Concept Induction Using LLMs

Adrita Barua[1]

[1]Kansas State University, Manhattan, KS, USA

Abstract

In this study, the capability of Large Language Models (LLMs) is explored to automate Concept Induction, a process traditionally reliant on formal logical reasoning using description logic ontologies, within the context of explainable AI (XAI). Initially, a pre-trained LLM like GPT-4 is employed to assess its ability to generate high-level concepts describing data differentials for a scene classification task via prompting. A human assessment study was conducted which revealed that concepts produced by GPT-4 are preferred over those from logical concept induction systems in terms of human understandability, despite some limitations in neuron activation analysis. Building on these insights, further research aims to automate the concept induction system using LLMs, potentially addressing the shortcomings of traditional logical reasoners. This approach has the potential to scale and provide a significant avenue for concept discovery in complex AI models.

Keywords

Concept Induction, LLM, Explainable AI, GPT-4

1. Problem statement

Concept Induction [1, 2] is a symbolic reasoning task that involves generating complex class descriptions from instance examples using deductive reasoning algorithms over Description Logic knowledge bases. It can be used to depict meaningful explanations by identifying patterns from complex data. Previous studies [3, 4] have explored the potential of using concept induction in the context of explainable AI (XAI) to provide human-understandable explanations of machine learning classifications. However, traditional concept induction systems face several limitations in adaptability, scalability, and capturing complex data relationships due to their reliance on predefined rules and limited background knowledge and may not capture the full scope of human-like reasoning. In contrast, research in XAI aims to improve the understandability of AI models without compromising accuracy [5]. Current techniques often rely on post-hoc algorithms [6], which encounter challenges like visualization and adversarial attacks [7]. Concept-based models [8, 9] offers a promising alternative by incorporating explicit representations of concepts aligned with human intuition to explain the model's behavior. However, generating context-specific meaningful concepts from complex data remains challenging. This research aims to explore the feasibility of replacing conventional concept induction systems with Large Language Models (LLMs) to overcome these limitations and enhance the interpretability of AI models.

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

adrita@ksu.edu (A. Barua)

0000-0002-3287-7443 (A. Barua)

2. Importance

Concept induction plays a crucial role in various domains including XAI, enabling the generation of interpretable and meaningful insights from complex data. Transitioning to LLM-based methods for concept induction can improve symbolic reasoning tasks at scale across different domains such as information retrieval, knowledge extraction, etc. Furthermore, automating concept discovery through LLMs can make black box models more explainable, aligning with the ongoing efforts to map network activations with meaningful explanations [10]. This addresses critical issues of transparency and trust in AI decisions, crucial for stakeholders across industries impacted by AI. The significance of this work extends to the broader AI community by potentially advancing neurosymbolic AI, bridging the gap between traditional AI and symbolic reasoning approaches. We project that the outcomes of this research can contribute to overcoming the limitations of symbolic concept induction systems and contribute to advancing XAI techniques, enabling safer and more accountable AI systems.

3. Related work

There are different approaches that utilizes traditional concept induction systems using provably correct [1] or heuristic [11] deduction algorithms over description logic knowledge bases. Various applications [12] stand to benefit from a concept induction system that is not constrained by background knowledge and predefined rules. In the context of XAI, concept induction has shown significant results to generate human-understandable explanations through post-hoc analysis of input data to explain the machine learning classification outputs [4, 13]. However, these methods are limited by their reliance on background knowledge and heuristic nature of explanation generation, potentially overlooking common-sense interpretations that are evident to humans. Leveraging LLMs has the potential to bridge this gap by automating higher-level concept generation by utilizing minimal text-based information. Methods like TCAV [8] focus on global explanations by employing high-level concepts to estimate their importance for predictions, but relys on human-provided concepts. Alternatively, ACE [14] leverages image segmentation and clustering to curate automated concepts that may result in some information loss. Other approaches, such as Concept Bottleneck Models (CBM) [15] and Post-hoc CBM [16], map DNN models to human-understandable concepts but often depend on hand-picked concepts, highlighting the need for automated methods to generate higher-level concepts. Another study [17] employing a similar approach utilizes GPT-3 with a few-shot method to produce automated concepts. But none of these methods cater to the generation of complex description logic concepts. Our study delves into LLMs' ability to generate such explanations that can replace the symbolic reasoners at scale, to be used as a stand alone system.

4. Research question(s) and hypotheses

The objective of our research is to assess whether LLMs, leveraging their vast domain knowledge and reasoning capabilities, can outperform or at least match concept induction systems in producing accurate and understandable explanations aligned with human intuition, while also

being capable of explaining hidden neuron activations in the domain of XAI. Previous research [4] has explored the effectiveness of concept induction for creating explanations that ”make sense” to humans, indicating that while concept induction can explain data differentials in machine learning classifications, human-generated explanations are generally superior. This work employed the ECII heuristic concept induction system [11] and utilized the Wikipedia category hierarchy [18] as background knowledge. Building upon their findings, our study extends their work by replacing the ECII model with an LLM to generate meaningful and coherent explanations. Primarily, we seek to identify ”good” concepts that are understandable to humans and evaluate their alignment with human-generated explanations to potentially surpass concept induction in terms of accuracy and comprehensibility. Furthermore, we seek to understand if LLM explanations are preferred over logical concept induction systems in terms of ”meaningfulness to humans”, whether they will still remain effective in demonstrating neuron activations when mapped to a neural network architecture. There could be a trade-off between the two approaches; for example, the type of concepts that work well for humans might not always be useful to depict what the neuron 'sees' in a DNN architecture. The primary goal is to utilize pre-trained LLMs like GPT-4 [19] to achieve satisfactory results via prompting [20] and subsequently fine-tune an LLM to mimic the output of a symbolic reasoner (e.g., generating complex concepts) that could be verifiable using description logics while making use of the common-sense capability of LLMs.

5. Research methods

We begin by employing an initial prompting technique to asses the effectiveness of concepts generated by LLMs in terms of human understandability and their applicability to hidden neuron activation. This initial assessment serves as a foundation for our broader objective of fine-tuning it further to automate the system of concept induction.

Prompting method In preliminary investigations [21], we utilized GPT-4 to generate concepts for distinguishing between different image classes as an initial assessment of LLM's concept induction capability. Object tags from the ADE20K dataset [22, 23] were used as input for the GPT-4 model via the OpenAI API, using zero-shot prompting. This dataset comprises around 20,000 images annotated with scene categories and object tags. We selected 45 image set pairs, each containing two groups of images representing distinct scene categories(e.g., Bathroom vs Park). Our objective was to generate explanations that describe what distinguished category A from category B in each image set pair. To prompt the GPT-4 model effectively, we experimented with different techniques, ultimately leveraging only the object labels from each image set category. The model was instructed to differentiate between the two categories based on their object tags. The generated concepts were compared with those produced by the ECII system, which also used the same object tags. Object tags can be any items physically present in the images, such as stands, food, walls, etc. The process and the prompt used for interacting with the GPT-4 model are illustrated in Figure 1. The latest version of the GPT-4 model was used with specific parameter settings, including a temperature of 0.5 and top_p of 1, to ensure consistent and reproducible answers. We came up with the specific prompt(1) to generate

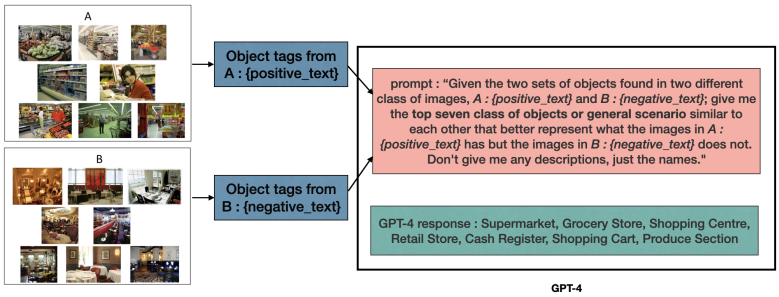

Figure 1: Prompting Method

generic concepts or object classes that mimic the ontology classes positioned somewhere in the middle of the hierarchy used by ECII, aiming to provide a balance between more general concepts and highly specific subclasses within the ontology structure. Each set produced a list of seven concepts following this method. A detailed description of the experimental setup and prompting method can be found in [21].

Hidden neuron activation analysis To evaluate whether concepts generated by LLMs can offer insights into hidden layer activation space, we do a preliminary investigation mentioned in [10], using two evaluation methods: Statistical Evaluation and Concept Activation Analysis. In this work, three approaches are compared for generating concepts: GPT-4, ECII, and CLIP-Dissect [9]. To begin, label hypotheses are obtained to determine which neurons are activated for specific concept labels. Initially, a trained ResNet50V2 is fed with ADE20K images, and the activations of the dense layer's 64 neurons are analyzed individually. For each neuron, positive examples (P) consist of images that activate the neuron with at least 80% of the maximum activation value, while negative examples (N) are images that activate the neuron with at most 20% or not at all. ECII generates concept-label hypotheses for each neuron based on P, N, and background knowledge. Similarly, GPT-4 uses the same sets P and N but with adjustments. Due to input constraints, only one image per class is selected for set N. GPT-4 identifies concepts present in P but not in N, using a prompting method described earlier in this section. This yields a list of three concepts per neuron, but only one concept per neuron is chosen at random for the analysis. To compare with other XAI methods, target labels are also generated using CLIP-Dissect, a label-free method that associates high-level concepts with individual neurons using a pre-trained multimodal model. To confirm these label hypotheses, images corresponding to each concept-label are retrieved from Google Images using the label as a keyword. 80% of the obtained images are used for hypothesis confirmation, and the remaining 20% for statistical evaluation. The images are fed to the network to check if the target neuron activates for the retrieved label and if any other neurons activate. A target label for a neuron is confirmed if it activates for $\geq$ 80% of its target images. In total, 19, 5, and 14 distinct confirmed concepts are obtained from Concept Induction, CLIP-Dissect, and GPT-4, respectively.

Fine-tuning an LLM After reviewing the initial results generated from the prompting technique, our next step is to fine-tune an open-sourced LLM, to automatically generate meaningful

concepts based on input data, that can effectively explain the reasoning behind specific outputs from neural network architectures. We want to fine-tune the LLM in a manner that captures the logical reasoning structure of a symbolic deductive system, ensuring it remains both explainable and verifiable. This approach aims to address the challenge of using another black box model, like an LLM, to explain a neural network system, while also mitigating the uncontrolled nature of a generic LLM by providing a more controlled system for concept generation.

6. Evaluation

Human Assessment We conducted a human assessment study on Amazon Mechanical Turk using the Cloud Research platform to evaluate the human understandability of concepts generated from GPT-4. 300 participants were recruited, with each compensated $5 for completing the task. The study aimed to evaluate the quality of explanations generated by LLM (GPT-4) compared to human-generated and ECII explanations. Participants were presented with 45 image set pairs and asked to choose the more accurate explanation among three types: human vs. ECII, human vs. LLM, and LLM vs. ECII. Each participant completed all three comparisons, with only two explanation types compared in any given question. Human and ECII explanations were crafted in the previous study [4], while LLM explanations were generated following the prompting method specified in section 5. Participants preferred human explanations over ECII explanations (83% preference) and LLM (GPT-4) explanations (69% preference). However, LLM explanations were preferred over ECII explanations (63% preference). Ability scores derived from Bradley-Terry analysis revealed that human explanations had the highest scores (M = 1.77), followed by LLM explanations (M = 0.724), with a significant overall difference ($p < 0.001$, $\eta^2 = 0.41$). Tukey's Honestly Significant Difference (HSD) test confirmed significant differences in ability scores between human vs. ECII explanations, human vs. LLM explanations (both $p < 0.0001$), as well as between LLM vs. ECII explanations ($p = 0.0004$). It indicates that the observed differences in ability scores are highly significant. Detailed ability scores for each image set pair and a discussion of the nature of the resulting concepts can be found in [21].

Statistical Evaluation To do a statistical analysis on the confirmed labels generated in hidden neuron activation method described in section 5, we consider each neuron-label pair as a hypothesis, using the remaining 20% images retrieved from Google Images. For example, the hypothesis for neuron 1 is that it activates more strongly for images related to "crosswalk" than for images related to other keywords. The corresponding null hypothesis is that activation values are not different. We test 20 hypotheses from Concept Induction, 8 from CLIP-Dissect, and 27 from GPT-4. Since activation values may not follow a normal distribution, we use the Mann-Whitney U test [24] for statistical assessment. Among the 20 null hypotheses from Concept Induction, 19 are rejected at $p < 0.05$. For CLIP-Dissect, all 8 null hypotheses are rejected are rejected at $p < 0.05$, and for GPT-4, 25 out of 27 null hypotheses are rejected. Considering unique concepts, Concept Induction validates 18 hypotheses statistically, CLIP-Dissect validates 5, and GPT-4 validates 12. Mann-Whitney U results demonstrate that for most neurons (with $p < 0.00001$), activation values of target images are significantly higher than those of non-target images.

Concept Activation Analysis We utilize Concept Activation [25, 26], an XAI technique that measures the presence of predefined concepts in hidden-layer activations of images. We evaluate label hypotheses obtained from all three methods using this analysis, and unlike previous methods, this analysis doesn't restrict itself to confirmed concepts. Images for each concept are collected from Google, and a concept classifier is trained using a Support Vector Machine (SVM). The dataset for each classifier consists of images showing the presence (label=1) and absence (label=0) of the concept. This dataset is passed through a pre-trained ResNet50V2 model, and the activation values of each image in the dense layer are saved. The transformed dataset is split into train (80%) and test (20%) datasets, and an SVM classifier is trained using the train split. Both linear (Concept Activation Vector, CAV) and non-linear (Concept Activation Region, CAR) kernels are used to assess the decision boundary separating the presence/absence of a concept. Finally, the test dataset is used to evaluate the concept classifier's ability to classify the existence of concepts. All concepts analyzed using Concept Activation achieved a p-value of less than 0.05 in k-fold cross-validation tests. CLIP-Dissect outperformed GPT-4 on CAR, and Concept Induction surpassed GPT-4 on CAV. However, there was no statistically significant difference between Concept Induction and CLIP-Dissect. A detailed result and discussion of both the neuron activation analysis can be found in [10].

7. Limitations and future work

The human assessment study of concepts generated by LLMs such as GPT-4 has shown that they have great potential in automating the system for concept induction to provide meaningful insights into data differentials. However, the evaluation using hidden neuron activation methods did not yield promising results. It is understandable as the evaluation method of neuron activations has its own constraints (e.g., verification using the Google image dataset can have anomalies and does not always depict the accurate concepts that are originally true to the neuron) and is still under development. Despite these limitations, there is room for improvement in LLM's concept generation pipeline to better align with the nature of activated neurons. Efforts to fully automate XAI systems for concept discovery within DNN are crucial and further refinement of LLM-based approaches is necessary. While challenges persist, LLMs demonstrate the capacity to produce human-understandable high-level concepts. Developing standalone systems by fine-tuning LLMs to leverage their common sense capabilities could potentially replace traditional Concept Induction systems at scale, offering significant value across various domains, including XAI. This study underscores the efficient utilization of LLMs in Concept Induction and paves the way for future research to harness these models to enhance the explainability of AI systems.

Acknowledgments This research acknowledges Dr. Pascal Hitzler, professor in the Department of Computer Science, Kansas State University, director of the Data Semantics (DaSe) Lab, for his supervision and guidance throughout this study. The study received partial funding from the National Science Foundation grant 2333782 "Proto-OKN Theme 1: Safe Agricultural Products and Water Graph (SAWGraph): An OKN to Monitor and Trace PFAS and Other Contaminants in the Nation's Food and Water Systems."

References

[1] 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.

[2] L. Bühmann, J. Lehmann, P. Westphal, Dl-learner - A framework for inductive learning on the semantic web, J. Web Semant. 39 (2016) 15–24. URL: https://doi.org/10.1016/j.websem.2016.06.001. doi:10.1016/J.WEBSEM.2016.06.001.

[3] 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 2017, London, UK, July 17-18, 2017, volume 2003 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-2003/NeSy17_paper4.pdf.

[4] C. L. Widmer, M. K. Sarker, S. Nadella, J. Fiechter, I. Juvina, B. Minnery, P. Hitzler, J. Schwartz, M. Raymer, Towards human-compatible XAI: Explaining data differentials with concept induction over background knowledge, Journal of Web Semantics 79 (2023) 100807.

[5] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable AI: A brief survey on history, research areas, approaches and challenges, in: J. Tang, M. Kan, D. Zhao, S. Li, H. Zan (Eds.), Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II, volume 11839 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 563–574. URL: https://doi.org/10.1007/978-3-030-32236-6_51.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information fusion 58 (2020) 82–115.

[7] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, arXiv preprint arXiv:2006.11371 (2020).

[8] 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 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2673–2682. URL: http://proceedings.mlr.press/v80/kim18d.html.

[9] T. P. Oikarinen, T. Weng, Clip-dissect: Automatic description of neuron representations in deep vision networks, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/pdf?id=iPWiwWHc1V.

[10] 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.

[11] M. K. Sarker, P. Hitzler, Efficient concept induction for description logics, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 3036–3043. URL: https://doi.org/10.

1609/aaai.v33i01.33013036.

[12] T. Procko, T. Elvira, O. Ochoa, N. D. Rio, An exploration of explainable machine learning using semantic web technology, in: 16th IEEE International Conference on Semantic Computing, ICSC 2022, Laguna Hills, CA, USA, January 26-28, 2022, IEEE, 2022, pp. 143–146. URL: https://doi.org/10.1109/ICSC52841.2022.00029. doi:10.1109/ICSC52841.2022.00029.

[13] 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.

[14] A. Ghorbani, J. Wexler, J. Y. Zou, B. Kim, Towards automatic concept-based explanations, Advances in neural information processing systems 32 (2019).

[15] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 5338–5348. URL: http://proceedings.mlr.press/v119/koh20a.html.

[16] M. Yüksekgönül, M. Wang, J. Zou, Post-hoc Concept Bottleneck Models, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/pdf?id=nA5AZ8CEyow.

[17] T. P. Oikarinen, S. Das, L. M. Nguyen, T. Weng, Label-free concept bottleneck models, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/pdf?id=FlCg47MNvBA.

[18] 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.), Knowledge Graphs and Semantic Web - Second Iberoamerican Conference and First Indo-American Conference, KGSWC 2020, Mérida, Mexico, November 26-27, 2020, Proceedings, 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.

[19] 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).

[20] S. Ekin, Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices, Authorea Preprints (2023). URL: https://www.techrxiv.org/doi/full/10.36227/techrxiv.22683919.v2.

[21] 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, submitted to NeSy 2024.

[22] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ADE20K dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 5122–5130. URL: https://doi.org/10.1109/CVPR.2017.544.

[23] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic under-

standing of scenes through the ADE20k dataset, International Journal of Computer Vision 127 (2019) 302–321.

[24] P. E. McKnight, J. Najab, Mann-whitney u test, in: The Corsini Encyclopedia of Psychology, Wiley, 2010.

[25] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2668–2677. URL: https://proceedings.mlr.press/v80/kim18d.html.

[26] J. Crabbé, M. van der Schaar, Concept activation regions: A generalized framework for concept-based explanations, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 2590–2607.