Automating interpretability: discovering and testing visual concepts learned by neural networks

Amirata Ghorbani*
Stanford University
amiratag@stanford.edu

James Wexler
Google Brain
jwexler@google.com

Been Kim
Google Brain
beenkim@google.com

Abstract

Interpretability has become an important topic of research as more machine learning (ML) models are deployed and widely used to make important decisions. For high-stakes domains such as medical, providing intuitive explanations that can be consumed by domain experts without ML expertise becomes crucial. To this demand, concept-based methods (e.g., TCAV) were introduced to provide explanations using user-chosen high-level concepts rather than individual input features. While these methods successfully leverage rich representations learned by the networks to reveal how human-defined concepts are related to the prediction, they require users to select concepts of their choice and collect labeled examples of those concepts. In this work, we introduce DTCAV (Discovery TCAV) a global concept-based interpretability method that can automatically discover concepts as image segments, along with each concept's estimated importance for a deep neural network's predictions. We validate that discovered concepts are as coherent to humans as hand-labeled concepts. We also show that the discovered concepts carry significant signal for prediction by analyzing a network's performance with stitched/added/deleted concepts. DTCAV results revealed a number of undesirable correlations (e.g., a basketball player's jersey was a more important concept for predicting the basketball class than the ball itself) and show the potential shallow reasoning of these networks.

1. Introduction

As machine learning (ML) has become a widely used tool in many applications from medical (e.g., [13]) to commercial [27], gaining insights into ML models' predictions has become one of the most important topics of study, and sometimes even a legal requirement [12]. The industry is also recognizing interpretability as one of the main components of responsible use of ML[1]; not just a nice-to-have component but a must-have one. The ability to understand and interact with ML tools is one of the crucial factors to decide whether ML should be implemented in high risk domains with potentially severe consequences (e.g. medicine).

One of the unique challenges of interpretability in high stakes domains is that the users may not be very familiar with ML. This calls for using more intuitive interpretability language designed for laypersons. However, most of the developments in interpretability methods have been using less intuitive language, mostly focused on estimating how important each input feature is for prediction [24, 25, 33]. While this is a useful tool for explaining the prediction of a single data point (local explanation [7]), the limitations of this method has been repeatedly shown to be unreliable. These limitations include potential methodological weakness (e.g., the importance measure has little to do with the prediction, contradicting its promise[14]), vulnerability to adversarial attacks [11], and susceptibility to human confirmation biases [16]. In other words, using pixels as a medium requires the subjective judgment of humans, and some studies have shown that humans are able to find evidence for completely contradicting conclusions [16]. We argue that this might be partially due to the fact that humans do not think or communicate using pixels.

Some recent interpretability methods aim to overcome this by generating quantitative explanations using high-level 'concepts' (e.g., diagnostic concepts, gender, race) instead of input features. TCAV [16] uses a user-chosen set of example data points to form a concept activation vector (CAV), which then is used to calculate the importance of the concept for a prediction. This method uses intuitive language for a layperson to express concepts of interest and to understand their model through those concepts. However, the user has to have a set of concepts in mind for testing and provide examples of such concepts. What if users do not have candidates concepts and/or have ways to provide examples? What if the space of plausible concepts to test is exponentially large?

In this work, we introduce DTCAV (Discovery TCAV) which automatically discovers concepts by collecting connected parts of images (segments) that together form im-

*Work performed while interning at Google Brain.

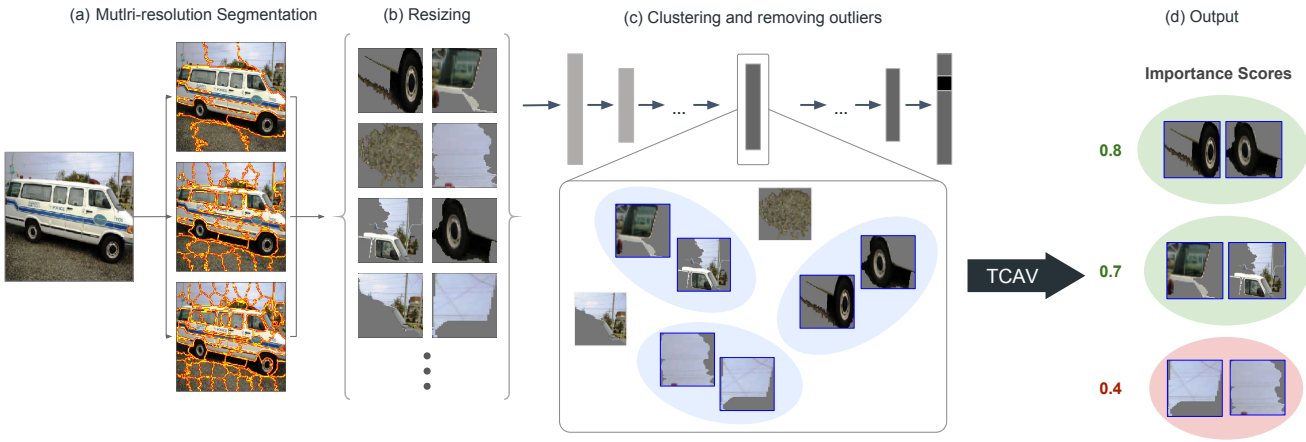

Figure 1: **DTCAV algorithm** (a) Given a set of concept discovery images, each image is segmented with different resolutions to find concepts that are captured best at different sizes. (b) After removing duplicate segments, each segment is resized to the original input size resulting in a pool of resized segments of the discovery images. (c) Resized segments are mapped to a model's activation space at a bottleneck layer. To discover the concepts associated with the target class, clustering with outlier removal is performed. (d) The output of our method is a set of discovered concepts for each class, sorted by their importance in prediction

portant concepts. We validate via human experiment that the learned segments form concepts as coherent as human-labeled concepts. We further validate the learned concepts by showing that these concepts segments alone often carry sufficient information to be predicted as the corresponding class. We also add and remove sets of segments sorted by their importance in prediction and show the resulting significant impacts on the prediction.

2. Related work

This work focuses on post-training interpretability methods - finding explanations given an already trained network. While there is a line of research on building inherently interpretable models [32, 15, 29], we focus on scenarios where we cannot modify the model. Most common post-training interpretability methods provide explanations by estimating the importance of each input feature or training sample for the prediction of a particular data point [24, 25, 33, 17]. Naturally, these methods can only explain one data point at a time (local explanation).

While this is useful when only specific data points matter, these methods have been shown to come with many limitations, both methodologically and fundamentally. For example, [14] showed that some input feature-based explanations are qualitatively and quantitatively similar for a trained model (i.e., making superhuman performance prediction) and a randomized model (i.e., making random predictions). This shows that the explanation may have little to do with prediction, contradicting its goal of explaining predictions. Other work proved that some of these methods are in fact trying to reconstruct the input image, rather than estimating pixels' importance for prediction [28]. In addition, it's been shown that these explanations are susceptible to humans' confirmation biases [16]. For example, [16] showed that given identical input feature-based explanations, human subjects confidently find evidence for completely contradicting conclusions. Using input features as explanations also introduces challenges in scaling this method to high dimensional datasets (e.g., health records). Humans typically reason in higher abstracted concepts ([20]) than a particular input feature (e.g., lab results, a particular hospital visit).

A recently developed method uses high-level concepts, instead of input features. Given a set of examples of a concept of user's choice, TCAV [16] produces estimates of how important that a concept was for the prediction. However, users have to provide examples of the concept, limiting this method to cases when users have a set of concepts in mind and are interested their importance measures.

Our method leverages multi-resolution image segmentation methods to There has been a long line of research on multiscale and hierarchical segmentation of images ([23, 30, 4]). In this work, we use the SLIC [2] superpixel segmentation method for its simplicity, memory efficiency, speed, and high quality performance, as shown in [3].

3. Methods

In this section, we first review TCAV, a concept based interpretation method for interpreting deep neural networks. We then introduce our method, Discovering and Testing Concept Activation Vectors (DTCAV), by first describing what

we define as concepts and how we discover them and then completing the description by testing discovered concepts.

3.1. Testing Concept Activation Vectors (TCAV)

TCAV [16] is a post-training interpretability method that calculates how important a user-chosen concept is for a deep neural network's prediction of a target class, e.g. how important is stripedness for predicting the zebra class. A user first provides a set of example data points of the chosen concept together with random data points that do not belong to the concept (i.e., a random counterpart). Then data points are then mapped to the activation space of a bottleneck layer of the user's choice. Concept Activation Vectors (CAVs) are defined as the direction orthogonal to a linear classifier trained to distinguish the concept activations from the random activations. The importance of the concept is generated using the directional derivative of the prediction unit of a particular class with respect to the calculated CAV. The final TCAV score is simply an aggregated statistic of these directional derivatives for many images from the target class. Intuitively, the TCAV score measures how important the concept (represented as a CAV) is for a class prediction by conducting a form of sensitivity test [22]. In order to reject any concepts that were "not learned" by the network, a statistical testing between TCAV scores with multiple CAVs of the same concept (using different random counterparts) and that with random CAVs (using random images in the place of concept images) is conducted. TCAV output only includes concepts that pass this test.

While TCAV allows a layperson to express their concept of interest and conduct hypothesis testing of their choice, users are in charge of selecting domains where users have a clear set of concepts in mind, it may not be suitable for domains where users simply do not know which concepts to test or have the resources to collect concept examples.

Our method was inspired by the following questions: If we limit the space of concepts to a set of target class images, can we automatically discover concepts contained in them? Can we discover concepts that are coherent to humans while sufficient for prediction?

3.2. Discovering and Testing Concept Activation Vectors (DTCAV)

At a high level, the DTCAV method first segments target class images (i.e., "discovery images") and applies simple clustering methods followed by outlier removal to finalize the set of discovered concepts. The output of the method is a number of sets of segments of the discovery images, each set representing a concept, together with a quantitative measure how important each set of concepts is.

Fig. 1 shows the overall algorithm in detail. First, we create a set of images belonging to the target class that we call "discovery image". For each discovery image, segmentation

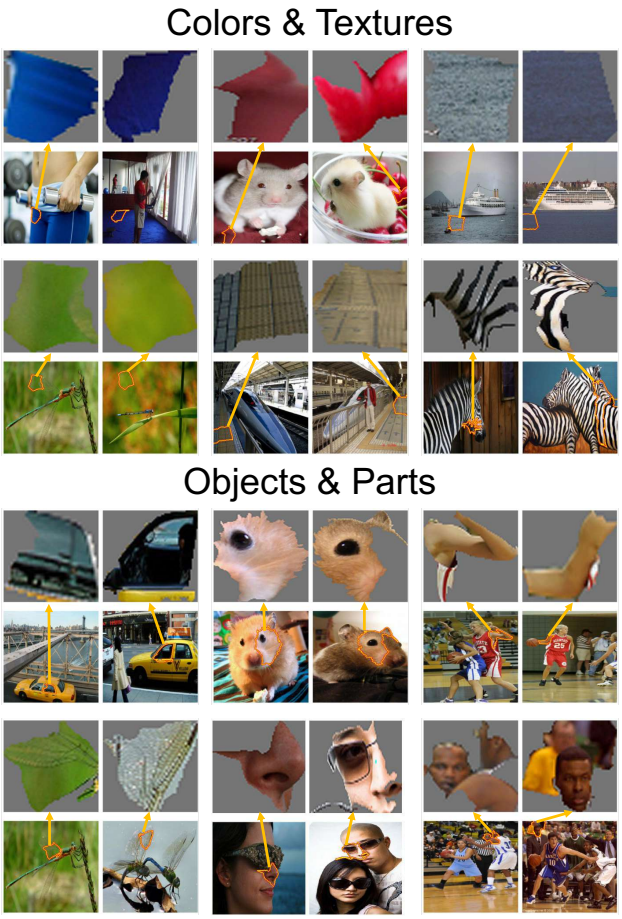

Figure 2: **Examples of discovered concepts.** A wide range of concepts like blue color, asphalt texture, car window, and human face are detected through the algorithm. Multi-resolution segmentation helped discovering concepts with varying sizes. For example, two car windows with different sizes (one twice as big as the other) were identified as the same concept.

is applied several times with different levels of resolution; for instance, using superpixel segmentation with various parameters resulting in different number of segments.(Fig. 1(a)) Each segment is then resized to the original input size of the network and mapped to a chosen bottleneck layer's activation space.(Fig. 1(b)) Clustering with outlier removal is then applied to the activations of the segments to discover the concepts. (Fig. 1(c)) A new set of images of the target class is used to calculate TCAV importance scores using the method described in [16]. (Fig. 1(d)) While the final method above is simple, each piece in the method address many inherent challenges of concept discovery. The first challenge is that discovered concepts must be location and scale invariant, since the same concept may appear multiple times in different scales and locations in images. An

efficient multi-resolution segmentation method (SLIC superpixel segmentation[2]) is crucial as one image is segmented with multiple resolutions (Fig. 1(a)). Since doing so may create duplicated segments, we use Jaccard similarity to remove potential duplicates. The second challenge is effective filtering for potentially important concepts that are coherent. In other words, we want to filter out potentially irrelevant segments, e.g. a human face appearing in one zebra image. We empirically identified three simple but important factors: distance, frequency and popularity. The distance factor is intuitive - we remove segments that are far from all of the clusters. The frequency factor means that segments in a cluster must occur across many images (frequency) and not just small number of images. The popularity factor simply means that the cluster also must be big enough to be a good candidate. We filter clusters where neither of these factors are satisfied (details in Section 4).

After filtering we have set of candidate concepts then are used to compute the concept activation vectors (CAVs) and perform statistical testing to obtain TCAV scores.

A number of previous literature supports our assumption that clustering method is surprisingly effective in distinguishing concepts. An experiment of [16] where in the right bottleneck layer of an image classification network, images of a concept are linearly separable from random images using various sets of random images. [5] also verify that simple linear classifiers were sufficient for discovering concepts. Another evidence is by [34] that pointed out striking similarities in deep neural network's learned representations with that in human perception. A comprehensive discussion of linearity in deep neural networks activation space is provided in [16].

Note that DTCAV is not limited to using TCAV scores to measure the importance. Once the concepts are discovered, for example, one could use gradient-based importance measures like saliency maps. The averaged value of importance scores for all pixels that fall into the segment could be used as a proxy for importance measure. One can also use cosine similarity between CAVs of each concept and CAVs of target images as a measure.

4. Experiments & Results

4.1. Datasets and Implementation Details

All experiments were performed using Inception-V3 model [26] trained on the ILSVRC2012 data set (Imagenet) [21]. We randomly chose 100 out of 1000 classes in the data set for our experiments. We used "mixed 8" bottleneck for this section.

As described, the first step of the DTCAV algorithm involves multi-resolution segmentation. Several superpixel segmentation methods were examined ([2, 9, 19, 31]). The simple SLIC[2] method was chosen as it strikes good balance

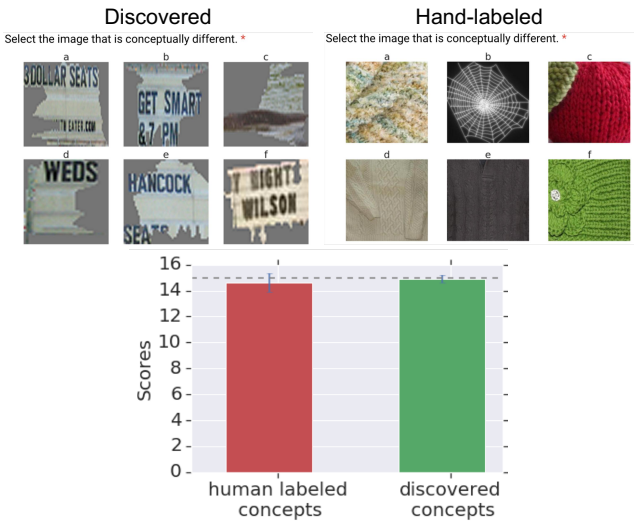

Figure 3: **Human subject experiment questionnaire and results.** 30 human subjects were asked to identify one image out of six that is conceptually different from the rest, mirroring interpretability literature [6]. We show a combined set of discovered and hand-labeled concepts for comparison. On average, participants answer the hand-labeled dataset 97% (14.6/15, ± 0.7) correctly, while discovered concepts were answered 99% (14.9/15, ± 0.3) correctly. The grey dotted line represents a perfect score.

between quality of segments and efficiency. We performed three-resolution segmentation by changing SLIC's number of segments parameter to 15, 50, and 80. After resizing each segment, since segments are in irregular shapes, we fill in the empty part of the image with the zero pixel value (117 in our network, after post-processing). For the choice of cluster, we performed concept discovery using several clustering methods including K-means [18], Affinity Propagation [10], and DBSCAN [8]. When Affinity Propagation was used, typically a large number of clusters (30-70) were produced, which was then simplified by another hierarchical clustering step. The best results, however, were acquired using k-means clustering followed by removing all points but the n points that have the smallest ℓ_2 distance from the cluster center. For filtering, as described in Section 3, we remove all but a) high frequency (segments come from more than $1/2$ of discovery images) b) medium frequency with medium popularity (more than $1/4$ of discovery images and the cluster size is larger than $S/2$) and c) high popularity (cluster size is larger than S). In all the following experiments, $n = 40$ and $S = 80$, and we use k-means with $k = 25$; 50 images of training set were used for concept discovery.

In what follows, we first show examples of the discovered concepts using DTCAV algorithm. We first verify that our method returns coherent sets of concepts via human subject

Figure 4: **Examples of DTCAV.** For each target class, four discovered concepts with high TCAV scores and the one with the lowest score are shown. We randomly chose three segments in each concept. Notice that a TCAV score of 0.91 means that 91% of the images in the target class returned a positive directional derivative, indicating that the discovered concept was important for the network's prediction. For example, the fins of the lionfish, the letters on a police van, and basketball jerseys are all highly associated with their respective classes, while asphalt roads or floors have low or no association. (* stands for a concept that did not shown to be statistically different from random concepts, meaning that TCAV did not find the concept to be meaningful for the target class.)

experiment. The results indicate that the discovered concepts are as coherent to humans as hand-labeled concepts. We show that our method is able to learn various abstract levels of concepts; from simple concepts (e.g., color, texture) to more complex ones (e.g., objects, parts). We also quantitatively verify that these concepts were in fact crucial for prediction. First, we show that a set of important concepts are enough to predict the right class. Second, we show that adding or deleting important concepts significantly impacts the prediction performance.

4.2. DTCAV can discover simple to complex concepts

The multi-resolution segmentation step of DTCAV naturally returns segments that contain simple concepts such as color or texture and more complex concepts, such as parts of body or objects. Among those segments, DTCAV successfully identifies concepts with similar level of abstract-ness with similar semantic meaning (as verified via human experiment). Fig. 2 shows some examples of the discovered concepts. Note that each segment is re-sized for display.

4.3. Discovered concepts are semantically coherent to humans

We designed an intruder detection human experiment following interpretability literature [6] to verify the quality of the discovered concepts. At each question, a subject is asked to identify one image out of six that is conceptually different from the rest. We created a questionnaire of 34 questions, such as shown in Fig. 3 Among 34 randomly ordered questions, 15 of them include a set of randomly chosen DTCAV concepts, and the other 15 questions are human-labeled concepts from Broaden dataset [5]. The first four questions were used as training and discarded. On average, participants answered the hand-labeled dataset 97% (14.6/15) ($\pm$0.7) correctly, while discovered concepts were answered 99% (14.9/15) ($\pm$0.3) correctly. This experiment confirms that while a discovered concept is only sets of segments of images, DTCAV was able to identify segments with coherent concepts.

4.4. Examples of DTCAV with high and low TCAV scores

For each discovered concept, we compute its CAV and then test the CAVs using a set of held-out images of the target class to get the TCAV score of each discovered concept. Note that a TCAV score of 0.9 means that 90%of the target class images have positive sensitivity to the concept; intuitively, this means that increasing the presence of that concept increases the prediction score of the class. Fig. 4 shows the result for running DTCAV on a subset of target classes. For each class we show four concepts, some with high and low TCAV scores and some concepts that did not pass the statistical test (i.e., the concept was not relevant to the prediction). Three randomly selected segments are shown for each concept. More examples are provided in Appendix B.

4.5. Insights from discovered concepts

Reviewing discovered concepts with high TCAV scores shows what the network pays attention to, which reveals some surprising correlations. In some cases, we see that the network picked up on appropriate related concepts. The letters in police van were correctly identified as important in Fig. 4, while the asphalt road in the background was identified as not important. Fig. 5(a) shows more examples of this case. Not surprisingly, we discover some undesirable correlations. For example, the lionfish prediction considers the background reef to be important, and basketball predictions consider player jerseys and the wooden floor important instead of the ball, as seen in Fig. 4. Some classes such as European gallinule Fig. 4, network considers the background (grass) much more important (TCAV score 0.73) then parts of the bird (TCAV score 0.46). This indicates that this classifier may not be great for robustly detecting this bird, and

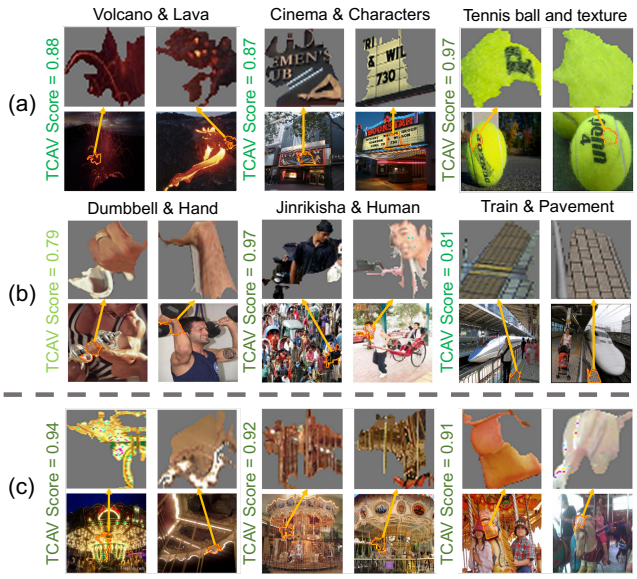

Figure 5: (a) desirable correlations (b) undesirable correlations (c) different parts of an object identified as separate but important concepts

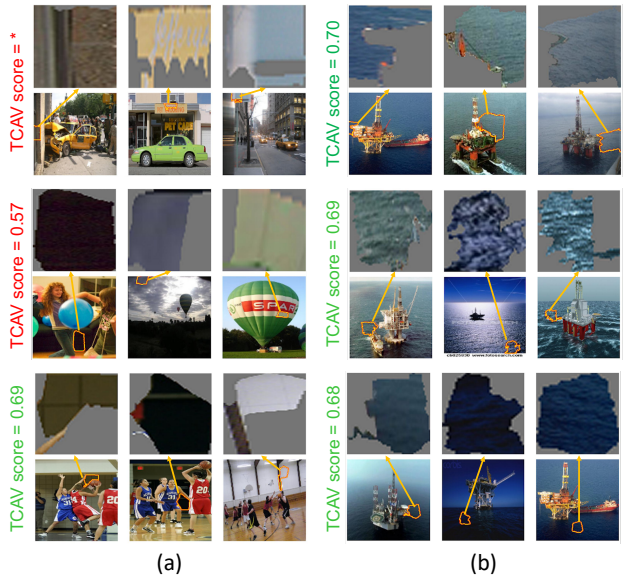

Figure 6: (a) Semantically inconsistent concepts achieve low or no TCAV scores (b) Seemingly duplicated concepts (to humans) may be discovered

that gathering more training data with various background might improve the result. Similarly undesirable correlations are shown in Fig. 5(b).

Another insight we gained was that in some cases when the object is complex, the network identifies parts of the object as separate concepts, and some parts are more important than others. For example, in Fig. 5(c), carousel lights, poles

structure, and seats (horses). It is interesting to learn that the lights were more indicative of the carousel than seats.

Note that some of the discovered concepts may seem duplicated to humans. For example, in Fig. 6(b), three different ocean surfaces (wavy, calm, and shiny) are discovered separately and all of them have similarly high TCAV scores. Future work remains to see whether this is because the network represents these ocean surfaces differently, or whether we can further combine these concepts into one 'ocean' concept.

While rare, there are concepts that are less coherent to humans. This may be due to limitations of our method or because things that are similar to the neural network are not similar to humans. However, the incoherent concepts were never in top-5 most important concepts among the 100 classes used for experiments.

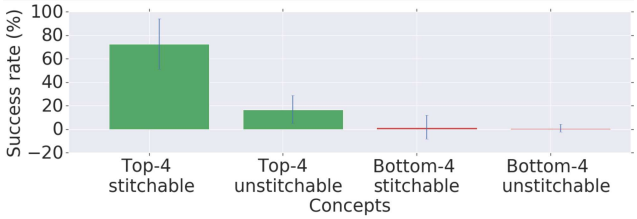

Figure 7: **Average results of stitching experiment.** The success rate (the ratio of stitched images with correct predictions) of pattern-based(stitchable) classes along with others. When concepts with low TCAV scores are stitched, network's success rate hovers around 1%. For the 100 classes, about half of the classes are defined as "stitchable", obtaining more than a 70% average success rate. Even when classes are defined as "unstitchable", the success rate is much higher than the bottom-4 concepts.

4.6. What does the network see if we stitch concepts together?

The discovered segments only contain a part of the story of the target class, especially since it loses the global structure of the object (e.g., shape). However, it is plausible that sometimes the mere presence of important concepts of a target class is sufficient for classification without considering the global structure of the class images. For example, zebra pattern could be distinctive enough that stitching together zebra skin textures may convince the network that it is a zebra, without having to see the anatomy of the zebra. To this end, we designed a concept stitching experiment where we randomly place concept segments on a blank image in a sorted order of importance.

For each target class, we experimented stitching top-k highly important concepts and generated 100 stitched images for each experiment. We then picked the experiment yielding the highest "success rate", which is the percentage

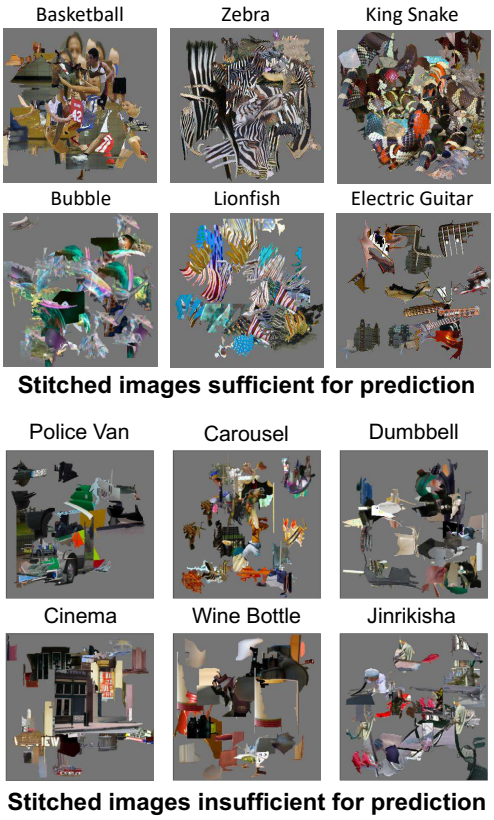

Figure 8: This experiment tests whether important concepts are sufficient by stitching together randomly chosen concept segments with high TCAV scores. We discover that for classes with a strong pattern, that is the case (top). For instance, basketball jerseys, zebra skin, lionfish, and king snake patterns all seem to be enough for the Inception-V3 network to classify them as elements of their target class. On the other hand, the network was not able to predict some classes via stitched images which may indicate that they require more 'structure' to be predicted correctly (bottom).

of stitched images classified as that target class (i.e., accuracy with stitched images). We choose k in top-k via greedy cross-validation. An average success rate of 39.6% was obtained (note that random chance is 0.001%). As a control, we also ran the experiment of stitching the bottom-5 concepts yielding a 1% success rate. Interestingly, for zebra, leopard, and drilling platform classes, the success rate is relatively high (more than 80%) which shows that the network is only looking at important concepts (Fig. 8). On the other side, police van, jinrikisha, and bullet train classes, the success rate is close to zero which means that the general structure of the class images is also necessary for correct classification. Examples of these classes are shown in Fig. 8. Aggregate results are shown in Fig. 7 where we group the classes into the ones with a success rate more than 40%, which we consider

"stitchable" classes, and the other classes which we consider "unstitchable". This experiment shed insights on whether the global structure of the class was crucial for the prediction or not, revealing potentially shallow reasoning of the network.

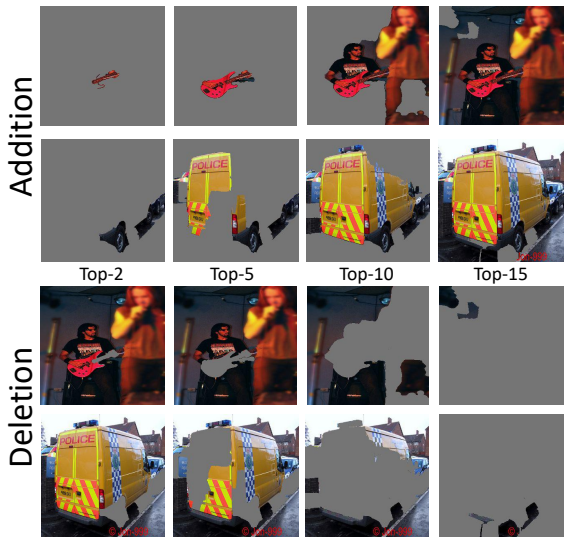

Figure 9: Example of sequential addition and deletion (from left to right) experiments for the guitar and police van classes.

4.7. Adding or deleting discovered segments significantly impacts the prediction.

In this experiment, we show the effect of adding or deleting important concepts from images. The idea is that if a segment's respective concept is indeed important, then deleting/adding that segment should decrease/increase the network's ability to predict more so than random deletion/addition.

For a set of test images, we add or delete segments with respect to their associated concept's TCAV scores. Then we track the prediction accuracy, one from highest scores (blue line in Fig. 9) and one from lowest scores (red line from Fig. 9). To find each segment's associated concept, we find its nearest neighbor concept cluster in the bottleneck layer's activation space space. Fig. 9 shows two examples of such addition/deletion.

The results in Fig. 10 show that the discovered concepts carry important signal for prediction; a small number (5) of top concepts are sufficient to predict 90% of images correctly, while the bottom concepts only achieve 20% accuracy. Note that the relative performance with randomly-ordered concepts further supports these results. Deleting top/bottom concepts also lead to the same conclusion. When the top-5 concepts are removed, more than 60% of originally correctly predicted images were no longer correctly predicted.

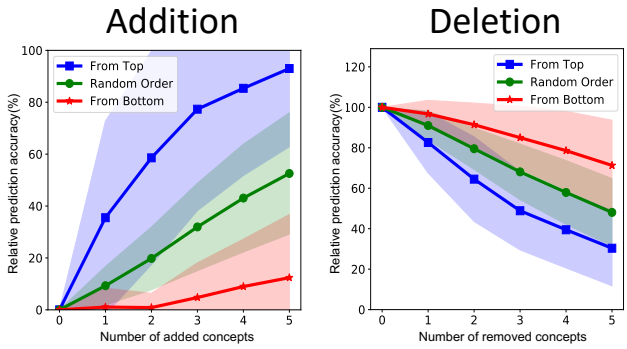

Figure 10: **Addition-deletion experiment results** Addition-deletion experiments seek to verify the importance of concepts with high TCAV scores for the network's prediction task. As the addition test shows, the presence of only top-5 important concepts is enough for reaching 90% of the original accuracy. The deletion experiment shows that around 70% of the original accuracy is lost by removing the top-5 important concepts (blue lines). When we add or delete the bottom-5 concepts (red lines), we see minimal impact on the accuracy. Random order of random concepts (green line) have an effect between the top and bottom concepts.

5. Discussion and conclusion

We note a couple of limitations of our method. This work is based on image data sets, as the super-pixel segmentation method is limited to images. While the general idea of discovering and testing concepts does apply to many other data types such as texts, it was not tested. Additionally, our method only can discover concepts that can be expressed with pixels. While we still discover plenty of insights based on pixel segments, there might be more complex and abstract concepts that we are unable to discover. Future work includes better optimizing our method's performance by tuning the multi-resolution segmentation parameter per class. This may better capture the inherent granularity of objects; nature scenes may have a smaller number of concepts than city scenes. For example, the "European gallinule" class in Fig 4 could have been benefited from a segment of the entire bird itself.

In conclusion, DTCAV is a post-training interpretability method that automatically discovers high-level important concepts in the form of image segments. We verified that the diverse set of discovered concepts are coherent to human via human experiment, and further validated that discovered concepts are indeed carry important signals for prediction. The discovered concepts reveal insights into surprising and sometimes undesirable correlations that the network has learned, highlighting networks' frequent shallow reasoning. Such insights may help to promote safer use of this powerful tool, machine learning.

References

[1] Google AI principles. https://www.blog.google/technology/ai/ai-principles/. Accessed: 2018-11-15.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, et al. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, et al. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[4] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3378–3385. IEEE, 2012.

[5] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

[6] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

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

[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[10] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[11] A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. *arXiv preprint arXiv:1710.10547*, 2017.

[12] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

[13] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[14] M. M. I. G. M. H. B. K. Julius Adebayo, Justin Gilmer. Sanity checks for saliency maps. *NIPS*, 2018.

[15] B. Kim, C. Rudin, and J. A. Shah. The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960, 2014.

[16] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2673–2682, 2018.

[17] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

[18] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[19] P. Neubert and P. Protzel. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 996–1001. IEEE, 2014.

[20] E. Rosch. Principles of categorization. *Concepts: core readings*, 189, 1999.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[22] A. Saltelli. Sensitivity analysis for importance assessment. *Risk analysis*, 22(3):579–590, 2002.

[23] E. Sharon, A. Brandt, and R. Basri. Fast multiscale image segmentation. In *cvpr*, page 1070. IEEE, 2000.

[24] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[25] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[27] N. Tintarev and J. Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook*. Springer, 2011.

[28] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Deep image prior. *CoRR*, abs/1711.10925, 2017.

[29] B. Ustun and C. Rudin. Methods and models for interpretable linear classification. *ArXiv*, 2014.

[30] D. Varas, M. Alfaro, and F. Marques. Multiresolution hierarchy co-clustering for semantic segmentation in sequences with small variations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4579–4587, 2015.

[31] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision*, pages 705–718. Springer, 2008.

[32] F. Wang and C. Rudin. Falling rule lists. In *AISTATS*, 2015.

[33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint*, 2018.

A. Extra details about the DTCAV algorithm implementation

As mentioned in Section 3 and Section 4, in order discover concepts that are frequently present in a target class, we perform unsupervised clustering of reszied image segments in the bottleneck layer's activation space. Outlier removal is then performed to remove unrelated members of a cluster. In order to make sure that a discovered concept is actually present in target class images, we introduced three different types of concepts that are acceptable:

- If the members of a concept's cluster come from majority of discover images, in other words, the segments forming that concept's cluster are parts of a large number of discovery images, it could be said that the discovered concepts frequently appears in the target class images. In the experiments, any concept appearing in more than 50% of discovery images is considered to be acceptable. On example would be the ball in the basketball class. It's present in every image and usually there is one of it. As a result, its respective cluster is not large but it has members coming from a large number of discovery images.

- If the segments in a concept cluster, appear in a reasonable number of discovery images but the cluster size is large, it means that the concept has a significant presence in part of the images belonging to the target class. In our experiments, if a concept appears in 25% to 50% of images but its cluster has more than 40 members, it is an acceptable concept. One example would be the hand object in the basketball class. Many of the basketball images do not have a hand in them but hand is highly related concept to the basketball class that appears constantly in a portion of the images.

- If the segments in a cluster come from a small number of images but still the cluster is large, it could be deduced that the concept has a very distinctive presence in those small portion of discovery images. One example would be the human crowd concept in the basketball class. A small percentage of images have that concept but when its present, because it covers a large area of its corresponding image, it will be partitioned into large number of segments; each of which belonging to the same concept. In our experiments, a cluster with more than 80 segments in it coming from more than 10% of the discovery images is acceptable.

Any concept that is not satisfying one of the aforementioned criteria is removed.

B. More examples of DTCAV

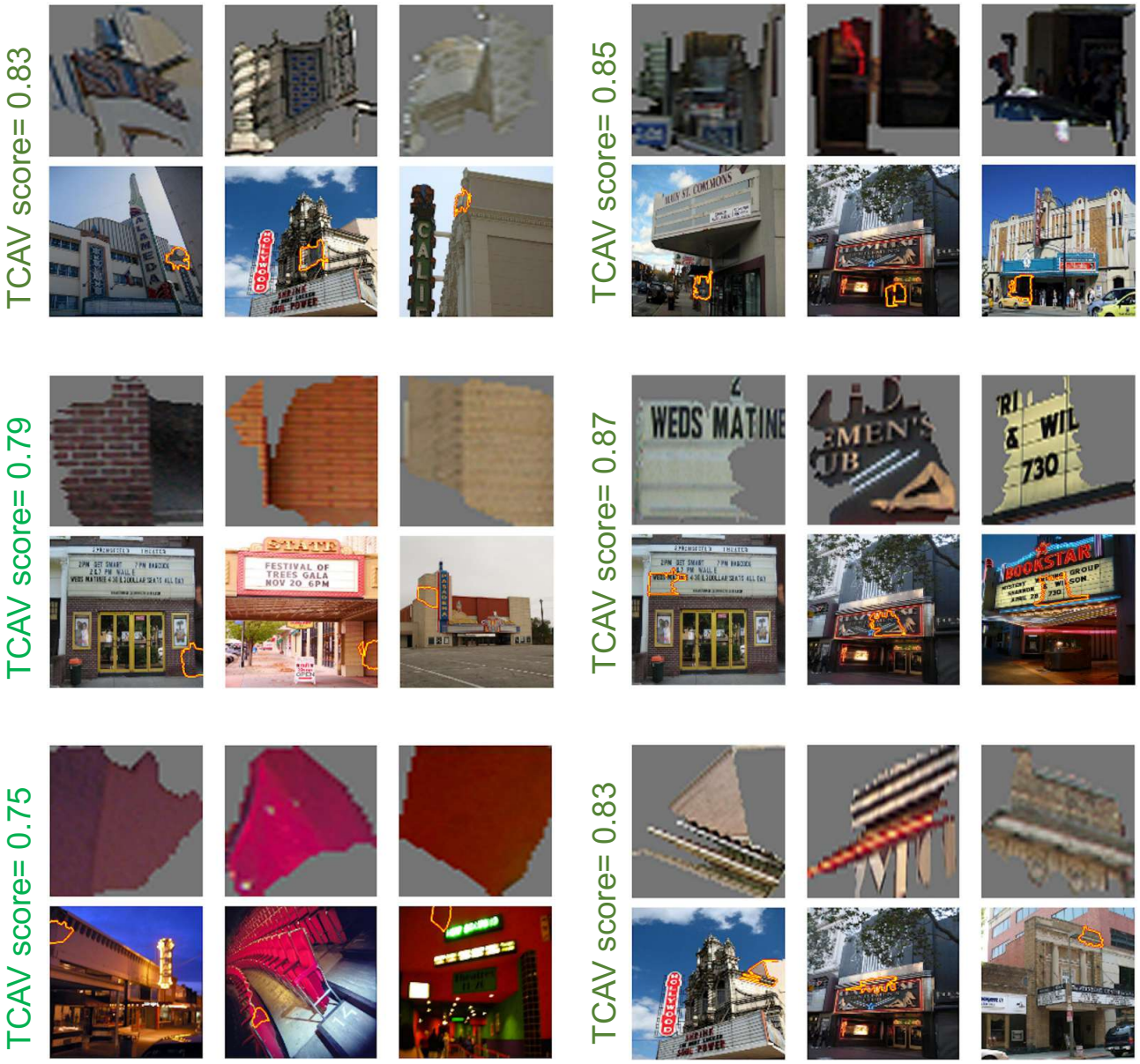

Figure 11: Examples of discovered concepts and their respective TCAV scores for class cinema.

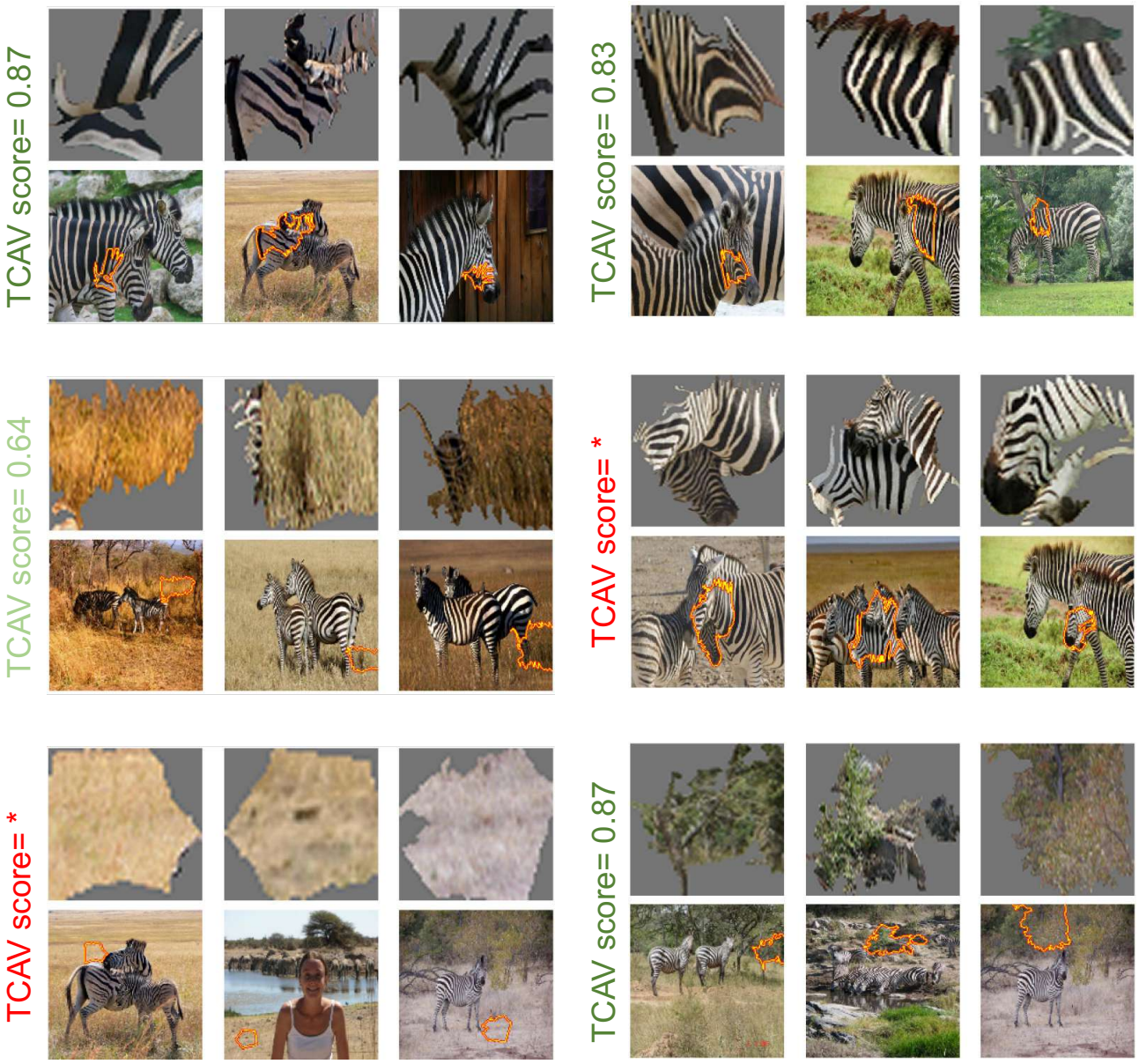

Figure 12: Examples of discovered concepts and their respective TCAV scores for class zebra.

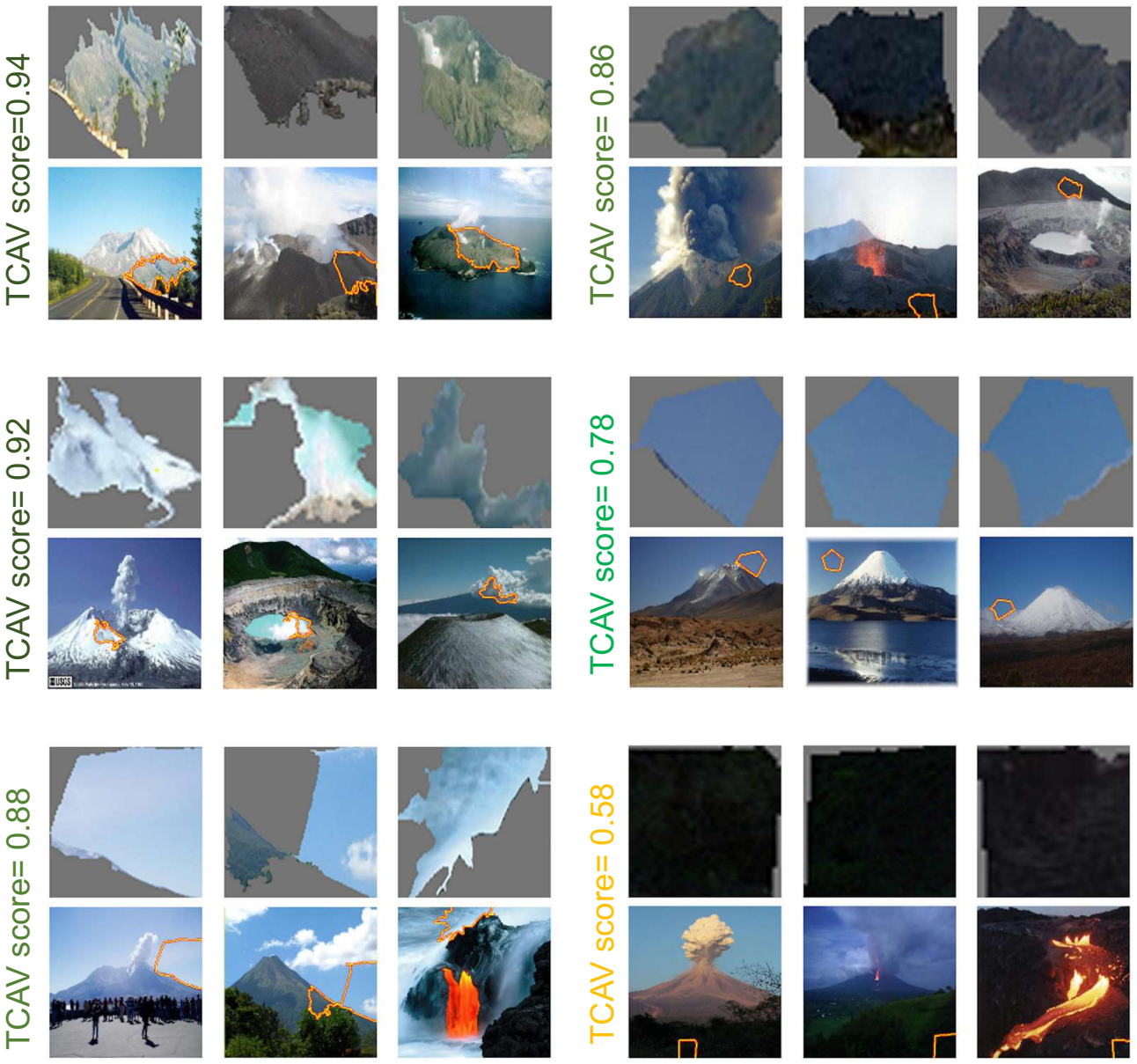

Figure 13: Examples of discovered concepts and their respective TCAV scores for class volcano.

Figure 14: Examples of discovered concepts and their respective TCAV scores for class unicycle.

Figure 15: Examples of discovered concepts and their respective TCAV scores for class hippo.

Figure 16: Examples of discovered concepts and their respective TCAV scores for class hamster.

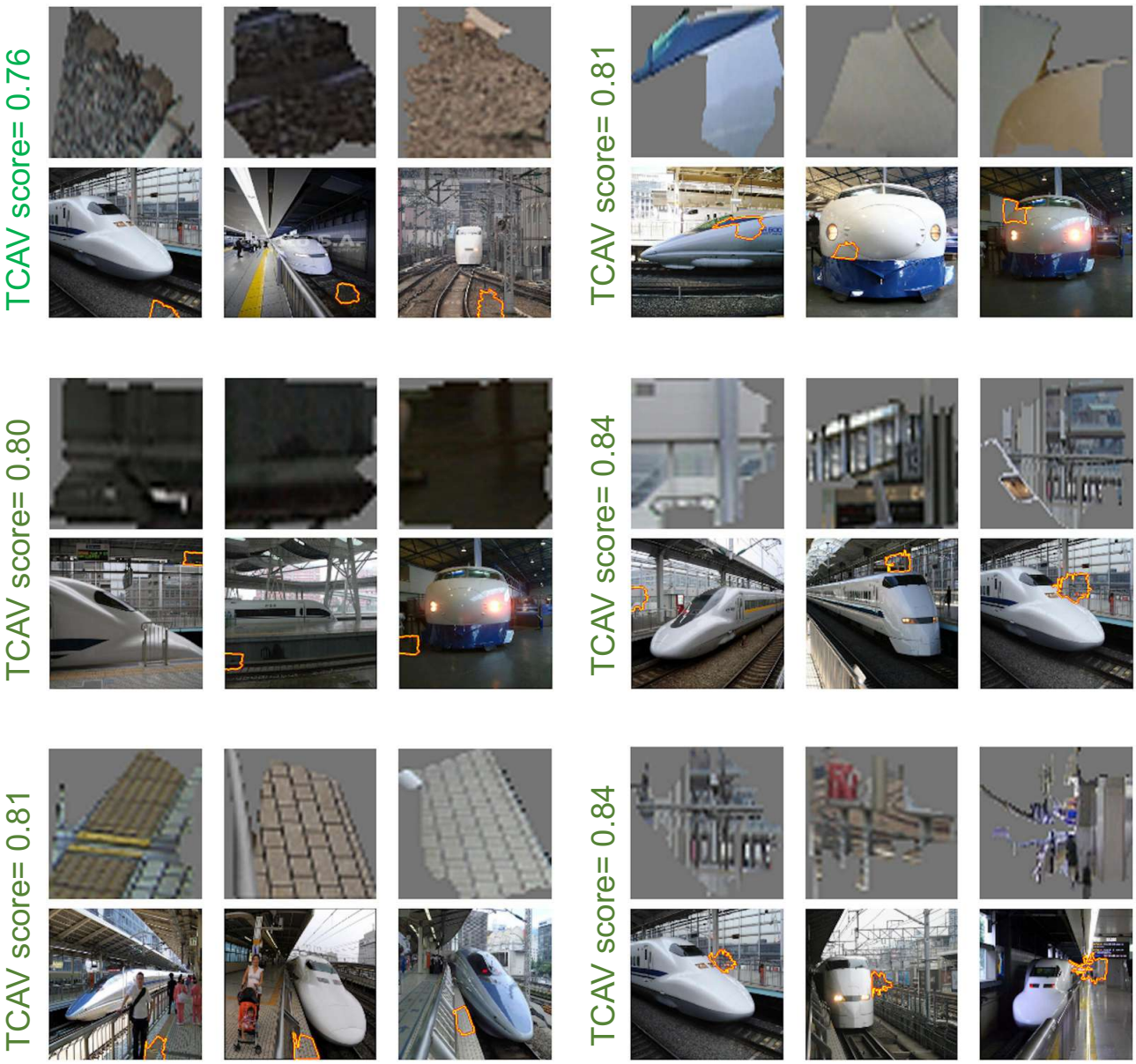

Figure 17: Examples of discovered concepts and their respective TCAV scores for class bullet train.

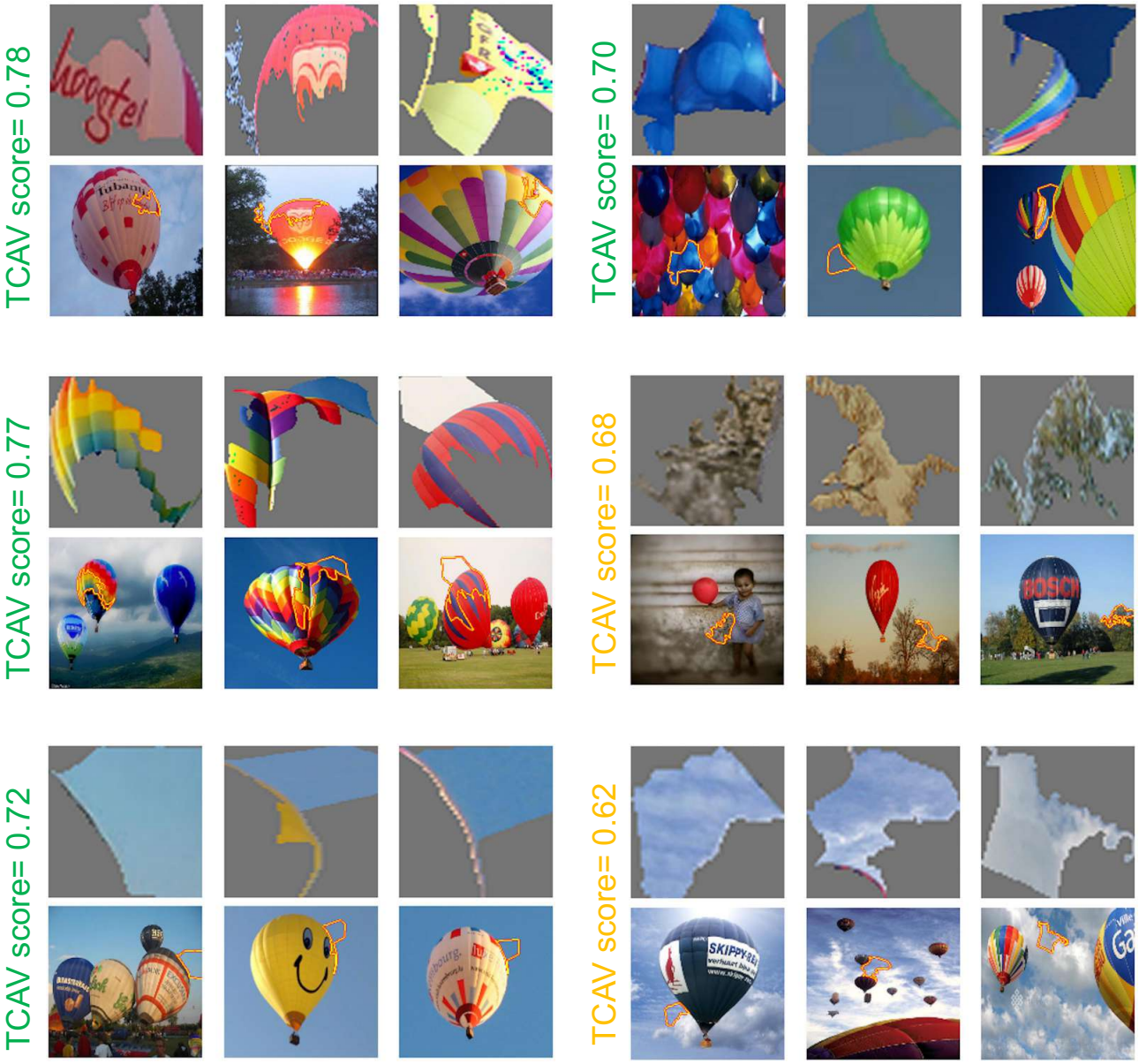

Figure 18: Examples of discovered concepts and their respective TCAV scores for class balloon.

Figure 19: Examples of discovered concepts and their respective TCAV scores for class liner.

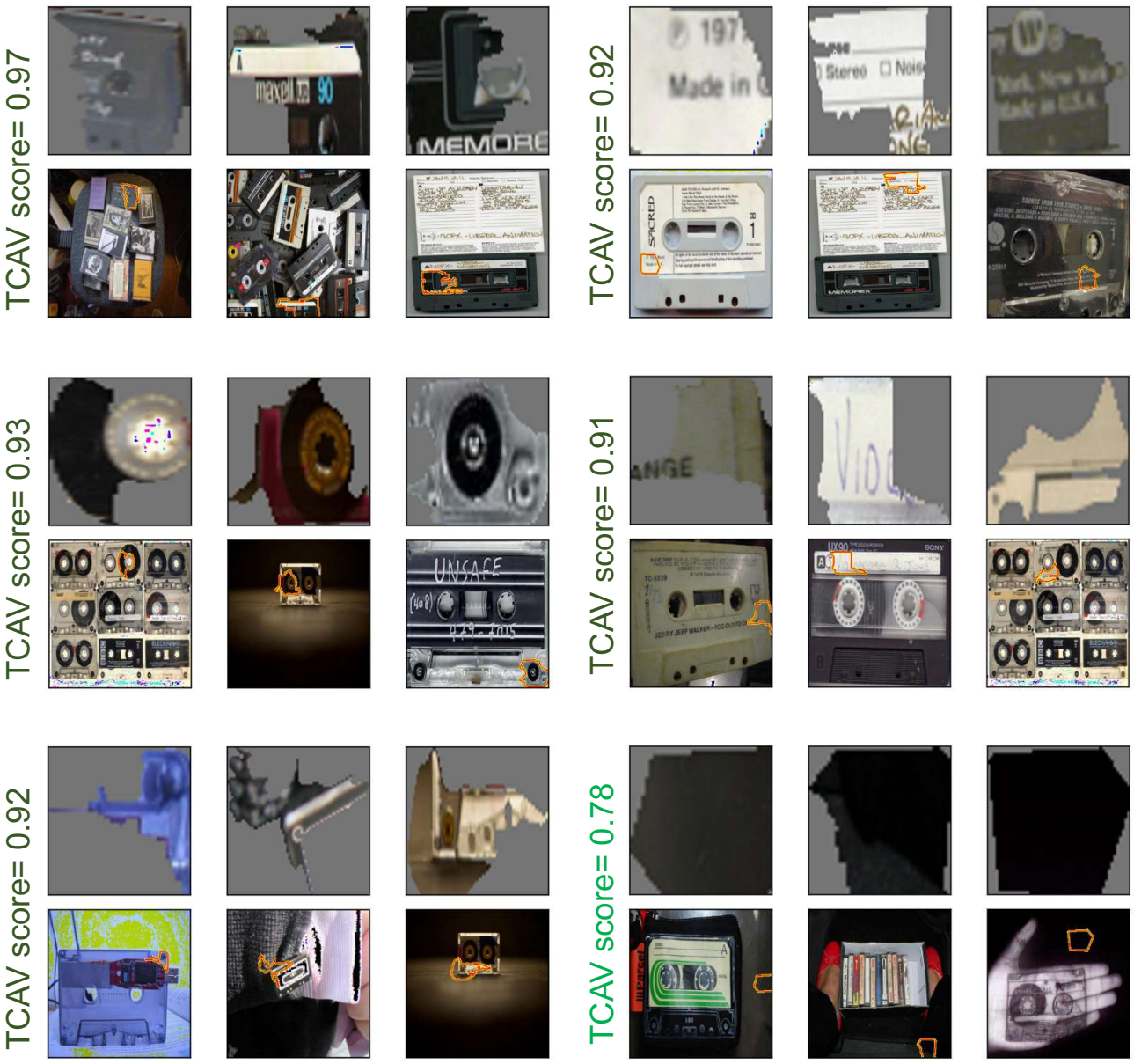

Figure 20: Examples of discovered concepts and their respective TCAV scores for class cassette.

C. More examples of stitched images

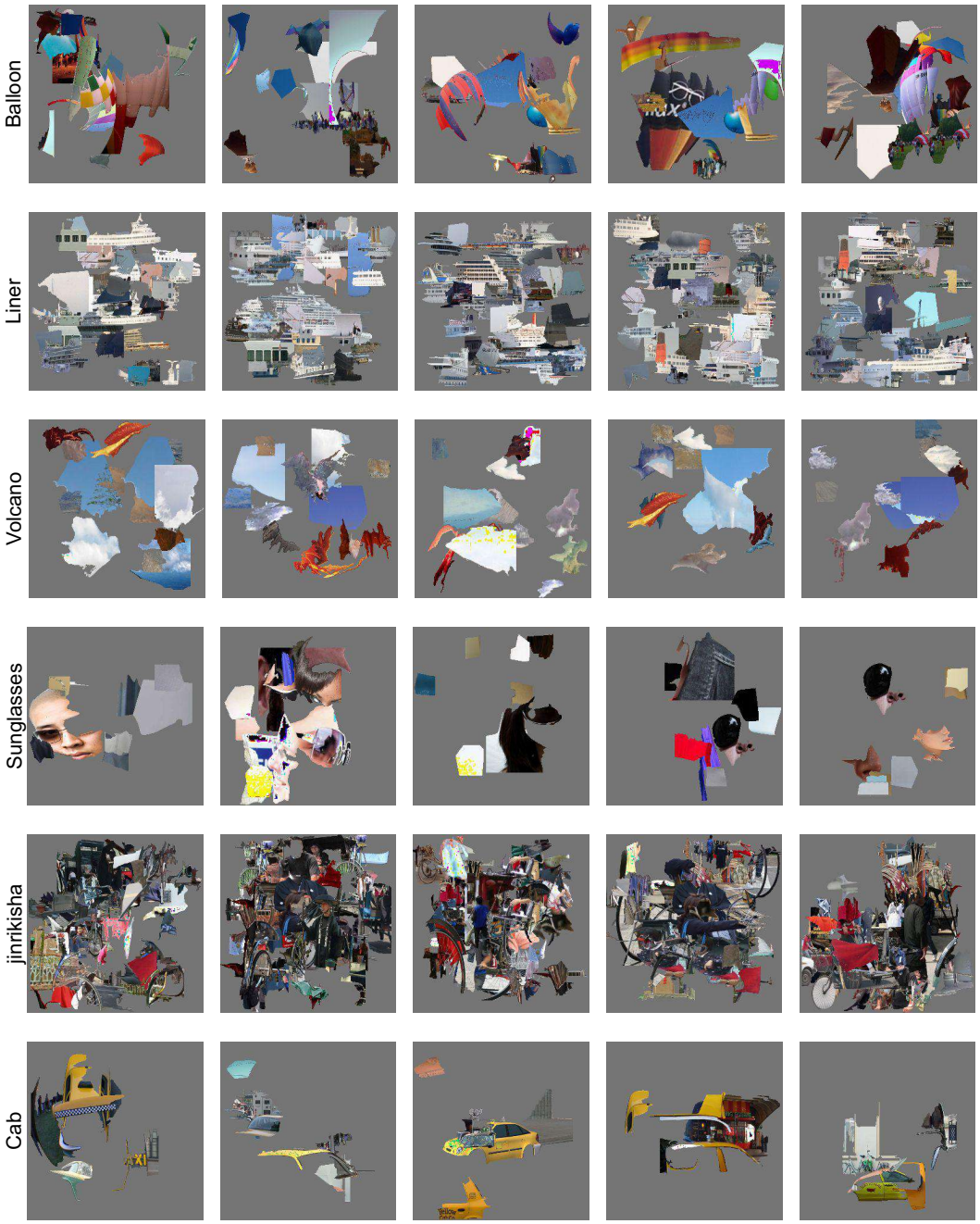

Figure 21: Examples of stitched images classified correctly by the Inception-V3 network.

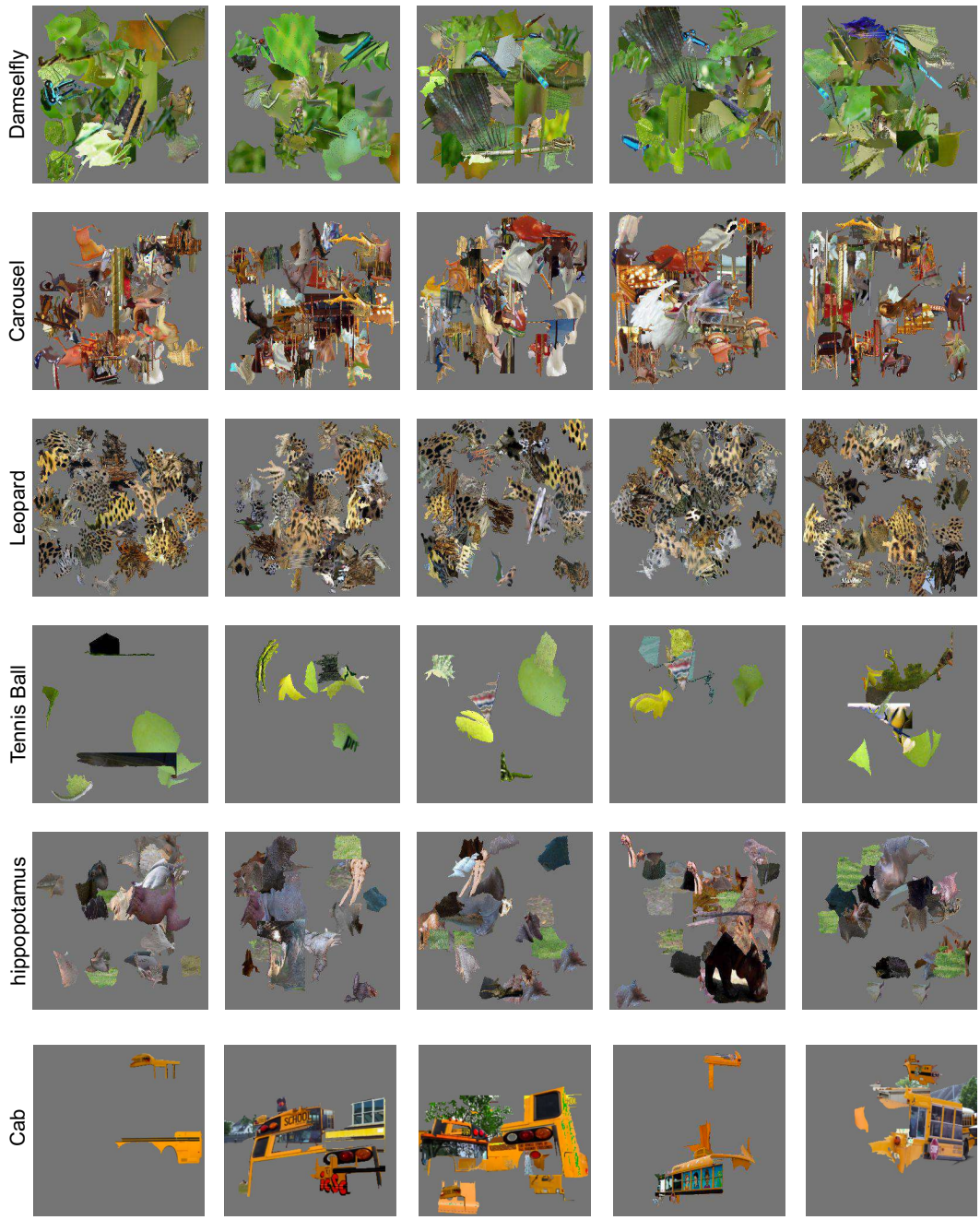

Figure 22: Examples of stitched images classified correctly by the Inception-V3 network.