On the Value of Labeled Data and Symbolic Methods for Hidden Neuron Activation Analysis

Abhilekha Dalal[1], Rushrukh Rayan[1], Adrita Barua[1], Eugene Y. Vasserman[1], Md Kamruzzaman Sarker[2], and Pascal Hitzler[1]

[1] Kansas State University, Manhattan, KS, USA
{adalal,rushrukh,adrita,eyv,hitzler}@ksu.edu
[2] Bowie State University, Prince George's County, MD, USA
ksarker@bowiestate.edu

Abstract. We introduce a novel model-agnostic post-hoc Explainable AI method that provides meaningful interpretations for hidden neuron activations in a Convolutional Neural Network. Our approach uses a Wikipedia-derived concept hierarchy with approx. 2 million classes as background knowledge, and deductive reasoning based Concept Induction for explanation generation. Additionally, we explore and compare the capabilities of off-the-shelf pre-trained multimodal-based explainable methods. Our evaluation shows that our neurosymbolic method holds a competitive edge in both quantitative and qualitative aspects.

Keywords: Explainable AI · Concept Induction · CNN · LLM

1 Introduction

While there has been significant progress on Explainable AI, the current state of the art is mostly restricted to explanation analyses based on a relatively small number of predefined explanation categories. This is problematic from a principled perspective, as it relies on the assumption that explanation categories pre-selected by humans would be viable explanation categories for deep learning systems – an as-yet unfounded conjecture. Other approaches rely on deep learning itself, e.g., Large Language Models (LLMs), to produce explanations [25] – which means that the explanation generation method in turn is yet another black box. Others rely on modified deep learning architectures, usually leading to a decrease in system performance compared to unmodified systems [43].

Others rely on low-level features such as Pixel attribution [3,38,36], thus not providing clear and explicit explanation concepts for human users.

For concept-based explanations, the lack of systematic approaches to consider a wide range of potential concepts that may influence the model appears to be a bottleneck. In some techniques [25], a list of frequently occurring English words has been utilized to represent a broad concept pool, which may suffice for general applications but lacks granularity for specialized fields.

Herein, we address several common shortcomings in the state of the art:

i. Concepts should not be hand-picked in light of completeness.
ii. Concept extraction methods should be inherently explainable.
iii. Explanations should be understandable without deep learning expertise.
iv. Candidate concepts pools should include meaningful relationships between concepts, that are made use of by the explanation approach.

We address these points by using *Concept Induction* as core mechanism, which is based on formal logic reasoning (in the Web Ontology Language OWL) and has originally been developed for Semantic Web applications [19].

We hypothesize that background knowledge coupled with inherently explainable deductive reasoning (here, Concept Induction) should be capable of generating meaningful explanations for the deep learning model we wish to explain.

To show that our approach can indeed provide meaningful explanations for hidden neuron activation, we instantiate it with a Convolutional Neural Network (CNN) architecture for image scene classification and a class hierarchy (i.e., a simple ontology) of approx. $2 \cdot 10^6$ classes derived from Wikipedia as the pool of explanation categories [32].[3] We demonstrate that our method performs competitively, as assessed through two separate evaluation methods, one statistical, one using Concept Activation analysis [17,7], when compared with other techniques such as CLIP-Dissect [25] and GPT-4 [1] as concept generation methods.

Core contributions of the paper are as follows.

1. A novel zero-shot model-agnostic Explainable AI method that explains existing pre-trained deep learning models through high-level concepts, utilizing symbolic reasoning over background knowledge as the source of explanation, which achieves state-of-the-art performance and is explainable by its nature.
2. A method to automatically extract *relevant concepts* through Concept Induction for any concept-based Explainable AI method, eliminating the need for manual selection of candidate concepts.
3. An in-depth comparison of explanation sources using statistical analysis for the hidden neuron perspective and Concept Activation analysis for the hidden layer perspective of our approach, namely with a pre-trained multimodal Explainable AI method (CLIP-Dissect [25]), and an LLM (GPT-4 [1]).

A significantly longer version of this paper is available online [8], which also includes a discussion of related work.

2 Concept Extraction

We explore and evaluate three concrete methods to generate high-level concepts for explaining hidden neuron activations. Fig. 1 is a high-level depiction of our workflow. Components are further discussed below and throughout the paper.

[3] See `https://anonymous.4open.science/r/xai-using-wikidataAndEcii-91D9/` for source code, input data, raw result files, and parameter settings for replication.

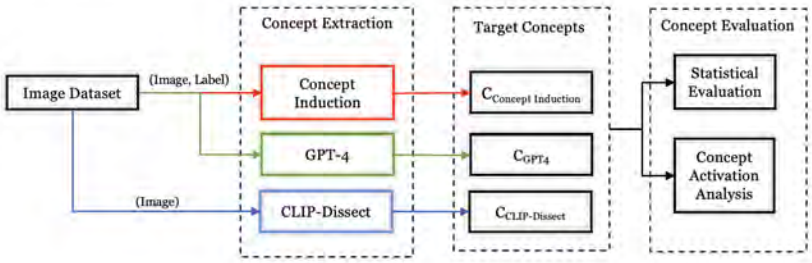

Fig. 1: An overview of the complete pipeline explored in this paper where Concept Extraction outlines the methods used to extract Target Concepts and Concept Evaluation outlines the evaluation methods.

Preparations: Scenario and CNN Training We use a scene classification from images scenario to demonstrate our approach, using the ADE20K dataset [46] which contains more than 27,000 images over 365 scenes, annotated with pixel-level objects and object part labels. This dataset provides a comprehensive resource for scene parsing, encompassing a diverse range of environments including urban, rural, indoor, and outdoor scenes. Though the distribution of these annotations naturally varies based on the prevalence of objects in real-world environments, this diversity in annotations still provides a rich foundation for our concept induction method. *The annotations are not used for CNN training*, but only for generating label hypotheses as described below.

We trained Resnet50V2 for the scene categories "bathroom", "bedroom", "building facade", "conference room", "dining room", "highway", "kitchen", "living room", "skyscraper", and "street". We chose scene categories with high numbers of images and such that some scene categories would have overlapping annotated objects – as this should make the hidden node activation analysis more interesting. We did not conduct any experiments on any other scene selections, i.e., *we did not change our scene selection based on any preliminary analyses.* It is important to note that the ResNet50V2 model employed in our study is pre-trained on the ImageNet dataset [9], allowing us to leverage transfer learning and to benefit from its comprehensive feature recognition capabilities. We then utilize this same pre-trained model to extract activation values for evaluating concept relevance.

We used Resnet50V2 because it achieved the highest accuracy (**86.46%**) among the networks we tested. Note that for our investigations of explainability of hidden neuron activations, achieving a very high accuracy for the scene classification task is not essential, but a reasonably high accuracy is necessary when considering models which would be useful in practice.

In the following, we detail the components shown in Fig. 1. We explain our use of Concept Induction for generating explanatory concepts, followed by our

utilization of CLIP-Dissect and GPT-4 for the same. We describe our two evaluation approaches in Section 3.

Concept Induction [19] is based on deductive reasoning over description logics, i.e., over logics relevant to ontologies, knowledge graphs, and generally the Semantic Web field [16,15] including the W3C OWL standard [30]. Concept Induction has already been shown, in other scenarios, to be capable of producing labels that are meaningful for humans inspecting the data [41]. A Concept Induction system accepts three inputs: (1) a set of positive examples P, (2) a set of negative examples N, and (3) a knowledge base (or ontology) K, all expressed as description logic theories, and all examples $x \in P \cup N$ occur as individuals (constants) in K. It returns description logic class expressions E such that $K \models E(p)$ for all $p \in P$ and $K \not\models E(q)$ for all $q \in N$. If no such class expressions exist, then it returns approximations for E together with a number of accuracy measures.

For scalability reasons, we use the heuristic Concept Induction system ECII [31] together with a background knowledge base that consists only of a hierarchy of approximately 2 million classes, curated from the Wikipedia concept hierarchy and presented in [32]. We use *coverage* as accuracy measure, defined as $\text{coverage}(E) = \frac{|Z_1|+|Z_2|}{|P \cup N|}$, where $Z_1 = \{p \in P \mid K \models E(p)\}$, $Z_2 = \{n \in N \mid K \not\models E(n)\}$, and P, N, K as above.

For our setting, positive and negative example sets contain images from ADE20K, i.e., we include the images in the background knowledge by linking them to the class hierarchy. For this, we use the object annotations available for the ADE20K images, but only part of the annotations for simplicity and scalability. More precisely, we only use the information that certain objects (such as windows) occur in certain images, and we do not make use of any of the richer annotations such as those related to segmentation.[4] All objects from all images are then mapped to classes in the class hierarchy using the Levenshtein string similarity metric [20] with edit distance 0. Mapping is in fact automated using the "combine ontologies" function of ECII.

The general idea for generating label hypotheses using Concept Induction is as follows: given a hidden neuron, P is a set of inputs (i.e., in this case, images) to the deep learning system that activate the neuron, and N is a set of inputs that do not activate the neuron (where P and N are the sets of positive and negative examples, respectively). As mentioned above, inputs are annotated with classes from the background knowledge for Concept Induction, but these annotations and the background knowledge are not part of the input to the deep learning system. ECII generates a label hypothesis[5] for the given neuron on inputs P, N, and the background knowledge.

[4] In principle, complex annotations in the form of sets of OWL axioms could of course be used, if a Concept Induction system is used that can deal with them, such as DL-Learner [19]. However DL-Learner does not quite scale to our size of background knowledge and task [33].

[5] In fact, it generates several, ranked, but we use only the highest ranked one for now.

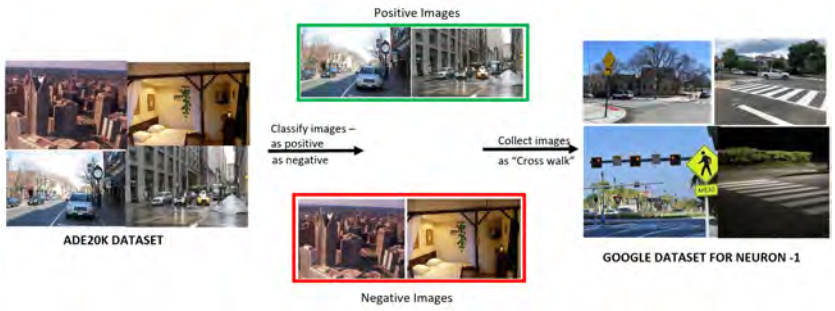

Fig. 2: Example of images that were used for generating and confirming the label hypothesis for neuron 1.

We first feed 1,370 ADE20K images to our trained Resnet50V2 and retrieve the activations of the dense layer. We chose to look at the dense layer because previous studies indicate [26] that earlier layers of a CNN respond to low level features such as lines, stripes, textures, colors, while layers near the final layer respond to higher-level features such as face, box, road, etc. The higher-level features align better with the nature of our background knowledge. The dense layer consists of 64 neurons, and we analyze each separately. Activation patterns involving more than one neuron are likely also informative in the sense that information may be distributed among several neurons, but this will be part of future investigations.

For each neuron, we calculate the maximum activation value across all images. We then take the positive example set P to consist of all images that activate the neuron with *at least* 80% of the maximum activation value, and the negative example set N to consist of all images that activate the neuron with *at most* 20% of the maximum activation value (or do not activate it at all). The highest scoring response of running ECII on these sets, together with the background knowledge described above, is shown in Table 1 for each neuron, together with the coverage of the ECII response. For each neuron, we call its corresponding label the *target label*, e.g., neuron 0 has target label "building." Note that some target labels consist of two concepts, e.g., "footboard, chain" for neuron 49 – this occurs if the corresponding ECII response carries two class expressions joined by a logical conjunction, i.e., in this example "footboard $\sqcap$ chain" (as description logic expression) or footboard$(x) \wedge$ chain(x) expressed in first-order predicate logic.

We give an example, depicted in Figure 2, for neuron 1. The green and red boxed images show positive and negative examples for neuron 1. Concept Induction yields "cross_walk" as target label. The example is continued below.

CLIP-Dissect [25] is a zero-shot Explainable AI method that associates high-level concepts with individual neurons in a designated layer. It utilizes the pre-trained multimodal model CLIP [28] to project a set of concepts and a set of

images into shared embedding space. Using Weighted Pointwise Mutual Information, it assesses the similarities between concepts and images in the hidden layer activation space to assign a concept to a neuron.

First, CLIP-Dissect uses a set of the most common 20,000 English vocabulary words as concepts. Then, we collect activations from our ResNet50v2 trained model for the ADE20K test images. This results in a matrix of dimensions (Number of Images $\times$ 64), where each row in the matrix represents an image through its 64 hidden neuron activation values. With these two sets of input, CLIP-Dissect assigns a label to each neuron such that the neuron is most activated when the corresponding concept is present in the image. This yields 22 unique concepts for the 64 neurons, with duplicate concepts for several neurons.

GPT-4 We leverage GPT-4's ability in generating concepts to distinguish between different image classes [24]. We use the same positive (P) and negative (N) example sets as for the Concept Induction approach, with some minor adjustments: while for Concept Induction, the negative example set (N) includes all images that activate the neuron with at most 20% of the maximum activation value, here we select only one image per class of images for each neuron to create the negative example set (N) due to GPT-4's input constraints.

The annotations of these images are passed to GPT-4 using prompts to identify concepts present in P but such that they are absent in N. We obtain a list of 3 concepts per neuron (wherein, 1 concept per class is randomly selected for evaluation) using the following zero-shot prompt: *Generate the top three classes of objects or general scenario that better represent what images in the positive set (P) have but the images in the negative set (N) do not.*

We acknowledge the influence of parameters such as: Temperature (set to 0) and Cumulative Probability Threshold (top_p, set to 1) in the output diversity of GPT-4. More detailed information regarding the experimental setup and prompt can be found in [4].

3 Evaluation

Confirming Label Hypotheses The three approaches described above produce label hypotheses for all investigated neurons – hypotheses that we will confirm or reject by testing the labels with new images. We use each of the target labels to search Google Images with the labels as keywords (requiring responses to be returns for *both* keywords if the label is a conjunction of classes, for Concept Induction). We call each such image a *target image* for the corresponding label or neuron. We use Imageye[6] to automatically retrieve the images, collecting up to 200 images that appear first in the Google Images search results, filtering for images in JPEG format and with a minimum size of 224x224 pixels (conforming to the size and format of ADE20K images).

[6] `https://chrome.google.com/webstore/detail/image-downloader-imageye/agionbommeaifngbhincahgmoflcikhm`

Table 1: Selected representative data from all three approaches as discussed throughout the text (the full version can be found in Appendix A). Images: Number of images used per label. Target %: Percentage of target images activating the neuron. Non-Target %: The same, but for all other images. **Bold** denotes neurons whose labels are considered confirmed.

Concept Induction					
Neuron	Obtained Label(s)	Images	Coverage	Target %	Non-Target %
0	**building**	**164**	**0.997**	**89.024**	**72.328**
1	**cross_walk**	**186**	**0.994**	**88.710**	**28.923**
3	**night_table**	**157**	**0.987**	**90.446**	**56.714**
6	dishcloth, toaster	106	0.999	16.038	39.078
11	river_water	157	0.995	31.847	22.309
CLIP-Dissect					
0	restaurants	140		55.000	59.295
3	**dresser**	**171**		**95.322**	**66.199**
6	dining	153		7.190	50.195
7	**bathroom**	**153**		**93.333**	**44.113**
11	highway	153		14.063	25.153
GPT-4					
0	Urban Landscape	176		54.545	59.078
1	**Street Scene**	**164**		**92.073**	**29.884**
3	**Bedroom**	**165**		**97.576**	**62.967**
8	**Bathroom**	**164**		**98.780**	**47.897**
12	Indoor Home Setting	164		62.805	47.205

For each retrieval label, we use 80% of the obtained images, reserving the remaining 20% for the statistical evaluation described later in the section. The number of images used in the hypothesis confirmation step, for each label, is given in the tables. These images are fed to the network to check (a) whether the target neuron (with the retrieval label as target label) activates, and (b) whether any other neurons activate. The Target % column of Tables 1 show the percentage of the target images that activate each neuron.

Returning to our example neuron 1 in the Concept Induction case (Fig. 2), 88.710% of the images retrieved with the label "cross_walk" activate it. However, this neuron activates only for 28.923% (indicated in the Non-Target % column) of images retrieved using all other labels excluding "cross_walk."

We define a target label for a neuron to be *confirmed* if it activates for $\geq 80\%$ of its target images regardless of how much or how often it activates for non-target images. The cut-offs for neuron activation and label hypothesis confirmation are chosen to ensure strong association and responsiveness to images retrieved under the target label, but 80% is somewhat arbitrary and could be chosen differently. For our example neuron 1, we consider the label "cross_walk" confirmed for neuron 1 since $88.710 \geq 80$.

We obtain **19, 5, and 14 (distinct)** confirmed concepts from Concept Induction, CLIP-Dissect, and GPT-4, respectively; see Table 2 and Appendix A.2.

Statistical Evaluation After generating the confirmed labels (as above), we evaluate the node labeling using the remaining images from those retrieved from Google Images as described earlier. Table 2 shows the results, omitting neurons that were not activated by any image, i.e., their maximum activation was 0.

We consider each neuron-label pair (rows in Table 2) to be a hypothesis, e.g., for neuron 1, from Concept Induction the hypothesis is that it activates more strongly for images retrieved using the keyword "cross_walk" than for images retrieved using other keywords. The corresponding null hypothesis is that activation values are *not* different. Table 2 shows partial results, from Concept Induction we get 20 hypotheses to test, corresponding to the 20 neurons with confirmed labels (recall that a double label such as neuron 16's "mountain, bushes" is treated as one label consisting of the conjunction of the two keywords). For CLIP-Dissect, we get 8 hypotheses to test, reflecting 8 confirmed labels. GPT-4 yields 27 hypotheses, representing 27 confirmed labels. Full details of Table 2 are in the Appendix A.2.

There is no reason to assume that activation values would follow a normal distribution, or that the preconditions of the central limit theorem would be satisfied. We therefore base our statistical assessment on the Mann-Whitney U test [22] which is a non-parametric test that does not require a normal distribution. Essentially, by comparing the ranks of the observations in the two groups, the test allows us to determine if there is a statistically significant difference in the activation percentages between the target and non-target labels.

The resulting z-scores and p-values are shown in Table 2 and are further discussed in Section 4. For our running example (neuron 1), we analyze the remaining 47 target images (20% of the images retrieved during the label hypothesis confirmation step). Of these, 43 (91.49%) activate the neuron with a mean and median activation of 4.17 and 4.13, respectively. Of the remaining (non-target) images in the evaluation (the sum of the image column in Table 2 minus 47), only 28.94% activate neuron 1 for a mean of 0.67 and a median of 0.00. The Mann-Whitney U test yields a z-score of **-8.92 and** $p < 0.00001$. The negative z-score indicates that the activation values for non-target images are indeed lower than for the target images, rejecting the null hypothesis.

Concept Activation Analysis We employ *Concept Activation* [17,7], a concept-based explainable AI technique which works with a *pre-defined* set of concepts. It explains a pre-trained model by measuring the presence of *concepts* in hidden-layer activations of a given image for a particular layer. We evaluate the label hypotheses obtained from all three methods using Concept Activation Analysis. Note that we do not restrict this analysis to only confirmed concepts, as the Concept Activation Analysis approach has not been developed with such a confirmation step as part of it.

For each concept, a set of images are collected using Imageye (exactly as described above) and a concept classifier (Support Vector Machine (SVM)) is trained. The dataset given to the concept classifier requires some pre-processing: (a) The dataset for each concept classifier consists of images that exhibit the presence of the concept (label=1) and images where the concept is absent (label=0),

Table 2: Statistical Evaluation details for all three approaches(full version can be found in Appendix A.2). Images: Number of images.# Activations: (targ(et)): Percentage of target images activating the neuron;(non-t):Same for all other images used in the evaluation. Mean/Median (targ(et)/non-t(arget)): Mean/median activation value for target and non-target images, respectively.

Concept Induction										
Neuron	Label(s)	Images	# Activations (%)		Mean		Median		z-score	p-value
			targ	non-t	targ	non-t	targ	non-t		
0	building	42	80.95	73.40	2.08	1.81	2.00	1.50	-1.28	0.0995
1	cross_walk	47	91.49	28.94	4.17	0.67	4.13	0.00	-8.92	<.00001
18	slope	35	91.43	68.85	1.59	1.37	1.44	1.00	-2.03	0.0209
19	wardrobe, air_conditioning	28	89.29	65.81	2.30	1.28	2.30	0.84	-4.00	<.00001
48	road	42	100.00	74.46	6.15	2.68	6.65	2.30	-7.78	<.00001
49	footboard, chain	32	84.38	66.41	2.63	1.67	2.30	1.17	-2.58	0.0049
CLIP-Dissect										
3	dresser	43	93.02	64.61	2.59	1.42	2.62	0.68	5.01	<0.0001
7	bathroom	46	89.47	41.56	2.02	1.01	2.15	0.00	5.45	<0.0001
18	dining	36	94.87	76.82	3.01	1.85	3.11	1.44	4.52	<0.0001
GPT-4										
1	Street Scene	42	90.50	30.40	3.80	0.70	4.20	0.00	-9.62	<0.0001
14	Living Room	41	78.00	67.50	1.40	1.30	1.20	0.90	-0.77	0.4413
17	Dining Room	40	97.50	45.90	2.20	0.60	2.50	0.00	-8.29	<0.0001
18	Outdoor Scenery	41	100.00	76.10	2.30	1.50	2.20	1.20	-3.96	<0.0001
30	Kitchen	43	86.00	38.60	2.60	0.80	2.70	0.00	-7.22	<0.0001
31	Urban Street Scene	41	80.50	65.70	1.80	1.30	1.70	0.90	-2.4	0.164

(b) as we are interested in validating the concepts in the hidden layer activation space the dataset is passed through the ResNet50V2 pre-trained model. The activation values of each image in the dense layer is saved.

The transformed dataset is split into train (80%) and test (20%) datasets. Thereafter, an SVM classifier is trained using the train split. We have used both linear (Concept Activation Vector, CAV) and non-linear (Concept Activation Region, CAR) kernels to assess which decision boundary separates the presence/absence of a concept best. Subsequently, the test dataset is used to see to what extent the concept classifier can classify the existence of concepts.

4 Results

We evaluate the concepts extracted by Concept Induction, CLIP-Dissect, and GPT-4 on ADE20K Test dataset split from two different perspectives (Section 3):

i. For each neuron of the dense layer, we identify the concepts that activate them the most (Statistical Evaluation).

ii. For each concept, we measure its degree of relevance across the entire dense layer activation space (Concept Activation Analysis).

Our findings suggest that Concept Induction consistently performs well in both sets of evaluations. From the statistical evaluation, it is evident that Concept Induction achieves better performance over the other methods. In the Concept Activation Analysis, quantitative measures reveal that Concept Induction achieves comparable performance to CLIP-Dissect, with GPT-4 exhibiting the

Table 3: Concept Accuracy in Hidden Layer Activation Space of selected Concepts (the full version can be found in Appendix A.3) extracted using Concept Induction, CLIP-Dissect, and GPT-4

Concept Induction				
Concept Name	CAR		CAV	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
Air Conditioner	0.8994	0.8415	0.811	0.8659
Baseboard	0.875	0.8717	0.8846	0.9102
Body	0.9035	0.8857	0.8642	0.9
CLIP-Dissect				
Bathroom	0.9700	0.9474	0.9400	0.9474
Bed	0.9587	0.9500	0.9437	0.9125
Bedroom	0.9167	0.9167	0.9137	0.9048
GPT-4				
Bedroom	0.9851	0.9761	0.9660	0.9523
Bathroom	0.9176	0.9024	0.9068	0.8902
Bathroom Interior	0.9273	0.9146	0.9241	0.9268

lowest performance. Moreover, the Concept Induction approach demonstrates several notable qualitative advantages over both CLIP-Dissect and GPT-4:

- CLIP-Dissect and GPT-4 are black-box models used as a concept extraction method to explain a probing network. This approach to explainability is itself not readily explainable. In contrast, Concept Induction, serving as a concept extraction method, inherently offers explainability as it operates on deductive reasoning principles.
- CLIP-Dissect relies on a common 20K English vocabulary as the pool of concepts, whereas Concept Induction is supported by a carefully constructed background knowledge (with about 2M concepts), affording greater control over the pool of possible explanations through hierarchical relationships.
- While GPT-4/CLIP-Dissect emulate intuitive and rapid decision-making processes, Concept Induction follows a systematic and logical decision-making approach – thereby rendering our approach to be explainable by nature.

Table 2 shows that Concept Induction analysis with large-scale background knowledge yields meaningful labels that stably explain neuron activation. Of the 20 null hypotheses from Concept Induction, 19 are rejected at $p < 0.05$, but most (all except neurons 0, 18 and 49) are rejected at much lower p-values. Only neuron 0's null hypothesis could not be rejected. With CLIP-Dissect, all 8 null hypotheses are rejected at $p < 0.05$, and with GPT-4, 25 out of 27 null hypotheses are rejected at $p < 0.05$, with exceptions for neurons 14 and 31. Excluding repeating concepts, Concept Induction yields **18** statistically validated hypotheses, CLIP-Dissect yields **5**, and GPT-4 yields **12**.

Mann-Whitney U results show that, for most neurons listed in Tables 2 (with $p < 0.00001$), activation values of target images are *overwhelmingly* higher than

Method	90-100%	80-89%	<80%
Concept Induction	14	6	0
GPT-4	10	4	0
CLIP-Dissect	4	1	0

Table 4: Count of statistically confirmed Concepts from each method (Table 2) such that their percentage of target activation is binned into 3 regions based on their degree of relevance.

Method	CAV			CAR			Count of Concepts		
	Mean	Median	SD	Mean	Median	SD	90-100%	80-89%	<80%
Concept Induction	0.9154	0.9230	0.0449	0.9150	0.9310	0.0465	46	22	1
CLIP-Dissect	0.9160	0.9146	0.0389	0.9259	0.9293	0.0443	17	5	0
GPT-4	0.8757	0.8863	0.0817	0.8887	0.9024	0.0690	11	9	1

Table 5: Mean, Median, and Standard Deviation (SD) of Concept Activation Analysis Test Accuracies, and Count of Concepts with their Concept Classifier Test Accuracies binned into 3 regions – High (90-100%), Medium (80-89%), and Low (<80%) relevance

that of non-target images. The negative z-scores with high absolute values informally indicate the same, as do the mean and median values. Neurons 16 and 49 of Concept Induction method in Table 2, for which the hypotheses also hold but with $p < 0.05$ and $p < 0.01$, respectively, still exhibit statistically significant higher activation values for target than for non-target images, but not overwhelmingly so. This can also be informally seen from lower absolute values of the z-scores, and from smaller differences between the means and the medians.

Although, solely based on the values of Mean Test Accuracy, CLIP-Dissect demonstrates a slightly superior performance compared to Concept Induction, and GPT-4 performs the least (in Table 5), we contend that the substantially higher number of concepts generated by Concept Induction allows CLIP-Dissect to achieve a marginally higher mean test accuracy. In the top 22 (equal to the number of concepts generated by CLIP-Dissect) test accuracies of concepts extracted by Concept Induction, the Mean Test Accuracies are **0.9599** (CAV) and **0.9584** (CAR). In k-fold cross validation tests, all concepts in Concept Activation analysis achieve $p < 0.05$. Using Mann-Whitney U test, we ascertain that CLIP-Dissect outperforms GPT-4 on CAR, and Concept Induction surpasses GPT-4 on CAV, while Concept Induction and CLIP-Dissect show now statistically significant difference (see Table 7).

5 Discussion

From the statistical evaluation, based on the percentage of target activation and from Concept Activation Analysis, based on the concepts' test accuracies, we categorize all confirmed concepts into three regions: high (90-100%), medium (80-89%), and low (< 80%) relevance concepts. Tables 4 and 5 show that Concept Induction produces a notably larger number of high-relevance concepts compared to other methods. Table 2, shows 8 and 27 statistically confirmed concepts from the CLIP-Dissect and GPT-4 method, respectively. However, upon

closer examination, it becomes evident that some concepts are duplicated across the tables.

Disregarding the duplicates, we have only 5 and 14 confirmed concepts, respectively, as opposed to 18 from Concept Induction.

This difference is likely due to Concept Induction's reliance on rich background knowledge, necessitating additional preprocessing but offering additional value. While a candidate concept pool of 20K English vocabulary words or off-the-shelf GPT-4 may not be universally effective, Concept Induction's ability to generate extensive, high-relevance concepts underscores the importance of well-engineered background knowledge.

If an application does not require comprehensive concept-based explanations, CLIP-Dissect/GPT-4 may serve as a useful solution, especially when time is limited. However, for detailed concept-based analysis, preparing background knowledge and leveraging Concept Induction is crucial. For CLIP-Dissect/GPT-4, it is unclear how to meticulously craft the pool of candidate concepts. By employing a background knowledge base, it is possible to define a large pool of potential explanations, tailored to the application scenario, with additional relationships among concepts. Concept Induction facilitates deductive reasoning utilizing this background knowledge, inherently offering transparency and flexibility in shaping the candidate concept pool.

While it is important to investigate methods that assess the relevance of concepts in hidden layer computations within a given candidate pool, it is equally, if not more, vital to thoughtfully design this pool. Neglecting this aspect could result in overlooking crucial concepts essential for gaining insights into hidden layer computations. Our approach offers a way to integrate rich background knowledge and extract meaningful concepts from it.

Our focus on dense layer activations, while providing valuable insights, represents only a part of what the deep representation encodes. The dense layer likely relates to clear-cut concepts that separate output classes, aligning well with our goal of identifying high-level, interpretable concepts. However, these concepts are influenced by combinations of features from previous layers. This limitation underscores the complex nature of deep neural networks, where concepts identified at the dense layer result from hierarchical feature compositions throughout the network. While our method offers meaningful insights into these high-level concepts, it may not fully capture the nuanced feature interactions in earlier layers. Nonetheless, focusing on the dense layer allows us to extract concepts more directly relevant to the network's final decision-making process, balancing interpretability with the complexity of internal representations. Future work could explore extending our method to analyze concept formation across multiple layers, potentially revealing a more comprehensive picture of the network's decision-making process.

One drawback of utilizing Concept Induction (and GPT-4) is its dependency on object annotations, which serve as data points in the background knowledge. In contrast, CLIP-Dissect operates without the need for labels and can function with any provided set of images.

We view this as a trade-off that must be carefully considered based on the application scenario. If the application is broad and does not demand a meticulous design of candidate concepts, then employing approaches like CLIP-Dissect can be advantageous. Conversely, for applications that are focused or specialized, CLIP-Dissect may only provide broadly relevant concepts.

Our focus has been primarily on assessing the comparative effectiveness of Concept Induction within the confines of Convolutional Neural Network architecture using ADE20K Image data. Nevertheless, it is imperative to investigate its suitability across different architectures and with diverse datasets. Given the model-agnostic nature of our approach, our results suggest its potential applicability across a range of neural network architectures, datasets, and modalities. While we utilized a Wikipedia Concept Hierarchy comprising 2 million concepts, it would be intriguing to observe the outcomes of our approach when powered by a domain-specific Knowledge Graph in specialized domains such as Medical Diagnosis.

6 Conclusion

Concept Induction on background knowledge results in meaningful labeling of hidden neuron activations, confirmed by statistical analysis. This enables us to identify concepts that trigger pronounced responses from neurons, thus "explaining" neuron activations. Additionally, Concept Activation Analysis measures the relevance of each concept across the dense layer activation space. This combined approach provides a comprehensive understanding of hidden layer computations. To our knowledge, this approach, particularly the use of large-scale background knowledge, is novel, allowing for diverse label categories. Our research compares the performance of approaches like CLIP-Dissect and GPT-4, demonstrating that Concept Induction has an edge in our setting where labeled data is available. However, trade-offs between methods are acknowledged (Section 5). Overall, our line of work aims for comprehensive hidden layer analysis in deep learning systems, facilitating interpretation of activations as implicit input features, thus explaining system input-output behavior.

Acknowledgement. The authors acknowledge partial funding under National Science Foundation grants 2119753 and 2333782.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
3. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods (2018), `http://arxiv.org/abs/1806.08049`

4. Barua, A., Widmer, C., Hitzler, P.: Concept induction using LLMs: a user experiment for assessment (2024), https://arxiv.org/abs/2404.11875, submitted to NeSy 2024
5. Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B., Torralba, A.: Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences **117**(48), 30071–30078 (2020)
6. Confalonieri, R., Weyde, T., Besold, T.R., del Prado Martín, F.M.: Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. Artificial Intelligence **296**, 103471 (2021)
7. Crabbé, J., van der Schaar, M.: Concept activation regions: A generalized framework for concept-based explanations. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 2590–2607. Curran Associates, Inc. (2022)
8. Dalal, A., Rayan, R., Barua, A., Vasserman, E.Y., Sarker, M.K., Hitzler, P.: On the value of labeled data and symbolic methods for hidden neuron activation analysis (2024)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
10. Díaz-Rodríguez, N., Lamas, A., Sanchez, J., Franchi, G., Donadello, I., Tabik, S., Filliat, D., Cruz, P., Montes, R., Herrera, F.: Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. Information Fusion **79**, 58–83 (2022)
11. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189 – 1232 (2001). https://doi.org/10.1214/aos/1013203451
12. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
13. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics **24**(1), 44–65 (2015). https://doi.org/10.1080/10618600.2014.907095
14. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: XAI – Explainable artificial intelligence. Science robotics **4**(37) (2019). https://doi.org/10.1126/scirobotics.aay7120
15. Hitzler, P.: A review of the semantic web field. Commun. ACM **64**(2), 76–83 (2021). https://doi.org/10.1145/3397512, https://doi.org/10.1145/3397512
16. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman and Hall/CRC Press (2010), http://www.semantic-web-book.org/
17. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2668–2677. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/kim18d.html
18. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (Un)Reliability of Saliency Methods, p.

267–280. Springer-Verlag, Berlin, Heidelberg (2022), https://doi.org/10.1007/978-3-030-28954-6_14

19. Lehmann, J., Hitzler, P.: Concept learning in description logics using refinement operators. Mach. Learn. **78**(1-2), 203–250 (2010). https://doi.org/10.1007/s10994-009-5146-2
20. Levenshtein, V.I.: On the minimal redundancy of binary error-correcting codes. Inf. Control. **28**(4), 268–291 (1975). https://doi.org/10.1016/S0019-9958(75)90300-9
21. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
22. McKnight, P.E., Najab, J.: Mann-whitney u test. In: The Corsini Encyclopedia of Psychology. Wiley (2010)
23. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: A comprehensive review. Artificial Intelligence Review pp. 1–66 (2022)
24. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: The Eleventh International Conference on Learning Representations. ICLR (2023), https://openreview.net/forum?id=FlCg47MNvBA
25. Oikarinen, T., Weng, T.W.: CLIP-Dissect: Automatic description of neuron representations in deep vision networks. In: International Conference on Learning Representations. ICLR (2023), https://openreview.net/forum?id=iPWiwWHc1V
26. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization: How neural networks build up their understanding of images. Distill (November 2017). https://doi.org/10.23915/distill.00007
27. Procko, T., Elvira, T., Ochoa, O., Rio, N.D.: An exploration of explainable machine learning using semantic web technology. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC). pp. 143–146. IEEE Computer Society (jan 2022). https://doi.org/10.1109/ICSC52841.2022.00029
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, vol. 139 (2021)
29. Ribeiro, M.T., Singh, S., Guestrin, C.: "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. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778
30. Rudolph, S., Krötzsch, M., Patel-Schneider, P., Hitzler, P., Parsia, B.: OWL 2 web ontology language primer (second edition). W3C recommendation, W3C (Dec 2012), https://www.w3.org/TR/2012/REC-owl2-primer-20121211/
31. Sarker, M.K., Hitzler, P.: Efficient concept induction for description logics. In: The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI) The Thirty-First Innovative Applications of Artificial Intelligence Conference (IAAI), The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI). pp. 3036–3043. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33013036
32. Sarker, M.K., Schwartz, J., Hitzler, P., Zhou, L., Nadella, S., Minnery, B.S., Juvina, I., Raymer, M.L., Aue, W.R.: Wikipedia knowledge graph for explainable AI. In: Villazón-Terrazas, B., Ortiz-Rodríguez, F., Tiwari, S.M., Shandilya, S.K. (eds.)

Proceedings of the Knowledge Graphs and Semantic Web Second Iberoamerican Conference and First Indo-American Conference (KGSWC). Communications in Computer and Information Science, vol. 1232, pp. 72–87. Springer (2020). https://doi.org/10.1007/978-3-030-65384-2_6

33. Sarker, M.K., Xie, N., Doran, D., Raymer, M.L., Hitzler, P.: Explaining trained neural networks with semantic web technologies: First steps. In: Besold, T.R., d'Avila Garcez, A.S., Noble, I. (eds.) Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy). CEUR Workshop Proceedings, vol. 2003. CEUR-WS.org (2017), https://ceur-ws.org/Vol-2003/NeSy17_paper4.pdf
34. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
35. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Why did you say that? (2016), http://arxiv.org/abs/1611.07450
36. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/shrikumar17a.html
37. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR **abs/1312.6034** (2013), https://api.semanticscholar.org/CorpusID:1450294
38. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. p. 180–186. AIES '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3375627.3375830
39. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. CoRR **abs/1706.03825** (2017), http://arxiv.org/abs/1706.03825
40. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. CoRR **abs/1711.00399** (2017), http://arxiv.org/abs/1711.00399
41. Widmer, C., Sarker, M.K., Nadella, S., Fiechter, J., Juvina, I., Minnery, B.S., Hitzler, P., Schwartz, J., Raymer, M.L.: Towards human-compatible XAI: Explaining data differentials with concept induction over background knowledge (2022). https://doi.org/10.48550/arXiv.2209.13710
42. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
43. Zarlenga, M.E., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., Jamnik, M.: Concept embedding models: Beyond the accuracy-explainability trade-off. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
44. Zhang, R., Madumal, P., Miller, T., Ehinger, K.A., Rubinstein, B.I.P.: 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). https://doi.org/10.1609/aaai.v35i13.17389

45. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2131–2145 (2018)
46. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)

A Appendices

A.1 Related work

Efforts to demystify deep learning [14,2,23] are underway. Explainable methods can be categorized based on understanding input data e.g., feature summarizing [35,29] or the model's internal representation e.g., node summarizing [45,5]. These methods further classify into model-specific [35] or model-agnostic [29] approaches. Some methods rely on human interpretation of explanatory data, such as counterfactual questions [40].

Model-agnostic feature attribution techniques, like LIME [29] and SHAP [21], aim to clarify model predictions by assessing individual feature influence. However, they face challenges such as explanation instability [3] and susceptibility to biased classifiers [38]. Pixel attribution, seeks to understand predictions by attributing significance to individual pixels [37,34,39]. However, it has notable limitations, especially with ReLU activation [36] and adversarial perturbations [18], leading to interpretability inconsistencies.

[17,7] developed explanations using supervised learning and hand-picked concepts. These methods utilize classifiers on target concepts, with weights serving as Concept Activation Vectors (CAVs). [12] employs image segmentation and clustering for concept curation, though this approach may lose information and only works with visible concepts. [44] proposed enhancements using Non-negative Matrix Factorization to address information loss. Individual Conditional Expectation (ICE) plots [13] and Partial Dependency Plots [11] offer local and global perspectives on prediction-feature relationships but struggle with complex feature interactions.

Previous studies suggest that hidden neurons may represent high-level concepts [45,5], but these methods often require semantic segmentation [42] (resource-intensive) or explicit concept annotations [17]. Some research have utilized Semantic Web data for explaining deep learning models [6,10], and Concept Induction for providing explanations [33,27]. However, their focus was on analyzing input-output behavior, generating explanations for the overall system.

CLIP-Dissect [25] represents work similar to ours, employing a different approach. They utilize the CLIP pre-trained model, employing zero-shot learning to associate images with labels. Label-Free Concept Bottleneck Models [24], building upon CLIP-Dissect, use GPT-4 [1] for concept set generation. However, CLIP-Dissect has limitations that may be challenging to overcome without significant changes to the approach. These include limited accuracy in predicting output labels based on concepts in the last hidden layer and difficulty in transferring to other modalities or domain-specific applications. The Label-Free approach inherits these limitations and may compromise explainability, as it uses a concept derivation method that is not inherently explainable.

A.2 Detailed results of Statistical Evaluation

Table 6: Evaluation details for all three approaches as discussed in Section 3. Images: Number of images used for evaluation. # Activations: (targ(et)): Percentage of target images activating the neuron;(non-t):Same for all other images used in the evaluation. Mean/Median (targ(et)/non-t(arget)): Mean/median activation value for target and non-target images, respectively.

Concept Induction										
Neuron	Label(s)	Images	# Activations (%)		Mean		Median		z-score	p-value
			targ	non-t	targ	non-t	targ	non-t		
0	building	42	80.95	73.40	2.08	1.81	2.00	1.50	-1.28	0.0995
1	cross_walk	47	91.49	28.94	4.17	0.67	4.13	0.00	-8.92	<.00001
3	night_table	40	100.00	55.71	2.52	1.05	2.50	0.35	-6.84	<.00001
8	shower_stall, cistern	35	100.00	54.40	5.26	1.35	5.34	0.32	-8.30	<.00001
16	mountain, bushes	27	100.00	25.42	2.33	0.67	2.17	0.00	-6.72	<.00001
18	slope	35	91.43	68.85	1.59	1.37	1.44	1.00	-2.03	0.0209
19	wardrobe, air_conditioning	28	89.29	65.81	2.30	1.28	2.30	0.84	-4.00	<.00001
22	skyscraper	39	97.44	56.16	3.97	1.28	4.42	0.33	-7.74	<.00001
29	lid, soap_dispenser	33	100.00	80.47	4.38	2.14	4.15	1.74	-5.92	<.00001
30	teapot, saucepan	27	85.19	49.93	2.52	1.05	2.23	0.00	-4.28	<.00001
36	tap, crapper	23	91.30	70.78	3.24	1.75	2.82	1.29	-3.59	<.00001
41	open_fireplace, coffee_table	31	80.65	15.11	2.03	0.14	2.12	0.00	-7.15	<.00001
43	central_reservation	40	97.50	85.42	7.43	3.71	8.08	3.60	-5.94	<.00001
48	road	42	100.00	74.46	6.15	2.68	6.65	2.30	-7.78	<.00001
49	footboard, chain	32	84.38	66.41	2.63	1.67	2.30	1.17	-2.58	0.0049
51	road, car	21	100.00	47.65	5.32	1.52	5.62	0.00	-6.03	<.00001
54	skyscraper	39	100.00	71.78	4.14	1.61	4.08	1.12	-7.60	<.00001
56	flusher, soap_dish	53	92.45	64.29	3.47	1.48	3.08	0.86	-6.47	<.00001
57	shower_stall, screen_door	34	97.06	32.31	2.60	0.61	2.53	0.00	-7.55	<.00001
63	edifice, skyscraper	45	88.89	48.38	2.41	0.83	2.36	0.00	-6.73	<.00001
CLIP-Dissect										
3	dresser	43	93.02	64.61	2.59	1.42	2.62	0.68	5.01	<0.0001
7	bathroom	46	89.47	41.56	2.02	1.01	2.15	0.00	5.45	<0.0001
18	dining	36	94.87	76.82	3.01	1.85	3.11	1.44	4.52	<0.0001
33	bathroom	38	71.05	34.02	1.28	0.47	0.95	0.00	4.91	<0.0001
38	bathroom	38	84.21	31.71	1.79	0.54	1.83	0.00	7.14	<0.0001
43	highways	32	100.00	63.87	7.00	3.14	6.39	2.64	6.17	<0.0001
49	bedroom	40	97.50	55.77	3.48	1.63	3.43	0.63	6.05	<0.0001
50	bedroom	40	97.50	63.21	4.56	1.30	4.60	0.66	8.70	<0.0001
GPT-4										
1	Street Scene	42	90.50	30.40	3.80	0.70	4.20	0.00	-9.62	<0.0001
3	Bedroom	42	97.60	63.40	4.70	1.20	4.90	0.70	-9.05	<0.0001
6	Kitchen	43	83.70	52.00	2.40	1.00	2.00	0.10	-5.06	<0.0001
8	Bathroom	41	100.00	44.10	4.10	1.00	4.10	0.00	-9.57	<0.0001
14	Living Room	41	78.00	67.50	1.40	1.30	1.20	0.90	-0.77	0.4413
17	Dining Room	40	97.50	45.90	2.20	0.60	2.50	0.00	-8.29	<0.0001
18	Outdoor Scenery	41	100.00	76.10	2.30	1.50	2.20	1.20	-3.96	<0.0001
22	Street Scene	42	90.50	50.10	3.00	1.40	3.30	0.00	-5.95	<0.0001
23	Street Scene	42	85.70	20.70	2.40	0.30	2.10	0.00	-10.83	<0.0001

Neuron	Label(s)	Images	# Activations (%)		Mean		Median		z-score	p-value
			targ	non-t	targ	non-t	targ	non-t		
29	Bathroom	41	90.20	68.40	2.60	1.50	2.40	1.00	-4.05	<0.0001
30	Kitchen	43	86.00	38.60	2.60	0.80	2.70	0.00	-7.22	<0.0001
31	Urban Street Scene	41	80.50	65.70	1.80	1.30	1.70	0.90	-2.4	0.164
36	Bathroom	41	100.00	61.30	3.10	1.20	2.80	0.60	-7.48	<0.0001
38	Living Room	41	92.70	54.30	2.00	1.00	2.20	0.30	-5.53	<0.0001
39	Bicycle	39	84.60	47.40	2.10	0.90	2.40	0.00	-5.64	<0.0001
41	Living Room	41	97.60	42.00	2.60	0.60	2.30	0.00	-9.31	<0.0001
43	Outdoor Urban Scene	41	92.70	56.30	4.10	2.40	4.30	1.00	-4.42	<0.0001
44	Kitchen Scene	42	81.00	43.40	2.30	1.00	2.10	0.00	-5.43	<0.0001
48	Urban Street Scene	41	100.00	52.60	4.90	2.30	4.80	0.40	-6.03	<0.0001
49	Bedroom	42	95.20	35.00	3.80	0.70	4.00	0.00	-10.31	<0.0001
50	Living Room	41	97.60	63.90	3.00	1.20	2.60	0.60	-6.78	<0.0001
51	Street Scene	42	95.20	42.90	5.70	1.50	6.10	0.00	-9.05	<0.0001
56	Toilet Brush	42	97.60	34.60	3.60	0.70	3.60	0.00	-10.48	<0.0001
57	Bathroom Interior	41	92.70	40.50	3.00	0.80	2.90	0.00	-8.35	<0.0001
59	Urban Street Scene	41	82.90	26.30	2.70	0.50	2.50	0.00	-9.06	<0.0001
62	Dining Room	40	90.00	43.90	3.30	0.80	3.70	0.00	-8.64	<0.0001
63	Cityscape	39	97.40	48.50	2.80	0.70	2.40	0.00	-8.76	<0.0001

A.3 Detailed results of Concept Activation Analysis

Table 7: Mann-Whitney U Test results on Concept Activation Analysis Test Accuracies of Concept Induction, CLIP-Dissect, and GPT-4

Method	CAV		CAR	
	z-score	p-value	z-score	p-value
Concept Induction x CLIP-Dissect	0.1252	0.9004	-0.8717	0.3834
CLIP-Dissect x GPT-4	1.7494	0.0801	**1.9680**	**0.0488**
Concept Induction x GPT-4	**2.1560**	**0.0308**	1.7792	0.0751

Table 8: Concept Accuracy in Hidden Layer Activation Space of Concepts extracted using Concept Induction.

Concept Name	CAR		CAV	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
Air Conditioner	0.8994	0.8415	0.811	0.8659
Baseboard	0.875	0.8717	0.8846	0.9102
Body	0.9035	0.8857	0.8642	0.9
Building	0.9085	0.9404	0.8262	0.8690
Bushes	0.9150	0.9487	0.9477	0.9743
Car	0.9464	0.9571	0.925	0.9429
Casserole	0.9458	0.9375	0.9808	0.975
Central Reservation	0.8694	0.9	0.8917	0.9
Chain	0.9556	0.9677	0.9637	0.9677
Cistern	0.8734	0.8375	0.8449	0.8875
Coffee Table	0.9047	0.9523	0.8988	0.9166
Crapper	0.8516	0.8043	0.8571	0.8695
Cross Walk	0.9166	0.9468	0.9247	0.9361
Dishcloth	0.9055	0.9375	0.9685	0.9531
Dish Rack	0.9375	0.9583	0.9843	0.9375
Dishrag	0.8603	0.9285	0.9144	0.9464
Doorcase	0.8936	0.8611	0.8581	0.8194
Edifice	0.9487	0.9642	0.9548	0.9523
Fire Hydrant	0.9171	0.9625	0.9171	0.925
Fire Escape	0.8950	0.9146	0.9104	0.8902
Flooring	0.8841	0.9166	0.8871	0.9047
Flusher	0.8722	0.8285	0.9014	0.9285
Fluorescent Tube	0.9006	0.9625	0.9358	0.9125
Footboard	0.9268	0.9519	0.9585	0.9423
Go Cart	0.9378	0.9512	0.9254	0.9390
Jar	0.9059	0.9333	0.9572	0.9666
Left Arm	0.8549	0.8536	0.8858	0.8658
Left Foot	0.8734	0.8658	0.8703	0.8536
Letter Box	0.8901	0.8636	0.875	0.9242
Lid	0.8622	0.9047	0.8712	0.8809
Manhole	0.9349	0.8953	0.9349	0.9302
Mountain	0.9426	0.95	0.9745	0.9625
Mouth	0.8963	0.9268	0.9481	0.9512
Night Table	0.8917	0.875	0.9235	0.8875
Nuts	0.9223	0.9134	0.9417	0.9230
Open Fireplace	0.9129	0.9222	0.9101	0.9333
Ornament	0.8910	0.9375	0.9198	0.9625
Paper Towels	0.9021	0.9166	0.9239	0.9166
Pillar	0.8372	0.8837	0.7732	0.8372
Pipage	0.84239	0.7826	0.7826	0.7391
Plank	0.8719	0.9523	0.9146	0.9047
Posters	0.8806	0.9230	0.8806	0.9230
Pylon	0.8397	0.8125	0.8205	0.8375
River	0.9430	0.925	0.9399	0.925
River Water	0.9554	0.9375	0.9617	0.9375
Road	0.9221	0.9642	0.9461	0.9404
Rocker	0.8953	0.9545	0.9457	0.8939

Concept Name	CAR		CAV	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
Rocking Horse	0.9173	0.9310	0.9347	0.9655
Saucepan	0.9561	0.9827	1	0.9827
Screen Door	0.9076	0.9375	0.9235	0.925
Sculpture	0.8242	0.8333	0.8788	0.8571
Shower Stall	0.9409	0.9722	0.9652	0.9583
Sideboard	0.91	0.94	0.965	0.92
Side Rail	0.9054	0.9459	0.8986	0.9054
Skyscraper	0.9455	0.9743	0.9615	0.9743
Slipper	0.9262	0.9456	0.9617	0.9565
Slope	0.8705	0.8714	0.9208	0.8857
Soap Dish	0.8733	0.8589	0.8474	0.8589
Soap Dispenser	0.88	0.9375	0.916	0.9531
Spatula	0.9017	0.9431	0.9219	0.9204
Stem	0.8834	0.8676	0.8383	0.8382
Stretcher	0.89375	0.9375	0.9312	0.9375
Tank Lid	0.8947	0.8846	0.8848	0.8717
Tap	0.8198	0.8536	0.8354	0.8902
Teapot	0.9365	0.9411	0.9552	0.9779
Toaster	0.927	0.9714	0.9197	0.9736
Toothbrush	0.9198	0.9125	0.9198	0.9
Utensils Canister	0.9262	0.925	0.9487	0.9375
Wardrobe	0.9375	0.95	0.9188	0.9125

Table 9: Concept Accuracy in Hidden Layer Activation Space of Concepts extracted using CLIP-Dissect.

Concept Name	CAR		CAV	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
Bathroom	0.9700	0.9474	0.9400	0.9474
Bed	0.9587	0.9500	0.9437	0.9125
Bedroom	0.9167	0.9167	0.9137	0.9048
Buildings	0.9321	0.9230	0.8990	0.8974
Dallas	0.9447	0.9318	0.9750	0.9545
Dining	0.9294	0.9125	0.8907	0.9000
Dresser	0.9762	0.9625	0.9650	0.9500
File	0.9837	0.9750	0.9681	0.9500
Furnished	0.8843	0.8875	0.8762	0.8625
Highways	0.9396	0.9375	0.9679	0.9531
Interstate	0.9293	0.9268	0.8593	0.8536
Kitchen	9848	0.9743	0.9590	0.9487
Legislature	0.9149	0.9000	0.9156	0.9000
Microwave	0.9803	0.9807	0.9873	0.9807
Mississauga	0.9041	0.9054	0.9467	0.9324
Municipal	0.8679	0.8461	0.9298	0.9102
Restaurants	0.9850	0.9722	0.9692	0.9583
Road	0.9362	0.9250	0.9387	0.9250
Room	0.8653	0.8125	0.8273	0.8250
Roundtable	0.9405	0.9473	0.9136	0.8947
Valencia	0.8735	0.8625	0.8781	0.875
Street	0.9830	0.9722	0.9347	0.9167

Table 10: Concept Accuracy in Hidden Layer Activation Space of Concepts extracted using GPT-4.

Concept Name	CAR		CAV	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
Bedroom	0.9851	0.9761	0.9660	0.9523
Bathroom	0.9176	0.9024	0.9068	0.8902
Bathroom Interior	0.9273	0.9146	0.9241	0.9268
Bicycle	0.9787	0.9615	0.9887	0.9871
Cityscape	0.9438	0.9358	0.9894	0.9743
Classroom	0.8981	0.8780	0.9012	0.8536
Dining Room	0.9256	0.9125	0.8942	0.8875
Eyeglasses	0.9813	0.9883	0.9883	0.9883
Home Interior	0.8515	0.8452	0.8363	0.8214
Indoor Home Decor	0.8428	0.8333	0.8418	0.8222
Indoor Home Setting	0.6713	0.6785	0.6890	0.6666
Kitchen	0.9122	0.9302	0.9122	0.9186
Kitchen Scene	0.8562	0.8571	0.8022	0.7976
Living Room	0.8963	0.8658	0.8658	0.8414
Outdoor Scenery	0.9135	0.9024	0.9054	0.9024
Outdoor Urban Scene	0.8343	0.8170	0.7650	0.7317
Street Scene	0.8819	0.8809	0.8568	0.8690
Toilet Brush	0.9815	0.9761	0.9727	0.9642
Urban Landscape	0.8665	0.8636	0.8922	0.8863
Urban Street Scene	0.9140	0.9024	0.8757	0.8658
Urban Transportation	0.8412	0.8414	0.8251	0.8414