Evaluation of Concept Induction in Explainable AI Using Multiple Datasets

Samatha Ereshi Akkamahadevi, Abhilekha Dalal, Pascal Hitzler

Kansas State University Manhattan, KS, USA
samatha94@ksu.edu, adalal@ksu.edu, hitzler@ksu.edu

Abstract

Explainable AI aims to make artificial intelligence systems more transparent and trustworthy. A major challenge within this field involves deciphering the complex computational processes and understanding the activation patterns of hidden neurons. It was previously shown, in a single example, that an approach using formal logical reasoning in the form of so-called concept induction can meaningfully interpret hidden neuron activations in a Convolutional Neural Network (CNN) by assigning human-understandable labels. However, as this was only demonstrated for a single trained system, it is necessary to validate the approach using modified settings. In this paper, we replicate the results of the approach by employing an efficient automated pipeline which improved the processing speed and accuracy. We also extend the evaluation by selecting different training targets for the classifier. Our replication results are comparable to the previous ones and show that the approach can assign meaningful labels to individual neurons in the dense layer of a CNN, based on a statistical validation.

Introduction and Related Work

Deep learning (DL), a specialized subset of machine learning (ML), has driven significant advancements in artificial intelligence (AI), transforming various industries and domains by often surpassing human performance in several complex tasks (Narayanan et al. 2023). Its applications span a wide range of fields, including object detection (Farhadi and Redmon 2018; Dhillon and Verma 2020), speech recognition (Li 2021; Chiu et al. 2018), natural language processing (Vaswani et al. 2017; Young et al. 2018), autonomous systems (Bojarski et al. 2016), personalized recommendations (Covington, Adams, and Sargin 2016), genomics (Alipanahi et al. 2015), finance (Heaton, Polson, and Witte 2017) and healthcare (Askr et al. 2023; Ajagbe and Adigun 2024) contributing to breakthroughs in both research and industry (Alzubaidi et al. 2021).

Despite remarkable advances, the black-box nature of deep learning models poses significant challenges, particularly in high-stakes applications and safety-critical systems (Samek, Wiegand, and Müller 2017). A minor disturbance in input during training can cause deep learning models to form unintended associations, which are difficult for humans to identify due to the lack of transparency in the learning process (Ribeiro, Singh, and Guestrin 2016). This has raised concerns in sensitive fields such as healthcare (Yang, Ye, and Xia 2022), finance (Shah et al. 2024), and autonomous systems (Castelvecchi 2016), where understanding the rationale behind model predictions is crucial to prevent potentially life-threatening errors (Guidotti et al. 2018). For example, a widely reported incident involved an autonomous vehicle failing to accurately classify a pedestrian, resulting in a fatal accident (Stanton et al. 2019). Consequently, there is a growing need and expectation from users, society, and regulatory bodies that the actions and decisions made by these systems should be explainable to foster trust, accountability, and broader acceptance (Adadi and Berrada 2018; Ali et al. 2023).

In response to these concerns, Explainable AI (XAI) aims to make artificial intelligence systems more transparent and trustworthy (Embarak 2023). A major challenge in XAI involves deciphering the complex computational processes and understanding the activation patterns of hidden neurons. Various state-of-the-art techniques have been developed to enhance the interpretability and transparency of machine learning models. Key approaches include feature importance analysis, model decomposition, decision trees, Local Interpretable Model-agnostic Explanations (LIME), SHAP values, and counterfactual explanations. Despite significant progress, these methods face several challenges (Thakur, Vashisth, and Tripathi 2023; Shevskaya 2021).

Previously, an approach based on formal logical reasoning in the form of concept induction (Sarker and Hitzler 2019; Sarker et al. 2017) – borrowed from the Semantic Web field (Hitzler 2021) – was shown to be effective in assigning human-understandable labels to explain hidden neuron activations in a CNN image scene recognition scenario (Dalal et al. 2023, 2024b; Barua, Widmer, and Hitzler 2024; Dalal et al. 2024a). The approach relies on identifying activating (positive) and non-activating (negative) images for specific neurons and applying concept induction over large-scale background knowledge to discover common semantic features that activate the neuron. By leveraging a curated ontology derived from the Wikipedia concept hierarchy as background knowledge, this previous work demonstrated that the method could produce explanatory categories, such as "cross

walk" and "bushes" for neurons in the dense layer. However, the initial implementation faced limitations in terms of scalability and efficiency due to manual intervention and computational overhead in processing large datasets. In addition, it was a study based on only one trained CNN, in an image scene recognition scenario with 10 scene categories.

In this paper, we present a replication and extension of the previous approach by employing an automated pipeline (Akkamahadevi, Dalal, and Hitzler 2024) designed to streamline and improve efficiency. We applied the same methodology to newer and larger sets of scene categories to evaluate its robustness and generalizability. For the base study see primarily (Dalal et al. 2023, 2024b). As we discuss later, our new experiments led to similarly strong results as the original work. For complete results on all datasets and the top 3 solutions, please refer to `https://bit.ly/XaiReplication`

Replication Experiments

1. Dataset Selection. As in the base study, we used the ADE20k scene classification dataset (Zhou et al. 2019), which contains a diverse set of annotated images. Different sets of scene categories were selected and evaluated independently, with each set processed and analyzed separately to assess replicability of the base study.

In line with the base study, we selected the same baseline set consisting of 10 scene categories: bathrooms, bedrooms, building facades, conference rooms, dining rooms, highways, kitchens, living rooms, skyscrapers, and streets. This set, comprising 6,187 images, was previously evaluated manually, and we replicated the analysis to validate the original findings and assess whether the approach remains effective when re-trained using an automated pipeline. A second dataset consisting of 10 additional scene categories was selected: airport terminals, art studios, attics, corridors, game rooms, home offices, hotel rooms, snowy mountains, offices, and waiting rooms. This set included 983 images and was chosen to expand the evaluation by introducing new scene categories beyond those used in the baseline set. A third dataset consisting of 10 additional scene categories was selected and this set included alleys, art galleries, beaches, castles, children's rooms, closets, coasts, mountains, parks, and parlors, with a total of 676 images.

A fourth dataset, consisting of all 30 scene categories from the baseline, second, and third sets, was used. This set included a total of 7,846 images and served to evaluate the scalability and generalization of the model across a broader set of scene categories. A final dataset was constructed by extending the previous 30 scene categories with an additional 20 complex and varied scenes, resulting in a total of 50 scene categories. These additional categories included staircases, pastures, dorm rooms, nurseries, lobbies, receptions, bars, roundabouts, houses, bridges, classrooms, rivers, youth hostels, lighthouses, creeks, shoe shops, window seats, amusement parks, cockpits, and playrooms. This set contained 8,741 images and was used to evaluate the interpretability of the approach across a larger and more diverse set of scenes. Analysis for each dataset was conducted separately, and the results are presented individually.

Datasets	No. of scenes	Training accuracy	Validation accuracy
Previous study	10	87.60%	86.46%
Baseline set	10	87.56%	87.50%
Set 2	10	88.48%	75.00%
Set 3	10	90.72%	89.06%
Set 4	30	78.40%	79.05%
Set 5	50	77.99%	76.88%

Table 1: Training and Validation accuracies

2. Model Training. The images in the datasets were processed and used for training as described in the base study. All images were standardized to ensure consistent input quality for CNN training. This involved resizing the images to 224x224 pixels, normalizing pixel values, and applying data augmentation techniques. The datasets were split such that 80% of the images in each scene category were used for training, while the remaining 20% were reserved for extracting neuron activations in subsequent step. The ResNet50V2 model architecture was selected for its proven efficiency and robustness in handling complex image classification tasks. Training was performed using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy as the loss function. The model was trained for 30 epochs, with an early stopping applied to prevent overfitting. The training process was executed through an automated pipeline to ensure consistency across runs. The evaluation of the ResNet50V2 model on the ADE20K dataset demonstrated high training and validation accuracies across different sets, with minimal signs of overfitting (Table 1). Notably, the second set has the validation accuracy (75%) which is lower than the training accuracy (88.48%). However, the gap is not significant, and the slight overfitting does not compromise the overall model performance. The third set showed the best overall performance, with the highest accuracy across both training (90.72%) and validation (89.06%).

3. Neuron Activation Values and Config File Generation. Following the training of the ResNet50V2 model, the reserved 20% of images were passed through the model to extract neuron activation values from the dense layer. For each neuron, the activation values were analyzed to generate two distinct sets: positive and negative image activation sets. The positive image activation set included a list of all images that activated the neuron to at least 80% of its highest activation value among all images, while the negative image activation set included a list of all images that activated the neuron to at most 20% of its highest activation value or showed no activation. This information for each neuron was compiled into a configuration file containing the lists of positive and negative image sets, along with an Web Ontology Language (OWL) file containing machine-readable background knowledge for use in the subsequent concept induction step. Since 64 neurons were analyzed in the dense layer, a total of 64 configuration files were generated, each specifying which images strongly activated the neuron and which did not.

4. Concept Induction and Label Hypotheses Generation.

In this stage, we employed the Efficient Concept Induction Implementation (ECII) tool (Sarker and Hitzler 2019) to automatically generate semantic labels for each of the 64 neurons in the CNN's dense layer. The ECII tool utilized the config files generated in the previous step and assigned several human understandable semantic labels or concepts from the background knowledge, along with an accuracy (coverage) score for each neuron, and we termed them "target labels". Among these labels, the top three solutions were considered for the next step of label hypothesis confirmation.

5. Image Retrieval. After generating concept labels for each neuron using the ECII tool during the Concept Induction stage, we proceeded to collect image data using the PyGoogle library corresponding to the top three concept labels assigned to each neuron. For each neuron, the top three concept labels were used as search queries. For example, if neuron 0 had concept labels "building," "skyscraper," and "architecture," separate searches were conducted for each of these labels, and 100 images were downloaded for each concept label. Only images with a resolution of at least 224x224 pixels and in JPEG and PNG format were retained to ensure consistency in image quality.

Once the images were downloaded, they were split into two subsets: 80% of the images for each concept label were randomly selected and used at the label hypothesis confirmation step, and the remaining 20% of the images were assigned for validating the hypotheses by statistical analysis. This ensured that the model's performance was evaluated on previously unseen data, minimizing potential bias. Both the sets were processed through the dense layer of the trained CNN, and activation values were captured.

6. Confirmation of Label Hypotheses. For the label hypotheses confirmation step we used 80% of the downloaded images. The goal of this step was to calculate the activation values for each neuron when presented with images corresponding to its assigned concept labels (target image set) and unrelated labels (non-target image set). Each image was passed through the dense layer of the previously trained CNN, and activation values were captured for all neurons. This resulted in a table where each row represented an image and each column corresponded to the activation value of a specific neuron. For each neuron, "target activation values" were retrieved from images associated with its assigned concept labels, whereas "non-target activation values" were retrieved from images associated with the concept labels of all other neurons. Based on these values, neurons with a target value greater than or equal to 80% were selected for further analysis. This threshold ensured only neurons with consistently high activations were considered significant.

The Target% column of Table 2 represents the percentage of target images that activated each neuron, while the Non-Target% column indicates the percentage of non-target images (i.e., images corresponding to unrelated labels) that activated the neuron. For example, Neuron 37 exhibited activation for 90% of the images associated with the label "Hill". In contrast, only 47.734% of the images corresponding to unrelated labels (i.e., non-target images) activated the neuron, as shown in the Non-Target% column. Since the Target% for Neuron 37 was 90%, which is greater than the threshold of 80%, the label hypothesis for this neuron was confirmed. Following the same criterion, neurons with a Target% $\geq$ 80% were selected for further validation through statistical analysis. As a result of this process, a list of 19 confirmed labels was generated, as shown in Table 6. Furthermore, the same procedure was applied to the other two solutions for each neuron, generating corresponding lists of confirmed labels and will validate the label hypotheses for each solution through statistical analysis in the next step.

7. Validation of Label Hypotheses. After selecting neurons with target values greater than or equal to 80%, we proceeded to validate the label hypotheses using the remaining 20% (verification set images) of the downloaded images. This set of images was used to independently validate the hypotheses generated in the previous step. For each selected neuron, we used the verification set images corresponding to its assigned concept labels (target images) and unrelated labels (non-target images). These images were processed through the dense layer of the trained CNN, and the activation values were retrieved for each selected neuron. These activation values of target and non-target images were further processed and used for statistical analysis.

To assess whether the differences between target and non-target activations were statistically significant, we performed a Mann-Whitney U test (McKnight and Najab 2010) for each selected neuron using the activation values obtained from the verification set. This test was chosen because it is a non-parametric test, making it suitable for comparing two independent groups without assuming a normal distribution of activation values. For each selected neuron, the test compared the distribution of target activation values with that of non-target activation values to determine whether the neuron responded significantly more to target images than to non-target images. The null hypothesis for the test was that there was no significant difference between the target and non-target activation values, while the alternative hypothesis was that the neuron exhibited stronger activations for target images. A p-value under 0.05 was considered statistically significant, leading to the rejection of the null hypothesis. Neurons with significant p-values were deemed to have confirmed label hypotheses, meaning their assigned concept labels were strongly associated with their activations.

The results of the label hypothesis verification for Set 1, Solution 1 are summarized in Table 6. For each selected neuron, the percentage of activations for both target and non-target images, as well as the corresponding mean and median activation values and the resulting z-scores and p-values are provided. Neurons with a p-value less than 0.05 were considered to have statistically significant differences between target and non-target activations, confirming the validity of their assigned labels. For example, Neuron 37, associated with the label "Hill", exhibited activations for 90% of the target images and 40.09% of the non-target images. The Mann-Whitney U test yielded z-score of 4.07 and p-value of 5.179E-05, indicating a statistical significance between target and non-target activations ($p < 0.005$). Hence, the label hypothesis for Neuron 37 was confirmed.

Discussion

During model training, Set 1 demonstrated good generalization, likely due to its larger size and balanced categories. Although Sets 2 and 3 had smaller datasets, only Set 2 showed slight overfitting, attributed to higher variability and diverse scene categories, including visually overlapping indoor scenes. In contrast, Set 3 achieved better performance due to clearer and more distinct visual patterns (e.g., beaches, mountains, parks). These results indicate that data complexity and class variability can impact model performance more significantly than dataset size alone (Ghosh et al. 2021). Also, Set 5 demonstrates that while increasing the dataset size helps mitigate overfitting, the classification task becomes inherently more complex with a larger number of categories, leading to slightly lower overall accuracy.

Regarding the replication results for the explainability task, we observe overall that results are comparable to the original study. Tables 3 to 5 and 7 to 9 show the corresponding results. Detailed results for dataset 3 have been omitted due to lack of space but can be found in the appendix.[1] In the original study (Dalal et al. 2024b), the correctness of 19 target labels for dense layer neurons was statistically verified. Our replications provided 15 confimed labels for dataset 1 (i.e., the same dataset as in the original study), 17 for dataset 2, 15 for dataset 3, 19 for dataset 4, and 7 for dataset 5. While the number of confirmed labels for the larger dataset 5 is lower (see discussion further below), our experiments broadly confirm the findings.

We discuss some of the results in more detail: Despite replicating the same pipeline with the same dataset and model architecture, we observed differences in the concept labels generated for certain neurons. For example, Neuron 22, which was previously labeled as "Skyscraper," was assigned the label "Bus and Autobus" in our study (Table 1). This variation can be attributed to the inherent randomness in the CNN training processes (e.g., weight initialization, data augmentation, and batch shuffling) and the sensitivity of concept induction to minor differences in neuron activation patterns. Such differences highlight the importance of evaluating the robustness of XAI methods under slightly varying conditions, as even small changes can influence the interpretability outcomes.

During the analysis of neuron activations, distinct differences in neuron specificity were observed. As shown in Table 3, Neuron 8 (Mountain) exhibited a high target activation percentage (98.75%) alongside a low non-target activation percentage (23.045%). This indicates that the neuron primarily responds to images associated with its assigned concept, demonstrating high specificity and reliability. Conversely, Neuron 13 (Counter and Bulletin Board) showed high activation percentages for both target (97.5%) and non-target images (96.909%), suggesting that it responds to general features shared across multiple categories, thereby lacking specificity. These findings emphasize that neurons with high target activation but low non-target activation are more suitable for representing unique concepts, while neurons displaying high activations for both target and non-target images may be less reliable in distinguishing their assigned concept. Evaluating both target and non-target activation percentages is thus crucial for validating neuron-specific concept labels.

Notably, during neuron verification, Neuron 13, which was selected for verification based on its high target activation percentage, was ultimately excluded due to its high non-target activation percentage. This resulted in a low z-score and a non-significant p-value, indicating poor specificity for its assigned concept. In contrast, Neuron 3 (Mountain), with consistently low non-target activation during verification, produced a high z-score and statistically significant p-value, confirming its specificity for the "Mountain" concept. These results demonstrate that high target activation alone is insufficient for reliable concept labeling. The verification step is essential to ensure that only neurons with distinct and consistent responses to their assigned labels are validated. This filtering process enhances the robustness of the concept induction approach by excluding neurons that respond to general features across multiple categories, thereby improving the precision of neuron-specific concept representations. These observations holds true for other neurons also.

During neuron evaluation, we observed that Set 5, despite including all categories from Set 4 along with 20 additional complex scene categories, resulted in significantly fewer neurons being selected for verification (only 8, compared to 19 in Set 4). This can be attributed to increased complexity, higher intra-class variability (Yu et al. 2023), feature overlap (Ghosh et al. 2021), and fewer images available for the newly added categories (Rangel et al. 2024). The limited image count likely led to insufficient feature training and inconsistent neuron activations (Huesmann et al. 2021).

Additionally, the larger number of categories may have likely caused neurons to activate for multiple categories, reducing the number of neurons that met the criteria for selection during evaluation. Notably, 7 out of 8 neurons selected in Set 5 were successfully validated, indicating that despite fewer selections, the neurons that got validated were highly specific. Moreover, while the Mann-Whitney U test is generally suitable for comparing distinct distributions, its effectiveness may be reduced in scenarios with small sample sizes and overlapping activations, making it harder to detect significant differences. These findings underscore the importance of balancing dataset size, class representation, and model capacity, as well as considering alternative statistical analysis or threshold adjustments to enhance the robustness of concept induction in complex datasets.

Conclusion

Our study confirmed that the concept induction method can reliably produce meaningful labels for neurons, making it a valuable tool in XAI. We successfully replicated the original findings using an automated pipeline and demonstrated its robustness across different datasets. While larger and more complex datasets posed challenges, the approach proved effective in identifying highly specific neurons, effectively validating the original study.

[1]Detailed results can be found in the appendix, available at `https://bit.ly/XaiReplication`

Neuron Id	Label(s)	Images	Coverage	Target%	Non-Target%
2	Fence	80	0.965	50.750	51.923
5	Sideboard and Counter	80	0.976	40.468	45.060
8	Chain and Chandelier	80	0.997	**87.500**	64.270
9	Skyscraper	80	0.960	**97.500**	62.812
14	Knife_set and Toaster	80	0.976	**83.750**	71.145
17	Head and Right_arm	80	0.954	**95.000**	82.031
18	Bench and Night_table	80	0.983	**100.000**	80.911
21	Frying_pan and Ornament	80	0.964	1.250	7.734
25	Night_table	80	0.976	**85.000**	73.125
30	Pillow and Desk_lamp	80	0.948	**98.750**	64.322
31	Dormer	80	0.910	**82.500**	17.760
33	Washing_machine	80	0.954	75.000	61.979
37	Hill	80	0.958	**90.000**	47.734
39	Wheel	80	0.838	55.000	33.151
43	Paper_towels and Mopboard	80	0.985	66.250	40.078
44	Mountain and Bushes	80	0.964	**87.500**	39.244
55	Lid and Toilet_paper	80	0.993	**92.500**	45.104
57	Dishrag	80	0.974	20.000	35.494
58	Left_hand	80	0.976	8.750	7.161
60	Seat_cushion and Seat_base	80	0.947	**82.500**	60.104

Table 2: Representative data from dataset 1, solution 1, showing evaluation results. The table lists neurons, their assigned labels, and the percentage of target and non-target images that activated each neuron. For example, Neuron 37, associated with the label "Hill", was activated by 90% of target images and 47.734% of non-target images. Neurons with a target activation percentage greater than or equal to 80% (highlighted in bold) were selected for further hypothesis validation. A total of 19 confirmed labels were identified for this purpose.

Neuron Id	Label(s)	Images	Coverage	Target%	Non-Target%
6	Hassock and Puff	80	0.954	24.500	26.544
8	Mountain	80	0.800	**98.750**	23.045
13	Counter and Bulletin_board	80	0.912	**97.500**	96.909
15	Mountain	80	0.981	**100.000**	75.522
18	Keyboard and Computer	80	0.933	76.250	71.818
19	Billiard_table and Corner_pocket	80	0.902	**90.000**	54.340
26	Base and Base	80	0.965	**82.500**	65.568
27	Suitcase and Right_hand	80	0.963	71.250	67.25
29	Night_table and Pillow	80	0.980	48.750	41.977
31	Paddle and Television	80	1.000	**96.250**	86.795
43	Chairs	80	0.904	76.250	34.863
44	Data_processor and Computer	80	0.885	73.750	52.113
46	Flush_mount_light	80	0.882	52.500	37.931
48	Mountain	80	1.000	**96.250**	47.568
50	Seat_base and Seat_cushion	80	1.000	**85.000**	80.068
51	Sky and Trees	80	0.861	52.500	48.704
53	Easel	80	1.000	**83.750**	42.977
56	Mousepad and Speaker	80	1.000	0.000	0.4772
57	Ash-bin and Ventilation_shaft	80	0.942	0.000	29.25
63	Light_troffer and Poster	80	0.960	**92.500**	71.090

Table 3: Representative data from dataset 2, solution 1, showing evaluation results. The table lists neurons, their assigned labels, and the percentage of target and non-target images that activated each neuron. For example, Neuron 8, associated with the label "Mountain", was activated by 98.75% of target images and 23.045% of non-target images. Neurons with a target activation percentage greater than or equal to 80% (highlighted in bold) were and selected for further hypothesis validation. A total of 23 confirmed labels were identified for this purpose.

Neuron Id	Label(s)	Images	Coverage	Target%	Non-Target%
2	Washing_machine and Pepper	80	0.986	39.583	36.666
4	Skyscraper	80	0.929	25.781	29.021
8	Span and River_water	80	0.983	8.750	36.580
11	Rock	80	0.926	66.250	24.150
12	Mountain	80	0.940	**100.000**	31.273
15	Computer_case and Hair	80	0.991	37.500	47.570
19	Computer_case and Mouse_mat	80	0.995	77.500	57.169
21	Saucepan and Glass	80	0.993	**88.750**	60.801
24	Papers and Written_document	80	0.973	**98.750**	66.155
26	Field	80	0.990	**100.000**	72.311
28	Motorcoach and Bus	80	0.888	**96.250**	55.731
30	Pillow	80	0.963	**97.500**	62.264
31	Housing_lamp and Bus	80	0.995	58.750	57.735
33	Bathrobe and Bathrobe	80	0.952	**91.250**	47.971
34	Grip and Suitcase	80	0.913	42.500	61.179
36	Sky and Door	80	0.862	0.000	5.8490
39	Clock and Headboard	80	0.995	**90.000**	62.735
42	Skirting_board and Baseboard	80	0.944	**82.500**	45.353
50	Motorcoach and Rim	80	0.975	**81.250**	55.424
63	Slope and Truck	80	0.992	80.000	47.264

Table 4: Representative data from dataset 4, solution 1, showing evaluation results. The table lists neurons, their assigned labels, and the percentage of target and non-target images that activated each neuron. For example, Neuron 12, associated with the label "Mountain", was activated by 100% of target images and 31.273% of non-target images. Neurons with a target activation percentage greater than or equal to 80% (highlighted in bold) were and selected for further hypothesis validation. A total of 19 confirmed labels were identified for this purpose.

Neuron Id	Label(s)	Images	Coverage	Target%	Non-Target%
4	Bushes and Lighthouse	80	0.993	68.409	60.475
6	Lighthouse and Beacon_light	80	0.955	23.125	22.521
15	Power_pylon	80	0.944	**96.250**	56.520
23	Computer and Computer_case	80	0.978	77.500	47.583
24	Fauna	80	0.936	33.750	32.416
26	Fence and Sidewalk	80	1.000	31.250	38.000
28	Vale and Creature	80	0.947	**92.500**	74.479
30	Plates	80	0.964	28.750	32.104
31	Teacup and Buffet	80	0.982	**86.250**	48.083
33	Sea	80	0.950	**91.250**	34.145
38	Mountain_pass and Mountain_pass	80	0.966	**91.250**	60.270
39	Lock and Tapestry	80	0.984	71.250	42.479
40	Dishcloth and Soap_bottle	80	0.992	12.500	20.020
44	Rock and Mortal	80	1.000	5.000	1.000
46	Jacket and Apparel	80	0.983	**85.000**	46.333
48	Conveyor_belt and Transporter	80	0.948	11.250	15.958
52	Housing_lamp and Bus	80	0.985	**81.250**	69.166
54	Skyscraper	80	0.939	**98.750**	64.291
55	Stones	80	0.972	48.750	29.229
63	Bench and Ottoman	80	0.997	18.750	27.291

Table 5: Representative data from dataset 5, solution 1, showing evaluation results. The table lists neurons, their assigned labels, and the percentage of target and non-target images that activated each neuron. For example, Neuron 33, associated with the label "Sea", was activated by 91.25% of target images and 34.145% of non-target images. Neurons with a target activation percentage greater than or equal to 80% (highlighted in bold) were and selected for further hypothesis validation. A total of 8 confirmed labels were identified for this purpose.

Neuron Id	Label(s)	Images	Activations(%)		Mean		Median		z-score	p-value
			targ	non-t	targ	non-t	targ	non-t		
8	Chain and Chandelier	20	90.00	62.91	2.74	0.66	2.54	1.19	3.76	**0.000114188**
9	Skyscraper	20	95.00	62.08	4.56	0.77	4.03	1.47	4.81	**7.82917E-07**
13	Skyscraper	20	90.00	33.12	1.88	0.00	1.93	0.49	4.96	**4.5555E-09**
14	Knife_set and Toaster	20	75.00	69.27	0.68	0.97	1.44	1.45	-0.00	0.99224574
15	Flusher and Spigot	20	100.00	60.83	3.39	0.58	3.40	1.17	5.43	**2.15761E-08**
16	Button_panel and Oven	20	70.00	44.16	1.79	0.00	1.56	0.89	2.40	**0.008382643**
17	Head and Right_arm	20	95.00	80.10	2.97	2.34	3.12	2.86	0.62	0.529383471
18	Bench and Night_table	20	95.00	79.58	4.34	2.17	4.13	2.46	2.92	**0.003340697**
22	Bus and Autobus	20	90.00	68.95	2.18	0.86	2.28	1.47	2.76	**0.005037085**
25	Night_table	20	85.00	71.87	2.76	1.42	2.77	1.78	2.14	**0.030105389**
30	Pillow and Desk_lamp	20	90.00	61.77	3.73	0.64	3.62	1.32	4.76	**9.67604E-07**
31	Dormer	20	90.00	18.43	2.69	0.00	2.36	0.24	6.13	**1.36419E-18**
34	Shower and Crapper	20	85.00	80.20	1.81	1.86	1.80	2.14	-0.39	0.695491568
37	Hill	20	80.00	50.52	2.46	0.01	2.66	0.86	4.07	**1.4498E-05**
44	Mountain and Bushes	20	90.00	39.58	3.76	0.00	3.60	0.57	5.60	**2.86221E-10**
46	Air_conditioning and Desk_lamp	20	100.00	89.37	2.13	2.45	2.15	2.64	-1.05	0.293070236
55	Lid and Toilet_paper	20	85.00	45.20	2.85	0.00	2.78	0.72	4.38	**1.78707E-06**
60	Seat_cushion and Seat_base	20	70.00	59.68	2.17	0.55	2.87	1.25	2.35	**0.015119106**
61	Oven	20	100.00	73.75	5.23	1.31	4.93	1.84	5.87	**3.07996E-09**

Table 6: Verification results for dataset 1, solution 1. The table lists neurons, their assigned labels, and the percentage of target and non-target images that activated each neuron. Additionally, it includes the mean and median activation values for both target and non-target images, along with the corresponding z-scores and p-values obtained from the Mann-Whitney U test. For example, Neuron 37, associated with the label "Hill", was activated by 80% of target images and 50.52% of non-target images, with a mean activation of 2.46 for target images and 0.01 for non-target images. The Mann-Whitney U test yielded a z-score of 4.07 and p-value < 0.05, confirming the label hypothesis for this neuron. In total, the null hypothesis was rejected for 15 out of 19 neurons, indicating that their activations were significantly different for target and non-target images, thus confirming the corresponding label hypotheses.

Neuron Id	Label(s)	Images	Activations(%)		Mean		Median		z-score	p-value
			targ	non-t	targ	non-t	targ	non-t		
8	Mountain	20	100	22.18	5.61	0.00	5.79	0.96	6.50	**2.19996E-18**
10	Mountain	20	100	83.81	3.93	2.02	4.06	2.15	4.95	**6.77973E-07**
13	Counter and Bulletin_board	20	100	97.00	3.23	3.50	3.39	3.52	-0.41	0.67631451
14	Telephone_set and Bed	20	75	43.09	0.85	0.00	1.59	0.69	2.97	**0.001008803**
15	Mountain	20	100	76.27	3.84	1.03	3.70	1.45	5.90	**2.74598E-09**
19	Billiard_table and Corner_pocket	20	100	56.45	3.06	0.36	3.62	1.11	5.61	**4.90011E-09**
20	Mountain	20	100	63.54	2.82	0.53	2.82	1.04	5.61	**8.79696E-09**
26	Base and Base	20	75	68.81	0.93	0.87	1.29	1.22	0.40	0.680186327
28	Night_table	20	95	51.00	2.16	0.05	2.56	0.90	4.95	**1.47888E-07**
30	Pool_ball and Side_pocket	20	85	48.63	2.19	0.00	2.88	0.63	4.57	**9.50893E-07**
31	Paddle and Television	20	100	87.90	3.18	1.70	2.95	2.06	2.95	**0.003092391**
35	Mountain	20	100	58.54	5.43	0.39	5.44	1.45	6.32	**5.90916E-11**
36	Bed	20	100	80.54	4.45	1.24	4.13	1.51	6.18	**5.41529E-10**
38	Mountain	20	100	52.09	5.70	0.11	5.57	1.35	6.38	**1.56904E-11**
39	Mountain	20	100	77.54	5.00	1.27	5.09	1.80	6.09	**9.0411E-10**
47	Tree	20	95	84.27	2.37	1.94	2.48	2.03	1.53	0.123485702
48	Mountain	20	95	47.72	3.86	0.00	3.69	1.00	5.81	**3.84833E-10**
50	Seat_base and Seat_cushion	20	70	78.72	0.97	1.18	0.89	1.43	-1.71	0.085568449
53	Easel	20	85	42.72	0.96	0.00	1.04	0.46	3.73	**3.74019E-05**
54	Mountain	20	100	90.72	5.27	2.00	5.20	2.41	5.94	**2.68196E-09**
55	Sky	20	100	83.63	2.34	1.85	2.21	2.14	0.86	0.385680246
61	Mountain	20	100	40.90	3.34	0.00	3.43	0.79	6.29	**2.16575E-12**
63	Light_troffer and Poster	20	95	71.27	1.96	1.52	2.06	1.81	1.27	0.196252732

Table 7: Verification results for dataset 2, solution 1. For example, Neuron 8, associated with the label "Mountain", was activated by 100% of target images and 22.18% of non-target images, with a mean activation of 5.61 for target images and 0.00 for non-target images. The Mann-Whitney U test yielded a z-score of 6.50 and p-value < 0.05, confirming the label hypothesis for this neuron. The null hypothesis was rejected for 17 out of 23 neurons, indicating that their activations were significantly different for target and non-target images, thus confirming the corresponding label hypotheses.

Neuron Id	Label(s)	Images	Activations(%)		Mean		Median		z-score	p-value
			targ	non-t	targ	non-t	targ	non-t		
9	Tap and Crapper	20	100	65.75	3.72	0.91	3.73	1.54	4.67	**1.84907E-06**
12	Mountain	20	100	31.88	6.33	0.00	5.94	0.67	7.42	**7.85117E-19**
14	Dial and Microwave_oven	20	100	50.09	2.98	0.00	2.95	0.84	5.40	**8.57287E-09**
16	Mountain and Rock	20	85	52.92	3.35	0.11	3.03	0.82	4.43	**2.88646E-06**
17	Flyscreen and Flyscreen	20	95	51.22	2.08	0.11	2.32	1.12	4.00	**2.21041E-05**
21	Saucepan and Glass	20	90	60.75	3.55	0.54	3.49	1.19	4.50	**3.42252E-06**
24	Papers and Written_document	20	100	65.09	2.02	0.87	2.04	1.39	2.81	**0.004096266**
26	Field	20	100	73.39	5.03	1.36	4.79	1.84	5.34	**7.0909E-08**
28	Motorcoach and Bus	20	100	56.60	3.91	0.36	3.89	1.36	5.43	**1.49413E-08**
30	Pillow	20	100	60.66	5.97	0.54	5.52	1.29	6.52	**1.77967E-11**
33	Bathrobe and Bathrobe	20	90	47.07	1.55	0.00	1.84	0.91	3.81	**3.74309E-05**
35	Fog_bank and Mountains	20	95	72.83	4.99	1.66	4.89	2.12	4.94	**6.08394E-07**
37	Sea	20	85	58.49	6.31	0.58	5.90	1.73	4.57	**2.17573E-06**
38	Sand_beach	20	95	54.24	6.06	0.26	5.50	1.00	6.77	**1.20758E-12**
39	Clock and Headboard	20	90	62.07	4.75	0.68	4.31	1.25	4.47	**4.32903E-06**
42	Skirting_board and Baseboard	20	90	48.77	2.25	0.00	2.02	0.89	4.27	**4.58383E-06**
43	Shop_window	20	95	67.26	2.95	1.35	3.38	1.82	3.56	**0.000282876**
50	Motorcoach and Rim	20	70	55.66	1.37	0.41	2.35	1.15	2.18	**0.022226052**
57	Keyboard and Monitor	20	100	53.86	3.82	0.20	3.55	1.01	5.46	**9.33005E-09**

Table 8: Verification results for dataset 4, solution 1. For example, Neuron 57, associated with the label "Keyboard and Monitor", was activated by 100% of target images and 53.86% of non-target images, with a mean activation of 3.82 for target images and 0.20 for non-target images. The Mann-Whitney U test yielded a z-score of 5.46 and a p-value < 0.05, confirming the label hypothesis for this neuron. In total, the null hypothesis was rejected for 19 out of 19 neurons, indicating that their activations were significantly different for target and non-target images, thus confirming the corresponding label hypotheses.

Neuron Id	Label(s)	Images	Activations(%)		Mean		Median		z-score	p-value
			targ	non-t	targ	non-t	targ	non-t		
15	Power_pylon	20	95.00	54.33	2.39	0.34	2.43	1.55	3.04	**0.001425796**
28	Vale and Creature	20	100.00	73.91	4.30	2.07	4.07	2.46	3.44	**0.0005045**
31	Teacup and Buffet	20	85.00	45.41	2.38	0.00	2.50	0.96	3.78	**3.6779E-05**
33	Sea	20	75.00	33.25	4.82	0.00	3.91	0.75	4.32	**2.9295E-07**
38	Mountain_pass and Mountain_pass	20	85.00	61.08	3.73	0.76	3.67	1.56	3.07	**0.00157004**
46	Jacket and Apparel	20	75.00	46.58	2.35	0.00	2.04	1.14	2.71	**0.003227966**
52	Housing_lamp and Bus	20	85.00	68.08	1.30	1.30	2.39	2.02	1.08	0.269217753
54	Skyscraper	20	100.00	63.58	3.94	0.95	3.96	1.58	4.21	**1.6138E-05**

Table 9: Verification results for dataset 5, solution 1. For example, Neuron 54, associated with the label "Skyscraper", was activated by 100% of target images and 63.84% of non-target images, with a mean activation of 3.94 for target images and 0.95 for non-target images. The Mann-Whitney U test yielded a z-score of 4.21 and a p-value < 0.05, confirming the label hypothesis for this neuron. In total, the null hypothesis was rejected for 7 out of 8 neurons, indicating that their activations were significantly different for target and non-target images, thus confirming the corresponding label hypotheses.

References

Adadi, A.; and Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.

Ajagbe, S. A.; and Adigun, M. O. 2024. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. *Multimedia Tools and Applications*, 83(2): 5893–5927.

Akkamahadevi, S. E.; Dalal, A.; and Hitzler, P. 2024. Automating CNN Neuron Interpretation using Concept Induction. In Etcheverry, L.; Garcia, V. L.; Osborne, F.; and Pernisch, R., eds., *Proceedings of the ISWC 2024 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 23nd International Semantic Web Conference (ISWC 2024), Hanover, Maryland, USA, November 11-15, 2024*, volume 3828 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99: 101805.

Alipanahi, B.; Delong, A.; Weirauch, M. T.; and Frey, B. J. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8): 831–838. Publisher: Nature Publishing Group.

Alzubaidi, L.; Zhang, J.; Humaidi, A. J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M. A.; Al-Amidie, M.; and Farhan, L. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1): 53.

Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y. A. M. M.; Gomaa, M. M.; and Hassanien, A. E. 2023. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56(7): 5975–6037.

Barua, A.; Widmer, C.; and Hitzler, P. 2024. Concept Induction Using LLMs: A User Experiment for Assessment. In Besold, T. R.; d'Avila Garcez, A.; Jimenez-Ruiz, E.; Confalonieri, R.; Madhyastha, P.; and Wagner, B., eds., *Neural-Symbolic Learning and Reasoning*, 132–148. Cham: Springer Nature Switzerland. ISBN 978-3-031-71170-1.

Bojarski, M.; Testa, D. D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; Zhang, X.; Zhao, J.; and Zieba, K. 2016. End to End Learning for Self-Driving Cars. *CoRR*, abs/1604.07316.

Castelvecchi, D. 2016. Can we open the black box of AI? *Nature News*, 538(7623): 20. Cg_type: Nature News Section: News Feature.

Chiu, C.-C.; Sainath, T. N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R. J.; Rao, K.; Gonina, E.; Jaitly, N.; Li, B.; Chorowski, J.; and Bacchiani, M. 2018. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4774–4778. ISSN: 2379-190X.

Covington, P.; Adams, J.; and Sargin, E. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, 191–198. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4035-9.

Dalal, A.; Rayan, R.; Barua, A.; Akkamahadevi, S. E.; Das, A.; Widmer, C.; Vasserman, E. Y.; Sarker, M. K.; and Hitzler, P. 2024a. Towards a Neurosymbolic Understanding of Hidden Neuron Activations. Under review, available from https://neurosymbolic-ai-journal.com/paper/towards-neurosymbolic-understanding-hidden-neuron-activations.

Dalal, A.; Rayan, R.; Barua, A.; Vasserman, E. Y.; Sarker, M. K.; and Hitzler, P. 2024b. On the Value of Labeled Data and Symbolic Methods for Hidden Neuron Activation Analysis. In Besold, T. R.; d'Avila Garcez, A.; Jimenez-Ruiz, E.; Confalonieri, R.; Madhyastha, P.; and Wagner, B., eds., *Neural-Symbolic Learning and Reasoning*, 109–131. Cham: Springer Nature Switzerland. ISBN 978-3-031-71170-1.

Dalal, A.; Sarker, M. K.; Barua, A.; Vasserman, E. Y.; and Hitzler, P. 2023. Understanding CNN Hidden Neuron Activations Using Structured Background Knowledge and Deductive Reasoning. *arXiv*, abs/2308.03999.

Dhillon, A.; and Verma, G. K. 2020. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, 9(2): 85–112.

Embarak, O. 2023. Decoding the Black Box: A Comprehensive Review of Explainable Artificial Intelligence. In *2023 9th International Conference on Information Technology Trends (ITT)*, 108–113.

Farhadi, A.; and Redmon, J. 2018. Yolov3: An incremental improvement. In *Computer vision and pattern recognition*, volume 1804, 1–6. Springer Berlin/Heidelberg, Germany.

Ghosh, K.; Bellinger, C.; Corizzo, R.; Krawczyk, B.; and Japkowicz, N. 2021. On the combined effect of class imbalance and concept complexity in deep learning. In *2021 IEEE international conference on big data (big data)*, 4859–4868. IEEE.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5): 93:1–93:42.

Heaton, J. B.; Polson, N. G.; and Witte, J. H. 2017. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1): 3–12. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asmb.2209.

Hitzler, P. 2021. A review of the semantic web field. *Commun. ACM*, 64(2): 76–83.

Huesmann, K.; Rodriguez, L. G.; Linsen, L.; and Risse, B. 2021. The Impact of Activation Sparsity on Overfitting in Convolutional Neural Networks. In Del Bimbo, A.; Cucchiara, R.; Sclaroff, S.; Farinella, G. M.; Mei, T.; Bertini, M.; Escalante, H. J.; and Vezzani, R., eds., *Pattern Recognition. ICPR International Workshops and Challenges*, volume 12663, 130–145. Cham: Springer International Publishing. ISBN 978-3-030-68795-3 978-3-030-68796-0. Series Title: Lecture Notes in Computer Science.

Li, J. 2021. Recent Advances in End-to-End Automatic Speech Recognition. *CoRR*, abs/2111.01690.

McKnight, P. E.; and Najab, J. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology*, 1–1.

Narayanan, V.; Cao, Y.; Panda, P.; Reddy Challapalle, N.; Du, X.; Kim, Y.; Krishnan, G.; Lee, C.; Li, Y.; Sun, J.; Venkatesha, Y.; Wang, Z.; and Zheng, Y. 2023. Overview of Recent Advancements in Deep Learning and Artificial Intelligence. In *Advances in Electromagnetics Empowered by Artificial Intelligence and Deep Learning*, 23–79. John Wiley & Sons, Ltd. ISBN 978-1-119-85392-3. Section: 2 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119853923.ch2.

Rangel, G.; Cuevas-Tello, J. C.; Nunez-Varela, J.; Puente, C.; and Silva-Trujillo, A. G. 2024. A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks. *Journal of Sensors*, 2024(1): 2797320.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, 1135–1144. San Francisco California USA: ACM. ISBN 978-1-4503-4232-2.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ArXiv*, abs/1708.08296.

Sarker, M. K.; and Hitzler, P. 2019. Efficient Concept Induction for Description Logics. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 3036–3043. AAAI Press.

Sarker, M. K.; Xie, N.; Doran, D.; Raymer, M. L.; and Hitzler, P. 2017. Explaining Trained Neural Networks with Semantic Web Technologies: First Steps. In Besold, T. R.; d'Avila Garcez, A. S.; and Noble, I., eds., *Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy)*, volume 2003 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Shah, T.; Shekokar, K.; Barve, A.; and Khandare, P. 2024. An Analytical Review: Explainable AI for Decision Making in Finance Using Machine Learning. In *2024 Parul International Conference on Engineering and Technology (PICET)*, 1–5. IEEE.

Shevskaya, N. V. 2021. Explainable artificial intelligence approaches: Challenges and perspectives. In *2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, 540–543. IEEE.

Stanton, N. A.; Salmon, P. M.; Walker, G. H.; and Stanton, M. 2019. Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian. *Safety Science*, 120: 117–128. Publisher: Elsevier.

Thakur, A.; Vashisth, R.; and Tripathi, S. 2023. Explainable Artificial Intelligence: A Study of Current State-of-the-Art Techniques for Making ML Models Interpretable and Transparent. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 111–115. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Yang, G.; Ye, Q.; and Xia, J. 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77: 29–52. Publisher: Elsevier.

Young, T.; Hazarika, D.; Poria, S.; and Cambria, E. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3): 55–75. Publisher: ieee.

Yu, L.; Hu, T.; Hong, L.; Liu, Z.; Weller, A.; and Liu, W. 2023. Continual Learning by Modeling Intra-Class Variation. *Trans. Mach. Learn. Res.*, 2023.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.