A Case Study on Concept Induction for Neuron-Level Interpretability in CNN

Moumita Sen Sarma[1,*], Samatha Ereshi Akkamahadevi[1] and Pascal Hitzler[1,*]

[1]Department of Computer Science, Kansas State University, Manhattan, Kansas, USA

Abstract

Deep Neural Networks (DNNs) have advanced applications in domains such as healthcare, autonomous systems, and scene understanding, yet the internal semantics of their hidden neurons remain poorly understood. Prior work introduced a Concept Induction–based framework for hidden neuron analysis and demonstrated its effectiveness on the ADE20K dataset [1]. In this case study, we investigate whether the approach generalizes by applying it to the SUN2012 dataset [2], a large-scale scene recognition benchmark. Using the same workflow, we assign interpretable semantic labels to neurons and validate them through web-sourced images and statistical testing. Our findings confirm that the method transfers to SUN2012, showing its broader applicability.

Keywords

Explainable AI (XAI), Concept Induction, Hidden Neuron Analysis, CNNs

1. Introduction

Deep Neural Networks (DNNs), particularly Convolutional Neural Networks (CNNs), have achieved state-of-the-art performance in image classification and scene understanding. Yet, the hidden neurons remain opaque, limiting interpretability in domains where transparency is critical [3]. Explainable AI (XAI) techniques such as saliency maps and attribution methods (e.g., SHAP [4], LIME [5]) highlight input contributions but rarely capture what individual neurons semantically represent. Concept Induction [6] offers a neurosymbolic alternative by mapping neuron activations to high-level concepts grounded in knowledge graphs, which organize concepts into hierarchies of entities and relations. By contrasting positive and negative activation sets with a structured ontology, it induces logical class expressions for each neuron that yield candidate semantic labels. In previous work [1], this approach was applied to the ADE20K dataset [7] with the automation pipeline in [8], demonstrating that neurons could be systematically assigned interpretable labels with strong empirical support. Herein, we apply the same methodology to a different large-scale benchmark, SUN2012 [2]. Our objective is to assess whether the findings extend beyond ADE20K. By validating the approach on SUN2012, we indeed show that the method transfers and that robust neuron–concept associations emerge consistently across datasets.

2. Methodology

The hidden neuron analysis follows a structured workflow that connects hidden neuron activations to human-understandable concepts. We follow the approach from [1], consisting of data preparation, model training, activation extraction, concept induction, and evaluation of the induced concepts [1].

Data Selection & Preparation: SUN2012 contains 131,000 images across 908 scene categories with annotations for over 3,800 objects. For this study, the ten largest categories are selected: bathroom, bedroom, building facade, dining room, highway, kitchen, living room, mountain snowy, skyscraper, and street, yielding 3,157 images for training and validation and 793 for testing.

Model Training: As in [1], multiple CNN architectures (VGG16/19, InceptionV3,

K-CAP 2025 Companion Volume

*Corresponding author.

moumita@ksu.edu (M. S. Sarma); samatha94@ksu.edu (S. E. Akkamahadevi); hitzler@ksu.edu (P. Hitzler)

0009-0003-6644-6531 (M. S. Sarma); 0009-0001-6333-8004 (S. E. Akkamahadevi); 0000-0001-6192-3472 (P. Hitzler)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]Supporting data.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-10-26

CEUR-WS.org/Vol-4303/pdt-paper88.pdf

ResNet50/101/152/50V2) are fine-tuned. The models were trained using their standard input resolutions: 299 × 299 for InceptionV3 and 224 × 224 for the remaining architectures. Training was carried out with the Adam optimizer set to a learning rate of 0.001, using categorical cross-entropy as the loss function. A batch size of 32 was used, and training ran for up to 30 epochs. To prevent overfitting, early stopping was applied by monitoring the validation loss with a patience of three epochs, restoring the best-performing model weights. While the prior study on ADE20K reported ResNet50V2 as the best-performing model, in our experiments on SUN2012, InceptionV3 achieves the highest performance (96.83% training accuracy; 92.71% validation accuracy; 93.95% test accuracy) and is therefore selected for further analysis.

Neuron Activation Extraction: Following the process introduced in previous work [1], neuron activations are extracted from the dense layer of the trained network. Each test image is passed through the network, and activation values of all 64 dense-layer neurons are collected. To capture activation behavior, thresholds are defined relative to the maximum observed response: images at or above 80% form the positive set, while those at or below 20% form the negative set. These contrasting sets are then used in the Concept Induction stage to derive the semantic labels.

Concept Induction: The Efficient Concept Induction and Integration (ECII) system [9] is employed to derive concepts from the neuron activation sets. For every image, a minimal ontology is constructed by considering only the annotated objects and mapping them to exact lexical matches within a Wikipedia-based concept hierarchy [10]. These image-specific ontologies are then integrated with the Wikipedia hierarchy through ECII to form a single, robust background ontology. ECII then made use of this background knowledge to generate logical class expressions evaluated using coverage score that distinguishes positive from negative sets and that is defined by

$$\text{coverage}(E) = \frac{|Z_1| + |Z_2|}{|P \cup N|},$$

where $Z_1 = \{\, p \in P \mid K \models E(p) \,\}$ and $Z_2 = \{\, n \in N \mid K \not\models E(n) \,\}$, P and N refer to positive set and negative set, and K represents the knowledge base. Coverage measures how well the induced concept aligns with a neuron's activation pattern, with higher scores indicating a better match. This stage resulted in generating semantic labels (e.g., snowy mountain, toilet tissue, crosswalk) for neurons. The automated system [8] generates semantic label for 64 neurons in approximately 20 minutes through parallel execution.

Concept Evaluation: We adopt the same evaluation procedure as in the ADE20K study, combining web-sourced image confirmation with statistical testing. For each neuron label, up to 100 Google Images were retrieved and divided into two disjoint subsets: 80% of the successfully processed images were used for label confirmation, while the remaining 20% were reserved for statistical testing. An image was considered to reliably activate neuron j when its activation was at least 80% of the maximum activation observed for that neuron. Target Label Activation (TLA) is the percentage of images retrieved using the target label that activate the corresponding neuron above this threshold. A neuron label was confirmed when TLA was at least 80%. Non-Target Label Activation (Non-TLA) is the percentage of images retrieved using the other semantic labels that activate the same neuron above its threshold. Images from the other labels were pooled to calculate a single Non-TLA value for each neuron. For statistical validation, a Mann–Whitney U test was performed using the held-out 20% of images to compare the neuron's activation for target-label images against its activation for non-target-label images. Here, $p < 0.05$ with a negative z-score means statistical significance, which demonstrates that target images consistently trigger stronger activations than non-target images. This results in rejecting the null hypothesis, which states that there is no significant difference between activations for target images (retrieved with the neuron's label) and non-target images (retrieved with other labels).

3. Results and Discussion

The application of Concept Induction to the SUN2012 dataset produces a set of interpretable neuron labels that aligned strongly with semantic concepts. Out of the 64 dense-layer neurons analyzed, 32

Table 1
Label confirmation via Google Images results with confirmed neurons (TLA % $\geq$ 80).

Neuron ID	ECII Concepts	Coverage Score	TLA %	Non-TLA %
0	snowy_mountain	0.986	95.00	52.57
7	snow, mountain	0.998	80.00	39.55
9	field	0.971	93.75	74.96
13	dish_rack	0.892	81.25	31.12
15	pillow, ceiling_fan	0.992	88.75	57.15
16	skyscraper, river_water	0.997	80.00	42.05
18	city	0.993	87.50	68.62
19	snowy_mountain	0.975	97.50	46.58
21	bedroom, duvet	0.967	85.00	47.67
22	dishwasher	0.945	96.20	24.58
23	fence	0.965	98.75	77.85
24	sink	0.939	96.25	52.59
26	toilet_tissue	0.979	95.00	39.49
27	toilet	0.961	95.00	55.33
28	cars	0.903	97.50	72.79
31	snowy_mountain	0.948	97.50	67.00
32	shell	0.950	80.00	46.64
33	bed, pillow	0.945	80.00	34.41
36	snowy_mountain	0.973	98.75	68.33
40	plant	0.972	83.75	51.00
41	pole, handrail	0.997	85.00	70.35
42	skyscraper	0.969	93.75	49.47
43	skyscraper	0.978	100.00	66.88
47	crosswalk	0.983	81.25	23.42
48	bidet	0.966	98.75	57.35
49	ironing_board, shoe	0.969	93.75	63.09
50	field	0.976	83.75	66.54
54	snowy_mountain	0.965	92.50	40.80
58	snowy_mountain	0.969	97.50	55.05
60	brocoli, cheese	0.979	93.75	39.00
61	cars	0.976	90.00	45.51
62	air_conditioning, chest_of_drawers	0.982	87.50	61.18

are confirmed to exhibit stable concept associations with Target Label Activations (TLA) of at least 80% (depicted in table 1). Of these 32 neurons, 29 neurons also show statistically significant separation between target and non-target activations under the Mann–Whitney U test ($p < 0.05$), confirming that their responses were meaningfully stronger for concept-related images. Table 2 shows the result of statistical test, where confirmed labels include crosswalk, skyscraper, pillow, ceiling fan, and bidet,

Compared to ADE20K, which yielded 19 confirmed neurons, SUN2012 produces 32 under the same evaluation procedure. This shows that despite dataset and architectural differences (ResNet50V2 for ADE20K and InceptionV3 for SUN2012), the Concept Induction framework reliably identifies semantically coherent neurons across benchmarks. Therefore, it provides fine-grained, human-readable, and verifiable neuron-level explanations, supporting transparent analysis, greater trust, and practical debugging of deep models.

Acknowledgments

The authors acknowledge partial funding through the Kansas State University Game-changing Research Initiation Program (GRIP).

Declaration on Generative AI

For this work, the authors used ChatGPT-5 for grammar and spelling checks. Subsequently, the authors reviewed and edited the content as needed and take full responsibility for the content.

Table 2
Statistical evaluation results. **Bold** rows represents neurons with p-value ≥ 0.05, where we **cannot reject the null hypothesis**.

Neuron ID	ECII Concepts	TLA %	Non-TLA %	Target Median	Non-Target Median	Target Mean	Non-Target Mean	z-score	p-value
0	snowy_mountain	95	53.02	5.99	0.23	5.58	1.22	-6.18	< 0.00001
7	snow, mountain	70	39.68	2.10	0.00	1.74	0.68	-3.04	0.00061
9	field	100	75.48	4.38	1.43	4.03	1.97	-3.64	0.00025
13	dish_rack	90	32.62	1.65	0.00	1.68	0.56	-4.66	< 0.00001
15	pillow, ceiling_fan	90	54.21	3.03	0.21	2.36	0.85	-4.40	< 0.00001
16	skyscraper, river_water	70	43.41	1.54	0.00	1.44	0.45	-3.15	0.00052
18	city	80	71.11	2.17	0.82	2.03	1.10	-2.85	0.00398
19	snowy_mountain	100	45.95	5.91	0.00	5.39	1.09	-6.55	< 0.00001
21	bedroom, duvet	65	45.95	0.86	0.00	1.91	0.57	-2.70	0.00332
22	dishwasher	80	26.11	1.60	0.00	1.87	0.29	-5.04	< 0.00001
23	fence	100	78.73	3.07	1.61	2.96	1.80	-3.40	0.00063
24	sink	95	53.25	3.14	0.11	3.14	0.81	-5.60	< 0.00001
26	toilet_tissue	100	38.65	2.47	0.00	2.36	0.51	-6.29	< 0.00001
27	toilet	95	56.98	4.06	0.33	3.88	0.94	-6.10	< 0.00001
28	**cars**	**95**	**72.78**	**1.58**	**1.03**	**1.66**	**1.59**	**-1.10**	**0.26788**
31	snowy_mountain	90	67.62	4.58	0.62	3.97	1.35	-5.05	< 0.00001
32	shell	85	46.19	1.65	0.00	1.55	0.70	-3.77	0.00004
33	bed, pillow	75	35.87	3.19	0.00	2.60	0.43	-4.62	< 0.00001
36	snowy_mountain	100	67.62	4.72	0.93	4.60	1.48	-6.24	< 0.00001
40	**plant**	**70**	**51.90**	**1.00**	**0.07**	**0.86**	**0.68**	**-1.44**	**0.12808**
41	pole, handrail	90	69.21	2.23	1.05	2.05	1.51	-2.18	0.02735
42	skyscraper	95	50.79	2.91	0.05	2.78	1.01	-4.49	< 0.00001
43	skyscraper	100	66.51	2.94	0.81	3.01	1.10	-5.65	< 0.00001
47	crosswalk	95	22.94	3.20	0.00	3.20	0.29	-6.74	< 0.00001
48	bidet	100	58.02	3.00	0.43	2.91	1.01	-5.35	< 0.00001
49	ironing_board, shoe	90	62.06	1.98	0.55	2.20	1.02	-3.37	0.00053
50	**field**	**85**	**66.83**	**0.98**	**0.73**	**1.15**	**1.11**	**-0.89**	**0.36257**
54	snowy_mountain	100	39.68	4.53	0.00	4.01	0.72	-6.58	< 0.00001
58	snowy_mountain	95	52.94	4.56	0.18	4.77	1.12	-6.33	< 0.00001
60	brocoli, cheese	95	39.52	1.79	0.00	1.93	0.50	-5.89	< 0.00001
61	cars	85	45.95	1.41	0.00	1.32	0.55	-4.07	< 0.00001
62	air_conditioning, chest_of_drawers	85	62.22	2.12	0.50	2.59	0.95	-3.92	0.00006

References

[1] A. Dalal, R. Rayan, A. Barua, E. Y. Vasserman, M. K. Sarker, P. Hitzler, On the value of labeled data and symbolic methods for hidden neuron activation analysis, in: International Conference on Neural-Symbolic Learning and Reasoning, Springer, 2024, pp. 109–131. doi:10.1007/978-3-031-71170-1_12.

[2] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, SUN database: Large-scale scene recognition from abbey to zoo, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3485–3492. doi:10.1109/CVPR.2010.5539970.

[3] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, Explainable artificial intelligence: an analytical review, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1424. doi:10.1002/widm.1424.

[4] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NeurIPS), volume 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[5] M. T. Ribeiro, S. Singh, C. Guestrin, "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, ACM, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[6] J. Lehmann, P. Hitzler, Concept learning in description logics using refinement operators, Machine Learning 78 (2010) 203–250. doi:10.1007/s10994-009-5146-2.

[7] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ADE20K dataset, International Journal of Computer Vision 127 (2019) 302–321. doi:10.1007/s11263-018-1140-0.

[8] S. E. Akkamahadevi, A. Dalal, P. Hitzler, Automating cnn neuron interpretation using concept induction, in: Proceedings of the ISWC, 2024, pp. 11–15.

[9] M. K. Sarker, P. Hitzler, Efficient concept induction for description logics, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, AAAI Press, 2019, pp. 3036–3043. doi:10.1609/aaai.v33i01.33013036.

[10] M. K. Sarker, J. Schwartz, P. Hitzler, L. Zhou, S. Nadella, B. S. Minnery, I. Juvina, M. L. Raymer, W. R. Aue, Wikipedia knowledge graph for explainable AI, in: Knowledge Graphs and Semantic Web Conference, KGSWC, 2020, Proceedings, volume 1232 of *Communications in Computer and Information Science*, Springer, 2020, pp. 72–87. doi:10.1007/978-3-030-65384-2_6.