Understanding CNN Hidden Neuron Activations using Concept Induction over Background Knowledge

Abhilekha Dalal[1]

[1]Kansas State University, Manhattan KS, USA

Abstract

A major challenge in Explainable AI is interpreting hidden neuron activations accurately. These interpretations can reveal what a deep learning system perceives as relevant in the input data, thereby addressing the black-box nature of such systems. The state of the art indicates that hidden node activations can be interpretable by humans, but there's a lack of systematic automated methods to verify these interpretations, especially those that utilize substantial background knowledge and inherently explainable methods. In this proposal, we introduce a novel model-agnostic post-hoc Explainable AI method based on a Wikipedia-derived concept hierarchy with approximately 2 million classes. Our approach utilizes OWL-reasoning-based Concept Induction for explanation generation and compares with off-the-shelf pre-trained multimodal-based explainable methods. Our results demonstrate that our method automatically provides meaningful class expressions as explanations to individual neurons in the dense layer of a Convolutional Neural Network, outperforming prior work in both quantitative and qualitative aspects.

Keywords

Explainable AI, Concept Induction, Convolutional Neural Network, Knowledge Graph,

1. Introduction

Deep learning has revolutionized various fields such as image classification [1], speech recognition [2], translation [3], drug design [4], medical diagnosis [5], climate sciences [6]. However, the opaque nature of deep learning systems poses challenges in applications involving automated decisions and safety-critical systems. For instance, concerns arise from incidents like Steve Wozniak's accusation of gender discrimination in Apple Card credit limits and biased image search results for "CEOs" [7]. Safety-critical areas like self-driving cars [8] and [9, 10] are also vulnerable to adversarial attacks [11], including altering classification results [11] and manipulating the order of training images [12]. Some attacks are hard to detect post facto, posing significant risks [13, 14].

Problem Statement: While statistical evaluations are standard for assessing deep learning performance, they fall short in providing explanations for specific system behaviors [15]. Therefore, developing robust explanation methods for deep learning systems remains crucial. Despite significant progress in this area (see Section 4), current approaches often rely on a limited set of predefined explanation categories. This reliance on human-selected categories is problematic, as it assumes they are suitable for explaining deep learning systems, which lacks evidence. Some methods leverage deep learning models, such as LLMs, to generate explanations [16], introducing another layer of opacity. Additionally, state-of-the-art explanation systems often require modified deep learning architectures, which can lead to reduced system performance compared to unmodified versions [17].

Importance: The importance of solving this challenge cannot be overstated. Transparent and interpretable AI systems are crucial for building trust, especially in domains like healthcare, finance, and autonomous vehicles. By providing explanations, we empower users, including non-experts, to understand AI decisions, fostering better acceptance and adoption. Advancing explainable AI contributes to interdisciplinary collaboration and can enhance societal benefits while mitigating ethical

Proceedings of the Doctoral Consortium at ISWC 2024, co-located with the 23rd International Semantic Web Conference (ISWC 2024)

adalal@ksu.edu (A. Dalal)

0000-0002-7047-5074 (A. Dalal)

risks associated with AI deployment. Therefore, it is imperative to address the challenge of developing transparent and interpretable explanation methods for deep learning systems.

The subsequent section presents the research question and objectives, building on the above core principles. 2.1 describe the contributions we have made, focusing on methods we use or plan to use to support these contributions and then describing the results 3 thus far from them.

2. Research Question and Contributions

Research Question: How can we develop an effective approach to explainable deep learning that can be used to assign human-understandable interpretations to the activations of hidden neurons in the deep learning model?

This proposal outlines an approach to use *Concept Induction*, i.e., formal logical deductive reasoning [18] to automatically provide meaningful explanations for hidden neuron activation in a Convolutional Neural Network (CNN) architecture for image scene classification (on the ADE20K dataset [19]), using a class hierarchy consisting of about $2 \cdot 10^6$ classes, derived from Wikipedia, as the pool of categories [20]. Stating the hypothesis clearly that drives the work outlined in this proposal.

Hypothesis: Concept Induction analysis with large-scale background knowledge yields meaningful labels that stably explain neuron activation in the hidden layer of CNN architecture.

2.1. Contributions and Methodology

To achieve the above-stated hypothesis, the following objectives with the methodology followed or planned to follow are outlined:

Objective 1: Employing Concept Induction and a Wikipedia Knowledge Graph to Assign Meaningful Labels to Hidden Neurons' Activation.

We explored and evaluated three concrete methods (Concept Induction, CLIP-Dissect [16], GPT-4 [21]) to generate high-level concepts for explaining hidden neuron activations. Our comprehensive methodology for Objective 1 is detailed in our paper [22].

1. **Prep: Scenario and CNN Training** - Utilizing the annotated ADE20K dataset [19], we trained Resnet50V2 for scene classification, achieving an accuracy of (**86.46%**). *The annotations are only used for generating label hypotheses, not for CNN training.* While highest accuracy isn't critical for our investigation, it's important for models to be practically applicable.
2. **Concept Induction** - [18] system accepts three inputs: positive set P and negative set N of images from ADE20K, and a knowledge base 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 $q \in N$. For scalability, we used ECII [23] heuristic Concept Induction system with Wikipedia [20]. We included the images in the background knowledge by associating object annotations from ADE20K images with classes in the hierarchy, using the Levenshtein string similarity metric [24] with edit distance 0.
3. **Generating Label Hypotheses** -
    a) In Concept Induction, we used 1,370 ADE20K images with our trained ResNet50V2, extracting activations from the dense layer with 64 neurons. Positive examples (P) are images activating the neuron with $>$ 80% of its max activation, negative examples (N) are those activating it with $<$ 20% of its max or not at all. ECII generates the target label for each neuron based on these sets and background knowledge.
    b) **CLIP-Dissect** employs the top 20,000 English vocabulary words as concepts. Subsequently, activations from our trained ResNet50v2 model for ADE20K test images were collected, resulting in a matrix (Number of Images $\times$ 64). Utilizing these inputs, CLIP-Dissect assigns a label to each neuron such that the neuron is most activated when the corresponding concept is present in the image, resulting in 22 distinct concepts across 64 neurons.

c) **GPT-4** Leveraging GPT-4, we adopt a methodology akin to [25] for concept generation to differentiate image classes [26]. We input image annotations from positive (P) and negative (N) sets into GPT-4 with prompts to discern concepts unique to P. The prompt "Generate top three classes of objects/general scenarios that better represent what images in the positive set (P) have but the images in the negative set (N) do not," yields three concepts per neuron, from which we select one per class for assessment.

Objective 2: Automate Concept Label Association for Input Images using Neuron Ensembles and Non-target Activation Probabilities.

1. **Concept Associations and Non-Target Activations** - In pursuit of Objective 1, Step 3 generates labels for neuron activation. Each neuron's label is the target concept, with all other images considered as non-target concepts. This analysis focuses on the top three ECII responses, assessing neuron activation for non-target concepts at various cut-off values relative to each neuron's maximum activation value: > 0, $> 20\%$ of max, $> 40\%$ of max, and $> 60\%$ of max. The goal is to establish strong associations between concepts and neuron activations, understanding which concepts trigger specific neurons and to what extent.
2. **Neuron Ensembles for Concept Associations** - Input information can be distributed across simultaneously activated neurons, necessitating the examination of neuron ensemble activations using previously established cut-off values. However, the scale challenge arises with 2^{64} potential neuron ensembles for just 64 neurons. To address this, we propose combining neurons activated for semantically related labels (with top-3 responses from ECII). For instance, if "building" activates both neuron 0 and neuron 63. We assess all images activating both neurons 0 and 63 for specified cut-off values. In cases where a concept activates more than two neurons, our analysis encompasses all possible combinations of pairs, evaluating target and non-target activations. We proceed with concepts, including neuron ensembles, that exhibit target activation exceeding 80% for further analysis
3. **Validating Neuron-Concept Associations** - After completing Step 1 and Step 2, we obtain probabilities for non-target concepts across all concepts, including those activating single neurons as well as neuron ensembles. This allows for identifying potential concepts and assessing associated error margins. To verify or reject these concepts, we revisit the ADE20K dataset. Using a subset of 1050 randomly chosen images, we conduct a user study via Amazon Mechanical Turk (MTurk) [27] to annotate images with target concepts. We then cross-reference these designated concepts with image annotations obtained from the MTurk study. We evaluate the likelihood of neuron activations for non-target concepts.
4. **Developing an Automated System** - We propose developing an automated system to streamline the entire process, enabling scalability to larger datasets and exploration of a broader parameter range. The system would comprise: *Concept induction*: Generates class expressions/responses ranked by coverage score. *Neuron activation*: Calculates activation for target and non-target concepts (including neuron ensembles) at various cut-off values. *Concept validation*: Validates generated concepts. This automated system would analyze new images, generating a list of potential concepts with associated probabilities. Users could review the concepts and select the most relevant ones for the image. The automated approach offers several advantages, including speed, efficiency, scalability to larger datasets, and exploration of diverse parameter settings.

3. Evaluation and Results

Objective 1: The three approaches generate label hypotheses for all studied neurons, which we validated using new images. We search Google Images using each target label as keywords and collect 200 images per label with Imageye[1]. These images are split into 80% for evaluation and 20% for statistical analysis. We then determine if the target neuron activates when the retrieval label matches the target

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

Table 1
Generated label hypotheses from all three approaches,**Bold** denotes neurons whose labels are considered confirmed(the full version can be found in our work at [22]).

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**
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**
7	**bathroom**	**153**		**93.333**	**44.113**
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**

Table 2
Statistical Evaluation details for all three approaches(full version can be found in our work at [22]).

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
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
31	Urban Street Scene	41	80.50	65.70	1.80	1.30	1.70	0.90	-2.4	0.164

label and if any other neurons activate. Table 1(presents selective representation due to space constraints, complete version is available at [22].) show the percentage of target images that activated each neuron. A target label is confirmed if it activates for $\geq 80\%$ of its target images, regardless of its activation for non-target images. Detailed paper can be found at [22].

Statistical Evaluation and Result:- After generating confirmed labels from all three approaches, we assess node labeling using the remaining images, treating each neuron-label pair in Table 1 as a hypothesis. Concept Induction, CLIP-Dissect, and GPT-4 produce 20, 8, and 27 hypotheses, respectively, based on confirmed labels. Using the Mann-Whitney U test, we compared activation strengths between images retrieved using the target label and those retrieved using other keywords. Table 2 shows the selective representation of results obtained through Mann-Whitney U test. Concept Induction consistently outperforms other methods, as evidenced by Mann-Whitney U results and statistical analysis. For most neurons, activation values of target images significantly exceed those of non-target images (with $p < 0.00001$). Concept Induction rejects 19 out of 20 null hypotheses at $p < 0.05$, CLIP-Dissect rejects all 8 null hypotheses, and GPT-4 rejects 25 out of 27 null hypotheses at $p < 0.05$. More details in [22].

Objective 2: We will conduct a comprehensive statistical evaluation using the Mann-Whitney U (MWU) test for each concept across different cut-off values. This evaluation aims to compare the activation strengths of non-target concepts retrieved through Google Images(from Objective 1) with those retrieved from the ADE20K dataset. The hypothesis under consideration is that the activation strength of non-target concepts from Google Images exceeds that from the ADE20K dataset. Conversely, the null hypothesis (H0) posits that the activation strength of non-target concepts from Google Images

equals that from the ADE20K dataset. For each category of cut-off values, concepts exhibiting a significant difference in activation strengths (p-value < 0.005) will undergo further validation through the Wilcoxon signed-rank test across all cut-off values as a collective unit. We refine our approach and enhance concept label associations' accuracy by identifying concepts with significantly higher activation strengths.

4. Related Work

With the recent advances in deep learning [28], its wide usage in nearly every field, and its opaque nature make explainable AI more important than ever, and there are multiple ongoing efforts to demystify deep learning [29, 30, 31]. Existing explainable methods can be categorized based on input data (feature) understanding, e.g., feature summarizing [32, 33], or based on the model's internal unit representation, e.g., node summarizing [34, 11]. Those methods can be further categorized as model-specific [32] or model-agnostic [33]. Another kind of approach relies on human interpretation of explanatory data returned, such as counterfactual questions [35].

We focus on the understanding of internal units of the neural network-based deep learning models. Prior work has shown that internal units may indeed represent human-understandable concepts [34, 11], but these approaches often require resource-intensive methods like semantic segmentation [36] or explicit concept annotations [37]. There has been research utilizing Semantic Web data for explaining deep learning models [38, 39], and Concept Induction for generating explanations [40, 41]. However, they mainly focused on analyzing how inputs relate to outputs and generating explanations for the whole system, while we focused on understanding internal node activations.

CLIP-Dissect [16], similar to our work, takes a different approach. It utilizes the CLIP pre-trained model, employing zero-shot learning to associate images with labels. Another related work, Label-Free Concept Bottleneck Models [26], builds upon CLIP-Dissect, using GPT-4 [21] for concept set generation. However, CLIP-Dissect faces challenges in accurately predicting output labels based on concepts in the last hidden layer and transferring to other modalities or domain-specific applications. The Label-Free approach inherits these limitations and may compromise explainability due to its use of a concept derivation method that lacks inherent explainability.

5. Conclusion

Concept Induction, leveraging large-scale ontological background knowledge, provides meaningful labeling of hidden neuron activations, validated by statistical analysis. This allows us to pinpoint concepts that strongly trigger neuron responses, effectively explaining neuron activations. Our approach introduces novel possibilities for diverse label categories. Comparative analysis against CLIP-Dissect and GPT-4 showcases Concept Induction's superiority, especially in settings with labeled data. Ultimately, our work aims to thoroughly analyze hidden layers in deep learning systems, facilitating the interpretation of activations as implicit input features and explaining system input-output behavior. Moving forward, future work will focus on enhancing Concept Induction's scalability and efficiency, enabling its broader applicability across various domains.

Acknowledgments

The author acknowledge advisor Dr. Pascal Hitzler and partial funding under National Science Foundation grants 2119753 and 2333782.

References

[1] M. Ramprasath, M. V. Anand, S. Hariharan, Image classification using convolutional neural networks, International Journal of Pure and Applied Mathematics 119 (2018) 1307–1319.

[2] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in: International conference on machine learning, The Proceedings of Machine Learning Research, 2014, pp. 1764–1772.

[3] M. Auli, M. Galley, C. Quirk, G. Zweig, Joint language and translation modeling with recurrent neural networks, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1044–1054.

[4] M. H. Segler, T. Kogej, C. Tyrchan, M. P. Waller, Generating focused molecule libraries for drug discovery with recurrent neural networks, ACS central science 4 (2018) 120–131.

[5] H.-I. Choi, S.-K. Jung, S.-H. Baek, W. H. Lim, S.-J. Ahn, I.-H. Yang, T.-W. Kim, Artificial intelligent model with neural network machine learning for the diagnosis of orthognathic surgery, Journal of Craniofacial Surgery 30 (2019) 1986–1989.

[6] Y. Liu, E. Racah, Prabhat, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. F. Wehner, W. D. Collins, Application of deep convolutional neural networks for detecting extreme weather in climate datasets, 2016. URL: http://arxiv.org/abs/1605.01156. arXiv:1605.01156.

[7] I. A. Hamilton, Apple cofounder Steve Wozniak says Apple Card offered his wife a lower credit limit, Business Insider (2019).

[8] Z. Chen, X. Huang, End-to-end learning for lane keeping of self-driving cars, in: 2017 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2017, pp. 1856–1860.

[9] A. S. Rifaioglu, E. Nalbat, V. Atalay, M. J. Martin, R. Cetin-Atalay, T. Doğan, Deepscreen: high performance drug–target interaction prediction with convolutional neural networks using 2-d structural compound representations, Chemical science 11 (2020) 2531–2557.

[10] W. Hariri, A. Narin, Deep neural networks for covid-19 detection and diagnosis using images and acoustic-based techniques: a recent review, Soft computing 25 (2021) 15345–15362.

[11] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, A. Torralba, Understanding the role of individual units in a deep neural network, Proceedings of the National Academy of Sciences 117 (2020) 30071–30078.

[12] I. Shumailov, Z. Shumaylov, D. Kazhdan, Y. Zhao, N. Papernot, M. A. Erdogdu, R. J. Anderson, Manipulating SGD with data ordering attacks, in: Advances in Neural Information Processing Systems (NeurIPS), volume 34, 2021, pp. 18021–18032.

[13] S. Goldwasser, M. P. Kim, V. Vaikuntanathan, O. Zamir, Planting undetectable backdoors in machine learning models: [extended abstract], in: IEEE Annual Symposium on Foundations of Computer Science (FOCS), 2022, pp. 931–942. doi:10.1109/FOCS54457.2022.00092.

[14] T. Clifford, I. Shumailov, Y. Zhao, R. Anderson, R. Mullins, ImpNet: Imperceptible and blackbox-undetectable backdoors in compiled neural networks, 2022. arXiv:2210.00108.

[15] D. Doran, S. Schulz, T. R. Besold, What does explainable AI really mean? a new conceptualization of perspectives, in: CEUR Workshop Proceedings, volume 2071, CEUR, 2018.

[16] T. Oikarinen, T.-W. Weng, CLIP-Dissect: Automatic description of neuron representations in deep vision networks, in: International Conference on Learning Representations, ICLR, 2023. URL: https://openreview.net/forum?id=iPWiwWHc1V.

[17] M. E. Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, P. Lió, M. Jamnik, Concept embedding models: Beyond the accuracy-explainability trade-off, in: Advances in Neural Information Processing Systems (NeurIPS), 2022.

[18] J. Lehmann, P. Hitzler, Concept learning in description logics using refinement operators, Mach. Learn. 78 (2010) 203–250. URL: https://doi.org/10.1007/s10994-009-5146-2. doi:10.1007/s10994-009-5146-2.

[19] 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.

[20] 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: B. Villazón-Terrazas, F. Ortiz-Rodríguez, S. M. Tiwari, S. K. Shandilya (Eds.), Proceedings of the Knowledge Graphs and Semantic Web Second Iberoamerican Conference and First Indo-American Conference (KGSWC), volume 1232 of *Communications in Computer and Information Science*, Springer, 2020, pp. 72–87. URL: https://doi.org/10.1007/978-3-030-65384-2_6. doi:10.1007/978-3-030-65384-2_6.

[21] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[22] 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, 2024. arXiv:2404.13567.

[23] M. K. Sarker, P. Hitzler, 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), AAAI Press, 2019, pp. 3036–3043. URL: https://doi.org/10.1609/aaai.v33i01.33013036. doi:10.1609/aaai.v33i01.33013036.

[24] V. I. Levenshtein, On the minimal redundancy of binary error-correcting codes, Inf. Control. 28 (1975) 268–291. URL: https://doi.org/10.1016/S0019-9958(75)90300-9. doi:10.1016/S0019-9958(75)90300-9.

[25] A. Barua, C. Widmer, P. Hitzler, Concept induction using LLMs: a user experiment for assessment, 2024. URL: https://arxiv.org/abs/2404.11875. arXiv:2404.11875.

[26] T. Oikarinen, S. Das, L. M. Nguyen, T.-W. Weng, Label-free concept bottleneck models, in: The Eleventh International Conference on Learning Representations, ICLR, 2023. URL: https://openreview.net/forum?id=FlCg47MNvBA.

[27] K. Crowston, Amazon mechanical turk: A research tool for organizations and information systems scholars, in: A. Bhattacherjee, B. Fitzgerald (Eds.), Shaping the Future of ICT Research. Methods and Approaches, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 210–221.

[28] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[29] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI – explainable artificial intelligence, Science robotics 4 (2019) eaay7120.

[30] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[31] D. Minh, H. X. Wang, Y. F. Li, T. N. Nguyen, Explainable artificial intelligence: a comprehensive review, Artificial Intelligence Review (2022) 1–66.

[32] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: Why did you say that?, 2016. URL: http://arxiv.org/abs/1611.07450. arXiv:1611.07450.

[33] 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, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[34] B. Zhou, D. Bau, A. Oliva, A. Torralba, Interpreting deep visual representations via network dissection, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 2131–2145.

[35] S. Wachter, B. D. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, CoRR abs/1711.00399 (2017). URL: http://arxiv.org/abs/1711.00399. arXiv:1711.00399.

[36] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 418–434.

[37] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: J. G. Dy, A. Krause (Eds.), Proceedings of the International Conference on Machine Learning (ICML), volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2673–2682. URL: http://proceedings.mlr.press/v80/kim18d.html.

[38] R. Confalonieri, T. Weyde, T. R. Besold, F. M. del Prado Martín, Using ontologies to enhance human

understandability of global post-hoc explanations of black-box models, Artificial Intelligence 296 (2021) 103471.

[39] N. Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, F. Herrera, 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 (2022) 58–83.

[40] M. K. Sarker, N. Xie, D. Doran, M. L. Raymer, P. Hitzler, Explaining trained neural networks with semantic web technologies: First steps, in: T. R. Besold, A. S. d'Avila Garcez, I. Noble (Eds.), Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy), volume 2003 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-2003/NeSy17_paper4.pdf.

[41] T. Procko, T. Elvira, O. Ochoa, N. D. Rio, An exploration of explainable machine learning using semantic web technology, in: 2022 IEEE 16th International Conference on Semantic Computing (ICSC), IEEE Computer Society, 2022, pp. 143–146. doi:10.1109/ICSC52841.2022.00029.