Discover-then-Name Revisited: Enhancing Concept Bottle-Neck Models Interpretability

Anonymous authors
Paper under double-blind review

Abstract

This study aims to reproduce and extend the research on Discover-Then-Name Concept Bottleneck Models (DN-CBM) introduced by Rao et al. (2024). DN-CBM enhances traditional CBM models by incorporating sparse autoencoders (SAEs) to enable automatic concept discovery and improved concept generation and interpretability. We replicate the key experiments on CIFAR-10, CIFAR-100, Places365, and ImageNet, confirming the claims of automated concept discovery, task-agnostic applicability, and improved vocabulary leading to greater granularity. However, we find that the claim of superior interpretability over CLIP is inconclusive. Beyond replication, we introduce new experiments, including an analysis of color perturbations on concept robustness and the integration of Local Interpretable Model-Agnostic Explanations (LIME) to trace which features correspond to each concept. Our findings reveal the model's limited robustness to color variations and demonstrate how adding LIME results in increased interpretability and the ability to detect (spurious) correlations. The complete implementation of the original authors experiments as well as ours is available in our repository: `https://github.com/EKarasevnl/Reproducibility-DN-CBM`.

1 Introduction

Recent advances in deep learning have led to increased attention to the importance of explainability and interpretability, with mechanistic interpretability aiming to reveal how neural networks structurally process and represent information. The Concept-Bottleneck Model (CBM) promotes this by leveraging a set of human-interpretable concepts (e.g., "round," "yellow") to guide predictions; however, these concepts may not always be straightforward to detect or interpret. To address this, Rao et al. (2024) introduced Discover-Then-Name (DN-CBM), which uncovers the model's naturally learned concepts and assigns meaningful labels by aligning them with CLIP embeddings from a predefined vocabulary, thereby enabling a linear probe to classify images based on extracted concepts.

In this work, we reproduce the author's experiments across CIFAR-10, CIFAR-100, Places365, and ImageNet to verify their claims. Additionally, with the goal of both deepening the understanding of the method and improving it. We conduct further experiments such as investigating how modifying Places365's color map affects the SAE's color understanding and enhancing the interpretability of concept assignments using LIME. Section 2 outlines the scope of our study, while Section 3 details the proposed methodology and experimental setup. Section 4.1 presents our replicated qualitative and quantitative results, followed by additional experiments and a survey study in Section 4.2. The reproduction process is examined in Section 5, with concluding remarks presented in Section 6.

2 Scope of reproducibility

This paper's scope of reproducibility focuses on the author's framework for reproducing the DN-CBM results. The main claims presented by the author are as follows:

1. **Automated Concept Discovery.** The paper introduces a method for automatically discovering semantically meaningful concepts from CLIP features using Sparse Autoencoders (SAEs), where semantically related concepts cluster together in the latent concept space.
2. **Enhanced interpretability.** The use of SAEs leads to interpretable models by extracting and naming concepts that align well with human-understandable features, enhancing the model's transparency.
3. **Task-Agnostic Approach.** The feature extractor, the datasets used for concept discovery and downstream tasks, and the vocabulary used for naming concepts can be freely chosen.
4. **Impact of Vocabulary on Concept Name Granularity** The choice of vocabulary affects the granularity of concept name assignment. Expanding the vocabulary with more specific terms enhances the granularity of concept naming while removing terms leads to less descriptive concept names.

In this work, we reproduce the experiments on all the provided downstream datasets to validate and verify the author's claims.

3 Methodology

3.1 CBM Construction

The CBM pipeline proposed by Rao et al. (2024) consists of 3 parts: automated concept discovery via sparse autoencoders (SAEs), automated concept naming by aligning the CLIP embeddings of predefined vocabulary with concept embeddings, and training of the concept bottleneck model.

3.1.1 Concept Extraction

First, the images are converted to the CLIP embedding space. The concepts are then discovered using the SAE, which represents CLIP features in a high-dimensional space but with sparse activations to obtain interpretable concept representations (Bricken et al., 2023) (Cunningham et al., 2023). The SAE consists of a linear encoder $f(\cdot)$ with weights $W_E \in \mathbb{R}^{d \times h}$, a ReLU activation function ϕ, and a linear decoder $g(\cdot)$ with weights $W_D \in \mathbb{R}^{h \times d}$. For a given input a, the SAE computes:

$$SAE(a) = (g \circ \phi \circ f)(a) = W_D^T \phi(W_E^T a) \tag{1}$$

The hidden representation $f(a)$ is designed to have significantly higher dimensionality ($h \gg d$) compared to the CLIP embedding space, while being optimized to activate very sparsely. Specifically, the SAE is trained with a reconstruction loss L_2, as well as a sparsity regularization L_1 on CC3M dataset (Sharma et al., 2018):

$$\mathcal{L}_{SAE}(a) = \|SAE(a) - a\|_2^2 + \lambda_1 \|\phi(f(a))\|_1 \tag{2}$$

3.1.2 Concept Naming

Once the SAE is trained, the latent representation produced by the encoder corresponds to the concept space where each concept can be named to extract human-understandable interpretations of the concepts.

To align concept representations with a set of words, the authors employ the 20k most commonly used English words from the CLIP-Dissect vocabulary (Oikarinen and Weng, 2023). The SAE decoder is designed to map a concept representation back to the CLIP embedding space using weights $W_D \in \mathbb{R}^{h \times d}$. In this context, a vector p_c represents the concept c in the latent space:

$$\mathbf{p}_c = [\mathbf{W}_D]_c \in \mathbb{R}^d \tag{3}$$

The SAE is trained to transform CLIP embeddings into concept representations through a reconstruction objective. This means that the SAE decoder can reconstruct the original CLIP embeddings from the concept representations produced by the encoder. Consequently, a specific concept representation p_c can be decoded

back into a CLIP embedding, which can then be compared to the CLIP embeddings of the CLIP-Dissect vocabulary. By calculating the cosine similarity between the decoded embedding and the embeddings of the vocabulary words, we can associate the concept p_c with the word v_c from the vocabulary that has the highest similarity:

$$v_c = \arg\min_{v \in \mathcal{V}} \cos\left(\mathbf{p}_c, \mathcal{T}(v)\right). \tag{4}$$

3.1.3 Concept Bottleneck Model

To perform the classification, the authors use a linear model $h(\cdot)$ designed to operate on SAE concept activations. The DN-CBM predicts the class of an input image x_i as follows:

$$t(x_i) = \underbrace{h}_{\text{Probe}} \circ \underbrace{\phi \circ f}_{\text{SAE}} \circ \underbrace{\mathcal{I}}_{\text{CLIP}}(x_i). \tag{5}$$

The linear probe is trained using cross-entropy (CE) loss along with L1 loss to increase interpretability (Bricken et al., 2023).

$$\mathcal{L}_{\text{probe}}(x_i) = \text{CE}\left(t(x_i), y_i\right) + \lambda_2 \|\omega\|_1 \tag{6}$$

where λ_2 is L_1 sparsity coefficient, y_i is the ground truth label, and $\mathcal{I}(\cdot)$ is the CLIP model.

3.2 Datasets

The authors utilize the CC3M dataset to train the SAE model, while other datasets are used to train a linear probe for downstream tasks.

- **CC3M** is a dataset comprising approximately 3.3 million image-caption pairs, primarily designed for training models for image-text alignment tasks (Sharma et al., 2018).
- **CIFAR-10** is a dataset of 60,000 32x32 images distributed over 10 classes, with 6,000 images per class (Krizhevsky et al., a).
- **CIFAR-100** is a dataset similar to CIFAR-10, but features 100 classes, each containing 600 images, offering a finer-grained classification challenge (Krizhevsky et al., b).
- **ImageNet** is a large-scale dataset with more than 1 million labeled images spanning 1,000 categories, widely used to benchmark image classification models (Deng et al., 2009).
- **Places365** is a scene recognition dataset comprising 365 categories. For this study, we utilize the *Places365-Standard* version, which contains 1.8 million images (Zhou et al., 2017).

3.3 Experimental setup and code

We used the provided GitHub repository to replicate the authors' experiments, implementing minor bug fixes and adding a missing script for computing normalized CLIP embeddings for the CLIP-Dissect vocabulary (Oikarinen and Weng, 2023). We also modified the code to enable image sampling based on cosine similarity with vocabulary embeddings, providing a baseline for assessing the SAE's interpretability.

To verify that automated concept discovery produces semantically meaningful concepts, we apply K-Means clustering over concept activations on Places365 (Section 4.1.1). To assess classifier interpretability, we analyze DN-CBM's predictions by identifying top contributing concepts per prediction and conducting a survey comparing CLIP's top-activating images to DN-CBM concepts.

To assess the interpretability of the classifier, we examine DN-CBM's predictions by analyzing the top contributing concepts for each prediction and evaluating whether those concepts are relevant to the corresponding image. Additionally, we conduct a survey to compare the top-activating images for each vocabulary token in CLIP to those for the corresponding DN-CBM concepts, illustrating how the discovered concepts align with or differ from CLIP's learned representations.

Color can be both beneficial and detrimental to model performance. It aids in distinguishing elements like flowers or fruits but can also lead to over-reliance on irrelevant features. For example, color is unnecessary for differentiating between a car and a plane. Ideally, a model should also be invariant to intraclass color

changes, a car can be both red and blue. To assess whether generated concepts capture color we performed stratified sampling for 7300 Places365 validation images and creating three variations: original, grayscale, and inverted color map. Extracting the top 20 concepts, we scanned for basic colors, including "noir" to assess overall color scheme understanding (Figure 11). To evaluate color-invariance, we analyzed the top 10 concepts across variations, categorizing them as color-invariant if more than five overlapped and color-dependent if more than five were unique for the original image, while also examining true labels and predicted classes for qualitative assessment.

Finally, we also extend the original work by integrating LIME (Local Interpretable Model-Agnostic Explanations). This is doen to better understand where the generated concepts come from, which is especially useful for concepts that lack a clear relationship to the image. To achieve this we modify LIME to measure concept contributions rather than label confidence. We use the Python library `lime` (**?**) and configure the `LimeImageExplainer` with `top_labels=5`, `num_samples=5000`, and `batch_size=1`. For the `get_image_and_mask` function, which generates masks, we set `positive_only=True`, `num_features=3`, `hide_rest=False`, and `min_weight=0.0001`. All other hyperparameters remain at their default values.

3.4 Computational requirements

We perform all experiments on an NVIDIA A100 GPU. We replicate the experimental setup provided by the author, which utilizes Python 3.10. Carbon emissions are included and calculated using the Machine Learning Impact calculator (Lacoste et al., 2019). We conducted our experiments using a private infrastructure with a carbon efficiency of 0.555 kgCO2eq/kWh (Moro and Lonza, 2018). We summarize the total compute and emissions in Table 1. We trained an SAE for 200 epochs using CC3M (Sharma et al., 2018). Subsequently, for each of the five probe datasets, we conducted zero-shot classification and trained linear probes on both CLIP features and the concepts generated by the trained SAE. Furthermore, we performed a data augmentation experiment by adjusting color schemes on the validation set of the Places365 dataset (Zhou et al., 2017). Finally, we assessed how various image features contribute to activations in top concepts during inference using LIME on approximately 100 images.

	Reproducing Original Experiments					**Experiments Beyond Original Paper**		**Total**
	CC3M	CIFAR-10	CIFAR-100	Places365	ImageNet	Data Augmentations	LIME	
Total compute time (h)	1.5	0.14	0.35	4.6	9.24	0.67	4	**20.5**
kgCOeq Emissions	0.21	0.02	0.05	0.64	1.28	0.09	0.56	**2.85**

Table 1: Total compute time and carbon emissions for different experiments.

4 Results

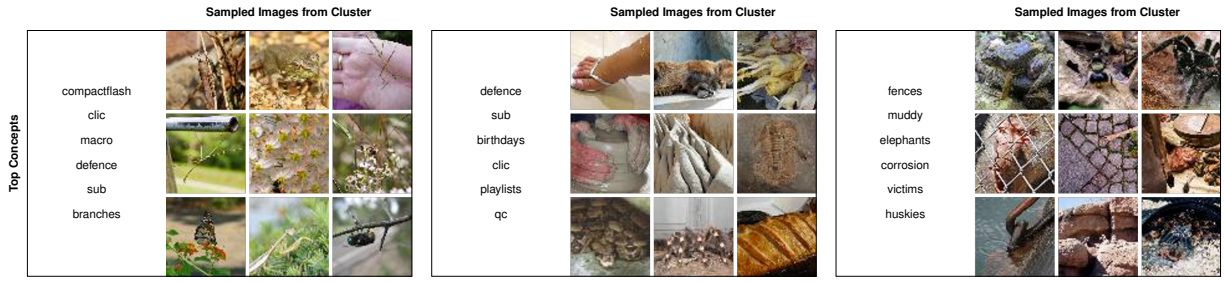
Our reproducibility study confirms the accuracy of the final two claims outlined in Section 2. The assertions regarding automated concept discovery and enhanced interpretability are generally valid; however, there are instances where they do not hold. In this section, we present supporting results and discuss situations where certain claims fail.

4.1 Results reproducing original paper

4.1.1 Automated concept discovery

Similarly to the authors, we demonstrate that the resulting cluster concepts are semantically meaningful, as illustrated in Figure 1a. The first set of images represents agriculture and crop production related concepts. The second cluster of features is about water-related activities. The final set depicts abandoned or demolished locations.

(a) Concept clusters for Places365.



(b) Concept clusters with non-descriptive names.

Figure 1: Comparison of concept clusters for Places365.

However, on numerous occasions, when clusters of images that could have been accurately described using appropriate words from the vocabulary were instead labeled with nonsensical in a given context words, as shown in Figure 1b. This problem frequently occurred with words assigned to multiple concepts. For instance, in the SAE we trained, 28 concepts were labeled with the word "clic", and 11 concepts were labeled with the word "sub."

Overall, despite the described issues, most identified clusters are semantically consistent, demonstrating that the concepts discovered by SAE are effective for representation learning and human interpretable.

4.1.2 Enhanced interpretability of SAE generated concepts

The results for individual concept contributions show that classifier predictions are indeed human-interpretable. As illustrated in Figure 2, the top contributing concepts strongly align with both the actual content of the image and the predicted class.

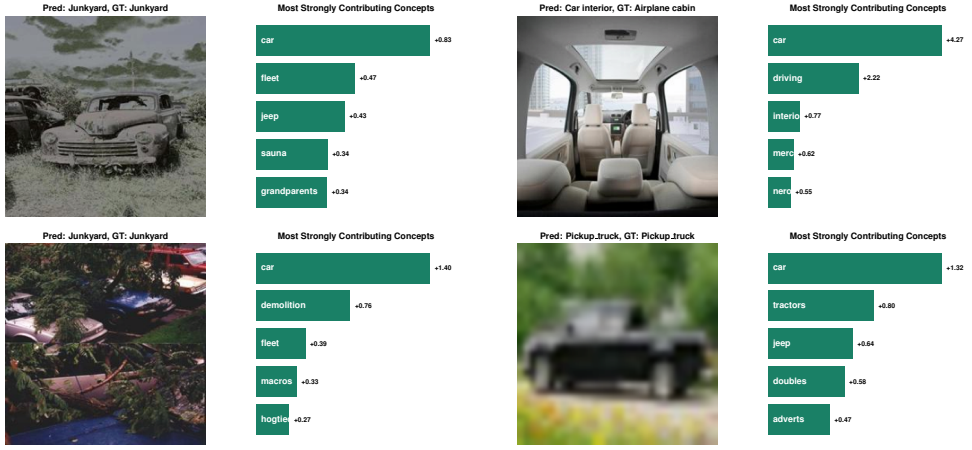


Figure 2: Local explanations for images related to cars.

Although most of the examples include the main concepts that contribute to the content of the images, there are examples when meaningless concepts appear as described in Appendix A.2. In general, most examples where the content of the image can be described using words from the vocabulary are well-represented, while cases where relevant concepts are present but the image is poorly described are relatively rare. This indicates that DN-CBM is not entirely inherently interpretable, as it occasionally fails to utilize relevant concepts when necessary. However, for most samples, the main contributing concepts accurately reflect the image content. To further evaluate whether the contributing concepts align with the objects in the image, we analyze the contributions of the top concepts using LIME in Section 4.2.2.

4.1.3 Task-Agnostic approach

Model	**ImageNet**	**Places365**	**CIFAR-10**	**CIFAR-100**
Linear Probe	71.0	52.0	88.0	69.0
Reported	73.3	53.4	88.7	70.3
Zero-Shot	59.8	34.2	71.5	41.9
Reported	59.6	38.7	75.6	41.6
DN-CBM	71.0	50.0	89.0	69.0
Reported	72.9	53.5	87.6	67.5

Table 2: Top-1 Validation Accuracy(%) comparison of models across downstream datasets.

Rao et al. (2024) emphasise the agnostic approach of the DN-CBM framework. Essentially, the concept bottleneck layer is agnostic to the downstream dataset used for classification. The concepts discovered by the SAE on a given dataset, such as CC3M (Ng et al., 2021), allow for seamless application across various datasets for classification tasks. We reproduce the classification experiments on the datasets mentioned in Section 3.2 to verify this claim. As baselines, Rao et al. (2024) use a linear probe trained on the corresponding dataset's CLIP-RN50 (ResNet-50) image embeddings (Linear Probe) and CLIP-RN50's zero-shot classification performance (Zero-Shot).

Table 2 reports the top-1 validation accuracies for the baselines and DN-CBM. The linear probe trained on CLIP embeddings are very similar to the reported scores. Similarly, the zero-shot performance of CLIP-RN50 broadly aligns with the reported baselines, aside from a minor (4%) shortfall on Places365 likely due to the absence of the exact CLIP prompt templates utilized by the author for this dataset. Our reproduced DN-CBM accuracies also closely match the reference values on ImageNet, Places365, CIFAR-10, and CIFAR-100, confirming the robustness of CBM approaches.

4.1.4 Impact of vocabulary on concept name granularity

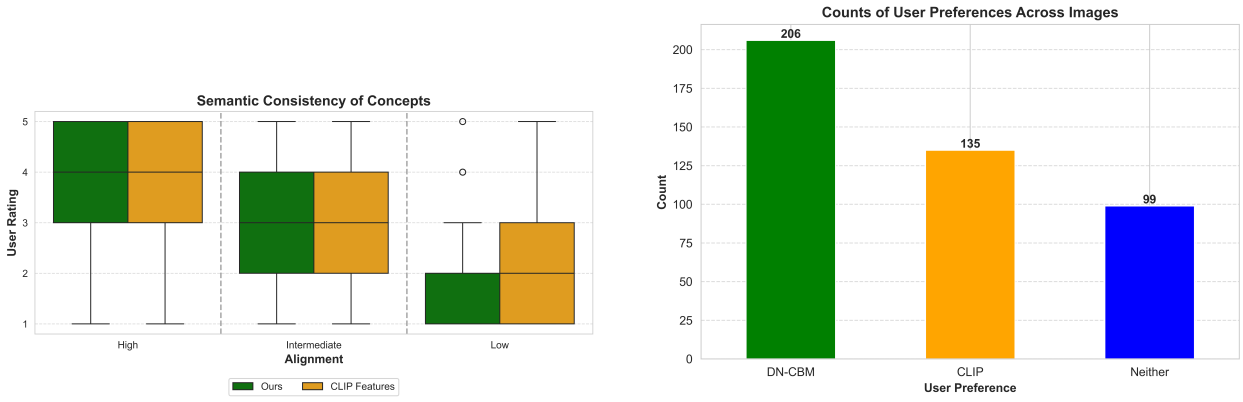
The original authors argue that the granularity and size of the vocabulary play a crucial role in the accuracy of obtained concept names. They further suggest that vocabulary design can be leveraged to control the level of granularity in assigned names, depending on the specific use cases.

Figure 3: Impact of vocabulary on granularity of concept names

Our experimental findings confirm the validity of the author's claim that refining vocabulary enhances concept assignment accuracy. As illustrated in Fig. 3, our analysis of the concepts "tree" and "driving" demonstrates this effect. For instance, in the case of "driving," images depicting car interiors were more accurately labelled after adding the term "car interior" to the vocabulary. Conversely, removing "driving" reduced accuracy, leading to inaccurate assignments such as "passenger." Therefore, these results validate the authors' claim regarding that a refined vocabulary improves the precision of the concept assignments.

4.1.5 Qualitative evaluation through a user-study

We conducted a survey with 20 randomly selected concepts, categorized by alignment with text embeddings. Participants ranked two image grids—one from CLIP and one from DN-CBM—on semantic consistency and name accuracy. Our internally published survey received 22 responses.



(a) Survey Results on Semantic Consistency

(b) Survey Results on User Preference

Figure 4: Comparison of survey results on semantic consistency and user preference.

Additionally, we introduced a preference question for each concept, asking users whether they preferred the concepts from Image 1 (CLIP), Image 2 (DN-CBM), or neither. While Rao et al. (2024) claim DN-CBM clusters are more semantically consistent and accurate, our results show minimal differences, with CLIP scoring slightly higher in some cases. However, user preferences aligned with the original study—despite concept names sometimes lacking semantic meaning, participants favored DN-CBM-generated images more often.

4.2 Results beyond original paper

4.2.1 Effect of color augmentation on concept representations

As mentioned previously, understanding intra-class color variations is crucial for a classifier. If the concepts drastically change when color changes it would indicate that the DN-CBM methodology is not robust. In this experiment, we assess how the concepts change when the colors of the images change. We also record how the classification performance changes and look at which classes are most impacted by the change in color.

The first part evaluates color recognition through generated concepts. Observations suggest the SAE effectively captures different colors. As shown in Figure 11, the original image associates "yellow" with a bus, the grayscale version lacks color-related concepts except for "noir", and the inverted image correctly identifies "blue", matching the bus color. This hypothesis is further supported by the distribution of color-related concepts across different image variations (Figure 12). Results show that original images utilize a balanced color distribution, while grayscale images predominantly use "noir" and inverted images emphasize bright colors like purple and turquoise. This suggests the SAE accurately associates colors with images. Notably, color concepts appear more critical in augmented images than in original ones, likely because the unusual color schemes force the model to focus on color for accurate representation.

In the second part, we measured the change in the classifier performance depending on different color schemes. Results in Table 4 indicate a performance drop when colors are altered, with the inverted dataset showing the largest decline (18.7% accuracy). Grayscale images also reduce accuracy, though less severely. This suggests the model relies on color-based concepts, with grayscale images simply removing color while inverted images introduce unnatural noise, disrupting the model's ability to interpret content effectively. Additional information on individual class performance can be seen in Table 4. We observe the largest accuracy drop for classes "swimming pool outdoor" (65% drop), "Garage outdoor" (65% drop), and "Legislative chamber" (65% drop). On the other hand, certain classes, such as "Cockpit" and "Boxing ring", maintain their accuracy despite the transformation. These results highlight how certain classes rely heavily on color (e.g. swimming pools often being blue), while others rely more on shape and texture-based features. Relying on color as a feature is not a bad thing on its own however for certain classes such as "Garage outdoor" it is likely to result in a far less robust classification during inference.

Transformation	**Original**	**Grayscale**	**Inverted**
Accuracy (%)	49.9	40.1	18.7

Table 3: Top-1 Validation Accuracy(%) comparison across transformations.

Finally, we explored color invariance in the generated concepts, focusing on two key research questions: (1) Which concepts remain consistent across different colors, and how frequently do they occur? (2) Which concepts are inherently dependent on color?

In our analysis, we found that instances of color independence were observed a total of 366 times, accounting for approximately 5% of all images. A common characteristic of these images is their focus on a primary subject with minimal competing elements. Example images are shown in Fig. 5. For color independence to occur, it is crucial that color variations neither alter the overall meaning of the image nor introduce significant ambiguity. This creates an ideal scenario in which concepts remain consistent across all sub-images. However, it is important to note that shared concepts do not always lead to correct predictions. For instance, in the attached car image, the true label was "auto showroom". Although all variations predicted a car-related class, only the original image produced the correct label.

The exact opposite can be said for color-dependent images. We have observed them slightly more often, with a total of 697 times, making up almost 10% of all sample data. This suggests that most images are in between significant color dependence and invariance. What characterized those images was that with the

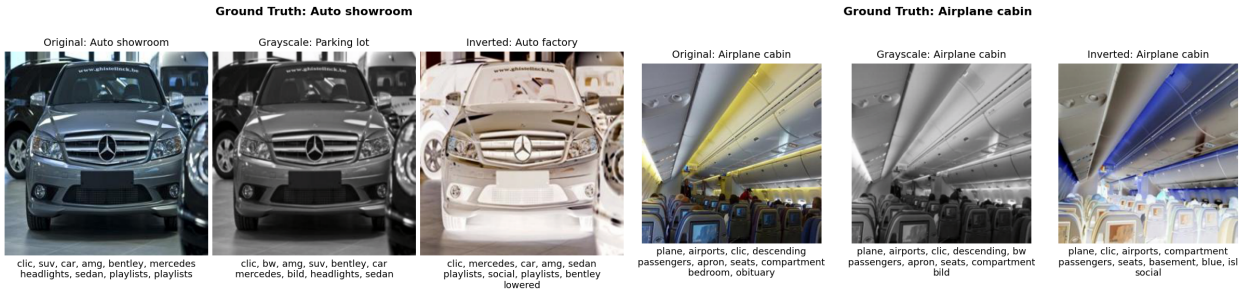


Figure 5: Example of two Color-Invariant Images with Top 10 concepts under images

color change, the image meaning was not entirely clear. Even a change to grayscale can introduce ambiguity, making it difficult to clearly infer what the initial image was referring to. An example can be seen in Fig 6.

Figure 6: Example of Color Dependent Images with Top 10 concepts listed under images

In the provided "Raft" example, the original image shares the concept of "rapids" with the grayscale image but otherwise tends to produce more activity-focused concepts, such as "kayaking" or "backpacking". In contrast, the grayscale image leans toward more serious themes, generating concepts like "military", "victims", and "airsoft". Finally, the inverted color map perturbs the image so that the resulting concepts appear entirely unrelated. Interestingly, even though the concepts are significantly different between the original and grayscale images the predicted output is the same for both. On the other hand, the case of "Bazaar indoor" is only detected for the original image as the vibrant colors make it far more possible to correctly interpret.

In general, these findings highlight that while the SAE method demonstrates an understanding of colors, it is not entirely color-independent, which was further supported by the perturbed validation accuracy. Certain images require color information to ensure successful recognition and accurate concept generation. This dependency is particularly apparent in cases where the alteration of color introduces ambiguity, leading to shifts in the perceived meaning of the image. However, the ability of the SAE to maintain consistent concepts across a subset of images showcases its potential robustness to color given a clear enough leading motive in the image.

4.2.2 Enhancing classification decision interpretability with LIME

As shown in Figure 10, the generated concepts often lack a clear relationship to the image. To better interpret these concepts, we apply LIME, a tool used to identify the most influential features in a model's prediction. For image classification, LIME: 1) Segments the image into hyperpixels, 2) Perturbs segments and observes changes in label confidence, and 3) Identifies the hyperpixels that most influence the model's decision.

The yellow-highlighted areas indicate the regions most influencing each concept. For instance, we see the concept "quarterback" corresponds to the player in the back as expected. We also see that the concept "arch"

Figure 7: LIME explanation for high-activating concepts from the Places365 dataset.

corresponds to the top of the rock which is also intuitive. Interestingly in the third image it is not clear at first glance where the concept "softball" is coming from, however when we look at the LIME output we see that it is focused on the bent over man resembling the catcher in softball. This highlights the usefulness of LIME in finding spurious correlations.

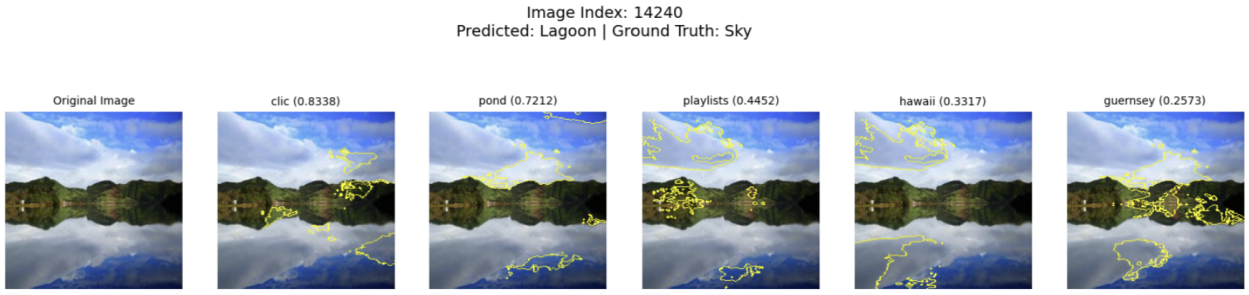


Figure 8: Explaining obscure top concepts using LIME.

We now revisit the examples from Figure 10 to analyze obscure top concepts, focusing specifically on the third image. Using LIME, we observe that the concept "pond" is associated with both the blue water and parts of the sky. The concept "playlists" appears to stem from a combination of hills, sky, and water, possibly reflecting common elements in playlist cover images. Interestingly, "hawaii" is entirely focused on the clouds, while "guernsey" primarily highlights the land. These insights illustrate how LIME helps reveal the specific image features influencing concept formation.

5 Discussion

In this study, we conducted a series of experiments to validate the main claims of the original paper. We find that the claims are mostly true apart from few exceptions. Specifically, the trained SAE effectively generates a human-interpretable latent concept space. Meaningful names are assigned to most concepts, enabling us to interpret the linear probe classifier's output based on concept activations, however, certain images contained uninformative explanations such as the ones detailed in Appendix A.2. Results across tested probe datasets confirm the task-agnostic nature of the DN-CBM approach. Additionally, we found that a more granular vocabulary improves the accuracy of concept naming. Lastly we add to the original work by experimenting with modifying the colors in the input images and observing the change in concepts. We were able to infer that the model understands color, however, that for certain classes it still sometimes relies on it too heavily

resulting in poor robustness. By incorporating LIME we were able to observe exactly where each concept comes from, enhancing understandability and giving us a tool for finding spurious correlations.

5.1 What was easy

The original paper was well-written and included numerous examples for clarity. Additionally, the original code was made available on GitHub, including the instructions for running the training and visualization scripts. The documentation within the codebase was also sufficient.

5.2 What was difficult

The main difficulty involved completing the code necessary for certain experiments, particularly implementing a module to sample images based on concepts derived from CLIP features. The original paper lacked details on the authors' implementation of this module, which may lead to differences in how the survey was conducted. Luckily after contacting the original authors to ask for clarification we received clarification.

References

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research).

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research).

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Moro, A. and Lonza, L. (2018). Electricity carbon intensity in european member states: Impacts on ghg emissions of electric vehicles. *Transportation Research Part D: Transport and Environment*, 64:5–14. The contribution of electric vehicles to environmental challenges in transport. WCTRS conference in summer.

Ng, E. G., Pang, B., Sharma, P., and Soricut, R. (2021). Understanding guided image captioning performance across domains.

Oikarinen, T. and Weng, T.-W. (2023). Clip-dissect: Automatic description of neuron representations in deep vision networks.

Rao, S., Mahajan, S., Böhle, M., and Schiele, B. (2024). Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Appendix

A.1 Additional results for concept clustering

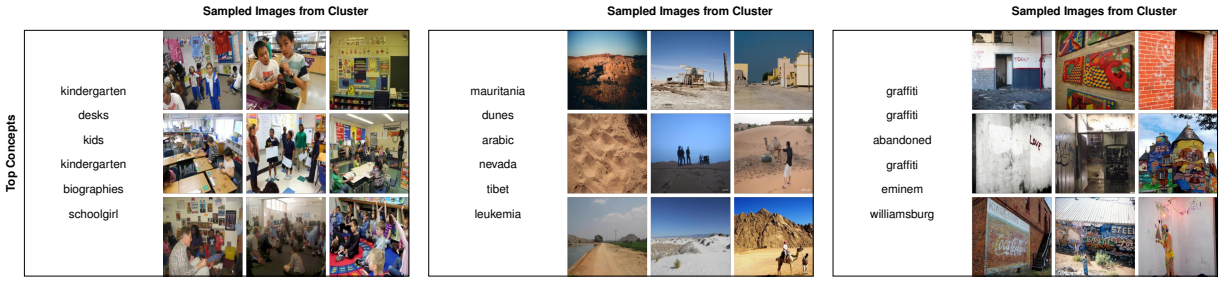


Figure 9: Concept clusters for Places365 with repeated concept names.

Often the clusters contained repeated words, due to mapping of different concepts to the same word in the vocabulary as described in Figure 9 . This problem can be easily addressed by simply using a larger vocabulary, as closely related concepts that do not have representative words in the vocabulary are going to be assigned to the most relevant word based on cosine similarity, which can coincide.

A.2 Additional results for local explanations

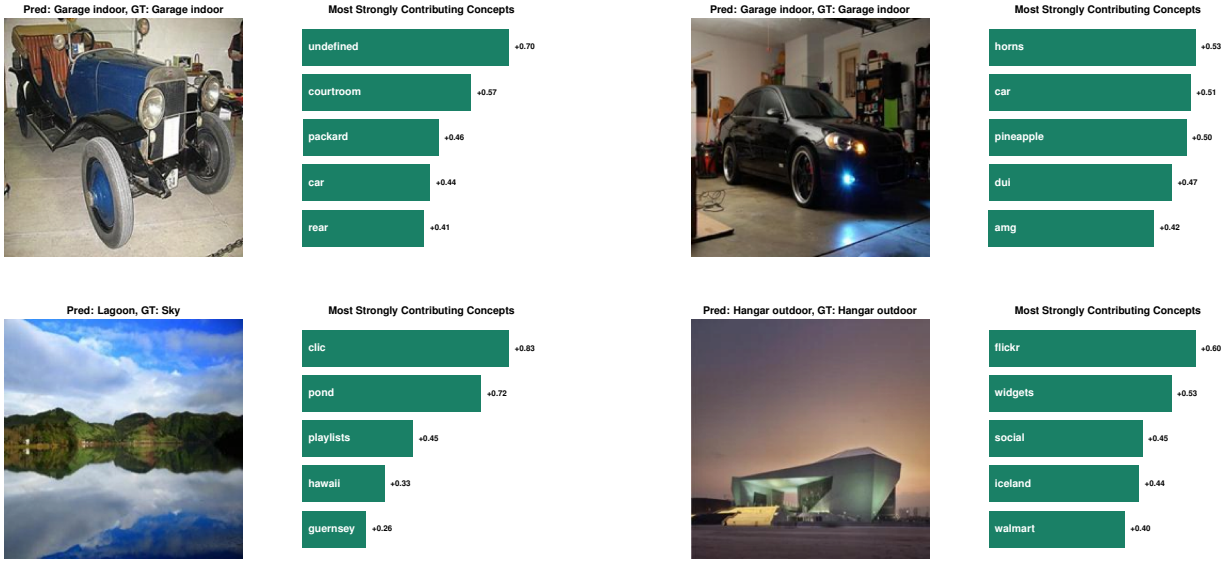


Figure 10: Local explanations for images with obscure top concepts contributions.

For example, words such as “undefined“, “clic“, “horns“, and “flickr“ appeared as top concepts for images that have no association with those concepts, which calls into question the robustness of this method. In Figure 10 for the top row, we expect the concept car to be the top concept and to have a high contribution score, since the main object in both examples is a car; however, the contribution of the concept "car" is not as prominent. The bottom left image, while having a relevant concept “pond“ has other contributing concepts that have little correlation with the content in the image. The bottom right example despite the correct prediction does not provide a meaningful explanation for the content of the image except possibly the word “walmart“ as it can be associated with a building which is present in the image, however, despite that it is quite a weak link, especially given that other concepts also provide little information about actual content of the image.

A.3 Additional visualizations for color augmentation experiment

Figure 11: Top 20 concepts extracted for: 1. Original Image 2. Grayscale Image 3. Inverted Color Map Image

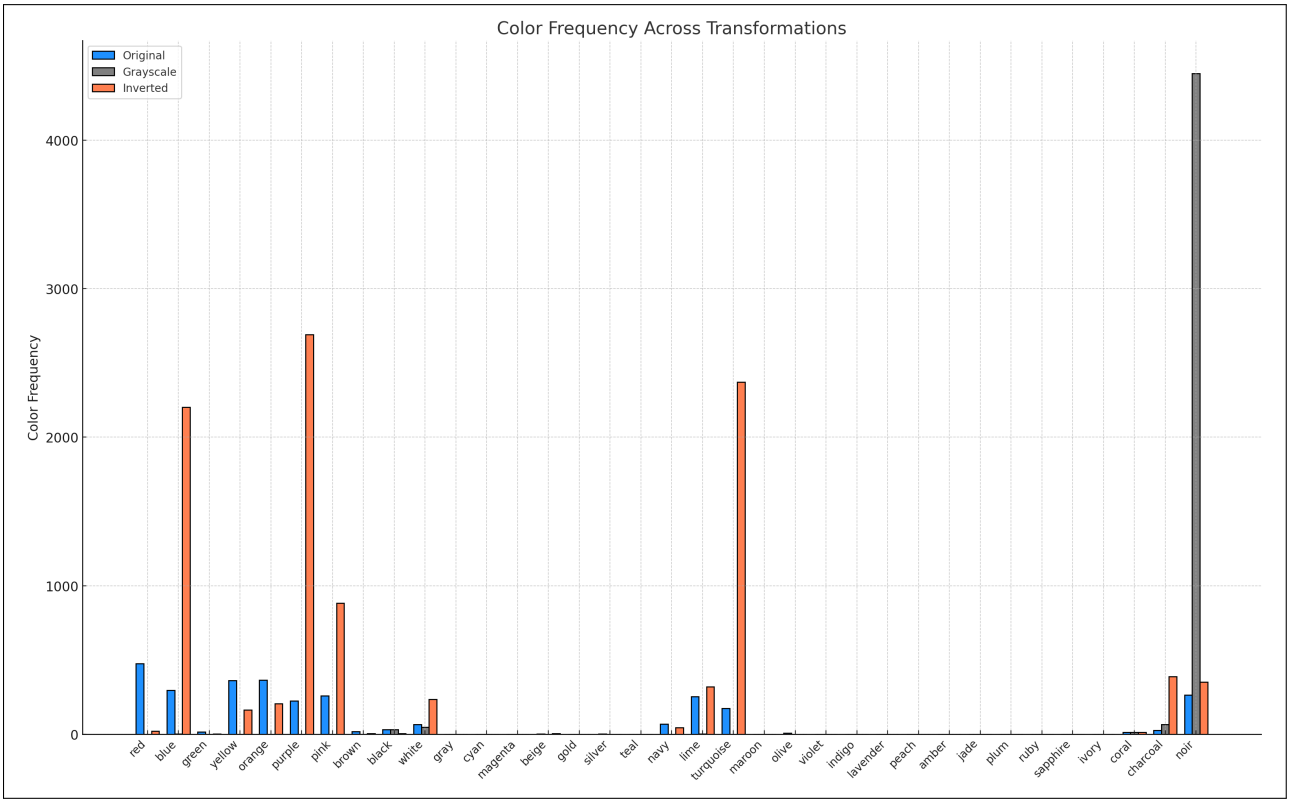


Figure 12: Quantity of colors observed in Top 20 concepts

Class	**Original Accuracy**	**Grayscale Accuracy**	**Difference**
Swimming pool outdoor	0.65	0.00	0.65
Garage outdoor	0.80	0.15	0.65
Legislative chamber	0.85	0.20	0.65
Greenhouse indoor	0.70	0.10	0.60
Excavation	0.85	0.25	0.60
Hotel room	0.60	0.60	0.00
Cockpit	0.95	0.95	0.00
Boxing ring	1.00	1.00	0.00
Ice skating rink outdoor	0.65	0.65	0.00
Natural history museum	0.70	0.70	0.00

Table 4: Accuracy Comparison for Classes with Largest Absolute Differences and No Changes between Original and Grayscale