Visual-TCAV: Concept-based Attribution and Saliency Maps for Post-hoc Explainability in Image Classification

Anonymous authors
Paper under double-blind review

Abstract

Convolutional Neural Networks (CNNs) have shown remarkable performance in image classification. However, interpreting their predictions is challenging due to the size and complexity of these models. State-of-the-art saliency methods generate local explanations highlighting the area in the input image where a class is identified but cannot explain how a concept of interest contributes to the prediction. On the other hand, concept-based methods, such as TCAV, provide insights into how sensitive the network is to a human-defined concept but cannot compute its attribution in a specific prediction nor show its location within the input image. We introduce Visual-TCAV, a novel explainability framework aiming to bridge the gap between these methods by providing both local and global explanations. Visual-TCAV uses Concept Activation Vectors (CAVs) to generate class-agnostic saliency maps that show where the network recognizes a certain concept. Moreover, it can estimate the attribution of these concepts to the output of any class using a generalization of Integrated Gradients. We evaluate the method's faithfulness via a controlled experiment where the ground truth for explanations is known, showing better ground truth alignment than TCAV. Our code is available at (see supplementary material .zip file).

1 Introduction

As the performance of deep learning models has grown significantly over recent years, their complexity has also increased, making it difficult for users to understand how these models make decisions. As a result, they are often referred to as *black-box* models, since only their inputs and outputs are known, while their internal mechanisms are too complex for humans to comprehend. This lack of algorithmic transparency (von Eschenbach, 2021) has been shown to reduce trust in AI-based systems (Lipton, 2016), particularly in critical fields such as healthcare or autonomous driving in which neural networks are becoming increasingly employed (Turay & Vladimirova, 2022; Cai et al., 2020). Additionally, debugging models becomes challenging without comprehending the process they use to make predictions. To this end, the field of Explainable AI (XAI) has made significant progress in developing techniques for explaining black-box models. However, determining whether a certain human-understandable concept is recognized by the network and how it influences the prediction remains a significant challenge. In computer vision, widely used explainability approaches use saliency maps to localize where a class is identified in an input image, but they can't explain which high-level features led the model to its prediction. For instance, these methods cannot determine whether a golf ball was recognized by the spherical shape, the dimples, or some other feature. To address this, Kim et al. (2018) introduced TCAV (Testing with Concept Activation Vectors), a method that can discern whether a user-defined concept (e.g., dimples, spherical) correlates positively with a selected class. However, TCAV cannot measure the importance of a concept in a specific prediction or show the locations within the input images where the concepts are recognized.

In this article, we introduce a novel explainability framework, namely Visual-TCAV, which integrates the core principles of saliency methods and concept-based approaches while aiming to overcome their respective limitations. Visual-TCAV can be applied to any layer of a CNN model whose output is a set of feature maps. Its main contributions are: (i) it provides visual explanations that show where the network identifies

concepts of interest; (ii) it can estimate the importance of these concepts to the output of any class; (iii) it can be used for both local and global explainability.

2 Related Work

The field of XAI has seen rapid growth in recent years. For image classification, early methods primarily provide explanations via saliency maps that highlight the most important regions in the input image to predict a certain class. A model-agnostic approach for generating such visualizations involves studying the model's input-output relationship by creating a set of perturbed versions of the input and analyzing how the output changes with each perturbation. Notable contributions to this approach include LIME (Ribeiro et al., 2016), which uses random perturbations, and SHAP (Lundberg & Lee, 2017), which estimates the importance of each pixel using Shapley values. A different approach that instead tries to access the internal workings of the model was originally proposed by Simonyan et al. (2014) and consists of generating saliency maps based on the gradients of the model output w.r.t. the input images. This idea led many researchers (Springenberg et al., 2014; Smilkov et al., 2017) to investigate how to exploit gradients to produce more accurate saliency maps. Among them, Selvaraju et al. (2017) proposed Grad-CAM, a generalization of CAM (Zhou et al., 2016) which extracts the gradients of the logits (i.e., raw pre-softmax predictions) w.r.t. the feature maps and uses a Global Average Pooling (GAP) operation to transform these gradients into class-specific weights for each feature map. It then performs a weighted sum of these feature maps to produce a class localization map. Grad-CAM has gained considerable attention and is extensively used to explain convolutional networks. However, Sundararajan et al. (2017) demonstrated that gradients can saturate, leading to an inaccurate assessment of feature importance. To address this issue, they introduced Integrated Gradients (IG), a method that calculates feature attribution by integrating the gradients along a path from a baseline (e.g., a black image) to the actual input image. A notable contribution of IG is the ability to provide fine-grained saliency maps and a strong theoretical foundation.

While saliency maps are widely used for image classification, they do not provide a complete picture of why a model made a certain decision. This is because these methods only perform class localization but cannot explain which high-level concepts led the model to recognize the highlighted class. Furthermore, these techniques rely on per-pixel importance which can't be generalized across multiple instances. Consequently, they can only explain one image at a time, preventing them from providing global explanations. To overcome these limitations, Kim et al. (2018) proposed TCAV, a method that investigates the correlations between user-defined concepts and the network's predictions using a set of example images representing a concept. For instance, example images of the concept "striped" can be used to determine whether the network is sensitive to this concept for predicting the "zebra" class. This is accomplished by learning a Concept Activation Vector (CAV), which is a vector orthogonal to the decision boundary of a linear classifier trained to differentiate between the feature maps of concept examples and random images. From this, a sensitivity score is computed using the signs of the dot products between the CAV and the gradients of a selected class w.r.t. the feature maps. This approach is effective in detecting specific biases and can be considered complementary to saliency maps. Indeed, while Grad-CAM and IG apply exclusively to individual predictions, TCAV provides only global explanations. However, TCAV does not show where concepts are identified within the input images, making it challenging to assess whether a high score can truly be attributed to the intended concept. Moreover, TCAV computes the network's sensitivity to a concept, but not the magnitude of its importance as the score only depends on the signs of the directional derivatives. For instance, the concepts "grass" and "dimples" might have identical TCAV scores for the "golf ball" class, even if one contributes substantially more to the prediction.

TCAV has received attention within the XAI community, leading to various applications (Lucieri et al., 2020; Cai et al., 2019) and extensions (Crabbé & van der Schaar, 2022; Graziani et al., 2018). A notable extension that tries to add localization for a concept of interest is CAVLI (Shukla et al., 2023), which uses LIME to decompose images into superpixels and uses CAVs to assess their similarity to the concepts. In parallel, several unsupervised approaches have also been proposed (Ghorbani et al., 2019; Zhang et al., 2021; Fel et al., 2023b; Akula et al., 2020) to discover concepts automatically. These methods typically involve cropping images of a class to generate patches, which are then re-inserted into the network, and activations are clustered, resulting in a set of extracted concepts. More recently, Sparse Autoencoders (SAEs) have also

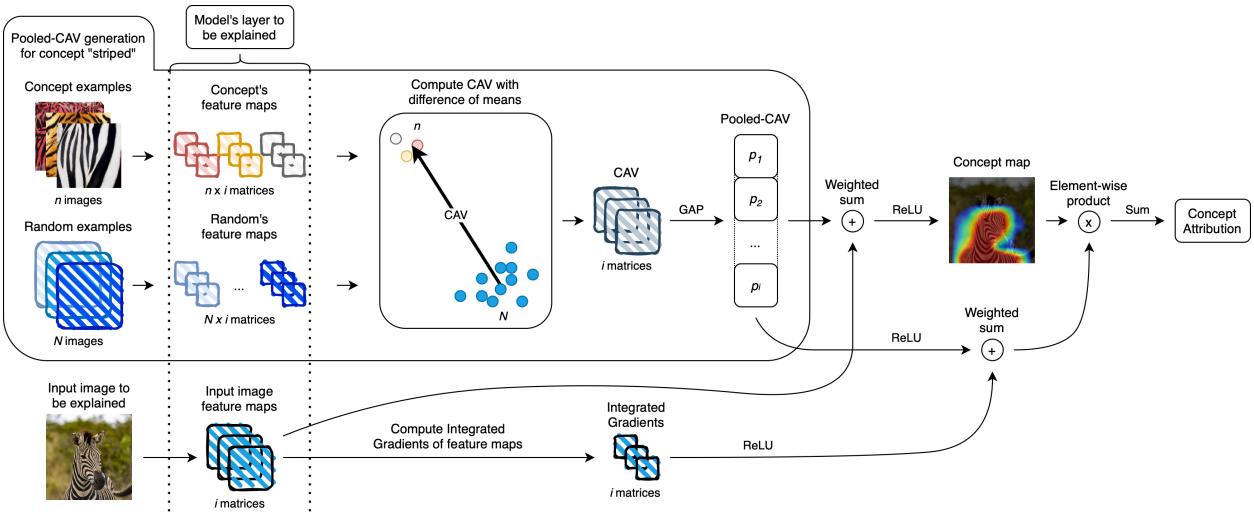


Figure 1: The Visual-TCAV pipeline for generating local explanations. A pooled-CAV is computed using the feature maps of user-defined concept examples and random images. A concept map is then produced through a weighted sum of the pooled-CAV and the image's feature maps. Finally, a concept attribution is obtained by extracting the IG attributions of the neurons that the concept activates using the pooled-CAV and the concept map, which is used as a spatial mask.

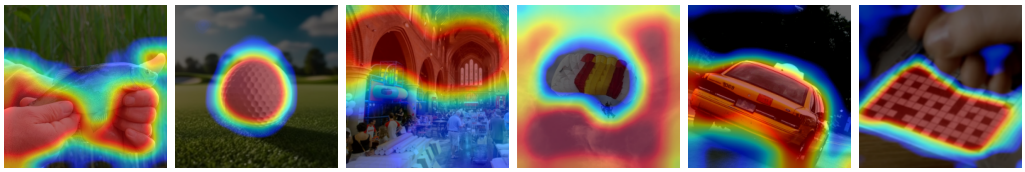
shown promising results in discovering interpretable features across both vision and language domains (Fel et al., 2023a; Bereska & Gavves, 2024). However, it is often the case that concepts of interest are not being extracted even though they are important for the task (Sharkey et al., 2025). For the scope of this paper, we will focus only on supervised scenarios in which the concept to study is manually defined by the user.

3 Visual-TCAV

This section presents Visual-TCAV, whose methodology is outlined in Fig. 1. Local explanations can be generated for any layer of interest and consist of two key components: (i) the *concept map*, a saliency map that serves as a visual representation of where the network recognized the selected concept in the input image and (ii) the *concept attribution*, a numerical value that estimates the importance of the concept for a selected class. To derive global explanations, the process is replicated across multiple input images and the concept attributions are averaged to quantify how the concept influences the network's decisions across a wide range of inputs.

3.1 CAV Generation and Spatial Pooling

Similarly to the original TCAV framework, the initial step of our method consists of computing a Concept Activation Vector (CAV) from a set of example images representing a user-defined concept, and a set of negative examples (e.g., random images). Specifically, we use the *Difference of Means* method proposed by Martin & Weller (2019) to compute the CAV, as they demonstrated that this approach produces more consistent CAVs than logistic classifiers or support vector machines. As the name suggests, this method uses the arithmetic mean to determine the centroids of both the concept's activations and the activations of random images. A CAV is then computed directly as the difference between the concept centroid and the random centroid, which is also mathematically equivalent to the pattern-CAV approach proposed by Pahde et al. (2025). Since we are interested in identifying which feature maps are activated by the concept, irrespective of its spatial location within the example images, we apply a GAP operation on the obtained CAV. The result is a vector of scalar values whose length is equal to the number of feature maps of the layer under consideration. Each element of this vector is associated with a feature map, and its raw value



(a) Hands (b) Dimples (c) Arches (d) Sky (e) Car (f) Chequered

Figure 2: Examples of class-independent concept maps for various input images and concepts.

approximates the degree of correlation between that feature map and the concept. Moving forward, we refer to this vector as the *pooled-CAV*.

3.2 Concept Map

Using the pooled-CAV, we derive a concept map $\boldsymbol{M}^c$ that spatially localizes a concept c within the input image by applying Eq. (1). Specifically, we compute a weighted sum of the feature maps $\mathbf{F}$ from the selected layer, where each channel i is weighted by the corresponding element p_i^c of the pooled-CAV. Subsequently, we apply a ReLU function to retain only the regions positively correlated with the concept.

$$\boldsymbol{M}^c = ReLU\left(\sum_i p_i^c \cdot \mathbf{F}_{:,:,i}\right) \tag{1}$$

To enable meaningful comparisons of concept activation across different input images, we apply a custom min-max normalization to the concept map based on a predefined range, as shown in Eq. (2). The normalization range boundaries (ℓ^c and u^c) are derived from the example images provided by the user to represent the concept. Specifically, the upper bound u^c is computed by first generating concept maps from positive example images $\mathcal{D}_{pos}$, then calculating the median of their contraharmonic means ($chmean$), which serve to heuristically extract values of high activation within each concept map. The lower bound ℓ^c is computed using the same method but with the concept maps of negative example images $\mathcal{D}_{neg}$ instead. The normalized concept map $\hat{\boldsymbol{M}}^c$ is then obtained by clipping any value outside this predefined range and then scaling it to $[0, 1]$. The notation $[\boldsymbol{A}]_a^b$ denotes clipping each element of $\boldsymbol{A}$ within the interval $[a, b]$, i.e., $[\boldsymbol{A}]_a^b = min(max(\boldsymbol{A}, a), b)$, where the min and max are applied element-wise.

$$\hat{\boldsymbol{M}}^c = \frac{[\boldsymbol{M}^c]_{\ell^c}^{u^c} - \ell^c}{u^c - \ell^c}, \quad \text{where} \quad \begin{cases} u^c = median\left(chmean_{i \in \mathcal{D}_{pos}}(\mathbf{M}^c_{:,:,i})\right) \\ \ell^c = median\left(chmean_{i \in \mathcal{D}_{neg}}(\mathbf{M}^c_{:,:,i})\right) \end{cases} \tag{2}$$

By overlaying the normalized concept map on the input image, we generate a class-independent visualization that highlights the region of the image where the network recognized the concept (examples are shown in Fig. 2). This allows us to know, for any input image, the concept's location and its degree of activation w.r.t an ideal concept defined by the user. Additionally, concept maps act as a qualitative validation for whether the CAV represents the intended concept, without requiring activation maximization techniques (Mordvintsev et al., 2015) or sorting images based on their similarity to the CAV.

3.3 Concept Attribution

We aim to gain insights into the network's decision-making process by estimating the importance of any concept of interest toward the predictions. To this end, given an input image, a layer, and a target class, we compute the attribution of each neuron's activation toward the output of that class and then use the pooled-CAV to approximate the correlation between the concept and the activations of each neuron. Since concepts may overlap or be hierarchically related, their attributions are not assumed to be additive. The neurons' attributions can be computed through a generalized variant of IG, which computes the integrated

gradients of a class's logit (i.e., pre-softmax prediction) w.r.t. the feature maps instead of the input image. Specifically, we compute the gradients along a straight-line path from zero-filled matrices to the actual feature maps, approximating the integral using the Riemann trapezoidal rule. Since IG satisfies the completeness axiom, the attributions sum to the difference between the target class's logit and the baseline's logit for that class, regardless of which layer is used as input. Then, for multiclass networks, we retain only the positive attributions. For binary class networks, we sum all the positive attributions to derive the score for the positive class, and the negative ones to obtain the score for the negative class. Given a target class k, we refer to the attributions towards that class as $\mathbf{IG}^k$. Since the scale of these attributions is the same as the raw logits, making interpretation difficult, we normalize the attributions so that their sum is between 0 and 1. Specifically, we begin by computing the difference between the logits of our input image $\boldsymbol{z}(x)$ and the logits of the baseline $\boldsymbol{z}(x')$ followed by a ReLU and $[0, 1]$ scaling as shown in Eqs. (3a) and (3b).

$$\Delta \boldsymbol{z} = ReLU\left(\boldsymbol{z}(x) - \boldsymbol{z}(x')\right) \tag{3a}$$

$$\Delta \hat{\boldsymbol{z}} = \frac{\Delta \boldsymbol{z}}{max(\Delta \boldsymbol{z})} \tag{3b}$$

The attributions for the target class (k) are then scaled so that their sum equals the normalized difference $\Delta \hat{z}_k$ for that class, as shown in Eq. (4a).

$$\hat{\mathbf{IG}}^k = \frac{\mathbf{IG}^k}{\sum \mathbf{IG}^k} \cdot \Delta \hat{z}_k \tag{4a}$$

$$\hat{\boldsymbol{p}}^c = \frac{ReLU(\boldsymbol{p}^c)}{max(\boldsymbol{p}^c)} \tag{4b}$$

Finally, to estimate the attribution of a concept c for a target class k, we utilize the pooled-CAV to perform a weighted sum of the normalized attributions $\hat{\mathbf{IG}}^k$ toward that target class. Before this summation, we also apply a ReLU and min-max normalization to the pooled-CAV, as shown in Eq. (4b), to progressively reduce attribution for feature maps that are less correlated with the concept. The rationale behind using the ReLU is to consider only attributions of feature maps that are activated by the concepts. We refer to the normalized pooled-CAV as $\hat{\boldsymbol{p}}^c$. Then, as shown in Eq. (5), we obtain the concept attribution $a^{c,k}$ for a concept c and a target class k by summing all values of an element-wise multiplication of this weighted sum and the normalized concept map $\hat{\boldsymbol{M}}^c$, which is used as a spatial mask. This enables us to discard the attributions of activations related to the regions within the input image where the concept is not present or was not recognized.

$$a^{c,k} = \sum \Big(\hat{\boldsymbol{M}}^c \odot \sum_i (\hat{p}_i^c \cdot \hat{\mathbf{IG}}^k_{:,:,i}) \Big) \tag{5}$$

The concept attribution is a per-concept metric, meaning that two concepts can have significantly different attributions even if they are recognized in the same location of the input image, resulting in similar concept maps. This distinction is achieved by focusing not on per-pixel attributions but on the attributions of the activations produced by the neurons responsible for recognizing the concepts. Moreover, we can average the concept attribution across multiple input images to measure the overall importance of a concept for a selected class, thus providing a global explanation.

4 Results

This section presents the results of applying Visual-TCAV to a range of widely used CNNs. We show that our method (i) can reveal how concepts contribute to mispredictions, (ii) provides insights into where concepts are detected within the network, and (iii) aligns with ground truth in controlled experiments.

4.1 Experimental Setup

Our experiments are conducted on ResNet50V2 (He et al., 2016), InceptionV3 (Szegedy et al., 2016), and VGG16 (Simonyan & Zisserman, 2015), all pre-trained on the ImageNet (Deng et al., 2009) dataset, as well as a ResNet50V2 model trained from scratch (i.e., without ImageNet weights) on the CelebA (Liu et al., 2015) dataset for celebrity gender classification. The tested concepts were sourced from the Describable Textures

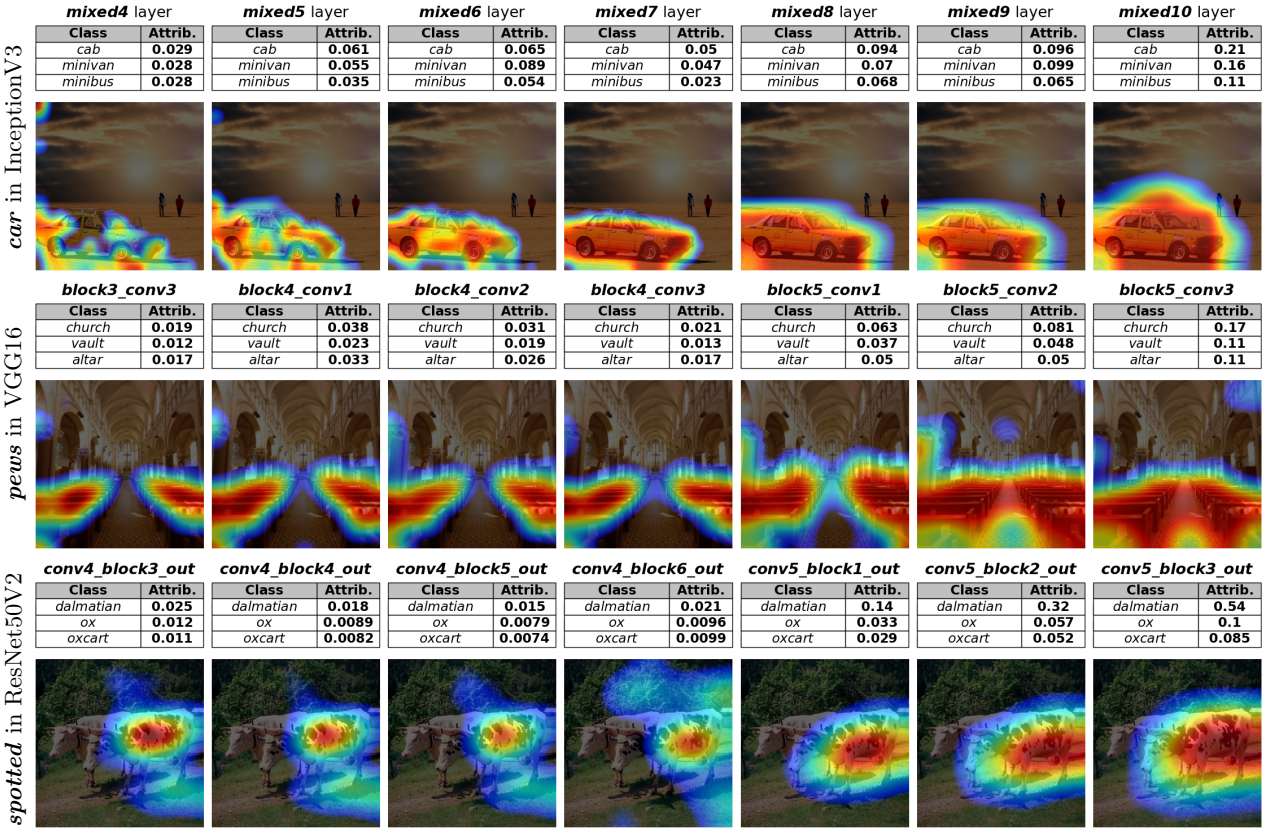


Figure 3: Examples of layer-wise local explanations for various concepts and networks. We compute the attribution of each concept for the top three predicted classes, ranked by class probability, and the last seven layers.

Dataset (DTD) (Cimpoi et al., 2014), obtained from popular image search engines, or generated with Stable Diffusion (Rombach et al., 2022) (more on this in Appendix E). For concepts collected via search engines, we use 30 example images per concept, which can be considered a reasonable amount as shown in the stability experiment provided in Appendix G. For DTD concepts, we use all 120 available images, and for generated concepts, we use 200 images. For ImageNet negative examples, we follow the approach recommended by Martin & Weller (2019), selecting 500 random images. In the case of CelebA, images containing the concept are used as positive examples, while those without it serve as negative examples. Regarding the computation of the integrated gradients, we used 300 steps, which are usually enough to approximate the integral within a 5% error margin (Sundararajan et al., 2017). Experiments are conducted on an Intel i7 13700k with an RTX 4060Ti 16GB and 32 GB of RAM, using TensorFlow 2.15.1, CUDA 12.2, and Python 3.11.5. With this setup, local explanations for 7 layers take 1 to 2 minutes, while global ones with 200 class images and 7 layers take from 5 to 30 minutes depending on the model. For global explanations, the computation time remains nearly constant regardless of the number of concepts processed simultaneously.

4.2 Local Explanations

While concept maps are class-independent, the attribution of each concept depends on the class considered. In our examples, we examine the top three predicted classes and apply Visual-TCAV to a subset of the CNN layers. The rationale behind these layer-wise explanations is that they enable us to visualize where specific concepts are learned within the network. For instance, the "car" concept in Fig. 3 is not well recognized in earlier layers, corroborating previous findings that higher-level features are typically extracted deeper into the network (Zeiler & Fergus, 2014; Olah et al., 2017; Bau et al., 2017; Bianchi et al., 2024). On the other

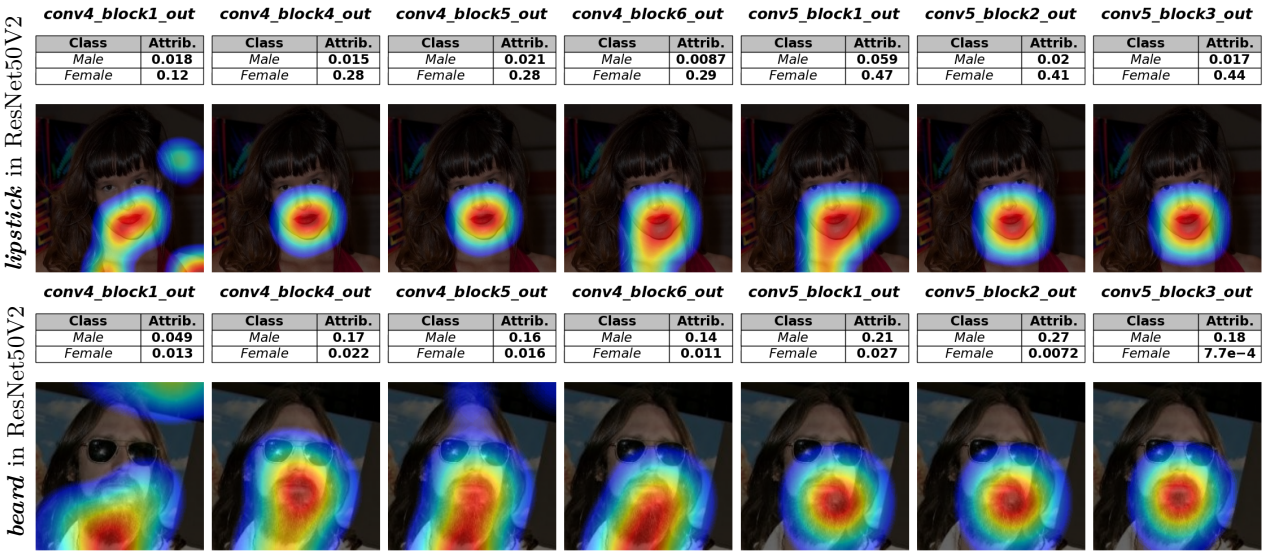


Figure 4: Examples of layer-wise local explanations for a ResNet50V2 (He et al., 2016) model trained on the CelebA (Liu et al., 2015) dataset for celebrity gender classification. The concepts tested are "lipstick" and "beard", both significantly contributing to the prediction.

hand, concept maps in earlier layers are more fine-grained (see the "pews" concept in Fig. 3) due to their neurons having smaller receptive fields. Additionally, we observe a substantial increase in attributions in deeper layers, which is aligned with other studies (Morcos et al., 2018; Amjad et al., 2022) showing deeper neurons exhibit higher class selectivity due to their proximity to the output. However, this is slightly less pronounced on the CelebA dataset (refer to Fig. 4), where the "lipstick" and "beard" concepts exhibit class selectivity even in earlier layers, which may be due to the lower task complexity. The utility of these explanations lies mainly in their ability to show not only which image regions the network is focusing on, but also what the model recognizes in those regions and how much this contributes to each class's output. This can be especially useful for revealing the model's reasoning behind mispredictions. For instance, the third image in Fig. 3 is an "ox" wrongly classified as "dalmatian" for which we observe that the network's decision is largely influenced by the "spotted" concept, which accounts for more than half the logit value of the "dalmatian" class.

4.3 Global Explanations

The concept attribution is a per-concept metric of importance, hence we can derive global explanations by aggregating this attribution across multiple input images of a selected class. In our experiments, we utilize 200 images per class for each global explanation. For concepts that are inherently part of the class (e.g., "striped" for "zebra", or "dimples" for "golf ball"), we can directly use any image representing that class. On the other hand, for concepts that appear sporadically, we only use images where the concept is present. For instance, we only use images of church interiors for "pews" and "fresco" concepts, and images of church exteriors for the "steeple" concept. This ensures that the explanations are independent of the frequency of the concept's appearance in the class images.

Results of global explanations for various concepts are provided in Fig. 5. The attributions match intuitive expectations, considering, for instance, the importance of "striped" for "zebra" or "spotted" for "dalmatian". However, such high attribution may suggest the model over-relies on certain concepts, using them as shortcuts for class discrimination. This may lead the network to misclassify out-of-distribution cases, such as the "ox" being incorrectly labeled as a dalmatian in Fig. 3. This may also pose issues for concepts like "lipstick" in gender classification, potentially leading to biased predictions. Furthermore, while the final layer typically has the highest attribution, which is expected for class discriminative concepts, there are instances, such as "dimples" and "spherical" for "golf ball", where concepts recognized in the earlier layers have a greater

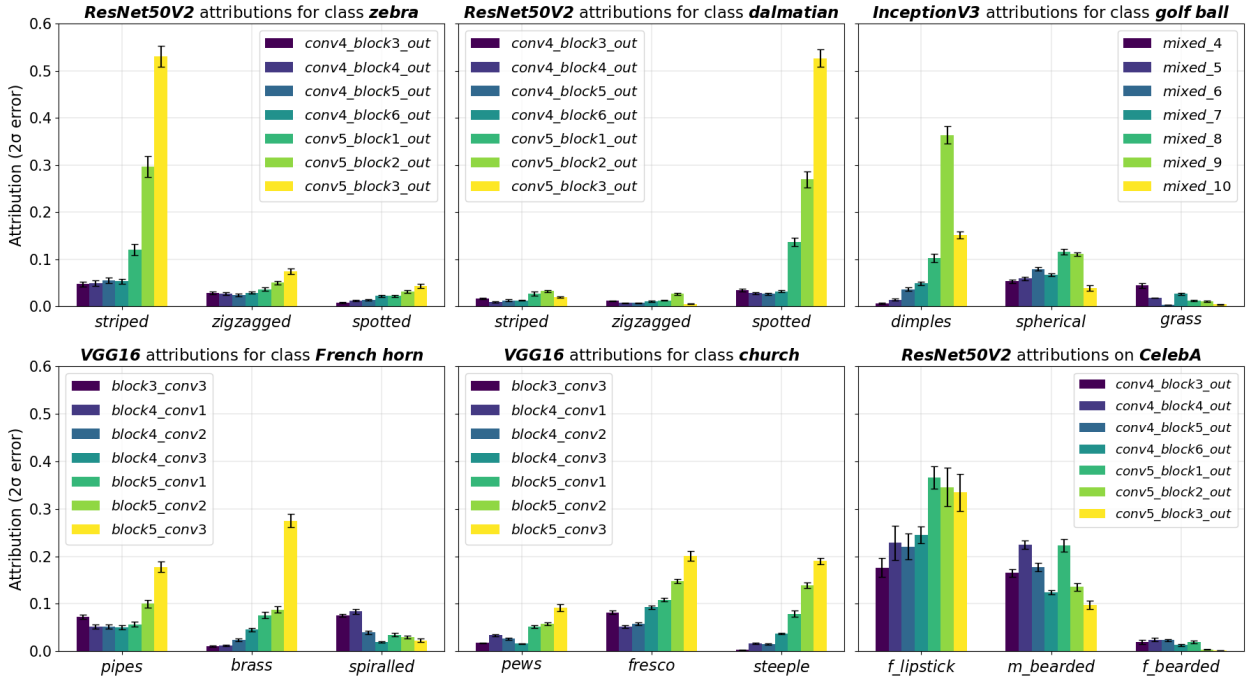


Figure 5: Global explanations for various concepts, classes, and networks. Each bar chart shows the attributions of three concepts for a specific class across the last seven layers of each network. In the CelebA dataset, "m_" denotes attribution toward the "male" class, while "f_" indicates attribution toward the "female" class. Each concept's attribution is averaged over 200 input images.

impact on the network's prediction. Empirically we found such behavior to be dependent not only on the concept but also on the model architecture. More examples of local and global explanations can be found in Appendices C and D.

4.4 Quantitative Evaluation with Ground Truth

We conducted a validation experiment to evaluate the faithfulness of Visual-TCAV. In this experiment, we train a series of convolutional networks in a controlled setting, where the ground truth for explanations is known, and assess whether the Visual-TCAV attributions match this ground truth. For this purpose, we create a dataset of three classes – cucumber, taxi, and zebra – which are the same classes used in the TCAV validation experiment. We then create multiple versions of this dataset by altering a percentage of the images with a tag, represented by a letter enclosed in a randomly sized square, added in a random location of the image. Specifically, zebra images are tagged with a "Z" in a purple square, taxi images with a "T" in a magenta square, and cucumber images with a "C" in a cyan square. From these tagged images, we create five datasets: one of images without tags, and four others with 25%, 50%, 75%, and 100% of tagged images respectively. Each dataset is then used to train a different model, each including six convolutional layers and a GAP layer. Depending on the dataset used for training, each model may learn to recognize either the entities (i.e., cucumbers, taxis, and zebras), the tags, or both and will decide which ones to give more importance. To obtain an approximated ground truth assessing which concept – entity or tag – is more important, we ask the models to classify a set of 200 incorrectly tagged test images per class. In this test set, taxis are tagged with the "Z", cucumbers are tagged with the "T" and zebras are tagged with the "C". If the network accuracy remains high, it indicates that the entity is more important than the tag, and thus we expect its attribution to be higher. On the other hand, if the performance deteriorates on these wrongly tagged images, it indicates that the tag is more important than the entity, hence the tag's attribution should be higher. We obtain the CAVs for entities using images of each class as concept examples and the other two classes with random tags as negative examples. For tags, we use random images containing that tag

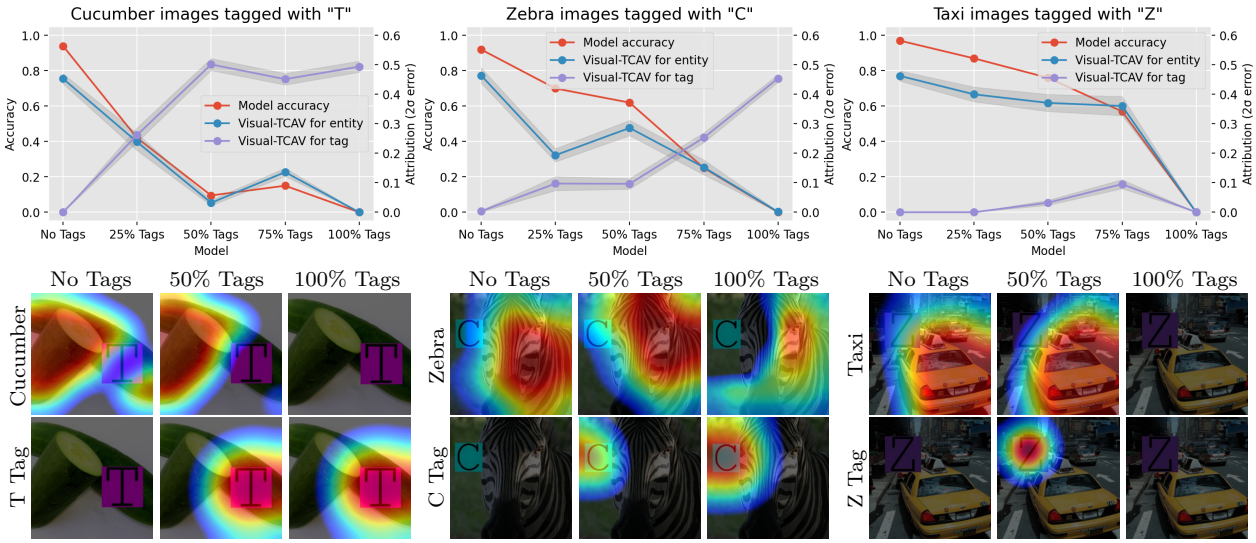


Figure 6: The results of the validation experiment. The upper section of the figure shows the models' accuracies and the Visual-TCAV attributions for both entities and tags across all models. The lower section provides examples of concept maps for the no tags model, 50% tags model, and 100% tags model.

as concept examples and images of cucumbers, taxis, and zebras containing the other two tags as negative examples. We use the same incorrectly tagged test set to compute the concept attributions for both entities and tags across the last convolutional layer of all models.

The results are shown in Fig. 6 and, as expected, an increase in the percentage of tagged images correlates with a decrease in accuracy for all classes. In particular, for the "cucumber" class, the accuracy declines much faster than other classes, with most images being incorrectly classified as taxis. This suggests that even the models trained on a small fraction of tagged images tend to overfit on the "T" tag. The concept attributions for both the "cucumber" entity and the "T" tag closely mirror this ground truth. The "zebra" entity and the "C" tag are also consistent with the ground truth: the attributions for "zebra" show a positive correlation with accuracy, whereas the attributions for the "C" tag demonstrate a clear inverse correlation. Notably, the networks did not pay much attention to the "Z" tag, focusing instead on the absence of the other two tags to classify zebras. Indeed, the model trained with 100% of images tagged classifies any image without a "C" or a "T" tag as "zebra", regardless of whether the "Z" tag is present or not. This is confirmed by our method, which assigns an attribution of nearly zero to both the "Z" tag and the "taxi" entity for the aforementioned model. We also tested saliency methods, such as Grad-CAM and IG, to further validate these findings. These methods do not highlight the "Z" tag either, but rather the entire image, in search of the "zebra" class (see Appendix B). For all models excluding the one trained with 100% of tags, the accuracy for "taxi" remains high, implying that these models are indeed capable of recognizing the "taxi" entity. The concept attribution for the "taxi" entity aligns with this observation. In the lower part of Fig. 6, we provide examples of concept maps for the models trained with a different percentage of tagged images. As expected, the model trained without tags can recognize the entities but not the tags, the model trained with 50% tagged images can recognize both, and the model trained with 100% tagged images struggles to recognize the entities but effectively identifies the "T" and "C" tags.

4.4.1 Comparative Analysis with TCAV

The primary difference between our concept attribution and the TCAV score is that our method considers the magnitude of the gradients, not just their direction. This allows us to measure the concept's impact on the predictions, beyond just the network's sensitivity to it. To demonstrate this, we compute the TCAV scores for tags and entities across each validation model (see Fig. 7). For the model trained without tags, the TCAV scores align with ground truth, showing high sensitivity to entities and no or negative sensitivity to tags (where 0.5 indicates no sensitivity and 0 signifies negative sensitivity). However, for the other models, TCAV

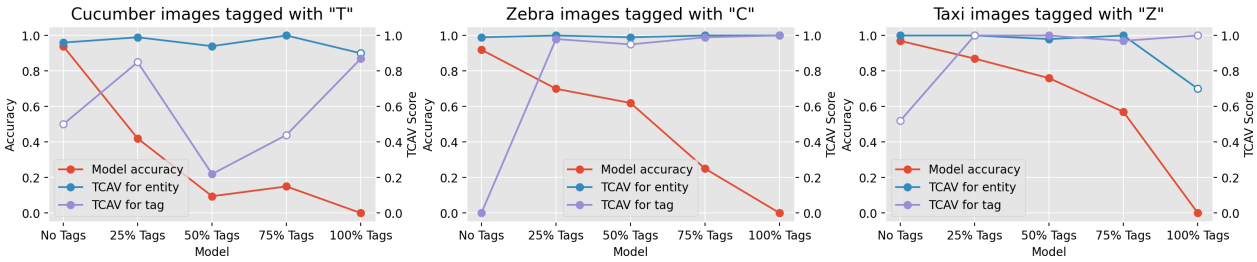


Figure 7: TCAV scores for tags and entities across models trained on a different percentage of tagged images. Results that did not reach statistical significance (p-value < 0.05) according to the TCAV test are indicated with a white dot.

struggles to capture variations in concept importance as defined by ground truth. These models yield high TCAV scores for both entities and tags, despite the decline in accuracy indicating that their contributions to predictions differ significantly. This limitation arises because TCAV relies solely on the signs of gradients, making it incapable of distinguishing the relative importance of these concepts. Therefore, while the TCAV score may be valuable in explaining whether the network learned to correlate a certain concept with a class, our concept attribution provides a more comprehensive explanation by quantifying the extent to which a given concept influences the network's decisions. A qualitative and quantitative comparison with saliency methods (Grad-CAM and Integrated Gradients) is provided in Appendix F.

5 Conclusion

In this article, we presented Visual-TCAV, a novel method that provides concept-based local and global explanations for image classification models by estimating the attribution of user-defined concepts to the network's predictions. Visual-TCAV also generates saliency maps to show where concepts are identified by the network, providing users with visual evidence that CAVs align with their intended concepts. Our method's effectiveness was demonstrated in widely used CNNs and a ground truth experiment, where it successfully identified the most important concept in each examined model.

5.1 Limitations and Future Work

One limitation is that collecting example images requires some effort and domain knowledge. To address this, future works could explore generating explanations directly from text using methods such as Text-To-Concept (Moayeri et al., 2023), which align the activations of multi-modal models like CLIP (Radford et al., 2021) with vision-only models. Additionally, generative approaches, such as DreamBooth (Ruiz et al., 2023), could be investigated to generate many concept images from a small set of examples. Finally, future work could explore adapting our method to other tasks like regression as well as other architectures like Vision Transformers (ViTs) (Dosovitskiy et al., 2021). It may also be interesting to study interconnections between concepts to determine not only the concept attribution toward a class but also toward other concepts in deeper layers.

References

Arjun Akula, Shuai Wang, and Song-Chun Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2594–2601, Apr. 2020. doi: 10.1609/aaai.v34i03.5643. URL `https://ojs.aaai.org/index.php/AAAI/article/view/5643`.

Rana Ali Amjad, Kairen Liu, and Bernhard C. Geiger. Understanding neural networks and individual neuron importance via information-ordered cumulative ablation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7842–7852, 2022. doi: 10.1109/TNNLS.2021.3088685.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017. doi: 10.1109/CVPR.2017.354.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=ePUVetPKu6`. Survey Certification, Expert Certification.

Matteo Bianchi, Antonio De Santis, Andrea Tocchetti, and Marco Brambilla. Interpretable network visualizations: A human-in-the-loop approach for post-hoc explainability of cnn-based image classification. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 3715–3723. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/411. URL `https://doi.org/10.24963/ijcai.2024/411`. Main Track.

Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 258–26. Association for Computing Machinery, 2019. ISBN 9781450362726. doi: 10.1145/3301275.3302289. URL `https://doi.org/10.1145/3301275.3302289`.

Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of Translational Medicine*, 8(11), 2020. ISSN 2305-5847. URL `https://atm.amegroups.com/article/view/36944`.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Jonathan Crabbé and Mihaela van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2590–2607. Curran Associates, Inc., 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2711–2721, June 2023b.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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 (eds.), *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pp. 124–132, Cham, 2018. Springer International Publishing. ISBN 978-3-030-02628-8.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 630–645, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

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 Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Zachary Lipton. The mythos of model interpretability. *Communications of the ACM*, 61, 10 2016. doi: 10.1145/3233231.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. pp. 1–10, 07 2020. doi: 10.1109/IJCNN48605.2020.9206946.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Tyler Martin and Adrian Weller. *Interpretable Machine Learning*. M.Phil. diss., Dept. of Engineering, University of Cambridge, August 2019. URL https://www.mlmi.eng.cam.ac.uk/files/tam_final_reduced.pdf.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Ari S. Morcos, David G.T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1iuQjxCZ.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Frederik Pahde, Maximilian Dreyer, Moritz Weckbecker, Leander Weber, Christopher J. Anders, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. Navigating neural space: Revisiting concept activation vectors to overcome directional divergence. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q95MaWfF4e.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, August 2016. doi: 10.1145/2939672.2939778. URL http://dx.doi.org/10.1145/2939672.2939778.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.02155. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02155.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. doi: 10.1109/iccv.2017.74. URL http://dx.doi.org/10.1109/ICCV.2017.74.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Pushkar Shukla, Sushil Bharati, and Matthew Turk. Cavli - using image associations to produce local concept-based explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3750–3755, 2023. doi: 10.1109/CVPRW59228.2023.00387.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

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.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

Tolga Turay and Tanya Vladimirova. Toward performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey. *IEEE Access*, 10:14076–14119, 2022. doi: 10.1109/ACCESS.2022.3147495.

Warren von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34, 12 2021. doi: 10.1007/s13347-021-00477-0.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11682–11690, May 2021. doi: 10.1609/aaai.v35i13.17389. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17389`.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, Los Alamitos, CA, USA, June 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.319. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.319`.

A Appendix Overview

In the appendix, we provide:

B. IG and Grad-CAM for 100% tags model

C. Additional results of Local Explanations

D. Additional results of Global Explanations

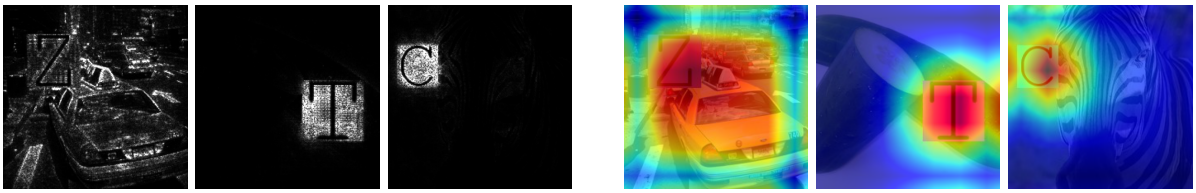
E. Example images for generated concepts

F. Comparison with IG and Grad-CAM

G. Experiments on Visual-TCAV stability

B Saliency methods for 100% tags model

We provide the results obtained by applying IG (Sundararajan et al., 2017) and Grad-CAM (Selvaraju et al., 2017) to the 100% tags model (see Fig. 8). These methods align with Visual-TCAV in showing that this model does not pay attention to the "Z", but rather to the absence of the "T" and the "C" for predicting the "zebra" class.



(a) Integrated Gradients (b) Grad-CAM

Figure 8: IG and Grad-CAM for the model with 100% tags, searching respectively for the classes "zebra", "taxi", and "cucumber". Both methods highlight the "T" for class "taxi" and the "C" for class "cucumber", but fail to recognize the "Z" for class "zebra".

C Additional results of Local Explanations

Continuing from the results presented in the main paper, we further provide additional local explanations for more input images and concepts in Figs. 9 to 11.

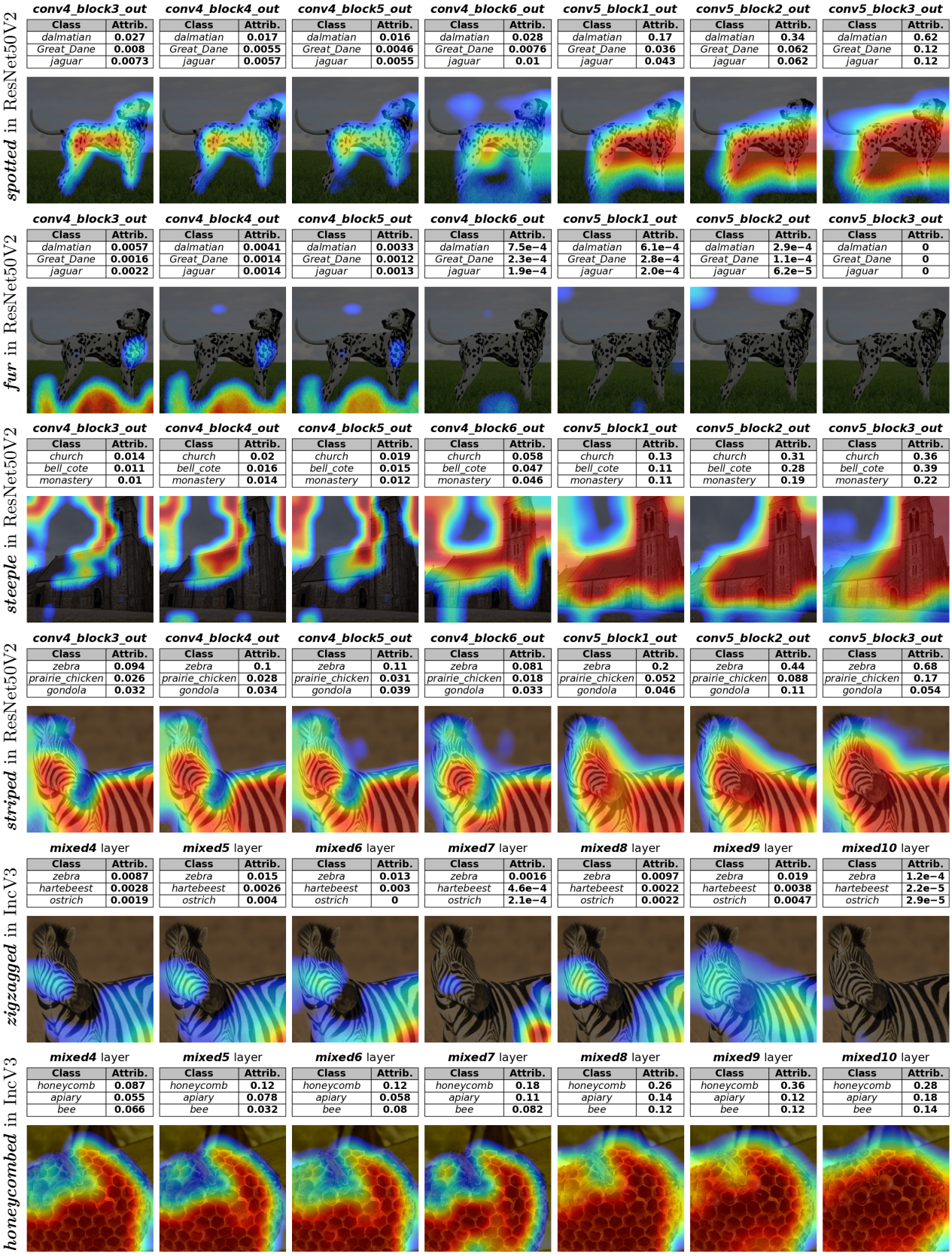


Figure 9: More examples of layer-wise local explanations for various concepts and networks. (Part 1)

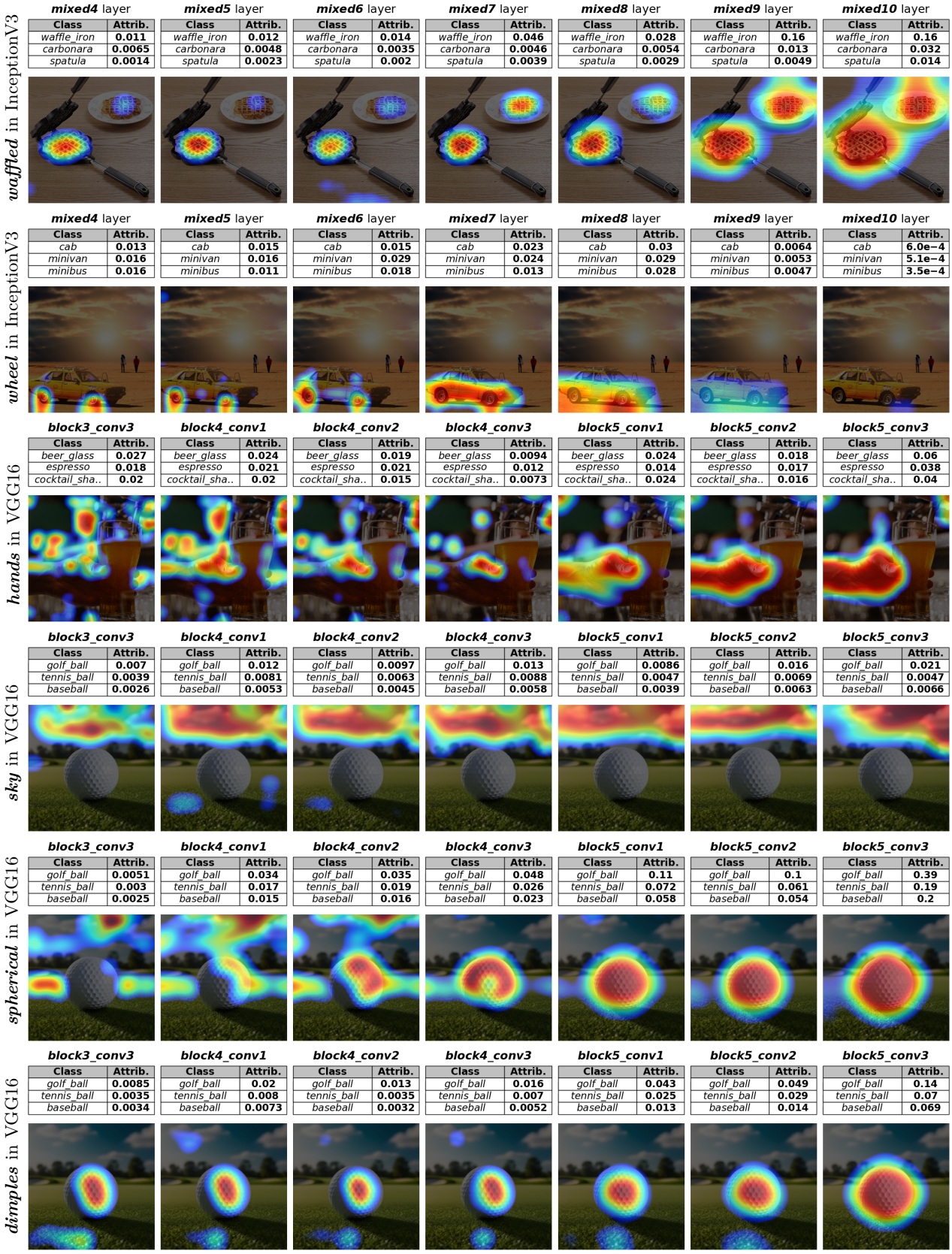


Figure 10: More examples of layer-wise local explanations for various concepts and networks. (Part 2)

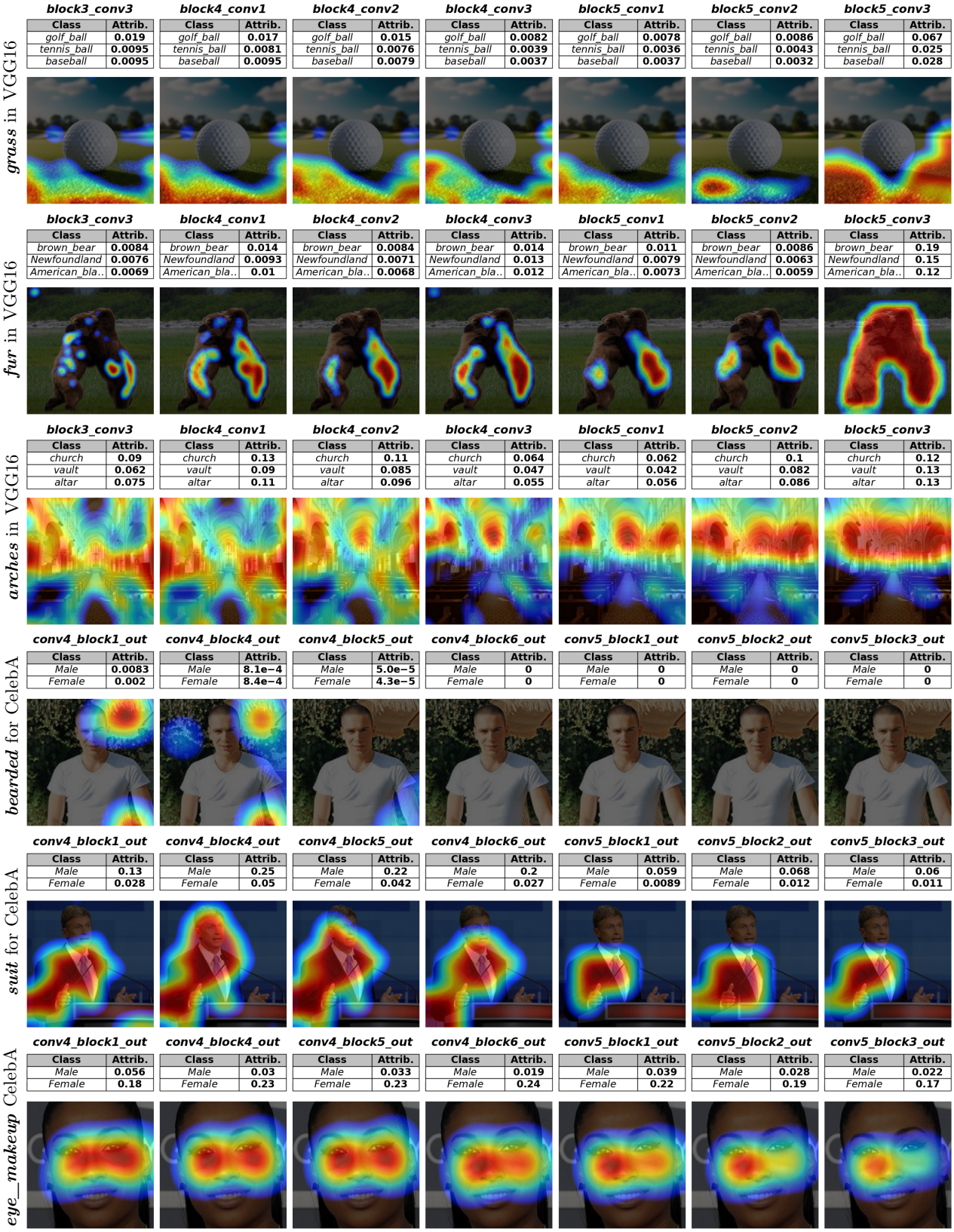


Figure 11: More examples of layer-wise local explanations for various concepts and networks. (Part 3)

D Additional results of Global Explanations

In Figure 12 we present additional global explanations for various classes and concepts.

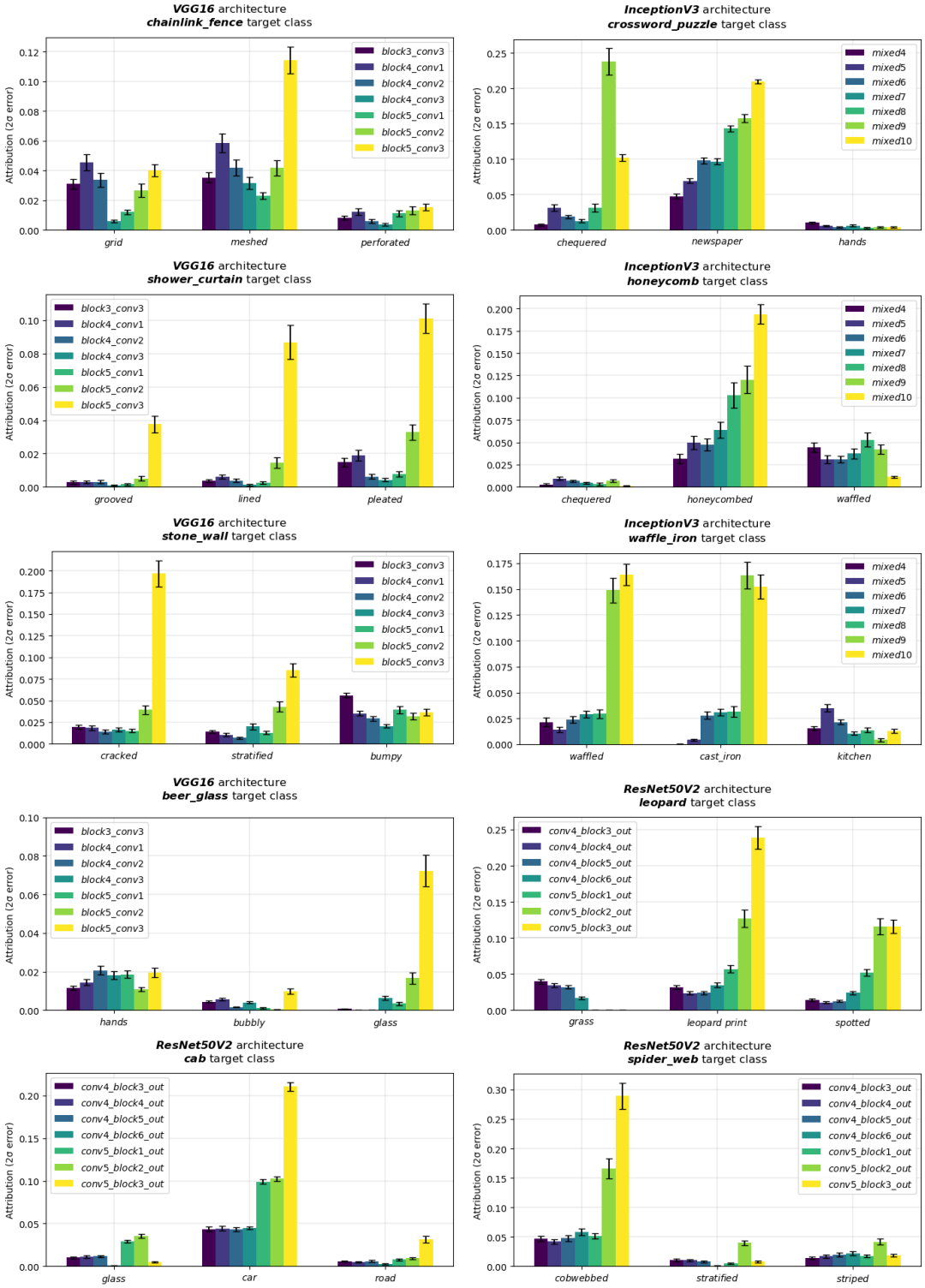


Figure 12: More examples of global explanations for various classes, concepts, and networks.

E Example images for generated concepts

Some of the concepts used in the paper were generated using Stable Diffusion v1.5 (Rombach et al., 2022) with default parameters. In particular, we generated the following concepts: "pews", "fresco", "arches", "sky", "pipes", and "brass". We used just the concept name as a prompt and generated 200 images per concept. A subsequent manual revision was still necessary to eliminate errors and artifacts. In Fig. 13, we provide three example images for each generated concept.

(a) Fresco (b) Arches (c) Pews (d) Sky (e) Brass (f) Pipes

Figure 13: We provide three example images for each concept generated with Stable Diffusion v1.5.

F Comparison with IG and Grad-CAM

F.1 Qualitative comparison

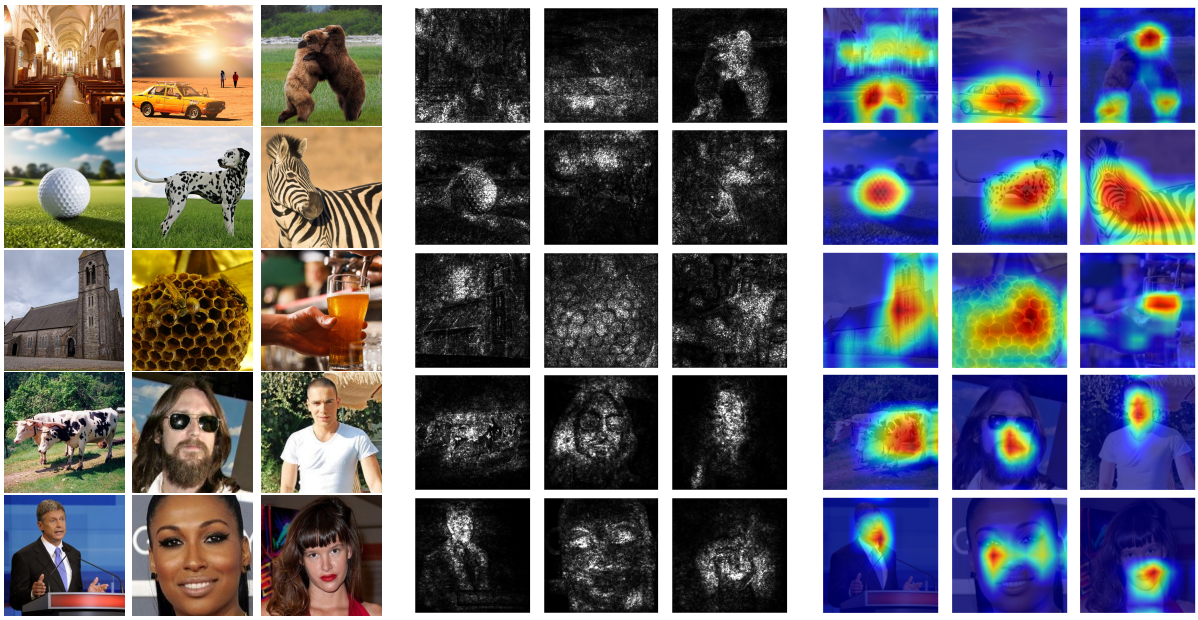
In Figure 14, we apply IG and Grad-CAM to some of the images analyzed in the paper, with the aim of qualitatively assessing the differences in output and scope between these methods and Visual-TCAV. As discussed in previous sections, saliency methods are among the most widely used approaches for explaining black-box image classifiers, as they enable the visualization of the most relevant regions of an image for a given prediction. However, these methods provide limited insights beyond identifying these regions. Specifically, they do not explain what features or concepts the model is identifying in those regions, how these features contribute to a specific prediction, and cannot provide global explanations.

Considering the "church (interiors)" image, IG reveals that the model primarily focuses on the pews, while Grad-CAM indicates attention to both the pews and the arches, although with lesser emphasis on the latter. These observations align with the results derived from Visual-TCAV (Figs. 3 and 11), which show that the concepts "pews" and "arches" are indeed important for the prediction, with "pews" more than "arches". Moreover, Visual-TCAV also provides a quantitative analysis of their contributions, explains "what" the network in seeing in the image in terms of the example images in Figs. 13b and 13c, and offers a global explanation by computing the average attribution of these concepts across multiple church images. Furthermore, from Visual-TCAV, we can also derive that the "arches" concept is not class discriminative in this case, as it has a similar importance for the classes "vault" and "altar".

Often key concepts may also remain undetected by saliency methods. For example, in the case of the "golf ball" and "bear" images, saliency methods fail to reveal that the "grass" concept is somewhat relevant for the

former and that the "fur" concept is highly significant for the latter. Instead, saliency maps mainly highlight the golf ball itself, while Grad-CAM focuses on the bear's head and feet. Additionally, these methods are unable to discern whether the model recognizes the dimples on the golf ball, its spherical shape, or both since these features visually overlap. In contrast, Visual-TCAV enables such distinctions by attributing these overlapping features to specific concepts with different importance.

Considering the saliency maps for the CelebA dataset, especially those generated by Grad-CAM, they are all very similar, making it difficult to understand whether the model learned to recognize specific attributes such as "beard", "lipstick" or "suit". However, while Visual-TCAV can provide explanations in terms of such concepts, other facial features can be very complex and are also difficult to define as human-understandable concepts, making this task challenging to interpret for any state-of-the-art XAI method.

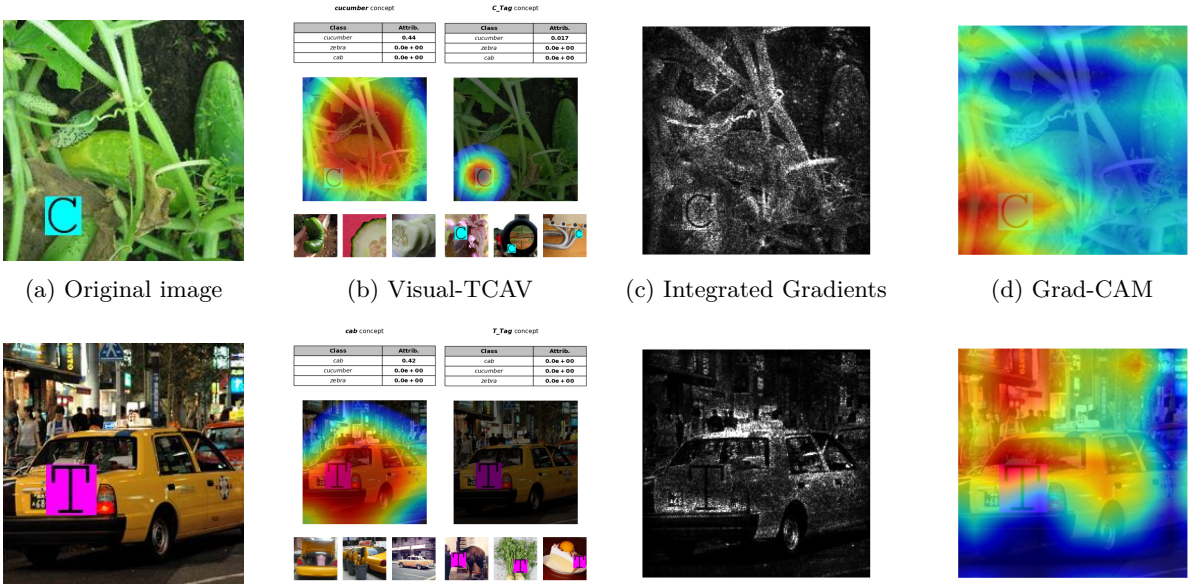


(a) Original Images (b) Integrated Gradients (c) Grad-CAM

Figure 14: IG and Grad-CAM for the images tested in the paper. Explanations of church (interiors), bear, and golf ball are from VGG-16. Explanations of church (outside), dalmatian, ox, and zebra are from Resnet50V2. Explanations of cab, honeycomb, and waffle iron are from InceptionV3. All explanations are computed in the last convolutional layer and considering the predicted class as target class.

F.2 Quantitative comparison

In this section, we present the results of a quantitative experiment designed to evaluate and compare the effectiveness of Visual-TCAV, IG, and Grad-CAM in communicating which concepts are more important in a given prediction. To perform this experiment, we took the "No Tags" and "100% Tags" models from the experiment in Section 4.4 and considered the classes "cucumber" and "taxi" with their respective "C" and "T" tags, as for these classes and models, we have a clear ground truth for explanations. Indeed, from Fig. 6, we know by construction that for the "No Tags" model, the entity itself (i.e., cucumbers and taxis) is more important than the tag, while for the "100% Tags" model, the tag is more important than the entity. We selected a total of 60 images, equally split across the two classes, and generated explanations using Visual-TCAV, IG, and Grad-CAM for the two models (for a total of 360 unique explanations). Each image-explanation pair was then presented to a Multimodal Large Language Model (GPT-5 in our case), which was asked to rate, based on the explanation, which concept between the entity and the tag appeared to be more important for the model's decision. A possible answer was also "Don't know" in case, based on the explanation, it was too difficult to tell which concept was more important.



(a) Original image (b) Visual-TCAV (c) Integrated Gradients (d) Grad-CAM

(e) Original image (f) Visual-TCAV (g) Integrated Gradients (h) Grad-CAM

Figure 15: Explanations from Visual-TCAV, IG, and Grad-CAM for the model trained without tags. In this model, the approximated ground truth is that the model pays much more attention to the entity than the tag (which was not present in the training set). While with Visual-TCAV the explanation matches this ground truth, this is not clear from IG and Grad-CAM.

For the "No Tags" model, the ground truth label was the entity, while for the "100% Tags" model, the ground truth label was the tag. "Don't know" responses were counted as incorrect. As for the metrics considered, for each method, we computed its accuracy (i.e., fraction of correct judgment over all images) and a statistical significance applying McNemar's test between Visual-TCAV and the two baselines (Grad-CAM and IG). The results in terms of accuracy are shown in Table 1, while the results of the McNemar's tests are shown in Table Table 2. Fig. 15 shows two examples where saliency maps communicated the wrong concept importance. Overall, Visual-TCAV was able to communicate correctly which concept was more important 98.33% of the time, while saliency methods (Grad-CAM and Integrated Gradients) achieved substantially lower accuracies of 71.67% and 65.83%, respectively. This suggests that explanations based solely on saliency maps may be misleading, which is the same conclusion reached in a similar experiment with human subjects by Kim et al. (2018). Statistical analysis using McNemar's test confirms that these results are statistically significant with p-values $\ll 0.05$. The exact prompts used in this experiment are provided below:

Visual-TCAV: You are an evaluator judging which visual concept a classifier relied on more strongly for a prediction based on the output of an explainability method.
The first image is the input to the classifier and was predicted as class <class>.

The explainability method used is called Visual-TCAV, and given a concept, it provides a heatmap showing whether the network recognized that concept and where, and the attribution of the concept towards each class.
The second image is the explanation for the concept <concept1>, and the third is the explanation for the concept <concept2>.
In the explanations, below the heatmap, are shown examples of images containing the concept being analyzed.
Question: Based only on these explanations, which concept appears more responsible for the model's decision?
Answer only with the name of the concept, or if you can't tell from the explanation, then answer "na".

IG & Grad-CAM: You are an evaluator judging which visual concept a classifier relied on more strongly for a prediction based on the output of an explainability method.
The first image is the input to the classifier and was predicted as class <class>.
The second image is the explanation.

The explainability method used is called <method>, and it provides a saliency map highlighting the most important image region for the prediction.
Question: Based only on the explanation, which feature appears more responsible for the model's decision between <concept1> and <concept2>?
Answer only with the name of the concept, or if you can't tell from the explanation, then answer "na".

Table 1: Comparison of accuracy and number of "don't know" responses across methods. Visual-TCAV achieves consistently higher accuracy than pure saliency approaches on both classes and overall.

Class	Method	Accuracy (%)	Don't know (%)
cucumber	Visual-TCAV	96.67	1.67
cucumber	Grad-CAM	60.00	23.33
cucumber	Integrated Gradients	51.67	0.00
taxi	Visual-TCAV	100.0	0.00
taxi	Grad-CAM	83.33	6.67
taxi	Integrated Gradients	80.00	0.00
Overall	**Visual-TCAV**	**98.33**	**0.83**
Overall	Grad-CAM	71.67	15.00
Overall	Integrated Gradients	65.83	0.00

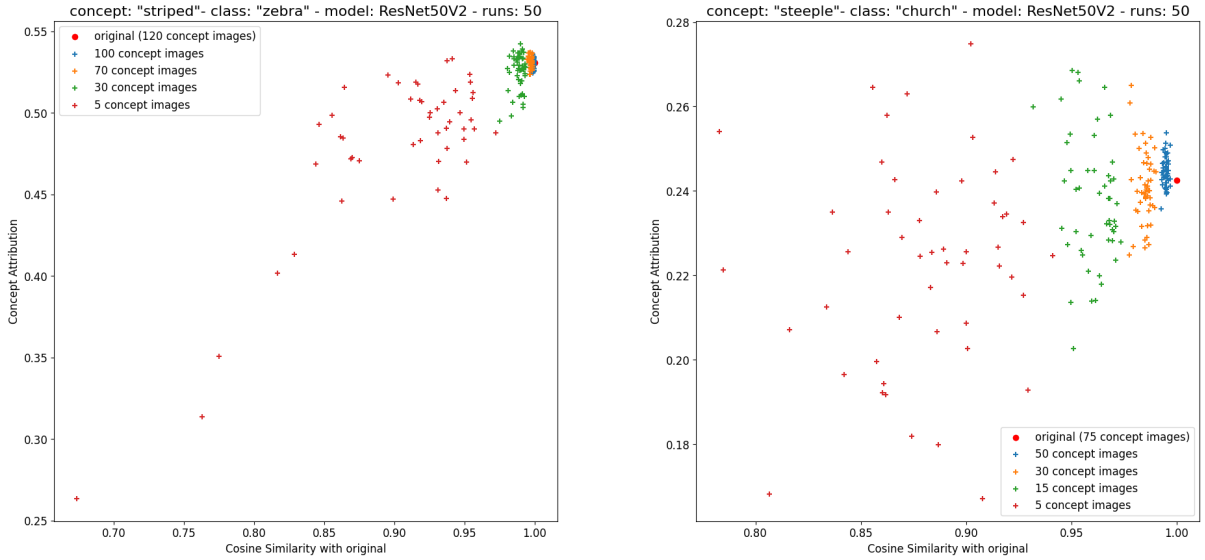
Table 2: McNemar tests comparing Visual-TCAV against Grad-CAM and IG. b = (Visual-TCAV correct, other method wrong) and c = (other method correct, Visual-TCAV wrong).

Class	Comparison	b	c	p-value
cucumber	Visual-TCAV vs Grad-CAM	24	2	1.05×10^{-5}
cucumber	Visual-TCAV vs IG	27	0	1.49×10^{-8}
taxi	Visual-TCAV vs Grad-CAM	10	0	1.95×10^{-3}
taxi	Visual-TCAV vs IG	12	0	4.88×10^{-4}
Overall	Visual-TCAV vs Grad-CAM	34	2	1.94×10^{-8}
Overall	Visual-TCAV vs IG	39	0	3.64×10^{-12}

G Experiments on Visual-TCAV stability

In this section, we experiment on the stability of Visual-TCAV. Specifically, we are interested in evaluating (i) the stability of Visual-TCAV with respect to the number of concept example images used for learning the pooled-CAVs, and (ii) the robustness of concepts when perturbed with random images. To address these points, we perform two separate experiments inspired by the procedure used by Martin & Weller (2019) to assess the stability of the *difference of means* and the TCAV Score. The experiments are all performed using the final layer of the same ResNet50V2 model, ensuring comparability of the results. Regarding the example images, we select two classes, "zebra" and "church", and perform an evaluation on the "striped" and "steeple" concepts respectively.

For the first experiment, starting from a concept defined through an arbitrary amount of training images (specifically, 120 images for "striped" and 75 images for "steeple"), we evaluate both the similarity of pooled-CAVs and the concept attribution when using a lower number of example images. To do this, we apply random sampling without replacement to the original concept images, and we apply Visual-TCAV to derive an updated pooled-CAV and the global concept attribution across the same 50 test images. We computed these metrics considering four different subsets of the concepts over 50 random runs, ultimately resulting in the plots of Fig. 16. From these plots, we observe that when considering both a texture-like concept such as striped, or a less homogeneous concept such as steeple, utilizing 30 example images instead of the original amount leads to a contained error for Visual-TCAV's concept attribution (i.e., with a mean absolute error lower than < 0.01 for both concepts). At the same time, the pooled-CAVs obtained with 30 concept images present a cosine similarity very close to the original. In general, these experiments suggest that 30-50 example images for defining a concept could be a reasonable amount. However, future studies on a large variety of concepts and networks are still necessary to draw definitive conclusions. Notably, in these experiments, even just 5 images were able to provide a rough estimate, even though the risk of misrepresenting the concept is slightly higher.

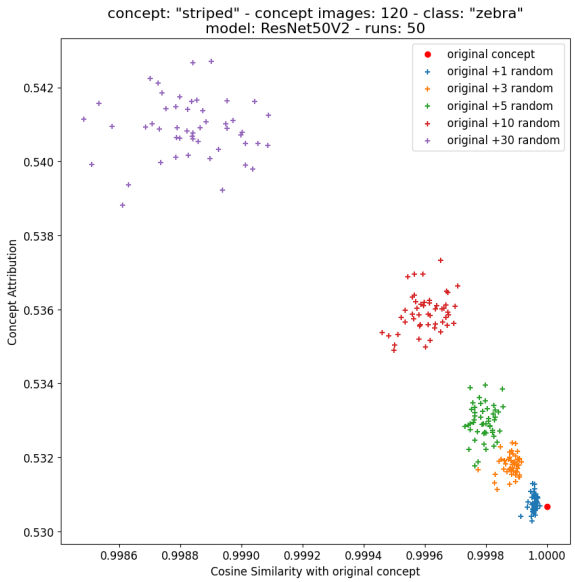


(a) "striped" concept

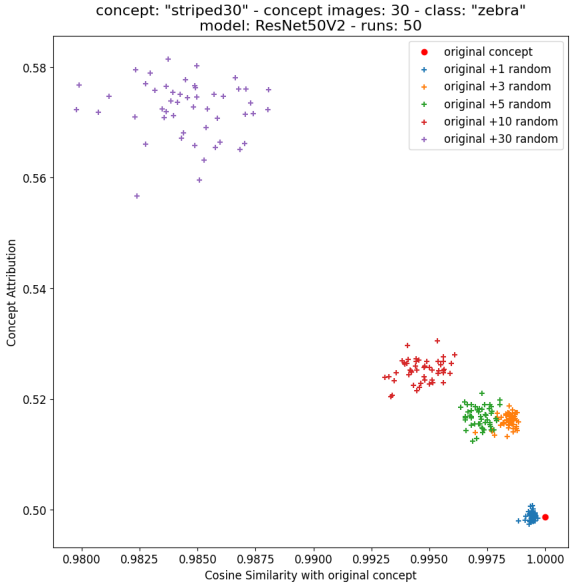
(b) "steeple" concept

Figure 16: Analysis of results for "striped" and "steeple" concepts computed with a different number of example images. Each point represents the concept attribution and pooled-CAV obtained by removing n example images from a concept set.

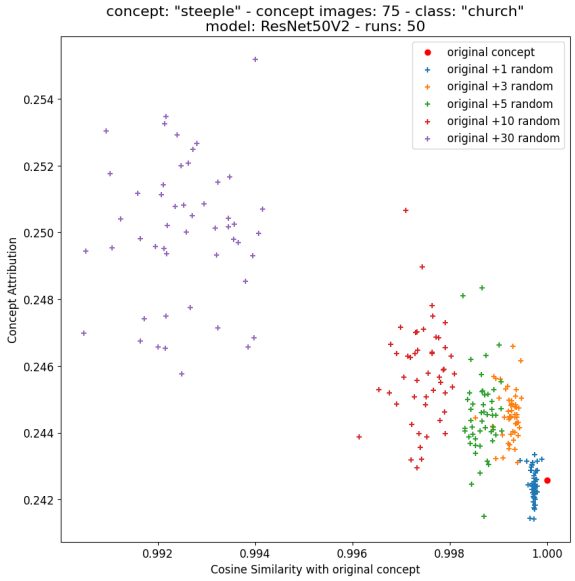
We perform a similar experiment to test Visual-TCAV's stability when the example images contain noise (i.e., random images not correlated with the concept). The purpose is to assess how careful an analyst should be when collecting example images. To this end, we apply random sampling without replacement to select a subset of images from the "random" concept and inject them inside the training images. By deriving the updated pooled-CAVs, the global concept attribution, and by repeating this procedure for an increasing number of noisy images, we can compare the deviation from the initial pooled-CAV representing the concept, and the variation of Visual-TCAV's concept attribution score. We perform this test both considering all the available example images and also with just 30 images. The results are shown in Figs. 17 and 18. From these plots, we observe that adding a few random images to the concept examples does not change the attribution significantly. The worst case is when adding 30 random images to 30 example images, in which we see a slight increase in attribution. This can be explained by the fact that we are also considering the attribution of the background, which also includes some noise accumulated by the integrated gradients. Another effect of adding many random images is that the upper bound to activate the concept map also decreases slightly, as it depends on the concept examples. Overall, from these experiments, we can conclude that results from our method are meaningful even when adding a significant amount of noise (up to 100%) to the example images. In Fig. 19, we also provide examples of concept maps computed with different amounts of example images and with added noise.



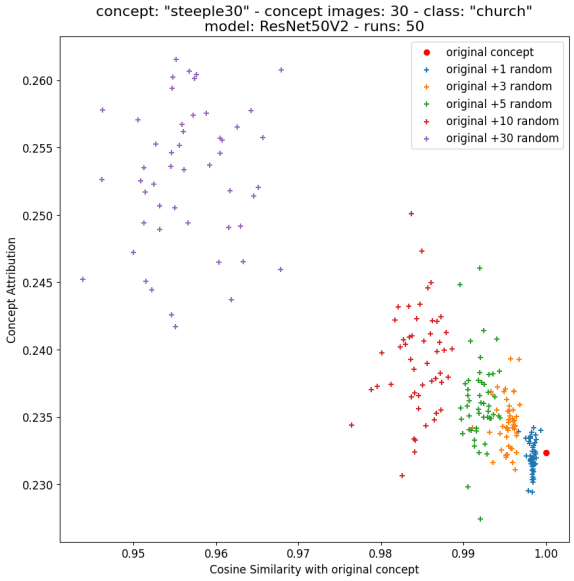
(a) "striped" with 120 example images



(b) "striped" with 30 example images



(c) "steeple" with 75 example images



(d) "steeple" with 30 example images

Figure 17: Analysis of results for "striped" and "steeple" concepts computed with a different number of example images and with random images added as noise to the concept set. Each point represents the concept attribution and pooled-CAV obtained by adding N random images to the example images of a concept.

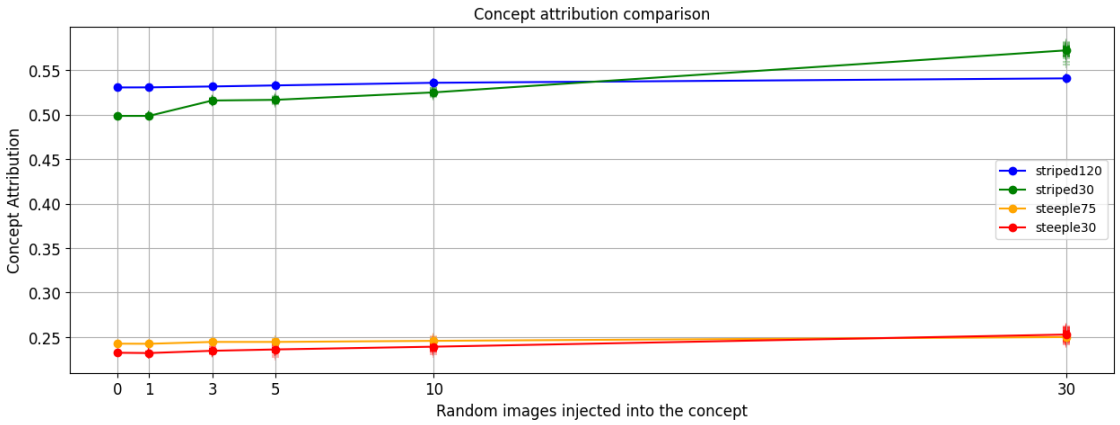
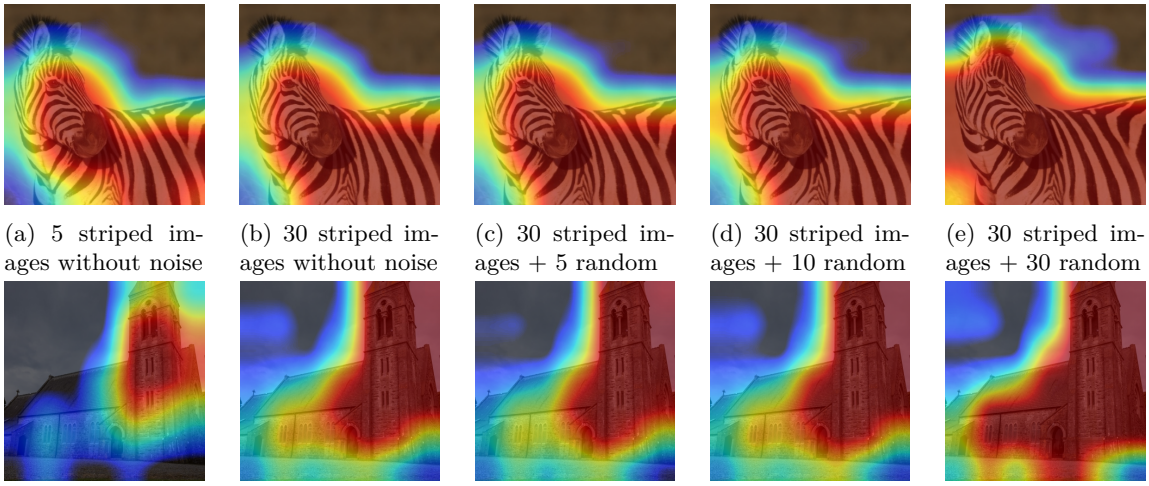


Figure 18: Summary of the concept attribution results for the second experiment. The line plots represent the variation of the average concept attribution when adding N random images to the example images of a concept.



(a) 5 striped images without noise (b) 30 striped images without noise (c) 30 striped images + 5 random (d) 30 striped images + 10 random (e) 30 striped images + 30 random

(f) 5 steeple images without noise (g) 30 steeple images without noise (h) 30 steeple images + 5 random (i) 30 steeple images + 10 random (j) 30 steeple images + 30 random

Figure 19: Examples of concept maps computed with different amounts of example images and with different amounts of added noise via random images.