Towards Human-Understandable Multi-Dimensional Concept Discovery

Supplementary Material

Evaluation Details

In Figure 5, we show examples that were shown to participants in the study. The figure shows an example of ACE (left), MCD (middle), and HU-MCD (right).

Figure 5. Example of outputs shown to participants: left ACE, middle MCD, right HU-MCD. Accuracy of descriptions can be found in Section 4.

Beyond our primary evaluations, we report the number of discovered clusters and their completeness values for the ten CIFAR-10-like classes used in our experiments. The cluster count for each class is derived from the average number of image segments generated by SAM. To ensure concept coherence, we consider clusters with more than 50 segments as concepts, following Ghorbani et al. [15]. Using SAM's segmentation capabilities, we account for class-specific variability, recognizing that different classes exhibit different conceptual structures, resulting in a varying cluster count. We refer to Vielhaben et al. [37] for detailed algorithmic descriptions regarding clustering and experiments that verify these hyperparameter settings.

Class	Number of Clusters	Completeness
Beach Wagon	19	0.73
Hummingbird	12	0.69
Police Van	20	0.79
Ox	15	0.67
Container Ship	13	0.67
Siamese Cat	15	0.75
Zebra	12	0.69
Golden Retriever	13	0.67
Tailed Frog	14	0.68
Airliner	11	0.65

Table 2. Number of clusters and completeness scores of the discovered concepts for the CIFAR-10-like classes used in our experiments.

(a) Original Image (b) Input Masking (original scale) (c) Inpainting (original scale) (d) Inpainting (cropped segment)

Figure 6. Different masking options. When employing Input Masking (option b), only the information within the highlighted region and its immediate surroundings is propagated through the model. In the case of inpainting (options c and d), the baseline color (grey) is propagated through the model.

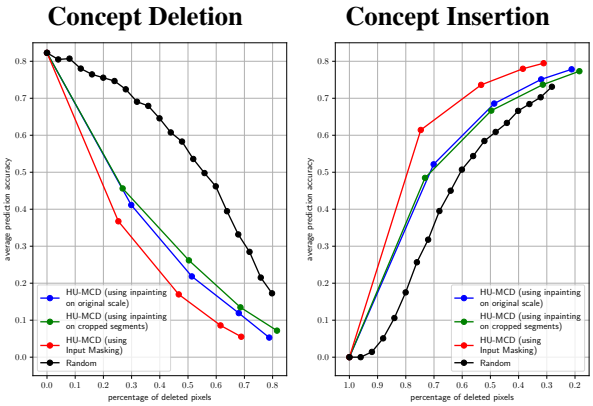

Figure 7. We delete (left) or insert (right) concepts in decreasing order of concept importance and measure the impact on model prediction accuracy, averaged over all validation images of ten *ImageNet1k* classes. Each point represents a discovered concept. Faithful concept importance scores are supposed to result in a sharp decline (left) or ascent (right). We compare the input masking scheme of HU-MCD to the simpler approach of masking with a baseline color, as used in previous methods, either at the original scale of the image or after cropping and resizing the image to fit the segment.

Effectiveness of the Input Masking Scheme

In our experiments, we demonstrate the superior faithfulness of HU-MCD's generated explanations compared to ACE and MCD, which is particularly enforced through the application of the input masking scheme, aimed at ensuring that neither the baseline color used for inpainting the image segment nor the shape of the segmentation mask introduces undesired artifacts that could distort the model's prediction.

We recalculate the C-Deletion and C-Insertion benchmarks for two alternative masking strategies: (1) masking

regions outside segments with a baseline value and (2) cropping segments to minimal bounding boxes and rescaling (as in ACE). Figure 6 shows the three different setups using an example image of class "*airliner*".

The results are illustrated in Figure 7, which shows the effectiveness of the Input Masking scheme in ensuring that the concept importance scores *faithfully* reflect the model's reasoning process. Interestingly, using inpainting at both the original scale and on the cropped segments yields similar performance, which is surprising, particularly considering that masks covering only small regions propagate much information regarding the baseline color through the network. Although CNNs are trained to be robust against scale variations, this result shows that scale information, as well as aspect ratio, plays a significant role, and distorting them negatively impacts the generated explanation quality.