]. Results should also be evaluated in terms of robustness to adversarial attacks [AJ18, DAA+19, GAZ17, YHS+19], as [DAA+19] have shown that integrated gradients are highly sensitive to this kind of attack. We leave this evaluation for future work.

At the local level on ImageNet, we observe that ICS can give intuitive results. However, these results are prone to confirmation bias and only a human-grounded evaluation [DVK18] can assess the usefulness of local TCAV explanations. Such user-study could further investigate how concept-based explanations could potentially be combined with feature-based attributions to provide a full picture to the user. Indeed, TCAV only covers a limited set of concepts. It is therefore possible that specific aspects of an image leading to a prediction (true or false) would not be covered by this set. In this case, complementing the concept-based by feature-based attributions could be interesting.

In conclusion, we hope that our work demonstrates the potential and pitfalls of combining local and global explanation techniques for model interpretability, and will encourage further work in this direction.

Acknowledgements

We would like to thank Mengjiao Yang for sharing BAM models, and Yash Goyal for sharing the synthetic data generation, as well as Mahima Pushkarna for helping with figures. We also thank collaborators in Google Research, Search and in Google Health.

References

[ACÖG18] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *Proceedings of the 2018 International Conference on Learning Representations (ICLR)*, 2018.

[AGM+18] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 9505–9515, 2018.

[AJ18] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

[ALSA+17] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *J. Mach. Learn. Res.*, 18(1):8753–8830, January 2017.

[AÖG19] Marco Ancona, Cengiz Öztireli, and Markus Gross. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[BBM+15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), jul 2015.

[BCB16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[BH20] Mohammad Taha Bahadori and David E. Heckerman. Debiasing concept bottleneck models with instrumental variables. *CoRR*, abs/2007.11500, 2020.

[BOGO15] François-Xavier Briol, Chris Oates, Mark Girolami, and Michael A Osborne. Frank-wolfe bayesian quadrature: Probabilistic integration with theoretical guarantees. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1162–1170. Curran Associates, Inc., 2015.

[CDH+20] Chen Chen, Xianzhi Du, Le Hou, Jaeyoun Kim, Jing Li, Yeqing Li, Abdullah Rashwan, Fan Yang, and Hongkun Yu. Tensorflow official model garden, 2020.

[CLT+19] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[COPA+19] James R. Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink, Andrew P. King, and Julia A. Schnabel. *Global and local interpretability for cardiac MRI classification*, pages 656–664. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). SPRINGER, January 2019. 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019 ; Conference date: 13-10-2019 Through 17-10-2019.

[DAA+19] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame, 2019.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[DV17] Been Doshi-Velez, Finale; Kim. Towards a rigorous science of interpretable machine learning. In *eprint arXiv:1702.08608*, 2017.

[DVK18] Finale Doshi-Velez and Been Kim. *Considerations for Evaluation and Generalization in Interpretable Machine Learning*, pages 3–17. Springer International Publishing, Cham, 2018.

[EEG+18] Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018.

[EJS+20] Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients, 2020.

[GAM18] Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression concept vectors for bidirectional explanations in histopathology. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva, Sergio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132, Cham, 2018. Springer International Publishing.

[GAZ17] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile, 2017.

[GLW+20] Gary S. W. Goh, Sebastian Lapuschkin, Leander Weber, Wojciech Samek, and Alexander Binder. Understanding integrated gradients with smoothtaylor for deep neural network attribution. *CoRR*, abs/2004.10484, 2020.

[GSK19] Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *CoRR*, abs/1907.07165, 2019.

[HPRPC20] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1096–1106, 2020.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[JW19] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 3543–3556, 2019.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[KHA+19] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham, 2019.

[KNT+20] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020.

[KWG+18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *35th International Conference on Machine Learning, ICML 2018*, volume 6, pages 4186–4195, 2018.

[LEC+19] Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable AI for trees: From local explanations to global understanding. *CoRR*, abs/1905.04610, 2019.

[LHK19] I. V. D. Linden, H. Haned, and E. Kanoulas. Global aggregations of local explanations for black box models. *ArXiv*, abs/1907.03039, 2019.

[Lip18] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[MLH+21] Diana Mincu, Eric Loreaux, Shaobo Hou, Sebastien Baur, Ivan Protsyuk, Martin Seneviratne, Anne Mottram, Nenad Tomasev, Alan Karthikesalingam, and Jessica Schrouff. *Concept-Based Model Explanations for Electronic Health Records*, page 36–46. Association for Computing Machinery, New York, NY, USA, 2021.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 1135–1144, 2016.

[SGK17] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pages 4844–4866. International Machine Learning Society (IMLS), apr 2017.

[SGL19] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why modified bp attribution fails. *arXiv preprint arXiv:1912.09818*, 2019.

[SGSK16] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv*, 1:0–5, may 2016.

[SHZ+18] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[SLL20] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the Impact of Feature Attribution Baselines. *Distill*, 5(1), jan 2020.

[STK+17] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[STR+19] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B Sood, Greg S Corrado, Lily Peng, and Dale R Webster. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*, 126(4):552–564, 2019.

[STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pages 5109–5118, 2017.

[SVZ14] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[TJMG19] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Proceedings of Machine Learning Research*, pages 1 – 21, 2019.

[TL20] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[UCS17] Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of tr(f(a)) via stochastic lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1075–1099, January 2017.

[WSC+20] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2020.

[YHS+19] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10967–10978. Curran Associates, Inc., 2019.

[YK19] Mengjiao Yang and Been Kim. BIM: towards quantitative evaluation of interpretability methods with ground truth. *CoRR*, abs/1907.09701, 2019.

[ZF14] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[ZLK+18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

A Methods

A.1 Choice of the baseline

It has been suggested [KHA+19, ACÖG18] that the attributions may not be invariant to the choice of the baseline used in the computation of integrated gradients (IG, [STY17]). In this section, we display the baselines considered for investigation throughout this work, while their influence on the proposed method is extensively studied in sections B.4 and C.4. We divide the baselines based on their 'informativeness'.

A.1.1 Uninformative baselines

As in [STY17], the baseline is typically chosen to be uninformative. This characteristic is fuzzy and can be defined in various ways. [SLL20] provides interactive visualization of the impact of this choice on the resulting saliency maps.

Below, you can find a list of various ways of defining 'uninformative'.

- *Zero image baseline*, i.e. black image: $\mathbf{a}' = f_l(\mathbf{0})$
- *One image baseline*, i.e. white image: $\mathbf{a}' = f_l(\mathbf{1})$
- *Noisy image baseline*: $\mathbf{a}' = f_l(\mathcal{N})$ where $\mathcal{N}$ is a noise distribution.
- *Pixel-wise average baseline*: $\mathbf{a}' = f_l(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})$
- *Pixel-wise median baseline*: $\mathbf{a}' = f_l(\text{median}_n\mathbf{x_n})$
- *Entropy-maximizing baseline*: $\mathbf{a}' = \text{argmin}_{\mathbf{x}}||\mathbf{a}-\mathbf{x}|| + \lambda\mathcal{H}(h(\mathbf{x}))$ where $\lambda \in \mathbb{R}+^*$ is some large number, $\mathcal{H}(h(\mathbf{x}))$ is the entropy of the output of the neural network when activations at hidden layer l are equal to $\mathbf{x}$.

The main text mostly reports results using the 'One image baseline' (i.e. a black or white image input).

A.1.2 Informative baselines

- *Concept-occluding baseline*: $\mathbf{a}' = \mathbf{a} - (\mathbf{a}^T\mathbf{v_C} + b)\frac{\mathbf{v_C}}{||\mathbf{v_C}||_2^2}$: the baseline with the concept removed, where b is the bias of the linear model. In that case, ICS is equal to the occluded concept prediction difference.
- *Concept-forgetting baseline*: Similarly to the previous baseline, we can define $\mathbf{a}' = \mathbf{a} - \lambda\mathbf{v_C}$ for some $\lambda \in \mathbb{R}^*+$, which is then a generalization.

Please note that $\mathbf{v_C}$ is not assumed to be unit-normed for informative baselines.

In this work, we present results for the 'Concept-forgetting baseline' as this baseline leads to a closed form solution for ICS (see Sec.A.2). We however note the presence of a hyper-parameter λ that influences the results: small values for λ mean that the prediction won't change much and hence lead to small ICS scores for all values, while large values for λ mean that predictions can be changed so much that all directions are considered as 'relevant'. We leave the exploration of the tuning of this hyperparameter for future work.

Finally, while some baselines 'remove' the concept, this is not equivalent to estimating the causal effect of the concept on the model's output, as suggested by [GSK19]. They propose to estimate the causal effect of a concept by generating counterfactuals with a conditional VAE to emulate the *do* operator. While the concept-forgetting baseline is the closest activation displaying as much opposing concept as there is concept in the original activation (for $\lambda = 2(a^T v_C + b)/||v_C||_2^2$), there is no guarantee that there exists a corresponding counterfactual input mapping to this baseline, nor that this input is a good approximation for the *do* operation. In our experiments, we observed that it is possible to find a visually identical input mapping to the symmetric activations (the concept prediction is flipped), with the model's prediction being unchanged.

A.2 Analytical forms of ICS

In some cases, it is possible to derive an analytical formula for ICS. We present the derivations of the formulations proposed in the main text below, using the notation defined in the main text.

A.2.1 Last layer of a binary classification model with entropy-maximizing baseline

In the binary case, we have

$$h : \mathbf{a} \to \sigma(\mathbf{w}^T\mathbf{a} + b)$$

where σ is the sigmoid function and $\mathbf{w}^T\mathbf{a} + b$ represents the logits. Therefore,

$$\nabla_{\mathbf{v_C}} h(\mathbf{a}) = \sigma'(\mathbf{w}^T\mathbf{a} + b)\mathbf{w}^T\mathbf{v_C}$$

where σ' is the derivative of σ. Note that the k index is dropped since the model has a single output.

We select a baseline $\mathbf{a}'$ that maximizes the entropy of the prediction, as suggested in [STY17]. This means that

$$h(\mathbf{a}') = 0.5$$

and $\mathbf{a}'$ is the orthogonal projection of $\mathbf{a}$ on the decision boundary $\mathbf{w}^T\mathbf{x} + b = 0$, i.e.

$$\mathbf{a}' := \mathbf{a} - (\mathbf{w}^T\mathbf{a} + b)\frac{\mathbf{w}}{||\mathbf{w}||_2^2}$$

With the change of variable $u := \mathbf{w}^T\mathbf{a}' + b + \alpha\mathbf{w}^T(\mathbf{a} - \mathbf{a}') = \alpha\mathbf{w}^T(\mathbf{a} - \mathbf{a}')$ in the ICS formulation, we obtain:

$$\begin{aligned}
\mathrm{ICS}_C(\mathbf{a}, \mathbf{a}') &= \mathbf{v_C}^T(\mathbf{a} - \mathbf{a}') \int_0^{\mathbf{w}^T\mathbf{a}+b} \sigma'(u)\frac{\mathbf{w}^T\mathbf{v_C}}{\mathbf{w}^T(\mathbf{a} - \mathbf{a}')}du \\
&= \frac{\mathbf{v_C}^T(\mathbf{a} - \mathbf{a}')\mathbf{w}^T\mathbf{v_C}}{\mathbf{w}^T(\mathbf{a} - \mathbf{a}')}\Big(\sigma(\mathbf{w}^T\mathbf{a} + b) - 0.5\Big) \qquad (7)
\end{aligned}$$

Noting that $\mathbf{a} - \mathbf{a}' = (\mathbf{w}^T\mathbf{a} + b)\frac{\mathbf{w}}{||\mathbf{w}||_2^2}$, Eq. 7 can be rewritten as:

$$\begin{aligned}
\mathrm{ICS}_C(\mathbf{a}, \mathbf{a}') &= \frac{\mathbf{v_C}^T\mathbf{w}(\mathbf{w}^T\mathbf{a} + b)\mathbf{w}^T\mathbf{v_C}}{||\mathbf{w}||_2^2(\mathbf{w}^T\mathbf{a} + b)}\Big(\sigma(\mathbf{w}^T\mathbf{a} + b) - 0.5\Big) \\
&= \left(\frac{\mathbf{v_C}^T\mathbf{w}}{||\mathbf{w}||_2}\right)^2\Big(\sigma(\mathbf{w}^T\mathbf{a} + b) - 0.5\Big)
\end{aligned}$$

The conceptual sensitivity in that situation is given by $\mathrm{CS}_C = \mathbf{w}^T\mathbf{v_C}$. In contrast, ICS depends on the predicted probability for the considered input while not depending on the norm of $\mathbf{w}$, hence alleviating the aforementioned concerns. Note that the dependency on the square of the cosine similarity between the CAV and the model's last layer weights means that a poor estimate of $\mathbf{v_C}$ could result in a dramatically small value.

A.2.2 Multi-class model, with concept-forgetting baseline

Let's consider a multi-class model (i.e. h_k is the k-th output of the softmax), with a concept-occluding baseline. In this section, $\mathbf{v_C}$ is not assumed unit-normed. This baseline is the orthogonal projection on the hyperplane orthogonal to the CAV, i.e., assuming that the bias is 0 for compactness, it is defined by

$$\mathbf{a}' := \mathbf{a} - \lambda\mathbf{v_C},\ \lambda \in \mathbb{R}$$

For $\lambda = 2$, it is akin to occlusion [ZF14] for concepts.

Let

$$\mathcal{B} := (\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_d})$$

be an orthonormal basis of $\mathbb{R}^d$ where $\mathbf{e_1}$ is chosen to be $\mathbf{v_C} \in \mathbb{R}^d$. We define $\mathrm{IG}_i^k(\mathbf{a}, \mathbf{a}')$ as the integrated gradient for the i-th feature. Similarly, $\mathrm{ICS}_C^k(\mathbf{a}, \mathbf{a}')$ designates the ICS.

As mentioned in Related works, integrated gradients verify completeness (Eq.8), which means that:

$$F_k(\mathbf{x}) - F_k(\mathbf{x}') = \sum_i \mathrm{IG}_i^k(\mathbf{x}, \mathbf{x}') \tag{8}$$

By virtue of completeness, we have that:

$$h_k(\mathbf{a}) - h_k(\mathbf{a}') = \sum_{1 \leq i \leq d} \mathrm{IG}_i^k(\mathbf{a}, \mathbf{a}') \tag{9}$$

Each IG_i^k is the product of two terms, one being the dot product between $\mathbf{a} - \mathbf{a}'$ and $\mathbf{e_i}$. By definition, $\mathbf{a} - \mathbf{a}'$ is colinear to $\mathbf{e_1}$ and $\mathcal{B}$ is orthonormal, therefore

$$(\mathbf{a} - \mathbf{a}')^T \mathbf{e_i} = 0, \ \forall i > 1$$

Therefore, $\mathrm{IG}_i^k = 0 \ \forall i > 1$ and Eq. 9 becomes:

$$h_k(\mathbf{a}) - h_k(\mathbf{a}') = \mathrm{IG}_1^k(\mathbf{a}, \mathbf{a}')$$

Given $\mathbf{e_1} = \mathbf{v_C}$, ICS can be rewritten:

$$\mathrm{ICS}_C^k(\mathbf{a}, \mathbf{a}') = \mathrm{IG}_1^k(\mathbf{a}, \mathbf{a}')$$

And by transitivity:

$$\mathrm{ICS}_C^k(\mathbf{a}, \mathbf{a}') = h_k(\mathbf{a}) - h_k(\mathbf{a}')$$

i.e. ICS is equal to the difference in model probability between the model's prediction for this sample and the prediction should the concept be removed.

B Additional results on BARS

B.1 MLP model F_o does not rely on the bar's color for its predictions

In the BARS dataset, it is possible to evaluate how much the model relies on a given concept by evaluating the model on counterfactual examples [GSK19]. We define g_C as follows:

$$g_C(\mathbf{x}) := \max_{(C_1, C_2) \in \mathcal{C}^2} |F(\mathbf{x}_{C_1}) - F(\mathbf{x}_{C_2})| \tag{10}$$

where x_C is a counterfactual version of x where its concept is set to C, and $\mathcal{C}$ is the set of possible values for the concept of interest. In the present case, C_1 and C_2 represent the color and orientation concepts.

This quantity can be aggregated over the entire test set to obtain the global influence of a concept on the model's output:

$$G_C := \mathbb{E}_{\mathbf{x}}\big(g_C(\mathbf{x})\big) \tag{11}$$

We verify that F_o is almost invariant to the color of the bar: on the test set, we obtain $G_{\mathcal{C}}(F_o) = 0.0011$.

Figure 6 shows the output of the model on two random samples (vertical and horizontal) when slowly changing the color from green to red.

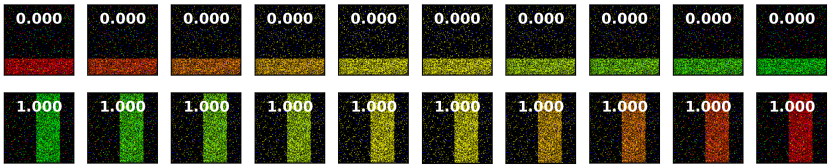

Figure 6: Vertical and horizontal bars with varying colors and the associated predicted probabilities (F_o) displayed.

B.2 Classification performance of F_o and F_c

The BARS dataset being very simple, our models (trained to detect either the color or the orientation) reach 100% test accuracy after only a few epochs.

B.3 Statistical significance and classification performance of trained CAVs

Table 1 displays the ROCAUC on held-out data (mean and standard deviation across bootstraps) of the CAVs for both concepts, and both models.

	F_o		F_c	
layer	Color	Orientation	Color	Orientation
1	100 (0)	100 (0)	100 (0)	59.1 (5.6)
2	100 (0)	100 (0)	100 (0)	54.0 (4.6)
3	100 (0)	100 (0)	100 (0)	50.8 (3.7)

Table 1: ROCAUC of the bootstrapped CAVs. Average value rounded to closest decimal, and standard deviation in parentheses.

Note that while the color concept is not being used by F_o, its CAVs have perfect discriminative performance for identifying that concept in the activation spaces. However, the orientation concept is impossible to identify with a linear classifier in the activation spaces of F_c.

B.4 Influence of the baselines

In the BARS dataset, we computed the predicted probability associated with all baselines (see Table 2 for a selected sample). Pixel-wise average and entropy-maximizing baselines seem to be truly uninformative: the output of the model is a tie. This results in bimodal ICS distributions clustering around -50% and 50% (Figure 7a,b).

If uninformative means having a 50% predicted risk, then it is clear that some of these baselines are not uninformative. The iid Gaussian noise $\mathcal{N}(0, 1)$ in the pixel space leads to a bimodal distribution that has two modes in 0 and in 1, meaning that this seemingly meaningless noise is interpreted very confidently by the model. Similarly, the white baseline leads to a very confident prediction of 'vertical' for F_o.

In terms of ICS, we observe that different 'informative' baselines (as interpreted by the model) lead to different supports for ICS. The difference across concepts, hence MCS, is however relatively consistent across baselines (Table 3).

	layer 1	layer 2	layer 3
i.i.d. $\mathcal{N}(0, 1)$ pixels	60% (41%)	60% (41%)	60% (41%)
Black image (zero)	62%	62%	62%
Average activation	1%	1%	6%
Average pixel-wise	66%	66%	66%
Entropy-maximizing	50% (0.1%)	50% (0.1%)	50% (0.1%)
White image (zero)	100%	100%	100%

Table 2: Predicted probability of the bar being vertical for several baselines. Standard deviation is provided in parenthesis when it is not 0.

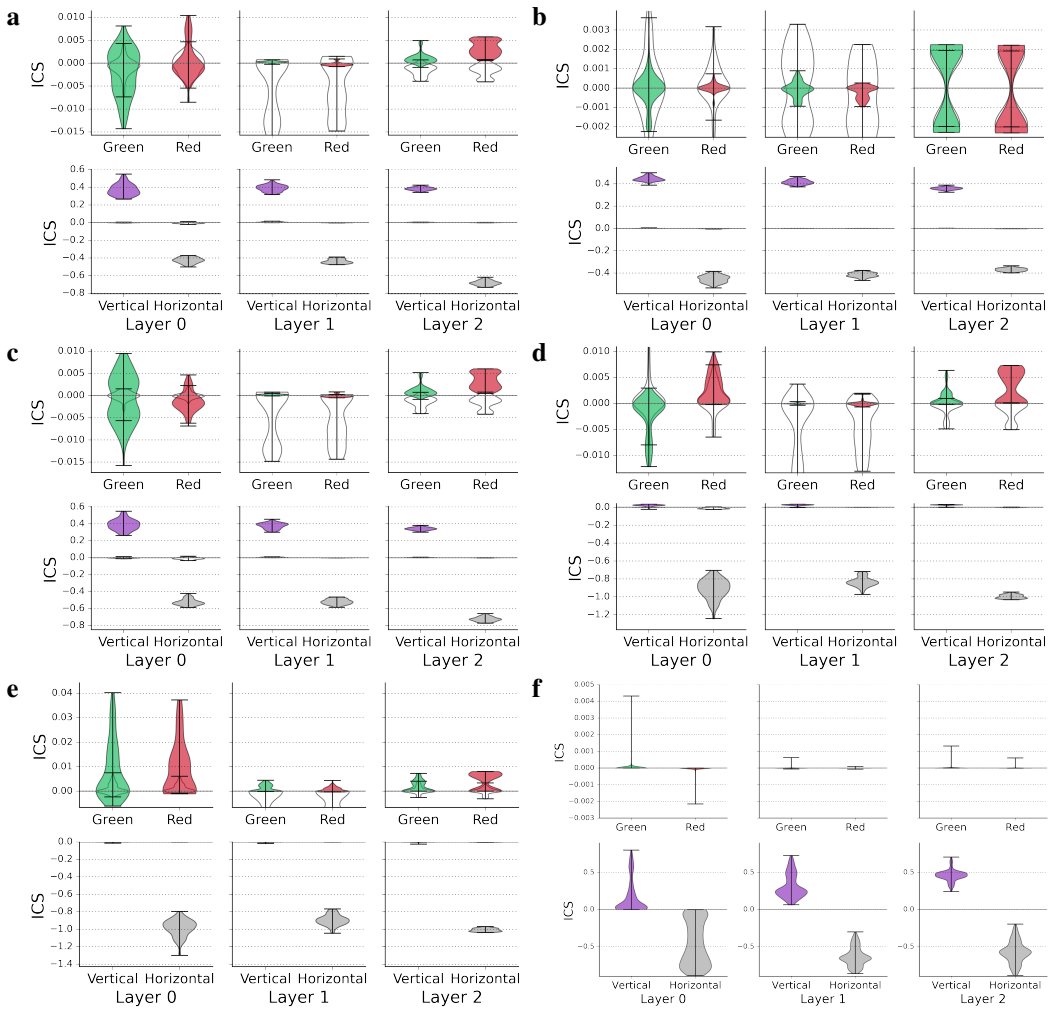

Figure 7: **Distributions of ICS for F_o^{BARS}. a** Black baseline. **b** Entropy-maximizing baseline. **c** Pixel-wise average baseline. **d** Noisy baseline, i.e. the baseline is a sample of $\mathcal{N}(0, 1)$. **e** White image baseline. **f** Concept-occluding baseline.

concept	layer	MCS(TCAV$^{sign(\text{CS})}$)	MCS(TCAV$^{\text{ICS}}$) black	MCS(TCAV$^{\text{ICS}}$) max ent.
orientation	0	0.00 [0.00, 1.00]	0.36 [0.31, 0.41]	0.46 [0.43, 0.47]
orientation	1	0.00 [0.00, 1.00]	0.33 [0.27, 0.39]	0.47 [0.45, 0.48]
orientation	2	1.00 [0.00, 1.00]	0.42 [0.37, 0.48]	0.35 [0.34, 0.36]
color	0	0.93 [0.00, 1.00]	0.41 [0.33, 0.47]	0.47 [0.45, 0.47]
color	1	0.00 [0.00, 0.00]	0.45 [0.38, 0.50]	0.36 [0.33, 0.37]
color	2	1.00 [0.00, 1.00]	0.56 [0.46, 0.64]	0.34 [0.33, 0.34]

Table 3: Median MCS scores (%) computed on the entire BARS test set, with relevant 95% bootstrapped non-parametric confidence intervals.

C Additional results on BAM

C.1 F_o^{BAM} does not rely on the scene background, and F_s^{BAM} does not rely on the pasted object

Using a similar approach [YK19] showed that the accuracy of the models is no better than a random guess when the relevant concept are removed from the images.

C.2 Classification performance of F_o and F_s

For BAM, we used pre-trained ResNet50 [HZRS15] (74.9% top-1 accuracy) and EfficientNet-B3 [TL20] (81.6% top-1 accuracy) available at `www.tensorflow.org`. They were fine-tuned on the BAM dataset to predict either scenes (reaching 91% test top-1 accuracy) or objects (reaching 80% test top-1 accuracy) by iteratively updating the weights of the final dense layers, followed by convolutional stacks, one after the other with Adam optimizer (learning rate 10^{-4}) [KB14] until convergence of the validation accuracy (about 1-5 epochs). Training is performed with classical data augmentations (random flips, contrast, brightness, hue, saturation changes). Standard pre-processing techniques were applied to the images (see `https://www.tensorflow.org/api_docs/python/tf/keras/applications/imagenet_utils/preprocess_input`).

C.3 Statistical significance and classification performance of trained CAVs

We trained CAVs using ElasticNet regularization [ZH05], which resulted in sparse weights and best classification performance on held-out samples (compared to Ridge and Lasso regression). Regularization coefficient is estimated via cross-validation. CAVs are not all statistically significant, as assessed by a permutation test. Figure 8 displays the predictive performance of CAVs for both EfficientNet-B3 models, 6 layers, and 6 concepts. Shallow CAVs, which weren't fine-tuned and kept Imagenet-learned weights, tend to be very good at identifying scenes for both models. They however cannot identify objects. Deeper CAVs keep a high discriminative performance for scenes for both F_o and F_s. Only deep CAVs of F_o reach high discriminative performance for objects. We observed similar results for ResNet50.

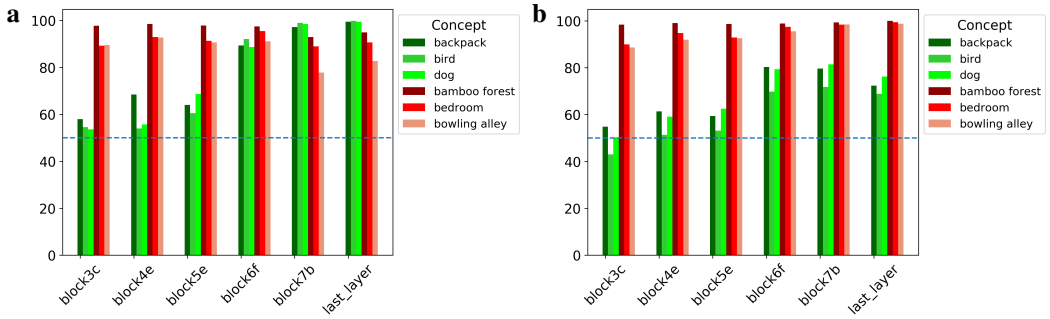

Figure 8: ROCAUC of EfficientNet's CAVs for 6 concepts, estimated on heldout data. **a.** For model F_o. **b.** For model F_s.

C.4 Influence of the baseline

While the choice of the baseline mostly affected the support of ICS for BARS, we observe that the scale of the results can be affected for BAM, whereby small values of ICS are observed for the relevant concepts. This phenomenon is more pronounced for shallow layers of the models (Table 4), and is variable across baselines (Table 5, Figure 9). We hypothesize that these results are due to (the combination of) two factors: the higher dimensionality of the CAVs, and a deterioration in the quality of the projection of IG onto the CAV (variable across baselines). We explore the former in section C.5.

$\mathrm{TCAV}^{\mathrm{ICS}}$ leads to higher scores for deeper layers of the network, with narrower confidence intervals. On the other hand, MCS scores for $\mathrm{TCAV}^{sign(\mathrm{CS})}$ have wider confidence intervals, especially for deeper layers.

C.5 Curse of dimensionality

We hypothesize that the divergence between the directionality of bootstrapped CAVs increases with the dimensionality d of the space, or more accurately, with the decrease in the ratio $\frac{n}{d}$ where n is the number of samples used to train the CAV. This increase in variance of the CAV directionality leads to smaller dot products (the average cosine similarity between two vectors scales with $d^{-\frac{1}{2}}$)

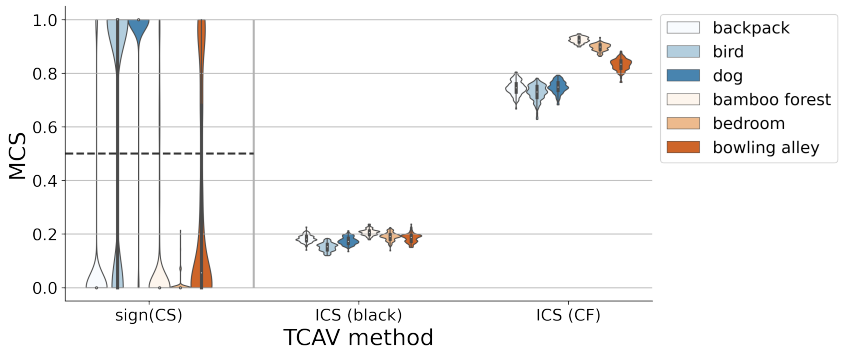

Figure 9: MCS for $\mathrm{TCAV}^{sign(\mathrm{CS})}$ and for $\mathrm{TCAV}^{\mathrm{ICS}}$ for 3 objects and 3 scenes in BAM using ResNet. The dashed line represents the 0.5 ground truth for $\mathrm{TCAV}^{sign(\mathrm{CS})}$.

concept	layer	$\mathrm{MCS}(\mathrm{TCAV}^{sign(\mathrm{CS})})$ (%)	$\mathrm{MCS}(\mathrm{TCAV}^{\mathrm{ICS}})$ (%)
backpack	conv3	0.21 (-0.11, 0.61)	-0.03 (-0.24, 0.17)
bird	conv3	0.15 (-0.26, 0.47)	0.09 (0.01, 0.21)
dog	conv3	0.04 (-0.36, 0.51)	-0.00 (-0.19, 0.23)
bamboo forest	conv3	0.36 (0.08, 0.62)	0.25 (-0.01, 0.54)
bedroom	conv3	0.55 (0.32, 0.69)	0.10 (-0.08, 0.29)
bowling alley	conv3	0.29 (-0.08, 0.52)	0.11 (-0.14, 0.33)
backpack	conv4	0.42 (-0.19, 0.86)	0.09 (-0.03, 0.24)
bird	conv4	0.70 (-0.02, 0.96)	0.06 (-0.07, 0.19)
dog	conv4	0.37 (-0.02, 0.91)	0.07 (-0.07, 0.21)
bamboo forest	conv4	0.52 (0.17, 0.76)	0.53 (0.35, 0.68)
bedroom	conv4	0.42 (0.17, 0.75)	0.44 (0.17, 0.68)
bowling alley	conv4	0.56 (0.26, 0.86)	0.31 (0.08, 0.57)
backpack	conv5	1.00 (-0.00, 1.00)	0.61 (0.45, 0.71)
bird	conv5	1.00 (-0.00, 1.00)	0.49 (0.30, 0.62)
dog	conv5	1.00 (-0.00, 1.00)	0.48 (0.36, 0.65)
bamboo forest	conv5	0.00 (0.00, 1.00)	0.75 (0.61, 0.86)
bedroom	conv5	0.85 (0.00, 1.00)	0.72 (0.57, 0.84)
bowling alley	conv5	1.00 (0.00, 1.00)	0.60 (0.45, 0.74)
backpack	last	-0.00 (-0.00, 1.00)	0.75 (0.70, 0.79)
bird	last	1.00 (-0.00, 1.00)	0.73 (0.66, 0.77)
dog	last	1.00 (-0.00, 1.00)	0.75 (0.70, 0.79)
bamboo forest	last	0.00 (0.00, 0.89)	0.92 (0.90, 0.94)
bedroom	last	0.00 (0.00, 0.07)	0.90 (0.87, 0.92)
bowling alley	last	0.06 (0.00, 1.00)	0.83 (0.79, 0.87)

Table 4: Median bootstrapped MCS scores for two concept-attribution methods: $\mathrm{TCAV}^{sign(\mathrm{CS})}$ and $\mathrm{TCAV}^{\mathrm{ICS}}$ with concept-forgetting baseline ($\lambda = 100$). They were computed for 3 scene and 3 object concepts. 95% nonparametric confidence intervals in parenthesis. conv3, conv4, conv5 designate the outputs of the stacks of 2d convolutions having respectively 512, 1024, and 2048 filters in the Resnet50 architecture (see Figure 3 in [HZRS15]).

and therefore ICS goes to 0. We verified on BARS that decreasing the $\frac{n}{d}$ ratio from 30 to 0.1 resulted in MCS scores 3 times smaller (45% vs 15%).

In this section, we evaluate different techniques that would lead to more consistent directions of the CAV. We first increase n by augmenting the CAV training set with image transformations such as random flips and changes of contrast, brightness, hue, and saturation. The figure in the main text shows how increasing $\frac{n}{d}$ improved MCS for $\mathrm{TCAV}^{\mathrm{ICS}}$ using a concept-forgetting baseline. For the zero baseline, the MCS scores benefit only marginally from increased $\frac{n}{d}$ from data augmentation (Figure 10). We also experimented with various regularization schemes for training CAVs, which did not lead to any improvement in terms of ICS score magnitude.

concept	layer	zero	entropy-maximizing
backpack	conv4	0.00 (-0.00, 0.00)	0.00 (0.00, 0.00)
bird	conv4	-0.00 (-0.00, 0.00)	0.00 (0.00, 0.00)
dog	conv4	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
bamboo forest	conv4	0.02 (0.01, 0.03)	0.00 (0.00, 0.00)
bedroom	conv4	0.01 (0.01, 0.02)	0.00 (0.00, 0.00)
bowling alley	conv4	0.01 (0.00, 0.02)	0.00 (0.00, 0.00)
backpack	conv5	0.08 (0.04, 0.11)	0.00 (0.00, 0.01)
bird	conv5	0.06 (0.03, 0.09)	0.00 (0.00, 0.01)
dog	conv5	0.06 (0.03, 0.08)	0.00 (0.00, 0.01)
bamboo forest	conv5	0.04 (0.02, 0.06)	0.00 (0.00, 0.00)
bedroom	conv5	0.06 (0.03, 0.08)	0.00 (0.00, 0.00)
bowling alley	conv5	0.06 (0.04, 0.08)	0.00 (0.00, 0.00)
backpack	last	0.18 (0.16, 0.21)	0.05 (0.04, 0.06)
bird	last	0.15 (0.12, 0.18)	0.06 (0.04, 0.07)
dog	last	0.17 (0.15, 0.20)	0.05 (0.03, 0.05)
bamboo forest	last	0.21 (0.18, 0.23)	0.03 (0.02, 0.03)
bedroom	last	0.19 (0.16, 0.22)	0.04 (0.03, 0.05)
bowling alley	last	0.19 (0.15, 0.22)	0.05 (0.04, 0.05)

Table 5: Median bootstrapped MCS for $\mathrm{TCAV}^{\mathrm{ICS}}$ with two baselines: zeros (black image) and entropy-maximizing. They were computed for 3 scene and 3 object concepts. 95% nonparametric confidence intervals in parenthesis. conv3, conv4, conv5 designate the outputs of the stacks of 2d convolutions having respectively 512, 1024, and 2048 filters in the Resnet50 architecture, and last is the output of the final layer (before the softmax) (see Figure 3 in [HZRS15]).

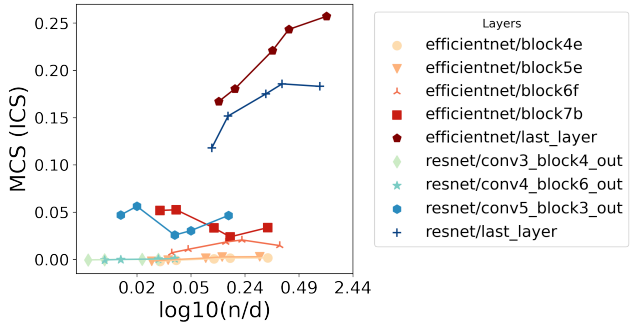

Figure 10: **MCS with varying n/d ratio.** Influence of $\frac{n}{d}$ ratio over MCS computed with concept "bird" for several layers of a EfficientNet-B3 model (in shades of orange to red) [TL20] and a ResNet50 (in shades of blue to green) [HZRS15] fine-tuned on BAM. Darker colors represent deeper layers. MCS(ICS) with black image baseline.

D ImageNet

D.1 Image selection and CAV building

The first three concepts (striped, dotted and zigzagged) rely on the same images as used in [KWG+18]. For the other concepts, we used Google search images, automatically downloaded using `https://github.com/hardikvasa/google-images-download` (MIT license). A manual review of those images was performed to discard images that included multiple concepts (e.g. "hoop" also displaying the wooden floor), or obvious confounders such as watermarks, drawings, captions, etc. We caveat the manual review of the Google search images, especially in the last 2 concepts as "gender" was assessed by the captions or web links of the images (e.g. "Women League") as well as physical appearance. For skin tone (referred to as the "race" concept), physical appearance was used. We also note that these images were directly scraped, and that no consent was obtained from e.g. basketball players. In a real-world use case of our method, these two aspects would need to be addressed with care, e.g. by obtaining consented images, with self-reported demographics information.

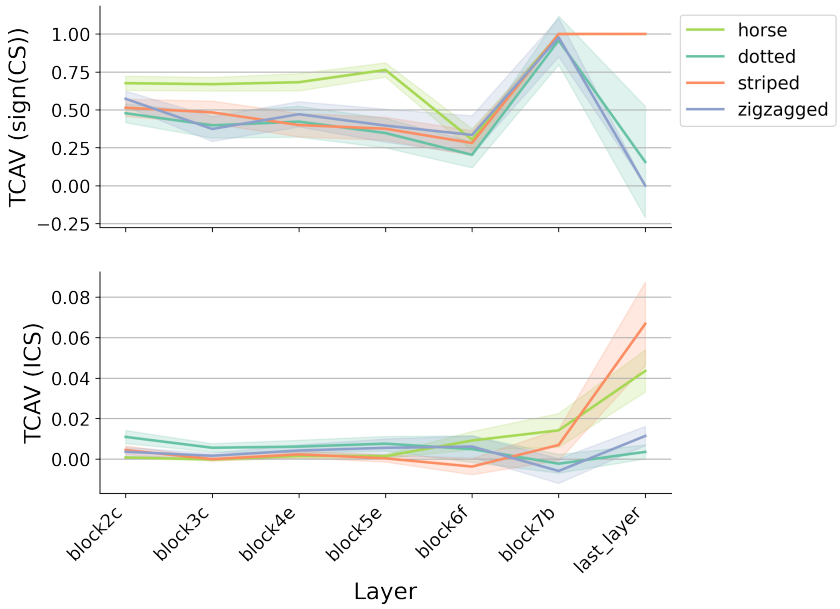

Figure 11: TCAV, CS and ICS scores of several concepts on the 'zebra' output of an EfficientNet-B3 model, for 100 pictures of zebras, computed on all layers, with a white image baseline.

Our selection resulted in ~ 100 images per concept. Some concepts were then built by distinguishing between the selected images and random images as defined in [KWG+18]. For the "race" and "gender" concepts, we compute *relative* concepts by contrasting pictures of men and women, and pictures of lighter skin tone and darker skin tone basketball players (in an attempt to control for clothing, background or occupation biases in our image search).

D.2 Global results for zebras

We perform the global analysis for zebras on different models and layers, for the white image baseline. Our results are consistent across both versions of TCAV: the "zebras" concept has highest scores (not displayed due to scale), followed by "stripes" and then "horse". Different architectures (EfficientNet, MobileNet, ResNet, Inception) display different distributions of the results across their layers (Figures 11, 12, 13 and 14). It however seems that the deepest layers are the most suited for TCAV analyses across architectures. We also observe that the standard deviation of $\mathrm{TCAV}^{sign(\mathrm{CS})}$ can be large for irrelevant concepts like dotted and zigzagged.

E Comparison of computational efficiency

While ICS seems to provide more reliable concept attribution scores, it is more computationally intensive, since it requires the evaluation of several gradients to estimate the integral. Modern hardware and software capabilities make this constraint not so restrictive, since the gradients can be computed in parallel. Computing ICS score for a single image with a given deep learning model only requires to be able to run a forward pass with batch size 100 (100 values in the sum that approximates the integral). It should work on most PC (e.g. any CPU and 8 - 16Go RAM). However, for high resolution inputs containing many features and large models (e.g. ResNet50 and Inception-v1 have hundreds-of-thousands dimensional activation spaces), all the interpolated samples may not fit in memory.

Efficient quadratures may help mitigate this problem. For example, Bayesian [BOGO15] and Gaussian [UCS17] quadratures may provide up to exponential convergence rates when the integrand is smooth enough. This means that far fewer samples (we used 50) are needed to achieve the same level of estimation error.

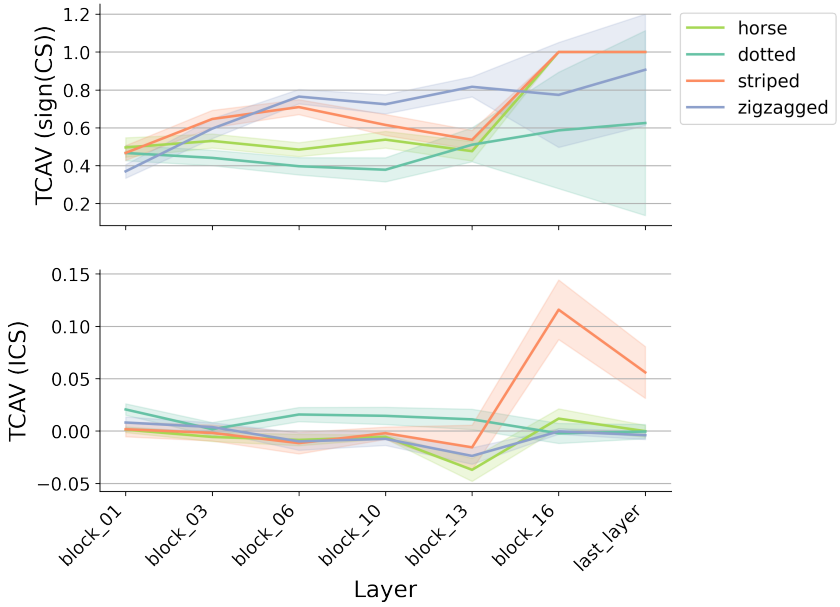

Figure 12: MobileNet V2.

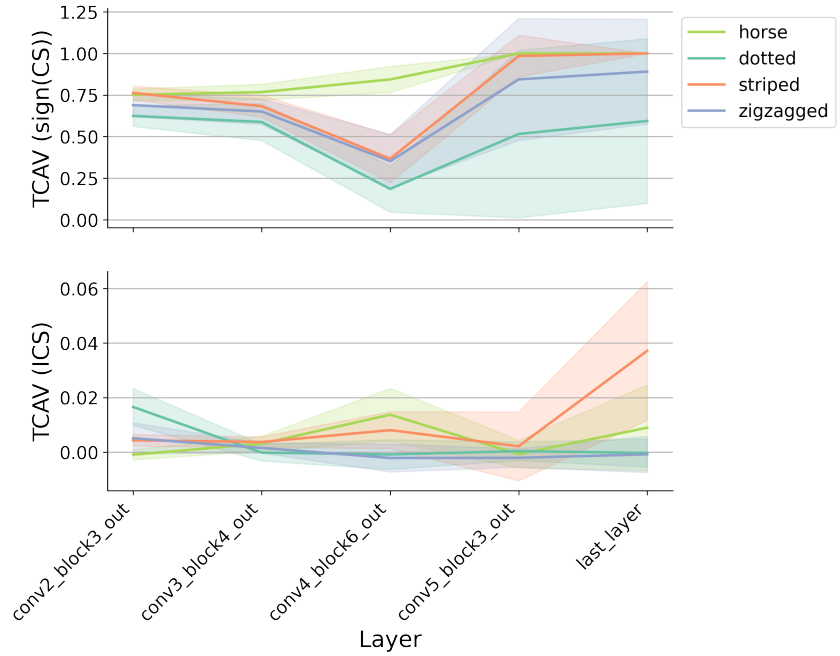

Figure 13: ResNet50.

We also note that some specific use cases can lead to closed form solutions (e.g. Sec A.2), which are inexpensive to compute.

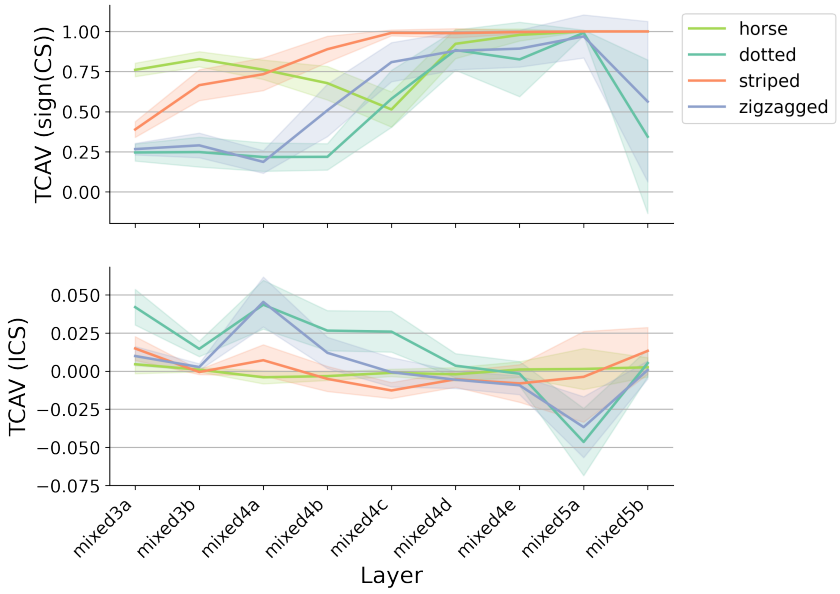

Figure 14: Inception.